Project 3 - A comprehensive study design reveals treatment- and transcript abundance–dependent concordance between RNA-seq and microarray data

Konrad Thorner, Aishwarya Deengar, Jia Liu, Morgan Rozman

INTRODUCTION

The ability to quantify genome-wide gene expression is an invaluable tool in research, made possible through advancements in microarrays and RNA-sequencing. There has been a widespread effort to determine the relative reliability of these two technologies, particularly in regards to detecting differentially expressed genes (DEGs).

Wang et al. sought to investigate this question by comparing gene expression data from the livers of rats exposed to chemicals with varying modes of action (MOAs) [1]. The primary goal was to measure the degree of concordance or overlap between these platforms for DEGs, MOAs, and pathways. Additionally, through use of a training set, predictive classifiers were generated to determine the MOA of other samples.

For our approach to this study, a subset of the conditions (a “toxgroup” of three chemicals with different modes of action) is to be assessed. Beginning with raw RNA-seq data, we process it to find differential gene expression between treatments and controls. A similar analysis is performed on microarray data for the same samples, allowing us to calculate and visualize concordance. We also attempt to find similar enriched pathways and clustering of samples as in the original study.

DATA

Samples Description

In the original study, they use same set of liver samples of rats under varying degrees of perturbation by 27 chemicals representing multiple modes of action (MOA).[1] With three rats per chemical with matched controls, they isolated RNA from the livers, and analyzed these samples using Affymetrix microarrays and Illumina RNA-seq.

In this project, we only focus on three MOAs which associated with well-defined receptor-mediated processes — orphan nuclear hormone receptors (CAR/PXR), aryl hydrocarbon receptor (AhR) and non-receptor-mediated—DNA damage (DNA_Damage). The chemicals analyzed are leflunomide (LEF), fluconazole (FLU), and ifosfamide (IFO).

Microarray Data

Microarray data for the toxgroup was obtained in the form of a pre-normalized expression matrix. We did not need to perform any further quality control or normalization on the data. This matrix and the treatment samples are used to determine differentially expressed genes.

RNA-Seq Data

RNA-Seq data for the toxgroup was obtained from a prepared dataset as 18 paired fastq files. We run fastqc on each of the fastq files and use multiqc to collect information from these 18 fastqc files and combine them into a single convenient report. After checking the quality of the fastq file, we align each of the samples against the rat genome using the STAR aligner.

Sequencing Datasets

Using multiqc, we obtained general statistics of samples. Table 1 shows the percentage of duplicate reads, GC percentage, length of each read and total sequences in millions. Duplications and GC percentage are effectively the same in all of the samples. However, the average sequence length in sample SRR1178014 is 50 base pairs which is different from the others. We checked Table 2 for unique mapped reads and Figure 1 for sequence counts and found no significant difference between SRR1178014 and other samples. We conclude that differences in length should not have an effect on our results.

Sample Name	% Dups	% GC	Length	Total sequences in M
SRR1177981_1	56.10%	48%	101 bp	14.2
SRR1177981_2	48.40%	48%	101 bp	14.2
SRR1177982_1	58.50%	48%	101 bp	17.2
SRR1177982_2	54.30%	48%	101 bp	17.2
SRR1177983_1	57.60%	48%	101 bp	16.2
SRR1177983_2	50.00%	48%	101 bp	16.2
SRR1178008_1	56.20%	49%	101 bp	15.2
SRR1178008_2	54.40%	49%	101 bp	15.2
SRR1178009_1	62.20%	49%	101 bp	18.1
SRR1178009_2	60.10%	49%	101 bp	18.1
SRR1178010_1	62.50%	49%	101 bp	18.6
SRR1178010_2	61.40%	49%	101 bp	18.6

SRR1178014_1	53.90%	49%	50 bp	17.5
SRR1178014_2	51.90%	49%	50 bp	17.5
SRR1178021_1	48.70%	49%	100 bp	17.5
SRR1178021_2	46.40%	49%	100 bp	17.5
SRR1178047_1	48.50%	49%	100 bp	17.1
SRR1178047_2	47.30%	49%	100 bp	17.1

Table 1. General statistics of each sample. SRR117-7981 to 7983 in orange stand for DNA_Damage MOA, 8008 to 8010 in green stand for AhR MOA, 8014 to 8047 in blue stand for CAR/PXR MOA.

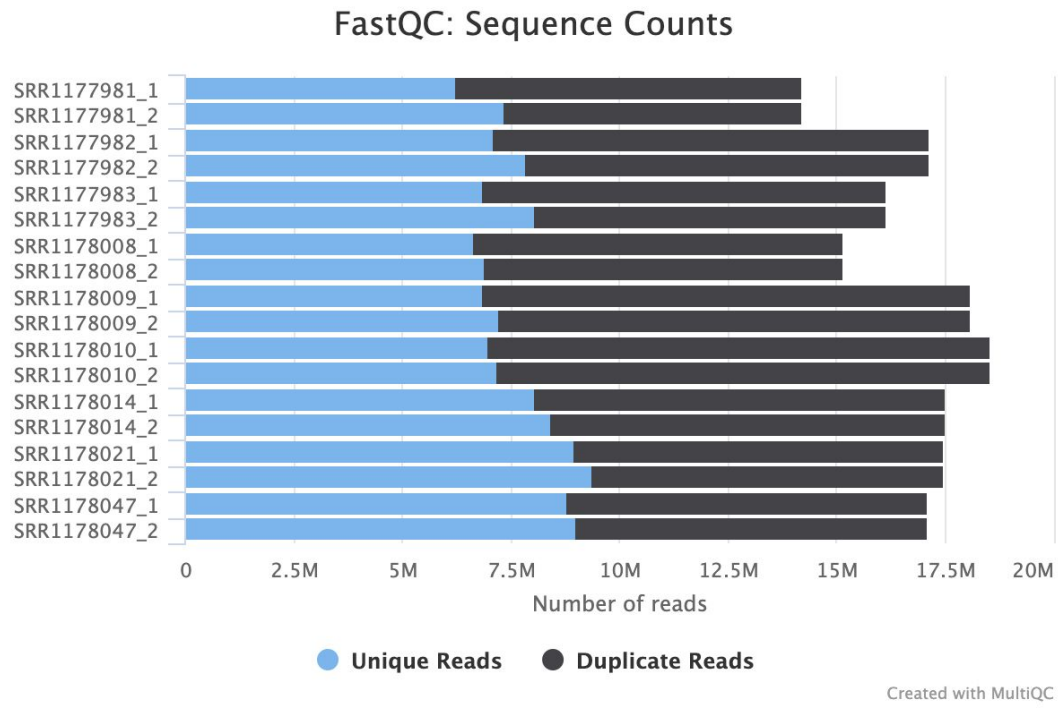

Figure 1. Number of reads in each sample.

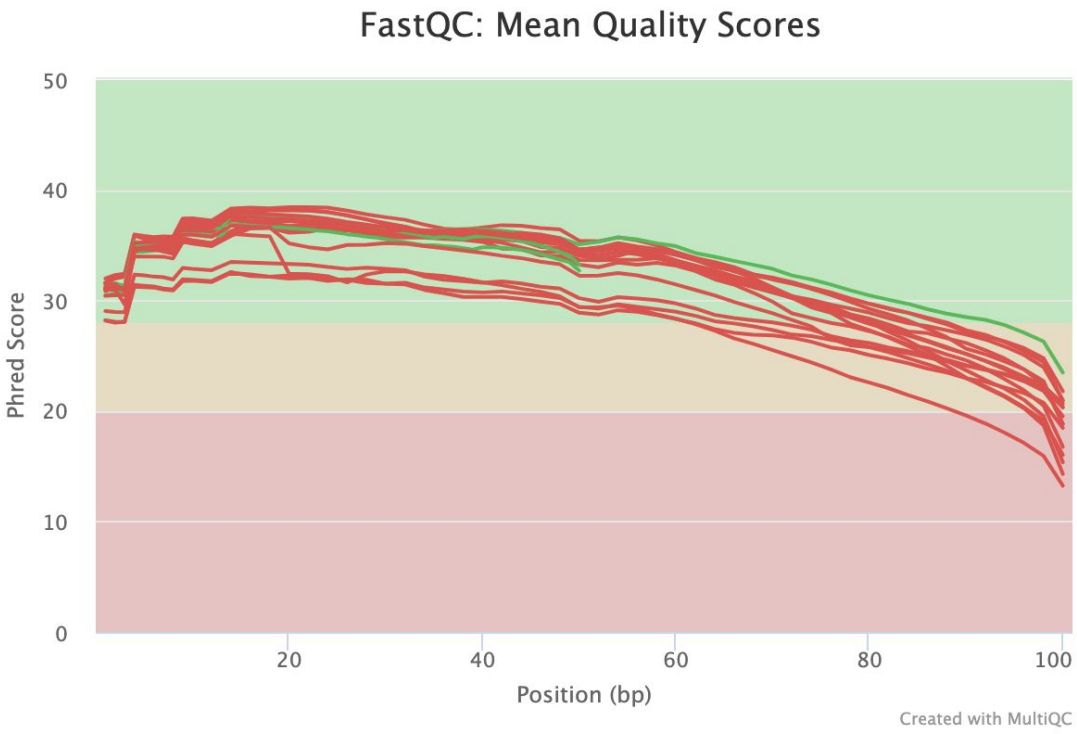

Figure 2. Mean quality scores of 9 samples.

