Final Project Report

Re-analysis of the datasets that were used in a previously published article named "A module of human peripheral blood mononuclear cell transcriptional network containing primitive and differentiation markers is related to specific cardiovascular health variables"

Sayed Muhammad Saifuddin, Jonathan Garnsey

ABSTRACT

In this project, we analyzed the genes from 26 healthy subjects and 20 hypertensive patients using GEO2R to compare the two groups (e.g., HealthySubject and HypertensionPatient) of Samples in order to identify the differentially expressed genes across experimental conditions. We chose this part of analysis because we learned these techniques from the labs and we decided to apply those techniques in this project. We didn't try any partial analysis due to the time constraints along with so many submissions. More importantly, one of our group members was missing throughout the semester so it was very difficult to conduct further analysis within a very short period of time.

INTRODUCTION

In order to determine the effects of experimental conditions in identifying differentially expressed genes while comparing a healthy subject against a patient with hypertension, we applied techniques learned in this course as well as independent research and testing. Our group selected this focus as we found that this is a relevant focus to an admissible problem. Our interest in this topic was spurred by the history of this condition in patients and the newly emerging data that is being found using newer technologies. This report will cover our findings, our thought processes throughout the research conducted, as well as our categorized findings that's split into two cohesive data sets.

BACKGROUND

We analyzed the whole task with GEO2R and RStudio in two different ways. Such as, at first, we defined the two groups named HealthySubject and HypertensionPatient and then applied Benjamini & Hochberg (False discovery rate) for the adjustment to the P values and limma precision weights to analyze the significance level of top differentially expressed genes. After that, we applied force normalization and reanalyzed. By comparing these two procedures, we observed

significant changes in the significance level of the top differentially expressed genes. This was a difficult task as determining the accuracy of a relevant p-value requires careful implementation.

FORMULATION PROCESS

We find a gene as differentially expressed if the difference or change between two experimental conditions that is observed in read counts or expression levels is statistically significant. One common objective in microarray experiments is to identify a subset of genes that express differentially among different experimental conditions, for example, between HealthySubject and HypertensionPatient. For identifying biological functions or predicting specific therapeutic outcomes, our goal is to determine the underlying relationship between the Healthy Subject's gene signature versus Hypertension Patient's gene signatures. In microarray data analysis, to enhance the relationships among the underlying biological structures or to improve prediction accuracy of clinical outcomes, selection of a subset of genes has been an important issue due to the complexity in studying a large number of genes in an experiment. Hence, the selection of differentially expressed genes is a two-step process. First, we have to select an appropriate test statistic to compute the P-value and the genes will be ranked according to their P-values as evidence of differential expression. Secondly, we have to assign a significance level to determine a cutoff threshold from the P-values as per the study objective.

MODEL DESCRIPTION

In this project, we used the limma (Linear Models for Microarray Analysis) R package with Benjamini & Hochberg (False discovery rate) and limma precision weights. We considered the commonly used t statistics to compute the P-values and adjusted P-values for gene ranking. Later, we applied force normalization to reanalyze the top differentially expressed genes.

ANALYSIS

To conduct the first part of differentially expressed genes analysis, we considered Benjamini & Hochberg (False discovery rate) and limma precision weights in GEO2R and we found the below results:

ID	adj.P.Val	P.Value	t	UniGene_ID	Description
CNN1	0.00000235	5.35e-08	6.445	Hs.465929	Calponin 1, basic, smoot…
NOTCH4	0.00000289	2.35e-07	6.024	Hs.436100	Notch 4
BGLAP	0.00000289	2.97e-07	5.957	Hs.654541	Bone gamma-carboxyglu…
SFTBP	0.00000289	3.26e-07	5.931	Hs.512690	Surfactant protein B
ALB	0.00000289	3.94e-07	5.876	Hs.418167	Albumin
NOS3	0.00000289	3.72e-07	5.893	Hs.707978	Nitric oxide synthase 3 (e…
THY1	0.00000778	1.41e-06	5.51	Hs.644697	Thy-1 cell surface antigen
PROM1	0.00000694	1.10e-06	5.581	Hs.614734	Prominin 1
TEK	0.00001026	2.32e-06	5.367	Hs.89640	TEK tyrosine kinase, end…
KRT14	0.00001219	3.60e-06	5.239	Hs.654380	Keratin 14
NES	0.00001026	2.33e-06	5.365	Hs.527971	Nestin
ADIPOQ	0.00001219	3.58e-06	5.241	Hs.80485	Adiponectin, C1Q and col…
COL1A1	0.00001102	2.76e-06	5.317	Hs.172928	Collagen, type I, alpha 1
ENO2	0.00001667	6.66e-06	5.059	Hs.511915	Enolase 2 (gamma, neur…
GATA4	0.00001585	5.04e-06	5.141	Hs.243987	GATA binding protein 4
NT5E	0.00001611	5.86e-06	5.097	Hs.153952	5'-nucleotidase, ecto (CD…
MYH6	0.00001667	7.35e-06	5.03	Hs.278432	Myosin, heavy chain 6, c…
MAP2	0.00001611	5.54e-06	5.114	Hs.368281	Microtubule-associated p…
VWF	0.00001667	6.92e-06	5.048	Hs.440848	Von Willebrand factor
KIT	0.00001667	7.58e-06	5.021	Hs.479754	V-kit Hardy-Zuckerman 4 …
MKI67	0.00003106	1.55e-05	4.809	Hs.689823	Antigen identified by mon…
CAV3	0.00002429	1.16e-05	4.896	Hs.98303	Caveolin 3
ACTA2	0.00005379	3.30e-05	4.584	Hs.500483	Actin, alpha 2, smooth m…
KDR	0.00004231	2.31e-05	4.691	Hs.479756	Kinase insert domain rec…
CD34	0.00004605	2.72e-05	4.642	Hs.374990	CD34 molecule

