Tutorial 5: OTU Clustering USearch Walkthrough

Written By: Jonathan Palko and Thomas Coard

1. Download USearch binary file from the following link and unzip it
    a. https://drive5.com/usearch/download.html
2. Put binary file in accessible folder after extracting it
    a. Linux file name will be something similar to usearch10.0.786_i86linux32
        i. NOTE: WSL 1 Doesn't support 32-bit binary files. USearch only has a 32-bit binary for free so you MUST use WSL 2 if using Linux through Windows.
3. Give the binary file read and execute permissions
    a. For Linux use chmod +x [PATH]/usearch
        i. [PATH] is the file path to the folder with the binary that changes for each machine
4. Set the $usearch environment variable by linking to the USearch bin
    a. https://drive5.com/usearch/manual/env_usearch.html
    b. Run export usearch=[PATH]/usearch10.0.786_i86linux32 in bash to set the environment variable
    c. Alternatively, go to the ~/.bashrc file and add the export usearch=[PATH]/usearch10.0.786_i86linux32 to link the binary to the PATH
        i. [PATH] is the file path to the folder with the binary that changes for each machine
5. Download the Misop tutorial data, scripts, and precomputed results: https://drive5.com/usearch/manual/upp_tut_misop.html
    a. The exercises use data from the mothur MiSeq SOP
6. Make sure to first run the run.bash and run_mock.bash in the misop/scripts folder

a. To run the bash scripts use the following two commands in the Terminal: ./run.bash and ./run_mock.bash while in the correct folder

b. NOTE: That there may be some issues with the version checking, which indicates either an out of date version, an issue with the $usearch environment variable or a problem with the run.bash & run_mock.bash code

   i. If it is the last case (bad error messaging) then simply go into the bash file and comment out the following lines for run.bash:

```
version=`$usearch -version | sed "-es/usearch //" | sed "-es/v10.*/v10/"`

if [ x$version != xv10 ] ; then
    echo "usearch version too old, need v10" >> /dev/stderr
    exit 1
fi
```

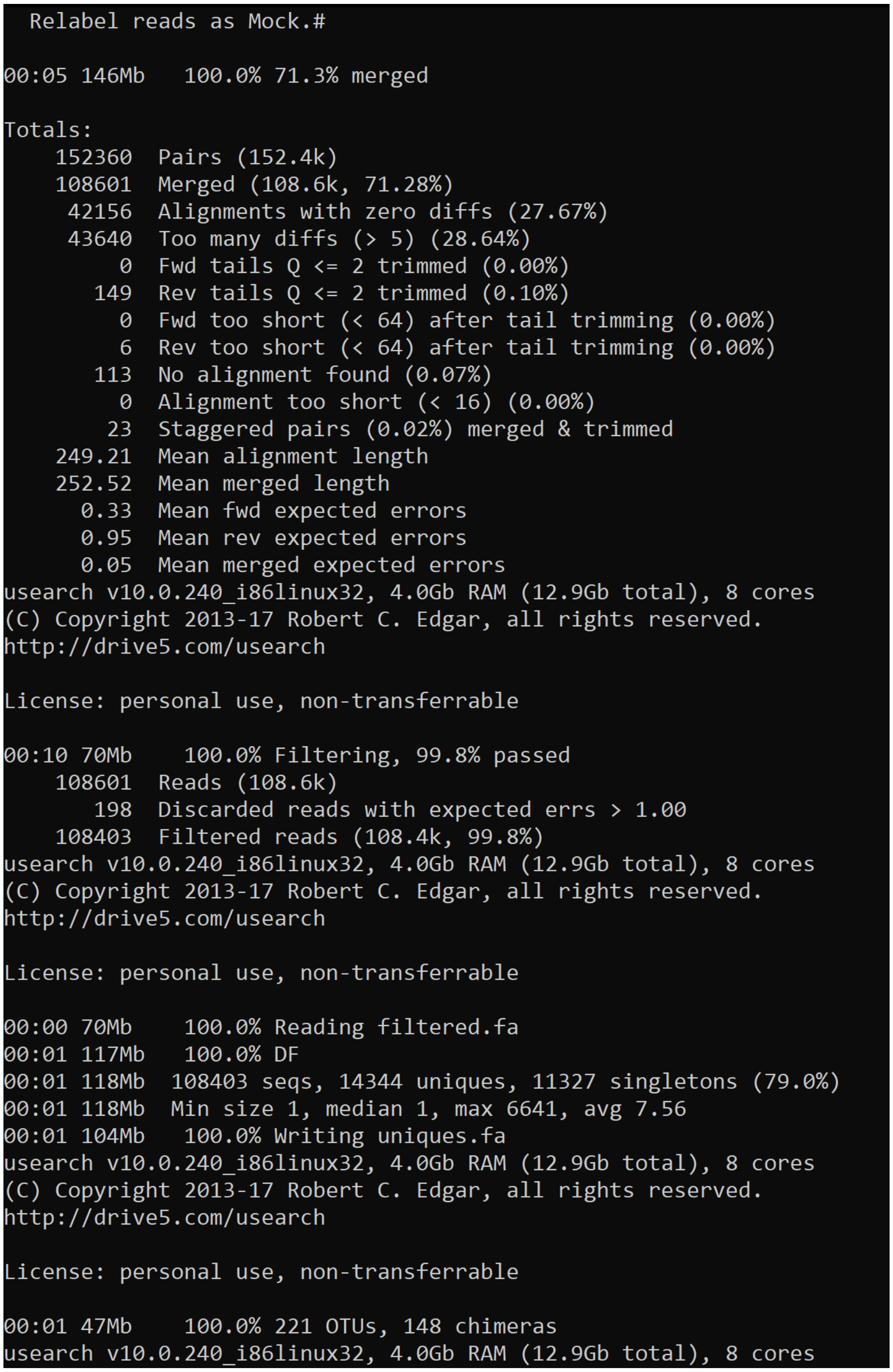
   ii. And comment out the following lines for run_mock.bash

```
version=`$usearch -version | sed "-es/usearch //" | sed "-es/v10.*/v10/"`

if [ x$version != xv10 ] ; then
    echo "Wrong usearch version, need v10" >> /dev/stderr
    exit 1
fi
```

