Project 3 - Concordance of microarray and RNA-Seq differential gene expression; replicating Wang et al. 2004

Group members: Emily Hughes, Simran Makwana, Sumiti Sandhu, and Michiel Smit
TA: Nick

Introduction

RNA-seq provides unbiased genome-wide gene expression profiling, but its concordance with the conventional microarray platform is a question which should be further assessed for confident uses in clinical and regulatory application. The authors designed their study to generate Illumina RNA-seq andAffymetrix microarray data from the same set of liver samples of rats[1]. The rats were exposed to different degrees of perturbation by 27 chemicals representing multiple modes of action (MOA). They generated differentially expressed genes (DEGs) of each chemical as well in order to assess the influence of the treatment effect on the concordance between RNA-seq and microarrays and on the performance of predictive models generated from each technique.

For this analysis, the original work by Wang et al. will be replicated, focusing on a subset of chemicals that each represent distinct modes of action. Microarray and RNA-seq results will be compared to determine concordance of the two methods. Differential gene expression between treatments will be used for pathway enrichment analysis and compared with the results from the original study.

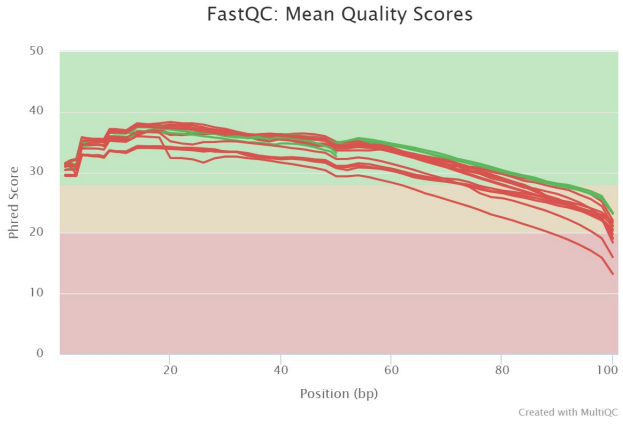
Data

For the purpose of this paper, a subset of samples was selected from tox-group 6 to analyze. All the datasets were downloaded and made available for us. The microarray (performed on Affymetrix whole genome GeneChip® Rat Genome 230 2.0 Array) and sequencing datasets (performed on Illumina HiScanSQ or HiSeq2000) are from the accessions SRP039021[2], GSE55347[3], and GSE47875[4] and were retrieved from NCBI. After identifying the samples to be analyzed based on the given tox-groups, FastQC was used to process the sample files. STAR aligner (v2.5.3a) was used to align each of the samples against the rat genome index, which has been built and provided for us. STAR aligner was run using the conda environment by installing miniconda and it was run on paired-end reads by joining the individual pairs. MultiQC tool was used to collect information from the alignment statistics produced by STAR.

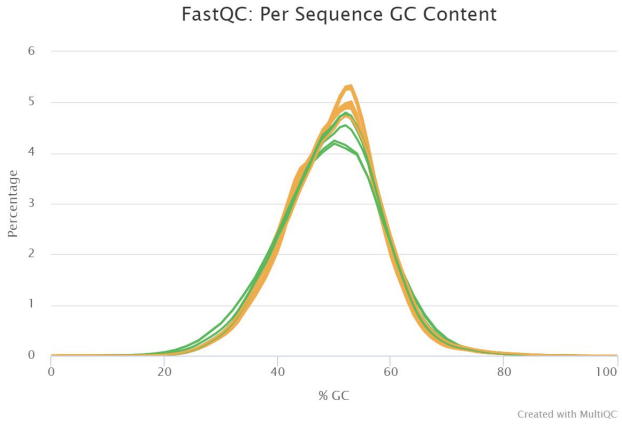
Sample Name	% Aligned	M Aligned	Chemical treatment	% Dups	% GC	Length	M Seqs
SRR1177963	84.8%	15.2	Pirinixic acid				
SRR1177963_1			Pirinixic acid	54.8%	48%	101 bp	17.9
SRR1177963_2			Pirinixic acid	52.3%	48%	101 bp	17.9
SRR1177964	85.3%	16.5	Pirinixic acid				
SRR1177964_1			Pirinixic acid	57.3%	48%	101 bp	19.3
SRR1177964_2			Pirinixic acid	54.8%	48%	101 bp	19.3
SRR1177965	85.1%	14.3	Pirinixic acid				
SRR1177965_1			Pirinixic acid	54.5%	48%	101 bp	16.8
SRR1177965_2			Pirinixic acid	51.8%	48%	101 bp	16.8
SRR1177997	89.2%	17.6	3-methylcholanthrene				
SRR1177997_1			3-methylcholanthrene	59.6%	49%	101 bp	19.7
SRR1177997_2			3-methylcholanthrene	58.6%	49%	101 bp	19.7
SRR1177999	88.7%	19.4	3-methylcholanthrene				
SRR1177999_1			3-methylcholanthrene	60.2%	49%	101 bp	21.8
SRR1177999_2			3-methylcholanthrene	58.9%	49%	101 bp	21.8
SRR1178002	89.1%	16.8	3-methylcholanthrene				
SRR1178002_1			3-methylcholanthrene	58.5%	49%	101 bp	18.8
SRR1178002_2			3-methylcholanthrene	57.6%	49%	101 bp	18.8
SRR1178014	83.5%	14.6	Fluconazole				
SRR1178014_1			Fluconazole	53.9%	49%	50 bp	17.5
SRR1178014_2			Fluconazole	51.9%	49%	50 bp	17.5
SRR1178021	82.0%	14.3	Fluconazole				
SRR1178021_1			Fluconazole	48.7%	49%	100 bp	17.5
SRR1178021_2			Fluconazole	46.4%	49%	100 bp	17.5
SRR1178047	84.0%	14.4	Fluconazole				
SRR1178047_1			Fluconazole	48.5%	49%	100 bp	17.1
SRR1178047_2			Fluconazole	47.3%	49%	100 bp	17.1

Table 1. General statistics from STAR alignment which includes the percent of uniquely mapped reads (% Aligned), the number of uniquely mapped reads in millions (M Aligned), percent of duplicate reads (% Dups), average percent of GC content (% GC), average sequence length in base pairs (Length), and total sequences in millions (M Seqs). The green colored entries represent the joined pairs while the orange colored entries represent individual pairs.

