BF528 Individual Project
Jason Yeung

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

Introduction

The motivation behind the research performed in Wang et al. was to quantify concordance between RNA expression datasets generated using microarray and RNA-seq methods. This report outlines the process of counting reads that have been aligned using STAR (Spliced Transcripts Alignment to a Reference) and subsequently performing differential expression analysis to identify genes that are differentially expressed in treatments groups compared to control groups.

Methods

The alignment output from STAR was passed to featureCounts, which quantifies RNA-seq reads and maps the reads to genomic features using a reference annotation file. MultiQC was then run to aggregate output files into a more readable format, as shown in Figure 1 below. The counts for each of the treatment samples, along with their respective control samples, were compiled into a counts matrix for differential expression (DE) analysis. Analysis was then performed using the DESeq2 package for each treatment group in toxgroup 2. This produced differential expression results for each of the beta-naphthoflavone (NAPH), econazole (ECON), and thioacetamide (THIO) treatment groups. Genes were filtered for significance at p-adjust < 0.05.

Results

Results of the featureCounts mapping to genomic features are shown below in Figure 1. The majority of reads were successfully assigned to their respective genes, with another large portion of reads unassigned to features. Visualization of the distribution of gene counts in Figure 2 shows fairly similar distributions across samples, suggesting no outliers with low gene assignments.

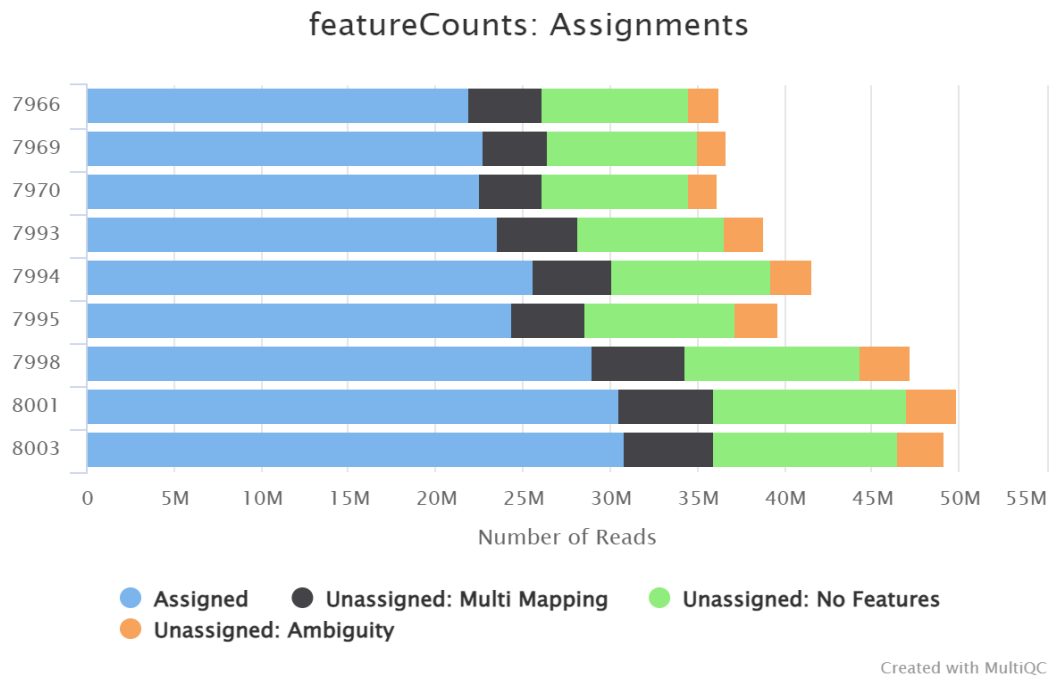

Figure 1. Statistics of featureCount output describing number of reads assigned to genomic features.

