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

Group Members:

Jinghan Huang (Data Curator)

Jiaming Zhang (Programmer)

Mikhail Kouzminov (Analyst)

TA:

Kritika Karri

Introduction

The original paper, 'A comprehensive study design reveals treatment- and transcript abundance–dependent concordance between RNA-seq and microarray data' [1], presented the comparison between two popular techniques for measuring mRNA abundance: microarray and RNA-seq. The goal of this study was to verify the concordance between RNA-seq and microarray data so that the level of the confidence of the results could be more accurately assessed. The author generated the data from the same set of rat liver samples which were treated with different toxicological chemicals representing the different modes of action (MOA) [1]. Since the differentially expressed genes (DEGs) were highly affected by the biological complexity of the MOA, the DEGs had been identified and the performance of the predictive models had been assessed to evaluate the differences and similarities between these two techniques. For the bioinformatics techniques, six bioinformatics pipelines had been used to process the raw RNA-seq data in the upstream analysis, and then three methods (limma, edgeR and DESeq) had been applied for identifying DEGs in downstream analysis. RMA and Microarray Analysis suite had been used for microarray data.

In this project, we are going to reproduce parts of the results in this paper, including Figure 2a and Figure 3b,c, which showed the between-platform concordance of DEGs against the number of DEGs identified by RNA-seq, and the concordance level related to highly/low expressed DEGs for each toxic group. The toxic group we selected is toxic group 3, which included Leflunomide, Fluconazole and Ifosfamide.

Data

Data Description

The data was obtained from male Sprague Dawley rats treated with 27 different chemicals and then the RNA was isolated from their livers. The samples had been divided into a training set and test set. In the training set, there were 45 treated rats and 18 controls in total. There were 15 chemicals with 3 rats per chemical. While for the test set, there were 36 treated rats and 6 controls in total and therefore 12 chemicals with 3 rats per chemical. The Affymetrix microarrays and Illumina HiScanSQ and Hiseq2000 systems had been applied to process these samples.

The average number of reads for the 9 samples we selected was about 16833935. These pair end reads were all 100bp long. The raw data could be accessed from GEO, NCBI with accession SRP039021 [2], GSE55347 [3] and GSE47875 [4], which were sequenced from rat livers. The samples we processed were SRR1178008, SRR1178009, SRR1178010, SRR1178014, SRR1178021, SRR1178047, SRR1177981, SRR1177982, SRR1177983, SRR1178050, SRR1178061, SRR1178063, SRR1178004, SRR1178006 and SRR1178013 from toxic group 3 that included Leflunomide, Fluconazole and Ifosfamide. These samples were then aligned using STAR and the quality control had also been conducted using fastqc and multiqc software.

Data quality control

The quality of the data was assessed by fastqc and multiqc software. In each section of fastqc and multiqc report, there was a standard for acceptable quality. The following table and figures were the results created from multiqc. There wasn't any sample removed from our dataset, though some of them did not meet the standard of some quality control metrics. The details were shown below.

Sample	**Total reads**	**Number of uniquely mapped reads**	**% of uniquely mapped reads**	**Number of multi-mapping reads**	**% of multi-mapping reads**	**% of reads unmapped**
SRR1177981	14229351	11410110	80.19%	537573	3.78%	15.91%
SRR1177982	17168681	14390023	83.82%	630047	3.67%	12.35%
SRR1177983	16152281	12859278	79.61%	568705	3.52%	16.72%

SRR1178008	15156072	13348761	88.08%	533695	3.52%	8.16%
SRR1178009	18110504	16315003	90.09%	460617	2.54%	7.22%
SRR1178010	18572519	16832537	90.63%	505301	2.72%	6.50%
SRR1178014	17524782	14637080	83.52%	1151194	6.57%	9.41%
SRR1178021	17497925	14340044	81.95%	1008867	5.77%	11.98%
SRR1178047	17093302	14355442	83.98%	999412	5.85%	9.73%

Table.1 STAR alignment statisticsTable.1 STAR alignment statistics of total reads, uniquely mapped reads, multi-mapping reads and unmapped reads for the 9 samples

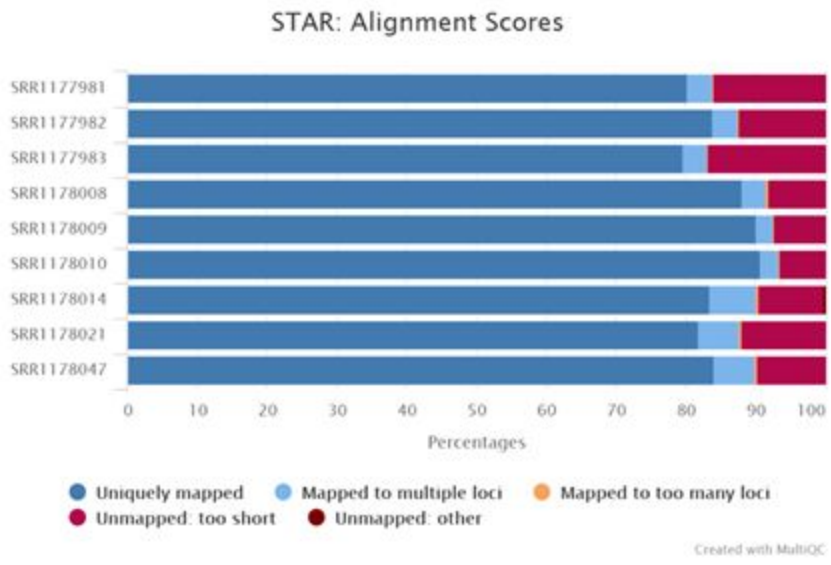

Fig.1 STAR alignment percentage of numbers of uniquely mapped reads, multi-mapping reads and unmapped reads for the 9 samples

The alignment statistics (Table.1) and the percentage bar chart for the statistics (Fig.1) were shown above. The average number of the total reads for the 9 samples was around 16833935.

Among these reads, there were about 80%-90% of them successfully uniquely mapped to the genome. There were only 2%-6% of the reads multi-mapped and 6%-16% of the reads unmapped, which could be treated as acceptable quality.

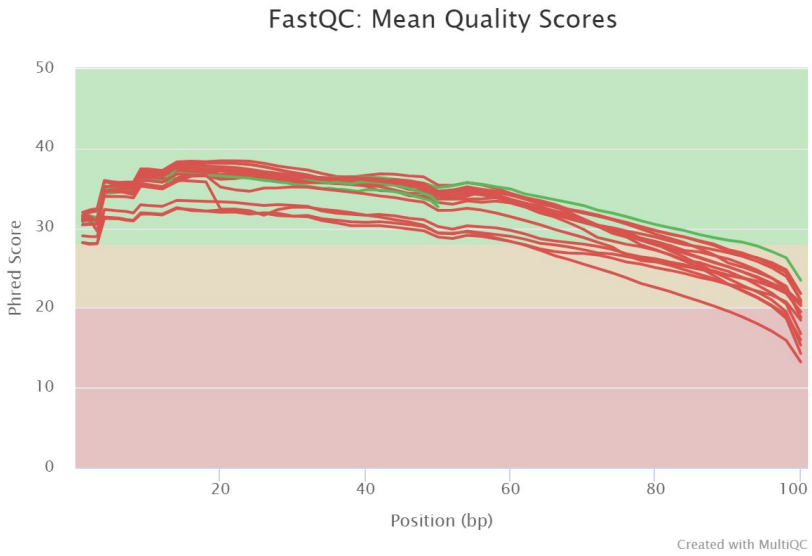

Fig.2 MultiQC report: FastQC—quality scores across all bases

The ‘per base sequence quality’ which showed the quality scores across all bases for the 9 samples (18 fastq files) was shown above (Fig.2). In this part, among the 18 fastq files, only 3 of them passed the quality test. There were 15 of them raising failures. From Fig.2, for most of the samples, the qualities of the bases were good until ~70bp (higher than 30). However, they were not satisfactory after 90bp and dropped below 20 at the end of the reads. The reason for these failures could be concluded that in general sequencing, chemistry degraded with increasing read length and for long runs the base quality would decrease to a level that would result in a failure [5]. One possible solution is to perform quality trimming where reads are truncated based on their average quality. Here the last 10 bp could be trimmed. However, we did not remove these data or perform any further steps to improve their qualities. On one hand, the author of the original paper did not mention if they performed any trimming step, though they did mention that they had used Fastqc for quality control. On the other hand, it was acceptable to see the relatively lower quality at the end of the reads (over 20 around 90bp). As we were aiming to reproduce their results, we kept these data for the following analysis.

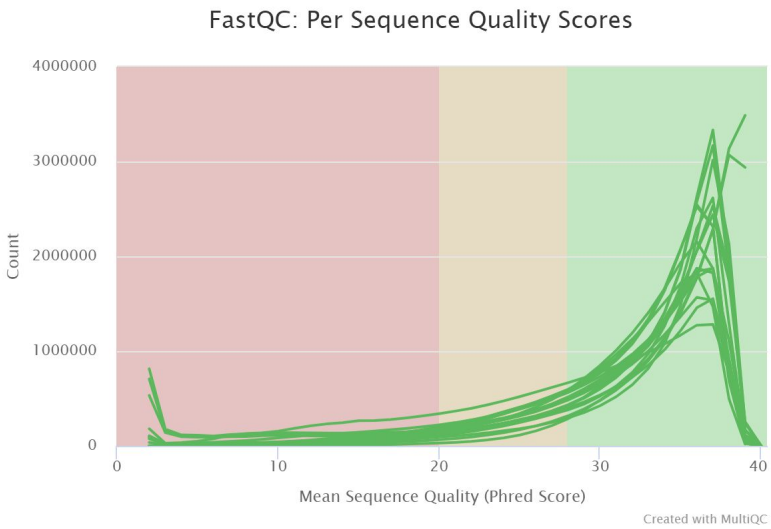

Fig.3 MultiQC report: FastQC—per sequence quality scores

The ‘per sequence quality scores’ which indicated the quality of the subsets of the sequences was shown in Fig.3. In this part, all the samples passed the quality test. From Fig.3, the most frequently average quality was around 36 and there is no other obvious peak or fluctuation, which indicated that a large proportion of the sequences had good average quality.

