Project 5 Report

Microarray Based Tumor Classification

Submitted by Urvy Mudgal

Introduction

Marisa et al. used an extended characterization series of colon cancer samples to build a molecular screening based on a genome-wide mRNA expression analysis to identify predictive biomarkers in their article. Their further assessments included associations between molecular subtypes and clinicopathological factors, common DNA alterations, and prognosis. This report will contain the parts that include Data preprocessing and Quality Control as well as the analysis of the results that start with Noise filtering and dimensionality reduction, and lead to Hierarchical clustering and subtype discovery. This report uses data as previously curated sample files for this project.

Data preprocessing and Quality Control

To guarantee that variations in intensities read by the scanner are attributable to differential gene expression rather than printing, hybridization, or scanning errors, microarray data must be normalized. There are several approaches for combining microarray samples. Quantile normalization, log-median centering, and the inclusion of control genes are just a few examples. Many scholars, fortunately, spend their work to making jobs like this as simple and basic as feasible. I helped write a R script that used the Robust Multiarray Averaging (RMA) technique to normalize all of the microarray data together. Then, using Principal Component Analysis, I generated standard quality control measures on the normalized data and visualized the sample distribution (PCA).

Noise filtering and Dimensionality Reduction

Clustering is a strong analytical approach that may be used with enormous sample sizes. Clustering is an unsupervised approach for grouping similar objects based on some criterion, often a collection of characteristics whose similarity is quantified by a distance function. Marisa et al. used a method called Consensus Clustering to determine the real number of clusters in their data for their research. Cluster analysis is effective because it does not require a class label like the disease status, allowing for the discovery of new associations. For the purposes of this course, we have utilized hierarchical clustering instead of Consensus clustering approach for computational feasibility.

Methods

All data wrangling and statistical analysis for this research was done on the BU Shared Computing Cluster using R 4.1.2 and BiocManager 3.14 after collecting array-based data from Gene Expression Omnibus in the form of files created by Affymetrix microarray image analysis software (CEL files)- which were readily available in the project directory

ReadAffy was used to read CEL files into an AffyBatch object, which was then transformed to an ExpressionSet object using the rma method. Background correction, log2 transformation, and quantile normalization are used in the RMA (robust multichip average) technique to normalize CEL data.

The AffyBatch object was also processed through fitPLM as a quality control step, resulting in a PLMset object that was subjected to relative log expression (RLE) and normalized unscaled standard error (NUSE) calculations. To summarize and readily illustrate variance between samples, the medians of each of these computations were shown in Figures 1A and 1B. An RLE plot is made by finding the median expression of a particular gene across all samples and then calculating the deviations from that median. The use of medians helps mitigate the effect of outliers. RLE calculations are based on the premise that the expression levels of the vast majority of genes in the dataset are unaffected by the biological trait under investigation.[1] As a result, in a perfect or near-ideal dataset with no unwanted variance across samples, the estimated median expression values would all be close to zero [2]. Similarly, the NUSE calculation requires that consistent expression across most probes should be assessed. The NUSE values indicate the predicted expression values' relative accuracy. The median NUSE value in a high-quality expression array should be one. [3].

The exprs method was used to extract expression data from the original ExpressionSet object after these quality control processes. Marisa et al. supplied clinical and batching annotation, which was imported into the R script to run ComBat to compensate for batch effects while keeping characteristics of interest. In smaller datasets, the ComBat technique employs a Bayes framework, which is resilient to outliers and similar to other batch correction methods in bigger datasets [4].
The data was then scaled and centered to a mean of zero and a standard deviation of one after being normalized and batch adjusted. Prcomp was used to perform principal component analysis on the normalized and batch corrected data.

The data produced from normalization and batch correction was put through a series of filters to undertake the studies of Noise filtering and Dimensionality reduction, followed by Hierarchical clustering and subtype identification. All of the filters remain the same as those specified in Marisa et aloriginal .'s study. Three filters were applied serially to the preprocessed data for noise filtering. There were 54675 probes and 134 samples in the input data. The first filter eluted only probes with expression greater than log2 in more than 20% of the samples (15). The second filter distinguished probes with variances that differed considerably from the median variance. The t-statistic for each probe was determined using the procedure in the supplementary material using a two-tailed chi-square t-test.

