Concordance of Microarray and RNA-Seq Differential Gene Expression

Independent Project Roles: Sriramteja Veerisetti (Analyst and Biologist)

Introduction

RNA-Sequencing is a next-generation sequencing methodology that is important in providing critical information such as which genes are highly expressed and which are not. This type of high-throughput sequencing can allow a more concrete understanding of underlying factors that can lead to potentially life threatening diseases. In 2014, Wang et al. generated illumina RNA-Sequencing and Affymetrix microarray data by utilizing a uniform set of liver samples from rats that were exposed to 27 different chemicals, each representing a different mode of action (MOA)[1]. The Wang et al. paper concluded that illumina RNA-Sequencing was much more precise and accurate, with a 90% differentially expressed genes (DEG) verification, in comparison to the 76% DEG with the affymetrix microarray[1].

Method

As a preface, in the Wang et al. paper a combination of both the microarray data as well the RNA-Sequencing data were utilized, however, in this project only the statistical information within the toxgroup_6_rna data set was utilized. More specifically, 15 samples were subjected to 1 of 3 chemicals: 3-Methylcholanthrene, Fluconazole, or Pirinixic Acid. These chemically induced samples were then further categorized via one of three MOA, which could have either been Aryl Hydrocarbon (ArH), CAR/PXR, or the peroxisome proliferator-activated receptor alpha (PPARA).

Onwards, the analyst role of this project had two key sections: conducting Microarray Differential Expression with Limma and producing concordance information between the differentially expressed genes in both microarray and RNA-Sequencing. After conducting differential expression analysis with Limma, the differentially expressed genes were then filtered so that only DE genes that have a p-adjust value less than 0.05 ($padj < 0.05$) remained. The top 10 differentially expressed genes post chemical analysis were compiled together and visualized in a tabular form (Table 2). In order to visualize the

statistical information, histograms of fold change values from differentially expressed genes and scatter plots of fold change versus the nominal p-value were developed.

Further, the concordance between the RNA-Seq and microarray expression was measured by performing statistical analysis on the differential expression results from both DESeq2 and Limma biostatistical packages. In order to calculate the concordance values, 2 separate mathematical formulas were utilized. The n_x value, which represents the "true" overlapping genes was calculated via the following equation:

$$n_x = \frac{Nn_0 - n_1 n_2}{n_0 + N - n_1 - n_2}$$

In this formula, other variables such as the n_0, n_1, and n_2 values all represent the number of differentially expressed genes that will overlap between the microarray and RNA-Sequencing processes. However, the N value is constant at 25,225 which is how many genes in total there are within the genome of the Sprague-Dawley rat genome. Finding the n_x value is crucial because it is used in order to find the concordance value. The concordance value can be calculated via the mathematical formula:

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

After the concordance values are calculated, two separate plots were generated: concordance vs number of DE genes from the RNA-Sequencing analysis as well as concordance vs the number of DE genes from the microarray analysis. Once the plots were made, the differentially expressed genes were split into "above-median" and "below-median" categories. In order to calculate the median, the baseMean variable was utilized for the RNA-Sequencing results, however, the AveExpr values were utilized for the microarray analysis. After computing the concordances for all DE genes, those in the "above-median" subset, and those in the "below-median" subset, the concordances were compiled in order to generate a

barplot. The barplot effectively displays the chemical utilized, concordance values, as well as whether the genes are in the above, below, or all DE genes subset.

To switch gears, the second role that was chosen was the biologist role. The differentially expressed genes in each MOA (AhR, CAR/PXR and PPARA) were filtered so that genes that meet the following conditions are included: p value < 0.05 and log2FoldChange > 1.5. The filtered gene set was utilized in a gene set enrichment analysis via the bioinformatics tool: Database for Annotation, Visualization, and Integrated Discovery (DAVID). DAVID is a bioinformatics tool that is used in order to garner functional information of a large gene set. The pathway enrichment results were then compared to the results that were computed in the Wang et al. paper. After performing the gene set enrichment analysis, the normalized expression counts matrix that was generated by the programmer was used in order to produce a clustered heat map that measured the counts. Prior to producing the heat map, the average count and coefficient of variance metrics were calculated and filtered so that genes with a coefficient of variation < 1 and average count > 100 are included in the heat map.

