Concordance of Microarray and RNA-Seq Differential Gene Expression

Aneeq Husain

Roles chosen: Project 3 Programmer and Biologist

Introduction:

In this study, I processed the RNA-Sequencing (RNA-Seq) data, performed differential expression analysis and further interpreted the results using functional enrichment analysis and heatmaps. Wang *et al.*'s [1] study was a large-scale study that attempted to study the concordance between microarray and RNA-Seq results. Their study is significant in that microarray was an industry standard that is being phased out in favour of RNA-Seq and hence it is necessary to understand how comparable the two methods are. An essential part of their study relies on processing the RNA-Seq data appropriately so that the interpretations made are accurate and scientifically valid.

Methods:

RNA-Seq processing was performed by computing the counts and performing differential expression analysis. Counting was done using the featureCounts [2] package (v1.6.2) on 9 sample BAM files. MULTIQC (v1.10.1) [3] was also run to check the quality of the resulting counts. Subsequently, the counts matrices were imported to R (v3.14) [4] for normalization and differential expression. Both steps were performed using the DESeq2 package (v1.34.0) [5]. Here, the samples were split into three sub-groups based on the mode of action. The subgroups are listed below in Table 1. The tidyverse (v1.31) [6] was used for data wrangling and plots were produced using the ggplot2 package (v3.3.5) [7].

For biological interpretation, the differentially expressed genes obtained from the previous steps were supplied to DAVID [8] for functional enrichment analysis. DAVID was run on each subgroup's DE genes with the identifier set to GENBANK_ACCESSION. The default annotation categories were selected for clustering. Since DAVID allows clustering only up to 3000 genes and the CAR/PXR subgroup had 3499, the genes with the highest p-values were omitted to perform functional annotation. The heatmaps were made with the help of pheatmap (v1.0.12) [9] in R using the normalized counts as produced by DESeq2.

Sample Name	Method of action
SRR1177981	DNA Damage
SRR1177982	DNA Damage
SRR1177983	DNA Damage
SRR1178008	AhR
SRR1178009	AhR
SRR1178010	AhR
SRR1178014	CAR/PXR
SRR1178021	CAR/PXR
SRR1178047	CAR/PXR

Table 1: The distribution of samples and their associated subgroups.

Results:

The featureCounts package was used to produce counts matrix of the genes for all samples. The resulting counts matrices appeared to be of good quality with high assignments (55% at the lowest) across all samples (Fig. 1). To further understand the quality, the distributions of the counts were plotted in the form of a box plot (Fig. 2). The distribution of counts is fairly consistent across the samples.

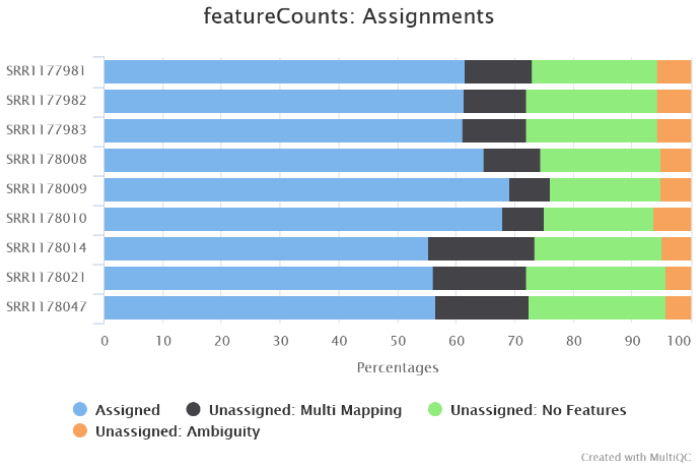

Figure 1: Assignments across the nine samples in percentage.

