**Project 3 Writeup**
**Team Frazzled**

## Introduction

In their 2014 paper, “A comprehensive study design reveals treatment- and transcript abundance–dependent concordance between RNA-seq and microarray data,” Wang el al sought to understand the difference between RNA-seq and microarray platforms in terms of recognizing differentially expressed genes (DEGs). In this study, the authors generated RNA-seq and microarray data from liver samples of rats that were exposed to perturbation by 27 chemicals representing Multiple models of action (MOA). In this study, they treated three male rats per chemical and RNA were collected from liver tissue for further analysis. They investigated the difference between RNA-seq and microarray across the chemical and analysis DEGs on gene expression.

## Data

Data was generated as follows: Rats were exposed to one of 27 chemicals and RNA was isolated from their livers. Next, these samples were analyzed by Affymetrix microarrays and Illumina HiSeq for RNA-seq. Paired-end RNA-seq data was deposited at NCBI with accession SRP024314. For our analysis, samples from toxgroup 3 were chosen, representing Leflunomide, Fluconazole, and the relevant controls. Sample IDs for the RNA-seq data were extracted from the toxgroup 3 rna data table, representing the experimental and control samples, and are included in Supplementary Table 1. FastQC (version 0.11.7) was run on files with these ids. Next, an alignment to the provided genome was performed using STAR (version 2.6.0c) to generate BAM files for the nine experimental samples. MultiQC (version 1.10.1) was performed on fastq files and STAR alignments.

FastQC results showed an average of 15 million reads per sample (approximately half of those unique). Samples associated with SRR1178061 and SRR1178063 had a total number of reads that was 3 and 4 times the average, respectively. All sequences maintained a Phred score of approximately 30 at least through 50 bp. FastQC reports flagged 22 of these sequences for poor quality, but because the low Phred score began late in the read and STAR aligner takes into account the Phred score, reads were not trimmed before alignment. Other flagged FastQC steps, like overrepresented sequences and sequence duplications were ignored because RNA-Seq by nature yields non-homogenous sequencing results (tied to expression differences, and thereby, differences in RNA template prevalence in the samples). STAR alignment QC showed that for each sample, > 80% of all reads were mapped to the genome, and 11-16 million reads were aligned per sample. These statistics reflect high quality RNA-Seq reads and a good starting material for analysis. See Supplementary Tables for selected data from MultiQC output.

**Methods**

All programs and analysis were performed using BU's shared computing cluster and RStudio (R version 4.0.2). Packages utilized for analysis included; featureCounts (version 1.6.2), MultiQC (version 1.6), DESeq2 (version 1.30.1) and apeglm (version 1.12.0). Following STAR alignment of the nine samples we utilized the featureCounts package on the previously produced bam files. In addition to the bam files, the featureCounts package was also given the respective annotated gtf file. As a result of this package we produced counts for each sample and proceeded to perform MultiQC. MultiQC reports ensured featureCounts successfully ran count assignments for each sample (Supplemental Image 3).

Following MultiQC all nine samples along with the metadata and controls were loaded into R for further analysis. Each mode of action pathway samples and controls were merged together; AhR samples SRR1178008, SRR1178009, SRR1178010 with their respective controls SRR1178050, SRR1178061, SRR1178063 followed by, CAR/PXR samples SRR1178014, SRR1178021, SRR1178047 with their respective controls SRR1178050, SRR1178061, SRR1178063 and lastly DNA Damage samples SRR1177981, SRR1177982, SRR1177983 with their respective controls SRR1178004, SRR1178006, SRR1178013. Once merged, we removed any zeros and 'NA's' from each of the sample counts data frames and generated boxplots(Figure 1).

Subsequently, we performed RNA-seq analysis using the DESeq2 package. In order to perform this analysis we needed to create DESeq2 objects for each of the three different modes of actions; AhR, CAR/PXR, and DNA Damage. With each object we needed to provide their respective information; count matrix, sample information, and mode of action. With the DESeq2 package, we had to recognize the controls or R will do it on its own alphabetically. In this case the controls were the modes of action; AhR, CAR/PXR, and DNA Damage. DESeq2 was run for each mode of action calculating the mean, log2 fold changes, standard error of the log2 fold change, p-values, and adjusted p-values. In addition we extracted normalized counts from the DESeq2 object. From these calculations, we identified differentially expressed genes with p-values adjusted to (padj) < 0.05, the top ten differentially expressed genes for each mode of action (Table 1). Furthermore, we create histograms of fold change values (Figure 2) and scatter plots of fold change versus nominal p-values for each mode of action (Figure 3).

After performing and analyzing our RNA-Seq results, we performed a microarray analysis using the Limma bioconductor package (Ritchie). In RStudio the limma package implemented analytical methodology to determine the differentially expressed genes from microarray analyses. Each of our samples was run through limma compared to the controls from the RMA matrix available to us.

Using the differentially expressed genes collected from the RNA-Seq analysis and microarray analysis, we were able to calculate the concordance. Following the paper's methodology of concordance calculations, each determined a concordance measure for each of the analyses.

Certain aspects of the concordance calculation methodology was vague, as far as defining variables. The paper's main formula for concordance was:

$2 * \text{intersect}(DEG_{microarray}, DEG_{RNA\text{-}Seq}) / DEG_{microarray} + DEG_{RNA\text{-}Seq} = C.$

In order to use this main formula, we had to determine the various sub-variables prior. Our calculations used the intersection between the DEGs (n_x), the dimensions of each set of genes (n_1 and n_2), and the total sum of the DEGs (total). We also had to use a variable, N, as the number of items in the whole set of genes. This definition of N could vary in different ways. For our calculation we used N as the dimension of the affymetrix map we used a reference to match the DEGs. The variables used in our calculations are listed below. Our final main formula used was:

$2 * n\_x / total = C.$

Note:

$n\_x = abs((n\_0*N-n\_1*n\_2)/(n\_0+N-n\_1-n\_2))$

$n\_0 = sum(intersection)$

$total = sum(matches)$

$n\_1 = dim(microarray\ set)$

$n\_2 = dim(RNA\text{-}Seq\ set)$

$N = dim(affymap)$

To break down our results further, we separated each analysis between the above and below median values and provided concordance computations for each (Results Table 4).

Lastly, We performed enrichment analyses using one of the online bioinformatics gene set enrichment methods DAVID (Huang et al., 2009). The list of genes similar to Wang et al.'s MOAs were filtered utilizing the threshold padj <0.05. Furthermore, the DE genes results were collected for the enrichment analysis in David tool.