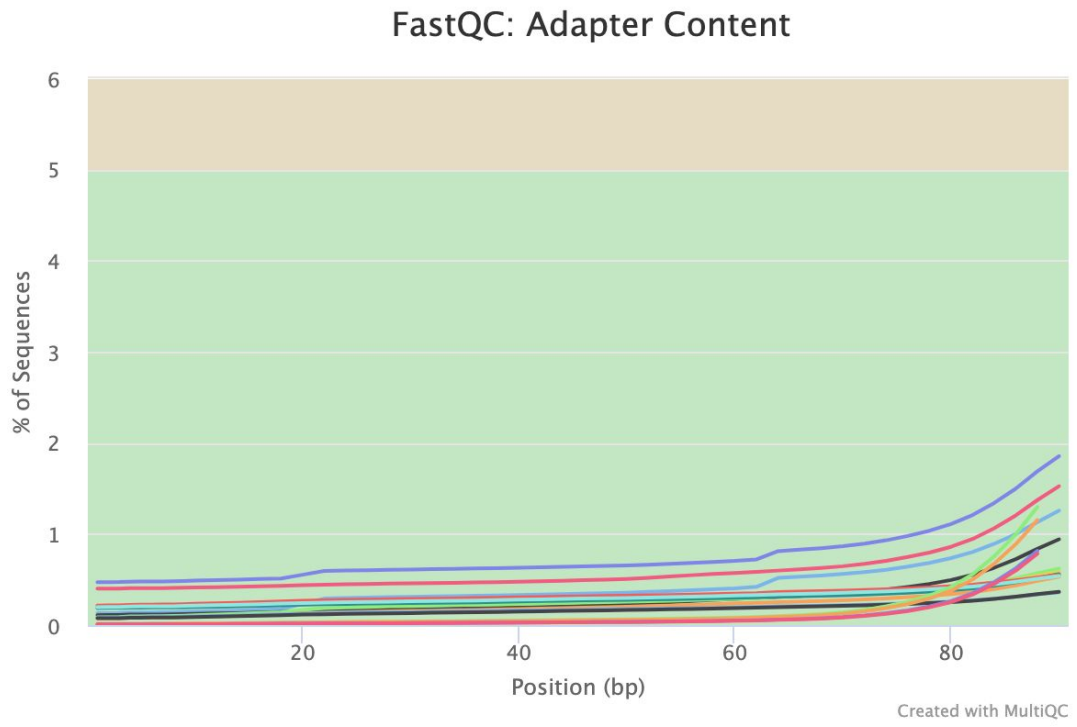

Figure 3. The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.

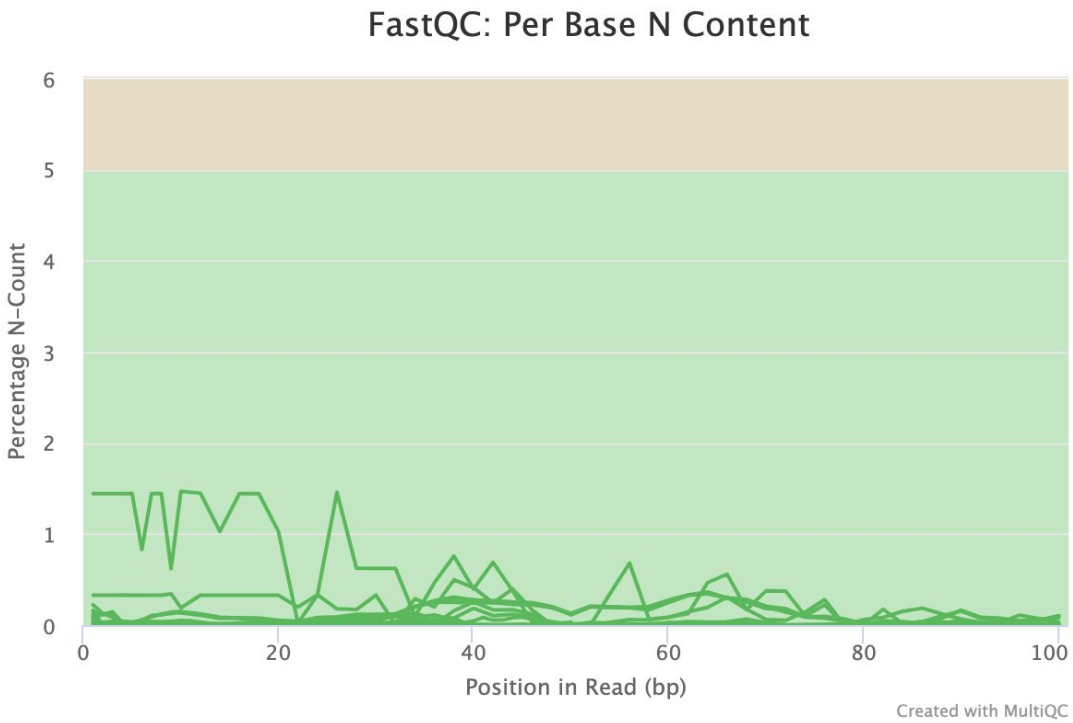

Figure 4. The percentage of base calls at each position for which an N was called.

Figure 1 shows the number of reads in each sample. In each sample, the ratio of unique and duplicated reads is shown in different colors. From the figure, the results of all the samples are very similar. In Figure 2, mean quality scores of each sample shows that towards the end of the sequence, the quality of all samples decreases. The reason for the quality loss may be the degradation of RNA in the tail. However, in this project, the low quality of the tail has little impact on our analysis given that the first part of sequences are of high quality. Adaptor content and per base N content shown in Figures 3 and 4 all indicate a high quality in each sample.

Sample Name	Input reads number	Uniquely mapped reads, %	Mismatch rate per base, %	Deletion rate per base, %	Insertion rate per base, %	Too short reads unmapped, %
SRR1177981	14,229,351	80.19%	0.76%	0.01%	0.01%	15.82%
SRR1177982	17,168,681	83.82%	0.77%	0.01%	0.01%	12.26%
SRR1177983	16,152,281	79.61%	0.86%	0.01%	0.01%	16.63%
SRR1178008	15,156,072	88.08%	0.59%	0.01%	0.01%	8.11%
SRR1178009	18,110,504	90.09%	0.66%	0.01%	0.01%	7.19%

SRR1178010	18,572,519	90.63%	0.73%	0.01%	0.01%	6.47%
SRR1178014	17,524,782	83.52%	0.44%	0.00%	0.00%	9.20%
SRR1178021	17,497,925	81.95%	1.10%	0.01%	0.01%	11.92%
SRR1178047	17,093,302	83.98%	0.88%	0.01%	0.01%	9.67%

Table 2. Statistic result from STAR

The statistics in Table 2 indicate that the quality of alignment is also acceptable. In unique reads, the percentage of mismatch, deletion and insertion rate per base are relatively similar and low which means the quality of each alignment is similar and high.

