Understanding Genetic Causes of Irritable Bowel Syndrome Through Learned Models of Stress-Related and Inflammatory Bowel Syndrome Diseases

Jean Feng

June 5, 2012

Abstract

Irritable bowel syndrome is a complex disorder that has a high comorbidity with stress-related disorders and inflammatory bowel diseases. In this study, our goal was to determine how IBS is related to these two different classes of diseases. Our first goal was to learn models for the genes related to these two disease classes, which was performed via a hierarchical clustering over the phenotype feature vectors correlated with the genes. Our results revealed two clusters of genes, one highly correlated with autoimmune diseases and the other with stress-related disorders. We then used these models to predict the categories the genes correlated with IBS belonged to. The results revealed that of the six genes under consideration, only TAP2 was significantly related to stress while the other five genes had a high likelihood of being associated with inflammatory bowel diseases, but did not meet the threshold ($p < 0.05$). In addition, we constructed indirect protein-protein interaction networks from the genes from each cluster and determined that the genes were not only related by phenotype but also through their biomolecular function, confirming the validity of our methodology.

Introduction

Irritable bowel syndrome (IBS) is a highly prevalent disorder affecting 10-20% of people in Western countries. A symptom-based diagnosis, IBS is characterized by recurrent abdominal pain and alterations in bowel function. Patients experience diarrhea, constipation, or even alternations between the two. IBS is most commonly classified a neurogastroentological disorder. Common triggers for IBS include diet, stress, and abdominal surgery, therefore leading doctors to recommend palliative treatments such as dietary adjustments and psychotherapy. In addition, IBS has a high comorbidity with depression and inflammatory bowel diseases. However, even with the large population of IBS patients, the exact pathophysiology remains unclear.

The goal of this study is to develop a better understanding of the underlying causes of IBS. If some of the genes associated with IBS are also significantly associated with other diseases, doctors might be able to adopt those treatments to possibly cure or relieve IBS. Aerssens, Camilleri, et. al. have already made headway on identifying genes associated with IBS [1]. Using gene expression profiling, they identified twelve genes that were significantly associated with IBS. In our study, we investigated if these genes were associated with conditions that have high comorbidity with IBS. That is, using genetic analysis, we tried to determine how IBS is associated with inflammatory bowel disorders, which are primarily autoimmune diseases, and stress-induced conditions.

We based our experiments on a paper recently published by Cotsapas, Voight, et. al [2]. In their study, they analyzed how autoimmune diseases were related by aggregating the genome-wide association studies from various autoimmune diseases. They constructed a feature vector for each SNP using the association p-value of the SNP with each disease. A hierarchical clustering was then performed over the SNP feature vectors to determine which ones were related by phenotype. From this, they discovered four clusters of genes, where three of the four clusters were associated with a distinct subset of autoimmune diseases. Next, to determine how closely the SNPs from each cluster were related by their biomolecular interactions, they

Disease	Dataset	Number of Samples	Affymetrix Chip Type
IBS	E-TABM-176	36 cases, 25 control	U133 Plus 2.0
Ulcerative Colotis	GSE3365	59 cases, 42 control	U133 Plus 2.0
Crohn's Disease	GSE3365	26 cases, 42 control	U133 Plus 2.0
Chronic Fatigue Syndrome	GSE14577	8 cases, 7 control	U133A
Depression/Bipolar/Schizophrenia	SMRI: Study ID 9*	60 cases, 40 control	U133A

Table 1: Datasets used for this study, (*) SMRI is the Stanley Medical Research Institute. The data can be accessed by visiting www.stanleygenomics.com

constructed indirect protein-protein interaction maps. Their results confirmed that the hierarchical clustering was effective at predicting which SNPs were related biomolecularly.

