Concordance of Microarray and RNA-seq Differential Gene Expression

Data Curator: Teresa Rice
Programmer: Maha Naim
Analyst: Arushi Shrivastava

INTRODUCTION

Differentially expressed genes (DEGs) are genes which have statistically significant changes in expression levels between treatment groups. Research on DEGs is expanding with regards to investigation of gene markers for specific disease outcomes. RNA-seq and microarrays are two common methods used to measure gene expression and analyze DEGs. Microarrays have historically been the go-to technology, however, emerging RNA-seq technologies allow for analysis of DEGs at high-throughput volumes and whole-transcriptome analyses. In order for new technologies to become standard in clinical and regulatory applications, these technologies must be thoroughly assessed. The U.S. Food and Drug Administration (FDA) has already begun investigating the reliability of soon-to-be outdated microarrays. There are some studies which directly conflict others when comparing RNA-seq analyses such as sensitivity [7,8]. Wang et al. addresses the question about whether microarray technologies or RNA-seq are superior to one another when predicting toxicity outcomes [1]. The study looks at Affymetrix microarray and Illumina RNA-seq gene expression profiles from rat livers that have been exposed to 27 chemicals with varying modes of action (MOAs). In our analysis, we will look into a fraction of these profiles using the tools and methods outlined in this paper in order to examine the concordance between RNA-seq and microarray technologies among three different chemical groups with distinct MOAs.

DATA

Three male Sprague-Dawley rats were exposed to one of 27 chemicals, RNA was isolated from these mice and analysed with microarray chips (Affymetrix) and RNA-seq. Our analysis will only look at three of these chemicals for a total of 9 mice. The microarray data was from the Affymetrix GeneChip® Rat Genome 230 2.0 array[1]. The RNA reads are paired ends and input reads range from 15559784 to 19627402 with an average of 17293468 reads. Each single read is around 100 bp except for one sample which has 50 bp for each read in the pair. Microarray and RNA-seq data are from NCBI services and databases from accessions SRP039021, GSE55347, and GSE47875[2]. According to the bioinformatics tool, MultiQC, all samples pass 4/10 metrics and 14 or more samples fail 3/10 metrics[5](Figure 1).

METHODS

Quality Control and Sample Alignment

Samples are downloaded from NCBI; RNA-seq data for three conditions are grouped together. Our group will be reporting on tox group 4 which includes N-nitrosodimethylamine, beta-estradiol, and bezafibrate. FastQC is run on each of the files, producing FastQC Reports for each read for review[3]. Simultaneously, the alignment program STAR(v 2.6.0c) is run on each sample against the rat genome using the provided reference genome, this tool will also provide .bam files for other analysis[4]. A summary of STAR output can be found in Table 6 and consists of uniquely mapped reads, reads mapped to multiple loci, and unmapped reads. Additionally, the tool MultiQC is used for a variety of extra QC analysis[5](Figure 1). The FastQC and MultiQC results reveal samples with failed QC measures. All samples failed Per base sequence content and Sequence Duplication Levels from the FastQC. All or most samples failed Per Base Sequence Content, Sequence Duplication Levels, and Sequence Quality Histograms from MultiQC (Figure 1). Samples were not trimmed despite some QC metrics in order to keep read lengths.

