Individual Project

Introduction

For the first task, I have decided to perform the data curator task from project 3. This project focuses on studying the concordance between microarray and RNA-Seq methods, using gene expression data from a toxicological treatment study using rat livers. The data curator role consists of checking the quality of the files, aligning the files to a genome, and using both sets of results to run multiqc. Using fastqc to check the quality of the files is important to the overall study, because it makes sure that the data doesn't contain any problems before running any analysis. Since the study is using rats as a model organism, it is necessary to align any data to the rat genome in order to identify the genes in the sequencing data. Finally, multiqc allows us to visualize all of the summary statistics across multiple samples at once.

Methods

To begin with, I had to select a toxgroup to perform the data curation on. Since my group had previously worked with toxgroup 1, I chose to use the samples from toxgroup 2 instead. Toxgroup 2 consisted of three toxins (Beta-Naphthoflavone, Econazole, and Thioacetamide), with three samples for each toxin. In order to obtain access to the sample files, I linked each of the 9 samples (SRR1177966, SRR1177969, SRR1177970, SRR1177993, SRR1177994, SRR1177995, SRR1177998, SRR1178001, and SRR1178003) from the project 3 directory to my scc directory. The first step in quality control was to run fastqc on each of the 18 fastq files through the terminal. In order to do this, I used a qsub script which called fastqc on each sample, and outputted an html file and a zip file with all the summary statistics into a new samples directory.

Next, STAR was used to map the RNA-seq data to the rn4_STAR rat reference genome. The paired end reads alignment produced one BAM file for each sample. According to the log.final.out files, there is on average a 198 mapped length. Multiqc was also executed on the command line using a qsub script. The fastqc outputs were combined and summarized into a new fastq directory, while the STAR outputs were combined and the summarized statistics were outputted to a bam directory.

Results

Looking at the fastqc html files, we are able to evaluate the quality of the sequencing data. All of the files have good base statistics and good Per sequence quality scores. However, both the Per base sequencing quality and Per base sequence content show a failure.

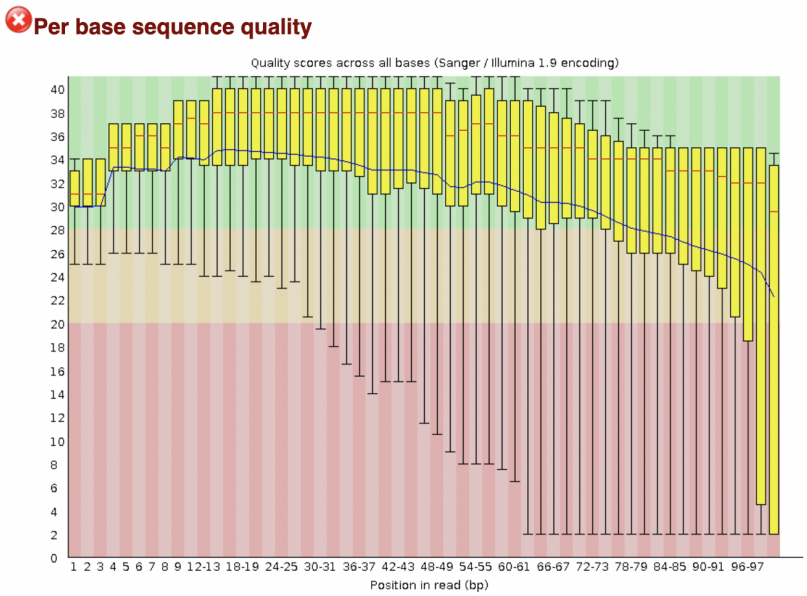

Figure 1: Fastqc Per base sequence quality box and whisker plot of the Phred quality scores at different read positions.

