**Code employed in the analysis.**

**NG-Tax code.**

1. **Build libraries**

# Download sample files for ENA repository (accession number PRJEB11702, decompress them and concatenate them to build the different libraries.

cat \*l01\_1.fastq > lib01\_1.fastq

cat \*l01\_2.fastq > lib01\_2.fastq

cat \*l02\_1.fastq > lib02\_1.fastq

cat \*l02\_2.fastq > lib02\_2.fastq

cat \*l03\_1.fastq > lib03\_1.fastq

cat \*l03\_2.fastq > lib03\_2.fastq

cat \*l04\_1.fastq > lib04\_1.fastq

cat \*l04\_2.fastq > lib04\_2.fastq

cat \*l05\_1.fastq > lib05\_1.fastq

cat \*l05\_2.fastq > lib05\_2.fastq

cat \*l06\_1.fastq > lib06\_1.fastq

cat \*l06\_2.fastq > lib06\_2.fastq

cat \*l07\_1.fastq > lib07\_1.fastq

cat \*l07\_2.fastq > lib07\_2.fastq

**2. Create customized 16S databases**

**2.1 Customized V5-V6 database**

customized\_database\_generator -d Silva\_111\_full\_unique.fasta -t SILVA\_db/tax\_slv\_ssu\_128.txt -x Silva\_taxonomic\_table\_v56 -k [AG]GGATTAGATACCC -p CGAC[AG][AG]CCATGCA[ACGT]CACCT -q AGGTG[ACGT]TGCATGG[TC][TC]GTCG -f primer\_BSF784\_76\_nt\_1mm\_db -r primer\_R1064\_71\_nt\_1mm\_db -o 76 -e 71 -y primer\_BSF784\_1mm -z primer\_R1064\_1mm

**2.2 Customized V4 database**

customized\_database\_generator -d Silva\_111\_full\_unique.fasta -t SILVA\_db/tax\_slv\_ssu\_128.txt -x Silva\_taxonomic\_table\_v4 -k GTGCCAGC[AC]GCCGCGGTAA -p GGACTAC[ACT][ACG]GGGT[AT]TCTAAT -q ATTAGA[AT]ACCC[TCG][ATG]GTAGTCC -f primer\_F515\_71\_nt\_1mm\_db -r primer\_R806\_70\_nt\_1mm\_db -o 71 -e 70 -y primer\_F515\_1mm -z primer\_R806\_1mm

**3. Filtering libraries 01 to 07.**

library\_filtering -a lib01\_1.fastq -b lib01\_2.fastq -p Mock\_communities -n 01 -f [AG]GGATTAGATACCC -r CGAC[AG][AG]CCATGCA[ACGT]CACCT -l 8

library\_filtering -a lib02\_1.fastq -b lib02\_2.fastq -p Mock\_communities -n 02 -f GTGCCAGC[AC]GCCGCGGTAA -r GGACTAC[ACT][ACG]GGGT[AT]TCTAAT -l 8

library\_filtering.sh -a lib03\_1.fastq -b lib03\_2.fastq -p Mock\_communities -n 03 -f GTGCCAGC[AC]GCCGCGGTAA -r GGACTAC[ACT][ACG]GGGT[AT]TCTAAT -l 8

library\_filtering.sh -a lib04\_1.fastq -b lib04\_2.fastq -p Mock\_communities -n 04 -f GTGCCAGC[AC]GCCGCGGTAA -r GGACTAC[ACT][ACG]GGGT[AT]TCTAAT -l 8

library\_filtering.sh -a lib05\_1.fastq -b lib05\_2.fastq -p Mock\_communities -n 05 -f GTGCCAGC[AC]GCCGCGGTAA -r GGACTAC[ACT][ACG]GGGT[AT]TCTAAT -l 8

library\_filtering.sh -a lib06\_1.fastq -b lib06\_2.fastq -p Mock\_communities -n 06 -f GTGCCAGC[AC]GCCGCGGTAA -r GGACTAC[ACT][ACG]GGGT[AT]TCTAAT -l 8

library\_filtering.sh -a lib07\_1.fastq -b lib07\_2.fastq -p Mock\_communities -n 07 -f GTGCCAGC[AC]GCCGCGGTAA -r GGACTAC[ACT][ACG]GGGT[AT]TCTAAT -l 8