Results

To begin, after conducting microarray differential expression with Limma on the data sets, the differential expression results of the samples that were subjected to either 3-methylcholanthrene, Fluconazole, or Pirinixic Acid were written out as .csv files. The 3 separate .csv files were then filtered so that only genes that have padj values < 0.05 are included. The number of genes significant at padj values less than 0.05 are summarized in Table 1. It is clear that based on Table 1, there were 33 DE genes in the 3-Methylcholanthrene chemical analysis group that had a padj value of less than 0.05. There were 6016 DE genes in the Fluconazole chemical analysis group. Furthermore, there were 8790 DE genes in the Pirinixic Acid chemical analysis group.

Chemical	Number of Differentially Expressed (DE) Genes at padj < 0.05
3-Methylcholanthrene	33
Fluconazole	6016
Pirinixic Acid	8789

Table 1. Summary of the Number of DE Genes for each Chemical at padj < 0.05

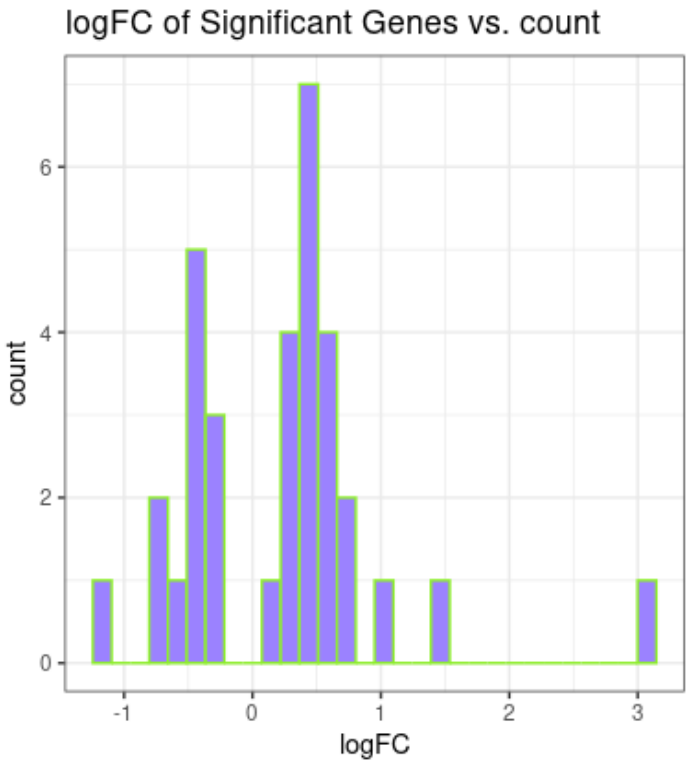
After garnering the number of genes that are significant at the filter threshold of padj < 0.05, the top then differentially expressed genes from each analysis was gathered and organized into Table 2.

