

## Introduction à la bioanalyse sur jupyter

```
#introduction au projet
# on commence avec la fonction pwd pour se situer dans les fichiers
pwd
#on veut travailler dans le dossier "nobackup" car il est partagé et
avec une grande capacité de stockage
cd /nesi/nobackup/uoo03555
pwd # pour afficher dans quel dossier on est et pour verifier si la
commande cd a fonctionnée
# on va lister les fichiers qui se trouvent à l'intérieur du projet
uoo03555
ls
# on va créer un nouveau fichier à mon nom
mkdir celia
ls # pour vérifier que le fichier a bien été crée
cd celia # pour travailler dans le fichier à mon nom
pwd # pour voir si on est bien dans le fichier a mon nom
# on veut copier tout ce que je fais dans le dossier Celia avec la
commande cp
cp -r /nesi/project/uoo03555/Celia .
# on veut travailler dans le dossier à mon nom
cd /nesi/nobackup/uoo03555/celia/Celia
#commande ls pour vérifier que le projet a été copié dans le dossier à
mon nom
ls
# on obtient 0G8254-390638912
# on va demander à travailler dans le projet 0G8254-390638912
cd 0G8254-390638912/
# on veut voir ce qu'il y a déjà dans le projet
ls
# on obtient :
#230609_M01795_1092_000000000-L647R run quality report.pdf
0G8254_sample_info.xlsx
#FASTQ_Generation_2023-06-11_02_48_04Z-676096427
Project 0G8254 report_MultiQC Report.pdf
#0G8254_libQC_High Sensitivity DNA Assay_DE04103523_2023-05-30_15-58-
29.pdf summary_Illumina_run.pdf
# on veut voir combien de fichiers et dossiers sont contenus dans
FASTQ_Generation[...]
cd FASTQ_Generation_2023-06-11_02_48_04Z-676096427/
ls -l | wc
# on affiche ensuite le nom de tous les dossiers
ls# ce qui, ici, nous donne la liste de tous les fichiers zip des
sequences fatsq
# pour quitter la liste on tape sur "q" du clavier

#si on veut détailler ce qu'il y a dans un des fichiers on fait :
cd 8254-070-00-01_L001-ds.f999119d0e82441791897e29543ccaba/# cd le nom
```

```

du dossier
# ce qui, ici nous affiche les 2 fichiers zip fatsq contenu dans le
dossier
pwd #
/nesi/nobackup/uo003555/celia/Celia/0G8254-390638912/FASTQ_Generation_
2023-06-11_02_48_04Z-676096427/8254-070-00-01_L001-
ds.f999119d0e82441791897e29543ccaba
# si on veut enlever des morceaux du chemin d'accès on fait ;;
cd ../../ # 1x ../ enlève 1 chemin d'accès | 2x ../ (../../) enlève 2
chemins d'accès
pwd # on obtient /nesi/nobackup/uo003555/celia/Celia/0G8254-390638912

Error in parse(text = x, srcfile = src): <text>:10:7: unexpected
symbol
9: # on va créer un nouveau fichier à mon nom
10: mkdir celia
      ^
Traceback:

```

Unzipper les fichiers

```

# On veut unzipper tous les fichiers
# on doit créer un fichier de destination unzip
mkdir unzip
ls # pour vérifier qu'il est créé
cd FASTQ_Generation_2023-06-11_02_48_04Z-676096427/
ls
# on veut unzipper TOUS les fichiers
find . -type f -name '*.gz*' -exec gunzip "{}" \; # tout ce qui est
contenu entre *...* est recherché au sein de tous les fichiers
# copier tous les fichiers unzipés dans le dossier unzip
find . -type f -name '*.fastq*' -exec cp "{}" ../unzip \;

# on veut voir si tous les fichiers ont bien été copiés
cd ../unzip # change directory
ls # afficher
# ensuite on va programmer le rapport d'analyse
fastqc.sh # c'est mieux de mettre .sh car ca peut être lu par la
plupart des logiciels d'exploitation
#!/bin/bash -e # -e veut dire que si il détecte une erreur il va
s'arreter et signaler quelle est l'erreur
#SBATCH --job-name=fastqc
#SBATCH --output=fastqc_%j.out
#SBATCH --error=fastqc_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END
#SBATCH --time=5:00:00
#SBATCH --mem=1G

```

```

#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

module load FastQC/0.11.9 #ca c'est pour lancer le logiciel nommé
0.11.9

cd /nesi/nobackup/uoo03555/celia/Celia/OG8254-390638912/unzip

fastqc *.fastq*
less fastqc.sh # pour en sortir on tape sur "q" du clavier
sbatch fastqc.sh #permet de montrer le nom job du projet : Submitted
batch job 36611539

sacct # permet de donner un aperçu de comment se déroule le projet
sacct -u koece533 #permet de donner un aperçu de comment se déroule le
projet de koece533
sacct -j 36611539 #permet de donner un aperçu de comment se déroule le
projet 36611539

ls -l # va nous donner tous les fichiers zip et unzip du dossier en
ligne (-l) ; total 14550277
# à la fin on retrouve :
#fastqc_36611539.err
#fastqc_36611539.out
#fastqc.sh
# qu'on a crée avec SBATCH

```

importer les fichiers fastq dans qiime

```

#IMPORTER LES FICHIERS FASTQ DANS QIIME
# verifier si on a un package
module spider BLAST #pour les packages BLAST
module spider Qiime # pour les packages Qiime

# créer un fichier qiime
cd /nesi/nobackup/uoo03555/celia/Celia
mkdir qiime
# on copie tous les fichiers fastq dans le dossier qiime
cp -r /nesi/project/uoo03555/Celia/OG8254-
390638912/FASTQ_Generation_2023-06-11_02_48_04Z-676096427 .
# on crée le dossier "all"
mkdir all
# on copie tous les dossiers fastq du projet 8254 avec leurs double
cp 8254*/*.* ../all
# verifier ce qu'on a dans all
cd ../all
ls

--type 'SampleData[PairedEndSequencesWithQuality]' \

```

```

--input-path ./ \ #on est deja dans all donc pas besoin de le préciser

--input-format CasavaOneEightSingleLanePerSampleDirFmt \
--output-path demux-raw-82.qza # on change P7 pour le n° 82 - debut de
notre projet 8254
# on veut ensuite visualiser le qza
qiime demux summarize \
--i-data demux-raw-82.qza \
--o-visualization demux-raw-82.qzv
# enregistrer le fichier demux-raw-82.qzv normalement dans le dossier
qiime
# aller sur le site https://view.qiime2.org/
# importer le fichier demux-raw-82.qzv dans qiime view et on a la vue
sur tous les résultats
# on obtient ca :
forward reads      reverse reads
Minimum           1          1
Median           30174.0      30174.0
Mean      81473.051613      81473.051613
Maximum          1264144      1264144
Total 12628323      12628323

```

## DADA 2 and cutadapt

```

# Dada 2 et cutadapt
# tout ce qu'on veut faire dans qiime est en qza
# on crée le dossier raw_qza
mkdir raw_qza
#on copie donc le fichier qza dans le dossier crée raw_qza
cp demux-raw-82.qza raw_qza

# on crée dans nano notre dossier
nano metadata.qiime.sh

#!/bin/bash -e
#SBATCH --job-name=metadata
#SBATCH --output=metadata_%j.out
#SBATCH --error=metadata_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END
#SBATCH --time=24:00:00
#SBATCH --mem=10G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

module load QIIME2/2022.2
cd /nesi/nobackup/uoo03555/celia/Celia/qiime/raw_qza/

```

```

qiime dada2 denoise-paired \
  --i-demultiplexed-seqs demux-raw-82.qza \
  --p-trim-left-f 13 \
  --p-trim-left-r 13 \
  --p-trunc-len-f 240 \
  --p-trunc-len-r 240 \
  --o-table table-82.qza \
  --o-representative-sequences rep-seqs-82.qza \
  --o-denoising-stats denoising-stats-82.qza

qiime feature-table summarize \
  --i-table table-82.qza \
  --o-visualization table-82.qzv \
  --m-sample-metadata-file ../metadata_tsv.tsv

qiime feature-table tabulate-seqs \
  --i-data rep-seqs-82.qza \
  --o-visualization rep-seqs-82.qzv
#ctrl+X
# Y
# enter
sbatch metadata.qiime.sh
#36633325 job name
sacct # pour voir si tout est ok et a bien été complété

```

cutadapt

```

#cutadapt
# on veut aller dans le dossier unzip
cd /nesi/nobackup/uoo03555/celia/Celia/OG8254-390638912/unzip/
pwd
ls
# on veut voir si on peut trouver les adapteurs de séquences à l'aide
de la fonction grep
grep AATGATACGGCGACCACCGAGATCTACACGCT -c *.fastq # nos fichiers sont
terminés par .fastq

less cutadapt.sh # job 3665806
#!/bin/bash -e
#SBATCH --job-name=cutadapt
#SBATCH --output=cutadapt_%j.out
#SBATCH --error=cutadapt_%j.err
#SBATCH --mail-user=priscila.madisalloum@otago.ac.nz
#SBATCH --mail-type=END
#SBATCH --time=2:00:00
#SBATCH --mem=10G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

```

```

module load QIIME2/2022.2

cd /nesi/nobackup/uoo03555/celia/Celia/qiime

echo "Step 1"
echo "remove index from read 2, write log into cutadapt_step1.out and
make a qza file with the output"
qiime cutadapt trim-paired \
--i-demultiplexed-sequences demux-raw-82.qza \
--p-adapter-r AGCGTGTAGATCTCGGTGGTCGCCGTATCATT \
--p-error-rate 0 \
--p-minimum-length 240 \
--p-no-discard-untrimmed \
--o-trimmed-sequences step1-adapter-trimmed-demux-82.qza \
--verbose > cutadapt_step1.out

echo "generate visualization file with summary from the cutadapt step
1"
qiime demux summarize \
--i-data step1-adapter-trimmed-demux-82.qza \
--o-visualization step1-adapter-trimmed-demux-82.qzv

echo "Step 2"
echo " Remove the Reverse Complement (RC) of the reverse adaptor and
primer from read 1, and the incomplete RC of forward adaptor and
primer from read 2" # note that the input to this step is the output
from step 1 (demultiplexed-sequences iadapter-trimmed-demux-82.qza)
qiime cutadapt trim-paired \
--i-demultiplexed-sequences step1-adapter-trimmed-demux-82.qza \
--p-adapter-f ATTAGAWACCCBNGTAGTCCGGCTGGCTGACTATCTCGTATGCCGTCTTCTGCTTG
\
--p-adapter-r TTACCGCGGCKGCTGRCACACAATTACCATA \
--p-error-rate 0 \
--p-minimum-length 240 \
--p-no-discard-untrimmed \
--o-trimmed-sequences step2-adapter-trimmed-demux-82.qza \
--verbose > cutadapt_step2.out

echo "generate visualization file with summary from the cutadapt step
2"
qiime demux summarize \
--i-data step2-adapter-trimmed-demux-82.qza \
--o-visualization step2-adapter-trimmed-demux-82.qzv

# on télécharge les fichiers "rep-seq-82.qzv" et "table-82.qzv" du
dossier raw_qza pour les lire ensuite dans view qiime2

