Path to working directory: `/nfs/turbo/schloss-lab/dotsonga`

SILVA Full Length (`/nfs/turbo/schloss-lab/dotsonga/data/references/silva`)

- Download SILVA reference

```
> wget https://www.mothur.org/w/images/3/32/Silva.nr_v132.tgz
> tar xvzf Silva.nr_v132.tgz
```

  - Output files: `silva.nr_v132.align` and `silva.nr_v132.tax`

- Select for bacteria (i.e. remove archaea and eukaryotic sequences)

```
> nfs/turbo/schloss-lab/bin/mothur "#get.lineage(fasta =
silva.nr_v132.align, taxonomy=silva.nr_v132.tax, taxon=Bacteria)"
```

  - Output files: `silva.nr_v132.pick.tax, silva.nr_v132.pick.align`

```
> mv silva.nr_v132.pick.align silva_bacteria.align
> mv silva.nr_v132.pick.tax silva_bacteria.tax
```

- Select for full length sequences

```
> nfs/turbo/schloss-lab/bin/ mothur
"#summary.seqs(fasta=silva_bacteria.align)"

                Start   End     NBases  Ambigs  Polymer NumSeqs
Minimum:        1045    41028   1202    0       4       1
2.5%-tile:      1046    43116   1386    0       5       4707
25%-tile:       1046    43116   1434    0       5       47062
Median:         1046    43116   1452    0       5       94124
75%-tile:       1046    43116   1463    0       6       141186
97.5%-tile:     1046    43116   1490    2       7       183541
Maximum:        1120    43116   2839    5       24      188247
Mean:   1046.01 43115.7 1447.66 0.141909        5.57426
# of Seqs:      188247
```

Output File Names:
silva_bacteria.summary

It took 364 secs to summarize 188247 sequences.

```
> nfs/turbo/schloss-lab/bin/ mothur
"#screen.seqs(fasta=silva_bacteria.align, start=1046, end=43116)"
```

Output File Names:
silva_bacteria.good.align
silva_bacteria.bad.accnos

It took 482 secs to screen 188247 sequences.

```
> nfs/turbo/schloss-lab/bin/ mothur
"#summary.seqs(fasta=silva_bacteria.good.align)"
```

```
                Start   End     NBases  Ambigs  Polymer NumSeqs
Minimum:        1045    43116   1202    0       4       1
2.5%-tile:      1046    43116   1386    0       5       4676
25%-tile:       1046    43116   1434    0       5       46760
Median:         1046    43116   1452    0       5       93519
75%-tile:       1046    43116   1463    0       6       140278
97.5%-tile:     1046    43116   1490    2       7       182361
Maximum:        1046    43116   2839    5       24      187036
Mean:   1046    43116   1447.68 0.141358        5.57431
# of Seqs:      187036
```

Output File Names:
silva_bacteria.good.summary

It took 380 secs to summarize 187036 sequences.

```
> /nfs/turbo/schloss-lab/bin/mothur
"#filter.seqs(fasta=silva_bacteria.good.align, trump=., vertical=T)"
```

It took 51 secs to filter 187036 sequences.

Length of filtered alignment: 13368
Number of columns removed: 36632
Length of the original alignment: 50000
Number of sequences used to construct filter: 187036

Output File Names:
silva_bacteria.filter
silva_bacteria.good.filter.fasta

```
> /nfs/turbo/schloss-lab/bin/mothur
"#unique.seqs(fasta=silva_bacteria.good.filter.fasta)"
```

Output File Names:
silva_bacteria.good.filter.names
silva_bacteria.good.filter.unique.fasta

```
> /nfs/turbo/schloss-lab/bin/mothur
"#pre.cluster(fasta=silva_bacteria.good.filter.unique.fasta,
name=silva_bacteria.good.filter.names, diffs=10)"
```

Total number of sequences before precluster was 187029.
pre.cluster removed 927 sequences.

It took 614 secs to cluster 187029 sequences.