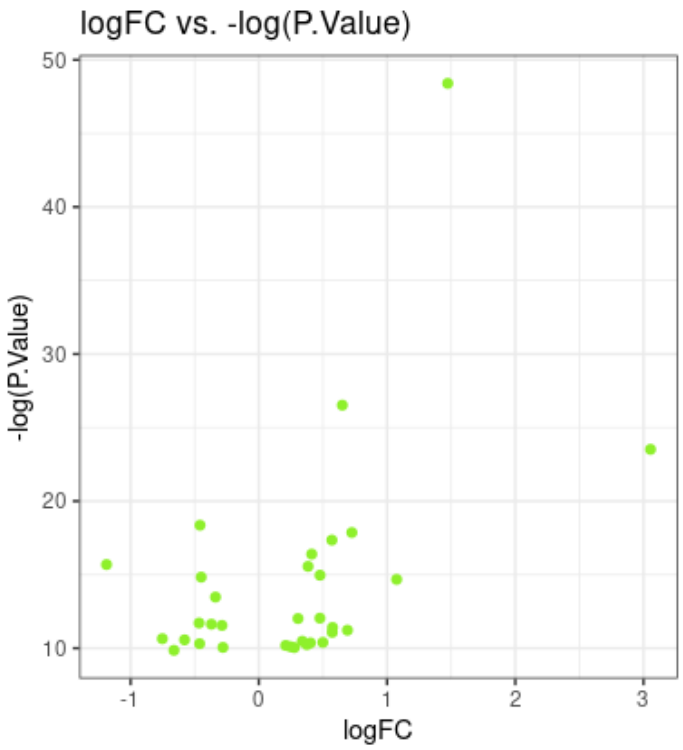
Chemical	Top Ten Differentially Expressed (DE)
3-Methylcholanthrene	1. 1383075_at 2. 1371381_at 3. 1386953_at 4. 1383136_a_at 5. 1370040_at 6. 1395411_at 7. 1370363_at 8. 1371646_at 9. 1369177_at 10. 1383066_at
Fluconazole	1. 1380027_at 2. 1370008_at 3. 1385710_at 4. 1389263_at 5. 1372988_at 6. 1373565_at 7. 1368520_at 8. 1389725_at 9. 1382126_at 10. 1373085_at
Pirinixic Acid	1. 1390436_at 2. 1391226_at 3. 1385527_at 4. 1371961_at 5. 1398916_at 6. 1381902_at 7. 1384166_a_at 8. 1386279_at

	9. 1374400_at 10. 1384355_at

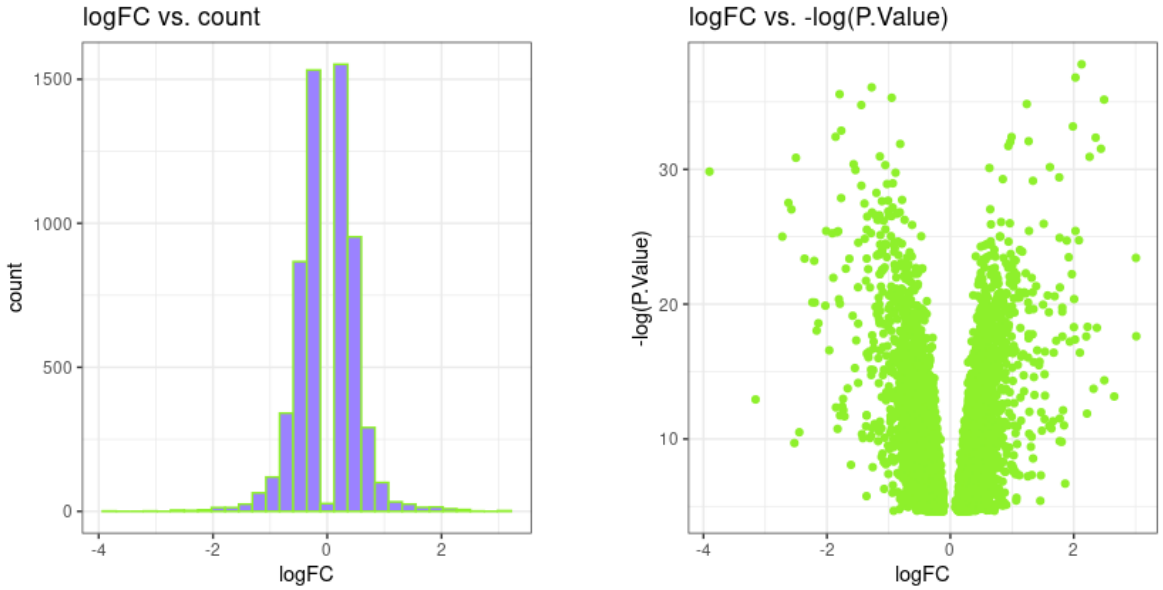
Table 2. Summary of Top Ten DE Genes for each Chemical

Once the top differentially expressed genes were identified and gathered together, histograms that plot the fold change values from the significant DE genes in each analysis, subjection to either 3-methylcholanthrene, Fluconazole, or Pirinixic Acid, were produced. On a similar note, scatter plots that visualize the relationships between the fold change vs. -log (nominal p-value) were produced for each analysis as well.





