First steps:
Prepare reference databases and trees

TRAITS project meeting: 7-10 September 2021 – CNB, Madrid

Install software:

Install Miniconda:

https://docs.conda.io/projects/conda/en/latest/user-guide/install/linux.html

configure channels

```
$ conda config --add channels defaults
$ conda config --add channels bioconda
$ conda config --add channels conda-forge
```

create a new environment and install programs:

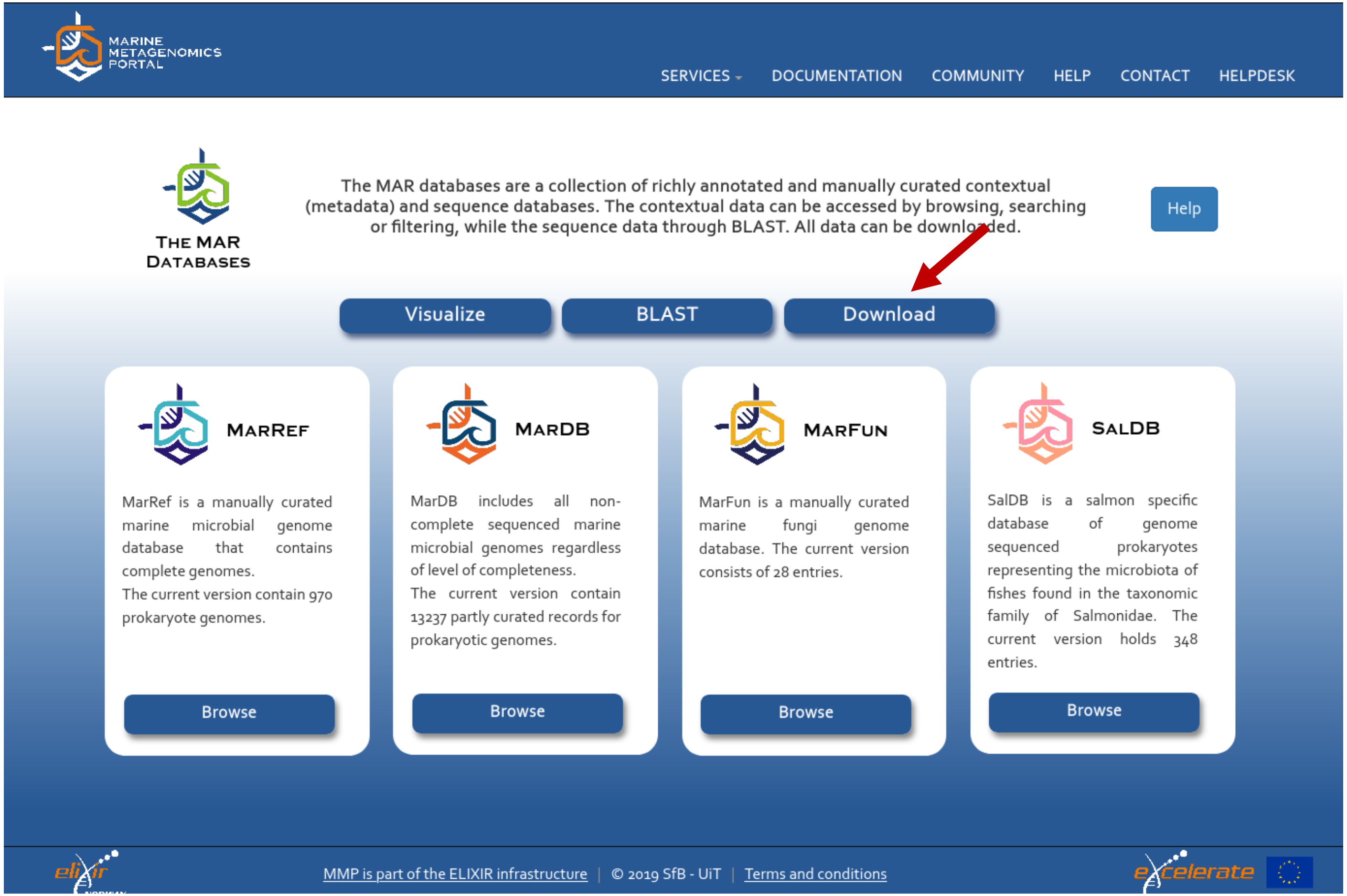
```
$ conda create -n traits
$ conda install python=3.9.6
$ conda install -c conda-forge biopython
$ conda install -c bioconda seqkit prodigal hmmer muscl iqtree epa-ng gappa
```

Install Papara

https://cme.h-its.org/exelixis/web/software/papara/index.html

Download MAR databases:

https://mmp.sfb.uit.no/databases/



The MAR databases are a collection of richly annotated and manually curated contextual (metadata) and sequence databases. The contextual data can be accessed by browsing, searching or filtering, while the sequence data through BLAST. All data can be downloaded.

MarRef is a manually curated marine microbial genome database that contains complete genomes. The current version contain 970 prokaryote genomes.

MarDB includes all non-complete sequenced marine microbial genomes regardless of level of completeness. The current version contain 13237 partly curated records for prokaryotic genomes.

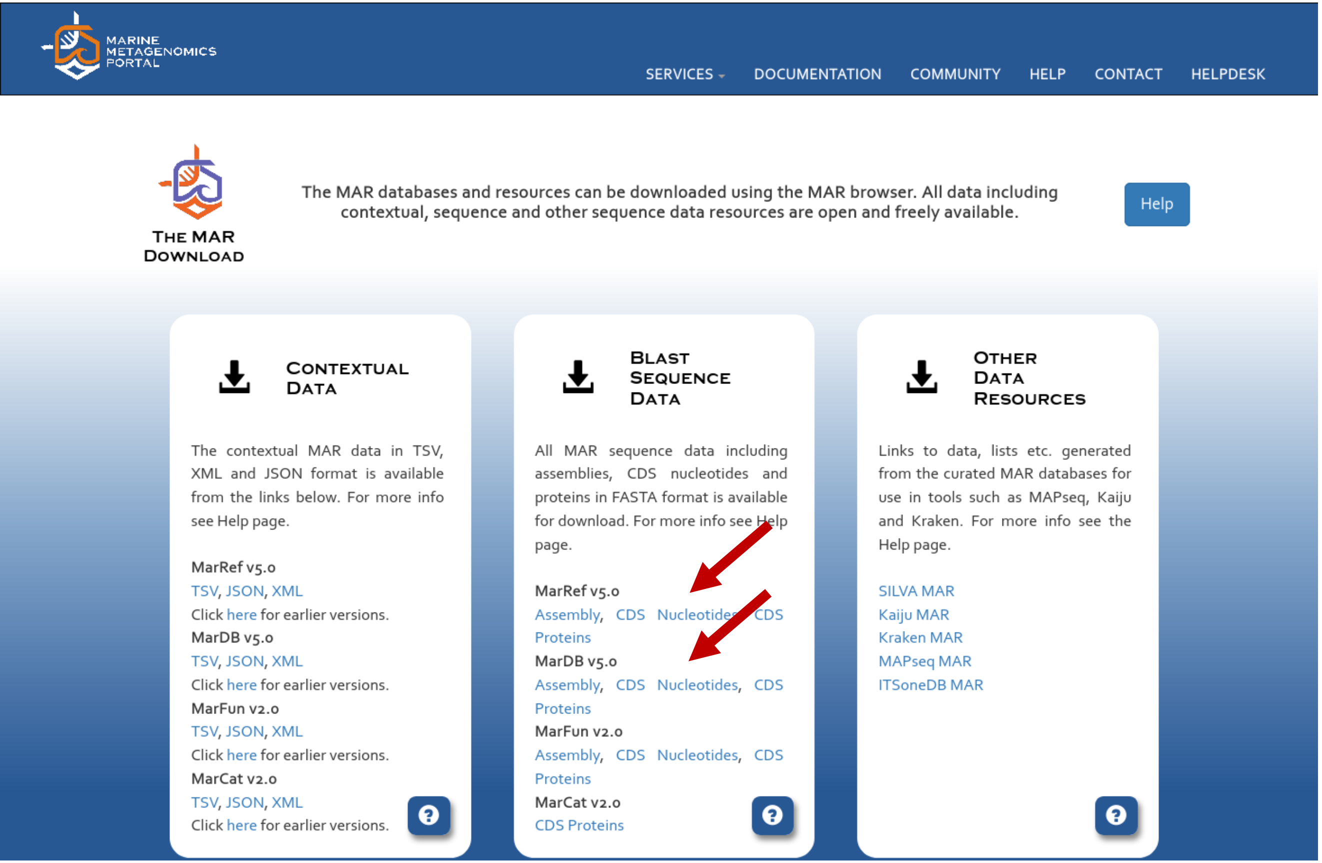
MarFun is a manually curated marine fungi genome database. The current version consists of 28 entries.

SalDB is a salmon specific database of genome sequenced prokaryotes representing the microbiota of fishes found in the taxonomic family of Salmonidae. The current version holds 348 entries.

MMP is part of the ELIXIR infrastructure | © 2019 SfB - UiT | Terms and conditions

Download DNA versions

- Click on CDS Nucleotides
- Next page: look for most recent version and copy link



MARINE METAGENOMICS PORTAL

SERVICES ▾ DOCUMENTATION COMMUNITY HELP CONTACT HELPDESK

THE MAR DOWNLOAD

The MAR databases and resources can be downloaded using the MAR browser. All data including contextual, sequence and other sequence data resources are open and freely available.

Help

CONTEXTUAL DATA

The contextual MAR data in TSV, XML and JSON format is available from the links below. For more info see Help page.

MarRef v5.0
TSV, JSON, XML
Click here for earlier versions.
MarDB v5.0
TSV, JSON, XML
Click here for earlier versions.
MarFun v2.0
TSV, JSON, XML
Click here for earlier versions.
MarCat v2.0
TSV, JSON, XML
Click here for earlier versions.

BLAST SEQUENCE DATA

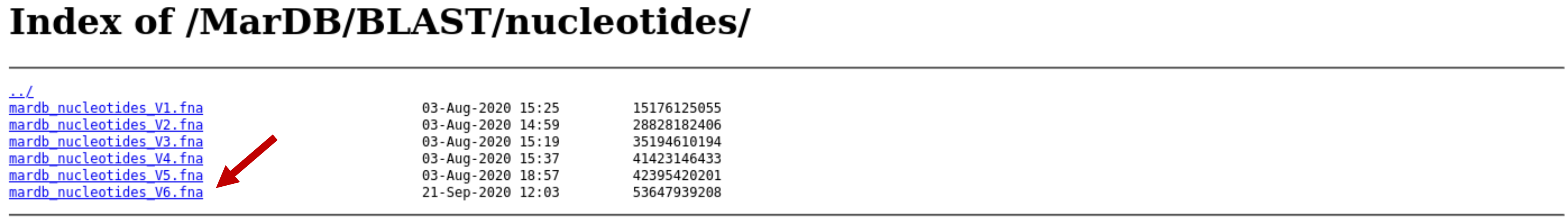
All MAR sequence data including assemblies, CDS nucleotides and proteins in FASTA format is available for download. For more info see Help page.

MarRef v5.0
Assembly, CDS Nucleotides, CDS Proteins
MarDB v5.0
Assembly, CDS Nucleotides, CDS Proteins
MarFun v2.0
Assembly, CDS Nucleotides, CDS Proteins
MarCat v2.0
CDS Proteins

OTHER DATA RESOURCES

Links to data, lists etc. generated from the curated MAR databases for use in tools such as MAPseq, Kaiju and Kraken. For more info see the Help page.

SILVA MAR
Kaiju MAR
Kraken MAR
MAPseq MAR
ITSoneDB MAR

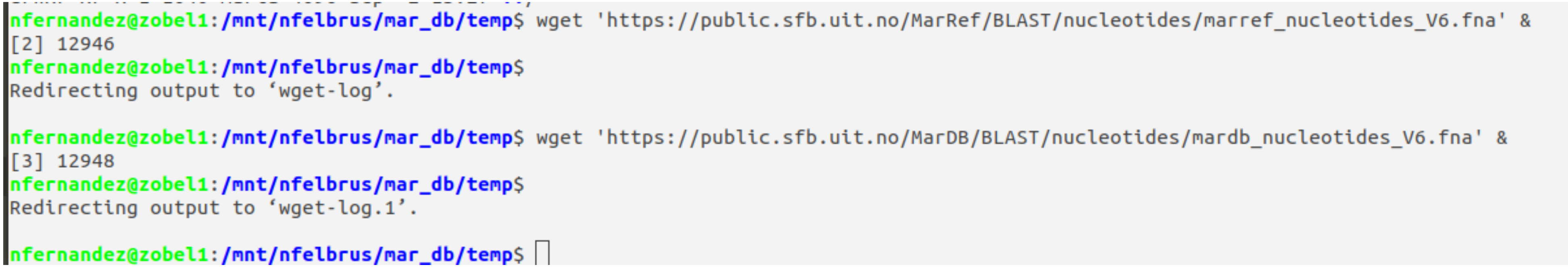


Index of /MarDB/BLAST/nucleotides/

```
../
mardb_nucleotides_V1.fna                   03-Aug-2020 15:25        15176125055
mardb_nucleotides_V2.fna                   03-Aug-2020 14:59        28828182406
mardb_nucleotides_V3.fna                   03-Aug-2020 15:19        35194610194
mardb_nucleotides_V4.fna                   03-Aug-2020 15:37        41423146433
mardb_nucleotides_V5.fna                   03-Aug-2020 18:57        42395420201
mardb_nucleotides_V6.fna                   21-Sep-2020 12:03        53647939208
```

Back to your terminal:

```
$ wget 'https://public.sfb.uit.no/MarDB/BLAST/nucleotides/mardb_nucleotides_V6.fna' &

$ wget 'https://public.sfb.uit.no/MarRef/BLAST/nucleotides/marref_nucleotides_V6.fna' &
```



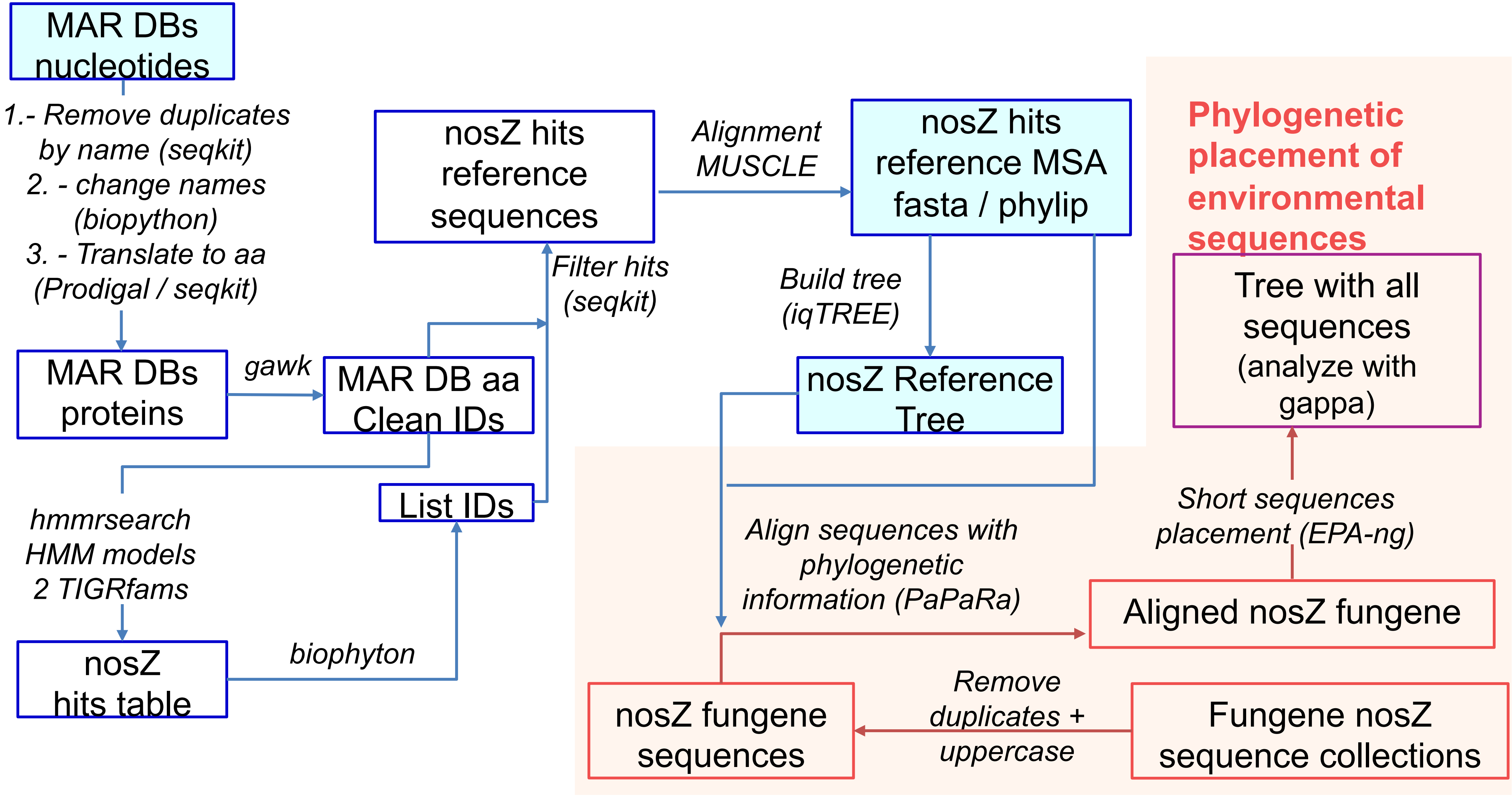
```
nfernandez@zobel1:/mnt/nfelbrus/mar_db/temp$ wget 'https://public.sfb.uit.no/MarRef/BLAST/nucleotides/marref_nucleotides_V6.fna' &
[2] 12946
nfernandez@zobel1:/mnt/nfelbrus/mar_db/temp$
Redirecting output to ‘wget-log’.

nfernandez@zobel1:/mnt/nfelbrus/mar_db/temp$ wget 'https://public.sfb.uit.no/MarDB/BLAST/nucleotides/mardb_nucleotides_V6.fna' &
[3] 12948
nfernandez@zobel1:/mnt/nfelbrus/mar_db/temp$
Redirecting output to ‘wget-log.1’.

nfernandez@zobel1:/mnt/nfelbrus/mar_db/temp$
```

Other databases:

Paoli’s dataset contains MAR databases – discuss

Use of JGI-GOLD

Overview



MAR DBs nucleotides

1.- Remove duplicates by name (seqkit)
2. - change names (biopython)
3. - Translate to aa (Prodigal / seqkit)

MAR DBs proteins

gawk

MAR DB aa Clean IDs

hmmrsearch HMM models 2 TIGRfams

nosZ hits table

biophyton

List IDs

Filter hits (seqkit)

nosZ hits reference sequences

Alignment MUSCLE

nosZ hits reference MSA fasta / phylip

Build tree (iqTREE)

nosZ Reference Tree

Phylogenetic placement of environmental sequences

Fungene nosZ sequence collections

Remove duplicates + uppercase

nosZ fungene sequences

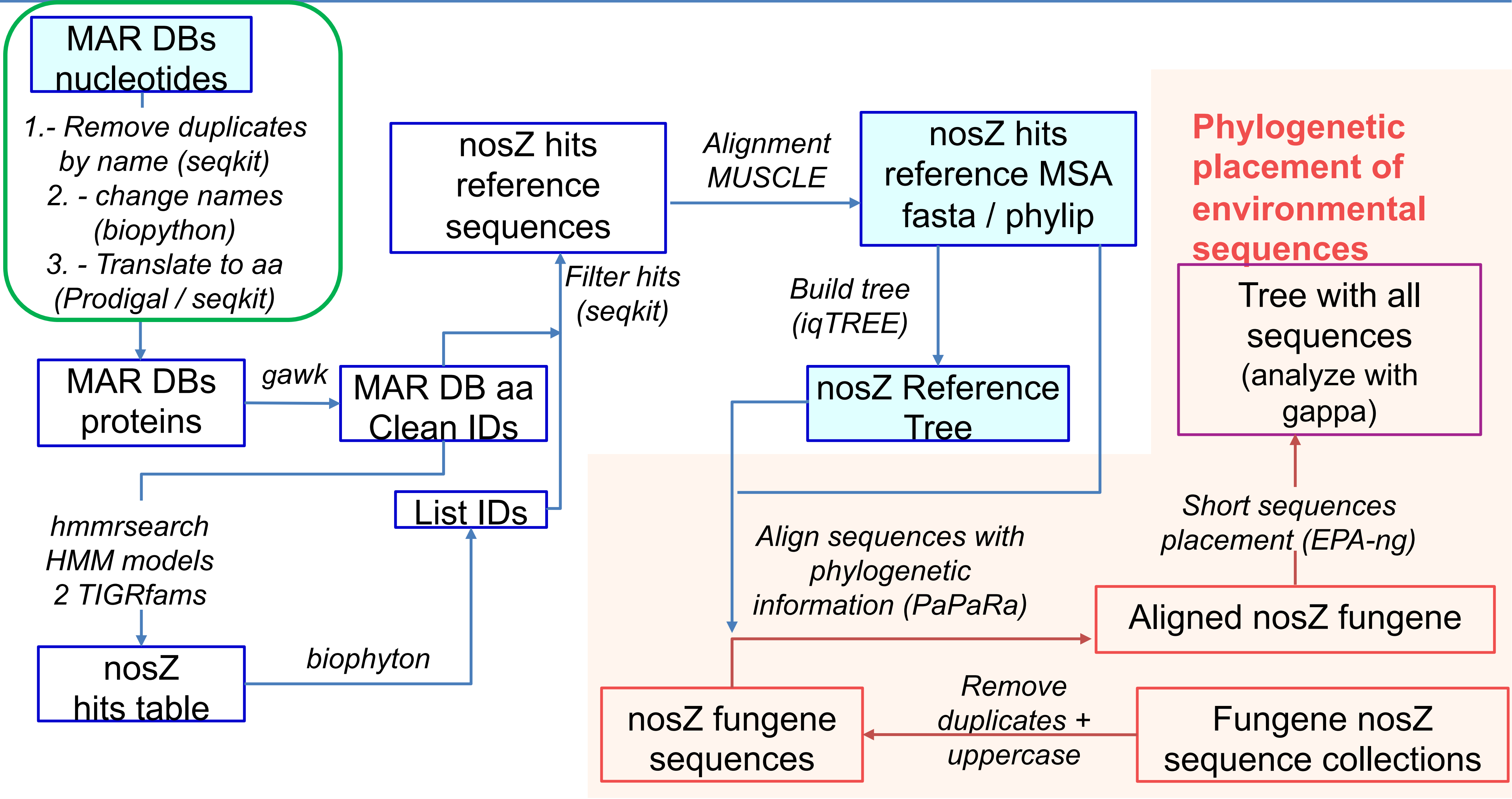
Align sequences with phylogenetic information (PaPaRa)

Aligned nosZ fungene

Short sequences placement (EPA-ng)

Tree with all sequences (analyze with gappa)

Overview



MAR DBs nucleotides

1.- Remove duplicates by name (seqkit)
2. - change names (biopython)
3. - Translate to aa (Prodigal / seqkit)

MAR DBs proteins

gawk

MAR DB aa Clean IDs

hmmrsearch HMM models 2 TIGRfams

nosZ hits table

biophyton

List IDs

Filter hits (seqkit)

nosZ hits reference sequences

Alignment MUSCLE

nosZ hits reference MSA fasta / phylip

Build tree (iqTREE)

nosZ Reference Tree

Phylogenetic placement of environmental sequences

Align sequences with phylogenetic information (PaPaRa)

Aligned nosZ fungene

Short sequences placement (EPA-ng)

Tree with all sequences (analyze with gappa)

Fungene nosZ sequence collections

Remove duplicates + uppercase

nosZ fungene sequences

Prepare databases:

Databases seem to contain duplicates – removing them

```
$ conda activate traits
```

