## **Introduction**

Colorectal cancer (CRC) is a heterogeneous disease which is the fourth leading cause of death worldwide. Pathological staging is the existing solution for the prognostic classification to select patients for adjuvant chemotherapy; yet, colorectal cancer remains the third most common among other cancers. This is because of the failures in recurrence prediction accuracy through pathological staging. [1] Early detection of CRC with reliable biomarkers could enhance prognosis, treatment response prediction, and recurrence risk. [2]

In the paper, authors of *Merisa et al*. have exploited an extensively characterized series of colon cancer samples to develop a molecular screening based on a genome-wide mRNA expression analysis to investigate prognostic biomarkers. Their further assessments included associations between molecular subtypes and clinicopathological factors, common DNA alterations, and prognosis. We have reproduced the findings in this paper using the same datasets.

## **Data**

Fresh-frozen tumor tissue samples were collected from 750 CRC patients ranging from stage I to IV from various institutes and hospitals for Marisa, et al. Only 566 of the 750 samples satisfied the RNA quality criteria, therefore they were divided into two groups: discovery (n = 443) and validation (n = 123). To generate a more rigorous validation set, the datasets were chosen based on a similar chip platform (Affymetrix U133 Plus 2.0), tumor site, and either common DNA alteration or patient outcome.

The final dataset had 359 and 416 patients from the discovery and validation sets, respectively for individuals who had stage II–III CC, and had a confirmed relapse-free survival. Hence, they were eligible for survival analysis.

The validation set for the DNA alteration annotations provided by the 152 CC samples comprised the Cancer Genome Atlas dataset, which was obtained on a non-Affymetrix technology and processed separately. To replicate Marisa, et al.’s study of C3 and C4 tumor subtypes, we downloaded these samples from GEO (Gene Expression Omnibus) under the entry number GSE39582. All The data for the analysis was stored in a common repository, excluding one sample, which was then downloaded from the GEO using the sample ID: GSM971958.

## 

## **Methods**

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 was performed on the BU Shared Computing Cluster using R 4.1.2 and BiocManager 3.14. Total runtime was about 357 seconds.

CEL files were first read into an AffyBatch object using *ReadAffy* and subsequently converted into an ExpressionSet object using the *rma* method. The RMA (robust multichip average) algorithm normalizes CEL files together through background correction, log2 transformation, and quantile normalization.

As a quality control step, the AffyBatch object was also passed into *fitPLM* to generate a PLMset object that underwent relative log expression (RLE) and normalized unscaled standard error (NUSE) calculation [3]. The medians of each of these calculations was plotted in Figures 1A and 1B to summarize and 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 the large majority of genes in the dataset are not affected by whichever biological feature is being studied. Therefore, with an ideal or close to 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 *exprs* method was next used to extract expression data from the original ExpressionSet object. Clinical and batching annotation provided by Marisa et al. was loaded into the R script in order 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].

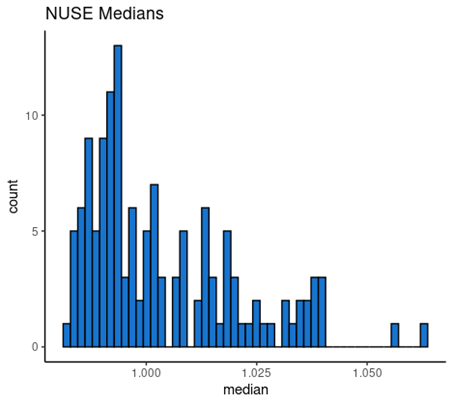
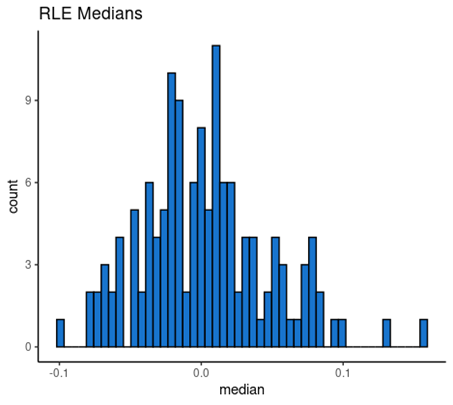
The normalized and batch corrected data was then scaled and centered to a mean of zero and standard deviation of one. Principal component analysis was performed on the normalized and batch corrected data using *prcomp*.

