Microarray Based Tumor Classification

Programmer: & Biologist: Manas Dhanuka

Introduction

Colorectal cancer (CRC) or commonly known as Colon cancer (CC) is an extremely prevalent cancer type, as well as extremely fatal [2]. Pathological staging or biopsies were the only clinical method used to treat people with CRC, these though were not capable of predicting reoccurrence accurately, as they did not account for the heterogeneous nature of the disease.

In 2013, Marisa et al. successfully predicted clinical prognostic factors of colon cancer (CC) using mRNA gene expression profiles. This approach had been taken before too, using microarray technology to analyze gene expression profiles, albeit the results were not promising enough to warrant use in the prognosis of CRC. Marisa et al, to comprehensively classify samples, used genome-wide mRNA analysis. Specifically, they used an array-based comparative genomic hybridization. Additionally, to find common clusters in the data, consensus unsupervised analysis of gene expression profiles revealed six molecular subtypes. Through the classification process, they formed significant associations between these molecular subtypes and clinicopathological factors using the Chi-squared test, logistic regression, and the Kyoto Encyclopedia of Genes and Genomes to observe associated signaling pathways. Lastly, they tested the strength and validity of their classified molecular subtypes in a large independent dataset. Compared to previous studies, they were able to reveal a comprehensive classification process for CC molecular subtypes that can be used in clinical practices to predict prognosis and recurrence more accurately.

Marisa et al. used 750 tumor samples from patients with stage I to IV colon cancer from the French CIT program, of which 566 tumor samples satisfied the RNA quality control measurements for gene expression profiling (GEP) analysis. These samples were further separated into a discovery set which comprised 443 samples and a validation set which comprised the remaining 123 samples. In this analysis project, there were a total of 134 samples utilized from the data described above . These included data from the combined discovery and validation datasets. The data for these samples can be accessed at [http://www.ncbi.nlm.nih.gov/geo /; accession number GSE39582].

This report is an attempt to replicate parts of the process followed by Marisa et. al. By performing the data processing and explaining the biological interpretation of the analysis performed earlier in the semester.

Methods

Data Preprocessing and Quality Control

After retrieving array-based data for this report from Gene Expression Omnibus in the form of files generated by Affymetrix microarray image analysis software (CEL files), all data wrangling and statistical analysis were performed on R and libraries from BiocManager. The preprocessing started by first reading the CEL files into an AffyBatch object and subsequently converting them into an ExpressionSet object. The RMA (robust multiarray average) algorithm normalizes CEL files together through background correction, log2 transformation, and quantile normalization.

As a quality control step, the AffyBatch object was used to generate a PLMset (This is a class representation for Probe level Linear Models fitted to Affymetrix GeneChip probe level data.) object that underwent relative log expression (RLE) and normalized unscaled standard error (NUSE) calculation [3]. The medians of each of these calculations were plotted in Figures 1 and 2 to quickly visualize variation between samples. An RLE plot is constructed by calculating a single gene's

median expression across all samples, then calculating the deviations from this median. The use of medians helps mitigate the effect of outliers. The underlying assumption in RLE calculations is that expression levels of most genes in the dataset are not affected by whichever biological feature is being studied. Therefore, with an ideal or close to an ideal dataset where there is no undesirable variation between samples, the calculated median expression values would all lie around zero [4]. Similarly, NUSE calculation also assumes that consistent expression should be measured across most probes. NUSE values represent the relative precision of the estimated expression values. In a good quality expression array, the median NUSE value should be one. [5].

Following these quality control steps; the expression data were extracted from the original ExpressionSet object. Clinical and batching annotation provided by Marisa et al. was loaded into the R script to run *ComBat* to correct for batch effects while retaining features of interest. The ComBat method uses a Bayes framework that is robust for outliers in smaller datasets and comparable to other batch correction methods in larger datasets [6].

The normalized and batch corrected data were then scaled and centered to a mean of zero and a standard deviation of one. Principal component analysis was performed on the normalized and batch corrected data and PC1 and PC2 were scatter plotted to examine the presence of outliers.

In-depth Analysis

Welch's t-test was performed on the probe sets passing the noise filtering and dimensionality reduction filters which is part of the analysis section. These were then mapped to gene symbols using the Bioconductor package, hgu133plus2.db. Adding the mapped symbols to the probe set IDs resulted in "NA" values and duplicated mapped probeset IDs to gene symbols, which were both removed before continuing the analysis. The researchers in Marisa et al. chose the probeset with the greatest variance out of the ones mapped to a single gene symbol, which is what was also implemented in this analysis by filtering results by the adjusted p-value and grouping the gene symbols.

