Xinyu Sun - Data curator

Will Mischler - Programmer

Reva Shenwai - Analyst

Nitsueh Kebere - Biologist

**TA: Dakota Hawkins**

**Introduction**

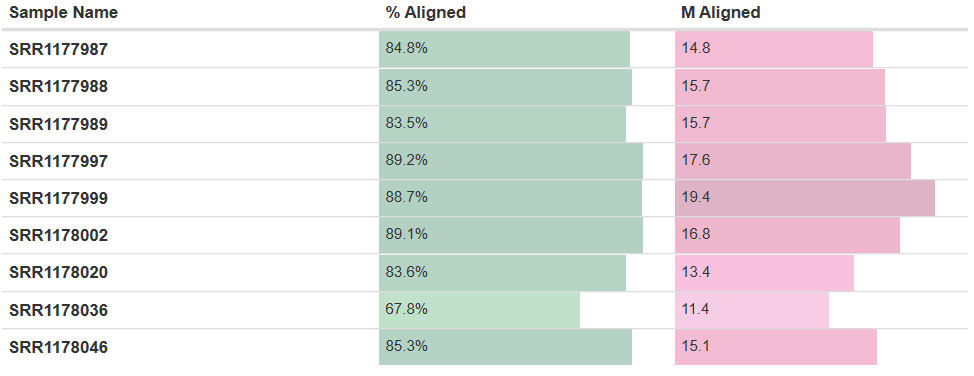
When Wang et al. published their paper in 2014, RNA-seq was an emerging technology, and not many studies had adequately addressed the possibility of RNA-seq enhancing the prediction of toxicity outcomes based on gene-expression data compared to microarray. Wang et al. (2014), used Affymetrix microarrays and Illumina RNA-seq to analyze liver samples collected from male Sprague-Dawley rats, which were treated with one of 27 chemicals (three rats per chemical with matched controls). According to Wang’s paper’s result, the concordance between array and sequencing platforms were positively correlated. In addition, gene expression–based predictive models generated from RNA-seq and microarray data were similar. However, RNA-seq performed better at detecting weakly expressed genes. In this project, we would like to compare and reproduce the concordance between RNA-seq and microarray data’s gene expression responses from Wang’s paper (2014).

**Data**

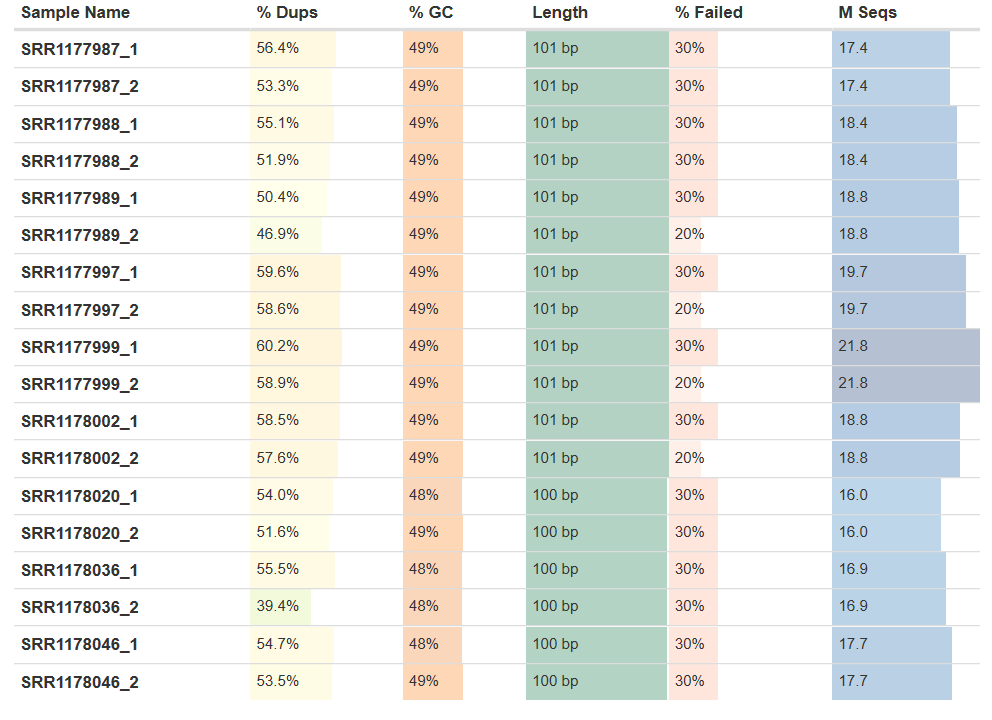
In this project, we chose toxgroup 1 to analyze. In the following analysis, we analyzed both Illumina RNA-seq and Affymetrix microarray data from liver samples of rates under three chemical conditions (AhR, CAR/PXR, and Cytotoxic). Affymetrix microarray data has been already prepared, so we just produced RNA-seq alignments. There are 9 fastq files available for toxgroup 1, in total, and three replicates for each condition (AhR, CAR/PXR, and Cytotoxic). 9 fastq files are quality checked by Fastqc, and then these reads are aligned to transcripts, using star/2.5.3a. Finally, multiqc was run to collect statistics and information from star alignment and fastqc.

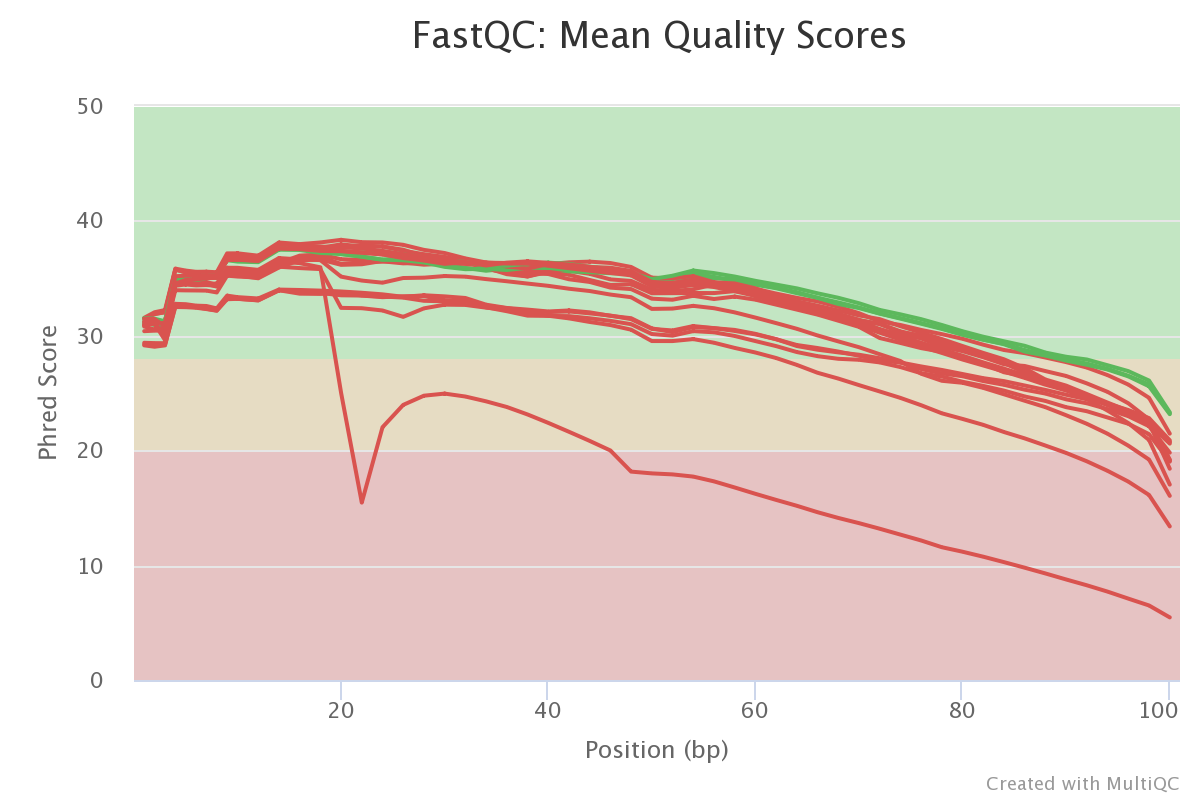
From the table of star alignment statistics, all alignments showed a high percentage of aligned reads, except for SRR1178036(Table 1). The similar pattern was also observed in fastqc statistics table, which SRR1178036\_2(8036\_2) has less duplications than others (Table 2). After looking into FASTQC results, We noticed that 8036\_2 has a lower duplication percentage and lower phred score at bp position 20(Table 2 & Figure 1). In addition, the percentage of base N counts is extremely high at bp position 20. Based on the above observation, we suggested that the low quality for 8036\_2 is probably due to the fact that most of the reads around bp position 20 all had ambiguous base calls.