Figure 1: Top Differentially Expressed Genes

We observed from figure 1 that it ranked the genes according to their P-values as evidence of differential expression. To get these top genes, we considered FDR and limma precision weights without any force normalization. And considering the same, we noticed from the Venn diagram and Histogram (Figure 2) below that there were 35 significant genes which were differentially expressed, and 9 genes were not significant. We used 0.05 as significance level to determine a cutoff threshold from the adjusted P-values.

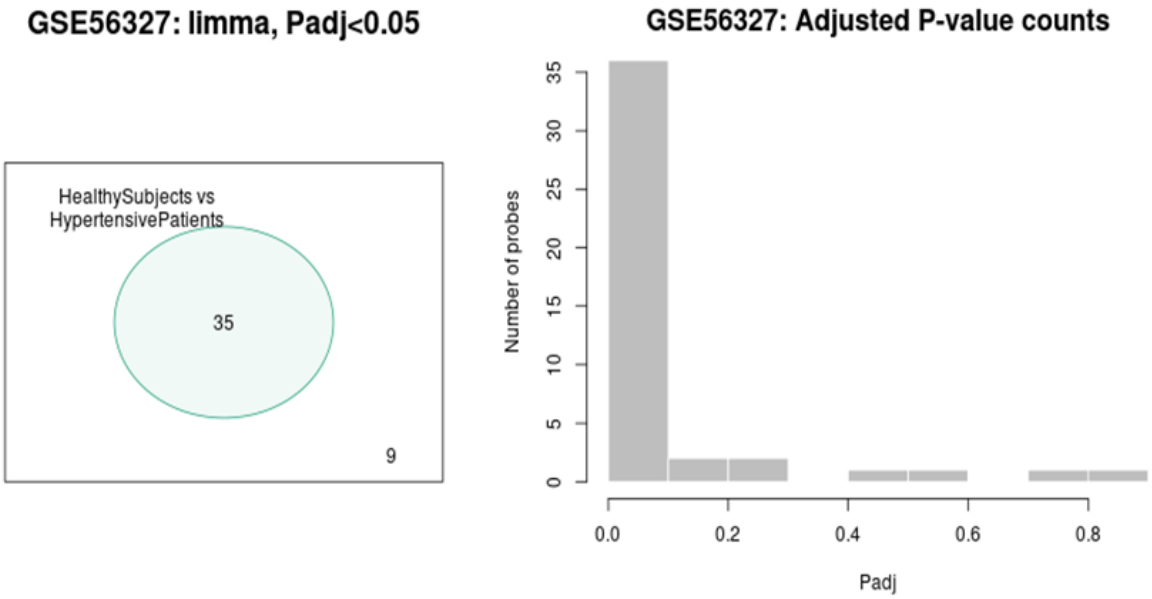

Figure 2: Venn diagram and Histogram for Adjusted P-value counts

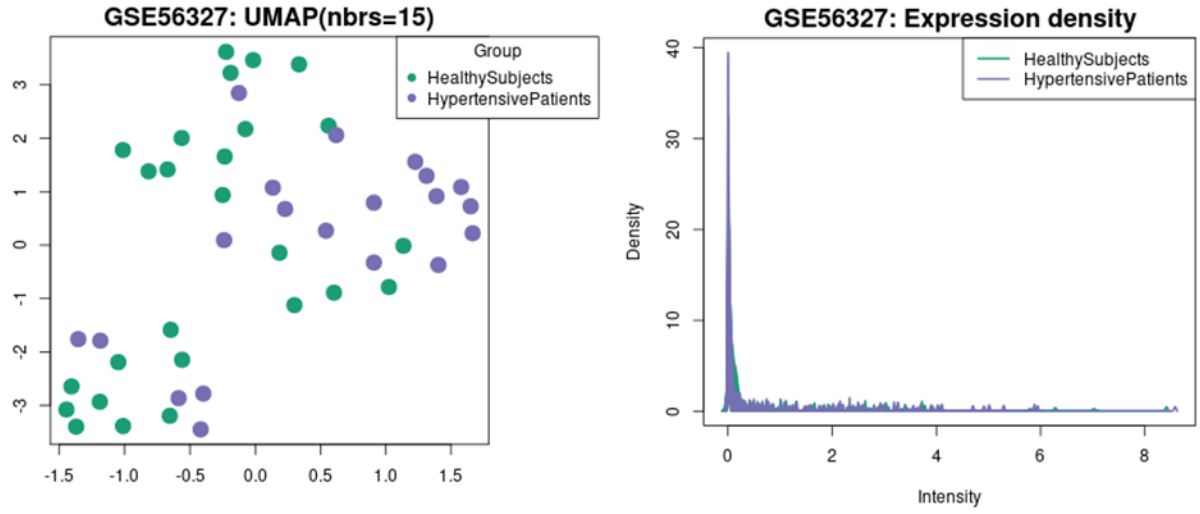

Figure 3: Uniform Manifold Approximation and Projection (UMAP) for visualizing how Samples are related to each other and Expression Density plot for viewing the distribution of the values of the selected Samples. Different colors represent different groups

We examined from the above UMAP plot (Figure 3) that the nearest neighbors were not calculated clearly and by observing the Expression Density plot (Figure 3), we analyzed the distribution of the values for the selected samples. In addition, we considered a boxplot (figure 4) for