Appendix

Kraken2

BUILD FUNGI DATABASE

kraken2-build --download-taxonomy --db $DBNAME

kraken2-build --download-library fungi --db $DBNAME

(assembly_summary.txt generated - assembly levels ‘full’ and major)

alternatives:
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/fungi/assembly_summary.txt
or
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/fungi/assembly_summary.txt

Download latest version of all RefSeq/Genbank fungal sequences

List FTP directory paths from assembly_summary.txt in new file ftpdirpaths

awk -F "\t" '$11=="latest"{print $20}' assembly_summary.txt > ftpdirpaths

Append file extension genomic.fna.gz to all FTP directory names

awk 'BEGIN{FS=OFS="/";filesuffix="genomic.fna.gz"}
{ftpdir=$0;asm=$10;file=asm"_"filesuffix;print ftpdir,file}' ftpdirpaths > ftpfilepaths

Download all fna.gz files from list in ftpfilepaths, output to references directory

wget -i $DBNAME/library/fungi/references

Move additionally downloaded ATCC genomes to same references directory

Decompress downloaded fna.gz files to get fna files

gunzip references/*fna.gz

(or submit as script)

Add all fna/fasta files in references directory to library

find references/ -name '*.fna' -print0 | xargs -0 -I{} -n1 kraken2-build --add-to-library {} --db $DBNAME

Build database

kraken2-build --build --db $DBNAME

PREPROCESSING SEQUENCE DATA

concatenate all fastq files from each barcode

cat *.fastq > output.fastq

convert fastq files to fasta files

sed -n '1~4s/^@/>/p;2~4p' $FILE > $FILE.fasta

trim adapters with Porechop

porechop --min_split_read_size 400 -i $INPUT.fasta -o $OUTPUT.fasta

discard reads with lengths under 500 bases

seqtk seq -L 500 $INPUT.fasta > $OUTPUT.fasta

CLASSIFY/ASSIGN TAXONOMY

classify sequence reads (fasta format) and generate report with read abundances and standard output

kraken2 --db $DBNAME $QUERY.fasta --use-names --report $QUERY_OUTPUT.report.txt --output $OUTPUT_PATH

ABUNDANCE ESTIMATION AND DIVERSITY COMPUTATIONS

estimate species abundances and generate Bracken reports using Kraken2 report.txt files

bracken -d $DBNAME -i $KRAKEN2_OUTPUT.report.txt -o $OUTPUT.bracken -w $OUTPUT.breport -r 500 -l S -t 10

calculate alpha diversity using Diversity Tools python script and standard bracken output

python KrakenTools/DiversityTools/alpha_diversity.py -f $BRACKEN_OUTPUT.bracken -a BP

python KrakenTools/DiversityTools/alpha_diversity.py -f $BRACKEN_OUTPUT.bracken -a Sh

python KrakenTools/DiversityTools/alpha_diversity.py -f $BRACKEN_OUTPUT.bracken -a F

python KrakenTools/DiversityTools/alpha_diversity.py -f $BRACKEN_OUTPUT.bracken -a Si

calculate beta diversity and generate matrix using Diversity Tools python script and all standard bracken output files as shown for 5 files:

python KrakenTools/DiversityTools/beta_diversity.py -i $BRACKEN_OUTPUT_1.bracken $BRACKEN_OUTPUT_2.bracken $BRACKEN_OUTPUT_3.bracken $BRACKEN_OUTPUT_4.bracken $BRACKEN_OUTPUT_5.bracken

Scripts

Build DB

List FTP directory paths from assembly_summary.txt in new file ftpdirpaths

awk -F "\t" '$11=="latest"{print $20}' assembly_summary.txt > ftpdirpaths

---

Append file extension genomic.fna.gz to all FTP directory names

awk 'BEGIN{FS=OFS="/";filesuffix="genomic.fna.gz"}{ftpdir=$0;asm=$10;file=asm"_"filesuffix;print ftpdir,file}' ftpdirpaths > ftpfilepaths

