Concordance of microarray and RNA-Seq differential gene expression

Data curator- Neha Gupta
Programmer - Rachel Thomas
Analyst - Yichi Zhang
Biologist- Varun Raghuraman
TA- Jackie

Introduction:

Before RNA sequencing, microarray analysis was considered the standard platform for genome expression profiling. Number of studies have been performed based on the microarray platforms, but the advancement of new technologies brings the promise of improvement. Through this study Wang et al.[1] investigate both these platforms of genome expression. The study design was to compare differentially expressed genes (DEGs) from Illumina RNA-seq and Affymetrix microarray data from the same set of liver samples of rats under varying degrees of perturbation by 27 chemicals. These chemicals represent multiple modes of action (MOA). The primary goal was to measure the degree of concordance or overlap between these platforms for DEGs, MOAs and pathways. The scope of this project was to study a subset of the conditions (‘tox group’ of three chemicals with different modes of action). After selecting the subset of samples to analyze, the rat short reads were aligned to it’s reference genome. Then differential gene expression analysis on RNA-seq and microarray data was performed and results from our analysis were compared to the reference article

Data:

For the purpose of this project, we choose tox-group 2 to analyze. All the datasets and samples were made available to us. The microarray experiment was performed on Affymetrix whole genome GeneChip RatGenome 230.2.0 Array and sequencing was carried out 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. Illumina TruSeq RNA sample preparation kit and SBS kit v3 was used for the experiment. The data were deposited and made available for public use in the Sequence Read Archive(SRA) database under accession number SRP024314, GSE55347, and GSE47875. For each sample, depth of around 23-25 million paired-end 101bp reads were generated.

The selected tox group and relevant information is cited in Table 1. The MOA's (mode of action) are Cytotoxic (cytotoxicity, toxic to cells), Car/Pxr (orphan nuclear hormone receptors) and AhR (aryl hydrocarbon receptor) which represent the mediator/receptor process. The mice were given listed chemical treatments during the experiment. For this group, the only chemicals used were Thioacetamide, Econazole and Beta-naphthoflavone respectively. The vehicle column represents substance used to house the chemical for injection into the animal. The vehicles used were 100% saline, 100% corn oil and 0.5% CMC. Route represents how the chemical along with the vehicle substance was administered into the mouse. Oral gavage means that it was introduced into the mouse orally by force whereas intraperitoneal means it was given by needle injection around the abdomen area of the animal.

RUN	MOA	Chemical	Vehicle	Route
SRR1177966	Cytotoxic	Thioacetamide	SALINE_100_%	INTRAPERITONEAL
SRR1177969	Cytotoxic	Thioacetamide	SALINE_100_%	INTRAPERITONEAL
SRR1177970	Cytotoxic	Thioacetamide	SALINE_100_%	INTRAPERITONEAL
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
SRR1177998	AhR	Beta-Naphthoflavone	CMC_.5_%	ORAL_GAVAGE
SRR1178001	AhR	Beta-Naphthoflavone	CMC_.5_%	ORAL_GAVAGE
SRR1178003	AhR	Beta-Naphthoflavone	CMC_.5_%	ORAL_GAVAGE
SRR1178030	Control	Vehicle	CMC_.5_%	ORAL_GAVAGE
SRR1178040	Comtrol	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: Tox group 2 RNA sample information.

After selecting the tox-group, FastQc was performed on all the 9 samples for analyzing the quality of the reads. To align the reads, STAR [9] aligner was executed against the rat reference genome (rattus norvegicus). The STAR aligner used paired end samples fastq files as input and indexed the reference rat genome. As an output SAM files were generated that were converted to aligned BAM files. Further the BAM files along with the sample fastq files were used by the multiqc to give a general overall quality and alignment report of the reads with appropriate statistical results.

Samples	%aligned	M aligned (millions)	Chemical Treatment	GC%	Seq 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 from STAR alignment which includes the percent of uniquely mapped reads(% aligned), the number of uniquely mapped reads in millions(M aligned), average percent of GC content(GC %), chemical treatment and average sequence length in base pairs(length).

As shown in Table 2 above, the percent aligned for all the sample reads are above 84%, indicating that the quality of the library reads are good to be considered for alignment. The percentage of GC content is mostly 49% for the sample reads which is similar to the expected range of rats [2]. The average sequence length was 101 base pairs. All the reads were paired end i.e. each sample had 2 fastq files(one from 5’ and other from 3’ end). Sample SRR1177966 had the least mapped reads with 16 million base pairs whereas SRR1178001 had the highest of 22.2 millions mapped reads base pairs.

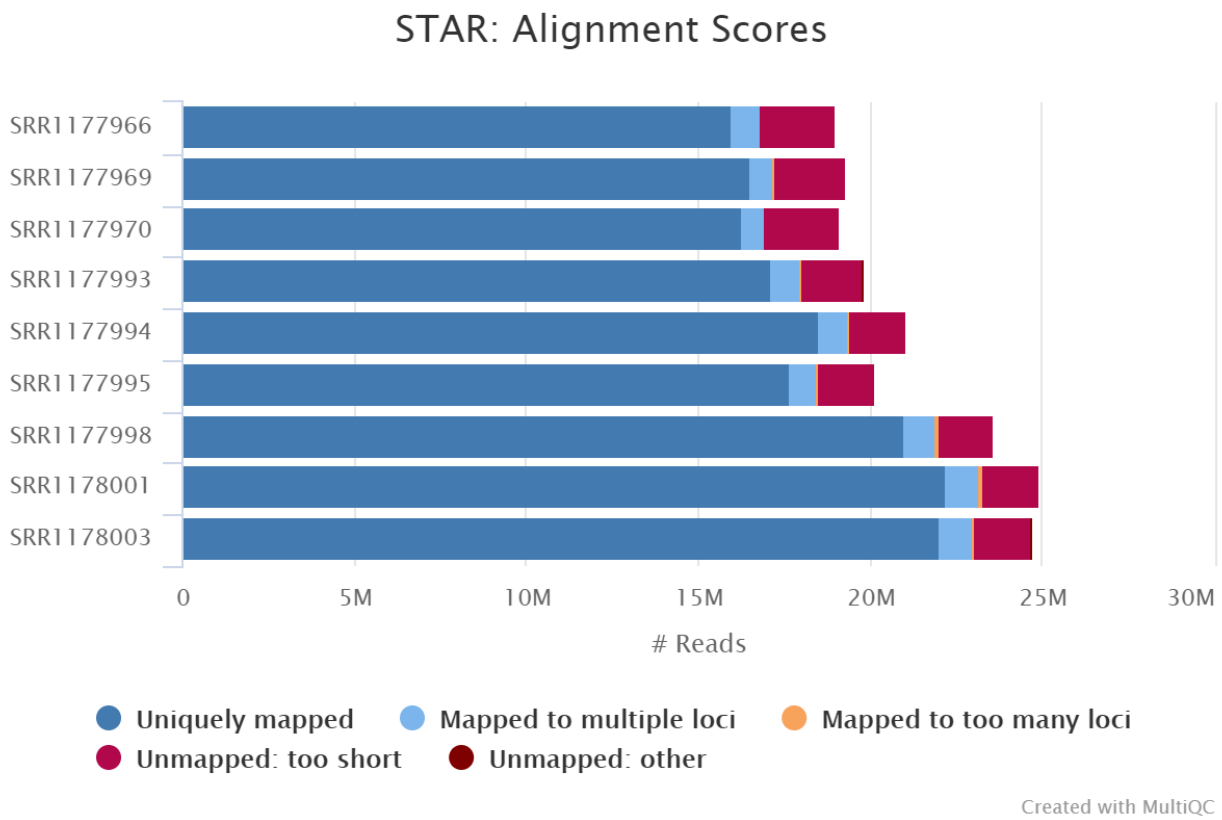


Figure 1: Result from multiqc for each sample which includes STAR alignment score.

