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

Group 1

TA : Kritika Karri

Divya Venkatraman (Analyst)
Garima Lohani (Programmer)
Marlene Tejeda (Biologist)
Xudong Han (Data Curator)

INTRODUCTION:

Gene expression can be measured by two processes RNA-seq and microarray. To study drug evaluation and safety evaluation, microarray was the principal technology in the past decade. However, in the current studies high throughput sequencing technology is used to analyse gene expression. The aim of Wang et al was to study similarities and differences between RNA-seq and microarray analysis[3]. In this study, there was comparison of liver transcriptional responses in rats exposed to different chemicals which were analyzed by microarray and RNA-seq technology. In this project, we try to identify which of the sequencing technology has advantage over the other in identifying differentially expressed genes.[3]

DATA:

In this project, the data we need to process is partly from the Illumina raw RNA-seq data generated from the liver samples of rats under varying degrees of perturbation by 27 chemicals representing multiple modes of action(MOA) such as cytotoxicity (Cytotoxic), orphan nuclear hormone receptors(CAR/PXR), and aryl hydrocarbon receptor(AhR) [3]. Each chemical was administered to male Sprague-Dawley rats with 5 day maximum tolerated doses(MTD) proposal[4] and the RNA samples for RNA-seq were derived from the NTP DrugMatrix Frozen Tissue Library[3]. The data we select is the toxgroup1 which includes three MOAs: AhR which was treated by 3-Methylcholanthrene, CAR/PXR treated by Clotrimazole, Cytotoxic treated by Chloroform(9 samples totally). The Affymetrix microarray data we used to compare were generated from the same liver samples [3] and already processed by the instructor. The average number of reads in each sample is around 18 million, ranging from 16 million and 21.7 million. The average length of each read is between 100 to 101 base pairs. The average aligned rate is around 84% and the average rate of duplicate reads is around 54%. The CG content of these samples is all between 48% and 49%. All the detailed information for the sequencing quality can be found in the result part.

METHODS:

Sequencing quality control statistics and alignment

The RNA-seq sample statistics and alignment were separated into three steps. In the first step, the raw paired-end sequencing data (fastq files)were processed by the fastqc tool installed on SCC. The second step was to align each of the samples against the rat genome using the STAR aligner. Lastly, the processed output from the first two steps was collected and integrated by the tool multiqc. Multiqc generated a final report containing all the information about RNA-seq sample statistics and alignment in an HTML file.

Quantification of gene expression using featureCounts

After the STAR alignment generated the BAM files, the featureCounts were run on these samples. featureCounts is accurate, fast and easy to use tool that counts reads that map to a single location [1]. The 9 BAM file and a rat reference gene annotation file were used as input for featureCounts. The output produced 9 count matrix files with summary files. The count matrix file contains counts read at the genomic feature level. The summary files summarize the assigned and unassigned reads.

Multiqc was run on the 9 samples to check the quality of the assignments. As stated previously, it combines reports from different tools into a single report. The individual count matrix files from each sample were combined into single CSV files. After that the boxplot was produced corresponding to each sample.

Differential expression analysis using DESeq2

DESeq2 is a statistical tool to analyse count data that uses negative binomial regression [2]. To estimate the differential expression between treated samples and their controls, the single csv (the combined count matrix file from nine samples) were combined with counts of the control. Differential gene expression was studied for three different treatment conditions namely ahr, car/pxr and cytotoxic.

The DESeq2 analysis was run thrice for each of the three treatments and corresponding DESeq results were obtained from each analysis. The top 10 differentially expressed genes were filtered based on p-value and the number significant genes with p-adjusted value <0.05 were estimated. The scatter plot and histogram was plotted for the significant genes.

Differential Expression in microarray samples using limma

We used the limma package in R [14] to compute the differentially expressed genes in each of the MOAs by using samples of the chemical corresponding to the MOA and controls of the same vehicle as the samples. Limma uses a linear model fit for the microarray samples and computes moderated t-statistics, moderated F-statistic, and log-odds of differential expression by empirical Bayes moderation of the standard errors. The p-value is adjusted using Bejamini-Hochberg [15] method. We get the significantly differentially expressed genes by using a cutoff of 0.05 for the adjusted p-value.

Concordance among the DEGs of both platforms

We then computed the concordance between the microarray and RNA-Seq platforms using the results we obtained from DESeq and limma. To do so, we used the genes that were differentially expressed with nominal p-value < 0.05. We mapped the refseq ids to the probeset id using a matrix provided that mapped refseq id to probe ids and gene symbols.

Some refseq ids mapped to multiple probe ids and vice versa. Since we needed to use the number of DEGs in computing the concordance, there was no requirement of collapsing the fold change or p-values. Hence, we chose only the unique gene symbols represented in each set.

As mentioned in the reference paper [3], concordance is computed using the equation below.

$$\frac{2 * intersect(DEGs\ of\ microarray, DEGs\ of\ RNA{-}Seq)}{(DEGs\ of\ microarray) + (DEGs\ of\ RNA{-}Seq)}$$

We found the intersection of symbols in both sets , and computed concordance between microarray and RNA-Seq for each MOA.

We also computed the concordance between genes of above median expression and below median expression for both platforms . The median expression for microarray was found using the AveExpr returned by the limma results. For RNA-Seq the median was computed using the baseMean values returned by the DESeq results.

Comparison of pathways enriched in different MOA

The top 100 differentially expressed genes were filtered based on p-value and the number significant genes with p-adjusted value <0.05 were used to find enriched pathways with GATHER.

Heatmap of different MOA

The normalized DEseq2 counts for each MOA were used to create a heatmap to show clustering between the different MOA using a function called heatmap.

RESULTS:

Quantity control statistics for the RNA-seq samples

The raw RNA-seq data for 9 samples were processed by fastqc, STAR and multiqc tools in silico. Table 1 shows the general statistics of alignment. Each MOA was treated with one drug and has three replications. The average rate of aligned reads, duplicate reads, GC content, the number of uniquely mapped reads and total sequences are all listed in the table. The SRR1178036_2 sample shows a significant inconsistency with other samples due to its abnormal aligned rate, mapped reads and duplicate reads rate. Additionally, this sample also represents poor quality in many other attributes (Figure1). For example, the Per Sequence Quality Scores of this sample is relatively lower than other samples, and its distribution of Per Sequence GC Content is quite dissimilar with others(Fig1 D and E). For the other samples, only two samples have qualified Mean Quality Scores(Fig1 C) in the level of 100 bp-length read, but all of them have good scores before 60bp, which indicates the sequencing for the first 60 bps in each read is very reliable. Except for the SRR1178036_2 sample, all the other samples have good qualities in Per Base N Content and Adapter Content(Fig1 F, H), while for the sequence duplication levels, none of them have good quality which may indicate some kind of enrichment bias (eg PCR over amplification) (Fig1 G). Since the fastqc analysis was performed to rna-seq data, a little higher duplicate rate is acceptable.

