Project 3: Concordance of microarray and RNA-Seq differential gene expression

Group: Saxophone
TA: Jing Zhang

Sooyoun Lee: Data Curator
Jason Rose: Programmer
Daniel Goldstein: Analyst
Sunny Yang: Biologist

INTRODUCTION

Gene sequencing can be profiled by the traditional microarray or the more advanced and powerful method of RNA seq. RNA-seq can help identify more differentially modulated transcripts of toxicological relevance, splice variants, and non-coding transcripts such as micro RNA (miRNA), long non-coding RNA (lncRNA), pseudogenes. These additional data may be informative for toxicity prediction, mechanistic investigations, or could be the biomarker discovery [1]. In this study, Wang et al were to study and result in the comparative analysis of gene expression responses profiled by Affymetrix microarray and Illumina RNA-seq in the liver tissues from rats that are exposed to diverse chemicals.

DATA

Wang et al (2014) obtained the data from the male Sprague-Dawley rats to generate Illumina RNA-seq and Affymetrix microarray data from the same set of liver samples under varying degrees of perturbation by 27 chemicals representing multiple mode of action (MOA). The study was conducted by dividing the study into two different groups such as a training set and a test set. As a result, three MOAs are associated with well-defined receptor-mediated processes such as peroxisome proliferator-activated receptor alpha (PPARA), orphan nuclear hormone receptor (CAR/PXR), and aryl hydrocarbon receptor (AhR). Two MOAs were found to be non-receptor-mediated such as DNA damage and cytotoxicity [1].

In this project, we focused on three chemicals from toxgroup 4 (Beta-Estradiol, Bezafibrate, and N-nitrosodiethylamine), which represented three MOAs (DNA_Damage, Estrogen Receptor Agonist; ER, and Peroxisome Proliferator; PPARA). The samples we processed were SRR1177984, SRR1177985, SRR1177986, SRR1177960, SRR1178023, SRR1178049, SRR1177967, SRR1177968, and SRR1177971. The symbolic link was created for these nine samples from toxgroup 4 to our group's directory. Then the FastQC was run for each sample to provide quality control checks on raw sequencing data. [2] After running the FastQC, 9 different html and .zip files were created. Then, in order to align each of the samples against the rat genome, the STAR was used. Few details were made such as --outSAM type BAM SortedByCoordinate will output by the bam files and the directory and file names were specified. Also for each file name prefix, a specific directory and the file name were given. Lastly, the multiqc was run to collect all the statistics and information from the fastqc and the STAR alignment.

METHODS

RNASeq Differential Expression Analysis

After annotating the STAR results, MultiQC [3] summary analysis was done to detect any sample bias. Differential gene expression analysis was performed using the DESeq2 package [4] from Bioconductor. These were set up individually against controls with modes of actions as factors for the three chemical treatments: DNA Damage for N-nitrosodiethylamine, ER for beta-estradiol, and PPARA for bezafibrate. Any genes with no counts among them are removed prior to analysis. Normalization is done within the DESeq2 package correcting for composition bias and library size (read depth and gene length). Using the ggplot2 package in R, a histogram of log2 FC against DEGs was created. Additionally, volcano plots were made with the same R package plotting log2 FC against the negative log of nominal p-values.

Microarray Differential Expression Analysis

Differential expression of microarray data was performed for three treatment groups in toxgroup 4 using the limma package of Bioconductor [5]. The full RMA expression matrix was provided in the bf528 project directory on SCC. Microarray data for toxgroup 4 was subset into tables for each treatment group, and the full RMA matrix was subset for the probes used in toxgroup 4. Design matrices for each treatment group were developed to compare gene expression level to the control group. In the limma package, we utilized the function lmFit to fit the subset RMA expression matrix to the design matrix with a linear model. We also utilized the function eBayes to compute statistical analyses of our data with the Empirical Bayes method. Significant differentially expressed genes were reported with an adjusted p-value < 0.05 and log fold change >1.5. Differential expression was visualized with a histogram of log fold change from the significant differentially expressed genes and a volcano plot of log fold change vs. nominal p-value.

Concordance between RNA-Seq and microarray differentially expressed genes

Using differential expression results from DESeq2 and limma analyses, concordance between RNA-Seq and microarray technologies was examined for each treatment group as described by Wang *et al.* [1]. The Affymetrix probe IDs from microarray analysis were mapped to the refSeq IDs from RNA-Seq analysis using the refseq_affy_map.csv file from the bf528 project directory on SCC. Concordance was calculated for genes with an agreement in directionality of fold change; therefore, the mapped data was filtered based on whether the sign of fold change between RNA-Seq and microarray were the same. Concordance was determined using the background-corrected intersection (x) between DESeq2 and limma datasets and the number of differentially expressed genes for each technology. The following equation was used to estimate background-corrected intersection:

$$x = \frac{n_0 * N - n_1 * n_2}{n_0 + N - n_1 - n_2}$$

where is the number of differentially expressed genes in the observed intersection, and are the number of differentially expressed genes in two independent datasets, and is the total number of genes in the rat genome. The total number of genes in Sprague-Dawley rats was estimated to be 22,229 from the Rat Genome Project [6]. From the background-corrected intersection, we computed concordance as:

$$Concordance = \frac{2 * x(n_1, n_2)}{n_1 + n_2}$$

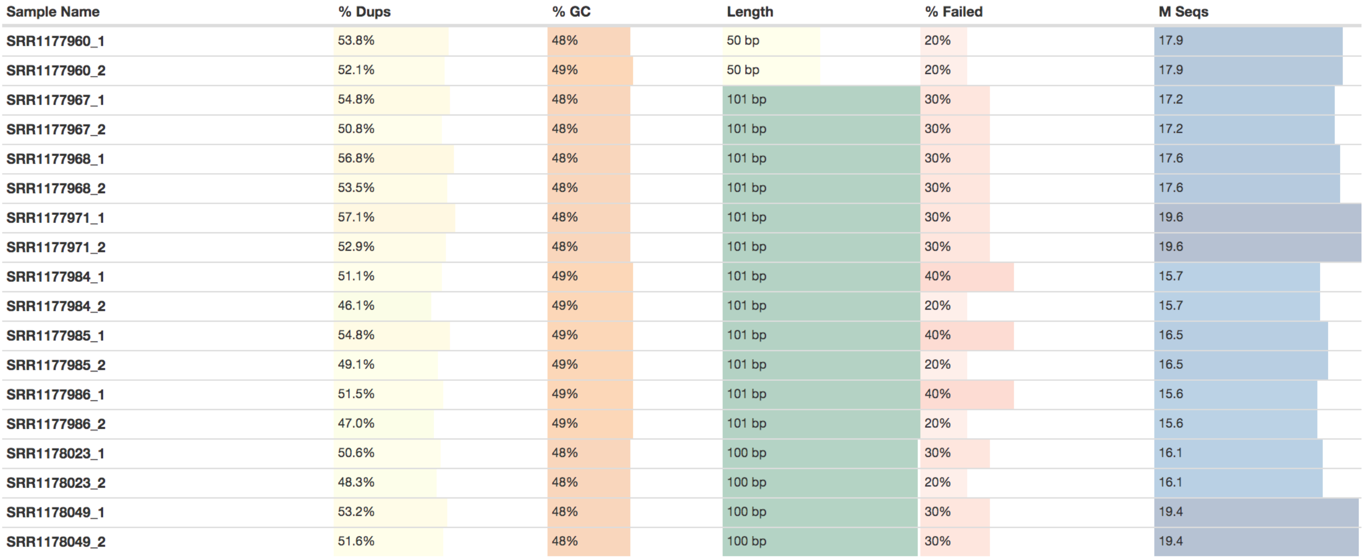
Significant differentially expressed genes were divided into above-median and below-median subsets for each treatment group. These subsets were mapped, filtered for fold change agreement, then an above-median and below-median concordance was calculated. Overall, above-median, and below-median concordance was plotted as a histogram for each treatment group.

