Sophia Bevans
BF528: Transcriptional Regulation of Mammalian Cardiac Regeneration with mRNA-Seq

PROGRAMMER

Introduction

O'Meara et al. (2015) used RNA-Seq analysis to derive a transcriptional signature of cardiac myocyte regeneration by comparing the transcription patterns of cardiac myocyte samples under different conditions. All but one sample from O'Meara et al. (2015) was prepared for analysis already, so in this task I prepared the final sample (vP0_1) for use in differential expression analysis with the previously prepared samples. After FASTQ files are extracted from the study samples, they must be aligned to the mouse genome to quantify gene expression within the samples. Since paired end reads were used, two FASTQ files were aligned to the reference genome for this sample. After aligning the reads, the results were checked for quality. Next, the alignments were quantified in terms of fragments per kilobase of sequence per million (fpkm) mapped reads, which normalizes the aligned reads for further analysis. Finally, differentially expressed genes between the samples were identified.

Methods

To align the paired FASTQ files to the mouse genome, samtools, bowtie2, boost, python2 and tophat were loaded on the SCC (module load samtools bowtie2 boost tophat python2), and tophat was used to align the paired P01 reads to the available reference mouse genome (mm9) using the arguments provided by O'Meara et al. (2015). Tophat was run as a batch job on the scc. Tophat results were examined using samtools (samtools flagstat P0_1_tophat/accepted_hits.bam). Next, rseqc were loaded on the SCC, and the bam file resulting from tophat alignment was indexed using samtools (samtools index P0_1_tophat/accepted_hits.bam). Finally, quality control checks were run on the accepted hits using the programs geneBody_coverage.py (geneBody_coverage.py -i P0_1_tophat/accepted_hits.bam -r /project/bf528/project_2/reference/annot/mm9.bed -f "png" -o geneBody), inner_distance.py (inner_distance.py -i P0_1_tophat/accepted_hits.bam -r /project/bf528/project_2/reference/annot/mm9.bed), and bam_stat.py (bam_stat.py -i P0_1_tophat/accepted_hits.bam).

To count the reads mapped to annotated genome regions, cufflinks was loaded on the SCC, and run on the accepted hits from tophat, using the arguments provided by O'Meara et al. (2015). The quantified alignments in fpkm format were compared to the remaining three samples (Ad_1, Ad_2, P0_2) using cuffdiff to identify differentially expressed genes.

Results

The tophat alignment produced 49,706,999 quality check-passed reads (100%), and 0 reads which failed the quality check. All reads were mapped to the genome, and 41,389,334 reads were paired in sequencing (83%), with 29,422,646 (71.09%) reads properly paired (mapped to the same chromosome and oriented correctly). 39,946,472 reads were mapped with the paired read mapped as well (80%). Overall, these are good alignment results for these reads, and this alignment can be dependably used for differential expression analysis.

The output from geneBody_coverage.py is shown in Figure 1. From this figure, we can see that the majority of the gene body has an RNA-seq coverage over 50%, with only the bottom 0-20 percentile having lower coverage than 50% (Figure 1). The output from inner_distance.py

shows the distribution of inner distance (insert size between read 1 and read 2) of RNA-seq fragments. The insert size distribution is centered around roughly 75 base pairs, and is slightly right skewed, with a range of -50bp to 250bp (Figure 2). This distance between reads should not have any negative impact on analysis of the data and passes the quality control check. The output from bam_stat.py showed no reads which failed the quality check, no duplicated reads, and no unmapped reads. Of the total (49,706,999) reads, 27,972,916 were mapped in proper pairs, mapped to four different chromosomes.

