Taxonomic distribution of metabolic functions underpins nutrient cycling in Trichodesmium consortia

Authors: Coco Koedooder

March 2023

SCRIPT Metagenomic analysis of functions within the Red Sea Trichodesmium consortium.

1. **Metagenomic Pipeline:**

- ATLAS2 Pipeline - assembly, binning and read coverage per bin

2. **Phylogenetic Analysis:**

- GTDB-tk taxonomy
- GTO-tree

3. **Functional trait Analysis:**

- MAG annotation - Ghost-Koala (KEGG) and HMMScan (PFAM)
- Vitamin B12 biosynthesis and uptake
- Particle-related Traits
- Siderophore biosynthesis
- Phosphate-uptake
- Nitrogen Metabolism

Metagenomic Pipeline: ATLAS2

Metagenomes were assembled and binned using the ATLAS (v2) pipeline (Kieser et al., 2020) which involved the following steps:

- Quality Control through BBTools
- Assembly via metaSPAdes - k-mer length 21, 33, 55, 99, 121 bp
- Binning using MetaBAT 2, MaxBin 2.0 and VAMB.
- Each bin was assessed for completeness and redundancy using CheckM2 v0.1.2
- A non-redundant set of bins is produced using DAS-Tool and dRep - 97.5% ANI cutoff
- Genes from each MAG are predicted using Prodigal.

ATLAS2 pipeline was performed seperately for each of the following datasets.

1. MAG -XX (Gs0149370)
2. MAG R-XX (PRJNA804487)
3. MAG T-XX (PRJNA435427, PRJNA358796, PRJNA330990)

```
conda activate atlasenv
cd /home/bioinf/Desktop/Data/Coco/FICUS_2 #Raw Data

# important:
# do not have other fastq files present in folders within your folder
# ATLAS will attempt to use these as well when creating a tsv file

# important:
# do not skip fastq  (error message/ bug)

atlas init -d /media/bioinf/Data/ATLAS -w /media/bioinf/WD/coco/Morphotypes/reads/ATLAS2
--threads 54 --assembler spades --interleaved-fastq /media/bioinf/WD/coco/Morphotypes/reads/ATLAS2

# interleaved fastq is not necessary for MORPH Dataset

# make the necessary changes to your .yaml file

ATLAS Ficus - kmer 21,33,55,77,99
            - ANI 97
# run atlas
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

#output MAGs
```

output MAGs for further downstream analysis.

Phylogenetic analysis (taxonomy and phylogenetic tree)

The 52 assembled genomes from this dataset were taxonomically characterized using the genome taxonomy database tool kit GTDB-tk. Similarly, we performed this on the genomes assembled from other metagenomic datasets.

```
#GTDB-tk
conda activate gtd
gtdbtk classify_wf --cpus 24 --extension fasta --genome_dir PATHWAY --out_dir PATHWAY/gtdb
```

The MAGs were further placed in a phylogenetic tree using GTo-Tree which uses a concatenated list of 74 conserved single-copy HMM markers for bacteria.The tree was visualized using FigTree (v1.4.4). The genome of Fuseobacterium nucleatum (PRJNA1419) were downloaded and added to the analysis as an outgroup to re-root the tree (see Coleman et al., 2021). The tree table was visualised using IQ-Tree.

```
#GTO-tree

# Make a list (metadatafile - with all fasta names - doublecheck for strange sampels)
# The sequences MUST be in fasta format
# Save from excel to txt file for correct format
# GTO-tree aligns the FASTA files using a selection of HMMS (e.g 74 HMMs for Bacteria)

conda activate gtotree

gtt-hmms # to know the HMM groups possible

GToTree -f GTO.txt -H Bacteria -j 8 -o Tree #74 genes

conda activate gblocks

#o. file name

Aligned_SCGs.faa

#b. get blocks
# Original alignment: 11237 positions
# Gblocks alignment:  106 positions (0%) in 2 selected block(s)
# Too much - will continue with unfilted alignment

# Construct phylogenetic tree using IQtree2 (without gblocks)
/Users/dustbin/iqtree-2.1.3-MacOSX/bin/iqtree2 -s
/Users/dustbin/Desktop/T0_paper/1_Phylogeny/tree/Aligned_SCGs.faa --alrt 1000 -B 1000 -T 8
# Best-fit Model: Q.pfam+R7

iqtree -s /Users/dustbin/Desktop/T0_paper/4_Proteomics/old2new/combi/Tree/Aligned_SCGs.faa
--alrt 1000 -B 1000 -T 8

#visualize tree using figtree
#visualize tree using iTOL
```

