Microarray Based Tumor Classification

BF528 Project1
Yu Zhong, Zhiyu Zhang, Jianfeng Ke, Huisiyu Yu

Introduction

Colorectal cancer is the fourth most deadly cancer, with nearly 900,000 people dying from it each year [1]. The prognosis and treatment options for colorectal cancer depend on five pathological stages, yet pathological staging often fails to predict relapse in patients accurately. Previous gene expression profile (GEP) studies had identified gene expression profiles for predicting prognosis of CC, however due to lack of consistency and validation, no firm conclusions could be drawn from those studies to help clinical practice.

In their 2013 study, Marisa et al performed mRNA GEP analyses with Affymetrix U133 Plus 2.0 microarray, aiming to establish a comprehensive and reproducible molecular classification of Colon Cancer (CC). They also assessed possible associations between identified molecular subtypes and clinical and pathological factors, common DNA alterations, and prognosis [2].

In this study, we obtained a subset of 134 samples of tumor tissue microarray data with two molecular subtypes from the original study by Marisa et al, and performed normalization, quality inspection, noise filtering, hierarchical clustering, as well as enrichment analysis on these samples.

Methods

Normalization, quality inspection, batch effect correction and principal component analysis were performed through the following Bioconductor (3.1.2) packages: affy [3], affyPLM [4], sva [5], AnnotationDbi [6], and hgu133plus2.db [7] in RStudio version 4.0.2 [8].

Data

Gene Expression Profiles (GEPs) of different human colon cancer subtypes determined on Affymetrix chips using microarray technologies were obtained from the NCBI Gene Expression Omnibus (Accession Number: GSE39582). Only C3 (75 samples) and C4 subtypes (59 samples) were involved in our study. We collected all GEPs with CEL format from above that can be served as raw datasets for the downstream analyses. No error or contamination can be detected in this case.

Normalization and Quality Control

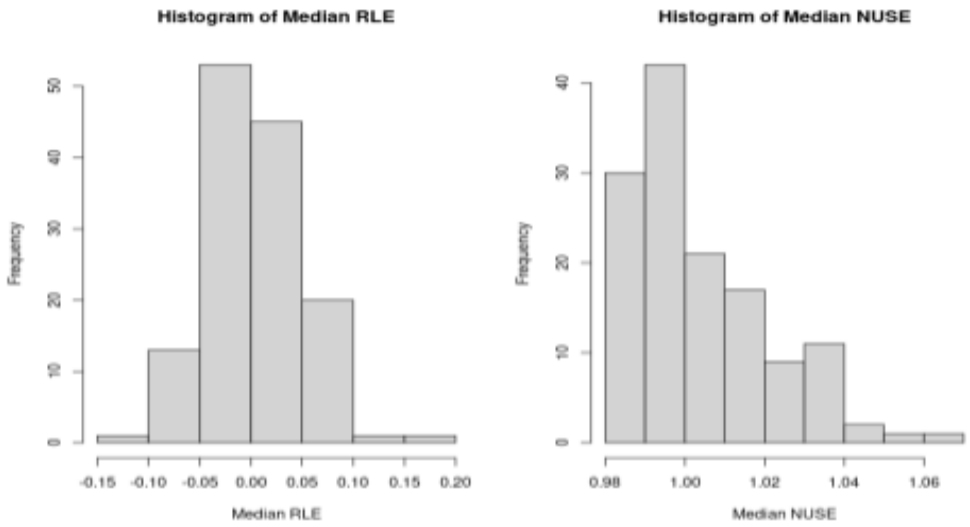

Figure 1. Histograms of medina RLE and median NUSE
Median NUSE and median RLE score calculated for 134 samples

Probe level data were normalized via the Robust Multiarray Averaging (RMA) algorithm using the rma function in package Affy, and then fitted with a robust linear model using the fitPLM [9] function of affyPLM package with quantile normalization and background correction enabled as additional arguments.

Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) scores were computed using the RLE and NUSE [10] functions of the affyPLM package, in order to detect unwanted variation and to eliminate problematic arrays. Median RLE and NUSE scores were used to filter out 4 samples of poor-quality data, with thresholds being median RLE < 0.10 and median NUSE < 1.05, as well as plotted in histograms (Figure 1) for the purpose of summarization and visual inspection.

Batch Effect

Normalized probeset data were converted to expression values by the exprs function, and batch effects were corrected using the ComBat function of the sva package. Two parameters provided with this function were extracted from an annotation file by the original authors: batch effects which include both Center and RNA extraction method, and features of interest which include both tumor and MMR status.

Principal Component Analysis

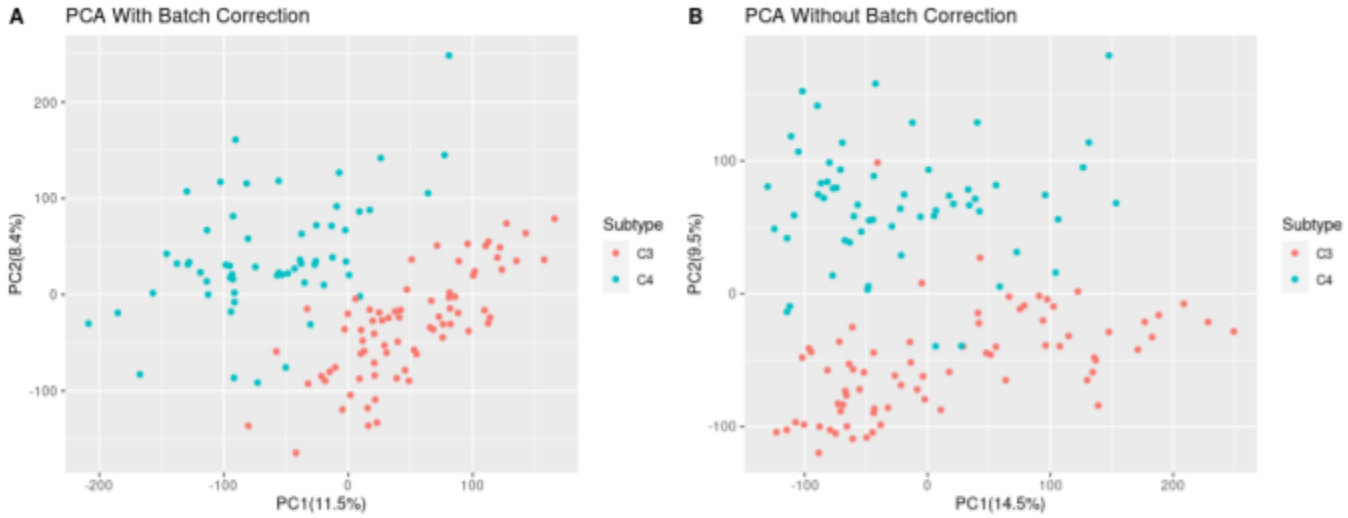

Figure 2. Principal Component Analysis With and Without Batch Correction
PCA with filtering and batch correction (A) data vs PCA without filtering and batch correction (B). Blue points were classified as subtype C4, red points were classified as subtype C3 according to Marisa et al.

Dimensionality reduction was achieved by performing Principal Component Analysis (PCA) on normalized, filtered and batch-corrected expression data. The expression matrix was centered and scaled within each gene using the scale function, after which the prcomp function was used to perform PCA. PC1 vs PC2 were plotted using ggplot [11] and data points were colored according to subtypes (C3 and C4).

Analyses above took 6 minutes and 22 seconds to run on the Interactive RStudio Server on the Shared Computing Cluster (SCC) of Boston University using 1 CPU core.

Results

There were 39522 genes expressed in at least 20% of samples. 22870 genes had a variance significantly different from the median variance of all probe sets under a threshold of $p<0.01$. Only 1731 probes had a coefficient of variation greater than 0.186. In total, 1544 probes passed all three of the filters. Hierarchical clustering was performed upon our fully filtered data matrix. 130 patients were divided into two clusters, 56 samples in one and 74 in the other (Figure 3.).

