Notes S4: Endobacteria

Natacha Bodenhausen, Alessandro Desiro and Klaus Schlaeppi (ENTER YEAR and JOURNAL HERE)

2018-08-23

Contents

Note: This code relies on output files from the main analysis (Notes S3).

Method

The overall approach is to pre-select in a first step candidate endobacteria OTUs, which are subsequently validate by phylogenetic placement in step 2.

STEP 1 Identification of candidate endobacteria OTUs:
We explored two approaches to pre-select candidate endobacteria OTUs. The first approach was based on sequence clustering and for the second we employed co-occurrence characteristics from network analysis.

STEP 2 Validation by phylogenetic placement:
We used phylogenetic placement to validate the representative sequences of candidate endobacteria OTUs by fine mapping to a reference tree of known endobacteria sequences. Here is the commandline code documented, while the approach is explained in detail in Methods S1.

STEP 1 Identification of candidate endobacteria OTUs

Clustering-based approach

Preparing endobacteria reference sequence set

A. Desiro prepared updated input fasta files for BRE/CaGg, MRE and Mycoavidus. They were converted to UPPER case if needed and combined to a endobacteria_ref.fasta file. This file was dereplicated and used as DB to map the filtered representative bOTU fasta sequences to. This step was done with the command '-usearch_global' allowing to build sequence clusters with a relaxed sequence similarity of 0.90.

```
## inspect endobacteria sequences
# less endobacteria/input/CaGg.fas                    # lower case
# less endobacteria/input/MRE.fas                     # lower case
# less endobacteria/input/Mycoavidus.fas            # UPPER case

## convert to UPPER case
awk 'BEGIN{FS=" "}{if(!/>/){print toupper($0)}else{print $1}}'
          endobacteria/input/CaGg.fas > endobacteria/CaGg_UPPER.fas
awk 'BEGIN{FS=" "}{if(!/>/){print toupper($0)}else{print $1}}'
          endobacteria/input/MRE.fas > endobacteria/MRE_UPPER.fas
cp endobacteria/input/Mycoavidus.fas endobacteria/Mycoavidus_UPPER.fas

## combine separate files to one
cat endobacteria/*UPPER.fas > endobacteria/endobacteria_ref.fasta

## manually "-" replaced by "" in file endobacteria/endobacteria_ref.fasta
```

clustering using usearch

```
## dereplicate reads
usearch -derep_fulllength endobacteria/endobacteria_ref.fasta
        -fastaout endobacteria/endobacteria_derep.fasta -sizeout
```

```
# map endobacteria sequences to all OTUs
usearch -usearch_global endobacteria/endobacteria_derep.fasta
        -db data_16S/p327_p1617-o2184-s1_16S_OTU_ab2_renamed.fasta
        -strand both -id 0.90 -uc endobacteria/endobacteria_derep_otu90.txt
```

Upload of the clustering results

bOTUs and number of matches in the endobacteria reference set

```
##
##         * bOTU_1020 bOTU_1046   bOTU_11  bOTU_134 bOTU_1623 bOTU_1682
##       167         2         1         1         7         3         2
## bOTU_1797 bOTU_1822 bOTU_1995 bOTU_2298 bOTU_2334   bOTU_29 bOTU_3291
##         5         1         3         1         1         2         9
##   bOTU_33  bOTU_330  bOTU_390  bOTU_529  bOTU_778  bOTU_800   bOTU_84
##         2        41         1         2         8         1         1
##  bOTU_914  bOTU_915
##         5        11
```

The clustering-based approached identified 22 candidate endobacteria OTUs. The representative fasta sequences of these candidate endobacteria OTUs were then exported to be analyzed in STEP2.

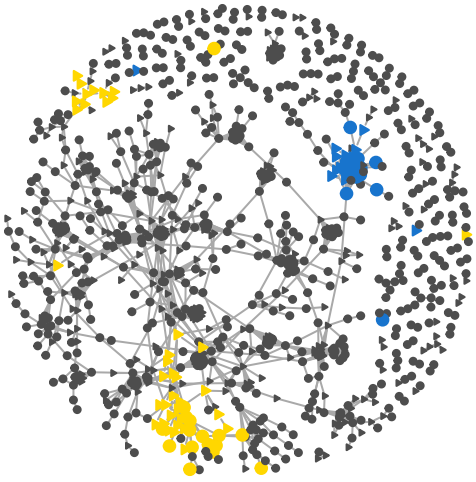
Network analysis based approach

Rebuild the networks using the same codes as in Notes S3 to . . .

