Project 3: Concordance of Microarray and RNA-Seq Differential Gene Expression

Alec Jacobsen, Daisy Wenyan Han, Divya Sundaresan, Emmanuel Saake
Data Curator Programmer Analyst Biologist

ENG BF528, Spring 2021

Introduction

New technology is critical to the advancement of scientific research, but the efficacy of this technology should be extensively tested before drawing conclusions from its results. Recently, RNA sequencing (RNA-Seq) has risen to prominence in studies on differential gene expression analysis, but its ability to detect differentially expressed genes (DEGs) is still open to scrutiny. Its predecessor, microarrays, have been the principal method used for detecting DEGs, and have been extensively tested to ensure the validity of their results through the Microarray Quality Control Project. Concordance between microarray and RNA-Seq data should, therefore, be indicative that RNA-Seq is a valid method for detecting DEGs as well.

Previous studies comparing microarrays and RNA-Seq produced mixed results, with some suggesting that RNA-Seq has a lower precision for finding weakly expressed genes (Łabaj et al. 2011, McIntyre et al. 2011), while others have found RNA-Seq to be more highly sensitive (Mooney et al. 2013, Sultan et al. 2008). The likely reason behind this discrepancy in results, as noted by Wang et al., is the limited treatment conditions used in the studies, failing to cover the wide range of biological complexity. Additionally, none of the studies have compared the ability of RNA-Seq or microarrays to predict toxicity outcomes based on gene expression data.

The purpose of Wang et al.'s *A Comprehensive Study Design Reveals Treatment- and Transcript Abundance–Dependent Concordance Between RNA-Seq And Microarray Data*, was therefore to rigorously test the differences in DEG detection between microarrays and RNA-Seq, as well as assess their usefulness in developing predictive models. To do this, the authors examined gene expression differences in rat livers due to extensive chemical treatments over multiple biological replicates. mRNA abundance was then detected with both Affymetrix microarrays and RNA-Seq. They found that the concordance between mRNA detection methods increased with increasing severity of chemical treatment, with RNA-Seq more sensitive to detection of weakly expressed genes than microarrays. They were therefore able to conclude that both techniques resulted in similar gene expression based models for toxicity outcomes.

The purpose of this report is to recapitulate the authors' bioinformatics methods in an attempt to replicate their findings, understand the techniques that they used, and assess the reproducibility of their results given the upgrades in softwares since Wang et al. 2014 was published.

Data

Data was sourced from NCBI's Sequence Read Archive (accession number SRP039021) and Genome Expression Omnibus (accession numbers GSE55347 and GSE47875). These data were generated from male Sprague-Dawley rats, ages 6-8 weeks, who had been administered test chemicals mixed with corn oil or water either orally or via injection. Animals received doses daily for three, five, or seven days depending on the administered drug. Livers were harvested within 24 hours of the last injection for analysis. Reads were prepared with the Illumina TruSeq RNA Sample Preparation Kit and SBS Kit v3 and sequenced on Illumina HiScanSQ or HiSeq2000 systems. Additionally, sample DNA was hybridized to the Affymetrix whole genome GeneChip® Rat Genome 230 2.0 Array for microarray analysis.

Microarray data were preprocessed prior to analysis, while read expression data were in raw sequence read form. Tox group 3 was chosen for analysis. This included samples SRR1177981, SRR1177982, SRR1177983, SRR1178008, SRR1178009, SRR1178010, SRR1178014, SRR1178021, SRR1178047. These comprised three biological replicates of each of the chemicals Leflunomide, Fluconazole and Ifosfamide, respectively. FastQC and MultiQC were used to assess read quality, and found no reads that failed to meet quality standards (*Figures 1 and 2*). Alignment of reads to the rat genome was performed with STAR aligner with the default settings. Alignment quality statistics are reported in *Table 1*. Again, none of the alignments failed to meet quality standards.

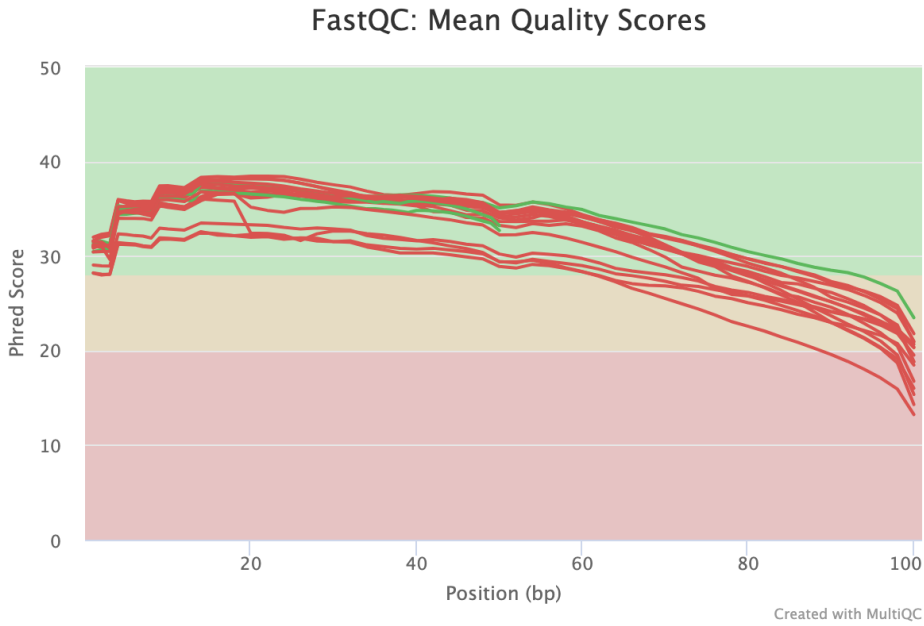


Figure 1. *Mean quality (Phred) scores across reads for all samples*

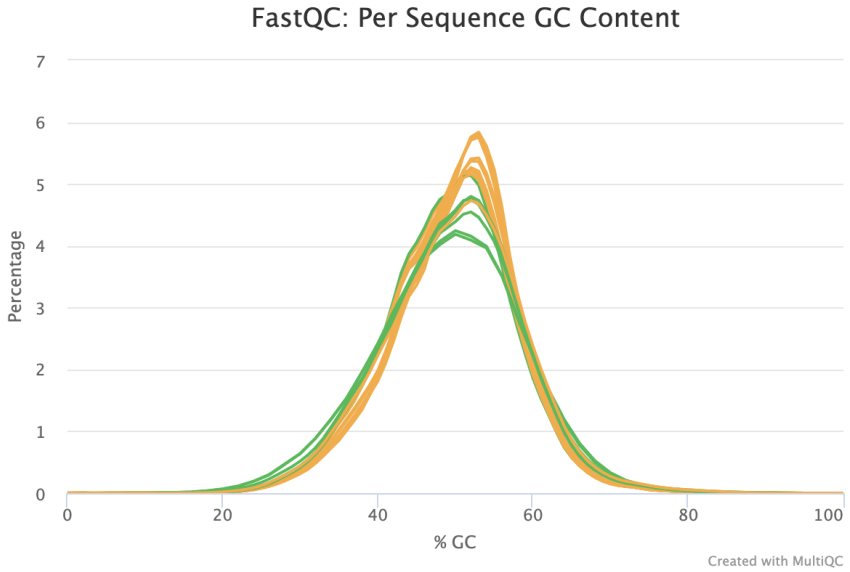


Figure 2. *The GC content (%) for all sequence reads in all of the samples.*

Table 1. *Alignment quality summary statistics for sample reads aligned by STAR aligner.*

Alignment Statistics				
Sample	% Uniquely Mapped Reads	Mismatch Rate per Base, %	% Multi-Mapped Reads	% Unmapped Reads - Too Short
SRR1177981	80.19	0.76	3.78	15.82
SRR1177982	83.82	0.77	3.67	12.26
SRR1177983	79.61	0.86	3.52	16.63
SRR1178008	88.08	0.59	3.52	8.11
SRR1178009	90.09	0.66	2.54	7.19
SRR1178010	90.63	0.73	2.72	6.47
SRR1178014	83.52	0.44	6.57	9.20
SRR1178021	81.95	1.10	5.77	11.92
SRR1178047	83.98	0.88	5.85	9.67

Methods

Read Counting Using featureCounts

Using the STAR-aligned BAM files generated in the previous section, count matrices for each of the nine samples, spanning three different treatment groups, were generated using the featureCounts program. The featureCounts program is a software program developed for counting reads to genomic gestures such as genes, exons, and promoters, and is contained within the Subread package, which can be executed on the command line. Since its initial release in 2014, it has been considered a highly efficient method for read summarization, operating an order of magnitude faster than existing protocols for gene-level analyses, and requiring far less computer memory (Yang, 2014). The reference GTF file used to generate these matrices was provided pre-downloaded in the class directory on the shared computing cluster. No additional arguments were provided for the featureCounts command, other than the number of threads required to run. The nine count matrices took less than a minute to generate, on sixteen threads. featureCounts v1.6.2 was used for the generation of these matrices.

