Concordance of Microarray and RNA-Seq Differential Gene Expression

Group Van Gogh

Data Curator: Lindsay Wang

Programmer: Andrew Gjelsteen

Analyst: Monil Gandhi

Biologist: Elysha Sameth

INTRODUCTION

For decades, microarray technology has been the forefront of gene expression studies. Through its ability to assay large amounts of biological material and detect the expression of thousands of genes simultaneously, microarrays have deepened our understanding of transcriptomics. However, this platform is limited by its ability to detect known sequences, and background hybridization and probe saturation interfere with low and high-level detection. With the emergence of RNA sequencing (RNA-seq), limitations of microarrays have been alleviated through the rapid profiling and deep investigation of the transcriptome for any species[1]. Due to the absence of species- or transcript-specific probes, RNA-seq has the ability to detect novel transcripts, and rare and low-abundance transcripts at a higher specificity and sensitivity.

Although RNA-seq is beginning to be widely used, the concordance between RNA-seq and microarray has not been thoroughly assessed. Previous studies have compared the platforms using two distinct biological conditions, such as normal versus cancer tissue[2] and liver versus muscle samples[3], however this design reveals only one aspect of comparative characteristics. In order to evaluate the concordance between methods using similar biological conditions, Wang et al[4] verified the cross-platform results from the same set of samples. In the study, differentially expressed genes (DEGs) and enrichment pathways of liver samples of rats under perturbation by 27 chemicals with 5 different modes of action (MOA) were assessed under Affymetrix microarray and Illumina RNA-seq. In doing so, three important findings were established: 1) the concordance between the two methods in detecting number of DEGs was positively correlated to the level of perturbation caused by the chemical treatment, 2) RNA-seq is better at detecting weakly expressed genes than microarrays, and 3) the prediction models generated from the two methods have similar performance.

Our study aims to reproduce the results of Wang et al using the data from a toxicity group (toxgroup) with MOAs of aryl hydrocarbon receptor (AhR), orphan nuclear hormone receptors (CAR/PXR), DNA damage, and their corresponding vehicle controls. We assessed the quality of RNA-seq data of the experimental group, performed differential expression analysis on the microarray and RNA-seq data, and compared the results obtained from the two methods. While our replication yielded different results from the original paper, we feel that we were successful in our reproduction and arrived to similar conclusions.

DATA

Within Wang, et al, researchers used male Sprague-Dawley rats of age 6 to 8 weeks, and treatment chemicals were administered orally or through injections in 5 day maximum tolerated doses. Depending on the chemical, the selected animals were dosed for either 3, 5, or 7 days, and the liver was harvested 24 hours after the last dose. RNA samples for RNA-seq analysis were obtained from the NTP DrugMatrix Frozen Tissue Library, and sequencing was performed on Illumina HiScanSQ or HiSeq2000 systems using the manufacturer's protocol. Samples were blinded during the entire sequencing process, and ~23-25 million paired-end reads with 100 bp read length were generated. Meanwhile, the microarray data was prepared using rat liver RNA hybridized to the Affymatrix whole genome GeneChip® Rat Genome 230 2.0 Array following the technical manual.

For our study, we processed and analyzed one part of the data from the paper. We performed differential analysis on the Affymetrix microarray and RNA-seq of the toxgroup with MOAs AhR, DNA damage and CAR/PXR with specific chemical treatments (Table 1). Pre-downloaded RNA-seq data from NCBI Gene Expression Omnibus was obtained with the accessions SRP039021, GSE55347, and GSE47875. The corresponding samples were identified using the metadata file with the run ID.

MIE/MOA	Chemical Name	Abbreviation	Structure Activity Group; Therapeutic indication
AHR	LEFLUNOMIDE	LEF	Inhibits pyrimidine /purine metabolism, dihydroorotase inhibitor; Antirheumatic Disease Modifying Agent
DNA DAMAGE	IFOSFAMIDE	IFO	DNA-alkylator, nitrogen mustard; Antineoplastic
CAR/PXR	FLUCONAZOLE	FLU	Sterol 14-demethylase inhibitor, fluconazole like; Antifungal azole

Table 1. MOA summary with the corresponding chemical treatment, abbreviation, and description.

METHODS

RNA-Seq Sample Statistics and Alignment

Nine compressed FASTQ files of the treatment group were obtained from the Shared Computing Cluster[5] (SCC), and the samples corresponding to the toxgroup were identified using the toxgroup metadata. *FastQC*[6] was used to assess sample quality, and then samples were aligned against the rat genome index provided on SCC using *STAR*[7] (2.6.0c). Since the original study used 15-60 million 100bp paired-end reads, two reads were aligned together for each sample, and the alignments were saved as *bam* files. After aligning the reads, *MultiQC*[8] was used to summarize the quality statistics for *FastQC* and *STAR* output for all samples (Table 2). Overall, the percentage of uniquely aligned and multi-mapped reads are within reasonable range to proceed with further analysis.

Summary of STAR Alignment Statistics								
Sample Id	**# of Input Reads**	**Uniquely Mapped Reads #**	**Uniquely Aligned Reads %**	**Multi-mapping Reads #**	**Multi-mapping Reads %**	**Read * mapped to too many loci #**	**Read * mapped to too many loci %**	**Unmapped Reads (too many mismatch + short + other)**
SRR1178008	15156072	13348761	88.08%	533695	3.52%	38106	0.25%	0% + 8.11% + 0.05%
SRR1178009	18110504	16315003	90.09%	460617	2.54%	25765	0.14%	0% + 7.19% + 0.03%
SRR1178010	18572519	16832537	90.63%	505301	2.72%	27095	0.15%	0% + 6.47% + 0.03%
SRR1178014	17524782	14637080	83.52%	1151194	6.57%	86184	0.49%	0% + 9.20% + 0.21%
SRR1178021	17497925	14340044	81.95%	1008867	5.77%	53853	0.31%	0% + 11.92% + 0.06%
SRR1178047	17093302	14355442	83.98%	999412	5.85%	76159	0.45%	0% + 9.67% + 0.06%
SRR1177981	14229351	11410110	80.19%	537573	3.78%	17876	0.13%	0% + 15.82% + 0.09%
SRR1177982	17168681	14390023	83.82%	630047	3.67%	28961	0.17%	0% + 12.26% + 0.09%
SRR1177983	16152281	12859278	79.61%	568705	3.52%	24389	0.15%	0% + 16.63% + 0.09%

Table 2. Summary of *STAR* alignment statistics for each sample. The columns with "Read mapped to too many loci" is defined as mapping to more than 10 locations.

Read Counting with *featureCounts*

In order to quantify the number of reads per sample that matched the *gtf* file containing annotations for the rat genome, the nine *bam* files outputted from *STAR* were compared to the *gtf* using the Subread aligner's *featureCounts*[9] tool. This function outputs count files containing data for start, end, strand, length and match count for each *bam* file passed into it. The match count and gene ID data was kept for each count file and the files were merged using R[10] and exported as a *csv* file.

RNA-Seq Differential Expression with *DESeq2*

With the *featureCount* data for each sample merged into a single R dataframe, sample quality was assessed using *MultiQC*[8]. The control sample data with the same vehicle and no treatment were added. Bioconductor's *DESeq2*[11] package was used to perform analyses between each treatment group and the control in order to test for differential expression. *DESeq2* also extracts the normalized counts, which was used in downstream analysis. The top 10 differentially-expressed genes found in this step were reported along with their corresponding *logFC* and *p*-values in Table 4.

Microarray Differential Expression with *limma*

Our analysis for differentially expressed genes on microarray data was performed using the *limma*[12] package in R[10]. For each individual chemical sample, we performed empirical bayes statistics to calculate differentially expressed genes. Controls to samples were matched based on the vehicle they are in, either 100% corn oil or 100% saline. During the analysis we determined 0 significantly differentially expressed probe sets for ifosfamide, compared to the non-zero significant DE probe sets found in the paper. Therefore, we decided to filter the significant DEGs based on $|logFC| > log_2(1.5)$ and $p < 0.05$ for the microarray and concordance analysis as defined by Wang et al. Histograms of log fold change as well as volcano plots of *logFC* versus nominal *p*-values for each of the chemical samples were plotted to get a better insight on the differential expression of the genes.

Concordance between Microarray and RNA-Seq DE Genes

Concordance was calculated between microarray and RNA-seq methods using the final results obtained from the *DESeq2* and *limma* analyses. This is the percentage of DEGs commonly shared by the two platforms with agreement with the *logFC* direction. The first step in calculation of the concordance was to filter the significant DEGs based on $|logFC| > log_2(1.5)$ and $p < 0.05$ for microarray as well as RNA-seq data. RefSeq IDs from RNA-seq were then mapped to probe set IDs based on a provided mapping matrix on SCC. There were instances where the probe IDs were mapped to multiple RefSeq IDs and vice versa, therefore we collapsed those instances into multiple rows. The RefSeq IDs were then grouped and summarized based on the median values of *logFC* and average expression.

As per the reference paper, concordance was then calculated as:

$$\mathbf{2* intersect(DEG_{microarray}, DEG_{RNA\text{-}seq}) \ / DEG_{microarray} + DEG_{RNA\text{-}seq}}$$

In the above formula, **DEG** represents the number of differentially expressed genes in both microarray and RNA-seq while **intersect** represents the intersection of symbols found in both platforms. We computed the DEGs only for values which have the same direction of fold change in both the microarray and RNA-seq results.

Concordance was also computed for genes of above median and below median expression for both microarray and RNA-seq methods. We took baseMean values into consideration to compute the median for *DESeq2* results and AveExpr to compute median for *limma* results.

Gene Set Enrichment and Hierarchical Clustering

To assess pathway concordance of differentially expressed genes identified by an individual chemical treatment and treatments in its MOA, gene set enrichment analysis (GSEA) was performed for each agent and gene expression technique. Using the results from *DESeq2* and *limma*, we found that filtering by adjusted $p < 0.05$ yielded no significant genes for IFO microarray, therefore GSEA would be unable to be performed using this standard. Instead, significantly differentially expressed genes were filtered using $| logFC | > log_2(1.5)$ and $p < 0.05$ based on the study's criteria for DEGs[4]. Gene set enrichment was performed using DAVID[13] and RGD[14], and common pathways enriched for each of the MOA that were shared by both RNA-seq and microarray were determined and compared to the results of Wang et al.

To evaluate the utility of the RNA-Seq platform in predicting the MOA of a given treatment, a heatmap-based hierarchical clustering of the samples was performed. Given the normalized counts from *DESeq2*, segregation by MOA was assessed with the counts of common control samples from AhR and PXR averaged together. While *DESeq2* uses the median of ratios to normalize for sequencing depth, this was not enough to induce a clustering that is consistent with MOA. To improve clustering, genes having a variance significantly different from the median variance of all genes using a threshold of $p < 0.01$[15] were filtered. This was done using a two-tailed chi-squared test, where the test statistic (T) for each gene (G) was calculated using

$((n - 1)\ x\ var(G)\ /\ var_{med})$, where n is the number of samples, and selecting those with $T > qchisq(1 - p/2,\ n - 1)$ or $T < qchisq(p/2,\ n - 1)$. These genes were further filtered for high coefficient of variation ($sd(P)/mean(P)$> 0.406) to eliminate the lowest expression values across samples. This value was chosen by brute-forcing a coefficient of variance that would show more consistent clusters. After applying the filters, a heatmap of the normalized counts across samples was visualized using *pheatmap()* to see how well MOAs group together while simultaneously visualizing differences in gene expressions that could define these clusters.

RESULTS

RNA-Seq Sample Statistics and Alignment

Sample w/ Treatment	**%Reads aligned from our analysis**	**%Reads aligned from the paper (Ref. supplement table 3)**
AhR	91.85%	96.10%
AhR	92.77%	96.90%
AhR	93.50%	96.10%
CAR/PXR	90.58%	96.40%
CAR/PXR	88.03%	95.80%
CAR/PXR	90.28%	95.80%
DNA_Damage	84.10%	93.70%
DNA_Damage	87.66%	90.20%
DNA_Damage	83.28%	90.90%

Table 3. Comparing RNA-Seq alignment statistics to numbers reported in the paper. Note: the samples do not correspond to each other, the data provided only represent the treatment category the sample belongs to.

Compared to the alignment done in the original paper, our analysis has a lower percent alignment rate (Table 3). There are two possible reasons for the difference, the first being the software. In the original paper, authors used multiple aligners for different pipelines, of which *STAR* was not included. Furthermore, the version of the software might also affect the downstream analysis. The other reason that might result in the difference is the RNA-seq quality.