Fig.4 MultiQC report: FastQC—Per Base Sequence Content

The ‘per base sequence content’ which indicated the percentage of four normal DNA bases in each base position was shown in Fig.4. In this part, most of the samples failed the test because an unexpected difference between the four DNA bases was greater than 20% at the start of the reads. From Fig.4, for each sample, there were obvious colors representing the DNA bases shown at the first several bases that indicated their higher percentage. Nevertheless, it was acceptable since nearly all RNA-Seq data will fail this test because of a biased selection of the

random primers that would affect around the first 12bp [5]. In this situation, it didn't represent biased sequences and would not influence the following processing and analysis.

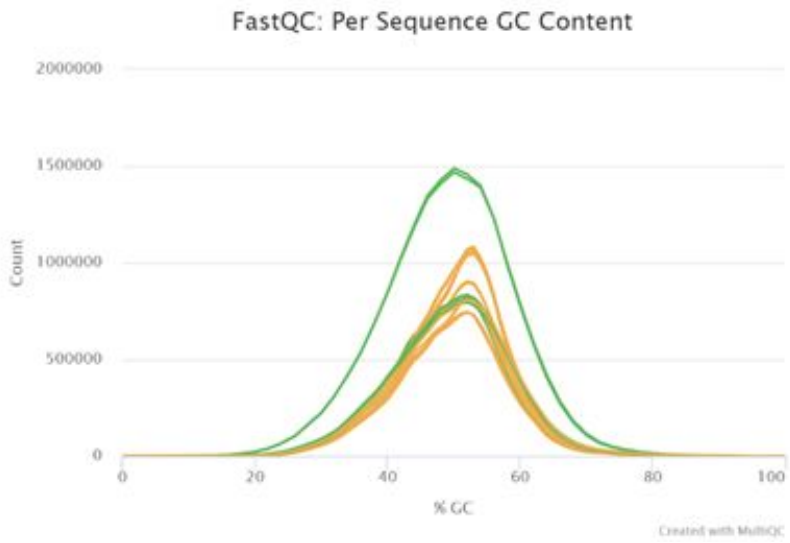

Fig.5 MultiQC report: FastQC—Per Sequence GC Content

The 'per sequence GC content' which indicated the average GC content of reads was shown in Fig.5. In this part, there were some warnings raised because the sum of the deviations from the normal distribution represents more than 15% of the reads [5] but they were still acceptable.

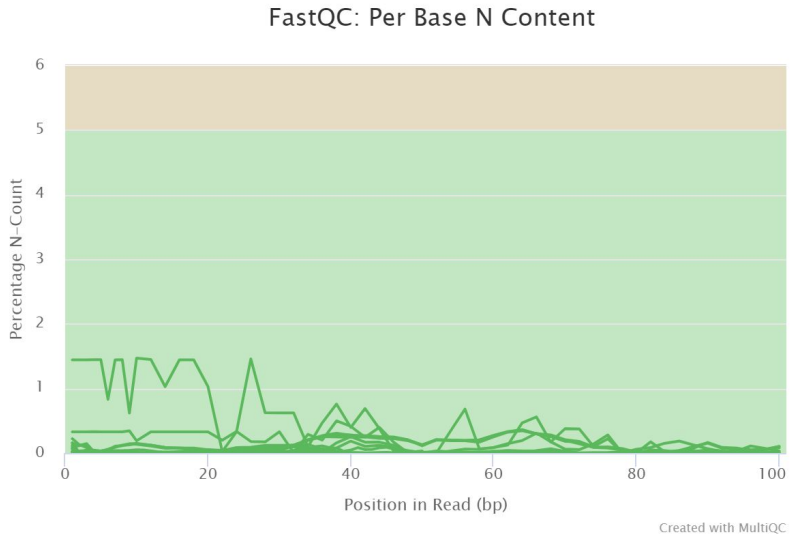

Fig.6 MultiQC report: FastQC—Per base N Content

The 'per base N content' which indicated the percentage of 'N' at each position was shown in Fig.6. When it was not confident to call any of the four DNA bases, the N would be substituted.

In this part, all the samples passed the test and the samples showed very low percent of ‘N’ in each position.

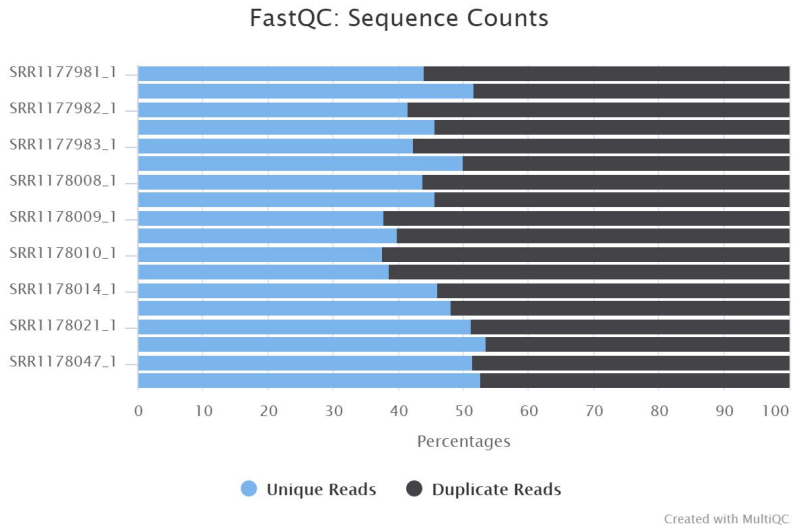

Fig.7 MultiQC report: FastQC—percentage of sequence counts

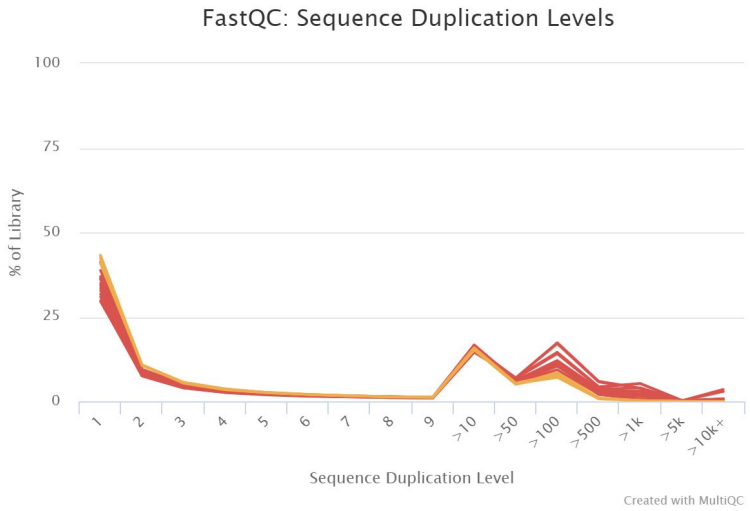

Fig.8 MultiQC report: FastQC—sequence duplication levels

The percentage of unique reads and duplicate reads and the ‘sequence duplication levels’ which indicated the relative number of sequences with different degrees of duplication were shown in Fig.7 and Fig.8. In this part, many of the samples failed the test due to their high duplication levels. However, for RNA-seq data, usually high expressed transcripts would be over-sequenced to obtain the lowly expressed transcript so that the large amounts of duplicates would always happen [5]. It was acceptable for RNA-seq data to have a high duplicated level and we concluded that the data could be used for following analysis.

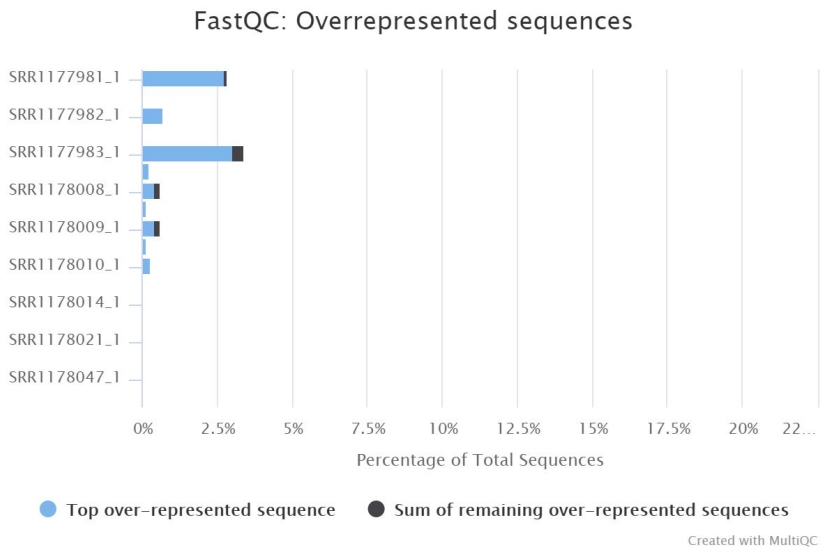

Fig.9 MultiQC report: FastQC—overrepresented sequence

The ‘overrepresented sequences’ which indicated the percentage of top overrepresented sequence and sum of remaining overrepresented sequences was shown in Fig.9. In this part, there are two failed tests among the samples due to the sequences representing over 1% of total. As mentioned previously, for RNA-seq some high expressed transcripts would be over-sequenced to obtain those lowly expressed transcripts. Therefore, it was common that overrepresented sequences were present and they were acceptable for RNA-seq data.