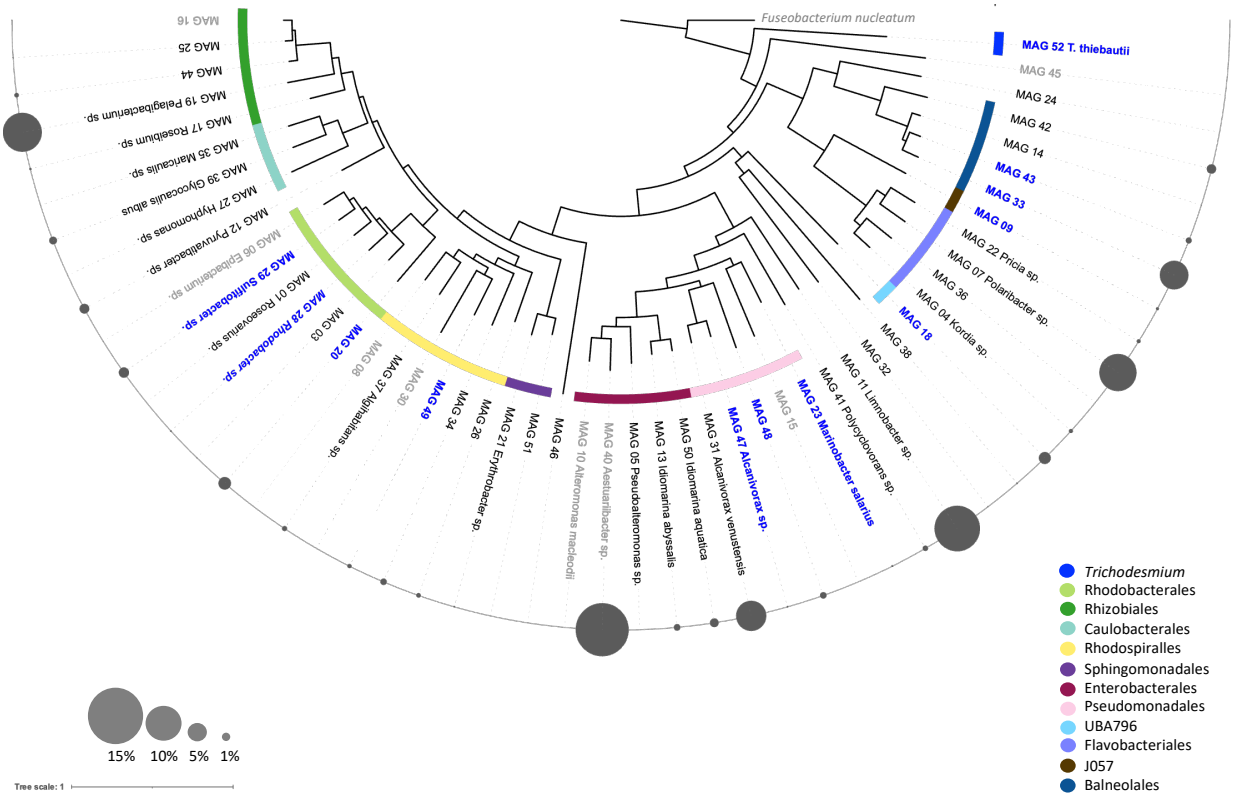


Figure 1: Visualization of iqtree - consensus tree file - in iTOL

MAG Annotation

The 52 MAGs were annotated to assess the presence of functional pathways within the consortium. All MAGs were annotated using the seed/RAST, pfam and kegg databases. For pfam and kegg annotation, MAGs were concatenated into one large text file.

```
## Rename MAG fasta files
for f in MAG*.fasta; do mv "$f" "${f/%.fasta/_OLD.fasta}"; done
## concatenate into one file
cat *.faa > concatenated.faa
```

hmmscan (pfam)

```
hmmscan --tblout OLD_hmm.tsv --cut_ga --cpu 8  /Users/dustbin/Bioinf/pfam/Pfam-A.hmm concatenated.faa
```

ghostkoala (kegg)

Upload the concatenated file to **ghostkoala**:https://www.kegg.jp/ghostkoala/):

RAST

Upload MAG fasta files individually to **RAST**:https://rast.nmpdr.org/):

