Concordance of microarray and RNA-Seq differential gene expression

Data Curator: Priyanka Chary

Programmer: Kyler Anderson

Analyst: Alice Zheng

Biologist: Poorva Juneja

Introduction

While both RNA-Seq and microarray analysis are promising, there have been conflicting findings on the benefits and weaknesses of each one, especially when it comes to detecting differentially expressed genes. Some studies have claimed that RNA-seq have lower precision while others believe it is highly sensitive in gene detection [1]. Previous studies, however, lacked complexity and thoroughness, testing only a small set of conditions.

For this reason, Charles Wang et al. investigate RNA samples from rat livers with 27 conditions through a process of both microarray analysis and RNA-seq. In order to perform the analyses, limma, edgeR and DESeq robust multi-array average (RMA), and MAS5 were used. While limma was originally created for microarray analysis, it has been proven useful to analyze many different forms of data including RNA-seq, which is how it was utilized by Wang et al.. Limma uses linear models to make use of all the data and is particularly advantageous in analyzing data with many conditions in a relatively small sample size, which is one of the greatest benefits it would have provided Wang et al.'s study. Both DESeq and edgeR are more common RNA-seq analysis tools and both function to perform normalization in different ways. RMA is one of the most common tools used to normalize and summarize microarray data. The combination of RMA and MAS5 provides two forms of normalization and expression data from the microarray sample data.

Here, we aim to reproduce a portion of Wang et al.'s study using a similar approach.

Data

In Charles Wang et al.'s study, samples were gathered from DrugMatrix tissue/RNA bank. Male Sprague-Dawley rats were purchased and chemicals provided either orally or through an

injection. Chemicals were administered over the course of 3-7 days and livers were harvested soon after.

RNA samples were collected from the NTP DrugMatrix Frozen Tissue Library. The samples were categorized into training and test groups and of the 63 samples in the training group, 45 were proven to be rats treated with the chemicals. [1] For each test chemical there were three rats treated and for each mode of action there were three chemicals. RNA seq was then conducted using Illumina. For Microarray analysis, cRNA was prepared from the liver RNA and was hybridized to the Rat genome's Affymetrix whole genome GeneChip.

In order to replicate a part of Wang et al.'s study, tox group 1 was chosen and treatment samples' (SRR1177987, SRR1177988, SRR1177989, SRR1177997, SRR1177999, SRR1178002, SRR11778020, SRR1178036, SRR1178046) datasets were collected. This included nine treatment samples, three replicates for three treatments. A combination of Fastqc, STAR, and multiqc were used to process the samples. Once fastqc was run from terminal, STAR was used to map the RNA-seq to the reference genome and provide BAM files. This was done as a paired end read with each read being 101 base pairs long. Figure 1 shows the distribution of the number of reads, with the average number of unique reads being about 8.5 - 9 million.

Using multiqc, data from fastq and bam files were combined for more efficient data analysis. All the results for the SRR1178036_2 sample were concerning as the Phred score for the sample was rather low, dropping into the 0-20 range (Figure 2a). Most importantly, the percentage of N counts in the fastq file was alarming as seen in Figure 2b. After repeating the fastqc and STAR portion for the sample, the results appeared to be the same, proving that the sample was of low quality

Arguably, one of the most important statistics produced by the multiqc report was the STAR alignment scores (Table 1) in which there is a representation of samples aligned to the genome, In Table 1, it is seen that the majority of each sample is uniquely mapped to the genome. With small percentages of each sample either mapping to multiple loci or being too short to map. Sample SRR1178036 has the largest number of reads that were too short to map which is

unsurprising since a large number of nucleotides from SRR1178036_2 were N counts leading to an ambiguous sequence and a resulting shorter usable read.

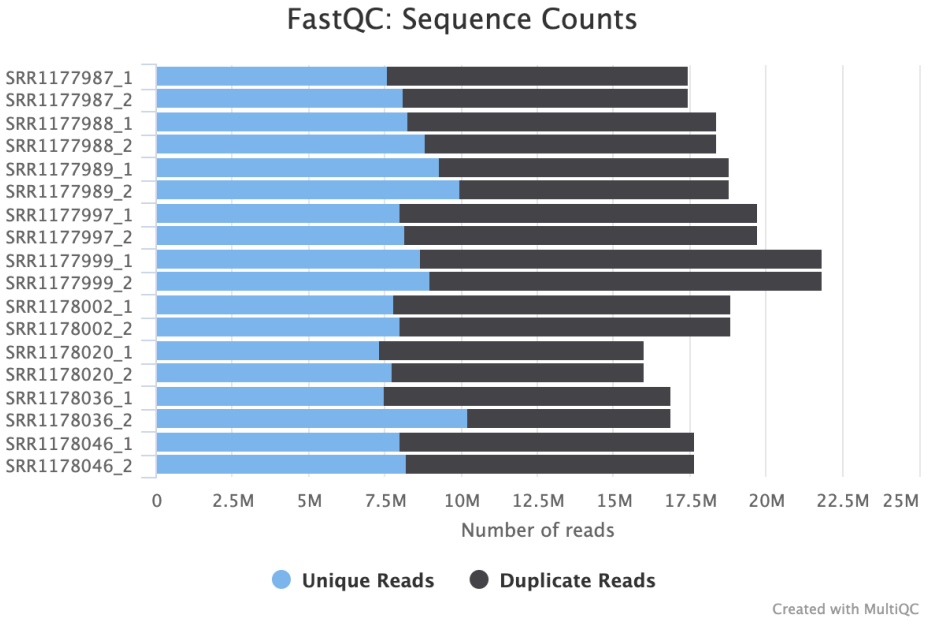

Figure 1. The bar graph above shows the number of unique and duplicate reads per sample.

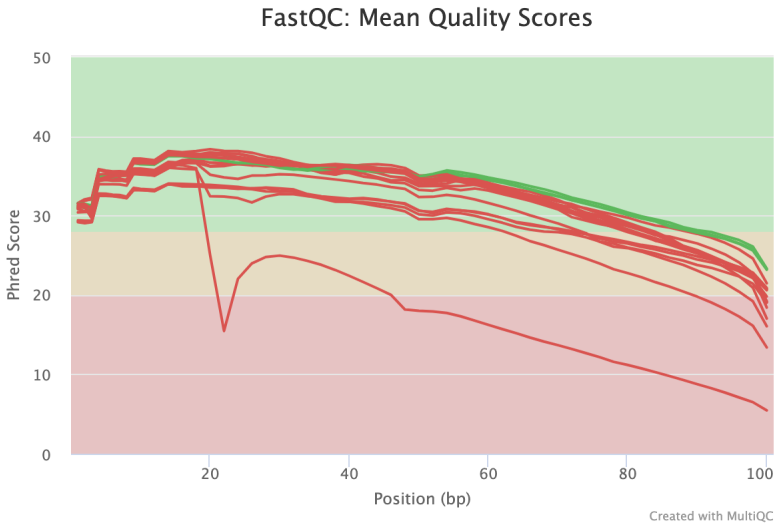

Figure 2a. The graph above shows the Phred score per sample. It can be seen that one sample (SRR1178036_2) has a low Phred score and can be considered low quality.

