Concordance of Microarray and RNA-Seq Differential Gene Expression

Project 3: Team wheeler

Data Curator: Reina Chau
Programmer: Vishala Mishra
Analyst: Jessica Fetterman
Biologist: Ariel Xue

Introduction

RNA-seq and microarray are two fundamentally different approaches to measure the mRNA abundance in order to evaluate the gene-expression profile. To ensure their successful uses in clinical and regulatory applications, data from these two approaches need to be in good concordance. Comparison of the results from microarrays and RNA-seq from several recent studies showed that RNA-seq exhibits lower precision for weakly expressed genes, owing to the nature of sampling, while other studies have suggested that RNA-seq has higher sensitivity for gene detection[1]. To further analyze the concordance of differential gene expression across platforms, Wang *et al* presented a comprehensive study design between data from Illumina RNA-seq and Affymetrix microarray using the same samples from rats under various toxicological treatments with known mode of action (**MOA**) and assessed the MOA prediction accuracy of each platform. Our study objective was to reproduce the concordance comparison between two platforms using 3 of the toxicological treatments, assessing how RNA-seq and microarray gene expression patterns performed across chemical treatment effects, above- and below-median expressed genes, and pathway enrichment. STAR was utilized to perform the sequence alignment as it can detect alternative splicing in RNA-seq data. To investigate how genes are differentially expressed (**DEGs**), DESeq2 as a good Bioconductor package using negative binomial regression was used to estimate the count differences between treatment and control groups. We used the limma, a Bioconductor package, to identify DEGs in the microarray datasets and calculated the concordance across microarray and RNA-seq by chemical exposure. To further identify which pathways were enriched, DEGs obtained from DESeq2 were analyzed by GATHER, a gene annotation tool that could help identify related KEGG pathways.

Methods

RNA-Seq Sample Statistics and Alignment

Our team selected toxgroup 3 to study the concordance in gene expression between microarray and RNA-seq data. There were 15 samples in the RNA-seq toxgroup: 3 treatment conditions (Fluconazole, Ifosamide, and Leflunomide) and two *vehicle* controls (Corn Oil 100% and Saline 100%). Each treatment condition had 3 replicates, so in total, we have nine samples to process. In order to map each sample against the rat reference genome, we utilized STAR, an alignment program similar to tophat that is designed specifically for spliced alignment for RNA-seq data.

One useful feature of the STAR aligner is that it reports the alignment statistics when it is finished running. These summary statistics usually include uniquely aligned reads, multimapped reads, unmapped reads, etc. **Table 1** and **Figure 1** shows the read and alignment statistics obtained from STAR for the nine treatment samples. To assess the sequencing quality of our library, we first look at the average mapped lengths of all samples. The average mapped lengths ranged from ~98-200bp as compared to the corresponding average input read length of 100-202bp, indicating a good quality alignment. Furthermore, we examined the percentage of uniquely mapped reads of all nine samples, which identified that ~80-90% of the reads that were uniquely mapped and ~3-6% of the reads were mapped to multiple loci, suggesting a high quality human RNA library.

Sample Name	Average Input Read Length	Average Mapped Length	% Uniquely Mapped Reads	% Mapped to Multiple Loci	% Unmapped Too short	% Unmapped : Other
SRR1177981	202	198.51	80.2%	3.8%	15.8%	0.1%
SRR1177982	202	198.65	83.8%	3.7%	12.3%	0.1%
SRR1177983	202	198.19	79.6%	3.5%	16.6%	0.1%
SRR1178008	202	198.13	88.1%	3.5%	8.1%	0.0%
SRR1178009	202	198.96	90.1%	2.5%	7.2%	0.0%
SRR1178010	202	198.81	90.6%	2.7%	6.5%	0.0%
SRR1178014	100	98.63	83.5%	6.6%	9.2%	0.2%
SRR1178021	200	195.15	82.0%	5.8%	11.9%	0.1%
SRR1178047	200	196.05	84.0%	5.8%	9.7%	0.1%

Table 1 – The read and alignment statistics obtained from STAR for the nine treatment samples