As part of the second analysis approach, hierarchical clustering was done along patient samples after the noise filtered data had been acquired. The data was clustered with the defaults in hclust(), and then divided into two clusters. Using the heatmap.2() function, a heatmap was created to highlight the clustering of cancer molecular subtypes, such as C3 and others, which were color-coded with red and yellow, respectively. The grouping also made it easier to see the changes in gene expression between samples. The next step was to compare the two clusters created earlier using a Welch's t-test and p.adjust()with FDR. This was done to see if the distributions of the two clusters were identical.

Results

RLE medians were mainly centered around 0, with a few outliers ranging around -0.1 and 0.1, according to relative log expression calculations on the Marisa et al. dataset. RLE values considerably greater than 0.1 were found in two outliers (Figure 1A). Using normalized unscaled standard error, similar findings were obtained. The majority of NUSE medians were at or slightly below 1.0, with two outliers at 1.050 being much higher (Figure 1B). RLE medians of 0.0 and NUSE medians of 1.0, as previously stated, are indicative of excellent quality and perfect datasets.

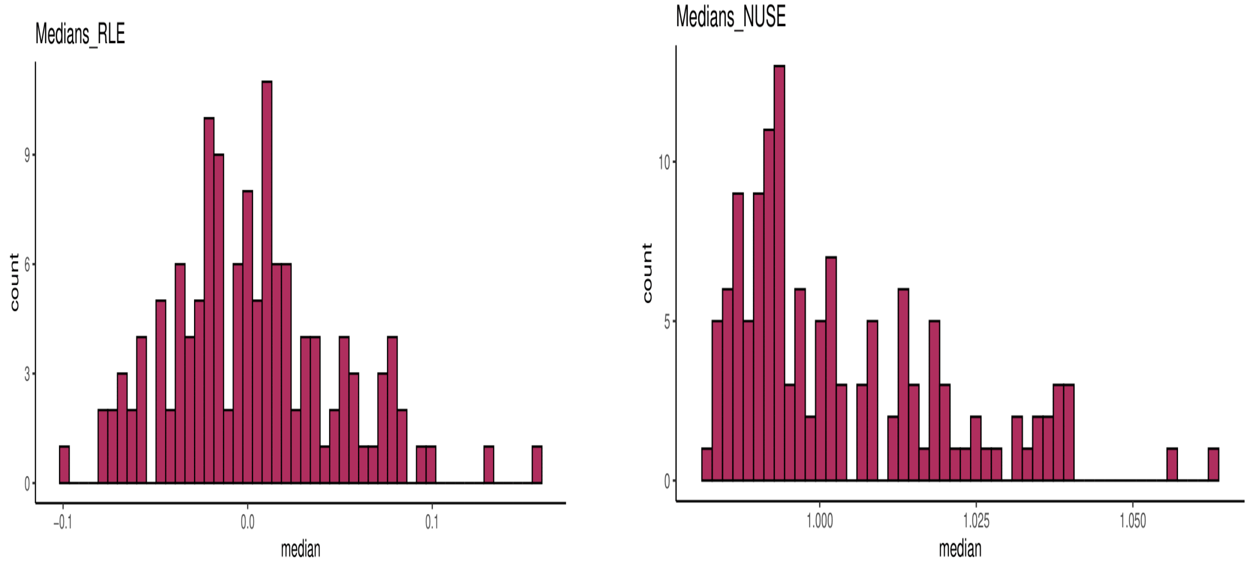

Fig. 1A (Left) Distribution of median relative log expression (RLE) values across 134 samples in the dataset. **Fig. 1B (Right)** Distribution of median normalized unscaled standard error (NUSE) values across samples.

The principal components (PC) 1 and 2 in Figure 2 seek to explain the dataset's highest percent variability. They account for 11.47 percent and 8.41 percent of variability, respectively, leaving nearly 80% of variability unaccounted for in the plot. Although the bulk of samples cluster towards the center, outliers may be seen on the plot.