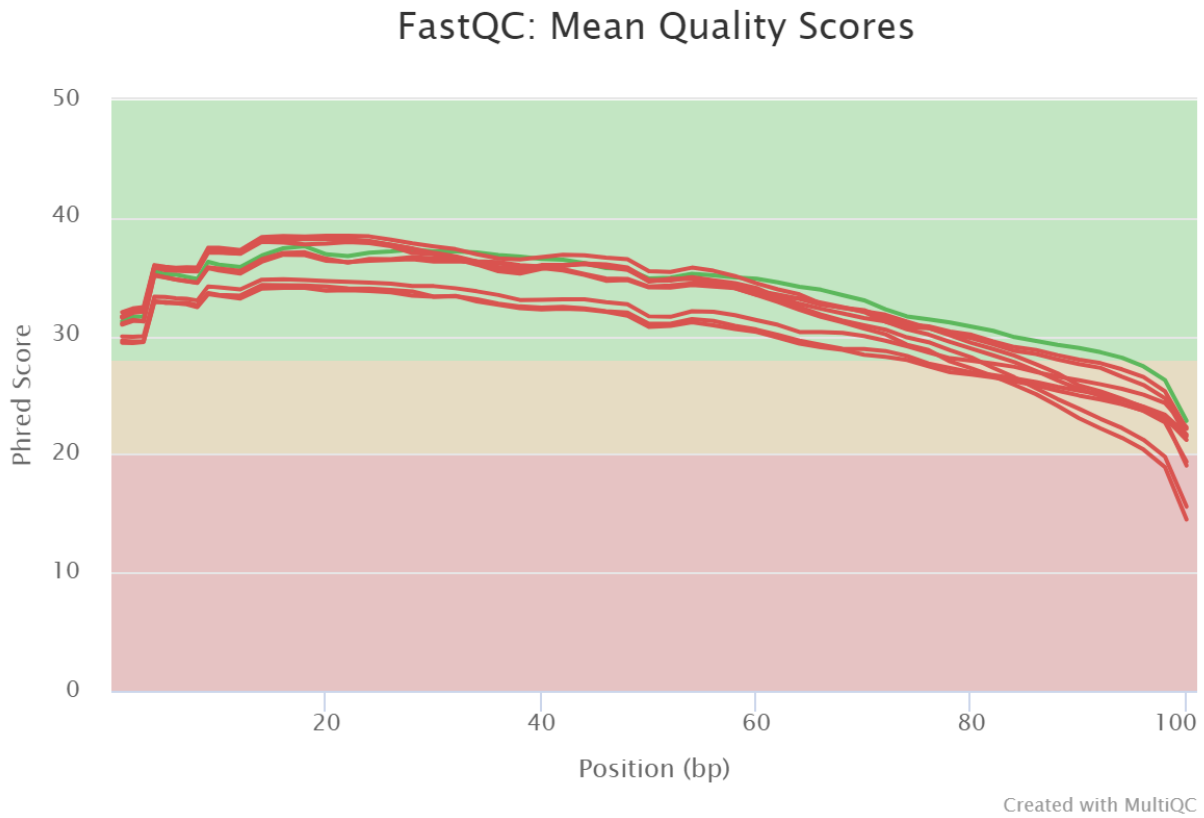


Figure 2: Result showing mean fastqc quality scores from multiqc for all 9 samples.

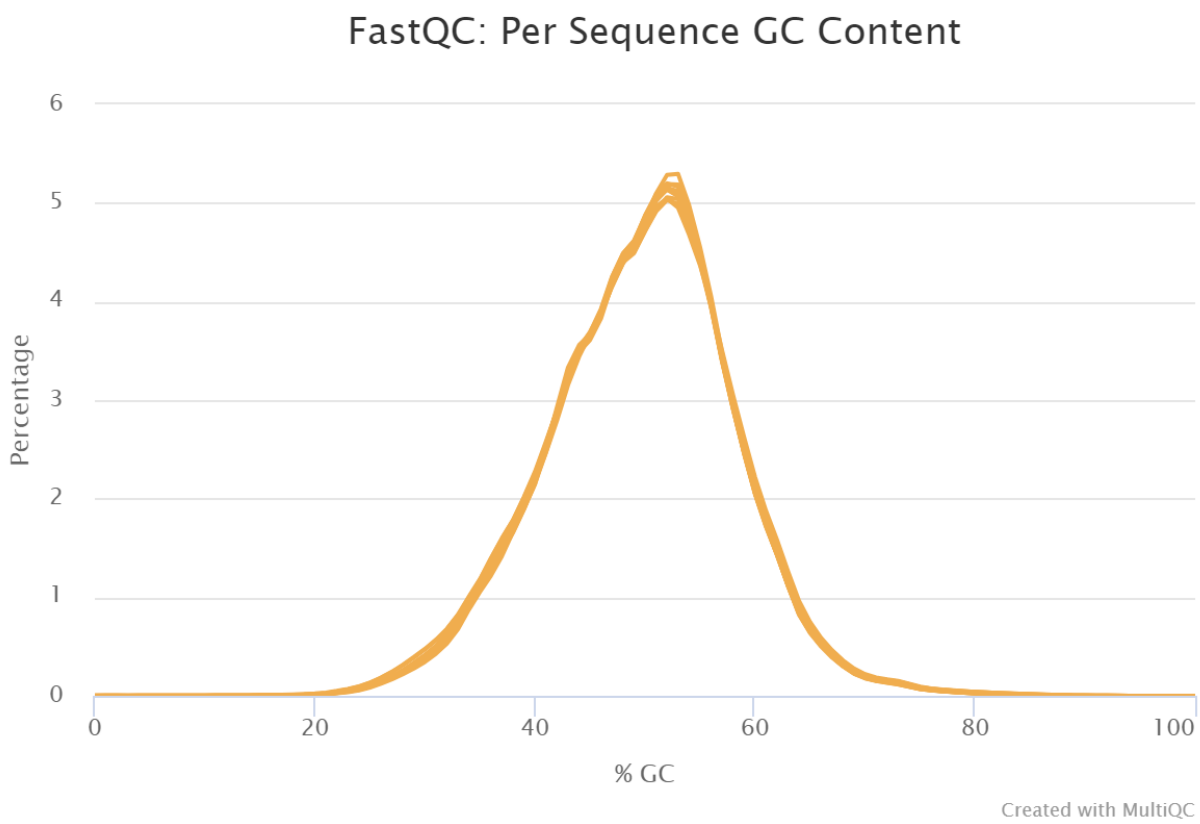


Figure 3: Result from multiqc depicting per sequence GC content for all samples.