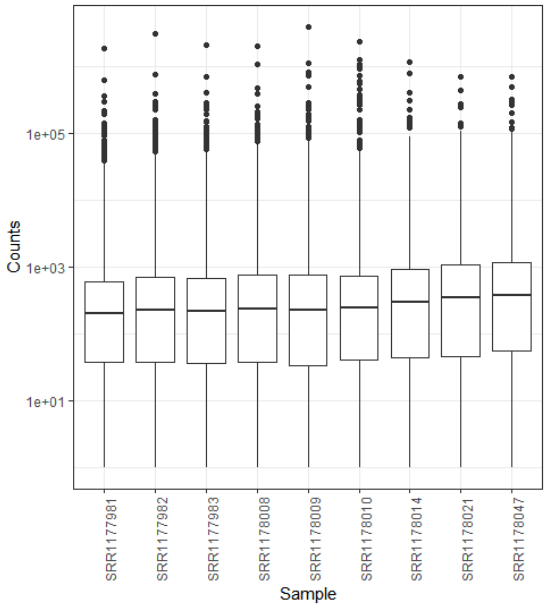

Figure 2: Boxplot showing distribution of counts of the nine samples. Y-axis is log10 scaled.

After normalization and differential expression analysis was performed on the counts, the results were filtered to contain only significant genes (Adjusted p-value < 0.05). This produced results that varied largely by subgroups. It was observed that the DNA damage subgroup contained the least number of significant genes with only 91 crossing the threshold. The AhR subgroup contained 1389 genes while the CAR/PXR subgroup was the largest with 3499 significant genes. The top ten differentially expressed genes are summarised below in tables 2 to 4 for each subgroup.

Gene	Adjust P-value	Log2 Fold Change
NM_033234	8.14×10^{-58}	-7.011
NM_001007722	5.03×10^{-57}	-6.876
NM_198776	1.16×10^{-31}	-0.007
NM_013096	3.18×10^{22}	-0.002
NM_012623	7.00×10^{-16}	3.781
NM_199113	2.45×10^{-13}	-2.418
NM_001107084	2.28×10^{-8}	-0.007
NM_001013057	9.03×10^{-7}	-2.589
NM_053962	2.20×10^{-6}	1.627
NM_001271152	3.84×10^{-6}	-1.403

Table 2: Top 10 differentially expressed genes for the DNA damage subgroup

Gene	Adjust P-value	Log2 Fold Change
NM_013096	1.06×10^{-55}	-9.918
NM_033234	7.98×10^{-54}	-10.145
NM_001007722	6.79×10^{-44}	-9.213
NM_001257095	1.99×10^{-40}	-4.455
NM_001130558	1.06×10^{-34}	-7.073
NM_012540	3.22×10^{-30}	9.969
NM_198776	7.50×10^{-30}	-7.500
NM_012541	1.89×10^{-28}	4.329
NM_130407	2.90×10^{-27}	4.023
NM_001012174	5.11×10^{-27}	2.267

Table 3: Top 10 differentially expressed genes for the AhR subgroup

Gene	Adjust P-value	Log2 Fold Change
NM_053288	1.33×10^{-134}	4.792
NM_001130558	1.35×10^{-87}	-6.639
NM_001134844	2.61×10^{-82}	6.923
NM_080581	1.96×10^{-51}	4.899
NM_013033	7.94×10^{-45}	5.587
NM_024127	1.68×10^{-44}	2.541
NM_053699	1.68×10^{-44}	5.079
NM_031048	2.85×10^{-41}	4.069
NM_013098	1.14×10^{-39}	-4.028
NM_017006	4.39×10^{-39}	2.935

Table 4: Top 10 differentially expressed genes for the CAR/PXR subgroup

To visualize the distribution of the differentially expressed genes, plots were made for all three subgroups. First, histograms of the log2 fold change values were made (Fig. 3). All three subgroups showed clear normal distribution of the log2 fold change values. Next, a volcano plot was made of the log2Fold change value versus the adjusted p-values (Fig. 4). It was observed that all the subgroups showed a number of genes that had log2 fold change values >1.5. The AhR subgroup also appeared to have a lot of downregulated genes. The distributions for the DNA damage subgroup are not as clear as a normal distribution but this is likely due to the lower number of significant DE genes.

