Metagenomes of Red Sea subpopulations challenge the use of marker genes and morphology to assess Trichodesmium diversity

Scripts for the Analysis of the Figures:

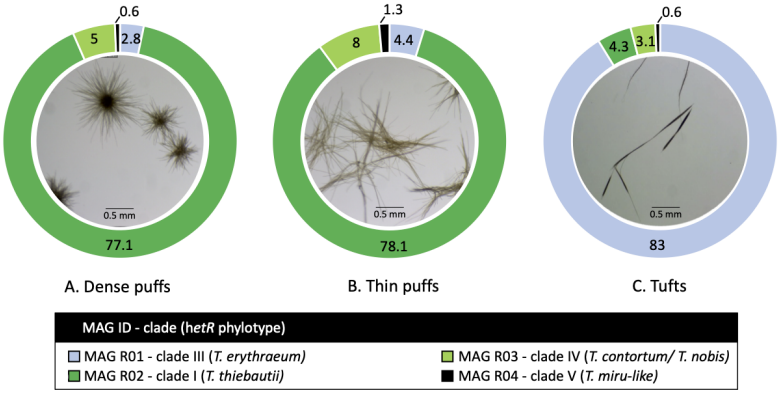
This Markdown File presents the scripts and analysis used to obtain each of the figures from the manuscript: "Metagenomes of Red Sea subpopulations challenge the use of marker genes and morphology to assess Trichodesmium diversity"

In this manuscript, Trichodesmium MAGs from a metagenomic dataset (MORPH) and additional datasets from previous publications (OCEAN) are phylogenetically analysed.

The MORPH dataset - concerns 3 morphotype sampels that were collected:

- Long Puffs
- Dense Puffs
- Tufts

(Figure 1) - Collected Samples for Metagenomics



We show the following forms of analysis:

- Metagenomic Pipeline to assemble and obtain our 5 Trichodesmium bins.
- Differential coverage plot of samples to MAGs
- **Figure 2:** Phylogenetic Analysis of the 5 Trichodesmium MAGs using a set of Cyanobacteria Marker Genes (GTO-tree) and Average Nucleotide Identity (ANI).
- **Figure 3:** Phylogenetic Analysis of the 5 Trichodesmium MAGs using hetR as a marker gene
- **Figure 4:** Visualisation of the raw reads mapped to the hetR containing contig
- **Supplementary Figure 1:** Phylogenetic Analysis of the 5 Trichodesmium MAGs using rcbL

METAGENOMIC PROCESSING

ATLAS 2:

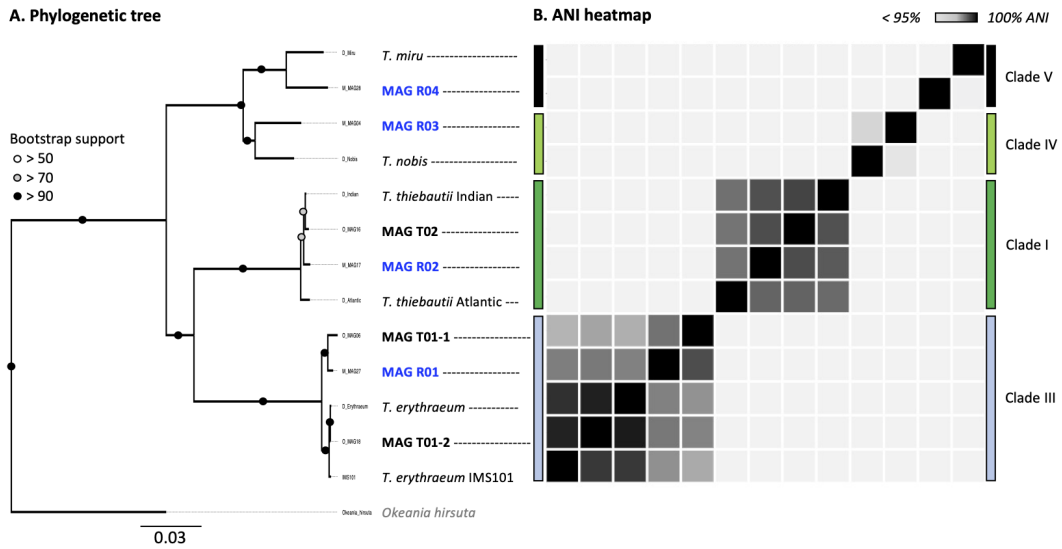
A snakemake assembly, annotation, and binning pipeline

Uses: BBTools; metaSPAdes; MetaBAT 2; MaxBin 2.0; VAMB; CheckM; DAS Tool; dRep; Prodigal; GTDB-tk

