Integrated Assignment : April 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

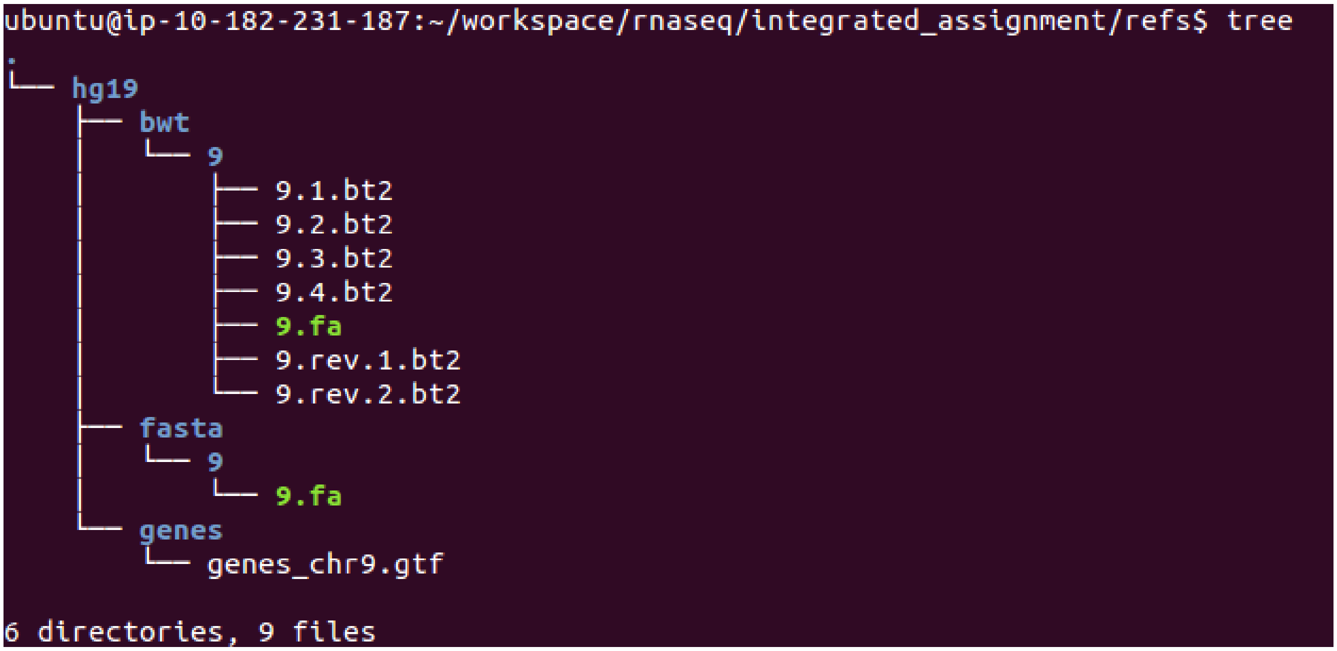
```
mkdir -p ~/workspace/rnaseq/integrated_assignment/
export RNA_HOME=~/workspace/rnaseq/integrated_assignment
```

#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

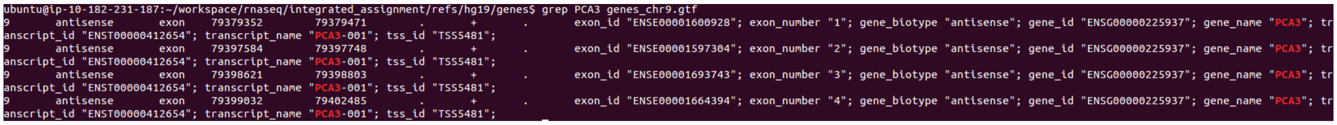
```
echo $RNA_HOME
cp -r ~/CourseData/RNA_data/integrated_assignment_files/* $RNA_HOME
cd $RNA_HOME
```

