6-iii. Integrated assignment answers

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
```

Q1.) How many items are there under the “refs” directory (counting all files in all sub-directories)?

A1.) The answer is 6. Review these files so that you are familiar with them.

```
cd $RNA_ASSIGNMENT/refs/
tree
find *
find * | wc -l
```

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?

A2.) The answer is 4. Review the GTF file so that you are familiar with it. What downstream steps will we need this file for? What is it used for?

```
cd $RNA_ASSIGNMENT/refs
grep -w "PCA3" Homo_sapiens.GRCh38.86.chr9.gtf
```

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

A3.) The answer is 12. 6 normal and 6 tumor.

```
cd $RNA_ASSIGNMENT/fasta/
ls -l
ls -1 | wc -l
```

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?

A4.) The answer is that 'carcinoma_C06' has the most reads (288428/2 = 144214 reads).

An easy way to figure out the number of reads is to make use of the command ‘wc’. This command

counts the number of lines in a file. Keep in mind that one sequence can be represented by multiple lines. Therefore, you need to first grep the read tag ">" and count those.

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

Running this command only give you 2 x read number:

```
cd $RNA_ASSIGNMENT/fasta/
wc -l YourFastaFile.fasta
wc -l *
```

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

```
echo $RNA_ALIGN_DIR
mkdir -p $RNA_ALIGN_DIR
cd $RNA_ALIGN_DIR

hisat2 -p 8 --rg-id=carcinoma_C02 --rg SM:carcinoma --rg LB:carcinoma_C02 -x
$RNA_REF_INDEX --dta -f -1 $RNA_DATA_DIR/carcinoma_C02_read1.fasta -2
$RNA_DATA_DIR/carcinoma_C02_read2.fasta -S ./carcinoma_C02.sam
hisat2 -p 8 --rg-id=carcinoma_C03 --rg SM:carcinoma --rg LB:carcinoma_C03 -x
$RNA_REF_INDEX --dta -f -1 $RNA_DATA_DIR/carcinoma_C03_read1.fasta -2
$RNA_DATA_DIR/carcinoma_C03_read2.fasta -S ./carcinoma_C03.sam
hisat2 -p 8 --rg-id=carcinoma_C06 --rg SM:carcinoma --rg LB:carcinoma_C06 -x
$RNA_REF_INDEX --dta -f -1 $RNA_DATA_DIR/carcinoma_C06_read1.fasta -2
$RNA_DATA_DIR/carcinoma_C06_read2.fasta -S ./carcinoma_C06.sam

hisat2 -p 8 --rg-id=normal_N02 --rg SM:normal --rg LB:normal_N02 -x $RNA_REF_INDEX
--dta -f -1 $RNA_DATA_DIR/normal_N02_read1.fasta -2
$RNA_DATA_DIR/normal_N02_read2.fasta -S ./normal_N02.sam
hisat2 -p 8 --rg-id=normal_N03 --rg SM:normal --rg LB:normal_N03 -x $RNA_REF_INDEX
--dta -f -1 $RNA_DATA_DIR/normal_N03_read1.fasta -2
$RNA_DATA_DIR/normal_N03_read2.fasta -S ./normal_N03.sam
hisat2 -p 8 --rg-id=normal_N06 --rg SM:normal --rg LB:normal_N06 -x $RNA_REF_INDEX
--dta -f -1 $RNA_DATA_DIR/normal_N06_read1.fasta -2
$RNA_DATA_DIR/normal_N06_read2.fasta -S ./normal_N06.sam

#convert sam alignments to bam..how much space did you save by performing this
conversion?

samtools sort -@ 8 -o carcinoma_C02.bam carcinoma_C02.sam
samtools sort -@ 8 -o carcinoma_C03.bam carcinoma_C03.sam
samtools sort -@ 8 -o carcinoma_C06.bam carcinoma_C06.sam
samtools sort -@ 8 -o normal_N02.bam normal_N02.sam
samtools sort -@ 8 -o normal_N03.bam normal_N03.sam
```

```
samtools sort -@ 8 -o normal_N06.bam normal_N06.sam

#merge the bams for visulization purposes
cd $RNA_ASSIGNMENT/alignments/hisat2
java -Xmx2g -jar /usr/local/picard/picard.jar MergeSamFiles OUTPUT=carcinoma.bam
INPUT=carcinoma_C02.bam INPUT=carcinoma_C03.bam INPUT=carcinoma_C06.bam
java -Xmx2g -jar /usr/local/picard/picard.jar MergeSamFiles OUTPUT=normal.bam
INPUT=normal_N02.bam INPUT=normal_N03.bam INPUT=normal_N06.bam
```

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

A6.) There are many RNA-seq QC tools available that can provide you with detailed information about the quality of the aligned sample (e.g. FastQC and RSeQC). However, for a simple summary of aligned reads counts you can use samtools flagstat. You can also look for the logs generated by TopHat. These logs provide a summary of the aligned reads.

