Concordance of microarray and RNA-Seq differential gene expression

Group: Hedgehog
Teaching Assistant: Joey Orofino

Data Curator: Dylan Beeber
Programmer: Merai Dandouch
Analyst: Qinrui Wu
Biologist: Rojashree Jayakumar

Introduction

Microarray analysis and RNA sequencing techniques are both used to measure the expression of genes in living organisms. As these technologies mature, questions have arisen about relative biases of each method. Some previous studies have suggested that RNA-seq is not as sensitive as microarray analysis in detecting genes with low levels of expression (Labaj et al., 2011), while other studies have determined it is more sensitive (Mooney et al., 2013). Even setting aside the limitations of the physical data collection, differences in data analysis pipelines can affect how the data is reported.

In order to distinguish the limitations of both microarray analysis and RNA-seq, Wang et al. set out to create a comprehensive study that compared the results of both techniques across several different samples and treatment conditions. In this study, Wang exposed male rats to 27 different chemicals, extracted RNA from the livers and analyzed the samples through both Affymetrix microarrays and Illumina RNA-seq. They established 3 important findings:

1. Concordance between RNA sequencing and microarray platforms are positively correlated with the level of perturbation elicited by the treatment.
2. RNA sequencing is better at detecting weakly expressed genes.
3. Gene expression models generated from both platforms are similar.

We attempt to partially recreate this study by focusing specifically on toxgroup 2 from the Wang study.

Data

Toxgroup 2 consists of samples collected from rats that were exposed to three different chemicals: beta-naphthoflavone, econazole, and thioacetamide, as well as samples from control groups for each of these chemicals. For each chemical sample, three biological replicates were included, resulting in 9 total experimental condition RNA-seq samples for toxgroup 2. Pre-processed microarray data was provided to us, as well as RNA-seq count files from the control samples. However, we processed and aligned all 9 toxgroup rna-seq samples. Sample quality on the 9 raw RNA-seq read files was assessed through the Fastqc package. Reads were aligned via the Star toolkit, using the standard arguments listed by the Encode project. Alignment for all 9 reads was completed in ~7 hours when submitted to a remote computing cluster. After all reads were aligned, Multiqc was run on all alignment and Fastqc files to obtain a summary report of the data.

The rate of successful mapping of reads to the reference genome was between 88-93% for all samples (Appendix Figure 1). However, when assessing sequence quality of the original reads, 14 out of the original 18 read files (2 paired ends per sample) were flagged for consistently degraded base quality toward the end of the read. The mean quality score by base position for each sample is provided in Figure 1. Additionally, we found that base composition towards the start of each read is biased toward a specific sequence in all samples, likely a technical artifact

from the methods used to generate the rna-seq data. Finally, we note that all samples were flagged for duplication levels (Appendix Figure 2), indicating that certain sequences may be over represented in the data. Five samples in particular (SRR1177966_1, SRR1177970_1, SRR77994_1, SRR1177994_1, and SRR77995_1) displayed a large proportion of overrepresented sequences, the largest of which was sample SRR1177993_1, in which over 1.05 percent of the reads in that sample were identical.

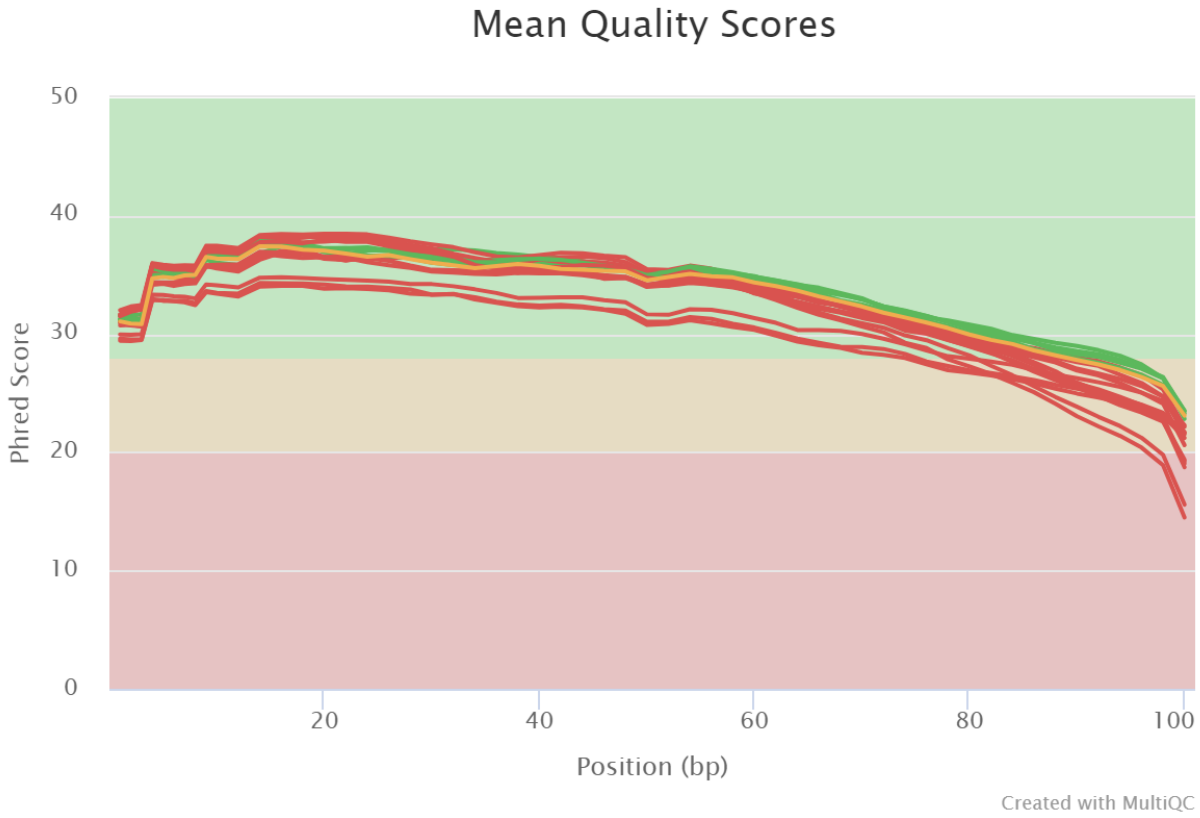


Figure 1: Mean quality scores for each of the 18 original read files. Samples flagged for poor quality are shown in red and orange, while samples that meet the quality criteria are shown in green. Sample quality loss is mostly due to a decrease in quality toward the end of the read.

Subsequently, all mapped reads for the 9 samples in Toxgroup 2 were counted against a reference annotation gtf file (rn4_refGene_20180308.gtf) using the featureCounts tool from the subread package (Liao et al., 2014). Multiqc was executed in the same directory as the file counts and statistics and plots describing the counts were outputted in a html file (Ewels et al., 2016). Below, tables and plots extracted from the html file are shown as table 1, and figure 2 in this paper. In table 1, SRR1177998, SRR1178001, SRR1178003 had the highest reads assigned with about 59 - 61% resolution with 28.9-30.7 million reads assigned against the reference annotation. All remaining samples had around 21.8-24.3 million reads assigned against the reference annotation displayed in the blue bar in figure 2.