Figure 1 shows that the STAR alignments were uniquely mapped with the reference genome whereas only a small percentage was unmapped due to their too short length. The green zone in the figure 2 showed mean quality scores across the bases depicting high phred scores. The graph also shows the declining trend in the read quality towards the end. This could be because of decay in signal strength, adapter sequences or phasing during sequencing run. In figure 3, it shows that all our samples fall in the warning zone i.e. all samples have GC% ranging from 48-49% which is similar to a rat's genetic makeup. Considering all our 9 samples, few were eliminated by multiqc itself that showed low quality reads and were not displayed in the above figures.

Methods:

- **Quantification of Gene Expression: featureCounts and MultiQC**

  Using the STAR alignment files, the number of read counts mapped to genes in the genome were generated using the featureCounts[4] tool on the shared computing cluster. FeatureCounts is a fast and efficient tool that counts reads that map to a single location. The 9 BAM files and a reference gene annotation

were used as featureCount inputs, and the output created 9 new count matrix files containing count reads at the genomic feature level. Next, MultiQC [5] was used to check the quality of the reads and to identify any outliers in our samples.The MultiQC summary is shown in Table 3. The individual count matrix files gathered from each sample were combined into one CSV file in Rstudio.

The counts for each sample are shown on a boxplot in Figure 5. The average counts number across all genes were found to be similar between samples. The multiqc report shows that the total number of reads varied between samples, as shown in Figure 4, but the percentage of assigned reads to unassigned reads remained relatively constant across all samples.

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

Table 3: MultiQC summary. Percentage of mapped reads assigned to the gene-ids in the reference genome. On average, 61.5% of reads were assigned to gene-ids. Sample SRR1178003 had the highest percentage of reads assigned to gene-ids (62.7%) and sample SRR1177966 had the lowest percentage of reads assigned to gene-ids (60.4%).

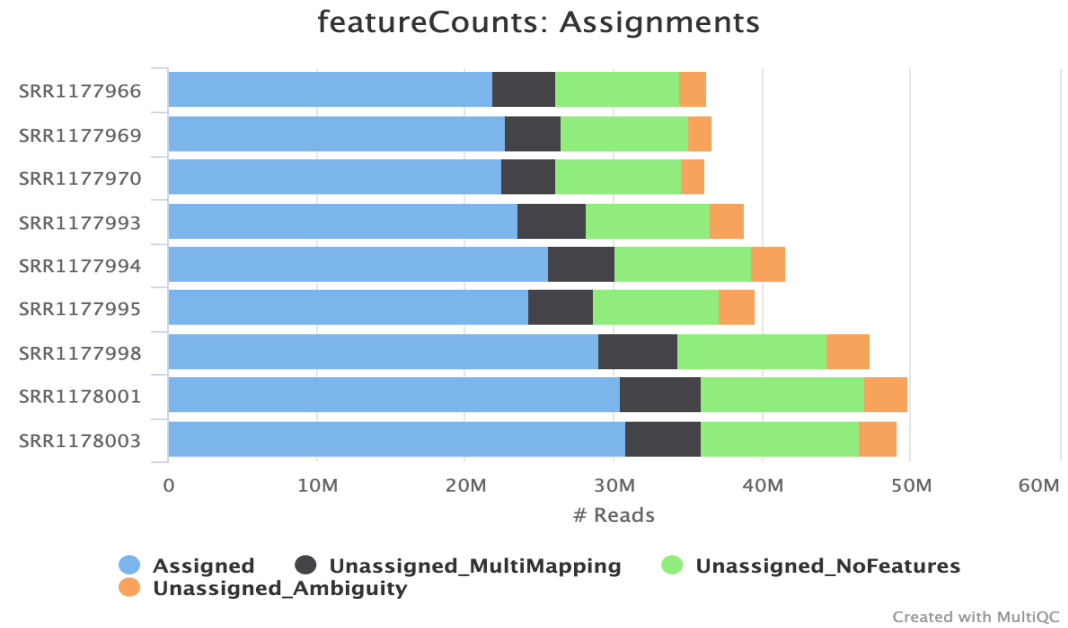


Figure 4: MultiQC report: Percentage of assigned reads. Multiqc results for featureCounts summarizing number of reads for treatment samples across four categories. The figure also shows the percentage of unassigned reads for every sample after running featureCounts.

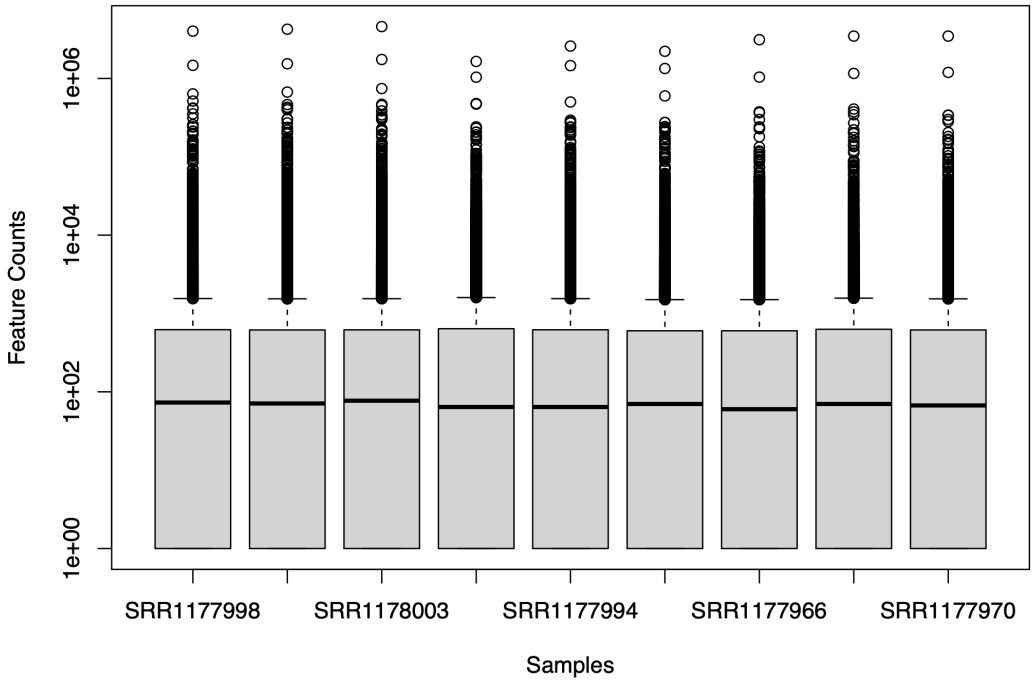


Figure 5: **Log Count distribution across samples.** The graph shows a slight positive skew in distribution, with similar median counts across all samples. The top whisker

represents the maximum value and black repeating circles represent the outliers found in the sample.