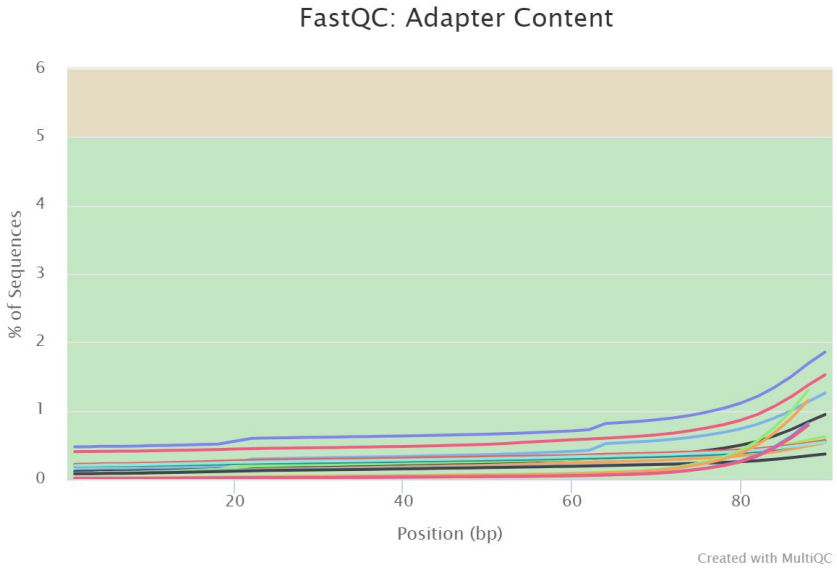

Fig.10 MultiQC report: FastQC—Adapter content

The ‘adapter content’ which indicated the cumulative percentage count of the proportion of the adapter sequence at each base position was shown in Fig.10. In this part, all the samples

passed the test and there was no adapter sequence present in more than 2% of all reads, which indicated their good quality.

Methods

Quality Control Tools

Fastqc/0.11.7 and MultiQC/1.6 had been used for raw data quality control to ensure the good quality of reads. For Fastqc, it provided some analyses that could be used as indices and indicators for the data quality before doing any further analysis [5]. The fastq files were the input and the output was the quality control metrics as html format. For MultiQC [6], it collected information created from multiple bioinformatics software such as Fastqc and STAR (described below) and made them into one single html report. The available information would be put in one directory as input and the output would be one single html report.

Sample alignment

Each sample was aligned against the rat genome using STAR/2.6.0c [7]. STAR was an alignment tool spliced RNA-seq data and was based on an alignment algorithm that applied sequential maximum mappable seed search in uncompressed suffix arrays followed by seed clustering and stitching procedure [8]. Running STAR aligner required the genome and index as parameters put in one directory, together with setting the fastq files as input. The other parameters we set include ‘runThreadN’ as 16, ‘readFilesCommand’ as zcat and ‘outSAMtype’ as Bam SortedByCoordinate. This allowed us to obtain the sorted Bam format files instead of Sam format files. Besides, the ‘outFileNamePrefix’ parameter was also specified to obtain a unique name for each sample. The output included the alignment bam files, together with alignment statistics including the information for uniquely mapped reads, multi-mapped reads as well as the unmapped reads.

Gene annotation

The nine samples aligned with STAR are then annotated against a gene annotation gtf file using featureCounts tool from the subread package. A qsub script was submitted on SCC, which ran featureCounts on each aligned sample BAM file against the gene annotation file. The script successfully executed and outuputted nine gene count files for each sample. Subsequent quality check on gene annotation output was performed by multiqc per samplewise. The gene counts were extracted from each featureCounts output file and combined into a single csv file for further analysis. Each samples’ gene counts across the entire annotation gene list was plotted as a boxplot using the combined csv file and R for a quick visual examination of the distribution of the counts. Due to irregular distribution and the large range of the gene counts, log scale of the gene counts were used instead.

RNA-seq differential expression analysis

DESeq2 that uses a negative binomial regression was employed to analyze the gene count difference between experimental samples groups and control groups. The samples from toxicological chemicals treatment group three contain three sets of mode of action treatment groups, including chemical targeting AhR, CAR/PXR nuclear receptors, and chemical that causes DNA damage. Depending on the delivery vehicle of the toxins, three sets of samples' gene counts were compared with the appropriate controls. DESeq2 package in R successfully produced 3 sets of output files for each toxin group. The result data frames were then sorted based on the adjusted p value in an ascending order, allowing the most significantly differentially expressed genes after toxin treatment to be examined. The number of genes with adjusted p value lower than 0.05 were counted for each treatment group and the top 10 most significant genes were reported and combined into a new dataframe and subsequently saved as a new csv file. Histograms of significant genes' fold change for each toxin treatment group were plotted using R. Volcano plots were generated using the ggplot package in R and the most significantly differentially expressed genes in each group were highlighted red.

LIMMA and Microarray Differential Expression

Limma uses linear models and differential expression to predict the Expression of various genes based on microarray analysis and using different probes and their relevance to predicting fold change. In the original paper a cutoff of 1.5 FC was used and a p value of 0.05, and our paper used a p of 0.05 for predicting the genes most differentially expressed due to the use of various probes.

Concordance:

We also attempted to calculate Concordance between Ref-Seq and microarray DE genes. Unfortunately, we realized too late that we had accidentally found all genes for microarray probesets where p was > 0.5, leading to what was an interesting image, but also took far too much time to calculate. If we had a few extra hours, we could get much better results. This would require running a batch job either way, since it requires over a million comparisons between 3 different memory locations in order to find that the gene sets are concordant in their direction of fold change.

The formula used to calculate concordance corrected for background, after finding which pairs of genes were concordant was: where n0 is the number of items observably concordant, n1 was the number of items in one gene set and n2 the number of items in the other. N is the total number of unique items.

$(n_0 \times N - n_1 \times n_2)/(n_0 + N - n_1 - n_2)$ = concordance

As for accounting for multiple probe sets per reference sequence, this would have been handled by counting both probe sets as different samples for the purpose of finding their concordance

and variance. They could have been taken under different conditions and resulted in different results. This would have separated them when finding the result by splitting into 3 groups based on the median.

Results

The multiQC report on gene annotation showed that the aligned reads from all nine samples has a good match percentage to the genes in the annotation file. Most of the samples have a 60% assigned rate.

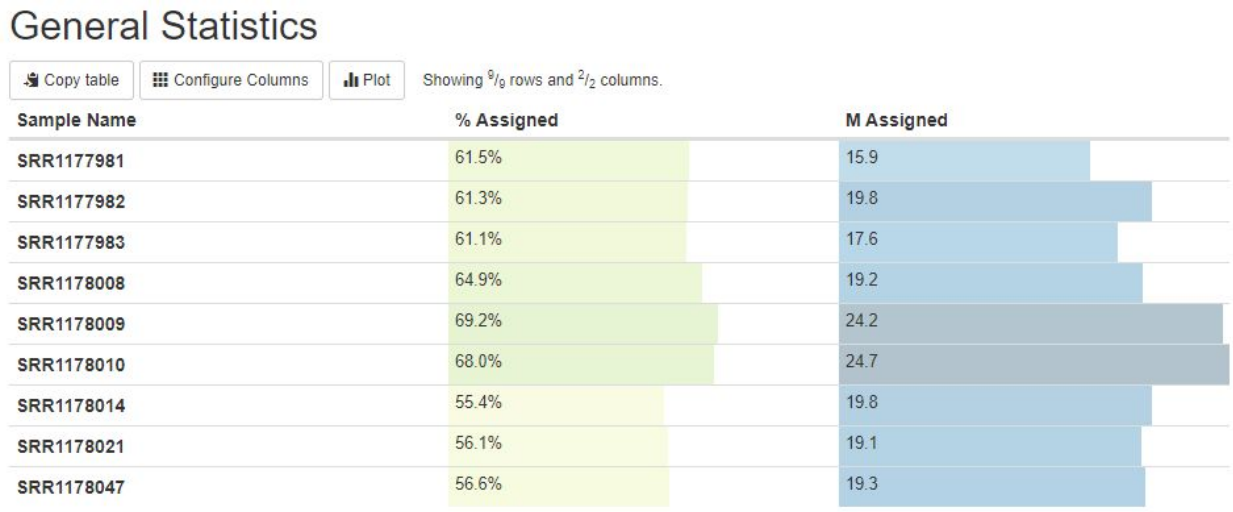

Fig.11. MultiQC result on featureCounts alignment result

According to the log scale boxplot, The range of the gene counts are really high even on a log scale. The distribution of gene expression level seems to be agreeable between samples.

