**BF528 Individual Project**

Daniel Golden

Project 3: Analyst role

**Introduction**

Wang et al.’s 2014 paper “A comprehensive study design reveals treatment- and transcript abundance–dependent concordance between RNA-seq and microarray data” reported the authors’ quantification of the concordance between microarray-based and RNA-seq-based measures of gene expression in the livers of rats exposed to 27 different chemical perturbants, as well as the authors’ observation of that concordance’s partial dependence on RNA transcript abundance and the type and mode of action of chemical perturbant. I have elected to complete the analyst role of the project associated with this paper, both out of intellectual curiosity and out of a desire to rectify a personal error that I made in the programmer role of that project several weeks ago. This role proved more daunting than anticipated, requiring an unusually long analysis with significant chunks of repetitive code, but it still held significant appeal, as it offered the chance to sharpen valuable skills, including the use of the *Limma* package for differential gene expression analysis [Ritchie et al., 2015]. The role of the analyst in this project was to (1) use *Limma* to conduct differential gene expression analysis on sample data associated with my group’s selected toxgroup relative to the appropriate controls using a provided RMA normalized expression matrix, (2) to calculate the concordance between the results of differential expression analysis based on *Limma* and the results of differential expression analysis based on *DESeq2* [Love et al., 2014], (3) to repeat the calculation of concordance separately for genes with above-median and below-median expression, and (4) to generate a variety of plots and tables representing the results of these methods. This role relates to the purpose of the paper in that it comprises the final stages of data analysis involved in completing the paper’s key objective: evaluating the concordance, or measure of agreement in identifying differentially expressed genes, between *Limma*, which is typically used for differential gene expression analysis with microarray data, and *DESeq2*, which is typically used for differential gene expression analysis with RNA-seq data. The deliverables of the analyst role for this project were, broadly speaking, the results for which Wang et al.’s paper was originally published.

**Methods**

An instructor-provided, pre-normalized RMA expression matrix with 31,099 genes and 2,218 samples was read into Rstudio and subset to include only the rows relevant to analysis for toxgroup 6. This resulted in the generation of three data frames: one containing the expression values for fluconazole treatment and control treatment in the same delivery vehicle as fluconazole, one containing the expression values for 3-methylcholanthrene treatment and control treatment in the same delivery vehicle as 3-methylcholanthrene, and one containing the expression values for pirinixic acid treatment and control treatment in the same delivery vehicle as pirinixic acid. The *Limma* package was used to conduct differential gene expression analysis comparing gene expression between each treatment and its associated control, per standard applications of *Limma* in the literature [Ritchie et al., 2015]. The results of this differential expression analysis were filtered to exclude those with nominal p-values greater than 0.05, but the absolute log2 fold change cutoff of 1.5 used in Wang et al.’s analysis was not applied, as its combination with the p-value cutoff resulted in an impractically small number of filtered results for the differential expression analysis of 3-methylcholanthrene treatment (only 5).

The filtered results data frames, representing 1,997 genes, 58 genes, and 8,761 genes respectively for the fluconazole, 3-methylcholanthrene, and pirinixic acid treatments, were written out as csv files arranged by adjusted p-value with the lowest values on top. From each of the filtered data frames, the top 10 most differentially expressed genes by nominal p-value were identified and separated into a new data frame (tables 1A, 1B, and 1C). Histograms of log2 fold change values and scatter plots of log2 fold change value versus nominal p-value for significant differentially expressed genes (DEGs) were also generated (figures 1 and 2).

The results of *DESeq2* differential gene expression analysis on the expression matrix for my samples and controls were read into Rstudio as three data frames. Corrected, pre-filtered versions of the *DESeq2* results CSV files that I generated in the programmer role of this project earlier in the semester were used, rather than instructor-provided results. As before, the top 10 most differentially expressed genes by nominal p-value for each treatment’s data frame were identified and separated into a new data frame (tables 2A 2B, and 2C). In order to calculate the concordance between *Limma* and *DESeq2* results for each treatment, Wang et al.’s approach was recreated.