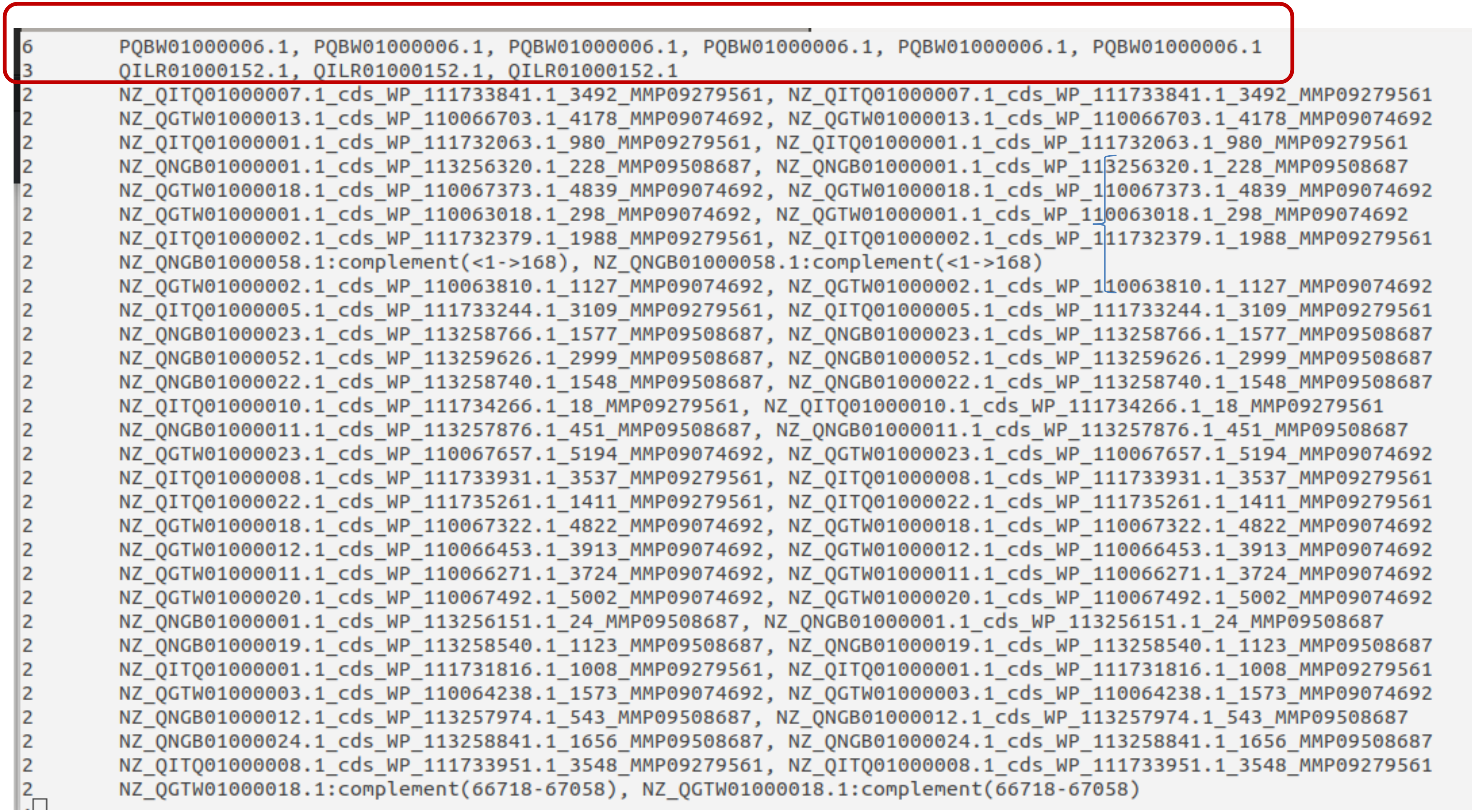
Do not run (MarRef – no duplicates):

```
$ cat marref_nucleotides_V6.fna | seqkit rmdup -n -o marrefnoDup_nucleotides.fna -d
duplicates_marref.fna -D duplicates_marref_info.txt
[INFO] 0 duplicated records removed
```

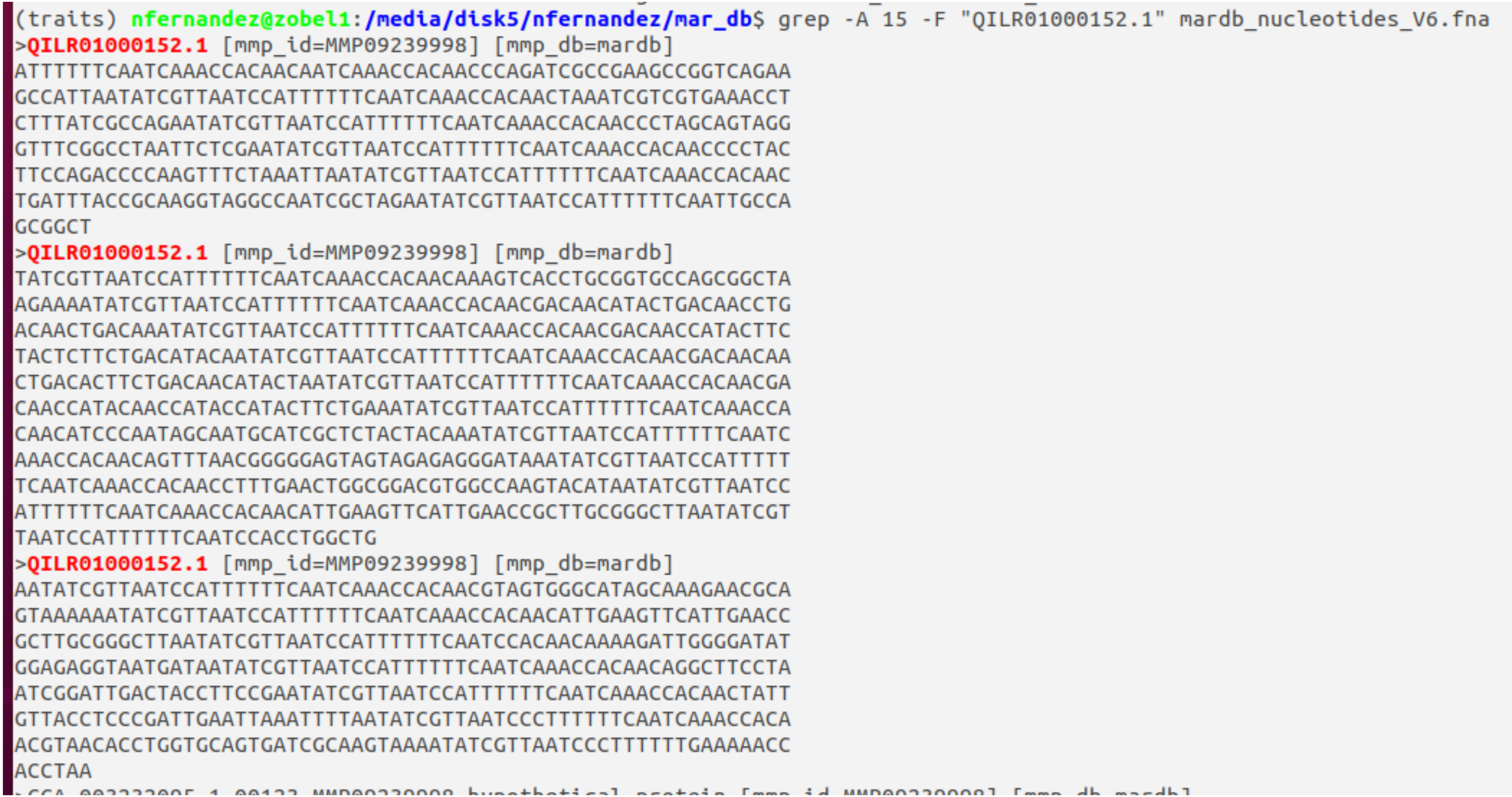
MarDB - seqkit tool:

- n: duplicates identification by name (not by sequence)
- d: fasta with duplicated sequences
- D: text file with information

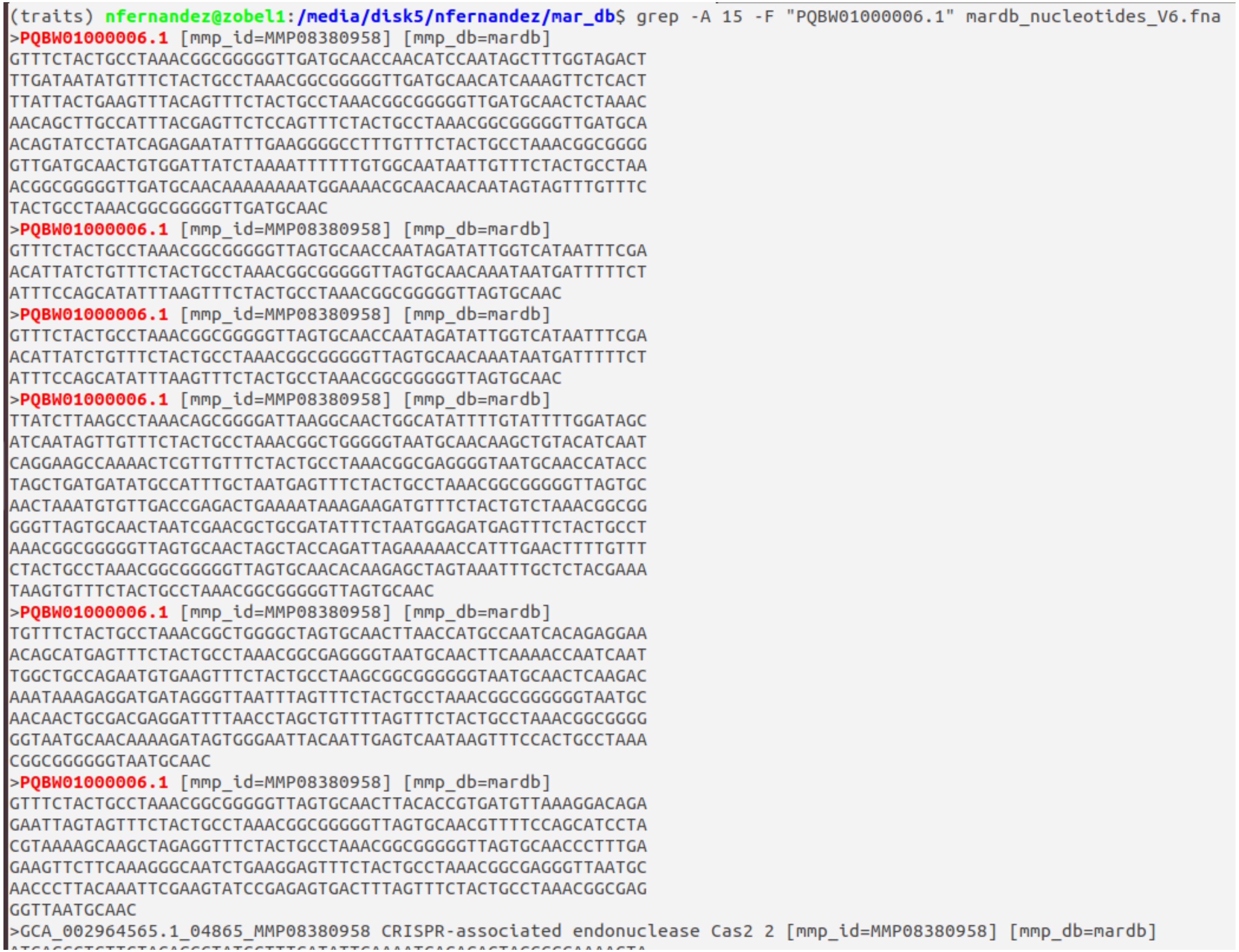
```
$ cat mardb_nucleotides_V6.fna | seqkit rmdup -n -o mardbnoDup_nucleotides.fna -d
duplicates_mardb.fna -D duplicates_mardb_info.txt 1>log_rmdup.txt 2>&1 &

$ cat log_rmdup.txt
[INFO] 12890 duplicated records removed

$ less duplicates_mardb_info.fna
```



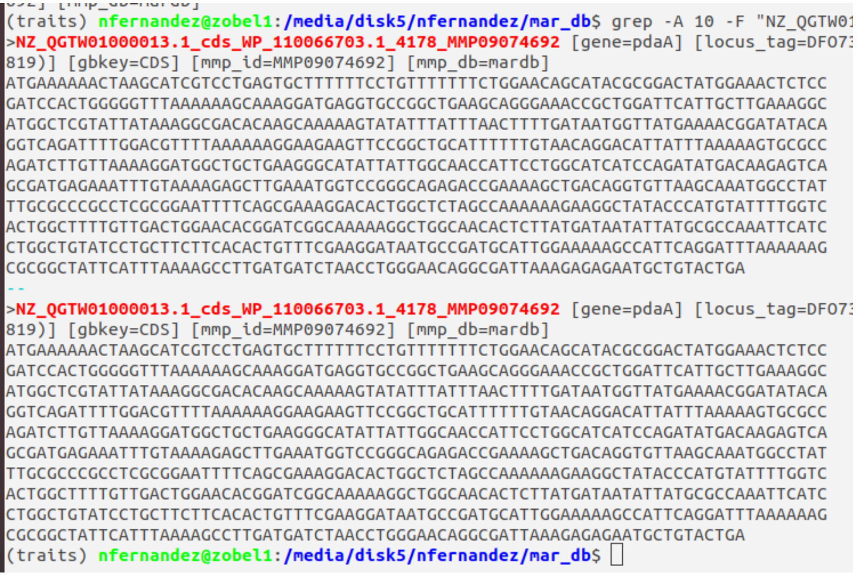
```
6       PQBW01000006.1, PQBW01000006.1, PQBW01000006.1, PQBW01000006.1, PQBW01000006.1, PQBW01000006.1
3       QILR01000152.1, QILR01000152.1, QILR01000152.1
2       NZ_QITQ01000007.1_cds_WP_111733841.1_3492_MMP09279561, NZ_QITQ01000007.1_cds_WP_111733841.1_3492_MMP09279561
2       NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692, NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692
2       NZ_QITQ01000001.1_cds_WP_111732063.1_980_MMP09279561, NZ_QITQ01000001.1_cds_WP_111732063.1_980_MMP09279561
2       NZ_QNGB01000001.1_cds_WP_113256320.1_228_MMP09508687, NZ_QNGB01000001.1_cds_WP_113256320.1_228_MMP09508687
2       NZ_QGTW01000018.1_cds_WP_110067373.1_4839_MMP09074692, NZ_QGTW01000018.1_cds_WP_110067373.1_4839_MMP09074692
2       NZ_QGTW01000001.1_cds_WP_110063018.1_298_MMP09074692, NZ_QGTW01000001.1_cds_WP_110063018.1_298_MMP09074692
2       NZ_QITQ01000002.1_cds_WP_111732379.1_1988_MMP09279561, NZ_QITQ01000002.1_cds_WP_111732379.1_1988_MMP09279561
2       NZ_QNGB01000058.1:complement(<1->168), NZ_QNGB01000058.1:complement(<1->168)
2       NZ_QGTW01000002.1_cds_WP_110063810.1_1127_MMP09074692, NZ_QGTW01000002.1_cds_WP_110063810.1_1127_MMP09074692
2       NZ_QITQ01000005.1_cds_WP_111733244.1_3109_MMP09279561, NZ_QITQ01000005.1_cds_WP_111733244.1_3109_MMP09279561
2       NZ_QNGB01000023.1_cds_WP_113258766.1_1577_MMP09508687, NZ_QNGB01000023.1_cds_WP_113258766.1_1577_MMP09508687
2       NZ_QNGB01000052.1_cds_WP_113259626.1_2999_MMP09508687, NZ_QNGB01000052.1_cds_WP_113259626.1_2999_MMP09508687
2       NZ_QNGB01000022.1_cds_WP_113258740.1_1548_MMP09508687, NZ_QNGB01000022.1_cds_WP_113258740.1_1548_MMP09508687
2       NZ_QITQ01000010.1_cds_WP_111734266.1_18_MMP09279561, NZ_QITQ01000010.1_cds_WP_111734266.1_18_MMP09279561
2       NZ_QNGB01000011.1_cds_WP_113257876.1_451_MMP09508687, NZ_QNGB01000011.1_cds_WP_113257876.1_451_MMP09508687
2       NZ_QGTW01000023.1_cds_WP_110067657.1_5194_MMP09074692, NZ_QGTW01000023.1_cds_WP_110067657.1_5194_MMP09074692
2       NZ_QITQ01000008.1_cds_WP_111733931.1_3537_MMP09279561, NZ_QITQ01000008.1_cds_WP_111733931.1_3537_MMP09279561
2       NZ_QITQ01000022.1_cds_WP_111735261.1_1411_MMP09279561, NZ_QITQ01000022.1_cds_WP_111735261.1_1411_MMP09279561
2       NZ_QGTW01000018.1_cds_WP_110067322.1_4822_MMP09074692, NZ_QGTW01000018.1_cds_WP_110067322.1_4822_MMP09074692
2       NZ_QGTW01000012.1_cds_WP_110066453.1_3913_MMP09074692, NZ_QGTW01000012.1_cds_WP_110066453.1_3913_MMP09074692
2       NZ_QGTW01000011.1_cds_WP_110066271.1_3724_MMP09074692, NZ_QGTW01000011.1_cds_WP_110066271.1_3724_MMP09074692
2       NZ_QGTW01000020.1_cds_WP_110067492.1_5002_MMP09074692, NZ_QGTW01000020.1_cds_WP_110067492.1_5002_MMP09074692
2       NZ_QNGB01000001.1_cds_WP_113256151.1_24_MMP09508687, NZ_QNGB01000001.1_cds_WP_113256151.1_24_MMP09508687
2       NZ_QNGB01000019.1_cds_WP_113258540.1_1123_MMP09508687, NZ_QNGB01000019.1_cds_WP_113258540.1_1123_MMP09508687
2       NZ_QITQ01000001.1_cds_WP_111731816.1_1008_MMP09279561, NZ_QITQ01000001.1_cds_WP_111731816.1_1008_MMP09279561
2       NZ_QGTW01000003.1_cds_WP_110064238.1_1573_MMP09074692, NZ_QGTW01000003.1_cds_WP_110064238.1_1573_MMP09074692
2       NZ_QNGB01000012.1_cds_WP_113257974.1_543_MMP09508687, NZ_QNGB01000012.1_cds_WP_113257974.1_543_MMP09508687
2       NZ_QNGB01000024.1_cds_WP_113258841.1_1656_MMP09508687, NZ_QNGB01000024.1_cds_WP_113258841.1_1656_MMP09508687
2       NZ_QITQ01000008.1_cds_WP_111733951.1_3548_MMP09279561, NZ_QITQ01000008.1_cds_WP_111733951.1_3548_MMP09279561
2       NZ_QGTW01000018.1:complement(66718-67058), NZ_QGTW01000018.1:complement(66718-67058)
```