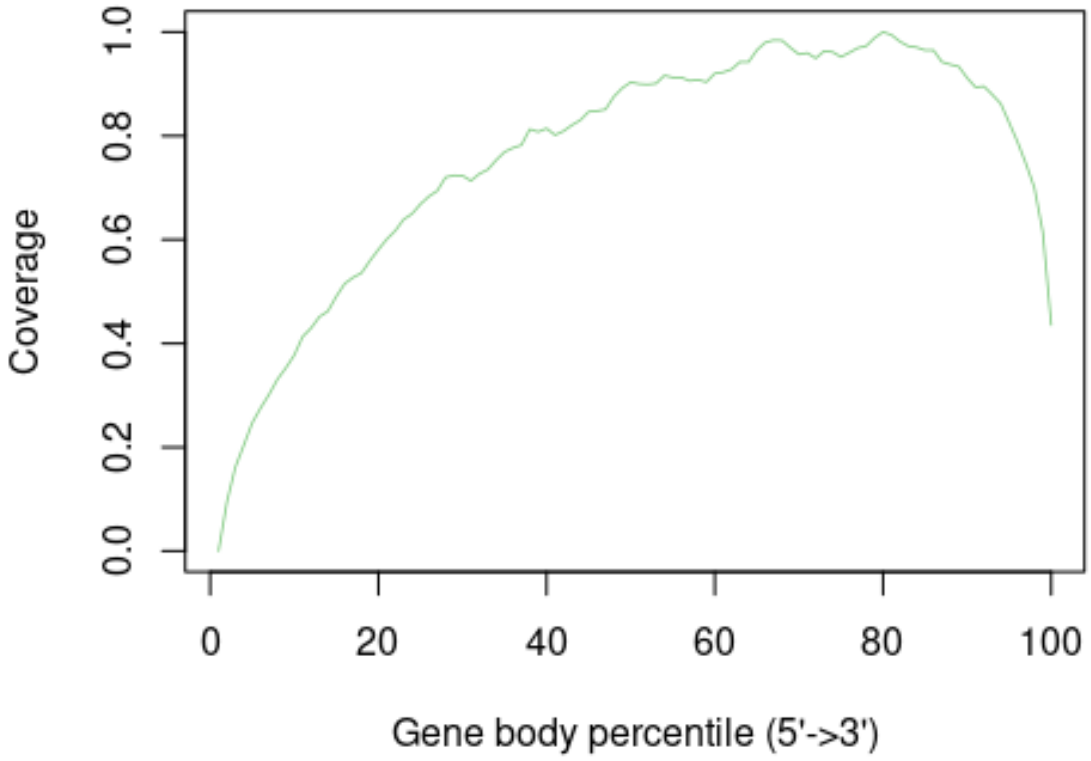

Figure 1. Results of geneBody_coverage Quality Check.

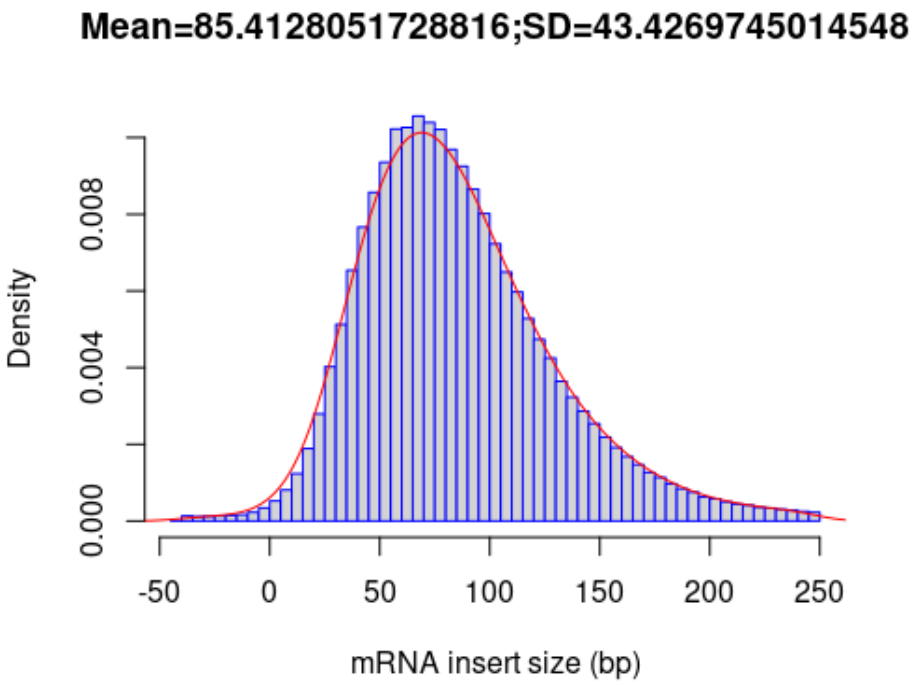

Figure 2. Results of inner_distance Quality Check.