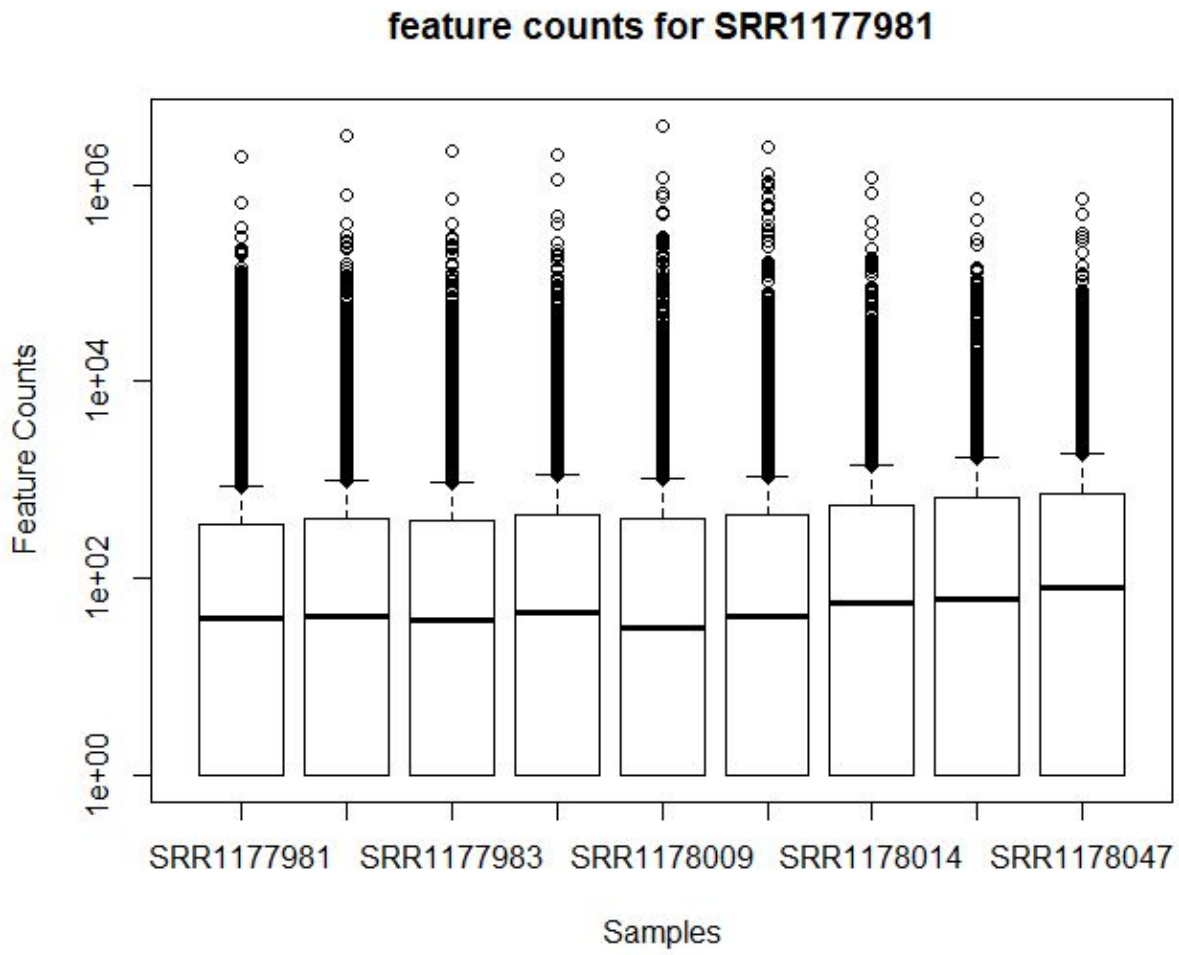

Fig.12 Boxplots of gene counts in log scale

The histograms showed that AhR samples had most genes with a log2 fold change of 1 comparing treatment samples and control. and CAR/PXR samples have a similar distribution as AhR samples. DNA damage treatment group on the other hand has very few significant genes and the log2 fold change are relatively low compared to AhR and CAR/PXR group.