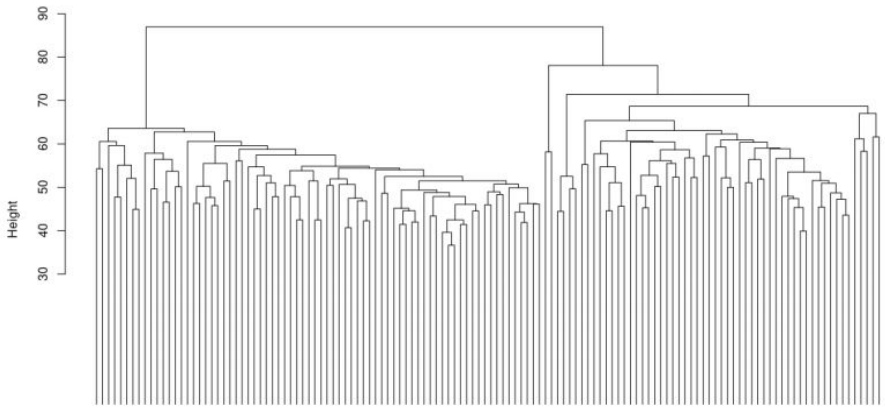

Figure 3. Hierarchical clustering of patients
Cluster Dendrogram of the 130 tumor sample GEPs obtained from Hierarchical clustering. Cluster 1 (left) has 56 and cluster 2 (right) has 74 GEPs.

A heatmap was created to visualize differential expression of each gene across all 130 samples (Figure 4.). Samples with C3 subtype were colored as red while C4 subtype were colored blue.

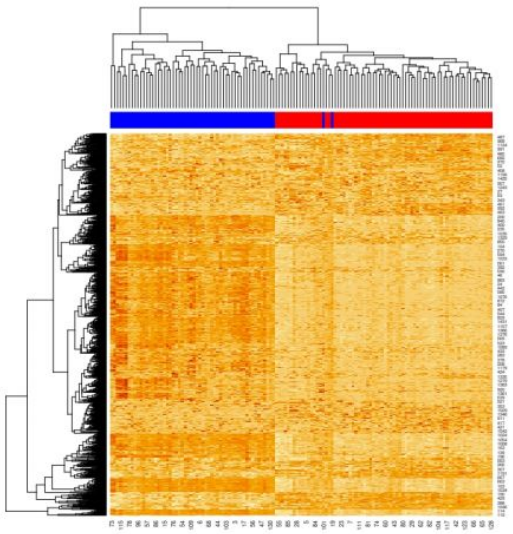

Figure 4. Heatmap of Gene Expression
DEG heatmap of the 130 tumor samples (x-axis) against 1544 probe sets (y-axis). Red refers to C3 subtype samples, and blue refers to C4.

It was evident from the heatmap that samples were divided into two clusters based on subtypes. Welch t-test was then applied to the two clusters on each probe. There were in total 1257 genes differentially expressed at adjusted $p<0.05$ (C4 as the reference).

