Concordance of Microarray and RNA-Seq Differential Gene Expression

Husain A.*, Ramaiah R.*, Choy A.*, Kafrawi R.*

INTRODUCTION

Prior to 2014, a variety of conclusions had been drawn regarding several key metrics produced by microarrays and RNA-seq, including reproducibility, expression levels, and gene isoform identification (Wang *et al*. 2014). A study conducted by Mortazavi *et al*. found that RNA-seq studies are less precise when it comes to detecting low expression transcripts (2008). This is in stark contrast to findings by later studies, which would suggest otherwise. In another study, the findings of Mooney *et al.* (2013) found that RNA-seq was more sensitive for low expression transcript detection when comparing the RNA-seq and microarray datasets generated from B-Cell Lymphomas of *Canis familiaris*.

Wang *et al.* suggests that the lack of parity in the established literature regarding the efficacy of RNA-seq results when compared to that of microarray datasets is largely a consequence of insufficient treatment conditions (2014). In an attempt to rectify this scientific lacuna, Wang *et al.* (2014) designed a study that would generate RNA-seq and microarray data from a set of rat liver samples that were subjected to 27 unique chemicals representative of several modes of action (MOA) – i.e. cellular mechanisms. In this paper, we hope to replicate some of the results and conclusions drawn by Wang *et al.* (2014) with a small subset of 3 chemicals.

DATA

Microarray and RNA-Seq data were produced and processed for the purpose of this project. This was produced using the Affymetrix microarray and Illumina RNA-Seq respectively. The data was downloaded from SRA and GEO using the accession numbers- SRP039021, GSE55347 and GSE47875. The microarray data was provided fully processed but the RNA-Seq data required processing. We used samples from tox-group 3 for performing analyses and recreating the figures from this paper. Tox-group 3 consisted of 9 samples- 3 treatments (Leflunomide, Fluconazole and Ifosfamide) with three replicates each. The data was aligned to the mouse genome using STAR (v2.6.0) and reports were generated using fastqc (v0.11.7) and multiqc (v1.10.1). These results are summarized below in Table 1 and 2 respectively.

❖ Bioinformatics Program, Boston University, Boston, MA

Sample Name	Mode of action	Chemical	% Duplicates	Average Sequence Length (bp)	Total Sequences (millions)
SRR1177981_1	DNA Damage	Ifosfamide	56.1	101	14.2
SRR1177981_2	DNA Damage	Ifosfamide	48.4	101	14.2
SRR1177982_1	DNA Damage	Ifosfamide	58.5	101	17.2
SRR1177982_2	DNA Damage	Ifosfamide	54.3	101	17.2
SRR1177983_1	DNA Damage	Ifosfamide	57.6	101	16.2
SRR1177983_2	DNA Damage	Ifosfamide	50.0	101	16.2
SRR1178008_1	AhR	Leflunomide	56.2	101	15.2
SRR1178008_2	AhR	Leflunomide	54.4	101	15.2
SRR1178009_1	AhR	Leflunomide	62.2	101	18.1
SRR1178009_2	AhR	Leflunomide	60.1	101	18.1
SRR1178010_1	AhR	Leflunomide	62.5	101	18.6
SRR1178010_2	AhR	Leflunomide	61.4	101	18.6
SRR1178014_1	CAR/PXR	Fluconazole	53.9	50	17.5
SRR1178014_2	CAR/PXR	Fluconazole	51.9	50	17.5
SRR1178021_1	CAR/PXR	Fluconazole	48.7	100	17.5
SRR1178021_2	CAR/PXR	Fluconazole	46.4	100	17.5
SRR1178047_1	CAR/PXR	Fluconazole	48.5	100	17.1
SRR1178047_2	CAR/PXR	Fluconazole	47.3	100	17.1

Table 1. Summarized report of read quality.

Sample Name	Mode of Action	Chemical	% Aligned	Uniquely Mapped Reads (millions)
SRR1177981	DNA Damage	Ifosfamide	80.2%	11.4
SRR1177982	DNA Damage	Ifosfamide	83.8%	14.4
SRR1177983	DNA Damage	Ifosfamide	79.6%	12.9
SRR1178008	AhR	Leflunomide	88.1%	13.3
SRR1178009	AhR	Leflunomide	90.1%	16.3
SRR1178010	AhR	Leflunomide	90.6%	16.8
SRR1178014	CAR/PXR	Fluconazole	83.5%	14.6
SRR1178021	CAR/PXR	Fluconazole	82.0%	14.3
SRR1178047	CAR/PXR	Fluconazole	84.0%	14.4

Table 2. Summarized results of mapping reads to mouse genome.

The summarized fastqc results as obtained from multiqc indicate mixed results about the data quality. It was observed in the sequence quality histograms (**Figure 1**) that the majority (15) of the samples failed and only three (SRR1178008_2, SRR1178014_1 and SRR1178014_2) had valid quality scores. This indicates that there could have been issues during sequencing but there are no extreme drops or unusual patterns. The samples followed the general trend that is expected with Illumina sequencing. Furthermore, it was also observed that the per sequence quality scores (**Figure 2**) for all samples were good with no unexpected bumps in the lower phred scores. The per sequence GC content (**Figure 3**) also roughly matched the expected % GC (~52.4%) for *Mus musculus* (Romiguier, 2010). The STAR alignment performed using the mouse genome resulted in good quality alignments (**Figure 4**) with all alignments being over 75% uniquely mapped.

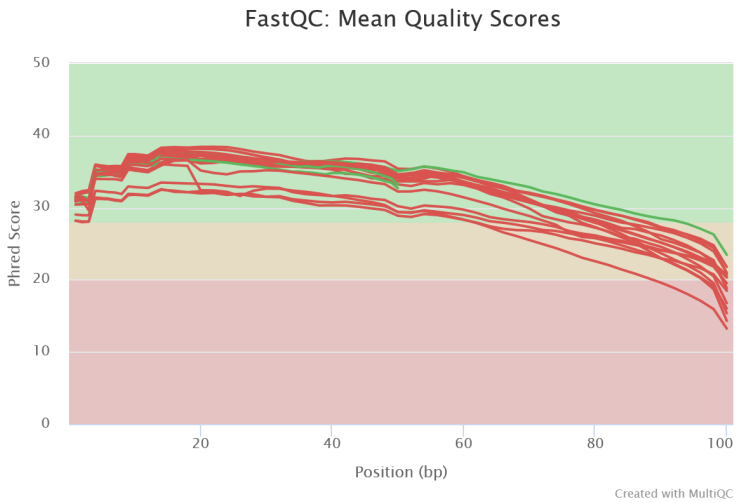


Figure 1. Plot showing mean quality scores across each base position in the read for the eighteen samples. Red indicates samples that failed and green lines indicate the samples that had good scores.

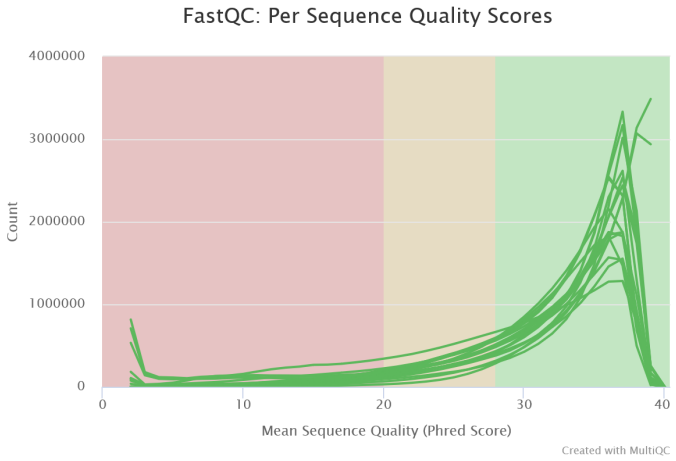


Figure 2. Line plot of number of number of reads vs the average quality score. Green indicates samples of good quality.

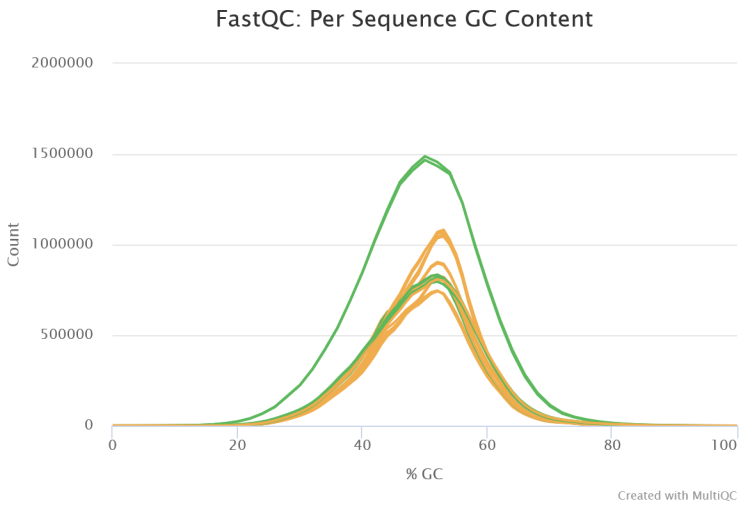


Figure 3. Plot showing the per sequence GC content in percentage across the whole length of each sequence in a file.

