Project 3

The Concordance Between RNA-Seq and Microarray Data Depends on Chemical Treatment and Transcript Abundance

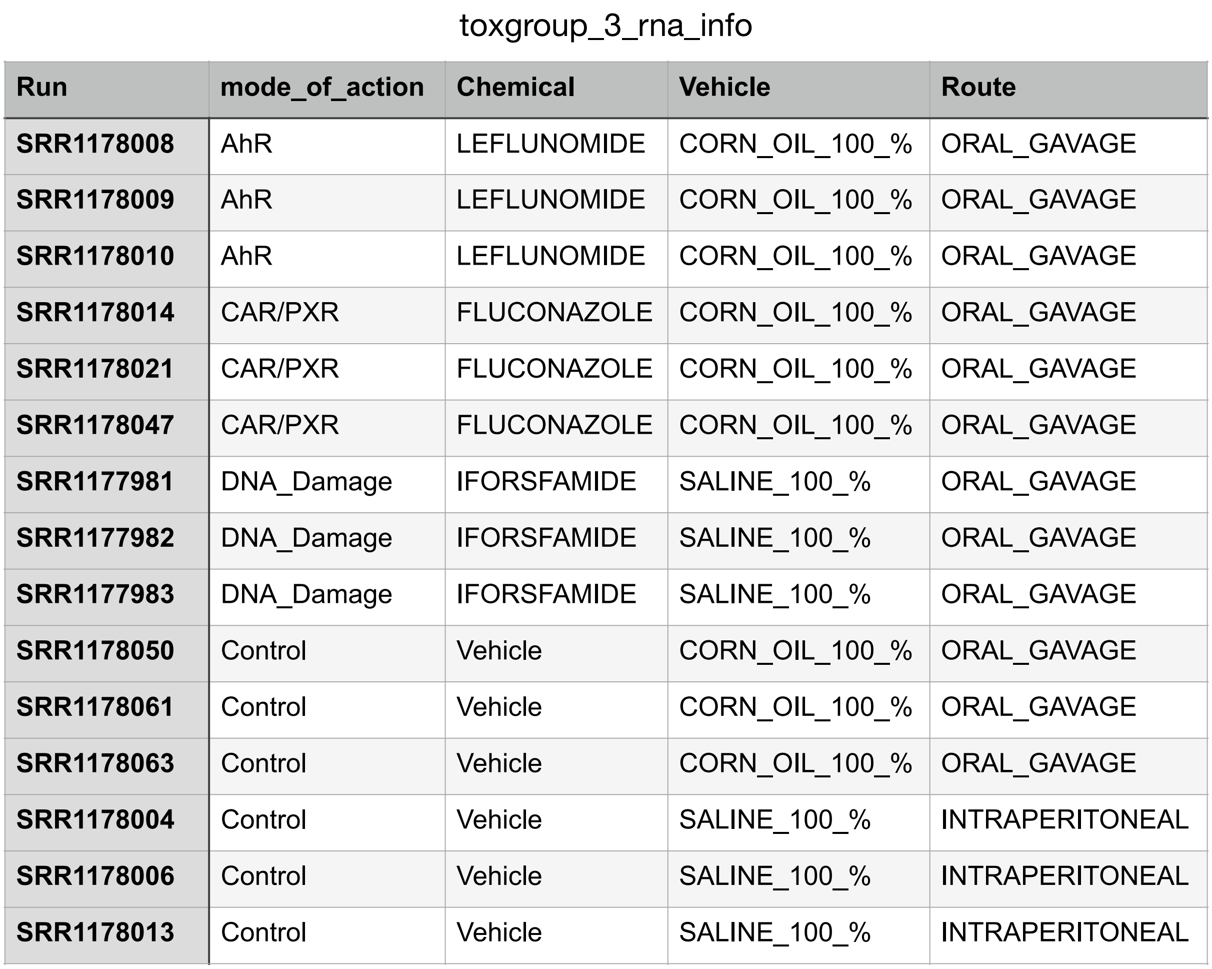
Data Curator: Cory Williams
Programmer: Caroline Muriithi
Analyst: Zeyuan Cao
Biologist: Nicholas Mosca
TA: Dakota

INTRODUCTION

Both RNA-seq and microarray are two known methods that are used to study and measure gene expression. The purpose of the Wang et. al study was to compare RNA-seq and microarray analysis. This was done by characterizing the concordance of differential gene expression across platforms, test and compare how effective each platform is at detecting expected pathway-level effects based on each treatments MOA, and assess the MOA prediction accuracy of each platform using a test set [1]. One of the advantages of RNA-seq is that it allows unbiased genome-wide gene-expression profiling. However, its concordance with the microarray must be meticulously assessed in order to be used in clinical application. RNA-seq is also superior to microarray, in regards to differentially expressed genes (DEG) verification by quantitative PCR. For low expressed genes, RNA-seq has improved accuracy as compared to microarray analysis. However, when determining predictive classifiers, both RNA-seq and microarray had similar performance. As a result, the intended application of transcriptomic research is essential for method decision-making. The findings of Wang et. al. determined that the cross-platform concordance with differentially expressed genes or enriched pathways is highly correlated with treatment effect size, gene-expression abundance and the biological complexity of the mode of action (MOA). Using similar analysis methods, the purpose of our study is to replicate Wang et. results [1]. One distinction to make is that in our study, a majority of the analysis was conducted using precomputed data.

Data

RNAsequencing was performed using Illumina HiSeq2000 system; these systems are relatively common; however, they have recently become obsolete and replaced with newer models.[2] System and application version 1.9 was used for encoding and proper analysis. The Illumina TruSeq RNA Sample Preparation kit and SBS kit v3 was used in tangent with Illumina manufactures protocol to prepare the samples for sequencing.[1] The data was generated post RNA extraction and Illumina sequencing and the Rattus norvegicus Rnor_6.0 genome was used for alignment.[1-3] There were a total of 15 samples in the tox group we have selected. Tox group 3 provided us with six controls and nine samples. The nine samples were categorized into three different modes of action (MOA), and each sample according to their MOA had the same chemical, vehicle, and route (Table 1). There were a total of 3 different MOA in the data observed named Aryl hydrocarbon receptor, orphan nuclear hormone receptors, and on receptor-mediated. These were used in addition to the 3 chemicals to induce a reaction, LEFUNOMIDE, FLUCONAZOLE, IFOSFAMIDE.



toxgroup_3_rna_info

Run	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	IFORSFAMIDE	SALINE_100_%	ORAL_GAVAGE
SRR1177982	DNA_Damage	IFORSFAMIDE	SALINE_100_%	ORAL_GAVAGE
SRR1177983	DNA_Damage	IFORSFAMIDE	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

Table 1. Tox group 3 sample information
The mode_of_action (MOA) represents the mediator or receptor process for each sample. The three different MOA's that are present in tox group 3 is aryl hydrocarbon receptor (AhR), orphan nuclear hormone receptors (CAR/PXR), and on receptor-mediated—DNA damage (DNA_Damage). Chemical represents the chemicals that the mouse was dosed with during the experiment. For tox group 3, the only chemicals that were used were LEFLUNOMIDE, FLUCONAZOLE, and IFOSFAMIDE. The vehicle represents the substance used to house the chemical for injection into the mouse. The two vehicles used were 100% corn oil (CORN_OIL_100_%) and 100% saline (SALINE_100_%). The route represents how the chemical along with the vehicle substance was administered into the mouse, ORAL_GAVAGE represents that it was introduced into the mouse orally by force while INTRAPERITONEAL means it was introduced by needle injection around the abdomen of the mouse.

For comparison to microarray data, the microarray that was used was an Affymetrix microarray. Each paired end read had a sequencing depth of 23-25 million. The reads are paired-end the output of two fastq files by the same sample confirms this. [5] The sample files were first run through fastqc and then through the STAR aligner against the Rattus norvegicus Rnor_6.0 genome. Multiqc was used to pool all of the samples together in a comprehensive report for streamline analysis (Table 2).

MultiQC Report

Sample Name	% Aligned	M Aligned	% Dups	% GC	Length
SRR1177981_1	85.5%	12.2	56.1%	48%	101 bp
SRR1177982_1	88.6%	15.2	58.5%	48%	101 bp
SRR1177983_1	85.0%	13.7	57.6%	48%	101 bp
SRR1178004_1_control	87.7%	17.2	61.1%	48%	101 bp
SRR1178006_1_control	88.8%	19.1	59.2%	49%	101 bp
SRR1178008_1	88.1%	13.3	56.2%	49%	101 bp
SRR1178009_1	90.2%	16.3	62.2%	49%	101 bp
SRR1178010_1	90.7%	16.9	62.5%	49%	101 bp
SRR1178013_1_control	89.5%	14.4	58.4%	49%	101 bp
SRR1178014_1	81.6%	14.3	53.9%	49%	50 bp
SRR1178021_1	83.6%	14.6	48.7%	49%	100 bp
SRR1178047_1	84.2%	14.4	48.5%	49%	100 bp
SRR1178050_1_control	87.6%	14.1	54.6%	48%	100 bp
SRR1178061_1_control	89.2%	56.5			
SRR1178063_1_control	88.3%	39.3			

Table 2. MultiQC table
Table of the group analysis produced by MulitQC. % aligned represents the percentage of sequences that were successfully uniquely aligned to the reference genome. M aligned is the number of uniquely aligned sequences by the millions. % dups is the percentage of duplications per sample. % GC is the percentage of GC base pair (bp) per sample. Length is the length of bp per read, M seq is the total sequence number by million base pairs.