```

DADA2

```

# dada 2
# on télécharge les fichiers "step1-adapter-trimmed-demux-82.qzv" et
"step2-adapter-trimmed-demux-82.qzv" pour aller voir dans view qiime 2
# on télécharge la table de séquence donnée par view qiime 2 pour rep-
seq-82.qzv
# on l'appel sequences.fasta et on va charger la table dans le
terminal
cd raw_qza/ # pour être sûre d'etre dans le bon dossier
grep GTAGTCCGGCTGGCTGACTATCTCGTATGCCGTCTTCTGCTTG sequences.fasta >
id_overRep_bases.txt # pour obtenir un fichier txt
# le fichier id_overRep_bases.txt est créé dans le dossier raw_qza
# on le télécharge sur l'ordinateur et on va l'ouvrir avec notepad++
# on fait ctr+F pour rechercher la séquence :
GTAGTCCGGCTGGCTGACTATCTCGTATGCCGTCTTCTGCTTG
# quand on a trouvé la séquence on met une tabulation juste devant
pour séparer cette partie du reste de la séquence visuellement
TTTAATGATACGGCGACCAACGAGATCTACACGCTACTAAAGCAAATATGGTAATTGTGTGTCAGCAGC
CGCGGTAACACCTACAGGGCCGAGTTATCAGATTAGGTAATAAAGGCAGTTAGGCTAAGGCTTAATAATT
GTAATATTAAGGGTGCTTTAGGGGTAGAATCCTAAAAAGCGTTTTAAACTTGAAATTAACAATTAAGCC
GAATCTGCGAAGCCAAAAAGGGAACCTGGGATTAGAAACCCCTA
      GTAGTCCGGCTGGCTGACTATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAAAAAAAAAA
ATACGGCGACCAACGAGATCTACACGCTCGTATAAATGCGTATGGTAATTGTGTGTCAGCAGCCGCGGTA
ACACGATTAACCCAAGTCAATAGAAGCCGGCGTAAAGAGTGTTTTAGATCACCCCCTCCCAATAAAGCT
AAAACTCACCTGAGTTGTAAAAAACTCCAGTTGACACAAAATAGACTACGAAAGTGGCTTTAACATATCT
GAACACACAATAGCTAAGACCCAACTGGGATTAGAAACCCCT
      GTAGTCCGGCTGGCTGACTATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAAAAAAAAAA
ACCGAGATCTACACGCTCTGTCTATACTATATGGTAATTGTGTGTCAGCAGCCGCGGTAAATGGCAAAGA
TGGACGCTTTTACTAGTGAGAGGAAGTTCCTCGGCCTTTTTTAGGGGCCTCGGGATCTTCATAAAGTT
AATACGTATTAAGAGTTTCAACGCTGGCCTCGTGAGGTATTTAATATTTTTTAATTATTTTCAAATTTAA
TACAAAGAAAAAAAAATTTACAACGCAAGGATTAGAAACCCGC
      GTAGTCCGGCTGGCTGACTATCTCGTATGCCGTCTTCTGCTTGAAAA
GATACGGCGACCAACGAGATCTACACGCTACCTTACACCTTTATGGTAATTGTGTGCCAGCAGCCGCGGT
AATCTTAAACCCGAAAAAATGCGCACGGTTGATATCTTTAAGGGCAAATGAGGAAAGTAATAATAATAAT
GATAATAATAATGATTGGTAAAGTCAACCTTGCAACTCAGAAACAATTAAGTCCTTTATTCTTCCACCTG
TCCAAGCTATTTTAAAGAAACCTTCAATGTATTAGAAACCCGT
      GTAGTCCGGCTGGCTGACTATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAAAAAAAAAA
CGGCGACCAACGAGATCTACACGCTTACTAACGCGGTTATGGTAATTGTGTGCCAGCAGCCGCGGTAAGA
TGACCATATCTTGCAAATCGTGTAGCGTGGGCCGATAGGAACAACCAATCGGCATTGGACTGGCAAGACA
TCGTCAATTTGGCTGTTGCTATCGGCTTACGCTACACGATATACAAAAATATAAACGAAGCATAAAAAACG
AATTTCATAGTGTCATGAATTTATTGAACAATTAGATACCCCG
      GTAGTCCGGCTGGCTGACTATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAAAAAAAAAA
ACCACCGAGATCTACACGCTACCTTACACCTTTATGGTAATTGTGTGCCAGCCGCGCGGTAAAAATGGT
TATTCTTTATAGACAGACTGAGAATGTTTGCGGAGTTGAAAAGTTTAATTACAGTAGACTCGTACAACAA
GCATTAACGATTCTGCAATAGAGATCAGAAGACATCCACAGGAACGTGTAGTCCCAAGATATGCTGACAA
GGACTTTGTTTGTTGACTCTTACTTCGTCCATTAGAAACCCCT
      GTAGTCCGGCTGGCTGACTATCTCGTATGCCGTCTTCTGCTTGAAAAAA
CACCGAGATCTACACGCTACTAAAGCAAATAGGGTAATTGTGTGCCAGCAGCCGCGGTAATACGGAGGG
TGCAAGCGTTATCCGGATTTATTGGGTTTAAAGGCTAGAGTAGGTGAGAGGGGAAGTAGAATTTAGGTGT
AGCGGTGAAATGTGTAGAGATCTGAAGGAATACCGATGGCGAAGGCAGCTTCCTGGCATCATACTGACAC
TGAGGCTCGAAAGCGTGGGTAGCAAACAGGATTAGAAACCCCT

```

```
GTAGTCCGGCTGGCTGACTATCTCGTATGCCGTCTTCTGCTTGAAAAA
```

```
less cutadapt_step1.out  
nano cutadapt_remove_over-represented.sh
```

```
less cutadapt_remove_over-represented.sh
```

```
## Warning isn't present, so no extra trimming was performed (code for  
extra below, should be repeated by each seq to trim)
```

```
####qiime cutadapt trim-paired \
```

```
#!/bin/bash -e #-e veut dire que si il détecte une erreur il va  
s'arreter et signaler quelle est l'erreur
```

```
#SBATCH --job-name=cutadaptstep3
```

```
#SBATCH --output=cutadapts3_%j.out
```

```
#SBATCH --error=cutadapts3_%j.err
```

```
#SBATCH --mail-user=cece.koellsch@gmail.com
```

```
#SBATCH --mail-type=END
```

```
#SBATCH --time=2:00:00
```

```
#SBATCH --mem=10G
```

```
#SBATCH --ntasks 1
```

```
#SBATCH --profile=task
```

```
#SBATCH --account=uoo03555
```

```
module load QIIME2/2022.2
```

```
cd /nesi/nobackup/uoo03555/celia/Celia/qiime/
```

```
qiime cutadapt trim-paired \
```

```
--i-demultiplexed-sequences step1-adapter-trimmed-demux-82.qza \
```

```
--p-adapter-f CTAGTAGTCCGGCTGGCTGACTATCTCGTATGCCGTCTTCTGCTTG \
```

```
--p-error-rate 0 \
```

```
--p-minimum-length 240 \
```

```
--p-no-discard-untrimmed \
```

```
--o-trimmed-sequences step2-adapter-trimmed-demux-82.qza \
```

```
--output-dir removeCTA/ \
```

```
--verbose > cutadapt_step3.out
```

```
qiime cutadapt trim-paired \
```

```
--i-demultiplexed-sequences step2-adapter-trimmed-demux-82.qza \
```

```
--p-adapter-f CCCGTAGTCCGGCTGGCTGACTATCTCGTATGCCGTCTTCTGCTTG \
```

```
--p-error-rate 0 \
```

```
--p-minimum-length 240 \
```

```
--p-no-discard-untrimmed \
```

```
--o-trimmed-sequences step2-adapter-trimmed-demux-82.qza \
```

```
--output-dir removeCCC/ \
```

```
--verbose >> cutadapt_step3.out
```

```
qiime cutadapt trim-paired \
```

```
--i-demultiplexed-sequences step2-adapter-trimmed-demux-82.qza \
```

```
--p-adapter-f CGCGTAGTCCGGCTGGCTGACTATCTCGTATGCCGTCTTCTGCTTG \
```



```

--p-error-rate 0 \
--p-minimum-length 240 \
--p-no-discard-untrimmed \
--o-trimmed-sequences step2-adapter-trimmed-demux-82.qza \
--output-dir removeCGC/ \
--verbose >> cutadapt_step3.out

qiime cutadapt trim-paired \
--i-demultiplexed-sequences step2-adapter-trimmed-demux-82.qza \
--p-adapter-f CGTGTAGTCCGGCTGGCTGACTATCTCGTATGCCGTCTTCTGCTTG \
--p-error-rate 0 \
--p-minimum-length 240 \
--p-no-discard-untrimmed \
--o-trimmed-sequences step2-adapter-trimmed-demux-82.qza \
--output-dir removeCGT/ \
--verbose >> cutadapt_step3.out

qiime cutadapt trim-paired \
--i-demultiplexed-sequences step2-adapter-trimmed-demux-82.qza \
--p-adapter-f CCGGTAGTCCGGCTGGCTGACTATCTCGTATGCCGTCTTCTGCTTG \
--p-error-rate 0 \
--p-minimum-length 240 \
--p-no-discard-untrimmed \
--o-trimmed-sequences step2-adapter-trimmed-demux-82.qza \
--output-dir removeCCG/ \
--verbose >> cutadapt_step3.out

qiime cutadapt trim-paired \
--i-demultiplexed-sequences step2-adapter-trimmed-demux-82.qza \
--p-adapter-f CTTGTAGTCCGGCTGGCTGACTATCTCGTATGCCGTCTTCTGCTTG \
--p-error-rate 0 \
--p-minimum-length 240 \
--p-no-discard-untrimmed \
--o-trimmed-sequences step2-adapter-trimmed-demux-82.qza \
--output-dir removeCTT/ \
--verbose >> cutadapt_step3.out

qiime cutadapt trim-paired \
--i-demultiplexed-sequences step2-adapter-trimmed-demux-82.qza \
--p-adapter-f CTCGTAGTCCGGCTGGCTGACTATCTCGTATGCCGTCTTCTGCTTG \
--p-error-rate 0 \
--p-minimum-length 240 \
--p-no-discard-untrimmed \
--o-trimmed-sequences step2-adapter-trimmed-demux-82.qza \
--output-dir removeCTC/ \
--verbose >> cutadapt_step3.out

##generate visualization file with summary from the cutadapt step
above
qiime demux summarize \

```

```
--i-data step2-adapter-trimmed-demux-82.qza \  
--o-visualization step2-adapter-trimmed-demux-82.qzv
```

```
sbatch cutadapt_remove_over-represented.sh # job_id : 36690237
```

```
grep 'Pairs written (passing filters):' cutadapt_step2.out
```

```
Pairs written (passing filters):          49,568 (99.8%)  
Pairs written (passing filters):          66,059 (99.1%)  
Pairs written (passing filters):          82,553 (97.8%)  
Pairs written (passing filters):        239,889 (99.9%)  
Pairs written (passing filters):          46,272 (97.2%)  
Pairs written (passing filters):          32,481 (99.4%)  
Pairs written (passing filters):        219,743 (91.0%)  
Pairs written (passing filters):          82,418 (99.6%)  
Pairs written (passing filters):          29,624 (99.6%)  
Pairs written (passing filters):          29,343 (99.6%)  
Pairs written (passing filters):           144 (100.0%)  
Pairs written (passing filters):    1,028,517 (99.3%)  
Pairs written (passing filters):           3,194 (99.7%)  
Pairs written (passing filters):           7,334 (99.8%)  
Pairs written (passing filters):           2,173 (99.1%)  
Pairs written (passing filters):           3,753 (99.7%)  
Pairs written (passing filters):           2,723 (99.7%)  
Pairs written (passing filters):           943 (100.0%)  
Pairs written (passing filters):         80,847 (99.2%)  
Pairs written (passing filters):           6,852 (99.5%)  
Pairs written (passing filters):           6,752 (99.8%)  
Pairs written (passing filters):        13,002 (99.8%)  
Pairs written (passing filters):           4,412 (99.5%)  
Pairs written (passing filters):           6,187 (100.0%)  
Pairs written (passing filters):           753 (100.0%)  
Pairs written (passing filters):        36,621 (99.6%)  
Pairs written (passing filters):           1,584 (99.2%)  
Pairs written (passing filters):        30,055 (99.8%)  
Pairs written (passing filters):       132,338 (97.3%)  
Pairs written (passing filters):        29,091 (99.6%)  
Pairs written (passing filters):           5,206 (99.8%)  
Pairs written (passing filters):           267 (100.0%)  
Pairs written (passing filters):           1,802 (100.0%)  
Pairs written (passing filters):           3,646 (98.8%)  
Pairs written (passing filters):       302,481 (100.0%)  
Pairs written (passing filters):           5,424 (100.0%)  
Pairs written (passing filters):          9,846 (100.0%)  
Pairs written (passing filters):          4,961 (100.0%)  
Pairs written (passing filters):          9,873 (99.9%)  
Pairs written (passing filters):          7,372 (99.9%)  
Pairs written (passing filters):          9,396 (100.0%)  
Pairs written (passing filters):        11,064 (99.9%)  
Pairs written (passing filters):        19,184 (100.0%)
```