Using the differential expression results and Chi-squared filtered results, the top 1,000 up-and down-regulated genes were selected based on the t-statistic. Additionally, the top 10 up-and down-regulated genes were also sliced out.

To compare overlaps between the gene sets and top up and down-regulated genes. The KEGG, GO, and Hallmark gene sets were downloaded from MSigDB.The KEGG gene set collection contains canonical pathways derived from the Kyoto Encyclopedia of Genes and Genomes. The GO gene set contains the gene ontology terms for biological processes, cellular components, and molecular functions. Lastly, the hallmark gene set represents well-defined biological states and processes.

Fisher t-test was then used to compute hypergeometric statistics and p-values comparing overlap for each gene set and each of the top 1000 increased and 1000 decreased genes. The results from the differential expression matrix with removed duplicated probeset IDs were used to build contingency tables, which were needed to implement Fisher's Exact Test. The contingency table for this analysis requires four numbers: number of differentially expressed genes in the gene set, number of differentially expressed genes not in the gene set, number of not differentially expressed genes in the gene set, and number of not differentially expressed gene not in the gene set.

The Fisher's Exact Test was performed on the KEGG, GO, and hallmark contingency tables using a pre-defined function in R. The yielded results included statistic estimate and p-value for each gene set used. Using the results from the Fisher's Test for each gene set (KEGG, GO, and Hallmark), the p-values were adjusted using the Benjamin-Hochberg (FDR) procedure. To find the statistically enriched gene sets, the nominal p-values were filtered using a 5% significance level. The top three gene sets from each gene set type were sliced out. These results were then compared to the results found in Marisa et al. (2013).

Results

Based on relative log expression calculation on the Marisa et al. dataset, RLE medians were generally centered around 0, with a few outliers hovering around -0.1 and 0.1. Two outliers were observed that had RLE values significantly higher than 0.1 (Figure 1). Similar results were observed using normalized unscaled standard error. Most NUSE medians were around or slightly less than 1.0, with two outliers that were significantly higher at 1.050 (Figure 2). Thigh quality and ideal datasets would have RLE medians at 0.0 and NUSE medians at 1.0, so ours is not too far away from a good dataset.

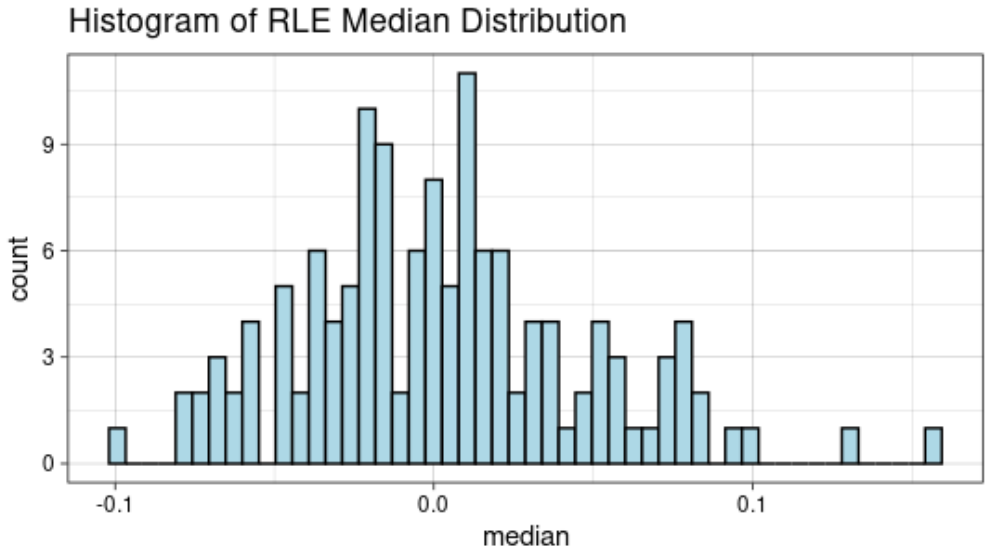

Figure 1: Distribution of median relative log expression (RLE) values across 134 samples in the dataset.