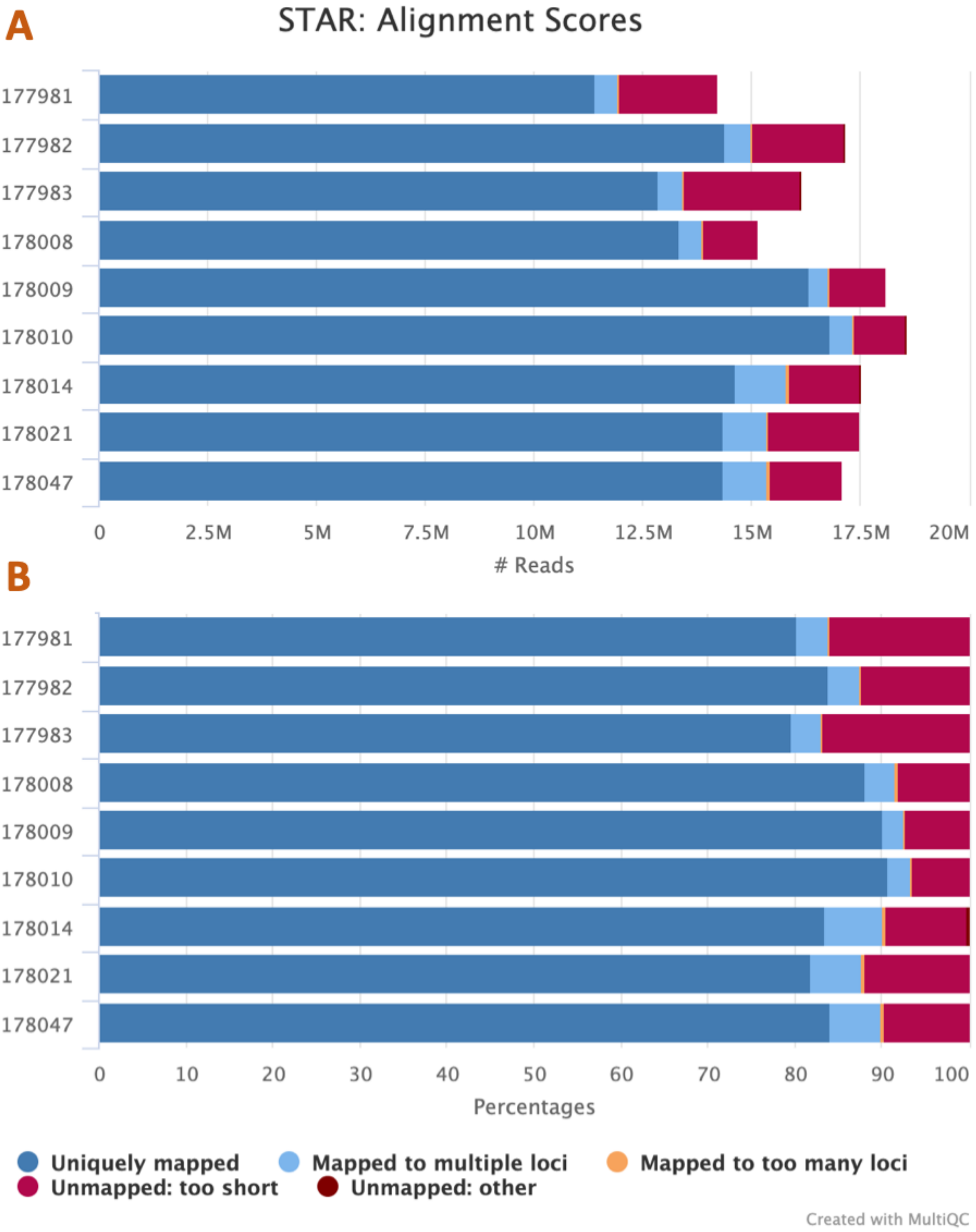


Figure 1. Bar plot of the alignment scores obtained from the STAR alignments for the nine treatment samples. The vertical axis is the samples. The horizontal-axis is the alignment scores displayed in number of reads (**A**) and in percentages (**B**).

Distribution of Read Counts

After the alignments were processed by STAR, we used the featureCounts program in the *subread* package to count the reads against a gene annotation. The featureCounts program is a tool that is developed for counting reads to genomic features such as genes, exons, promoters and genomic bins. Once the count file was generated for each sample, we ran Multiqc to collect relevant information about the counts data. Multiqc is a reporting tool that collects information from many different bioinformatics programs for a set of samples and combining them into a single convenient report. **Figure 2** show the bar plots of count assignments for all samples obtained from Multiqc. Similarly, **Figure 3** shows the box plots of distribution of counts for all nine samples. Looking at the distribution of counts for all samples, it seems that the mean and variance of the sample count are relatively similar to each other. Hence, the samples must be roughly independent from each other as we would expect.

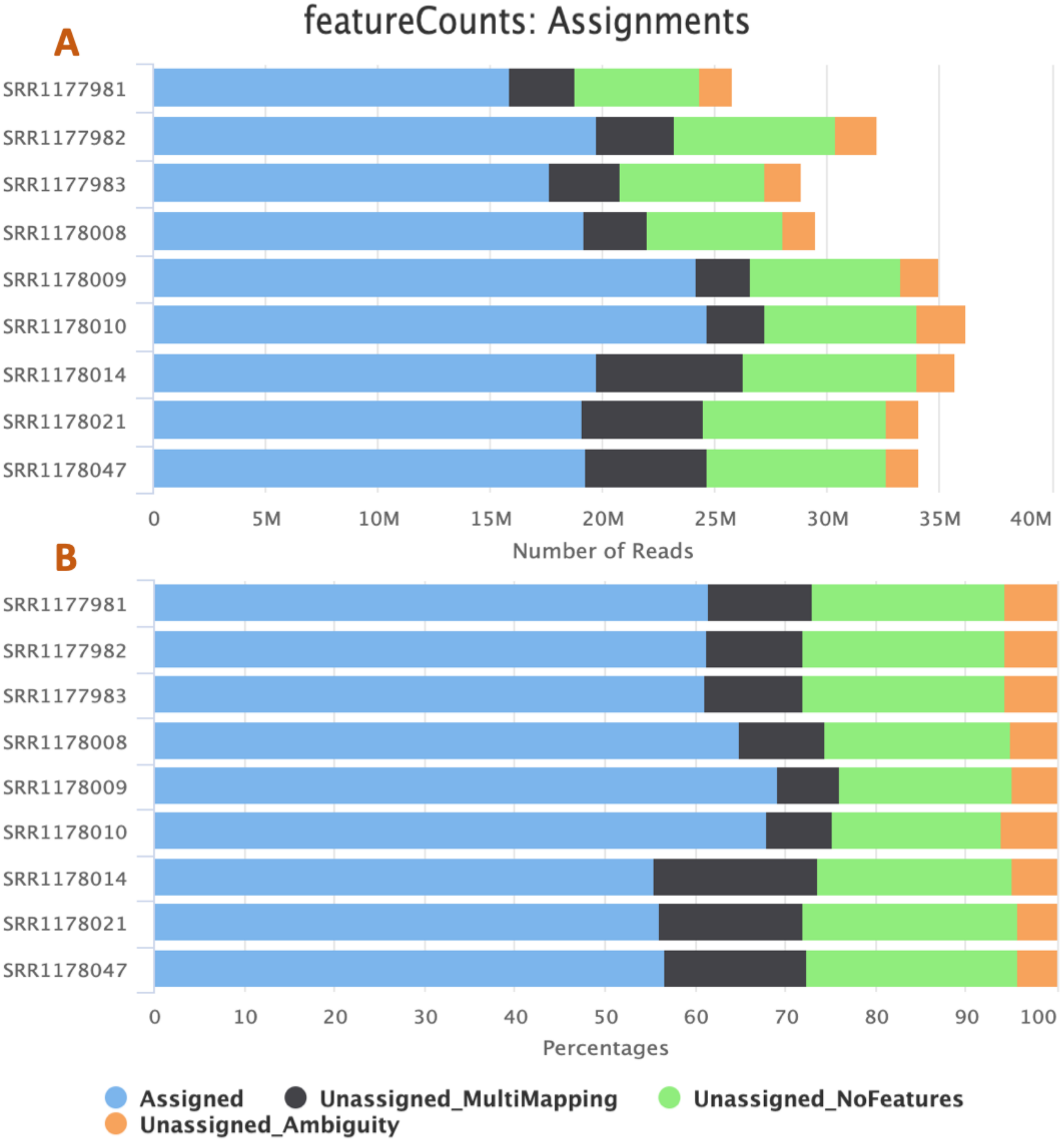


Figure 2 – Bar plot of assignments displayed by number of reads (**A**) and by percentages (**B**) obtained from Multiqc report.