Important! - when performing ATLAS - do not have other fastq files present in folders within your folder - ATLAS will attempt to use these to create a tsv file

```
cd /home/bioinf/Desktop/Data/Coco/MORPH #MORPHOTYPES
cd /home/bioinf/Desktop/Data/Coco/OCEAN #OCEAN

#conda activate atlasenv

atlas init -d /media/bioinf/Data/ATLAS -w /media/bioinf/WD/coco/Morphotypes/reads/ATLAS2
--threads 48 --assembler spades /media/bioinf/WD/coco/Morphotypes/reads/ATLAS2
#interleaved fastq is not necessary for MORPH

atlas run -w /home/bioinf/Desktop/Data/Coco/OCEAN/SRR -c
/home/bioinf/Desktop/Data/Coco/OCEAN/SRR/config_OCEAN.yaml all --resources mem=680
--keep-going --skip-qc

# make the necessary changes to your .yaml file (high enough RAM)

#ATLAS Morphotypes - kmer 21,33,55,77,99,121

#Select MAGs that meet 90% completeness and 10% contamination on the CheckM quality-control

#bin_stats can additionally be found as output for quality control

#97.5 % ANI binning cut off.

# 5 Trichodesmium MAGs were present within the MORPH dataset
# More than the 3x morphotype samples isolated.
```

(Figure 2) - GTO- phylogenetic Tree and ANI-heatmap



Multi-locus (251 HMMs) phylogenomic tree (a) and average nucleotide identity (ANI) heatmap (b) of Trichodesmium MAGs. All MAGs assembled in this study are marked in bold. Red Sea MAGs (our samples) are marked in blue. The phylogeny includes five MAGs from the TARA Oceans dataset and the laboratory culture T. erythraeum IMS101. The tree was rooted at Okeania hirsuta (grey). The accession numbers for each MAG can be found in Supplementary Table 1 and 2, and their ANI values in Supplementary Table 3a and Table 3b.

Figure 2 is composed of the following 2 analyses:

1. Phylogenetic Tree based on an alignement of 250 selected HMM domains from Cyanobacteria in GTO-tree.

2. Heatmap comparing the average-nucleotide-identity (ANI) between the different Trichodesmium MAGs in PyANI.

Phylogenetic Tree (GTO-Tree)

- A 251 single-copy gene-set hidden Markov Models (HMMs) for Cyanobacteria is concatenated and aligned using GToTree (v.1.16.12; default settings).

- The alignment is refined using Gblocks (0.91b; default settings) to eliminate poorly aligned positions and divergent regions.

- A phylogenetic tree is constucted from the cleaned alignment using IQtree2 with ModelFind to estimate the best-fit model. Branch supports are estimated from 1000 bootstraps.

- The constructed phylogenetic tree is visualized using FigTree (v1.4.4).
- The tree is rooted with Okeania hirsuta (GCA_003838225) as an outgroup.

Before starting: make a .txt file (metadata file) listing all fasta names (double check for strange sample names) of the sequences we want to include in our analysis. All sequences must be in FASTA format.

```
conda activate gtotree

gtt-hmms # to know the HMM groups possible (Cyanobacteria - 251 HMMs)

GToTree -f /Users/dustbin/Desktop/COCO_HETR/MAGS_Analysis/1_Phylogeny/GTO/GTO_MAGs.txt -H Cyanobacteria

# Alignment of GTO (Aligned_SCGs.faa) is then cleaned in GBlocks.
# Other alignment options are MAFFT or ClustalW

gblocks # interactive program on the terminal

# o. file name
Aligned_SCGs.faa
# b. get gblocks
# Aligned_SCGs.faa
# Original alignment: 60609 positions
# Gblocks alignment:  13700 positions (22 %) in 78 selected block(s)

# Construct phylogenetic tree using IQtree2 (without gblocks)
/Users/dustbin/iqtree-2.1.3-MacOSX/bin/iqtree2 -s
/Users/dustbin/Desktop/COCO_HETR/MAGS_Analysis/1_Phylogeny/GTO/Tree/gblocks/Aligned_SCGs.faa
--alrt 1000 -B 1000 -T 8
# Best-fit Model: JTT+F+I+G4

# Construct phylogenetic tree using IQtree2 (with gblocks)
/Users/dustbin/iqtree-2.1.3-MacOSX/bin/iqtree2 -s
/Users/dustbin/Desktop/COCO_HETR/MAGS_Analysis/1_Phylogeny/GTO/Tree2/Aligned_SCGs-gb.faa
--alrt 1000 -B 1000 -T 8
# Best-fit Model: Q.plant+F+I+G4

# Alignment (before and after gblocks) can be compared using Geneious
# Visualize constructed tree using FigTree (branch support estimates - seen as branch-'lengths')
```

ANI HeatMap (PyANI)

- Diversity of our Trichodesmium MAGs was assessed phylogenetically using their average nucleotide identity (ANI)

- PyANI was performed using ANVIO 7.1 software

- Values sharing an ANI higher than 95 corresponds to a widely-used convention for bacterial species boundaries.

