Project 1 Report

Reva Shenwai
Nitsueh Kebere
Xinyu Sun
Will Mischler

Introduction

This study attempts to replicate gene expression analyses conducted by Marisa et al. (2013) on fresh-frozen primary colorectal cancer (CC) tumor samples. We conducted our analyses on a subset of their original dataset, using 134 of the 500+ samples found at NCBI Gene Expression Omnibus: GSE39582.

Prior to the Colorectal Cancer study conducted by Marisa et al. (2013), pathological staging was mainly used to determine whether a cancer patient would undergo adjuvant chemotherapy after tumor removal surgery (Lippincott-Raven, 1997). However, pathological staging does not always accurately predict CC recurrence for patients who have undergone such surgery (Popat et al., 2005)(Hutchins et al., 2011). Results from studies on gene expression profiling of CC tumor DNA have also been difficult to reproduce, likely since CC is a genetically heterogeneous disease involving different molecular pathways (Jass, 2007)(Shen et al., 2007)(Kang, 2011). Studies conducted prior to Marisa et al. (2013), including one involving high-throughput methylome data (Hinoue et al., 2012), found at least three molecular subtypes. Through their study, Marisa et al. (2013) aimed to further refine this molecular classification of CC based on mRNA expression profile analyses, in a reproducible manner. They also explored relationships between these molecular classifications and the associated clinicopathological factors, genetic variants, and patient prognosis.

Marisa et al.'s (2013) main dataset consisted of tumor samples from 750 CC patients (stage I - IV) treated in 7 centers between 1987 and 2007. About 566 of these were found to be high-quality samples, and 433 of these were used as the discovery cohort in their gene expression analysis. The remaining 123 high-quality samples, along with another independent dataset (total n=1,029), formed their validation dataset. In this project, we specifically attempted to reproduce their gene expression analysis from a subset of their discovery cohort samples.

Marisa et al. (2013) used the multi array average method from affy R package (Bioconductor) to normalize their datasets independently in batches, since their gene expression data was originally gathered on Affymetrix chips. They then used the ComBat method from the sva R package to correct for batch

effects. This was followed by unsupervised consensus hierarchical clustering on gene expression data to produce a robust classification of distinct molecular subtypes for CC. They found 6 subtypes with different clinicopathological features, named C1 - C6.

Data

Marisa et al. used microarrays to study gene expression profiles (GEPs) in CC tumors. Specifically, they used Affymetrix U133 Plus 2.0 chips to analyze their 566 high-quality samples.

Their protocols for RNA purification, quality control, fluorescent probe production, hybridization, and raw data processing were described in de Reyniès et al. (2009). The fresh tumor samples were subjected to snap-freezing right after surgery, and then stored in liquid nitrogen. This was followed by powdering using the liquid nitrogen, RNA extraction using RNAble and clean-up using RNAeasy columns. Quality control involved electrophoresis on a Bioanalyser 2100 and quantification using NanoDrop ND-1000, with stringent criteria to rule out RNA degradation. Next, microarray analysis was conducted with 3 µg RNA from each sample and 10 µg cRNA per hybridization. Total RNA was amplified and labeled using Affymetrix's one-cycle target labeling protocol. Then, the labeled cDNA was hybridized to HG-U133 Plus 2.0 Affymetrix GeneChip arrays. Finally, the chips were scanned with a GCOS 1.4.

The data was quality controlled by screening with whole genome and transcriptome arrays. As stated above in Introduction, Marisa et al. (2013) found that 566 of their original 750 patient samples fulfilled RNA quality requirements, and the remainder were not used in their analysis. While they added additional datasets to their analyses, we used a subset of these 566 high-quality samples (n=134) in our study. Our analysis later revealed that our n=134 dataset consisted of 2 subsets; 75 of these were of C3 molecular subtype (of CC) and 59 were of C4 molecular subtype.

