Project 2: Analyst(Test Data)

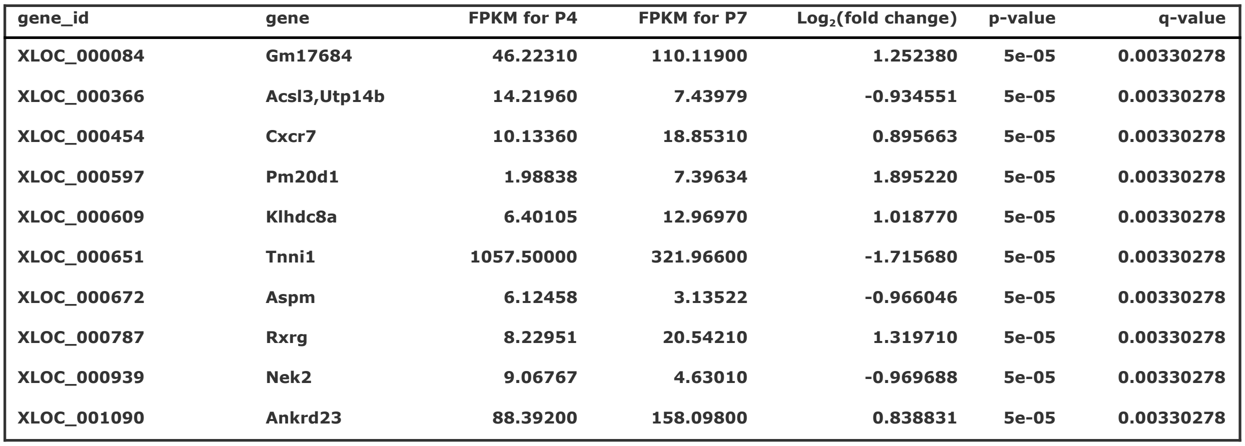
Introduction:

This section of the project was tasked to take the gene expression data produced by the programmer and analyze the differential expression data that was produced by the samples. The first step is identifying what genes were produced which is done by comparing the FPKM values produced by the different groups. Then using DAVID Functional Annotation Clustering groups we can observe the overall biological process related to the data produced by the overall experiment. From this data we can make predictions of how Myocytes differentiate themselves between P4 and P7 groups.

Methods:

After downloading the gene_exp.diff produced from the earlier step. We first obtained the top 10 differentially expressed genes to have a quick summary of the results of the Differential expression analysis, and then graphed all significant $\log_2$(fold change) on a histogram using ggplot2. For the gene enrichment analysis, all significantly differentially expressed genes were split into up-regulated and down-regulated sets to have a greater understanding of specific changes of the GO terms that we observe. These sets were uploaded to DAVID Functional Annotation Clustering with the analysis based on *Mus Musculus*.

Results:



gene_id	gene	FPKM for P4	FPKM for P7	Log$_2$(fold change)	p-value	q-value
XLOC_000084	Gm17684	46.22310	110.11900	1.252380	5e-05	0.00330278
XLOC_000366	Acsl3,Utp14b	14.21960	7.43979	-0.934551	5e-05	0.00330278
XLOC_000454	Cxcr7	10.13360	18.85310	0.895663	5e-05	0.00330278
XLOC_000597	Pm20d1	1.98838	7.39634	1.895220	5e-05	0.00330278
XLOC_000609	Klhdc8a	6.40105	12.96970	1.018770	5e-05	0.00330278
XLOC_000651	Tnni1	1057.50000	321.96600	-1.715680	5e-05	0.00330278
XLOC_000672	Aspm	6.12458	3.13522	-0.966046	5e-05	0.00330278
XLOC_000787	Rxrg	8.22951	20.54210	1.319710	5e-05	0.00330278
XLOC_000939	Nek2	9.06767	4.63010	-0.969688	5e-05	0.00330278
XLOC_001090	Ankrd23	88.39200	158.09800	0.838831	5e-05	0.00330278

Table 1 Top 10 most significant expressed genes. We see the top 10 most differentially expressed genes in the group with low q-values and both up and down regulated genes.

The gene expression analysis appears to have proper results as the q-values show significance in the differentiation and some of the top genes are associated with processes that are relevant. *Acsl3* is the 2nd top hit, and is associated with cell differentiation and establishment of localization. Additionally, *Ankrd23* is associated with cytoskeletal protein binding which is expected for a growing heart.

