Project 1 Writeup

Daniel Glukhov (Data Curator), Ashley Morrison (Programmer), Muxi Wang (Analyst)

Abstract

Current pathological staging methods for stage II and stage III colorectal cancer (CRC) fail to predict the rate of recurrence accurately. As a result, a new molecular classification model is needed for CRC diagnosis based on mRNA expression profiles. Prior work by Marisa et. al. has concluded that CRC can be classified along six molecular subtypes, each linked to both differing biological pathways as well as prognostic outcomes. The aim of this study is to validate their discovery of the C3 and C4 subtypes of stage II and stage III CRC through replicating their data processing and analysis techniques. Our analysis first demonstrates a workflow for normalization of microarray data without loss of gene expression profile definition. Our analysis also demonstrates the robust nature of Marisa et al's original subtyping scheme, as alterations to both probe selection and clustering methods resulted in only one of the C4 samples clustering with the C3 group.

Introduction

Colorectal cancer is currently the third most common cancer diagnosis (3) and second leading cause of cancer deaths (2). At the time of publication of Marisa et al, pathological staging was the only clinically relevant classification for CRC adjuvant therapy selection (1). A limit to staging is the inability to predict recurrence, which occurs in 10-20% of patients with stage II CRC, and 30-40% of patients with stage III CRC (1). Previous attempts to isolate the source of inter-tumoral heterogeneity using unsupervised hierarchical clustering have identified three distinct CRC subtypes, but the limited DNA markers used in the analysis and the molecular complexity of CRC has resulted in low subtype reproducibility. The only reoccurrence marker found to be significantly prognostic in both meta-analysis and clinical trials for colon cancer (CC) is microsatellite instability (MSI) (1).

To generate robust CC subtypes, Marisa et al analyzed genome wide mRNA expression on a large cohort of well characterized patients (n=566). Using unsupervised consensus clustering (discovery set n=443), 6 molecular subtypes of CC were identified (C1 through C6) and then validated on the remaining test set (n=123) and an independent dataset of 1,058 CC cancers. In stage II to III, the C4 and C6 subtypes were independently associated with relapse-free survival.

In our analysis of the microarray data from Marisa et al, we focused on the C3 and C4 subtypes. The C3 subtype, referred to as KRASm, were enriched for KRAS mutation tumors and displayed down regulation in most signaling pathways. The C4 subtype displays a cancer stem cell (CSC) gene expression profiles (GEP) with upregulated cell communication pathways and increased metastasis and is referred to as CSC. C4 is downregulated in cell growth, death, and cycle pathways, consistent with its CSC phenotype. Both C3 and C4 are associated with the CpG island methylator phenotype (CIMP+) and BRAF-mutant-like tumors with a serrated CC phenotype (1).

We first normalize and batch correct expression data downloaded from the original paper following the author's original workflow. We then use Principal Component Analysis for outlier identification as well as demonstrating a separation of the two tumor subtypes after normalization. Using the normalized expression data, we then perform probe filtering and agglomerative hierarchical clustering to show the differential clustering of the C3 and C4 subtypes. Although one C4 sample switched to clustering with the C3 subgroup, our analysis demonstrates the robust nature of Marisa et al's original subtyping scheme for the C3 and C4.

Methods

Data Collection

The analytical technique used to acquire patient gene expression profile (GEP) data from tumor samples was a microarray technique, specifically through the use of Affymetrix U133 Plus 2.0 chips. Given the paper chose to reference a separate publication for their wet lab steps, the summary for the other study's analytical methodology is as follows (6):

1. Tumor samples were powdered under liquid nitrogen and aliquoted
2. RNA were extracted using RNAble, followed by a clean-up step on RNAesy columns
3. These were analyzed by electrophoresis and quantified using a Bioanalyzer 2100 and NanoDrop ND-1000 respectively.
4. RNA, now aliquoted, was used with cRNA in a 3:10 ratio per hybridization step.
5. RNA was amplified and labeled, then hybridized to Affymetrix U133 Plus 2.0 chips
6. Signals were scanned using a GCOS 1.4.
7. Signal data is converted into a CEL file proprietary format (with instrumental QC criteria baked into the output) for downstream analysis

