Project 5: Concordance of microarray and RNA-Seq differential gene expression

Data Curator and Programmer: Poorva Juneja

Introduction

While RNA-Seq and microarray analysis are both promising, there have been conflicting results regarding the strengths and weaknesses of each, especially in regards to detecting differentially expressed genes.

In the 2014 study by Charles Wang et al. [1] RNA-Seq and microarray technology were compared to assess their performance in identifying differentially expressed genes across three treatment groups with known mechanisms of action by investigating RNA samples from rat livers with 27 conditions. In order to perform the analyses, limma, edgeR and DESeq2 robust multi-array average (RMA), and MAS5 were used. While limma was originally created for microarray analysis, it has been proven useful to analyze many different forms of data including RNA-seq, which is how it was utilized by Wang et al. Both DESeq2 and edgeR are more common RNA-seq analysis tools and both functions to perform normalization in different ways. RMA is one of the most common tools used to normalize and summarize microarray data. The combination of RMA and MAS5 provides two forms of normalization and expression data from the microarray sample data.

Here, in this analysis I have tried to reproduce a portion of Wang et al.'s study with RNA-seq analysis using a similar approach.

Data

In the 2014 study of Charles Wang et al., data were generated from male Sprague-Dawley rats, ages 6-8 weeks, who had been administered test chemicals mixed with corn oil or water either orally or via injection. The RNA samples were obtained from the NTP DrugMatrix Frozen Tissue Library and were categorized into training and test groups.

For each test chemical there were three rats treated and for each mode of action there were three chemicals. Chemicals were administered over the course of 3-7 days and livers were harvested within 24 hours of the last injection for analysis. Illumina RNA-seq (performed on HiScanSQ or HiSeq2000) and Affymetrix microarray (performed on Affymetrix whole genome GeneChip® Rat Genome 230 2.0 Array) data was generated from these liver samples.

For the purpose of this project, a subset of samples was selected for analysis belonging to toxgroup 4. All the microarray and sequencing datasets were made available on BU's Shared Compute Cluster (SCC) which was sourced from NCBI's Sequence Read Archive (accession number SRP039021) and Genome Expression Omnibus (accession numbers GSE55347 and GSE47875).

This included 9 treatment samples (SRR1177984, SRR1177985, SRR1177986, SRR1177960, SRR1178023, SRR1178049, SRR1177967, SRR1177968, SRR1177971) and 6 control (SRR1178004, SRR1178006, SRR1178013, SRR1178064, SRR1178074, SRR1178075) on which differential analysis was performed with mode of action (MOA) involving DNA damage, ER (Estrogen Receptor-mediated cytotoxicity), and PPARA with specific chemical treatments.

Method

Alignment of each sample against Rat genome using STAR

Firstly the 9 treatment FASTQ sample files from toxgroup 4 were assessed for quality control of using FastQC, then reads were then aligned against the rat genome which was provided on the SCC. For alignment STAR aligner (v2.6.0c) [2] was used, it takes in two reads to be aligned together for each sample and save the alignment output as bam files. STAR also generates read alignment statistics.
Lastly, MultiQC (v1.10.1) was run on the results from both FastQC and STAR to generate a summary report containing quality control metrics and alignment statistics for all the samples. MultiQC checks all samples at once and synthesizes the results into a single report to identify global trends and biases. Both star and multiqc commands were run on scc1 command line as qsub scripts.

Read counting using Feature counts

With the output from the previous step the STAR-aligned BAM files count matrices for each of the 9 samples, spanning three different treatment groups, were generated using the featureCounts program from the subread module(v1.6.2) using rn4_refGene_20180308.gtf as the annotated reference which was provided on SCC. Subread is a general-purpose read aligner and featureCounts [3] counts the reads to genomic features such as genes, exons, promoters, and genomic bins. MultiQC was run again in the directory with the counts files for a summary report.

Then, we combined the all the feature count from each sample into a single CSV file and plotted a boxplots showing the distribution of counts in R using packages tidyverse and janitor.

RNA-Seq Differential Expression with DESeq2

For detection of differentially expressed genes within our sample for 3 different modes of action, DESeq2 an R package within Bioconductor was used. It makes use of negative binomial generalized linear models in the detection of differentially expressed genes. First the control and treatment counts were all combined into one tibble. Then all sample id were matched to the respective mode of action. Then all with all 3 MOA (DNA damage, ER and PPRARA) were subset and DESeq2 was ran on reach separately.
This outputs a data frame, consisting of the gene name, nominal and adjusted p-values, and log fold change for each of the conditions. Genes were determined to be differentially expressed if adjusted p-value was less than 0.05. This was then exported to CSV for each of the three MOA groups.
Histograms of fold change values and scatter plots of fold change vs nominal-pvalue from the significant DE genes was then plotted.