MultiQC General Statistics

Sample Name	% Assigned	M Assigned
SRR1177966	59.7%	21.8
SRR1177969	61.7%	22.7
SRR1177970	61.7%	22.4
SRR1177993	59.5%	23.5
SRR1177994	60.7%	25.5
SRR1177995	60.4%	24.3
SRR1177998	59.7%	28.9
SRR1178001	60.1%	30.4
SRR1178003	61.7%	30.7

Table 1: General statistics table for MultiQC mapping for all experimental samples found in Toxgroup 2

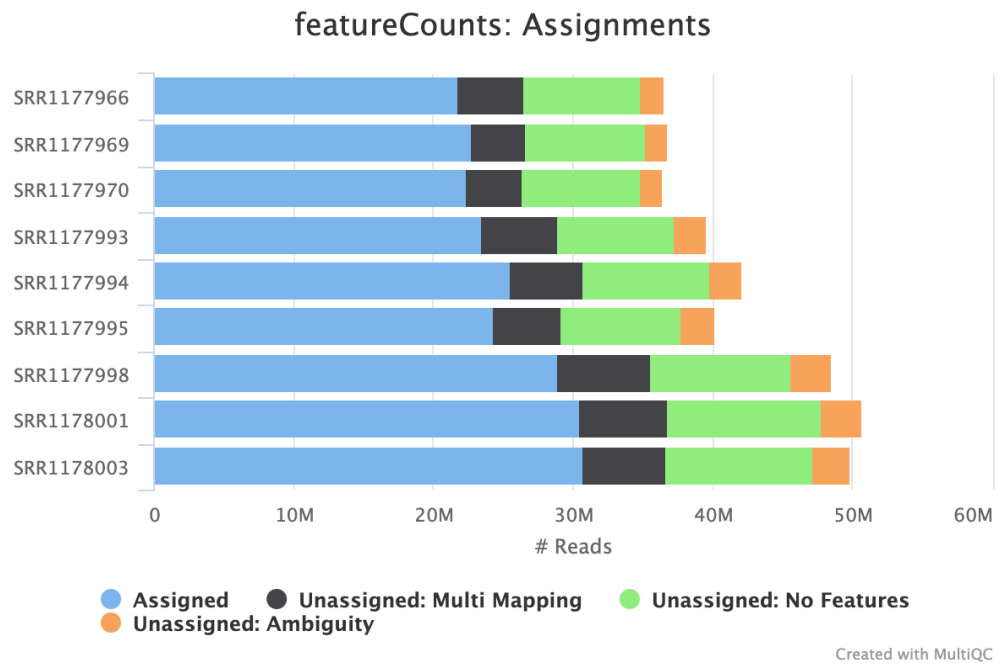


Figure 2: MultiQC mapping report for all experimental samples found in Toxgroup 2

Methods

Read Counts and RNA Seq Differential Expression using DeSeq2

The read counts for each sample in Toxgroup 2 were merged into a single dataframe using gene ID row names and sample column names. Count samples belonging to treatment and

control groups were combined and subsetted to only include samples corresponding to chemical exposure treatments such as Beta-Naphthoflavone (AhR MOA), Econazole (CAR/PXR MOA), Thioacetamide (Cytotoxic MOA). All read count samples for tox group 2 are distributed in a box plot below. It is shown that all sample read counts range between 0 to 750 counts and normalization prior to DeSeq2 was not necessary. The box plot shows that expression levels for all samples are comparable between each other. However, a combination of manual filtering and further optimization was required before analysis. For example, rows containing zero variance or 0 reads along all columns were removed. Additionally, outliers detected in the boxplot figure beyond maximum and minimum range were removed from the dataframe. Through the bioconductor DeSeq2 package, it was possible to normalize the reads applying median-of-ratios method where a pseudo-reference sample is created and counts are divided by sample-specific size factors determined by median ratio of gene counts relative geometric mean per gene (Love and Anders, 2014). Moreover, DeSeq2 utilizes binomial regression to estimate count difference given a statistical design. Results and statistical information including log2FoldChange, pvalue, padj were obtained and analyzed through histogram and scatter plot figures. Significant differentially expressed genes for each chemical exposure group were found by applying a filter that removed p-adjusted values greater than 0.05 and the following results were arranged by p-value. The top ten differentially expressed genes for each chemical exposure group are displayed in a table. All scripts utilizing R objects and bioconductor packages required little to no time to run.

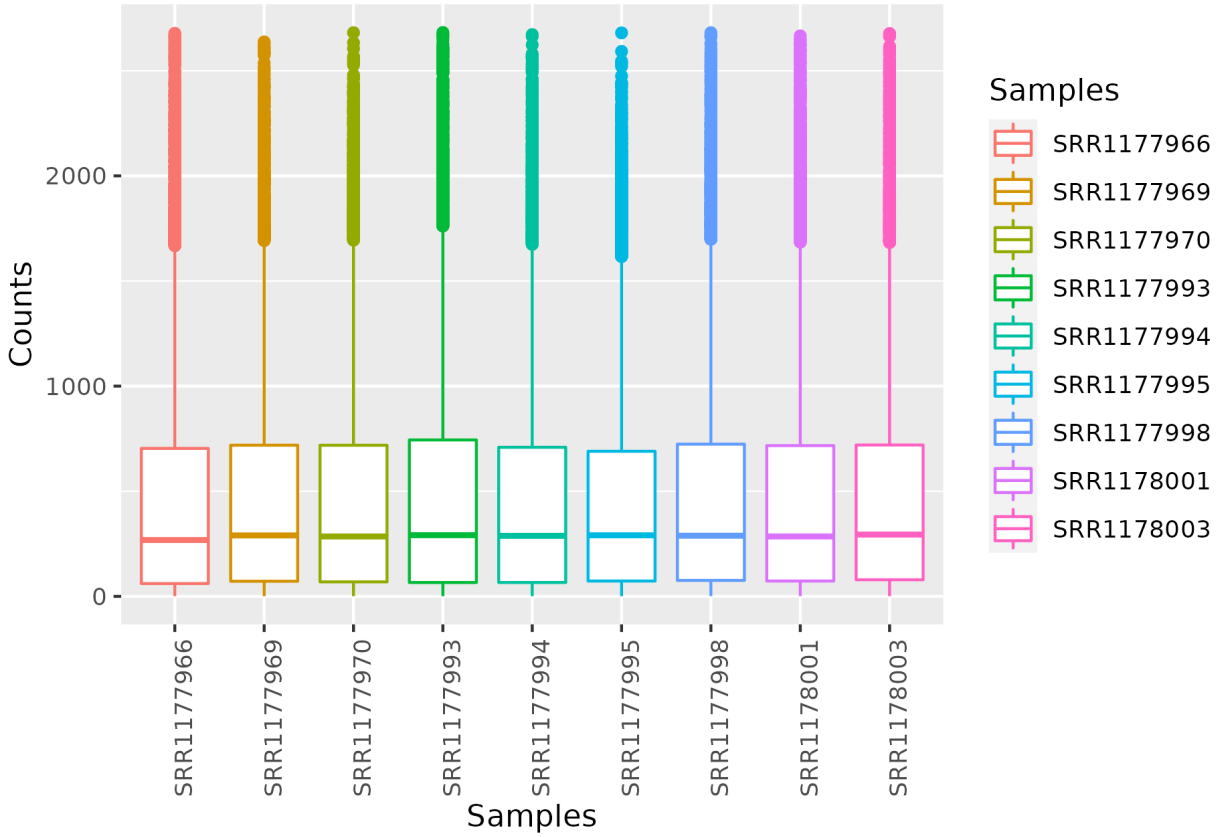


Figure 3: Boxplot distribution of all 9 read count samples from Toxgroup 2

LiMMA Differential Expression Analysis

Both Pre-normalized expression matrix and the sequencing data treatment from toxgroup 2 were used to run microarray differential expression analysis by LIMMA in R. Three chemicals(BETA-NAPHTHOFLAVONE, ECONAZOLE, THIOACETAMIDE) and control data were used to construct the design matrix in LIMMA. After differential expression results were generated, sorted all files by p-values, reported the number of genes significant at p-value < 0.05 and top 10 differential expression genes for all three files.

Concordance Computation

The concordance calculation equation that provided by Wang et al.'s paper:

$$\frac{2\times intersect(DEGs_{microarray}, DEGs_{RNA-Seq})}{DEGs_{microarray} + DEGs_{RNA-Seq}} \quad (1)$$

$$x + \frac{(n1-x)(n2-x)}{N-x} = n_0 \quad (2)$$

Equation (1) was used to compute the concordance result, equation (2) was used to calculate the intersect result in equation (1). n1 and n2 are the number of items in the microarray(LIMMA) and RNA-Seq(DESeq) result, separately. n0 is the intersection number between n1 and n2. N is the number of genes in mice, which is 54,879 (Bult CJ et al., 2015).

