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

Authors: Preshita Dave (Data Curator), Italo Duran (Programmer), Monica Roberts (Analyst)

Introduction

Microarrays have been at the forefront of analyzing transcriptomes for evaluating drug safety. It helps in identifying Differentially Expressed Genes (DEGs) and predicting patient/toxicity outcomes based on gene-expression data. The advent of newer technologies like high throughput sequencing technologies provides new methods for whole-transcriptome analyses of gene expression like RNA-seq. Some studies go on to compare the technical reproducibility of the results obtained from microarrays and RNA-seq experiments. Some have reported that RNA-seq reports lower precision for weakly expressed genes owing to the nature of sampling or higher sensitivity of RNA-seq for gene detection. Based on these varied conclusions, Wang et al [1] conducted a comprehensive study to evaluate RNA-seq in its differences and similarities to microarrays in terms of identifying DEGs and developing predictive models. The design of this study was to generate Illumina RNA-seq and Affymetrix microarray data from the same set of liver samples of rats under varying degrees of perturbation by 27 chemicals representing multiple modes of action (MOA). In doing so, three important findings were established: 1) the concordance between the two methods in detecting the number of DEGs was positively correlated to the level of perturbation caused by the chemical treatment, 2) RNA-seq is better at detecting weakly expressed genes than microarrays, and 3) the prediction models generated from the two methods have similar performance. Our study aims to reproduce the results of Wang et al [1] using the data from a toxicity group (toxgroup 4) with MOAs of DNA damage, peroxisome proliferator-activated receptor alpha (PPARA), ER-mediated cytotoxicity; and their corresponding vehicle controls. We assessed the quality of RNA-seq data of the experimental group, performed differential expression analysis on the microarray and RNA-seq data, and compared the results obtained from the two methods.

Data

The study design involved exposing male Sprague-Dawley rats (6-8 months) to one of 27 chemicals (three rats per chemical with matched controls), isolating RNA from the livers, and analyzing these samples using Affymetrix microarrays and Illumina RNA-seq. Furthermore, sets of three chemicals share one of five MOAs. The MOAs serve as endpoints for the investigation of the predictive models and cross-platform concordance. Test chemicals were administered orally or injected. In order to ensure a maximal transcriptional response, 5-day maximum tolerated doses (MTD) of test chemicals were administered to the study animals over 3,5 or 7 days depending on the chemical. RNA samples for RNA-seq were derived from the NTP DrugMatrix Frozen Tissue Library. Sequencing was performed on Illumina HiScanSQ or HiSeq2000 systems using the manufacturer's protocol. Samples were blinded during the entire

sequencing process, and ~23-25 million paired-end reads with 100 bp read length were generated. Data was deposited in the Sequence Read Archive (NCBI) under accession number SRP024314. Meanwhile, the microarray data was prepared using rat liver RNA hybridized to the Affymetrix whole-genome GeneChip® Rat Genome 230 2.0 Array following the procedure described in the technical manual.

For our study, we processed and analyzed one part of the data from the paper. We performed differential analysis on the Affymetrix microarray and RNA-seq of the toxgroup with MOAs ER (Estrogen Receptor-mediated cytotoxicity), DNA damage, and PPARA with specific chemical treatments.

Methods

RNA-Seq Sample Statistics and Alignment

FASTQ files for the nine treatment samples of the chosen toxgroup 4 were obtained from a common location on the Shared Computing Cluster (SCC) and symbolic links were created in the group folder for our personal analysis. FastQC [3] was utilized to assess the quality of sequencing. Each of the samples was aligned against the rat genome (provided in the SCC) using the STAR aligner (2.6.0c) [4]. STAR is an alignment program, like tophat, that is designed specifically for spliced alignment of RNA-Seq data. STAR also requires a genome index to be created or used, which was already provided for in the SCC. The study design included 15–60 million 100bp paired-end reads for each sample. Hence, STAR takes in two reads to be aligned together for each sample, and the alignments were saved as bam files. STAR also generates read alignment statistics, which have been summarized in Table 1. After aligning the reads, MultiQC [5] was used to summarize the quality statistics for FastQC and STAR output for all samples. MultiQC is a helpful tool for collecting information from many different bioinformatics programs for a set of samples and combining them into a single convenient report.

Read Continuing with feature counts

We used the alignments made with STAR, to count reads against a gene annotation using the program featureCounts. With these counts we quantified the abundance of genes to compare with the microarray data. Since featureCounts needs GTF annotation we used a reference annotation, rn4_refGene_20180308.gtf. We can ran this against each of the nine samples from the tox group 4 to generate counts for each sample using a qsub script made for the specific samples. The results were gathered to run a multiqc [6] analyses to examine for the feature counts across all samples, as shown in figure 4. We used R studio to build a single comma-delimited text file that we combined from the counts files from the samples. At last the purpose of this feature counts was to extend the analyses and build a box plots; figure 5; for each of your samples showing the distribution of counts [2].

RNA-Seq Differential expression with DESeq2

Basic reading counts were processed with RNA-Seq Differential expression with the use of DESeq2 by using R. Created new counts matrix that includes the control counts from last step. Sampled a metadata file group_4_rna_info.csv, that contained the sample ID's of the control samples for vehicles that correspond to the type of treatment for the samples. As well as Identification of the columns of the control samples within the sample matrix provided from the control_counts.csv file. DESeq2 was installed via Bioconductor package in R, to build a script a script that would compare each of the groups in the treatment against the control groups with it's matching vehicles and conditions, producing a three separate DE genen files, The files were subset to include the most relevant samples in the analyses [7]. Differential expression results were made to sort the files by the adjusted p-value. The number of genes that are significant at p-adjust < 0.05, were reported up to the top 10 DE genes from each analysis by p-value, shown by table 2. Constructed scatter plots of fold change vs nominal p-value s shown in figures, 6 to 10.

Concordance between microarray and RNA-Seq DE genes

Concordance between platforms was calculated with the following equation[1]:

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

The number of genes that intersected both platforms was adjusted for background noise. In other words, the actual number accounted for and removed genes that intersected by chance. The following equation was used:

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

Where n_x was the true number of intersecting genes, N was the total number of genes evaluated, n1 was the total number of DE genes in the microarray analysis, and n2 was the total number of DE genes in the RNA-Seq analysis. Each probe in the microarray analysis was mapped to the corresponding refseq identification using a master spreadsheet that contained all identifiers. Probes that did not match any genes were removed. Some probes matched multiple genes and all genes were kept in the analysis. Lastly, some probes matched the same genes and the probe with the lowest p-value was kept. Concordance was calculated for each group between the platforms.

