PROJECT 3 BF528

Concordance of microarray and RNA- Seq differential gene expression

Vishwa Talati (Data Curator), Kyrah Kotary (Programmer), Marina Natividad (Analyst), Brad Fortunato (Biologist)

Introduction:

RNA-seq is a well known method to study and measure gene expression however, its concordance with well established microarray must be accessed for confident uses in clinical and regulatory application. The study design in Wang et al.,[1] investigates and compares these methods by generation of Illumina RNA-seq and Affymetrix microarray data from the same set of liver samples of rats using 27 chemical treatments with multiple modes of action (MOA). RNA- seq has proved better than microarray in DEG verification by quantitative PCR and its improved accuracy for low expressed genes although the predictive classifiers derived from both the studies performed similarly. As a result, the endpoint studied, the biological complexity and transcript abundance determined by the study are important factors in transcriptomic research for decision making. The purpose of our study was to replicate the results of Wang et al. using precomputed data and considering a subset (toxo group) of 3 chemicals namely Econazole, Thioacetamide and Beta Naphthoflavone with 3 different modes of action.The goal was to process the data by aligning short reads to the rat genome, perform differential expression of RNA seq, perform differential expression of pre-normalized microarray expression data and finally map the affymetrix and refseq identifier systems.

Data:

For this project, we chose a subset of samples from toxo group 2 having Thioacetamide, Beta- naphthoflavone, Econazole chemicals with 3 different modes of action namely aryl hydrocarbon receptor (AhR), orphan nuclear hormone receptor(CAR/PXR) and cytotoxicity. This group had 9 samples and 9 controls. Each sample had its chemical, MOA, vehicle and route which is explained in Table 1. The microarray was performed on the Affymetrix whole genome GeneChip Rat Genome 230 2.0 Array and sequencing was performed on Illumina 1.9 using Sanger sequencing. RNA-seq of 63 training and 42 test set samples on Illumina HiScanSQ or HiSeq2000 systems was performed according to the manufacturer's protocol using the Illumina TruSeq RNA Sample Preparation Kit and SBS Kit v3. Depths of ~23 – 25 million paired-end 100 bp reads were generated for each sample. Data was deposited in the Sequence Read Archive (NCBI) under accession number SRP024314 for public use.

Sample	Mode of Action	Chemical	Vehicle	Route
SRR1177998	AhR	Beta- naphthoflavone	CMC_.5_%	Oral_Gavage
SRR1178001	AhR	Beta- naphthoflavone	CMC_.5_%	Oral_Gavage
SRR1178003	AhR	Beta- naphthoflavone	CMC_.5_%	Oral_Gavage
SRR1177993	CAR/PXR	Econazole	CORN_OIL_100_%	Oral_Gavage
SRR1177994	CAR/PXR	Econazole	CORN_OIL_100_%	Oral_Gavage
SRR1177995	CAR/PXR	Econazole	CORN_OIL_100_%	Oral_Gavage
SRR1177966	Cytotoxic	Thioacetamide	SALINE_100_%	Intraperitoneal
SRR1177969	Cytotoxic	Thioacetamide	SALINE_100_%	Intraperitoneal
SRR1177970	Cytotoxic	Thioacetamide	SALINE_100_%	Intraperitoneal
SRR1178030	Control	Vehicle	CMC_.5_%	Oral_Gavage
SRR1178040	Control	Vehicle	CMC_.5_%	Oral_Gavage
SRR1178056	Control	Vehicle	CMC_.5_%	Oral_Gavage
SRR1178024	Control	Vehicle	CORN_OIL_100_%	Oral_Gavage
SRR1178035	Control	Vehicle	CORN_OIL_100_%	Oral_Gavage
SRR1178045	Control	Vehicle	CORN_OIL_100_%	Oral_Gavage
SRR1178004	Control	Vehicle	SALINE_100_%	Intraperitoneal
SRR1178006	Control	Vehicle	SALINE_100_%	Intraperitoneal
SRR1178013	Control	Vehicle	SALINE_100_%	Intraperitoneal

Table 1: Toxo- group 2 information table where Mode of Action includes AhR i.e. aryl hydrocarbon receptor, CAR/PXR i.e. orphan nuclear hormone receptors, cytotoxic i.e. cytotoxicity. There are three different chemicals for samples and vehicles for controls. Vehicle represents the substance used for injection of chemicals and it includes saline_100_%, corn_oil_100_% and CMC_.5_%. The route of administration includes oral and intraperitoneal

For this project, all the datasets were downloaded and made available for us from accession numbers SRP039021, GSE55347, and GSE47875. After the identification of samples based on the toxo- group 2, FastQC was done to process and check the quality of sample files. STAR aligner was used to align the reads against the rat reference genome index. STAR aligner was run on the paired end reads and aligned bam files were generated. MultiQC was run on these bam files along with the fastq files of the samples to get the information of the alignment statistics.

Samples	**% Aligned**	**M Aligned (million)**	**Chemical Treatment**	**GC %**	**Sequence length (bp)**
SRR1177966	84.2	16.0	Thioacetamide		
SRR1177966_1			Thioacetamide	48	101
SRR1177966_2			Thioacetamide	48	101
SRR1177969	85.4	16.5	Thioacetamide		
SRR1177969_1			Thioacetamide	49	101
SRR1177969_2			Thioacetamide	49	101
SRR1177970	85.0	16.3	Thioacetamide		
SRR1177970_1			Thioacetamide	49	101
SRR1177970_2			Thioacetamide	49	101
SRR1177993	86.2	17.1	Econazole		
SRR1177993_1			Econazole	49	101
SRR1177993_2			Econazole	49	101
SRR1177994	88.0	18.5	Econazole		
SRR1177994_1			Econazole	49	101
SRR1177994_2			Econazole	49	101
SRR1177995	87.6	17.7	Econazole		
SRR1177995_1			Econazole	49	101
SRR1177995_2			Econazole	49	101
SRR1177998	88.8	21.0	Beta- Naphthoflavone		
SRR1177998_1			Beta- Naphthoflavone	49	101
SRR1177998_2			Beta- Naphthoflavone	49	101
SRR1178001	89.1	22.2	Beta- Naphthoflavone		
SRR1178001_1			Beta- Naphthoflavone	49	101

SRR1178001_2			Beta- Naphthoflavone	49	101
SRR1178003	89.2	22.0	Beta- Naphthoflavone		
SRR1178003_1			Beta- Naphthoflavone	49	101
SRR1178003_2			Beta- Naphthoflavone	49	101

Table 2: General Statistics for STAR alignment which includes percent of uniquely mapped reads (%Aligned), number of uniquely mapped reads in millions (M Aligned), chemical treatment, Average percentage of GC content (GC%) and Average sequence length in base pairs(sequence length) from Multiqc and fastqc

From Table 2, we can say that the percentage of uniquely mapped reads of all samples was above 80% with highest (89.2%) in SRR1178003 and lowest (84.2%) in SRR1177966. This indicates that the quality of sample reads to be considered for alignment is good. Number of uniquely mapped reads was lowest in SRR1177966 (16 million) and highest in SRR1178001 (22.2 million). %GC content was 49% for all samples except SRR1177966_1 and SRR1177966_2 which had 48% GC content which is in the expected range for rats. Sequence was of 101 base pairs for all samples.

