Guowei Review (bioinformatic analysis)

Coco Koedooder

Cyanobacteria FASTA Files

34 representative cyanobacteria FASTA files were downloaded for this Fe-analysis (see Table). The idea was to construct a phylogenetic tree coupled with a presence/absence list for different Fe-related genes within the text.

The full genome names, genebank accession number, bioproject and additional info can be found compiled in a metadata file:

```
guowei <- read.csv("/Users/mrblab/Desktop/Guowei/guowei_metadata.csv")
guowei <- guowei[c(1:5)]
colnames(guowei)
```

```
## [1] "X...Genomes.in.Tree" "Full.Genome.Names"   "GenBank.Accesion"
## [4] "Bioproject"          "Isolation"
```

Construction of a Phylogenetic Tree (GTO-Tree)

All genomes were aligned using GToTree.
GTO-Tree was used to construct a phylogenetic tree based on specified HMM profiles. In this case the alignment of the 34 FASTA files was done using a selection of 251 HMMs for Cyanobacteria.
The resulting **.treefile** was opened in **FigTree (v1.4.4)** to form the final phylogenetic tree which was rooted using the distant ancestor strain **Gloeoceobacter violaceaous PCC 7421**.

A list was constructed with all the fasta names (remove spaces or strange characters).

```
guowei <- read.csv("/Users/mrblab/Desktop/Guowei/guowei_list.csv")
head(guowei)
```

```
##             Acaryochloris_MBIC11017.fasta
## 1     Anabaena_Trichomus_ATCC_29413.fasta
## 2         Atelocyanobacterium_ALOHA.fasta
## 3      Atelocyanobacterium_SI064986.fasta
## 4           Coleofasciculus_PCC7420.fasta
## 5               Crocosphaera_WH8501.fasta
## 6 Cyanothece_crocosphaera_ATCC51142.fasta
```

From this list and fasta files gtotree was run. We selected (-H) the HMM profile for Cyanobacteria (251 HMMs).
The resulting alignment from GToTree can be viewed in the software program **Geneious**.

```
#Alignment of FASTA files using a selection of 251 HMMs for Cyanobacteria (24 threads)
conda acctivate gtotree
GToTree -f list.csv -H Cyanobacteria -j  24 -o Tree
#View the alignment in geneious (.faa)
```