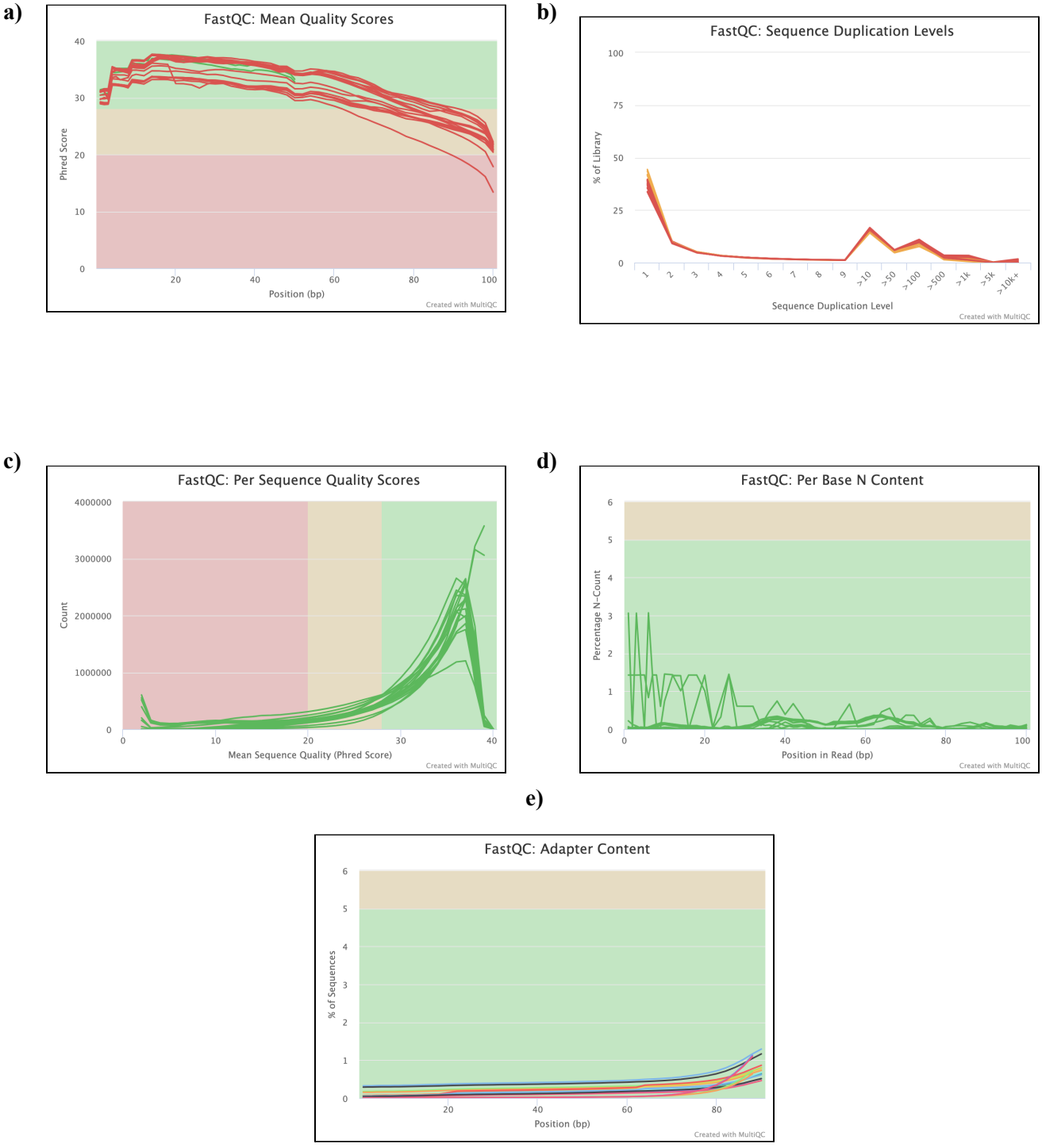


Figure 1. MultiQC Report. Samples failed Mean Quality Scores (a), and Sequence Duplication Levels (b) but passed Per Sequence Quality Scores (c), Per Base N Content (d), and Adapter Content from MultiQC Report (e). (Other metrics from MultiQC report had varied amounts of failed, warning, and passed samples)

Transcript Abundance Quantification Using featureCounts

Quantified transcript abundance measurements were collected to compare with the microarray data. Using the files generated by STAR, count files were produced using the featureCounts tool from the subread package (version 2.0.1). A batch job was performed to run featureCounts on each of the 9 aligned BAM files against the genomic features in the rat reference gene annotation file. Read counts for each sample were compiled into a single CSV file. The resulting count files

were then subject to quality assessment by MultiQC (version 1.6) which outputted an HTML report with plots displaying the read assignment measurements as seen in Figures 2 and 3 [5]. The same information is reported by both figures, however Figure 3 allows for enhanced inference of the quality of the samples where a read assignment percentage over 50% is preferable for a reliable and good quality dataset. A box plot was generated demonstrating the count distributions of each of our samples which count files were produced for as shown in Figure 4. No significant distribution differences are observable between samples. The density distributions of count intensities of the samples do not deviate notably from the median.

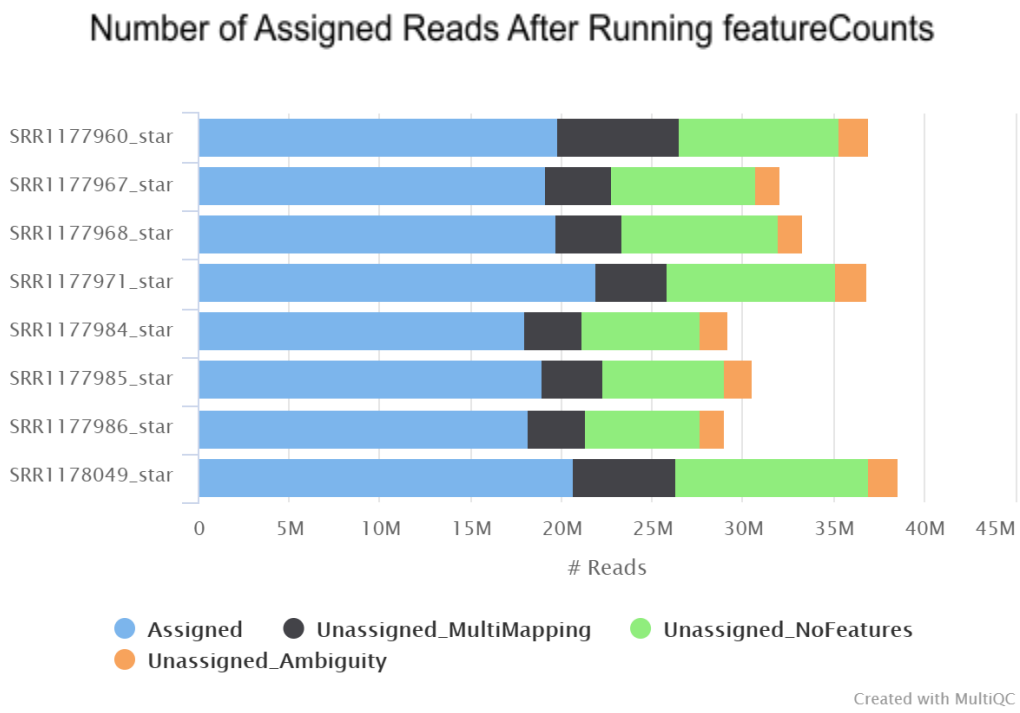


Figure 2: MultiQC plot summary of the count of assigned reads for the treatment samples after running feature Counts. The measurements are reported according to four different categories. The number of assigned reads, in millions, varies for each sample. A range of about 20 to 22 million reads is indicated by the blue bar coloring.

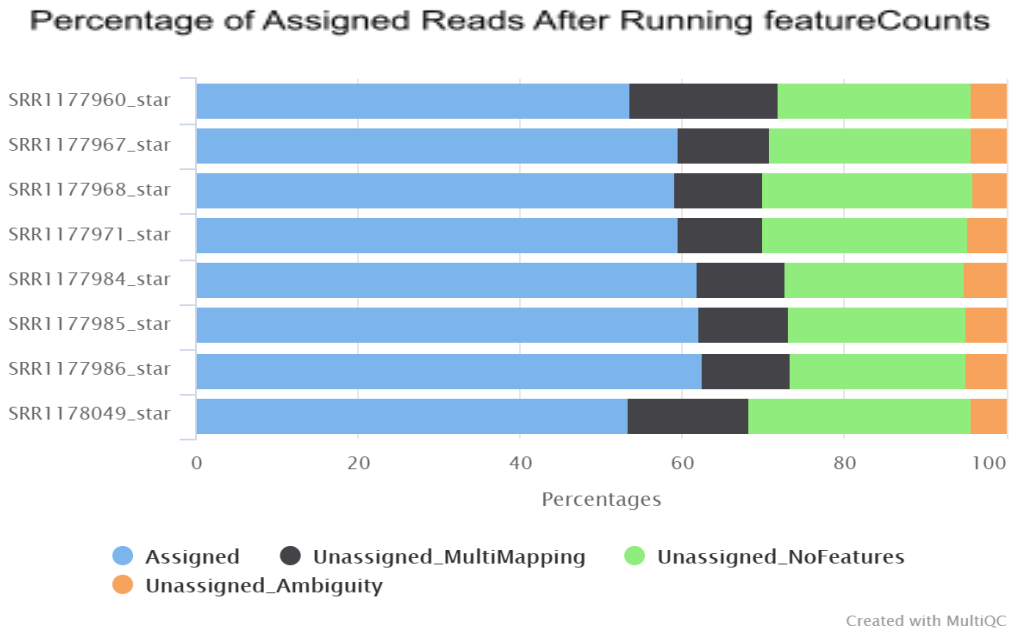


Figure 3: MultiQC plot assessing featureCounts summary of the percentage of reads for treatment samples across four categories. The percentage of assigned reads ranges between approximately 53% to 63% as indicated by the blue bar

coloring. Unassigned read percentages are also reported.