Results

After aligning the reads from the 9 treatment samples, STAR outputted the alignment statistics for each sample run (Table 1). The summary also reported 0 % too many mismatched Unmapped read. Only Sample SRR1177985 had slightly higher % section of unmapped reads. From the percentage of uniquely aligned and multi-mapped reads they indicated to be in a good alignment and within reasonable range to proceed for further analysis.

Sample	Total number of reads	Uniquely mapped reads (with %)	Mismatch Rate per Base (%)	Reads mapped to Multiple Loci (with %)	Reads mapped to *too many loci (with %)	% of Unmapped Reads (too short)
SRR1177960	17947778	15080805 (84.03%)	0.45%	1189810 (6.63%)	42948 (0.24%)	8.96%
SRR1177967	17157413	14253911 (83.08%)	0.77%	673010 (3.92%)	31242 (0.18%)	12.78%
SRR1177968	17604520	14857439 (84.40%)	0.77%	686056 (3.90%)	29288 (0.17%)	11.48%
SRR1177971	19627402	16518020 (84.16%)	0.83%	732186 (3.73%)	51903 (0.26%)	11.78%
SRR1177984	15732068	12999695 (82.63%)	0.77%	586314 (3.73%)	29840 (0.19%)	13.41%
SRR1177985	16537337	13601134 (82.25%)	0.69%	630871 (3.81%)	24455 (0.15%)	13.75%
SRR1177986	15559784	12936410 (83.14%)	0.72%	600101 (3.86%)	21894 (0.14%)	12.82%
SRR1178023	16076957	13486519 (83.89%)	1.13%	915705 (5.70%)	18535 (0.12%)	10.25%
SRR1178049	19397953	16448173 (84.79%)	0.92%	1041082 (5.37%)	49116 (0.25%)	9.53%

*Table1: STAR read and alignment summary statistics. *Mapped reads with too many loci indicate mapping with more than 10*

To process the results after the STAR alignment, MultiQC html provided with relevant figures. Figure 1 represents the STAR Alignment scores which indicates SRR1178049 has the highest number of uniquely mapped reads, followed by SRR1177971. From the Figure 2 in the bottom left figure which represents the number of reads/average quality scores indicates a phred score greater than 28 which is the threshold for good quality. In Figure 2 top left which represents the mean quality value across each base position in the read has around base pair position 70 and SRR1178023_2 has the lowest mean quality score as the phred score drops below 28 after base pair position 60.

