Project 1: Microarray Based Tumor Classification Replication Analysis of Marisa et. al 1

Data Curator: Zeyuan Cao
Programmer: Nicholas Mosca
Analyst: Cory Williams
Biologist:Caroline Muriithi
TA: Dakota

Introduction:

Colorectal Cancer(CRC) is considered to be one of the most prevalent cancers in the world: third most diagnosed in both men and women each year in the United States and leading cause of death worldwide. Pathological staging of the disease has in the past failed to precisely predict the recurrence. Some of the causal factors established over the years pertain to CRC being a heterogenous disease that is caused by different genes and alleles rather than a homogenous entity. In the early stage of CRC, microsatellite instability(MSI) is the only reproducible molecular marker to have a significant prognostic factor[4]. Additionally, no clinical significance has been established by using microarray gene expression profiles (GEP) to study and to establish clinical biomarkers for CRC [3]. However, GEP studies that exploit use of high-throughput technology and unsupervised clustering in recent years have successfully identified at least 3 distinct stages of colorectal cancer. These studies still need refining and augmentation to establish standard and reproducible classification of colorectal cancer [4].

Purpose and Methodology:

The major goal of the study is to successfully create a robust and reproducible molecular classification of CRC. This would contribute to the establishment of a robust standard molecular subtyping of CRC, in order to diagnose and predict the prognosis in CRC patients that go through curative surgery for localized CRC [1]. The study was performed with genome-wide mRNA expression profiling and clustering, CRC molecular classification with clinical findings, DNA alterations and prognosis were assessed for correlations. 750 Samples that ranged from Stage I to Stage IV CC were used in this study and unsupervised classification and hierarchical clustering was performed on the samples. 566 of the samples fulfilled RNA quality requirements. Out of the 566 samples, 433 CC samples were used for genome-wide mRNA expression analysis. The classification was then further done on the remaining 123 samples(from the 750 batch) and an additional independent set of 1,058 CC samples from eight public datasets. Classification was further validated on the remaining 123 CC samples and independent set of 1,058 CC samples from eight public datasets[1].

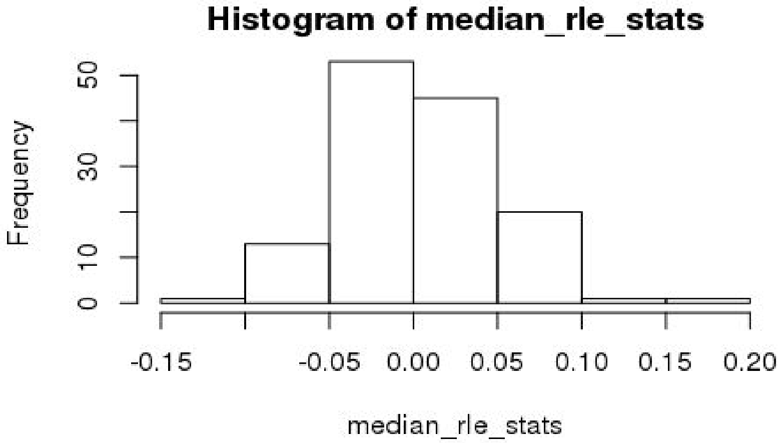
Data:

Human colon cancer cells were taken as samples. The gene expression profiles (GEP) were analyzed using Affymetrix U133 Plus 2.0 chips. RNA molecules were extracted using RNAble followed by RNAeasy columns. Electrophoresis and Nanodrop were used to assess RNA quality. RNA molecules were then reverse-transcribed into cDNA using RT-qPCR. A 28s/18s ratio of above 1.8 was used as an indication of high RNA quality[2]. 750 Samples that ranged from Stage I to Stage IV CRC were used in this study and unsupervised classification and hierarchical clustering was performed on the samples. 566 of the samples fulfilled RNA quality requirements. Out of the 566 samples, 433 CRC samples were used for genome-wide mRNA expression analysis. The classification done was then further on the remaining 123 samples(from the 750 batch) and an additional independent set of 1,058 CC samples from eight public datasets. Classification was further validated on the remaining 123 CC samples and independent set of 1,058 CC samples from eight public datasets. [1]All CEL files containing raw data were downloaded from the Gene Expression Omnibus website at https://www.ncbi.nlm.nih.gov/geo/ under the accession number GSE39582. KEGG, GO and Hallmark gene sets with .gmt extensions were downloaded from MSigDB