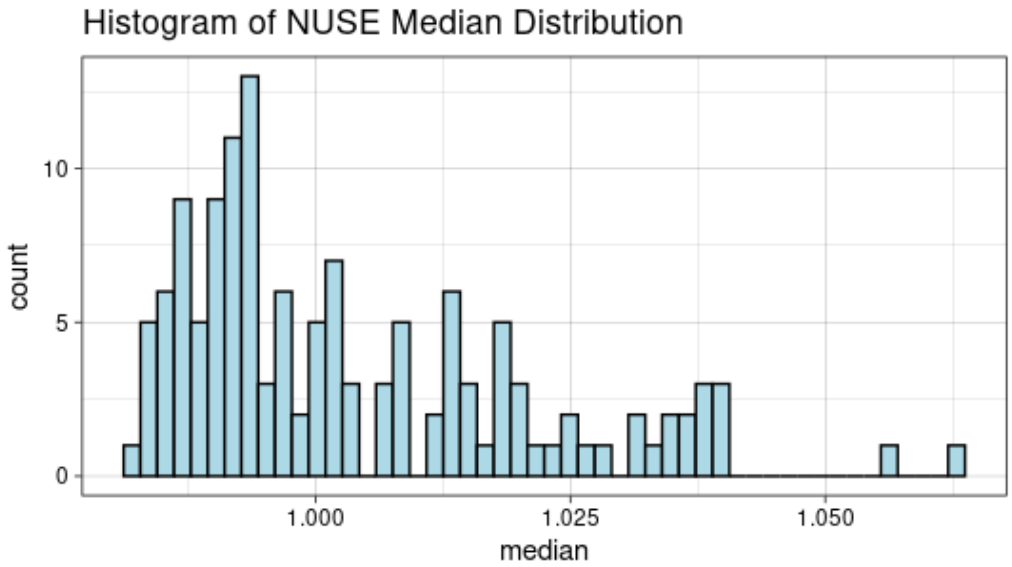

Figure 2: Distribution of median normalized unscaled standard error (NUSE) values across samples.

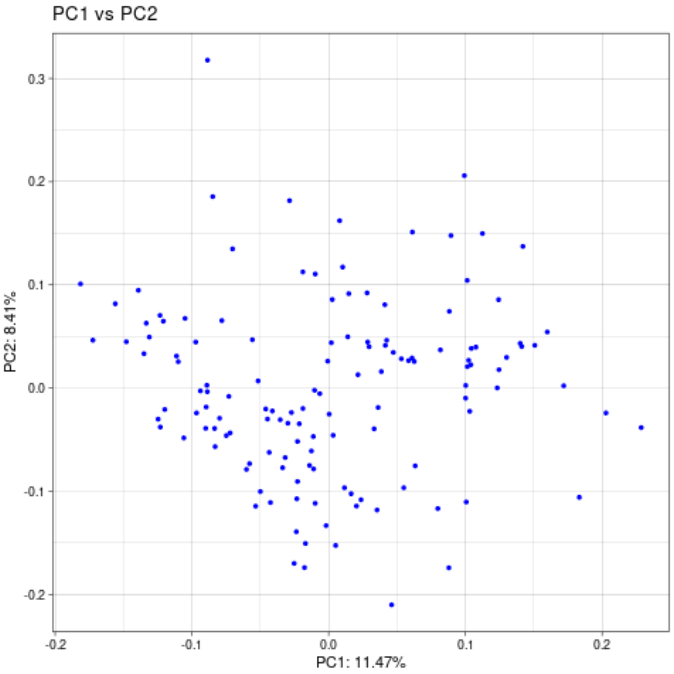

Figure 3: Scatterplot of samples with axes as principal components PC1 and PC2, capturing the highest degrees of variability.

Figure 3 plots principal components (PC) 1 and 2, which attempt to explain the highest percent variability in the dataset. Respectively, they make up 11.47% and 8.41% of variability, so there is still about 80% of variability that is not captured in the plot. Outliers are observed on the plot although the majority of samples cluster towards the center.