**Table 1. Most BAM files have most of the reads aligned.** The table shows the percentage of aligned reads in each sample. Only SRR1178036 showed a lower percentage than the others.

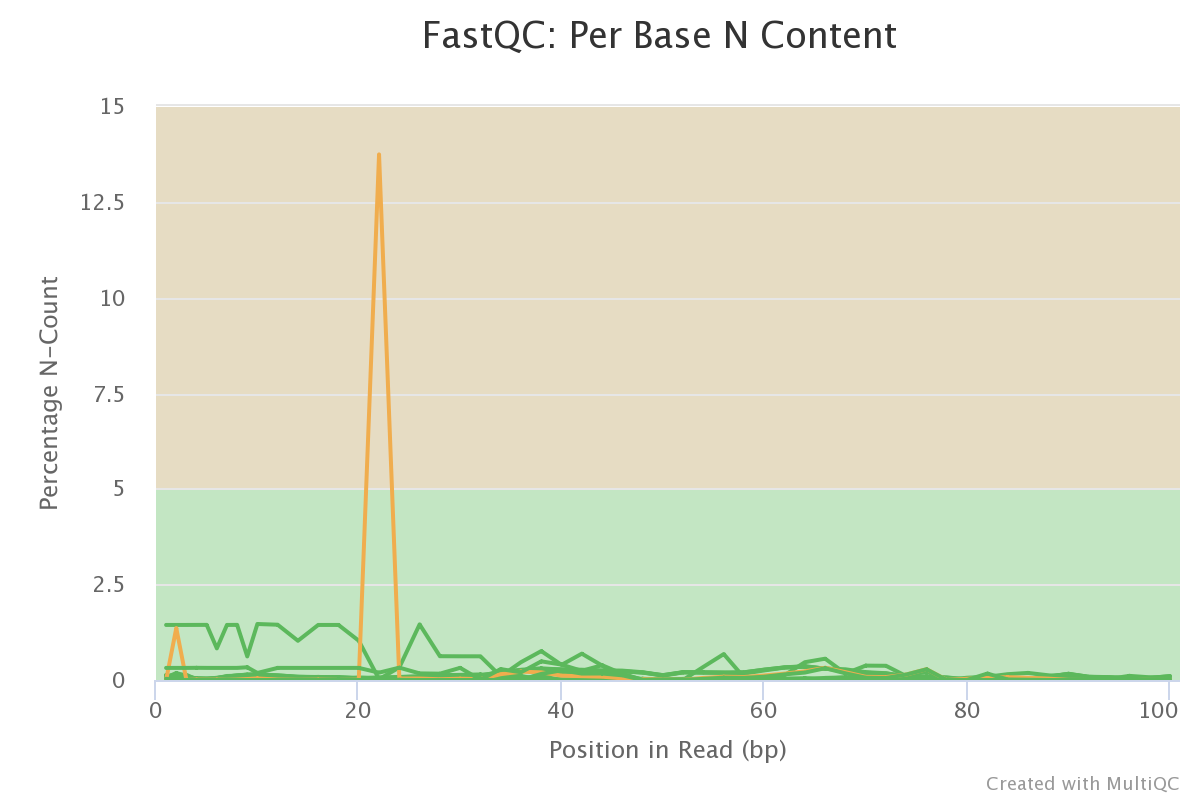


**Table 2. Most fastq files showed similar overall statistics.** Only SRR1178036\_2 showed a lower percentage in duplications than the others.





**Figure 1. SRR1178036\_2 had a lower phred score.** The plot showed the mean quality value across each base position in the read. At bp position around 20, the score drops dramatically to 15.46.

