A Reanalysis of “Microarray based Tumor Classification”

Bettenia Cole[1]*, Taylor Falk[1]*, Mackenzie Knox[1]*, Saket Pandit[1]*
1: Boston University Bioinformatics Program, Boston MA, USA
* Lead co-authors

Analysis code available at: https://github.com/BF528/project-1-hedgehog/tree/master

Introduction

We sought to replicate a portion of the experimental methods performed by Marisa et al. in order to test the efficacy of their reported methods, as well as to improve our own understanding of microarray studies in the fields of pathology and oncology [1]. Our repeat analysis can help to authenticate not only the original paper’s results, but also the clarity and robustness of the methods described.

Marisa et al. intended to improve the prognostic evaluation of patients with colorectal cancer. In colon cancer, one the leading causes of cancer death, typical clinical methods only utilize staging to indicate prognosis [2]. However, Marisa et. al. shows that the pathological staging process is inaccurate in terms of predicting cancer recurrence. Therefore, the researchers sought to determine whether gene expression analysis for persons with stages I-IV colorectal cancer may lead to more accurate patient prognoses.

The researchers looked to identify standard signatures based on the gene expression profiles of patients from a large, multi-center tumor collection study, including the patients’ relevant clinical data. For our analysis, we used a subset of this data with the aim of reproducing their results. Specifically, we looked to replicate their results concerning two of the six tumor subtypes that they identified, C3 and C4. Microarray data was used in order to develop reproducible tests that could be used in a clinical environment, and potentially aid in prognosis of future colon cancer patients [3].

Data

The gene expression data used in our analysis came from multiple sources. The first source was the Cartes d'Identité des Tumeurs (CIT) program, which collected tumor samples from 750 different patients with stage I to IV colon cancer. Of these samples, 566 samples satisfied the RNA quality criteria as previously described [4]. The gene expression profiles (GEPs) of these samples were determined via Affymetrix U133 Plus 2.0 chips. In the original study, these samples were split into discovery and validation sets containing 443 and 123 samples, respectively. However, in our analysis, we used a subset 134 samples for our analyses, and we did not split them into discovery and validation sets.

The original authors sought to further validate their data on independent datasets. These included 906 samples coming from seven publicly available datasets. These specific datasets were selected by the authors because they satisfied several criteria. First, the GEP for the samples had to have been generated on a similar platform (the Affymetrix U133 Plus 2.0 chip). The data also had to contain tumor locations, DNA alteration annotations, and/or patient outcome data. Furthermore, 152 additional samples from The Cancer Genome Atlas were added to the validation set. Although the GEPs for these samples were generated on a different platform (Agilent), this data was chosen for its extensive annotations on DNA alterations. Overall, the authors' validation set contains 1181 samples spanning several independent datasets, and the discovery set contains 443 samples. All of these datasets, and the original studies using them, can be found using the information provided by **Table 1.**

Paper Title	**Accession Number**
DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers [5]	GSE13067 + GSE13294
Metastasis-Associated Gene Expression Changes Predict Poor Outcomes in Patients with Dukes Stage B and C Colorectal Cancer [6]	GSE14333
Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer [7]	GSE17536 / GSE17536
Molecular profiles and clinical outcome of stage UICC II colon cancer patients [8]	GSE18088
MRE11 deficiency increases sensitivity to poly(ADP-ribose) polymerase inhibition in microsatellite unstable colorectal cancers [9]	GSE26682
Methylation of cancer-stem-cell-associated Wnt target genes predicts poor prognosis in colorectal cancer patients [10]	GSE33113
Comprehensive Molecular Characterization of Human Colon and Rectal Cancer [11]	TCGA dataset, PMID = 22810696

Table 1: Publications associated with datasets used in study.

Methods

The analyses were completed using an R script running R 3.5.1 on a shared computing cluster through Boston University. The data was read into the file as a phenoData object using the function ReadAffy from the AffyBatch package [12]. The data was then normalized via a Robust Multi-Array Average Expression Measure using the rma function from AffyBatch [13]. Batch effects were corrected using the function ComBat from the R package sva [14]. Batch effects for the both Center and RNA extraction were merged into a single batch variable, and features of

interest like tumor of MMR status (as per Marisa et. al.) were merged into a single modification variable [1]. These two variables were used to run ComBat to correct for batch effects while preserving features of interest.

Data was summarized by computing the Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE). Non-normalized data was fitted using a robust linear model through the function fitPLM from AffyBatch, then summarized using the functions RLE and NUSE from AffyBatch. Results of these functions were then plotted in a histogram.