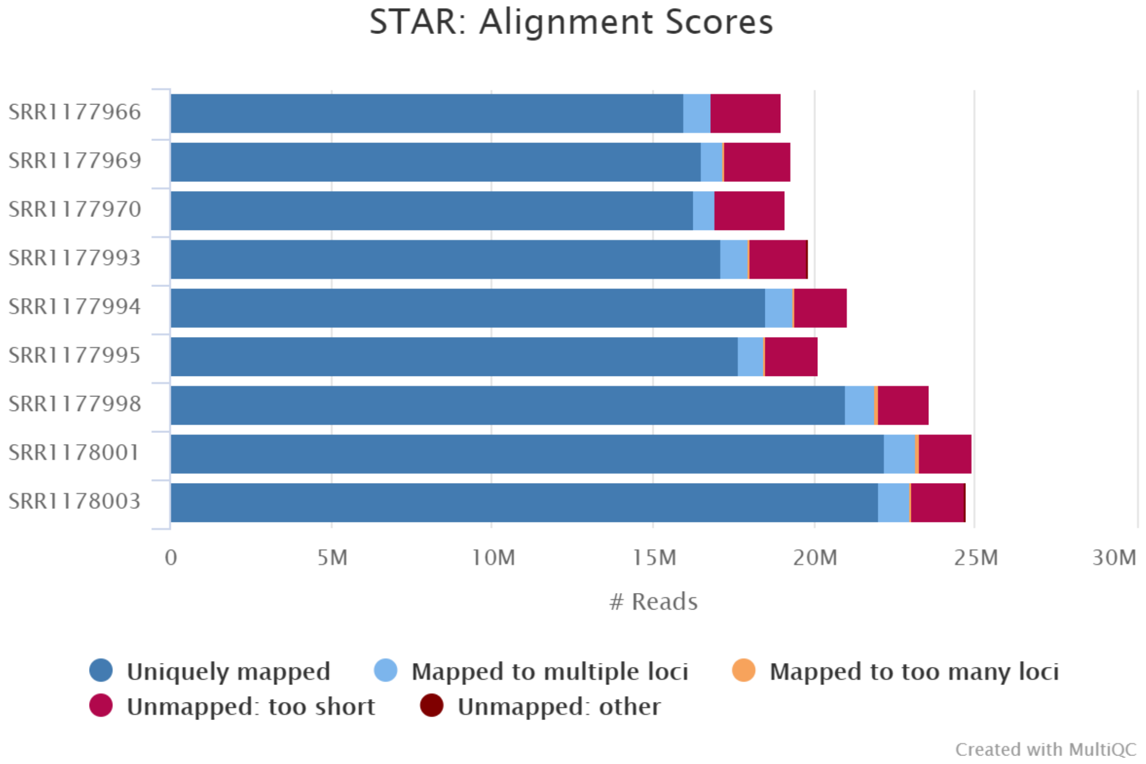


Figure 1: Results from Multiqc for each sample which include STAR alignment scores with sample names on y-axis and the reads distribution on x-axis

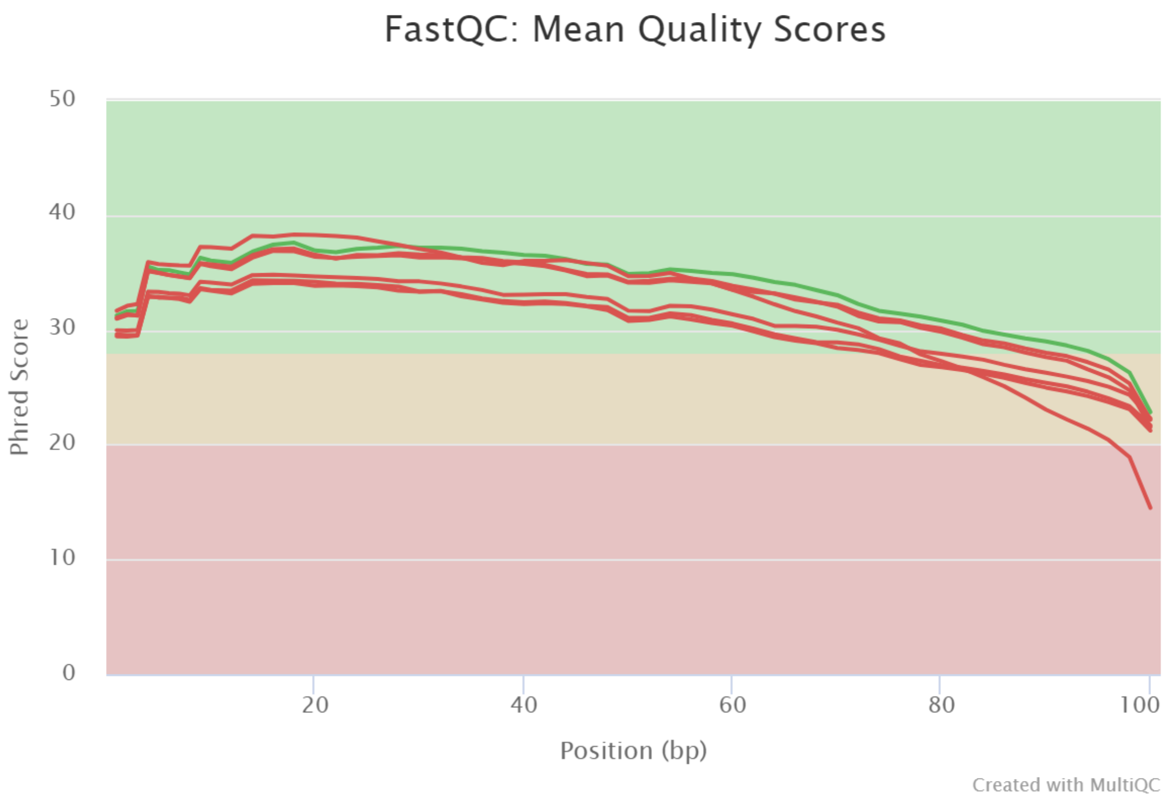


Figure 2: Results from Multiqc report which include fastqc mean quality scores with Phred scores on y-axis and base pair count on x-axis.

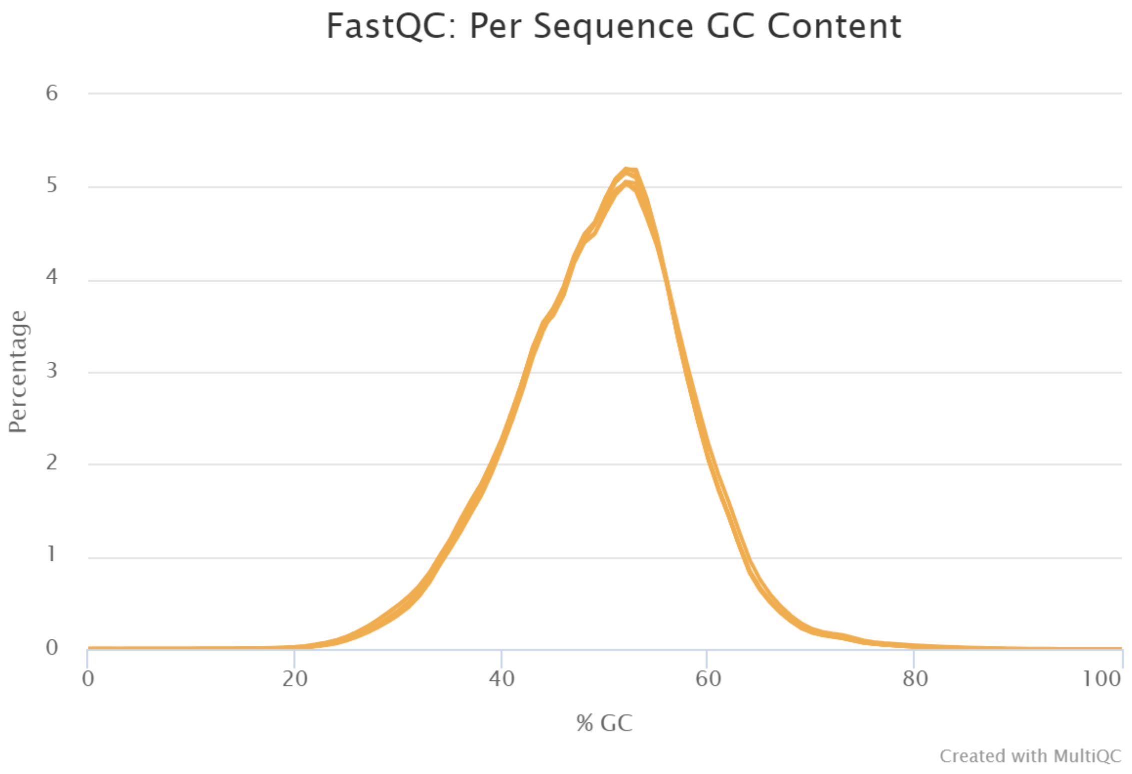


Figure 3: Results of fastqc per sequence GC content from multiqc report

Figure 1, Figure 2 and Figure 3 are the plots from the MultiQC report. Figure 1 shows that most of the STAR alignments were uniquely mapped with a small proportion of unmapped reads which were too short. As per Figure 2, there is declining quality of reads towards the end however no quality score appeared to be too low since all were in the green zone which depicted a higher Phred score. A possible reason for this declining trend could be due to signal decay/phasing during sequence run or adapter sequences. Figure 3 shows that %GC content falls in the warning zone(yellow) with all samples having 48-49% GC content which is similar to the expected rat's genetic makeup[2]. Also some of the samples were eliminated by multiqc due to low quality of reads and were not shown in Figure 3.

