Reanalysis of Wang et. al.’s “The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance”

B. Cole*, T. Falk*, M. Knox*, S. Pandit*
*Bioinformatics Program, Boston University, Boston MA

Introduction

We sought to perform a partial reanalysis of Wang et. al.’s “The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance”, a study comparing RNA-seq and microarray technologies on brown rats (*Rattus norvegicus*) dosed with various toxins (1,2). While both RNA-seq and microarray technology seek to quantify gene expression, the two have not been compared using a range of chemical treatment conditions in an organism.

Comparing the two technologies is crucial due to the fact that both are widely used in the medical and scientific fields to quantify gene expression information, whether to provide medical diagnoses, support drug development, or discover novel transcripts [16]. Quantifying and understanding the limits as well as the advantages of the two technologies allows for researchers and clinicians to best apply the correct technology.

Data

While the study we were trying to replicate considered both microarray and RNA-seq data, we only sourced and processed the data for the RNA-seq portion of the experiment. Further, we selected only toxgroup number two as the focus of our replication. This toxgroup represents three rats (*Rattus norvegicus*) dosed with beta-napthoflavone, three with econazole, three with thioacetamide, and nine controls that were not dosed with any additional chemicals. All RNA-seq samples were sequences on the Illumina HiScan SQ platform [2]. The sequenced samples were aligned to the *Rattus norvegicus* RGSC3.4.66 EnsEMBL genome. For toxgroup 2, all reads were sequenced to a length of 101 bp, and the average library size was 21.3 million paired end reads, ranging from 19 million to almost 25 million reads.

After performing quality control with FastQC and combining the results using MultiQC, we found that all samples were of high enough quality to proceed with the analysis recreation [10, 11]. The samples passed the default FastQC metrics on a number of measures, such as the per sequence quality scores and per base N content. Some metrics were less certain, and likely failed due to the RNA nature of this sequencing. For instance, Figure 1 displays the mean quality scores for all 18 runs. While only a couple of runs were above

the passing level, we felt that a phred score higher than 30 for more than 80% of the read was a suitable level for our analysis.

We also had a failure for all samples in the per base sequence content metric, Figure 2, in which the first 12 base pairs showed significant disruption. This is consistent with the other RNA-seq data we have studied, and likely results from a “biased selection of random primers”, but is not thought to impact later expression analyses [12]. No samples were found to be of low enough quality to exclude, and metrics for the STAR alignment (Supplemental Figure 1) appear to be consistent amongst samples as well.

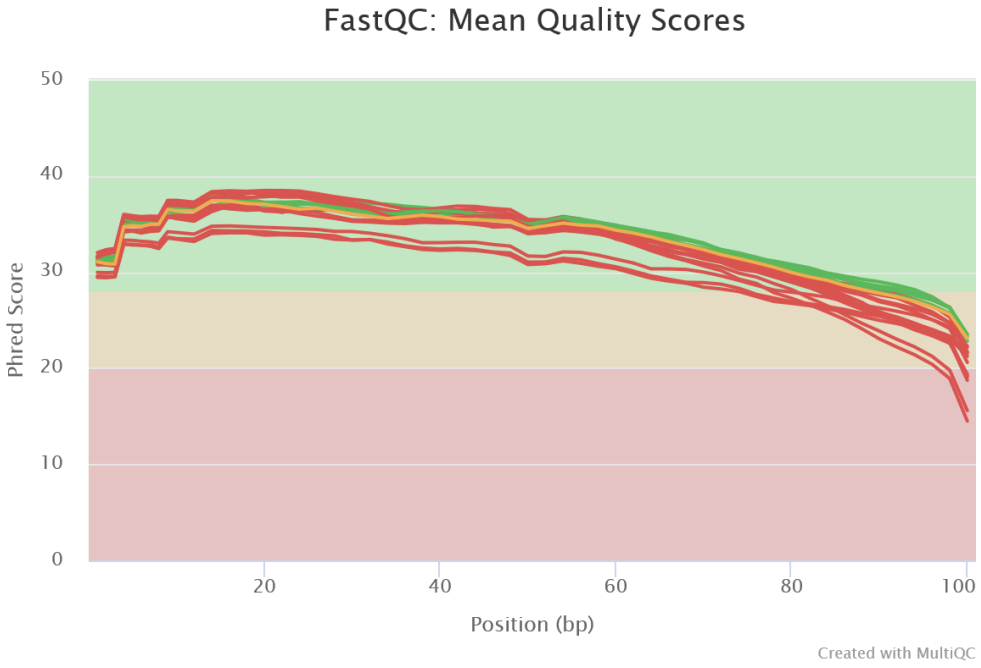


Figure 1 The mean quality scores for each of the 9 paired samples. While the lines in red denote samples that failed under the default FastQC parameters, many of the samples were close to or above a phred score of 30 until the 80 bp mark (3).

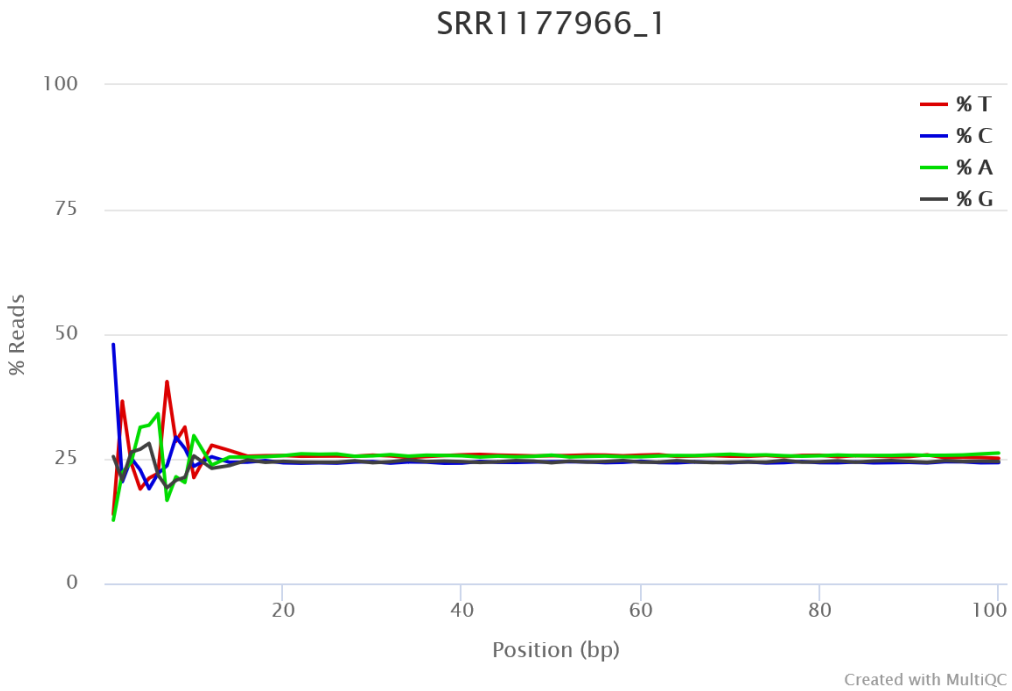


Figure 2 The per base sequence content for sample SRR1177966_1. All other samples exhibited a similar pattern and failed for the same reasons.

Methods

In order to initially process the data, we first retrieved the 18 paired end fastq files for our 9 samples, and processed them using fastqc to determine the sequencing statistics for each sample [10]. We then aligned both files using the STAR RNA-seq aligner, producing 9 bam files used in the later analysis [13]. Finally, we used MultiQC to examine summary statistics for both the STAR alignment and to concatenate the information from the FastQC process [11].

First, we took a closer look at the counts data via MultiQC [3]. FeatureCounts, contained within MultiQC assisted with summarizing the gene counts data combing annotated file *.gtf and STAR aligned *.bam files [4]. Here we see the reads range from just under 40 million to about 52 million per run with about 60% of reads being assigned/mapped (figures 3 & 4, respectively).