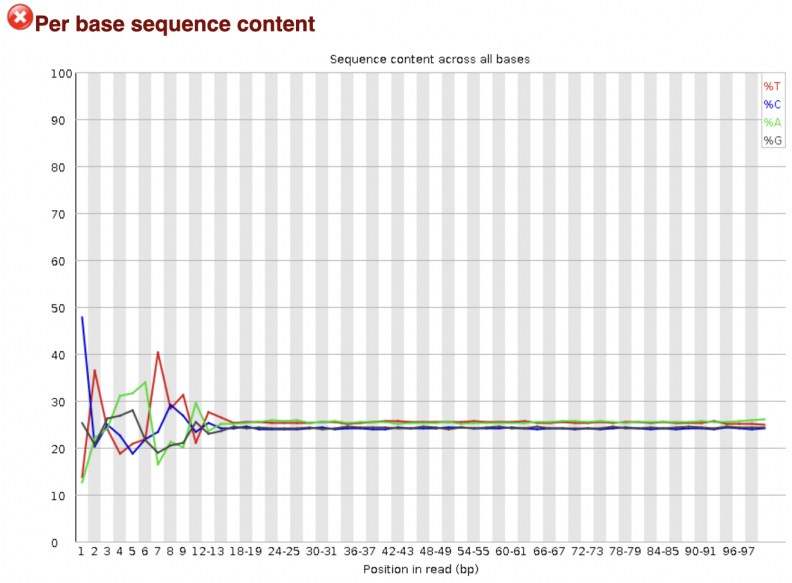

Figure 2: Per base sequence content plot, showing the percentage of each base at different read positions

One of the plots from the fastqc portion of the multiqc output is a histogram of sequence counts (Figure 3). Here we can see that for six of the samples, there are on average 20 million sequence counts for each, and around 25 million sequence counts for the other three samples. For all nine

samples there are around 10 million unique reads. Looking at the Fastqc Mean quality scores plot in Figure 4A, we can see that 4 samples passed, 1 had a warning, and 13 samples failed. However, since the majority of the samples are within the green Phred score region, we can conclude that they are good quality samples. Figure 4B graphs the Per sequence Quality scores, and the distribution shows that the majority of sequences have a high Phred quality score between 30 to 40. The Per sequence GC content graph in Figure 4C shows a distribution of GC percentages. All of the samples have a close to normal distribution with a median at 50%, which is good. The last fastqc plot of the Per base N content shows the percentage of bases at each position with no base call. Since all of these values are low for each position, there are relatively low no base calls indicating good sequence quality.

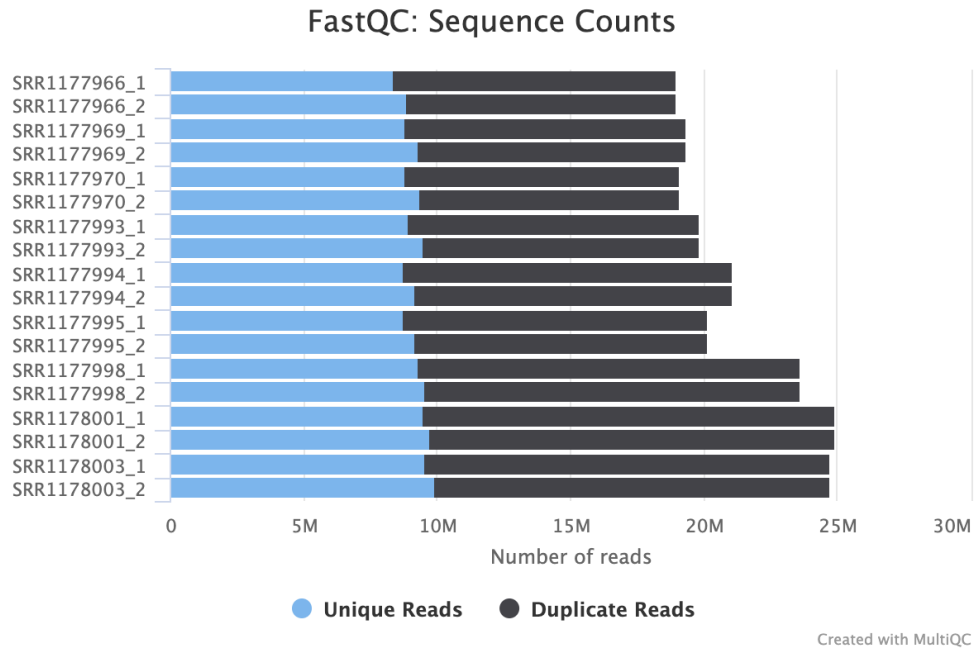

Figure 3: Fastqc sequence counts histogram showing the number of unique and duplicate reads per sample