No samples appeared to have a quality score that was alone too low; each sample seemed to follow the same trend in terms of the higher the bp position count, the more the quality declined (Figure 1). As opposed to excluding any samples, a better approach would be to trim the samples from the beginning bp to about the 70th bp position. The quality score graph does, however, raise concern considering that the controls are also included and appear to have the same trend line. I do not believe that all of the samples having the same trend line to be 100% accurate due to the assumption of the control to be of higher quality throughout the whole sample as opposed to decreasing along with the samples.

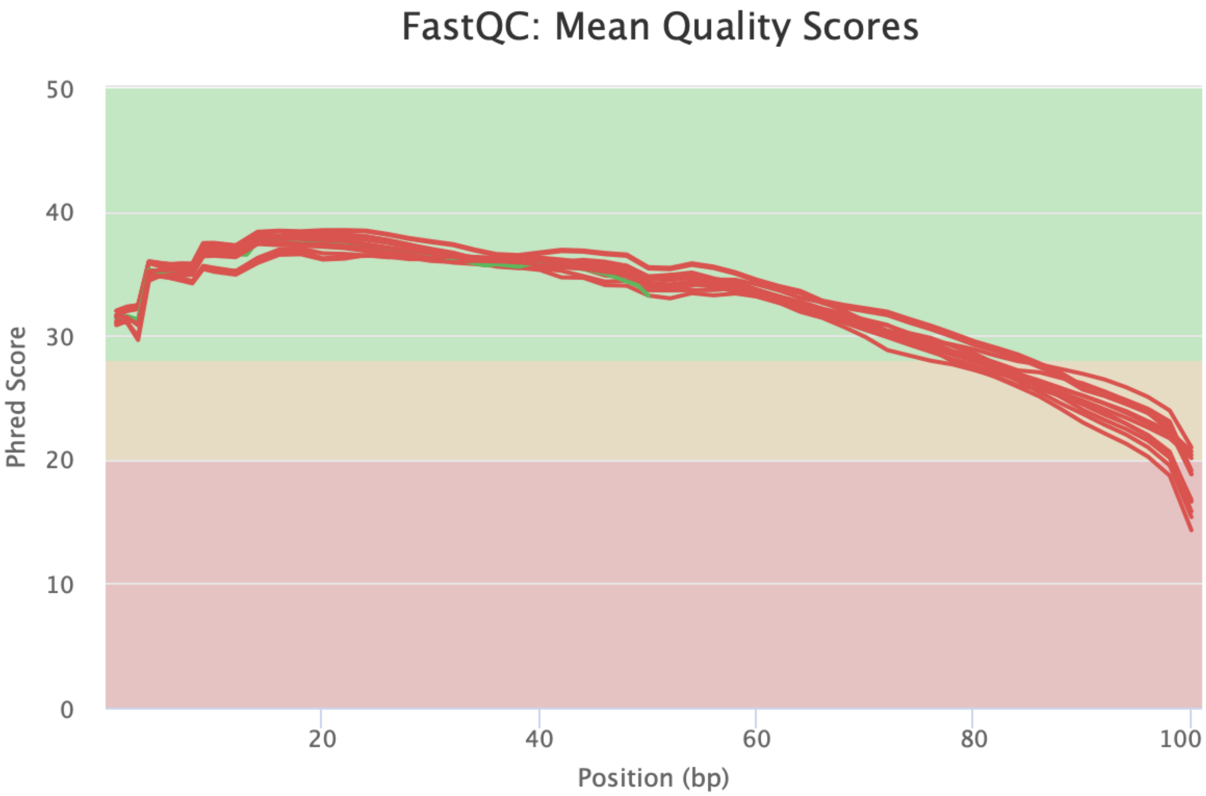


Figure 1.
Created by MultiQC, a chart containing all the samples processed with an x-axis representing the base pair count of 1 sequence. The max on the x-axis is 100bp, which is the max average base pair per sequence. The Phred score on the y-axis is a quality meter with starting at 20, and going up represents preferred good reads.

Two main issues with that data that were observed are overrepresented sequences and duplication. Overrepresented sequence refers to a sequence that seemed to reappear for alignment more often than expected (Figure 2). The potential source of error that was provided by the FastQC analysis program seemed to be due to TruSeq adapter index 4.[4,5] The adapter is what binds to the flow cell provided in the TruSeq RNA Sample Preparation kit used for sample prep, which implies that the issue lies within the adapter and most likely occurred during the preparation step. [2]

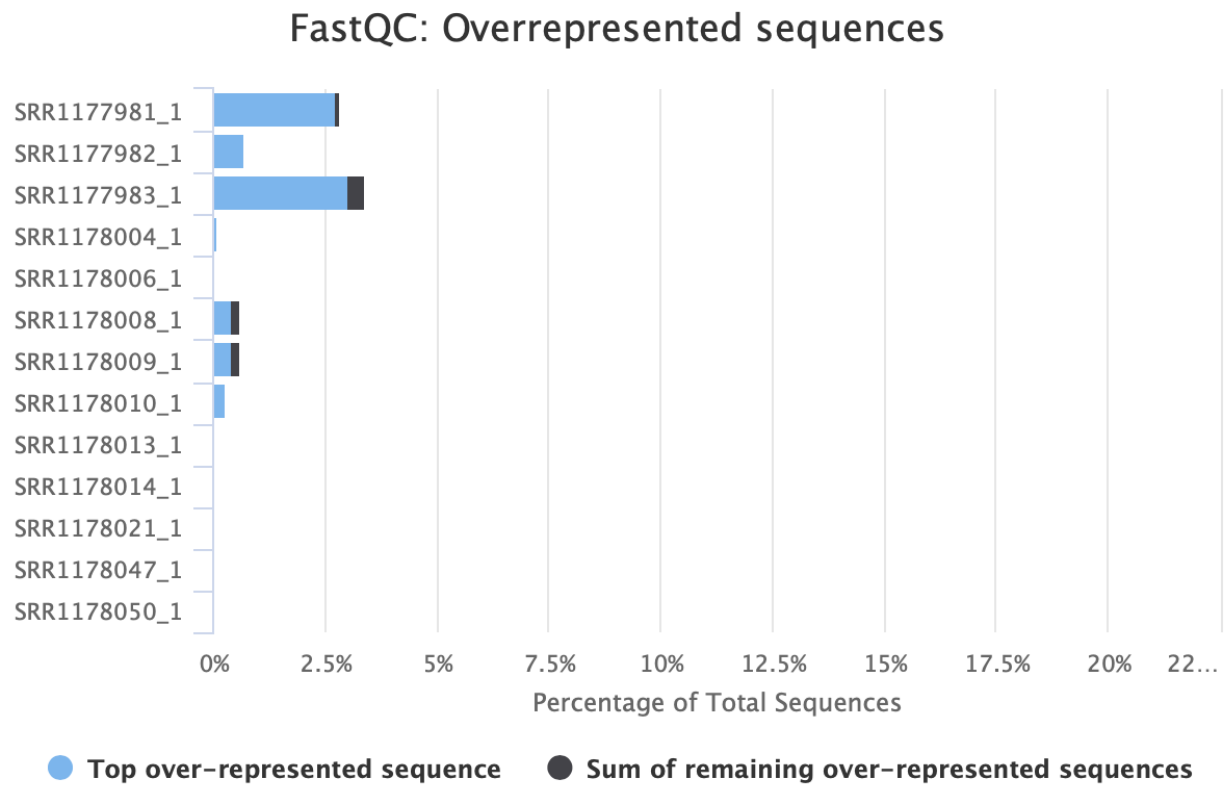


Figure 2.
Produced by FastQC showing the percentage of total sequences overrepresented on the x-axis and the sample names on the y-axis.

Only six samples appeared to have overrepresented sequences with an average percentage of about 2%, which does not lead to recommend a complete do-over of the experiment starting at the peeping stage, but it is something to be noted.