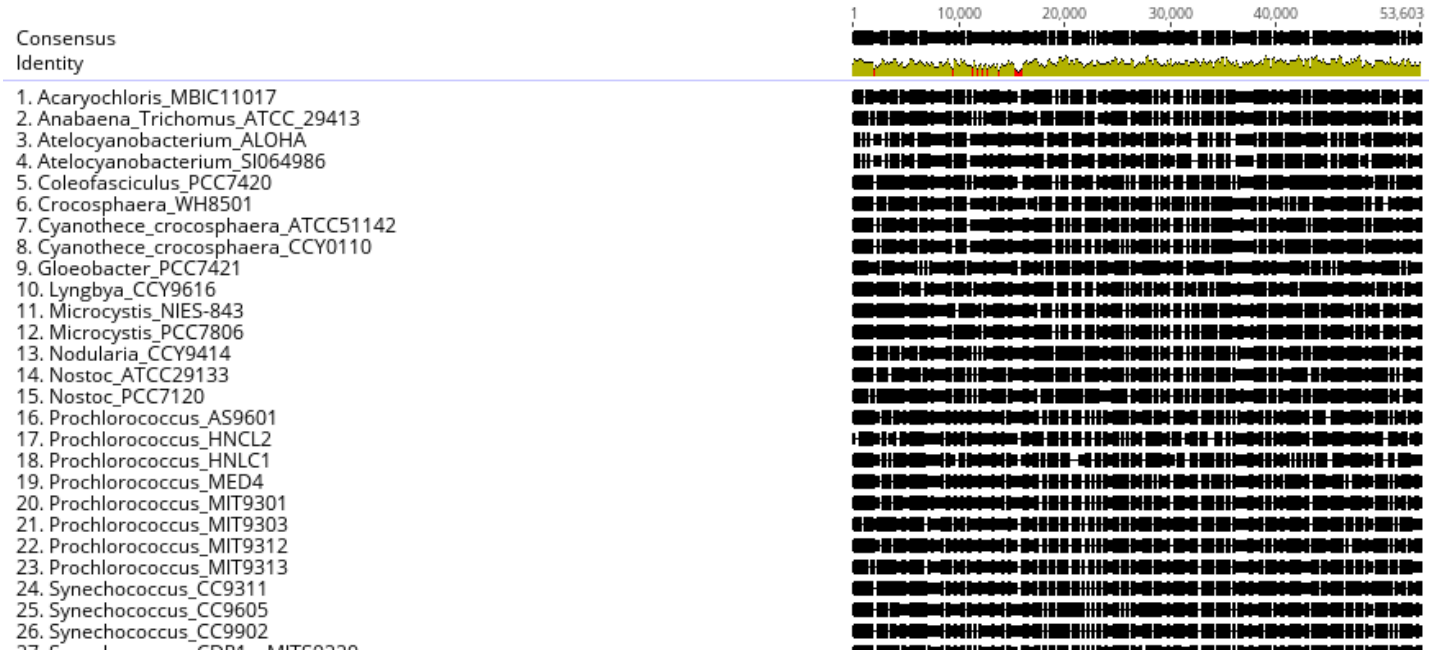

Figure 1.

From this alignment we can see several gaps which are hypervariable regions within the 251 HMMs. The program **gblocks** trims and truncates a multiple sequence alignment to obtain a better alignment. This results in a more reliable region from which to compare evolutionary rates (as is the case for tree building). Gblocks is relatively old (so newer more advanced software can be used for this).

```
#Trimming/ Truncate alignment for better alignment using Gblocks - old, find new softwar
e?
gblocks (interactive)
#Formation of (.faa-gb) file
```

Figure 2.

Finally from here we can conduct the phylogenetic tree using **IQtree** which form a NEWICK tree file (.treefile) that can be visualized by tree viewer programs such as **FigTree (v1.4.4)** where we form the final phylogenetic plot that is rooted by the phylogenetically different Gloeceobacter cyanobacteria strain.

```
#Phylogenetic tree construction (1000 bootstraps, 24 threads)
iqtree2 -s *-gb -o RiftiaPh051 -alrt 1000 -bb 1000 -nt 24

#Open (.treefile) in FigTree v1.4.4 to form your plot - root with distant ancestor strai
n (Gloeoceobacter).
```

Overview of the Comparison Analysis

Several different comparison methods were used to address the presence or absence of certain pathways

- **Fe-Genie search** - gives a general overview of Fe-related pathways present in each strain. Used for confirmation.
- **HMM search** - presence or absence is based on a specific pfam domain.
  HMM searches were used to decide the presence or absence of **NRAMP, ZIP, FTR1, NIS-type siderophores, FeoA,B, TonB dependent transport**.
- **BLAST search** - sequence identity (>200 bitscore) matches to genes of interest. Is not useful when a specific gene shows a large diversity between samples.
- **AntiSmash** - identifies secondary metabolites and is important for identifying siderophores and **NRPS-like pathways**.

General Information

- **Habitat**
  The habitat of each strain was recorded as the location where each strain was isolated from as mentioned in the bioproject.

- **Nitrogen Fixation**
  N2 fixation is based on the presence of a NifH gene. The Pfam (HMM) domain **PF00142** represents the presence of the nifH domain, however this domain is also found in the chloroplast encoded ChlL/ frxC. As cyanobacteria are photsynthetic, inherently, **all** cyanobacteria will be positive for this domain. N2 fixation was instead dependent on a <200 e-value score of NifH using BLAST.

FeGenie Analysis of Fe-related Pathways

FeGenie is a program that prints out Fe-related pathways for genome batches and was used to verify a lof of our HMM and Blast searches. It was also used to select which genomes to upload into **ANTISMASH 5.0** to further look into the siderophore biosynthesis pathways.

```
conda activate fegenie
FeGenie.py -bin_dir /Users/Cockeydooder/Desktop/FASTA_Guowei -bin_ext fasta -out /Users/
Cockeydooder/Desktop/FASTA_Guowei/FeGenie_output --makeplots
```