Methods

All data preprocessing, statistical analysis, and quality control was done in R 3.6.1 and BiocManager 3.9 using packages affy, affyPLM, sva, AnnotationDbi, hgu133plus2.db, and ggplot2. Probe level data (CEL files) was read and stored into an AffyBatch object using the ReadAffy function of the affy package. To examine the data's distribution, the AffyBatch object was converted into a PLM dataset, which is a class representation for Probe Level Linear Models fitted to Affymetrix GeneChip probe level data. This was done using the fitPLM function of the affyPLM package. The dataset was then summarized by computing the median RLE and NUSE for each sample using the RLE function and NUSE function of the affyPLM package respectively. Since RLE is computed based on the assumption that most of the genes are not changing in expression across samples, the ideal median RLE values should be near 0. Meanwhile NUSE

is computed by standardizing the standard error estimates across samples, so that the ideal median standard error for the genes is close to 1 across all samples.

The data contained in the AffyBatch object was then normalized using the Robust Multi-Array Average expression measure (RMA) function of the affy package. This function corrects the background based on a convolution model, converts the affybatch object into an expression dataset in log base 2 scale, and normalizes the data using quantile normalization. Next, to obtain a ComBat adjusted expression dataset, the RMA normalized data was corrected for batch effects using the ComBat function of the sva package. The batch and mod arguments used in this function were obtained from the annotation file that contains clinical and and batching annotation used by Marisa et al. (2013) in their study. Principal component analysis (PCA) was then performed to determine the percent variation that each principle component (PC) accounts for in the dataset. In this analysis, the data was scaled and centered within each gene, and then the prcomp function was used to obtain the principal components. After this, standard deviation obtained from prcomp was used to calculate how much variation in the original data each principal component accounts for in the PC1 vs PC2 graph plotted using the ggplot function of the ggplot2 package.

The Combat adjusted expression dataset obtained was first filtered with intensity greater than log2(15) for more than 20% of total samples to exclude low expression probe sets. Then, a chi-squared test was performed to differentiate probe sets that were significantly different from the median variance of all probe sets. Furthermore, the coefficient of variation(CV) was measured for probe sets that were significantly different to determine the dispersion of the probability distribution of each probset. Then hierarchical clustering was performed to group samples into two groups, where a two sample Welch T test was used to examine if there was any significant difference between genes.

The top 10 up and down regulated differentially expressed probesets, with matching gene symbols, were selected based on T-test statistics to determine the most differentially expressed probe sets were mapped to gene symbols using the *hgu133plus2.db*.

A Fisher Exact test was performed on our observed differentially expressed gene list compared to three gene set collections using the *fisher.test* function. These included Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Hallmark gene sets, containing 9,996, 186, 50 gene sets respectively. Each collection was downloaded from the Molecular Signatures Database (MSigDB).

Results

The median RLE and NUSE values were computed for each sample, and the distribution of the medians was plotted in a histogram (Fig. 1, Fig. 2). As seen in Figures 1 and 2, all the samples have median RLE and median NUSE close to 0 and 1 respectively, indicating they are of good quality. Furthermore, this suggests that none of the samples stand out as being extremely different; i.e. no significant outliers. Thus, there was no reason to exclude any samples from the analysis.

The PCA performed helped determine the percentage of variation represented by n-1 PCs. According to Figure 3, the first two PCs account for more variability than other PCs. However, the two PCs do not cover enough variabilities, and this can be seen in Figure 4, where the variations among samples was not significantly evident. Moreover, the PC1 vs PC2 plot(Fig. 4) did not depict any significant outliers, further strengthening our reasoning for not removing outliers.

There were 1531 genes that passed all of the threshold in noise filtering and dimensionality reduction procedure. From the hierarchical clustering, 134 samples were divided into two clusters. There are 57 samples in cluster 1(C4) and 77 samples in cluster 2(C3).