Methods:

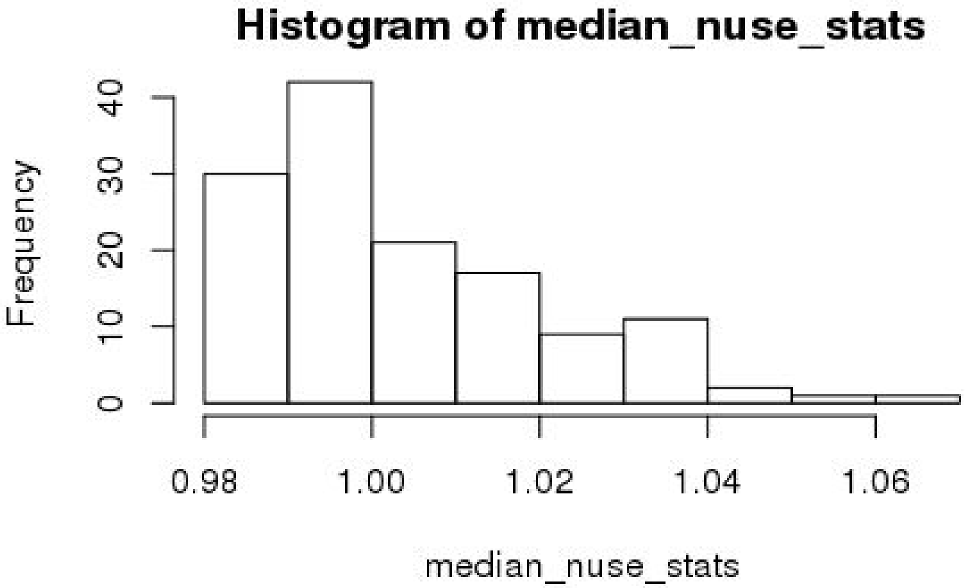
All of the following methods are a product of Bioconductor open source software. All of the patient samples were gathered and organized using ReadAffy. The intensity values of each CEL file were normalized via the quantile normalization method. This method was applied by the fitPLM function. Outliers were removed by converting AffyBatch files into an expression measurement[4]. The (robust multi-array) function removes unnecessary expression values and summarizes the data into a large expression set . Batch effects were corrected using the empirical Bayes method (ComBat) applied by the SVA R package[5].

Applying the following methods is an effort to perform quality control on the Affymetrix Oligonucleotide Array samples(n=134). After data normalization, Relative Log Expression and Normalized Unscaled Standard Error were computed on the patient sample dataset. Median Relative Log Expression was calculated and plotted below:

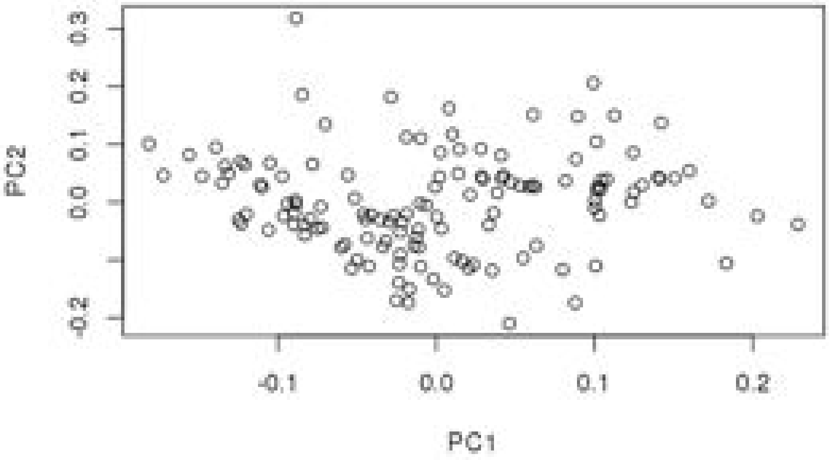


Majority of the patient samples displayed a RLE score between -0.05 and 0.05. This can be interrupted as most of the gene intensity values are between similar genes across patients. Values closer to zero represent differential expression in similar genes among all patients. The patients in this study were all associated with Colon cancer. I would expect a RLE score close to zero because we have a large overlap of expressed genes. Similar genes are expressed in multiple samples.