Microarray Differential Expression Analysis with Limma

RMA (robust multichip average) normalization and the software package Limma[2] was used to determine differentially expressed genes in the microarray data. The RMA corrected, normalized, and summarized the probe level information to account for differences in gene expression that could be due to confounding factors such as technical error, library size, and PCR amplification bias. The RMA expression matrix made available by Wang et al was used to carry out our analysis. Since the matrix was pre-processed by the authors of the paper, no quality control measures were taken.

Three treatments were compared to controls in tox group 4: beta-estradiol, bezafibrate, and N-nitrosodiethylamine. Each of these corresponded to the mechanisms of action ER, PPARA, and DNA Damage. The sample IDs corresponding to tox group 4 were subsetted from the RMA matrix and the design matrix was constructed. The design matrix included information on the model parameters, which describes the chemical group names. The purpose of the design matrix was to indicate which group corresponds to each sample in the expression matrix. The RMA expression matrix was then passed into the limma linear model function (lmFit) along with the design matrix. When fitted to the data, the model produces an object that contains coefficients, standard errors, and residual standard errors for each probe. These coefficients describe the relationship between the response variable, the expression of each gene, and the explanatory variable, the chemical group.

After the linear model was fitted, it was passed into the function eBayes, which uses a Bayes model to moderate the standard errors. This function was used to rank the genes for differential expression. The output of the Bayes model included a t-statistic and log odds of differential expression for each gene. The function assessed differential expression by minimizing the probe-wise variance.

The object created by eBayes was then passed into the topTable function, which summarized the model. This included log fold change values, p-values, adjusted p-values, and average expression per gene. The p-value was adjusted using the ‘BH’ method, which controlled the false-discovery rate that resulted from multiple testing.

Results

RNA-Seq Sample Statistics and Alignment

After aligning the reads, STAR outputted the alignment statistics for each sample run. Table 1 summarizes these statistics.

Sample ID	# Input reads	# Uniquely mapped reads	% Uniquely mapped reads	% Mismatch rate per base	# reads mapped to multiple loci	% of reads mapped to multiple loci	# reads mapped to too many loci	% of reads mapped to too many loci	Unmapped reads (too many mismatches + too short + others)
SRR1177960	17947778	15080805	84.03%	0.45%	1189810	6.63%	42948	0.24%	0% + 8.96% + 1.5%

SRR1177967	17157413	14253911	83.08%	0.77%	673010	3.92%	31242	0.18%	0% + 12.78% + 0.04%
SRR1177968	17604520	14857439	84.40%	0.77%	686056	3.90%	29288	0.17%	0% + 11.48% + 0.06%
SRR1177971	19627402	16518020	84.16%	0.83%	732186	3.73%	51903	0.26%	0% + 11.78% + 0.07%
SRR1177984	15732068	12999695	82.63%	0.77%	586314	3.73%	29840	0.19%	0% + 13.41% + 0.04%
SRR1177985	16537337	13601134	82.25%	0.69%	630871	3.81%	24455	0.15%	0% + 13.75% + 0.04%
SRR1177986	15559784	12936410	83.14%	0.72%	600101	3.86%	21894	0.14%	0% + 12.82% + 0.04%
SRR1178023	16076957	13486519	83.89%	1.13%	915705	5.70%	18535	0.12%	0% + 10.25% + 0.05%
SRR1178049	19397953	16448173	84.79%	0.92%	1041082	5.37%	49116	0.25%	0% + 9.53% + 0.06%

Table 1: *STAR read and alignment summary statistics.*

The columns with “Read mapped to too many loci” are defined as mapping to more than 10 locations. Overall, the percentage of uniquely aligned and multi-mapped reads were within reasonable range to proceed with further analysis.

MultiQC helped to process results from both FastQC and STAR. Figure 1 represents the STAR Alignment scores. SRR1178049 has the highest number of uniquely mapped reads, followed by SRR1177971. Based on Table 1 and Figure 1, the percentage of unmapped reads which are too short lie in the range of 10-13%.

Figure 2 represents the number of reads with average quality scores. Based on the figure, a majority of the reads have a phred score greater than 28, which is the threshold for good quality.

Figure 3 represents the mean quality value across each base position in the read. Around base pair position 70, many of the reads drop below the phred score of 28, indicating that trimming might be required for better quality of reads. SRR1178023_2 has the lowest mean quality score as the phred score drops below 28 after base pair position 60.