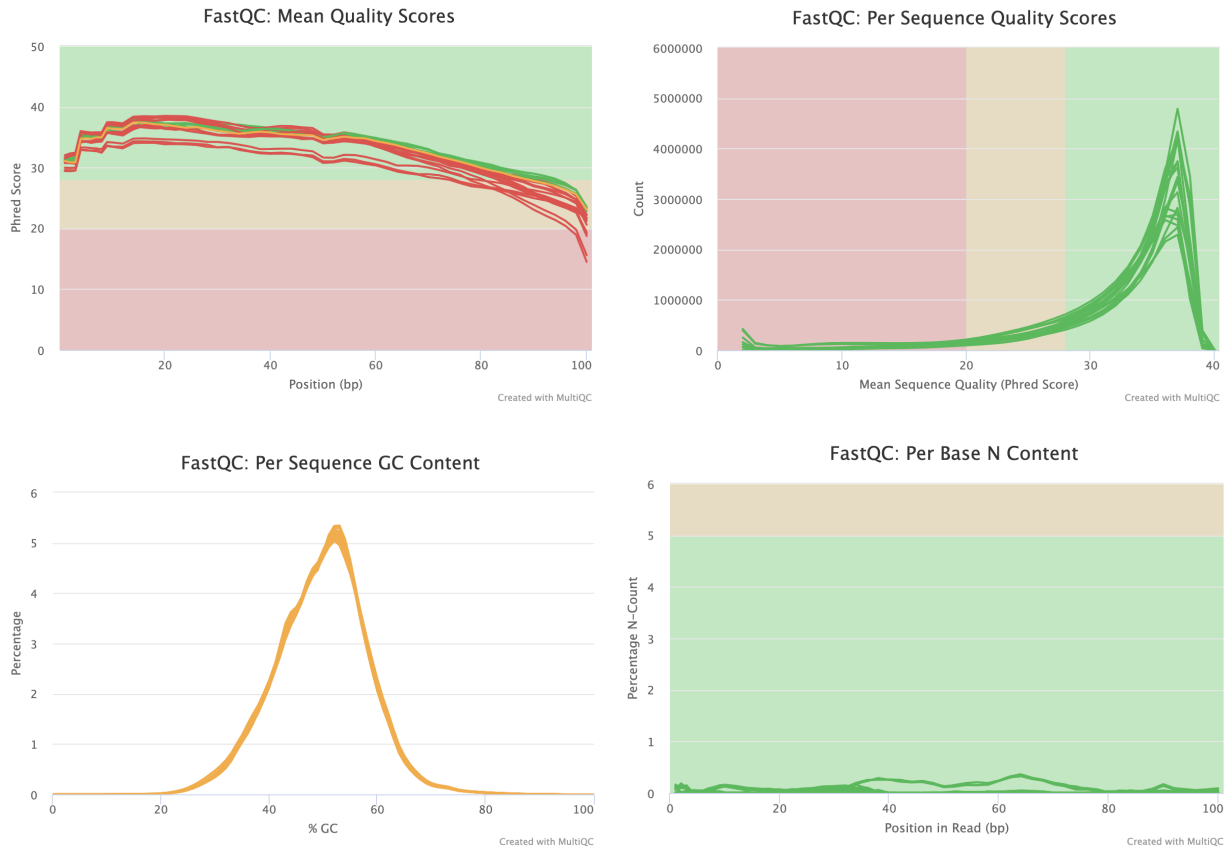

Figure 4 (A, B, C, and D): A: multiqc output of mean quality scores, which plots the Phred quality scores of all the samples. B: Per sequence quality scores which plots the count of each quality score for all the samples. C: Per sequence GC content plots the distribution of GC content percentages. D: Per base N content plots the percentage of Ns at each position.