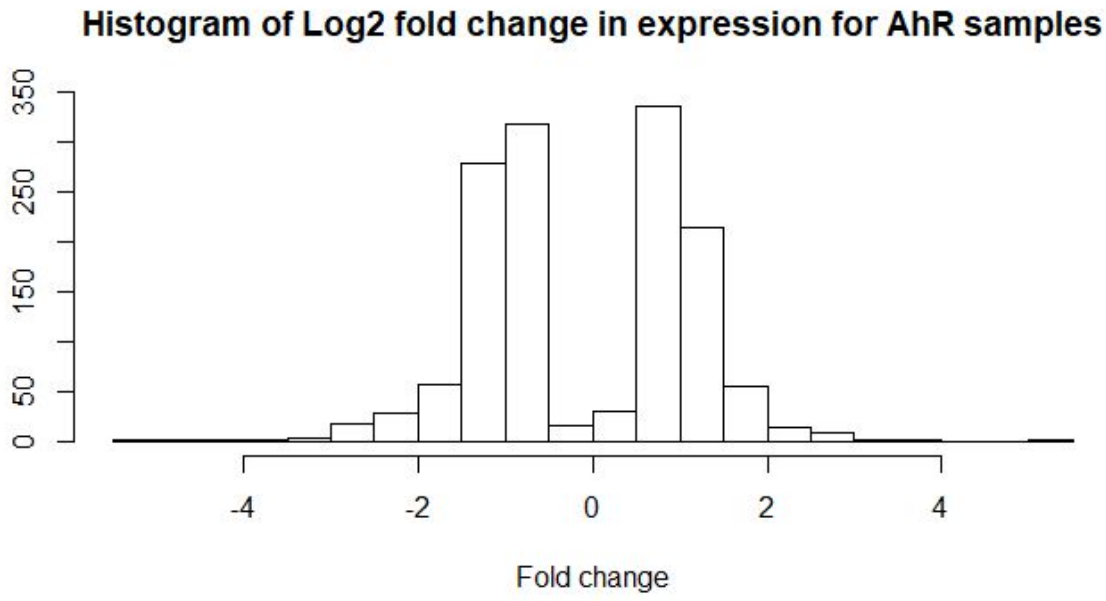

Fig.13. Histogram of expression fold change for AhR group samples

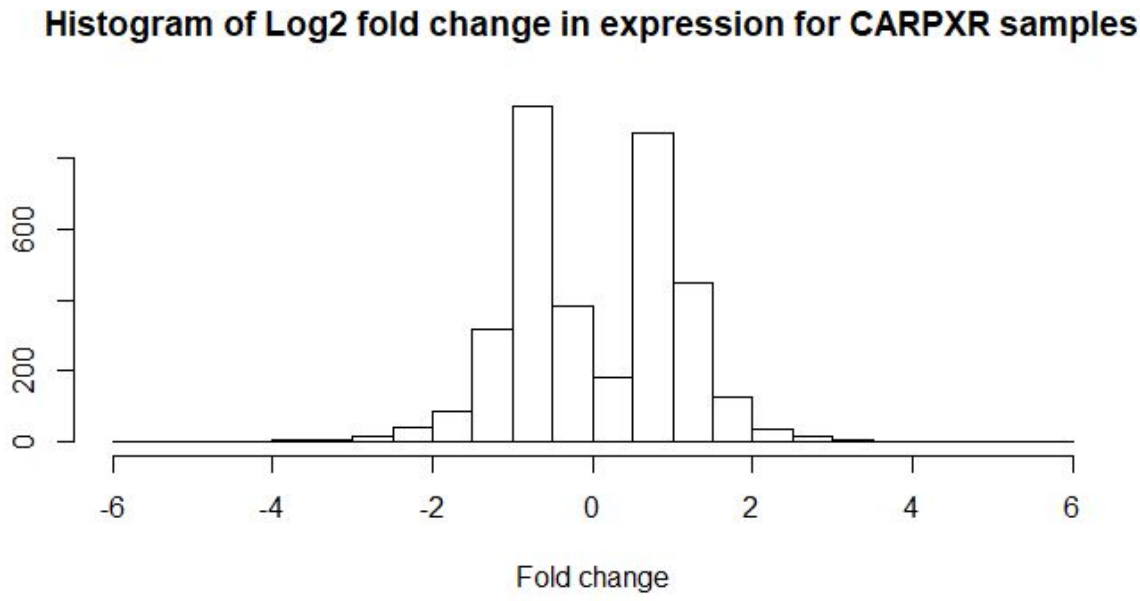

Fig.14 Histogram of expression fold change for CAR/PXR group samples

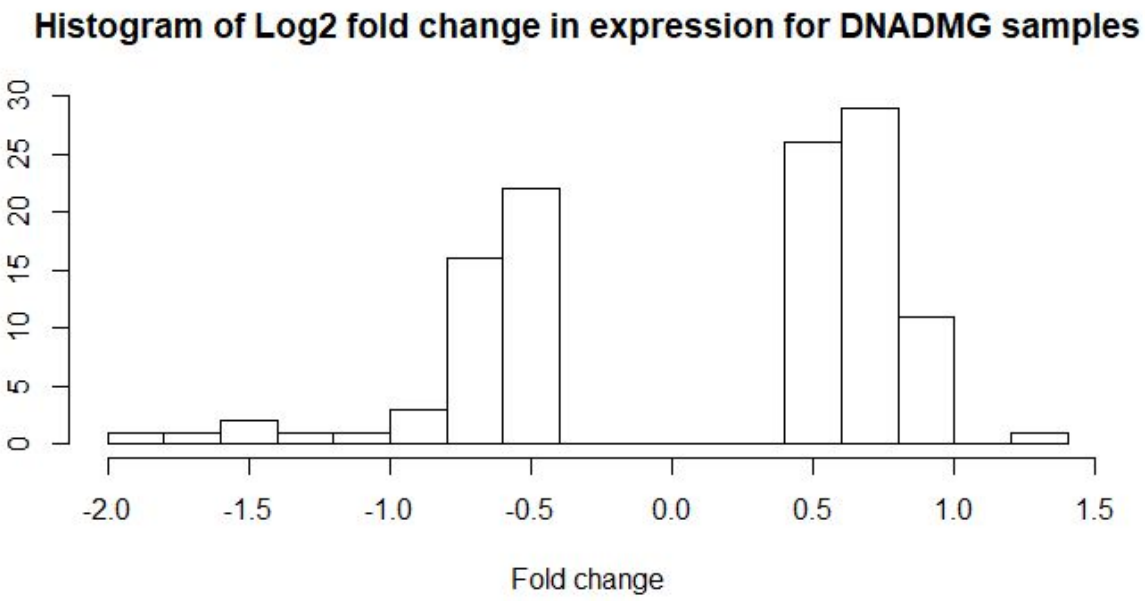

Fig.15 Histogram of expression fold change for DNA-dmg group samples

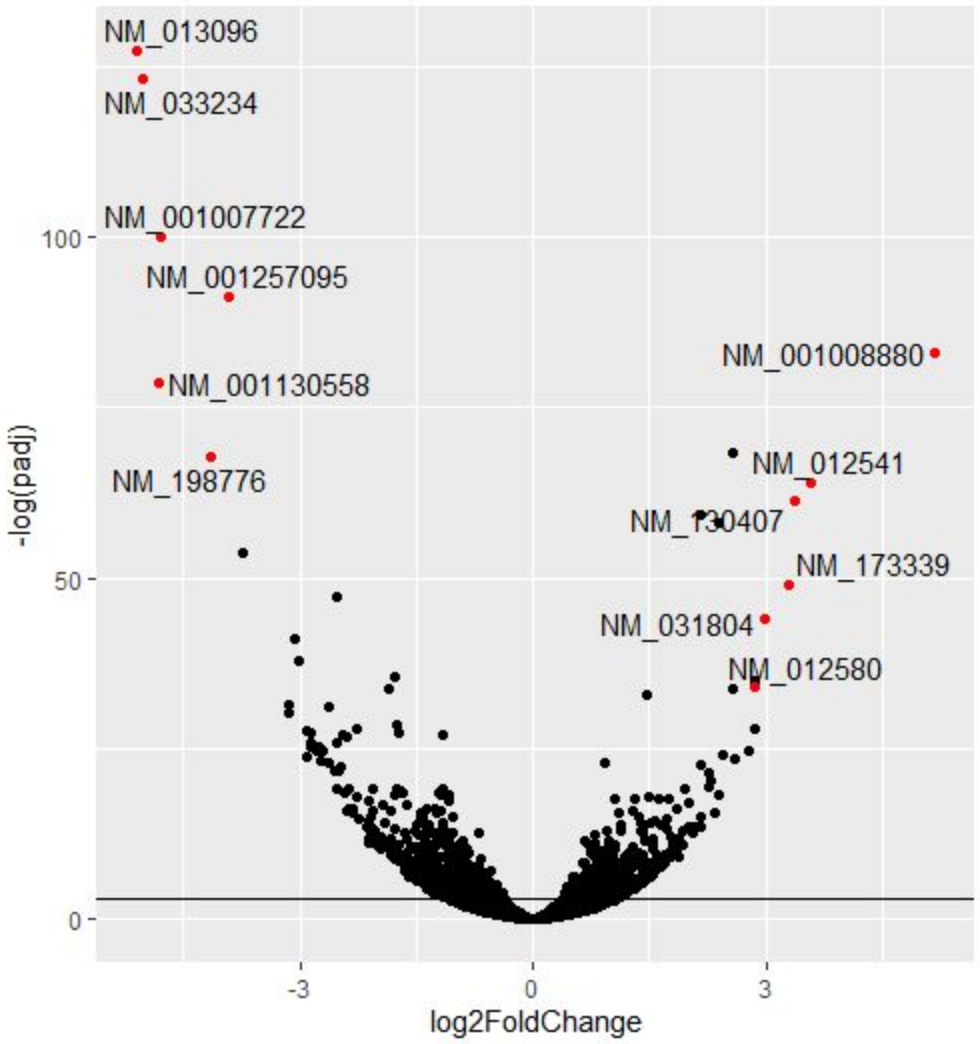

Fig.16. Expression log-fold change against p-value for AhR samples

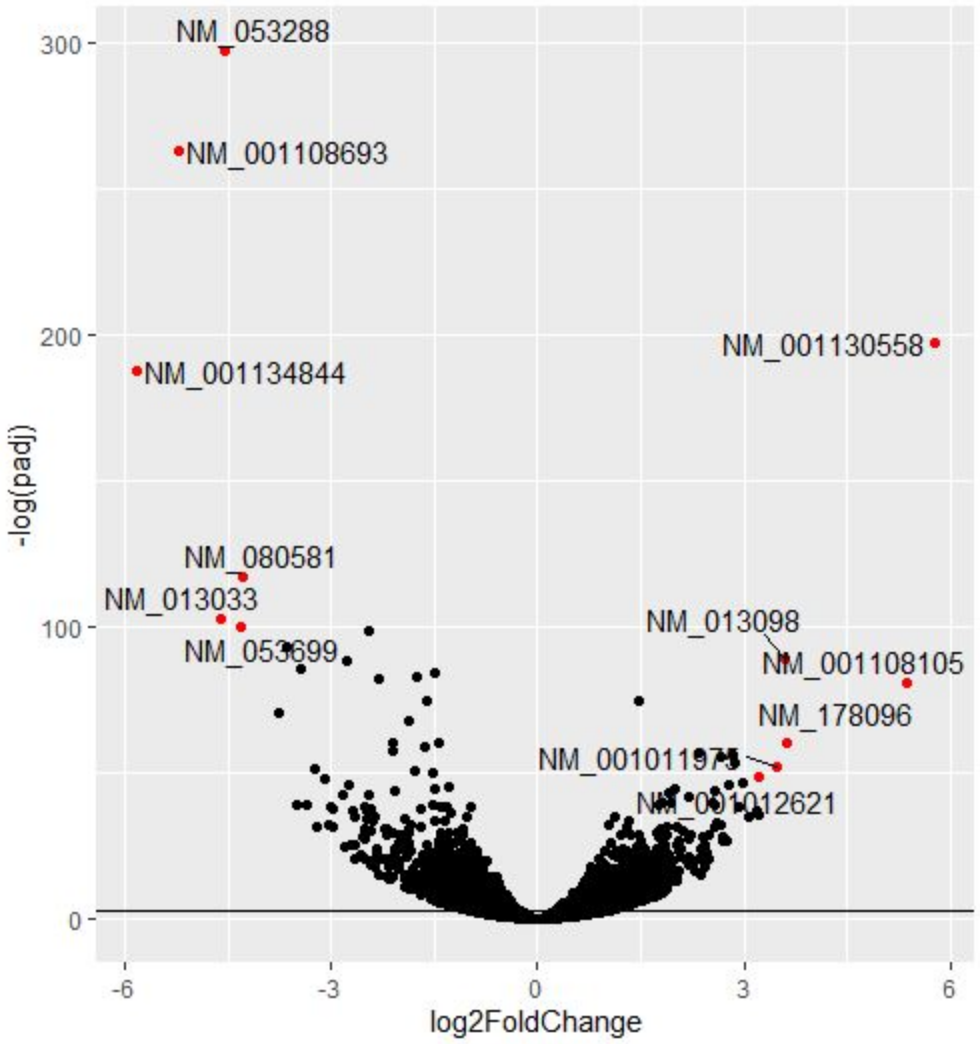

Fig.17 Expression log-fold change against p-value for CAR/PXR samples

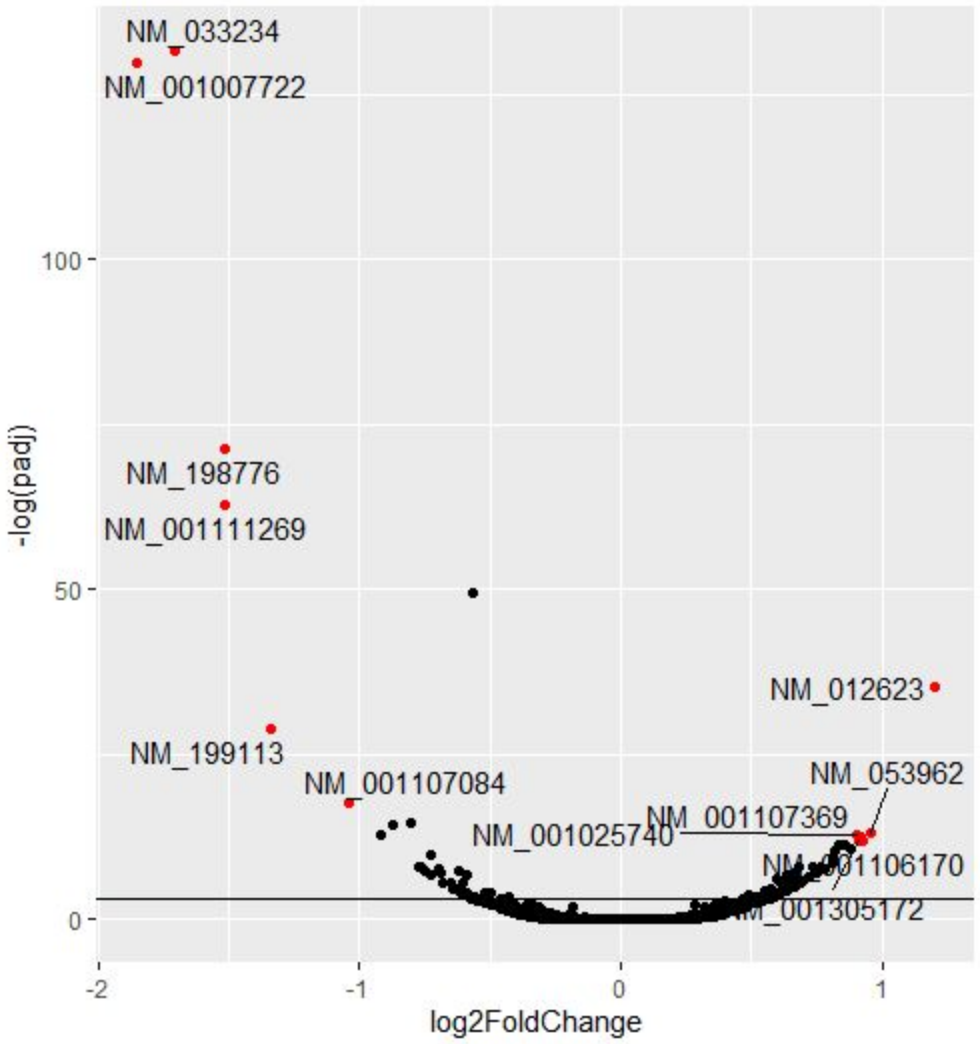

Fig.18 Expression log-fold change against p-value for DNA damage samples

	AhR	CAR/PXR	DNA DMG
Significant genes count	1392	3503	114
1st	NM_013096	NM_053288	NM_033234
2nd	NM_033234	NM_001108693	NM_001007722
3rd	NM_001007722	NM_001130558	NM_198776
4th	NM_001257095	NM_001134844	NM_001111269

5th	NM_001008880	NM_080581	NM_013096
6th	NM_001130558	NM_013033	NM_012623
7th	NM_012540	NM_053699	NM_199113
8th	NM_198776	NM_024127	NM_001107084
9th	NM_012541	NM_031048	NM_001113223
10th	NM_130407	NM_013098	NM_001013057

Table.2 Significant genes and top hits report

Analysis of the probeset showed that probes got the most significant results, with THIOACETATE having 5865 total passing with a p value lower than 0.05, while ECONAZOLE had 106, and BETA-NAPHTHOFLAVONE had 106. Unlike in the paper, we did not have an absolute 1.5 Fold Count cutoff. We also made a histogram of Fold Counts, as indicated in the summary of steps, as opposed to log of fold count, though this might have been in error.