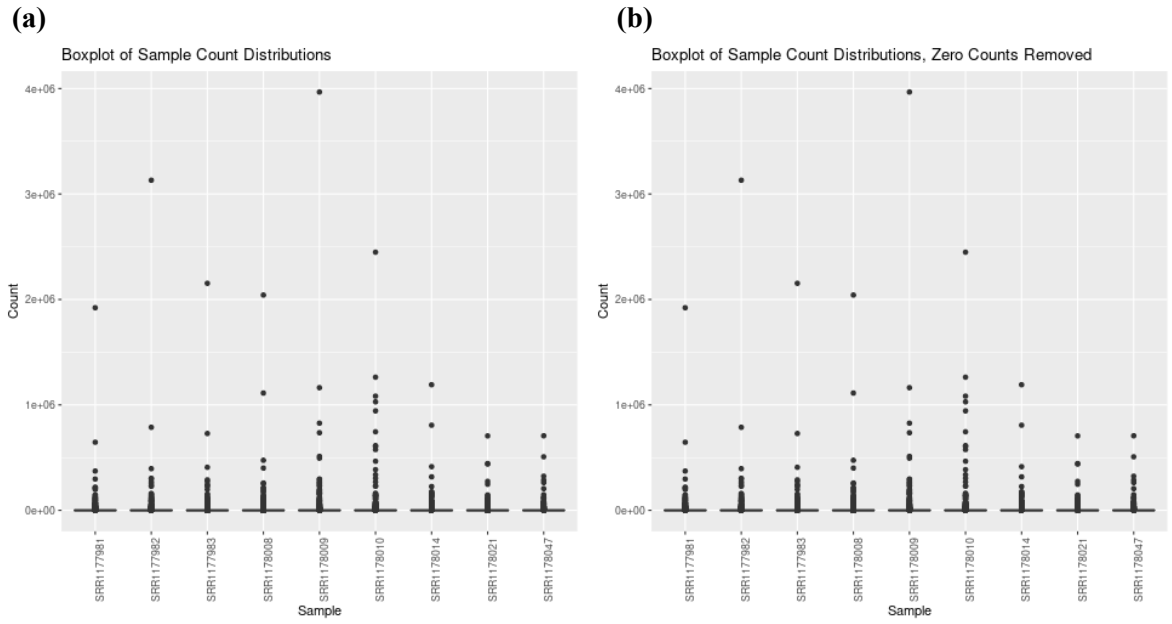


Figure 3*. Box Plots illustrating distribution of count values, for each of the nine treatment samples, before (a) and after (b) removal of rows containing zeros across all rows (genes).*

The counts for each of the samples varied greatly, as can be seen in *Figure 3*, as shown above. Namely, the samples appear to consist primarily of genes having extremely low counts, with the occasionally overexpressed gene in each sample. These genes are overexpressed so much so that the interquartile range, which is typically represented as the "box" in a box plot, does not appear to be visible. This suggests that the majority of the genes present in each sample are present at extremely low counts. Interestingly, when removing rows containing all zeros across the matrix (corresponding to genes that were not present in any of the samples), the distributions of the count samples did not appear to change significantly. This means that while each sample contains many zero-count genes, these genes happen to appear in other samples, and can therefore not be removed from the analysis in general. This then suggests that multiple genes are differentially expressed across the different samples.

To further explore the distribution of these counts, an arbitrarily selected upper limit was placed on the box plot to examine the change in count distributions following the removal of zero-count genes. These are shown in *Figure 4*.

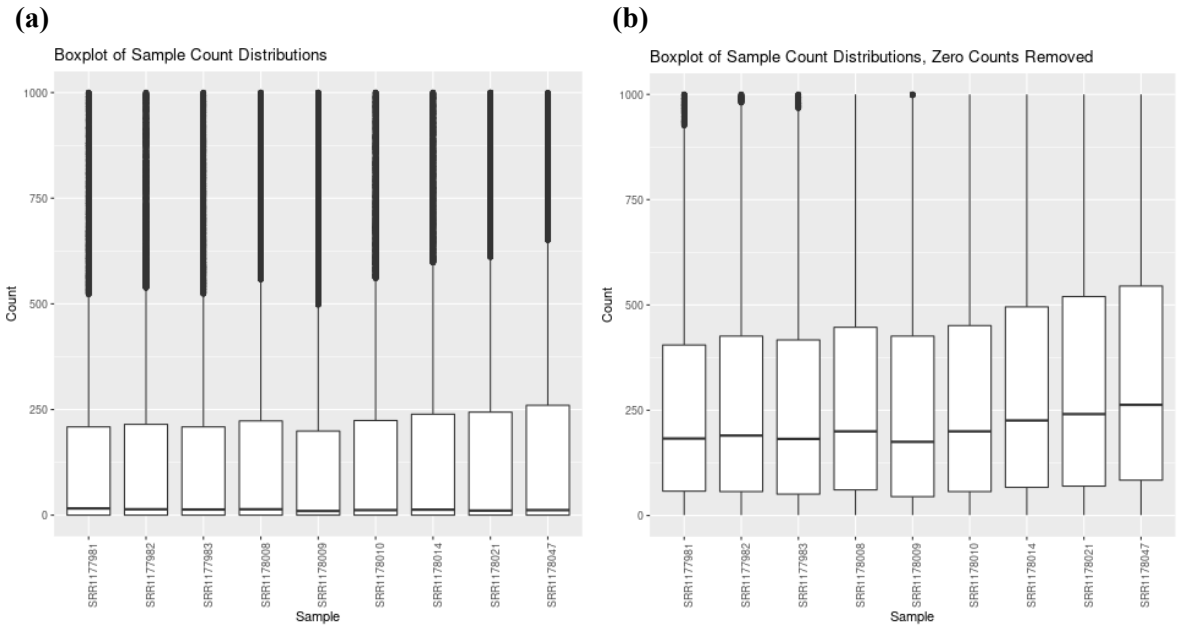


Figure 4. *Boxplot of the count distributions count values for the nine treatment samples, both (a) before, and (b) following the removal of genes will zero-counts across all samples. An arbitrary upper limit of 1000 was selected to better visualize the change in distribution following the removal of genes containing zero counts across all samples. A rescaled boxplot is available in the Supplementary Data (Supplementary Figure 1).*

Figure 4, above, demonstrates that while the change in distributions is difficult to illustrate in comparison to the extreme outliers within the samples as per *Figure 3*, there is a noticeable difference in the count distributions of each sample following the removal of zero-valued genes when zooming in. The median and interquartile ranges have increased significantly with the removal of these genes; such drastic change in the data can therefore be expected to have a tangible effect on downstream analysis and the differential gene expression analysis to be performed.

The featureCounts program also generated a summary file for each count matrix, outlining the number of assigned and unassigned reads. These summary files were then loaded into MultiQC for quality assurance, as outlined below.

Quality Control on Count Matrices

MultiQC is a reporting tool that parses through the summary statistics generated from previously run bioinformatics tools, in order to provide a single HTML file with all relevant summary statistics. In our analysis, MultiQC was provided the nine summary files generated when running featureCounts on the nine corresponding BAM files.

The full table was generated, showing the percentage of each of the reads that were assigned to a genomic location, and the number of reads (in millions) that were assigned can be found in the Supplementary Data (*Supplementary Table 1*). These same results can also be summarized by the plots in *Figure 5*.

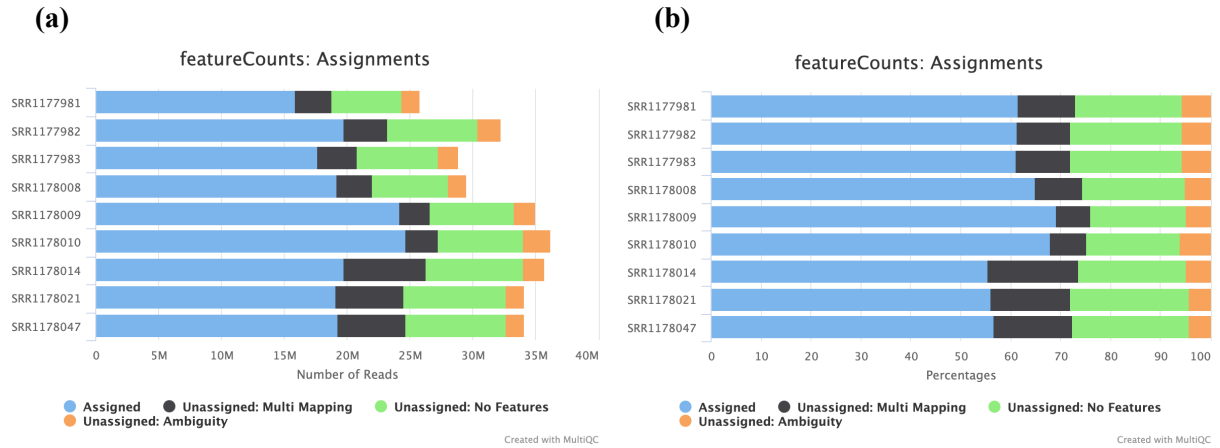


Figure 5. *MultiQC results on summary reports generated from featureCounts count matrix generation of the nine BAM samples, showing (a) the number of reads aligned per sample, and (b) the percentage breakdown of the assignment status of reads for each sample.*

From the above figure, it appears that the reads do not differ significantly from each other, in regards to the percentage of reads assigned, or the number of reads (in millions) themselves that were able to be aligned to the reference GTF annotation. The percentage assigned ranged from 55.4% aligned to 69.2% aligned - overall, a fairly small difference across samples.

It is important to note that the samples were loaded with their respective treatment groups in mind. This means that the first three samples represented (SRR1177981, SRR1177982 and SRR1177983) will be compared to each other, and the next three following, and so on. When looking at the values in this sense, the values are all very similar across treatment groups. For example, the first treatment group (AhR MOA), comprising the aforementioned three samples, have percentages of reads ranging from 61.1% to 61.5%. The following treatment group (CAR/PXR MOA) has assignment percentages ranging from 64.9% to 69.2%. The last treatment group has percentage assignments ranging from 55.4% to 56.6%. This is best illustrated in *Figure 3(b)*, where the groups of three have very similar colorings within their respective MOA grouping. For each MOA to be examined, the percentages of assigned reads is within a 5% difference amongst the three samples. Because all the samples had similar percentages of assigned reads, with no significant outliers detected, no samples were removed from the analysis.

For this project, MultiQC was loaded via conda, and was run using version v.1.10. The nine samples, run on eight cores, took less than a minute to generate the MultiQC results.