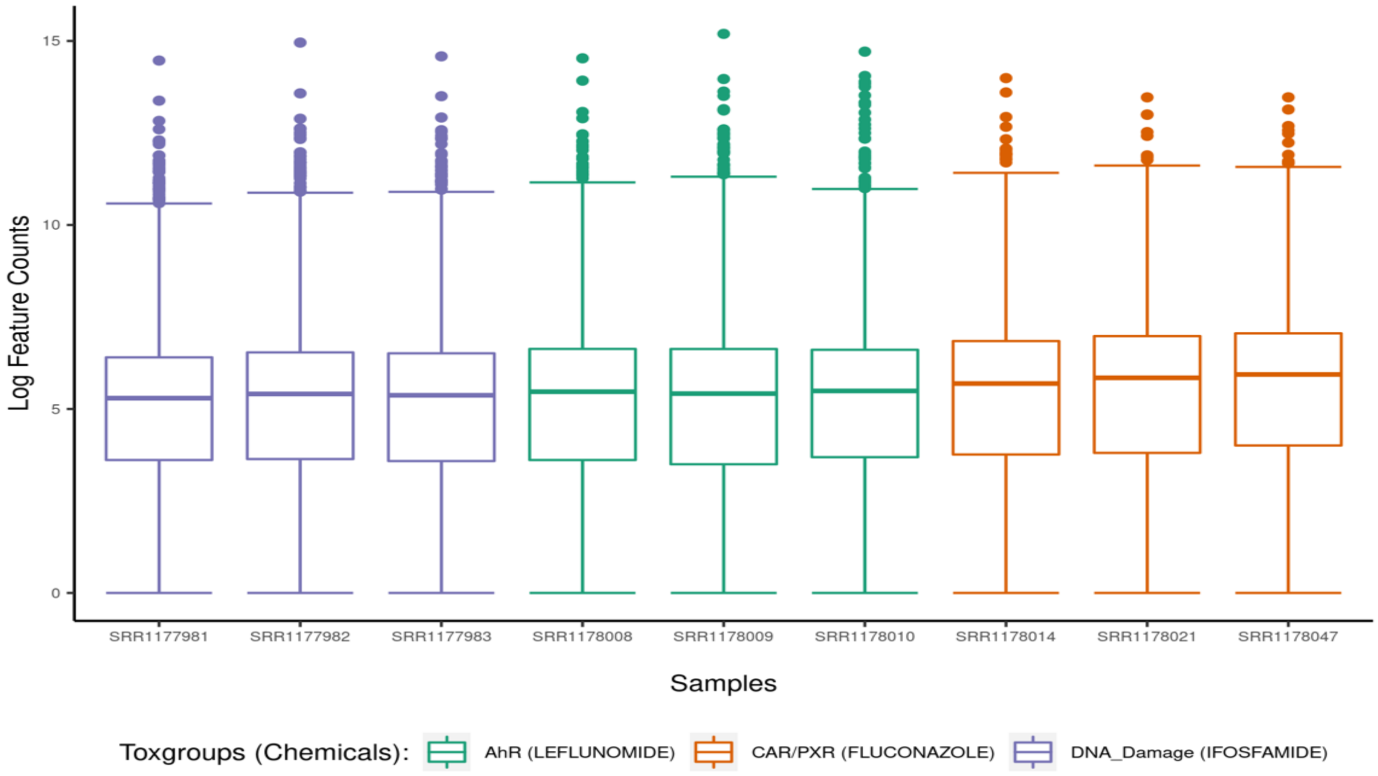


Figure 3: Box plot of distribution of feature counts for nine treatment samples. The counts in the vertical axis was computed in log scale, and the horizontal axis represents the samples.

RNA-Seq Differential Expression

Read counts are the most basic form of processed data that can be obtained from the RNA-seq experiments. The distribution of count data has three unique properties that often make them unsuitable for standard statistical analysis like linear models. First, count data is discrete as they only occur as integers. Second, they do not follow a normal distribution as it cannot have a negative value for count. Third, the mean and variance of a given gene's count distribution are often not independent, as the greater the mean count of a gene, the higher the variance. Thus, for these reasons, a generalized linear model, especially negative binomial regression, was commonly used as a statistical tool to analyze the count data. In this study, we utilized DESeq2, a Bioconductor package that uses negative binomial regression to estimate the count differences between a given treatment and an appropriate control group. **Table 2** shows the top 10 differentially expressed genes (**DEGs**) between a given treatment and an appropriate control that have the same *vehicle* control. The number of DEGs were reported at p-adjusted < 0.05 using Benjamin-Hochberg correction. A histogram in **Figure 4A,C,E** and a volcano plot in **Figure 4B,D,F** also show the distribution of log2 fold change values and the log2 fold change versus normal-p value for the significant DE genes, respectively. Based on the volcano plots for each chemical group, we observed that there are more DEGs in Fluconazole and Leflunomide as compare to Ifosamide. These findings indicate toxic groups of AhR and CAR/PXR which associate with Fluconazole and Leflunomide respectively, are more likely expressed in Rat gene expression as a way to treat and prevent fungal infections and rheumatoid arthritis as opposed to Ifosamide which use to treat cancer of the testicles.