**Results**

Through the use of featureCounts described above we generated boxplots which display the distribution of read counts for each sample (Figure 1). The figure demonstrates samples have a similar distribution read count. Following featureCounts, differential gene expression analysis shows there are 1,431 differentially expressed genes within AhR mode of action at adjusted p-values of < 0.05. In addition, CAR/PXR shows there are 3,519 differentially expressed genes at adjusted p-values of < 0.05. The mode of action DNA Damage shows the least amount of differentially expressed genes, 335 at adjusted p-values of < 0.05. Furthermore, we generated histograms for each mode of action samples displaying fold change values of differentially expressed genes (Figures 2A, 2B, 2C). In addition, scatter plots demonstrate the fold changes versus the nominal p-values of significantly differentially expressed genes for each mode of action condition (Figures 3A, 3B, 3C). The top ten differentially expressed genes were also identified for each mode of action (Table 1A, 1B, 1C).

Limma analysis allowed us to collect the differentially expressed genes from the microarray analysis. Each chemical sample had a different amount of differentially expressed genes according to the adjusted p.value < 0.05, as reported in Table 2. Interestingly, the ifosfamide chemical had no differentially expressed genes at adjusted p.value < 0.05. Though, the flucozonale chemical showed 1997 DEGs and leflunomide showed 466 DEGs (Table 2). In Table X, we report the top 10 DEGs for each chemical sample. Noting that the table shows ifosfamide DEGs starting at a p adjusted value of 0.1092657 (Table 3). Using the limma results we were able to create visual plots highlighting the data. The fold change values of the significantly expressed genes were used to create histograms (Figure 4A, 4B, 4C). Scatter plots were also generated using the log fold change values versus the p values (Figure 5A, 5B, 5C).

Following both limma and RNA-Seq analyses, concordance values were calculated and reported in Table 4. Each chemical sample concordance measure was determined using the overall, above, and below median values (Table 4). Notably, the ifosfamide concordance values were extremely low. This may have been due to the trend of low DEGs in the previous microarray analysis. To visualize our results, a scatter plot was generated for each of the overall concordance values (Figure 6A and 6B). Furthermore, a combined bar plot shows the values for all the concordance measures graphically (Figure 7).

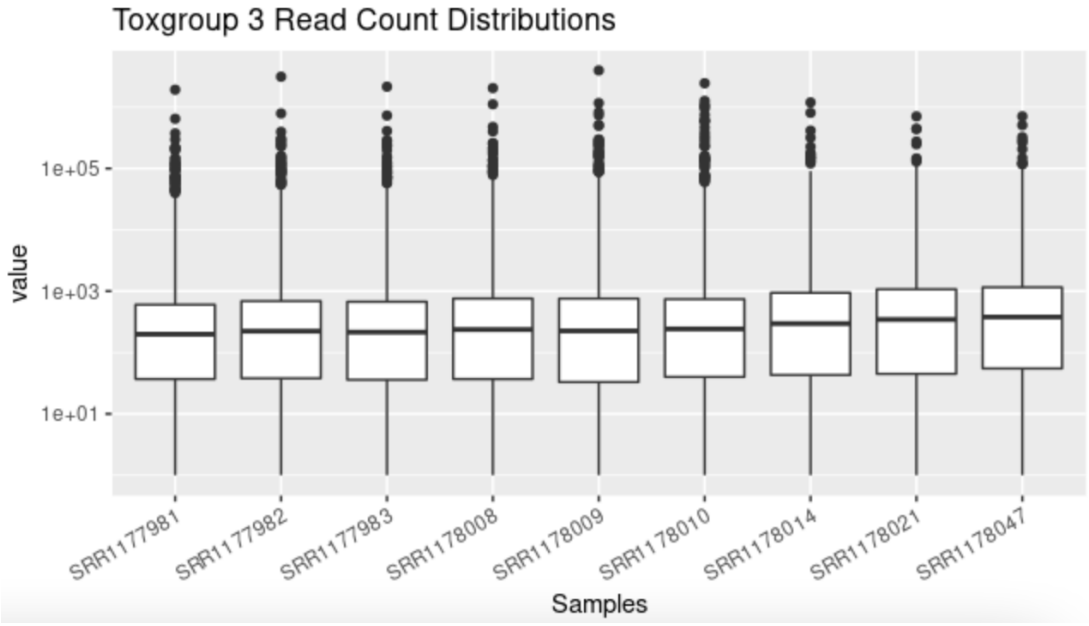


Figure 1: featureCounts read count distributions for each toxgroup 3 samples

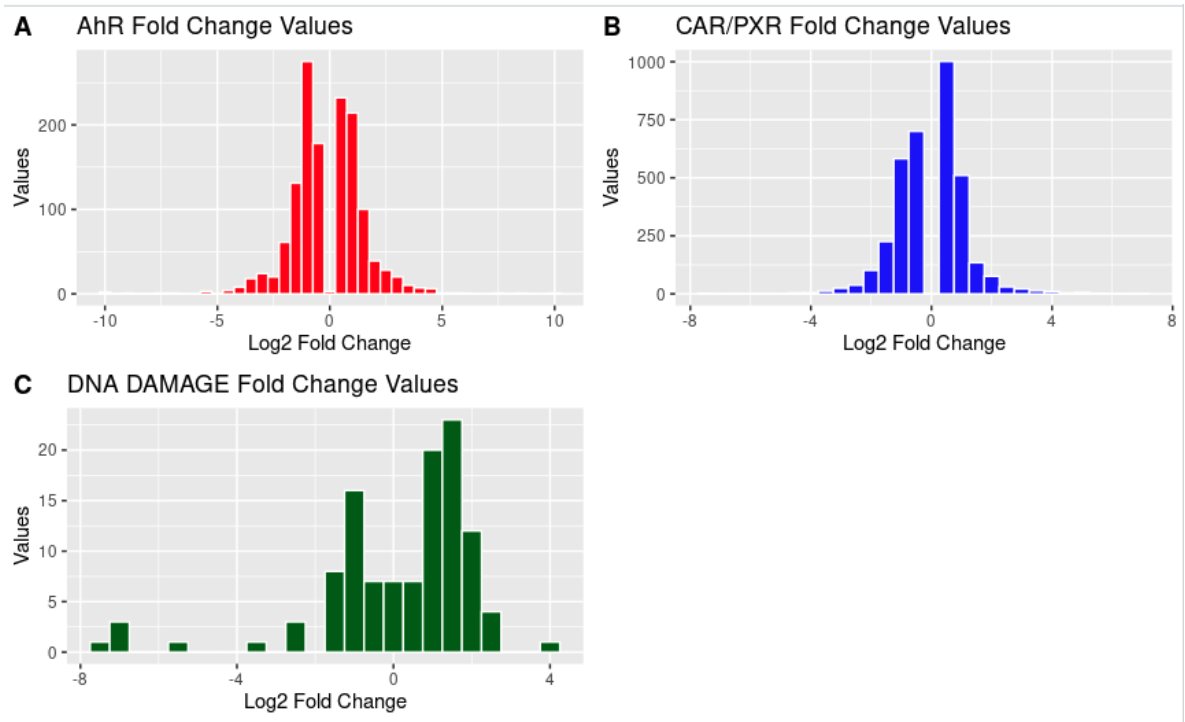



Figure 2: A) AhR log2 fold change values of differentially expressed genes. B)CAR/PXR log2 fold change values of differentially expressed genes. C) DNA Damage log2 fold change values of differentially expressed genes

