Project 3 Write-up

Abhishek Thakar (Analyst), Allison Nau (Programmer), Mae Rose Gott (Biologist), and Sheila Yee (Data Curator)

Introduction

While RNA-sequencing and microarrays are common methods used to measure whole transcriptome gene expression, there remains a need to determine a concordance between these two technologies in order to establish confidence in their usage in clinical and regulatory applications. In order to define this concordance, Wang et al. (2014) compared liver gene expression profiles in rats that had been treated with different chemicals with varying modes of action (MOA) using Illumina RNA-seq and Affymetrix microarrays. The authors determined that there was a positive correlation between RNA-seq and microarrays in terms of differentially expressed genes (DEGs) as a result of treatment effect size, gene-expression abundance, and the biological complexity of the MOA. However, RNA-seq was found to perform better at identifying DEGs.

In this analysis, we sought to reproduce the concordance between RNA-seq and microarray data from Wang et al. By using a subset of samples from the original data, we focused on three different chemical treatments, each of which had a distinct MOA. Differential expression was performed on RNA-seq and microarray data. A comparative analysis and pathway enrichment analysis was subsequently conducted between the results of these two methods.

Data

Illumina RNA-seq (performed on HiScanSQ or HiSeq2000) and Affymetrix microarray (performed on Affymetrix whole genome GeneChip® Rat Genome 230 2.0 Array) data was generated from rat liver samples and were accessible using NCBI accessions SRP039021, GSE55347, and GSE47875. For the purpose of this project, a subset of samples was selected for analysis belonging to toxgroup 1 that had been generated from Illumina raw RNA-seq data. These samples had been exposed to different degrees of chemical perturbation, each representing a different MOA. Each chemical was administered to male Sprague-Dawley rats with the maximum tolerated dose (MTD) for five days (Wang et al., 2014). The RNA samples were obtained from the NTP DrugMatrix Frozen Tissue Library (NTP, https://ntp.niehs.nih.gov/drugmatrix/index.html).

In toxgroup 1, the three different chemical agents and their known associated MOAs were as follows: chloroform and toxicity, 3-methylcholanthrene and the aryl hydrocarbon receptor (AhR), and clotrimazole and the orphan nuclear hormone receptors (CAR/PXR). In total, there were nine treatment samples with the three treatment conditions having three replicates each and six control samples. For the treatment samples, the length of each paired end read ranged from 100 - 101 base pairs (bp) long. The average number of reads in each sample

was around 18.3 million with the range being 16 - 21.8 million reads. The average percentage of reads aligned per sample was 84.1%.

The Affymetrix microarray data used for comparison with RNA-seq data had already been processed by the instructor.

Methods

Raw paired end RNA-seq data from toxgroup 1 were selected for analysis. In order to access the quality of the raw sequence data, the sample fastq files were processed using FastQC v0.11.7 (Andrews, 2010). Next, each of the samples was aligned to a reference rat genome using the STAR aligner v2.6.0c (Dobin et al., 2013). Lastly, MultiQC v1.6 was run on the results from both FastQC and STAR in order to generate a summary report containing quality control metrics and alignment statistics for all samples (Ewels et al., 2016).

Next, we ran featureCounts from the module subread v1.6.2 on each of the aligned and sorted RNAseq bam files, using rn4_refGene_20180308.gtf as the annotated reference (Liao 2013, Liao 2014). featureCounts takes advantage of chromosome hashing, feature blocking, and multithreaded processing to improve speed and memory efficiency compared to earlier read assignment algorithms. We then ran multiqc, from the module multiqc available through Anaconda, on each of the resulting count files (Ewels 2016). MultiQC performs quality control on all samples at once, synthesizing the results into a single report to help identify global trends and biases. Both steps took less than 10 minutes each as submitted jobs on a shared high performance computing cluster. Then, we combined the feature count information into a single CSV file and created boxplots of count distributions using R packages tidyverse and janitor (<2 minutes on desktop; Wickham 2019, Firke 2021).

We then loaded the count information for each of our experimental samples, the provided count data for all the control samples, and the metadata for each sample into R. Data was loaded and some processing was done using tidyverse, hardhat, and stringr (Wickham 2019, Davis 2020, Hadley 2019). For each experimental group (AhR, CAR/PXR, Cytotoxic), we selected the experimental samples for that group and the relevant controls (which matched the “vehicle” of the experimental group), then ran differential expression analysis using DESeq2 (Love 2014, Zhu 2018). DESeq2 performs differential expression analysis on RNAseq count data. RNAseq expression data requires special statistical considerations as RNAseq has discrete counts, large dynamic ranges, outliers, and generally has small replicate numbers. DEseq2 represents an improvement on earlier differential expression algorithms by using statistical techniques to evaluate the strength of the differential expression results, such as shrinkage estimation, improving the stability and interpretability of the results. Normalized counts were also calculated using DESeq2. Genes were determined to be differentially expressed if adjusted p-value was less than 0.05. Differential expression results and normalized counts were exported for use in downstream analysis. After packages were installed, DESeq2 for the three experimental groups took less than 2 minutes total when run locally.

The microarray data was provided for in the corresponding project directory as a .csv file as well as the RMA expression matrix that is used for mapping the differential expression. Limma, a bioconductor analytical package, was implemented to determine differential expression analysis for microarrays. The example `run_limma.R` script provided a basis to use limma on tox group 1 (AhR:3ME, CAR/PXR: CLO, Cytotoxic: CHL). The genes were filtered based on the threshold of adjusted p-value < 0.05 for each chemical to create a significant genes list. Using these lists and logFC values, histograms for each chemical as well as volcano plots were produced to help identify the most significant genes. The concordance of RNA-Seq and microarray gene expression was compared using a refSeq-to-probe id as a mapping reference to create concordance plots. Concordance is calculated as the intersection between set 1 (microarray of significantly differentially expressed genes) and set 2 (RNA-seq of significantly differentially expressed genes) with the agreement between both sets determined by fold change (positive or negative comparison). The equation that was used can be seen in the paper under the cross-platform concordance analysis section for greater detail on the methodology. However, with larger set sizes, the background intersection increases and needs correction. The corrected intersecting values between two sets were calculated and then the corrected concordance was determined for all genes, above the median, below the median, per chemical.

A heatmap was generated in RStudio using the DESeq norm counts. Data from all three experimental groups were merged together by gene and the final data table was used for the heatmap after filtering. Filtering was applied by using all of the genes where the coefficient of variance was less than 0.189 and then filtering for higher than average means.

The top ten differentially expressed genes for each MOA were then run through DAVID (Database for Annotation, Visualization, and Integrated Discovery) v6.8 for KEGG pathway analysis.