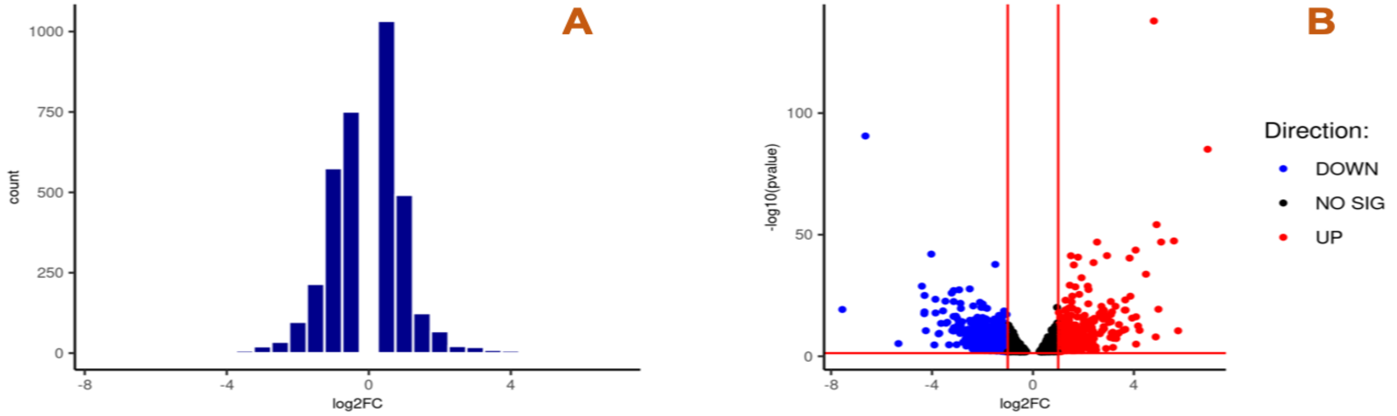
<table>
<tr><th colspan="4">Table 2 - Top 10 Differentially Expressed Gene by Chemical</th></tr>
<tr><th>Gene</th><th>Log2 Fold Change</th><th>P-value</th><th>Adjusted P-value</th></tr>
<tr><td colspan="4">Fluconazole (n = 3,499)</td></tr>
</table>

ORM1	4.79	1.25E-138	1.34E-134
STAC3	-6.64	2.53E-91	1.35E-87
CYP2B1	6.92	7.33E-86	2.61E-82
ABCC3	4.90	7.37E-55	1.97E-51
SLC5A1	5.59	3.72E-48	7.94E-45
GADD45A	2.54	1.10E-47	1.68E-44
CITED4	5.08	1.06E-47	1.68E-44
LIFR	4.07	2.14E-44	2.86E-41
G6PC	-4.03	9.63E-43	1.14E-39
G6PD	2.94	4.11E-42	4.39E-39
Ifosamide (n = 91)			
HBB	-7.01	7.76E-62	8.14E-58
HBA2	-6.88	9.60E-61	5.03E-57
NA	-0.01	3.33E-35	1.16E-31
HBA1	0.00	1.21E-25	3.18E-22
ABCB1B	3.78	3.34E-19	7.00E-16
POPDC2	-2.42	1.40E-16	2.45E-13
BCL6	-0.01	1.53E-11	2.29E-08
NA	-2.59	6.89E-10	9.03E-07
SDS	1.63	1.90E-09	2.21E-06
PAQR9	-1.40	3.67E-09	3.84E-06
Leflunomide (n = 1,389)			
HBA1	-9.92	9.97E-60	1.06E-55
HBB	-10.15	1.50E-57	7.99E-54
HBA2	-9.21	1.91E-47	6.79E-44
NA	-4.56	7.49E-44	2.00E-40
STAC3	-7.07	5.01E-38	1.07E-34
CYP1A1	9.97	1.82E-33	3.23E-30

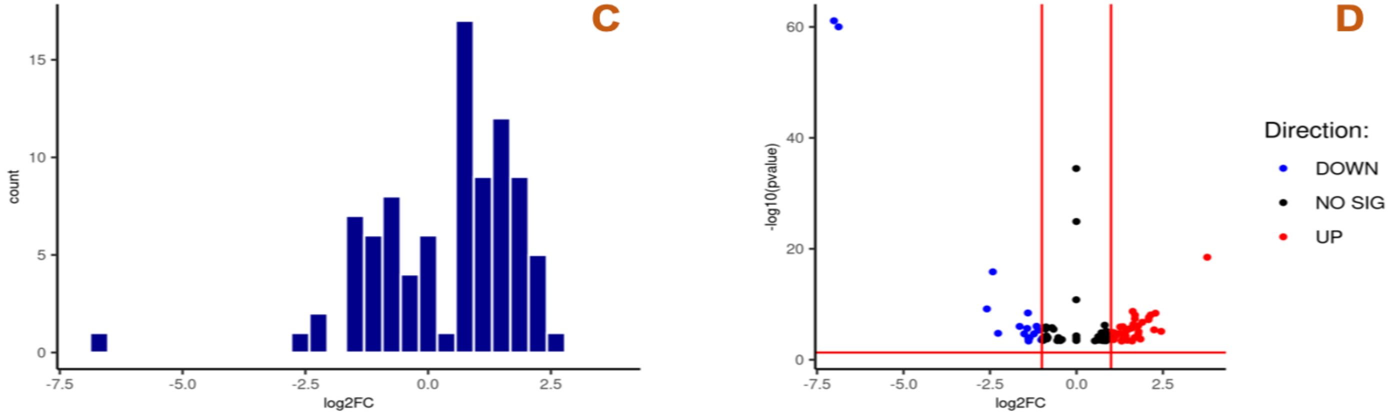
NA	-7.50	4.93E-33	7.50E-30
CYP1A2	4.33	1.42E-31	1.89E-28
UGT1A7C	4.02	2.46E-30	2.91E-27
FKBP5	2.27	4.79E-30	5.11E-27