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



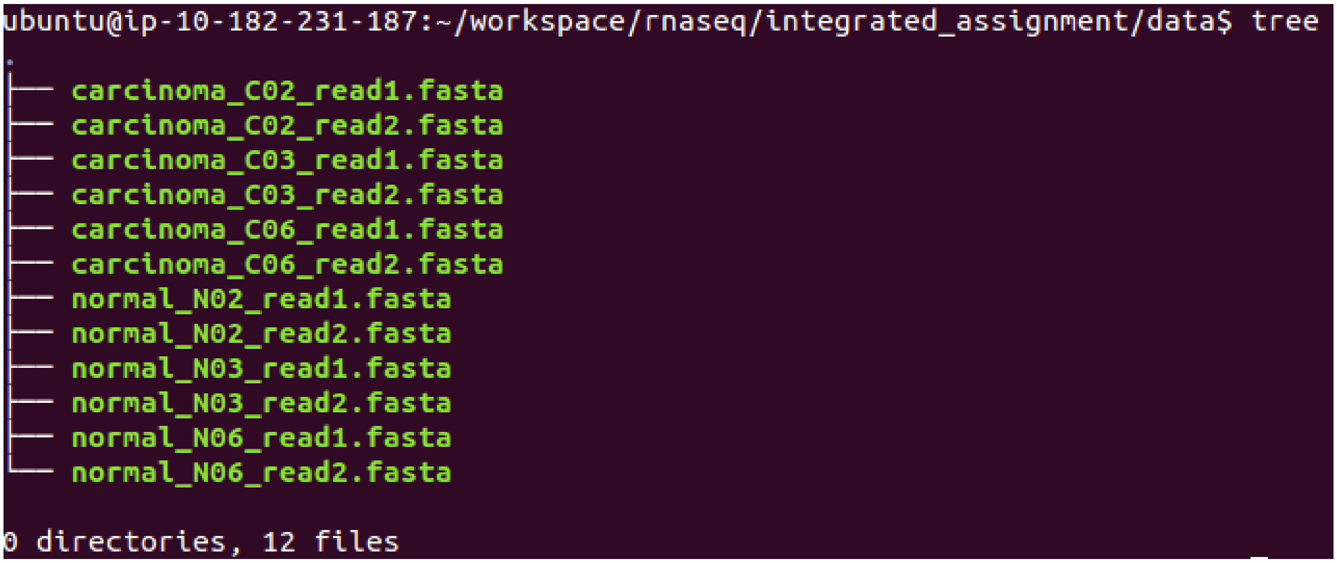
```
ubuntu@ip-10-182-231-187:~/workspace/rnaseq/integrated_assignment/refs$ tree
.
└── hg19
    ├── bwt
    │   └── 9
    │       ├── 9.1.bt2
    │       ├── 9.2.bt2
    │       ├── 9.3.bt2
    │       ├── 9.4.bt2
    │       ├── 9.fa
    │       ├── 9.rev.1.bt2
    │       └── 9.rev.2.bt2
    ├── fasta
    │   └── 9
    │       └── 9.fa
    └── genes
        └── genes_chr9.gtf

6 directories, 9 files
```

Q2) How many exons does the gene PCA3 have?



```
ubuntu@ip-10-182-231-187:~/workspace/rnaseq/integrated_assignment/refs/hg19/genes$ grep PCA3 genes_chr9.gtf
9       antisense       exon    79379352        79379471        .       +       .       exon_id "ENSE00001600928"; exon_number "1"; gene_biotype "antisense"; gene_id "ENSG00000225937"; gene_name "PCA3"; tr
anscript_id "ENST00000412654"; transcript_name "PCA3-001"; tss_id "TSS5481";
9       antisense       exon    79397584        79397748        .       +       .       exon_id "ENSE00001597304"; exon_number "2"; gene_biotype "antisense"; gene_id "ENSG00000225937"; gene_name "PCA3"; tr
anscript_id "ENST00000412654"; transcript_name "PCA3-001"; tss_id "TSS5481";
9       antisense       exon    79398621        79398803        .       +       .       exon_id "ENSE00001693743"; exon_number "3"; gene_biotype "antisense"; gene_id "ENSG00000225937"; gene_name "PCA3"; tr
anscript_id "ENST00000412654"; transcript_name "PCA3-001"; tss_id "TSS5481";
9       antisense       exon    79399032        79402485        .       +       .       exon_id "ENSE00001664394"; exon_number "4"; gene_biotype "antisense"; gene_id "ENSG00000225937"; gene_name "PCA3"; tr
anscript_id "ENST00000412654"; transcript_name "PCA3-001"; tss_id "TSS5481";
```

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



```
ubuntu@ip-10-182-231-187:~/workspace/rnaseq/integrated_assignment/data$ tree
.
├── carcinoma_C02_read1.fasta
├── carcinoma_C02_read2.fasta
├── carcinoma_C03_read1.fasta
├── carcinoma_C03_read2.fasta
├── carcinoma_C06_read1.fasta
├── carcinoma_C06_read2.fasta
├── normal_N02_read1.fasta
├── normal_N02_read2.fasta
├── normal_N03_read1.fasta
├── normal_N03_read2.fasta
├── normal_N06_read1.fasta
└── normal_N06_read2.fasta

0 directories, 12 files
```

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?