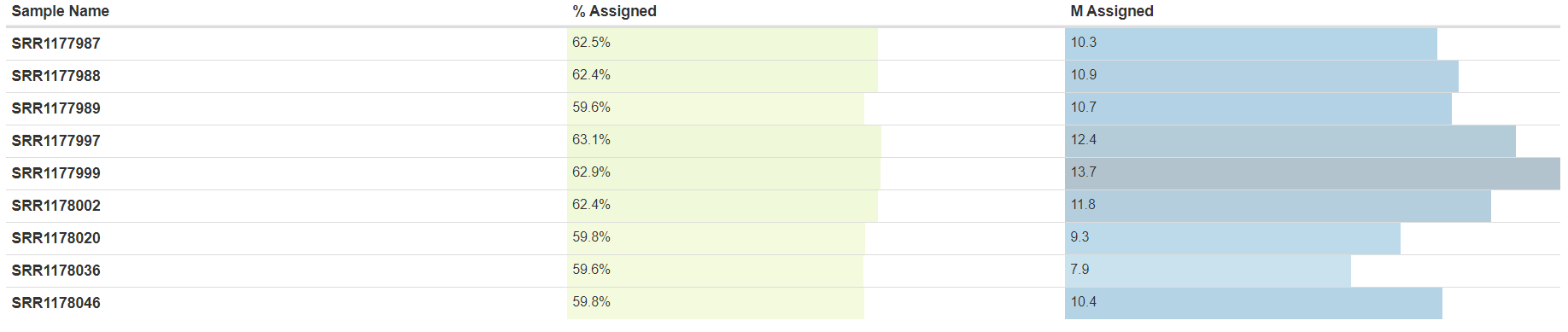


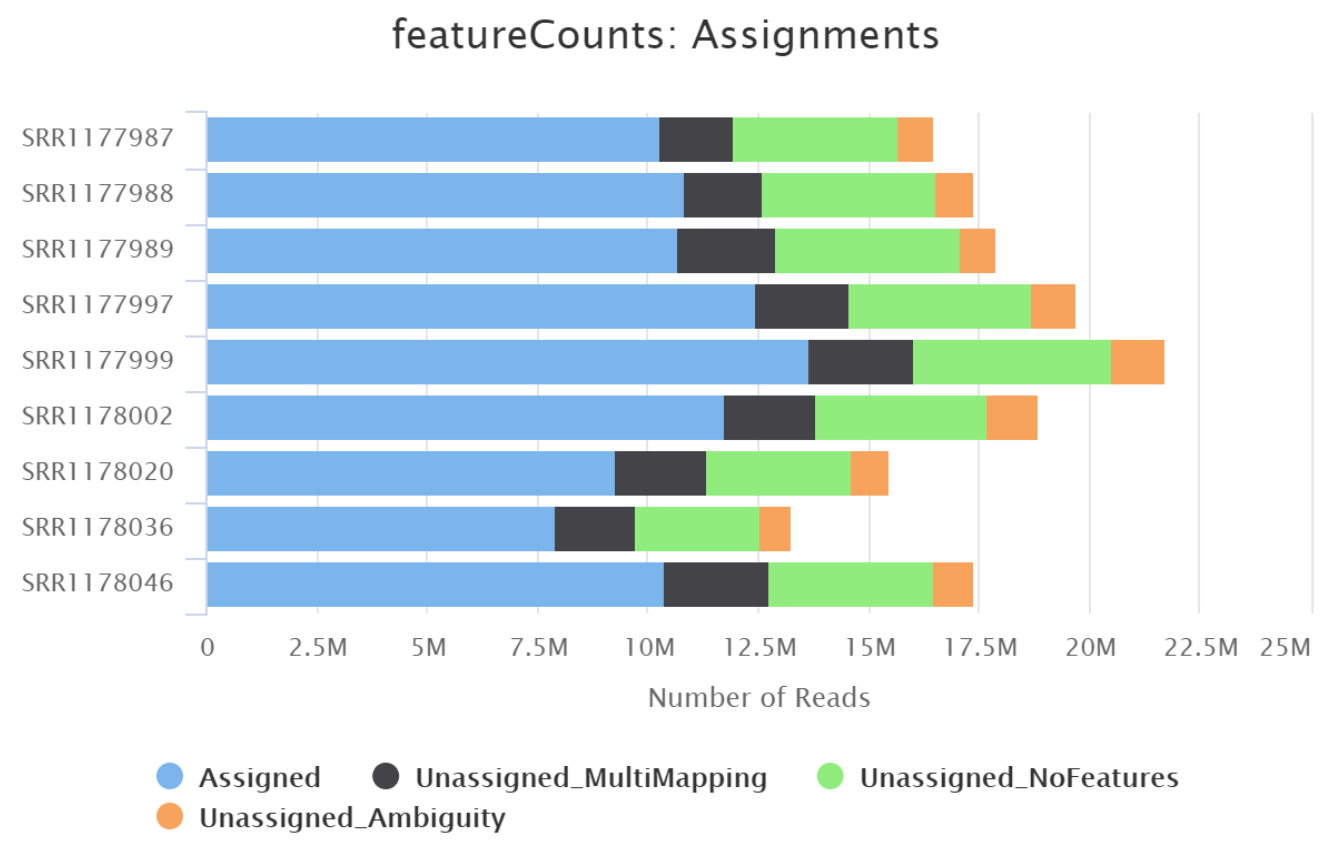
**Figure 2. SRR1178036\_2 had a Base N count at bp position 20.** This plot indicates the percentage of base calls at each position for which an N was called.

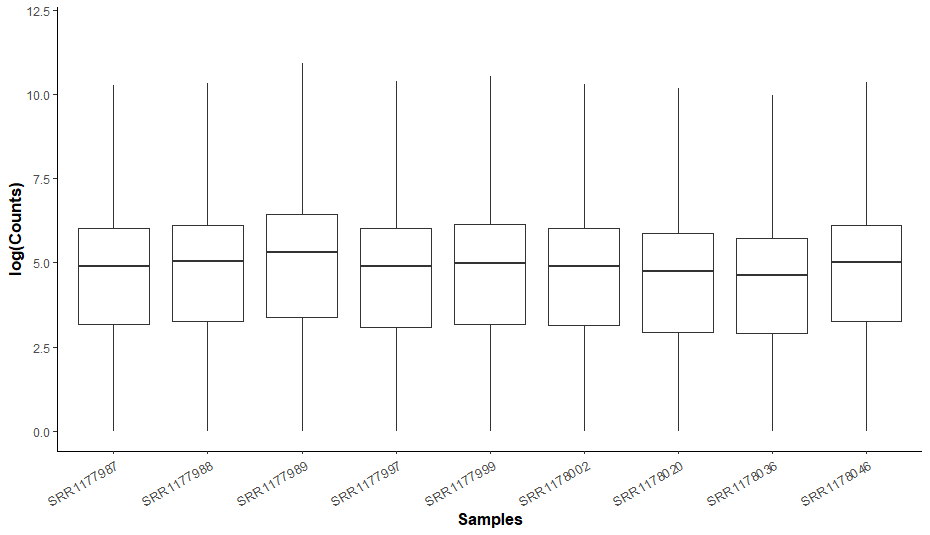
**Methods**

Using the featureCounts tool on the cluster with annotations in a GTF file, individual count files were generated for each of our nine samples. Sample counts were examined for outliers and other statistics using the multiqc tool. Samples show a good overall percentage assignment with no samples as outliers (Table 3). It can be seen that samples, SRR1178036 and SRR1178020, have lower numbers of overall reads (Figure 3). Samples were then combined to examine the overall count distribution between the samples. Overall the samples show the same distribution of counts across the whole geneset (Figure 4).

Control samples were added to the sample files. The samples were split into 3 different groups based on their mode of action (MOA), AhR, CAR/PXR and Cytotoxic, which contain 3 samples and control each. Differential expression analysis was performed on these groups separately using the DESeq2 package from Bioconductor. Normalized counts were obtained through DESeq2 using the median of ratios method. P-values for significance were adjusted using the Benjamin-Hochberg procedure and the cutoff was set to padj < 0.05.

**Table 3. Multiqc quality control.** Shows the percent assigned of all of the samples in the study. All samples hover around 60% assigned.

**Figure 3. Assigned and Unassigned reads in all samples.** Shows the breakdown of the number of reads in each sample. Samples SRR1178036 and SRR1178020 both have less overall and assigned reads than the rest of the samples.