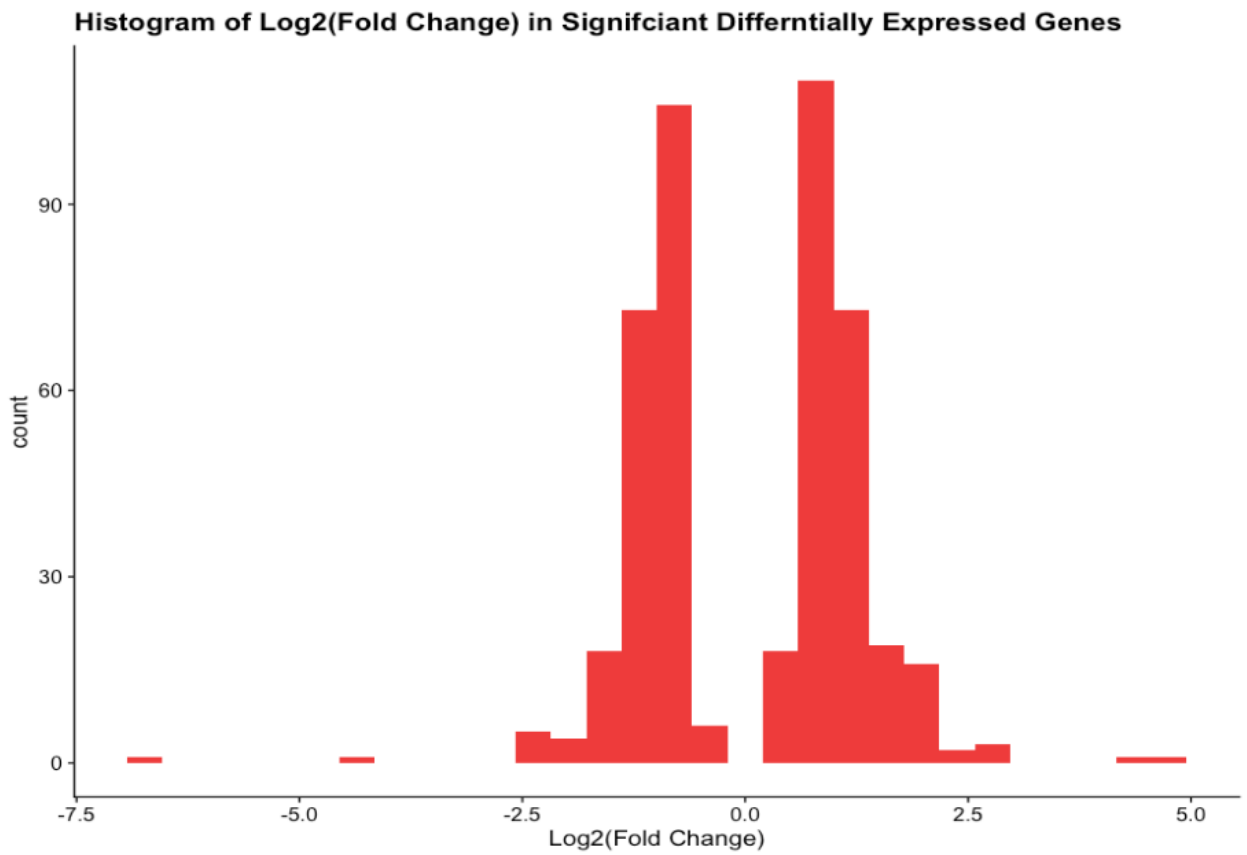


Figure 1 Histogram of Log$_2$(Fold Change) in Significantly Differentially Expressed Genes The histogram depicts 459 significantly differentially expressed genes. Log$_2$(Fold Change) greater than 0 represents genes more expressed in P4 and less than 0 represents the genes that are less expressed in P4 compared to P7.

The significantly differentially expressed genes(DEG) were graphed on a histogram. We can see that there is a comparable amount of up-regulated(243 genes) and down-regulated(216 genes) in the graph, all clustered above and below log$_2$(0). While most DEGs appear to be clustered in the center there appears to be a few genes that have high Log$_2$(Fold Changes).