Prior to filtering by cancer subtype, etc. the original cohort of tumor samples chosen consisted of n=750 patients collected from seven clinical sites around France whose surgeries were conducted between 1987 to 2007. Only stage II and stage III patients were included for further study due to prognostic need. Because stage I patients have excellent prognoses, and almost all stage IV patients die from cancer, the stage II and stage III patients are the only cohorts that can benefit from the information gleaned from this study. Of these, only 566 samples passed the preliminary quality control standards used to monitor the analytical technique which acquired patient gene expression data.

Sample QC

Downloaded CEL files were read together using the ReadAffy (affy (version 1.76.0)) to make a AffyBatch object. To assess sample quality the normalized unscaled standard error (NUSE) and relative log expression (RLE) were calculated. For both functions, the raw AffyBatch data was converted to a PLMseq object using fitPLM (affyPLM package (version 1.74.2)) with quantile normalization and background correction using Robust Multichip Average (RMA).

By standardizing the median standard error (SE) for each sample around 1, NUSE boxplots help visualize the distribution of SE. Samples with a median SE above 1 or with a large interquartile range can indicate lower quality. NUSE statistics and histograms were generated using the NUSE function (xps package (version 1.32.0)) (Figure 1a). The mean SE was 1.003691 with a range of (0.9817812, 1.0626528) indicating low standard error across samples (Figure 1a).

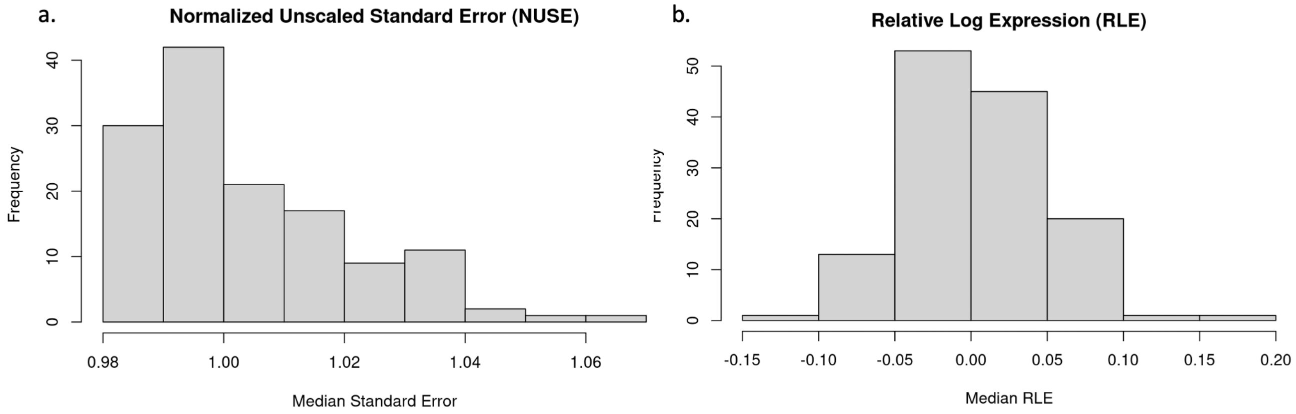

Figure 1 a) Histogram of NUSE statistics for normalized CEL files. All samples pass with low standard error. b) Histogram with median RLE for normalized CEL files. All samples pass with median RLE around 0.

RLE is used to detect unwanted variation in data caused by batch effects and can indicate if previous normalization methods have been successful. RLE is calculated by subtracting an artificial median expression level across all chips from each probe for a given gene. As it is assumed that most genes are not differentially expressed during an experiment, the median RLE should be centered at 0. RLE for each sample was calculated using the RLE function and plotted (Figure 1b). As the samples had a median standard error with a mean of 0.002654726 and range (-0.1005867, 0.1554819), we can conclude that there is limited unwanted variation between samples (Figure 1b). As all samples pass initial quality control metrics and are used for further processing.

