Meghana Peshwa as Data Curator
Haroon Qureshi as Programmer
Liz Murphy as Analyst
Go Ogata as Biologist

Concordance of Microarray and RNA-Seq Differential Gene Expression

Introduction

Microarray technology has been widely used to detect the expression of multiple genes at the same time. RNA-Seq uses high throughput sequencing methods which can also be used to quantify gene expression. The study conducted by Wang et al sought to find the concordance between these two methods when it comes to identifying differentially expressed genes (DEGs) and developing predictive models. They reported the results of a comparative analysis of gene expression responses profiled by Affymetrix microarray and Illumina RNA-seq in liver tissue from rats exposed to 27 chemicals representing multiple modes of action (MOA). They found that the concordance between array and sequencing platforms for detecting the number of DEGs was linearly correlated to the treatment effect. Also, they reported a much higher concordance for the above-median expressed genes than for the below-median expressed genes. The purpose of this project is to recreate these findings of the study and also to compare the pathway enrichment results reported in the paper using a subset of the samples.

Data

In Wang et al, RNA was isolated from the livers of male Sprague-Dawley rats after being exposed to one of 27 chemicals (three rats per chemical with matched controls). These samples were then analyzed using Affymetrix microarrays and Illumina RNA-seq. The samples were split into training (63 samples) and test set (42 samples). RNA seq of these was performed and the paired-end reads were deposited under accession number SRP024314. The reads underwent analysis through multiple pipelines (magic-pipeline as mentioned in the supplementary material) which included

QC, alignment, annotation, quantification, and normalization. Fragmented cRNA prepared from liver RNA was hybridized and all arrays met the minimal recommended quality parameters as described in the Affymetrix Data Analysis Fundamentals Guide.

For this project, the microarray and sequencing datasets have been downloaded from the accessions SRP039021, GSE55347, and GSE47875. The microarray data was already processed as described in the paper. For sequencing data, we picked toxgroup 1 which consists of 9 toxic samples and 6 control samples. The control samples have already been processed. FastQC was run on the 9 toxic samples. Then, 9 BAM files were created by aligning the samples to the provided rat genome using STAR. MultiQC was performed on both the fastqc files and the STAR alignments to combine results into convenient reports.

The FastQC report showed that the GC content was around 49% for all samples. The number of reads for all samples ranged from 16M to 22M reads with the average proportion of unique reads being around 50%. The majority of the sequences had a Phred score greater than or equal to 20 even though 17 of the 18 samples showed failure. This could be due to degradation of quality over the duration of long runs or due to a short loss of quality earlier in the run, which then recovers to produce a good quality sequence. In such cases, it is not recommended to trim the sequences. The per sequence quality score showed good results. The STAR alignment QC report showed that for each sample, >80% of the reads mapped to the reference genome except for SRR1178036 which was 68%. The uniquely mapped reads ranged from 11M to 19M for all the samples. Overall, these statistics show that the sequence data are of good quality for analysis. The MultiQC reports are available in our git repository.

Table 1: Summary of read and alignment statistics

Sample names	Input reads	Avg read length	Uniquely aligned reads %	Multimapped reads %	Unmapped reads %
SRR1177987	17433393	202	84.81	3.79	11.4
SRR1177988	18368879	202	85.29	3.62	11.09
SRR1177989	18804974	202	83.53	4.87	11.6
SRR1177997	19746775	202	89.17	4.06	6.77
SRR1177999	21838440	202	88.72	4.18	7.1
SRR1178002	18844950	202	89.13	4.14	6.73
SRR1178020	16002905	200	83.57	5.22	11.2

SRR1178036	16867274	200	67.85	4.2	27.96
SRR1178046	17652475	200	85.35	5.19	9.44

Methods:

The alignments produced by STAR were processed using featureCounts, a program that uses a reference gene to count the reads mapped to genomic features. In this reproductive study, rn4_refGene_20180308.gtf was used as the reference. Once the reads were mapped, the results were assessed using MultiQC, a program that compiles an HTML report using the statistics generated by featureCounts. MultiQC performed quality control checks on the 9 samples and produced a report indicating any significant biases and trends.

The Biocunductor package DESeq2 (version 1.34.0) was installed and used to analyze the differential expression between the control samples and the samples treated with their respective treatments. The samples chosen in this study were grouped by their treatment method, being either AhR, CAR/PXR, or Cytotoxic, respectively. The samples in each group, along with the control samples that were treated through the same vehicle and the genes found with an adjusted p-value lower than 0.05 were determined to be differentially expressed. Along with the differential expression, normalized counts were also calculated and further analyzed downstream.

Gene expression was experimentally measured via the Affymetrix whole genome GeneChip® Rat Genome 230 2.0 Array. The counts were normalized by robust multi-array average before being analyzed by limma. The samples of interest were treated with chemicals from toxgroup 1: 3-methylcholanthrene (3ME) as the representative aryl hydrocarbon receptor (AhR) ligand, clotrimazole (CLO) as the representative orphan nuclear hormone receptor (CAR/PXR) ligand, and chloroform (CHR) as the representative cytotoxic (Cyto) treatment. Three separate limma analyses were performed, comparing each drug treatment to their respective vehicle controls. Limma results were sorted by adjusted p value ascending, and genes with an adjusted p value < 0.05 were considered significant.