Normalized unscaled standard error (NUSE) determines the standard error from the probe- level. Larger NUSE scores can be correlated with poor quality samples. Median NUSE scores of usable quality should be centered around 1. Below is a histogram displaying the median NUSE scores for the 134 patient samples.



Principal component analysis was computed below using prcomp function on transposed robust multi array data after normalization.



Based on the relationship between PCA1 and PCA2 no clear direction of variance was Identified . This could be a result of a upstream formatting error with the normalized data. If correct I would expect a slight difference between data before and after quality control.

All computational analysis was conducted on the Boston University Shared Computing Cluster. Multiple 12 interactive sessions in RStudio IDE were used to accurately generate a normalized dataset. Boston University's RCS consultants were also used as a resource.

Results:

Through normalization of the data, one major CSV was produced and had to be filtered through to obtain relevant genes of interest that met the qualifications of multiple significance filters. The CSV file was formatted where the header rows contained the gene names while the header columns contained the sample names, and the intensity values filled up the remaining spots for each gene crossed by the sample name. The first filters were for each gene; at least 20% of the gene expression values across multiple patients (sample name) must be higher than log2(15). The first filter was done by using a simple R script that looped through each row of genes and saved all of the genes that had a value higher than log2(15) to a separate CSV file; this was followed by obtaining the p-value of the first filtered samples. For the p-value, an R script that loops through each row and produces the p-value and adjusted p-value along with the gene names (row names) associated was used and produced a separate CSV file. This newly made CSV file was then looped through using a similar script as the first filter to obtain the p-values that were significant (less than 0.01) and was saved to an additional separate CSV file. With these separate CSV files containing the p-values that were significant only containing gene names and not crossed to the sample names, a cross-reference had to be performed. This cross-reference was needed in order to match and extract the significant genes with the sample names from the CSV used for the initial p-value test. This was saved to an additional separate CSV file. Finally the genes had to have a coefficient greater than 0.186 which was determined using a similar script to the one used in the first filter. After the first filter, the relevant genes were cut almost in half to about 30,000 from the original 50,000 genes obtained. The significant p-value filter brought the 30,00 thousand genes to about the mid 20,00 while the last filter lowered the amount of genes to around the lower end of 20,000.

A significant issue with utilizing the necessary data was in the area of using R to convert the files into the correct format, which is explained in the challenges section of this article. Multiple packages are needed to complete the various filters needed before analysis, and not all of the packages work correctly with CSV files. A fix however for the issue would be an alternative to creating multiple CSV files. It seems it would be best to obtain the affybatch file directly after normalization and convert that into an expression set, which is more friendly with the package limma in R that was designed to analyze microarray datasets thoroughly. The expression set not only works well with the limma package but also with a majority of the base bioclite packages. Expression set files can also comfortably house all of the data not directly related to the current analysis being performed, such as the date the sample was obtained. This is beneficial because it reduces the amount of CSV files needed and produced. Due to issues with the produced files, a formal heatmap of the filtered was not produced; however, a very large heat map was produced with the normalized CSV file after being converted to an expression set for easy use, and this produced irrelevant information; however, it was done to show the ease of files being in an expression set and proof of using the proper packages and codes.

