Guowei Review (bioinformatic analysis)

Coco Koedooder

Cyanobacteria FASTA Files

Phylogenetic tree of 34 representative and diverse cyanobacterial strains and the presence and absence of several different Fe-uptake pathways. The figure includes additional information on the habitat they were isolated from, whether a strain fixes nitrogen (thus having a higher Fe-requireiment), and if they can synthesize siderophores. The tree is rooted at the genomically distinct Gloeobacter violensis PCC 7421. More information on its construction and analysis can be found at the following github repository link:

The tree was constructed from a concatenated gene-alignment of a 251 single-copy gene-set HMM for Cyanobacteria using GToTree that was truncated using gblocks and inserted into IQtree2 with 1000 bootstraps using the best-fit model LG+F+R6. The tree is rooted by the evolutionary distant Gloeobacter violensis PCC 7421.

Construction of a Phylogenetic Tree (GTO-Tree)

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"
```

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 (be sure to 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 csv list and associated 34 fasta files Gtotree was run. We selected (-H) the HMM profile for Cyanobacteria (contains 251 HMMs). The resulting alignment from GToTree can be viewed using 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)
```
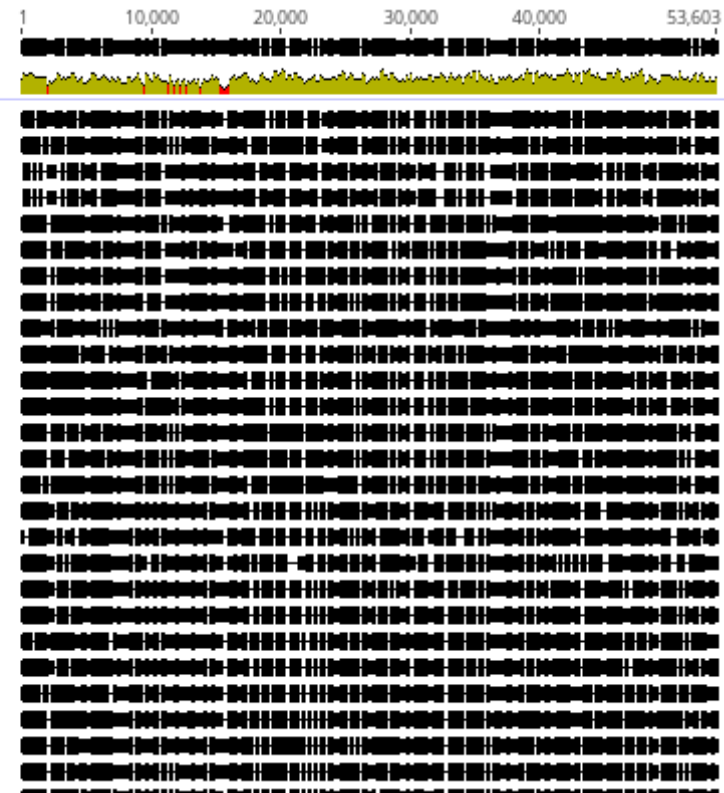
Figure 1. GTO-Tree not truncated.

From this alignment we can see several gaps which represent hypervariable regions within the 251 aligned HMMs. The program **gblocks** trims and truncates a multiple sequence alignment, resulting in a more reliable region from which to compare evolutionary rates (as is the case for tree building) and will (typically) result in a better allignment. Gblocks is a relatively old software (so newer and more advanced techniques may be used for this).

