my checklist

thumb drive

double check

data

mothur

formattted utax database

utax fasta file

mothur windows

where is dos again?

utax fasta file -- is it compatible?

---- must formatdb in pc land

ancient pizza: blast/bin/formatdb -i ../pr2\_UTAX.fast -p F -o T

blast\bin\formatdb.exe -i pr2\_UTAX.fast -p F -o T

blast\bin\megablast.exe

part one

bit of background on the harbor project

use the aquarium presentation

the great sewer re-alignment

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

then switch to tara oceans

http://taraoceans.sb-roscoff.fr/EukDiv/

look at their renderings

what is their taxonomy?

what can we infer from the samples?

Test dataset: should have a go:

I downloaded

Station TARA\_048

ERS489336 an identifier

ERR1718062 a batch of sequences

ERR1718062.fastq is the file I put in the drive

This is a nice baby file so we should all work with this.

we want to look at this file in something like wordpad, notepad, textwrangler, text edit,

or in unix (for mac users Applications->Utilities->Terminal

WYSIWYG

open source software

\*\*\*\*often fairly disorienting to work on the command line. \*\*\*\*

Look at the format

look at wikipedia for a guide

https://en.wikipedia.org/wiki/FASTQ\_format

Take the first sequence and run blast on ncbi

https://blast.ncbi.nlm.nih.gov

choose 'nucleotide'

cut and paste a sequence into the window this will be the line that says ACTG...

...

What kind of results do you get?

Click through a bit

Define a couple of terms here -- % identity and distance

Basically what we will do is this process but somewhat faster and more in bulk.

Also since we are doing bulk we start to get our feet muddy with uniqueness and other stuff.

try running mothur

for mac go to terminal and cd into the directory

cd 18S/mothur

to run mothur type

./mothur

type help() and look at all the text

looks like this command.name(key=value,key2=value2)

make mothur angry by typing help ()

The remaining lines with parentheses in them are commands within mothur

you can type at the mothur command prompt

mothur > fastq.info(fastq=ERR1718062.fastq)

this converts fastq to fasta

mothur > unique.seqs(fasta=ERR1718062.fasta)

this looks for and clusters sequences that are unique

if you look at the file ending with .names (in a text editor) it looks sort of like a histogram

mothur > count.seqs(name=ERR1718062.names )

this makes a table for you

You could put the table into excel and sort or other things

mothur > pre.cluster(fasta=ERR1718062.unique.fasta, count=ERR1718062.count\_table, diffs=3)

this will now cluster your reads with 3 differences

and produce more count tables or count tables that take this clustering into account

This is a form of alpha (how many different things) and beta (their abundance) diversity.

How do we figure out what they are? What are the organisms?

In mac os you exit mothur and type the following command

\*We will have to work on getting things working in the pc world.\*

I have grabbed one of the databases for eukaryotic diversity

https://github.com/pr2database/pr2database

Another option

https://www.arb-silva.de

we can use a sequence comparison tool

This is run outside of mothur herself on the command line

in mac from the mothur directory we need to find blast and then tell it which input file

-i input

and which database -d file to use

blast/bin/megablast -i ERR1718062.unique.precluster.fasta -d pr2\_version\_4.11.1\_mothur.db -m 9

*In Linux, use blastn:*

*$ blastn -query ERR1718062.unique.precluster.fasta -db pr2\_version\_4.11.1\_mothur.db -outfmt 6 > blastn.out*

try again by adding an extra parameter

blast/bin/megablast -i ERR1718062.unique.precluster.fasta -d pr2\_version\_4.11.1\_mothur.db -m 9 -e 1e-20

or this is the best way to see what they are.

but if you

blast/bin/megablast -i ERR1718062.unique.precluster.fasta -d pr2\_UTAX.fast -D 3 > output\_blast

look carefully at this table you will find that many of these hits are incomplete

If you want cut and paste into excel

or try running without -D

We could've skipped everything else and simply run 'megablast' on all the input sequences.

Why would we want to avoid this?

blast/bin/megablast -i ERR1718062.unique.precluster.fasta -d pr2\_UTAX.fast -D 3 -e 1e-20

Now divide and conquer

split taxa into groups and look them up?

work with students to see what makes sense to them

Consider a general topic of eukaryotic evolution as a 'lecture' next time no

I would normally assign a group per person....

Or we could look at hazards like the Naeglaria?

Or both?

Usearch version of above?

Next idea is to use rc reads and a bigger dataset

Bridge too far? Yes

This is from

**https://mothur.org/wiki/MiSeq\_SOP**

change to a directory one level up

run mothur by typing

this directory has four fastq files (the source is detailed below)

\_1 and \_2 are reverse compliments of each other

These are much bigger files.

We can talk about some tricks here for handling these files.

Take the first one (or last) or any one and try running blast

mothur > make.file(inputdir=., type=fastq, prefix=stability)

then look at the file called 'stability.files'

mothur > make.contigs(file=stability.files)

It took 403 secs to assemble 2,469,613 reads.

It took 575 secs to assemble 3572313 reads.

Group count:

ERR562383 2469613

ERR562560 3572313

Total of all groups is 6,041,926

So it might take a while...