RNA-Seq Differential Gene Expression Using DESeq2

DESeq2 is an R package, available through Bioconductor, that makes use of negative binomial generalized linear models in the detection of differentially expressed genes. We made use of this package

in our analysis to identify and examine the differentially expressed genes across our samples and their corresponding “control” methods.

More specifically, we were examining the differences between the Leflunomide, Fluconazole and Ifosfamide treatments. Their corresponding control treatments were those with the same vehicle, indicating they had been delivered to the cells in the same fashion; this was corn oil (100%) for Leflunomide and Fluconazole, and saline (100%) for Ifosfamide. By doing so, this enabled us to evaluate the differentially expressed genes by mode of action (MOA), which for our samples, were AhR, CAR/PTX and DNA Damage, respectively. The sample information, containing the treatment chemical, targeted mode of action and corresponding control samples, were provided for us as a CSV file on the SCC.

The input values of the DESeq2 package consists of raw count data, in the form of a matrix. It is important that these be un-normalized values, as the DESeq2 model internally corrects for library size. Our counts matrices for this analysis were those obtained from featureCounts. The DESeq package also expects a dataframe consisting of sample information, in the same order as the columns of the counts matrix being inputted. This was obtained by subsetting the sample information CSV provided on the SCC. This created a *DESeqDataSet* object. Of note, genes were pre-filtered such that only rows containing one or more reads (corresponding to one read across all samples) were kept, in order to reduce the memory size of the DESeqDataSet object. Minimal filtering was performed, as a stricter, independent filtering step is applied by DESeq in later analysis. This decreased the number of genes in our analysis by 7391 genes, from 18014 genes pre-filtering to 10695 genes post-filtering.

DESeq2 outputs a dataframe, consisting of the gene name, nominal and adjusted p-values, and log fold change for each of the conditions. This was exported to CSV for each of the three conditions for further analysis, and to be used in the evaluation of concordance between RNA-Seq and microarray data. DESeq2 version 1.26.0 was used in this analysis, and took less than a minute to compute the differentially expressed genes for each of the three MOA groups.

Microarray Differential Gene Expression Using Limma

Limma stands for linear models of microarray data. It is a tool that is used for taking microarray data and converting it into output of differentially expressed genes. For this project, since we were given pre-normalized data, we did not need to conduct any QC or normalization. We were therefore able to proceed straight to the differential gene expression analysis. The authors used a combination of RMA normalization and Limma to determine differential expression between the different treatments and control samples, which is what I followed to achieve the differentially expressed genes. I split samples into their respective treatments (Leflunomide, Fluconazole, and Ifosfamide) and paired them with their respective controls which made sure the vehicles (corn oil, saline) were consistent between the control and experimental groups.

The central idea is to fit a linear model to the expression data for each gene. Empirical Bayes was used to borrow information across genes making the analyses stable even for experiments with a small number of arrays. We used two matrices, a design matrix which provides a representation of different targets, and the second a contrast matrix which allows the coefficients defined by the design matrix to be combined into

contrasts of interest. We then fit a linear model, where each row of the design matrix array in the experiment and each column corresponds to a coefficient. One purpose of this step is to estimate the variability in the data. The Limma package contains the topTable function, which allows us to view summary statistics of our linear model that includes the t-statistic, p-value, logFC, and adjusted p-value, which we can save as a CSV. From here I was able to select my top differentially expressed genes using the adjusted p-value. (Smyth 2005).

We then used these results to calculate concordance. Concordance is the overlap of the DEGs from the two platforms with agreement in the direction of fold change. The concordance was adjusted to remove the contribution of random chance. Since n_0 (observed overlapped genes), n_1 (DEG micro), n_2 (RNA-seq), and N (total genes in common set: 13079) are known, we can solve this equation for n_x to find the background corrected number of "true" overlapping genes.

$$n_x = \frac{Nn_0 - n_1 n_2}{n_0 + N - n_1 - n_2}$$

Once we solve for n_x we can place that in as the intersect and use the equation to solve for concordance between the two platforms.

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

Gene Set Enrichment Analysis

Wang et al. derived statistically significant pathways from their DEGs for each chemical, and summarised key pathways associated with xenobiotic activities, as shown in their Supplementary Table 4. Similarly, in this project, the differentially expressed genes from our analysis were mapped to pathways by undertaking pathway enrichment analysis via the online bioinformatics tool, the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Huang et al., 2009).

First, each list of genes (from RNAseq) corresponding to MOA were filtered using the threshold *padj* < *0.05* to capture only statistically significant genes for the enrichment analysis. After applying this filtering threshold, we had 1389 DEGs for AhR, 3499 DEGs for CAR, and 91 DEGs for DNA. The resulting list in each case was then ordered in descending order by *padj* (adjusted P). This process created a list of significant GeneIDs (i.e. the Genbank_accession identifiers of each gene), which was then copied and pasted into the online DAVID tool. From the pathway section of the output, both KEGG_pathway and REACTOME_pathway were considered in comparing our enriched pathways to that identified in the Supplementary Table 4 by Wang *et al* (2014). Emphasis was placed on these three MOA in our tox group: AhR, CAR and DNA. The aforementioned *padj* filtering was also applied to the three sets of microarray based DEGs generated from Limma, and we had 1997 DEGs for Fluconazole(CAR), 446 DEGs for Leflunomide treated samples (AhR), and zero significant DEGs for Ifosfamide (DNA) treated samples.

Clustered Heatmap

The three normalized count matrices from the DESeq2 object were combined into one larger matrix, by GeneID. The coalesced data was scaled using the "average" method. A distance matrix was computed with the 'euclidean' option. Combining *hclust()* and *cutree()*, the normalized count was clustered and limited to 3 clusters. From this process a clustered data object was created, and used to produce the clustered heatmap.

Of note, the total number of samples is fifteen. However in the clustered heatmap we have eighteen. That is because we had three overlapping control samples, SRR1178050, SR1178061and SRR1178063 common to both AhR and CAR . Because their normalized counts are different, both sets of overlapping samples were maintained in the coalesced data for the clustering purpose in the heatmap.

Results

Differentially Expressed Genes Using RNA-Seq

The top three differentially expressed genes for each mode of action (AhR, CAR/PXR and DNA Damage, respectively) are shown below, with their respective GeneIDs, Log2FC, and nominal and adjusted p-values. The top differentially expressed genes were selected as those having the lowest adjusted p-values following differential gene expression analysis. A full table of the top ten differentially expressed genes for each treatment, as produced by the DESeq2 program, can be found in the Supplementary Data.

Table 2A. *Top three differentially expressed genes for AhR (Leflunomide Treatment), ordered by increasing adjusted p-value.*

Gene ID	Log2 Fold Change	p value	Adjusted p-value
NM_001007722	-4.71810604478048	1.91169121334766e-47	6.79287611142868e-44
NM_001012174	2.15742300128457	4.79425556173557e-30	5.11067642881011e-27
NM_001130558	-4.77594855706857	5.01467541818446e-38	1.06912879915693e-34

Table 2B. *Top three differentially expressed genes for CAR/PXR (Fluconazole Treatment), ordered by increasing adjusted p-value.*

Gene ID	Log2 Fold Change	p value	Adjusted p-value
NM_001130558	-5.76775475610783	2.53243944371547e-91	1.35282915083281e-87
NM_001134844	5.78744880535407	7.33497489621203e-86	2.61222905970431e-82
NM_013033	4.55216854516338	3.71628354855532e-48	7.940954686553e-45

Table 2C. *Top three differentially expressed genes for DNA Damage (Ifosfamide Treatment), ordered by increasing adjusted p-value.*

Gene ID	Log2 Fold Change	p value	Adjusted p-value
NM_001007722	-1.7663002728252	9.60234085535284e-61	5.03306695933319e-57
NM_001013057	-0.840211636329588	6.89478341343649e-10	9.03475181538184e-07
NM_001107084	-1.0093195609329	1.52820941662706e-11	2.2886027592145e-08

Of the inputted genes, 1389 were found to be significantly different in the AhR treatment group, 3499 in the CAR/PXR treatment group, and 91 in the DNA Damage treatment group. Genes were identified as significant if their adjusted p-values fell below the threshold of 0.05, as set by Wang et al.. Their respective Log2 Fold Change frequencies are shown in *Figure 6* below, and can be used to examine the distribution of up- and down- regulated genes in each of the treatments. From the figures below, it would appear the CAR/PXR treatment group has the most up- and down- regulated genes, while the DNA Damage group has the least.