METHODS

Using the raw data obtained in bam files, the “featurecounts” tool (a software package designed for counting reads to genomic structures such as genes, exons, etc) was run. Featurecounts produces a count matrix and a summary for each sample. Next, multiqc was run using the summar files generated in the previous step. Multiqc is a bioinformatics tool which aggregates the results from analyses and combines them in a single report.

Microarray Analysis

To determine differentially expressed genes based on the microarray data, we used the R limma [2] package from Bioconductor. For each chemical in our tox group, we performed limma analysis via empirical bayes statistics for differential expression. p-value was adjusted using the Benjamini & Hochberg correction. We matched controls to samples based on the vehicle they were in (100% corn oil or 100% saline).

Probe ID to RefSeq Mapping

Probe ids are mapped from the microarray dataset to gene names and to RefSeq ids using a mapping matrix. Probe ids with no corresponding RefSeq id are filtered out in concordance analyses. If a probe id has more than one corresponding RefSeq id, each RefSeq id is counted with the same fold change and expression values that resulted from the limma analysis for that probe id. If a RefSeq id is mapped more than once, these entries are collapsed to one entry by taking the median of the results from limma for each entry.

Concordance

Concordance between microarray and RNA-seq is the percentage of differentially expressed genes shared by the two platforms with agreement in the direction of fold change. Concordance is computed as:

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

This is the same equation used in the original paper where DEG represents the number of differentially expressed genes identified in each sample type. DEGs are genes with an unadjusted p-value < 0.05 and a log fold change < -1.5 or log fold change > 1.5 to match the definitions in the original research. We also only take into account the DEGs which have the same direction of fold change in both the microarray and RNA-seq results.

Pathway Enrichment and MOA Clustering

The top ten DEGs for each condition are entered into the DAVID software to determine functional enrichment [3]. A clustered heatmap of normalized counts is also generated, combining the 50 DEGs with the highest log fold change for each condition.

RESULTS

The resultant plot of the multiqc run on the featurecounts summary file is shown in Figure 5 below..

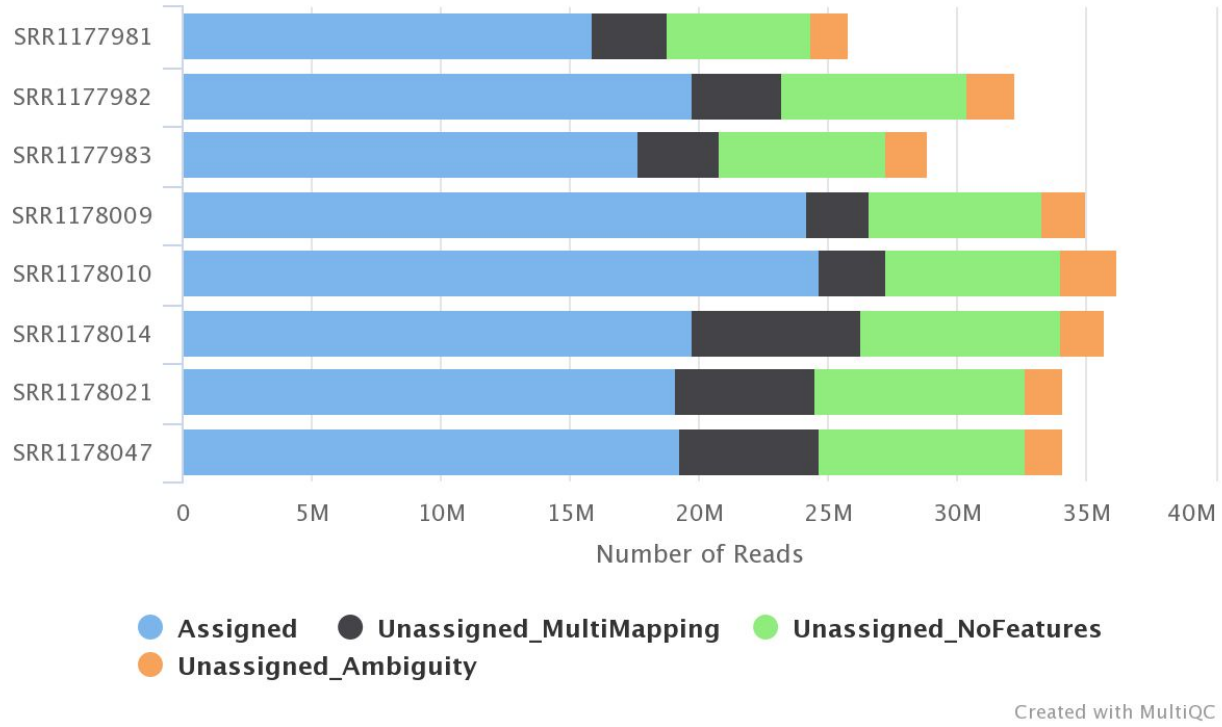

Figure 5. Multiqc plot depicting reads associated with counts of experiment samples

The multiqc reported a plot which shows the number of reads in each experiment group which could be mapped along with those regions which could not be mapped either due to multi-mapping, lack of features or due to ambiguity.

Sample Name	% Assigned	M Assigned
SRR1177981	61.5%	15.9
SRR1177982	61.3%	19.8
SRR1177983	61.1%	17.6
SRR1178009	69.2%	24.2
SRR1178010	68.0%	24.7
SRR1178014	55.4%	19.8
SRR1178021	56.1%	19.1
SRR1178047	56.6%	19.3

Table 3. Multiqc report table depicting sample names and their % counts assigned and M assigned

The first three samples are from the Ifosfamide treatment, the second two from Leflunomide and the final three from Fluconazole. Using the counts matrix generated from feature counts, DESeq2, a bioinformatics tool which uses negative binomial regression to estimate count differences between treated samples and appropriate controls to determine differential expression. The number of genes having $p < 0.05$ are as follows:

AhR-Leflunomide samples : 2794 genes

CAR/PXR-Fluconazole samples: 3503 genes

DNA Damage-Ifosfamide samples: 740 genes

The boxplots for the 8 experimental groups were generated and shown below. Additionally , the top 10 differentially expressed genes (DEGs) obtained by sorting by p-value are described in Tables 4-6.

Gene ID	p-value
15085	2.09e-189
5531	4.47e-101
2458	4.77 e-98
14503	1.16 e-81
2129	9.94 e-76
8230	3.57 e-72
16838	1.63 e-69
13442	1.41 e-66
13343	2.89 e-62
12363	1.59 e-56

Table 4. Table depicting the p values of AhR-Leflunomide samples when compared with control (Corn oil - 100%) samples using DESeq2.

GeneID	p-value
14945	4.57e-134
8231	1.19e-118
9638	3.72 e-90
9920	9.19 e-86
15711	5.60 e-55
12752	1.20 e-48
15319	1.66 e-47
14157	7.46 e-47
14368	2.35 e-44
12810	1.40 e-42

Table 5. Table depicting the p values of CAR/PXR-Fluconazole samples when compared with control (Corn oil - 100%) samples using DESeq2.

Gene ID	P value
12850	3.02e-118
12283	1.02 e-59
9278	8.80 e-52
14132	1.71 e-46
8632	3.96 e-40
14208	4.59 e-38
14795	4.63 e-34
14902	1.56 e-32
14236	3.34 e-30
13353	1.12 e-28

Table 6. Table depicting the p values of DNA Damage-Ifosfamide samples when compared with control (Corn oil 100%) samples using DESeq2.

The fold change values from the significantly differentially expressed genes were used to create the following histograms and scatter plots as shown below.

