BIOSTATE.AI BIOINFORMATICS CODING CHALLENGE REPORT

by Harsha Pincha

NOVEMBER 8, 2024

PH: 9804221115

pinchaharsha@gmail.com

Tools and dependencies:

Task	Tools	Version
Quality control	FastQC	0.12.1
	MultiQC	1.25
Adapter trimming	Fastp	0.23.4
Alignment	STAR aligner	2.7.8a
	Genome assembly (given)	GRCm39.primary_assembly.genome.fa
	GTF annotation (given)	gencode.vM35.basic.annotation.gtf
DGE analysis	R	4.3.3
	annotation	org.Mm.eg.db

1.a. Quality Control

Attached-

Raw reads QC: multiqc_report_raw_reads.html

1.b. Adapter trimming

Sample	No. of reads before trimming	No. of reads after trimming	Reads with adapter trimmed	Percentage of reads trimmed	Percentage of Reads with adapter trimmed
Heart_ZT0_1	39655191	39246723	5665258	**1.030049256**	**14.28629609**
Heart_ZT0_2	32275135	32088345	4251542	**0.578742738**	**13.17280935**
Heart_ZT12_1	41767679	41450425	3245220	**0.759568182**	**7.769691967**
Heart_ZT12_2	15930786	15772581	12529958	**0.993077178**	**78.65247829**
Liver_ZT0_1	26335934	26164105	1889172	**0.652450754**	**7.173362448**
Liver_ZT0_2	39706568	38991826	3329194	**1.800059879**	**8.384491956**
Liver_ZT12_1	18807831	18733004	1954022	**0.397850236**	**10.38940641**
Liver_ZT12_2	13787431	13731249	662586	**0.40748708**	**4.805724866**

The trimming process has resulted in better per sequence GC content, in addition to trimming and removing adapter contamination.

Attached-

QC after adapter trimming: multiqc_report_postTrimming.html

1. c & d Genome preparation, Alignment and Mapping

Attached-

Final QC report as in: multiqc_report_postAlignment.html

Although the duplication rates are high for the samples and the number of sequences are low for 3 samples, the percentage of mapping and the percentage of uniquely mapped reads (~80%) is high. This may be beneficial for other downstream analyses like CNV calling and Global Splice variant analyses, etc.

2. Statistical analysis

2.a) Data reproducibility and pattern of variation

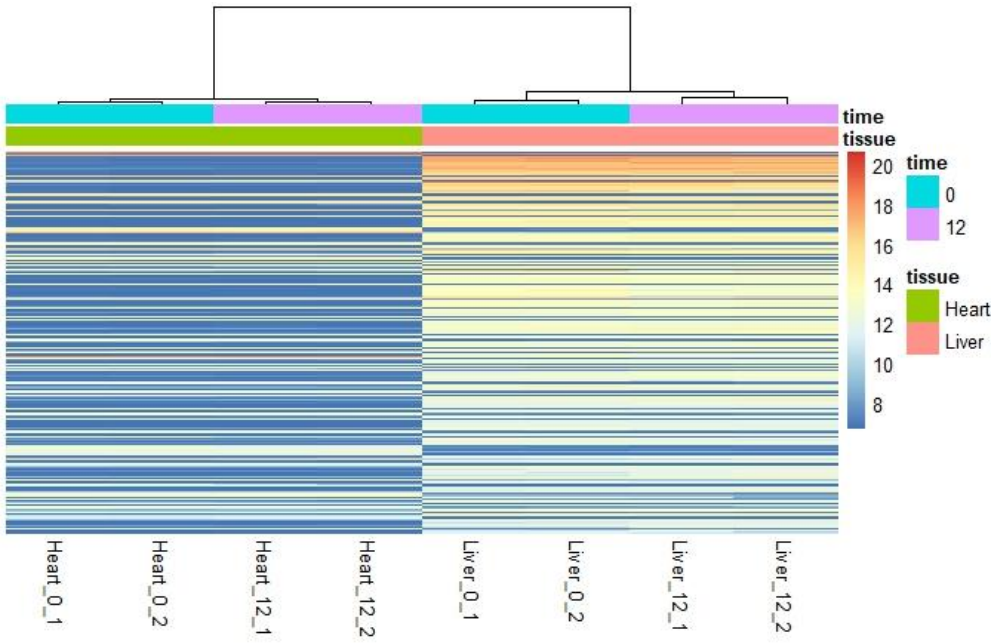

Figure 1: Hierarchical clustering of samples based on top 500 variable gene counts

The above heatmap clusters the samples based on the top 500 variable genes. This indicates that the clustering or transcriptomic differences in the samples are primarily dictated by the tissue type.