Normalization and Batch Correction

To normalize all CEL files together, the rma function (affy package (version 1.76.0)) was run on the raw AffyBatch data to compute the RMA. Then ComBat (sva package (version 3.46.0)) was used for batch correction. First, the normalized expression data was stored as an eSet using exprs (Biobase (version 2.32.0)). In batch correction we are interested in adjustment variables (i.e. batch effects) and variables of interest. Retaining these variables is important for adjusting for batch effects with an empirical Bayesian framework without losing the correlation between variables of interest and gene expression.

Metadata for batch correction was obtained from Marisa et al. Metrics used were Center and RNA Extraction Method, while variables of interest were Mismatch repair status (MMR) and tumor subtype. MMR describes the mutations in genes involved in proofreading during DNA replication (4). A panel of five different microsatellite loci from the Bethesda reference panel were used to determine MMR status. Tumors were characterized as deficient MMR (dMMR) if two or more markers showed instability and proficient MMR (pMMR) if one or less showed instability (1). Tumor subtype was determined through unsupervised probe set selection and then consensus unsupervised class discovery. First, probes used for determination were selected if they were expressed in at least 55 of the samples, had significantly different variance compared to the mean variance of all probe sets, and had a high robust coefficient of variation. Chosen probes were then used with the R package ConsensusClusterPlus (5) using hierarchical clustering with (1-Person correlation) distance and Ward Linkage and 1000 resampling with 90% of probes used each time. k=2 through 8 were tested and k=6 clusters was chosen based on Cumulative distribution (CDF) (1)

The metadata provided was used to make a model with the variables of interest using the model.matrix function (stats (version 3.6.2)). Using this model and the batch effects recorded in the metadata, ComBat was run on the expression data. The normalized and batch corrected output used in all future analysis. Pre-processing was done with R 4.2.1 on the Boston University Shared Computing Cluster (3 hours run time, 1 core, 0gpus).

Principal Component Analysis

Principal Component Analysis (PCA) reduces dimensionality in data and helps identify the “direction of variance”. The basic workflow for PCA is first to standardize the variance between genes to prevent bias towards those with larger ranges. Next, to test for relationships between genes, a covariance matrix is made to store how gene expression varies from the mean with respect to one another. Finally, eigenvectors and eigenvalues are computed from the covariance matrix to find the principal components (PC), or the direction of data that explain the maximum amount of variance in the data.

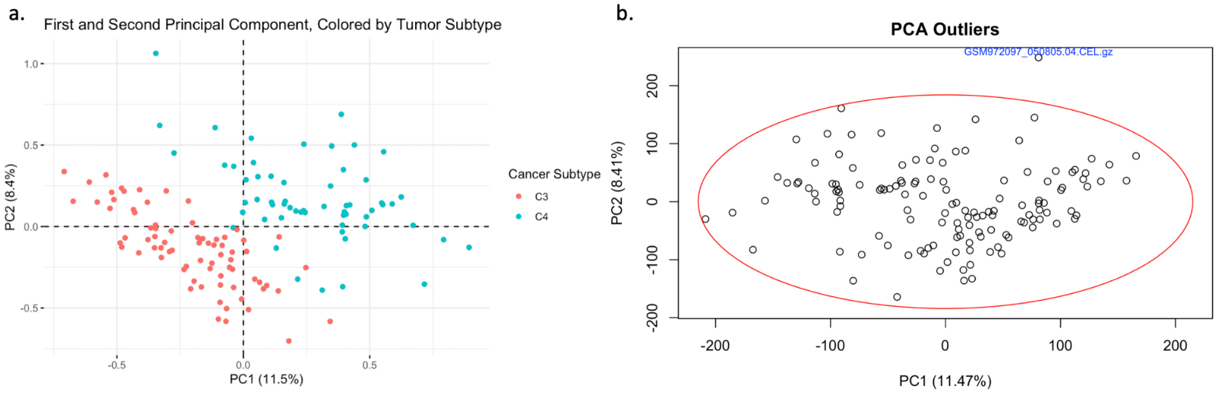

Figure 2 a) First and second principal components, colored by tumor subtype. b) PCA outliers determined using Mahalanobis distances. One outlier was isolated (GSM972097).

