BF528 Project 3: Group Frizzled

Zhuorui Sun - *Data Curator*
Camilla Belamarich -*Programmer*
Janvee Patel - *Analyst*
Yashrajsinh Jadeja - *Biologist*

Concordance of microarray and RNA-Seq differential gene expression

Introduction

Microarrays have been synonymous with gene expression studies over the past years as they were considered the preferred method of choice at the time. They provide a robust way to quantify gene expression and to measure differences in gene expression between different samples. However, due to certain limitations and the advent of next-generation sequencing technologies and increasing accessibility due to the costs lowering significantly each year, they are seeing a significant decrease in utilization in modern studies. One of the biggest limitations of microarrays is the limited amount of genes that can be tested at once depending on the size of the microarray slide. This problem is alleviated by sequencing technologies where there are no limitations based on probes and almost all mRNA content in the sample has a chance of being quantified as a result. This results in more accurate data with a higher degree of confidence as increased coverage could significantly increase the chances of a read being determined correctly.[1]

This study by Wang et. al (2014) aims to determine the concordance between microarrays and RNA-seq differential gene expression by comparing the results from both techniques. The study primarily tests the differences between the two methods by comparing the results from both methods in differential gene expression and creating predictive classifiers. The group devised a comprehensive study design to generate Illumina RNA-seq and Affymetrix microarray data from the same group of liver samples of rats under varying degrees of perturbation by 27 chemicals that represented multiple modes of action (MOA). The cross-platform concordance in terms of differentially expressed genes (DEGs) or enriched pathways was found to be highly correlated with treatment effect size, gene-expression abundance and the biological complexity of the MOA. It was observed in the study that RNA-seq outperformed microarray (90% vs. 76%) in DEG verification by quantitative PCR. This was primarily due to the improved accuracy of RNA-seq for detecting low expressed genes. However, predictive classifiers derived from both platforms were observed to perform similarly. The study's experimental design also facilitated the identification of positive correlations in differentially expressed RNA elements (mRNA, splice variants, non-coding RNA and exon-exon junction) with the extensive perturbation elicited by the treatment, and the examination of treatment-induced alternative splicing and shortening of 3' untranslated regions (UTRs). Therefore, it was concluded by the study that the biological complexity, transcript abundance, and intended application were important factors in transcriptomic research and for decision-making.[2]

Our goal for this project is to reanalyze the data from the original study by Wang et. al. (2014) and to replicate some of the results that are presented in the paper. We focus on analyzing a specific toxgroup consisting of samples that were treated with chemicals belonging to three different modes of action (MoA) named AhR (aryl hydrocarbon receptor), CAR/PXR (orphan nuclear hormone receptors) and Cytotoxic. The chemicals that were used as treatment that we have in the toxgroup we analyze are: 3-methylcholoanthrene, clotrimazole and chloroform. 3-methylcholoanthrene is an organic cyclic

compound that is highly carcinogenic in nature, clotrimazole is a common anti-fungal medication for humans and chloroform is an organic chemical solvent that has shown to have carcinogenic properties. The primary objective of this study is to analyze the concordance between RNA-seq and microarrays and to compare it to the results detailed in the original paper.

Data

Data Description

The study by Wang et. al.[2] used liver samples of rats under varying degrees of perturbation by 27 chemicals that represented multiple modes of action (MOA). They generated RNA-seq data by Illumina RNA-seq and microarray data by Affymetrix microarray for each sample. In this project, there are 6 tox groups and we only focused on one tox group to test three different MOA. We selected the tox group one with fifteen samples (six controls and nine samples). The nine samples were categorized into three different MOA, AhR, CAR/PXR and Cytotoxic. Each MOA repeated three times with the same condition of chemical, vehicle and route.

RNA-seq data was performed with the Illumina HiSeq2000 system using the Illumina TruSeq RNA Sample Preparation Kit and SBS Kit v3 (San Diego, CA) at the depths of about 23 – 25 million paired-end 100 bp reads. The row data can be downloaded from NCBI through the ascension number SPR039021.[5] The RNA information for tox group one is shown in Table 1. We have 30 paired fastq files, 2 paired fastq files for each sample (12 fastq files for controls and other 18 fastq files for the three different MOA).

The Microarray data was performed by Affymetrix Rat Genome 230 2.0 Array and downloaded from GEO using ascension number GSE55347, GSE47875 (Wang et. al repository to NCBI GEO).[6] In this project, the microarray data have already been processed and we didn't need to do the pre-processing or quality control for the microarray data.

Table 1. RNA Information for tox group 1. The information of the tox group one contained the sample name, mode of action(MOA), chemical, vehicle and route. Tox group one contained three MOA: AhR with 3-METHYLCHOLANTHRENE, CAR/PXR with CLOTRIMAZOLE, and Cytotoxic with CHLOROFORM.

Run	Mode of Action	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
SRR1178020	CAR/PXR	CLOTRIMAZOLE	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178036	CAR/PXR	CLOTRIMAZOLE	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178046	CAR/PXR	CLOTRIMAZOLE	CORN_OIL_100_%	ORAL_GAVAGE
SRR1177987	Cytotoxic	CHLOROFORM	CORN_OIL_100_%	ORAL_GAVAGE
SRR1177988	Cytotoxic	CHLOROFORM	CORN_OIL_100_%	ORAL_GAVAGE
SRR1177989	Cytotoxic	CHLOROFORM	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178000	Control	Vehicle	CMC_.5_%	ORAL_GAVAGE

SRR1178005	Control	Vehicle	CMC_.5_%	ORAL_GAVAGE
SRR1178007	Control	Vehicle	CMC_.5_%	ORAL_GAVAGE
SRR1177973	Control	Vehicle	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178016	Control	Vehicle	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178019	Control	Vehicle	CORN_OIL_100_%	ORAL_GAVAGE

Data Quality Control

In this project, we just need to process the RNA-seq data in one tox group mentioned above. We have been given a prepared dataset as 30 paired fastq files. For each sample, we had 2 fastq.gz files such as SRR1177997_1.fastq.gz and SRR1177997_2.fastq.gz. We used the FastQC package to run fastqc first on the command line and generated the fastqc report. This package is used as a quality control tool for high throughput sequence data.[3]

To compare the gene differential expression of RNA-seq to microarrays, we aligned each of the samples against the rat genome (the Rattus norvegicus Rnor_6.0 genome) using the STAR aligner which is an alignment program and designed specifically for alignment of RNA-Seq data.[4] We run STAR on command line for each sample by the reference genome index we were given. The bam files generated by STAR were used in the following MultiQC part and DEG part.

To examine results more closely, MultiQC was performed on the experimental counts files produced from featureCounts. MultiQC is a Bioinformatics tool which combines results into a single report that is easily understandable.[7] Each sample's counts file was processed using the "multiqc" feature and a table and figure was produced revealing the percentage of assigned and unassigned reads. Specifically, the unassigned reads were determined to be unassigned-ambiguous, unassigned-multi-mapped, and unassigned with no features.

Methods

Feature Counts and Read Counting

From the Subread package, the featureCounts program was used in combination with the STAR alignment reads and GTF reference annotation file. Subread is a package used for next-generation sequencing techniques. Within Subread, feautureCounts is a program that counts reads to genomic features.[8] Fifteen of the samples (nine experimental and six control) were individually processed using "feautureCounts" and a parameter within featureCounts that allows the input of multiple features. Each bam file associated to each sample was then outputted to corresponding counts files and a summary of the counts file. Each counts file was then merged into one counts matrix, which was used for DESeq analysis. These results were used as transcript abundance measurements for the comparison to the microarray analysis. The count distribution for each sample was plotted in a box plot and grouped by chemical treatment. This was generated using the ggplot2 package in R.[9]

DESeq2 of Gene Counts

In order to estimate count differences between the samples and between the experimental and control groups, DESeq2 was performed on the counts matrix generated from feature counts after merging

the control counts file. Before DESeq2 was performed, all rows containing zeros in the combined counts matrix were removed. If the rows containing zero gene counts remained, the DESeq results would have been skewed. After, the first column in the combined counts matrix, containing Refseq gene IDs, was subset to be the row names. DESeq2 is a Bioconductor package in R that estimates the variance mean dependence on high-throughput sequencing data.[10] DESeq2 was run on each treatment group of the samples and the control counts. For each treatment group and corresponding control, a list of differentially expressed genes was produced. The counts matrix, information table of samples to corresponding treatments (Table 1), and design choice were inputted into a function of DESeq2 that created an object that was applied to each sample group and generated a list of differentially expressed genes for the treatment of 3-Methylcholanthrene, Clotrimazole, and Chloroform. For optimization, vectors of each experimental treatment and control was created. DESeq2 was run to loop through each experimental group and output results by each sample and for each group. The output file for each group contained statistical measures for the corresponding Refseq gene ID. Lastly, for subsequent clustering, the normalization of counts for all samples was performed, including control, using DESeq2's normalized counts feature. This provided counts scaled by size, which was an important step to perform due to the comparison of different samples.

Statistical Analysis of DESeq2 Results

The output of each sample group contained the mean, log2 fold change, lg2 fold change standard error, Wald statistic, p-value, and adjusted p-value for each corresponding Refseq gene ID. In order to locate the statistically significant differential expressed genes, the adjusted p-values were filtered by a threshold of 0.05. Additionally, "NA" values were removed from the significant genes list. All genes below this threshold were considered statistically significant and used for subsequent analyses. For each chemical treatment group, the top ten significant DE genes were extracted for further observation. The Refseq gene IDs were converted into formal gene names using DAVID.[11] A histogram of fold change values of the significant DE genes was plotted by each treatment group, excluding control. Additionally, a volcano plot for each treatment group was plotted using the number of not significant genes and significant DE genes with positive fold change, and negative fold change values to represent the expression of genes that increased and decreased, respectively. The log fold change threshold used to determine increased and decreased expression was values above and below zero, respectively. The plotting was done using the ggplot2 package in R.[9]

Limma Analysis and Differential Expression

Affymetrix Microarray data was analyzed using the Limma package. There were 3 comparisons where 3-Methylcholanthrene, Clotrimazole, and Chloroform were each analyzed against the appropriate control samples which were determined by samples with corresponding vehicle values. Data corresponding to the above samples' Affymetrix probe ids were found within the RMA normalized expression matrix and used for Limma analysis. Using the Bioconductor Limma package, a microarray linear model was fit, and empirical Bayes statistics were computed[29]. For multiple testing process, the Benjamini-Hochberg method was used to adjust p-values. This methodology was used for the comparison of 3-Methylcholanthrene, Clotrimazole, and Chloroform each against the appropriate controls. For determining significant differentially expressed (DE) genes from the Limma results, a cutoff of Adjusted P-value < 0.05 was implemented for the downstream processes. For the top 10 differentially expressed genes in each treatment analysis, the Affymetrix probe ids were converted to official gene symbols using the DAVID database.[11] Histograms of the log fold change values of the significant DE genes were plotted

from each of the treatment analyses. For each treatment analysis, a volcano plot was generated using the metrics of log fold change versus -log10(P-value), and significant DE genes with positive and negative log fold changes were colored red and blue respectively to represent the increased and decreased gene expression. The plotting was completed using the ggplot2 package in R[9].

Concordance Analysis

For the concordance analysis, a filter of Adjusted P-value < 0.05 was implemented for each of the treatment analyses from RNA-Seq and Microarray for determining significant differentially expressed genes. A fold change > 1.5 cutoff was not implemented as the number of DE genes for 3-Methylcholanthrene was substantially low. The concordance method from Wang et al. (2014) utilized a background corrected intersection formula:

(1)

$$x = \frac{(n_0 \times N) - (n_1 \times n_2)}{(n_0 + N - n_1 - n_2)}$$

In equation 1, *N* is the number of items in the whole sets; *n1* and *n2* are the number of items in the individual sets; *n0* is the number of items in the observed intersection, and *x* is the background corrected intersection[2]. Wang et al. (2014) utilized equation 2 for computing concordance in which $DEGs_{microarray}$ represents the number of DE genes from the microarray analysis and $DEGs_{rna\text{-}seq}$ represents the number of DE genes from the RNA-Seq analysis.