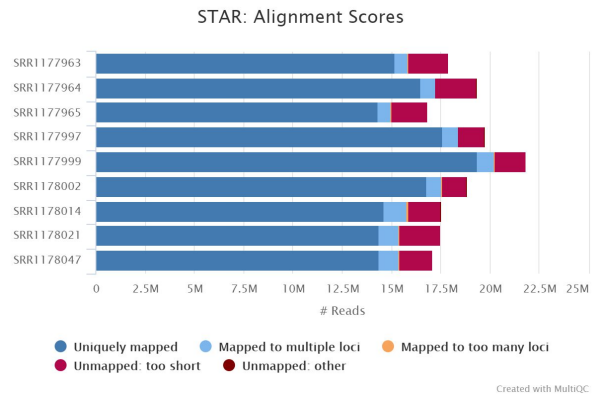
As shown in Table 1, the percentage of uniquely mapped reads are all above 80%, indicating that the quality of the libraries is good[5]. All of the average percentages of the GC content are either 48% or 49%, which is in the expected range of rats[6]. Most of the samples have an average sequence length of about 100 bp. The majority of the mean quality scores across the bases are in the green area, indicating a high Phred score (Figure 1A). Although the quality of reads appear to drop at the end of the sequence, this is common due to signal decay or phasing during the sequencing run. As seen in Figure 1B, five samples (in green) passed the threshold of MultiQC while 13 samples (in yellow) were given warnings. Despite the slight deviation from the expected normal distribution, visual inspection of the graph revealed that the GC content for each of the samples was still in the expected range. Figure 1C shows that the vast majority of the STAR alignments were uniquely mapped while a very small portion were unmapped due to being too short. The average number of reads across all of the samples is roughly 18 million. Given the report of MultiQC as a whole, it is not necessary to eliminate any particular sample due to low quality.



A)



B)



C)

Figure 1. Results from multiqc for each sample which include **(A)** fastqc mean quality score, **(B)** per sequence GC content percentage, and **(C)** STAR alignment scores.

Methods

The featureCounts tool from the subread (v1.5.2) was used to generate nine count files from the aligned reads in bam files containing treatment samples. Genomic features from the Gene Transfer Format (GTF) file were used. MultiQC (v1.6) was used to check the quality on the featureCounts results. An html report was compiled summarizing the nine samples, shown in Figure 2. The nine counts files were combined in a single comma-delimited text file using gene id, excluding gene ids with zero counts for all samples.

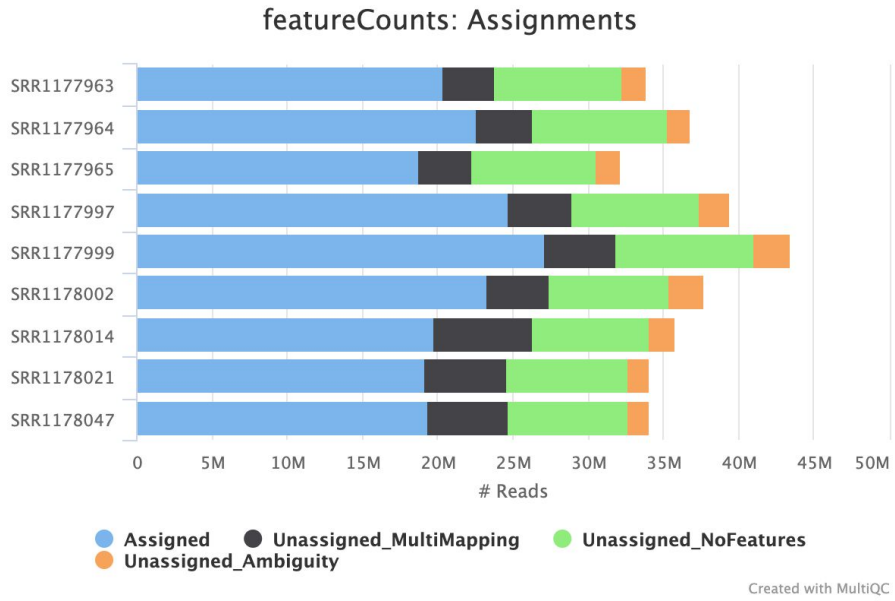


Figure 2: Multiqc results for featureCounts summarizing number of reads for treatment samples across four categories.

Box plots representing the distribution of the counts for each sample are shown in Figure 3. The boxplots in Figure 3 show that samples for all three groups of chemicals showed similar variability for the log of counts and had similar median values. The distribution for samples belonging to pirinixic acid and 3-methylcholanthrene is greater than that of fluconazole.