The second issue appears to be duplication, with a majority of the samples achieving above 40% duplication numbers (Figure 3). A majority of reasons can cause duplication, with a major one being cluster miscalling. Another possible cause for duplication is an error in the PCR process.[5]

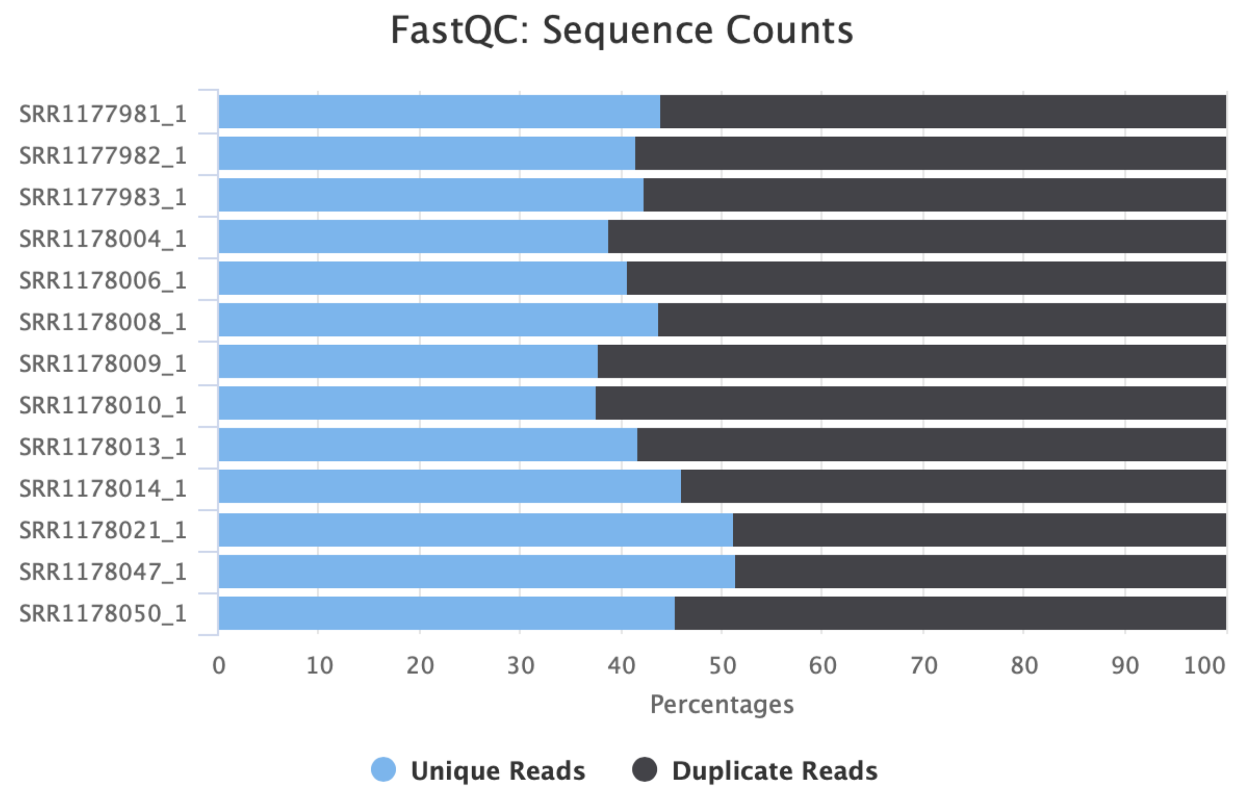


Figure 3.
Chart produced by FastQC representing sequence counts per sample with the y-axis showing the percentage of unique reads vs. duplicate reads and the x-axis representing sample names.

Methods

Data obtained post RNA extraction using RNA extraction protocol and RNA sequencing using the Illumina system, was checked for quality using the terminal module FastQC version 0.11.7. Paired-end sequencing reads from FastQC were aligned to the Rattus norvegicus Rnor_6.0 genome using the STAR aligner version 2.6.0c. Terminal module MultiQC version 1.6 was used with the assistance of python version 2 to pool together output results from both FastQC and STAR aligner output into one comprehensive report for downstream analysis.

Quantification of transcript abundance using FeatureCounts

Alignments from STAR were used to count reads against a gene annotation using the FeatureCounts program (present within the subread module Version 1.6.2). This was done by assigning the reads from the 6 bam alignment files generated using the STAR alignment to genes in a reference genome. The counts were generated to serve as a measure of transcript abundance. FeatureCounts was run on all 6 files as a batch job, using the bam file as the input and a rat reference gene annotation file in the gtf format. MultiQC was used to was run on the count files produced from the 6 samples. This was to assess the quality of these assignments and any outliers samples. The read counts from all the 6 samples were then combined into a single CSV file using the read.csv function in R. Boxplots of the count distribution for each of the samples were generated. log2(counts)+1 was done in order to generate the plot. 1 was added to log2(counts) to prevent performing log2(0) since the counts file had multiple values as zero.

Differential expression analysis using DEseq2

DEseq2 is a Bioconductor package that uses negative binomial regression to estimate/calculate count differences given a statistical design. The count matrix generated using FeatureCounts was combined with the counts of the control samples to estimate the differential expression between treated samples and their controls.

The Differential gene expression was studied for 1 different treatment condition with chemical simvastatin and 100% corn oil vehicle. The control counts matrix and treatment counts were subset to only incorporate the treatment counts for a specific treatment and the corresponding controls with the same vehicle before DESeq2 is run. The analysis using DESeq2 was conducted once and a matrix obtained that consisted of the log2FoldChange, p-value and adjusted p-value. The number of significant genes with p-adjusted value <0.05 were estimated. Subsequently, the top ten differentially expressed genes(DE) from the analysis were sorted based on the p-value. Histograms showing the fold change values from the significant DE genes for each analysis were created for each analysis. Additionally, scatter plots of fold change vs nominal p-value were created.

DAVID functional annotation clustering

Full list of upregulated and downregulated genes were analyzed together to find biological pathway enrichment. The gene list was uploaded using REFSEQ_MRNA gene names. Top Gene Ontology terms and from the highest enrichment custer are displayed in table#. Genes were filtered before enrichment analysis by Log2-fold change absolute value >=1 and adjusted P-value <0.05.

Results

STAR Alignment Scores

For STAR alignment, the average percentage of sequences aligned was 92 percent, with an average read length of 100bp per sequence. This average score of 92 percent is relatively good for an alignment score considering this represents that most samples had above an 80% alignment score with no abnormal outliers in either direction (low alignment score, high alignment score). However, the average score of duplication across the samples was concerning 56 percent. This average duplication score implies that most samples had a duplication rate of around 50 percent, meaning a lot of the paired-end sequence read showed up multiple times in the alignment. A cause for this duplication can range from cluster miscalling to an error in the PCR process. A proper conclusion for the cause of the duplication requires more analysis that was not done for the reproducing of this experiment.

FastQC

Upon evaluation of the FastQC output file on the samples, all of the files processed returned results following the same trend line. A good quality Phred score of between 30 to 40 from 0 base pairs to about 70 base pairs was observed. From 70 bp to the max of 100bp per sequence, the quality dramatically dropped to a Phred score of between 10 and 25. There are many reasons for a quality score decline like this. One common reason would be adapter issues in the RNA extraction protocol; however, just like the STAR alignment duplication issue discussed above, a further analysis that was not conducted would be needed for accurate evaluation. This data is still usable and can be corrected for this error by using a trimming protocol that will cut off at the 70th bp position eliminating the poor quality results if needed.

The Distribution and Analysis of Read Counts

Sample Name	% Assigned	M Assigned
SRR1178075__rn4_STAR__sortedByCoord	63.7%	50.9
SRR1178074__rn4_STAR__sortedByCoord	60.8%	61.7
SRR1178055__rn4_STAR__sortedByCoord	62.0%	17.5
SRR1178039__rn4_STAR__sortedByCoord	61.8%	21.0
SRR1178035__rn4_STAR__sortedByCoord	57.7%	13.4
SRR1178029__rn4_STAR__sortedByCoord	60.8%	17.6

Table 3: MultiQC Summary Table. *Percentage of mapped reads assigned to the gene-ids in the reference genome*

The number of read counts mapped to genes in the rat genome were generated using the featureCounts program which is a highly efficient general-purpose read summarization program that counts mapped reads for genomic features. Subsequently, MultiQC was used to check the quality of the reads and identify outliers samples. It was found that on a whole, 60.8% of all the reads were assigned to gene-ids. Control sample SRR1178075_rn4_START_sortedByCoord had the highest of reads assigned to gene-ids (63.7%) followed by the SRR1178055_rn4_STAR_sortedByCoord(62.0%) sample treated with simvastatin. Control sample SRR1178035_rn4_START_sortedByCoord treated with the vehicle had the lowest reads assigned to gene-ids(57.7%) as shown in the table 3 above.