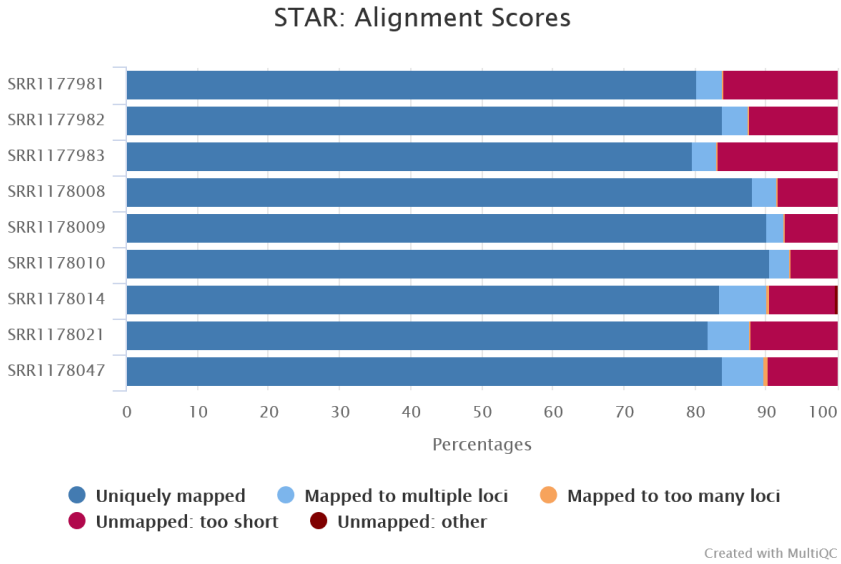


Figure 4. STAR Alignment results of samples shown in percentage across the samples.

METHODS

FeatureCounts was run on each of the aligned and sorted RNAseq bam files, using rn4_refGene_20180308.gtf as the annotated reference (Liao, 2013; Liao, 2014). Then multiqc reports were generated on each of the resulting count files (Ewels, 2016). In MultiQC, all samples are subjected to quality control at once, and the results are compiled into a single report to help identify any global trends or biases. Feature count information was combined into a single CSV file from which boxplots of count distributions were created using R packages tidyverse and janitor.

Differential expression analysis was run on each experimental group (AhR, CAR/PXR, DNA Damage) using DESeq2. Relevant controls that matched with "vehicle" were selected for the experimental groups prior to DESeq2 analysis. DESeq2 performs differential expression analysis on RNAseq count data. Normalized counts were calculated and differential expressed genes were determined with an adjusted p-value below 0.05.

To determine the microarray differential gene expression (DEG) between treatment and controls samples in our tox-group, the Bioconductor (Morgan, 2021) package limma (Ritchie, 2015) was used along with the pre-normalized gene expression matrix provided by the authors that can also be found on SCC. The sample metadata detailing the array ids, chemical, vehicle, and route for our chosen tox-group were also provided for us on SCC. Using the sample ids for our tox-group, the RMA normalized matrix of all experiments were subset for each treatment group with controls that have the same vehicle for treatment. Thus, in tox-group 3, Leflunomide and Fluconazole were subset with controls with 100% corn oil, whereas Ifosfamide was subset with 100% saline in controls. For each gene, in the limma package, the function lmFit estimates the fold changes and standard errors to fit a linear model to the design matrix. Then empirical Bayes helps apply statistics to our linear model, and topTable returns the top gene statistics that were generated. By default, it returns top 10, but in this study, we returned all statistics. Significant differentially expressed genes were ranked and reported out by adjusted p-value < 0.05. To be consistent with authors of the original study, further analysis of DEGs for each treatment were performed using a cutoff of unadjusted p-value < 0.05 and absolute fold change of > 1.5. DEG analysis was visualized with histograms and volcano plots for each treatment group. Information regarding abbreviations for chemicals and their mode of action can be found in **Table 5**. R packages tidyverse (Wickham, 2019) and ggpubr (Kassambara, 2020) were used for all plots. All microarray analysis procedures were conducted on the SCC platform in R version 4.1.1 (R Core Team, 2013).

Concordance is the level of agreement between two entities. In this study, the concordance is determined between Microarray and RNA-Seq platforms by the number of overlapping DE genes and fold change directionality that result from their respective analysis for each treatment group, as described by Wang *et al.* (2014). Concordance can simply be calculated as follows:

$$\frac{2 \times intersect(DEGs_{Microarray}, DEGs_{RNA-Seq})}{DEGs_{Microarray} + DEGs_{RNA-Seq}} \qquad Eq.1$$

As there is a probability that a gene falls into the region of overlap by chance, this calculation needs to be adjusted to account for the randomness, and can be derived to the following:

$$n_x = \frac{Nn_0 - n_1 n_2}{n_0 + N - n_1 - n_2} \qquad Eq.2$$

Where n_x is the adjusted overlap of genes with randomness, N is the pool of total genes in the rat genome, n_0 is the observed unadjusted overlap, and n_1 and n_2 are the DEGs from each platform respectively. Putting these two equations together, the adjusted concordance is calculated as follows:

$$Concordance = \frac{2 \times n_x}{n_1 + n_2} \qquad Eq. 3$$

In order to determine the number of genes that overlap, a file, refseq_affy_map.csv, was provided for us on SCC. It consists of Affymetrix probe IDs that were previously mapped to their respective refSeq IDs. Due to incomplete mapping or mapping to multiple IDs of our DE gene results with this file, we took a brief look at the rat database (Carlson, 2021) from Bioconductor using AnnotateDbi (Pagès, 2021), but for reasons discussed later, we primarily used the ID map that was provided. For simplicity, the pools of n were estimated based on the unique genes the probes were mapped to since each probe confers to a specific gene.

As with microarray analysis, probes from each set were filtered for significance by p-value and absolute fold change of > 1.5 before mapping for the overall concordance for each treatment. However, concordance was also calculated for the above-median and below-median gene expression subset for both methodologies. These subsets were ranked by fold change before dividing by its median, which subsequently results in a subset of positive and negative fold changes for each treatment across both platforms. These three measures of concordance for each treatment were then visualized in a bar plot.

Gene set enrichment analysis was conducted on the output of the limma analysis using DAVID. In order to do so, we first had to filter out results with a p-value > 0.05, which yielded a list of 2151 gene probes. This list was input into DAVID's "Gene List" query. The AFFYMETRIX_3PRIME_IVT_ID identifier was selected, and the list type was set to "Gene List". For our particular dataset, not all of the IDs were mapped by DAVID for reasons that will be discussed later. Thus, we proceeded with the recommended option, which was to continue to submit the IDs that DAVID was able to actually map. This split the initial gene list into two distinct lists labeled "*Rattus norvegicus*" and "Unknown". We proceeded to conduct the enrichment analysis using the "*Rattus norvegicus*" list, which consisted of a total of 1609 genes.

A clustered heatmap of the normalized counts was created from the expression matrix. This was done using the pheatmap function. Row names were not included in the figure generated from the expression matrix due to the sheer number of expression values included in the input data (n > 10695). Additional arguments were used to refine the presentation of the expression values. Most importantly, the coloring scheme was set using the colorRampPalette argument, the cluster_rows argument was set to "TRUE" and the scale argument was set to "row."

RESULTS

After running featureCounts on the RNAseq data, proportion of reads assigned, unassigned, and multi-mapped was roughly consistent across the 9 experimental samples, with between 55.4% -69.2% of reads assigned, 6.8%-18.1% of reads unassigned due to multimapping, 18.7%-23.7% unassigned due to no features being identified, and 4.3%-6.1% of reads unassigned due to ambiguity (**Figure 5**). Distribution of counts varied across samples, with the cytotoxic sample SRR1177989 having the widest distribution of counts (**Figure 6**). Distribution of counts for the DNA_Damage and AhR experimental group were relatively uniform when compared to the distribution of counts for CAR/PXR group.

