Concordance of Microarray and RNA-Seq Differential Gene

Group: Dachsund
Sana Majid (Data Curator) ♦ Katherine Tu (Programmer) ♦ Shreen Katyan (Analyst) ♦ Vamshi Mallepalli (Biologist)

Introduction

There have been numerous advancements in the realm of basic research, and government based regulatory institutes have welcomed and promoted the translation of these findings to support drug discovery and precision medicine. High-throughput sequencing has been an eminent challenge as microarrays have been replaced by RNA-sequencing for transcriptome level analysis of differential gene expression. The FDA launched the Microarray Quality Control Consortium initially to test the reliability of microarray data. Through their sponsorship, Charles Wang and group sought to compare Illumina RNA-seq and Affymetrix microarray data in terms of differential gene expression through the study of a range of chemical treatment conditions and their effects on the rat liver. Liver samples of rats were exposed to 27 different chemicals with multiple modes of action (MOA) [1].

In this class project, we sought to reproduce select findings to determine the concordance between RNA-seq and microarray, as well as compare pathway enrichment results reported by Wang et al [1].

Data

We chose to analyze the subset of RNA samples listed under toxgroup 6, which consisted of the following fastq files generated from rat liver samples exposed to select chemicals in triplicates: SRR1177997, SRR1177999, SRR1178002 (chemical: 3-methylcholanthrene, MOA: AhR); SRR1178014, SRR1178021, SRR1178047 (chemical: fluconazole, MOA: CAR/PXR), and SRR1177963, SRR1177964, SRR1177965 (chemical: pirinixic acid, MOA: PPARA). These files were copied into the Dachsund group folder, which had previously been from the accessions **SRP039021**, **GSE55347**, and **GSE47875** by the course instructors.

Following the acquisition of the fastq files, quality control was individually assessed for each using FastQC [2]. MultiQC was then performed to compile the individual FastQC results in order to assess read quality. As shown in **Figure 1**, the treatment run samples averaged 40-50% unique reads. Only five of eighteen samples had passing mean quality Phred scores (**Figure 2**); however, all samples had passing per sequence quality scores (figure not shown). All samples had a relatively normal distribution of GC content (**Figure 3**), although five reads passed the quality thresholds.

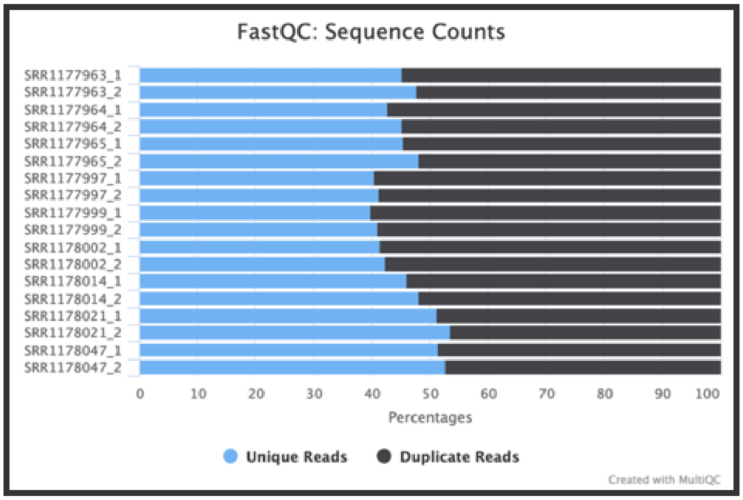


Figure 1. Percentage counts of unique vs. duplicate reads in run samples.

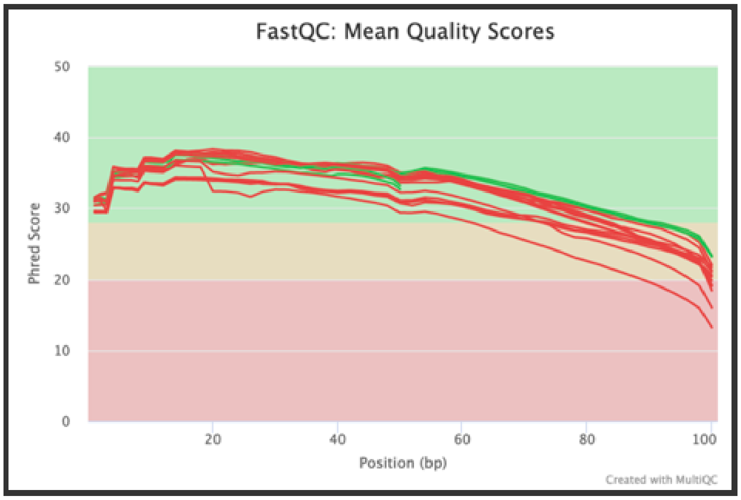


Figure 2. Phred score for run samples.

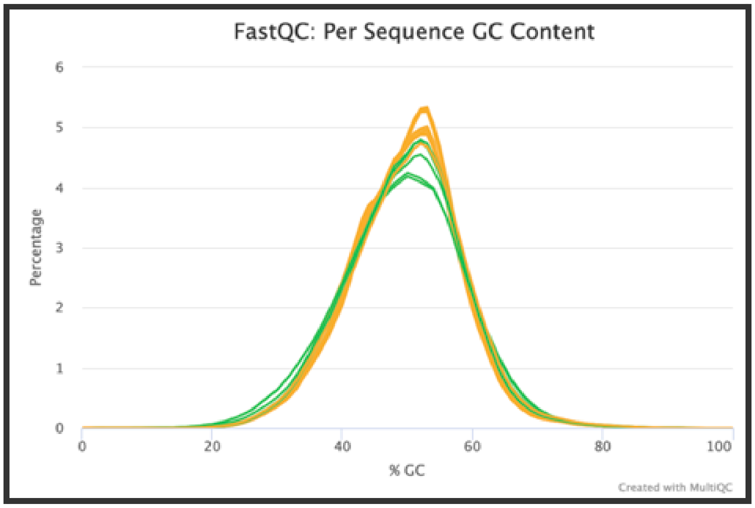
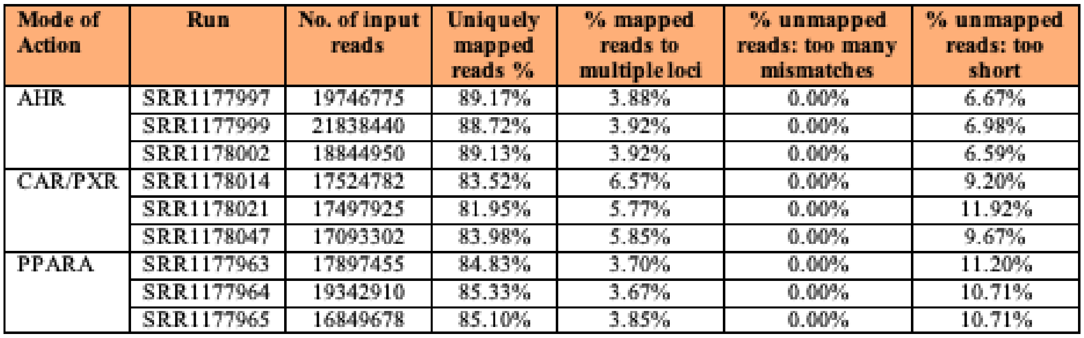


Figure 3. GC content for run samples.