   iii. You may also get this the “wrong version” error if the current user does not have the correct permissions to run the program. This can be fixed with chmod +x run.bash

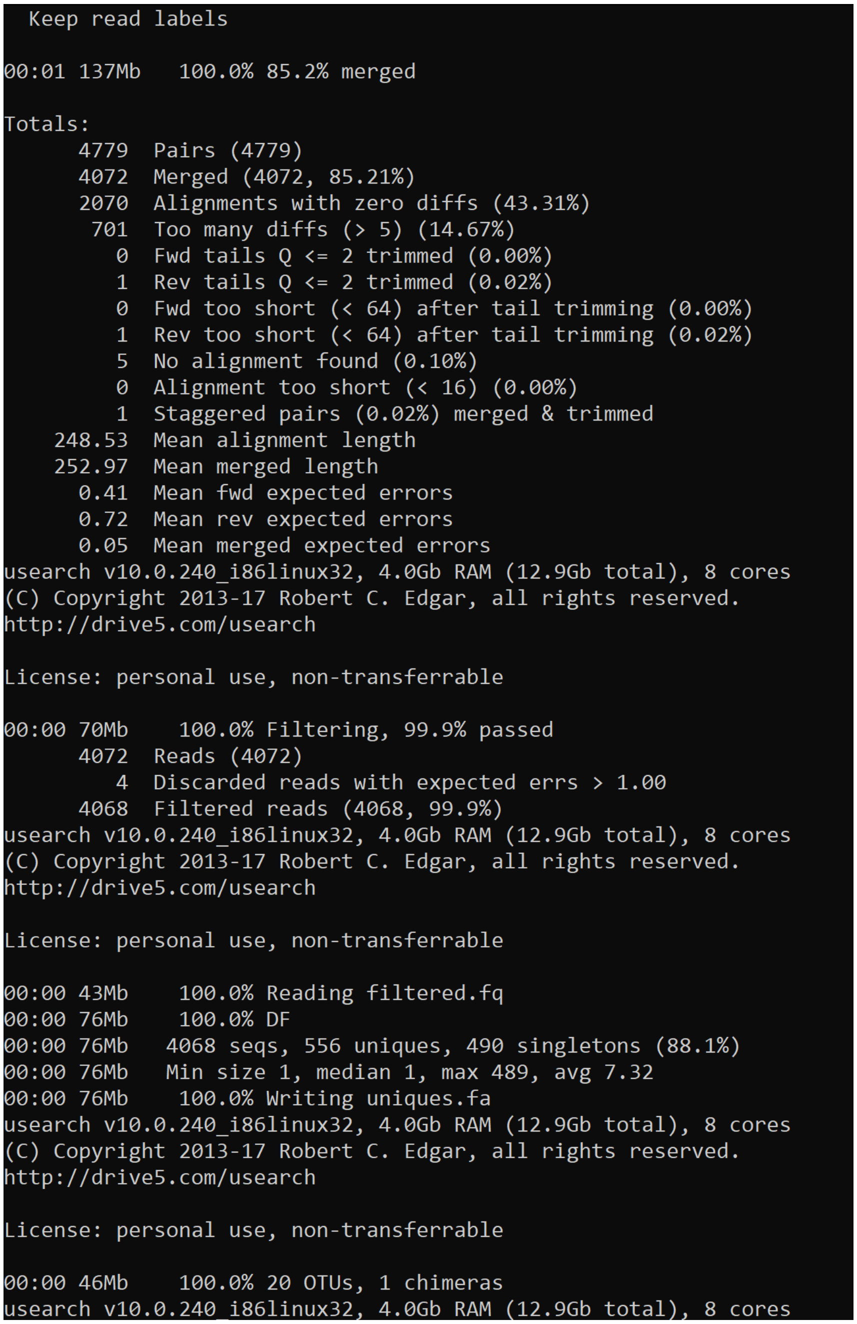
7. After running the run.bash and run_mock.bash files a lot of information should be shown in the Terminal. The following are two images showing some of that information.