DAVID functional annotation clustering

The differential expression analysis of the RNA-Seq data with DESeq2 resulted in three differential expression results and three normalized counts matrices for each treatment. Significantly differentially expressed genes from each treatment was obtained by establishing a p-value cutoff of less than 0.05, and an absolute value log2FoldChange of greater than 1.5. These upregulated and downregulated genes were sent to Database for Annotation, Visualization and Integrated Discovery (DAVID) for pathway enrichment analysis. The gene list was uploaded using GENBANK_ACCESSION gene identifier names. These results were compared to those reported in Wang et al. Supplementary Table 4.

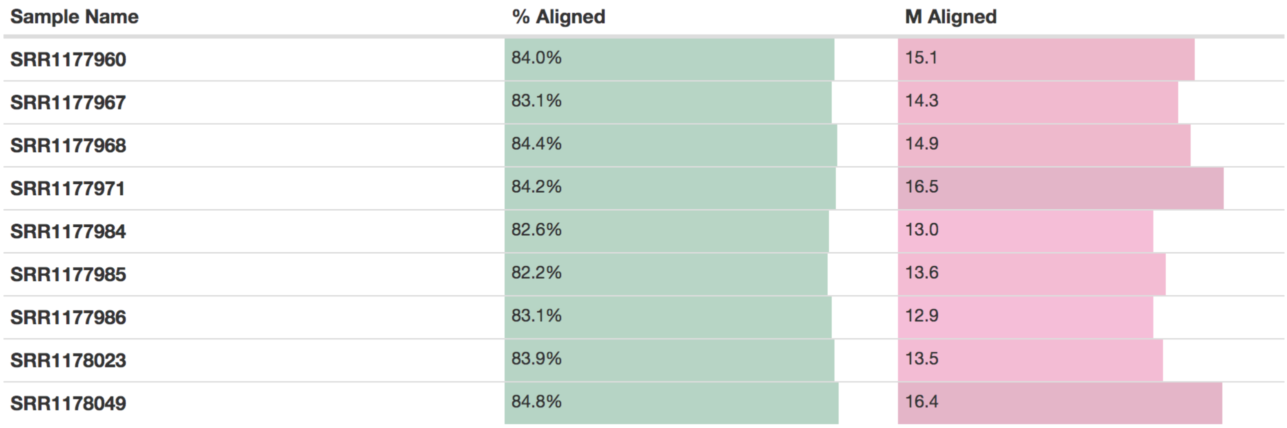
A heatmap-based hierarchical clustering of the RNA-Seq samples was performed to look for segregation by MOA. Each normalized expression matrix had nine columns: four sets for treatment and five sets for control. The heatmap matrix was constructed by merging all three normalized count matrices at the gene-level, and took the union of differentially expressed genes that were significant in each treatment. The final subsetted heatmap matrix was plotted using the heatmap() function in R.

RESULTS



Sample Name	% Dups	% GC	Length	% Failed	M Seqs
SRR1177960_1	53.8%	48%	50 bp	20%	17.9
SRR1177960_2	52.1%	49%	50 bp	20%	17.9
SRR1177967_1	54.8%	48%	101 bp	30%	17.2
SRR1177967_2	50.8%	48%	101 bp	30%	17.2
SRR1177968_1	56.8%	48%	101 bp	30%	17.6
SRR1177968_2	53.5%	48%	101 bp	30%	17.6
SRR1177971_1	57.1%	48%	101 bp	30%	19.6
SRR1177971_2	52.9%	48%	101 bp	30%	19.6
SRR1177984_1	51.1%	49%	101 bp	40%	15.7
SRR1177984_2	46.1%	49%	101 bp	20%	15.7
SRR1177985_1	54.8%	49%	101 bp	40%	16.5
SRR1177985_2	49.1%	49%	101 bp	20%	16.5
SRR1177986_1	51.5%	49%	101 bp	40%	15.6
SRR1177986_2	47.0%	49%	101 bp	20%	15.6
SRR1178023_1	50.6%	48%	100 bp	30%	16.1
SRR1178023_2	48.3%	48%	100 bp	20%	16.1
SRR1178049_1	53.2%	48%	100 bp	30%	19.4
SRR1178049_2	51.6%	48%	100 bp	30%	19.4

Table 1. General Statistics of each read Most of the fastq files showed similar statistics results. Samples SRR1177984 to SRR1177986 represents DNA_Damage, SRR1177960 to SRR1178049 represents ER, and SRR1177967 to SRR1177971 represents PPARA.



Sample Name	% Aligned	M Aligned
SRR1177960	84.0%	15.1
SRR1177967	83.1%	14.3
SRR1177968	84.4%	14.9
SRR1177971	84.2%	16.5
SRR1177984	82.6%	13.0
SRR1177985	82.2%	13.6
SRR1177986	83.1%	12.9
SRR1178023	83.9%	13.5
SRR1178049	84.8%	16.4

Table 2. Summary of MultiQC table Describes the percentage of uniquely mapped reads and the uniquely mapped reads in millions. Most of the STAR files showed similar statistics results.

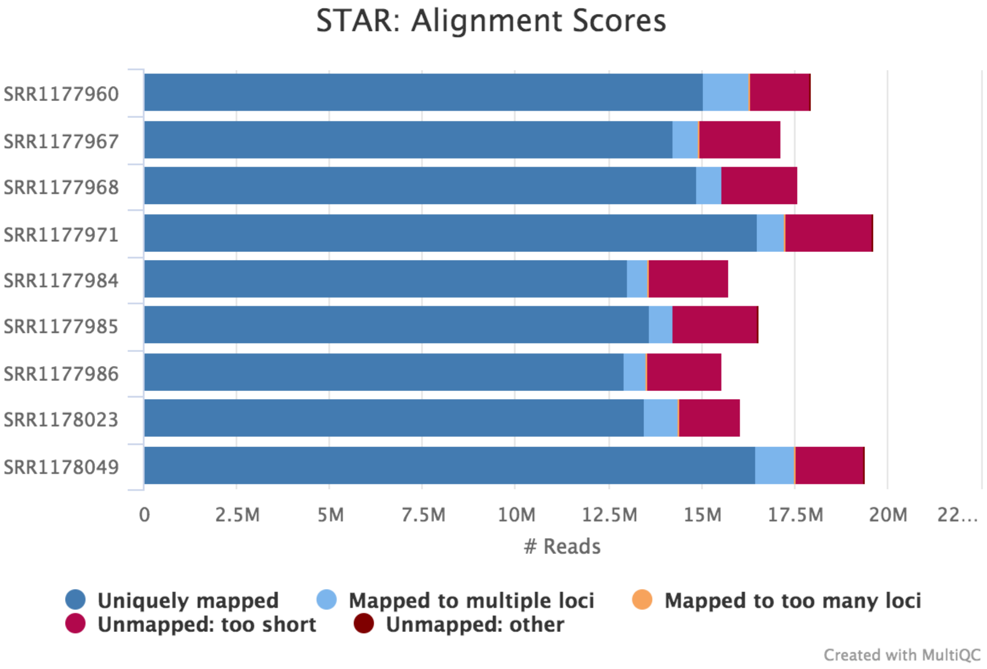


Figure 1. STAR Alignment Scores Showing the number of uniquely mapped reads, multiple mapped reads, and unmapped reads for the 9 samples.