**Figure 4. Log Count distribution across samples.** Each sample has a similar distribution of counts across the genes expressed. Log used to expand view of different counts

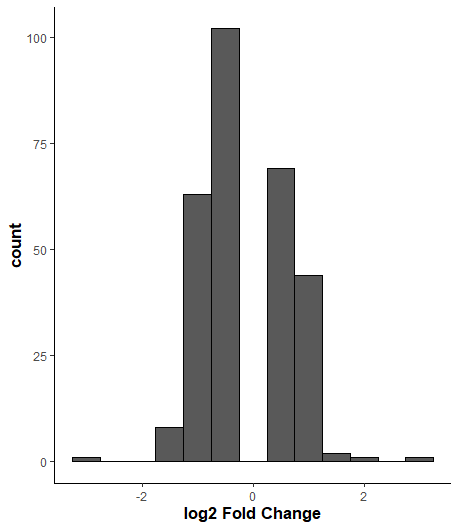
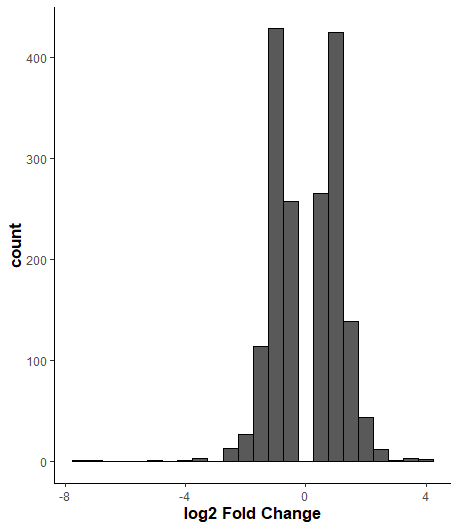
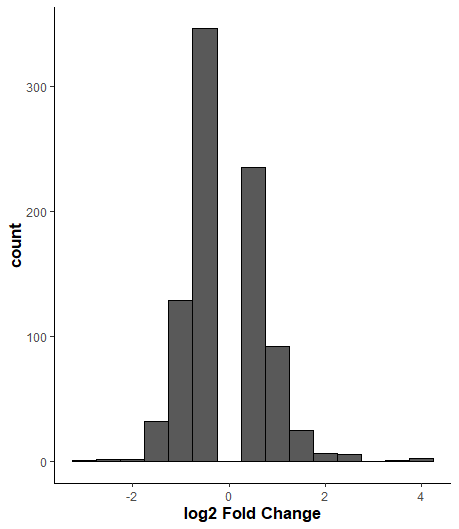
For the microarray data, RMA normalization and limma were used to perform differential expression analysis on the microarray data for each chemical in toxgroup 1. These chemicals were 3-Methylcholanthrene (3ME), Clotrimazole (CLO) and Chloroform (CHL), and they correspond to AhR, CAR/PXR and Cytotoxic MOAs respectively. After differential expression analysis, genes with insignificant adjusted p-values of >= 0.05 were filtered out for each chemical. All further downstream analysis was performed using only significant DEGs with p-adjust < 0.05.

We calculated concordance of microarray DE data with RNA-seq DE data for each of our three chemicals in toxgroup 1. For this, we first mapped Affymetrix probe IDs from the microarray analysis to refSeq IDs from the RNA-Seq analysis using a provided refSeq-to-probe ID mapping reference. In the case of duplicate IDs, only one copy of the probe ID or refSeq ID was mapped. Arbitrarily, the first copy of an ID encountered by the code was mapped. We feel that this will not impact the concordance results significantly, since only a small fraction of IDs from one platform maps to multiple IDs in the other platform. We then performed our concordance analysis using the same formula as Wang et al. (2014), as outlined in their methods section. We also performed a background-correction on our concordance value to account for increases in background intersection with set size, following the same formula as Wang et al. (2014).

**Results**

Differential expression analysis revealed 291 significantly differentially expressed genes (DEGs) in the AhR MOA, 881 significantly differentially expressed genes in the CAR/PXR group and 1736 significantly differentially expressed genes in the Cytotoxic group (Table 3). Histograms plotted of the significant log2 Fold Change show the range across each group (Figure 5a-c). Each histogram centers around zero as expected and shows relatively even split in all of the groups between upregulated and downregulated differential expression.

The top ten genes from each group were selected and are shown in Tables 4-6. Volcano plots were generated using all of the expression data regardless of significance. Like in the histogram, each group shows a fairly even distribution between up and down regulated genes. The plots show all of the significantly differentially expressed, everything above the red line, as well as highlight the most up and down regulated genes in each group (Figure 6a-c).

**A B C**

**Figure 5. Histograms of the three different groups.** Histograms show the range of log2 Fold Changes in the groups. All center around zero and show an even distribution of upregulated and downregulated differentially expressed genes. A) AhR group B) CAR/PXR and C) Cytotoxic

**Table 4. Top 10 Differentially Expressed Genes in AhR group**

|  |  |  |
| --- | --- | --- |
| **Gene Symbol** | **Log2 Fold Change** | **Padj** |
| Cyp1a2 | 3.0729716 | 4.545951e-77 |
| LOC298111 | -2.8382060 | 1.899089e-29 |
| Ugt1a7c | 2.1081757 | 3.267826e-20 |
| Oat | 1.3881420 | 1.743764e-10 |
| Inmt | -1.4762090 | 3.291060e-10 |
| Ugt1a6 | 1.4649857 | 4.349518e-10 |
| Adh7 | -1.3175358 | 1.700408e-09 |
| Hsp90aa1 | -1.1882444 | 2.177815e-09 |
| Rn45s | 1.1957548 | 2.177815e-09 |
| Slc13a3 | -1.4568771 | 1.209058e-08 |

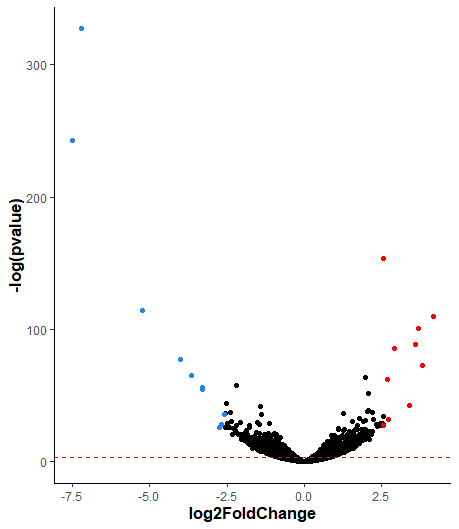
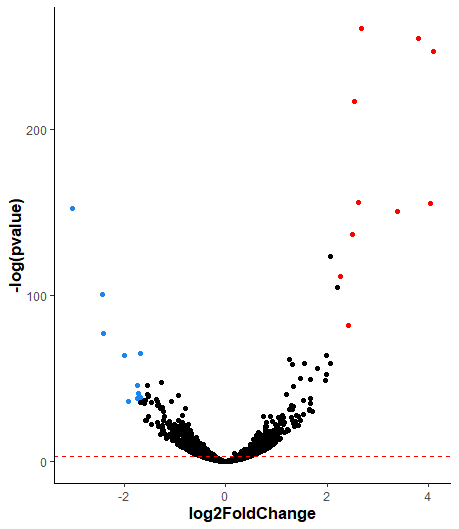
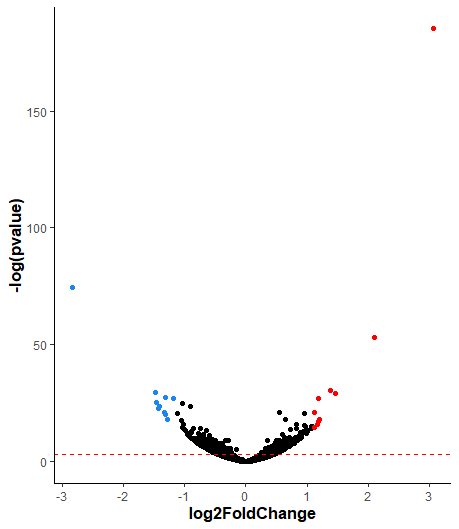
**Table 5. Top 10 Differentially Expressed Genes in CAR/PXR group**