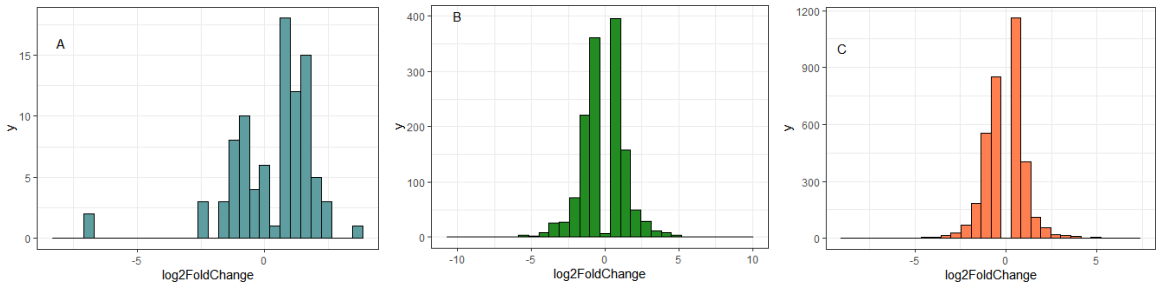

Figure 3: Distribution of the Log2 Fold change values for the three subgroups. Fig. A- Histogram of the log2 fold change values of the DNA damage subgroup. Fig. B- Histogram of the log2 fold change values of the AhR subgroup. Fig. C- Histogram of the log2 fold change values of the CAR/PXR subgroup.

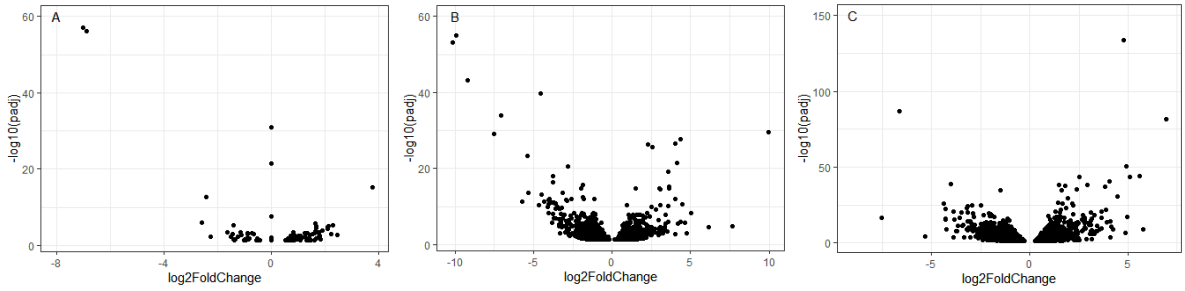

Figure 4: Volcano plots showing distribution of Adjusted P values. Fig A. Distribution of the DNA damage subgroup. Fig. B. Distribution of AhR subgroup. Fig. C. Distribution of CAR/PXR subgroup.

The results of the DAVID analysis are summarized below in the tables 5 to 7. The DNA damage, AhR, CAR/PXR subgroups produced 22,190 and 378 clusters respectively. However, for brevity, the tables only show the top 10 sorted based on the enrichment scores. From the DNA damage subgroups, Wang *et al.* had only two common pathways enriched in both microarrays and RNA-Seq. Of these two, one of them (cell cycle) was found to be enriched in my results as well. The other pathways in Table 5 are also involved in cell cycle. However, no direct overlaps were observed in the AhR and CAR/PXR subgroups. The closest potential pathway overlap observed in the CAR/PXR subgroup was between Drug metabolism in my results and Nicotine degradation II in theirs. The differences in results can likely be explained by the difference in the software and the annotation categories used for functional annotation.

Cluster Term	Enrichment Score	Count	Benjamini corrected p-value
Microtubule binding	3.97	10	3.5×10^{-4}
Erythrocyte development	3.28	6	1.5×10^{-4}
Cell cycle	3.27	15	7.5×10^{-7}
ATP binding	3.13	22	2.5×10^{-4}
Cellular response to bile acid	2.88	3	3.5×10^{-2}
Condensed chromosome	2.34	5	7.1×10^{-4}
Cell division	1.94	7	2.8×10^{-2}
Substrate binding	1.91	8	6.3×10^{-2}
Response to drug	1.54	8	1.9×10^{-1}

Transcription regulation	1.29	12	9×10^{-2}

Table 5: Summarized table of the functionally enriched pathways as produced by DAVID for the DNA damage subgroup.