A Principal Component Analysis (PCA) was also performed on the normalized data using the prcomp function from AffyBatch. Before performing the PCA, the normalized data was centered and scaled using the function scale. PC1 was plotted against PC2 (the two components with the largest variance) to examine the dimensionality.

An R script was used to filter the normalized, ComBat adjusted expression data in three ways: gene expression among samples, variance of a probe set compared to the entire expression matrix, and the coefficient of variation of the gene [15]. Genes needed to be expressed in at least 20% of the 134 samples in filter one, and to be expressed they must exceed $log_2(15)$. Filter two compared a probe set's variance to the median variance of the entire expression matrix, and excluded probe sets with a chi-squared test result $p < 0.01$. Finally, filter three excluded probe sets with a coefficient of variation less than or equal to 0.186 ($CV = \frac{\sigma}{\mu}$). These filters were applied separately, and also combined to return an expression matrix without the probe sets that did not exceed the filters.

Using the expression data from the combined filters, we then clustered the expression data. Using the cluster and gplots R packages, we first measured the euclidean distance between all of the points, and then used the "ward.D" method of hierarchical clustering [16,17]. The effectiveness of "ward.D" was measured using the agnes test to determine the agglomerative coefficient [18]. The trees were cut into two clusters, as we are looking to compare the C3 and C4 samples, represented in **Figure 4a** as red and orange rectangles. We also used the heatmap.2() function to plot the clustered data and expression levels, and used red and blue column markers to indicate whether samples were a part of the C3 or C4 tumor subtype, respectively (**Figure 5**). Finally, we used t.test() to compare the groups generated as a part of the clustering to one another, with respect to each probe set. The results of these t-tests were associated with their respective probe set IDs, and exported to a CSV file.

Figures were generated using a combination of ggplot2, ggdendro, grid, gridExtra, and tidyverse [19–22]. All data gathering and analysis took place using Boston University's Shared Compute Cluster, https://www.bu.edu/tech/support/research/computing-resources/scc/. Analysis and plotting code available on our GitHub repository.

Results

Relative log expression (RLE) values are computed by comparing the expression value of each array against the median expression value for that set across all arrays, and so indicate how much the expression of a gene is changed from the median. Lower RLE scores indicate that genes expressed are not changing much across arrays. **Figure 1** is a histogram of the all median RLE scores across the data set. The RLE scores have a normal distribution, centered near 0.

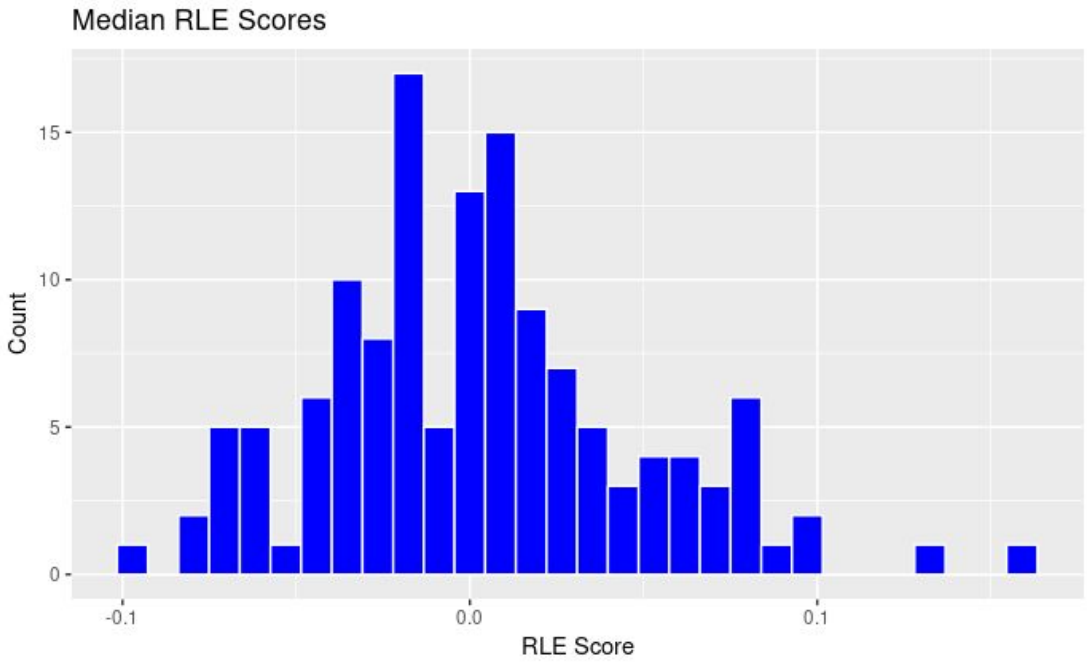

Figure 1 Median Relative Log Expression for all samples. A RLE score of 0 indicates little to no deviance from the median expression value.