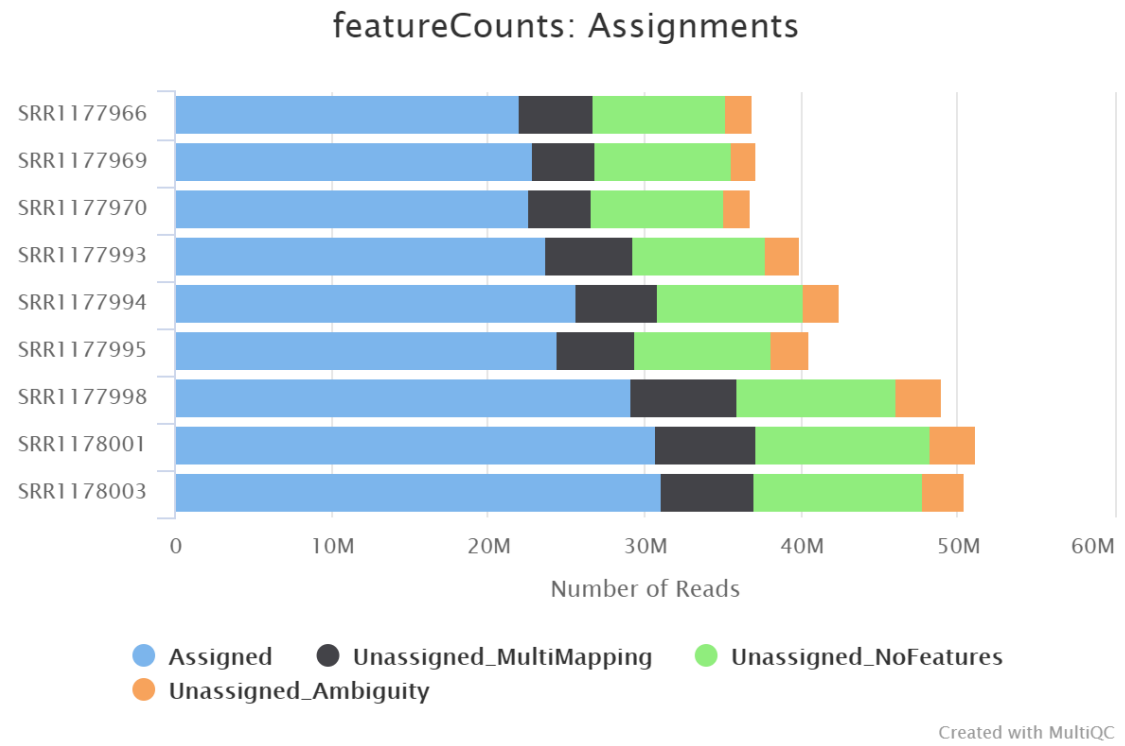


Figure 3 featuresCounts: Number of reads per run.

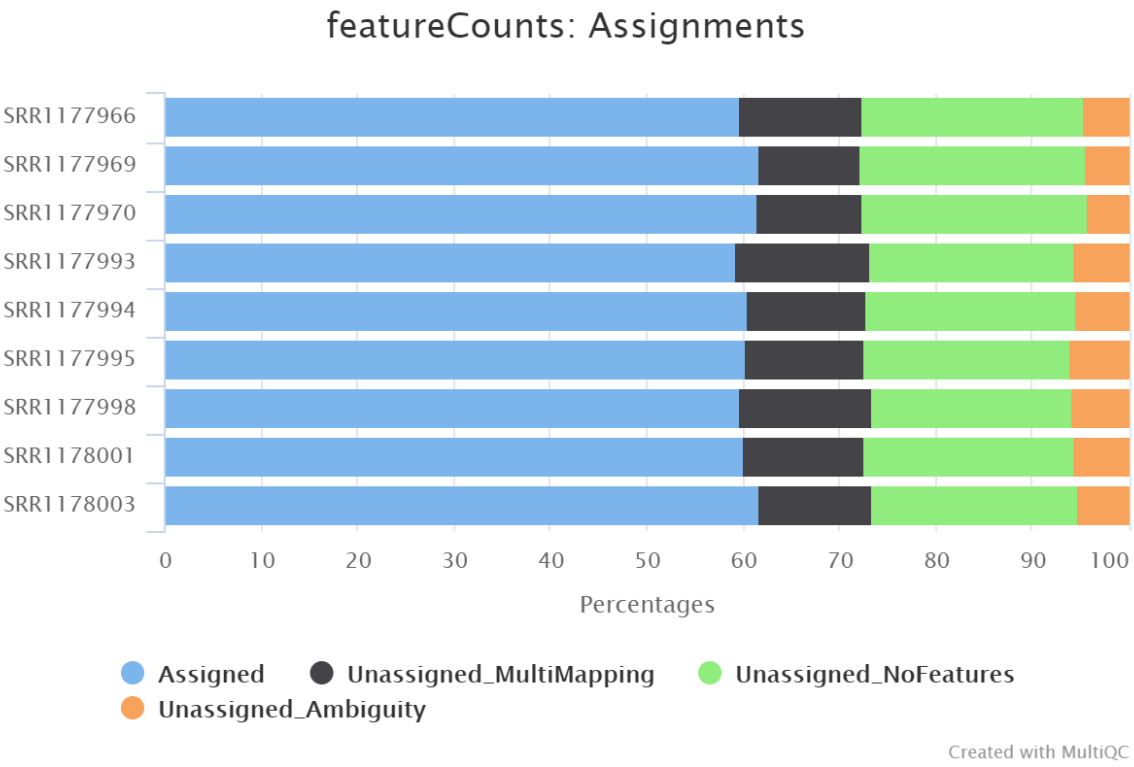


Figure 4 Percent Assigned reads (featureCounts)

We begin with 18,014 genes (observations) then we remove outliers (genes without counts or all zero rows) to get 11,380 genes, roughly 63% as observed in the featuresCount analysis.

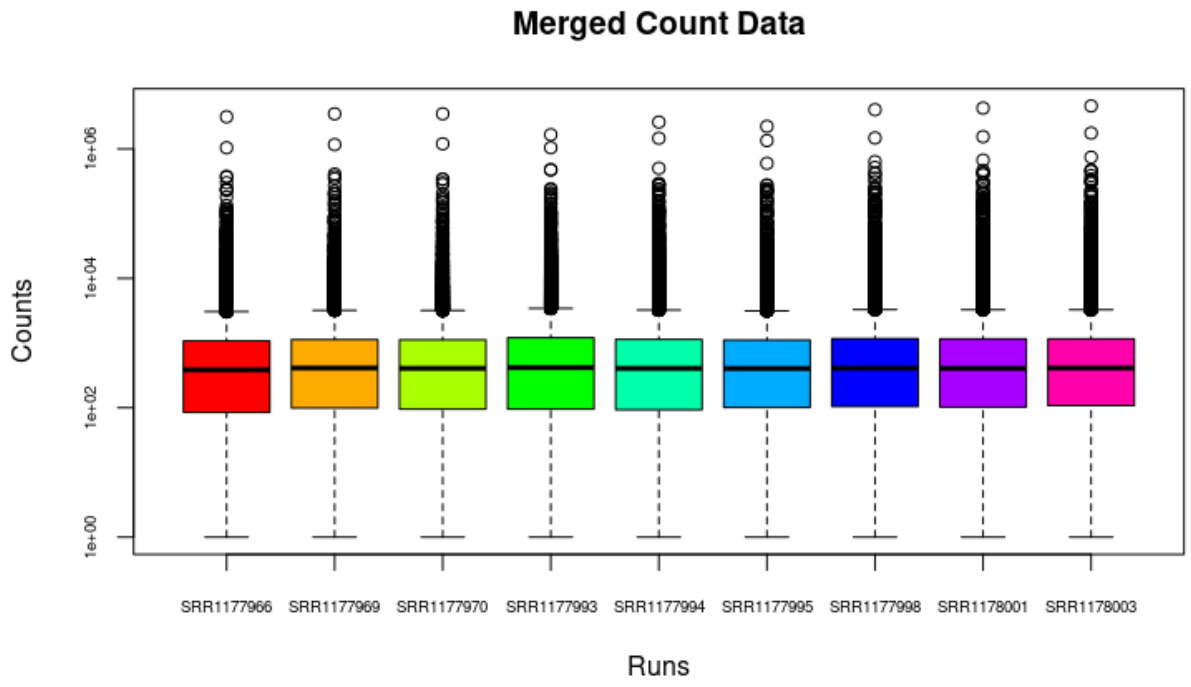


Figure 5 Merged Count Data (boxplot)

We then merge and analyze the counts data which is depicted here in Figure 5. This histogram shows the relative count distribution across the nine runs. There is no apparent evidence of wide disparity by viewing the histogram, as the runs have an almost even relative distribution in normalized, log-scale (y). However, when investigating the data item-wise we see the lowest counts for a given observation (gene) is 8,927 and the highest is 4,612,568.