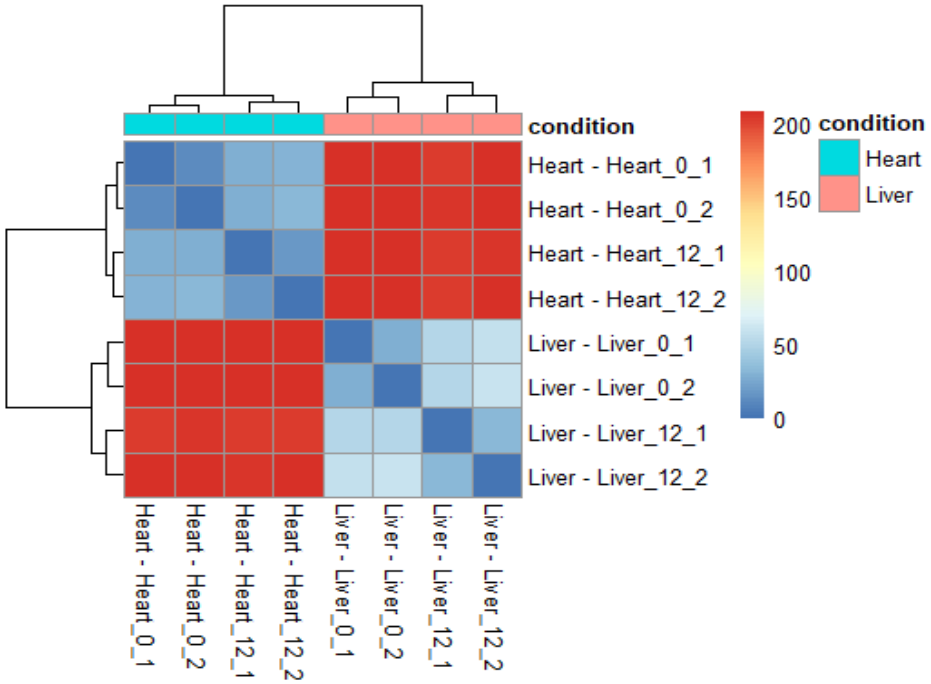

Figure 2: Sample distance matrix heatmap

The sample distance matrix depicts high correlation between the biological replicates, thus indicating efficient reproducibility of data.

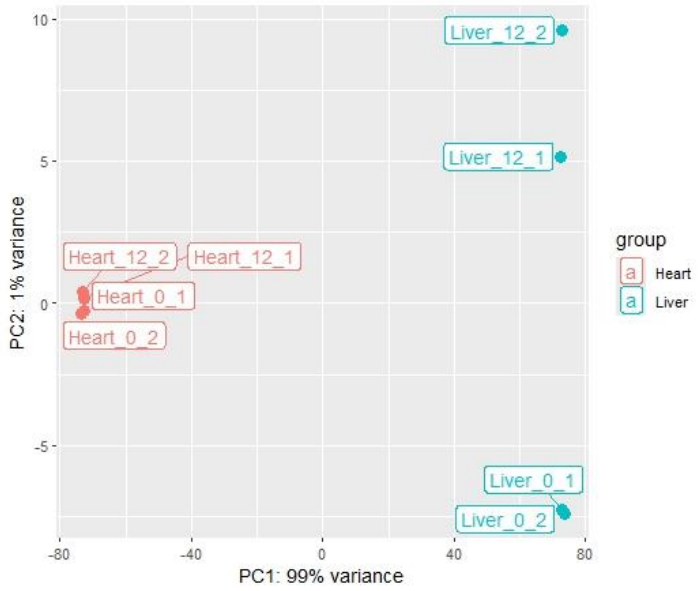

Figure 3: Principal component Analysis (PCA plot) based on raw counts

The PCA plot indicates that the highest variation in the transcriptomic profile of the samples is due to the difference in tissue type. Further, the effect of Sampling time as a cause of variation is more prominent in the liver tissue, than in the heart, making it the principal component with second highest variation.

2.b) Differential Expression Analysis

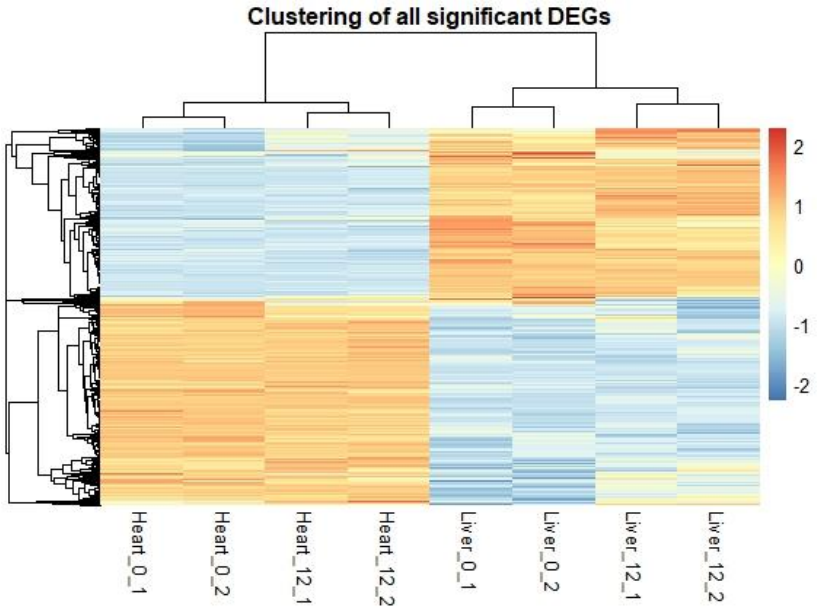

Figure 4: Hierarchical clustering of significantly differentially expressed genes, across all contrasts

All the significantly differentially expressed genes collectively depict prominent separation and clustering of tissue-wise samples. The difference between the transcriptomics profile between the sampling times is also observed.