**4. OTU picking**

**4.1 OTU picking V5V6 samples.**

otu\_picking\_pair\_end\_read.sh -m Mock\_communities\_V5V6.txt -p Mock\_communities -a 0.1 -c 0.985 -f primer\_BSF784\_76\_nt\_1mm\_db -r primer\_R1064\_71\_nt\_1mm\_db -o 76 -e 71 -t Silva\_taxonomic\_table\_v56 -q 2 -k 100 -n 24

**4.2 OTU picking V4 samples.**

otu\_picking\_pair\_end\_read.sh -m Mock\_communities\_V4.txt -p Mock\_communities -a 0.1 -c 0.985 -f primer\_F515\_71\_nt\_1mm\_db -r primer\_R806\_70\_nt\_1\_mm\_db -o 71 -e 70 -t Silva\_taxonomic\_table\_v4 txt -q 2 -k 100 -n 24

**5. OTU recovery for reassignment or removal.**

**5.1 OTU recovery for reassignment or removal for V5V6 samples.**

otu\_recovery\_by\_pattern.sh -t tax\_files -n non\_assigned\_V5V6 -p NA -s taxonomy

otu\_recovery\_by\_pattern.sh -t tax\_files -n Parabacteroides\_V5V6 -p Porphyromonadaceae -s taxonomy

**# 3 otus from V5V6 samples (containing only 12 reads in total) could not be assigned. They were blasted against the NCBI NR repository and the hits obtained did not allow to reclassify the OTUs, therefore they were considered artefacts and removed.**

**# Files Mock\_communities\_V5V6\_non\_assigned\_alternative\_taxonomy\_file and Mock\_communities\_V5V6\_Parabacteroides\_alternative\_taxonomy\_file were concatenated.**

cat otu\_retrievement\_files/non\_assigned\_V5V6\_alternative\_taxonomy\_file otu\_retrievement\_files/Parabacteroides\_V5V6\_alternative\_taxonomy\_file > otu\_retrievement\_files/Mock\_communities\_V5V6\_alternative\_taxonomy\_file

**5.2 OTU recovery for reassignment or removal for V4 samples.**

**# Recover Parabacteroides reads from V4 samples.**

otu\_recovery\_by\_pattern.sh -t tax\_files -n Parabacteroides\_V4 -p TACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGCCTTTTAAGTCAGCGGCCTGTTTGATCCCCACGCTTTCG -s sequence

**6. Taxonomical reassignments**

**# Removal of Parabacteroides and NA OTUs from V5V6 samples. The word remove was written in the 3rd column of Mock\_communities\_V56\_alternative\_taxonomy\_file**

taxonomical\_reassignment.sh -a otu\_retrievement\_files/ Mock\_communities\_V5V6\_alternative\_taxonomy\_file -t tax\_files

**# Removal of Parabacteroides OTUs from V4 samples. The word remove was written in the 3rd column of Mock\_communities\_V4\_alternative\_Parabacteroide\_taxonomy\_file**

taxonomical\_reassignment.sh -a out\_retrievement\_files/Parabacteroides\_V4\_alternative\_taxonomy\_file -t tax\_files

**7. Add theoretical tax files.**

**# Theoretical tax files for Mc.1.t.l44, Mc.1.t.l56, Mc.2.t.l56, Mc.2.t.l44, Mc.3.t.l44, Mc.3.t.l56, Mc.4.t.l44 and Mc.4.t.l56 are added to the folder alternative\_reassigned\_tax\_files for their respective region.**

**8. Region comparison**

**8.1 Region comparison V4 samples.**

region\_16S\_comparator.sh -m Mock\_communities\_V4\_plus\_theoretical.txt -t alternative\_reassigned\_tax\_files -f primer\_BSF784\_76\_nt\_1mm\_db -r primer\_R1064\_71\_nt\_1mm\_db -p 1 &

**8.2 Region comparison V5V6 samples.**

region\_16S\_comparator.sh -m Mock\_communities\_V5V6\_plus\_theorical.txt -t alternative\_reassigned\_tax\_files -f primer\_F515\_71\_nt\_1mm\_db -r primer\_R806\_70\_nt\_1mm\_db -p 2 &