Normalized unscaled standard error (NUSE) values are calculated by taking the standard error rate (generated using the function fitPLM) for each gene and standardized across all arrays so the median standard error for all genes is 1. This process accounts for differences in variability between genes, and scores closer to 1 indicate good quality samples. **Figure 2** depicts the median NUSE score across arrays in a histogram. The NUSE scores are skewed to the right close to a score of 1, indicating that the majority of samples are of high quality.

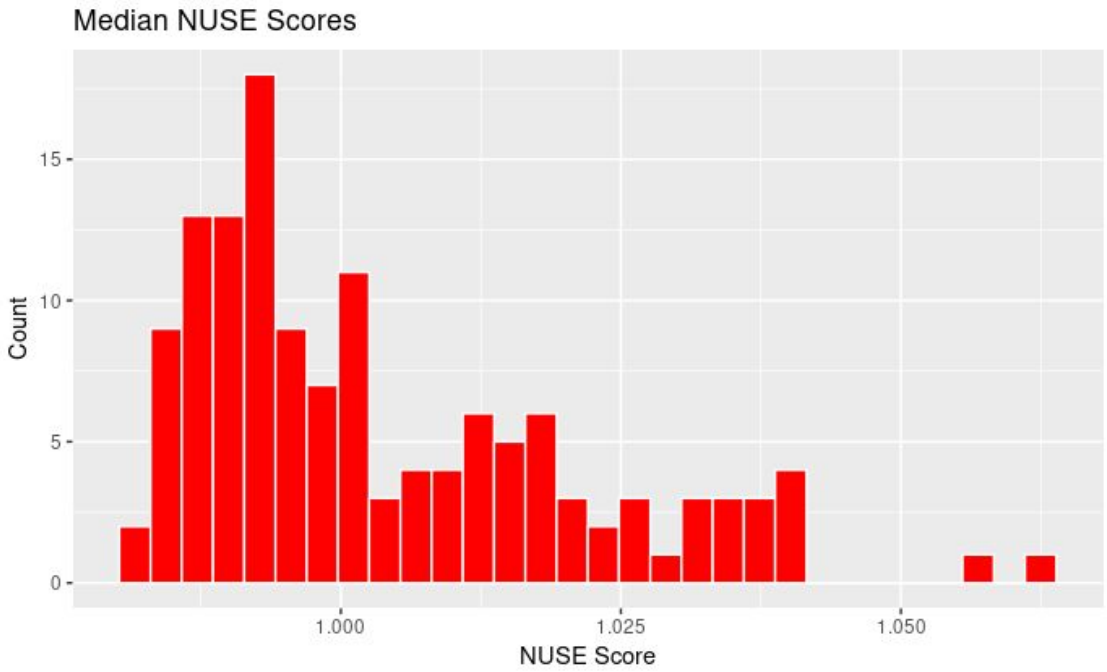

Figure 2 Median Normalized Unscaled Standard Errors for all samples. A NUSE score of 1 indicates good quality samples.

Principal component analysis (PCA) allows two-dimensional data to be reduced to one-dimension while conserving as much of the variation as possible. PCA is especially useful when variables are highly correlated, which indicates high redundancy. **Figure 3** shows the relationship between PC1 and PC2, the 2 components with the highest proportions of variability (14.52% and 9.54% respectively). The points are colored according to their subtype (C3 or C4), which indicates a separation of clusters along principal component 2.