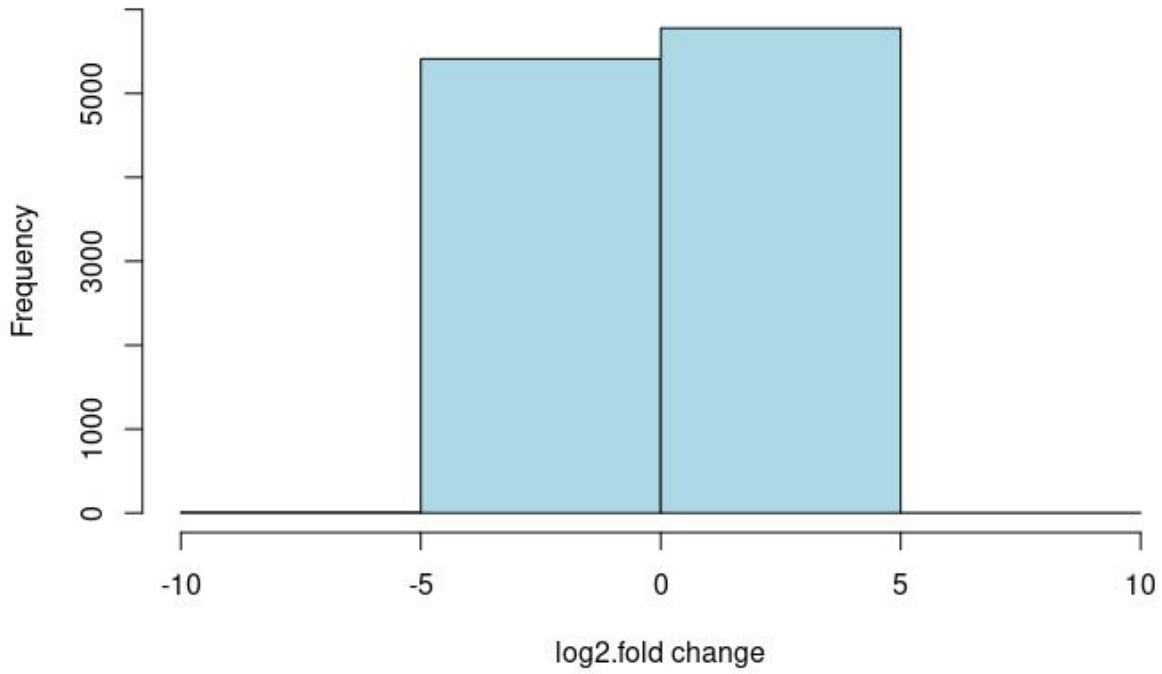

Figure 6. Log2 fold change vs frequency histogram of AhR-Leflunomide samples when compared with control.

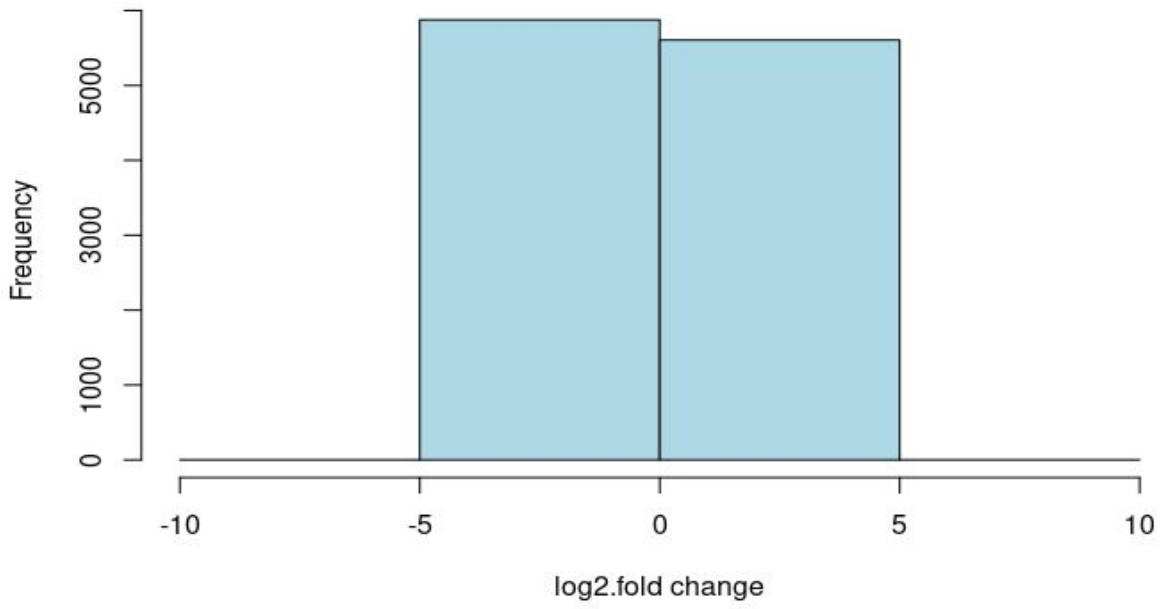

Figure 7. Log2 fold change vs frequency histogram of CAR/PXR-Fluconazole samples when compared with control (Corn oil - 100%) samples using DESeq2.

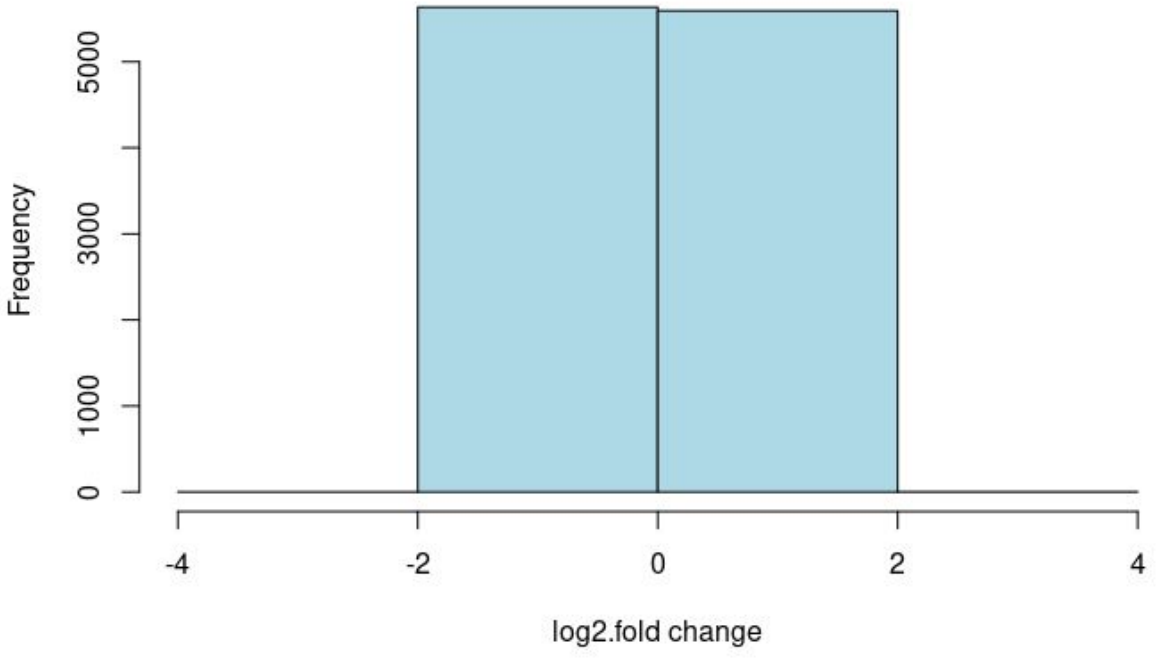

Figure 8. Log2 fold change vs frequency histogram of DNA Damage-Ifosfamide samples when compared with control (Corn oil 100%) samples using DESeq2.

The scatterplots were generated using the log2 fold change and nominal p values obtained from the feature counts result.

