Concordance of microarray and RNA-Seq differential gene expression

Group Roles: Sriramteja Veerisetti (Data Curator), Mano Ranaweera (Programmer), Benyu Zhou (Analyst), Lucas Zhang (Biologist)

Introduction

RNA-Sequencing utilizes next-generation sequencing mechanisms in order to quantify the presence of RNA within a biological sample. This sequencing mechanism produces data that allows scientists to analyze which genes are active within a cell, thus allowing them to better understand the transcription levels within the cell[1]. However, in order to confidently utilize RNA-Sequencing in a clinical and regulatory environment, its concordance with microarray analysis needs to be further researched. In 2014, Wang et al. utilized a common set of liver samples from rats, that were each exposed to different degrees of perturbation with 27 chemicals that represented multiple modes of action (MOA), in order to produce Illumina RNA-seq and Affymetrix microarray data[2]. A cross-platform analysis was conducted between Illumina RNA-seq and Affymetrix microarray methodologies in order to determine differentially expressed gene (DEG) verification. At the conclusion of the study, it was determined that performance of RNA-Seq was far more precise and accurate at 90% DEG verification compared to the 76% DEG verification via microarray[2].

The purpose of the project was to reproduce the concordance analysis within the Wang et. al paper by specifically targeting the toxgroup_6_rna containing 3 chemicals (3-Methylcholanthrene, Fluconazole, Pirinixic Acid) as well as to compare the results from the pathway enrichment process to the paper.

Data

In the Wang et al. study, both the microarray data and RNA-Seq components were valued because of the objective of the study. However in this particular project, only the RNA-Seq data was viewed. 9 samples, which were subject to either 3-Methylcholanthrene, Fluconazole, Pirinixic Acid, were chosen from the toxgroup_6_rna group. This specific subset of samples was further categorized based on a mode of action, which could have been either Aryl hydrocarbon receptor (AhR), CAR/PXR, or peroxisome proliferator-activated receptor alpha (PPARA).

After the treatment samples from the toxgroup_6_rna were selected and symbolic links to each sample were made, they underwent a FastQC process in order to measure the quality of the data. The yielding FASTQ files were then aligned against the genome of a Sprague-Dawley rat via the STAR aligner tool in order to produce 9 separate bam files[5]. In order to garner statistical information on the biological samples, a MultiQC process was run on the bam files[9].

Sample	Mode of Action (MOA)	Chemical	Vehicle	Route
SRR1177997	AhR	3-Methylcholanthrene	CMC_.5_%?%	Oral_Gavage
SRR1177999	AhR	3-Methylcholanthrene	CMC_.5_%?%	Oral_Gavage
SRR1178002	AhR	3-Methylcholanthrene	CMC_.5_%?%	Oral_Gavage
SRR1178014	CAR/PXR	Fluconazole	Corn_Oil_100_%	Oral_Gavage
SRR1178021	CAR/PXR	Fluconazole	Corn_Oil_100_%	Oral_Gavage
SRR1178047	CAR/PXR	Fluconazole	Corn_Oil_100_%	Oral_Gavage
SRR1177963	PPARA	Pirinixic Acid	CMC_.5_%?%	Oral_Gavage
SRR1177964	PPARA	Pirinixic Acid	CMC_.5_%?%	Oral_Gavage
SRR1177965	PPARA	Pirinixic Acid	CMC_.5_%?%	Oral_Gavage

Table 1. Summary table that includes the MOA, Chemical, Vehicle, and Route that each biological sample was exposed to.

STAR Analysis					
Sample ID	Average Read Sequence Length	Uniquely mapped reads %	Duplicate Reads %	% of reads mapped to multiple loci	% of reads unmapped:
SRR1177963_1	101 bp	84.83%	54.8%	3.70%	11.20%
SRR1177963_2	101 bp		52.3%		
SRR1177964_1	101 bp	85.33%	57.3%	3.67%	10.71%
SRR1177964_2	101 bp		54.8%		
SRR1177965_1	101 bp	85.10%	54.5%	3.85%	10.71%
SRR1177965_2	101 bp		51.8%		
SRR1177997_1	101 bp	89.17%	59.6%	3.88%	6.67%
SRR1177997_2	101 bp		58.6%		
SRR1177999_1	101 bp	88.72%	60.2%	3.92%	6.98%
SRR1177999_2	101 bp		58.9%		
SRR1178002_1	101 bp	89.13%	58.5%	3.92%	6.59%
SRR1178002_2	101 bp		57.6%		
SRR1178014_1	50 bp	83.52%	53.9%	6.57%	9.20%
SRR1178014_2	50 bp		51.9%		
SRR1178021_1	200 bp	81.95%	48.7%	5.77%	11.92%
SRR1178021_2	100 bp		46.4%		
SRR1178047_1	100 bp	83.98%	48.5%	5.85%	9.67%
SRR1178047_2	100 bp		47.3%		

Table 2. Summarization of the STAR alignment statistics. The table contains the biological sample, average read sequence length, uniquely mapped reads %, duplicate reads %, % of reads mapped to multiple loci, and the % of reads unmapped.

When comparing the genome of the Sprague-Dawley rat to the sequences of the toxgroup_6_rna subset there were some key points that were specifically analyzed. For example, the average read sequence length is 100.66, while the average uniquely mapped percentage is 85.75%. The high uniquely mapped percentage average indicates that the quality of the sample reads is good. The average uniquely mapped reads was 15.9 million and the range of the uniquely mapped reads was 14.3 to 19.4 million.

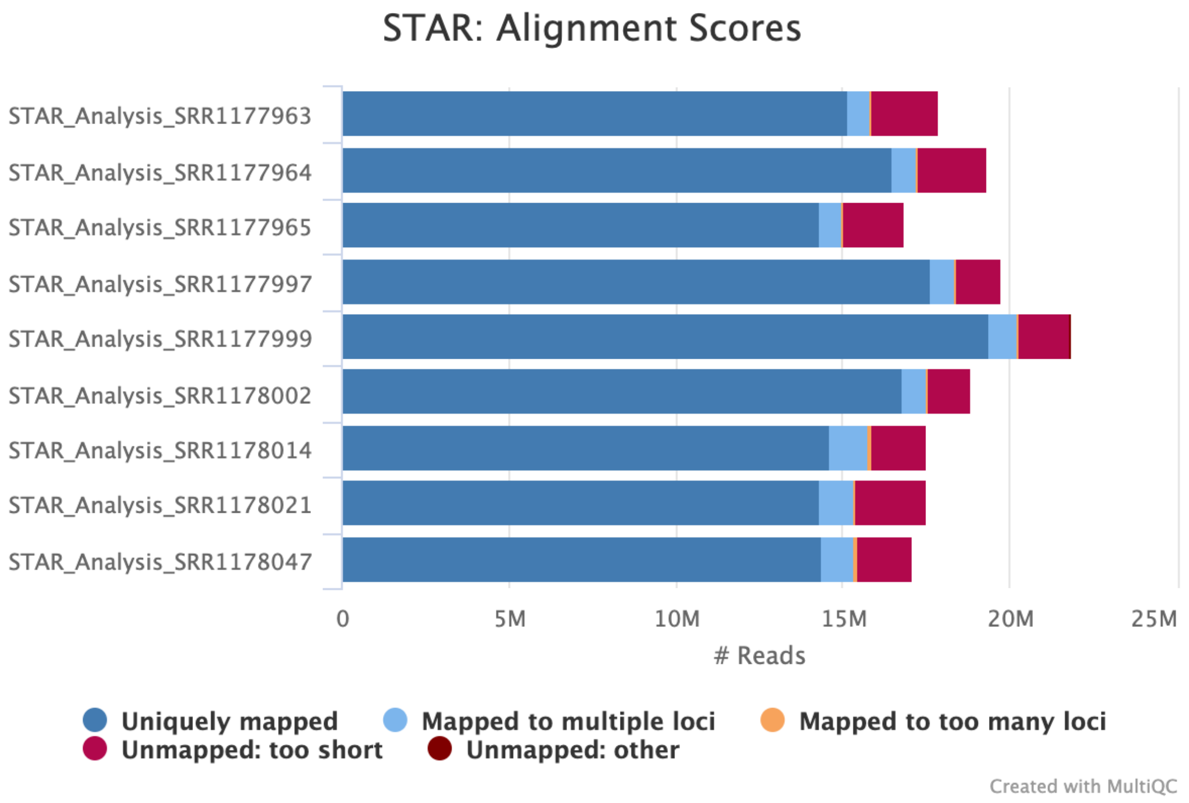

Figure 1. STAR: Alignment Scores for each biological sample. The x-axis represents the number of reads. The y-axis represents the samples. Accumulate to mapped to multiple loci, mapped to too many loci, unmapped: too short, or unmapped: other.

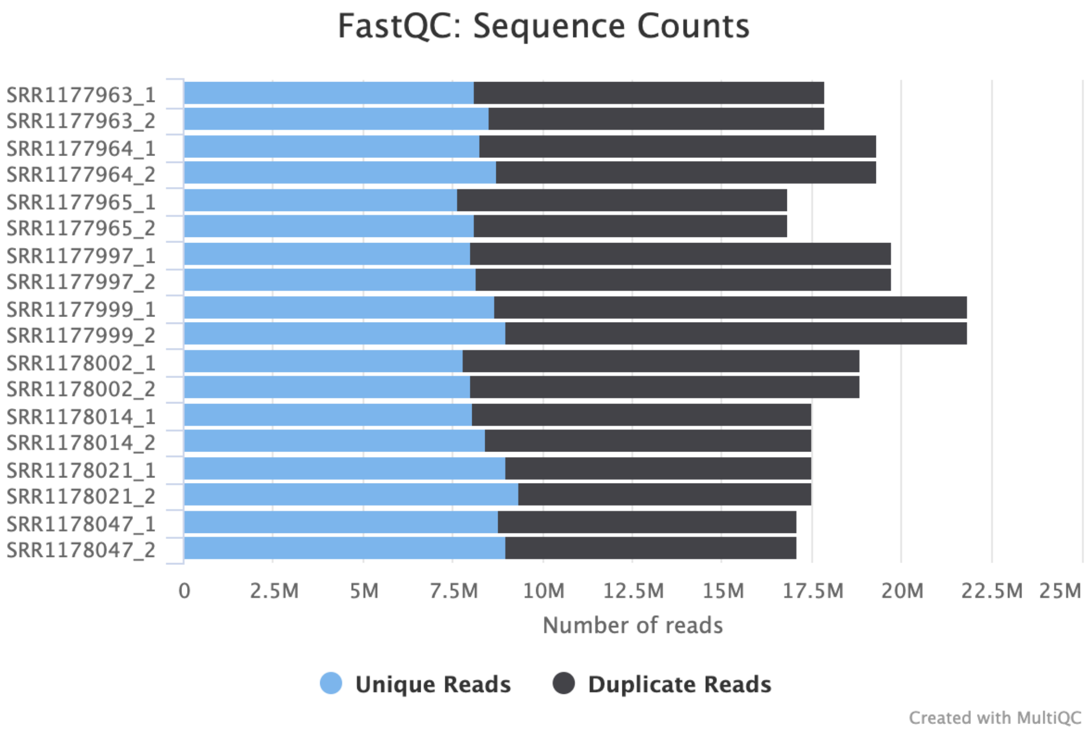

Figure 2. The Number of Reads are on the x-axis and the biological samples are on the y-axis. The number of reads are listed from 0M to 25M.

