HetR Paper

- Genomes of natural Trichodesmium colonies isolated from the Red Sea.
- Comparison of 3 genomic datasets (OCEAN, FICUS, MORPH)
- Several genomes (97.5 % binning cut off) present (5) - which is more than the morphotypes isolated (3).
- HetR paralogs observed within our genomes.

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 as well to create a tsv file

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

#conda activate atlasenv

atlas init .......

atlas run -w /home/bioinf/Desktop/Data/Coco/OCEAN/SRR -c /home/bioinf/Desktop/Data/Co
co/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
#ANI 97.5

#Select MAGs that meet 90% completeness and 10% contamination on the CheckM quality-c
ontrol.
```

Phylogenetic Analysis (GTo-Tree)

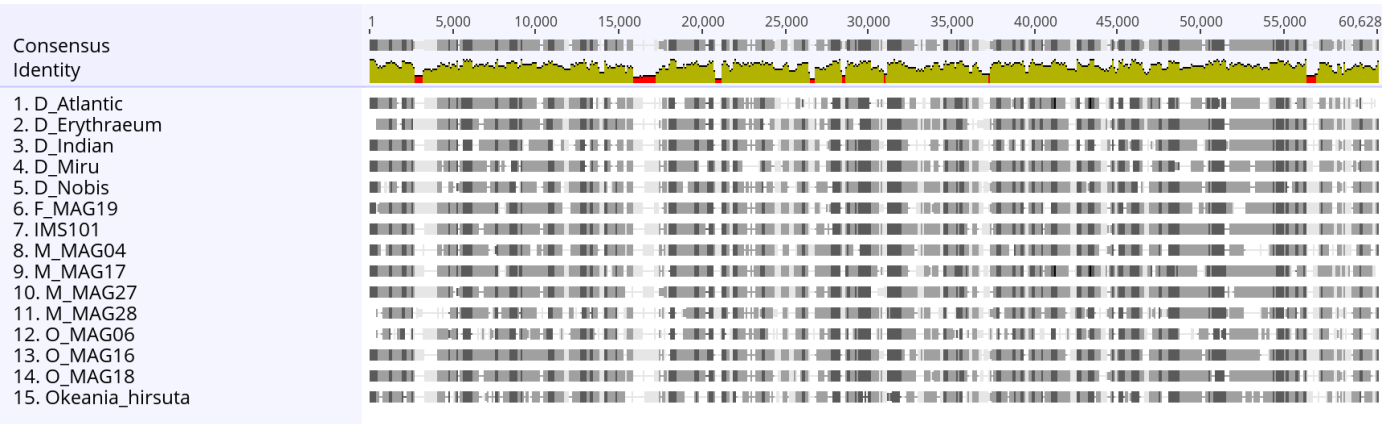
- Diversity of our Trichodesmium MAGs was assessed phylogenetically using GTo-Tree
- A 251 single-copy gene-set hidden Markov Models (HMMs) for Cyanobacteria is concatenated and aligned using GToTree (v.1.16.12; default settings).
- The alignment is refined using Gblocks (0.91b; default settings) to 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.tx
t -H Cyanobacteria -j 8 -o ./1_Phylogeny/GTO/Tree

# 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/MA
GS_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')
```