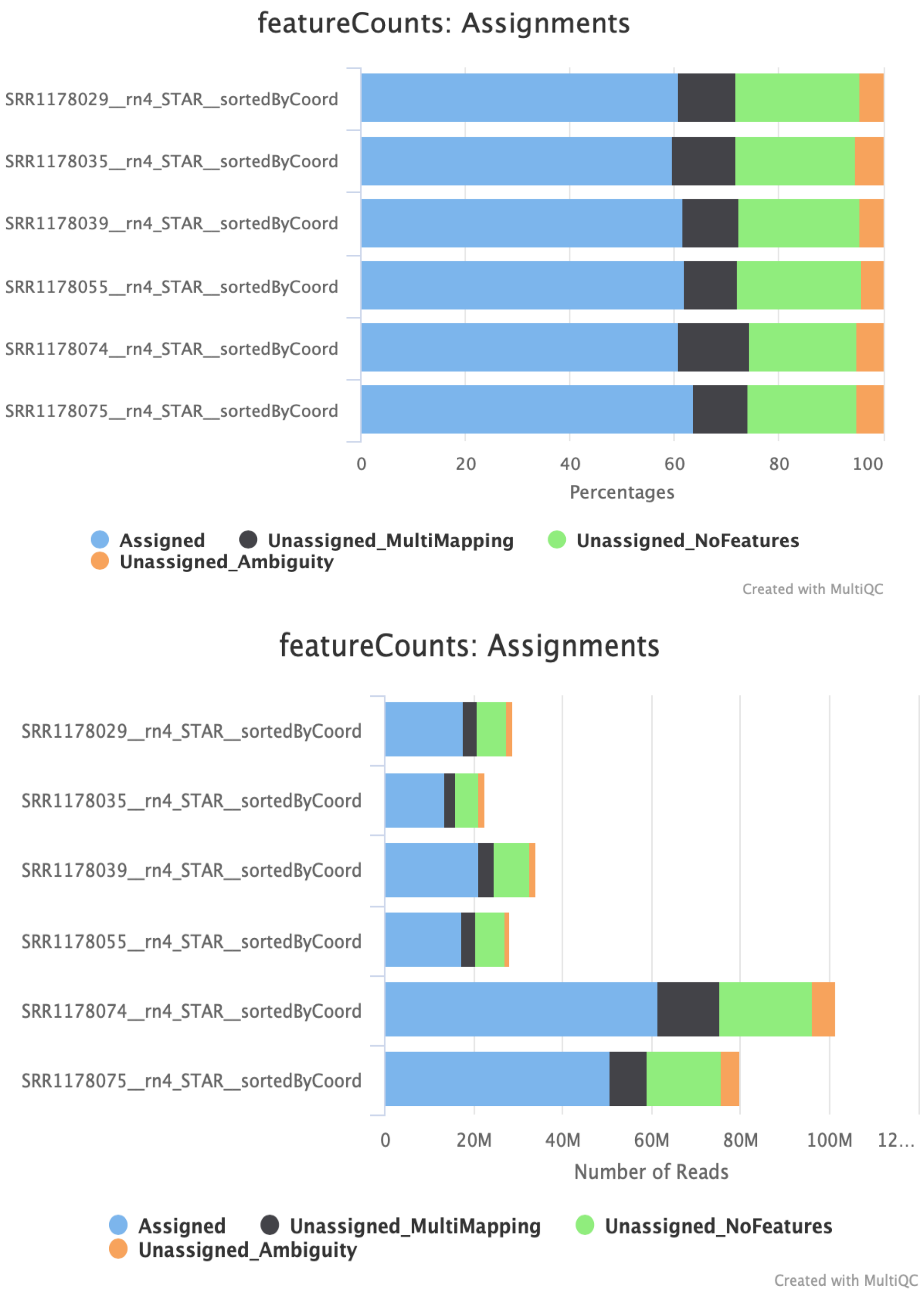


Figure 4: Percentage of Assigned reads as shown by MultiQC report. The figure also shows the percentage of unassigned reads for every sample after running featureCounts. 60.8% of all the reads were assigned to gene-ids as shown by the blue color.

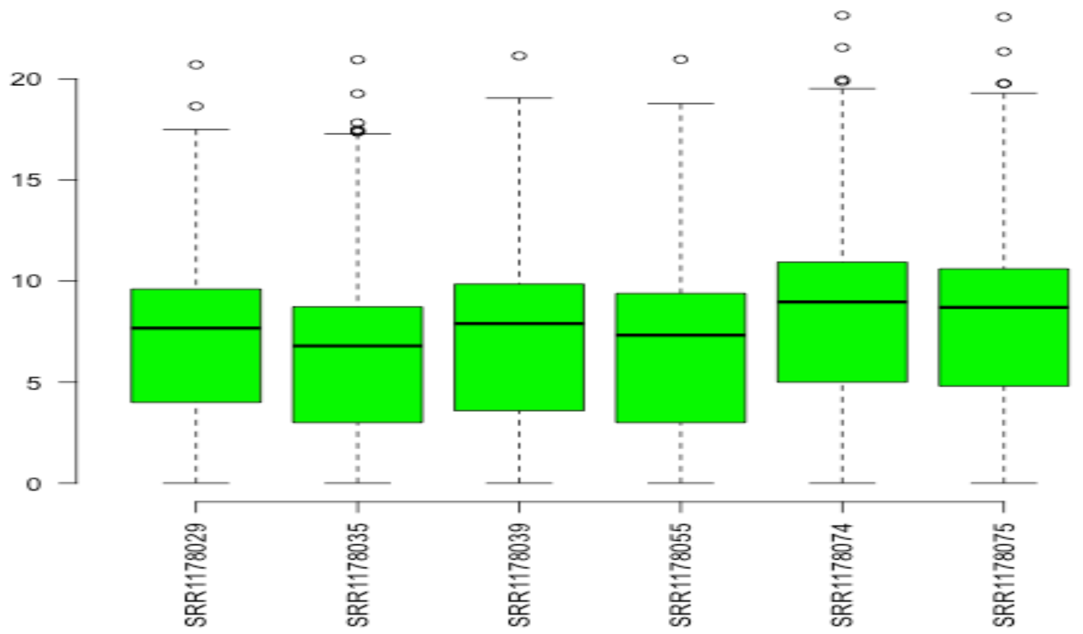


Figure 5: Boxplot showing the count distribution of each sample in the box plot

Box plots were generated to analyze the distribution of the read counts of the 6 samples generated by the featureCounts program. The horizontal black line across each box shown in the figure represent the median of log counts. The overall density distributions of raw log-intensities of the samples do not deviate much from the median although they are not very identical. The samples SRR1178029, SRR1178039, SRR1178074 SRR1178075 showed some slight deviation from the median as shown in figure. There were no major observable distribution differences between the samples.

Differential expression Analysis for RNA-seq using DESeq2

DESeq2 analysis was used to identify genes that were differentially expressed between the samples that were subject to treatment with different chemicals (Simvastatin) and their corresponding controls.   It was found that the number of genes differentially expressed at an adjusted p-value < 0.05 for the sample groups treated with

Simvastatin were 4308. The top 10 differentially expressed genes that were filtered based on the p-value are shown in table below.

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
NM_001000750	79.85	-2.92	0.36	-7.84	4.34E-15	9.29E-14
NM_001001505	4513.99	-0.94	0.17	-5.43	5.59E-08	5.15E-07
NM_001001509	5022.74	-1.21	0.23	-5.19	2.16E-07	1.78E-06
NM_001001511	184.14	1.15	0.27	4.32	1.55E-05	9.13E-05
NM_001001512	3565.32	-0.41	0.12	-3.39	0.000704	0.002796
NM_001001718	542.68	0.39	0.16	2.37	0.018023	0.045571
NM_001001719	76.32	-2.42	0.32	-7.42	1.20E-13	2.24E-12
NM_001001800	808.73	-0.90	0.18	-5.04	4.69E-07	3.67E-06
NM_001002016	2614.60	0.79	0.12	6.39	1.70E-10	2.19E-09
NM_001002253	562.60	-0.74	0.17	-4.33	1.52E-05	8.96E-05

Table 4 Top 10 genes from the DESeq2 run for samples treated with Simvastatin. The top 10 DE were filtered based on p-values