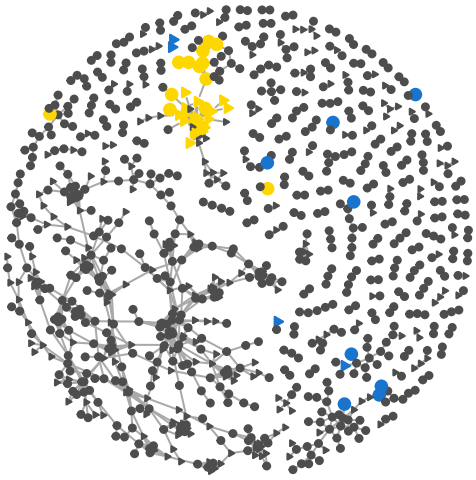
. . . repeat edgeR to identify P-sensitive OTUs (also TMM transformation)

. . . repeat network analysis

Arabidopsis



Petunia

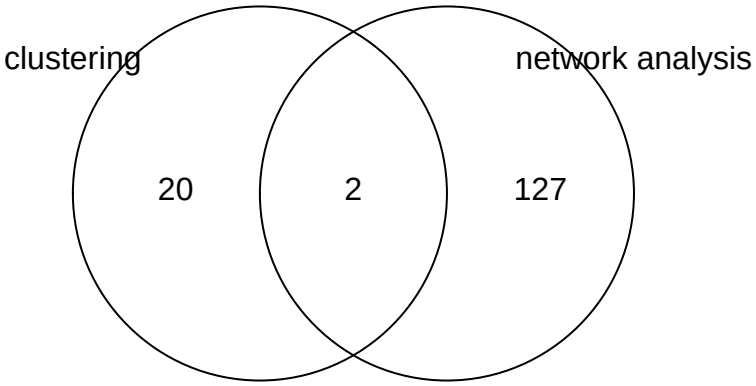


Use network characteristics to identify candidate endobacteria bOTUs

Aproach: we first subsetted the microbiota data to fOTUs with taxonomy assignments to Glomeromycotina, Mortierellomycotina or Endogonaceae. These were then confirmed to be present in the network graphs. Then all their co-occuring bOTUs were identified in the Arabidopsis and the Petunia networks.

As potential hosts we found a total of 67 fOTUs with taxonomy assignments to Glomeromycotina, Mortierellomycotina or Endogonaceae in the network graphs. In the Arabidopsis and Petunia networks, we found 90 and 49 bOTUs that significantly correlated with the potential host fOTUs, respectively. There was some overlap between the two species so that the network analysis based approach identified a total of 129 candidate endobacteria bOTUs. The representative fasta sequences of these candidate endobacteria OTUs were then exported to be analyzed in STEP2.

Candidate endobacteria OTUs by clustering and network analysis



The clustering and network analysis-based approaches identified 22 and 129 candidate endobacteria OTUs, respectively. Only two candidate endobacteria OTUs were identified with both approaches (bOTU_134, bOTU_778).

STEP 2 Validation by phylogenetic placement

This script above produced two output files (cand_endoB_byClustering.fasta, cand_endoB_byNetwork.fasta), which were used for subsequent phylogenetic placement analysis. The fasta sequences of the candidate endobacteria OTUs were aligned and placed into the tree of the reference endobacteria sequences (same as used for clustering). Candidate endobacteria OTUs, which align close to a confirmed endobacterium sequence, we then refer to as *endobacteria bOTU*.

Clustering-based approach

```
## Bayesian (MrBayes) Analysis
begin mrbayes;

exe infile_name;
    set autoclose=yes; lset nst=6 rates=gamma;
  prset statefreqpr=fixed(0.2864,0.1913,0.2630,0.2593)
        revmatpr=fixed(1.0000,3.0856,1.0000,1.0000,4.4832,1.0000)
        shapepr=fixed(0.6970) pinvarpr=fixed(0.2330);
  mcmc ngen=10000000 samplefreq=1000;
    sump burnin=2500;   sumt burnin=2500;
end;

## Maximum Likelihood (RAXML) Analysis
raxmlHPC -s infile_name -n outfile_name -m GTRCAT -x random_number -f a
         -N autoMR -p random_number
```

Phylogenetic placement of the *22* candidate endobacteria OTUs (as identified by the clustering approach) confirmed *six* bOTUs to be phylogenetically close to Burkholderia-related endobacteria and *four* bOTUs phylogenetically close to Mycoplasma-related endobacteria (see Figure S6 and Figure 7). The four Mycoplasma-related endobacteria include bOTU330, bOTU778, bOTU1797 and bOTU3291. The six Burkholderia-related endobacteria are bOTU134, bOTU1020, bOTU1046, bOTU1682, bOTU1995 and bOTU2298. bOTUs 33 and 390 clustered close to environmental Burkholderia sequences.