- **RNA-Seq Differential Expression**

  The DESeq2 package is designed for normalization, visualization, and differential analysis of high-dimensional count data. It makes use of empirical Bayes techniques to estimate priors for log fold change and dispersion, and to calculate posterior estimates for these quantities [7].

  In order to find the differentially expressed genes, first the counts matrix was combined with the read counts from a control dataset. Then, each sample was annotated for its mode of action (MOA) and chemical vehicle. There are three modes of action in our samples: AhR, CAR/PXR, and Cytotoxic. Each MOA was represented by 3 samples, and 3 controls were matched to each MOA through finding shared vehicle annotations. The AhR samples correspond to use of Beta-Naphthoflavone as the chemical, the CARPXR samples correspond to usage of Econazole, and the Cytotoxic samples correspond to Thioacetamide.

  Next, a separate counts matrix was made for each of the three MOAs, composed of 6 columns with 3 samples and 3 controls. DESeq2 was performed on the counts matrices for each MOA to produce a normalized counts matrix and differential expression statistics. The differentially expressed genes were filtered by p-adjusted values < 0.05. Genes with an adjusted p value < 0.05 were counted and the top 10 most significant genes were reported in Table 4. Lastly, three matrices were generated with the differential expression analysis produced from the RNA-Seq data along with three normalized count matrices that correspond to each treatment.

Results:

RNA-Seq Differential Expression Analysis

Differential expression analysis revealed 212 significantly differentially expressed genes (DEGs) in the AhR MOA, 1643 significantly differentially expressed genes in the CAR/PXR group and 3113 significantly differentially expressed genes in the Cytotoxic group (Table 4). Figure 6 displays histograms plotted, representing the significant $\log_2$ Fold Change range across each group. Each histogram centers around zero as

expected. The histograms show an almost even split of upregulated and downregulated differential expression in each treatment group.

The top ten genes from each group were selected and are shown in Table 4. Volcano plots were generated using all of the expression data for each treatment regardless of significance. Like in the histogram, all treatment groups show a fairly even distribution between up and down regulated genes. The red horizontal lines denote significantly differentially expressed genes. Upregulated and downregulated genes are represented in red and blue respectively.

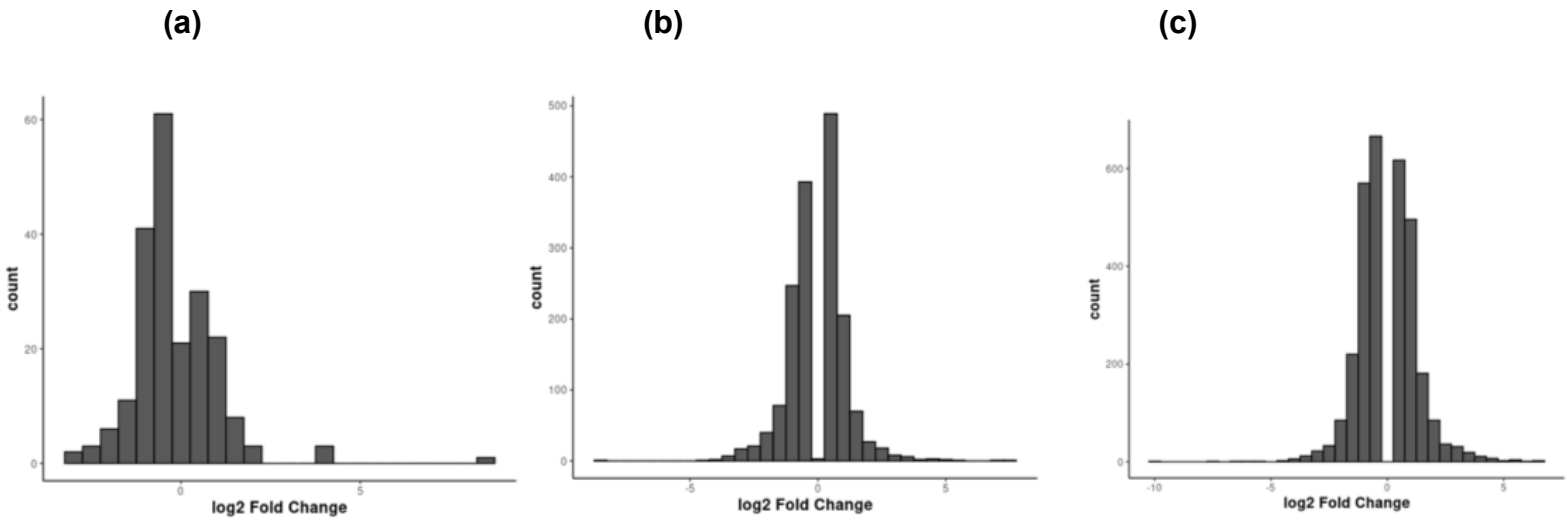


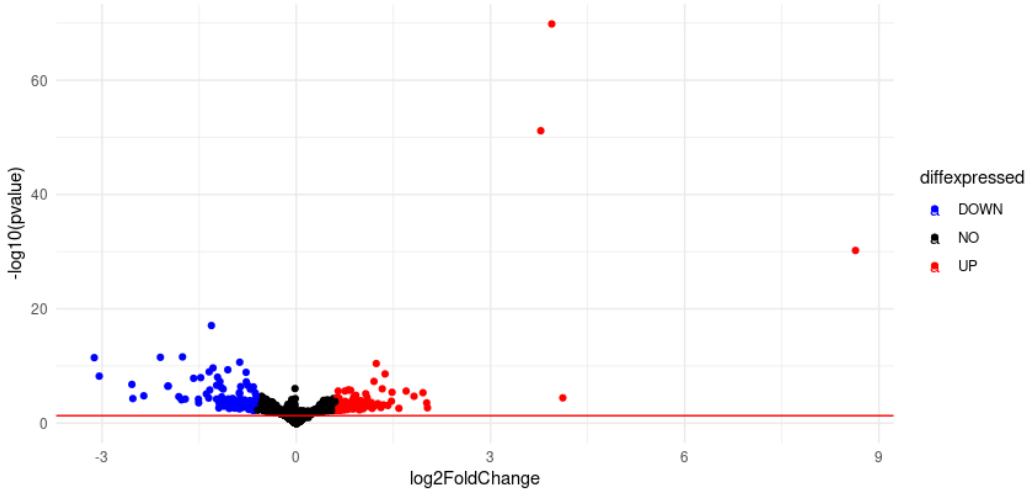
Figure 6: The distribution of log fold change values of the significant differentially expressed genes at adjusted p-value < 0.05 for (a) Beta-Naphthoflavone, (b) Econazole, (c) Thioacetamide.