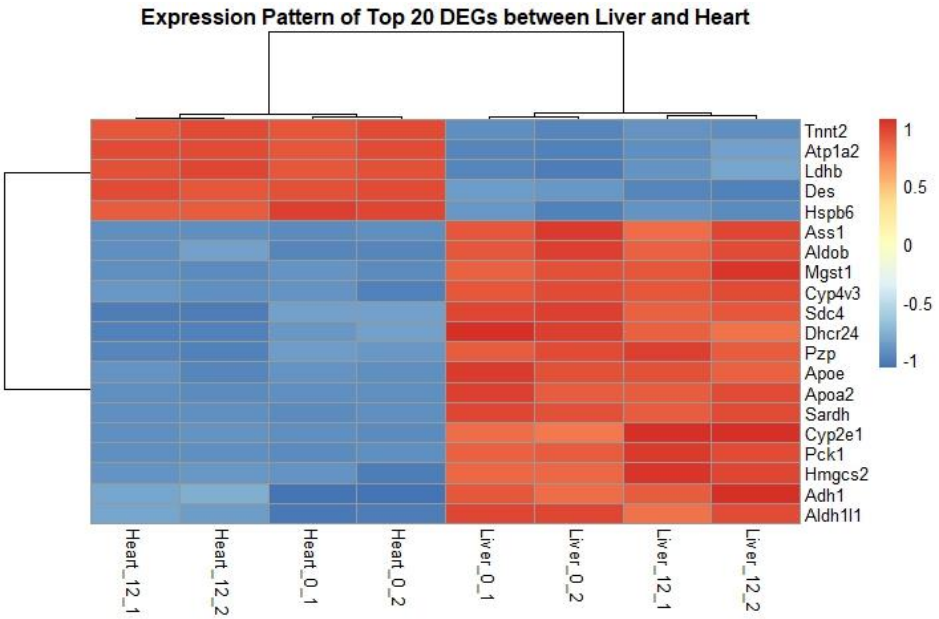

Figure 5: Heirarchical clustering heatmap of Top 20 DEGs between Liver and Heart

Figure 5 depicts prominent clustering of Heart and Liver samples based on only the Top 20 significant DEGs. The genes have been annotated from org.Mm.eg.db. Additionally, fewer genes are upregulated in Heart in comparison to that in Liver.

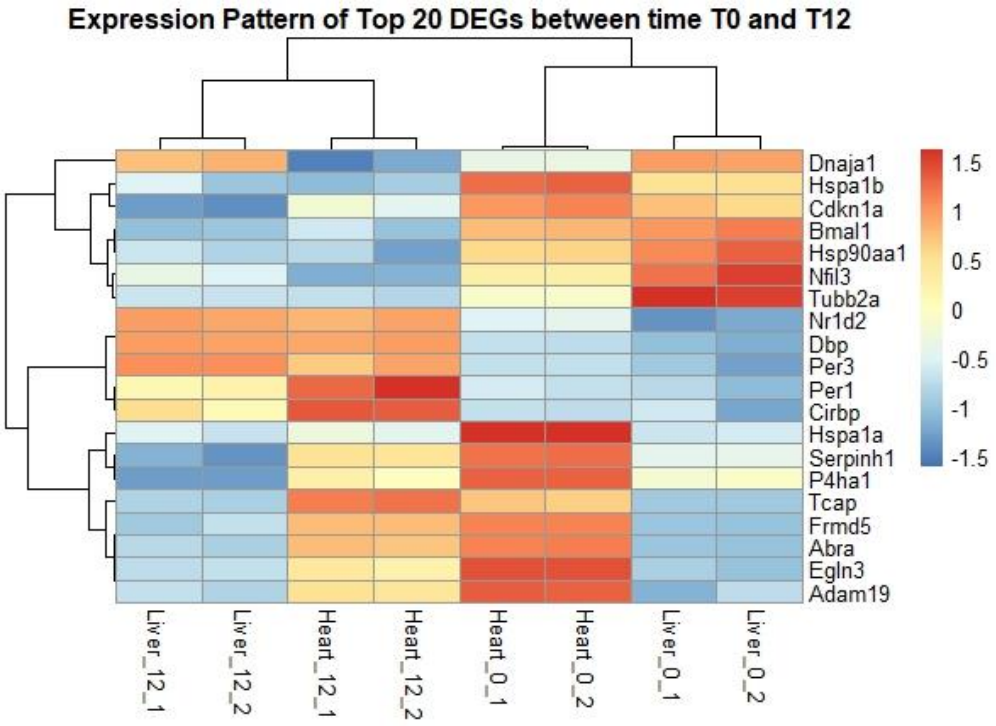

Figure 6: Heirarchical clustering based on top 20 DEGs between sampling time T0 and T12 across

While the samples cluster prominently based on the sampling time, the differential expression between tissues is more prominent at T0 in comparison to that at T12

