0. SETTING UP ENVIRONMENT

```
#checking files quality on desktop version of FastQC Version
0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/
fastqc/)
SoilFenceLib-UNIS_S1_L001_R2_001.fastq.gz
SoilFenceLib-UNIS_S1_L001_R1_001.fastq.gz

#setting up Abel interactive session
qlogin --account=nn9320k --ntasks-per-node=16

pwd
    /usit/abel/u1/magdalenaw/fences

ls
    -rw-r--r--  1 magdalenaw users 1.5G Jan 28  2016
SoilFenceLib-UNIS_S1_L001_R2_001.fastq.gz
    -rw-r--r--  1 magdalenaw users 1.2G Jan 28  2016
SoilFenceLib-UNIS_S1_L001_R1_001.fastq.gz

#unzipping files, perlscript does not work with zipped files
gunzip SoilFenceLib-UNIS_S1_L001_R1_001.fastq.gz
gunzip SoilFenceLib-UNIS_S1_L001_R2_001.fastq.gz

#creating symbolic link to fastq files
ln -s SoilFenceLib-UNIS_S1_L001_R1_001.fastq forward.fastq
ln -s SoilFenceLib-UNIS_S1_L001_R2_001.fastq reverse.fastq



#in total: 8413098 PE reads
```

1. QUALITY FILTERING

```
#loading perlmodules/5.10_2
module load perlmodules

#running Reads_Quality_Length_distribution.pl (supplemented in
Bálint et al., 2014)
perl Reads_Quality_Length_distribution.pl -fw forward.fastq -
rw reverse.fastq -sc 33 -q 26 -l 150 -ld Y


    FILES CREATED IN THIS ANALYSIS
    -rw-r--r--  1 magdalenaw users 4.8G Jun  5 15:31
Filtered_reads_without_Ns_quality_threshold_26_length_threshol
d_150_R1.fastq
```

```
    -rw-r--r--  1 magdalenaw users 4.8G Jun  5 15:31
Filtered_reads_without_Ns_quality_threshold_26_length_threshol
d_150_R2.fastq
    -rw-r--r--  1 magdalenaw users  652 Jun  5 15:31
Reads_quality_and_reads_having_Ns_summary.txt
    -rw-r--r--  1 magdalenaw users    0 Jun  5 14:47
Reads_Average_quality_distribution_table.tab
    -rw-r--r--  1 magdalenaw users 937M Jun  5 14:47
Reads_length_distribution_table.tab
    -rw-r--r--  1 magdalenaw users  476 Jun  5 14:47
Reads_length_summary.txt


FastQC reported 7779879 PE reads (~92% of sequences)
```

---

2. Paired - end assembly

```
#!/bin/sh
#SBATCH --job-name=2step
#SBATCH --account=nn9320k
#SBATCH --output=slurm-%j.base
#SBATCH --time=02:00:00
#SBATCH --mem-per-cpu=12G
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --mail-type=ALL
#SBATCH --mail-user=magdalena.wutkowska@unis.no

module purge
set -o errexit
module load pandaseq/2.11

pandaseq -f /usit/abel/u1/magdalenaw/fences/
Filtered_reads_without_Ns_quality_threshold_26_length_threshol
d_150_R1.fastq -r /usit/abel/u1/magdalenaw/fences/
Filtered_reads_without_Ns_quality_threshold_26_length_threshol
d_150_R2.fastq -F N -o 5 > /usit/abel/u1/magdalenaw/fences/
paired_assembled.fastq


grep '@M01610' paired_assembled.fastq | wc -l
7 450 729 (~88% of raw reads)
```

---

3. Removing sequences with primer artifacts

```
module load python2
```

```
python remove_multiprimer.py -i paired_assembled.fastq -o
paired_assembled_no_primer_artifacts.fastq -f
"TCCTCCGCTTATTGATATGC" -r "GTGAATCATCGAATCTTTG"
#remove_multiprimer.py comes from Balint et al., 2014


grep '@M01610' paired_assembled_no_primer_artifacts.fastq | wc
-l
7 028 992 (~84% of raw reads)

gzip paired_assembled_no_primer_artifacts.fastq
```

