**Concordance of Microarray and RNA-Seq Differential Gene Expression**

**Introduction:**

Microarray and RNA-Seq, in principle, measure the same thing: the genome-wide abundance of mRNA molecules. The technologies used in RNA-Seq and microarrays are different, and even though the same experiment is being assayed, there can be substantial differences in the results. Wang et al. created a comprehensive study designed to generate Illumina RNA-seq and Affymetrix microarray data from the same set of liver samples from rats. The rats were under varying degrees of perturbation by 27 chemicals which represented multiple modes of action (MOA) (Wang et al 2014). The goal of the study was to characterize the concordance of differential gene expression across platforms then assess how effective each platform was at detecting expected pathway level effects based on each treatment’s MOA. Ultimately, the MOA prediction accuracy of each of the platforms was assessed.

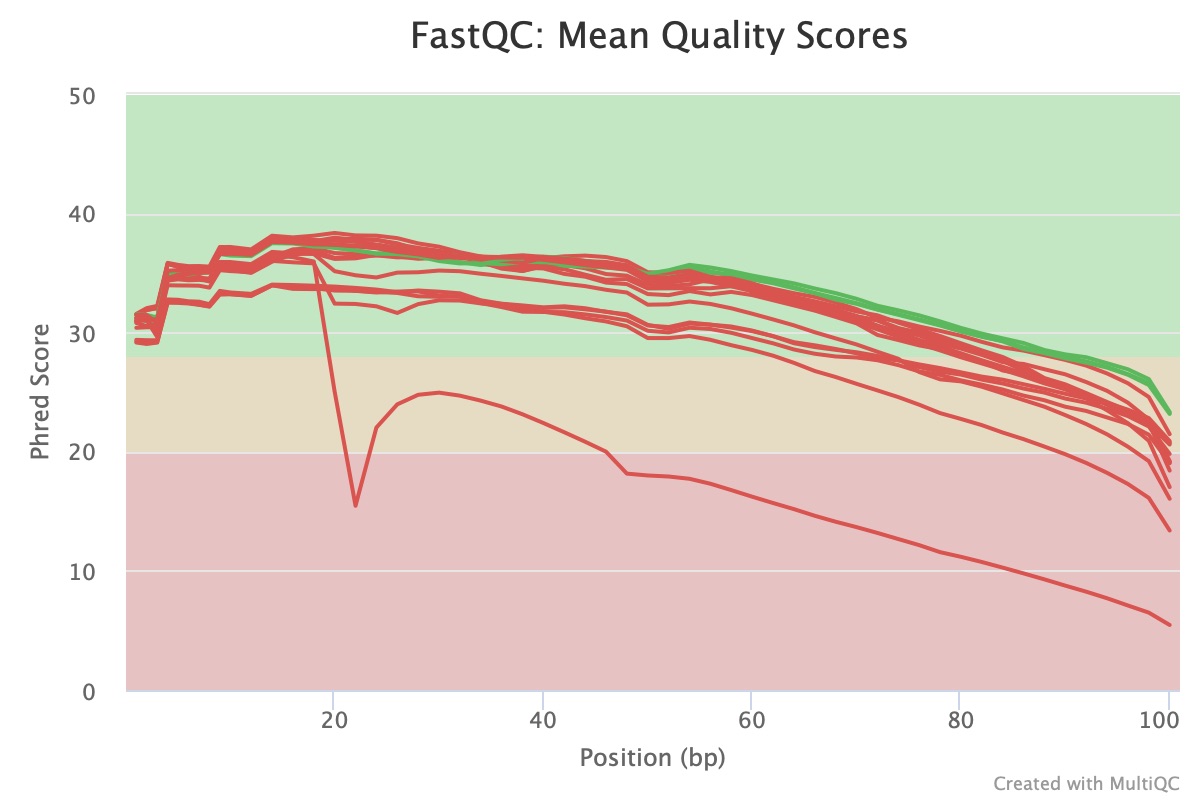
The authors of the experiment implemented bioinformatic techniques by using limma, edgeR, and DESeq in order to perform the RNA-seq analysis. The limma package (Ritchie et al. 2015) was originally developed for microarray analysis, and the authors added support for RNASeq after the high-throughput sequencing technique became available. Limma uses count data by performing a count transformation with a voom transformation procedure. Voom transforms counts by performing a counts per million (CPM) normalization then taking the logarithm of the CPM values and finally estimating the mean variance relationship across genes with the Bayes statistical framework. Limma performs statistical inference using a linear framework in order to make differential expression results with arbitrarily complex statistical models.

DESeq (Love 2014) and EdgeR (Robinson et. al 2010) are both bioconductor packages that take negative binomial regression to perform differential expression on the RNASeq data. DESeq and EdgeR both perform raw counts normalization, but the method they each use differs. EdgeR uses trimmed mean of M-values in order to do the normalization. DESeq uses the median of ratio method in order to normalize the counts. Both DESeq and EdgeR assume that the majority of the genes are not differentially expressed. All three methods -Limma, DESeq, and EdgeR- are used by the authors in order to identify differentially expressed genes from the RNA-seq data.

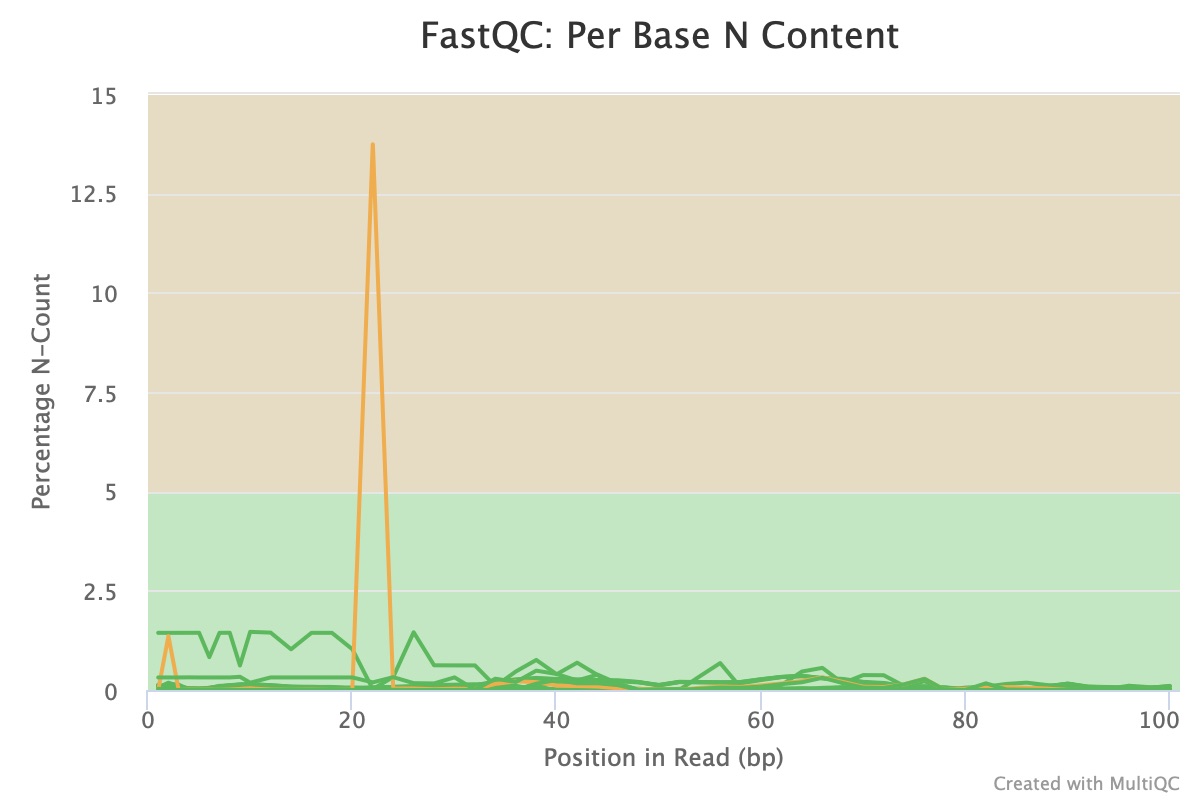
**Data:**

Male Sprague-Dawley rats, around the age of 6-8 weeks, were selected for experimentation. The test chemicals were administered either orally or by injection. Each of the subjects was dosed once a day for 3, 5, or 7 days depending on the particular chemical used. The livers were harvested 24 hours after the last dose (Wang et al 2014). RNA samples for RNA-seq were derived from the NTP DrugMatrix Frozen Tissue library. The samples were divided into two sets - training and test - which were sequenced and run. There were a total of 63 samples in the training set and 42 samples in the testing set. In the training set, 45 samples were derived from rats treated with the test chemicals while 18 were control samples. In the testing set, 39 samples were derived from the rats treated with test chemicals and 6 were from vehicle and route matched controls. Each test chemical had three rats, and for each MOA (mode of action), there were three test chemicals. There were a total of 5 MOAs in the test set and 4 MOAs in the training set. The RNA-seq of the test and training set were performed using Illumina. In the microarray analysis, the fragmented cRNA created from the rats liver RNA was hybridized to the Affymetrix gene chip.

From the metadata containing the sample information for each toxgroup, the first tox group was chosen to analyze further. The main focus was to perform sequencing data processing. The 15 samples from the first tox group contained 9 samples with 6 controls. In the nine samples, there were three different treatments: AhR, CAR/PXR, and Cytotoxic, with three replicates each. The nine samples were: SRR1177997, SRR1177999, SRR1178002, SRR1178020, SRR1178036, SRR1178046, SRR1177987, SRR1177988, and SRR1177989, all in fastq format. The samples were all paired end samples which were used when STAR alignment (Liao et al., 2014) was performed. FastQC (Andrews, 2010) was run on all the fastq samples in order to provide an overview of the basic quality control metric for the raw sequencing data. Each of the samples was aligned against the rat genome using the STAR aligner. From the output of STAR, bam files were created for each of the samples, which were used for analysis later on. The tool multiQC (Ewels 2016) was run on the fastQC and STAR results in order to combine the results into a single report.



**Figure 1: Mean Quality Scores.** A line graph containing all nine samples with Phred Score on the y-axis and bp position on the axis.