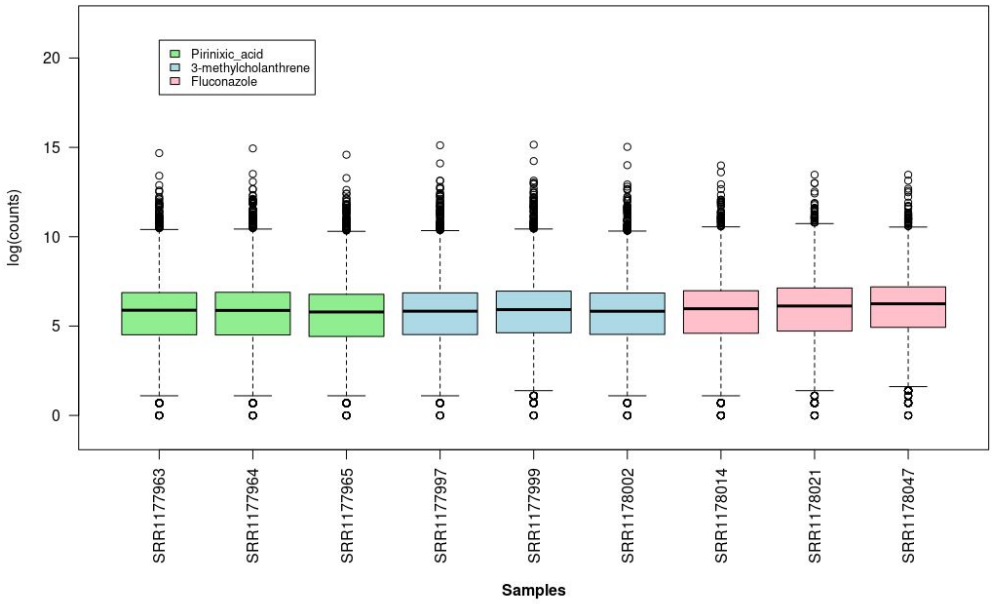


Figure 3: Box plot containing log of counts for nine treatment samples where each color represents different chemicals.

The DESeq2 package (v1.26.0) from Bioconductor was used to compare treatment and control samples using a file containing information about these samples[7]. A DESeq object was created using DESeqDataSetFromMatrix() for three groups, each having three treatment samples for a chemical (3-methylcholanthrene,fluconazole or pirinixic acid) and three controls having same vehicle value (CMC or corn oil) as the treatment samples for a particular chemical. Each chemical has a MOA associated with it. These three MOAs are associated with well-defined receptor-mediated processes—peroxisome proliferator-activated receptor alpha (PPARA), orphan nuclear hormone receptors (CAR/PXR) and aryl hydrocarbon receptor (AhR). MOAs were used as the design to model the DESeq matrix. DESeq() was run on the matrix created above, giving us three differentially expressed gene lists. Normalized counts were extracted out of the DESeq2 object for each comparison.

The limma package (Version 3.10) from Bioconductor was used to analyze the microarray data[8]. The results from the tox-group data were extracted from normalized results of the microarray chip, which was provided. Each of the chemicals was matched to controls in the same vehicle, either CMC or corn oil. After creating a design matrix for each chemical, differential analysis was performed. The lmFit function was used to generate a linear model for each gene. The eBayes function was used to obtain statistics based on the linear model, and the topTable function was used to extract the most significant differentially expressed genes using the Benjamini-Hochberg adjustment of p-values. Concordance between the microarray and RNA-Seq data was calculated using the same equation in Wang et al.[1] (displayed below).

$$\frac{2 \times intersect(DEGs_{microarray}, DEGs_{RNA-Seq})}{DEGs_{microarray} + DEGs_{RNA-Seq}}$$

In the equation, the variable DEGs represents the number of differentially expressed genes that have a nominal p-value < 0.05 and an absolute fold change greater than 1.5. The intersect between the two groups is the number of differentially expressed genes that have a nominal p-value < 0.05 and have the same sign of log fold change.

Results

RNA-Seq data

The differentially expressed results from the DESeq2 analysis were sorted by adjusted p-value and the top ten significant genes at adjusted p-value < 0.05 are reported in Table 2 along with their nominal p-value.

	3-methylcholanthrene		Fluconazole		Pirinixic acid	
Rank	**Gene**	**P-value**	**Gene**	**P-value**	**Gene**	**P-value**
1	Cyp1a2	3.64e-88	Cited4	2.87e-117	Fabp3	1.02e-192
2	Ugt1a9	2.75e-50	Stac3	2.65e-87	Apoa4	2.41e-166
3	Oat	8.60e-10	Cyp4a8	1.36e-64	Aqp7	1.30e-89
4	Mme	3.34e-9	Slc5a1	4.12e-59	Sult2a1	1.08e-89
5	Dusp6	4.87e-9	Trib3	8.81-47	Cyp2c7	1.07e-86
6	Lox	1.66e-8	Osmr	1.21e-44	Dusp6	4.54e-81
7	Adh7	3.30e-8	Lifr	6.60e-41	Acnat2	8.37e-74
8	Ddah1	1.25e-7	Cyp3a23/3a1	5.68e-40	Me1	2.72e-73
9	Nat8	1.52e-7	Il33	4.40e-38	Dhrs7l1	1.16e-68
10	Slc13a3	1.39e-7	Orm1	1.96e-36	Notum	6.54e-67

Table 2: Top ten differentially expressed genes (at adjusted p-value < 0.05) for each chemical along with their nominal p-value.