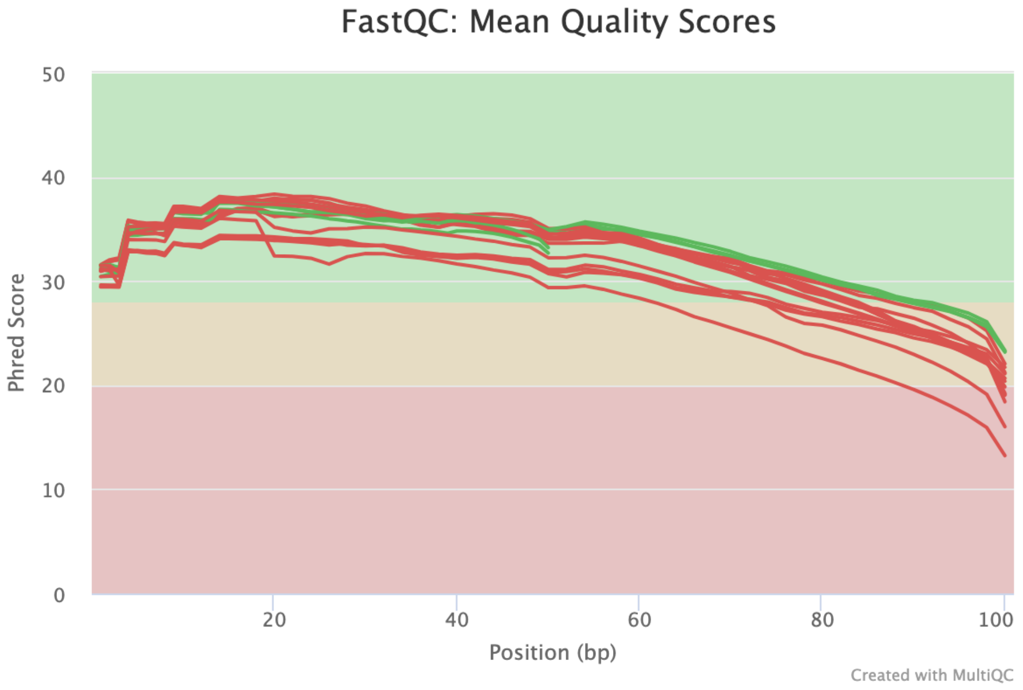

Figure 3. FastQC Mean Quality Scores with the Position (bp) on the x-axis and the Phred Score on the y-axis.

Figures 1, 2, and 3 are all made from the MultiQC tool. Based on Figure 1, most of the reads for each biological sample are uniquely mapped, while only a small piece of each biological sample accumulates to: mapped to multiple loci, mapped to too many loci, unmapped: too short, or unmapped: other. On the other hand, Figure 2 shows the number of reads for each biological sample, with x-axis being scaled from 0M to 25M. Figure 3 visualizes FASTQC mean quality scores for the samples and has the position in bp on the x-axis and the Phred Score on the y-axis. The Phred Score measures the quality of the nucleotide bases that are generated from the sequencing tool. A key source of error, based on Figure 3, was that at first a majority of the Phred scores were high, indicating quality reads, but slowly took a dip into poor quality near the end of the plot. Roughly 60-70% of the reads were in the green passing zone, which gave the group confidence to move forward through the project. After concluding extensive analysis on each of the biological samples, it was concluded that no samples would need to be eliminated due to low quality.

Methods

Through featureCounts, a summarization tool, the bam files from the STAR alignments for each of the nine samples were used to count the number of reads against a gene annotation, with a count file generated for each sample. MultiQC was run on each of the counts files to visualize the distribution of reads that were either assigned or unassigned to the reference genome, as seen below Figure 4.

Each of the nine counts files were then combined into one matrix, with all the counts for each gene in the same row. The boxplot seen below in Figure 5 shows the distribution of counts plotted on a log scale. Each sample has a similar box-and-whiskers plot, as we should be seeing. A sample with a significantly different distribution would require further attention.

To analyze the count data that is put together, the control sample counts in the same group as the treatment samples were added to the data, and DESEQ2 was used for negative binomial regression to estimate count differences. Negative binomial regression is appropriate for this count data due to there not being a normal distribution, only discrete values in the data, and the mean and variance of a gene’s count distribution have a dependent relationship with each other.

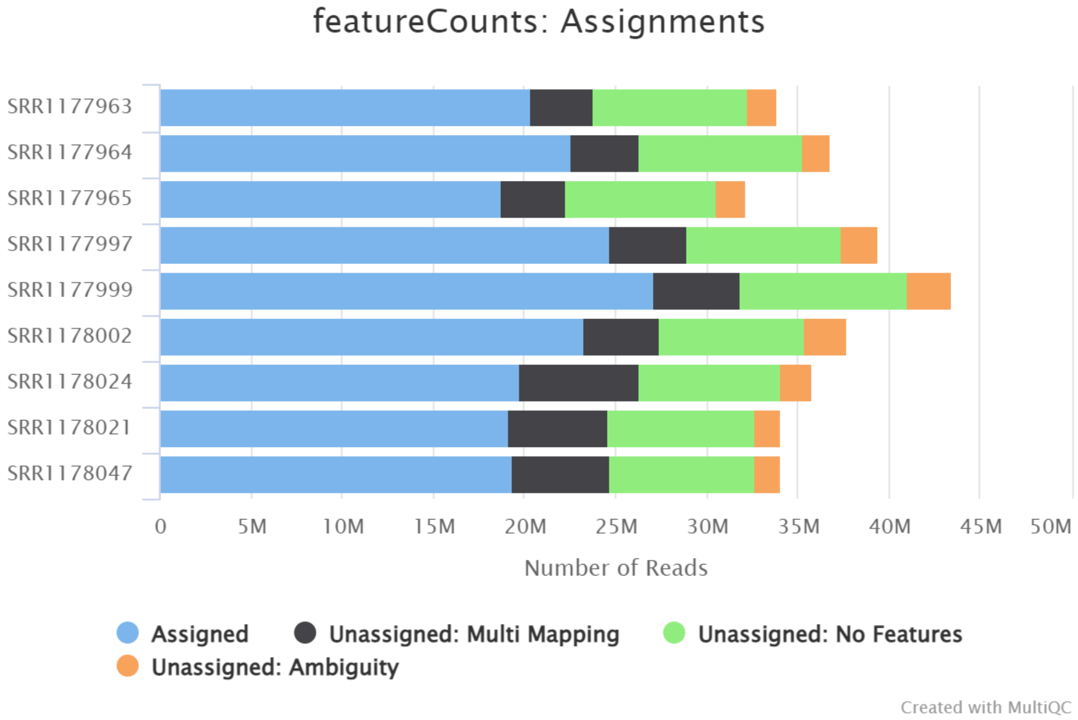

Figure 4. FastQC read assignment distribution with number of reads on the x-axis and samples on the y-axis.

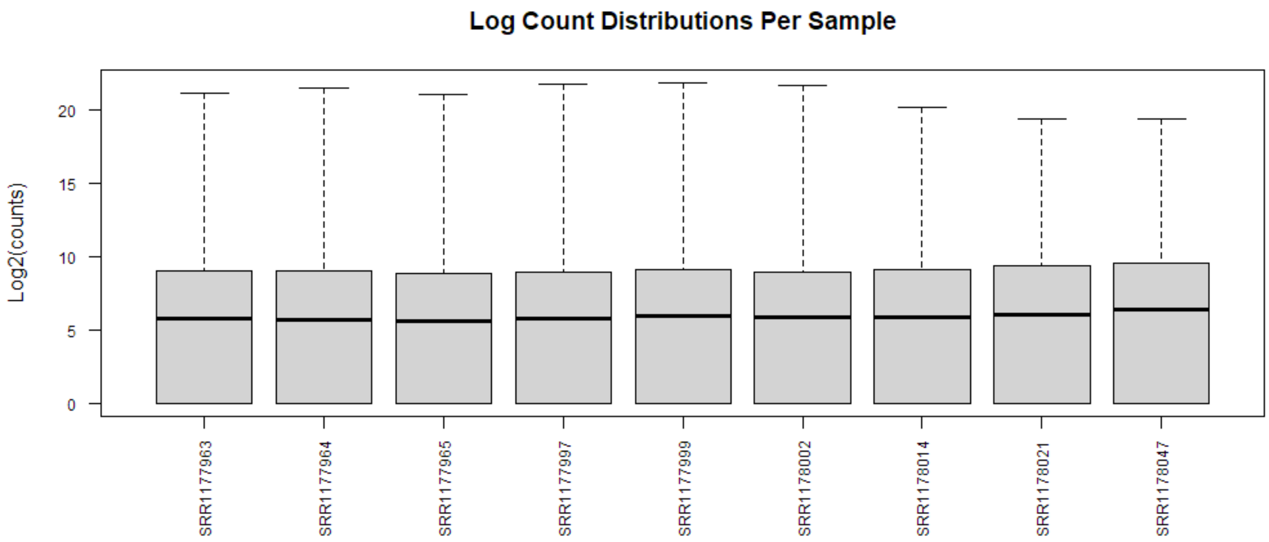

Figure 5. Boxplot of count distributions for each sample.

Running DESEQ2 would result in a row of the genes, with the ability to identify which genes are differentially and significantly expressed. For each of the three treatments, the counts matrix was subsetted into three groups, with the samples from each treatment(3-methylcholanthrene, fluconazole, and pirinixic acid) and the appropriate controls, which had to have the same vehicle(solvent for dissolving treatment) as the treatment. With the DESEQ results, the top ten significant genes from each treatment were able to be identified. The DESEQ data was normalized by dividing the counts sample-specific size factors. These factors are determined by the median ratio of gene counts relative to the geometric mean per gene. Tidyverse, ggplot2, and Deseq2 from Bioconductor were the packages used for this analysis. The documentation for Deseq2 was needed to refer to many times, as well as referring to ggplot features for proper data visualization.

The Limma package was used to analyze the microarray differential expression results data of 3 chemicals (3-Methylcholanthrene, Fluconazole, Pirinixic Acid) vs control group in Rstudio. Then reports of DE genes with p-value <0.05 were generated. The limma results were displayed in histograms and scatter plots.