****

**Figure 2: Per Base N Content.** A line graph containing all nine samples with percentage N-count on the y-axis and bp read position on the x-axis

In Figure 1, the phred scores for each of the samples reported from fastQC are plotted. For the majority of the samples, the phred score ranged from 30-40 with an increase in the beginning positions and then a drop off towards the end, which is considered an expected pattern. In the figure, it shows that 15 out of the 18 samples failed with only 3 out of the 18 passing. Even though most of the samples technically failed, they seem to follow the same pattern. The phred score range is not as concerning, since it still falls in the phred score range of 30-40 which indicates an accuracy of 99.99 -99.9%. A phred score of 20 indicates an accuracy of 99%. It is unclear why the majority of the samples which still have the same pattern and around the same phred score failed, but this could be due to the way the samples were processed. The phred scores of the majority of the samples were in a range where it is still an acceptable sample. In the per sequence quality scores, all the samples passed with the exception of one, which is why it is not as concerning why they did not pass the mean quality score. The most concerning aspect of this plot is the line that does not follow the same pattern as the rest, sample SRR1170836\_2. Sample SRR1170836 also failed the per sequence quality scores. Since the phred score is so low for the sample, it raises concern since there is a higher chance that a base could have been called incorrectly. Figure 2 (shows the per base N content of each of the samples, which is the percent of bases at each position or bin with no base call. Sample SRR117036 was the only sample that caused a warning to be raised because it should have an N content of >5%, yet the sample ended up having an N content of around 13.73%. In order to fail the test, the N content would need to be >20%, so the sample was approximately 7% off from failing. Overall, SRR117036 was the only sample that raised concern in the overall data quality for this experiment. SRR117036 could have been improperly downloaded or sequenced incorrectly, which could have resulted in these failures. Since it is only one of the 18 samples that raised concern, the data is still suitable for the analysis; it is just an area to look out for if sample SRR117036 raises issues further down in the analysis. If time permitted, it would be ideal to look into how sample SRR117036 was derived and perhaps take it again to see if the quality results change.

**Table 1: STAR Alignment Statistics.** A table containing the statistical outputs of STAR for each sample including number of input reads, uniquely aligned reads, multi-mapped reads, unmapped reads, and chimeric reads.

| **Sample Name** | **Number of Input Reads** | **Uniquely Aligned Reads** | **Multi Mapped Reads:**  Mapped to multiple loci | **Multi Mapped Reads:**  Mapped to too many loci | **Unmapped Reads:** too many mismatches (%) | **Unmapped Reads:** too short (%) | **Chimeric Reads** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SRR1177987 | 14733393 | 14785610 (84.81%) | 634883  (3.64%) | 26136  (0.15%) | 0.0% | 11.36% | 0  (0.0%) |
| SRR1177988 | 18368879 | 15667286  (85.29%) | 650689  (3.54) | 15207  (0.08%) | 0.0% | 11.05% | 0  (0.0%) |
| SRR1177989 | 18804974 | 15707078  (83.53%) | 818451  (4.35%) | 96880  (0.52%) | 0.0% | 11.55% | 0  (0.0%) |
| SRR1177997 | 19746775 | 17608043  (89.17%) | 767134  (3.88%) | 35390  (0.18%) | 0.0% | 6.67% | 0  (0.0%) |
| SRR1177999 | 21838440 | 19374545  (88.72%) | 855549  (3.92%) | 56592  (0.26%) | 0.0% | 6.98% | 0  (0.0%) |
| SRR1178002 | 18844950 | 16796763  (89.13%) | 737790  (3.92%) | 41352  (0.22%) | 0.0% | 6.59% | 0  (0.0%) |
| SRR1178020 | 13374032 | 13374032  (83.57%) | 784376  (4.90%) | 51179  (0.32%) | 0.0% | 11.10% | 0  (0.0%) |
| SRR1178036 | 16867274 | 11443947  (67.85%) | 676001  (4.01%) | 31941  (0.19%) | 0.0% | 27.08% | 0  (0.0%) |
| SRR1178046 | 17652475 | 15067177  (85.35%) | 876408  (4.96%) | 41360  (0.23%) | 0.0% | 9.36% | 0  (0.0%) |

Table 1 shows the statistical output from the STAR alignment for each sample. Overall, in each of the samples the statistics are roughly around the same values. The majority of the reads were uniquely aligned reads with only under 5% of the reads for each sample being mapped to multiple loci. In the unmapped reads, there were no reads that were not matched due to too many mismatches with the reference genome. The only possible concern was reads that were too short. For the most part, the majority of the samples did not raise any concerns on overall data quality. Sample SRR117036 does seem to deviate from the other sample values, especially in the uniquely aligned reads. SRR117036 did raise concern as previously mentioned in the fastQC analysis, so it is not surprising that it is also raised concern in the STAR alignment analysis. The rest of the samples tend to have uniquely aligned reads in the range of 80-90% while SRR117036 is only at 67.85%. It is not a significantly low match up rate but it does have a high unmapped reads percentage for the too short category at 27.08%. This does raise concern in whether it plays a role in why the uniquely mapped reads are lower. One reason could be that when SRR117036 was sequenced, more shorter reads were created which altered the way that the STAR alignment tool aligned it with the genome. It would be interesting to look into the sample and see why the data results are different from the rest of the samples. It would have been expected that all the samples would have roughly around the same values if they were all processed the same way.

**Methods:**

Sample data and bioinformatics tools mirroring those employed by the authors of the previous experiment were used to quantify reads and assess differential gene expression in this experiment. For each of the nine samples retrieved, the program featureCounts was applied to the raw reads generated by the STAR aligner to quantify raw RNA counts per gene (Liao et al., 2014). featureCounts computes the number of reads in an alignment file that align to a given genome annotation. Due to the high demand on computational resources, read quantification was performed on the scc2 remote server. The quality of read quantification was assessed using multiqc. multiqc conveniently aggregates quality control metrics across multiple sample files to facilitate sample quality determinations (Ewels et al., 2016). Sample counts were aggregated into a single comma-delimited count matrix in R. The distribution of counts across samples was visualized in a boxplot using the open-source data visualization ggplot2 (Wickham, 2016). A log10 scaling was applied to the counts for visualization to reduce the range of y-values, and consequently improve visibility of the count distribution. Control sample counts were added to the aggregated count matrix for downstream differential expression analysis. Genes with a count less than ten across all samples were removed from the count matrix. This removed undesirable signals from the data and improved the power of downstream statistical testing. Differential expression analysis was performed within treatment groups between treatment sample counts and control sample counts. Differential expression analysis was performed using DESeq2. DESeq2 is a well established software for differential expression analysis that uses a model based on the negative binomial distribution of gene expression counts (Love et al. 2016). Three lists of differentially expressed genes and their associated test statistics were generated from the differential expression analyses. The top ten genes from each analysis were aggregated into a single data table. Histograms of the fold change values from each analysis were generated. Additionally, scatterplots of the fold change values versus nominal p-values were generated for each analysis.

Next, Limma was used to run differential expression analysis of the toxgroup 1 samples vs controls found in the normalized expression matrix. Limma (Ritchie et al., 2015) is a very mature Bioconductor package that implements the analytical methodology for microarray DE analysis. As mentioned in the introduction, Limma accomplishes differential expression analysis by performing a count transformation using a voom transformation procedure. Voom transforms counts by performing a counts per million normalization then takes the logarithm of the CPM values and estimates the mean variance relationship across genes using the Bayes statistical framework. Pre–normalized RMA normalization values were obtained from the SCC and used in conjunction with limma to determine the differential expression between the tox group 1 treatments and control samples. Since the pre-normalized control values were already available for use, it was not necessary to perform normalization and quality control prior to Limma differential expression analysis. The first step of this procedure was to subset the control values found in the full RMA expression matrix to those just found in the tox group 1 csv file. A design matrix was created that modeled the treatment versus control groups for use in Limma differential expression analysis. Limma differential expression analysis was performed with the voom parameter enabled, and the results were written out to a csv file for further analysis. Next, the limma results were sorted in descending order by adjusted p value. Tables showing the number of genes in each treatment significant at an adjusted p-value < 0.05 and the top 10 differentially expressed genes for each treatment were generated from the results of the Limma DE analysis. Histograms of fold change values from the significant DE genes were created for each of the three treatment groups, and scatter plots of fold change vs nominal p-value were created using the open source data visualization package, ggplot (Wickham).

The next portion of the analysis involved calculating the concordance between microarray and RNA-Seq DE genes. An equation used by Wang et al. to calculate cross-platform concordance was used for this analysis and is displayed below (Wang et al., 2014) :

The first step of implementing the concordance method was to map Affymetrix probe IDs from the microarray analysis to refSeq identifiers used by the RNA-Seq analysis. This was accomplished by merging the affy table with the limma table, subsetting only the REFSEQ column of the affy table, and removing any duplicate entries. Once probe IDs were mapped to refSeq identifiers, refSeq identifiers were converted to genes for both the microarray data and the RNA-Seq data. The number of genes that had the same DE directionality were found for each of the three treatments (3ME, CLO, CHL), and this metric was used to calculate concordance along with the total number of genes in the microarray data, the total number of genes in the RNA-seq data, and the total number of genes in the genome (30,000). This provided a cross platform concordance score for each of the three treatments. Scatterplots for concordance vs the number of DE genes for both DEseq and Limma analyses were created using the ggplot package (Wickham). Finally, DE genes were subdivided into “above-median” and “below-median” groups for each of the sample groups. Concordance of each of these separate groups of genes was calculated using the median value, which was obtained from the baseMean column in the DESeq2 results and the AveExpr column of the microarray results. This value corresponds to the overall mean count of each gene across all samples in the comparison. For each treatment, the concordance score was calculated using “all the genes”, “above-median genes”, and “below-median genes”. A barplot that showed the combined overall concordance measures and the above/below-median subsets for each of the three treatment groups was generated using the ggplot package (Wickham, 2016) .

This portion of the analysis was run locally, and took several minutes to complete. The very long run time can be attributed to initial reading of the rma-matrix file, which was a very large 654.2 megabyte text file. In the future, it would be prudent to perform these analyses on the SCC, as the cluster is much better suited for carrying out computational tasks involving such large files than a local machine.