Network analysis-based approach

```
## Bayesian (MrBayes) Analysis
begin mrbayes;

exe infile_name;
    set autoclose=yes; lset nst=6 rates=gamma;
  prset statefreqpr=fixed(0.2656,0.2055,0.2740,0.2549)
        revmatpr=fixed(1.0000,2.9034,1.0000,1.0000,4.1949,1.0000)
        shapepr=fixed(0.6830) pinvarpr=fixed(0.2190);
  mcmc ngen=10000000 samplefreq=1000;
    sump burnin=2500;   sumt burnin=2500;
end;

## Maximum Likelihood (RAXML) Analysis
raxmlHPC -s infile_name -n outfile_name -m GTRCAT -x random_number -f a
         -N autoMR -p random_number
```

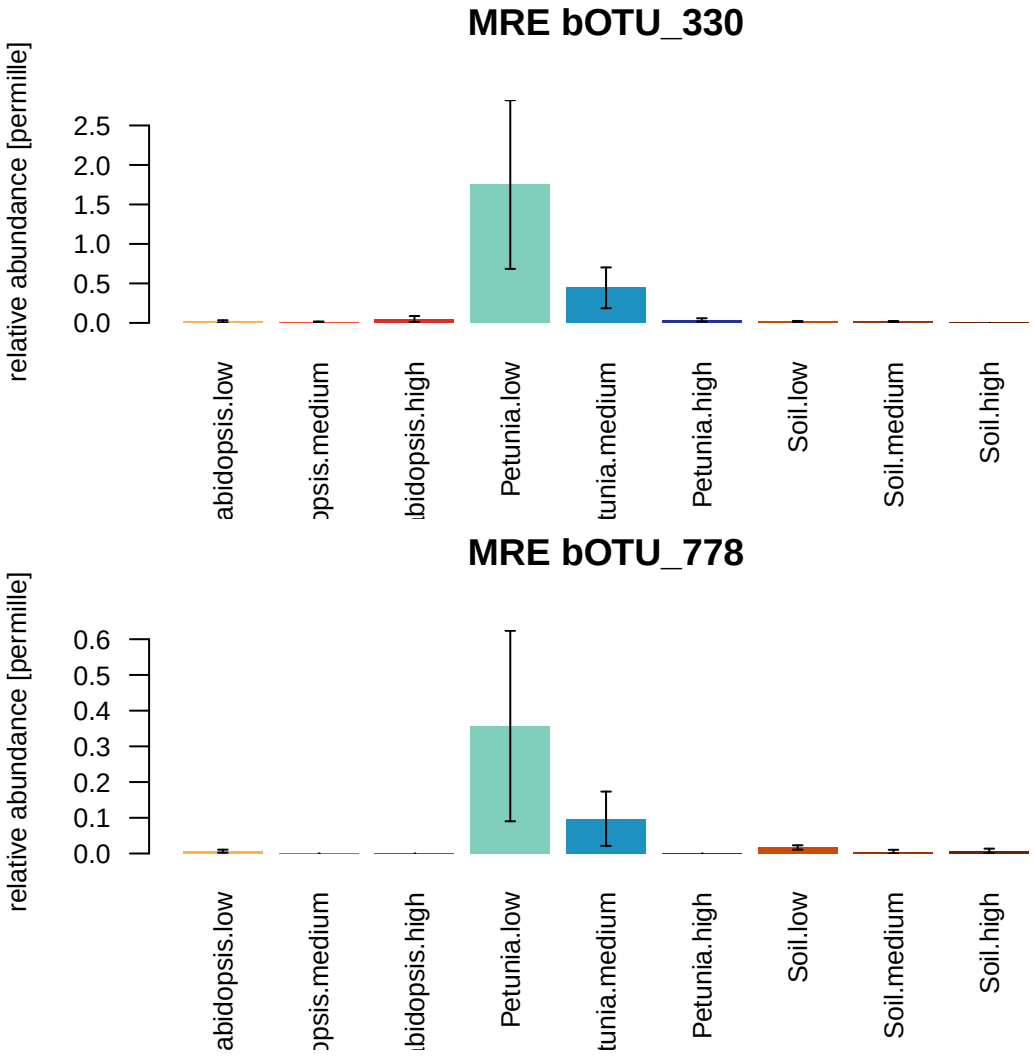
Phylogenetic placement of the 129 candidate endobacteria OTUs (as identified by the network analysis approach) confirmed *two* bOTUs to be phylogenetically close to Mycoplasma-related endobacteria (see Figure