Results

The raw RNA-seq data for the treatment samples were processed with FastQC, the STAR aligner, and MultiQC. General statistics from the alignment and quality of samples can be observed in Table 1. With the exception of sample SRR1178036, all samples showed a high percentage of aligned reads (Table 1). The majority of mean quality scores across the entire range of base positions are situated in the green zone which is indicative of a high Phred score (Figure 1A). Three samples (in green) passed the Phred score threshold, marking greater confidence in the base call during sequencing, while fifteen samples (in red) failed to meet the threshold. However, all of them have reliable Phred scores up until the 60 base pair position. With the exception of sample SRR1178036_2, all of the samples have low percentage per base N content (Figure 1B). Most of the samples (16/18 samples highlighted in red) failed to have good sequence duplication levels (Figure 1D). This could possibly be due to an enrichment bias (such as PCR over amplification or high abundance of certain genes). A higher rate of duplication is to be expected with RNA-seq.

Sample SRR1178036_2 was observed to be of lower quality than the other samples. Firstly, it was inconsistent relative to other samples as observed with its lower percentage of duplicated reads (Table 1). Additionally, at base positions 20 - 22, the mean quality score across base positions was markedly lower (Figure 1A) and percentage of N content was relatively higher than the other samples (Figure 1B). The number of uniquely mapped reads in sample SRR1178036 was also lower relative to the other samples that all had mapped reads over 83%, presumably due to SRR1178036_2 (Table 1 and Figure 2). Its per sequence GC content was also dissimilar relative to the other samples (Figure 1C).

Table 1. General statistics produced by MultiQC. Samples were perturbed by varying chemical treatments with different modes of action (MOA). The three different chemical treatments used in toxgroup 3 are chloroform, 3-Methylcholanthrene, and clotrimazole that had cytotoxicity, aryl hydrocarbon receptor (AhR), and orphan nuclear hormone receptors (CAR/PXR) as MOAs, respectively. From the STAR alignment, this includes the percentage of reads that were uniquely aligned against the reference genome (% aligned), the number of uniquely mapped reads in millions (M aligned). Results are highlighted in a darker color. From FastQC, this includes the percentage of duplicate sequences (% dups), the average percentage of GC content (% GC), and the total number of sequences in millions (M Seqs). Results are highlighted in a lighter color.

Sample name	MOA	Chemical treatment	% aligned	M aligned	% dups	% GC	M Seqs
SRR1177987	Cytotoxic	Chloroform	84.8%	14.8			
SRR1177987_1					56.4%	49%	17.4
SRR1177987_2					53.3%	49%	17.4
SRR1177988	Cytotoxic	Chloroform	85.3%	15.7			
SRR1177988_1					55.1%	49%	18.4
SRR1177988_2					51.9%	49%	18.4
SRR1177989	Cytotoxic	Chloroform	83.5%	15.7			
SRR1177989_1					50.4%	49%	18.8
SRR1177989_2					46.9%	49%	18.8
SRR1177997	AhR	3-Methylcholanthrene	89.2%	17.6			
SRR1177997_1					59.6%	49%	19.7
SRR1177997_2					58.6%	49%	19.7
SRR1177999	AhR	3-Methylcholanthrene	88.7%	19.4			
SRR1177999_1					60.2%	49%	21.8
SRR1177999_2					58.9%	49%	21.8
SRR1178002	AhR	3-Methylcholanthrene	89.2%	16.8			
SRR1178002_1					58.5%	49%	18.8
SRR1178002_2					57.6%	49%	18.8
SRR1178020	CAR/PXR	Clotrimazole	83.6%	13.4			
SRR1178020_1					54%	48%	16.0
SRR1178020_2					51.6%	49%	16.0
SRR1178036	CAR/PXR	Clotrimazole	67.8%	11.4			
SRR1178036_1					55.5%	48%	16.9
SRR1178036_2					39.4%	48%	16.9
SRR1178046	CAR/PXR	Clotrimazole	85.4%	15.1			
SRR1178046_1					54.7%	48%	17.7
SRR1178046_2					53.5%	49%	17.7

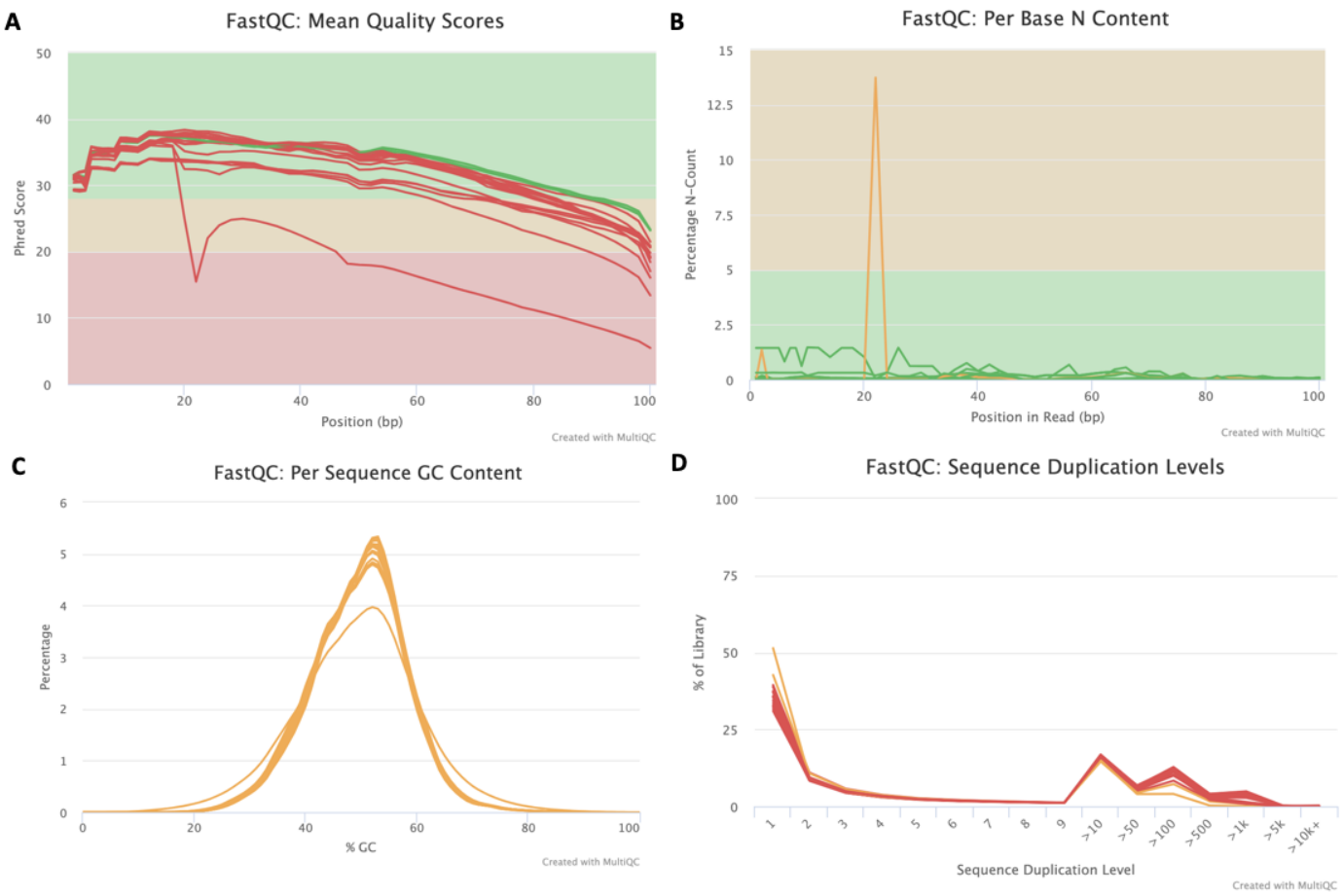

Figure 1. MultiQC report for FastQC. A) Mean quality score for each sample. It displays the mean quality value across the range of base positions in a read. B) The percentage of base calls at each given position in which an N was called, therefore meaning that the base is non-identified. C) The percentage of GC content per sequence in each sample. D) The sequence duplication level for each sample.