   a. This image shows a part of the output of the run.bash script



```
  Relabel reads as Mock.#

00:05 146Mb   100.0% 71.3% merged

Totals:
    152360  Pairs (152.4k)
    108601  Merged (108.6k, 71.28%)
     42156  Alignments with zero diffs (27.67%)
     43640  Too many diffs (> 5) (28.64%)
         0  Fwd tails Q <= 2 trimmed (0.00%)
       149  Rev tails Q <= 2 trimmed (0.10%)
         0  Fwd too short (< 64) after tail trimming (0.00%)
         6  Rev too short (< 64) after tail trimming (0.00%)
       113  No alignment found (0.07%)
         0  Alignment too short (< 16) (0.00%)
        23  Staggered pairs (0.02%) merged & trimmed
    249.21  Mean alignment length
    252.52  Mean merged length
      0.33  Mean fwd expected errors
      0.95  Mean rev expected errors
      0.05  Mean merged expected errors
usearch v10.0.240_i86linux32, 4.0Gb RAM (12.9Gb total), 8 cores
(C) Copyright 2013-17 Robert C. Edgar, all rights reserved.
http://drive5.com/usearch

License: personal use, non-transferrable

00:10 70Mb    100.0% Filtering, 99.8% passed
    108601  Reads (108.6k)
       198  Discarded reads with expected errs > 1.00
    108403  Filtered reads (108.4k, 99.8%)
usearch v10.0.240_i86linux32, 4.0Gb RAM (12.9Gb total), 8 cores
(C) Copyright 2013-17 Robert C. Edgar, all rights reserved.
http://drive5.com/usearch

License: personal use, non-transferrable

00:00 70Mb    100.0% Reading filtered.fa
00:01 117Mb   100.0% DF
00:01 118Mb  108403 seqs, 14344 uniques, 11327 singletons (79.0%)
00:01 118Mb  Min size 1, median 1, max 6641, avg 7.56
00:01 104Mb   100.0% Writing uniques.fa
usearch v10.0.240_i86linux32, 4.0Gb RAM (12.9Gb total), 8 cores
(C) Copyright 2013-17 Robert C. Edgar, all rights reserved.
http://drive5.com/usearch

License: personal use, non-transferrable

00:01 47Mb    100.0% 221 OTUs, 148 chimeras
usearch v10.0.240_i86linux32, 4.0Gb RAM (12.9Gb total), 8 cores
```

