Integrated Assignment : June 2015

Day1: Parts 1-2
Day2: Parts 3-4

Background: 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's 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 tests 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.

Things to keep in mind:

- The libraries are polyA selected.
- The libraries are prepared as paired end.
- The samples are sequenced on Illumina's Genome Analyzer II.
- 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.
- Dataset is located here: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=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 ------Obtainning 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

#set your working directory

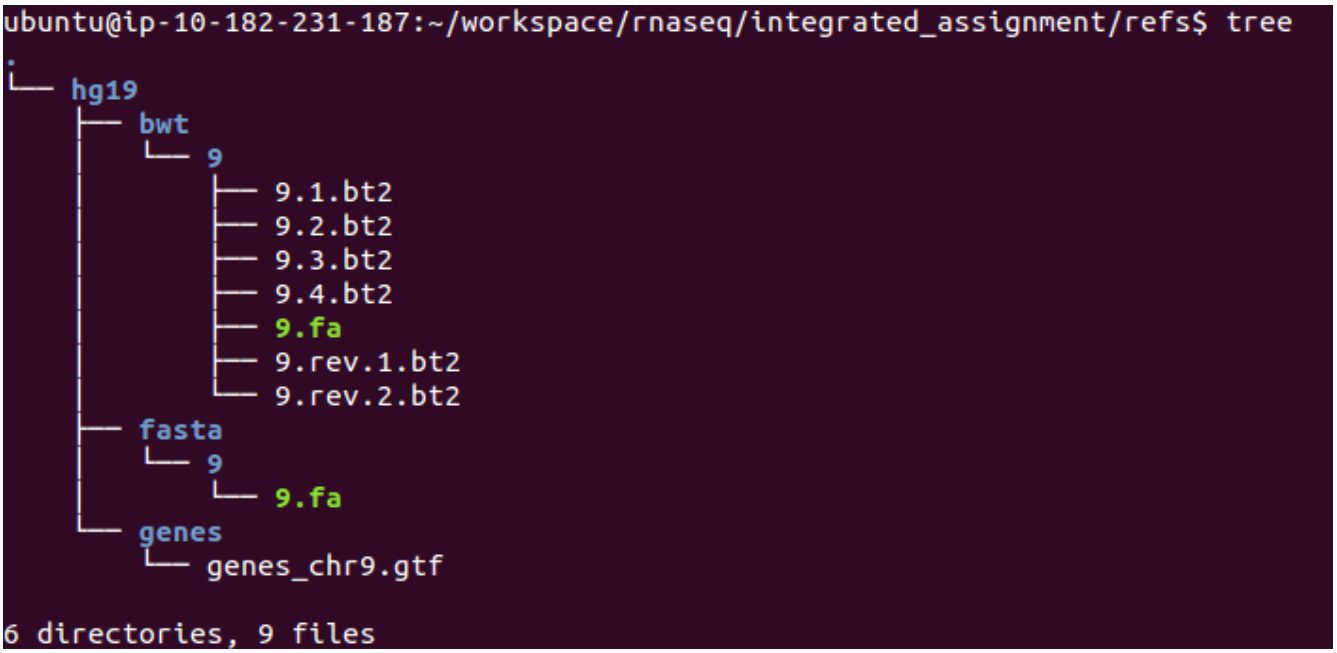
```
mkdir -p ~/workspace/rnaseq/integrated_assignment/
export RNA_ASSIGNMENT=~/workspace/rnaseq/integrated_assignment
```

#copy the necessary reference and annotation files. Note, when initiating an environment variable, we don't need the $; however, everytime we call the variable, it needs to be preceeded by a $.