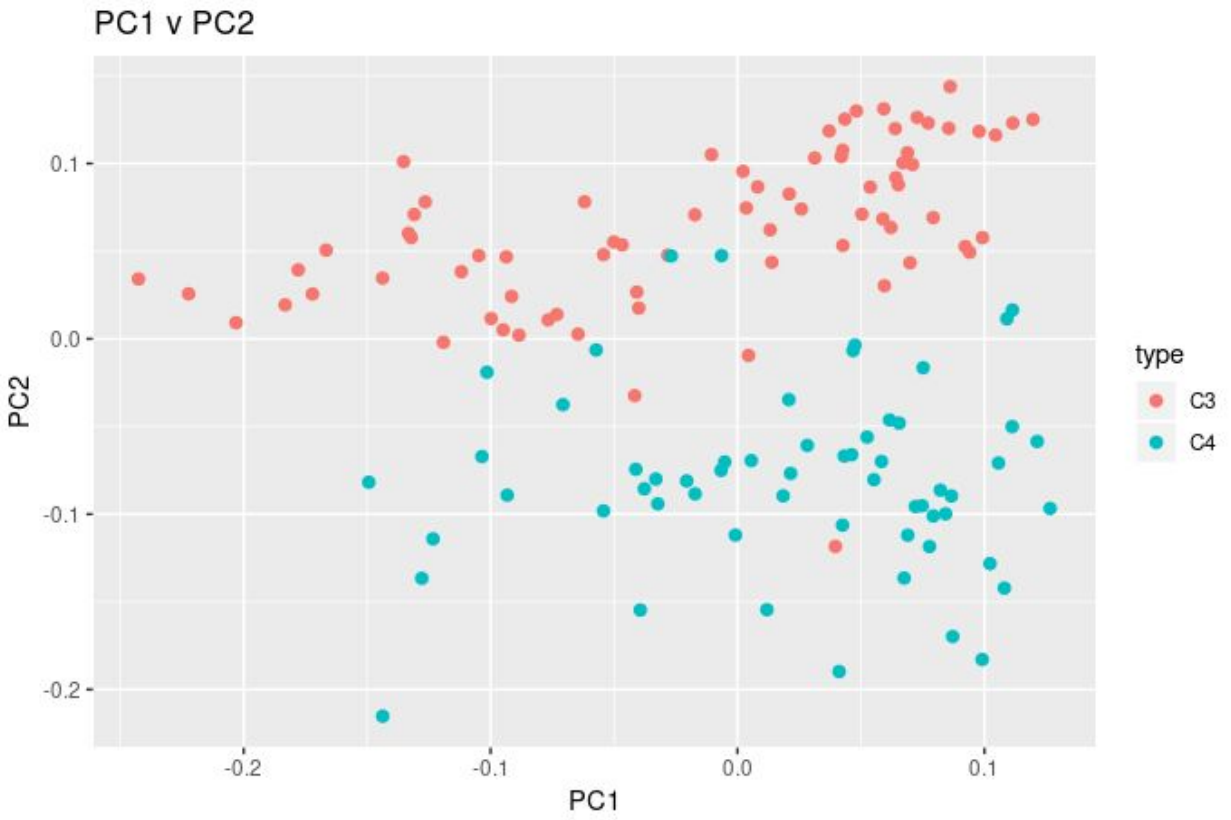

Figure 3 PC1 vs PC2 from the Principal Component Analysis of all samples, color indicating subtype.

The initial gene expression set, after normalization and ComBat adjustment, consisted of 134 samples across 54,675 probe sets. Filter one removes 15,014 probe sets, filter two removes 13,157 probe sets, and filter three removes 53,144 probe sets. There is overlap between which probe sets were filtered by which filter, and so overall we were able to filter down to 1,531 total probe sets. The original authors were able to filter down to 1,459 probe sets, so we are less than 100 probe sets away from their filtered results. However, we were able to compare the probe sets filtered to those provided from the original set, and 1,051 out of the 1,459 probe sets overlap between our data and the data from Marisa et al. The differing numbers of probe sets filtered could arise from the nuance to the filtering methods, as changing the thresholds even slightly may include or exclude a significant number of probe sets. For instance, adjusting the coefficient of variation threshold to 0.1 instead only removes 43,000 probe sets from the sample. Marisa et al.'s supplementary methods do not go into exact detail as to which statistical methods were used, so we are left to use our best judgment in selecting the correct filters for probe sets.

For our hierarchical clustering, we split the 134 patient samples into a dendrogram into two groups, where group one has 57 samples and group two has 77 (**Figure 4**).