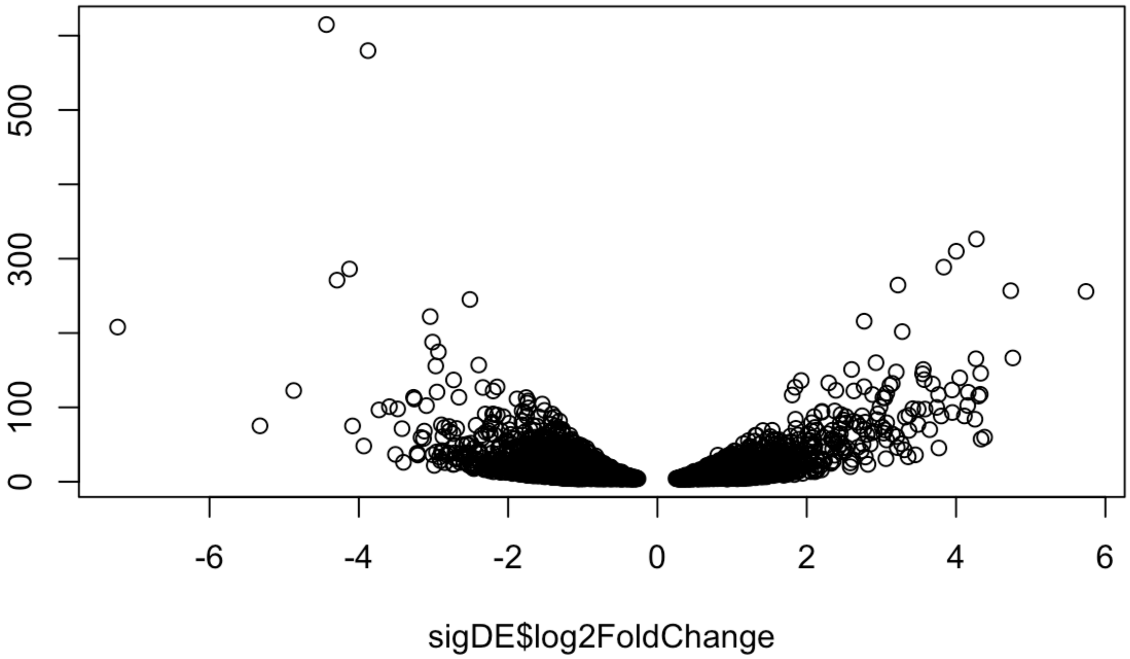
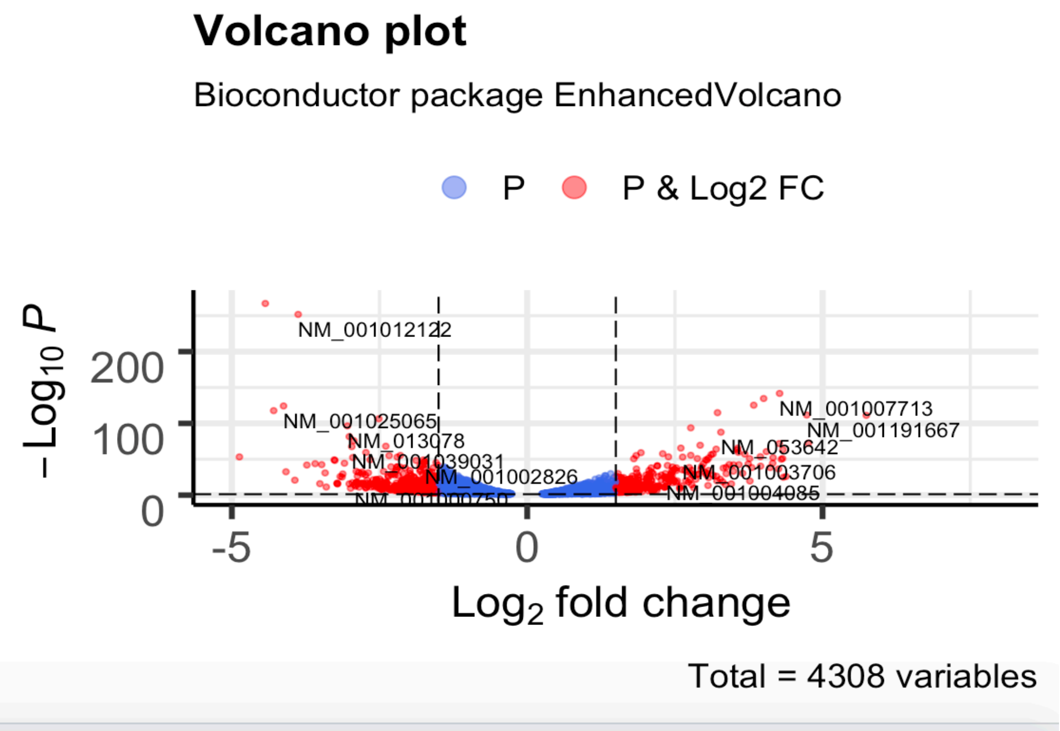
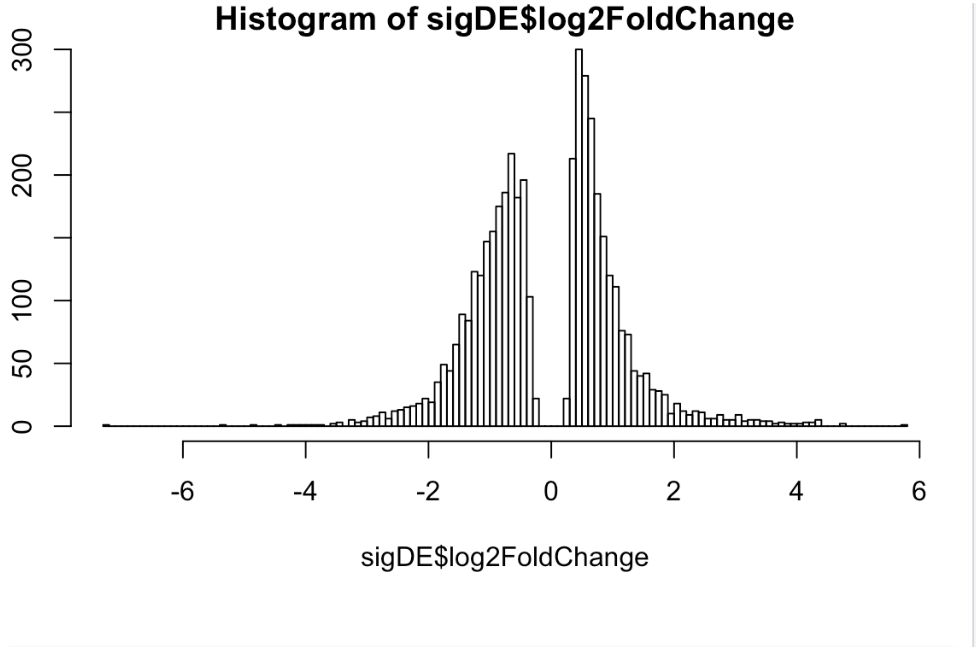


Figure 6 represents the scatterplot of fold change versus log p-value of differentially expressed genes: The number of genes downregulated are almost equal to the number of genes upregulated. All samples were treated with Simvastatin



Figures 7 represents volcano plot of fold change versus log p-value: The total variables are the number of genes differentially expressed at an adjusted p-value < 0.05 for the sample groups treated with Simvastatin



Figures 8 represents histograms of log fold change for the significant differentially expressed genes: All samples were treated with Simvastatin. The histogram showed a slight distinction between the number of unregulated and downregulated genes.

Histograms of the fold change values and scatter plots of fold change versus nominal p-value were also generated for the most significant genes as shown in the figure 8. It was found that for the samples treated with simvastatin, the number of genes downregulated are almost equal to the number of genes upregulated as seen from the scatter plot in figure 6. However, the histogram shows a more clear distinction where the samples treated with simvastatin had a slightly higher number of upregulated genes.

Analysis of microarray data

We used limma to analyze the normalized expression data obtained from microarray in order to determine differential expression between the treatments and control samples. There are 31099 probesets for each of the treatment we analyzed, leflunomide, fluconazole and ifosfamide. 1348, 4174 and 26 of them were detected as significant(i.e. having an adjusted p-value of less than 0.05), respectively. We sorted the results by adjusted p-value and mapped the top 10 significantly differential expressed probesets from each treatment to their Refseq IDs. The Refseq IDs, log fold change values and adjusted p-values are summarized in the tables below.

While the leflunomide and fluconazole treatments yielded similar level of significance at comparable number of genes, isosfamide yielded a much smaller number of significant genes at a much larger level of significance. Additionally, several probesets under the isosfamide treatment were unable to be matched to any Refseq ID.

Refseq ID	log fold change value	adjusted p-value
NM_012540	7.44	1.52*10^-18
NM_012541	1.38	2.87*10^-16
NM_013123	1.60	1.14*10^-13
NM_001015008	-0.69	3.79*10^-12
NM_001108008	0.86	6.20*10^-12
NM_001044259	1.15	2.71*10^-11
NM_013156	0.55	5.56*10^-11
NM_001130558	-3.10	1.16*10^-10
NM_013156*	0.87	1.38*10^-10
NM_175846	0.90	1.54*10^-10

Table 5: Top 10 differentially expressed genes under leflunomide treatment ranked by adj.p-value. Log fold change and adj.p-value were rounded to 2 decimal places. *: two probesets were matched to NM_013156.