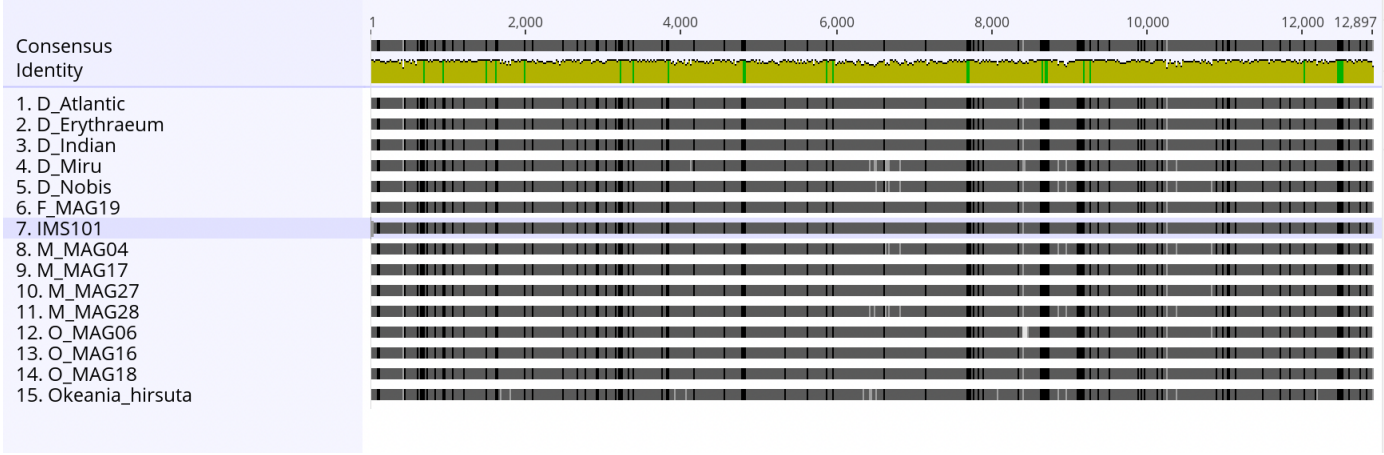
GTO Alignment Results

- **Without gblocks**



Markdown Figure 1: Alignment

- **With gblocks**



Markdown Figure 2: Alignment_gb

We can see that gblocks gives a cleaner alignment of the 251 HMMS in comparison to the absence of gblocks.
From the original alignment (60628 positions) only 21 % was used (12897 positions) in 76 selected blocks).
We continue with the cleaned alignment - and plot the resulting IQ-Tree:

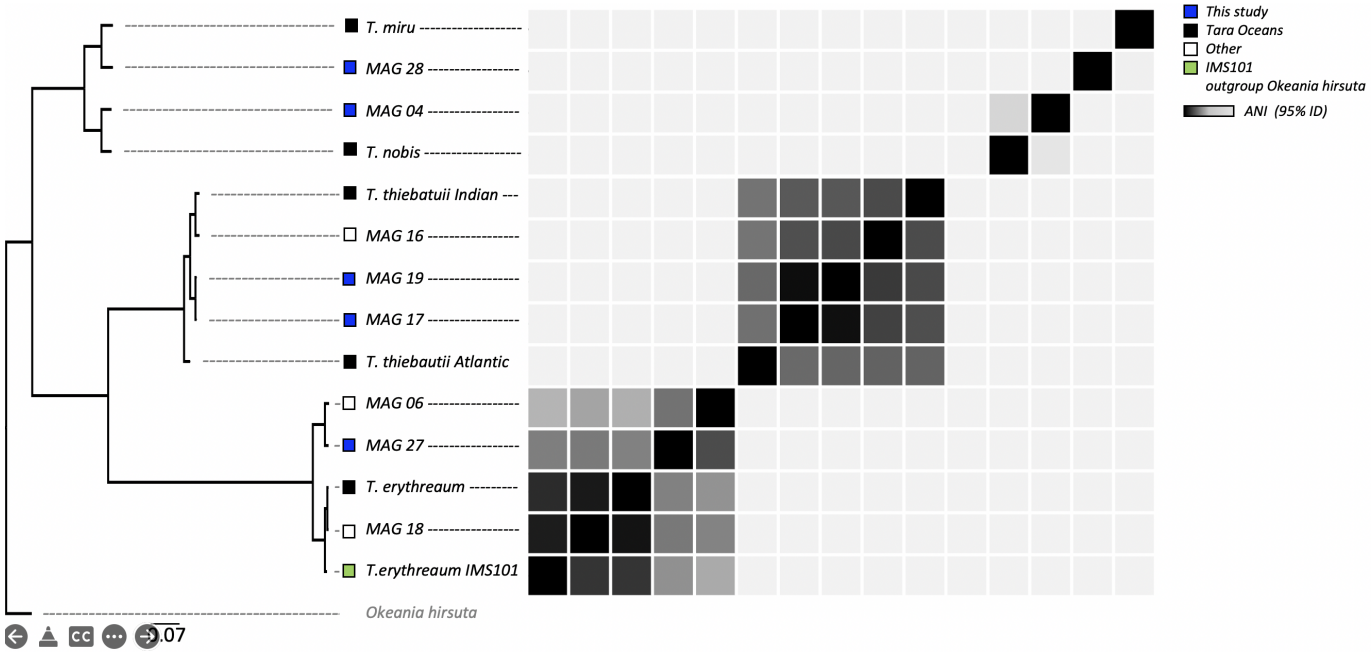
Main Figure 2a. Multi-locus Phylogenetic tree of Trichodesmium MAGs based on a truncated alignment of 251 HMMs selected for Cyanobacteria using Gto-Tools. Squares represents the genomes derived from this study (white), other metagenomic studies (green), Delmont (black) and the culturable strain T. erythreaum IMS101. Bootstrap confidence levels lower than <90% are shown. The genome of Okeania hirsuta was used as an outgroup.