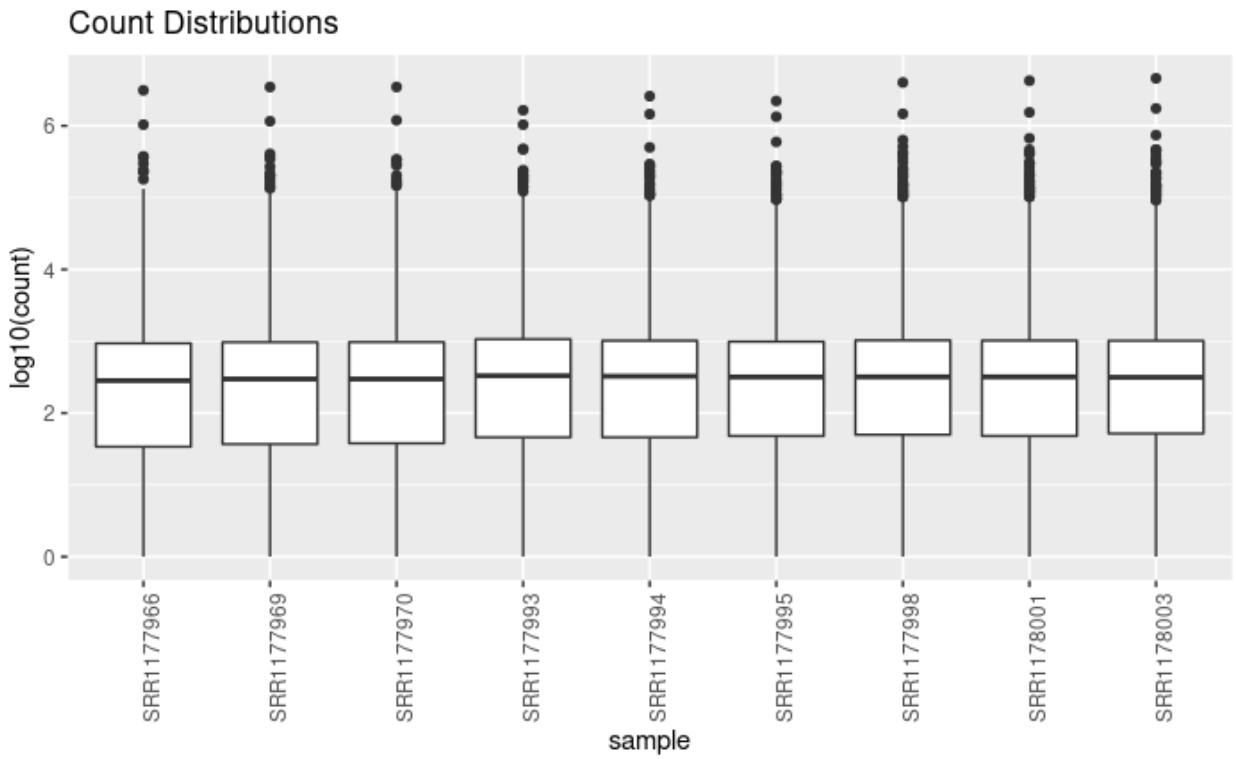

Figure 2. Distribution of number of counts across genes for each sample in the treatment group. Counts are visualized on log 10 scale.

Differential expression analysis of the processed counts matrix encompassing all nine treatment samples and their respective nine control samples revealed 212 significant DE genes in the NAPH group, 1666 in the ECON group, and 3199 in the THIO group. The top ten differentially expressed genes by adjusted p-value in each group are described below in Tables 1-3. Histograms of logFC values for each of the treatment groups are plotted below in Figures 3A-C. Volcano plots of logFC against the negative log of nominal p-values are also visualized in Figures 4A-C below.