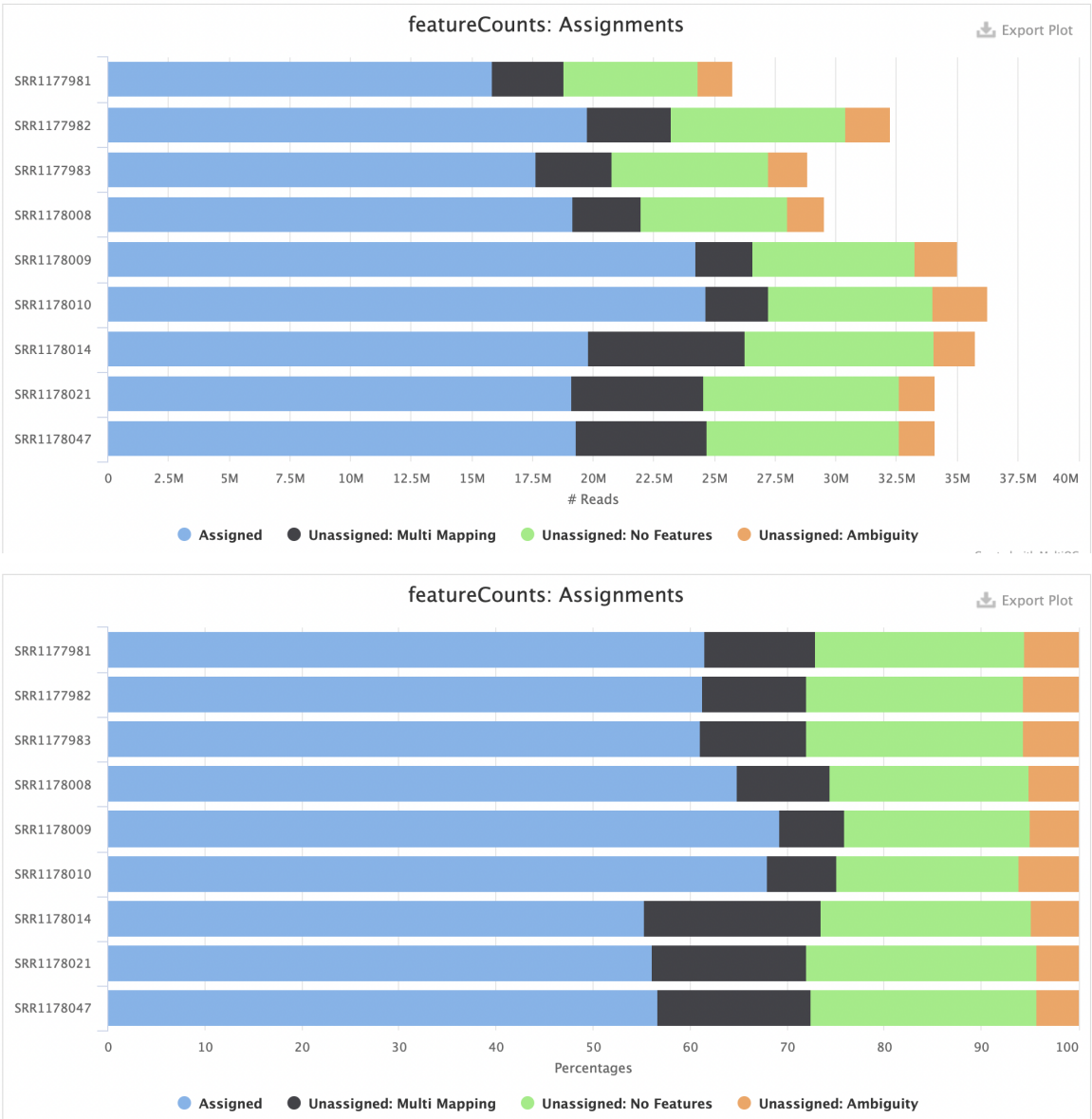


Figure 5. MultiQC results: A) Number of reads and B) Percentage of reads that were assigned, multi-mapped, ambiguous, and unassigned for each sample.

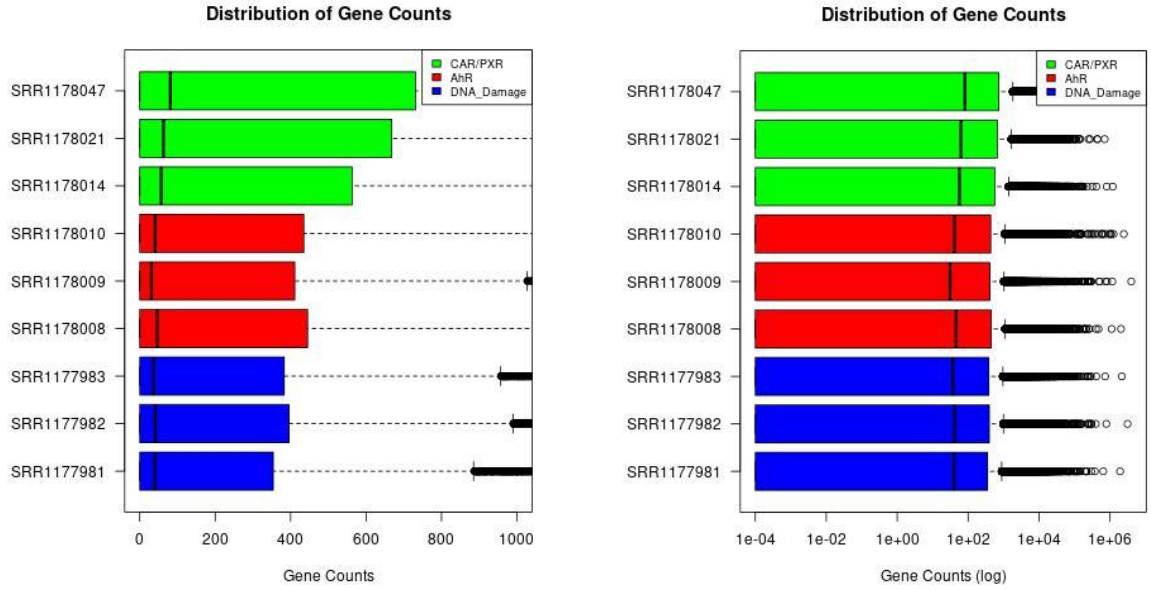


Figure 6. A) Read count distribution for each gene for each sample. B) Read count distribution for each gene for each sample in log scale.

AhR Geneid	AhR padj	CAR/PXR Geneid	CAR/PXR padj	DNA_Damage Geneid	DNA_Damage padj
NM_013096	1.06275535067091E-55	NM_053288	1.33970557922521E-134	NM_033234	8.14014648166088E-58
NM_033234	7.98689588122619E-54	NM_001130558	1.3528291553548E-87	NM_001007722	5.03327217737575E-57
NM_001007722	6.79287792629248E-44	NM_001134844	2.61222911112625E-82	NM_198776	1.16355138867511E-31
NM_001257095	1.99597758132921E-40	NM_080581	1.96938882518137E-51	NM_013096	3.18000617398426E-22
NM_001130558	1.06912923810329E-34	NM_013033	7.94095476378078E-45	NM_012623	7.00170485143715E-16
NM_012540	3.22682208679589E-30	NM_024127	1.68102067696134E-44	NM_199113	2.4518902116208E-13
NM_198776	7.5011713726223E-30	NM_053699	1.68102067696134E-44	NM_001107084	2.28861719083956E-08
NM_012541	1.89321761165946E-28	NM_031048	2.85676542628105E-41	NM_001013057	9.03481454916215E-07
NM_130407	2.90997935201269E-27	NM_013098	1.142944904318E-39	NM_053962	2.20847037792734E-06
NM_001012174	5.11067783734695E-27	NM_017006	4.3902440897451E-39	NM_001271152	3.84388822324998E-06

Table 3. Top 10 differentially expressed genes from each experimental group by adjusted p-value ($\alpha < 0.05$).

The top 10 differentially expressed genes, by adjusted p-value, were identified for each experimental group for RNAseq data (**Table 3**) and Microarray data (**Table 4**).

AhR					
	probeid	SYMBOL	logFC	P.Value	adj.P.Val
1	1370269_at	Cyp1a1	8.013842439	1.28E-15	3.97E-11
2	1387243_at	Cyp1a2	1.383694417	3.09E-12	4.80E-08
3	1372600_at	Fbxo31	1.226486222	3.84E-10	3.98E-06
4	1392946_at	Il1r1	1.627798917	3.14E-09	2.44E-05
5	1388611_at	Tcea3	-0.684972983	1.80E-08	0.000111724
6	1370244_at	Ctsl	0.589706542	2.86E-08	0.00014009
7	1373810_at	Pla2g12a	1.871151708	3.15E-08	0.00014009
8	1376827_at	Eml4	0.783970675	6.42E-08	0.000221687
9	1373814_at	R3hdm2	-0.761206761	7.75E-08	0.000229011
10	1378016_at	Eml4	0.897531783	8.10E-08	0.000229011
CAR/PXR					
	probeid	SYMBOL	logFC	P.Value	adj.P.Val
1	1368731_at	Orm1	1.392496383	7.62E-12	2.37E-07
2	1371076_at	Cyp2b1	2.42087715	3.53E-10	3.28E-06
3	1390255_at	Ablim3	1.782758115	4.67E-10	3.28E-06
4	1394022_at	Id4	-1.403790192	1.14E-09	5.89E-06
5	1380336_at	Irak3	1.327452197	2.68E-09	9.64E-06
6	1398597_at	Rnf144a	-1.307324715	2.79E-09	9.64E-06
7	1377192_a_at	Clpx	-1.180870625	4.62E-09	1.34E-05
8	1390801_at	Tmem252	-1.093168008	1.16E-08	2.59E-05
9	1367957_at	Rgs3	1.944340542	1.56E-08	3.22E-05
10	1387109_at	Por	1.928520075	2.23E-08	4.09E-05
DNA Damage					
	probeid	SYMBOL	logFC	P.Value	adj.P.Val
1	1374475_at	Abhd1	-0.347716802	3.09E-05	0.300665944
2	1368273_at	Mapk6	-0.319897142	4.30E-05	0.300665944
3	1371266_at	Afm	-0.259292317	5.52E-05	0.300665944
4	1373217_at	Ehbp1	-0.356506295	5.82E-05	0.300665944
5	1368718_at	Aldh1a7	0.9779115	6.54E-05	0.300665944
6	1376481_at	Adamts9	-0.527078327	6.77E-05	0.300665944
7	1383137_at	Sox4	-0.543869178	0.00011977	0.335540117
8	1390019_at	H3f3b	-0.341711733	0.000139808	0.335540117
9	1387391_at	Cdkn1a	-0.45584271	0.000146892	0.335540117
10	1383046_at	Cfh	0.234185417	0.000182375	0.37742293