Pathway Enrichment Analysis and Clustering of normalized counts based on MOA

The genes in the results obtained from DESeq2 and LIMMA analysis were used for the pathway enrichment analysis. Only genes with p adjusted value < 0.05 and absolute $\log_2$ fold change > 1.5 were considered. DAVID functional annotation tool was used to find the common pathways enriched for each of the MOA chemical groups that are shared by both RNA-seq and microarray platforms. Only pathways with FDR < 0.05 were selected as significant for AhR and cytotoxic modes of action, whereas the threshold (FDR<0.1) was less strict for the CAR/PXR mode of action, as a stricter threshold did not yield significant pathways, especially from the DESeq2 results.

From the normalized count matrices of each tox group from RNA sequencing, a heatmap was constructed to induce a clustering consistent with mode of action. The filtering metrics to induce such a clustering were

1. Mean Filter: Mean of counts for each gene is greater than 20
2. Variance Filter: Variance significantly different from the median variance of all genes using a threshold of p>0.01 using two tailed chi squared test

3. Coefficient of Variance (CV) filter: Coefficient of variation of each gene is greater than 0.2

Results

Figure 4 displays three different histograms of log2foldchange for each chemical exposure group obtained from the DeSeq2 package. The Beta-Naphthoflavone histogram follows a bell shaped curve whereas Econazole and Thioacetamide have a pivot around 0. Beta-Naphthoflavone contains more genes that are negatively expressed compared to the remaining chemical exposure groups. However, most down regulated genes in Beta-Naphthoflavone only have a 1 fold change difference between all other genes. In contrast, Thioacetamide gene expression is more evenly distributed between positive and negative regulated log2foldchange (gene expression). A higher frequency for fold change of 2 - 4 for both positively and negatively expressed genes are seen in Thioacetamide compared to Econazole and Beta-Naphthoflavone. A similar trend can be found for Econazole except most genes only have a 0-2 fold change in positive/negative expression.

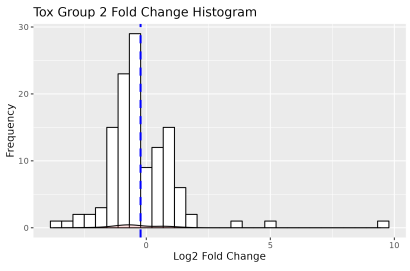
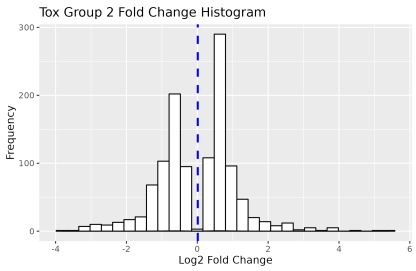
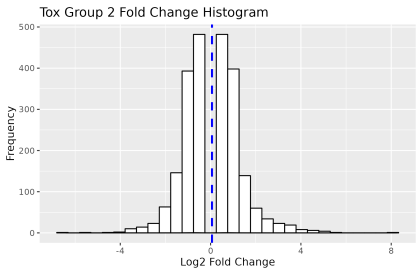
Beta-Naphthoflavone	Econazole	Thioacetamide
		

Figure 4: Histograms of log2foldchange for all genes in chemical groups AhR, CAR/PXR, and Cytotoxic

Beta-Naphthoflavone, Econazole, Thioacetamide are plotted using log2FoldChange against -log10(pvalue) in figure 5. The scatterplots were made after filtering the DeSeq expression results for pad < 0.05. Although the scatterplot for Beta-Naphthoflavone contains genes that are highly positively expressed, most of the genes are negatively expressed in the log2foldchange variable. It can also be found that Thioacetamide points are more scattered vertically, whereas data points in Econazole are spread horizontally. This is due to Thioacetamide having more genes that contain highly significant p-values for gene expression.

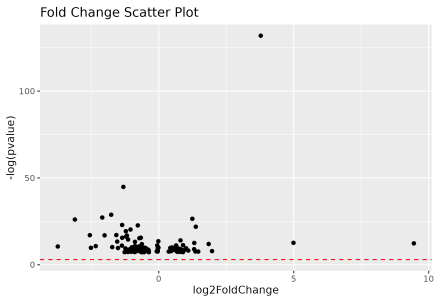
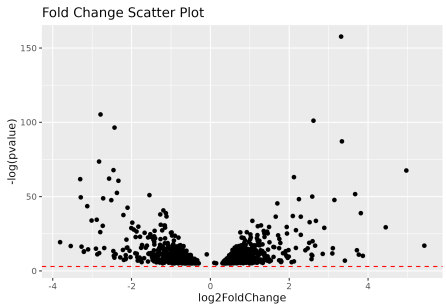
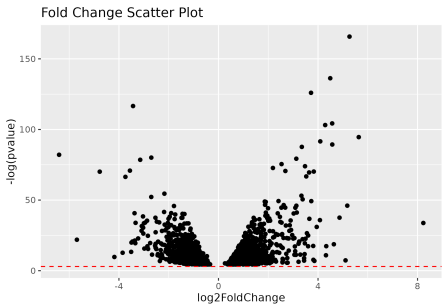
Beta-Naphthoflavone	Econazole	Thioacetamide
		

Figure 5: Scatterplots of log2foldchange against nominal -log10(p-value) for each chemical group containing Beta-Naphthoflavone, Econazole, and Thioacetamide

According to table 2, Thioacetamide contains the most differentially expressed genes with 2311 observations. In contrast, only 123 genes are differentially expressed under Beta-Naphthoflavone experimental conditions. Table 3 represents the top 10 differentially expressed genes for each chemical exposure group. It is found that NM_130407, NM_138502, NM_001109430 are the three top highly expressed genes for Beta-Naphthoflavone with padj values of 4.37E-54, 1.42E-16 and 8.65E-10. The three top highly expressed genes for Econazole are NM_001190380, NM_013141, NM_001108565 with padj values of 2.97E-65, 8.49E-43, 3.69E-41. Lastly the three top highly expressed genes for Thioacetamide are NM_001109260, NM_001130500, NM_001039344 with padj values of 1.03E-72, 6.54E-60, 1.97E-55. Thiacetamide's top three genes have the most significant padj values compared to beta-naphthoflavone and econazole. According to table 2, thioacetamide has the highest number of differentially expressed genes as well.

Treatment (Chemical)	**Total number of DEGs (adjusted p-value < 0.05)**
BETA-NAPHTHOFLAVONE	123 Observations
ECONAZOLE	1161 Observations
THIOACETAMIDE	2311 Observations

Table 2: Summary of results for three chemical exposure experiments using DeSeq2

	baseMean	log2FoldChange	lfcSE	pvalue	padj	Group
NM_130407	797.606396337528	3.777718292451	0.237805800065332	5.13391495024378E-58	4.37460892910273E-54	BETA-NAPHTHOFLAVONE
NM_138502	1661.1044345163	-1.31144125732263	0.147716413194155	3.33459716182838E-20	1.42070512079698E-16	BETA-NAPHTHOFLAVONE
NM_001109430	745.884470668853	-1.76209123437786	0.256673176498491	3.04559980478343E-13	8.65051864551987E-10	BETA-NAPHTHOFLAVONE
NM_001191751	140.584983141309	-2.0934309087456	0.315858828061884	1.53296487017614E-12	3.26559841469273E-09	BETA-NAPHTHOFLAVONE
NM_172019	674.623788787319	1.24341582308974	0.190146789939784	3.07681173567768E-12	5.24350255994 19E-09	BETA-NAPHTHOFLAVONE
NM_033352	62.0052686452008	-3.10916863966349	0.48099571591846	4.84468214931718E-12	6.88025609905528E-09	BETA-NAPHTHOFLAVONE
NM_001013889	668.564527513134	-1.35728575784498	0.22745528831142	1.10193926087025E-10	1.34137492026791E-07	BETA-NAPHTHOFLAVONE
NM_138911	1677.46048814073	-0.777366567933147	0.130651162945071	1.43165527381E-10	1.52489182351687E-07	BETA-NAPHTHOFLAVONE
NM_053293	1333.92422051551	1.37140470755215	0.23626302101028	3.03884689451931E-10	2.87711270979989E-07	BETA-NAPHTHOFLAVONE
NM_013059	239.525251730185	-1.04785192230938	0.189473940338575	1.51332361882597E-09	1.28950305560161E-06	BETA-NAPHTHOFLAVONE
NM_001190380	916.181394757529	3.31653131519965	0.190198313308798	3.27583773723889E-69	2.97609858428153E-65	ECONAZOLE
NM_013141	1008.65663526191	-2.78753189744465	0.197618620293927	1.86837636878647E-46	8.48709965521254E-43	ECONAZOLE
NM_001108565	555.05090330111	2.61528318699797	0.189046271861011	1.21811666838011E-44	3.68886331074444E-41	ECONAZOLE
NM_012508	764.34152869341	-2.43240642364531	0.180596154227497	1.28764584866838E-42	2.92456563378805E-39	ECONAZOLE
NM_139115	219.859464422783	3.33716767152029	0.261679360092941	1.36542907437964E-38	2.48098462814781E-35	ECONAZOLE
NM_198750	554.97141927038	-2.82471405479551	0.242307597496467	1.09075348641079E-32	1.65158257067368E-29	ECONAZOLE
NM_017128	349.099761961133	-2.45876029729272	0.220353299915708	3.46124927936629E-30	4.49220710043468E-27	ECONAZOLE
NM_001113422	147.929798744364	4.97140255122376	0.440014888016891	4.49966837392911E-30	5.10993589714325E-27	ECONAZOLE
NM_001277694	381.092848619105	2.12004557049739	0.197648675740237	3.89628924439384E-28	3.93308753170201E-25	ECONAZOLE
NM_001111269	489.728773429988	-2.57031775048333	0.241595291326389	1.05902696028E-27	9.62125993414382E-25	ECONAZOLE
NM_001109260	548.806231962047	5.26525621186151	0.293684579617026	1.02621854040711E-72	9.70700117371086E-69	THIOACETAMIDE
NM_001130500	403.533834096833	4.49105113865487	0.278769963795928	6.54265702419031E-60	3.09434963959081E-56	THIOACETAMIDE