By employing Bioconductor's **DESeq2** package, we were able to glean more information on differential expression of RNASeq data as this package accounts for both the library depth and varying gene expression to normalize data. In addition, it also offers a way to account for non-zero outliers through log-scaling [5,6]. DESeq2 uses a negative binomial distribution model.

There were just three steps to installing the DESeq2 software: install BiocManager on R, install DESeq2 using BiocManager, and finally load DESeq2. RStudio was utilized to run the DESeq2 package. Once all the packages were installed and libraries loaded, the run took under 10 minutes.

After installation, we had to prepare the files that would go through the DESeq analysis. The metadata, control counts, and counts files were all participants. First, we joined the control counts with the merged counts files by GeneID. This created a subset of 10,753 observations

(genes). Then we further subset by mode of action selecting the runs which correspond to our toxicology treatment (meta data): AHR, Cytotoxic or CAR/PXR.

Next, we created a DESeq object, releveled by mode of action and finally ran our DESeq analysis.

> *DESeq object:* Where we store and identify objects used in DESeq analysis. Here we specified counts, metadata and design (mode of action) for each treatment.
>
> *Relevel:* Subsetting analysis based on specific mode of action
>
> *DESeq Analysis:* Normalization to identify differentially expressed genes of significance.

Alongside the RNA-Seq data used in the original study, we also analyzed the microarray data for each of the treatment conditions in toxgroup 2. As this data had already been curated for us, we ran a differential expression analysis using the limma package on Bioconductor. These results were compared to those from the RNA-Seq data in our concordance calculations, as discussed below.

Using the normalized counts from the DESeq2 analysis, a clustered heatmap was produced (Figure 9). This normalized expression matrix was clustered by MOA.

Using GATHER [15], significant genes identified through the DESeq2 analysis were clustered into gene ontology annotation groups and compared to the common pathways enriched for MOA chemical groups shared by RNA-seq and microarray results from Wang et. al.

Results

Using DESeq2 and limma, we analyzed the RNA-Seq and microarray data, respectively. There were three different analyses per platform, one for each mode of action-chemical pair: beta-naphthoflavone (NAP) in CMC, econazole (ECO) in Corn Oil, and thioacetamide (THI) in saline.

After running the DESeq analysis, we see that 4,932 genes have significant adjusted p-values of less than 0.05%. Within this analysis 214 were AHR treatment, 1,660 CAR/PXR, and 3,058 were Cytotoxic treated (Table 1).

	FALSE	TRUE
AHR & CMC .5%	10114	214
CAR/PXR & CORN OIL 100%	8881	1660
Cytotoxic & SALINE 100%	7694	3058

Table 1 (adjusted p-value < 0.05%).

Similarly, after running the limma analysis we found 8016 genes have adjusted p-values less than 0.05. Of this, 97 were from NAP, 1,981 were from ECO, and 5,938 were from THI.

	FALSE	TRUE
NAP	31002	97
ECO	29118	1981
THI	25161	5938

Table 2 for limma

In addition, we found the following genes in the DESeq analysis were in the top 10 for adjusted p-value for modes of action AHR, CAR/PXR and Cytotoxic by vehicles CMC .5%, corn oil 100%, and saline 100% respectively. All three MOAs were included in the training set and just CAR/PXR were included in the test set [1]. We also report the top 10 genes by chemical treatment from the limma analysis.

Programmer_HH.R* × top_10_CMC_AHR ×

Filter

	Gene Id	baseMean	log2FoldChange	lfcSE	pvalue	padj
1	NM_012540	4679.64715	8.982004	0.5069209	1.926766e-71	1.989964e-67
2	NM_012541	106467.01253	3.945413	0.2243530	1.716943e-70	8.866291e-67
3	NM_130407	800.44989	3.773379	0.2529250	1.494770e-51	5.145995e-48
4	NM_001109459	881.05679	3.133023	0.3419348	2.522184e-21	6.512279e-18
5	NM_138502	1670.24372	-1.310381	0.1569323	3.442680e-18	7.111199e-15
6	NM_001170562	74.77380	2.277188	0.3090889	1.064144e-14	1.831746e-11
7	NM_001008833	1185.17559	1.372187	0.1960710	1.285094e-13	1.896064e-10
8	NM_001109430	749.74322	-1.756092	0.2672023	2.278556e-12	2.739059e-09
9	NM_001191751	141.25694	-2.107686	0.3211555	2.386864e-12	2.739059e-09
10	NM_033352	62.35996	-3.128492	0.4795576	3.296962e-12	3.405102e-09

Table 3 Top 10 DE Genes for AHR (DESeq Analysis)

ProbeID	logFC	AveExpr	t	P.Value	adj.P.Val	B
1370269_at	7.093900137	6.672310397	30.8872737	9.24E-29	2.88E-24	39.16998186
1387243_at	1.559409241	13.07063518	19.74248187	9.94E-22	1.55E-17	31.56841625
1370613_s_at	0.787144278	12.71755407	11.82665579	2.21E-14	2.29E-10	20.10211874
1368990_at	1.46345707	5.600121386	8.73854838	1.14E-10	8.83E-07	13.21498769
1387759_s_at	0.855201593	11.88477136	8.194806116	5.78E-10	3.60E-06	11.84081571

1387811_at	-0.449123444	11.69881747	-7.066754789	1.87E-08	9.70E-05	8.854768393
1367673_at	0.591638778	10.84719669	6.798838459	4.34E-08	0.000192799	8.122790094
1367856_at	1.224284281	8.965154289	6.645583132	7.04E-08	0.00027353	7.700920642
1387901_at	-0.508053389	8.654721042	-6.43067122	1.39E-07	0.000479788	7.105956735
1388122_at	0.849847678	8.769210242	6.106652199	3.88E-07	0.001207011	6.20287192

Table 4 Top 10 DE genes for NAP (limma)

Programmer_HH.R* | top_10_CMC_AHR | top_10_CORN_CARPXR

Filter

	Gene Id	baseMean	log2FoldChange	lfcSE	pvalue	padj
1	NM_001130558	2605.0176	-8.472085	0.3070718	1.741528e-168	1.835745e-164
2	NM_144743	16922.5800	3.110593	0.1804783	8.669174e-68	4.569088e-64
3	NM_001190380	925.6488	3.326633	0.2049255	2.029357e-60	7.130486e-57
4	NM_001013904	59056.4739	2.334218	0.1493404	2.079510e-56	5.480029e-53
5	NM_013105	132527.0451	4.190685	0.2820420	3.459232e-51	7.292752e-48
6	NM_017272	12407.0179	4.348935	0.2935181	7.234592e-51	1.270997e-47
7	NM_019184	38972.9528	-2.867722	0.1954693	6.305290e-50	9.494866e-47
8	NM_001109459	805.4593	2.675370	0.1928207	7.542513e-45	9.938204e-42
9	NM_001108565	577.6315	2.677897	0.1992887	2.757688e-42	3.229866e-39
10	NM_013141	1012.1496	-2.759778	0.2109249	2.186719e-40	2.305020e-37

Table 5 Top 10 DE genes for CAR-PXR (DESeq Analysis)