Output File Names:
silva_bacteria.good.filter.unique.precluster.fasta

silva_bacteria.good.filter.unique.precluster.names
silva_bacteria.good.filter.unique.precluster.map

- Next steps...
  - Generate distance matrix (use mothur dist.seqs function)
  - Cluster using OptiClust (use mothur cluster function)
  - Identify taxonomy for each OTU (use mothur classify.otu function)
  - Find representative sequence from each OTU (use mothur get.oturep function)

SILVA V4 `(/nfs/turbo/schloss-lab/dotsonga/data/references/silva)`

- Trim SILVA to V4 region

```
> /nfs/turbo/schloss-lab/bin/mothur
"#pcr.seqs(fasta=silva_bacteria.align, start=11894, end=25319, keepdots=F,
processors=8)"
```

[NOTE]: no sequences were bad, removing silva_bacteria.bad.accnos

It took 57 secs to screen 188247 sequences.

Output File Names:
silva_bacteria.pcr.align

- Verify trim

```
> /nfs/turbo/schloss-lab/bin/mothur
"#summary.seqs(fasta=silva_bacteria.pcr.align)"
```

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1	10216	219	0	3	1
2.5%-tile:	1	13425	292	0	3	4707
25%-tile:	1	13425	293	0	4	47062
Median:	1	13425	293	0	5	94124
75%-tile:	1	13425	293	0	5	141186
97.5%-tile:	1	13425	295	1	6	183541
Maximum:	1982	13425	1467	5	16	188247
Mean: 1	13424	293	0	4		
# of Seqs:	188247					

It took 11 secs to summarize 188247 sequences.

Output File Names:
ilva_bacteria.pcr.summary

```
> mv silva_bacteria.pcr.align silva.bac.v4.align
```

- Filter sequences

```
> /nfs/turbo/schloss-lab/bin/mothur
"#filter.seqs(fasta=silva.bac.v4.align, trump=., vertical=T)"
```

It took 10 secs to filter 188247 sequences.

Length of filtered alignment: 3447
Number of columns removed: 9978
Length of the original alignment: 13425
Number of sequences used to construct filter: 188247

Output File Names:
silva.filter
silva.bac.v4.filter.fasta

```
> /nfs/turbo/schloss-lab/bin/mothur
"#unique.seqs(fasta=silva.bac.v4.filter.fasta)"
```

Output File Names:
silva.bac.v4.filter.names
silva.bac.v4.filter.unique.fasta

```
> /nfs/turbo/schloss-lab/bin/mothur
"#pre.cluster(fasta=silva.bac.v4.filter.unique.fasta,
name=silva.bac.v4.filter.names, diffs=2)"
```

Total number of sequences before precluster was 117381.
pre.cluster removed 32536 sequences.

It took 1560 secs to cluster 117381 sequences.

Output File Names:
silva.bac.v4.filter.unique.precluster.fasta
silva.bac.v4.filter.unique.precluster.names
silva.bac.v4.filter.unique.precluster.map

```
> /nfs/turbo/schloss-lab/bin/mothur
"#dist.seqs(fasta=silva.bac.v4.filter.unique.precluster.fasta, cutoff=0.03)"
```

It took 4497 secs to find distances for 84845 sequences. 471204 distances below cutoff 0.03.

Output File Names:
silva.bac.v4.filter.unique.precluster.dist

```
> /nfs/turbo/schloss-lab/bin/mothur
"#cluster(column=silva.bac.v4.filter.unique.precluster.dist,
name=silva.bac.v4.filter.unique.precluster.names, cutoff=0.03)"
```

It took 10 seconds to cluster

Output File Names:

silva.bac.v4.filter.unique.precluster.opti_mcc.list
silva.bac.v4.filter.unique.precluster.opti_mcc.steps
silva.bac.v4.filter.unique.precluster.opti_mcc.sensspec

```
> /nfs/turbo/schloss-lab/bin/mothur
"#classify.otu(taxonomy=silva_bacteria.tax,
list=silva.bac.v4.filter.unique.precluster.opti_mcc.list,
name=silva.bac.v4.filter.unique.precluster.names)"
```

Output File Names:
silva.bac.v4.filter.unique.precluster.opti_mcc.0.03.cons.taxonomy
silva.bac.v4.filter.unique.precluster.opti_mcc.0.03.cons.tax.summary

```
> /nfs/turbo/schloss-lab/bin/mothur "#get.oturep(method=abundance,
list=silva.bac.v4.filter.unique.precluster.opti_mcc.list,
name=silva.bac.v4.filter.unique.precluster.names)"
```