The Affymetrix probe IDs denoting the genes in the data frame of *Limma* results were mapped to RefSeq identifiers, and the RefSeq identifiers denoting the genes in the data frame of *DESeq2* results were mapped to Affymetrix probe IDs. This was accomplished via the use of an instructor-provided mapping CSV file. As a result, it became possible to determine which genes were common to the two sets of significant DEGs generated by the two differential expression analysis methods. After this point, concordance could be calculated using the following trio of equations, the first to calculate concordance, the second to calculate the number of DEGs shared between methods, and the third to calculate background-adjusted intersection in order to compute a more accurate concordance value.

The values represented by these variables were applied with relative ease. N represented the total number of genes in the instructor-provided mapping CSV file, n0 represented the observed intersection between the two sets of DEGs, n1 represented the number of DEGs in one set, n2 represented the number of DEGs in the other set, and X represented the background-corrected intersection between the two. X was simplified to equal (n0 × N − n1 × n2)/(n0 + N − n1 − n2), per Wang et al.’s algebra.

Concordances between the two differential gene expression analysis methods were computed for all significant DEGs, and then separately for significant DEGs with log2 fold change values above the median and below the median. Two scatter plots akin to Wang et al.’s figure 2a were generated, showing overall concordance between methods versus the number of DEGs identified with each method (figures 3A and 3B). Finally, a bar plot was generated comparing the concordance between differential gene expression analysis methods for each treatment in toxgroup 6 with above-median, below-median, and overall DEGs (figure 4).

**Results**

As shown in tables 1A, 1B, and 1C, the DEGs identified by *Limma* results were represented by Affymetrix probe IDs specific to their transcripts. This was unsurprising, given that *Limma* is typically used for microarray data analysis. Noticeably, the tables representing the top 10 genes from the significant results from 3-methylcholanthrene treatment had the smallest average absolute value of log2 fold change, likely due to the smaller number of DEGs for that treatment.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Log2FC | AveExpr | t | p-value | adjusted p-value | B |
| 1368731\_at | 1.392496 | 13.416336 | 11.06255 | 7.62E-12 | 2.37E-07 | 16.48128 |
| 1377014\_at | -2.30625 | 4.7417069 | -9.6827 | 1.58E-10 | 2.45E-06 | 13.77057 |
| 1371076\_at | 2.420877 | 12.407155 | 9.333393 | 3.53E-10 | 3.28E-06 | 13.039 |
| 1390255\_at | 1.782758 | 7.2589869 | 9.213977 | 4.67E-10 | 3.28E-06 | 12.78462 |
| 1391570\_at | 1.641168 | 6.5054849 | 9.162804 | 5.27E-10 | 3.28E-06 | 12.67495 |
| 1394022\_at | -1.40379 | 9.7636904 | -8.83901 | 1.14E-09 | 5.89E-06 | 11.97168 |
| 1380336\_at | 1.327452 | 6.9153326 | 8.484594 | 2.68E-09 | 9.64E-06 | 11.18351 |
| 1372136\_at | -0.86663 | 8.9412926 | -8.47373 | 2.75E-09 | 9.64E-06 | 11.15905 |
| 1398597\_at | -1.30732 | 5.9650597 | -8.46819 | 2.79E-09 | 9.64E-06 | 11.14658 |
| 1377192\_a\_at | -1.18087 | 10.454968 | -8.26283 | 4.62E-09 | 1.34E-05 | 10.68063 |