Phylogenetic Analysis (ANI)

- Diversity of our Trichodesmium MAGs was assessed phylogenetically using their average nucleotide identity (ANI)
- ANVIO-7.1 software

```
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-n
ames; 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 writte
n in the txt file.

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

ANI-ANVIO Results



Main Figure 2b: HetR Alignment

- **Possible to merge GTO-tree plot together with the HeatMap (create a single phylogenetic figure!)**

Phylogenetic Analysis (HetR)

- Diversity of our Trichodesmium MAGs was assessed phylogenetically using the HetR marker gene.
- Sequences taken from each Trichodesmium sp. MAG by 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).

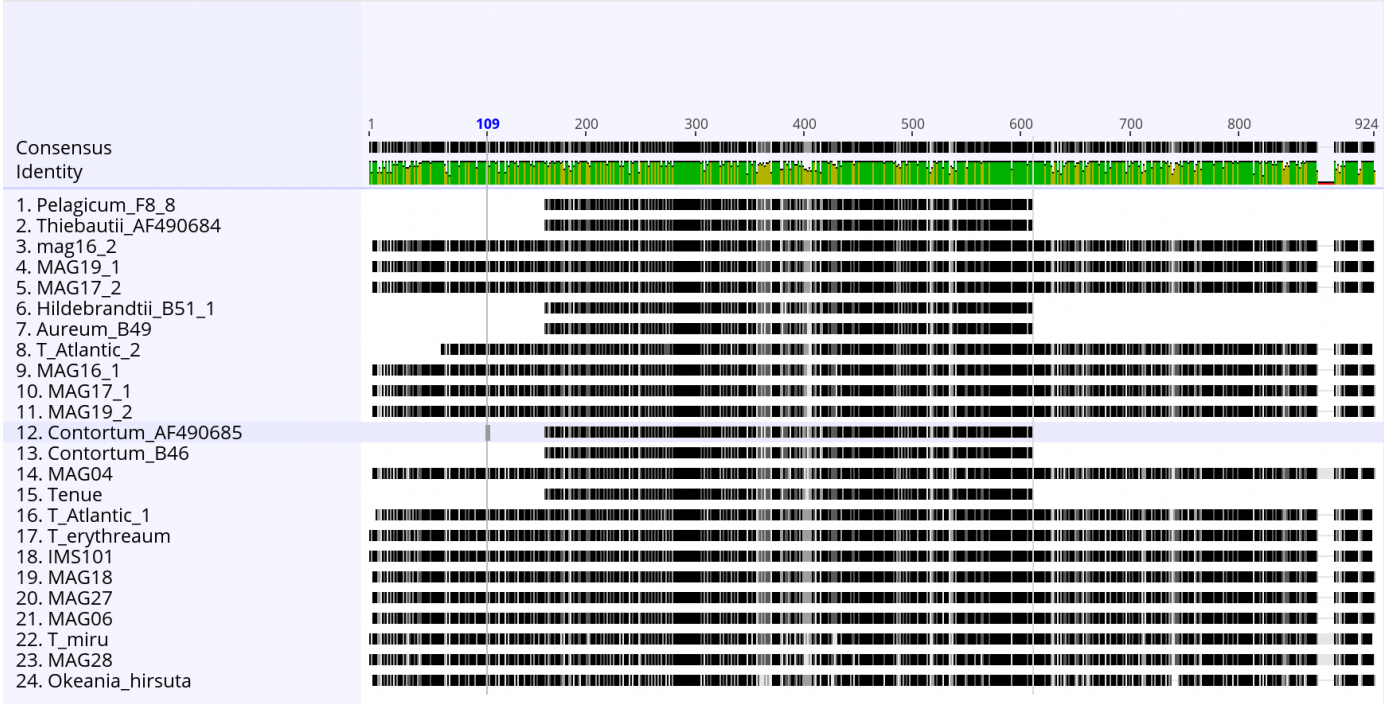
Do we use GBlocks or not for analysis?

- 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 100
0 -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
```