A) 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*readNumber

```
>wc -l YourFastaFile.fasta
```

running this command will give you the proper readNumber

```
>grep ">" YourFastaFile.fasta | wc -l
```

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?

A) Mate inner distance is the approximate distance between the reads. You can get this number by:
1) Using **insert size** estimates provided from the library preparation step. --mate-inner-distance= **insert size**-2x(ReadLength)
2) If you don’t have that information, then you can subset the FASTA file and run a quick alignment. Plot the fragment distribution from this subset and use those numbers for the full alignment
3) We were told that the average **insert size** for these samples is 150 bp and the reads are 36bp long. so --mate-inner-distance= 150-2x(36)=78=~80bp

-remember this from our notes?

```
PE reads        R1--------->                                <---------R2
fragment    ~~~=========================================~~~
insert           =========================================
inner mate                ............................................
```

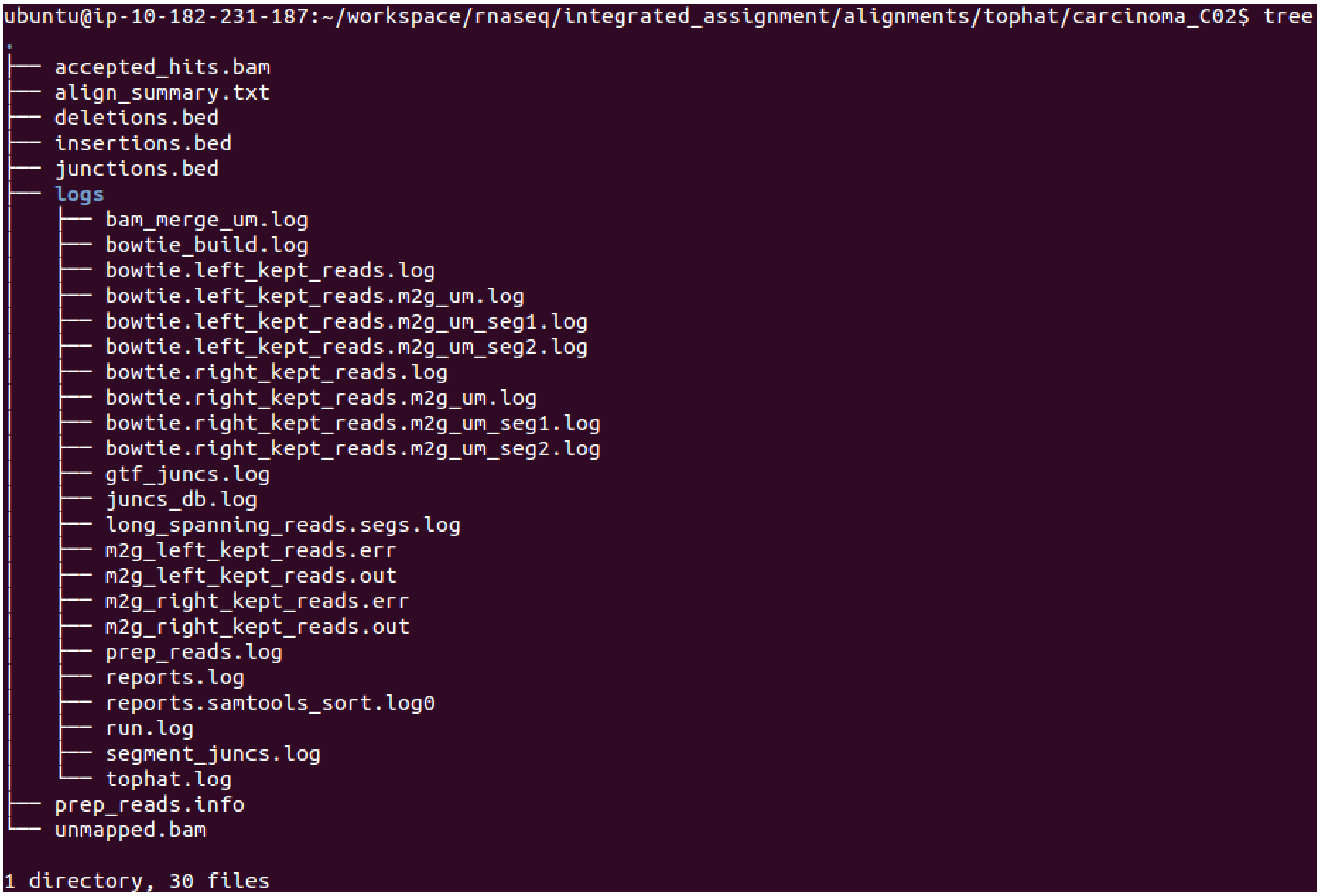
Q6) Considering that the read length in this exercise is 36bp, what should you set the --segment-length to (default is 25bp)?