Our most significant results for each probe follow:

BETA-NAPHTHOFLAVONE:

	logFC	AveExpr	t	P.Value	adj.P.val	B
1387243_at	1.6741589	13.403113	25.693972	3.72E-17	1.16E-12	20.849144
1370269_at	7.1823046	8.401583	20.653905	2.87E-15	4.46E-11	19.060397
1370613_s_at	0.9243164	12.845754	14.85764	1.70E-12	1.77E-08	15.58545
1387759_s_at	1.1355707	11.958387	12.707636	3.17E-11	2.46E-07	13.664844
1376252_at	-0.5658033	6.477249	-8.895145	1.66E-08	1.04E-04	8.944298
1391981_at	-0.5563853	6.959432	-7.741988	1.55E-07	7.24E-04	7.097065
1368168_at	-1.3557268	8.508688	-7.718005	1.63E-07	7.24E-04	7.056321
1371643_at	-0.6242631	7.659838	-7.414486	3.03E-07	1.18E-03	6.532368
1388271_at	2.4220777	9.881536	7.248775	4.26E-07	1.35E-03	6.239806
1371237_a_at	2.2390596	11.819897	7.227518	4.46E-07	1.35E-03	6.201946

ECONAZOLE:

ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
1386944_a_at	-1.5092537	11.74805	-6.992738	4.46394E-08	0.001388241	7.897598
1398265_at	-0.7598861	6.809951	-6.31982	3.26179E-07	0.002170647	6.215498
1384408_at	1.1688692	6.894928	6.310908	3.34954E-07	0.002170647	6.192886
1398597_at	-0.8714352	6.146704	-6.310356	3.35506E-07	0.002170647	6.191484
1390801_at	-0.8721717	5.239252	-6.297132	3.4899E-07	0.002170647	6.157914
1388874_at	0.6053027	10.95278	6.129144	5.76159E-07	0.0024811	5.730051
1369012_at	-0.6412944	6.022861	-6.098994	6.30497E-07	0.0024811	5.652996
1370725_a_at	-1.7552623	10.545257	-6.094908	6.38246E-07	0.0024811	5.642548
1396901_at	-0.5014107	4.884597	-5.965127	9.41156E-07	0.003193538	5.309983
1370698_at	1.1978244	11.932738	5.936022	1.02689E-06	0.003193538	5.235225

THIOACETAMIDE

Column1	logFC	AveExpr	t	P.Value	adj.P.Val	B
1369268_at	5.392573	6.454692	51.73372	1.14094E-14	3.54822E-10	20.64125
1395557_at	3.681872	6.62649	34.31727	1.09337E-12	1.70013E-08	18.18192
1367764_at	2.765759	8.676252	30.15617	4.57382E-12	3.98971E-08	17.20685
1368072_at	3.451321	7.159355	29.84391	5.13163E-12	3.98971E-08	17.12446
1391032_at	-4.208859	8.505141	-28.79483	7.62037E-12	4.73972E-08	16.83699
1382490_at	-2.798262	7.2466	-26.13409	2.22047E-11	1.01639E-07	16.02669
1371785_at	3.726813	7.177624	25.94253	2.40789E-11	1.01639E-07	15.96341
1373767_at	4.335179	7.692554	25.74924	2.61459E-11	1.01639E-07	15.89884
1370583_s_at	4.305828	7.267736	24.46541	4.58962E-11	1.58592E-07	15.45071
1385132_at	3.868307	5.89858	22.5115	1.14432E-10	3.11761E-07	14.69842

The histograms show that FoldChange is mostly a little under 1 for all 3 types of probes, though some are much larger. There is a “valley” in the center since though there are many genes in that area, they have very high p values, as we found out by accidentally making these with all the wrong genes first.