Methods:

Sample Statistics and Alignment:

FastQC and MultiQC were used for quality control on the raw data to check the quality of the reads. Each sample was aligned against the rat genome using the STAR alignment tool (this step produced 9 bam files which would be used in the feature counts step).

Read Counting:

The featureCounts tool[3] was used to count reads from the bam files generated by the STAR alignment tool against a provided gene annotation file (rn4_refGene_20180308.gtf). MultiQC was then used to assess the quality of the featureCounts output. The counts for each sample were combined into a single count matrix, and a box plot was created to show the distribution of the counts in each sample.

RNA-Seq Differential Expression Analysis:

DESeq2 is a Bioconductor package that estimates variance-mean dependence in count data and tests for differential expression using a negative binomial distribution-based model[4]. DESeq2 package was used to analyze the gene count differences between the experimental samples and control samples. The samples were divided into three groups based on the modes of action: AhR, CAR/PXR, and Cytotoxic. The sample groups were paired with the appropriate controls based on the delivery vehicles (i.e. corn oil, saline). The outputs of the DESeq2 program were sorted based on adjusted p-value, so that the top 10 differentially expressed genes from each group could be selected. The data was visualized using histograms and scatter plots.

Microarray Differential Expression Analysis:

The limma Bioconductor package was used in R to determine which genes were differentially expressed based on the microarray data for each sample. The differential expression results were sorted by adjusted p-value, and the top 10 differentially expressed genes from each sample group were selected for comparison to the results from the RNA-Seq differential expression analysis. Data was visualized with histograms and scatter plots.

Determination of Concordance:

Concordance between the two sets of differentially expressed genes (from the RNA-seq analysis and Microarray analysis) was calculated as follows:

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

Results:

Summary statistics from MultiQC obtained after counting the reads with FeatureCounts can be seen in Table 3 and Figure 4. Table 3 shows that all samples were reported to have around 60% assigned. Figure 4 shows the breakdown of assigned reads, unassigned multimapped reads, unassigned reads with no features, and unassigned ambiguous reads. The boxplot in Figure 5 shows that all samples have a similar distribution of counts.

Table 3. MultiQC Quality Statistics showing the percept assigned of the experimental samples used in our analyses

Sample Name	% Assigned	M Assigned
SRR1177966	60.4%	21.9
SRR1177969	62.1%	22.8
SRR1177970	62.2%	22.5
SRR1177993	60.8%	23.6
SRR1177994	61.5%	25.6
SRR1177995	61.4%	24.3
SRR1177998	61.4%	29.0
SRR1178001	61.2%	30.5
SRR1178003	62.7%	30.8