Table 4. Top 10 differentially expressed genes from Microarray analysis

The number of differentially expressed genes for each experimental group with a p-adjusted value less than 0.05 from RNA-Seq analysis were 1424, 3510, and 303 for AhR (aryl hydrocarbon receptor, 3-methylcholanthrene, 3ME), CAR/PXR, and DNA_Damage experimental groups, respectively. Microarray analysis had much less differentially expressed genes at 466, 1997, and 0 respectively for AhR, CAR/PXR, and DNA Damage as described in **Table 5.**

Mode of Action (MOA)	Treatment	Abbreviation	RNA-Seq DEGs	Microarray DEGs
AhR	Leflunomide	LEF	1424	466
CAR/PXR	Fluconazole	FLU	3510	1997
DNA Damage	Ifosfamide	IFO	303	0

Table 5. Summary of the resulting DEGs found by each platform, RNA-Seq and Microarray, for each chemical treatment and its corresponding mode of action (MOA). DEGs were filtered based on adjusted p-value < 0.05.

For the significant differentially expressed genes (α<0.05), log2 fold change distribution do not vary much between AhR and CAR/PXR in RNA-Seq results (**Figure 7A, 7C, 7E**). Log2 fold change vs nominal p-value also do not vary much between AhR and CAR/PXR (**Figure 7B, 7D, 7F**). Likewise it did not vary significantly between AhR and CAR/PXR in Microarray analysis (**Figure 8A,B**). However, there appears to be significantly much less differentially expressed genes with Ifosfamide, or with DNA Damage as mode of action (**Figure 8C, 9C**). This may be reflected in the lack of DEG findings for the microarray platform when filtering for adjusted p-values as observed in **Table 5**.

Although also part of common practice, p-values, for this reason, instead of adjusted p-values, were used as a cutoff for significance for the histograms and volcano plots. As observed in **Figure 9**, the significant genes by only p-value < 0.05 does not meet the significance level of an absolute log fold change < 1.5 (red dots in **Figure 9C** and p-values from **Table 4**). This is only observed for DNA Damage; the other modes of action have agreeing p-values that also meet the fold change threshold. Overall, the total DEGs found in their respective platforms and cutoff values can be summarized in **Table 6**.

Mode of Action (MOA)	DE Genes	RNA-Seq	Microarray
AhR	Adj.P.Val <0.05	1424	466
	$\lvert\log_2 FC\rvert > 1.5$ + P.Val < 0.05	1385	288
CAR/PXR	Adj.P.Val <0.05	3510	1997
	$\lvert\log_2 FC\rvert > 1.5$ + P.Val < 0.05	2282	619
DNA Damage	Adj.P.Val <0.05	303	0
	$\lvert\log_2 FC\rvert > 1.5$ + P.Val < 0.05	119	46

Table 6. Summary of DEGs found in both platforms for different cutoff levels used.

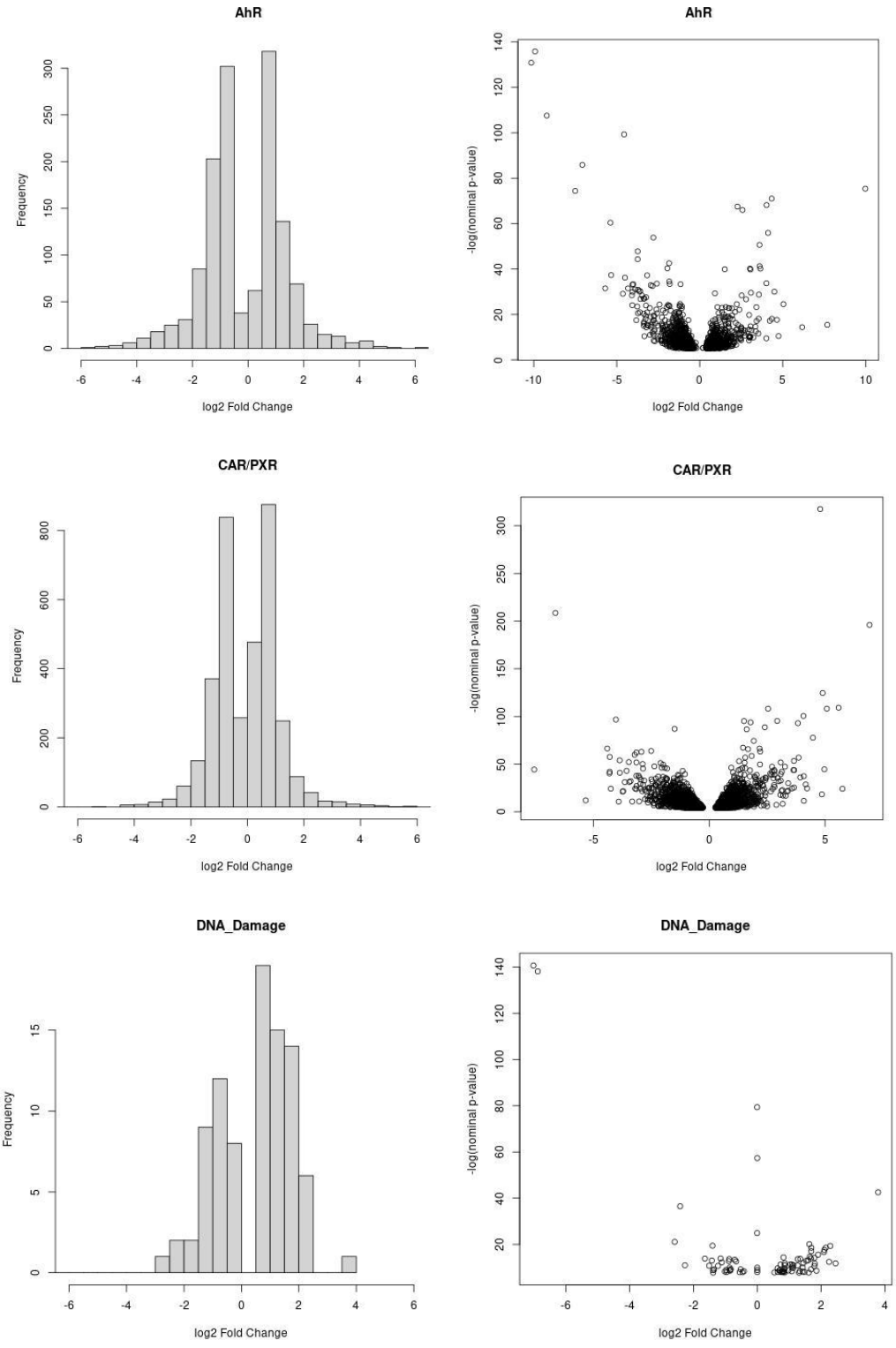


Figure 7. Log2 fold change and scatter plots for significant differentially expressed genes ($\alpha < 0.05$) for experimental groups AhR, CAR/PXR, and DNA_Damage respectively.

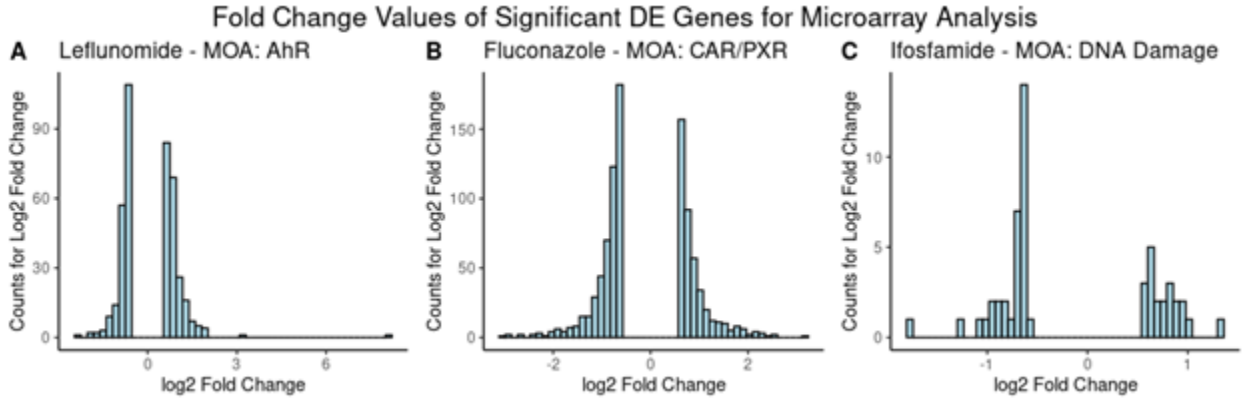


Figure 8. Histograms of the fold change values of significant DE genes for each Mode of Action (MOA) as determined by Microarray analysis with Limma. Counts of fold change values were filtered for significance at unadjusted p-value < 0.05 and |log2 Fold Change| > $\log_2(1.5)$ for Leflunomide-AhR (**A**), Fluconazole-CAR/PXR (**B**), and Ifosfamide-DNA Damage (**C**).