Cross-platform concordance scores were calculated between differentially expressed genes from RNA-Seq and microarray analyses.

$$(1) \qquad x + \frac{(n_1 - x) \times (n_2 - x)}{N - x} = n_0$$

Equation 1 shows the relationship between n_0, the observed intersection of differentially expressed genes from the two analyses, n_1 and n_2, the total number of differentially expressed genes in each analysis, $N,$ the total number of genes in the rat genome, and the background corrected intersection, x. After calculating the background corrected intersection, concordance scores were calculated using Equation 2.

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

Using the Deseq2 results, the differential expressed gene with the lowest p-value was used for the pathway enrichment analysis. The pathway enrichment tool that was used was the web-based DAVID gene database, and the data was queried using KEGG and Reactome pathway database. The heatmap and sample clustering was created using the base R graphing package. The 3 different MOA sample normalized gene counts were combined into one dataset, and each gene was filtered. The filtering involved removing genes that were in less than 20% of the samples, an expression level greater than 20 count, and the coefficient of variation > 0.2.

Results:

Once the STAR alignments were prepared, featureCounts was used for mapping the reads. As seen in Figure 1, between all samples, 59.3% -62.6% of reads were assigned, 9.9%-13.7% of reads unassigned due to multimapping, 21.1%-23.9% unassigned due to no identification of features, and 4.7%-6.2% of reads unassigned due to feature ambiguity.

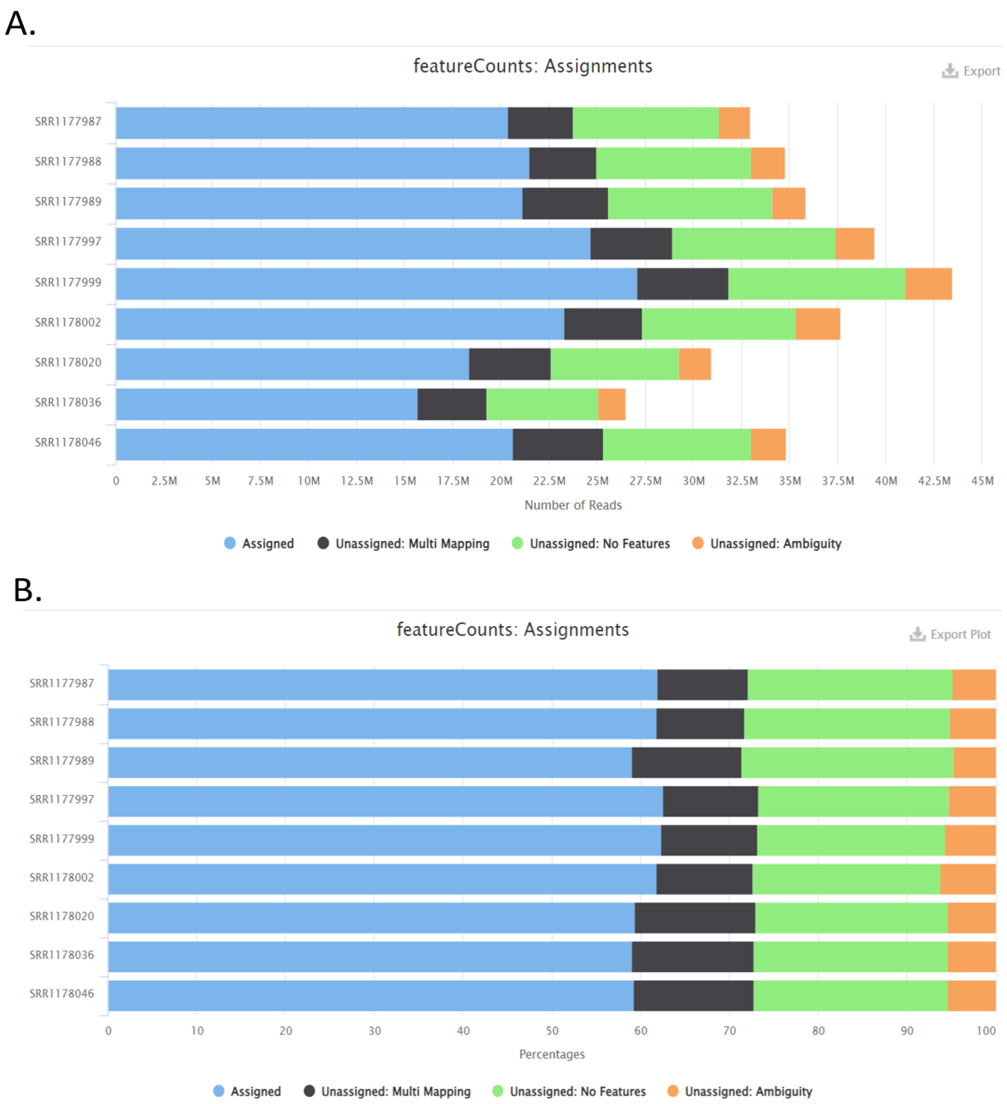


Figure 1. Results from MultiQC of featureCounts mappings. **A.** Reads aligned per sample **B.** Percentage of assigned and unassigned reads

Boxplots were made to further visualize the distribution of the counts for each sample. The distribution, as seen in Figure 2, was relatively varied between all samples, although sample SRR1177989, in the Cytotoxic group, had the widest distribution of counts. Looking at each group, the distributions were mostly in the same ranges, but the CAR/PXR group showed the greatest range in distribution. In this group, the sample SRR1178046 had the largest distribution, while in the AhR group, sample SRR1177999 had the greatest distribution. The black line in the boxplot indicated the median of each count distribution.