With both RNA-seq and MIcroarray differential expression analysis done, concordance is calculated with the following equation.

$$n_x = \frac{Nn_0 - n_1 n_2}{n_0 + N - n_1 - n_2}$$

In the equation above, n0, n1, n2 indicate the DE genes number of overlap between microarray and RNA-seq, microarray and RNA-seq respectively. N indicates the total gene numbers in rats, which is 54879 genes with nucleotide sequence data, according to the article *Mouse genome database 2016,* by Bult C.J. et al. [3] nx is the background corrected number of “true” overlapping genes. With nc calculated, concordance can be calculated using the equation below.

$$concordance = \frac{2 * n_x}{n_1 + n_2}$$

With the overall concordance calculated for each of the chemicals, the Microarray results and RNA-seq results were further divided into above median and below median groups. The medians were determined using baseMean value for the RNA-seq results and AveExpr value for the Microarray results. The same equations were used to calculate the concordance again. A bar chart comparing overall vs above-median vs below-median for all 3 chemicals was generated.

Enriched pathway analysis was then performed for the genes found to be differentially expressed for each MoA (AhR, CAR/PXR, PPARA). The differentially expressed gene names (genes with <0.05 p-value) were first pulled from the CSV file for each of the analyses, totaling six different lists. The gene name lists were then inputted into DAVID, with the resulting text file run through a python program in order to extract the resulting enriched KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways. The enriched pathways from the six individual lists were then compared in pairs based on the MoA, outputting shared enriched pathways based on both RNA-seq and microarray based platform analysis.

Following the enriched pathway analysis, hierarchical clustering was done on a collated table with all the normalized counts from samples for each MoA (AhR, CAR/PXR, PPARA) as well as a control to determine whether aforementioned samples could accurately be clustered by MoA. These samples included three from each MoA listed above (SRR1177997, SR1177999, SR1178002, SR1178014, SR1178021, SR11778047, SR1177963, SR1177964, SR1177965) as well as six control samples. After filtering the unmodified data to remove genes with a coefficient of variation greater than 0.2, and average counts of less than a hundred, samples from the normalized counts matrix were correctly clustered into each corresponding MoA, with control samples correctly clustered as well. The heatmap function in R was used to perform hierarchical clustering, with data processing done in R as well.

Results

Figures 6-11 show visualizations of the DE expression data. The histograms show the counts of log 2 fold change values, and the scatter plots plot fold change vs nominal p-value for DE genes. The plots for the 3-Methylcholanthrene treatment are more spread out than the others, due to there being much less DE genes found. It can be concluded that this treatment has a low effect.

3-Methylcholanthrene	Fluconazole	Pirinixic Acid
NM_022521	NM_001014166	NM_017158
NM_012608	NM_001005384	NM_131903
NM_022297	NM_031048	NM_019157
NM_134329	NM_144755	NM_001014063
NM_022866	NM_053288	NM_024162
NM_022635	NM_031605	NM_001013098
NM_053883	NM_053699	NM_053883
NM_012541	NM_001130558	NM_012600
NM_130407	NM_013105	NM_012737
NM_017061	NM_013033	NM_001013975

Table 3. The table contains the top ten DE genes from each of the three analyses, in terms of p-value.

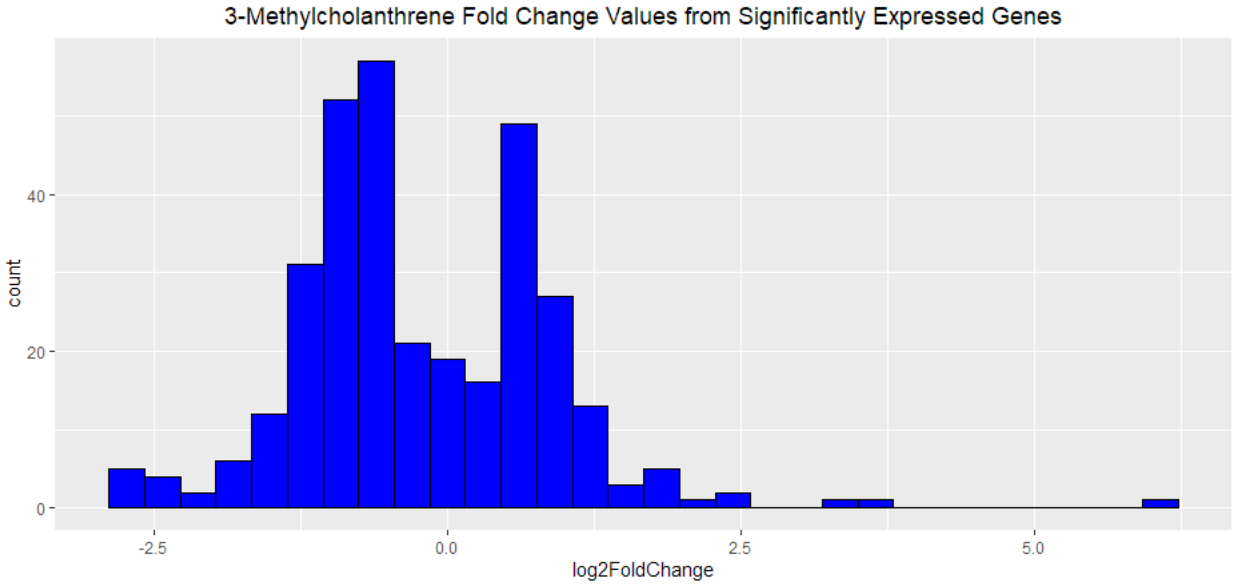

Figure 6. Histograms of log fold change values from the significant DE genes in the 3-Methylcholanthrene treatment group.

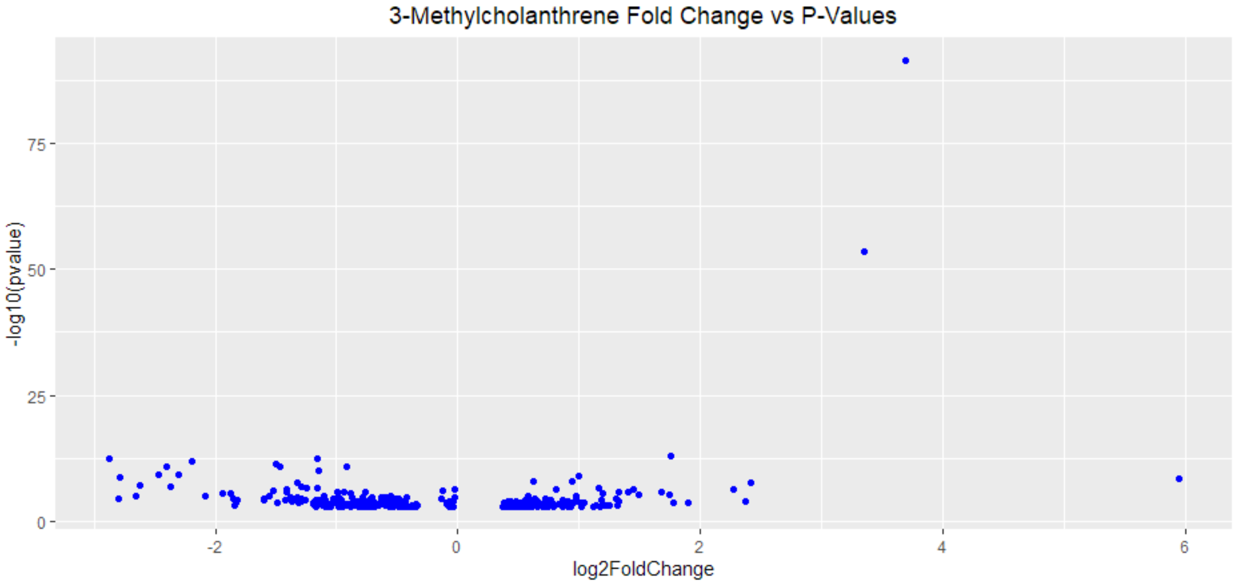

Figure 7. Scatter plots of log fold change vs nominal p-value from the significant DE genes in the 3-Methylcholanthrene treatment group

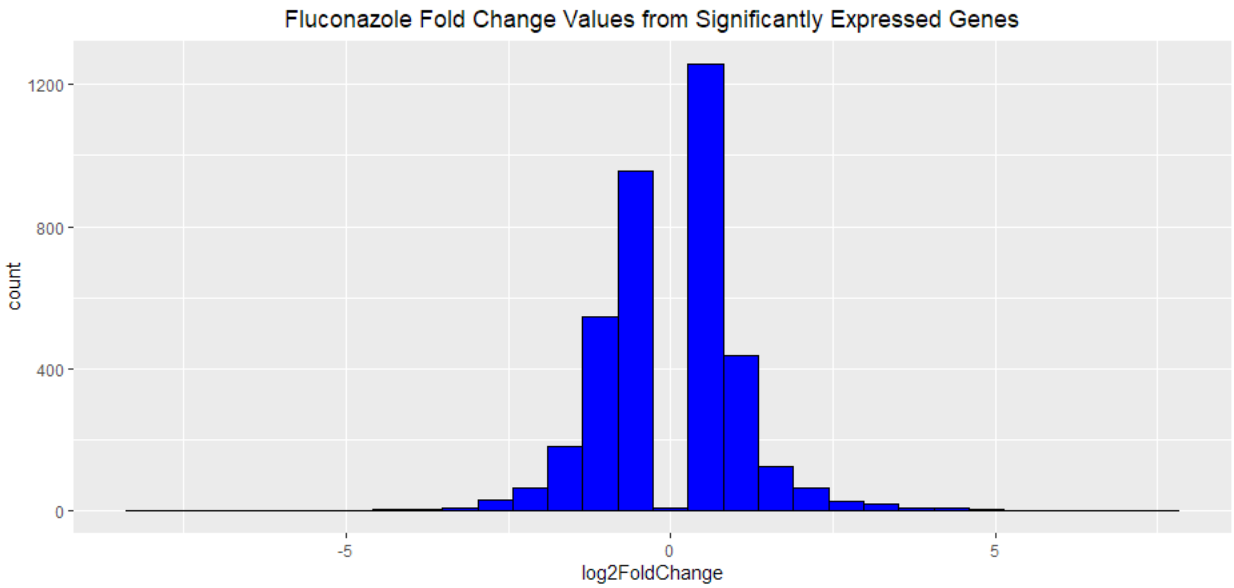

Figure 8. Histograms of log fold change values from the significant DE genes in the Fluconazole treatment group.