Pairs written (passing filters):	48,724	(99.8%)
Pairs written (passing filters):	10,480	(99.8%)
Pairs written (passing filters):	10,113	(99.9%)
Pairs written (passing filters):	169	(99.4%)
Pairs written (passing filters):	967	(99.8%)
Pairs written (passing filters):	1,242,850	(99.8%)
Pairs written (passing filters):	67,529	(99.9%)
Pairs written (passing filters):	57,053	(99.9%)
Pairs written (passing filters):	15,765	(99.3%)
Pairs written (passing filters):	19,034	(99.5%)
Pairs written (passing filters):	63,005	(99.7%)
Pairs written (passing filters):	75,956	(100.0%)
Pairs written (passing filters):	10,154	(100.0%)
Pairs written (passing filters):	2,314	(99.9%)
Pairs written (passing filters):	8,539	(99.8%)
Pairs written (passing filters):	2,136	(99.7%)
Pairs written (passing filters):	8,550	(99.8%)
Pairs written (passing filters):	9,170	(100.0%)
Pairs written (passing filters):	3,098	(99.9%)
Pairs written (passing filters):	17,571	(100.0%)
Pairs written (passing filters):	1,302	(99.9%)
Pairs written (passing filters):	1,188	(99.5%)
Pairs written (passing filters):	1,732	(99.8%)
Pairs written (passing filters):	3,521	(99.6%)
Pairs written (passing filters):	1,804	(99.8%)
Pairs written (passing filters):	2,493	(99.7%)
Pairs written (passing filters):	1,541	(99.8%)
Pairs written (passing filters):	2,324	(99.8%)
Pairs written (passing filters):	1,422	(99.9%)
Pairs written (passing filters):	855	(99.8%)
Pairs written (passing filters):	7,692	(99.8%)
Pairs written (passing filters):	4,736	(99.9%)
Pairs written (passing filters):	1	(100.0%)
Pairs written (passing filters):	2,729	(99.5%)
Pairs written (passing filters):	2,721	(99.9%)
Pairs written (passing filters):	3,281	(99.8%)
Pairs written (passing filters):	15,423	(100.0%)
Pairs written (passing filters):	33,905	(100.0%)
Pairs written (passing filters):	1,160	(100.0%)
Pairs written (passing filters):	3,940	(99.8%)
Pairs written (passing filters):	28,487	(99.9%)
Pairs written (passing filters):	2,292	(100.0%)
Pairs written (passing filters):	1,321	(100.0%)
Pairs written (passing filters):	411	(100.0%)
Pairs written (passing filters):	12,372	(100.0%)
Pairs written (passing filters):	98,342	(100.0%)
Pairs written (passing filters):	17,194	(99.9%)
Pairs written (passing filters):	585,529	(100.0%)
Pairs written (passing filters):	15,361	(99.9%)
Pairs written (passing filters):	59,671	(100.0%)

Pairs written (passing filters):	39,928	(99.9%)
Pairs written (passing filters):	12,817	(100.0%)
Pairs written (passing filters):	1	(100.0%)
Pairs written (passing filters):	3	(100.0%)
Pairs written (passing filters):	2	(100.0%)
Pairs written (passing filters):	15,115	(96.0%)
Pairs written (passing filters):	22,088	(98.7%)
Pairs written (passing filters):	20,543	(72.5%)
Pairs written (passing filters):	16,695	(78.4%)
Pairs written (passing filters):	24,772	(85.8%)
Pairs written (passing filters):	14,365	(66.7%)
Pairs written (passing filters):	30,535	(81.8%)
Pairs written (passing filters):	82,894	(99.1%)
Pairs written (passing filters):	9,869	(85.2%)
Pairs written (passing filters):	7,844	(84.8%)
Pairs written (passing filters):	59,948	(100.0%)
Pairs written (passing filters):	76,536	(99.6%)
Pairs written (passing filters):	32,402	(99.8%)
Pairs written (passing filters):	21,946	(97.5%)
Pairs written (passing filters):	4,159	(78.5%)
Pairs written (passing filters):	74,363	(96.1%)
Pairs written (passing filters):	58,154	(86.4%)
Pairs written (passing filters):	50,539	(98.0%)
Pairs written (passing filters):	34,990	(75.4%)
Pairs written (passing filters):	74,973	(99.4%)
Pairs written (passing filters):	127,553	(97.7%)
Pairs written (passing filters):	9,464	(84.7%)
Pairs written (passing filters):	5,356	(74.0%)
Pairs written (passing filters):	7,423	(83.3%)
Pairs written (passing filters):	46,452	(82.6%)
Pairs written (passing filters):	26,040	(91.9%)
Pairs written (passing filters):	46,828	(98.8%)
Pairs written (passing filters):	177,898	(100.0%)
Pairs written (passing filters):	82,740	(100.0%)
Pairs written (passing filters):	61,415	(99.9%)
Pairs written (passing filters):	45,946	(99.0%)
Pairs written (passing filters):	12,237	(67.5%)
Pairs written (passing filters):	2,893	(51.7%)
Pairs written (passing filters):	18,264	(78.2%)
Pairs written (passing filters):	56,287	(99.1%)
Pairs written (passing filters):	130,097	(99.9%)
Pairs written (passing filters):	1,045,670	(99.9%)
Pairs written (passing filters):	46,653	(99.9%)
Pairs written (passing filters):	20,611	(87.2%)
Pairs written (passing filters):	119,336	(99.9%)
Pairs written (passing filters):	6,354	(74.9%)
Pairs written (passing filters):	7,472	(99.9%)
Pairs written (passing filters):	14,149	(68.5%)
Pairs written (passing filters):	33,598	(99.9%)
Pairs written (passing filters):	50,439	(85.1%)

```
Pairs written (passing filters):      45,420 (73.1%)
Pairs written (passing filters):      8,725 (57.4%)
Pairs written (passing filters):      30,878 (99.9%)
Pairs written (passing filters):      8,991 (96.7%)
Pairs written (passing filters):     11,379 (90.7%)
Pairs written (passing filters):      8,944 (84.5%)
Pairs written (passing filters):     167,151 (99.3%)
Pairs written (passing filters):      30,256 (96.3%)
Pairs written (passing filters):      42,000 (100.0%)
Pairs written (passing filters):      24,299 (99.9%)
Pairs written (passing filters):       2,503 (94.9%)
Pairs written (passing filters):         6 (100.0%)
```

```
#on verifie si le programme fonctionne
sbatch cutadapt_remove_over-represented.sh
sacct
# ca n'a pas fonctionné, on regarde pourquoi :
less cutadapts3_36690066.err
```

---

```
#si on veut voir les paramètres des précédentes programmation pour
voir le timing ou les GB :
nn_seff 36633325 #nn_seff n° du job id
# ne fonctionne qu'avec les programmations finies pour voir les
paramètres qu'il a nécessité
# ca donne ca :
Cluster: mahuika
Job ID: 36633325
State: TIMEOUT
Cores: 1
Tasks: 1
Nodes: 1
Job Wall-time: 100.1% 06:00:19 of 06:00:00 time limit
CPU Efficiency: 99.7% 05:59:17 of 06:00:19 core-walltime
Mem Efficiency: 53.8% 5.38 GB of 10.00 GB
```

Training silva

```
# silva est le site de référence pour les séquences bactériennes
https://www.arb-silva.de/

# créer un fichier silva pour mettre nos trucs
mkdir training_silva
cd training_silva

nano build_silva_classifier.sh

#!/bin/bash -e
#SBATCH --job-name=silva_class
```

```

#SBATCH --output=classifier_%j.out
#SBATCH --error=classifier_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END
#SBATCH --qos=DEBUG #time=36:00:00
#SBATCH --mem=40G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

#download the data using Silva, and get ready-to-use sequence and
taxonomy files
module load QIIME2/2022.2

cd /nesi/nobackup/uoo03555/celia/Celia/qiime/

qiime rescript get-silva-data \
  --p-version '138.1' \
  --p-target 'SSURef_NR99' \
  --p-include-species-labels \
  --o-silva-sequences silva-138.1-ssu-nr99-rna-seqs.qza \
  --o-silva-taxonomy silva-138.1-ssu-nr99-tax.qza

#“Culling” low-quality sequences with cull-seqs
qiime rescript cull-seqs \
  --i-sequences silva-138.1-ssu-nr99-rna-seqs.qza \
  --o-clean-sequences silva-138.1-ssu-nr99-seqs-cleaned.qza

#Filtering sequences by length and taxonomy - We will remove rRNA gene
sequences that do not meet the following criteria:
#Archaea (16S) >= 900 bp, Bacteria (16S) >= 1200 bp, and any Eukaryota
(18S) >= 1400 bp
qiime rescript filter-seqs-length-by-taxon \
  --i-sequences silva-138.1-ssu-nr99-seqs-cleaned.qza \
  --i-taxonomy silva-138.1-ssu-nr99-tax.qza \
  --p-labels Archaea Bacteria Eukaryota \
  --p-min-lens 900 1200 1400 \
  --o-filtered-seqs silva-138.1-ssu-nr99-seqs-filt.qza \
  --o-discarded-seqs silva-138.1-ssu-nr99-seqs-discard.qza

# Dereplication of sequences and taxonomy (method uniq will retain
identical sequence records that have differing taxonomies)
#Uniq is the default methods, but there are other options -
https://forum.qiime2.org/t/processing-filtering-and-evaluating-the-silva-database-and-other-reference-sequence-data-with-rescript/15494
qiime rescript dereplicate \
  --i-sequences silva-138.1-ssu-nr99-seqs-filt.qza \
  --i-taxa silva-138.1-ssu-nr99-tax.qza \
  --p-rank-handles 'silva' \
  --p-mode 'uniq' \

```



```

--o-dereplicated-sequences silva-138.1-ssu-nr99-seqs-derep-
uniq.qza \
--o-dereplicated-taxa silva-138.1-ssu-nr99-tax-derep-uniq.qza

# make classifier
qiime feature-classifier fit-classifier-naive-bayes \
--i-reference-reads silva-138.1-ssu-nr99-seqs-derep-uniq.qza \
--i-reference-taxonomy silva-138.1-ssu-nr99-tax-derep-uniq.qza \
--o-classifier silva-138.1-ssu-nr99-classifier.qza

# Use the classifier above and make an amplicon-region specific
classifier (specific for the primers used)
# For the UORG, we used the EMP primers for the v4 region of the 16S
rRNA gene (http://earthmicrobiome.org/protocols-and-standards/16s/)
#The next step will extract the relevant region (v4) from the
previously generated classifier

qiime feature-classifier extract-reads \
--i-sequences silva-138.1-ssu-nr99-seqs-derep-uniq.qza \
--p-f-primer GTGYCAGCMGCCGCGGTAA \
--p-r-primer GGACTACNVGGGTWTCTAAT \
--p-n-jobs 2 \
--p-read-orientation 'forward' \
--o-reads silva-138.1-ssu-nr99-seqs-515f-806r.qza

#De-replicate the extracted region
qiime rescript dereplicate \
--i-sequences silva-138.1-ssu-nr99-seqs-515f-806r.qza \
--i-taxa silva-138.1-ssu-nr99-tax-derep-uniq.qza \
--p-rank-handles 'silva' \
--p-mode 'uniq' \
--o-dereplicated-sequences silva-138.1-ssu-nr99-seqs-515f-806r-
uniq.qza \
--o-dereplicated-taxa silva-138.1-ssu-nr99-tax-515f-806r-derep-
uniq.qza

#make amplicon-specific classifier
qiime feature-classifier fit-classifier-naive-bayes \
--i-reference-reads silva-138.1-ssu-nr99-seqs-515f-806r-uniq.qza \
--i-reference-taxonomy silva-138.1-ssu-nr99-tax-515f-806r-derep-
uniq.qza \
--o-classifier silva-138.1-ssu-nr99-515f-806r-classifier.qza

# Evaluate the classifier fit

qiime rescript evaluate-fit-classifier \
--i-sequences silva-138.1-ssu-nr99-seqs-515f-806r-uniq.qza \
--i-taxonomy silva-138.1-ssu-nr99-tax-515f-806r-derep-uniq.qza \
--o-classifier silva-138.1-ssu-nr99-515f-806r-classifier.qza \

```

```

--o-observed-taxonomy silva-138.1-ssu-nr99-515f-806r-predicted-
taxonomy.qza \
--o-evaluation silva-138.1-ssu-nr99-515f-806r-fit-classifier-
evaluation.qzv

sbatch build_silva_classifier.sh # job id : 36690456
sacct

# si on voit que ça fonctionne dans les premières minutes alors on
# peut retirer la fonction qos=DEBUG #time
# pour ça on fait
scancel 36690456 #scancel num du projet
# le programme arrête de tourner
# on reva dans
nano build_silva_classifier.sh # on retire la fonction qos=DEBUG #time
sbatch build_silva_classifier.sh # job id : 36690505
sacct

sacct -j 36690505 #on veut voir où en est le projet
#le projet manquait de mémoire alors on en a fait un nouveau avec
100GB :
sbatch build_silva_classifier_moreMemory.sh #etc...

Error in parse(text = x, srcfile = src): <text>:4:7: unexpected symbol
3: # créer un fichier silva pour mettre nos trucs
4: mkdir training_silva
   ^
Traceback:

```

removing primers (19.06.2023)

```

nano cutadaptP_dada2.sh # nano pour créer la boucle de programmation
#!/bin/bash -e
#SBATCH --job-name=cut_dada2
#SBATCH --output=cut_dada2_%j.out
#SBATCH --error=cut_dada2_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END
#SBATCH --time=24:00:00
#SBATCH --mem=10G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

module load QIIME2/2022.2

cd /nesi/nobackup/uoo03555/celia/Celia/qiime/

## Primers removal ##