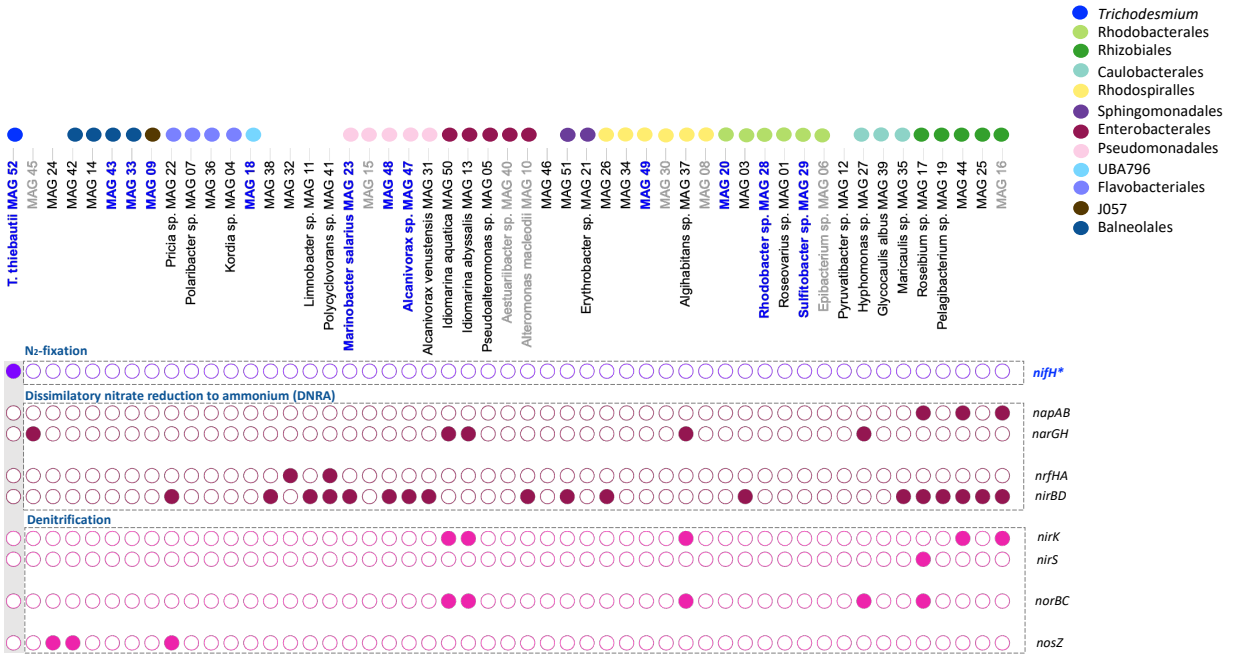
Metabolic

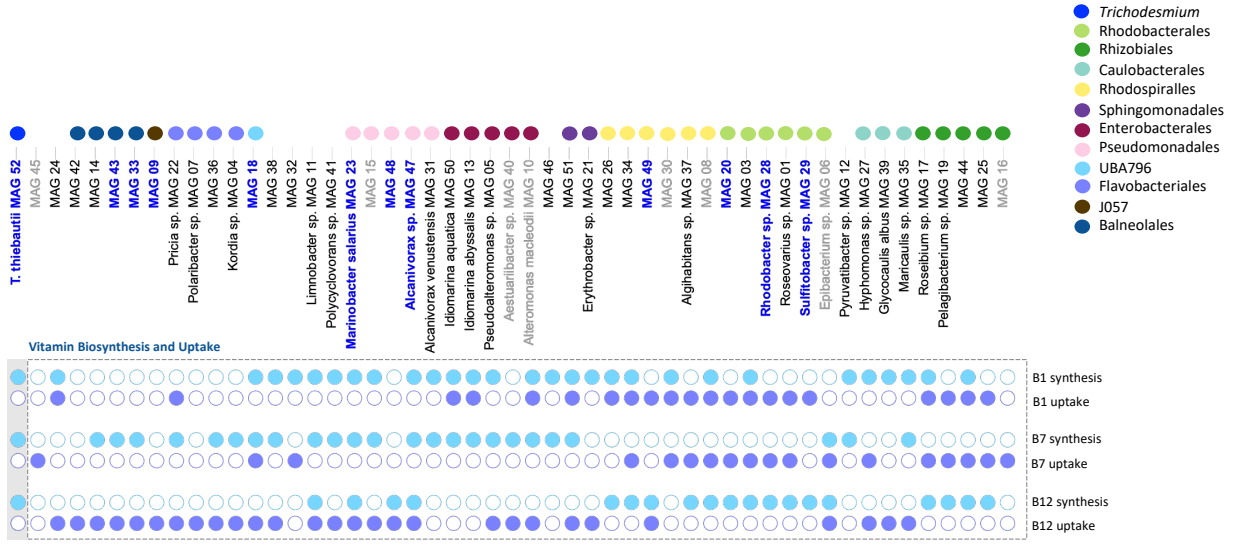
METABOLIC offers a standardized approach for the annotation of protein function for large datasets, and was used to assess nitrogen metabolism (denitrification, DNRA and nitrogen fixation) and vitamin B12 biosynthesis.

Due to the automatic presence/absence cutoff - some pathways would be listed as absent when they were in fact present (e.g. Vitamin B12 biosynthesis pathways in T. thiebautii MAG 52). To counter the presence of false-negatives, KEGG IDs were obtained from the METABOLIC output files to manually cross-check the presence of a pathway using KEGGs metabolic maps.

```
#script to print pathway and filename into a text file (that can then be modified)

printf '%s\n' "$PWD"/* > ../filenames.txt

# metabolic-c
# Place T01 QC-fastq.gz reads in folder - FW and RV
# Run scripts on the raw reads to get abundances

conda activate metabolic

perl METABOLIC-C.pl -t 40 -m-cutoff 0.75  -in-gn  -r omic_reads_parameters.txt -o METABOLIC_out
# make a txt file with the path of the FW and RV reads separated by a ','
# m-cutoff pathway completeness higher than 0.75%
# taxonomy gives additional output - summarising presence according to the 'order' level
# 2nd_run allows you to repeat for metaT data from your genomics data
```

```
# or play with parameters such as your m-cutoff

perl /home/bioinf/bin/metabolic/METABOLIC/METABOLIC-C.pl -t 32 -m-cutoff 0.75  -in-gn
/home/bioinf/Desktop/Data/Coco/FICUS/MetabolicC/genomes -kofam-db full -r filenames.txt
-st "illumina" -tax "order"  -o METABOLIC_out

perl /home/bioinf/bin/metabolic/METABOLIC/METABOLIC-C.2nd_run.pl -t 32 -in-gn
/media/bioinf/Data12New/Metagenomics22/Seep1/Idas2/metabolic2/genomes -r filenamesTR.txt
-rt "metaT" -st "illumina" -tax "genus" -o METABOLIC_out -2nd-run true -2nd-run-suffix 2nd_run_T
```