**Table 1A.** The top ten most differentially expressed genes for fluconazole treatment by nominal p-value, as determined by *Limma* analysis.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Log2FC | AveExpr | t | p-value | adjusted p-value | B |
| 1387243\_at | 1.578477 | 13.355272 | 24.90666 | 2.54E-17 | 7.90E-13 | 20.04472 |
| 1370613\_s\_at | 0.784688 | 12.77594 | 14.55946 | 1.35E-12 | 2.09E-08 | 15.11115 |
| 1387759\_s\_at | 0.992617 | 11.88691 | 14.14747 | 2.36E-12 | 2.45E-08 | 14.78118 |
| 1383325\_at | 0.473877 | 5.9909993 | 9.837972 | 2.08E-09 | 1.62E-05 | 10.23091 |
| 1387901\_at | -0.46683 | 8.5132321 | -6.98844 | 5.94E-07 | 0.00369187 | 5.77417 |
| 1372297\_at | 0.420597 | 11.195 | 6.835296 | 8.29E-07 | 0.00429717 | 5.495955 |
| 1384544\_at | 0.345375 | 11.854305 | 6.763178 | 9.71E-07 | 0.00431553 | 5.363576 |
| 1368168\_at | -1.2359 | 8.5686023 | -6.46192 | 1.90E-06 | 0.00737186 | 4.801246 |
| 1380888\_at | 0.538434 | 5.9829003 | 6.316619 | 2.63E-06 | 0.00908494 | 4.524751 |
| 1367669\_a\_at | 0.381983 | 9.6810073 | 6.236323 | 3.15E-06 | 0.00980539 | 4.370507 |

**Table 1B.** The top ten most differentially expressed genes for 3-methylcholanthrene treatment by nominal p-value, as determined by *Limma* analysis.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Log2FC | AveExpr | t | p-value | adjusted p-value | B |
| 1398250\_at | 9.542023 | 8.3242267 | 146.6296 | 2.17E-32 | 6.74E-28 | 52.73941 |
| 1388211\_s\_at | 7.116135 | 9.8270342 | 72.37328 | 3.94E-26 | 6.12E-22 | 46.37621 |
| 1391433\_at | 3.870568 | 9.9352763 | 69.31753 | 9.48E-26 | 9.83E-22 | 45.82093 |
| 1387740\_at | 4.476151 | 8.0719471 | 63.19284 | 6.24E-25 | 4.85E-21 | 44.56811 |
| 1375845\_at | 4.843399 | 9.3107213 | 58.73205 | 2.77E-24 | 1.72E-20 | 43.51918 |
| 1374187\_at | 4.200635 | 5.3743097 | 55.8921 | 7.58E-24 | 3.93E-20 | 42.78166 |
| 1386885\_at | 3.607191 | 11.761231 | 55.08521 | 1.02E-23 | 4.52E-20 | 42.56124 |
| 1389253\_at | 5.163534 | 9.583448 | 51.82174 | 3.52E-23 | 1.37E-19 | 41.61636 |
| 1384244\_at | 3.221774 | 10.787974 | 49.49834 | 8.93E-23 | 3.09E-19 | 40.88711 |
| 1367680\_at | 2.066581 | 13.02148 | 45.27228 | 5.46E-22 | 1.70E-18 | 39.4237 |

**Table 1C.** The top ten most differentially expressed genes for pirinixic acid treatment by nominal p-value, as determined by *Limma* analysis.

While it would have been preferable to implement the absolute log2 fold change cutoff of 1.5 in order to focus more exclusively on genes with larger differences in expression between treatment and control, the absence of this cutoff did not skew results to exclude relevant genes; it simply resulted in the inclusion of genes with low but still significant differences in expression. As shown below in figure 1, there were no filtered results that had absolute log2 fold change values of zero, although the majority had values between 0.5 and 1.0. Notably, due to the smaller total number of significant DEGs in 3-methylcholanthrene treatment and the resultantly larger effect of stochasticity, the histogram of log2 fold change values for that treatment was less symmetrically distributed than that of other treatments.

Chart, histogram

Description automatically generated

**Figure 1A.** Histogram of significant results for *Limma* analysis of fluconazole treatment.

Chart, histogram

Description automatically generated

**Figure 1B.** Histogram of significant results for *Limma* analysis of 3-methylcholanthrene treatment.

Histogram

Description automatically generated with low confidence

**Table 1C.** Histogram of significant results for *Limma* analysis of pirinixic acid treatment.

The results shown in figure 1 were recapitulated by the generation of scatter plots representing log2 fold change values for the DEGs for each treatment versus the nominal p-values for those DEGs, as shown below.