***n is the number of DE genes using p-adjusted < 0.05 from Benjamin-Hochberg procedure

Fluconazole



Ifosamide



Leflunomide

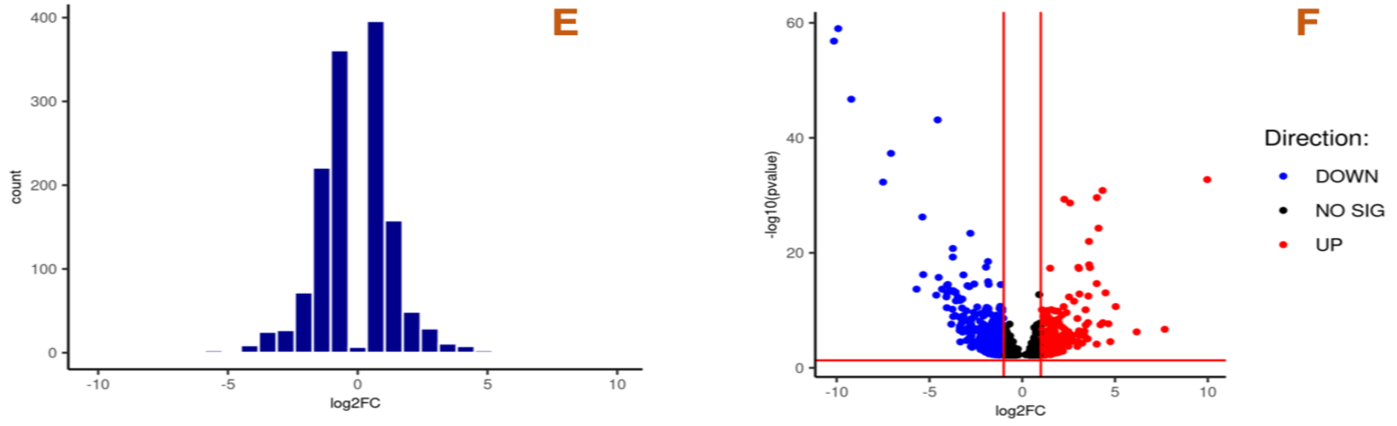


Figure 4 – Histograms of log2 fold changes values and volcano plots of log2 fold changes versus normal p-value for three treatment conditions (Fluconazole, Ifosamide, and Leflunomide). For A, C, E, vertical axis is count distribution of log2 fold change while horizontal axis is the log2 fold change values. For B, D, F, the vertical axis is the -log10 of norminal p-value and the horizontal axis is the log2 fold changes. Two red vertical lines represents the decision boundary that separates the up- and down-regulated genes. The red horizontal line shows the significant boundary of p-value with -log10 of alpha = 0.05.

Microarray Differential Expression

Following normalization of the microarray data with the Bioconductor package RMA, the limma package (in Bioconductor) was used to determine the differential gene expression between a treatment group (leflunomide, fluconazole, and ifosamide) and the control group with the same vehicle as the treatment (corn oil control for leflunomide and fluconazole, and saline control for ifosfamide). In the limma package, the lmFit function was used to estimate the fold change and standard error by fitting a linear model for each gene. The eBayes function in the limma package was used to apply empirical Bayes smoothing to the standard errors and the topTable function in limma was used to calculate the statistics for the top 10 genes, adjusting for multiple testing. All microarray differential expression analyses were conducted in R version 4.0.2.

Concordance of Microarray and RNA-Seq Differential Gene Expression

The concordance of the significant ($p<0.05$) DEGs detected by microarray and RNA-Seq were first both mapped by Refseq in order to identify the DEGs that were detected by both platforms for each chemical. The probability of observing the DEGs detected for each individual platform was calculated as n_1 and n_2, taking into account the entire gene set (N) as $P_{microarray}=n_1/N$ and $P_{RNA\text{-}Seq}=n_2/N$. The concordance was calculated as:

$$\frac{2 \text{ x intersection}(\text{DEGs}_{\text{microarray}}, \text{DEGs}_{\text{RNA-Seq}})}{\text{DEGs}_{\text{microarray}} + \text{DEGs}_{\text{RNA-Seq}}}$$

We calculated the number of DEGs observed by chance (n_x):

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

We used the number of DEGs observed by chance to adjust the concordance.

Results

A total number of 1997 genes were found to be differentially expressed with fluconazole treatment at an adjusted p value of <0.05. The volcano plot (**Figure 5A**) and histogram (**Figure 5B**) of the differentially expressed genes for fluconazole showed a normal distribution of differentially expressed genes spread on either side of 0. In contrast, many of the genes differentially expressed with ifosamide treatment had large adjusted p-values (**Figure 6**) with none of the genes reaching the adjusted p value threshold of <0.05 of significance. For leflunomide, 466 genes were differentially expressed with an adjusted p value of <0.05. The differentially expressed genes for leflunomide were relatively normally distributed around a log fold change of 0, although several genes had high fold changes but did not meet statistical significance based upon an adjusted p value <0.05 (**Figure 7**).

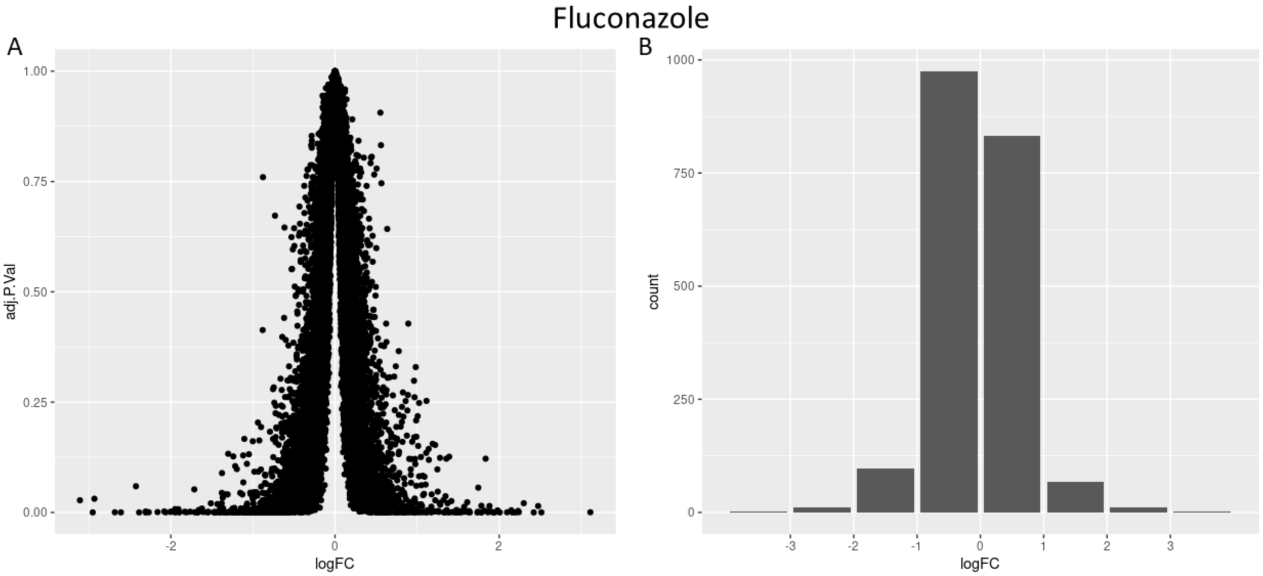


Figure 5. Volcano plot (**A**) and histogram (**B**) of the differentially expressed genes with Fluconazole treatment.