Presence/ Absence table of MAGs for a list of KEGG IDs of interest

R script that reformats output files of GhostKoala and Excel-Table list MAGs and KEGG IDs of interest

```
#Format a annotation table into a presence and absence table

library(reshape) #melt/transpose data
library(stringr) #get subset
library(dplyr)

# upload files
#KOs of interest file
#MAG names
#MAGs <- read.delim("Project_Julia.csv", header=TRUE, sep=",", dec=".") #data table with (240) KOs of i
MAGs <- read.delim("Project_Ficus.csv", header=TRUE, sep=",", dec=".") #KO annotations of all MAGs (wit

#KO annotation files
Koala <- read.delim("GhostKOALA_FICUS.csv", header=TRUE, sep=",", dec=".") #KO annotations of all MAGs

#reformat data for R - by transposing (melting) with columns 1 and 2 untouched
MAGs_long <- melt(MAGs, id = c(1, 2),
                  measured = c(3, ncol(MAGs)))

#rename column to KO
names(MAGs_long)[3] <- "KO"

#remove 'empty' value column
MAGs_long <- MAGs_long[, -4]

#make a new column "MAG ID dataset" - where column 1 and column 2 are pasted together, separated by "_"
#this column will be used to match names with the Koala data-table
MAGs_long$MAG_ID_dataset <- paste(MAGs_long$MAG_ID, MAGs_long$Dataset, sep="_")
#transposed data - results in 11424 rows

#In the Koala data-table - make a new column - that selects the first 11 characters
#Now both tables have a matching column that can be used as common identifier
Koala$MAG_ID <- str_sub(Koala$MAG,1,5)

#Make a new table from KOALA that only keeps unique identifiers
Koala_unique <- unique(Koala[, c(2, 3)])
#Koala goes from 194619 rows down to 72625

#Make a column called "presence" where those that have a KOALA ID match - get a 1 (present)
Koala_unique$presence <- 1
Koala_unique$KO <- as.factor(Koala_unique$KO)

#Now we match the KOs with the KO IDs of interest present in MAGs_long (column 3 - KO ID number)
Koala_unique_goodKOs <- unique(inner_join(MAGs_long[3], Koala_unique))
#2648 matches of the 72625

#Now we add a column of presence/ absence
MAGs_final <- full_join(MAGs_long, Koala_unique_goodKOs)
#to change NAs to Os (see script below)
#MAGs_final$presence <- ifelse(is.na(MAGs_final$presence), 0, MAGs_final$presence)
```

```
#now we get back the transposed data to its original format (where each KO represents a column)
#we should get back the 241 KOs of interest in each column
MAGs_final_transposed <- cast(MAGs_final,
                              MAG_ID + Dataset ~ KO,
                              value = "presence")

#write table to csv
write.csv(MAGs_final_transposed, "List_FICUS_FINAL.csv")
```

Siderophores

siderophore biosynthesis pathways were analysed using several different tools:

- FeGenie
- KEGG maps
- HMM - iucA/iucC domains
- AntiSMASH 7.0

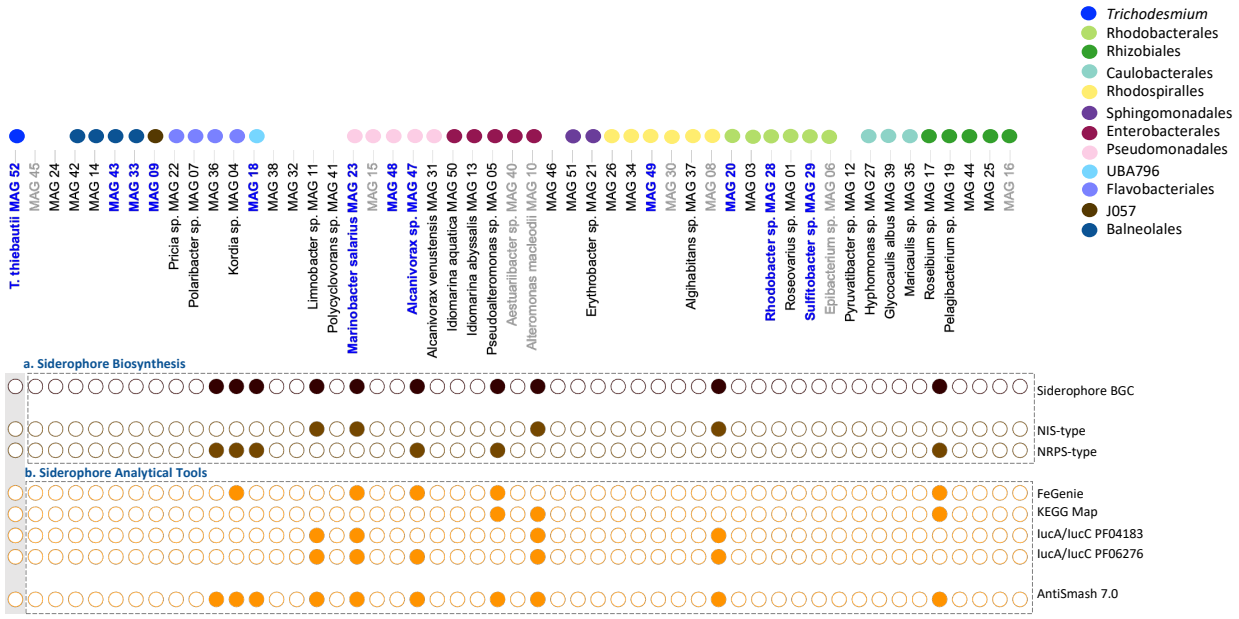


Figure 2: Tools to asses Siderophore Biosynthesis Pathways

NIS Siderophores

NIS domains were assessed through the presence of an IucA/IucC domain.

```
# NIS-Type (IUCA/IUCC domain)
# PF04183
# confirm with ANTISMASH for IucC/IucA and RAST

grep -c PF04183 /Users/dustbin/Desktop/T0_paper/2_Iron_Uptake/FICUS_hmm.tsv #7
grep PF04183 /Users/dustbin/Desktop/T0_paper/2_Iron_Uptake/FICUS_hmm.tsv | cut -f 3,5,18
> /Users/dustbin/Desktop/T0_paper/2_Iron_Uptake/siderophore_NIS_hits.tsv
```

NIS-genes (containin an iucA/iucC domain) from MAGs were clustered with known NIS-genes into one fasta file. A phylogenetic tree allowed us to group NIS genes into several different types as described by Carroll and Moore (2018).

MAFFT alignment

```
## MAFFT ############
https://www.ebi.ac.uk/Tools/msa/mafft/
NIS_all_aligned.fasta
```

A phylogentic tree was constructed from the alignment using IQtree2 (without gblocks)

```
# Construct phylogenetic tree using IQtree2 (without gblocks)

/Users/dustbin/iqtree-2.1.3-MacOSX/bin/iqtree2 -s
/Users/dustbin/Desktop/T0_paper/2_Iron_Siderophore/NIS_all_aligned.fasta --alrt 1000 -B 1000 -T 8

# Best-fit Model: Q.pfam+F+R4 chosen according to BIC
# Branch selection using Shimodira-Hasegawa
# 1000 Bootstraps
```