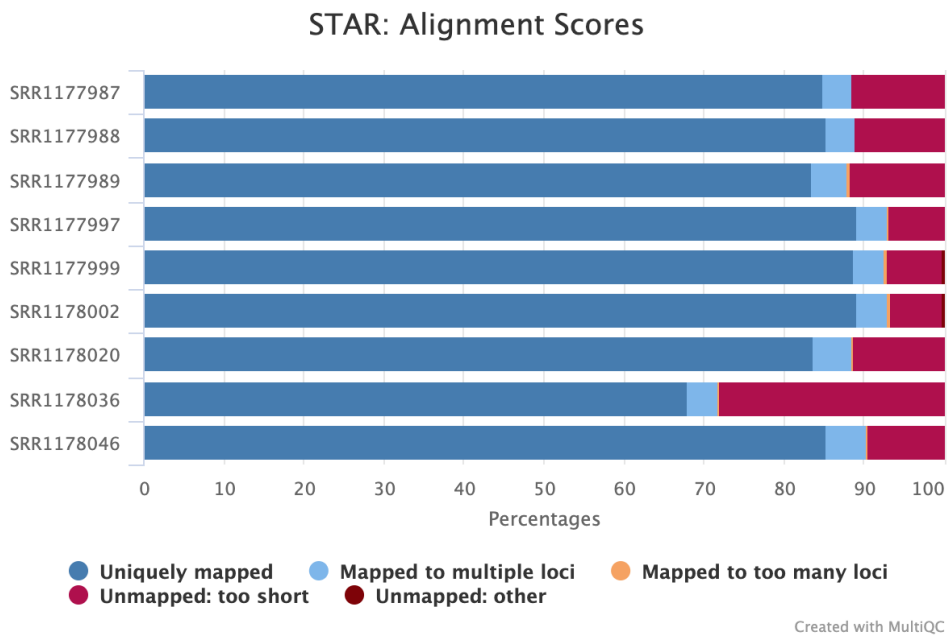

Figure 2. MultiQC report for STAR alignment. Alignment scores were determined after aligning the samples to a reference rat genome. The percentage of mapped and unmapped reads are displayed.

After running featureCounts on the RNAseq data, proportion of reads assigned, unassigned, and multimapped was roughly consistent across the 9 experimental samples, with between 59.2% -62.7% of reads assigned, 9.8%-13.6% of reads unassigned due to multimapping, 21.2%-23.9% unassigned due to no features being identified, and 4.7%-6.2% of reads unassigned due to ambiguity (Figure 3). Distribution of counts varied across samples, with the cytotoxic sample SRR1177989 having the widest distribution of counts (Figure 4). Distribution of counts for the AhR experimental group were relatively uniform when compared to the distribution of counts for CAR/PXR and cytotoxic experimental groups.

The top 10 differentially expressed genes, by adjusted p-value, were identified for each experimental group for the RNAseq data (Table 2). The number of differentially expressed genes for each experimental group with a p-adjusted value less than 0.05 were 530, 1140, and 1740 for AhR, CAR/PXR, and cytotoxic experimental groups respectively. AhR (aryl hydrocarbon receptor, 3-methylcholanthrene, 3ME) has similar results to Figure 2A in Wang et al, but we found fewer significant differentially expressed cytotoxic genes (cytotoxicity, chloroform, CHL). We found 530 differentially expressed genes in 3ME and 1740 in CHL in comparison to 578 and 3850 respectively in Wang et al supplementary table 9. (Note, the numbers in Wang et al Figure 2A and supplementary table 9 do not match). For the significant differentially expressed genes (α<0.05), log2 fold change distribution varied across experimental group (Figure 5A, 5C, 5E). log2 fold change vs nominal p-value also varied across experimental groups (Figure 5B, 5D, 5F).