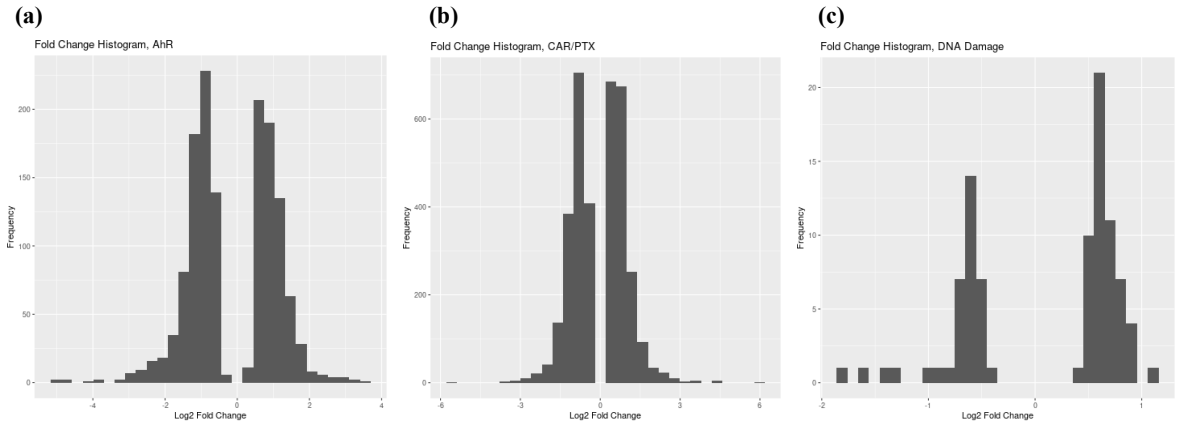
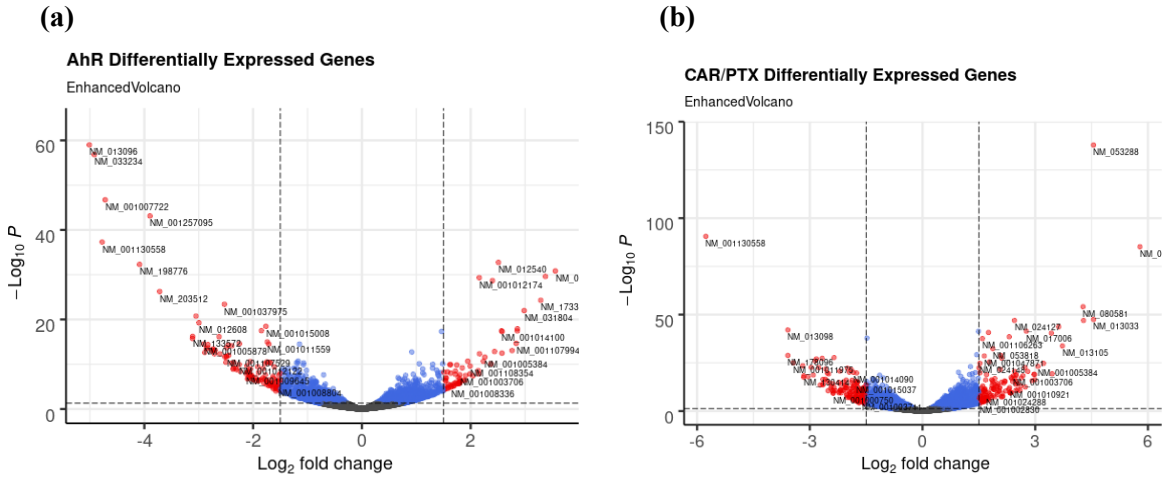


Figure 6. *Histograms of Log2 Fold Change values for each of the three treatments: (a) AhR, treated with Leflunomide, (b) CAR/PXR, treated with Fluconazole, and (c) DNA Damage, treated with Ifosfamide. Only genes meeting the p-value significance cutoff of $p < 0.05$ are represented in the above plots.*

The number of differentially expressed genes also becomes increasingly apparent in the above plots. The DNA Damage MOA plot, for example, shows markedly fewer differentially expressed genes than its two counterparts. There also appears to be more up-regulated genes than down-regulated genes for this particular treatment, in comparison to the other two treatments, which appear to have more equally distributed up- and down- regulated genes. This can also be visualized in the volcano plots in *Figure 7* below.



(c)

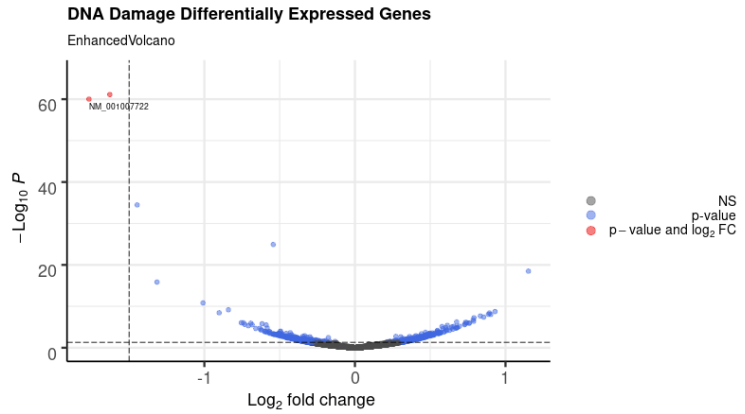


Figure 7. *Volcano Plots of the differentially expressed genes for each of the three treatments, plotting the Log2 Fold Change against the negative log of the nominal p-value. The p-value cutoff for significance was set at $p < 0.05$, and the LogFC cutoff for significance was set at 1.5, as per the Wang et al. paper. The red points indicate GeneIDs which satisfy both the LogFC and p-value significance cutoffs, blue points indicate GeneIDs which satisfy only the p-value cutoff, but not the LogFC cutoff, while the grey points are not significant by either metric.*

The difference in the numbers of genes meeting specific cutoffs for each of the three treatments is made more apparent when plotting the number of genes meeting the significance cutoff. For our analysis, we used the original cut-offs set by Wang et al., with a p-value significance of p-adjusted < 0.05, and LogFC < 1.5. These volcano plots differ from the histograms in *Figure 6,* as they represent not only just the significant genes, but all genes used in the differential gene expression analysis. The number of genes that were deemed insignificant, therefore, are also visualized.

As illustrated above, a large number of genes were able to satisfy these criteria in the Leflunomide and Fluconazole treatments (AhR and CAR/PXR Mode of Actions, respectively), while only two were able to meet these criteria in the Ifosfamide treatment (DNA Damage MOA). Of note, there were no genes in any of the three MOA categories in which the p-value significance cutoff was not met, yet the LogFC cutoff was.

Differentially Expressed Genes Using Microarray

The top ten differentially expressed genes for each mode of action AhR (Leflunomide), CAR/PXR (Fluconazole) are shown below *Figure 8, 9* with gene name, logFC, and nominal and adjusted p-values. The top differentially expressed genes were selected as those having the lowest adjusted p-values as well as |logFC| > log2(1.5) following differential gene expression analysis. As DNA damage (Ifosfamide) has 0 DEG there is no table for it.

This is the top 10 genes adj.p-val <0.05 LEFLUNOMIDE

PROBEID	SYMBOL	logFC	t	P.Value	adj.P.Val
1370269_at	Cyp1a1	8.0138424	14.584463	1.275747e-15	3.967445e-11
1387243_at	Cyp1a2	1.3836944	10.876477	3.089979e-12	4.804763e-08
1372600_at	Fbxo31	1.2264862	8.905528	3.836092e-10	3.976621e-06
1392946_at	Il1r1	1.6277989	8.105409	3.137490e-09	2.439320e-05
1388611_at	Tcea3	-0.6849730	-7.462977	1.796257e-08	1.117236e-04
1370244_at	Ctsl	0.5897065	7.294601	2.860294e-08	1.400901e-04
1373810_at	Pla2g12a	1.8711517	7.259451	3.153255e-08	1.400901e-04
1376827_at	Eml4	0.7839707	7.004813	6.415588e-08	2.216871e-04
1373814_at	R3hdm2	-0.7612068	-6.937368	7.752145e-08	2.290109e-04
1378016_at	Eml4	0.8975318	6.921732	8.100324e-08	2.290109e-04

Figure 8
On the right are the top 10 DEG for Leflunomide based on adj. P-val. I have included the symbols so we can see which specific genes are the top expressed. The max adj. P-value is 3.96 e -11 and goes up to 2.29 e-04. The total amount of DEG based on adj. p-value with added filtering of |logFC| > log2(1.5) cutoff is **183** total DEG.

This is the top 10 genes adj.p-val <0.05 FLUCONAZOLE					
PROBEID	SYMBOL	logFC	t	P.Value	adj.P.Val
1368731_at	Orm1	1.392496	11.062552	7.618510e-12	2.369280e-07
1371076_at	Cyp2b2	2.420877	9.333393	3.534737e-10	3.278103e-06
1371076_at	Cyp2b1	2.420877	9.333393	3.534737e-10	3.278103e-06
1390255_at	Ablim3	1.782758	9.213977	4.673357e-10	3.278103e-06
1391570_at	Ablim3	1.641168	9.162804	5.270432e-10	3.278103e-06
1394022_at	Id4	-1.403790	-8.839012	1.136907e-09	5.892779e-06
1380336_at	Irak3	1.327452	8.484594	2.679584e-09	9.637781e-06
1398597_at	Rnf144a	-1.307325	-8.468191	2.789158e-09	9.637781e-06
1377192_a_at	Clpx	-1.180871	-8.262826	4.620568e-09	1.335853e-05
1390801_at	Tmem252	-1.093168	-7.892586	1.163930e-08	2.585504e-05

Figure 9
On the right are the top 10 DEG for Leflunomide based on adj. P-val. The max adj. P-value is 2.36 e -07 and goes up to 2.58 e-05. The total amount of DEG based on adj. p-value and with |logFC| > log2(1.5) cutoff is **726** DEG.

*NOTE the tables were selected on adj. P-value and |logFC| > log2(1.5).

Of the genes imputed into LIMMA just based on adjusted p-value the number of DEG at p-adjust < 0.05 and |logFC| > log2(1.5): *Leflunomide: 183, Fluconazole: 726, Ifosfamide: 0*. The MOA CAR/PXR (Fluconazole) had the most differentially expressed genes and AhR followed in second. Down below you can see the histograms and scatter plot which further show the distribution of logFC and individual genes vs. their logFC.