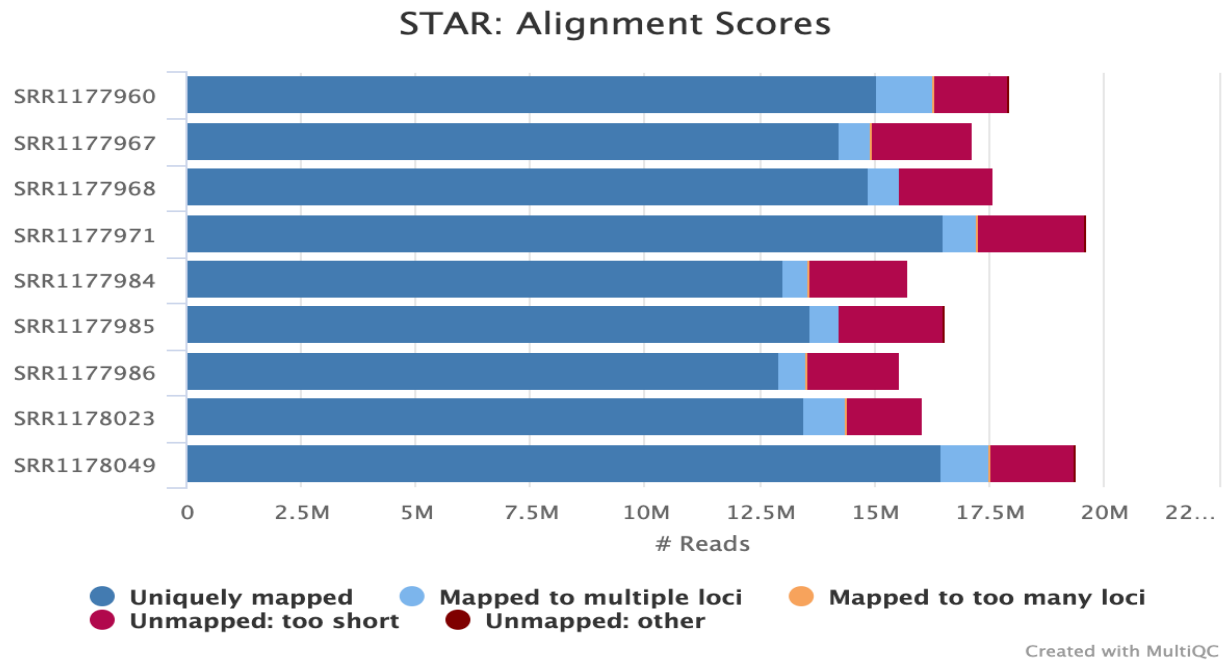

Figure 2:MultiQC: STAR alignment statistics

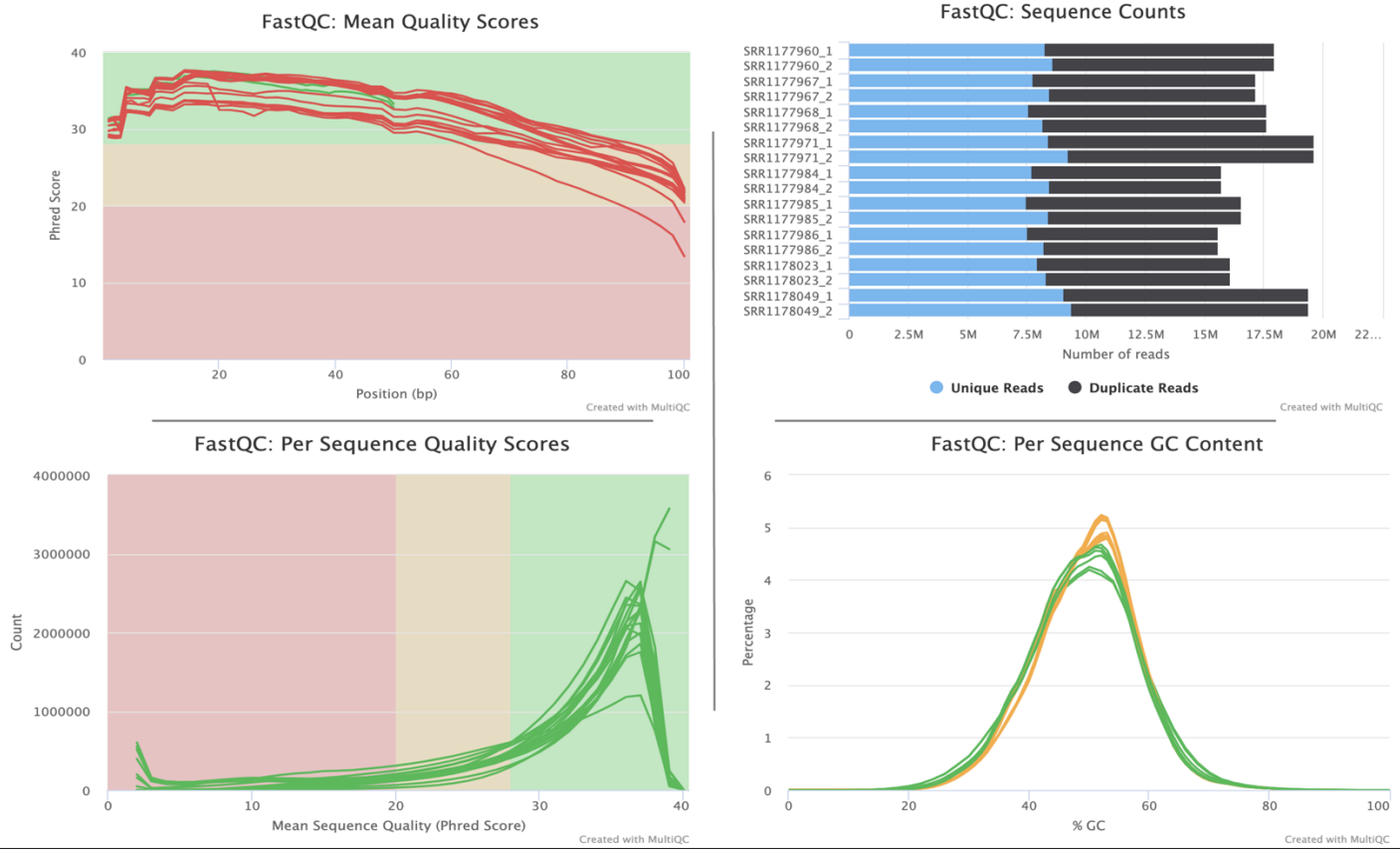

Figure 2 : MultiQC: FASTAQC summary statistics.