Chart, scatter chart

Description automatically generated

**Figure 2A.** Scatter plot of significant results for *Limma* analysis of fluconazole treatment.

Chart

Description automatically generated

**Figure 2B.** Scatter plot of significant results for *Limma* analysis of 3-methylcholanthrene treatment.

Chart, histogram

Description automatically generated

**Figure 2C.** Scatter plot of significant results for *Limma* analysis of pirinixic acid treatment.

As shown in tables 2A, 2B, and 2C, the top 10 most differentially expressed genes as identified by *DESeq2* analysis were represented by RefSeq identifiers, rather than the Affymetrix probe IDs used in *Limma* results. Notably, log2 fold change values seem

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | baseMean | log2FoldChange | lfcSE | p-value | adjusted p-value |
| NM\_001108693 | 573.5890483 | 5.17685378 | 0.214808 | 8.02E-130 | 9.30E-126 |
| NM\_053699 | 426.5724459 | 6.606094471 | 0.28389 | 2.00E-120 | 1.16E-116 |
| NM\_001130558 | 2377.22854 | -7.88653594 | 0.393327 | 9.76E-91 | 3.77E-87 |
| NM\_031605 | 1076.962744 | 3.820959809 | 0.221821 | 1.33E-67 | 3.86E-64 |
| NM\_013033 | 1000.389495 | 6.002591717 | 0.36174 | 6.16E-63 | 1.43E-59 |
| NM\_144755 | 2060.777948 | 4.101083227 | 0.275333 | 2.11E-51 | 4.07E-48 |
| NM\_001005384 | 1522.844313 | 3.912838821 | 0.270924 | 1.70E-48 | 2.81E-45 |
| NM\_031048 | 6984.513451 | 4.220337156 | 0.306013 | 1.80E-44 | 2.60E-41 |
| NM\_013105 | 300199.8119 | 4.900401787 | 0.356258 | 1.82E-43 | 2.35E-40 |
| NM\_001014166 | 1614.185995 | -2.908111508 | 0.219556 | 2.61E-41 | 3.03E-38 |

**Table 2A.** The top ten most differentially expressed genes for fluconazole treatment by nominal p-value, as determined by *DESeq2* analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | baseMean | log2FoldChange | lfcSE | p-value | adjusted p-value |
| NM\_012541 | 80326.80848 | 3.689871988 | 0.182507 | 6.17E-92 | 6.43E-88 |
| NM\_130407 | 546.3575796 | 3.352482997 | 0.218852 | 3.89E-54 | 2.02E-50 |
| NM\_022521 | 13003.77943 | 1.754133885 | 0.247668 | 7.83E-14 | 2.72E-10 |
| NM\_012608 | 64.24548234 | -2.871329606 | 0.419767 | 3.85E-13 | 1.00E-09 |
| NM\_053883 | 1310.725154 | -1.161600999 | 0.171182 | 5.97E-13 | 1.24E-09 |
| NM\_017061 | 127.0197977 | -2.187135316 | 0.330759 | 1.80E-12 | 3.13E-09 |
| NM\_134329 | 285.7950497 | -1.502129818 | 0.231136 | 3.88E-12 | 5.77E-09 |
| NM\_022866 | 1108.166598 | -2.404996022 | 0.381378 | 1.23E-11 | 1.60E-08 |
| NM\_022297 | 2888.890563 | -0.921196332 | 0.146194 | 1.65E-11 | 1.79E-08 |
| NM\_022635 | 368.0667426 | -1.469284758 | 0.234314 | 1.72E-11 | 1.79E-08 |