ProbeID	logFC	AveExpr	t	P.Value	adj.P.Val	B
1384408_at	1.330039703	6.569188694	9.147779962	3.31E-12	6.61E-08	17.36083892
1378027_at	-0.590386151	9.780591777	-8.960161914	6.31E-12	6.61E-08	16.75961935
1371781_at	0.683638366	8.649026102	8.95700576	6.38E-12	6.61E-08	16.74946905
1369012_at	-0.75797501	6.217807757	-8.812668085	1.05E-11	8.17E-08	16.2840174
1388874_at	0.641016406	10.81152546	8.70955398	1.50E-11	9.35E-08	15.95002582
1395403_at	-2.949794015	9.639710504	-8.459180572	3.60E-11	1.87E-07	15.1341649
1370698_at	1.222950441	11.6793762	8.30343516	6.22E-11	2.76E-07	14.62334461
1391570_at	1.151515431	5.853130911	7.945870044	2.19E-10	8.53E-07	13.44186909
1380805_at	-0.520333952	4.010581349	-7.777577931	3.98E-10	1.38E-06	12.88202331
1386944_a_at	-1.331333548	11.94688873	-7.707798447	5.10E-10	1.45E-06	12.64926286

Table 6 Top 10 DE genes for ECO (limma)

Programmer_HH.R* | top_10_CMC_AHR | top_10_CORN_CARPXR | top_10_SALINE_Cytotoxic

Filter

	X	baseMean	log2FoldChange	lfcSE	pvalue	padj
1	NM_001008363	4054.7655	4.334019	0.1908093	2.065721e-115	2.221064e-111
2	NM_031642	4283.5969	3.814636	0.2136649	2.169768e-72	1.166467e-68
3	NM_001109260	561.5203	5.315903	0.2984479	8.048981e-72	2.884755e-68
4	NM_001130573	1055.9723	6.536379	0.3743071	2.477193e-69	6.658694e-66
5	NM_031821	5213.4418	2.613082	0.1562935	6.756622e-64	1.452944e-60
6	NM_012623	4726.9722	6.615227	0.4025631	3.901035e-62	6.990654e-59
7	NM_001130500	412.4234	4.540378	0.2828652	1.711302e-59	2.628560e-56
8	NM_001108099	1793.5187	2.339376	0.1466959	2.170377e-58	2.916986e-55
9	NM_001039344	551.8756	3.771658	0.2420643	6.811214e-56	8.137131e-53
10	NM_012923	3556.3030	2.763886	0.1796018	1.189385e-54	1.278827e-51

Table 7 Top 10 DE genes Cytotoxic (DESeq Analysis)

ProbeID	logFC	AveExpr	t	P.Value	adj.P.Val	B
1373767_at	4.31261533	6.70131987	37.27956534	3.08E-28	9.59E-24	51.61957255
1370177_at	4.567961156	5.750820293	31.73207348	5.05E-26	7.85E-22	47.51688409
1380229_at	4.293741567	4.09943149	29.63742445	4.32E-25	4.47E-21	45.71025697
1369519_at	2.510462863	4.22293039	23.96611209	3.19E-22	2.48E-18	39.90647136
1385706_at	2.189917481	3.929071367	23.76161231	4.15E-22	2.58E-18	39.66796277
1367764_at	2.91006527	7.901032257	23.22208586	8.40E-22	4.36E-18	39.02755028
1396236_at	2.320035859	3.73673026	22.57064643	2.01E-21	8.93E-18	38.23204284
1368072_at	3.465269504	6.341493113	21.59620131	7.73E-21	3.00E-17	36.99436265
1397564_at	2.109776226	4.158714463	21.22329919	1.31E-20	4.53E-17	36.50487881
1387269_s_at	2.406844967	4.348370363	20.30026715	5.04E-20	1.57E-16	35.25349674

Table 8 Top 10 DE genes for THI (limma)

The most significant genes based on adjusted p-values for each treatment were as follows:

- NM_012540, adjusted p-value of $1.99e\text{-}^{67}$ AHR MOA by CMC by .5% CMC (Table 3).
- NM_001130558, adjusted p-value of $1.8e\text{-}^{164}$ CAR/PXR MOA by 100% corn oil (Table 5)
- NM_001008386, adjusted p-value of $2.22e\text{-}^{111}$ Cytotoxic MOA by 100% saline (Table 7)

NM_012540 rat (CYP1A1 human) gene is responsible for protein transcription specific to conversion of vitamins, fatty acids and select steroids [7]. Expression is mostly in the liver, blood and gallbladder among a couple others.

NM_01130558 rat (STAC3 human) gene is involved in protein transcription specific to contraction in muscles [8].

NM_001008386 rat (CRPPA human) gene is also involved in protein transcription specific to catalization in sugars, yeast and alcohol compounds [9]. Expression is systemic.

Next, we reviewed the fold change within the DESeq results (Figure 6) and limma results (Figure 8). Fold change (absolute log2) measure of when genes change and correlates with significant differences in gene expression [10]. Further, we correlate this measure with normalized p-value (Figure 7). There does not seem to be much correlation between these two measures. These results also hold true when looking at the results from the limma analysis.

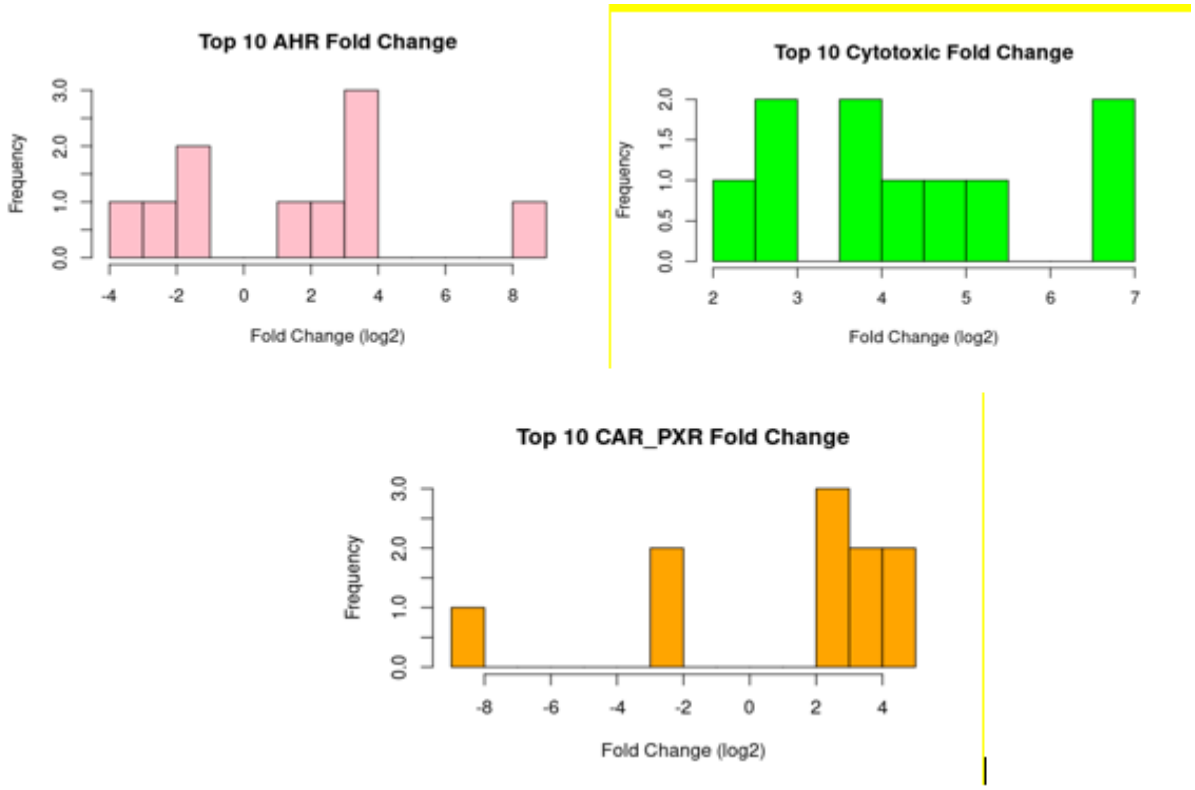


Figure 6 Fold Change (DESeq Anaysis)