**Results:**

Table 2 shows read quantification statistics from multiqc. These statistics indicated a tight range of read assignments across samples (59%-62%). The count matrix compiled from featureCounts featured 18,014 genes. A boxplot as shown in Figure 3 revealed the raw count distribution on a log10 scale. This boxplot was generated from the compiled count matrix. The count distribution box plot showed that each sample had a small portion of some genes very highly expressed relative to the median expression of genes in samples. Similarly, the count distribution box plot showed that each sample had some genes very lowly expressed relative to the median expression of genes in samples. 5164 low abundance genes were filtered from the count matrix after adding control counts. After differential expression analysis, 317 in the AhR,3-METHYLCHOLANTHRENE treatment group, 1812 genes in the Cytotoxic treatment group, and 948 genes in the CAR/PXR treatment group were reported to have an adjusted p-value less than 0.05. These results are shown in Table 3. The ten genes from each treatment group with the lowest adjusted p-values are summarized in Table 4. Histograms of the fold change values for each experiment, as shown in Figure 4, revealed a high frequency of fold change values close to zero. Scatterplots of the fold change values versus nominal p values, as shown in Figure 5, revealed a trend of decreasing p-values as fold change values gained distance from 0.

**Table 2. featureCounts multiqc Report.** A table of the percent of unique reads and count of unique reads assigned to the given genome annotation was generated by multiqc.

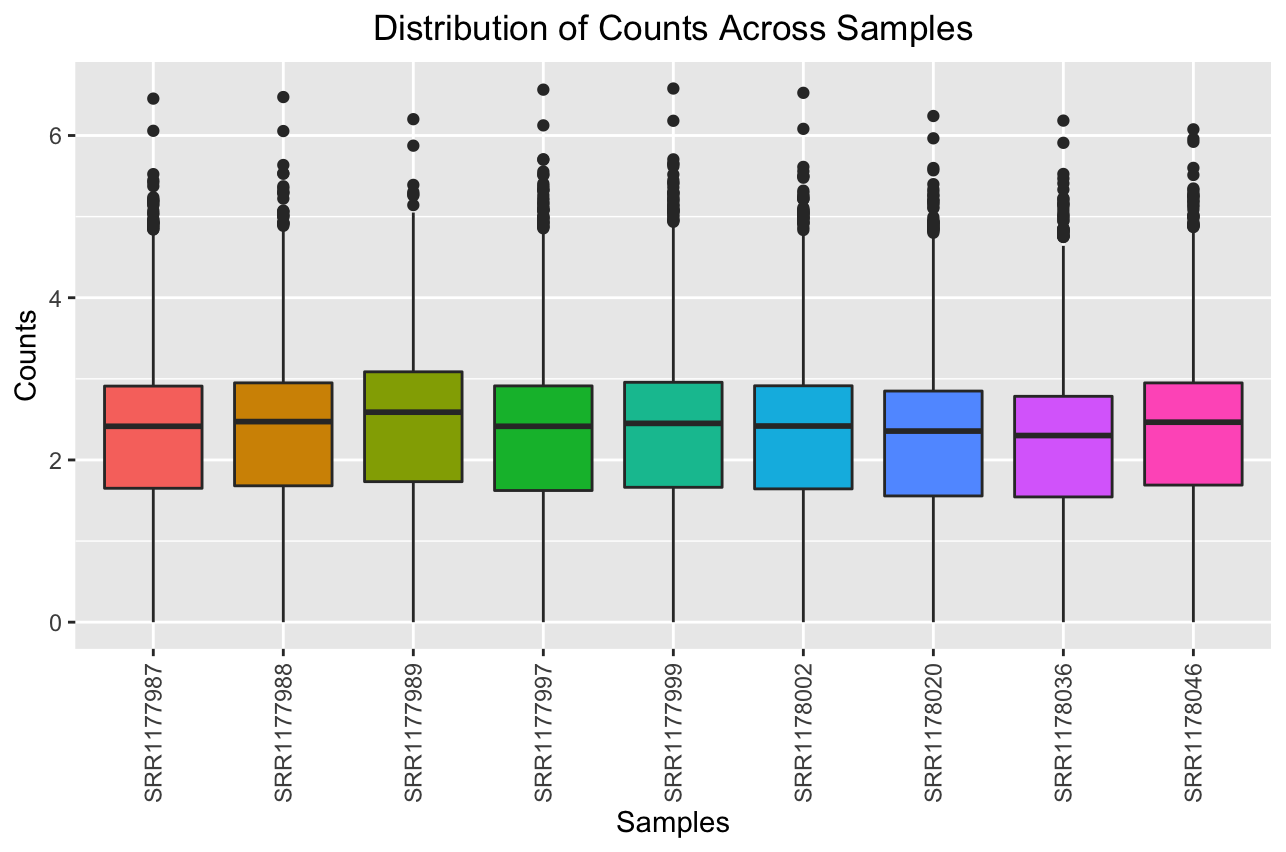
| **Sample Name** | **% Assigned** | **M Assigned** |
| --- | --- | --- |
| SRR1177987 | 61.90% | 20.4 |
| SRR1177988 | 61.90% | 21.5 |
| SRR1177989 | 59.10% | 21.2 |
| SRR1177997 | 62.60% | 24.7 |
| SRR1177999 | 62.40% | 27.1 |
| SRR1178002 | 61.80% | 23.3 |
| SRR1178020 | 59.40% | 18.4 |
| SRR1178036 | 59.10% | 15.7 |
| SRR1178046 | 59.30% | 20.7 |

**Table 3. Differentially expressed genes with adjusted p-value < 0.05.** A table of each treatment group tested for differential gene expression against associated control sample counts and the resulting number of genes with adjusted p-values < 0.05.

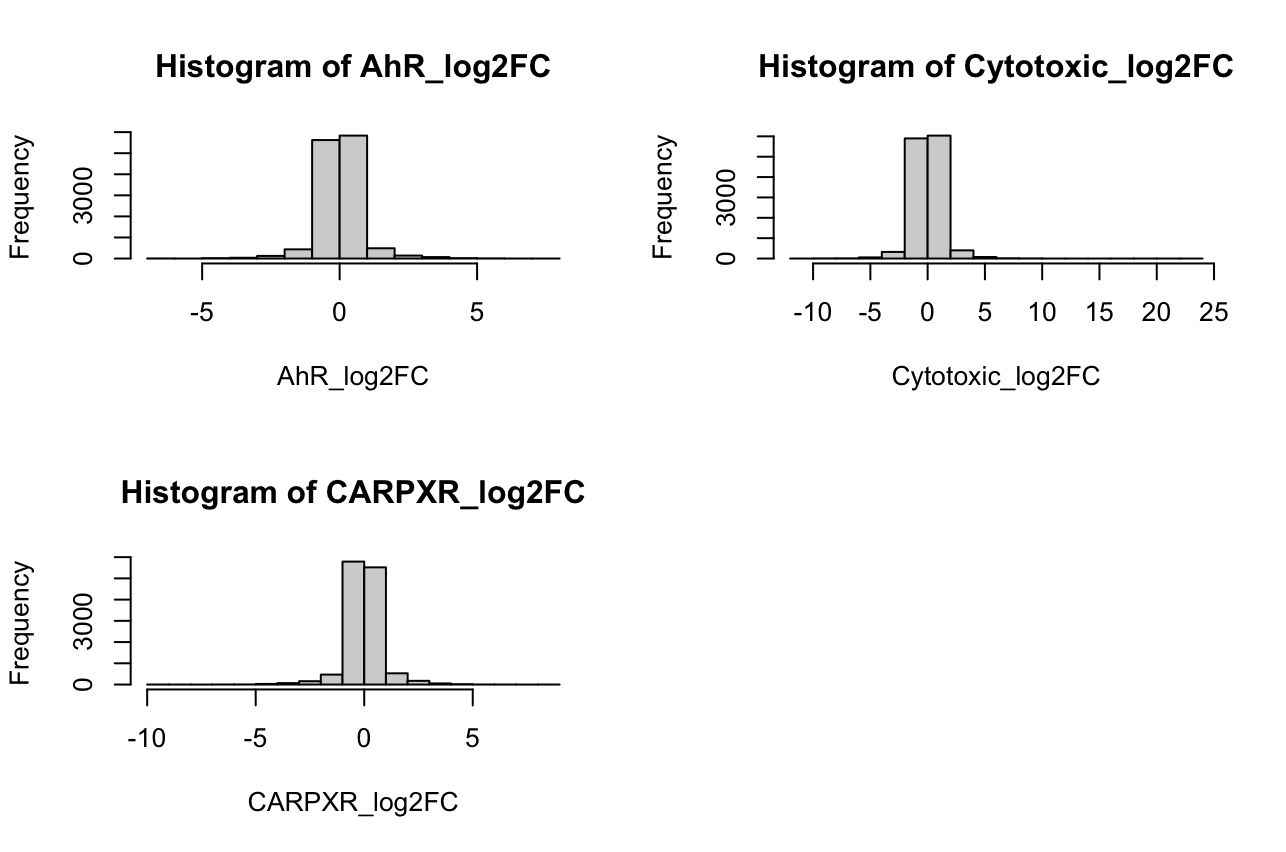
| mode\_of\_action | padj<0.05 |
| --- | --- |
| AhR,3-METHYLCHOLANTHRENE | 317 |
| Cytotoxic | 1812 |
| CAR/PXR | 948 |

**Table 4. Top Ten Genes with Lowest Adjusted P-values per Treatment Group.** A table showing the top ten genes with the lowest adjusted p-values in each treatment group determined by differential expression analysis against control sample counts using DESeq2.

