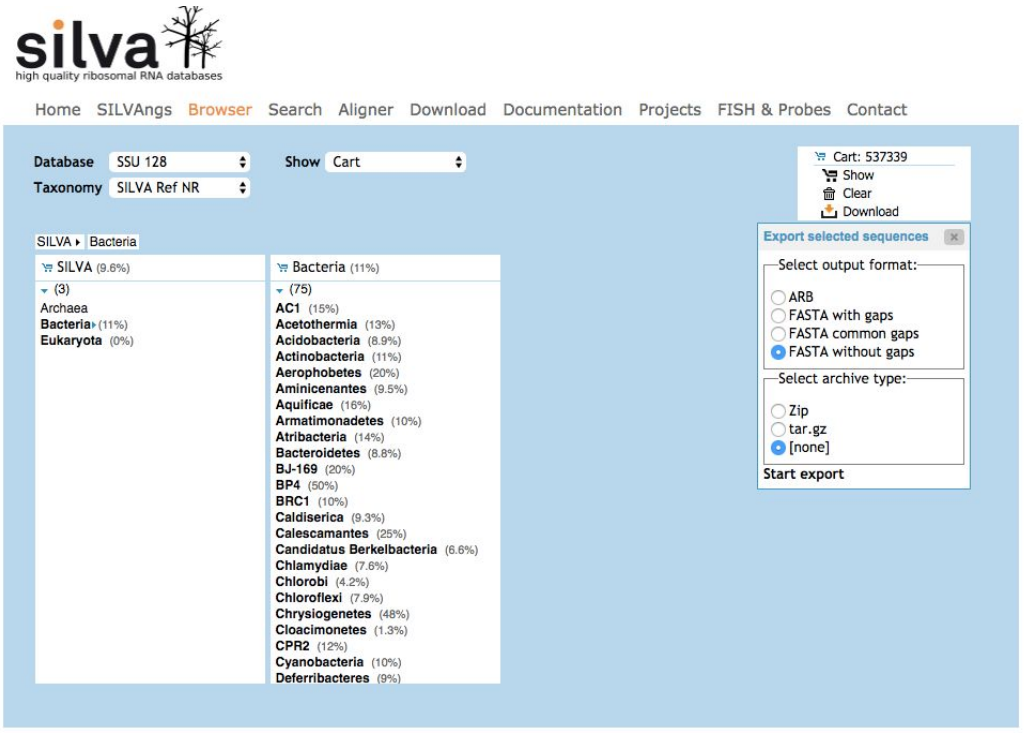
**EXTRA INFORMATION ABOUT 16S rRNA SEQUENCES**

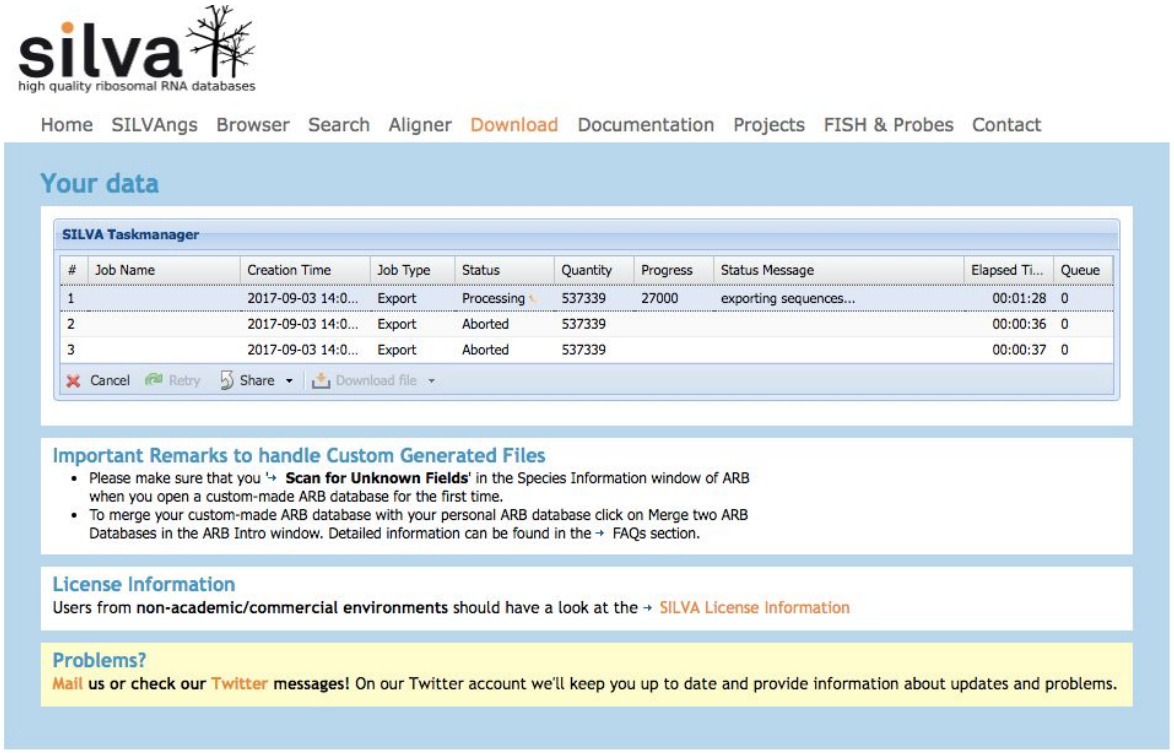
**Because of the limited amount of time, we previously downloaded the references sequences. Here are the step that were taken to retrieve and cluster them:**

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




Click on “Start Export”




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
SILVARefNR_2017-09-03_c80.fasta > SILVARefNR_2017-09-03_c80-nouncultured.fasta
```