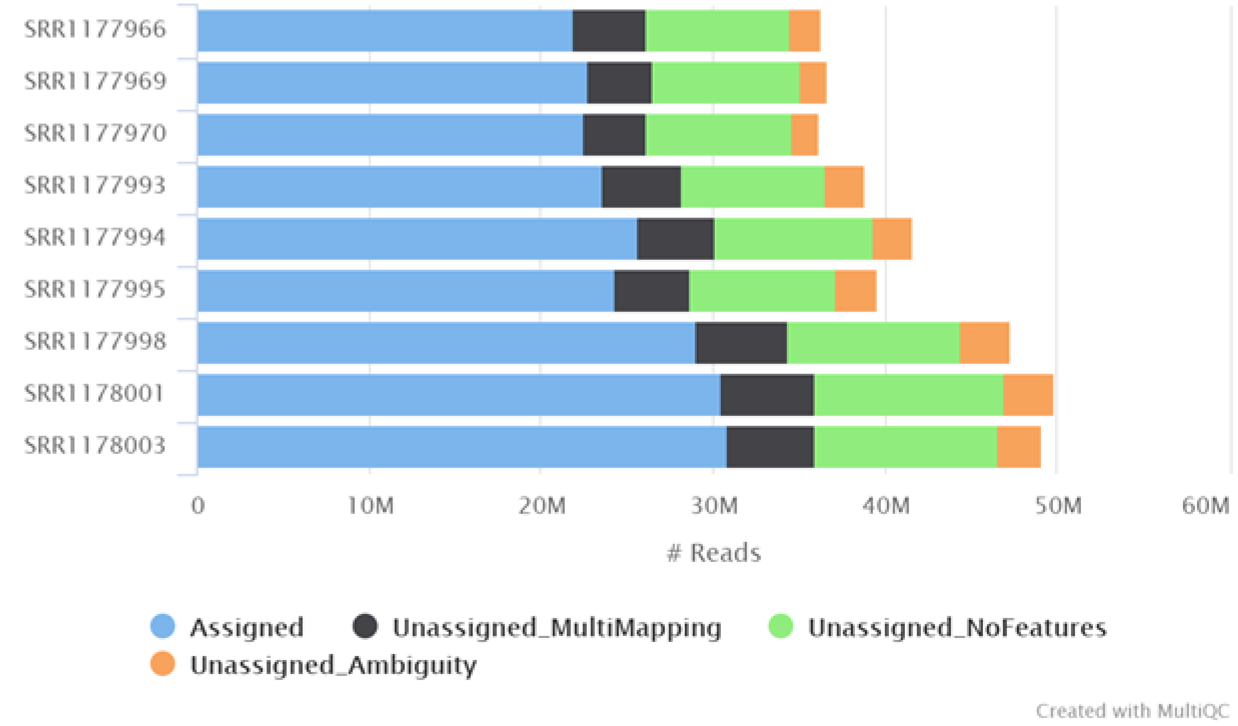


Figure 4. Assigned and Unassigned reads in experimental samples

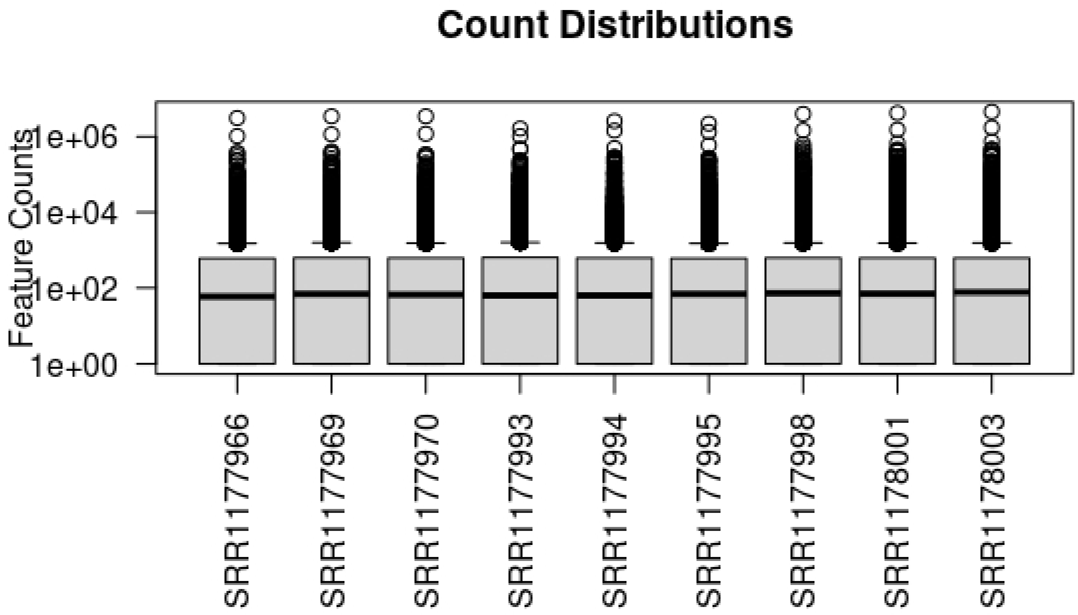
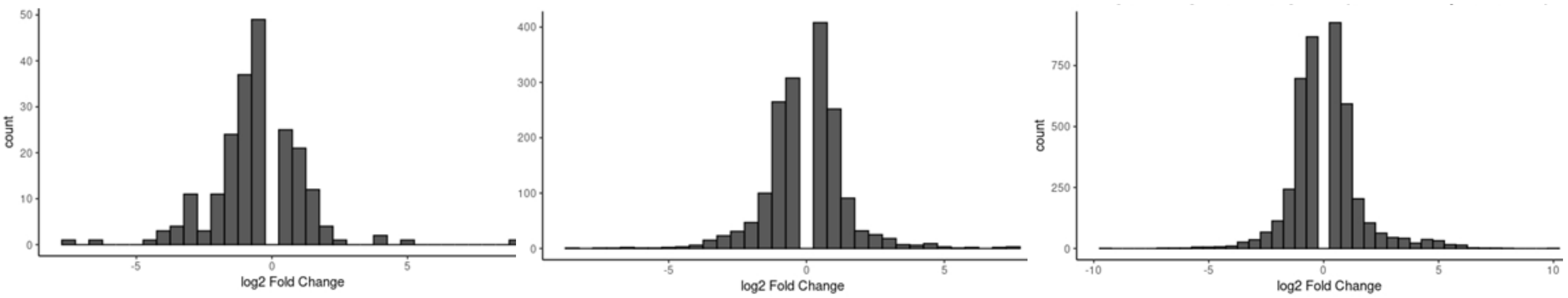


Figure 5. Boxplot showing the count distribution of each sample, with samples on the x axis and log(counts) on the y axis

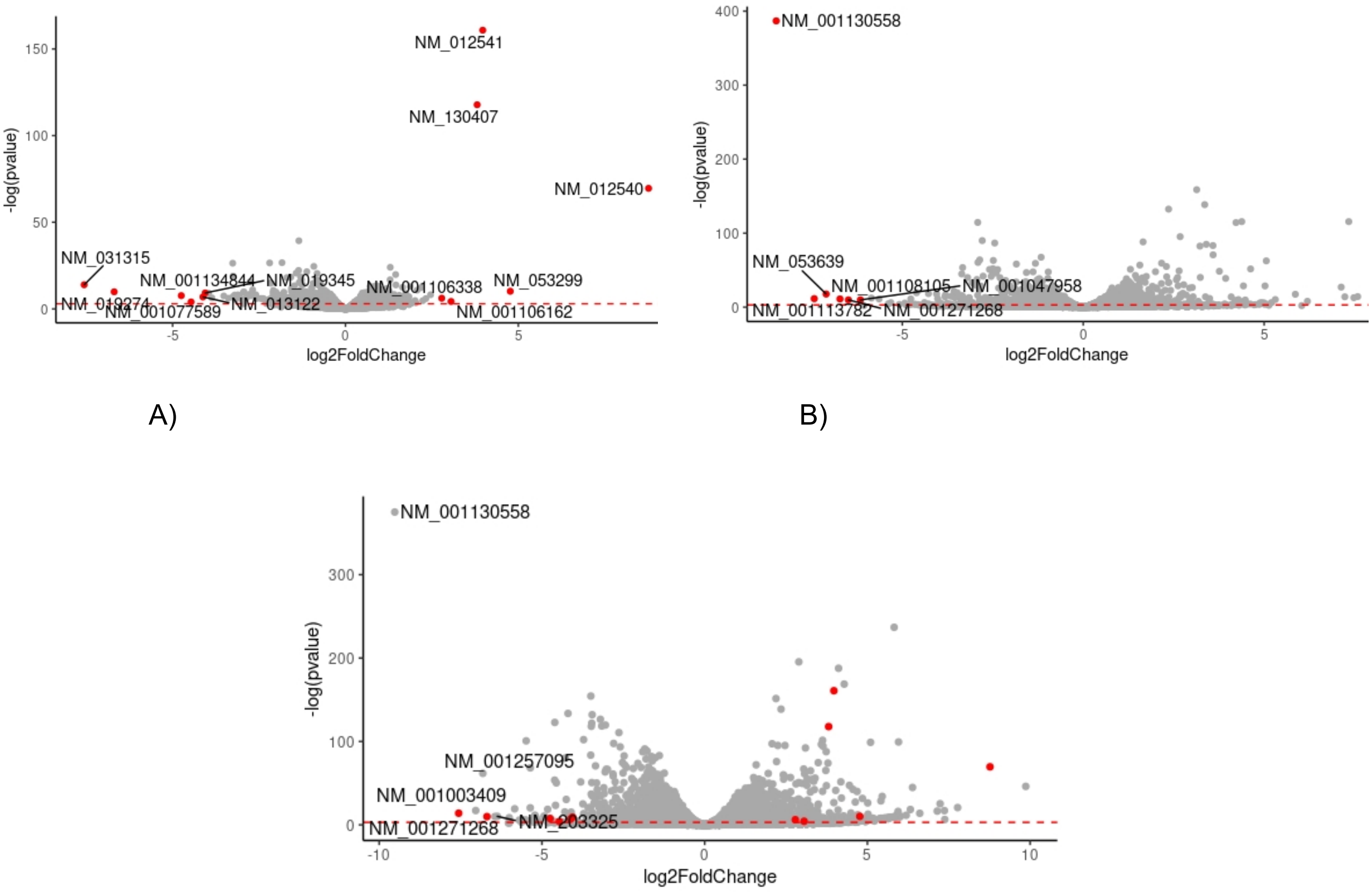


A)

B)

C)

Figure 6. Histograms showing log fold change values for each treatment group. A) AhR, B) CAR/PXR, C) Cytotoxic



A)

B)

C)

Figure 7. Volcano plots of differential gene expression data. The red line shows the significance cutoff of 0.05. The most differentially expressed genes are marked with red points.

The limma analysis returned the amount of genes listed in Table 4 with an adjusted P-value < 0.05. Tables 5-7 list the top DE genes from the analysis as well as their P-values. For each MOA, we got a different order of magnitude in the amount of DE genes. Therefore, it is expected to have variation in the subsequent analyses. Figures 10 and 11 show the concordance plotted vs the number of DE genes from each platform.

Table 4. Number of DE genes by MOA.

MOA	# DE genes
AhR	106
CAR/PXR	58
Cytotoxic	5865

Table 5. Top DE genes for CAR/PXR. Ordered by increasing adjusted P value.

SYMBOL	GENENAME	Adj.P.Val
Aoc3	amine oxidase, copper containing 3	0.00138824
G6pc	glucose-6-phosphatase, catalytic subunit	0.00138824
Psme3	proteasome activator subunit 3	0.00138824
Tmem252	transmembrane protein 252	0.00217065
Abcc9	ATP binding cassette subfamily C member 9	0.00217065
Rnf144a	ring finger protein 144A	0.00217065
Inhba	inhibin subunit beta A	0.0024811
Aoc3	amine oxidase, copper containing 3	0.0024811
G6pc	glucose-6-phosphatase, catalytic subunit	0.0024811
Psme3	proteasome activator subunit 3	0.0024811
Mtss1	MTSS I-BAR domain containing 1	0.0024811
Ugt2b1	UDP glucuronosyltransferase 2 family, polypeptide B1	0.00319354
Sesn2	sestrin 2	0.00319354

Table 6. Top DE genes for Cytotoxic. Ordered by increasing adjusted P value

SYMBOL	GENENAME	Adj.P.Val
Atf3	activating transcription factor 3	3.55E-10
Klf6	Kruppel-like factor 6	1.70E-08
Ccng1	cyclin G1	3.99E-08
Btg3	BTG anti-proliferation factor 3	3.99E-08
Sez6	seizure related 6 homolog	4.74E-08
Tnfrsf12a	TNF receptor superfamily member 12A	1.02E-07
Zfand2a	zinc finger AN1-type containing 2A	1.02E-07
Abcb1b	ATP-binding cassette, subfamily B (MDR/TAP), member 1B	1.59E-07
Abcb1a	ATP binding cassette subfamily B member 1A	1.59E-07
Mybl1	MYB proto-oncogene like 1	3.12E-07

Table 7. Top DE genes for AhR. Ordered by increasing adjusted P value.

