CONCORDANCE OF MICROARRAY AND RNA-Seq DIFFERENTIAL GENE EXPRESSION

From Wang et al. (2014)

BF528 final project

Introduction

Microarray and high-throughput RNA-sequencing are both common measurements of RNA sequences. Both technologies aim to attain the same biological etiology, although the technologies are fundamentally different. Due to the difference, how coordinating the two technologies are essential.

The original paper compared the two methods and tried to determine their concordance. the study design investigated concordance between the aforementioned technologies using 27 unique toxicology treatments from a range of known modes of action (MOA)

In our group tasks before, I played the biologist role and analyzed the result from my teammates. This time, I reproduced the result of the data curator and programmer to deeply understand the RNA sequence analysis methods. In the data curator part, I should do quality control and STAR alignment on the original SRA data. In the next programmer part, I will use featureCounts to annotate the RNA sequence result and generate the count matrix. After that, the differential expression is performed on the count matrix using DESeq2. At last, I will draw out the distribution plot through R ggplot.

I choose this project because STAR alignment and DESeq2 are the most wide-used bioinformatics methods. I wished I could be more familiar with the two methods. Similarly, STAR alignment for the original paper is essential since it provides the most efficient and fast way to align short RNA reads. DESeq2 allows us to have a reliable differential expression result.

Method

The data was obtained for Affymetrix microarray and Illumina RNA-sequencing datasets, downloaded from accessions SRP039021, GSE55347, and GSE47875. The group chosen is based on the course instruction. To avoid using too much disk space, I directly used the tox group 5 data downloaded by my team before. The specific information is in the table below.

Sample	Mode of Action (MOA)	Treatment	Treatment Vehicle	Treatment Delivery
SRR1177978	DNA_Damage	Aflatoxin B1	CMC 0.5%	Oral Gavage
SRR1177979	DNA_Damage	Aflatoxin B1	CMC 0.5%	Oral Gavage
SRR1177980	DNA_Damage	Aflatoxin B1	CMC 0.5%	Oral Gavage
SRR1178015	CAR/PXR	Miconazole	Corn Oil 100%	Oral Gavage
SRR1178022	CAR/PXR	Miconazole	Corn Oil 100%	Oral Gavage
SRR1178048	CAR/PXR	Miconazole	Corn Oil 100%	Oral Gavage
SRR1177963	PPARA	Pirinixic Acid	CMC 0.5%	Oral Gavage
SRR1177964	PPARA	Pirinixic Acid	CMC 0.5%	Oral Gavage
SRR1177965	PPARA	Pirinixic Acid	CMC 0.5%	Oral Gavage

SRR1178067	Control	None	CMC 0.5%	Oral Gavage
SRR1178068	Control	None	CMC 0.5%	Oral Gavage
SRR1178069	Control	None	CMC 0.5%	Oral Gavage
SRR1178035	Control	None	Corn Oil 100%	Oral Gavage
SRR1178045	Control	None	Corn Oil 100%	Oral Gavage
SRR1178050	Control	None	Corn Oil 100%	Oral Gavage

Table 1. Mice RNA-seq Treatment Samples Utilized in Concordance Analysis (Toxgroup 5)

Then, the quality control was performed using fastqc. STAR is used to do alignment on the samples with the default setting. Samples on average contained 48.5% GC content with no outliers. Samples averaged 83.91% unique alignment with unalignment primarily due to too short of reads (11.02% on average).