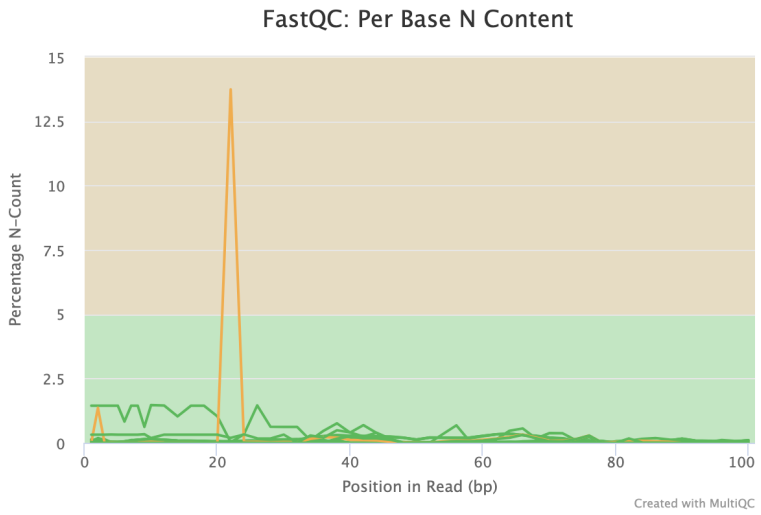

Figure 2b. The graph above shows Percentage of N Count over the read. Sample SRR1178036_2 has a peak in percentage N leading to ambiguity in its sequence and resulting shorter usable read.

Category	Uniquely mapped	Mapped to multiple loci	Mapped to too many loci	Unmapped: too short	Unmapped: other
SRR1177987	14785610 (84.8%)	634883 (3.6%)	26136 (0.1%)	1979793 (11.4%)	6971 (0.0%)
SRR1177988	15667286 (85.3%)	650689 (3.5%)	15207 (0.1%)	2028355 (11%)	7342 0.0%)
SRR1177989	15707078 (83.5%)	818451 (4.4%)	96880 (0.5%)	2173157 (11.6%)	9408 (0.1%)
SRR1177997	17608043 (89.2%)	767134 (3.9%)	35390 (0.2%)	1316471 (6.7%)	19737 (0.1%)
SRR1177999	19374545 (88.7%)	855549 (3.9%)	56592 (0.3%)	1525527 (7.0%)	26227 (0.1%)
SRR1178002	16796763 (89.1%)	737790 (3.9%)	41352 (0.2%)	1242646 (6.6%)	26399 (0.1%)
SRR1178020	13374032 (83.6%)	784376 (4.9%)	51179 (0.3%)	1777306 (11.1%)	16012 (0.1%)
SRR1178036	11443947 (67.8%)	676001 (4.0%)	31941 (0.2%)	4701893 (27.9%)	13492 (0.1%)
SRR1178046	15067177 (85.4%)	876408 (5.0%)	41360 (0.2%)	1653398 (9.4%)	14132 (0.1%)

Table 1. The table above shows the number of reads per sample that mapped uniquely, mapped to multiple loci, mapped to too many loci, or remained unmapped. Sample SRR1178036 had the largest section of unmapped reads which is not surprising based on the results in Figures 2a and 2b.

Methods

After alignment, gene counts were extracted from the bam file using the Subread package's featureCounts tool [6], supplied with a standard annotation of the rat genome. Each sample was provided 16 threads, and processing took no longer than 5 minutes for each. Multiqc was used again here to inspect the performance of counting across the 9 chemically treated samples. As shown in **Figure 3**, most of the reads were matched to a gene annotation; the proportion is consistent across samples. Around 20% of each aligned read set was not matched to a gene, and 10-15% mapped to multiple genes disqualifying them from the count. Less than 10% of aligned reads were lost to ambiguity.

The distribution of counts for each sample is shown in **Figure 4**, now including counts for pre-processed control samples. The counts are log scaled (base-10 here), since the distribution is otherwise extremely left skewed. Individual log-log histograms for each sample are provided in the supplementary **Figure S1**. The number of outliers for each sample is on the order of tens as the distribution falls off above counts of 45,000. Roughly 5,500 genes in the annotation were not matched in each sample; these zeros cannot be represented on the log scale. These counts are used as the measure of per sample gene expression. Supplementary **Table S1** gives the experiment design, including mode of action, chemical treatment, and vehicle; the latter indicates the appropriate control sample for a treatment group.

From these counts, the first tool for identifying DE genes was applied, DESeq2 [7]. The data was loaded into R and sliced into 3 treatment-control sets based on chemical treatment. Differential expression analysis results and normalized counts were exported. For 3-methylcholanthrene, 313 genes were DE at (Benajmini-Hochberg adjusted) $p < 0.05$. For clotrimazole there were 931 DE genes, and for chloroform there were 1810. The top 10 DE genes by adjusted p are given in **Table 2**, ties included. Log_2 fold change (L2FC) for these genes is visualized in **Figures 5** and **6**, as a histogram and plotted against nominal p-value, respectively. Both plots show the clustering of fold change at small magnitude. The histograms indicate slight rightward (up-regulation) skewness. These graphs are expected to be bimodal around zero, since L2FC near zero indicates the treatment expression levels and control levels are nearly equal; here it is harder to show significant DE. Corroborating this, the scatters show nominal p-value is more sporadic for L2FC near 0.

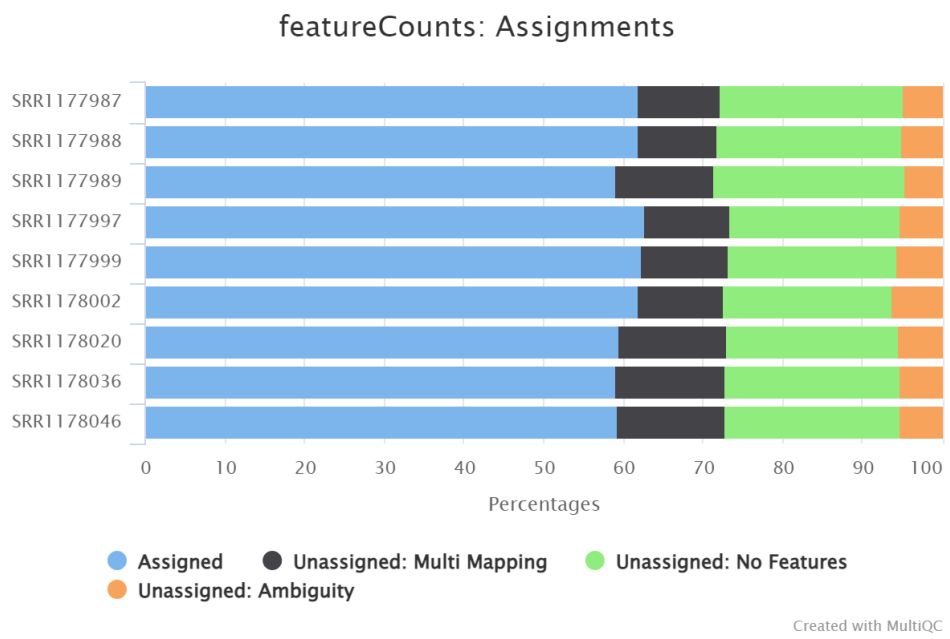

Figure 3. Summary of subread results for matching aligned reads to a rat genome annotation. Each sample was composed similarly, despite the poor quality of SRR1178036_2.