We applied a similar process to determine how the genes associated with IBS are related with inflammatory bowel disorders and stress-related conditions. In particular, we analyzed at Crohn's disease and Ulcerative Colotis from the former category and Chronic Fatigue Syndrome, Bipolar Disorder, Schizophrenia, and Depression for the latter.[1] From our results, we found that out of the twelve genes associated with IBS, one gene was significantly associated with psychological conditions and the rest were more highly associated with inflammatory bowel disorders, but failed to meet the threshold of $p<0.05$.

Materials and Methods

The datasets used for this research paper are shown in Table 1. The data was collated for a total of five phenotypes, where depression, bipolar disorder, and schizophrenia were grouped as one phenotype and the other four phenotypes were IBS, Ulcerative Colotis, Crohn's Disease, and Chronic Fatigue Syndrome [1, 7, 8, 9]. As discussed previously, CD and UC are both inflammatory bowel diseases, while CFS, depression, bipolar disorder, and schizophrenia are considered to be stress-related disorders partially caused by mental stress. All of these diseases have high comorbidity with IBS.

For each gene present on both the Affymetrix Chip U133 Plus 2.0 and U133A, we constructed a corresponding five-element feature vector using the q-values for each phenotype. To limit the number of genes in the hierarchical clustering process, we only clustered genes with q-values less than 0.03 for at least two of the phenotypes. Furthermore, we filtered out genes with a q-value for IBS greater than 0.1, leaving a total of 87 genes.

The first step was to learn models over the genes. We performed hierarchical clustering on the gene feature vectors to form two major clusters. Because the level of association between a gene and a phenotype is not linearly proportional to the q-value, we did not perform clustering on the raw q-values. Instead, we used buckets for the q-values. For each element in the feature vectors, if the q-value was less than 1e-5, it was replaced by a label of 0. Similarly, q-values between 1e-5 and 1e-3 were set as 1, q-values between 1e-3 and 0.05 were set as 2, 0.05 and 0.1 were set as 3, and anything higher had a label of 4. Instead of using Euclidean distance between the feature vectors, we used the "Gower metric," which is meant for calculating the dissimilarities between mixed data types [10]. After calculating the Gower distance between all pairs of gene feature vectors, we performed hierarchical clustering using Ward's method and extracted the final two major clusters.

To assess the hypothesis that the genes in each cluster were also highly associated in their biomolecular interactions, we used DAPPLE, a Disease Association Protein-Protein Link Evaluator developed by Rossin EJ, Lage Ky, et. al [11]. DAPPLE looks for significant physical connectivity among proteins encoded for by genes according to protein-protein interactions reported in publicly available literature. In addition to the networks, DAPPLE automatically calculates three different metrics to measure the connectivity of the indirect protein networks: associated protein direct connectivity, associated protein indirect connectivity, and common interactor connectivity. As stated in DAPPLE, the connectivity is proportional to the sum of

[1]There have been many studies that include stress in the etiology of bipolar disorder, CFS, schizophrenia, and depression. We include relevant papers in the reference section [3, 4, 5, 6].

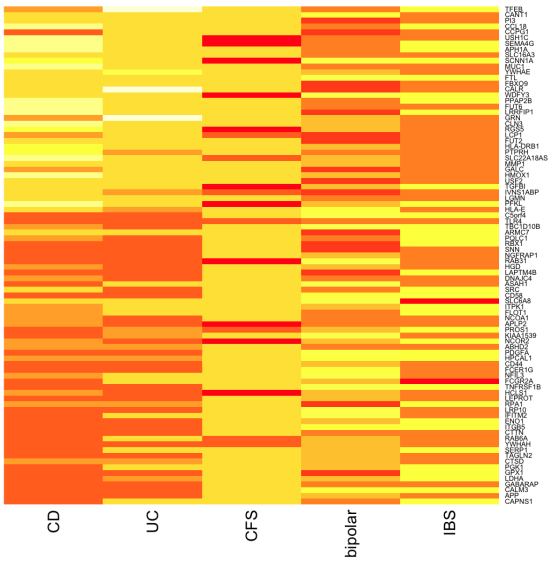

Figure 1: Heat map of from the hierarchical clustering of the filtered genes. Rows correspond to each gene. One can see a distinct different between the bottom half and the top half. The bottom half corresponds to genes in the first cluster, which is more associated with CD and UC. The top half is more associated with bipolar depression, and schizophrenia.