(2)

$$\frac{2 \times intersect(DEGs_{microarry}, DEGs_{rna-seq})}{(DEGs_{microarry} + DEGs_{rna-seq})}$$

For each treatment, the significant DE genes from the microarray analysis were merged with the Affymetrix to RefSeq id conversion resource provided using the probe id, and then, the significant DE genes from the RNA-Seq analysis were merged using the RefSeq id. To address fold change directionality, for each of the intersected genes between the 2 platforms, if the log2 fold change values did not agree in directionality, meaning that one platform had positive log2 fold change and the second platform had negative log2 fold change, then this mapping was removed from the intersected genes list. Duplicated gene symbols within the intersected genes list in which the same RefSeq id was mapping to multiple different Affymetrix probe ids were included as this could possibly indicate that there was mapping to multiple different regions within the same gene. For the concordance calculation, *N* was set as the number of coding genes in the rat genome (Rnor_6.0) which was approximately 22,250 according to the Ensembl Annotation System; *n1* and *n2* were set as the number of significant DE genes in the microarray and RNA-Seq analyses respectively for each of the treatments, and *n0* was set as the size of the intersected genes list[30,31]. The background corrected intersection *x* was computed using these values and equation 1. The concordance was computed using equation 2, and the *x* value as the intersect value, and *n1* and *n2* as the $DEGs_{microarray}$ and $DEGs_{rna\text{-}seq}$ values. Above and below-median subsets were determined by subsetting the DE genes from microarray on the median of the Average Expression, and the DE genes from RNA-Seq on the median of the baseMean. This concordance analysis was performed on the overall, above-median, and below-median sets for all 3 treatments. Scatter plots of the overall concordance versus the number of DE genes from RNA-Seq and microarray respectively were generated,

and a grouped barplot of concordance of the overall, above-median, and below-median sets were generated for each treatment. The plotting was completed using the ggplot2 package in R[9].

In addition, a supplemental concordance analysis was performed using a different cutoff for determining significant DE genes: P-value < 0.05 and absolute fold change > 1.5. The concordance results from this analysis are shown in Supplemental Table 1.

GSEA and Hierarchical Clustering of RNA-seq Samples

Gene Set Enrichment Analysis (GSEA) was performed using DAVID for the differentially expressed genes observed in microarray and RNA-seq analyses.[11] The analyses were performed for each MoA (mode of action) groups : AhR (aryl hydrocarbon receptor), CAR/PXR (orphan nuclear hormone receptors) and Cytotoxic respectively. Differentially expressed genes were selected for analysis based on the adjusted p-values (FDR corrected) that were statistically significant at a p-value less than 0.05 and an absolute log fold-change value threshold of 1.5. The adjusted p-values account for multiple testing and the fold-change threshold selected was consistent with what was used in the Wang et. al. study. The functional annotation results from DAVID were filtered based on a p-value threshold of 0.05.[12]

Hierarchical clustering was performed on the RNA-seq normalized expression matrix obtained from DESeq2.[13] R package “gplots” was used to perform the clustering and to visualize the results by generating a heatmap of the expression matrix. ‘Euclidean’ distance method was used for hierarchical clustering and ‘complete’ method was used as the clustering algorithm that helps identify similar clusters based on the data.[14] Coefficient of variation (standard deviation divided by mean) was calculated for every gene in the dataset and the genes having a coefficient of variation greater than 0.3567 were selected for analysis. This threshold value was selected as it was the best cutoff where the hierarchical clustering of samples was ideal as each sample clustered with it’s MoA with 100% accuracy (all samples correctly clustered with one another). This helped filter the genes with low variance and to only retain the genes that gave each sample it’s unique signature. Filtering genes with low coefficients of variation made the data more variable for each sample as genes with low variance were not biologically relevant in our analysis. Genes with low variance reduce the clustering power as the expression values are similar across all samples and there is nothing inherently unique about those genes that would help differentiate the clusters. Thus, removing those low variance genes allowed the clustering power to be significantly increased which resulted in accurate clustering of the data.

Results

Data Quality Control by FastQC and STAR

We first ran FastQC on fastq files for each sample, the results of Mean Quality Scores were shown in Figure 1. The Phred score on the y-axis started from about 30, and the score of the samples were above 10 from 0 to 100bp except one sample, the SRR187036. The phred score of sample SRR187036 was less than 10 in the last 10 base pairs. For the other samples, the Phred scores were greater than 25 until the last 10bp. The quality of data decreased in the last 10bp may because of the Illumina machine.

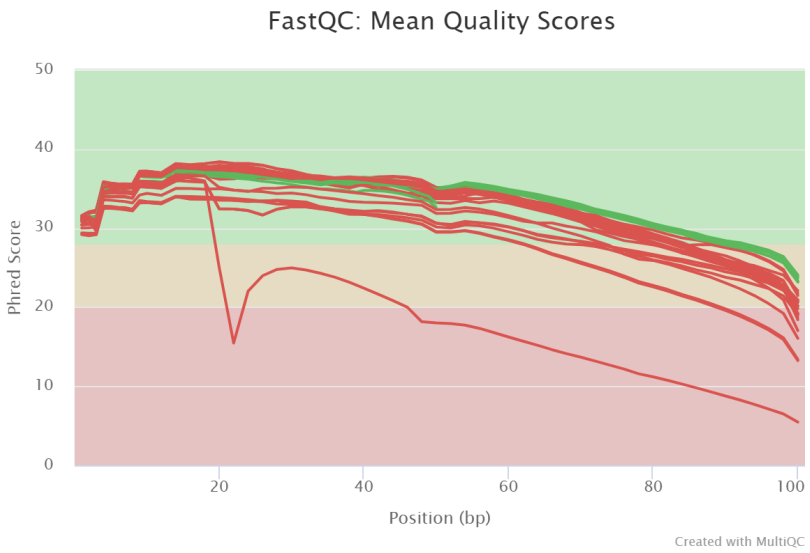

Figure 1. FastQC: Mean Quality Scores. From the result, 8 samples passed the test with the green lines in the figure and the other 22 failed. For the 22 failed samples, the 21 of them failed due to the lower score for the last 10 bp. Only one sample (SRR187036) had an obviously lower phred score compared to others.

We aligned each of the samples against the rat genome using the STAR. By running on the command line, each time we input two paired fastq files and aligned by the provided reference genome index. The alignment scores by STAR were shown in Figure 2 and the detailed counts and information are provided in Table 2. The alignment scores were around 88% for all samples. The overall quality of the data was good. Only SRR178036 had lower mapped reads for 67.8%.

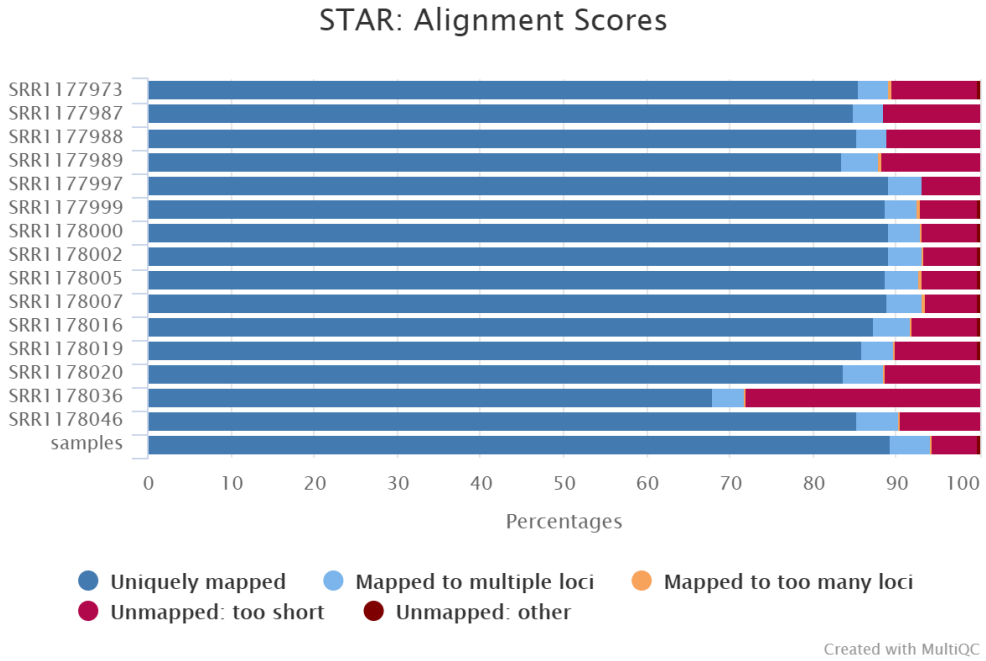

Figure 2. Alignment Scores for STAR. The mapped reads were about 89.2% and unmapped reads about 5.4% for the samples.

Table 2. Read quality statistics. (Generated by MultiQC) Including Sample name, Aligned percentage, Reads number, Uniquely mapped reads and Unmapping too short reads. Across the all samples, the mapped reads was about 89.2% and the unmapping reads was about 5,4%.

Sample Name	Aligned	M Aligned	Reads number	Uniquely mapped reads	Unmapping too short
SRR1177973	85.4%	13.3	13,315,766	85.4%	10.4%
SRR1177987	84.8%	14.8	14,785,610	84.8%	11.4%
SRR1177988	85.3%	15.7	15,667,286	85.3%	11.0%
SRR1177989	83.5%	15.7	15,707,078	83.5%	11.6%
SRR1177997	89.2%	17.6	17,608,043	89.2%	6.7%
SRR1177999	88.7%	19.4	19,374,545	88.7%	7.0%
SRR1178000	89.1%	17.1	17,104,741	89.1%	6.6%
SRR1178002	89.1%	16.8	16,796,763	89.1%	6.6%
SRR1178005	88.8%	21.1	21,088,168	88.8%	6.8%
SRR1178007	89.0%	15.1	15,084,762	89.0%	6.3%
SRR1178016	87.2%	16.4	16,418,388	87.3%	7.9%
SRR1178019	85.8%	13.7	13,689,332	85.9%	10.0%
SRR1178020	83.6%	13.4	13,374,032	83.6%	11.1%
SRR1178036	67.8%	11.4	11,443,947	67.8%	27.9%
SRR1178046	85.3%	15.1	15,067,177	85.4%	9.4%
Samples	89.2%	13.9	13,910,155	89.2%	5.4%

RNA-seq and Differential Expression Analysis Results

After FastQC and alignment with STAR, we ran MultiQCto determine the quality of the reads and generated a report based on the previous results from STAR and FastQC. The General quality statistics of the data were shown in Table 3.

Table 3. General quality statistics. General quality statistics included the percentage of duplicate reads, GC percentage, length of each read and total sequences in million for each sample. The sample SRR11778016 had length for 50bp, the other samples had 101bp. Only the SRR11778036_2 had a lower percentage of duplicate reads less than 40%.