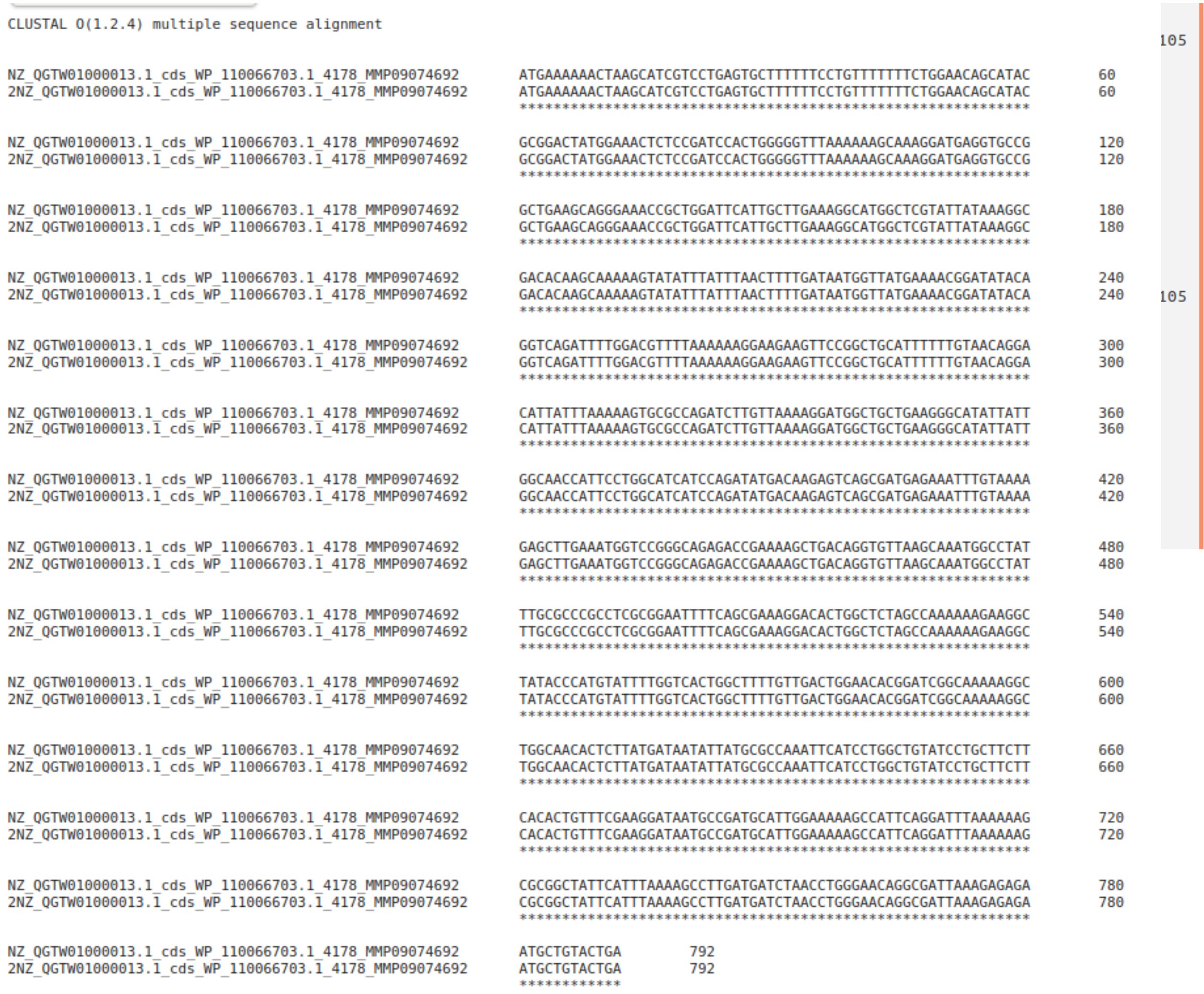
```
(traits) nfernandez@zobel1:/media/disk5/nfernandez/mar_db$ grep -A 15 -F "QILR01000152.1" mardb_nucleotides_V6.fna
>QILR01000152.1 [mmp_id=MMP09239998] [mmp_db=mardb]
ATTTTTTCAATCAAACCACAACAATCAAACCACAACCCAGATCGCCGAAGCCGGTCAGAA
GCCATTAATATCGTTAATCCATTTTTTCAATCAAACCACAACTAAATCGTCGTGAAACCT
CTTTATCGCCAGAATATCGTTAATCCATTTTTTCAATCAAACCACAACCCTAGCAGTAGG
GTTTCGGCCTAATTCTCGAATATCGTTAATCCATTTTTTCAATCAAACCACAACCCCTAC
TTCCAGACCCCAAGTTTCTAAATTAATATCGTTAATCCATTTTTTCAATCAAACCACAAC
TGATTTACCGCAAGGTAGGCCAATCGCTAGAATATCGTTAATCCATTTTTTCAATTGCCA
GCGGCT
>QILR01000152.1 [mmp_id=MMP09239998] [mmp_db=mardb]
TATCGTTAATCCATTTTTTCAATCAAACCACAACAAAGTCACCTGCGGTGCCAGCGGCTA
AGAAAATATCGTTAATCCATTTTTTCAATCAAACCACAACGACAACATACTGACAACCTG
ACAACTGACAAATATCGTTAATCCATTTTTTCAATCAAACCACAACGACAACCATACTTC
TACTCTTCTGACATACAATATCGTTAATCCATTTTTTCAATCAAACCACAACGACAACAA
CTGACACTTCTGACAACATACTAATATCGTTAATCCATTTTTTCAATCAAACCACAACGA
CAACCATACAACCATACCATACTTCTGAAATATCGTTAATCCATTTTTTCAATCAAACCA
CAACATCCCAATAGCAATGCATCGCTCTACTACAAATATCGTTAATCCATTTTTTCAATC
AAACCACAACAGTTTAACGGGGGAGTAGTAGAGAGGGATAAATATCGTTAATCCATTTTT
TCAATCAAACCACAACCTTTGAACTGGCGGACGTGGCCAAGTACATAATATCGTTAATCC
ATTTTTTCAATCAAACCACAACATTGAAGTTCATTGAACCGCTTGCGGGCTTAATATCGT
TAATCCATTTTTTCAATCCACCTGGCTG
>QILR01000152.1 [mmp_id=MMP09239998] [mmp_db=mardb]
AATATCGTTAATCCATTTTTTCAATCAAACCACAACGTAGTGGGCATAGCAAAGAACGCA
GTAAAAAATATCGTTAATCCATTTTTTCAATCAAACCACAACATTGAAGTTCATTGAACC
GCTTGCGGGCTTAATATCGTTAATCCATTTTTTCAATCCACAACAAAAGATTGGGGATAT
GGAGAGGTAATGATAATATCGTTAATCCATTTTTTCAATCAAACCACAACAGGCTTCCTA
ATCGGATTGACTACCTTCCGAATATCGTTAATCCATTTTTTCAATCAAACCACAACTATT
GTTACCTCCCGATTGAATTAAATTTTAATATCGTTAATCCCTTTTTTCAATCAAACCACA
ACGTAACACCTGGTGCAGTGATCGCAAGTAAAATATCGTTAATCCCTTTTTTGAAAAACC
ACCTAA
```



```
(traits) nfernandez@zobel1:/media/disk5/nfernandez/mar_db$ grep -A 15 -F "PQBW01000006.1" mardb_nucleotides_V6.fna
>PQBW01000006.1 [mmp_id=MMP08380958] [mmp_db=mardb]
GTTTCTACTGCCTAAACGGCGGGGGTTGATGCAACCAACATCCAATAGCTTTGGTAGACT
TTGATAATATGTTTCTACTGCCTAAACGGCGGGGGTTGATGCAACATCAAAGTTCTCACT
TTATTACTGAAGTTTACAGTTTCTACTGCCTAAACGGCGGGGGTTGATGCAACTCTAAAC
AACAGCTTGCCATTTACGAGTTCTCCAGTTTCTACTGCCTAAACGGCGGGGGTTGATGCA
ACAGTATCCTATCAGAGAATATTTGAAGGGGCCTTTGTTTCTACTGCCTAAACGGCGGGG
GTTGATGCAACTGTGGATTATCTAAAATTTTTTGTGGCAATAATTGTTTCTACTGCCTAA
ACGGCGGGGGTTGATGCAACAAAAAAAATGGAAAACGCAACAACAATAGTAGTTTGTTTC
TACTGCCTAAACGGCGGGGGTTGATGCAAC
>PQBW01000006.1 [mmp_id=MMP08380958] [mmp_db=mardb]
GTTTCTACTGCCTAAACGGCGGGGGTTAGTGCAACCAATAGATATTGGTCATAATTTCGA
ACATTATCTGTTTCTACTGCCTAAACGGCGGGGGTTAGTGCAACAAATAATGATTTTTCT
ATTTCCAGCATATTTAAGTTTCTACTGCCTAAACGGCGGGGGTTAGTGCAAC
>PQBW01000006.1 [mmp_id=MMP08380958] [mmp_db=mardb]
GTTTCTACTGCCTAAACGGCGGGGGTTAGTGCAACCAATAGATATTGGTCATAATTTCGA
ACATTATCTGTTTCTACTGCCTAAACGGCGGGGGTTAGTGCAACAAATAATGATTTTTCT
ATTTCCAGCATATTTAAGTTTCTACTGCCTAAACGGCGGGGGTTAGTGCAAC
>PQBW01000006.1 [mmp_id=MMP08380958] [mmp_db=mardb]
TTATCTTAAGCCTAAACAGCGGGGATTAAGGCAACTGGCATATTTTGTATTTTGGATAGC
ATCAATAGTTGTTTCTACTGCCTAAACGGCTGGGGGTAATGCAACAAGCTGTACATCAAT
CAGGAAGCCAAAACTCGTTGTTTCTACTGCCTAAACGGCGAGGGGTAATGCAACCATACC
TAGCTGATGATATGCCATTTGCTAATGAGTTTCTACTGCCTAAACGGCGGGGGTTAGTGC
AACTAAATGTGTTGACCGAGACTGAAAATAAAGAAGATGTTTCTACTGTCTAAACGGCGG
GGGTTAGTGCAACTAATCGAACGCTGCGATATTTCTAATGGAGATGAGTTTCTACTGCCT
AAACGGCGGGGGTTAGTGCAACTAGCTACCAGATTAGAAAAACCATTTGAACTTTTGTTT
CTACTGCCTAAACGGCGGGGGTTAGTGCAACACAAGAGCTAGTAAATTTGCTCTACGAAA
TAAGTGTTTCTACTGCCTAAACGGCGGGGGTTAGTGCAAC
>PQBW01000006.1 [mmp_id=MMP08380958] [mmp_db=mardb]
TGTTTCTACTGCCTAAACGGCTGGGGCTAGTGCAACTTAACCATGCCAATCACAGAGGAA
ACAGCATGAGTTTCTACTGCCTAAACGGCGAGGGGTAATGCAACTTCAAAACCAATCAAT
TGGCTGCCAGAATGTGAAGTTTCTACTGCCTAAGCGGCGGGGGGTAATGCAACTCAAGAC
AAATAAAGAGGATGATAGGGTTAATTTAGTTTCTACTGCCTAAACGGCGGGGGGTAATGC
AACAACTGCGACGAGGATTTTAACCTAGCTGTTTTAGTTTCTACTGCCTAAACGGCGGGG
GGTAATGCAACAAAAGATAGTGGGAATTACAATTGAGTCAATAAGTTTCCACTGCCTAAA
CGGCGGGGGGTAATGCAAC
>PQBW01000006.1 [mmp_id=MMP08380958] [mmp_db=mardb]
GTTTCTACTGCCTAAACGGCGGGGGTTAGTGCAACTTACACCGTGATGTTAAAGGACAGA
GAATTAGTAGTTTCTACTGCCTAAACGGCGGGGGTTAGTGCAACGTTTTCCAGCATCCTA
CGTAAAAGCAAGCTAGAGGTTTCTACTGCCTAAACGGCGGGGGTTAGTGCAACCCTTTGA
GAAGTTCTTCAAAGGGCAATCTGAAGGAGTTTCTACTGCCTAAACGGCGAGGGTTAATGC
AACCCTTACAAATTCGAAGTATCCGAGAGTGACTTTAGTTTCTACTGCCTAAACGGCGAG
GGTTAATGCAAC
>GCA_002964565.1_04865_MMP08380958 CRISPR-associated endonuclease Cas2 2 [mmp_id=MMP08380958] [mmp_db=mardb]
ATCAGCCTCTTCTACAGCCTATGCTTTCATATTCAAAATCAGACACTAGCCCCAAAACTA
```



```
092] [mmp_db=mardb]
(traits) nfernandez@zobel1:/media/disk5/nfernandez/mar_db$ grep -A 10 -F "NZ_QGTW01
>NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692 [gene=pdaA] [locus_tag=DFO73
819)] [gbkey=CDS] [mmp_id=MMP09074692] [mmp_db=mardb]
ATGAAAAAACTAAGCATCGTCCTGAGTGCTTTTTTCCTGTTTTTTTCTGGAACAGCATACGCGGACTATGGAAACTCTCC
GATCCACTGGGGGTTTAAAAAAGCAAAGGATGAGGTGCCGGCTGAAGCAGGGAAACCGCTGGATTCATTGCTTGAAAGGC
ATGGCTCGTATTATAAAGGCGACACAAGCAAAAAGTATATTTATTTAACTTTTGATAATGGTTATGAAAACGGATATACA
GGTCAGATTTTGGACGTTTTAAAAAAGGAAGAAGTTCCGGCTGCATTTTTTGTAACAGGACATTATTTAAAAAGTGCGCC
AGATCTTGTTAAAAGGATGGCTGCTGAAGGGCATATTATTGGCAACCATTCCTGGCATCATCCAGATATGACAAGAGTCA
GCGATGAGAAATTTGTAAAAGAGCTTGAAATGGTCCGGGCAGAGACCGAAAAGCTGACAGGTGTTAAGCAAATGGCCTAT
TTGCGCCCGCCTCGCGGAATTTTCAGCGAAAGGACACTGGCTCTAGCCAAAAAAGAAGGCTATACCCATGTATTTTGGTC
ACTGGCTTTTGTTGACTGGAACACGGATCGGCAAAAAGGCTGGCAACACTCTTATGATAATATTATGCGCCAAATTCATC
CTGGCTGTATCCTGCTTCTTCACACTGTTTCGAAGGATAATGCCGATGCATTGGAAAAAGCCATTCAGGATTTAAAAAAG
CGCGGCTATTCATTTAAAAGCCTTGATGATCTAACCTGGGAACAGGCGATTAAAGAGAGAATGCTGTACTGA
--
>NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692 [gene=pdaA] [locus_tag=DFO73
819)] [gbkey=CDS] [mmp_id=MMP09074692] [mmp_db=mardb]
ATGAAAAAACTAAGCATCGTCCTGAGTGCTTTTTTCCTGTTTTTTTCTGGAACAGCATACGCGGACTATGGAAACTCTCC
GATCCACTGGGGGTTTAAAAAAGCAAAGGATGAGGTGCCGGCTGAAGCAGGGAAACCGCTGGATTCATTGCTTGAAAGGC
ATGGCTCGTATTATAAAGGCGACACAAGCAAAAAGTATATTTATTTAACTTTTGATAATGGTTATGAAAACGGATATACA
GGTCAGATTTTGGACGTTTTAAAAAAGGAAGAAGTTCCGGCTGCATTTTTTGTAACAGGACATTATTTAAAAAGTGCGCC
AGATCTTGTTAAAAGGATGGCTGCTGAAGGGCATATTATTGGCAACCATTCCTGGCATCATCCAGATATGACAAGAGTCA
GCGATGAGAAATTTGTAAAAGAGCTTGAAATGGTCCGGGCAGAGACCGAAAAGCTGACAGGTGTTAAGCAAATGGCCTAT
TTGCGCCCGCCTCGCGGAATTTTCAGCGAAAGGACACTGGCTCTAGCCAAAAAAGAAGGCTATACCCATGTATTTTGGTC
ACTGGCTTTTGTTGACTGGAACACGGATCGGCAAAAAGGCTGGCAACACTCTTATGATAATATTATGCGCCAAATTCATC
CTGGCTGTATCCTGCTTCTTCACACTGTTTCGAAGGATAATGCCGATGCATTGGAAAAAGCCATTCAGGATTTAAAAAAG
CGCGGCTATTCATTTAAAAGCCTTGATGATCTAACCTGGGAACAGGCGATTAAAGAGAGAATGCTGTACTGA
(traits) nfernandez@zobel1:/media/disk5/nfernandez/mar_db$
```



```
CLUSTAL O(1.2.4) multiple sequence alignment


NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692      ATGAAAAAACTAAGCATCGTCCTGAGTGCTTTTTTCCTGTTTTTTTCTGGAACAGCATAC	60
2NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692     ATGAAAAAACTAAGCATCGTCCTGAGTGCTTTTTTCCTGTTTTTTTCTGGAACAGCATAC	60
                                                           ************************************************************

NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692      GCGGACTATGGAAACTCTCCGATCCACTGGGGGTTTAAAAAAGCAAAGGATGAGGTGCCG	120
2NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692     GCGGACTATGGAAACTCTCCGATCCACTGGGGGTTTAAAAAAGCAAAGGATGAGGTGCCG	120
                                                           ************************************************************

NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692      GCTGAAGCAGGGAAACCGCTGGATTCATTGCTTGAAAGGCATGGCTCGTATTATAAAGGC	180
2NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692     GCTGAAGCAGGGAAACCGCTGGATTCATTGCTTGAAAGGCATGGCTCGTATTATAAAGGC	180
                                                           ************************************************************

NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692      GACACAAGCAAAAAGTATATTTATTTAACTTTTGATAATGGTTATGAAAACGGATATACA	240
2NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692     GACACAAGCAAAAAGTATATTTATTTAACTTTTGATAATGGTTATGAAAACGGATATACA	240
                                                           ************************************************************

NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692      GGTCAGATTTTGGACGTTTTAAAAAAGGAAGAAGTTCCGGCTGCATTTTTTGTAACAGGA	300
2NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692     GGTCAGATTTTGGACGTTTTAAAAAAGGAAGAAGTTCCGGCTGCATTTTTTGTAACAGGA	300
                                                           ************************************************************

NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692      CATTATTTAAAAAGTGCGCCAGATCTTGTTAAAAGGATGGCTGCTGAAGGGCATATTATT	360
2NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692     CATTATTTAAAAAGTGCGCCAGATCTTGTTAAAAGGATGGCTGCTGAAGGGCATATTATT	360
                                                           ************************************************************

NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692      GGCAACCATTCCTGGCATCATCCAGATATGACAAGAGTCAGCGATGAGAAATTTGTAAAA	420
2NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692     GGCAACCATTCCTGGCATCATCCAGATATGACAAGAGTCAGCGATGAGAAATTTGTAAAA	420
                                                           ************************************************************

NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692      GAGCTTGAAATGGTCCGGGCAGAGACCGAAAAGCTGACAGGTGTTAAGCAAATGGCCTAT	480
2NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692     GAGCTTGAAATGGTCCGGGCAGAGACCGAAAAGCTGACAGGTGTTAAGCAAATGGCCTAT	480
                                                           ************************************************************

NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692      TTGCGCCCGCCTCGCGGAATTTTCAGCGAAAGGACACTGGCTCTAGCCAAAAAAGAAGGC	540
2NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692     TTGCGCCCGCCTCGCGGAATTTTCAGCGAAAGGACACTGGCTCTAGCCAAAAAAGAAGGC	540
                                                           ************************************************************

NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692      TATACCCATGTATTTTGGTCACTGGCTTTTGTTGACTGGAACACGGATCGGCAAAAAGGC	600
2NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692     TATACCCATGTATTTTGGTCACTGGCTTTTGTTGACTGGAACACGGATCGGCAAAAAGGC	600
                                                           ************************************************************

NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692      TGGCAACACTCTTATGATAATATTATGCGCCAAATTCATCCTGGCTGTATCCTGCTTCTT	660
2NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692     TGGCAACACTCTTATGATAATATTATGCGCCAAATTCATCCTGGCTGTATCCTGCTTCTT	660
                                                           ************************************************************

NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692      CACACTGTTTCGAAGGATAATGCCGATGCATTGGAAAAAGCCATTCAGGATTTAAAAAAG	720
2NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692     CACACTGTTTCGAAGGATAATGCCGATGCATTGGAAAAAGCCATTCAGGATTTAAAAAAG	720
                                                           ************************************************************

NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692      CGCGGCTATTCATTTAAAAGCCTTGATGATCTAACCTGGGAACAGGCGATTAAAGAGAGA	780
2NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692     CGCGGCTATTCATTTAAAAGCCTTGATGATCTAACCTGGGAACAGGCGATTAAAGAGAGA	780
                                                           ************************************************************

NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692      ATGCTGTACTGA	792
2NZ_QGTW01000013.1_cds_WP_110066703.1_4178_MMP09074692     ATGCTGTACTGA	792
                                                           ************
```

105

105

Almost all sequences are real duplicates

Join MarRef and MarDB without nucleotides in a single fasta:

```
$ cat marref_nucleotides_V6.fna mardbnoDup_nucleotides.fna > mar_nucleotides.fna &
```

Rename sequences to avoid long names:

Custom python script `rename_mar_seqs.py`
Copy to folder where sequences are, check permissions, and run it:

```
$ chmod u+x rename_mar_seqs.py
$ python rename_mar_seqs.py 1>log_rename.txt 2>&1 &
```

```
from Bio.SeqIO.FastaIO import SimpleFastaParser

newseqs = {}
equivalent = {}
# parse fasta file with the low-level SimpleFastaParser, reads it as a tuple
with open("marref_sample10pc.fna") as sequences:
for k, seq in enumerate(SimpleFastaParser(sequences)):
newseqs[k]=seq[1]
equivalent[k]=seq[0]

ofile = open("marsample.fna", "w")
for i in newseqs.keys():
ofile.write(">{}\n{}\n".format(i, newseqs[i]))
ofile.close()

ofile = open("marsample_names_equivalence.txt", "w")
for i in equivalent.keys():
ofile.write("{}\t{}\n".format(i, equivalent[i]))
ofile.close()
```

Translate from DNA to proteins:

With gene prediction - using Prodigal (full database takes ~3 days):

```
$ prodigal -i mar_renamed_nucleotides.fna -o mar_gene_coords.gbk -a mar_proteintrans.faa -p meta
2>log_prodigal.err &
```

Small dataset as example for this meeting - 10% of MarRef database:

```
$ prodigal -i marsample.fna -o marsample_gene_coords.gbk -a marsample_proteintrans.faa -p meta
2>log_sample_prod.txt &
```

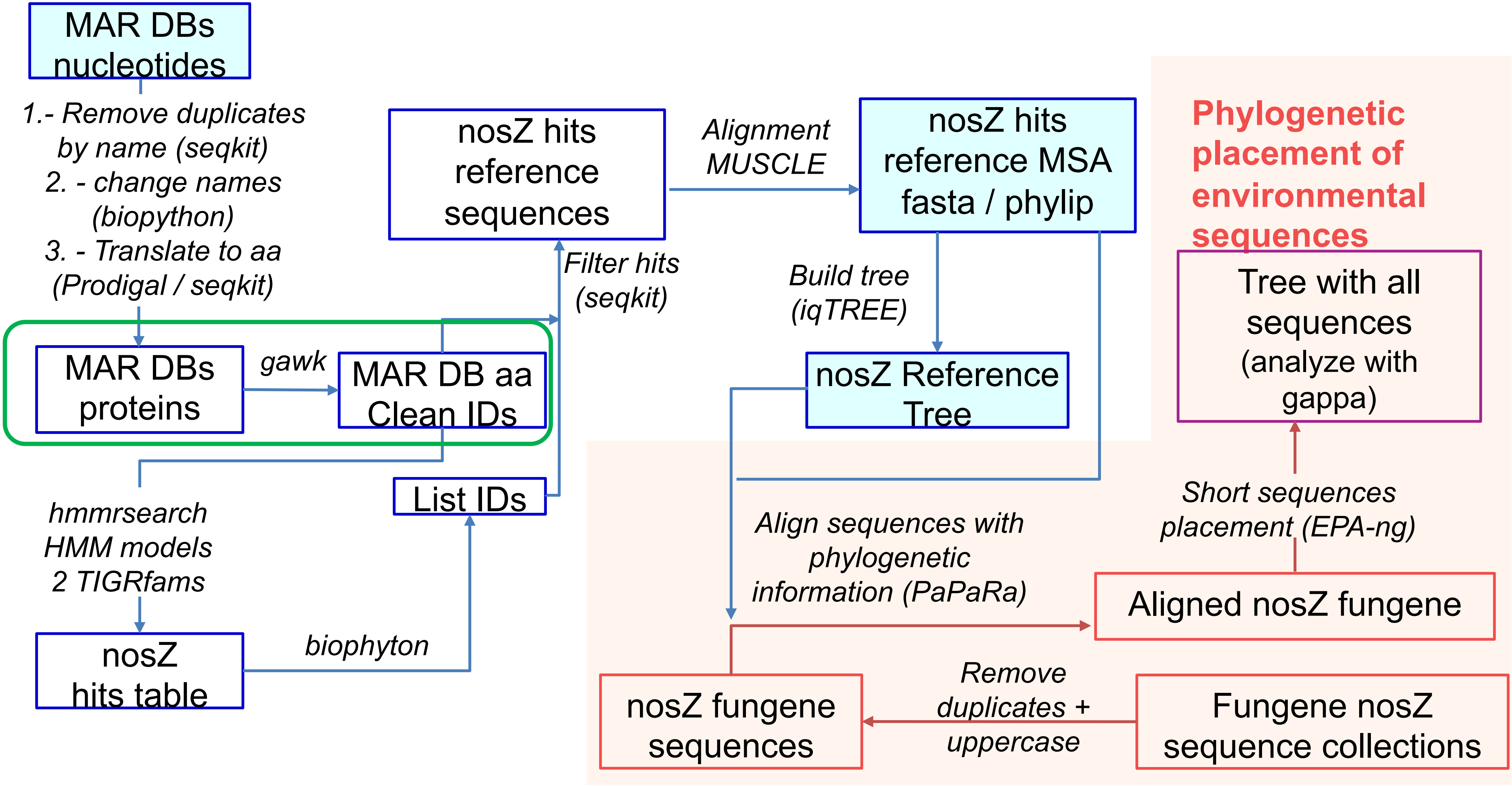
Without gene prediction – using seqkit:

-T 11: genetic code to use: 11 is The Bacterial, Archaeal and Plant Plastid Code

--trim: remove the stop codón (just done because MUSCLE complains)

```
$ seqkit translate -T 11 --trim mar_renamed_nucleotides.fna > mar_proteintrans_v2.faa
2>log_seqkit.txt &
```

Overview



MAR DBs nucleotides

1.- Remove duplicates by name (seqkit)
2. - change names (biopython)
3. - Translate to aa (Prodigal / seqkit)

MAR DBs proteins

gawk

MAR DB aa Clean IDs

hmmrsearch HMM models 2 TIGRfams

nosZ hits table

biophyton

List IDs

Filter hits (seqkit)

nosZ hits reference sequences

Alignment MUSCLE

nosZ hits reference MSA fasta / phylip

Build tree (iqTREE)

nosZ Reference Tree

Phylogenetic placement of environmental sequences

Fungene nosZ sequence collections

Remove duplicates + uppercase

nosZ fungene sequences

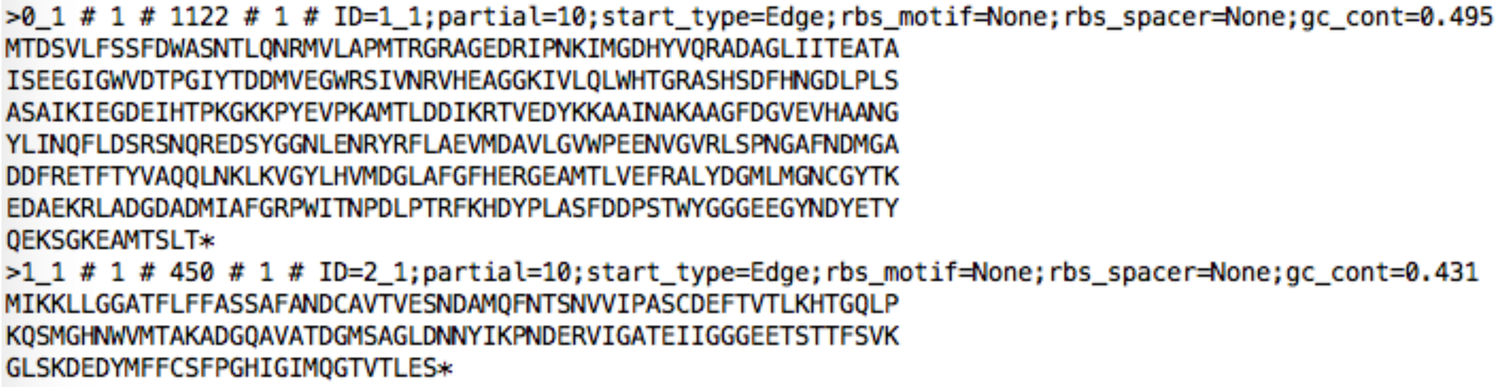
Align sequences with phylogenetic information (PaPaRa)

Aligned nosZ fungene

Short sequences placement (EPA-ng)

Tree with all sequences (analyze with gappa)