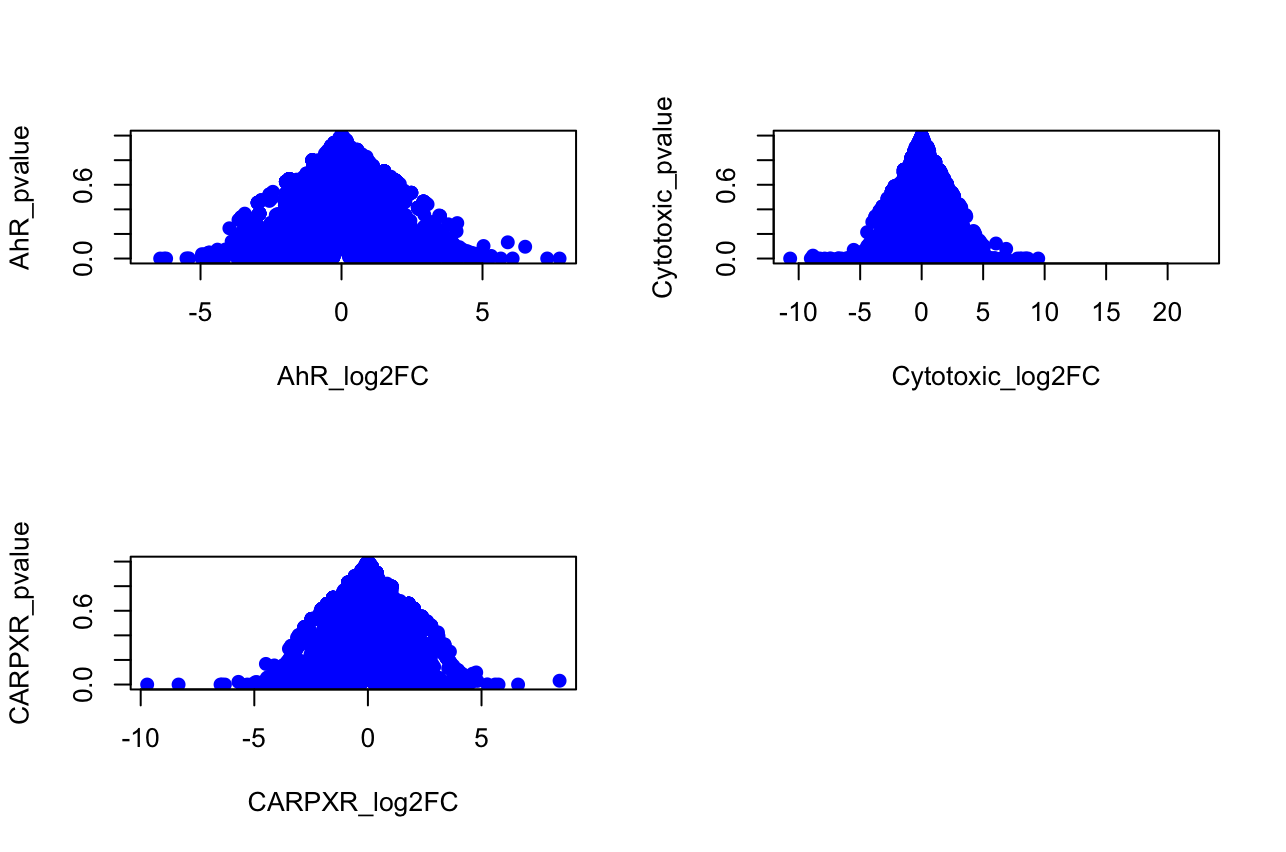
| Gene | mode\_of\_action |
| --- | --- |
| NM\_012541 | AhR,3-METHYLCHOLANTHRENE |
| NM\_130407 | AhR,3-METHYLCHOLANTHRENE |
| NM\_022521 | AhR,3-METHYLCHOLANTHRENE |
| NM\_023094 | AhR,3-METHYLCHOLANTHRENE |
| NR\_046239 | AhR,3-METHYLCHOLANTHRENE |
| NM\_001109022 | AhR,3-METHYLCHOLANTHRENE |
| NM\_031972 | AhR,3-METHYLCHOLANTHRENE |
| NM\_134329 | AhR,3-METHYLCHOLANTHRENE |
| NM\_175761 | AhR,3-METHYLCHOLANTHRENE |
| NM\_053883 | AhR,3-METHYLCHOLANTHRENE |
| NM\_203512 | Cytotoxic |
| NM\_001257095 | Cytotoxic |
| NM\_080581 | Cytotoxic |
| NM\_023978 | Cytotoxic |
| NM\_012844 | Cytotoxic |
| NM\_013215 | Cytotoxic |
| NM\_139115 | Cytotoxic |
| NM\_012540 | Cytotoxic |
| NM\_001010921 | Cytotoxic |
| NM\_130741 | Cytotoxic |
| NM\_001010921 | CAR/PXR |
| NM\_080581 | CAR/PXR |
| NM\_013105 | CAR/PXR |
| NM\_012844 | CAR/PXR |
| NM\_131903 | CAR/PXR |
| NM\_173295 | CAR/PXR |
| NM\_017272 | CAR/PXR |
| NM\_013215 | CAR/PXR |
| NM\_001134844 | CAR/PXR |
| NM\_133586 | CAR/PXR |



**Figure 3. Boxplot of Count Distribution Across Samples.** A boxplot showing the distribution of counts in each sample on log10 scale. A wide distribution of the expression across genes within samples is shown in this boxplot.



**Figure 4. Histograms of Fold Change Values Across Treatment Groups.** Histograms of the fold change values for each treatment group show a majority of fold change values close to 0 for each treatment group.



**Figure 5. Scatterplot of Fold Change Values versus Nominal P-values Across Treatment Groups.** Scatterplot of the fold change values against nominal p-value for each treatment group show an increase in p-value as fold change approaches 0.

Table 5 shows the number of DE genes at p-adjust < 0.05 for each toxin/chemical. There were a total of 235 differentially expressed genes for 3-methylcholanthrene, 10587 for chloroform, and 3975 for clotrimazole. All these genes met the criteria of having an adjusted p-value of <0.05. Table 6 shows the top 10 differentially expressed genes for each treatment, sorted by p-value in descending order. The p-values for the top 10 DE genes ranged from 0.00377 to 0.000334 for 3-Methylcholanthrene, 0.006384 to 0.006332 for clotrimazole, and 0.01701 to 0.01694 for chloroform.

Figure 6 consists of three histograms, which each show the relationship between fold change values and frequency for each of the three toxins/chemicals after Limma analysis. The shape of each histogram was relatively similar, with a high frequency of fold change values close to zero. The distribution of fold change values differed somewhat between each of the three toxins/chemicals, with the CHL plot showing the widest variance in LogFC relative to the other two.

Scatter plots of the fold change values versus nominal p values after Limma analysis, as shown in Figure 7, revealed a trend of decreasing p-values as fold change values gained distance from 0. The density of these scatter plots also varied, with the 3ME plot being noticeably sparser than the other two. This can be attributed to the very low total number of DE genes in the 3ME group (235) relative to the other two toxins/chemicals (3975 for CLO and 10587 for CHL).

Table 7 shows the concordance between overall, above/below median genes for all three analyses. The values for above median genes were: 0.029 (3ME), 0.011 (CLO), and 0.47 (CHL). The values for below-median genes were: 0.017 (3ME), 0.0089 (CLO), and 0.28 (CHL). Finally, the cross-platform concordance values for the all genes group were: 0.033 (3ME), 0.017 (CLO), and 0.4686 (CHL). Above-median concordance values were higher across the board compared to the below-median concordance values, which is similar to what Wang et al. discovered in their study.

Figure 8 is a plot of overall concordance vs number of DE genes from both the Limma and RNAseq analysis**.** CHL/cytotoxic MOA showed the highest concordance, by far, in both analyses. Scores were very similar in each, at roughly 47%. This was the one chemical/toxin that showed a concordance value that was extremely similar to the result found in the Wang et al. study. Both 3ME and CLO had extremely low concordance values, ranging from 1 to 4%. This was expected for 3ME, considering its very small treatment effect size, but not for CLO. Aside from CLO, there is a clear correlation between the number of DE genes and concordance, as indicated by the trendline in each plot. This corroborates the findings of Wang et al. Finally, figure 9 is a combined bar plot that displays overall, above, and below median genes for each of the three chemicals/toxins. Cross platform concordance was consistently higher for above-median groups when compared to below-median groups, regardless of the particular toxin/chemical. Cross platform concordance of the above-median group was very similar to that of the “all genes” group, with a few chemicals/toxins falling very slightly below the concordance score of the all genes category (3ME and CLO) and one falling slightly above (CHL). All tables and figures are shown below:

**Table 5. Differentially expressed genes with adjusted p-value < 0.05 (Limma).** A table of the number of differentially expressed genes (padj < 0.05) for each chemical

| **Chemical** | **padj < 0.05** |
| --- | --- |
| 3-methylcholanthrene | 235 |
| Chloroform | 10587 |
| Clotrimazole | 3975 |