GO Term	Enrichment Score	Benjamini
response to oxygen-containing compound	6.64	5.4E-7
extracellular space	6.14	1.5E-5
response to oxygen-containing compound	6.03	5.4E-7
response to organic substance	5.46	2.6E-5
response to endogenous stimulus	3.7	6.7E-3

Table 2 Gene Enrichment Analysis of Up-Regulated Genes

GO Term	Enrichment Score	Benjamini
microtubule cytoskeleton	4.96	9.8E-8
mitotic cell cycle	4.87	2.5E-19
carbohydrate derivative binding	4.49	2.4E-5
chromosome organization	3.84	2.0E-6
mitotic spindle organization	3.2	1.7E-5

Table 3 Gene Enrichment Analysis of Down-Regulated Genes

The up and down regulated genes produced a similar amount of clusters that described GO terms related to cell growths. The up-regulated genes appeared to describe clusters that were responding to possible changes as many terms were describing responses to different chemical changes or external stimuli (Table 2). The down-regulated genes clustering described cellular processes for cell proliferation, as genes related to microtubule cytoskeleton and mitotic cell cycle were less expressed in P4 than in P7.

Discussion:

The goal of this section was to capture any changes in gene expression that can possibly explain the *Mus Musculus* to repair heart cells during infancy. Observing the gene enrichment analysis, we saw that the up-regulation mainly related to response to different chemicals and stimuli which do not appear to be found in the the main paper. This relationship may be capturing the gene expression naturally associated with the development of the heart. The down-regulated clustering showed a more definitive relationship as the GO term related more on the cell Cycle, DNA repair which was found in the original paper. While we were able to replicate a section of the results, we were unable to capture the changes in heart cell mechanism within the snapshot of P4 to P7.

Project 2: Biologist(Test Data)

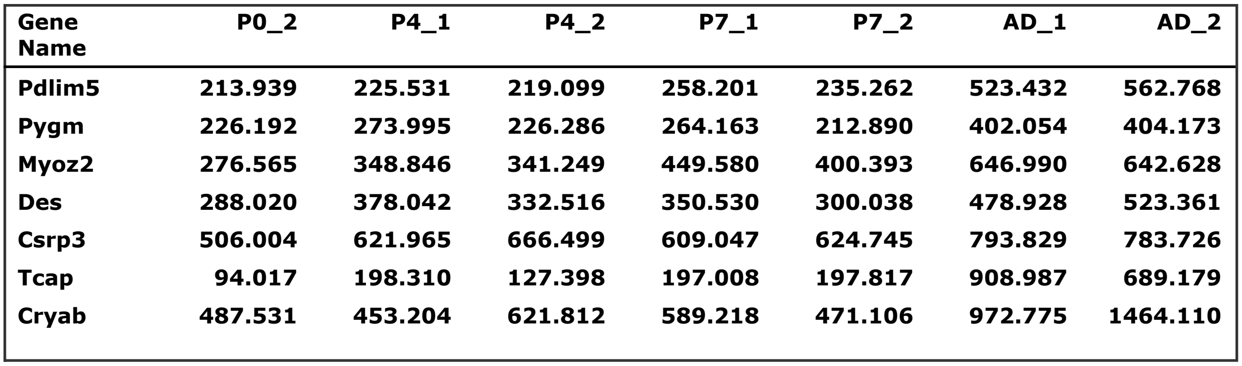
Introduction:

This section of the analysis supplements the previous results by attempting to capture the same DEG but instead of DEG and gene enrichment, used FPKM tables and displaying the FPKM of the most expressed gene. The process involves obtaining the FPKM data for all of the tables and comparing the trend as the heart develops. Then we create a heatmap of the data that includes all the genes.