Sample Name	MOA	Drug	% Aligned	M Aligned	% Dups	% GC	M Seqs
SRR1177987	Cytotoxic	Clotrimazole	84.8%	14.8			
SRR1177987_1					56.4%	49%	17.4
SRR1177987_2					53.3%	49%	17.4
SRR1177988	Cytotoxic	Clotrimazole	85.3%	15.7			
SRR1177988_1					55.1%	49%	18.4
SRR1177988_2					51.9%	49%	18.4
SRR1177989	Cytotoxic	Clotrimazole	83.5%	15.7			
SRR1177989_1					50.4%	49%	18.8
SRR1177989_2					46.9%	49%	18.8
SRR1177997	AhR	3-Methylcholanthrene	89.2%	17.6			
SRR1177997_1					59.6%	49%	19.7
SRR1177997_2					58.6%	49%	19.7
SRR1177999	AhR	3-Methylcholanthrene	88.7%	19.4			
SRR1177999_1					60.2%	49%	21.8
SRR1177999_2					58.9%	49%	21.8
SRR1178002	AhR	3-Methylcholanthrene	89.1%	16.8			
SRR1178002_1					58.5%	49%	18.8
SRR1178002_2					57.6%	49%	18.8
SRR1178020	CAR/PXR	Clotrimazole	83.6%	13.4			
SRR1178020_1					54.0%	48%	16.0
SRR1178020_2					51.6%	49%	16.0
SRR1178036	CAR/PXR	Clotrimazole	67.8%	11.4			
SRR1178036_1					55.5%	48%	16.9
SRR1178036_2					39.4%	48%	16.9

SRR1178046	CAR/PXR	Clotrimazole	85.3%	15.1			
SRR1178046_1					54.7%	48%	17.7
SRR1178046_2					53.5%	49%	17.7

Table 1: General Statistics of Alignment Quality

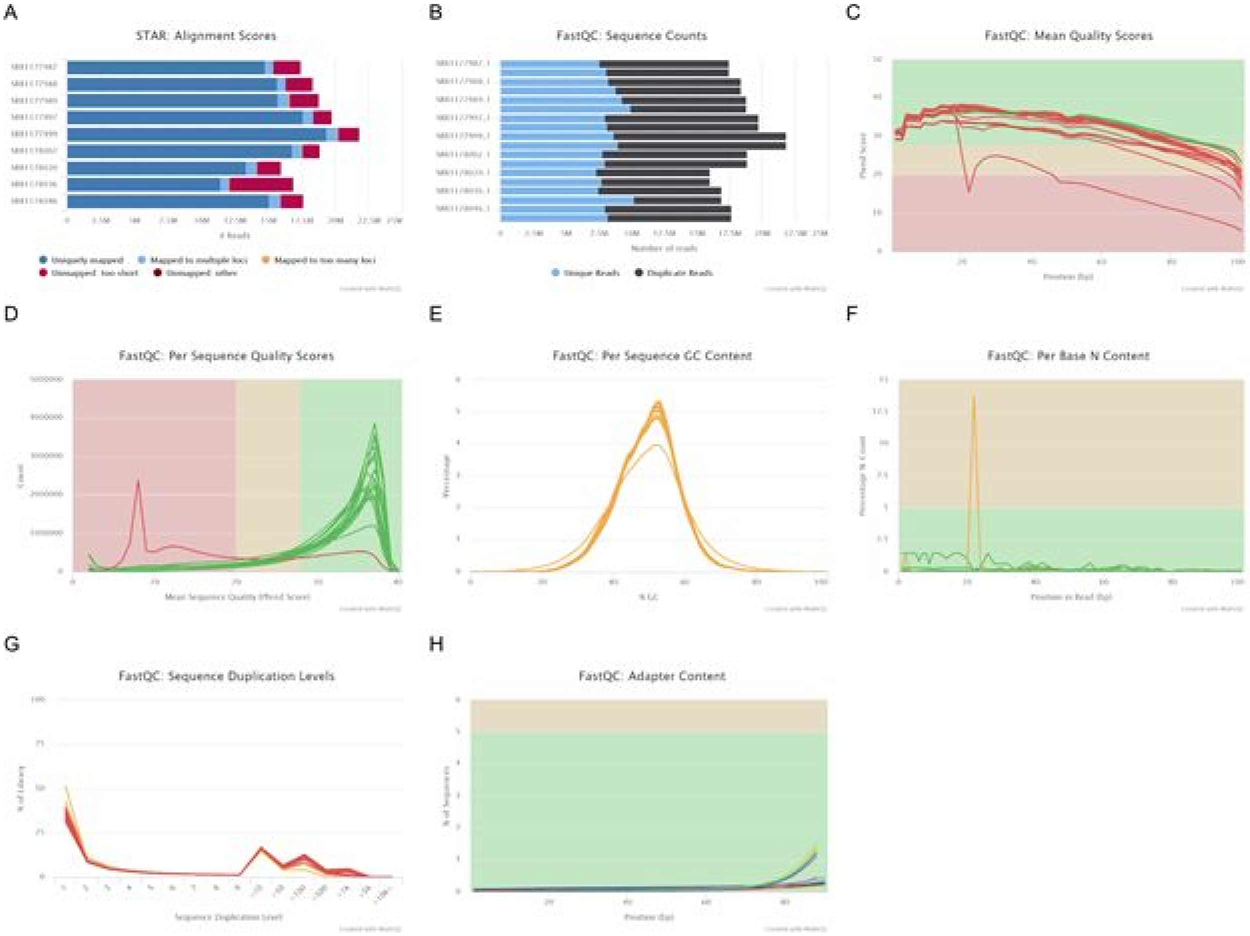


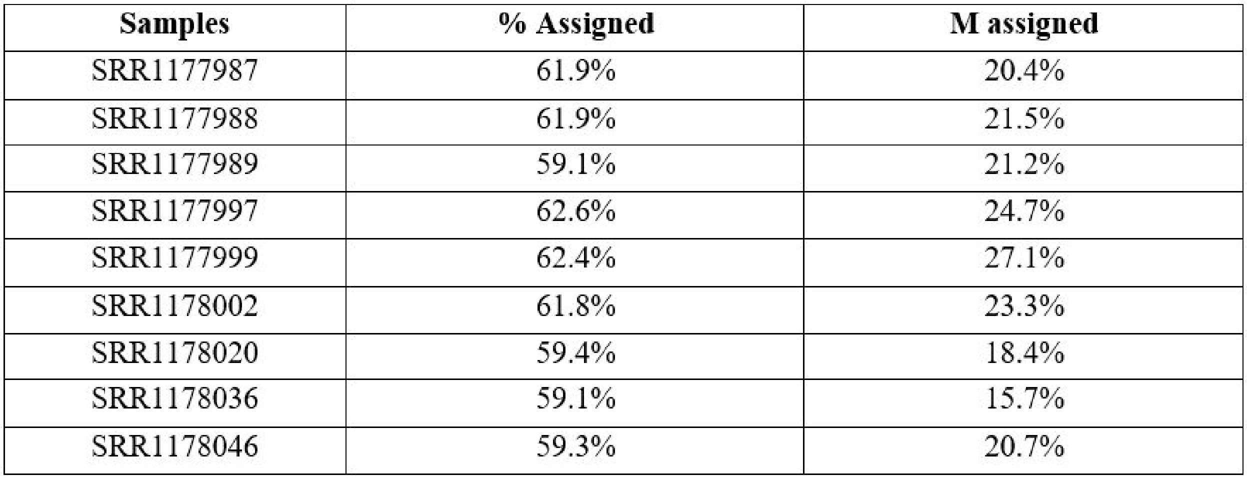
Figure 1: MultiQC reports for the RNA-seq statistics and alignment. The green, red and yellow lines in C-G represent the tracts of samples which are passed, failed, or with warnings respectively in each aspect of assessment. A: STAR alignment scores. All the samples have uniquely mapped reads over 83% except for SRR1178036. B: The read counts for all samples. The blue bar represents the unique reads and the black represents the duplicate reads. C: The mean quality scores. All samples except for SRR1178036 have good quality before 60 bps of each read. D: Per sequence quality scores. All samples have good quality except for SRR1178036. E: Per

sequence GC content. The distribution of SRR1178036 is dissimilar with others. F: Per base N content. G: Sequence duplication levels. All samples have higher duplicate rates. H: Adapter content.

Quantification of gene expression result

We found that on a whole, 60.8% of all the reads were assigned to gene-ids. The highest of reads assigned (nearly 62%) was observed in samples treated with AHR and those samples treated with CAR/PXR had the lowest (close to 59.2%) as shown in Table 2. Further, the green bars indicate the reads that were not assigned to any gene-id in Figure 2a and 2b.

From the boxplots in Figure 3, we found that the density distributions of raw log-intensities are not the same but still not very different. The horizontal blue line represents the median of log counts. We would need to investigate that sample further which is really far above or below the blue horizontal line.The samples SRR117989 showed slight deviation from the median. Altogether, we observed no major distribution differences between the samples.