Cluster Term	Enrichment Score	Count	Benjamini corrected p-value
Cell division	4.75	35	1.2×10^{-3}
Peroxisome	3.83	11	6.7×10^{-3}
Extracellular space	3.46	150	1.9×10^{-3}
Steroid hormone biosynthesis	3.33	20	1.5×10^{-3}
ATP binding	2.45	142	4.4×10^{-3}
Flavoprotein	2.4	21	1.2×10^{-2}
Protein phosphorylation	2.25	56	2.3×10^{-1}
Maturation of 5.8S rRNA	2.18	4	1.6×10^{-1}
Mitotic cell cycle phase transition	2.03	10	1.5×10^{-3}
Thrombospondin	1.93	12	2.4×10^{-1}

Table 6: Summarized table of the functionally enriched pathways as produced by DAVID for the AhR subgroup.

Cluster Term	Enrichment Score	Count	Benjamini corrected p-value
Mitochondrion	13.25	221	5.5×10^{-15}
Protein biosynthesis	7.04	49	1.4×10^{-8}
Isopeptide bond	6.29	116	6.9×10^{-10}
ATP binding	6.22	308	9.7×10^{-10}
Cytoplasm	5.62	338	3.2×10^{-6}
Blood coagulation	5.55	30	6.7×10^{-5}
Flavoprotein	5.31	42	5.2×10^{-6}
Protein folding	4.99	38	1.8×10^{-3}
Drug metabolism	4.95	41	6.4×10^{-7}
Aminoacyl tRNA biosynthesis	4.85	47	3.4×10^{-7}

Table 7: Summarized table of the functionally enriched pathways as produced by DAVID for the CAR/PXR subgroup.

Three heatmaps were produced, with one for each subgroup to observe the clustering of the subgroups. In the DNA damage subgroup, the samples clustered separately with the controls. However, in the AhR subgroup, two of the samples clustered together while the third clustered with the controls. The CAR/PXR subgroup showed the best clustering with all three samples clustering together separate from the controls.

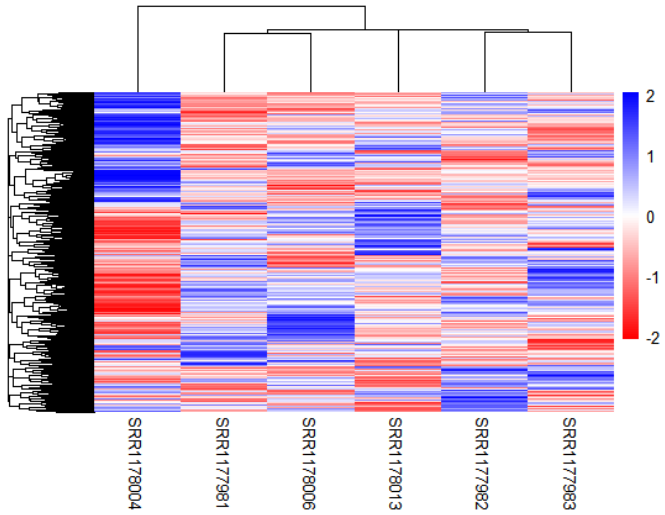

Figure 5: Heatmap showing clustering of DNA damage subgroup with the control samples. Blue indicates upregulation while red indicates down-regulated genes.