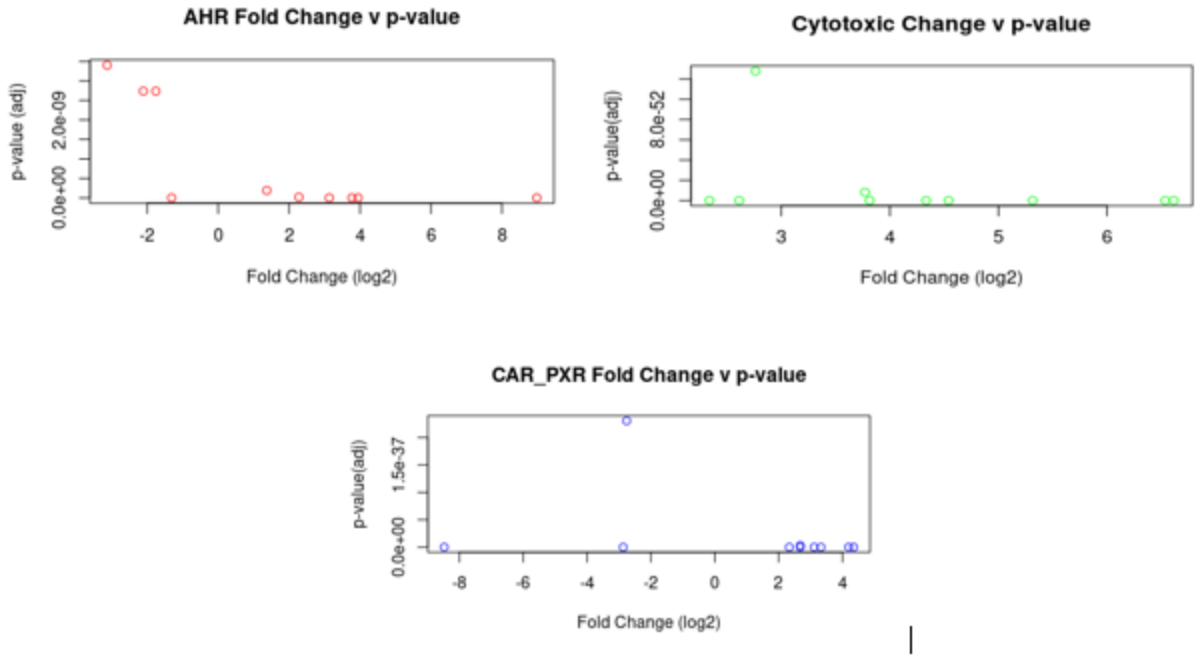
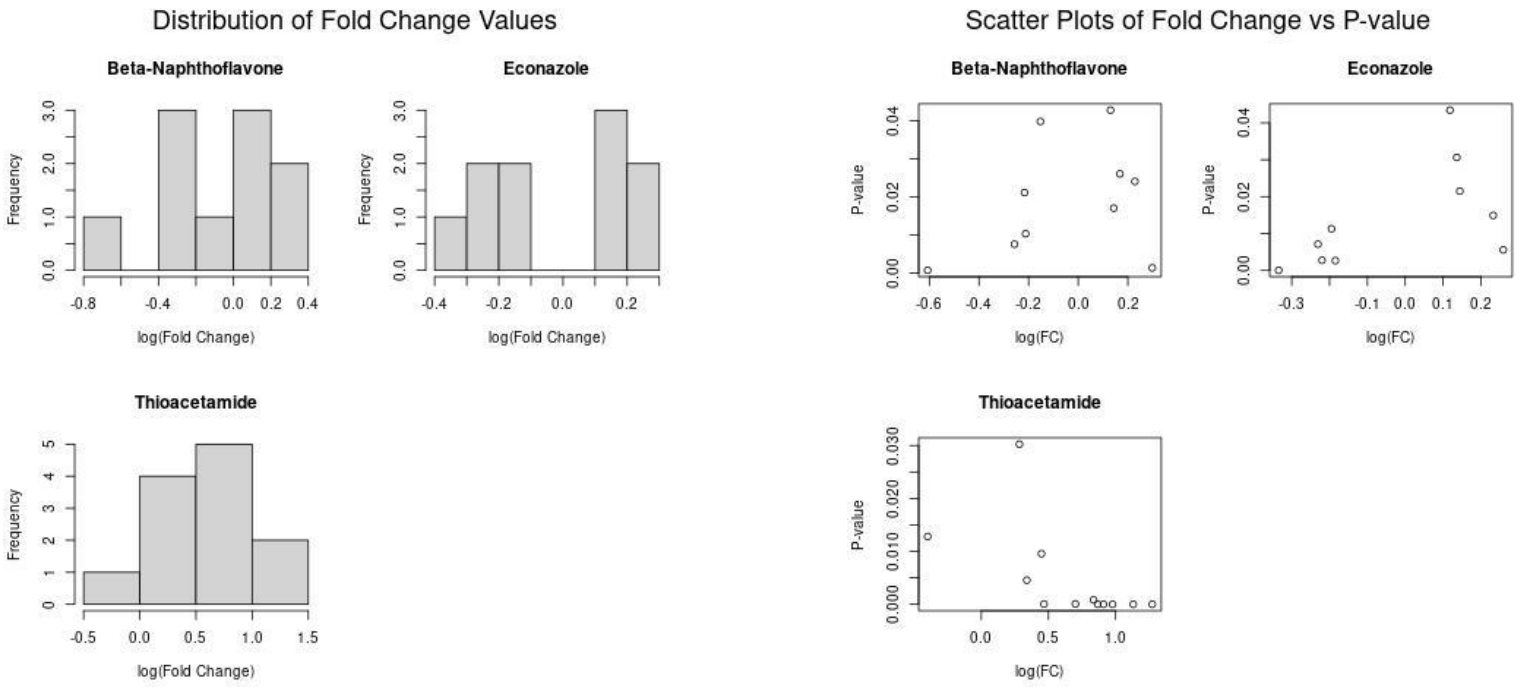


Figure 7 Fold change vs. p-value

Figure 8 Distribution of fold change values and scatter plots of fold change vs p-value (limma)