Figures 1 and 2. Figure 1 Histogram (Left) displays the log fold change (logFC) values on the x-axis and the count on the y-axis of the significant DE genes that were subjected to the 3-methylcholanthrene chemical. Figure 2 Scatter Plot (Right) displays the logFC on the x-axis and the -log(P.Value) on the y-axis of the significant DE genes that were also subjected to the 3-methylcholanthrene chemical.



Figures 3 and 4. Figure 3 Histogram (Left) displays the log fold change (logFC) values on the x-axis and the count on the y-axis of the significant DE genes that were subjected to the Fluconazole chemical. Figure 4 Scatter Plot (Right) displays the logFC on the x-axis and the -log(P.Value) on the y-axis of the significant DE genes that were also subjected to the Fluconazole chemical.

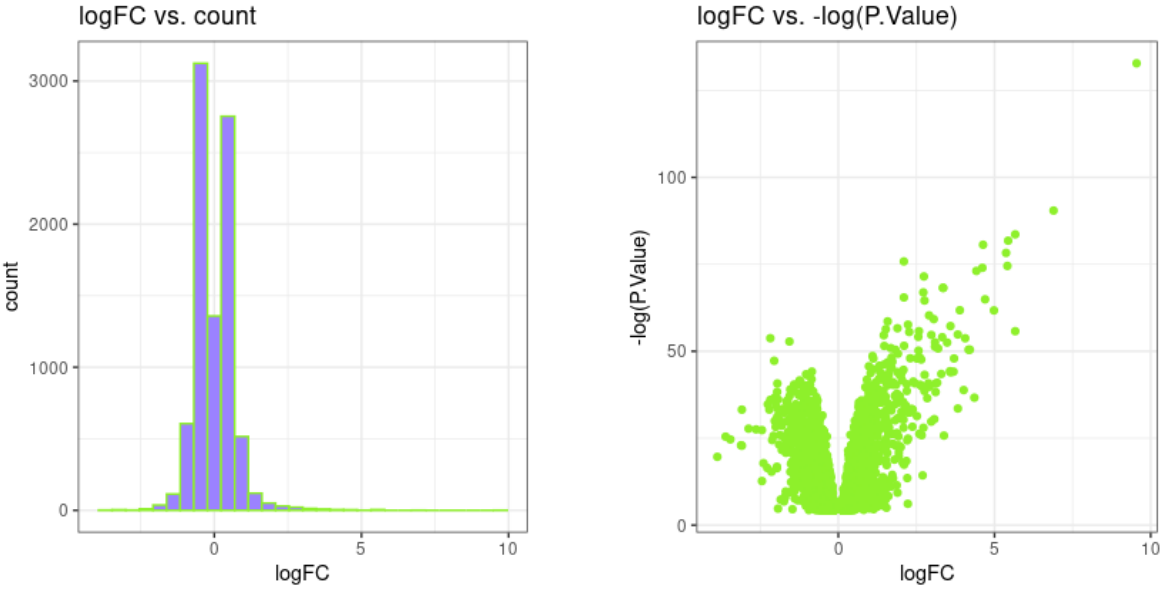


Figure 5 and 6. Figure 5 Histogram (left) displays the logFC on the x-axis and the count on the y-axis of the significant DE genes that were subjected to the Pirinixic Acid chemical. Figure 6 Scatter Plot (Right)

displays the logFC on the x-axis and the -log(P.Value) on the y-axis of the significant DE genes that were also subjected to the Pirinixic Acid chemical.

After the “microarray differential expression with Limma” section was completed, the attention shifted over to the concordance between microarray and RNA-Seq DE genes section. Here, the concordance values overall, above, and below the median value in the baseMean variable for the RNA-Sequencing results and the AveExpr values in the microarray analysis were calculated and summarized in Table 3.