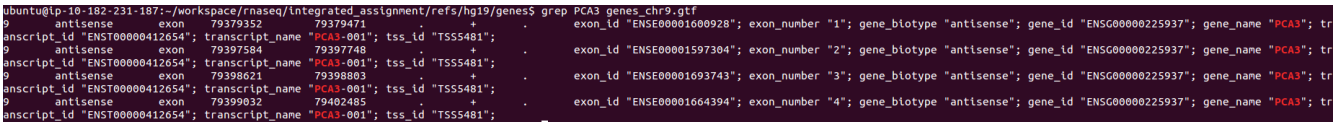
#make sure that the environment variable is set correctly

```
echo $RNA_ASSIGNMENT
cp -r ~/CourseData/RNA_data/integrated_assignment_files/* $RNA_ASSIGNMENT
cd $RNA_ASSIGNMENT
```

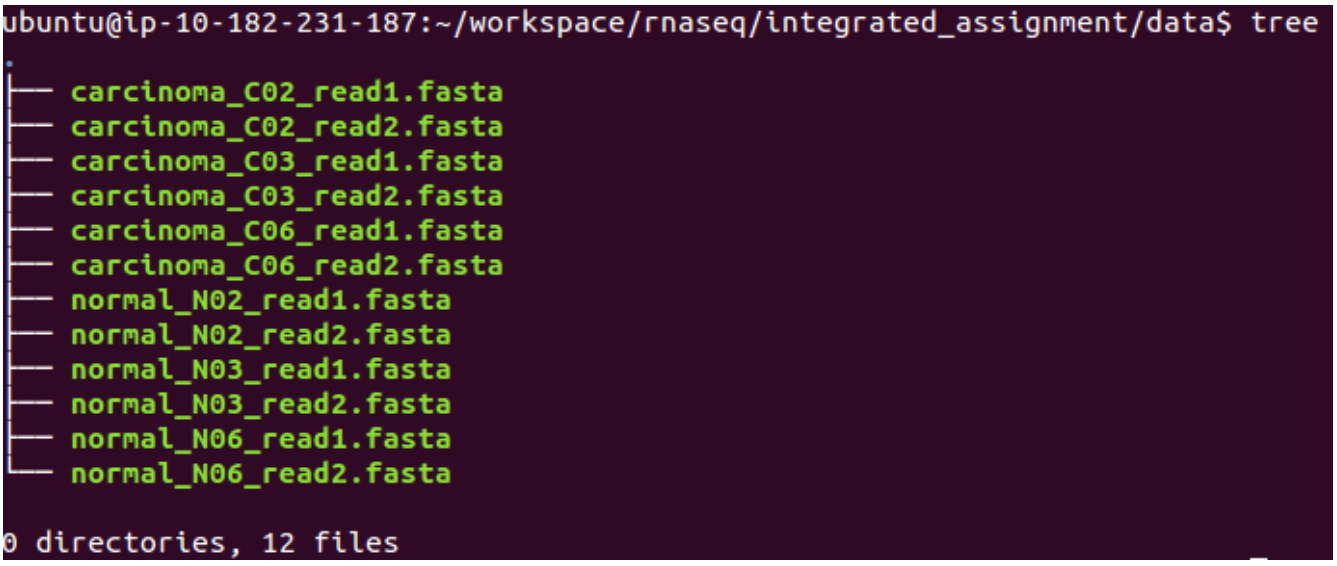
Q1) How many directories are there under the “refs” directory?



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. I 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; However, I will explain below the process I went through to get them to this point.

############################ FYI ################################
- Access the following link: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22260. Scroll down to select the files you want to download.

-The raw data in GEO is provided as _map.txt. After you download the files, you can run the following command to convert them to FASTA:

```
cat GSM554076_C02_read1_map.txt | grep chr9 | cut -f1,2 | awk '{print ">"$1"\n"$2}' > GSM554076_C02_read1_map.chr9.fasta

cat GSM554076_C02_read2_map.txt | grep chr9 | cut -f1,2 | awk '{print ">"$1"\n"$2}' > GSM554076_C02_read2_map.chr9.fasta
```