Following quality assessment, each of the paired sample runs was aligned using the STAR aligner (star 2.6.0c) against the reference rat genome provided at /project/bf528/project_3/reference/rn4_STAR. More than 80% of the input reads were aligned uniquely to the reference genome, with 3-7% aligning to multiple loci, and the main reason for unalignment being due to reads being too short (**Table 1**).



Mode of Action	Run	No. of input reads	Uniquely mapped reads %	% mapped reads to multiple loci	% unmapped reads: too many mismatches	% unmapped reads: too short
AHR	SRR1177997	19746775	89.17%	3.88%	0.00%	6.67%
	SRR1177999	21838440	88.72%	3.92%	0.00%	6.98%
	SRR1178002	18844950	89.13%	3.92%	0.00%	6.59%
CAR/PXR	SRR1178014	17524782	83.52%	6.57%	0.00%	9.20%
	SRR1178021	17497925	81.95%	5.77%	0.00%	11.92%
	SRR1178047	17093302	83.98%	5.85%	0.00%	9.67%
PPARA	SRR1177963	17897455	84.83%	3.70%	0.00%	11.20%
	SRR1177964	19342910	85.33%	3.67%	0.00%	10.71%
	SRR1177965	16849678	85.10%	3.85%	0.00%	10.71%

Table 1. Read and alignment statistics with the STAR aligner.

Methods

Read Counting with featureCounts

The BAM files obtained from applying STAR from the previous step were then processed through the featureCounts program from the subread package to generate count matrices for each of the samples. In our analysis, we cover three different treatment groups, each with three samples treated with chemicals 3-methylcholanthrene, fluconazole, and pirinixic acid, respectively. The featureCount program count reads to genomic features such as genes, exons, promoters, and genomic bins [3]. The program requires not only bam files that state the genomic coordinates of where the read is mapped, but also an annotation in GTF format that serves as a cross-reference to the genomic coordinates of the features we are interested in counting expression of. The run results in a count matrix that includes the following columns: geneid, chromosome, start position, end position, strand, length, and count. We further extracted just the geneid and count, first and seventh column, respectively, in each sample to plot a boxplot to visualize the sample count distribution. The resulting count values for each sample are shown in figure 1 and further explained in the result section. The featureCounts program also produced a summary file that includes the number of assigned and unassigned reads, and other information such as unassigned ambiguity and multimapping. This information is then evaluated through MultiQC for quality control.

RNA-Seq Differential Expression with DESeq2

To estimate the count difference in each sample, we further process the counts obtained in the previous step through the DESeq2 package from Bioconductor. DESeq2 performs this estimate using a negative binomial regression. A negative binomial regression is used instead of linear models because of the following three properties of the count. First, the counts are discrete. Second, there are no negative values and thus do not follow a normal distribution. Lastly, the mean and variance of a given gene's count distribution are not independent.

DESeq2 was performed to analyze the differentially expressed genes compared to its control group and across different treatment groups. Each treatment group was analyzed with DESeq2 separately and compared with its own control group, which are those that had the same vehicle. The vehicle indicates how the chemicals have been delivered to the cells and are CMC_._% and corn oil (100%) in this study.

By comparing the control group information, we can make sure that the differentially expressed genes are caused because of the different modes of action. The read counts of the control group were provided and the only step we had to perform was to merge the featureCount results from the previous step to its control group. In addition, we removed genes that had counts equal to zero, reducing the number of genes from 18014 to 11011.

DESeq2 requires the input to be a matrix and so we had to process the function DESeqDataSetFromMatrix. DESeqDataSetFromMatrix also requires the input of countData to be a data.frame format, including column names and row names. It is important to also include the row names here as they provide information about which gene the row is about and will be helpful in further analysis. We also inputted the colData from the RNA information csv file provided. Lastly, the design attribute is referring to the mode of action column in the RNA information csv file. After running DESeq, the generated output included columns base mean, log2 fold change, lfcSE, p-value, and adjusted p-value. This file was exported as a csv and further analyzed for the concordance between RNA-seq and microarray data.

Results

Read Counting with featureCounts

Here in **figure 4** shows the boxplot of the sample count distribution of each sample with different treatments. Counting three as a group from left to right, the group is treated with 3-methylcholanthrene, fluconazole, and pirinixic acid, respectively. Figure 4a plots the raw count distribution. The reason why we cannot see a box in this boxplot is that most counts are low or zero in the samples that the interquartile range became almost flat in the plot. Although most counts are low, there appear to have some outliers with a larger number of counts in each sample. In addition, it is interesting to see how the outliers in each sample are similar to other samples within the same group. In order to get a better visualization of the count distribution, we converted the number of counts into a logarithmic scale shown in figure 4b. Because log(0) is undefined, the plot is generated by the counts that are not equal to zero. This plot shows a slightly better change in count distribution in each sample. The difference in count distribution makes us believe that there will be differential gene expression differences in each sample.

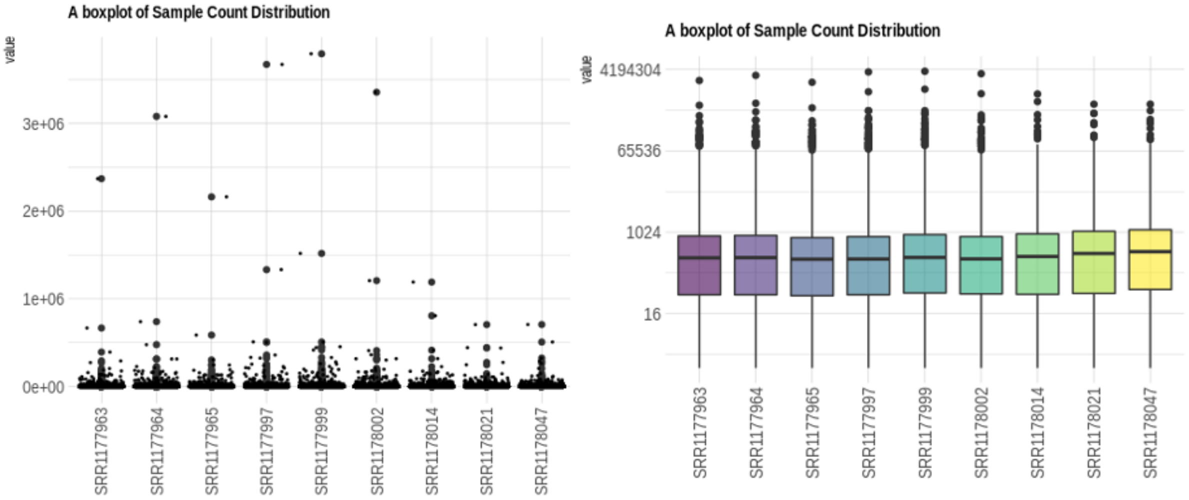
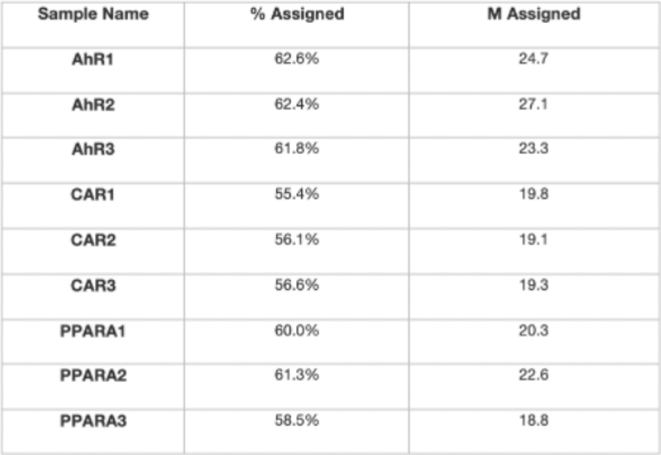


Figure 4. Boxplot of sample count distribution (a) on the left, shows the raw sample count distribution with most counts equal to zero. (b) on the right, the sample count distribution is shown in a logarithmic scale with better visualization of the distribution in each sample.

