Counting Reads

Author: Angela Garibaldi
Bioinformatics Support Group
April 4, 2018

HTSeq manual: https://htseq.readthedocs.io/en/release_0.9.1/count.html
NOTE: This tutorial assumes you completed the **Review tutorial** from last week. Run those scripts, so that everything is in the directories that are described below. If you did not complete the Review tutorial, but DID complete the **Alignment tutorial**, just utilize the path to your alignment files that you already have. But you will need to change the directory paths in the Griffith_HTSeq.sh script yourself. Great practice!

Getting Started

- Log on to the HPC
  - `$ ssh user_name@hpc.oit.uci.edu`
  - `$ qrsh`
- Download the scripts for today, change permissions to make them executable
  - `$ cd /data/users/$USER/BioinformaticsSG/Review/griffith_analysis_demo/scripts`
  - `$ git clone https://github.com/bioinformaticssg/Counting-Reads.git`
  - `$ cd Counting-Reads`
  - `$ chmod 700 *sh`

Remove Stats rows from either HTSeq or STAR counts:

- Change to the alignment directory to edit the read count files
  - `$ cd /data/users/$USER/BioinformaticsSG/Review/griffith_analysis_demo/alignments`
- Take a look at the top of the file. Note there are some stats at the top that do not align with the read count columns. We need to get rid of this in order to move forward in DESEQ.
  - `$ head /data/users/$USER/BioinformaticsSG/Review/griffith_analysis_demo/alignments/*starReadsPerGene.out.tab`
- Use nano to create the following script:
  - `$ nano cleaning_star_counts.sh`

```
#!/bin/bash
#Trim off bottom count text on all files in a directory

set -euxo pipefail

for file in `ls *starReadsPerGene.out.tab`; do
    cat ${file} | grep -v -E 'N_unmapped|N_ambiguous|N_multimapping|N_noFeature' > ${file}.txt; #for star
done
```

- Save your new nano script.
  - `$ ^X`
  - `$ y`
  - `$ <ENTER>`
- Make your script executable
  - `$ chmod 700 cleaning_star_counts.sh`
- Run your script
  - `$ ./cleaning_star_counts.sh`
- Copy your cleaned counts files to your counts directory

    - `$ cp *.tab.txt /data/users/$USER/BioinformaticsSG/Review/griffith_analysis_demo/counts/`

Add the counts columns of each of your files into 1 file:

- Change to the counts directory to further munge your count files
    - `$ cd /data/users/$USER/BioinformaticsSG/Review/griffith_analysis_demo/counts/`
- Confirm that the stats have been taken off
    - `$ head *tab.txt`
- Confirm that the files are all the same length
    - `$ for file in *tab.txt; do wc -l ${file}; done`
- Paste all of the files together, then check to make sure the gene names match
    - `$ paste HBR_1* HBR_2* HBR_3* | head`
    - `$ paste HBR_1* HBR_2* HBR_3* | tail`
- Now you want to combine just the first column of genes, and the first read column of each sample. First, confirm that you got it right before saving it out.
    - `$ paste HBR_1* HBR_2* HBR_3* | awk '{print $1 "\t" $2 "\t" $6 "\t" $10}' | head`
    - `$ paste HBR_1* HBR_2* HBR_3* | awk '{print $1 "\t" $2 "\t" $6 "\t" $10}' | tail`
- Now actually combine them together, just the first column of genes, and the first read column of each sample. Save it out. You can run head and tail on the combined_out.txt file, if you would like to admire your handy work.
    - `$ paste HBR_1* HBR_2* HBR_3* | awk '{print $1 "\t" $2 "\t" $6 "\t" $10}' > combined_out.txt`

Running HTseq:

If you want to run HTseq in order to utilize some of its alternative counting methods (not union), I have included a script that you may edit. You would edit the “--mode” option

Generate the aligned sample list:

- `$ cd /data/users/$USER/BioinformaticsSG/Review/griffith_analysis_demo/alignments/`

- `$ printf '%s\n' `pwd`/HBR_[1-3].starAligned.sortedByCoord.out.bam > /data/users/$USER/BioinformaticsSG/Review/griffith_analysis_demo/alignments/HBR_alignments_filenames.txt`

- Take a look at the contents of the filename txt files. By providing the full path to the file we do not need to be in the same directory to execute this command.
    - `$ cat /data/users/$USER/BioinformaticsSG/Review/griffith_analysis_demo/alignments/HBR_alignments_filenames.txt`

Get the reads for each alignment file:

    - `$ ./data/users/$USER/BioinformaticsSG/Review/griffith_analysis_demo/scripts/Counting-Reads/`Griffith_HTSeq.sh