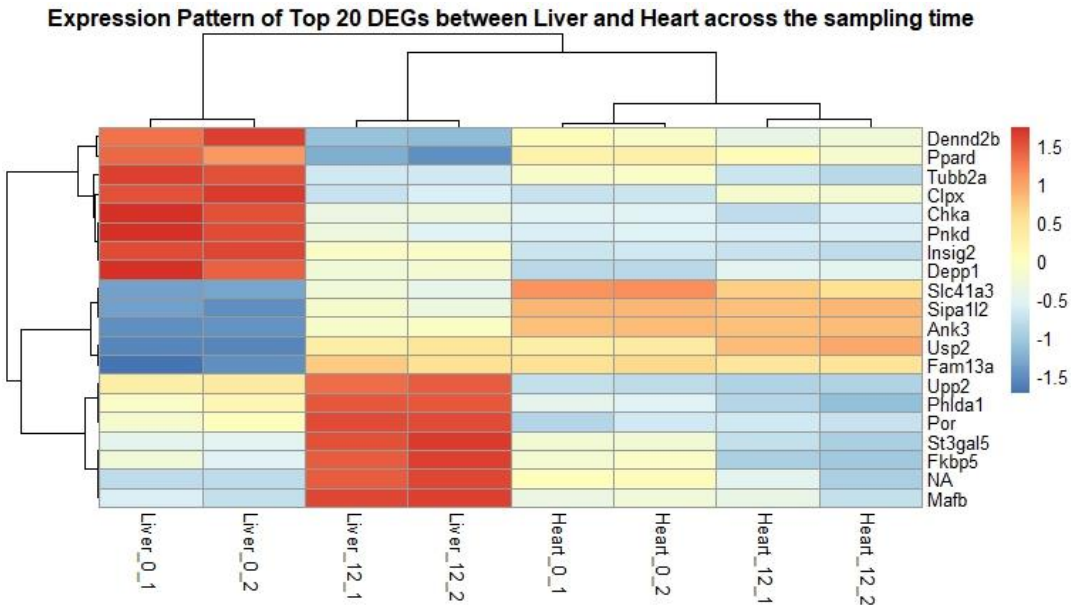

Figure 7: Hierarchical clustering based on Top 20 DEGs between Liver and Heart across the sampling times

The clustering based on the differentially expressed genes from the interaction of contrasting tissue and time, again emphasis that the samples predominantly cluster based on the tissue type. Also, the difference due to sampling time is more prominent in Liver tissue than in Heart.

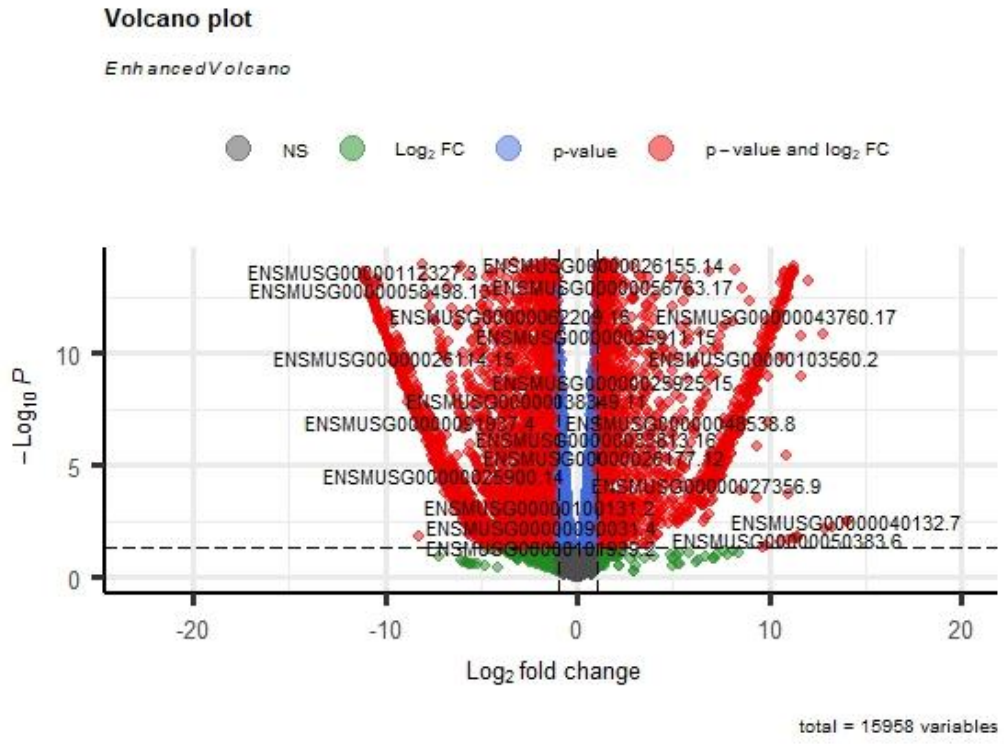

Figure 8: Volcano plot depicting the fold change and confidence of the DEGs between the two tissues

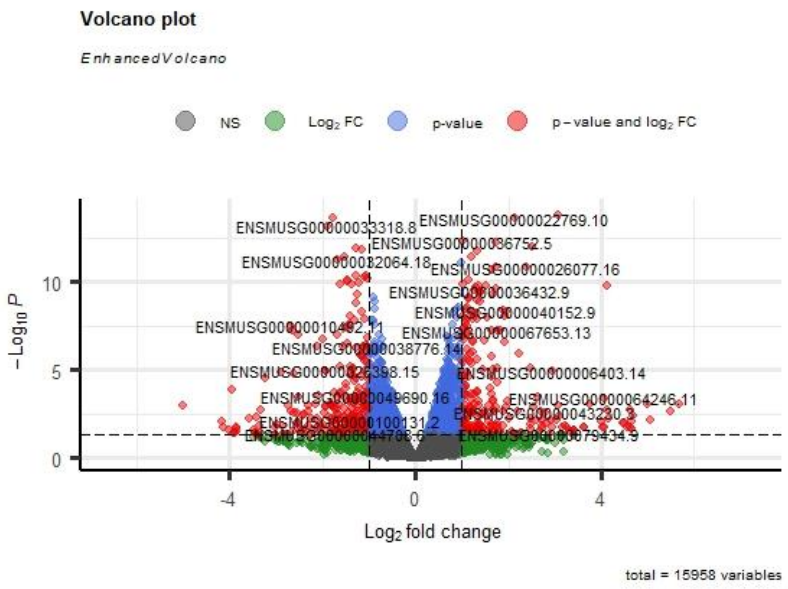

Figure 9: Volcano plot depicting the fold change and confidence of the DEGs between the two sampling time points