SYMBOL	GENENAME	adj.P.Val
Cyp1a2	cytochrome P450, family 1, subfamily a, polypeptide 2	1.16E-12
Cyp1a1	cytochrome P450, family 1, subfamily a, polypeptide 1	4.46E-11
Ugt1a3	UDP glycosyltransferase 1 family, polypeptide A3	1.77E-08
Ugt1a2	UDP glucuronosyltransferase 1 family, polypeptide A2	1.77E-08
Ugt1a8	UDP glucuronosyltransferase family 1 member A8	1.77E-08
Ugt1a1	UDP glucuronosyltransferase family 1 member A1	1.77E-08

Ugt1a9	UDP glucuronosyltransferase family 1 member A9	1.77E-08
Ugt1a6	UDP glucuronosyltransferase family 1 member A6	1.77E-08
Ugt1a7c	UDP glucuronosyltransferase 1 family, polypeptide A7C	1.77E-08
Ugt1a5	UDP glucuronosyltransferase family 1 member A5	1.77E-08
Ugt1a3	UDP glycosyltransferase 1 family, polypeptide A3	2.46E-07
Ugt1a1	UDP glucuronosyltransferase family 1 member A1	2.46E-07
Ugt1a6	UDP glucuronosyltransferase family 1 member A6	2.46E-07
Ugt1a2	UDP glucuronosyltransferase 1 family, polypeptide A2	2.46E-07
Ugt1a9	UDP glucuronosyltransferase family 1 member A9	2.46E-07
Ugt1a5	UDP glucuronosyltransferase family 1 member A5	2.46E-07
Ugt1a7c	UDP glucuronosyltransferase 1 family, polypeptide A7C	2.46E-07
Ugt1a8	UDP glucuronosyltransferase family 1 member A8	2.46E-07
LOC100910660	serine/arginine-rich splicing factor 3-like	0.00010353
Srsf3	serine and arginine rich splicing factor 3	0.00010353
Slc34a2	solute carrier family 34 member 2	0.00072432
Ccnd1	cyclin D1	0.00117608
Ttr	transthyretin	0.00135474
Mt1	metallothionein 1	0.00135474
Mt2A	metallothionein 2A	0.00135474

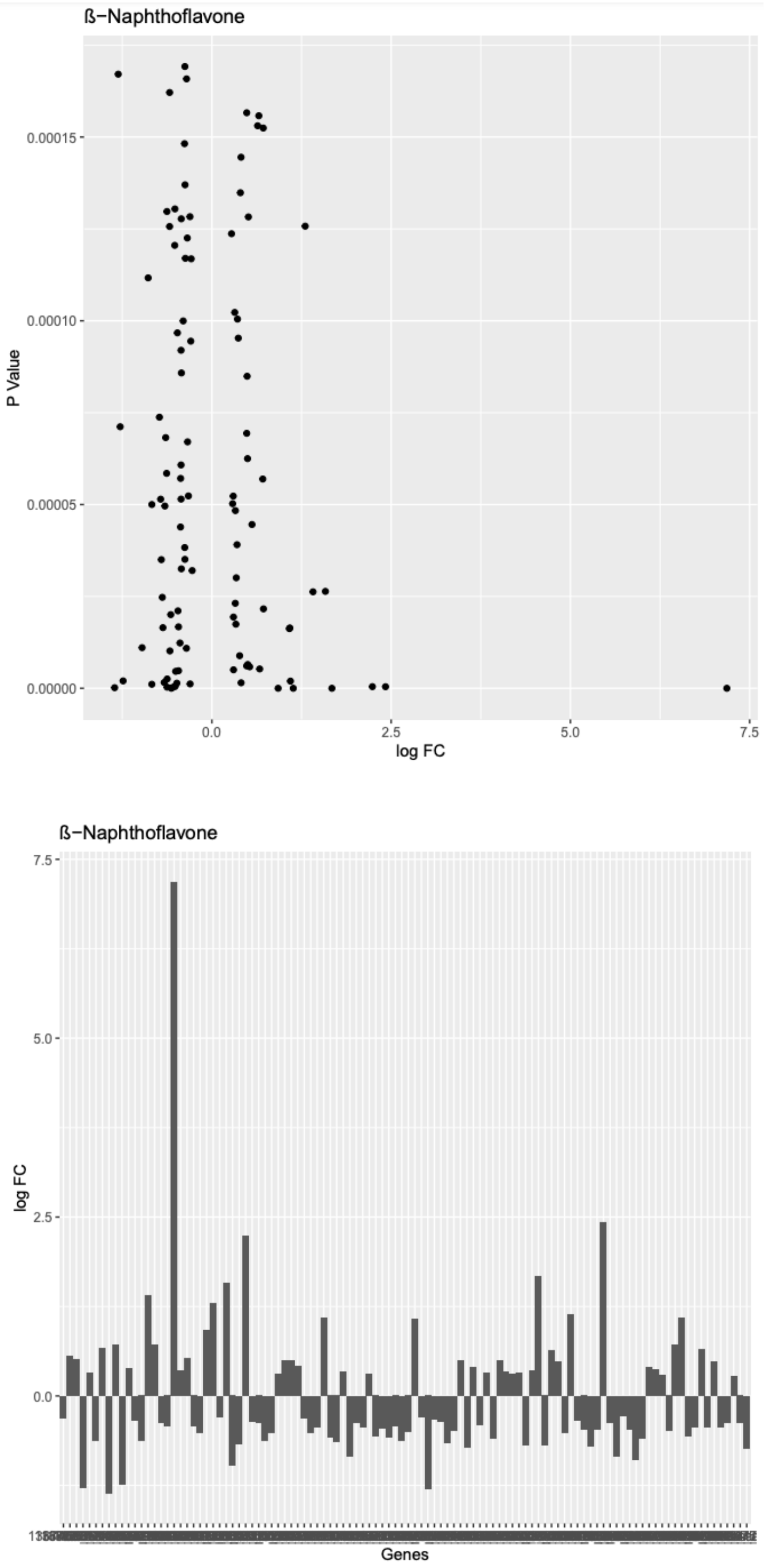


Figure 8. Scatter plot and Histogram of AhR.