b. This image shows a part of the output of the run_mock.bash script



```
  Keep read labels

00:01 137Mb   100.0% 85.2% merged

Totals:
      4779  Pairs (4779)
      4072  Merged (4072, 85.21%)
      2070  Alignments with zero diffs (43.31%)
       701  Too many diffs (> 5) (14.67%)
         0  Fwd tails Q <= 2 trimmed (0.00%)
         1  Rev tails Q <= 2 trimmed (0.02%)
         0  Fwd too short (< 64) after tail trimming (0.00%)
         1  Rev too short (< 64) after tail trimming (0.02%)
         5  No alignment found (0.10%)
         0  Alignment too short (< 16) (0.00%)
         1  Staggered pairs (0.02%) merged & trimmed
    248.53  Mean alignment length
    252.97  Mean merged length
      0.41  Mean fwd expected errors
      0.72  Mean rev expected errors
      0.05  Mean merged expected errors
usearch v10.0.240_i86linux32, 4.0Gb RAM (12.9Gb total), 8 cores
(C) Copyright 2013-17 Robert C. Edgar, all rights reserved.
http://drive5.com/usearch

License: personal use, non-transferrable

00:00 70Mb    100.0% Filtering, 99.9% passed
      4072  Reads (4072)
         4  Discarded reads with expected errs > 1.00
      4068  Filtered reads (4068, 99.9%)
usearch v10.0.240_i86linux32, 4.0Gb RAM (12.9Gb total), 8 cores
(C) Copyright 2013-17 Robert C. Edgar, all rights reserved.
http://drive5.com/usearch

License: personal use, non-transferrable

00:00 43Mb    100.0% Reading filtered.fq
00:00 76Mb    100.0% DF
00:00 76Mb   4068 seqs, 556 uniques, 490 singletons (88.1%)
00:00 76Mb   Min size 1, median 1, max 489, avg 7.32
00:00 76Mb    100.0% Writing uniques.fa
usearch v10.0.240_i86linux32, 4.0Gb RAM (12.9Gb total), 8 cores
(C) Copyright 2013-17 Robert C. Edgar, all rights reserved.
http://drive5.com/usearch

License: personal use, non-transferrable

00:00 46Mb    100.0% 20 OTUs, 1 chimeras
usearch v10.0.240_i86linux32, 4.0Gb RAM (12.9Gb total), 8 cores
```

8. Here are the script output locations for run.bash and run_mock.bash

    i. ./run.bash → misop/out

    ii. ./run_mock.bash → misop/out_mock

9. Now you can go through the tutorials one by one listed on the USearch site using the provided scripts for each located in the misop/exercises folder

a. The outputs for most of the tutorials will be in the misop/exout folder

b. To run each bash script you must currently be in the misop/exercises directly (via cd) and then can use the following command: ./ex#.bash to run each exercise, changing the # with the # exercise you are trying to run

10. Here are the script output locations for all of the exercises

    i. ./run ex1.bash → misop/exout/otus_sintax.txt & misop/exout/sintax_summary.txt

    ii. ./run ex2.bash → misop/exout/alpha.txt

    iii. ./run ex3.bash → misop/exout/alpha_5k.txt & misop/exout/otutab_5k.txt

    iv. ./run ex4.bash → misop/exout/otu20.fa & misop/exout/otu20_hit.uc

    v. Exercise 5 Uses NCBI BLAST on the OTU sequence stored as misop/exercises/out20.fa