- Coverage Values sharing a coverage higher than 50 corresponds to a strict majority of each genome in the comparison being alignable (a plausible ad hoc minimum requirement for two sequences being considered the same thing).

For more information on how to interpret the PyANI output:
https://pyani.readthedocs.io/en/latest/interpreting_plots.html

```
conda activate anvio-7

#rename the fasta file - each contig needs to be "simple"
for f in *.fasta; do anvi-script-reformat-fasta $f -o ./rename/$f -l 800 --simplify-names; done

#construct a database from each renamed fasta file
cd ./rename
for f in *.fasta; do tag=${f%.fasta}; anvi-gen-contigs-database -f $f -o $tag.db -T 8; done

#run HMMs
for f in *.db; do anvi-run-hmms -c $f -T 8; done

#can also run NCBI COGs (version)
anvi-setup-ncbi-cogs -h #install diamond
for f in *.db; do anvi-run-ncbi-cogs -c $f -T 8; done

#start the pangenomic analysis
#includes Okeania_hirsuta
anvi-gen-genomes-storage -e MAGS.csv -o MAGS-GENOMES.db

anvi-pan-genome -g MAGS-GENOMES.db \
                --project-name "MAGS" \
                --output-dir ANI3 \
                --num-threads 8 \
                --minbit 0.5 \
                --mcl-inflation 2 \
                --use-ncbi-blast

anvi-compute-genome-similarity --external-genomes MAGS.csv \
                               --program pyANI \
                               --output-dir ./ANI3/ANI \
                               --num-threads 8 \
                               --pan-db ./ANI3/MAGS-PAN.db

#modify the phylogenetic tree using FigTree of the newick file
#% identity is written in the txt file.

anvi-display-pan -g MAGS-GENOMES.db \
                 -p ./ANI3/MAGS-PAN.db

#pyANI uses an ANI 95% cutoff as default
```

```
#the values of the Average Nucleotide Identity (ANI)
#the values of the Percentage ANI Coverage
#can be found in the output files.
```

Differential Coverage Plot of MAGR02 (T. thiebautii) between samples

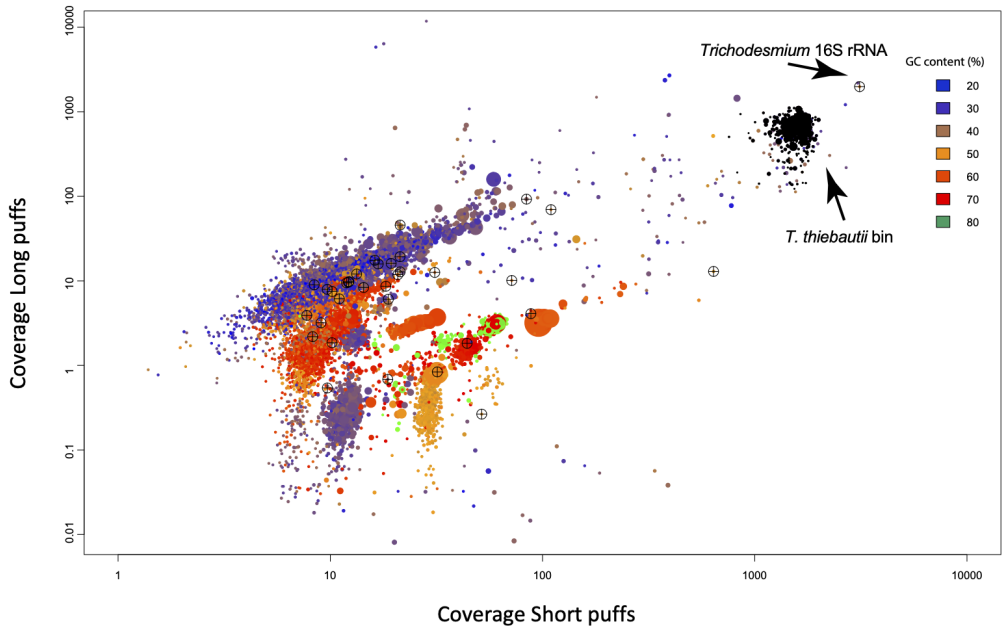


Figure 1: **Differential Coverage Plot of MAGR02 (T. thiebautii) in Long Puff and Dense Puff samples**

The plot depicts the total gene scaffolds present within the dense-puff sample and compares their coverage between the dense and long puff sample. The presence of 16S genes are marked by black circle-crosses.Colours depict GC differences.