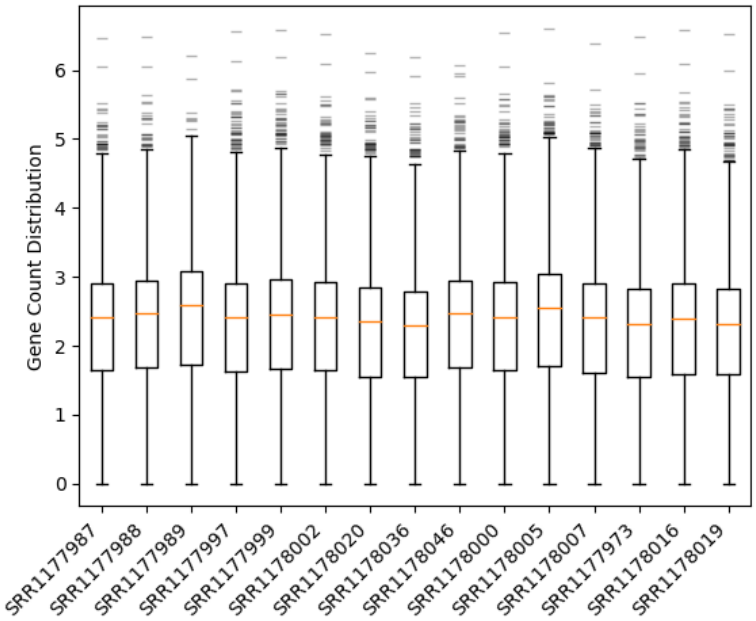

Figure 4. Boxplots of log-scale gene counts of each sample. Dashes above indicate outliers, orange indicates median. All samples of the same cell type, we expect to see horizontal consistency here; even significant expression effects would not dramatically shift such a log-scale distribution.

Chemical	Geneid	Log_2 FC	p_{adj}
3-methylcholanthrene	NM_012541	-3.671817	2.030124e-81
	NM_130407	-3.274867	2.985848e-19
	NM_022521	-1.772560	3.931974e-11
	NM_023094	-6.209278	4.142248e-11
	NR_046239	-1.434612	3.751066e-10
	NM_031972	-5.974377	1.042190e-09
	NM_134329	1.614242	1.964462e-09
	NM_175761	1.351148	1.984931e-09
	NM_053883	1.133717	1.020826e-08
	NM_012608	3.051650	1.638126e-08
	NM_022866	2.377431	1.638126e-08
	NM_024351	1.480423	1.638126e-08
clotrimazole	NM_001010921	-2.896350	3.427892e-121
	NM_080581	-4.558250	7.409958e-112
	NM_013105	-5.287681	3.085898e-106
	NM_131903	3.708974	7.545986e-69
	NM_173295	-2.982861	5.821773e-68
	NM_012844	-2.281175	1.522418e-67
	NM_017272	-4.654501	2.578046e-64
	NM_013215	-2.883042	9.389098e-60
	NM_001134844	-8.260472	4.729852e-52
	NM_133586	-5.095642	3.946980e-48
chloroform	NM_203512	-8.791042	2.766279e-132
	NM_001257095	-10.643050	1.137775e-99
	NM_080581	5.090019	7.507553e-47
	NM_023978	4.224402	5.364592e-40
	NM_013215	4.303240	2.276175e-37
	NM_012844	2.023735	4.953590e-37
	NM_139115	3.146194	1.171939e-33
	NM_012540	5.116621	9.359446e-30
	NM_001010921	2.112211	2.022918e-26
	NM_130741	3.042627	4.491139e-25

Table 2. List of top 10 differentially expressed genes by Banzamini-Hochberg adjusted p-value for each chemical treatment group. There are 12 in the first set for the 3-way tie for 10th.

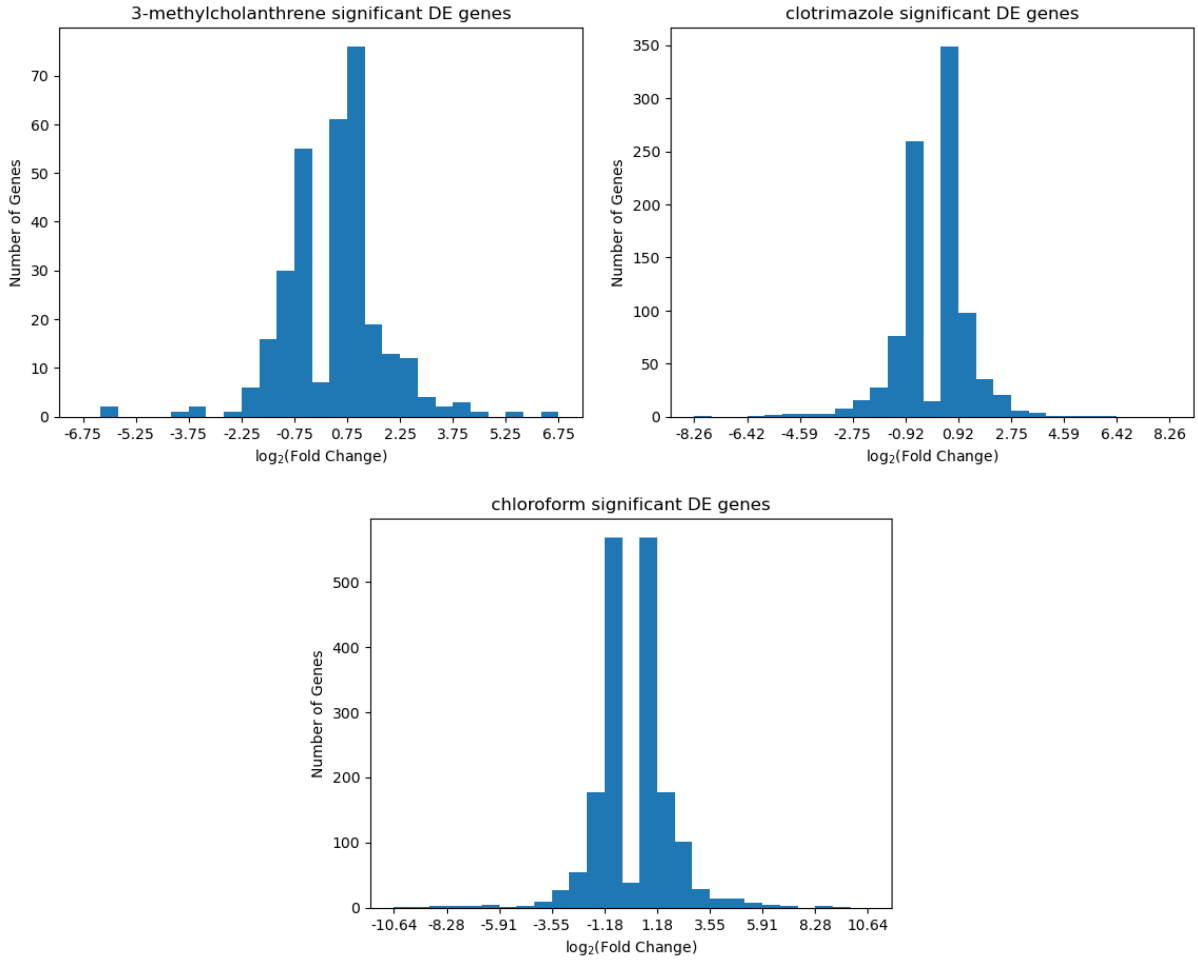

Figure 5. Histograms of gene counts by L2FC for each chemical treatment. Only significantly DE genes are counted.

