ITBI2024-2025, Exercise-lecture5, 7-11-2024, M-Reczko

1.

Add required environment variables:

echo "export IGV=/usr/share/java" >> ~/.bashrc

echo "export PATH=$PATH:/home/reczko/tools/samtools-0.1.19" >> ~/.bashrc

Unpack and run the RNAseq-tutorial using

cd ~

tar -xzvf /mnt/hdd_b/home/reczko/tools/naseq_workshop.taz

cd rnaseq_workshop

 ./runTuxedoDemo.pl

Answer 'y' to all steps until the IGV visualization starts.

For the data in the RNAseq-tutorial, using the IGV tool and reporting each location as "genome:<start>-<end>":

1.1 find a read with unmapped pair

1.2 find a read that has 2 mismatches
1.3 find a read-pair with pair orientation: F2R1
1.4 find a gene that agrees in 6 splice-sites with its annotation
1.5 find a location where less than 96% of the reads agree with the annotation and
at least 2 other nucleotides occur.

1.6 find a gene where the peak with the highest number of reads has at least 3 times the
reads of the highest peak in a different sample (with non-zero reads for that gene)

2.
Using the cuffcompare tool (see
http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/index.html),
calculate the statistics describing the match between the reconstructed transcriptome (in
merged_asm/merged.gtf) and the reference annotation (in genes.gff3). How many novel
exons and novel introns were discovered?

3.
Using tophat's align_summary.txt, what is the mapping percentage for left and right reads,
averaged over all 4 samples?

4.

Using the commands in the tutorial Tuxedo_workshop_activities.pdf (load from eclass) generate a (bash) shell script to execute the tuxedo pipeline. Modify all relevant program invocations to use 4 CPUs. Help for shell scripts can be found e.g. at http://tldp.org/HOWTO/Bash-Prog-Intro-HOWTO.html . An efficient solution will have one script to analyze one sample and a second script calling the first script for the 4 samples.

5.

Bonus: Modify a copy of the script(s) in question 4 to use the Hisat2 mapper instead of tophat and compare the obtained results in terms of the number of significantly differentially expresssed genes.