S7). These two bOTUs were bOTU134 and bOTU778, which were found in the overlap of both approaches (see Venndiagramm above).

Characterization of endobacteria bOTUs

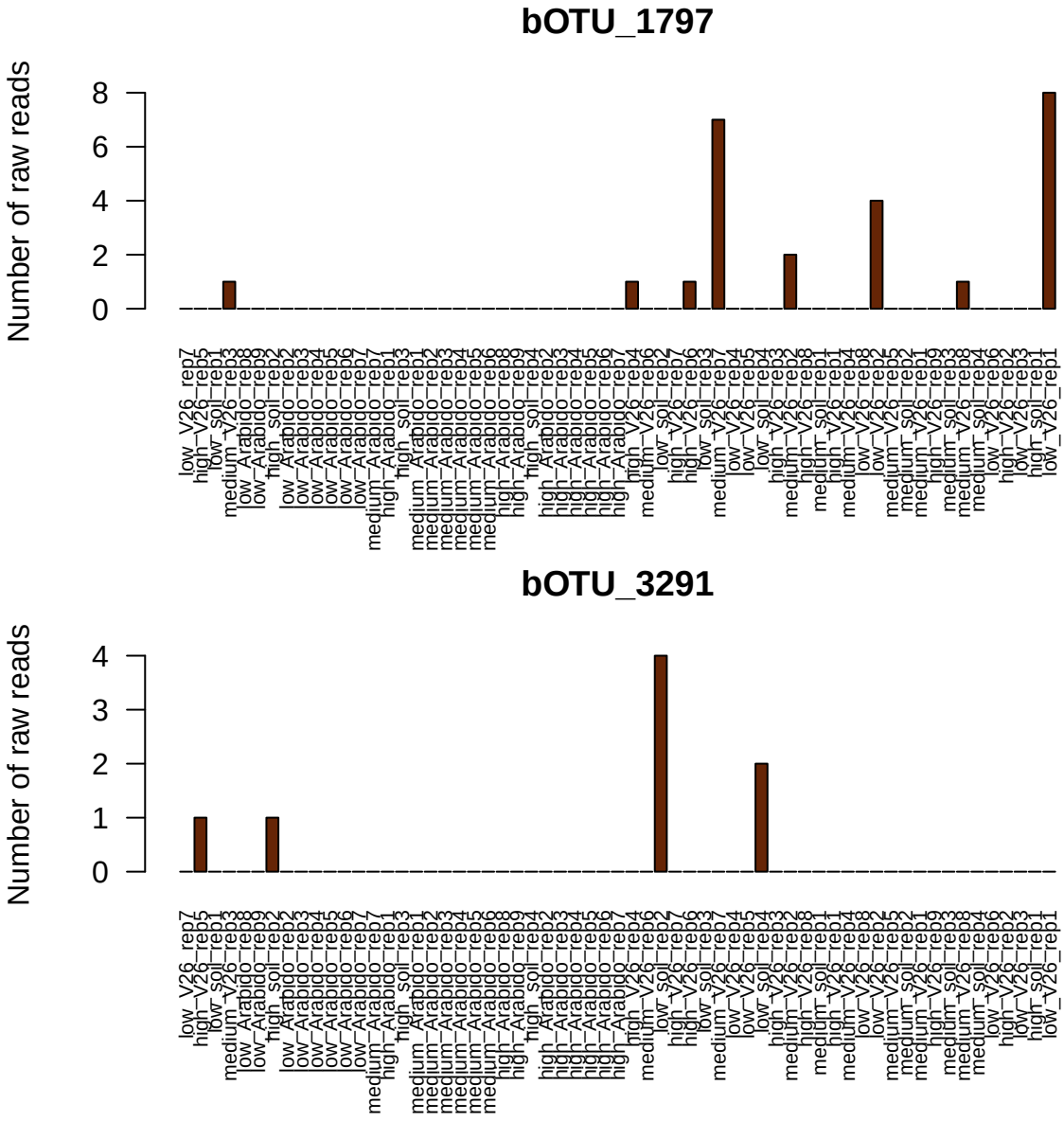
Abundance of MRE bOTUs

The four Mycoplasma-related endobacteria include bOTU330, bOTU778, bOTU1797 and bOTU3291.