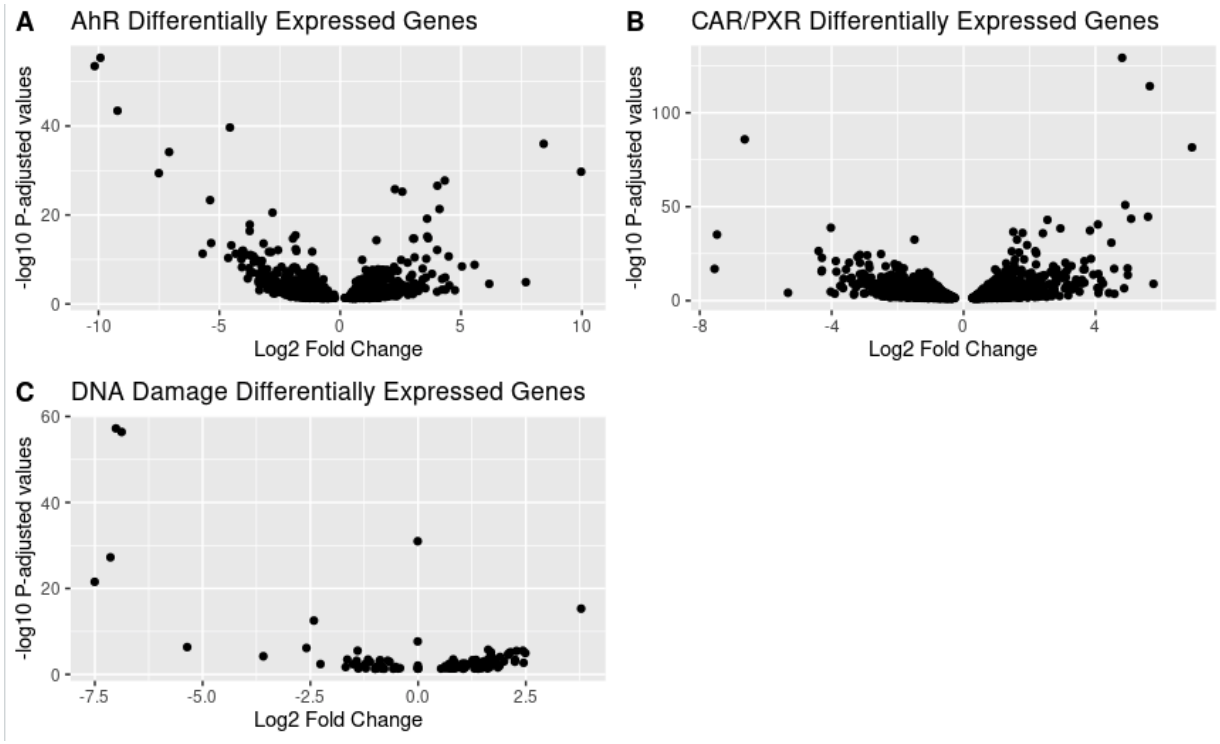


Figure 3: A) AhR differentially expressed genes log2 fold change vs nominal p values. B) CAR/PXR differentially expressed genes log2 fold change vs nominal p values. C) DNA Damage differentially expressed genes log2 fold change vs nominal p values

| | baseMean | log2FoldChange | lfcSE | pvalue | padj |
|---|---|---|---|---|---|
| **NM_013096** | 10715.0347 | -9.919301 | 0.6128536 | 4.078732e-60 | 4.513525e-56 |
| **NM_033234** | 8518.1577 | -10.157233 | 0.6400207 | 6.024004e-58 | 3.333081e-54 |
| **NM_001007722** | 4144.7321 | -9.212939 | 0.6431716 | 9.955665e-48 | 3.672313e-44 |
| **NM_001257095** | 6962.6328 | -4.561071 | 0.3336887 | 7.786360e-44 | 2.154096e-40 |
| **NM_001008880** | 386.9592 | 8.417538 | 0.6609357 | 4.512408e-40 | 9.986862e-37 |
| **NM_001130558** | 683.0783 | -7.076736 | 0.5551097 | 3.832728e-38 | 7.068828e-35 |
| **NM_012540** | 67201.2289 | 9.967511 | 0.8411282 | 1.200217e-33 | 1.897371e-30 |
| **NM_198776** | 1813.8355 | -7.502313 | 0.6382769 | 2.886615e-33 | 3.992910e-30 |
| **NM_012541** | 207867.1448 | 4.325954 | 0.3785058 | 1.471770e-31 | 1.809623e-28 |
| **NM_130407** | 2957.2763 | 4.021023 | 0.3589651 | 2.392713e-30 | 2.647776e-27 |

Table 1A: AhR top ten differentially expressed genes arranged by pvalue (row names are the gene ids).

| | baseMean | log2FoldChange | lfcSE | pvalue | padj |
|---|---|---|---|---|---|
| **NM_053288** | 156253.0954 | 4.808324 | 0.1951989 | 4.573665e-134 | 5.245994e-130 |
| **NM_001108693** | 480.4558 | 5.648195 | 0.2451090 | 1.194013e-118 | 6.847666e-115 |
| **NM_001130558** | 839.9042 | -6.635127 | 0.3315353 | 3.721804e-90 | 1.422970e-86 |
| **NM_001134844** | 52723.4566 | 6.927276 | 0.3556080 | 9.194549e-86 | 2.636537e-82 |
| **NM_080581** | 6237.6721 | 4.902755 | 0.3173311 | 5.604229e-55 | 1.285610e-51 |
| **NM_013033** | 848.9781 | 5.592743 | 0.3898193 | 1.207400e-48 | 2.308146e-45 |
| **NM_053699** | 367.4442 | 5.081106 | 0.3604318 | 1.667319e-47 | 2.732022e-44 |
| **NM_024127** | 1688.0383 | 2.543686 | 0.1791996 | 7.465167e-47 | 1.070318e-43 |
| **NM_031048** | 5926.6843 | 4.073899 | 0.2960558 | 2.350319e-44 | 2.995351e-41 |
| **NM_013098** | 15250.0645 | -4.023552 | 0.2989010 | 1.400275e-42 | 1.606116e-39 |