Samples	% Assigned	M assigned
SRR1177987	61.9%	20.4%
SRR1177988	61.9%	21.5%
SRR1177989	59.1%	21.2%
SRR1177997	62.6%	24.7%
SRR1177999	62.4%	27.1%
SRR1178002	61.8%	23.3%
SRR1178020	59.4%	18.4%
SRR1178036	59.1%	15.7%
SRR1178046	59.3%	20.7%

Table 2: MultiQC summary table which shows percentage of mapped reads assigned to the gene-ids.

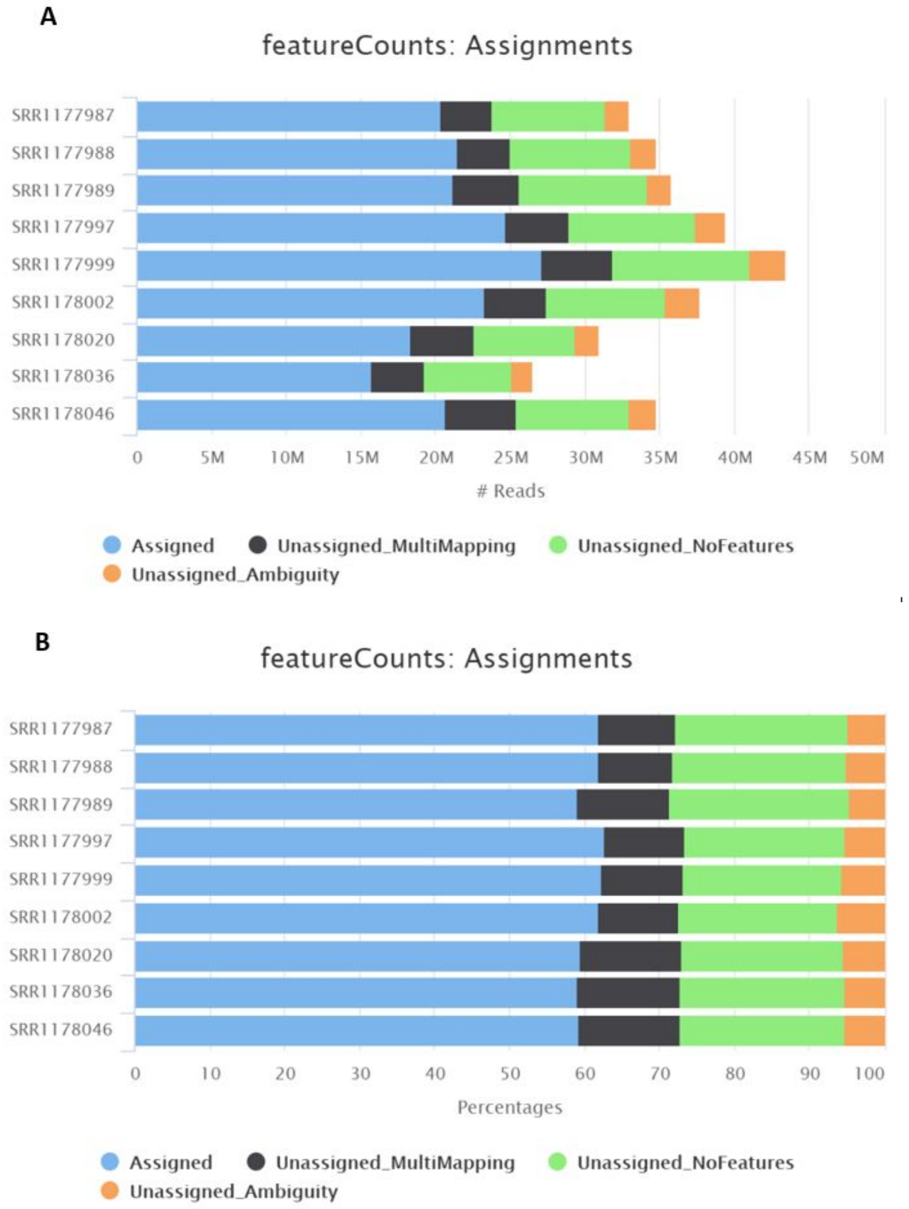


Figure 2: a) number of reads assigned, and b) the percentage of assigned reads

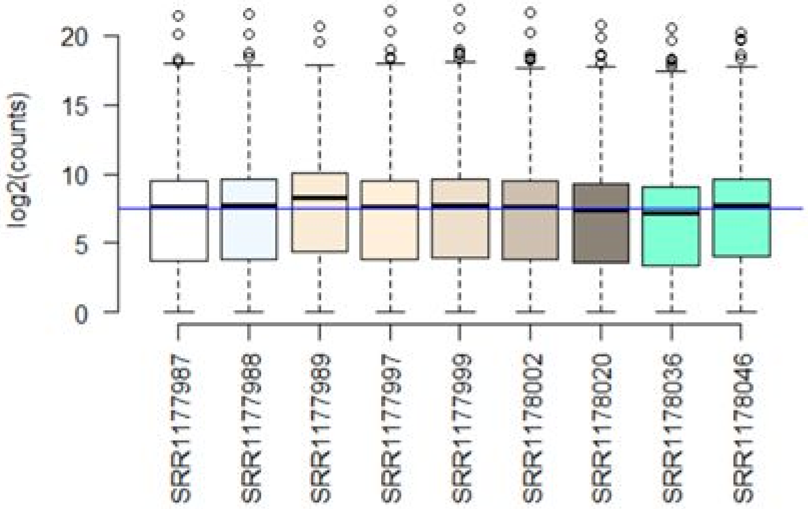
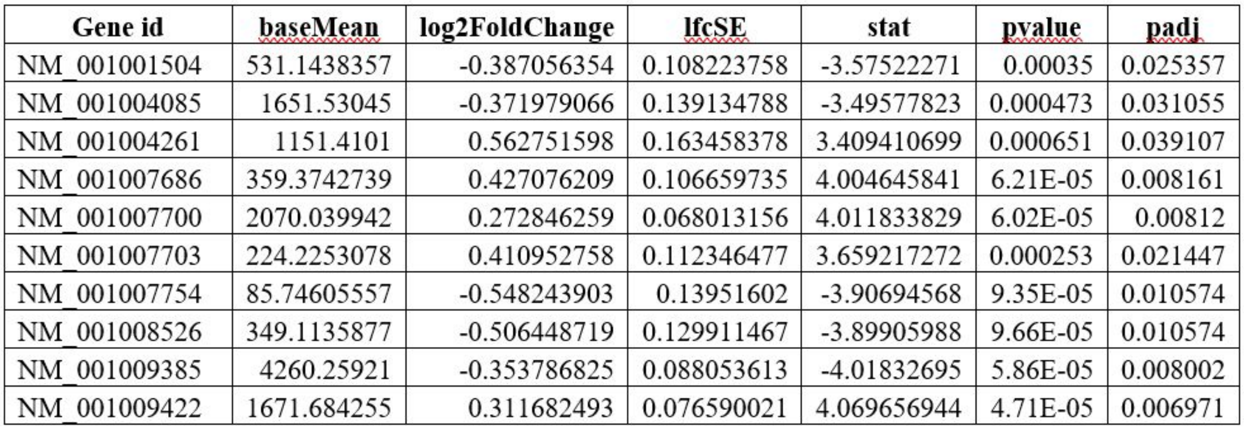


Figure 3: The count distribution of each sample in the box plot

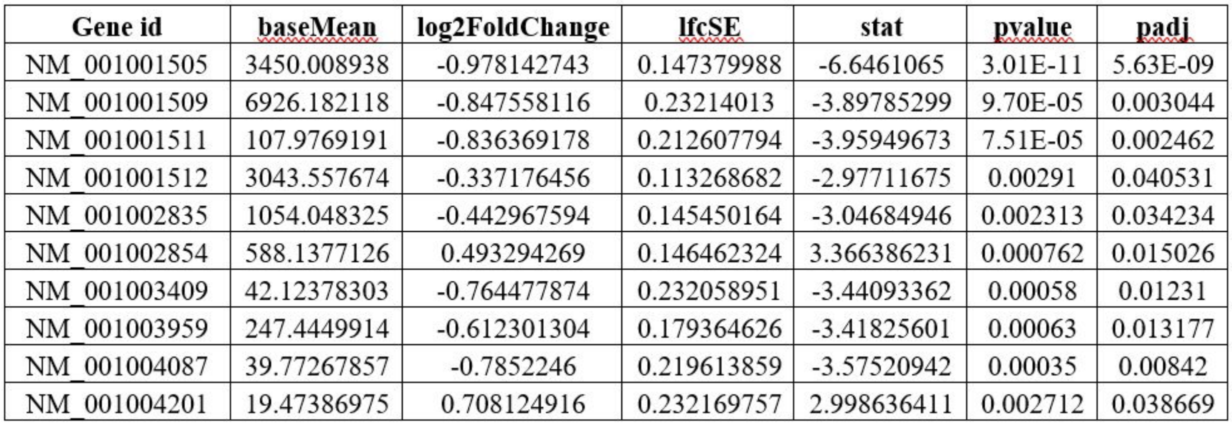
DESeq2 result :

The top 10 differentially expressed genes that were filtered based on the p-value are shown in Table 3 for AHR, Table 4 for CAR/PXR , and Table 5 for Cytotoxic.It was found that the number of genes differentially expressed at an adjusted p-value < 0.05 for the three sample groups treated with AHR, CAR/PXR and Cytotoxic were 193, 860, and 2554 respectively.