The 82%-84% of samples were uniquely mapped to the genome which is a highly acceptable percentage of quality. 3%-6% of the reads were mapped to multiple loci and 9%-13% of the reads unmapped.

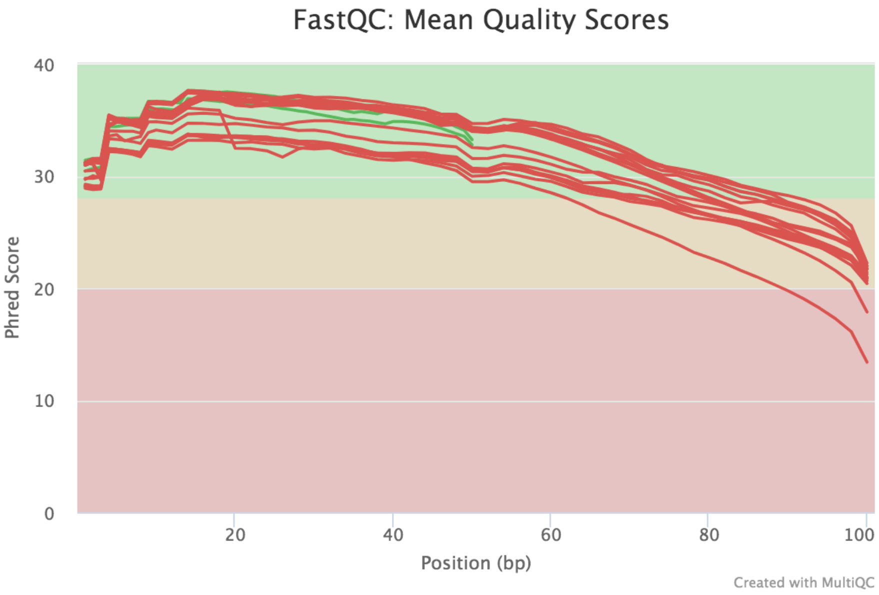


Figure 2. FastQC Mean Quality Scores from MultiQC output The plot showed the mean quality value across each base position in the read. Out of 18 samples, only two samples passed the quality scores which are SRR1177960_1 and SRR1177960_2 and the rest of the samples have failed. The sample SRR 1178023_2 had the lowest phred score.

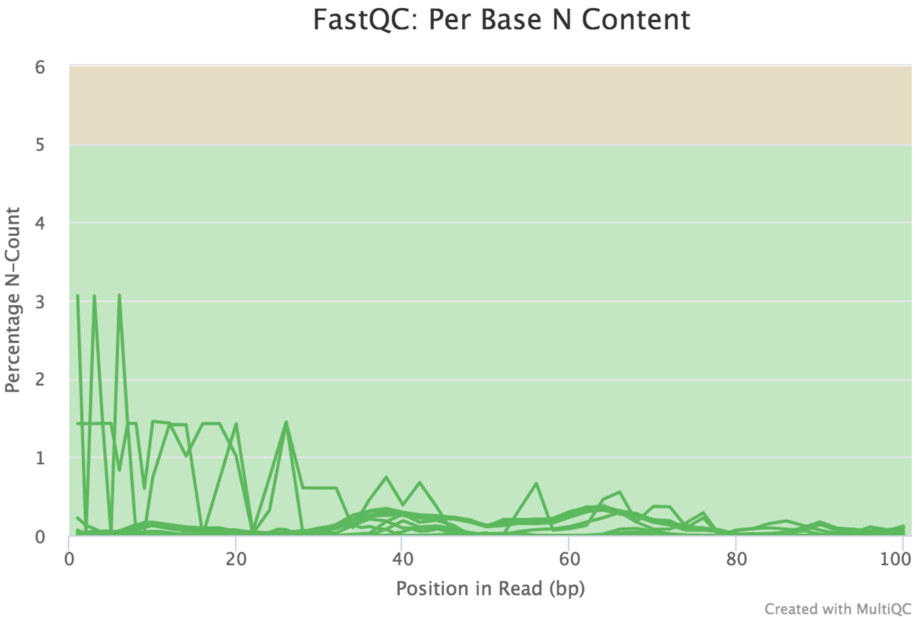


Figure 3. FastQC Per Base N Content Showing the Position in Read(bp) on the x-axis and the percentage N-count on the y-axis

In Figure 3, most of the N content depicts lower than 4% but also any point where the curve rises noticeably above zero. This figure represents that all of the 18 fastq files passed the test and shows different percentages of 'N's in their positions.

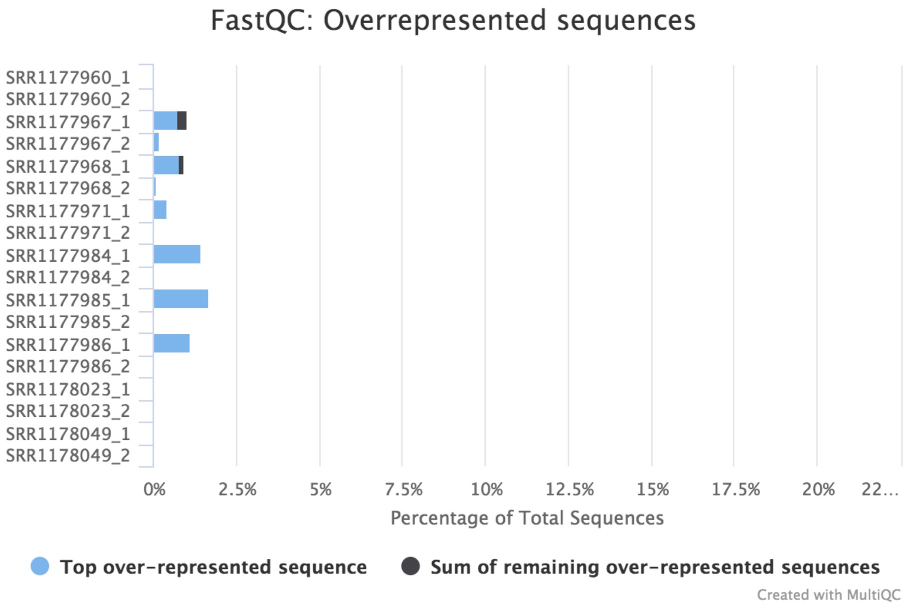


Figure 4. FastQC Overrepresented Sequences Showing the percentage of total sequences overrepresented on the x-axis and the samples on the y-axis

Figure 4 shows the percentage of the top overrepresented sequences and the sum of the remaining overrepresented sequences. Among the 18 fastq files, three samples failed to test because sequences are represented over 1% and those samples are SRR 1177984_1, SRR1177985_1, and SRR1177986_1.