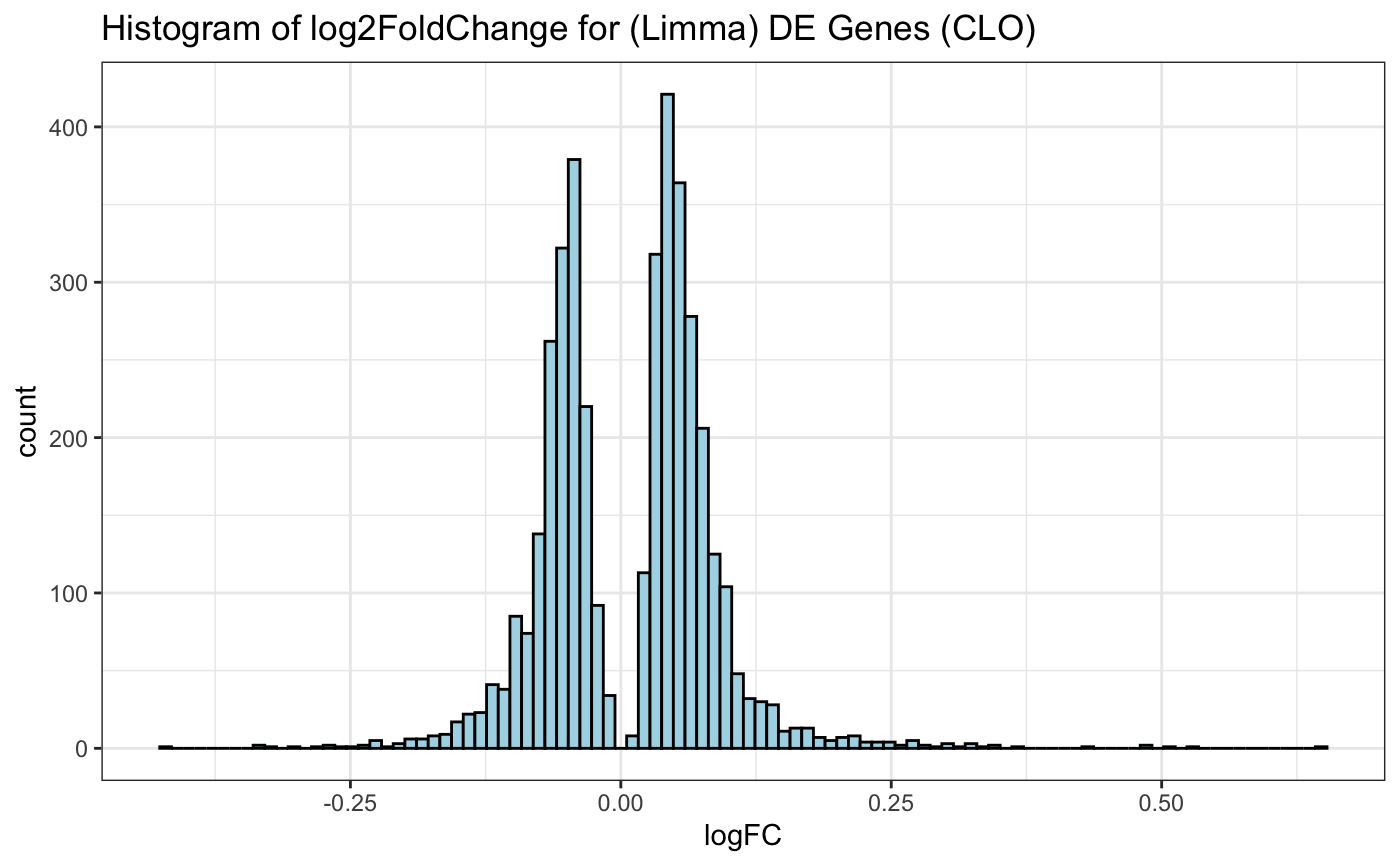
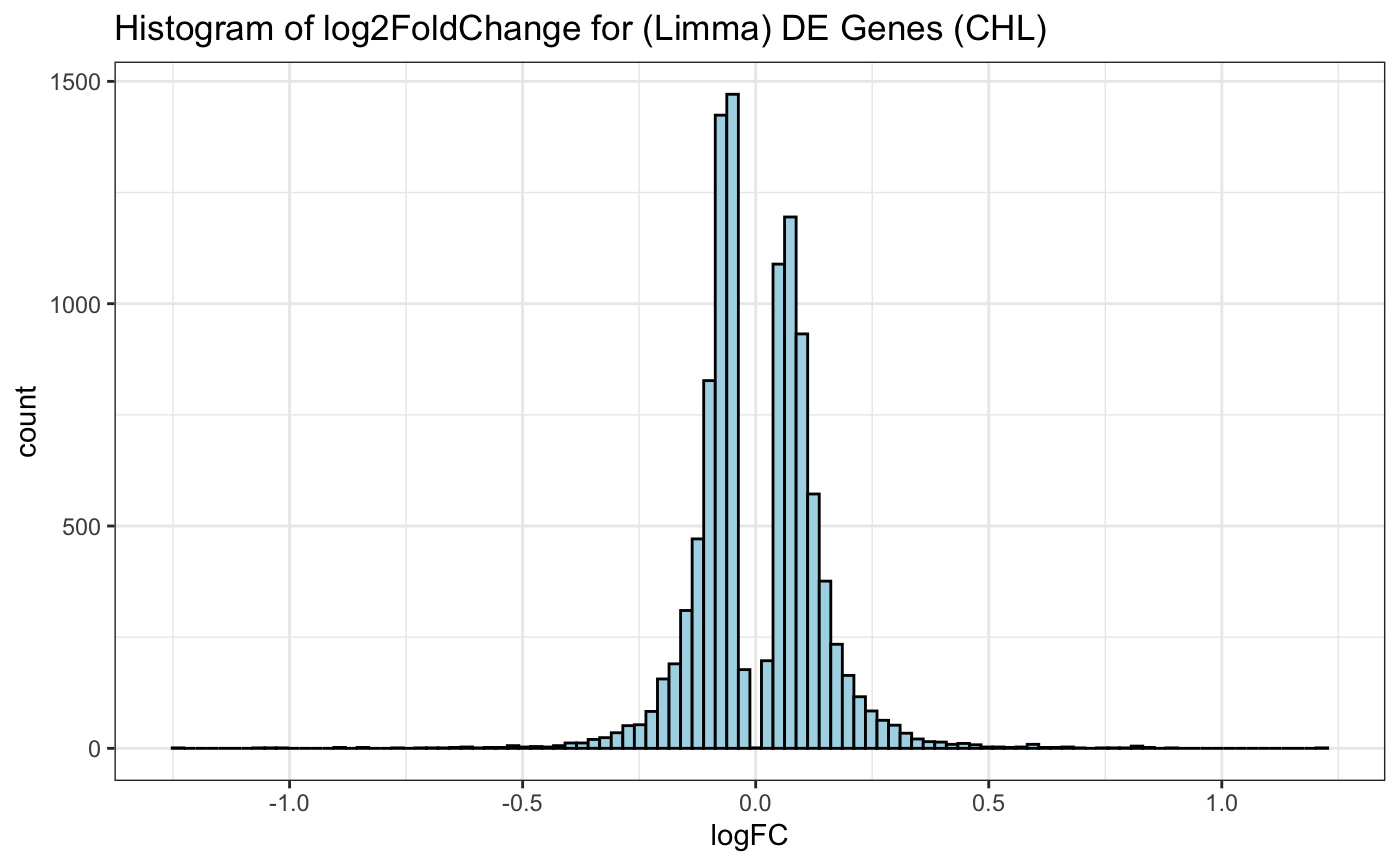
**Table 6. Differentially expressed genes with adjusted p-value < 0.05 (Limma).** A table of each treatment group tested for differential gene expression against associated control sample counts and the resulting number of genes with adjusted p-values < 0.05.

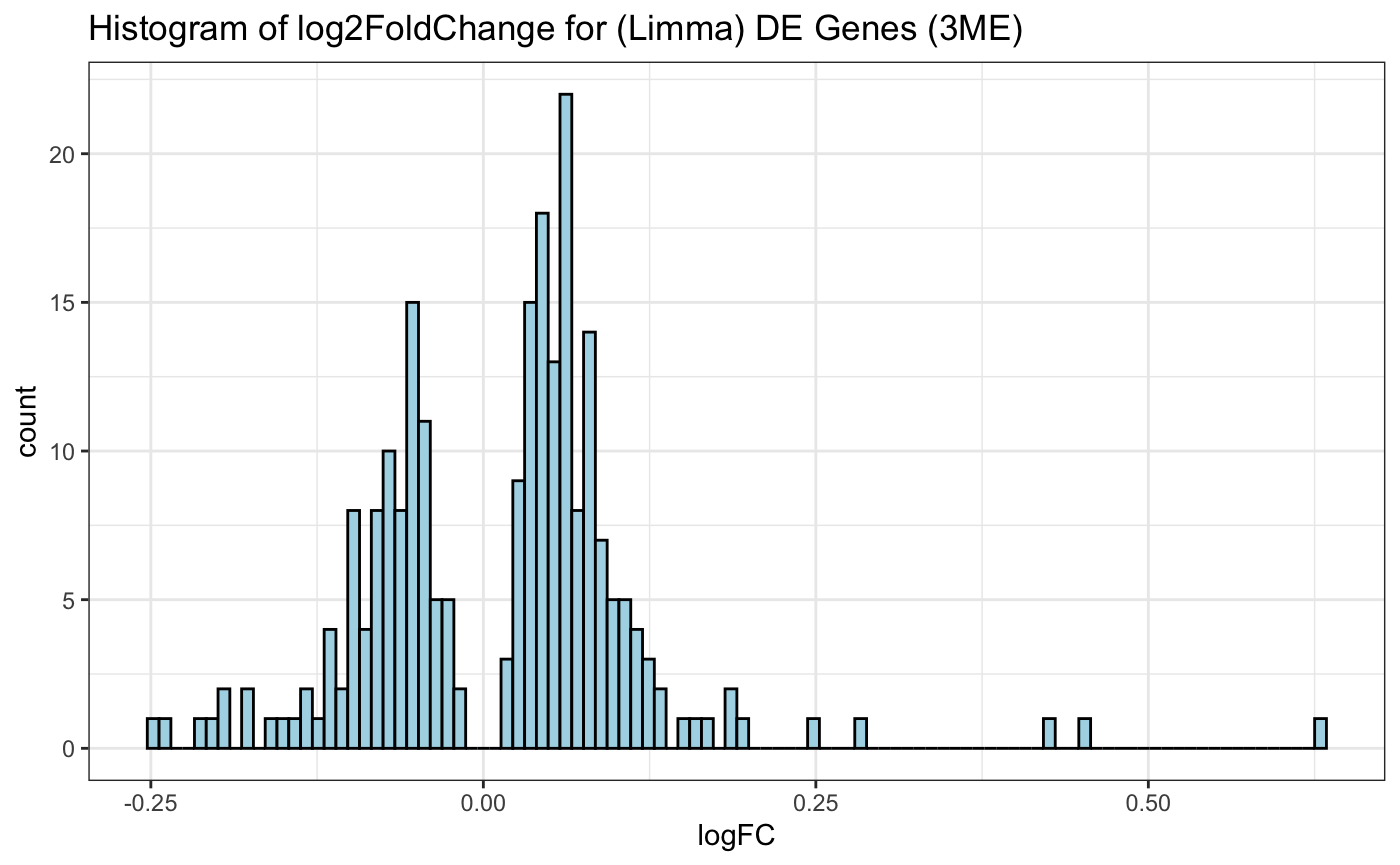
| **Probe ID** | **Toxin/chemical** |
| --- | --- |
| 1367507\_at | 3-METHYLCHOLANTHRENE |
| 1385240\_at | 3-METHYLCHOLANTHRENE |
| 1373337\_at | 3-METHYLCHOLANTHRENE |
| 1376795\_at | 3-METHYLCHOLANTHRENE |
| 1380030\_at | 3-METHYLCHOLANTHRENE |
| 1378504\_at | 3-METHYLCHOLANTHRENE |
| 1373756\_at | 3-METHYLCHOLANTHRENE |
| 1383732\_at | 3-METHYLCHOLANTHRENE |
| 1380125\_at | 3-METHYLCHOLANTHRENE |
| 1393719\_at | 3-METHYLCHOLANTHRENE |
| 1395889\_at | Chloroform |
| 1398970\_at | Chloroform |
| 1381392\_at | Chloroform |
| 1372917\_at | Chloroform |
| 1381460\_at | Chloroform |
| 1376865\_at | Chloroform |
| 1380397\_at | Chloroform |
| 1389207\_at | Chloroform |
| 1385505\_at | Chloroform |
| 1374077\_at | Chloroform |
| 1386530\_at | Clotrimazole |
| 1389520\_at | Clotrimazole |
| 1371651\_at | Clotrimazole |
| 1377222\_at | Clotrimazole |
| 1391639\_at | Clotrimazole |
| 1388519\_at | Clotrimazole |
| 1372562\_at | Clotrimazole |
| 1378493\_at | Clotrimazole |
| 1385428\_at | Clotrimazole |
| 1375551\_at | Clotrimazole |