```
cd $RNA_ASSIGNMENT/alignments/hisat2/

samtools flagstat carcinoma_C02.bam > carcinoma_C02.flagstat.txt
samtools flagstat carcinoma_C03.bam > carcinoma_C03.flagstat.txt
samtools flagstat carcinoma_C06.bam > carcinoma_C06.flagstat.txt

samtools flagstat normal_N02.bam > normal_N02.flagstat.txt
samtools flagstat normal_N03.bam > normal_N03.flagstat.txt
samtools flagstat normal_N06.bam > normal_N06.flagstat.txt

grep "mapped (" *.flagstat.txt
```

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

```
cd $RNA_ASSIGNMENT/
mkdir -p expression/stringtie/ref_only/
cd expression/stringtie/ref_only/

stringtie -p 8 -G $RNA_REF_GTF -e -B -o carcinoma_C02/transcripts.gtf
$RNA_ALIGN_DIR/carcinoma_C02.bam
stringtie -p 8 -G $RNA_REF_GTF -e -B -o carcinoma_C03/transcripts.gtf
$RNA_ALIGN_DIR/carcinoma_C03.bam
stringtie -p 8 -G $RNA_REF_GTF -e -B -o carcinoma_C06/transcripts.gtf
$RNA_ALIGN_DIR/carcinoma_C06.bam
stringtie -p 8 -G $RNA_REF_GTF -e -B -o normal_N02/transcripts.gtf
$RNA_ALIGN_DIR/normal_N02.bam
stringtie -p 8 -G $RNA_REF_GTF -e -B -o normal_N03/transcripts.gtf
$RNA_ALIGN_DIR/normal_N03.bam
stringtie -p 8 -G $RNA_REF_GTF -e -B -o normal_N06/transcripts.gtf
```

```
$RNA_ALIGN_DIR/normal_N06.bam
```

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

A7.) To look for the expression value of a specific gene, you can use the command 'grep' followed by the gene name and the path to the expression file

```
cd $RNA_ASSIGNMENT/expression/stringtie/ref_only
grep ENSG00000225937 ./*/transcripts.gtf | cut -f1,9 | grep FPKM
```

PART 4: Differential Expression Analysis

Goals:

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

```
mkdir -p $RNA_ASSIGNMENT/de/ballgown/ref_only/
cd $RNA_ASSIGNMENT/de/ballgown/ref_only/
```

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 where we will examine these results:

```
printf
"\"ids\",\"type\",\"path\"\n\"carcinoma_C02\",\"carcinoma\",\"$RNA_ASSIGNMENT/expre
ssion/stringtie/ref_only/carcinoma_C02\"\n\"carcinoma_C03\",\"carcinoma\",\"$RNA_AS
SIGNMENT/expression/stringtie/ref_only/carcinoma_C03\"\n\"carcinoma_C06\",\"carcino
ma\",\"$RNA_ASSIGNMENT/expression/stringtie/ref_only/carcinoma_C06\"\n\"normal_N02\
",\"normal\",\"$RNA_ASSIGNMENT/expression/stringtie/ref_only/normal_N02\"\n\"normal
_N03\",\"normal\",\"$RNA_ASSIGNMENT/expression/stringtie/ref_only/normal_N03\"\n\"n
ormal_N06\",\"normal\",\"$RNA_ASSIGNMENT/expression/stringtie/ref_only/normal_N06\"
\n" > carcinoma_vs_normal.csv

R
```

***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 PCA3?

A8.) Due to the small sample size, the *PCA3* signal is not significant at the adjusted p-value level. You can try re-running the above exercise on your own by using all of the samples in the original data set. Does including more samples change the results?

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

A9.) The CummerBund package provides a wide variety of plots that can be used to visualize a gene's

expression profile or genes that are differentially expressed. Some of these plots include heatmaps, boxplots, and volcano plots. Alternatively you can use custom plots using ggplot2 command or base R plotting commands such as those provided in the supplementary tutorials. Start with something very simple such as a scatter plot of tumor vs. normal FPKM values.

***see attached assignment_Supplementary.R for plotting options**