Microarray-based Tumor Classification

Monil Gandhi, Andrew Gjelsteen, Elysha Sameth, Lindsay Wang
Group Van Gogh

Introduction

Colorectal cancer (CRC) is considered to be the third most common cancer worldwide with more than 200,000 cases per year in the US alone. Pathological staging is the only classification practice used to select patients for chemotherapy, however it fails to accurately predict recurrence risk. With 10-20% of stage II and 30-40% of stage III patients developing recurrence [1], studies have been looking towards microarray technology to investigate probe gene expression profiles in CRC and predict prognosis. However, no signature for predicting recurrence risk has been established due to CRC being treated as a homogeneous entity.

This study focuses on creating a robust molecular classification based on gene expression profiles of CRC patients to investigate prognostic biomarkers. While CRC was once thought of as a homogenous entity, previous studies that have used unsupervised hierarchical clustering have identified at least three distinct molecular subtypes of CRC. By applying such methods with mRNA expression analysis, Marisa, et al. has created a reproducible system for CRC classification and molecular subtype refinement.

Methods

Within Marisa, et al., fresh-frozen tumor tissue samples were taken from 750 CRC patients ranging from stage I to IV from various institutes and hospitals. Of the 750 samples, 566 met the RNA quality requirements and were further split into a discovery (n = 443) and validation set (n = 123). In addition to the 123 validation set samples, 906 samples gathered from the following datasets were included: GSE18088, GSE33113, GSE13294, GSE14333, GSE13067, GSE17536/17537, and GSE26682. The datasets were selected on the basis of a similar chip platform (Affymetrix U133 Plus 2.0), and tumor location and either common DNA alteration or patient outcome to create a more robust validation set.

The final dataset included 359 and 416 patients, respectively, from the discovery and validation sets. These patients were diagnosed with stage II–III CC and had documented relapse free survival, making them available for survival analysis. The Cancer Genome Atlas dataset, obtained using a non-Affymetrix platform and analyzed separately, was included in the

validation set for the DNA alteration annotations given by the 152 CC samples. We downloaded these samples from GEO (Gene Expression Omnibus) under the accession number GSE39582 to reproduce the comparison of C3 and C4 tumor subtypes from Marisa, et al.

The dataset containing 134 CC patient samples were normalized together using the rma function in the affy R package, and the residual batch effects were corrected based on the metadata for the samples using the *ComBat()* method in the sva R package. To check the quality of the raw data, the Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) scores were computed using the *affyPLM* package in R for each sample. To further examine the data quality, PCA was performed on the normalized and batch-corrected data by first using the *scale()* function then the *prcomp()* function.

After normalizing the data, probe sets were selected based on three metrics suggested by Marisa, et al.: (1) expressed in at least 20% of samples, (2) have a variance significantly different from the median variance of all probe sets using a threshold of $p < 0.01$, and (3) have a coefficient of variation > 0.186. The first filter was applied to ensure that the genes are expressed in at least 27 samples (20% of all samples) so rarely expressed genes are not included in the consensus. This was done by summing the number of expression values > $log_2(15)$ for each probe set. Then, a two-tailed chi-squared test was performed to filter genes with a significantly higher variance than the average gene, i.e. genes with average expression were removed. To determine this, the test statistic (T) for each probe set (P) was calculated using $((n - 1)\ x\ var(P)\ /\ var_{med})$, where n is the number of samples, and selecting those with $T > qchisq(1 - p/2,\ n - 1)$ or $T < qchisq(p/2,\ n - 1)$. These probe sets were further filtered for high coefficient of variation ($sd(P)/mean(P)$> 0.186) to eliminate the highest and lowest expression values across samples.

After applying the filters, a hierarchical clustering of samples was performed using the *hclust()* function. A heatmap of the gene-expression across samples was visualized using *heatmap.2()* and shows the clusters with samples color-coded by their cancer molecular subtype. This was done to see how well the subtypes group together while simultaneously visualizing differences in gene expressions that could define these clusters. Finally, a Welch t-test was used to determine if the gene expressions between clusters have similar distributions by applying *t.test()* for each gene and adjusting p-values using the FDR method in *p.adjust()*. These results

were further analyzed to understand the biological significance of the different gene expression profiles for each tumor subtype.