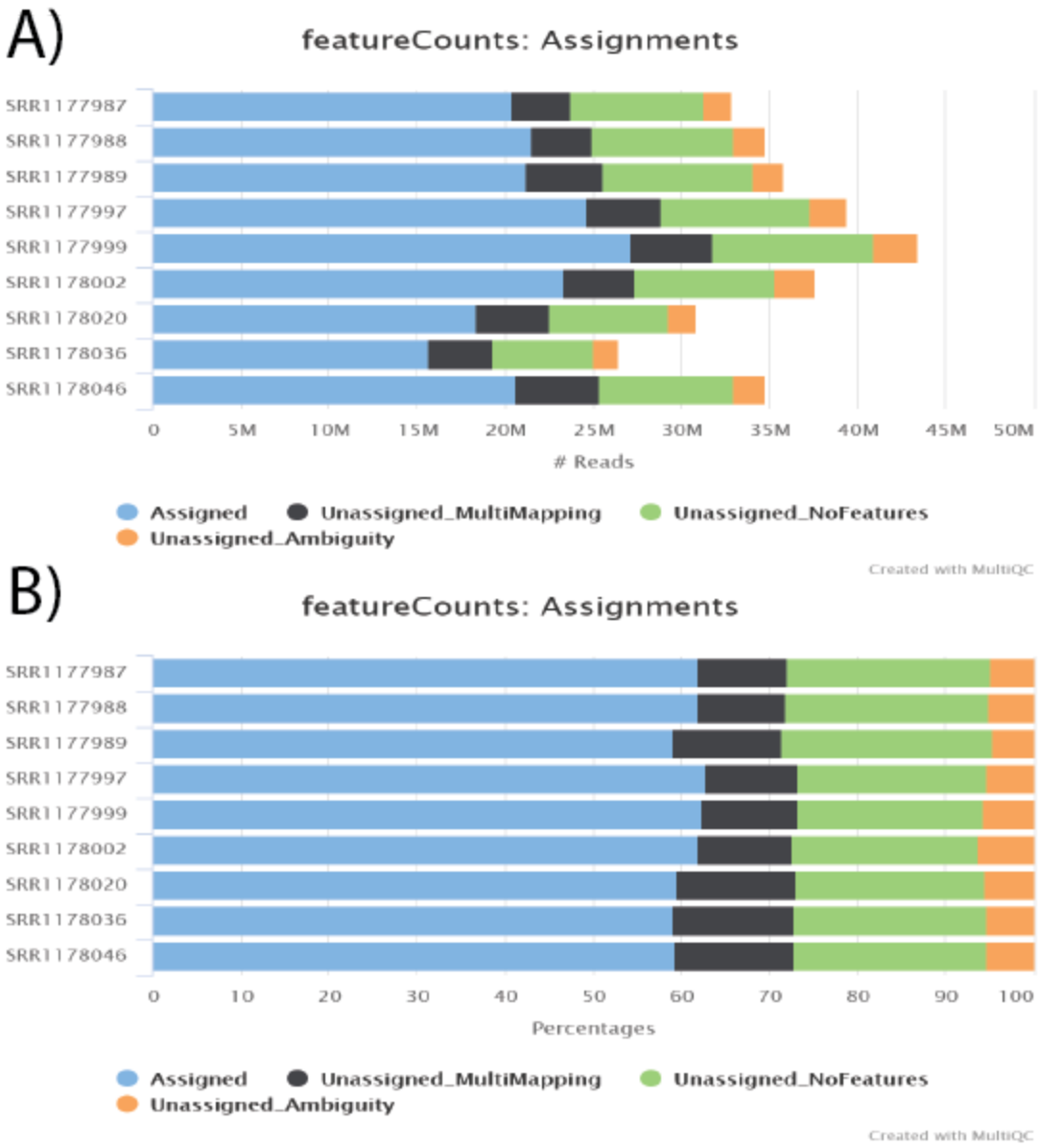

Figure 3. A) Number of reads and B) Percentage of reads that were assigned, multimapped, ambiguous, and unassigned for each sample.

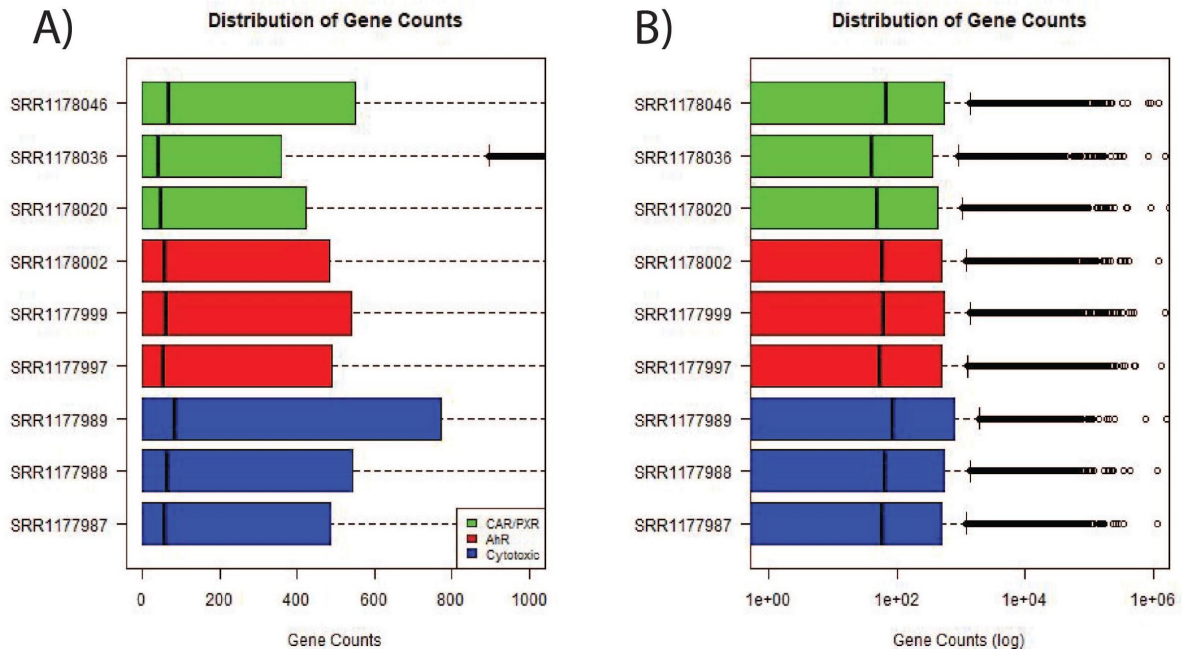

Figure 4. A) Read count distribution for each gene for each sample. B) Read count distribution for each gene for each sample, log scale.

Table 2. Table of top 10 differentially expressed genes from each experimental group by adjusted p-value ($\alpha < 0.05$).

AhR_Geneid	AhR_padj	CAR/PXR_Geneid	CAR/PXR_padj	Cytotoxic_Geneid	Cytotoxic_padj
NM_012541	1.36E-79	NM_001010921	1.63E-114	NM_203512	3.56E-126
NM_130407	4.52E-19	NM_080581	1.63E-109	NM_001257095	6.65E-96
NM_022521	7.49E-11	NM_013105	9.36E-104	NM_080581	1.12E-45
NR_046239	8.51E-10	NM_131903	6.77E-66	NM_023978	1.56E-40
NM_134329	1.56E-09	NM_173295	1.70E-65	NM_012844	1.60E-37
NM_175761	3.07E-09	NM_012844	7.72E-65	NM_013215	2.15E-36
NM_012608	8.68E-09	NM_017272	4.26E-63	NM_139115	5.29E-35
NM_053883	1.19E-08	NM_013215	6.50E-58	NM_012540	1.70E-29
NM_024351	1.19E-08	NM_001134844	3.27E-52	NM_001013098	1.16E-28
NM_022866	2.35E-08	NM_133586	1.84E-49	NM_001010921	4.33E-26

