Huntington's Disease Biomarkers: Peripheral Blood Gene Expression Profiling

Samantha Hartner

Introduction

Huntington's Disease (HD) is an autosomal dominant genetic condition that results in progressive neurodegeneration as patients age. Biomarkers for HD that can be tracked throughout disease progression can help improve diagnosis of disease stage, and aid in the assessment of treatment effects during clinical trials[1]. Although HD targets the central nervous system (CNS), the mutant *huntingtin* protein is known to be ubiquitous, meaning that biomarkers for HD may be found in any tissue throughout the body. The goal of the study presenting this dataset was to identify potential biomarkers for HD in human peripheral blood using mRNA assays to track changes in gene expression among symptomatic HD patients, pre-symptomatic HD patients, and healthy control patients.

Methods:

GEO2R was used to import the GSE1751 dataset into R Studio[2]. For this simplified analysis, the pre-symptomatic HD group was excluded, and only the symptomatic HD and control groups were analyzed. First the expression set data was normalized using a log2 transformation. Then, the *limma* package for R was used the test the null hypothesis that genes were not differentially expressed between HD and control patients. After performing the empirical Bayes moderated t-statistics, a table was produced containing the top 322 genes ranked by adjusted p-value. These 322 genes were used for the remaining analyses.

Principal component analysis (PCA) was performed using the *prcomp* function and according to a protocol published on VIB Bioinformatics Core Wiki[3]. Hierarchical and k-means clustering were performed according to a machine learning protocol[4]. The list of top genes was further narrowed down to include only those genes that were differentially expressed with an adjusted p-value less than 0.05, and absolute fold change greater than 2. The 81 genes that met these criteria were compared to the 12 genes indicated as biomarkers by the original study. Finally, basic pathway analysis was performed using Panther, Reactome, and KEGG[5-7].

Data Description

The GSE1751 dataset included 31 samples (12 HD, 5 pre-symptomatic HD, and 14 healthy controls), 26 of which were analyzed in this project. Gene expression was profiled using the Affymetrix Human Genome U133A Array. The dataset contained gene expression information from 22,283 human genes.

Results

In my analysis, I found that only 8 of the 12 potential biomarkers identified by the original publication were included in my top 322 genes, while only 2 of the 12 genes were included in my top 81 (Table 1). A volcano plot including all genes from the microarray identifies 193 genes bearing both significant adjusted p-value and absolute fold change (Figure 1). It is evident in the volcano plot that there is a bias showing far more genes with a negative fold change than genes with a positive fold change. Since control samples were compared to the HD samples as a reference, this indicates that the genes analyzed tend to be more over-expressed than under-expressed in HD patients.

ANXA1	†
AX0T	
CAPZA1	†
HIF1A	†
JJAZ1	
P2Y5	
PCNP	†*
ROCK1	†
SF3B1	†
SP3	†*
TAF7	†
YIPPEE	

Table 1. List of 12 biomarkers selected in original publication. † indicates inclusion in the 322 top genes in this analysis, while * indicates inclusion in the top 81 differentially expressed genes.

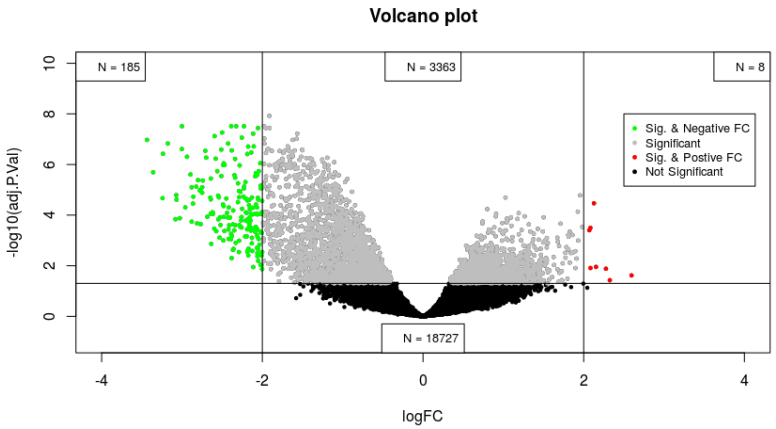

Figure 1. Volcano plot showing fold change compared to adjusted p-value for all genes.

A heatmap comparing gene expression between HD patients and control patients shows a similar trend; gene expression values tend to be lower in the control samples, indicated by a trend toward more blue coloring in the control columns and more red coloring in the HD columns (Figure 2).