We analyzed the dataset of probes corresponding to differentially expressed results by matching each probe to its corresponding gene symbol using Bioconductor's hgu133plus2.db package. From here, a comparison was done between these two sets of 1000 probes and each gene set in the KEGG, GO, and Hallmark gene sets in order to find how many of each up- and down-regulated genes appeared in each gene set. These gene sets contain 186, 10271, and 50 gene sets, respectively. We attempted to perform Fisher's Exact test the comparisons between the top 1000 up- and down-regulated genes and each of the gene sets in order to generate p-value and false discovery rates to compare them with those in Marisa, et. al. and validate our findings as being consistent with those of the paper. Due to time constraints, this was not accomplished.

Results

Based on the results of RLE, there are two samples with median scores greater than 0.10. Similarly, from the result of NUSE, two samples with median scores greater than 1.05 were detected (Fig. 1).

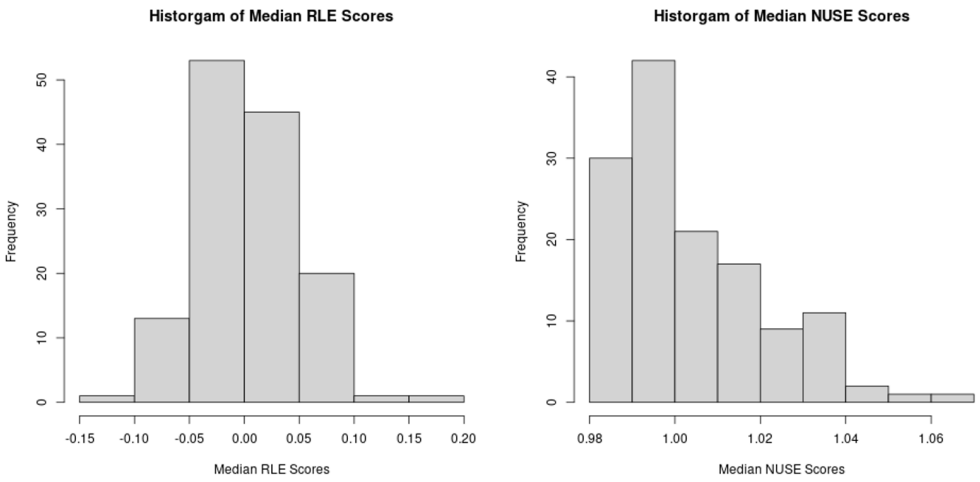

Fig. 1 (Left) Histogram of median RLE scores for 134 normalized and background corrected CC samples. Two outliers were detected (GSM971993, GSM972390). **(Right) Histogram of median NUSE scores for the same dataset.** Two outliers were detected (GSM972113, GSM972269).

Fig. 2 shows the PCA result for the two principal components that have the highest percent variabilities contributing to the result. The samples with the same subtype of CC were

clustered together based on the sample metadata. Since PC1 and PC2 only represent about 20% of the data variability, the distribution of the clustering is reasonable. The four samples with abnormal RLE and NUSE scores are all clustered within the subtype boundaries. Based on the result from RLE, NUSE and PCA analysis, none of the samples in the dataset have been identified as outliers by all three tests. Thus, the quality of the dataset was assured for further analysis.

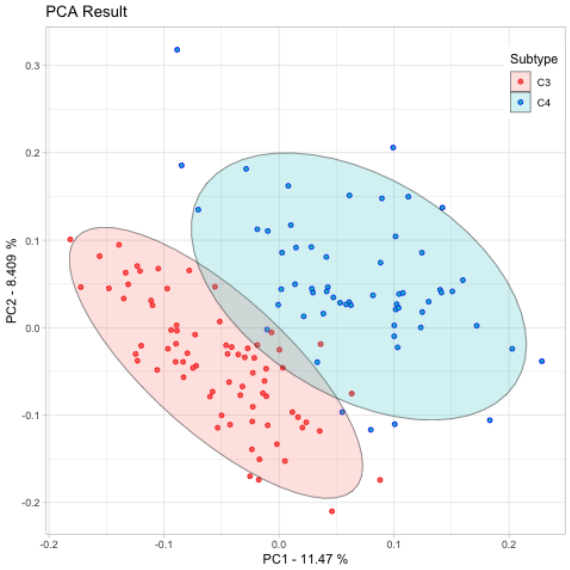

Fig. 2 PCA plot for the two principal components with the highest variability. Samples were clustered according to the CC subtypes from the sample metadata.