The consensus tree was visualised using FigTree

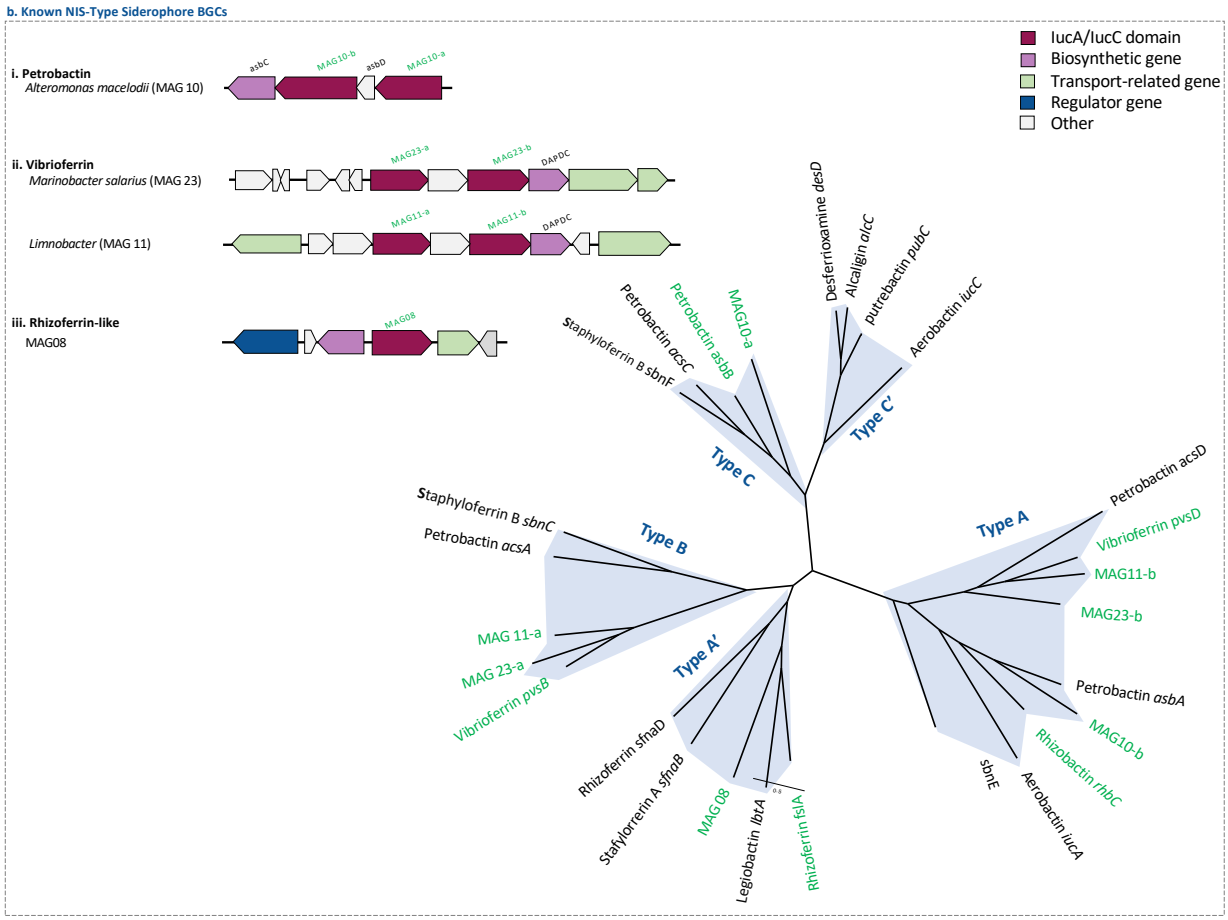


Figure 3: NIS Siderophore Biosynthesis Pathways

NRPS Siderophores

FeGenie

FeGenie was used to assess the NRPS-type siderophore biosynthesis pathways

```
#### FeGenie

conda activate fegenie
pwd
FeGenie.py -bin_dir /Users/dustbin/Desktop/T0_paper/2_Iron_Siderophore/FeGenie -bin_ext fasta
-out /Users/dustbin/Desktop/T0_paper/2_Iron_Siderophore/FeGenie/out -t 6 --makeplots
```

AntiSMASH 7.0

Upload fasta files and assess NRPS-pathways using MiBIG database and the presence of a siderophore TonB receptor.