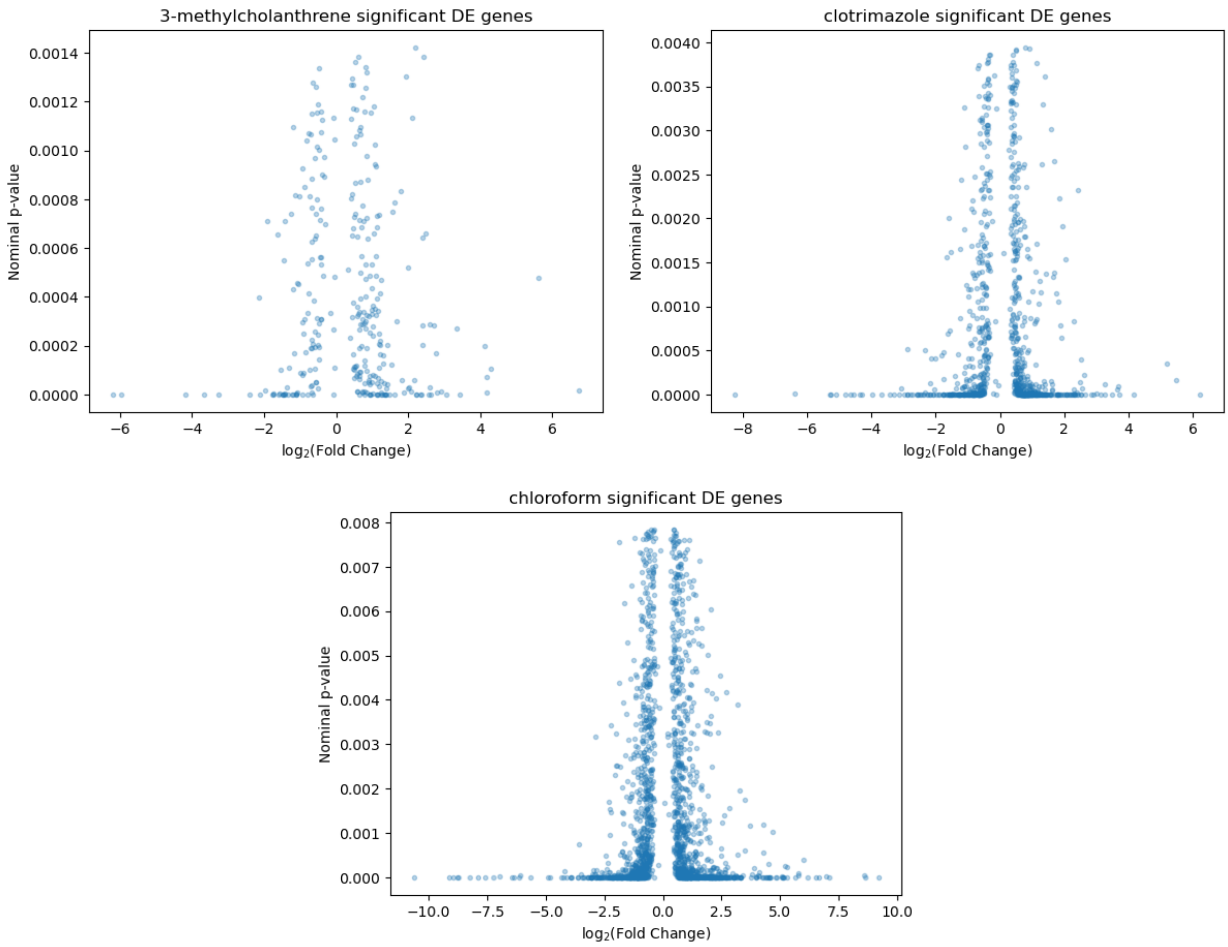

Figure 6. Scatter plots of Nominal p-value vs. L2FC for each chemical treatment. Only significantly DE genes are plotted.

Results

Limma was applied with a pre-normalized expression matrix provided by the paper to determine differential expression between the different treatments and control samples. Limma was performed three times to generate a microarray DE analysis for each of the toxins. The results were then filtered for differentially expressed genes using an adjusted p-value less than 0.05.

	Methylcholanthrene	Clotrimazole	Chloroform
1	58	5803	11407

Table 3. The total number of DE genes at p-adjust < 0.05 for each toxin in the microarray analysis.

	Methylcholanthrene		Clotrimazole		Chloroform
1	Cyp1a2	1	Cyp2b1	1	Abcc3
2	Cyp1a1	2	Cyp2b2	2	Gstp1
3	Ugt1a9	3	Cyp3a23/3a1	3	Akr1b8
4	Ugt1a8	4	Ugt2b1	4	Car3
5	Ugt1a7c	5	Ces2c	5	NA
6	Ugt1a6	6	Ugt1a9	6	Akr7a3
7	Ugt1a5	7	Ugt1a8	7	RGD1566134
8	Ugt1a3	8	Ugt1a7c	8	LOC259246
9	Ugt1a2	9	Ugt1a6	9	Dap
10	Ugt1a1	10	Ugt1a5	10	Abcb1b

Table 4. The gene short names for the top 10 differentially expressed genes in the microarray analysis for each toxin.

After filtering the samples to obtain the differentially expressed genes, the gene names of the samples were mapped onto the original data frame through joining a refSeq-to-probe id mapping table. Histograms and scatter plots were also created for each of the toxin DE gene results. The scatter plots show the relationship between the p-values and the fold change of the differentially expressed genes. The histograms show the distribution of DE gene counts at different fold change values for each of the toxins.

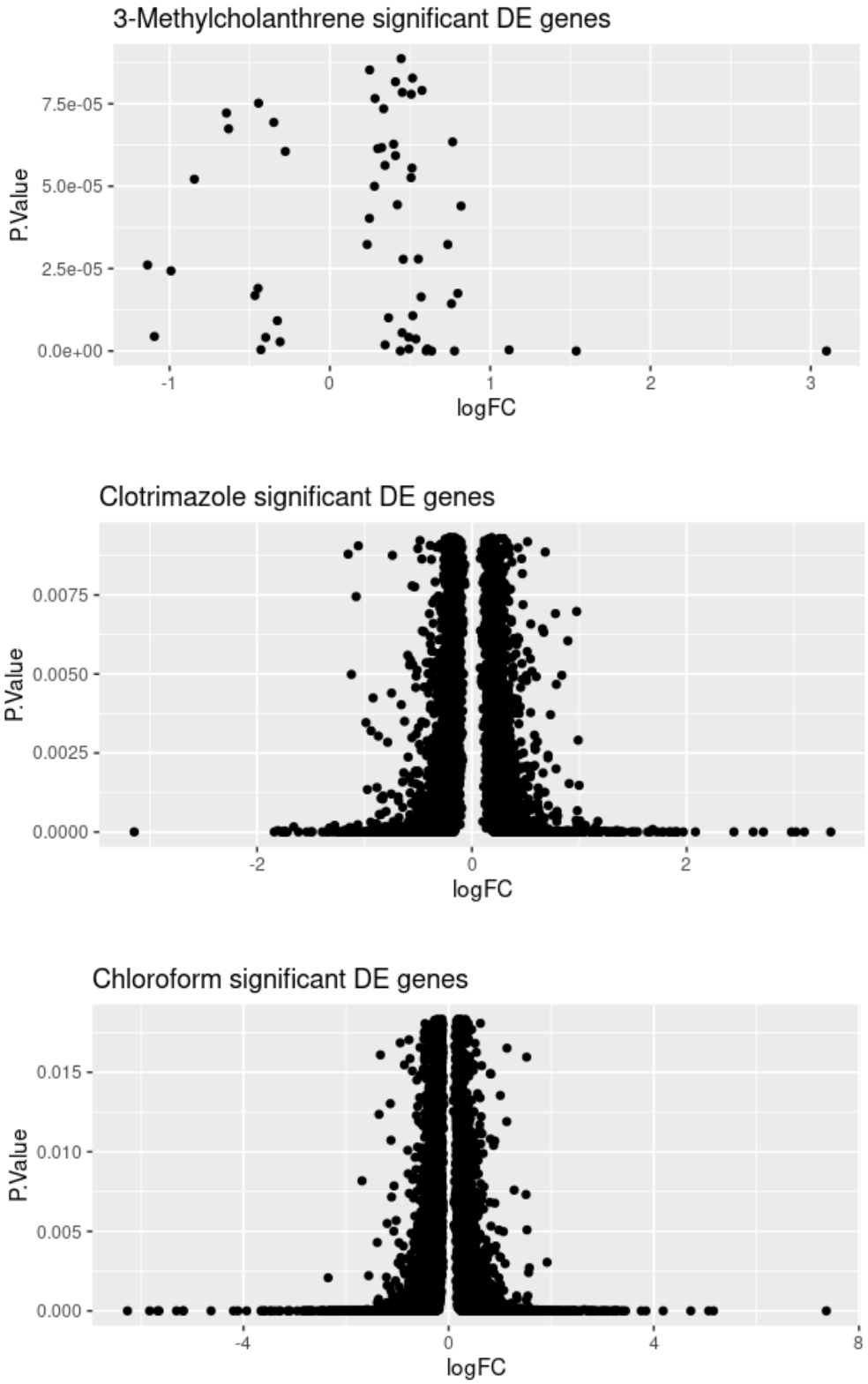

Figure 7. Scatter plots of the log fold change values versus the p-values for the differentially expressed genes in the 3-Methylcholanthrene, Clotrimazole, and Chloroform microarray analysis results.