```



```
qiime cutadapt trim-paired \
--i-demultiplexed-sequences step2-adapter-trimmed-demux-82.qza \
--p-front-f TATGGTAATTGTGTGYCAGCMGCCGCGGTAA \
--p-front-r AGTCAGCCAGCCGGACTACNVGGGTWTCTAAT \
--p-error-rate 0 \
--p-minimum-length 240 \
--p-no-discard-untrimmed \
--o-trimmed-sequences primer-adapter-trimmed-demux-82.qza \
--verbose
```

```
qiime demux summarize \
--i-data primer-adapter-trimmed-demux-82.qza \
--o-visualization primer-adapter-trimmed-demux-82.qzv
```

*#Run dada2 on the reads cleaned with cutadapt (this will generate a feature table)*

```
qiime dada2 denoise-paired \
--i-demultiplexed-seqs primer-adapter-trimmed-demux-82.qza \
--p-trim-left-f 13 \
--p-trim-left-r 13 \
--p-trunc-len-f 240 \
--p-trunc-len-r 240 \
--o-table table-82.qza \
--o-representative-sequences rep-seqs-82.qza \
--o-denoising-stats denoising-stats-82.qza \
--output-dir cutadaptP/
```

*#Create visualization (summary file) of the feature table*

*#Required files:*

*#- table-82.qza #file generated with the dada2 denoise-paired command above*

*#- metadata\_tsv.tsv #tsv file with metadata*

```
qiime feature-table summarize \
--i-table table-82.qza \
--o-visualization table-82.qzv \
--m-sample-metadata-file metadata_tsv.tsv
```

```
qiime feature-table tabulate-seqs \
--i-data rep-seqs-82.qza \
--o-visualization rep-seqs-82.qzv
```

```
qiime metadata tabulate \
--m-input-file denoising-stats-82.qza \
--o-visualization denoising-stats-82.qzv
```

```
sbatch cutadaptP_dada2.sh
```

*# job id : 36768187*

*sacct -j 36768187 # pour voir où en est le programme*

remove non bacterial sequences 20.06

```
nano bacteria-only.sh
#!/bin/bash -e
#SBATCH --job-name=bac-only
#SBATCH --output=bac-only_%j.out
#SBATCH --error=bac-only_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END
#SBATCH --time=4:00:00
#SBATCH --mem=40G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

module load QIIME2/2022.2

cd /nesi/nobackup/uoo03555/celia/Celia/qiime/

qiime quality-control exclude-seqs \
  --i-query-sequences rep-seqs-08.qza \
  --i-reference-sequences silva_reference_Priscila/silva-138.1-ssu-
nr99-seqs-515f-806r.qza \
  --p-method blast \
  --p-perc-identity 0.8 \
  --p-perc-query-aligned 0.8 \
  --o-sequence-hits hits-08-82.qza \
  --o-sequence-misses misses-08-82.qza

qiime feature-table tabulate-seqs \
  --i-data hits-08-82.qza \
  --o-visualization hits-08-82.qzv

qiime feature-table tabulate-seqs \
  --i-data misses-08-82.qza \
  --o-visualization misses-08-82.qzv

echo 'step2'
echo 'taxa filter'

qiime feature-classifier classify-sklearn \
  --i-classifier silva_reference_Priscila/silva-138.1-ssu-nr99-515f-
806r-classifier.qza \
  --i-reads hits-08-82.qza \
  --p-n-jobs 2 \
  --o-classification hits-08-82-observed-taxonomy-silva.qza

qiime metadata tabulate \
  --m-input-file hits-08-82-observed-taxonomy-silva.qza \
  --o-visualization hits-08-82-observed-taxonomy-silva.qzv
```

```
sbatch bacteria-only.sh # job id : 36800568
sacct
```

22.06

```
nano feature_taxaID-filter.sh
#!/bin/bash -e
#SBATCH --job-name=filter
#SBATCH --output=feat_taxID_filter_%j.out
#SBATCH --error=feat_taxID_filter_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --time=3:00:00
#SBATCH --mem=10G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

module load QIIME2/2022.2

cd /nesi/nobackup/uoo03555/celia/Celia/qiime/

echo 'filtering features to output hits08-filtered-table-P7'
qiime feature-table filter-features \
  --i-table table-82.qza \
  --m-metadata-file hits-08-82.qza \
  --o-filtered-table hits08-filtered-table-82.qza
qiime feature-table summarize \
  --i-table hits08-filtered-table-82.qza \
  --o-visualization hits08-filtered-table-82.qzv

# view the taxonomic composition of our samples with interactive bar
plots
echo 'Using SILVA'
echo 'making bar plot of taxonomic composition of samples'
qiime taxa barplot \
  --i-table hits08-filtered-table-82.qza \
  --i-taxonomy hits-08-82-observed-taxonomy-silva.qza \
  --m-metadata-file metadata_tsv.tsv \
  --o-visualization hits08-taxa-bar-plots-silva.qzv

echo 'isolating features without phyla'
# try to isolate hits without phyla
qiime taxa filter-seqs \
  --i-sequences hits-08-82.qza \
  --i-taxonomy hits-08-82-observed-taxonomy-silva.qza \
  --p-exclude p__ \
  --o-filtered-sequences hits-08-82-without-phyla-silva.qza
qiime metadata tabulate \
  --m-input-file hits-08-82-without-phyla-silva.qza \
```

```
--o-visualization hits-08-82-without-phyla-silva.qzv
```

```
sbatch feature_taxaID-filter.sh # job id : 36845731  
sacct
```

filtre taxonomique 22.06

```
nano taxa-remove.sh  
#!/bin/bash -e  
#SBATCH --job-name=miss_filter  
#SBATCH --output=filtermiss_%j.out  
#SBATCH --error=filtermiss_%j.err  
#SBATCH --mail-user=cece.koellsch@gmail.com  
#SBATCH --mail-type=END  
#SBATCH --time=4:00:00  
#SBATCH --mem=50G  
#SBATCH --ntasks 1  
#SBATCH --profile=task  
#SBATCH --account=uoo03555  
  
module load QIIME2/2022.2  
  
cd /nesi/nobackup/uoo03555/celia/Celia/qiime/  
  
echo 'running silva classifier'  
qiime feature-classifier classify-sklearn \  
  --i-classifier silva_reference_Priscila/silva-138.1-ssu-nr99-515f-  
806r-classifier.qza \  
  --i-reads misses-08-82.qza \  
  --p-n-jobs 2 \  
  --o-classification misses-08-82-observed-taxonomy-silva.qza  
  
qiime metadata tabulate \  
  --m-input-file misses-08-82-observed-taxonomy-silva.qza \  
  --o-visualization misses-08-82-observed-taxonomy-silva.qzv  
  
echo 'filtering features to output misses-08-filtered-table-82'  
qiime feature-table filter-features \  
  --i-table table-82.qza \  
  --m-metadata-file misses-08-82.qza \  
  --o-filtered-table misses-08-filtered-table-82.qza  
qiime feature-table summarize \  
  --i-table misses-08-filtered-table-82.qza \  
  --o-visualization misses-08-filtered-table-82.qzv  
  
echo 'making bar plot of taxonomic composition of samples'  
qiime taxa barplot \  
  --i-table misses-08-filtered-table-82.qza \  
  --i-taxonomy misses-08-82-observed-taxonomy-silva.qza \  
  --m-metadata-file metadata_tsv.tsv \  
  --o-visualization misses-08-filtered-table-82.qzv
```

```
--o-visualization misses-08-taxa-bar-plots-silva.qzv  
sbatch taxa-remove.sh # job id : 36846610
```

filtrer par phylum 22.06

```
# --p-include to retain only features that contain a phylum-level  
# Are there are Chloroplast (level 3) and mitochondria (level 5)  
# sequences?  
# if yes, remove it together with unclassified reads:  
  
nano mtchlot_filter.sh  
#!/bin/bash -e  
#SBATCH --job-name=mtchlot  
#SBATCH --output=filtermtchlot_%j.out  
#SBATCH --error=filtermtchlot_%j.err  
#SBATCH --mail-user=cece.koelisch@gmail.com  
#SBATCH --mail-type=END  
#SBATCH --time=00:30:00  
#SBATCH --mem=10G  
#SBATCH --ntasks 1  
#SBATCH --profile=task  
#SBATCH --account=uoo03555  
  
module load QIIME2/2022.2  
  
cd /nesi/nobackup/uoo03555/celia/Celia/qiime/  
  
echo 'filtering arthropoda out based on silva'  
qiime taxa filter-table \  
  --i-table hits08-filtered-table-82.qza \  
  --i-taxonomy hits-08-82-observed-taxonomy-silva.qza \  
  --p-include p__ \  
  --p-exclude p__Arthropoda \  
  --o-filtered-table hits-08-82-no-mtchlot-withP-noArthropoda-  
silva.qza \  
  
echo 'generating qzv file'  
qiime feature-table summarize \  
  --i-table hits-08-82-no-mtchlot-withP-noArthropoda-silva.qza \  
  --o-visualization hits-08-82-no-mtchlot-withP-silva.qzv  
  
echo 'generating silva barplot'  
qiime taxa barplot \  
  --i-table hits-08-82-no-mtchlot-withP-noArthropoda-silva.qza \  
  --i-taxonomy hits-08-82-observed-taxonomy-silva.qza \  
  --m-metadata-file metadata_tsv.tsv \  
  --o-visualization hits-08-82-no-mtchlot-withP-bar-plots-silva.qzv
```

```
sbatch mtchlot_filter.sh #job id : 36847554
sacct
```

filtrer en retirant tout l'embranchement des malacostracés 22.06

```
#Excluding Malacostraca
https://www.arb-silva.de/browser/ssu/silva/L78902/#15

nano malacostraca_filter.sh
#!/bin/bash -e
#SBATCH --job-name=molFilter
#SBATCH --output=filterMalacostraca_%j.out
#SBATCH --error=filterMalacostraca_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END
#SBATCH --qos=debug #time=00:30:00
#SBATCH --mem=10G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

module load QIIME2/2022.2

cd /nesi/nobackup/uoo03555/celia/Celia/qiime/

echo 'filtering malacostraca out based on silva'
qiime taxa filter-table \
  --i-table hits08-filtered-table-82.qza \
  --i-taxonomy hits-08-82-observed-taxonomy-silva.qza \
  --p-include c__ \
  --p-exclude c__Malacostraca \
  --o-filtered-table hits-08-82-no-mtchlor-withP-noMalacostraca-
silva.qza \

echo 'generating qzv file'
qiime feature-table summarize \
  --i-table hits-08-82-no-mtchlor-withP-noMalacostraca-silva.qza \
  --o-visualization hits-08-82-no-mtchlor-withP-noMalacostraca-
silva.qzv

echo 'generating silva barplot'
qiime taxa barplot \
  --i-table hits-08-82-no-mtchlor-withP-noMalacostraca-silva.qza \
  --i-taxonomy hits-08-82-observed-taxonomy-silva.qza \
  --m-metadata-file metadata_tsv.tsv \
  --o-visualization hits-08-82-no-mtchlor-withP-noMalacostraca-bar-
plots-silva.qzv
```

```
sbatch malacostraca_filter.sh #job id : 36849178
```

filtrer en retirant tout l'embranchement des eukaryotes 22.06

```
# Filter all eukaryotes out
```

```
nano eukaryota_filter.sh
```

```
#!/bin/bash -e
```

```
#SBATCH --job-name=molFilter
```

```
#SBATCH --output=filterMalacostraca_%j.out
```

```
#SBATCH --error=filterMalacostraca_%j.err
```

```
#SBATCH --mail-user=cece.koellsch@gmail.com
```

```
#SBATCH --mail-type=END
```

```
#SBATCH --qos=debug #time=00:15:00
```

```
#SBATCH --mem=10G
```

```
#SBATCH --ntasks 1
```

```
#SBATCH --profile=task
```

```
#SBATCH --account=uoo03555
```

```
module load QIIME2/2022.2
```

```
cd /nesi/nobackup/uoo03555/celia/Celia/qiime/
```

```
echo 'filtering malacostraca out based on silva'
```

```
qiime taxa filter-table \
```

```
--i-table hits08-filtered-table-82.qza \
```

```
--i-taxonomy hits-08-82-observed-taxonomy-silva.qza \
```

```
--p-include p__ \
```

```
--p-exclude d__Eukaryota \
```

```
--o-filtered-table hits-08-82-no-mtchlr-withP-noMalacostraca-silva.qza \
```

```
echo 'generating qzv file'
```

```
qiime feature-table summarize \
```

```
--i-table hits-08-82-no-mtchlr-withP-noMalacostraca-silva.qza \
```

```
--o-visualization hits-08-82-no-mtchlr-withP-noMalacostraca-silva.qzv
```

```
echo 'generating silva barplot'
```

```
qiime taxa barplot \
```

```
--i-table hits-08-82-no-mtchlr-withP-noMalacostraca-silva.qza \
```

```
--i-taxonomy hits-08-82-observed-taxonomy-silva.qza \
```

```
--m-metadata-file metadata_tsv.tsv \
```

```
--o-visualization hits-08-82-no-mtchlr-withP-noMalacostraca-barplots-silva.qzv
```

```
sbatch eukaryota_filter.sh #job id : 36849265
```

```
sacct
```