After performing the Welch T test, there were 1236 genes that had adjusted p-values(FDR) less than 0.05. Among the 1236 genes, all genes were considered statistically significant. 10 genes were chosen after ranking the 1236 genes from lowest to highest adjusted p-value. The 10 genes are: GAS1, RBMS1, SFRP2, CCDC80, MGP, ARMCX1, GPC6, MSRB3, FNDC1, and SPOCK1. According to Welch T test, these 10 most significant genes have the least type-I error-rate α, which means they are very unlikely to be false positives. Therefore, they are the best representative genes to differentiate two clusters. The top 10 up and down regulated differentially expressed genes, not based on adjusted p-values, are shown in Table 2 and 3, respectively.

There should be 75 C3 and 59 C4 subtype patients, but the hierarchical clustering gives 77 and 57 for each group respectively. It is possible that some patients are outliers. From the heatmap (Fig. 5), it is clear that two C4 subtypes are mixed with C3 subtypes.

Gene set enrichment analysis did not show any gene sets significantly enriched indicated by an adjusted p-value of less than 0.05. Based off of the nominal p-values, three groups were inferred from each gene set collection as potentially the top 3 enriched gene sets. (Tables 4, 5 and 6). The gene sets mainly corresponded to cell adhesion, metabolism and receptor complexes.

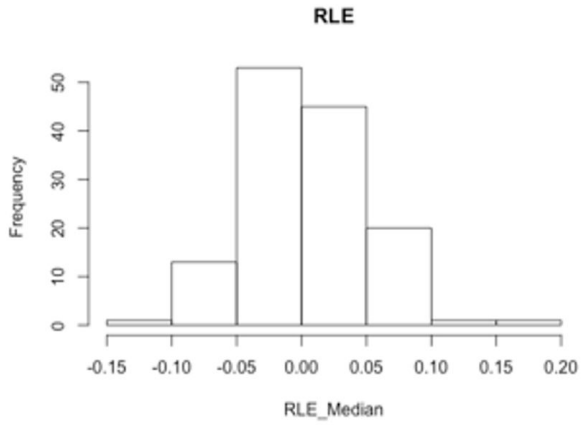

***Figure 1 RLE median distribution histogram**. RLE medians are calculated across samples. Over 95% of the RLE median values are within -0.05-0.00.*

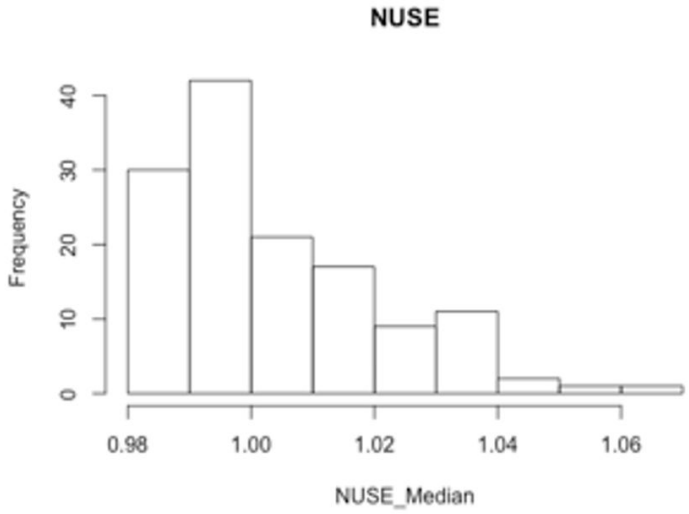

***Figure 2 NUSE median distribution**. NUSE medians are calculated across samples. Over 95% of the NUSE median values are around 1.00.*