-The second challenge was to match the reads for both read1 and read2, since the two FASTA files have different number of records.

```
for i in `cat GSM554076_C02_read2_map.chr9.fasta | grep ">"`;do R1=`echo ${i} | sed 's/0\/2/0\/1/g'`; grep -A1 $R1 GSM554076_C02_read1_map.chr9.fasta >> carcinoma_C02_read1.fasta;done;

for i in `cat carcinoma_C02._read1.fasta | grep ">"`;do R2=`echo ${i} | sed 's/0\/1/0\/2/g'`; grep -A1 $R2 GSM554076_C02_read2_map.chr9.fasta >> carcinoma_C02_read2.fasta;done;
```

- Now you have two FASTA files with the same number of reads at the each end
############################ FYI ################################

Q4) What sample has the highest number of reads?

PART 2 ----- Data alignment -----

Goals:
- Familiarize yourself with Tophat/Bowtie alignment options
- Perform alignments
- Obtain alignment summary

Q5) What is the value of --mate-inner-dist? What calculation did you do to get that answer?

Q6) Considering that the read length in this exercise is 36bp, what should you set the --segment-length to (default is 25bp)?

```
cd $RNA_ASSIGNMENT/
export RNA_DATA_DIR=$RNA_ASSIGNMENT/data/
echo $RNA_DATA_DIR
mkdir -p alignments/tophat/trans_idx
cd alignments/tophat
export TRANS_IDX_DIR=$RNA_ASSIGNMENT/alignments/tophat/trans_idx/
echo $TRANS_IDX_DIR
```

#take a minute and try to figure out what each parameter means and how we go the numbers.

tophat2 -p 8 --mate-inner-dist 80 --mate-std-dev 38 --segment-length 18 --rg-id=normal --rg-sample=normal_N02 -o normal_N02 -G $RNA_ASSIGNMENT/refs/hg19/genes/genes_chr9.gtf --transcriptome-index $TRANS_IDX_DIR/ENSG_Genes $RNA_ASSIGNMENT/refs/hg19/bwt/9/9 $RNA_DATA_DIR/normal_N02_read1.fasta $RNA_DATA_DIR/normal_N02_read2.fasta

tophat2 -p 8 --mate-inner-dist 80 --mate-std-dev 38 --segment-length 18 --rg-id=normal --rg-sample=normal_N03 -o normal_N03 -G $RNA_ASSIGNMENT/refs/hg19/genes/genes_chr9.gtf --transcriptome-index $TRANS_IDX_DIR/ENSG_Genes $RNA_ASSIGNMENT/refs/hg19/bwt/9/9 $RNA_DATA_DIR/normal_N03_read1.fasta $RNA_DATA_DIR/normal_N03_read2.fasta