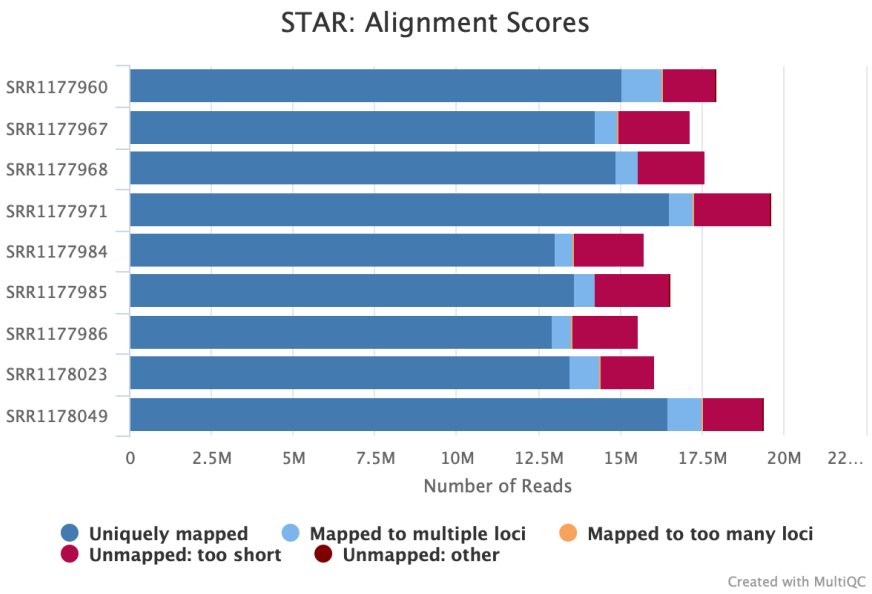

Figure 1: *STAR alignment statistics by MultiQC.*

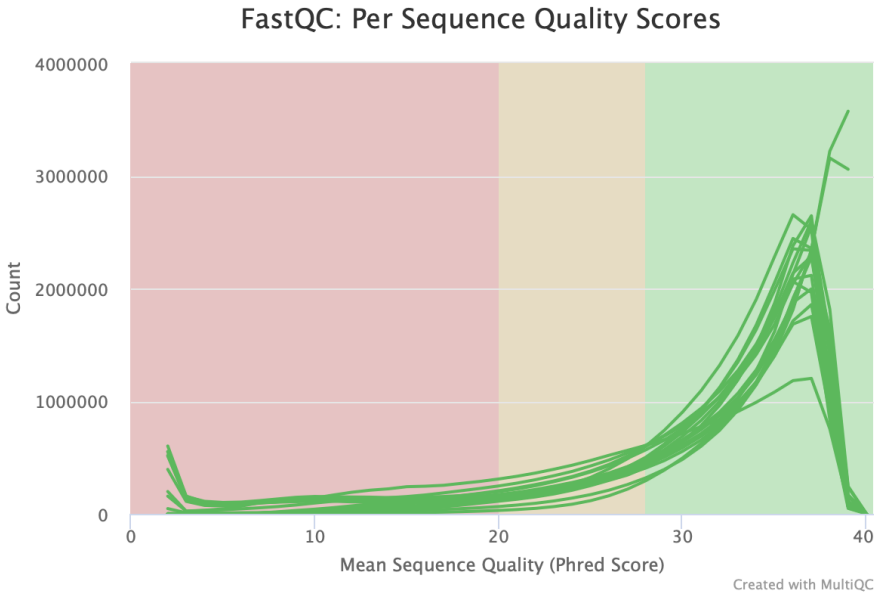

Figure 2: *FastQC Per Sequence Quality Scores through MultiQC.*

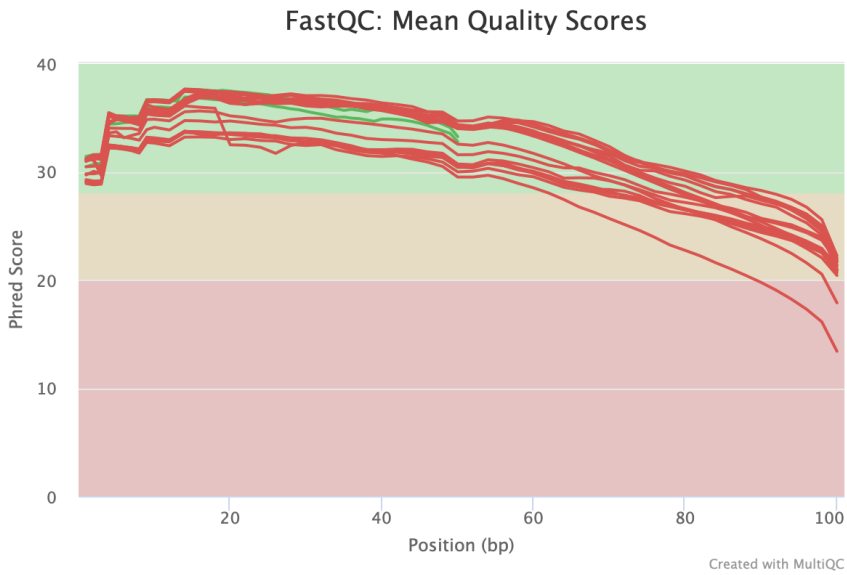

***Figure 3**: FastQC Mean Quality Scores through MultiQC.*

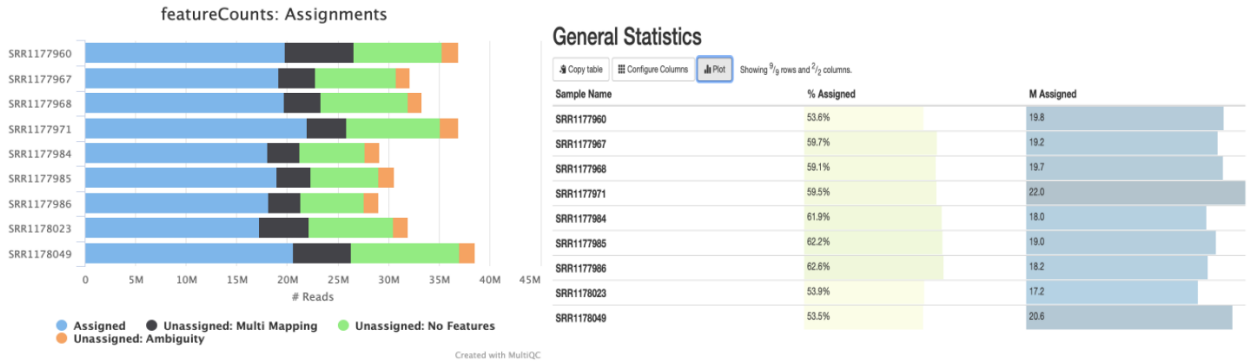

***Figure 4**: MultiQC Report*

MultiQC, 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 62.6%, a 9% difference.

After running MultiQC, we can see that most of the reads are somewhat consistent, when we compare the feature counts vs the general statistics we can see there is not much difference again between the reads. With a 9% difference. Once we even go further into the analyses between the three different groups, as we see in figure 5 there is almost no difference between the feature counts of the genes whether it was adjusted by using log or not, there was no significance difference between them.