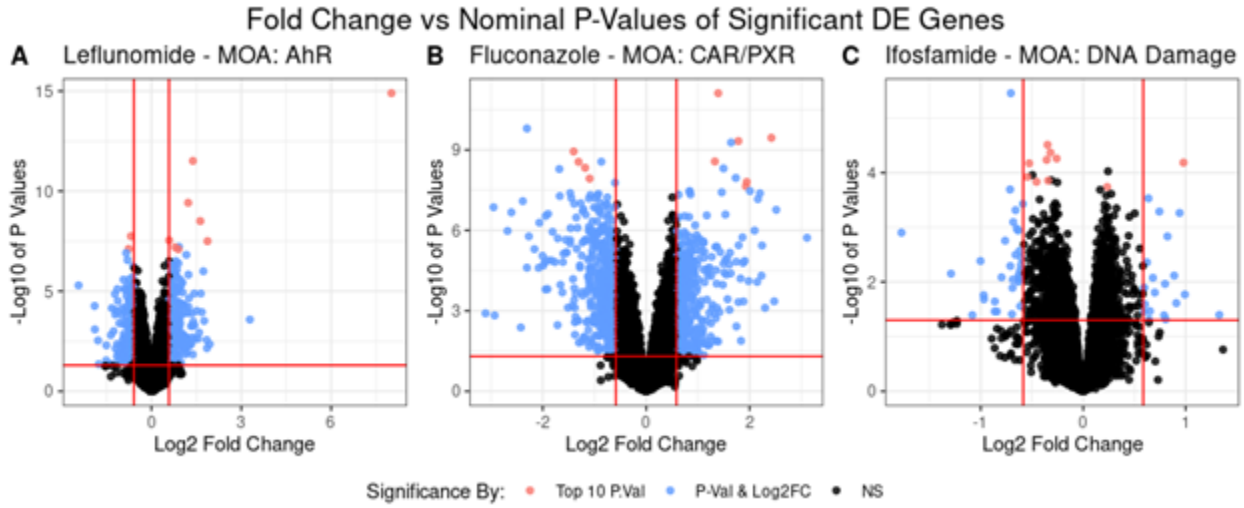


Figure 9. Scatter plot of fold changes vs -Log10 of the nominal p-values of significant DE genes for each Mode of Action (MOA) from Microarray analysis with Limma. Each panel refers to a different chemical treatment with mode of action of AhR (**A**), CAR/PXR (**B**), and DNA Damage (**C**). The level of significance is demarcated by the red lines in each panel: horizontal for unadjusted p-value threshold of 0.05, and vertical lines for fold change of > 1.5 or |log2 Fold Change| > $\log_2(1.5)$. Genes that meet both these thresholds are colored in blue, whereas points colored in red are within the top 10 most significant genes by unadjusted p-values.

In calculating the concordance between the two platforms, an attempt has been made to replicate the author's method of at least 8 probes matching unanimously to a gene to be considered a match (Wang, 2014), but this has yielded low if not the absence of significant genes (*data not shown, please see R script for details*). In exploring alternatives, multiple probes were found to be mapping to the same gene, which would be expected, but also to multiple genes

across the dataset. As a result, the rat genome database (rat2302.db) was consulted and the initial results were similarly disappointing (**Supplementary Figure 1A**) with only one gene that mapped twice. When compared to the affymetrix map provided, there were more genes with the same threshold for matches at a minimum of 2, since 8 yields only one gene (**Supplementary Figure 1B-D**). For this reason, the provided refSeq-to-probeid map was used. As a result, the overall number of genes in the pool, N, in the concordance calculation, was based on the unique number of genes in this mapping file, which is 12,748 genes as opposed to 13,079 genes in Wang *et al.* (2014).

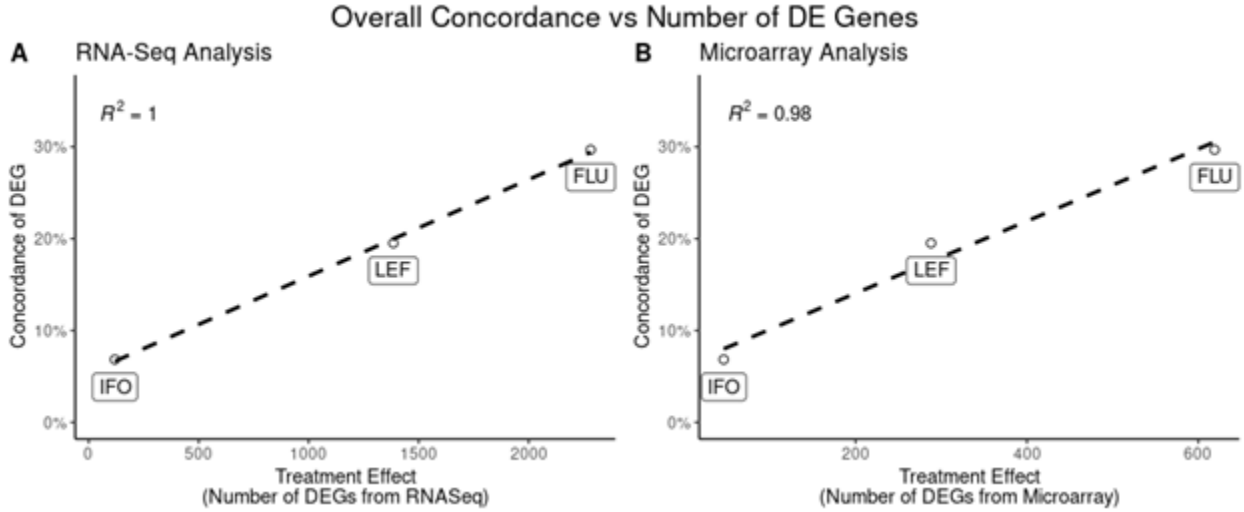


Figure 10. Plot of the overall concordance between RNA-Seq (**A**) and Microarray (**B**) analysis against the number of DEGs found in their respective platform. Level of concordance, measured in percentage, represents the extent of overlapping DE genes from both methodologies that are also in agreement of the direction of fold change. Concordance was also adjusted to account for the random chance of overlap (see **Methods**). Each labeled point represents a chemical treatment: IFO for Ifosfamide (MOA DNA Damage), LEF for Leflunomide (MOA AhR), and FLU for Fluconazole (MOA CAR/PXR).

Despite more than a 10 fold difference in the number of DEGs for the three chemicals used in our tox-group, there is a linear relationship between the concordance between these two platforms and the treatment effect. Agreement is higher at higher numbers of DEGs, but between the two platforms, the pairwise correlation (R^2) of 0.98-1.0 is almost perfectly linear as shown in **Figure 10**.

Having observed a strong correlation, we divided the probe sets into their respective directionality in fold change to determine if we still see similar concordance. As the summary bar plot indicates in **Figure 11**, the overall level of concordance correlates to the above median gene expression across all modes of action. Below median subsets correlates at roughly 50% of the overall concordance. Similarly enough, these results concur with those obtained from using the rat genome database (**Supplementary Figures 2 and 3**), although the overall concordance calculations are slightly lower than those in **Figure 11**.

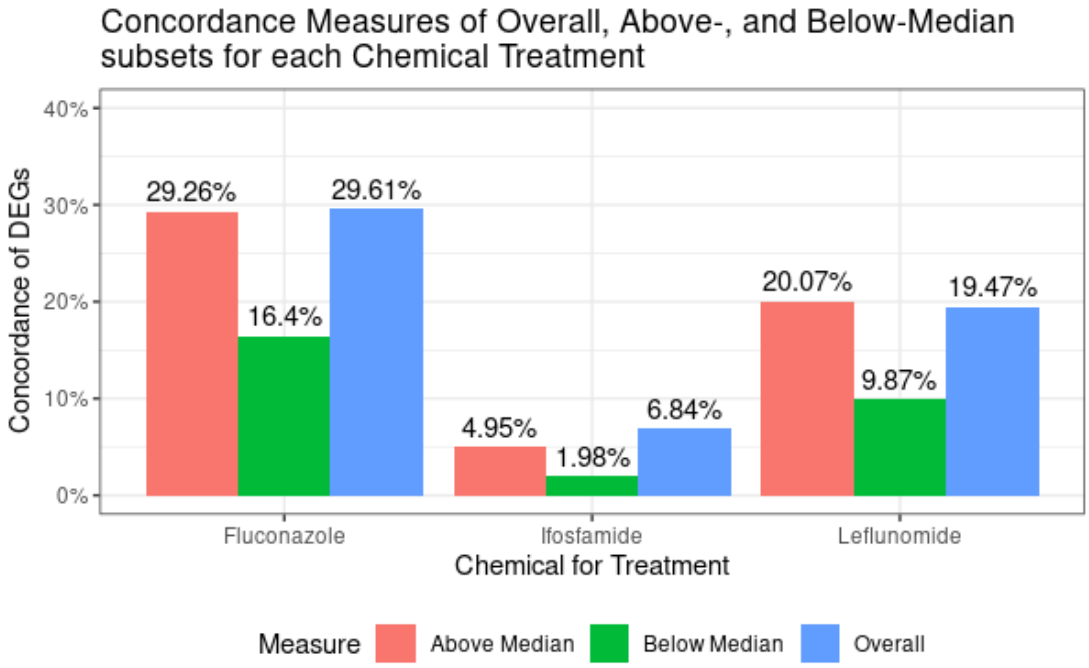


Figure 11. Summary of all concordance studies with the three chemical treatments: Fluconazole (MOA CAR/PXR), Ifosfamide (MOA DNA Damage), and Leflunomide (MOA AhR). Measure denotes whether the concordance calculated is for the above- or below-median subsets, or the overall DE gene list, respectively. Subsets are divided into half for above and below median expressed genes.