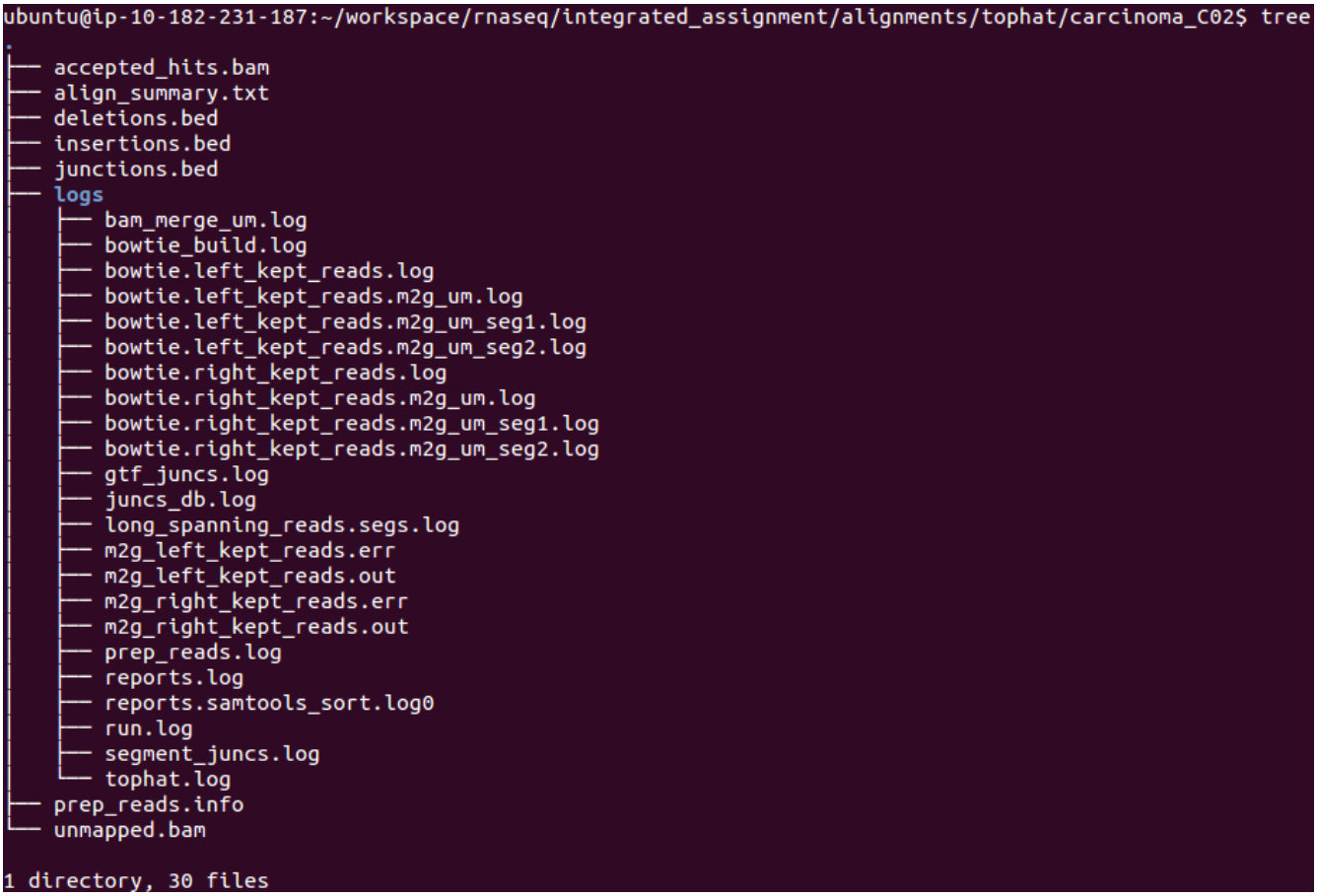
tophat2 -p 8 --mate-inner-dist 80 --mate-std-dev 38 --segment-length 18 --rg-id=normal --rg-sample=normal_N06 -o normal_N06 -G $RNA_ASSIGNMENT/refs/hg19/genes/genes_chr9.gtf --transcriptome-index $TRANS_IDX_DIR/ENSG_Genes $RNA_ASSIGNMENT/refs/hg19/bwt/9/9 $RNA_DATA_DIR/normal_N06_read1.fasta $RNA_DATA_DIR/normal_N06_read2.fasta

tophat2 -p 8 --mate-inner-dist 80 --mate-std-dev 38 --segment-length 18 --rg-id=carcinoma --rg-sample=carcinoma_C02 -o carcinoma_C02 -G $RNA_ASSIGNMENT/refs/hg19/genes/genes_chr9.gtf --transcriptome-index $TRANS_IDX_DIR/ENSG_Genes $RNA_ASSIGNMENT/refs/hg19/bwt/9/9 $RNA_DATA_DIR/carcinoma_C02_read1.fasta $RNA_DATA_DIR/carcinoma_C02_read2.fasta