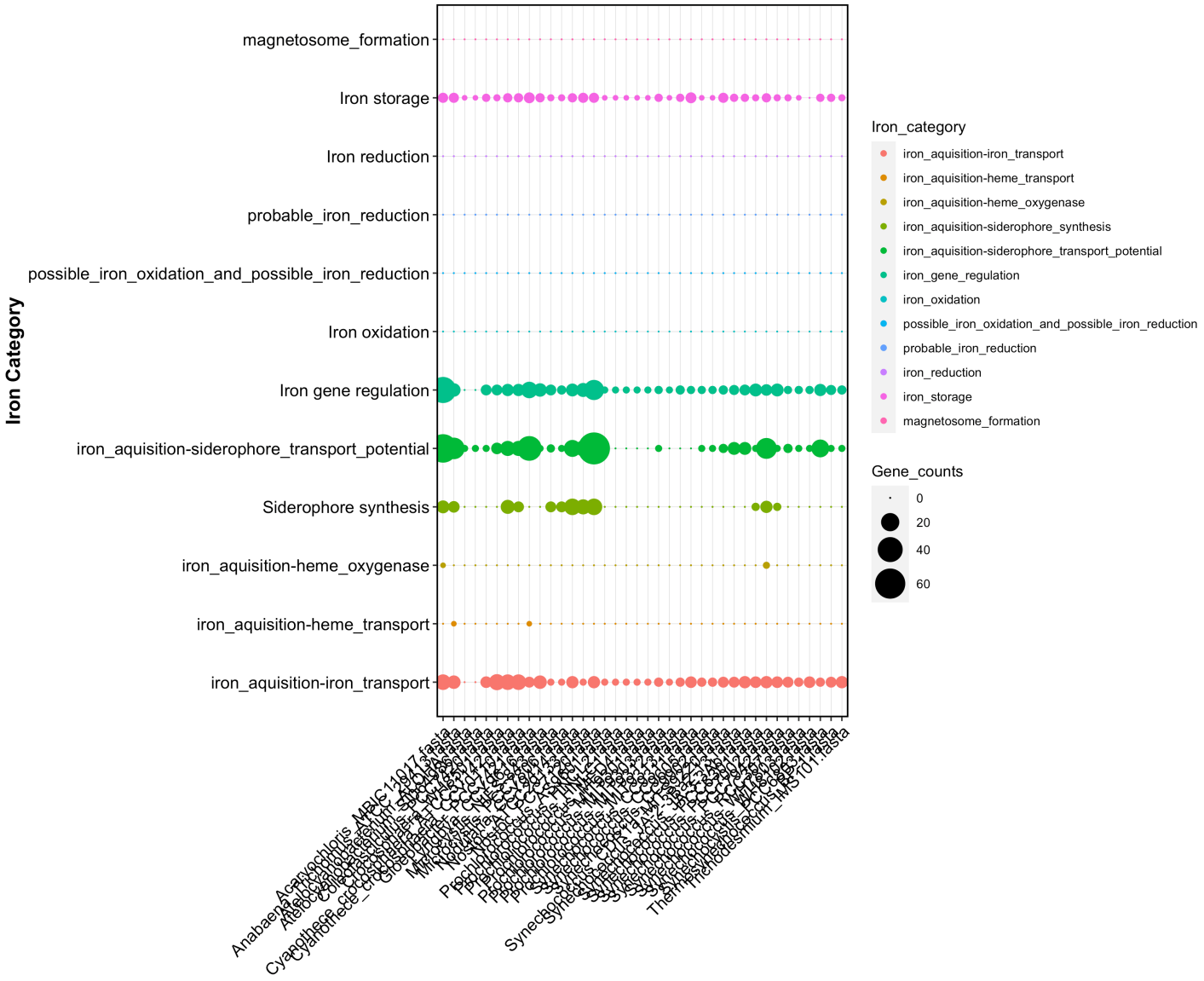

Figure 3.

The figure shows that from most cyanobacteria, big potential differences between strains lies in their ability to pick up Fe via specific TONB-dependent transporters. The specificity of TonB-dependent transporters can not be verified in this analysis and remains putative (e.g. TonB-dependent transport is also required for cobalamin uptake).