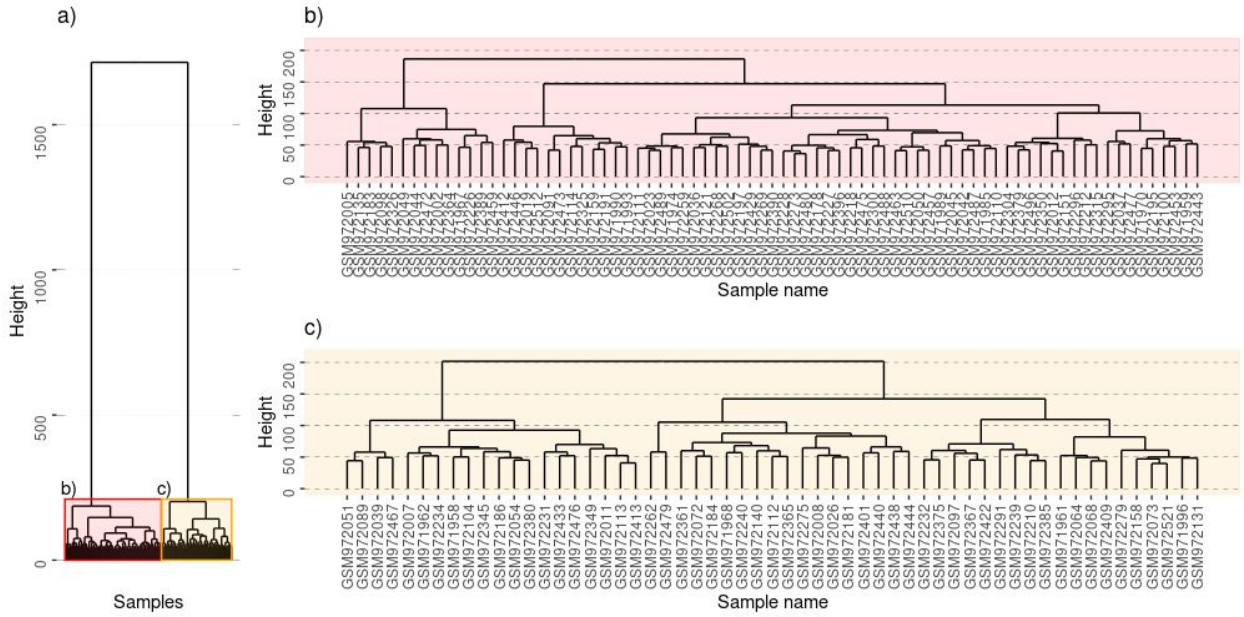

Figure 4 A dendrogram of the clustered samples we examined. **A)** All 134 samples represented, and the two clusters represented by red and orange rectangles. **B-C)** The split clusters expanded with sample names listed.

In order to judge the distance calculation methods, we calculated the agglomerative coefficient for the “ward” method at 0.9 [18]. Comparatively, other methods such as “average” had an agglomerative coefficient of 0.43. In order to visually compare the gene expression and cancer subtype relationship, we used a heatmap to plot all 134 samples and 1,531 probe sets (**Figure 5**). Finally, we used the results of our t-tests to compare the two groups, and found there were 1,070 probe sets that were significantly different from one another ($p < 0.05$).