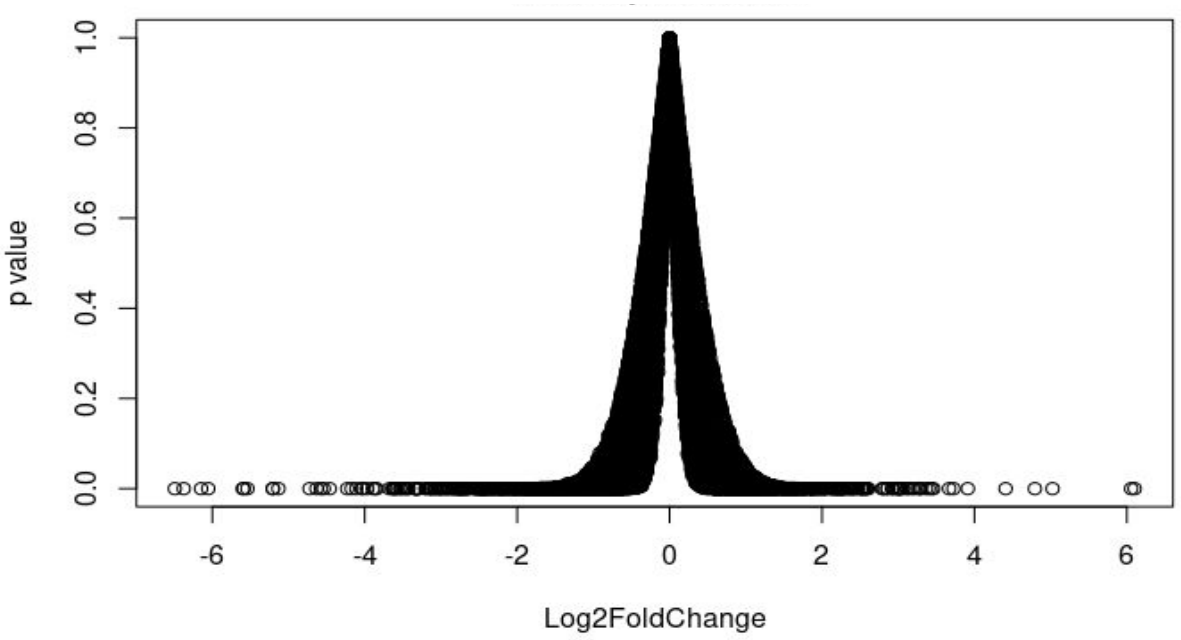

Figure 9. Log2 fold change vs frequency scatter plot of AhR-Leflunomide samples when compared with control.

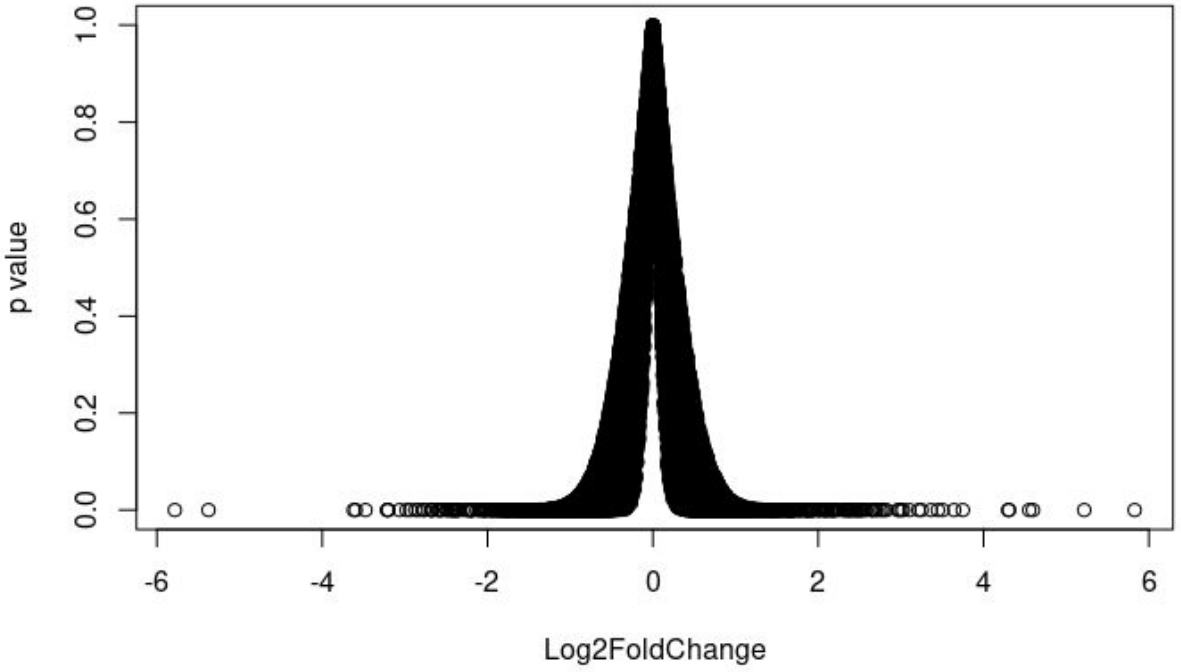

Figure 10. Log2 fold change vs frequency scatter plot of CAR/PXR-Fluconazole samples when compared with control (Corn oil - 100%) samples using DESeq2.

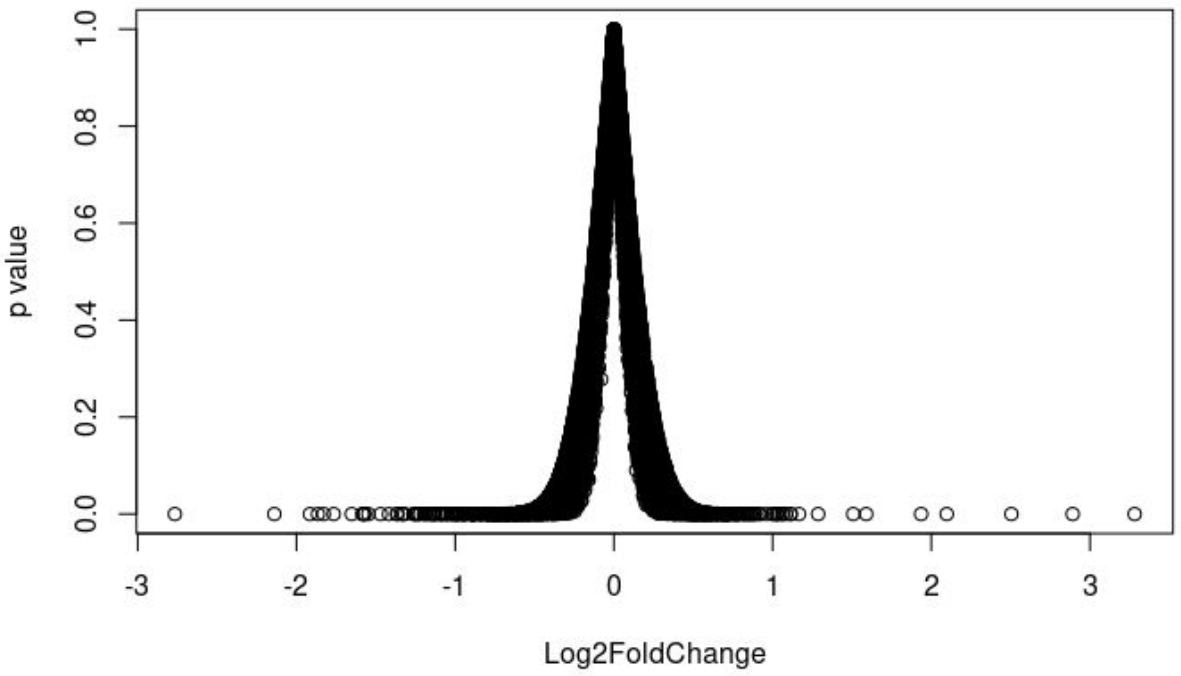

Figure 11. Log2 fold change vs frequency scatter plot of DNA Damage-Ifosfamide samples when compared with control (Corn oil 100%) samples using DESeq2

From the above histograms and scatter plots we can see that only minor perturbations were seen in the RNA counts between control and experimental group which is contrary to what was seen by the authors in their paper.

Microarray DEGs

There were 466, 1,997, and 0 probesets significant in the leflunomide, fluconazole, and ifosfamide chemical sample groups, respectively, at an adjusted p-value < 0.05 (Table 7). At an unadjusted p-value < 0.05, there were 4,351, 5,774, and 2,151 probesets found significant in the same groups.

The top ten differentially expressed genes by adjusted p-value for each chemical group are found in Table 8. There are genes that mapped to more than one probeset id, including Ablim3. There are also probesets in each group that mapped to multiple genes (1371076_at mapped to Cyp2b1 and Cyp2b2) ans probesets that mapped to no genes.

Chemical	Significant Probsets
Leflunomide	466
Fluconazole	1997
Ifosfamide	0

Table 7. Number of probesets identified as significant at an adjusted p-value of less than 0.05 for each chemical group.