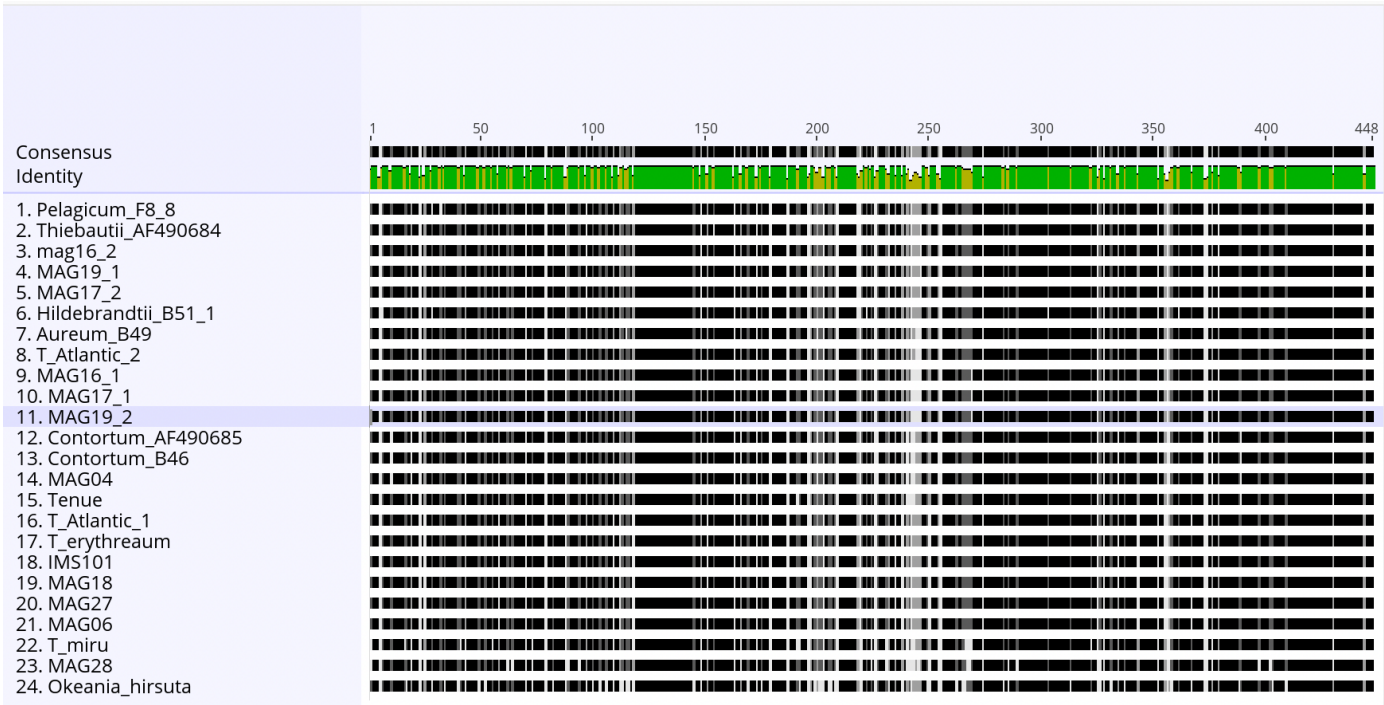
HetR Alignment Results

- **Without gblocks**



Markdown Figure 3: HetR Alignment

- **With gblocks**



Markdown Figure 4: HetR Alignment-gb

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% was used (448 positions) in 1 selected block.
We continued with the cleaned alignment resulting in the following IQ-Tree:

HetR Results