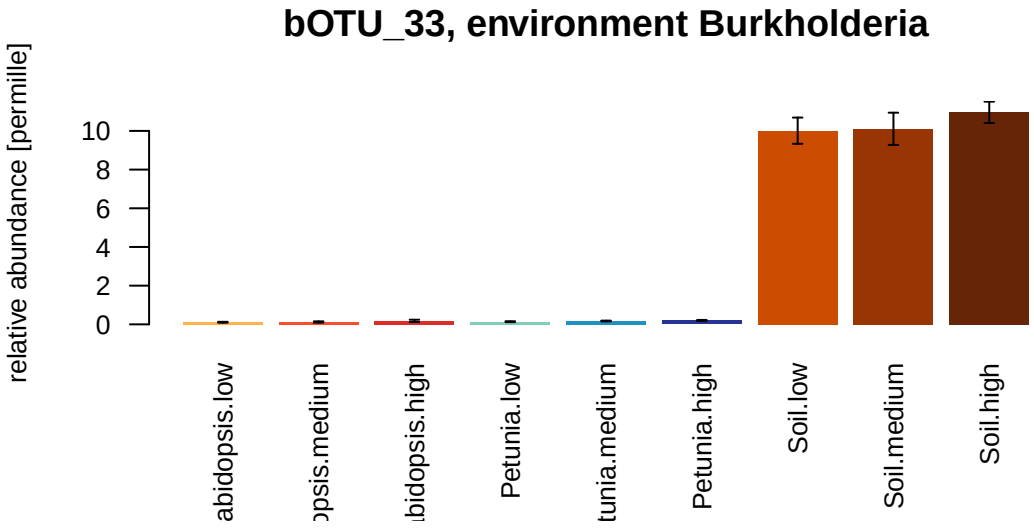
Abundances of MRE bOTUs 1797 and 3291

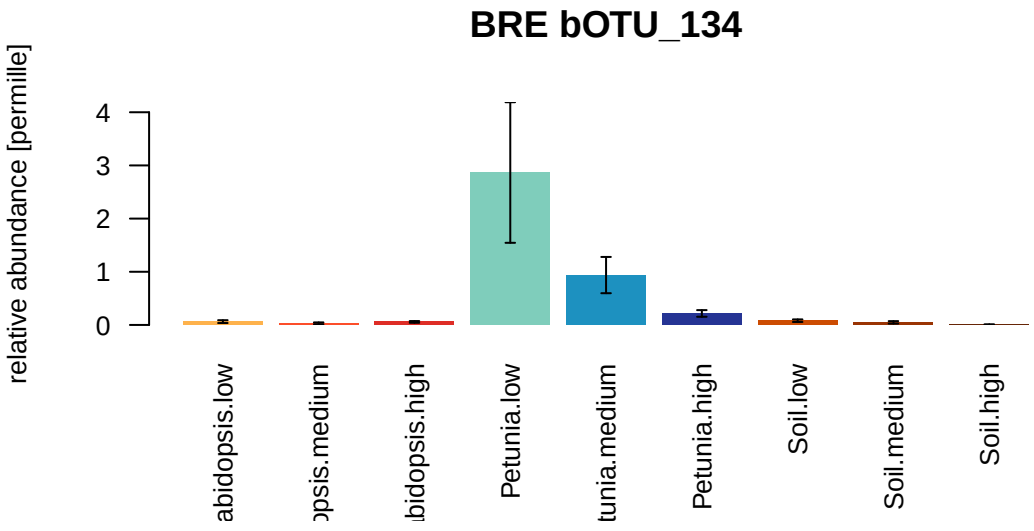
These MRE bOTUs are very low abundant, they fail to pass our filtering strategy (present in at least 4 samples per replicate group). Non-normalized raw reads are shown.