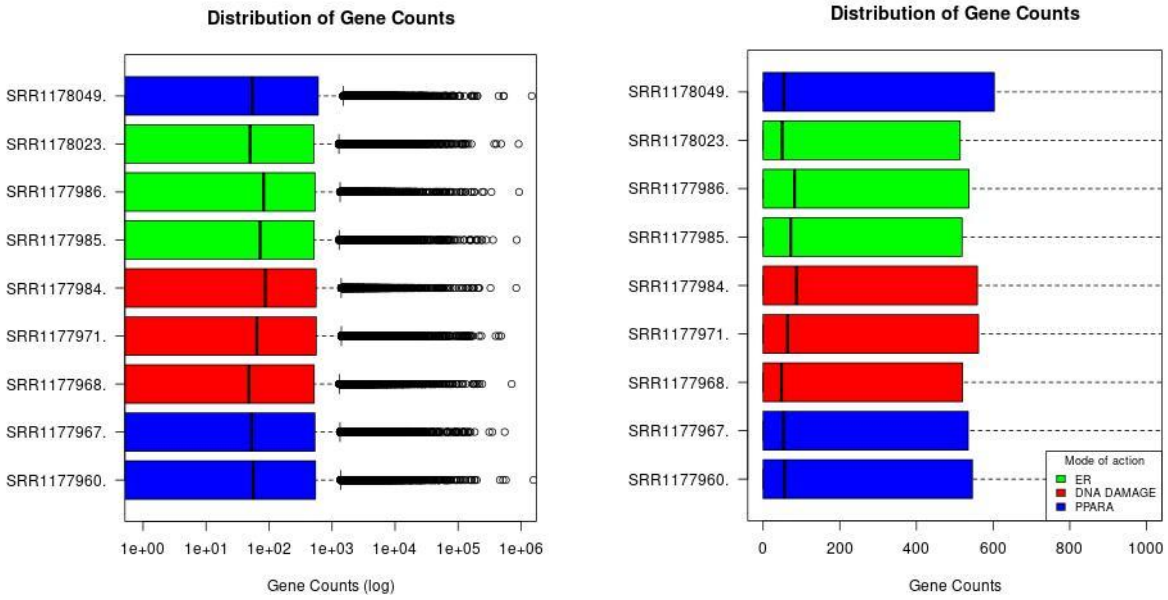

***Figure 5**: Gene count distributions for each sample.*
We can observe that there is no major significant difference between the distribution of gene counts with or without a log scale. Green representing ER, red DNA_Damage and blue for PPARA.

Through the generation of feature counts, as seen above, we can concur that the gene count generated by the program featureCounts and that general statistics from MultiQC share the same distribution pattern.

	DNA_Damage	ER	PPARA
Significant genes count	4369	3252	2834
1st	NM_031533	NM_019292	NM_012737
2nd	NM_053365	NM_013033	NM_133606
3rd	NM_022521	NM_053699	NM_024162
4th	NM_001009353	NM_001271220	NM_012508
5th	NM_012623	NM_001100661	NM_175837
6th	NM_001191609	NM_001108542	NM_012575
7th	NM_012580	NM_173136	NM_012598
8th	NM_001025675	NM_031048	NM_001040019
9th	NM_145672	NM_053445	NM_017158
10th	NM_053734	NM_057100	NM_057197

***Table 2**: Top 10 DE genes identified in the RNA-Seq analysis for each treatment group.*

The top 10 differentially expressed genes for the three different groups with their corresponding significant gene count. The table shows that DNA-Damage has a higher significant gene count and it's more consistent than the rest, but when we adjust the p value, we can see that there is almost no significant difference between the groups, when we look at figures.

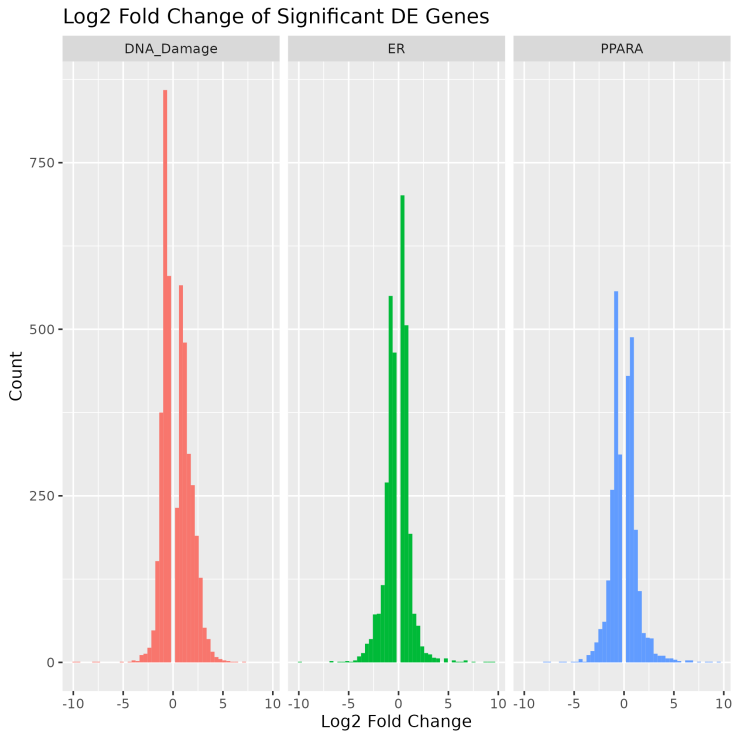

***Figure 6:** Distribution of Log_2 Fold Change values for each treatment group in the RNA-Seq differential expression analysis. On red it's DNA_Damage, green is ER and blue for PPARA. Distribution of log fold change values changes significantly, it shows the distribution in expressed genes as identified by RNA-seq analyses at the adjusted p-value.*