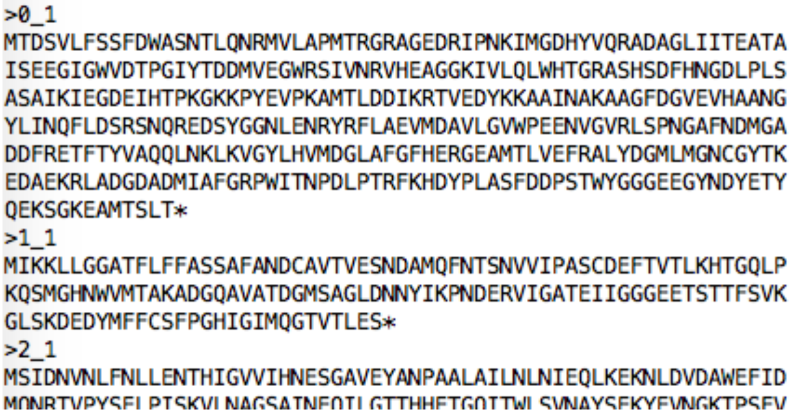
Clean fasta headers of Prodigal output



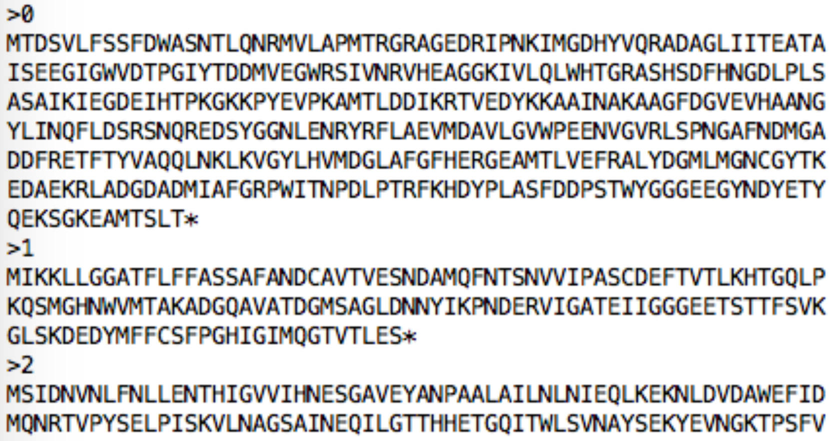
```
>0_1 # 1 # 1122 # 1 # ID=1_1;partial=10;start_type=Edge;rbs_motif=None;rbs_spacer=None;gc_cont=0.495
MTDSVLFSSFDWASNTLQNRMVLAPMTRGRAGEDRIPNKIMGDHYVQRADAGLIITEATA
ISEEGIGWVDTPGIYTDDMVEGWRSIVNRVHEAGGKIVLQLWHTGRASHSDFHNGDLPLS
ASAIKIEGDEIHTPKGKKPYEVPKAMTLDDIKRTVEDYKKAAINAKAAGFDGVEVHAANG
YLINQFLDSRSNQREDSYGGNLENRYRFLAEVMDAVLGVWPEENVGVRLSPNGAFNDMGA
DDFRETFTYVAQQLNKLKVGYLHVMDGLAFGFHERGEAMTLVEFRALYDGMLMGNCGYTK
EDAEKRLADGDADMIAFGRPWITNPDLPTRFKHDYPLASFDDPSTWYGGGEEGYNDYETY
QEKSGKEAMTSLT*
>1_1 # 1 # 450 # 1 # ID=2_1;partial=10;start_type=Edge;rbs_motif=None;rbs_spacer=None;gc_cont=0.431
MIKKLLGGATFLFFASSAFANDCAVTVESNDAMQFNTSNVVIPASCDEFTVTLKHTGQLP
KQSMGHNWVMTAKADGQAVATDGMSAGLDNNYIKPNDERVIGATEIIGGGEETSTTFSVK
GLSKDEDYMFFCSFPGHIGIMQGTVTLES*
```

```
$ gawk 'BEGUIN{FS="#"}{if ($1 ~ /^>/) print $1; else print $0}' marsample_proteintrans.faa |\
 sed 's/_1//g' >marsample_proteintrans_clean.faa
```

```
$ gawk 'BEGUIN{FS="#"}{if ($1 ~ /^>/) print $1; else print $0}' mar_proteintrans.faa |\
 sed 's/_1//g' > mar_proteintrans_clean.faa 2>err_gawk.txt &
```

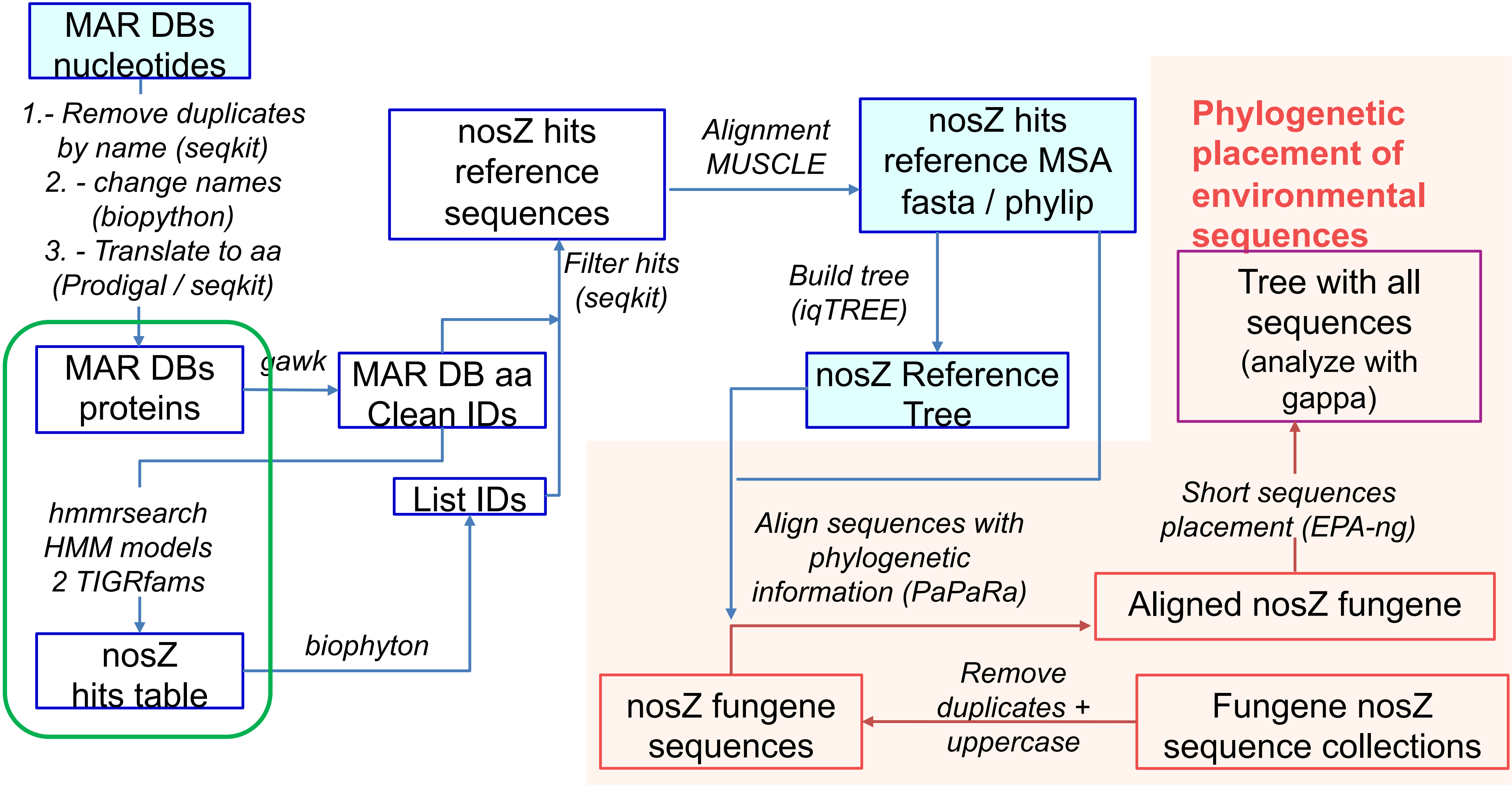


```
>0_1
MTDSVLFSSFDWASNTLQNRMVLAPMTRGRAGEDRIPNKIMGDHYVQRADAGLIITEATA
ISEEGIGWVDTPGIYTDDMVEGWRSIVNRVHEAGGKIVLQLWHTGRASHSDFHNGDLPLS
ASAIKIEGDEIHTPKGKKPYEVPKAMTLDDIKRTVEDYKKAAINAKAAGFDGVEVHAANG
YLINQFLDSRSNQREDSYGGNLENRYRFLAEVMDAVLGVWPEENVGVRLSPNGAFNDMGA
DDFRETFTYVAQQLNKLKVGYLHVMDGLAFGFHERGEAMTLVEFRALYDGMLMGNCGYTK
EDAEKRLADGDADMIAFGRPWITNPDLPTRFKHDYPLASFDDPSTWYGGGEEGYNDYETY
QEKSGKEAMTSLT*
>1_1
MIKKLLGGATFLFFASSAFANDCAVTVESNDAMQFNTSNVVIPASCDEFTVTLKHTGQLP
KQSMGHNWVMTAKADGQAVATDGMSAGLDNNYIKPNDERVIGATEIIGGGEETSTTFSVK
GLSKDEDYMFFCSFPGHIGIMQGTVTLES*
>2_1
MSIDNVNLFNLLENTHIGVVIHNESGAVEYANPAALAILNLNIEQLKEKNLDVDAWEFID
MQNRTVPYSELPISKVLNAGSAINEQILGTTHHETGQITWLSVNAYSEKYEVNGKTPSFV
```



```
>0
MTDSVLFSSFDWASNTLQNRMVLAPMTRGRAGEDRIPNKIMGDHYVQRADAGLIITEATA
ISEEGIGWVDTPGIYTDDMVEGWRSIVNRVHEAGGKIVLQLWHTGRASHSDFHNGDLPLS
ASAIKIEGDEIHTPKGKKPYEVPKAMTLDDIKRTVEDYKKAAINAKAAGFDGVEVHAANG
YLINQFLDSRSNQREDSYGGNLENRYRFLAEVMDAVLGVWPEENVGVRLSPNGAFNDMGA
DDFRETFTYVAQQLNKLKVGYLHVMDGLAFGFHERGEAMTLVEFRALYDGMLMGNCGYTK
EDAEKRLADGDADMIAFGRPWITNPDLPTRFKHDYPLASFDDPSTWYGGGEEGYNDYETY
QEKSGKEAMTSLT*
>1
MIKKLLGGATFLFFASSAFANDCAVTVESNDAMQFNTSNVVIPASCDEFTVTLKHTGQLP
KQSMGHNWVMTAKADGQAVATDGMSAGLDNNYIKPNDERVIGATEIIGGGEETSTTFSVK
GLSKDEDYMFFCSFPGHIGIMQGTVTLES*
>2
MSIDNVNLFNLLENTHIGVVIHNESGAVEYANPAALAILNLNIEQLKEKNLDVDAWEFID
MQNRTVPYSELPISKVLNAGSAINEQILGTTHHETGQITWLSVNAYSEKYEVNGKTPSFV
```

Overview



MAR DBs nucleotides

1.- Remove duplicates by name (seqkit)

2. - change names (biopython)

3. - Translate to aa (Prodigal / seqkit)

MAR DBs proteins

gawk

MAR DB aa Clean IDs

hmmrsearch HMM models 2 TIGRfams

nosZ hits table

biophyton

List IDs

Filter hits (seqkit)

nosZ hits reference sequences

Alignment MUSCLE

nosZ hits reference MSA fasta / phylip

Build tree (iqTREE)

nosZ Reference Tree

Phylogenetic placement of environmental sequences

Tree with all sequences (analyze with gappa)

Short sequences placement (EPA-ng)

Aligned nosZ fungene

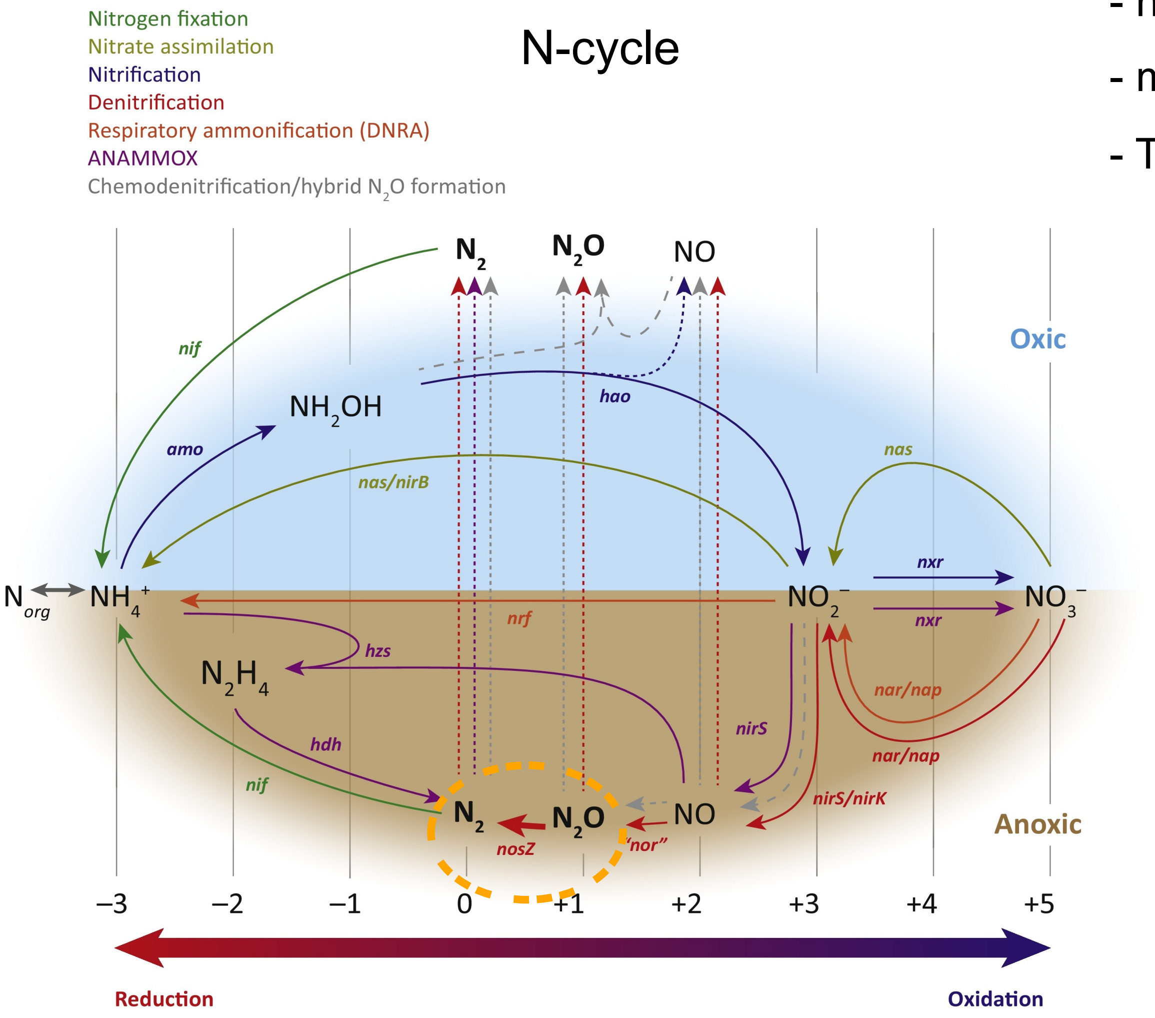
Align sequences with phylogenetic information (PaPaRa)

nosZ fungene sequences

Remove duplicates + uppercase

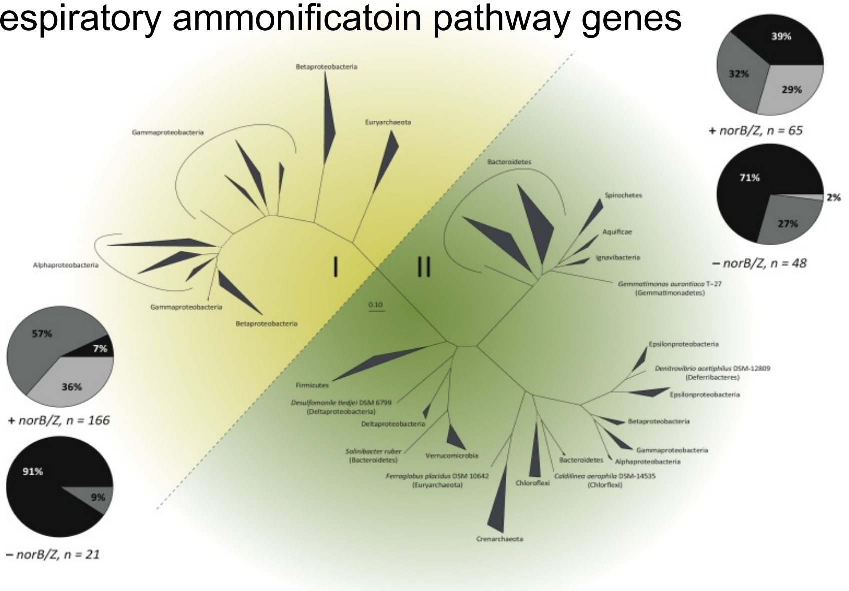
Fungene nosZ sequence collections

Search for a specific protein – Nitrous oxide reductase (NosZ) – N_2O reduction trait



Trends in Microbiology

- high diversity (12 phyla)
- mostly vertical inheritance
- Two clades (I and II). Differences:
  - associated N_2O membrane translocation pathway
  - *nos* gene cluster organization
  - frequencies of co-ocurrence with other denitrification or respiratory ammonificatoin pathway genes



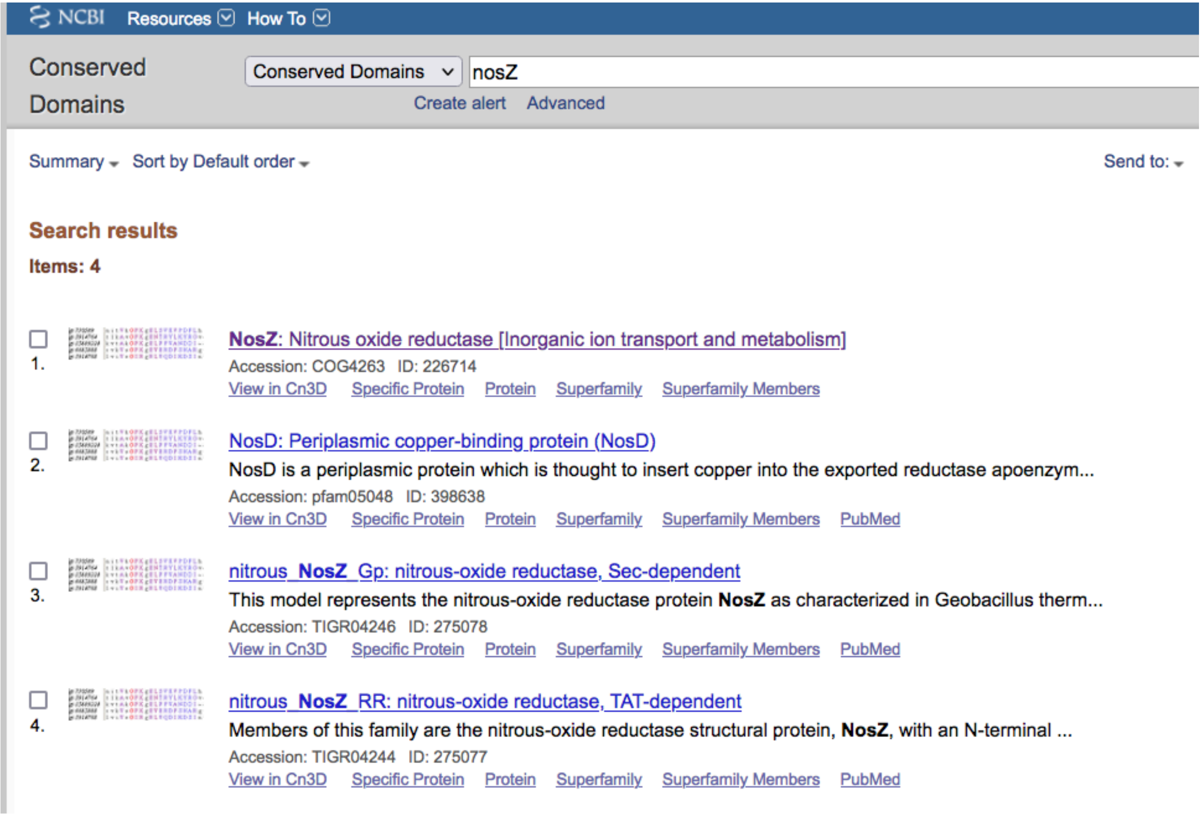
Hallin et al. 2018

Trends in Microbiology

Build a reliable nosZ database and reference tree: search nosZ in the full database

Hiden Markov Models (HMMs) – look for them:

- NCBI conserved domain database: https://www.ncbi.nlm.nih.gov/cdd/
- Fungene database: http://fungene.cme.msu.edu/



Search results

Items: 4

☐ 1. NosZ: Nitrous oxide reductase [Inorganic ion transport and metabolism]
Accession: COG4263 ID: 226714
View in Cn3D Specific Protein Protein Superfamily Superfamily Members

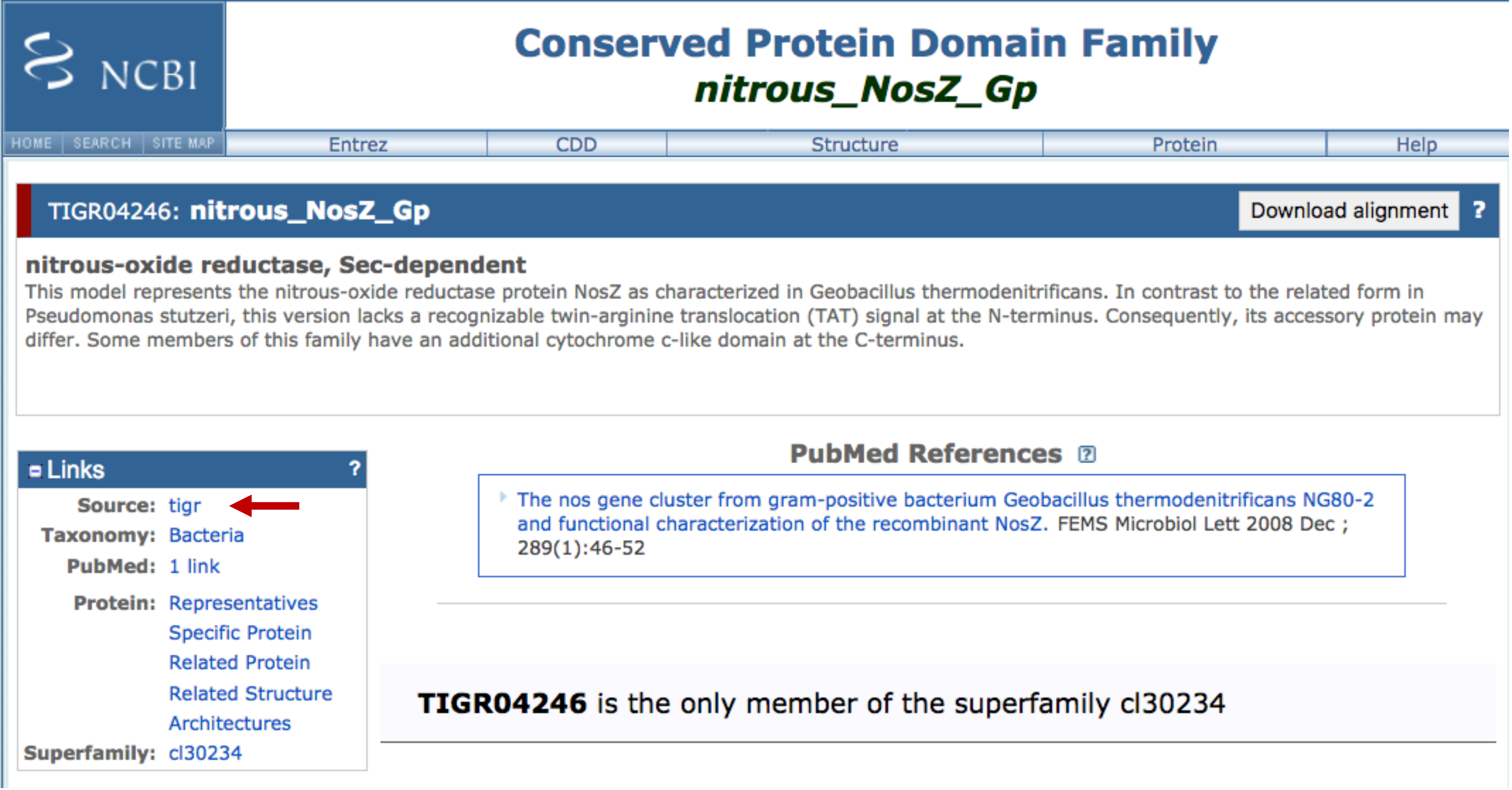
☐ 2. NosD: Periplasmic copper-binding protein (NosD)
NosD is a periplasmic protein which is thought to insert copper into the exported reductase apoenzym...
Accession: pfam05048 ID: 398638
View in Cn3D Specific Protein Protein Superfamily Superfamily Members PubMed

☐ 3. nitrous_NosZ_Gp: nitrous-oxide reductase, Sec-dependent
This model represents the nitrous-oxide reductase protein **NosZ** as characterized in Geobacillus therm...
Accession: TIGR04246 ID: 275078
View in Cn3D Specific Protein Protein Superfamily Superfamily Members PubMed

☐ 4. nitrous_NosZ_RR: nitrous-oxide reductase, TAT-dependent
Members of this family are the nitrous-oxide reductase structural protein, **NosZ**, with an N-terminal ...
Accession: TIGR04244 ID: 275077
View in Cn3D Specific Protein Protein Superfamily Superfamily Members PubMed

Two TIGRfams:
type I and II

Protocol for building our own HMMs??



Conserved Protein Domain Family
nitrous_NosZ_Gp

TIGR04246: **nitrous_NosZ_Gp**

nitrous-oxide reductase, Sec-dependent

This model represents the nitrous-oxide reductase protein NosZ as characterized in Geobacillus thermodenitrificans. In contrast to the related form in Pseudomonas stutzeri, this version lacks a recognizable twin-arginine translocation (TAT) signal at the N-terminus. Consequently, its accessory protein may differ. Some members of this family have an additional cytochrome c-like domain at the C-terminus.

Links ?