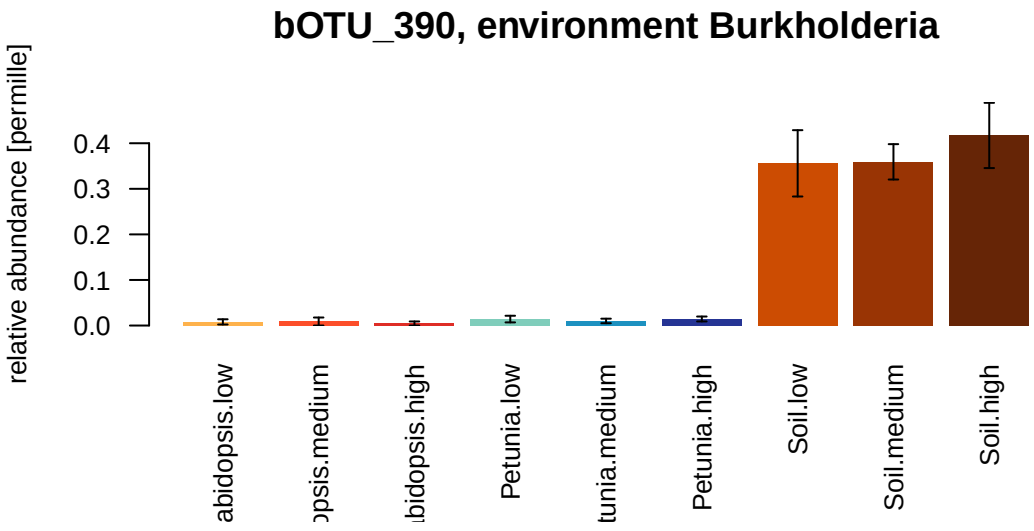


Abundance of BRE bOTUs

The eight Burkholderia-related endobacteria include bOTU134, bOTU390, bOTU1020, bOTU1046, bOTU1682, bOTU1995 and bOTU2298. bOTUs 33 and 390 clustered close to environmental Burkholderia sequences.