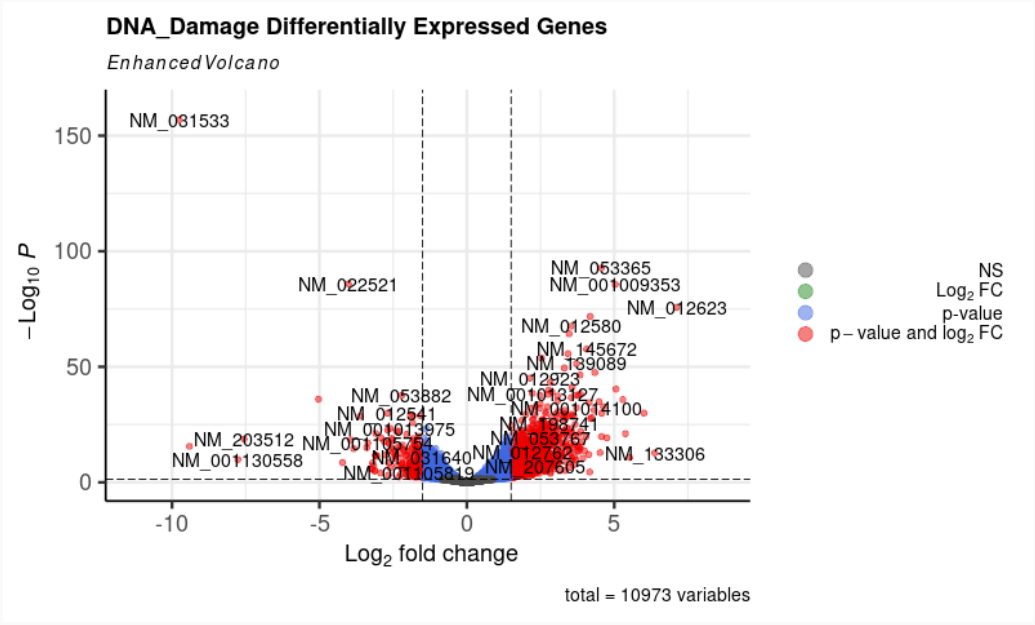

Figure 7: *Volcano plot of differentially expressed genes in the DNA Damage treatment group. The plot shows the relationship between -log$_{10}$(p-value) and Log$_2$ Fold Change.*

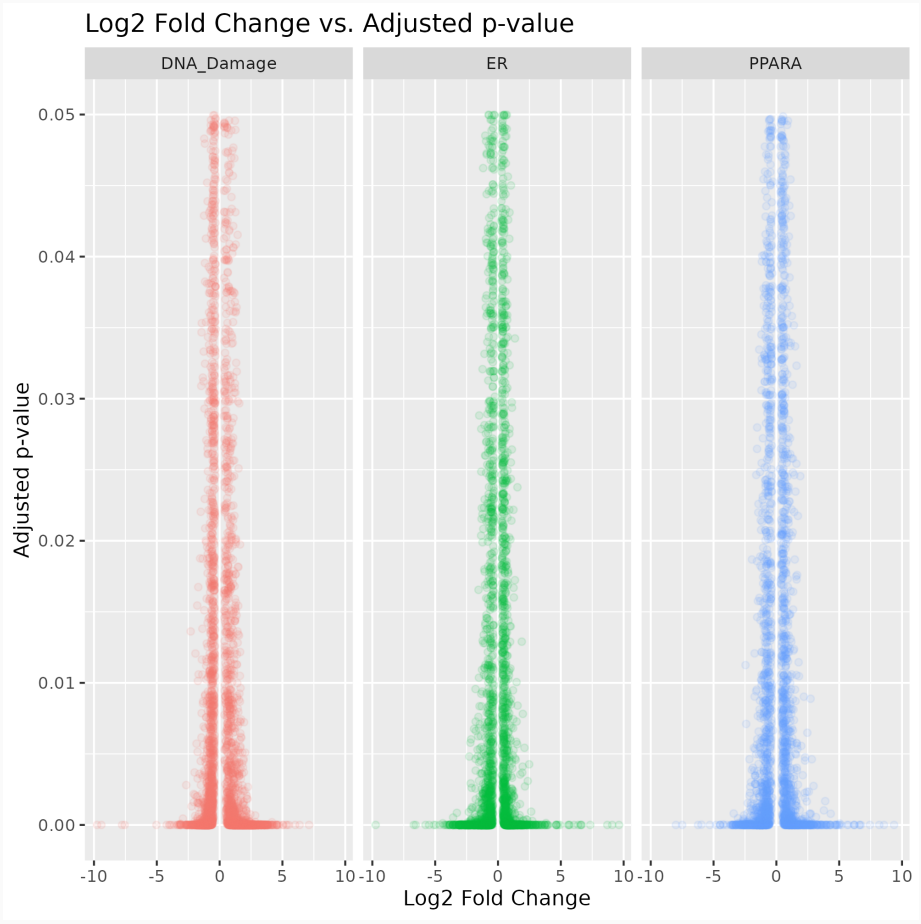

Figure 8: *Relationship between gene Log_2 Fold Change and adjusted p-values for each treatment group from the RNA-Seq differential expression analysis. On red it's DNA_Damage, green is ER and blue for PPARA.*

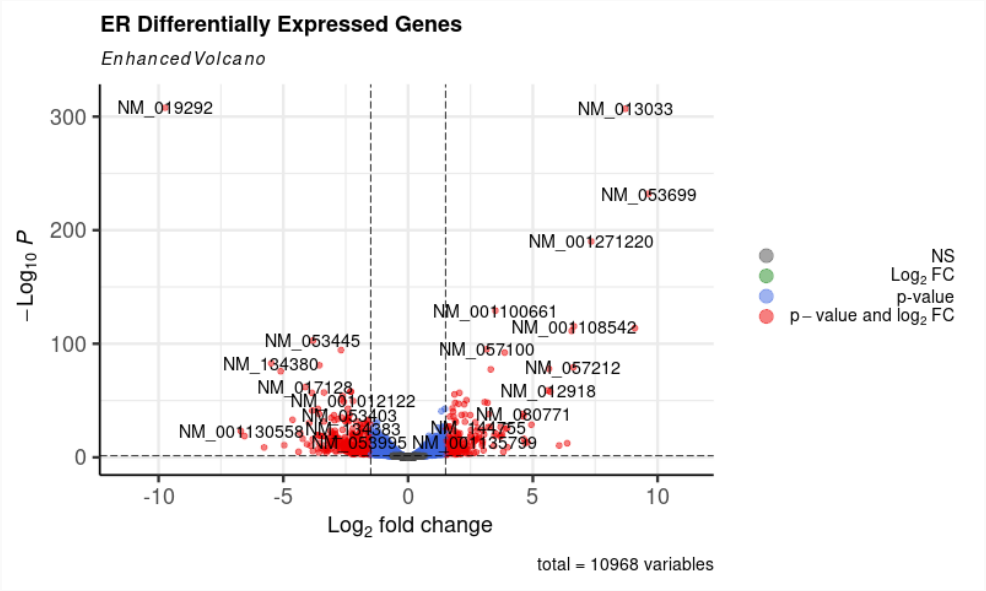

Figure 9: *Volcano plot of differentially expressed genes in the ER treatment group. The plot shows the relationship between -log_{10}(p-value) and Log_2 Fold Change.*