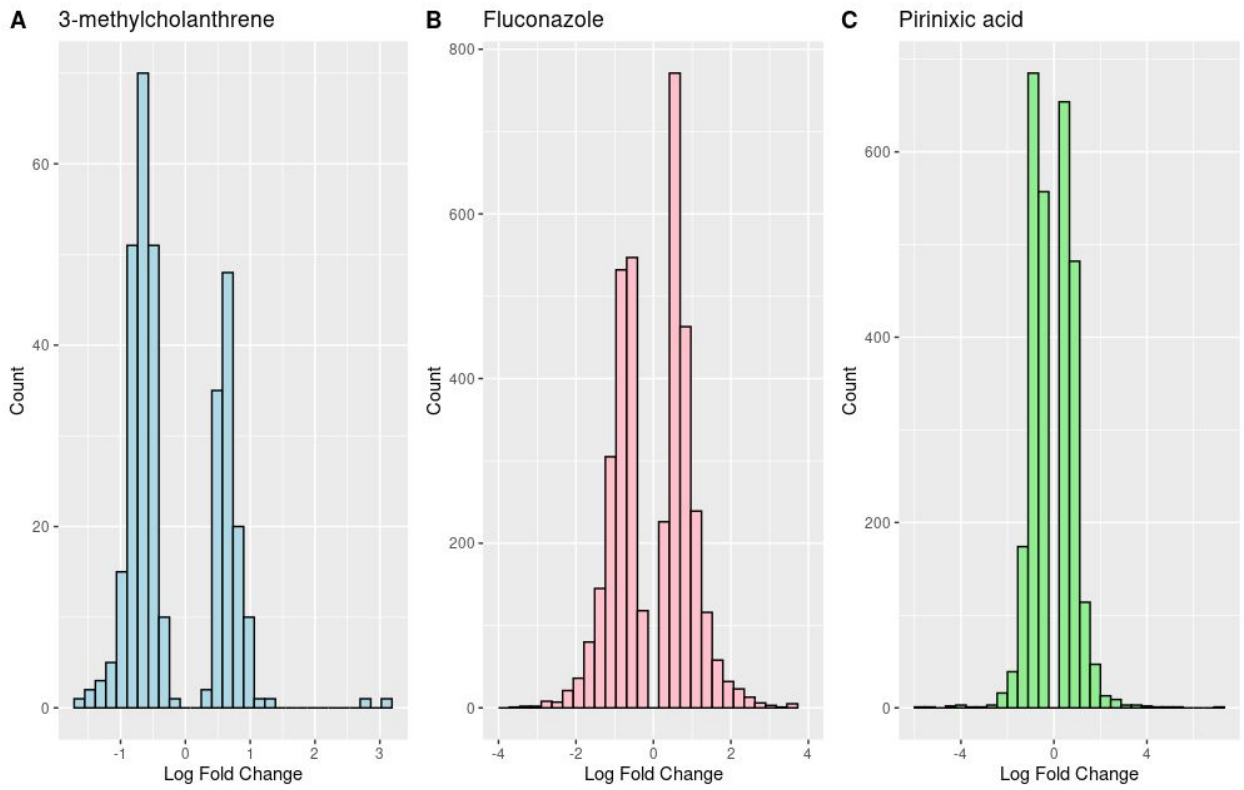


Figure 4: The distribution of the log fold change values of the significant differentially expressed genes at adjusted p-value < 0.05 for **(A)** 3-methylcholanthrene, **(B)** fluconazole, **(C)** pirinixic acid.

There were 328 genes for 3-methylcholanthrene, 3766 genes for fluconazole and 2814 genes pirinixic acid that were significant at adjusted p-value < 0.05. The log fold change and associated p-values of this subset of differentially expressed genes is shown in Figure 5.

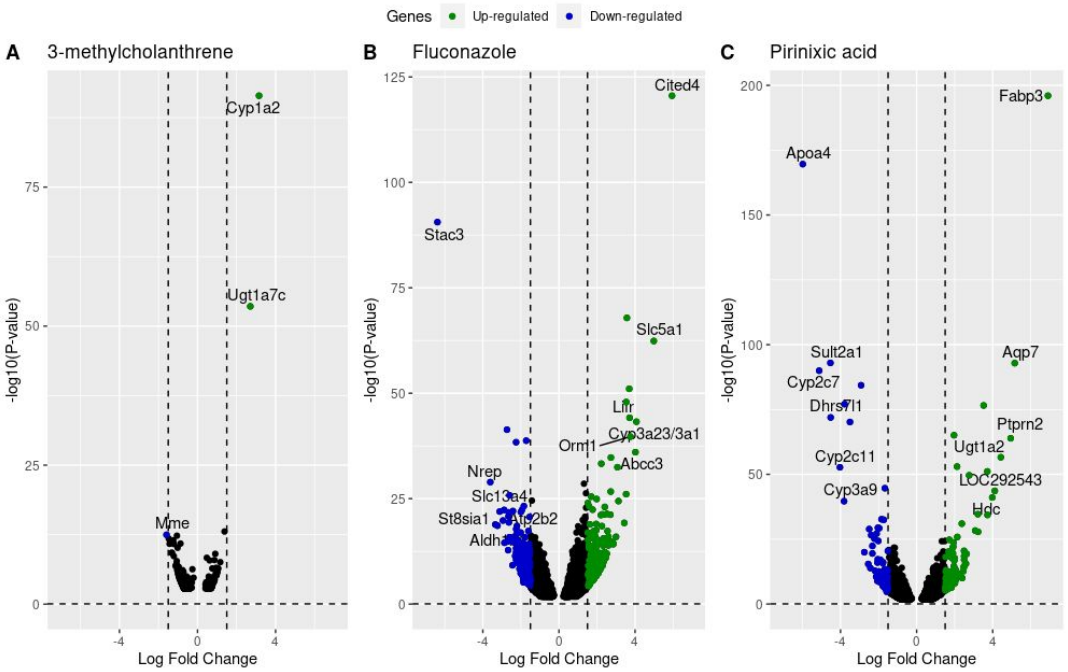


Figure 5: Volcano plots of the significant genes from each chemical. The horizontal dashed line represents the p-value threshold of 0.05. The vertical lines represent the log fold change values of ±(1.5). The up-regulated and down- regulated genes are colored green and blue with top up and down regulated genes represented by their symbols.

A heatmap was plotted to demonstrate trends in transcript abundance across and within microarray samples. In the differential expression analyses, fluconazole was matched to the corn oil control, while 3-methylcholanthrene and pirinixic acid were matched to the CMC control.