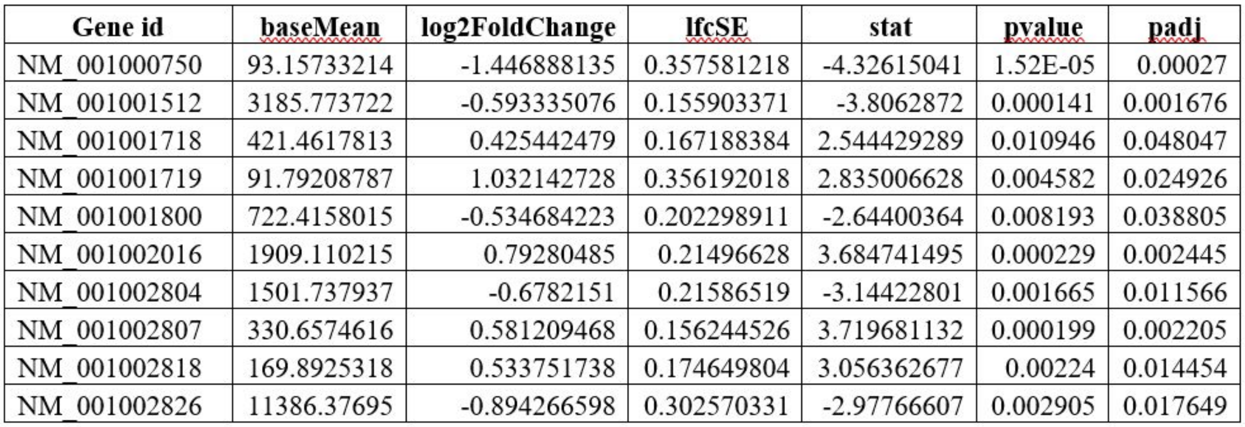
Gene id	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
NM_001001504	531.1438357	-0.387056354	0.108223758	-3.57522271	0.00035	0.025357
NM_001004085	1651.53045	-0.371979066	0.139134788	-3.49577823	0.000473	0.031055
NM_001004261	1151.4101	0.562751598	0.163458378	3.409410699	0.000651	0.039107
NM_001007686	359.3742739	0.427076209	0.106659735	4.004645841	6.21E-05	0.008161
NM_001007700	2070.039942	0.272846259	0.068013156	4.011833829	6.02E-05	0.00812
NM_001007703	224.2253078	0.410952758	0.112346477	3.659217272	0.000253	0.021447
NM_001007754	85.74605557	-0.548243903	0.13951602	-3.90694568	9.35E-05	0.010574
NM_001008526	349.1135877	-0.506448719	0.129911467	-3.89905988	9.66E-05	0.010574
NM_001009385	4260.25921	-0.353786825	0.088053613	-4.01832695	5.86E-05	0.008002
NM_001009422	1671.684255	0.311682493	0.076590021	4.069656944	4.71E-05	0.006971

Table 3 :The top ten significant genes treated with chemical ahr based on p value



Gene id	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
NM_001001505	3450.008938	-0.978142743	0.147379988	-6.6461065	3.01E-11	5.63E-09
NM_001001509	6926.182118	-0.847558116	0.23214013	-3.89785299	9.70E-05	0.003044
NM_001001511	107.9769191	-0.836369178	0.212607794	-3.95949673	7.51E-05	0.002462
NM_001001512	3043.557674	-0.337176456	0.113268682	-2.97711675	0.00291	0.040531
NM_001002835	1054.048325	-0.442967594	0.145450164	-3.04684946	0.002313	0.034234
NM_001002854	588.1377126	0.493294269	0.146462324	3.366386231	0.000762	0.015026
NM_001003409	42.12378303	-0.764477874	0.232058951	-3.44093362	0.00058	0.01231
NM_001003959	247.4449914	-0.612301304	0.179364626	-3.41825601	0.00063	0.013177
NM_001004087	39.77267857	-0.7852246	0.219613859	-3.57520942	0.00035	0.00842
NM_001004201	19.47386975	0.708124916	0.232169757	2.998636411	0.002712	0.038669

Table 4**:**The top ten significant genes treated with chemical car/pxr based on p value



Gene id	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
NM_001000750	93.15733214	-1.446888135	0.357581218	-4.32615041	1.52E-05	0.00027
NM_001001512	3185.773722	-0.593335076	0.155903371	-3.8062872	0.000141	0.001676
NM_001001718	421.4617813	0.425442479	0.167188384	2.544429289	0.010946	0.048047
NM_001001719	91.79208787	1.032142728	0.356192018	2.835006628	0.004582	0.024926
NM_001001800	722.4158015	-0.534684223	0.202298911	-2.64400364	0.008193	0.038805
NM_001002016	1909.110215	0.79280485	0.21496628	3.684741495	0.000229	0.002445
NM_001002804	1501.737937	-0.6782151	0.21586519	-3.14422801	0.001665	0.011566
NM_001002807	330.6574616	0.581209468	0.156244526	3.719681132	0.000199	0.002205
NM_001002818	169.8925318	0.533751738	0.174649804	3.056362677	0.00224	0.014454
NM_001002826	11386.37695	-0.894266598	0.302570331	-2.97766607	0.002905	0.017649

Table 5:The top ten significant genes treated with chemical cytotoxic based on p value

Histogram of log2FoldChange treated with chemicals ahr,car/pxr and cytotoxic is shown in Figure 4a,4b, and 4c respectively.A scatterplot that shows statistical significance (P value) versus magnitude of fold change is a volcano plot. Volcano plots are widely used to display the results of RNA-seq experiments. In a volcano plot (Figure 4a, 4b and 4c), the most upregulated genes are towards the right, the most downregulated genes are towards the left, and the most statistically significant genes are towards the top.