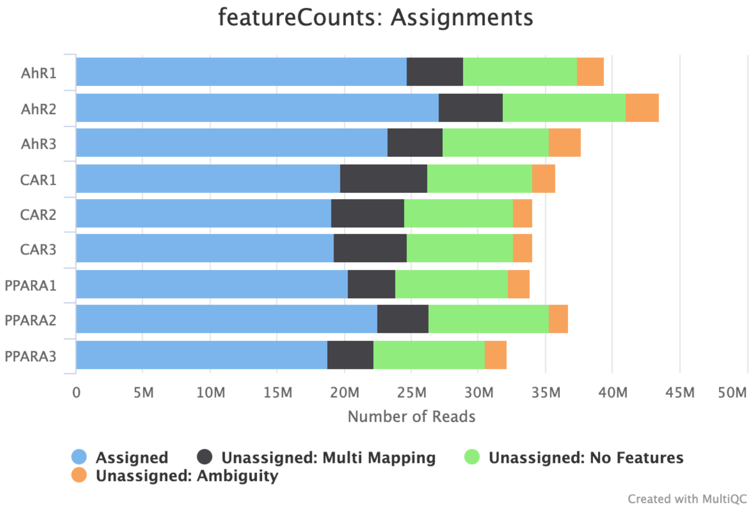
Quality Control

MultiQC is a modular tool to aggregate results from bioinformatics analyses across many samples into a single report. It creates an HTML file that further summarizes all the summary files produced in the previous step and provides all relevant summaries. The percentage of assigned reads and assigned reads in million for each sample are shown in **table 2**. From the table, we can see that the reads aligned in the nine samples did not differ much, falling between 55% to 62% and 19 to 27 million reads. This is even better understood if we look at **figure 5**. Figure 5a shows the distribution of assigned and unassigned featureCounts. We can see from the plot that most genes are assigned, a portion are unassigned due to having no features or multi-mapping, and only a small portion are unassigned due to ambiguity. Figure 5b further shows the percentage of each assigned and unassigned category. It is interesting to note that samples within the same treatment group seem to show similar assignment and unassignment percentage as indicated by the similarity of their coloring bar length. The treatment group with the AhR mode of action (MOA) had assigned reads percentage ranging from 61.8% to 62.6%. CAR/PXR MOA had assigned reads percentage falling between 55.4% and 56.6%. The PPARA MOA treatment group had assigned reads percentage ranging from 58.5% to 61.3%. Because the assigned unassigned reads distribution showed a similar trend, all samples were used for the following RNA-Seq differential expression analysis.



Sample Name	% Assigned	M Assigned
AhR1	62.6%	24.7
AhR2	62.4%	27.1
AhR3	61.8%	23.3
CAR1	55.4%	19.8
CAR2	56.1%	19.1
CAR3	56.6%	19.3
PPARA1	60.0%	20.3
PPARA2	61.3%	22.6
PPARA3	58.5%	18.8

Table 2. General statistics of the count matrices summary. The sample name indicates the mode of action of each treatment group with three samples in each group. The second and third column indicate the percentage of assigned reads and assigned reads in million, respectively.



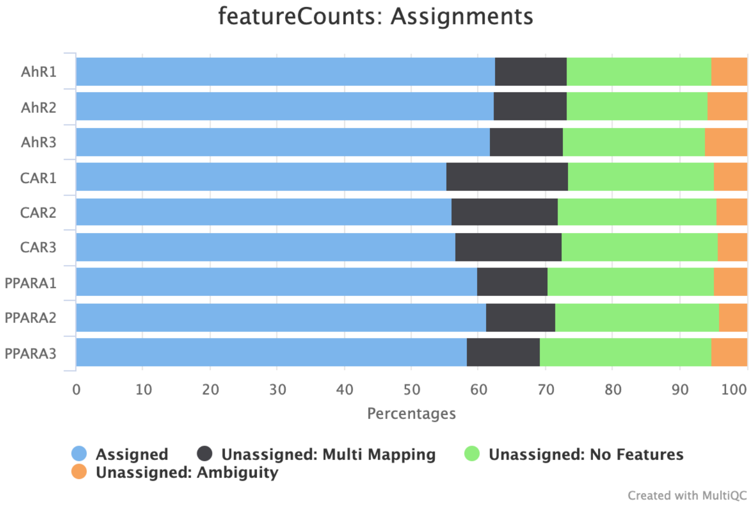
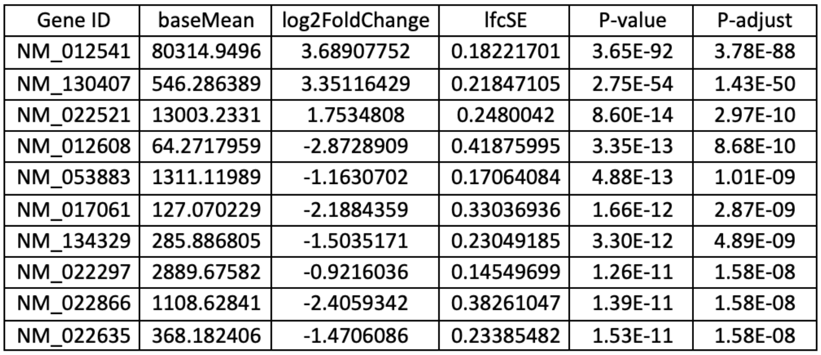


Figure 5. MultiQC summary results on the nine featureCounts summary report. (a) the top shows the number of reads (in unit million) assigned and unassigned. (b) the bottom shows the percentage distribution of the assigned and unassigned reads in each sample.

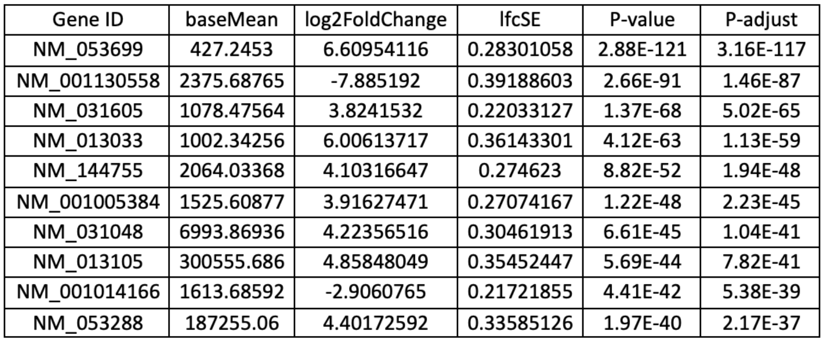
RNA-Seq Differential Expression with DESeq2

The differential expression results sorted by adjusted p-value are provided in our Github directory. The number of differentially expressed genes significant at p-adjust < 0.05 in each treatment group AhR, CAR/PXR, PPARA is 328, 3766, and 2814, respectively. The top 10 differentially expressed genes evaluated by p-value from each treatment group are shown in **Tables 3, 4, and 5**.