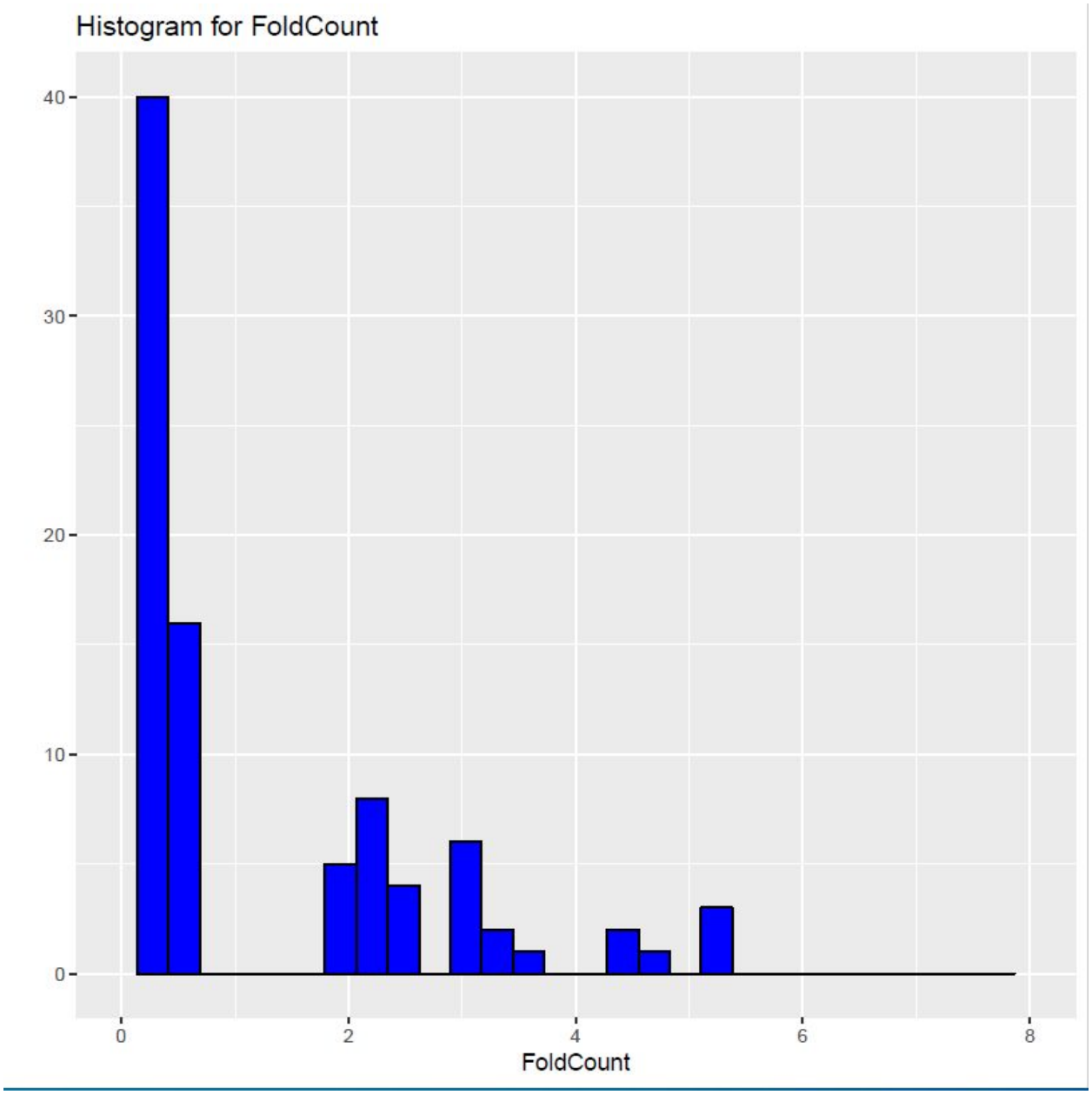

Fig. 19: Histogram of fold change for BETA-NAPHTHOFLAVONE

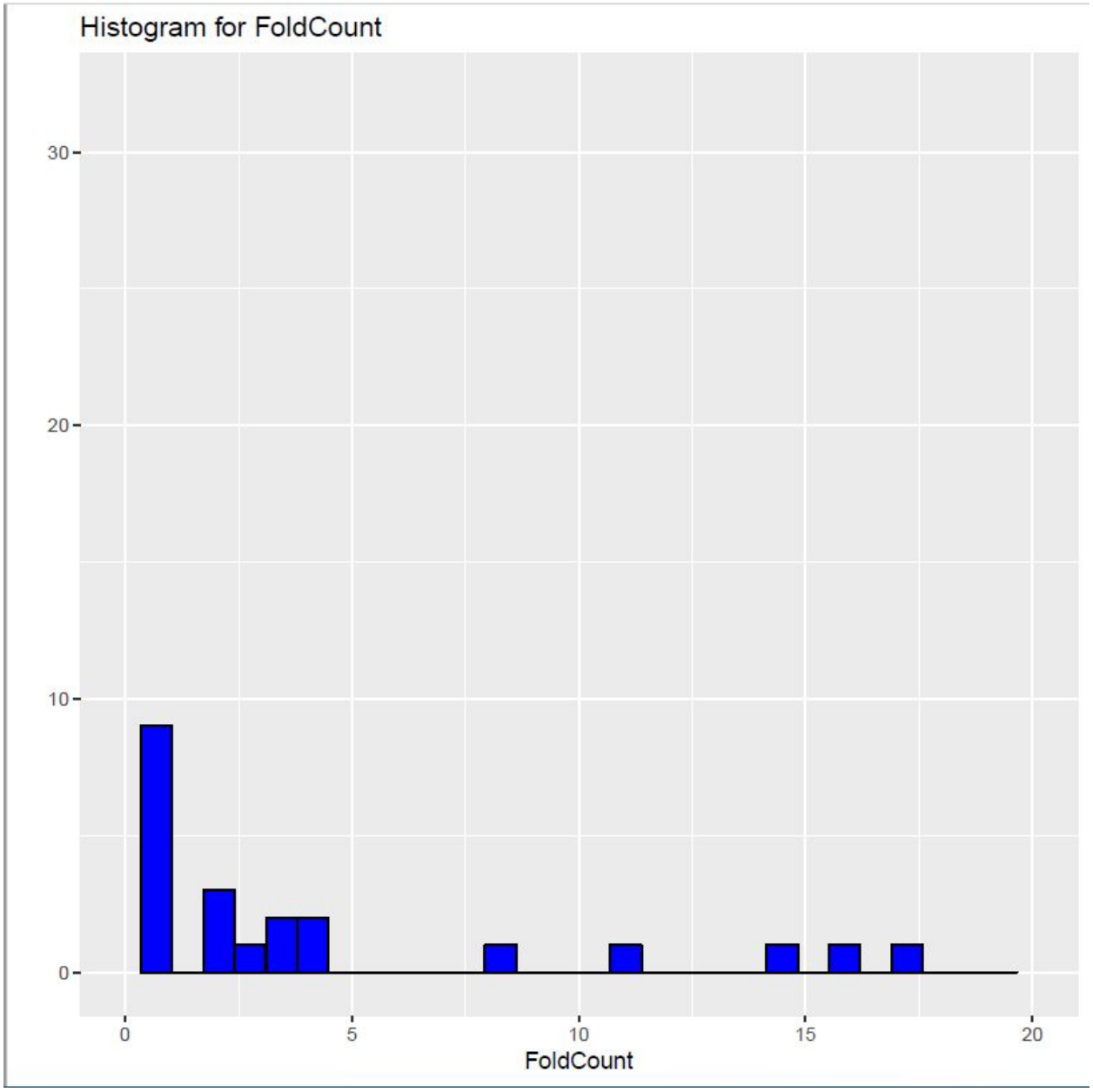

Figure 20:Histogram of fold change for ECONAZOLE:

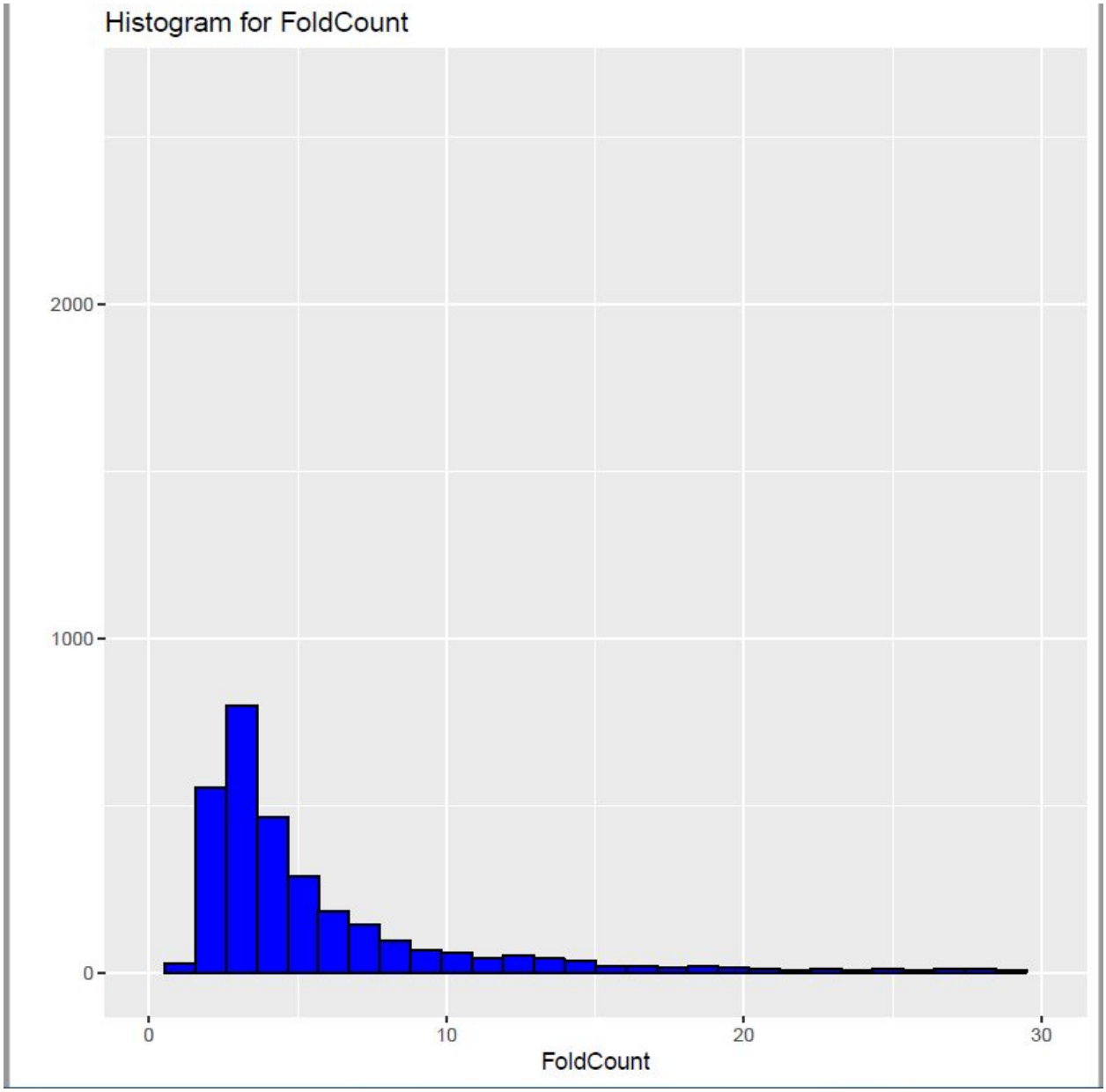

Figure 21:Histogram of fold change for THIOACETAMIDE:

The scatterplots of fold c vs P-Value also all showed that as fold change got further from 1, p value tended to decrease, and that point with a high p-value tended to be scattered in a very narrow range between a 1 and 2 (or ½ and 1 on the lower end) fold change ratio.

