Good websites to reference:

<https://sites.tufts.edu/biotools/files/2019/04/bioinformatics_for_rnaseq_day1.pdf>

<https://genehub.wordpress.com/2016/02/29/rnaseq-pipeline-alignment-to-de-analysis/>

Sources for program downloads:

FastQC: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

STAR: <https://github.com/alexdobin/STAR>

HTseq-Count: <https://htseq.readthedocs.io/en/master/>

DESeq2: <https://bioconductor.org/packages/release/bioc/html/DESeq2.html>

1. Download Fastq files onto external hard drive or remote server
   1. Follow GTAC documentation
2. Run QC (this may have already been done by GTAC but good to run on raw fastqs)
   1. Module load fastqc/0.11.8 **(to find conda env: “conda info –envs”)**
   2. mkdir fastQC
   3. fastqc \*.fastq.gz -o fastQC –extract
   4. cat summary.txt to make sure passed
3. STAR Alignment to genome
   1. Website to download genomes, copy link address for desired genome: <https://support.illumina.com/sequencing/sequencing_software/igenome.html>
   2. wget link address
   3. tar -xzf genome.file
   4. Generate Genome (**use star\_index script for reference)**

# load the module

module load STAR/2.6.1d

#create a directory to store the index in REF\_DIR=/path/to/dir

# Run STAR in "genomeGenerate" mode

STAR --runMode genomeGenerate \ --genomeDir ${REF\_DIR}/Sequence/STAR \ --genomeFastaFiles ${REF\_DIR}/Sequence/WholeGenomeFasta/genome.fa \ --runThreadN 16

* 1. Align to genome (**use star\_align script for reference)**

bin/bash

#SBATCH -J STARalign # Can be changed

#SBATCH -o STARalign-%j.out # MODIFY WITH CAUTION

#SBATCH -e STARalign-%j.err # MODIFY WITH CAUTION

#SBATCH --partition=genomics # DO NOT DELETE

GENOMEDIR="/xfs2/dipaololab/Bulk\_RNA/Mouse\_IL13/mm10\_STAR/"

for i in \*\_R1.fastq.gz; do

echo $i

describe=$(echo ${i} | awk -F'[.]' '{print $1}')

echo $describe

STAR --genomeDir=$GENOMEDIR \

--readFilesIn $i ${i%\_R1.fastq.gz}\_R2.fastq.gz \

--readFilesCommand zcat \

--outSAMtype BAM SortedByCoordinate \

--outFileNamePrefix ${describe}. \

--runThreadN 32

done ;

* 1. HTSeq-count (**use htseq\_script script for reference)**

#!/bin/bash

#SBATCH -J HTseq-count # Can be changed

#SBATCH -o HTseq-count-%j.out # MODIFY WITH CAUTION

#SBATCH -e HTseq-count-%j.err # MODIFY WITH CAUTION

#SBATCH --partition=genomics # DO NOT DELETE

samples=$( ls \*.bam)

for m in $samples ; do

gtf="/xfs2/dipaololab/Bulk\_RNA/Mouse\_IL13/mm10\_STAR/Mus\_musculus/UCSC/mm10/Annotation/Genes/genes.gtf"

echo $m

describe=$(echo ${m} | awk -F'[.]' '{print $1}')

echo $describe

htseq-count -m union -r pos -i gene\_name -a 10 --stranded=no $m $gtf > output\_${describe}.counts.txt

done ;

1. DESeq2 in R
   1. <http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#htseq-count-input>