Sample Name	% Aligned	M Aligned
SRR1177960	84.00%	15.1
SRR1177967	83.10%	14.3
SRR1177968	84.40%	14.9
SRR1177971	84.20%	16.5
SRR1177984	82.60%	13
SRR1177985	82.20%	13.6
SRR1177986	83.10%	12.9
SRR1178023	83.90%	13.5
SRR1178049	84.80%	16.4

Table 2: MultiQC: STAR percentage alignment summary

After using featureCounts after filtering the zero values boxplots were generated which display the distribution of read counts for each sample (Figure 3 and 4). The figure demonstrates samples having a somewhat consistent, distribution read count.
After running MultiQC (Table 3) on the feature count, we can see that most of the reads have similar values, when we compare the feature counts vs the general statistics we can see there not much difference again between the reads. From 53.5% being the lowest and the highest 62.60%, a 9% difference.
The gene count generated by the program featureCounts shares the same distribution pattern as the general statistics obtained from MultiQC.

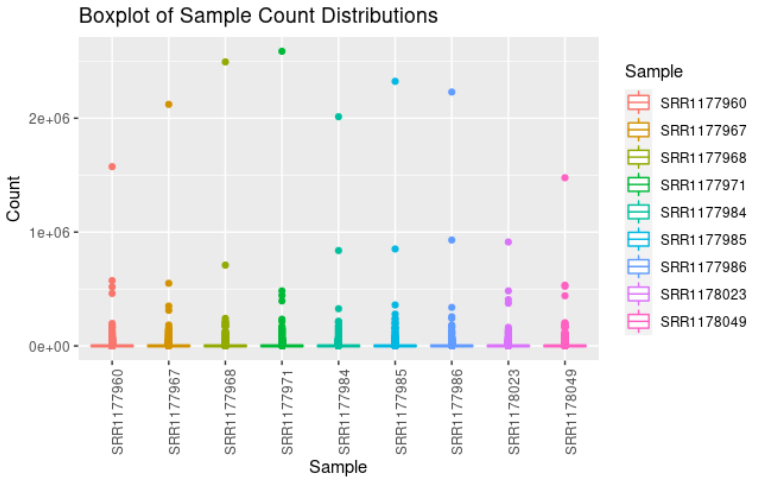

Figure 3: Box plot of sample count distribution

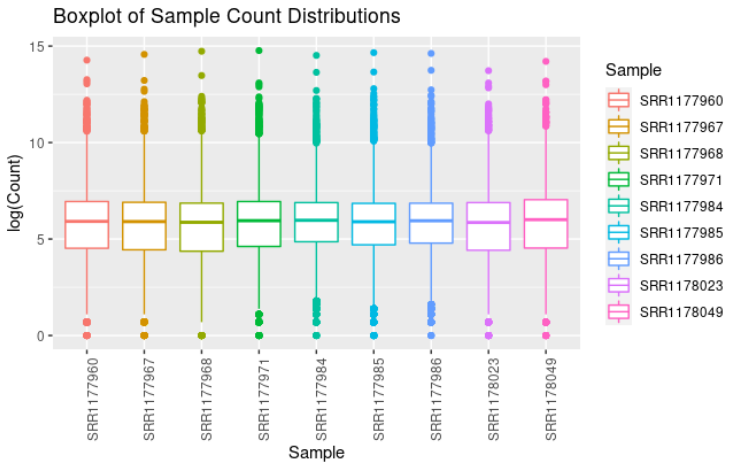

Figure 4 :Box plot of sample count distribution on log scale

Sample Name	% Assigned	M Assigned
SRR1177960	53.60%	19.8
SRR1177967	59.70%	19.2
SRR1177968	59.10%	19.7
SRR1177971	59.50%	22
SRR1177984	61.90%	18
SRR1177985	62.20%	19
SRR1177986	62.60%	18.2
SRR1178023	53.90%	17.2
SRR1178049	53.50%	20.6

Table 3: MultiQC: Feature Count summary statistics

Following featureCounts, differential gene expression analysis with DESeq2 shows there are 4369 differentially expressed genes within DNA Damage mode of action at adjusted p-values of < 0.05, Estrogen Receptor-mediated cytotoxicity shows there are 3252 differentially expressed genes at adjusted p-values of < 0.05 and PPARA mode of action has 2834 differentially expressed genes, at adjusted p-values of < 0.05.