the degrees of the nodes, and the p-value of the connectivity value is the probability of a random network attaining the connectivity metric value (random networks are generated by replacing nodes in the original network with genes with the same total number of reported protein-protein interactions in the public literature). Statistically significant connectivity suggests that the proteins in the constructed network are involved closely in some molecular mechanism. Details regarding the calculations for these metrics are given in the supplementary section of their paper.

Using these models, we classified the genes reported by Aerssens that were significantly associated. We used two prediction methods to classify the IBS genes. The first method adds the feature vector for each IBS-associated gene to the set of feature vector set of the original 87 genes and re-performs hierarchical clustering.[2] The second method calculates the gower distance from each IBS gene feature vector to the median feature vector of each cluster. Therefore, we then classified each gene based on the cluster it ended up in.

To determine the accuracy of this method, we used DAPPLE again and built another two indirect protein-protein interaction networks by adding all the genes highly associated with IBS to the the pool of genes in the cluster 1 and cluster 2 networks. We then compared the p-values for the seed scores for each IBS-associated gene to determine which cluster it was more closely connected to. The seed score, like the connectivity metric, is also generated by DAPPLE and measures the connectivity and integration of that particular gene in the network. The metric is proportional to the degree of the node of the gene, and even more importantly, its p-value is the probability of attaining that metric value by chance. Each gene is assigned to the network cluster with lower p-value for the seed score.

Results

After performing Ward's method on the 87 genes, 51 of the genes were grouped into the first cluster and 36 were in the second by hierarchical clustering (Fig. 1 shows the heat map, Fig. 3 in Appendix A contains the dendrogram). The median feature vector for cluster 1 was $(0, 0, 4, 2, 2)$, which indicates that these genes

[2] Adding a gene at a time to the filtered genes did not affect clustering of the original set of genes, which makes this method feasible.