|  |  |  |
| --- | --- | --- |
| **Gene Symbol** | **Log2 Fold Change** | **Padj** |
| Gsta5 | 2.6681939 | 1.080565e-109 |
| Abcc3 | 3.7916839 | 1.662618e-107 |
| Cyp3a23/3a1 | 4.0969591 | 2.504238e-104 |
| Ephx1 | 2.5444434 | 2.780543e-91 |
| Ugt2b1 | 2.6148548 | 4.886264e-65 |
| Ces2c | 4.0289043 | 7.391968e-65 |
| Sult2a1 | -3.0221374 | 1.317770e-63 |
| Aldh1a7 | 3.3879571 | 4.988355e-63 |
| Akr7a3 | 2.5060891 | 5.170728e-57 |
| Cyp2b1 | 2.0669067 | 2.270177e-51 |

**Table 6. Top 10 Differentially Expressed Genes in Cytotoxic group**

|  |  |  |
| --- | --- | --- |
| **Gene Symbol** | **Log2 Fold Change** | **Padj** |
| RGD1566134 | -7.2071492 | 8.374828e-139 |
| LOC100360095 | -7.4973967 | 1.746701e-102 |
| Ephx1 | 2.5787214 | 7.021818e-64 |
| Stac3 | -5.2345200 | 5.269257e-47 |
| Abcc3 | 4.1860695 | 3.081858e-45 |
| Per3 | 3.6908223 | 2.516543e-41 |
| Akr7a3 | 3.6102210 | 5.206704e-36 |
| Coro6 | 2.9121089 | 7.847942e-35 |
| Dhrs7l1 | -4.0139041 | 3.022858e-31 |
| Cyp1a1 | 3.8271731 | 2.883016e-29 |

**A B C**

**Figure 6. 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) expression are shown. A) AhR B) CAR/PXR C) Cytotoxic

Now we will look at the microarray analysis. The numbers of significant DEGs for Microarray analysis are summarized in table 7. Of the total 31100 genes analysed for each chemical, 3M has the lowest number of significant DEGs at 59, while CHLR has the highest at 11408. The top 10 highest expressed DEGs for each chemical from the microarray analysis can be found in tables 8-10. Overall, it appears that CHL has the highest log fold-change in expression of its top 10 DEGs. In addition, it is interesting that the three groups do not appear to share many top 10 DEGs, but this makes sense given their different MOAs.

**Table 7. Numbers of significant DEGs for each chemical in Microarray analysis.**

|  |  |  |
| --- | --- | --- |
| **Chemical** | **RNA-seq significant DEGs** | **Microarray significant DEGs** |
| 3ME | 292 | 59 |
| CLO | 882 | 5804 |
| CHL | 1737 | 11408 |

**Table 8. Top 10 DE genes from the Microarray analysis for 3ME.**

|  |  |
| --- | --- |
| **ProbeID** | **Log fold-change** |
| 1370269\_at | 3.0976 |
| 1387243\_at | 1.5365 |
| 1367856\_at | 1.1177 |
| 1369150\_at | 0.8183 |
| 1390097\_at | 0.7978 |
| 1387759\_s\_at | 0.7773 |
| 1377934\_at | 0.7657 |
| 1373778\_at | 0.7586 |
| 1370371\_a\_at | 0.7353 |
| 1370613\_s\_at | 0.6348 |

**Table 9. Top 10 DE genes from the Microarray analysis for CLO.**

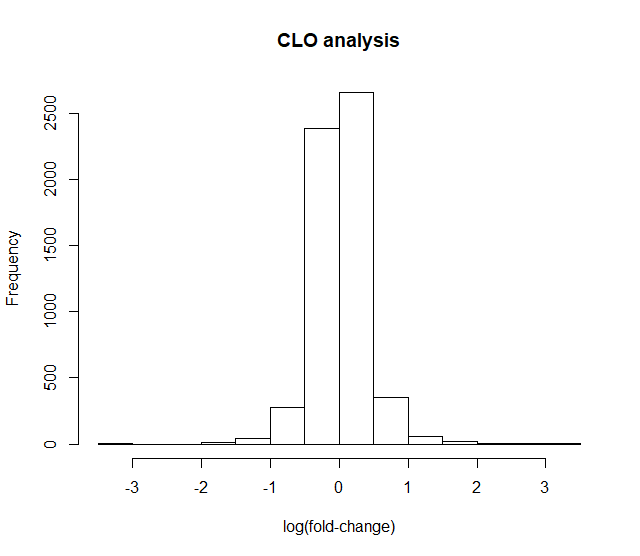
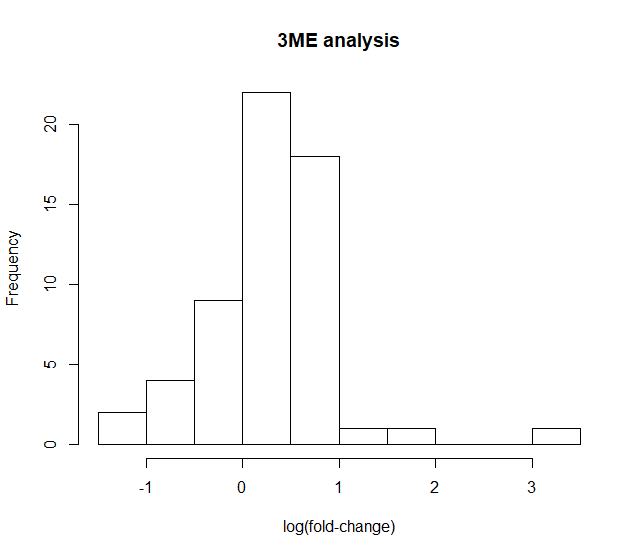
|  |  |
| --- | --- |
| **ProbeID** | **Log fold-change** |
| 1368905\_at | 3.3418 |
| 1369698\_at | 3.0958 |
| 1371076\_at | 3.0189 |
| 1370269\_at | 2.9758 |
| 1369921\_at | 2.7149 |
| 1372016\_at | 2.6202 |
| 1368718\_at | 2.4399 |
| 1394128\_at | 2.0817 |
| 1370698\_at | 1.9683 |
| 1373810\_at | 1.9069 |