SYMBOL	probeids	t statistic	p-value	p-adjusted
Upregulated Genes				
SFRP2	223121_s_at	21.66559687	1.60E-42	4.89E-39
ARMCX1	218694_at	21.17798024	1.01E-42	4.14E-39
SPOCK1	202363_at	20.92993163	4.29E-43	2.12E-39
GSDME	203695_s_at	20.58033378	6.02E-38	3.72E-35
FNDC1	226930_at	20.34076416	3.76E-42	9.28E-39
NXN	219489_s_at	20.2042052	1.43E-40	1.93E-37
EFEMP2	206580_s_at	20.00689035	2.05E-38	1.41E-35
SERPING1	200986_at	19.89713331	1.29E-40	1.82E-37
GLI3	227376_at	19.53477224	9.82E-37	3.78E-34
ZFPM2	219778_at	19.46432965	2.29E-34	4.93E-32
Downregulated Genes				
FCGBP	203240_at	-13.9024065	1.79E-25	9.79E-24
LRRC31	220622_at	-13.7613625	5.92E-27	3.84E-25
ST6GALNAC1	227725_at	-13.309567	1.21E-22	4.66E-21
CRYM	205489_at	-12.2074948	4.15E-23	1.71E-21
C4orf19	219450_at	-12.1752437	1.30E-20	3.95E-19
PRELID2	236513_at	-11.9872344	3.20E-22	1.17E-20
GSKIP	223239_at	-11.9819115	2.67E-21	8.82E-20
NR3C2	205259_at	-11.8619453	1.73E-20	5.19E-19
MRAP2	227226_at	-11.8530612	8.05E-22	2.82E-20
NXPE1	1561387_a_at	-11.8521222	7.04E-22	2.49E-20

Table 1: Top and Bottom 10 Up- and Down-Regulated Genes

The t-test results of the differential expression matrix consisted of 29,645 probeset IDs and their corresponding calculated t-test statistic, p-value, and adjusted p-value. After mapping the probeset IDs to gene symbols from the hgu133plus2.db, there were a total of 37,147 mapped probeset IDs to gene symbols. Subsequently, after removing all the “NA” values from the matrix, a total of 27125 probes were left. Like the findings in the paper, multiple probe sets mapped to the same gene symbol. To correct this, the probes with the greatest variance were removed, consistent with what was done in the paper.

After selecting the most variant probes, a total of 15,804 unique probes to gene symbol matches were obtained. Specifically, a total of 8,080 differentially expressed genes and 7724 not differentially expressed genes. Both of which were statistically significant based on the adjusted p-value. Using the uniquely mapped probe-gene symbols, the top 10 up- and down-regulated genes were selected by using the t-statistic as the filtering parameter (Table 1). The top up-regulated and down-regulated genes were consistent with the most positive and negative t-statistic, respectively.

gene_set	statistic_estimate	p_value	pvalue_adjusted
KEGG Gene Sets			
KEGG_METABOLISM_OF_XENOBIOTICS_BY_CYTOCHROME_P450	2.946533561	0.000262892	0.01262827
KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION	3.71417729	0.000328159	0.01262827
KEGG_UBIQUITIN_MEDIATED_PROTEOLYSIS	0.488931687	0.00033947	0.01262827
GO Gene Sets			
GO_ORGANIC_ACID_CATABOLIC_PROCESS	1.71489283	0.000100328	0.013345401
HP_DELAYED_GROSS_MOTOR_DEVELOPMENT	0.510945017	0.000101687	0.013405492
GO_MITOCHONDRIAL_PROTEIN_COMPLEX	1.888919149	0.000103412	0.01345373
Hallmark Gene Sets			
HALLMARK_ADIPOGENESIS	1.888043988	0.000123539	0.001853769
HALLMARK_XENOBIOTIC_METABOLISM	1.853476059	0.000179993	0.001853769
HALLMARK_OXIDATIVE_PHOSPHORYLATION	1.943222864	0.000185377	0.001853769

Table 2: Top Three Significantly Enriched Gene Sets for each Set Type. KEGG, GO, and Hallmark. The gene set in dark green is one that corroborated with Marisa et al.

Next gene sets were downloaded from MSigDB for the next part of the analysis. The KEGG, GO, and hallmark gene set databases contained 186, 14,765, and 50 gene sets, respectively. Contingency tables were generated for each gene set within each gene set type and underwent Fisher's Exact Test. From these results, the total number of significantly enriched gene sets within each gene set type was found. The total number of significantly enriched gene sets in KEGG, GO, and Hallmark gene set types were 38, 1305, and 15, respectively. This added up to 1358 significantly enriched gene sets. From each gene set type of significantly enriched gene sets, the top three statistically significant (<0.05) enriched gene sets were selected (Table 2).