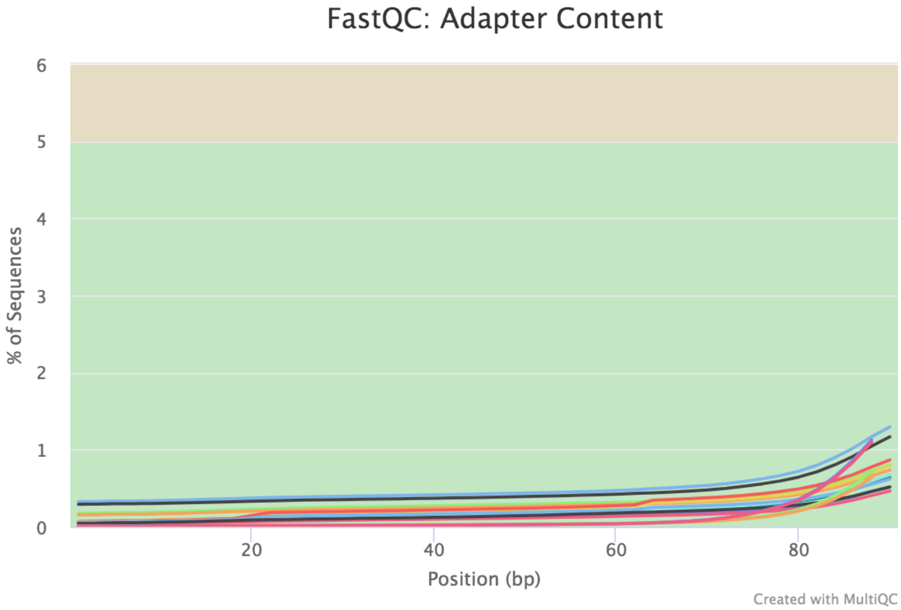


Figure 5. FastQC Adapter Content The position (bp) on the x-axis and the percentage of the sequences on the y-axis.

The adapter sequences at each position are represented. All 18 fastq file samples passed the test and the highest adapter samples were SRR1177967_1 with 1.29%. Overall, there were no other samples that went beyond 2% of sequences which is showing that all the fastq files are of good quality.

Annotated feature counts and RNA-seq results

Prior to running DESeq, MultiQC was run on the annotated STAR results to detect bias and trends among the samples (figure 6). The report shows a high degree of similarity among the samples. The samples made of three groups can be visually distinguished, such as SRR1177967, SRR1177968, and SRR1177971. They share a narrow range of assigned genes just under 60% and are for the bezafibrate chemical group. The N-nitrosodiethylamine group (SRR1177984, SRR1177985, and SRR1177986) have a similar range just above 60%. The final group for beta-estradiol (SRR1177960, SRR1178023, and SRR1178049) have the lowest assigned percentage of around 53%.

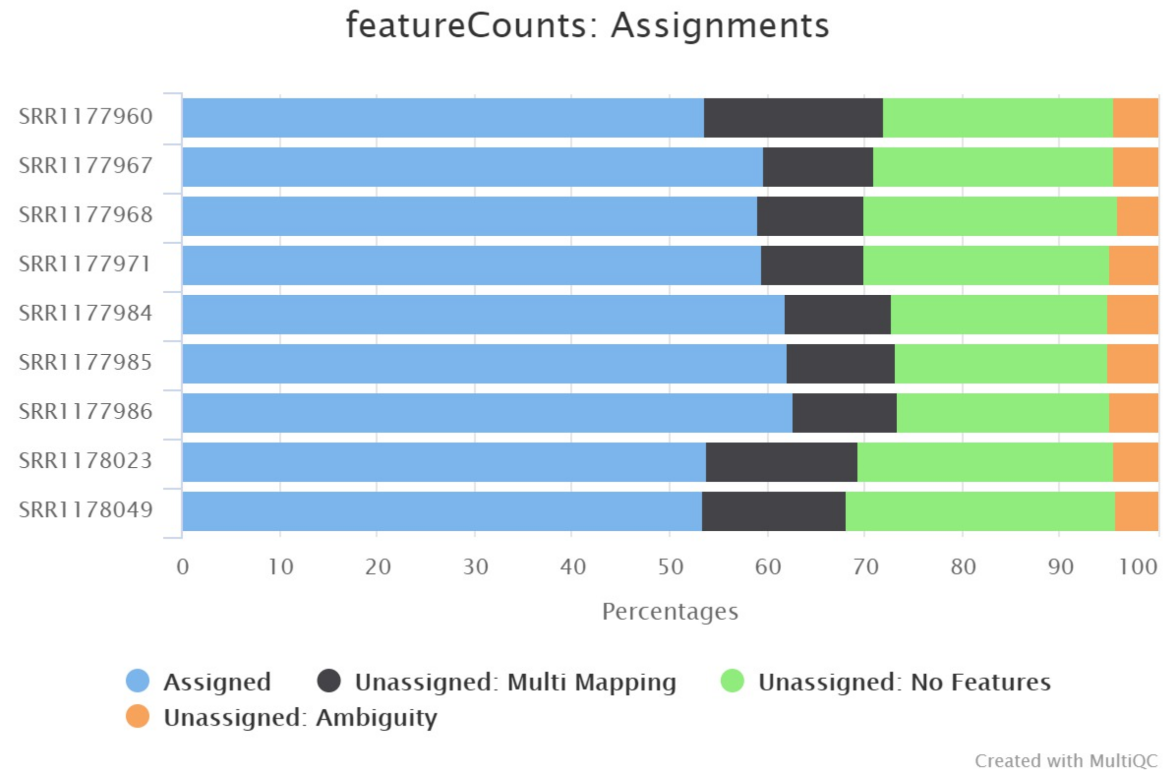


Figure 6. featureCounts Assignments for RNA-seq annotated data MultiQC report of each annotated sample feature counts file from STAR. Shows the similarity of genes assigned by percent with three distinct groupings among them that correlate to the mode of action used.

The visualization of the DESeq results shows similarity among modes of action, but more coverage of expression in the DNA damage mode (figure 7) than the other two. The volcano plots (A, B, and C) give insight into the density of DEGs (differentially expressed genes) among samples with a much higher density exhibited in the DNA. The histograms (D, E, and F) provide a look into the expression levels they capture in their group. The DNA mode of action group (D) exhibits more expression than the other two modes, which show very similar levels of expression in both breadth of coverage.