Output File Names:
silva.bac.v4.filter.unique.precluster.opti_mcc.0.03.rep.names

RDP Full Length `(/nfs/turbo/schloss-lab/dotsonga/data/references/rdp/trainset16_022016.rdp)`

- Download RDP database

```
> wget https://www.mothur.org/w/images/d/dc/Trainset16_022016.rdp.tgz
> tar xvzf Trainset16_022016.rdp.tgz
```

- Select for bacteria

```
> /nfs/turbo/schloss-lab/bin/mothur "#get.lineage(fasta=rdp.fasta,
taxonomy=rdp.tax, taxon=Bacteria)"
```

Output File Names: `rdp.pick.tax, rdp.pick.fasta`

```
> mv rdp.pick.fasta rdp.bacteria.fasta
> mv rdp.pick.tax rdp.bacteria.tax
```

- Align to SILVA SEED reference alignment
  - Download SILVA SEED reference

```
> wget https://www.mothur.org/w/images/7/71/Silva.seed_v132.tgz
```

  - Align

```
> /nfs/turbo/schloss-lab/bin/mothur
"#align.seqs(candidate=rdp.bacteria.fasta, template=silva.seed_v132.align)"
```

Output File Names:
rdp.bacteria.align
rdp.bacteria.align.report

- Filter sequences for full length sequences

```
        > /nfs/turbo/schloss-lab/bin/mothur
"#summary.seqs(fasta=rdp.bacteria.align)"

                Start   End     NBases  Ambigs  Polymer NumSeqs
Minimum:        1045    8600    320     0       4       1
2.5%-tile:      1046    41513   1270    0       5       318
25%-tile:       1046    43100   1409    0       5       3171
Median:         1046    43116   1439    0       5       6341
75%-tile:       1060    43116   1455    0       6       9511
97.5%-tile:     1463    43116   1486    9       7       12364
Maximum:        14957   43116   1709    129     68      12681
Mean:   1094.11 42669.7 1417.67 0.87635 5.52732
# of Seqs:      12681
```

Output File Names:
rdp.bacteria.summary

It took 24 secs to summarize 12681 sequences.

```
        > /nfs/turbo/schloss-lab/bin/mothur
"#screen.seqs(fasta=rdp.bacteria.align, start=1046, end=43116)"
```

Output File Names:
rdp.bacteria.good.align
rdp.bacteria.bad.accnos

It took 24 secs to screen 12681 sequences.

```
        > /nfs/turbo/schloss-lab/bin/mothur
"#summary.seqs(fasta=rdp.bacteria.good.align)"

                Start   End     NBases  Ambigs  Polymer NumSeqs
Minimum:        1046    43116   1361    0       4       1
2.5%-tile:      1046    43116   1386    0       5       175
25%-tile:       1046    43116   1437    0       5       1743
Median:         1046    43116   1448    0       5       3485
75%-tile:       1046    43116   1462    0       6       5227
97.5%-tile:     1046    43116   1488    5       7       6795
Maximum:        1046    43116   1617    59      48      6969
Mean:   1046    43116   1446.35 0.559191        5.49003
# of Seqs:      6969
```

Output File Names:
rdp.bacteria.good.summary

It took 14 secs to summarize 6969 sequences.

```
        > /nfs/turbo/schloss-lab/bin/mothur
"#filter.seqs(fasta=rdp.bacteria.good.align, trump=., vertical=T)"
```

It took 1 secs to filter 6969 sequences.