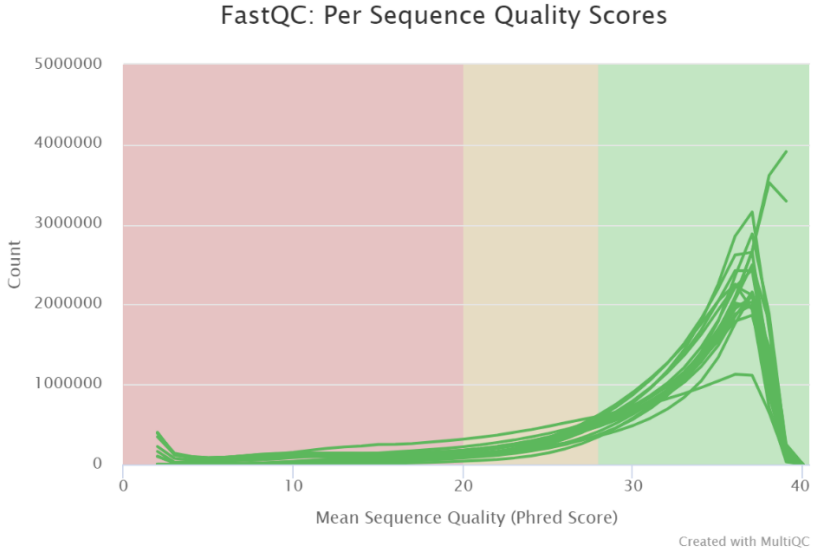

Figure 1: Average Phred quality scores for each sequence read per sample

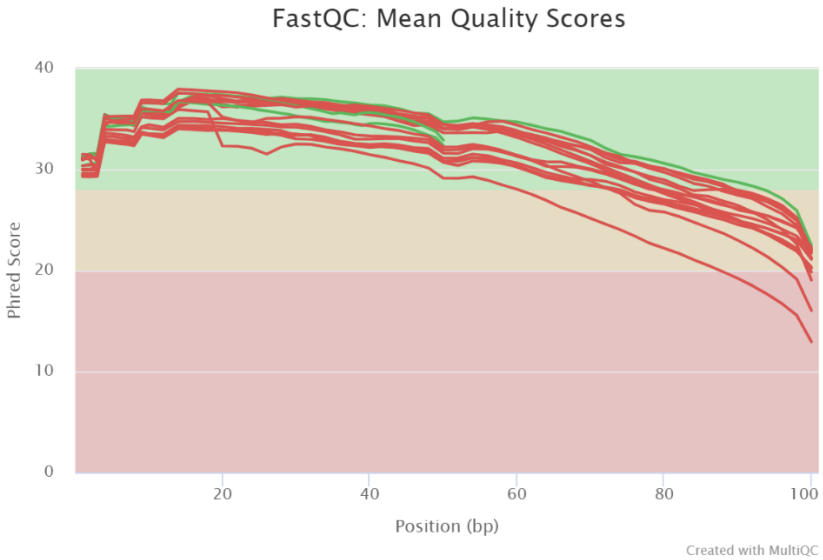

Figure 2: The mean quality scores of samples.