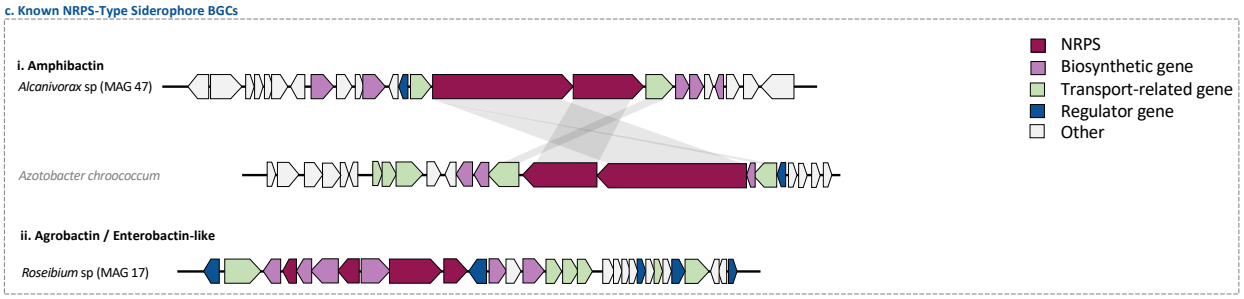


Figure 4: Known - NRPS Siderophore Biosynthesis Pathways

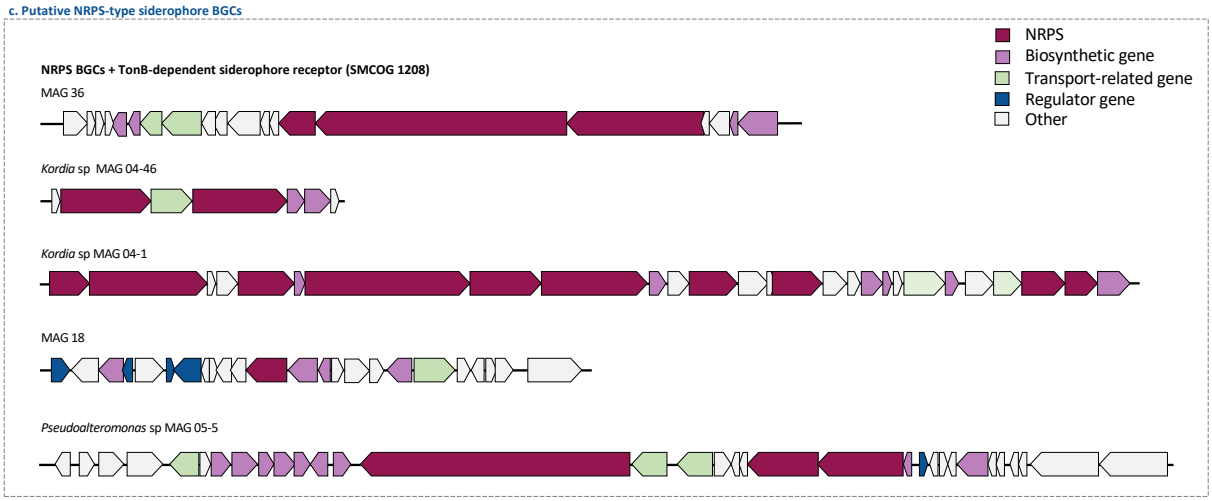


Figure 5: Putative NRPS Siderophore Biosynthesis Pathways