Refseq ID	log fold change value	adjusted p-value
NM_053288	1.32	5.20*10^-15
NM_001134844 **NM_001198676***	2.47	2.64*10^-14
NM_030845	2.79	1.58*10^-12
NM_001130558	-3.58	2.17*10^-12
NM_019340	2.02	8.34*10^-12
NM_199391	-0.94	9.03*10^-12
NM_173295	1.38	1.38*10^-11
NM_019216	2.21	2.81*10^-11
NM_001191698	1.54	3.96*10^-11
NM_001105883	-0.78	1.20*10^-10

Table 6: Top 10 differentially expressed genes under fluconazole treatment ranked by adj.p-value. Log fold change and adj.p-value were rounded to 2 decimal places. *: The same probeset was matched to two different Refseq IDs.

Refseq ID	log fold change value	adjusted p-value
NM_001107877	-0.64	1.75*10^-7
NM_022220	-0.56	1.40*10^-6
NM_022627	0.58	2.58*10^-6
NA	0.89	3.00*10^-6
NM_001191992	-0.71	3.03*10^-6
NA	0.34	3.20*10^-6
NM_012580	0.64	3.21*10^-6
NM_053985	-0.34	3.43*10^-6
NM_053587	-0.76	4.33*10^-6
NM_001008367	0.28	5.22*10^-6

Table 7: Top 10 differentially expressed genes under isosfamide treatment ranked by adj.p-value. Log fold change and adj.p-value were rounded to 2 decimal places. NA: Not available.

We found that leflunomide induced the widest range of log fold change values. Majority of the log fold change values fall between -2.5 and 2.5. Fluconazole have a smaller range of effect than lefulnomide, with majority of genes having log fold change values between -2 and 2. The Ifosfamide have the smallest range of effect at between -1 and 1. Also, The histogram shows that for all three treaments, the mode value of log fold change is roughly -0.25.

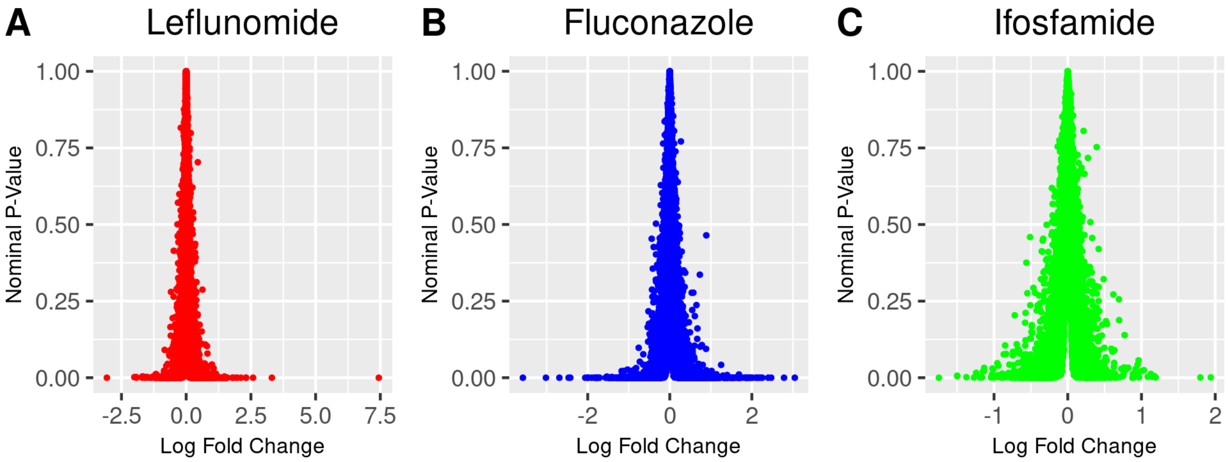


Figure 9: Scatter plots showing nominal p-value versus log fold change for each treatment.

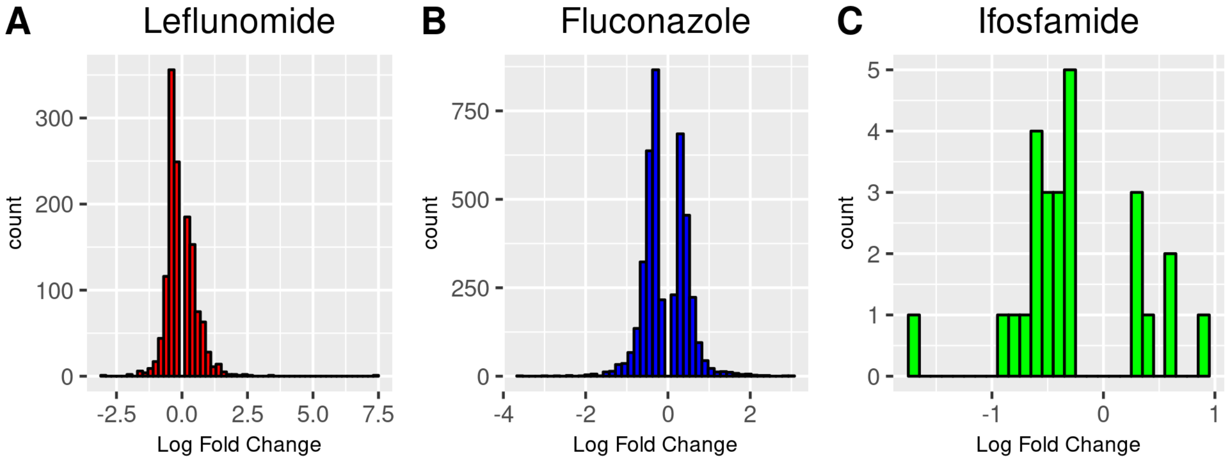


Figure 10: histograms of log fold change values for each treatment.