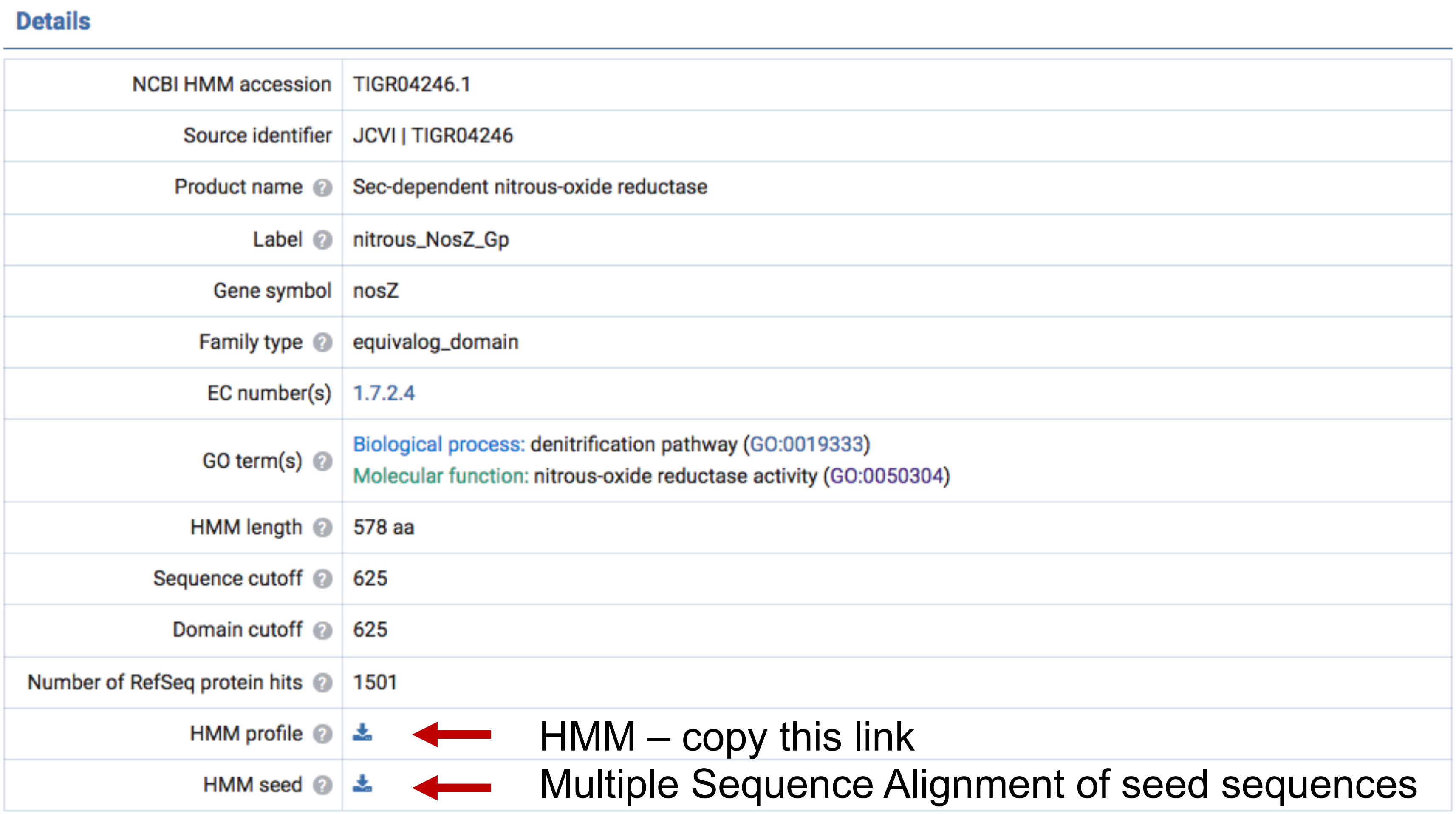
Source: tigr
Taxonomy: Bacteria
PubMed: 1 link

Protein: Representatives
Specific Protein
Related Protein
Related Structure
Architectures
Superfamily: cl30234

PubMed References

The nos gene cluster from gram-positive bacterium Geobacillus thermodenitrificans NG80-2 and functional characterization of the recombinant NosZ. FEMS Microbiol Lett 2008 Dec ; 289(1):46-52

TIGR04246 is the only member of the superfamily cl30234



Details

NCBI HMM accession	TIGR04246.1
Source identifier	JCVI | TIGR04246
Product name ?	Sec-dependent nitrous-oxide reductase
Label ?	nitrous_NosZ_Gp
Gene symbol	nosZ
Family type ?	equivalog_domain
EC number(s)	1.7.2.4
GO term(s) ?	Biological process: denitrification pathway (GO:0019333) Molecular function: nitrous-oxide reductase activity (GO:0050304)
HMM length ?	578 aa
Sequence cutoff ?	625
Domain cutoff ?	625
Number of RefSeq protein hits ?	1501
HMM profile ?	⬇
HMM seed ?	⬇

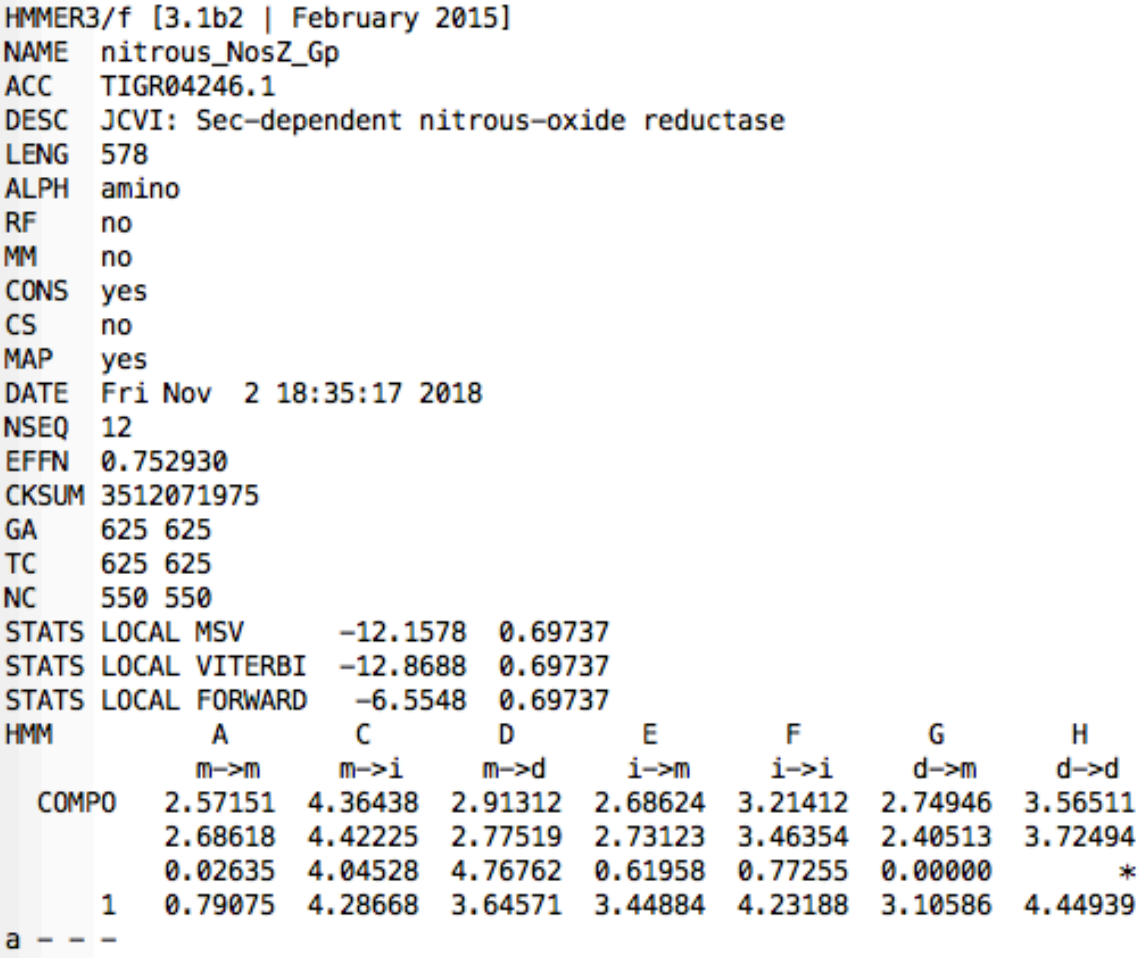
HMM – copy this link

Multiple Sequence Alignment of seed sequences

```
$ wget https://ftp.ncbi.nlm.nih.gov/hmm/current/hmm_PGAP.HMM/TIGR04246.1.HMM
```

HMMs beginning

```
$ less TIGR04246.1.HMM
```

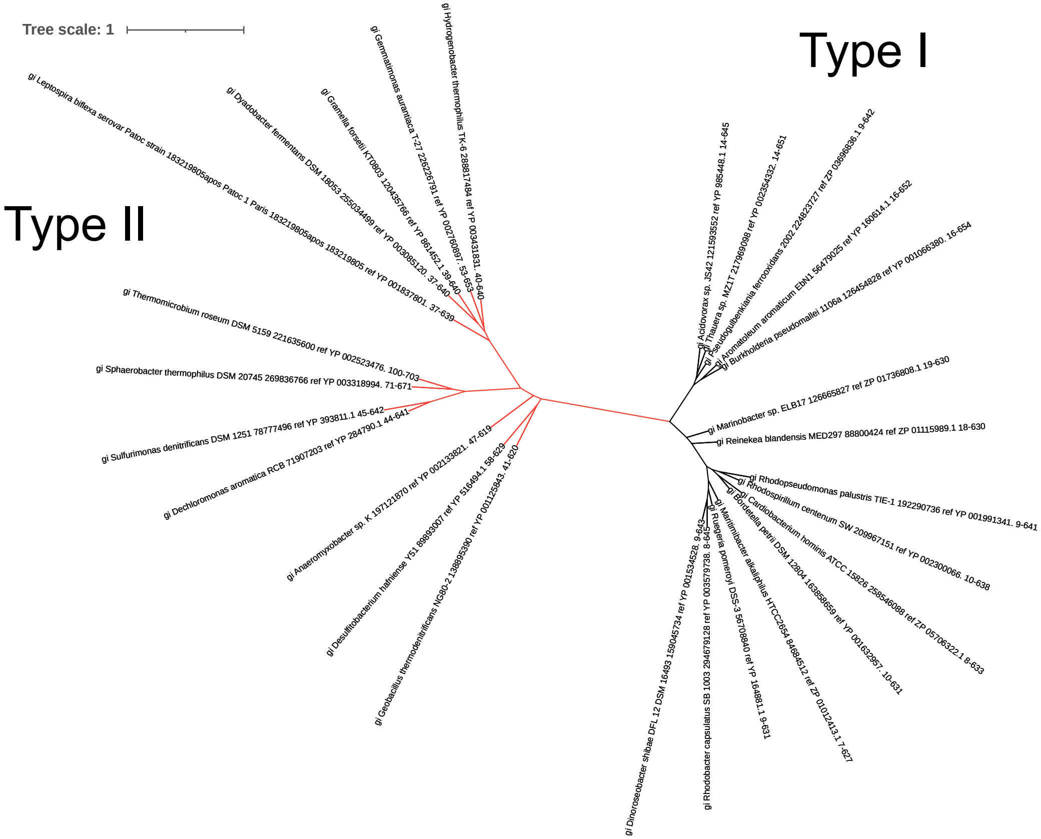


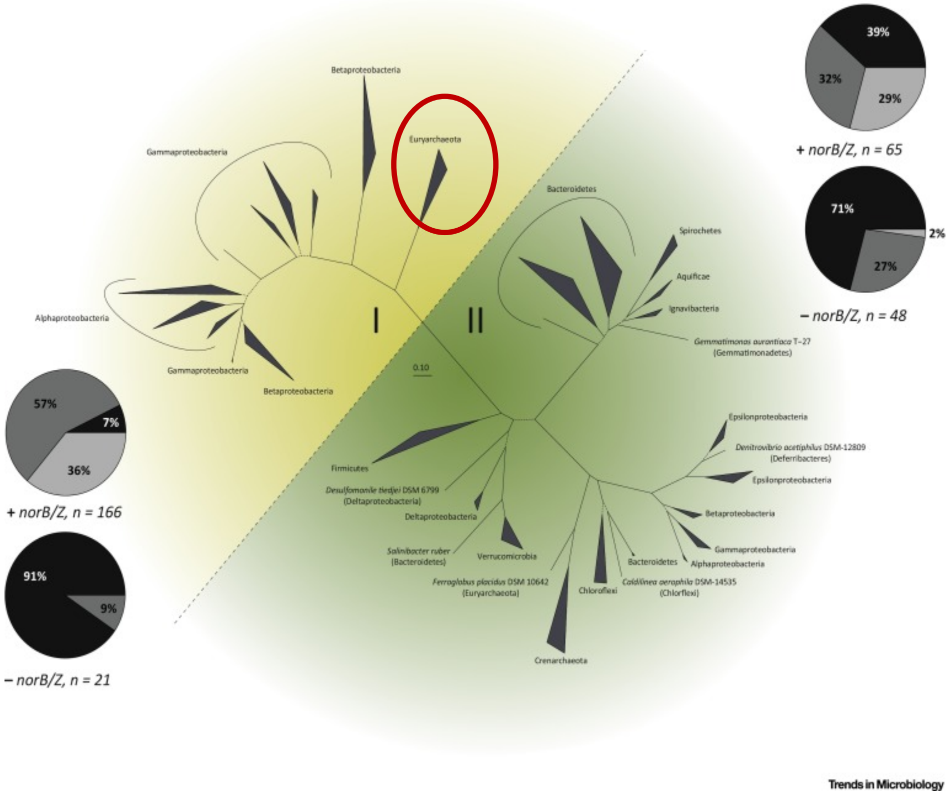
```
HMMER3/f [3.1b2 | February 2015]
NAME  nitrous_NosZ_Gp
ACC   TIGR04246.1
DESC  JCVI: Sec-dependent nitrous-oxide reductase
LENG  578
ALPH  amino
RF    no
MM    no
CONS  yes
CS    no
MAP   yes
DATE  Fri Nov  2 18:35:17 2018
NSEQ  12
EFFN  0.752930
CKSUM 3512071975
GA    625 625
TC    625 625
NC    550 550
STATS LOCAL MSV       -12.1578  0.69737
STATS LOCAL VITERBI   -12.8688  0.69737
STATS LOCAL FORWARD    -6.5548  0.69737
HMM          A        C        D        E        F        G        H
            m->m     m->i     m->d     i->m     i->i     d->m     d->d
  COMPO   2.57151  4.36438  2.91312  2.68624  3.21412  2.74946  3.56511
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494
          0.02635  4.04528  4.76762  0.61958  0.77255  0.00000        *
      1   0.79075  4.28668  3.64571  3.44884  4.23188  3.10586  4.44939
a - - -
```

2 HMMs: put them toguether in a single file:

```
$ cat TIGR04244.1.HMM TIGR04246.1.HMM > TIGRnosZ.HMM
```

I used the 27 seed sequences to check the diversity used to build those HMMs:
Missing a lot of known diversity – i.e. *Archaea*


Tree scale: 1
Type I
Type II
gi Hydrogenobacter thermophilus TK-6 288817484 ref YP 003431831. 40-640
gi Gemmatimonas aurantiaca T-27 226226791 ref YP 002760897. 53-653
gi Gramella forsetii KT0803 120435766 ref YP 861452.1 39-640
gi Dyadobacter fermentans DSM 18053 255034499 ref YP 003085120. 37-640
gi Leptospira biflexa serovar Patoc strain 183219805apos Patoc 1 Paris 183219805apos 183219805 ref YP 001837801. 37-639
gi Thermomicrobium roseum DSM 5159 221635600 ref YP 002523476. 100-703
gi Sphaerobacter thermophilus DSM 20745 269836766 ref YP 003318994. 71-671
gi Sulfurimonas denitrificans DSM 1251 78777496 ref YP 393811.1 45-642
gi Dechloromonas aromatica RCB 71907203 ref YP 284790.1 44-641
gi Anaeromyxobacter sp. K 197121870 ref YP 002133821. 47-619
gi Desulfitobacterium hafniense Y51 89893007 ref YP 516494.1 58-629
gi Geobacillus thermodenitrificans NG80-2 138895390 ref YP 001125843. 41-620
gi Acidovorax sp. JS42 121593552 ref YP 985448.1 14-645
gi Thauera sp. MZ1T 217969098 ref YP 002354332. 14-651
gi Pseudogulbenkiania ferrooxidans 2002 224823727 ref ZP 03696836.1 9-642
gi Aromatoleum aromaticum EbN1 56479025 ref YP 160614.1 16-652
gi Burkholderia pseudomallei 1106a 126454828 ref YP 001066380. 16-654
gi Marinobacter sp. ELB17 126665827 ref ZP 01736808.1 19-630
gi Reinekea blandensis MED297 88800424 ref ZP 01115989.1 18-630
gi Rhodopseudomonas palustris TIE-1 192290736 ref YP 001991341. 9-641
gi Rhodospirillum centenum SW 209967151 ref YP 002300066. 10-638
gi Cardiobacterium hominis ATCC 15826 258546088 ref ZP 05706322.1 8-633
gi Bordetella petrii DSM 12804 163858659 ref YP 001632957. 10-631
gi Maritimibacter alkaliphilus HTCC2654 84684512 ref ZP 01012413.1 7-627
gi Ruegeria pomeroyi DSS-3 56708840 ref YP 164881.1 9-631
gi Rhodobacter capsulatus SB 1003 294679128 ref YP 003579738. 8-645
gi Dinoroseobacter shibae DFL 12 DSM 16493 159045734 ref YP 001534528. 9-643



Trends in Microbiology

Need to discuss a QC protocol for third-part HMMs??

Search nosZ in MAR databases - HMMER:

Command: *hmmsearch*
--noali: don´t keep alignment
--domtblout: output type – domain hits table (when there is more than 1 domain)
--cut-ga: use HMM gathering threshold as cutoff
--cpu: number of threads

<u>Small dataset</u>:

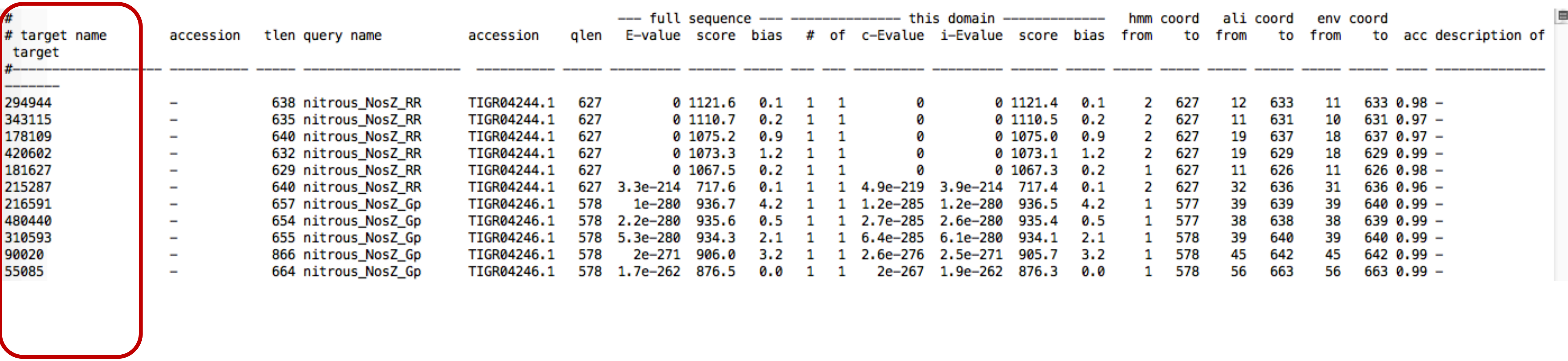
```
$ hmmsearch --noali --domtblout marsample_nosZ_tigrfam.hmm --cut_ga --cpu 15 TIGRnosZ.HMM
marsample_proteintrans_clean.faa > /dev/null 2>hmmerrlog.txt &
```

<u>Full database:</u>

```
$ hmmsearch --noali --domtblout /mnt/nfelbrus2/mar_db/mar_nosZ_tigrfam.hmm --cut_ga --cpu 15
/media/disk5/nfernandez/nosZ/TIGRnosZ/TIGRnosZ.HMM /mnt/nfelbrus2/mar_db/mar_proteintrans_clean.faa > /dev/null
2>/mnt/nfelbrus2/mar_db/hmmerrlog.txt &
```

```
$ wc -l mar_nosZ_tigrfam.hmm
1482 mar_nosZ_tigrfam.hmm. (-13 lines from file header and end): 1469 nosZ sequences
```

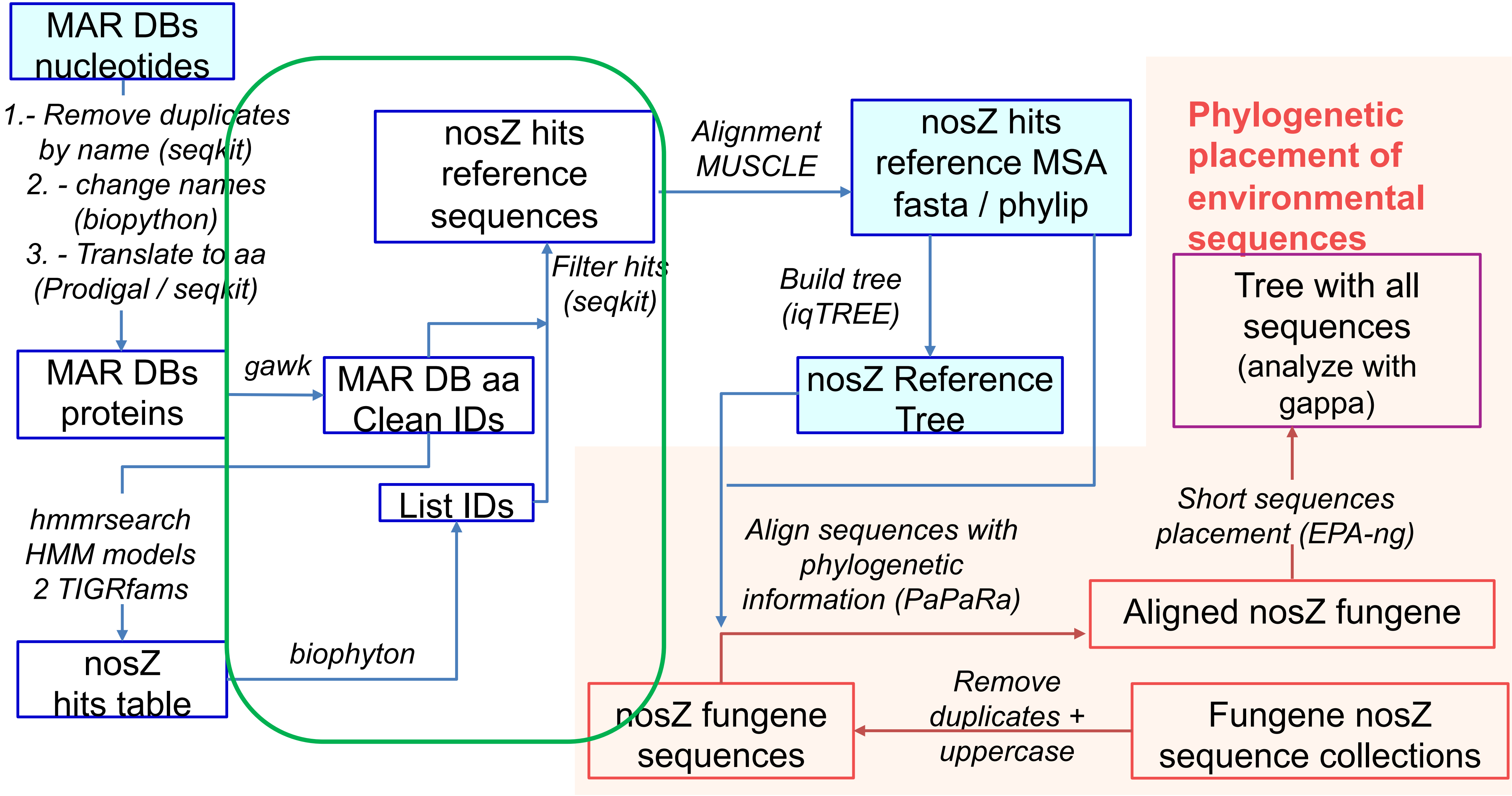
HMMER output table:



```
#                                                                             --- full sequence --- -------------- this domain -------------   hmm coord   ali coord   env coord
# target name        accession   tlen query name           accession    qlen   E-value  score  bias   #  of  c-Evalue  i-Evalue  score  bias  from    to  from    to  from    to  acc description of
 target
#------------------- ---------- ----- -------------------- ----------  ----- --------- ------ ----- --- --- --------- --------- ------ ----- ----- ----- ----- ----- ----- ----- ---- ---------------
--------
294944               -            638 nitrous_NosZ_RR      TIGR04244.1   627         0 1121.6   0.1   1   1         0         0 1121.4   0.1     2   627    12   633    11   633 0.98 -
343115               -            635 nitrous_NosZ_RR      TIGR04244.1   627         0 1110.7   0.2   1   1         0         0 1110.5   0.2     2   627    11   631    10   631 0.97 -
178109               -            640 nitrous_NosZ_RR      TIGR04244.1   627         0 1075.2   0.9   1   1         0         0 1075.0   0.9     2   627    19   637    18   637 0.97 -
420602               -            632 nitrous_NosZ_RR      TIGR04244.1   627         0 1073.3   1.2   1   1         0         0 1073.1   1.2     2   627    19   629    18   629 0.99 -
181627               -            629 nitrous_NosZ_RR      TIGR04244.1   627         0 1067.5   0.2   1   1         0         0 1067.3   0.2     1   627    11   626    11   626 0.98 -
215287               -            640 nitrous_NosZ_RR      TIGR04244.1   627 3.3e-214  717.6   0.1   1   1 4.9e-219 3.9e-214  717.4   0.1     2   627    32   636    31   636 0.96 -
216591               -            657 nitrous_NosZ_Gp      TIGR04246.1   578   1e-280  936.7   4.2   1   1 1.2e-285 1.2e-280  936.5   4.2     1   577    39   639    39   640 0.99 -
480440               -            654 nitrous_NosZ_Gp      TIGR04246.1   578 2.2e-280  935.6   0.5   1   1 2.7e-285 2.6e-280  935.4   0.5     1   577    38   638    38   639 0.99 -
310593               -            655 nitrous_NosZ_Gp      TIGR04246.1   578 5.3e-280  934.3   2.1   1   1 6.4e-285 6.1e-280  934.1   2.1     1   578    39   640    39   640 0.99 -
90020                -            866 nitrous_NosZ_Gp      TIGR04246.1   578   2e-271  906.0   3.2   1   1 2.6e-276 2.5e-271  905.7   3.2     1   578    45   642    45   642 0.99 -
55085                -            664 nitrous_NosZ_Gp      TIGR04246.1   578 1.7e-262  876.5   0.0   1   1   2e-267 1.9e-262  876.3   0.0     1   578    56   663    56   663 0.99 -
```

Hits: potential NosZ sequences IDs

Overview



MAR DBs nucleotides

1.- Remove duplicates by name (seqkit)
2. - change names (biopython)
3. - Translate to aa (Prodigal / seqkit)

MAR DBs proteins

gawk

MAR DB aa Clean IDs

hmmrsearch HMM models 2 TIGRfams

nosZ hits table

biophyton

List IDs

Filter hits (seqkit)

nosZ hits reference sequences

Alignment MUSCLE

nosZ hits reference MSA fasta / phylip

Build tree (iqTREE)

nosZ Reference Tree

Phylogenetic placement of environmental sequences

Tree with all sequences (analyze with gappa)

Short sequences placement (EPA-ng)

Aligned nosZ fungene

Align sequences with phylogenetic information (PaPaRa)

nosZ fungene sequences

Remove duplicates + uppercase

Fungene nosZ sequence collections

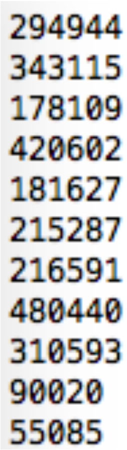
Retrieve found nosZ sequences.