Table 1B: CAR/PXR top ten differentially expressed genes arranged by pvalue (row names are the gene ids).

| | baseMean | log2FoldChange | lfcSE | pvalue | padj |
|---|---|---|---|---|---|
| NM_033234 | 3408.95474 | -7.010396582 | 0.42656295 | 5.518027e-62 | 6.018512e-58 |
| NM_001007722 | 1699.42794 | -6.877276039 | 0.42228207 | 7.365893e-61 | 4.016990e-57 |
| NM_198776 | 796.41894 | -0.009105007 | 0.02571932 | 2.834518e-35 | 1.030536e-31 |
| NM_001111269 | 277.18992 | -7.137415973 | 0.61852725 | 2.105490e-31 | 5.741144e-28 |
| NM_013096 | 4753.93897 | -7.503075605 | 0.73846248 | 1.336718e-25 | 2.915918e-22 |
| NM_012623 | 575.23124 | 3.782740877 | 0.43878448 | 2.747854e-19 | 4.995141e-16 |
| NM_199113 | 81.43245 | -2.417383710 | 0.30724166 | 1.905853e-16 | 2.969592e-13 |
| NM_001107084 | 632.81388 | -0.008636215 | 0.02539185 | 1.536084e-11 | 2.094258e-08 |
| NM_001113223 | 37.47479 | -5.357098387 | 0.89028197 | 3.619247e-10 | 4.386125e-07 |
| NM_001013057 | 793.60405 | -2.590080552 | 0.45821472 | 6.444442e-10 | 7.028953e-07 |

Table 1C: DNA Damage top ten differentially expressed genes arranged by pvalue (row names are the gene ids).

| **Chemical** | **Number of DEGs** |
|---|---|
| Leflunomide | 466 genes |
| Ifosfamide | 0 genes |
| Fluconazole | 1997 genes |

Table 2: Report of DE genes at p-adjust < 0.05. (Determined by Limma Analysis)

| Leflunomide: | | Ifosfamide: | | Fluconazole: | |
|---|---|---|---|---|---|
| Top 10 DE Genes | Adj. P. Value | Top 10 DE Genes | Adj. P. Value | Top 10 DE Genes | Adj. P. Value |
| 1370269_at | 3.967445e-11 | 1379397_at | 0.1092657 | 1368731_at | 2.369280e-07 |
| 1387243_at | 4.804763e-08 | 1374475_at | 0.3006659 | 1377014_at | 2.454702e-06 |
| 1372600_at | 3.976621e-06 | 1368273_at | 0.3006659 | 1371076_at | 3.278103e-06 |
| 1392946_at | 2.439320e-05 | 1371266_at | 0.3006659 | 1390255_at | 3.278103e-06 |
| 1388611_at | 1.117236e-04 | 1373217_at | 0.3006659 | 1391570_at | 3.278103e-06 |
| 1370244_at | 1.400901e-04 | 1368718_at | 0.3006659 | 1394022_at | 5.892779e-06 |
| 1373810_at | 1.400901e-04 | 1376481_at | 0.3006659 | 1380336_at | 9.637781e-06 |

| | | | | | |
|---|---|---|---|---|---|
| 1398598_at | 2.216871e-04 | 1378596_at | 0.3355401 | 1372136_at | 9.637781e-06 |
| 1376827_at | 2.216871e-04 | 1377029_at | 0.3355401 | 1398597_at | 9.637781e-06 |
| 1373814_at | 2.290109e-04 | 1383137_at | 0.3355401 | 1377192_a_at | 1.335853e-05 |

Table 3: Report of top 10 DEGs from Limma Results

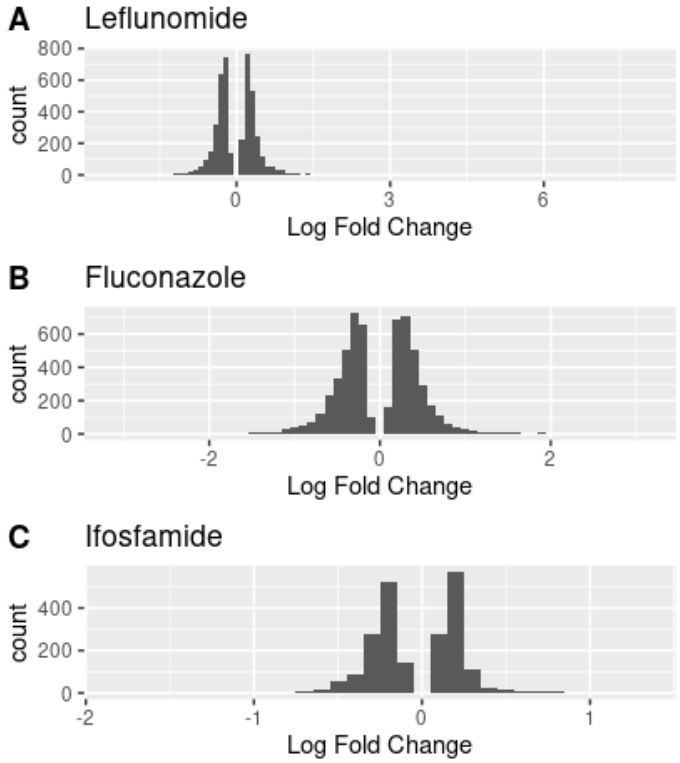



Figure 4: Histogram of fold change values for each sample chemical (A, B, C) (determined by limma analyses)

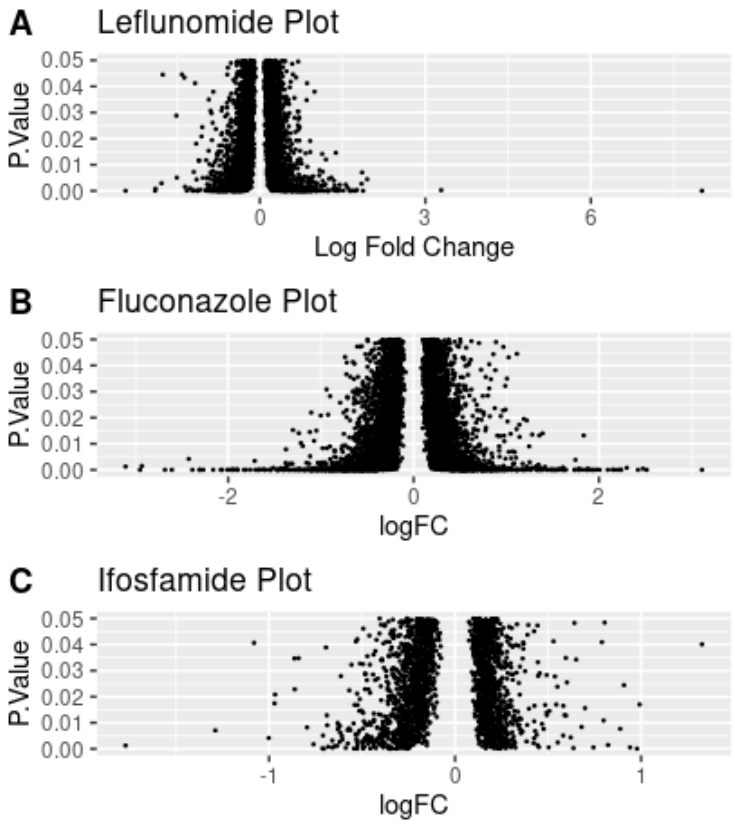



Figure 5: Scatter plots of fold changes vs. nominal p values from each sample chemical (A, B, C) (determined by limma analyses)