From the STAR outputs, multiqc compiles a report of alignment scores, which has been converted to a table format in Table 1. Here, we can see how the sequences have aligned to the genome. The majority of each sample is uniquely mapped to the genome, with percentages between 80% and 90%. For cases of mapping to multiple loci and being too short to map, there are much lower percentages between 3% and 11%.

Category	Uniquely mapped	Mapped to multiple loci	Mapped to too many loci	Unmapped: too short	Unmapped: other
SRR1177966	15983143 (84.2%)	807481 (4.3%)	38159 (0.2%)	2134238 (11.2%)	11383 (0.1%)
SRR1177969	16494573 (85.4%)	695194 (3.6%)	23159 (0.1%)	2087700 (10.8%)	9656 0.1%)
SRR1177970	16262555 (85.1%)	670339 (3.5%)	30790 (0.2%)	2148781 (11.2%)	7654 (0.0%)
SRR1177993	17094859 (86.2%)	858349 (4.3%)	56194 (0.3%)	1798631 (9.1%)	15864 (0.1%)
SRR1177994	18514298 (88.0%)	859510 (4.1%)	39809 (0.2%)	1612036 (7.7%)	18904 (0.1%)
SRR1177995	17661373 (87.6%)	794497 (3.9%)	40093 (0.2%)	1633442 (8.1%)	22183 (0.1%)
SRR1177998	13374032 (83.6%)	784376 (3.9%)	51179 (0.3%)	1777306 (11.1%)	16012 (0.1%)
SRR1178001	22291635 (89.1%)	974756 (3.9%)	64118 (0.3%)	1642973 (6.6%)	34957 (0.1%)
SRR1178003	22041739 (89.2%)	926439 (3.7%)	57984 (0.2%)	1670458 (6.8%)	24711 (0.1%)

Table 1: This table shows the number of reads per sample that are uniquely mapped, mapped to multiple loci, mapped to too many loci, or are unmapped.

Discussion

Fastqc, STAR, and multiqc were used to process the samples. Fastqc was used to check the quality of the data, STAR was used to align the data to a reference genome, and multiqc was used to compile the results for analysis.

Taking a closer look at the base sequencing quality plots in Figure 1, there is always a drop in quality at the end which is indicative of either signal decay or phasing. Additionally, in RNA-seq library preparation, random hexamer priming occurs in the first 10-12 bases, resulting in poor Per base sequence content (Figure 2). Since both of these failures are to be expected, and all of the fastqc plots in Figure 4 show good results in the green regions, we can conclude that our sample data is of good quality.

The alignment table shows high percentages for uniquely mapped sequences, which can be interpreted as good alignment. Since we have determined that our samples are of good quality it is no surprise that there is also good alignment, and no high percentages of unmapped sequences.

Introduction

For the second part, I have decided to do the Biologist role for project 3. This task consists of performing gene enrichment on the differentially expressed genes from the treatments, and creating a clustered heatmap using the normalized expression matrix. I will be using the previously produced DE genes and normalized expression matrix from my project group. This analysis is important to look at enriched pathways which are expressed for different treatments, and the clustered heatmap allows us to see if different modes of action (MOA) cluster together.

Methods

In order to look at the enriched KEGG pathways of the differentially expressed genes, first I had to obtain a list of DE genes for each of the three treatments. To do this, I downloaded the three csv files for the AhR, CAR/PXR, and Cytotoxic mode of action toxin treatments. By filtering for p-values <.05, I got a list of DE genes, which were entered into the DAVID annotation tool for analysis. The AhR mode of action had a total of 1647 DE genes, the CAR/PXR mode of action had 2115 DE genes, and the Cytotoxic mode of action had 3309 DE genes. The RefSeq RNA identifier was used, and the KEGG pathway was selected for the pathway type, allowing us to view the enriched pathways.

For the heatmap, the csv file containing the normalized count matrix for all three treatments was used to produce the visualization. In order to do this, I converted the csv file into a data frame and filtered for rows which did not contain all zeros. After filtering, I converted the data frame into a matrix and used the heatmap function to plot a clustered heatmap with the Gene ids as the y-axis labels and the sample ids as the x-axis labels.