The fpkm quantified alignments from cufflinks were visualized with a histogram (Figure 3). Fpkm values of 0 were removed, since a value of 0 means essentially no expression, which filtered out 3,017 genes. A total of 20,487 identified genes remained after filtering. I additionally removed fpkm values above 500 for visualization, which filtered out another 166 genes.

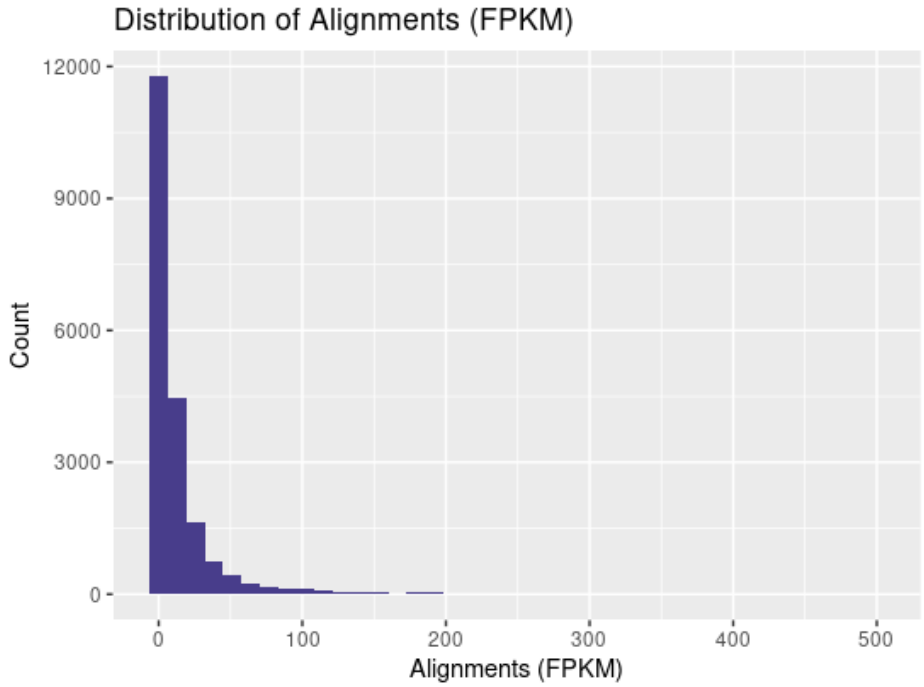

Figure 3. Distribution of FPKM Quantified Alignment Values.

Sophia Bevans
BF528: Transcriptional Regulation of Mammalian Cardiac Regeneration with mRNA-Seq

Discussion

The purpose of this task was to prepare samples in FASTQ format for use in differential expression analysis. The reads were aligned to the mouse genome, the alignment was assessed for quality, and then quantified in terms of fpkm mapped reads, a method of normalization for enrichment analysis downstream. The quality checks showed a good alignment to the genome, decent sequence coverage in the gene body, a relatively normal distribution of inner distances between fragment reads. The reads stood up to all quality metrics and had no issues with duplication or unmapped reads. Therefore, the genes identified through the alignment and counting of mapped reads are identified with high confidence. The distribution of fpkm mapped reads was visualized, another quality check, and finally differential expression analysis between samples was performed using the genes identified. The quality metrics assessed throughout this process demonstrate the trustworthiness of the differential expression results.

Sophia Bevans
BF528: Transcriptional Regulation of Mammalian Cardiac Regeneration with mRNA-Seq

ANALYST

Introduction

Once differentially expressed genes are identified, many questions remain about the quantity and magnitude of differential expression between the samples, and the function of the differentially expressed genes. In this section, I explored the differential expression results generated above by identifying the top differentially expressed genes, the number of significantly differentially expressed genes, and visualizing the distribution of log2 fold changes among these genes. Furthermore, I used functional annotation clustering to group significantly up- and down-regulated genes into functionally similar clusters. This analysis replicates the findings of O'Meara et al. figure 1B (2015) where differentially expressed genes during postnatal and adult cardiac myocyte differentiation are compared.