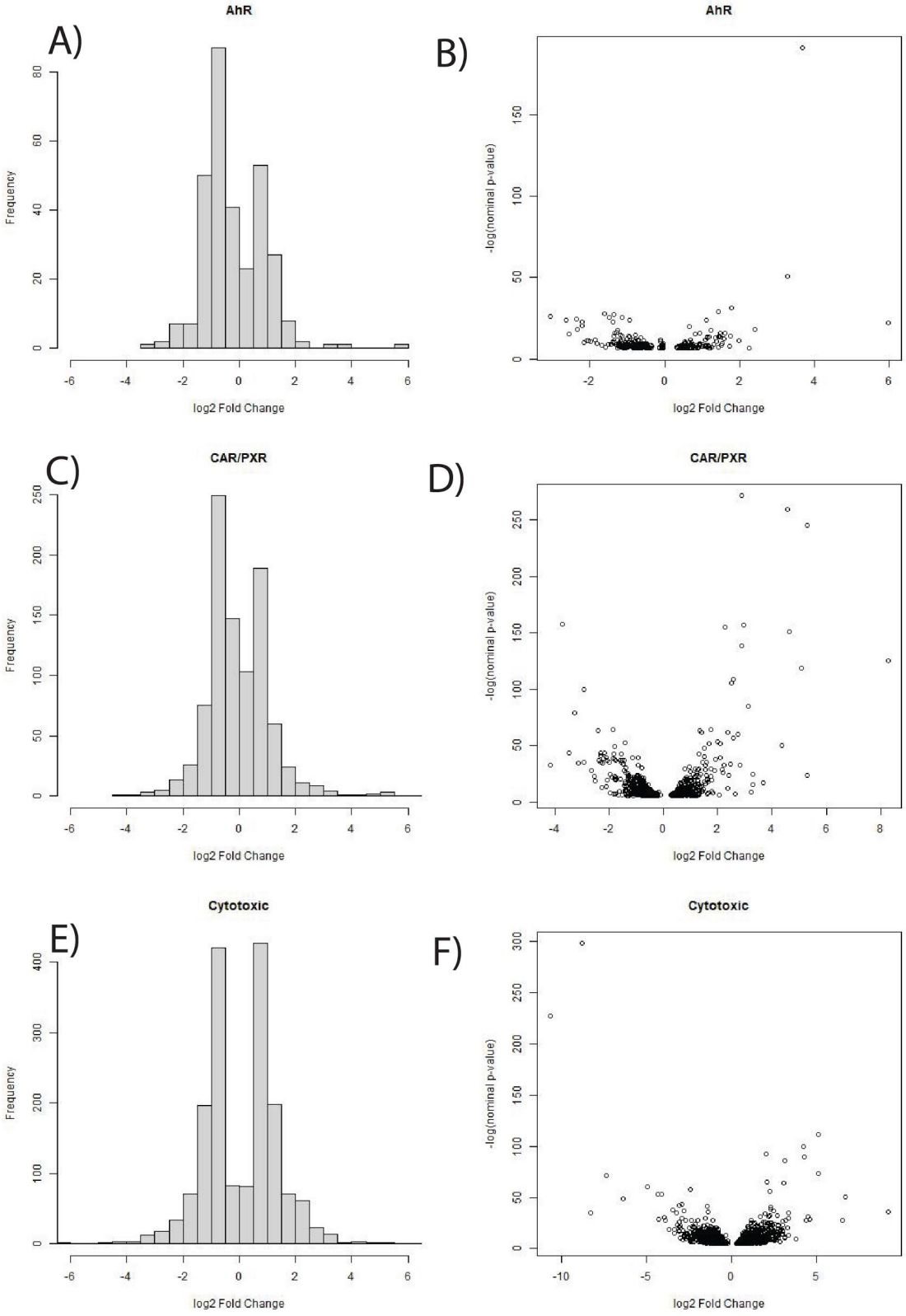

Figure 5. Log2 fold change for significant differentially expressed genes ($\alpha < 0.05$). Histogram of log2 fold change for significant differentially expressed genes for experimental groups A)

AhR, C) CAR/PXR, and E) Cytotoxic experimental groups. Scatter plots for log2 fold change compared with -log nominal p-value for B) AhR, D) CAR/PXR, and F) Cytotoxic experimental groups.

Looking at the microarray analysis, we can see the numbers of significant differentially expressed genes as compared to the RNA-seq analysis in Table P2, we see an increasing trend in significant DEGs in both RNA-seq and microarray with the largest group in chloroform. We also see the highest log fold change in the cytotoxic MOA when looking at the top 10 DEGs from both RNA-seq and microarray analysis. In Figure 6, we see that in A) the 3-methylcholanthrene diagram shows a center between 0 and 1, indicating that there is a larger overall upregulation in the differentially expressed genes. However, it is important to note the very small sample size that fulfilled the filtering threshold of p-value <0.05. Looking at the volcano plots for clotrimazole and chloroform in E) and F), we see a well formed shape with highlighted points that show the most up and downregulated genes with statistical significance.

Looking at the concordance plots of each of the three chemicals (Figure 7), we see a linear trend in increasing number of DEGs and concordance (%). There was a higher concordance and greater amount of significant genes for chloroform than for 3-methylcholanthrene and clotrimazole. Clotrimazole appeared to be an intermediate between 3-methylcholanthrene and chloroform.

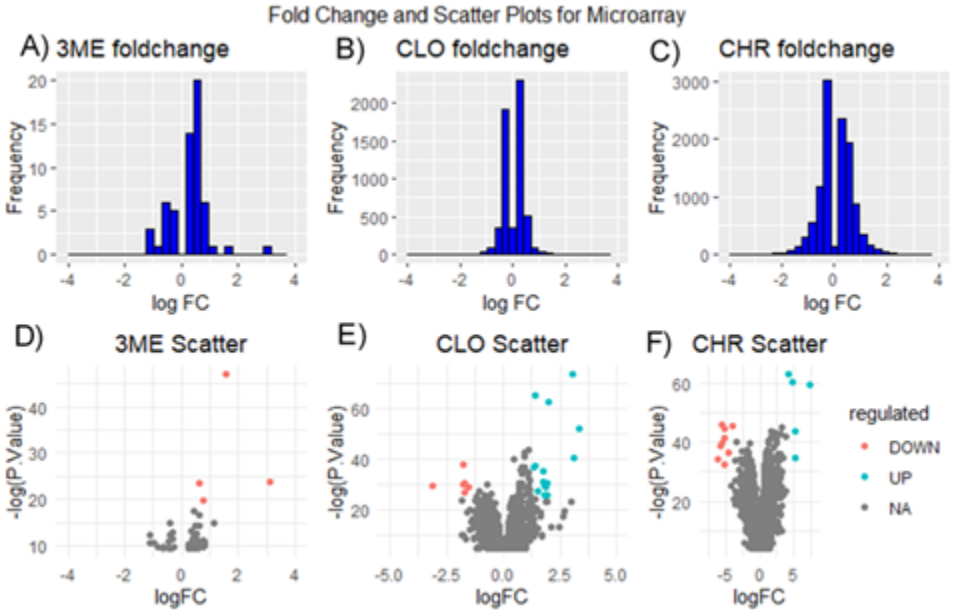

Figure 6. Log fold change for significant differentially expressed genes from limma ($\alpha < 0.05$). Histogram of log fold change for significant differentially expressed genes for MOA group A) AhR:3ME, B) CAR/PXR:CLO, and C) Cytotoxic:CHR experimental groups. Scatter plots for log fold change compared with log(p-value) for D) AhR, E) CAR/PXR, and F) Cytotoxic experimental groups. Highlighted dots signify self determined logFC and log(p-value) thresholds to help clarity.

Table 3. Number of differentially expressed genes for each experimental MOA group by adjusted p-value ($\alpha < 0.05$).

Chemical	RNA-seq significant DEGs	Microarray significant DEGs
3ME	530	58
CLO	1140	5803
CHL	1740	11407

Table 4. Table of top 10 differentially expressed genes from AhR group by adjusted p-value ($\alpha < 0.05$) and fold change.