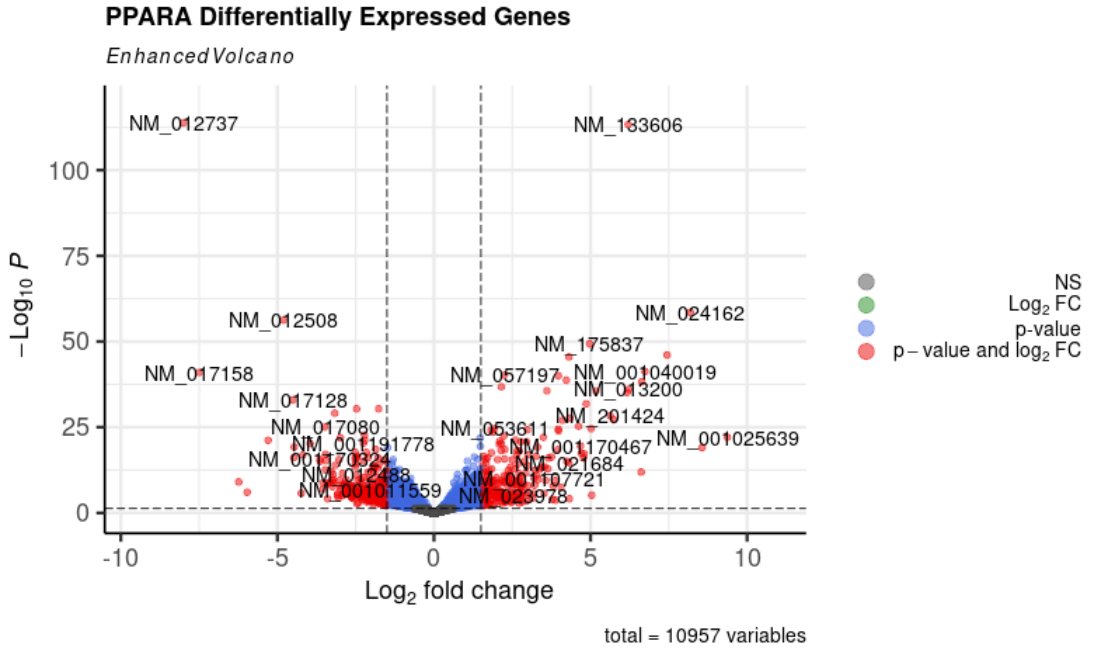

Figure 10: *Volcano plot of differentially expressed genes in the PPARA treatment group. The plot shows the relationship between -log_{10}(p-value) and Log_2 Fold Change.*

Microarray Differential Expression Analysis with Limma

Three analyses were conducted on the microarray data using Limma[2], one for each treatment group against the control. Each chemical in the analysis corresponded to a known mechanism of action. Beta-Estradiol corresponded to ER, bezafibrate to PPARA, and N-nitrosodiethylamine to DNA damage. There were 31,099 probes that mapped to 12,747 unique genes measured in each analysis. Genes were deemed differentially expressed if they had an adjusted p-value less than 0.05. Some probes mapped to several genes and some genes had several probes mapped to them. The duplicate probes were kept in the analysis since it was not possible to discern which mapped gene was actually measured by the probe at the time of the experiment. If several probes matched to a gene, the most significant probe was kept in the analysis. The number of DE genes per group from the microarray analysis are listed in Table 3.

Chemical	**# DE Genes**
Beta-Estradiol	1,372
Bezafibrate	2,904
N-Nitrosodiethylamine	80

Table 3: *Number of differentially expressed genes in each chemical vs. control group. Differentially expressed genes have an adjusted p-value less than 0.05 when compared to the control group.*

The top 10 differentially expressed genes for each group had the lowest p-values. The top 10 for beta-estradiol can be found in table 4, the top 10 for bezafibrate can be found in table 5, and the top 10 for N-nitrosodiethylamine can be found in table 6.

Probe ID	**Gene Symbol**	**P-value**
1368731_at	ORM1	2.275664e-12
1367957_at	RGS3	3.337102e-11
1374863_at	RBP7	4.262222e-11
1375900_at	TNFRSF9	1.078392e-10
1387435_at	ST8SIA3	6.534058e-09
1371089_at	GSTA3	6.932008e-09
1373369_at	ZMIZ1	8.122641e-09
1372468_at	ADGRE5	9.931612e-09

1393245_at	PHYH	2.024858e-08
1390918_at	GRTP1	2.422657e-08

Table 4: *Top 10 differentially expressed genes between the beta-estradiol and control groups in the microarray analysis.*

Probe ID	Gene Symbol	P-value
1368934_at	CYP4A1	2.491924e-38
1368934_at	CYP4B1	2.491924e-38
1378169_at	ACOT3	2.598095e-35
1398250_at	ACOT1	8.746752e-34
1391433_at	ACOT2	1.066194e-33
1367680_at	ACOX1	1.522672e-32
1368283_at	EHHADH	7.594547e-30
1371886_at	CRAT	1.996521e-28
1387740_at	PEX11A	5.431953e-28
1377037_at	ACOT4	2.901054e-27

Table 5: *Top 10 differentially expressed genes between the Bezafibrate and control groups in the microarray analysis.*

Probe ID	Gene Symbol	P-value
1390317_at	RGD1561849	7.435604e-25
1385132_at	MYBL1	1.512726e-16
1371036_at	NRCAM	3.012110e-16
1374245_at	NHEJ1	7.845193e-16
1370157_at	PLN	1.987797e-15
1383288_at	MDM2	1.939545e-11
1368691_at	GRIA3	2.407431e-11

1370583_s_at	ABCB1A	3.292723e-10
1370583_s_at	ABCB1B	3.292723e-10
1374251_at	KCNJ15	1.202223e-09

Table 6: *Top 10 differentially expressed genes between N-nitrosodiethylamine and control groups.*

The distribution of log fold-change values for differentially expressed genes in each treatment and the relationship between the nominal p-value and log fold-change of the differentially expressed genes for each treatment group is shown in figures 3-5.