A) If you keep the default value of 25 bases, Tophat will split each read into 2 segments of 25bp and 11bp lengths. It is preferred to split the read into segments of equal length. Therefore, assigning —segment-length a value of 18 for a 36bp read is recommended. When deciding on a number, try avoiding a split that will result in a very short segment. Short segments might not be uniquely mapped and this can affect your transcript assembly process.

```
cd $RNA_HOME/
export RNA_DATA_DIR=$RNA_HOME/data/
echo $RNA_DATA_DIR
mkdir -p alignments/tophat/trans_idx
cd alignments/tophat
export TRANS_IDX_DIR=$RNA_HOME/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_HOME/refs/hg19/genes/genes_chr9.gtf
--transcriptome-index $TRANS_IDX_DIR/ENSG_Genes $RNA_HOME/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_HOME/refs/hg19/genes/genes_chr9.gtf
--transcriptome-index $TRANS_IDX_DIR/ENSG_Genes $RNA_HOME/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_HOME/refs/hg19/genes/genes_chr9.gtf
--transcriptome-index $TRANS_IDX_DIR/ENSG_Genes $RNA_HOME/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_HOME/refs/hg19/genes/genes_chr9.gtf
--transcriptome-index $TRANS_IDX_DIR/ENSG_Genes $RNA_HOME/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_HOME/refs/hg19/genes/genes_chr9.gtf
--transcriptome-index $TRANS_IDX_DIR/ENSG_Genes $RNA_HOME/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_HOME/refs/hg19/genes/genes_chr9.gtf
--transcriptome-index $TRANS_IDX_DIR/ENSG_Genes $RNA_HOME/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:



```
ubuntu@ip-10-182-231-187:~/workspace/rnaseq/integrated_assignment/alignments/tophat/carcinoma_C02$ tree
.
├── accepted_hits.bam
├── align_summary.txt
├── deletions.bed
├── insertions.bed
├── junctions.bed
├── logs
│   ├── bam_merge_um.log
│   ├── bowtie_build.log
│   ├── bowtie.left_kept_reads.log
│   ├── bowtie.left_kept_reads.m2g_um.log
│   ├── bowtie.left_kept_reads.m2g_um_seg1.log
│   ├── bowtie.left_kept_reads.m2g_um_seg2.log
│   ├── bowtie.right_kept_reads.log
│   ├── bowtie.right_kept_reads.m2g_um.log
│   ├── bowtie.right_kept_reads.m2g_um_seg1.log
│   ├── bowtie.right_kept_reads.m2g_um_seg2.log
│   ├── gtf_juncs.log
│   ├── juncs_db.log
│   ├── long_spanning_reads.segs.log
│   ├── m2g_left_kept_reads.err
│   ├── m2g_left_kept_reads.out
│   ├── m2g_right_kept_reads.err
│   ├── m2g_right_kept_reads.out
│   ├── prep_reads.log
│   ├── reports.log
│   ├── reports.samtools_sort.log0
│   ├── run.log
│   ├── segment_juncs.log
│   └── tophat.log
├── prep_reads.info
└── unmapped.bam