NM_001039344	539.867734390671	3.71886927917815	0.239743762312042	1.96864744954018E-55	6.2071454084002E-52	THIOACETAMIDE
NM_053923	700.958031977909	-3.43289249113378	0.230607707405893	2.30820269484443E-51	5.45832232263336E-48	THIOACETAMIDE
NM_001106632	328.072814104857	4.56893584653975	0.325959489910713	5.09569518272518E-46	9.6400361466795E-43	THIOACETAMIDE
NM_019290	319.969867750619	4.28806510953524	0.307098534338708	1.62451412008574E-45	2.56104651031516E-42	THIOACETAMIDE
NM_012548	274.28533742798	5.63669239584693	0.422735706634394	8.3296606206195E-42	1.12557514014914E-38	THIOACETAMIDE
NM_181086	752.499084012169	4.09051183558071	0.31231534422166	1.71462335571184E-40	2.02732779020979E-37	THIOACETAMIDE
NM_001107909	256.435225336388	4.57266328086867	0.353812094266959	1.56696151626622E-39	1.6468765535958E-36	THIOACETAMIDE
NM_139089	373.490490350759	3.34926440341293	0.26117730093279	8.47040099328276E-39	8.01215229954617E-36	THIOACETAMIDE

Table 3: Top 10 differentially expressed genes and results from DeSeq2 package for each group (Beta-Naphthoflavone, Econazole, Thioacetamide). NM_130407, NM_001190380, NM_001109260 have the highest significant p-value and padj value for Beta-Naphthoflavone, Econazole, and Thioacetadmide.

Microarray Differential Expression with Limma

Limma analysis allows the generation of microarray differential expression results for different chemicals with corresponding modes of action (MOA). In this study, we primarily analyzed toxgroup 2 includes three chemicals: BETA-NAPHTHOFLAVONE, ECONAZOLE, and THIOACETAMIDE. Each chemical had different amounts of differential expression genes(DEGs) at adjusted p-values at 0.05 (table 4). After sorting the results by p-value, we created three tables containing the top 10 DEGs for each chemical (table 5A, 5B, 5C). Furthermore, we visualized log2 fold change values from the significant DEG from three analyses as histogram (figure 6) and log2 fold change values versus nominal p-value as scatter plot (figure 7).

After completing both RNA-Seq and LIMMA analyses, overall concordances for three chemicals were calculated and showed the relationship between two analyses and the treatment effect (figure 8). In addition, a combined bar plot shows overall, above and below concordance for three MOA (figure 9).

Treatment (Chemical)	Total number of DEGs (adjusted p-value < 0.05)
BETA-NAPHTHOFLAVONE	100
ECONAZOLE	1,641
THIOACETAMIDE	4,935

Table 4: Summary of DEGs for three treatments on microarray analysis by LIMMA

(A)

	logFC	AveExpr	t	P.Value	adj.P.Val	B
1370269_at	7.0939001	6.672310	30.887274	9.244819e-29	2.875046e-24	39.169982
1387243_at	1.5594092	13.070635	19.742482	9.944964e-22	1.546392e-17	31.568416
1370613_s_at	0.7871443	12.717554	11.826656	2.211144e-14	2.292146e-10	20.102119
1368990_at	1.4634571	5.600121	8.738548	1.135145e-10	8.825466e-07	13.214988
1387759_s_at	0.8552016	11.884771	8.194806	5.783745e-10	3.597374e-06	11.840816
1387811_at	-0.4491234	11.698817	-7.066755	1.870860e-08	9.696981e-05	8.854768
1367673_at	0.5916388	10.847197	6.798838	4.339663e-08	1.927988e-04	8.122790
1367856_at	1.2242843	8.965154	6.645583	7.036370e-08	2.735301e-04	7.700921
1387901_at	-0.5080534	8.654721	-6.430671	1.388498e-07	4.797878e-04	7.105957
1388122_at	0.8498477	8.769210	6.106652	3.881190e-07	1.207011e-03	6.202872

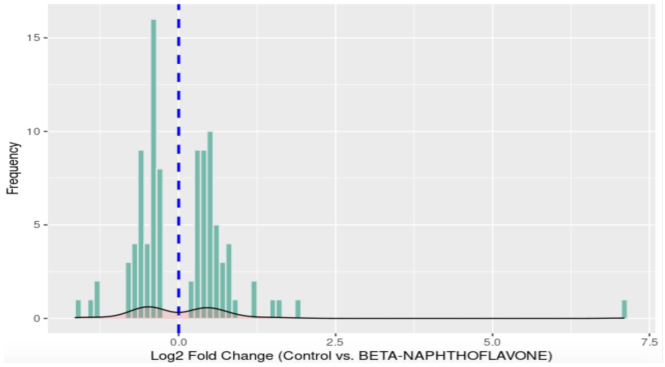
(B)

	logFC	AveExpr	t	P.Value	adj.P.Val	B
1384408_at	1.3300397	6.569189	9.147780	3.308190e-12	6.614064e-08	17.36084
1378027_at	-0.5903862	9.780592	-8.960162	6.311185e-12	6.614064e-08	16.75962
1371781_at	0.6836384	8.649026	8.957006	6.380331e-12	6.614064e-08	16.74947
1369012_at	-0.7579750	6.217808	-8.812668	1.051287e-11	8.173497e-08	16.28402
1388874_at	0.6410164	10.811525	8.709554	1.503822e-11	9.353469e-08	15.95003
1395403_at	-2.9497940	9.639711	-8.459181	3.601515e-11	1.866725e-07	15.13416
1370698_at	1.2229504	11.679376	8.303435	6.217580e-11	2.762293e-07	14.62334
1391570_at	1.1515154	5.853131	7.945870	2.193684e-10	8.527671e-07	13.44187
1380805_at	-0.5203340	4.010581	-7.777578	3.983125e-10	1.376347e-06	12.88202
1386944_a_at	-1.3313335	11.946889	-7.707798	5.103347e-10	1.453543e-06	12.64926