AhR differentially expressed genes					
RNA-Seq DE data			Microarray DE data		
gene symbols	log2FoldChange	padj	gene symbols	logFC	adj.P.Val
Cyp1a2	3.680469139	1.36E-79	**Cyp1a2**	1.536513	1.33E-16
Ugt1a7c	3.274370997	4.52E-19	**Cyp1a1**	3.097586	6.38E-07
Oat	1.769382178	7.49E-11	**Ugt1a9**	0.634759	6.84E-07
NA	1.434247893	8.51E-10	**Ugt1a9**	0.777306	1.99E-05
Adh7	-1.614664926	1.56E-09	**Pon3**	0.439345	0.000181
Hsp90aa1	-1.349402758	3.07E-09	**Cyp2a1**	0.60019	0.000284
Mme	-3.055765906	8.68E-09	**G6pd**	1.117664	0.001494
Dusp6	-1.131724314	1.19E-08	**Ptprs**	-0.43005	0.001494
Hspa8	-1.489133627	1.19E-08	**Gsta4**	0.492776	0.001989
Slc13a3	-2.368625585	2.35E-08	**#N/A**	0.608276	0.001989

Table 5. Table of top 10 differentially expressed genes from each CAR/PXR group by adjusted p-value ($\alpha < 0.05$) and fold change.

CAR/PXR differentially expressed genes					
RNA-Seq DE data			Microarray DE data		
gene symbols	log2FoldChange	padj	gene symbols	logFC	adj.P.Val
Gsta5	2.8986221	1.63E-114	Cyp2b1	3.018918	3.75E-28
Abcc3	4.55976132	1.63E-109	Cyp3a23/3a1	1.375837	6.06E-25
Cyp3a23/3a1	5.290784056	9.36E-104	Ugt2b1	1.968334	5.71E-24
Sult2a1	-3.697760954	6.77E-66	Ces2c	3.341845	1.96E-19
Ugt2b1	2.972724186	1.70E-65	Ugt1a9	1.122778	5.49E-16
Ephx1	2.282735715	7.72E-65	Ugt1a9	0.920436	1.21E-15
Aldh1a7	4.652455235	4.26E-63	#N/A	0.915251	5.25E-15
Akr7a3	2.884161155	6.50E-58	Abcc3	3.095757	9.66E-15
Cyp2b1	8.276719466	3.27E-52	Cyp2c6v1	0.484028	1.66E-14
Ces2c	5.09777124	1.84E-49	Tsc22d1	-1.75325	9.39E-14

Table 6. Table of top 10 differentially expressed genes from Cytotoxic group by adjusted p-value ($\alpha < 0.05$) and fold change.

Cytotoxic differentially expressed genes					
RNA-Seq DE data			Microarray DE data		
gene symbols	log2FoldChange	padj	gene symbols	logFC	adj.P.Val
RGD1566134	-8.780120896	3.56E-126	Abcc3	4.182743	1.68E-23
NA	-10.64010351	6.65E-96	Gstp1	4.72061	1.23E-22
Abcc3	5.090745375	1.12E-45	Akr1b8	7.364298	1.32E-22
Per3	4.225594795	1.56E-40	Car3	-5.67521	8.29E-17
Ephx1	2.026094116	1.60E-37	#N/A	-4.10915	1.20E-16
Akr7a3	4.301546695	2.15E-36	Akr7a3	3.271531	1.20E-16
Coro6	3.148263377	5.29E-35	RGD1566134	-5.30724	2.31E-16
Cyp1a1	5.117056466	1.70E-29	Dap	1.576835	3.94E-16
Dhrs7l1	-7.375994861	1.16E-28	Abcb1b	5.06929	3.94E-16
Gsta5	2.103250319	4.33E-26	Krt8	2.125194	4.37E-16

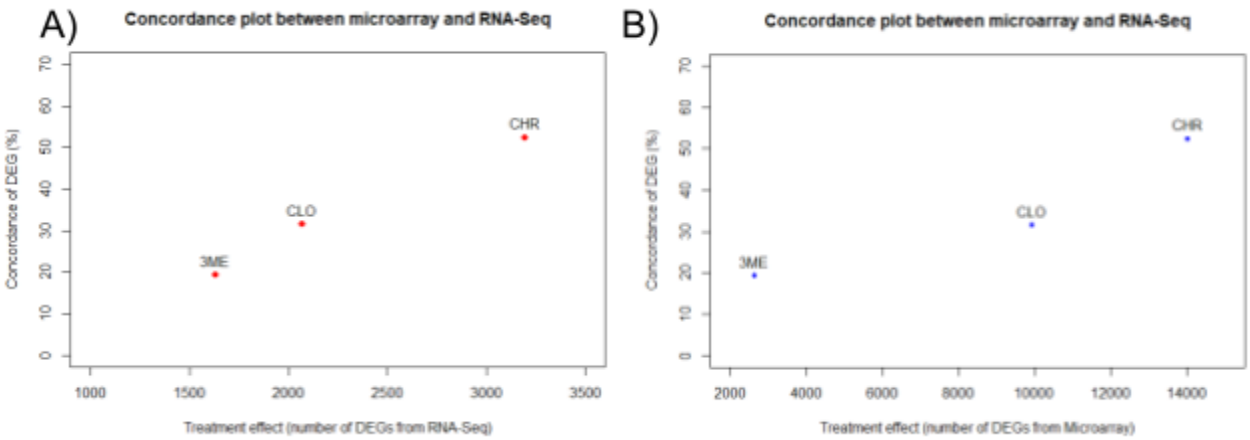

Figure 7. Concordance vs the number of differentially expressed genes where A) uses the number of differentially expressed genes from the RNA-seq analysis and B) uses the number of differentially expressed genes from microarray analysis.

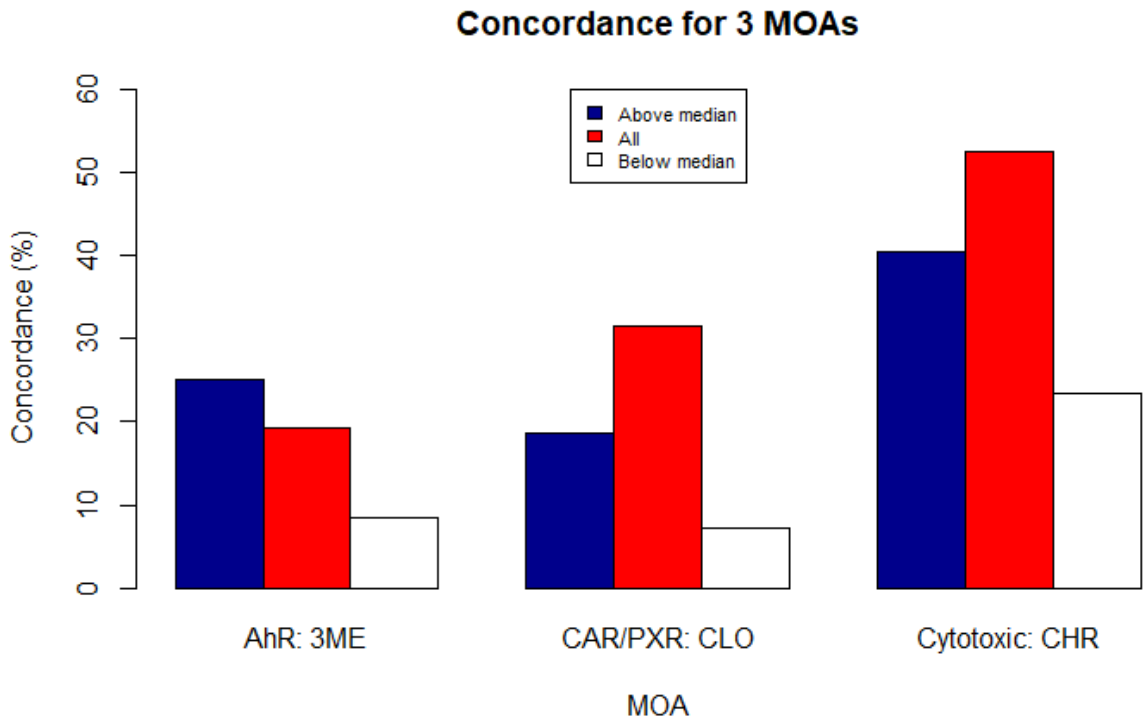

Figure 8. Concordance vs the number of differentially expressed genes grouped by the MOA and using subsets for each chemical above and below the median of significantly differentially expressed genes using the baseMean and AveExp.

