RNA-Seq
Data processing
03/07/2019

University of Massachusetts Medical School
Alper Kucukural & Onur Yukselen
Bioinformatics Core
Introduction
Transcript alignment and quantification

For genomic + transcriptomic alignments use Tophat

For transcriptomic alignments and quantifications use RSEM

$ rsync -vau /local/folder/*fastq.gz user@ghpcc06.umassrc.org:/remote/path/.

Use FTP Client to transfer the files:

FileZilla, WinSCP, Cyberduck
Typical Pipeline for RNA-Seq Analysis

1. Barcode Separation
2. QC Checks
3. Adapter Removal
4. 5', 3' Trimming
5. Quality Trimming
6. Quality Removal
7. rRNA, smallRNA
   spikeln, quantification & elimination
8. mRNA quantification
9. mRNA quantifications
10. Produce files for Visualization
    bw, tdf, bam
11. Produce QC & Mapping reports
12. Generate Cluster, Heatmap
    and Scatter plots
13. Differential Expression Analysis
Data transfer
Transfer fastq files to the cluster

Create a directory:
$ mkdir /full/path/of/your_folder
Copy or move your files (same machine):
$ cp -R /source/dir/*_fastq.gz /full/path/of/your_folder/
$ mv /source/dir/*_fastq.gz /full/path/of/your_folder/
$ rsync -azu /source/dir/*_fastq.gz /full/path/of/your_folder/

Copy your files (remote machine):
$ scp /local/folder/*_fastq.gz user@ghpcc06.umassrc.org:/remote/path/
$ rsync -azu /local/folder/*_fastq.gz user@ghpcc06.umassrc.org:/remote/path/

Use FTP Client to transfer the files:
Index files

Tophat and RSEM

http://bioinfo.umassmed.edu/index.php?p=33

- Index files for Tophat:
  - mm10.1.bt2
  - mm11
  - mm11_howto2-build -f mm10.fa mm10
  - mm10
  - mm10.rev.1.bt2
  - mm10.rev.2.bt2

  Output file:
  Alignments in BAM format
  only for genomic coordinates

- Index files for RSEM:
  - mm10.rnn.gz
  - mm10.rse.gz
  - mm10.rev.1.rse.gz
  - mm10.rev.2.rse.gz

  Output files:
  Alignments in BAM format
  for both genomic coordinates
  and transcriptomic coordinates
  Gene and isoform
  quantification results
Expression
Quantify with the RSEM program

- **RSEM** depends on an existing annotation and will only scores transcripts that are present in the given annotation file.
- **RSEM** generates 2 result files:
  - rsem.genes.results
  - rsem.isoforms.results
Expression matrices

Create consolidated table

We use rsem outputs to join the rsem.genes.results files side-by-side, that contains the expected_count or tpm information for all samples, and place them into a final output file. Repeat the same step for isoforms.
Visualization
Alignment visualization using IGV

Download the IGV program from the IGV site and transfer your bam files from the hpcc to your laptop.

http://software.broadinstitute.org/software/igv/download