The MAG R02 bin scaffolds, highlighted in black, cover the majority of the high-coverage scaffolds in both dense and long puff samples (top-right corner).

The plot further shows that manual binning using differential coverage will likely produce a bin coinciding with one constructed from automatic binning, and further confirms the quality of automatic binning of MAGR02 in this case. From this plot, we gather that it is therefore unlikely that there were two distinct high-abundance Trichodesmium populations present within our dense-puff sample. This is important when discussing the presence of 2 hetR genes within this particular genome.

*Note that Trichodesmium thiebautii 16S is outside of the bin, likely due to being present in multicopy, as was shown to be the case for T. erythraeum IMS101.

Before Starting:

- Install bbmap to get coverage files (covstats) from each sample. Make sure to remove the hastag in the cov plot files as it will confuse gbtools.

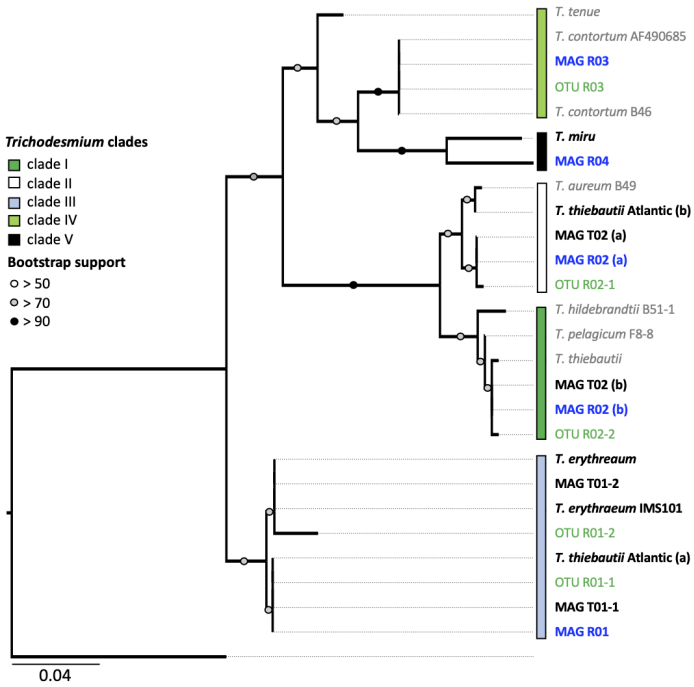
- Install barrnap to get 16S annotations using Silva.

- Download the full set of contigs of a sample (ATLAS2 output) fasta file format.

```
#get the coverage of each SAMPLE
bbmap.sh threads=48 minid=1 ref=SPC.fasta nodisk
in1=/media/bioinf/Data12/Coco/MORPH/ATLAS2/SPC_R1_001.fq.gz
in2=/media/bioinf/Data12/Coco/MORPH/ATLAS2/SPC_R2_001.fq.gz covstats=SPC.cov

conda activate barrnap

perl -/bin/genome-bin-tools/accessory_scripts/get_ssu_for_genome_bin_tools3.pl -d
/media/bioinf/Data/Pfdata/138.1/SILVA_SSU.noLSU.masked.trimmed.fasta -c 24 -a
idas16_final_contigs.fasta -o 16

## Visualise coverage in RStudio using Gbtools

#install.packages("sp")
#install.packages("plyr")
#install.packages("gbtools_2.6.0.tar.gz",repos=NULL,type="source")
#install.packages("devtools")
#install.packages("usethis")
#install_github("kbseah/genome-bin-tools/gbtools") # Install latest version of the R package

library(gbtools)
library(devtools)
library (gbtools)

# make sure to remove the # in the cov plot files!
d <- gbt(covstats=c("SPC.cov","LPC_2.cov"), ssu="spc.ssu.tab") # SSU gene annotations

plot(d,slice=c(1,2), cutoff=4000, taxon="Order", textlabel=F, gc=T,ssu=T, marker=F,
ylim=c(0.001,9000),xlim = c(1,10000),legend=TRUE)

#highlight the bin SPC_metabat_11 in black
bin1.contigNames <- scan(file="mag11.csv",what=character())
d.bin1 <- gbtbin(shortlist=bin1.contigNames,x=d,slice=NA)
points(d.bin1,col="black",slice=c(1,2))
```

(Figure 3) HetR Phylogenetic Analysis

- Diversity of our Trichodesmium MAGs was assessed phylogenetically using the HetR marker gene.