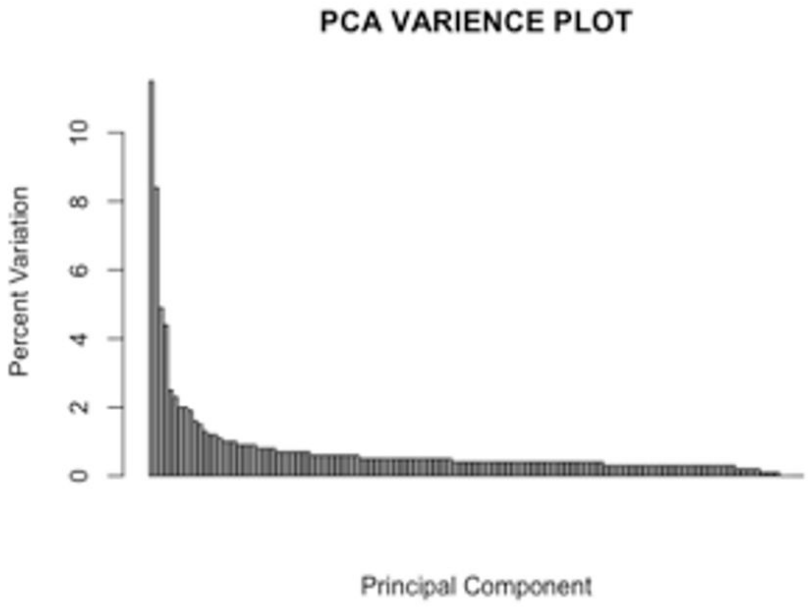

Figure 3 PCA Variance Plot. *The plot shows the percentage of variation for each principle component. The first two component only represents about 20% of the data.*

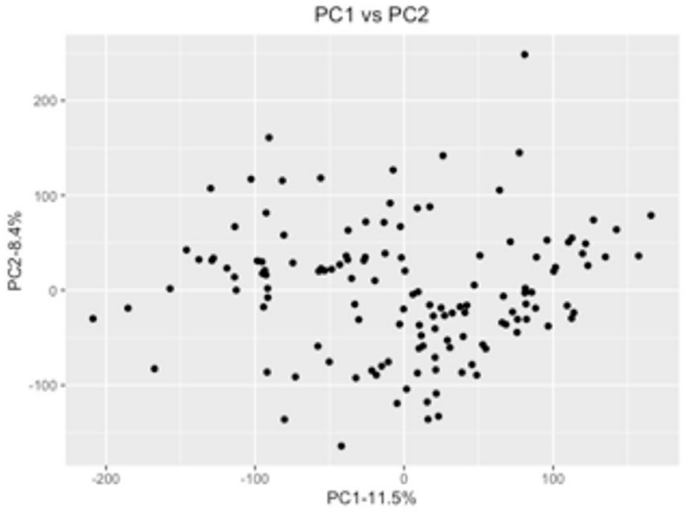

Figure 4 The PCA plot of samples. *The x axis is the first principle component and y axis is the second principle component. The two axes represent only about 20% of the variation.*