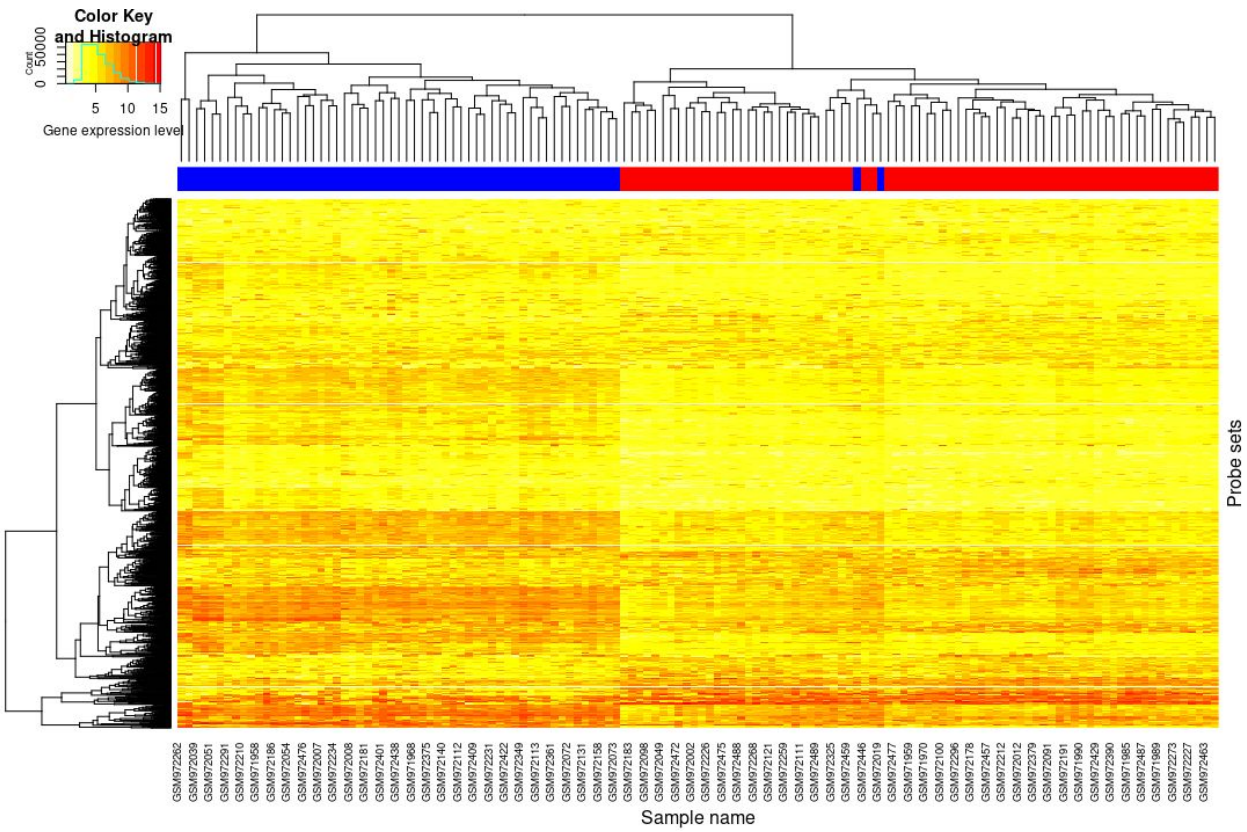

Figure 5 Heatmap describing the expression level of 134 samples plotted across 1,531 filtered probe sets. The blue bars represent subtype C4, while the red bars represent subtype C3. Dendrograms for samples and probe sets are plotted as well. Sample names shortened for brevity, but all 134 samples are represented in the heatmap.

X	probeID	t_stat	p_valu	p_adju	SYMBOL
t234	204457_s_at	2.43E+01	4.43E-49	5.94E-47	GAS1
t563	209868_s_at	2.27E+01	3.84E-47	5.15E-45	RBMS1
t945	223122_s_at	2.24E+01	1.07E-46	1.43E-44	SFRP2
t1000	225242_s_at	2.27E+01	3.10E-46	4.16E-44	CCDC80
t128	202291_s_at	2.21E+01	9.09E-46	1.22E-43	MGP
t1072	227059_at	2.21E+01	4.68E-45	6.28E-43	GPC6
t815	218694_at	2.25E+01	4.99E-45	6.68E-43	ARMCX1
t1018	225782_at	2.16E+01	6.36E-45	8.53E-43	MSRB3
t1068	226930_at	2.17E+01	6.48E-45	8.68E-43	FNDC1
t130	202363_at	2.15E+01	3.28E-44	4.39E-42	SPOCK1

Table 2 Gene Sets by Lowest p-value (adjusted).

X	probeID	t_stat	p_valu	p_adju	SYMBOL
t	1552283_s_at	-1.50E+00	1.35E-01	1.00E+00	ZDHHC11
t	1552283_s_at	-1.50E+00	1.35E-01	1.00E+00	ZDHHC11B
t6	1552511_a_at	5.80E-01	5.63E-01	1.00E+00	CPA6
t7	1552766_at	1.00E+00	3.19E-01	1.00E+00	HS6ST2
t8	1552767_a_at	2.35E+00	2.02E-02	1.00E+00	HS6ST2
t27	1554242_a_at	-2.03E+00	4.46E-02	1.00E+00	COCH
t29	1554411_at	4.63E-01	6.44E-01	1.00E+00	CTNNB1
t33	1554726_at	-1.32E-01	8.95E-01	1.00E+00	ZNF655
t37	1554997_a_at	1.77E+00	7.92E-02	1.00E+00	PTGS2
t42	1555731_a_at	-2.20E+00	3.06E-02	1.00E+00	AP1S3

Table 3 Gene Sets by Highest p-value (adjusted).

Discussion

Data was collected from a variety of sources and studies. This was to ensure that the results were reproducible across independent datasets. The normalized and batch-corrected data was then “denoised” using criteria and filters outlined by Marisa et al. A PCA plot of the normalized data was produced in order to properly examine outliers. Colors were assigned based on subtypes, revealing clustering along the PC2 axis. Distribution of the median RLE and NUSE scores indicate that the samples were of good quality and that gene expression generally did not dramatically shift across arrays.

In order to test the filtering steps outlined in the supplementary methods, we also developed three filters in R to reduce the number of probe sets studied. Marisa et al. ultimately filtered their probe sets down to 1,459, while we were only able to reach 1,531 probe sets. Furthermore, compared to their published list of probe sets, only 1,051 of our filtered probe sets match the original set. Possibilities for the difference in number and accuracy include the input expression matrix, whether due to inaccuracies in the batch correction and normalization steps or due to the fact that our analysis uses only a small subset of the data used in the original study. That being said, the adjusted filtering methods do pare the number of probe sets to a level very similar to that found in the original data. We may want to consider a slightly larger degree of uncertainty when it comes to interpreting results, but we still have a majority of probe sets in common with Marisa et al. after this filtering step.