| Chemicals: | Overall Concordance: | Above Median Concordance: | Below Median Concordance: |
|---|---|---|---|
| Fluconazole | 0.1213545 | 0.1260429 | 0.009909024 |
| Ifosfamide | 5.087758e-05 | 0 | 0 |
| Leflunomide | 0.04488275 | 0.03879421 | 0.008110292 |

Table 4: The Overall, Above Median, and Below Median Concordance Values

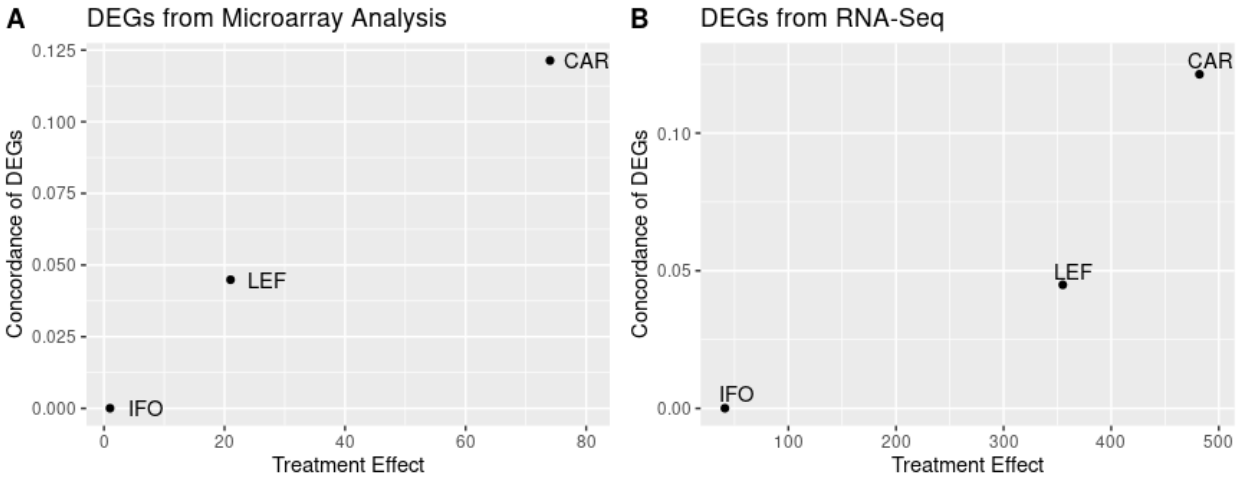



Figure 6: Scatter plots of overall concordance vs. number of DEGs for each analysis. 6A: The DEGs from microarray analysis. 6B: the DEGs from RNA-Seq analysis.

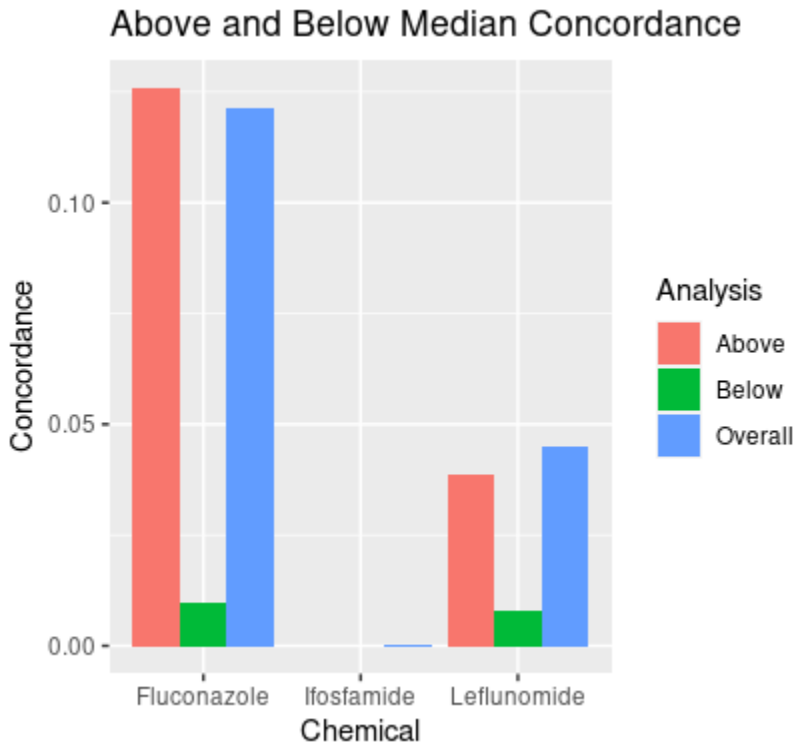



Figure 7: Bar plot of all combined concordance measures

**Discussion**

Using a determined set of normalized counts in our analysis, we were able to pursue a further analysis to make biological interpretation and comparison. The Wang et al.'s paper created a foundation to understand the differences between RNA-seq and microarray platforms. We were

able to identify some similarities in our results and these similarities were critical for indication of some biological conclusion.

Our measured concordance values did not follow similar high percentage trends for the chemicals as the Wang et al.'s paper did. This is likely due to the differences in methods to calculate concordance between our calculations and the papers. Though, the scatter plots generated (Results Figure 6) were compared to the paper's Figure 2A, and showed a similar trend between the three chemicals. Overall, the concordance measures were necessary to reproduce the paper's results, though our results were not the same from the original paper.

To compare the enrichment pathways of DE genes from our RNA-seq and microarray analysis, we compared our result from DAVID (Huang et al., 2009) results to (Wang et al., 2014) paper's analysis of MOA chemical groups that are shared by both RNA-seq and microarray platforms that found in the online supplementary Table 4. In this comparison, our groups used three MOA chemical groups; CAR/PXR, AhR, and DNA damage and Wang et al listed seven MOA chemical groups enriched pathways. Using the DAVID gene set enrichment method we gathered our result Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways terms. Based on the result, we noticed various similarities between the Wang et al. paper's findings and ours. For instance, the paper and our results on DNA damage indicates significant similarities on enriched pathways such as cell cycle and metabolism of xenobiotics. On the other hand, we were only able to find very few of similar pathways in CAR/PXR and AhR. Overall, we produced some similarities in the KEGG pathways between our result and the papers to the same biological interpretation. However, some of the terms in our results were not found in the paper and this might be due to DE gene expression method.