To complete PCA, the normalized and batch corrected data was standardized across genes using the scale function (base R 4.2.1). The principal components were calculated using prcomp (stats (version 3.6.2)). The first two principal components were plotted using fvis_pca_var (factoextra (version 1.0.7)), and colored by tumor subtype from the metadata. Outliers were identified using pca.outlier (mt (version 2.0.1.19)). using the Mahalanobis distances of PC1 and PC2 with a confidence level of 97.5%. Mahalanobis distances measure the distance between a point and a distribution, allowing for multivariate anomaly identification. There was one outlier, GSM972097, which was removed from future analysis. GSM972097 had a median NUSE of 1.03756 with an IQR of 0.05880782, and RLE of 0.0822556 with IQR of 0.7013769.

Noise Filtering & Dimensionality Reduction

Based on 3 variability metrics suggested in Marisa et.al, we used similar selection methods to remove noise and reduce the dimensionality of data to prepare for subtype prediction.First,, genes were selected to have at least log2(15) expressed reads for 20% of the sample. Second, in order to remove noise, filtered genes were required to have significantly higher variance than average. By selecting p-value at 0.01, we set up the confidence interval threshold through the chi square comparison. We also introduced the coefficient of variation, which should be greater than 0.186. The cut-off point was defined using Gaussian mixture model clustering approach, stated in Marisa et. al (1). We eliminated the highest and lowest expression value across the samples for a third time using this parameter.

Hierarchical Clustering & Subtype Discovery

We first performed hierarchical clustering into 2 sets of data, then we cut the dendrogram produced by the data into C3 and C4 subtype clusters. We created heatmaps for all samples for visualization in conjunction with the dendrogram (Figure 3). A Welch t-test is also used to identify the differentially expressed genes. For another filter of adjusted p-value < 0.05, differentially expressed genes are calculated across both subtypes and the five most differentially expressed genes were picked. A table consisting of probeset ID, t-statistic, p-value, and adjusted p-value was created (Table 1). Number of differentially expressed genes was thus calculated using adjusted p-value as the parameter.

Results

Data Processing and Principal Component Analysis

All samples passed NUSE and RLE metrics, with low standard error and RLE centered at zero. There was clear separation of the C3 and C4 subtype after PCA across both the first and second principal component. This indicates that our normalization and batch correction methods for expression data did not result in loss of sample resolution. Further, the PCA highlighted the existence of an unclustered sample GSM972097, indicating perhaps the existence of an outlier worthy of removal.

Noise filtering & dimensionality reduction

The original 54,975 were subject to various filtering criteria before biological analysis. This was done by first filtering out probes which expressed less than a log2(15) cutoff threshold among 20% of samples. Next a chi-square test was performed to filter out probes whose variance varied significantly from the median probe variance ($p<0.01$). Finally, probes which had a coefficient of variation larger than 0.183 were also filtered.

Hierarchical Clustering & Subtype Discovery

Hierarchical clustering of the remaining 1614 among 133 samples (after outlier removal) revealed a total of 57 samples within the C3 subtype and 76 samples within the C4 subtype, as indicated in Figure 3.