Table 1. Top ten Up-Regulated and Down-Regulated Probesets

ProbeID	Statistic	P_value	P_adj	Symbol
Up-regulated				
204457_s_at	23.78899	1.02E-47	2.33E-43	GAS1
203748_x_at	22.67979	1.87E-46	1.95E-42	RBMS1
225242_s_at	22.56048	1.39E-45	6.35E-42	CCDC80
207266_x_at	22.5185	2.55E-46	1.95E-42	RBMS1
209868_s_at	22.33961	1.12E-45	6.35E-42	RBMS1
218694_at	22.21934	5.60E-44	1.42E-40	ARMCX1
227059_at	22.12753	2.36E-44	6.75E-41	GPC6
223122_s_at	21.90289	4.35E-45	1.66E-41	SFRP2
202291_s_at	21.65315	1.92E-44	6.26E-41	MGP
223121_s_at	21.57428	1.06E-40	1.01E-37	SFRP2
Down-regulated				
203240_at	-16.05877	1.28E-28	8.37E-27	FCGBP
235350_at	-13.81086	1.19E-24	4.51E-23	C4orf19
220622_at	-13.54217	2.84E-26	1.33E-24	LRRC31
227725_at	-13.47097	5.39E-23	1.69E-21	ST6GALNAC1
210107_at	-13.26033	5.90E-25	2.34E-23	CLCA1
204673_at	-13.16611	7.46E-24	2.65E-22	MUC2
228463_at	-12.74869	8.74E-23	2.68E-21	FOXA3
1553828_at	-12.72928	4.21E-24	1.52E-22	NXPE1
219450_at	-12.66731	1.13E-21	3.00E-20	C4orf19
236513_at	-12.66231	5.23E-24	1.88E-22	PRELID2

We then recalculated the t-statistics of differential expression for probe sets with a variance significantly different from the median variance of all probes (Table 1.) Provided with GO, KEGG and Hallmark annotation collections having 10271, 186 and 50 genesets respectively, gene enrichment analysis was performed on each geneset using fisher test, resulting in 1229 GO terms (Up-regulated: 385; Down-regulated: 844), 30 KEGG terms (Up-regulated: 22; Down-regulated: 8) and 33 Hallmark terms (Up-regulated: 16; Down-regulated: 17). We selected the top three enriched terms for each gene set type in order to acquire a global view on the GEPs (Table 2.).

Table 2. Top three Enriched Gene Sets For Each Gene Set Type

Term	Odds_ratio	P_value	P_adj
GO: Up-regulated			
RNA_binding	0.29648966	2.48E-19	2.55E-15
RNA_processing	0.18872766	1.20E-15	6.14E-12
Cell_cycle	0.36006319	8.25E-15	2.83E-11
GO: Down_regulated			
Collagen_containing_extracellular_matrix	8.064772	5.03E-74	5.17E-70
Extracellular_matrix	7.164827	6.70E-73	3.44E-69
Extracellular_matrix_structural_constituent	12.417991	6.19E-61	2.12E-57
KEGG: Up-regulated			
Amino_sugar_and_nucleotide_sugar_metabolism	8.4116322	4.14E-08	7.71E-06
Drug_metabolism_cytochrome_P450	5.2136447	9.74E-08	9.06E-06
Glycosphingolipid_biosynthesis_lacto_and_neolacto_series	7.9691709	1.41E-06	8.73E-05
KEGG: Down-regulated			
Ecm_receptor_interaction	5.5575971	2.97E-15	5.52E-13
Focal_adhesion	3.1821083	5.23E-12	4.87E-10
Leukocyte_transendothelial_migration	2.9008141	2.82E-05	1.75E-03
Hallmark: Up-regulated			
Estrogen_response_late	2.9346118	6.80E-08	2.30E-06
E2F_targets	0.1401626	9.19E-08	2.30E-06
G2M_checkpoint	0.2175325	3.97E-06	4.96E-05
Hallmark: Down-regulated			
Epithelial_mesenchymalL_transition	9.1824068	1.40E-75	7.01E-74
Coagulation	4.8788506	1.27E-14	3.18E-13
UV_response_dn	3.7314597	3.51E-14	5.85E-13

Note: C4 subtype were selected as reference set

Discussion

Histograms of median RLE and median NUSE (Figure 1.) showed that the majority of samples were in good quality, evidence of which normalization was successful. Removing samples according to relative log expression scores rely on the assumption that sample heterogeneity detected by RLE scores was a sign of technical artifacts, instead of biological factors of interest [12].

Systematic non-biological differences, also known as batch effects, in high throughput experiments can lead to indirect comparison between samples in different batches [13]. Principal component analysis revealed that some samples separated from their cluster can be statistically adjusted by removing unwanted variation, such that two biologically different groups can be separated further. This demonstrated the effectiveness of batch effect correction in reducing non-biological noise while emphasizing features of interests underlying two different subtypes in our study (Figure 2.).