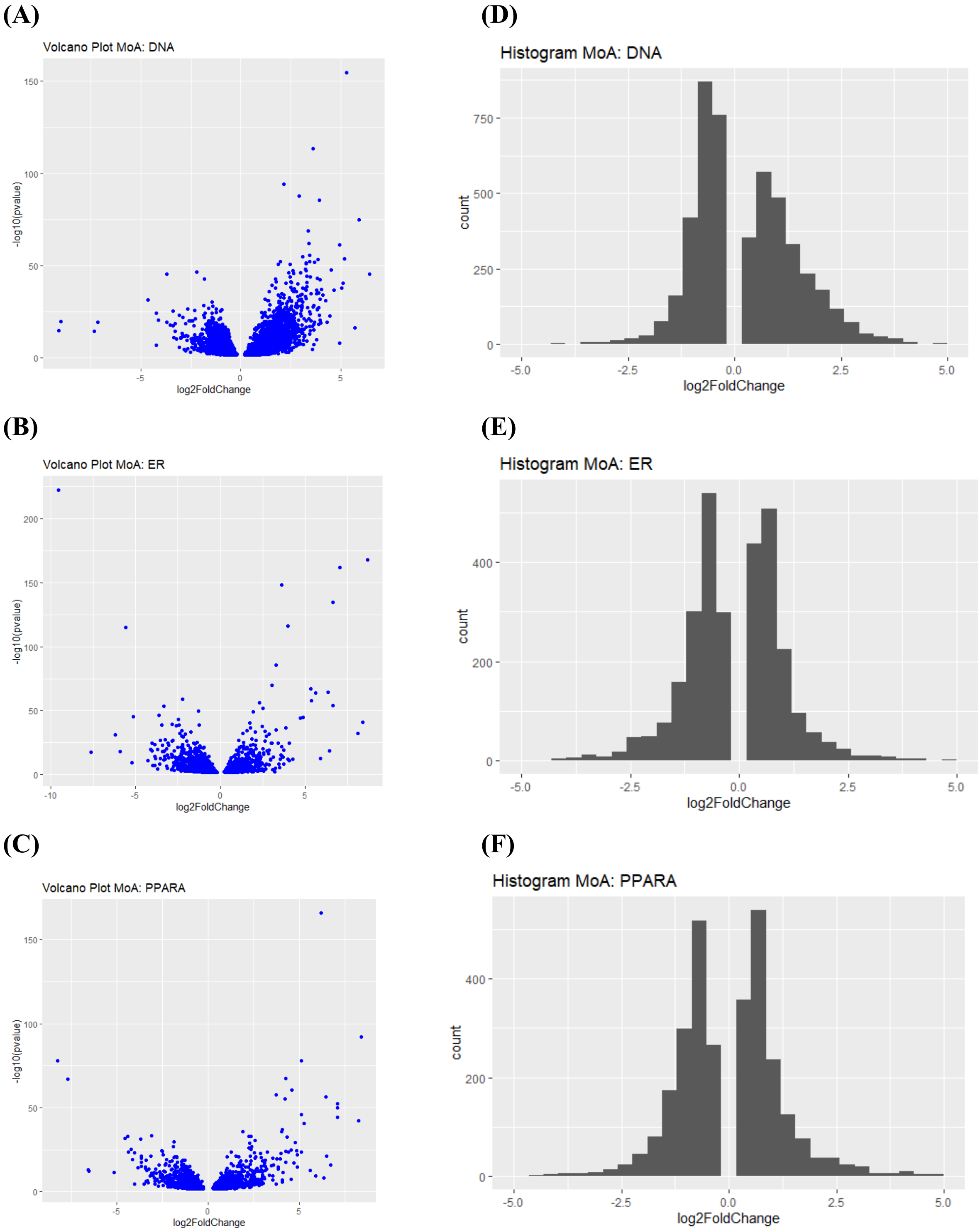
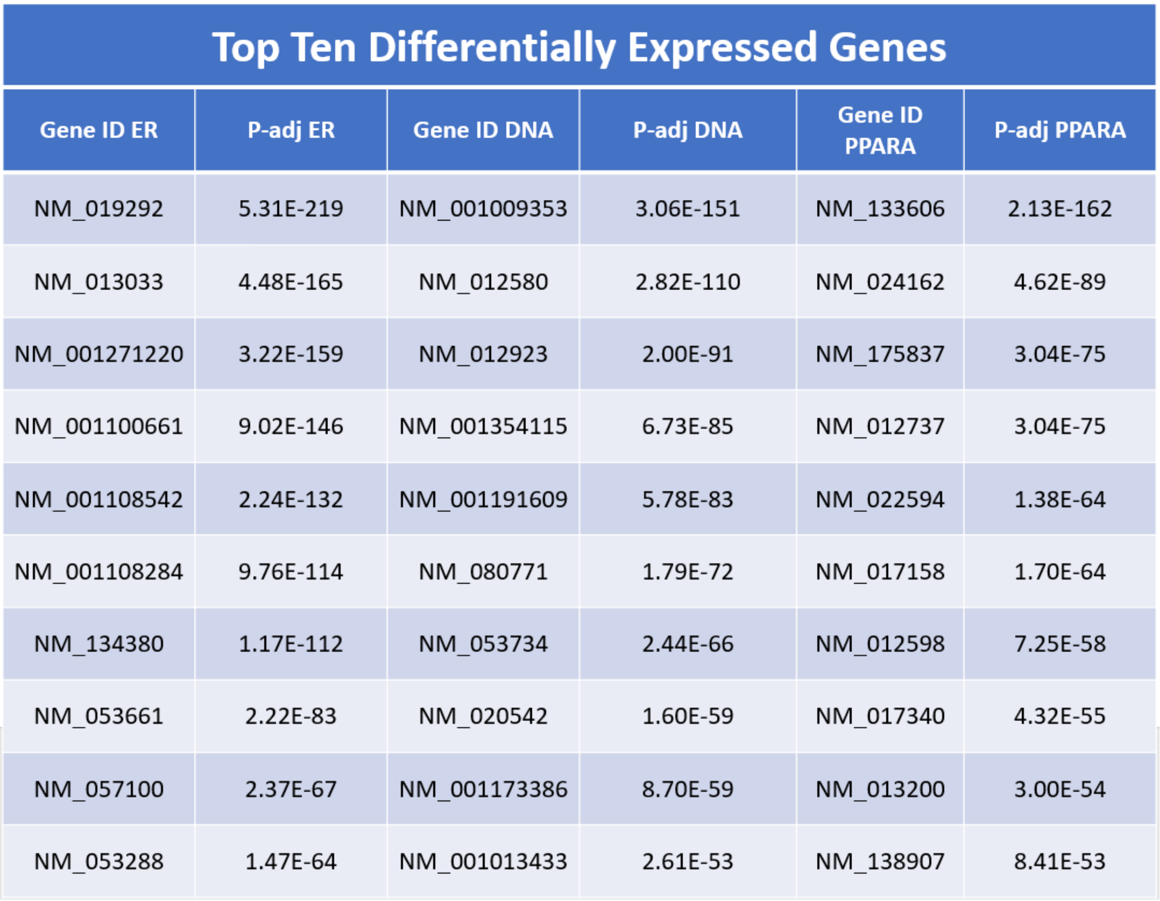


Figure 7. Volcano Plots and Histograms of RNA-seq Results Volcano plots A (DNA damage), B (ER), and C (PPARA) of DESeq2 results based on the log2 FC against -log of the nominal p-value. Histograms E (DNA damage), F (ER), and G (PPARA)

The top DEGs among each mode can be seen in table 1, which contain no shared genes in expression. The number of significant genes with an adjusted p-value less than 0.05 are greater in the DNA damage mode than others at 4816. ER contains 2962 significant DEGs and PPARA contains 2952.