gene_id	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
NM_012541	106141.1524	3.970764	0.2206642	17.994599	2.15E-72	2.43E-68
NM_130407	799.73658	3.809638	0.2503607	15.216599	2.74E-52	1.55E-48
NM_012540	3875.30571	8.767585	0.7580131	11.566534	6.09E-31	2.29E-27
NM_138502	1663.9076	-1.347575	0.1553038	-8.677025	4.06E-18	1.15E-14
NM_001109430	746.89572	-1.83455	0.2612099	-7.023278	2.17E-12	4.90E-09
NM_001191751	140.9109	-2.192855	0.3199919	-6.852847	7.24E-12	1.36E-08
NM_001191863	5465.28812	-0.914487	0.1345986	-6.794179	1.09E-11	1.76E-08
NM_033352	62.17474	-3.261619	0.4839877	-6.739053	1.59E-11	2.25E-08
NM_172019	675.51778	1.29619	0.1955572	6.628188	3.40E-11	4.27E-08
NM_175761	2431.23565	-1.355903	0.2115879	-6.408227	1.47E-10	1.66E-07

Table 1. Top ten differentially expressed genes sorted by adjusted p-value in beta-naphthoflavone treatment group.

gene_id	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
NM_001130558	2595.8445	-8.48595	0.30651	-27.68572	1.04E-168	1.17E-164
NM_144743	16917.6817	3.135977	0.1777583	17.6418	1.18E-69	6.65E-66
NM_001190380	925.8734	3.357025	0.2039397	16.46087	7.01E-61	2.64E-57
NM_001013904	59017.5126	2.359218	0.1466062	16.09221	2.89E-58	8.18E-55
NM_017272	12327.388	4.381226	0.2918414	15.01235	6.09E-51	1.18E-47
NM_012575	579.305	7.339207	0.488937	15.01054	6.26E-51	1.18E-47
NM_019184	38768.0729	-2.916012	0.195208	-14.93797	1.87E-50	3.01E-47
NM_013105	131661.5771	4.222425	0.2829365	14.92358	2.32E-50	3.27E-47
NM_001108565	566.1927	2.678593	0.1970049	13.59658	4.20E-42	5.27E-39
NM_013141	1008.5585	-2.791877	0.2114884	-13.20109	8.65E-40	9.78E-37

Table 2. Top ten differentially expressed genes sorted by adjusted p-value in econazole treatment group.

gene_id	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
NM_001008363	4031.7004	4.342363	0.1881206	23.08287	6.88E-118	8.67E-114
NM_031642	4256.4716	3.826319	0.2109337	18.13991	1.54E-73	9.72E-70
NM_001109260	557.1175	5.336941	0.2979705	17.91097	9.68E-72	4.07E-68
NM_001130573	1046.0876	6.56463	0.371435	17.6737	6.69E-70	2.11E-66
NM_031821	5188.3394	2.623246	0.1550237	16.92157	3.12E-64	6.55E-61
NM_012623	4692.1544	6.653292	0.3929473	16.93176	2.62E-64	6.55E-61
NM_001130500	409.607	4.56512	0.2818883	16.19478	5.49E-59	9.88E-56
NM_001108099	1783.6505	2.347949	0.146209	16.05885	4.96E-58	7.81E-55
NM_001039344	547.2088	3.788931	0.2416687	15.6782	2.13E-55	2.99E-52
NM_012923	3535.0123	2.77718	0.1773627	15.65819	2.92E-55	3.68E-52

Table 3. Top ten differentially expressed genes sorted by adjusted p-value in thioacetamide treatment group.