Results

From Table 2, we can see the enriched KEGG pathways found using DAVID for each of the three modes of action. The top KEGG pathway for each of the MOAs is Metabolic pathways, followed by chemical carcinogenesis for the AhR and CAR/PXR MOAs.

The clustered heatmap in Figure 5 shows three distinct clusters based on the dendrogram at the top. There are two columns which are significantly darker than the rest, indicating that these samples have more differential expression of the DE genes.

AhR KEGG Pathway

Category	Term
KEGG_PATHWAY	Metabolic pathways
KEGG_PATHWAY	Chemical carcinogenesis - reactive oxygen species
KEGG_PATHWAY	Diabetic cardiomyopathy
KEGG_PATHWAY	Metabolism of xenobiotics by cytochrome P450
KEGG_PATHWAY	Chemical carcinogenesis - DNA adducts
KEGG_PATHWAY	Drug metabolism - cytochrome P450
KEGG_PATHWAY	Fatty acid degradation
KEGG_PATHWAY	Biosynthesis of cofactors

CAR/PXR KEGG Pathway

Category	Term
KEGG_PATHWAY	Metabolic pathways
KEGG_PATHWAY	Chemical carcinogenesis - DNA adducts
KEGG_PATHWAY	Drug metabolism - other enzymes
KEGG_PATHWAY	Metabolism of xenobiotics by cytochrome P450
KEGG_PATHWAY	Bile secretion
KEGG_PATHWAY	Steroid hormone biosynthesis
KEGG_PATHWAY	Ascorbate and aldarate metabolism
KEGG_PATHWAY	Retinol metabolism

Cytotoxic KEGG Pathway

Category	Term
KEGG_PATHWAY	Metabolic pathways
KEGG_PATHWAY	Coronavirus disease - COVID-19
KEGG_PATHWAY	Chemical carcinogenesis - DNA adducts
KEGG_PATHWAY	Steroid hormone biosynthesis
KEGG_PATHWAY	Complement and coagulation cascades
KEGG_PATHWAY	Biosynthesis of cofactors
KEGG_PATHWAY	AMPK signaling pathway
KEGG_PATHWAY	Drug metabolism - other enzymes

Table 2 (A, B, and C): KEGG pathway annotation tables for the MOA group 1 (AhR, CAR/PXR and Cytotoxic) from DAVID