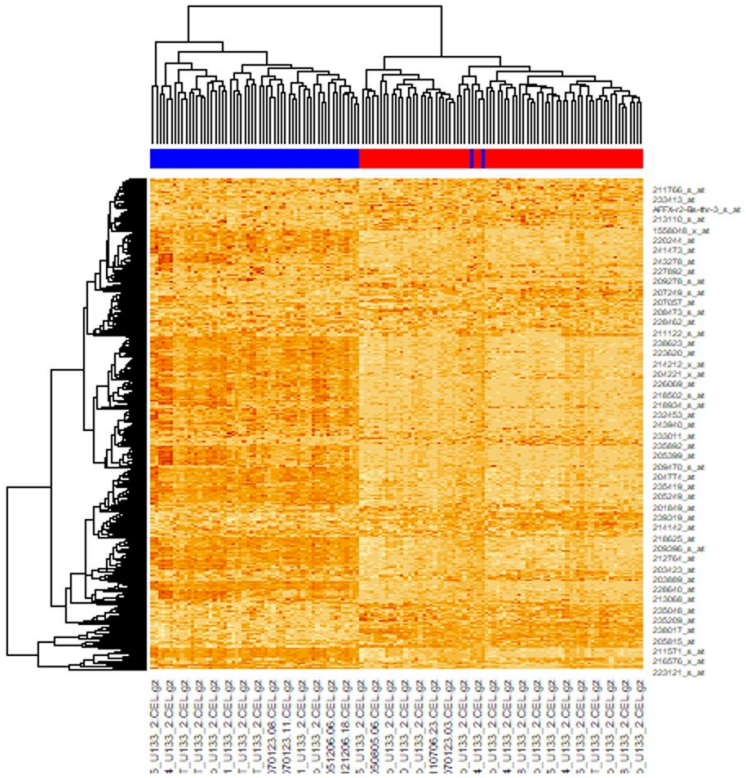

Figure 5 Hierarchical clustering heatmap of 1531 probe sets, using spearman distance matrix, defines two clusters of 134 samples, C3 and C4 subtype CRC. Red is C3, and blue is C4.

Table 1: Top 10 Up-Regulated Probesets Differentially expressed genes up-regulated in cluster 1 vs cluster 2 through the Welch's T-Test based on Probeset IDs.

Probeset ID	T-Test Statistic	P value	Adjusted p value	Gene Symbol
204457_s_at	24.26186958	4.43E-49	6.78E-46	GAS1
225242_s_at	22.68808868	3.10E-46	1.19E-43	CCDC80
209868_s_at	22.66264234	3.84E-47	2.94E-44	RBMS1
218694_at	22.51415916	4.99E-45	1.09E-42	ARMCX1
223122_s_at	22.35166133	1.07E-46	5.44E-44	SFRP2
227059_at	22.1272134	4.68E-45	1.09E-42	GPC6
202291_s_at	22.09309835	9.09E-46	2.78E-43	MGP
223121_s_at	21.97947713	1.21E-41	1.22E-39	SFRP2
226930_at	21.74075571	6.48E-45	1.10E-42	FNDC1
225782_at	21.59857435	6.36E-45	1.10E-42	MSRB3

Table 2: Top 10 Down-Regulated Probesets Differentially expressed genes down-regulated in cluster 1 vs cluster 2 through the Welch's T-Test based on Probeset IDs.

Probeset ID	T-Test Statistic	P value	Adjusted p value	Gene Symbol
203240_at	-16.21911271	4.69E-29	3.99E-28	FCGBP
227725_at	-13.49233909	2.82E-23	1.43E-22	ST6GALNAC1
204673_at	-13.40696633	1.77E-24	9.80E-24	MUC2
210107_at	-13.27444621	2.90E-25	1.70E-24	CLCA1
220622_at	-13.13460682	1.47E-25	9.00E-25	LRRC31
238750_at	-12.72528827	1.09E-24	6.11E-24	CCL28
242601_at	-12.58634394	5.03E-24	2.68E-23	HEPACAM2
1553828_at	-12.51860178	8.62E-24	4.47E-23	NXPE1
230615_at	-12.29261929	5.89E-23	2.90E-22	DUOXA2
207214_at	-12.21224576	7.67E-22	3.43E-21	SPINK4

Table 3: Top 3 Enriched GO Gene Sets Fisher test performed on GO gene set collection. Odds ratio is statistic estimate

Gene Sets	Statistic Estimate	P value	Adjusted p value
GO_RECEPTOR_COMPLEX	1.898811038	0.0001005619516	1
GO_MITOCHONDRIAL_MATRIX	1.812251084	0.0001010046828	1
GO_CELL_CELL_ADHESION	1.496747976	0.0001054489712	1

Table 4: Top 3 Enriched Hallmark Gene Sets Fisher test performed on Hallmark gene set collection. Odds ratio is statistic estimate

Gene Set	Statistic Estimate	P value	Adjusted p value
HALLMARK_BILE_ACID_METABOLISM	3.160640925	0.000539375147	1
HALLMARK_COAGULATION	2.371654573	0.0008183073284	1
HALLMARK_ADIPOGENESIS	2.038724448	0.0008762024348	1

Table 5: Top 3 Enriched KEGG Gene Sets Fisher test performed on KEGG gene set collection. Odds ratio is statistic estimate