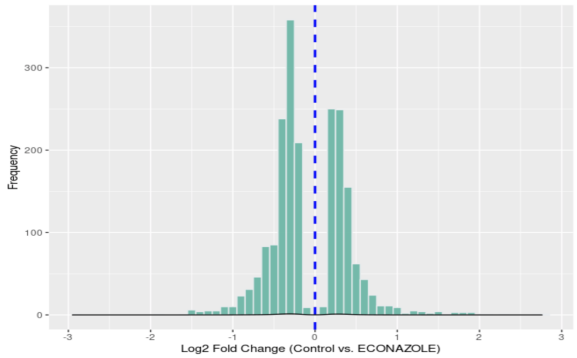
	logFC	AveExpr	t	P.Value	adj.P.Val	B
1373767_at	4.312615	6.701320	37.27957	3.083974e-28	9.590850e-24	51.61957
1370177_at	4.567961	5.750820	31.73207	5.045444e-26	7.845413e-22	47.51688
1380229_at	4.293742	4.099431	29.63742	4.315841e-25	4.473944e-21	45.71026
1369519_at	2.510463	4.222930	23.96611	3.187374e-22	2.478104e-18	39.90647
1385706_at	2.189917	3.929071	23.76161	4.149485e-22	2.580897e-18	39.66796
1367764_at	2.910065	7.901032	23.22209	8.404674e-22	4.356283e-18	39.02755
1396236_at	2.320036	3.736730	22.57065	2.009956e-21	8.929659e-18	38.23204
1368072_at	3.465270	6.341493	21.59620	7.727115e-21	3.003819e-17	36.99436
1397564_at	2.109776	4.158714	21.22330	1.312052e-20	4.533724e-17	36.50488
1387269_s_at	2.406845	4.348370	20.30027	5.041272e-20	1.567785e-16	35.25350

Table 5: Top 10 differential expression genes arranged by p-value (a) BETA-NAPHTHOFLAVONE (b)ECONAZOLE (c)THIOACETAMIDE.

(a)



(b)



(c)

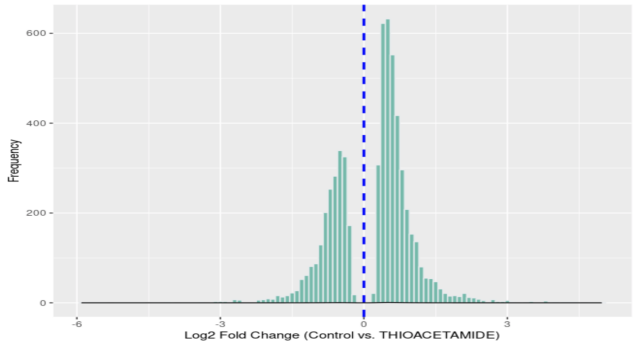
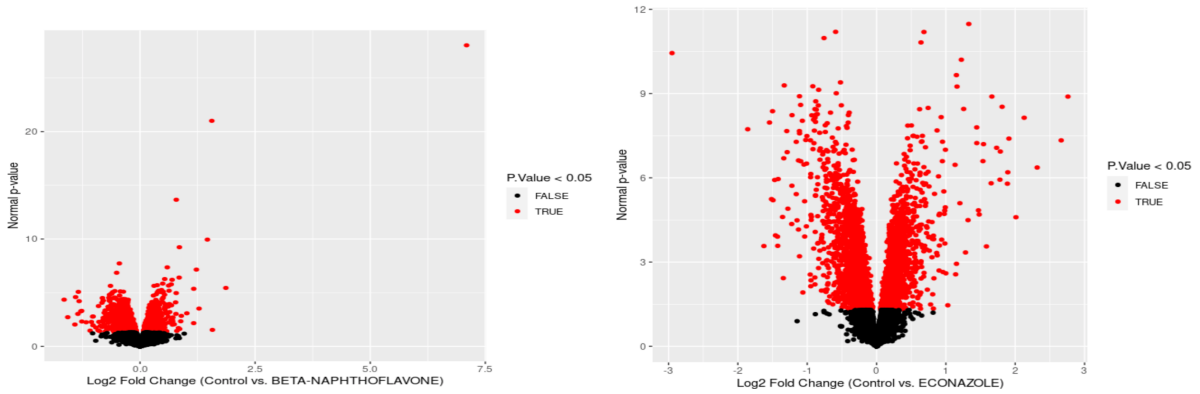


Figure 6: Histograms of log2 fold change for all significant differential expression gene in a) BETA-NAPHTHOFLAVONE b)ECONAZOLE c)THIOACETAMIDE

(a)

(b)



(c)

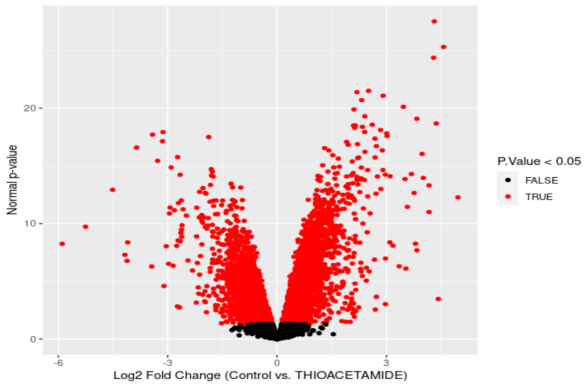


Figure 7: Scatter plots of fold change vs nominal p-value from three analyses a) BETA-NAPHTHOFLAVONE b)ECONAZOLE c)THIOACETAMIDE. Red points show genes significant at nominal value < 0.05.

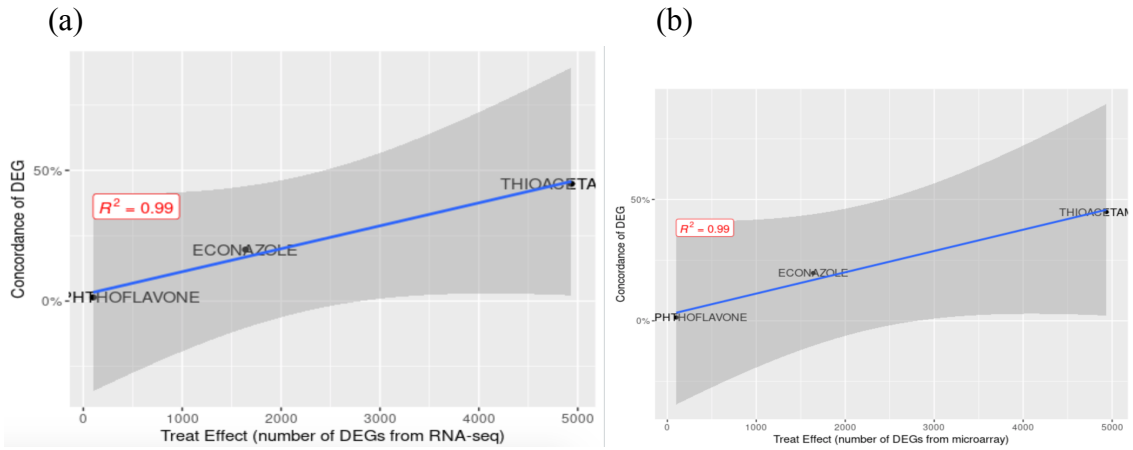


Figure 8: concordance vs the number of DE genes from the (a) RNA-Seq analysis and (b) microarray analysis.