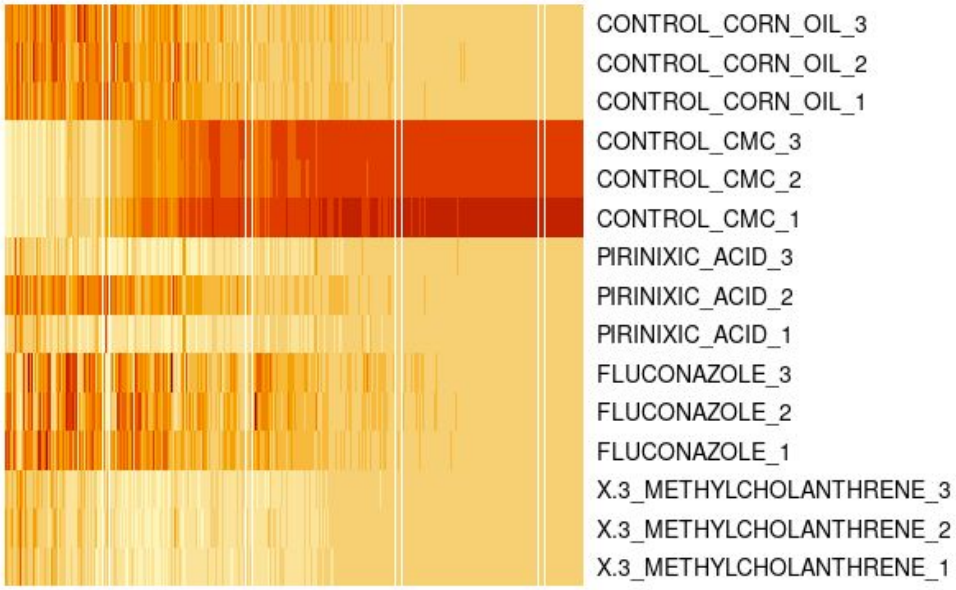


Figure 6. Heatmap of normalized counts for each treatment condition and control, clustered by gene expression. Red color indicates higher count while yellow indicates lower count.

Microarray analysis

The number of significant genes (adjusted p-value < 0.05) after the differential expression analysis on the microarray data was 58 for 3-methylcholanthrene, 1997 for fluconazole, and 8761for pirinixic acid. Table 3 presents the top ten most significant genes and the associated nominal p-values for each chemical group. The distribution of the log fold change values of the significant genes for each chemical can be seen in Figure 7.

	3-methylcholanthrene		**Fluconazole**		**Pirinixic acid**	
Rank	**Gene**	**P-value**	**Gene**	**P-value**	**Gene**	**P-value**
1	Cyp1a2	2.54E-17	Orm1	7.62E-12	Acot1	2.17E-32
2	Ugt1a9	1.35E-12	Cyp2b1	3.53E-10	Acot2	3.94E-26
3	Smim13	2.08E-09	Ablim3	4.67E-10	Pex11a	6.24E-25
4	Ptprs	5.94E-07	Id4	1.14E-09	Aig1	2.77E-24
5	Gsta4	8.29E-07	Irak3	2.68E-09	Ech1	1.02E-23
6	Pon3	9.71E-07	Rnf144a	2.79E-09	Vnn1	3.52E-23
7	Slc34a2	1.90E-06	Clpx	4.62E-09	Hsdl2	8.93E-23
8	Map1lc3b	3.15E-06	Tmem252	1.16E-08	Acox1	5.46E-22
9	Pla2g12a	4.29E-06	Rgs3	1.56E-08	Pex11a	2.08E-21
10	Cyp1a1	4.78E-06	Por	2.23E-08	Hdc	5.79E-21

Table 3. Top ten of the most significant genes based on p-value for each chemical. Note: if the probe ID did not match to a gene in the map or the same gene was mapped multiple times, the row was omitted.

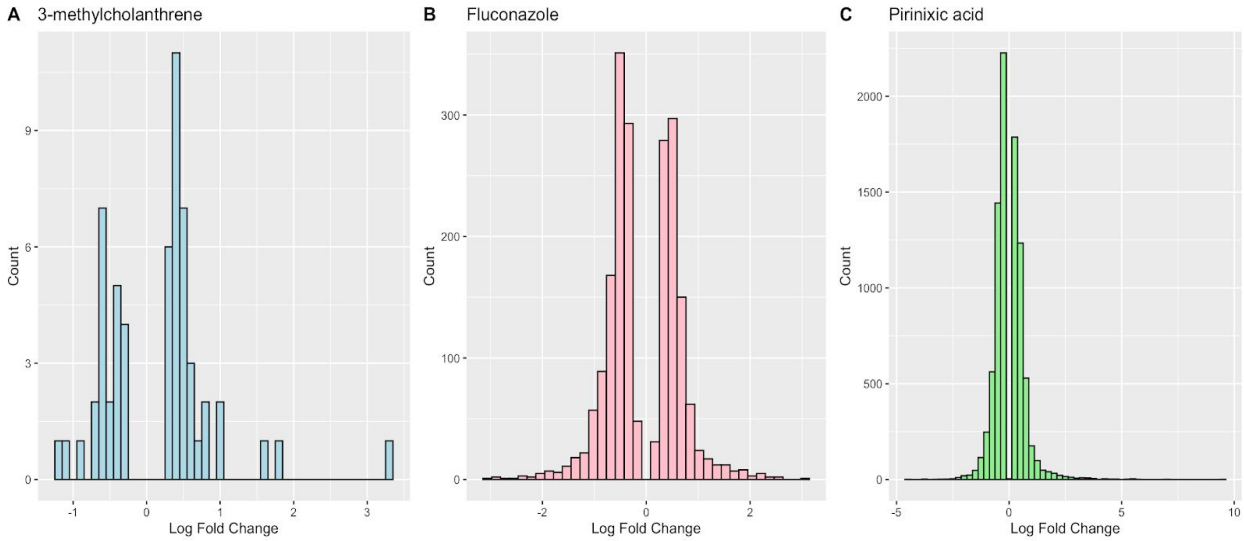


Figure 7. Distribution of the log fold change of the significant differentially expressed genes for each chemical. Genes were deemed significant if there was an associated adjusted p-value < 0.05.