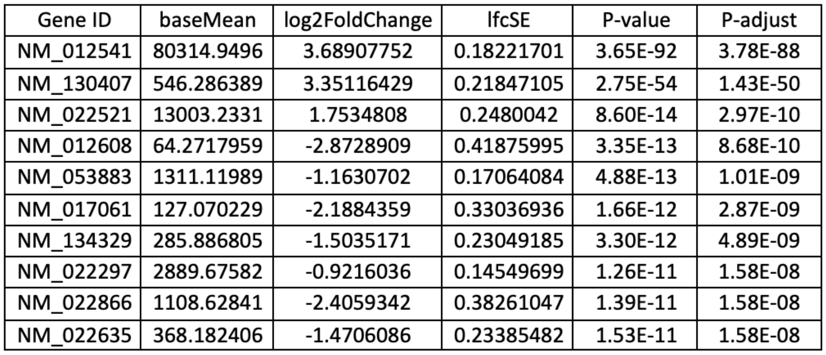
Gene ID	baseMean	log2FoldChange	lfcSE	P-value	P-adjust
NM_012541	80314.9496	3.68907752	0.18221701	3.65E-92	3.78E-88
NM_130407	546.286389	3.35116429	0.21847105	2.75E-54	1.43E-50
NM_022521	13003.2331	1.7534808	0.2480042	8.60E-14	2.97E-10
NM_012608	64.2717959	-2.8728909	0.41875995	3.35E-13	8.68E-10
NM_053883	1311.11989	-1.1630702	0.17064084	4.88E-13	1.01E-09
NM_017061	127.070229	-2.1884359	0.33036936	1.66E-12	2.87E-09
NM_134329	285.886805	-1.5035171	0.23049185	3.30E-12	4.89E-09
NM_022297	2889.67582	-0.9216036	0.14549699	1.26E-11	1.58E-08
NM_022866	1108.62841	-2.4059342	0.38261047	1.39E-11	1.58E-08
NM_022635	368.182406	-1.4706086	0.23385482	1.53E-11	1.58E-08

Table 3. Top 10 differentially expressed genes for AhR treatment, ordered by increasing p-value.



Gene ID	baseMean	log2FoldChange	lfcSE	P-value	P-adjust
NM_053699	427.2453	6.60954116	0.28301058	2.88E-121	3.16E-117
NM_001130558	2375.68765	-7.885192	0.39188603	2.66E-91	1.46E-87
NM_031605	1078.47564	3.8241532	0.22033127	1.37E-68	5.02E-65
NM_013033	1002.34256	6.00613717	0.36143301	4.12E-63	1.13E-59
NM_144755	2064.03368	4.10316647	0.274623	8.82E-52	1.94E-48
NM_001005384	1525.60877	3.91627471	0.27074167	1.22E-48	2.23E-45
NM_031048	6993.86936	4.22356516	0.30461913	6.61E-45	1.04E-41
NM_013105	300555.686	4.85848049	0.35452447	5.69E-44	7.82E-41
NM_001014166	1613.68592	-2.9060765	0.21721855	4.41E-42	5.38E-39
NM_053288	187255.06	4.40172592	0.33585126	1.97E-40	2.17E-37

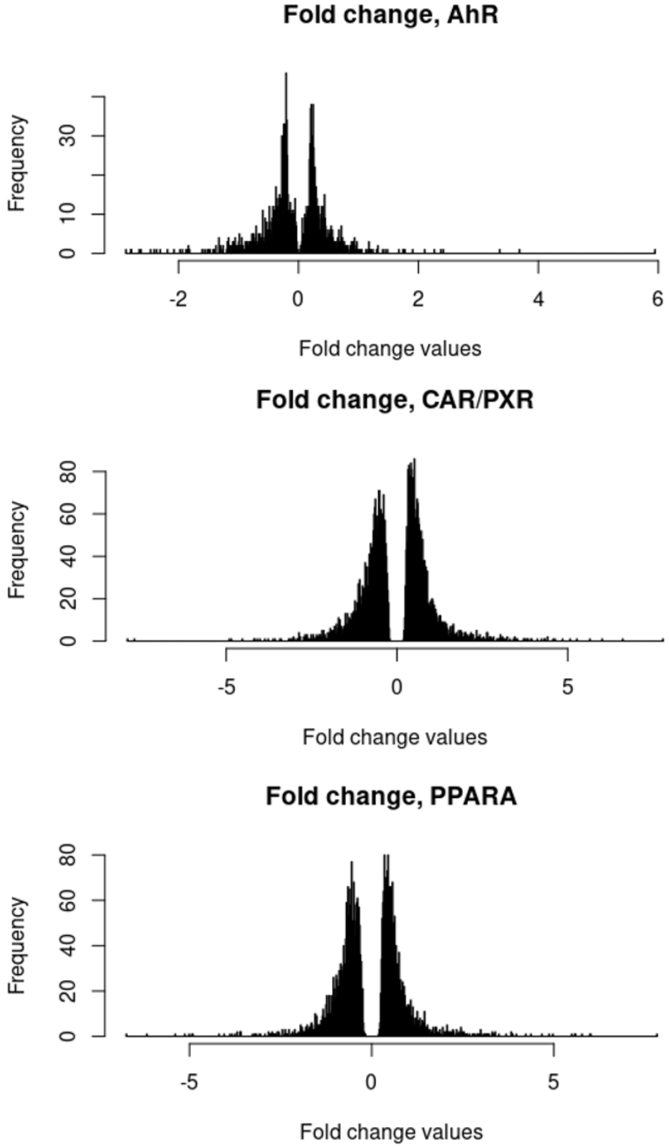
Table 4. Top 10 differentially expressed genes for CAR/PXR treatment, ordered by increasing p-value.



Gene ID	baseMean	log2FoldChange	lfcSE	P-value	P-adjust
NM_012541	80314.9496	3.68907752	0.18221701	3.65E-92	3.78E-88
NM_130407	546.286389	3.35116429	0.21847105	2.75E-54	1.43E-50
NM_022521	13003.2331	1.7534808	0.2480042	8.60E-14	2.97E-10
NM_012608	64.2717959	-2.8728909	0.41875995	3.35E-13	8.68E-10
NM_053883	1311.11989	-1.1630702	0.17064084	4.88E-13	1.01E-09
NM_017061	127.070229	-2.1884359	0.33036936	1.66E-12	2.87E-09
NM_134329	285.886805	-1.5035171	0.23049185	3.30E-12	4.89E-09
NM_022297	2889.67582	-0.9216036	0.14549699	1.26E-11	1.58E-08
NM_022866	1108.62841	-2.4059342	0.38261047	1.39E-11	1.58E-08
NM_022635	368.182406	-1.4706086	0.23385482	1.53E-11	1.58E-08

Table 5. Top 10 differentially expressed genes for PPARA treatment, ordered by increasing p-value.

The respective log2 fold change frequencies of the differentially expressed genes in each treatment group are presented in **figure 6**. Only those genes that meet the p-value significance cutoff at $p < 0.05$ were considered. The up-regulated and down-regulated genes in each treatment group are indicated in the histogram with fold change values greater than and less than zero, respectively. Samples treated with the mode of action AhR and treated with 3-methylcholanthrene seem to have the least up and down-regulated genes as the frequency only ranged from 0 to about 50 and had slightly more down-regulated genes than up-regulated genes. Samples with the mode of action CAR/PXR, treated with fluconazole showed the opposite, with more up-regulated genes than down-regulated genes. The up and down-regulated genes appear to be more even in those samples with mode of action PPARA, treated with pirinixic acid.