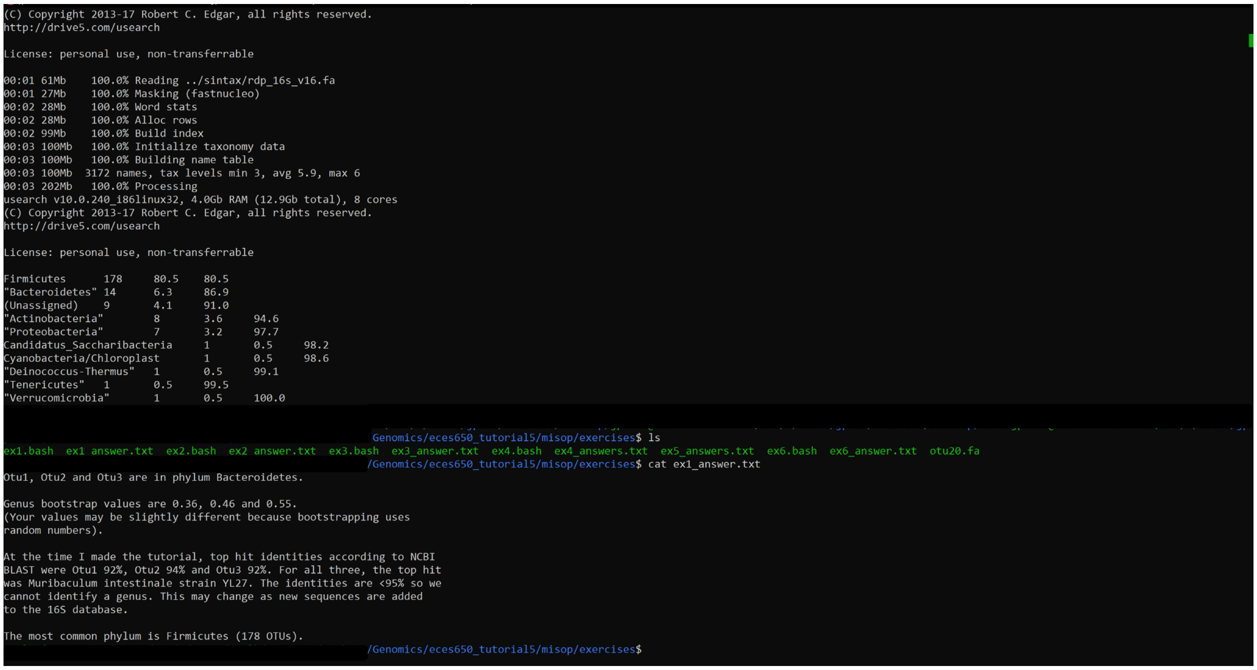
    vi. ./run ex6.bash → misop/exout/phylum_summary.txt

11. After running the provided scripts the results can be compared with the exercise answers text files also located in the misop/exercises folder

    a. There may be some minor differences due to the impact of random number generation, but overall the results shouldn't differ by much

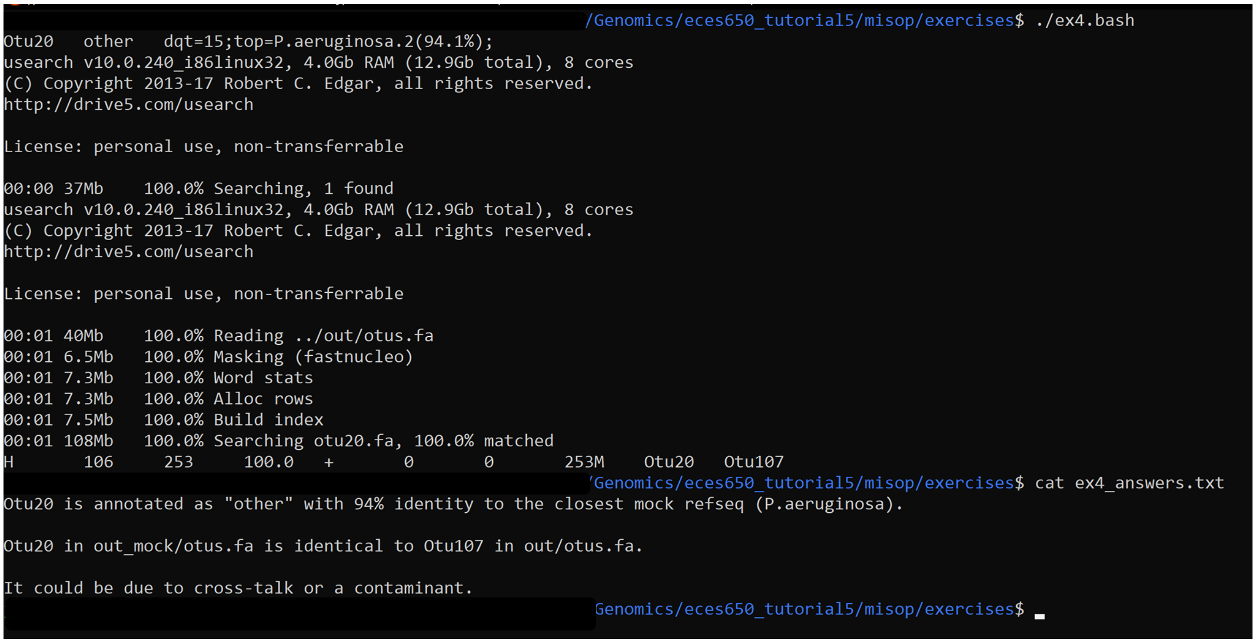
12. Below are some of the results from the scripts as they are shown in the Terminal. NOTE: Not all the script outputs are shown as some of the scripts outputs are very large or deposited directly into a file without being shown with the cat command (or both).