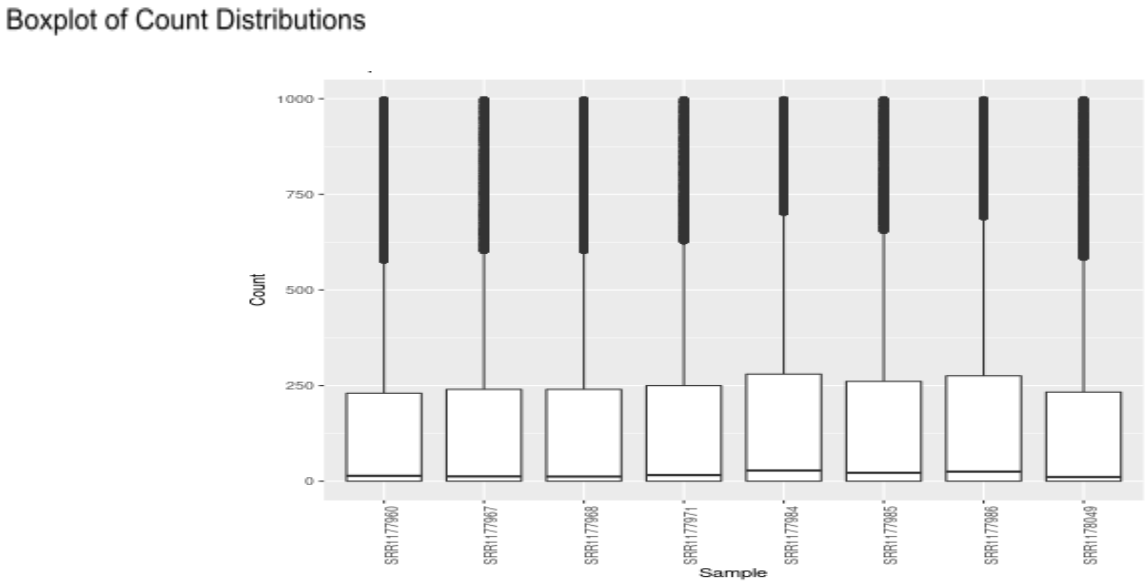


Figure 4: Boxplot of count distribution among the 8 samples for which count files were produced. Counts are largely uniform among each of the samples. The horizontal black line across each box shown in the figure represents the median of counts.

RNA-Seq Differential Expression Analysis Using DESeq2

The DESeq2 Bioconductor R package (version 1.30.1) was used to examine differential expression between both our treatment and control samples. A combined counts matrix was created with corresponding featureCounts and control samples. DESeq2 objects were established for our three different treatment toxin types: N-nitrosodimethylamine, beta-estradiol, and bezafibrate, and their appropriate controls depending on shared vehicle type: saline or corn oil. To model the DESeq2 matrix, the modes of action (MOA) associated with each of these three toxins were utilized. The three MOA are DNA damage (DNA), estrogen receptor (ER), and peroxisome proliferator-activated receptor alpha (PPARA) and correspond to the defined treatment chemicals, respectively. After running DESeq2 on this design matrix, three differentially expressed gene lists were generated and sorted by adjusted p-value. Gene counts were calculated for all three conditions at an adjusted p-value threshold of less than 0.05. These counts are reported in Table 1. The top ten genes for each MOA condition as sorted by p-value were also identified and reported in Table 2. Histograms demonstrating log fold changes among significant genes were plotted for each MOA group. The ggplot2 package (version 3.3.3) was used to create histograms and scatter plots for significantly differentially expressed genes according to fold change in each MOA group.

Microarray Differential Expression Analysis Using limma

limma is a Bioconductor package that helps in performing the standard analytical methodology for microarray differential expression analysis. The authors from the paper used RMA normalization which is a normalisation procedure for microarrays that background corrects, normalises and summarises the probe level information without the use of the information obtained in the MM probes. After that used limma to determine differential expression between the different treatments and control samples.

By utilizing the limma package for the group 4, containing chemicals such as n-nitrosodimethylamine, beta-estradiol, and bezafibrate. The list of differentially expressed genes were generated. As limma is the R package, thus R --version(4.0.2) was used. The data was then filtered using the adjusted_P-value< 0.05.

To identify the significant top 10 differentially expressed hits from the filtered dataset, the data was sorted based on the p value and results in the desired genes. This process was being utilized for all the three chemicals. The following information is reported in Table 4. Moreover, histograms were generated for visualizing the distribution of log fold change across all the samples.The graphical representation is reported in Figure7 Along with that scatter plot was constructed to understand the relationship between P_value and the log fold change.The visualization for the scatter plot can be found in Figure8**.**

Concordance between Microarray and RNA-Seq DE genes

The main objective for conducting this research is to find the concordance between two methods i.e. RNA-Seq and Microarray and these estimates were dependent on a number of factors, including biological effect size and gene expression level. Thus deseq and limma, differential expression results were utilized for finding the concordance between these two methods.

Concordance was calculated using the following formula with agreement in the directionality of the fold change.