Toxgroup significant gene count and top 10 most significant genes

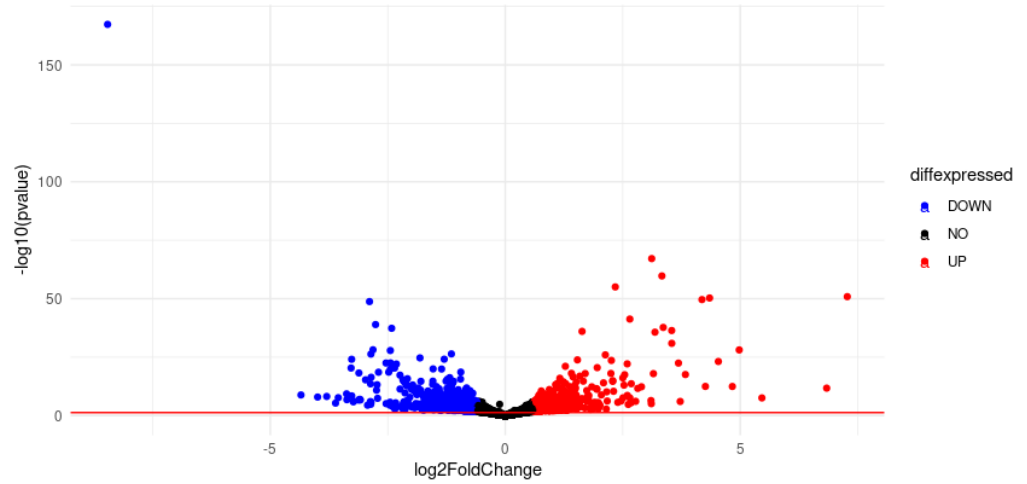
	AhR	CARPXR	Cytotoxic
Significant genes count	212	1643	3113
1st	Cpy1a2	Stac3	Zfand2a
2nd	Ugt1a7c	Ces2a	Vxn
3rd	Cyp1a1	Ces2j	Klf6
4th	Mgll	Cyp2c6v1	Maff
5th	Cacna2d4	Grin2c	Abcb1b
6th	Lrtm2	Aldh1a7	Plk2
7th	Abcd2	Cyp3a23-3a1	Gtse1
8th	Fkbp4	Cyp2c11	Mdm2
9th	Lfi47	Pla2g12a	Tes
10th	Hsp90aa1	Ppard	Ccng1

Table 4: Significant Differentially Expressed Genes. Number of DEG at p.adjust < 0.05 from RNA-seq analysis and top 10 DEGs from each treatment using DESeq2.

(a)



(b)



(c)

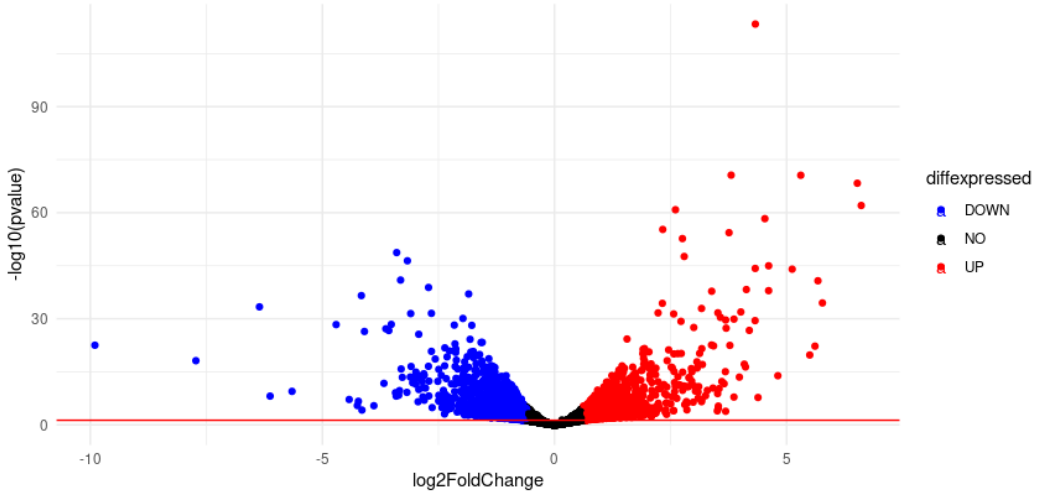


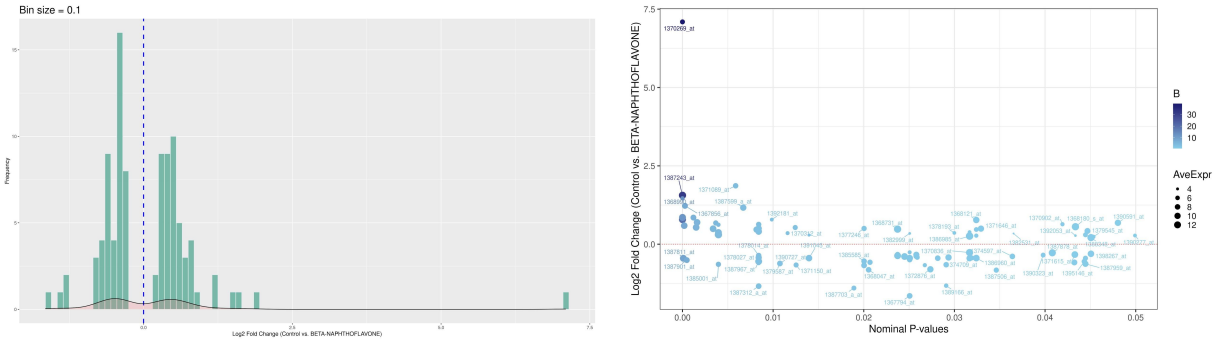
Figure 7: Volcano plots of differential expression data. Nominal values plotted against the Log2 Fold Change data of each gene. Red line shows the nominal significance cutoff of 0.05. Genes with the greatest log2 fold change for up (red) and downregulated (blue) expressions are shown. (a) AhR, (b) CAR/PXR, (c) Cytotoxic

- **Microarray Differential Expression with Limma**
  LIMMA[3] is a library for the analysis of gene expression microarray data, especially the use of linear models for analysing designed experiments and the assessment of differential expression. *LIMMA* provides the ability to analyse comparisons between many RNA targets simultaneously in arbitrary complicated designed experiments. Empirical Bayesian methods are used to provide stable results even when the number of arrays is small.