**9. Make biom file.**

**# Samples for both regions are copied to the folder both\_regions\_tax\_files.**

make\_biom\_file.sh -m Mock\_communities.txt -t both\_regions\_tax\_files

**10. Make dnd tree file using uclust with Mock\_communities\_otu\_database.fa**

**11. Downstream analysis with QIIME.**

biom summarize-table -i Mock\_communities.biom -o sample\_size\_mock\_communities

summarize\_taxa\_through\_plots.py -i Mock\_communities.biom -o taxa\_summary\_mc

echo "alpha\_diversity:metrics shannon,simpson,PD\_whole\_tree,chao1,observed\_species" > alpha\_params.txt

echo "beta\_diversity:metrics unweighted\_unifrac,weighted\_unifrac,bray\_curtis" > beta\_params.txt

alpha\_rarefaction.py -i Mock\_communities.biom -m Mock\_communities.txt -o alpha\_rarefaction\_1911\_mc -t Mock\_communities\_otu\_database.dnd -p alpha\_params.txt -e 1911

beta\_diversity\_through\_plots.py -i Mock\_communities.biom -m Mock\_communities.txt -o beta\_diversity\_mc -t Mock\_communities\_otu\_database.dnd -p beta\_params.txt

jackknifed\_beta\_diversity.py -i Mock\_communities.biom -t Mock\_communities\_otu\_database.dnd -m Mock\_communities.txt -o jack\_beta\_diversity\_1911\_mc -e 1911 -p beta\_params.txt

**QIIME Code**

1. **Extract the barcodes. For each library.**

extract\_barcodes.py -f library\_3\_1.fastq -c barcode\_single\_end -m mapping\_file\_forQIIME\_comparison\_MC\_lib1.1.txt -l 8 -o processed\_seqs1.1 -a

1. **Extract the linker. For each library.**

extract\_barcodes.py -f processed\_seqs1.1/reads.fastq -c barcode\_single\_end -m mapping\_file\_forQIIME\_comparison\_MC\_lib1.1.txt -l 2 -o processed\_seqs1.1/wo\_linker -a

1. **Demultiplex using the reads without linker and the 8nt barcode. For each library.**

split\_libraries\_fastq.py -i processed\_seqs1.1/wo\_linker/reads.fastq -o Split\_samples1.1/wo\_linker/ -b processed\_seqs1.1/barcodes.fastq -m mapping\_file\_forQIIME\_comparison\_MC\_lib1.1.txt --barcode\_type 8

1. **Pick closed reference OTUs against the default GreenGeengenes database using the "pick\_otus:enable\_rev\_strand\_match True" line in the parameter file. For each library.**

pick\_closed\_reference\_otus.py -i Split\_samples1.1/wo\_linker/seqs.fna -o otus/otus1.1/wo\_linker/ -p params\_OTU\_picking.txt -a -O 16

1. **Merged all resulting OTU tables**

merge\_otu\_tables.py -i otu\_table1.1.biom,otu\_table2.1.biom,otu\_table3.1.biom,otu\_table4.1.biom,otu\_table5.1.biom,otu\_table6.1.biom,otu\_table7.1.biom,otu\_table1.2.biom,otu\_table2.2.biom,otu\_table3.2.biom,otu\_table4.2.biom,otu\_table5.2.biom,otu\_table6.2.biom,otu\_table7.2.biom -o merged\_otu\_table\_wo\_linker.biom

1. **Apply advised default filtering parameters**

filter\_otus\_from\_otu\_table.py -i ~/qiime/otu\_tables/wo\_linker/merged\_otu\_table\_wo\_linker.biom -o ~/qiime/wo\_linker/merged\_otu\_table\_filtered\_wo\_linker.biom --min\_count\_fraction 0.00005

1. **Due to the quality issues with the Parabacteroides sequences these (family: Porphyromonadaceae) were filtered out manually because the normal command does not work with special characters.**

filter\_taxa\_from\_otu\_table.py -i ~/qiime/wo\_linker/merged\_otu\_table\_filtered\_wo\_linker.biom -o ~/qiime/wo\_linker/otu\_table\_no\_Porphyromonadaceae.biom -n k\_\_Bacteria;p\_\_Bacteroidetes;c\_\_Bacteroidia;o\_\_Bacteroidales;f\_\_Porphyromonadaceae