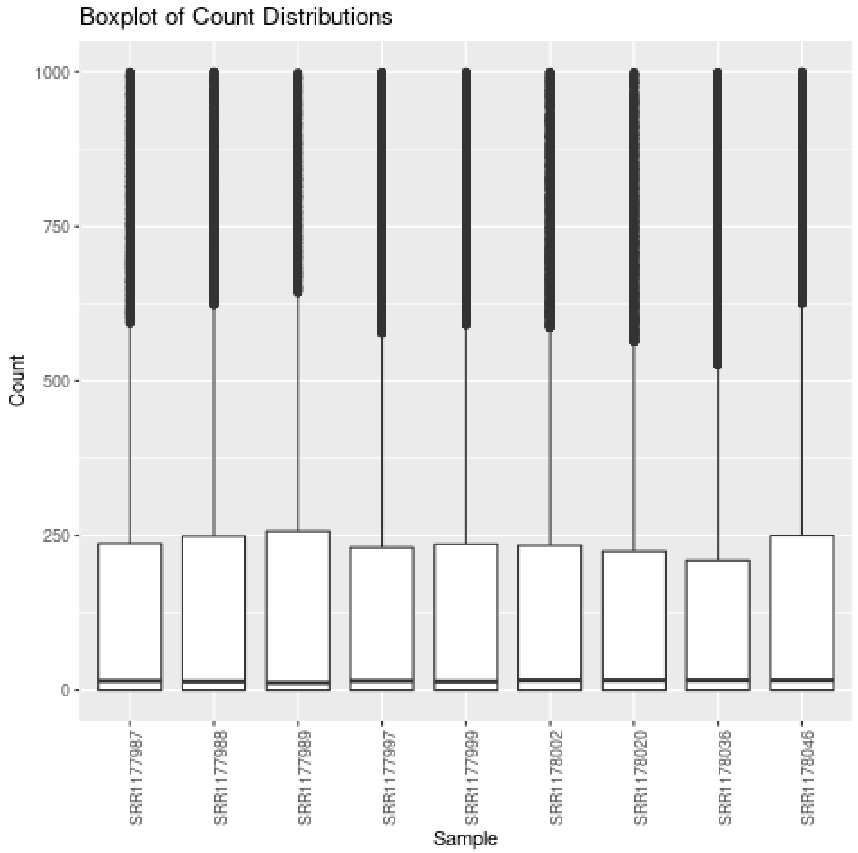


Figure 2. Boxplot distribution for each sample. The nine samples are ordered by groups of treatment applied to each sample.

Using the DESeq2 package in R, differentially expressed genes were identified by searching for genes with an adjusted p-value less than 0.05. Respectively, the AhR, CAR/PXR, and Cytotoxic sample groups had 313, 930, and 1,728 differentially expressed genes. The top 10 most differentially expressed genes for each group can be seen in Table 2. In Wang et al. these trends were mirrored, as the AhR group had the fewest differentially expressed genes, while the Cytotoxic group had the most, although there were fewer of these genes in number found in each group compared to what Wang et al. had discovered.

Table 2. The top 10 most significant differentially expressed genes in each treatment group.

AhR Gene ID	Padj	CAR/PXR Gene ID	Padj	Cytotoxic Gene ID	Padj
NM_012541	1.52E-79	NM_001010921	2.82E-114	NM_203512	2.86E-128
NM_130407	4.00E-19	NM_080581	8.63E-110	NM_001257095	2.72E-96
NM_022521	7.61E-11	NM_013105	4.42E-104	NM_080581	1.14E-45
NR_046239	8.54E-10	NM_131903	2.97E-67	NM_023978	2.57E-40
NM_134329	1.68E-09	NM_173295	1.65E-65	NM_012844	1.23E-36
NM_175761	3.56E-09	NM_012844	1.69E-64	NM_013215	2.18E-36
NM_012608	8.85E-09	NM_017272	2.60E-63	NM_139115	1.39E-34
NM_024351	1.12E-08	NM_013215	6.82E-58	NM_012540	1.52E-29
NM_053883	1.20E-08	NM_001134844	4.16E-52	NM_001013098	7.43E-29
NM_022866	2.30E-08	NM_133586	2.21E-49	NM_001010921	9.06E-26

Histograms were made to capture the log2 fold change for the significant differentially expressed genes (where $\alpha < 0.05$) and the frequency of such genes in each treatment group (Figure 3 A-C), followed by volcano plots, a type of scatter plot, that compared the log2 fold change against the -log nominal p-value (Figure 3 D-F).