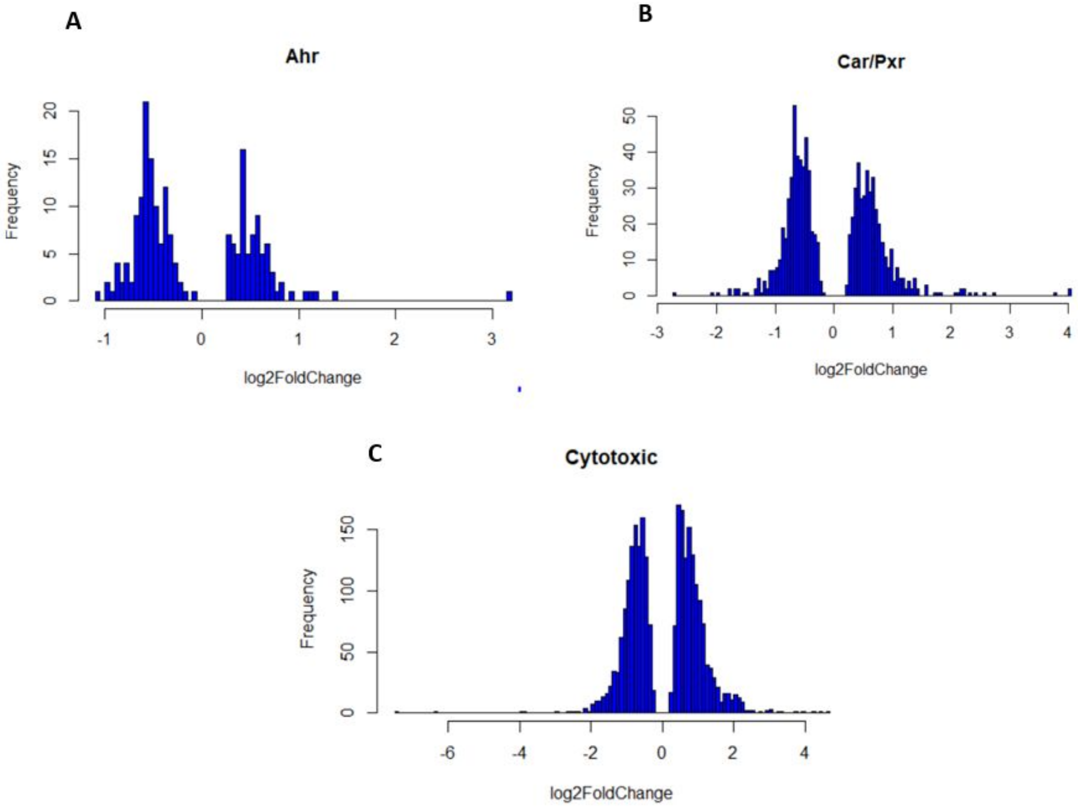


Figure 4 Histogram of log2FoldChange in case of a) ahr, b) car/pxr , and c) cytotoxic

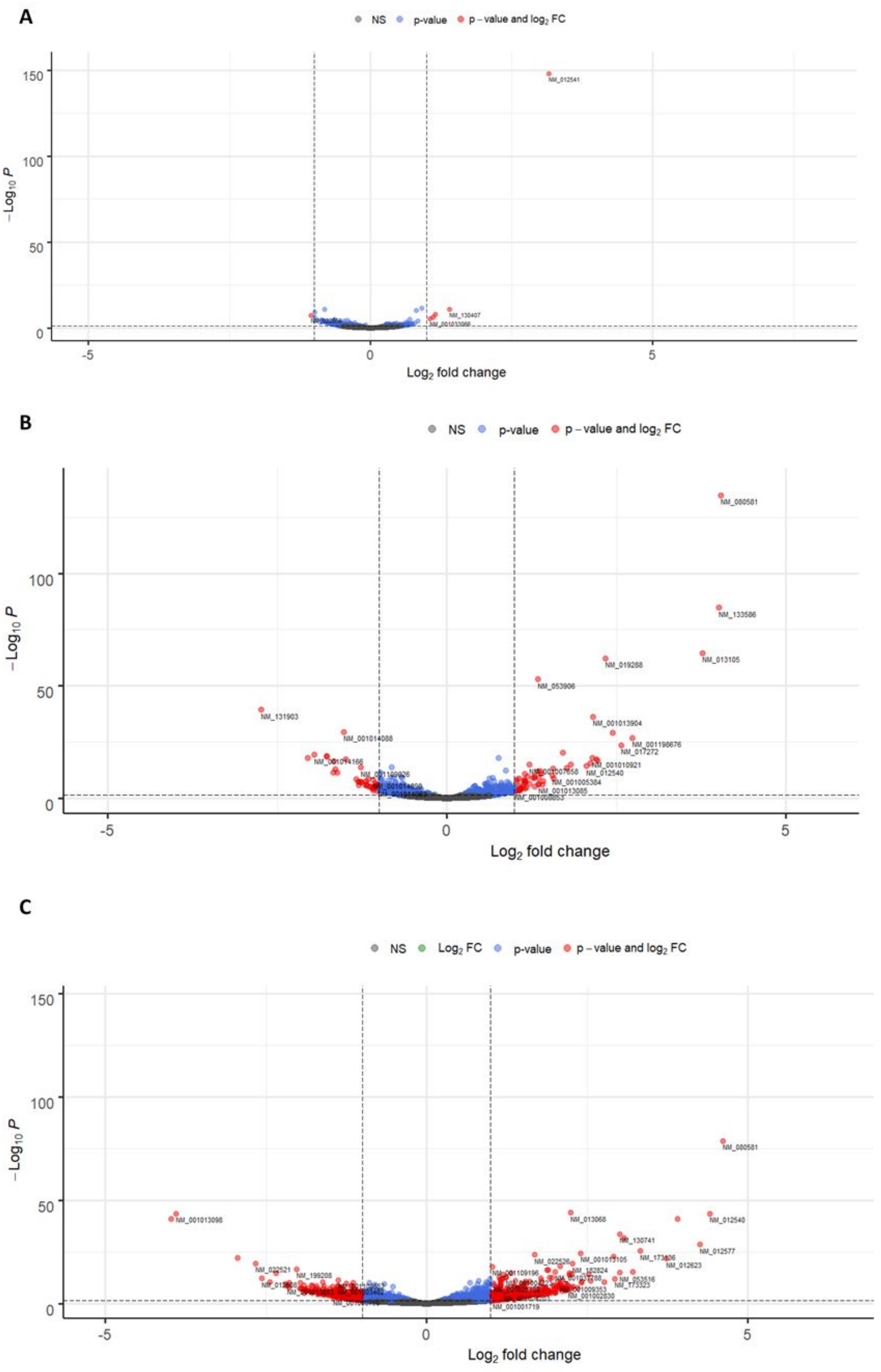


Figure 5: Volcano plots for differentially expressed genes in case of a) ahr, b)car/pxr, and c)cytotoxic

Limma Results for Differentially Expressed Genes

For adjusted p-value < 0.05 we obtained 58 DEGs for Ahr, 9458 DEGs for cytotoxic and 2692 DEGs for CAR/PXR treatment conditions. The top 10 DEGs for Ahr, Cytotoxic and CAR/PXR conditions are given in tables 6, 7 and 8 respectively.

Probeset ID	**log 2 Fold Change**	**P Value**	**Adjusted P Value**
1387243_at	1.57847694444445	2.54068663441576E-17	7.901E-13
1370613_s_at	0.784688111111114	1.34706529985649E-12	2.094E-08
1387759_s_at	0.992616555555556	2.35883352122916E-12	2.445E-08
1383325_at	0.473877344444445	2.08325597494283E-09	1.619E-05
1387901_at	-0.466830944444444	5.9356698902014E-07	0.003691867958307
1372297_at	0.42059677777778	8.29063230240519E-07	0.004297172899542
1384544_at	0.345374555555556	9.71372036460206E-07	0.004315528423125
1368168_at	-1.23589822222222	1.89635858836285E-06	0.007371856967437
1380888_at	0.538433611111111	2.62916683211214E-06	0.00908493992354
1367669_a_at	0.38198288888889	3.1529607191816E-06	0.009805392540583

Table 6: Top 10 differentially expressed genes for Ahr treatment condition

Probeset ID	**log 2 Fold Change**	**P Value**	**Adjusted P Value**
1369698_at	4.13685572222222	3.85452925347886E-21	1.198E-16
1388122_at	4.73158402777778	3.57702962043757E-20	4.726E-16
1370902_at	7.39365077777777	4.55926808521675E-20	4.726E-16
1395403_at	-4.65823335555555	1.33148246452036E-17	1.035E-13
1393508_at	-4.17825707777777	3.73329427055713E-17	2.322E-13
1370425_at	-5.1796098	2.1253341213192E-16	9.777E-13

1368121_at	3.13012655555555	2.20072949109352E-16	9.777E-13
1374070_at	3.78793116111111	1.01714930136108E-15	3.729E-12
1397205_at	-5.65554448333333	1.07934288392385E-15	3.729E-12
1369941_at	1.55573938888889	1.74708434364678E-15	5.433E-12

Table 7: Top 10 differentially expressed genes for Cytotoxic treatment condition

Probeset ID	**log 2 Fold Change**	**P Value**	**Adjusted P Value**
1371076_at	2.99937746551724	1.40158402503419E-27	4.358E-23
1387118_at	1.38106331609195	8.69321016046323E-24	1.3513E-19
1370698_at	1.93690391762452	2.62808477837141E-23	2.724E-19
1368905_at	3.21850014942529	1.31675937002466E-17	1.023E-13
1387759_s_at	1.0151228697318	7.87795906700881E-15	4.899E-11
1370580_a_at	0.492910219348659	1.99600346322376E-14	1.034E-10
1370613_s_at	0.84547180842912	2.68152256791276E-14	1.191E-10
1378126_at	0.89661920421456	4.81030219040579E-14	1.869E-10
1369698_at	3.04987015478927	9.24399971730072E-14	3.19E-10
1398307_at	0.777418454980843	1.95824457783964E-13	6.089E-10

Table 8: Top 10 differentially expressed genes for CAR/PXR treatment condition

The frequencies of the DEGs for different fold changes is shown in figure 6 for each treatment condition.

