Laboratory Journal

Cody Glickman

cody.glickman@ucdenver.edu

Beginning 27 February 2018

Contents

Pre 2018

To add directory to path in MAC

```
export PATH=/path_goes_here/folder:$PATH
```

Get HMM Contig IDS

```
sed ’s/ˆ>.*\>/>/g’ $f | grep -o ’ˆ\S*’ | grep ’>’ | sed ’s/>//g’ | uniq > hmm_headers_$f.txt
```

File3 will contain items not in control (count lines)

```
comm -23 file1 file2 > file3
wc -l file3
```

Count number of Bases in Fasta

```
cat non_paired.fastq | paste - - | cut -f 2 | tr -d ’\n’ | wc -c
wc -l file3
```

Convert fastq to fasta

```
cat test.fastq | paste - - - - | sed ’s/ˆ@/>/g’| cut -f1-2 | tr ’\t’ ’\n’ > Output.fasta
```

BAM file to fasta

```
samtools view filename.bam | awk ’{OFS="\t"; print ">"$1"\n"$10}’  > filename.fasta
```

To filter significant contigs by headers (in QIIME)

```
filter_fasta.py -f final.contigs.fa -o test.fasta -s Headers.txt
grep -vFf file_with_patterns other_file
```

Assembly of significant reads

```
for f in *.fastq; do megahit -r $f -o ${f%%.*}; done;
```

Align reads back to contigs

```
bwa index contigs.fasta
bwa aln -t NUMBER_OF_THREADS contigs.fasta short_reads.fastq > alignment.sai
bwa samse contigs.fasta alignment.sai short_reads.fastq > alignment.sam
```

Split Combined Fastq File

```
awk ’{++count ; if (count<=4) {print > "F3.fastq"} else {print > "R3.fastq" } if (count==8) count=
```

Filter reads that mapped to contigs

```
samtools view -b -F 4 file.bam > file_unmapped.bam
```

Trim Reads

```
fastx_trimmer -Q33 -f [Length] -i input.fastq -o output.fastq
```

GreenGenes Database Level Munging

```
sed ’s/Other//g’ test3.txt > removed_other.txt
sed ’s/[a-z]__//g’ removed_other.txt > removed_taxonomy_levels_test2.txt
sed ’s/[;]\{2,\}//g’ removed_taxonomy_levels_test2.txt > middle_1.txt
sed ’s/$;//g’ middle_1.txt > middle_2.txt
sed ’s/\[//g;s/\]//g’ removed_end_semi_colons_with_counts.txt > middle_3.txt
sed ’s/ //g’ middle_3.txt > middle_4.txt
```

Run Kracken kmer indentifier

```
kraken --db /Users/stronglab2/Downloads/minikraken_DB/ final.contigs.fa > sequences.kraken
kraken-translate --db /Users/stronglab2/Downloads/minikraken_DB/ sequences.kraken > sequences.labe
```

Kraken Viral Data Processing

```
cat $f/sequences.labels | grep Viruses > $f/viral_labels.txt
sed ’s!.*;!!’ $f/viral_labels.txt > $f/viral_species.txt
uniq -u $f/viral_species.txt > unique_species_$f.txt
```

Randomly Sample Fasta File

```
reformat.sh in=input.fasta out=output.fasta samplereadstarget=[number of reads]
```

Random Shuffling and Spliting Fasta

```
seqkit shuffle [input.fa] --two-pass > output.shuffled.fa
partition.sh in=shuffled_ebola_reads.fasta out=ebola_part%.fasta ways=36
```

Simulate metagenome / Create Coverage Map / Stats for Synthetic Metagenome

```
randomreads.sh ref=Synthetic_Metagenome_Genomes.fasta out=reads.fastq reads=10M paired metagenome
bbmap.sh ref=Combined_synthetic_Metagenome.fasta in=reads.fastq covstats=covstats.txt scafstats=sc
stats.sh in=$f/final.contigs.fa out=$f/assembly_stats_$f.txt
```

Regex to get contain lengths file

```
sed ’s/.*\=//g’ final.contigs.fa | grep ’[0-9]’ > test.txt
```