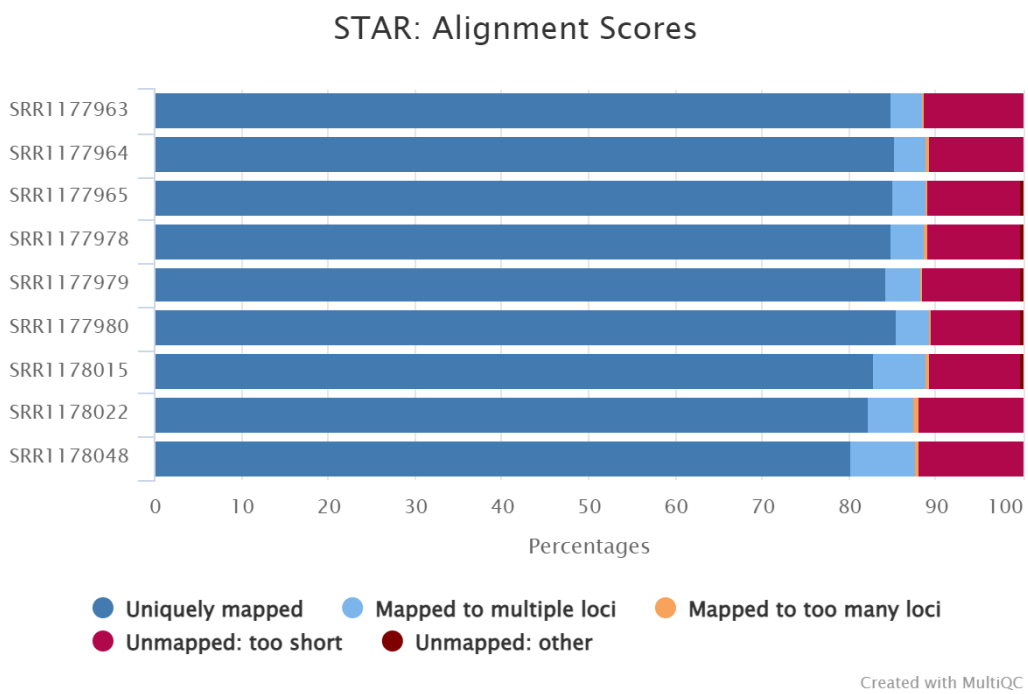

Figure 3, The result of fastqc on STAR.

In the programming section, the program featureCounts was used to generate count matrix according to the course instruction. And Multiqc was run to combine results from all nine samples. DESeq2, a software using negative binomial regression, is run to estimate count differences given a statistical design. I installed the the DESeq2 package from Bioconductor. According to its manual, I created the DESeq Object. After performing DESeq on the object, results were extracted using DESeq and results functions, with function lfcShrink used to add shrunken log2 fold change results to the tables. The threshold of significance was set as padj<0.05 and the differential expression result was saved into a csv file.

Result

Most results of data part were shown in the method part.

The percentage of featureCounts assigned reads ranged between 54% (sample SRR1178048) and 61.5% (sample SRR1177979) (Figure 4). Figure 5 was the distribution of raw counts of the nine samples and there was no significance between the three groups.

In the DESeq2 analysis part, there were 10,839 differentially expressed genes in total, with 6,513 differentially expressed genes between the CAR/PXR mode of action samples and controls, 1,242 between DNA_Damage samples and controls, and 3,152 genes differentially expressed between PPARA samples and controls. The distribution of all those genes in each group was shown in the figure 6. The table of top10 genes in DESeq2 result were shown in the table 2.

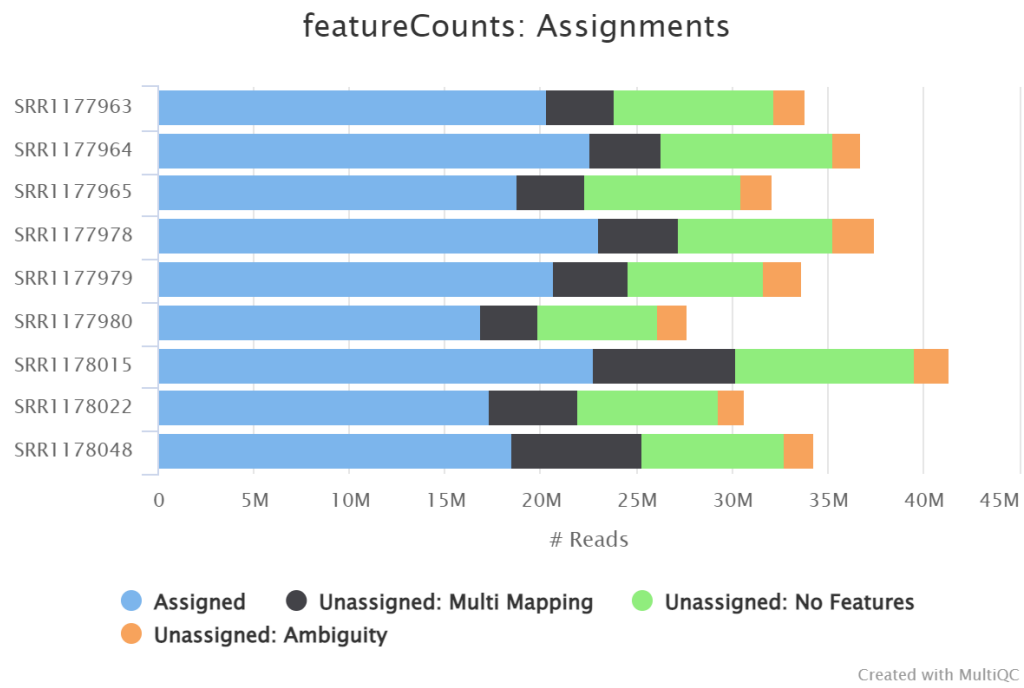

Figure 4, the assigned reads using featureCounts.