Main Figure 3: Trichodesmium hetR phylogeny. The tree is based on the alignment of 312 nucleotide sequences. Bold text indicates the hetR sequences from MAGs (Red Sea sequences are marked in blue). Amplicon hetR sequences are shown in green (Supplementary Table 4). The five proposed Trichodemium clades are shown with colored boxes. Note that clades IV and V are polyphyletic and are therefore not completely resolved. MAGs R02 and T03 contain two hetR sequences, clustering within clades I and II, respectively. The average percentage identity (ANI) matrix between the hetR gene sequences can be found in Supplementary Table 4.

Before Starting:

- MAGs of interested were uploaded in RAST.
- HetR sequences taken from each Trichodesmium sp. MAG were obatined using a protein BLAST of the HetR AA sequence in RAST using the SEED server.
- HetR gene sequences from each Trichodesmium MAG was aligned with other published hetR sequences using Multiple Alignment using Fast Fourier Transformation (MAFFT; default settings).

Do we use GBlocks or not for analysis?

- HetR1 (w/out Gblocks)
- HetR1 (w/ Gblocks)

```
# HETR

# MAFFT alignment (online - default L-INS-i)
# https://mafft.cbrc.jp/alignment/server/

# Upload alignment in Genious - select consensus sequence and save as a fasta file

# Conduct an IQTree (raw vs consensus vs gblocks)
# HetR_alignment.fasta
# HetR_alignment_con.fasta
# HetR_alignment_gb.fasta

cd /Users/dustbin/Desktop/COCO_HETR/MAGS_Analysis/1_Phylogeny/HetR/
/Users/dustbin/iqtree-2.1.3-MacOSX/bin/iqtree2 -s HetR_Alignment.fasta
--alrt 1000 -B 1000 -T 8
#Best-fit Model: TPM3+F+G4

/Users/dustbin/iqtree-2.1.3-MacOSX/bin/iqtree2 -s HetR_Alignment_con.fasta
--alrt 1000 -B 1000 -T 8
#Best-fit Model: TPM3+F+I

gblocks # HetR_All_alignment.fasta
# Original alignment: 924 positions
# Gblocks alignment:  448 positions (48 %) in 1 selected block(s)
/Users/dustbin/iqtree-2.1.3-MacOSX/bin/iqtree2 -s HetR_alignment-gb.fasta
--alrt 1000 -B 1000 -T 8
# Best-fit Model: TPM3+F+I

gblocks # HetR_Alignment_con.fasta
#Original alignment: 442 positions
#Gblocks alignment:  442 positions (100 %) in 1 selected block(s)
/Users/dustbin/iqtree-2.1.3-MacOSX/bin/iqtree2 -s HetR_Alignment_con-gb.fasta
--alrt 1000 -B 1000 -T 8

# Best-fit Model: TPM3+F+I

# Trees are visualized using Genious or FigTree
```

HetR Alignment Results

- **Without gblocks**

- **With gblocks**

Gblocks chops of the non consensus regions (flanking regions) giving a much cleaner alignment of the HetR gene than without. From the original alignment (924 positions) around 48 percent was used (448 positions) in 1 selected block.

The hetR phylogeny tree in Figure 3 therefore represents the analysis conducted on hetR sequences that were aligned and cleaned in Gblocks.