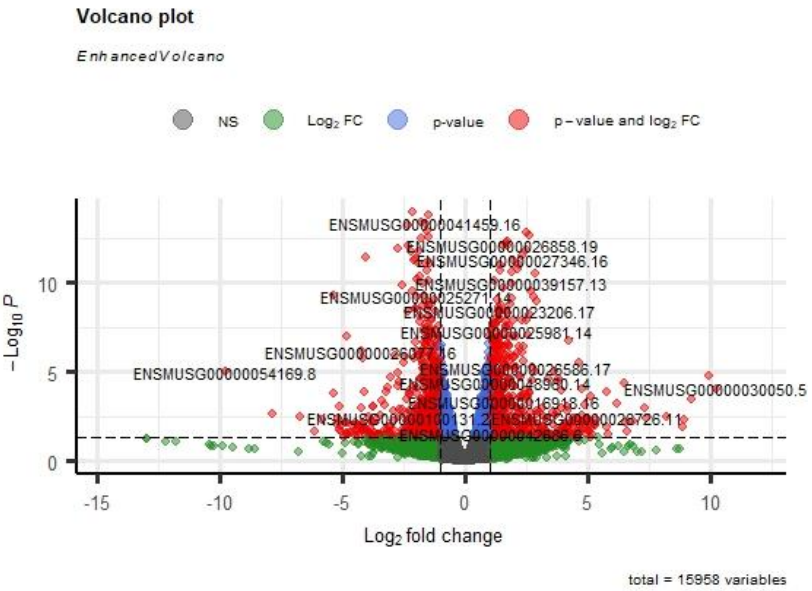

Figure 10: Volcano plot depicting the fold change and confidence of the DEGs across the tissues over the two sampling time points

A comparison of the volcano plots between all the groups depicts that with the same cutoff for significance of Differentially Expressed Genes (DEGs) was analysed. There are more number of genes upregulated or downregulated between the tissue types – Heart and Liver, in comparison to the contrast between the two sampling times – 0 and 12. In addition, the contrast between samples, upon considering the transcriptome differences between Heart and Liver over the sampling time difference, also depicts more number of significantly differentially expressed genes than the contrast between sampling time alone.

2.c) Functional Enrichment Analysis

a) GO Molecular functions analysis:

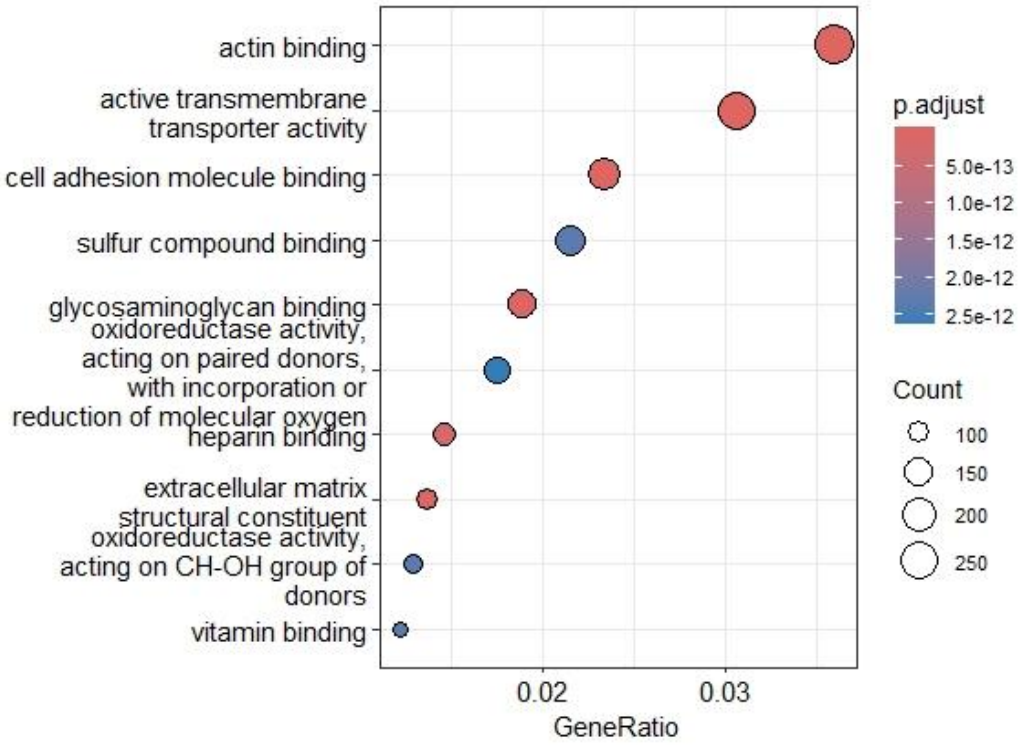

Figure 11: Bubble plot of Molecular Functions from GO that are enriched from the DEGs between the tissues