Chemical & MOA	All, Above, or Below Median	Concordance Value
3-Methylcholanthrene/AhR	All	-0.024645527
Fluconazole/CAR-PXR	All	4.100449664
Pirinixic Acid/PPARA	All	1.310816693
3-Methylcholanthrene/AhR	Above	-0.031537916
Fluconazole/CAR-PXR	Above	4.256695416
Pirinixic Acid/PPARA	Above	1.312986801
3-Methylcholanthrene/AhR	Below	-0.003721693
Fluconazole/CAR-PXR	Below	2.032449565
Pirinixic Acid/PPARA	Below	0.667111616

Table 3. Summary of concordance values for each chemical and MOA combination as well as whether the genes within the analysis are above or below the mean, or all the DE genes.

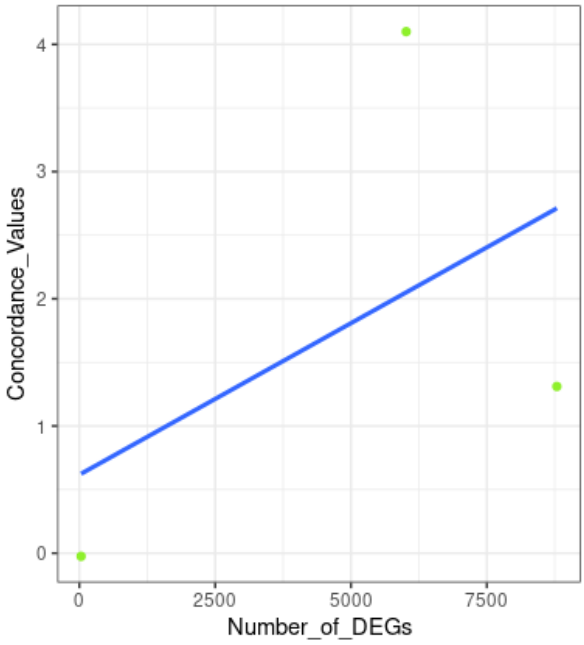


Figure 7. This plot summarizes the relationship between the number of differentially expressed genes from the Affymetrix microarray analysis with the total concordance.

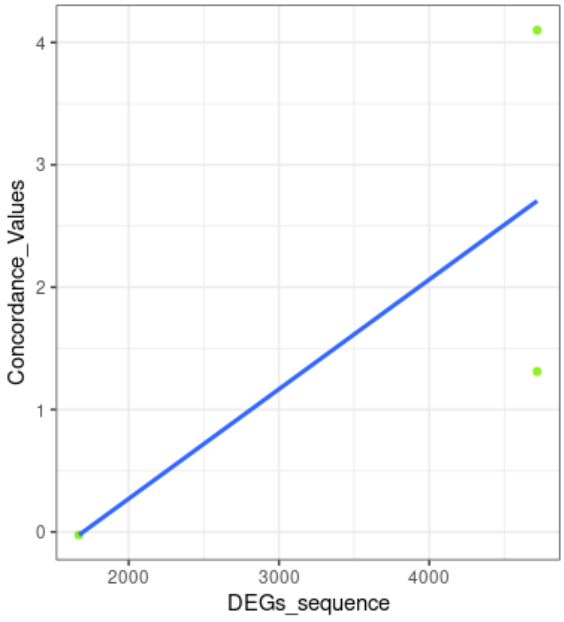


Figure 8. This plot summarizes the relationship between the number of differentially expressed genes from the illumina RNA-Sequence analysis with the total concordance.