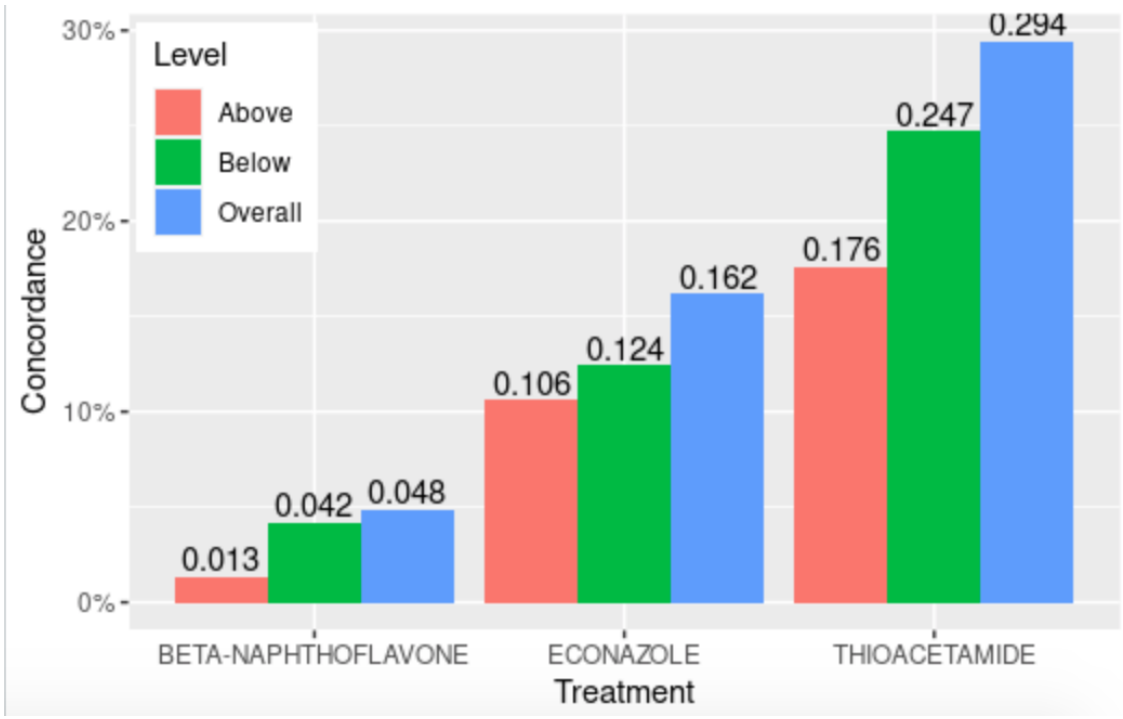


Figure 9: bar plot combining overall concordance measures you obtained for the overall DE gene list and the above- and below-median subsets.

Pathway Enrichment Analysis and Clustering of normalized counts based on MOA

For the pathway enrichment analysis, DAVID functional annotation chart was used. The default parameters were modified to include only pathways and all other ontology terms were

excluded, since we wanted to check the common pathways that get affected with respect to the transcriptional profile generated by the chemical treatments. Wang et al., too have compared the transcription response for rat livers in terms of extensive chemical treatments, biologic replication and shared mode of action of the chemical. However the results obtained were not entirely similar to the results by Wang et al. Table 2 summarizes the common pathways (with FDR value below 0.1 for CAR/PXR and 0.5 for AhR and cytotoxic MOA), shared by both the platforms for each mode of action.

AhR (6)
rno05207:Chemical carcinogenesis - receptor activation
R-RNO-211859~Biological oxidations
WP1302~Estrogen metabolism
rno00140:Steroid hormone biosynthesis
rno00830:Retinol metabolism
WP1286~Metapathway biotransformation
CAR/PXR (1)
rno01100:Metabolic pathways
Cytotoxic (7)
rno03008:Ribosome biogenesis in eukaryotes
R-RNO-6791226~Major pathway of rRNA processing in the nucleolus and cytosol
R-RNO-72312~rRNA processing
R-RNO-8868773~rRNA processing in the nucleus and cytosol
rno05160:Hepatitis C
rno04110:Cell cycle
rno04115:p53 signaling pathway

Table 2: List of common pathways enriched for each of the MOA chemical groups that are shared by both RNA-seq and microarray platforms. Only pathways with FDR values < 0.1 are included in the table.

A heatmap (Figure 10) was constructed to characterize the clustering based on the Mode of Action. Each mode of action formed a separate cluster. Euclidean distance was used for hierarchical clustering on the Mode of Action (MOA). The cytotoxic MOA clustered separately from the rest of the samples.

Figure 10: Heatmap of the Differential expression normalized counts. Clustering is done based on the mode of action of the tox groups. The annotation color on the top of the heatmap shows that there is a clear distinction between the treated groups

Discussion

In our reproduction of the study by Wang et al.,the differentially expressed genes for each of the treatment groups from the microarray platform using Limma and from RNA sequencing using DESeq2 packages in R was determined and the concordance between these two platforms was calculated.

All three chemical exposure groups showed highly differentially expressed gene results from the DeSeq2 and Limma package. However, Thioacetamide findings for differential expression results using both packages show that the Thioacetamide chemical exposure group contains more differentially expressed genes that are highly significant than Econazole and Beta-Naphthoflavone. A study found that rats who were subjected to acute thioacetamide exposure for around 8 or 24 hours experienced an inflammation stress response and increase in activated liver gene modules associated with inflammation and fibrosis (Schyman et al., 2018).

Thioacetamide could potentially induce more genes through acute exposure than Beta-Naphthoflavone and Econazole. Another reason for this could be due to discrepancy between mode of actions between microarray and RNA-seq platforms. According to Wang et al., some discrepancies occurred between receptor-mediated mode of actions (MOA) where a high overlap between AhR and CAR/PXR was observed for both platforms and a lower overlap for non-specific toxicity using Cytotoxic mode of action was encountered. This suggests that the difference in the number of highly expressed genes found in Thioacetamide (MOA using Cytotoxic) compared to Econazole and Beta-Naphthoflavone could be due to the MOA mechanism used for each chemical exposure group. Future studies should include both chronic and acute toxicity studies for both platforms, all MOAs and toxic chemicals. Nonetheless, a similar trend can be found for all three chemical exposure groups regardless of the number of differentially expressed genes obtained.

The trend of our measured overall concordance value for three chemicals versus treat effect is similar to figure 2a in Wang et al.'s paper. Additionally, figure 8A and 8B representing the number of DEGSs from RNA-Seq and LIMMA analysis also have similar trends for all three chemicals. This indicated we were able to reproduce the finding that differentially expressed genes in the cross-platform concordance is highly correlated with treatment effect size. In addition, compared to figure 3b+c in Wang et al.'s paper, the concordance trend of above-median and below-median groups (figure 9) was different. The concordance of the above-median group is supposed to be higher than the below-median group, while in our study, each chemical has a higher concordance score in the below-median group rather than the above-median group. It may be due to the below group having a higher n0 value, that is, the below groups in RNA-Seq and LIMMA analysis have more intersections compared to that in the above group. Based on our study, splitting the result into above-median and below-median may not be an ideal way to analyze cross-platform concordance. Picking the mean of baseMean value in RNA-Seq and the mean of AveExpr value in LIMMA as standard to split result should not be used as a method to determine up and down regulated genes.

Gene set enrichment analysis was performed to detect the common pathways between the two platforms for each mode of action. The enriched pathways obtained were different from the pathways detected by Wang et al. This is because in this analysis a different annotation tool (DAVID) was used, whereas the method used by Wang et al. was GeneGo:MetaCore. Furthermore, many of the pathways especially in CAR/PXR had a FDR higher than 1 and hence were filtered out in the analysis, which is the reason why there is only a very generic metabolic pathway shared between the two platforms and the results are different. AhR MOA in the paper has some pathways involved with Nicotine degradation and Retionoate biosynthesis. Our analysis retrieved the following related pathways such as Retinol metabolism and chemical carcinogenesis - receptor activation. Hence with respect to AhR MOA there was an overlap in the results, however it was not complete. The cytotoxic pathway in our analysis retrieved known

cytotoxic pathways such as p53 signaling pathway. However since this MOA is general, non-specific toxicity, it is expected that the results obtained will be vast, involving many signaling and metabolic pathways and was not similar to the results obtained by Wang et al. Another reason why there is dissimilarity in the results between our analysis and by Wang et al., may be because Wang et al.'s study design involved three chemicals per MOA whereas in our study we have utilized only one chemical per mode of action, hence there will be a difference in the number of significant pathways enriched.

To check if there were any similarities between our gene set enriched pathways and the results in the paper, all the pathways without filtering based on FDR values were examined. Appendix table 1 shows the common pathways when no filtering based on FDR was done on the DAVID functional annotation chart results. As expected, the cytotoxic MOA has around 75 enriched pathways and this may be because the cytotoxic MOA is non-specific in nature.

On examination, there were some pathways which overlapped with the results of Wang et al. For example, Wang et al. reported xenobiotic metabolic signaling as a pathway common in all three modes of action. Similarly drug metabolism was identified in all three MOA in this study. This was one significant overlap, as all three MOA are based on the drugs essentially xenobiotics that were introduced into the rat livers to assess the transcriptional data. However none of the pathways were significant.