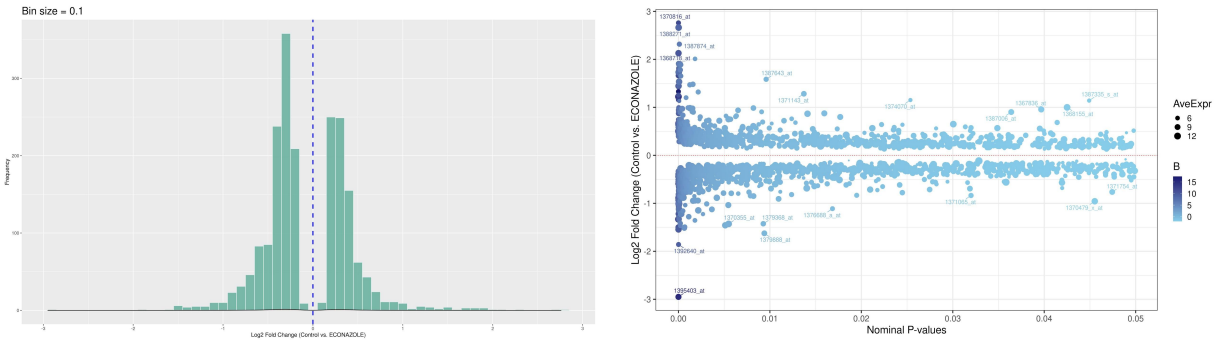
For each chemical treatment, a matrix-like data object containing log-ratios or log-expression values for a series of arrays, with rows corresponding to genes and columns to samples was fitted to a linear model, given the design matrix of the microarray experiment, with rows corresponding to arrays and columns to coefficients to be estimated.

Genes were then ranked in order of evidence for differential expression via an empirical Bayes method to squeeze the genewise-wise residual variances towards a common value(or towards a global trend). The degrees of freedom for the individual variances were increased to reflect the extra information gained from the empirical Bayes moderation, resulting in increased statistical power to detect differential expression.

(a)



(b)



(c)

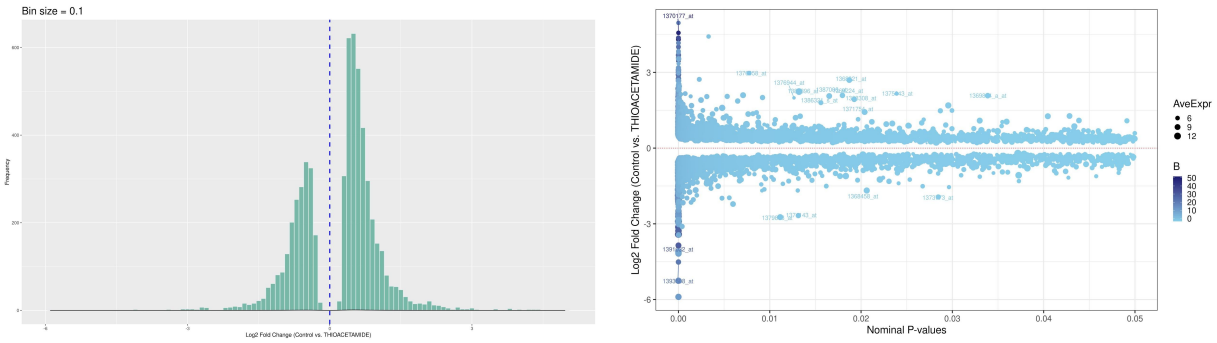


Figure 8: Histograms of fold change values and scatter plots of fold change vs. nominal-p value from the significant DE genes. (a): Beta-Naphthoflavone. (b): Econazol. (c): Thioacetamide. All histograms with densities shared the same bin size of 0.1. Zero bars were plotted as blue vertical dash lines. All scatter plots used dot sizes to indicate the average expression levels of probes. Color scales were introduced to indicate log-odds that the gene is differentially expressed; the darker the dot, the greater the odds. 0-Fold-Change lines were plotted as red horizontal dash lines.

Chemicals	Total number of DEGs (adjusted p-value < 0.05)	Top 10 DEGs
Beta-Naphthoflavone	73	Slc2a9 Akap1 Abcc3 Nqo1 Nectin3 Tmem150a Cyp1a2 Slc17a1 Cyb5a Atad3a
Econazol	1209	Slc13a4 Dync2h1 Slc25a10 Ppp2r5a Plxna2 Marcks Acot2 Sult2a6 Tppp

		Lnx2
Thioacetamide	3669	Cluap1 Slco2a1 Plxna2 Ddit3 Rogdi Fgf1 Thoc3 Spag9 Mbp Rpl10

Table 5: Summary of DEGs for each treatment on the microarray platform analyzed by *LIMMA*. For all probe IDs involved in each chemical treatment, a mapper was used to convert them into gene symbols and **REPEATS WERE ELIMINATED**. Total number of DEGs were reported and top 10 DEGs for each treatment were listed, sorted based on adjusted p-values.

- **Concordance Between Microarray and RNA-Seq DEGs**
  Results from the RNA-seq analysis and microarray analysis were used to calculate cross-platform concordances. Based on the reference paper, the concordance was described as

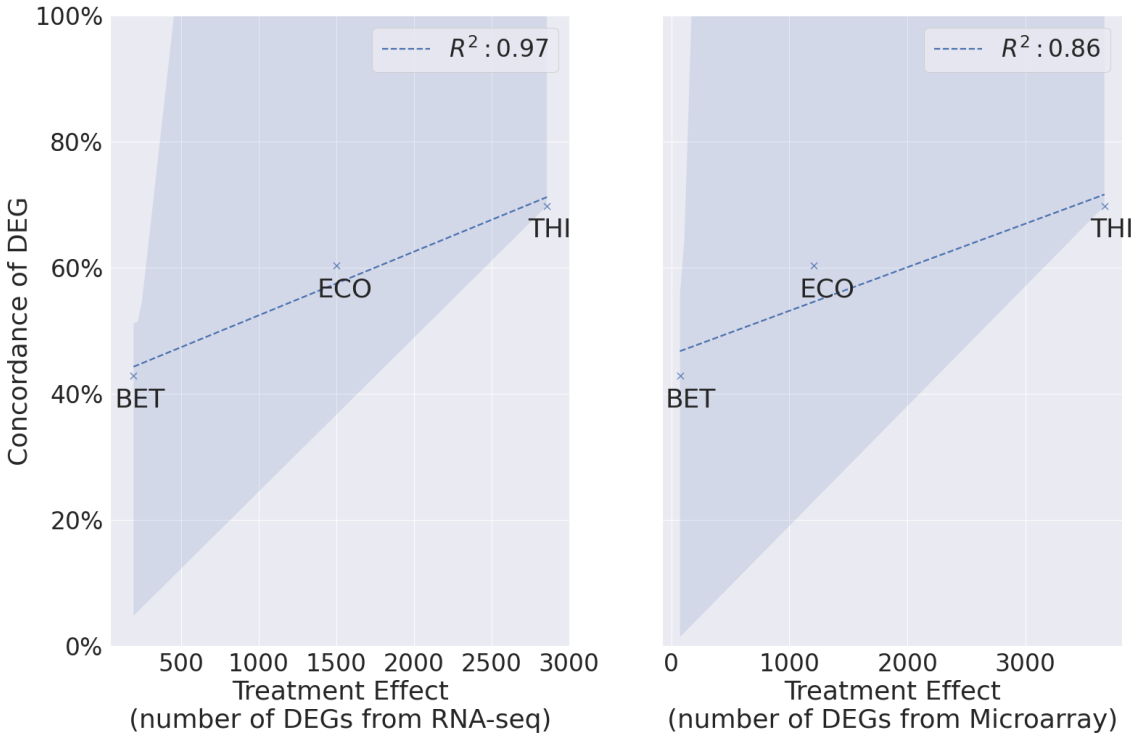
$$\frac{2 \times intersect(DEGs_{microarray}, DEGs_{RNA-Seq})}{DEGs_{microarray} + DEGs_{RNA-Seq}}$$