Methods:

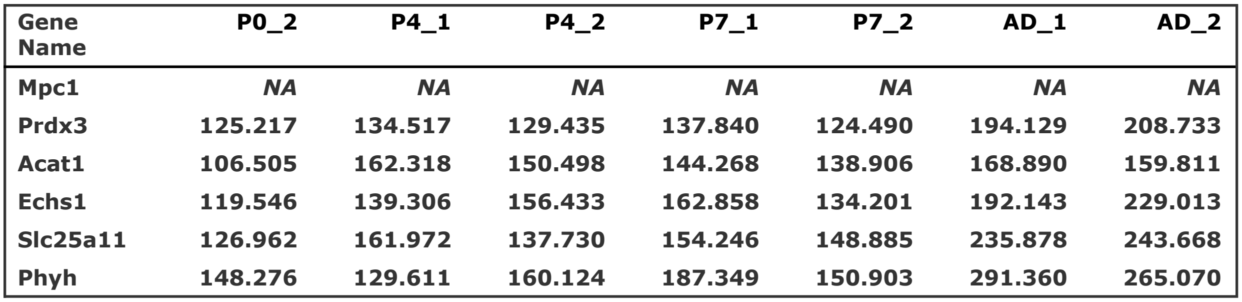
After the steps from the analyst, we obtain all the fpkm_tracking files from every single sample file available. From this point we filtered to only the genes of interest as seen in the original paper and we used ggplot to graph the changes of FPKM in the different genes as the heart developed. We additionally added FPKMs from the different 7 samples and combined them to graph using the built in heatmap function.

Results:



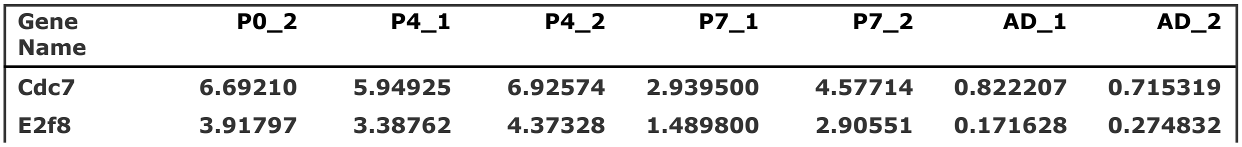
Gene Name	P0_2	P4_1	P4_2	P7_1	P7_2	AD_1	AD_2
Pdlim5	213.939	225.531	219.099	258.201	235.262	523.432	562.768
Pygm	226.192	273.995	226.286	264.163	212.890	402.054	404.173
Myoz2	276.565	348.846	341.249	449.580	400.393	646.990	642.628
Des	288.020	378.042	332.516	350.530	300.038	478.928	523.361
Csrp3	506.004	621.965	666.499	609.047	624.745	793.829	783.726
Tcap	94.017	198.310	127.398	197.008	197.817	908.987	689.179
Cryab	487.531	453.204	621.812	589.218	471.106	972.775	1464.110

Table 1 FPKM values for genes from each sample for Sarcomere



Gene Name	P0_2	P4_1	P4_2	P7_1	P7_2	AD_1	AD_2
Mpc1	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
Prdx3	125.217	134.517	129.435	137.840	124.490	194.129	208.733
Acat1	106.505	162.318	150.498	144.268	138.906	168.890	159.811
Echs1	119.546	139.306	156.433	162.858	134.201	192.143	229.013
Slc25a11	126.962	161.972	137.730	154.246	148.885	235.878	243.668
Phyh	148.276	129.611	160.124	187.349	150.903	291.360	265.070

Table 2 FPKM values for genes from each sample for Mitochondria



Gene Name	P0_2	P4_1	P4_2	P7_1	P7_2	AD_1	AD_2
Cdc7	6.69210	5.94925	6.92574	2.939500	4.57714	0.822207	0.715319
E2f8	3.91797	3.38762	4.37328	1.489800	2.90551	0.171628	0.274832

Cdk7	12.31520	9.53774	10.83330	9.221130	8.73875	6.946230	5.692480
Cdc26	17.86200	16.80100	18.78760	13.60640	15.59540	11.67280	11.82180
Cdc6	2.38139	2.16122	3.85812	0.971499	1.48514	0.230830	0.036513
E2f1	6.54143	7.38801	8.42559	3.742090	5.58000	0.319143	0.425882
Cdc27	18.68470	16.61770	18.46850	13.16270	15.59470	7.967580	6.448150
Bora	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
Cdc45	4.85574	5.01777	6.55647	2.660450	4.20611	1.081390	1.243620
Rad51	3.62869	4.03168	5.80323	2.939520	2.49621	0.217335	0.466008
Aurkb	7.77629	8.49992	12.99720	5.671570	6.92785	0.232604	0.288508
Cdc23	27.14160	24.49960	25.60690	17.64970	20.99760	11.69960	10.54000

Table 3 FPKM values for genes from each sample for Cell Cycle

Table 1,2,3 are the looking expressed genes focused on the paper associated with specific functions. GO term associated with Sarcomere we see a general increase in the expression level of the different genes. The greatest increase seen in this gene set was *Cryab*. The GO term of Mitochondria, we see a general up trend similar to Sarcomere with one of the major changes seen in *Phyh*. Unlike the other tables, the genes associated with Cell Cycle has a general downtrend seen in its value and additionally has an overall lower FPKM level compared to other 2 tables.