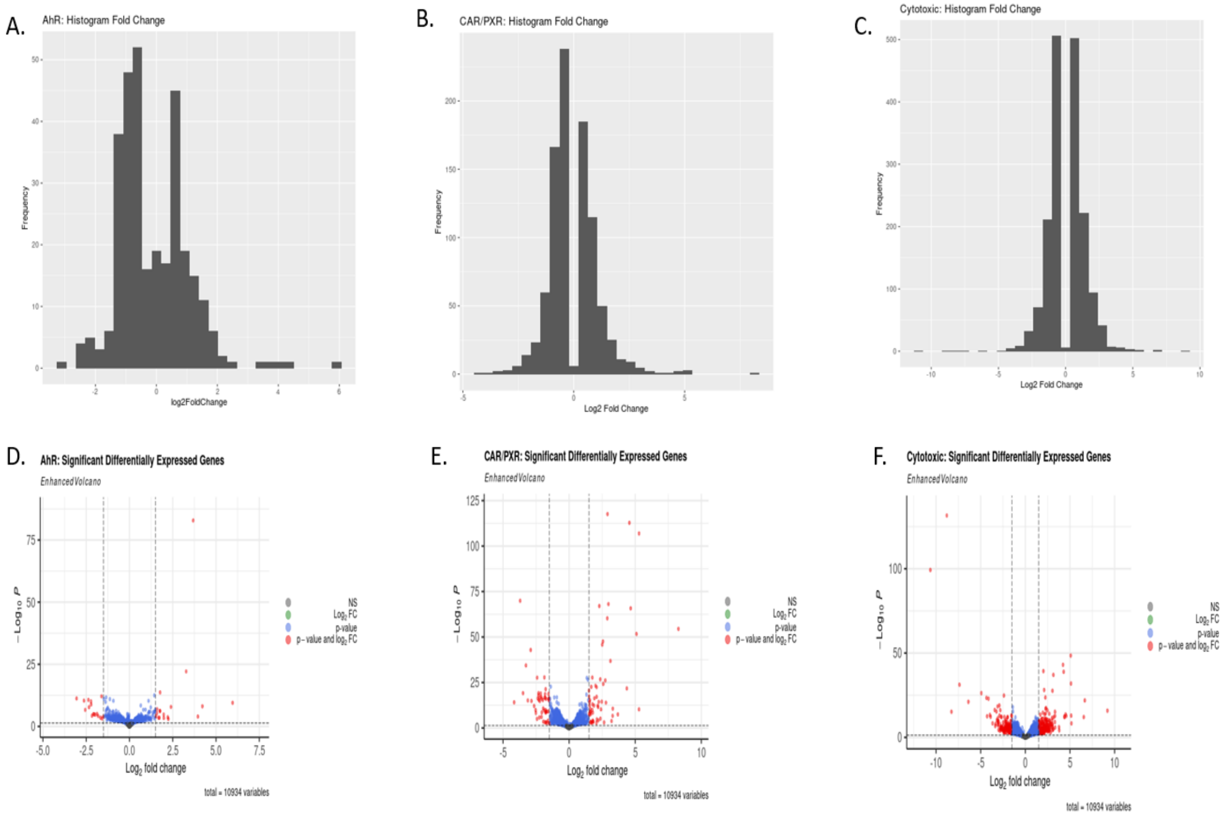
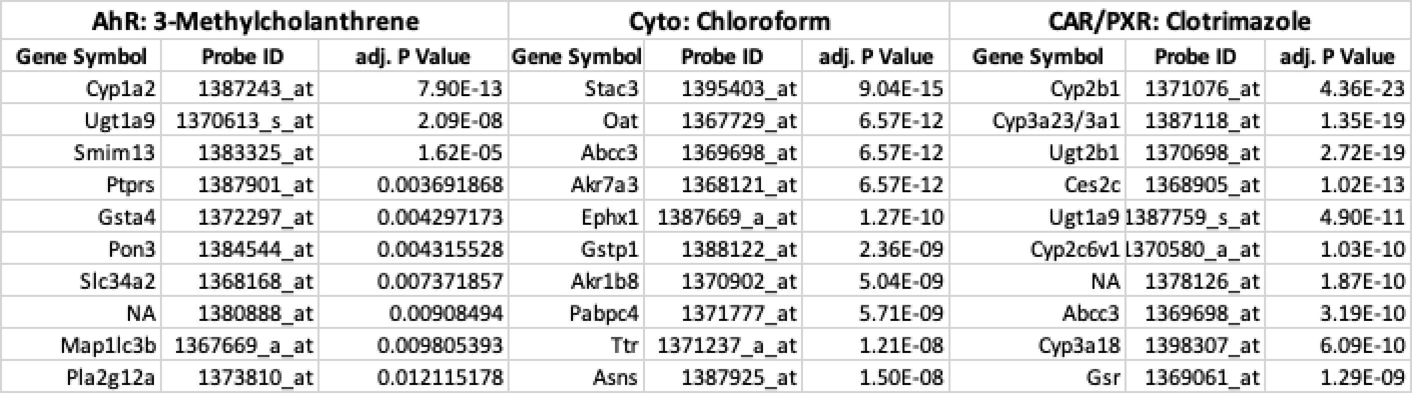


Figure 3. Histogram and Volcano Plots depicting log2 fold change for the significant differentially expressed genes for each treatment group: **A-C:** Histograms depicting log2

fold change and frequency of **A.** AhR **B.** CAR/PXR and **C.** Cytotoxic treatment groups. **D-F:** Volcano Plots depicting log2 fold change against the -log nominal P-value for **D.** AhR **E.** CAR/PXR and **F.** Cytotoxic treatment groups.