From the *FastQC* analysis, we found that most samples have significantly low Phred scores (below 20) on the 3' end. Since *STAR* aligner would clip the last few bases at each end, we did not trim the sequence output, which could also lead to inconsistency in the data.

RNA-Seq Differential Expression

The *csv* file of the combined featureCount output data for each of the nine *bam* files was assessed for each of the toxicology groups. These samples were assessed for quality using MultiQC[8], showing that the Leflunomide samples had the highest read assignment counts (Figure 2).

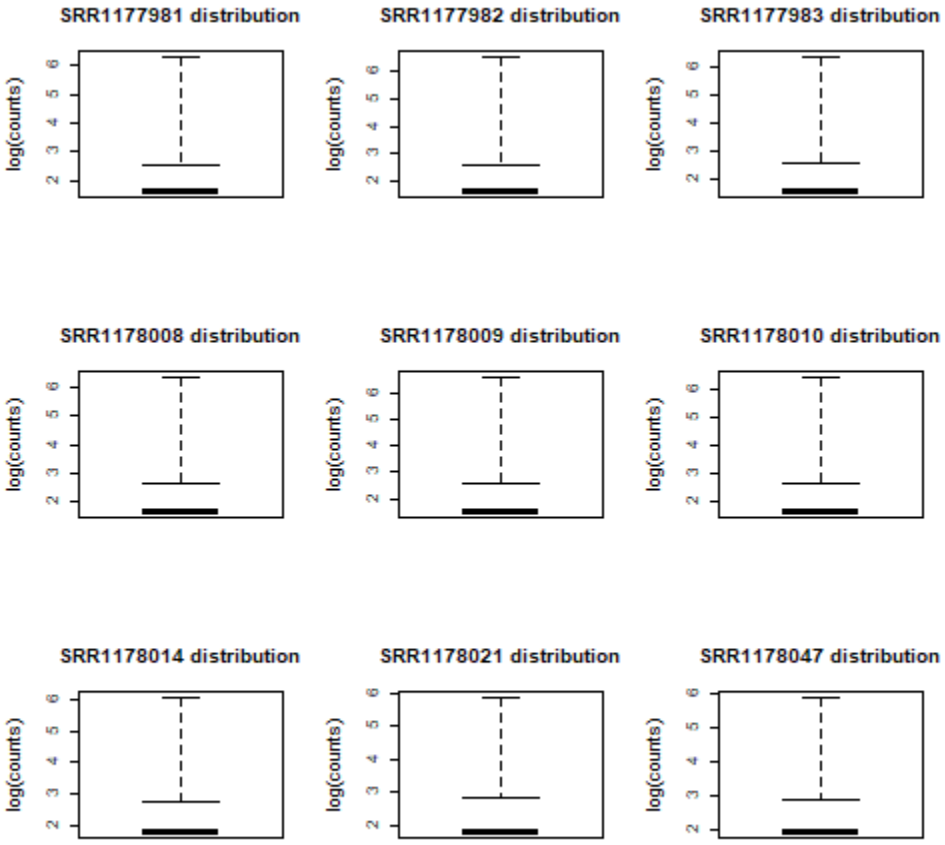


Figure 1: Box plots of each of the nine samples' count distributions across all gene IDs. The results showed that the Fluconazole (PXR) group had the lowest count mean, while the Leflunomide (AhR) appeared to have the highest count mean.

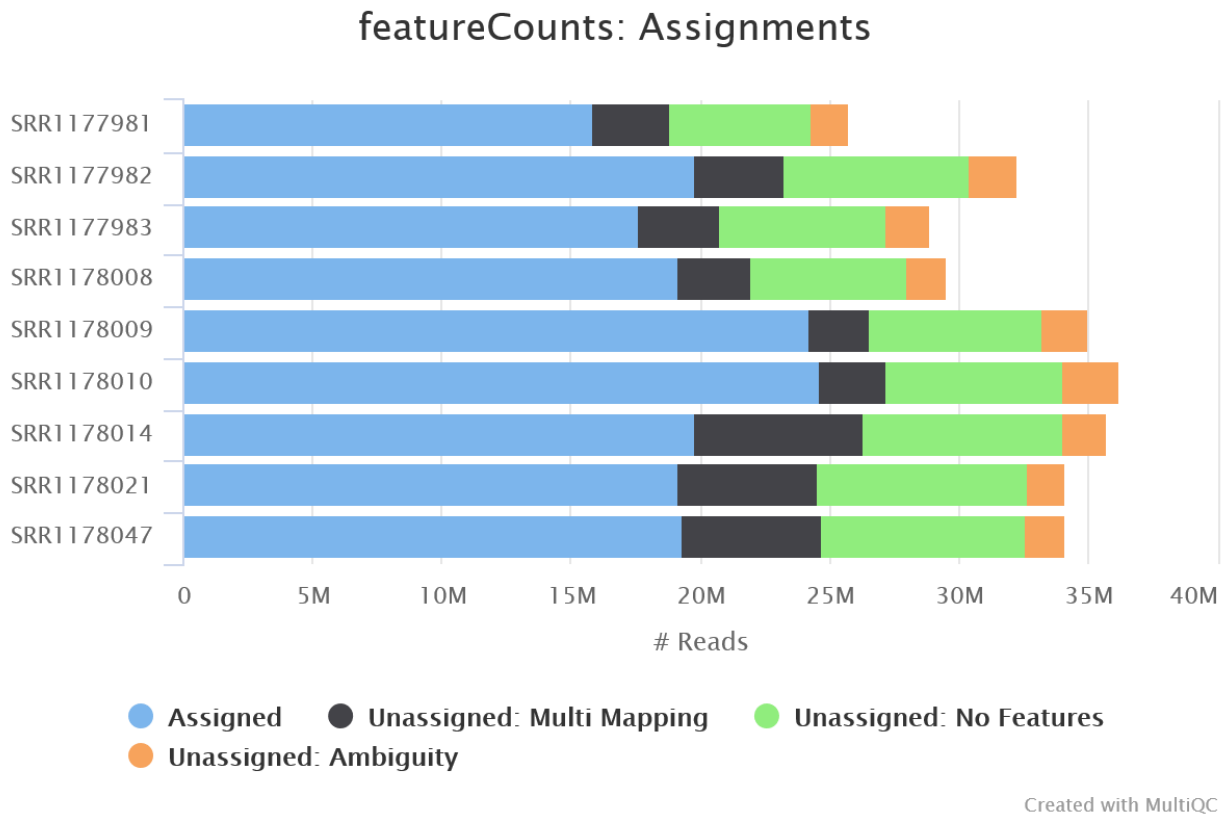


Figure 2: featureCounts assignments across the nine samples. Note that the samples corresponding to Leflunomide (SRR1178008, SRR1178009, and SRR1178010) had average assignment counts. Figure generated using MultiQC[8].

After count quality was assessed, the next step was to identify the top differentially expressed genes in each treatment group. Genes were filtered by adjusted p-values in R to identify the top 10 DE genes for each treatment group (Table 4). The Leflunomide treatment group had the lowest significance in its top DE genes, indicating a weaker effect on gene expression than the other two chemicals.

Top 10 Differentially Expressed Genes per Analysis						
Analysis	Gene ID	log_2FC	padj	Gene ID	log_2FC	padj
DNA Damage (Ifosfamide)	NM_033234	-7.24	1.78E-120	NM_013141	-3.07	1.67E-46
	NM_001007722	-7.01	1.07E-109	NM_013096	-7.59	1.83E-40
	NM_198776	-5.47	5.05E-55	NM_001109189	-1.67	1.06E-35
	NM_012543	5.77	1.15E-49	NM_024362	-2.73	2.51E-31
	NM_001109119	-2.26	2.36E-47	NM_019194	1.92	2.18E-24

AhR (Leflunomide)	NM_080783	-1.52	3.25E-12	NM_001100560	-1.27	5.05E-08
	NM_001113422	5.87	3.45E-12	NM_001106321	-0.65	5.26E-08
	NM_012623	2.23	1.72E-11	NM_012651	-1.76	8.86E-08
	NM_001134730	-1.52	1.90E-10	NM_001108692	-1.20	1.30E-07
	NM_138533	-1.51	7.94E-10	NM_134383	-1.54	4.86E-07
CAR/PXR (Fluconazole)	NM_053699	6.42	1.07E-84	NM_001005384	3.91	5.96E-43
	NM_001130558	-8.24	5.12E-83	NM_017272	4.86	3.20E-40
	NM_144755	4.37	1.06E-58	NM_013105	4.76	1.89E-38
	NM_013033	6.15	2.51E-58	NM_053288	4.41	1.55E-37
	NM_001014166	-3.05	2.89E-44	NM_019288	2.27	2.41E-34

Table 4: Top differentially expressed genes per analysis. The DNA damaging agent (Ifosfamide) resulted in the most significant treatment effect, resulting in a significant log fold decrease in 8 of its 10 most significant genes.

We found 399, 754, and 2002 significant ($p < 0.05$) differentially expressed genes for Ifosfamide, Leflunomide, and Fluconazole, respectively. A distribution of the log fold change of gene expression for each is displayed in Figure 2.

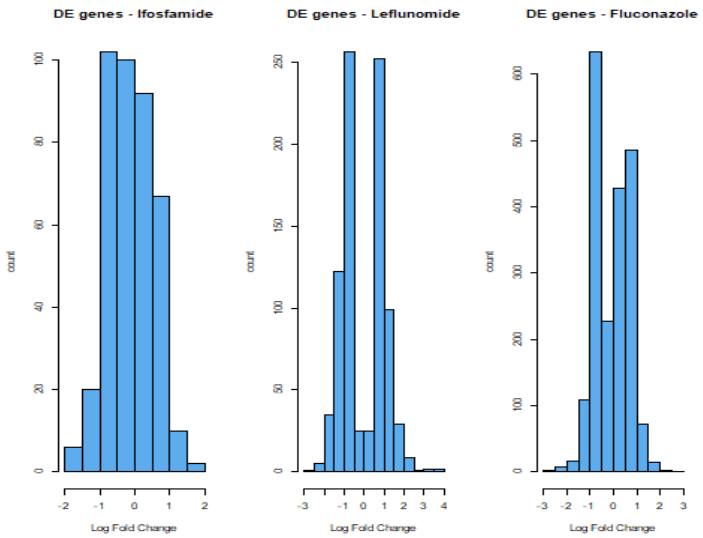


Figure 2: Distribution of *logFC* of significant DEGs for each treatment.

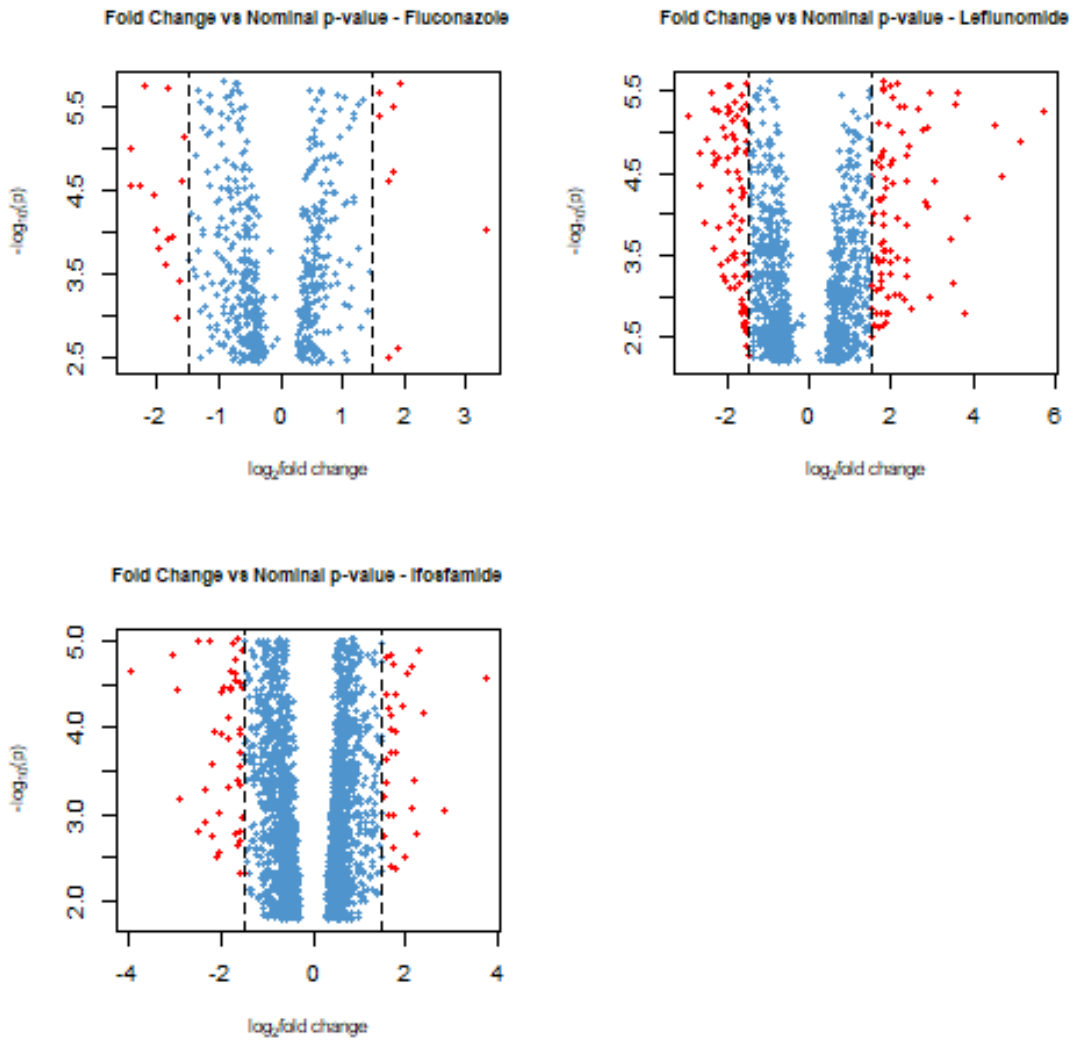


Figure 3: Scatterplot of the $\log_2$FC vs. $-\log_{10}$(p-value) for each treatment group. The treatment values with both significant p-values ($-\log_{10}$(p-value) > 13) and $|\log_2 FC| > 1.5$ are highlighted in red.