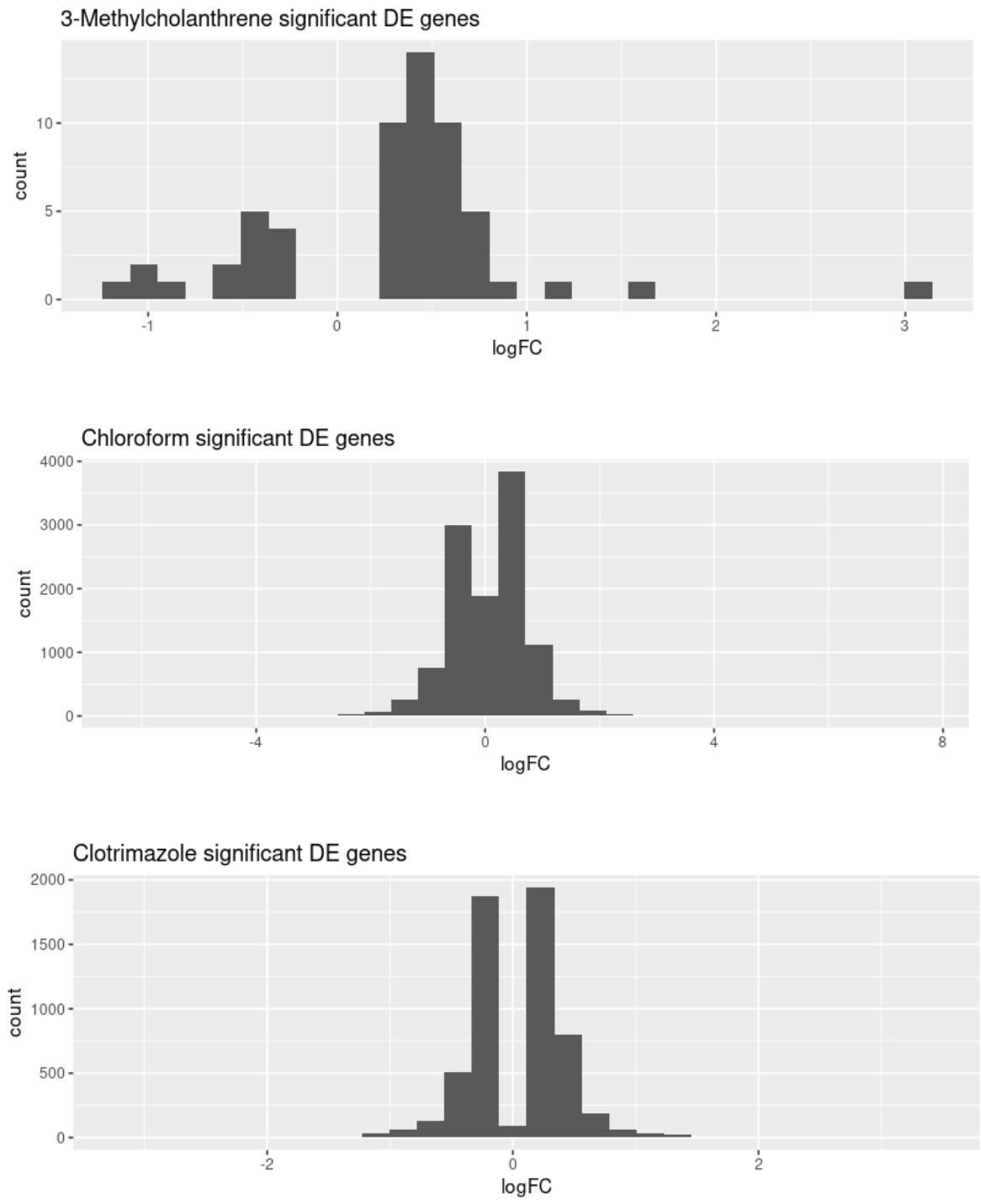

Figure 8. Histograms of the log fold change values versus the DE gene counts in the 3-Methylcholanthrene, Clotrimazole, and Chloroform microarray analysis results.

The concordance between the DESeq2 and limma platforms was calculated using the two sets of previously generated differential expression results. The equation to calculate the concordance used was (n0*N-n1*n2)/(n0+N-n1-n2). N is the number of all possible genes in the genome, n1 is the number of DE genes in the first analysis, n2 is the number of DE genes in the second analysis, and n0 is the amount of DE genes which are overlapping in both analysis results.

The first analysis is defined as the microarray analysis, and the second as the RNA-Seq analysis. The DE genes for each analysis were obtained through filtering the adjusted p-value less than 0.05. The overlapping DE genes were obtained by merging the two tables together using the sample names. Since the probe id from the microarray analysis is different from the refseq id from the DESeq analysis, the affymetrix map was required to join the two different data frames together.

``{r} f.x(21,58,313,20000) ```	```{r} f.x(768,5803,913,20000) ```	```{r} f.x(1787,11407,1810,20000) ```
[1] 20.45018	[1] 716.0448	[1] 1761.182

Table 5. The values plugged into the concordance function where concordance = f.x(n0, n1, n2, N). The concordance values from left to right correspond to 3-Methylcolantherene, Clotrimazole, and Chloroform.

The concordance values were also plotted against the total number of DE genes detected for each of the analyses. These plots also contain a linear regression line to better visualize the relationship between concordance and DE gene count between toxins.

Additionally, the DE genes were subdivided into above and below median groups, based on the median generated from the baseMean values in the DESeq results. After separating the genes into groups for both the DESeq and limma analyses, the concordance values were computed once again using the previous equations. The above median and below median concordance results were then combined with those from the median, or total DE gene concordance results. A bar plot was created containing all three sets of data, with color coding corresponding to the toxin.