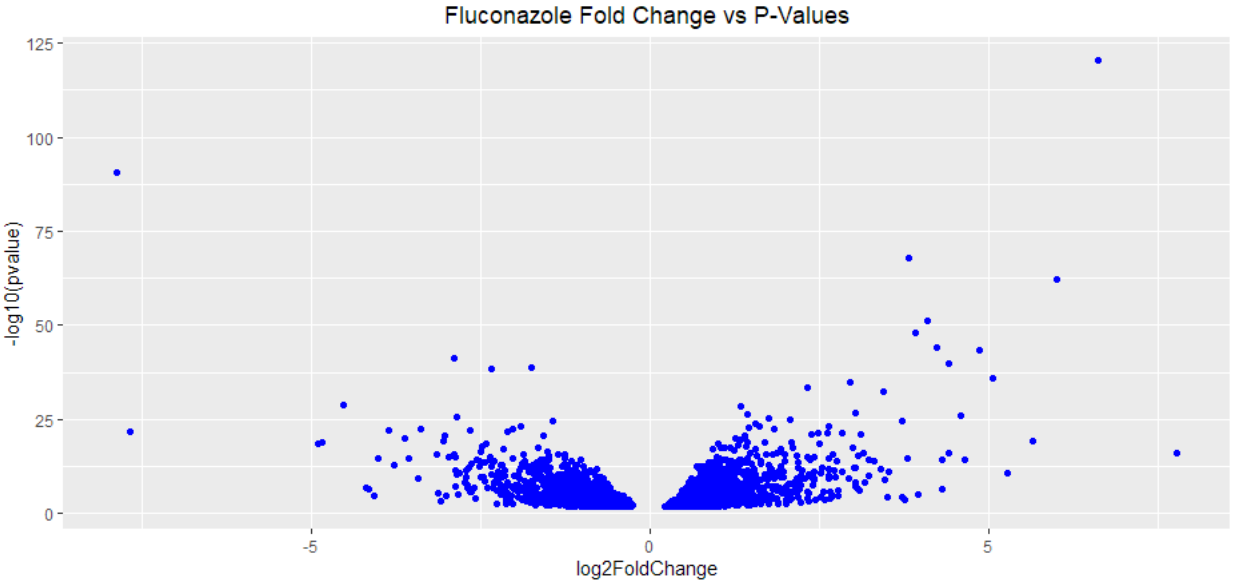

Figure 9. Scatter plots of log fold change vs nominal p-value from the significant DE genes in the Fluconazole treatment group.

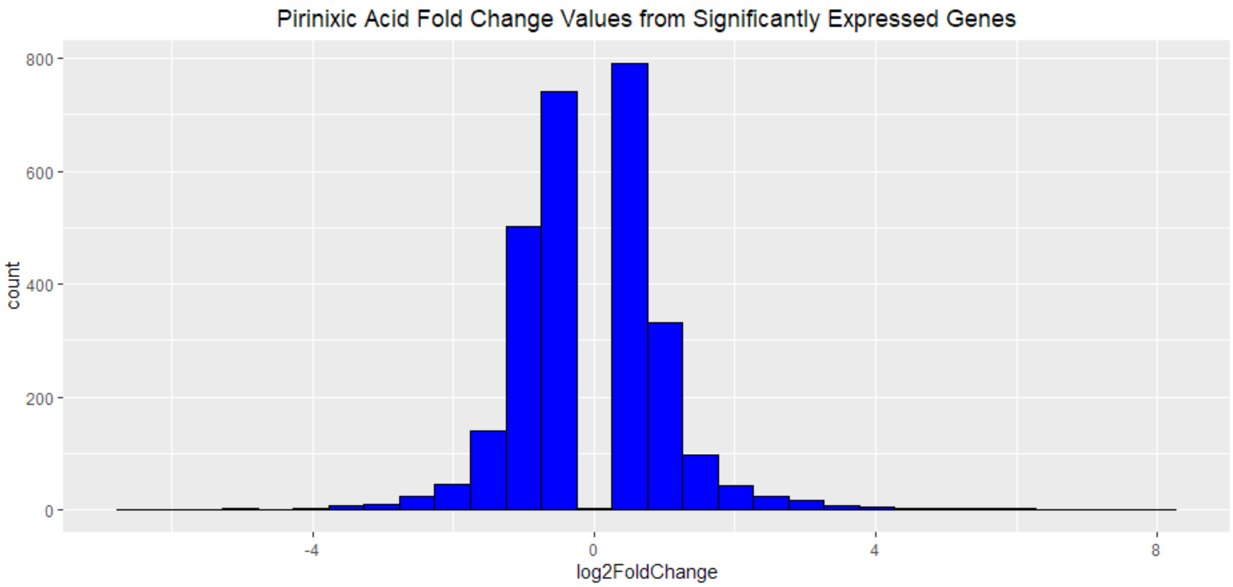

Figure 10. Histograms of log fold change values from the significant DE genes in the 3-Methylcholanthrene treatment group.

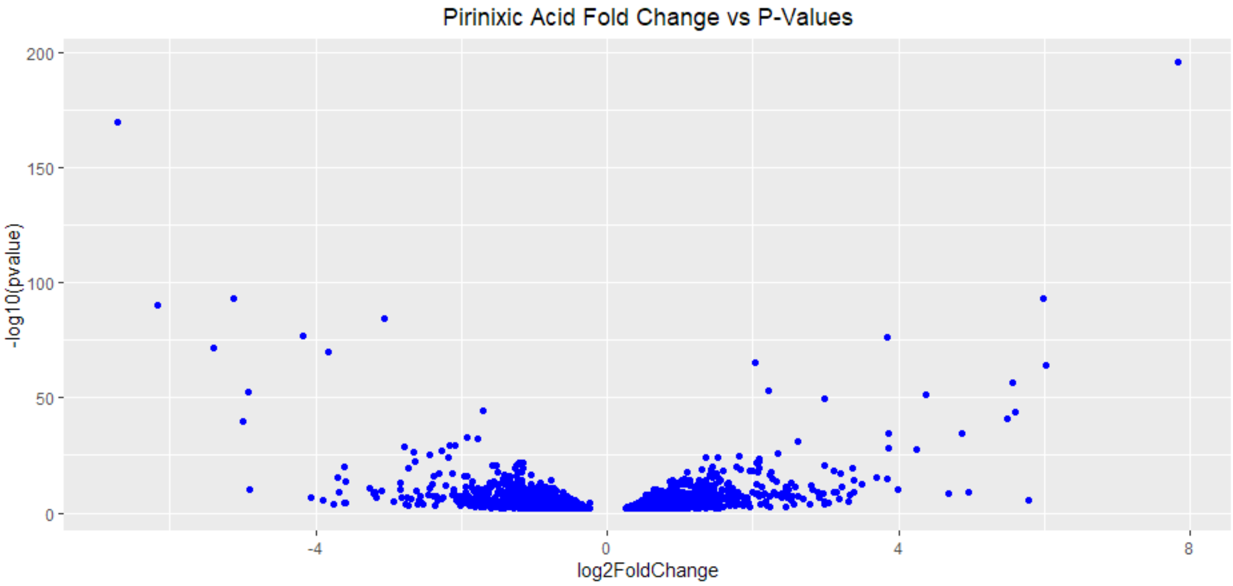

Figure 11. Scatter plots of log fold change vs nominal p-value from the significant DE genes in the Pirinixic Acid treatment group.

Limma analysis on the microarray differential expression found 33 DE genes in the 3-Methylcholanthrene treatment group, 6016 DE genes in the Fluconazole treatment group, and 8789 genes in the Pirinixic Acid treatment group. The table 4 below shows the top 10 DE genes ordered by p-value for each of the 3 chemicals. The csv files in the repository have complete DE genes information. (Analysis of the data was done using the limma package v3.46.0 [7]).

3-Methylcholanthrene	Fluconazole	Pirinixic Acid
1387243_at	1390255_at	1398250_at
1370613_s_at	1391570_at	1388211_s_at
1370269_at	1377192_a_at	1370491_a_at
1387901_at	1378029_at	1367742_at
1387759_s_at	1372136_at	1390358_at
1387511_at	1371076_at	1377867_at
1384544_at	1368731_at	1367680_at
1368047_at	1394022_at	1380536_at
1383325_at	1384408_at	1375845_at
1372297_at	1371840_at	1384474_at

Table 4 Top 10 DE genes ordered by p-value for each of the 3 chemicals

Figure12-14 are histograms of fold change values and scatter plots of fold change vs nominal p-value from the significant DE genes for each of the 3 chemicals. The histogram for 3-Methylcholanthrene looks spread out due to low DE genes number, but the peaks are still centered around logFC=0/FC=1 The histogram for Fluconazole looks pretty normally distributed with peaks centered around logFC=0/FC=1. The histogram for Pirinixic Acid is heavily skewed right. There are DE genes with extremely high Fold change in this group. Still, the third histogram is also peaked around logFC=0/FC=1

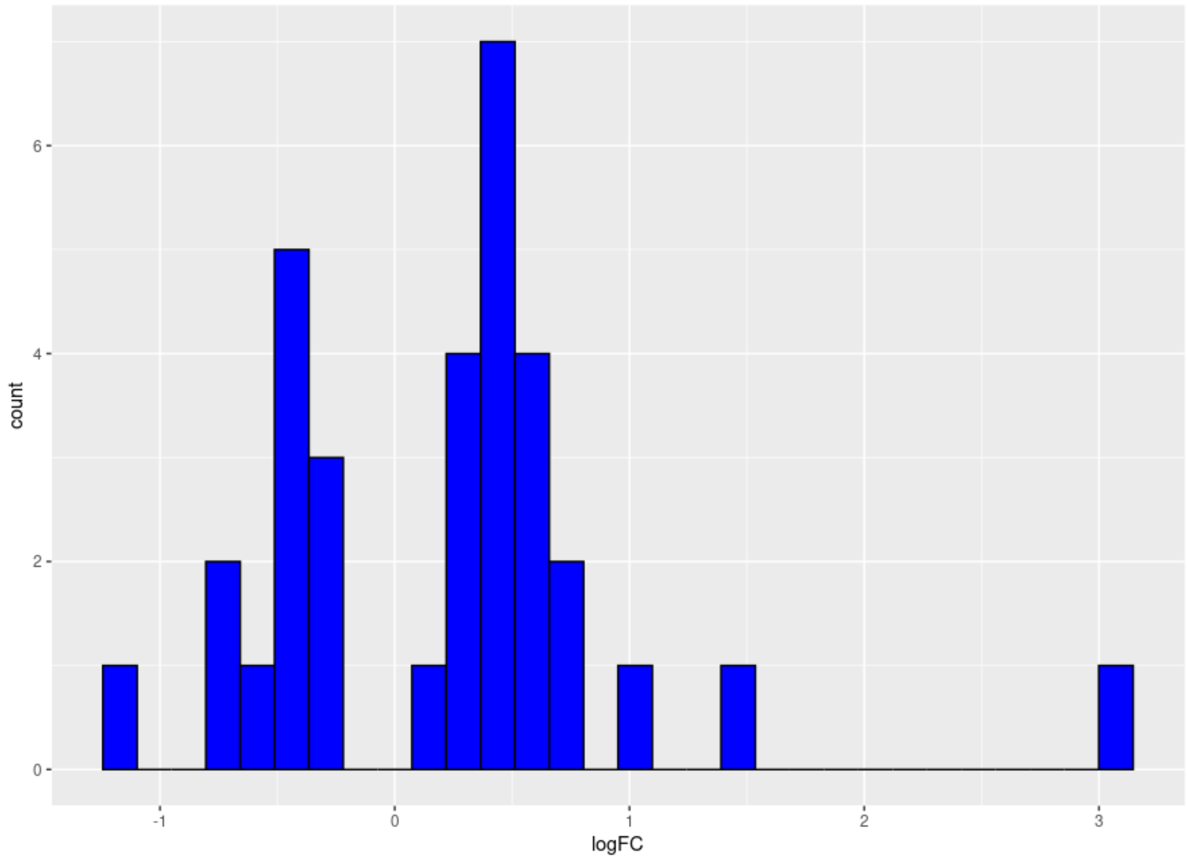

Figure 12 Histograms of log fold change values from the significant DE genes in the 3-Methylcholanthrene treatment group