Microarray Differential Expression

MOA	Chemical Sample	Significant Probesets	Total Genes
AhR	Leflunomide	410	31099
DNA Damage	Ifosfamide	54	31099
PXR	Fluconazole	943	31099

Table 5. Summary of number of significant probe sets for each chemical sample. The significance was identified by $|logFC| > log_2(1.5)$ and unadjusted $p < 0.05$ for each chemical sample.

Based on the $|logFC| > log_2(1.5)$ and unadjusted $p < 0.05$, we found 410 significant probe sets for leflunomide, 943 significant probe sets for fluconazole and 54 significant probe sets for ifosfamide (Table 5). The probe IDs, *logFC*, as well as adjusted and unadjusted *p*-values for the top 10 most differentially expressed genes for each chemical are listed in Table 6.

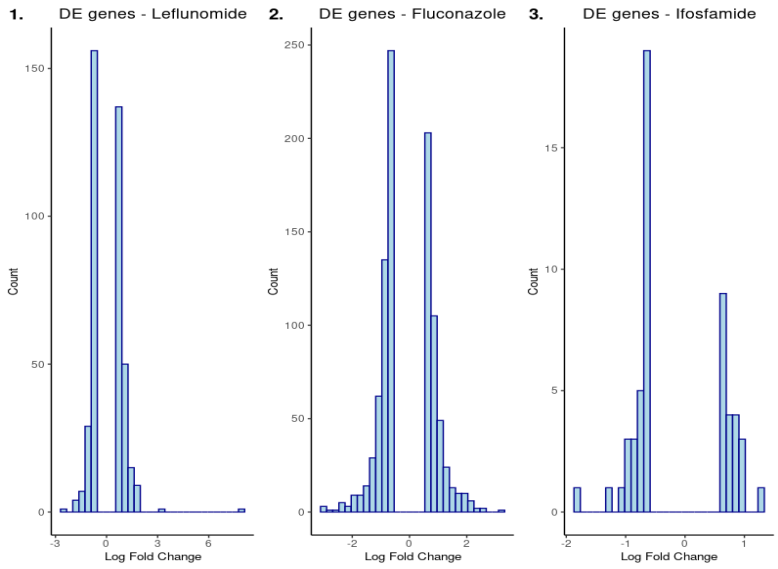
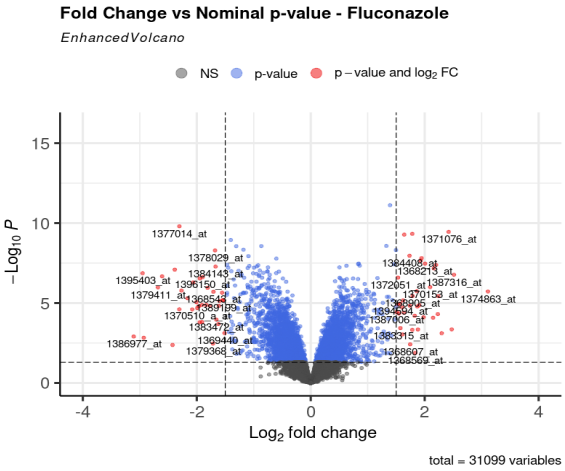


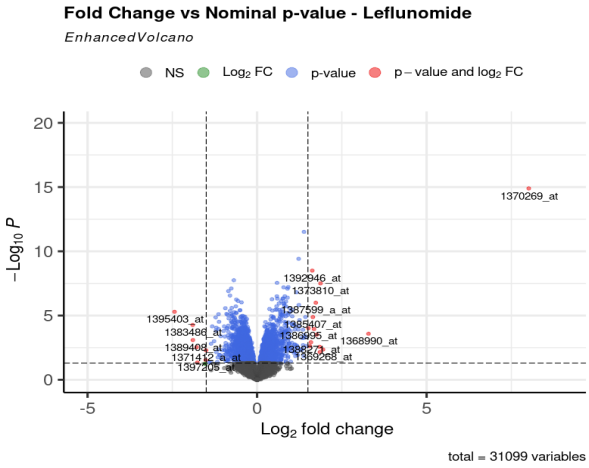
Figure 3. Histograms of log fold changes of the differentially expressed genes in **1)** Leflunomide, **2)** Fluconazole and **3)** Ifosfamide. Probe Sets are included only for $|logFC| > log_2(1.5)$ and unadjusted $p < 0.05$ for each chemical sample.

From the histograms (Figure 3), we can see that all the genes have non-zero log fold change value. The distributions seem normal for all the three chemicals despite the difference in the significant probe sets for ifosfamide as compared to leflunomide and fluconazole. It can also be observed that the log fold change scale for leflunomide extends to around 8 because of one probeset having a log fold change value of 8.013. These results suggest there is an equal number of up-regulated and down-regulated genes in the AhR MOA, the CAR/PXR may be more up-regulated, and the DNA Damage group has less differentially expressed genes, but the DEGs it contains are more down-regulated.

A.



B.



C.

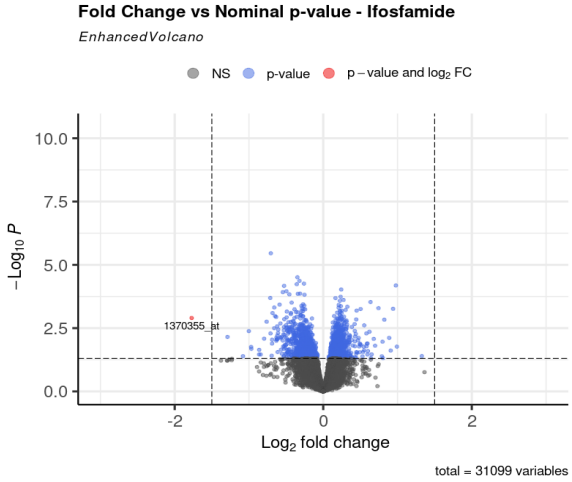


Figure 4. Volcano plots of log fold change versus -log 10 nominal p-value for each of the three chemical samples, **A)** Leflunomide, **B)** Fluconazole and **C)** Ifosfamide. The two vertical bars represent log fold change of ± 1.5 and the horizontal bar represents the p-value threshold of less than 0.05. The red dots with the probeset ids labelled, are the most significant id satisfying the log fold change threshold. NS in the plot the represents non significant genes

Volcano plots of *logFC* versus nominal *p*-values (Figure 4) reflect the sparsity of the number of significant genes in each chemical sample. Ifosfamide has only one significant probeset, whereas fluconazole has more significant probe sets than that in leflunomide. One

probeset (1370269_at), which belongs to the leflunomide sample, has a very high log fold change and is highly significant as compared to the other probe sets in all the three groups, which might indicate the chemical's significant effect on the gene.

1. AhR: Leflunomide

Probeset ID	Log FC	P-value	Adj P-value
1370269_at	8.013	1.28E-15	3.97E-11
1387243_at	1.383	3.09E-12	4.80E-08
1372600_at	1.226	3.84E-10	3.98E-06
1392946_at	1.627	3.14E-09	2.44E-05
1388611_at	-0.684	1.80E-08	0.000111724
1370244_at	0.589	2.86E-08	0.00014009
1373810_at	1.871	3.15E-08	0.00014009
1398598_at	0.919	6.07E-08	0.000221687
1376827_at	0.783	6.42E-08	0.000221687
1373814_at	-0.761	7.75E-08	0.000229011

2. CAR/PXR: Fluconazole

Probeset ID	Log FC	P-value	Adj P-value
1368731_at	1.392	7.62E-12	2.37E-07
1377014_at	-2.306	1.58E-10	2.45E-06
1371076_at	2.420	3.53E-10	3.28E-06
1390255_at	1.782	4.67E-10	3.28E-06
1391570_at	1.641	5.27E-10	3.28E-06
1394022_at	-1.403	1.14E-09	5.89E-06
1380336_at	1.327	2.68E-09	9.64E-06
1372136_at	-0.866	2.75E-09	9.64E-06
1398597_at	-1.307	2.79E-09	9.64E-06
1377192_a_at	-1.180	4.62E-09	1.34E-05

3. DNA Damage: Ifosfamide

Probeset ID	Log FC	P-value	Adj P-value
1379397_at	-0.705993025	3.51E-06	0.109265725
1374475_at	-0.347716802	3.09E-05	0.300665944
1368273_at	-0.319897142	4.30E-05	0.300665944
1371266_at	-0.259292317	5.52E-05	0.300665944
1373217_at	-0.356506295	5.82E-05	0.300665944
1368718_at	0.9779115	6.54E-05	0.300665944
1376481_at	-0.527078327	6.77E-05	0.300665944
1378596_at	0.242376338	9.44E-05	0.335540117
1377029_at	-0.495171708	0.000110862	0.335540117
1383137_at	-0.543869178	0.00011977	0.335540117

Table 6. Above are the top 10 differentially expressed genes for each of the three chemical samples **1)** Leflunomide, **2)** Fluconazole and **3)** Ifosfamide, based on the $|\ logFC\ | > log_2(1.5)$ and unadjusted $p < 0.05$ for each chemical sample

Concordance between Microarray and RNA-Seq DE Genes

Comparisons between the sets of RNA-seq and microarray were established by mapping the probe set IDs to the corresponding RefSeq IDs. The DEGs were found significant at $|logFC| > log_2(1.5)$ and unadjusted $p < 0.05$ for both microarray and RNA-seq data. For both the data sets, fluconazole seems to have the highest DEGs based on the significance threshold, followed by leflunomide and ifosfamide (Table 7).

	Leflunomide	Fluconazole	Ifosfamide
Microarray	311	659	52
RNA-seq	1950	2742	56

Table 7. Number of DEG in each of the three chemicals for both the Microarray and RNA-seq data. The results are further used to calculate the concordance between the sample chemicals for each group.

Concordance between microarray and RNA-Seq was calculated for genes of each chemical group and above- and below-median expression values by using the formula in the Methods section. In doing so, we found that the total concordance was 16.5% for leflunomide, 27.6% for fluconazole, and 7.4% for ifosfamide (Table 8). The results differ from the paper significantly, as the reported concordance values range between 20% to 60%, which may be due to an unsuccessful replication of the background correction technique that was outlined in Wang et al. In observation, leflunomide and fluconazole had higher concordance values compared to ifosfamide, as these two groups had higher numbers of significant DEGs for RNA-seq as well as microarray data.

Furthermore, we plotted the histograms for concordance of each sample for the overall DEGs as well as the DEGs for above and below median expression levels for that sample (Figure 6) . From it, we can observe that the values for the concordance for above-median expression levels are higher than that for below-median, which means that the genes with higher levels of expression values are likely to be common across the DE analysis of both the RNA-seq as well as microarray data. In comparison with the paper, we had a similarly higher concordance for the above-median expressed genes than for the below-median expressed genes (Table 8). While our concordance values differed significantly, we were able to replicate the similarity in that the above-median concordance values are greater than below-median concordance values.