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

Figure 2. GTO-tree truncated

From this GTO-tree we can construct a phylogenetic tree using **IQtree** which forms 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 a phylogenetically/ evolutionarily distinct 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 strain (Gloeoceobacter).
```

Comparative Analysis (overview)

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

- **BLAST search** - Using sequence identity (and cut-off value) to select for the presence or absence of a gene of interest. BLAST search is not useful when a specific gene shows a large diversity between samples.
- **HMM search** - presence or absence of a specific **protein-family** (Pfam) domain. HMM search is useful when a gene of interest contains a relatively unique and conserved domain that can be used to identify the presence of a certain function. It can, to some degree, provide a relatively robust way to locate the presence (or absence) of a particular gene of interest that can have a large diversity between samples.

- The analysis was conducted with the help of the **Transporter Classification Database (TCDB)** which is a curated reference database for transport protein research. The database can provide both FASTA sequences (BLAST search) and PFAM codes (HMM search) for all transport related families. The TCDB database was hereby used for the selection of both FASTA and PFAM analysis for each Fe-uptake system.
  https://pubmed.ncbi.nlm.nih.gov/26546518/

- A Blast search was used to decide the presence or absence of:
  NifH, FutABC, Ftr1, ARTO

- An HMM search was used to decide the presence or absence of:
  NRAMP, ZIP, FTR1, ARTO, NIS-type siderophores, FeoA,B, TonB dependent transport.

Additionally:

- **Fe-Genie search** - Software that provides a general overview of Fe-related pathways present in each strain. Used for confirmation.

- **AntiSmash** - identifies secondary metabolites and is important for the identification of siderophores (NRPS-like pathways).

BLAST preparation and script

Save the FASTA file (AA) of the gene of interest (Gene_name.faa). Contruct a database of the 34 fasta file that your sequence of interest can be compared to.

```
#concatenate all fasta files to form one giant fasta file
cat *.faa > guowei_database.faa
#construct your database from your giant fasta file
makeblastdb -in guowei_database.fasta -dbtype prot -out guowei_db #prot or nucl
#blast (protein sequence to nucleotide file) for multiple genes
ls *.faa | parallel -a - blastp -query {} -db guowei_db -out {.}.tsv #protein blast
#alternatively - perform each separately.
ls test_guowei.faa | parallel -a - blastp -query {} -db guowei_db -out {.}.tsv
```

HMM preparation and script

Fasta files (contain DNA sequences) 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.

```
#renaming scaffolds (1,2,3)
for file in *.fasta; do tag=${file%.fasta}; awk '/^>/{print ">" ++i; next}{print}' < "$tag".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" --increment 10 --cpus 12 --mincont
```

```
iglen 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
```

HMM-scan

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

hmmscan --tblout hmm.tsv --cut_ga --cpu 48  /media/bioinf/Data/pfam/Pfam-A.hmm contactenated.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 PF_number /Users/Cockeydooder/Desktop/guowei_hmm.tsv # number of hits
    grep PF_number /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Cockeydooder/Desktop/Guowe
i_PFnumber_hits.tsv # selects several columns as printed output (save)
```