Discussion:

Welch's t-test

Welch's t-test was performed to identify the differentially expressed genes among the clusters. The t-test between clusters was performed and a file containing the p-value, adjusted p-value, t statistic and probeset ID was generated. The lower the adjusted p-value, the more statistically significant the observation. All this was done in the differential expression csv file provided. Data from parts 4.5 and 5.6 was not available on time so as to do the analysis in a timely manner. The additional gene sets used for the biologist role were obtained from The Molecular Signatures Database (MSigDB): KEGG, GO and Hallmark gene sets. The files downloaded specifically did not have entrezid. GSEABase bioconductor package was used to load the genesets after they were downloaded as a gene set collection. Upon analysis of the .gmt files, I found the KEGG gene collection ("c2.cp.kegg.v7.0.symbols.gmt") to have 186 genes sets. The

GO (("c5.all.v7.0.symbols.gmt") gene collection had 9996 genes sets and the Hallmark gene(("h.all.v7.0.symbols.gmt") sets had 50 gene sets.

Description and relevance of Enriched Gene Sets(KEGG, GO and Hallmark collection)

The additional gene sets used for the biologist role were obtained from The Molecular Signatures Database MSigDB: KEGG, GO and Hallmark gene sets. The files downloaded specifically did not have entrezid. The Kyoto Encyclopedia of Genes and Genomics (KEGG) gene set collection is a collection of manually drawn pathways that shows the molecular interaction, reaction and relation networks for pathways listed below [6]:

1. Organism Systems
2. Metabolism: Carbohydrate, Energy, Lipid, Nucleotides, Amino Acids, Glycan, Cofactor Vitamin, Terpenoid/PK, Xenobiotics.
3. Drug Development
4. Cellular Processes

Environmental Information Processes

5.Organismal Systems

6.Human Diseases

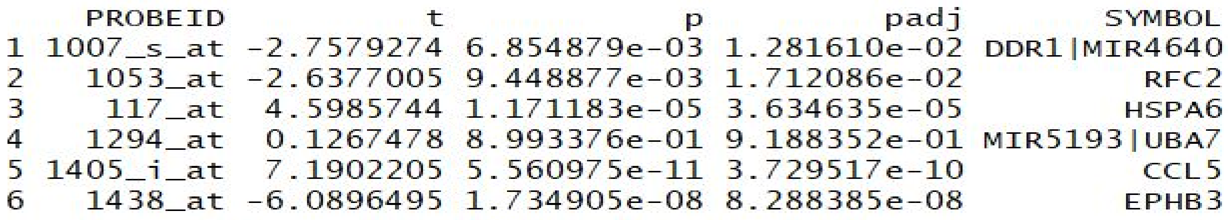
7. Genetic Processes

The Hallmark gene set collection

Each hallmark in this gene set collection is made up of a refined gene set. The gene sets are obtained from multiple founder sets, that conveys a specific biological state or process and displays coherent expression. The collection has genes involved in epithelial mesenchymal transition, metastasis, UV response genes. [6]