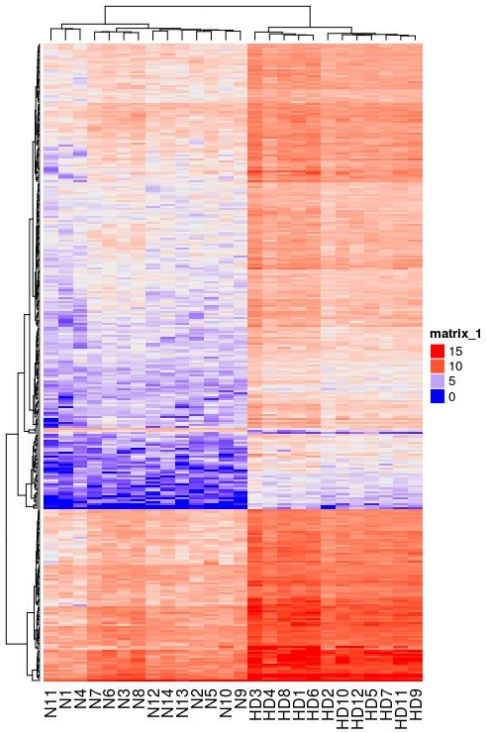

Figure 2. Heatmap showing differential gene expression between control (N1 – N14) and Huntington's Disease (HD1 – HD12) samples (columns) for the 322 genes (rows) analyzed.

Both the PCA and k-means analyses were successful at clustering the samples into HD and control groups. Figure 3 shows the results of PCA and k-means plotted in two dimensions. The PCA comparing components 1 and 2 and components 1 and 3 was successful in grouping the two sets of samples, while the PCA comparing components 2 and 3 was less successful, and does not show 2 clear groups of samples. In all PCA plots, it is evident that the control samples (shown in blue) exhibit a tighter grouping than the HD samples. Results plotted from the k-means analysis also shows two groups of samples successfully clustered around two centers. A cluster dendrogram, shown in Figure 4, shows that the two groups of G0 and G1 samples (HD and control, respectively) were successfully partitioned into different branches of the tree.

Pathway analyses performed via Reactome and KEGG indicated that the 322 differentially-expressed genes may be significantly involved in cell signaling, transcription, cell cycle, and apoptosis pathways, among others. Functional analysis performed via the Panther database show that the differentially-expressed genes may play a large roll in cell binding, catalytic activity, and cellular and metabolic processes (Figure 5).