Length of filtered alignment: 3709
Number of columns removed: 46291
Length of the original alignment: 50000
Number of sequences used to construct filter: 6969

```
        > /nfs/turbo/schloss-lab/bin/mothur
"#unique.seqs(fasta=rdp.bacteria.good.filter.fasta)"
```

Output File Names:
rdp.bacteria.good.filter.names
rdp.bacteria.good.filter.unique.fasta

```
        > /nfs/turbo/schloss-lab/bin/mothur
"#pre.cluster(fasta=rdp.bacteria.good.filter.unique.fasta,
name=rdp.bacteria.good.filter.names, diffs=2)"
```

Output File Names:
rdp.bacteria.good.filter.unique.precluster.fasta
rdp.bacteria.good.filter.unique.precluster.names
rdp.bacteria.good.filter.unique.precluster.map

```
        > /nfs/turbo/schloss-lab/bin/mothur
"#dist.seqs(fasta=rdp.bacteria.good.filter.unique.precluster.fasta,
cutoff=0.03)"
```

Output File Names:
rdp.bacteria.good.filter.unique.precluster.dist

```
        > /nfs/turbo/schloss-lab/bin/mothur
"#cluster(column=rdp.bacteria.good.filter.unique.precluster.dist,
name=rdp.bacteria.good.filter.unique.precluster.names, cutoff=0.03)"
```

Output File Names:
rdp.bacteria.good.filter.unique.precluster.opti_mcc.list
rdp.bacteria.good.filter.unique.precluster.opti_mcc.steps
rdp.bacteria.good.filter.unique.precluster.opti_mcc.sensspec

```
        > /nfs/turbo/schloss-lab/bin/mothur
"#classify.otu(taxonomy=rdp.bacteria.tax,
list=rdp.bacteria.good.filter.unique.precluster.opti_mcc.list,
name=rdp.bacteria.good.filter.unique.precluster.names)"
```

Output File Names:

rdp.bacteria.good.filter.unique.precluster.opti_mcc.0.03.cons.taxonomy
rdp.bacteria.good.filter.unique.precluster.opti_mcc.0.03.cons.tax.summary

```
> /nfs/turbo/schloss-lab/bin/mothur "#get.oturep(method=abundance,
list=rdp.bacteria.good.filter.unique.precluster.opti_mcc.list,
name=rdp.bacteria.good.filter.unique.precluster.names)"
```

Output File Names:
rdp.bacteria.good.filter.unique.precluster.opti_mcc.0.03.rep.names

RDP V4 (/nfs/turbo/schloss-lab/dotsonga/data/references/rdp/trainset16_022016.rdp)

- Trim RDP to V4 region

```
> /nfs/turbo/schloss-lab/bin/mothur
"#pcr.seqs(fasta=rdp.bacteria.align, start=11894, end=25319, keepdots=F,
processors=8)"
```

It took 6 secs to screen 12681 sequences.

Output File Names:
rdp.bacteria.pcr.align
rdp.bacteria.bad.accnos
rdp.bacteria.scrap.pcr.align

```
> /nfs/turbo/schloss-lab/bin/mothur
"#summary.seqs(fasta=rdp.bacteria.pcr.align)"
```

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1	1235	6	0	2	1
2.5%-tile:	1	13425	291	0	3	317
25%-tile:	1	13425	293	0	4	3170
Median:	1	13425	293	0	4	6339
75%-tile:	1	13425	293	0	5	9508
97.5%-tile:	1	13425	294	2	6	12361
Maximum:	3063	13425	363	44	14	12677
Mean: 2	13387	292	0	4		
# of Seqs:	12677					

It took 1 secs to summarize 12677 sequences.

Output File Names:
rdp.bacteria.pcr.summary

```
> mv rdp.bacteria.pcr.align rdp.bac.v4.align
```

- Next steps...
  - Filter sequences (use mothur filter.seqs(), unique.seqs(), and pre.cluster() functions)

- Generate distance matrix (use mothur dist.seqs function)
- Cluster using OptiClust (use mothur cluster function)
- Identify taxonomy for each OTU (use mothur classify.otu function)
- Find representative sequence from each OTU (use mothur get.oturep function)

Greengenes Full Length `(/nfs/turbo/schloss-lab/dotsonga/data/references/greengenes)`

- Download Greengenes database

```
> wget http://www.mothur.org/w/images/6/68/Gg_13_8_99.taxonomy.tgz
> wget http://www.mothur.org/w/images/1/19/Gg_13_8_99.refalign.tgz
> tar xvzf Gg_13_8_99.refalign.tgz
> tar xvzf Gg_13_8_99.taxonomy.tgz
```

- Select for bacteria

```
> /nfs/turbo/schloss-lab/bin/mothur "#get.lineage(fasta=gg_13_8_99.fasta, taxonomy=gg_13_8_99.gg.tax, taxon=Bacteria)"
```