	Leflunomide	Fluconazole	Ifosfamide
Total	0.165	0.276	0.074
Above	0.146	0.255	0.037
Below	0.084	0.163	0

Table 8. Calculated concordance table for chemical group based on values of both RNA-seq and microarray. Total - aggregated DEGs, Above - DEG's for above median expression level, Below - DEG's for below median expression level

An assessment of the concordance of the DEGs versus the size of the DEGs that were present in the sample (treatment effect) shows that as the DEGs increase, the concordance also increases (Figure 5). The pattern seems to be common for both sets of data where there is a linear relationship between the number of DEGs in each set versus the concordance between them. In comparison with the paper, even though the number values for the concordance may not be the same, we were able to produce the same linear increasing pattern for the relationship between the number of DEGs in each set and the concordance between them. This pattern seems to replicate for both the RNA-seq as well as the microarray platform.

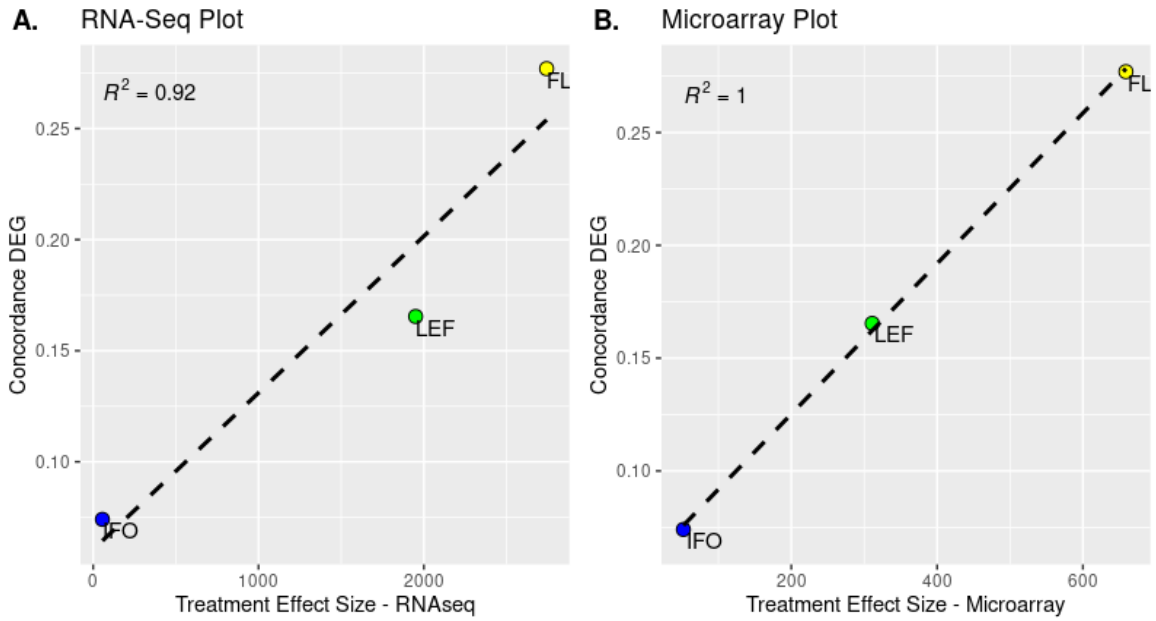


Figure 5. Scatter plot for number of DEGs found using microarray and RNA-seq analysis. The plot is a measurement of concordance for individual chemical groups as a function of treatment effect size. The plot labels for IFO is Ifosfamide, LEF is Leflunomide, and FL is Fluconazole. The dashed line is the trend line with the corresponding R-squared value

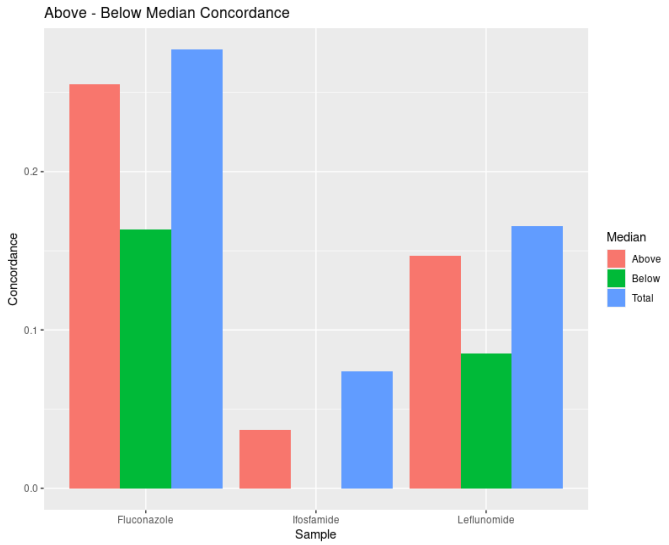


Figure 6. Plot for concordance values versus individual chemicals for both microarray and RNA-seq analysis. The 3 groups of analysis include total - all DEGs, Above median - DEGs above median expression level and Below median - DEGs below median expression level.

Gene Set Enrichment

Further analysis of the concordance of RNA-seq and microarray platforms was performed using differentially expressed genes defined by $|logFC| > log_2(1.5)$ and unadjusted $p < 0.05$. In doing so, we obtained a sufficient number of genes to be used for GSEA that could define each MOA in contrast to defining DEGs by adjusted $p < 0.05$ (Table 9).

MOA	Technique	# DEG (*adj. p < 0.05*)	# DEG (*|FC| > 1.5, p < 0.05*)	Total Genes
AhR	Microarray	466	410	31099
	RNA-Seq	1389	1950	10695
DNA Damage	Microarray	0	54	31099
	RNA-Seq	91	56	10695
PXR	Microarray	1997	943	31099
	RNA-Seq	3499	2742	10695

Table 9. Total number of differentially expressed genes for each mechanism of action (MOA) and gene expression technique. The number of significantly differentially expressed genes were identified using $|log_2FC| > log_2(1.5)$ and unadjusted $p < 0.05$, and compared to the number using adjusted $p < 0.05$.

From the GSEA, we observed little overlap between our study and the common pathways shared by both RNA-seq and microarray platforms defined in Wang et al (Table 10). Our results from DAVID reflected only one common term, "xenobiotic metabolism signaling", which was observed across all MOAs. This is as expected since our samples were exposed to chemicals and as a result, their cells elicited a xenobiotic-related response. While one annotation overlapped, several others did not. This may be due to different GSEA methods, where the study used GeneGo:MetaCore[16] for their analysis whereas we used DAVID[13] and RGD[14]. GeneGo is based on a proprietary manually curated database, supported by specific ontologies and controlled vocabulary, which may not align to our enrichment tools. As a result, inconsistencies between annotations are expected. If the DEGs described in the paper were to be made available, a comparison of their defined DEGs and ours using the same GSEA method would provide a better comparison of the enrichment terms. Despite the inconsistencies and lack of a common controlled vocabulary however, we have found that the underlying biological terms remain true. To obtain a more robust analysis, we suggest performing GSEA using various tools due to differing pathway analysis algorithms and updates to annotations. Consensus processes identified to be enriched may be used to characterize the MOA for each platform, and compared to the results of Wang et al. By using multiple GSE methods, processes may be validated or relevant annotations may be given that one tool did not provide.

2. CAR/PXR (7)
Aryl hydrocarbon receptor signaling
Glutathione-mediated detoxification
LPS/IL-1 mediated inhibition of RXR function
NRF2-mediated oxidative stress response
Nicotine degradation II
PXR/RXR activation
Xenobiotic metabolism signaling
3. AhR (10)
Acetone degradation I (to methylglyoxal)
Aryl hydrocarbon receptor signaling
Bupropion degradation
LPS/IL-1 mediated inhibition of RXR function
Melatonin degradation I
Nicotine degradation II
Nicotine degradation III
Retinoate biosynthesis I
Superpathway of melatonin degradation
Xenobiotic metabolism signaling
5. DNA damage (2)
Cell cycle: G2/M DNA damage checkpoint regulation
Xenobiotic metabolism signaling

Table 10. List of common pathways enriched for each of the MOA chemical groups that are shared by both RNA-seq and microarray platforms from Wang et al.

Hierarchical Clustering

The normalized counts from *DESeq2* contained 10695 genes that were filtered to obtain a hierarchical clustering that better represented sample clusters. The two-tailed chi-squared test resulted in 7887 genes that were further filtered based on coefficient of variance, which led to 3986 genes remaining for clustering. The dendrogram of these clusters show that each sample successfully clustered according to their MOA, however a control sample (SRRR1178004) from the 100% saline group (Figure 7) was incorrectly grouped but remained in the appropriate clade. A PCA plot of these samples (Appendix 8) shows that this is as expected since the control samples cluster with the DNA Damage MOA. Therefore, the gene expression of controls and those treated with ifosfamide is similar, while those treated with leflunomide is significantly distinct. This is supported by the number of differentially expressed genes of each MOA, which shows CAR/PXR containing the most DEGs in both microarray and RNA-seq technologies. With a significantly more number of genes considered differentially expressed compared to the other MOAs, we would expect CAR/PXR samples to have a more distinct expression signature. Similarly, with DNA Damage having less DEGs, we would expect the group to be more closely related to the control samples. Lastly, our results suggest that AhR and CAR/PXR are more closely related than with DNA Damage MOAs. This is further supported by common Pathway Ontology and GO: Biological Process Ontology terms given by RGD (Appendix 2 and 6).

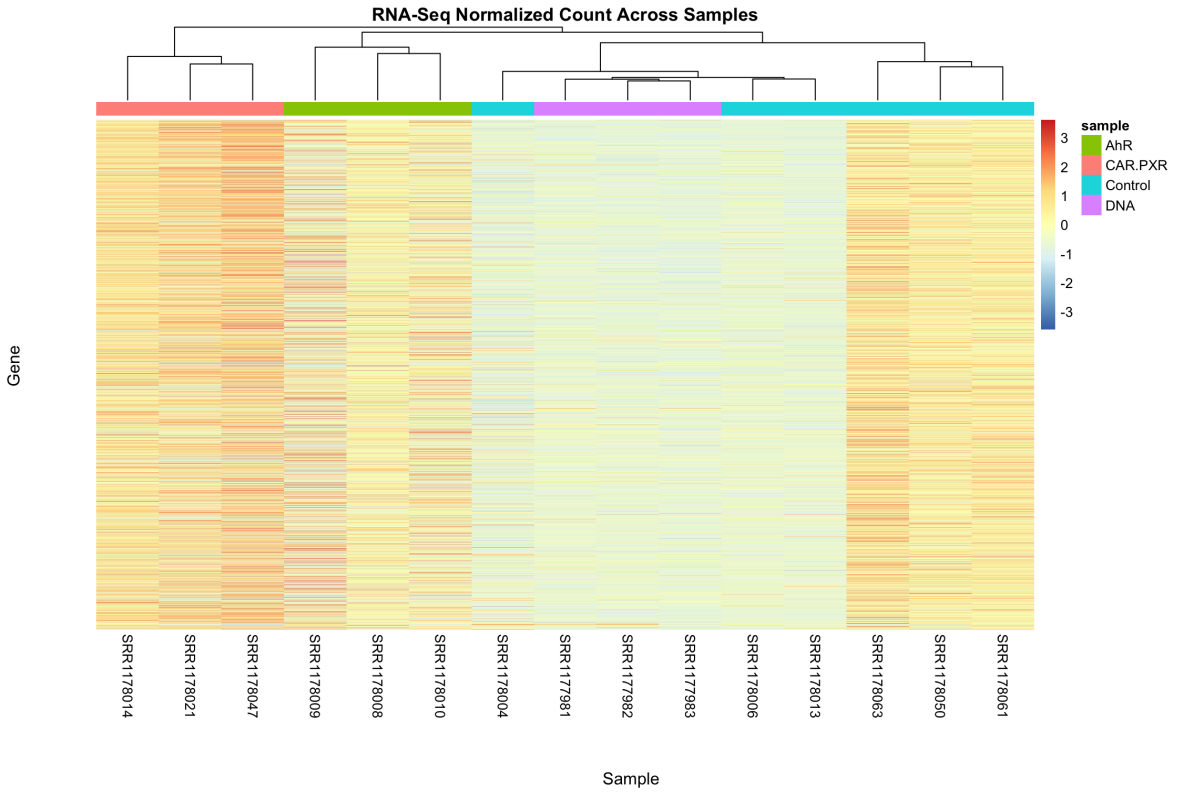


Figure 7. Heatmap hierarchical clustering of samples based on *DESeq2* normalized counts using *pheatmap*. Sample type is identified by the column colors - AhR in green, CAR/PXR in pink, DNA Damage in purple, and the controls in blue. One control sample (SRRR1178004) from the 100% saline group was incorrectly grouped but remained in the appropriate clade.