Lastly, we created a clustered heatmap of count using normalized expression matrix. The normalized counts in CAR/PXR, AhR, and DNA damage samples as detected by RNA-seq differential expressions were visually contrasted using a clustered heatmap. First, we clustered all the MOAs normalized expression samples together into one file. Hence, We observed that our clustering showed 60% accuracy, it clustered 3/4 samples correctly in each of the cases AhR, CAR/PXR, and DNA damage. We observed that our final result had only 12 samples instead of 15 samples which can be attributed to in processing of normal count together.

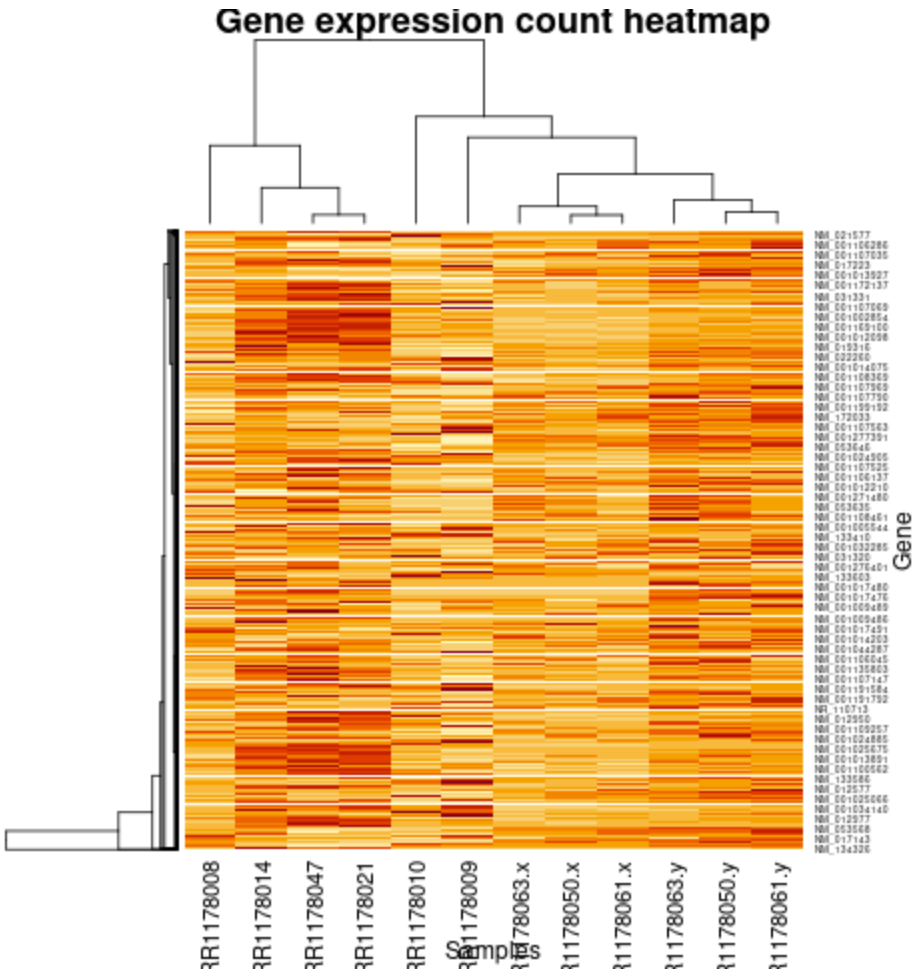



Figure 8: Clustered heatmap generated from lists of genes corresponding to MOAs

**Conclusion**

Overall, our findings did not fully reproduce the paper's findings. Though, we were able to draw some similar biological conclusions. Through our efforts to reproduce the Wang et al. paper we did experience a few problems.

An issue we encountered was the vague method description from the original paper. The methods we were aiming to follow, to compute concordance, did not give extremely clear information regarding their calculations. To solve this issue, we had to use assumptions to define the variables in our calculations. These variable definitions were noted and explained in the Methods section of our paper. This issue is likely the reason our concordance values were unable to match the results of the original paper.

Another challenge encountered was finding 0 differentially expressed genes according to adjusted p value for the ifosfamide sample. After collaborating with other groups, we determined

that others were also having the same findings. To move forward with the analyses, we used the top DEGs from our results but did not use adjusted p value < 0.05 as a parameter. This may not have been the best solution to this challenge, but we were able to make the rest of our analyses work.

**References**


Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635

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

Liao Y, Smyth GK, Shi W (2019). "The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads." Nucleic Acids Research, 47, e47. doi: 10.1093/nar/gkz114.

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.

Philip Ewels, Måns Magnusson, Sverker Lundin, Max Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report, Bioinformatics, Volume 32, Issue 19, 1 October 2016, Pages 3047–3048, https://doi.org/10.1093/bioinformatics/btw354

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), e47. doi: 10.1093/nar/gkv007.

Zhu, A., Ibrahim, J.G., Love, M.I. (2018) Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences
Bioinformatics. https://doi.org/10.1093/bioinformatics/bty895