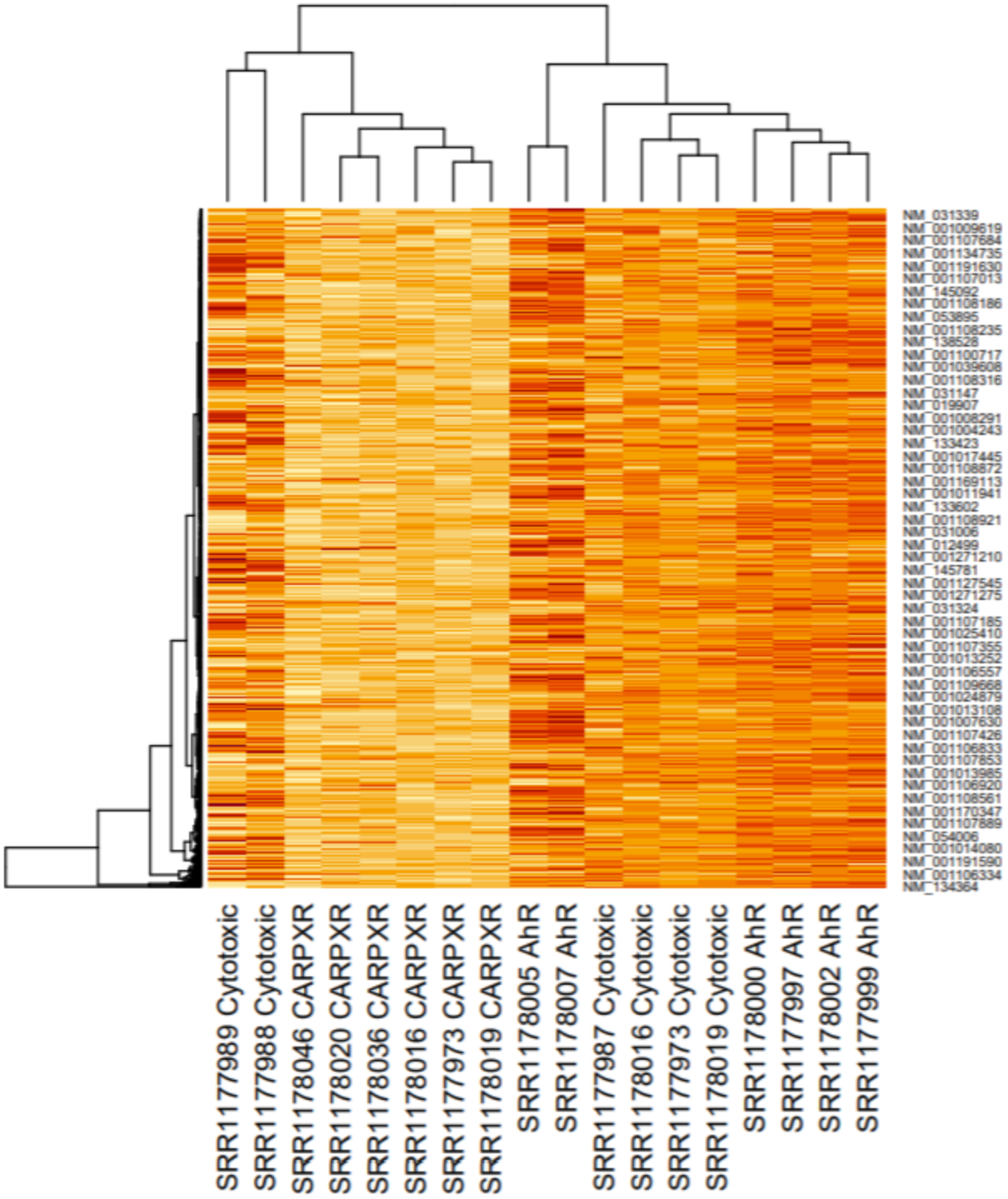

Figure 9. Heatmap and clustering of the differentially expressed genes from each member of the 3 tox groups after filtering

Even after filtering, the heatmap was unable to cluster all of the MOAs together. Figure 9 depicts heatmap clustering after filtering for lower than 0.186 coefficient of variation and then selecting the genes with means higher than the median. Selecting for genes with a cov higher than 0.186 yielded a heatmap that was unable to cluster the MOAs together.

Table 7. Wang et al's Supplementary Table 4, which details common enriched pathways for each MOA using GeneGo's pathway analysis software.

CAR/PXR (7)
Aryl hydrocarbon receptor signaling
Glutathione-mediated detoxification
LPS/IL-1 mediated inhibition of RXR function
NRF2-mediated oxidative stress response
Nicotine degradation II
PXR/RXR activation
Xenobiotic metabolism signaling
AhR (10)
Acetone degradation I (to methylglyoxal)
Aryl hydrocarbon receptor signaling
Bupropion degradation
LPS/IL-1 mediated inhibition of RXR function
Melatonin degradation I
Nicotine degradation II
Nicotine degradation III
Retinoate biosynthesis I
Superpathway of melatonin degradation
Xenobiotic metabolism signaling
Cytotoxic (15)
Acetone Degradation I (to Methylglyoxal)
Bupropion degradation
Cell cycle: G2/M DNA damage checkpoint regulation
Citrulline biosynthesis
Estrogen biosynthesis
LPS/IL-1 mediated inhibition of RXR function
Melatonin degradation I
Methylglyoxal degradation III
NRF2-mediated oxidative stress response
Nicotine degradation II
Pyrimidine ribonucleotides de novo biosynthesis
Regulation of eIF4 and p70S6K signaling
Superpathway of melatonin degradation
Superpathway of methionine degradation
Xenobiotic metabolism signaling

Table 8. AhR KEGG pathway annotation of top 10 differentially expressed genes via DAVID.