Methods

For this analysis, I used the R package tidyverse. I read in the differential expression results from cuffdiff and arranged the rows by ascending q value. I additionally cleaned the data by fixing column names and data types. Next, I created a table with the top ten differentially expressed genes, by finding the top 10 log2fold changes of highest magnitude within the rows with the smallest q value. I plotted a histogram of the log2 fold change values from the differentially expressed genes. Then, I created a subset of the differential expression results which included only significant genes (these genes all had $q < 0.05$ and $p < 0.005$). I visualized the distribution of the log2 fold change values of these genes by plotting another histogram. Using the subset of significant genes, I split this subset into upregulated and downregulated genes, and saved the two sets of gene names as csv files. Finally, I used DAVID functional annotation clustering to cluster each enriched gene set. I summarized the top five annotation clusters for upregulated and downregulated genes by selecting an enrichment term that best fit the cluster and reporting the enrichment score.

Results

The top ten differentially expressed genes between postnatal and adult samples are reported in Table 1, organized by log2 fold change value (magnitude of differential expression).

gene	sample_1	sample_2	value_1	value_2	log2(fold_change)	p_value	q_value
Tnni1	P0	Ad	1339.75	1.47349	-9.82851	5E-05	3.18974E-04
Gm2078,Mir1895	P0	Ad	1.25433	362.473	8.17481	5E-05	3.18974E-04
Myl7	P0	Ad	331.604	1.15256	-8.16848	5E-05	3.18974E-04
Xist	P0	Ad	36.2544	0.246985	-7.19759	5E-05	3.18974E-04
Ces1d	P0	Ad	0.294449	33.7372	6.84018	5E-05	3.18974E-04
Myl4	P0	Ad	283.642	2.58695	-6.77668	5E-05	3.18974E-04
2210408F21Rik	P0	Ad	0.444102	48.0354	6.75706	5E-05	3.18974E-04
H19,Mir675	P0	Ad	1176.81	11.14	-6.72299	5E-05	3.18974E-04
Tuba8	P0	Ad	0.855118	79.8964	6.54586	5E-05	3.18974E-04
C7	P0	Ad	0.228574	20.3859	6.47877	5E-05	3.18974E-04

Table 1. Top Ten Differentially Expressed Genes Between Postnatal and Adult Samples During Cardiac Myocyte Differentiation.

The distribution of log2 fold change values for all differentially expressed genes is centered around 0, with most values falling at or near 0 (Figure 4). Conversely, the distribution of significantly differentially expressed genes is bimodal, with no log2 fold change values falling at 0 and a peak of values at approximately -1 and 1 (Figure 5).

Figure 4. Distribution of Log2 Fold Changes of All Differentially Expressed Genes