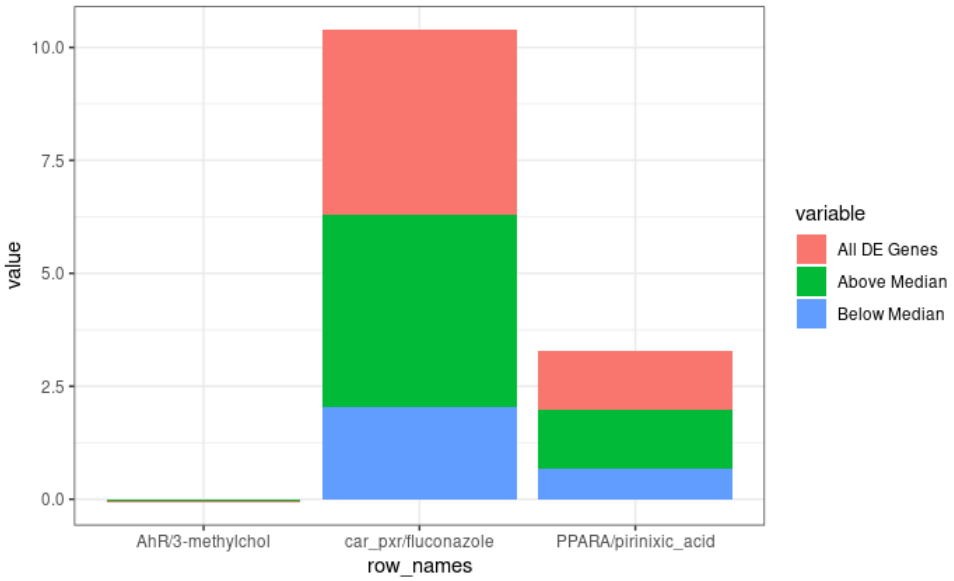


Figure 9. Combined plot of the concordances including the overall, above and below median genes, for each analysis was produced.

Next, for the biologist section, a table of enriched pathways that are identified with the differentially expressed genes from the analyses was produced after conducting DAVID analysis.

DAVID Analysis on Genes in Different Mode of Actions (MOA) *Indication that Term Overlaps with Supplementary Table 4 in Wang et al.				
MOA	Chemical	Enrichment Score	Cluster Number	Examples Terms in Clusters
PPARA	Pirinixic Acid	14.07	1	• Fatty Acid Metabolism* • Lipid Metabolism*
PPARA	Pirinixic Acid	4.59	2	• Peroxisome • Peroxisomal Membrane
PPARA	Pirinixic Acid	4.39	3	• long -chain fatty acid transport* • long -chain fatty acid transporter* activity
CAR/PXR	Fluconazole	5.8	1	• GO:0014070~Response to organic cyclic compound • GO:0009410~Response to drug*

CAR/PXR	Fluconazole	4.32	2	• Retinal Metabolism • Chemical carcinogenesis - DNA adducts • Drug metabolism - other enzymes
CAR/PXR	Fluconazole	3.86	3	• Retinal Metabolism • Xenobiotic metabolic process*
AhR	3-methylcholanthrene	2.36	1	• Metabolism of xenobiotics by cytochrome P450* • Chemical carcinogenesis - DNA adducts • Chemical carcinogenesis - receptor activation
AhR	3-methylcholanthrene	0.29	2	• Transmembrane • Integral Component of membrane

Table 4. Summarization of the enriched pathways that identify with the group's differentially expressed genes from analysis. The table contains the MOA, chemical used, enrichment score, cluster number, as well as common GO terms that are located within the cluster.

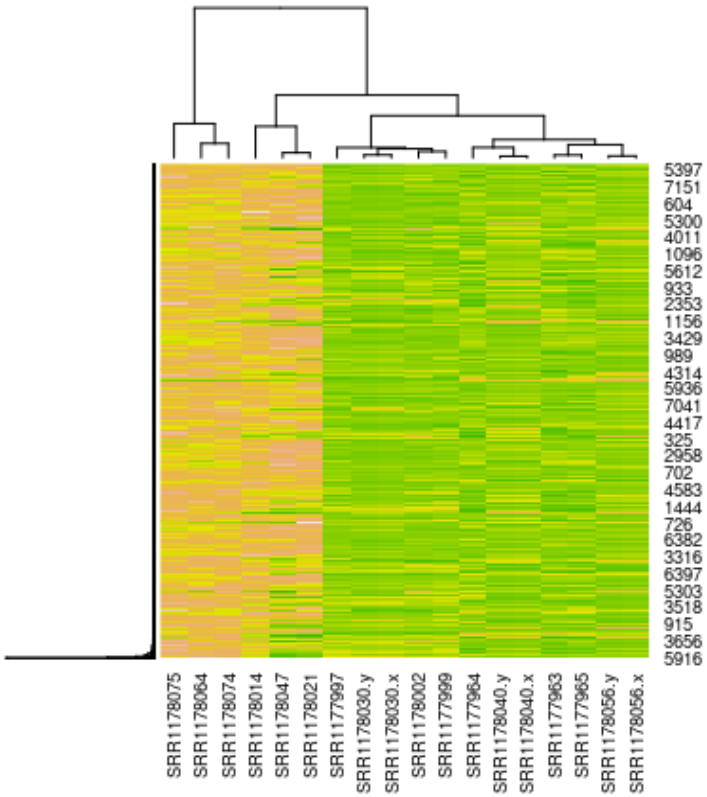


Figure 10. Heatmap that shows the 15 samples on the x-axis and the counts on y-axis