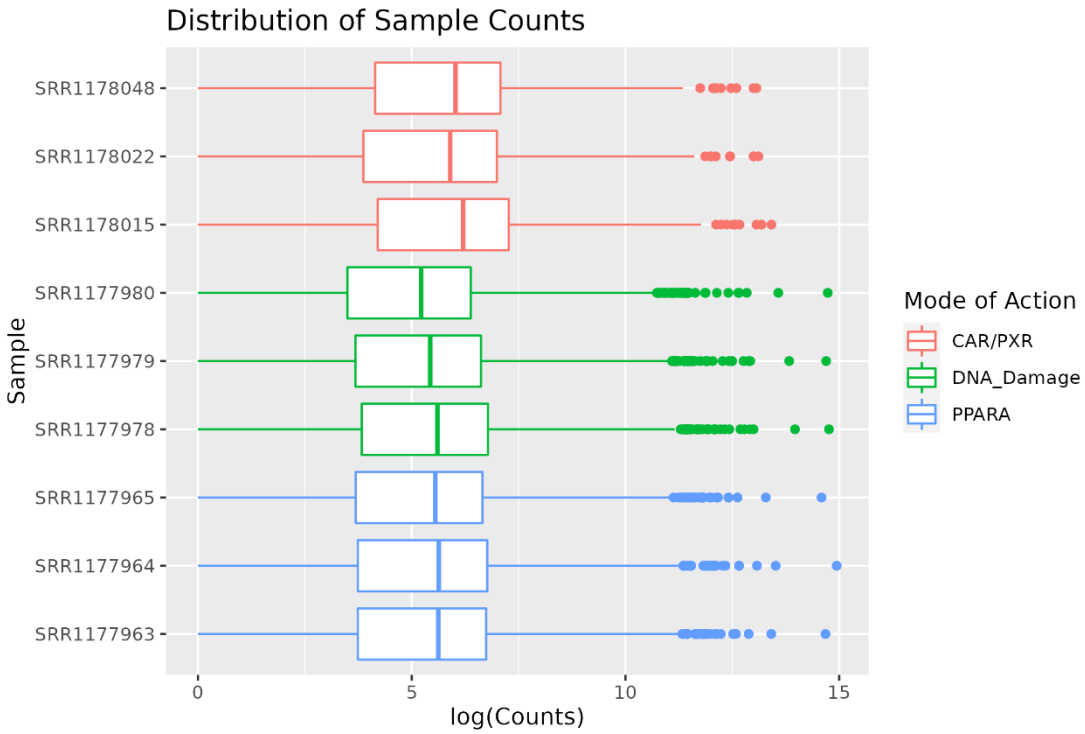

Figure 5, the distribution of raw counts per sameple.