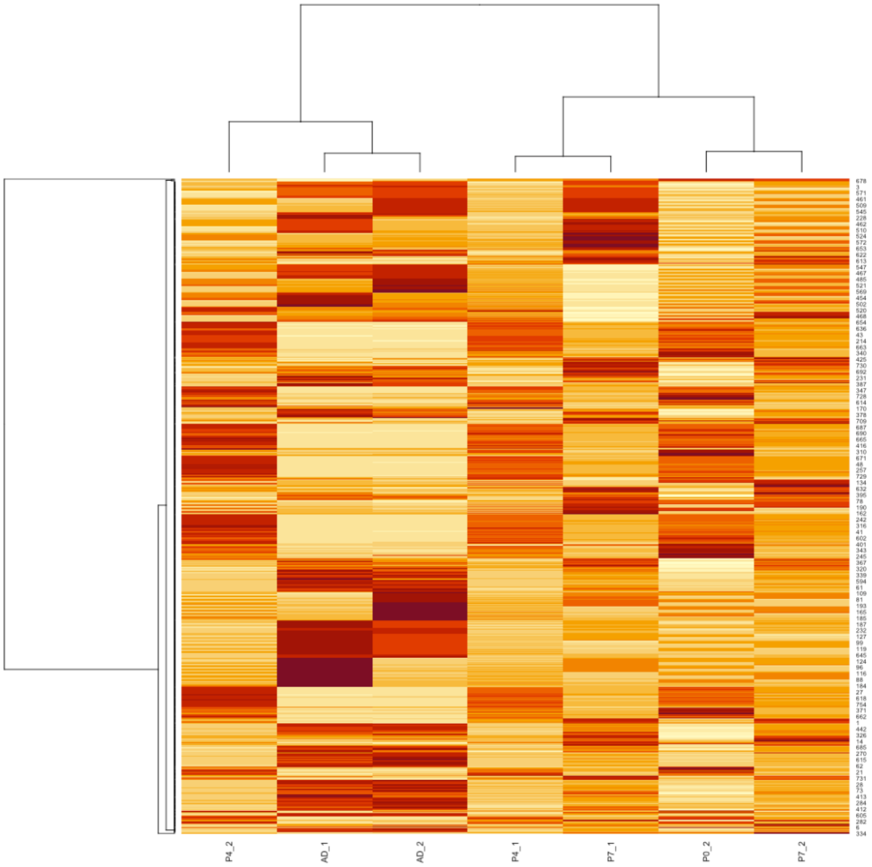


Figure 1 Heat map of the top 1000 differentially expressed gene from P4 vs P7

Using the top 1000 DEG found in P4 vs P7, the heatmap was unable to differentiate between the different groups. P4_2 was clustered with the adults while P7 were clustered in different groups. With the top 1000 DEGs found in P4 vs P7 the gene expression does not seem to have any overall differences in its gene expression.

Discussion:

The main goal of this analysis was to capture the changes of gene expression that allows for the recovery of heart cells in young mice. With the data we were unable to capture any significant differences between the different groups. We were able to replicate the results seen in the FPKM tables, where there are overall uptrends within Sarcomere and Mitochondria related genes and overall downtrend within Cell Cycle related genes. This can possibly be explained by the shift from developing the heart to focusing on its function. One of these mechanisms can possibly explain the loss of heart repair mechanisms. While the FPKM table effectively captured the changes of gene expression throughout the development process, the heatmap of the top 1000 DGE from P4 vs P7 was unable to capture the same relationship. The clustering was unable to properly separate the different samples in its proper groups and the heatmap does not capture any major trends. This may be due to the differences in P4 and P7 don’t capture the overall changes and only captures more nuanced differentiations for cell development and not differentiation overall.