## Verif rarefaction des courbes de la taxo silva 22.06

```
# Vérifiez d'abord la relation entre les blancs et les échantillons.  
# Une contamination bizarre ?  
# Générer un arbre pour l'analyse de la diversité phylogénétique  
# Les diagrammes de raréfaction alpha indiquent l'impact de la  
# profondeur d'échantillonnage sur la diversité alpha.  
#(qui sera tangentielle liée aux impacts sur la diversité bêta et  
# à d'autres analyses en aval, c'est donc un bon repère approximatif à  
# utiliser ici).  
# un bon point de repère approximatif à utiliser ici). Nous  
# recherchons idéalement une profondeur de séquençage au point où ces #  
# courbes de raréfaction commencent à s'estomper.  
# courbes de raréfaction commencent à se stabiliser (ce qui indique  
# que la majeure partie de la diversité pertinente a été capturée).
```

```
nano check-sample-blanks.sh  
#!/bin/bash -e  
#SBATCH --job-name=checksamblan  
#SBATCH --output=sample-blanks_%j.out  
#SBATCH --error=sample-blanks_%j.err  
#SBATCH --mail-user=cece.koellsch@gmail.com  
#SBATCH --mail-type=END  
#SBATCH --time=4:00:00  
#SBATCH --mem=25G  
#SBATCH --ntasks 1  
#SBATCH --profile=task  
#SBATCH --account=uoo03555
```

```
module load QIIME2/2022.2
```

```
cd /nesi/nobackup/uoo03555/celia/Celia/qiime/
```

```
# filter sequences with silva  
echo 'filter sequences with silva'  
qiime feature-table filter-seqs \  
--i-data hits-08-82.qza \  
--i-table hits-08-82-no-mtchclor-withP-noMalacostraca-silva.qza \  
--o-filtered-data hits-08-82-no-mtchclor-withP-noMalacostraca-silva-  
rep-seqs.qza
```

```
qiime feature-table tabulate-seqs \  
--i-data hits-08-82-no-mtchclor-withP-noMalacostraca-silva-rep-  
seqs.qza \  
--o-visualization hits-08-82-no-mtchclor-withP-noMalacostraca-silva-  
rep-seqs.qzv
```

```
echo 'aligning and estimating tree with silva'  
# align and estimate tree  
qiime phylogeny align-to-tree-mafft-fasttree \  

```



```

--i-sequences hits-08-82-no-mtchlor-withP-noMalacostraca-silva-rep-
seqs.qza \
--o-alignment aligned-rep-seqs-ALLraw-82-silva.qza \
--o-masked-alignment masked-align-rep-seqs-ALLraw-82-silva.qza \
--o-tree unrooted-tree-ALLraw-82-silva.qza \
--o-rooted-tree rooted-tree-ALLraw-82-silva.qza

echo 'alpha rarefaction with silva'
qiime diversity alpha-rarefaction \
--i-table hits-08-82-no-mtchlor-withP-noMalacostraca-silva.qza \
--i-phylogeny rooted-tree-ALLraw-82-silva.qza \
--p-max-depth 4000 \
--m-metadata-file metadata_tsv.tsv \
--o-visualization ALLraw-alpha-rarefaction-82-silva.qzv

sbatch check-sample-blanks.sh #job id : 36849762
sacct

```

Pour l'étape suivante, il faut choisir un seuil de profondeur d'échantillonnage. Il doit s'agir d'un équilibre entre la conservation du plus grand nombre d'échantillons possible (donc pas trop élevé) et la capture d'une grande diversité (pas trop faible). Jetez un coup d'œil au fichier : hits-08-82-no-mtchlor-withP-noMalacostraca-silva.qzv, en particulier à la fréquence minimale (ici c'est 84.0) de la vue d'ensemble, puis au "feature number" dans la vue interactive des détails de l'échantillon (= Retenir 7 339 266 (100,00 %) features dans 150 (100,00 %) échantillons à la profondeur d'échantillonnage spécifiée.). Jetez également un coup d'œil aux courbes de fractions rares, qui devraient se stabiliser lorsque la profondeur n'a plus d'impact sur la découverte d'éléments.

Le fichier contenant les courbes de fractions rares est le suivant : ALLraw-alpha-rarefaction-82-silva.qzv Ici, la fréquence minimale est de 84 (avec silva). Si l'on examine le graphique des fractions rares, la plupart des échantillons se stabilisent juste avant une profondeur de 600 (le max est à 1200, sûrement les échantillons environnementaux), à moins qu'il ne s'agisse de contrôles environnementaux, qui n'ont pas montré de plateau même à une profondeur de 4000. Pour éviter de perdre trop d'échantillons, et compte tenu du fait que la plupart des échantillons se stabilisent à 600, j'aurai deux ensembles de données, l'un avec une coupure à 600 et l'autre avec une coupure à 1300.

```

# diversité core-metrics-phylogenetic
# J'ai sélectionné la profondeur d'échantillonnage en fonction des
détails de l'échantillon et du texte dans la cellule ci-dessus.

```

```

nano core-metrics.sh
#!/bin/bash -e
#SBATCH --job-name=coremetrics
#SBATCH --output=coremetrics_%j.out
#SBATCH --error=coremetrics_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END

```

```

#SBATCH --time=4:00:00
#SBATCH --mem=25G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

module load QIIME2/2022.2

cd /nesi/nobackup/uoo03555/celia/Celia/qiime/

echo 'depth-filtering with silva, output core metrics'
qiime diversity core-metrics-phylogenetic \
  --i-phylogeny rooted-tree-ALLraw-82-silva.qza \
  --i-table hits-08-82-no-mtchlr-withP-noMalacostraca-silva.qza \
  --p-sampling-depth 500 \
  --m-metadata-file metadata_tsv.tsv \
  --output-dir core-metrics-results-withBlanks-silva

#In the output directory, find:
#rarefied_table.qza
#faith_pd_vector.qza
#observed_features_vector.qza
#shannon_vector.qza
#evenness_vector.qza
#unweighted_unifrac_distance_matrix.qza
#weighted_unifrac_distance_matrix.qza
#jaccard_distance_matrix.qza
#bray_curtis_distance_matrix.qza
#unweighted_unifrac_pcoa_results.qza
#weighted_unifrac_pcoa_results.qza
#jaccard_pcoa_results.qza
#bray_curtis_pcoa_results.qza
#unweighted_unifrac_emperor.qzv
#weighted_unifrac_emperor.qzv
#jaccard_emperor.qzv
#bray_curtis_emperor.qzv

sbatch core-metrics.sh# job id: 36850033
sacct

```

D'après les graphiques générés ci-dessus avec greengenes et silva :

- Les PBS se trouvent parmi les échantillons ; les PBS de la bouteille sont plus proches les uns des autres, les PBS des lavages sont plus distribués parmi les échantillons ;
- l'ADN et les blancs de PCR sont relativement groupés et se retrouvent parmi les échantillons.
- l'environnement du laboratoire se regroupe
- Les MCS sont regroupées ;
- le substrat est regroupé ;

- les échantillons d'hôtes se chevauchent avec les échantillons de parasites ;

vérifier la diversité alpha 22.06

```
nano alpha-signif.sh
#!/bin/bash -e
#SBATCH --job-name=alpha
#SBATCH --output=alphasignif_%j.out
#SBATCH --error=alphasignif_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END
#SBATCH --time=1:00:00
#SBATCH --mem=10G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

module load QIIME2/2022.2

cd /nesi/nobackup/uoo03555/celia/Celia/qiime/

echo 'alpha-group-significance with silva'
# check alpha diversity: how diverse or less diverse are blank samples
qiime diversity alpha-group-significance \
  --i-alpha-diversity core-metrics-results-withBlanks-
silva/shannon_vector.qza \
  --m-metadata-file metadata_tsv.tsv \
  --o-visualization core-metrics-results-withBlanks-silva/shannon-
group-significance.qzv

sbatch alpha-signif.sh #job id 36850074
sacct

# Importer dans view qiime
# D'après les résultats ci-dessus :
# Les échantillons environnementaux ont la plus grande diversité
alpha.
# Les acantho sont significativement différents des amphipodes
```

Élimination des contaminants Il est difficile de déterminer ce qu'est une contamination, et de nombreux éléments doivent être pris en compte. 23.06

Une contamination croisée dans le séquenceur (Illumina) peut se produire et les amplicons les plus abondants finiront par être présents même dans les "blancs parfaits". Si tout ce qui se trouve dans les blancs est éliminé en tant que contamination, le signal réel sera probablement éliminé également. Il existe des logiciels tels que Decontam (paquet R, <https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0605-2>) et Re-centrifuge (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6472834/>). Re-centrifuge utilise des classifications taxonomiques (liées à Kraken) et attribue un score aux lectures, qui est utilisé pour une analyse phylogénétique ultérieure. Decontam part du principe que les contaminants

seront plus fréquents dans les échantillons à faible concentration d'ADN et moins fréquents dans les échantillons à forte concentration d'ADN (inversement proportionnels). Le code ci-dessous est également une option qui permet de filtrer les caractéristiques trouvées dans les blancs.

```
# Filtering features found in blanks

nano correct_blanks_composition.sh
#!/bin/bash -e
#SBATCH --job-name=blanks
#SBATCH --output=blankscmp_%j.out
#SBATCH --error=blankscmp_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END
#SBATCH --time=4:00:00
#SBATCH --mem=50G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

module load QIIME2/2022.2

cd /nesi/nobackup/uoo03555/celia/Celia/qiime/

#How many features in blanks and how abundant are they in comparison
with all other features?
echo 'counting features in DNA blank'
qiime feature-table filter-samples \
--i-table hits-08-82-no-mtchlr-withP-noMalacostraca-silva.qza \
--m-metadata-file metadata_tsv.tsv \
--p-where "sample type='negativcontrol'" \
--o-filtered-table table-blank-DNA-silva.qza
qiime feature-table summarize \
--i-table table-blank-DNA-silva.qza \
--o-visualization table-blank-DNA-silva.qzv

echo ' taxonomic composition of DNA blanks'
qiime feature-table filter-seqs \
--i-data hits-08-82-no-mtchlr-withP-noMalacostraca-silva-rep-seqs.qza \
--i-table table-blank-DNA-silva.qza \
--o-filtered-data blank-DNA-silva-rep-seqs.qza
qiime feature-classifier classify-sklearn \
--i-classifier silva_reference_Priscila/silva-138.1-ssu-nr99-515f-806r-classifier.qza \
--i-reads blank-DNA-silva-rep-seqs.qza \
--p-n-jobs -2 \
--o-classification blank-DNA-observed-taxonomy-silva.qza
qiime metadata tabulate \
--m-input-file blank-DNA-observed-taxonomy-silva.qza \
```

```
--o-visualization blank-DNA-observed-taxonomy-silva.qzv
```

```
sbatch correct_blanks_composition.sh #job id: 36864515  
sacct
```

*#D'après les tableaux ci-dessus, quelques caractéristiques sont présentes dans plus d'une case (jusqu'à trois),  
#mais beaucoup n'étaient présentes que dans une seule case.  
# A partir de cela, nous pouvons déjà voir que l'utilisation de la prévalence dans les blancs dans le décontam laissera de nombreuses caractéristiques dans le décontam.*

*# sauvegarder à partir de la 'table-blanks-DNA-silva' la fréquence par détail de caractéristique en csv  
# éditer le tableau :  
# la première colonne est le 'FeatureID', la seconde observée 'Frequency',  
# et le convertir en TSV via notepad++ : "allBlanks\_filter.tsv"*

Vérifier la composition de TOUS les "blancs" et filtrage des contrôles négatifs

Exclut tous les contrôles négatifs de l'ensemble de données et toutes les fonctionnalités qui sont sur liste noire (noMalacostraca, les blancs, etc), supprimant ainsi toute contamination potentielle de nos échantillons réels. Le résultat est une table de caractéristiques sans contrôles négatifs et sans aucune des caractéristiques correspondant à celles du contrôle négatif.