Output File Names:
gg_13_8_99.gg.pick.tax
gg_13_8_99.pick.fasta

```
> mv gg_13_8_99.pick.fasta gg_13_8_99.bacteria.fasta
> mv gg_13_8_99.gg.pick.tax gg_13_8_99.bacteria.tax
```

- Align to SILVA SEED reference alignment

```
> /nfs/turbo/schloss-lab/bin/mothur "#align.seqs(fasta=gg_13_8_99.bacteria.fasta, reference=silva.seed_v132.align)"
```

[WARNING]: Some of your sequences generated alignments that eliminated too many bases, a list is provided in gg_13_8_99.bacteria.flip.accnos. If you set the flip parameter to true mothur will try aligning the reverse compliment as well.

It took 7059 secs to align 198510 sequences.

Output File Names:
gg_13_8_99.bacteria.align
gg_13_8_99.bacteria.align.report
gg_13_8_99.bacteria.flip.accnos

- Filter for full length sequence

```
> /nfs/turbo/schloss-lab/bin/mothur "#screen.seqs(fasta=gg_13_8_99.bacteria.align, start=2000, end=41788)"
```

It took 112 secs to screen 198510 sequences, removed 5786.

Output File Names:
gg_13_8_99.bacteria.good.align
gg_13_8_99.bacteria.bad.accnos

It took 117 secs to screen 198510 sequences.

```
> /nfs/turbo/schloss-lab/bin/mothur
"#summary.seqs(fasta=gg_13_8_99.bacteria.good.align)"

                Start   End     NBases  Ambigs  Polymer NumSeqs
Minimum:        1045    41788   1254    0       4       1
2.5%-tile:      1046    41788   1330    0       5       4819
25%-tile:       1046    42554   1372    0       5       48182
Median:         1046    43115   1418    0       5       96363
75%-tile:       1051    43116   1452    0       6       144544
97.5%-tile:     1776    43116   1484    5       8       187906
Maximum:        1817    43116   1891    213     213     192724
Mean:   1089    42848   1412    0       5
# of Seqs:      192724
```

It took 101 secs to summarize 192724 sequences.

Output File Names:
gg_13_8_99.bacteria.good.summary

```
> /nfs/turbo/schloss-lab/bin/mothur
"#filter.seqs(fasta=gg_13_8_99.bacteria.good.align, trump=., vertical=T)"
```

It took 259 secs to filter 192724 sequences.

Length of filtered alignment: 6736
Number of columns removed: 43264
Length of the original alignment: 50000
Number of sequences used to construct filter: 192724

Output File Names:
gg_13_8_99.filter
gg_13_8_99.bacteria.good.filter.fasta

```
> /nfs/turbo/schloss-lab/bin/mothur
"#unique.seqs(fasta=gg_13_8_99.bacteria.good.filter.fasta)"
```

Output File Names:
gg_13_8_99.bacteria.good.filter.names
gg_13_8_99.bacteria.good.filter.unique.fasta

```
> /nfs/turbo/schloss-lab/bin/mothur
"#pre.cluster(fasta=gg_13_8_99.bacteria.good.filter.unique.fasta,
name=gg_13_8_99.bacteria.good.filter.names, diffs=2)"
```

Total number of sequences before precluster was 192428.
pre.cluster removed 42 sequences.

It took 590 secs to cluster 192428 sequences.

Output File Names:
gg_13_8_99.bacteria.good.filter.unique.precluster.fasta
gg_13_8_99.bacteria.good.filter.unique.precluster.names
gg_13_8_99.bacteria.good.filter.unique.precluster.map

- Next steps...
  - Generate distance matrix (use mothur dist.seqs function)
  - Cluster using OptiClust (use mothur cluster function)
  - Identify taxonomy for each OTU (use mothur classify.otu function)
  - Find representative sequence from each OTU (use mothur get.oturep function)

Greengenes V4 `(/nfs/turbo/schloss-lab/dotsonga/data/references/greengenes)`

- Trim to V4 region

```
> /nfs/turbo/schloss-lab/bin/mothur
"#pcr.seqs(fasta=gg_13_8_99.bacteria.align, start=11894, end=25319,
keepdots=F, processors=8)"
```

It took 60 secs to screen 198510 sequences.

