### Introduction and class layout

**Files you will need for today:**

usearch (or usearch....win32.exe or the linux version)

You can download this from drive5.com if you register (best).

Three more files all on thumb drive:

oriented\_primers.stripped

pr2\_version\_4.11.1\_UTAX.fasta

SILVAngs\_User\_Guide\_2018\_03\_02.pdf

**Part one -- presentations and discussion**

Outreach and crowdsourcing

Scary bug example

The leaky pipes (I have the website)

An interesting article about potential troublesome eukaryotes in sewage

<https://link.springer.com/article/10.1007/s00248-017-0996-9>

Remember to show off my water bill pamphlet, talk about open house at IMET!

I end with pictures and a bit of lecture on approaches (local or server) and closed or to reference for OTU.

Then we go over what we did last week with our sample data.

**Part two -- Command line portion**

Then we use usearch -- drive5.com

I have linux, pc and mac versions, but it is easy to download and more accurate.

The sample files you will need are:

**oriented\_primers.stripped**

two samples all facing the same way with primers removed -- the reads file.

A reference file:

**pr2\_version\_4.11.1\_UTAX.fasta**

which I downloaded from the github for pr2 (so you could too) or it is on the drive.

We will only do a few commands, just to get a feel for a real world dataset.

**Part three -- Online analysis**

movie from Illumina, exploring results from arb-NGS platform (uploaded data). More on visualizing data with fingerprints and customizing output (see the SILVAngs\_User\_Guide\_2018\_03\_02.pdf). Also they claim I can share my output with you -- expect this to take some time to work out.

**Part four -- stories about reference sequences?**

Will show off a current analysis of morphotypes and ribotypes.

------------

### Opening presentations:

Fill in here

Outreach and crowdsourcing

Scary bug example

Leaky pipes

**My 'lecture'**

Remember to talk about importance of filter feeders too!

In theory you could **walk in with a laptop and walk out with a ‘stand alone’ analysis platform**.

Resources needed

1. Cloud computing space for larger genomic analyses
   1. Amazon Web Services (AWS)
   2. Linode
   3. Google Cloud Computing
2. Experiment design & Sequencing analysis
   1. James White, [Resphera](https://www.researchgate.net/profile/James_White21) [respherabio.com]
3. Genomics programs
   1. [Qiime2](https://qiime2.org/): reference data
      1. Analyze, share, and visualize genetic data
   2. Command line-based programs
      1. Usearch
      2. Mothur
4. Public data repositories -- ncbi, ebi

Presentation Notes

<https://fold.it/portal/>

<https://www.bluewaterbaltimore.org/>

Compendium of life in harbor / Maryland / USA

---------

### Usearch command line portion

usearch is a bit faster than the mothur pipeline.

Four window idea: You'll have to keep track of what you are running and there are often several windows worth of information to keep open.

Also you don't return to an internal command prompt but back to your DOS or unix prompt.

I have supplied you with a set of sequences (oriented\_primers.stripped) that have been 'oriented' and the primers stripped out.

then you need the program (usearch) and a reference file (pr2\_version\_4.11.1\_UTAX.fasta).

Here it is in list format in the directory

usearch

oriented\_primers.stripped

pr2\_version\_4.11.1\_UTAX.fasta

I got here from raw data with the usearch -orient, a little creative sequence collection to get those with full length primers, and usearch -fastx\_truncate to take away the primers.

We're going to do three basic steps here. First reduce unique sequences

1) **pooled-167-59:18S tsetso$ usearch -fastx\_uniques oriented\_primers.stripped -fastaout uniques\_both\_subsample.fasta -sizeout**

usearch v10.0.240\_i86osx32, 4.0Gb RAM (17.2Gb total), 8 cores

(C) Copyright 2013-17 Robert C. Edgar, all rights reserved.

http://drive5.com/usearch

License: bachvarofft@umces.edu

00:01 32Mb 100.0% Reading oriented\_primers.stripped

00:01 22Mb 100.0% DF

00:01 24Mb 91815 seqs, **21168** uniques, **18632** singletons (88.0%)

00:01 24Mb Min size 1, median 1, max 16401, avg 4.34

00:01 23Mb 100.0% Writing uniques\_both\_subsample.fasta

pooled-167-59:18S tsetso$

pooled-167-59:18S tsetso$ grep -c '>' uniques\_both\_subsample.fasta

21168

Second cluster with a minimum otu size of 10 -- 10 sequences per otu at least; all are at 97% identity or better

2) **pooled-167-59:18S tsetso$ usearch -cluster\_otus uniques\_both\_subsample.fasta -minsize 10 -otus otus.fa -uparseout uparse.txt**

usearch v10.0.240\_i86osx32, 4.0Gb RAM (17.2Gb total), 8 cores

(C) Copyright 2013-17 Robert C. Edgar, all rights reserved.

http://drive5.com/usearch

License: bachvarofft@umces.edu

00:00 8.5Mb 100.0% **162** OTUs, 0 chimeras

Ok so 162 OTU, let's try with a much lower number and overwrite our previous files

2a) **pooled-167-59:18S tsetso$ usearch -cluster\_otus uniques\_both\_subsample.fasta -minsize 2 -otus otus.fa -uparseout uparse.txt**

In most systems if you press the up arrow and adjust just the minsize parameter it is easiest.

usearch v10.0.240\_i86osx32, 4.0Gb RAM (17.2Gb total), 8 cores

(C) Copyright 2013-17 Robert C. Edgar, all rights reserved.

http://drive5.com/usearch

License: bachvarofft@umces.edu

00:01 9.9Mb 100.0% **445** OTUs, 12 chimeras

More OTU and more chimeras (mixed OTU). Let's go back to 10 (or not).