Discussion

The main takeaway from the Wang et al. paper is that the illumina RNA-Sequencing methodology is more precise and accurate in terms of differentially expressed gene verification by PCR quantification. Statistically, the RNA sequencing method produced a 90% DEG verification, while the Affymetrix microarray method produced a 76% DEG verification. In essence, the cross-platform concordance between the illumina RNA-Sequencing and Affymetrix microarray methodologies is dependent on the gene expression level, treatment effect size, and also the mode of action. Figures 7 and 8 both support this conclusion because they demonstrate the notion that the larger the treatment effect of the number of differentially expressed genes, the greater the concordance value will be. When comparing the number of differentially expressed genes from RNA-Sequencing (treatment effect) between the Wang et. al paper and the individual project statistics, it can be noticed that there is a general trend of 3-Methylcholanthrene being the lowest treatment effect, Flucanazole having the second highest treatment effect, and Pirinixic Acid having the largest treatment effect. Although this is the general trend, the number of DEGs that fall under each of these categories is not consistent with the paper. This could be because only statistical information within the toxgroup_6_rna data was utilized. In the future, multiple tox groups should be analyzed at the same time in order to garner a more accurate representation of the Wang et. al paper.

Onwards, Figure 9 truly illustrates how important the chemical treatment group/MOA combination has on the concordance value of the differentially expressed genes. It can be consistently seen that the concordance values of DE genes that are in the below-median category have a lower concordance value regardless of what chemical group or MOA was utilized. The incredibly low concordance value of the 3-Methylcholanthrene/AhR MOA combination should be noted. This low concordance value could be due to the large discrepancy between the number of differentially expressed genes post limma analysis (33 genes) and number of differentially expressed genes post DESeq analysis (1669 genes). If anything, this finding can further support the conclusion that a low treatment effect can lead to a lower concordance value between the microarray and RNA-Seq data. On the other hand, the

concordance value of the overall (all DE Genes) can be seen to be the greatest value regardless of the chemical and MOA.

Furthermore, the enriched pathways that were produced in the individual process via DAVID had some overlapping common pathways with the paper. For example, the Fatty Acid Metabolism, Lipid Metabolism, and long chain fatty acid transport terminologies under the PPARA mode of action resembled some of the terms in the Supplementary Table 4 located within the Wang et. al paper. However, very few terminologies for the CAR/PXR and AhR modes of actions were similar to those in the Wang et. al paper. For example, the xenobiotic metabolic process was one of the only similar pathways that was shared between the paper and the individual project results for both the CAR/PXR and AhR MOAs. A big reason for the discrepancies between the paper and the individual project could be due to the fact that the Wang et al. paper used 27 chemicals, with sets of three chemicals sharing a common MOA (there were 5 MOA). Furthermore, errors in analysis and terminology clustering could have caused discrepancies between the paper and the project as well.

When constructing the heatmap, it was absolutely necessary to filter the data so that only genes with a coefficient of variation greater than 1 and an average count value greater than 100 are included. By undergoing this process, thousands of genes were filtered out so that only significant genes and their counts are included within the heatmap. The samples were properly plotted on the horizontal axis and their respective counts were located on the vertical axis.

Conclusion

Although the enriched pathway analysis via DAVID did not match the results of the paper, the individual project was able to support the conclusion that the concordance between RNA-Seq and microarray analysis is largely dependent on the treatment effect size, MOA, and even the gene expression levels. In the future including all 27 chemicals and modes of action could help produce results that are more similar to the paper.

References

1. Wang, Charles et al. “The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance.” Nature biotechnology vol. 32,9 (2014): 926-32. doi:10.1038/nbt.3001