Microarray Based Tumor Classification- Programmer Role

Introduction

The task being performed here is normalization of microarray data using the Robust Multiarray Averaging algorithm. Quality control metrics are then computed on the normalized data, and samples are visualized by Principal Component Analysis. Normalization is an important role for ensuring that the intensities measured are distinguishable by differential gene expression only, and not from other factors.

Methods

The packages needed to perform this task include affy, affyPLM, sva, AnnotationDbi, hgu133plus2.db, tidyverse, and ggplot. Most of these packages came from the CRAN and Bioconductor repositories. Each of the 134 microarray data samples were saved in a .CEL file, and read into the R program using the ReadAffy() function. Robust Multiarray Averaging(RMA) was used to normalized all the .CEL files.

On the data read in before normalization, the Relative Log Expression(RLE) and Normalized Unscaled Standard Error(NUSE) were computed and summarized. Median calculations for each of the samples were extracted. Both of these methods compute important quality control methods to keep track of. Median RLE values should be close to zero to ensure that the data is of good quality. Secondly, median NUSE values show good quality data if they are near or below 1.

The sva package was used to correct for batch effects on the normalized microarray data. This process required the use of a metadata CSV file, which contained information on RNA extraction and centering methods as batch effects. Features of interest were tumor and MMR status, and needed to be preserved for the batch effect correction. The output of this was a properly filtered expression data file.

Principal Component Analysis(PCA) was performed on the expression data after the expression data was scaled and centered, and the variability was able to be visualized and calculated.

Results

The results of the median RLE are represented by a histogram that is centered around zero, indicating good quality results. This can be seen below in Figure 1.

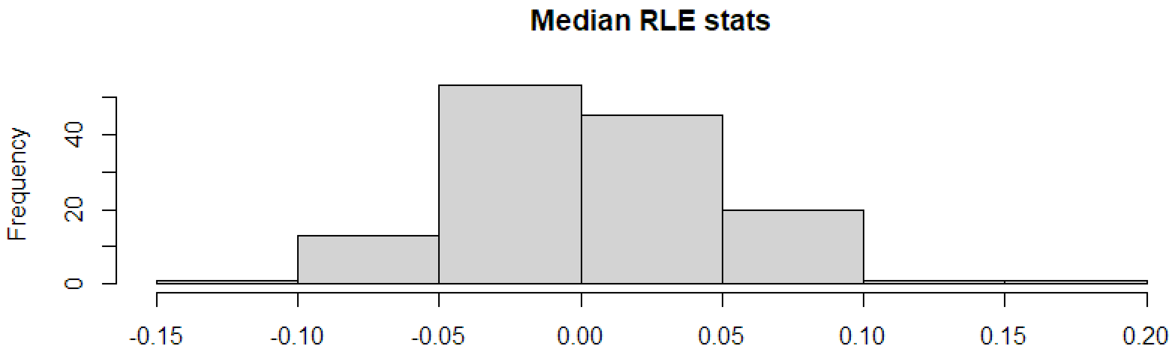

Figure 1. Relative Log Expression statistics specifically showing median values.

The values for the median NUSE are also represented by a histogram, and the quality of these values are not as strong since a fair amount of values are above 1.

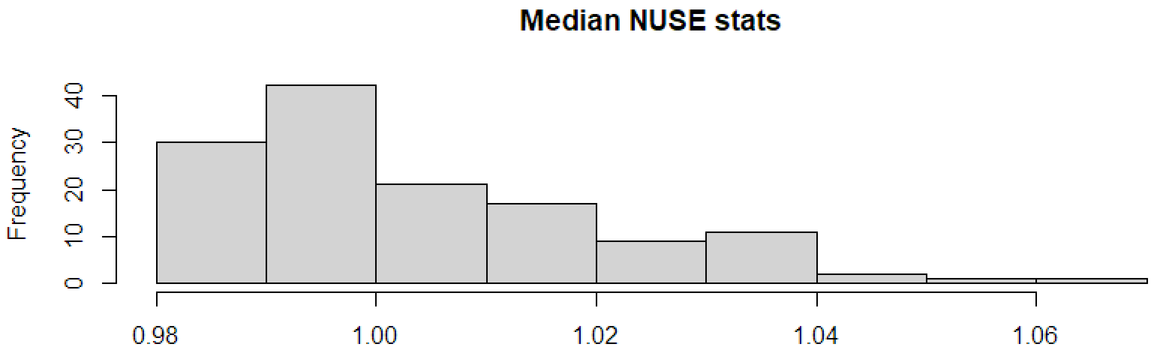

Figure 2. Normalized Unscaled Standard Error statistics specifically showing median values.

Principal component analysis(PCA) was computed for the normalized data. The first two are plotted below. PC1 had a variance of 11.47%, and PC2 had a variance of 8.41%.

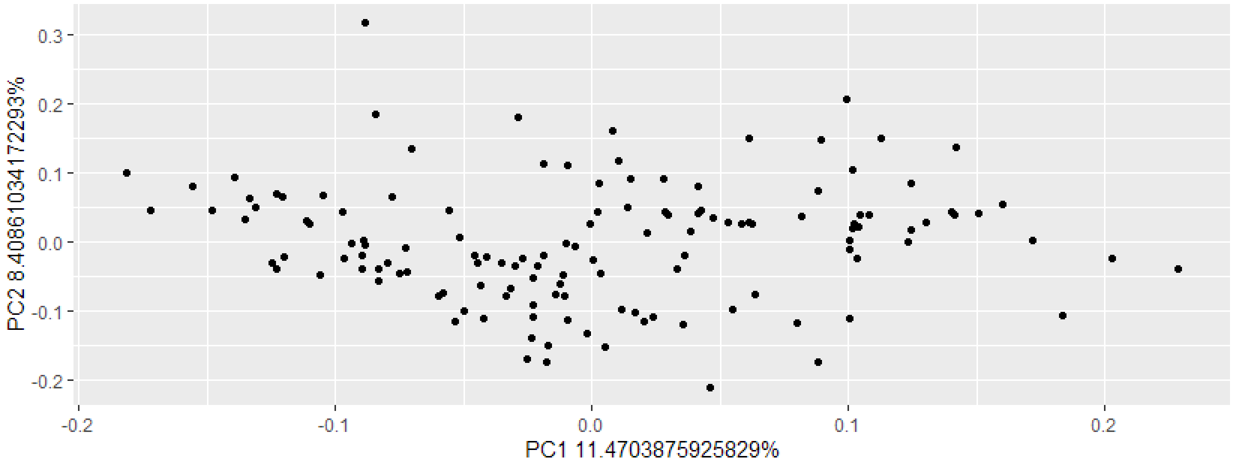

Figure 3. PCA plot of PC1 and PC2.

Discussion

The goal of the analysis was to normalize the microarray data, compute quality control metrics to ensure the data is of good quality, and visualize the variability of the data through Principal Component Analysis.

The RLE statistics showed good quality data and the NUSE statistics showed fair quality. There were many principal components, with PC1 having a percent variance of 11.4%, and PC2 having a percent variance of 8.4%.

References

1. Greenlee RT, Murray T, Bolden S, Wingo PA (2000) Cancer statistics, 2000. *CA Cancer J Clin* 50: 7–33. [PubMed] [Google Scholar]

2. 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