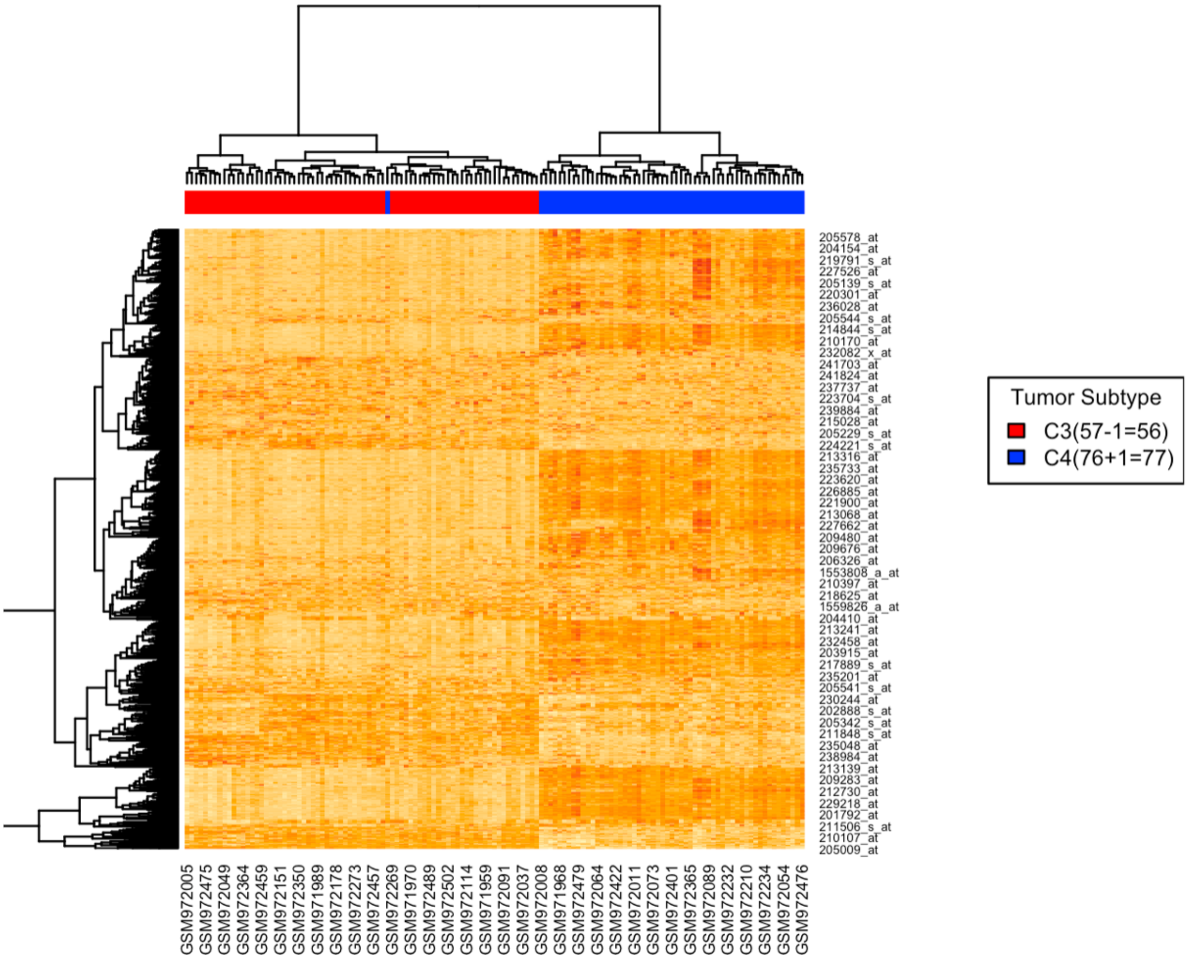

Figure 3 Heatmap of gene expression across all samples of in conjunction with the dendrogram shows clustering between the C3 and C4 subtypes with the exception of one C4 sample.