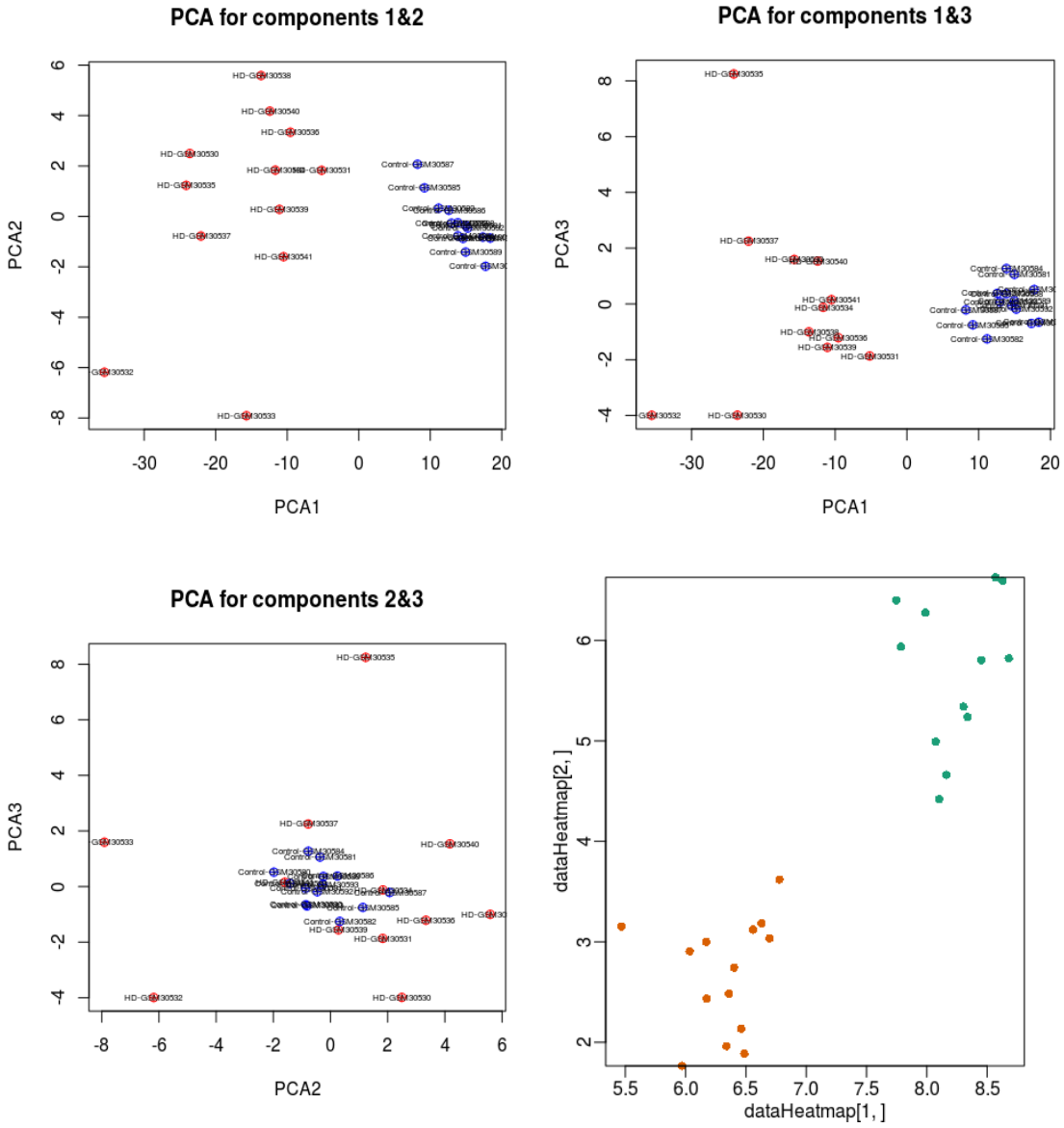

Figure 3. PCA component plots showing successful clustering of control and HD samples using components 1&2 as well as components 1&3. Clustering was not successful using components 2&3. A data heatmap from the k-means clustering shows the 14 control samples and 12 HD samples successfully grouped around 2 centers.

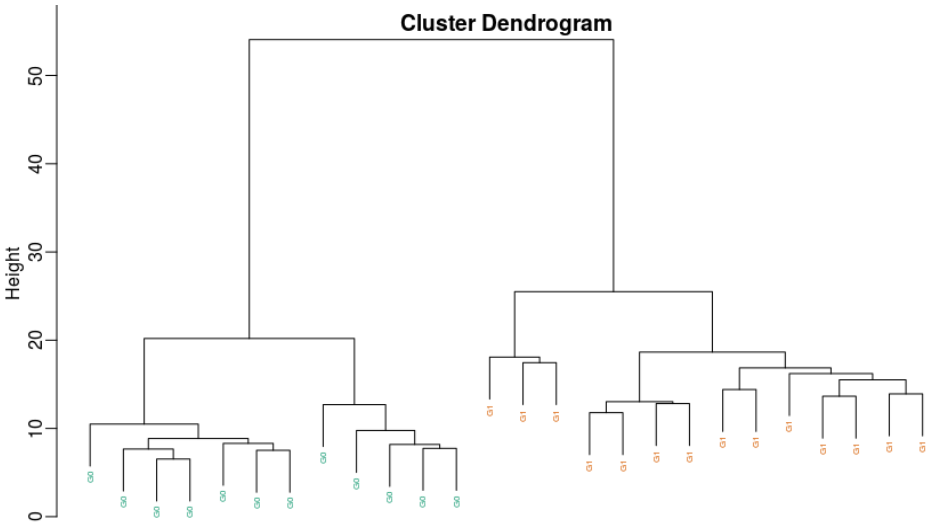

Figure 4. Cluster dendrogram showing the 12 HD samples (G0, green) and 14 control samples (G1, orange), partitioned into separate branches of the tree.