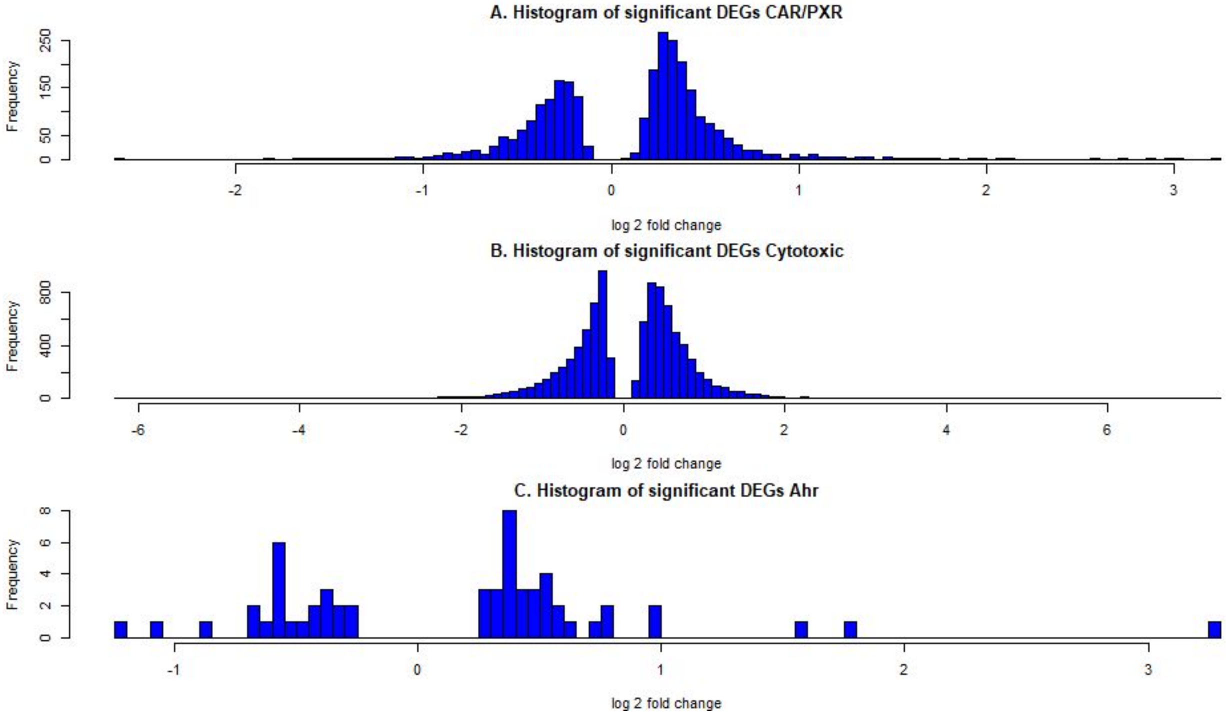


Figure 6: Histogram representing the frequencies of the different log 2 fold changes across the significant DEGs for (A) CAR/PXR (B) Cytotoxic (C) Ahr treatment conditions

The figure 7 shows a volcano plot of the log 2 fold change against the -log10 p-value for all the genes in each treatment condition

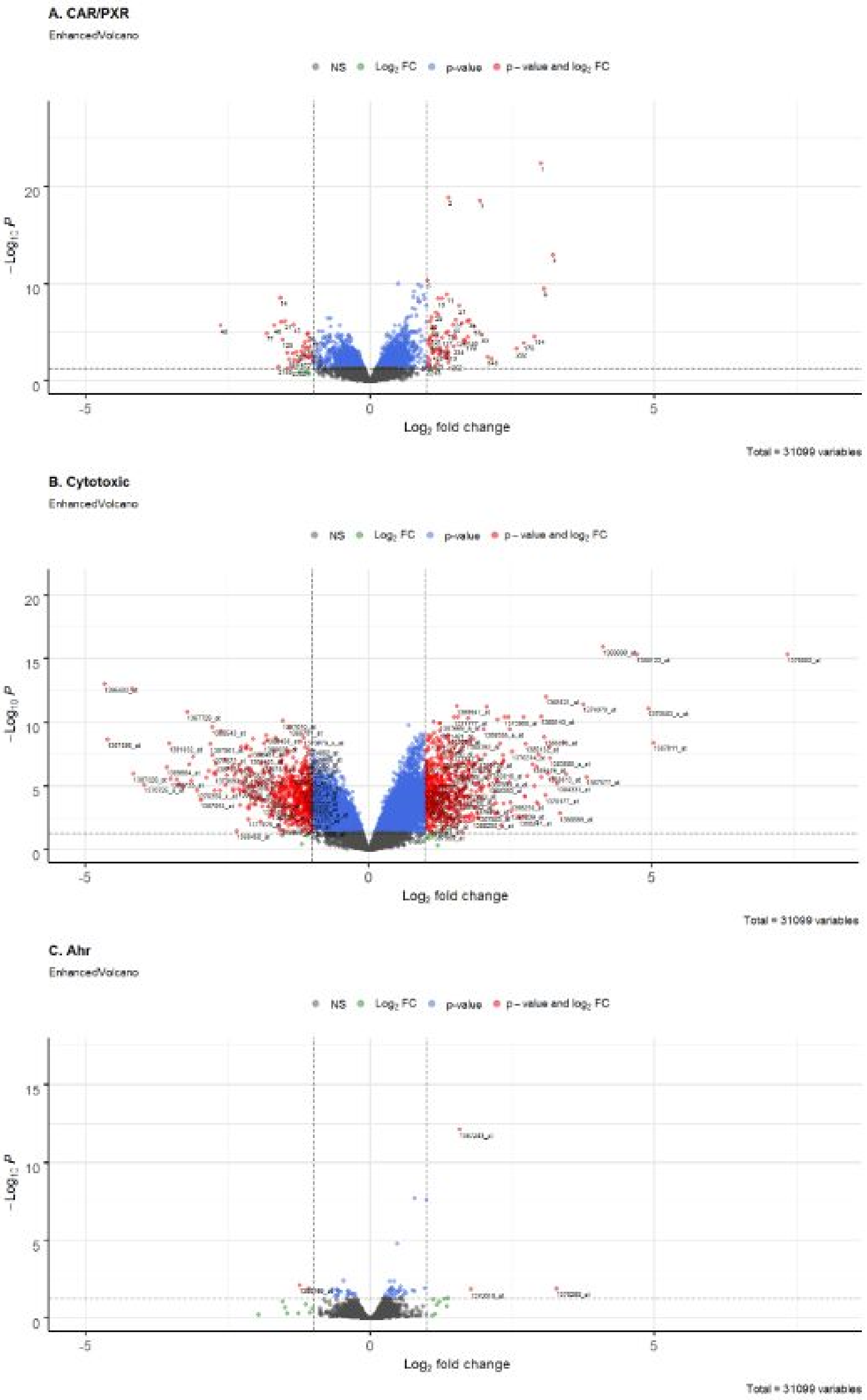


Figure 7:Volcano plots for differentially expressed genes in case of a) CAR/PXR, b) Cytotoxic, and c) Ahr treatment conditions. NS are the non significant genes.