---

*make ‘references’ directory to keep downloaded fna files

---

ftp_download.sh

```
#!/bin/sh

#PBS -l walltime=72:00:00
#PBS -l select=1:ncpus=16:mem=32gb

wget -i /rds/general/project/fisher-aspergillus-analysis/live/clarisse/dog_ear_kraken2_scripts/kraken2_fungi_db/library/fungi/ftpfilepaths -P /rds/general/project/fisher-aspergillus-analysis/live/clarisse/dog_ear_kraken2_scripts/kraken2_fungi_db/library/fungi/references
```

---

*gunzip *.fna.gzfiles

---

add_library.sh

```
#!/bin/sh

#PBS -l walltime=48:00:00
#PBS -l select=1:ncpus=16:mem=32gb


## OTHER OPTIONAL PBS DIRECTIVES

module load anaconda3/personal
source activate kraken2_env


find /rds/general/project/fisher-aspergillus-analysis/live/clarisse/dog_ear_kraken2_scripts/kraken2_fungi_db/library/fungi/references -name '*.fna' -print0 | xargs -0 -I{} -n1 kraken2-build --add-to-library {} --db /rds/general/project/fisher-aspergillus-analysis/live/clarisse/dog_ear_kraken2_scripts/kraken2_fungi_db/
```

---

kraken2_build_fungi.sh

```
#!/bin/sh

#PBS -l walltime=48:00:00
#PBS -l select=1:ncpus=16:mem=32gb

## OTHER OPTIONAL PBS DIRECTIVES

module load anaconda3/personal
source activate kraken2_env

kraken2-build --build --db /rds/general/project/fisher-aspergillus-analysis/live/clarisse/
dog_ear_kraken2_scripts/kraken2_fungi_db/
```

Classify/assign

query and output name list format (dog_kraken2_list.txt):

```
/rds/general/project/fisher-aspergillus-rawdata/live/clarisse/dog_ear_fasta/Alternaria_3,5Kb.fasta
Alternaria_3,5Kb
/rds/general/project/fisher-aspergillus-rawdata/live/clarisse/dog_ear_fasta/Alternaria_6Kb.fasta
Alternaria_6Kb
/rds/general/project/fisher-aspergillus-rawdata/live/clarisse/dog_ear_fasta/Aspergillus_3,5Kb.fasta
Aspergillus_3,5Kb
/rds/general/project/fisher-aspergillus-rawdata/live/clarisse/dog_ear_fasta/Aspergillus_6Kb.fasta
Aspergillus_6Kb
```

---

run kraken2 (dog_kraken2.sh):

```
#!/bin/sh

#PBS -l walltime=48:00:00
#PBS -l select=1:ncpus=16:mem=32gb

module load anaconda3/personal
source activate kraken2_env

kraken2 --db /rds/general/project/fisher-aspergillus-analysis/live/clarisse/dog_ear_kraken2_scripts/kraken2_fungi_db/ $1 --use-names --report /rds/general/project/fisher-aspergillus-results/live/Clarisse/dog_ear_kraken2/$2.report.txt --output /rds/general/project/fisher-aspergillus-results/live/Clarisse/dog_ear_kraken2/$2
```

---

batch script (qsub dog_kraken2_batch.sh):

```
#!/bin/sh

#PBS -l walltime=72:00:00
#PBS -l select=1:ncpus=16:mem=32gb
## This tells the batch manager to re-run job with parameter varying from 1 to N in steps on step-size
#PBS -J 1-14

/rds/general/project/fisher-aspergillus-analysis/live/clarisse/dog_ear_kraken2_scripts/dog_kraken2.sh $(head -$PBS_ARRAY_INDEX /rds/general/project/fisher-aspergillus-analysis/live/clarisse/dog_ear_kraken2_scripts/dog_kraken2_list.txt | tail -1)
```