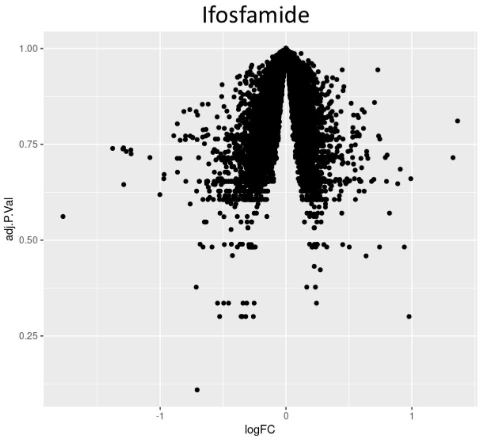


Figure 6. Volcano plot of the differentially expressed genes with Ifosamide treatment.

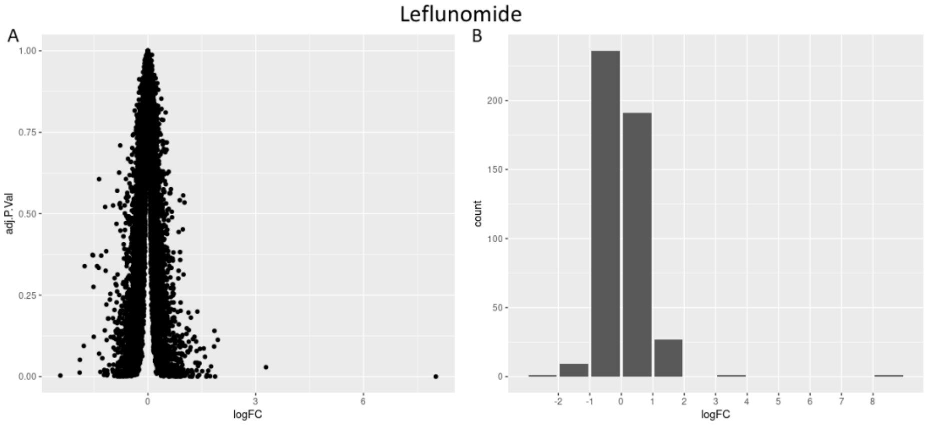


Figure 7. Volcano plot (**A**) and histogram (**B**) of the differentially expressed genes with Leflunomide treatment.

In order to better understand the genes that were differentially expressed between each treatment and the controls, we further evaluated the top 10 differentially expressed genes (**Table 3**).

Table 3. Top 10 Differentially Expressed Genes by Chemical

		Fluconazole		
Gene	**Log Fold Change**	**P value**	**Adjusted P Value**	**B**
ORM1	1.39	7.62E-12	2.37E-07	16.48
NIMK1	-2.31	1.58E-10	2.45E-06	13.77
CYP2B1	2.42	3.53E-10	3.28E-06	13.04
ABLIM3	1.78	4.67E-10	3.28E-06	12.78
ABLIM3	1.64	5.27E-10	3.28E-06	12.67
ID4	-1.40	1.14E-09	5.89E-06	11.97
IRAK3	1.33	2.68E-09	9.64E-06	11.18
TSPAN14	-0.87	2.75E-09	9.64E-06	11.16
WOYTEE	-1.31	2.79E-09	9.64E-06	11.15
CLPX	-1.18	4.62E-09	1.34E-05	10.68
		Ifosamide		
RORA	-0.71	3.51E-06	0.11	2.53
ABHD1	-0.35	3.09E-05	0.30	1.16
MAPK6	-0.32	4.30E-05	0.30	0.94
AFM	-0.26	5.52E-05	0.30	0.78
EHBP1	-0.36	5.82E-05	0.30	0.75
ALDH1A7	0.98	6.54E-05	0.30	0.67
ADAMTS9	-0.53	6.77E-05	0.30	0.65
DYRK1A	0.24	9.44E-05	0.34	0.43
RORA	-0.50	1.11E-04	0.34	0.32
SOX4	-0.54	1.20E-04	0.34	0.27
		Leflunomide		

CYP1A1	8.01	1.28E-15	3.97E-11	22.82
CYP1A2	1.38	3.09E-12	4.80E-08	16.66
FBXO31	1.23	3.84E-10	3.98E-06	12.56
SHEEPU	1.63	3.14E-09	2.44E-05	10.72
TCEA3	-0.68	1.80E-08	1.12E-04	9.17
CTSL1	0.59	2.86E-08	1.40E-04	8.75
PLA2G12A	1.87	3.15E-08	1.40E-04	8.66
DST	0.92	6.07E-08	2.22E-04	8.08
EML4	0.78	6.42E-08	2.22E-04	8.03
R3HDM2	-0.76	7.75E-08	2.29E-04	7.86

We evaluated the concordance between the array and sequencing platforms for the ifosfamide, fluconazole, and leflunomide chemical treatments (**Figures 8 and 9**). We found that for leflunomide, the concordance in DEGs between the platforms was -23.7%. For ifosfamide, the concordance was 55.2% whereas for Fluconazole, the concordance was 32.1%. In order to determine whether the concordance between the microarray and RNA-seq platforms was similar for DEGs in high and low abundance, we compared the overall concordance to the DEGs below and above the median for the three chemicals (**Figure 10**). We found that the low abundance DEGs (below the median) had a lower concordance compared to the DEGs above the median.