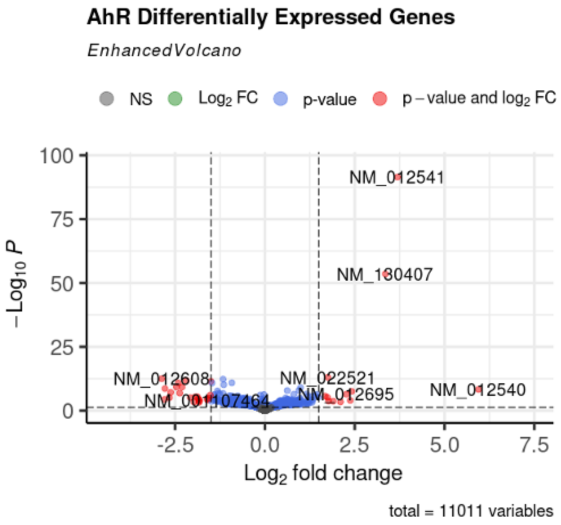

Fold change, AhR

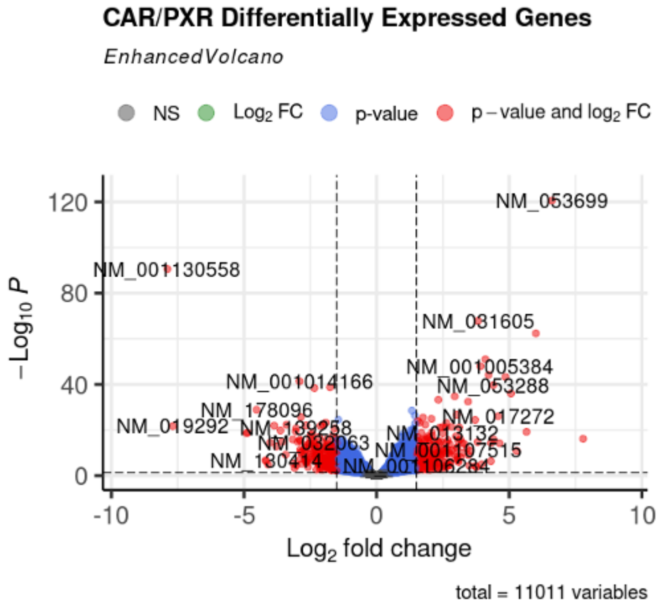
Fold change, CAR/PXR

Fold change, PPARA

Figure 6. Log2 fold change histogram for each of the treatment groups. (a) on the top, Ahr, treated with 3-methylcholanthrene, (b) middle, CAR/PXR, treated with fluconazole, (c) on the bottom, PPARA, treated with pirinixic acid. Genes with fold change value greater than zero indicate up-regulated genes and those less than zero indicated down-regulated genes.

The relationship between the log2 fold change and the p-value is also explored and evaluated through volcano plots as shown in **figure 7**. We defined a gene as differentially expressed genes (DEGs) the same way the authors in the paper did, with p-values < 0.05 and log2FC > 1.5. Genes with larger fold change appear to have a higher negative log p-value. These volcano plots provide information and visualization of genes that are insignificant and were not included in the histograms in figure 6. Many genes in the treatment groups fluconazole and pirinixic acid with mode of action CAR/PXR and PPARA, respectively, were able to meet both the p-value and log2 fold change cutoff that we set, shown as the red dots in the volcano plots. It is interesting to see that there were no genes that pass the log2FC criteria but did not pass the p-value cutoff, no green dots were shown.





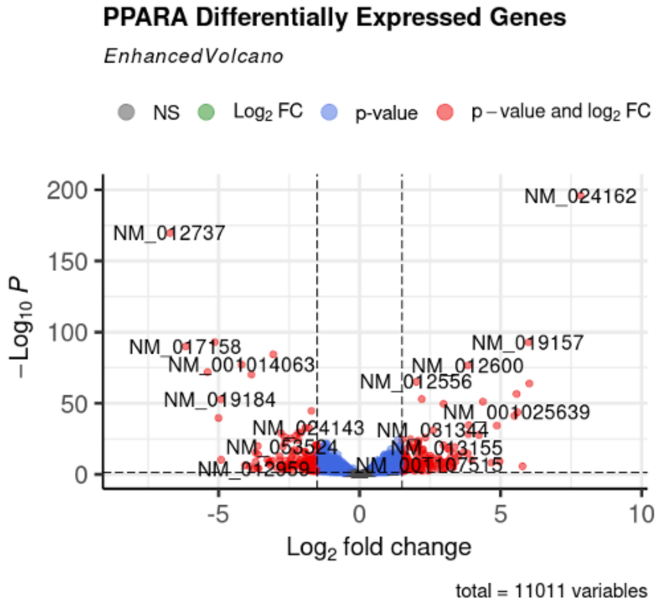


Figure 7. Volcano plot of log2 fold change versus p-value. The volcano plots have a cutoff set at 0.05 for p-value and 15 for log2FC, indicated as the vertical and horizontal dashed lines in each plot. Red points indicate those genes that meet both the p-value and log2FC criteria. Blue points indicate genes that only meet the p-value cutoff, and the grey points indicate those that meet neither criterion. The green point, not seen in either of the three plots, indicates those that only meet the log2FC criteria.

Microarray Differential Expression with limma

We load the samples that correspond to the sequencing data treatments as group_6_mic_info.csv file. The pre-normalized RMA expression matrix is used to run differential expression analysis of our samples versus the appropriate controls using the Limma Bioconductor package. We subset the RMA matrix according to the 3 chemicals we used for the analysis along with their controls. For limma analysis, we first create the design matrix. After fitting the data to the design matrix model, we run eBayes. Finally, we run toptable to extract differentially expressed genes(DEGs) for all 3 different chemical analyses. Thus we obtained 58 significant DEGs for 3methylcholanthrene, 8761 DEGs for fluconazole, and 1997 DEGs for pirinixic acid respectively. The top 10 DEGs for each chemical group are shown in **Tables 6, 7,** and **8**.