The top ten differentially expressed genes were also identified for each mode of action as shown in Table 4. For tables with all respective values refer supplementary table 2,3 and 4.
Furthermore, we generated histograms for each mode of action samples displaying fold change values of differentially expressed genes (Figures 5). In addition, scatter plots were also plotted to demonstrate the fold changes versus the nominal p-values of significantly differentially expressed genes for each mode of action condition (Figures 6).
While from above counts of significant differentially expressed gene from each MOA shows that DNA-Damage has the highest gene count and is more consistent than the others but after adjusting the p value, we can see that there is not a significant difference between the groups, when we examine the figures 5 & 6.
Final part of the analysis involved creating volcano plots for each MOA for visual identification of genes with large fold changes that are also statistically significant. (Figure 7).

	DNA_DAMAGE	ER	PPARA
1	NM_031533	NM_019292	NM_012737
2	NM_053365	NM_013033	NM_133606
3	NM_022521	NM_053699	NM_024162
4	NM_001009353	NM_001271220	NM_012508
5	NM_012623	NM_001100661	NM_175837
6	NM_001191609	NM_001108542	NM_012575
7	NM_012580	NM_173136	NM_012598
8	NM_001025675	NM_031048	NM_001040019
9	NM_145672	NM_053445	NM_017158
10	NM_053734	NM_057100	NM_057197

Table 4: The top 10 Differentially Expressed genes for all three MOA

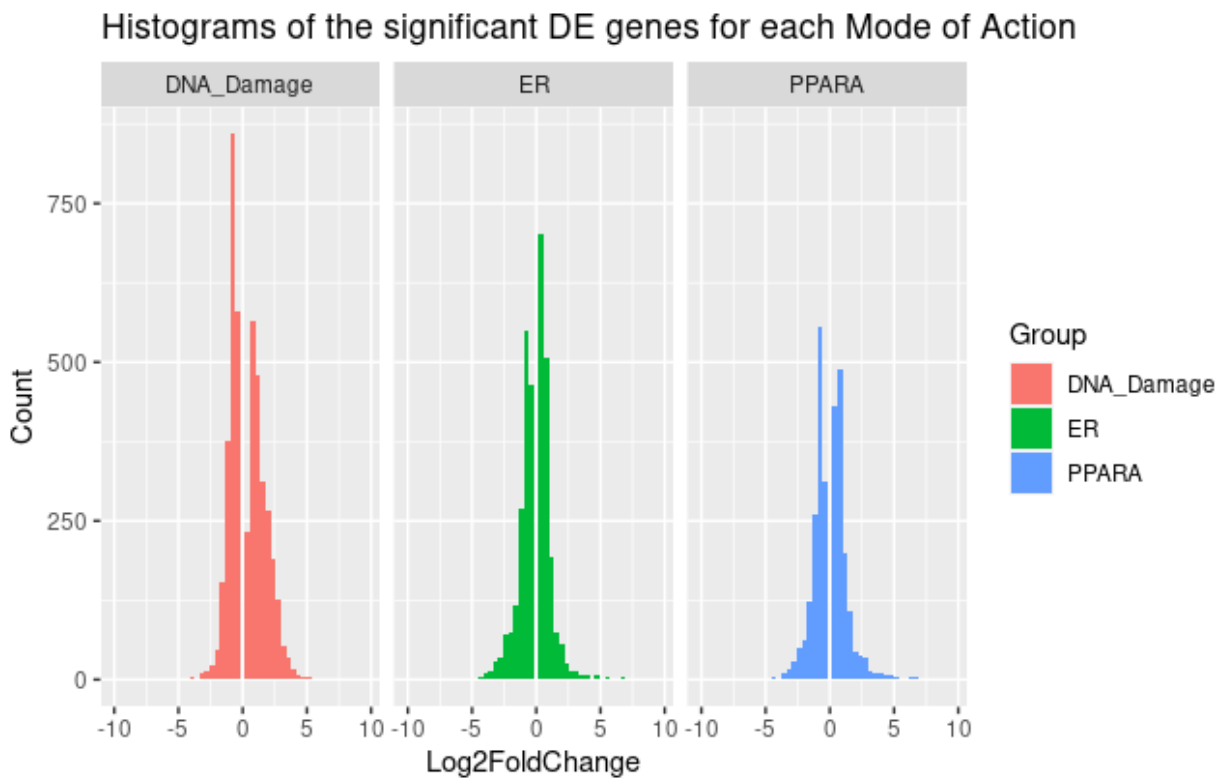

Figure 5: Histogram for Log2 fold change values of Differentially expressed genes