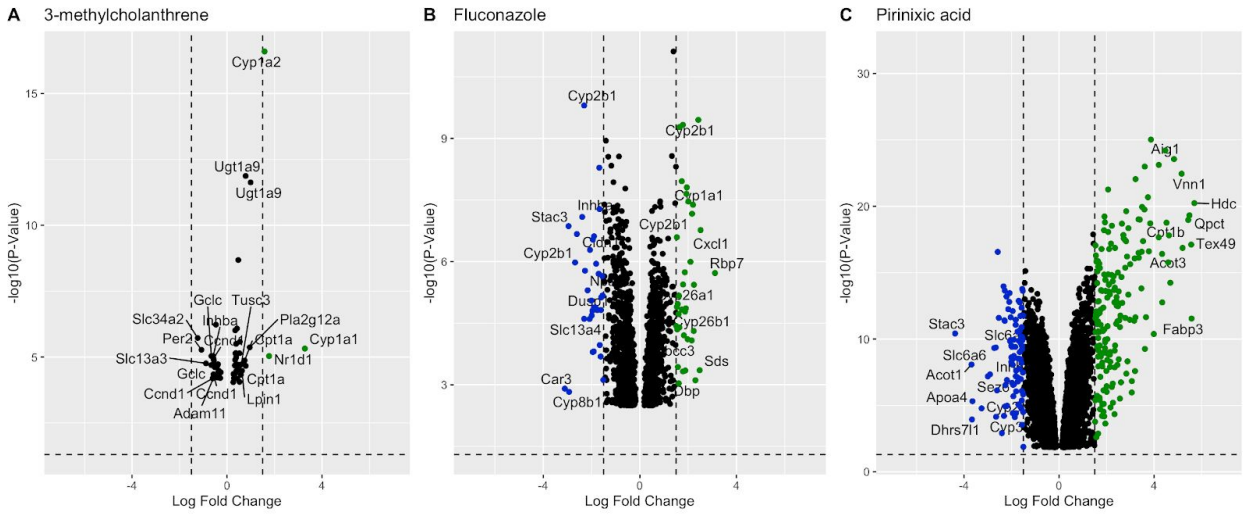


Figure 8. Volcano plots of the significant genes from each chemical. The horizontal dashed line represents the p-value threshold of 0.05. The vertical lines represent the threshold of ±1.5 log fold change. The blue and green dots represent those that pass the 1.5 absolute log fold change threshold. The top ten genes with highest and lowest log fold change were labeled.

Concordance

The data from the RNA-Seq and microarray analyses were filtered using an adjusted p-value threshold of < 0.05 and an absolute log fold change > 1.5. Concordance was calculated for the genes of each chemical group using the equation described in the Methods section. The concordance for 3-methylcholanthrene was 0%, fluconazole was 17.8%, and pirinixic acid 0.9%. Additionally, concordance was calculated for the genes with above and below median expression values, which are shown in Figure 9. Concordance values were plotted as a function of the number of differentially expressed genes overall, which were determined based on the thresholds previously mentioned (Figure 10).