One accession GSM972269 which was classified among the C4 subtype ended up clustering among C3 samples. A Welch T-test was performed which identified 1213 differentially expressed genes ($p<0.05$). The top five of these are illustrated in the table below (Table 1).

probeset_ID	t_statistic	p_value	adjusted_p_value
204457_s_at	24.24398	2.813018e-48 4.	4.227967e-45
223122_s_at	22.69212	3.216136e-47	2.416926e-44
209868_s_at	22.20894	3.367887e-46	1.687311e-43
226930_at	22.28632	1.544708e-45	5.804240e-43
202291_s_at	21.98959	3.523008e-45	1.059016e-42

Table 1 Describes the top five differentially expressed genes according to the Welch T-test

Discussion

Raw expression data was normalized using RMA, then the quality assessed using NUSE and RLE. All samples passed initial QC checks and so underwent batch correction using ComBat. Here RNA extraction method and Center the samples were processed at were corrected for, while retaining tumor subtype and MMR status. The final processed expression data was then used for PCA, where one outlier was identified and we visualized the separation of the C3 and C4 subtype across the first two principal components. All samples passed initial quality control metrics except one outlier by PCA, indicating high quality data for use in subtype generation.

With the exception of one C4 sample, the C3 and C4 subtypes cluster together (Figure 3). The switch could be a result of the change in probe selection and clustering between our analysis and Marisa et al. In our analysis we used a stricter cutoff for probe selection (5% of total genes vs 20% in our analysis) as well as using agglomerative hierarchical clustering instead of consensus clustering. If the C4 sample that now clusters with the C3 group was on the cusp of the C4 subtype, then the change in genes used for clustering could have resulted in the GEP resembling a C3. It is important to note that this does not indicate a switch in subtype for this sample, but instead highlights a potential weakness with clustering methods, especially consensus clustering, that are highly dependent on the sample group being sorted. However, as there was only one sample that switched subtype groups from the original paper, this also demonstrates the robustness of Marisa et al's subtype scheme and gene selection, as after heavy modification of the original clustering method the C3 and C4 subtypes still clustered together in general.

Conclusions

In our analysis, we demonstrate a method for normalizing and batch correction of microarray data without loss of sample definition. After normalization and batch correction, there is still clear subtype separation with PCA and agglomerative hierarchical clustering, indicating that normalization methods did not flatten gene expression profiles. We also demonstrate the robust nature of the subtyping scheme used by Marisa et al, as changes in probe selection and clustering resulted in the change of clustering of a single sample out of 133 total samples.

The analyst has a minimal understanding of R code and implementation of R script took way longer than expected. The analyst encountered a problem with ggdendrogram(), which is a powerful tool for dendrogram plotting: variables that it passes through are not compatible with the heatmap() function (8). As a result, the data analysis quality was lower than expected due to time constraints.. To make sure further projects don't counter a similar problem again with time constraints, our group decided to make a detailed time schedule for projects in the future.

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. Centers for Disease Control and Prevention. (2022, February 17). *Basic information about colorectal cancer*. Centers for Disease Control and Prevention. Retrieved February 17, 2023, from https://www.cdc.gov/cancer/colorectal/basic_info/
3. *Colorectal cancer early detection, diagnosis, and staging*. (n.d.). Retrieved February 17, 2023, from https://www.cancer.org/content/dam/CRC/PDF/Public/8606.00.pdf
4. *NCI Dictionary of Cancer terms*. National Cancer Institute. (n.d.). Retrieved February 17, 2023, from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/mismatch-repair-deficiency
5. Matt Wilkerson (2011). ConsensusClusterPlus: ConsensusClusterPlus. R package version 1.6.0.
6. Lièvre, A., Bachet, J.-B., Boige, V., Cayre, A., Le Corre, D., Buc, E., Ychou, M., Bouché, O., Landi, B., Louvet, C., André, T., Bibeau, F., Diebold, M.-D., Rougier, P., Ducreux, M., Tomasic, G., Emile, J.-F., Penault-Llorca, F., &amp; Laurent-Puig, P. (2008). kras mutations as an independent prognostic factor in patients with advanced colorectal cancer treated with Cetuximab. Journal of Clinical Oncology, 26(3), 374–379.
7. *KRAS mutations as an independent prognostic factor in patients with ...* (n.d.). Retrieved February 17, 2023, from https://ascopubs.org/doi/10.1200/JCO.2007.12.5906
8. *Dendrograms*. Dendrograms in ggplot2. (n.d.). Retrieved February 17, 2023, from https://plotly.com/ggplot2/dendrogram/