HMM Analysis of Fe-related Pathways

HMM analysis, to some degree, provides a relatively more robust way to locate the presence (or absence) of a particular pathway in comparison to a BLAST search. Fasta files (contain DNA sequences) and are translated into AA-sequences using Prokka which will be needed to conduct a HMM search. Prior to performing Prokka in batch, the file name (genome identity) of each strain needed to be placed in the scaffold in order to be able to identify it later on in the results.

Prokka result in the formation of several **.faa** files.
These .faa files are then **concatenated** into one giant .faa file.
This concatenated.faa file subsequently undergoes **HMMscan**.
The result is the formation of a **hmm.tsv** file which can now be searched using the **grep** function in the terminal.

Obtaining a hmm.tsv file for all strains using Prokka and HMM Scan

```
#renaming scaffolds (1,2,3)
for file in *.fasta; do tag=${file%.fasta}; awk '/^>/{print ">" ++i; next}{print}' < "$t
ag".fasta > ./rename/"$tag".fa; done

#adding file name to the scaffold
for file in *.fa; do fname="${file##*/}";  awk '/>/{sub(">","&"FILENAME"_");sub(/\.fa/,
x)}1' "$file" > ./rename/"$file"; done

#Prokka for batch genomes
for file in *.fa; do tag=${file%.fa}; prokka --prefix "$tag" --locustag "$tag" --increme
nt 10 --cpus 12 --mincontiglen 200 --outdir "$tag"_prokka --force --addgenes --gcode 11
"$file"; done

### the resulting .faa files are then concatenated into one big file.
cat ./*/*.faa > contactenated.faa

#### hmmscan is then conducted against Pfam databse (hmmscan [-options] <hmmdb> <seqfile
>)

hmmscan --tblout hmm.tsv --cut_ga --cpu 48  /media/bioinf/Data/pfam/Pfam-A.hmm contacten
ated.faa

#The result is the formation of a hmm.tsv file as output of all HMMS in one list.
# the hmm.tsv file can now be scanned for the presence of specific pfam domains from the
following command:
grep -c PF00593 /Users/Cockeydooder/Desktop/guowei_hmm.tsv # number of hits
    grep PF00593 /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Coc
keydooder/Desktop/Guowei_PF00593_hits.tsv # selects several columns as printed output (s
ave)
```

Fe-porin

slr1908 protein present in cyanobacteria (see Guowei).
Table S3: A unique porin meditates Fe-selective transport through cyanobacterial outer membranes, supplementary info).

TONB-transporters

There are multiple TonB receptors and additional components of the TonB-complex:

- **PF00593**
- **PF07660**
- **PF07715**
- **PF03544**
- **PF13103**
- **PF01618**
- **PF02472**

YfeABCD (similar to FecA - Ferric-citrate uptake)

- **PF01032**
- Mn and Fe2+ uptake

The presence of TonB-dependent transport was further confirmed with FeGenie.

```
#TBDT
grep -c PF00593 /Users/Cockeydooder/Desktop/guowei_hmm.tsv #103 hits
    grep PF00593 /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Coc
keydooder/Desktop/Guowei_PF00593_hits.tsv

grep -c PF07660 /Users/Cockeydooder/Desktop/guowei_hmm.tsv #17 hits
    grep PF07660 /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Coc
keydooder/Desktop/Guowei_PF07660_hits.tsv

grep -c PF07715 /Users/Cockeydooder/Desktop/guowei_hmm.tsv #103 hits
    grep PF07715 /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Coc
keydooder/Desktop/Guowei_PF07715_hits.tsv

grep -c PF03544 /Users/Cockeydooder/Desktop/guowei_hmm.tsv #27 hits
    grep PF03544 /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Coc
keydooder/Desktop/Guowei_PF03544_hits.tsv

grep -c PF13103 /Users/Cockeydooder/Desktop/guowei_hmm.tsv #12 hits
    grep PF13103 /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Coc
keydooder/Desktop/Guowei_PF13103_hits.tsv

grep -c PF01618 /Users/Cockeydooder/Desktop/guowei_hmm.tsv #55 hits
    grep PF01618 /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Coc
keydooder/Desktop/Guowei_PF01618_hits.tsv

grep -c PF02472 /Users/Cockeydooder/Desktop/guowei_hmm.tsv #49 hits
    grep PF02472 /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Coc
keydooder/Desktop/Guowei_PF02472_hits.tsv
```

Ferrous uptake: FeoABC

FeoB Review in bacteria: **Lau et al., 2015**
https://pubmed.ncbi.nlm.nih.gov/26684538/ (https://pubmed.ncbi.nlm.nih.gov/26684538/)
Ferrous iron (Fe2+) is more abundant under anaerobic conditions or at low pH in comparison to ferric Fe (Fe3+).