**Table 7.** **Concordance between overall and above/below median genes for all three analyses.**

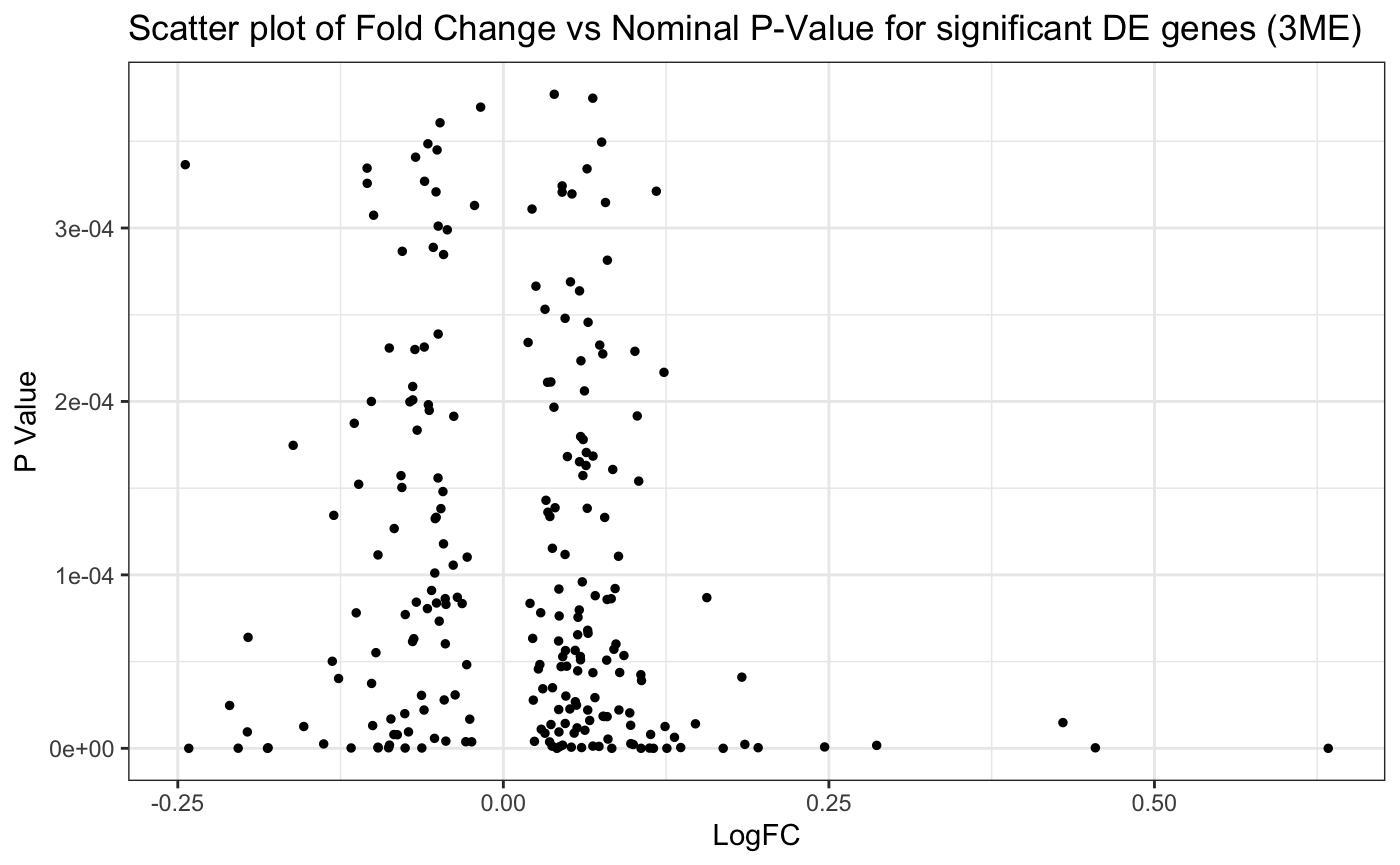
Genes in each treatment group were subsetted into three categories: all genes, above-median, and below-median. One of the main findings of the Wang et al. study was that concordance tends to be higher for genes with higher than average gene expression than those with lower than average gene expression, which is exactly what is seen here.

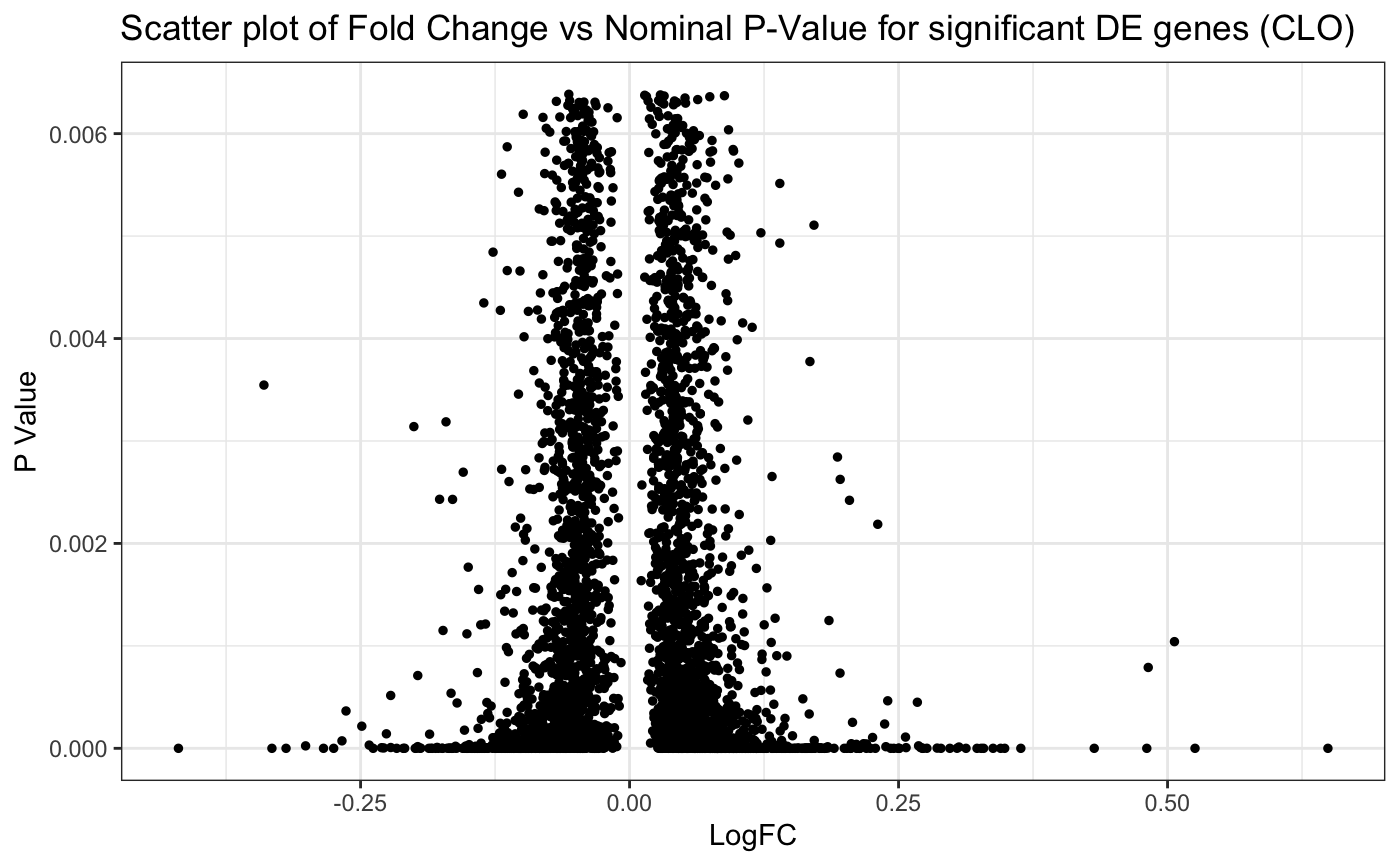
| **Toxin/Chemical** | **Gene subset** | **Concordance score** |
| --- | --- | --- |
| 3-METHYLCHOLANTHRENE | All genes | 0.0333 |
| 3-METHYLCHOLANTHRENE | Above-median genes | 0.02913 |
| 3-METHYLCHOLANTHRENE | Below-median genes | 0.017 |
| Chloroform | All genes | 0.4686 |
| Chloroform | Above-median genes | 0.4736 |
| Chloroform | Below-median genes | 0.2852 |
| Clotrimazole | All genes | 0.0141 |
| Clotrimazole | Above-median genes | 0.0110 |
| Clotrimazole | Below-median genes | 0.0089 |