Sample Name	%Dups	%GC	Length	M seqs
SRR1177993_1	56.7%	48%	101bp	15.6
SRR1177993_2	54.5%	49%	101bp	15.6
SRR1177987_1	56.4%	49%	101bp	17.4
SRR1177987_2	53.3%	49%	101bp	17.4
SRR1177988_1	55.1%	49%	101bp	18.4
SRR1177988_2	51.9%	49%	101bp	18.4
SRR1177989_1	50.4%	49%	101bp	18.8
SRR1177989_2	46.9%	49%	101bp	18.8
SRR1177997_1	59.6%	49%	101bp	19.7

SRR1177997_2	58.6%	49%	101bp	19.7
SRR1177999_1	60.2%	49%	101bp	21.8
SRR1177999_2	58.9%	49%	101bp	21.8
SRR1178000_1	59.0%	49%	101bp	19.2
SRR1178000_2	57.7%	49%	101bp	19.2
SRR1178002_1	58.5%	49%	101bp	18.8
SRR1178002_2	57.6%	49%	101bp	18.8
SRR1178005_1	57.8%	49%	101bp	23.8
SRR1178005_2	56.7%	49%	101bp	23.8
SRR1178007_1	56.0%	48%	101bp	17.0
SRR1178007_2	54.8%	48%	101bp	17.0
SRR1178016_1	57.8%	49%	50bp	18.8
SRR1178016_2	55.6%	49%	50bp	18.8
SRR1178019_1	55.9%	49%	101bp	15.9
SRR1178019_2	53.3%	49%	101bp	15.9
SRR1178020_1	54.0%	48%	101bp	16.0
SRR1178020_2	51.6%	49%	101bp	16.0
SRR1178036_1	55.5%	48%	101bp	16.9
SRR1178036_2	39.4%	48%	101bp	16.9
SRR1178046_1	54.7%	48%	101bp	17.7
SRR1178046_2	53.5%	49%	101bp	17.7

The MultiQC program generated basic statistics of the assigned and unassigned gene counts for each sample. Data from the featureCounts files for each sample was used to generate this plot. The percentage of assigned and unassigned gene counts is displayed in Figure 3. This plot shows generally the same amount of read counts were assigned for every sample. Specifically, the percentage of assigned counts ranged from 58.4% to 62.6%. The percent unassigned due to ambiguity was low for all samples at about 5%. The samples (SRR1178046, SRR1178036, SRR1178020) are the samples treated with Clotrimazole were all assigned/unassigned similarly. This was also consistent with the 3-Methylcholanthrene and Chloroform treatment groups. Samples and their corresponding treatment group are displayed in Table 1. The most variability with assigned/unassigned statistics was seen in the control groups.

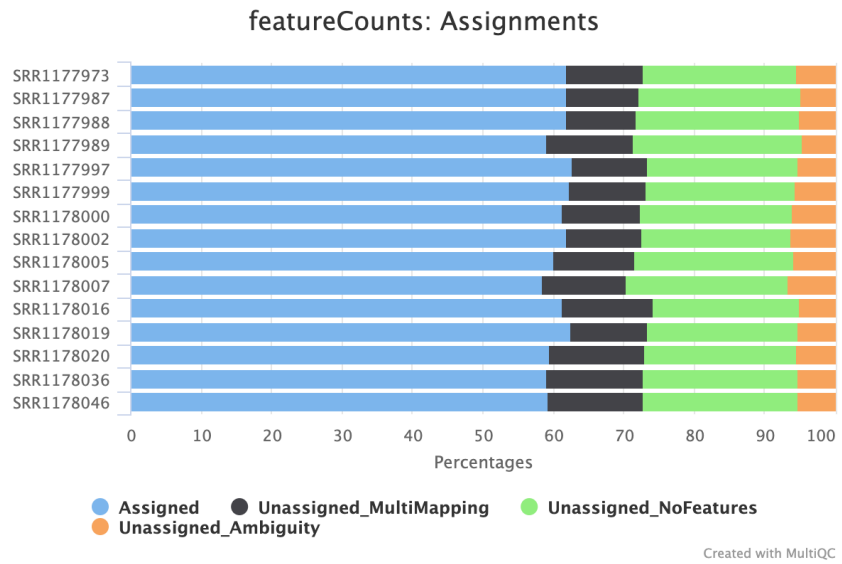

Figure 3. MultiQC Feature Counts Results. Percentage of assigned and unassigned gene counts for each sample. Both experimental and control group sample gene counts are displayed with corresponding labels indicating mapped features. Sample name is associated with a specific mode of action listed in Table 1.

From the featureCounts files, the number of gene counts was extracted and displayed in Figure 4. The distribution of counts varied between and within treatment groups (Fig. 4). However, the medians were generally the same across all samples. The number of gene counts for SRR11778036 was noticeably smaller than the other samples in the treatment group. Specifically, this sample was treated with Clotrimazole. Conversely, sample SRR1177989 treated with Chloroform has noticeably higher gene counts compared to the other samples.

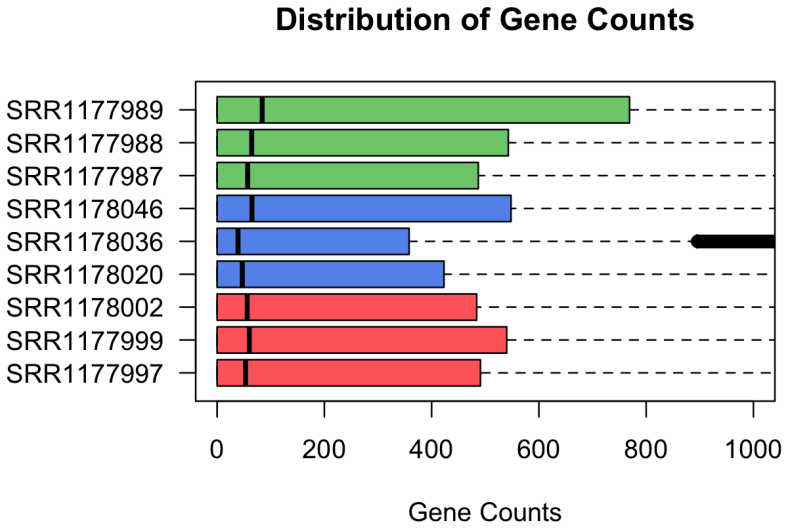

Figure 4. Distribution of Counts for each Chemical Treatment Group. The distribution of gene counts for each sample processed. Each color represents the chemical treatment used. Green) Chloroform, Blue) Clotrimazole, and Red) 3-Methylcholanthrene.

From the counts matrix, DESeq2 was performed on all the samples. The output from this analysis was a file of DE genes grouped by chemical treatment (3-Methylcholanthrene, Chloroform, and Clotrimazole) with statistical measurements corresponding to each gene ID. In order to extract and observe the significant differentially expressed genes, The results for each group were filtered by adjusted

p-value, and the genes with an adjusted p-value smaller than the 0.05 significance level were extracted and counted for each chemical group (Table 4). Chloroform reported the highest number of significant DE genes, while 3-Methylcholanthrene reported the lowest.

Table 4. Total Number of Differentially Expressed Genes at Adjusted P-value < 0.05 for the 3 Treatments from Differential Expression Analysis

Treatment	Number of Differentially Expressed Genes (Adjusted P-value < 0.05)
3-Methylcholanthrene	313
Clotrimazole	930
Chloroform	1,728

The top ten significant DE genes from each chemical treatment group were extracted (Table 5-7). Each Refseq gene ID was converted into formal gene names with DAVID.[11] The adjusted p-value for all top ten significant DE genes were observed to be extremely small. In Table 7, the gene ID, ENSROG00000042543, was not recognized when inserted into DAVID for proper conversion. Further research was necessary to better classify the gene. Ensembl recognized it as RGD1566134 found in *Rattus norvegicus*.[17]

Table 5. Top 10 Differentially Expressed Genes in 3-Methylcholanthrene Experimental Group

3-Methylcholanthrene			
Gene Name	Log2 Fold Change	P-value	Adjusted P-value
CYP1A2	3.0943	1.4166E-83	1.5173E-79
UGT1A7	2.0417	7.4642E-23	3.9975E-19
OAT	1.4049	2.1305E-14	7.6067E-11
RN45S	1.2220	3.1903E-13	8.5429E-10
ADH7	-1.3045	7.8547E-13	1.6826E-09
HSP90AA1	-1.1638	1.9945E-12	3.5606E-09
MME	-1.4715	5.7825E-12	8.8480E-09
HSPA8	-1.2212	8.3885E-12	1.1231E-08
DUSP6	-1.0224	1.0063E-11	1.1976E-08
SLC13A3	-1.4182	2.1499E-11	2.3028E-08

Table 6. Top 10 Differentially Expressed Genes in Clotrimazole Experimental Group

Clotrimazole			
Gene Name	Log2 Fold Change	P-value	Adjusted P-value
GSTA5	2.6956	2.6279E-118	2.8166E-114
ABCC3	3.8115	1.6105E-113	8.6305E-110
CYP3A23/3A1	4.091	1.2367E-107	4.4183E-104
SULT2A1	-3.0656	1.1086E-70	2.9704E-67
UGT2B1	2.6268	7.7104E-69	1.6528E-65
EPHX1	2.1172	9.4878E-68	1.6948E-64

ALDH1A7	3.3906	1.7002E-66	2.6033E-63
AKR7A3	2.5221	5.0898E-61	6.8191E-58
CYP2B1	2.0606	3.4930E-55	4.1597E-52
CES2C	3.5050	2.0650E-52	2.2133E-49

Table 7. Top 10 Differentially Expressed Genes in Chloroform Experimental Group

Chloroform			
Gene Name	**Log2 Fold Change**	**P-value**	**Adjusted P-value**
ENSROG00000042543 (RGD1566134)	-7.0160	2.6171E-132	2.8576E-128
LOC100360095	-6.7132	4.9737E-100	2.7154E-96
ABCC3	4.1987	3.1423E-49	1.1437E-45
PER3	3.6536	9.4268E-44	2.5733E-40
EPHX1	1.9617	5.6279E-40	1.2290E-36
AKR7A3	3.6238	1.1958E-39	2.1762E-36
CORO6	2.8712	8.8874E-38	1.3863E-34
CYP1A1	3.8380	1.1134E-32	1.5197E-29
DHRS7	-3.6439	6.1209E-32	7.4260E-29
GSTA5	2.0118	8.2941E-29	9.0563E-26

Additionally from the DESeq2 results, the distribution of significant DE genes for each chemical treatment group was observed through histogram plots of log fold change against frequency (Figure 5). Up- and down-regulated significant DE genes were observed for each treatment group (Fig. 5). Genes that were larger in log fold change value above zero were observed to be up-regulated genes. Conversely, genes below the log fold change threshold of zero, were observed to be down-regulated. Both 3-Methylcholanthrene (Fig. 5A) and Clotrimazole (Fig. 5B) showed a higher number of down-regulated DE genes. However, Chloroform (Fig. 5C) showed an equal number of up- and down-regulated genes.

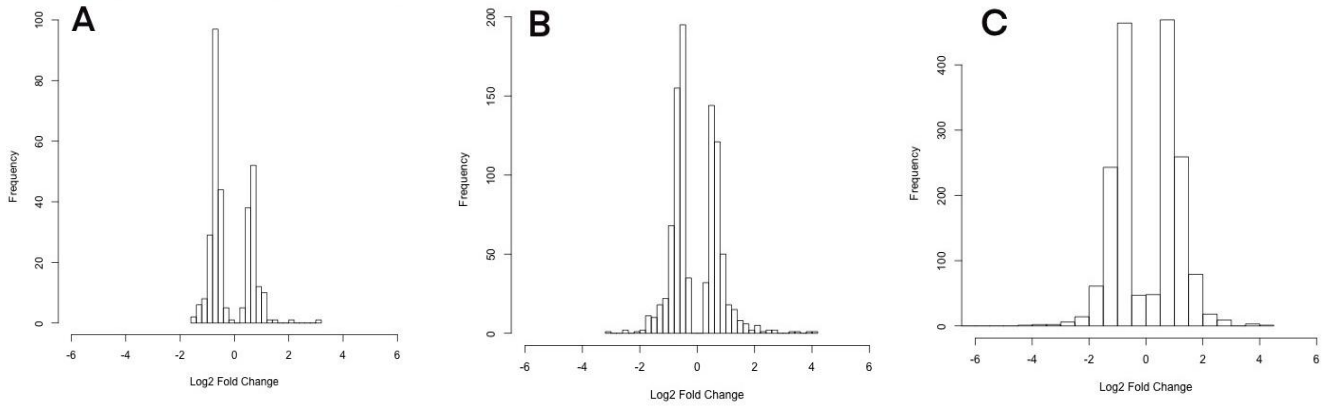

Figure 5. Distribution of Significant Differentially Expressed Genes for each Chemical. Histogram displays distribution of A) 3-Methylcholanthrene, B) Clotrimazole and C) Chloroform chemicals from DESeq analysis.