A Leflunomide

Probeset ID	Gene	Log FC	p-value
1370269_at	Cyp1a1	8.01	< 0.001
1387243_at	Cyp1a2	1.38	< 0.001
1372600_at	Fbxo31	1.22	< 0.001
1392946_at	Il1r1	1.63	< 0.001
1388611_at	Tcea3	-0.69	< 0.001
1370244_at	Ctsl	0.59	< 0.001
1373810_at	Pla2g12a	1.87	< 0.001
1398598_at		0.92	< 0.001
1376827_at	Eml4	0.78	< 0.001
1373814_at	R3hdm2	-0.76	< 0.001

B Fluconazole

Probeset ID	Gene	Log FC	p-value
1368731_at	Orm1	1.33	< 0.001
1377014_at		-2.31	< 0.001
1371076_at	Cyp2b1/2	2.42	< 0.001
1390255_at	Ablim3	1.79	< 0.001
1391570_at	Ablim3	1.54	< 0.001
1394022_at	Id4	-1.40	< 0.001
1380336_at	Irak3	1.33	< 0.001
1372136_at		-0.87	< 0.001
1398597_at	Rnf144a	-1.31	< 0.001
1377192_a_at	Clpx	-1.18	< 0.001

C Ifosfamide

Probeset ID	Gene	Log FC	p-value
1379397_at		-0.71	0.11
1374475_at	Abhd1	-0.35	0.30
1368273_at	Mapk6	-0.32	0.30
1371266_at	Afm	-0.26	0.30

1373217_at	Ehbp1	-0.36	0.30
1368718_at	Aldh1a7	0.97	0.30
1376481_at	Adamts9	-0.53	0.30
1378596_at		0.24	0.34
1377029_at		-0.50	0.34
1383137_at	Sox4	-0.54	0.34

Table 8. Table of the top ten differentially expressed genes for each chemical based on adjusted p-value. "Log FC" represents log fold change as computed by limma. Missing values in Gene correspond to probeset ids that are not present in the refSeq-to-probe id mapping matrix.

Histograms of log fold change for the chemical groups (Figure 12) show that all genes have a non-zero log fold change. The distributions for each are very similar, despite the large difference in number of genes found significant in each group. We filter by unadjusted p-value here because there were no probesets found significant at an adjusted p-value < 0.05 in the Ifosfamide chemical group. There is a probeset in the leflunomide group that had a log fold change of 8.01 (probeset id 1370269_at), which is why we see a long tail in the histogram.

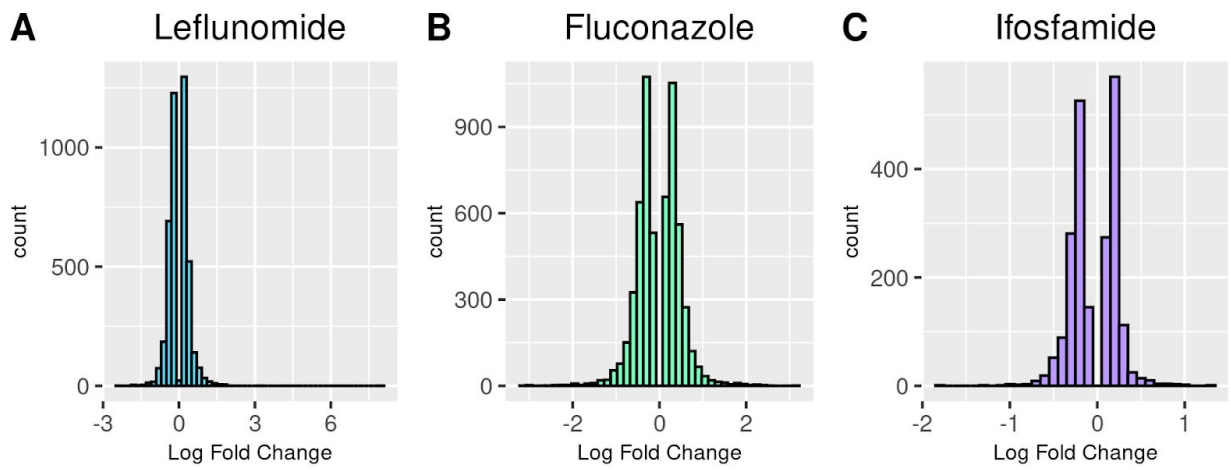

Figure 12 Histograms of log fold change for each chemical. We only include probesets with an unadjusted p-value < 0.05. Count is the number of genes with log fold change at a given value.

Volcano plots of log fold change vs. nominal (unadjusted) p-value are shown in Figure 13. Like in the histograms, there are no genes found in any group with a log fold change of zero. The scatter plot for ifosfamide is very sparse, which is a reflection of the small number of probesets that were found significant by the limma analysis. Only one probeset was found significant with an absolute log fold change of at least 1.5 in the ifosfamide group. In the leflunomide chemical group, 21 probesets met these criteria and in the fluconazole group, 74 probesets met this criteria. The top three most significant genes that meet a log fold change threshold of 1.5 for

each chemical group are summarized in table 9. Probeset id 1370269_at (gene Cyp1a1) in the leflunomide group stands out as very highly significant and has a relatively large fold change. This indicates that the chemical may have a significant effect on this gene in particular.

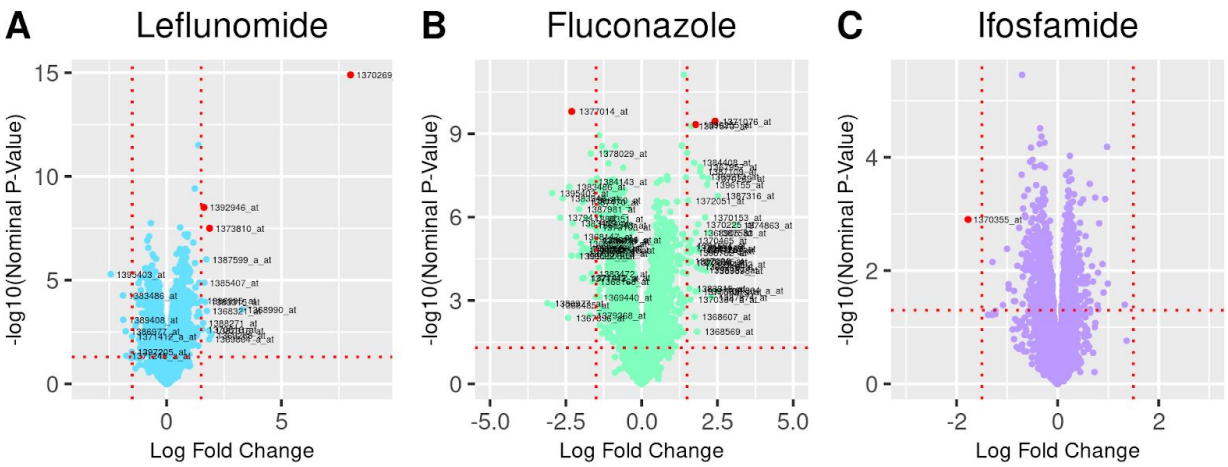

Figure 13. Volcano plots of log fold change vs. -log10 adjusted nominal p-value for each chemical. Horizontal line at p-value = 0.05 and the vertical bars are at log fold change ±1.5. Significant probesets that also meet the fold change requirement are labeled with their probe ID. Red dots are the most significant that meet the log fold change threshold and are found in Table _.

Leflunomide		**Fluconazole**		**Ifosfamide**	
1370269_at	Cyp1a1	1377014_at		1370355_at	Scd
1392946_at	Il1r1	1371076_at	Cyp2b1/2		
1373810_at	Pla2g12a	1390255_at	Ablim3		

Table 9. The top 3 significantly differentially expressed probesets for each chemical group with nominal p-value < 0.05 and absolute log fold change > 1.5. Probeset ID is on the left and the corresponding gene symbol is on the right.

Concordance

Probeset ids were mapped to their corresponding RefSeq ids to facilitate comparisons between the two sets, as outlined in the methods section. Table 10 summarizes the number of non-redundant DEGs that were found significant at an unadjusted p-value of 0.05 and absolute log fold change of 1.5 for each toxgroup and for both the microarray and RNA-seq data. Only one gene was significantly expressed at these thresholds in the ifosfamide group from the microarray data, which is further evidence that this chemical did not change expression levels for most of the mRNAs. For the microarray data, more DEGs were identified in the leflunomide group, but in the RNA-Seq data, more DEGs were identified in the fluconazole group.