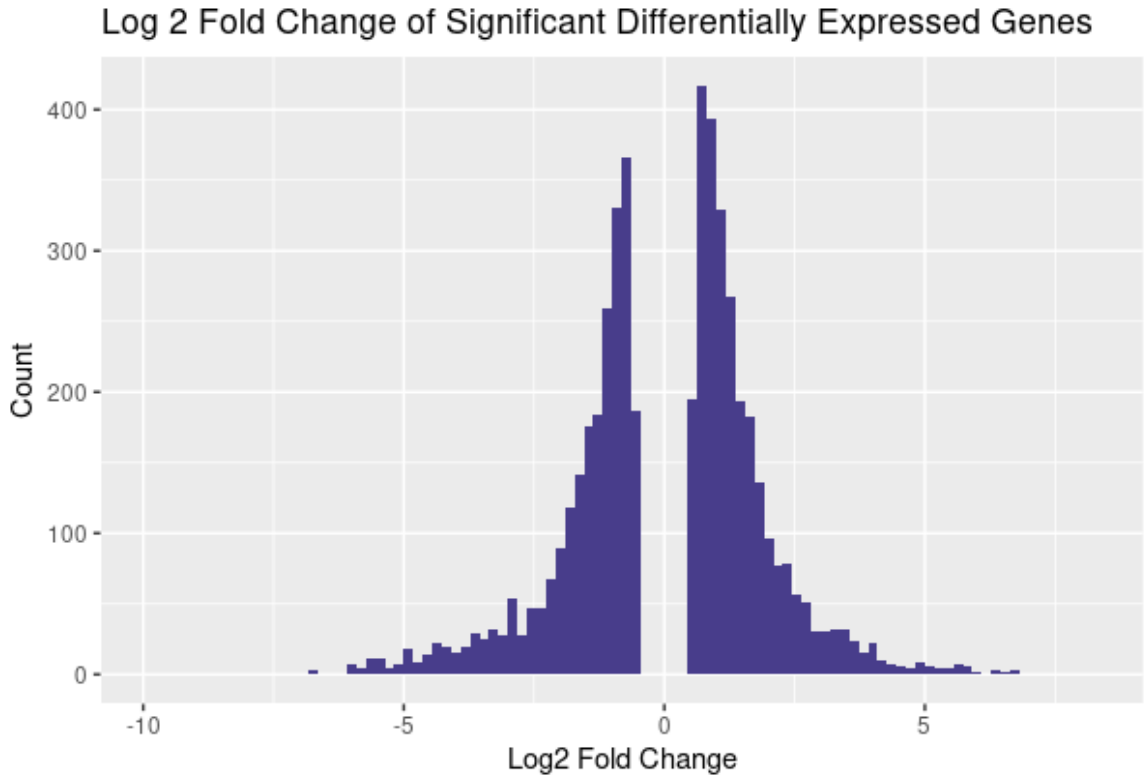

Figure 5. Distribution of Log2 Fold Changes of Significantly Differentially Expressed Genes.

Of the significantly differentially expressed genes, there were 2,757 significantly upregulated genes and 2,531 significantly downregulated genes. DAVID functional annotation clustering found 677 up-regulated clusters and 635 down-regulated clusters. The top five clusters and enrichment scores are reported in Table 2.

Up-Regulated		**Down-Regulated**	
Enrichment Term	*Score*	*Enrichment Term*	*Score*
Mitochondria	54.67	Cell cycle / mitosis	31.26
Nucleoside Metabolic Process	24.48	chromosome	21.06
Respiration	23.42	Regulation processes	20.57
Metabolic processes	22.94	Chromosome organization	17.57
Extracellular	14.92	Cell organization	16.11

Table 2. DAVID Functional Annotation Cluster Results for Significantly Up-Regulated and Down-Regulated Differentially Expressed Genes.

Discussion

Differential expression analysis allows us to use RNA-seq results to compare gene expression patterns between samples under different conditions. O'Meara et al. (2015) compared the common and uniquely up-regulated and down-regulated differentially expressed genes between postnatal and adult mice during cardiac myocyte differentiation in Figure 1B, and I replicate this

portion of their analysis here. As expected, I found the distribution of all Log2 fold changes values to be near or at 0, with significant expression having Log2 fold change values greater or less than 0, depending on the direction of regulation. My cluster results from DAVID analysis did not precisely replicate the results of O'Meara et al (2015, Figure 1B), but there was lots of overlap between my findings and these results. Specifically, my top enrichment term for upregulated genes was mitochondria, which is the same as the top result reported in Figure 1B (O'Meara et al 2015). Additionally, respiration and metabolism both appear in my upregulated enrichment terms and the terms reported by O'Meara et al. (2015), and cell cycle was among the enrichment terms for my downregulated genes and those reported in the paper. Interestingly, the enrichment scores for my top enrichment terms were much higher than the scores reported by O'Meara et al (2015, Figure 1B). Ultimately, these differential expression analysis results largely replicate the findings of the paper, with some discrepancies that may be due to a different process for selecting the top enrichment terms.

Sophia Bevans
BF528: Transcriptional Regulation of Mammalian Cardiac Regeneration with mRNA-Seq

References

O'Meara CC, Wamstad JA, Gladstone RA, Fomovsky GM, Butty VL, Shrikumar A, Gannon JB, Boyer LA, Lee RT. Transcriptional reversion of cardiac myocyte fate during mammalian cardiac regeneration. Circ Res. 2015 Feb 27;116(5):804-15. doi: 10.1161/CIRCRESAHA.116.304269. Epub 2014 Dec 4. PMID: 25477501; PMCID: PMC4344930.