    a. This image is the Terminal output for the Exercise 1 script along with the information stored in the Exercise 1 Answer text file



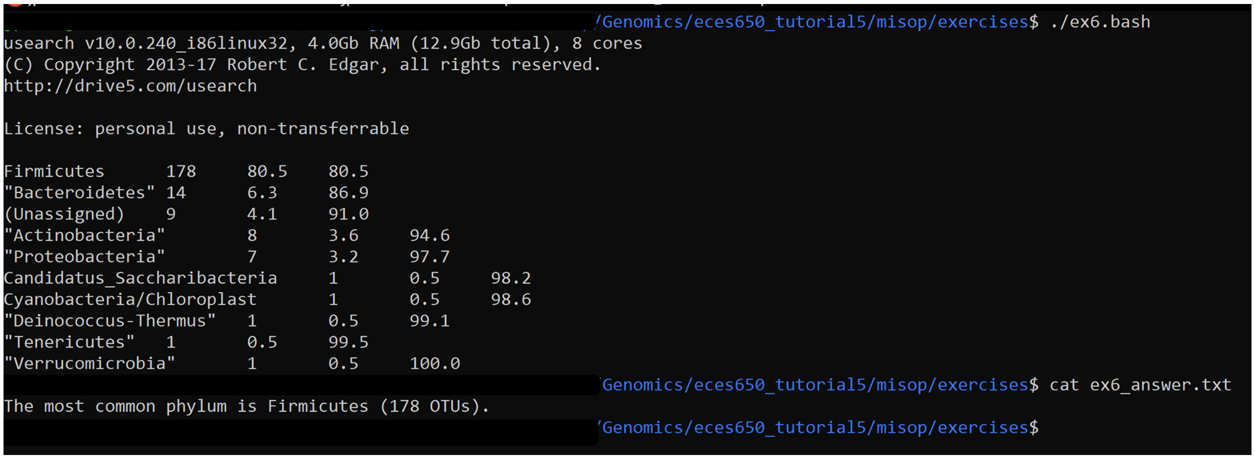
```
(C) Copyright 2013-17 Robert C. Edgar, all rights reserved.
http://drive5.com/usearch

License: personal use, non-transferrable

00:01 61Mb    100.0% Reading ../sintax/rdp_16s_v16.fa
00:01 27Mb    100.0% Masking (fastnucleo)
00:02 28Mb    100.0% Word stats
00:02 28Mb    100.0% Alloc rows
00:02 99Mb    100.0% Build index
00:03 100Mb   100.0% Initialize taxonomy data
00:03 100Mb   100.0% Building name table
00:03 100Mb  3172 names, tax levels min 3, avg 5.9, max 6
00:03 202Mb   100.0% Processing
usearch v10.0.240_i86linux32, 4.0Gb RAM (12.9Gb total), 8 cores
(C) Copyright 2013-17 Robert C. Edgar, all rights reserved.
http://drive5.com/usearch

License: personal use, non-transferrable

Firmicutes      178     80.5    80.5
"Bacteroidetes" 14      6.3     86.9
(Unassigned)    9       4.1     91.0
"Actinobacteria"        8       3.6     94.6
"Proteobacteria"        7       3.2     97.7
Candidatus_Saccharibacteria     1       0.5     98.2
Cyanobacteria/Chloroplast       1       0.5     98.6
"Deinococcus-Thermus"   1       0.5     99.1
"Tenericutes"   1       0.5     99.5
"Verrucomicrobia"       1       0.5     100.0
[redacted]Genomics/eces650_tutorial5/misop/exercises$ ls
ex1.bash  ex1 answer.txt  ex2.bash  ex2 answer.txt  ex3.bash  ex3_answer.txt  ex4.bash  ex4_answers.txt  ex5_answers.txt  ex6.bash  ex6_answer.txt  otu20.fa
[redacted]/Genomics/eces650_tutorial5/misop/exercises$ cat ex1_answer.txt
Otu1, Otu2 and Otu3 are in phylum Bacteroidetes.

Genus bootstrap values are 0.36, 0.46 and 0.55.
(Your values may be slightly different because bootstrapping uses
random numbers).

At the time I made the tutorial, top hit identities according to NCBI
BLAST were Otu1 92%, Otu2 94% and Otu3 92%. For all three, the top hit
was Muribaculum intestinale strain YL27. The identities are <95% so we
cannot identify a genus. This may change as new sequences are added
to the 16S database.

The most common phylum is Firmicutes (178 OTUs).
[redacted]/Genomics/eces650_tutorial5/misop/exercises$
```