Leflunomide	Microarray	22
	RNA-seq	640
Fluconazole	Microarray	56
	RNA-seq	345
Ifosfamide	Microarray	1
	RNA-seq	19

Table 10. The number of DEGs identified in each chemical group for both the microarray and RNA-seq results. All DEGs have an unadjusted p-value < 0.05 and an absolute log fold change of at least 1.5. These numbers are used to compute concordance between the sample types for each group.

Concordance between microarray and RNA-Seq is summarized in table 11. RefSeq ids were mapped based on row number to the RNA-Seq data. All concordance values are 0. This is most likely due to incorrect mapping. None of the genes in the microarray DEG groups were found in the corresponding RNA-seq DEG group. We also stratified the analyses based on DEGs that fell above or below the median expression level in each group. The concordances were also zero because no genes matched between the groups overall.

Leflunomide	Overall	0
	Above	0
	Below	0
Fluconazole	Overall	0
	Above	0
	Below	0
Ifosfamide	Overall	0
	Above	0
	Below	0

Table 11. Concordance between microarray and RNA-Seq data for each chemical group. Overall: all DEGs; Above: DEGs above median expression level for that group; Below: DEGs below median expression level for that group.

In Figure 14, concordance is shown as a function of the total number of DEGs identified in both the microarray analysis and the RNA-Seq analysis. Note that the chemical groups show in different orders due to the inconsistent relative number of DEGs identified in each. Figure 15 shows a histogram of concordance for each group for overall DEGs and DEGs that fall above and below median expression levels for that group.

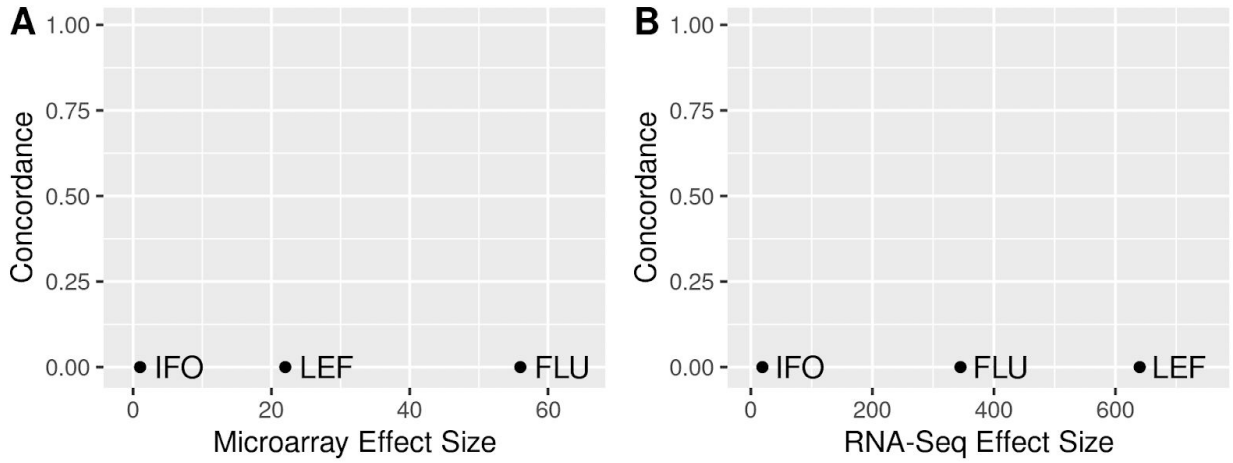

Figure 14. Concordance for each chemical group as a function of effect size, or the number of DEGs found in the microarray and RNA-Seq analyses.

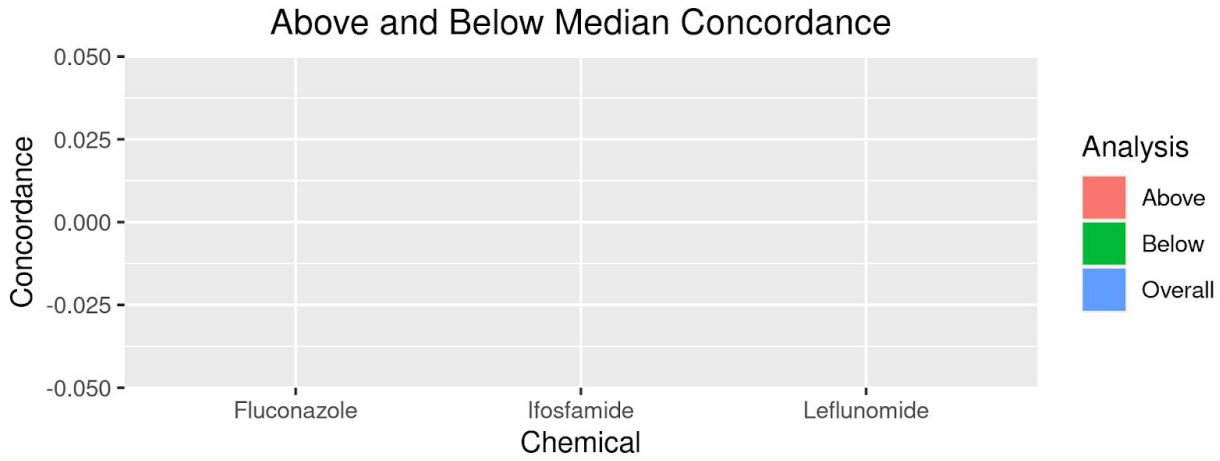

Figure 15. Concordance between microarray and RNA-Seq analyses for each chemical group. Overall: all DEGs; Above: DEGs above median expression level for that group; Below: DEGs below median expression level for that group.

Concordance on Example Data

Performing concordance analysis on the example data provided, there are 49 DEGs identified in the microarray set and 708 DEGs identified in the RNA-Seq set. Overall concordance is computed as described in the methods, and was 10.0% for the example data. Concordance by effect size for the microarray and RNA-Seq sets is shown in

Figure 16. When DEGs are stratified by above and below median expression, the above set has similar concordance to overall (10.1%), but the below set has a much smaller concordance (2.1%) (Figure 17).

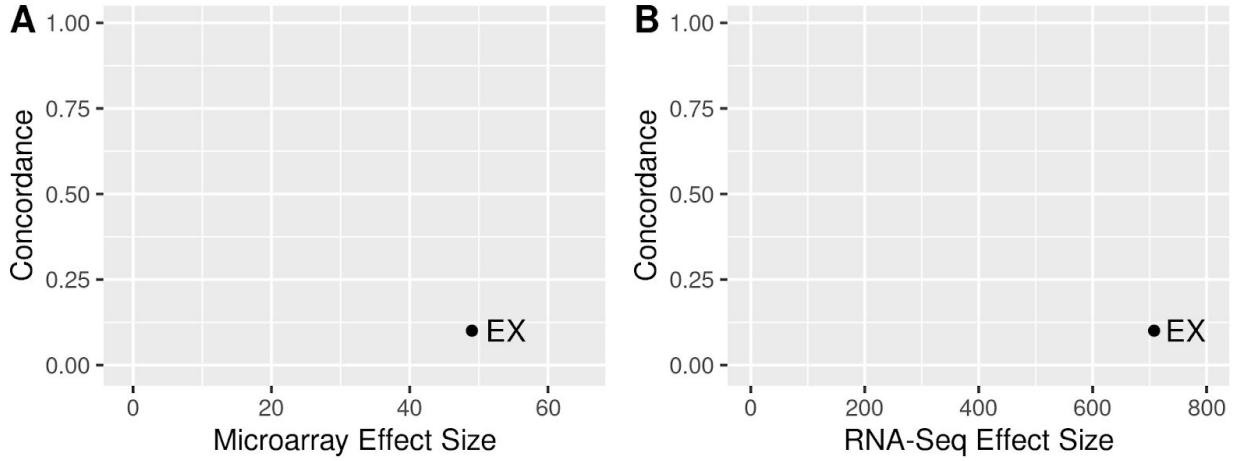

Figure 16. Concordance for the example data set as a function of effect size, or the number of DEGs found in the microarray and RNA-Seq analyses.