1. **Manual removal of k\_\_Bacteria;p\_\_Bacteroidetes;c\_\_Bacteroidia;o\_\_Bacteroidales;f\_\_Porphyromonadaceae**

biom convert -i ~/qiime/wo\_linker/merged\_otu\_table\_filtered\_wo\_linker.biom -o ~/qiime/wo\_linker/merged\_otu\_table.filtered.txt --to-tsv --header-key taxonomy

biom convert -i ~/qiime/wo\_linker/merged\_otu\_table.filtered.txt -o ~/qiime/wo\_linker/otu\_table\_no\_Porphyromonadaceae.biom --to-hdf5 --table-type="OTU table" --process-obs-metadata taxonomy

1. **OTU picking for the theoretical sequences**

pick\_closed\_reference\_otus.py -i ~/qiime/wo\_linker/mc\_theoretical\_for\_qiime/Mc.1.t.lfV56.fna -o ~/qiime/wo\_linker/full\_length\_reference/1.fV56 -p ~/qiime/params\_OTU\_picking.txt -a -O 16

1. **Merge all theoretical OTU tables**

merge\_otu\_tables.py -i qiime/wo\_linker/full\_length\_reference/1.fV56/otu\_table.biom,qiime/wo\_linker/full\_length\_reference/2.fV56/otu\_table.biom,qiime/wo\_linker/full\_length\_reference/3.fV56/otu\_table.biom,qiime/wo\_linker/full\_length\_reference/4.fV56/otu\_table.biom,qiime/wo\_linker/full\_length\_reference/1.fV44/otu\_table.biom,qiime/wo\_linker/full\_length\_reference/2.fV44/otu\_table.biom,qiime/wo\_linker/full\_length\_reference/3.fV44/otu\_table.biom,qiime/wo\_linker/full\_length\_reference/4.fV44/otu\_table.biom,qiime/wo\_linker/full\_length\_reference/1.rV56/otu\_table.biom,qiime/wo\_linker/full\_length\_reference/2.rV56/otu\_table.biom,qiime/wo\_linker/full\_length\_reference/3.rV56/otu\_table.biom,qiime/wo\_linker/full\_length\_reference/4.rV56/otu\_table.biom,qiime/wo\_linker/full\_length\_reference/1.rV44/otu\_table.biom,qiime/wo\_linker/full\_length\_reference/2.rV44/otu\_table.biom,qiime/wo\_linker/full\_length\_reference/3.rV44/otu\_table.biom,qiime/wo\_linker/full\_length\_reference/4.rV44/otu\_table.biom -o ~/qiime/wo\_linker/merged\_reference.biom

1. **Merge the sequenced and theoretical data for analysis**

merge\_otu\_tables.py –i ~/qiime/wo\_linker/merged\_reference.biom,qiime/wo\_linker/otu\_table\_no\_Porphyromonadaceae.biom –o ~/qiime/wo\_linker/Reference\_and\_sequenced\_data.biom

1. **Downstream analysis.**

summarize\_taxa\_through\_plots.py -i ~/qiime/wo\_linker/Reference\_and\_sequenced\_data.biom -o taxa\_summary\_qiime\_mc

beta\_diversity\_through\_plots.py -i ~/qiime/wo\_linker/Reference\_and\_sequenced\_data.biom -m ~/qiime/mapping\_all.txt -a -O 16 -e 10000 -t/usr/local/lib/python2.7/dist-packages/qiime\_default\_reference/gg\_13\_8\_otus/trees/97\_otus.tree -o ~/qiime/wo\_linker/beta\_div\_even10000\_reference\_and\_sequenced\_data -f

alpha\_rarefaction.py -i ~/qiime/wo\_linker/Reference\_and\_sequenced\_data.biom -n 30 -e 20000 -o ~/qiime/wo\_linker/a\_rare/all\_no\_porphyr\_with\_reference -m ~/qiime/mapping\_all.txt -t /usr/local/lib/python2.7/dist-packages/qiime\_default\_reference/gg\_13\_8\_otus/trees/97\_otus.tree -f -a -O 16 -p ~/qiime/params\_alpha\_diversity.txt​