The Ferrous uptake system FeoABC (together with the EfeUOB - present in pathogenic bacteria) is the only known uptake system solely dedicated to ferrous Fe uptake. It was first discovered in E. coli K12 (Hantke, 1987), where:

- **FeoA** is a hydrophillic receptor
- **FeoB** is a ferrous permease
- **FeoC** is considered to be a ferrous receptor within the cytoplasm

In this study FeoABC was considered present if a FeoA and a FeoB domain was present within the genome:

- **FeoA (PF04023)** (used)
- **FeoB-C (PF07664)** (used)
- FeoB-N (PF02421) (not used) was too variable with 365 HMM hits.
- FeoB-Cyto (PF17910) (not used) too few hits present with only 11 hits.
- FeoC (YhgG) was only present in y-proteobacter. Has an Fe-sensing function.

The presence of FeoA and B was confirmed with matching hits from FeGenie.
Synechococcus JA-3-3Ba and Synechococcus JA-3-3Ab were putative due to the presence of a FeoB but not a FeoB domain.

```
#FeoA and B (present = FeoA + 2x FeoB) --> PF02421 was too variable, PF17910 was not alw
ays present. (confirmed with FeGenie)

grep -c PF04023 /Users/Cockeydooder/Desktop/guowei_hmm.tsv #31 hits
    grep PF04023 /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Coc
keydooder/Desktop/Guowei_PF04023_hits.tsv

grep -c PF07664 /Users/Cockeydooder/Desktop/guowei_hmm.tsv #20 hits
    grep PF07664 /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Coc
keydooder/Desktop/Guowei_PF07664_hits.tsv

grep -c PF02421 /Users/Cockeydooder/Desktop/guowei_hmm.tsv #365 hits
    grep PF02421 /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Coc
keydooder/Desktop/Guowei_PF02421_hits.tsv

grep -c PF17910 /Users/Cockeydooder/Desktop/guowei_hmm.tsv #11 hits
    grep PF17910 /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Coc
keydooder/Desktop/Guowei_PF17910_hits.tsv
```

General Metal Uptake Transporters

NRAMP2

- **PF01566**
- natural resistance-associated macrophage protein (ferrous uptake)
- low e-values observed for Synechococcus and Prochloroccus

ZIP

- **PF02535**
- zinc transport proteins and many putative metal transporters(ferric uptake)
- presence and absence

FTR1

- **PF03239**
- membrane permease which translocates generated ferric iron
- presence and absence

```
#NRAMP2 --> low e-values for Synechococcus and Prochloroccus
grep -c PF01566 /Users/Cockeydooder/Desktop/guowei_hmm.tsv #13 hits
    grep PF01566 /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Coc
keydooder/Desktop/Guowei_PF01566_hits.tsv

#ZIP --> presence and absence
grep -c PF02535 /Users/Cockeydooder/Desktop/guowei_hmm.tsv #8 hits
    grep PF02535 /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Coc
keydooder/Desktop/Guowei_PF02535_hits.tsv

#FTR1 --> presence and absence
grep -c PF03239 /Users/Cockeydooder/Desktop/guowei_hmm.tsv #16 hits
    grep PF03239 /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Coc
keydooder/Desktop/Guowei_PF03239_hits.tsv
```

Ferric Uptake: FUTABC

The Fut ferric uptake system is related to the Sfu/Fbp family of iron transporters. It was first found in Synechocystis sp.PCC6803. While originally thought to be a ferric-uptake system, FutA1 and FutA2 preferentially binds to ferrous rather than ferric Fe and is therefore not exclusively a ferric uptake system. Strangely, an FeGenie analysis did not result in a positive hit for Synechocystis sp. PCC6803 genome.

