Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value

-- Individual Project

https://github.com/MogicianEik/BF528_individual_project

Analyst - Yichi Zhang
Team Swiss Cheese

Noise Filtering & Dimensionality Reduction

- ### Introduction

  Microarray technique is used to study the role of genetics involved in the development of diseases in an early stage and has made an enormous contribution to explore the diverse molecular mechanisms involved in tumorigenesis, like what the reference paper did. The end product of microarray, whose quality is often degraded by noise caused due to inherent experimental variability, relies on noise reduction to obtain high intensity gene expression results and to avoid biased results. Microarray analysis is characterized by the number of features far exceeding the number of samples, and many basic methods can be used to remove this noise. In the reference paper, Marisa et al. selected genes for subtype determination and eventually reported a new classification of CC into six molecular subtypes that arise through distinct biological pathways.

- ### Methods

  Several Filters were implemented on the RMA normalized, *ComBat* adjusted expression matrix obtained from the previous programmer Yichi Zhang.

  Probe sets used for subtype determination fulfill the three following criteria:

  (1) to be expressed in at least 20% of the samples (normalized intensities across samples> log2(15))
  (2) to have a variances significantly different from the median variance of all probesets(i.e. variance test p-value<0·01).
  (3) to have a high robust coefficient of variation (rCV>0·186). rCV for each probe set was calculated by dividing the standard deviation by the mean, eliminating the highest and lowest expression value across the samples for each probe set.

- ### Results
  A total of 39661 genes passed all 3 filters.

- ### Discussion
  The simple but effective filters have reduced the number of genes from 54675 to 39661 obtained from 134 samples, but unfortunately I could not directly compare my results against the reference paper where only 1459 probe sets were used for subtype determination. The Chi-squared test and logistic regression were used to study associations between anatomoclinical features, common DNA alterations, and subtypes. To confirm the robustness of the subtypes obtained, The authors further validated their molecular classification in a large independent set.

Hierarchical Clustering & Subtype Discovery

- ### Introduction
  The aim was to discover molecular subtypes of Colon Cancer. These subtypes were associated with distinct clinicopathological characteristics, molecular alterations, specific enrichments of supervised gene expression signatures (stem cell phenotype–like, normal-like, serrated CC phenotype–like), and deregulated signaling pathways. The filtered expression data of 134 samples would be classified unsupervised. Preliminary subtypes would be assigned to each sample from the validation series in the reference paper via a centroid-based predictor using the most discriminating probe sets (over and under expressed) of each subtype. The predictions would be examined for characteristics and robustness which will not be reproduced in the current analysis.

- ### Methods
  The class discovery approach mentioned by Marisa et al is implemented in the R package *ConsensusClusterPlus*. In brief, a clustering analysis is performed ***n*** times on subsets of the probe sets and of the samples selected randomly. Then all derived partitions for a given number of clusters ***k*** are summarized by clustering the (samples x samples) co-classification matrix. To reduce the complexity of the task, only hierarchical clustering was implemented in this analysis. It was expected that samples were separated into 2 major clusters. Probe Sets Significantly Differentially expressed in samples of the given subtype compared to samples of their subtypes according to the Limma moderated t-test or the Welch's t-test (adjusted p-value < 1e-5 and |log2 fold change| > 0·5) were retained.

- ### Results

Cut the dendrogram such that the samples are divided into 2 clusters. 70 samples were attributed to cluster 1 and 64 samples were attributed to cluster 2. By selecting DE genes with adjusted p-value (FDR) < 1e-5, 6422 genes were differentially expressed between the clusters for both lists.