**Table 10. Top 10 DE genes from the Microarray analysis for CHL.**

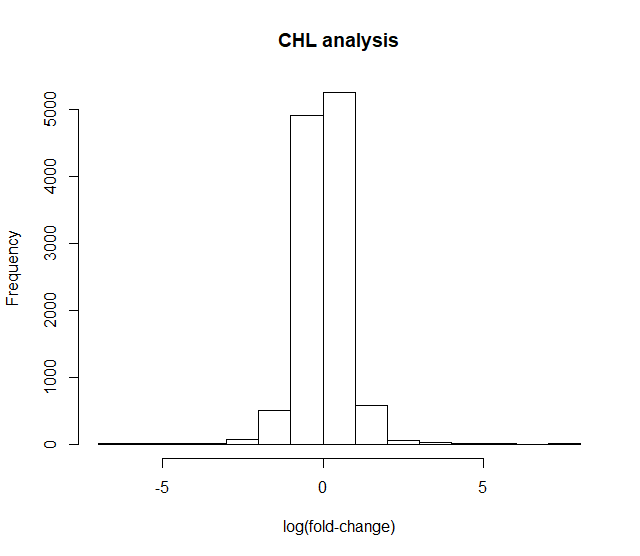
|  |  |
| --- | --- |
| **ProbeID** | **Log fold-change** |
| 1370902\_at | 7.3643 |
| 1387011\_at | 5.1641 |
| 1370583\_s\_at | 5.0693 |
| 1388122\_at | 4.7206 |
| 1369698\_at | 4.1827 |
| 1367577\_at | 3.8469 |
| 1374070\_at | 3.7551 |
| 1374610\_at | 3.4489 |
| 1367584\_at | 3.3983 |
| 1384331\_at | 3.3608 |

Histograms showing fold-change values from the significant DE genes in microarray analysis for each of the three chemicals can be seen in Figure 7. It shows that log fold change of DEGs for the three chemicals tends towards 0, indicating minimal changes in gene expression. The histogram for 3ME however, is centered at greater values between 0 and 1, indicating that it has more overall upregulation in DEGs. However, 3ME also has the smallest sample size of significant DEGs, and so this observation may possibly be an artifact of the data. Scatter plots showing the fold change vs nominal p-values for each chemical’s analysis can also be seen in Figure 8. These plots make the differences in sample sizes between the chemicals more evident, 3ME’s small number of datapoints and ill-defined shape intensifying its small sample size. The scatter plots for CLO and CHL both have well-defined shapes. The plot for CHL shows wider ranges of nominal p-values and log fold-change than CLO, but this can be explained by the former having twice as many DEGs than the latter. Besides this, both CLO and CHL graphs have well-defined characteristic shapes.

**A B**

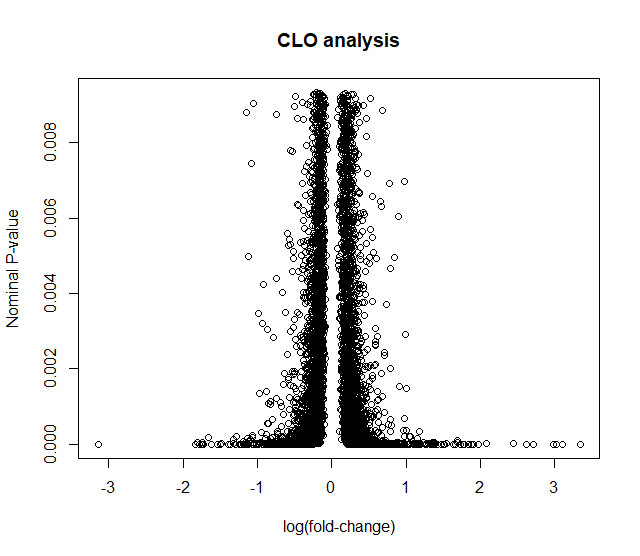
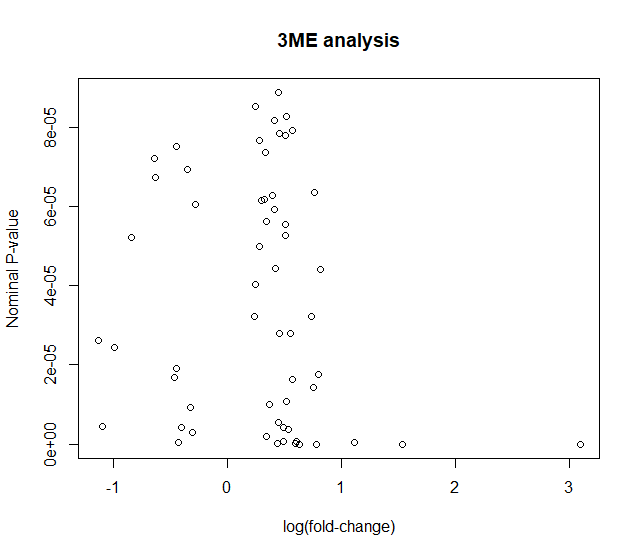


**C**

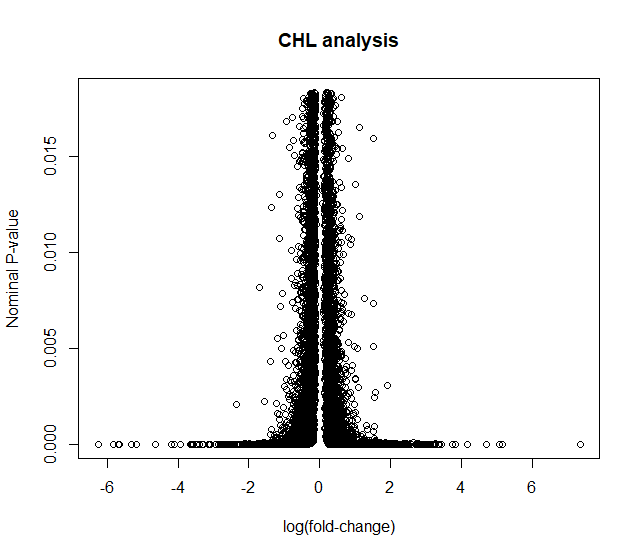


**Fig 7. Histograms showing log fold-change values of significant DEGs from Microarray analysis** for each of the three chemicals.