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/FAST
A_Guowei/FeGenie_output --makeplots
```
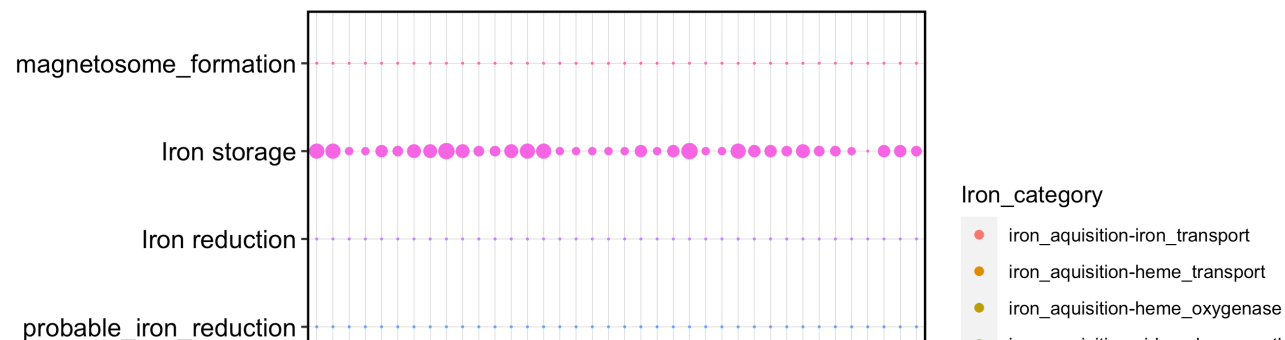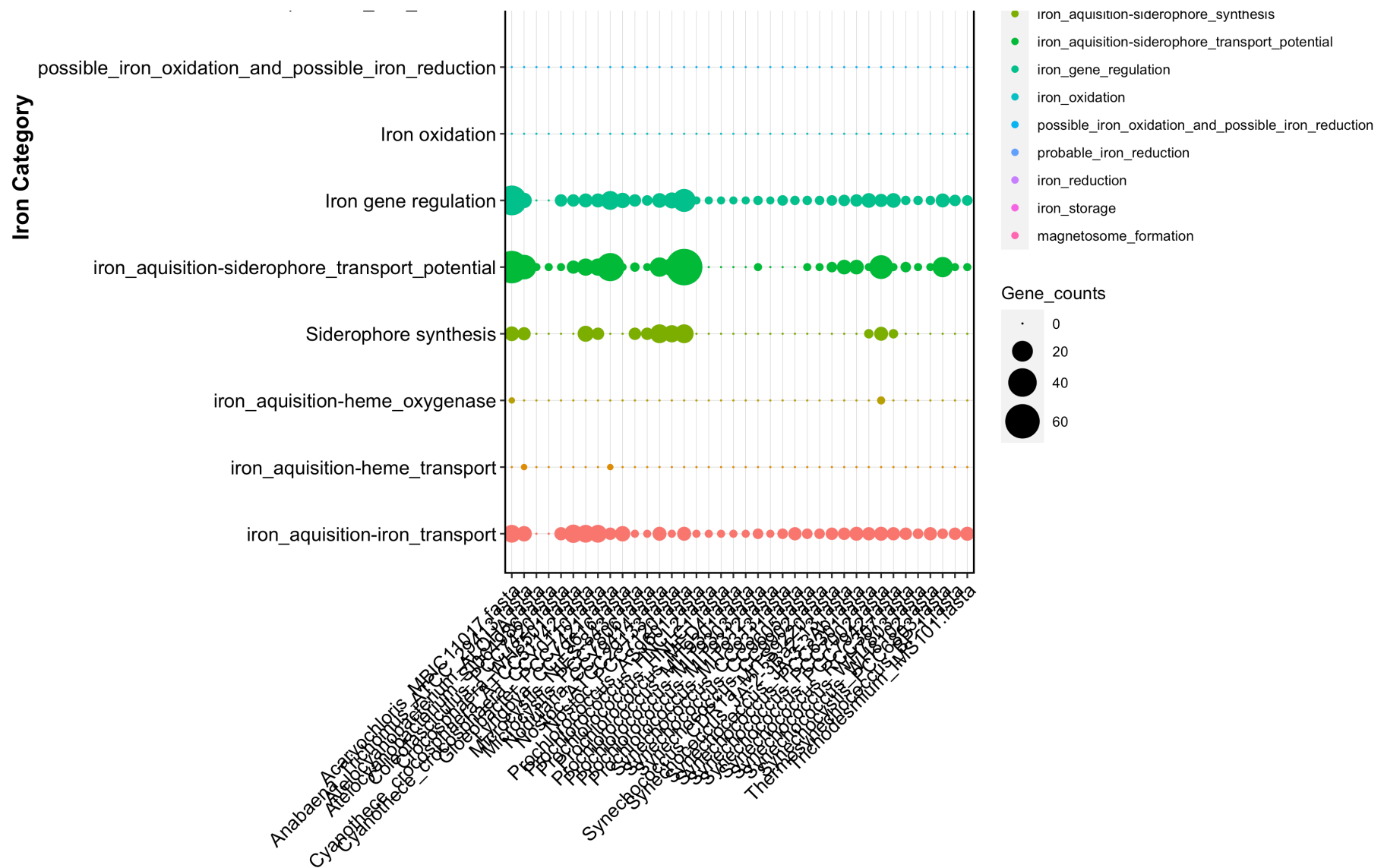
Figure 3.

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).

Analysis and Results:

1. General Information:

(Habitat, N2 fixation, Siderophore-biosynthesis)

- **Habitat**
  The habitat of each strain was recorded as the location where each strain was isolated from as mentioned in the bioproject and was confirmed with the literature. Habitats were subdivided according to extreme/thermal (cream), fresh water (green), and marine (purple). Marine habitat is further subdivided into open (dark purple) and coastal (light-purple). An asterisk denotes that the sample was collected from sediment.

- **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 ChIL/ frxC. As cyanobacteria are photosynthetic, inherently, **all** cyanobacteria will be positive for this domain. N2 fixation was instead dependent on ablast search of NifH - and verified within the literature (Latysheva et al., 2012).