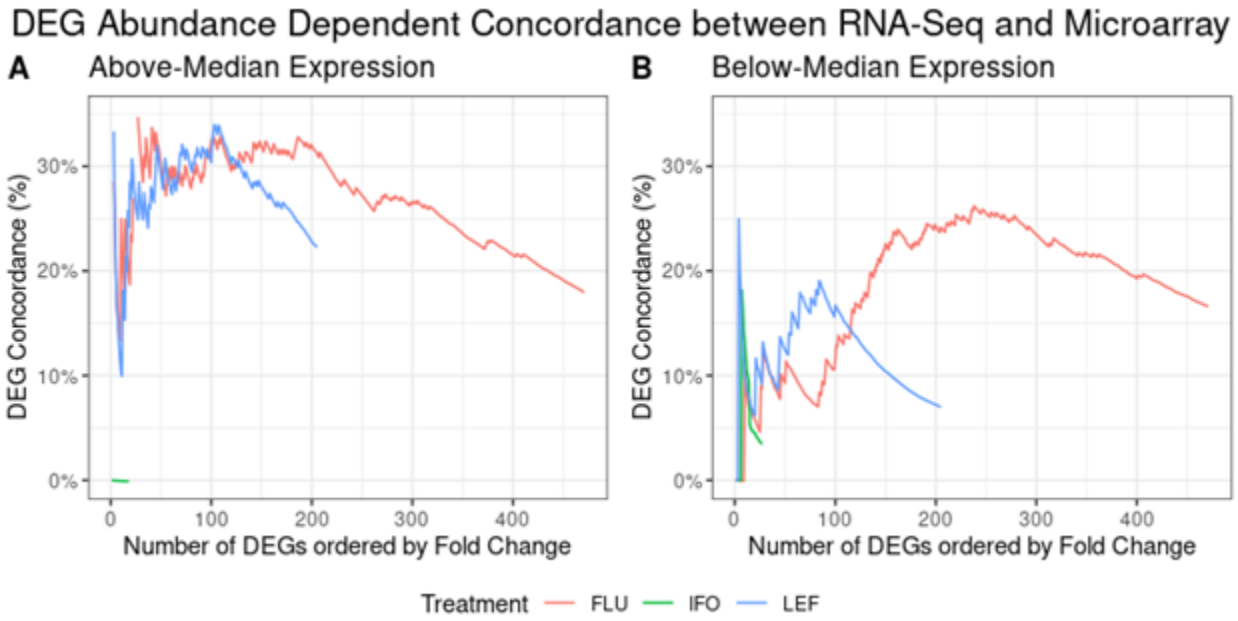


Figure 12. DEG Abundance Dependent Concordance between platforms. Above- and Below-Median subsets were ranked by decreasing fold change prior to concordance determinations. Degrees of concordance between the platforms, Microarray and RNA-Seq, were determined based on the increasing number of overlapping genes in both platforms. Only concordance for toxins in this study was determined: Fluconazole (FLU), Ifosfamide(IFO), and Leflunomide (LEF).

As an attempt to observe the effect of pool size of DEGs on concordance in different directionality of fold changes, concordance calculations were performed with increasing number of DEGs in each subset. As **Figure 12** indicates, the initial smaller pool of DEGs will have a varied, typically low, level of concordance, but as the pool of DEGs increases, so does the level of concordance and at a more stable rate before trailing off.

The pathway enrichment analysis yielded a total of 36 enriched pathways. For conciseness, only the top 10 enriched pathways were included in our analysis as represented in **Table 7**. Of these 10 pathways, the "metabolic pathways" pathway had the lowest p-value, at 1.66E-07, with a total of 174 genes being clustered together, for a total of 11.67% of the 1609 genes used as input in the DAVID analysis. This is in contrast to the lowest performing pathway enriched by the DAVID analysis— the thyroid hormone signaling pathway— that had a p-value of 0.089, and a total of 15 clustered genes, which accounts for a total of 1.006% of the query genes. The fold-enrichment values for these 2 pathways were 1.430 and 1.586 respectively. Unsurprisingly, the FDR for the highest performing pathway (5.42E-05) was much lower than that of the lowest performing pathway (0.804005).

Term	Count	PValue	Fold Enrichment	Bonferroni	Benjamini	FDR
rno01100:Metabolic pathways	174	1.66E-07	1.430071	5.42E-05	5.42E-05	5.42E-05
rno03040:Spliceosome	22	0.001151	2.152221	0.313764	0.188159	0.188159
rno00480:Glutathione metabolism	13	0.006871	2.400236	0.895095	0.472624	0.472624
rno00982:Drug metabolism - cytochrome P450	13	0.008588	2.334477	0.940409	0.472624	0.472624
rno01524:Platinum drug resistance	14	0.009562	2.211154	0.95679	0.472624	0.472624
rno05200:Pathways in cancer	57	0.009589	1.38887	0.957183	0.472624	0.472624
rno05202:Transcriptional misregulation in cancer	25	0.011498	1.698055	0.977214	0.472624	0.472624
rno04610:Complement and coagulation cascades	14	0.012765	2.134021	0.985021	0.472624	0.472624
rno05100:Bacterial invasion of epithelial cells	13	0.013008	2.213205	0.986178	0.472624	0.472624
rno00860:Porphyrin metabolism	9	0.014542	2.743741	0.991688	0.475515	0.475515
rno04919:Thyroid hormone signaling pathway	15	0.088514257	1.585764169	1	0.80400451	0.804004505

Table 7. Top 10 enriched pathways and the singular lowest performing enriched pathway generated by DAVID's functional annotation tool using an input of 1609 genes from *R. norvegicus*. Results are sorted by P-value (low at top high at bottom). All of these pathways are derived from the Kegg pathway under DAVID's functional annotation analysis. Highlighted in green is the highest performing pathway generated by the DAVID analysis. Conversely, red denotes the lowest performing pathway.

As seen in **Figure 13**, SRR1177981 and SRR1178006 were clustered together while SRR1177982 and SRR1177983 were clustered together. SRR1178013 were found in the same clade as the four previously mentioned samples but it was a standalone node. SRR1178004 was

an outgroup to the clade containing the six previously mentioned samples. When examining the expression value profiles that were the basis of this clustering, it is evident that the expression profile for the outgroup sample SRR1178004 was markedly different from the remaining five samples. For sample SRR1178004, the gene expression value patterns were almost always the opposite to that which was observed in the ingroup of five samples.