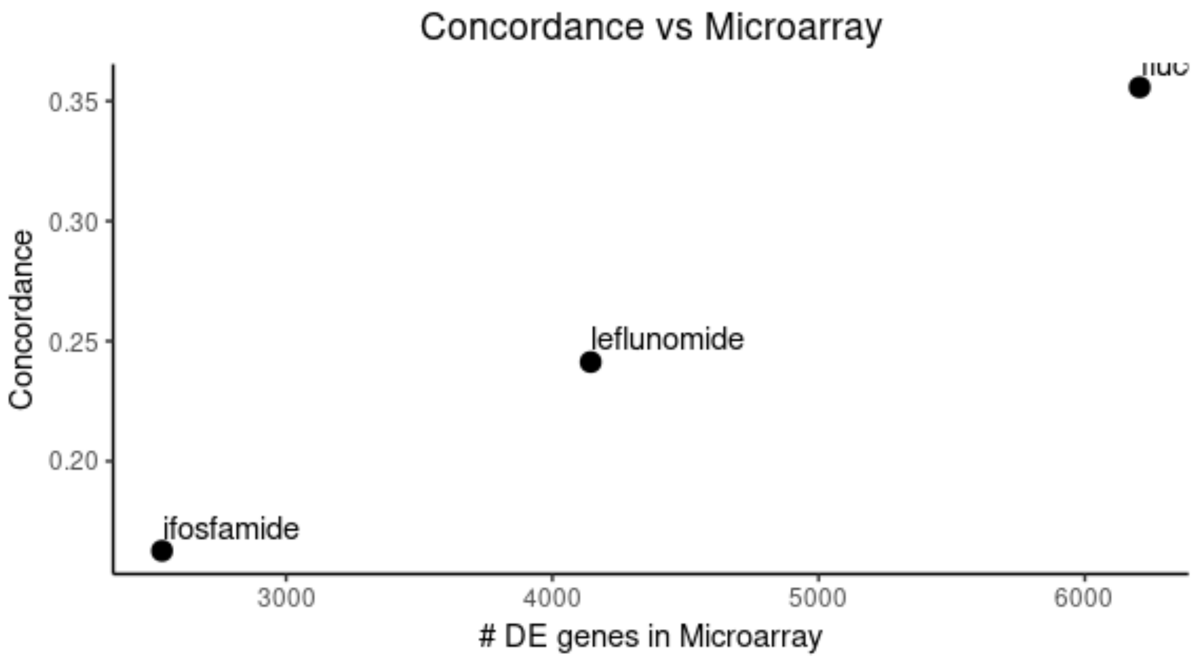


Figure 11 : Shows a concordance vs Microarray of DE genes in Microarray

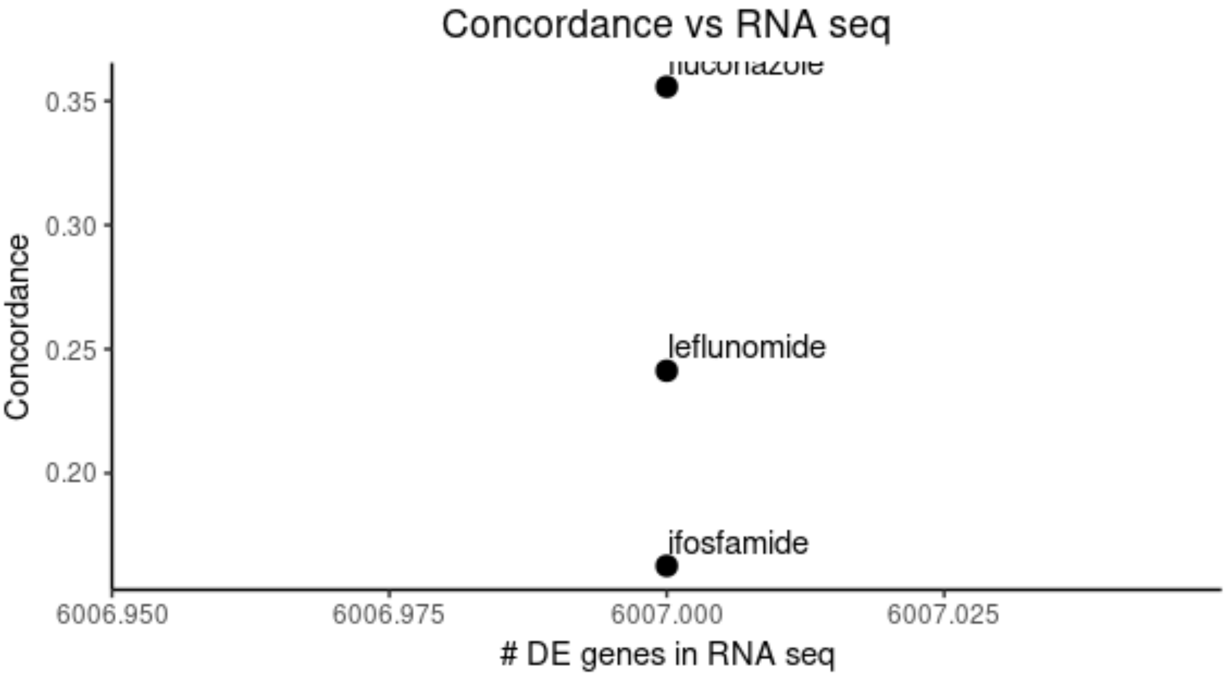


Figure 12: showing concordance vs RNA seq with differential genes in RNA seq on the x axis against the concordance in the y axis. All chemicals line up on x axis 6007.00.

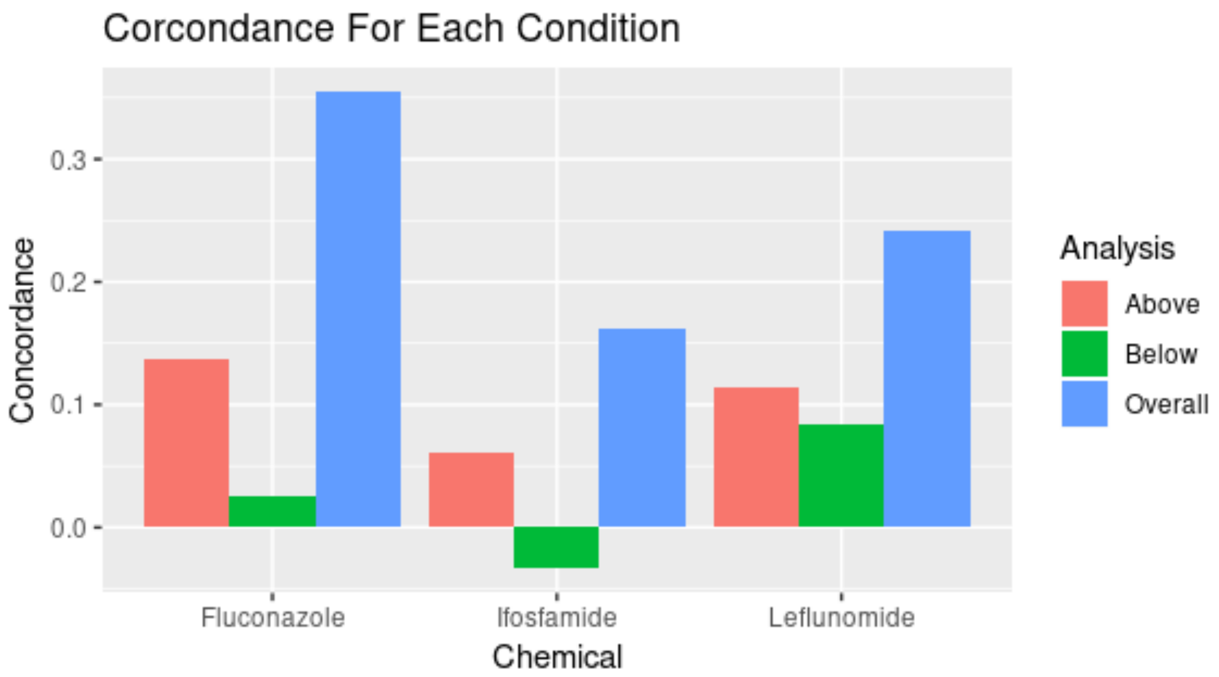


Figure 8 Bar plot of the overall, below-median and above-median concordance values for each chemical treatment

	Fluconazole	Leflunomide	Ifosfamide
Overall Concordance	0.366	0.241	0.162
Below-median Concordance	0.025	0.083	-0.033
Above-median Concordance	0.136	0.114	0.062

Table 8 Table summarizing the overall, below-median and above-median concordance values for each chemical treatment. All values are rounded to 3 decimal places

The chemical fluconazole showed the largest concordance value, in both overall and above-median groups. Ifosfamide showed the smallest concordance value among all three chemicals, with the below-median concordance value being negative. We were able to identify that above-median concordance values are larger than the below-median ones, in agreement with Wang et al. Concordance values of overall, below and above-median groups for each chemical are summarized in the table below.

Discussion

Sample files were obtained post-RNAseq were and decompressed and processed through FastQC for quality evaluation. Files were then aligned to the Rattus norvegicus Rnor_6.0 genome using the STAR aligner tool in a shared computer cluster (SCC) terminal. FastQC and STAR aligner output files were pooled together using the MultiQC module in an SCC terminal and produced a report for easy viewing.

The remaining analysis was conducted on a chemically treated group of controls and samples treated with Simvastatin. Simvastatin, also known as the brand name product Zocor, is a statin(lipid-lowering) drug derived synthetically from a fermentation product of Aspergillus terreus[3]. Statins are medications that work by reducing the amount of cholesterol made by the liver. They lower cholesterol by interfering/inhibiting the cholesterol biosynthetic pathway. Genes functions of each drug reflect the mode of action of each drug. It is classified as HMG-CoA reductase inhibitor, non-aromatic; Hypolipidemic Agent[1,4]. We would also expect simvastatin to cause more downregulation of genes. However, samples treated with simvastatin did not show a significant difference in gene expression results from the RNA-Seq Differential Expression analysis using DESeq2. The DESeq2 analysis was conducted so as to determine differential expression between treated samples and appropriate controls. A histogram, scatter plot and volcano plot were generated for the most significant genes to show if there was a difference in upregulation or downregulation genes as a result of treatment with the chemical simvastatin. All diagrams were almost in concordance and depicted the number of genes downregulated to almost be equal to the number of genes upregulated which was inconclusive or can be prematurely attributed to the chemical not having any effect. However it would have been quite interesting to see if there was a significant difference in the number of DE expressed genes if there was access to other samples treated with different chemicals. From the analysis, we were not able to establish if their different patterns in the number of differentially expressed genes if the sample were treated with different chemicals. Therefore we can not confidently state that we were able to reproduce DESeq2 results from the original paper.