Unsupervised gene expression analysis of two subtypes provided us with insights into the distance between two clusters, as well as the association relationship among samples within a cluster (Figure 3., 4.). It is evident that the majority of samples were clustered with its group except for a few samples, which were mixed up with their counterparts. This could be interpreted as either misclassification of samples due to experimental procedures, or actual outliers induced by aberrant gene expression within the cancer cells.

Our results of gene set enrichment in the C3 subtype were mostly consistent with Marisa et al's findings (Table 2.; [2]) Compared with GEPs of the C4 subtype, most cell communication pathways as well as pathways involving motility and angiogenesis were down-regulated in the C3 subtype, whereas biological processes associated with cell growth and death, and metabolism related processes were up-regulated in the C3 subtype. The fact that pathways associated with the C3 subtype were down-regulated relative to C4 in extracellular matrix (ECM) related processes and pathways also supported the original author's findings, which states that C4 subtype tumors were more prone to metastasis [2]. Interestingly, we found that biological processes linked with cell growth and death were up-regulated in the C3 subtype, however, were not statistically significant based on the results of the original study [2], illustrating that those genes differentially expressed in the C3 (C3 versus C4), were less significant in the comparison of the C3 and normal cell.

Conclusion

In our study, we recovered the gene expression profiles of the C3 and C4 subtype in human colon cancer. Unsupervised gene expression analysis demonstrated a biological difference underlying the two molecular subtypes, supported by the differential gene expression analysis as well. Referencing the profiles of the C4 subtype, cell communication, motility and angiogenesis related pathways were down-regulated in the C3 subtype, whereas metabolism was up-regulated in the C3 subtype.

Our results were largely consistent with the original authors' findings in terms of clustering of subtypes, except that two samples classified as C4 by Marisa et al clustered with C3 in our analysis. Our enrichment analyses also confirmed part of the biological processes and pathway associations involving C3 and C4 subtypes stated in the aforementioned research.

References

1. Dekker, E., Tanis, P. J., Vleugels, J. L., Kasi, P. M., & Wallace, M. B. (2019). Colorectal cancer. *The Lancet*, *394*(10207), 1467–1480. https://doi.org/10.1016/s0140-6736(19)32319-0
2. 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). https://doi.org/10.1371/journal.pmed.1001453
3. Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, *20*(3), 307–315. https://doi.org/10.1093/bioinformatics/btg405
4. Bolstad, B. M., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R. A., & Speed, T. P. (2005). Quality Assessment of Affymetrix GeneChip Data in Bioinformatics and Computational Biology Solutions Using R and Bioconductor. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, 33–47. https://doi.org/10.1007/0-387-29362-0_3
5. Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Zhang Y, Storey JD, Torres LC (2020). sva: Surrogate Variable Analysis. R package version 3.38.0.
6. Hervé Pagès, Marc Carlson, Seth Falcon and Nianhua Li (2020). AnnotationDbi:Manipulation of SQLite-based annotations in Bioconductor. R package version1.50.3.
7. Carlson M (2016). hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2). R package version 3.2.3
8. RStudio Team (2020). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.
9. Bolstad, B. M. (2004). *Low-level analysis of high-density oligonucleotide array data: background, normalization and summarization*. University of California, Berkeley.
10. Brettschneider, J., Collin, F., Bolstad, B. M., & Speed, T. P. (2008). Quality Assessment for Short Oligonucleotide Microarray Data. *Technometrics*, *50*(3), 241–264. https://doi.org/10.1198/004017008000000334
11. Wickham, H. (2016). ggplot2. *Use R!* https://doi.org/10.1007/978-3-319-24277-4
12. Gandolfo, L. C., & Speed, T. P. (2018). RLE plots: Visualizing unwanted variation in high dimensional data. *PLOS ONE*, *13*(2). https://doi.org/10.1371/journal.pone.0191629
13. Johnson, W. E., Li, C., & Rabinovic, A. (2006). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037