================ EXAMPLE Calculating expression ==============

We have used reads sequences obtained from a knockdown of the arm and smo gene in the drosophila S2 cell line to estimate gene expression. We used tophat to align the reads. We will now use the output from this step, to look at gene expression in these samples. We will use a program called Cufflinks to do this.

First, take the arm gene knockdown sample:

Task: Assemble expressed genes and transcripts in the ARM knockdown with CUFFLINKS

\* To get help

**cufflinks -h**

\* The output of tophat in step 1 can be found in your home directory at:

**ARM-1\_tophatOutput/accepted\_hits.bam**

\* The annotation Data is

**geneRef/dros\_BD5.25.gff**

\* Run the command:

**cufflinks -G geneRef/dros.gtf --upper-quartile-norm --compatible-hits-norm -p 2 -o ARM-1\_CufflinksOutput ARM-1\_tophatOutput/accepted\_hits.bam**

Review Questions:

- Why assemble transcripts with cufflinks?

- What do the options in the cufflinks command do?

- What is the gtf file and why do I need it?

- what data does the gtf file contain?

- what does the last column in the gtf file contain?

- what is the "tss\_id" tag in the gtf file and what is it used for?

- why is both transcript and gene information in the gtf file?

- What files does cufflinks output and what do the columns mean?

Next, we must estimate the expression in the smo gene knockdown:

Task: Assemble expressed genes and transcripts in the SMO knockdown with CUFFLINKS (similar to above)

=========== EXAMPLE Estimating differential Expression ==========

We have estimated gene expression in SMO and ARM gene knockdowns. We would now like to see if this is expressed differently as one would find in a control. We will now use the output from the cufflinks step, to look at differential gene expression by comparing it to a control. We will use a program called Cuffdiff to do this.

(To save time we have aligned and pre-calculated gene expression in the control samples)

There is 1 replicate of the ARM sample: ARM-1

There are 2 replicates of the CTRL sample: FL1-1, FL2-1

Task: Identify differentially expressed genes in the ARM gene knockdown compared to control.

* Use the drosophila gtf file:

**geneRef/dros\_BD5.25.gff**

\* Used the accepted\_hits.bam output from the tophat alignment

For ctrl these are:

FL1-1\_tophatOutput/accepted\_hits.bam

FL2-1\_tophatOutput/accepted\_hits.bam

And for ARM these are:

ARM-1\_tophatOuput/accepted\_hits.bam

**cuffdiff -o ARM\_vs\_CTRL\_diffOut -b genomeRef/dros\_BD5.25.fa**

**-p 8 –L FL1\_ctrl,arm -u geneRef/dros\_BD5.25.gff FL1-1\_tophatOutput/accepted\_hits.bam, FL2-1\_tophatOutput/accepted\_hits.bam**

**ARM-1\_tophatOuput/accepted\_hits.bam**

Review Questions:

- What normalization strategies are available with cuffdiff?

- What is the -L option for?

- Why are replicates important here?

- What are the output files from cuff\_diff?

Use your linux know how to:

- How many significantly differentially expressed genes are there?

- Can you find the expression of gene "CG7224"?

Task: Use your know-how to estimate estimate differential expression in the smo knockdown compared to control.

\* The output can be found at:

SMO-1\_tophatOuput/accepted\_hits.bam

The controls and other input files are as before.