tophat2 -p 8 --mate-inner-dist 80 --mate-std-dev 38 --segment-length 18 --rg-id=carcinoma --rg-sample=carcinoma_C03 -o carcinoma_C03 -G $RNA_ASSIGNMENT/refs/hg19/genes/genes_chr9.gtf --transcriptome-index $TRANS_IDX_DIR/ENSG_Genes $RNA_ASSIGNMENT/refs/hg19/bwt/9/9 $RNA_DATA_DIR/carcinoma_C03_read1.fasta $RNA_DATA_DIR/carcinoma_C03_read2.fasta

tophat2 -p 8 --mate-inner-dist 80 --mate-std-dev 38 --segment-length 18 --rg-id=carcinoma --rg-sample=carcinoma_C06 -o carcinoma_C06 -G $RNA_ASSIGNMENT/refs/hg19/genes/genes_chr9.gtf --transcriptome-index $TRANS_IDX_DIR/ENSG_Genes $RNA_ASSIGNMENT/refs/hg19/bwt/9/9 $RNA_DATA_DIR/carcinoma_C06_read1.fasta $RNA_DATA_DIR/carcinoma_C06_read2.fasta

At this point, each one of your samples should have the following files:



Q7) How would you obtain summary statistics for each aligned file?

PART 3 ---- Expression Estimation ---------

Goals:
- Familiarize yourself with Cufflinks options
- Run Cufflinks to obtain expression values
- Obtain expression values for the gene PCA3

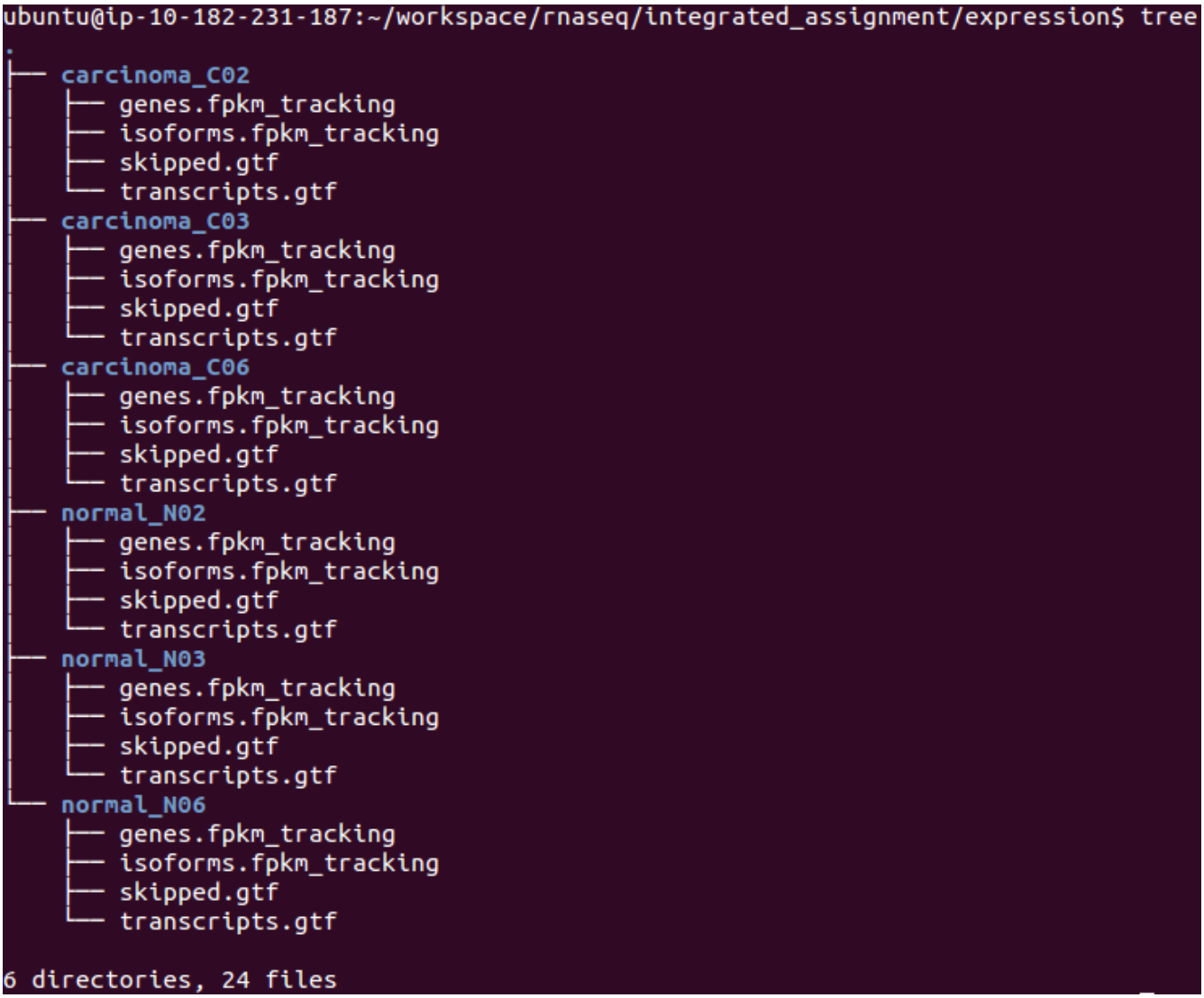
```
cd $RNA_ASSIGNMENT/
mkdir expression
cd expression
```