DISCUSSION

In our reproduction of Wang et al, we determined differentially expressed genes in RNA-seq and microarray platforms, calculated their concordance, and assessed pathways affected by each chemical and how MOAs cluster together. In doing so, we found that the concordance between the two methods in detecting number of DEGs was positively correlated to the level of perturbation caused by the chemical treatment, and RNA-seq is better at detecting weakly expressed genes than microarrays. Our results are therefore consistent with Wang et al. and we conclude that we were successful in our reproduction.

Although our results are similar to the paper, inconsistencies have been found in the number of significant genes, concordance, and enrichment terms. We started our limma analysis for the microarray data with significance threshold of $p < 0.05$ and found no probe sets significantly expressed in the ifosfamide samples. However, by changing our threshold to |

$logFC| > log_2(1.5)$ and unadjusted $p < 0.05$, we obtained 311, 659, and 52 DEGs in the leflunomide, fluconazole, and ifosfamide chemical samples respectively, and 1950, 2742, and 56 significant DEGs for RNA-seq data set. The relative differences in the DEGs for microarray and the RNA-seq data sets for the three groups are relatively low and might be the result of our filtering of the significant DEGs or it might be due the data itself since different alignment methods were used. Due to differences in the study design, downstream analysis may have been affected and explains inconsistencies.

Within Wang et al, it was found that concordance between platforms is dependent upon effect size and expression level. The results of our concordance analysis have varying degrees of similarity to that of Wang et al. The concordance values of our chemical samples increased as the number of DEGs increased, and observed that the more genes we were investigating the higher the similarity between the microarray and RNA-seq data sets. This indicates that RNA-seq and microarray have a high degree of similarity between them considering a large number of genes are expressed. Through our GSEA, we have also found that cross-platform concordance in terms of enriched pathways is highly correlated with treatment effect size. An analysis of the Disease Ontology, Pathway Ontology, and GO: Biological Process terms from *RGD* for each MOA and platform shows that as effect size increases, i.e. the number of DEGs increases, the percentage of microarray terms that are included in the intersection of RNA-seq and microarray terms also increases (Appendix 1). Therefore, the results of the concordance of RNA-seq and microarray datasets and enriched pathways support the paper's conclusion and suggest that when investing two similar biological conditions, a lower cross-platform concordance is expected, but a larger treatment effect would improve agreement.

Although gene enrichment analysis results did not overlap with the original study, GO terms from DAVID and RGD remain relevant to the study and reflect the biology of each chemical. For the RNA-Seq genes of the AhR MOA, DAVID gave an annotation cluster with an enrichment score of 1.5 that contains innate immunity, immunity, and innate immune response terms. Leflunomide contains immunosuppressive and anti-inflammatory properties by blocking key enzymes of de novo pyrimidine synthesis, thereby preventing the expansion of activated T lymphocytes[17]. The results from RNA-seq RGD contain GO: Biological Process Ontology (Appendix 3B) terms such as alpha-beta T cell differentiation and regulation of immature T cell

proliferation, which supports this property. This may be associated with the 'Aryl hydrocarbon receptor signaling' and 'Retinoiate biosynthesis I' pathways defined by the paper as studies have shown that AhR signaling induces retinoic acid production, which enhances T cell generation[18]. Furthermore, this chemical modulates biological processes highly relevant to tissue homeostasis and the development of pathological conditions ranging from inflammatory to neoplastic disorders[17]. An observation of the associated diseases (Appendix 3A) show leflunomide being involved in several cancers and is further supported by DAVID terms such as p53 signaling pathway and MAPK signaling pathway. Finally, the microarray terms from DAVID contain 'aging' (Appendix 2), which supports one study's findings that tobacco smoke induces extrinsic skin ageing via the AhR pathway[19]. Wang et al defined ten common pathways between RNA-seq and microarray platforms, three of which are associated with nicotine, therefore it is plausible to assume that these terms are related.

On the other hand, ifosfamide (DNA Damage) is known to prevent, inhibit or halt the development of a tumor (antineoplastic activity) by preventing DNA strand separation and DNA replication[20]. Our results from DAVID show key terms such as DNA unwinding involved in DNA replication, cellular response to DNA damage stimulus, and DNA helicase activity which reflect the mechanism of the chemical. Due to this property of ifosfamide, it is used in several forms of cancer. Our results from RGD report several cancers as top diseases identified by both RNA-seq and microarray platforms, including liver neoplasms, gastrointestinal system cancer, and hepatocellular carcinoma (Appendix 4). Furthermore, within Wang et al, one common pathway found between RNA-seq and microarray platforms was G2/M DNA damage checkpoint, which serves to prevent the cell from entering mitosis (M-phase) with genomic DNA damage. This is also reflected in the RGD Pathway and Biological Process Ontology (Appendix 4) in which we see terms related to the cell cycle pathway and DNA repair.

Finally, fluconazole (CAR/PXR) is associated with antifungal activity, inhibiting fungal cytochrome P-450 sterol C-14 alpha-demethylation[21]. Therefore, we can state that like AhR, CAR/PXR play a role in immunity, which is supported by one DAVID annotation cluster containing the same terms as AhR with an enrichment score of 0.07. We observed that CAR/PXR and AhR have similar pathway and biological process ontologies, which is a result of them being xenobiotic receptors (XRs). Although AhR is not typically classified as a nuclear receptor, it has similar functionality and follows the same overall paradigm as other XRs in a

pharmacological or toxicological perspective[22]. As a result, the common terms seen in Wang et al between the MOAs such as aryl hydrocarbon receptor signaling, LPS/IL-1 mediated inhibition of RXR function, and nicotine degradation II is expected. From our own study, we observe a possible association with the LPS/IL-1 mediated inhibition term, which has shown can lead to impaired metabolism, transport or biosynthesis of lipid, cholesterol, bile acid, and xenobiotics. Such terms include lipid biosynthetic process and lipid digestion (Appendix 7D), hypercholesterolemia and lipodystrophy (Appendix 7C), and bile secretion. Therefore, while our results do not have the exact LPS/IL-1 mediated inhibition term, we can see clear associations between GSEA methods.

Although one sample was incorrectly clustered, we believe that the hierarchical clustering was successful and have identified a set of gene expressions that may help identify each MOA. While Wang et al. performed a specific clustering analysis and we used a heatmap-based hierarchical clustering, an observation of the PCA plots of the normalized counts suggests that the control samples are similar to the DNA Damage group, therefore the inconsistency may not be a result of the clustering method. Instead, sample SRRR1178004 may have a gene expression signature more closely related to the ifosfamide treatment due to another DNA damaging condition, i.e. cancer, or there were not enough samples to provide a distinct signature for the DNA Damage group. Since hierarchical clustering was performed using fifteen samples -- three from each MOA and six controls -- including other chemicals that fall under each MOA may produce a more robust gene signature and clustering that can further assess this inconsistency.

Due to the small sample size and study focusing on one chemical for each MOA, it is not possible to conclude that chemicals with similar MOAs will induce a similar gene expression response. However, this assumption is reasonable given that all samples correctly cluster by their MOA (Appendix 9) without applying gene filters when removing the control samples from the analysis. This suggests distinct gene expressions across groups, which is reflected in the pathways and GO terms defined in the gene set enrichment analysis. The heatmap of the RNA-seq normalized counts without the control samples and *log2FC* volcano plots show that rats treated with ifosfamide have more down-regulated genes, leflunomide have more moderately expressed genes, and those treated with fluconazole are more up-regulated. Since this observation is seen within multiple samples and the clustering was successful, we can assume

that this may be true for other chemicals with the same MOA. An inclusion of other chemicals to the study with GSEA and heatmap-based clustering may confirm this assumption.

CONCLUSION

Wang et al. compared differential gene expressions obtained from RNA-seq and microarray analysis, and observed concordance between the two methods under similar biological conditions. We attempted to reproduce their results using the data from the toxicity group with MOAs of aryl hydrocarbon receptors (AhR), orphan nuclear hormone receptors (CAR/PXR), DNA damage, and their corresponding vehicle controls. Despite inconsistencies in our concordance for one MOA, GSEA, and hierarchical clustering, we feel that we have successfully arrived at the same conclusions as the paper.

In our reproduction of the study, we determined that concordance between platforms is dependent upon effect size and expression level and identified specific pathways that reflected the biology of the chemicals. However, we cannot say that chemicals with the same MOA will have the similar gene expressions, but given our clustering and GSEA we can say that it is highly likely. Overall, from the results we obtained from the replication analysis and the results in the original paper, we believe that microarrays are equally applicable to transcriptomics. Although RNA-seq performs better in identifying DEGs, especially for genes with low expression, microarrays perform almost as well to analyze pathways. As a result, we suggest that researchers choose the method according to their aim of study, taking into account effect size and expression level, as there are advantages and disadvantages to both platforms.

REFERENCES

[1] “RNA-Seq vs Microarrays: Compare Technologies.” RNA-Seq vs Microarrays, Illumina,www.illumina.com/science/technology/next-generation-sequencing/microarray-rna-seq-comparison.html.

[2] Bradford, J.R., Hey, Y., Yates, T. et al. A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. BMC Genomics 11, 282 (2010). https://doi.org/10.1186/1471-2164-11-282.

[3] Xu, W. et al. “Human transcriptome array for high-throughput clinical studies.” Proceedings of the National Academy of Sciences 108 (2011): 3707-3712. ©2011 by the National Academy of Sciences.

[4] 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.

[5] “SCC Quick Start Guide.” BU TechWeb RSS, www.bu.edu/tech/support/research/system-usage/scc-quickstart/.

[6] Babraham Bioinformatics. FastQC (Version 0.11.9) [Program documentation]. Retrieved March 3, 2021, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[7] Dobin, Alexander et al. “STAR: ultrafast universal RNA-seq aligner.” Bioinformatics (Oxford, England) vol. 29,1 (2013): 15-21. doi:10.1093/bioinformatics/bts635

[8] 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://doi.org/10.1093/bioinformatics/btw354

[9] Yang Liao, Gordon K. Smyth, Wei Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, Bioinformatics, Volume 30, Issue 7, 1 April 2014, Pages 923–930, https://doi.org/10.1093/bioinformatics/btt656

[10] “The R Project for Statistical Computing.” R, www.R-project.org/.

[11] Anders, S., Huber, W. Differential expression analysis for sequence count data. Genome Biol 11, R106 (2010). https://doi.org/10.1186/gb-2010-11-10-r106

[12] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, Gordon K. Smyth, limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Research, Volume 43, Issue 7, 20 April 2015, Page e47, https://doi.org/10.1093/nar/gkv007

[13] 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.

[14] Smith JR, Hayman GT, Wang SJ, Laulederkind SJF, Hoffman MJ, Kaldunski ML, Tutaj M, Thota J, Nalabolu HS, Ellanki SLR, Tutaj MA, De Pons JL, Kwitek AE, Dwinell MR, Shimoyama ME. The Year of the Rat: The Rat Genome Database at 20: a multi-species knowledgebase and analysis platform. Nucleic Acids Res. 2020 Jan 8;48(D1):D731-D742. doi: 10.1093/nar/gkz1041. PMID: 31713623; PMCID: PMC7145519.

[15] Marisa, L., De Reyniès, A., Duval, A., Selves, J., Gaub, M. P., Vescovo, L., . . . Boige, V. (2013). Gene expression classification of colon cancer into Molecular Subtypes: Characterization, validation, And Prognostic Value. *PLoS Medicine, 10*(5). doi:10.1371/journal.pmed.1001453

[16] “MetaCore A Cortellis Solution.” Clarivate, portal.genego.com/.

[17] National Center for Biotechnology Information. "PubChem Compound Summary for CID 3899, Leflunomide" PubChem, https://pubchem.ncbi.nlm.nih.gov/compound/Leflunomide. Accessed 5 April, 2021.