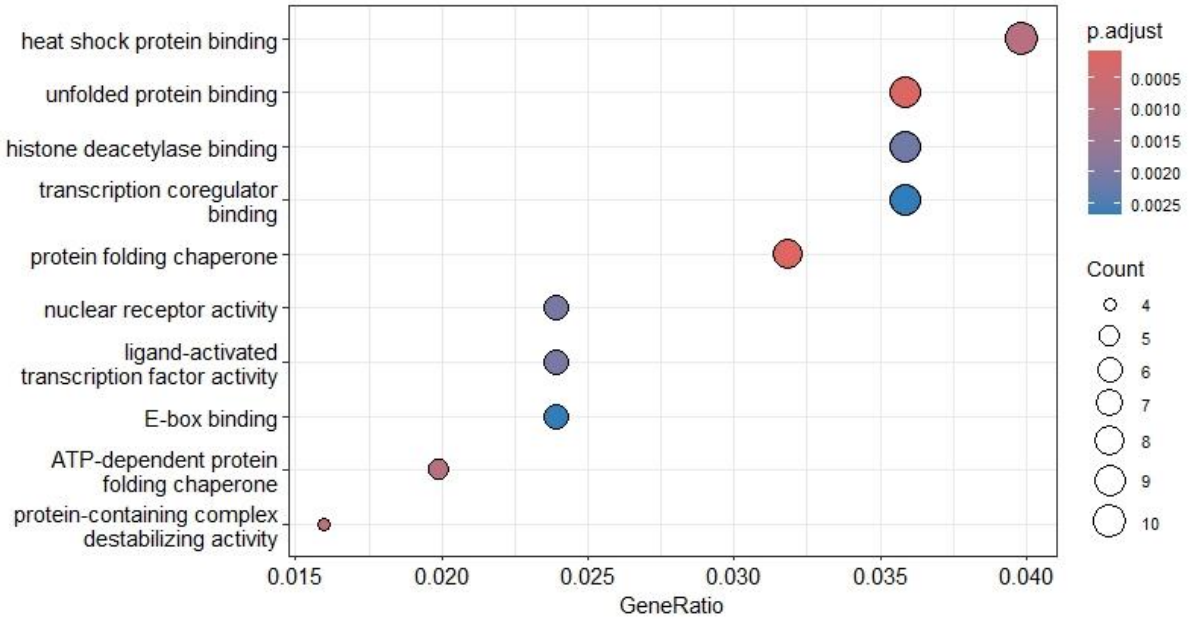

Figure 12: Bubble plot of Molecular Functions from GO that are enriched from the DEGs between the sampling time points

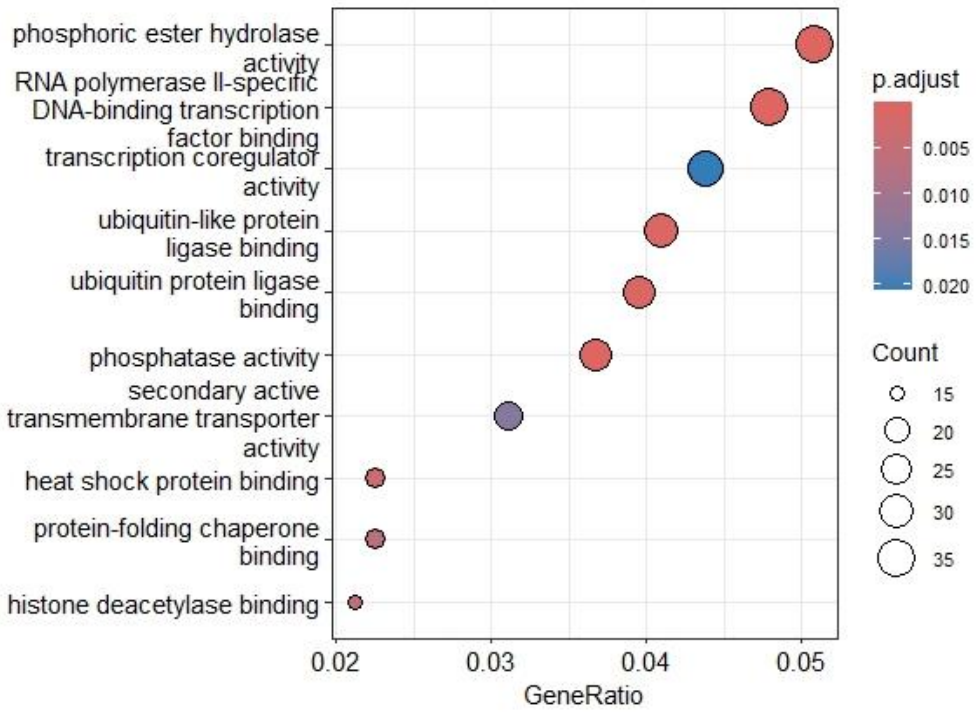

Figure 13: Bubble plot of Molecular Functions from GO that are enriched from the DEGs between across the interaction of tissue and time contrasts

The following observations can be made from the GO – Molecular Functions enrichment results:

1. Liver vs Heart:

Actin binding and active transmembrane transporter activity are highly enriched with highly significant p-values. This suggests that genes associated to cytoskeletal dynamics and transport across membranes are differently expressed between the mouse Liver and Heart.