1 directory, 30 files
```

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

A) There are many RNA-seq QC tools available that can provide you with detailed information about the quality of the aligned sample. However, for a simple summary of aligned reads counts you can use samtools flagstat:

samtools flagstat accepted_hits.bam

or

samstat accepted_hits.bam

hint: You can also look for the logs generated by Tophat. These logs provide a summary of the aligned reads.

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

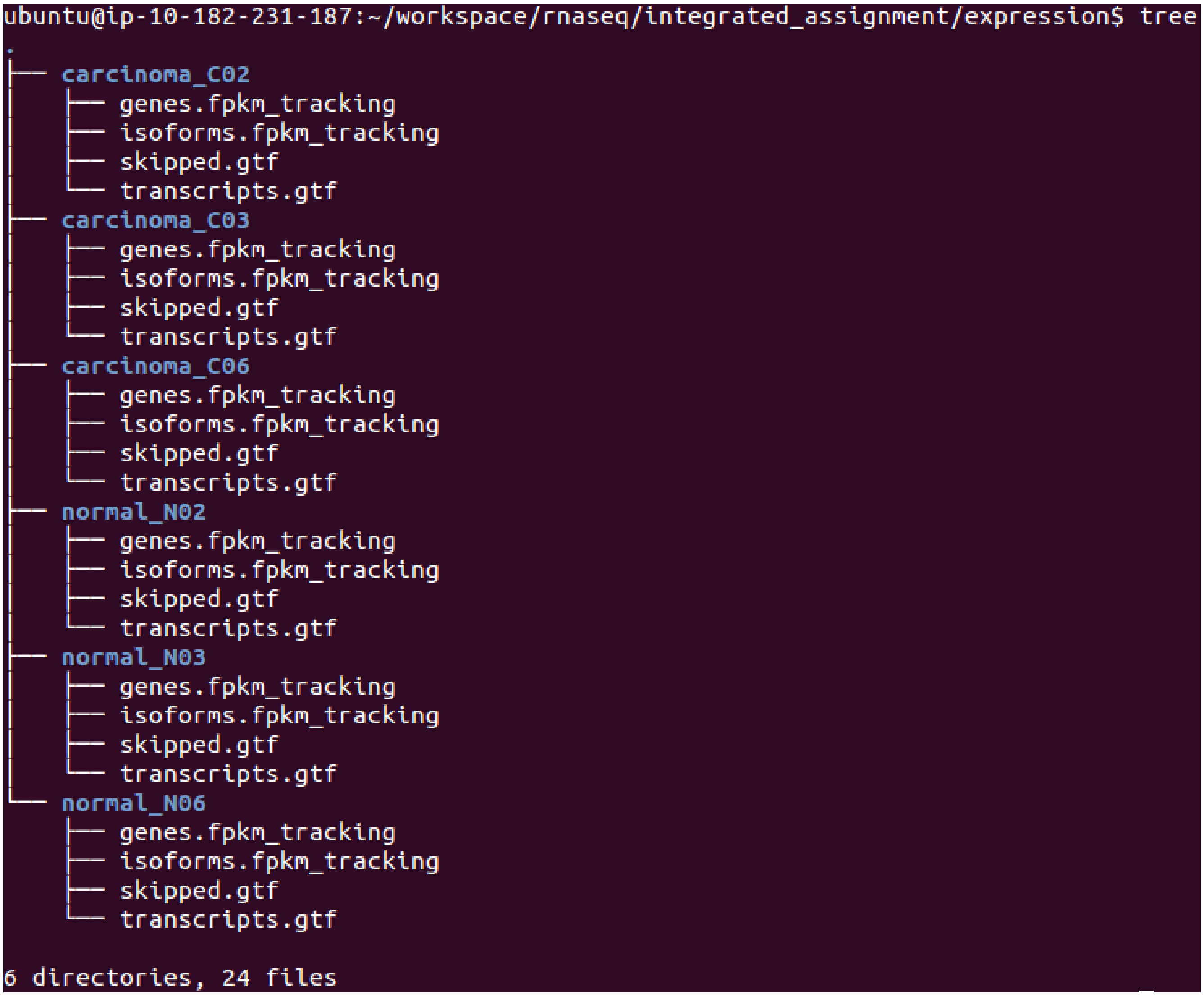
Goals:
- Familiarize yourself with Cufflinks options
- Run Cufflinks to obtain expression values
- Obtain expression values for the gene PCA3