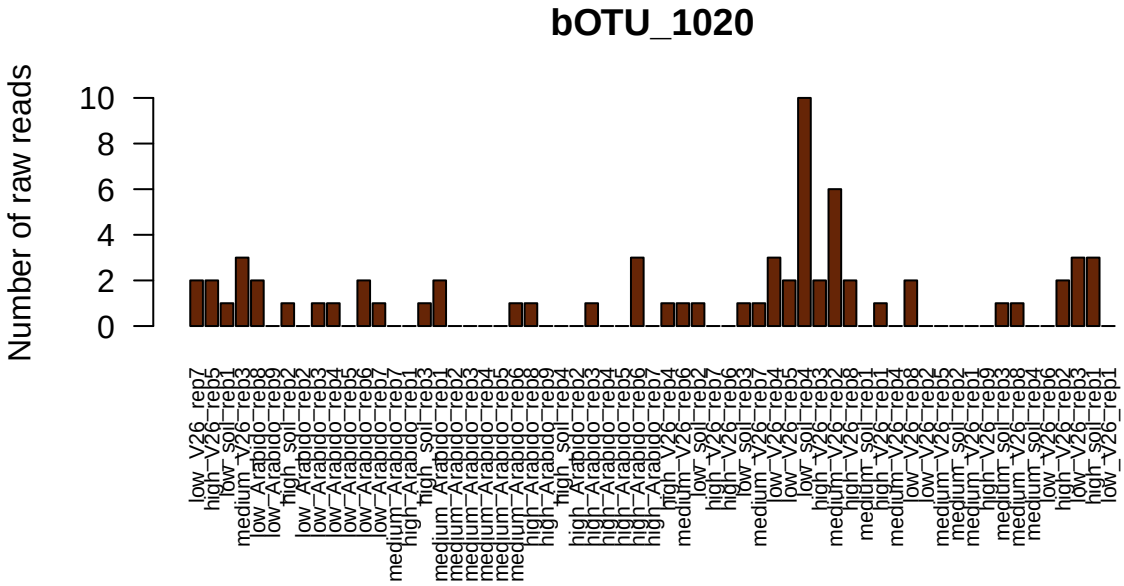


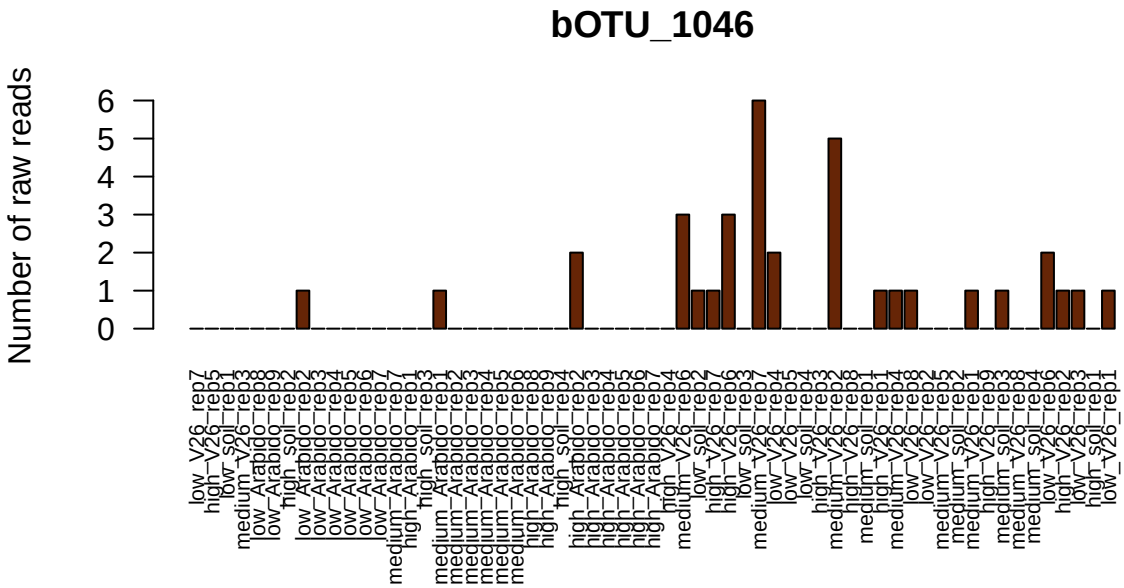


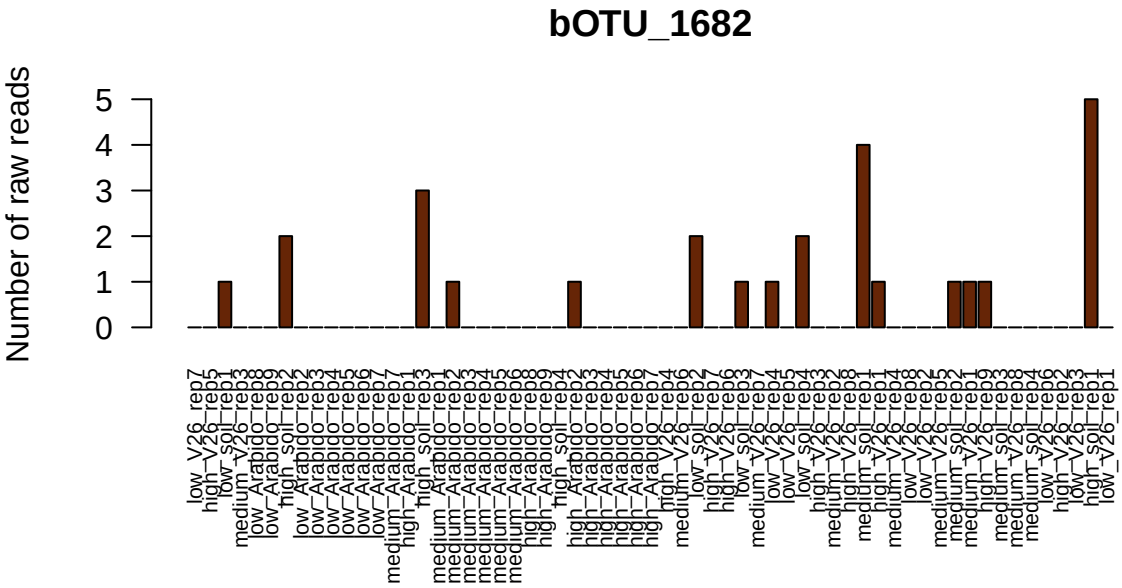


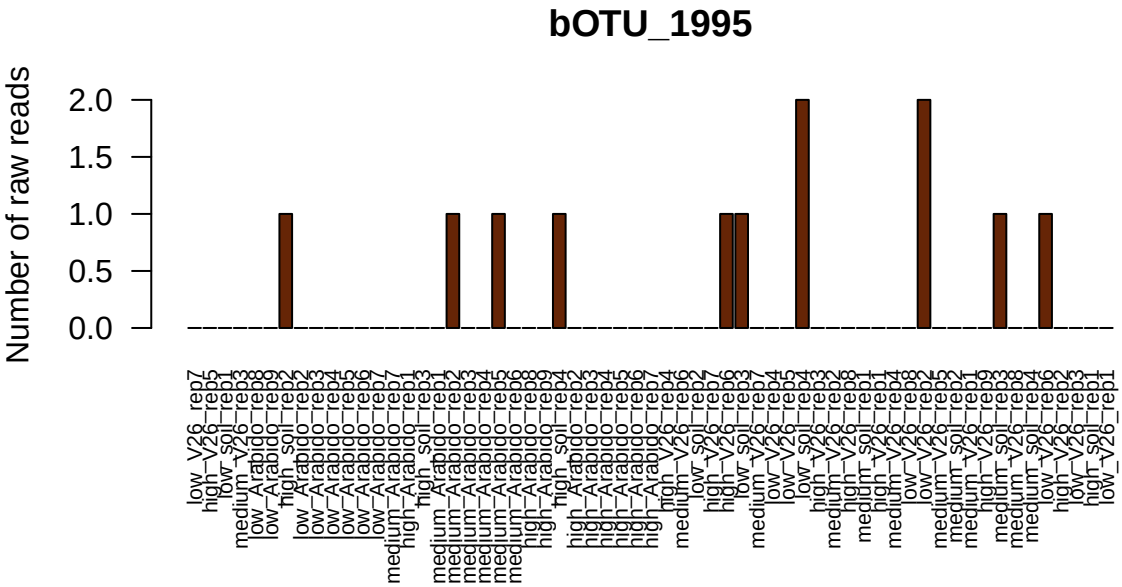
Abundances of BRE bOTUs 1020, 1046, 1682, 1995 and 2298

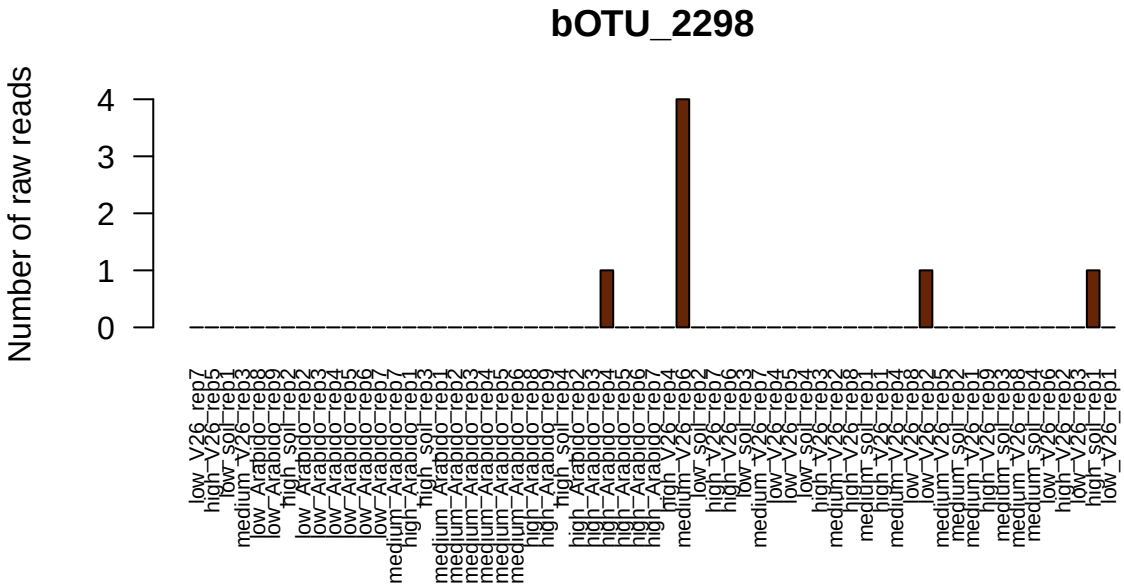
These BRE bOTUs are very low abundant, they fail to pass our filtering strategy (present in at least 4 samples per replicate group). Non-normalized raw reads are shown.






bOTU_1682
Number of raw reads
0
1
2
3
4
5


bOTU_1995
Number of raw reads
0.0
0.5
1.0
1.5
2.0



CONCLUSION

We only consider the bOTUs 134, 330 and 778 as endobacteria OTUs as the other low count bOTUs could not be robustly quantified.