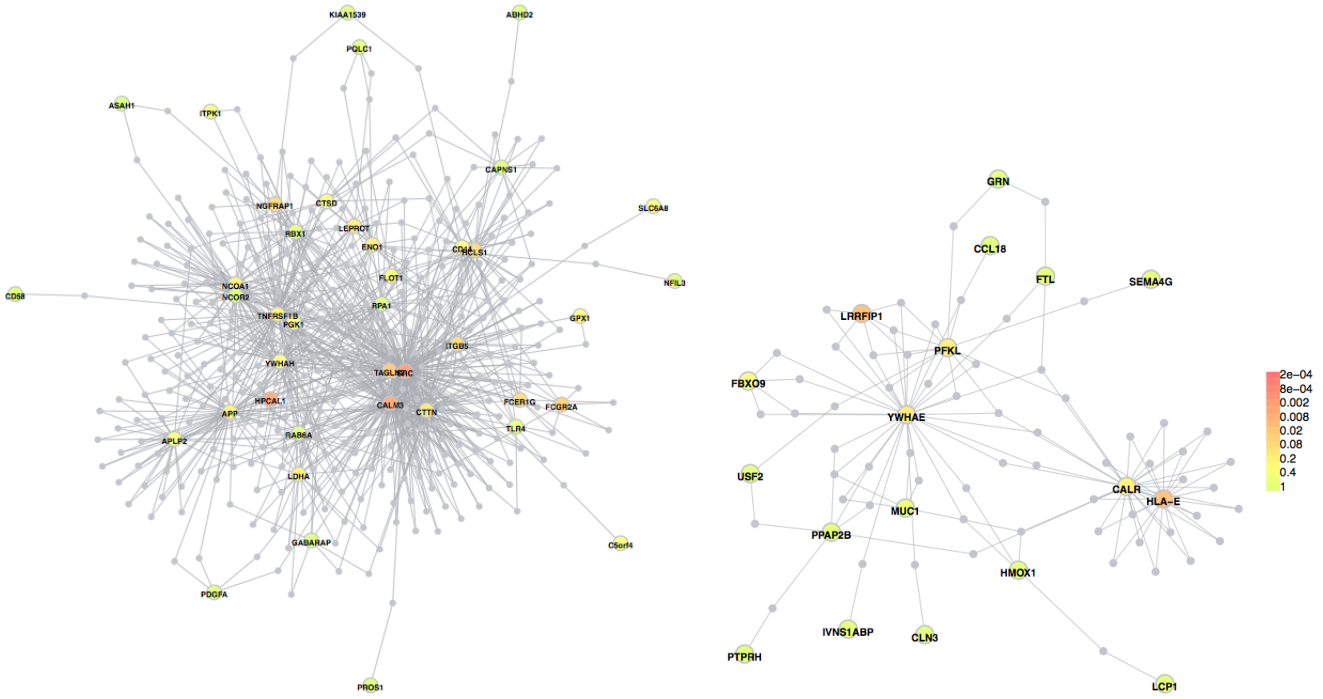

Figure 2: Left: The indirect protein-protein interaction network containing genes associated with UC and CD, Right: The indirect protein-protein interaction network containing genes associated with stress-related disorders

Cluster	Direct Associated Protein	Indirect Associated Protein	Common Interactors Protein
1	1.63, p=0.101	47.90, p=0.016	2.60, p=0.014
2	1.0, p=0.508	8.78, p=0.042	2.13, p=0.250

Table 2: Connectivity metric values of the indirect protein-protein interaction network for each cluster as calculated by DAPPLE

Gene	cluster 1 seed score p-value	cluster 2 seed score p-value	Network Prediction	Clustering Prediction
TAP2	0.536	0.002	2	2, 2
FCGR2A	0.065	0.653	1	1, 1
RCF4	0.250	0.831	1	1, 1
MCM5	0.289	0.625	1	1, 2
KCNS3	0.304	N/A	1	1, 2
CKB	0.356	0.534	1	2, 2

Table 3: Seed scores of genes significantly associated with IBS in clusters 1 and 2 from the networks and the classifications from the network seed scores and hierarchical clustering. The two numbers in the clustering prediction column are the predictions from the first and second classification methods, respectively.

are highly correlated to CD and UC. The median for cluster 2 was $(2, 2, 4, 1, 2)$, which means that these genes are highly correlated with mental disorders. Note that the ordering of the elements in the feature vector are (CD, UC, CFS, bipolar/depression/schizophrenia, IBS). The DAPPLE-constructed indirect protein networks are shown in Figure 2. The connectivity metrics of each network are shown in Table 2.

In addition, to determine the effects each cluster had on IBS, we compared the predictions based on hierarchical clustering and the indirect protein networks. The clustering predictions and the seed scores from networks are shown in Table 3. In general, we would consider a gene to be significantly related to a cluster if its p-value for that cluster was less that 0.05. We included the genes with seed score p-values up to 0.5, however, to show that the hierarchical clustering predictions using the first method were mostly correct up to an even higher p-value threshold. Therefore, our hierarchical clustering classification method was able to correctly classify 5 out of the 6 genes. The second method that just calculated the distance to the median of each cluster was able to correctly classify the genes that had a low p-value in one cluster and a p-value that was at least 3 times higher in the other cluster. In total, this second method correctly classified 3 out of the 6 genes. This is still useful since it is the most important that the clustering prediction match the network prediction for genes with a p-value that is significant in one of the clusters.This is because the network prediction for which cluster the IBS-associated gene is much more likely to be correct.[3]