**A B**



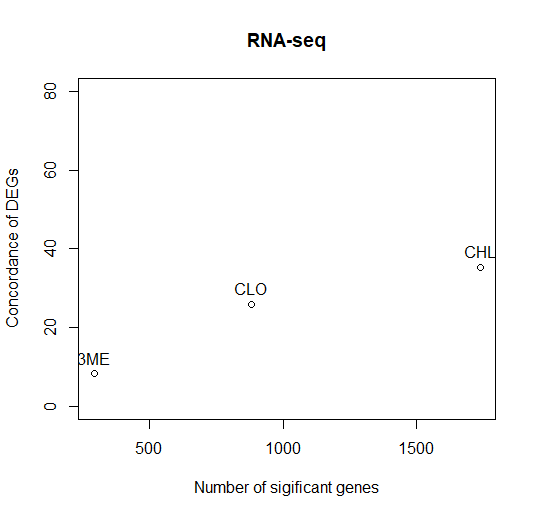
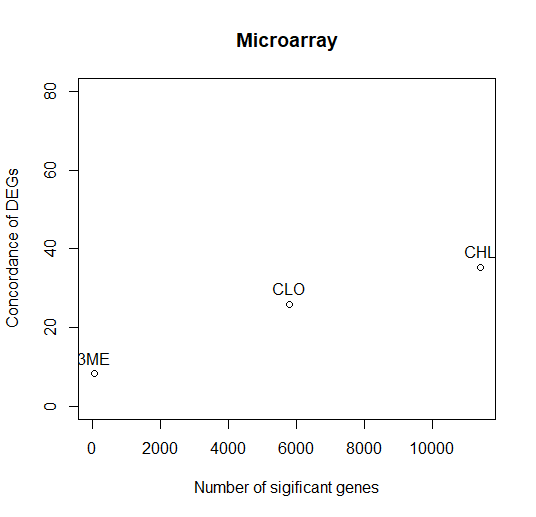
**C**



**Fig 8. Scatter plots showing the fold change vs. nominal p-values of significant DEGs from Microarray analysis** for each of the three chemicals.

Next, let’s look at the results from our concordance analysis, which are summarized in Figure 9 and table 11. Figure 9 contains plots showing concordance versus number of DEGs for each chemical analysis from both platforms. On comparing our Figure 9 with Figure 2a from Wang et al. (2014), which shows a similar plot for RNA-seq, it appears that our concordance value is lower for both 3ME and CHL. Since concordance is a measurement of relatedness between two datasets - with higher values indicating greater concordance - our analysis suggests the two platforms to be less related than Wang et al.’s analysis (2014). Our results also show fewer significant DEGs for CHL, but similar numbers for 3ME. The lower concordance can be attributed to our disclusion of duplicates while mapping refSeq-to-probe IDs. The fewer significant DEGs for CHL can be attributed to potential errors during our pre-processing steps for the RNA-seq data, which are outlined in the Methods section. However overall, our data follows a similar upward trend from the 3ME datapoint to CHL. This means that similar to Wang et al.’s data (2014), our results also had both higher concordance and larger number of significant genes for CHL than 3ME. CLO appears to be an intermediate between the 3ME and CHL, both in terms of number of significant genes and concordance. However since it was not plotted on Wang et al.’s (2014) Figure 2a, we are unable to compare our conclusions to their results.

**A B**

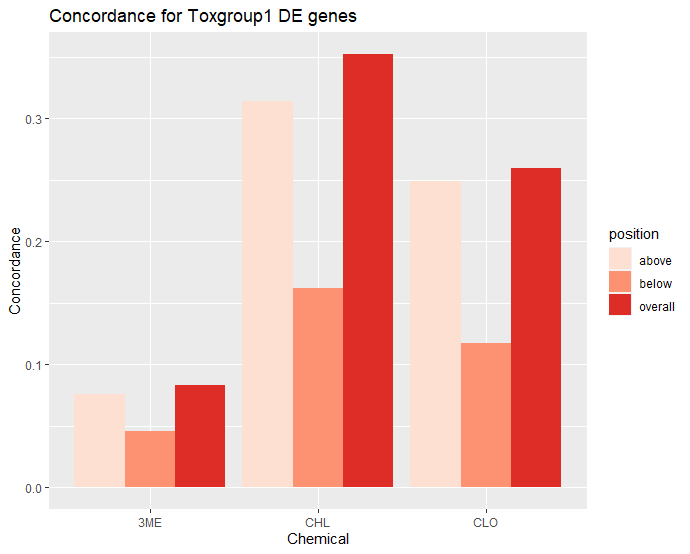


**Fig 9. Concordance vs number of DE genes from RNA-Seq and microarray analyses.**

Following this, DEGs for each chemical’s analysis were subdivided into two groups, one with gene expression above median, and the other below median. The median value for each chemical’s DE analysis was computed based on the average expression values for each gene in that set. Data was divided into these two sets for both RNA-seq and microarray data, using medians calculated for each of the 6 datasets. Concordance values were then computed for each of the resulting 6 pairs of DE genes data, and are summarized in table 11 and a bar plot (Figure 10), along with overall concordance values for each chemical. Figure 10 shows that in general, concordance is higher for each of the above-median and overall pairs, when comparing with the corresponding below-median sets for each chemical. This is similar to Wang et al.’s (2014) results in their Figure 3(b,c), which indicate that microarray and RNA-seq results tend to show higher concordance for above-median genes than below-median genes.

**Table 11. Concordance between RNA-seq and Microarray results** for overall DEGs, above-median subset and below-median subset.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Overall** | **Above median** | **Below median** |
| **3M** | 0.0833 | 0.07554876 | 0.04568921 |
| **CLOT** | 0.2593 | 0.2487223 | 0.117526 |
| **CHLR** | 0.3526 | 0.3141391 | 0.1623956 |



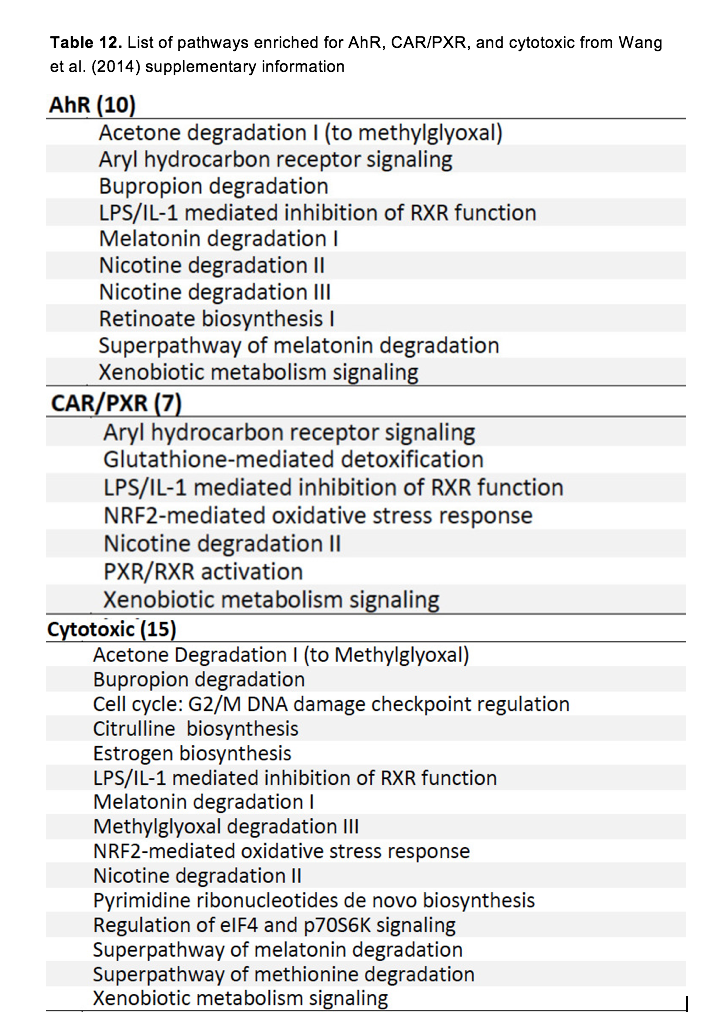
**Fig 10. Bar plot showing Concordance between RNA-seq and Microarray results** for overall DEGs, above-median subset and below-median subset.