Top Ten Differentially Expressed Genes

Gene ID ER	P-adj ER	Gene ID DNA	P-adj DNA	Gene ID PPARA	P-adj PPARA
NM_019292	5.31E-219	NM_001009353	3.06E-151	NM_133606	2.13E-162
NM_013033	4.48E-165	NM_012580	2.82E-110	NM_024162	4.62E-89
NM_001271220	3.22E-159	NM_012923	2.00E-91	NM_175837	3.04E-75
NM_001100661	9.02E-146	NM_001354115	6.73E-85	NM_012737	3.04E-75
NM_001108542	2.24E-132	NM_001191609	5.78E-83	NM_022594	1.38E-64
NM_001108284	9.76E-114	NM_080771	1.79E-72	NM_017158	1.70E-64
NM_134380	1.17E-112	NM_053734	2.44E-66	NM_012598	7.25E-58
NM_053661	2.22E-83	NM_020542	1.60E-59	NM_017340	4.32E-55
NM_057100	2.37E-67	NM_001173386	8.70E-59	NM_013200	3.00E-54
NM_053288	1.47E-64	NM_001013433	2.61E-53	NM_138907	8.41E-53

Table 1. Top Ten Differentially Expressed Genes from RNA-seq Based on mode of actions for each group. ER for beta-estradiol, DNA damage for N-nitrosodiethylamine, and PPARA for bezafibrate.

Microarray differential expression analysis using LIMMA

From differential expression analysis of microarray data, there was a range in the number of DE genes across the three treatment groups. We found the total number of significant differentially expressed genes for beta-estradiol, bezafibrate, and N-nitrosodiethylamine was 37 genes, 115 genes, and 36 genes, respectively. The top ten differentially expressed genes sorted by adjusted p-value for each treatment group are reported in Table 2. The number of significant upregulated and downregulated genes can be visualized as a frequency distribution and a volcano plot in Figure 8.

Top Ten Differentially Expressed Genes for each MOA								
Beta-Estradiol (MOA: ER)			Bezafibrate (MOA: PPARA)			N-nitrosodimethylamine (MOA: DNA damage)		
Gene	logFC	p_{adj}	Gene	logFC	p_{adj}	Gene	logFC	p_{adj}
Rgs3	2.15	5.91E-05	Cyp4b1	2.63	4.04E-26	RGD1561849	3.28	6.63E-09
Rbp7	3.20	4.94E-04	Acot1	8.43	4.04E-26	Nhej1	2.07	7.56E-07
Gsta3	2.27	9.91E-04	Acot2	2.95	3.64E-24	Rbp7	1.81	2.39E-06
Myc	1.54	1.02E-03	Acot3	4.45	9.52E-23	Nrcam	1.91	4.93E-06
Tsku	1.63	1.12E-03	Pex11a	2.93	1.23E-20	Mdm2	1.66	2.78E-05
Spink3	-1.64	1.18E-03	Acox1	1.79	3.12E-20	Gria3	2.36	1.29E-04
Bcl6	-2.02	1.18E-03	Ehhadh	3.15	9.71E-20	Slc20a1	1.57	2.63E-04
Gdf15	1.90	2.41E-03	Crat	2.61	2.58E-18	Ces2c	2.02	4.65E-04
Abcg8	-1.56	3.16E-03	Eci1	2.60	4.16E-18	Mgmt	1.68	5.50E-04
Ctr9	1.83	3.28E-03	Cpt2	1.58	7.56E-18	Ccng1	2.11	7.21E-04

Table 2. Limma Differential Expression Analysis. The top ten differentially expressed genes for three treatment groups: Beta-Estradiol, Bezafibrate, N-nitrosodimethylamine as determined by limma analysis of microarray data. Genes were considered differentially expressed with the absolute value of log fold change > 1.5.

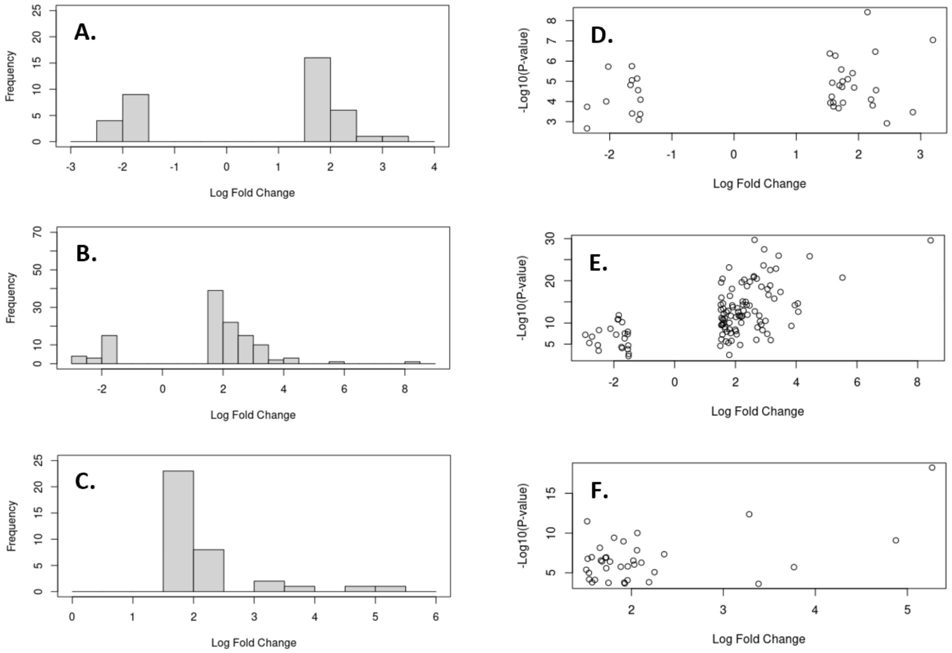


Figure 8. Histogram and Volcano Plot of Fold Change. Frequency distribution of log fold change for: (A) Beta-Estradiol, (B) Bezafibrate, (C) N-nitrosodimethylamine, with their associated volcano plots of log fold change vs. nominal p-value (D-F).

Concordance between RNA-Seq and microarray differentially expressed genes

Concordance of differentially expressed genes varied among ER, PPARA, and DNA damage modes of action; these values were computed as 0.107, 0.250, and 0.018, respectively. As shown in Figure 9, the relationship between concordance and number of differentially expressed genes was significantly different for each sequencing technology. For microarray technology, there is a positive linear correlation between concordance and number of differentially expressed genes. However, it is difficult to discern a relationship between concordance and number of differentially expressed genes from RNASeq with only three treatment groups.