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

Before computing the concordance the background-corrected intersection between the two sets was performed. As the background intersection increases with the increase in the size sets. Thus background-corrected intersection was utilized for the sets which may not be independent of each other. The formula used to find the background-corrected intersection:

$$x + \frac{(n_1 - x) \times (n_2 - x)}{N - x} = n_0$$

In the above equation: the differentially expressed genes were obtained by using Deseq for RNA-Seq and limma for Microarray. Then the filter was performed on the adjusted_P-value<0.05. The following steps were performed where merging of these two data sets was executed, leading to the generation of one table that only contains the genes that are common between these two methods. Further, filtering was done based on the directionality of the fold change. Finally, the data set now left with the genes that are common between the two methods, passed the filter and have the same direction based on logFC. Therefore, n0 represents the number of genes that are common between the RNA-Seq and the Microarray. n1 represents the total number of genes obtained from DESeq2 after applying the filter on adjusted_Pvalue<0.05, n2 represents the total number of genes obtained from limma after applying the filter on adjusted_Pvalue<0.05, N was the total number of genes in the rat genome which is equal to 25,000 genes.

As the value of n1, n2 , N is known. x can be calculated. Thus the following formula can be used to find the background-corrected intersection:

X = (*n*0 × *N* − *n*1 × *n*2)/(*n*0 + *N* − *n*1 − *n*2)

After getting the value of x, concordance in percentage was calculated using the formula:

Concordance = (2*(x))/(n1+n2)*100

The calculation mentioned above was performed on all the three chemical to find concordance: For the overall set, For the above median set, that was calculated based on the baseMean column in the DESeq2 results, which corresponds to the overall mean count of the gene across all samples in the comparison. Thus, the median was calculated on the basis of the baseMean column and the above set was generated with all genes that were greater than the median. In the same way concordance was also being calculated for the set containing genes which have baseMean below the median. The following information is described in the tabular format can be found in Table 5.

RESULTS & DISCUSSION

Transcript Abundance

The featureCounts tool was used to generate counts of the nine sample files generated by the STAR aligner. A total of eight sample files were produced by the featureCounts batch job. One sample, SRR1178023, did not output any count files, and consequently, no MultiQC quality assessment summary either. It is unclear why featureCounts failed for only this sample, however it is possible that foreign characters may have polluted the original FASTQ sample or that the SAM file was incomplete. The same problem was reported by others who are also working with this sample, so a problem in our methodology is unlikely. Otherwise, featureCounts ran successfully on the remainder of the samples and generated a total of eight count files. After running MultiQC, the quality report shows read percentages of about 60% for each sample against the gene annotation file and suggests a sufficient level of quality for downstream analysis.

RNA-Seq Differential Gene Expression

After running DESeq2 on the toxin treatment samples, significantly differentially expressed genes were identified. Both the total number of differentially expressed genes and significantly expressed genes with adjusted p-values less than 0.05 are reported in Table 1. The total number of differentially expressed genes was 11,019 for each MOA group. Gene counts were compared to those provided by the paper in Figure 2A [1]. The greatest amount of significantly differentially expressed genes is seen for the DNA group corresponding to the N-nitrosodimethylamine (NIT) chemical with a total of 4,384 genes. This value is slightly larger than the number of differentially expressed genes for NIT reported in the paper which ranges around approximately 3,700 genes. The number of significant genes found for ER which corresponds to the beta-estradiol chemical (BES) is 2,627, however a corresponding gene count was not found within the paper. Significantly expressed genes found for PPARA which corresponds to bezafibrate (BEZ) is approximately 2,838 which is within the same range as that reported by the paper. Consistencies among the significant gene counts with that of the paper offers some confidence that RNA-seq was performed successfully.

Mode of Action (MOA)	Total Differentially Expressed Genes	Significant Differentially Expressed Genes
DNA Damage (DNA)	11,019	4,384
Estrogen Receptor (ER)	11,019	2,627
Peroxisome Proliferator-Activated Receptor Alpha (PPARA)	11,019	2,838

Table 1: Total number of differentially expressed genes and those with adjusted p-values of less than 0.05 for each MOA group as determined by RNA-seq. DNA, ER, and PPARA correspond to the toxins N-nitrosodimethylamine, beta-estradiol, and bezafibrate, respectively. The largest count is seen for the DNA group, otherwise considered as the N-nitrosodimethylamine-treated sample.

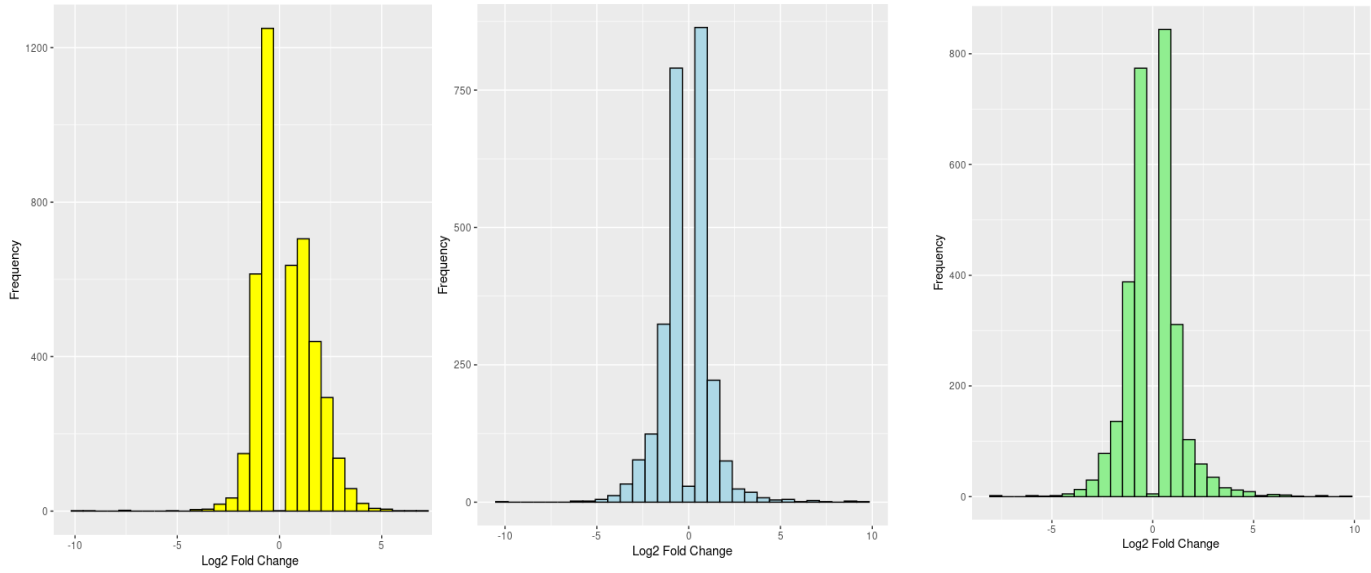
The top 10 differentially expressed genes and their nominal p-values were identified for each toxin or MOA treatment group and are reported in Table 2. RefSeq accessions for each transcript were searched for in external databases to map to a gene name, and then inputted into the table.