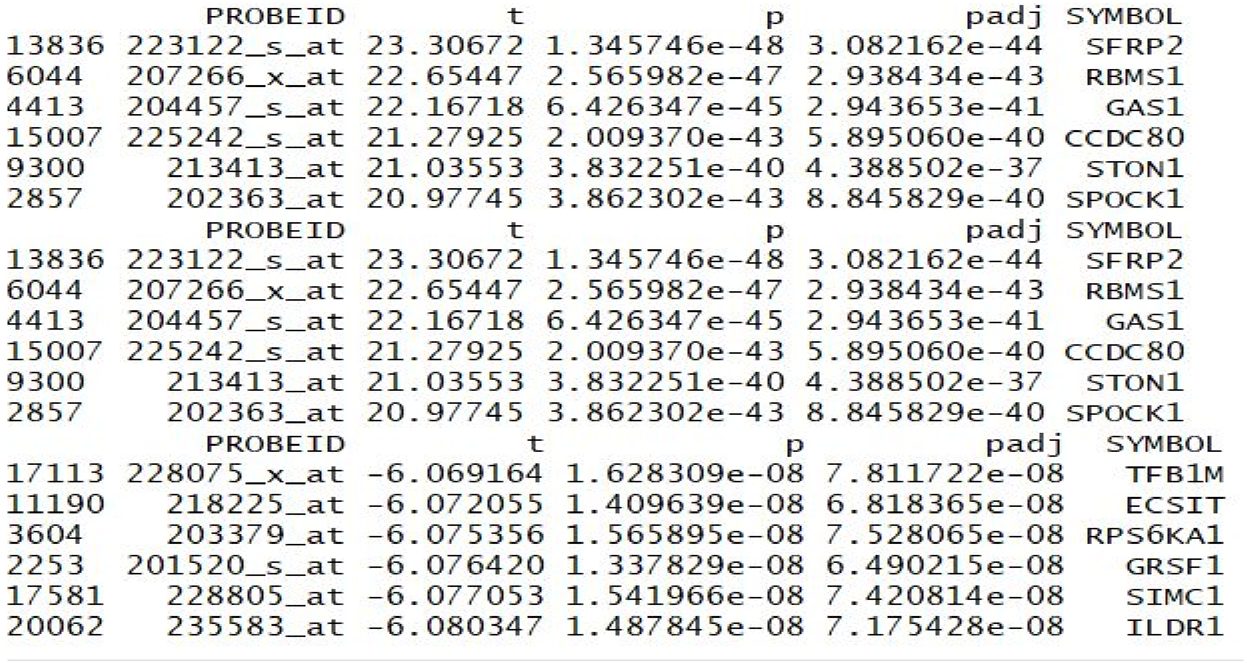
GO gene set collection

The GO gene set collection that we used has gene sets that are annotated using the same GO term.



```
    PROBEID          t            p         padj         SYMBOL
1 1007_s_at -2.7579274 6.854879e-03 1.281610e-02 DDR1|MIR4640
2   1053_at -2.6377005 9.448877e-03 1.712086e-02         RFC2
3    117_at  4.5985744 1.171183e-05 3.634635e-05        HSPA6
4   1294_at  0.1267478 8.993376e-01 9.188352e-01 MIR5193|UBA7
5 1405_i_at  7.1902205 5.560975e-11 3.729517e-10         CCL5
6   1438_at -6.0896495 1.734905e-08 8.288385e-08        EPHB3
```

Table 1: The table shows the proboset ID mapped to the gene symbols. An additional column that shows the SYMBOL was added to the differential expression csv.



```
            PROBEID        t            p         padj SYMBOL
13836 223122_s_at 23.30672 1.345746e-48 3.082162e-44  SFRP2
6044  207266_x_at 22.65447 2.565982e-47 2.938434e-43  RBMS1
4413  204457_s_at 22.16718 6.426347e-45 2.943653e-41   GAS1
15007 225242_s_at 21.27925 2.009370e-43 5.895060e-40 CCDC80
9300    213413_at 21.03553 3.832251e-40 4.388502e-37  STON1
2857    202363_at 20.97745 3.862302e-43 8.845829e-40 SPOCK1
            PROBEID        t            p         padj SYMBOL
13836 223122_s_at 23.30672 1.345746e-48 3.082162e-44  SFRP2
6044  207266_x_at 22.65447 2.565982e-47 2.938434e-43  RBMS1
4413  204457_s_at 22.16718 6.426347e-45 2.943653e-41   GAS1
15007 225242_s_at 21.27925 2.009370e-43 5.895060e-40 CCDC80
9300    213413_at 21.03553 3.832251e-40 4.388502e-37  STON1
2857    202363_at 20.97745 3.862302e-43 8.845829e-40 SPOCK1
            PROBEID         t            p         padj  SYMBOL
17113 228075_x_at -6.069164 1.628309e-08 7.811722e-08   TFB1M
11190   218225_at -6.072055 1.409639e-08 6.818365e-08   ECSIT
3604    203379_at -6.075356 1.565895e-08 7.528065e-08 RPS6KA1
2253  201520_s_at -6.076420 1.337829e-08 6.490215e-08   GRSF1
17581   228805_at -6.077053 1.541966e-08 7.420814e-08   SIMC1
20062   235583_at -6.080347 1.487845e-08 7.175428e-08   ILDR1
```

Table 2: Show the enriched gene sets for each gene set type. The top significantly enriched gene sets in the GO collection, KEGG Collection and Hallmark Collection. The data used was from the differential expression csv provided(precomputed data)