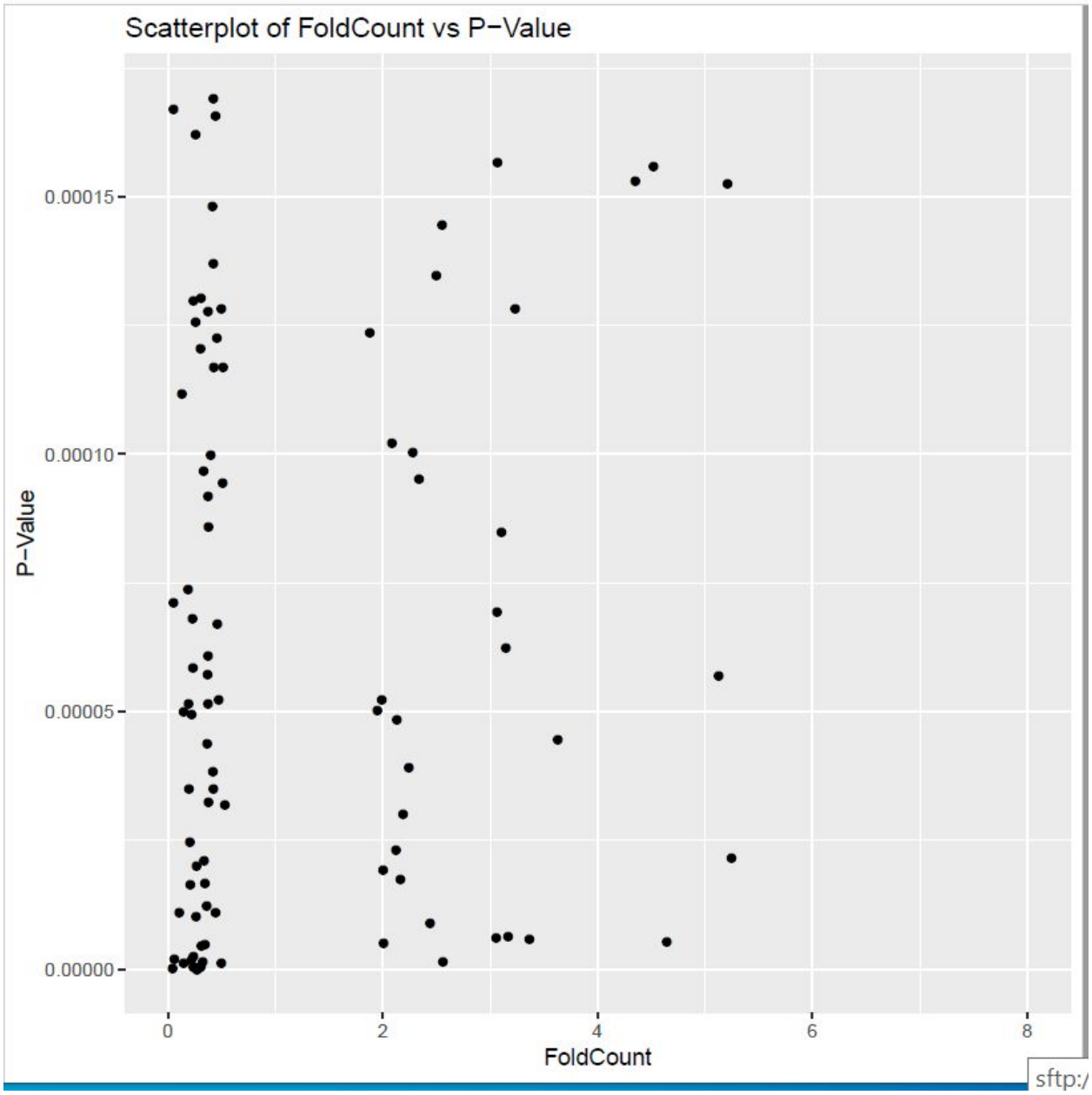

Figure 22:Scatterplot of fold change vs p-value for BETA-NAPHTHOFLAVONE

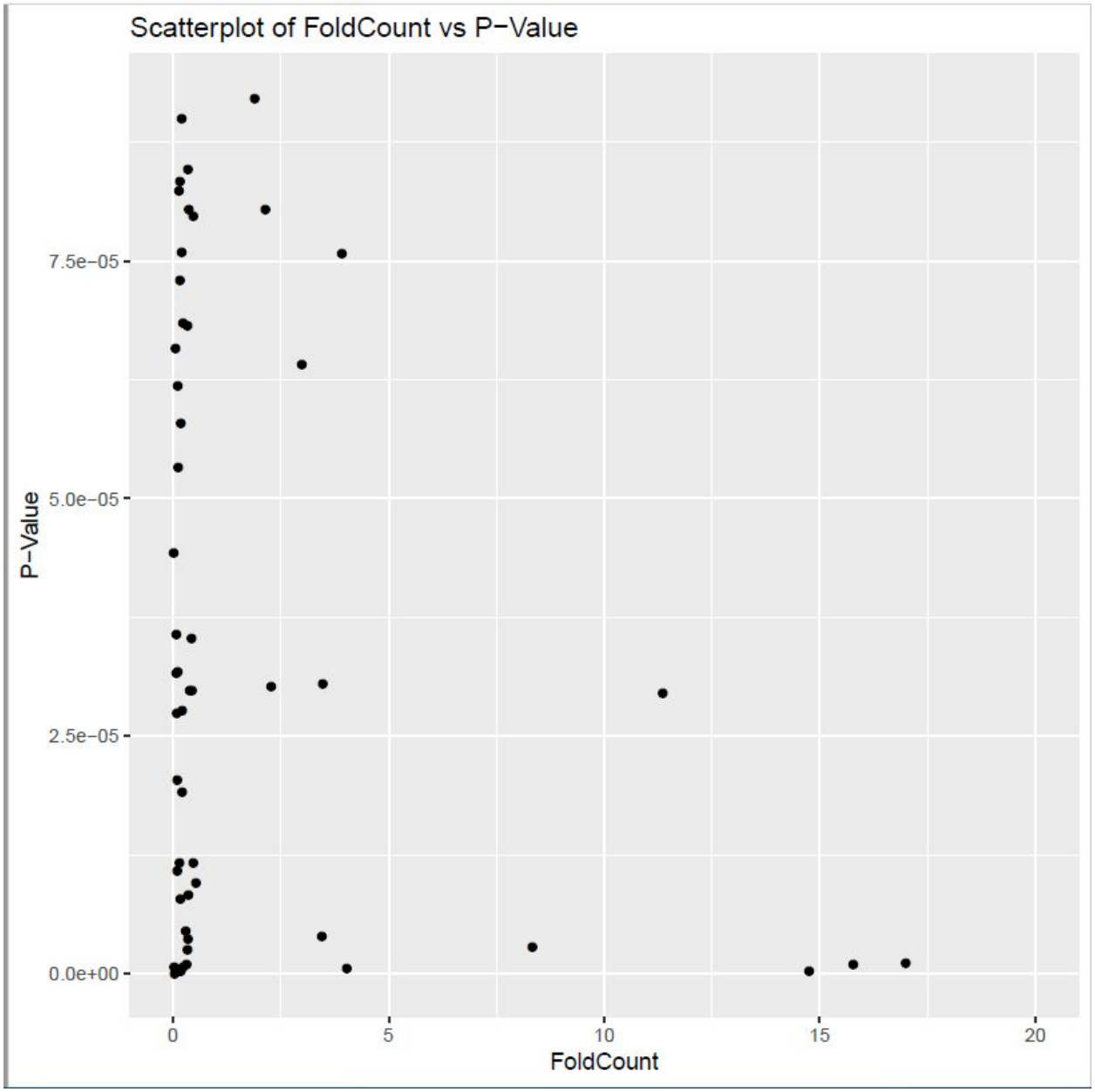

Figure 23:Scatterplot of fold change vs p-value for ECONAZOLE

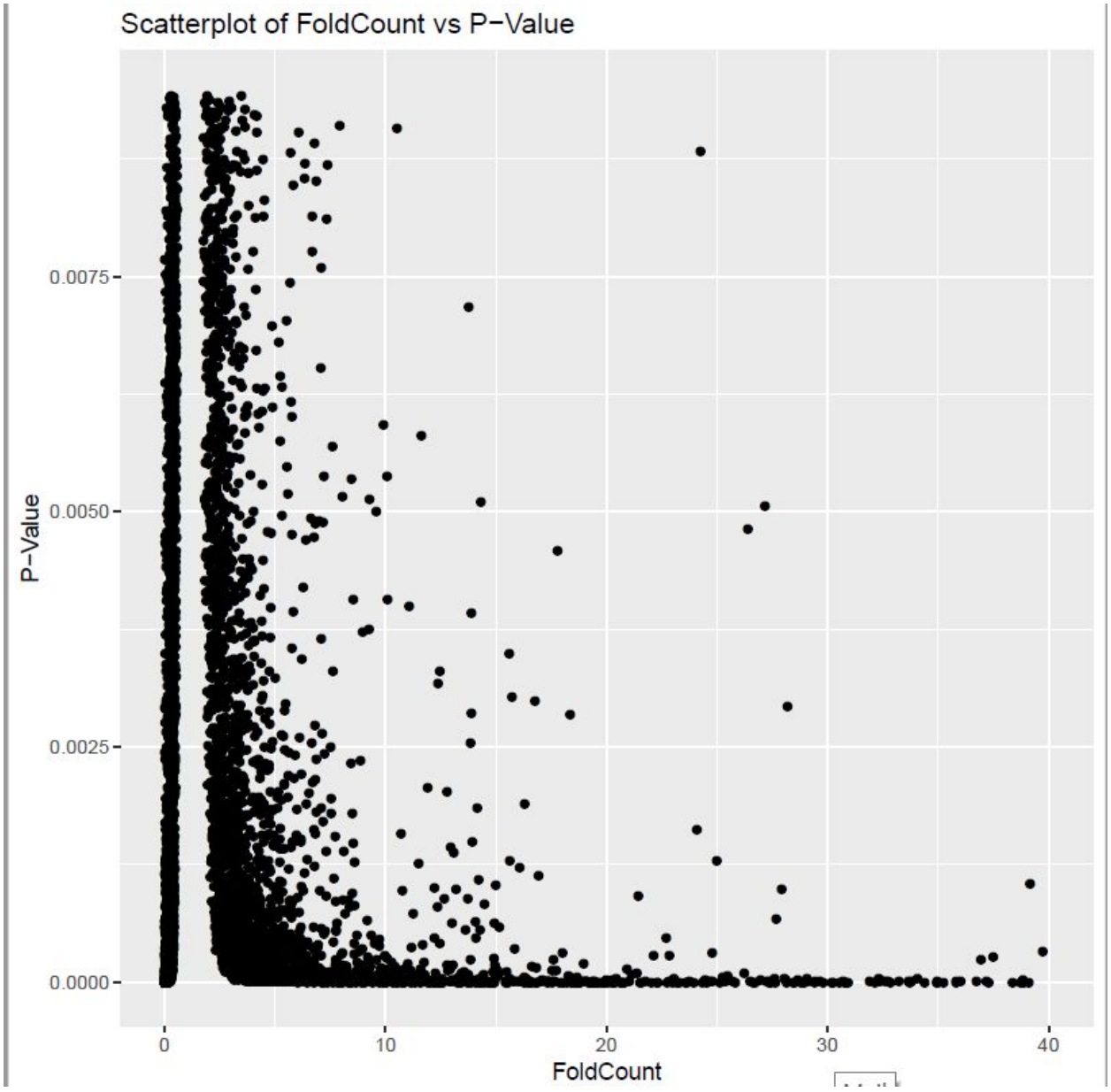

Figure 24:Scatterplot of fold change vs p-value for THIOACETAMIDE

Our concordace calculation produced no results

Discussion

The quality of the raw sequencing data was evaluated based on the quality control metrics from fastqc and had shown acceptable results. Some of the failed tests were due to the nature of RNA-seq data and it was not necessary to take any further steps to improve it. Overall, the

sequencing data was suitable for the following analysis. These sequencing data was then successfully aligned against the rat genome.

The variation of gene expression changes after the three different chemical treatments vary significantly. This can be observed from both the histograms and the volcano plots. Most significant genes with gene expression level change after treatment have a log2 fold change around 1 for both AhR group and CAR/PXR group. The difference being CAR/PXR has 3503 genes with a significant p value for gene expression change compared to the AhR samples' 1392 significant genes. DNA damage group has only 114 significant genes and for those genes, the fold change are not as high compared with the other two samples.

We attempted to replicate their analysis, but due to making a mistake early on in the process ended up choosing all the wrong genes during the cutoff, which made it impossible to calculate Concordance due to having too much data.

Even with the proper data, too many points pass the threshold and make it take far too long to calculate concordance for the system to work without a batch job. (Having over a million combinations of gene pairs to look up concordance in 3 different memory locations for).

Conclusion

For the data curator part, the quality control step for raw sequencing data and the sequence alignment step were successfully performed and showed acceptable results. One challenge for this part was to figure out the rationale behind each quality control metrics and explore the reason for the failed tests for RNA-seq samples. The note [5] explained them well.

The MultiQC showed good alignment results for gene annotation. The range of gene counts is high across all samples which is to be expected in an RNA seq experiment. Good number of genes has shown significant change in expression level across all three treatment groups.

The analyst section was also difficult, and not made easier by having to recalculate everything less than an hour before submitting.

References

[1] Wang C, Gong B, Bushel PR, et al. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nat Biotechnol. 2014;32(9):926–932. doi:10.1038/nbt.3001

[2] https://www.ncbi.nlm.nih.gov/sra/?term=SRP039021

[3] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55347

[4] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47875

[5] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[6] Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047–3048. doi:10.1093/bioinformatics/btw354

[7] https://github.com/alexdobin/STAR/blob/2.5.3a/doc/STARmanual.pdf

[8] Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21. doi:10.1093/bioinformatics/bts635