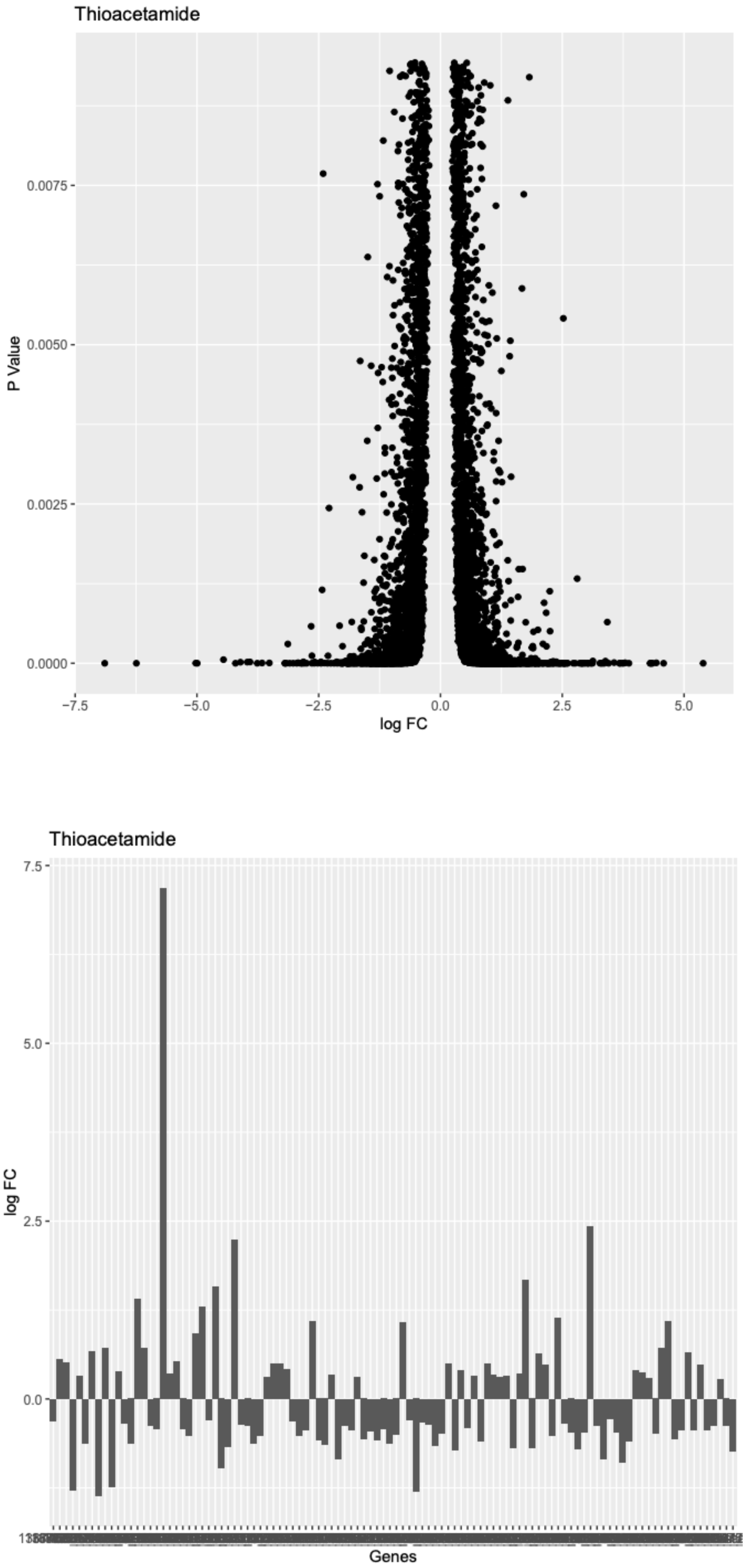


Figure 8. Scatter plot and Histogram of Cytotoxic.

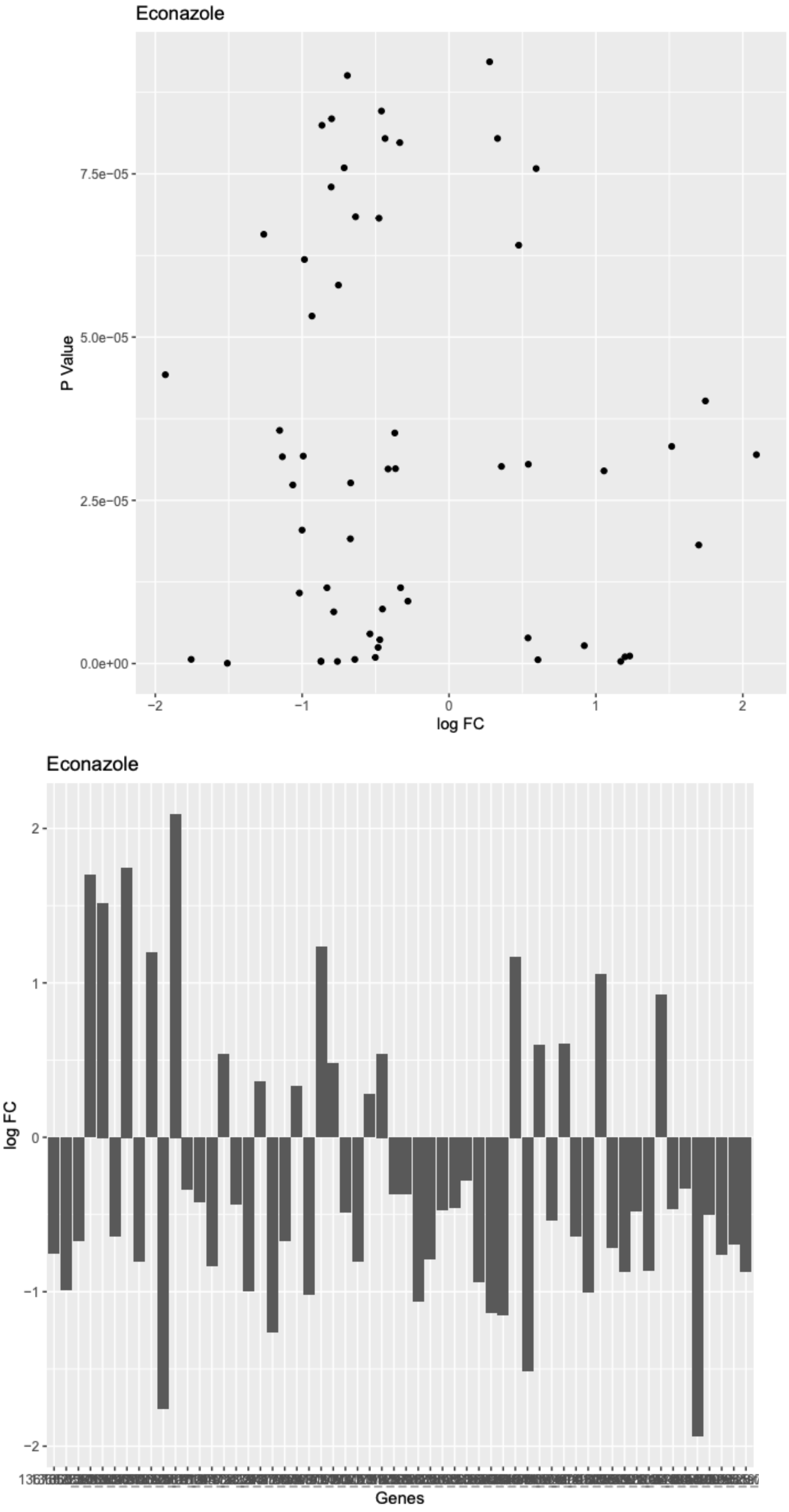


Figure 9. Scatter plot and Histogram of CAR/PXR.

Table 8. Concordance calculations for the three groups

AhR	overall	0.06843595
	above	0.0355849
	below	-0.0324335
CAR/PXR	overall	0.12454019
	above	0.05784784
	below	0.412632
Cytotoxic	overall	-0.0007473
	above	0.02210792
	below	-0.0188333

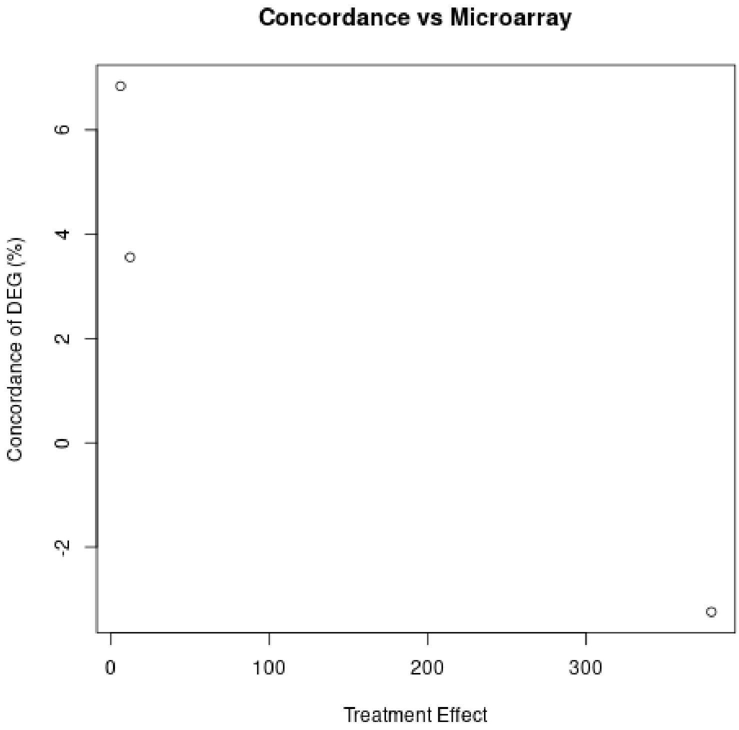


Figure 10. Number of DE genes in Microarray vs overall concordance.