The data analysis of this study involved two major methods. First, Noise filtering and Dimensionality reduction, and the second, Hierarchical clustering and subtype discovery. To conduct these analyses, the data obtained from the normalization and batch correction is passed through a series of filters. All of the filters are the same as mentioned in the original study by Marisa et al. For noise filtering, three filters were applied serially on the preprocessed data. The input data contained 54675 probes and 134 samples. The first filter eluted only probes where more than 20% of the samples had expression greater than log2(15). The second filter separated probes whose variance was significantly different than the median variance. A two-tailed chi-square t-test was employed for this, the t-statistic for each probe was calculated using the formula mentioned in the supplementary material of Marisa et al. publication as *((n-1)×Var(P) / Varmed)*. A p-value of 0.01 was used to calculate the thresholds and the t-statistics, the values lying in the ranges of t\_stat > qchisq(1 − p/2, n − 1) or t\_stat < qchisq(p/2, n − 1)were taken forward. The third filter removed probes whose coefficient of variance (standard deviation/mean) was less than 0.186.

After acquiring the noise filtered data, hierarchical clustering was performed along patient samples as part of the second analyses method. The data was clustered using hclust()with the defaults and then cut into two clusters. A heatmap was also produced using the heatmap.2()function to showcase the grouping of the cancer molecular subtypes i.e C3 and others, color-coded with red and blue respectively. The clustering also helped in visualizing the differences in expression of the genes across the samples. The last step was to perform a Welch’s t-test and p.adjust()with FDR between the two clusters that were obtained above. This was performed to check the similarity between the distribution of the two clusters.

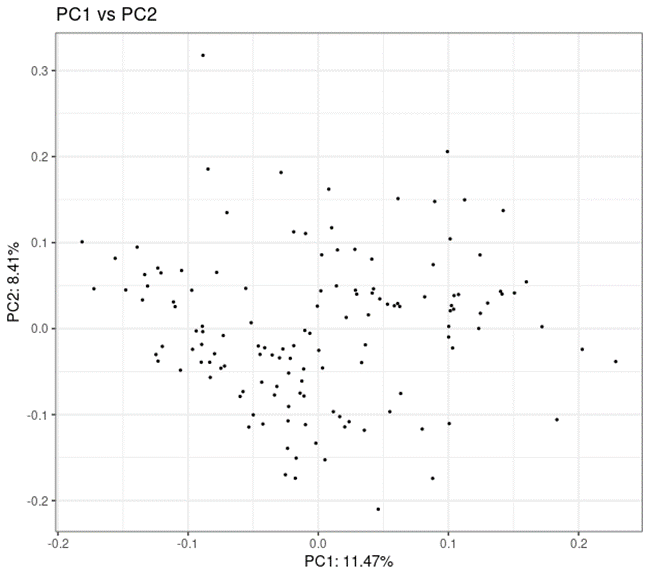
## **Results**

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 1A). 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 1B). As mentioned earlier, high quality and ideal datasets would have RLE medians at 0.0 and NUSE medians at 1.0.