A B

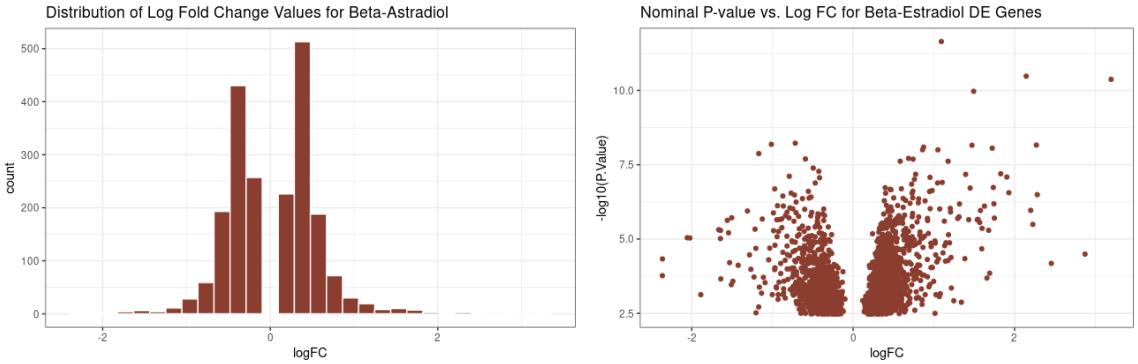

Figure 11: *Beta-Estradiol microarray differential expression results.* ***A:*** *Distribution of log fold-change values for differentially expressed genes in the Beta-Estradiol treatment group.* ***B:*** *Relationship between log fold-change values and nominal p-values of differentially expressed genes in the Beta-Estradiol treatment group.*

A B

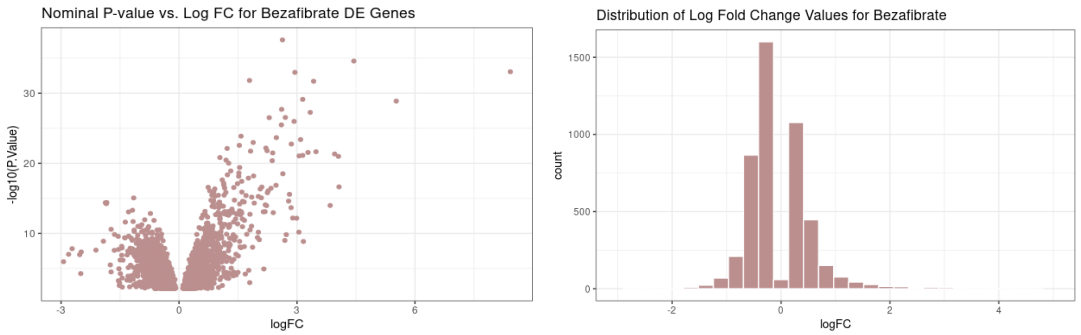

Figure 12: *Bezafibrate microarray differential expression results.* ***A:*** *Distribution of log fold-change values for differentially expressed genes in the Bezafibrate treatment group.* ***B:*** *Relationship between log fold-change values and nominal p-values of differentially expressed genes in the Bezafibrate treatment group.*

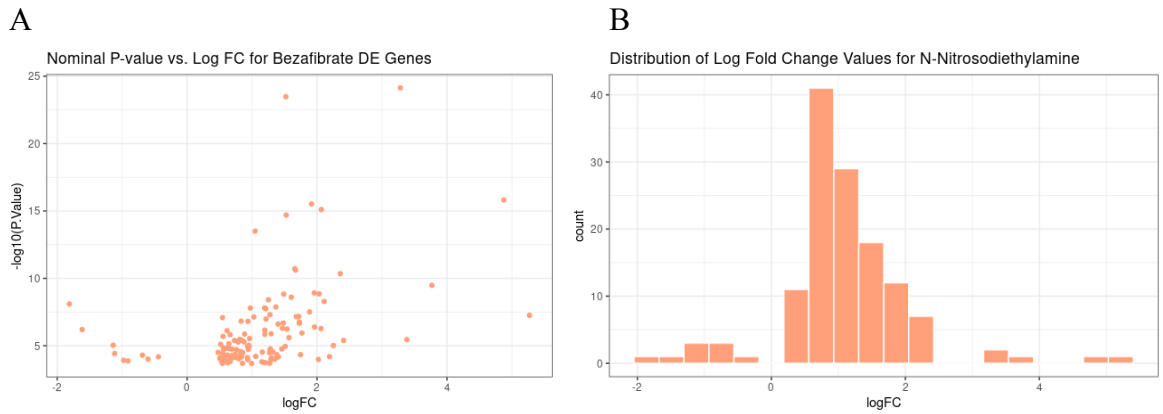

Figure 13: *N-Nitrosodiethylamine microarray differential expression results.* ***A:*** *Distribution of log fold-change values for differentially expressed genes in the N-Nitrosodiethylamine treatment group.* ***B:*** *Relationship between log fold-change values and nominal p-values of differentially expressed genes in the N-nitrosodiethylamine treatment group.*

Results from the RNA-Seq differential expression analysis and microarray differential expression analysis were compared to determine the number of genes identified as differentially expressed in both analyses. The number of common differentially expressed genes was adjusted for background noise, the genes that were differentially expressed in both groups by chance, using the equation described in the methods. The true value of intersecting genes was 761 for beta-estradiol, 1,273 for beta-fibrate, and 35 for N-nitrosodiethylamine. Concordance was calculated between the differentially expressed genes determined by microarray and RNA-Seq for each treatment group. For all genes, beta-estradiol had a concordance of 34.96, beza-fibrate had a concordance of 46.10, and N-nitrosodiethylamine had a concordance of 1.71. The relationship between the number of differentially expressed genes and concordance is shown in Figure 6.