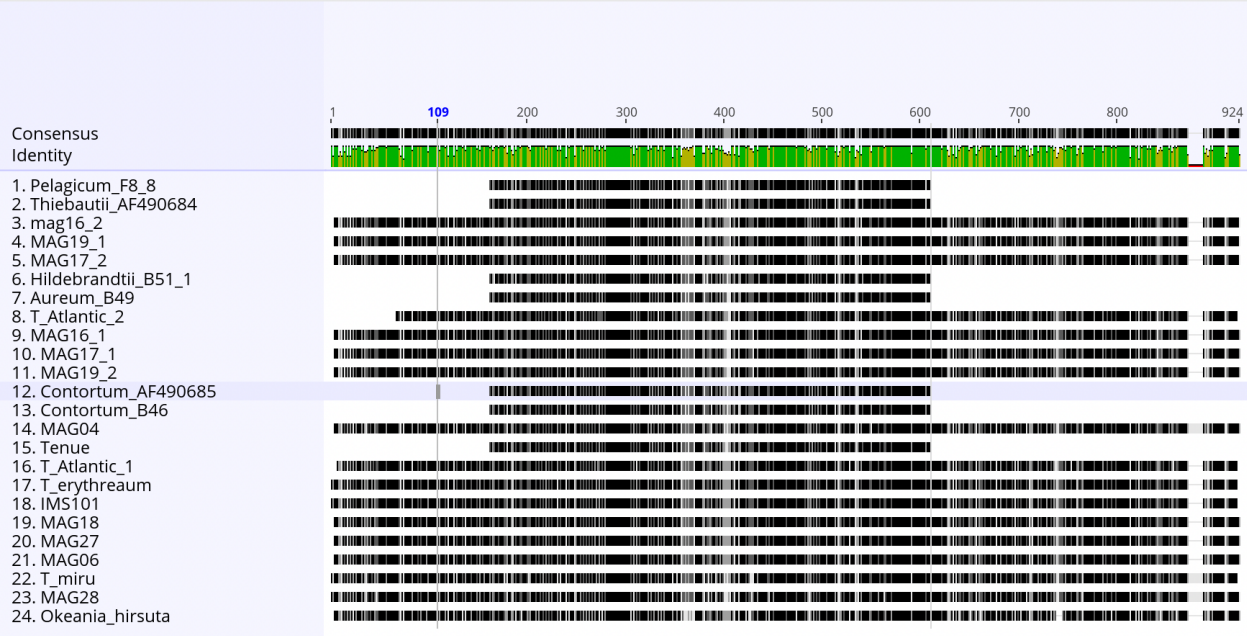


Figure 2: **Markdown Figure 3: HetR Alignment**

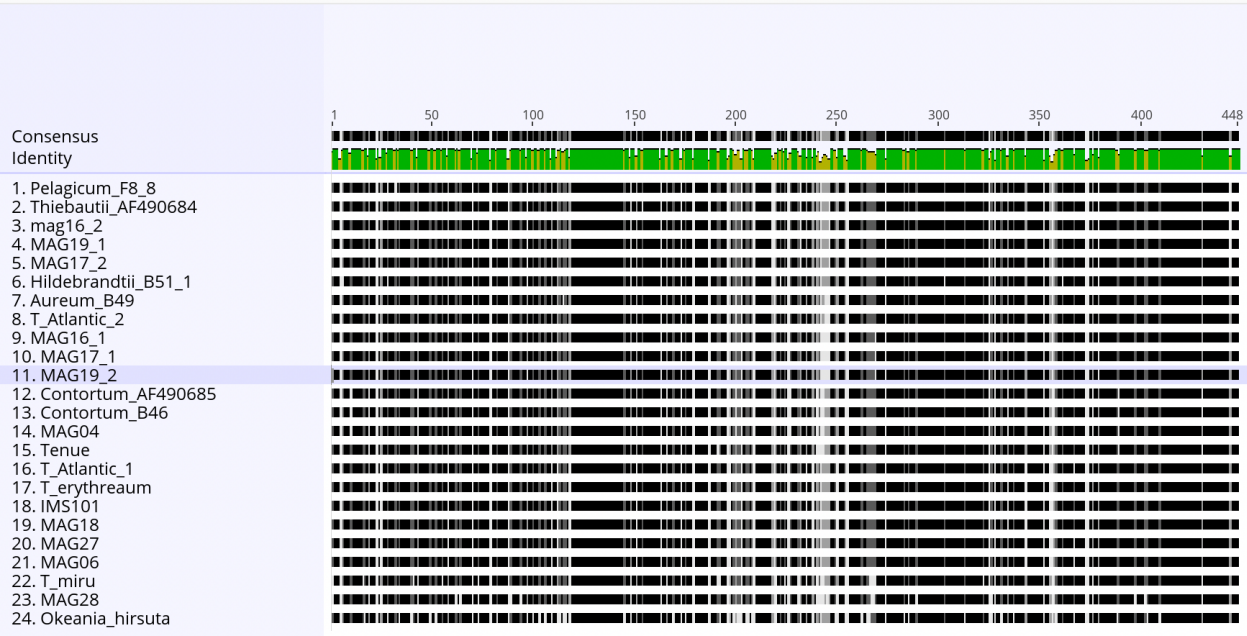


Figure 3: **Markdown Figure 4: HetR Alignment-gb**

(Figure 4) Visualization and Raw Read Counts of the HetR gene clusters

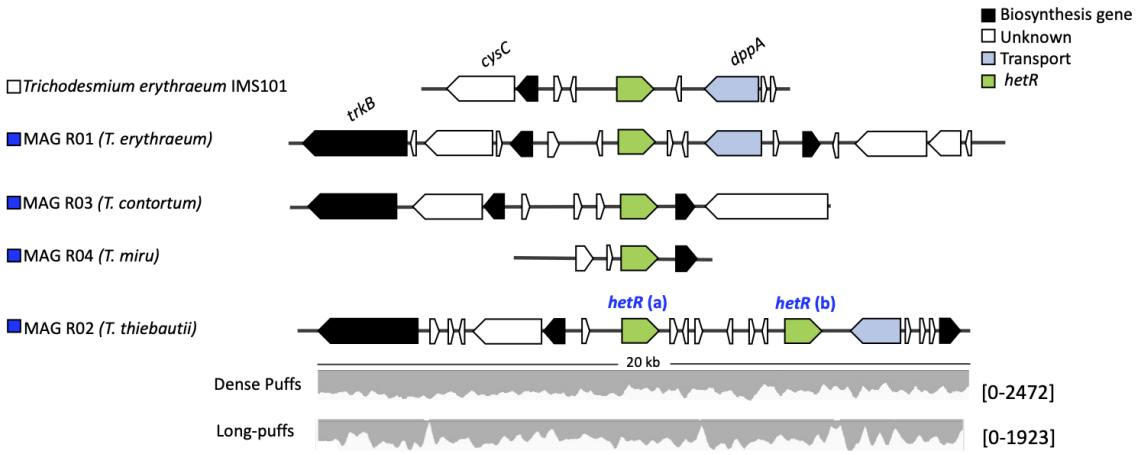


Figure 4: The hetR gene clusters in Trichodesmium sp. The synteny of the hetR genes (green) and the neighboring genes of the cluster is conserved between the Trichodesmium MAGs and T. erythraeum IMS101. MAG R02 (T. thiebautii) contains two hetR genes. Raw read counts, listed within the brackets, were mapped to the hetR gene cluster of MAG R02 from both the dense and long puff samples.

To assess the quality of the assembly - in light of these findings raw read counts of the three samples can be mapped back to MAG R02 (T. thiebautii containing the two hetR genes). If the raw read counts are not homogeneously aligned - it means that assembly contamination likely resulted (reflected by stark differences in the raw read counts).

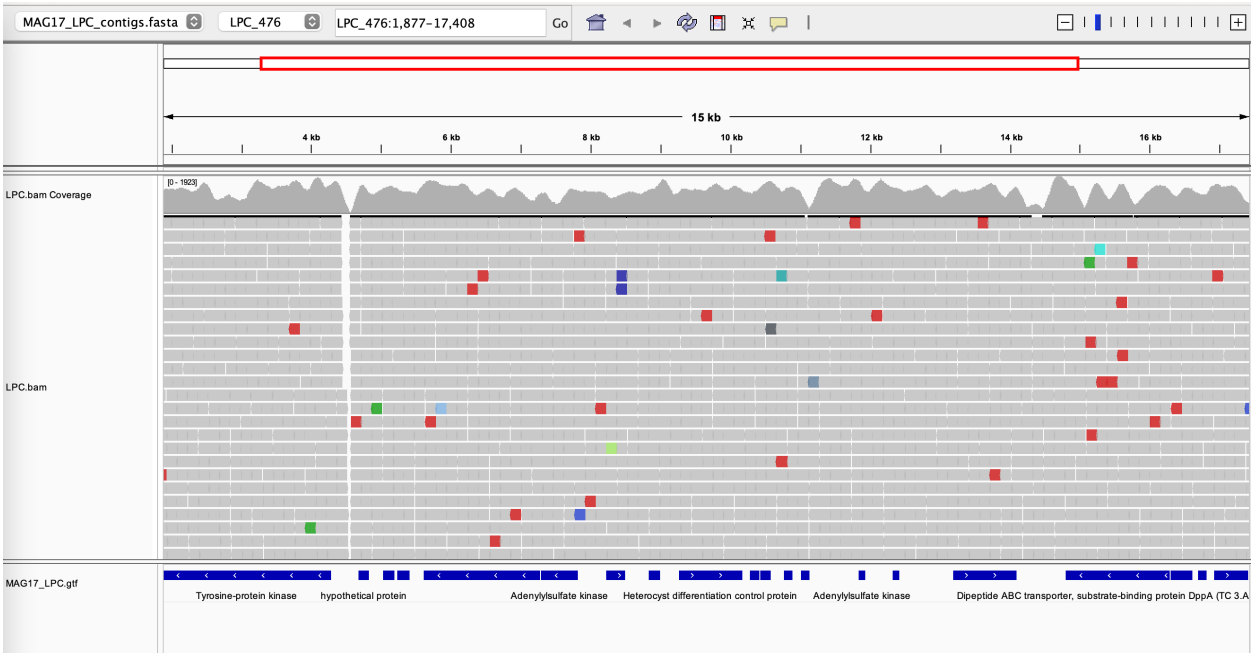
- Download the DNA_contig.fa file from RAST
- Download the gtf file (EC numbers stripped) from RAST