Note : Une fois la composition terminée,

- sauvegarder à partir de la "table-blank-DNA-silva.qzv" la fréquence par détail de caractéristique en csv
- éditer le tableau :
  - #La première colonne est le 'FeatureID',
  - #la seconde observée 'Fréquence',
  - #et le convertir en TSV
- ajouter le nom de la table tsv au fichier allBlanks\_filter.sh
- exécuter allBlanks\_filter.sh

```
nano negative_control_filter.sh  
#!/bin/bash -e  
#SBATCH --job-name=blanks  
#SBATCH --output=blankscomp_%j.out  
#SBATCH --error=blankscomp_%j.err  
#SBATCH --mail-user=cece.koellsch@gmail.com  
#SBATCH --mail-type=END  
#SBATCH --qos=debug #time=4:00:00  
#SBATCH --mem=25G  
#SBATCH --ntasks 1  
#SBATCH --profile=task  
#SBATCH --account=uoo03555
```

```

module load QIIME2/2022.2

cd /nesi/nobackup/uoo03555/celia/Celia/qiime/

# First, remove the negative controls from the feature table with
# 'filter samples', using p-exclude-ids
echo 'filter blanks from gg feature table'
qiime feature-table filter-samples \
  --i-table hits-08-82-no-mtchlr-withP-noMalacostraca-silva.qza \
  --m-metadata-file metadata_tsv.tsv \
  --p-where "samplotype='negativcontrol'" \
  --p-exclude-ids \
  --o-filtered-table table-no-negatives.qza

qiime feature-table summarize \
  --i-table table-no-negatives.qza \
  --o-visualization table-no-negatives.qzv

#Example output in qzv format below
#Metric                               Sample
#Number of samples                    40 #11 blanks removed (had 51 before)
#Number of features                   6,989 # Total was 7111, this is implies that
#Total frequency                      3,509,594
#Minimum frequency                   4,107.0
#Maximum frequency                   485,134.0

# Remove the features that are identified in 'allBlanks_filter.tsv',
# with filter-features and --p-exclude-ids
echo 'filter features found in blanks from feature table'
qiime feature-table filter-features \
  --i-table table-no-negatives.qza \
  --m-metadata-file allBlanks_filter.tsv \
  --p-exclude-ids \
  --o-filtered-table filtered-table-08-82.qza

qiime feature-table summarize \
  --i-table filtered-table-08-82.qza \
  --o-visualization filtered-table-08-82.qzv

#Example of qzv output below
#Table summary
#Metric                               Sample
#Number of samples                    40 #11 blanks removed
#Number of features                   6,936 #175 features found in blanks removed
#Total frequency                      2,366,578
#Minimum frequency                   13.0

```

```
#Maximum frequency      308,508.0
#Il y a 2 échantillons avec moins de 1000 caractéristiques
(l'échantillon 9 avec 39 caractéristiques et l'échantillon 34 avec 13
caractéristiques).

sbatch negative_control_filter.sh #job id : 37025807
sacct
```

## Filtre MCS

La contamination croisée dans le MCS serait commune, qu'elle provienne du laboratoire ou d'Illumina. Toutes les caractéristiques de MCS seront supprimées, et ce pour une table de caractéristiques avec des blancs et une autre sans blancs, afin que les résultats puissent être comparés pour l'analyse de la qualité.

Remarque : une fois que MCS\_composition a fini de s'exécuter,

- sauvegarder à partir de la "filtered-table-08-82.qzv" la fréquence par détail d'élément en csv
- éditer le tableau :
  - #La première colonne est le "FeatureID",
  - #la seconde observée 'Fréquence',
  - #et le convertir en TSV
- ajouter le nom de la table tsv au fichier MCS\_filter.sh
- exécuter MCS\_filter.sh

```
nano MCS_composition.sh
#!/bin/bash -e
#SBATCH --job-name=blanks
#SBATCH --output=blankscomp_%j.out
#SBATCH --error=blankscomp_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END
#SBATCH --qos=debug #time=4:00:00
#SBATCH --mem=25G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

module load QIIME2/2022.2

cd /nesi/nobackup/uoo03555/celia/Celia/qiime/

# Tout d'abord, supprimez les contrôles négatifs de la table des
caractéristiques avec "filter samples", en utilisant p-exclude-ids
echo 'filter blanks from gg feature table'
qiime feature-table filter-samples \
```

```

--i-table hits-08-82-no-mtchlor-withP-noMalacostraca-silva.qza \
--m-metadata-file metadata.tsv.tsv \
--p-where "samplotype='negativcontrol'" \
--p-exclude-ids \
--o-filtered-table table-no-negatives.qza

qiime feature-table summarize \
--i-table table-no-negatives.qza \
--o-visualization table-no-negatives.qzv

#Example output in qzv format below
#Metric      Sample
#Number of samples      40 #11 blanks removed (had 51 before)
#Number of features      6,989 # Total was 7111, this implies that
from the 175 features found in blanks, 53 are shared with non-bl$
#Total frequency      3,509,594
#Minimum frequency      4,107.0
#Maximum frequency      485,134.0

# Remove the features that are identified in 'allBlanks_filter.tsv',
with filter-features and --p-exclude-ids
echo 'filter features found in blanks from feature table'
qiime feature-table filter-features \
--i-table table-no-negatives.qza \
--m-metadata-file allBlanks_filter.tsv \
--p-exclude-ids \
--o-filtered-table filtered-table-08-82.qza

qiime feature-table summarize \
--i-table filtered-table-08-82.qza \
--o-visualization filtered-table-08-82.qzv

#Example of qzv output below
#Table summary
#Metric      Sample
#Number of samples      40 #11 blanks removed
#Number of features      6,936 #175 features found in blanks removed
#Total frequency      2,366,578
#Minimum frequency      13.0
#Maximum frequency      308,508.0

#Il y a 2 échantillons avec moins de 1000 caractéristiques
(l'échantillon 9 avec 39 caractéristiques et l'échantillon 34 avec 13
caractéristiques).
# La fréquence moyenne dans le tableau avec blancs est légèrement plus
élevée que dans le tableau sans blancs (les caractéristiques qui
étaient également présentes dans le MCS ont été supprimées avec les
blancs).
# Il y a 11 caractéristiques dans la MCS (pas de blancs), 9 trouvées
dans les deux échantillons MCS, 1 trouvée seulement dans une MCS (8

```



```
caractéristiques attendues).  
sbatch MCS_composition.sh #job_id: 37063612  
sacct
```

Le MCS (Mock Community Standard) est un tube d'ADN bactérien que nous achetons chez Zymo. Il se fait avec une proportion connue de chaque bactérie (il a aussi deux champignons, mais cela n'a pas d'importance pour nous car nous n'avons que des données d'ARNr 16S, qui ne détectent que les bactéries - voir tableau 1 de ce PDF

[https://files.zymoresearch.com/protocols/\\_d6300\\_zymobiomics\\_microbial\\_community\\_standard.pdf](https://files.zymoresearch.com/protocols/_d6300_zymobiomics_microbial_community_standard.pdf)).

Nous avons inclus l'ADN MCS dans nos PCR (contrôle positif) et nous l'avons séquencé. Nous avons donc fait au MCS tout ce que nous avons fait à nos échantillons. Maintenant, nous voulons savoir dans quelle mesure nous récupérons l'ADN de ces bactéries que Zymo y a ajoutées en premier lieu (nous voulons savoir dans quelle mesure la composition observée de nos contrôles positifs correspond à la composition attendue).

Pour ce faire, nous devons créer un tableau avec le nom et l'abondance des espèces de bactéries dans nos échantillons MCS (ces données sont fournies par Zymo, nous créons simplement un tableau avec). Ce tableau doit avoir un format spécifique pour que Qiime le comprenne. La première chose à faire est de lui donner l'ID de nos bibliothèques MCS (#OTU ID). J'ai examiné votre fichier metadata\_tsv.tsv et j'ai vu que vos échantillons MCS portent les ID 8254-130-00-01 et 8254-148-00-01. La deuxième chose est de lui donner une liste des huit espèces de bactéries dans le MCS que nous achetons à zymo (*B. subtilis*, *E. faecalis*, *E. coli*, *L. fermentum*, *L. monocytogenes*, *P. aeruginosa*, *S. enterica* et *S. aureus*) - mais nous le faisons en utilisant les notations Silva (d\_, p\_, c\_, o\_, f\_, g\_, s\_). Les quantités de chacun sont standard de Zymo (les valeurs du tableau 1 du PDF ci-dessus, mais divisées par 100, car nous transformons les pourcentages en proportions), et devraient être les mêmes dans MCS1 et MCS2.

## Utilisation des communautés modèles de Zymo Biomics pour estimer le biais d'extraction et de PCR

Les communautés fictives ne permettent jamais d'obtenir une correspondance à 100 % entre les compositions attendues et observées. Le biais d'extraction de l'ADN, le biais d'amorce, le biais d'amplification/séquençage, la variation du nombre de copies, sont autant de causes de déviations qui échappent à tout contrôle une fois que la communauté est séquencée.

Les standards de communautés microbiennes et les standards d'ADN de ZymoBIOMICS contiennent 12% de chaque composant bactérien. La composition attendue est la suivante

- *Pseudomonas aeruginosa* (4,2 %)
- *Escherichia coli* (10,1 %)
- *Salmonella enterica* (10,4 %)
- *Lactobacillus fermentum* (18,4 %)

- Enterococcus faecalis (9,9 %)
- Staphylococcus aureus (15,5 %)
- Listeria monocytogenes (14,1%)
- Bacillus subtilis (17,4 %)

```
mkdir MCS
```

```
cp table-no-negatives.qza MCS
cp filtered-table-08-82.qza MCS
cp rep-seqs-82.qza MCS
```

```
cd MCS
```

```
#####
```

```
#Créer le tableau prévu soit avec nano soit avec une application  
externe comme notepad++
```

```
cd /nesi/nobackup/uoo03555/celia/Celia/qiime/MCS
```

```
nano MCS-expected-table.tsv
```

```
#ID OTU 8254-130-00-01 8254-148-00-01
```

```
d__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Bacillaceae;g__B  
acillus;s__subtilis 0.174 0.174
```

```
d__Bacteria;p__Firmicutes;c__Bacilloos;o__Lactobacillales;f__Enterococ  
caceae;g__Enterococcus;s__faecalis 0.099 0.099
```

```
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteria  
les;f__Enterobacteriaceae;g__Escherichia;s__coli 0.101 0.101
```

```
d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacill  
aceae;g__Lactobacillus;s__fermentum 0.184 0.184
```

```
d__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Listeriaceae;g__  
Listeria;s__monocytogenes 0.141 0.141
```

```
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadale  
s;f__Pseudomonadaceae;g__Pseudomonas;s__aeruginosa 0.042
```

```
0.042
```

```
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteria  
les;f__Enterobacteriaceae;g__Salmonella;s__enterica 0.104
```

```
0.104
```

```
d__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Staphylococcacea  
e;g__Staphylococcus;s__aureus 0.155 0.155
```

```
#là où il y a des es
```

```
# convertir la table TSV OTU en un fichier .biom
```

```
# à faire en dehors de nano, le faire dans le terminal
```

```
module load QIIME2/2022.2
```

```
biom convert -i MCS-expected-table.tsv -o MCS-expected-table_hdf5.biom  
--table-type="OTU table" --to-hdf5
```

```
#importer le fichier .biom dans FeatureTable[RelativeFrequency]  
(format qza) dans qiime2
```

```
qiime tools import \
```

```

--input-path MCS-expected-table_hdf5.biom \
--type 'FeatureTable[RelativeFrequency]' \
--input-format BIOMV210Format \
--output-path MCS-expected-table.qza

#####
nano observedMCS.sh
#!/bin/bash -e
#SBATCH --job-name=Obs_MCS_noBlanks
#SBATCH --output=observedMCS_%j.out
#SBATCH --error=observedMCS_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END
#SBATCH --time=4:00:00
#SBATCH --mem=50G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

module load QIIME2/2022.2

cd /nesi/nobackup/uoo03555/celia/Celia/qiime/MCS

echo "Extract only the MCS representative sequences from the rep-seqs
table"
qiime feature-table filter-seqs \
--i-data rep-seqs-82.qza \
--i-table table-no-negatives.qza \
--o-filtered-data MCS-noBlanks-rep-seqs-silva.qza

qiime feature-table filter-seqs \
--i-data rep-seqs-82.qza \
--i-table filtered-table-08-82.qza \
--o-filtered-data MCS-withBlanks-rep-seqs-silva.qza

echo "running taxonomy classifier, silva"
qiime feature-classifier classify-sklearn \
--i-classifier
/nesi/nobackup/uoo03555/celia/Celia/qiime/silva_reference_Priscila/
silva-138.1-ssu-nr99-515f-806r-classifier.qza \
--i-reads MCS-noBlanks-rep-seqs-silva.qza \
--o-classification MCS-noBlanks-taxonomy-silva.qza
qiime metadata tabulate \
--m-input-file MCS-noBlanks-taxonomy-silva.qza \
--o-visualization MCS-noBlanks-taxonomy-silva.qzv

qiime feature-classifier classify-sklearn \
--i-classifier
/nesi/nobackup/uoo03555/celia/Celia/qiime/silva_reference_Priscila/
silva-138.1-ssu-nr99-515f-806r-classifier.qza \