We, and in turn the original paper, sought to find genes whose expression levels could be used to determine outcomes for sampled CRC tumors. In order to determine the most useful genes, we need to separate the “signal” from the “noise”; that is, we want to target the genes that are expressed at very high levels or very low levels. Filtering out genes that fall too close to the median of all the probe sets’ expression levels filters the genes that are unlikely to serve as good biomarkers. Even if their expression levels were correlated with prognosis or recurrence, their levels would fall too close to the other genes present, and would be indecipherable from the bulk of genes.

Once the data had been sufficiently filtered, we performed an unsupervised clustering analysis. Our analysis was different from that in the paper, where the original authors used consensus clustering, we instead performed hierarchical clustering. Samples were split such that they fell into one of two clusters, as defined by our hierarchical clustering analysis. Guided by these groupings, we then performed a differential expression analysis, and identified the most significantly differentially expressed genes between these two groups. We found that 1070 genes were differentially expressed, and 461 were not.

After we obtained a list of the most significantly differentially expressed genes, we did a gene set enrichment analysis. We compared our list of genes against three different gene ontology databases, KEGG, GO, and the Hallmark Cancer gene sets. We checked for any significant gene set enrichment with respect to these three databases using the **GSEA** package [23,24].

We realized the unique genes within the GeneCollection using BioConductor's Gene Set Enrichment Analysis or GSEA data [23,24]. We used the KEGG, GO, and Hallmark GeneSetCollection data for this analysis.

1. 5,245 KEGG GeneSetCollection
2. 19,276 GO GeneSetCollection
3. 4,383 Hallmark GeneSetCollection

In addition, we found that GAS1 had the lowest adjusted p-value ($p = 5.9e^{-47}$), and that ZDHHC11 had the highest value ($p = 1.0e^{10}$) . GAS1 is responsible for binding proteins phenotypically and is known for its expression in cell cytotoxicity [25]. ZDHHC11, composed of zinc, has binding and transferase actions and phenotypically is measured in carcinoma [26].

We used the Benjamini & Hochberg (FDR) procedure to help identify differentially expressed genes, FDR < 0.05. The purpose of this additional method is to further justify data "realness" [27].

X	probeID	t_stat	p_valu	p_adju	SYMBOL	FDR
t1	1552309_a_at	1.79E+01	3.75E-32	5.02E-30	NEXN	0
t	1552283_s_at	-1.50E+00	1.35E-01	1.00E+00	ZDHHC11	0.1588589
t	1552283_s_at	-1.50E+00	1.35E-01	1.00E+00	ZDHHC11B	0.1588589

Table 4 Top 3 Gene Sets (with BH-FDR) Analysis.

Within the set of genes analyzed, we notice that 1332 are differentially expressed while 337 are not.

FALSE	TRUE	TOTAL
337	1332	1669

Table 5 Differential expression

Conclusion

While our original intent of this report was to replicate a portion of the experimental methods of Marisa et. al. in their gene expression study, we found it difficult to interpret and represent the paper's original results and were not able to create the exact results of the original study. From our surface level investigation it appears that the original authors' interpretation of the microarray data combined with our own lead to a distinct set of roughly 1,000 genes that could be used to mark CRC stage and prognosis. Improvements to the paper's methodology and code availability would have gone a long way to helping us understand and recreate their analysis with more accuracy. Another consideration towards our results surrounds potential benign errors made early in data processing and analysis programming that may have transformed into more significant issues further down through the process. Details were also difficult to gather from the fischer test, and creating the contingency table was a challenge without outside aid.

More clarity in the instructions laid out for this report may have aided in smoothing these analyses. Increased collaboration between roles also would help with both the difficult analyses and confidence in correcting any benign errors.