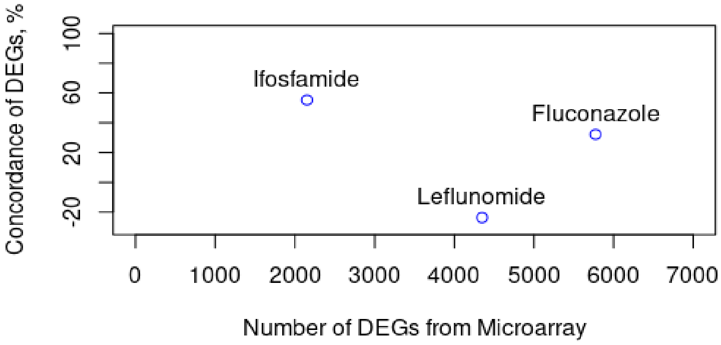


Figure 8. Comparison of the number of DEGs identified by Microarray

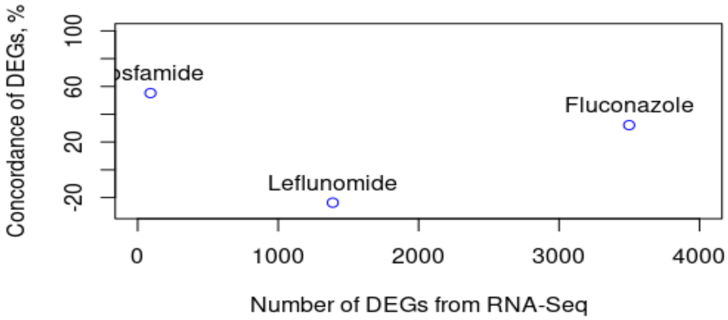


Figure 9. Comparison of the number of DEGs identified by RNA-Seq

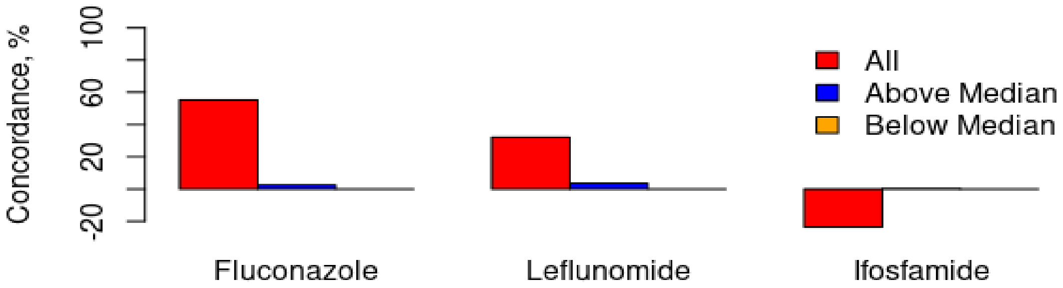


Figure 10. Concordance between Microarray and RNA-Seq

Pathway enrichment

The top 10 expressed genes for each chemical treatment were used to determine which pathways were enriched. GATHER was utilized to complete the top 10 DEGs as it is convenient and fast at identifying KEGG pathways. Results from GATHER for each chemical were compared with the listed common pathways enriched for each of the MOA chemical groups from the supplementary materials by Wang *et al*[1] (**Table 4**). Fluconazole was included as a chemical in the MOA of orphan nuclear hormone receptors (CAR/PXR), ifosfamide was included in DNA damage, and leflunomide was included in aryl hydrocarbon receptor (AhR). The results show that only the pathway of glutathione metabolism was mostly consistent with the listed common pathway enrichment of glutathione-mediated detoxification (**Table 4**). Other pathways cannot be considered as consistent between two lists.

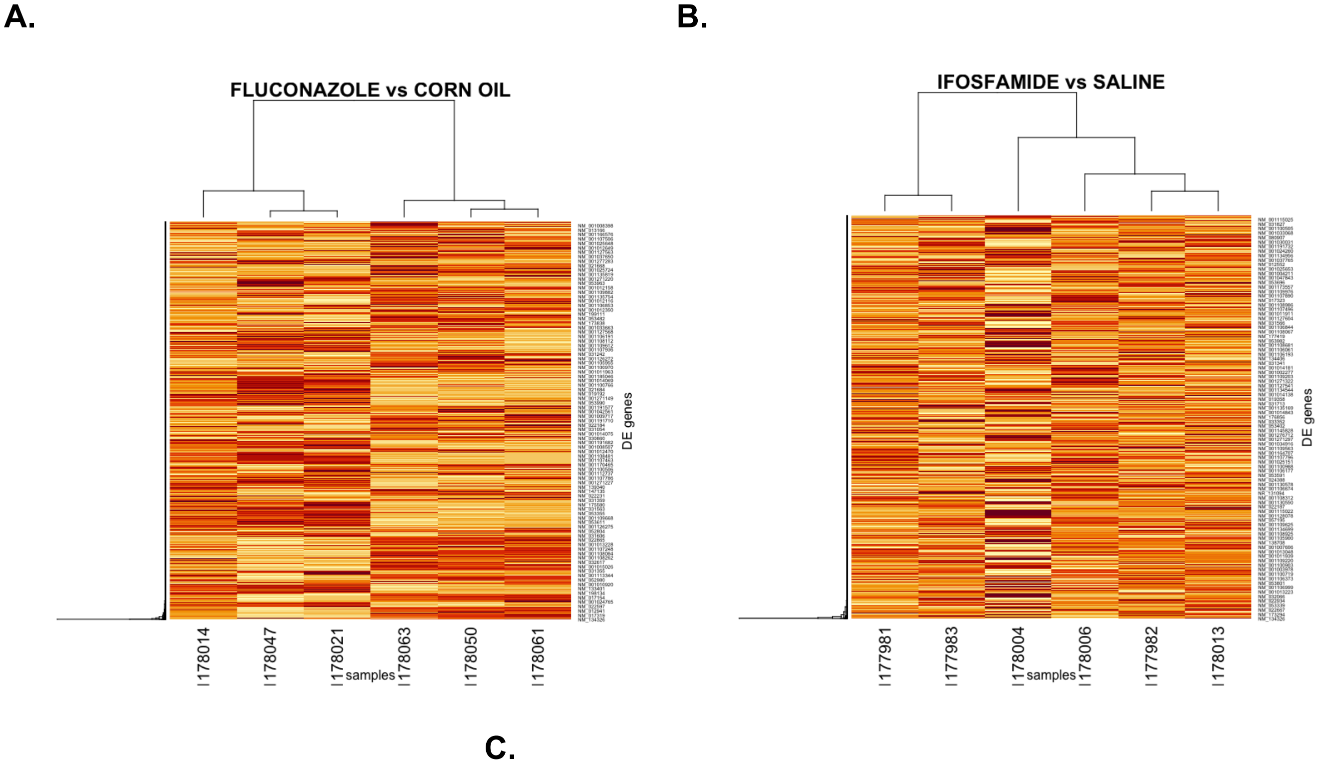
Table 4. Pathway Enrichment with top 10 expressed genes for each chemical treatment analyzed by GATHER