---

4. Reorienting the reads to 5’-3’

```
module load fqgrep 0.4.4
module load fastx-toolkit/0.0.14

#primers DNA: ACTCCTCCGCTTATTGATATGC, ACGTGAATCATCGAATCTTTG
#primers RNA: TGTCCTCCGCTTATTGATATGC, TGGTGAATCATCGAATCTTTG

fqgrep -m 1 -p ACTCCTCCGCTTATTGATATGC
paired_assembled_no_primer_artifacts.fastq.gz -o
good_5-3_DNA.fastq.gz
[zgrep '@M01610' good_5-3_DNA.fastq.gz | wc -l #1658547]

fqgrep -m 1 -p TGTCCTCCGCTTATTGATATGC
paired_assembled_no_primer_artifacts.fastq.gz -o
good_5-3_RNA.fastq.gz
[zgrep '@M01610' good_5-3_RNA.fastq.gz | wc -l #1773449]

     cat good_5-3_DNA.fastq.gz good_5-3_RNA.fastq.gz >
good_forward.fastq.gz
     [zgrep '@M01610' good_forward.fastq.gz | wc -l #3431996]

fqgrep -m 1 -p ACGTGAATCATCGAATCTTTG
paired_assembled_no_primer_artifacts.fastq.gz -o
good_3-5_DNA.fastq.gz
[zgrep '@M01610' good_3-5_DNA.fastq.gz | wc -l #1748976]

fqgrep -m 1 -p TGGTGAATCATCGAATCTTTG
paired_assembled_no_primer_artifacts.fastq.gz -o
good_3-5_RNA.fastq.gz
[grep '@M01610' good_3-5_RNA.fastq.gz | wc -l #1811564]

cat good_3-5_DNA.fastq.gz good_3-5_RNA.fastq.gz >
good_reverse.fastq.gz
[zgrep '@M01610' good_reverse.fastq.gz | wc -l #3560540]
```

```
module load fastx-toolkit/0.0.14
fastx_reverse_complement -Q 33 -i good_reverse.fastq.gz >>
good_forward.fastq.gz
zgrep '@M01610' good_forward.fastq.gz  | wc -l #6992536 (~83%
of raw reads)
```

---

5. Demultiplexing the dataset based on barcodes, because our barcodes are of variable lengths we could not use the demultiplexing script suggested by Balint et al., 2014, instead we use split_libraries.py (part of qiime Caporaso et al., 2010)

```
#first fastq -> fna
module load qiime/1.9.1
convert_fastaqual_fastq.py -c fastq_to_fastaqual -f
good_forward.fastq
     [just to check if things are going good:
     grep 'M01610' good_forward.fna  | wc -l #6992536]

#splitting/demultiplexing libraries and removing of barcodes/
primers
('truncate_only' option will remove the primer and subsequent
sequence data from the output read and will not alter output
of sequences where the primer cannot be found.)

module purge
module load python2/2.7.10
module load qiime/1.9.1

split_libraries.py -f good_forward.fna -m
forward_T_split_map.txt -o demultiplex_T_M1/ -M 1 -H 8 -l 200
-L 500 -b variable_length -z truncate_only

#this produced an error: 'ValueError: Duplicate ID found in
FASTA/qual file:  M01610:128:000000000-ADYRP:
1:1101:21878:1417:1', thus, removing the sequence from fna
file

removing this duplicated seq ID according to oneliner found in
http://www.filiphusnik.com/content/bioinformatics-one-liners
awk '/^>/{f=!d[$1];d[$1]=1}f' good_forward.fna >
readyfordemultiplex.fna

split_libraries.py -f readyfordemultiplex.fna -m
forward_T_split_map.txt -o demultiplex_T_M1/ -M 1 -H 8 -l 200
-L 500 -b variable_length -z truncate_only

5184214
```

---

6. Sort by length

```
module load vsearch/2.7.1

vsearch -sortbylength /cluster/home/magdalenaw/fences/seqs.fna
-output /cluster/home/magdalenaw/fences/Vsearch_200-500.fasta
--minseqlength 200 --maxseqlength 500

vsearch v2.7.1_linux_x86_64, 62.9GB RAM, 32 cores
https://github.com/torognes/vsearch

Reading file /cluster/home/magdalenaw/fences/seqs.fna 100%
1570361620 nt in 5184213 seqs, min 212, max 471, avg 303
minseqlength 200: 1 sequence discarded.
Getting lengths 100%
Sorting 100%
Median length: 302
Writing output 100%

5184213 seq
```

---

7. Dereplicating/grouping of replicate sequences

```
module load vsearch/2.7.1

vsearch -derep_fulllength /cluster/home/magdalenaw/fences/
Vsearch_200-500.fasta -output derep.fasta --sizeout
vsearch v2.7.1_linux_x86_64, 62.9GB RAM, 32 cores
https://github.com/torognes/vsearch

Reading file /cluster/home/magdalenaw/fences/
Vsearch_200-500.fasta 100%
1570361620 nt in 5184213 seqs, min 212, max 471, avg 303
Dereplicating 100%
Sorting 100%
1017958 unique sequences, avg cluster 5.1, median 1, max
181765
Writing output file 100%

grep '>' derep.fasta | wc -l
1017958
```