Concordance of DEGs across both platforms

The concordance values computed for Ahr, Cytotoxic and CAR/PXR were 0.214 ,0.202, 0.414 respectively. We observe that concordance is affected by the number of DEGs present in the RNA-Seq results as shown by figure 8. It is not dependent on the number of DEGs given by the microarray.

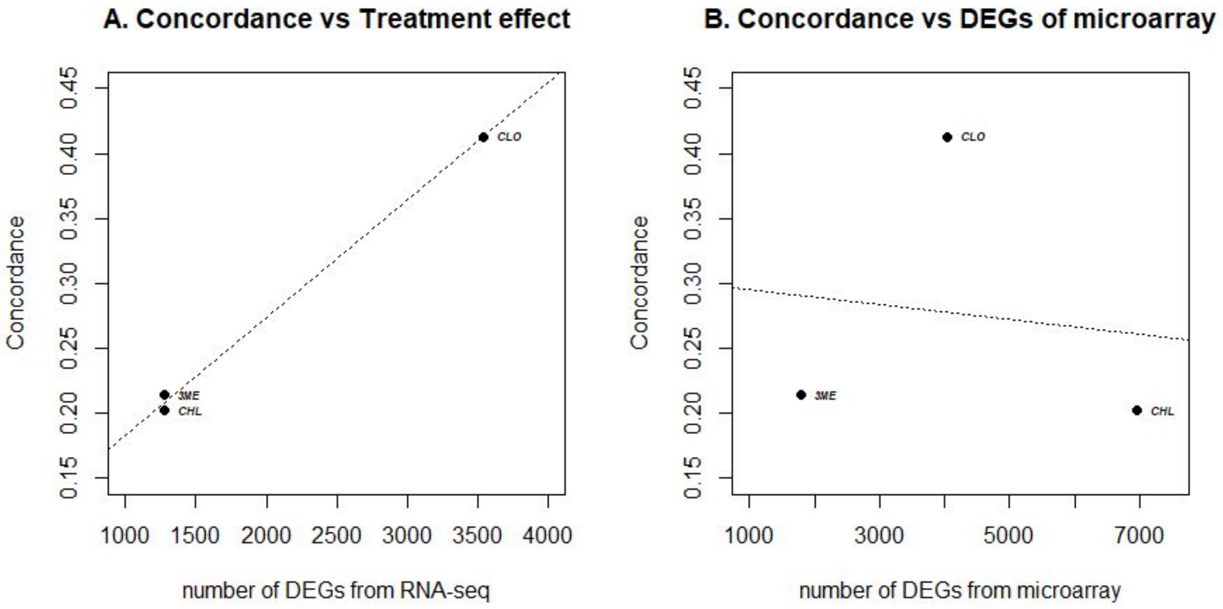


Figure 8: (A) Concordance value vs Treatment effect , i.e, number of DEGs found using RNA-Seq for the three chemicals (3ME - 3-METHYLCHOLANTHRENE which has Ahr MOA , CHL - Chloroform which has Cytotoxic MOA , CLO - Clotrimazole which has CAR/PXR MOA) , (B) Concordance value vs the number of DEGs found using microarray for the same three chemicals

The concordance computed for above median expression DEGs for Ahr, Cytotoxic and CAR/PXR were 0.195,0.172 and 0.416 respectively. Concordance values for below median expression genes for the above 3 MOAs were 0.088, 0.081 and 0.229 respectively. Figure 9 shows that the below median expression DEGs have low concordance values and might affect the overall concordance when considering all significant differentially expressed genes. Better concordance can be achieved with above median expression DEGs.

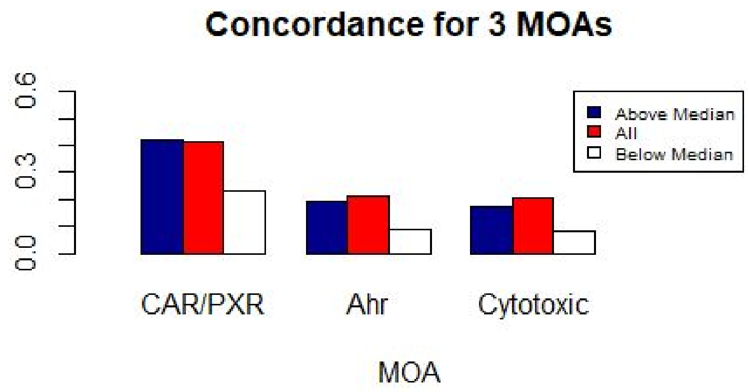


Figure 9: Concordance calculated for above median, all DEGs and below median expressed DEGs for three MOAs

Enriched Pathways:

MOA	**GATHER Pathways**
AhR	Cytokine-cytokine receptor interaction
Cytotoxic	Tight junction
	Adherens junction
CAR/PXR	Chondroitin/Heparan sulfate biosynthesis
	Adherens junction
	Tight junction

Table 9:The enriched pathways using GATHER

Gather Gene Annotation		**P Value**
GO:0006091	carboxylic-acid metabolism	0.003
GO:0006749	Gluthathione metabolism	0.01

GO:0000187	Activation of MAPK	0.01
GO:0018894	Dibenzo-p-dioxin metabolism	0.02

Table 10:The gene annotation of AhR MOA

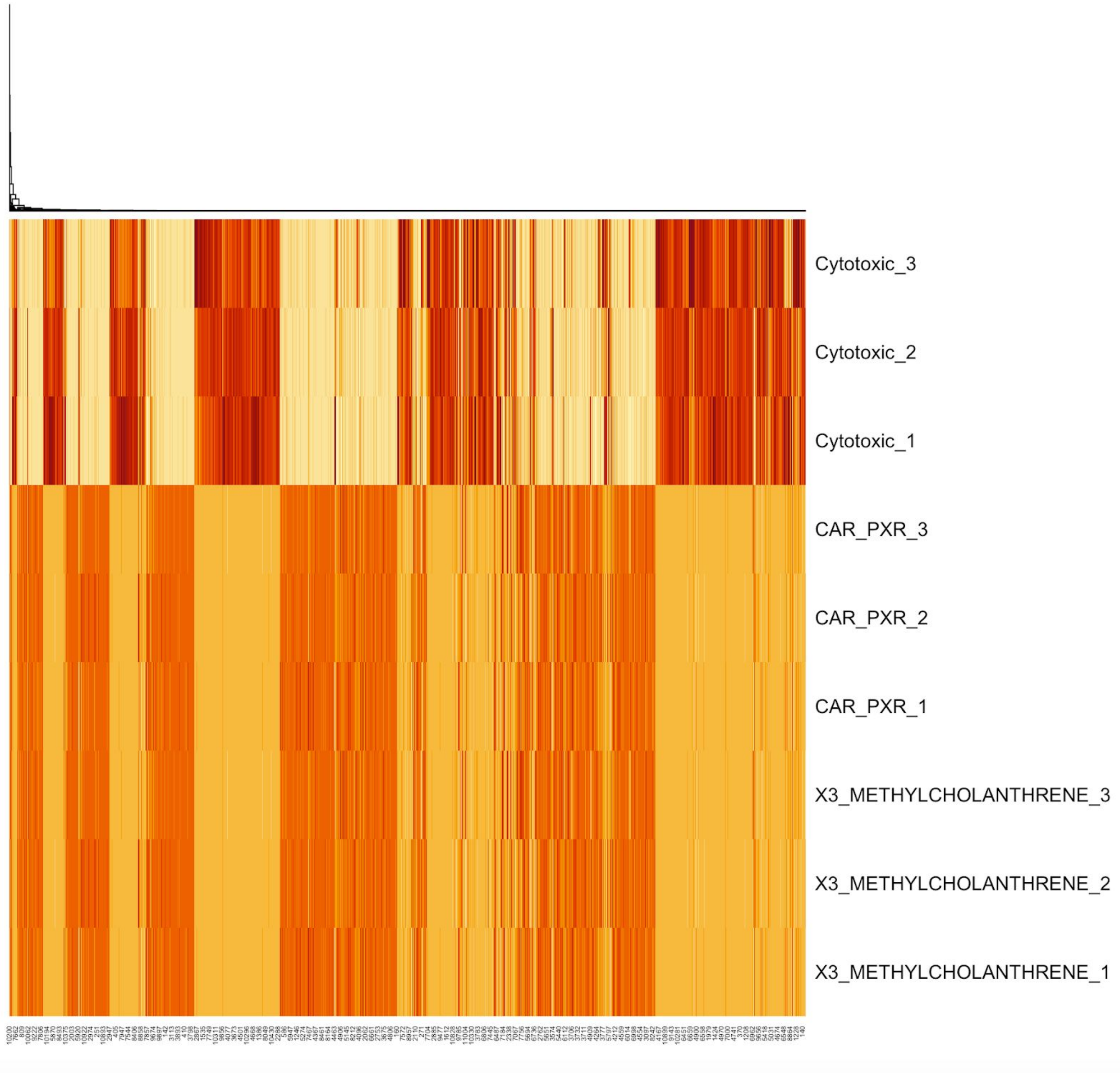


Figure 10: A clustered heatmap of the normalized counts.