However the results in this study are consistent with the mode of action used. For example, the AhR MOA involves treatment with beta-Naphthoflavone, which has estrogen receptor and steroid hormone receptors as targets (Wishart DS et al, 2017) and hence the pathways Estrogen metabolism and steroid hormone biosynthesis are enriched. Econazole is a broad spectrum antimycotic (Wishart DS et al, 2017) and the pathway signaling by Interleukins enriched, however the FDR was high to consider it significant. Thioacetamide is a carcinogen and hepatotoxin (Hunter AL et al, 2007) and pathway enrichment analysis of the differentially expressed genes confirms this as there are several carcinogenic pathways, cell cycle pathways and hepatitis pathways that are enriched.

To do hierarchical clustering of the samples based on MOA, filtering of the genes were done to remove noise. The mean filter was used to remove counts with mean lower than 20. The cutoff was set as 20, as incorrect clustering was obtained at a cutoff below 20. The filter retained 8497 genes from the total 9650 genes. The variance filter was used to retain only genes which have a variance different from the test statistic (p value < 0.01). The test statistic was calculated as a ratio of the product of the degrees of freedom and variance of genes to the median variance of all genes (Heckert N. et al, 2002). The number of genes retained after this filter was 7036. Finally the covariance filter retained genes with a covariance >0.2 and it retained 5530 genes. These genes were used for generating a heatmap and clustering of the samples was based on

Euclidean distance metrics. The samples clustered based on their mode of action. The cytotoxic MOA clustered separately from other samples, and this is convincing as genes upregulated in other groups are downregulated in the cytotoxic MOA and vice versa. This significant difference may be attributed to the cytotoxic MOA being non-receptor-mediated, whereas the others are receptor-mediated.

Conclusion

Overall, we were able to reproduce partial findings in Wang et al.'s paper. The number of DEGs in RNA-Seq and microarray analyses and the trend of cross platform differential expression gene concordance for three chemicals were similar to Wang et al.'s paper. While the concordance scores of the above-median were lower than that of the below-median group, this trend was different to the original paper.

The enriched pathways from our results were different from the original study by Wang et al. This is due to differences in the study design and functional annotation tools that were used. However, we were able to reproduce the clustering of the differentially expressed genes from both platforms on the basis of mode of action of the treatment groups.

References

Bult CJ, Eppig JT, Blake JA, Kadin JA, Richardson JE; Mouse Genome Database Group. Mouse genome database 2016. Nucleic Acids Res. 2016 Jan 4;44(D1):D840-7. doi: 10.1093/nar/gkv1211. Epub 2015 Nov 17. PMID: 26578600; PMCID: PMC4702860.

Heckert, N. , Filliben, J. , Croarkin, C. , Hembree, B. , Guthrie, W. , Tobias, P. and Prinz, J. (2002), Handbook 151: NIST/SEMATECH e-Handbook of Statistical Methods, NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD (Accessed April 9, 2022)

Hunter AL, Holscher MA, Neal RA. Thioacetamide-induced hepatic necrosis. I. Involvement of the mixed-function oxidase enzyme system. J Pharmacol Exp Ther. 1977 Feb;200(2):439-48. PMID: 839448.

Labaj PP, Leparc G, Linggi B, Markillie L, Wiley HS, Kreil D. Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. Bioinformatics. 2011; 27:383-391.

Liao Y, Smyth GK, Shi W: featureCounts: an efficient general purpose program for assigning sequence reads to genomic features . Bioinformatics. 2014, 30: 923-930. 10.1093/bioinformatics/btt656.

Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology*, **15**, 550. doi: 10.1186/s13059-014-0550-8.

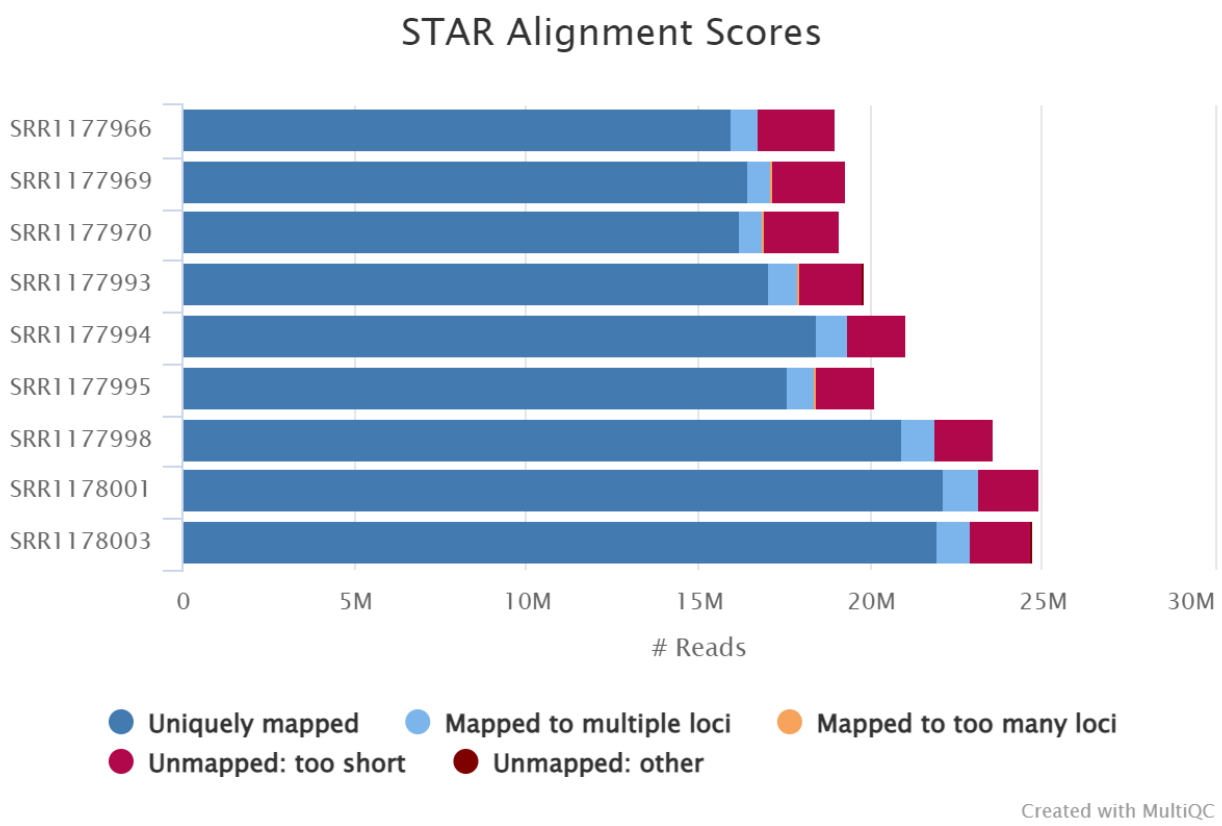
Mooney M, Bond J, Monks N, Eugster E, Cherba D, Berlinski P, Kamerling S, Marotti K, Simpson H, Rusk T, Tembe W, Legendre C, Benson H, Liang W, Webb CP. Comparative RNA-Seq and microarray analysis of gene expression changes in B-cell lymphomas of Canis familiaris. PloS one. 2013; 8:e61088.

MultiQC: Summarize analysis results for multiple tools and samples in a single report
Philip Ewels, Måns Magnusson, Sverker Lundin and Max Käller, Bioinformatics (2016),
doi: 10.1093/bioinformatics/btw354, PMID: 27312411
Schyman, P., Printz, R.L., Estes, S.K., Boyd, K.L., Shiota, M. and Wallqvist, A., 2018. Identification of the toxicity pathways associated with thioacetamide-induced injuries in rat liver and kidney. *Frontiers in Pharmacology*, p.1272.