	N-nitrosodimethylamine (NIT)		Beta-estradiol (BES)		Bezafibrate (BEZ)	
Rank	Gene	P-Value	Gene	P-Value	Gene	P-Value
1	UGT2B	2.33E-157	CAR3	0	APOA4	2.18E-114
2	FABP4	1.57E-93	SLC5A1	2.05E-217	EHHADH	6.49E-114
3	PLA2G7	1.69E-86	CITED4	1.12E-191	FABP3	3.65E-59
4	OAT	1.65E-86	ARHGEF28	1.29E-154	ATP2B2	5.22E-57
5	ABCB1B	2.94E-76	BAIAP2L2	9.52E-153	CYP4A1	5.11E-50
6	LAMA5	1.38E-72	LIFR	2.59E-111	GRIN2C	8.43E-47
7	HMOX1	2.62E-68	AKR1B8	3.16E-95	LPL	3.56E-46
8	TUBB6	4.32E-65	CTR9	1.74E-89	ACAA1B	5.84E-42
9	CXCL9	2.28E-58	CAR12	4.02E-89	CYP2C7	1.18E-41
10	NCF1	2.09E-56	FADS1	1.56E-88	DECR1	5.54E-41

Table 2. Top 10 differentially expressed genes found via RNA-seq analysis listed for each toxin treatment group as indicated by ascending nominal p-values < 0.05.

Histograms were generated from the significant differentially expressed genes for each treatment group. 3 histograms are shown in Figure 5 for each MOA group. Commonalities are seen among Figure 5B and 5C. Gene frequencies of about 750-800 are within close range to each other for both groups. Additionally, log fold distribution patterns are comparable among both groups. Alternatively, Figure 5A, which represents the DNA MOA, demonstrates a much larger significant gene frequency over 1200. Overall, the log-fold change distribution is varied across the experimental groups.

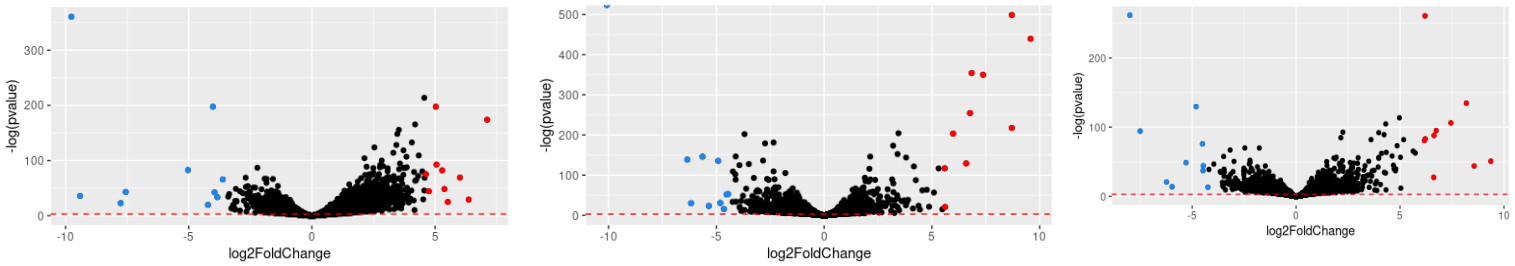
There are a number of steps that may have resulted in the variation between our findings and that of the paper's, with regards to the differential gene counts and distribution patterns. Given that the researchers mainly used limma for both the RNA-seq and microarray data while we used DESeq2, sources of variation may be likely in the RNA sequencing data measurements. Despite this discrepancy, it is not a concern for impacting later downstream analysis with concordance value calculation.



a) b) c)

Figure 5. Distribution of log fold change values for significantly differentially expressed genes as identified through RNA-seq analysis at adjusted p-value < 0.05 for **(A)** DNA, **(B)** ER, and **(C)** PPARA. Similarities can be observed between Graphs B and C, the ER and PPARA groups, in terms of both log fold change distribution and approximate gene frequency.

As an additional statistical graphic, scatter plots were also generated for each treatment group. Log fold changes were plotted against nominal p-values as shown in Figure 6. About the same amount of up- and down-regulated genes are observed in the ER and PPARA plots, Figure 6B and 6C. Alternatively, the N-nitrosodimethylamine-treated, or DNA damage, group demonstrates greater presence of up-regulated gene expression than the other groups. This is consistent with the histogram data in Figure 5 which also reflects the same findings as that suggested by the scatter plots.



a) b) c)

Figure 6. Scatter plots of the significant genes from each MOA as determined by RNA-Seq analysis: (A) DNA damage, (B) ER, and (C) PPARA. The red dashed line represents the p-value threshold of 0.05. The up-regulated and down-regulated genes are marked by red and blue colored indications, respectively.

Microarray Differential Expression Analysis

Counts for the significantly expressed genes varied between the microarray and RNA seq data.

Chemical Name	No. of Differentially Expressed (DE) Genes	Significant No. of Expressed Genes (adjusted P-value< 0.05)	Top 10 DE genes
N-nitrosodimethylamine	31,099	377	Mybl1 Abcb1a Abcb1b Ccng1 Lama5 RGD1561849 TR1-CE5 Atp6v1d Mdm2
Beta-estradiol	31,099	102	Rgs3 Orm1 Hist1h4b Rbp7 Tsku Hgf Zmiz1
Bezafibrate	31,099	2823	Cyp4a1 Cyp4b1 Acot1 Acot2 Acot2 Acot3 Acox1 Pex11a Ehhadh Pex11a Crat

Table 3: Overall differential gene results obtained using the limma package. As the Group 4 contains three chemicals. Each row represents one chemical, following the number of differential genes obtained as the end result of limma. To get the significant genes the filter was applied on adjusted P-value and the list of top 10 genes were obtained based on the P-value from the filtered dataset.

