Expression Profiling in Inflammatory Bowel Disease

Yue Sun[1], Yiming Zhang[2], Omar AlOmeir[3], and Abrar Wafa[4]

[1]Bioinformatics Program, [2]Electrical and Computer Engineering Department, [3]Department of Computer Science, [4]Electrical and Computer Engineering Department

Introduction

Question

Inflammatory bowel disease (IBD) is a group of inflammatory conditions of the colon and small intestines. It is a complex disease which arises as a result of the interaction of environmental and genetic factors. There are two main forms of IBD: Crohn's disease (CD) and ulcerative colitis (UC). There is an overlap between the two forms in several areas including clinical criteria and therapy.

Objective: Our goal is to identify novel unknown genes involved in perpetuating inflammatory disease progression.

Data

We worked on a public dataset (GEO accession number: GSE1710)[1]. Our aim was to find Differentially regulated genes in High-density cDNA microarray data from the GPL284, Human UniGene Set RZPD 1, platform.

The dataset contains 34560 probes of 31 samples. Biopsies were taken from the sigmoid colon.

Group	Sex	Total
NC	Male	4
	Female	7
UC	Male	8
	Female	2
CD	Male	5
	Female	5

Table 1. Experimental design, number of samples for each group.

Data Exploration

	row.names	GSM29595	GSM29596	GSM29597	GSM29598	GSM29599	GSM29600	GSM29601
1	01A01	0.214269041	0.141720994	0.153143777	0.2013757224	0.128302990	0.137858628	0.179688616
2	01A02	0.116874023	0.079949199	0.068256213	0.0997897974	0.057652420	0.061531542	0.100846051
3	01A03	0.119945868	0.076125740	0.069136939	0.1136518264	0.088503479	0.058384946	0.098192795
4	01A04	0.124453886	0.071602313	0.069181426	0.1093685914	0.107247277	0.063042267	0.071402176
5	01A05	0.113459273	0.061464241	0.058253242	0.0837338854	0.082051321	0.050408623	0.063358425

Table 2. Data excerpt showing first 5 probes of first 7 samples.

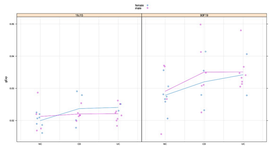

Fig 1. Stripplot of two random genes showing the effect of groups and sex.

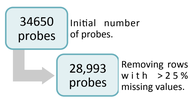

Fig 2. Number of probes after filtering.

- In the data exploration step, we found an outlier sample that has low correlation to other samples. The outlier was removed. The dataset also contained a large number of missing values.
- In the quality control steps, we removed rows with more than 25% missing values, and imputed the remaining missing values using k nearest neighbors and filled them with data from the 10 nearest neighbors.
- As a last step we performed quantile normalization to deal with technical variability.

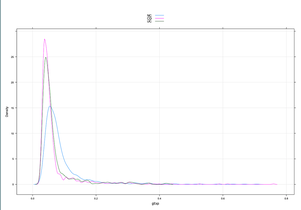

Fig 3. Distribution of first 100 probes for all samples in the three groups.

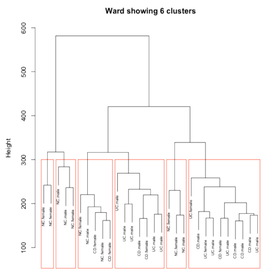

Fig 4. Hierarchical clustering using Ward method to explore the data where labels reflect groups and sex.

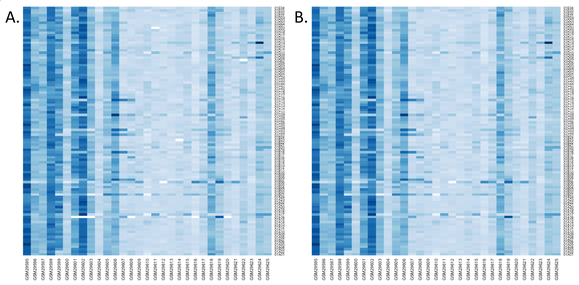

Fig 5. Imputation of missing values using k nearest neighbor. **A**. Heatmap of first 100 probes before imputation. **B**. Heatmap of first 100 probes after applying knn, using nearest neighbor averaging.

Differential Analysis

The used linear model is the ANOVA style, 'reference + treatment effects' parameterization.

$$Y_{ij} = \theta + \tau_i + \varepsilon_{ij} \text{ where } \tau_1 = 0.$$

$$\mu_{NC} = \theta, \mu_{CD} = \theta + \tau_2, \text{ and } \mu_{UC} = \theta + \tau_3.$$

i = 1, 2, 3.

j = 1, 2, 3, ..., 28933 probes.

Equation 1. ANOVA style linear model for group effect.