The RMA normalized gene expression matrix consisted of 54675 genes and 134 samples that were filtered to remove noise and select probe sets that met the criteria defined by Marisa, et al. When filtering the 20th decile of normalized intensities, 39661 genes were selected. The two-tailed chi-squared test resulted in 29645 genes, and after applying the last filter, 1531 genes remained. This is a higher number of probe sets than expected, surpassing the study's results by 72 genes. This may be due to our analysis focusing only on C3 and C4 tumor subtypes (compared to the six subtypes within Marisa, et al.) with the discovery and validation set samples combined into a single dataset. The addition of samples and removal of subtypes could affect the overall variance of the gene and the probe set median variance, thus affecting the results of our second filter. There is also ambiguity in the appropriate chi-squared test to use. While some only consider lower or upper extremes for microarray, others examine both ends of the distribution.

Marisa, et al. does not define if a lower one-tailed, upper one-tailed, or a two-tailed test was used, however we have found that the upper one-tailed and two-tailed tests give the same results.

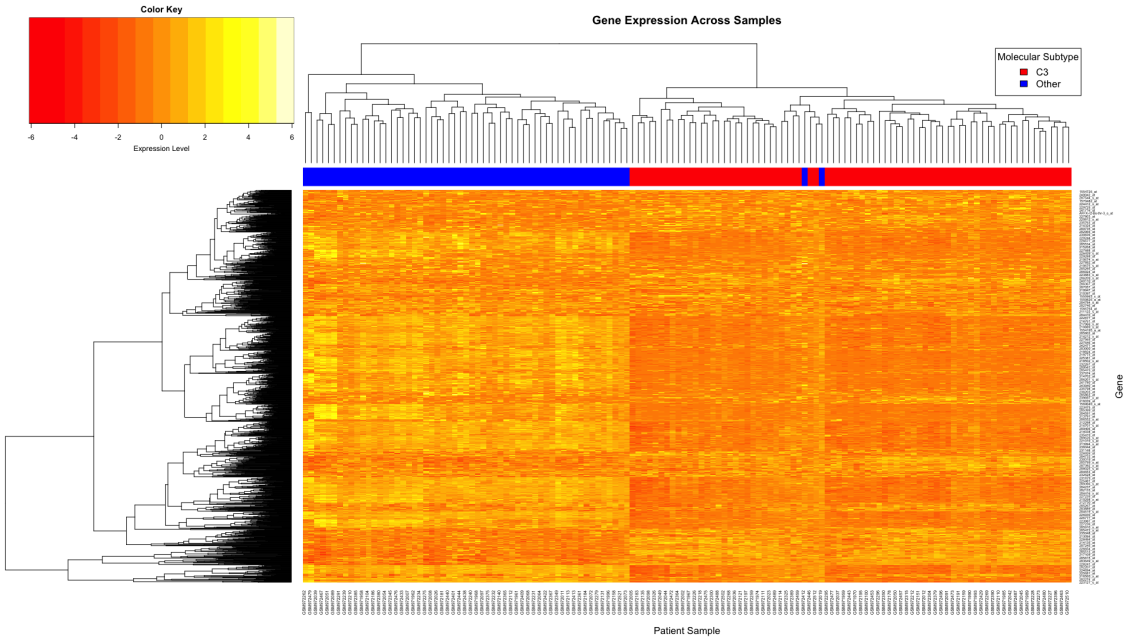

Fig. 3 Heat map of the gene-expression of the 1531 probesetes (y-axis) across 134 samples (x-axis). The results of the hierarchical clustering shows two distinct groups and their expression levels. The column colorbar is red if the sample belongs to the C3 subtype and blue otherwise.

The hierarchical clustering of the samples according to the remaining 1531 genes separated 57 samples into one group and 77 in the other. The heatmap (Fig. 3) shows that samples GSM972019 and GSM972412 incorrectly clustered with the C3 group. This may be due to the hierarchical method, which fails to get true associations and differs from the consensus clustering used in the paper, or the gene expressions in these samples being slightly down-regulated compared to the non-C3 subtypes. To explain the misclassified samples, we checked the classification for the two data points in the original PCA plot. The samples were located in the center of the C4 group, rather than on the border of the two groups. Since PCA used all the probes in the dataset for clustering, while hierarchical clustering only used a subset of the probes that have significant expression level, we think that the misclassification might be due to the probe set chosen for the clustering. By combining the validation and discovery sets, the filters may have removed important genes that helped to define the clusters, resulting in the two misclassified samples being seen as down-regulated and related to the C3 subtype.