Other enriched functions include cell adhesion molecule binding, sulfur compound binding and heparin binding. Extracellular matrix structural constituent and oxidoreductase activity terms indicate roles in structural support and redox reactions, which may be different for the tissue specific physiology.

2. Sampling time:

Protein folding and stress response functions are strongly enriched, as seen with terms like "heat shock protein binding," "unfolded protein binding," and "protein folding chaperone." These terms suggest an increased demand for protein folding and stabilization, possibly due to cellular stress responses that vary over the two sampling times.

Histone deacetylase binding and transcription coregulator binding enrichment indicates epigenetic and transcriptional regulation changes, implying that gene expression may be modulated in response to time-dependent factors. Protein maintenance and signalling molecule were also regulated between the sampling times.

3. Interaction between the tissue and time contrast:

The enrichment of GO molecular functions associated with transcription, translation and histones indicate differences in developmental signals along with environmental changes, which correlates with functional and anatomical differences between the two tissues, and the changes over sampling time.

b) KEGG Pathway analysis

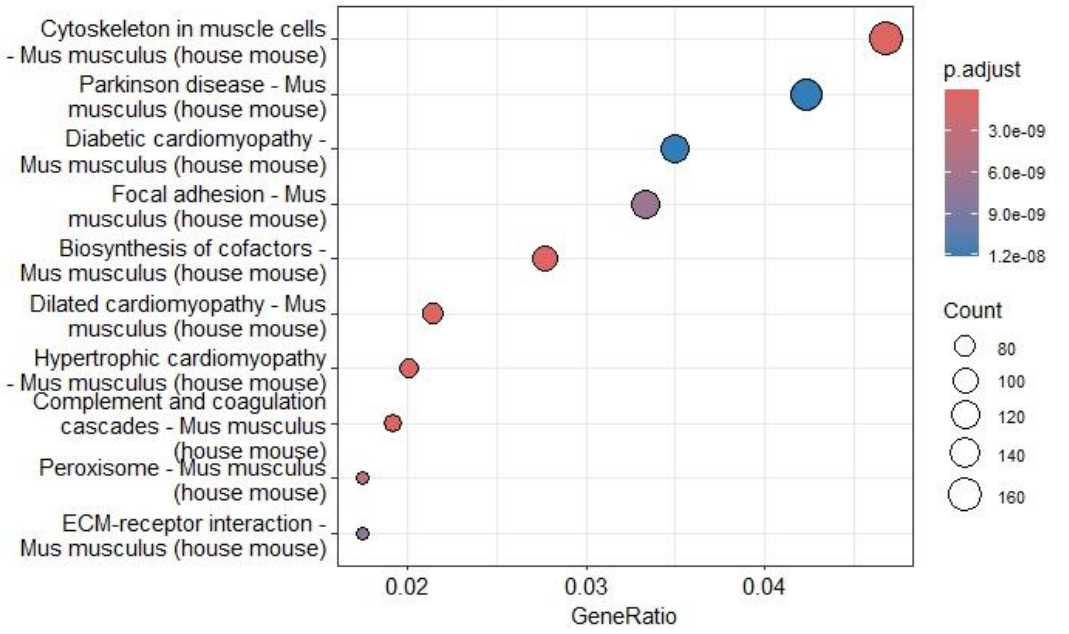