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


**Supplementary Data:**

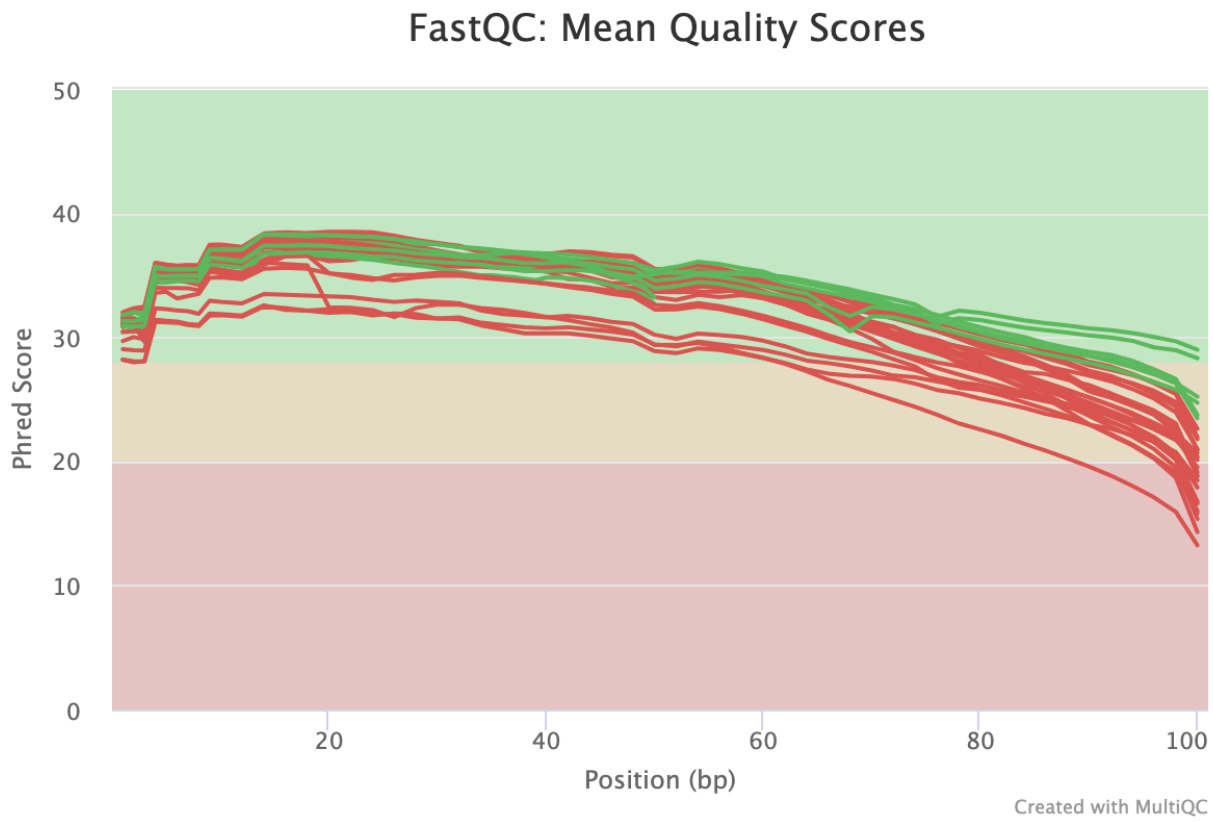
| Sample ID | mode_of_action | chemical | vehicle | route |
|---|---|---|---|---|
| SRR1178008 | AhR | LEFLUNOMIDE | CORN_OIL_100_% | ORAL_GAVAGE |
| SRR1178009 | AhR | LEFLUNOMIDE | CORN_OIL_100_% | ORAL_GAVAGE |
| SRR1178010 | AhR | LEFLUNOMIDE | CORN_OIL_100_% | ORAL_GAVAGE |
| SRR1178014 | CAR/PXR | FLUCONAZOLE | CORN_OIL_100_% | ORAL_GAVAGE |
| SRR1178021 | CAR/PXR | FLUCONAZOLE | CORN_OIL_100_% | ORAL_GAVAGE |
| SRR1178047 | CAR/PXR | FLUCONAZOLE | CORN_OIL_100_% | ORAL_GAVAGE |
| SRR1177981 | DNA_Damage | IFOSFAMIDE | SALINE_100_% | ORAL_GAVAGE |
| SRR1177982 | DNA_Damage | IFOSFAMIDE | SALINE_100_% | ORAL_GAVAGE |
| SRR1177983 | DNA_Damage | IFOSFAMIDE | SALINE_100_% | ORAL_GAVAGE |
| SRR1178050 | Control | Vehicle | CORN_OIL_100_% | ORAL_GAVAGE |
| SRR1178061 | Control | Vehicle | CORN_OIL_100_% | ORAL_GAVAGE |
| SRR1178063 | Control | Vehicle | CORN_OIL_100_% | ORAL_GAVAGE |
| SRR1178004 | Control | Vehicle | SALINE_100_% | INTRAPERITONEAL |
| SRR1178006 | Control | Vehicle | SALINE_100_% | INTRAPERITONEAL |
| SRR1178013 | Control | Vehicle | SALINE_100_% | INTRAPERITONEAL |

Supplementary Table 1: Toxgroup 3 sample metadata

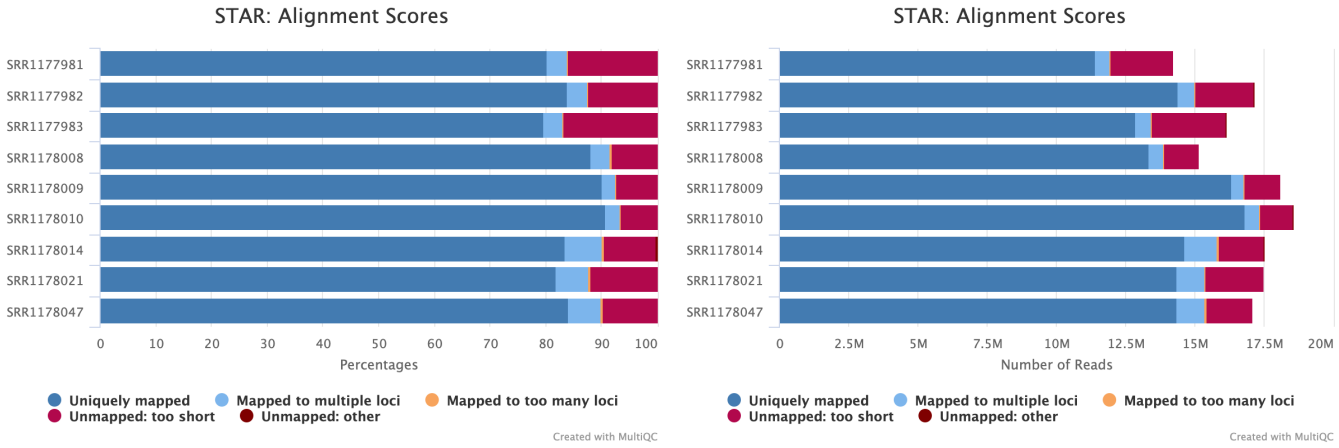
| Sample Name | % Dups | % GC | Length | M Seqs |
|---|---|---|---|---|
| **SRR1177981_1** | 56.1% | 48% | 101 bp | 14.2 |
| **SRR1177981_2** | 48.4% | 48% | 101 bp | 14.2 |
| **SRR1177982_1** | 58.5% | 48% | 101 bp | 17.2 |
| **SRR1177982_2** | 54.3% | 48% | 101 bp | 17.2 |
| **SRR1177983_1** | 57.6% | 48% | 101 bp | 16.2 |
| **SRR1177983_2** | 50.0% | 48% | 101 bp | 16.2 |
| **SRR1178004_1** | 61.1% | 48% | 101 bp | 19.6 |
| **SRR1178004_2** | 60.5% | 48% | 101 bp | 19.6 |
| **SRR1178006_1** | 59.2% | 49% | 101 bp | 21.5 |
| **SRR1178006_2** | 58.1% | 49% | 101 bp | 21.5 |
| **SRR1178008_1** | 56.2% | 49% | 101 bp | 15.2 |
| **SRR1178008_2** | 54.4% | 49% | 101 bp | 15.2 |
| **SRR1178009_1** | 62.2% | 49% | 101 bp | 18.1 |
| **SRR1178009_2** | 60.1% | 49% | 101 bp | 18.1 |
| **SRR1178010_1** | 62.5% | 49% | 101 bp | 18.6 |
| **SRR1178010_2** | 61.4% | 49% | 101 bp | 18.6 |
| **SRR1178013_1** | 58.4% | 49% | 101 bp | 16.1 |