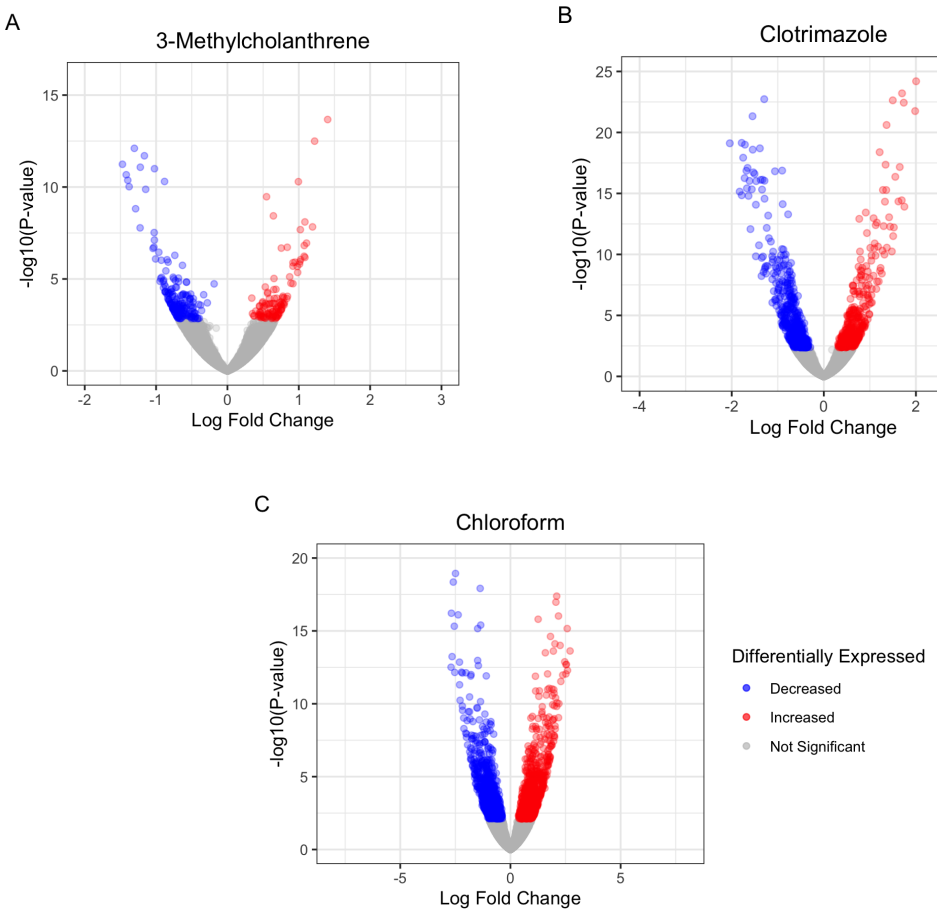

Figure 6. Volcano Plots of Log Fold Change Against -log10(P-Value) from RNA-seq Differential Expression Analysis. Volcano plots show the distribution of increased genes (red) and decreased expressed genes (blue) against -log10(p-value) results from DESeq analysis of A) 3-Methylcholanthrene, B) Clotrimazole and C) Chloroform chemicals.

In order to better visualize the distribution of genes that were not-significant, significant with increased expression, and significant with decreased expression for each treatment group, volcano plots were generated (Fig. 6). Both 3-Methylcholanthrene (Fig. 6A) and Clotrimazole (Fig. 6B) showed a wider distribution than Chloroform, which was observed to be more tightly centered around zero (Fig. 6c). 3-Methylcholanthrene showed the largest proportion of not-significantly DE genes compared to significant (Fig. 3A). Consistent with the information in Table 3, it was observed that Chloroform has the greatest number of significantly DE genes (Fig. 6C).

Microarray Analysis and Differential Expression Results

Table 8 displays the total number of differentially expressed genes determined by an adjusted P-value < 0.05 for each of the 3 treatments. A fold change > 1.5 cutoff was not implemented as the total number of DE genes for 3-Methylcholanthrene was substantially low. Chloroform had the highest number of 9,458 DE genes, while 3-Methylcholanthrene had the lowest number of 58 DE genes. Compared to the other treatments, 3-Methylcholanthrene had a smaller number of DE genes, and this could have been attributed to using an adjusted P-value filter.

Table 8. Total Number of Differentially Expressed Genes at Adjusted P-value < 0.05 for the 3 Treatments from Microarray Analysis

Treatment	Number of Differentially Expressed Genes (Adjusted P-value < 0.05)
3-Methylcholanthrene	58
Clotrimazole	2,692
Chloroform	9,458

Tables 9-11 display the top 10 differentially expressed genes from the microarray analysis for each of the treatments respectively: 3-Methylcholanthrene, Clotrimazole, and Chloroform. In Tables 9 and 10, the probe ids corresponding to UGT1A6 mapped to multiple genes within the same family UDP Glucuronosyltransferase. For this analysis, UGT1A6 was shown as the representative. In Table 9 for 3-Methylcholanthrene, 8 of the top 10 genes had positive log fold change values indicating increased expression, while 2 genes PTPRS and SLC34A2 had negative log fold changes indicating decreased expression. In Table 10 for Clotrimazole, all of the top 10 genes had increased expression. In Table 11 for Chloroform, 6 of the top 10 genes had increased expression, while 4 genes STAC3, NOX4, RGD1566134, and DHRS7I1 had decreased expression.

Table 9. Top 10 Differentially Expressed Genes in 3-Methylcholanthrene from Microarray Analysis

3-Methylcholanthrene			
Gene Name	**Log Fold Change**	**P-value**	**Adjusted P-value**
CYP1A2	1.5785	2.5407e-17	7.9013e-13
UGT1A6	0.7847	1.3471e-12	2.0946e-8
UGT1A6	0.9926	2.3588e-12	2.4452e-8
SMIM13	0.4739	2.0833e-9	1.6197e-5
PTPRS	-0.4668	5.9357e-7	3.6919e-3
GSTA4	0.4206	8.2906e-7	4.2972e-3
PON3	0.3454	9.7137e-7	4.3155e-3
SLC34A2	-1.2359	1.8964e-6	7.3719e-3
LOC684841 (CG31613-PA)	0.5384	2.6292e-6	9.0849e-3
MAP1LC3B	0.3820	3.1530e-6	9.8054e-3

Table 10. Top 10 Differentially Expressed Genes in Clotrimazole from Microarray Analysis

Clotrimazole			
Gene Name	**Log Fold Change**	**P-value**	**Adjusted P-value**
CYP2B1	2.9994	1.4016e-27	4.3588e-23
CYP3A23/3A1	1.3811	8.6932e-24	1.3518e-19
UGT2B1	1.9369	2.6281e-23	2.7244e-19
CES2C	3.2185	1.3168e-17	1.0237e-13

UGT1A6	1.0151	7.8780e-15	4.8999e-11
CYP2C6V1	0.4929	1.9960e-14	1.0346e-10
UGT1A6	0.8454	2.6815e-14	1.1913e-10
DENND2A	0.8966	4.8103e-14	1.8699e-10
ABCC3	3.0499	9.2440e-14	3.1942e-10
CYP3A18	0.7774	1.9582e-13	6.0899e-10

Table 11. Top 10 Differentially Expressed Genes in Chloroform from Microarray Analysis

Chloroform			
Gene Name	**Log Fold Change**	**P-value**	**Adjusted P-value**
ABCC3	4.1369	3.85453-21	1.1987e-16
GSTP1	4.7316	3.57703-20	4.7263e-16
AKR1B8	7.3937	4.5593e-20	4.7263e-16
STAC3	-4.6582	1.3315e-17	1.0352e-13
NOX4	-4.1783	3.7333e-17	2.3220e-13
RGD1566134	-5.1796	2.1253e-16	9.7772e-13
AKR7A3	3.1301	2.2007e-16	9.7772e-13
GPX2	3.7879	1.017e-15	3.7296e-12
DHRS7I1	-5.6555	1.0793e-15	3.7296e-12
DAP	1.5557	1.7471e-15	5.4333e-12

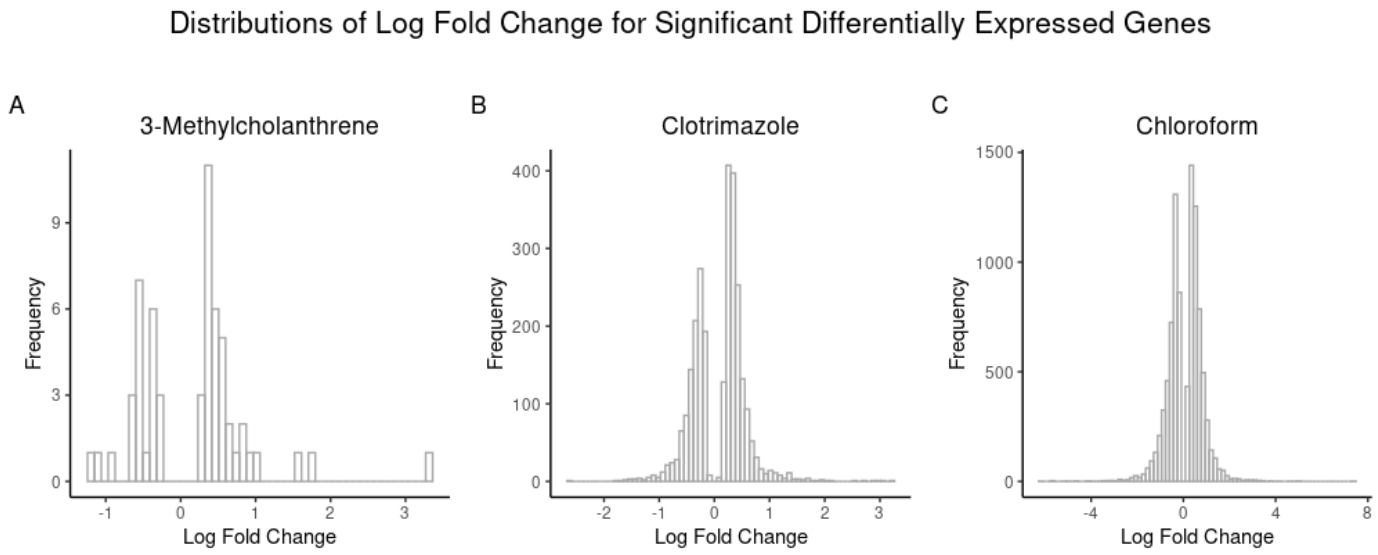

Figure 7. Distribution of Log Fold Change of Significant Differentially Expressed Genes From Microarray Analysis. The 3 treatments are shown in A) 3-Methylcholanthrene, B) Clotrimazole, C) Chloroform

Figure 7 shows the distributions of log fold change of significant differentially expressed genes for each of the 3 treatments. Genes with positive log fold change values were determined to have increased expression, while genes with negative log fold change values were determined to have decreased expression. In Figure 7A, there appears to be a greater number of up-regulated DE genes in the 3-Methylcholanthrene treatment. Figure 7B shows a greater number of up-regulated DE genes compared to down-regulated DE genes in the Clotrimazole treatment. Figure 7C shows approximately an equal number of up-regulated and down-regulated genes in the Chloroform treatment.