Kmerize Single Genome

```
pyfasta split -k 848 -o 0 Input.fasta -n 1
```

Week of 26 February 2018

1 Identifying Virulence Factors in Phages

Downloaded phage protein data searching by taxonomy from Uniprot/TrEmbL. Dowloaded three datasets Claudioviruses, Ligamenvirales, and Unclassified. Downloaded viral contigs from vHMM Earth Virome Project. Metagenomic gene prediction through prodigal downloaded from conda ”conda install prodigal”

```
prodigal -i mVGs_sequences_v2.fna -o my_genes -a my_proteins.faa -p meta
```

The myproteins.faa file contains the translated predicted genes. This set is applied to the VF HMM and similar to the other datasets, returned no hits with hmmsearch.

Creating virulence factor blast database to blast against viral contigs. Choosing a E value threshold:
Goals Finish written GRAB and Abstract for NLM Done (3/2/18)

```
makeblastdb -in Combined_VF.faa -dbtype prot -out Combined_VF -title "Combined_VF"
```

```
blastp -db /Users/stronglab2/blastdb/Combined_VF/Combined_VF -out results.txt
-outfmt 6 -query phage_proteins.faa
```

```
## Hello World
print(x)
```

Week of 5 March 2018

Goals

- Prophage Annotation and VF/ARG Pipeline (Wednesday)
- ML Pipeline Active Prophage
- CAMI Data With Conda for Reproducibility

1 Identifying Virulence Factors in Phages

Establishing the BASH pipeline

```
1. Prophage Prediction
Input: Contigs.fasta
Output: Prophage Zip Folder

2. Gene Prediction
Input: Sequences of Identified Prophages
Output: Protein Fasta

3. Virulence Factor Identification
Input: Protein Fasta
Output: Proteins called Virulence Factors
```

2 Lysogenic Pan Genome

Downloaded phage table from PhageDB. Parse temperate phages from Graham Hatful's List and those that infect Mycobacterium. Save Genbank ID numbers as Numbers.txt in script PhageDBProcessing.Rmd. Calling GBK Files from nuccore.

```
## Load GenBankIds and Remove Whitespace
with open("Numbers.txt") as f:
    content = f.readlines()
content = [x.strip() for x in content]

## Call Entrez for Genbank_Ids
for i in range(0,len(content)):
    handle = Entrez.efetch(db="nuccore", id=content[i], rettype="gb")
    filename = 'genbank_files/'+ content[i] + '.gbk'
    out_handle = open(filename, "w")
    out_handle.write(handle.read())
    out_handle.close()
    handle.close()
    print("Saved " + filename)
```

Run Core Genome Analysis

Convert to GFF3

`bp_genbank2gff3.pl --dir pathtofiles`

Run Roary

`roary -e --maft -p 8 *.gff`

Week of 12 March 2018

1 Genomic Retrieval and BLAST Database Creation

Submitted GRAB document to bioRxiv: received resubmission request
Resupply manuscript to MS ID 246553
Include data about functionality compared to other tools or show with an example dataset

Week of 19 March 2018

1 Classifying predicted prophages as active or degraded

Met Monday with Graham Hatfull and members of his lab. Confirmed genes to predict lysogenic life cycle. Holins are hard to predict as they are transmembrane proteins and may have small amount of conservation. Holins may be better predicted by k-mer protein groups.

- Integrases
  Involved in phage insertion into host genome
- ParA - ParB - ParS
  Involved in extra chromosomal arrangement and replication
- Repressors
  Inhibits replication until stimulus

Additional Take-Aways: Portal genes may be targets of bacterial resistance to phages and Abscessus excised phages can be engineered to become lytic. Excised phages resemble cluster P or N, Abscessus infected by cluster A3

Review Kclust and MMseq2 for clustering sequences

Check for non synonomous mutations in presence of clustered gene

Installed MMSeqs2 for sequence clustering
Downloaded viral HMMs from EggNog 4.5 — VOGDB — pVOG

Process to Filter Viral Lysogeny HMMs from Pfam
Split the complete Pfam database and rename each file to the Pfam ID

```
python -c "import sys
for i, c in enumerate(sys.stdin.read().split('//')):
    open('out' + str(i), 'w').write(c)" < Pfam.hmm
```