  https://academic.oup.com/bioinformatics/article/28/5/603/247739

2. Fe-selective porin

All cyanobacteria strains contain porins.

slr1908 proteins represent a special Fe-selective porin that is present within a distinct monophyletic clade. It was hereby presumed that all strains within this clade that contain an slr-1908-like protein can perform a similar function which has been proved through knock-out and complementary experiments performed in Synechocysistis PCC 6803 (see Qui et al., 2020).

>slr1908

MNKLTSHLLKLFPLALGSSLAIVPGAMAQSTGELATPGDFPRISNQGDSLELMRRRQNAG
TFNAATPDITDMSQVTSVSELRDVQPTAWAYEALKSLVERYGCIVGYPDRTFRGDRALSR
WEFAAGLNACMNVMERLIQENVAVLREDIDKLKRLMQEFEAELAALGARIDNLETRTSFL
EDHQFSTTTKLNGVAVFALVDQWGGDKAVDWRQQDNIDNFGAAAPAPVEENATLSSRVRL
NFDTSFTGKDLLRTRLQAGSVPNLSGPTGTNMARLSFDGSSPDNNVDINKLFYRFPMGNL
TTWIGGTGLALDDVFKTYNPYLESGDSGALSRFSRYSPFVNRGPEGTGGALRYKFNDVFT
VSAAYLADTGQASTPSDDVFTSGGNTFRSGNGFFNGSYSTGVQFDIKPVDNFSFGISYLH
KYYSQGDVNLTGSTGSRIASNPFYQAATTMDTYNLQATWQITDKFNLSGWFGYANATAQG
FNTGGNPQNRDGLGADLWTWNAALSVIDVFKEGAVLSLSGGLMPYAPYVGSLSGDRISND
RNSPYIIEAQYQFPVNKNIQITPGAYVILSPEANSNNSAIWVGVLRTTFKF