The statistical test performed is limma F-test to compare models where the following null hypothesis is tested:

$$\mu_{NC} = \mu_{CD} = \mu_{UC}$$

Equation 2. Null hypothesis to test if all the means of the three groups are equal.

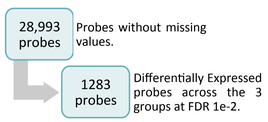

Fig 6. Number of probes after DEA.

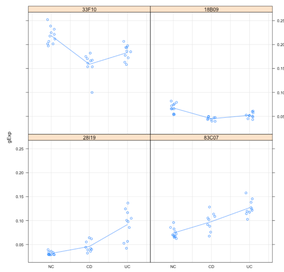

Fig 6. Four random hits where the mean of the three different groups are NOT equal.

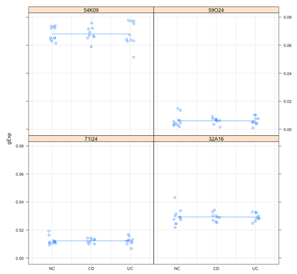

Fig 7. Four random non-hits where the mean of the three different groups are equal.

Thus, testing for differentially expressed genes between normal and IBD samples resulted in 1283 significant probes. While, testing for DE genes between Crohn's disease and ulcerative colitis samples did not show significant results.

Further tests were done to explore the interaction effect between group and sex, and group and age. However, age and sex proved to be nuisance factors at best.

Principle Component Analysis

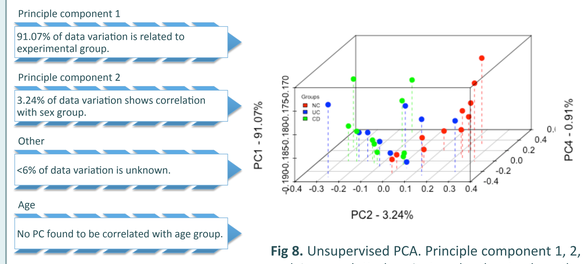

Fig 8. Unsupervised PCA. Principle component 1, 2, and 4 were plotted against each other to show the difference between samples.

Principal Component Analysis shows that from PC1, PC2 and PC4, we can tell the differences between normal control group (NC) and IBD patient group (CD and UC), but we cannot tell the difference between the Crohn's disease (CD) group and the ulcerative colitis (UC) group.

Gene clustering

In gene clustering analysis, 1283 genes which were found to be differentially expressed were analyzed. Setting k as 5, we performed k-means clustering to get the results.

In k-means gene clustering results, with the differentially expressed genes partitioned into 5 clusters, there is one cluster that shows significantly higher differential expression levels than the other clusters.

- This cluster only contains 7 genes.
- The differential expression level of this cluster is 10 times higher than of other clusters.
- Genes in this cluster have much higher expression levels in the IBD patients group than in the control group.
- Differences between UC and CD in this cluster are still not significant.

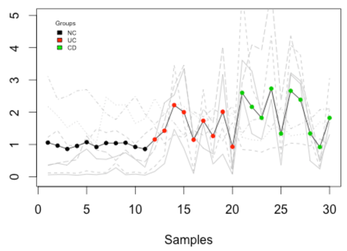

Fig 9. Gene clustering. The different genes and cluster centers were plotted against each other to show the difference between samples.

Classification

- In classification analysis, we performed and compared 6 different classification methods with different numbers of features.
- The 30 samples were separated into a training set (18 samples) and a testing set (12 samples), after doing classification analysis, 2-fold cross validation was performed to get the error rate.
- Features were chosen from 1283 differentially expressed probes identified in the differential analysis stage, the number of features was set as 100, 300, 500, and 1000.
- Finally, the Linear discriminant analysis (LDA) methods with 300 top hits genes as features had the lowest error rate of 11.11%.

Gene Enrichment Analysis

Conclusions

- A number of genes were found to have statistically significant expression values across all three groups.
- Age and sex had no apparent effect in our experiments even though age could be a factor for Crohn's disease and ulcerative colitis.
- In overall analysis, significant differences between the genes of normal people and IBD patients' genes were found, while any differences between the Crohn's disease group and the ulcerative colitis group still need to be confirmed.
- Classification showed promising results, the classifier we built could be used for prediction.

Improvements

- Increased sample size would increase the reliability of the conclusion, especially for classification analysis.

References

[1] Expression profiling in inflammatory bowel disease data set.

[2] Costello, Christine M., et al. "Dissection of the inflammatory bowel disease transcriptome using genome-wide cDNA microarrays." PLoS medicine 2.8 (2005): e199.

[3] Chen, Chao, et al. "Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods." PloS one 6.2 (2011): e17238.

[4] Eickhoff, H. et al. Tissue gene expression analysis using arrayed normalized cDNA libraries. Genome Res 10, 1230-40. (2000).

RESEARCH POSTER PRESENTATION DESIGN © 2015
www.PosterPresentations.com