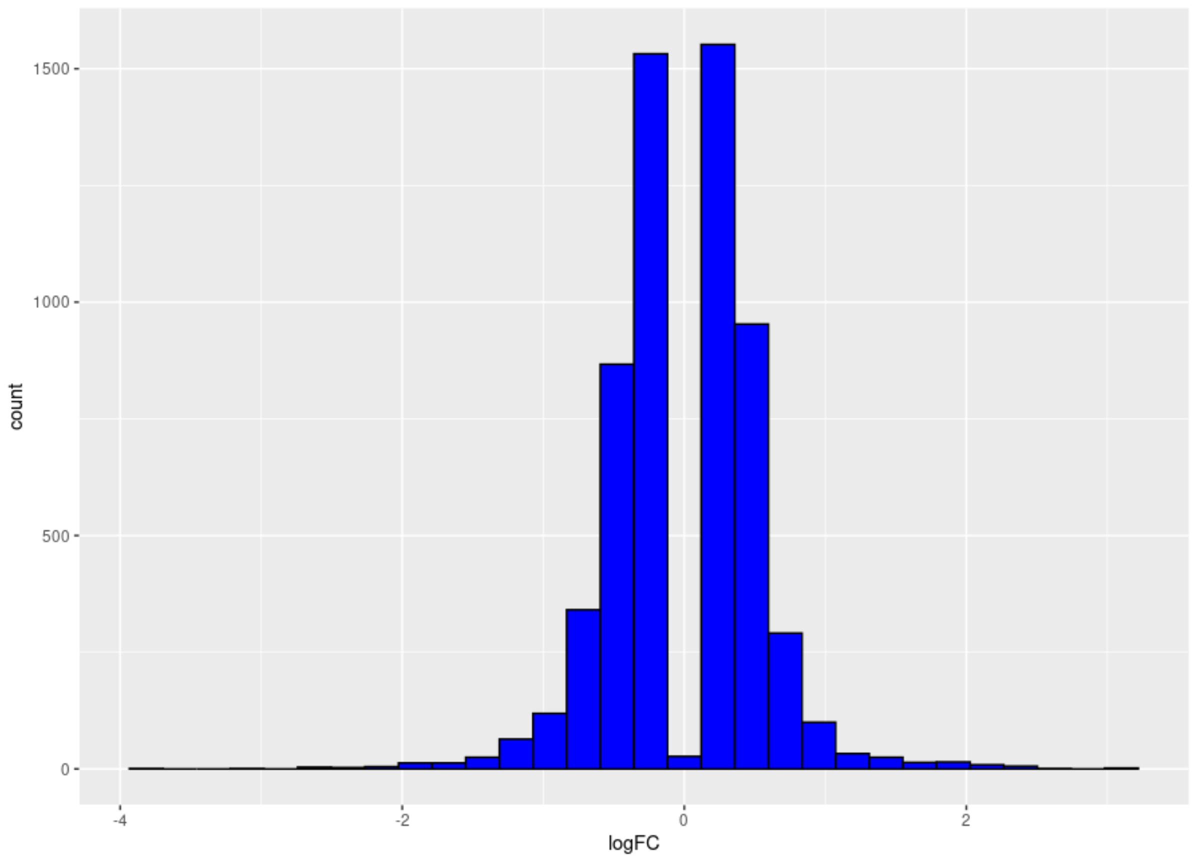

Figure 13 Histograms of log fold change values from the significant DE genes in the Fluconazole treatment group

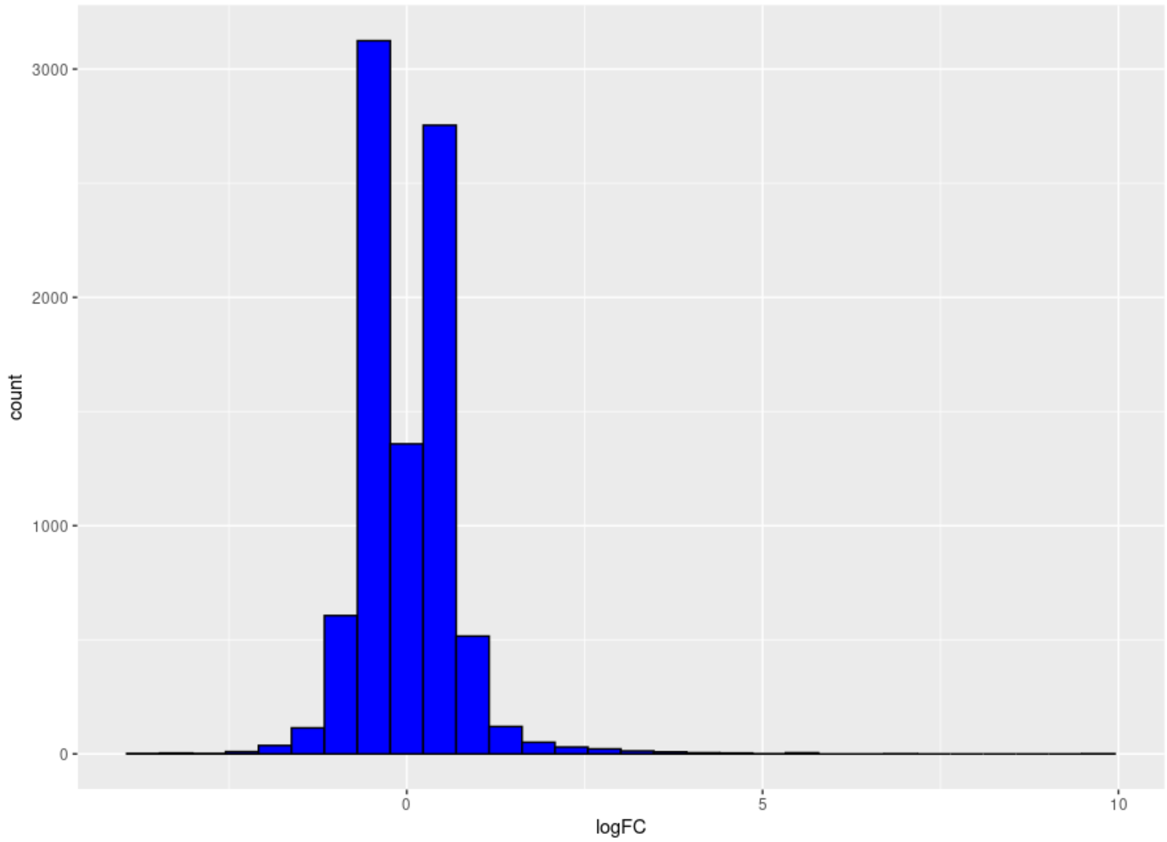

Figure 14 Histograms of log fold change values from the significant DE genes in the Pirinixic Acid treatment group

Despite the fact that The first group,3-Methylcholanthrene treatment group, has far fewer DE genes in the results, the 3 scatter plots (figure 15-17) all show the trend that the higher the fold change, the lower the p-value is.

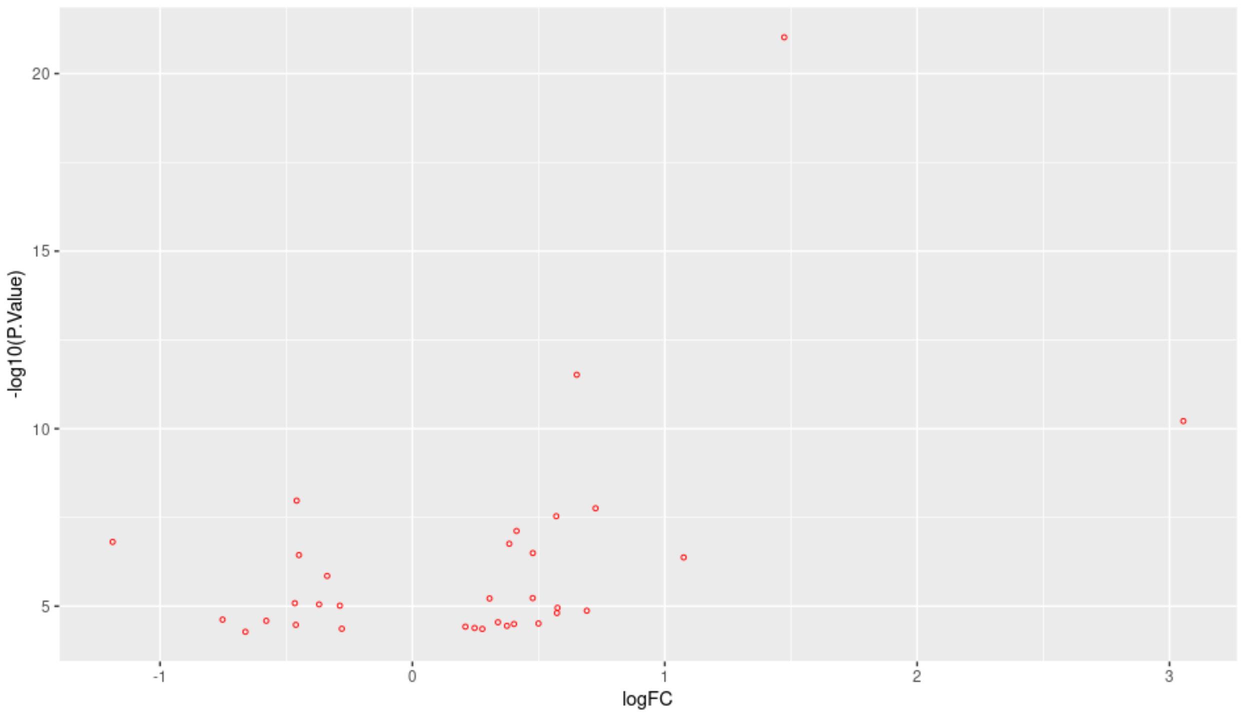

Figure 15 scatter plots of log fold change vs nominal p-value from the significant DE genes in the 3-Methylcholanthrene treatment group