Conclusion:

The gene expression(GEP) data that was used for colon cancer was obtained from GEO and NCBI databases. fitPLM function was used to normalize raw data. Additionally, RMA background correction was done. RLE and NUSE were carried out as a part of data quality control and histograms from both these methods were generated**.** Outliers were then detected and removed using Principal component analysis(PCA) and batch effects in the data were removed using ComBat. Genes with similar functions were grouped by hierarchical clustering and the dendrogram was cut into 2, yielding 2 clusters with 57 and 77 samples respectively. The gene expression across different samples was studied by generating a heatmap with colorbars. Differentially expressed genes were determined using the Welch's t-test.

Challenges: Programmer

Specific challenges for this project were centered around generating a script to actually qualify and normalize the cell files. One major problem was learning how to use R as a tool.

Not having any filmarity with how to use R or even do simple tasks was a challenge within itself. The second was manipulating data in the correct way and in the correct format. A large portion of this project was performing functions that act as quality control for our dataset, understanding if the output of these functions was correct was also a major challenge. Once the proper functions were conducted and converted to csv files, the format was not allowing further analysis. Not understanding how to check results or having reference to check if each function was used correctly was a major issue. Throughout troubleshooting and building the script a lot was learned about how R works as a tool for computational analysis. The lack of confident data quality control resulted in analysis downstream to be performed on the sample dataset. Overall these problems were not fully solved. To solve these challenges more hands on experience with datasets would be useful. Examples of correct function results would help guide programmers through the array of quality control tasks.

Most of this project was trying to get R to work on anything. Following the instructions gave an idea of what major functions to use but being unaware of data types forced me to try and change the data to fit arguments. Many of my functions would work but produce a blank or null result. It would have been great to see what results for each step should look like rather than moving forward and hoping it worked just because no errors were displayed.

Challenges: Data Curator

The biggest challenge so far is to always be sure that correct directory is used so that files are created or uploaded to the wrong place. Manipulating the files on the command line can be frustrating since sometimes there is no visual feedback. The problem was solved by using FileZilla to work with the files rather than directly on the command line. This is immensely helpful since it provides an intuitive graphical interface which is much easier to work with than the black screen of command line.

Challenges: Analyst

Some of the challenges that were faced when recreating the analysis is the way the data is formatted after normalization, with the most prevalent issue consisting of the row names not being generated after normalization. This formatting issue caused the r package for p-value analysis to produce either a completely blank table or empty values with no way of linking the values back to the original gene name due to the row names being absent. This proved to be extremely difficult to solve, and due to this, a lot of the scripts had to be tested on the sample data provided by the instructor, with the actual data that we normalized we could get it to produce the results we were expecting from the readings. A solution for this seemed to be housed in the creation of an expression set. Creating an expression set proved to be more challenging with a CSV file as opposed to the affybacth file created right after normalization, which we un, unfortunately, did not have access too after the file was converted to a CSV.

Challenges: Biologist

Some of the challenges I faced was an initial breakdown of communication(mostly not being aware because of my late start of the class) that we were supposed to begin the role with the sample data provided on SCC until Friday and not wait until the data generation of project data by the programmer and the analysis which left not so much time left to do the biologist analysis. Generating differentially expressed genes from the results proved to be a challenge because of the lack of generation of data on time by both part 4.5 and 5.6 and a resort tried to work on the differential expression csv provided. This proved to be harder to do because it was easier to write a function using data obtained from the clusters that were determined using classification to have the most upregulated and downregulated genes.

References

1. Marisa, L., Reyniès, A. D., 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).

2. 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.

3. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31: e15.

4. Popat, S., Hubner, R., & Houlston, R. (2005). Systematic Review of Microsatellite Instability and Colorectal Cancer Prognosis. Journal of Clinical Oncology,23(3), 609-618.

5. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8: 118–127.

6. Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, Jill P. Mesirov. 2011. “Molecular signatures database (MSigDB) 3.0”. *Bioinformatics* 27(12): 1739–1740

- https://support.bioconductor.org/p/67297/
- https://support.bioconductor.org/p/76067/