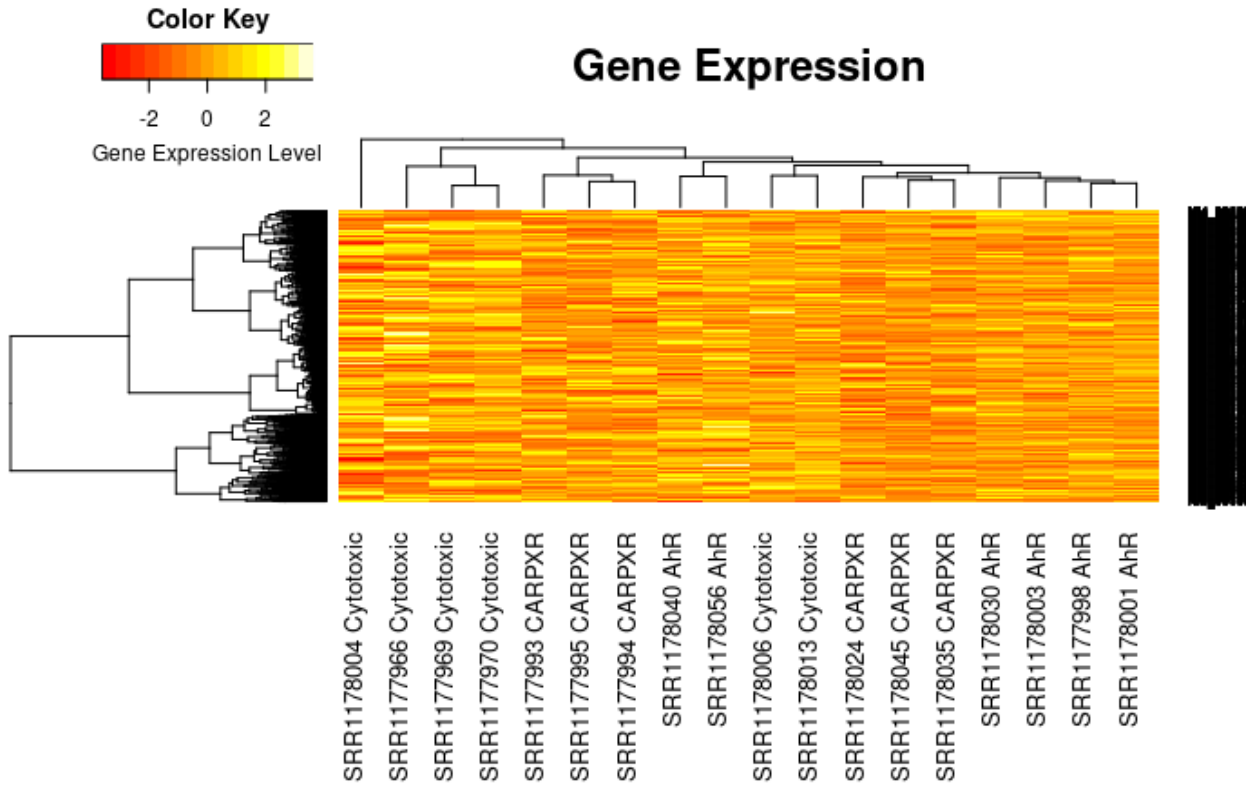


Figure 9 Clustered heatmap of counts from DESeq2 analysis.

From the DESeq2 analysis, a heatmap (Figure 9) clustered by MOA was produced. There is no obvious trend in expression level between clusters of genes and MOA clusters.

Significant genes from each MOA were identified and clustered into enrichment pathways using the online tool GATHER. Tables 9, 10, and 11 compare the results of the GATHER analysis for each MOA with the results from Wang et. al. Of the top 10 GO terms for each MOA, only 2 terms in the total 30 were found to overlap with the original paper's results. The original paper saw common pathways mostly in degradation for all 3 MOAs, while the GATHER results we generated saw no terms of degradation, even beyond the top 10 terms for each MOA.

Cytotoxic

GO Annotation	Total Genes With Ann	Your Genes (With Ann)	Your Genes (No Ann)	Genome (With Ann)	Genome (No Ann)	ln(Bayes factor)	neg ln(p value)	FE: neg ln(p value)	FE: neg ln(FDR)	Genes	Similar term to Wang et al.
GO:0001558 [4]: regulation of cell growth	9	9	150	49	4188	3.69	7.18	8.54	2.11	Alox15 Cgref1 Dtr Esm1 Igfbp1 Igfbp6 Pla2g2a Pttg1 S100a10	
GO:0016049 [4]: cell growth	9	9	150	59	4178	2.5	5.92	7.31	2.11	Alox15 Cgref1 Dtr Esm1 Igfbp1 Igfbp6 Pla2g2a Pttg1 S100a10	
GO:0008361 [5]: regulation of cell size	9	9	150	60	4177	2.4	5.8	7.2	2.11	Alox15 Cgref1 Dtr Esm1 Igfbp1 Igfbp6 Pla2g2a Pttg1 S100a10	
GO:0019748 [4]: secondary metabolism	5	5	154	16	4221	2.27	5.55	7.21	2.11	Akr7a3 Alas2 Aldh1a1 Hmgcr Hmox1	
GO:0007267 [4]: cell-cell signaling	1	1	158	246	3991	2.05	5.32	0	0	Nr4a2	
GO:0050877 [4]: neurophysiological process	2	2	157	295	3942	2.03	5.32	0	0	Nr4a2 Olr59	
GO:0040008 [4]: regulation of growth	9	9	150	65	4172	1.91	5.22	6.69	1.95	Alox15 Cgref1 Dtr Esm1 Igfbp1 Igfbp6 Pla2g2a Pttg1 S100a10	
GO:0001836 [9]: release of cytochrome c from mitochondria	2	2	157	0	4237	1.61	4.91	6.65	1.95	Bax Myc	
GO:0044255 [5]: cellular lipid metabolism	17	17	142	205	4032	1.46	4.74	6.01	1.76	Acaca Aldh1a1 Alox15 Baat Fabp2 Fads1 Galr2 Hmgcr Hmox1 Hsd11b1 Hsd17b2 Hsd17b9 Hsd3b Pc Sah Scd1 Ste	
GO:0046907 [5]: intracellular transport	1	1	158	227	4010	1.39	4.65	0	0	Arl4	

Table 9 Top 10 GATHER GO terms for the MOA “Cytotoxic”.

CARPXR

GO Annotation	Total Genes With Ann	Your Genes (With Ann)	Your Genes (No Ann)	Genome (With Ann)	Genome (No Ann)	ln(Bayes factor)	neg ln(p value)	FE: neg ln(p value)	FE: neg ln(FDR)	Genes	Similar term to Wang et al.
GO:0046483 [5]: heterocycle metabolism	5	5	76	42	4273	2.2	5.47	6.46	1.73	Alas2 Cyp1a1 Egln3 Hmox1 Ugt1a1	
GO:0006778 [6]: porphyrin metabolism	3	3	78	11	4304	1.93	5.22	6.27	1.73	Alas2 Hmox1 Ugt1a1	
GO:0006805 [5]: xenobiotic metabolism	4	4	77	26	4289	1.9	5.17	6.2	1.73	Cyp1a1 Cyp2c Hdc Lbp	Xenobiotic metabolism signaling
GO:0015669 [5]: gas transport	2	2	79	2	4313	1.86	5.17	6.23	1.73	Hba-a1 Hbb	
GO:0015671 [6]: oxygen transport	2	2	79	2	4313	1.86	5.17	6.23	1.73	Hba-a1 Hbb	
GO:0009410 [7]: response to xenobiotic stimulus	4	4	77	27	4288	1.78	5.06	6.07	1.73	Cyp1a1 Cyp2c Hdc Lbp	
GO:0040007 [3]: growth	7	7	74	101	4214	1.51	4.78	5.66	1.7	Alox15 Cdh13 Esr1 Gpam Igfbp2 Inhba Socs3	
GO:0006787 [7]: porphyrin catabolism	2	2	79	3	4312	1.36	4.65	5.73	1.7	Hmox1 Ugt1a1	
GO:0044255 [5]: cellular lipid metabolism	10	10	71	212	4103	0.95	4.25	4.95	1.7	Aldh1a1 Alox15 Gpam Hmgcs1 Hmox1 Hsd17b2 Hsd17b9 LOC246252 Prlr Scd1	
GO:0005977 [8]: glycogen metabolism	3	3	78	17	4298	0.9	4.21	5.21	1.7	G6pc Gaa Ppp1r3b	

Table 10 Top 10 GATHER GO terms for the MOA “CAR/PXR”.

AhR

GO Annotation	Total Genes With Ann	Your Genes (With Ann)	Your Genes (No Ann)	Genome (With Ann)	Genome (No Ann)	ln(Bayes factor)	neg ln(p value)	FE: neg ln(p value)	FE: neg ln(FDR)	Genes	Similar term to Wang et al.
GO:0006805 [5]: xenobiotic metabolism	2	2	7	28	4359	4.17	7.91	6.45	3.65	Cyp1a1 Cyp1a2	Xenobiotic metabolism signaling
GO:0009410 [7]: response to xenobiotic stimulus	2	2	7	29	4358	4.11	7.91	6.39	3.65	Cyp1a1 Cyp1a2	
GO:0018894 [6]: dibenzo-p-dioxin metabolism	1	1	8	0	4387	3.89	7.45	6.19	3.65	Cyp1a1	
GO:0018904 [6]: organic ether metabolism	1	1	8	0	4387	3.89	7.45	6.19	3.65	Cyp1a1	
GO:0006118 [6]: electron transport	3	3	6	164	4223	3.33	6.77	5.57	3.25	Cyp1a1 Cyp1a2 Cyp4a14	
GO:0015909 [7]: long-chain fatty acid transport	1	1	8	3	4384	2.51	5.92	4.81	2.67	Cpt1b	
GO:0044237 [4]: cellular metabolism	8	8	1	2057	2330	2.19	5.47	4.39	2.55	Abcd2 Cpt1b Cyp1a1 Cyp1a2 Cyp4a14 Hspb1 Mme Usp2	
GO:0015908 [6]: fatty acid transport	1	1	8	5	4382	2.11	5.38	4.4	2.55	Cpt1b	
GO:0006091 [5]: generation of precursor metabolites and energy	3	3	6	272	4115	1.98	5.27	4.18	2.55	Cyp1a1 Cyp1a2 Cyp4a14	
GO:0006631 [6]: fatty acid metabolism	2	2	7	92	4295	1.96	5.27	4.22	2.55	Abcd2 Cpt1b	

Table 11 Top 10 GATHER GO terms for the MOA “CAR/PXR”.