Output File Names:
gg_13_8_99.bacteria.pcr.align
gg_13_8_99.bacteria.bad.accnos
gg_13_8_99.bacteria.scrap.pcr.align

```
> /nfs/turbo/schloss-lab/bin/mothur
"#summary.seqs(fasta=gg_13_8_99.bacteria.pcr.align)"
```

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1	13400	170	0	2	1
2.5%-tile:	1	13425	292	0	3	4963
25%-tile:	1	13425	293	0	4	49624
Median:	1	13425	293	0	4	99247
75%-tile:	1	13425	293	0	5	148870
97.5%-tile:	1	13425	297	1	6	193531
Maximum:	1236	13425	469	126	125	198493
Mean:	1	13424	293	0	4	
# of Seqs:	198493					

It took 9 secs to summarize 198493 sequences.

Output File Names:
gg_13_8_99.bacteria.pcr.summary

- Filter sequences

```
> /nfs/turbo/schloss-lab/bin/mothur
"#filter.seqs(fasta=gg.bac.v4.align, trump=., vertical=T)"
```

It took 12 secs to filter 198493 sequences.

Length of filtered alignment: 1358
Number of columns removed: 12067
Length of the original alignment: 13425
Number of sequences used to construct filter: 198493

Output File Names:
gg.filter
gg.bac.v4.filter.fasta

```
> /nfs/turbo/schloss-lab/bin/mothur
"#unique.seqs(fasta=gg.bac.v4.filter.fasta)"
```

Output File Names:
gg.bac.v4.filter.names
gg.bac.v4.filter.unique.fasta

```
> /nfs/turbo/schloss-lab/bin/mothur
"#pre.cluster(fasta=gg.bac.v4.filter.unique.fasta,
name=gg.bac.v4.filter.names, diffs=2)"
```

Total number of sequences before precluster was 154119.
pre.cluster removed 30769 sequences.

It took 1117 secs to cluster 154119 sequences.

Output File Names:
gg.bac.v4.filter.unique.precluster.fasta
gg.bac.v4.filter.unique.precluster.names
gg.bac.v4.filter.unique.precluster.map

- Next steps...
  - Generate distance matrix (use mothur dist.seqs function)
  - Cluster using OptiClust (use mothur cluster function)
  - Identify taxonomy for each OTU (use mothur classify.otu function)
  - Find representative sequence from each OTU (use mothur get.oturep function)

OptiClust on Samples

- Soil `(/nfs/turbo/schloss-lab/dotsonga/data/soil)`

```
> /nfs/turbo/schloss-lab/bin/mothur "#unique.seqs(fasta=soil.fasta)"
```

Output File Names:
soil.names
soil.unique.fasta

```
> /nfs/turbo/schloss-lab/bin/mothur "#cluster(column=soil.dist,
name=soil.names)"
```

It took 190 seconds to cluster

Output File Names:
soil.opti_mcc.list
soil.opti_mcc.steps
soil.opti_mcc.sensspec

- Mice `(/nfs/turbo/schloss-lab/dotsonga/data/mice)`

```
> /nfs/turbo/schloss-lab/bin/mothur "#unique.seqs(fasta=mice.fasta)"
```

Output File Names:
mice.names
mice.unique.fasta

```
> /nfs/turbo/schloss-lab/bin/mothur
"#cluster(column=mice.dist,name=mice.names)"
```

It took 87 seconds to cluster

Output File Names:
mice.opti_mcc.list
mice.opti_mcc.steps
mice.opti_mcc.sensspec

- Marine `(/nfs/turbo/schloss-lab/dotsonga/data/marine)`

```
> /nfs/turbo/schloss-lab/bin/mothur "#unique.seqs(fasta=marine.fasta)"
```

Output File Names:
marine.names
marine.unique.fasta

```
> /nfs/turbo/schloss-lab/bin/mothur
"#cluster(column=marine.dist,name=marine.names)"
```

It took 224 seconds to cluster

Output File Names:
marine.opti_mcc.list
marine.opti_mcc.steps
marine.opti_mcc.sensspec