```
#Feporin

ls Feporin.faa | parallel -a - blastp -query {} -db guowei_db -out {.}.tsv
```

RESULTS: As all cyanobacteria contain porins some contain an Fe-selective porin which was identified in Synechocystis PCC 6803 (slr1908). A putative list was constucted - where all hits to slr1908 with an e-value of 0 were listed as present, while all hits within the same monophyletic clade were marked as putative.

3. TBDT-transporters

TonB

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

- **PF03544** TonB (used)
- **PF01032** FeCD (used) - but includes both B12 and Fe-citrate.

The TonB-ExbB-ExbD/TolA-TolQ-TolR Outer Membrane Receptor Energizers and Stabilizers (TonB/TolA) Family. **PF03544** was compared as it represents the well defined TonB/EbdB and indicates the potential of active transport into the cell of substrates requiring a TBDT (>500 da) which

may be Fe-siderophore related - from simple citrate molecules, to more complex heme - although it has been shown that Fe' can also be taken up by TonB-dependent transporters.

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

```
#TonB
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/Cockeydooder/Desktop/Guowei_
PF03544_hits.tsv

    #FeCD
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/Cockeydooder/Desktop/Guowei_
PF03544_hits.tsv
```

FecA - Ferric-citrate uptake)

- **PF01032**
- Putative Ferric-citrate uptake
- Majority of TonB-containing cyanobacteria also contain a positive hit for FecCD pfam domain for Fe-citrate uptake.

RESULTS: More TonB-dependent transporter systems are present in comparison to Siderophore-producing systems indicating the siderophore production is rarer than its potential uptake by cyanobacteria. The majority of Ton-B containing cyanobacteria also contained a Ferric-citrate transporter (not shown). Interestingly both UCYNA produce NIS-siderophores but do not contain a TonB-dependent transporter system..

4. Ferrous uptake: FeoABC

FeoB Review in bacteria: **Lau et al., 2015**
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).

FeoABC is marked present based on PFAM domains - where a FeoA and a FeoB domain was located within the genome. The presence of FeoA and B was confirmed with matching hits from FeGenie.

FeoA (PF04023) (used) is a hydrophillic interacting with FeoB
FeoB (PF07664)(PF07670)(PF01926) is a ferrous permease, the main transporter component where:
* PF07664 - FeoB C-terminus (used)
* PF07670 - Membrane Pore gate (but also nucleoside transporter) - too general
* PF01926 - 50S ribosome-binding GTP-ase - too general **FeoC** is considered to be a ferrous receptor within the cytoplasm

```
#FeoA and B (present = FeoA + 2x FeoB) --> PF02421 was too variable, PF17910 was not always present. (confirmed w
ith 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/Cockeydooder/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/Cockeydooder/Desktop/Guowei_
PF07664_hits.tsv
```

Results: Quite a few cyanobacteria take up Fe(II) specifically via the FeoAB uptake system - which is largely absent for the more streamlined Prochlorococcus and Synechococcus strains. Nostoc sp. PCC 73102, Synechococcus JA-3-3Ba and Synechococcus JA-3-3Ab were putative due to the presence of only FeoA (nosotoc) or FeoB (synechococcus) but not the other domain. Interestingly, several cyanobacteria contained multiple FeoAB-like domains.

5. General Metal Uptake Transporters

NRAMP2

- **PF01566**
- Natural Resistance-Associated Macrophage Protein (ferrous uptake)
- The Metal Ion (Mn2+-iron) Transporter (Nramp) Family.

```
#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/Cockeydooder/Desktop/Guowei_
PF01566_hits.tsv
```

ZIP

- **PF02535**
- The Zinc (Zn2+)-Iron (Fe2+) Permease (ZIP) Family
- Zinc transport proteins and many putative metal transporters(ferric uptake)

```
#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/Cockeydooder/Desktop/Guowei_
PF02535_hits.tsv
```

Results: Several Prochlorococcus and Synechococcus strains take up Fe(II) via a general divalent metal uptake system rather than the Fe(II) specific FeoAB uptake system (Kranzler et al., 2013). Only A few cyanobacteria can take up Fe through a ZIP permease. .

6. Ferric Uptake: FUTABC

The Fut ABC ferric uptake system is related to the Sfu/Fbp family of iron transporters and was first discovered in Synechocystis sp.PCC6803. It belongs to the ATP-Binding Cassette (ABC) superfamil and is composed of several components that are not distinct enough to allow for a specific distinction via PFAM IDs (as is the case for the Feo Fe-uptake system).
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 - therefore, the analysis was conducted by blasting against the curated FutABC FASTA sequences from the TCDB-database. A BLAST analysis was hereby conducted on FutA1, FutA2, and FutB. The ferric uptake system was considered present if a FutB HMM domain was found, and putative if only a match to a receptor (FutA1, FutA2) was found.

- **FutA1 (slr1295):** ferric binding periplasmic recepto (PF01547)
- **FutA2 (slr0513):** ferric binding periplasmic receptor (PF01547)
- **FutB (slr0327):** ferric permease (allows Ferric Fe to pass through the membrane) (PF00528)
- **FutC (sll1879):** membrane associated ATPase (PF00005)(PF08402)

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

Results: FutABC is an important (and essential) Fe-uptake system for cyanobacteria. As shown in the bioinformatic analysis most cyanobacteria were shown to contain a FUT-like uptake system which coincided with the literature (Kranzler et al., 2013) where, for our representative strains, Trichormus sp ATCC 29413 does not contain this uptake system, while Prochlorococcus HNCL1 does not seem to contain a FutA receptor.

7. 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/Cockeydooder/Desktop/Guowei_
PF04183_hits.tsv
```