[18] Rothhammer, V., Quintana, F.J. The aryl hydrocarbon receptor: an environmental sensor integrating immune responses in health and disease. Nat Rev Immunol 19, 184–197 (2019). https://doi.org/10.1038/s41577-019-0125-8

[19] Ono Y, Torii K, Fritsche E, Shintani Y, Nishida E, Nakamura M, Shirakata Y, Haarmann-Stemmann T, Abel J, Krutmann J, Morita A. Role of the aryl hydrocarbon receptor in tobacco smoke extract-induced matrix metalloproteinase-1 expression. Exp Dermatol. 2013 May;22(5):349-53. doi: 10.1111/exd.12148. PMID: 23614742.

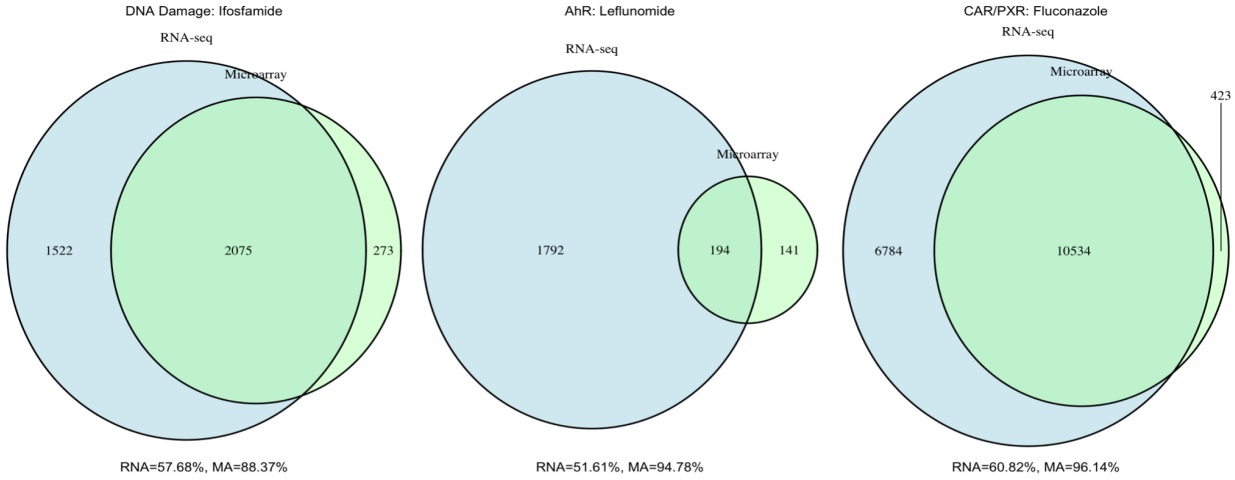
[20] National Center for Biotechnology Information. PubChem Compound Summary for CID 3690, Ifosfamide. https://pubchem.ncbi.nlm.nih.gov/compound/Ifosfamide. Accessed Apr. 5, 2021.

[21] National Center for Biotechnology Information. "PubChem Compound Summary for CID 3365, Fluconazole" PubChem, https://pubchem.ncbi.nlm.nih.gov/compound/Fluconazole. Accessed 5 April, 2021.

[22] Mackowiak, Bryan, and Hongbing Wang. “Mechanisms of xenobiotic receptor activation: Direct vs. indirect.” Biochimica et biophysica acta vol. 1859,9 (2016): 1130-1140. doi:10.1016/j.bbagrm.2016.02.006

[23] Blighe K, Lun A (2020). PCAtools: PCAtools: Everything Principal Components Analysis. R package version 2.2.0, https://github.com/kevinblighe/PCAtools.

APPENDIX



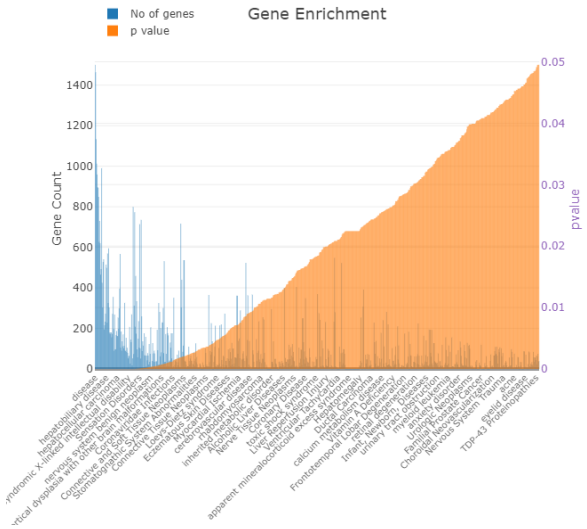
Appendix 1. Venn diagram of the number of common terms (Disease Ontology, Pathway Ontology, GO: Biological Process) from *RGD* shared between microarray and RNA-seq for each chemical. Blue represents the number of RNA-seq terms, light green the number of microarray terms, and green is the number of terms in common between the two. Below is the percentage of shared annotations belonging to each technology, i.e. the proportion of shared to the total for RNA-Seq or microarray.

AhR			
DAVID			
Microarray		**RNA-Seq**	
Term	**P-Value**	**Term**	**P-Value**
Response to organic cyclic compound	4.9e-11	Response to drug	6.6e-16
Aging	4.8e-9	Metabolic pathways	1.3e-11
Response to drug	1.7e-8	Liver development	9.5e-10
Cytoplasm	1.7e-7	Oxidation-reduction process	1.0e-9
Acetylation	2.5e-7	Response to organic cyclic compound	3.8e-9
RGD			
Microarray		**RNA-Seq**	
Disease Ontology			
Term	**P-Value**	**Term**	**P-Value**
Hepatobiliary disease	3.47e-30	Nervous system disease	2.07e-118

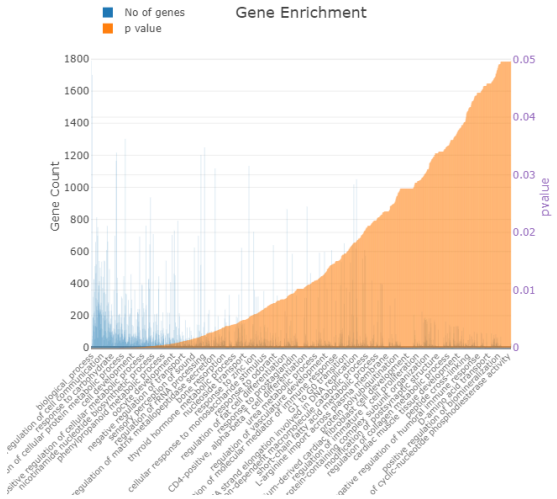
Liver disease	9.40e-30	Developmental disease	3.24e-88
Liver cirrhosis	1.53e-25	Central nervous system disease	1.32e-67
Experimental liver cirrhosis	2.13e-24	Congenital, hereditary, and neonatal diseases and abnormalities	3.07e-60
Fibrosis	9.07e-21	Brain disease	1.25e-54
Pathway Ontology			
Term	**P-Value**	**Term**	**P-Value**
Regulatory pathway	2.80e-30	Signaling pathway	3.68e-182
Signaling pathway	1.80e-15	Pathway pertinent to DNA replication and repair, cell cycle, maintenance of genomic integrity, RNA and protein biosynthesis	1.52e-121
Pathway pertinent to DNA replication and repair, cell cycle, maintenance of genomic integrity, RNA and protein biosynthesis	6.42e-13	Disease pathway	1.64e-98
Phase I biotransformation via cytochrome P450	2.58e-10	Homeostasis pathway	9.87e-95
Phase I biotransformation pathway	2.58e-10	Classic metabolic pathway	8.12e-71
GO: Biological Process Ontology			
Term	**P-Value**	**Term**	**P-Value**
Response to oxygen-containing compound	1.34e-21	Detection of chemical stimulus involved in sensory perception	2.87e-64
Cellular response to chemical stimulus	2.27e-21	Detection of chemical stimulus involved in sensory perception of smell	3.60e-63
Small molecule metabolic process	1.86e-19	Sensory perception of smell	4.63e-60
Response to organic cyclic compound	4.22e-19	Sensory perception of chemical stimulus	6.29-60
Response to organic substance	4.47e-18	Detection of stimulus involved in sensory perception	4.06e-56

Appendix 2. Table of enriched pathways for AhR identified with the DE genes from each of the analyses. For the MOA, the top 5 terms from DAVID and the top 5 terms for Disease Ontology, Pathway Ontology and GO: Biological Process from RGD are shown with their p-values.

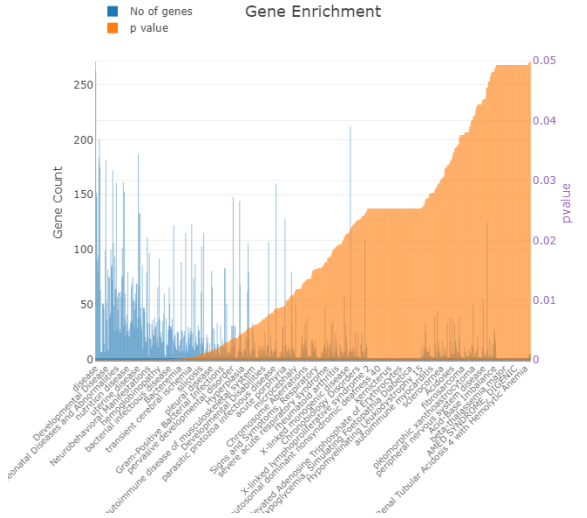
A. **Disease Ontology (RNA-seq)**



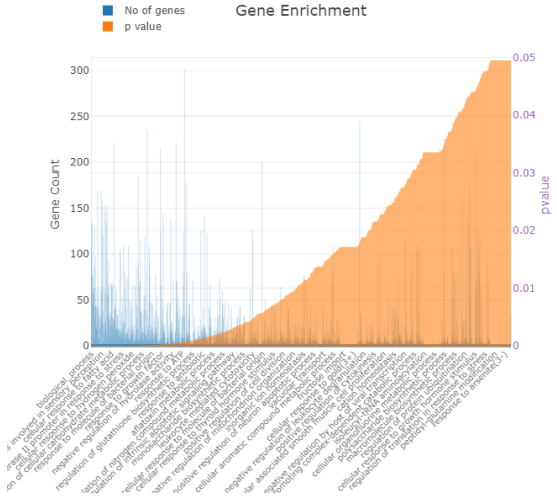
B. **GO: Biological Process (RNA-seq)**



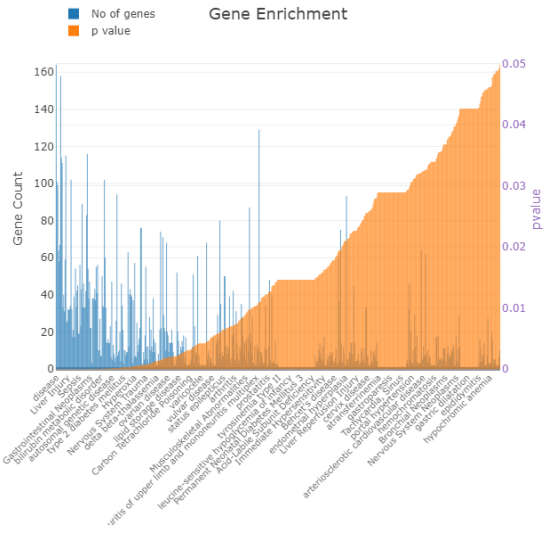
C. **Disease Ontology (Microarray)**



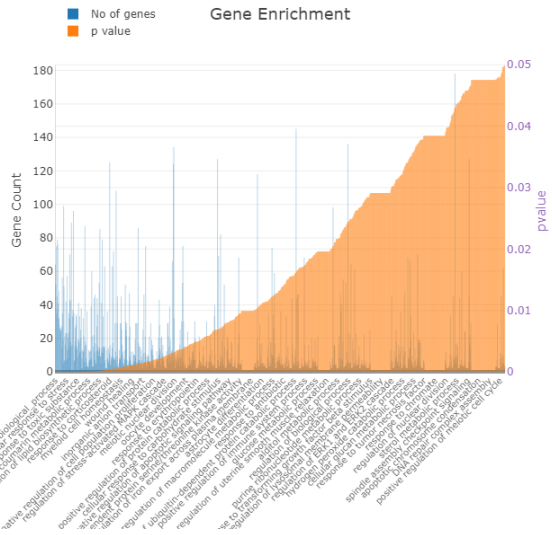
D. **GO: Biological Process (Microarray)**



E. **Disease Ontology (Common)**



F. **GO: Biological Process (Common)**



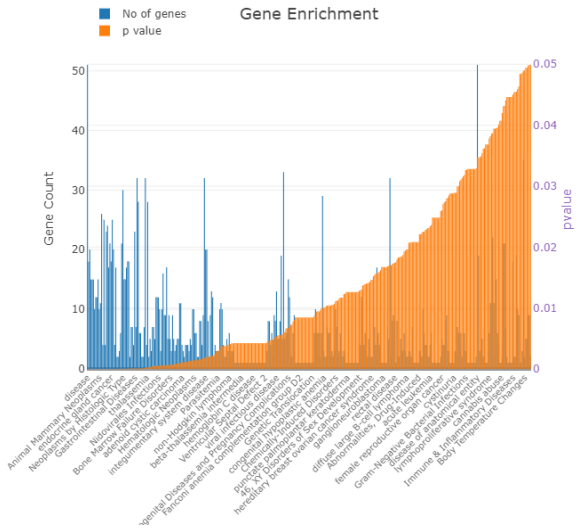
Appendix 3. Terms for RNA-seq and microarray genes for leflunomide using *RGD* with the number of genes in blue and the p-value in orange. Significance cutoff is 0.05 with **A and B)** Disease Ontology and GO: Biological Processes Ontology terms for RNA-seq data only, **C and D)** for microarray data only, and **E and F)** genes in common between RNA-seq and microarray.