| | | | | |
|---|---|---|---|---|
| **SRR1178013_2** | 57.0% | 49% | 101 bp | 16.1 |
| **SRR1178014_1** | 53.9% | 49% | 50 bp | 17.5 |
| **SRR1178014_2** | 51.9% | 49% | 50 bp | 17.5 |
| **SRR1178021_1** | 48.7% | 49% | 100 bp | 17.5 |
| **SRR1178021_2** | 46.4% | 49% | 100 bp | 17.5 |
| **SRR1178047_1** | 48.5% | 49% | 100 bp | 17.1 |
| **SRR1178047_2** | 47.3% | 49% | 100 bp | 17.1 |
| **SRR1178050_1** | 54.6% | 48% | 100 bp | 16.1 |
| **SRR1178050_2** | 53.1% | 49% | 100 bp | 16.1 |
| **SRR1178061_1** | 69.1% | 49% | 101 bp | 63.3 |
| **SRR1178061_2** | 68.6% | 49% | 101 bp | 63.3 |
| **SRR1178063_1** | 64.8% | 49% | 101 bp | 44.5 |
| **SRR1178063_2** | 62.3% | 49% | 101 bp | 44.5 |

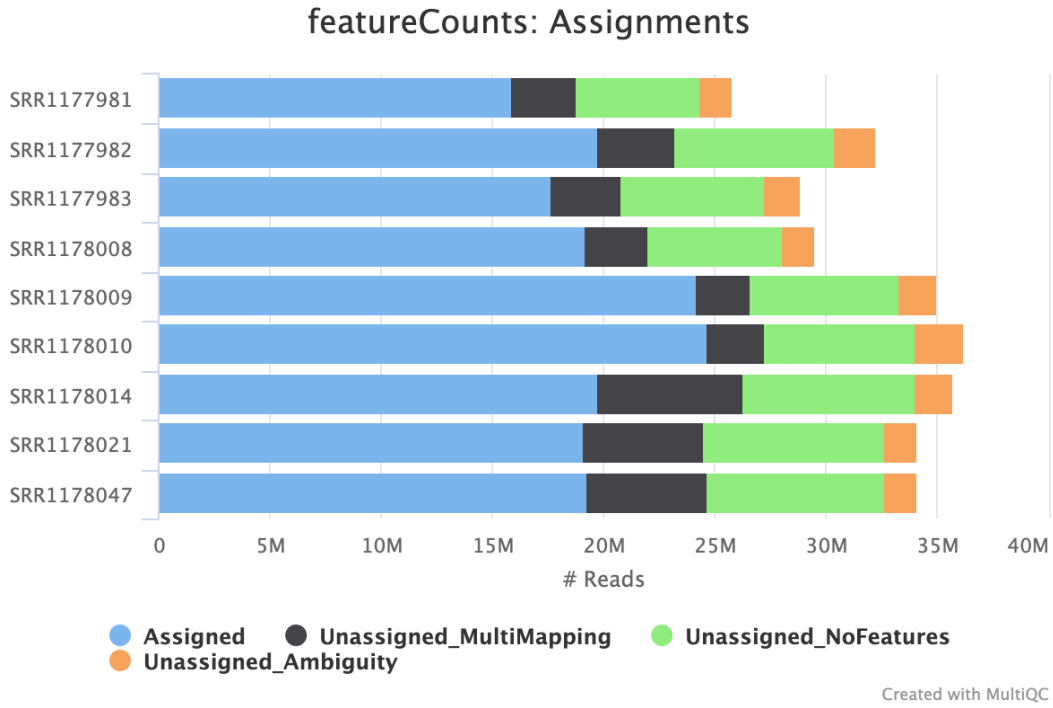
Supplementary Table 2: FastQC statistics for RNA-Seq files




Supplementary Figure 1: Mean Phred Score by sequence position




Supplementary Figure 2: STAR alignment scores for each of nine sample alignments. (left) by percentage of reads, (right) by number of reads




Supplementary Figure 3: MultiQC featureCounts assignments for nine samples

| 1. AhR (25) |
| --- |
| Metabolic pathways<br>Arginine and proline metabolism<br>Steroid hormone biosynthesis<br>Cell cycle |

Tryptophan metabolism
Retinol metabolism
ABC transporters
p53 signaling pathway
Bile secretion
Ascorbate and aldarate metabolism
MicroRNAs in cancer
Chemical carcinogenesis
Malaria
Glycerolipid metabolism
DNA replication
Nicotinate and nicotinamide metabolism
African trypanosomiasis
Ribosome biogenesis in eukaryotes
TNF signaling pathway
Peroxisome
Metabolism of xenobiotics by cytochrome P450
Chagas disease (American trypanosomiasis)
Drug metabolism - cytochrome P450
Biosynthesis of antibiotics
Insulin resistance

2. CAR/PXR (21)

Metabolic pathways
Arginine and proline metabolism
Steroid hormone biosynthesis
Cell cycle
Tryptophan metabolism
Retinol metabolism
ABC transporters
p53 signaling pathway
Bile secretion
Ascorbate and aldarate metabolism
MicroRNAs in cancer
Chemical carcinogenesis
Glycerolipid metabolism
Malaria
DNA replication
Nicotinate and nicotinamide metabolism
African trypanosomiasis
Ribosome biogenesis in eukaryotes
TNF signaling pathway
Peroxisome

| |
|---|
| Metabolism of xenobiotics by cytochrome P450 |
| 3. DNA damage (15) |
| Metabolic pathways<br>Arginine and proline metabolism<br>Steroid hormone biosynthesis<br>Cell cycle<br>Tryptophan metabolism<br>Retinol metabolism<br>ABC transporters<br>p53 signaling pathway<br>Bile secretion<br>Ascorbate and aldarate metabolism<br>MicroRNAs in cancer<br>Chemical carcinogenesis<br>Nicotinate and nicotinamide metabolism<br>TNF signaling pathway<br>Metabolism of xenobiotics by cytochrome P450 |

Supplementary Table 3. List of significant enriched pathways by KEGG analysis for AhR, CAR/PXR, and DNA damage groups that's shared by RNA-seq platform