Removed empty first space from Split:

```
for f in *; do grep . $f > $f.hmm; done;
```

Rename HMMs by Families Identifiers (3rd line of HMMs):

```
for f in *.hmm;
do
output=$(awk 'NR == 3 {print $2}' "$f" | cut -f1 -d '.')
mv "$f" "$output".hmm
done
```

Filter HMMs by those present in list:

```
while read file
do
mv -v -i "$file".hmm matched/
done < File_with_names.txt
```

Format filtered hmms for search against combined lysogenic profiles
Note: Needed to add ending '//' to hmm files to hmmpress effectively
Perform HMM Search and Create Table Separated

```
hmmsearch --tblout [table].txt [model].hmm [sequences].fasta > /dev/null

## Convert tblout to table
chmod +x convert_hmm_tblout_to_tab_seperated_table.sh

./convert_hmm_tblout_to_tab_seperated_table.sh -t [table].txt
```

2 Creation of rpoB, HSP65, and 16S Databases

Query:

```
(("Mycobacterium"[Organism] OR ("Mycobacterium"[Organism]
OR Mycobacterium[All Fields])) AND rpoB[Title]) AND
(bacteria[filter] AND biomol_genomic[PROP] AND
ddbj_embl_genbank[filter] AND ("500"[SLEN] : "5000"[SLEN]))
```

Muscle Sequence aligner downloaded via bioconda
usage: More usage examples at Drive5

```
muscle -msf -in seqs.fa -out seqs.msf
```

Week of 26 March 2018

1 Classifying predicted prophages as active or degraded

Thought: determine if focued kmer profile for mycobacteriophage would be ammendable.

```
for f in Pfam_Viral_HMMs/*.hmm; do filename=$(echo "${f\%\%.*}");
hmmsearch --tblout $filename.txt $f combined_lysogenic_phages_proteins.fasta >
/dev/null; done;
```

Move output to new folder titled SearchOutput and move into active directory

```
for f in Search_Output/*.txt;
do ./convert_hmm_tblout_to_tab_seperated_table.sh -t $f; done;
```

Week of 9 April 2018

1 Genomic Retrieval and BLAST Database Creation

Update Database Process

- Download NCBI Linages from Here as of March 12th.
- Download Assembly Summary

  ```
  wget ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/assembly_summary.txt
  ```

- Filter Taxa Linage by Kingdom

  The script Filtertaxlinage.R filters both the NCBI taxa lineages and the assembly summary files to include relevant information. **Note:** included no.rank2 as column in taxa lineages to include strain level identity.

Updating GRAB to include filtering by Strain Level

Week of 16 April 2018

1 Identifying Virulence Factors in Phages

Blasted VF Database against Self to identify redundant records between VFDB and Patric VF

```
blastp -db /Users/stronglab2/blastdb/Combined_VF/Combined_VF -out results.txt
 -outfmt 6 -query all_viral_protein.faa
```

Week of 23 April 2018

1 Identifying Virulence Factors in Phages

Established baseline presence of virulence genes in phage genomes.
Next steps: find number of virulence factors in vHMM contigs and perform Chi-square testing.

Week of 30 April 2018

1 Creation of rpoB, HSP65, and 16S Databases

Downloading rpoB data from NCBI Nucleotide (Copy and Paste into Search Window)

1780 Sequences as of May 1st, 2018

```
(("Mycobacterium"[Organism] OR ("Mycobacterium"[Organism]
OR Mycobacterium[All Fields])) AND rpoB[Title]) AND
(bacteria[filter] AND biomol_genomic[PROP] AND
ddbj_embl_genbank[filter] AND ("500"[SLEN] : "5000"[SLEN]))
```

Downloading HSP65 data from NCBI Nucleotide (Copy and Paste into Search Window)

1710 Sequences as of May 1st, 2018

```
(("Mycobacterium"[Organism] OR ("Mycobacterium"[Organism]
OR Mycobacterium[All Fields])) AND hsp65[Title]) AND
(bacteria[filter] AND biomol_genomic[PROP] AND
ddbj_embl_genbank[filter] AND ("400"[SLEN] : "5000"[SLEN]))
```