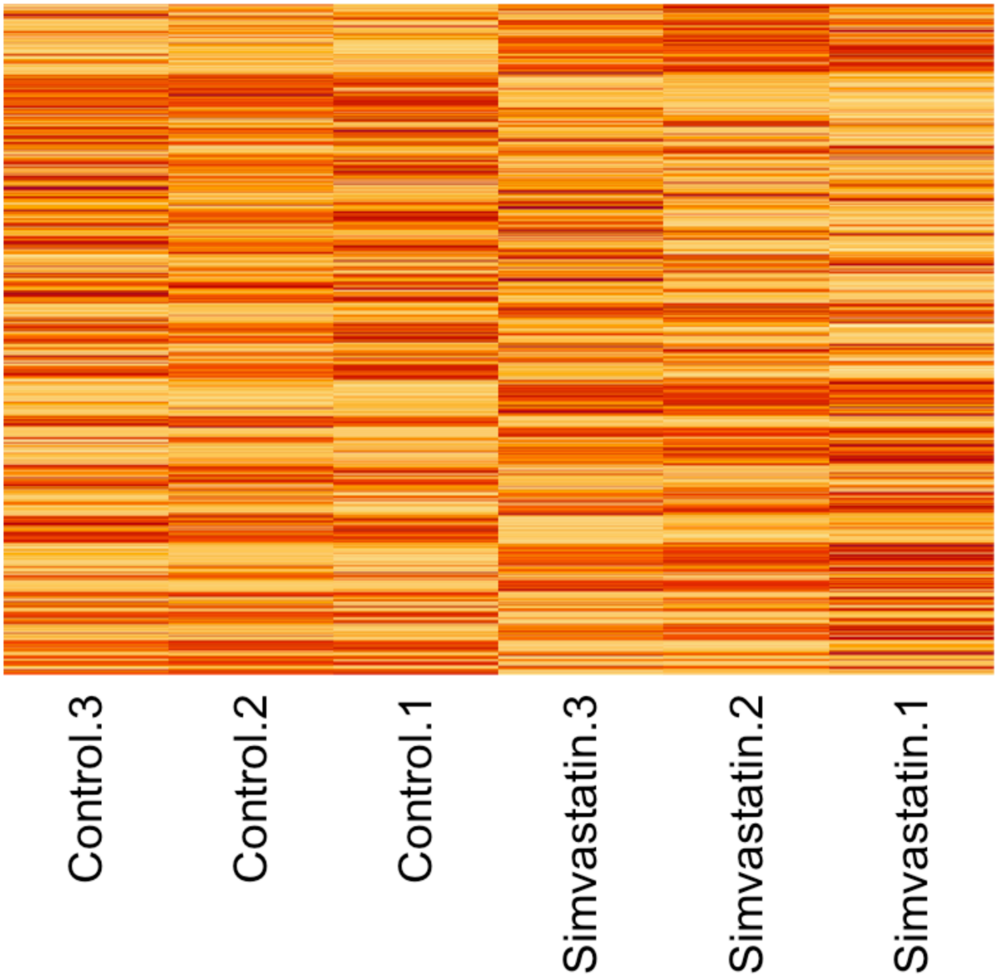


Figure 9: Clustered Heatmap of Differentially expressed genes from control and Simvastatin treated samples
A total of 11269 genes were expressed. Displayed are expression results for 3 control technical replicates and 3 technical replicates of Simvastatin treated rat samples. Route of delivery for treatment was orally with the vehicle being corn oil. No clear clustering groups developed although 627 genes were upregulated and 930 downregulated based on Log2fold change >= |1| and adjusted P-value <0.05.

Annotation Cluster 1	**Enrichment Score: 7.4**
Category	Term
UP_KEYWORDS	Lipid biosynthesis
UP_KEYWORDS	Sterol metabolism
UP_KEYWORDS	Cholesterol metabolism
UP_KEYWORDS	Sterol biosynthesis
UP_KEYWORDS	Steroid metabolism
UP_KEYWORDS	Cholesterol biosynthesis
Annotation Cluster 2	Enrichment Score: 6.5
KEGG_PATHWAY	Fatty acid degradation
KEGG_PATHWAY	Fatty acid metabolism

Table 9: Gene set enrichment pathways clustered by Enrichment score
A total of 1557 pathways were mapped based on the Simvastatin treated rat samples. Displayed is the top metabolic pathways associated with a variation or lipid metabolism. Simvastatin is a lipid statin that impacts cholesterol metabolism. Both Annotation Cluster 1&2 display pathways related to lipid metabolism. It is evident that the top differentially expressed genes impact pathways that resemble lipid metabolism.

Results are based on a single chemically treated group of controls and technical replicates. Three rat liver samples were treated chemically with Simvastatin. Gene set enrichment analysis was conducted via DAVID and biological pathways clusters were generated. The biological pathways displayed in table 9 directly correspond with the function of Simvastatin.The mode of action (MOA) was determined as HMGCOA.[1] Simvastatin is a lipid statin that impacts cholesterol metabolism[2]. The primary gene set enrichment clusters identified pathways were consistent with the pathways described in Wang et al., 2014 .

Our main findings do suggest a correlation between chemically treated samples and an associated mode of action (MOA) reported by Wang et al., 2014. Gene set enrichment analysis using DAVID identified biological pathways associated Simvastatin show in table 9.[1] However the number of differentially expressed genes did differ as a result of the use of precomputed data for the majority of the analysis. If given a full dataset we would have shown differentially expressed genes for multiple chemicals that would have correlated with multiple MOAs.

A clustered heatmap was produced after filtration of the top differentially expressed genes. Only genes with Log2-fold change absolute value >=1 and adjusted P-value <0.05 were plotted. No clear clustering pattern was displayed in figure 9. This can be attributed to the use of sample data or a mismatch between the chemical used and sample expression norm counts.

The premise of this study is reasonable, results indicated that RNA-Seq is more effective for evaluating genes with low expression. The correlation between chemical and associated MOA is strong based on the results previously discussed. Although we do see a correlation with limited data this experimental design of using chemicals could result in a number of unpredictable factors. Pharmacogenetic impacts on these chemicals on rats may result in unexpected gene up-regulation or deregulation.

Unfortunately we were not fully able to completely reproduce the entire study. We did use sample data which is a portion of what a full replication would have been. Limited differential expression data was used to perform gene set enrichment analysis via DAVID. This set of genes was solely related to one chemical treatment (Simvastatin). Gene set enrichment results identified two primary clusters related to lipid metabolism shown in table 9 which are consistent with the chemically treated samples. Although we had a fraction of samples We did draw a biological link to the HMGCOA MOA from supplement table 2 in Wang et al., 2014.

Conclusions

The findings of Wang et. al. determined that the cross-platform concordance with differentially expressed genes or enriched pathways is highly correlated with treatment effect size, gene-expression abundance and the biological complexity of the mode of action (MOA). The toxicology group sample we used for our analysis had samples treated with simvastatin. Our gene set enrichment analysis identified a link between simvastatin and the HMGCOA moa. RNA-seq differential expression analysis using DESeq2 did not show a clear distinction in upregulation or downregulation genes as a result of treatment with the chemical simvastatin based off of the scatter plots. Upon extra filtering RNA-seq differential expression results from DESeq2 analysis did show 627 upregulated genes and 930 downregulated genes as a result of treatment with the chemical simvastatin.

Our results were almost in concordance and depicted the number of genes downregulated to almost be equal to the number of genes upregulated. Although up and down regulated genes were identified in Figure 9 we do not show a clear concordance. We can not confidently assess whether this part of the analysis replicated Wang et al., 2014.

For the data curation, some of the problems encountered were selecting the correct version of the module tools used for post-analysis. For example, STAR aligner's newest version 2.7 was not compatible with the data set in the format that it was provided in. MultiQC also had a hard time pooling together the reports due to MultiQC not being compatible with the newest version of python version 3 that is used as a backbone to process the files. Other issues were file types and conversions. STAR aligner was able to only process the RNA seq data in ".fq" format as opposed to the ".gz" format it was provided to us in. The ".gz" format is a way of compressing the ".fq" format. The ".fq" format is the raw file data format produced post-sequencing. The compression is used for faster file handling and transfer. In order to run the files through STAR aligner, the gunzip module was used on the files in order to unzip (decompress) them to their original format.

For the programmer role, I did not encounter problems for my analysis with precomputed data. However, I could not make concrete results from my results and it would have been good to work with more samples as opposed to only one sample (tox group) available in the precomputed data so as to evaluate different responses caused by different chemical treatments on the liver samples. I was not able to work on the real data because I did not receive data from the data curator in a timely manner. The new idea as to how we summarize and write the report(not simultaneously) left me a bit weary especially when an individual also has to depend on one person to also partially input individual parts. Also not being able to have an overall summary of all the written parts including figures from other roles like before made it a bit difficult to evaluate and come to a conclusion if there was concordance between different analysis methods.

The challenges for the biologist role revolved around drawing a clear biological conclusion using the sample data. While using sample data we still could draw the link between one chemical and one MOA used in the original paper. As a result we could not draw a link between supplement table 4 in Wang et al., 2014 because we used data from one chemical and three replicates. We did however link to a MOA in supplement table 2 in Wang et al., 2014. I was partially able to reproduce supplement table 4 with gene enrichment pathways from one chemical. The clustered heat map was produced but I could not see a correlation between control and sample expressed genes.

Writing for this project was the most difficult because using a subsection of data did not give much of an overall conclusion. I think if we would have had a full dataset set used throughout the project my heatmaps and gene set enrichment would have had a stronger impact on our results.