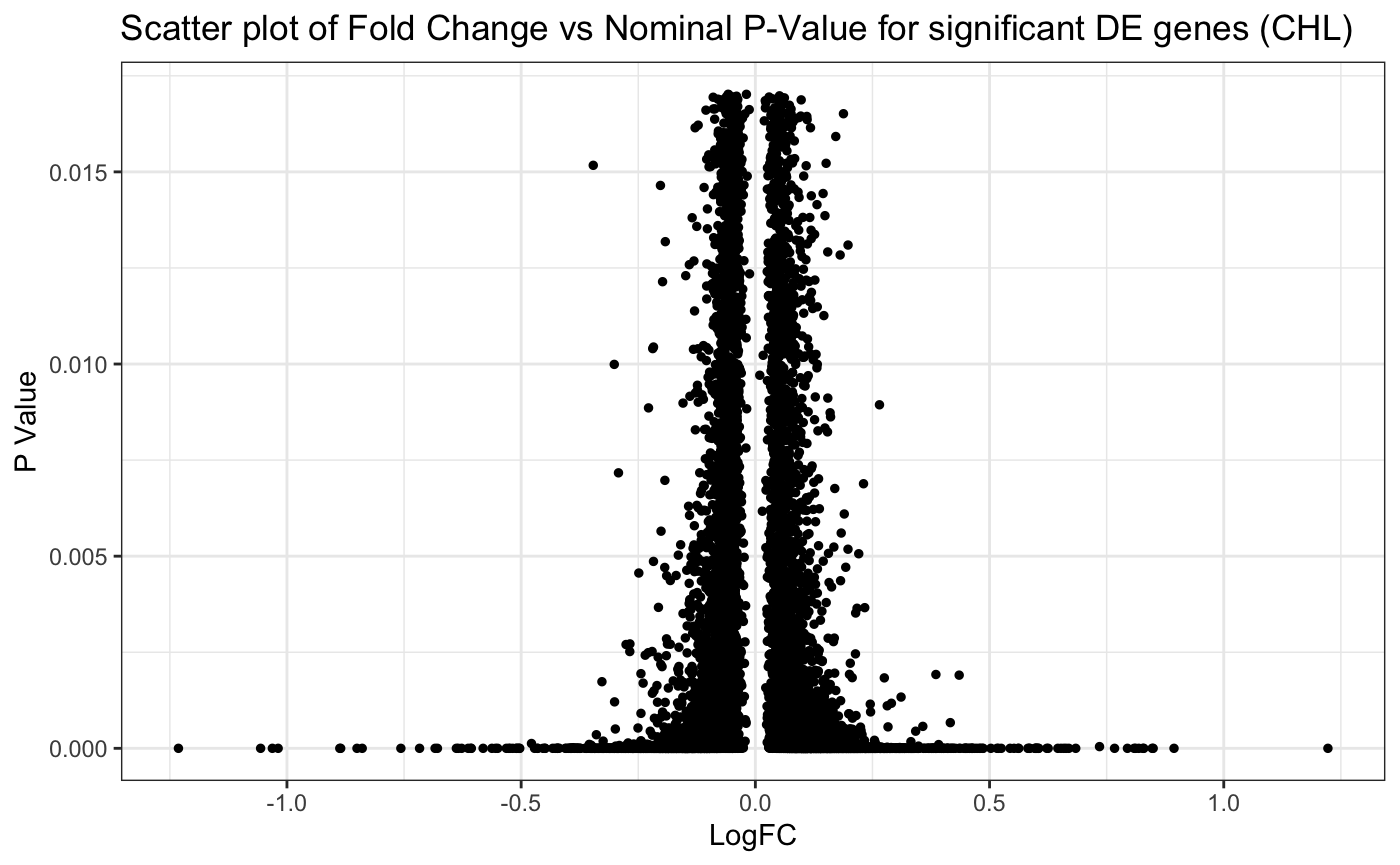
****

****

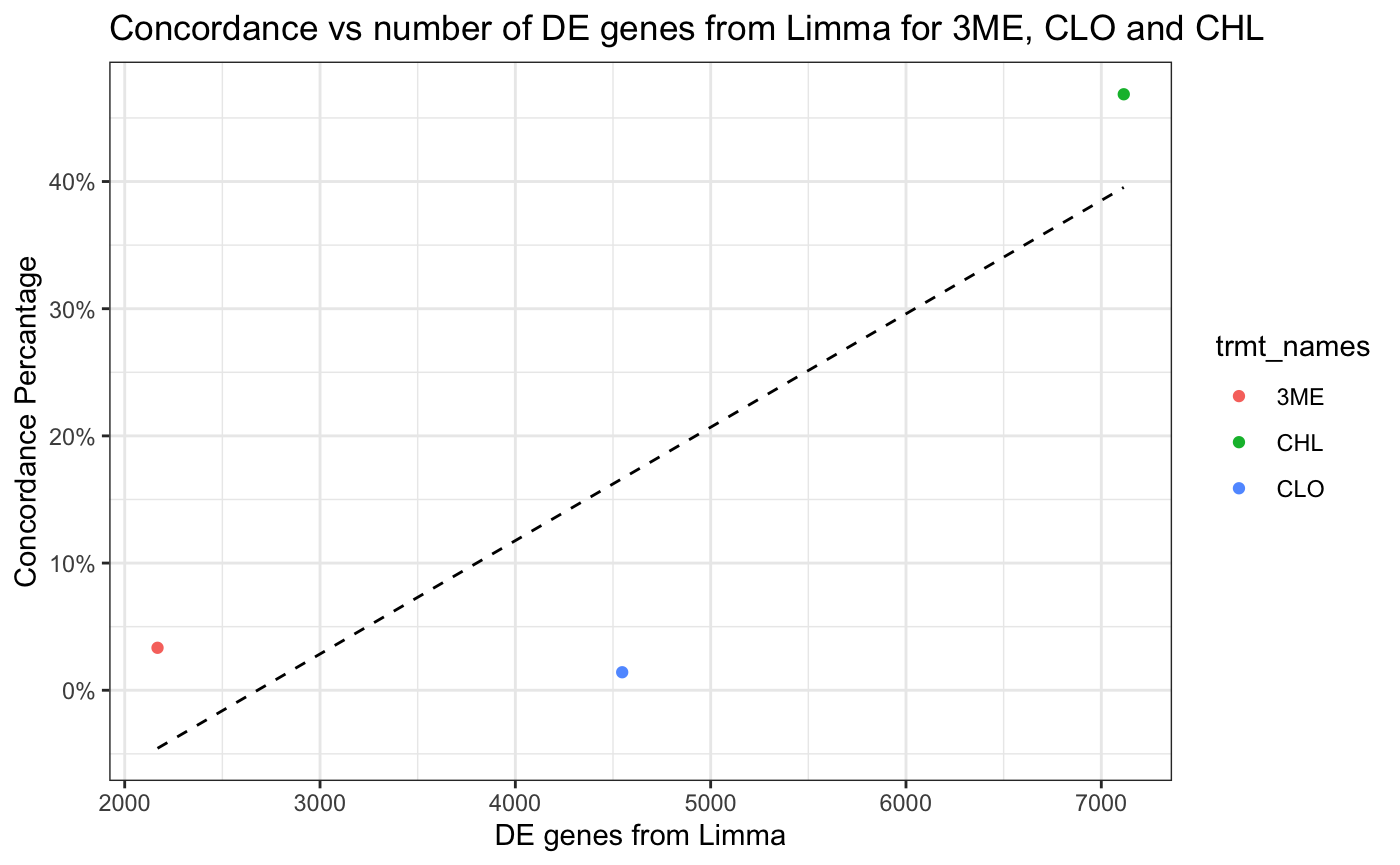
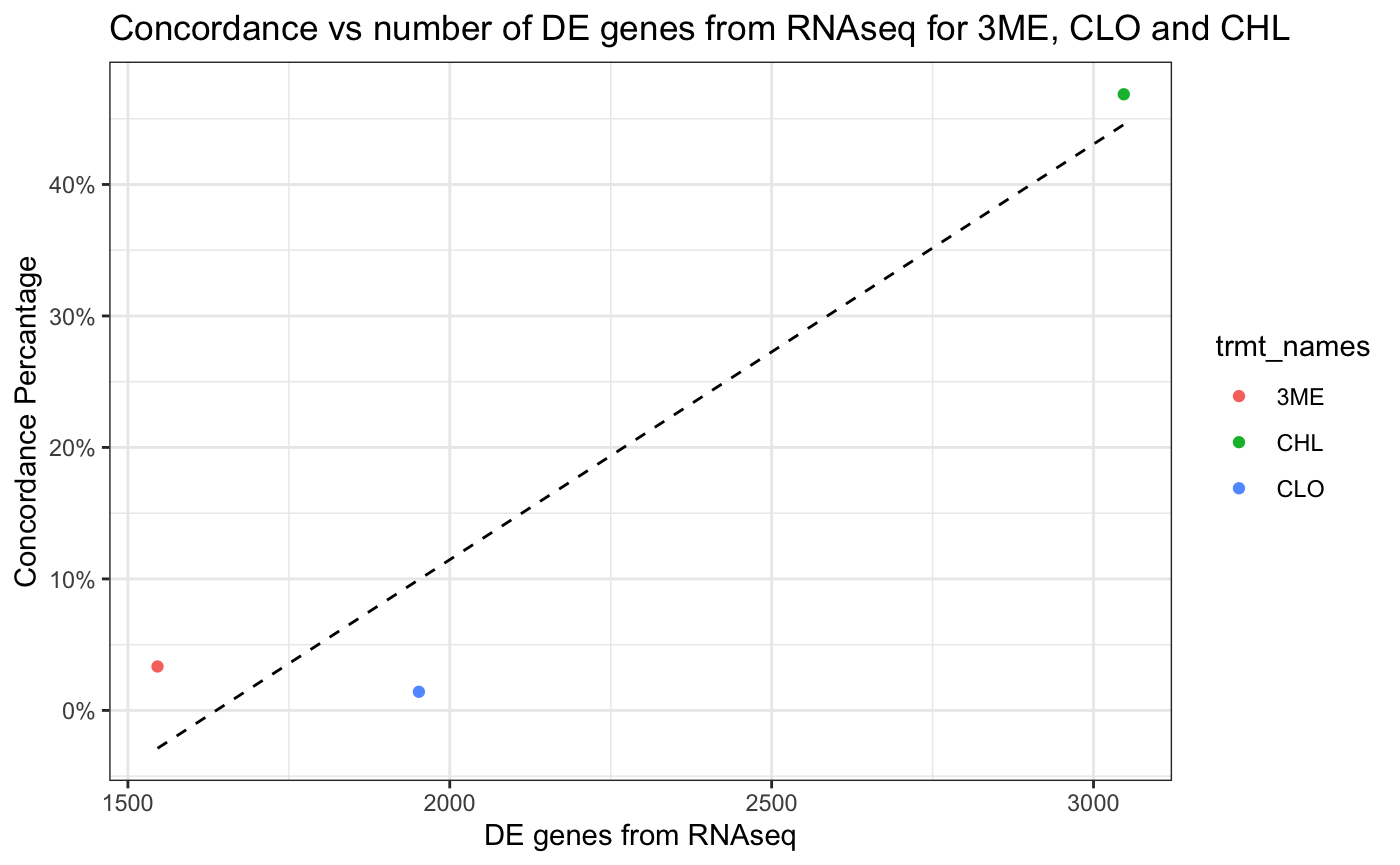
**Figure 6. Histograms of Fold Change Values Across Treatment Groups (Limma).** Histograms of the fold change values for each treatment group show that the majority of fold change values are close to 0 for each treatment group.