a)

probe ID	gene	logFC	AveExpr	t	P.Value	adj.P.Val	B
1368934_at	Cyp4a1, Cyp4b1	2.528	12.733	25.282	1.31E-25	4.09E-21	45.091
1398250_at	Acot1	8.372	8.776	23.793	1.19E-24	1.85E-20	43.345
1391433_at	Acot2	2.823	10.370	22.519	8.65E-24	8.96E-20	41.735
1388210_at	Acot2	3.368	8.151	19.803	8.35E-22	6.49E-18	37.904
1378169_at	Acot3	4.396	8.357	19.453	1.56E-21	9.70E-18	37.368

1367680_at	Acox1	1.838	13.026	17.495	6.21E-20	3.22E-16	34.154
1379361_at	Pex11a	2.853	10.091	17.047	1.51E-19	6.70E-16	33.367
1368283_at	Ehhadh	3.214	12.845	16.771	2.63E-19	1.02E-15	32.871
1387740_at	Pex11a	3.202	8.513	16.692	3.09E-19	1.07E-15	32.727
1371886_at	Crat	2.403	10.048	15.716	2.36E-18	7.34E-15	30.901

b)

probeID	gene	logFC	AveExpr	t	P.Value	adj.P.Val	B
1392754_at	NA	5.322	6.000	41.169	1.67E-13	5.21E-09	14.206
1385132_at	Mybl1	5.034	6.287	29.923	5.63E-12	8.76E-08	13.265
1370583_s_at	Abcb1a, Abcb1b	4.418	7.305	22.530	1.26E-10	1.30E-06	12.014
1367764_at	Ccng1	2.092	8.452	18.572	1.02E-09	7.94E-06	10.914
1388932_at	Lama5	1.809	5.368	17.431	2.02E-09	1.26E-05	10.509
1390317_at	RGD1561849	3.438	5.664	17.097	2.49E-09	1.29E-05	10.381
1388255_x_at	TR1-CE5	2.518	6.076	14.657	1.29E-08	5.72E-05	9.301
1388325_at	Atp6v1d	1.496	9.970	13.934	2.20E-08	8.10E-05	8.923
1384449_at	NA	2.306	5.332	13.851	2.34E-08	8.10E-05	8.878
1383288_at	Mdm2	1.656736	8.260799	13.43678	3.23E-08	1.00E-04	8.64484

c)

probeID	gene	logFC	AveExpr	t	P.Value	adj.P.Val	B
1367957_at	Rgs3	2.309	9.564	6.978	2.16E-08	5.11E-04	8.638
1368731_at	Orm1	1.188	13.544	6.846	3.29E-08	5.11E-04	8.277
1396155_at	NA	2.164	10.016	5.856	7.89E-07	8.18E-03	5.519
1376089_at	NA	1.034	10.859	5.589	1.86E-06	1.45E-02	4.770
1369728_at	Hist1h4b	-0.382	4.689	-5.517	2.34E-06	1.46E-02	4.568
1377222_at	NA	1.010	6.582	5.410	3.30E-06	1.49E-02	4.266
1374863_at	Rbp7	3.246	6.059	5.404	3.36E-06	1.49E-02	4.250
1383315_at	Tsku	1.720	10.225	5.362	3.85E-06	1.50E-02	4.132
1387701_at	Hgf	-0.745	5.358	-5.272	5.14E-06	1.74E-02	3.879
1373369_at	Zmiz1	0.932	8.498	5.244	5.61E-06	1.74E-02	3.801

Table 4: Top 10 significant differentially expressed genes (affyID) with the lowest P-values: A) The list of genes obtained from the chemical Bezafibrate. B) The list of genes obtained from the chemical N-nitrosodimethylamine. C) The list of genes obtained from the chemical Beta-estradiol.

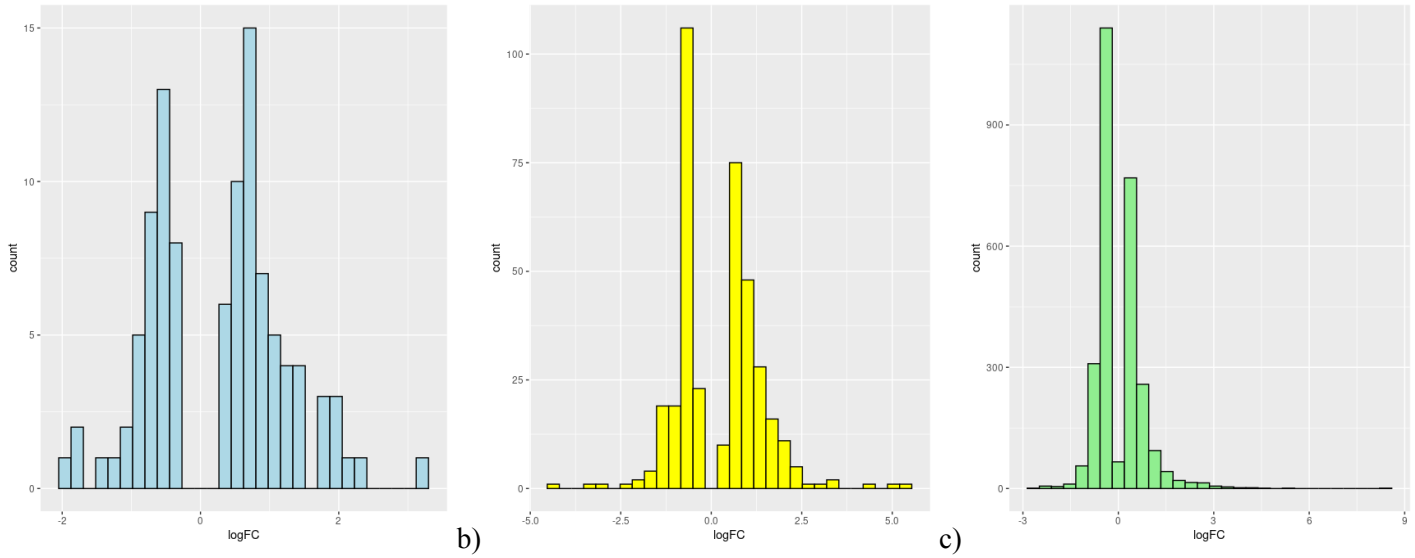


Figure 7:: Histogram representing the fold change values obtained from the filtered data set(adjusted_P-value<0.05) data was obtained using the table generated from the limma output and then filters were applied a) Chemical Beta-estradiol. b) Chemical N-nitrosodimethylamine. c) Chemical Bezafibrate.