Discussion

Once we had the data from the DESeq2 and limma analyses, we were able to compute the concordance. The authors defined concordance as the number of differentially expressed genes that were shared by both platforms, RNA-Seq and microarray, divided by the total set of genes.

Because these platforms use different methods of identifying the genes (affymetrix identifiers versus RefSeq identifiers), they had to be translated into a common language so that they could be compared. We were able to accomplish this by using a table that had mappings of probe IDs to gene names, as well as Refseq IDs to gene names. By scanning these tables and pairing them to the appropriate identifiers, we were able to get the names of the differentially expressed genes (DEGs) from both.

The first analysis that we sought to replicate from the original paper was the concordance across different treatment groups. For toxgroup 2, we had to analyze the effects that beta-naphthoflavone (NAP), econazole (ECO), and thioacetamide (THI) had on gene expression, and therefore on concordance. Generally, our results agree with those of the paper. We found that as the treatment effect of a chemical increased, its concordance did as well. This is illustrated in Figure 10, which plots the concordance of our three chemicals versus the number of DEGs, which serves as a barometer for the magnitude of the treatment effect. This figure also shows the concordance against the number of DEGs identified by each platform. Because these results are roughly consistent with each other (in that the points follow roughly the same pattern) it further validates the phenomenon observed in the paper.

However, there are several key areas where our results deviate from those in the paper. First, the number of differentially expressed genes between each platform is significantly different from that observed in the paper. This can be noted by the discrepancy in the range of the x-axis in our Figure 11 versus that of the paper's Figure 2a. Further, the concordance estimates for some of the chemicals are significantly lower than those attained by the paper. This is especially true for those chemicals with higher treatment effects, such as THI and ECO, while those with lower effects, such as NAP, are closer in magnitude to those in the paper.

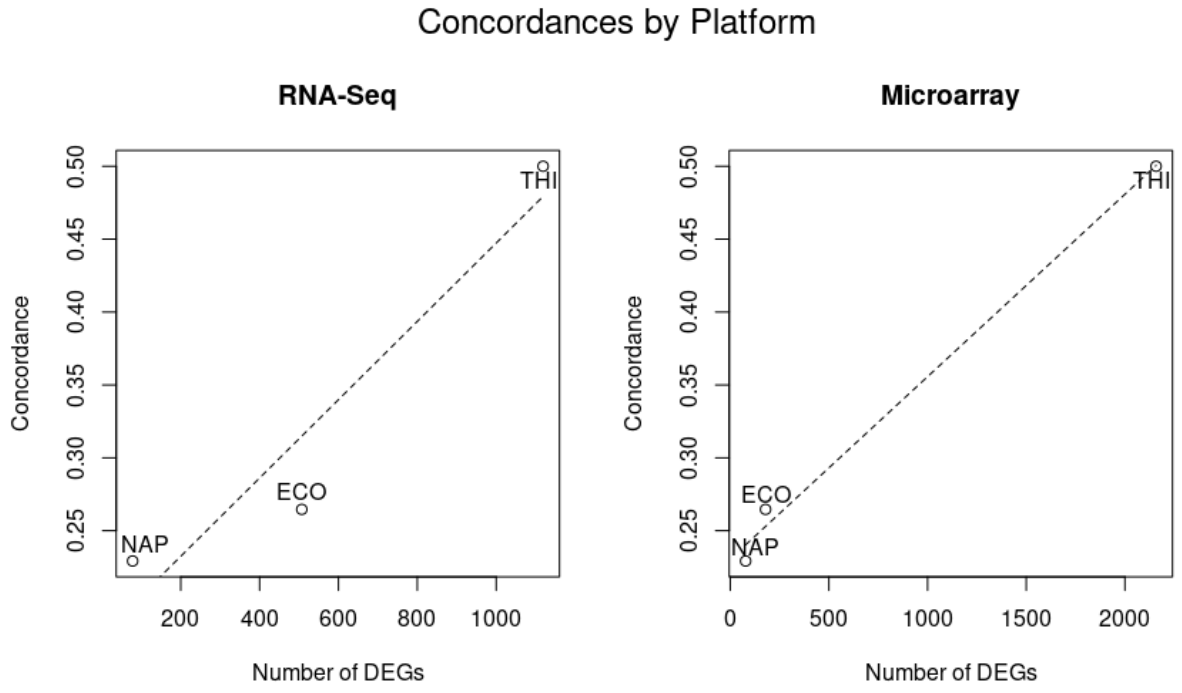


Figure 10 Concordance by platform. Number of DEGs identified by the platform is on the x axis, and the concordance is on the y axis.

In the original study, the degree of concordance was not only measured across platforms, but also across gene expression levels. This analysis was performed by splitting the differential expression data into high and low values, be they from DESeq2 or limma, and to see how the concordance is affected. The data was split into high and low values based on whether the average expression level of a gene was above or below the gene with the median average expression level.

The pathways enriched for each MOA chemical group were also very different in our results from the results of the original paper (Tables 9, 10, 11). Degradation pathways (including nicotine degradation, melatonin degradation, and bupropion degradation) were often listed in the common enriched terms from Wang et. al., but were not in the results produced using GATHER. This could be downstream effects of the difference in differentially expressed genes, where these more common terms might have been lost to. Only one term (Xenobiotic metabolism signaling in Ahr and CAR/PXR) was found to be common between our top 10 results and the common results of Wang et. a.

When the authors of the paper conducted this analysis, they found that the concordance for the genes with high average expression was highest, followed by the concordance of all genes, followed by the concordance for the genes with low expression. We had mixed results from our attempt to replicate this, as shown in Figure 11. While this pattern was observed for NAP and ECO, where the concordance is higher for the highly expressed genes, we did not observe this pattern for THI.

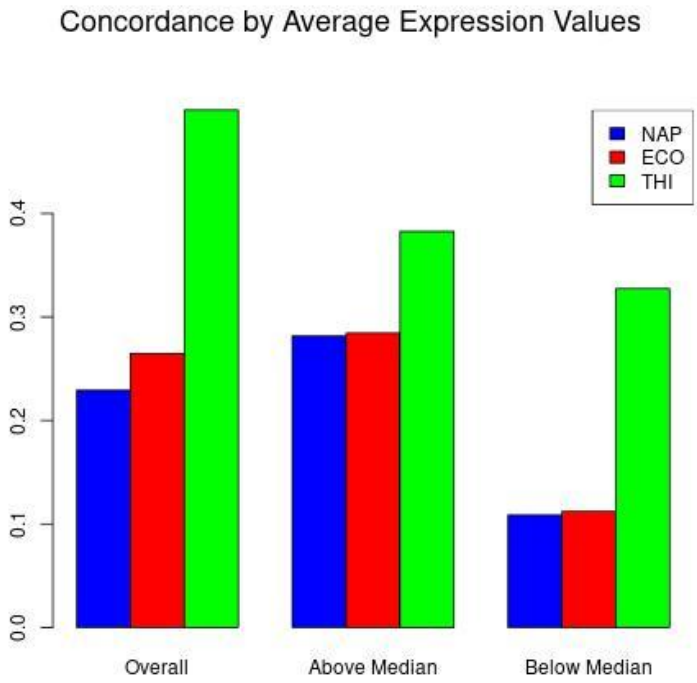


Figure 11 Concordance across gene expression levels. Median was determined by the median average expression level for both DESeq and limma results.