Downloading Silva-ARB Database: download SilvaSSUParctaxsilvatrunc.fasta.gz

See Silva Release information

Week of 7 May 2018

1 Identifying Virulence Factors in Phages

Compile against virome niches

```
module load megahit
for f in *; do megahit -1 $f/*_1.fastq -2 $f/*_2.fastq -o $f/megahit_out;done;
```

Week of 14 May 2018

NLP Twitter Info

```
consumer_key = "KjwOs3xP0egOG4zKzhAOIdrhP"
consumer_secret = "UzV7gRWvNtxLwj4u7Lgn3afhRwnkEk3MieEPaMXcJCCoUKaq5D"
access_token = "4784605044-ulRZNSGDGjeTyponvwK26TqimbC6ayYcmptQGnX"
access_secret = "D8yPVlZBXS2D79WLo8wx31YwqYH97S1bxIWUTdHHpt6Xw"
```

1 Identifying Virulence Factors in Phages

Updating HMM Filtering Protocol
Download Domains

- Get headers from files
- Add header

  ```
  awk '{ print $0 ".hmm" }' < Headers.txt > hmm_headers.txt
  ```

- Filter PFam Models

  ```
  for file in $(cat hmm_headers.txt); do mv $file Virulence/$f; done
  ```

- Added backslashes for hmmpress and hmmsearch

  ```
  for f in *.hmm; do echo -n "//" >> $f; done;
  cat *.hmm > PFAM_Combined.hmm
  ```

Week of 11 June 2018

1 BRASS Asthmatic Microbiome Study

BRASS Samples vary large, subsampled largest paired reads to shrink size of intermediates

```
reformat.sh in1=reads1.fq.gz in2=reads2.fq.gz
out=sampled1.fq.gz out2=sample2.fq.gz samplereadstarget=100000000
```

BRASS Samples too large to complete run on local PC, running QIIME Dada2 script on cluster
Making a manifest file for Qiime Demultiplexing

```
find /path/to/data -type f \( -iname "*.fastq.gz" \) >>
/path/to/manifest_prelim.txt
```

Then run manifest spliter.py
Running Qiime on cluster

```
module load qiime2
source activate qiime2-2017.12
```

Dada2 processing removing a majority of reads (18S?!) trying a less stringent thresholding for BRASS study

Week of 25 June 2018

1 BRASS Asthmatic Microbiome Study

The dada2 filtering using qiime implemented max-ee filter which removed a large percent of the reads. Went from 95K to 6K. Removed MaxEE by setting to –p-max-ee Inf

Rerunning BRASS Tables

```
module load qiime2
source activate qiime2-2017.12

bsub qiime dada2 denoise-paired --i-demultiplexed-seqs paired-end-demux.qza
--p-trunc-len-f 280 --p-trunc-len-r 279 --p-max-ee Inf --p-chimera-method
pooled --p-n-threads 0 --o-representative-sequences rep-seqs2.qza
--o-table table2.qza
```

Job failed to finish, most likely due to the large amounts of memory required, trying to run on the fat node now. May need to split data into multiple parts and run in parallel.
TO-DO Thursday:

Week of July 2018

Run assembly through Phaster (filter ¿1500 base pairs)

filter_contigs $ffiltered$_f.fasta –min_contig 1499

Number of Contigs (Make a table?) 143 1266 164 52 51

Number of Contigs Greater than 150 107 421 124 37 39

Week of July 2018

```
vsearch -cluster_fast  $inputfile  -id 0.995 \
-centroids    ${filename}.centroids \
-uc           ${filename}.clusters \
-consout      ${filename}.consesus \
-alnout       ${filename}.aln \
-clusters     $clustdir/${filename}.c- \
-msaout       $msadir/${filename}.c- \
```

Week of September 10 2018

Squares (analysis)
Projects in motion:

Projects	Status	Goals (To-Do)	Done This Week	Comments
Hawaiian Soils	Drafting Manuscript	Submit		Waiting on 3 samples
Duobiome	Testing (Formulating Pipeline)	Establish Pipelines on Github		Pam is creating testing dataset
vHMM	Development	Null Model — Truncate HMM		
GRAB	Development	Package Development		Incorporate SequenceServer