```
cd $RNA_HOME/
mkdir expression
cd expression
```

example (how to run cufflinks for one sample):

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

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

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

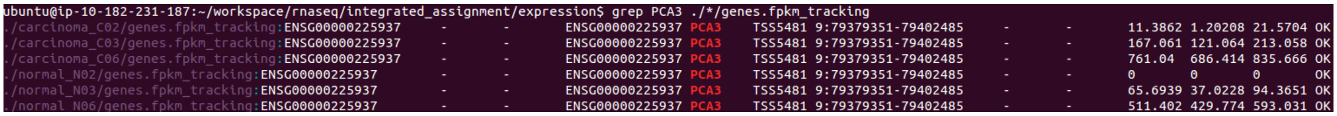


```
ubuntu@ip-10-182-231-187:~/workspace/rnaseq/integrated_assignment/expression$ tree
.
├── carcinoma_C02
│   ├── genes.fpkm_tracking
│   ├── isoforms.fpkm_tracking
│   ├── skipped.gtf
│   └── transcripts.gtf
├── carcinoma_C03
│   ├── genes.fpkm_tracking
│   ├── isoforms.fpkm_tracking
│   ├── skipped.gtf
│   └── transcripts.gtf
├── carcinoma_C06
│   ├── genes.fpkm_tracking
│   ├── isoforms.fpkm_tracking
│   ├── skipped.gtf
│   └── transcripts.gtf
├── normal_N02
│   ├── genes.fpkm_tracking
│   ├── isoforms.fpkm_tracking
│   ├── skipped.gtf
│   └── transcripts.gtf
├── normal_N03
│   ├── genes.fpkm_tracking
│   ├── isoforms.fpkm_tracking
│   ├── skipped.gtf
│   └── transcripts.gtf
└── normal_N06
    ├── genes.fpkm_tracking
    ├── isoforms.fpkm_tracking
    ├── skipped.gtf
    └── transcripts.gtf

6 directories, 24 files
```

Q8) How do you get the expression of PCA3 across the normal and carcinoma samples?

A) Cufflinks generates two expression files: gene level expression and isoform level expression. 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

```
grep PCA3 ./*/genes.fpkm_tracking
```



```
ubuntu@ip-10-182-231-187:~/workspace/rnaseq/integrated_assignment/expression$ grep PCA3 ./*/genes.fpkm_tracking
./carcinoma_C02/genes.fpkm_tracking:ENSG00000225937     -       -       ENSG00000225937 PCA3    TSS5481 9:79379351-79402485     -       -       11.3862 1.20208 21.5704 OK
./carcinoma_C03/genes.fpkm_tracking:ENSG00000225937     -       -       ENSG00000225937 PCA3    TSS5481 9:79379351-79402485     -       -       167.061 121.064 213.058 OK
./carcinoma_C06/genes.fpkm_tracking:ENSG00000225937     -       -       ENSG00000225937 PCA3    TSS5481 9:79379351-79402485     -       -       761.04  686.414 835.666 OK
./normal_N02/genes.fpkm_tracking:ENSG00000225937        -       -       ENSG00000225937 PCA3    TSS5481 9:79379351-79402485     -       -       0       0       0       OK
./normal_N03/genes.fpkm_tracking:ENSG00000225937        -       -       ENSG00000225937 PCA3    TSS5481 9:79379351-79402485     -       -       65.6939 37.0228 94.3651 OK
./normal_N06/genes.fpkm_tracking:ENSG00000225937        -       -       ENSG00000225937 PCA3    TSS5481 9:79379351-79402485     -       -       511.402 429.774 593.031 OK
```

PART 4 -- Differential Expression Analysis ---

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