As can be seen in both the histograms and the scatter plots, there seemed to be a similar amount of up-regulated genes than down-regulated genes, and this distribution remained consistent between the experimental groups.

Table 3. Top 10 differentially expressed genes by treatment group from Microarray by adjusted P value.



AhR: 3-Methylcholanthrene			Cyto: Chloroform			CAR/PXR: Clotrimazole		
Gene Symbol	**Probe ID**	**adj. P Value**	**Gene Symbol**	**Probe ID**	**adj. P Value**	**Gene Symbol**	**Probe ID**	**adj. P Value**
Cyp1a2	1387243_at	7.90E-13	Stac3	1395403_at	9.04E-15	Cyp2b1	1371076_at	4.36E-23
Ugt1a9	1370613_s_at	2.09E-08	Oat	1367729_at	6.57E-12	Cyp3a23/3a1	1387118_at	1.35E-19
Smim13	1383325_at	1.62E-05	Abcc3	1369698_at	6.57E-12	Ugt2b1	1370698_at	2.72E-19
Ptprs	1387901_at	0.003691868	Akr7a3	1368121_at	6.57E-12	Ces2c	1368905_at	1.02E-13
Gsta4	1372297_at	0.004297173	Ephx1	1387669_a_at	1.27E-10	Ugt1a9	1387759_s_at	4.90E-11
Pon3	1384544_at	0.004315528	Gstp1	1388122_at	2.36E-09	Cyp2c6v1	1370580_a_at	1.03E-10
Slc34a2	1368168_at	0.007371857	Akr1b8	1370902_at	5.04E-09	NA	1378126_at	1.87E-10
NA	1380888_at	0.00908494	Pabpc4	1371777_at	5.71E-09	Abcc3	1369698_at	3.19E-10
Map1lc3b	1367669_a_at	0.009805393	Ttr	1371237_a_at	1.21E-08	Cyp3a18	1398307_at	6.09E-10
Pla2g12a	1373810_at	0.012115178	Asns	1387925_at	1.50E-08	Gsr	1369061_at	1.29E-09

Analysis of the microarray data found 46 genes differentially expressed (adjusted p value <0.05) in the AhR treatment group, 1674 in the CAR/PXR treatment group, and 5643 in the Cytotoxic treatment group. Figure 4 presents the limma results in volcano plots and histograms of log2FoldChange for each treatment group. The increase in differentially expressed genes from 3ME, to CLO, to CHR is visible when comparing the three volcano plots in Figure 4 a, c, and e. The majority of log2FoldChanges for 3ME were between -1 and 1, whereas CHR had the widest distribution of logFold2Changes, as seen in Figure 4f.

Concordance scores increased as the number of differentially expressed genes increased across treatments as shown in the scatter plots from Figure 5. CHR and CLO had very similar concordance scores, with CHR having the higher score overall. Below-median genes had lower concordance than above-median genes for all three treatment groups.