Figure 14: KEGG analysis of DEGs between the tissues

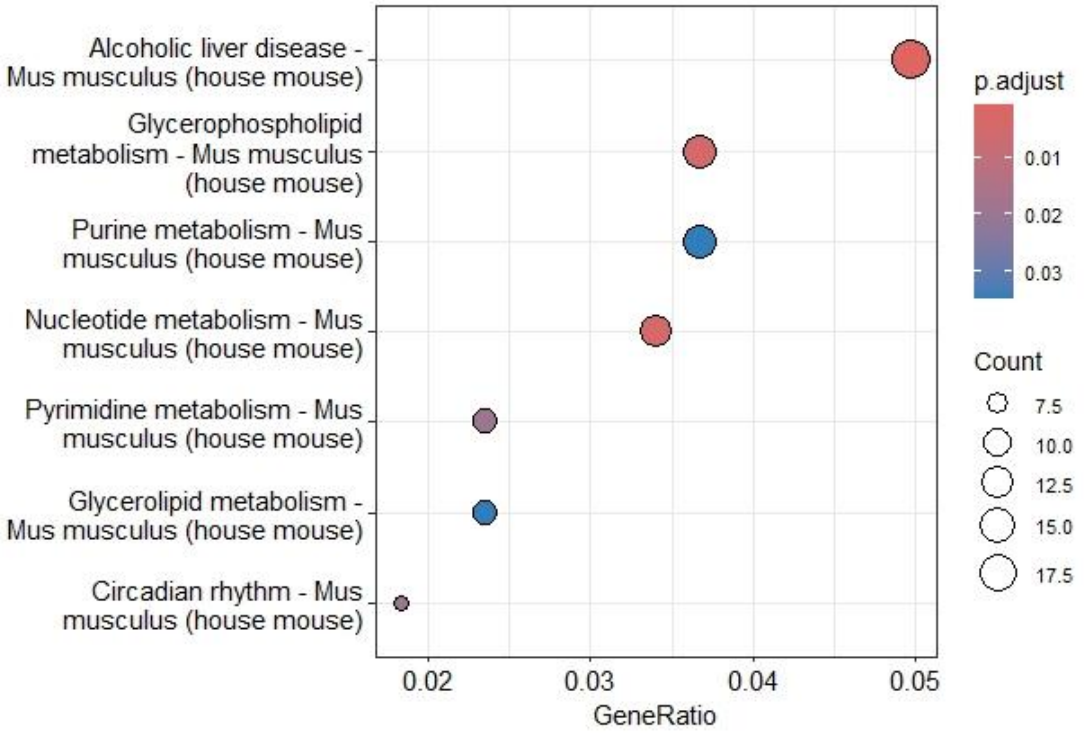

Figure 15: KEGG analysis of DEGs between the sampling time points

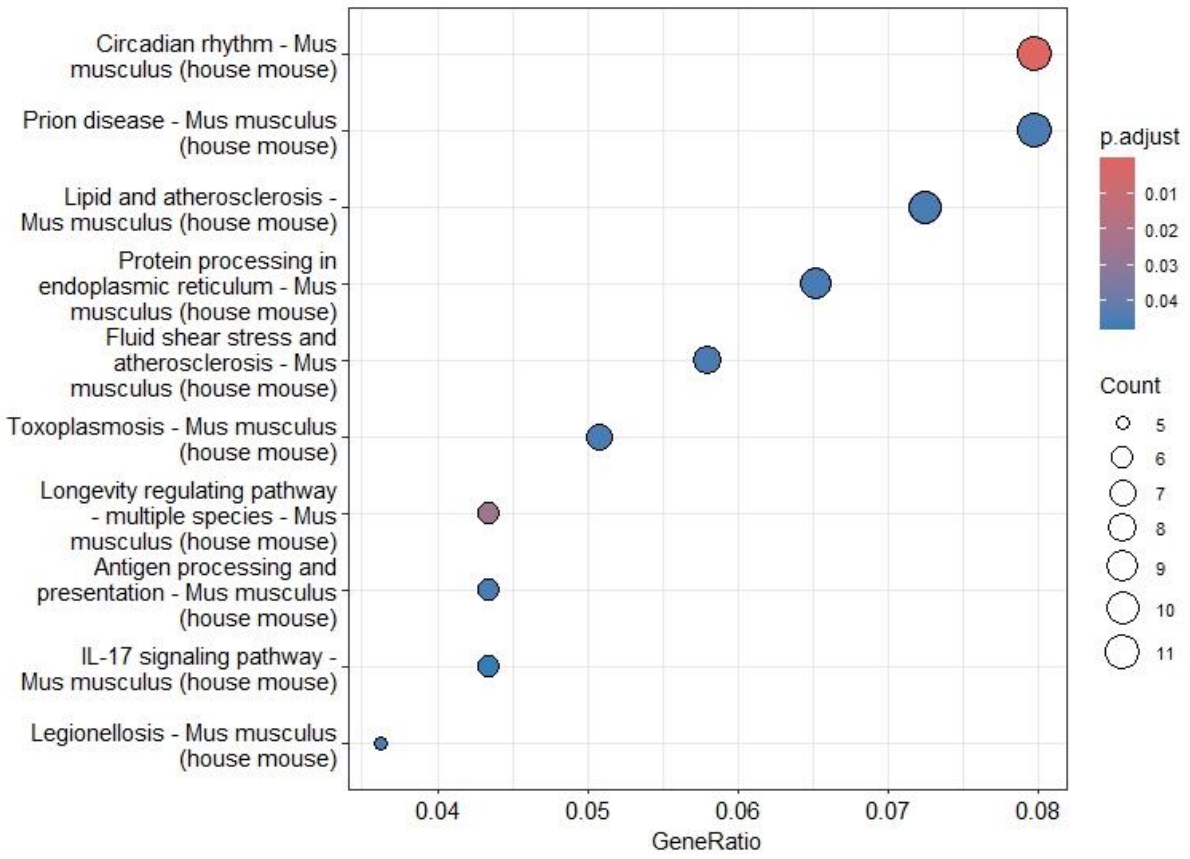

Figure 16: KEGG analysis of DEGs between the across the interaction of tissue and time contrasts

KEGG analysis represented by the top 20 significantly differentially expressed genes were analysed for each group of contrasts – Heart vs Liver, Sampling time (0 vs 12), and a contrast of the tissues over the sampling time. The following observations were made from the analysis:

1. Tissue: Liver vs Heart :-

The KEGG category involving the highest number of significant DEGs was Cytoskeleton in muscle cells, along with the lowest pvalue, indicating a high significance. In addition, Cardiomyopathy and Parkinsons disease were also highlighted as being enriched differentially between the mouse Liver and Heart.

2. Sampling time: 0 and 12:-

Over the sampling time, common metabolic pathways associated with the liver and Alcoholic liver disease were found enriched.

3. Tissue and time interaction:-

While observing the transcriptomic profile changes between the tissues over the sampling time, Functions such as the circadian rhythm, protein processing in endoplasmic reticulum, Diseases like atherosclerosis and prion disease, and immunological pathways like the IL-17 signalling pathway were enriched.