References

https://astrostatistics.psu.edu/su07/R/html/cluster/html/coef.hclust.html

1. Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med. 2013;10: e1001453. doi:10.1371/journal.pmed.1001453
2. Greenlee RT, Murray T, Bolden S, Wingo PA. Cancer statistics, 2000. CA: A Cancer Journal for Clinicians. 2000;50: 7–33. doi:https://doi.org/10.3322/canjclin.50.1.7
3. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics. 2005;21: 631–643. doi:10.1093/bioinformatics/bti033
4. de Reyniès A, Assié G, Rickman DS, Tissier F, Groussin L, René-Corail F, et al. Gene expression profiling reveals a new classification of adrenocortical tumors and identifies molecular predictors of malignancy and survival. J Clin Oncol. 2009;27: 1108–1115. doi:10.1200/JCO.2008.18.5678
5. Jorissen RN, Lipton L, Gibbs P, Chapman M, Desai J, Jones IT, et al. DNA Copy-Number Alterations Underlie Gene Expression Differences between Microsatellite Stable and Unstable Colorectal Cancers. Clin Cancer Res. 2008;14: 8061–8069. doi:10.1158/1078-0432.CCR-08-1431
6. Jorissen RN, Gibbs P, Christie M, Prakash S, Lipton L, Desai J, et al. Metastasis-Associated Gene Expression Changes Predict Poor Outcomes in Patients with Dukes Stage B and C Colorectal Cancer. Clin Cancer Res. 2009;15: 7642–7651. doi:10.1158/1078-0432.CCR-09-1431
7. Smith JJ, Deane NG, Wu F, Merchant NB, Zhang B, Jiang A, et al. Experimentally Derived Metastasis Gene Expression Profile Predicts Recurrence and Death in Patients With Colon Cancer. Gastroenterology. 2010;138: 958–968. doi:10.1053/j.gastro.2009.11.005
8. Gröne J, Lenze D, Jurinovic V, Hummel M, Seidel H, Leder G, et al. Molecular profiles and clinical outcome of stage UICC II colon cancer patients. Int J Colorectal Dis. 2011;26: 847–858. doi:10.1007/s00384-011-1176-x
9. MRE11 Deficiency Increases Sensitivity to Poly(ADP-ribose) Polymerase Inhibition in Microsatellite Unstable Colorectal Cancers | Cancer Research. [cited 23 Feb 2021]. Available: https://cancerres.aacrjournals.org/content/71/7/2632.short
10. de Sousa E Melo F, Colak S, Buikhuisen J, Koster J, Cameron K, de Jong JH, et al. Methylation of Cancer-Stem-Cell-Associated Wnt Target Genes Predicts Poor Prognosis in Colorectal Cancer Patients. Cell Stem Cell. 2011;9: 476–485. doi:10.1016/j.stem.2011.10.008
11. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487: 330–337. doi:10.1038/nature11252
12. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20: 307–315. doi:10.1093/bioinformatics/btg405
13. Bolstad B. affyPLM: Methods for fitting probe-level models. Bioconductor version: Release (3.12); 2021. doi:10.18129/B9.bioc.affyPLM
14. Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Zhang Y, et al. sva: Surrogate Variable Analysis. Bioconductor version: Release (3.12); 2021. doi:10.18129/B9.bioc.sva

15. R: The R Project for Statistical Computing. [cited 22 Feb 2021]. Available: https://www.r-project.org/
16. Maechler M, original) PR (Fortran, original) AS (S, original) MH (S, Hornik [trl K, maintenance(1999-2000)) ctb] (port to R, et al. cluster: “Finding Groups in Data”: Cluster Analysis Extended Rousseeuw et al. 2021. Available: https://CRAN.R-project.org/package=cluster
17. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, et al. gplots: Various R Programming Tools for Plotting Data. 2020. Available: https://CRAN.R-project.org/package=gplots
18. R: Agglomerative Coefficient for “hclust” Objects. [cited 22 Feb 2021]. Available: https://astrostatistics.psu.edu/su07/R/html/cluster/html/coef.hclust.html
19. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2nd ed. Springer International Publishing; 2016. doi:10.1007/978-3-319-24277-4
20. Vries A de, Ripley BD. ggdendro: Create Dendrograms and Tree Diagrams Using “ggplot2.” 2020. Available: https://CRAN.R-project.org/package=ggdendro
21. Auguie B, Antonov A. gridExtra: Miscellaneous Functions for “Grid” Graphics. 2017. Available: https://CRAN.R-project.org/package=gridExtra
22. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the Tidyverse. Journal of Open Source Software. 2019;4: 1686. doi:10.21105/joss.01686
23. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics. 2003;34: 267–273. doi:10.1038/ng1180
24. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS. 2005;102: 15545–15550. doi:10.1073/pnas.0506580102
25. Del Sal G, Collavin L, Ruaro ME, Edomi P, Saccone S, Valle GD, et al. Structure, function, and chromosome mapping of the growth-suppressing human homologue of the murine gas1 gene. Proceedings of the National Academy of Sciences. 1994;91: 1848–1852. doi:10.1073/pnas.91.5.1848
26. Dzikiewicz-Krawczyk A, Kok K, Slezak-Prochazka I, Robertus J-L, Bruining J, Tayari MM, et al. ZDHHC11 and ZDHHC11B are critical novel components of the oncogenic MYC-miR-150-MYB network in Burkitt lymphoma. Leukemia. 2017;31: 1470–1473. doi:10.1038/leu.2017.94
27. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57: 289–300.