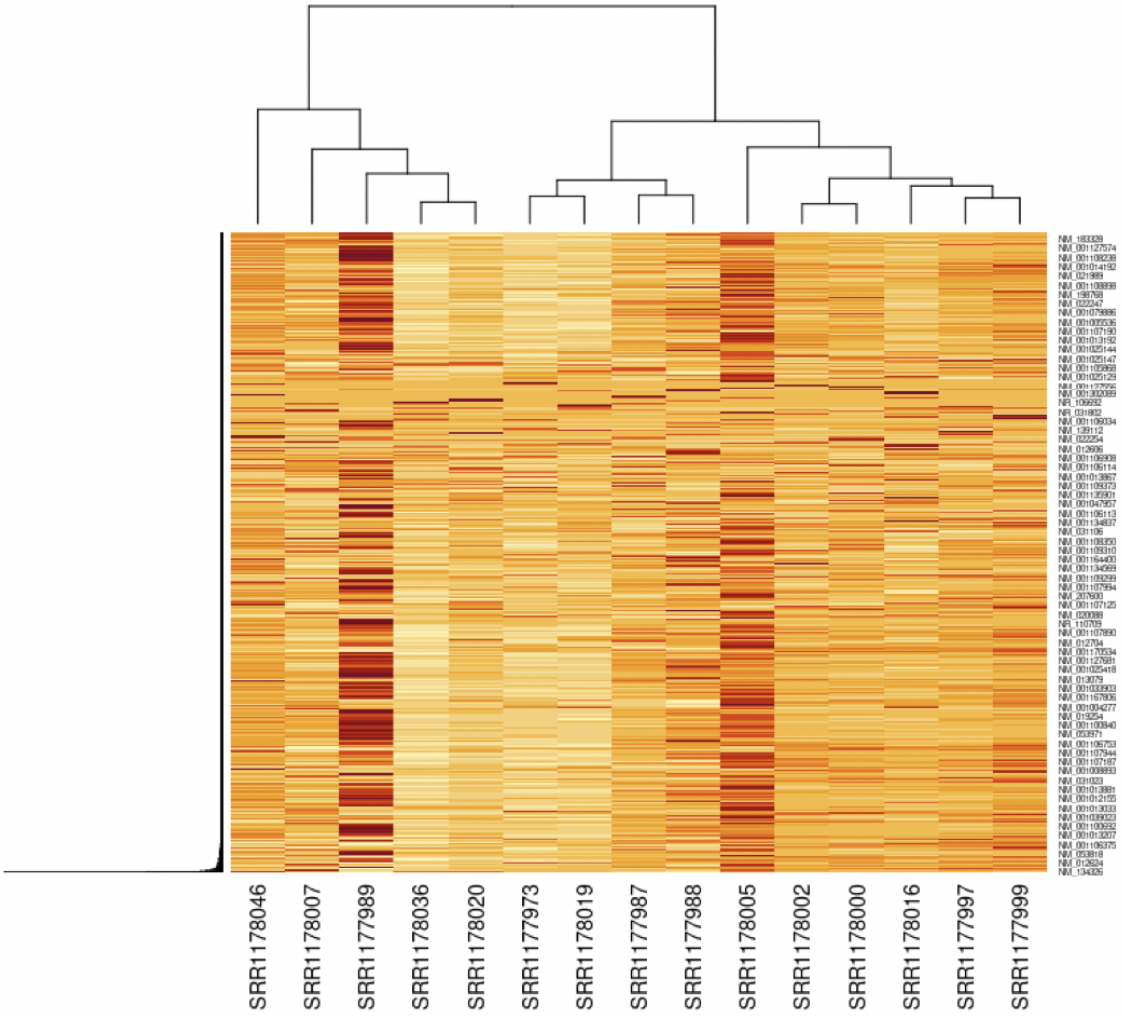

Figure 5: Cluttered heatmap of gene expression found using the normalized counts matrix for each MOA in toxgroup 1 (AhR, CAR/PXR, Cytotoxic).

Discussion

The goal of this analysis was to look at any upregulated KEGG pathways and clustering of DE genes for the samples from AhR, CAR/PXR, Cytotoxic MOAs. The results show that there are some shared pathways between the MOAs, but little clustering within the MOAs. The upregulated pathways found for the AhR MOA consists of many melatonin, nicotine, and acetone degradation pathways. Although there are also degradation pathways present in the annotated KEGG pathway results for our AhR samples, they are focused on degrading other

molecules. The upregulated pathways for the CAR/PXR MOA determined in the paper were metabolism, detoxification, and signaling pathways. Comparing these to our CAR/PXR pathways, there are more pathways in common, specifically xenobiotic metabolism pathways. The upregulated pathways for the Cytotoxic MOA in the paper consist of biosynthesis, degradation, and metabolism signaling pathways. Comparing these to the pathways in our analysis, we can see that many of the metabolism and degradation pathways are shared. There is not a complete overlap between the KEGG pathways for each MOA in the paper with our list of pathways since we only have samples for one toxin for each of the MOAs. Although, it makes sense that toxins with the same modes of action will result in similar pathways to be differentially expressed due to how the toxin interacts with the organism.

The first group of the heatmap dendrogram on the left contains three samples from the Cytotoxic MOA and two CAR/PXR samples. The second group contains two AhR and three CAR/PXR MOA samples. The last group contains three AhR and three Cytotoxic MOA samples. Although none of the MOAs have completely clustered separate from another, we can also see that this clustering shows the similarities in DE genes between the three MOAs. The two columns with the darkest colors are a CAR/PXR and a Cytotoxic MOA sample. Although the different MOAs may have differing upregulated pathways, it is not surprising that there are similarities between the DE genes of the different MOAs. This is because although the toxins are different, the organism is still required to react to a foriegn harmful substance which requires specific reactions from the body of the organism.

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. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc*. 2009;4(1):44-57.

3. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009;37(1):1-13.