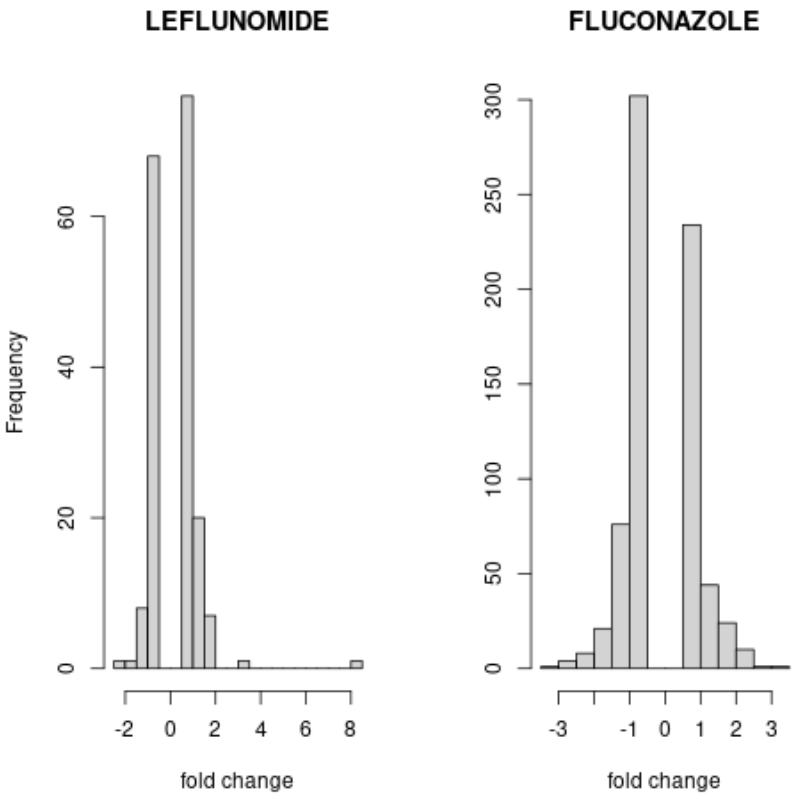


Figure 10 Here are the distributions for the logFC for differentially expressed genes of AhR (Leflunomide), CAR/PXR (Fluconazole). We are looking at the significant DEG's by adjusted p-value < 0.05 as well as the 1.5 logFC cutoff. We can see that there is a much higher amount of DEG's for the CAR/PXR MOA compared to the Ahr, and DNA (Ifosfamide) which is lowest at 0 DEGs. For leflunomide we can see a bit more up regulated genes and for fluconazole a bit more down regulated genes. However there is not that big of a difference as both treatment groups seem close to be close to symmetrical.

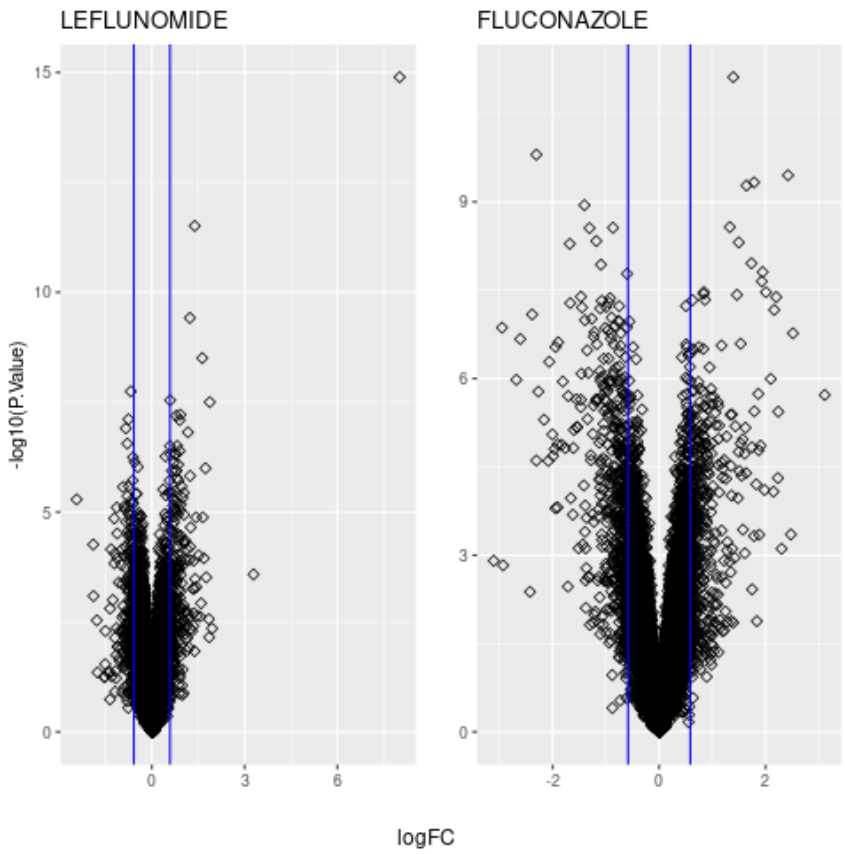
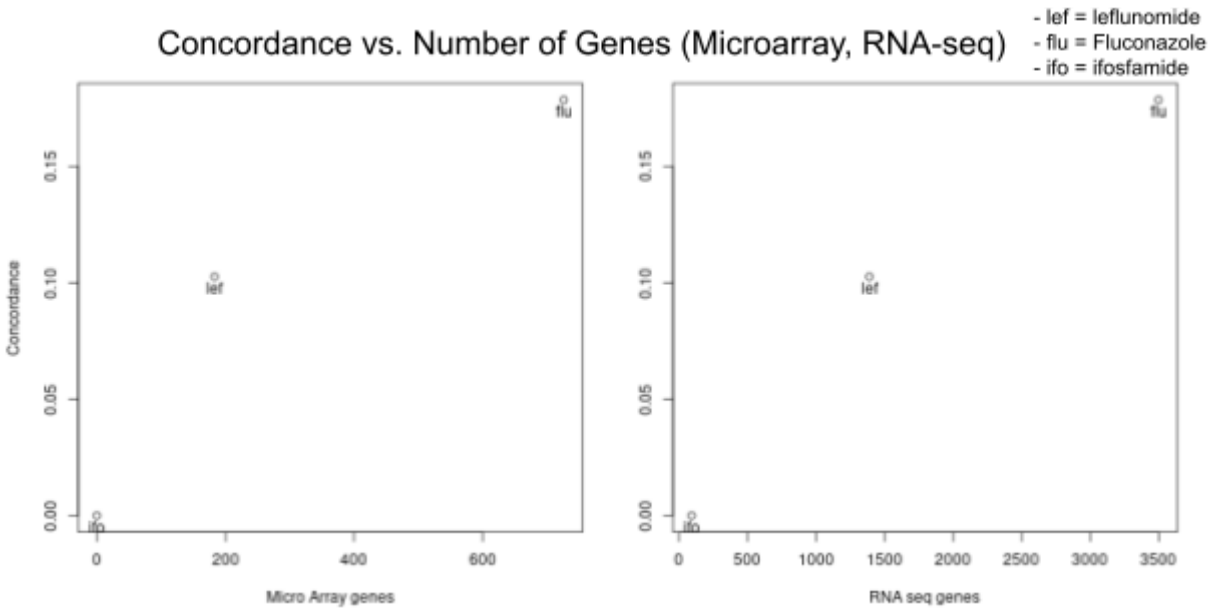


Figure 11 The volcano plots below show the same as the histograms above. I have added all genes in this plot with those that are significant on the outside of the |logFC| > log2(1.5) cutoff. Here we can see that the differentially expressed genes for Leflunomide treatment is lower than Fluconazole. Leflunomide has quite a few up regulated genes that have much higher logFC than the norm, while fluconazole seems to have similar max logFC for both up and down regulated genes.

Figure 12 ***Concordance Between Microarray and RNA-Seq Genes***



Above are the scatter plots of overall concordance vs. the amount of genes in microarray and RNA-Seq respectively. I got 0.1026765 ~ 10.2% concordance for Leflunomide (AhR), and a 0.1786035 ~ 17.8% concordance for Fluconazole (CAR/PXR). I used N= 13,079 total genes, this common set was based on the AceView gene model where they picked total genes that could be sampled by both platforms. (Wang et al., 2014).

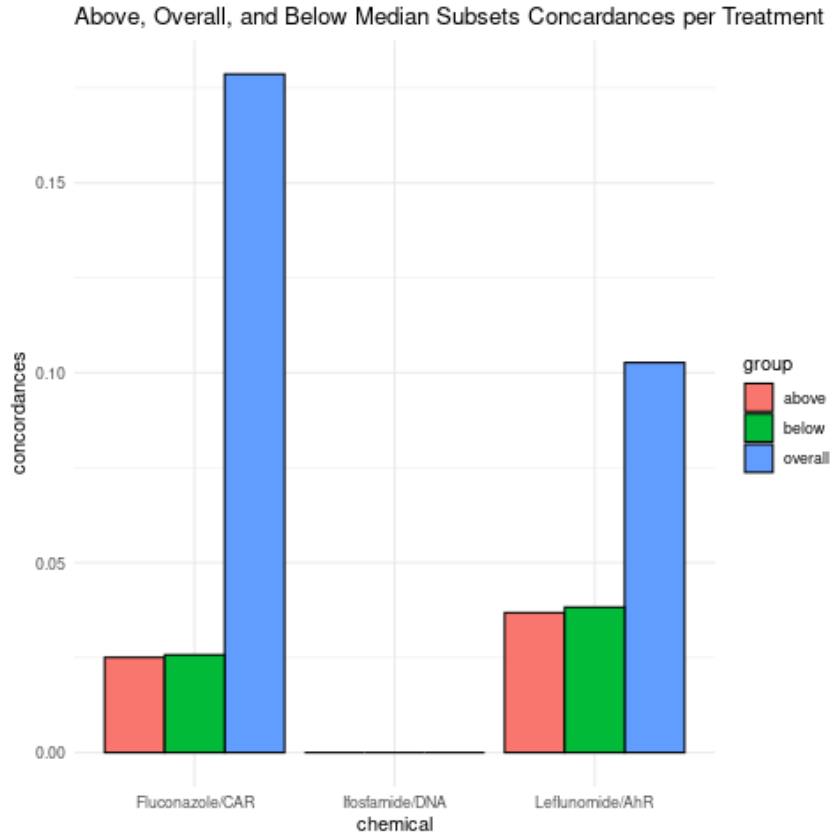


Figure 13
On the left I plotted the concordances of each subset above median, below median, and overall expression for each treatment. The concordances were:

Fluconazole:
Above- 2.51%
Overall-17.8%
Below-2.57%

Leflunomide: Above-3.68%,
Overall-10.2%,
Below- 3.83%

Ifosfamide:
0-ALL

Gene Set Enrichment Analysis

For each type of data source (RNAseq and microarray) there are three different sets of results from DAVID; each corresponding to one of the three MOAs (AhR, CAR and DNA). For each MOA result we also perused both KEGG_pathway and REACTOME_pathway.