example (how to run cufflinks for one sample):

```
cufflinks -p 8 -o normal_N02 --GTF $RNA_ASSIGNMENT/refs/hg19/genes/genes_chr9.gtf
--no-update-check $RNA_ASSIGNMENT/alignments/tophat/normal_N02/accepted_hits.bam
cufflinks -p 8 -o normal_N03 --GTF $RNA_ASSIGNMENT/refs/hg19/genes/genes_chr9.gtf
--no-update-check $RNA_ASSIGNMENT/alignments/tophat/normal_N03/accepted_hits.bam
cufflinks -p 8 -o normal_N06 --GTF $RNA_ASSIGNMENT/refs/hg19/genes/genes_chr9.gtf
--no-update-check $RNA_ASSIGNMENT/alignments/tophat/normal_N06/accepted_hits.bam

cufflinks -p 8 -o carcinoma_C02 --GTF $RNA_ASSIGNMENT/refs/hg19/genes/genes_chr9.gtf
--no-update-check $RNA_ASSIGNMENT/alignments/tophat/carcinoma_C02/accepted_hits.bam
cufflinks -p 8 -o carcinoma_C03 --GTF $RNA_ASSIGNMENT/refs/hg19/genes/genes_chr9.gtf
--no-update-check $RNA_ASSIGNMENT/alignments/tophat/carcinoma_C03/accepted_hits.bam
cufflinks -p 8 -o carcinoma_C06 --GTF $RNA_ASSIGNMENT/refs/hg19/genes/genes_chr9.gtf
--no-update-check $RNA_ASSIGNMENT/alignments/tophat/carcinoma_C06/accepted_hits.bam
```

At this point, you should have the following files in your “expression” directory:



Q8) How do you get the expression of 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