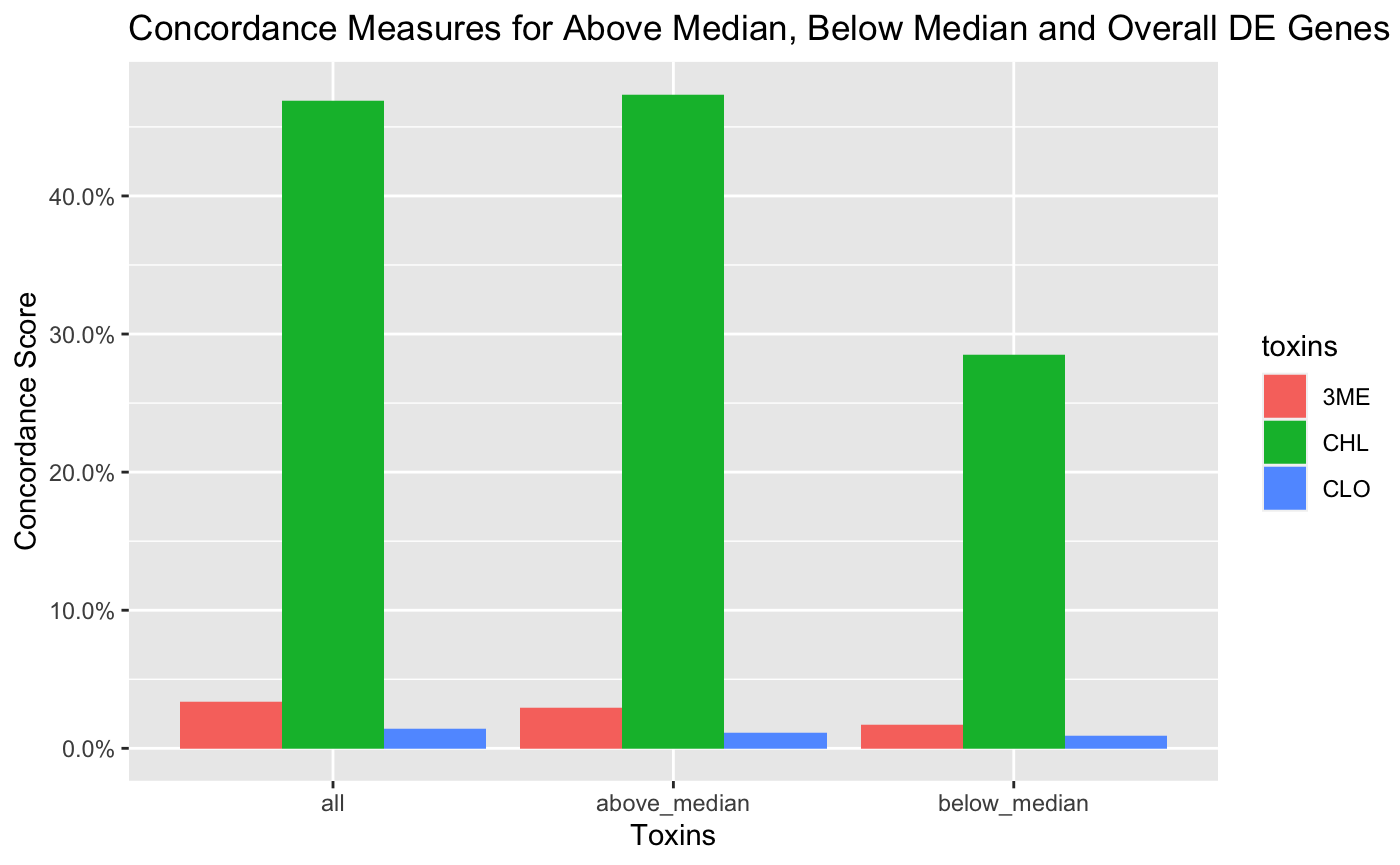




**Figure 7. Scatter plots of fold change vs nominal-p value for significant DE genes (Limma). S**catterplot of the fold change values against nominal p-value for each treatment group show an increase in p-value as fold change approaches 0.



**Figure 8**. **A plot of overall concordance vs number of DE genes from each analysis.** Aside from clotrimazole, there is a clear correlation between the number of DE genes and concordance. This corroborates the findings of Wang et al.



**Figure 9**. **A combined plot of overall, above, and below median genes for each analysis**. Cross platform concordance was consistently higher for above-median groups when compared to below-median groups, regardless of the particular toxin/chemical.

**Discussion:**

This experiment replicated a publication in literature that interrogated the concordance of RNA-seq and microarray data using the DESeq2, limma, and edgeR differential expression analysis methods. For this replicate experiment, the DESeq2 and limma differential expression analysis methods were used with RNA-seq and microarray data, respectively. Three of the treatment groups featured in the original experiment were featured in this experiment. The dependencies of concordance in microarray and RNA-seq data analysis results identified in this replicate analysis were equivalent to those identified in the original experiment. Specifically, concordance between sequencing platforms increased as the number of significantly differentially expressed genes increased across sample treatment groups. This replicated the finding from the original publication that treatment effect size is correlated with cross platform concordance. Additionally, concordance in treatment groups among genes with relatively high expression levels were found to be greater than concordance in treatment groups among genes with relatively low expression levels. This replicated the finding from the original publication that transcript abundance is correlated with cross platform concordance.

These results indicate that comparisons of differential expression analysis results for lowly expressed genes between microarray and RNA-seq data may be confounded by technical variance. Consequently, it would be advisable for researchers to consider the biological conditions of features of interest before evaluating experimental consensus. In the case that a gene is observed to be lowly expressed in a given condition, our results suggested that intra- sequencing platform comparisons of differential expression results would be more accurate than inter- sequencing platform comparisons. Alternatively, inter- sequencing platform comparisons of differential expression results may be considered with greater confidence in the case that a gene is observed to be highly expressed in a given condition. These results also indicate that inter- sequencing platform comparisons of differential expression results may be considered with greater confidence if an experimental condition features many significantly differentially expressed genes. Furthermore, the correlation between concordance and treatment effect size indicated that treatment effect size may overcome the technical variance introduced by transcript abundance in cross platform concordance values. For example, a constructive experiment may be to interrogate the concordance of RNA-seq and microarray data analysis results for highly differentially expressing conditions inflated with lowly expressed genes. If the differential expression analysis results show high concordance despite having many lowly expressed genes, then it may be possible to circumvent inter- sequencing platform comparison limitations due to low expressing genes in the case that the experimental conditions are highly differentially expressed.

Although the main findings of Wang et al. were reproduced to a significant extent, specific values such as the quantity of differentially expressed genes produced by each treatment group and the actual concordance values differed greatly from theirs. This was not a huge problem, considering that the overall results of this analysis still supported their findings. There was a very clear correlation between both treatment effect size and gene expression levels with concordance, despite the actual concordance values and number of differentially expressed genes for each treatment group diverging significantly from theirs. The latter could be attributed to differences in how Wang et al. filtered their differential expression results. Very subtle differences such as using slightly different gene count filter parameters could lead to dramatic changes in the number of DE genes obtained from each treatment group. This difference is not concerning, since the trend still holds that treatment effect size is directly proportional to concordance values across both platforms. Overall, the assertion that chemicals with similar MOAs will induce a similar gene expression response is a very reasonable one. Each mode of action is associated with a well-defined receptor mediated or non-receptor mediated process, and it would make sense that different chemicals that share a similar mode of action would elicit similar gene expression responses. Figure 2 in the Wang et al. analysis clearly shows different chemicals with the same mode of action clustering together in the cross platform concordance vs treatment effect scatterplots. Although only three chemicals were used in this analysis, Each chemical had a specific mode of action, and ultimately a very different concordance value.

**Conclusion:**

The main purpose of the original experiment was to compare microarray and RNA-seq gene expression data from a set of toxicology groups from rat livers. The goal was to determine how concordance between those platforms was affected by various factors such as the number of differentially expressed genes, enriched pathways, treatment effect size, and the biological complexity of the modes of action. The goal of this analysis was the same, though a much smaller subset of the original 27 chemicals was used. This subset consisted of three different tox group 1 samples, which were 3-Methylcholanthrene (AhR MOA), clotrimazole (CAR/PXR MOA), and chloroform (cytotoxic MOA). Although there were differences in the results between this analysis and the original study performed by Wang et. al, the overall findings were very similar. There was a correlation between cross platform concordance and both the overall gene expression levels and treatment effect size. Concordance in the above-median group was consistently higher than concordance in the below-median group, regardless of the method of analysis or MOA/treatment group. With one exception, concordance was significantly higher for MOAs/treatment groups with a higher number of differentially expressed genes.

One challenge that was encountered during the analysis was replicating the exact cross concordance levels for most of the MOA/treatment groups. This did not appear to have a large impact on the efficacy of the analysis, considering that the general trends remained the same, but most of the concordance values (aside from CHL/cytotoxic MOA) were significantly lower than those found in the Wang et al. study. The differences in concordance levels between the original study and the results of this analysis could be attributed to differences in how the two analyses were performed.

An additional challenge presented by this experiment was the collaborative element in the performance of the differential expression analyses. Slight variations in analysis design between the limma and DESeq2 methods may have introduced some undesirable technical variance in differential expression analysis results across sequencing platforms. Specifically, differences in low abundance gene filtering were identified between analysis methods after conducting the experiment. These undesirable differences could have been mitigated by better communication between researchers, which is important to keep in mind for future projects.

**References:**

1. Andrews, S. (2010). FastQC: A Quality Control Tool for high Throughput Sequencing Data [Online]. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
2. Ewels, Philip, et al. “MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report.” *Bioinformatics*, vol. 32, no. 19, Oct. 2016, pp. 3047–48. *DOI.org (Crossref)*, https://doi.org/10.1093/bioinformatics/btw354.
3. Liao, Y., et al. “FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics*, vol. 30, no. 7, Apr. 2014, pp. 923–30. *DOI.org (Crossref)*, https://doi.org/10.1093/bioinformatics/btt656.
4. Love, Michael I., et al. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology*, vol. 15, no. 12, 2014, p. 550. *PubMed*, <https://doi.org/10.1186/s13059-014-0550-8>.
5. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). “Limma powers differential expression analyses for RNA-sequencing and microarray studies.” *Nucleic Acids Research*, 43(7), e57. Doi: [10.1093/nar/gkv007](https://doi.org/10.1093/nar/gkv007).
6. Robinson MD, McCarthy DJ, Smyth GK (2010). “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” Bioinformatics, 26(1), 139-140. Doi: 10.1093/bioinformatics/btp616.
7. Wickham, Hadley. *Ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. 2016, Springer International Publishing : Imprint: Springer, 2016. *Library of Congress ISBN*, https://doi.org/10.1007/978-3-319-24277-4.