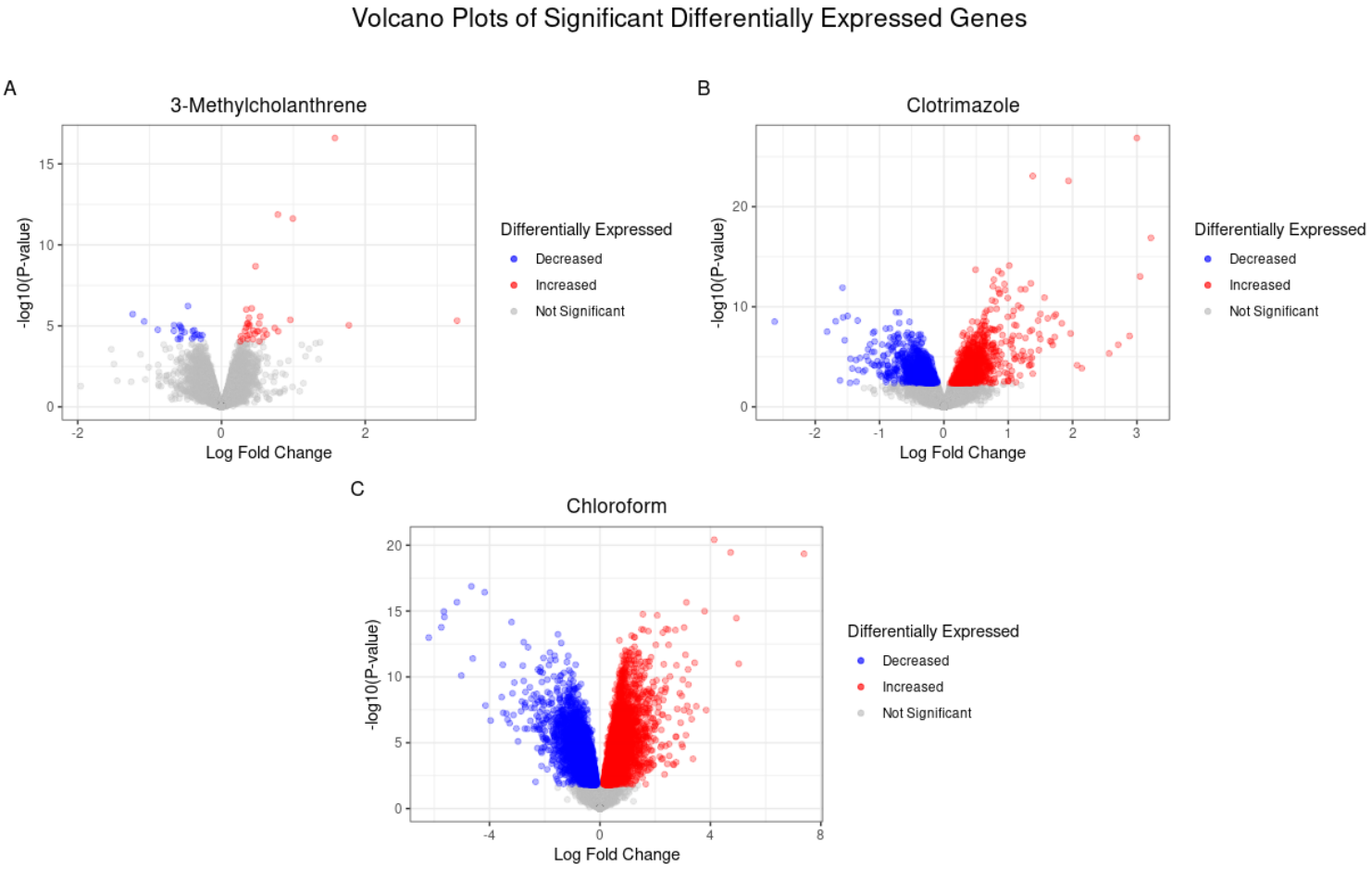

Figure 8. Volcano Plots of Log Fold Change Against -log10(P-Value) From Microarray Analysis. The 3 treatments are shown in A) 3-Methylcholanthrene, B) Clotrimazole, C) Chloroform. The color legend indicates differential expression changes. Red indicates increased expression, and blue indicates decreased expression. Grey indicates not significant genes. Significant genes were determined using Adjusted P-value < 0.05 cutoff.

Figure 8 displays volcano plots of log fold change against -log10(P-value) for genes within each of the treatments. Figure 8A shows a substantially smaller proportion of significant genes, determined by the red and blue colors, compared to not significant genes using the cutoff metric described above. This is consistent with the value of 58 for the total number of DE genes for 3-Methylcholanthrene shown in Table 8. Figure 8A also indicates some up-regulated genes shown in red with greater log fold changes that are the most statistically significant by P-value. Figure 8B shows a greater proportion of significant DE genes compared to not significant genes, and there appears to be up-regulated genes shown in red with greater log fold changes that are also the most statistically significant by P-value. Figure 8C shows approximately an equal number of up-regulated (red) and down-regulated (blue) DE genes. There also

appears to be DE genes with greater fold changes in up-regulated and down-regulated that are the most statistically significant by P-value.

Concordance Analysis Results

Tables 12-14 display the concordances of DE genes between the RNA-Seq and Microarray platforms for each of the 3 treatments.

Table 12. Percent concordance between RNA-Seq and Microarray Analyses for the 3 Treatments. Significant differentially expressed genes from the RNA-Seq and Microarray analyses were determined by adjusted P-value < 0.05. The background corrected intersection value was used as the intersect value in the concordance equation.

Overall	
Treatment	**Concordance**
3-Methylcholanthrene	**8.866%**
Clotrimazole	**28.277%**
Chloroform	**27.596%**

Overall concordance is shown in Table 12. Treatment 3-Methylcholanthrene had the smallest concordance value of approximately 8.866%, while Clotrimazole had the highest concordance value of 28.277%, and Chloroform had slightly smaller value of 27.596%.

Table 13. Percent concordance between RNA-Seq and Microarray Analyses for the 3 Treatments in the Above-Median Groups. Significant differentially expressed genes were determined by adjusted P-value < 0.05. The background corrected intersection value was used as the intersect value in the concordance equation.

Above-Median	
Treatment	**Concordance**
3-Methylcholanthrene	**8.450%**
Clotrimazole	**33.614%**
Chloroform	**29.409%**

Concordance for the above-median subsets in each of the treatment groups is shown in Table 13. The above-median subsets contained DE genes that were determined to have a value greater than the median of the Average Expression metric in microarray analysis and a value greater than the baseMean metric in RNA-Seq. Therefore, these subsets contain genes that were highly expressed. Treatment 3-Methylcholanthrene had the smallest concordance of 8.450%, while Clotrimazole had the highest concordance of 33.614%, and Chloroform had a value of 29.409%.

Table 14. Percent concordance between RNA-Seq and Microarray Analyses for the 3 Treatments in the Below-Median Groups. Significant differentially expressed genes were determined by adjusted P-value < 0.05. The background corrected intersection value was used as the intersect value in the concordance equation.

Below-Median	
Treatment	**Concordance**
3-Methylcholanthrene	**1.958%**
Clotrimazole	**12.764%**
Chloroform	**11.793%**

Concordance for the below-median subsets in each of the treatment groups is shown in Table 14. The below-median subsets contained DE genes that were determined to have a value less than the median of the Average Expression metric in microarray analysis and a value less than the baseMean metric in RNA-Seq. These subsets contain genes that had low expression in comparison to the highly expressed genes discussed above. The below-median concordance values are substantially smaller compared to the overall and above-median values. Treatment 3-Methylcholanthrene had the smallest concordance of 1.958%, while Clotrimazole had the highest concordance of 12.764%, and Chloroform had a slightly smaller value of 11.793%.

In addition, a supplemental analysis to compute concordance from the overall, above-median, and below-median sets for each of the treatments was performed using a filter of P-value < 0.05 and absolute FC > 1.5. A similar filtering metric was used in Wang et al. (2014). From Supplemental Table 1A, the overall set, the concordance for 3-Methylcholanthrene was 5.052%; Clotrimazole was 37.284%, and Chloroform was 45.646%. Except for 3-Methylcholanthrene, these concordance values were greater than the values shown in Table 12. The concordance for Chloroform was approximately similar to the corresponding value (concordance ≈45%, treatment effect ≈2,400 DE genes RNA-Seq) shown in Figure 2a in Wang et al. (2014). The concordance for 3-Methylcholanthrene was substantially smaller compared to the corresponding value (concordance ≈30%, treatment effect ≈400 DE genes RNA-Seq) shown in Figure 2a in Wang et al. (2014). The approximately similar concordance of Chloroform could be due to the use of similar filtering metrics. However, the filtering and determination of DE genes using these metrics was not exact to the paper as they implemented additional tests such as Levene's Test for testing variance[2]. From Supplemental Table 1B, the above-median set, the concordance for 3-Methylcholanthrene was 5.749%; Clotrimazole was 38.584%, and Chloroform was 50.434%. Except for 3-Methylcholanthrene, these values were also greater than the concordance values shown in Table 13. From Supplemental Table 1C, the below-median set, the concordance for 3-Methylcholanthrene was 2.774%; Clotrimazole was 15.459%, and Chloroform was 20.674%. These values were greater than the concordance values shown in Table 14, but followed a similar pattern in that the values were smaller compared to the overall and above-median sets.

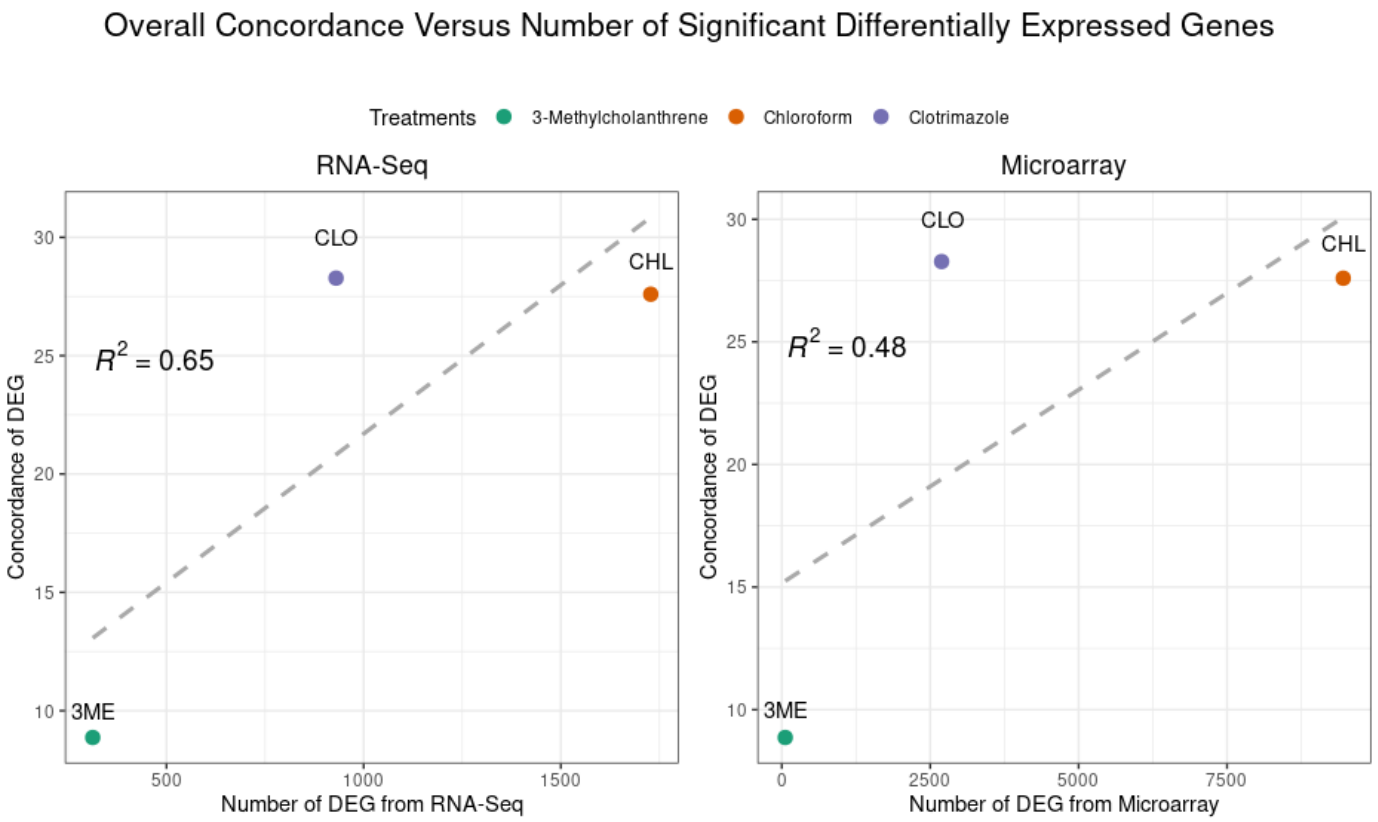

Figure 9. Number of Significant Differentially Expressed Genes Against Overall Concordance for RNA-Seq (left) and Microarray (right). Significant differentially expressed genes were determined by adjusted P-value < 0.05.