cd $RNA_ASSIGNMENT/expression

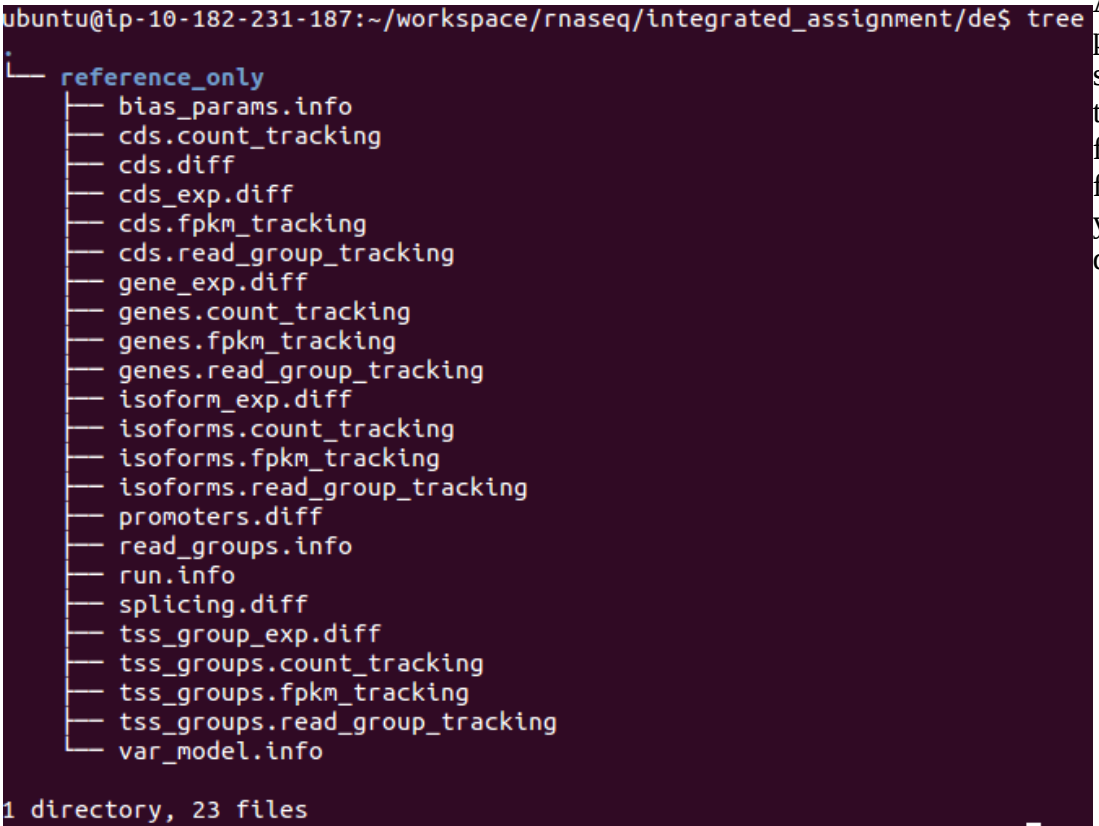
ls -1 */transcripts.gtf > assembly_GTF_list.txt

cuffmerge -p 8 -o merged -g $RNA_ASSIGNMENT/refs/hg19/genes/genes_chr9.gtf -s $RNA_ASSIGNMENT/refs/hg19/bwt/9/ assembly_GTF_list.txt

cd $RNA_ASSIGNMENT/
mkdir de
mkdir de/reference_only
cd $RNA_ASSIGNMENT/alignments/tophat

#run cuffdiff to perform comparison

cuffdiff -p 8 -L Normal,Carcinoma -o $RNA_ASSIGNMENT/de/reference_only/ --no-update-check $RNA_ASSIGNMENT/expression/merged/merged.gtf normal_N02/accepted_hits.bam,normal_N03/accepted_hits.bam,normal_N06/accepted_hits.bam carcinoma_C02/accepted_hits.bam,carcinoma_C03/accepted_hits.bam,carcinoma_C06/accepted_hits.bam



At this point, you should have the following files under your “de” directory:

Q9) any significant genes that are differentially expressed? what about PCA3?

NOTE: Make a copy of the data to use in generateCummerbund plots generation

```
cd $RNA_ASSIGNMENT/
mkdir final_results
cd $RNA_ASSIGNMENT/final_results
mkdir reference_only
cp $RNA_ASSIGNMENT/de/reference_only/isoform* reference_only/
cp $RNA_ASSIGNMENT/de/reference_only/read_groups.info reference_only/
```

NOTE: Rerun Obi's CummerBund Script focusing on PCA3 genes.

Q10) What plots can you generate to help you visualize this gene's expression profile?

Q11) List the reasons why the differential expression of PCA3 might not have been properly assessed in this analysis? Analysis weeknesses ?