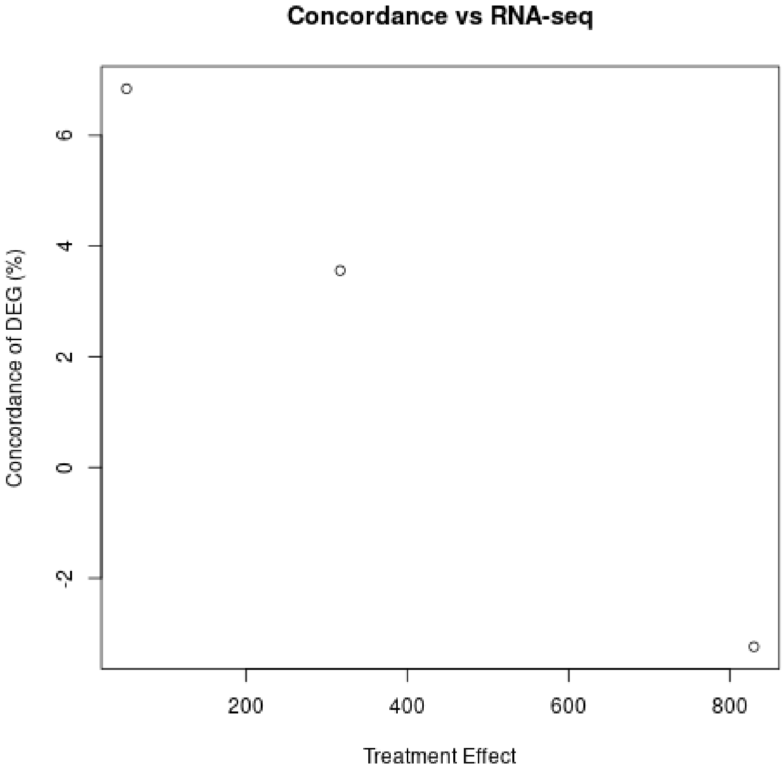


Figure 11. Number of DE genes in Microarray vs overall concordance.

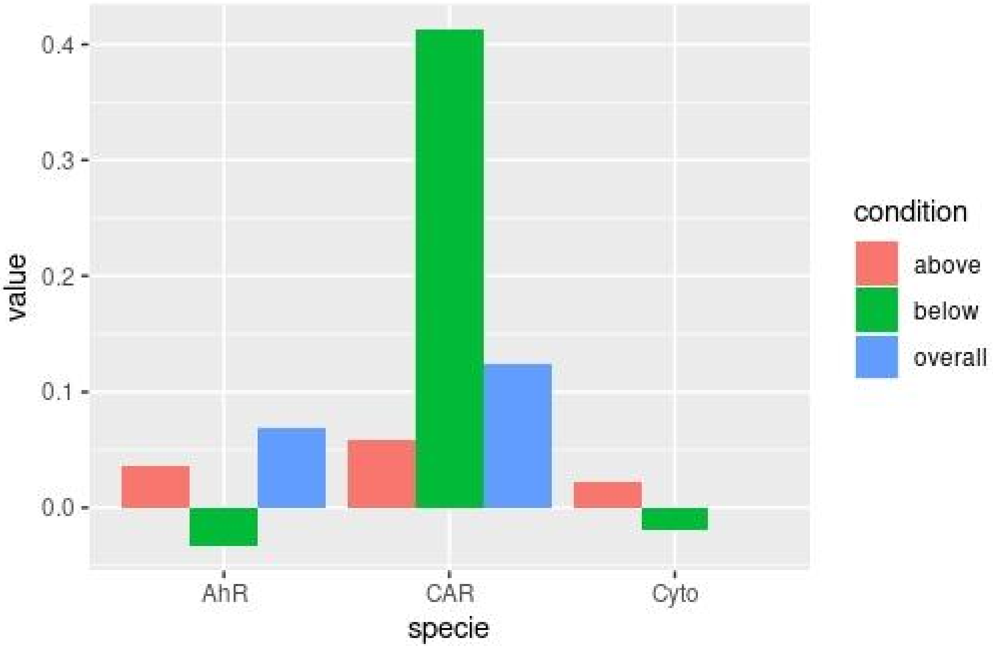


Figure 12. Overall, above and below concordance for each group.

Discussion:

The different orders of magnitude in the different MOAs led to a large variation in the results of the analysis for each MOA. When listing the top DE genes, Table 7 shows various transcripts for *UGT1A*, a gene whose protein converts bilirubin, a byproduct of the dissolution of red blood cells, from its toxic to its non-toxic form. With 16 alternative transcripts of this gene, it would be worth further investigating its effects in the laboratory. In Figure 8, the gene with the highest FC(7.1) was identified as cytochrome p450, a superfamily of proteins that oxidises steroids, important in the dissolution of other harmful compounds. For the concordance calculations, x <0 on multiple occasions, suggesting the overlap was a byproduct of chance alone. As can be seen in Table 8 and in Figure 12, consistent with the authors' results, the concordance is expected at its highest if only genes above the median are considered. It is also consistent with the author's results that the degree of concordance varies between the chemicals to which the rats' livers were subjected to. As a further insight, it would be interesting to calculate the concordance between different platforms with chemicals that have similar effects on the rats' livers, and calculate the concordance between one chemical in microarray and another chemical in RNA-seq. The cross-platform concordance from our analysis is different from that of the authors and the reasons for this remain unknown. There is no

Using the gene enrichment analysis results from Supplementary Table 10 in Wang et al., a comparison was made between the pathway enrichment results of our selected genes to those reported in the paper. The online gene set enrichment platform DAVID was used in making the following comparisons between the reported enriched pathways;

CAR/PXR (7)(Wang et al.)	CAR/PXR(7)(Our Analyses Results) (Sorted by Top 7 Ranked by P-Value with minimum of 25 Associated Genes)
Aryl hydrocarbon receptor signaling	Chemical carcinogenesis
Glutathione-mediated detoxification	Metabolic pathways
LPS/IL-1 mediated inhibition of RXR function	Steroid hormone biosynthesis
NRF2-mediated oxidative stress response	Retinol metabolism
Nicotine degradation II	Metabolism of xenobiotics by cytochrome P450
PXR/RXR activation	Biosynthesis of antibiotics
Xenobiotic metabolism signaling	MicroRNAs in cancer

Table 9: Differing enriched pathways between our's and Wang et al.'s gene enrichment analysis for the CAR/PXR MOA. Yellowighlighted cells show gene enrichment pathways related either to metabolism of

xenobiotics or of direct regulation/ activation of CAR/PXR (or its affiliated family members), blue cells show pathways related to compound biosynthesis.

CAR/PXR and AhR are both xenobiotic receptors dealing with the detection of xenobiotics in the body. Our analysis displayed some of these associated pathways, while the remaining dealt with cancer pathways and biosynthesis of various compounds. The pathways associated with biosynthesis of new compounds could be directly related to the metabolism of xenobiotic compounds, as said xenobiotics are detected by the body and then broken down into raw materials. An example of this found in the Cytotoxic table come in the form of the enriched pathways; “Valine, leucine and isoleucine degradation” and “Biosynthesis of amino acids”, in which the degradation of the ingested foreign amino acids leads into the body’s natural production of other amino acids.