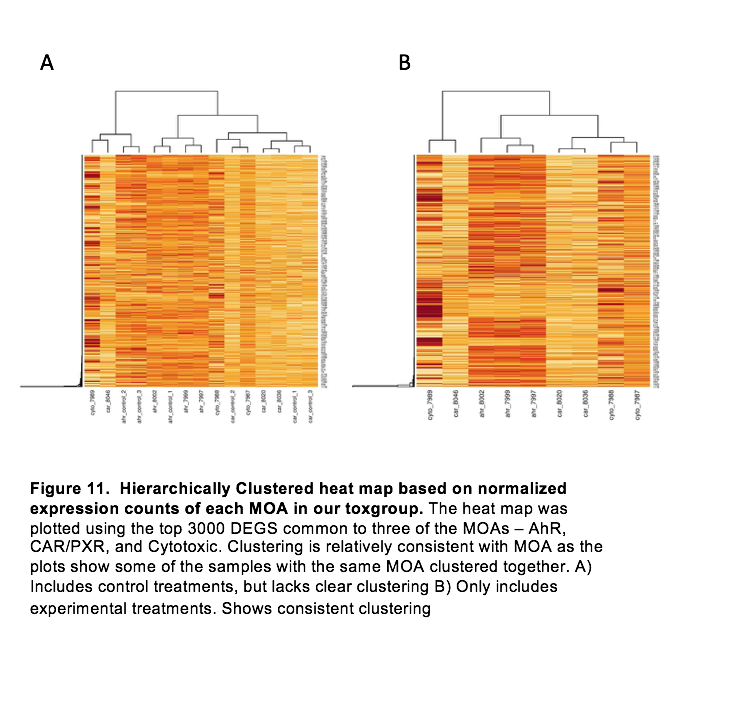
**Discussion**

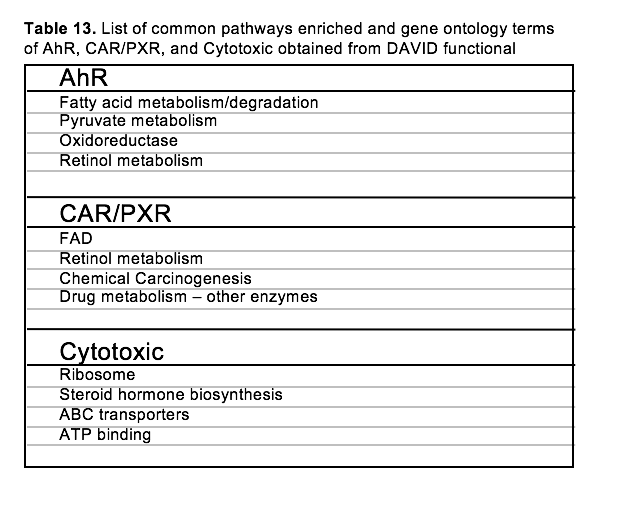
Concordance analysis (Figure 9) between the RNA-seq and microarray data revealed lower concordance than Wang et al.’s analysis (2014), but similar relative trends were observed between the CHL and 3ME. This can possibly be attributed to our disclusion of duplicates while mapping refSeq-to-probe IDs, and potential errors during data pre-processing. The similarity in our data’s trends lends further support to CHL showing higher concordance than 3ME. CLO appears to be an intermediate between the two along this trend, with concordance lying between those of 3ME and CHL. However, due to CLO being excluded from a key figure in Wang et al.’s paper (2014), we are unable to corroborate this observation with their results. Our results from the above- and below-median concordance analyses (Figure 10; Table 11) also suggests that concordance is higher for above-median expressed DEGs than the corresponding below-median sets. This is corroborated by results from Wang et al.’s (2014) results, lending further credibility to both sets of results.

Wang et al. (2014) performed a pathway enrichment analysis for each of the MOA chemical groups (Table 12), and found key pathways associated with xenobiotic activities in the liver. To perform a similar gene set enrichment analysis for the MOAs in our toxgroup – AhR, CAR/PXR, and Cytotoxic, the DEGs at p-adjust < 0.05 were filtered out from the DESeq2 result, and a functional annotation clustering was performed using the database for annotation, visualization, and integrated discovery (DAVID). In the results discovered from the DAVID analysis (Table 13) most of the enriched pathways/terms obtained are not similar to the enriched pathways in Wang et al. (Table 12). However, most of the enriched pathways obtained from the DAVID analysis are metabolic processes/biosynthesis involving an oxidation-reduction (redox) reaction, which are indicative of phase I of xenobiotic metabolism and activity, thus making the DAVID results relatively consistent with Wang et al. in regards to the pathway functionality (Croom, 2012).

To identify any clustering consistent with MOA, a clustered heat map was plotted using the top 3000 DEGs common to the three MOAs – AhR, CAR/PXR, and cytotoxic (Figure 11). In Figure 11A, the clustering between/among treatment and control groups is unclear. However, when control treatment groups were removed (Figure 11B), we observe a better and consistent clustering pattern, where all AhR treatment groups are clustered together, and a similar pattern of gene expression of the 3000 DEGs is present across the AhR treatment groups. Furthermore, a relatively similar clustering and gene expression pattern is present in the two CAR/PXR treatment groups, and the two Cytotoxic treatments groups.







**Conclusions**

In general, we believe this project was successful at reproducing most of the study, with exception to the concordance and clustered heat map analysis. In the concordance analysis, we had trouble calculating background corrected concordance that matches the paper’s result, but we worked around the caveats by spending time on understanding the concordance formula in the methods section of Wang et al. (2014). Although we were finally able to calculate the concordance in a reasonable range, we were unable to get the exact concordance as Wang et al, but overall, we observed a positive correlation between RNA-seq and microarray. Additionally, as mentioned in the discussion section, although we observed a better clustering when removing control treatments from the heatmap, we did not obtain a clear clustering consistent with MOA when control treatment groups were included in the heatmap (Figure 11A). This could be due to some error in the DESeq2 data or any upstream analysis.

**References**

Croom, E. (2012). Xenobiotic Metabolism. Retrieved from <https://www.sciencedirect.com/topics/medicine-and-dentistry/xenobiotic-metabolism>

Liao Y, Smyth GK and Shi W (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7):923-30. <http://www.ncbi.nlm.nih.gov/pubmed/24227677>

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](https://doi.org/10.1186/s13059-014-0550-8).

Philip Ewels, Måns Magnusson, Sverker Lundin and Max Käller Bioinformatics (2016). “MultiQC: Summarize analysis results for multiple tools and samples in a single report.”

doi: 10.1093/bioinformatics/btw354 PMID: 27312411

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W and Smyth GK (2015). “limma powers differential expression analyses for RNA-sequencing and microarray studies.” Nucleic Acids Research, 43, pp. doi: 10.1093/nar/gkv007.

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>