- Record (for yourself) the peg and contig the two hetR genes are in

```
#Dense Puff Samples - minid=1
bbmap.sh minid=1 threads = 48 ambiguous=best sam=1.3 ref=MAGR02_contig.fa nodisk
in1=SPC_R1.fastq.gz in2=SPC_R2.fastq.gz outm=MAGR02.sam
#map raw reads of dense puff samples to MAGR02

samtools view -bSh1 MAGR02.sam | samtools sort -m 20G -@ 4 -o MAGR02_SPC.bam
#convert to bam

rm MAGR02.sam #remove sam file

samtools index MAGR02_SPC.bam # open later in IGV

featureCounts -C -T 48 -t CDS -g ID -a MAGR02.gtf
-o MAGR02_SPC_counts.txt MAGR02_SPC.bam
#count table from bam file (*.bam also possible)
```

```
#Long Puff Samples - minid=1
bbmap.sh minid=1 threads = 48 ambiguous=best sam=1.3 ref=MAGR02_contig.fa nodisk
in1=LPC_R1.fastq.gz in2=LPC_R2.fastq.gz outm=MAGR02_LPC.sam
#map raw reads of long puff samples to MAGR02

samtools view -bSh1 MAGR02.sam | samtools sort -m 20G -@ 4
-o MAGR02_LPC.bam #convert to bam

rm MAGR02.sam #remove sam file

samtools index MAGR02_LPC.bam # open later in IGV

featureCounts -C -T 48 -t CDS -g ID -a MAGR02.gtf
-o MAGR02_LPC_counts.txt MAGR02_LPC.bam
#count table from bam file (*.bam also possible)
```