- **FutA1 (slr1295):** ferric binding periplasmic receptor.
- **FutA2 (slr0513):** ferric binding periplasmic receptor.
- **FutB (slr0327):** ferric permease (allows Ferric Fe to pass through the membrane)
- **FutC (sll1879):** membrane associated ATPase
  For this analysis, a ferric uptake mechanism was considered present if a FutB HMM domain was found, and putative if only a receptor (FutA1, FutA2 - may be linked to ferrous uptake instead) or an ATPase (FutC) was found.
- A BLAST analysis however of the FUTABC genes present in Synechocystis sp. PCC6803 highlighted that all cyanobacteria contained a FUT-ABC-like pathway.
- Use FeGenie (presence of A+B or B+C - putative if only FutB is present).

FASTA FILES FOR FUTABC

“**>FutA1_slr1295**
MVQKLSRRLFLSIGTAFTVVVGSQLLSSCGQSPDAPIADTPGEQQEINLYSSRHYNTDNE
LYAKFTAETGIKVNLIEGKADELLERIKSEGANSPADVLLTVDLARLWRAEEDGIFQPVQ
SEILETNVPEYLRSPDGMWFGFTKRARVIMYNKGKVKPEELSTYEELADPKWKGRVIIRS
SSNEYNQSLVASLVVADGEESTLAWAKGFVSNFAREPQGNDTAQIEAVSSGEADLTLANT
YYMGRLLESEDPAQKAIAENVGVFFPNQEGRGTHVNVSGVGVVKTAPNREGAVKFIEFLV
SEPAQAFLAQNNYEYPVLAGVPLNKSVASFGEFKSDTTSLDKLGPALAPATKIMNEAGWK”

“**>FutA2_slr0513**
MTTKISRRTFFVGGTALTALVVANLPRRASAQSRTINLYSSRHYNTDDALYDAFGEVNLI
EASAEELIERIQSEGANSPGDILFTVDAGMLWRAEQAGLFQPVRSGKLNERIPENLRHPD
GLWYGFTQRARVLYYSRDRVNPADLSTYEALADPQWRGKILVRPSSNVYNLSLTASRIAI

HGEPETRRWLQGLVGNFARQPEGNDTAQIRAIAAGIGDVAIANSYYYIRLQKSTDPADQE
VVEKVSLFFPNTGSGERGTHVNVSGAGVLKNAPNRDAAIAFLEYLASDDAQRYFAEGNNE
YPVIPGVPIDPVLAAHGQLKGDPLNVSNLGRYQPDSARLMNEVGWQ”

“**>FutB_slr0327**
MFNFLTVLPSPPKVLLNFWVLTSLLIAVWIAVPVIFVFLGIFSWQGEIFSHLWATVLGEY
IRNSLALMLGVGAGVFVLGVGTAWLVTMCRFPGCRWLEWALLLPLSAPAYLLAYGYSNLL
DFYGPVQTLLRSIFGWQSAPEYWFPQIRSLWGAIAILALVSYPYVYLLARIAFLEQGVCT
LEASRSLGCNPWQSFSRVALPLARPAIAAGLALVMMETLNDFGTVQYFGVNTFTTGIYST
WFGFGERQGATQLAAFLMIFVFLLVVLERWSRRQAKFYQSSSPHQNLPRYQLRGLRAIGA
LAFCLFPFLLGFLIPASYLLYLTVSYAQEVRNNNFFQLASHSLILSFLTAAIALVIGLIL
VYGQRLSRQPLTSFAVKVASMGYAIPGSVIAVGVLIPAGNFDNWLADWWENMWGVKIGLL
LSGTIAILVYAYLVRFLAVALGSLEGSLGKIKPTLDDAARSLGKSPSQILWQVHTPLMTG
GLLTAVMLVFVDVMKELPATLVIRPFNFDTLAIRVYQYASDERLIEAAAPALTIILAGML PVIFLSVQIARSRPSEG”

“**>FutC_sll1878**
MTVAQFSPVARLSIEDSVLTVQDLGKSFRGQSTPVLQKINFNLAPGEILGLLGPSGCGKT
TLLRIIAGFETPTSGTVHLEGDCVSGENGLTPPEQRQTGMVFQDYALFPHLTITDNIAFG
LRHKSQKLNRQQIQGRVAEVLHLVGLTGLEKRYPHELSGGQQQRIALARALAPKPNLILL
DEPLSNLDVQVRQRLRHEIRHILKATGTAAIFVTHDQEEAMAISDRIGVMYRGNLEQIGT
PEEIYRSPASRFVAEFVTQANFVPAQRQGTLWATEFGQWPLTFQGIQPELPSVGELMLRE
EEIELSPASDGPVVIRDRQFLGREYRYCLETPAGRQIHARTSLQTVIPVGSRVNLTPTNP CPPLFAQG”