DNA Damage			
DAVID			
Microarray		**RNA-Seq**	
Term	**P-Value**	**Term**	**P-Value**
Circadian rhythm	4.0e-8	Chromosome segregation	3.1e-8
Core promoter sequence-specific DNA binding	6.0e-5	Erythrocyte development	6.0e-7
Regulation of circadian rhythm	6.7e-5	Response to estradiol	2.2e-5
Response to fatty acid	1.9e-3	ATP binding	3.4e-5
Circadian rhythm	2.9e-3	Cell cycle	1.1e-4
RGD			
Microarray		**RNA-Seq**	
Disease Ontology			
Term	**P-Value**	**Term**	**P-Value**

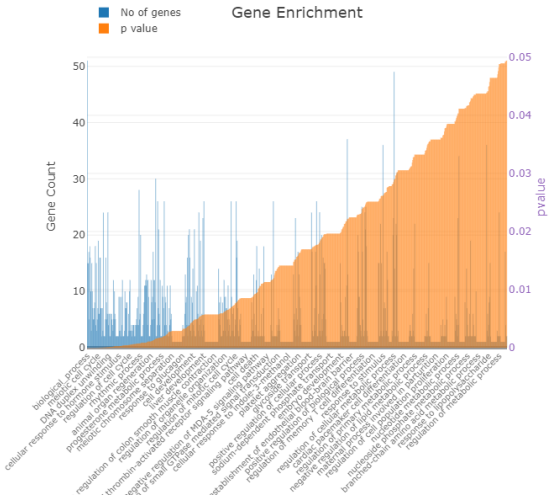
Hirschsprung's disease 1	6.75e-25	Liver neoplasms	1.73e-8
Thrombocytosis	9.92e-17	Gastrointestinal system cancer	7.86e-8
Hirschsprung's disease	8.23e-16	Hepatocellular carcinoma	1.35e-7
Megacolon	8.23e-16	Liver carcinoma	1.35e-7
Bone marrow cancer	3.59e-11	Liver cancer	1.37e-7
Pathway Ontology			
Term	**P-Value**	**Term**	**P-Value**
Regulatory pathway	1.03e-9	Regulatory pathway	1.16e-4
Pathway pertinent to DNA replication and repair, cell cycle, maintenance of genomic integrity, RNA and protein biosynthesis	1.17e-4	Glycogen storage disease type VII pathway	3.66e-4
Signaling pathway	3.21e-4	Fanconi syndrome pathway	3.66e-4
Hypertension pathway	4.37e-4	Cell cycle pathway, mitotic	4.93e-4
Homeostasis pathway	1.29e-3	Cell cycle pathway	5.45e-4
GO: Biological Process Ontology			
Term	**P-Value**	**Term**	**P-Value**
Homophilic cell adhesion via plasma membrane adhesion molecules	3.10e-13	Mitotic cell cycle process	7.28e-10
Cell-cell adhesion via plasma-membrane adhesion molecules	6.17e-11	Mitotic cell cycle	1.06e-8
Cell-cell adhesion	4.10e-9	Cell cycle process	1.17e-8
Cell adhesion	9.26e-9	Erythrocyte development	1.07e-7
Biological adhesion	1.06e-8	Cell cycle	1.18e-7

Appendix 4. Table of enriched pathways for DNA Damage identified with the DE genes from each of the analyses. For the MOA, the top 5 terms from DAVID and the top 5 terms for Disease Ontology, Pathway Ontology and GO: Biological Process from RGD are shown with their p-values.

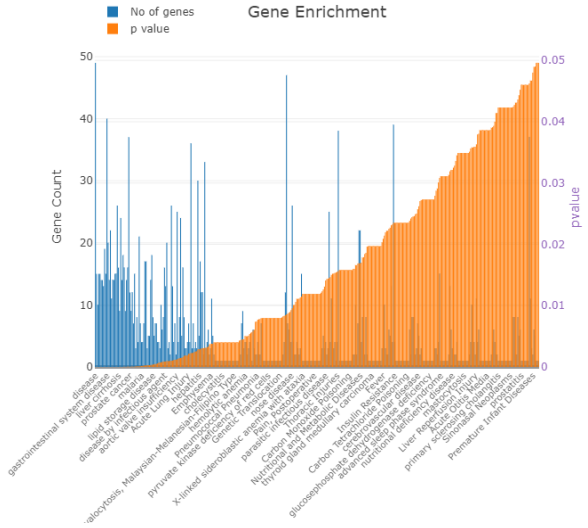
A. **Disease Ontology (RNA-seq)**



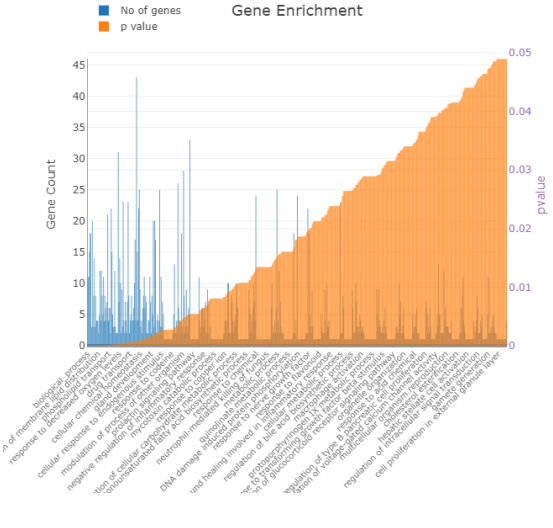
B. **GO: Biological Process (RNA-seq)**



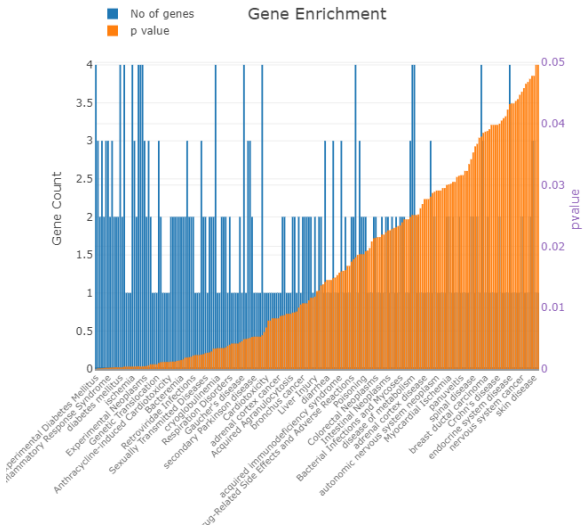
C. **Disease Ontology (Microarray)**



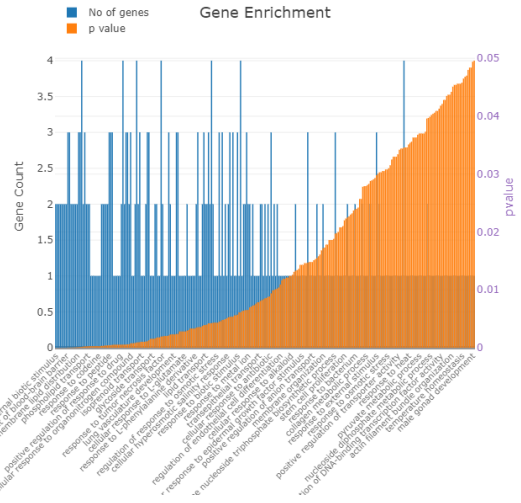
D. **GO: Biological Process (Microarray)**



E. **Disease Ontology (Common)**



F. **GO: Biological Process (Common)**



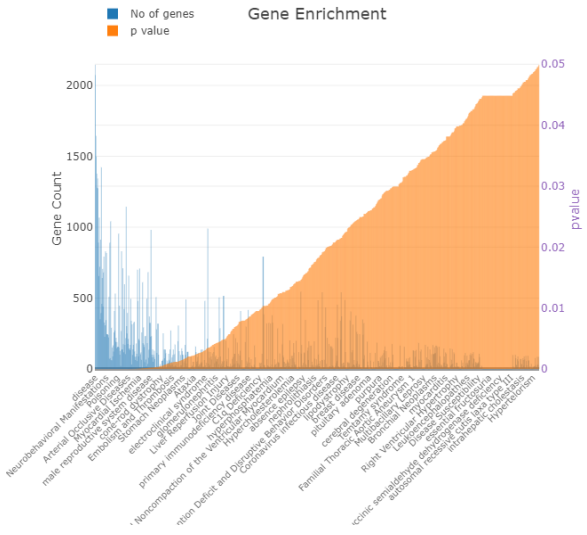
Appendix 5. Terms for RNA-seq and microarray genes for ifosfamide using *RGD* with the number of genes in blue and the p-value in orange. Significance cutoff is 0.05 with **A and B)** Disease Ontology and GO: Biological Processes Ontology terms for RNA-seq data only, **C and D)** for microarray data only, and **E and F)** genes in common between RNA-seq and microarray.

CAR/PXR			
DAVID			
Microarray		**RNA-Seq**	
Term	**P-Value**	**Term**	**P-Value**
Response to drug	3.2e-9	Metabolic pathways	1.0e-19
Heme binding	9.1e-9	Oxidation-reduction process	3.2e-19
Oxidation-reduction process	1.6e-8	Complement and coagulation cascades	1.3e-17
Regulation of cell cycle	2.0e-8	Chemical carcinogenesis	5.5e-12
Response to organic substance	3.7e-8	Retinol metabolism	1.6e-11
RGD			
Microarray		**RNA-Seq**	
Disease Ontology			
Term	**P-Value**	**Term**	**P-Value**
Disease of anatomical entity	4.65e-93	Nervous system disease	9.5e-159

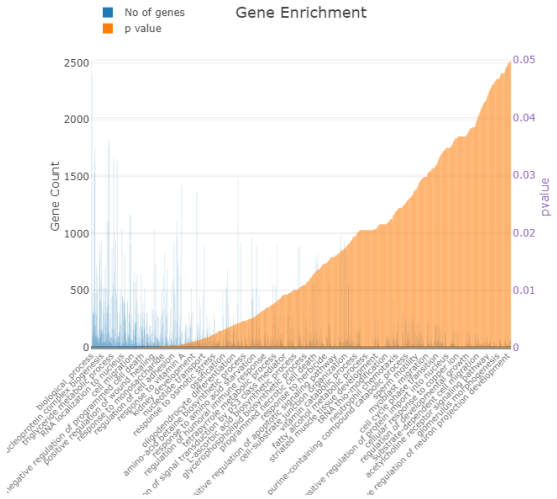
Experimental liver cirrhosis	1.54e-37	Developmental disease	5.57e-105
Nervous system disease	8.85e-36	Central nervous system disease	1.05e-93
Liver cirrhosis	2.66e-33	Brain disease	9.29e-76
Liver disease	1.92e-28	Congenital, hereditary and neonatal diseases and abnormalities	1.91e-64
Pathway Ontology			
Term	**P-Value**	**Term**	**P-Value**
Regulatory pathway	9.97e-87	Pathway pertinent to DNA replication and repair, cell cycle, maintenance of genomic integrity, RNA and protein biosynthesis	1.82e-160
Signaling pathway	1.54e-41	Disease pathway	2.50e-157
Homeostasis pathway	6.61e-28	Homeostasis pathway	9.89e-153
Pathway pertinent to DNA replication and repair, cell cycle, maintenance of genomic integrity, RNA and protein biosynthesis	1.32e-24	Transcription pathway	5.78e-107
Disease pathway	8.16e-24	Classic metabolic pathway	1.89e-99
GO: Biological Process Ontology			
Term	**P-Value**	**Term**	**P-Value**
Small molecule metabolic process	2.26e-25	Detection of chemical stimulus involved in sensory perception of smell	1.24e-92
Organic acid metabolic process	2.86e-22	Sensory perception of smell	7.77e-90
Detection of chemical stimulus involved in sensory perception of smell	5.57e-22	Detection of chemical stimulus involved in sensory perception	1.33e-86
Detection of chemical stimulus involved in sensory perception	1.69e-21	Sensory perception of chemical stimulus	1.97e-85
Oxoacid metabolic process	7.38e-21	Detection of chemical stimulus	3.47e-82

Appendix 6. Table of enriched pathways for CAR/PXRidentified with the DE genes from each of the analyses. For the MOA, the top 5 terms from DAVID and the top 5 terms for Disease Ontology, Pathway Ontology and GO: Biological Process from RGD are shown with their p-values.