Discussion

As expected, the hierarchical clustering of genes on the feature vectors formed one cluster highly correlated with CD and UC and another cluster highly correlated with depression, bipolar disorder, and schizophrenia. Tight biomolecular interactions in each cluster are not guaranteed by the phenotype-driven clustering, but from DAPPLE, we were able to confirm that each cluster were "significantly connected." For cluster 1, the connectivities for indirectly associated proteins and the common interactor proteins were significant. For cluster 2, the indirectly associated protein connectivity was significant. In general, as claimed by Rossin et. al., indirectly associated protein connectivity and the common interactor connectivity metrics are good indicators that a cluster contains genes participate in a complex biomolecular process. Therefore, our hierarchical clustering method proved effective at clustering genes by biomolecular interactions using phenotype-based evidence. This is similar to the observation by Cotsapas in their study only focused on autoimmune diseases. From both our study and Cotsapas, one cannot make a confident assertion that hierarchical clustering is an accurate method in predicting biomolecular interactions, but the results from both our studies do warrant further investigation into this method.

The second part of our study was to determine if we could predict the relationship IBS has with inflammatory bowel syndrome and stress-related disorders through hierarchical clustering. From the hierarchical clustering, we see that only two of the six genes, TAP2 and CKB, were classified as highly associated with stress-related diseases. The indirect protein network results from DAPPLE produce similar predictions, with the difference that CKB is classified into cluster 1 instead of 2. However, given the the high p-values of CKB for both clusters, one cannot confidently conclude which cluster CKB actually belongs to. Regardless, the correspondence between the results from DAPPLE and hierarchical clustering are significant, agreeing on five of the six genes. This gives us more confidence that our hierarchical clustering method can be used to predict what types of biomolecular interactions the proteins would be involved in.

Interestingly, when reviewing the GO terms for TAP2, there are no terms that mention its relation with stress or the nervous system. Instead, after searching through scholarly publications, we find a publication in 2012 that concluded that TAP2 was significantly correlated with depressive syndrome and alcohol dependence, which agrees with our results [12]. Therefore, clustering method may in fact help uncover certain functions of genes that are not well-known.

The rest of the proteins that IBS might be associated with are more tightly related to autoimmune diseases, although they are not significantly associated. While we cannot make a definite determination as to whether IBS is a disorder more correlated to one's mental stress or to one's immune system, this study suggests that genetically, IBS does have genetic connections with stress-related disorders like depression,

[3]The other 6 genes significantly associated with IBS are not displayed because DAPPLE was unable to find the interactions for those proteins in publicly available literature.

bipolar disorder, and schizophrenia.

Future work could include building bigger feature vectors involving more autoimmune and stress-related disorders, which would provide more fine-grained classifications and possibly reveal three or more clusters. In addition, comparisons between hierarchical clustering and indirect protein networks should be done for other classes of diseases to determine if this phenotype-based method can indeed reveal genes tightly connected in complex biomolecular connections.

References

1. Aerssens J, Camilleri M, Talloen W, et al. Alterations in mucosal immunity identified in the colon of patients with irritable bowel syndrome. Clin Gastroenterol Hepatol. 2008;6:194–205. (http://www.ncbi.nlm.nih.gov/pmc/

2. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, et al. (2011) Pervasive Sharing of Genetic Effects in Autoimmune Disease. PLoS Genet 7(8): e1002254.doi:10.1371/journal.pgen.1002254 (http://www.plosgenetics.org/ar

3. Post, R M. "The role of psychosocial stress in the onset and progression of bipolar disorder and its comorbidities: the need for earlier and alternative modes of therapeutic intervention." Development and psychopathology 18.04 (2006):1181.