Lastly, an observation of the Welch's t-test for the hierarchical clustering showed that 1236 genes were significantly differentially expressed between the clusters using adjusted $p < 0.05$. Some of the most differentially expressed genes were 204457_s_at, 209868_s_at, 223122_s_at, 225242_s_at, 202291_s_at, and 218694_at. We determined that the genes that best represent cluster 1 are 204457_s_at, 225242_s_at, and 209868_s_at, while genes 228004_at, 242601_at, and 238750_at best represent cluster 2. This conclusion was based on the highest *abs*(t-test) since a larger t-value indicates higher differences in gene expressions between the clusters. These locations of these genes were found in Fig. 3 and the heatmap gene expression levels of the region it resides (Fig. 4) collaborates with the results from Welch's t-test.

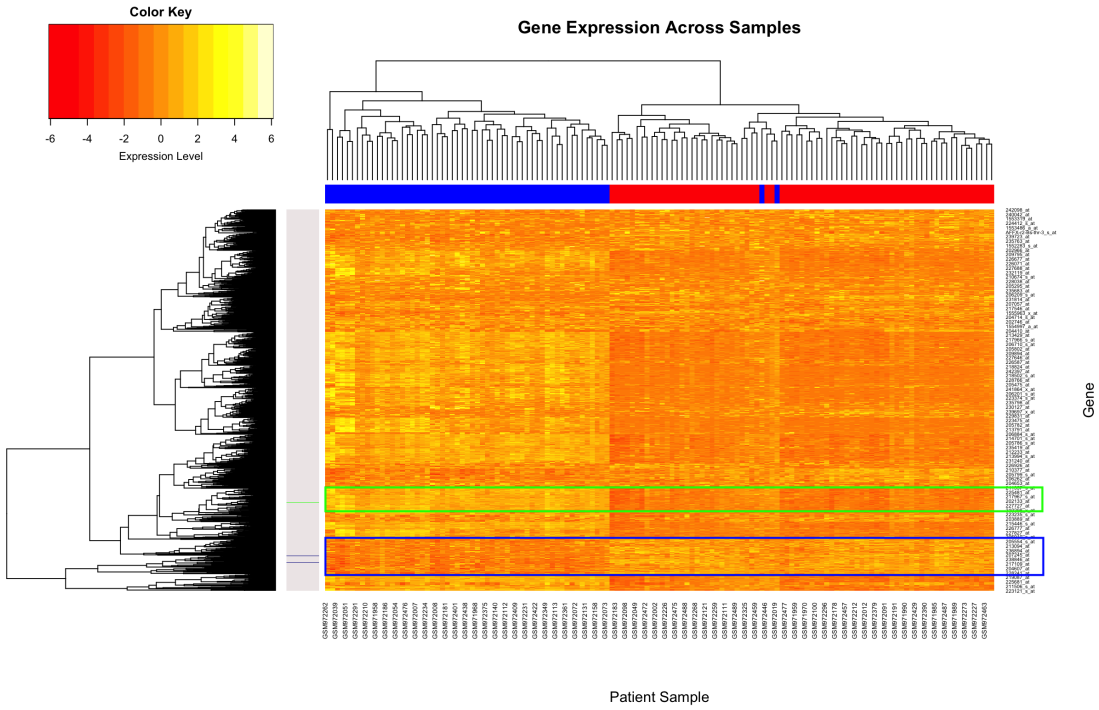

Fig. 4 Selected regions of the genes (x-axis) best representing each cluster. The selected area in green represents the genes that are more highly expressed in the C3 subtype clusters compared to the non-C3 cluster. The region in blue represents the genes less expressed in the C3 subtype clusters compared to the non-C3 cluster. These genes were determined by high Welch's t-test values, which signify higher differences in gene expressions between the clusters.

In analyzing the 1000 most up- and down-regulated genes in the enrichment samples, we found that the top three genes in each category were consistent with what we would expect from a cancer tissue genotype. The resulting dataset was filtered for the highest 1000 up- and down-regulated genes, using t-statistic and p-values as criteria. The top 10 up- and down-regulated genes are displayed in Fig. 5. Finally, performing Fisher's Exact Test on the

comparisons between the 1000 most up- and down-regulated genes and the gene sets in the collections and obtaining false discovery rates corresponding to each would help provide statistical validation that our findings were consistent with those of Marisa, et al.