2b) **pooled-167-59:18S tsetso$ usearch -cluster\_otus uniques\_both\_subsample.fasta -minsize 10 -otus otus.fa -uparseout uparse.txt**

usearch v10.0.240\_i86osx32, 4.0Gb RAM (17.2Gb total), 8 cores

(C) Copyright 2013-17 Robert C. Edgar, all rights reserved.

http://drive5.com/usearch

License: bachvarofft@umces.edu

00:00 8.5Mb 100.0% 162 OTUs, 0 chimeras

Alright now we've written to the otus.fa file three times over!

Now we want to take our original set of sequences and map them to the otus we just created.

3) **pooled-167-59:18S tsetso$ usearch -otutab oriented\_primers.stripped -otus otus.fa -otutabout otutab\_out.txt**

usearch v10.0.240\_i86osx32, 4.0Gb RAM (17.2Gb total), 8 cores

(C) Copyright 2013-17 Robert C. Edgar, all rights reserved.

http://drive5.com/usearch

License: bachvarofft@umces.edu

00:00 1.5Mb 100.0% Reading otus.fa

00:00 1.5Mb 100.0% Masking (fastnucleo)

00:00 2.3Mb 100.0% Word stats

00:00 2.3Mb 100.0% Alloc rows

00:00 2.4Mb 100.0% Build index

03:32 20Mb 100.0% Searching, 91.4% matched

83917 / 91815 mapped to OTUs (**91**.4%)

03:32 20Mb Writing otutab\_out.txt

03:32 20Mb Writing otutab\_out.txt ...done.

Good success rate -- 91% of our reads match an OTU.

pooled-167-59:18S tsetso$ less otutab\_out.txt

or text editor. This is three columns, OTU name (actually the representative sequence, then number of times seen in the two samples).

This can be used directly in excel for example.

The next command allows you to put it into a taxonomic context

4) **pooled-167-59:18S tsetso$ usearch -sintax otus.fa -db pr2\_version\_4.11.1\_UTAX.fasta -tabbedout otus\_sintaxtable -strand both -sintax\_cutoff 0.8**

usearch v10.0.240\_i86osx32, 4.0Gb RAM (17.2Gb total), 8 cores

(C) Copyright 2013-17 Robert C. Edgar, all rights reserved.

http://drive5.com/usearch

License: bachvarofft@umces.edu

00:02 371Mb 100.0% Reading pr2\_version\_4.11.1\_UTAX.fasta

00:10 337Mb 100.0% Masking (fastnucleo)

00:19 310Mb 100.0% Word stats

00:19 310Mb 100.0% Alloc rows

00:37 1.3Gb 100.0% Build index

00:41 1.3Gb 100.0% Initialize taxonomy data

00:41 1.3Gb 100.0% Building name table

00:41 1.3Gb 70588 names, tax levels min 8, avg 8.0, max 8

00:48 1.3Gb 100.0% Processing

pooled-167-59:18S tsetso$ less otus\_sintaxtable

this takes a few minutes.

pooled-167-59:18S tsetso$ wc -l otus\_sintaxtable

162 otus\_sintaxtable

But this may be misleading -- have to double check assignments to be certain. Or actually parse the ‘scores’ to be sure.

What you could do -- or I would do is create a spreadsheet combining the reads from each sample per otu with the taxonomic information for our 162 OTU.

-------------

### Online analysis

A slick video about sequencing

<https://www.youtube.com/watch?v=fCd6B5HRaZ8>

Now we are going to go to

<https://www.arb-silva.de/ngs/>

These guys will want you to register to get an account. I can share my project with you then. This is the same project that we used for the usearch -- open ocean samples from Tara Oceans.

If you are reluctant I have all the files zipped up on my computer.

We will want to look (I think) especially at fingerprints and fiddle with the settings. This might be better led by someone else. The krona view is also hard to put down.

### Reference data

Update on current state of Dinophysis in the bay.

My dinos?

This is how the mothur blog describes rarefaction and reasoning for it.

“ Alas, rarefaction is not a measure of richness, but a measure of diversity. If you consider two communities with the same richness, but different evenness then after sampling a large number of individuals their rarefaction curves will asymptote to the same value. Since they have different evennesses the shapes of the curves will differ. Therefore, selecting a number of individuals to cutoff the rarefaction curve isn't allowing a researcher to compare samples based on richness, but their diversity. “

How Tsetso Made our Data

1. Went to look at Tara Oceans, a European BioInformatics Institute project, and downloaded sequences from there.
2. He “merged pairs” of complementary DNA sequences (strand 1: \_1, strand 2: \_2)
   1. DNA amplification copies **both** strands of DNA
   2. So, in theory, both strands would show up afterwards
3. He subsampled the ebi data, selecting only the first 100,000 (i.e., not randomly sampled). There are 2 files: ERR562560 & ERR562383
4. Then he “oriented” the pairs using “usearch”
   1. You “orient” against the database you have
5. Double checked the orientation to make sure he could find the primers are the beginning and the end of the sequences
6. Then he stripped the primers from the sequences so they would not affect estimates of “diversity.”
   1. Leaving the primers in would result in ~60 nucleotides that would match all the time, which we don’t want.
7. Then, he took a sample of that and shared it with us.