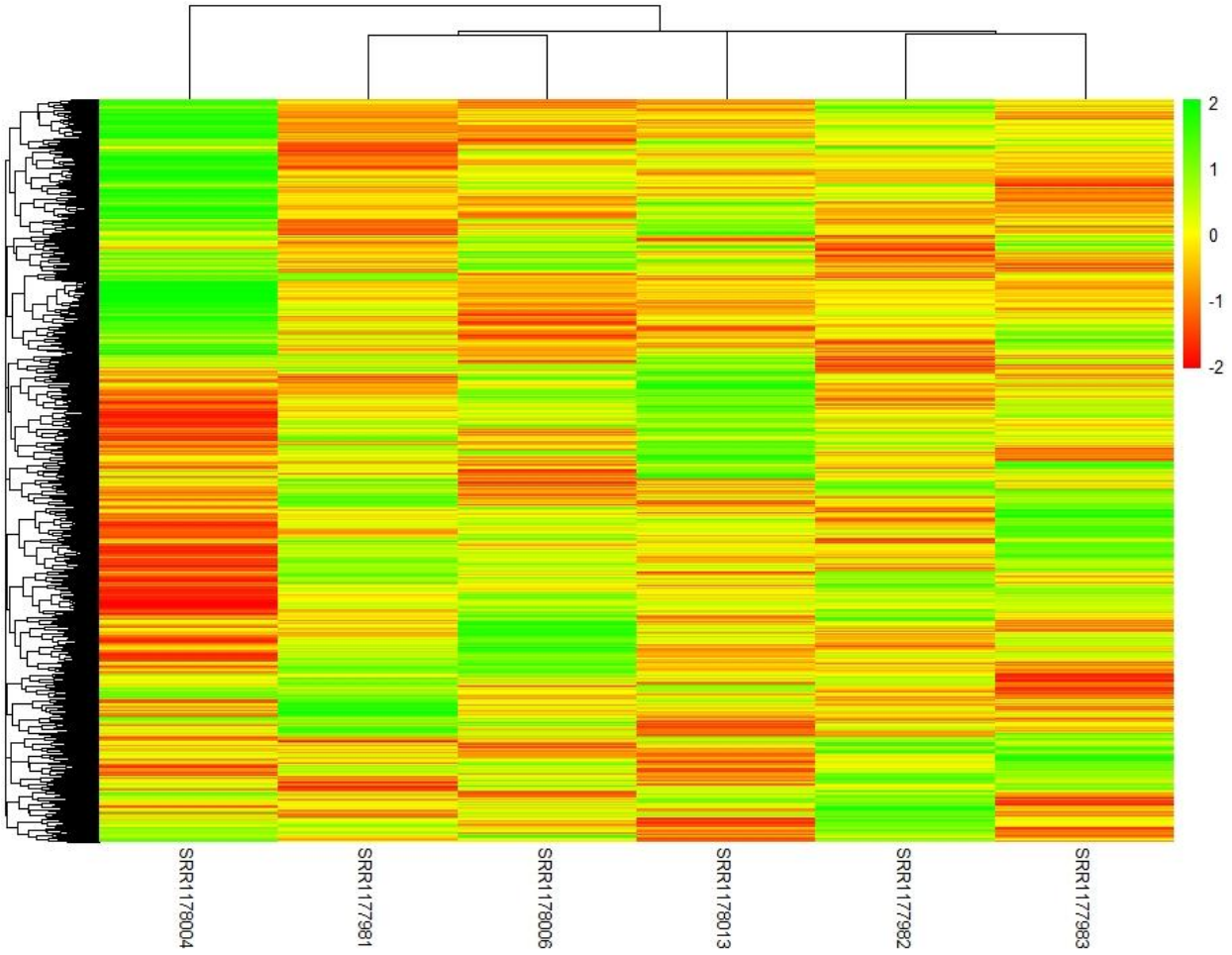


Figure 13. Clustered heatmap for samples subjected to different treatments representing different MOA. A total of 10695 genes and 6 samples were included in this clustered heatmap. Upregulated genes are denoted by green while downregulated genes are denoted by red. Genes with no variable expressions are represented by yellow cells.

DISCUSSION

In our attempt to replicate the concordance study on RNA-Seq and Microarray analysis, we have determined that the agreement between the two platforms are positively correlated to the level of chemical treatments, and hence the level of DEGs that were present. This is confirmed in the subset studies, where a higher concordant observation in the above median measures is an indication of better agreement between platforms in differential gene analysis when biological

processes are upregulated. Although one of our chemical treatments, Ifosfamide, had a lower level of perturbation, RNA-Seq still detected DEGs, and was thus better on this lower spectrum than Microarray in this respect. A similar study in rat liver toxins also came to a similar conclusion that RNA-Seq is a good alternative for downstream studies in lieu of microarrays, but also pointed out that RNA-Seq can also identify non-coding DEGs (Rao, 2019). This implies that chemical treatments in the DNA damage mode of action may simply be initiating a lower level of gene expression compared to others. As this study is only with a small subset of chemical treatments, the increasing measure of concordance when ranked by fold changes (**Figure 12**) may not be a reliable measure of real effect size, but the similarities in the general trend is worth noting.

Despite the concordance of our results to Wang *et al.* (2014), there were inconsistencies with the number of significant genes as well as how these genes were mapped. Unlike the original study, we found no significant genes when filtering with adjusted p-values for the chemical treatment involved in DNA damage, Ifosfamide, but this may be the result of the design differences in the use of limma, as this was not clearly specified in the paper. This may also relate to the different levels of filtering before obtaining these results. Nonetheless, the levels of DEGs detected in both studies were considerably low in stark contrast to the other two modes of action. This may indicate that the chemical perturbations are more pathway specific in the liver than previously expected since DNA damage encompasses a wider range of cellular pathways.

As Stalteri *et al.* (2007) had mentioned in their study of mapping probes in microarray studies, probes may not map to the same gene consistently due to the variations of gene expression as well as biological conditions (i.e. alternative splicing) that occurs at the cellular level. For this reason, we may be seeing the discrepancies in the number of genes that are mapped in our concordant calculations with those in Wang *et al.* (2014). While we are ideally using the same set of data, the annotations for these probes can improve and change over time, as we obtain more accurate data or as the technologies improve, so the mapped genes may be different depending on the version of the map that is used. This was seen with the use of the current rat genome database, which is updated biannually. Yet, despite that, there is not a big difference in the concordant observations nor the overall bi-directional fold change concordance (**Supplementary Figure 2 and 3**).

When conducting the gene enrichment analysis, it is worth noting that even though 2151 genes were input into the gene list query, only 1609 were included by the DAVID analysis. A total of 542 gene probes were not recognized by the DAVID webpage, and this likely had to do with the fact that several of the probes were named in a format that was not consistent with the probe type selected when initially uploading the gene list that we were trying to conduct the enrichment analysis on.

When looking at the highest performing upregulated genes in **Table 7**, it is evident that many of the enriched pathways were related to metabolism, drug resistance, and cancer regulation. This is largely consistent with what we see in the supplementary materials of Wang *et al*. (2014). Cell cycle regulation– specifically G2/M DNA damage checkpoint regulation– and Xenobiotic metabolism signaling are two of the most commonly enriched pathways for the DNA damage MOA, which is also one of the selected treatment groups in our study. Based on this

observation, it seems probable that chemicals that operate under analogous MOAs will likely have similar effects on the gene expression profiles of rat cells. The use of *R. norvegicus* cells in this instance would suggest that this finding is also applicable to humans to a certain extent, especially when considering how approximately 95% of the genes between the two species– as well as mice– are shared (Bryda E. C., 2013).

CONCLUSION

RNA-Sequencing and Microarrays are two commonly used platforms for measuring differential gene expression. To establish confidence in the use of either interchangeably, Wang *et al.* performed a concordance study on gene expression levels of rat livers post chemical treatments that encompass various modes of action (2014). In this replication analysis, we have found that the overall conclusion is in agreement with Wang *et al.* (2014) despite minor inconsistencies due to propagating differences in the filtering and selection process. We have determined that the concordance between platforms is overall dependent on the effect size of differentially expressed genes as well as the level of expression.

Based on the results of this study and our personal findings, it is not entirely unreasonable to assume that chemicals of similar MOAs would induce similar effects in rats and the organisms it is intended to model after. Overall, RNA-Seq and Microarrays are both equally sufficient in transcriptomic measurements.

ADDITIONAL INFORMATION

All scripts used in this analysis can be found at our public github repository at https://github.com/BF528/project-3-tinman-1.

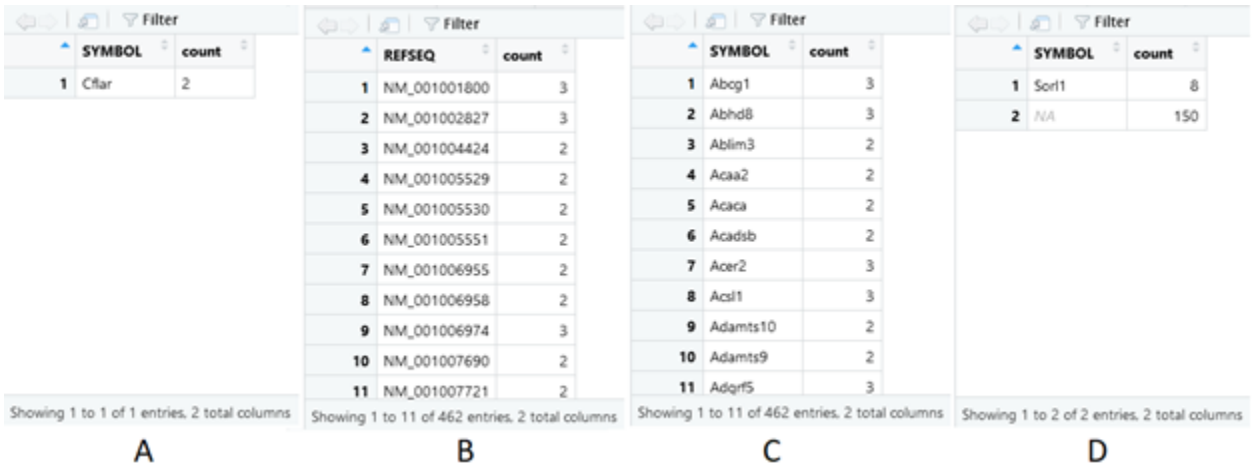
REFERENCES (APA)

1. Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., Meehan, J., Li, X., Yang, L., Li, H., Łabaj, P. P., Kreil, D. P., Megherbi, D., Gaj, S., Caiment, F., van Delft, J., … Tong, W. (2014). The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nature biotechnology, 32(9), 926–932. https://doi.org/10.1038/nbt.3001
2. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods, 5(7), 621–628. https://doi.org/10.1038/nmeth.1226
3. Alexander D., Carrie A. D., Felix S., Jorg D., Chris Z., Sonali J., Philippe B., Mark C., Thomas R. G., STAR: ultrafast universal RNA-seq aligner, Bioinformatics, Volume 29, Issue 1, January 2013, Pages 15–21, https://doi.org/10.1093/bioinformatics/bts635
4. Romiguier, J., Ranwez, V., Douzery, E. J., & Galtier, N. (2010). Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and

chromosome sizes. Genome research, *20*(8), 1001–1009. https://doi.org/10.1101/gr.104372.109

5. Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, *32*(19), 3047-3048.
6. FASTQC- https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
7. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/
8. Morgan, M. (2021). BiocManager: Access the Bioconductor Project Package Repository. R package version 1.30.16. https://CRAN.R-project.org/package=BiocManager
9. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7), e47.
10. Pagès, H., Carlson, M., Falcon, S., Li, N. (2021). AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. R package version 1.56.2. https://bioconductor.org/packages/AnnotationDbi
11. Carlson, M. (2021). rat2302.db: Affymetrix Affymetrix Rat230_2 Array annotation data (chip rat2302). R package version 3.13.0.
12. Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
13. Kassambara, A.(2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr
14. Bryda E. C. (2013). The Mighty Mouse: the impact of rodents on advances in biomedical research. Missouri medicine, 110(3), 207–211.
15. Rao, M. S., Van Vleet, T.R., Ciurlionis R., Buck, W.R., Mittelstadt, S.W., Blomme, E.A.G., Liguori M.J. (2019) Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies. Frontiers in Genetics, 9. https://doi.org/10.3389/fgene.2018.00636
16. Stalteri, M. A., & Harrison, A. P. (2007). Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. BMC bioinformatics, 8, 13. https://doi.org/10.1186/1471-2105-8-13

SUPPLEMENTARY FIGURES



A

	SYMBOL	count
1	Cflar	2

Showing 1 to 1 of 1 entries, 2 total columns

B

	REFSEQ	count
1	NM_001001800	3
2	NM_001002827	3
3	NM_001004424	2
4	NM_001005529	2
5	NM_001005530	2
6	NM_001005551	2
7	NM_001006955	2
8	NM_001006958	2
9	NM_001006974	3
10	NM_001007690	2
11	NM_001007721	2

Showing 1 to 11 of 462 entries, 2 total columns

C

	SYMBOL	count
1	Abcg1	3
2	Abhd8	3
3	Ablim3	2
4	Acaa2	2
5	Acaca	2
6	Acadsb	2
7	Acer2	3
8	Acsl1	3
9	Adamts10	2
10	Adamts9	2
11	Adgrf5	3

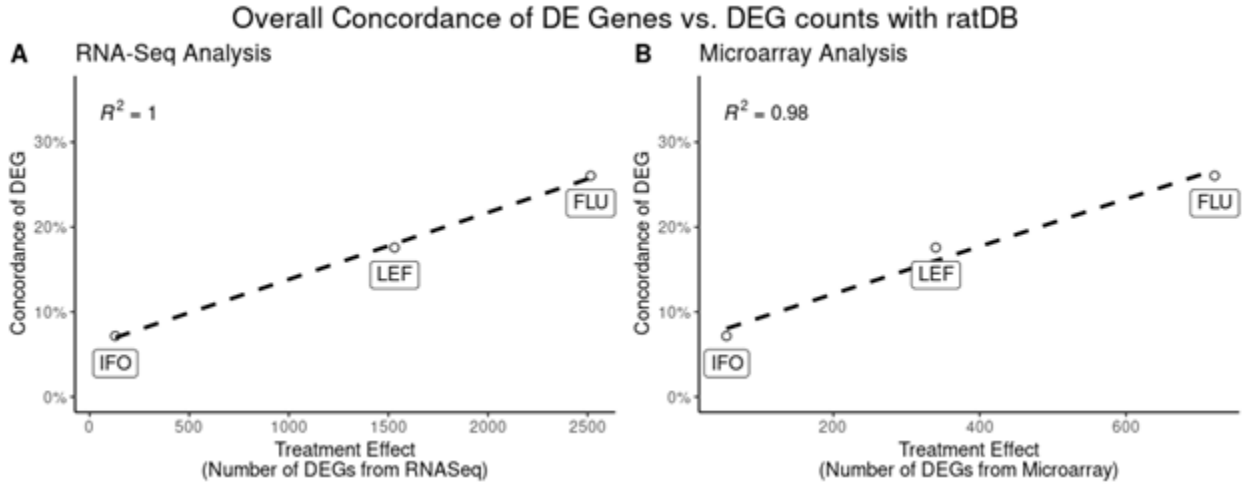
Showing 1 to 11 of 462 entries, 2 total columns

D

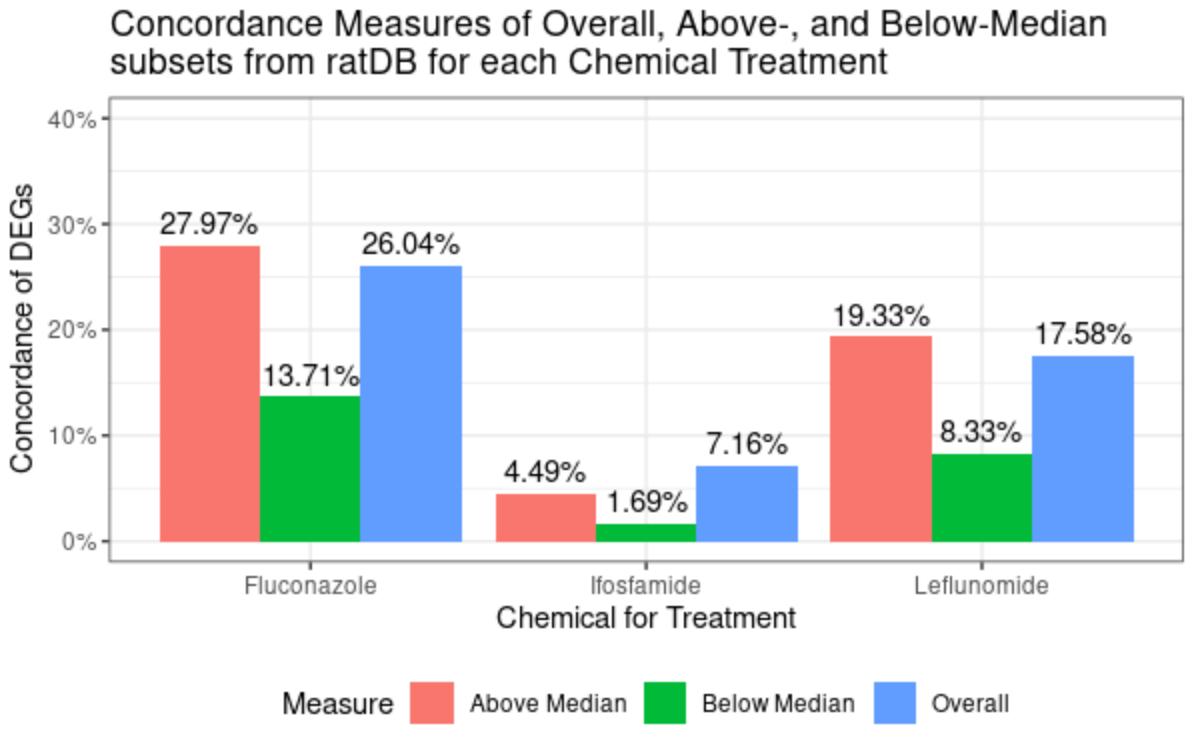
	SYMBOL	count
1	Sorl1	8
2	NA	150

Showing 1 to 2 of 2 entries, 2 total columns

Supplementary Figure 1. Results from attempting to filter RNA-Seq probes as described in original study: Genes were considered a match when mapped with 8 or more probes. (**A**) Gene symbols from filtering with rat database (rat2302.db) with map count of ≥ 2, (**B**) REFSEQ probes and its respective SYMBOLS (**C**) from filtering with the refseq-to-probeid map provided for our study with a map count of ≥ 2, or ≥ 8 (**D**).



Supplementary Figure 2. Concordance of DE Genes between Microarray and RNA-Seq using rat2302.db for mapping. For concordance calculation information, see Methods. As with **Figure 10**, the pairwise correlation is between 0.98-1.00, conferring to a perfectly linear correlation between platforms.



Supplementary Figure 3. Summary of concordance studies for the three chemical treatments: Fluconazole (MOA CAR/PXR), Ifosfamide (MOA DNA Damage), and Leflunomide (MOA AhR). Concordance calculations were determined using rat2302.db for mapping. Measure indicates the above-, below-median subset of expressed genes, or overall DE gene list of probe sets used, respectively.