And so we look at what this spawned.

What are the output files?

Which ones would you want to use as input for the next steps?

The ones ending in .fasta for sure

we could then go on and run the series of commands from above

I may be missing a step in here, so look at the wiki https://mothur.org/wiki/MiSeq\_SOP

mothur >get.current()

mothur > summary.seqs()

or

mothur >summary.seqs(fasta=stability.trim.contigs.fasta)

you might or might not want to do quality control now as in the wiki

Just remember that the length of these sequences is in the range of 144-170 so you will want to adjust

your parameters

so for example the wiki says

mothur > screen.seqs(fasta=stability.trim.contigs.fasta, group=stability.contigs.groups, summary=stability.trim.contigs.summary, maxambig=0, maxlength=275)

and all you need to do is change the length to a value that fits our data --- something like 170 bases

mothur > screen.seqs(fasta=stability.trim.contigs.fasta, group=stability.contigs.groups, summary=stability.trim.contigs.summary, maxambig=0, maxlength=170)

old command

mothur > unique.seqs(fasta=ERR1718062.fasta)

new command

√mothur > unique.seqs(fasta=stability.trim.contigs.good.fasta)

Skipping a series of commands that are based on the alignment file.

....

Going back to the wiki

√mothur > count.seqs(name=stability.trim.contigs.good.names, group=stability.contigs.good.groups)

The wiki command

mothur > pre.cluster(fasta=stability.trim.contigs.good.unique.good.filter.unique.fasta, count=stability.trim.contigs.good.unique.good.filter.count\_table, diffs=2)

An example of one that works

mothur > pre.cluster(fasta=stability.trim.contigs.good.unique.fasta, count=stability.trim.contigs.good.count\_table , diffs=8)

But this takes a while....

•Vocabulary from this session

fastq and fasta format

OTU

eukaryotic taxonomy

clustering

?for next session

alpha & beta diversity formal definitions?

more on ecology

more on change in environment and hx of sewage?

Why do we want a pca/pcoa?

# 

# ----sample information for second batch of data----

http://taraoceans.sb-roscoff.fr/EukDiv/

Search results for ERS490473 DCM 180-2000 microns TARA\_845

Show more data from EMBL-EBI

Read

Experiment (5)

Run (5)

Sample

Sample (1)

Experiment (5 results found)

Showing results 1 - 5 of 5 results

ERX1782453 Illumina HiSeq 2000 paired end sequencing; environmental samples cDNA

ERX521720 Illumina Genome Analyzer IIx paired end sequencing; Marine... Library Construction Protocol

Library preparations from V9-18S rDNA amplicons

ERR562383\_1.fastq.gz

ERR562383\_2.fastq.gz

ERX1782161Illumina HiSeq 2000 paired end sequencing; environmental samples cDNA

ERX556277Illumina HiSeq 2000 paired end sequencing; Marine metagenome. DNA

ERX1788174Illumina HiSeq 2000 sequencing; Illumina HiSeq 2000 sequencing amplicon

Series of others from the same sample

looks like they took a DCM sample

same sample smaller size fraction

INSDC ACCESSION NUMBER ERS490503 5-20 microns

Remember this for troubleshooting making the blast db

Experiment: ERX521429Tsvetans-MacBook-Pro-2:mothur tsetso$ blast/bin/formatdb -

formatdb 2.2.25 arguments:

-t Title for database file [String] Optional

-i Input file(s) for formatting [File In] Optional

-l Logfile name: [File Out] Optional

default = formatdb.log

-p Type of file

T - protein

F - nucleotide [T/F] Optional

default = T

-o Parse options

T - True: Parse SeqId and create indexes.

F - False: Do not parse SeqId. Do not create indexes.

[T/F] Optional

default = F

-a Input file is database in ASN.1 format (otherwise FASTA is expected)

T - True,

F - False.

[T/F] Optional

default = F

-b ASN.1 database in binary mode

T - binary,

F - text mode.

[T/F] Optional

default = F

-e Input is a Seq-entry [T/F] Optional

default = F

-n Base name for BLAST files [String] Optional

-v Database volume size in millions of letters [Integer] Optional

default = 4000

-s Create indexes limited only to accessions - sparse [T/F] Optional

default = F

-V Verbose: check for non-unique string ids in the database [T/F] Optional

default = F

-L Create an alias file with this name

use the gifile arg (below) if set to calculate db size

use the BLAST db specified with -i (above) [File Out] Optional

-F Gifile (file containing list of gi's) [File In] Optional

-B Binary Gifile produced from the Gifile specified above [File Out] Optional

-T Taxid file to set the taxonomy ids in ASN.1 deflines [File In] Optional

Tsvetans-MacBook-Pro-2:mothur tsetso$

ERR562560\_1.fastq.gz

ERR562560\_2.fastq.gz

# For next week

Me

phylogeny with rRNA

include SSU rRNA V9

sample or read data on google ready to be done

reference alignment

Specific topics:

*Naeglaria*

The ‘realignment’ of the city’s septic system

Crowdsourcing, gaming, outreach for science -- example of HIV