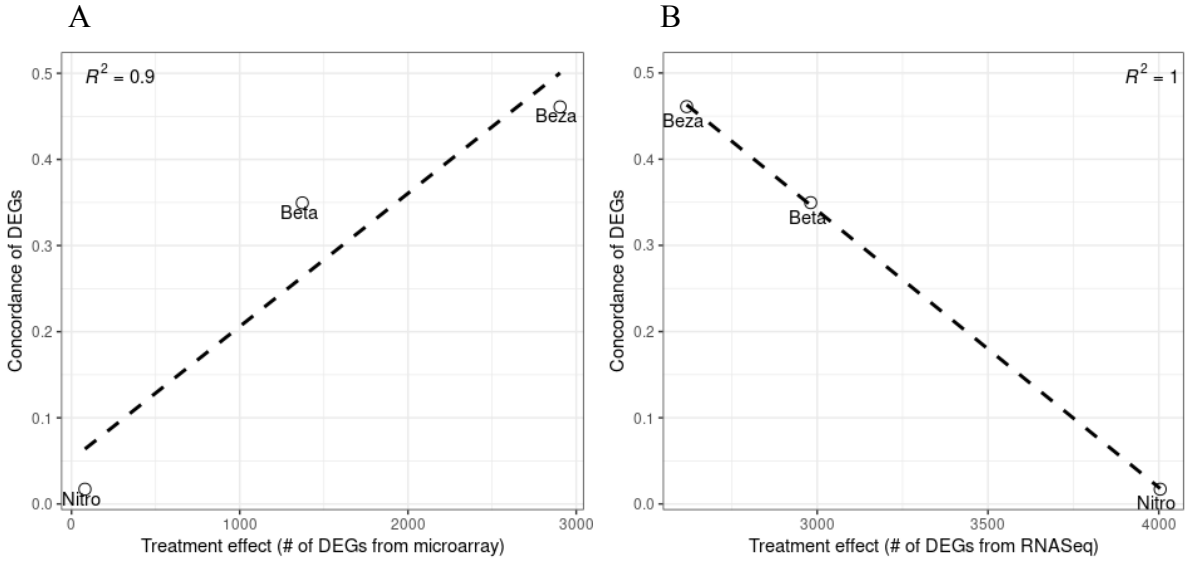

Figure 14: *Linear relationship between number of differentially expressed genes and concordance score.* ***A:*** *Microarray results.* ***B:*** *RNASeq results.*

Each treatment group was then split into two groups of genes: genes with above median expression and genes with below median expression. True intersection of genes was recalculated along with concordance for each subgroup in the treatment groups. The comparison between concordance of the subgroups across the treatment groups is shown in Figure 7.

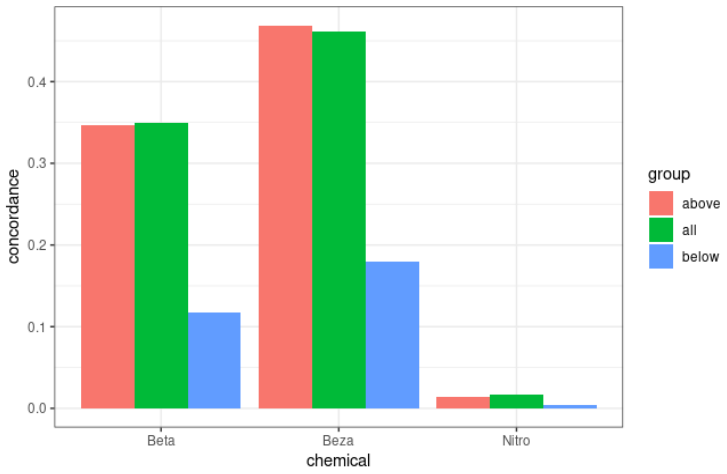

Figure 15: *Comparison of concordance between all genes, above median expression, and below median expression between treatment groups.*

Discussion

In this analysis, 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. The main findings of the analysis revealed some discrepancies in differential gene expression analysis between the two methods.

The relationship between the degree of perturbed gene expression and concordance was measured and shown in figure 14. There is a strong positive linear relationship in the microarray data, however the RNA-Seq data shows a strong negative correlation. The strong positive correlation for the microarray data agrees with the findings in the Wang et. al study that showed that lower concordance is expected in more similar conditions. This is due to the fact that RNA-Seq performs much better than microarrays at discerning lowly expressed genes and can therefore identify if they are differentially expressed. The inverse relationship seen in the RNAseq data could be attributed to a small number of data points or the effect the background correction had on the number of true intersected genes. The larger the difference between the number of genes in both sets, the smaller the proportion of intersected genes are deemed true. Overall, RNA-Seq identified more differentially expressed genes than the microarray did, especially for the DNA damage treatment group, which had the lowest concordance.

Comparing the concordance between above-median expressed genes and below-median expressed genes, figure 15 shows both methods perform better with above-median expressed genes. The agreement between both methods did not drastically change when only above-median

expressed genes were included, but it did dramatically drop when only below-median expressed genes were included. This reiterates the known fact that microarray technology has lower resolution for lowly expressed genes compared to RNA-Seq.

RNA-Seq and microarray are increasingly being used in a clinical setting to identify biomarkers for disease. The findings suggest there are differences in the method that should be taken into consideration when deciding protocol for diagnostics using gene expression data. Microarray could be used when biological conditions are substantially different, however, if the conditions are similar, RNA-Seq outperforms microarray greatly. 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 better suited.

Some of the findings were successfully replicated from Wang et. al, while others were not, likely due to less mechanisms of action being evaluated. To gain more insight into the performance of both platforms on predicting mechanisms of action given a transcriptional profile, gene set enrichment analysis and hierarchical clustering could be performed for future analyses. This could evaluate each platform's ability to diagnose a condition based on the given gene expression profile and further validate the use of both methods in a clinical setting.

Conclusion

Overall, with the advent of RNA-seq analysis, a conclusive comparison can be made regarding specific pathways, genes expression, more accurately than microarrays.
We could say that based on our results that there is high concordance found in the data set. As well for the three sample groups, when compared with the rest of the data sets. This concludes that our analysis has a parallel similarity to the original research. We can also observe the major differences between microarrays and how RNA-seq has an upper hand in terms of sensitivity for low expression, size and time it takes to process datasets for higher quality reads. Consequently, microarrays are substantially still being used for their practicality and price, when profiling smaller, non sensitive datasets.

References

1. Wang, Charles, Binsheng Gong, Pierre R. Bushel, Jean Thierry-Mieg, Danielle Thierry-Mieg, Joshua Xu, Hong Fang, et al. 2014. "A comprehensive study design reveals treatment- and transcript abundance–dependent concordance between RNA-seq and microarray data" Nature Biotechnology 32 (9): 926–32.
2. Smyth G. Limma: linear models for microarray data. *Bioinformatics and Computational Biology Solutions using R and Bioconductor.* 2005:397–420.

3. Babraham Bioinformatics. FastQC (Version 0.11.9) [Program documentation]. Retrieved March 3, 2021, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
4. Dobin, Alexander et al. “STAR: ultrafast universal RNA-seq aligner.” Bioinformatics (Oxford, England) vol. 29,1 (2013): 15-21. doi:10.1093/bioinformatics/bts635
5. 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
6. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Philip Ewels, Måns Magnusson, Sverker Lundin and Max Käller.* Bioinformatics (2016) doi: 10.1093/bioinformatics/btw354 PMID: 27312411
7. 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.