```

```

--i-reads MCS-withBlanks-rep-seqs-silva.qza \
--o-classification MCS-withBlanks-taxonomy-silva.qza
qiime metadata tabulate \
--m-input-file MCS-withBlanks-taxonomy-silva.qza \
--o-visualization MCS-withBlanks-taxonomy-silva.qzv

echo "generate interactive taxonomy bar plots"
qiime taxa barplot \
--i-table table-no-negatives.qza \
--i-taxonomy MCS-noBlanks-taxonomy-silva.qza \
--m-metadata-file ../metadata_tsv.tsv \
--o-visualization MCS-noBlanks-taxa-bar-plots-silva.qzv

qiime taxa barplot \
--i-table filtered-table-08-82.qza \
--i-taxonomy MCS-withBlanks-taxonomy-silva.qza \
--m-metadata-file ../metadata_tsv.tsv \
--o-visualization MCS-withBlanks-taxa-bar-plots-silva.qzv

echo "generate a tree for phylogenetic diversity analysis"
qiime phylogeny align-to-tree-mafft-fasttree \
--i-sequences MCS-noBlanks-rep-seqs-silva.qza \
--o-alignment MCS-noBlanks-aligned-silva.qza \
--o-masked-alignment MCS-noBlanks-masked-aligned-silva.qza \
--o-tree MCS-noBlanks-unrooted-tree.qza \
--o-rooted-tree MCS-noBlanks-rooted-tree.qza

qiime phylogeny align-to-tree-mafft-fasttree \
--i-sequences MCS-withBlanks-rep-seqs-silva.qza \
--o-alignment MCS-withBlanks-aligned-silva.qza \
--o-masked-alignment MCS-withBlanks-masked-aligned-silva.qza \
--o-tree MCS-withBlanks-unrooted-tree.qza \
--o-rooted-tree MCS-withBlanks-rooted-tree.qza

echo "explore alpha diversity as a function of sampling depth with an
alpha rarefaction plot"
qiime diversity alpha-rarefaction \
--i-table table-no-negatives.qza \
--i-phylogeny MCS-noBlanks-rooted-tree.qza \
--p-max-depth 4000 \
--m-metadata-file ../metadata_tsv.tsv \
--o-visualization MCS-noBlanks-alpha-rarefaction.qzv

qiime diversity alpha-rarefaction \
--i-table filtered-table-08-82.qza \
--i-phylogeny MCS-withBlanks-rooted-tree.qza \
--p-max-depth 4000 \
--m-metadata-file ../metadata_tsv.tsv \
--o-visualization MCS-withBlanks-alpha-rarefaction.qzv

```

```

echo "collapse the taxonomy to species"
qiime taxa collapse \
  --i-table table-no-negatives.qza \
  --i-taxonomy MCS-noBlanks-taxonomy-silva.qza \
  --p-level 7 \
  --o-collapsed-table species-table-MCS-noBlanks-silva.qza

qiime taxa collapse \
  --i-table filtered-table-08-82.qza \
  --i-taxonomy MCS-withBlanks-taxonomy-silva.qza \
  --p-level 7 \
  --o-collapsed-table species-table-MCS-withBlanks-silva.qza

echo "convert this new frequency table to relative-frequency" #THIS
COMMAND
qiime feature-table relative-frequency \
  --i-table species-table-MCS-noBlanks-silva.qza \
  --o-relative-frequency-table observed-species-table-MCS-noBlanks-
silva.qza
qiime metadata tabulate \
  --m-input-file observed-species-table-MCS-noBlanks-silva.qza \
  --o-visualization observed-species-table-MCS-noBlanks-silva.qzv

qiime feature-table relative-frequency \
  --i-table species-table-MCS-withBlanks-silva.qza \
  --o-relative-frequency-table observed-species-table-MCS-withBlanks-
silva.qza
qiime metadata tabulate \
  --m-input-file observed-species-table-MCS-withBlanks-silva.qza \
  --o-visualization observed-species-table-MCS-withBlanks-silva.qzv

sbatch observedMCS.sh #job id : 37482258
sacct
#####

nano qcMCS.sh
#!/bin/bash -e
#SBATCH --job-name=MCS_qc
#SBATCH --output=qcMCS_%j.out
#SBATCH --error=qcMCS_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END
#SBATCH --time=1:00:00
#SBATCH --mem=25G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

module load QIIME2/2022.2

```

```

cd /nesi/nobackup/uoo03555/celia/Celia/qiime/MCS

qiime quality-control evaluate-composition \
  --i-expected-features MCS-expected-table.qza \
  --i-observed-features observed-species-table-MCS-noBlanks-silva.qza \
  --o-visualization qc-MCS-comparison-noBlanks.qzv

qiime quality-control evaluate-composition \
  --i-expected-features MCS-expected-table.qza \
  --i-observed-features observed-species-table-MCS-withBlanks-silva.qza \
  --o-visualization qc-MCS-comparison-withBlanks.qzv

qiime quality-control evaluate-seqs \
  --i-query-sequences MCS-noBlanks-rep-seqs-silva.qza \
  --i-reference-sequences reference-seqs-expected.qza \ # on essaie de  

changer pour "rep-seqs-82.qza"  

  --o-visualization eval-seqs-noBlanks.qzv
#télécharger le qzv, le visualiser dans view.qiime et télécharger le  

fichier tsv  

#table-MCS-noBlanks-silva.tsv  

#penser à enlever la première colonne si on a que des chiffres et à  

mettre seulement ID au lieu de Query ID

qiime quality-control evaluate-seqs \
  --i-query-sequences MCS-withBlanks-rep-seqs-silva.qza \
  --i-reference-sequences reference-seqs-expected.qza \ # on essaie de  

changer pour "rep-seqs-82.qza" ; job id : 37511206  

  --o-visualization eval-seqs-withBlanks.qzv
#télécharger le qzv et le convertir aussi en tsv : table-MCS-  

withBlanks-silva.tsv  

#penser à enlever la première colonne si on a que des chiffres

sbatch qcMCS.sh # job id : 37510807 # pour rep-seqs-82.qza, job id :  

37511206  

sacct

# on recoit l'erreur suivante :  

# Invalid value for '--i-reference-sequences': 'reference-seqs-  

expected.qza' is not a valid filepath  

## ca fonctionne avec "rep-seqs-82.qza"

nano filter_MCS.sh
#!/bin/bash -e
#SBATCH --job-name=MCS
#SBATCH --output=MCSfilt_%j.out
#SBATCH --error=MCSfilt_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END

```

```

#SBATCH --time=00:20:00
#SBATCH --mem=10G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

module load QIIME2/2022.2

cd /nesi/nobackup/uoo03555/celia/Celia/qiime

## Filtrer les caractéristiques trouvées dans le "MCS" au cas où il y
aurait des contaminations croisées non trouvées dans les blancs
# Tout d'abord, supprimez la MCS du tableau des caractéristiques
echo 'supprimer les caractéristiques MCS du tableau des
caractéristiques sans blanc, taxonomie de Silva'
qiime feature-table filter-samples \
  --i-table filtered-table-08-82.qza \
  --m-metadata-file metadata_tsv.tsv \
  --p-where "AMPID='MCS'" \
  --p-exclude-ids \
  --o-filtered-table table-no-MCS-silva.qza
qiime feature-table summarize \
  --i-table table-no-MCS-silva.qza \
  --o-visualization table-no-MCS-silva.qzv

echo 'Filtrer les caractéristiques MCS, taxonomie Silva'
qiime feature-table filter-features \
  --i-table table-no-MCS-silva.qza \
  --m-metadata-file table-MCS-noBlanks-silva.tsv \
  --p-exclude-ids \sa
  --o-filtered-table clean-filtered-table-silva.qza
qiime feature-table summarize \
  --i-table clean-filtered-table-silva.qza \
  --o-visualization clean-filtered-table-silva.qzv

echo 'filtrer les séquences MCS, taxonomie de Silva'
qiime feature-table filter-seqs \
  --i-data hits-08-82-no-mtchlr-withP-noMalacostraca-silva-rep-
seqs.qza \
  --i-table clean-filtered-table-silva.qza \
  --o-filtered-data clean-filtered-rep-seqs-silva.qza
qiime feature-table tabulate-seqs \
  --i-data clean-filtered-rep-seqs-silva.qza \
  --o-visualization clean-filtered-rep-seqs-silva.qzv

sbatch filter_MCS.sh # job id : 37512254
sacct

```

# Vérifier la profondeur du séquençage et la diversité taxonomique des échantillons

```
nano check-depth-taxa.sh
```

```
#!/bin/bash -e
#SBATCH --job-name=depth-tax
#SBATCH --output=depth-taxa_%j.out
#SBATCH --error=depth-taxa_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END
#SBATCH --time=4:00:00
#SBATCH --mem=50G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

module load QIIME2/2022.2

cd /nesi/nobackup/uoo03555/celia/Celia/qiime

## déficiences du séquençage > --p-min-features & --p-min-frequency
# vérifier les échantillons pour la profondeur de séquençage et la
# diversité taxonomique ressemble à
echo 'seq depth and taxonomic diversity with silva'
qiime feature-classifier classify-sklearn \
  --i-classifier silva_reference_Priscila/silva-138.1-ssu-nr99-515f-
806r-classifier.qza \
  --i-reads clean-filtered-rep-seqs-silva.qza \
  --p-n-jobs 2 \
  --o-classification clean-filtered-observed-taxonomy-silva.qza
qiime metadata tabulate \
  --m-input-file clean-filtered-observed-taxonomy-silva.qza \
  --o-visualization clean-filtered-observed-taxonomy-silva.qzv

echo 'bar plots of taxonomic composition based on silva'
qiime taxa barplot \
  --i-table clean-filtered-table-silva.qza \
  --i-taxonomy clean-filtered-observed-taxonomy-silva.qza \
  --m-metadata-file metadata_tsv.tsv \
  --o-visualization clean-filtered-taxa-bar-plots-silva.qzv

echo 'alignment for rarefaction curve, silva'
qiime phylogeny align-to-tree-mafft-fasttree \
  --i-sequences clean-filtered-rep-seqs-silva.qza \
  --o-alignment clean-filtered-aligned-rep-seqs-silva.qza \
  --o-masked-alignment clean-filtered-masked-aligned-rep-seqs-
silva.qza \
  --o-tree clean-filtered-unrooted-tree-silva.qza \
```



```

--o-rooted-tree clean-filtered-rooted-tree-silva.qza

echo 'rarefaction curve with silva'
qiime diversity alpha-rarefaction \
  --i-table clean-filtered-table-silva.qza \
  --i-phylogeny clean-filtered-rooted-tree-silva.qza \
  --p-max-depth 4000 \
  --m-metadata-file metadata_tsv.tsv \
  --o-visualization clean-filtered-alpha-rarefaction-test-silva.qzv

sbatch check-depth-taxa.sh #job id: 37512337
sacct

```

## Filtrer la profondeur, générer une classification taxonomique filtrée et des diagrammes à barres

#regarder la fréquence minimale et le nombre d'éléments dans les tableaux pour décider d'un seuil de profondeur pour les étapes ci-dessous clean-filtered-table-silva.qzv --> min freq 5, 34 échantillons en dessous de 500, 46 échantillons en dessous de 1000. La fraction rare s'arrête avant 500, à moins qu'il ne s'agisse d'un échantillon environnemental. Le nombre minimal de caractéristiques est de 1, le filtrage se fera avec un minimum de 8.