In Figure 9, the number of differentially expressed genes from RNA-Seq and the concordance of DEG for each of the 3 treatments is shown in the left RNA-Seq plot. For 3-Methylcholanthrene, the number of DEGs was 313 with a concordance of 8.866%. The corresponding value for 3-Methylcholanthrene in Figure 2a in Wang et al. (2014) had approximately 400 DEGs from RNA-Seq with a concordance of approximately 30%. For Clotrimazole, the number of DEGs was 930 with a concordance of 28.277%. A corresponding value for Clotrimazole was not shown in Figure 2a. For Chloroform, the number of DEGs was 1,728 with a concordance of 27.596%. The corresponding value for Chloroform in Figure 2a in Wang et al. (2014) had approximately 2,400 DEGs from RNA-Seq with a concordance of approximately 45%. The lower numbers of DEGs and concordance values could be attributed to the different filtering metrics used for determining significant DEGs. In addition, the linear regression line shows an upward trend in which concordance of DEGs between the two platforms increases as the number of DEGs increases. There was greater agreement of DEGs for chemicals such as Chloroform and Clotrimazole which had a greater number of DEGs compared to 3-Methylcholanthrene.

In the right Microarray plot, the number of differentially expressed genes from Microarray and the concordance for each of the 3 treatments is shown. For 3-Methylcholanthrene, there were 58 DEGs with a concordance of 8.866%. For Clotrimazole, there were 2,692 DEGs with a concordance of 28.277%. For Chloroform, there were 9,458 DEGs with a concordance of 27.596%. The corresponding values for microarray analysis were not shown as a figure in Wang et al. (2014). Similar to RNA-Seq, the linear regression line within this plot shows an upward trend where concordance of DEGs between the two platforms increases as the number of DEGs increases.

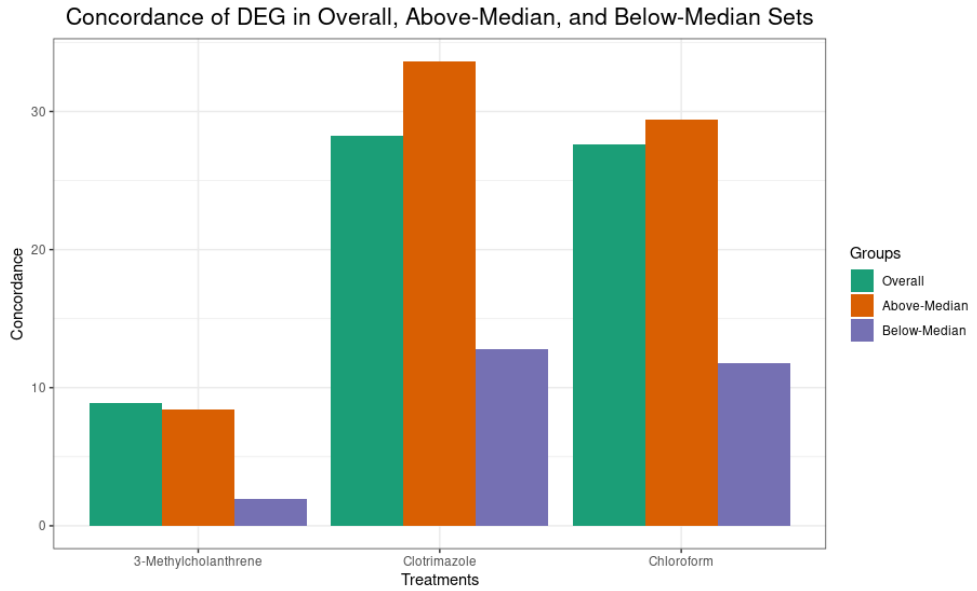

Figure 10. Bar Plot of Percent Concordance Among Overall, Above-Median, and Below Median Groups for the 3 Treatments

Figure 10 shows the concordance of DEGs in the overall, above-median, and below-median sets for each of the treatments. Clotrimazole and Chloroform showed greater overall concordance in comparison to 3-Methylcholanthrene. The above-median subset represented the more highly expressed DEGs. For 3-Methylcholanthrene, the above-median concordance value was approximately similar to the overall concordance value, while for Clotrimazole and Chloroform, the above-median values are greater than the overall. The above-median values were not as large as the results found in Wang et al. (2014), but the overall trend that agreement between the two platforms was higher for the above-median sets agreed with the results depicted here. The below-median subset represented the more lowly expressed DEGs. The below-median concordance values were comparatively less among all 3 treatments. The overall trend that there was less agreement between the two platforms for the below-median genes discussed in Wang et al. (2014) agreed with the results depicted here as well. This could also be due to the low presence of below-median DEGs that intersected between both platforms.

The results obtained from the Gene Set Enrichment Analysis (GSEA) are detailed in the tables 15-20.

Table 15. Results of GSEA from DAVID for Microarray DEGs (AhR MoA). Overlapping terms between the Wang et. al. paper and this analysis are highlighted in red boundaries and bold text.

GSEA with DAVID (Microarray DEGs) for AhR MoA				
Category	**Term**	**Term in Paper**	**PValue**	**FDR**
GOTERM_BP_DIRECT	**GO:0014070~response to organic cyclic compound**	**Aryl hydrocarbon receptor signaling**	**3.63E-05**	**0.0079**
INTERPRO	IPR008066:Cytochrome P450, E-class, group I, CYP1	-	0.0013	0.0415
UP_KEYWORDS	Microsome	-	0.0024	0.0827
GOTERM_BP_DIRECT	GO:0006778~porphyrin-containing compound metabolic process	-	0.0029	0.1743
GOTERM_MF_DIRECT	GO:0019899~enzyme binding	-	0.0030	0.1039

KEGG_PATHWAY	**rno00980:Metabolism of xenobiotics by cytochrome P450**	**Xenobiotic metabolism signaling**	**0.0041**	**0.0924**
KEGG_PATHWAY	**rno00830:Retinol metabolism**	**Retinoate biosynthesis I**	**0.0058**	**0.0924**
GOTERM_BP_DIRECT	**GO:0042493~response to drug**	**Nicotine degradation II**	**0.0061**	**0.1743**
GOTERM_BP_DIRECT	**GO:0032496~response to lipopolysaccharide**	**LPS/IL-1 mediated inhibition of RXR function**	**0.0176**	**0.2262**
GOTERM_BP_DIRECT	**GO:0006805~xenobiotic metabolic process**	**Xenobiotic metabolism signaling**	**0.0220**	**0.2400**

Table 16. Results of GSEA from DAVID for RNA-seq DEGs (AhR MoA). Overlapping terms between the Wang et. al. paper and this analysis are highlighted in red boundaries and bold text.

GSEA with DAVID (RNA Seq DEGs) for AhR MoA				
Category	**Term**	**Term in Paper**	**PValue**	**FDR**
KEGG_PATHWAY	rno05204:Chemical carcinogenesis	-	2.98E-13	5.54E-11
KEGG_PATHWAY	**rno00980:Metabolism of xenobiotics by cytochrome P450**	**Xenobiotic metabolism signaling**	**1.56E-12**	**1.45E-10**
UP_KEYWORDS	Oxidoreductase	-	1.57E-10	2.80E-08
KEGG_PATHWAY	**rno00830:Retinol metabolism**	**Retinoate biosynthesis I**	**2.62E-10**	**1.62E-08**
KEGG_PATHWAY	rno00982:Drug metabolism - cytochrome P450	-	5.31E-10	2.47E-08
UP_KEYWORDS	Microsome	-	5.89E-10	5.24E-08
GOTERM_BP_DIRECT	**GO:0042493~response to drug**	**Nicotine degradation II**	**4.04E-09**	**5.18E-06**
UP_KEYWORDS	Fatty acid metabolism	-	1.34E-08	7.94E-07
GOTERM_BP_DIRECT	**GO:0014070~response to organic cyclic compound**	**Aryl hydrocarbon receptor signaling**	**1.64E-07**	**1.06E-04**
GOTERM_BP_DIRECT	**GO:0032496~response to lipopolysaccharide**	**LPS/IL-1 mediated inhibition of RXR function**	**0.0187**	**0.6187**

Table 17. Results of GSEA from DAVID for Microarray DEGs (CAR.PXR MoA). Overlapping terms between the Wang et. al. paper and this analysis are highlighted in red boundaries and bold text.

GSEA with DAVID (Microarray DEGs) for CAR.PXR MoA				
Category	**Term**	**Term in Paper**	**PValue**	**FDR**
KEGG_PATHWAY	rno05204:Chemical carcinogenesis	-	1.65E-13	3.06E-11
KEGG_PATHWAY	**rno00830:Retinol metabolism**	**Retinoate biosynthesis I**	**6.01E-12**	**5.59E-10**
KEGG_PATHWAY	**rno00980:Metabolism of xenobiotics by cytochrome P450**	**Xenobiotic metabolism signaling**	**7.97E-11**	**4.94E-09**
UP_KEYWORDS	Endoplasmic reticulum	-	2.03E-10	4.31E-08
UP_KEYWORDS	Microsome	-	6.27E-10	6.65E-08
GOTERM_BP_DIRECT	**GO:0014070~response to organic cyclic compound**	**Aryl hydrocarbon receptor signaling**	**9.01E-08**	**7.53E-05**
GOTERM_BP_DIREC	**GO:0042493~response to drug**	**Nicotine degradation II**	**7.45E-07**	**3.11E-04**

T				
GOTERM_BP_DIRECT	**GO:0006979~response to oxidative stress**	**NRF2-mediated oxidative stress response**	**2.73E-04**	**0.0350**
GOTERM_BP_DIRECT	**GO:0032496~response to lipopolysaccharide**	**LPS/IL-1 mediated inhibition of RXR function**	**0.0207**	**0.7093**
GOTERM_BP_DIRECT process	**GO:0006749~glutathione metabolic**	**Glutathione-mediated detoxification**	**9.93E-05**	**0.0166**

Table 18. Results of GSEA from DAVID for RNA-seq DEGs (CAR.PXR MoA). Overlapping terms between the Wang et. al. paper and this analysis are highlighted in red boundaries and bold text.

GSEA with DAVID (RNA Seq DEGs) for CAR.PXR MoA				
Category	**Term**	**Term from Paper**	**PValue**	**FDR**
KEGG_PATHWAY	rno05204:Chemical carcinogenesis	-	1.76E-22	4.09E-20
UP_KEYWORDS	Microsome	-	3.34E-21	9.22E-19
KEGG_PATHWAY	rno00140:Steroid hormone biosynthesis	-	3.72E-17	4.34E-15
UP_KEYWORDS	Oxidoreductase	-	5.74E-17	7.92E-15
KEGG_PATHWAY	**rno00980:Metabolism of xenobiotics by cytochrome P450**	**Xenobiotic metabolism signaling**	**1.49E-16**	**8.69E-15**
GOTERM_BP_DIRECT	**GO:0014070~response to organic cyclic compound**	**Aryl hydrocarbon receptor signaling**	**5.64E-11**	**7.42E-08**
GOTERM_BP_DIRECT	**GO:0042493~response to drug**	**Nicotine degradation II**	**2.08E-10**	**1.82E-07**
GOTERM_BP_DIRECT	**GO:0006749~glutathione metabolic process**	**Glutathione-mediated detoxification**	**1.05E-04**	**0.0137**
GOTERM_BP_DIRECT	**GO:0006979~response to oxidative stress**	**NRF2-mediated oxidative stress response**	**2.89E-04**	**0.0266**
GOTERM_BP_DIRECT	**GO:0032496~response to lipopolysaccharide**	**LPS/IL-1 mediated inhibition of RXR function**	**0.0012**	**0.0765**

Table 19. Results of GSEA from DAVID for Microarray DEGs (Cytotoxic MoA). Overlapping terms between the Wang et. al. paper and this analysis are highlighted in red boundaries and bold text.