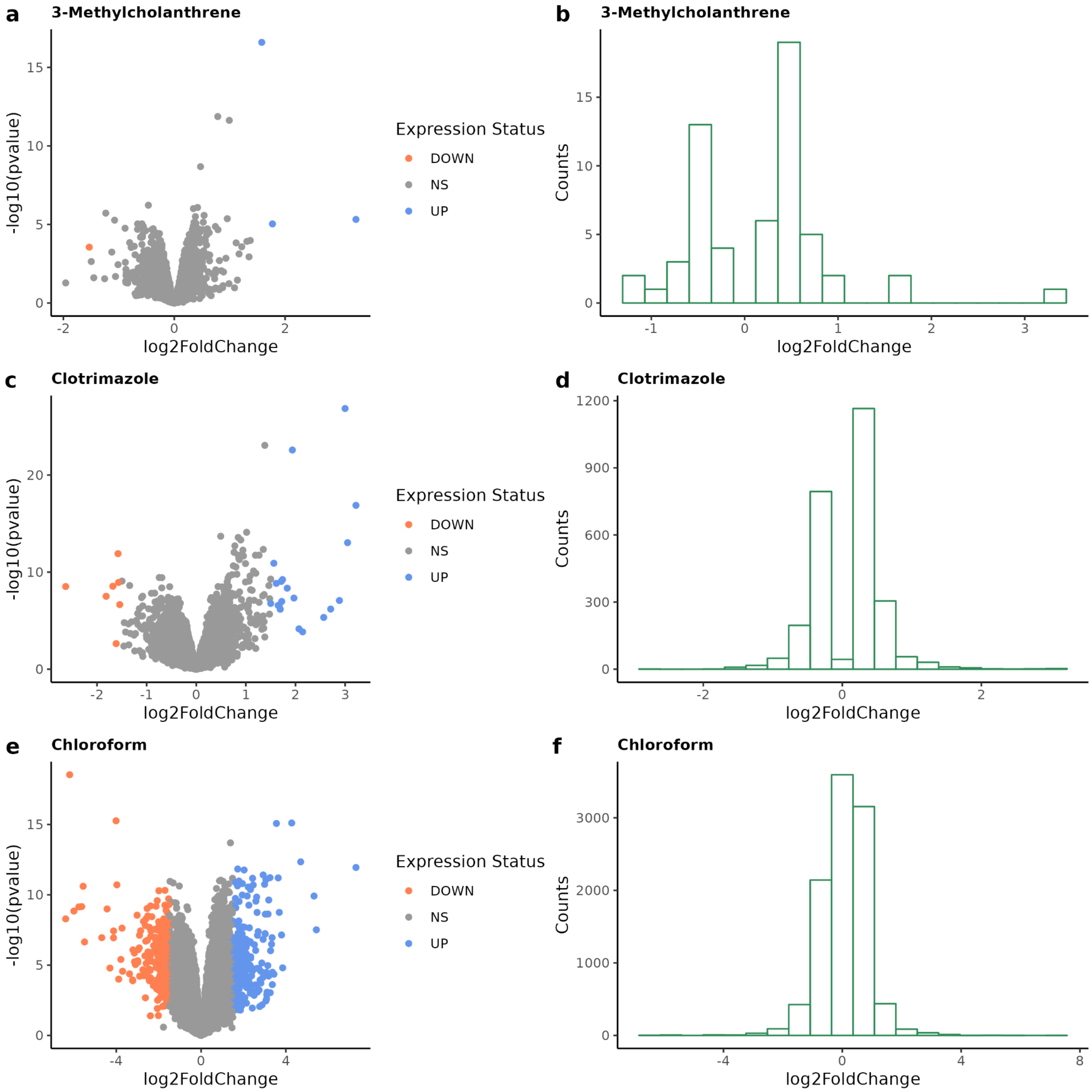


Figure 4. Volcano Plots and Histograms of log2FoldChange from Microarray analyses of 3 different treatment groups: 3-methylcholanthrene (a, b), clotrimazole (c, d) and chloroform (e, f). Volcano plots are colored by expression status: UP, p value < 0.05 and log2FoldChange > 1.5, DOWN, p value < 0.05 and log2FoldChange < -1.5, or NS, not significant.