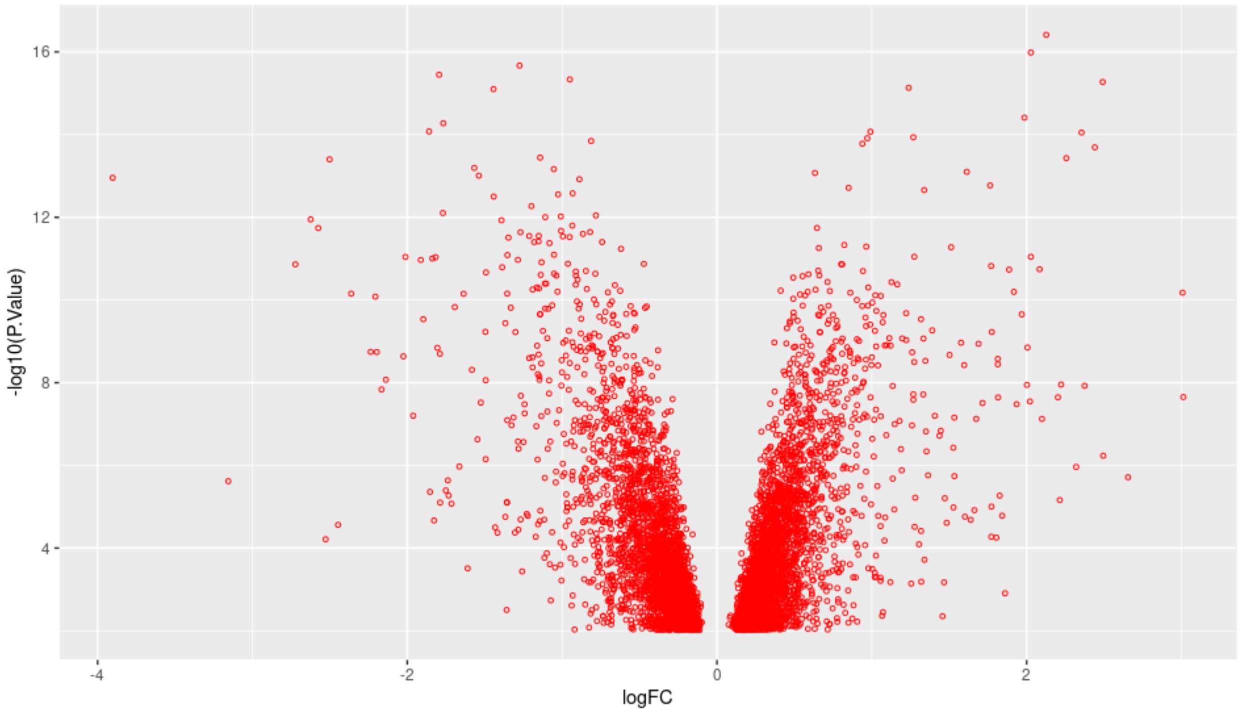

Figure 16 scatter plots of log fold change vs nominal p-value from the significant DE genes in the 3-Methylcholanthrene treatment group

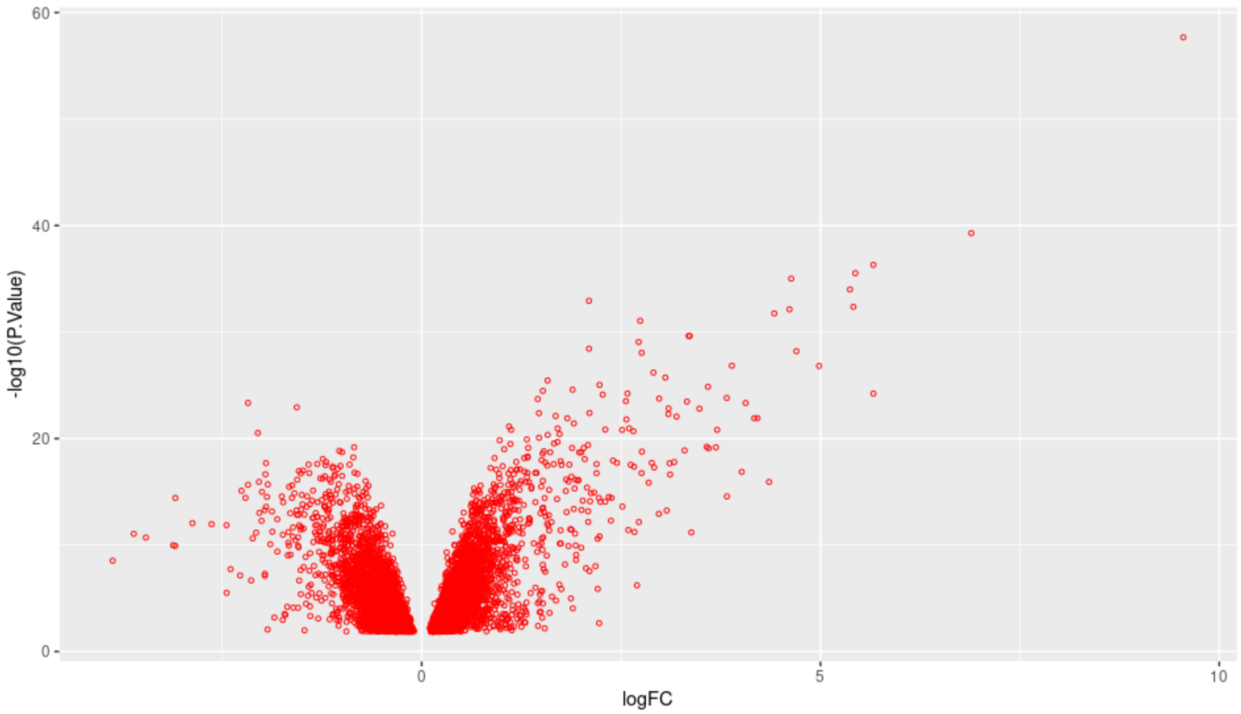

Figure 17 scatter plots of log fold change vs nominal p-value from the significant DE genes in the Pirinixic Acid treatment group

Figure 18 is the plot of concordance vs DE genes number for all 3 chemicals with regression added.Figure is for the microarray while figure 19 is for RNA-seq Both plots show the same trend that with higher number of DE genes found, there will be higher concordance Also the R^2 are very high (0.89 and 0.96) for both microarray analysis and RNA-seq analysis.

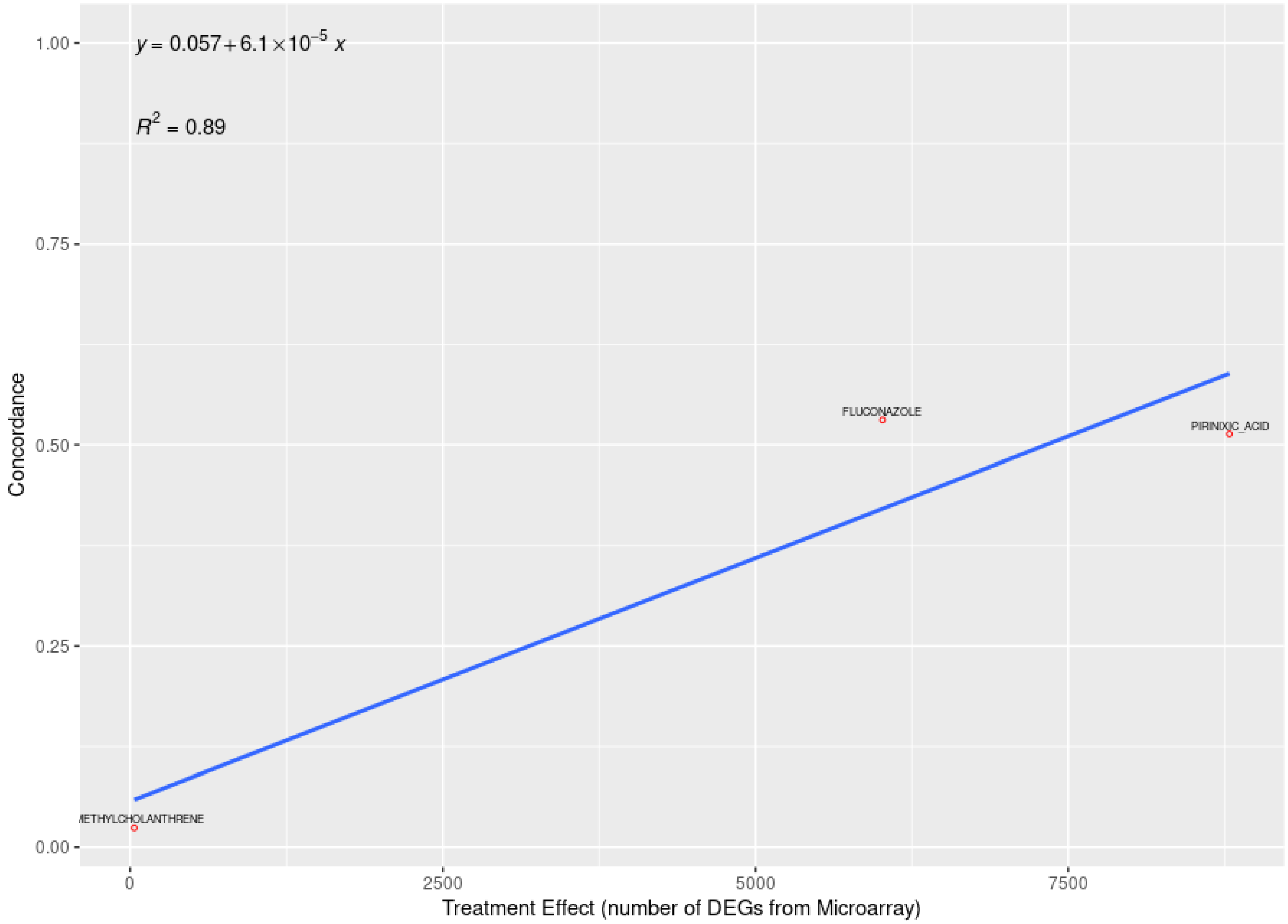

Figure 18 overall concordance vs number of DE genes from each Microarray analysis with regression. From left to right, each point represents