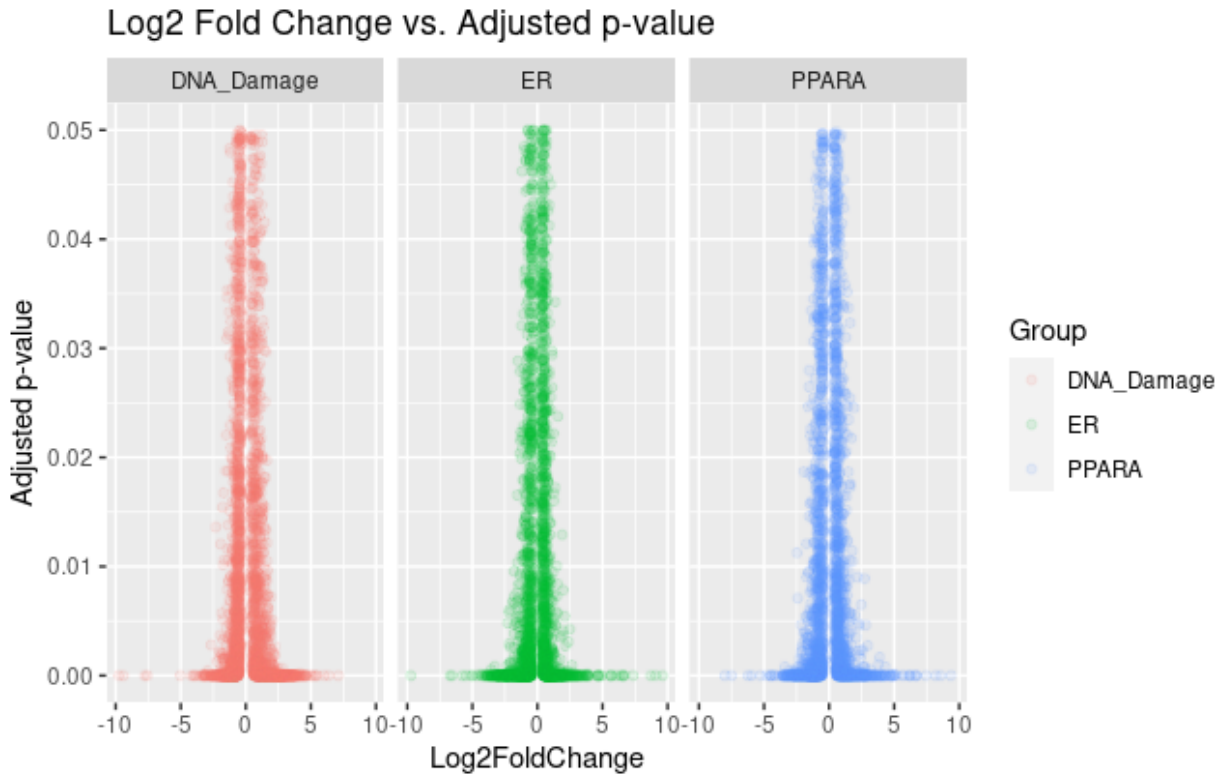

Figure 6 : Scatter plot for log2 fold change values of Differentially expressed genes