**Table 2B.** The top ten most differentially expressed genes for 3-methylcholanthrene treatment by nominal p-value, as determined by *DESeq2* analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | baseMean | log2FoldChange | lfcSE | p-value | adjusted p-value |
| NM\_024162 | 2298.61223 | 7.83672221 | 0.26235 | 5.62E-196 | 6.31E-192 |
| NM\_012737 | 2516.803823 | -6.722284601 | 0.242624 | 6.56E-170 | 3.68E-166 |
| NM\_131903 | 9516.988733 | -5.13410172 | 0.251588 | 8.63E-94 | 3.23E-90 |
| NM\_019157 | 439.4460982 | 5.989680332 | 0.292848 | 2.37E-93 | 6.64E-90 |
| NM\_017158 | 2723.415934 | -6.163576764 | 0.30728 | 1.10E-90 | 2.46E-87 |
| NM\_053883 | 994.2520835 | -3.059832378 | 0.15817 | 1.35E-84 | 2.53E-81 |
| NM\_001014063 | 2844.494836 | -4.181611487 | 0.226329 | 1.78E-77 | 2.85E-74 |
| NM\_012600 | 13167.95193 | 3.837349632 | 0.207484 | 4.05E-77 | 5.68E-74 |
| NM\_001013098 | 2748.692917 | -5.39288389 | 0.301901 | 1.15E-72 | 1.43E-69 |
| NM\_001013975 | 954.7396869 | -3.835691638 | 0.217722 | 1.08E-70 | 1.21E-67 |

**Table 2C.** The top ten most differentially expressed genes for pirinixic acid treatment by nominal p-value, as determined by *DESeq2* analysis.

Concordance between results for *Limma* and *DESeq2* was calculated, resulting in concordance values of 0.70 for fluconazole treatment, 0.14 for 3-methylcholanthrene treatment, and 0.82 for pirinixic acid. These values were plotted against the number of DEGs for each method of differential expression analysis, as shown in figures 3 below. Concordance values were also calculated between *Limma* and *DESeq2* results with above-median and below-median log2 fold change values, as shown in figure 4 below. Notably, concordance was far lower for below-median log2 fold change values than for above-median values. For fluconazole treatment, the above-median and below-median concordances were 0.65 and 0.08. For 3-methylcholanthrene treatment, the above-median and below-median concordances were 0.14 and 0.04. Lastly, for pirinixic acid treatment, the above-median and below-median concordances were 0.72 and 0.11. Most likely, the lower concordances for below-median results were at least partly a result of my decision to refrain from using the absolute log2 fold change cutoff of 1.5, since results with lower absolute log2 fold change values were less likely to be identified as significant by both *Limma* and *DESeq2*. Still, even with an absolute log2 fold change cutoff, concordance was higher for results with high log2 fold change values, as those were more likely to be detected by both analysis methods.

Chart

Description automatically generated

**Figure 3A.** Scatter plot of concordance versus number of DEGs resulting from *DESeq2* analysis.

Chart, scatter chart

Description automatically generated

**Figure 3B.** Scatter plot of concordance versus number of DEGs resulting from *Limma* analysis.

Chart, bar chart

Description automatically generated

**Figure 4.** Bar plot demonstrating the relationship between *Limma*-*DESeq2* concordance and the absolute value of log2 fold change.

**Discussion**

The goals of the analyst in this project were, in summary, to conduct differential gene expression analysis on an instructor-provided, pre-normalized RMA expression matrix using the *Limma* package, to identify significant DEGs from that analysis, and to evaluate the concordance between the set of *Limma*-identified, significant DEGs and the programmer’s set of *DESeq2*-identified significant DEGs. These goals were met successfully. The results of *Limma* analysis showed that, of the 31,099 genes with expression data in the RMA expression matrix, 1,997 genes, 58 genes, and 8,761 genes were identified as DEGs for treatments with fluconazole, 3-methylcholanthrene, and pirinixic acid, respectively. Concordance was found to be strong (greater than 0.65) for overall and above-median results of fluconazole and pirinixic acid treatments, but concordance was weak (below 0.15) for 3-methylcholanthrene treatment as well as for all below-median results. This suggests that concordance depends both on the degree of differential gene expression in significant DEGs and the sample size of DEGs to be analyzed.

References:

1. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, <https://doi.org/10.21105/joss.01686>
2. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.8. https://CRAN.R-project.org/package=dplyr
3. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
4. Love, M.I., Huber, W., Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 Genome Biology 15(12):550 (2014)
5. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7), e47.
6. 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