---

8. Sorting by size of groups and removing these who have <n reads

---

```
module load vsearch/2.7.1

vsearch -sortbysize /cluster/home/magdalenaw/fences/
derep.fasta -output /cluster/home/magdalenaw/fences/
sorted_minsize5.fasta -minsize 5

grep '>' sorted_minsize5.fasta | wc -l
66372
```

---

9. Picking OTUs

```
module load usearch/9.2.64

usearch -cluster_otus /cluster/home/magdalenaw/fences/
sorted_minsize5.fasta -otus otus_minsize5.fasta -
otu_radius_pct 0.97
usearch v9.2.64_i86linux32, 4.0Gb RAM (65.9Gb total), 32 cores
100.0% 2185 OTUs, 425 chimeras
```

---

10. Reference based chimera check (UNITE_public_01.12.2017.fasta, https://doi.org/10.15156/BIO/587474)
R. C. Edgar (2016), UCHIME2: Improved chimera detection for amplicon sequences, http://dx.doi.org/10.1101/074252

```
module load usearch/8.1.1861

#First step: converting databased into suitable format:
usearch -makeudb_usearch /cluster/home/magdalenaw/fences/
UNITE_public_01.12.2017.fasta -output /cluster/home/
magdalenaw/fences/UNITE_public_01.12.2017.udb

#Second step: match this with representative sequences:

module load usearch/8.1.1861

usearch -uchime2_ref /cluster/home/magdalenaw/fences/
otus_minsize5.fasta -db /cluster/home/magdalenaw/fences/
UNITE_public_01.12.2017.udb -notmatched /cluster/home/
magdalenaw/fences/otus_minsize5_good.fasta -chimeras /cluster/
home/magdalenaw/fences/otus_minsize5_chimeras.fasta -strand
plus  -mode balanced

#the last option  -mode balanced Attempts to balance false
```

negatives and false positives to minimize the overall error rate on typical data. Of course, the rates are highly data-dependent.

100.0% Chimeras 232/2185 (10.6%), in db 21 (1.0%), not matched 1932 (88.4%)

---

11. Removing non-fungal ITS2 by ITSx | ITSx v. 1.1b (Bengtsson-Palme et al., 2013)

```
module load perlmodules/5.10_2
module load hmmer/3.1b2

export PATH=$PATH:$HOME/ITS/ITS_1.1b1/

ITSx -i /cluster/home/magdalenaw/fences/
otus_minsize5_good.fasta -t F -summary T -save_regions ITS2 -
preserve T -partial 100 -minlen 200

Output:
Number of sequences in input file:              1953
Sequences detected as ITS by ITSx:      1473
  On main strand:               0
  On complementary strand:      1473
Sequences detected as chimeric by ITSx: 0
ITS sequences by preliminary origin:
  Alveolates:           0
  Amoebozoa:            0
  Bacillariophyta:      0
  Brown algae:          0
  Bryophytes:           0
  Euglenozoa:           0
  Eustigmatophytes:     0
  Fungi:                1473
  Green algae:          0
  Liverworts:           0
  Metazoa:              0
  Microsporidia:        0
  Oomycetes:            0
  Prymnesiophytes:      0
  Raphidophytes:        0
  Red algae:            0
  Rhizaria:             0
  Synurophyceae:        0
  Tracheophyta:         0
```

12. idenfitying fungal OTUs

```
#!/bin/sh
#SBATCH --job-name=bl_xml
#SBATCH --account=nn9320k
#SBATCH --output=slurm-%j.base
#SBATCH --time=02-00
#SBATCH --mem-per-cpu=60G
#SBATCH --ntasks=1
#SBATCH --mail-type=ALL
#SBATCH --mail-user=magdalena.wutkowska@unis.no

set -o errexit
module purge
module load blast+/2.6.0

blastn -num_threads $OMP_NUM_THREADS -db /work/databases/bio/
NCBI/blast/13DEC2017/nt \
-query /cluster/home/magdalenaw/fences/ITSx_out.ITS2.fasta \
-outfmt 5 -out /cluster/home/magdalenaw/fences/blastn.xml -
evalue 0.001
```

then blast xml was opened in megan: (excerpt from Balint)
‘The xml, and the blasted fasta file, is imported into ME-
GAN. The lowest common ancestor assignments depend on several
options, our choices for Illumina paired-end reads are minimum
reads 1, minimum score 170, upper percentage 5, no minimum
complexity, no min complex- ity (0). We uncollapse all
branches, select Fungi, and from the Select menu select
Subtree. Reads should be exported from the File menu (File/
Export/Reads).
Output: xml file containing the BLAST hits of the OTU cen-
troid sequences. rma file containing the parsed BLAST results.
fasta file containing the exported fungal OTU sequences’