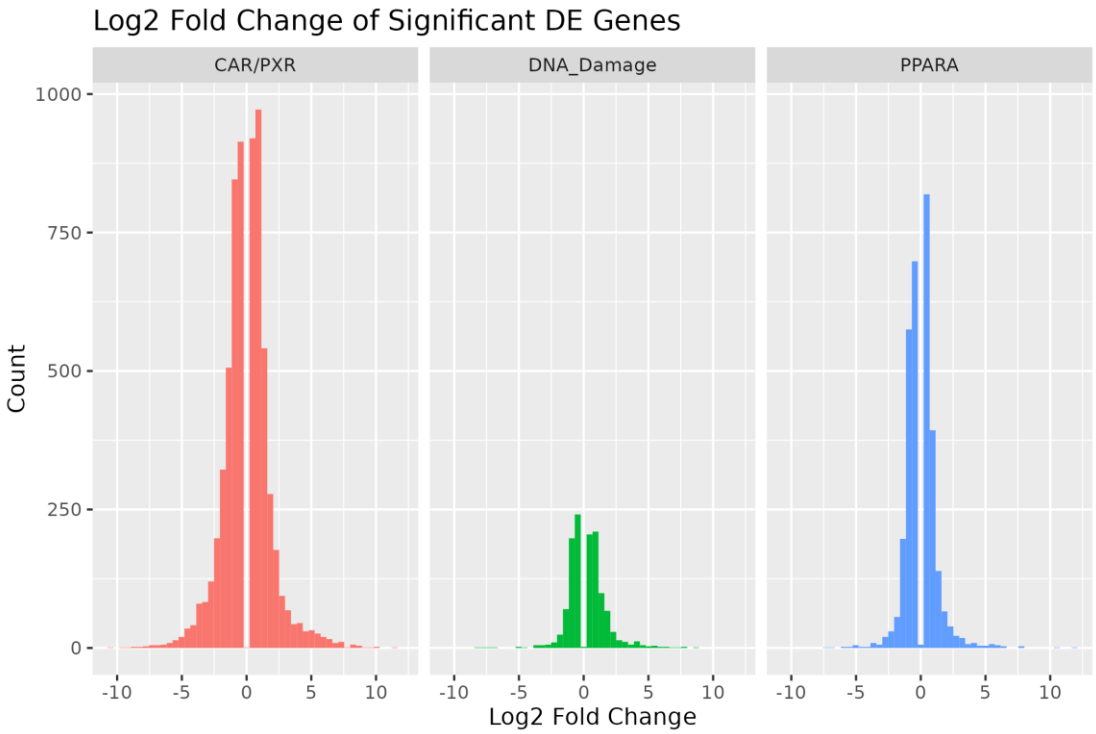

Figure 6 The distribution of differential gene expression in each group.

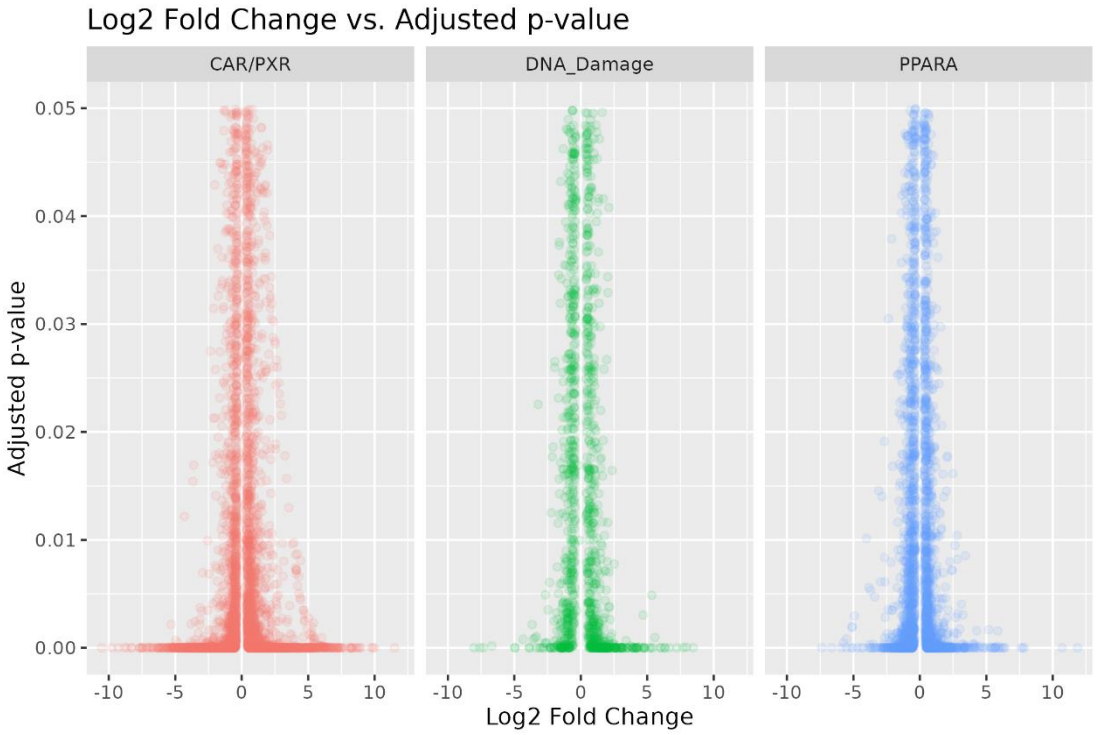

Figure 7 The fold Change vs padj. In DESeq2 result.