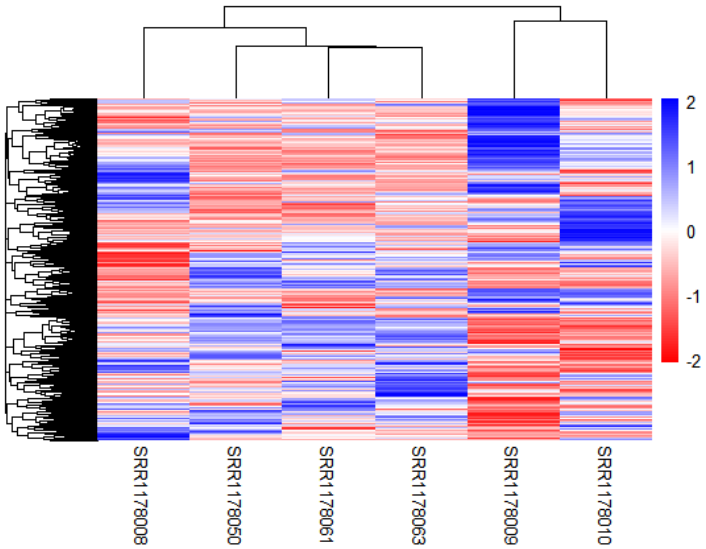

Figure 6: Heatmap showing clustering of AhR subgroup with the control samples. Blue indicates upregulation while red indicates down-regulated genes.

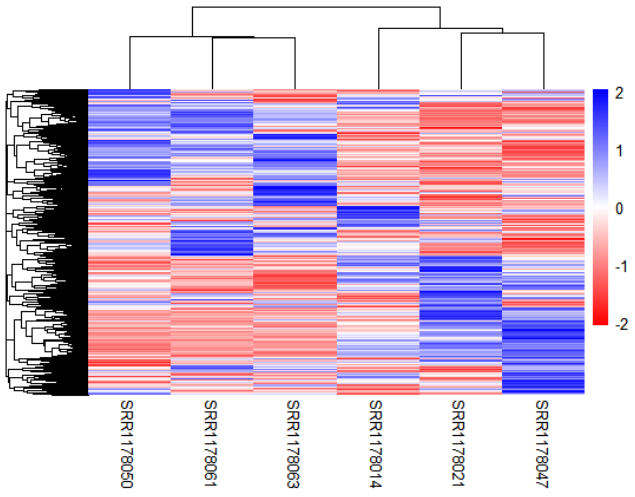

Figure 7: Heatmap showing clustering of CAR/PXR subgroup with the control samples. Blue indicates upregulation while red indicates down-regulated genes.

Discussion:

This project aimed to process the RNA-seq data, perform differential expression analysis on it and understand the results with the help of functional enrichment and clustering. Read counting was performed successfully using featureCounts with good mapping, with the lowest mapped sample still at over 55% uniquely mapped. Normalization and differential expression were then performed using DESeq2 and resulted in the identification of 91, 1389 and 3499 differentially expressed genes for the three subgroups (DNA damage, AhR and CAR/PXR respectively) in the chosen tox-group.

Functional enrichment analysis performed using DAVID did not result in the identification of many pathways that overlapped with Wang *et al.*'s results. But, while the enriched pathways did not show direct matches, it was observed that Wang *et al.*'s results had pathways largely associated with detoxification, drug metabolism and cell cycle. This agrees with the results produced by DAVID. A likely reason for the lack of more direct matches is due to the differences in the processing pipeline. Modifying the pipeline to account for this should result in more overlapping results.

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. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014 Apr 1;30(7):923-30. doi: 10.1093/bioinformatics/btt656. Epub 2013 Nov 13. PMID: 24227677.
3. 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.
4. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
5. Love, M., Anders, S., & Huber, W. (2014). Differential analysis of count data–the DESeq2 package. *Genome Biol*, *15*(550), 10-1186.
6. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of open-source software*, *4*(43), 1686.
7. Wickham, H. (2011). ggplot2. *Wiley interdisciplinary reviews: computational statistics*, *3*(2), 180-185.
8. B.T. Sherman, M. Hao, J. Qiu, X. Jiao, M.W. Baseler, H.C. Lane, T. Imamichi and W. Chang. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Research*. 23 March 2022. doi:10.1093/nar/gkac194.
9. 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.*
10. Kolde, R., & Kolde, M. R. (2015). Package 'pheatmap'. *R package*, *1*(7), 790.