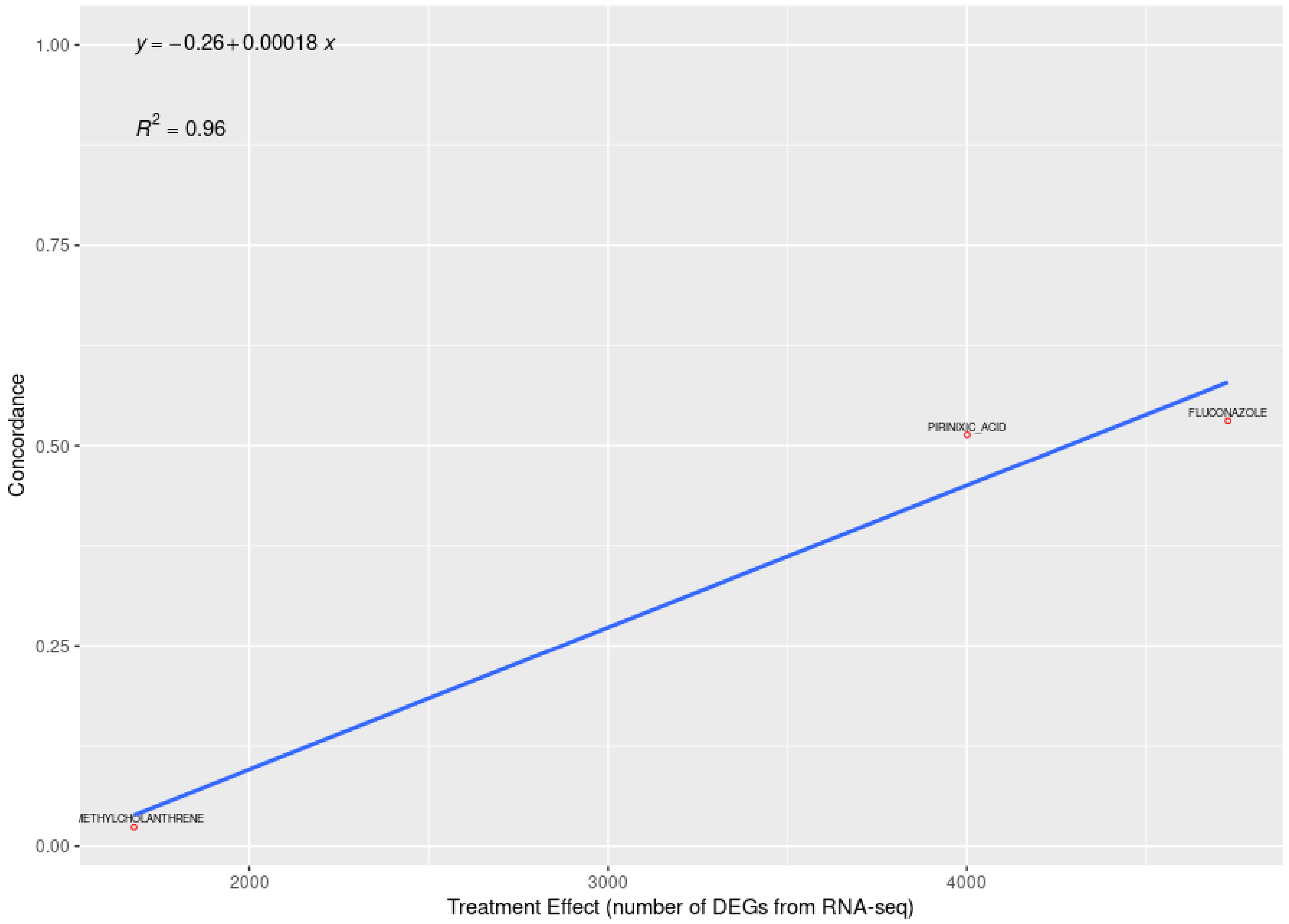

Figure 19 overall concordance vs number of DE genes from each RNA-seq analysis with regression

The combined bar plot of overall, above-, and below-median genes for each of the 3 chemicals (Figure 20) show that above-median concordance is similar to overall concordance while below-median concordance is way lower than those two. This observation applies to all 3 chemicals. Additionally, when effect size is really small /DEGs number found is very low, the concordance is very low for all groups (overall, above and below median).

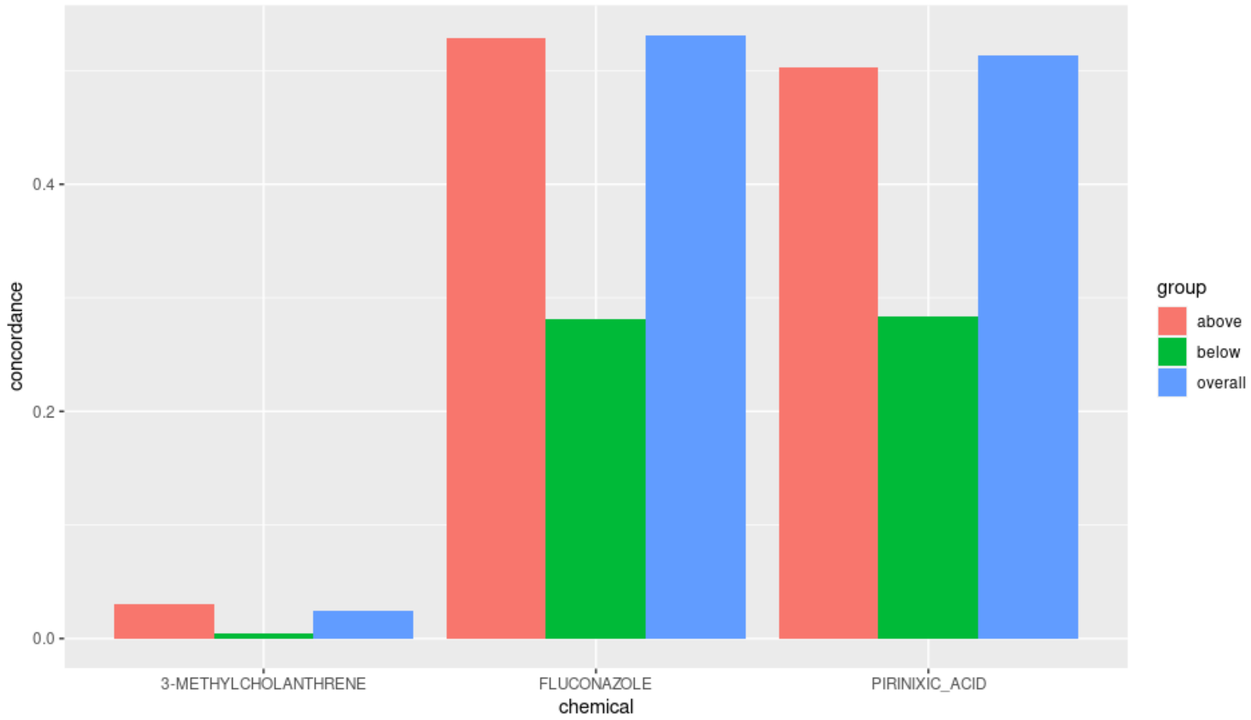

Figure 20 combined bar plot of overall, above-, and below-median genes for each of the 3 chemicals

The resulting shared enriched pathways, as seen in Table 5, were surprisingly dissimilar to the ones found in the original paper under Supplementary Table 4. KEGG pathway analysis showed that only shared enriched pathways found for PPARA were similar in any way, with AhR and CAR/PXR pathways sharing no similarities at all. It is also interesting to note that the KEGG pathways found using DAVID contained broad terminology, such as cell cycle, citric cycle, and biosynthesis, which may contain the more specific pathways reported in the original paper under its umbrella. The differences between the pathways reported in this analysis compared to that of the original paper may be due to differences in methodology, especially since differentially expressed gene analysis was performed differently in the original paper, which may have resulted in a different set of genes for pathway analysis and thus different enriched pathways. The original paper also provides enriched pathways for more chemical agents/MoAs while this analysis only focuses on the ones mentioned above.

The result from hierarchical clustering was a heatmap correctly clustering a total of fifteen samples to each respective MoA/Control. As seen in Figure 21, each of the three samples from their respective MoA's were clustered together correctly, while the control samples were identified as well. Removing the samples with high coefficients of variation, as well as low average counts were essential to producing the correct clustering results, as before data filtering some of the samples from the AhR and PPARA MoAs were incorrectly clustered with the control samples. Data filtering also removed approximately seven thousand genes, with the original matrix containing a little over eleven thousand and the resulting matrix after filtering containing a little over four thousand.

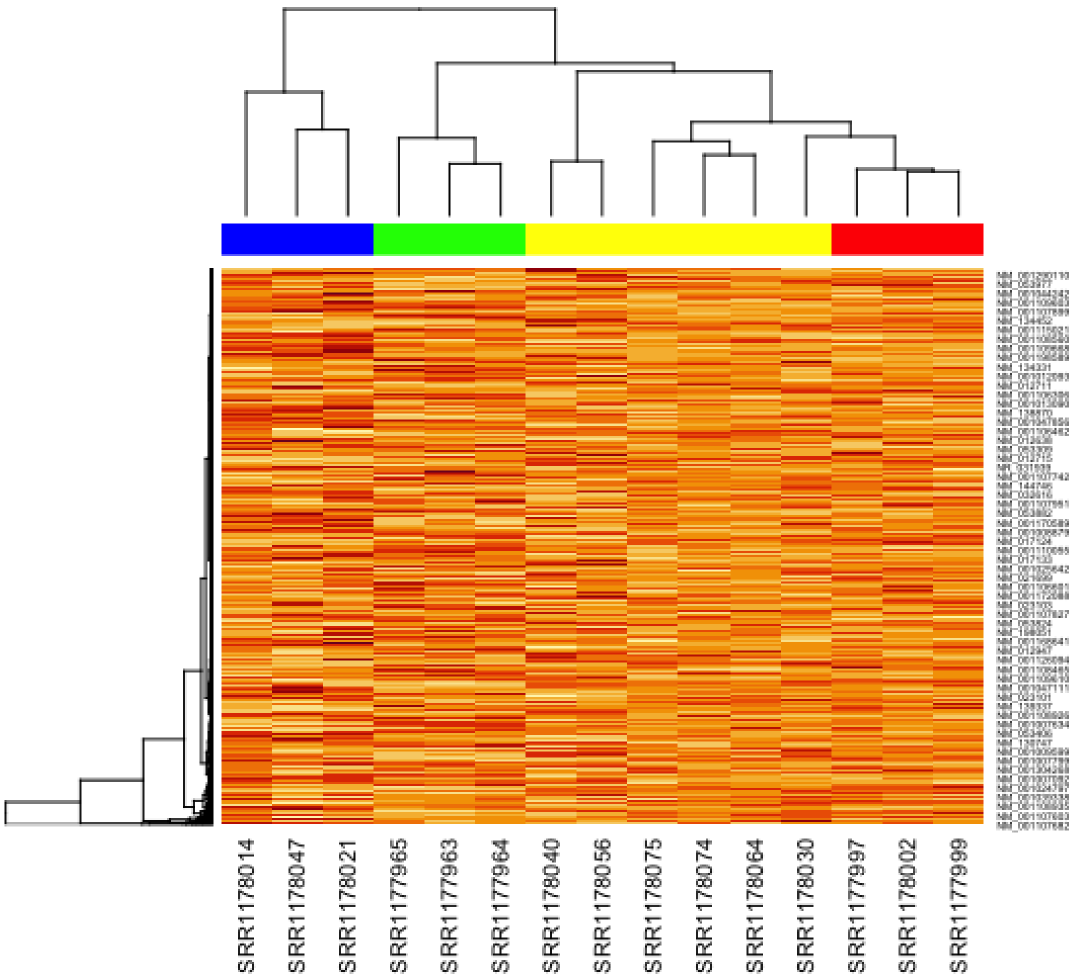

Figure 21 This heatmap shows samples which are correctly clustered by either their respective MoA/Chemical or as a control, shown by the color bar above the heatmap. Blue represents AhR/3-Methylcholanthrene samples, green represents CAR/PXR/Fluconazole samples, red represents PPARA/Pirinixic Acid samples, and yellow represents control samples. The sample names comprise the x-axis while gene names are on the y-axis.