Fluconazole	Ifosfamide	Leflunomide
Pentose phosphate pathway Galactose metabolism Glutathione metabolism Starch and sucrose metabolism	Cysteine metabolism Glycine, serine and threonine metabolism	gamma-Hexachlorocyclohexane degradation Fatty acid metabolism Tryptophan metabolism

Glycolysis / Gluconeogenesis Cell cycle Insulin signaling pathway Jak-STAT signaling pathway		

Clustered Heatmap of Counts

Normalized expression matrices are used to create heatmaps for each chemical treatment (**Figure 11**).



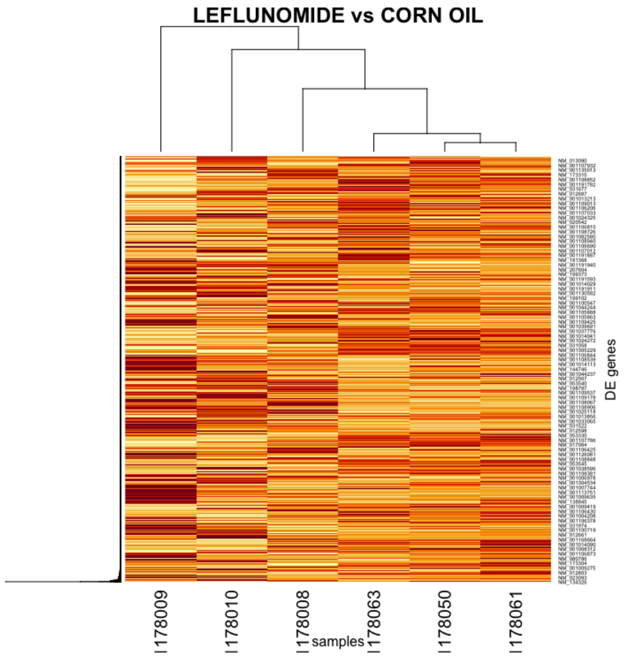


Figure 11. Clustered heatmap of the counts. (**A**) Comparing fluconazole as chemical treatment to corn oil as the control. (**B**) Comparing fluconazole as chemical treatment to saline as the control. (**C**) Comparing leflunomide as chemical treatment to corn oil as the control.

Discussion

In this study, a high quality RNA library was examined by STAR aligner and used to identify DEGs between rat genome samples from different chemical treatments and control groups and the limma package in Bioconductor was used to identify DEGs from the microarrays. By utilizing DESeq2, top 10 differentially expressed genes for each 3 chemicals - fluconazole, ifosfamide, and leflunomide were able to be identified and listed separately. After analyzing DEGs from both platforms, concordance of the significant DEGs detected by microarray and RNA-Seq was calculated for each chemical treatment. Results show that in general, the concordance in DEGs between platforms is not relatively high, with -23.7%, 55.2%, and 32.1% for each chemical, respectively. And the negative concordance for leflunomide indicates that the DEGs detected were due to chance rather than biology. Further, the pathway enrichment of the top 10 DEGs for each chemical treatment was not consistent with common pathway enrichment for their corresponding MOA chemical groups.

Our results suggest that the two platforms do not concordant with each other well. In general they showed low concordance scores, and that the concordance also depends upon the abundance of the DEGs- the low abundance DEGs had a lower concordance compared to the DEGs above the median. This is consistent with the study by Wang *et al* that suggested microarray and RNA-seq perform differently on the precision of low abundance gene detection.[1]

With the exception of the glutathione metabolism pathway, the enriched pathways were not consistent with the common pathway enrichment for the MOA chemical group. The only one that can be considered related is glutathione metabolism, which could be seen as a larger concept of glutathione-mediated oxidative stress detoxification. Our results could be due to one chemical within a MOA group that was not represented in the whole group's common pathway enrichment due to a lack of DEGs that met the threshold of significance (p<0.05) after adjusting for multiple testing. For example, the MOA of AhR consists of 3 chemicals - 3ME, LEF, NAP, so the top 10 enrichment genes for Leflunomide(LEF) alone is not a good representation of the whole AhR group; hence the difference in pathway involved. Secondly, it could be due to a small sample size of 10 top differentially expressed genes. Ten DEGs are such a small sample size that it could introduce a lot of uncertainty while running in GATHER, so some pathways that actually affected are not viewed as significant enough to be reported while some irrelevant pathways seemed to be of importance. If we included the top 100 genes, the pathways identified would likely be more reliable. Thirdly, some terms for the pathways shown in this study and the common pathway enrichment list are related in some degrees or might have intersections. Because of the difference in tools utilized or sample size used, different terms might be selected to represent the same general process, hence the difference between two lists.

Conclusion

Overall, our study findings are consistent with the original paper on the concordance dependency of treatment effect between RNA-Seq and microarrays, but did not successfully reproduce the same pathway enrichment and clustered heatmap. Our data suggests that the RNA expression methodology used should be considered carefully based upon whether low abundance genes are likely to be biologically or clinically meaningful.

One challenge we encountered was in making clustered heatmaps for all the DEGs from the normalized matrix, and trying to make it consistent with each MOA group. The needed information is which sample represented what kind of treatment or control, and how could the matrix be changed and filtered so that it could show desired clustering but meanwhile it is still objectively scientific?

References

1. Wang C, et al. A comprehensive study design reveals treatment- and transcript abundance-dependent concordance between RNA-seq and microarray data. Nat Biotechnol. 2014 Sep; 32(9): 926-932.