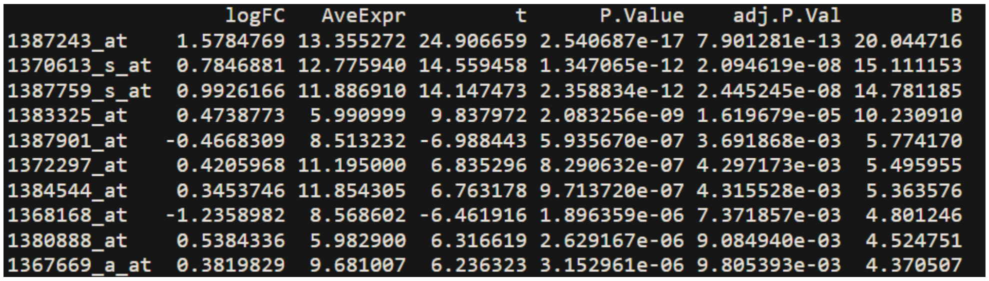
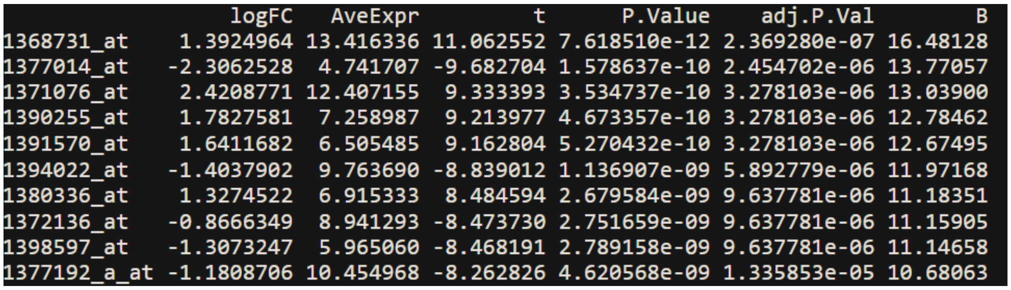
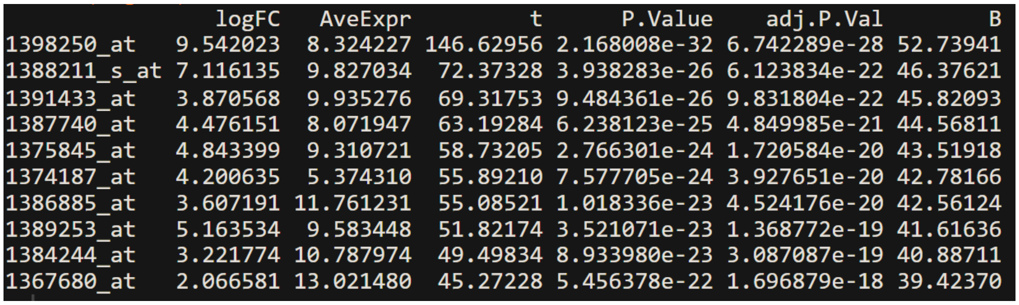

```
                  logFC   AveExpr         t      P.Value    adj.P.Val         B
1387243_at    1.5784769 13.355272 24.906659 2.540687e-17 7.901281e-13 20.044716
1370613_s_at  0.7846881 12.775940 14.559458 1.347065e-12 2.094619e-08 15.111153
1387759_s_at  0.9926166 11.886910 14.147473 2.358834e-12 2.445245e-08 14.781185
1383325_at    0.4738773  5.990999  9.837972 2.083256e-09 1.619679e-05 10.230910
1387901_at   -0.4668309  8.513232 -6.988443 5.935670e-07 3.691868e-03  5.774170
1372297_at    0.4205968 11.195000  6.835296 8.290632e-07 4.297173e-03  5.495955
1384544_at    0.3453746 11.854305  6.763178 9.713720e-07 4.315528e-03  5.363576
1368168_at   -1.2358982  8.568602 -6.461916 1.896359e-06 7.371857e-03  4.801246
1380888_at    0.5384336  5.982900  6.316619 2.629167e-06 9.084940e-03  4.524751
1367669_a_at  0.3819829  9.681007  6.236323 3.152961e-06 9.805393e-03  4.370507
```

Table 6. Top 10 genes significant at p-adjust < 0.05 for 3methylcholanthrene chemical



```
                 logFC   AveExpr         t      P.Value    adj.P.Val        B
1368731_at    1.3924964 13.416336 11.062552 7.618510e-12 2.369280e-07 16.48128
1377014_at   -2.3062528  4.741707 -9.682704 1.578637e-10 2.454702e-06 13.77057
1371076_at    2.4208771 12.407155  9.333393 3.534737e-10 3.278103e-06 13.03900
1390255_at    1.7827581  7.258987  9.213977 4.673357e-10 3.278103e-06 12.78462
1391570_at    1.6411682  6.505485  9.162804 5.270432e-10 3.278103e-06 12.67495
1394022_at   -1.4037902  9.763690 -8.839012 1.136907e-09 5.892779e-06 11.97168
1380336_at    1.3274522  6.915333  8.484594 2.679584e-09 9.637781e-06 11.18351
1372136_at   -0.8666349  8.941293 -8.473730 2.751659e-09 9.637781e-06 11.15905
1398597_at   -1.3073247  5.965060 -8.468191 2.789158e-09 9.637781e-06 11.14658
1377192_a_at -1.1808706 10.454968 -8.262826 4.620568e-09 1.335853e-05 10.68063
```

Table 7. Top 10 genes significant at p-adjust < 0.05 for fluconazole chemical



```
                logFC   AveExpr         t      P.Value    adj.P.Val        B
1398250_at   9.542023  8.324227 146.62956 2.168008e-32 6.742289e-28 52.73941
1388211_s_at 7.116135  9.827034  72.37328 3.938283e-26 6.123834e-22 46.37621
1391433_at   3.870568  9.935276  69.31753 9.484361e-26 9.831804e-22 45.82093
1387740_at   4.476151  8.071947  63.19284 6.238123e-25 4.849985e-21 44.56811
1375845_at   4.843399  9.310721  58.73205 2.766301e-24 1.720584e-20 43.51918
1374187_at   4.200635  5.374310  55.89210 7.577705e-24 3.927651e-20 42.78166
1386885_at   3.607191 11.761231  55.08521 1.018336e-23 4.524176e-20 42.56124
1389253_at   5.163534  9.583448  51.82174 3.521071e-23 1.368772e-19 41.61636
1384244_at   3.221774 10.787974  49.49834 8.933980e-23 3.087087e-19 40.88711
1367680_at   2.066581 13.021480  45.27228 5.456378e-22 1.696879e-18 39.42370
```

Table 8. Top 10 genes significant at p-adjust < 0.05 for pirinixic acid chemical

Now we create histograms of fold change values (**Figure 8**) and scatter plots (**Figure 9**) for fold changes vs p nominal value from the significant DE genes from each of our analyses.

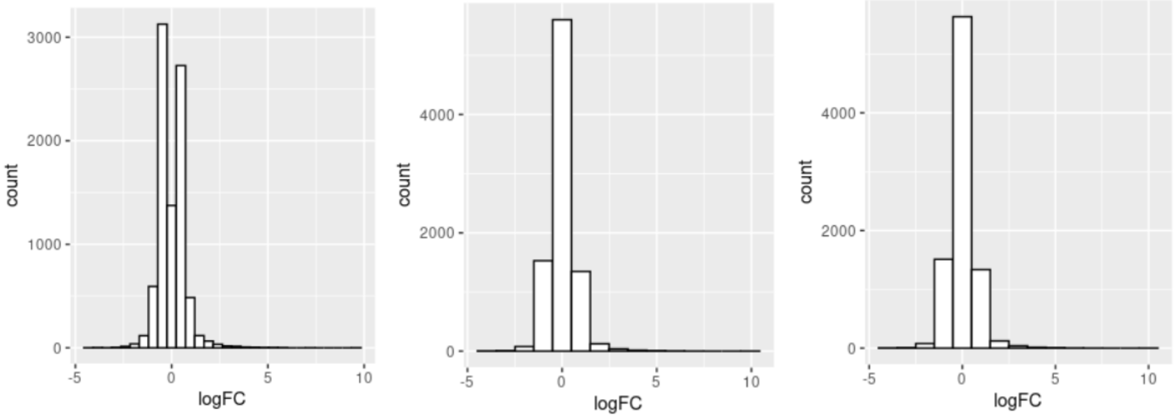


Figure 8. depicts a histogram for count vs log2Foldchange for each of the chemical analyses (3methylcholanthrene, fluconazole, and pirinixic acid) for the DEGs obtained from Microarray Differential Expression with Limma