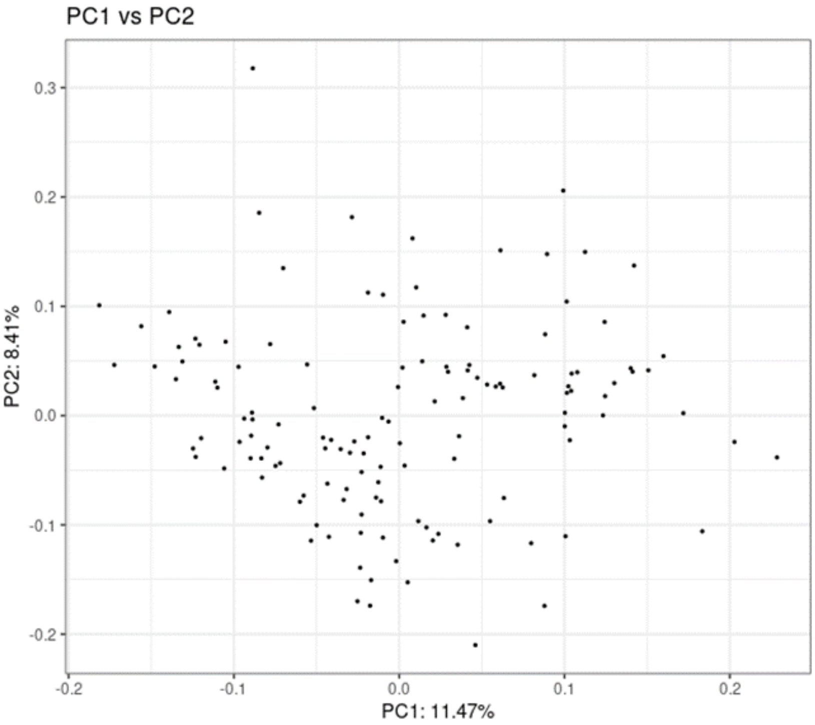

Fig. 2 Scatterplot of samples with axes as principal components PC1 and PC2, capturing the highest degrees of variability.

There were 54675 probes and 134 samples in the RMA normalized, ComBat adjusted gene expression levels collected. After going through the above-mentioned noise filters, we were left with 1531 probes. The first one produced 39661 probes by filtering the 20th decile of normalized intensity. Only carrying forward samples whose variance was substantially different from the median variance of the entire dataset were mentioned in the second filter of the chi-square t-test from the supplemental of Maris et al. A two-tailed t-test was considered as appropriate due to a lack of clarity and the wording utilized. We also performed the one-tailed upper and lower t-tests as a sanity check, and because the findings were identical, we went with the lower one.

The fully noise filtered data were then clustered by the patients and cut into two groups. The two clusters/groups comprised 57 and 77 elements respectively. The heat mapping of the filtered 1531 genes (Fig. 3) along with categorizing cancer molecular subtypes generated results according to expectations.