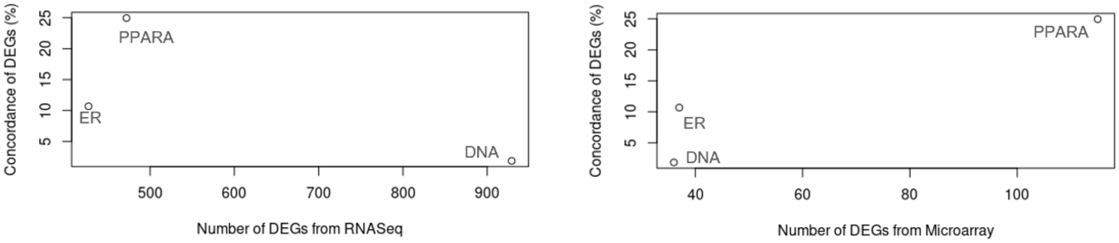


Figure 9. Scatterplot of Number of Differentially Expressed Genes vs. Concordance. Concordance of each treatment group, representing the ER, PPARA, and DNA modes of action plotted against the number of differentially expressed genes computed from RNASeq (left) and microarray (right).

Differential gene expression data obtained from DESeq2 and limma analyses was subset into above-median and below-median differentially expressed genes based on their mean expression levels to compare how concordance of these subsets differs from overall concordance. The above-median subsets of beta-estradiol, bezafibrate, and N-nitrosodiethylamine consisted of 10 genes, 47 genes, and 5 genes. For these differentially expressed genes, the respective concordances were computed as 0.0859, 0.320, and 0.0196. On the other hand, there was only 1 gene within the below-median subset for beta-estradiol, only 2 genes for N-nitrosodiethylamine, and only 4 genes for the bezafibrate group. The respective below-median concordances were computed as 0.00724, 0.00689, and 0.0235. Generally, above-median concordance was greater than overall concordance for bezafibrate and N-nitrosodiethylamine treatment groups, while above-median concordance was less than, but comparable to overall concordance for the beta-estradiol treatment group (Figure 10). For all below-median subsets, concordance was significantly lower than the overall and above-median values (Figure 10).

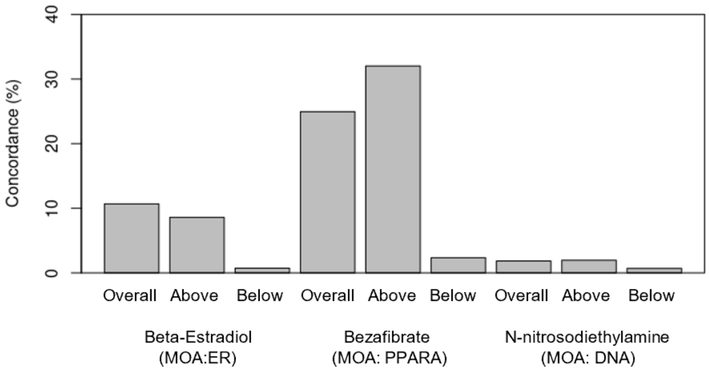


Figure 10. Bar plot of overall, above-median, and below-median concordance of treatment groups. Overall, above-median, and below-median concordance compared among the treatment groups.

Biological Interpretation

Enrichment Analysis

Enriched Pathway	**P-value**
PPARA - Bezafibrate	
Fatty acid beta-oxidation*	2.5E-12
Peroxisome	3.2E-10
Retinol metabolism	2.1E-12
Steroid hormone biosynthesis	1.4E-10
Arachidonic acid metabolism	1.3E-4
Linoleic acid metabolism	3.2E-2
Serotonergic synapse	7.2E-1
Ascorbate and aldarate metabolism	2.4E-9
Chemical carcinogenesis	7.8E-9
Metabolism of xenobiotics by cytochrome P450	1.2E-8
Drug metabolism - cytochrome P450	6.6E-6
Porphyrin and chlorophyll metabolism	1.8E-4

ER - Beta-estradiol (test set)	
Retinol metabolism	5.2E-9
Steroid hormone biosynthesis	3.1E-8
Linoleic acid metabolism	1.1E-4
Chemical carcinogenesis	8.1E-4
Arachidonic acid metabolism	1.6E-3
DNA damage - NIT	
Xenobiotic metabolism process	1.7E-1
Transcription regulation	5.1E-9
Rheumatoid arthritis	6.1E-6
Steroid biosynthesis	1.2E-4
Chemokine signaling pathway	5.2E-9
Regulation of cell cycle*	3.6E-1

Table 3: Enriched pathways and corresponding p-values identified using DE genes from each treatment analyses. Starred pathways are those similar to Wang et al, Highlighted in bold are exact matches to Wang et al.

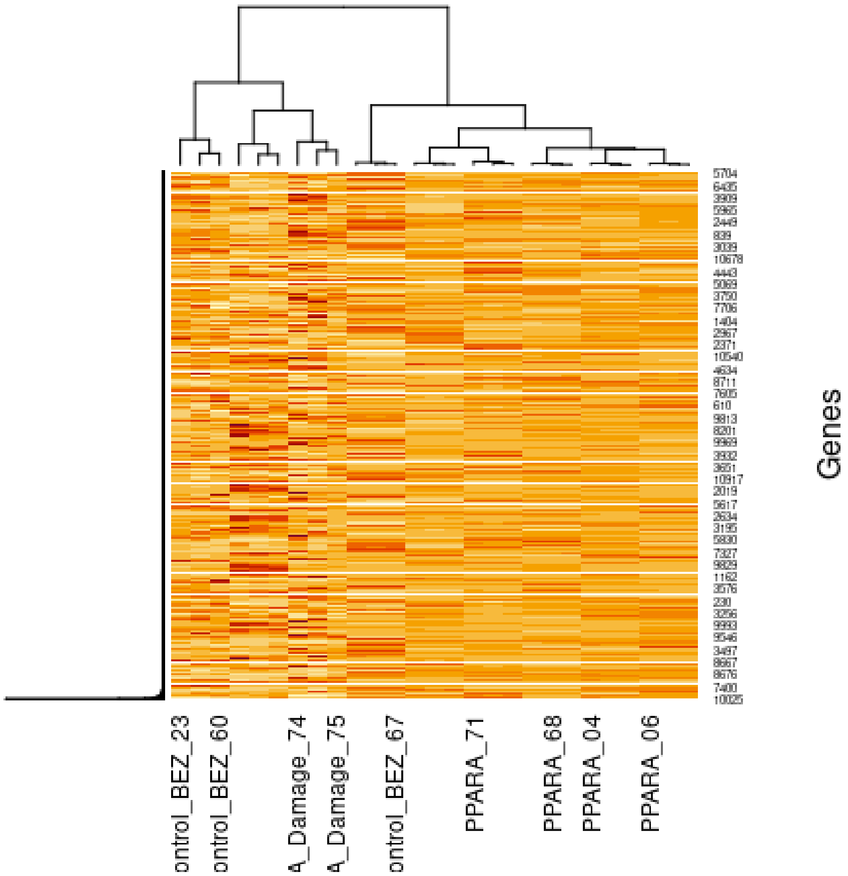


Figure 8. Heatmap-based hierarchical clustering of 12 RNA-Seq samples: ER(4), PPARA (4), DNA damage (4) based on DESeq2 normalized counts matrices. Samples from each MOA clustered together.