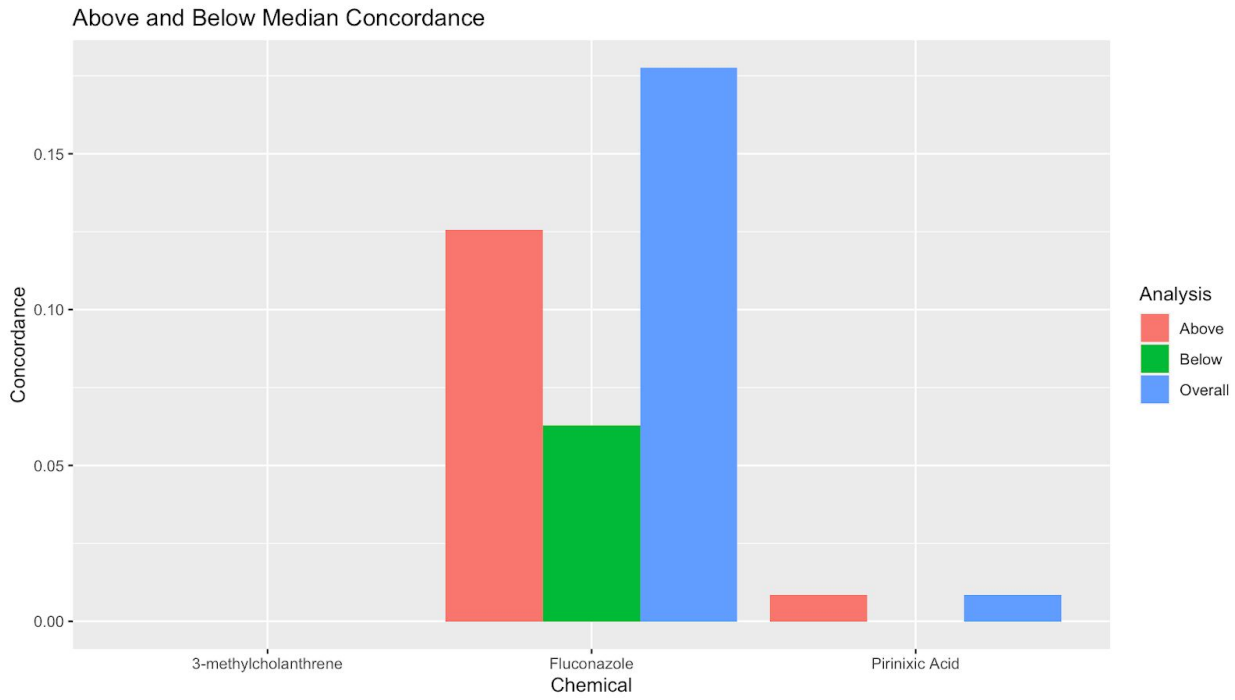


Figure 9. Concordance values for each chemical. The analysis was performed for genes above median expression, below median expression, and all genes. Note: after filtering of the data based on adjusted p-value < 0.05 and absolute log fold change > 1.5, the 3-methylcholanthrene group did not have any genes overlapping the microarray and RNA-Seq analyses.

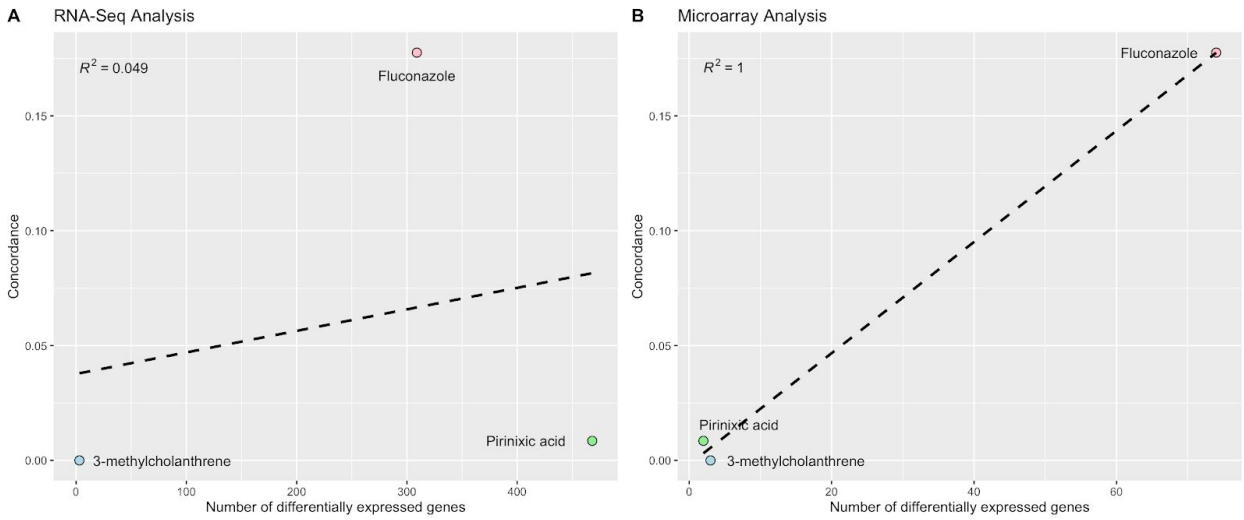


Figure 10. Scatter plots displaying concordance and number of differentially expressed genes for each chemical. The dashed line represents the trendline of the data with the associated R-squared value in the top left.

Enrichment Analysis

Of the significantly differentially expressed genes (adjusted p-value < 0.05), those that had an absolute log fold change value greater than one were selected for pathway analysis. DAVID was used for enrichment analysis for the gene list corresponding to each MOA using functional annotation clustering. The significantly enriched pathways (those with an FDR > 0.05) were selected as those representative for each respective mode of action, shown in Table 4.

Pathway	**P-Value**	**FDR**
PPARA		
PPAR signaling pathway	1.12e-15	1.45e-12
Fatty acid metabolism*	7.36e-15	9.64e-12
Fatty acid degradation*	4.49e-13	5.91e-10
Metabolic pathways	3.49e-11	4.58e-08
Chemical carcinogenesis	5.26e-08	6.91e-05
Steroid hormone biosynthesis	1.73e-07	2.28e-04
Staphylococcus aureus infection	3.32e-07	4.36e-04
Biosynthesis of unsaturated fatty acids*	3.82e-07	5.02e-04
Peroxisome	3.9e-07	5.13e-04

Metabolism of xenobiotics by cytochrome P450	4.56e-07	6e-04
Cell adhesion molecules (CAMs)	8.94e-07	1.18e-03
Retinol metabolism*	1.27e-06	1.67e-03
Drug metabolism - cytochrome P450	1.37e-05	1.8e-02
Valine, leucine and isoleucine degradation*	1.43e-05	1.88e-02
Biosynthesis of antibiotics	3.7e-05	4.86e-02
CAR/PXR		
Metabolic pathways*	1.68e-14	2.2e-11
Chemical carcinogenesis	5.46e-13	7.17e-10
Retinol metabolism*	2.37e-12	3.1e-09
Steroid hormone biosynthesis	9.14e-12	1.2e-08
Drug metabolism - other enzymes*	3.77e-09	4.95e-06
Metabolism of xenobiotics by cytochrome P450	3.16e-08	4.15e-05
Drug metabolism - cytochrome P450*	2.28e-07	3e-04
Bile secretion	2.28e-07	3e-04
Complement and coagulation cascades	2.87e-07	3.77e-04
ABC transporters	2.24e-05	2.94e-02
AhR		
Chemical carcinogenesis	2.04e-06	2.04e-03
Metabolism of xenobiotics by cytochrome P450	2.14e-05	2.13e-02
Drug metabolism - cytochrome P450*	2.26e-05	2.26e-02
Retinol metabolism*	4.2e-05	4.19e-02

Table 4. Significantly enriched pathways (FDR < 0.05) for each mode of action investigated. Similarities with Supplementary Table 4 from are marked with a (*) while exact matches are bolded.