MoA	**Common Pathways**
AhR	Peroxisome Fatty acid elongation Fatty acid metabolism Alcoholic liver disease Valine, leucine and isoleucine degradation PPAR signaling pathway Lysine degradation Biosynthesis of cofactors Oxidative phosphorylation Pantothenate and CoA biosynthesis Citrate cycle (TCA cycle) Chemical carcinogenesis - reactive oxygen species Alanine, aspartate and glutamate metabolism Metabolic pathways AMPK signaling pathway Huntington disease Prion disease Propanoate metabolism Parkinson disease Cysteine and methionine metabolism Fatty acid degradation Carbon metabolism Protein processing in endoplasmic reticulum p53 signaling pathway Hepatitis C Non-alcoholic fatty liver disease Glycine, serine and threonine metabolism Apoptosis Diabetic cardiomyopathy Lysosome Alzheimer disease Insulin resistance Arginine and proline metabolism Drug metabolism - other enzymes Tryptophan metabolism Ubiquinone and other terpenoid-quinone biosynthesis Biosynthesis of unsaturated fatty acids Platinum drug resistance Biosynthesis of amino acids

	Bile secretion Fluid shear stress and atherosclerosis Ferroptosis Ascorbate and aldarate metabolism Glycolysis / Gluconeogenesis Glutathione metabolism Pathways of neurodegeneration - multiple diseases Amyotrophic lateral sclerosis Pyruvate metabolism Thermogenesis beta-Alanine metabolism Glyoxylate and dicarboxylate metabolism
PPARA	Peroxisome Fatty acid elongation Coronavirus disease - COVID-19 HIF-1 signaling pathway Leukocyte transendothelial migration Adherens junction Pantothenate and CoA biosynthesis Oxidative phosphorylation Chemical carcinogenesis - reactive oxygen species Cytosolic DNA-sensing pathway N-Glycan biosynthesis AMPK signaling pathway Pertussis Diabetic cardiomyopathy Th17 cell differentiation Epstein-Barr virus infection Insulin resistance Aldosterone-regulated sodium reabsorption Inflammatory mediator regulation of TRP channels Drug metabolism - other enzymes Ubiquinone and other terpenoid-quinone biosynthesis Ribosome Pathways of neurodegeneration - multiple diseases Pathways in cancer Thermogenesis NOD-like receptor signaling pathway Natural killer cell mediated cytotoxicity Alcoholic liver disease Glucagon signaling pathway Pyrimidine metabolism Human T-cell leukemia virus 1 infection Purine metabolism Propanoate metabolism Progesterone-mediated oocyte maturation Parkinson disease Fatty acid degradation Fc gamma R-mediated phagocytosis

	Thyroid hormone signaling pathway Non-alcoholic fatty liver disease Proteasome Various types of N-glycan biosynthesis Biosynthesis of unsaturated fatty acids Biosynthesis of amino acids Platinum drug resistance Leishmaniasis Lipid and atherosclerosis Bile secretion Fluid shear stress and atherosclerosis 2-Oxocarboxylic acid metabolism AGE-RAGE signaling pathway in diabetic complications Amyotrophic lateral sclerosis Choline metabolism in cancer Fatty acid metabolism Tuberculosis Biosynthesis of cofactors Citrate cycle (TCA cycle) Metabolic pathways Huntington disease Amoebiasis Measles Rheumatoid arthritis Lysosome Alzheimer disease Osteoclast differentiation Phagosome Rap1 signaling pathway Central carbon metabolism in cancer Ferroptosis Ascorbate and aldarate metabolism Regulation of actin cytoskeleton Ribosome biogenesis in eukaryotes Cellular senescence Valine, leucine and isoleucine degradation PPAR signaling pathway Cholesterol metabolism Human cytomegalovirus infection Prion disease Cell cycle Carbon metabolism Platelet activation Proteoglycans in cancer Glycine, serine and threonine metabolism Apoptosis Porphyrin metabolism Influenza A FoxO signaling pathway Fatty acid biosynthesis Tryptophan metabolism

	Yersinia infection Legionellosis Spinocerebellar ataxia Pyruvate metabolism Complement and coagulation cascades beta-Alanine metabolism
CAR/PXR	Peroxisome Coronavirus disease - COVID-19 Retinol metabolism MAPK signaling pathway Cytosolic DNA-sensing pathway N-Glycan biosynthesis AMPK signaling pathway Pertussis Sphingolipid signaling pathway Glycerophospholipid metabolism Thyroid cancer cGMP-PKG signaling pathway Chemical carcinogenesis - DNA adducts Growth hormone synthesis, secretion and action Insulin resistance Arginine and proline metabolism Malaria Drug metabolism - other enzymes Autophagy - animal Ribosome Pathways in cancer Protein export Mitophagy - animal Alcoholic liver disease Alanine, aspartate and glutamate metabolism Purine metabolism Pyrimidine metabolism Non-small cell lung cancer Fatty acid degradation Nucleotide excision repair Various types of N-glycan biosynthesis Biosynthesis of amino acids Biosynthesis of unsaturated fatty acids Bile secretion Lipid and atherosclerosis Steroid hormone biosynthesis AGE-RAGE signaling pathway in diabetic complications Toxoplasmosis Choline metabolism in cancer Sulfur metabolism Fatty acid metabolism Biosynthesis of cofactors NF-kappa B signaling pathway

	Citrate cycle (TCA cycle) Metabolic pathways Cysteine and methionine metabolism Rheumatoid arthritis Acute myeloid leukemia Lysosome Fructose and mannose metabolism Ferroptosis Glycolysis / Gluconeogenesis Insulin signaling pathway Glyoxylate and dicarboxylate metabolism RNA polymerase Ribosome biogenesis in eukaryotes Cellular senescence Valine, leucine and isoleucine degradation PPAR signaling pathway Cholesterol metabolism Cell cycle Carbon metabolism Pancreatic cancer Adipocytokine signaling pathway Nucleocytoplasmic transport Protein processing in endoplasmic reticulum One carbon pool by folate p53 signaling pathway Proteoglycans in cancer Glycine, serine and threonine metabolism Apoptosis FoxO signaling pathway ABC transporters Bacterial invasion of epithelial cells Legionellosis Yersinia infection Glutathione metabolism Complement and coagulation cascades

Table 5 This table reports the shared enriched pathways for each MOA chemical group found with DAVID for differentially expressed genes found with DESeq2 and Limma. The paired MoA/Chemicals are AhR with 3-METHYLCHOLANTHRENE, CAR/PXR with FLUCONAZOLE, and PPARA with PIRINIXIC ACID.

Discussion

The paper had a similar DEGs number range for the 3 chemicals, especially for Fluconazole and Pirinixic Acid. Though, the DEGs number of microarray results and RNA-seq results in the 3-Methylcholanthrene treatment group was found to be extremely large in this project (33 vs 1679). It could be due to the low treatment effect/ low abundance causing high randomness. Also the fact that the Microarray data had different vehicle settings for the control group, which also did not get included in the design of the limma analysis design, could contribute to the huge difference in DEGs number.

The main finding from this project is that the concordance between platforms is dependent upon effect size and expression level, which is the same as the paper. As shown in the two plots of overall concordance vs number of DE genes, the higher the treatment effect (represented by the number of DEGs), the higher the concordance is. The high R^2 values also indicate that this is the main factor affecting the concordance. Looking at the barcharts, it can also be seen that the 3-Methylcholanthrene treatment group (lowest treatment effect group of the 3) has very low concordance, no matter the expression level. Maybe the huge difference between the DEGs number of microarray results and RNA-seq results (33 vs 1679) itself can lead to low concordance, but this also could indicate that low abundance contributes to discrepancies, which is also a finding of the paper.

Still, expression level is also an important factor. As can be seen in the bar chart, the below-median group has significantly lower concordance than other 2 groups while the overall groups have similar concordance to above-median groups. And effect size enlarges the difference more, as for the 3-Methylcholanthrene treatment group, the below-median concordance is only about 15% of above-median concordance, while for the other 2 chemicals, the reduction is about 50%.

The premise of the study, that similar MoAs will induce similar gene expression responses, is one that is reasonable in broad terms, but likely not true in certain cases. While it stands to reason that similar MoAs will likely elicit similar gene expression patterns in organisms, subtle differences in the chemical’s makeup will likely cause a number of genes to be expressed differently, which may have significant physiological effects. Furthermore, subtle differences in the chemical’s MoA may result in significant differences in gene expression if those manifest in biological pathways which cascade downstream to a large number of other pathways. In conclusion, while for the majority of MoAs this premise should be acceptable, this is likely not true for all MoAs.

Conclusion

Overall, this project successfully replicates the main finding that concordance between platforms is dependent upon effect size and expression level. And low abundance contributing to discrepancies is also an observation in this project. In other words, differences between microarray and RNA-seq would be larger due to larger random error with low treatment effect or expression level. So, it is suggested that apart from situations where treatment effect and expression level are low, either microarray or RNA-seq is suitable and either DESeq or Limma is suitable and can provide similar results as well.

As both this analysis and the original paper have found that low treatment effect or low expression levels result in significant differences in expression level between microarray and RNA-seq, the biological implication for this would be that one methodology is significantly less sensitive than the other. Differences in the way the two methodologies analyze biological material and biological outputs would be the key reason for this discrepancy for instrument sensitivity.

The clinical implications of these results would be that these methodologies are not yet ready to produce actionable results in a clinical setting for samples which do not fulfill the optimal prerequisites. To use these methods on samples with low treatment effect or low expression levels would result in an inconclusive study due to the errors thrown by both methodologies.

While the replication of the results of the original paper were largely successful, the enriched pathway analysis did not match the ones presented in the original. As mentioned before, this was likely due to a difference in methodology in this analysis, as the paper may have queried different publicly available pathway annotations or used different software to perform enriched pathway analysis. These differences, however, are not detrimental as the rest of the results were largely consistent and matched the original paper.

References

1. Mackenzie, R. (n.d.). RNA-seq: Basics, applications and Protocol. Genomics Research from Technology Networks. Retrieved April 10, 2022, from https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461#:~:text=RNA%2Dseq%20can%20tell%20us,are%20activated%20or%20shut%20off.&amp;text=This%20allows%20scientists%20to%20understand,changes%20that%20may%20indicate%20disease.
2. 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
3. Carol J. Bult, Janan T. Eppig, Judith A. Blake, James A. Kadin, Joel E. Richardson, the Mouse Genome Database Group, Mouse genome database 2016, *Nucleic Acids Research*, Volume 44, Issue D1, 4 January 2016, Pages D840–D847, https://doi.org/10.1093/nar/gkv1211
4. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, Imamichi T, Chang W. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. 2022 Mar 23:gkac194. doi: 10.1093/nar/gkac194. Epub ahead of print. PMID: 35325185.
5. Dobin, A. (n.d.). *STAR/STARmanual.pdf at master · alexdobin/star · GITHUB*. GitHub. Retrieved April 11, 2022, from https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf
6. *DESEQ2*. Bioconductor. (n.d.). Retrieved April 11, 2022, from https://bioconductor.org/packages/release/bioc/html/DESeq2.html
7. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47–e47.
8. Meeta Mistry, R. K. (2017, April 26). *Count normalization with deseq2*. Introduction to DGE - ARCHIVED. Retrieved April 11, 2022, from https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

9. Ewels, P. (2016, June). *Multiqc*. MultiQC. Retrieved April 11, 2022, from https://multiqc.info/