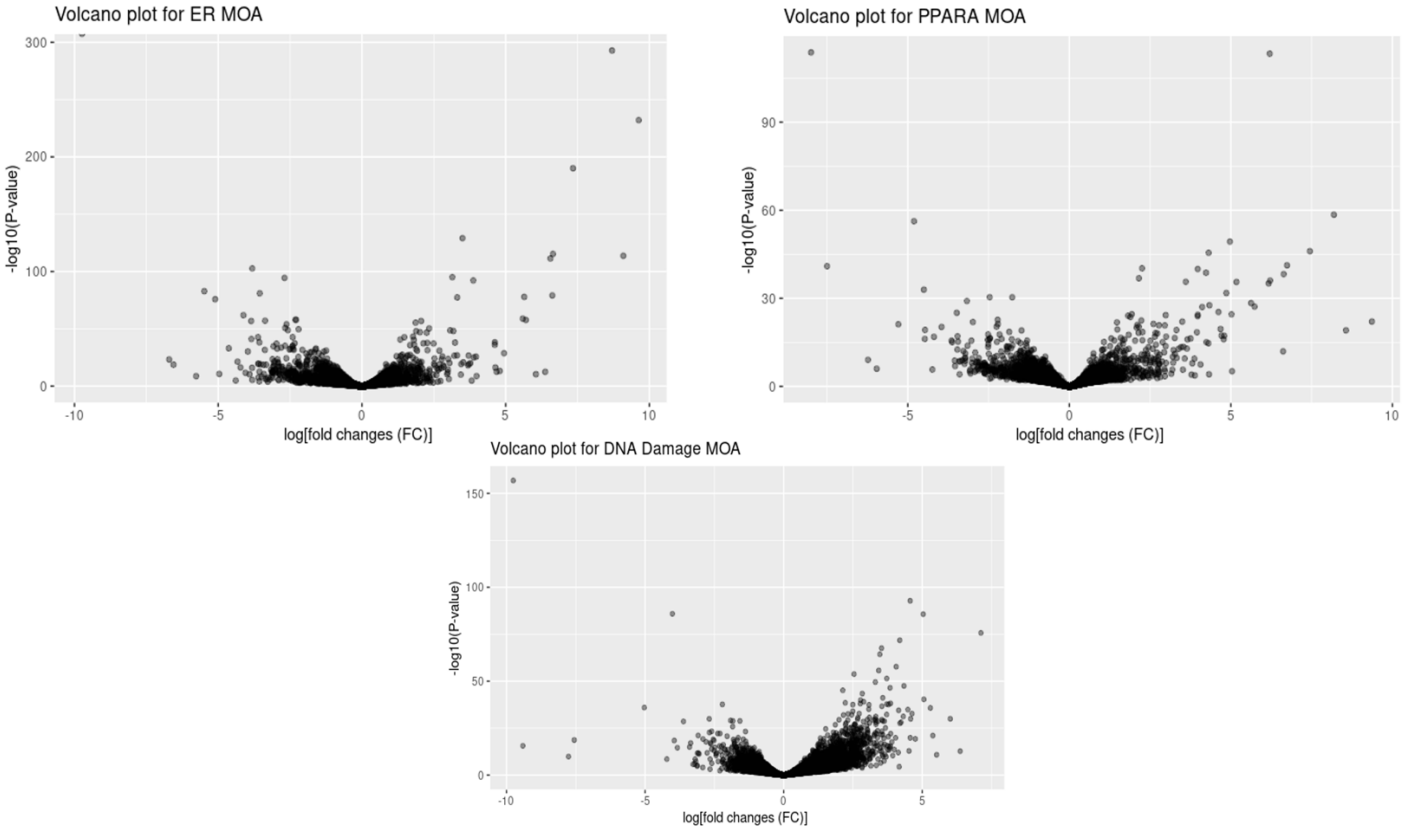

Figure 7: Volcano plot for all three different MOA Differentially expressed genes

Discussion

The data was obtained by aligning short reads to the rat genome and quantifying the expression through read counting using STAR. DESeq2 differential expression analyses was performed on the data. While the Wang et al.'s paper created a foundation to understand the differences between RNA-seq and microarray platforms. In our findings we only performed the RNA-seq part of the analysis in which after DESeq2 analysis we found there is almost no significant difference between the three MOA groups.

From the scatter plots of the p-value versus the fold change of the Differentially Expressed genes, it is observed that, there was no significance difference between them in the treatment sets compared to the control. Some studies have claimed that RNA-seq have lower precision while others believe it is highly sensitive in gene detection [1], but when considering factors such as cost and difficulty, microarray may be the better choice for clinical diagnostics and can be used in instances where conditions are drastically different, such as cancer and normal tissue. If the goal is to discover novel differentially expressed genes, the sensitivity of RNA-Seq is well suited. To gain further knowledge and draw comparison a microarray analysis of the same tox group will be a good way to approach as done by Wang et al.