Results: Only a few strains of cyanobacteria produce siderophores, the majority are putative as they involve the presence of an NRPS-biosynthesis pathway which can also be required for the production of toxins (as is the case for microcystis and crocosphaera). More cyamobacteria contain TonB dependent uptake pathways, suggesting the possibility that they may take up xenosiderophores from the environment directly (rather than through photooxidation processes). Interestingly, UCYN are the only NIS-like siderophore biosynthesis pathway and do not contain a putative TonB uptake system.

8. Fe-reduction

ARTO

The alternative respiratory oxidase (ARTO) is a putative Fe-reducing protein that sits on the cell membrane (rather than the thylakoid membrane) and consists of 3 subunits (ctaCII, CtaDII, ctaDII).

- **PF00115:**
- Subunit I is indistinguishable from photosynthesis
- **PF00116, PF02790:**
- ctaCII - subunit II can be separated from ctaCI by the presence of a C[AT]ELC motif rather than a D[AS]X[FY]S motif.
- **PF00510:**
- subunit III.

The ctaCII - subunit II can be separated from ctaCI by the presence of a C[AT]ELC motif rather than a D[AS]X[FY]S motif. We conducted a blast search of Tery_0278 (the ctacII sequence of Trichodesmium erythreaum IMS101). All cyanobacteria containing more than one hit was considered to have an ARTO Fe-uptake system. This was confirmed by the presence of the C[AT]ELC motif via Interpro (Lea-Smith, 2013).

The BLAST search was further confirmed using a Pfam search of the COX2 (ctacII) domain PF00115 - where the presence of multiple COX2 genes indicated the possibility of an ARTO gene.

```
#BLAST ARTO
ls ARTO.faa | parallel -a - blastp -query {} -db

#HMM ARTO
grep -c PF00116 /Users/Cockeydooder/Desktop/guowei_hmm.tsv #16 hits
    grep PF00116 /Users/Cockeydooder/Desktop/guowei_hmm.tsv | cut -f 3,5,18 > /Users/Cockeydooder/Desktop/Guowei_
PF00116_hits.tsv
```

FTR1

- **PF03239**
- THE Fe/Pb transporter (ILT) Family
- High-affinity oxidase-dependent plasma membrane Fe2+-Fe3+ uptake transporter
- Found in Synechococcus (slr0964) to be part of an reduction reoxidation Fe-uptake pathway.
- The Ftr1 (integral membrane transporter)/ P19 (periplasmic iron binding protein) Fe uptake system acting as a memmbrane permease which translocates generated ferric iron.

>FTR1_slr0964
MDFASGLPIFIVTLREALEASLVVGIVLACLARAQQMQLKGWVYRGISAGVVASVLVGCL
LAGVLQGVERLPGPYTPILKALLAALLGAIAVGMLSWMLLWMTKQARSLRGEIQGQINQA
VEKEGGGKAIAIVVFIAVVREGFEMVLFLAAQQNMANPAAIGAALAGIGTAVVMAFLIFR
LGVKLNLKLFFQVMGTLLLIIVGGLVIGVLKNLDLAVSMMGLANLGLGYLCFVPGDSCLL
GPLLWNLAPWLPDNQFPGIVLKTLAGYRDHLYLFQAIAYGIFLSVIGSLYFRGLAGKGDA PQAVAQKS

```
#FTR1 -BLAST
ls Ftr1p.faa | parallel -a - blastp -query {} -db

#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/Cockeydooder/Desktop/Guowei_
PF03239_hits.tsv
```

Results: A large number of cyanobacteria strains contain putative Fe-reduction proteins (either FRTR1 or ARTO), where FTR1 is more prevalent by synechococcus and prochlorococcus and ARTO is more prevalent in other cyanobacteria.