sequences are in meganfiltered.fasta #845 OTUs

13. Fungal OTU abundance table

```
#renaming headers in representative sequences (output from
megan):
bash /usit/abel/u1/magdalenaw/apps/bbmap/rename.sh
in=meganfiltered.fasta out=meganfiltered_header.fasta
prefix=OTU

cp seqs.fna seqs.fasta
sed 's/bc_diffs=0//g' seqs.fasta > seqs_mod.fasta
```

```
#removing M01610:128:000000000-ADYRP:1:1101:8264:1415:1 - part
of the header
sed 's/M01610[^ ]* //g' seqs_mod.fasta > seqs_mod2.fasta

removing new_bc=…
sed 's/new_bc[^ ]* //g' seqs_mod2.fasta > seqs_mod3.fasta

renaming barcode label
sed 's/orig_bc=/barcodelabel=/g' seqs_mod3.fasta >
seqs_formapping.fasta

sed 's/barcodelabel=[^ ]*//g' seqs_formapping.fasta >
seqs_formap2.fasta

sed -e 's/_1_/_1\./g' seqs_formap2.fasta >
seqs_halfready.fasta
sed -e 's/_2_/_2\./g' seqs_halfready.fasta >
seqs_halfready2.fasta

sed -e 's/1_C1_/1_C1\./g' seqs_halfready2.fasta >
seqs_temp1.fasta
sed -e 's/2_C1_/2_C1\./g' seqs_temp1.fasta > seqs_temp2.fasta
sed -e 's/3_C1_/3_C1\./g' seqs_temp2.fasta > seqs_temp3.fasta

sed -e 's/1_C2_/1_C2\./g' seqs_temp3.fasta > seqs_temp4.fasta
sed -e 's/2_C2_/2_C2\./g' seqs_temp4.fasta > seqs_temp5.fasta
sed -e 's/3_C2_/3_C2\./g' seqs_temp5.fasta > seqs_temp6.fasta

sed -e 's/_//g' seqs_temp6.fasta > seqs_forotutable.fasta

/usit/abel/u1/magdalenaw/apps/usearch/
usearch8.1.1861_i86linux32 -usearch_global
seqs_forotutable.fasta -db meganfiltered_header.fasta -strand
both -id 0.98 -otutabout otu_map.txt -biomout otu_map.biom
```

---

14. Rarefaction (number 42488 comes from demultiplex results)

```
single_rarefaction.py -i otu_map.biom -o otu_map_RARIF.biom -d
42488 #this step will remove all the samples that had less
that -d reads

biom convert -i otu_map_RARIF.biom -o otu_table_rarified.txt
--to-tsv

retaining OTUs that were kept in rarefied OTUs
filter_fasta.py -f meganfiltered_header.fasta -o le -s
otus_inrarified_table.txt
```

---

15. Assign taxonomy based on UNITE: https://plutof.ut.ee/#/datacite/10.15156%2FBIO%2F587475

```
makeblastdb -in sh_general_release_dynamic_s_01.12.2017.fasta
-parse_seqids -dbtype nucl


#!/bin/sh
#SBATCH --job-name=bl_xml
#SBATCH --account=nn9320k
#SBATCH --output=slurm-%j.base
#SBATCH --time=02-00
#SBATCH --mem-per-cpu=60G
#SBATCH --ntasks=1
#SBATCH --mail-type=ALL
#SBATCH --mail-user=magdalena.wutkowska@unis.no

set -o errexit
module purge
module load blast+

blastn -num_threads $OMP_NUM_THREADS -db
sh_general_release_dynamic_s_01.12.2017.fasta \
-query /cluster/home/magdalenaw/fences/
meganfiltered_rarifiedOTU.fasta \
-outfmt 6 -out /cluster/home/magdalenaw/fences/
blastn_unite.txt -evalue 0.00001 -max_target_seqs 1
```

---

16. Assign taxonomy from NCBI database (to use when there is no taxonomy assigned using step above (blast with UNITE database))

```
#!/bin/sh
#SBATCH --job-name=bl_xml
#SBATCH --account=nn9320k
#SBATCH --output=slurm-%j.base
#SBATCH --time=02-00
#SBATCH --mem-per-cpu=60G
#SBATCH --ntasks=1
#SBATCH --mail-type=ALL
#SBATCH --mail-user=magdalena.wutkowska@unis.no
```

```
set -o errexit
module purge
module load blast+

blastn -num_threads $OMP_NUM_THREADS -db /work/databases/bio/
NCBI/blast/13DEC2017/nt \
-query /cluster/home/magdalenaw/fences/
meganfiltered_rarifiedOTU.fasta \
-outfmt 6 -out /cluster/home/magdalenaw/fences/blastn.txt -
evalue 0.00001 -max_target_seqs 1
```