Table 3A shows the Top five enriched pathways for the *3499* significant DE genes from the RNAseq CAR MOA. *Table 3A1* represents RNAseq CAR pathways identified from KEGG_pathway and *Table 3A2* is for RNAseq CAR pathways from REACTOME_pathway.

Table 3B shows the Top five enriched pathways for the 1997 significant DE genes from the microarray Fluconazole/CAR MOA data. *Table 3B1* corresponds to KEGG_pathway and *Table 3B2* shows pathways from REACTOME_pathway

Again, Table 4A shows the set of enriched pathways obtained from the 1389 significant RNAseq DEGs for the AhR MOA, and *Table 4B* shows the enriched pathways obtained from the 446 significant microarray DE genes for Leflunomide/Ahr MOA.

Finally, *Table 5A* shows the set of enriched pathways from the 91 significant RNAseq DEGs for the DNA MOA. There were no pathways obtained from the Ifosfamide/DNA microarray DEGs because none of these DEGs were significant based on the padj < 0.05 filter.

To sum it up, we show the enriched pathways that overlap between both the two platforms (RNAseq and microarray).

Of note, enriched pathways that matches those reported by Wang *et al* (2014) are highlighted in yellow or those that match across platforms have their cell borders marked red.

CAR:

*Table 3A1 : **RNA-seq** - CAR (Top 5) list of enriched pathways from KEGG_pathway (online DAVID tool). Although **xenobiotics metabolism** is not in the top ten, it is the **#13**.*

Category	Term	Count	%	P-Value
KEGG	Metabolic pathways	382	10.9	6.90E-20
KEGG	Complement and coagulation cascades	45	1.3	2.60E-14
KEGG	Ribosome biogenesis in eukaryotes	42	1.2	1.70E-08
KEGG	RNA transport	64	1.8	3.60E-08
KEGG	Retinol metabolism	38	1.1	3.40E-07

*Table 3A2: **RNA-seq** - CAR (Top 5) list of enriched pathways from REACTOME_pathway (online DAVID tool)*

Category	Term	Name	Count	%	P-Value
REACTOME	R-RNO-382556	Transport of small molecules	38	1.1	3.80E-07

REACTOME	R-RNO-72649	Metabolism of proteins	34	1	2.40E-05
REACTOME	R-RNO-5687128	Signal Transduction (MAPK Family signaling)	32	0.9	2.90E-05
REACTOME	R-RNO-4641257	Signal Transduction	25	0.7	3.40E-05
REACTOME	R-RNO-72702	Metabolism of proteins	34	1	4.30E-05

*Table 3B1: **Microarray** - CAR (Top 5) list of enriched pathways from KEGG_pathway (online DAVID tool)*

Category	Term	Count	%	P-Value
KEGG	Ribosome biogenesis in eukaryotes	26	1.9	1.80E-09
KEGG	Metabolic pathways	136	10.1	2.40E-06
KEGG	RNA transport	29	2.1	1.80E-05
KEGG	Retinol metabolism	18	1.3	8.20E-05
KEGG	Ribosome	27	2	2.80E-04

*Table 3B2: **Microarray** - CAR (Top 5) list of enriched pathways from REACTOME_pathway (online DAVID tool). **Xenobiotic metabolism** is #8*

Category	Term	Names	Count	%	P-Value
REACTOME	R-RNO-72649	Metabolism of proteins: Translation initiation complex formation	22	1.6	1.80E-07
REACTOME	R-RNO-72702	Metabolism of proteins: Ribosomal scanning and start codon recognition	22	1.6	2.80E-07
REACTOME	R-RNO-156827	Metabolism of proteins: L13a-mediated translational silencing of Ceruloplasmin expression	31	2.3	2.40E-06
REACTOME	R-RNO-72706	Metabolism of proteins: GTP hydrolysis and joining of the 60S ribosomal subunit	31	2.3	3.10E-06
REACTOME	R-RNO-72695	Metabolism of proteins: Formation of the ternary complex, and subsequently, the 43S complex	18	1.3	1.50E-05

AhR:

*Table 4A1 : **RNA-seq** - AhR List (Top 5) of enriched pathways from KEGG_pathway (online DAVID). Although **Xenobiotic metabolism** does not make the top five, it is #18*

Category	**Term**	**Count**	**%**	**P-Value**
KEGG	Metabolic pathways	149	10.7	2.60E-07
KEGG	Steroid hormone biosynthesis	19	1.4	4.30E-05
KEGG	Cell cycle	25	1.8	4.30E-05
KEGG	Arginine and proline metabolism	14	1	1.10E-04
KEGG	Tryptophan metabolism	13	0.9	1.60E-04

*Table 4A2 : **RNA-seq** AhR List (Top 5) of enriched pathways from REACTOME_pathway (online DAVID tool)*

Category	**Term**	**Exact term**	**Count**	**%**	**P-Value**
REACTOME	R-RNO-211981	Xenobiotics Metabolism	10	0.7	2.70E-04

REACTOME	R-RNO-69273	cell cycle: Cyclin A/B1/B2 during G2/M transition	8	0.6	3.50E-04
REACTOME	R-RNO-2408508	Metabolism (Metabolism of ingested SeMet, Sec, MeSec into H2Se)	5	0.4	2.40E-03
REACTOME	R-RNO-1614635	Sulfur Amino Acid metabolism	5	0.4	4.10E-03
REACTOME	R-RNO-433137	Sodium-coupled sulphate, di- and tri-carboxylate transporters	4	0.3	4.90E-03

*Table 4B1 : **Microarray -** AhR List (Top 5) of enriched pathways from KEGG_pathway (online DAVID tool)*

Category	Term	RT	Count	%	P-Value
KEGG	Tryptophan metabolism	RT	5	1.4	1.60E-02
KEGG	Rheumatoid arthritis	RT	6	1.7	4.20E-02
KEGG	Chemical carcinogenesis	RT	6	1.7	4.30E-02
KEGG	Cytokine-cytokine receptor interaction	RT	9	2.6	6.00E-02
KEGG	Metabolism of xenobiotics by cytochrome P450	RT	5	1.4	6.00E-02

*Table 4B2 : **Microarray -** AhR List (Top 5) of enriched pathways from REACTOME_pathway (online DAVID tool)*

Category	Term	Names	Count	%	P-Value
REACTOME	R-RNO-2142816	Metabolism:Synthesis of (16-20)-hydroxyeicosatetraenoic acids (HETE)	4	1.1	5.40E-03
REACTOME	R-RNO-2142670	Metabolism:Synthesis of epoxy (EET) and hydroxyeicosatetraenoic acids (DHET)	3	0.9	3.10E-02
REACTOME	R-RNO-2559585	Cellular response to external stimuli: Cellular responses to external stimuli	3	0.9	3.50E-02
REACTOME	R-RNO-1296025	Neuronal System: ATP sensitive Potassium channels	2	0.6	8.40E-02
REACTOME	R-RNO-499943	Interconversion of nucleotide di- and triphosphates	3	0.9	9.40E-02

DNA:

*Table 5A1 : **RNA-seq** - DNA List (Top 5) of enriched pathways from KEGG_pathway (online DAVID tool). Even though xenobiotics metabolism didn't make top five of this list, it is #8.*

Category	**Term**	**Gene Count**	**%**	**P-Value**
KEGG	Cell cycle	8	8.8	7.30E-06
KEGG	African trypanosomiasis	4	4.4	1.30E-03
KEGG	Malaria	4	4.4	4.70E-03
KEGG	DNA replication	3	3.3	1.80E-02
KEGG	Fanconi anemia pathway	3	3.3	3.60E-02

*Table 5A2 : **RNA-seq** DNA List (Top 5) of enriched pathways from REACTOME_pathway (online DAVID tool)*

Category	**Term**	**Name**	**Gene Count**	**%**	**P-Value**
REACTOME	R-RNO-1538133	Cell Cycle	4	4.4	7.40E-04

REACTOME	R-RNO-69052	Cell Cycle, DNA replication	3	3.3	7.60E-04
REACTOME	R-RNO-68949	Cell Cycle, DNA replication	5	5.5	1.60E-03
REACTOME	R-RNO-1247673	Transport of small Molecules (Bicarbonate transport)	3	3.3	4.40E-03
REACTOME	R-RNO-68867	DNA replication	3	3.3	5.10E-03

Clustered Heatmap of Normalized Gene Count

Figure 14 below shows the clustered heatmap produced from the normalized count matrices generated from the DESeq2 object. A rowside-color bar shows a successful clustering based on the three MOA of interest (AhR, CAR, and DNA) .