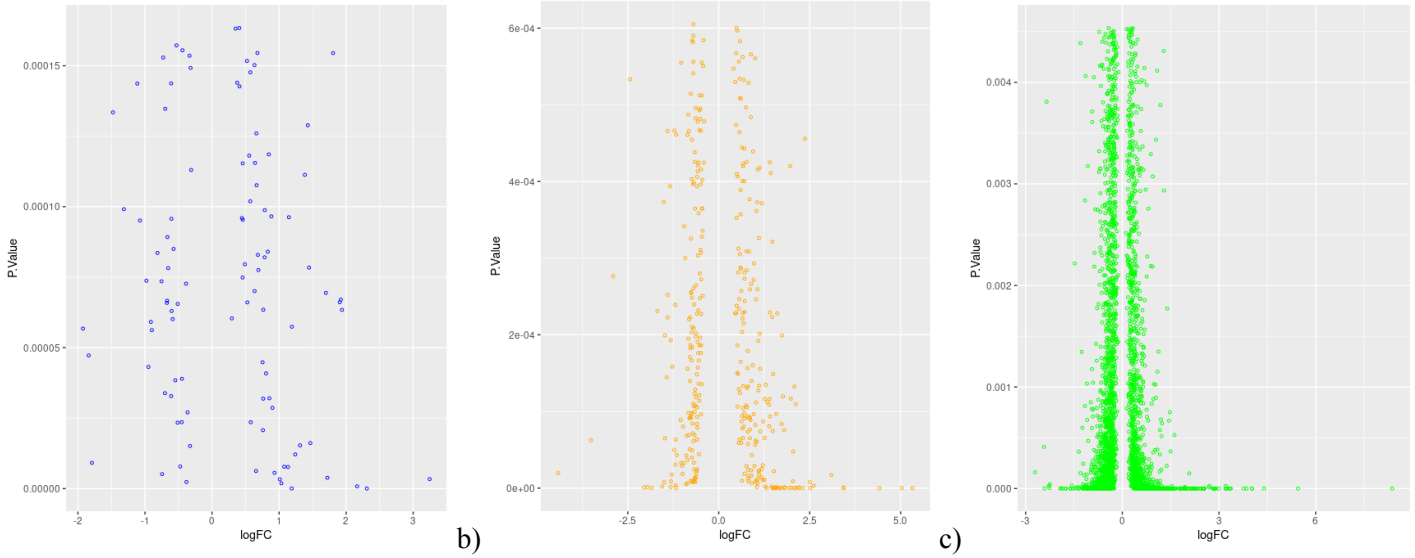


Figure 8: The scatter plot was generated using the filtered(adjusted_P-value) data set from the limma package for the logFold Change V/S the P-value. a) Plot represents the Chemical Beta-estradiol as we can see the plot follows a similar pattern like other two with some outliers. However, the density is low. This is because after applying the filter the original data got reduced to 102 total number of genes. b) Plot represents the Chemical N-nitrosodimethylamine, as the density of the points is good this is because of the 377 total number of genes obtained after applying the filter. c) Plot represents the Chemical Bezafibrate, highly dense among all the three chemicals because the number of points that pass the fiter is 2823.

Chemical Name	Above-Median Concordance(%)	OverallConcordance(%)	Below-MedianConcordance(%)
N-nitrosodimethylamine	8	7	4
Beta-estradiol	4	4	1
Bezafibrate	51	43	20

Table 5: Representation of the concordance for all the chemicals present in selected group 4. The overall result based on the dataset selected referring to the median. Concordance Results in the above and below median datasets and overall dataset. As, all the three chemicals follow a very similar pattern with greater concordance in above-median group and lowest concordance in the below-median group.

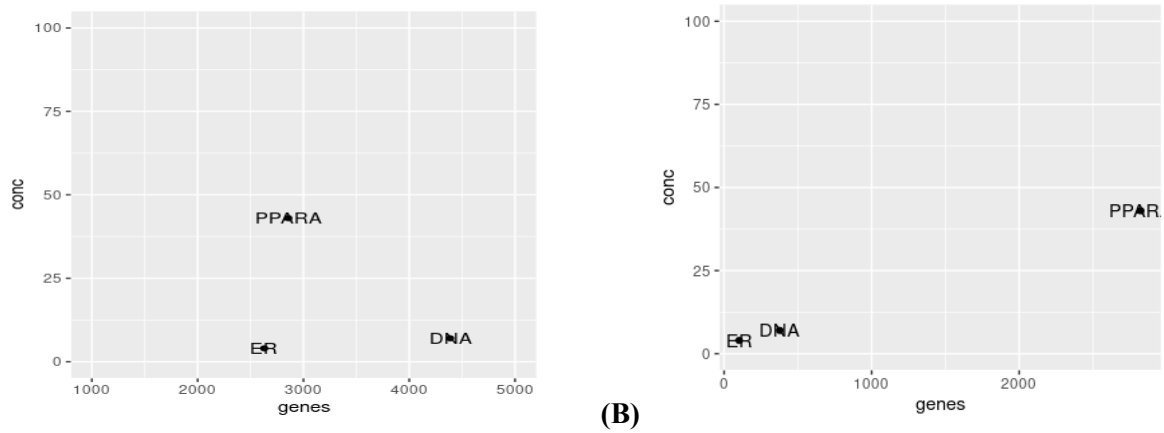


Figure 9: (A) Plots representing the concordance vs the number of DE genes from the RNA-Seq analysis. (B) Plot representing concordance vs number of DE genes from the microarray analysis.

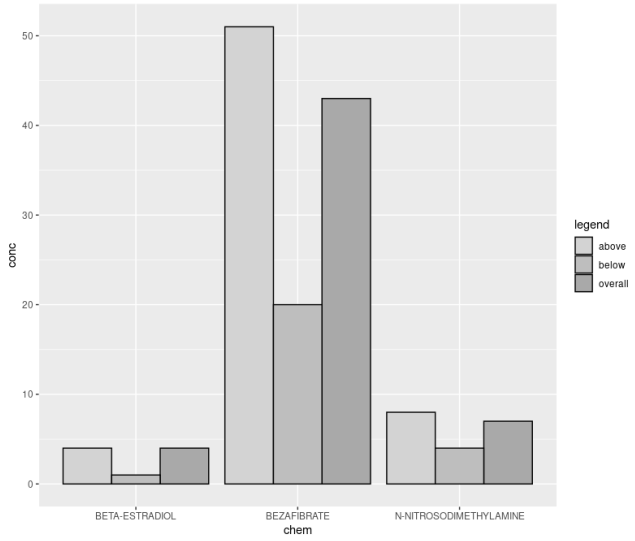


Figure 10: Histogram representing the concordance for all the three chemicals with overall, above-median and below-median concordance.