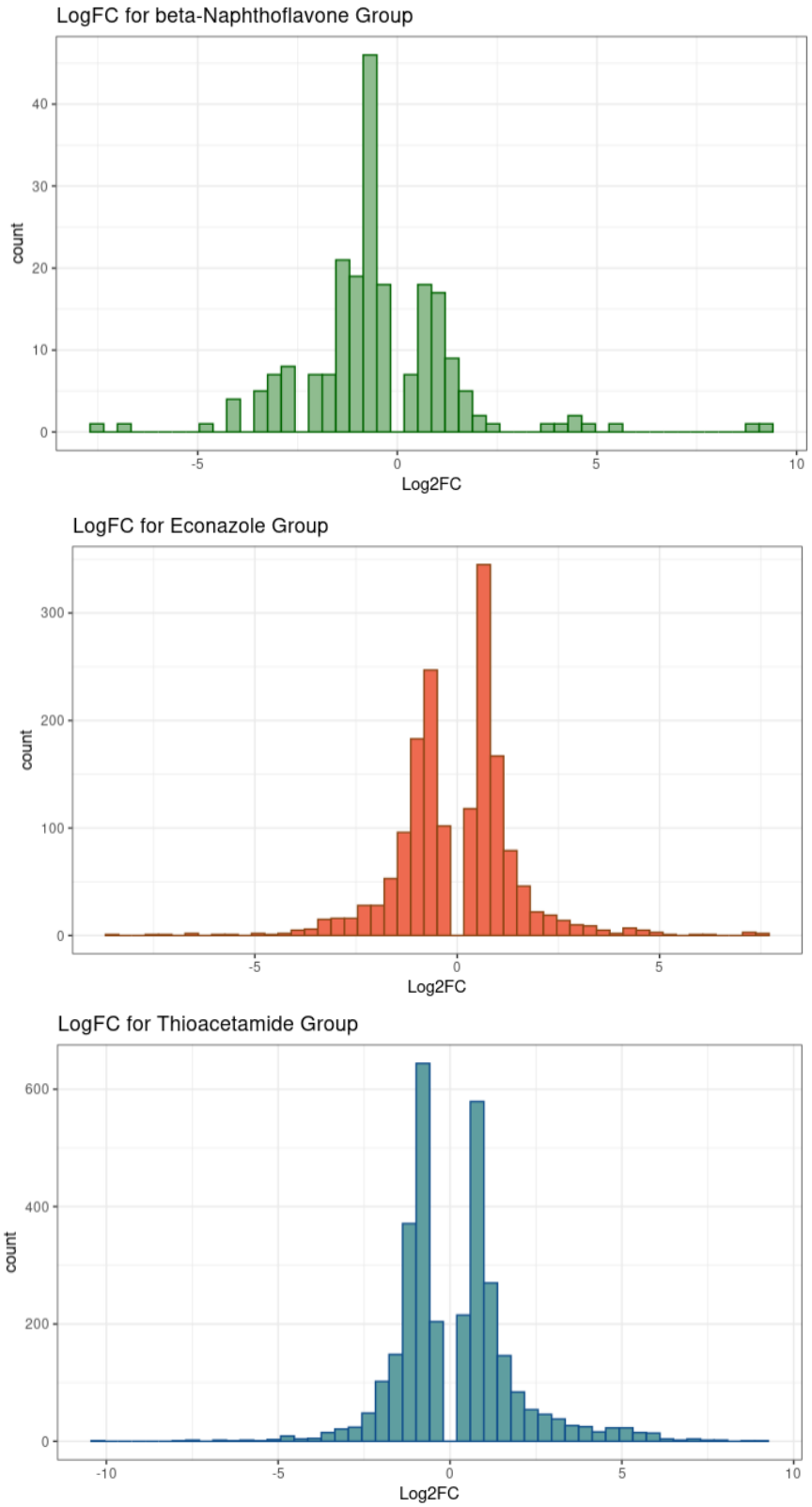

Figure 3. Histograms of logFoldChange values across significant differentially expressed genes for **(A)** beta-naphthoflavone, **(B)** econazole, and **(C)** thioacetamide groups.