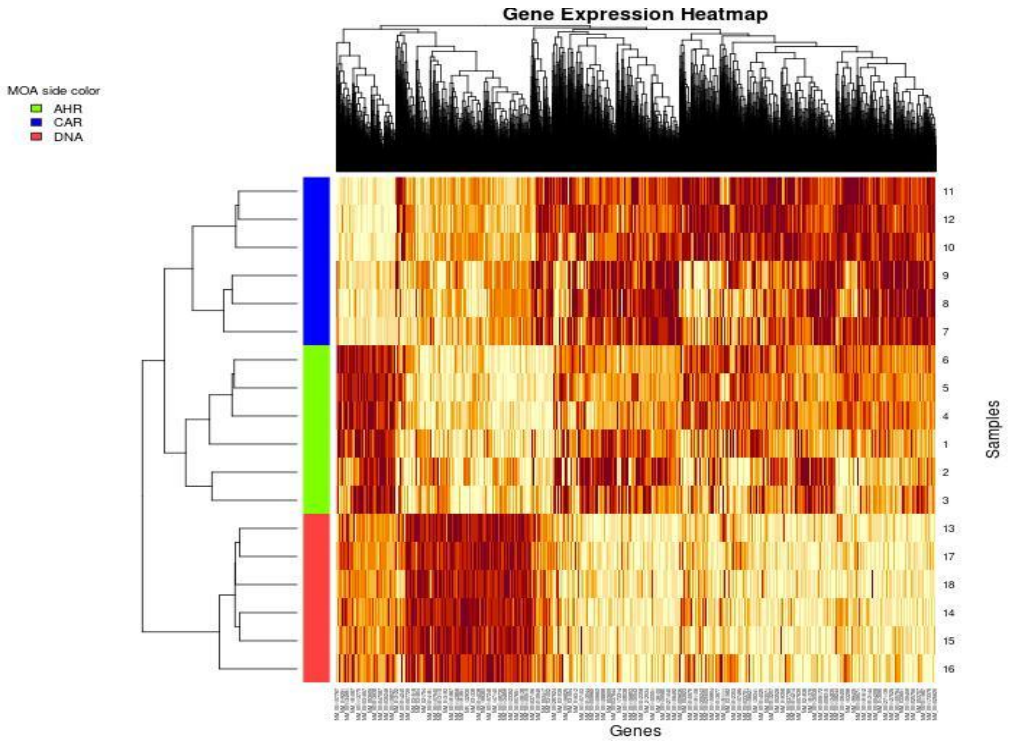


Figure 14: Clustered heatmap of normalized count based on MOA. (We have eighteen samples instead of fifteen because of , SRR1178050, SR1178061and SRR1178063 common to both AhR and CAR)

Discussion

Differentially Expressed Genes in RNA-Seq Data

The fifteen different treatment groups in the Wang et al. paper were categorized into one of five MOAs. Two of the MOAs we used in our analysis are associated with well-defined receptor-mediated processes - orphan nuclear hormone receptors (CAR/PXR) and aryl hydrocarbon receptor (AhR). The last treatment, DNA damage, is non-receptor-mediated. It was therefore expected that the AhR and CAR/PXR groups have significantly more differentially expressed genes in our analysis, as they were expected to induce a much stronger treatment effect, as was observed.

It is important to note that in the original Wang et al. paper, three different methods of identifying differentially expressed genes were tested: Limma, EdgeR and DESeq. Because all three were determined to be highly correlated, the majority of the differential gene expression analysis was done using Limma. While originally developed for use on microarray data, it has also been found to be equally applicable to RNA-Seq data. This is significant, as the RNA-Seq data for our analysis was processed using the DESeq2 package. While we do not suspect this accounts for a significant amount of variation between our results and those obtained by Wang et al., further analysis using the same packages used in the original paper may yield different results.

Additionally, Wang et al. used the Magic-pipeline, developed at NCBI, for the quality control, alignment, annotation, quantification, and normalization of the majority of their RNA-Seq data. This may have contributed to inherent differences in the generation of the raw count matrices that were used for differential gene expression analysis. However, because it is not specified which samples underwent which type of analysis in the original paper, we are unable to confirm, or to account for, these differences.

Concordance Between Microarray and RNA-Seq Genes

The main finding of the paper was that the concordance of RNA-Seq and microarray expression estimates depends on a number of factors, including biological effect size and gene expression level. Using the differential expression results from both DESeq2 and limma above, I measured concordance which was calculated as the similarity between the differentially expressed genes of both the microarray and RNA-seq experiments.

The concordance calculated in the paper and the one that I have are quite different. I used N as 13079 which was given in the paper as the common set of genes that were measured by both the RNA-seq as well as the microarray. I'm not quite sure why my results are not close to the paper, but I believe it could have something to with the computational analysis which resulted in a massive difference of differentially expressed genes in the microarray vs. RNA-seq analysis. It could also have been a user error where I used a different N as the paper did or incorrectly calculated the log2FC cutoff for the limma analysis. When I compared my results to the paper, the authors have plotted concordance of ifosfamide however I got 0 differentially expressed genes for it so concordance could not even be calculated. They also have not plotted fluconazole, yet that was the treatment I got the highest concordance for. Leflunomide was the only treatment I could compare and while the paper got around 60% concordance I got around 10%. Disregarding the actual numbers values if we look relatively, my results do match with what was observed in our RNA-seq as well as the paper. The Ahr and CAR/PXR were expected to have more differentially

expressed genes as they were expected to induce a much stronger treatment effect due to their well defined receptor mediated process, and that was also shown by my limma analysis with the higher amount of DEG's for those treatment groups. However based on my concordance analysis I cannot confirm what the authors found which was the three different methods of identifying differentially expressed genes: Limma, EdgeR and DESeq are highly correlated.

Gene Set Enrichment Analysis

The list of enriched pathways identified by our analysis based on three MOAs (two receptor-mediated MOAs [AhR, CAR] and one non-specific toxicity MOA [DNA-damage]) did not fully match the significant pathways reported by Wang *et al* (2014) in the Supplementary Table 4.

However, we matched xenobiotic metabolism signalling across RNA-seq and microarrays platforms for CAR and AhR MOA. Although our identified enriched pathways for DNA MOA fully matched the Supplementary Table 4, this does not significantly contribute to the objective of the project, since both the *cell cycle* and *xenobiotic metabolism* enriched pathway was obtained only from the RNAseq data; which does not allow for cross platform analysis or any form of intuition on the concordance between RNA-seq and microarrays. *Table 6A and 6B* below shows the MOAs, the enriched pathways that matched with those reported in the Supplementary Table 4 of Wang et al(2014), and the list of pathways we couldn't obtain in this analysis.

Although we couldn't reproduce the exact results reported in the study performed by Wang et al (2014), the observation from our analysis in terms of enriched pathways is that, both RNA-seq and microarrays provide the basis for DEG analysis, but the overlapping between the two platforms is lower in general, just as observed by Wang et al(2014). We cannot on the same level make the exact claims of Wang et al, 2014, since our analysis involved only three of the five MOAs used in the original study: one of the non-specific toxicity based MOA (which is DNA), and two of the receptor-mediated MOAs (Ahr and CAR).

Table 6A : A summary of pathways that matched with Wang et al (Supplementary Table 4)

MOA	RNA-Seq (Matching Wang et al)	Microarray (Matching Wang et al)	RNA-seq - Microarray Matches	RNA-seq-Microarray-Wang et al(2014) Matches
CAR	Xenobiotic metabolism	Xenobiotic Metabolism	Xenobiotic metabolism, Metabolic pathways	Xenobiotic Metabolism
AhR	Xenobiotic Metabolism	Xenobiotic metabolism	Xenobiotic metabolism	Xenobiotic metabolism
DNA	Cell cycle, Xenobiotic Metabolism	N/A (No gene passed padj < 0.05)		

Table 6B: Enriched pathways that were present in Wang et al (2014) but absent from our analysis

MOA	Matches with *Wang et al(2014)*	Pathways missed in our analysis
CAR	Xenobiotic metabolism	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

AHR	Xenobiotic metabolism	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
DNA	Cell cycle (Only RNA-seq)	
	Xenobiotic metabolism (Only RNA-seq)	

Additionally, our clustered analysis resulted in a clustered heatmap which is consistent with the MOA, which suggests that the chemical treatments or MOA has effects which are significant enough to be captured in the resulting DESeq2 counts.

Conclusion

Collective Conclusion

While we used the same data as used in the original Wang et al. study, we were unable to reproduce the results showing concordance between the microarray and RNA-Seq data results, across all three of the conditions that we examined. This is not to mean that these two methods do not produce comparable results, but rather that further analysis may be needed to determine the exact genes for which the microarray was able to detect and RNA-Seq was not, and vice versa, and how this may have contributed to the disparity noted between our results and those obtained in the original experiment.

Of note, our results represent only a small subset of the treatment groups ("tox groups") used in the original experiment. Therefore, additional testing in other groups may be required to detect a more noticeable trend, or to explain the results we observed.

Additionally, while no significant concordance was observed in our analysis, microarray and RNA-Seq data should, in effect, measure the same biological effects. This becomes increasingly important as current methodologies shift towards a preference for RNA-Seq over microarray for DEG detection. Thus, more comprehensive studies, evaluating multiple bioinformatics pipelines may be required in the future, not only when evaluating the concordance between microarray and RNA-Seq data, but also in reproducing and validating previously conducted experiments.

Data Curator

For the data curator role, I chose to deviate somewhat from the written project instructions to instead follow the advice given during the lecture about making a reproducible virtual environment. This involved using anaconda and an install_packages.sh script to set up the environment for the steps involved in read alignment and quality control. Additionally, I decided to use Snakemake to run the tools required for the data processing steps, which presented a steep learning curve. Though this undoubtedly made the role more challenging, it was a great learning experience as the number of processing steps and files were both low. As far as using STAR, FastQC, and MultiQC themselves, I had no trouble running the tools and producing the required outputs.

Programmer

For this project, the programmer role was fairly straight forward, and I encountered no significant challenges or problems during the process. I did notice, as mentioned above, that Wang et al. conducted most of their RNA-Seq data analysis using Limma rather than DESeq2. This meant that the specific parameters to be input into DESeq2, such as the method used for effect size shrinkage, weren't able to be matched exactly, and could have therefore led to differing results. For example, in our analysis, we used the "normal" Log2FC-shrinkage method when evaluating our DEG results, which was the default setting in DESeq2. However, newer methods have become available in recent years, such as the "apeglm" method, amongst others. Because Wang et al. used Limma, and did not specify the parameters used when testing each of the different bioinformatics pipelines against each other, we have no way to distinguish which method should be used in our own analysis - and whether this may have influenced the final results.