Additional TCDB - FASTA files

>FeoA_E_coli_K12

MQYTPDTAWKITGFSREISPAYRQKLLSLGMLPGSSFNVVRVAPLGDPIHIETRRVSLVLRKKDLALLEVEAVSC

>FeoB_E_coli_K12

MKKLTIGLIGNPNSGKTTLFNQLTGSRQRVGNWAGVTVERKEGQFSTTDHQVTLVDLPGTYSLTTISSQTSLDEQIACHY
ILSGDADLLINVVDASNLERNLYLTLQLLELGIPCIVALNMLDIAEKQNIRIEIDALSARLGCPVIPLVSTRGRGIEALK
LAIDRYKANENVELVHYAQPLLNEADSLAKVMPSDIPLKQRRWLGLQMLEGDIYSRAYAGEASQHLDAALARLRNEMDDP
ALHIADARYQCIAAICDVVSNTLTAEPSRFTTAVDKIVLNRFLGLPIFLFVMYLMFLLAINIGGALQPLFDVGSVALFVH
GIQWIGYTLHFPDWLTIFLAQGLGGGINTVLPLVPQIGMMYLFLSFLEDSGYMARAAFVMDRLMQALGLPGKSFVPLIVG
FGCNVPSVMGARTLDAPRERLMTIMMAPFMSCGARLAIFAVFAAAFFGQNGALAVFSLYMLGIVMAVLTGLMLKYTIMRG
EATPFVMELPVYHVPHVKSLIIQTWQRLKGFVLRAGKVIIIVSIFLSAFNSFSLSGKIVDNINDSALASVSRVITPVFKP
IGVHEDNWQATVGLFTGAMAKEVVVGTLNTLYTAENIQDEEFNPAEFNLGEELFSAIDETWQSLKDTFSLSVLMNPIEAS
KGDGEMGTGAMGVMDQKFGSAAAAYSYLIFVLLYVPCISVMGAIARESSRGWMGFSILWGLNIAYSLATLFYQVASYSQH
PTYSLVCILAVILFNIVVIGLLRRARSRVDIELLATRKSVSSCCAASTTGDCH

>FeoB_slr1392

MVSHCQRGSVQSSRPDVKKRVAFIGQPNTGKSTFFNRITKANAAIANWPGLTVDLFRAVV
PLQGELIEFVDLPGIYDLNGFSEDERVVQRFLANYAVNLVVVVVNAAQIDRQIRLLLQVQ
TLGIPAITLLNLADEAKRYGVQIDVAALQERLGLPLYPISAKYGTGCSRAMDAIGRAVKD
QPEAYQIPNLVNVLSDHPVAIADMETALAGVVQMPSPNARTLTNVIDGVMLHPVFGLPIF
FASMFGVFWVIWHVGLPSADPVDAVTGWVQSNILEPLFSPLPTILQGLLLDGIWTGFAAL
LSFVPLVAIFFIVMGILEGSGYLSRAAYLMDALMGRLGLDGRSFVLQMMGFGCNVPAIMG
TRVMRSRGMRLLSMLVIPFSLCSARLQVFVFILAAVMPGTQGAIALFLLYLMSFVAAFTV
AAILSRFHYFQARDPFVLELPPYRLPTFKQVFLRVWGEMREFVARLSMFMVIGSSLIWFL
TSFPQGSTGLETFAGRIGSVFQPLMNPLGINPFLTISLIFGFVAKEVQIAALTVIYGLNN
SEAVSDQIHSTVTFAQGFSYCLFSLIYIPCLTTLGAIWGESKSLAYTAISVATPLVTAWL FSFIFYQSFSWLGW