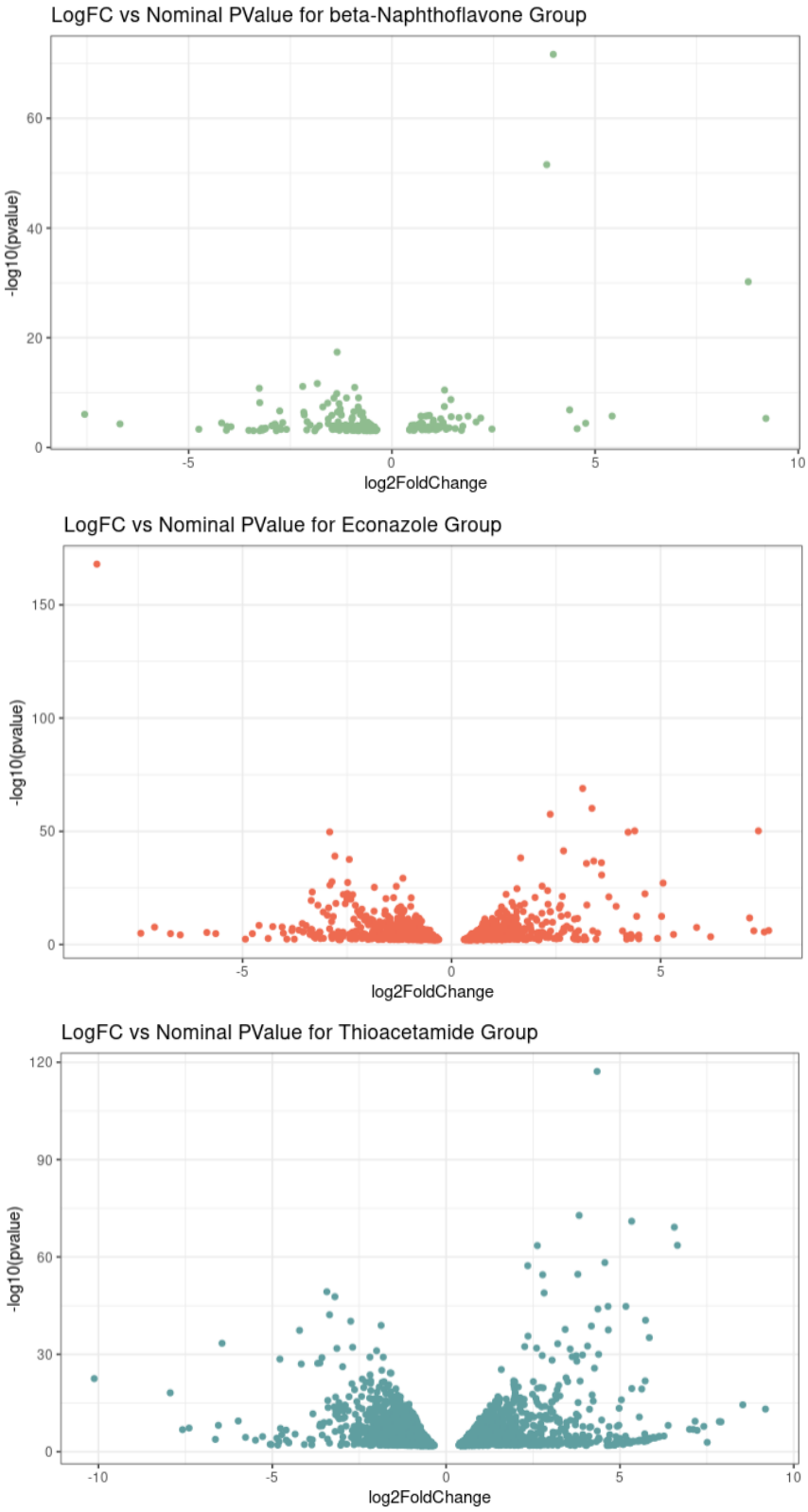

Figure 4. Volcano plots of logFoldChange values against nominal p-values for significant DE genes in **(A)** beta-naphthoflavone, **(B)** econazole, and **(C)** thioacetamide groups.

Discussion

The analyses performed here follow common differential expression analysis workflows using DESeq2. Reads were counted and mapped to gene features to generate a counts matrix. DE analysis of the counts matrix revealed 212 significant DE genes in the beta-naphthoflavone group, 1666 in the econazole group, and 3199 in the thioacetamide group, together which composed toxgroup 2 in the paper. Based on this expression data alone, it seems NAPH elicited the weakest response, with THIO eliciting the strongest response. The NAPH group also exhibited DE genes that skewed towards negative logFC expression, compared to the other two groups that exhibited a more even distribution of positive and negative logFC expression.

References

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

Yang Liao, Gordon K Smyth and Wei Shi. featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7):923-30, 2014.

Ewels, P., Magnusson, M., Lundin, S., Käller M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics (2016). https://doi.org/10.1093/bioinformatics/btw354