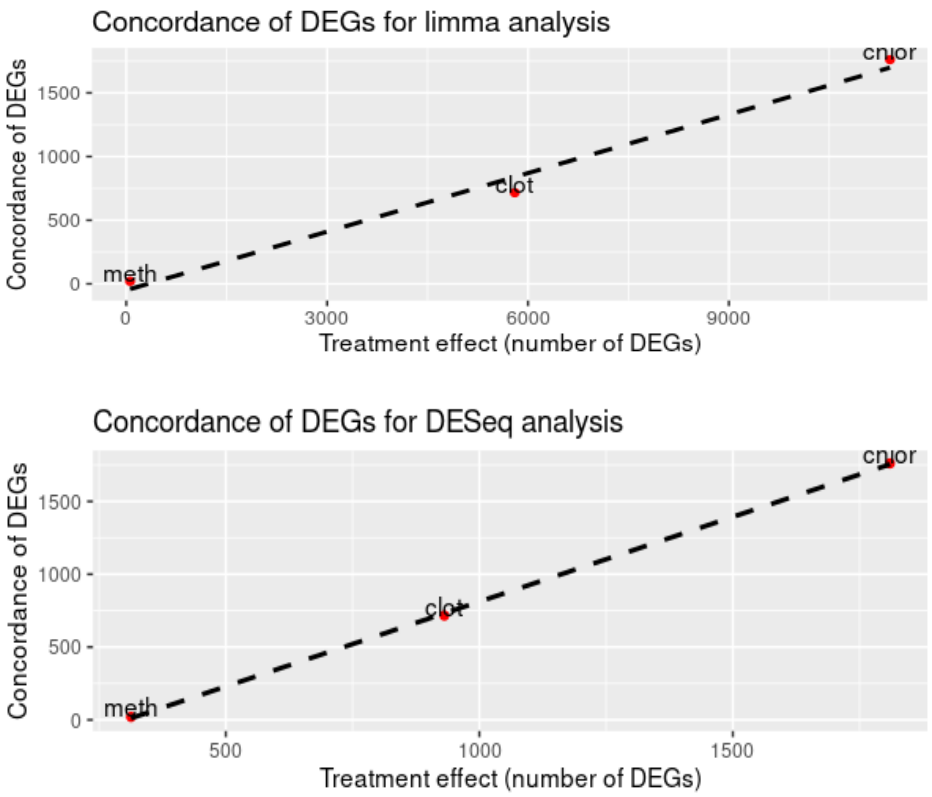

Figure 9. Plot with the concordance values for each of the toxins against the number of differentially expressed genes. There is one plot for both the DESeq and limma analyses

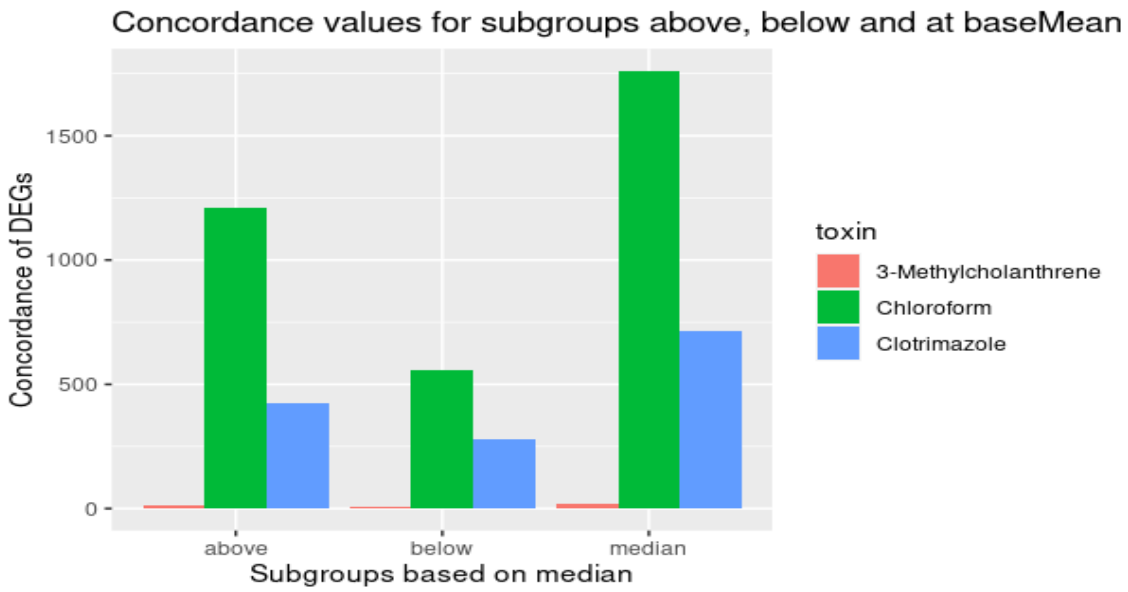

Figure 10. Barplots of concordance values for above median, below median, and total (median) subgroups of DE genes. Colors on the plots correspond to the toxins in the table.

For gene set enrichment analysis to compare the pathway enrichment for each of the MOA chemical groups Wang et al found a common pathway xenobiotic metabolism signaling which was common. In our analysis for the MOA in toxgroup 1-(AhR, CAR/PXR and Cytotoxic) all three results files were first filtered out at p adjusted <0.05 which were then run-on DAVID functional annotation tool to look for enriched pathways. In comparison with pathways found in Wang et al with our results from the DAVID though some of the enriched pathways referred to similar functions (metabolism of xenobiotics by cytochrome P450), most did not match the exact pathway described in the paper.

Table 6. AhR KEGG pathway annotation from DAVID.

Category	Term
KEGG_PATHWAY	Metabolic pathways
KEGG_PATHWAY	Metabolism of xenobiotics by cytochrome P450
KEGG_PATHWAY	Retinol metabolism
KEGG_PATHWAY	Chemical carcinogenesis - DNA adducts
KEGG_PATHWAY	Bile secretion
KEGG_PATHWAY	Chemical carcinogenesis - receptor activation
KEGG_PATHWAY	Drug metabolism - cytochrome P450
KEGG_PATHWAY	Fatty acid degradation
KEGG_PATHWAY	PPAR signaling pathway
KEGG_PATHWAY	Biosynthesis of unsaturated fatty acids
KEGG_PATHWAY	Steroid hormone biosynthesis

Table 7. CAR/PXR KEGG pathway annotation from DAVID.

Category	Term
KEGG_PATHWAY	Metabolic pathways
KEGG_PATHWAY	Steroid hormone biosynthesis
KEGG_PATHWAY	Chemical carcinogenesis - DNA adducts
KEGG_PATHWAY	Ribosome
KEGG_PATHWAY	Bile secretion
KEGG_PATHWAY	Coronavirus disease - COVID-19
KEGG_PATHWAY	DNA replication
KEGG_PATHWAY	ABC transporters
KEGG_PATHWAY	Cysteine and methionine metabolism
KEGG_PATHWAY	Ferroptosis
KEGG_PATHWAY	Peroxisome
KEGG_PATHWAY	Metabolism of xenobiotics by cytochrome P450
KEGG_PATHWAY	Glutathione metabolism
KEGG_PATHWAY	Drug metabolism - other enzymes
KEGG_PATHWAY	Retinol metabolism

Table 8. Cytotoxic KEGG pathway annotation from DAVID.

Category	Term
KEGG_PATHWAY	Metabolic pathways
KEGG_PATHWAY	Chemical carcinogenesis - DNA adducts
KEGG_PATHWAY	Steroid hormone biosynthesis
KEGG_PATHWAY	Retinol metabolism
KEGG_PATHWAY	Metabolism of xenobiotics by cytochrome P450
KEGG_PATHWAY	Ascorbate and aldarate metabolism
KEGG_PATHWAY	Drug metabolism - other enzymes
KEGG_PATHWAY	Chemical carcinogenesis - receptor activation
KEGG_PATHWAY	Bile secretion
KEGG_PATHWAY	Biosynthesis of cofactors
KEGG_PATHWAY	Drug metabolism - cytochrome P450

The heatmap failed to cluster by distinct mechanisms of action (MOA) together even after filtering the normalized expression matrix. Figure 11 depicts the heatmap that was yielded after filtering the matrix by coefficient of variation with cutoff set at 0.186 and then selecting the genes with means higher than the median.