```
cd $RNA_HOME/expression
```

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

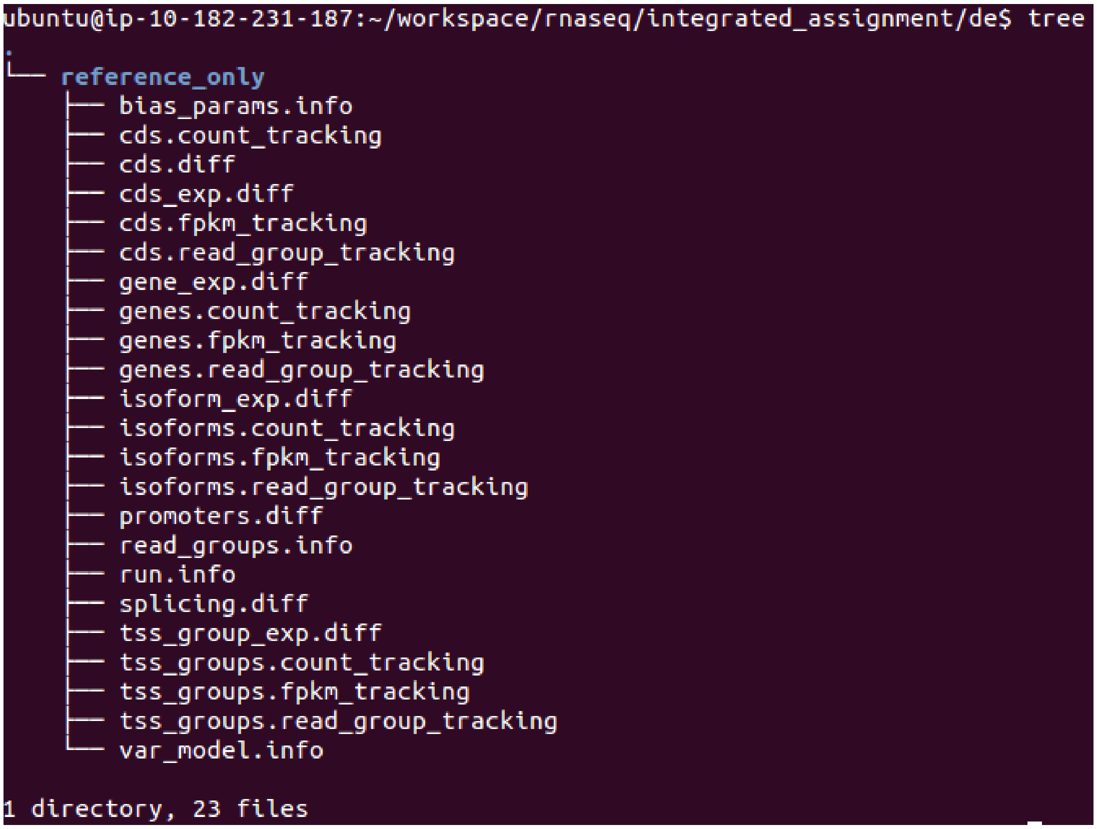
```
cuffmerge -p 8 -o merged -g $RNA_HOME/refs/hg19/genes/genes_chr9.gtf -s
$RNA_HOME/refs/hg19/bwt/9/ assembly_GTF_list.txt
```

```
cd $RNA_HOME/
mkdir de
mkdir de/reference_only
cd $RNA_HOME/alignments/tophat
```

#run cuffdiff to perform comparison

```
cuffdiff -p 8 -L Normal,Carcinoma -o $RNA_HOME/de/reference_only/ --no-update-check
$RNA_HOME/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:



```
ubuntu@ip-10-182-231-187:~/workspace/rnaseq/integrated_assignment/de$ tree
.
└── reference_only
    ├── bias_params.info
    ├── cds.count_tracking
    ├── cds.diff
    ├── cds_exp.diff
    ├── cds.fpkm_tracking
    ├── cds.read_group_tracking
    ├── gene_exp.diff
    ├── genes.count_tracking
    ├── genes.fpkm_tracking
    ├── genes.read_group_tracking
    ├── isoform_exp.diff
    ├── isoforms.count_tracking
    ├── isoforms.fpkm_tracking
    ├── isoforms.read_group_tracking
    ├── promoters.diff
    ├── read_groups.info
    ├── run.info
    ├── splicing.diff
    ├── tss_group_exp.diff
    ├── tss_groups.count_tracking
    ├── tss_groups.fpkm_tracking
    ├── tss_groups.read_group_tracking
    └── var_model.info

1 directory, 23 files
```

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

A) 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?

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

```
cd $RNA_HOME/
mkdir final_results
cd $RNA_HOME/final_results
mkdir reference_only
cp $RNA_HOME/de/reference_only/isoform* reference_only/
cp $RNA_HOME/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?

A) 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.