**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.

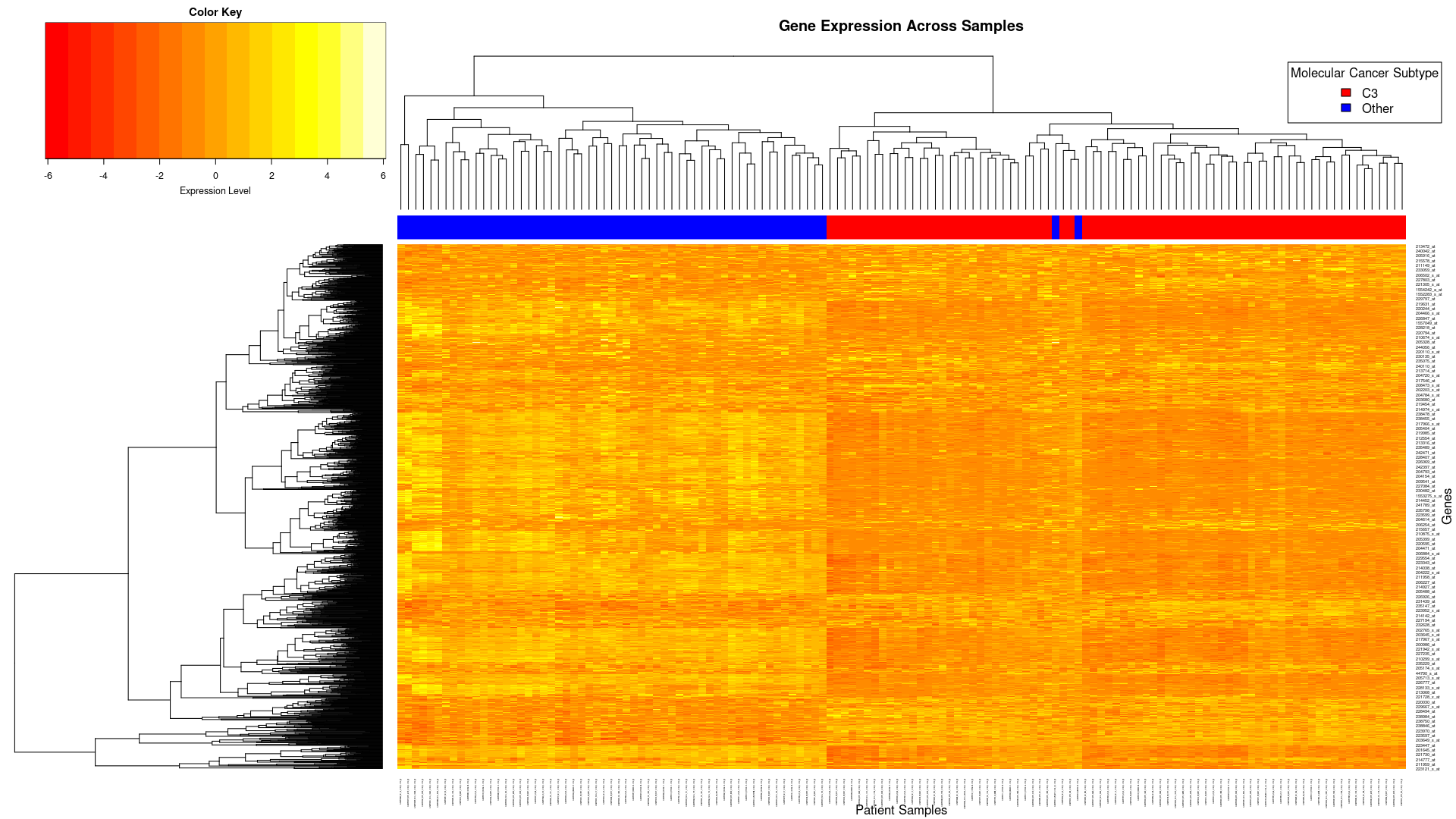
Figure 2 plots principal components (PC) 1 and 2, which attempt to explain the highest percent variability in the dataset. Respectively, they make up 11.47% and 8.41% of variability, so there is still about 80% of variability that is not captured in the plot. Outliers are observed on the plot although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



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

The RMA normalized, ComBat adjusted gene expression values obtained consisted of 54675 probes and 134 samples. After passing through a series of noise filters described above we finally obtained 1531 probes. The first one for filtering the 20th decile of the normalized intensities resulted in 39661 probes. The second filter of the chi-square t-test from the supplementary of the Maris et. al. only mentioned carrying forward samples which had their variance significantly different than the median variance of the whole dataset. Due to lack of clarity and the language used, it was interpreted that a two-tailed t-test would be fitting. Also as a sanity check, we ran the one-tailed upper and lower t-tests and since the results were identical, we chose to go with the two-tailed test. The t-test yielded 29645 probes. And as mentioned above the last filter of the coefficient of variance finally gave 1531 probes to be clustered.