Get list of IDs into a text file:

```
$ python parse_marsampleHMM.py
$ less hits_marsample.txt
```

```
$ python parse_marNosZhmm.py
$ wc -l hits_mar_nosZ.txt
1469 hits_mar_nosZ.txt
```

Output:



```
294944
343115
178109
420602
181627
215287
216591
480440
310593
90020
55085
```

```
from Bio import SearchIO
import csv

# Read the output table of database search with TIGRFAMs HMMs for nosZ
hmm_qresult = SearchIO.parse('marsample_nosZ_tigrfam.hmm',  'hmmsearch3-domtab')

# Filter table to make a list of hits IDs
hit_ids = []
for qresult in hmm_qresult:
    for i in range(len(qresult)):
        hit_ids.append(qresult[i].id)

# Write list to a text file
file = open('hits_nosZ_marsample.txt', 'w')
for index in range(len(hit_ids)):
    file.write(str(hit_ids[index]) + "\n")
file.close()
```

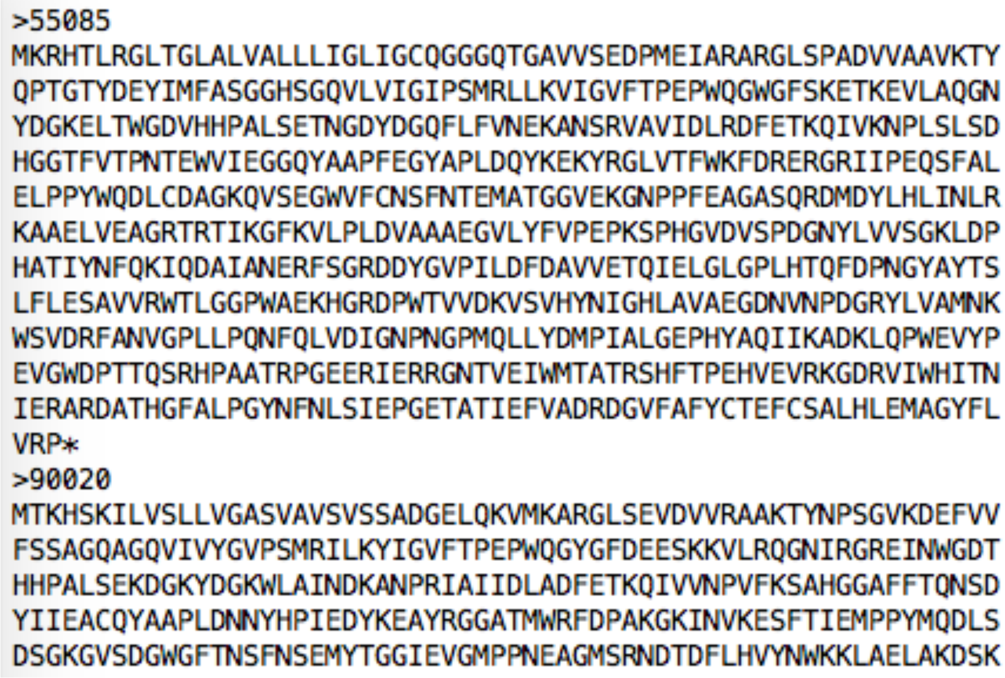
Retrieve found nosZ sequences.

Create fasta file with MAR nosZ sequences

Filter database to keep only nosZ sequences by their ID:

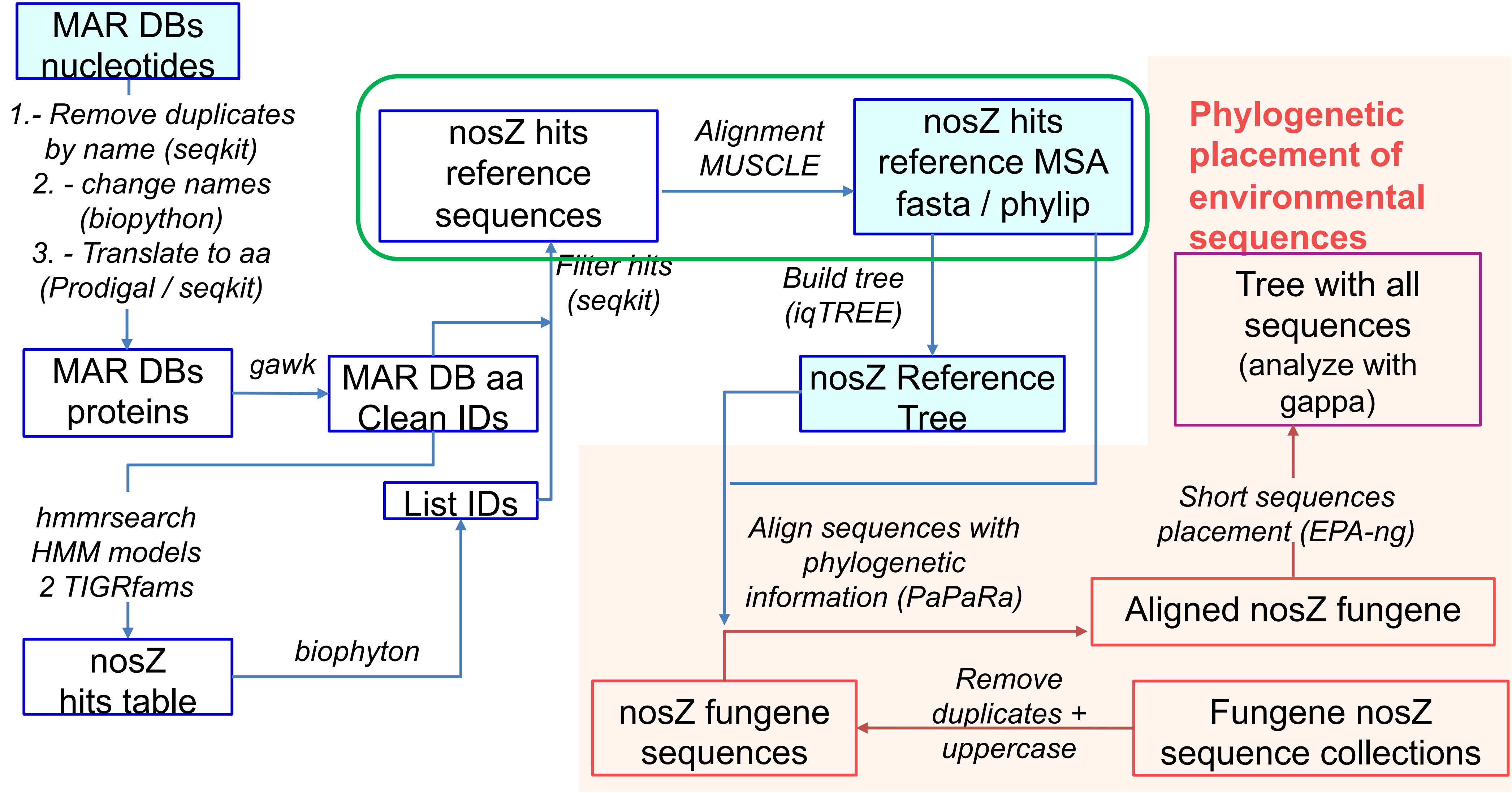
```
$ seqkit grep -f hits_nosZ_marsample.txt marsample_proteintrans_clean.faa -o hits_nosZ_marsample.faa
```

```
$ seqkit grep -f hits_mar_nosZ.txt mar_proteintrans_clean.faa -o mar_nosZ.faa &
```



```
>55085
MKRHTLRGLTGLALVALLLIGLIGCQGGGQTGAVVSEDPMEIARARGLSPADVVAAVKTY
QPTGTYDEYIMFASGGHSGQVLVIGIPSMRLLKVIGVFTPEPWQGWGFSKETKEVLAQGN
YDGKELTWGDVHHPALSETNGDYDGQFLFVNEKANSRVAVIDLRDFETKQIVKNPLSLSD
HGGTFVTPNTEWVIEGGQYAAPFEGYAPLDQYKEKYRGLVTFWKFDRERGRIIPEQSFAL
ELPPYWQDLCDAGKQVSEGWVFCNSFNTEMATGGVEKGNPPFEAGASQRDMDYLHLINLR
KAAELVEAGRTRTIKGFKVLPLDVAAAEGVLYFVPEPKSPHGVDVSPDGNYLVVSGKLDP
HATIYNFQKIQDAIANERFSGRDDYGVPILDFDAVVETQIELGLGPLHTQFDPNGYAYTS
LFLESAVVRWTLGGPWAEKHGRDPWTVVDKVSVHYNIGHLAVAEGDNVNPDGRYLVAMNK
WSVDRFANVGPLLPQNFQLVDIGNPNGPMQLLYDMPIALGEPHYAQIIKADKLQPWEVYP
EVGWDPTTQSRHPAATRPGEERIERRGNTVEIWMTATRSHFTPEHVEVRKGDRVIWHITN
IERARDATHGFALPGYNFNLSIEPGETATIEFVADRDGVFAFYCTEFCSALHLEMAGYFL
VRP*
>90020
MTKHSKILVSLLVGASVAVSVSSADGELQKVMKARGLSEVDVVRAAKTYNPSGVKDEFVV
FSSAGQAGQVIVYGVPSMRILKYIGVFTPEPWQGYGFDEESKKVLRQGNIRGREINWGDT
HHPALSEKDGKYDGKWLAINDKANPRIAIIDLADFETKQIVVNPVFKSAHGGAFFTQNSD
YIIEACQYAAPLDNNYHPIEDYKEAYRGGATMWRFDPAKGKINVKESFTIEMPPYMQDLS
DSGKGVSDGWGFTNSFNSEMYTGGIEVGMPPNEAGMSRNDTDFLHVYNWKKLAELAKDSK
```

Overview



MAR DBs nucleotides

1.- Remove duplicates by name (seqkit)
2. - change names (biopython)
3. - Translate to aa (Prodigal / seqkit)

MAR DBs proteins

gawk

MAR DB aa Clean IDs

hmmrsearch HMM models 2 TIGRfams

nosZ hits table

biophyton

List IDs

Filter hits (seqkit)

nosZ hits reference sequences

Alignment MUSCLE

nosZ hits reference MSA fasta / phylip

Build tree (iqTREE)

nosZ Reference Tree

Phylogenetic placement of environmental sequences

Fungene nosZ sequence collections

Remove duplicates + uppercase

nosZ fungene sequences

Align sequences with phylogenetic information (PaPaRa)

Aligned nosZ fungene

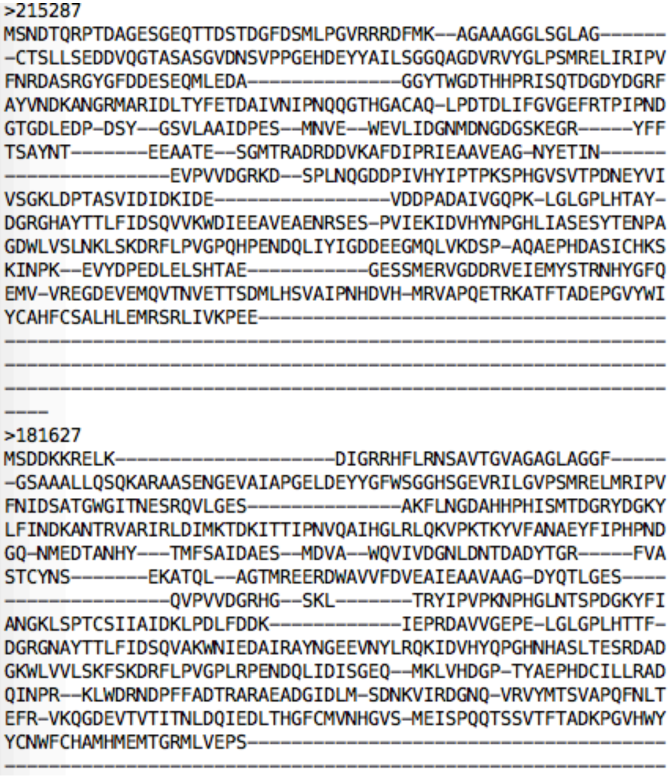
Short sequences placement (EPA-ng)

Tree with all sequences (analyze with gappa)

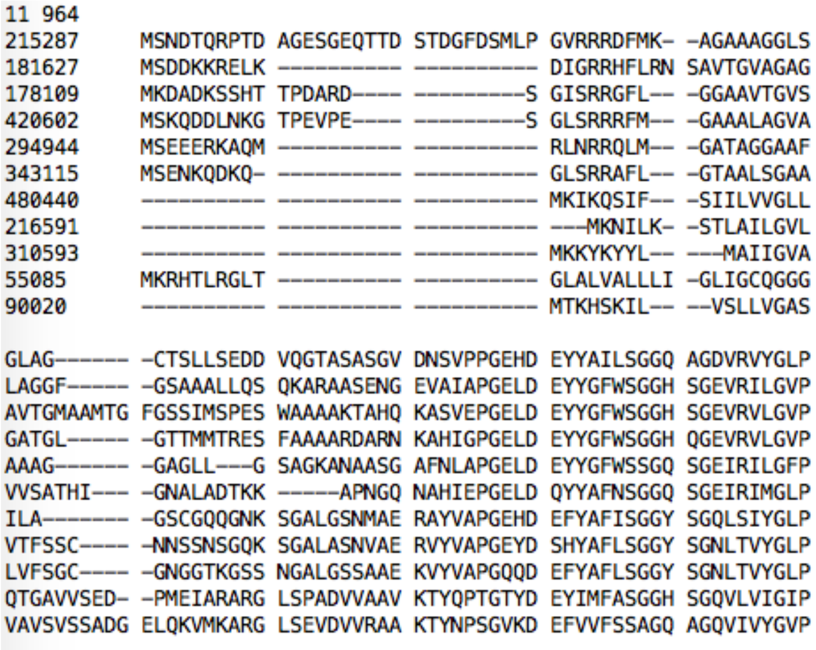
Multiple sequence alignment of MAR nosZ sequences – MUSCLE – small dataset

```
$ mkdir alignments
$ muscle -in hits_nosZ_marsample.faa -phyiout alignments/hits_nosZ_aligned.phy -fastaout
alignments/hits_nosZ_aligned.faa
```

Save output in two formats: fasta and phylip (we might need only one)

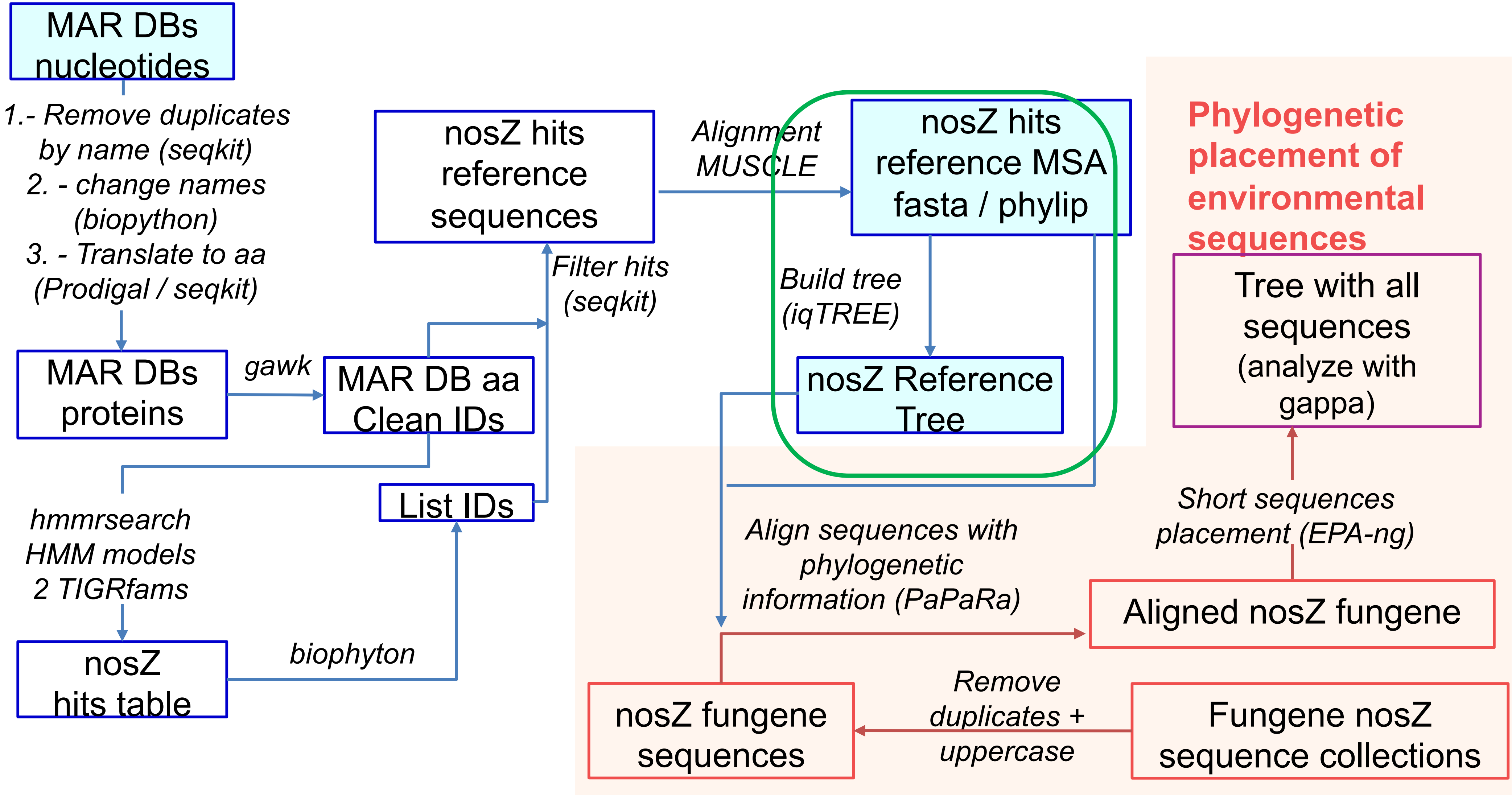


```
>215287
MSNDTQRPTDAGESGEQTTDSTDGFDSMLPGVRRRDFMK--AGAAAGGLSGLAG------
-CTSLLSEDDVQGTASASGVDNSVPPGEHDEYYAILSGGQAGDVRVYGLPSMRELIRIPV
FNRDASRGYGFDDESEQMLEDA--------------GGYTWGDTHHPRISQTDGDYDGRF
AYVNDKANGRMARIDLTYFETDAIVNIPNQQGTHGACAQ-LPDTDLIFGVGEFRTPIPND
GTGDLEDP-DSY--GSVLAAIDPES--MNVE--WEVLIDGNMDNGDGSKEGR-----YFF
TSAYNT-------EEAATE--SGMTRADRDDVKAFDIPRIEAAVEAG-NYETIN------
---------------EVPVVDGRKD--SPLNQGDDPIVHYIPTPKSPHGVSVTPDNEYVI
VSGKLDPTASVIDIDKIDE----------------VDDPADAIVGQPK-LGLGPLHTAY-
DGRGHAYTTLFIDSQVVKWDIEEAVEAENRSES-PVIEKIDVHYNPGHLIASESYTENPA
GDWLVSLNKLSKDRFLPVGPQHPENDQLIYIGDDEEGMQLVKDSP-AQAEPHDASICHKS
KINPK--EVYDPEDLELSHTAE-----------GESSMERVGDDRVEIEMYSTRNHYGFQ
EMV-VREGDEVEMQVTNVETTSDMLHSVAIPNHDVH-MRVAPQETRKATFTADEPGVYWI
YCAHFCSALHLEMRSRLIVKPEE-------------------------------------
------------------------------------------------------------
------------------------------------------------------------
------------------------------------------------------------
----
>181627
MSDDKKRELK--------------------DIGRRHFLRNSAVTGVAGAGLAGGF-----
-GSAAALLQSQKARAASENGEVAIAPGELDEYYGFWSGGHSGEVRILGVPSMRELMRIPV
FNIDSATGWGITNESRQVLGES--------------AKFLNGDAHHPHISMTDGRYDGKY
LFINDKANTRVARIRLDIMKTDKITTIPNVQAIHGLRLQKVPKTKYVFANAEYFIPHPND
GQ-NMEDTANHY---TMFSAIDAES--MDVA--WQVIVDGNLDNTDADYTGR-----FVA
STCYNS-------EKATQL--AGTMREERDWAVVFDVEAIEAAVAAG-DYQTLGES----
---------------QVPVVDGRHG--SKL-------TRYIPVPKNPHGLNTSPDGKYFI
ANGKLSPTCSIIAIDKLPDLFDDK------------IEPRDAVVGEPE-LGLGPLHTTF-
DGRGNAYTTLFIDSQVAKWNIEDAIRAYNGEEVNYLRQKIDVHYQPGHNHASLTESRDAD
GKWLVVLSKFSKDRFLPVGPLRPENDQLIDISGEQ--MKLVHDGP-TYAEPHDCILLRAD
QINPR--KLWDRNDPFFADTRARAEADGIDLM-SDNKVIRDGNQ-VRVYMTSVAPQFNLT
EFR-VKQGDEVTVTITNLDQIEDLTHGFCMVNHGVS-MEISPQQTSSVTFTADKPGVHWY
YCNWFCHAMHMEMTGRMLVEPS--------------------------------------
------------------------------------------------------------
```



```
11 964
215287    MSNDTQRPTD AGESGEQTTD STDGFDSMLP GVRRRDFMK- -AGAAAGGLS
181627    MSDDKKRELK ---------- ---------- DIGRRHFLRN SAVTGVAGAG
178109    MKDADKSSHT TPDARD---- ---------S GISRRGFL-- -GGAAVTGVS
420602    MSKQDDLNKG TPEVPE---- ---------S GLSRRRFM-- -GAAALAGVA
294944    MSEEERKAQM ---------- ---------- RLNRRQLM-- -GATAGGAAF
343115    MSENKQDKQ- ---------- ---------- GLSRRAFL-- -GTAALSGAA
480440    ---------- ---------- ---------- MKIKQSIF-- -SIILVVGLL
216591    ---------- ---------- ---------- ---MKNILK- -STLAILGVL
310593    ---------- ---------- ---------- MKKYKYYL-- ---MAIIGVA
55085     MKRHTLRGLT ---------- ---------- GLALVALLLI -GLIGCQGGG
90020     ---------- ---------- ---------- MTKHSKIL-- --VSLLVGAS

GLAG------ -CTSLLSEDD VQGTASASGV DNSVPPGEHD EYYAILSGGQ AGDVRVYGLP
LAGGF----- -GSAAALLQS QKARAASENG EVAIAPGELD EYYGFWSGGH SGEVRILGVP
AVTGMAAMTG FGSSIMSPES WAAAAKTAHQ KASVEPGELD EYYGFWSGGH SGEVRVLGVP
GATGL----- -GTTMMTRES FAAAARDARN KAHIGPGELD EYYGFWSGGH QGEVRVLGVP
AAAG------ -GAGLL---G SAGKANAASG AFNLAPGELD EYYGFWSSGQ SGEIRILGFP
VVSATHI--- -GNALADTKK -----APNGQ NAHIEPGELD QYYAFNSGGQ SGEIRIMGLP
ILA------- -GSCGQQGNK SGALGSNMAE RAYVAPGEHD EFYAFISGGY SGQLSIYGLP
VTFSSC---- -NNSSNSGQK SGALASNVAE RVYVAPGEYD SHYAFLSGGY SGNLTVYGLP
LVFSGC---- -GNGGTKGSS NGALGSSAAE KVYVAPGQQD EFYAFLSGGY SGNLTVYGLP
QTGAVVSED- -PMEIARARG LSPADVVAAV KTYQPTGTYD EYIMFASGGH SGQVLVIGIP
VAVSVSSADG ELQKVMKARG LSEVDVVRAA KTYNPSGVKD EFVVFSSAGQ AGQVIVYGVP
```