Visualise the raw reads onto the contig containing the hetR repeat.

- Open the MAGRO2_contig.fa in IGV (under genomes) /
- Open the (indexed) BAM file /
- Open the MAGR02.gtf file and see the annotated genes /



Note the high coverage of the genome (around2000 reads) - also shown when looking at the featurecount table. Possible importance - to normalise to read length (not done for this analysis)

(Supplementary Figure 1) RbcL Phylogenetic Analysis

Supplementary Figure 1: Trichodesmium rbcL phylogeny. Bold text indicates the hetR sequences from MAGs (Red Sea sequences are marked in blue). The five proposed Trichodemium clades are shown with

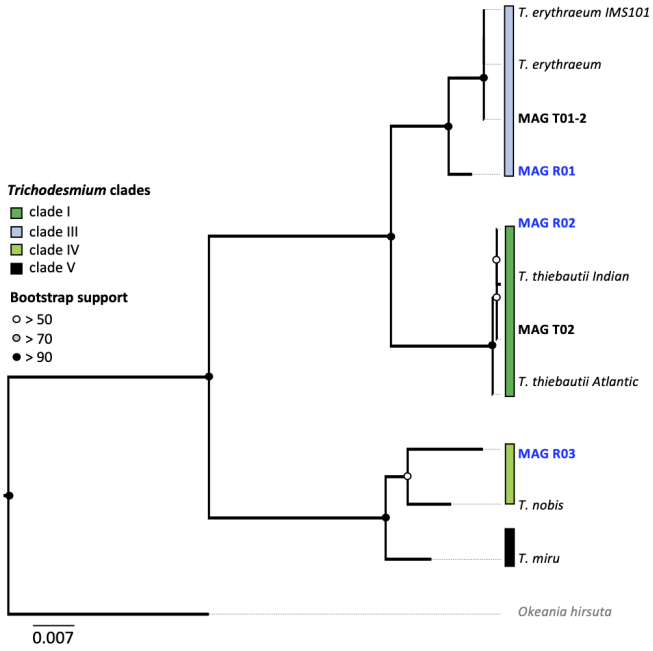


Figure 4: **Main Figure 6: rbcL Phylogenetic Tree Tree**

colored boxes and is congruent with phylogenomics, although an rbcL gene was missing in MAG T01-1. Nonetheless, rbcL may serve as an alternative to address the diversity of Trichodesmium populations by amplicon sequencing, although will need to be verified experimentally.

```
# MAFFT alignment (online - default L-INS-i)
# https://mafft.cbrc.jp/alignment/server/

# Upload alignment in Genious - select consensus sequence and save as a fasta file

# Conduct an IQTree (raw vs consensus vs gblocks)
# RbclX_alignment.fasta

cd /Users/dustbin/Desktop/COCO_HETR/MAGS_Analysis/1_Phylogeny/RbclX/
/Users/dustbin/iqtree-2.1.3-MacOSX/bin/iqtree2 -s RbclX_alignment-gb.fasta
--alrt 1000 -B 1000 -T 8

# Best-fit model: TN+F+G4
# Trees are visualized using Genious or FigTree
```