Discussion

Enrichment Analysis

The selected differentially expressed genes were found to be enriched in certain pathways (Table 4). While these pathways did not correspond well with those found from Wang et al. in their Supplementary Figure 4, there were certain potential overlaps. Our enrichment analysis found the fatty acid metabolism, fatty acid degradation, and fatty acid pathways to be significantly enriched for the PPARA mode of action (Table 4). While Wang et al. did not have these exact pathways, they did find fatty acid activation, fatty acid alpha-oxidation, fatty acid beta-oxidation I, and fatty acid beta-oxidation III to be significantly enriched[1]. All three of our modes of actions investigated found the retinol metabolism pathway to be significant (Table 4), which may correspond with the retinoate biosynthesis I pathway that Wang et al. found significant for each of the modes of actions as well[1]. Additionally, our analysis with DAVID highlighted metabolic pathways for the CAR/PXR mode of action (Table 4), while Wang et al. found xenobiotic metabolism signaling enriched for this mode of action[1].

The discrepancy between our results and those of Wang et al. may be due to differences in enrichment analysis methods. While we used DAVID for this analysis, Wang et al. used two different pathway analysis strategies (mapping Entrez gene IDs for differentially expressed genes to GeneGo's canonical pathway maps and to the pathway collection of the Ingenuity Knowledge Base[1]. It is possible that the different methods of enrichment analysis highlight different pathways, therefore producing different results. It is also possible that the different databases used invoke different names for each pathway, which would explain how some of our results appeared slightly similar to those of Wang et al. but did not exactly match.

Investigating the normalized counts, which are representative of sample transcript abundance, revealed similarities in abundance within chemical samples. The three fluconazole samples show a similar pattern of abundance, just as the three 3_methylcholanthrene samples do (Figure 6). This sanity check suggests that each of the replicated samples show the same trend. Additionally, the marked difference in abundance between the control samples and the rest of the chemical treatment samples shows the expected change due to treatment. A surprising result is the discrepancy between abundance for pirinixic acid samples; we would expect a similar trend for each of these but find the second pirinixic acid sample (SRR1178064) to more closely mirror the fluconazole samples. This might reflect an error in sampling or sample development. However, the microarray samples do tend to generally cluster by chemical (Figure 6).

Concordance

Calculation of the concordance values for each chemical resulted in 0% for 3-methylcholanthrene, 17.8% for fluconazole, and 0.9% for pirinixic acid. In the original study, the thresholds were of adjusted p-value < 0.05 and absolute log fold change > 1.5[1]. Because this is a replicate study, the same thresholds were applied for each tox-group. The chemical group 3-methylcholanthrene had 58 and 328 significant differentially expressed genes for microarray and RNA-Seq analyses, respectively. These values were much lower than those observed in the other two chemical groups. Due to the relatively low number of genes, there were no genes that intersected between these two analyses resulting in a concordance calculation of zero. Fluconazole had a much higher concordance calculation than the other two chemicals. Interestingly, this was the only chemical that was administered to the rats using corn oil as the vehicle; this may have played a role in the final results.

As seen with Figure 9, the differentially expressed genes with above-median expression had a higher concordance than those with below-median expression. This means that genes with higher expression values were more likely to be shared across the differential expression analyses of both RNA-Seq and microarray data. Figure 10 shows the general trend that concordance increases as the number of differentially expressed genes increases, which was also observed in the original study in Figure 2A[1].

Conclusion

Overall, the results from this replicated study concur with those of Wang et al. with varying degrees of similarity. Certain pathways were found to be enriched for certain chemicals and modes of action, which could be potential targets when using these chemicals in clinical applications. Generally speaking, the greater the overall number of significant differentially expressed genes found, the greater the concordance between the methods investigated. The greater the magnitude of differential expression, the more likely the gene is to be found as significantly differentially expressed in both the RNA-Seq and microarray analyses.

One complication for our replicate study was that the map matching probes to genes was incomplete. Multiple probes matched to the same gene and some probes had no genes associated with them. This made it extremely difficult to match the RNA-Seq data to the microarray data and to find genes associated with the microarray data. Additionally, comparing the results from our enrichment analysis to that of the paper was difficult due to the lack of similarities between pathway names. This may be due to the limited number of samples used in this analysis or due to differences in enrichment pathway databases.

References

1. 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
2. Wang, C., Auerbach, S.S., Shi, L., Tong, W. (2014). SEQC Toxicogenomics Study: RNA-Seq data set (data accessible at NCBI GEO database, accession SRP039021).
3. Wang, C., Auerbach, S.S., Shi, L., Tong, W. (2014). In Vivo Rat Liver Carcinogen (data accessible at NCBI GEO database, accession GSE55347).
4. Wang, C., Auerbach, S.S., Shi, L., Tong, W. (2014). SEQC Toxicogenomics Study: microarray data set (data accessible at NCBI GEO database, accession GSE47792).
5. Dobin, A., & Gingeras, T. R. (2015). Mapping RNA-seq Reads with STAR. *Current protocols in bioinformatics*, *51*, 11.14.1–11.14.19. https://doi.org/10.1002/0471250953.bi1114s51
6. Zhang, L., Kasif, S., Cantor, C.R., Broude, N.E. (2004). GC/AT-content spikes as genomic punctuation marks. *Proc Natl Acad Sci USA* 101:16855-16860.
7. Love, M., Anders, S., & Huber, W. (2014). Beginner's guide to using the DESeq2 package. *Genome biology*, *15*, 550.
8. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7), e47.