Overview



MAR DBs nucleotides

1.- Remove duplicates by name (seqkit)
2. - change names (biopython)
3. - Translate to aa (Prodigal / seqkit)

MAR DBs proteins

gawk

MAR DB aa Clean IDs

hmmrsearch HMM models 2 TIGRfams

nosZ hits table

biophyton

List IDs

Filter hits (seqkit)

nosZ hits reference sequences

Alignment MUSCLE

nosZ hits reference MSA fasta / phylip

Build tree (iqTREE)

nosZ Reference Tree

Phylogenetic placement of environmental sequences

Fungene nosZ sequence collections

Remove duplicates + uppercase

nosZ fungene sequences

Align sequences with phylogenetic information (PaPaRa)

Aligned nosZ fungene

Short sequences placement (EPA-ng)

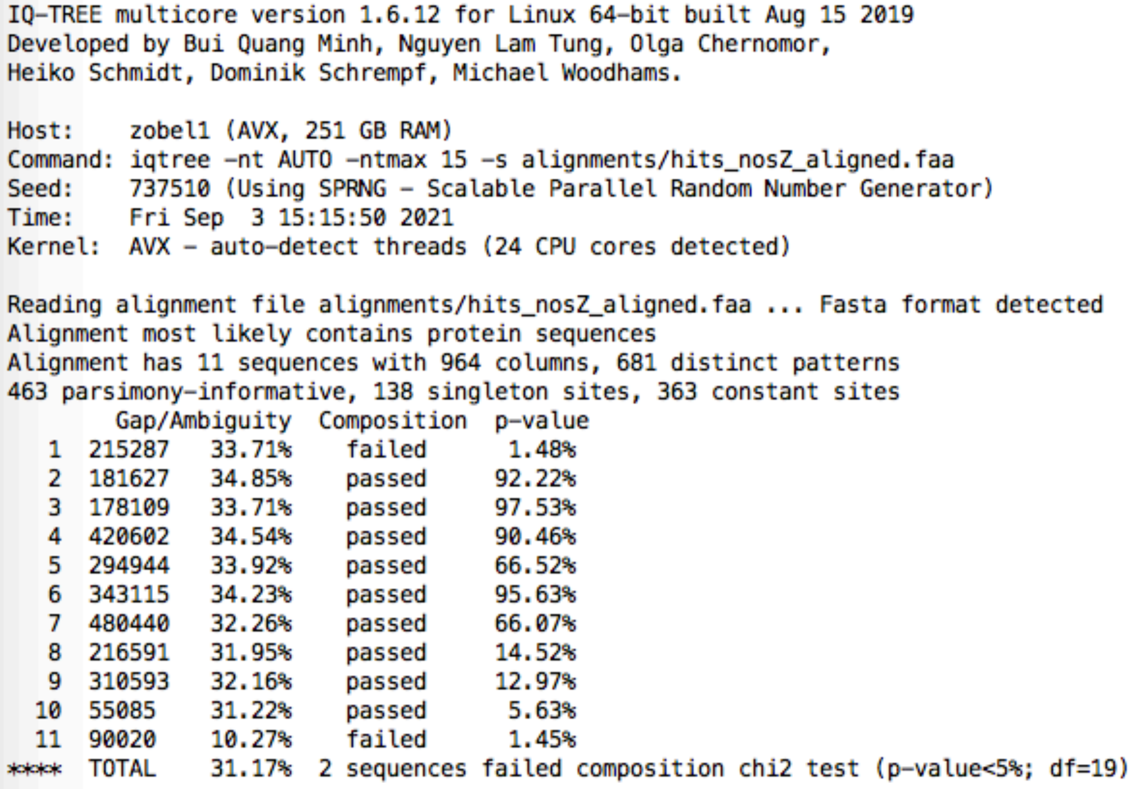
Tree with all sequences (analyze with gappa)

Build a tree of nosZ MAR sequences – iqTREE – small dataset

Two main steps:
1. Find the best substitution model for the dataset (user can specify it)
2. Build the ML tree

```
$ iqtree -nt AUTO -ntmax 15 -s alignments/hits_nosZ_aligned.faa
```

Options:
-nt AUTO : number of threads
-ntmax 15: If AUTO, max number of threads
-s: aligned sequences to build the tree
-pre: prefix for output files (not used here)
-m: model (not used here)



```
IQ-TREE multicore version 1.6.12 for Linux 64-bit built Aug 15 2019
Developed by Bui Quang Minh, Nguyen Lam Tung, Olga Chernomor,
Heiko Schmidt, Dominik Schrempf, Michael Woodhams.

Host:     zobel1 (AVX, 251 GB RAM)
Command: iqtree -nt AUTO -ntmax 15 -s alignments/hits_nosZ_aligned.faa
Seed:     737510 (Using SPRNG - Scalable Parallel Random Number Generator)
Time:     Fri Sep  3 15:15:50 2021
Kernel:  AVX - auto-detect threads (24 CPU cores detected)

Reading alignment file alignments/hits_nosZ_aligned.faa ... Fasta format detected
Alignment most likely contains protein sequences
Alignment has 11 sequences with 964 columns, 681 distinct patterns
463 parsimony-informative, 138 singleton sites, 363 constant sites
        Gap/Ambiguity  Composition  p-value
   1  215287   33.71%    failed      1.48%
   2  181627   34.85%    passed     92.22%
   3  178109   33.71%    passed     97.53%
   4  420602   34.54%    passed     90.46%
   5  294944   33.92%    passed     66.52%
   6  343115   34.23%    passed     95.63%
   7  480440   32.26%    passed     66.07%
   8  216591   31.95%    passed     14.52%
   9  310593   32.16%    passed     12.97%
  10  55085    31.22%    passed      5.63%
  11  90020    10.27%    failed      1.45%
****  TOTAL    31.17%  2 sequences failed composition chi2 test (p-value<5%; df=19)
```

MSA of MAR nosZ sequences and tree - full database

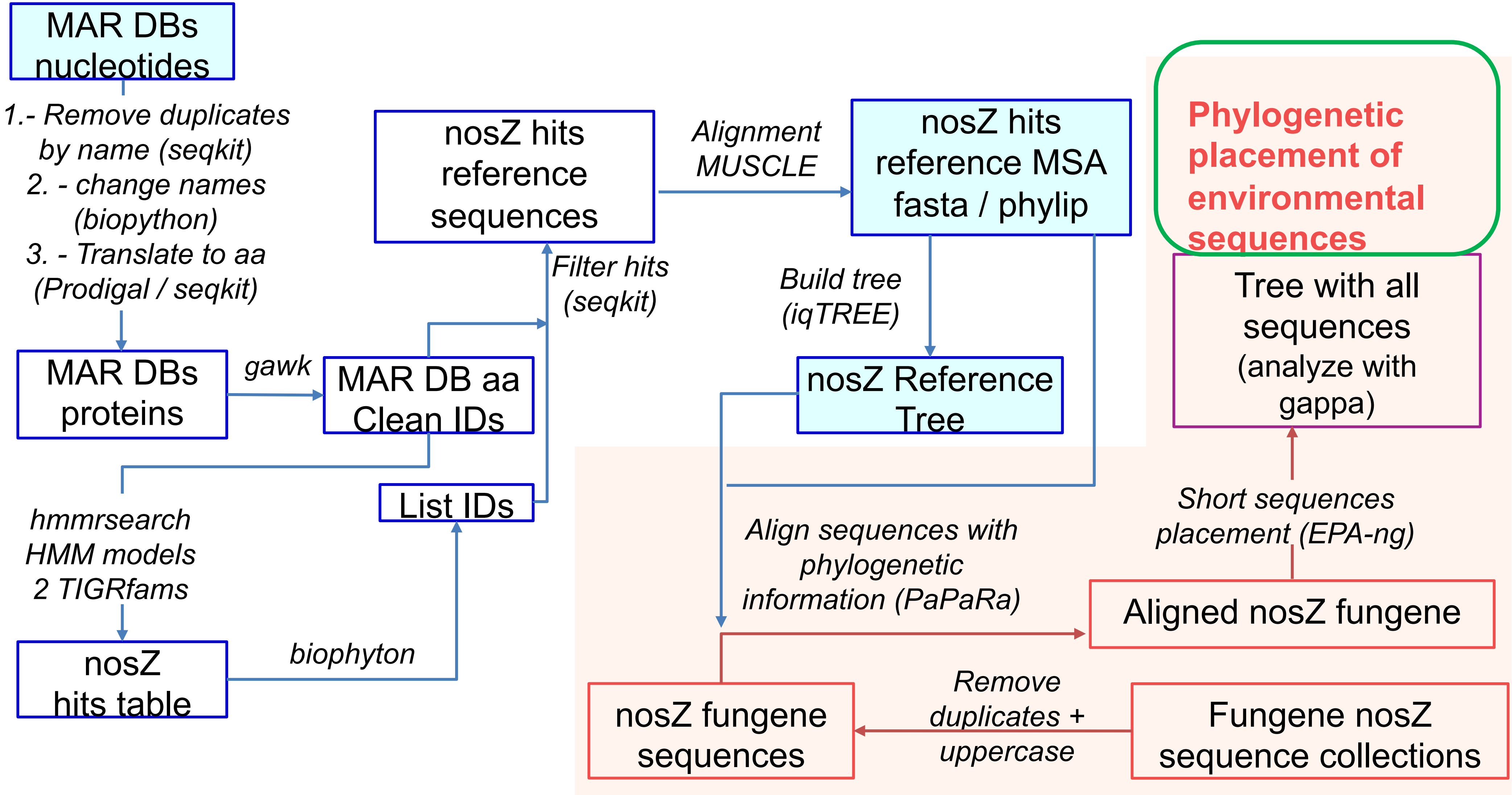
Align with MUSCLE

```
$ mkdir alignments
$ muscle -in mar_nosZ.faa -phyiout alignments/mar_nosZ_aligned.phy -fastaout
alignments/mar_nosZ_aligned.faa 1>muscle_log.txt 2>&1 &
```

Tree with iqTREE

```
$ iqtree -nt AUTO -ntmax 18 -s alignments/mar_nosZ_aligned.phy 1>log_iqtree.txt 2>&1 &
```

Overview



MAR DBs nucleotides

1.- Remove duplicates by name (seqkit)
2. - change names (biopython)
3. - Translate to aa (Prodigal / seqkit)

MAR DBs proteins

gawk

MAR DB aa Clean IDs

hmmrsearch HMM models 2 TIGRfams

nosZ hits table

biophyton

List IDs

Filter hits (seqkit)

nosZ hits reference sequences

Alignment MUSCLE

nosZ hits reference MSA fasta / phylip

Build tree (iqTREE)

nosZ Reference Tree

Fungene nosZ sequence collections

Remove duplicates + uppercase

nosZ fungene sequences

Align sequences with phylogenetic information (PaPaRa)

Aligned nosZ fungene

Short sequences placement (EPA-ng)

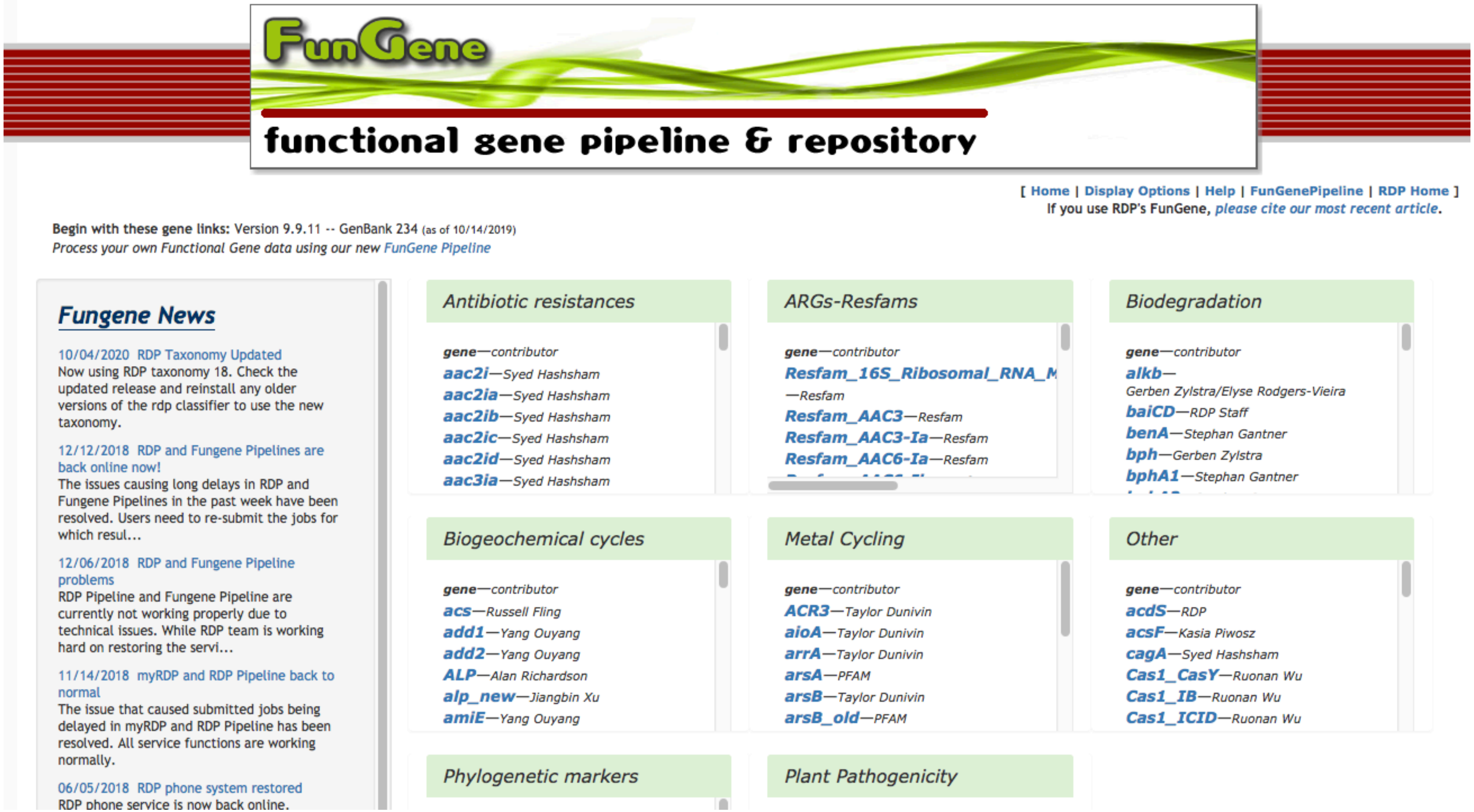
Tree with all sequences (analyze with gappa)

Phylogenetic placement of environmental sequences

Retrieve and prepare environmental sequences

I have cheated…protocol not developed yet.
My turnaround:

http://fungene.cme.msu.edu/



FunGene
functional gene pipeline & repository

[Home | Display Options | Help | FunGenePipeline | RDP Home]
If you use RDP's FunGene, *please cite our most recent article*.

Begin with these gene links: Version 9.9.11 -- GenBank 234 (as of 10/14/2019)
Process your own Functional Gene data using our new FunGene Pipeline

Fungene News

10/04/2020 RDP Taxonomy Updated
Now using RDP taxonomy 18. Check the updated release and reinstall any older versions of the rdp classifier to use the new taxonomy.

12/12/2018 RDP and Fungene Pipelines are back online now!
The issues causing long delays in RDP and Fungene Pipelines in the past week have been resolved. Users need to re-submit the jobs for which resul...

12/06/2018 RDP and Fungene Pipeline problems
RDP Pipeline and Fungene Pipeline are currently not working properly due to technical issues. While RDP team is working hard on restoring the servi...

11/14/2018 myRDP and RDP Pipeline back to normal
The issue that caused submitted jobs being delayed in myRDP and RDP Pipeline has been resolved. All service functions are working normally.

06/05/2018 RDP phone system restored
RDP phone service is now back online.

Antibiotic resistances

gene—*contributor*
aac2i—*Syed Hashsham*
aac2ia—*Syed Hashsham*
aac2ib—*Syed Hashsham*
aac2ic—*Syed Hashsham*
aac2id—*Syed Hashsham*
aac3ia—*Syed Hashsham*

ARGs-Resfams

gene—*contributor*
Resfam_16S_Ribosomal_RNA_M
—*Resfam*
Resfam_AAC3—*Resfam*
Resfam_AAC3-Ia—*Resfam*
Resfam_AAC6-Ia—*Resfam*

Biodegradation

gene—*contributor*
alkb—
Gerben Zylstra/Elyse Rodgers-Vieira
baiCD—*RDP Staff*
benA—*Stephan Gantner*
bph—*Gerben Zylstra*
bphA1—*Stephan Gantner*

Biogeochemical cycles

gene—*contributor*
acs—*Russell Fling*
add1—*Yang Ouyang*
add2—*Yang Ouyang*
ALP—*Alan Richardson*
alp_new—*Jiangbin Xu*
amiE—*Yang Ouyang*

Metal Cycling

gene—*contributor*
ACR3—*Taylor Dunivin*
aioA—*Taylor Dunivin*
arrA—*Taylor Dunivin*
arsA—*PFAM*
arsB—*Taylor Dunivin*
arsB_old—*PFAM*

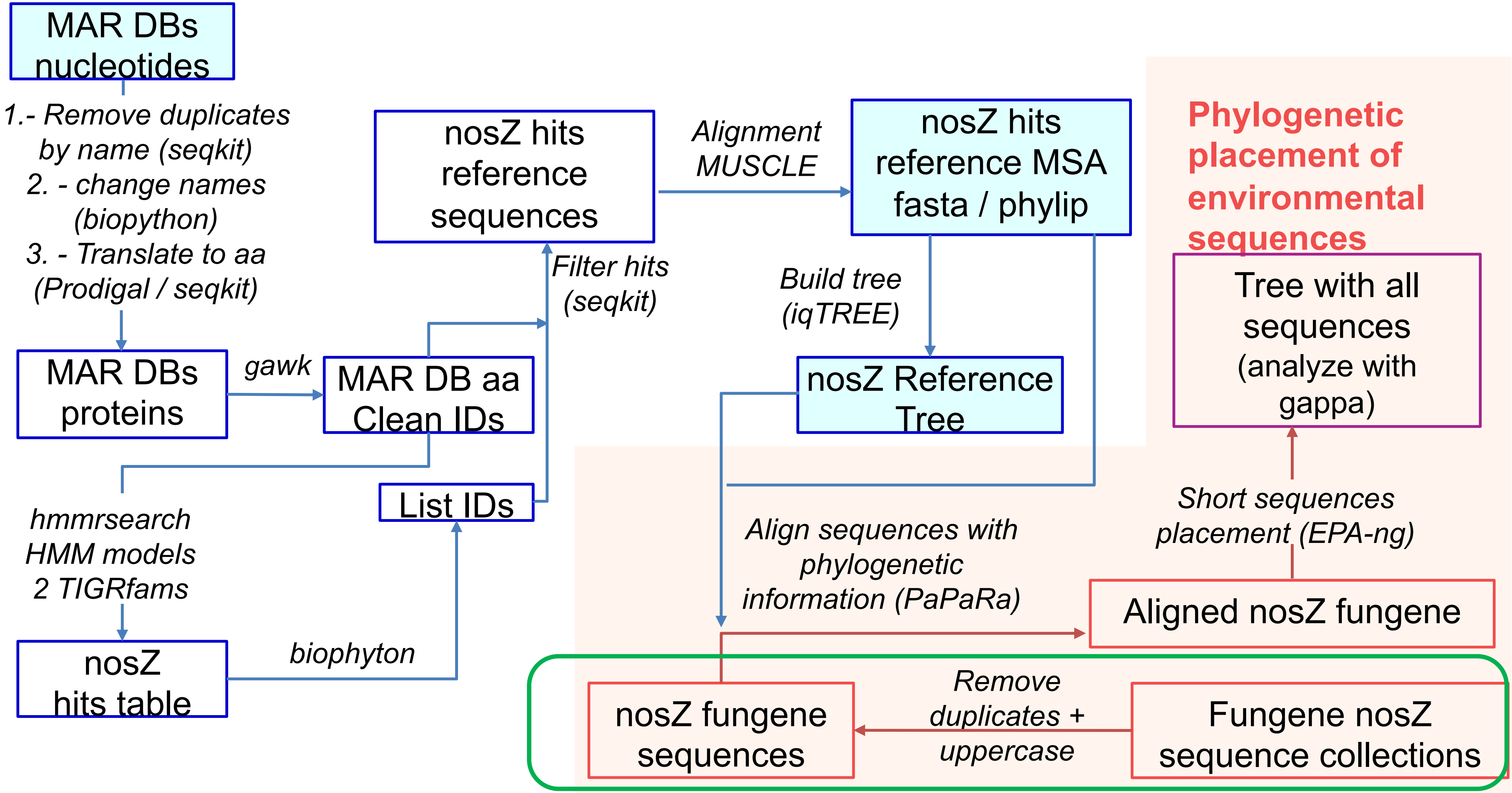
Other

gene—*contributor*
acdS—*RDP*
acsF—*Kasia Piwosz*
cagA—*Syed Hashsham*
Cas1_CasY—*Ruonan Wu*
Cas1_IB—*Ruonan Wu*
Cas1_ICID—*Ruonan Wu*

Phylogenetic markers

Plant Pathogenicity

Overview



MAR DBs nucleotides

1.- Remove duplicates by name (seqkit)
2. - change names (biopython)
3. - Translate to aa (Prodigal / seqkit)

MAR DBs proteins

gawk

MAR DB aa Clean IDs

hmmrsearch HMM models 2 TIGRfams

nosZ hits table

biophyton

List IDs

Filter hits (seqkit)

nosZ hits reference sequences

Alignment MUSCLE

nosZ hits reference MSA fasta / phylip

Build tree (iqTREE)

nosZ Reference Tree

Phylogenetic placement of environmental sequences

Fungene nosZ sequence collections

Remove duplicates + uppercase

nosZ fungene sequences

Align sequences with phylogenetic information (PaPaRa)

Aligned nosZ fungene

Short sequences placement (EPA-ng)

Tree with all sequences (analyze with gappa)