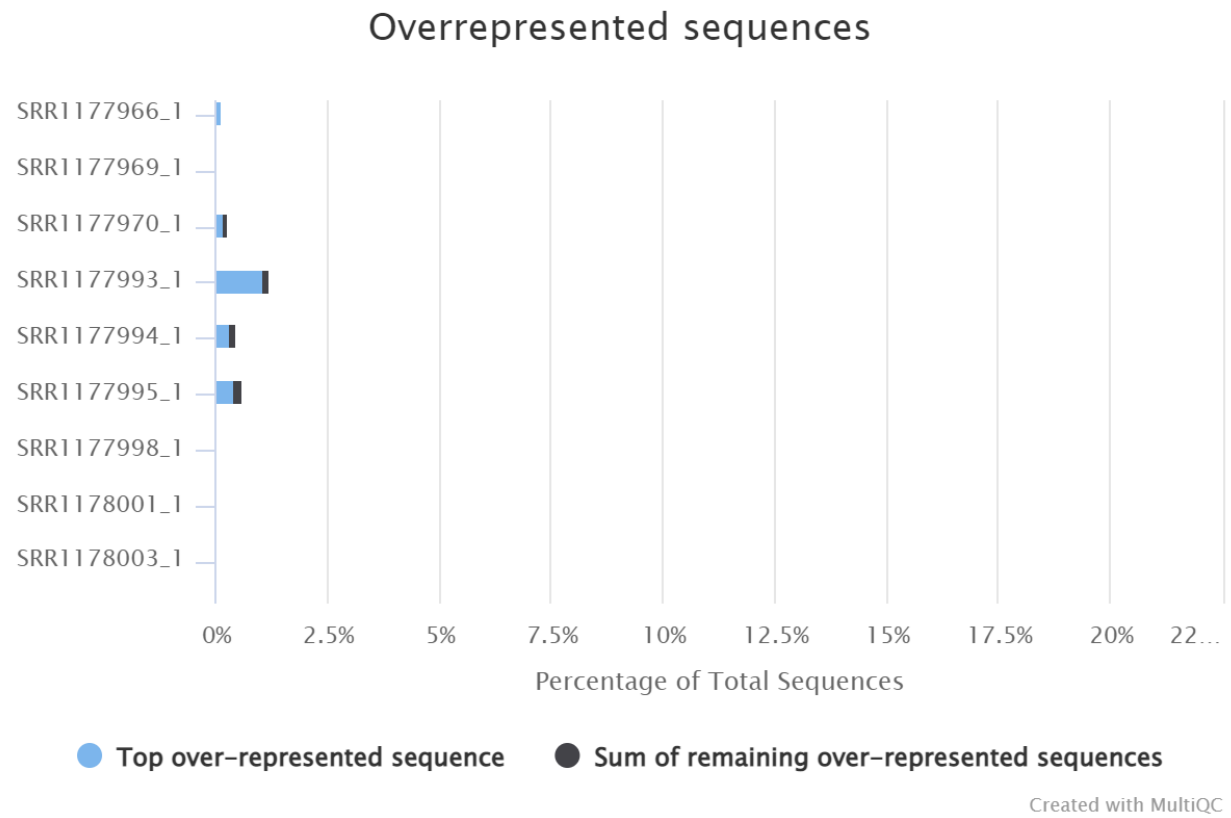
Wang C, Gong B, Bushel P, Thierry-Mieg J, Thierry-Miegg D, et al. A comprehensive study design reveals treatment- and transcript abundance–dependent concordance between RNA-seq and microarray data. Nat Biotechnol. 2014; 32(9):926–932.

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017 Nov 8. doi: 10.1093/nar/gkx1037.

Appendix



Appendix Figure 1: Displays the results of the star alignment of the 9 rna-seq non-control samples from toxgroup 2 during data curation.



Appendix Figure 2: Shows the original toxgroup 2 rna-seq read files that contained a large number of overrepresented sequences, likely due to some sort of bias during data collection.

Appendix Table 1: List of common pathways enriched for each of the MOA chemical groups that are shared by both RNA-seq and microarray platforms. Pathways highlighted in bold were also enriched in the study by Wang et al.

AhR (11)
rno01100:Metabolic pathways
rno00983:Drug metabolism - other enzymes
rno05207:Chemical carcinogenesis - receptor activation
R-RNO-211859~Biological oxidations
WP1302~Estrogen metabolism
rno00040:Pentose and glucuronate interconversions
R-RNO-156580~Phase II - Conjugation of compounds
rno00480:Glutathione metabolism
rno00140:Steroid hormone biosynthesis

rno00830:Retinol metabolism
WP1286~Metapathway biotransformation
CAR/PXR (12)
R-RNO-449147~Signaling by Interleukins
R-RNO-8957322~Metabolism of steroids
rno00983:Drug metabolism - other enzymes
rno00140:Steroid hormone biosynthesis
R-RNO-556833~Metabolism of lipids
rno00982:Drug metabolism - cytochrome P450
rno05204:Chemical carcinogenesis - DNA adducts
rno04710:Circadian rhythm
R-RNO-383280~Nuclear Receptor transcription pathway
rno04976:Bile secretion
rno05207:Chemical carcinogenesis - receptor activation
rno04152:AMPK signaling pathway
Cytotoxic (75)
R-RNO-8868773~rRNA processing in the nucleus and cytosol
rno05160:Hepatitis C
rno04110:Cell cycle
rno04115:p53 signaling pathway
rno05217:Basal cell carcinoma
rno04390:Hippo signaling pathway
rno01100:Metabolic pathways
R-RNO-73863~RNA Polymerase I Transcription Termination
rno05169:Epstein-Barr virus infection
rno05161:Hepatitis B
rno05203:Viral carcinogenesis
rno04015:Rap1 signaling pathway
rno05165:Human papillomavirus infection
rno03020:RNA polymerase
rno05200:Pathways in cancer

WP160~Glycogen metabolism
rno05226:Gastric cancer
R-RNO-69273~Cyclin A/B1/B2 associated events during G2/M transition
rno05213:Endometrial cancer
rno04010:MAPK signaling pathway
rno04152:AMPK signaling pathway
rno04071:Sphingolipid signaling pathway
R-RNO-8953854~Metabolism of RNA
rno01240:Biosynthesis of cofactors
rno05223:Non-small cell lung cancer
rno05410:Hypertrophic cardiomyopathy
rno05222:Small cell lung cancer
rno04914:Progesterone-mediated oocyte maturation
rno05166:Human T-cell leukemia virus 1 infection
R-RNO-73762~RNA Polymerase I Transcription Initiation
rno05142:Chagas disease
rno00983:Drug metabolism - other enzymes
rno00051:Fructose and mannose metabolism
rno04068:FoxO signaling pathway
R-RNO-73887~Death Receptor Signalling
rno05132:Salmonella infection
R-RNO-6804114~TP53 Regulates Transcription of Genes Involved in G2 Cell Cycle Arrest
rno04933:AGE-RAGE signaling pathway in diabetic complications
rno05225:Hepatocellular carcinoma
rno05145:Toxoplasmosis
rno05212:Pancreatic cancer
R-RNO-6804756~Regulation of TP53 Activity through Phosphorylation
R-RNO-3108232~SUMO E3 ligases SUMOylate target proteins
R-RNO-176408~Regulation of APC/C activators between G1/S and early anaphase
rno00600:Sphingolipid metabolism
rno04530:Tight junction

rno00040:Pentose and glucuronate interconversions
WP1290~Apoptosis
rno05210:Colorectal cancer
R-RNO-212165~Epigenetic regulation of gene expression
rno04114:Oocyte meiosis
R-RNO-2990846~SUMOylation
R-RNO-5250913~Positive epigenetic regulation of rRNA expression
R-RNO-5250924~B-WICH complex positively regulates rRNA expression
rno04550:Signaling pathways regulating pluripotency of stem cells
R-RNO-3700989~Transcriptional Regulation by TP53
rno04218:Cellular senescence
rno04976:Bile secretion
R-RNO-5633007~Regulation of TP53 Activity
rno04922:Glucagon signaling pathway
R-RNO-5656169~Termination of translesion DNA synthesis
rno01523:Antifolate resistance
rno05216:Thyroid cancer
R-RNO-110313~Translesion synthesis by Y family DNA polymerases bypasses lesions on DNA template
R-RNO-174048~APC/C:Cdc20 mediated degradation of Cyclin B
R-RNO-168928~DDX58/IFIH1-mediated induction of interferon-alpha/beta
rno04210:Apoptosis
R-RNO-69278~Cell Cycle, Mitotic
R-RNO-109581~Apoptosis
R-RNO-5357801~Programmed Cell Death