Combined Metagenomic / Metatranscriptomic Pipeline for Host-associated Microbiomes

Scott Daniel

Abstract

Overview:

1. Filter DNA reads by quality and host (e.g. for mouse gut bacteria we would filter low-quality reads AND reads that belong to mice)

2. Align DNA reads (fastq, fasta) to Patric database (bacteria genomes) with taxoner (a bowtie2-based linux program) >> Get bacterial genomes

3. Filter out low-quality alignments >> Bacterial genomes

4. Filter RNA reads by quality

5. Align RNA reads to bacterial genomes from step 3 >> RNA counts of genes

6. Get pathway information >> RNA counts per KEGG pathway

If you have treatment groups, we will be able to see changes in bacterial abundance as well as changes to RNA counts

Citation: Scott Daniel Combined Metagenomic / Metatranscriptomic Pipeline for Host-associated Microbiomes. **protocols.io**
dx.doi.org/10.17504/protocols.io.d7m9k5
Published: 23 May 2018

Before start

Unless you have a small set of reads and/or genomes, it is recommended that the pipeline be run on university supercomputers.

Protocol

Setup

Step 1.

Get sample data from github (requires git to be installed on your system).

cmd COMMAND

```
git clone https://github.com/hurwitzlab/mg-sample-data.git
```

NOTES

Scott Daniel 05 Sep 2017

This can be downloaded to your shared supercomputers (job scheduler managed) or your own laptop. It's small enough data (< 10mb) that it should be trivial to run on most systems.

Setup

Step 2.

Edit config.sh with your favorite text editor to set variables (directories mainly)

cmd COMMAND

```
Example: "vim config.sh"
```

QC

Step 3.

Get recipe for singularity image containing fastqc and solexaqa:

(This must be downloaded to a system where you have sudo privileges)

cmd COMMAND

```
git clone https://github.com/hurwitzlab/singularity-fastqc.git
```

QC

Step 4.

From within /singularity-fastqc run ./create_and_bootstrap.sh (see example). This should create a 'fastqc.img'.

cmd COMMAND

```
cd singularity-fastqc/ && ./create_and_bootstrap.sh
```

SAFETY INFORMATION

Make sure there are no errors! E.g. "ABORT: Aborting with RETVAL=255"

QC

Step 5.

sftp or scp the created "fastqc.img" to your shared supercomputer directory OR put it in your "/bin" directory (or other dir in your PATH)

cmd COMMAND

```
Example: "scp fastqc.img username@sftp.hpc.arizona.edu:/rsgrps/usergroup/username/singularity-images"
```

QC

Step 6.

Generate quality reports with fastqc (optional).

cmd COMMAND

```
Example script is mg-sample-data/scripts/00-fastqc-reports.sh

This script is meant to be submitted to a PBS job scheduler by issuing the command "./00-
fastqc-
reports.sh" but can be edited to run with other job schedulers (must also edit ./workers/fa
stqc.sh).
```

NOTES

Scott Daniel 12 Sep 2017

If you want to manually generate fastqc reports, the script you want is in 'scripts/workers/fastqc.sh'

This is true for most other steps.

QC

Step 7.

Run trim_galore (a wrapper script for cutadapt) to cut off low-quality bases and adapters (default is to cut off bases until the probability that the base call was correct is 99% or greater)

cmd COMMAND

```
Example script is mg-sample-data/scripts/01-trim-galore.sh

This script is meant to be submitted to a PBS job array scheduler by running the command ".
/01-trim-
galore.sh" but can be edited to run with other job schedulers (must also edit "./workers/tr
imgalore.sh")
```

QC

Step 8.

Compare quality reports between steps 5 and 6. If you don't see much improvement, consider re-running step 6 with different parameters.

Bowtie2 Prep

Step 9.

We need to build a couple more singularity images for this section: singularity-taxoner and singularity-tuxedo (named after the tuxedo suite of tools: bowtie2, cufflinks, etc.)

Bowtie2 Prep

Step 10.

Get the github repos for building the images

cmd COMMAND

```
git clone https://github.com/hurwitzlab/singularity-tuxedo.git
git clone https://github.com/hurwitzlab/singularity-taxoner.git
```

Bowtie2 Prep

Step 11.

From within /singularity-tuxedo (Tuxedo is a suite of tools that contains bowtie and cuffdiff see http://software.broadinstitute.org/cancer/software/genepattern/rna-seq-analysis#tuxedo)

run ./create_and_bootstrap.sh (see example). This should create a 'bowcuff.img'.

From within /singularity-taxoner run ./create_and_bootstrap.sh. This should create a 'taxoner.img'

cmd COMMAND

```
cd singularity-tuxedo/ && ./create_and_bootstrap.sh
```

SAFETY INFORMATION

Make sure there are no errors! E.g. "ABORT: Aborting with RETVAL=255"

Bowtie2 Prep

Step 12.

sftp or scp the created 'bowcuff.img' and 'taxoner.img' to your shared supercomputer directory OR put it in your '/bin' directory (or other dir in your PATH)

cmd COMMAND

```
Example: "scp bowcuff.img username@sftp.hpc.arizona.edu:/rsgrps/usergroup/username/singularity-images"
```

Bowtie2 Prep

Step 13.

Run 02-taxadb-prep.pbs to prepare the lineage file and genome fasta's for bowtie2 indexing

cmd COMMAND

```
Example script is mg-sample-data/scripts/02-taxadb-prep.pbs

This script is meant to be submitted to a PBS job scheduler by issuing the command "qsub 02-taxadb-prep.pbs" but can be edited to run with other job schedulers.
```

Bowtie2 Prep

Step 14.

Make Bowtie2 indices of the split multi-genome fasta files.

The example script will launch bowtie2 to run index building on two computers in parallel. This is why you split the giant fasta into files of 4 Gb each. So if you had 40 gb worth of genomes in one big file, you could have 10 computers doing the bowtie2 indexing at once.

cmd COMMAND

```
Example script is mg-sample-data/scripts/03-bowtie2-build.sh

This script is meant to be submitted to a PBS job scheduler by issuing the command "./03-bowtie2-
build.sh" but can be edited to run with other job schedulers (must also edit ./workers/bowtie2-build.sh).
```

Taxoner

Step 15.

Run taxoner for the DNA reads to determine species composition

cmd COMMAND

```
Example script is mg-sample-data/scripts/04-taxoner.sh

This script is meant to be submitted to a PBS job scheduler by issuing the command "./04-
taxoner.sh" but can be edited to run with other job schedulers (must also edit ./workers/ru
nTaxoner.sh).
```

Centrifuge

Step 16.

Get the singularity container for centrifuge a classifier program for bacterial/viral species (https://ccb.jhu.edu/software/centrifuge/)

cmd COMMAND

```
git clone https://github.com/scottdaniel/singularity-centrifuge.git
```

SAFETY INFORMATION

If doing this step for the species identification you can optionally skip QC and definitely skip Bowtie2 Prep

Centrifuge

Step 17.

Build the singularity container as before

cmd COMMAND

```
cd singularity-centrifuge && make img
```

Step 18.

Warnings

None