Furthermore, similar to our previous results, we observed concordances lower in magnitude than those in the original study. Because this effect was observed in both of our analyses, we were forced to take a closer look at our methodology, and how it compared with that of the paper. How is it that we can observe the same pattern, but with “dampened” concordances? Looking at the calculation for concordance, one way that the concordance estimate can change is through denominator. The concordance calculation detailed in the Online Methods of the paper states that one of the terms in the denominator, *N*, should be “the set of all genes.” This is a somewhat ambiguous definition, and so because of this we *N* to be 30,000, the total number of genes in the genome. If our *N* is significantly higher than that used by the paper, this would explain why our concordances are significantly lower than those reported in the original study.

All of the scripts used for this analysis can be found on our github:
https://github.com/BF528/project-3-hedgehog

Conclusion

Overall, we were mostly able to replicate the main findings from the paper. We observed that the concordance between RNA-seq and microarray data is proportional to the treatment effect. As the magnitude of the treatment effect increases, as measured by the number of differentially expressed genes, the concordance between the two platforms also increases. The main difference between our results and that of the original paper is in the enriched pathways for significantly differentially expressed genes.

One of the major challenges we faced during this project was concerning use of the differential expression analysis software. This interface was new to us, and we faced a steep learning curve when trying to get the results we needed. This learning curve unfortunately affected downstream analyses, which itself created another obstacle for us to navigate around and made comparison of results difficult. Given an opportunity to learn and apply this analysis software before attempting this recreation of Wang et. al.'s analysis would have been useful. Another challenge that we encountered was in the translation of Affymetrix and RefSeq IDs to genes. Doing this computation efficiently was the most challenging aspect, as the datasets we were working with were relatively large and so computational complexity became an issue that we had to keep in mind while running our scripts.

References

1. Wang, Charles, et al. “A Comprehensive Study Design Reveals Treatment- and Transcript Abundance–Dependent Concordance between RNA-Seq and Microarray Data.” Nature Biotechnology, vol. 32, no. 9, Sept. 2014, pp. 926–32. PubMed Central, doi:10.1038/nbt.3001.

2. GEO Accession Viewer. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55347. Accessed 5 Apr. 2021.

3. 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

4. WEHI Bioinformatics - FeatureCounts. http://bioinf.wehi.edu.au/featureCounts/. Accessed 6 Apr. 2021.

5. 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.

6. StatQuest: DESeq2, Part 1, Library Normalization. www.youtube.com, https://www.youtube.com/watch?v=UFB993xufUU. Accessed 6 Apr. 2021.

7. CYP1A1 Gene - GeneCards | CP1A1 Protein | CP1A1 Antibody. https://www.genecards.org/cgi-bin/carddisp.pl?gene=CYP1A1. Accessed 5 Apr. 2021.

8. STAC3 Gene - GeneCards | STAC3 Protein | STAC3 Antibody. https://www.genecards.org/cgi-bin/carddisp.pl?gene=STAC3#function. Accessed 5 Apr. 2021.

9. CRPPA Gene - GeneCards | ISPD Protein | ISPD Antibody. https://www.genecards.org/cgi-bin/carddisp.pl?gene=CRPPA. Accessed 5 Apr. 2021.

10. “How Many Differentially Expressed Genes: A Perspective from the Comparison of Genotypic and Phenotypic Distances.” Genomics, vol. 110, no. 1, Jan. 2018, pp. 67–73. www.sciencedirect.com, doi:10.1016/j.ygeno.2017.08.007.

11. Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 6 Apr. 2021.

12. Ewels, Philip, et al. “MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report.” *Bioinformatics*, vol. 32, no. 19, Oct. 2016, pp. 3047–48. *DOI.org (Crossref)*, doi:10.1093/bioinformatics/btw354.

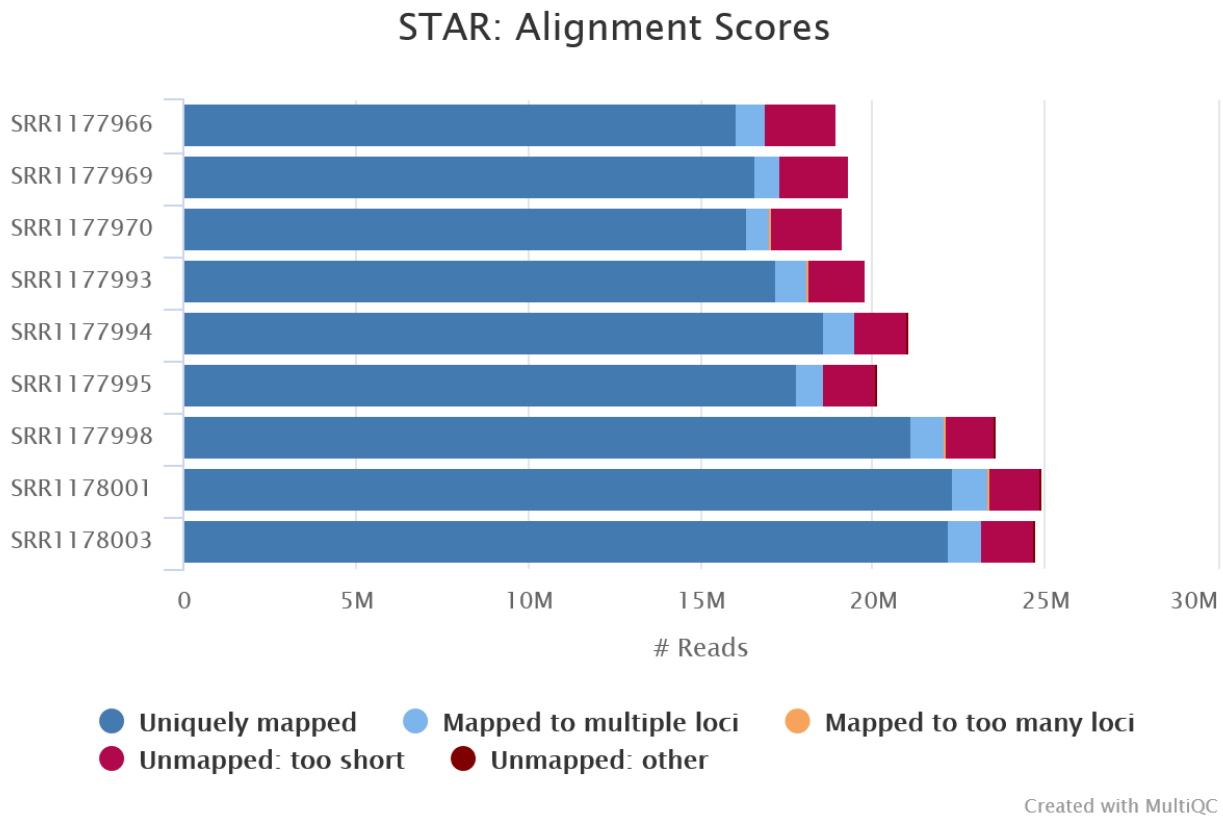
13. *VEuPathDB Workshop*. https://workshop.eupathdb.org/. Accessed 6 Apr. 2021.

14. Dobin, Alexander, et al. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics (Oxford, England)*, vol. 29, no. 1, Jan. 2013, pp. 15–21. *PubMed*, doi:10.1093/bioinformatics/bts635.

15. JT Chang and JR Nevins. "GATHER: A Systems Approach to Interpreting Genomic Signatures." *Bioinformatics* 22(23), 2006.

16. Rna-sequencing vs microarrays. (n.d.). Retrieved April 05, 2021, from https://geneviatechnologies.com/blog/rnasequencing/

Supplemental



Supplemental Figure 1 The alignment scores after our nine samples were mapped to the rat genome using STAR. No notable differences were found between the nine samples.