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 blue otherwise.

The probes were correctly colored red if belonging to the C3 subtype and blue if any other. Although there were two very clear blue streaks present on the C3 cluster side. These are sample ids GSM972019 and GSM972412. There can be several possible reasons for this wrongful association, from faults in the hierarchical methods to the gene expressions in these samples being slightly down-regulated compared to the non-C3 subtypes or because of the fact that in our analysis we have used a mix of the validation and discovery sets. This may have led to the removal of certain genes which could have been crucial for defining the cluster identity hence leading to the wrongful classification.

The final step of the analysis was to perform the Welch's t-test between the two clusters of hierarchical clustering. This test showed that 1236 genes were significantly differentially expressed between the two clusters with an adjusted p<0.05. To find the most differentially expressed genes we looked at the t-statistic as the greater the magnitude or absolute value of the t-statistic the greater is the difference between the two sets. We found that the most differentially expressed genes were at the probes 204457\_s\_at,225242\_s\_at,209868\_s\_at,218694\_at and 223122\_s\_at.Our results signify that the genes that best represent cluster 1 are 204457\_s\_at, 225242\_s\_at, and 209868\_s\_at, while genes 218694\_at and 223122\_s\_at best represent cluster 2. This conclusion was based on the highest abs(t\_stat) since a larger t-value indicates higher differences in gene expressions between the clusters as mentioned above and the position of these genes on the heatmap.

## **Discussion**

Microarray CEL files were provided by Marisa et al. on 134 colon cancer samples, which were normalized together for downstream analyses. Relative log expression and normalized unscaled standard error calculations were made to assess and confirm the quality of the dataset. Optimal datasets have RLE values consistently around zero, relying on the assumption that expression across a large set of genes should vary very little relative to the few biological features of interest. NUSE values should also lie around one because the expectation is that each probe set has uniform expression. Aside from a few outliers, the quality of this dataset was high enough to proceed with further analysis. Batch correction was also performed based on annotation data provided, to mitigate artifacts introduced by batching of samples.

Principal component analysis was performed on the batch corrected samples. Based on plotting of PC1 and PC2, there was no clearly defined clustering of samples. This was expected because PC1 and PC2 combined to explain about 20% of the variation in the dataset, which is relatively low.

As for the analysis, we started out with 54675 genes and ended up with 1236 genes. A Lot of these genes were the same as the results obtained from the original paper by MArisa et. al. The clustering was impressively successful with an accuracy of 98% as it only wrongly clustered two non-C3 subtype cells with the C3 cell. From the above we can conclude our analysis is fruitful and promising.

## 

## **Conclusion**

The methods employed in this paper to reproduce the results of “Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value” by Marisa et. al. using hierarchical clustering with probe gene expression profiles for the classification of Colorectal cancer molecular subtype have shown promise. The results were not identically replicated but were close enough to draw strong conclusions.

The small similarities can be attributed to the difference in the data, as mentioned previously, or a mixture of the discovery and validation sets used in our analysis, which was not the case with Marisa et. al.

For CRC molecular subtype categorization and prognosis prediction, unsupervised hierarchical clustering with probe gene expression patterns provides encouraging results. Additional analysis of the underlying biology behind these patterns could be done using gene set enrichment analysis. Future studies could also dive deeper into studying varied gene expression in subtypes C3 and C4 or alternatively look at how gene expression patterns cluster in other subtypes of colon cancer.

## 

## References

[1] Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M. P., Vescovo, L., ... & Boige, V. (2013). Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS medicine, 10(5), e1001453.

[2] Langan, R. C., Mullinax, J. E., Raiji, M. T., Upham, T., Summers, T., Stojadinovic, A., & Avital, I. (2013). Colorectal cancer biomarkers and the potential role of cancer stem cells. Journal of Cancer, 4(3), 241.

[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