Trapnell Data Set - Differential Expression (cuffdiff)

BIOL550 - Lab 3 Weekly Report (Week 3)

What I accomplished since the previous report

I produced a set of STAR BAM files that include the `XS` strand tag required by `cuffdiff`, and then ran `cuffdiff` on all six samples (3 C1 replicates and 3 C2 replicates) to complete the Trapnell differential expression step. I exported the `cuffdiff` output directory locally, summarized gene-level results from `gene_exp.diff` using `q_value <= 0.05`, and generated two sanity-check figures (volcano plot + top-DE bar chart).

Results summary

```
Lab 3: Cuffdiff gene-level DE summary
OK tests: 8289
Significant genes (q<=0.05): 265
Top up (by log2FC): Fatp, crc, scf, CTPsyn, Df31
Top down (by log2FC): Nep2, RpS19b, Amy-d, CG6847, Aplip1
```

Methods used (commands + parameters)

STAR was used to generate sorted coordinate BAMs, and `cuffdiff` was used for differential expression on the aligned reads. Then, downstream filtering and plotting were done locally from the `cuffdiff` output tables.

STAR re-alignment (to ensure `XS` tags)

STAR was re-run for all six samples with `--outSAMstrandField intronMotif` so that spliced alignments include `XS` tags. Representative command:

```
STAR \
  --genomeDir /home/pzg8794/star_index/classref_trapnell_zip_bdgp6_84_v2 \
  --runThreadN 4 \
  --sjdbGTFfile "/home/pzg8794/BIOL550/Lab1/Trapnell_Data/Trapnell Data/Drosophila
reference/Drosophila_melanogaster.BDGP6.84.gtf" \
  --readFilesIn \
    "/home/pzg8794/BIOL550/Lab1/Trapnell_Data/Trapnell Data/Raw reads/GSM794483_C1_R1_1.fq.gz" \
    "/home/pzg8794/BIOL550/Lab1/Trapnell_Data/Trapnell Data/Raw reads/GSM794483_C1_R1_2.fq.gz" \
  --readFilesCommand zcat \
  --outSAMtype BAM SortedByCoordinate \
  --outSAMstrandField intronMotif \
  --limitBAMsortRAM 6000000000 \
  --outFileNamePrefix /home/pzg8794/BIOL550/Lab1/star_align_classref_v2_all_xs/GSM794483_C1_R1/
```

Differential expression (`cuffdiff`)

The Drosophila reference directory includes both a GTF and a GFF3; the GTF (`Drosophila_melanogaster.BDGP6.84.gtf`) was used because `cuffdiff` expects GTF2 annotation (GFF3 would require conversion before use). `Cuffdiff` was run on the XS-tagged BAMs (3 replicates per condition) with bias correction enabled via `-b`.

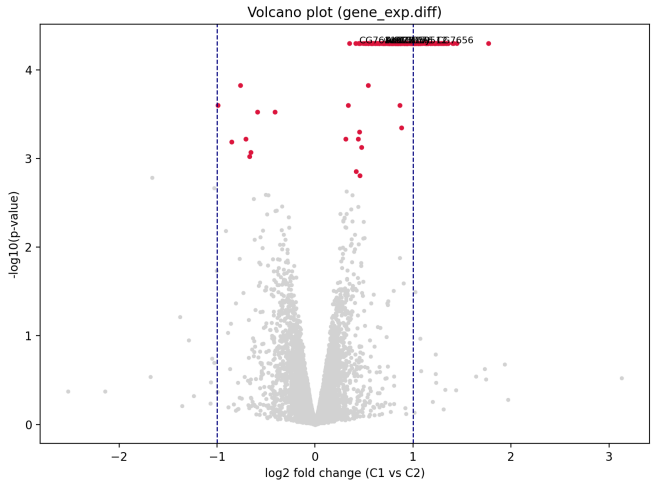
```
/usr/local/bin/cufflinks/cuffdiff \
  -o /home/pzg8794/BIOL550/Lab1/cuffdiff_classref_v2_xs \
  -p 4 \
  -L C1,C2 \
  -b /home/pzg8794/refs/classref_bdgp6_84_ids_v2.fa \
  "/home/pzg8794/BIOL550/Lab1/Trapnell_Data/Trapnell Data/Drosophila reference/Drosophila_melanogaster.BDGP6.84.gtf" \

/home/pzg8794/BIOL550/Lab1/star_align_classref_v2_all_xs/GSM794483_C1_R1/Aligned.sortedByCoord.out.bam,/home/pzg8794/BIOL550/Lab
1/star_align_classref_v2_all_xs/GSM794484_C1_R2/Aligned.sortedByCoord.out.bam,/home/pzg8794/BIOL550/Lab1/star_align_classref_v2_
all_xs/GSM794485_C1_R3/Aligned.sortedByCoord.out.bam \

/home/pzg8794/BIOL550/Lab1/star_align_classref_v2_all_xs/GSM794486_C2_R1/Aligned.sortedByCoord.out.bam,/home/pzg8794/BIOL550/Lab
1/star_align_classref_v2_all_xs/GSM794487_C2_R2/Aligned.sortedByCoord.out.bam,/home/pzg8794/BIOL550/Lab1/star_align_classref_v2_
all_xs/GSM794488_C2_R3/Aligned.sortedByCoord.out.bam
```

Downstream summary (local)

From `gene_exp.diff`, I filtered valid tests by keeping rows where `status == "OK"`, then defined significance at `q_value <= 0.05`. Then, I generated a volcano plot (log2FC vs. −log10 p-value) and a top-gene bar chart (ranked by log2FC) as sanity checks on effect-size distribution and signal presence.

Volcano plot

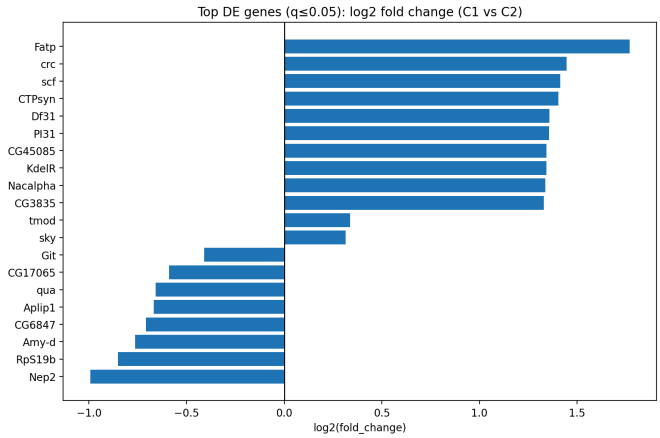
Generated from `gene_exp.diff`



Volcano plot: `gene_exp.diff`

Top 20 DE genes bar chart

Top up/down genes ranked by log2FC



Top DE genes (by log2FC)

Problems encountered

The main issue was that the initial STAR BAMs from Lab 2 did not contain `XS` strand tags on spliced reads, which caused `cuffdiff` to fail. This required going back and re-aligning all six samples with `--outSAMstrandField intronMotif` to produce XS-tagged BAMs.

A second point that required attention was annotation choice: the reference directory contains both GTF and GFF3. Since `cuffdiff` expects GTF2, I used `Drosophila_melanogaster.BDGP6.84.gtf` (rather than the GFF3 file) to avoid format incompatibility and conversion.

Goals for the coming week

Next week I will interpret the `cuffdiff` outputs more thoroughly by reviewing the strongest DE genes, understanding and checking directionality for the C1 vs. C2 contrast, and confirming replicate consistency. Also, I will further explore CummeRbund output to summarize results, and begin first steps on the project dataset pipeline for the group project.