	PROBEID	t	p	padj	SYMBOL
	Top 10 up-regulated genes				
6404	207266_x_at	22.09929155	3.426E-46	8.44E-42	RBMS1
15025	223122_s_at	21.94447855	8.715E-46	1.07E-41	SFRP2
4137	203748_x_at	21.83621773	1.376E-45	1.13E-41	RBMS1
7829	209868_s_at	21.08402831	6.199E-44	3.82E-40	RBMS1
4712	204457_s_at	21.3983704	1.43E-43	7.04E-40	GAS1
3007	202363_at	20.89248191	5.492E-43	2.25E-39	SPOCK1
15024	223121_s_at	21.85605081	9.123E-43	3.21E-39	SFRP2
16325	225242_s_at	20.33175915	1.14E-41	3.51E-38	CCDC80
17736	226930_at	20.09426598	1.354E-41	3.71E-38	FNDC1
2958	202291_s_at	19.76465248	4.091E-41	1.01E-37	MGP
	Top 10 down-regulated genes				
13697	220622_at	-13.4762128	2.641E-26	1.36E-24	LRRC31
8710	211715_s_at	-13.3804537	4.48E-25	1.99E-23	BDH1
12073	218189_s_at	-13.0012334	1.125E-24	4.81E-23	NANS
3719	203240_at	-13.3423834	2.688E-24	1.1E-22	FCGBP
21184	234008_s_at	-12.3952009	7.114E-24	2.78E-22	CES3
14789	222764_at	-12.7977107	7.642E-24	2.97E-22	ASRGL1
21564	235350_at	-13.1422364	1.388E-23	5.26E-22	C4orf19
5502	205489_at	-12.3063434	1.98E-23	7.29E-22	CRYM
1492	1568598_at	-12.3180267	3.28E-23	1.17E-21	KAZALD1
12618	218857_s_at	-12.4314957	3.277E-23	1.17E-21	ASRGL1

Fig. 5 Top 10 up- and down-regulated genes. The results of matching gene symbols to probe IDs and sorting by t statistic and adjusted p-values to yield the top 10 most up- and down-regulated genes. Displayed for each probe ID is its corresponding t-statistic, p-value, adjusted p-value, and gene symbol.

Discussion

In the reproduction of the Marisa, et al. study, we retrieved gene expression values from GEO, normalized and conducted gene quality analysis, performed hierarchical clustering of samples, and determined significantly enriched gene sets. Although we did not get the same results as Marisa, et al., we consider our analyses successful. The final filtered dataset contained 1052 of 1459 genes in the original study, and despite two samples incorrectly clustering with the C3 group, 132 of 134 samples were successfully grouped into their cancer molecular subtype. While gene set enrichment was not completed, the PCA plots and hierarchical clustering show promising results.

The analysis of the 1000 most up- and down-regulated genes in the enrichment samples found that the top 3-enriched probes are: RBMS1, SFRP2, and GAS1, which encode an RNA

Binding Motif Single-Stranded Interacting Protein [2], a protein that acts as a soluble modulator of Wnt signaling, a protein which binds to single stranded RNA/DNA [3], and a protein involved in cellular growth [4], respectively. These functions are involved in DNA replication and cellular signaling, and, as such, their enrichment seems consistent with that of a cancer genotype. In addition, the 3-most down-regulated probes corresponded to LRRC31, BDH1, and NANS, genes which are associated with DNA repair [5], fatty acid catabolism in the mitochondrial membrane [6], and sialic acid synthesis [7], respectively. While the first two are consistent with negative consequences for a cell's function, the down-regulation of NANS is less-obvious and is a subject of study [8].

Conclusion

Unsupervised hierarchical clustering with probe gene expression profiles shows promising results for CRC molecular subtype classification and prognosis prediction. We were able to successfully reproduce the results in Marisa, et al. with some error, however this could be contributed to looking at only C3 and C4 molecular subtypes and combining the discovery and validation sets. Despite differing results, we can say that the C3 tumor subtype from the original paper is a valid classification for CRC recurrence prediction. However, this analysis has limited applicability as all subjects were from France. Future studies should include a more diverse set of samples and take into account lifestyle and other host factors that may play a role in gene expression. Until then, our results conclude that Marisa, et al. was successful in creating a reproducible system for CRC classification and molecular subtype refinement.

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). doi:10.1371/journal.pmed.1001453

[2] https://www.genecards.org/cgi-bin/carddisp.pl?gene=RBMS1

[3] https://www.genecards.org/cgi-bin/carddisp.pl?gene=SFRP2

[4]https://www.genecards.org/cgi-bin/carddisp.pl?gene=GAS1

[5] https://www.genecards.org/cgi-bin/carddisp.pl?gene=LRRC31

[6] https://www.genecards.org/cgi-bin/carddisp.pl?gene=BDH1

[7] https://www.genecards.org/cgi-bin/carddisp.pl?gene=NANS

[8] Varki A. Sialic acids in human health and disease. *Trends Mol Med*. 2008;14(8):351-360. doi:10.1016/j.molmed.2008.06.002