GSEA with DAVID (Microarray DEGs) for Cytotoxic MoA				
Category	**Term**	**Term in Paper**	**PValue**	**FDR**
UP_KEYWORDS	**Acetylation**	**Acetone Degradation I (to Methylglyoxal)**	**1.40E-82**	**4.90E-80**
GOTERM_BP_DIRECT	**GO:0042493~response to drug**	**Nicotine degradation II**	**4.12E-08**	**3.02E-05**
GOTERM_BP_DIRECT	**GO:0006979~response to oxidative stress**	**NRF2-mediated oxidative stress response**	**3.25E-07**	**1.65E-04**
KEGG_PATHWAY	**rno00140:Steroid hormone biosynthesis**	**Estrogen biosynthesis**	**6.89E-06**	**3.43E-04**
GOTERM_BP_DIRECT	**GO:0032496~response to lipopolysaccharide**	**LPS/IL-1 mediated inhibition of RXR function**	**0.0039**	**0.2913**
KEGG_PATHWAY	**rno00980:Metabolism of xenobiotics by cytochrome P450**	**Xenobiotic metabolism signaling**	**0.0075**	**0.0719**
KEGG_PATHWAY	**rno00270:Cysteine and methionine metabolism**	**Superpathway of methionine degradation**	**0.0114**	**0.1054**

KEGG_PATHWAY	**rno00240:Pyrimidine metabolism**	**Pyrimidine ribonucleotides de novo biosynthesis**	**0.0127**	**0.1133**
GOTERM_BP_DIRECT	**GO:0019240~citrulline biosynthetic process**	**Citrulline biosynthesis**	**0.0339**	**0.9182**
GOTERM_BP_DIRECT	**GO:0043516~regulation of DNA damage response, signal transduction by p53 class mediator**	**Cell cycle: G2/M DNA damage checkpoint regulation**	**0.0339**	**0.9182**

Table 20. Results of GSEA from DAVID for RNA-seq DEGs (Cytotoxic MoA). Overlapping terms between the Wang et. al. paper and this analysis are highlighted in red boundaries and bold text.

GSEA with DAVID (RNA Seq DEGs) for Cytotoxic MoA				
Category	**Term**	**Term in Paper**	**PValue**	**FDR**
UP_KEYWORDS	**Acetylation**	**Acetone Degradation I (to Methylglyoxal)**	**3.15E-32**	**1.09E-29**
GOTERM_CC_DIRECT	GO:0070062~extracellular exosome	-	3.82E-30	2.35E-27
UP_KEYWORDS	Phosphoprotein	-	1.40E-28	2.42E-26
GOTERM_BP_DIRECT	**GO:0042493~response to drug**	**Nicotine degradation II**	**1.80E-12**	**4.11E-09**
KEGG_PATHWAY	**rno00140:Steroid hormone biosynthesis**	**Estrogen biosynthesis**	**2.46E-10**	**3.17E-08**
KEGG_PATHWAY	**rno00980:Metabolism of xenobiotics by cytochrome P450**	**Xenobiotic metabolism signaling**	**4.41E-05**	**0.0013**
GOTERM_BP_DIRECT	**GO:0006979~response to oxidative stress**	**NRF2-mediated oxidative stress response**	**1.38E-04**	**0.0263**
KEGG_PATHWAY	**rno00270:Cysteine and methionine metabolism**	**Superpathway of methionine degradation**	**0.0027**	**0.0394**
UP_KEYWORDS	**DNA damage**	**Cell cycle: G2/M DNA damage checkpoint regulation**	**0.0155**	**0.0845**
GOTERM_BP_DIRECT	**GO:0032496~response to lipopolysaccharide**	**LPS/IL-1 mediated inhibition of RXR function**	**0.0435**	**0.9205**

From the GSEA, we observe that there is not a consistent overlap between many of the terms in the original paper and our study. This could be due to the original paper employing a different method to perform GSEA (GeneGo and Ingenuity) and our analysis for this study being performed using DAVID. Due to the lack of a common controlled vocabulary for the terms, it is expected that various discrepancies would be present. However, a deeper look into the terms obtained after our analysis in DAVID does show that there are connections between both terms in the original paper and our study as detailed by the overlapping terms in the tables.

We observe that certain terms like “xenobiotic metabolism signaling”, “nicotine degradation” and “LPS/IL-1 mediated inhibition of RXR function” are common across all the MoAs. These are also found in the DAVID GSEA analysis results from our study. This is expected as the samples were exposed to chemicals and as a result, the cells elicited a xenobiotic-related response. Although some of the FDR p-values are not significant, they are still highlighted in the table as they showed overlap with the terms from the original study. The non-significant FDR values could possibly be due to our study utilizing different enrichment methods than the one in the original paper as different enrichment tools use different algorithms and would inherently lead to similar discrepancies. We also observe terms that are associated

with carcinogenesis in the AhR and CAR.PXR MoAs. The treatment chemical that corresponds to AhR is 3-methylcholanthrene - which is known to be highly carcinogenic, therefore it is expected for the enrichment to have cancer related terms.[15] However, the treatment chemical corresponding to CAR.PXR MoA - Clotrimazole is a common medication used to treat fungal infections in humans and is non-carcinogenic. However, Clotrimazole has been extensively studied for its potential as an agent that prevents cancer-cell proliferation therefore it is expected to play a role in cancer pathways.[16]

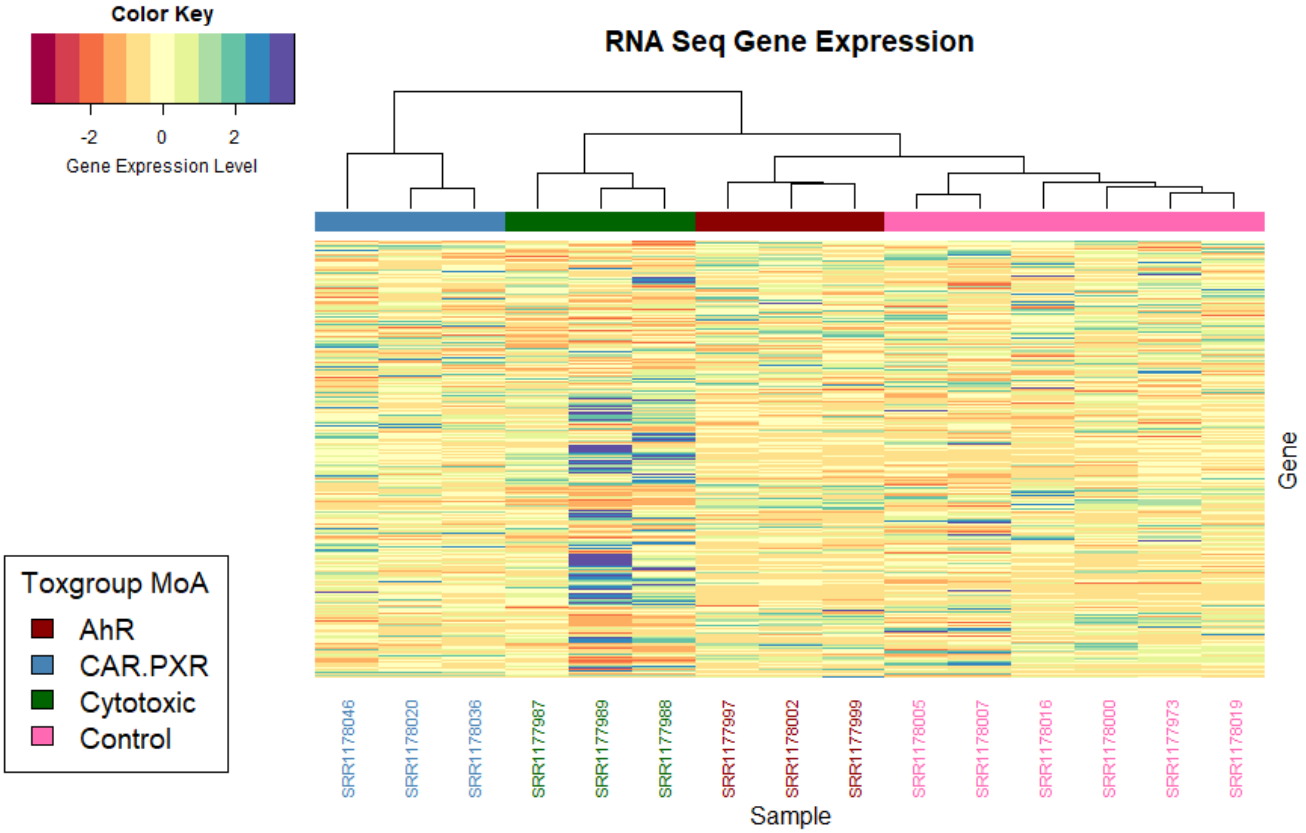

Figure 10. Hierarchically clustered gene-expression heatmap of RNA-seq samples in accordance with their MoA. The gene-expression heatmap for each sample along with the respective color coded label indicating MoA. Cluster dendrogram of the samples is seen on the top and the color key represents the gene expression level.

Hierarchically clustered heatmap for normalized RNA-Seq gene expression matrix is illustrated below in Figure 10. The clustered heatmap facilitates visualization of the normalized RNA-seq gene expression values along with the clustering of samples according to their toxgroup MoAs. We observed that all the samples clustered with a 100% accuracy with their respective MoAs. This indicates that the coefficient of variation filter successfully filtered out the genes that were insignificant and only retained the genes that give each cluster its unique signature. Based on the heatmap, we can observe that the samples belonging to the Cytotoxic MoA group have higher median expression levels when compared to the other MoAs. Samples belonging to the CAR.PXR MoA group had moderate median expression levels while the samples belonging to AhR MoA had the lowest median expression level amongst the three. The accurate clustering of samples according to their MoA suggests that genes could elicit a similar response when exposed to similar MoAs.[17]

Discussion

By following a similar workflow as Wang et al. (2014), we observed treatment and transcript abundance results by comparing RNA-seq differential expression and microarray data. We performed RNA-seq and microarray differential expression analyses to obtain significant DEGs and a subsequent concordance analysis assessing the agreement in DEGs between both platforms. Assessing concordance has implications in clinical and regulatory applications within the healthcare industry.

The top three genes that were differentially expressed for 3-Methylcholanthrene for RNA-seq analysis are CYP1A2, UGT1A7, and OAT. CYP1A2, or cytochrome P450, is a protein coding gene and is commonly associated with Porphyria Cutanea Tarda and Acetaminophen Metabolism.[18] Both Porphyria Cutanea Tarda and Acetaminophen Metabolism are categorized as metabolic diseases; however, Porphyria Cutanea Tarda is also categorized as a liver disease, which is consistent with what we would expect to find when observing the effects of chemicals on rat liver.[19] UGT1A7, or UDP-Glucuronosyltransferase, is also a protein coding gene and is associated with Gilbert Syndrome and Bilirubin Metabolic Disorder, which are both metabolic diseases.[20] Lastly OAT, or Ornithine Aminotransferases, is also a protein coding gene and is associated with metabolic diseases.[21] The top three genes that were differentially expressed for Clotrimazole for RNA-seq analysis are GSTA5, ABCC3, and CYP3A23/3A1. All three genes are protein coding genes and all associated with genetic, metabolic, and liver diseases.[22-24] Lastly, the top three genes that were differentially expressed for Chloroform for RNA-seq analysis are RGD1566134, LOC100360095, and ABCC3. Again, all three genes are protein coding genes. Specifically, both RGD1566134 and LOC100360095 are expressed in rat livers, which is consistent with our experiment.[25,26] ABCC3 is highly differentially expressed in both Clotrimazole and Chloroform. Overall, metabolic pathways and diseases are found to be associated with the top differentially expressed genes found using the RNA-seq. These findings could suggest an association between chemical perturbation of the liver and metabolic diseases in rats. Studies suggest that metabolic diseases are commonly linked to dysfunctionality of the liver.[27,28] This could explain our consistent metabolic results for the differentially expressed genes.