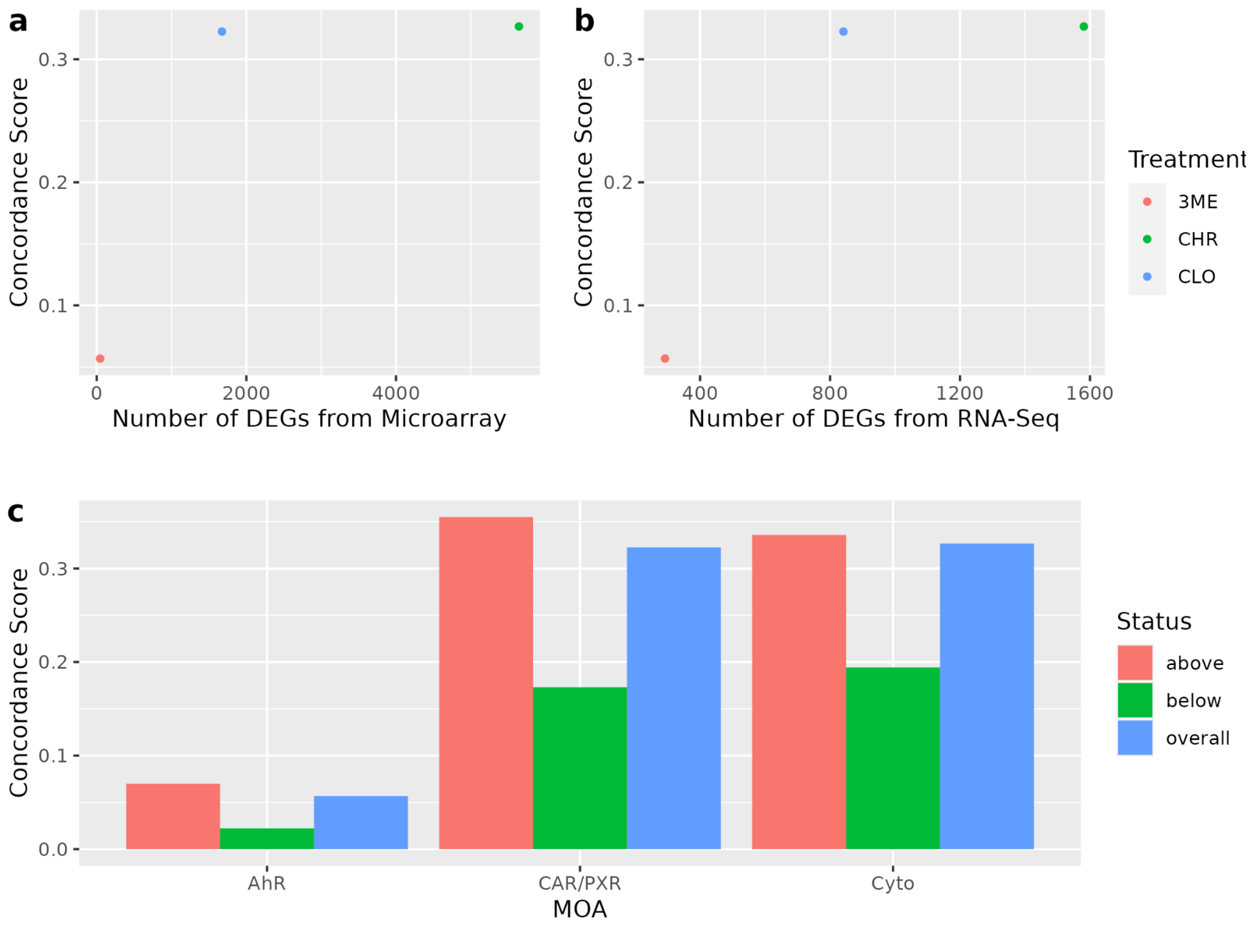


Figure 5. Cross-platform concordance analysis of differential gene expression between RNA-Seq and Microarray. a) Concordance score vs. number of DEGs identified via microarray for each treatment. b) Concordance score vs. number of DEGs identified via RNA-Seq for each treatment. c) Histogram of concordance scores from each treatment overall (blue), below-median genes only (green) and above-median genes only (red).

AhR Enriched Pathway	Car Enriched Pathway	Cytotoxic Enriched Pathway
Metabolism of xenobiotics by cytochrome P450	Drug metabolism - other enzymes	Drug metabolism - other enzymes
Chemical carcinogenesis - receptor activation	xenobiotic metabolic process	Retinol metabolism
Lipid metabolism	Porphyrin metabolism	negative regulation of cell growth
Retinol metabolism	Glutathion metabolism	estrogen metabolic process
Ascorbate and aldarate metabolism	Glutathione metabolism	xenobiotic metabolic process

fatty acid beta-oxidation	Metabolism of xenobiotics by cytochrome P450	Glutathione metabolism

Table 4 Enriched pathways for MOA Enriched pathways that were produced using DAVID with KEGG and Reactome databases. The displayed pathways were collected from the top most significant pathway in each cluster.

The overall trend showed different metabolic pathways involving Drug metabolism, other metabolic pathways, and several pathways involving cell growth. In addition to the general trend, each MOA has different trends. When compared to the original paper, there was very little overlap between the enriched pathways found by the reproduced data and the enriched pathways generated by the original paper. The only consistent overlap of the two enriched pathway analyses were the xenobiotic metabolic pathways. Most of the differences were found within the differences on drug related metabolic pathways as most of the pathways mentioned by the original papers are involved with certain types of drugs. To further support this finding, the enriched pathway that was recreated by us included an annotation describing “Drug metabolism - other enzymes” which can possibly describe the same relationship found in the original paper but only with its limited data.

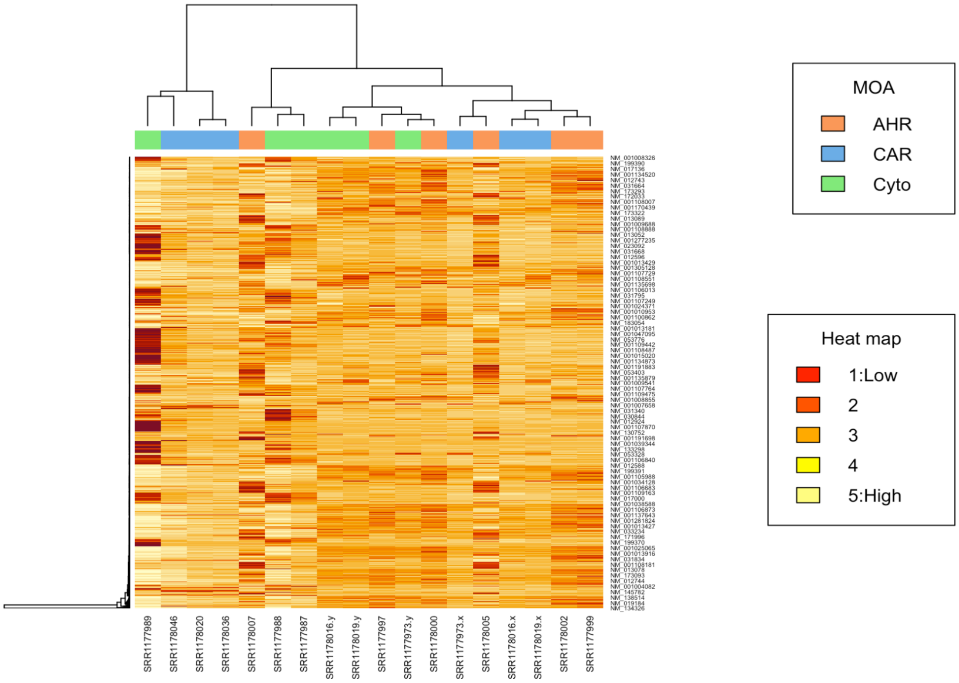


Figure 6 The heatmap of the different MOA and their gene expression. While there was general grouping of the 3 MOA, the gene expression was not sufficient to cluster the three different MOA.