Metabolism of xenobiotics by cytochrome P450
Drug metabolism - cytochrome P450
Retinol metabolism
Chemical carcinogenesis
Steroid hormone biosynthesis
Estrogen signaling pathway
Antigen processing and presentation
Metabolic pathways

8 pathways were identified in the AhR KEGG pathway analysis (Table 8). 6 of the 10 top differentially expressed genes were represented in this table. The list of top 10 differentially expressed genes can be found in Table 4. 4 of the 10 genes were excluded from the analysis by DAVID. These genes were listed in Table 4 as NA, Mme, Dusp6, and Slc13a3. 8 pathways represented by the 6 differentially expressed genes included in the analysis. 62.5% of these pathways were also found in the CA/PXR pathway output and 25% were found in the Cytotoxin pathway output. None of these pathways matched the exact pathway output found by Wang et al, though some alluded to similar functions.

Table 9. CAR/PXR KEGG pathway annotation of top 10 differentially expressed genes via DAVID

Chemical carcinogenesis
Metabolism of xenobiotics by cytochrome P450
Retinol metabolism
Bile secretion
Steroid hormone biosynthesis
Drug metabolism - other enzymes
Drug metabolism - cytochrome P450

All of the top 10 differentially expressed genes, which can be found in Table 5, were represented in Table 9 using the DAVID KEGG pathway annotation. 7 pathways were found in this analysis. 71.4% of these pathways were found in the AhR output while 42.9% were found in the Cytotoxin output. None of these pathways matched the exact pathway output found by Wang et al, though some alluded to similar functions.

Table 10. Cytotoxin KEGG pathway annotation of top 10 differentially expressed genes via DAVID

Metabolism of xenobiotics by cytochrome P450
Chemical carcinogenesis
Bile secretion

KEGG pathway analysis on the top 10 cytotoxin differentially expressed genes, found in Table 6, had 3 pathways with 5 genes not included in the DAVID output. The 5 genes not included in the DAVID output in Table 10 were listed in Table 3 as Abcc3, Gstp1, Car3, RGD1566134, and Abcb1b. ⅔ of these were also in the AhR pathway output and all of them were in the CAR/PXR pathway output. None of these pathways matched the exact pathway output found by Wang et al, though some alluded to similar functions.

Discussion

The number of genes we found to be differentially expressed in RNAseq for 3ME (AhR) were similar to the numbers presented in Wang et al, but our results were different for CHL (Cytotoxic). We found 530 differentially expressed genes in 3ME and 1740 in CHL in comparison to 578 and 3850 respectively in Wang et al. There were multiple steps that could have caused the variation seen between our results and Wang et al results. The main figures in the paper generally used limma for both RNAseq and microarray data, to facilitate more direct comparisons. We used DESeq2 for the RNAseq differential expression analysis instead. We also used adjusted p-values to screen for differentially expressed genes. Given that we see this discrepancy for the chemical with the largest number of differentially expressed genes, how we corrected for the higher number of comparisons seems a likely contributor to the discrepancy. This discrepancy between our results and Wang et al results is not necessarily a cause for concern, provided the concordance results and overall conclusions align.

We found 530 DEGs for 3ME in RNAseq, in comparison with only 58 DEGs in microarray. In contrast, we saw 1740 DEGs for CHL in RNAseq and 11407 DEGs in the microarray data. Perhaps 3ME has a smaller effect size, and the increased sensitivity of RNAseq allowed us to pick up more DEGs. In contrast, CHL belonged in the Cytotoxic group and we would expect a larger effect post CHL treatment. The fact that we found 6.5x more DEGs in microarray data compared to RNAseq for CHL suggests that for the higher end of the effect size, RNAseq was no longer able to effectively identify all DEGs. This suggests that the RNAseq data may not have been sequenced deeply enough to pick up all the DEGs involved.

While we were able to reproduce the finding that concordance between RNA-seq and microarray platforms is dependent on the effect size and expression level, we only know the information for 3ME and CHL. It is not known why clotrimazole (CLO) was not included in their figures and so we can only confirm that the trend truly exists between 3ME and CHL. The analysis of above and below median sets per chemical revealed an increasing trend between 3ME

and CHL, which makes sense. It seems that splitting DEGs into two subsets is not a great way to compare between RNA-seq and microarray as it does not seem to add any value. A median of expression should not be used as a method to separate up and down regulated genes if that was the comparison to be made. The difficulty of comparing genesets between RNA-seq and microarray techniques is still an evolving territory and requires a stronger effort in improving and defining concordance relationships.

Wang et al used chemicals with different modes of action to evaluate the concordance between microarray and RNAseq experiments. They made the simplifying assumption that chemicals with similar modes of action would have similar gene expression responses. The amount of similarity that would be seen between chemicals of similar mode of action will vary from chemical to chemical, based on the pathways and mechanisms involved. Regardless of the validity of this assumption, concordance of microarray and RNAseq data can still be evaluated as long as the data was produced using the same chemicals at the same dose.

The pathway annotations between our study and Wang et al's did not have similar results. This is probably due to the difference in databases used for pathway analysis. We used DAVID while Wang et al used GeneGo. Assuming each of these had its separate ontology and algorithm, that would account for the differences in pathway analysis results.

Conclusion

Overall we were able to reproduce similar numbers of DEGs in RNA-seq and microarray analyses and the same trend in concordance compared to what the authors found in the paper. One region of confusion is why clotrimazole was not included in much of the study as it limited our comparison to only two chemicals. In addition, why was the median chosen to separate overall DEGs sets? This is potentially introducing error in the interpretation of results as in the case of 3ME microarray we have a small set of significant DEGs to compare to RNA-seq. In addition, it was difficult to locate any information on the number of DEGs found through microarray on each MOA or chemical to truly compare if our sample sets aligned and the trends we see are statistically correct. The data processing and filtering during analysis could lead to differences in overall DEGs between the paper and our work. We feel that it is difficult to know what a good concordance truly is based on this work as stating that we see between 20%-50% overlap in differentially expressed genes between the two techniques does not seem significant. We found similar concordance between microarray and RNAseq as Wang et al, but whether RNAseq, microarray, or the overlap between the two identified the most biologically relevant genes would require further investigation.

References

Andrews, S. (2010). FASTQC. A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Davis Vaughan and Max Kuhn (2020). hardhat: Construct Modeling Packages. R package version 0.1.5. https://CRAN.R-project.org/package=hardhat

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Firke, Sam (2021). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.1.0. https://CRAN.R-project.org/package=janitor

Huang DW, Sherman BT, Lempicki RA. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37(1):1-13.

Huang DW, Sherman BT, Lempicki RA. (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 4(1):44-57.

Liao, Y., Smyth, G. K., & Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic acids research, 41(10), e108. https://doi.org/10.1093/nar/gkt214

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

Love, M.I., Huber, W., Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 Genome Biology 15(12):550 (2014).

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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.

Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Wickham, H. (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., and Müller, K. (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.1. https://CRAN.R-project.org/package=dplyr

Wickham, H., Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr

Wang, Charles, Binsheng Gong, Pierre R. Bushel, Jean Thierry-Mieg, Danielle Thierry-Mieg, Joshua Xu, Hong Fang, et al. (2014). A comprehensive study design reveals treatment- and transcript abundance–dependent concordance between RNA-seq and microarray data. Nature Biotechnology 32 (9): 926–32. PMID: 4243706

Zhu, A., Ibrahim, J.G., Love, M.I. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences Bioinformatics (2018).