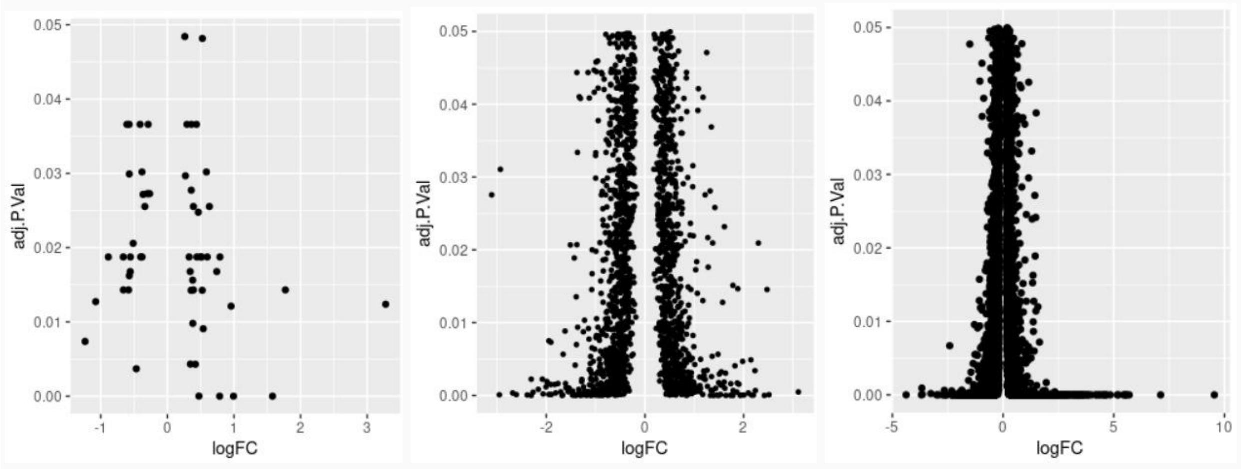


Figure 9. depicts scatter plots for Fold Change vs nominal p-value for each of the 3 chemicals (3methylcholanthrene, fluconazole, and pirinixic acid) for the DEGs obtained from Microarray Differential Expression with Limma

Now we take differential expression results from both DESeq2 and limma and we measure and examine concordance between platforms. We will load the limma results files which we saved earlier as csv files and we would map the data frame from the reference Affy map by matching the probe ids from the limma results to affy maps probe ids. For each chemical analysis, we take the median of the log2foldchange, read the DESeq2 data files and then filter DEG where Log2foldchange>1.5 and nominal p-value < 0.05. Then we Calculate concordance for each chemical analysis after obtaining counts of the number of differentially expressed genes that match between limma and deseq (**Figure 10**).

The formula to calculate the concordance is as follows

chemical_concordance = (2 * chemical_count)/(nrow(chemical) + nrow(chemical_map))

The concordance observed was 0.2303143 for 3methylcholanthrene, 0.5534543 for fluconazole, and 0.5214058 for pirinixic acid. Furthermore, we create a scatter plot with treatment on the x-axis and concordance of the y-axis for RNA-Seq analysis and Microarray Analysis.

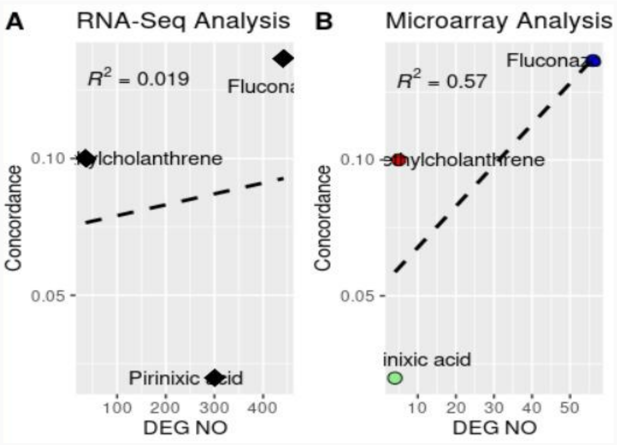


Figure 10. Concordance between RNA-seq Analysis and Microarray Analysis.

We use the DESEQ to find out the median of the base mean column in the DESeq2 results. The DE genes get divided into “above-median” and “below-median” groups. From limma results, we find the median of the average expression. We recompute concordance for each of the above and below groups (**Figure 11**). For below the median subset, concordance for 3 methylcholanthrene groups is 0.1120797, 0.3616364 for fluconazole,0.3516524 for pirinixic acid respectively. For the above median subset, concordance for the 3 methylcholanthrene group is 0.2515567,0.5152451 for fluconazole, 0.482746 for pirinixic acid respectively.

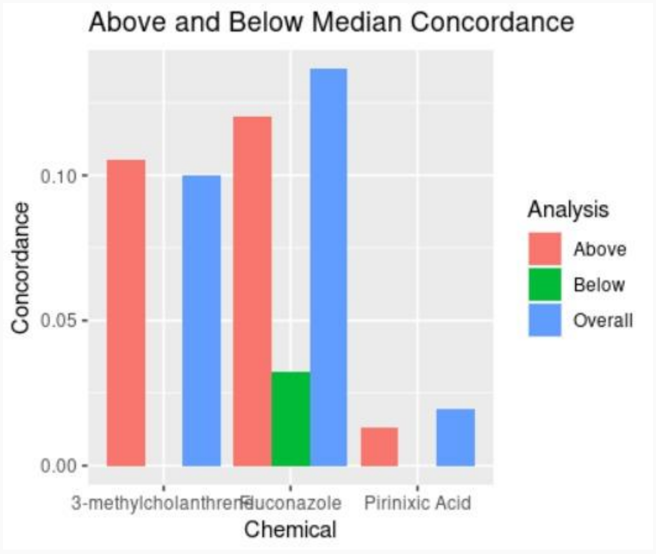
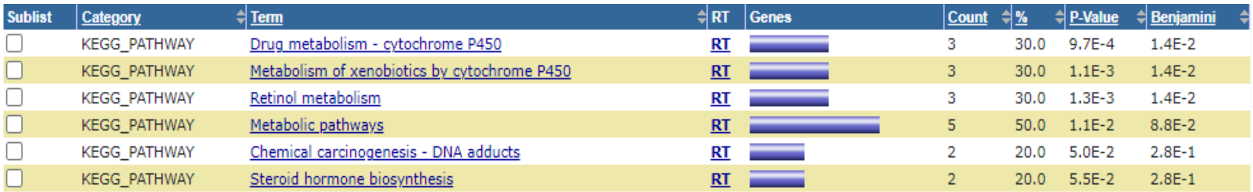


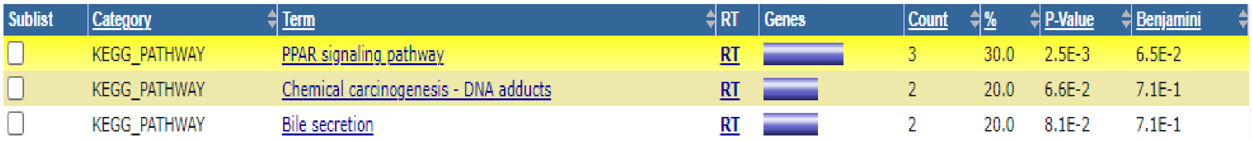
Figure 11. A bar plot combining overall concordance measures obtained for the overall DE gene list and the above- and below-median subsets.