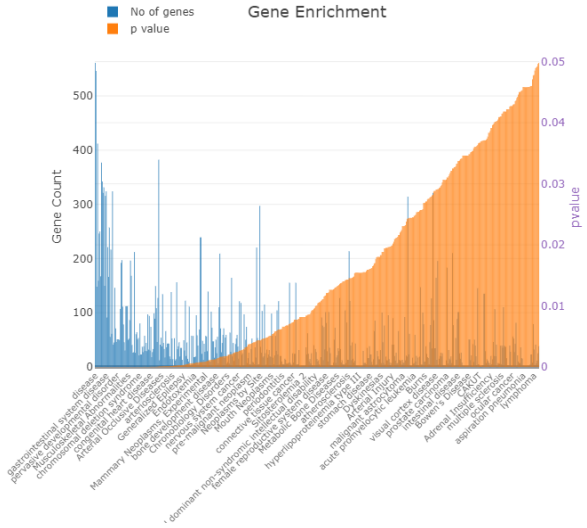
A. **Disease Ontology (RNA-seq)**



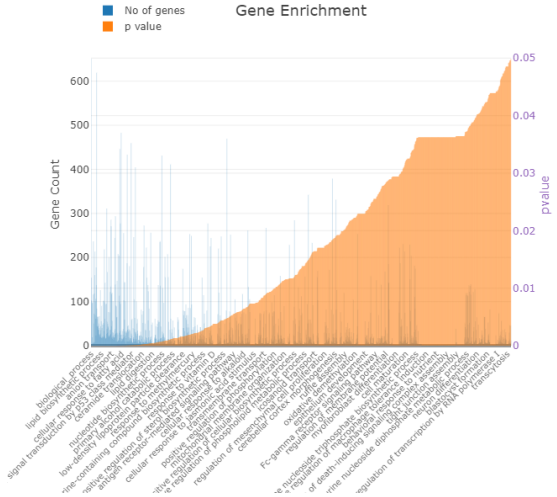
B. **GO: Biological Process (RNA-seq)**



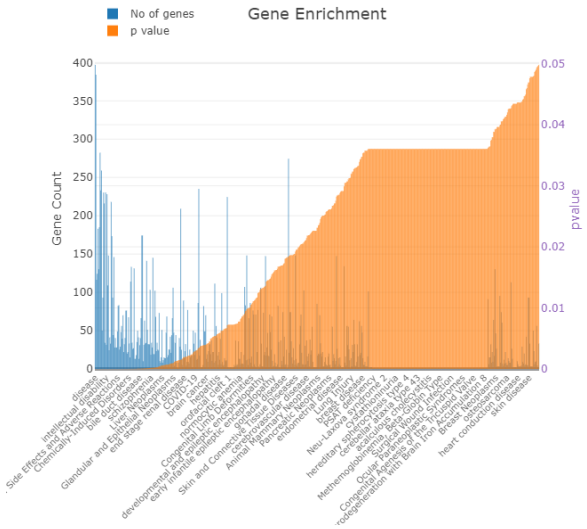
C. **Disease Ontology (Microarray)**



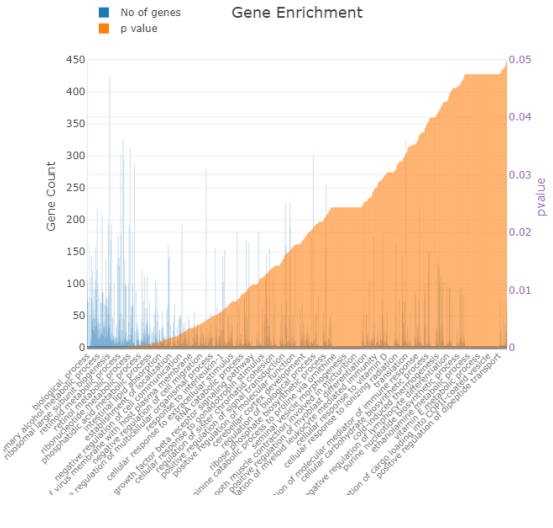
D. **GO: Biological Process (Microarray)**



E. **Disease Ontology (Common)**

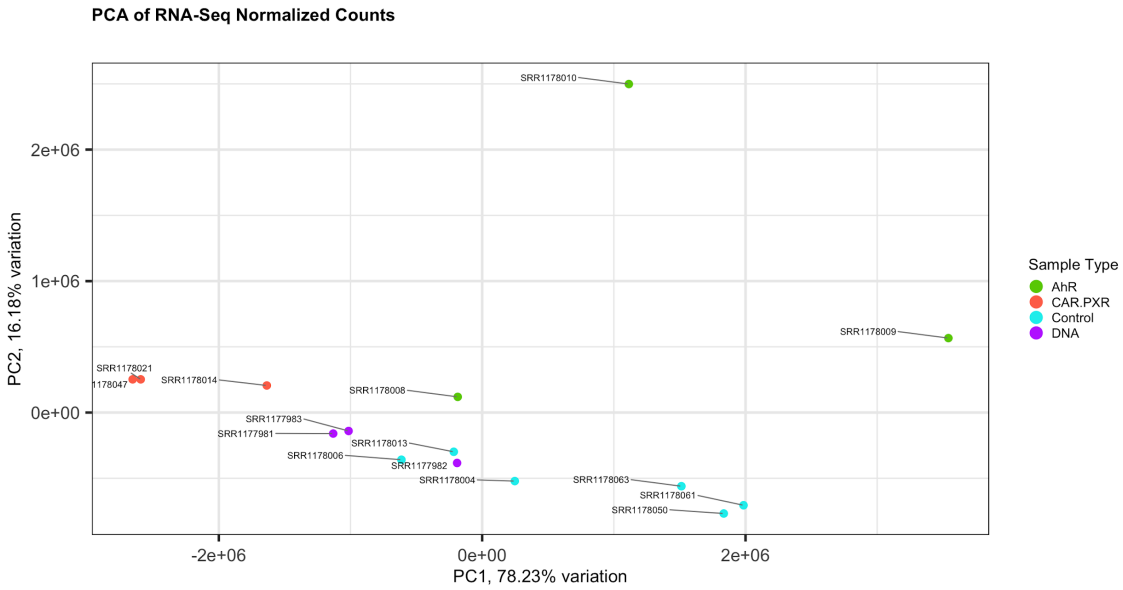


F. **GO: Biological Process (Common)**

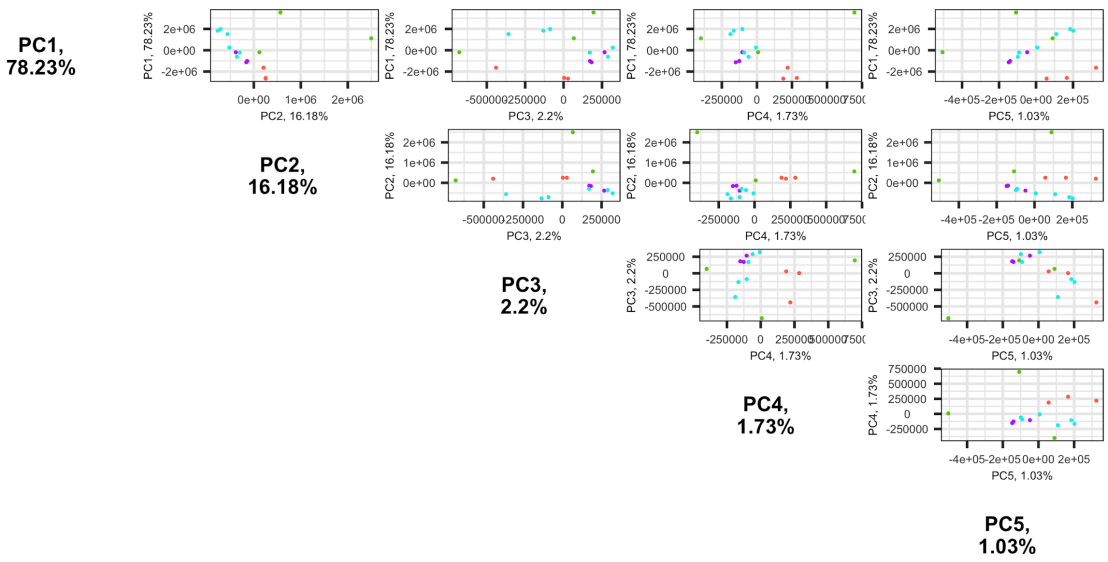


Appendix 7. Terms for RNA-seq and microarray genes for fluconazole using *RGD* with the number of genes in blue and the p-value in orange. Significance cutoff is 0.05 with **A and B)** Disease Ontology and GO: Biological Processes Ontology terms for RNA-seq data only, **C and D)** for microarray data only, and **E and F)** genes in common between RNA-seq and microarray.

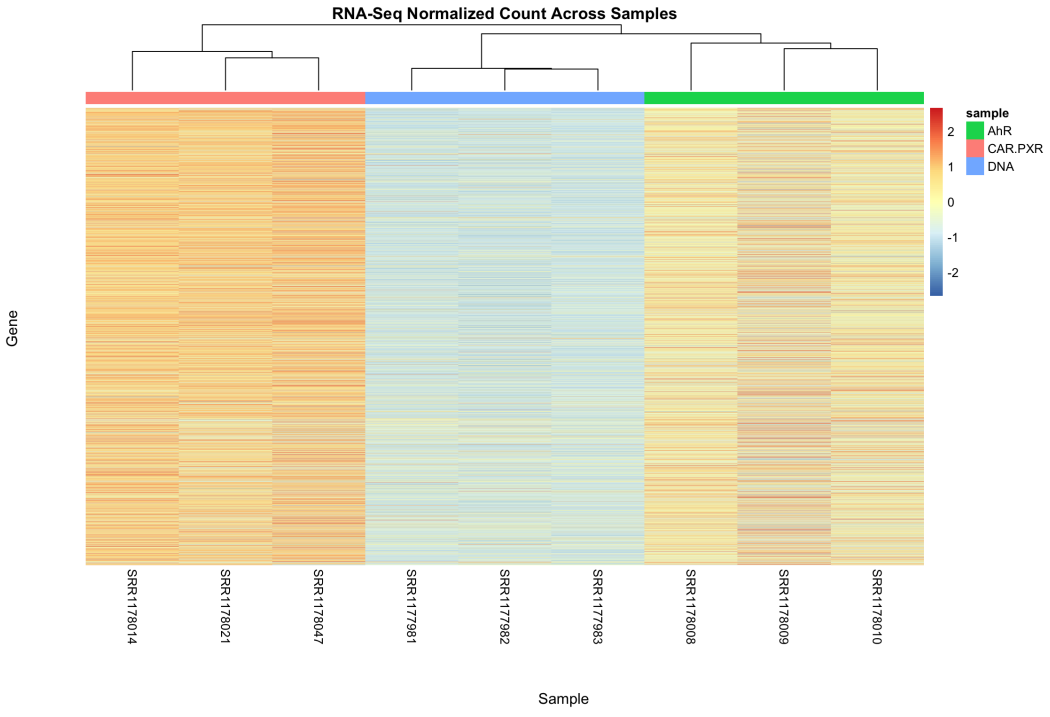
A.



B.



Appendix 8. PCA and pairs plot of the normalized RNA-seq samples made using the *PCAtools*[22] package. Sample type is identified by the column colors -- AhR in green, CAR/PXR in pink, DNA Damage in purple, and the controls in blue.



Appendix 9. Heatmap-based hierarchical clustering of samples without the control group. Sample type is identified by the column colors -- AhR in green, CAR/PXR in pink, DNA Damage in purple, and the controls in blue.