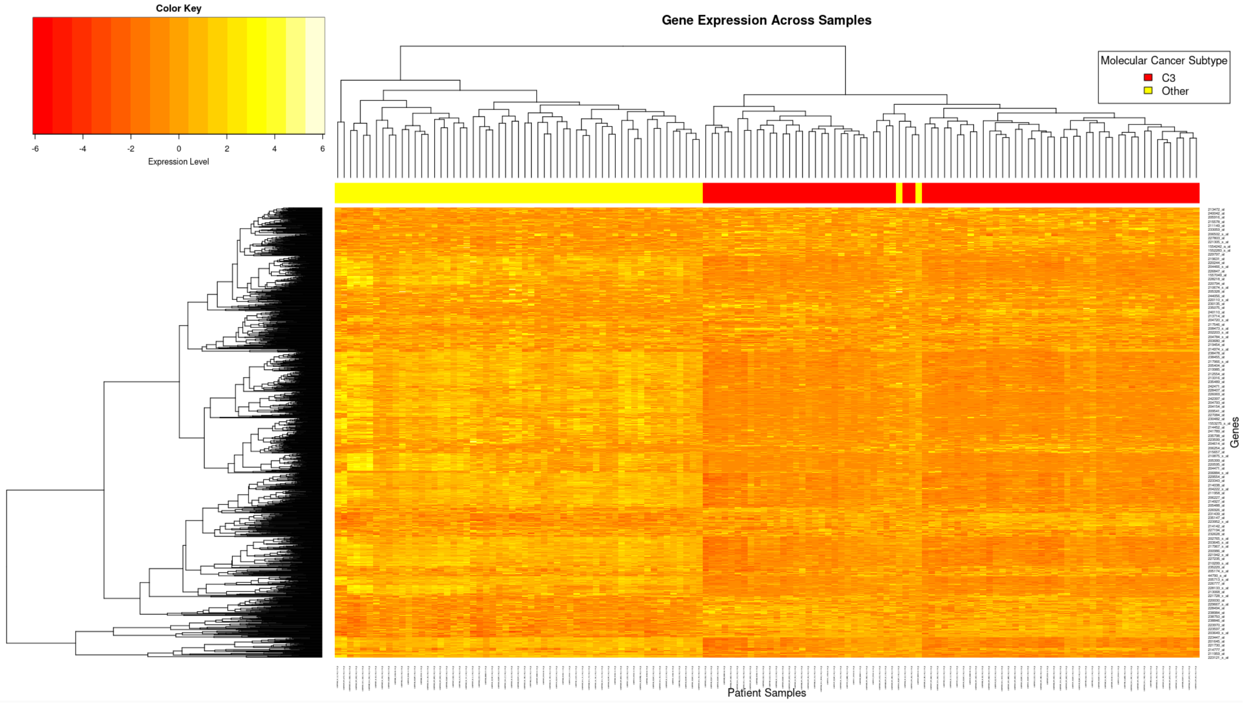

Fig. 3 Heat map of the gene expression of the 1531 probesets (y-axis) across 134 samples (x-axis). The results of the hierarchical clustering show two distinct groups and their expression levels. The column color bar is red if the sample belongs to the C3 subtype and yellow otherwise.

If the probes belonged to the C3 subtype, they were colored red, and if they belonged to any other subtype, they were colored yellow. On the C3 cluster side, however, there were two extremely distinct yellow streaks. GSM972019 and GSM972412 are example ids. There might be a variety of explanations for this erroneous correlation, ranging from flaws in the hierarchical approaches to the fact that gene expression in these samples is somewhat lower than in non-C3 subtypes, or the fact that we employed a combination of validation and discovery sets in our study. This may have resulted in the deletion of some genes that were critical for identifying the cluster identity, resulting in the incorrect categorization.

The Welch's t-test was used to compare the two hierarchical clustering clusters as the last phase of the investigation. With an adjusted p0.05, this test revealed that 1236 genes were substantially differently expressed between the two groups. The t-statistic was used to determine the most differentially expressed genes because the larger the magnitude or absolute value of the t-statistic, the greater the difference between the two sets. The probes 204457_s_at, 225242_s_at, 209868_s_at, 218694_at, and 223122_s_at revealed the most differentially expressed genes. Our findings indicate that genes 204457_s_at, 225242_s_at, and 209868_s_at best represent cluster 1, whereas genes 218694_at and 223122_s_at best represent cluster 2. The highest abs(t_stat) led to this result, as a bigger t-value suggests more gene differences.

Discussion

To verify and confirm the dataset's quality, relative log expression and normalized unscaled standard error estimations were used. RLE values in optimal datasets are always near zero, based on the idea that expression across a large number of genes should fluctuate very little in relation to the few biological variables of relevance. Because each probe set is expected to have uniform expression, NUSE values should likewise be around one. With the exception of a few outliers, the dataset was of sufficient quality to proceed with further research. Batch correction was also conducted based on the annotation data given in order to reduce artifacts caused by sample batching.

The batch adjusted samples were subjected to principal component analysis. There was no clearly identifiable grouping of samples based on the charting of PC1 and PC2. This was expected because PC1 and PC2 together explained just around 20% of the variance in the sample, which is a small amount.

The data analysis began with 54675 genes and ended up with 1236. Many of these genes were found to be identical to those found in the original Marisa et al. research. Only two non-C3 subtype cells were incorrectly associated with the C3 cell, indicating that the clustering was 98 percent accurate. As a result of the foregoing, we may conclude that our analysis was successful and promising.

References

[1] Abbas-Aghababazadeh, Farnoosh, Qian Li, and Brooke L. Fridley. "Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing." *PloS one* 13.10 (2018): e0206312.

[2]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

[3] 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

[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] 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

[5] Guinney, Justin, et al. "The consensus molecular subtypes of colorectal cancer." *Nature medicine* 21.11 (2015): 1350-1356.