6-iii. Integrated assignment

Preamble: Note that the following integrated assignment asks you to download new RNA-seq data and apply the concepts you have learned up to this point. To complete this assignment you will need to review commands we performed in many of the earlier sections. Try to construct these commands on your own and get all the way to the end of the assignment. If you get very stuck or would like to compare your solutions to those suggested by the instructors, refer to the answers page. The integrated assignment answers page is an expanded version of this page with all of the questions plus detailed code solutions to all problems. The answer page is available in the git repository for this wiki. It is slightly hidden to reduce temptation to look at it without trying on your own. Ask an instructor if you have trouble finding it.

Background: The *PCA3* gene plays a role in Prostate Cancer detection due to its localized expression in prostate tissues and its over-expression in tumour tissues. This gene expression profile makes it a useful marker that can complement the most frequently used biomarker for prostate cancer, PSA. There are cancer assays available that test the presence of *PCA3* in urine.

Objectives: In this assignment, we will be using a subset of the GSE22260 dataset, which consists of 30 RNA-seq tumour/normal pairs, to assess the prostate cancer specific expression of the PCA3 gene.

Experimental information and other things to keep in mind:

- The libraries are polyA selected.
- The libraries are prepared as paired end.
- The samples are sequenced on a Illumina Genome Analyzer II (this data is now quite old).
- Each read is 36 bp long
- The average insert size is 150 bp with standard deviation of 38bp.
- We will only look at chromosome 9 in this exercise.
- The dataset is located here: GSE22260
- 20 tumour and 10 normal samples are available
- For this exercise we will pick 3 matched pairs (C02,C03,C06 for tumour and N02,N03,N06 for normal). We can do more if we have time.

PART 1 : Obtaining Data and References

Goals:

- Obtain the files necessary for data processing
- Familiarize yourself with reference and annotation file format
- Familiarize yourself with sequence FASTQ format

Create a working directory `~/workspace/rnaseq/integrated_assignment/` to store this

exercise. Then create a unix environment variable named RNA_ASSIGNMENT that stores this path for convenience in later commands.

```
cd $RNA_HOME
mkdir -p ~/workspace/rnaseq/integrated_assignment/
export RNA_ASSIGNMENT=~/workspace/rnaseq/integrated_assignment/
```

You will also need the following environment variables througout the assignment:

```
export RNA_DATA_DIR=$RNA_ASSIGNMENT/fasta
export RNA_REFS_DIR=$RNA_ASSIGNMENT/refs
export RNA_REF_INDEX=$RNA_REFS_DIR/Homo_sapiens.GRCh38.dna.chromosome.9
export RNA_REF_FASTA=$RNA_REF_INDEX.fa
export RNA_REF_GTF=$RNA_REFS_DIR/Homo_sapiens.GRCh38.86.chr9.gtf
```

Obtain reference, annotation and data files and place them in the integrated assignment directory

Note: when initiating an environment variable, we do not need the $; however, everytime we call the variable, it needs to be preceeded by a $.

```
echo $RNA_ASSIGNMENT
cd $RNA_ASSIGNMENT
cp ~/CourseData/RNA_data/integrated_assignment/data.zip .
unzip data.zip
```

Answer all questions and follow all instructions below:

Q1.) How many items are there under the “refs” directory ?

What if this reference file was not provided for you? How would you obtain/create a reference genome fasta file for chromosome 9 only. How about the GTF transcripts file from Ensembl? How would you create one that contained only transcripts on chromosome 9?

Q2.) How many exons does the gene *PCA3* have?

Q3.) How many cancer/normal samples do you see under the data directory?

NOTE: The fasta files you have copied above contain sequences for chr9 only. We have pre-processed those fasta files to obtain chr9 and also matched read1/read2 sequences for each of the samples. You do not need to redo this.

Q4.) What sample has the highest number of reads?

Remember that a read record looks like this:

```
>HWUSI-EAS230-R:6:58:12:550#0/1
TTTGTTTGTTTGCTTCTGTTTCCCCCCAATGACTGA
```

PART 2: Data alignment

Goals:

- Familiarize yourself with HISAT2 alignment options
- Perform alignments
- Obtain alignment summary

Q5.) Create HISAT2 alignment commands for all of the six samples and run alignments

Q6.) How would you obtain summary statistics for each aligned file?

PART 3: Expression Estimation

Goals:

- Familiarize yourself with Stringtie options
- Run Stringtie to obtain expression values
- Obtain expression values for the gene *PCA3*

Create an expression results directory, run Stringtie on all samples, and store the results in appropriately named subdirectories in this results dir

Q7.) How do you get the expression of the gene *PCA3* across the normal and carcinoma samples?

PART 4: Differential Expression Analysis

Goals:

- Perform differential analysis between tumor and normal samples
- Check if PCA3 is differentially expressed

Perform carcinoma vs. normal comparison, using all samples, for known (reference only mode) transcripts. First create a file that lists our 6 expression files, then view that file, then start an R session. **Adapt the R tutorial file has been provided in the github repo for part 1 of the tutorial: Tutorial_Module4_Part1_ballgown.R. Modify it to fit the goals of this assignment then run it.**

Q8.) Are there any significant differentially expressed genes? What about the PCA3?

Q9.) What plots can you generate to help you visualize this gene expression profile