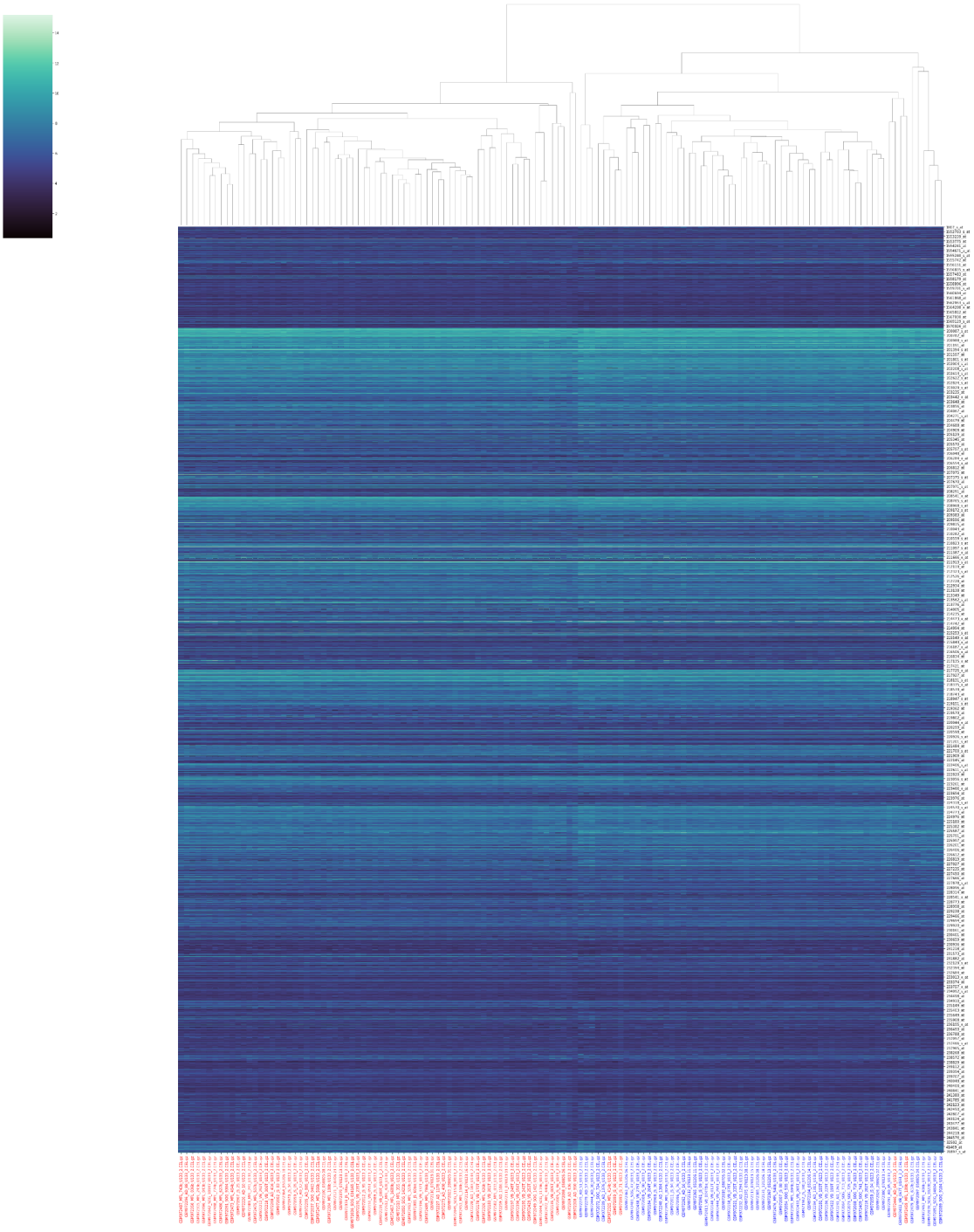

Figure 1: Heatmap of the gene-expression of each gene across all samples. The dendrogram of clustered samples shows on the top of the figure where corresponding labels show on the bottom. Red labels donate samples of C3

subtype and blue represent other subtypes. In the heatmap, levels of expression were represented as colors ranging from light to dark. The large and clear version of heatmap is available at the GitHub repository.

	t-statistic	p-value	fdr
203748_x_at	-20.897585	1.637247e-43	6.493486e-39
209868_s_at	-20.621450	5.152561e-43	1.021779e-38
207266_x_at	-20.636753	1.494585e-42	1.975892e-38
227059_at	-21.081130	3.498417e-42	3.468768e-38
225782_at	-20.360242	6.408415e-42	5.083283e-38
225651_at	-19.885346	2.052618e-41	1.356814e-37
209210_s_at	-19.701912	4.993155e-41	2.829050e-37
215127_s_at	-19.446701	1.425369e-40	7.066446e-37
216321_s_at	-19.345204	2.377952e-40	1.047911e-36
217764_s_at	-19.363191	3.267564e-40	1.295948e-36
225442_at	-19.587488	4.322409e-40	1.558464e-36
226237_at	-19.157342	7.450759e-40	2.462538e-36
227561_at	-19.902558	1.147266e-39	3.500131e-36
212607_at	-20.329536	1.521072e-39	4.309088e-36
212298_at	-19.074713	1.754751e-39	4.639677e-36
202363_at	-19.652069	2.361502e-39	5.853721e-36
211671_s_at	-18.907131	3.176186e-39	7.410042e-36
202766_s_at	-18.867618	3.949281e-39	8.701801e-36
202133_at	-18.883866	4.181587e-39	8.728733e-36
226930_at	-19.234803	6.149406e-39	1.219458e-35

Table 1: Top 20 DE genes (probe set IDs) in 2 clusters in ascending order of adjusted p-values.

Table 1 summarizes the most differentially expressed genes that define the clusters. Welch's t-test, or unequal variances t-test, is a two-sample location test which is used to test the hypothesis that two populations have equal means. Table 1's results have the most confidence, measured by p-values and FDR, that those genes significantly differentially expressed in two clusters and therefore characterized those clusters.

- ## Discussion

  The hierarchical clustering strategy has revealed 2 major clusters from the samples in the current study, the one containing the majority of C3 subtype samples and the one with the rest. It proves the naive cluster discovery approach, the hierarchical clustering, was moderately effective. The top DE genes in 2 clusters are informative. In the following analysis, those probe IDs can be converted into gene symbols to continue on a pathway analysis, and thus build a profile to the gene-expression classification of samples.

Reference

1. Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, Etienne-Grimaldi MC, Schiappa R, Guenot D, Ayadi M, Kirzin S, Chazal M, Fléjou JF, Benchimol D, Berger A, Lagarde A, Pencreach E, Piard F, Elias D, Parc Y, Olschwang S, Milano G, Laurent-Puig P, Boige V. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med. 2013;10(5):e1001453. doi: 10.1371/journal.pmed.1001453. Epub 2013 May 21. PMID: 23700391; PMCID: PMC3660251.