- Essayez de filtrer --feature min frequency basé sur la fréquence minimale de la table de fréquence par caractéristique du clean-filtered-table-silva.qzv (min freq de 2.0), pour laisser tomber moins de caractéristiques.

```

## supprimer les échantillons sur la base de la profondeur de
séquençage suggérant des déficiences de séquençage possibles
# --p-min-features 8 (identique à celui utilisé dans l'étude
précédente, pas sûr de savoir comment choisir ce seuil)
# --p-min-frequency 500

```

```
nano filter-depth-taxa.sh
```

```

#!/bin/bash -e
#SBATCH --job-name=depth-tax
#SBATCH --output=filter-depth-taxa_%j.out
#SBATCH --error=filter-depth-taxa_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END
#SBATCH --time=4:00:00
#SBATCH --mem=50G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

```

```
module load QIIME2/2022.2
```

```

cd /nesi/nobackup/uoo03555/celia/Celia/qiime

echo 'profondeur de filtrage, silva'
qiime feature-table filter-samples \
  --i-table clean-filtered-table-silva.qza \
  --p-min-features 8 \
  --p-min-frequency 1000 \
  --o-filtered-table clean-filtered-table-8-1000-silva.qza
qiime feature-table summarize \
  --i-table clean-filtered-table-8-1000-silva.qza \
  --o-visualization clean-filtered-table-8-1000-silva.qzv

echo 'tableau rep seqs, profondeur filtrée, silva'
qiime feature-table filter-seqs \
  --i-data clean-filtered-rep-seqs-silva.qza \
  --i-table clean-filtered-table-8-1000-silva.qza \
  --o-filtered-data clean-filtered-table-8-1000-rep-seqs-silva.qza
qiime feature-table tabulate-seqs \
  --i-data clean-filtered-table-8-1000-rep-seqs-silva.qza \
  --o-visualization clean-filtered-table-8-1000-rep-seqs-silva.qzv

echo 'verifier la taxonomie, silva'
qiime feature-classifier classify-sklearn \
  --i-classifier silva_reference_Priscila/silva-138.1-ssu-nr99-515f-806r-classifier.qza \
  --i-reads clean-filtered-table-8-1000-rep-seqs-silva.qza \
  --p-n-jobs 2 \
  --o-classification clean-filtered-8-1000-observed-taxonomy-silva.qza

echo 'bar plots de la composition taxonomique, silva'
qiime taxa barplot \
  --i-table clean-filtered-table-8-1000-silva.qza \
  --i-taxonomy clean-filtered-8-1000-observed-taxonomy-silva.qza \
  --m-metadata-file metadata_tsv.tsv \
  --o-visualization clean-filtered-8-1000-taxa-bar-plots-silva.qzv

sbatch filter-depth-taxa.sh #job id : 37526474
sacct

```

Analyses de diversité

```

nano alpha-beta-diversity.sh

#!/bin/bash -e
#SBATCH --job-name=alpha-prep
#SBATCH --output=diversity_%j.out
#SBATCH --error=diversity_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END

```

```

#SBATCH --time=04:00:00
#SBATCH --mem=50G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

module load QIIME2/2022.2

cd /nesi/nobackup/uoo03555/celia/Celia/qiime

echo 'alignment for rarefaction curve, silva'
qiime phylogeny align-to-tree-mafft-fasttree \
  --i-sequences clean-filtered-table-8-1000-rep-seqs-silva.qza \
  --o-alignment clean-filtered-table-8-1000-aligned-rep-seqs-silva.qza \
  --o-masked-alignment clean-filtered-table-8-1000-masked-aligned-rep-seqs-silva.qza \
  --o-tree clean-filtered-table-8-1000-unrooted-tree-silva.qza \
  --o-rooted-tree clean-filtered-table-8-1000-rooted-tree-silva.qza

echo 'rarefaction curve with silva'
qiime diversity alpha-rarefaction \
  --i-table clean-filtered-table-8-1000-silva.qza \
  --i-phylogeny clean-filtered-table-8-1000-rooted-tree-silva.qza \
  --p-max-depth 4000 \
  --m-metadata-file metadata_tsv.tsv \
  --o-visualization clean-filtered-8-1000-alpha-rarefaction-silva.qzv

echo 'core metrics, silva'
qiime diversity core-metrics-phylogenetic \
  --i-phylogeny clean-filtered-table-8-1000-rooted-tree-silva.qza \
  --i-table clean-filtered-table-8-1000-silva.qza \
  --p-sampling-depth 1000 \
  --m-metadata-file metadata_tsv.tsv \
  --output-dir 8-1000_main-q2f-core-metrics-results-silva

echo "#####ALPHA DIVERSITY#####"

echo 'alpha diversity: Is there differences in alpha diversity between treatments?'
echo 'alpha-group-significance, faith_pd_vector, silva'
qiime diversity alpha-group-significance \
  --i-alpha-diversity 8-1000_main-q2f-core-metrics-results-silva/faith_pd_vector.qza \
  --m-metadata-file metadata_tsv.tsv \
  --o-visualization 8-1000_main-q2f-core-metrics-results-silva/faith-pd-group-significance.qzv

echo 'evenness-group-significance, silva'

```

```

qiime diversity alpha-group-significance \
  --i-alpha-diversity 8-1000_main-q2f-core-metrics-results-
silva/evenness_vector.qza \
  --m-metadata-file metadata_tsv.tsv \
  --o-visualization 8-1000_main-q2f-core-metrics-results-
silva/evenness-group-significance.qzv

echo 'shannon-group-significance, silva'
qiime diversity alpha-group-significance \
  --i-alpha-diversity 8-1000_main-q2f-core-metrics-results-
silva/shannon_vector.qza \
  --m-metadata-file metadata_tsv.tsv \
  --o-visualization 8-1000_main-q2f-core-metrics-results-
silva/shannon-group-significance.qzv

echo "#####BETA DIVERSITY#####"

#Sampletype
echo 'Sample type unweighted-unifrac b-div, silva'
qiime diversity beta-group-significance \
  --i-distance-matrix 8-1000_main-q2f-core-metrics-results-
silva/unweighted_unifrac_distance_matrix.qza \
  --m-metadata-file metadata_tsv.tsv \
  --m-metadata-column sampletype \
  --o-visualization 8-1000_main-q2f-core-metrics-results-
silva/unweighted-SampleOrigin-significance.qzv \
  --p-pairwise \
  --p-permutations 9999

echo 'Sample type wheighted-unifrac b-div, silva'
qiime diversity beta-group-significance \
  --i-distance-matrix 8-1000_main-q2f-core-metrics-results-
silva/weighted_unifrac_distance_matrix.qza \
  --m-metadata-file metadata_tsv.tsv \
  --m-metadata-column sampletype \
  --o-visualization 8-1000_main-q2f-core-metrics-results-
silva/weighted-SampleOrigin-significance.qzv \
  --p-pairwise \
  --p-permutations 9999

sbatch alpha-beta-diversity.sh #job id: 37536714
sacct

```

## Profondeur de filtrage

avec --p-min-features 2 (fréquence minimale des caractéristiques dans le clean-filtered-table-silva.qzv) et --p-min-frequency 500

```
## supprimer les échantillons sur la base de la profondeur de
séquençage suggérant des déficiences possibles de séquençage
# --p-min-features 2 (fréquence minimale des caractéristiques dans le
tableau)
# --p-min-frequency 500
```

```
nano filter-2-500depth-taxa.sh
```

```
#!/bin/bash -e
#SBATCH --job-name=depth-tax
#SBATCH --output=filter-depth-taxa_%j.out
#SBATCH --error=filter-depth-taxa_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END
#SBATCH --time=4:00:00
#SBATCH --mem=50G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555
```

```
module load QIIME2/2022.2
```

```
cd /nesi/nobackup/uoo03555/celia/Celia/qiime
```

```
echo 'filtering depth, silva'
qiime feature-table filter-samples \
  --i-table clean-filtered-table-silva.qza \
  --p-min-features 2 \
  --p-min-frequency 500 \
  --o-filtered-table clean-filtered-table-2-500-silva.qza
qiime feature-table summarize \
  --i-table clean-filtered-table-2-500-silva.qza \
  --o-visualization clean-filtered-table-2-500-silva.qzv
```

```
echo 'rep seqs table, depth filtered, silva'
qiime feature-table filter-seqs \
  --i-data clean-filtered-rep-seqs-silva.qza \
  --i-table clean-filtered-table-2-500-silva.qza \
  --o-filtered-data clean-filtered-table-2-500-rep-seqs-silva.qza
qiime feature-table tabulate-seqs \
  --i-data clean-filtered-table-2-500-rep-seqs-silva.qza \
  --o-visualization clean-filtered-table-2-500-rep-seqs-silva.qzv
```

```
echo 'check taxonomy, silva'
qiime feature-classifier classify-sklearn \
  --i-classifier silva_reference_Priscila/silva-138.1-ssu-nr99-515f-
806r-classifier.qza \
  --i-reads clean-filtered-table-2-500-rep-seqs-silva.qza \
  --p-n-jobs 2 \
  --o-classification clean-filtered-2-500-observed-taxonomy-silva.qza
```

```

echo 'bar plots of taxonomic composition, silva'
qiime taxa barplot \
  --i-table clean-filtered-table-2-500-silva.qza \
  --i-taxonomy clean-filtered-2-500-observed-taxonomy-silva.qza \
  --m-metadata-file metadata_tsv.tsv \
  --o-visualization clean-filtered-2-500-taxa-bar-plots-silva.qzv

sbatch filter-2-500depth-taxa.sh #job id: 37545886
sacct

```

## Analyses de la diversité avec un jeu de données filtrées de 2 à 500 de profondeur

```

nano 2-500alpha-beta-diversity.sh

#!/bin/bash -e
#SBATCH --job-name=alpha-div
#SBATCH --output=diversity_%j.out
#SBATCH --error=diversity_%j.err
#SBATCH --mail-user=cece.koellsch@gmail.com
#SBATCH --mail-type=END
#SBATCH --time=04:00:00
#SBATCH --mem=50G
#SBATCH --ntasks 1
#SBATCH --profile=task
#SBATCH --account=uoo03555

module load QIIME2/2022.2

cd /nesi/nobackup/uoo03555/celia/Celia/qiime

echo 'alignment for rarefaction curve, silva'
qiime phylogeny align-to-tree-mafft-fasttree \
  --i-sequences clean-filtered-table-2-500-rep-seqs-silva.qza \
  --o-alignment clean-filtered-table-2-500-aligned-rep-seqs-
silva.qza \
  --o-masked-alignment clean-filtered-table-2-500-masked-aligned-rep-
seqs-silva.qza \
  --o-tree clean-filtered-table-2-500-unrooted-tree-silva.qza \
  --o-rooted-tree clean-filtered-table-2-500-rooted-tree-silva.qza

echo 'rarefaction curve with silva'
qiime diversity alpha-rarefaction \
  --i-table clean-filtered-table-2-500-silva.qza \
  --i-phylogeny clean-filtered-table-2-500-rooted-tree-silva.qza \
  --p-max-depth 4000 \

```

```

--m-metadata-file metadata_tsv.tsv \
--o-visualization clean-filtered-2-500-alpha-rarefaction-silva.qzv

echo 'core metrics, silva'
qiime diversity core-metrics-phylogenetic \
  --i-phylogeny clean-filtered-table-2-500-rooted-tree-silva.qza \
  --i-table clean-filtered-table-2-500-silva.qza \
  --p-sampling-depth 500 \
  --m-metadata-file metadata_tsv.tsv \
  --output-dir 2-500main-bis

echo "##### _____ ALPHA DIVERSITY _____ #####"

echo 'alpha diversity: Is there differences in alpha diversity between
treatments?'
echo 'alpha-group-significance, faith_pd_vector, silva'
qiime diversity alpha-group-significance \
  --i-alpha-diversity 2-500main-bis/faith_pd_vector.qza \
  --m-metadata-file metadata_tsv.tsv \
  --o-visualization 2-500main-bis/faith-pd-group-significance.qzv

echo 'evenness-group-significance, silva'
qiime diversity alpha-group-significance \
  --i-alpha-diversity 2-500mainbis/evenness_vector.qza \
  --m-metadata-file metadata_tsv.tsv \
  --o-visualization 2-500main-bis/evenness-group-significance.qzv

echo 'shannon-group-significance, silva'
qiime diversity alpha-group-significance \
  --i-alpha-diversity 2-500main-bis/shannon_vector.qza \
  --m-metadata-file metadata_tsv.tsv \
  --o-visualization 2-500main-bis/shannon-group-significance.qzv

echo "##### _____ BETA DIVERSITY _____ #####"

#sample type
echo 'sample type unweighted-unifrac b-div, silva'
qiime diversity beta-group-significance \
  --i-distance-matrix 2-500main-
bis/unweighted_unifrac_distance_matrix.qza \
  --m-metadata-file metadata_tsv.tsv \
  --m-metadata-column sampletype \
  --o-visualization 2-500main-bis/unweighted-SampleOrigin-
significance.qzv \
  --p-pairwise \
  --p-permutations 9999

echo 'sample type wheighted-unifrac b-div, silva'

```

```
qiime diversity beta-group-significance \
  --i-distance-matrix 2-500main-
bis/weighted_unifrac_distance_matrix.qza \
  --m-metadata-file metadata_tsv.tsv \
  --m-metadata-column samplotype \
  --o-visualization 2-500main-bis/weighted-SampleOrigin-
significance.qzv \
  --p-pairwise \
  --p-permutations 9999

sbatch 2-500alpha-beta-diversity.sh #job id: 37559019
sacct
```