Intersect was calculated as an expectation of the intersection of independent events A and B, here, the number of DEGs from microarray and RNA-seq, not the observed intersection directly obtained from two sets. This background intersection increases as the sets get enlarged. Utilizing this property, the equation below was used to estimate the background-corrected intersection between two sets which may not be independent.

$$x + \frac{(n_1 - x) \times (n_2 - x)}{N - x} = n_0$$

The observed intersection, ***n0***, was constructed by two parts: ***x***, the background-corrected result and the true intersection. In this analysis, we assumed DEGs observed on two platforms were independent thus ***x*** was settled to 0.

(a).



(b)

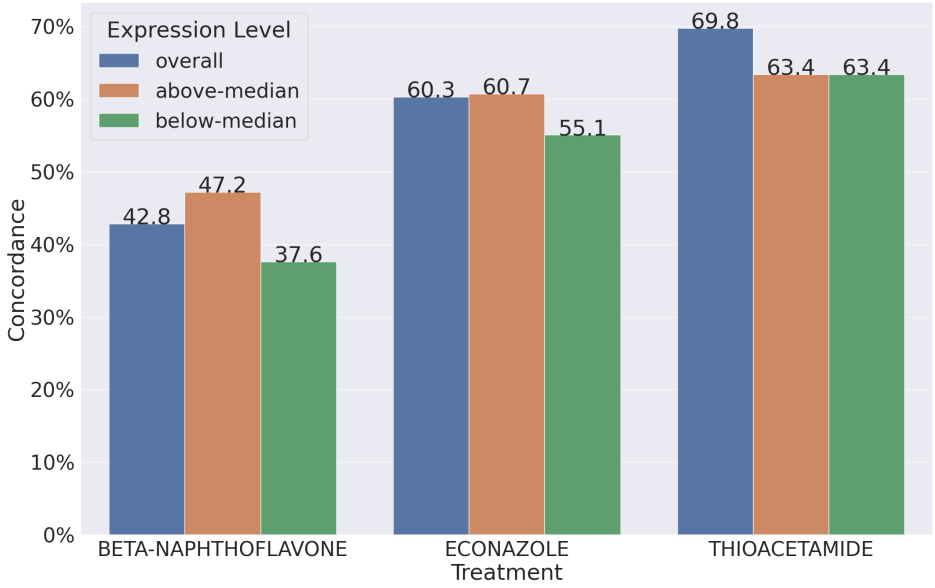


Figure 9: (a): Overall concordances against the number of DEGs from each analysis. (b): Concordance measures obtained for the overall DE gene list and the

above- and below-median subsets. Numerical values stand for accurate values of concordances. More details see supplement 1.

- Additional analysis was performed on both sets of differentially expressed gene data using the LIMMA and DESeq2 gene sets of differentially expressed genes. For genes differentially expressed by LIMMA in particular, the most differentially expressed genes were fed into the DAVID [7,8] functional annotation tool to locate enriched pathways within the differentially expressed gene sets for each chemical; Thioacetamide, Econazole, and Beta-Naphthoflavone. All three chemicals come from the peroxisome proliferator-activated receptor alpha, or PPARA mode of action (MOA). There are 45 gene enrichment terms noted in the reference paper; very few directly match to the ones located in the DAVID analysis [Table 6], which reports the top 10 terms for each chemical's analysis. The most notable term that does match is Metabolism of Xenobiotics by cytochrome P450, or just the Xenobiotics term in general--it is an almost direct match to the term "Xenobiotic metabolism signaling" located in the reference paper [1]. Most of the terms within the paper referenced "degradation" of assorted compounds, but there is no specific term from the DAVID analysis that matches any such term. It is suspected that the reference paper authors used a specialized enrichment analysis workflow using terms from the PubChem Database from NIH, as evidenced by the entry "Valine degradation I" being recognized by that database[10] .

Chemical	Top Ten DAVID Terms
ECONAZOLE	Circadian Rhythm
	Transcription Regulation
	Flavoprotein
	Metabolism of Xenobiotics
	Organelle Membrane
	Transit Peptide Mitochondria
	Glutathione Metabolism
	BRLZ (Basic Leucine Zipper)
	Pleckstrin Homology Domain
	Nucleotide Binding
THIOACETAMIDE	Acetylation

	Phosphoprotein
	Extracellular Exosome
	Ribosome Biogenesis in Eukaryotes
	Spliceosome
	Translation Initiation Factor Activition
	Ubiquitin Protein Ligase Binding
	Apoptosis
	Transferase
	Cell division
BETA-NAPHTHOFLAVONE	**Metabolism of Xenobiotics by cytochrome P450**
	Major Facilitator Superfamily Domain
	Response to organic Cyclic Compound
	NADP Metabolic Process
	Carbon metabolism
	Membrane
	Glycoprotein
	Mitochondrion
	ATP-binding
	Transmembrane Region

Table 6: Top ten distinct enrichment terms taken from DAVID analysis.