Geneid	baseMean	log2FoldChange	lfcSE	pvalue	padj
5540	10375.71	-5.91726	0.131065	0	0
6431	7678.699	-4.68724	0.119987	0	0
11980	6896.618	-4.05091	0.10384	0	0
573	9511.8	-6.12876	0.165085	8.21E-293	2.59E-293
6850	7747.961	-9.77127	0.271809	2.69E-282	6.80E-279
8818	2491.354	-7.43348	0.2101	4.12E-275	8.67E-272
7296	6236.852	-5.62273	0.165432	3.11E-254	5.61E-251
2050	27236.57	-8.65153	0.254876	8.65E-254	1.37E-250
5296	61494.95	-7.69091	0.227412	9.10E-251	1.28E-247
7682	3068.32	7.357956	0.217805	1.93E-248	2.44E-245

Table 2-1, MIC (CAR/PXR) top 10 genes in DESeq2 Result

Geneid	baseMean	log2FoldChange	lfcSE	pvalue	padj
4927	3523.55	7.74653	0.192956	0	0
6920	3618.296	-6.63052	0.237001	3.55E-173	2.17E-169
4595	671.8191	6.315234	0.244219	2.59E-149	1.06E-145
9243	1328.539	-3.6925	0.175379	1.21E-99	3.71E-96
1784	564.0718	7.612765	0.369274	6.03E-94	1.48E-90
573	11146.44	-4.73382	0.236974	4.31E-90	8.80E-87
521	1388.714	6.098656	0.32032	3.28E-82	5.74E-79
5245	4579.771	-5.51228	0.320455	1.24E-67	1.89E-64
63	3461.768	-5.88062	0.342168	1.97E-67	2.68E-64
4658	5116.842	-4.41863	0.259032	1.75E-66	2.14E-63

Table 2-2, MIC (PPARA) top 10 genes in DESeq2 Result

Geneid	baseMean	log2FoldChange	lfcSE	pvalue	padj
2411	1230.256	3.953129	0.178318	4.19E-110	4.80E-106
4521	1398.649	5.269827	0.247782	1.60E-101	9.15E-98
4726	15185.38	4.827738	0.238332	3.15E-92	1.21E-88
3550	25424.59	7.243137	0.364739	9.45E-89	2.71E-85
10205	10099.38	4.751057	0.259873	5.37E-76	1.23E-72
8501	10705.84	4.028852	0.23953	8.10E-65	1.55E-61
4227	1221.682	3.89951	0.235194	5.03E-63	8.25E-60
5527	3537.187	2.636853	0.165839	3.86E-58	5.53E-55
2588	826.2949	-3.11928	0.20084	1.04E-55	1.32E-52
5157	793.5419	4.150566	0.273517	2.74E-53	3.14E-50

Table 2-3, MIC (CAR/PXR) top 10 genes in DESeq2 Result

Discussion

This research aims to reproduce parts of the Charles Wang et al study. I re-analysis the data downloaded by group members before and replicated the result of the data curator and programmer.

Outlier samples in the read counts generated from featureCounts and multiqc results were not found according to the quality control result. The treatment groups were made up of three modes of action namely, DNA-Damage (Aflatoxin), CAR/PXR (Miconazole), and PPARA (Pirinixic acid). A total of 10,839 differentially expressed genes are found, similar to our group result.

Some parts of my scripts might look similar to the codes submitted by our group because every team member participated in the scripts writing before. Meanwhile, those scripts work very well, and there is no reason to revise them too much.