The heat map was created to cluster the different MOA using the gene counts generated by DESeq2. While some samples were clustered into MOA groups, they were not able to perfectly cluster into the proper MOA groups. Observing the gene expression and how they are clustered, the different MOA did not contain enough difference between each other and were unable to be clustered appropriately. While there are long batches of clustering in the colors, those are only how they were labeled but there are at most two samples that were clustered together in each MOA. This clustering shows that rats had similar responses to the MOA, and any clustering is due to confounding factors

Discussion:

In this replicate study, the number of differentially expressed genes between experimental groups was found to follow similar trends as indicated in Wang et al. There were the fewest differentially expressed genes by number in the AhR group, and the most in the Cytotoxic group, although the Cytotoxic group in Wang et al. found 3,850 total significantly differentially expressed genes than our 1,728 genes. This discrepancy could have occurred due to an up-stream processing method, or could have been caused due to a differentiation in the methods used for finding the number of differentially expressed genes.

Microarray analysis found similar trends to RNA-Seq analysis, with 3ME having the fewest differentially expressed genes and CHR having the most. However the two approaches differed in the actual numbers of differentially expressed genes in each treatment group. This is consistent with the results of Wang et al., and likely related to treatment effect. RNA-Seq may be better for evaluating DEGs for drugs like 3ME with a smaller treatment effect, whereas microarrays are better for evaluating DEGs for drugs like CHR with a larger treatment effect. Microarrays were able to identify 5,643 DEGs in CHR treated samples versus only 1,728 identified from RNA-Seq analysis.

Concordance scores for each of the three treatments studied were lower in this repeat analysis compared to that of Wang et al., however the same overall trends were present. This is most likely due to varying approaches to identifying differentially expressed genes, from both the DESeq2 and limma results. An important result seen both in this repeat analysis and in Wang et al. is the reduced concordance between the two approaches for below-median expression genes. From their qPCR analysis, Wang et al. concluded that RNA-Seq was the better approach for detecting DEGs at lower expression levels, and this result is consistent with the hypothesis that RNA-Seq is also better at detecting DEGs under treatments with smaller effects.

The enriched pathways were not identical to the original papers that the data was taken from. While the xenobiotic metabolic process was similar in our and original enriched pathways, the original paper showed larger links between drug related pathways compared to the results that were reproduced. A possible explanation for the differences may be due to the difference in data sets used for the enriched pathways analysis. Our pathway analysis generated through KEGG and Reactome data, and showed metabolism pathways related to different drugs and contained several general drug metabolism annotations. The database used by the original authors was GeneGo's Canonical pathways, a database that can not be accessed, and this database may have included more annotation describing drug related mechanisms on top of all the other biological mechanisms. While the KEGG and Reactome data appear to be limited in the scope of drug related annotation, we can see that the different enriched pathways are pathways related to some kind of metabolic pathways instead of regulations we can speculate that the treatments of the different MOA are having similar pathways of metabolizing foreign chemicals in a similar way as found in the original study.

The relationship described in enriched pathways can also be seen within the heatmap sample clustering. Within the different normalized sample count there doesn't appear to be any big difference besides the lower count found in sample SSR1177989. Further supporting this finding is that the clustering is not consistent with the MOA, regardless of different types of filtering. The enriched pathway showed a large overlap on pathways that were related to metabolizing different drugs. KEGG and Reactome may not have the annotations to be able to effectively describe all drug metabolic pathways, however it did show that there were similar commonalities between the different MOA. These results suggest that within the toxicology group, the MOAs do not change gene expression to the extent and are clustered together.

Conclusion:

In conclusion, we were able to replicate the results found in Wang et al. and came to a similar conclusion from the data. We were able to find similar levels of DEGs within the replicated differential expression results and in the original studies for both RNA-Seq analysis and microarray. When comparing Concordance scores, the general trend found in the original paper was also found within our analysis. From these two analyses we were able to conclude that RNA-Seq was the better approach for detecting DEGs at lower expression levels than in microarray data. During the analysis we found slight discrepancies within our data, which we concluded were due to small differences in programs, and methodology that can’t be perfectly matched. The differences however were not significant enough to mask the overall conclusion found in the original paper.

References:

1. Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., Meehan, J., Li, X., Yang, L., Li, H., Łabaj, P. P., Kreil, D. P., Megherbi, D., Gaj, S., Caiment, F., van Delft, J., … Tong, W. (2014). The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nature biotechnology*, *32*(9), 926–932. https://doi.org/10.1038/nbt.3001