Gene Set	Statistic Estimate	P value	Adjusted p value
KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS	8.476521726	0.0001249817649	1
KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	2.858749592	0.0004693363792	1

KEGG_FOCAL_ADHESION	1.905662947	0.001977298476	1

Discussion

The hierarchical clustering of 134 samples shows a split into two distinct clusters, one for C4 subtype and another for C3 subtype, as visualized by the heatmap (Figure. 5). This clustering, based on the differential gene expression of 1,236, indicates two samples, originally classified as C4 subtypes that are misclassified as C3 subtypes according to the gene expression displayed. This indicates that the classification method used to subtype the data is largely accurate but can still misidentify some outliers as a different subtype.

Through the use of Welch's T-Test, 1,236 genes were identified as differentially expressed. A positive T-statistic indicates that the gene expression is higher in cluster 1 than in cluster 2 while a negative T-statistic indicates that the gene expression is higher in cluster 2 than cluster 1. This is shown in Table 2 and Table 3 with *GAS1* being the highest differentially expressed gene in cluster 1 (T-Test statistic = 24.26) and *FCGBP* being the highest differentially expressed gene in cluster 2 (T-Test statistic = -16.22). This indicates that there are drastically different levels of gene expression across these two subtypes. Furthermore, the same genes indicated to be markers of poor prognosis in C4 subtype were still indicated as upregulated in C4, as shown in our subset of information.

In our GSEA (Gene Set Enrichment Analysis), we found a similar pathway upregulated as Marisa et al. (2013) This pathway was the focal adhesion in KEGG annotation, involved in cell communication. Originally, it was found to be upregulated in C4 subtypes and downregulated in C3 subtypes. While we did not find a p-value with as much power as the original paper, possibly due to analyzing only a subset of the data available, we did still see a correlation and possible enrichment in the same pathway. Other pathways seen in the GSEA are also involved in the cell communication/signaling pathways and metabolism in a similar fashion as the original classifications. These include cell-cell adhesion and receptor complex in GO annotations and adipogenesis and bile acid metabolism in Hallmark.

Conclusion

Through our analysis on a subset (n=134) of samples from the Marisa et al. (2013) paper, we found that they were able to further refine the classification of colorectal cancer into six different colorectal cancer subtypes with high relative accuracy. Specifically, we confirmed their classification of the C3 and C4 colorectal cancer subtypes and were able to confirm the presence of differentially expressed genes and enriched pathways based on our own expression analysis and GSEA, to potentially be used as markers and/or therapeutic targets in the future.

References (APA format)

American Joint Committee on Cancer (1997) AJCC cancer staging manual, 5th edition. Philadelphia: Lippincott-Raven.

de Reyniès A, Assié G, Rickman DS, Tissier F, Groussin L, et al. (2009) Gene expression profiling reveals a new classification of adrenocortical tumors and identifies molecular predictors of malignancy and survival. J Clin Oncol 27: 1108–1115.

Hinoue T, Weisenberger DJ, Lange CP, Shen H, Byun HM, et al. (2012) Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res 22: 271–282.

Hutchins G, Southward K, Handley K, Magill L, Beaumont C, et al. (2011) Value of mismatch repair, KRAS, and BRAF mutations in predicting recurrence and benefits from chemotherapy in colorectal cancer. J Clin Oncol 29: 1261–1270.

Jass JR (2007) Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. Histopathology 50: 113–130.

Kang GH (2011) Four molecular subtypes of colorectal cancer and their precursor lesions. Arch Pathol Lab Med 135: 698–703.

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., Olschwang, S., Milano, G., Laurent-Puig, P., 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

Popat S, Hubner R, Houlston RS (2005) Systematic review of microsatellite instability and colorectal cancer prognosis. J Clin Oncol 23: 609–618.

Shen L, Toyota M, Kondo Y, Lin E, Zhang L, et al. (2007) Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. Proc Natl Acad Sci U S A 104: 18654–18659.