>NRAMP_Escerichia coli

MTNYRVESSSGRAARKMRLALMGPAFIAAIGYIDPGNFATNIQAGASFGYQLLWVVVWANLMAMLIQILSAKLGIATGKN
LAEQIRDHYPRPVVWFYWVQAEIIAMATDLAEFIGAAIGFKLILGVSLLQGAVLTGIATFLILMLQRRGQKPLEKVIGGL
LLFVAAAYIVELIFSQPNLAQLGKGMVIPSLPTSEAVFLAAGVLGATIMPHVIYLHSSLTQHLHGGSRQQRYSATKWDVA
IAMTIAGFVNLAMMATAAAAFHFSGHTGVADLDEAYLTLQPLLSHAAATVFGLSLVAAGLSSTVVGTLAGQVVMQGFIRF
HIPLWVRRTVTMLPSFIVILMGLDPTRILVMSQVLLSFGIALALVPLLIFTSDSKLMGDLVNSKRVKQTGWVIVVLVVAL NIWLLVGTALGL

>ZIP_Escerichia coli

MSVPLILTILAGAATFIGAFLGVLGQKPSNRLLAFSLGFAAGIMLLISLMEMLPAALAAEGMSPVLGYGMFIFGLLGYFG

LDRMLPHAHPQDLMQKSVQPLPKSIKRTAILLTLGISLHNFPEGIATFVTASSNLELGFGIALAVALHNIPEGLAVAGPV

YAATGSKRTAILWAGISGLAEILGGVLAWLILGSMISPVVMAAIMAAVAGIMVALSVDELMPLAKEIDPNNNPSYGVLCG MSVMGFSLVLLQTAGIG

>FTR1_slr0964

MDFASGLPIFIVTLREALEASLVVGIVLACLARAQQMQLKGWVYRGISAGVVASVLVGCL

LAGVLQGVERLPGPYTPILKALLAALLGAIAVGMLSWMLLWMTKQARSLRGEIQGQINQA

VEKEGGGKAIAIVVFIAVVREGFEMVLFLAAQQNMANPAAIGAALAGIGTAVVMAFLIFR

LGVKLNLKLFFQVMGTLLLIIVGGLVIGVLKNLDLAVSMMGLANLGLGYLCFVPGDSCLL

GPLLWNLAPWLPDNQFPGIVLKTLAGYRDHLYLFQAIAYGIFLSVIGSLYFRGLAGKGDA PQAVAQKS

>TonB-ExbB_E.coli

MTLDLPRRFPWPTLLSVCIHGAVVAGLLYTSVHQVIELPAPAQPISVTMVTPADLEPPQAVQPPPEPVVEPEPEPEPIPE

PPKEAPVVIEKPKPKPKPKPKPVKKVQEQPKRDVKPVESRPASPFENTAPARLTSSTATAATSKPVTSVASGPRALSRNQ

PQYPARAQALRIEGQVKVKFDVTPDGRVDNVQILSAKPANMFEREVKNAMRRWRYEPGKPGSGIVVNILFKINGTTEIQ

>TonB_Synechocystis PCC 6803

MSISNFCLTQRSQQYQNSQKIILAGATCSLLIHGAIGAFWRFPLDGDTPAVEPIEFIVVDPSPIPEVTPPPTEQKTEQKK

VEPLPNISPPPPTPVTQAVLPPLPQAIIPPVPSVNSLPSPVVENPISLPNSLPPPPAAPVPNAVSLPVDQAITPPTPNLN

PLPNPVTESPASLTQPAQSAFRLPFSSLQTPEFAPIPEQAPAVSPGIRPATPQSLANAPSRQPVFGRISSSAPSQGNPLT

NQDSFTAIANGNAVSQGAPPMKPGTATKPIASGGGSSGPSLGSGQPRAERPGGGNAGANVGPIAANPVASSAPPKPKPTP

SSPAKPDPLKCVSQCKPSYPSILQGEEGSATVLISVNDSGGVTSVTITNAHGNSEVNRQALLAARKMQFTAPASGQSKSV

PVVIHFTVAGSDFDRQARERQQQQEELRQAARRAEEEKANQARQRQLEEERQARQRQLEKEREERLKQFSVESQPSPVPA

VKVSPSPSPTPAVRVSPSPSPEPPPTPAAAPKTEPAVGIDELAPDLEQSSDS