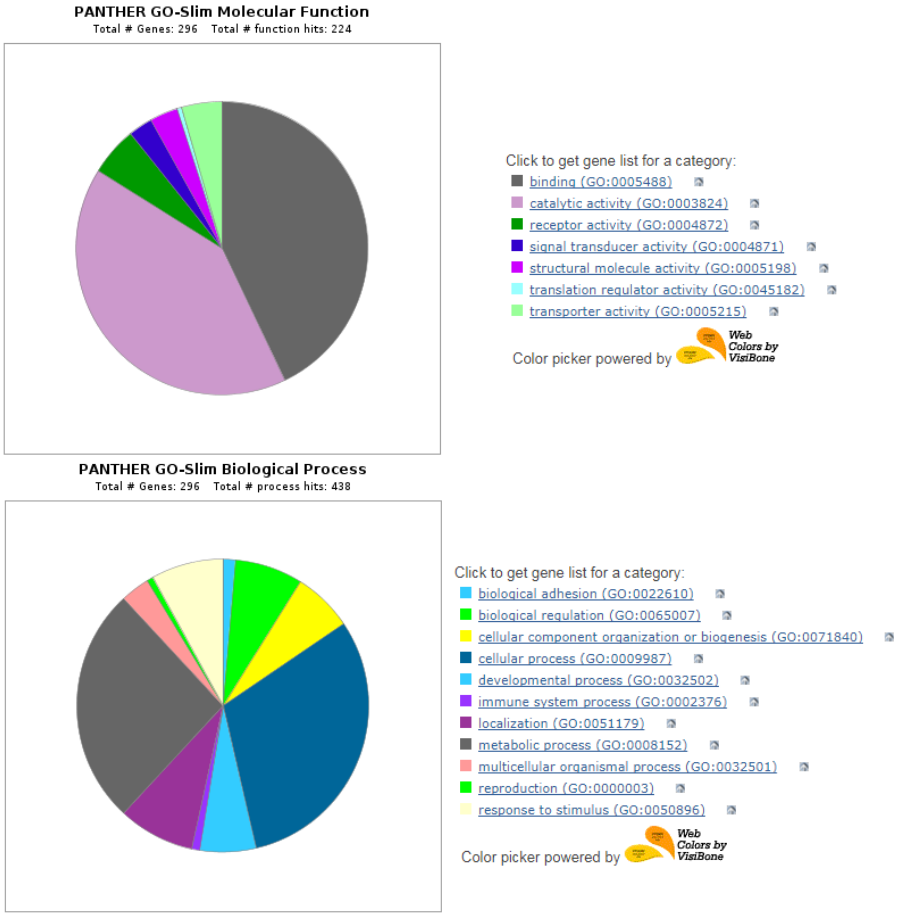

Figure 5. Panther analysis showing major molecular and biological functions correlated to the 322 genes analyzed.

Discussion

Overall, my analysis was successful at identifying differentially-expressed genes in the dataset. In the final step of the analysis, the list was narrowed down to 81 genes of interest that were differentially expressed, with both an adjusted p-value and absolute fold change that were deemed significant. In addition, the clustering and hierarchical analysis were successful at grouping the HD and control samples when using the 322 genes with the most significant p-values for analysis. Both the heatmap and the volcano plot show that a notable bias exists in the data, indicating that the genes analyzed tend to be over-expressed in HD patients rather than under-expressed. From my analysis, I would conclude that there are many potential biomarkers for Huntington's Disease in the peripheral blood, but further study would be needed to identify the most effective biomarkers.

In some ways, my data agreed with the reference study, but there were many notable differences. Many of these differences could have arisen from the different methods used in the two analyses. The main difference in analysis methods was that I excluded the middle group, pre-symptomatic HD patients, from my dataset, while the original authors included these samples in the HD group. For my analysis, I only used data from the Affymetrix GeneChip, while the reference study also included data from the Amersham microarray, and cross-referenced the two sets of expression data. While all of my analyses were performed in R Studio, the reference study used other statistics software, as well as some proprietary software used with their lab equipment. The original authors performed different filtering and normalization to their data before proceeding with the analysis, and they performed quantitative reverse-transcription (QRT)-PCR to further analyze differential expression of genes of interest. They identified 12 genes that can potentially be used as biomarkers to track the progression of Huntington's Disease. In my analysis, 8 of these genes were included in my top 322 genes, and only 2 of them were included in my top 81 genes. The reference study cited transcription, signaling, ubiquitin, and vesicle trafficking pathways as those potentially affected by these 12 differentially-expressed genes. My pathway analyses also identified these pathways of interest, as well as numerous other pathways. After performing pathway analyses on my 322 genes of interest, I found it nearly impossible to narrow down the number of pathways involved into a meaningful number.

The differences between my analysis and the analysis performed in the reference study highlight the importance of performing quality analyses and documenting the steps involved. Studies performed on the same dataset can have vastly different results depending on the type of analysis that is done. Documenting the steps is important because it allows other researchers to test the results against new data to further support or contest developing theories. It is important for the integrity of scientific data for all researches to make their best effort to perform accurate, unbiased analyses of their data. I achieved different results in my analysis, but I believe that if I were to follow the protocols and analysis methods of the original authors exactly, I would be much closer to replicating their results and conclusions.

References

1. Borovecki F, Lovrecic, L, Zhou, J, Jeong, H et al. Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. Proc Natl Acad Sci U S A 2005 Aug 2; 102(31):11023-8. PMID: 16043692
2. https://www.ncbi.nlm.nih.gov/geo/geo2r/
3. http://wiki.bits.vib.be/index.php/Analyse_GEO2R_data_with_R_and_Bioconductor
4. http://genomicsclass.github.io/book/pages/clustering_and_heatmaps.html
5. http://pantherdb.org/
6. http://reactome.org/
7. http://www.genome.jp/kegg/kegg1.html