AhR (10)(Wang et al.)	**AhR (10)(Analyses Results) (Sorted by Top 7 Ranked by P-Value with minimum of 5 Associated Genes)**
Acetone degradation I (to methylglyoxal)	Metabolism of xenobiotics by cytochrome P450
Aryl hydrocarbon receptor signaling	Chemical carcinogenesis
Bupropion degradation	Drug metabolism - cytochrome P450
LPS/IL-1 mediated inhibition of RXR function	Fatty acid elongation
Melatonin degradation I	Fatty acid degradation
Nicotine degradation II	Arginine and proline metabolism
Nicotine degradation III	Glutathione metabolism
Retinoate biosynthesis I	Metabolic pathways (35)
Superpathway of melatonin degradation	Bile secretion
Xenobiotic metabolism signaling	Biosynthesis of unsaturated fatty acids

Table 10: Differing enriched pathways between our’s and Wang et al.’s gene enrichment analysis for the AhR MOA. Highlighted cells show gene enrichment pathways related either to metabolism of xenobiotics or of direct regulation/ activation of AhR (or its affiliated family members), blue cells show pathways related to compound biosynthesis.

Cytotoxic (15)(Wang et al.)	Cytotoxic(15)(Analyses Results) (Sorted by Top 15 Ranked by P-Value with minimum of 25 Associated Genes)
Acetone Degradation I (to Methylglyoxal)	Metabolic pathways
Bupropion degradation	Complement and coagulation cascades
Cell cycle: G2/M DNA damage checkpoint regulation	RNA transport
Citrulline biosynthesis	Ribosome biogenesis in eukaryotes
Estrogen biosynthesis	Cell cycle
LPS/IL-1 mediated inhibition of RXR function	Fatty acid metabolism
Melatonin degradation I	Valine, leucine and isoleucine degradation
Methylglyoxal degradation III	Biosynthesis of antibiotics
NRF2-mediated oxidative stress response	Steroid hormone biosynthesis
Nicotine degradation II	Peroxisome
Pyrimidine ribonucleotides de novo biosynthesis	Retinol metabolism
Regulation of eIF4 and p70S6K signaling	p53 signaling pathway
Superpathway of melatonin degradation	Carbon metabolism
Superpathway of methionine degradation	Purine metabolism
Xenobiotic metabolism signaling	Biosynthesis of amino acids

Table 11: Differing enriched pathways between our's and Wang et al.'s gene enrichment analysis for the Cytotoxic MOA. Highlighted cells show gene enrichment pathways related either to metabolism of xenobiotics, blue cells show pathways related to compound biosynthesis.

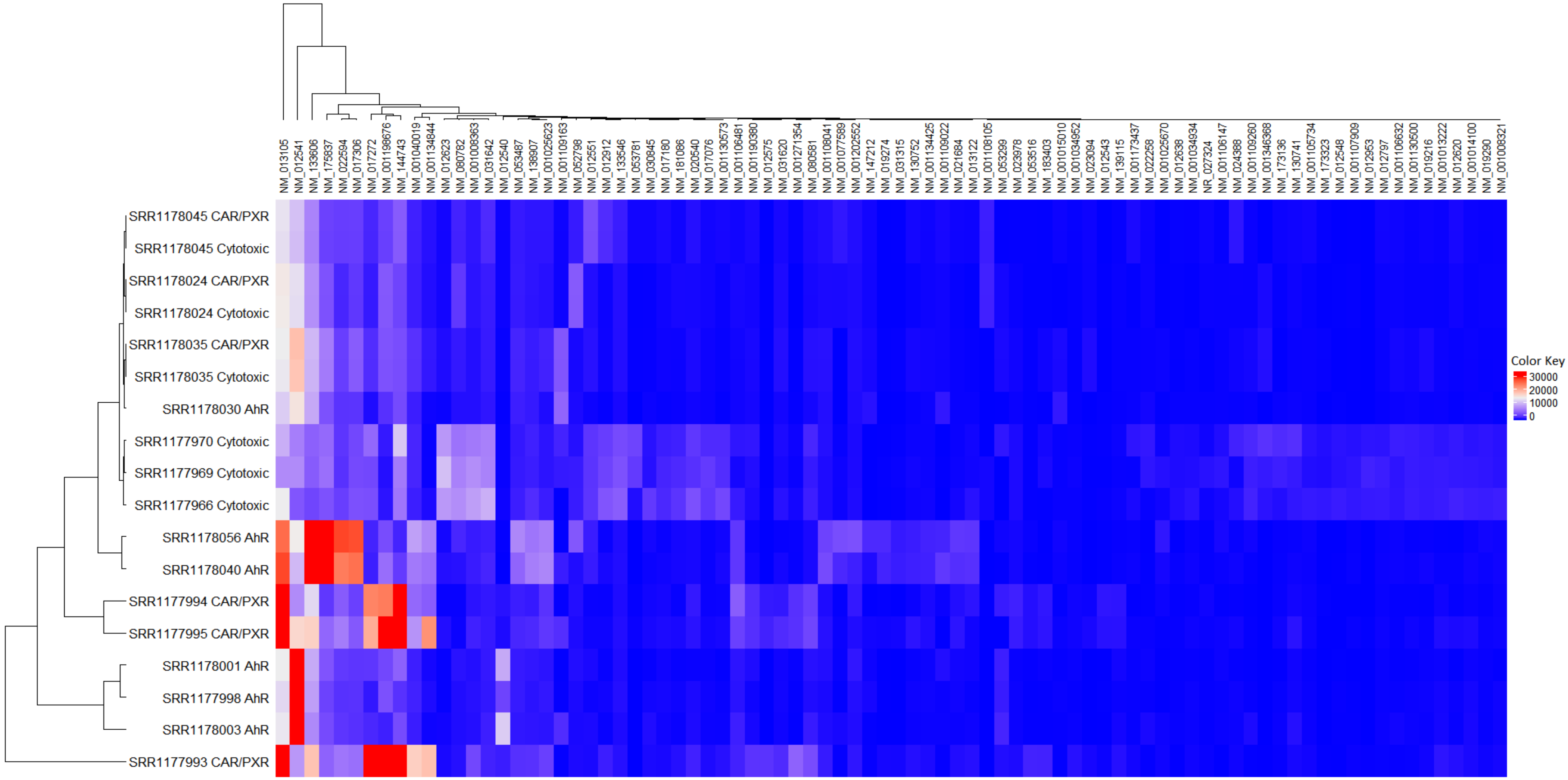


Figure 13: Hierarchical Cluster Heatmap of gene expression found using the normalized differentially expressed (DE) genes of each MOA (CAR/PXR, AhR, Cytotoxic). The matrix of DE genes was very large; we utilized the coefficient of variation to filter through the dataset to find a presentable medium of data. (CV cutoff of >1.2).

Conclusion:

Although consistent with the authors' results, the magnitude of our concordance calculations differed from that of the main text. Our reproduction of the gene pathway analysis was relatively successful; the DAVID tool was able to replicate similar metabolic, degradation, biosynthetic, and general xenobiotic detection pathways.

References

1. Wang, Charles et al. "The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance." Nature biotechnology vol. 32,9 (2014): 926-32. doi:10.1038/nbt.3001
2. Freeman, B. (2017, June 7). Alignment with STAR. Retrieved April 7, 2021, from Introduction to RNA-Seq using high-performance computing - ARCHIVED website: https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html
3. Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics (Oxford, England), 30(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656
4. 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.
5. MultiQC. (2013). Retrieved April 7, 2021, from Multiqc.info website: https://multiqc.info/
6. Zhang L, Kasif S, Cantor CR, Broude NE. GC/AT-content spikes as genomic punctuation marks. Proc Natl Acad Sci U S A. 2004 Nov 30;101(48):16855-60. doi: 10.1073/pnas.0407821101. Epub 2004 Nov 17. PMID: 15548610; PMCID: PMC534751.
7. DAVID Functional Annotation Bioinformatics Microarray Analysis. David.ncifcrf.gov. Retrieved from https://david.ncifcrf.gov/home.jsp
8. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). "**limma** powers differential expression analyses for RNA-sequencing and **microarray** studies." Nucleic Acids Research, 43(7), e47