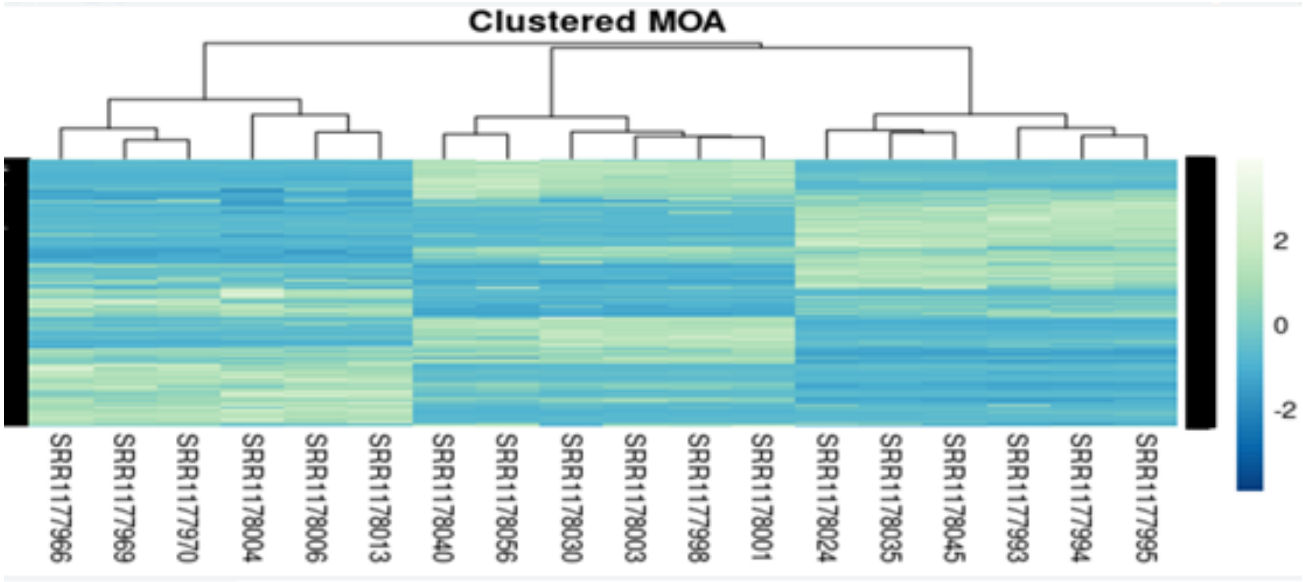


Figure 10: Clustered Heatmap of Gene Expression Data of selected Toxgroup.

Discussion:

In this project, we analyzed and compared the RNA-Seq and microarray data with reference to three modes of action, two receptor mediated processes (orphan nuclear hormone receptors (CAR/PXR) and aryl hydrocarbon receptor (AhR)) in the liver cells of rats treated with the three treatments Econazole, Beta-naphthoflavone, and Thioacetamide respectively. We calculated the concordance between treatment groups using the DEGs results produced from DESeq2 and Limma. The most significant DEGs were further examined using the DAVID annotation tool to identify enriched pathways in each treatment group.

The primary finding of the paper was that the concordance of RNA-Seq and microarray expression estimates depends on a number of factors, including biological effect size and gene expression level. In this study, it was concluded that (i) the concordance between array and sequencing platforms for detecting the number of DEGs was positively correlated with the extensive perturbation elicited by the treatment (Fig 9a), and (ii) gene expression–based predictive models generated from RNA-seq and microarray data were similar[1]. When investigating two similar biological conditions, a lower concordance is expected as the reference paper demonstrated between the two platforms and the discrepancy is derived from the measurement of the low expressed genes for which RNA-seq performs better. This finding supports the assumption made for the calculation of concordance, that the observed intersection was a result of background reads and true intersections, though the estimation of background reads was not mentioned in the reference paper. When set sizes increased, the theoretical true intersection increased as well instead of the observation intersection. The background-corrected results produced stronger “stopping effects” on this evaluation. In

summary, the treatment effects and the abundance of the genes dictate many observations in RNA-seq and its comparison with the microarray followed a systematic trend correlated to the strength of perturbation of the samples.

A heatmap [Figure 11] examining expression in the DESeq2 gene expression data was produced and served as a reasonable guide in determining that samples exposed to different chemicals within the PPARA group affected expression in similar fashions. Three distinct clustering regions are noted for Econazole, Beta-naphthoflavone, and Thioacetamide samples, and are in fact identified in that order, from left to right. While the exact type of clustering analysis performed by the reference paper [1] was not replicated, this was a reasonable approximation that helps support the overall approach of this study, and supports the idea proposed that the selected chemical for analysis impacted gene expression through similar methods .

Conclusion:

Overall, our study is in agreement with findings Wang et al. in evaluating the performance of RNA-seq and Microarray techniques on gene expression analysis. Different enrichment pathways were found for the three MOAs and chemical treatments. As a result, these pathways could be potential targets for different clinical applications. In general, we observed that the more significant DEGs found, the greater the concordance between the treatments investigated. Additionally, the greater the magnitude of differential expression in a gene, the more likely it is to be selected as a significant DEG in both RNA-Seq and microarray analyses. Due to limitations in the availability of gene enrichment tools, as well as limitations in the level of expertise in interpreting the links between pathways, it is unclear how many pathways from the reference paper were correctly identified in this study. There were likely differences in the databases used for locating annotations from the paper, and an added level of error from DAVID due to redundancy or other limitations in annotation curation.

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. 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.
3. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7), e47.
4. Yang Liao, Gordon K. Smyth, Wei Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, *Bioinformatics*, Volume 30, Issue 7, 1 April 2014, Pages 923–930. doi.org/10.1093/bioinformatics/btt656
5. 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
6. 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.
7. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. ***Nature Protoc. 2009;4(1):44-57.*** [PubMed]
8. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. ***Nucleic Acids Res. 2009;37(1):1-13.*** [PubMed]
9. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR "STAR: ultrafast universal RNA-seq aligner." Bioinformatics. 2013 Jan 1;29(1):15-21. 2012 Oct 25. https://doi.org/10.1093/bioinformatics/bts635
10. National Center for Biotechnology Information. "PubChem Pathway Summary for Pathway VALDEG-PWY, valine degradation I, Source: BioCyc" *PubChem*, https://pubchem.ncbi.nlm.nih.gov/pathway/BioCyc:CAULONA1000_VALDEG-PWY. Accessed 7 April, 2021.

Supplement:

1.

```
for treatment BETA-NAPHTHOFLAVONE:
overall concordance: 0.42835219065960517
# of DEGs in RNA-seq: 195
# of DEGs in microarray: 73
above-median concordance: 0.47232174395720844
# of DEGs in RNA-seq: 104
# of DEGs in microarray: 45
below-median concordance: 0.37592592592592594
# of DEGs in RNA-seq: 91
# of DEGs in microarray: 29


for treatment ECONAZOLE:
overall concordance: 0.6033665758871959
# of DEGs in RNA-seq: 1500
# of DEGs in microarray: 1209
above-median concordance: 0.6072376316278756
# of DEGs in RNA-seq: 781
# of DEGs in microarray: 736
below-median concordance: 0.5508227231740307
# of DEGs in RNA-seq: 719
# of DEGs in microarray: 531


for treatment THIOACETAMIDE:
overall concordance: 0.698062428852437
# of DEGs in RNA-seq: 2857
# of DEGs in microarray: 3669
above-median concordance: 0.6341895543645112
# of DEGs in RNA-seq: 1496
# of DEGs in microarray: 2219
below-median concordance: 0.6340892291375878
# of DEGs in RNA-seq: 1363
# of DEGs in microarray: 1648
```