In microarray analysis, the top two genes that were differentially expressed for 3-Methylcholanthrene are CYP1A2 and UGT1A6. CYP1A2 is a common gene between the RNA-Seq and microarray analyses, and the metabolic diseases that are commonly associated are previously discussed. UGT1A6 is part of the UDP Glucuronosyltransferase family, and the protein encoded by this gene functions in removing of possible toxic xenobiotics which is consistent with the liver's function in xenobiotic metabolism[32]. An associated disease is Crigler-Najjar Syndrome which is characterized by high concentrations of bilirubin in the body[33]. The top three genes that were differentially expressed for Clotrimazole are CYP2B1, CYP3A23/3A1, and UGT2B1. CTP2B1 encodes the Cytochrome P450 B1 protein and CYP3A23/3A1 encodes a monooxygenase whose functions include oxidizing compounds such as steroids and xenobiotics[34]. UGT2B1 encodes the UDP-glucuronosyltransferase 2B1 that has functions relating to detoxification of drugs and xenobiotics[34]. This is consistent with our study as rat liver tissue was perturbed with varying chemicals. The top three genes that were differentially expressed for Chloroform are ABCC3, GSTP1, AKR1B8. An associated disease with ABCC3 is Dubin-Johnson Syndrome which is characterized by an accumulation of bilirubin and symptoms similar to Jaundice[35]. GSTP1 encodes a Glutathione-S-transferase which is vital to the chemical detoxification process[32]. Finally, AKR1B8 encodes the Aldo-keto reductase family 1, member B8 protein which is part of an enzyme family that catalyze redox transformations for detoxification processes[34]. The functions of the top differentially expressed genes appear to be related to detoxification processes of drugs and xenobiotics

which is relevant with the study as there was chemical perturbation of rat liver tissue. Consistent with the results mentioned for RNA-Seq, there appears to be metabolic diseases that are associated with the top differentially expressed genes determined from microarray analysis as well.

One of the major findings in Wang et al. (2014) was that concordance between the two platforms increased with the degree of perturbation. The treatment effect which included the number of DEG from RNA-Seq or microarray was important to determining the agreement between these two platforms. While the concordance values are approximately similar between Clotrimazole and Chloroform at differing numbers of DEGs, concordance analysis of the three treatments reveals this overall linear trend. In addition, concordance analysis of the above-median subsets reveals that there is overall a greater agreement between the two platforms when highly expressed genes are utilized which is consistent with the findings in Wang et al. (2014). Concordance analysis of the below-median subsets shows that there is substantially less agreement between the platforms when low expressed DEGs are utilized which is consistent with the findings in Wang et al. (2014). A possible reason for the discrepancy in the concordance values could be the filtering metrics such as Adjusted P-value, P-value, and fold change used for determining significant DEGs. In addition, the variable parameters used for concordance computation could be causing the discrepancy such as the *N* total number of genes in rat genome used within this analysis might be different to the value used in Wang et al. (2014)

Gene Set Enrichment Analysis (GSEA) revealed that the common gene ontology terms across all MoAs were related to xenobiotic responses which is expected as all of the samples were treated with different chemicals. Although there was some overlap between the terms mentioned in the paper and our analysis, there was no major consensus as different methods were used to perform GSEA in both the paper and our study. One of the reasons for this discrepancy is the limited availability of data regarding DEGs in the paper. If the DEGs described in the original paper were to be made available, we would have a better understanding of gene enrichment and as a result could perform a better comparison of the enrichment terms between the paper and our study. The samples after filtration based on coefficient of variation clustered accurately with their MoAs which indicates that only representative genes were retained after filtration. This also suggests that it could be expected that chemicals with similar MoA may elicit a similar gene response.

This study bolsters the fact that microarrays are not outdated technology. Even though there are a significant number of advantages for using microarrays, RNA-Seq tended to perform better for detecting low expressed DEGs, and both methods performed equally well when creating predictive classifiers. Both microarrays and RNA-seq have their own set of advantages and disadvantages and it is suggested in the study by Wang et. al. (2014) that there are a lot of factors that go into deciding the ideal method for a specific study. At times, it could be more advantageous to use microarrays while at other times, RNA-seq could yield better results. Therefore, the type of study and its biological complexity, transcript abundance, and intended application are important factors one should be considering while performing transcriptomic research and for decision-making. Additionally, this also invites for discussion over best practices, methods, and technologies being used in healthcare for transcriptomic studies. This has direct implications for the healthcare industry where the wrong choice of technology may change the outcome of the treatment. It could also glean different insights into disease mechanisms which could guide researchers into developing and selecting appropriate therapies.

Conclusion

Overall, in this study, we reproduced some similar results from Wang et al. (2014). We observed explained differences due to possible different filtering metrics for determining differentially expressed genes, or different variable values within the concordance computation. This analysis showed that there is still a significant difference in terms of determining genes that are differentially expressed between both platforms. Although our values for overall concordance were different to Wang et al.(2014), we were able to see the same overall trends discussed in the paper. Our study reinforces the conclusion from the original study by Wang et. al (2014) that concordance between both platforms increases with the abundance of differentially expressed genes. In addition, concordance between both platforms is greater for highly expressed genes, while there are discrepancies between the platforms when low expressed DEGs are used. Both technologies have their own advantages and disadvantages in different biological contexts and complexities, and selection should be appropriately considered for transcriptomic research. Hopefully in the future, there will be widespread use of technologies that combine the efficacies of both RNA-Seq and microarrays and enhance the prospects of transcriptomic research.

References

1. Bolón-Canedo, Verónica, et al. "Challenges and future trends for microarray analysis." Microarray bioinformatics. Humana, New York, NY, 2019. 283-293.
2. Wang, Charles, Binsheng Gong, Pierre R. Bushel, Jean Thierry-Mieg, Danielle Thierry-Mieg, Joshua Xu, Hong Fang, et al. 2014. "A comprehensive study design reveals treatment- and transcript abundance–dependent concordance between RNA-seq and microarray data" Nature Biotechnology 32 (9): 926–32. PMID: 4243706
3. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
4. Dobin, Alexander et al. "STAR: ultrafast universal RNA-seq aligner." Bioinformatics (Oxford, England) vol. 29,1 (2013): 15-21. doi:10.1093/bioinformatics/bts635
5. RNA-seq data from SRA (NCBI) with number SPR039021 https://www.ncbi.nlm.nih.gov/sra/?term=SRP039021
6. DNA data from GSO (NCBI): GSE55347 and GSE47875 GSE55347:https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55347 GSE47875:https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47875
7. Philip Ewels, Måns Magnusson, Sverker Lundin, Max Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report, Bioinformatics, Volume 32, Issue 19, 1 October 2016, Pages 3047–3048. https://multiqc.info/
8. Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 2013 May 1;41(10):e108. doi: 10.1093/nar/gkt214. Epub 2013 Apr 4. PMID: 23558742; PMCID: PMC3664803.http://subread.sourceforge.net/
9. Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.
10. Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biology, 15, 550. doi: 10.1186/s13059-014-0550-8. https://bioconductor.org/packages/release/bioc/html/DESeq2.html
11. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009;4(1):44-57. https://david.ncifcrf.gov/tools.jsp

12. Wang, Charles, et al. "The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance." *Nature biotechnology* 32.9 (2014): 926-932.
13. Love, Michael I., Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome biology 15.12 (2014): 1-21.
14. Warnes, Maintainer Gregory R., et al. "Package 'gplots'." *Various R programming tools for plotting data* (2016).
15. Alfred, L. J., et al. "A chemical carcinogen, 3-methylcholanthrene, alters T-cell function and induces T-suppressor cells in a mouse model system." Immunology 50.2 (1983): 207.
16. Kadavakollu, S., et al. "Clotrimazole as a cancer drug: a short review." *Medicinal chemistry* 4.11 (2014): 722.
17. Gene RGD1566134 from ensembl genome browser: http://useast.ensembl.org/Rattus_norvegicus/Gene/Summary?db=core;g=ENSRNOG00000042543;r=5:77430401-77433847;t=ENSRNOT00000076906
18. Gene CYP1A2 from GeneCards: https://www.genecards.org/cgi-bin/carddisp.pl?gene=CYP1A2
19. GeneCards Malacards: https://www.malacards.org/card/porphyria_cutanea_tarda#disorders
20. Gene UGT1A7 from GeneCards: https://www.genecards.org/cgi-bin/carddisp.pl?gene=UGT1A7
21. Gene OAT from GeneCards: https://www.genecards.org/cgi-bin/carddisp.pl?gene=OAT
22. Gene GSTA5 from GeneCards: https://www.genecards.org/cgi-bin/carddisp.pl?gene=GSTA5&keywords=GSTA5
23. Gene ABCC3 from GeneCards: https://www.genecards.org/cgi-bin/carddisp.pl?gene=ABCC3&keywords=ABCC3
24. NCBI Gene Cyp3a23-3a1: https://www.ncbi.nlm.nih.gov/gene/25642
25. NCBI Gene RGD1566134: https://www.ncbi.nlm.nih.gov/gene/298109#gene-expression
26. NCBI Gene RGD2322852: https://www.ncbi.nlm.nih.gov/gene/100360095
27. Watanabe, Sumio et al. "Liver diseases and metabolic syndrome." *Journal of gastroenterology* vol. 43,7 (2008): 509-18. doi:10.1007/s00535-008-2193-6
28. Rafiq, Nila, and Zobair M Younossi. "Interaction of metabolic syndrome, nonalcoholic fatty liver disease and chronic hepatitis C." *Expert review of gastroenterology & hepatology* vol. 2,2 (2008): 207-15. doi:10.1586/17474124.2.2.207
29. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." *Nucleic Acids Research*, **43**(7), e47. doi: 10.1093/nar/gkv007.
30. Sarah E Hunt, William McLaren, Laurent Gil, Anja Thormann, Helen Schuilenburg, Dan Sheppard, Andrew Parton, Irina M Armean, Stephen J Trevanion, Paul Flicek, Fiona Cunningham Ensembl variation resources *Database* Volume 2018 doi:10.1093/database/bay119
31. Twigger, S., Pruitt, K., Fernandez-Suarez, X., Karolchik, D., Worley, K., Maglott, D., Brown, G., Weinstock, G., Gibbs, R., Kent, J., Birney, E., & Jacob, H. (2008). What everybody should know about the rat genome and its online resources. *Nat Genet*, *40*(5). https://doi.org/10.1038/ng0508-523
32. Stelzer G, Rosen R, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Iny Stein T, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary D, Warshawsky D, Guan - Golan Y, Kohn A, Rappaport N, Safran M, and Lancet D. *The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analysis* , Current Protocols in Bioinformatics(2016), 54:1.30.1 - 1.30.33.doi: 10.1002 / cpbi.5.
33. Crigler-najjar syndome Medlineplus genetics (n.d) https://medlineplus.gov/genetics/condition/crigler-najjar-syndrome/

34. Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., ... & UniProt Consortium. (2020). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*. Chicago

35. Dubin-johnson syndrome: Medlineplus genetics. (n.d.) https://medlineplus.gov/genetics/condition/dubin-johnson-syndrome/

Supplementary Materials

Supplemental Table 1. Concordance Values for Differentially Expressed Genes (P-value < 0.05 and FC > 1.5) A filter of P-value < 0.05 and absolute Fold Change (FC) > 1.5 (absolute log2 FC > 0.58). This additional analysis was performed since a similar method was used for determining DEG in Wang et al. They used absolute FC > 1.5 cutoff and uncorrected p-value < 0.05. They did additional tests such as Levene's test for testing variances. In addition, there were slightly more genes (≈400) in 3-Methylcholanthrene for overall concordance analysis in this method.

A) Overall

Overall	
Treatment	**Concordance**
3-Methylcholanthrene	**5.052%**
Clotrimazole	**37.284%**
Chloroform	**45.646%**

B) Above-Median Subset

Above-Median	
Treatment	**Concordance**
3-Methylcholanthrene	**5.749%**
Clotrimazole	**38.584%**
Chloroform	**50.434%**

C) Below-Median Subset

Below-Median	
Treatment	**Concordance**
3-Methylcholanthrene	**2.774%**
Clotrimazole	**15.459%**
Chloroform	**20.674%**