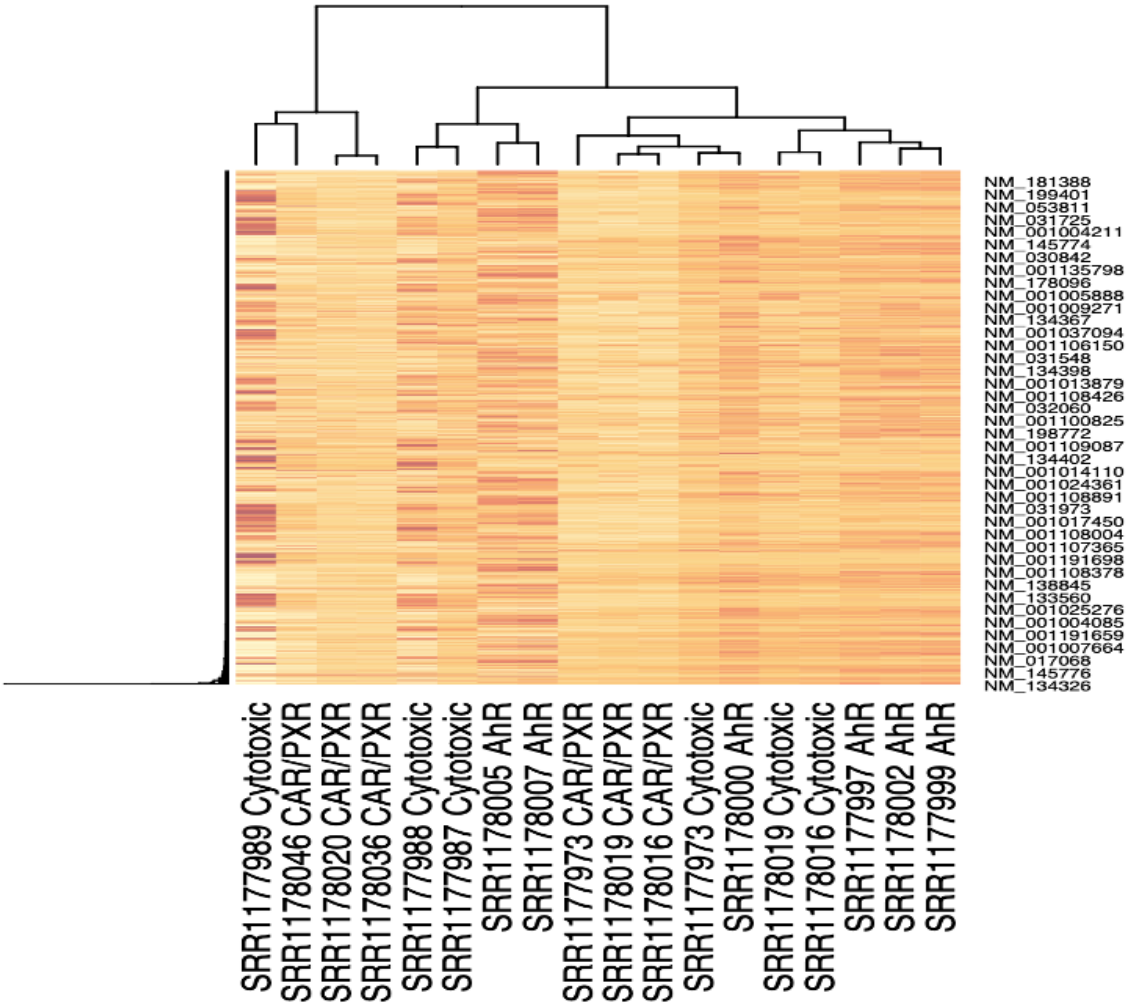

Figure 11. Hierarchical Cluster Heatmap of gene expression found using the normalized counts for each MOA (CAR/PXR, AhR, Cytotoxic). Since the matrix was large; we utilized the coefficient of variation to filter through the dataset. (CV cutoff >0.186).

Discussion

The data was obtained by aligning short reads to the rat genome, and quantifying the expression through read counting using STAR. Two separate differential expression analyses were performed on the data to later compare the results using concordance calculations. The first analysis was DESeq2, and the second was limma. These results were then mapped between the separate Affymetrix and refSeq identifiers using a provided table.

Looking at the number of differentially expressed genes in the limma analysis, the rats exposed to Chloroform had the highest count with 11,407, followed by Clotrimazole with 5,803, and 3-Methylcholanthrene with 58. After filtering the DE genes for the top 10 differentially expressed genes for each toxin, joining the affymetrix map table showed the names of the genes in question.

For rats exposed to 3-Methylcholanthrene, the top 10 differentially expressed genes contained two genes from the cytochrome superfamily, and eight glucuronosyltransferase genes. Cytochrome genes are responsible for clearance of certain compounds, and glucuronosyltransferase genes are responsible for removal of foreign chemicals and drugs from the body. Clotrimazole exposed rats had similar top 10 DE genes to 3-Methylcholanthrene exposed rats. The top DE genes consisted of 6 glucuronosyltransferase genes, 3 cytochrome genes, and a carboxylesterase gene which is responsible for metabolism of foreign chemicals. The rats that were exposed to Chloroform had a much different variety of top 10 differentially expressed genes, with seven different gene types. Two of which had duplicates were the ATP binding cassette family and the aldo-keto reductase family. ATP binding cassette genes are responsible for transportation, and aldo-keto reductase genes are responsible for drug metabolism.

From the scatter plots of the p-value versus the fold change of the DE genes, it is possible to see whether there is more or less expression in the treatment sets compared to the control. In the 3-Methylcholanthrene plot, there are more points located to the right of the 0 on the x-axis. This indicates that more genes are overexpressed rather than underexpressed in the experimentals compared to the control rats. For the Clotrimazole and Chloroform plots though, the high density of the points makes it difficult to identify whether the positive or negative side of the fold change

axis contains more genes. One thing that is possible to observe is that the genes which are most differentially expressed, meaning with the highest p-values, have relatively low fold change values. This means that these DE genes are only slightly over or under expressed compared to in the control conditions.

The histograms of the counts of fold change values for each of the toxins also reflect similar conclusions. All of the histograms have bars which cluster around zero, although there are no genes with fold changes of 0, because they would have the same expression as the controls and therefore not qualify as differentially expressed genes.

The concordance values which were calculated for 3-Methylcolantherene, Clotrimazole, and Chloroform were 20, 716, and 1761. This follows a trend with the number of DE genes found in each experimental condition, and as the number of DE genes increases, the concordance also increases. This relationship can be further supported by the plotting of the concordance values against the number of DE genes. Although the limma analysis has a maximum treatment effect, both of the analyses yield the same regression results. There is a positive relationship in the linear model of the number of DE genes and concordance values. In addition, the toxin points are located at relatively similar positions in both the limma and DESeq plots.

The bar plots of the different subgroups show that for Chloroform, there is higher concordance in the above median subgroup compared to the below median subgroup. This same result also holds true for the Clotrimazole and 3-Methylcholanthrene toxins as well. This means that there is a higher concordance between the two analysis methods for genes which have above median expressions.

Many of the figures and concordance results align with the conclusions of the original paper. Although we have only used a subset of three toxins, the concordance and DE gene counts mirror the relationship in the paper's results, that higher numbers of DE genes yield higher concordance between microarray and RNA-Seq analysis. Along that line, we were also able to replicate and confirm another one of their findings using subset groups of above and below median DE gene groups to find that there is higher concordance for above median expressed genes.