Analyst

The analyst role was quite complicated in terms of reproduction of the paper results. Running the Limma analysis was quite straightforward however when it came to computing the concordance, I ran into a lot of trouble. My results seemed to not align with the paper and I also had quite a bit of confusion on if I was doing the correct calculations. I tried multiple different calculations, and it seems that my issue is I got very few DEGs with my original microarray analysis for my treatment groups. Even if my results were off, I understand the importance of the original paper where they found high correlation among the limma, deseq2 analysis even though I could not confirm it.

Biologist

In this study, we attempt to investigate treatment-and-transcript dependent concordance between RNA-seq and microarrays. While we were not able to reproduce the original study, we did observe some treatment dependent influences which were captured from the successful MOA-based clustering of the DESeq2 normalized count matrix. Pathway enrichment analysis was undertaken using the online DAVID method; from which we identified at least one key pathway associated with xenobiotic activity in the liver across both RNA-seq and microarrays platforms.

References

Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 2009;37(1):1-13. doi:10.1093/nar/gkn923

Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 2009;4(1):44-57. doi:10.1038/nprot.2008.211

Łabaj PP, Leparc GG, Linggi BE, Markillie LM, Wiley HS, Kreil DP. Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics*. 2011;27(13):i383-i391. doi:10.1093/bioinformatics/btr247

McIntyre LM, Lopiano KK, Morse AM, et al. RNA-seq: technical variability and sampling. *BMC Genomics*. 2011;12:293. Published 2011 Jun 6. doi:10.1186/1471-2164-12-293

Mooney M, Bond J, Monks N, et al. Comparative RNA-Seq and microarray analysis of gene expression changes in B-cell lymphomas of Canis familiaris. *PLoS One*. 2013;8(4):e61088. Published 2013 Apr 4. doi:10.1371/journal.pone.0061088

RCoreTeam. R:A language and environment for statistical computing.Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/. Published 2013.

Smyth, G. K et. al. Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, Published 2005. pages 397–420.

Sultan M, Schulz MH, Richard H, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*. 2008;321(5891):956-960. doi:10.1126/science.1160342

Wang C, Gong B, Bushel PR, et al. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat Biotechnol.* 2014;32(9):926-932. doi:10.1038/nbt.3001

Wickham, H. dplyr: A Grammar of Data Manipulation. R package version 1.0.5. Published 2021. https://CRAN.R-project.org/package=dplyr Published 2021

Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Published 2016. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

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

Data Availability

All data for this project is publicly available at github.com/BF528/project-3-lava-lamp

Supplementary Data

The data in this section was generated, as specified in the project outline. However, because we did not feel it necessary to be commented on in the body of the report, it has been added as a supplement. A brief note is included for each figure or table, detailing why it was placed in the supplement rather than in the report itself.

Supplementary Table 1

General statistics of MultiQC alignment on featureCounts count matrices generated. These statistics were generated as part of the summary files of the featureCounts algorithm, and aggregated into a single report via MultiQC. From this table, it does not appear that there is significant deviation in the read alignment for any single sample, nor does any sample warrant removal from the analysis due to poor alignment.

This was included in the Supplementary Data, as the data represented is better depicted in Figures 5(a) and (b). Additionally, an HTML report with these figures, as well as this table, can be found in the GitHub repository for this project.

Sample Name	% Assigned	M Assigned
SRR1177981	61.5%	15.9
SRR1177982	61.3%	19.8
SRR1177983	61.1%	17.6
SRR1178008	64.9%	19.2
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

Supplementary Table 2

Top ten differentially expressed genes for (a) AhR, (b) CAR/PXR and (c) DNA Damage, respectively, ordered by increasing adjusted p-value.

These tables were included in the Supplementary Data because while the differentially expressed genes themselves are relevant when observing the effects of the particular chemical treatment, we are more interested in the concordance between the RNA-Seq and microarray data. Therefore, the top three differentially expressed genes for each MOA group were included in the body of the report to give readers an overview of the significance of the genes (via adjusted p-value and Log2FC). The remainder of the DESeq2 results are included below.

(a) DESeq2 Results Output - AhR MOA Treatment Group

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
NM_001007722	4147.89283311862	-4.71810604478048	0.375303674991664	-14.4686567368656	1.91169121334766e-47	6.79287611142868e-44
NM_001012174	2079.22833780418	2.15742300128457	0.189405064540447	11.3880821998601	4.79425556173557e-30	5.11067642881011e-27
NM_001130558	682.784007022878	-4.77594855706857	0.349161094531459	-12.8916688691639	5.01467541818446e-38	1.06912879915693e-34
NM_001257095	6959.56697081144	-3.89538665925631	0.28090143130591	-13.8880015647104	7.48959421519402e-44	1.99597685834921e-40
NM_012540	67262.752512119	2.50985429052831	0.391016056446114	12.0554119136719	1.81622216422583e-33	3.22682137844122e-30
NM_012541	208144.889528179	3.55661714348694	0.305834697525241	11.6907041103645	1.42202756026918e-31	1.89485172405868e-28
NM_013096	10722.3155824696	-5.01103316518716	0.37353815298518	-16.2993816729119	9.96955853817444e-60	1.0627549401694e-55
NM_033234	8523.65844555515	-4.92037032334757	0.37515609960076	-15.9900760798441	1.49847896045164e-57	7.98689285920724e-54
NM_130407	2961.00613160926	3.37268500008758	0.295276539009525	11.4461993292213	2.4568300247023e-30	2.90997867370295e-27
NM_198776	1815.0002383578	-4.08903221057932	0.377768443370816	-11.9729310888382	4.9257212687635e-33	7.50116981785985e-30

(b) DESeq2 Results Output - CAR/PXR Treatment Group

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
NM_001130558	840.763745421887	-5.76775475610783	0.262681046264808	-20.2666658139224	2.53243944371547e-91	1.35282915083281e-87
NM_001134844	52637.0745180869	5.78744880535407	0.296965250121064	19.6379140504152	7.33497489621203e-86	2.61222905970431e-82
NM_013033	848.44218590862	4.55216854516338	0.310269834588121	14.5808921214068	3.71628354855532e-48	7.940954686553e-45
NM_013098	15275.732039444	-3.58341063699305	0.26169225885102	-13.7038508727464	9.62795213480514e-43	1.14294489564731e-39
NM_017006	3785.10784528665	2.76393715622452	0.203178777486533	13.5981095972367	4.10917641451605e-42	4.39024408126895e-39
NM_024127	1687.34908786326	2.44684098801871	0.168564471425686	14.5065397429787	1.10138008205507e-47	1.68102068523949e-44
NM_031048	5919.01891423894	3.62520841012141	0.259476690090546	13.9774866096089	2.13909801941453e-44	2.8567654049281e-41
NM_053288	156111.387567242	4.54962723962857	0.181538795669266	25.0631872684266	1.25393633509966e-138	1.33970558042048e-134
NM_053699	367.141712271886	4.28482982108025	0.289801059463596	14.5090876575419	1.0612265380679e-47	1.68102068523949e-44
NM_080581	6232.17217995792	4.27279865698561	0.274083633498812	15.5991824576843	7.37322652306242e-55	1.96938880430997e-51

(c) DESeq2 Results Output - DNA Damage MOA Treatment Group

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
NM_001007722	1699.57899009657	-1.7663002728252	0.153728141817449	-16.4418003863722	9.60234085535284e-61	5.03306695933319e-57
NM_001013057	793.57943683368	-0.840211636329588	0.148509087235793	-6.16848397932181	6.89478341343649e-10	9.03475181538184e-07
NM_001107084	632.852426913837	-1.0093195609329	0.155895599358422	-6.74519566223506	1.52820941662706e-11	2.2886027592145e-08
NM_001271152	2172.98457217621	-0.901688245752628	0.153662060736836	-5.89856460860981	3.66677386685025e-09	3.84387904461912e-06
NM_012623	575.359035065484	1.15182730683128	0.150249362580659	8.95689470044982	3.3395035133278e-19	7.00160306604306e-16
NM_013096	4753.94856675902	-0.542495892795011	0.115371631069839	-10.467872365025	1.21339546846617e-25	3.18000617398271e-22
NM_033234	3409.00007193038	-1.62820168749264	0.152032002068913	-16.5935096378219	7.7647348994499e-62	8.13977159509333e-58
NM_053962	3593.16896741911	0.930216006824764	0.156621679257478	6.00647203806597	1.89603657094991e-09	2.20846126369644e-06
NM_198776	796.482045865268	-1.44674413921364	0.151468804359947	-12.3805688913475	3.329741618343e-35	1.16352271283632e-31
NM_199113	81.4697655893032	-1.31432344701664	0.156317298532937	-8.26445307025121	1.40336657307369e-16	2.45191529758859e-13

Supplementary Figure 1

Rescaled box plot of count distributions, following featureCounts summarization, both with and without removal of zero-count genes. The y-axis has been rescaled to log10, such that the distribution of counts can be better visualized. As per the previously generated boxplots, the counts across all samples appear to be fairly evenly distributed, with similar medians, interquartile ranges, and numbers of outliers.

The removal of the zero-count genes, while shifting the count distributions moderately upwards, does not appear to have had a significant impact on the distributions, other than the upward shifting of the interquartile range. Because all samples appear to be concordance, it does not seem necessary that any of the samples be removed or otherwise adjusted prior to downstream analysis.

These figures were included in the Supplementary Data because two plots have already been dedicated to representing this data. Those plots appear to be more informative in representing the change in the counts matrix when removing the zeros from the counts, in that the interquartile range appeared to shift more drastically with the removal of the zero-count genes. However, the scaled axis below, as suggested by our TA, allows for a better comparison of count distribution across samples.