For each chemical in toxicology group 4, the differential expressed genes were found using limma as sorted by adjusted p-value < 0.05. The number of genes that pass the filter was lowest in Beta-estradiol as only 102 genes successfully passed through it. Moreover, the maximum number of genes passed the filter was in chemical Bezafibrate with 2823 genes. Figure- and - depict the Histogram representing the count with respect to logFC and scatter plot helps in visualizing the fold change with respect to the p-value from the significant differentially expressed gene set.

After performing the concordance methodology, it is observed that Bezafibrate has the highest concordance among all the other chemicals. As mentioned previously, this chemical corresponds to the PPARA MOA. Upon review of the paper, our finding is validated given that the concordance level for the PPARA MOA is very similar in range to the value reported by our analysis [1, Figure 2D]. However, it is also worth noting that the researchers from the paper considered 3 chemicals for each MOA which resulted in the reported concordance values. Given our project focuses on toxicology group four, we only had one chemical per MOA. Nonetheless, it is likely that the concordance values reported in the paper are averages

taken among all three chemicals for each MOA. Thus, our single concordance value for Bezafibrate is within the likely average range of expected values for PPARA. It is also observed that the concordance level was higher in the above-median for all the three chemicals whereas the below-median concordance showed lowest among all.

The differences in the concordance between the two methods might have occurred due to various reasons such as difference in handling the experiment, various environmental conditions like temperature, humidity, difference in following the protocol. These errors might be responsible for the difference in the concordance between RNA-Seq and Microarrays method for getting differential expressed genes. The findings we generated are very close to what is presented in the original paper. The one evidence to prove this is the concordance level for Bezafibrate.

It is worth noting that the researchers make a generalizing assumption while utilizing varying chemicals with distinct modes of action to assess the concordance between microarray and RNA-seq technologies. This assumption implies that chemicals with similar MOA would consequently have shared gene expression responses. However, the biological mechanisms which determine the likely molecular response very likely vary between chemicals. Nonetheless, concordance between RNA-seq and microarray data can still be examined with some reliability if the same chemicals are administered at the same dose. Had an enrichment analysis been performed, this concern would be better addressed and evaluated.

CONCLUSION

For our analysis, we examine the reproducibility of the findings by Wang et al. for toxicology group four. The subset of chemicals: N-nitrosodimethylamine, beta-estradiol, and bezafibrate make up our treatment groups. Differential gene expression between these groups was compared between the microarray and RNA-seq methodology.

To assess the comparability of both sequencing technologies, concordance calculations were performed. Moreover, there were differences observed in the concordance between both the methods that could have possibly occurred due to various reasons such as difference in handling the experiment, various environmental conditions like temperature, humidity, difference in following the protocol. These errors might be responsible for the difference in the concordance between RNA-Seq and Microarrays method for getting differential expressed genes.

We believe our findings reflect those which are presented in the original paper. Consistencies between the concordance level for Bezafibrate in our project and the research study suggest proper methodology steps were performed. However, given one of our samples did not produce a count file for one of the three beta-estradiol samples for later RNA-seq analysis, a conclusive comparison cannot be drawn with regards to the ER pathway.

Finally, we could conclude based on the results that there is high concordance found in the above-median data set for all the three same when compared with below-median and overall data sets. Which shows that our analysis has similar results with the original research.To sum up with,the major difference between RNA-seq and microarrays is their respective sensitivity for the low expressed genes where RNA-seq offers advantages over microarrays. Consequently, RNA-seq outperforms microarrays substantially when profiling two similar biological conditions compared to profiling two distinct biological conditions.

REFERENCES

1. Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., Meehan, J., Li, X., Yang, L., Li, H., Łabaj, P. P., Kreil, D. P., Megherbi, D., Gaj, S., Caiment, F., van Delft, J., … Tong, W. (2014). The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nature biotechnology, 32(9), 926–932. https://doi.org/10.1038/nbt.3001
2. National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2017 Apr 06]. Available from: https://www.ncbi.nlm.nih.gov/
3. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
4. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635
5. Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). Multiqc: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics, 32*(19), 3047-3048. doi:10.1093/bioinformatics/btw354.
6. 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.
7. Labaj PP, et al. *Bioinformatics.* Vol. 27. Oxford, England: 2011. Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling; pp. 383–391. [PMC free article] [PubMed] [Google Scholar]
8. Mooney M, et al. Comparative RNA-Seq and microarray analysis of gene expression changes in B-cell lymphomas of Canis familiaris. *PloS one.* 2013;8:e61088. [PMC free article] [PubMed] [Google Scholar]

sample	No Input Reads	Average Input Read Length	Uniquely Mapped Reads	Reads Mapped to Multiple Loci	Unmapped Reads	Mismatch rate per base	% unmapped reads (total)
SRR1178049	19397953	200	16448173	1041082	1860264	0.92%	9.59%
SRR1177960	17947778	100	15080805	1189810	1635043	0.45%	9.11%
SRR1177967	17157413	202	14253911	673010	2199580	0.77%	12.82%
SRR1177968	17604520	202	14857439	686056	2031562	0.77%	11.54%
SRR1177971	19627402	202	16518020	732186	2325847	0.83%	11.85%
SRR1177984	15732068	202	12999695	586314	2115963	0.77%	13.45%
SRR1177985	16537337	202	13601134	630871	2280499	0.69%	13.79%
SRR1177986	15559784	202	12936410	600101	2000988	0.72%	12.86%
SRR1178023	16076957	200	13486519	915705	1655927	1.13%	10.30%

Table 6. Summary of read and alignment statistics from STAR output. % unmapped reads (total) is a combined percentage of "% of reads unmapped: too short" and "% of reads unmapped: other" from STAR. Column labeled "Unmapped Reads" is calculated using the % unmapped reads (total) and No. Input Reads.