b. This image is the Terminal output for the Exercise 4 script along with the information stored in the Exercise 4 Answer text file



```
[redacted]/Genomics/eces650_tutorial5/misop/exercises$ ./ex4.bash
Otu20   other   dqt=15;top=P.aeruginosa.2(94.1%);
usearch v10.0.240_i86linux32, 4.0Gb RAM (12.9Gb total), 8 cores
(C) Copyright 2013-17 Robert C. Edgar, all rights reserved.
http://drive5.com/usearch

License: personal use, non-transferrable

00:00 37Mb    100.0% Searching, 1 found
usearch v10.0.240_i86linux32, 4.0Gb RAM (12.9Gb total), 8 cores
(C) Copyright 2013-17 Robert C. Edgar, all rights reserved.
http://drive5.com/usearch

License: personal use, non-transferrable

00:01 40Mb    100.0% Reading ../out/otus.fa
00:01 6.5Mb   100.0% Masking (fastnucleo)
00:01 7.3Mb   100.0% Word stats
00:01 7.3Mb   100.0% Alloc rows
00:01 7.5Mb   100.0% Build index
00:01 108Mb   100.0% Searching otu20.fa, 100.0% matched
H       106     253     100.0   +       0       0       253M    Otu20   Otu107
[redacted]Genomics/eces650_tutorial5/misop/exercises$ cat ex4_answers.txt
Otu20 is annotated as "other" with 94% identity to the closest mock refseq (P.aeruginosa).

Otu20 in out_mock/otus.fa is identical to Otu107 in out/otus.fa.

It could be due to cross-talk or a contaminant.
[redacted]Genomics/eces650_tutorial5/misop/exercises$
```

c. This image is the Terminal output for the Exercise 6 script along with the information stored in the Exercise 6 Answer text file



```
[redacted]/Genomics/eces650_tutorial5/misop/exercises$ ./ex6.bash
usearch v10.0.240_i86linux32, 4.0Gb RAM (12.9Gb total), 8 cores
(C) Copyright 2013-17 Robert C. Edgar, all rights reserved.
http://drive5.com/usearch

License: personal use, non-transferrable

Firmicutes      178     80.5    80.5
"Bacteroidetes" 14      6.3     86.9
(Unassigned)    9       4.1     91.0
"Actinobacteria"        8       3.6     94.6
"Proteobacteria"        7       3.2     97.7
Candidatus_Saccharibacteria     1       0.5     98.2
Cyanobacteria/Chloroplast       1       0.5     98.6
"Deinococcus-Thermus"   1       0.5     99.1
"Tenericutes"   1       0.5     99.5
"Verrucomicrobia"       1       0.5     100.0
[redacted]Genomics/eces650_tutorial5/misop/exercises$ cat ex6_answer.txt
The most common phylum is Firmicutes (178 OTUs).
[redacted]/Genomics/eces650_tutorial5/misop/exercises$
```