BLAST Script

```
#concatenate all fasta files to form one giant fasta file
cat *.fasta > guowei_database.fasta
#construct your database from your giant fasta file
makeblastdb -in guowei_database.fasta -dbtype nucl -out guowei_db
#blast (protein sequence to nucleotide file) for multiple genes
ls *.fasta | parallel -a - tblastn -query {} -db guowei_db -out {.}.tsv
#alternatively - save each gene of interest in a separate fasta file (for a better overv
iew).
ls test_guowei.fasta | parallel -a - tblastn -query {} -db guowei_db -out {.}.tsv
```

Results: BLAST shows that all cyanobacteria have a FUT-like system (multiple hits for FUTABC genes).

Siderophores

- **NIS-type siderophores**
  PF04183 represents the presence of an IucA/IucC domain
- **NRPS-type siderophores**
- all NRPS siderophores are putative. NRPS-like pathways are, for example, also required for the production of toxin (as is the case for Microcystis and Crocosphaera).

```
#SIDEROPHORES --> IucA/IucC domain, all NRPS are putative (confirmed with ANTISMASH)
grep -c PF04183 /Users/Cockeydooder/Desktop/guowei_hmm.tsv #6 hits
    grep PF04183 /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Coc
keydooder/Desktop/Guowei_PF04183_hits.tsv
```

Fe-reduction (ARTO genes)

ARTO consists of 3 subunits:

- **PF00115:**
- Subunit I is indistinguishable from photosynthesis
- **PF00116, PF02790:**
- subunit II
- **PF00510:**
- subunit III.

Other FASTA files

"**>FeoB_slr1392**
MVSHCQRGSVQSSRPDVKKRVAFIGQPNTGKSTFFNRITKANAAIANWPGLTVDLFRAVV
PLQGELIEFVDLPGIYDLNGFSEDERVVQRFLANYAVNLVVVVVNAAQIDRQIRLLLQVQ
TLGIPAITLLNLADEAKRYGVQIDVAALQERLGLPLYPISAKYGTGCSRAMDAIGRAVKD
QPEAYQIPNLVNVLSDHPVAIADMETALAGVVQMPSPNARTLTNVIDGVMLHPVFGLPIF
FASMFGVFWVIWHVGLPSADPVDAVTGWVQSNILEPLFSPLPTILQGLLLDGIWTGFAAL
LSFVPLVAIFFIVMGILEGSGYLSRAAYLMDALMGRLGLDGRSFVLQMMGFGCNVPAIMG
TRVMRSRGMRLLSMLVIPFSLCSARLQVFVFILAAVMPGTQGAIALFLLYLMSFVAAFTV
AAILSRFHYFQARDPFVLELPPYRLPTFKQVFLRVWGEMREFVARLSMFMVIGSSLIWFL
TSFPQGSTGLETFAGRIGSVFQPLMNPLGINPFLTISLIFGFVAKEVQIAALTVIYGLNN
SEAVSDQIHSTVTFAQGFSYCLFSLIYIPCLTTLGAIWGESKSLAYTAISVATPLVTAWL FSFIFYQSFSWLGW

“**>FTR1_slr0964**
MDFASGLPIFIVTLREALEASLVVGIVLACLARAQQMQLKGWVYRGISAGVVASVLVGCL
LAGVLQGVERLPGPYTPILKALLAALLGAIAVGMLSWMLLWMTKQARSLRGEIQGQINQA
VEKEGGGKAIAIVVFIAVVREGFEMVLFLAAQQNMANPAAIGAALAGIGTAVVMAFLIFR
LGVKLNLKLFFQVMGTLLLIIVGGLVIGVLKNLDLAVSMMGLANLGLGYLCFVPGDSCLL
GPLLWNLAPWLPDNQFPGIVLKTLAGYRDHLYLFQAIAYGIFLSVIGSLYFRGLAGKGDA PQAVAQKS”