DISCUSSIONS:

In this project, we analyzed and compared the rna-seq data and microarray data with respect to three MOA in liver cells of rats treated by three drugs respectively. We calculated the concordance between the platforms using results of limma and DESeq which showed us that concordance is affected by the treatment size as mentioned in the paper. We also observe that we get different numbers of DEGs for each MOA which can be because the cutoff we used differed from the paper and number of samples used to replicate the experiment.

In the rna-seq data, most samples we used in this project passed fastqc quality control. Looking at the pathways enriched using GATHER, cytotoxic and car_pxr MOA enriched pathways are tight and adherens junction. Adherens junctions play an indispensable role in maintaining tissue architecture[5]. Similarly, tight junctions help to maintain the polarity of cells which is crucial. It is not surprising that they are often involved in diseases. They are fundamental for the overall functionality of the cell, so in instances where the cell has been perturbed by a toxin it is no surprise that they will be involved among other similarly diresputive pathways such as cell cycle checkpoint regulation described in the literature. Car_pxr had an additional enriched pathway, Heparan/sulfate biosynthesis, heparan sulfate have been shown to alter processes such as cell adhesion, immune cell infiltration and angiogenesis[6]. This pathway might potentially explain the loss in functionality in cell adhesion in these cells.

Additionally, AhR MOA enriched pathways was Cytokine-Cytokine receptor interaction. Cytokines have been recently recognized as targets by the AhR signaling cascade and their involvement with tumorigenesis [7,9]. It has been demonstrated that cytokine-cytokine receptor interaction pathway may be involved in hepatocellular carcinoma(HCC) carcinogenesis[8]. Aryl hydrocarbon (dioxin) receptor (AhR) is a ligand-activated transcription factor that increases xenobiotic metabolism, histone modifications, and tumorigenesis [9]. AhR activation is followed by changes in the compartmentalization within the cell and upon ligand binding, the receptor translocates to the nucleus and hetrodimerizes with the aryl hydrocarbon nuclear translocator (ARNT). [10] This heterodimer binds to a set of co-activators and/or co-repressors resulting in complex interactions such as xenobiotic response elements[9]. Since the enriched pathway of AhR was different from the rest, we proceeded to analyze the gene annotations. One of the enrichment terms GATHER found was carboxylic acids which constitute a large and heterogeneous class of both endogenous and xenobiotic compounds(Table 7) [11] . Another term found was glutathione(GSH) metabolism, GHS occurs non-enzymatically through the action of GSH S-transferases [12]. One important function of GHS S-transferases is the conjunction of GSH and xenobiotic compounds, enhancing excretion of xenobiotics [12]. Which all play a role in AhR activation and tumorigenesis. Another term found was the activation of MAPK pathways. It has been shown that AhR activates transcription of proto-oncogene c-jun[13]. MAPK has been associated with many cancers and so it makes sense that it would be a pathway also activated. Similar to the literature, GATHER found very similar pathways involved in all three MOA.

In the heatmap(Figure 10) we notice that each MOA clusters together and that car_pxr and AhR cluster together which is to be expected because they have very similar enriched pathways in the literature. In our analysis with Gather they showed different pathways and more similarities between cytotoxic and car_pxr.

CONCLUSION:

In conclusion, our results show the same pattern of concordance as the paper. Although the enriched pathways weren’t the same between them they had similar functionalities.

REFERENCES:

[1]Yang Liao, Gordon K. Smyth, Wei Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, Bioinformatics, Volume 30, Issue 7, 1 April 2014, Pages 923–930, https://doi.org/10.1093/bioinformatics/btt656

[2] 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.

[3]Wang C, Gong B, Bushel P R, et al. A comprehensive study design reveals treatment-and transcript abundance–dependent concordance between rna-seq and microarray data[J]. Nature biotechnology, 2014, 32(9): 926.

[4]Ganter B, et al. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. Journal of Biotechnology. 2005; 119:219–244.

[5]Van Campenhout, R., Crespo Yanguas, S., Cooreman, A., Gijbels, E., Leroy, K., Vilas-Boas, V., Devoogdt, N., Muyldermans, S., Cogliati, B., & Vinken, M. (2019). Increased Expression of Adherens Junction Components in Mouse Liver following Bile Duct Ligation. *Biomolecules*, *9*(10), 636. https://doi.org/10.3390/biom9100636

[6]Nagarajan, A., Malvi, P., & Wajapeyee, N. (2018). Heparan Sulfate and Heparan Sulfate Proteoglycans in Cancer Initiation and Progression. *Frontiers in endocrinology*, *9*, 483. https://doi.org/10.3389/fendo.2018.00483

[7]Fardel O. Cytokines as molecular targets for aryl hydrocarbon receptor ligands: implications for toxicity and xenobiotic detoxification. *Expert Opin Drug Met* 2013; 9:141–152. doi: 10.1517/17425255.2013.738194.

[8]Jiang, X., & Hao, Y. (2018). Analysis of expression profile data identifies key genes and pathways in hepatocellular carcinoma. *Oncology letters*, *15*(2), 2625–2630. https://doi.org/10.3892/ol.2017.7534

[9].Larigot, L., Juricek, L., Dairou, J., & Coumoul, X. (2018). AhR signaling pathways and regulatory functions. *Biochimie open*, *7*, 1–9. https://doi.org/10.1016/j.biopen.2018.05.001

[10]Tsay, J. J., Tchou-Wong, K. M., Greenberg, A. K., Pass, H., & Rom, W. N. (2013). Aryl hydrocarbon receptor and lung cancer. Anticancer research, 33(4), 1247–1256.

[11]Skonberg C.; Olsen J.; Madsen K. G.; Hansen S. H.; Grillo M. P. Metabolic activation of carboxylic acids. Expert Opin. Drug Metab. Toxicol. 2008, 4, 425–438. 10.1517/17425255.4.4.425.

[12]Sipes I.G., Wiersma D.A., Armstrong D.J. (1986) The Role of Glutathione in the Toxicity of Xenobiotic Compounds: Metabolic Activation of 1,2-Dibromoethane by Glutathione. In: Kocsis J.J., Jollow D.J., Witmer C.M., Nelson J.O., Snyder R. (eds) Biological Reactive Intermediates III. Advances in Experimental Medicine and Biology, vol 197. Springer, Boston, MA

[13]Weiss C., Faust D., Durk H., Kolluri S.K., Pelzer A., Schneider S., Dietrich C., Oesch F., Gottlicher M. TCDD induces c-jun expression via a novel Ah (dioxin) receptor-mediated p38-MAPK-dependent pathway. Oncogene. 2005;24:4975–4983. doi: 10.1038/sj.onc.1208679.

[14] 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. doi: 10.1093/nar/gkv007.

[15] Haynes W. (2013) Benjamini–Hochberg Method. In: Dubitzky W., Wolkenhauer O., Cho KH., Yokota H. (eds) Encyclopedia of Systems Biology. Springer, New York, NY