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. 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

3. Liao, Yang et al. “featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features.” Bioinformatics (Oxford, England) vol. 30,7 (2014): 923-30. doi:10.1093/bioinformatics/btt656

Supplementary Materials

Run	mode_of_action	chemical	vehicle	route
SRR1177984	DNA_Damage	N-NITROSODIMETHYLAMINE	SALINE_100_%	INTRAPERITONEAL
SRR1177985	DNA_Damage	N-NITROSODIMETHYLAMINE	SALINE_100_%	INTRAPERITONEAL
SRR1177986	DNA_Damage	N-NITROSODIMETHYLAMINE	SALINE_100_%	INTRAPERITONEAL
SRR1177960	ER	BETA-ESTRADIOL	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178023	ER	BETA-ESTRADIOL	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178049	ER	BETA-ESTRADIOL	CORN_OIL_100_%	ORAL_GAVAGE
SRR1177967	PPARA	BEZAFIBRATE	CORN_OIL_100_%	ORAL_GAVAGE
SRR1177968	PPARA	BEZAFIBRATE	CORN_OIL_100_%	ORAL_GAVAGE
SRR1177971	PPARA	BEZAFIBRATE	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178004	Control	Vehicle	SALINE_100_%	INTRAPERITONEAL
SRR1178006	Control	Vehicle	SALINE_100_%	INTRAPERITONEAL
SRR1178013	Control	Vehicle	SALINE_100_%	INTRAPERITONEAL
SRR1178064	Control	Vehicle	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178074	Control	Vehicle	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178075	Control	Vehicle	CORN_OIL_100_%	ORAL_GAVAGE

Supplementary Table 1: Tox group 4

Gene	baseMean	log2FoldChange	lfcSE	pvalue	padj
NM_031533	15462.51751	-9.757026081	0.36631732	1.47E-157	1.61E-153
NM_053365	539.3618286	4.563337249	0.22377512	1.62E-93	8.89E-90
NM_022521	7054.096084	-4.015103711	0.20511679	1.33E-86	4.87E-83
NM_001009353	440.5430422	5.033522828	0.25679951	2.11E-86	5.78E-83
NM_012623	6150.28693	7.116248415	0.388465	1.95E-76	4.27E-73
NM_001191609	879.8855503	4.190032009	0.23482838	1.67E-72	3.05E-69
NM_012580	5180.96817	3.531811829	0.20424803	2.68E-68	4.19E-65
NM_001025675	756.6516071	3.469559213	0.20540709	4.94E-65	6.77E-62
NM_145672	1332.659737	4.058740311	0.25423514	1.95E-58	2.38E-55
NM_053734	1005.353768	3.430060994	0.21877869	1.98E-56	2.17E-53

Supplementary Table 2: Top 10 Differentially expressed genes for MOA-DNA Damage

Gene	baseMean	log2FoldChange	lfcSE	pvalue	padj
NM_019292	46138.39119	-9.737128905	0.25838485	0	0
NM_013033	5840.25493	8.700153013	0.23206628	1.53E-307	8.37E-304
NM_053699	3160.927831	9.628097778	0.29604498	7.69E-233	2.81E-229
NM_001271220	917.4173526	7.341916134	0.24903603	7.96E-191	2.18E-187
NM_001100661	20650.77307	3.499722242	0.14458458	7.93E-130	1.74E-126
NM_001108542	956.171642	6.645669793	0.29103509	4.36E-116	7.97E-113
NM_173136	5473.62458	9.090726829	0.40265649	1.97E-114	3.09E-111
NM_031048	30943.23544	6.555282923	0.29305078	4.78E-112	6.55E-109
NM_053445	27271.58681	-3.816830714	0.17753571	2.16E-103	2.63E-100
NM_057100	15580.53288	3.143973952	0.15206403	8.34E-96	9.15E-93

Supplementary Table 3: Top 10 Differentially expressed genes for MOA-ER

Gene	baseMean	log2FoldChange	lfcSE	pvalue	padj
NM_012737	2784.505115	-7.994797128	0.35224324	1.75E-114	1.91E-110
NM_133606	478803.6798	6.201998814	0.27442564	4.71E-114	2.58E-110
NM_024162	6422.872281	8.187832077	0.51022343	3.24E-59	1.18E-55
NM_012508	791.4169793	-4.807871276	0.30522823	5.38E-57	1.48E-53
NM_175837	120206.8442	4.970444038	0.33821306	5.05E-50	1.11E-46
NM_012575	435.0680953	7.444941738	0.52155196	8.85E-47	1.62E-43
NM_012598	4796.364017	4.310810776	0.30629353	3.10E-46	4.85E-43
NM_001040019	126371.1605	6.737466637	0.5040671	5.57E-42	7.63E-39
NM_017158	3242.747135	-7.500921471	0.56306506	1.12E-41	1.37E-38
NM_057197	19994.5107	2.250497004	0.16958294	5.91E-41	6.47E-38

Supplementary Table 4: Top 10 Differentially expressed genes for MOA-PPARA