We were able to undertake a further investigation to make biological interpretation and comparison using a determined set of normalized counts in our analysis. The paper by Wang et al. laid the groundwork for comprehending the differences between RNA-seq and microarray systems. We were able to find certain parallels in our results, which were important for indicating a biological conclusion. We tried to match our DAVID (Huang et al., 2009) results to (Wang et al., 2014) paper's analysis of MOA chemical groups that are shared by both RNA-seq and microarray platforms to evaluate the enrichment pathways of DE genes from our RNA-seq and microarray investigation.We selected the the top 10 DEG from all the three groups employed three MOA chemical groups in this comparison: CAR/PXR, AhR, and PPARA, while Wang et al specified seven MOA chemical groups enhanced pathways. We gathered our results using the DAVID gene set enrichment approach using Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways terminology. The AhR KEGG pathway analysis found 6 pathways (**Table 9**). None of them matched the original study results but similar functions. The PPARA KEGG pathway analysis found 3 pathways (**Table 10**). However, these pathways are matching with the Ahr Pathways with chemical carcinogen pathways. None of them matched the original study results. The CAR/PXR KEGG pathway analysis found 2 pathways (**Table 11**). Surprisingly, all the pathways are matching with the AHr KEGG pathways but did not match with the results of the original paper. We have found that Xenobiotic metabolism in AHr was the only pathway that found similarities without study and the original study. However, The number of differentially expressed genes we discovered in RNAseq for 3ME (AhR) was identical to the figures reported by Wang et al.



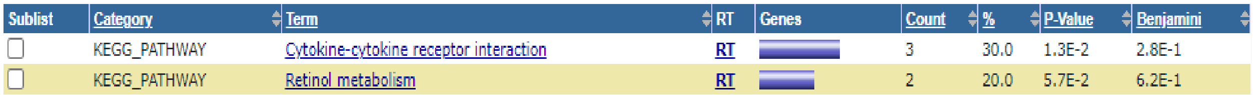
Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	KEGG_PATHWAY	Drug metabolism - cytochrome P450	RT	▬	3	30.0	9.7E-4	1.4E-2
□	KEGG_PATHWAY	Metabolism of xenobiotics by cytochrome P450	RT	▬	3	30.0	1.1E-3	1.4E-2
□	KEGG_PATHWAY	Retinol metabolism	RT	▬	3	30.0	1.3E-3	1.4E-2
□	KEGG_PATHWAY	Metabolic pathways	RT	▬	5	50.0	1.1E-2	8.8E-2
□	KEGG_PATHWAY	Chemical carcinogenesis - DNA adducts	RT	▬	2	20.0	5.0E-2	2.8E-1
□	KEGG_PATHWAY	Steroid hormone biosynthesis	RT	▬	2	20.0	5.5E-2	2.8E-1

Table 9. Ahr top DEG expressed pathway analysis.



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	KEGG_PATHWAY	PPAR signaling pathway	RT	▬	3	30.0	2.5E-3	6.5E-2
□	KEGG_PATHWAY	Chemical carcinogenesis - DNA adducts	RT	▬	2	20.0	6.6E-2	7.1E-1
□	KEGG_PATHWAY	Bile secretion	RT	▬	2	20.0	8.1E-2	7.1E-1

Table 10. PPARA top DEG expressed pathway analysis



Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	KEGG_PATHWAY	Cytokine-cytokine receptor interaction	RT	▬	3	30.0	1.3E-2	2.8E-1
□	KEGG_PATHWAY	Retinol metabolism	RT	▬	2	20.0	5.7E-2	6.2E-1

Table 11. CAR/PXR top DEG expressed pathway analysis

The heatmap was able to cluster all of the MOAs together after filtering. **Figure 12** shows heatmap clustering after filtering genes with means for coefficients of variation less than 0.186. Selecting genes with a covariant filter greater than 0.186 resulted in a heatmap that could group the MOAs together.

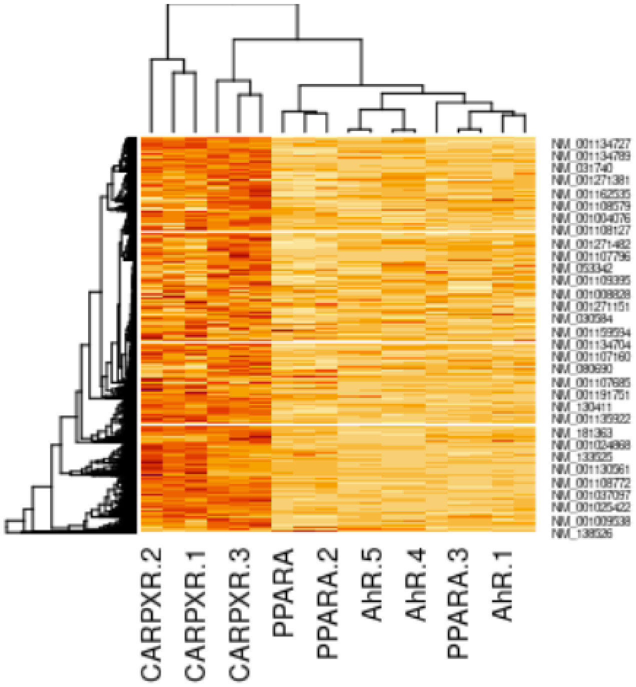


Figure 12. Heatmap and clustering of the differentially expressed genes from each member of the 3 tox groups after filtering

Discussion

The agreement in DEGs between the two platforms was linearly correlated to the treatment effect. Agreement is higher for chemicals with a marked treatment effect size (fluconazole >0.10) than for those where the response is weak (pirinixic acid i.e. less than 0.05) for both RNA Seq analysis and Microarray analysis. Limma is fairly consistent (R2 = 0.57) in determining the number of DEGs across the 3 chemicals whereas it is very less consistent (R2 =0.019)in the case of RNASeq analysis. We observed a slightly lower concordance for all chemicals except 3methylcholanthrene for the above-median expressed genes than for the below-median expressed genes. We found that the two platforms perform equally well for genes with expression levels above the median of all the assayed genes. Clearly, the discrepancy between the two platforms rests largely on the genes with below median expression. Thus in order to identify biomarkers transferable between two gene-expression measurement platforms, an emphasis should be placed on the above-median expressed genes. Thus our studies confirmed that RNA-seq performed better than microarrays at detecting weakly expressed genes which can be attributed to the fact that we used many treatment conditions and hence they covered a wide range of biological complexity.

Our pathway annotations and those of Wang et al.[5] did not yield the same results. This is most likely owing to the fact that the databases used for pathway analysis differ. DAVID[7] was employed by us, while GeneGo was used by Wang et al. The disparities in pathway analysis results would be explained if each of them had its own taxonomy and algorithm. Differences in overall DEGs between the study and our work could be due to data processing and filtering during analysis. We believe it is difficult to determine what constitutes a good concordance based on this research, as claiming that we detect between 20% and 50% overlap in differentially expressed genes across the two approaches does not appear to be significant. We discovered similar concordance between microarray and RNAseq as Wang et al, but more research is needed to determine if RNAseq, microarray, or the overlap between the two identified the most physiologically relevant genes. Overall, our data did not entirely replicate the conclusions of the paper. We were able to get to some comparable biological conclusions, nevertheless.

References

[1] Wang, C., Gong, B., Bushel, P. *et al.* The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat Biotechnol* **32,** 926–932 (2014). https://doi.org/10.1038/nbt.3001

[2] Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[3] Liao Y, Smyth GK, Shi W: featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014, 30: 923-930. 10.1093/bioinformatics/btt656.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

[4] 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.

[5] 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

[6] Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635

[7] 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. [PubMed]