Conclusion

Wang et al sought to establish the concordance between microarray and RNA-seq gene expression profiling platforms. Their study used liver samples from rats affected with 27 chemicals which represented many different modes of action. We aimed to replicate portions of their study using data from only one mode of action, which consisted of the chemicals 3-Methylcolantherene, Clotrimazole, and Chloroform. By calculating the concordance values of each chemical between the two analyses, and plotting against the total number of DE genes observed for the chemical, the resulting linear regression shows that concordance between the two analyses increases with the number of DE genes. Another discovery regarding the concordance which we replicated was that higher concordance is found using above median expressed genes. This was seen when splitting the DE genes into subgroups based on the mean occurrences of the gene. These results conclude that there is in fact strong concordance between the two analyses for the mode of action which we were specifically working with.

One challenge which we encountered during this project was the different identifications of the genes between the REFSeq results and the limma results. When trying to join and compare the two data sets, it was impossible due to the incompatibility of the gene identifications. This challenge was overcome by using the provided Affymetrix mapping table, which allowed for both the identification methods to be joined to either table. This way comparisons and calculations between the two analysis results would be simpler and made possible.

References

1. Wang, Charles et al. “The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance.” *Nature biotechnology* vol. 32,9 (2014): 926-32. doi:10.1038/nbt.3001
2. McDonnell, Anne M, and Cathyyen H Dang. “Basic review of the cytochrome p450 system.” *Journal of the advanced practitioner in oncology* vol. 4,4 (2013): 263-8. doi:10.6004/jadpro.2013.4.4.7
3. Rowland A, Miners JO, Mackenzie PI. The UDP-glucuronosyltransferases: their role in drug metabolism and detoxification. *Int J Biochem Cell Biol*. 2013 Jun;45(6):1121-32. doi: 10.1016/j.biocel.2013.02.019. Epub 2013 Mar 7. PMID: 23500526.

4. Laizure, S Casey et al. “The role of human carboxylesterases in drug metabolism: have we overlooked their importance?.” *Pharmacotherapy* vol. 33,2 (2013): 210-22. doi:10.1002/phar.1194
5. Locher, Kaspar P. “Review. Structure and mechanism of ATP-binding cassette transporters.” *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* vol. 364,1514 (2009): 239-45. doi:10.1098/rstb.2008.0125
6. Liao, Yang et al. “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.” *Bioinformatics (Oxford, England)* vol. 30,7 (2014): 923-30. doi:10.1093/bioinformatics/btt656
7. Love MI, Huber W, Anders S . “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” *Genome Biology* vol. 15,550 (2014). doi: 10.1186/s13059-014-0550-8.

Supplementary Materials

Table S1. Experiment design, as provided to DESeq2 for determining DE genes. Each set of conditions was performed in triplicate, grouped for analysis by DESeq2.

Sample	Mode of Action	Chemical	Vehicle
SRR1177997	AhR	3-METHYLCHOLANTHRENE	CMC_.5_%
SRR1177999	AhR	3-METHYLCHOLANTHRENE	CMC_.5_%
SRR1178002	AhR	3-METHYLCHOLANTHRENE	CMC_.5_%
SRR1178020	CAR/PXR	CLOTRIMAZOLE	CORN_OIL_100_%
SRR1178036	CAR/PXR	CLOTRIMAZOLE	CORN_OIL_100_%
SRR1178046	CAR/PXR	CLOTRIMAZOLE	CORN_OIL_100_%
SRR1177987	Cytotoxic	CHLOROFORM	CORN_OIL_100_%
SRR1177988	Cytotoxic	CHLOROFORM	CORN_OIL_100_%
SRR1177989	Cytotoxic	CHLOROFORM	CORN_OIL_100_%
SRR1178000	Control	Vehicle	CMC_.5_%
SRR1178005	Control	Vehicle	CMC_.5_%
SRR1178007	Control	Vehicle	CMC_.5_%
SRR1177973	Control	Vehicle	CORN_OIL_100_%
SRR1178016	Control	Vehicle	CORN_OIL_100_%
SRR1178019	Control	Vehicle	CORN_OIL_100_%

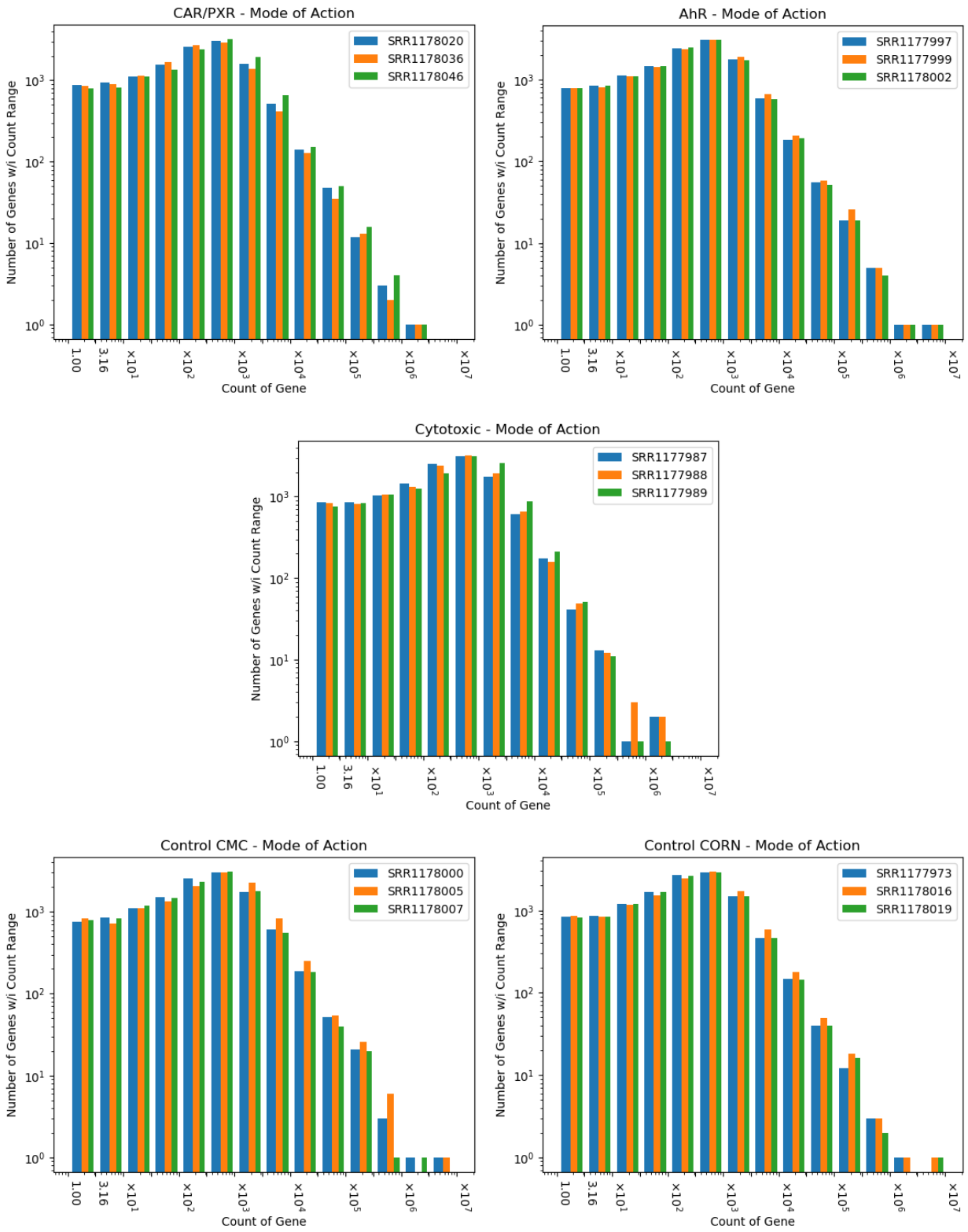

Figure S1. Per sample log-log histograms, stacked by mode of action, displaying the gene count distribution in greater detail.