4. Physical, behavioral and psychological risk factors for chronic fatigue syndrome: A central role for stress? Dobbins, James G.; Natelson, Benjamin H.; Brassloff, Ira; Drastal, Susan; Sisto, Sue-Ann Journal of Chronic Fatigue Syndrome, Vol 1(2), 1995, 43-58. doi: 10.1300/J092v01n02_04

5. Christine C Gispen-de Wied, Stress in schizophrenia: an integrative view, European Journal of Pharmacology, Volume 405, Issues 1–3, 29 September 2000, Pages 375-384, ISSN 0014-2999, 10.1016/S0014-2999(00)00567-7. (http://www.sciencedirect.com/science/article/pii/S0014299900005677)

6. Hammen, C. "Stress and depression." Annual review of clinical psychology 1.1 (2005):293.

7. Burczynski, M E. "Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells." The journal of molecular diagnostics 8.1 (2006):51.

8. Gow, J. "A gene signature for post-infectious chronic fatigue syndrome." BMC medical genomics 2.1 (2009):38.

9. V. Moskvina, N. Craddock, P. Holmans, I. Nikolov, J.S. Pahwa, E. Green, M.J. Owen, M.C. O'Donovan, Wellcome Trust Case Control Consortium. Gene-wide analyses of genome-wide association data sets: Evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk (http://www.nature.com/mp/journal/v14/n3/pdf/mp2008133a.pdf)

10. Gower J (1971) A general coefficient of similarity and some of its properties. Biometrics 27: 857–874. Ward Jr. J (1963) Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association. pp 236–244

11. Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, et al. (2011) Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. PLoS Genet 7(1): e1001273. doi:10.1371/journal.pgen.1001273

12. Edwards, A C. "Genome-wide association study of comorbid depressive syndrome and alcohol dependence." Psychiatric genetics 22.1 (2012):31.

Appendix A

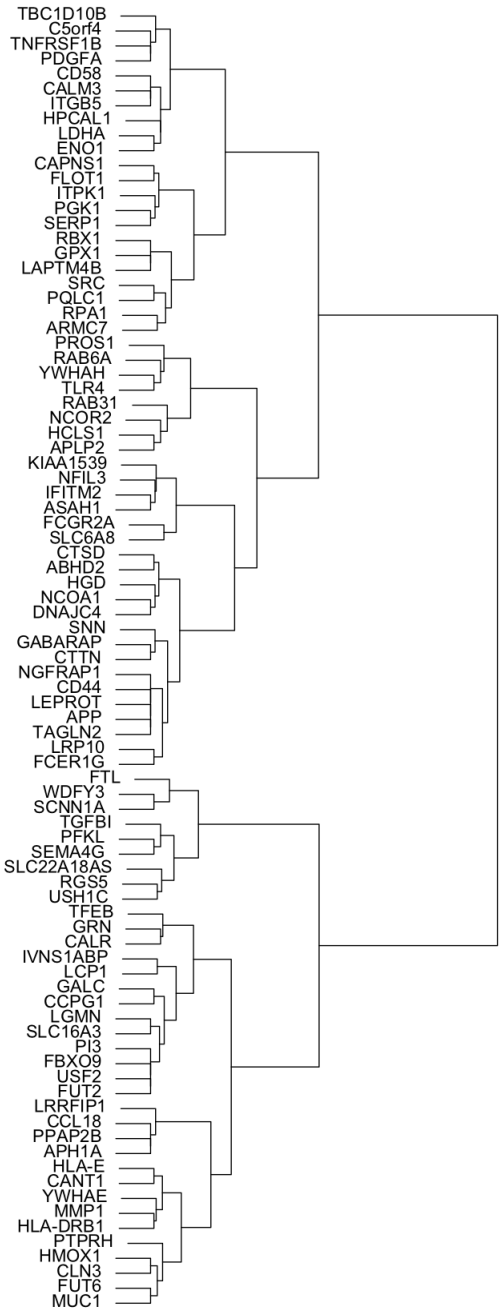

Figure 3: Dendrogram from hirearchical clustering