Discussion

For RLE, ideally, it is desired that the median RLE value be close to 0 and for NUSE, the standard error from the probe level should be close to 1 across the samples. By the result shown above the median value for RLE value around 0 and the median value of NUSE around 1, which indicates a good gene expression of the microarray data. With the Principal component analysis, the first two principal components PC1 and PC2 contributed about 20% variance, i.e., there is still 80% variance described by the other PCs.

In the paper, researchers found 57 probeset IDs corresponding to unique gene symbols compared to the 15,804 in this analysis. Just looking at the top 10 up-and down-regulated genes, none and two down-regulated are consistent with those found in the papers 57, respectively. The genes are FCGBP and C4orf9. A possible explanation for this discrepancy is that the dataset used here is merged; however, in the paper, researchers split their data into discovery and validation sets. Also, one of the top three enriched gene sets in our data matches what was found in the paper; however, looking at all our KEGG and GO enriched gene sets, there are multiple overlaps. This discrepancy could be due to the different statistical methods used in the paper or the merged dataset used here for this project.

In Summary, through this analysis, matches to biological pathways and processes to genes in the microarray data extracted from tumor samples was found. The functional enrichment analysis performed was somewhat consistent with the findings in the paper and overlap with some of the biological pathways and processes found was observed.

References

1. Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M. P., Vescovo, L., Etienne-Grimaldi, M. C., Schiappa, R., Guenot, D., Ayadi, M., Kirzin, S., Chazal, M., Fléjou, J. F., Benchimol, D., Berger, A., Lagarde, A., Pencreach, E., Piard, F., Elias, D., Parc, Y., … Boige, V. (2013). Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS medicine*, *10*(5), e1001453. https://doi.org/10.1371/journal.pmed.1001453
2. Greenlee, R. T., Murray, T., Bolden, S., & Wingo, P. A. (2000). Cancer statistics, 2000. *CA: a cancer journal for clinicians , 50* (1), 7–33. https://doi.org/10.3322/canjclin.50.1.7
3. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., Speed, T. P. (2003). Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Accepted for publication in Biostatistics.
4. Gandolfo, L. C., & Speed, T. P. (2018). RLE plots: Visualizing unwanted variation in high dimensional data. PloS one, 13(2), e0191629. https://doi.org/10.1371/journal.pone.0191629
5. Tang, H., & Therneau, T. M. (2010). Statistical metrics for quality assessment of high-density tiling array data. Biometrics, 66(2), 630–635. https://doi.org/10.1111/j.1541-0420.2009.01298.x
6. W. Evan Johnson, Cheng Li, Ariel Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods, Biostatistics, Volume 8, Issue 1, January 2007, Pages 118–127, https://doi.org/10.1093/biostatistics/kxj037
7. Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. 2004. affy---analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20, 3 (Feb. 2004), 307-315.
8. Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. 2004. affy---analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20, 3 (Feb. 2004), 307-315.
9. Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Elana J. Fertig, Andrew E. Jaffe, Yuqing Zhang, John D. Storey and Leonardo Collado Torres (2021). sva: Surrogate Variable Analysis. R package version 3.40.0.
10. Hervé Pagès, Marc Carlson, Seth Falcon and Nianhua Li (2021). AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. R package version 1.56.2. https://bioconductor.org/packages/AnnotationDbi
11. Marc Carlson (2021). hgu133plus2.db: Affymetrix Affymetrix HG-U133_Plus_2 Array annotation data (chip hgu133plus2). R package version 3.13.0.
12. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
13. uan Tang, Masaaki Horikoshi, and Wenxuan Li. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal 8.2 (2016): 478-489.
14. Masaaki Horikoshi and Yuan Tang (2016). ggfortify: Data Visualization Tools for Statistical Analysis Results. https://CRAN.R-project.org/package=ggfortify
15. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.9. https://CRAN.R-project.org/package=dplyr
16. Kirill Müller and Hadley Wickham (2021). tibble: Simple Data Frames. R package version 3.1.6. https://CRAN.R-project.org/package=tibble
17. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
18. Martin Morgan, Seth Falcon and Robert Gentleman (2021). GSEABase: Gene set enrichment data structures and methods. R package version 1.54.0.