

## EXTRA INFORMATION ABOUT 16S rRNA SEQUENCES

Because of the limited amount of time, we previously downloaded the references sequences. Here are the step that were taken to retrieve and cluster them:

1. Go to <https://www.arb-silva.de/> . It is a database for 16S rRNA sequences. This is a good DB because the sequences are curated and quality checked (as opposed to eg. regular Genbank (unless you go to the RefSeq category)).
2. In the top tabs, click on “Browser”
3. There are a few drop down menus you can choose from. For database
  - a. SSU: Small subunit: This is the phylogenetic marker used for phylogenies.
  - b. LSU : Large subunit:

Taxonomy:

- a. SILVA, Silva Ref, SILVA Ref NR
- b. LTP
- c. EMBL
- d. Greengenes:
- e. RDP: Ribosomal database

The screenshot shows the SILVA database Browser interface. At the top, there are navigation tabs: Home, SILVAngs, Browser (selected), Search, Aligner, Download, Documentation, Projects, FISH & Probes, and Contact. Below the tabs, there are dropdown menus for Database (SSU 128) and Taxonomy (SILVA Ref NR). A 'Show' dropdown is set to 'Cart'. In the top right corner, there is a 'Cart: 537339' with options to Show, Clear, or Download. The main content area displays a list of taxonomic categories and their percentages. On the right, an 'Export selected sequences' dialog box is open, showing options for output format (ARB, FASTA with gaps, FASTA common gaps, FASTA without gaps) and archive type (Zip, tar.gz, [none]). The 'FASTA without gaps' option is selected, and the 'Start export' button is visible at the bottom of the dialog.

**Database:** SSU 128 **Show:** Cart

**Taxonomy:** SILVA Ref NR

**Cart:** 537339  
Show  
Clear  
Download

**Export selected sequences**

Select output format:

- ☐ ARB
- ☐ FASTA with gaps
- ☐ FASTA common gaps
- ☒ FASTA without gaps

Select archive type:

- ☐ Zip
- ☐ tar.gz
- ☒ [none]

Start export

**SILVA** **Bacteria**

**SILVA** (9.6%)  
(3)  
Archaea  
**Bacteria** (11%)  
Eukaryota (0%)

**Bacteria** (11%)  
(75)  
AC1 (15%)  
Acetothermia (13%)  
Acidobacteria (8.9%)  
Actinobacteria (11%)  
Aerophobetes (20%)  
Aminicenantes (9.5%)  
Aquificae (16%)  
Armatimonadetes (10%)  
Atribacteria (14%)  
Bacteroidetes (8.8%)  
BJ-169 (20%)  
BP4 (50%)  
BRC1 (10%)  
Caldiserica (9.3%)  
Calescamantes (25%)  
Candidatus Berkelbacteria (6.6%)  
Chlamydiae (7.6%)  
Chlorobi (4.2%)  
Chloroflexi (7.9%)  
Chrysiogenetes (48%)  
Cloacimonetes (1.3%)  
CPR2 (1.2%)  
Cyanobacteria (10%)  
Deferribacteres (9%)

Click on “Start Export”

## Your data

SILVA Taskmanager									
#	Job Name	Creation Time	Job Type	Status	Quantity	Progress	Status Message	Elapsed Ti...	Queue
1		2017-09-03 14:0...	Export	Processing	537339	27000	exporting sequences...	00:01:28	0
2		2017-09-03 14:0...	Export	Aborted	537339			00:00:36	0
3		2017-09-03 14:0...	Export	Aborted	537339			00:00:37	0

### Important Remarks to handle Custom Generated Files

- Please make sure that you '→ **Scan for Unknown Fields**' in the Species Information window of ARB when you open a custom-made ARB database for the first time.
- To merge your custom-made ARB database with your personal ARB database click on Merge two ARB Databases in the ARB Intro window. Detailed information can be found in the → FAQs section.

### License Information

Users from non-academic/commercial environments should have a look at the → [SILVA License Information](#)

### Problems?

[Mail us](#) or check our [Twitter messages](#)! On our Twitter account we'll keep you up to date and provide information about updates and problems.

The Status message is at exporting sequences, you can see how many have been processed so far under "Progress".

Using the program cd-hit-est (<https://github.com/weizhongli/cdhit>) we clustered the 537 339 sequences downloaded from SILVA Ref NR using a cutoff of 80% identity to reduce the redundancy in the data. It ended up being 3044 sequences, but for this working this is still too much.

```
cd-hit-est -M [memory limit] -c [cutoff value] -i [input fasta file] -o [output file name]
```

Example:

```
> cd-hit-est -M 32000 -c 0.80 -i arb-silva.de_2017-09-03_id457204_tax_silva.fasta -o SILVAREfNR_2017-09-03_c80.fasta
```

To further reduce the number of sequences, we will only take sequences that are not "uncultured" using pyfasta (<https://pypi.python.org/pypi/pyfasta/>), a python package to deal with fasta sequences. Pyfasta can extract sequences from a multi-fasta file, by selecting or excluding sequences that match a certain header name:

```
pyfasta extract --header --exclude --file headers_uncultured_to_remove.txt --fasta Scenario1.fas > Scenario1_ref_croche_nuncultured.fasta
```

Then

- 1) remove gaps using sed (sed 's/>/g' in.fasta > out.fasta)
- 2) use cd-hit to cluster 35688 OTU sequences at 80% similarity just to reduce the number of sequences for the workshop. resulted in 657 clusters. (cd-hit-est -c .80 -i meso2016v3rep\_nogaps.fasta -o meso2016v3rep\_80cdhit.fasta)
- 3) Replace the headers >Otu for >Brazil Sequence Otu (just so it's easier to see in the tree).
- 3) use muscle to align the reference sequences + croche + brazil sequences = 1922 sequences in total.