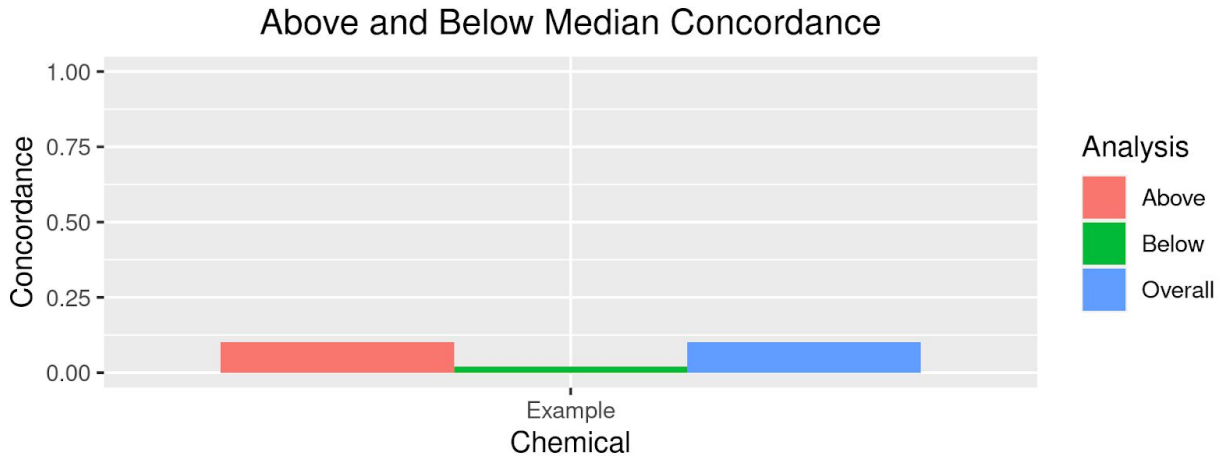

Figure 17. Concordance between microarray and RNA-Seq analyses for the example data set. Overall: all DEGs; Above: DEGs above median expression level for that group; Below: DEGs below median expression level for the example set. Above=10.1%; Below=2.1%, Overall=10.0%.

Biological Analysis

Functional enrichment was performed to assess pathway enrichment in our samples. Note that this analysis was performed on the example output from DESeq, so the samples it actually corresponds to are unknown. For the top 100 genes with the highest log fold change, the results show enrichment of genes related to steroid synthesis, but this is not consistent with any of the groups in the original study.

Type	GO Term	Adjusted p-value
Biological Process	cholesterol biosynthetic process	2.2e-8
	steroid biosynthetic process	9.4e-5
	sterol biosynthetic process	1.1e-4
	isoprenoid biosynthetic process	4.5e-4
	oxidation-reduction process	5.4e-4
Cellular component	endoplasmic reticulum membrane	6.9e-5
	intracellular membrane-bounded organelle	3.3e-2
	endoplasmic reticulum	2.8e-2
Molecular function	iron ion binding	2.0e-2
	steroid delta-isomerase activity	1.7e2

Table 12. Functional annotation results for example data. Up to the top five significant results for each type are shown.

Finally, a clustered heatmap was generated from the normalized counts of all samples in all three conditions. The genes used are the joining of the top 50 with the greatest log fold change in each condition. Qualitatively, there are gene expression patterns which suggest each chemical treatment may have a distinct MOA. The clustering also shows that replicates from the same treatment have greater similarity to one another than to the other samples.

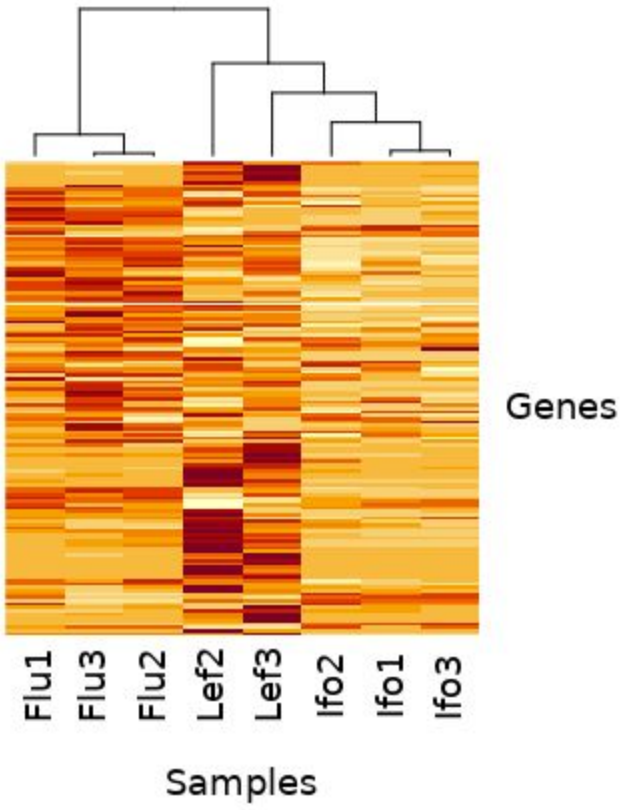

Figure 18. Clustered heatmap for the top 143 genes differentially expressed between each condition and their associated controls. The eight samples are labeled according to chemical treatment of either fluconazole, leflunomide, or ifosfamide.

DISCUSSION

The raw data from bam files was run in featurecounts for count-based differential expression analysis. However instead of the expected nine files we get only eight which caused discrepancy in the subsequent analyses. The multiqc report compiled the summary files of feature count to give the percent reads that were successfully mapped. With the obtained csv file containing the count matrix, DESeq2 analysis was undertaken to determine differential expression between your treated samples and appropriate controls.

In our limma analysis on the microarray data, we found no probesets significantly expressed in the ifosfamide samples. This could result from the ifosfamide group not inducing a change in

expression levels for our genes. For the microarray data, there were 22, 56, and 1 DEGs identified in the leflunomide, fluconazole, and ifosfamide chemical groups respectively. For the RNA-Seq data, there were 640, 345, and 19 DEGs identified in the same chemical groups. The relative number of DEGs identified in the microarray and RNA-seq data sets were inconsistent, which could be due to differences in the methods, or in our filtering of the sets to identify DEGs.

Concordance of the microarray and RNA-seq analyses was computed as 0 for all genes in the leflunomide chemical group, 0 for genes in the fluconazole group, and 0 for genes in the ifosfamide group. The RNA-Seq results were organized by “gene number” and we decided to map this to the corresponding row in the mapping matrix to determine the RefSeq id for comparison to the microarray data. This resulted in no intersection of DEGs between the two sample types. Therefore, this is most likely the incorrect mapping method and the reason we were unable to determine concordance.

We computed concordance for the example data to demonstrate results similar to what we expected for our toxgroup. We were able to compute concordance for this sedata using the same methods as attempted for our toxgroup. However, this data was successfully mapped between the microarray and RNA-Seq sets because we had the RefSeq ID available for the RNA-Seq results.

In constructing the heatmap, we successfully showed that samples from the same treatment and MOA clustered together. Many approaches were possible when filtering the data. Selecting genes by p-value would include both more highly and lowly expressed genes relative to controls. However, because we are plotting the normalized counts, the result was many genes with low counts, and the heatmap appeared random. By using the highest positive log fold change instead, those with high expression in the treatment, and therefore the most biological relevance, are selected. In addition, 50 genes per condition was chosen because at lower values, samples did not cluster as anticipated. Finally, because not all genes will be shared in common between conditions, the intersection only contained a total of 143 genes.

CONCLUSION

There was no concordance between microarray and RNA-Seq data among the samples for the leflunomide, fluconazole, or ifosfamide chemical groups. This is most likely due to incorrect mapping from RefSeq ids in the DESeq results. However, we computed concordance for the example data set to be 10.0% overall, 10.1% in the above median DEG set, and 2.1% in the below median DEG set. These results demonstrate that we were able to replicate the concordance methods in the original paper, but we were not able to replicate it with the raw data due to mapping inconsistencies.

REFERENCES

[1] Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., ... & Meehan, J. (2014). A comprehensive study design reveals treatment-and transcript abundance–dependent concordance between rna-seq and microarray data. *Nature biotechnology*, *32*(9), 926.
[2] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). “limma powers differential expression analyses for RNA-sequencing and microarray studies.” Nucleic Acids Research, 43(7), e47.
[3] 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.*