DISCUSSION

Quality Control Statistics for the RNA-seq Samples

Nine samples of raw RNA-seq data are processed by fastqc, STAR, and multiqc tools. Table 1 shows the general statistics of each read. It specifically shows the percentage of duplicate reads, GC percentage, length of the reads, percentage of modules failed, and total sequences in millions are represented. There were some commonalities found between the percentage of duplicate reads and the GC, and the total sequences in millions, that all samples were showing very similar result ranges. The only significant difference is shown in the length of the reads for samples SRR1177960_1 and SRR1177960_2. However, the percentage of modules that failed for these samples was 20 percent which is lower than the other samples and we can estimate that the length of the reads does not affect the results. Also by using the multiqc, table 2 was obtained which shows the percentage of uniquely mapped reads, and the uniquely mapped reads in millions for 9 samples.

All of the alignments are ranging from 82%-84% which is showing a high percentage of aligned reads. 9 samples are ranging from 12-16 uniquely mapped reads in millions which are tightly clustered. In the quality scores across the 9 samples (figure 2), among the 18 fastq files, only 2 of the samples passed the quality test and 16 samples have failed. This represents that most of the samples are steady and showing good quality until the 50bp position however starting from the 60bp the graph showed a decline and dropped below the phred score of 20 at the end of the reads.

The reason for the quality loss could be due to the illumina sequencing machine that the laser might have become less functional by the end of the position. However, since the general statistics of each reading and summary of the MultiQC table seem to show no major distribution differences, the quality control of the RNA-seq samples is in a good state.

After annotation of the STAR results, MultiQC was run again to affirm absence of bias within samples and to catch early trends. Right off we are able to see similarities of assignments within the group, but can see the highest is for the chemical associated with the DNA damage mode. This carries through some other visualization of the data for RNA-seq in the volcano plots and histograms from figure 7. There is a higher density of DEGs above the significance threshold in the DNA damage mode than the other two, while showing a similar pattern of results. In the histograms for DNA damage, there are higher levels of expression over the entirety of log2 FC exhibiting better coverage. There are nearly 50% more DEGs for DNA damage than either of the other modes based on adjusted p-value of less than 0.05.

Among the most significant DEGs from each mode (table 1), none were shared in the top 10 (because of mode applied), but the p-values are highest for the ER mode, with DNA and PPARA sharing similar levels in their top 10. This implies that there is higher confidence in that mode.

Microarray differential expression analysis

There was at least one order of magnitude fewer number of significantly differentially expressed genes reported in this project compared to *Wang et al.*, which had a negative influence on our "downstream" analysis of concordance with RNASeq differential expression. In one instance, the reported number of differentially expressed genes for the N-nitrosodiethylamine treatment group was 36, while the reported number of differentially expressed genes from the original study is on the order of 3,500 [1].

Several of the top differentially expressed genes were correlated with processes involved in the mode of action of the treatment groups. The bezafibrate treatment group represented the mode of action for peroxisome proliferator-activated receptor alpha (PPARA); therefore, the top differentially expressed genes are associated with perioxisome maintenance and proliferation (Pex11a), acetyl-CoA production for energy (Acot1, Acot2, Acot3), and drug metabolism (Cyp4b1). Similarly, the beta-estradiol group, which represented the estrogen receptor (ER) mode of action, demonstrated differential expression of genes related to signaling regulators (Rgs3), transporter activity (Rbp7), and drug metabolism (Gsta3).

Concordance of RNASeq and microarray data analysis

As previously stated, the values of concordance obtained from this project did not agree with the values reported in *Wang et al.*, which might have been attributed to the discrepancy between number of differentially expressed genes in the DESeq2 and limma analyses. The effect of this discrepancy could be observed on left plot in Figure 9, where there is no observable correlation between concordance of the three treatment groups and the number of differentially expressed genes from RNASeq. However, on the right plot in Figure 9, there could be a linear correlation between these treatment groups. Another explanation for the lack of a correlation between these two parameters is that for this project, we are only looking at a single toxgroup with three treatment groups. In the future, we could explore concordance with more than one toxgroup to more accurately represent the relationship between concordance and number of genes as shown in *Wang et al.* [1].

Due to the lack of correlation between concordance and number of genes, it was difficult to find a trend in the relationship between overall, above-median, and below-median concordance. In two of the treatment groups, there was a greater concordance in the above-median subset of genes and a significantly lesser concordance in the below-median subset of genes, when compared to overall concordance. The number of genes in the below-median subset ranged from 1-4 among treatment groups, which likely skewed the concordance value.

Biological Analysis

After establishing a p-value and log2FoldChange cutoffs, there were 926 differentially expressed genes for ER, 472 genes for PPARA, and 427 genes for DNA damage. These genes were used for DAVID functional annotation analysis of each treatment group. As seen in Table 2, some of the enriched pathways that were identified by our DE genes from each analysis are in agreement with the processes identified in Wang et al. Pathways such as fatty acid beta-oxidation in PPARA and xenobiotic metabolism, which is a common enriched pathway for DNA damage, all matched to those listed in Supplementary Table 4 in Wang et al. However, many pathways, including those for the ER(Beta-estradiol) group were not mentioned in the paper. This could be due to the fact that Wang et al. examined 27 chemicals in total and five MOAs while our analysis was based on three chemicals from three MOAs, which could potentially explain the difference that was seen between the pathways that were enriched in our analysis versus those in the paper.

In constructing the heatmap, we successfully illustrated that samples from the same treatment and MOA clustered together. Many approaches were possible when filtering the data, including using the highest positive log fold change. According to the clustered heatmap, samples from ER seem to have a lot more significant genes as compared to the other samples. This is correct since it had the most number of differentially expressed genes from the DESeq2 analysis of RNA-seq data.

CONCLUSION

In this project, our goal was to reproduce the results from Wang *et al.* to develop a method for comparing sequencing analysis between orthogonal technologies, specifically RNASeq and microarray. The results of our differential expression analysis were compared to the researcher's findings with varying degrees of success. From our microarray differential expression analysis, we found a fewer number of differentially expressed genes in comparison to the number reported in *Wang et al.*, which we believe influenced the downstream analysis of concordance between RNASeq and microarray technologies. Though the pathway enrichment results of our DE genes did not match exactly to those reported in the paper, there was still an agreement among the ones that we found through DAVID analysis and the paper's results. The heatmap provided additional visualization of the clustering among MOAs and confirmed that samples from each MOA cluster together.

In the future, we could explore multiple toxgroups to obtain a better understanding of the correlation between concordance and the number of differentially expressed genes.

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

[2]*FastQC*,www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/fastqc.html.

[3]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.

[4]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

[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), e47.

[6]Rat Genome Sequencing Project Consortium., (2004). DNA sequencing: Baylor College of Medicine., Gibbs, R. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521.