Mock environmental nosZ sequences

I have downloaded the three nosZ datasets available, put them together and removed duplicates:

```
$ cat fungene_nosZ.faa | seqkit rmdup -n -o fungene_nosZ_noDup.faa -d
duplicated_fungene_nosZ.faa -D duplicated_fungene_nosZ_detail.txt
[INFO] 12926 duplicated records removed
```

Avoid long names. Remove everything but the nucleotide GI that is the first field

```
$ gawk 'BEGIN{FS=" "}{if ($1 ~ /^>/) print $1; else print $0 }' fungene_nosZ_noDup.faa >
fungeneTest.faa
$ wc -l fungeneTest.faa
89856 fungeneTest.faa
```

Prepare a small subsample of Fungene sequences.
Notice that fungene sequences are lowercase, but mardbs are in uppercase. Fix it with seqkit

```
$ head -n 55 fungeneTest.faa | seqkit seq -u > query.faa
```

Phylogenetic placement of environmental sequences

1. – MSA of environmental sequences against reference database (PaPaRa)

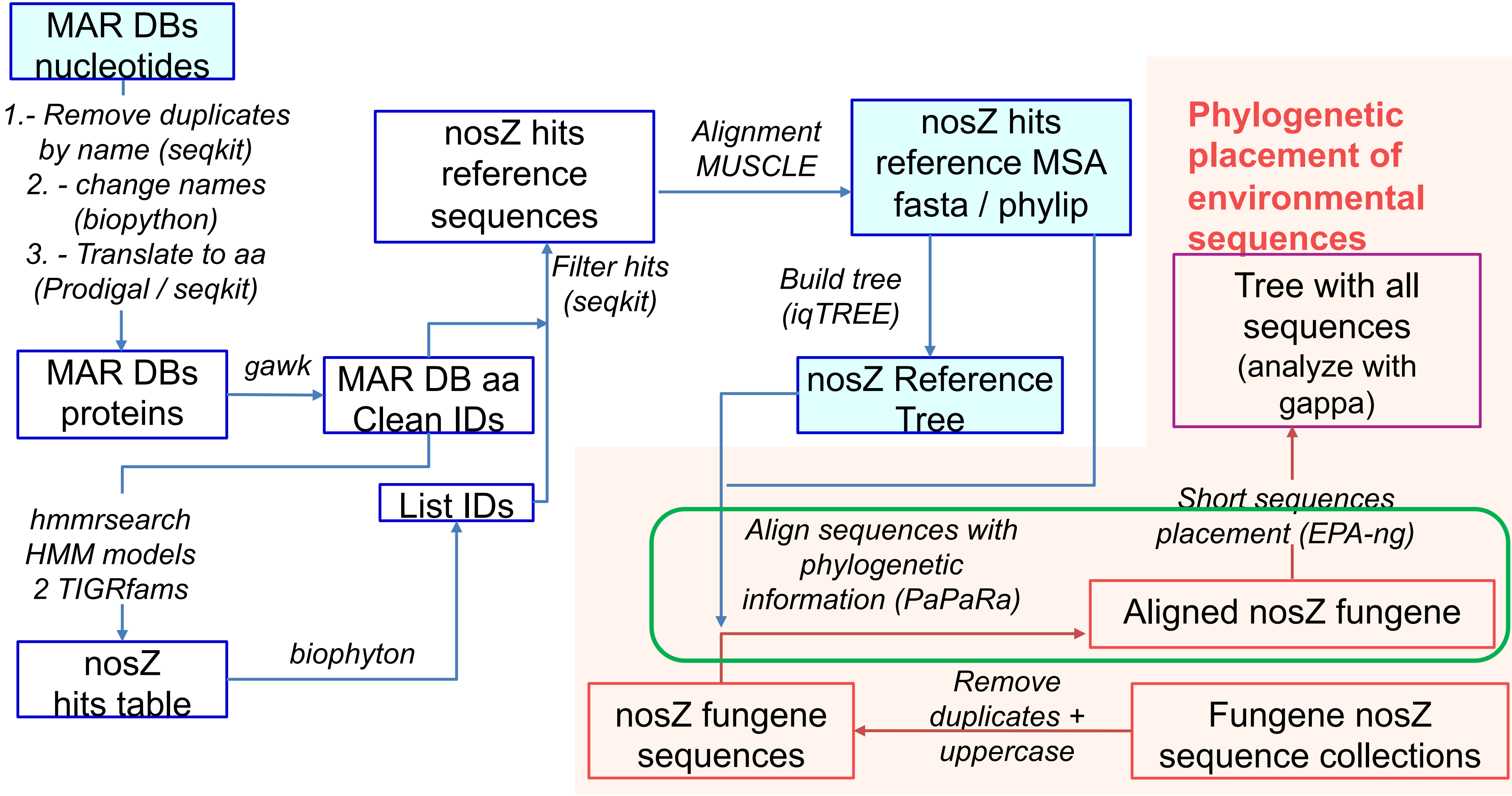
   PaPaRa requires:

   - aligned reference database in phylip format
   - reference tree in newick format (default output of iqTREE)

2. - Phylogenetic placement in the reference tree (EPA-ng)

3. - Analysis of results (gappa)

Overview



MAR DBs nucleotides

1.- Remove duplicates by name (seqkit)
2. - change names (biopython)
3. - Translate to aa (Prodigal / seqkit)

MAR DBs proteins

gawk

MAR DB aa Clean IDs

hmmrsearch HMM models 2 TIGRfams

nosZ hits table

biophyton

List IDs

Filter hits (seqkit)

nosZ hits reference sequences

Alignment MUSCLE

nosZ hits reference MSA fasta / phylip

Build tree (iqTREE)

nosZ Reference Tree

Phylogenetic placement of environmental sequences

Fungene nosZ sequence collections

Remove duplicates + uppercase

nosZ fungene sequences

Align sequences with phylogenetic information (PaPaRa)

Aligned nosZ fungene

Short sequences placement (EPA-ng)

Tree with all sequences (analyze with gappa)

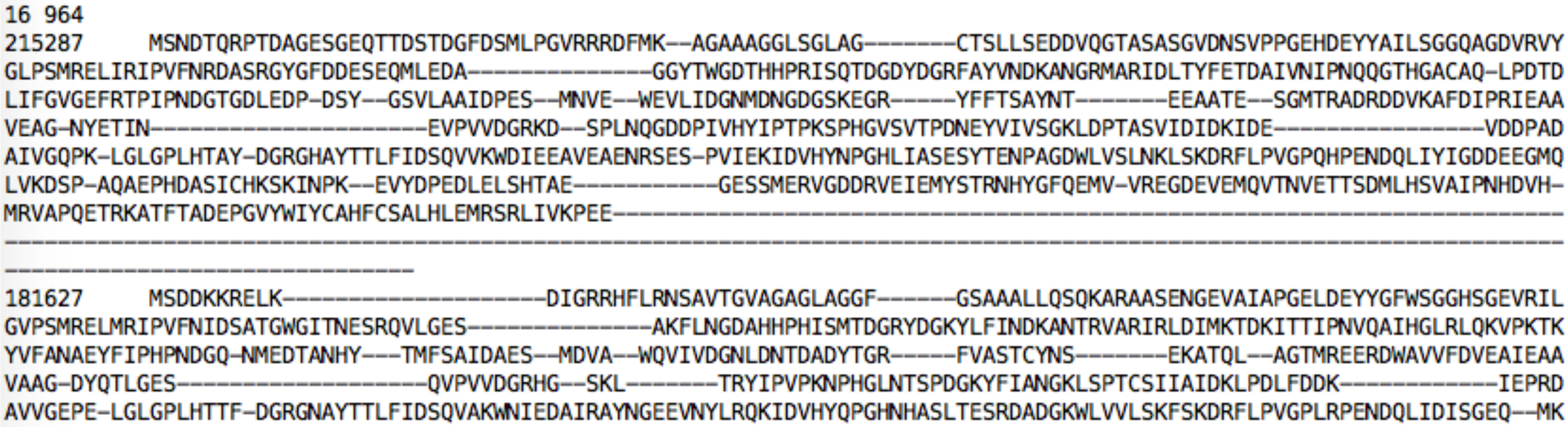
1 – Multiple Sequence Alignment with PaPaRA

PaPaRa options:
-r no additional gaps are included in the original MSA
-a sequences are aminoacid
-t reference tree,
-s reference alignment
-q sequences to align (query sequences),
-j threads
n sufix for output

```
$ cd alignments/
$ papara -t hits_nosZ_aligned.phy.treefile -a -r -s hits_nosZ_aligned.phy -q
../query.faa -j 4 -n papout 1>log_papara.txt 2>&1 &
```

1 – Multiple Sequence Alignment with PaPaRA

PaPaRa output, in phylip format:



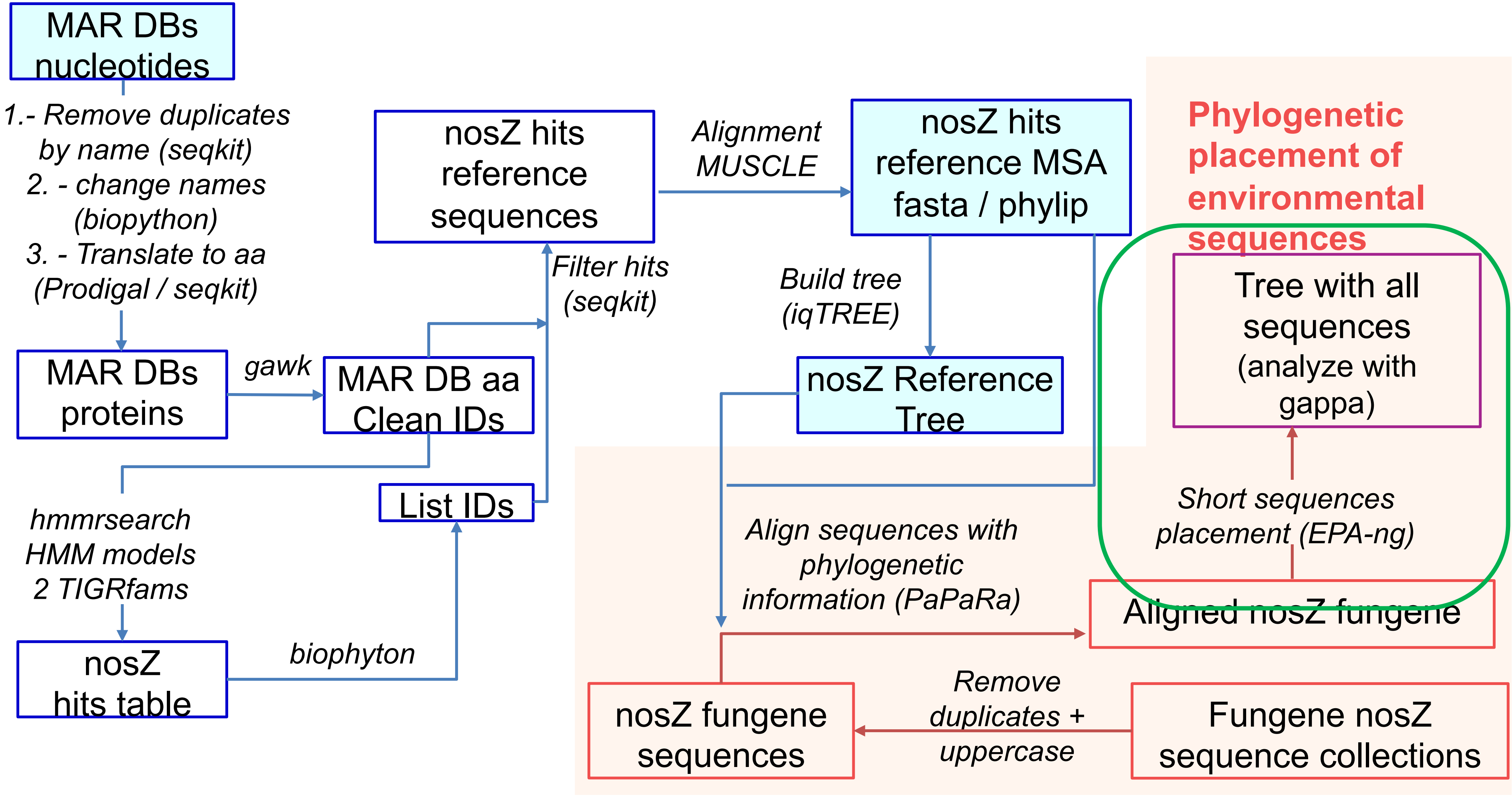
```
16 964
215287     MSNDTQRPTDAGESGEQTTDSTDGFDSMLPGVRRRDFMK--AGAAAGGLSGLAG-------CTSLLSEDDVQGTASASGVDNSVPPGEHDEYYAILSGGQAGDVRVY
GLPSMRELIRIPVFNRDASRGYGFDDESEQMLEDA--------------GGYTWGDTHHPRISQTDGDYDGRFAYVNDKANGRMARIDLTYFETDAIVNIPNQQGTHGACAQ-LPDTD
LIFGVGEFRTPIPNDGTGDLEDP-DSY--GSVLAAIDPES--MNVE--WEVLIDGNMDNGDGSKEGR-----YFFTSAYNT-------EEAATE--SGMTRADRDDVKAFDIPRIEAA
VEAG-NYETIN---------------------EVPVVDGRKD--SPLNQGDDPIVHYIPTPKSPHGVSVTPDNEYVIVSGKLDPTASVIDIDKIDE----------------VDDPAD
AIVGQPK-LGLGPLHTAY-DGRGHAYTTLFIDSQVVKWDIEEAVEAENRSES-PVIEKIDVHYNPGHLIASESYTENPAGDWLVSLNKLSKDRFLPVGPQHPENDQLIYIGDDEEGMQ
LVKDSP-AQAEPHDASICHKSKINPK--EVYDPEDLELSHTAE-----------GESSMERVGDDRVEIEMYSTRNHYGFQEMV-VREGDEVEMQVTNVETTSDMLHSVAIPNHDVH-
MRVAPQETRKATFTADEPGVYWIYCAHFCSALHLEMRSRLIVKPEE------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------------------------
-------------------------------

181627     MSDDKKRELK--------------------DIGRRHFLRNSAVTGVAGAGLAGGF------GSAAALLQSQKARAASENGEVAIAPGELDEYYGFWSGGHSGEVRIL
GVPSMRELMRIPVFNIDSATGWGITNESRQVLGES--------------AKFLNGDAHHPHISMTDGRYDGKYLFINDKANTRVARIRLDIMKTDKITTIPNVQAIHGLRLQKVPKTK
YVFANAEYFIPHPNDGQ-NMEDTANHY---TMFSAIDAES--MDVA--WQVIVDGNLDNTDADYTGR-----FVASTCYNS-------EKATQL--AGTMREERDWAVVFDVEAIEAA
VAAG-DYQTLGES-------------------QVPVVDGRHG--SKL-------TRYIPVPKNPHGLNTSPDGKYFIANGKLSPTCSIIAIDKLPDLFDDK------------IEPRD
AVVGEPE-LGLGPLHTTF-DGRGNAYTTLFIDSQVAKWNIEDAIRAYNGEEVNYLRQKIDVHYQPGHNHASLTESRDADGKWLVVLSKFSKDRFLPVGPLRPENDQLIDISGEQ--MK
```

Reference and query sequences are together in this output, separate them for EPA-ng

```
$ epa-ng --split hits_nosZ_aligned.phy papara_alignment.papout
                 (reference alignment)
```

It produces two files: `reference.fasta + query.fasta` → to EPA-ng

Overview



MAR DBs nucleotides

1.- Remove duplicates by name (seqkit)
2. - change names (biopython)
3. - Translate to aa (Prodigal / seqkit)

MAR DBs proteins

gawk

MAR DB aa Clean IDs

hmmrsearch HMM models 2 TIGRfams

nosZ hits table

biophyton

List IDs

Filter hits (seqkit)

nosZ hits reference sequences

Alignment MUSCLE

nosZ hits reference MSA fasta / phylip

Build tree (iqTREE)

nosZ Reference Tree

Phylogenetic placement of environmental sequences

Fungene nosZ sequence collections

Remove duplicates + uppercase

nosZ fungene sequences

Align sequences with phylogenetic information (PaPaRa)

Aligned nosZ fungene

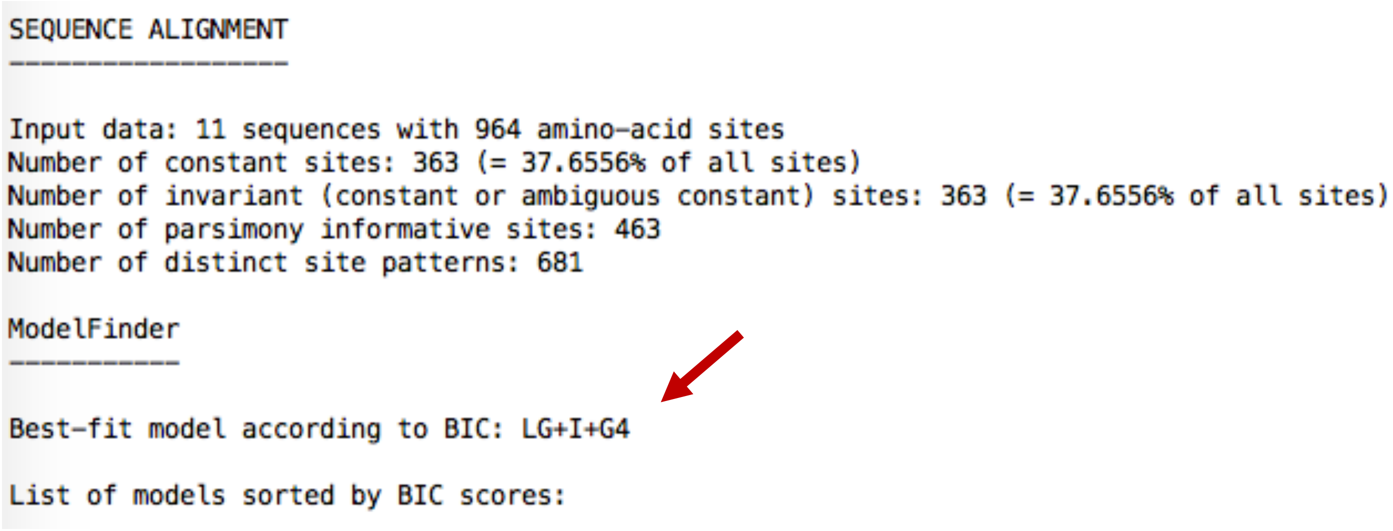
Short sequences placement (EPA-ng)

Tree with all sequences (analyze with gappa)

2. - Phylogenetic placement in the reference tree (EPA-ng)

Need to know the replacement model used in the tree

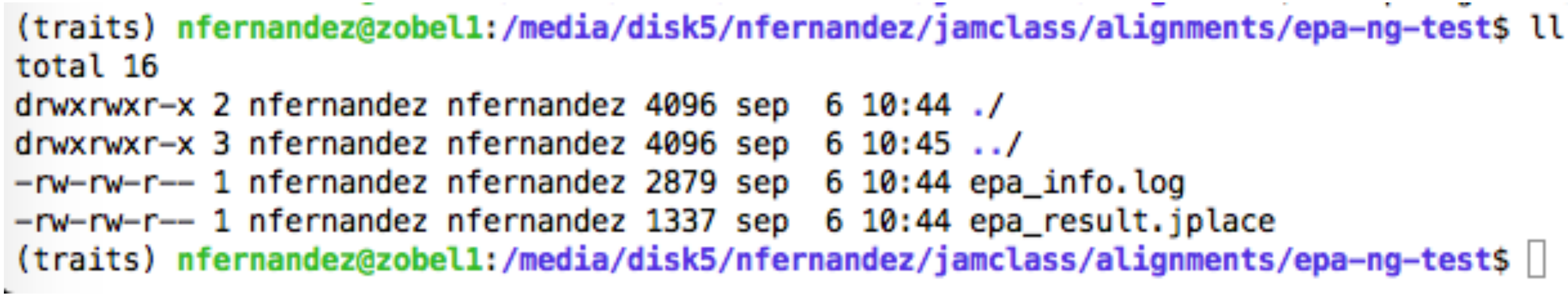
```
$ less hits_nosZ_aligned.phy.iqtree
```



```
SEQUENCE ALIGNMENT
------------------

Input data: 11 sequences with 964 amino-acid sites
Number of constant sites: 363 (= 37.6556% of all sites)
Number of invariant (constant or ambiguous constant) sites: 363 (= 37.6556% of all sites)
Number of parsimony informative sites: 463
Number of distinct site patterns: 681

ModelFinder
-----------

Best-fit model according to BIC: LG+I+G4

List of models sorted by BIC scores:
```

```
$ mkdir epa-ng-test
$ epa-ng --tree hits_nosZ_aligned.phy.treefile --ref-msa reference.fasta --query
query.fasta --out-dir epa-ng-test --model LG+F+R10 > log_epang.txt 2>&1 &
```

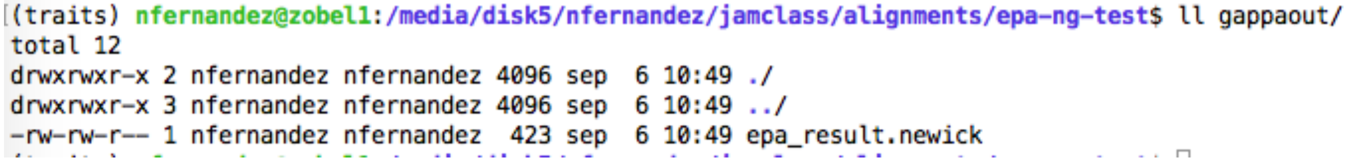
epa-ng options:
--tree: reference tree
--ref-msa: reference multiple sequence alignment, used for tree
--query: sequences to place in the tree, must be aligned against –ref-msa
--out-dir: output directory
--model: substitution model used to build the tree



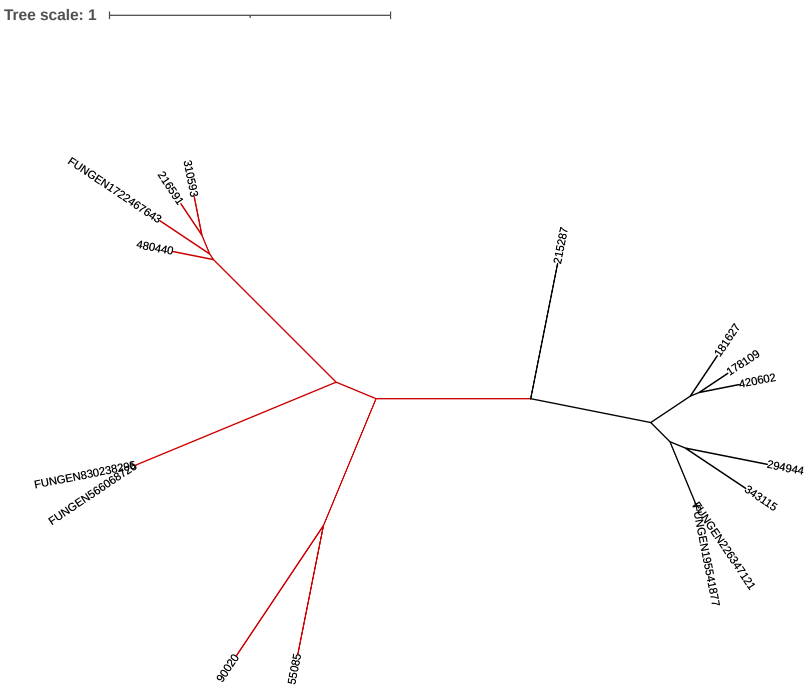
```
(traits) nfernandez@zobel1:/media/disk5/nfernandez/jamclass/alignments/epa-ng-test$ ll
total 16
drwxrwxr-x 2 nfernandez nfernandez 4096 sep  6 10:44 ./
drwxrwxr-x 3 nfernandez nfernandez 4096 sep  6 10:45 ../
-rw-rw-r-- 1 nfernandez nfernandez 2879 sep  6 10:44 epa_info.log
-rw-rw-r-- 1 nfernandez nfernandez 1337 sep  6 10:44 epa_result.jplace
(traits) nfernandez@zobel1:/media/disk5/nfernandez/jamclass/alignments/epa-ng-test$ 
```

3. - Analysis of results (gappa)

```
$ gappa examine graft --jplace-path epa_result.jplace --name-prefix FUNGEN --out-dir gappaout
```



```
(traits) nfernandez@zobel1:/media/disk5/nfernandez/jamclass/alignments/epa-ng-test$ ll gappaout/
total 12
drwxrwxr-x 2 nfernandez nfernandez 4096 sep  6 10:49 ./
drwxrwxr-x 3 nfernandez nfernandez 4096 sep  6 10:49 ../
-rw-rw-r-- 1 nfernandez nfernandez  423 sep  6 10:49 epa_result.newick
```



Type I

Type II