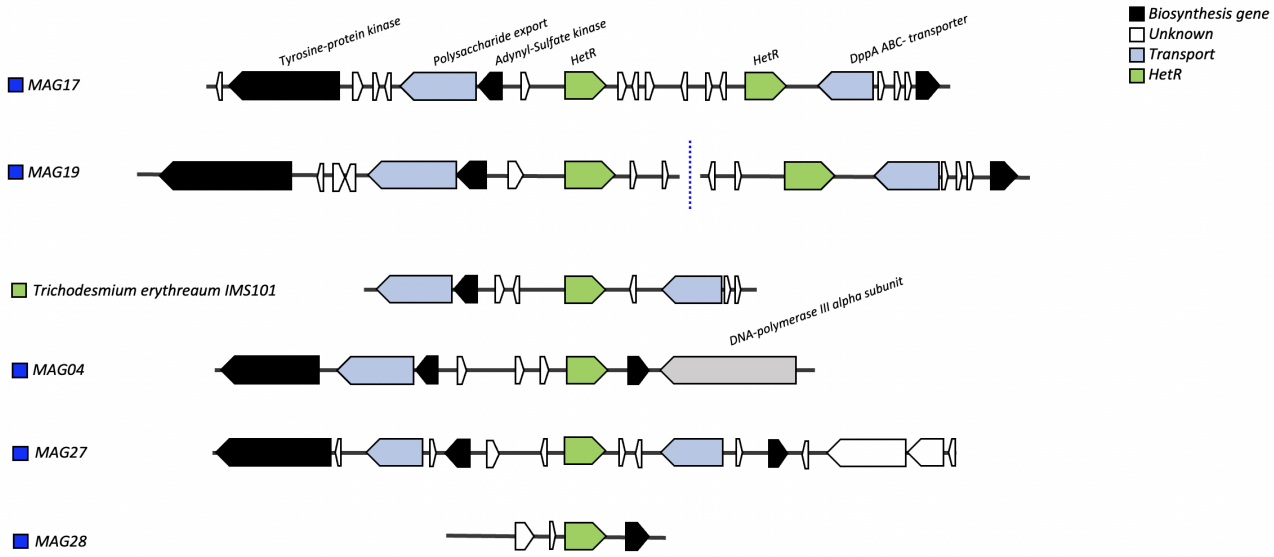
- **With gblocks**

Main Figure 3: HetR Phylogenetic Tree ## Phylogenetic Analysis (Rbcl-X)

```
# 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 100
0 -B 1000 -T 8
# Best-fit model: TN+F+G4
# Trees are visualized using Genious or FigTree
```

HetR gene clusters

- **from RAST server**



Main Figure 4: HetR Gene Clusters