two different groups to check the distribution whether it will be useful or not for determining the readiness of the selected Samples for differential expression analysis. Generally, median-centered values

indicate that the data are normalized and cross-comparable. But from figure 4, we observed significant variations in medians for both of the groups.

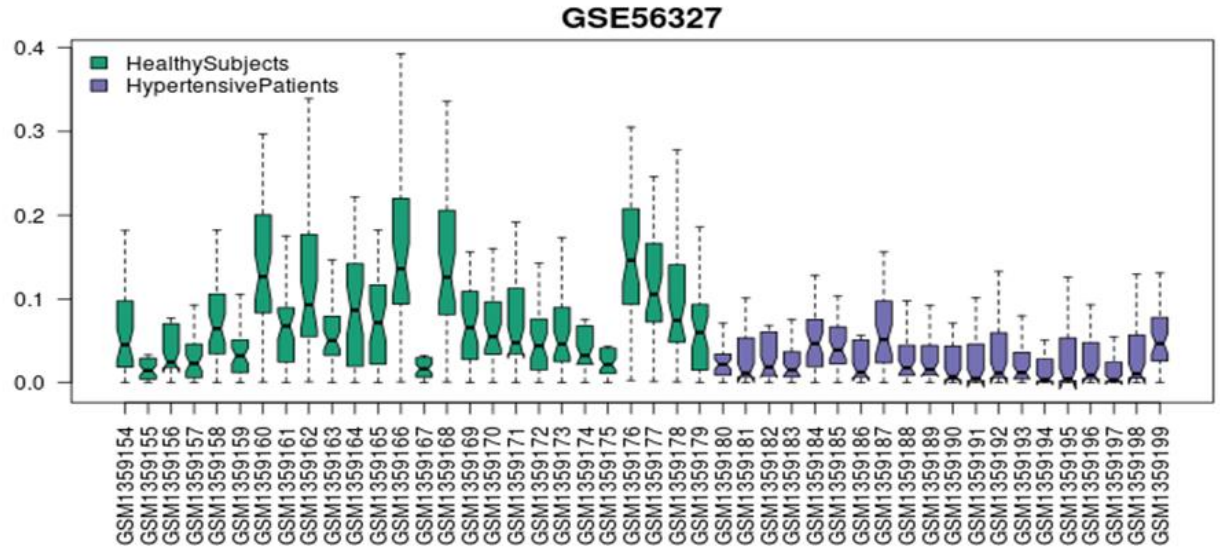

Figure 4: Boxplot for viewing the distribution of the values of the selected Samples. Two different colors represent the two different groups.

We considered both q-q plot and Mean-variance trend plot (Figure 5) to observe the quality of the limma test results and the mean-variance relationship of the expression data after fitting a linear model. From the q-q plot, we noticed that the points didn't lie along a straight line but ideally the points should lie along a straight line. Hence, we can say that the values for moderated t-statistics computed during the test didn't follow their theoretically predicted distribution. And from the Mean-variance trend plot, we noticed that an expression data was located far from the red line which is the mean-variance trend approximation that can be considered during differential gene expression analysis.

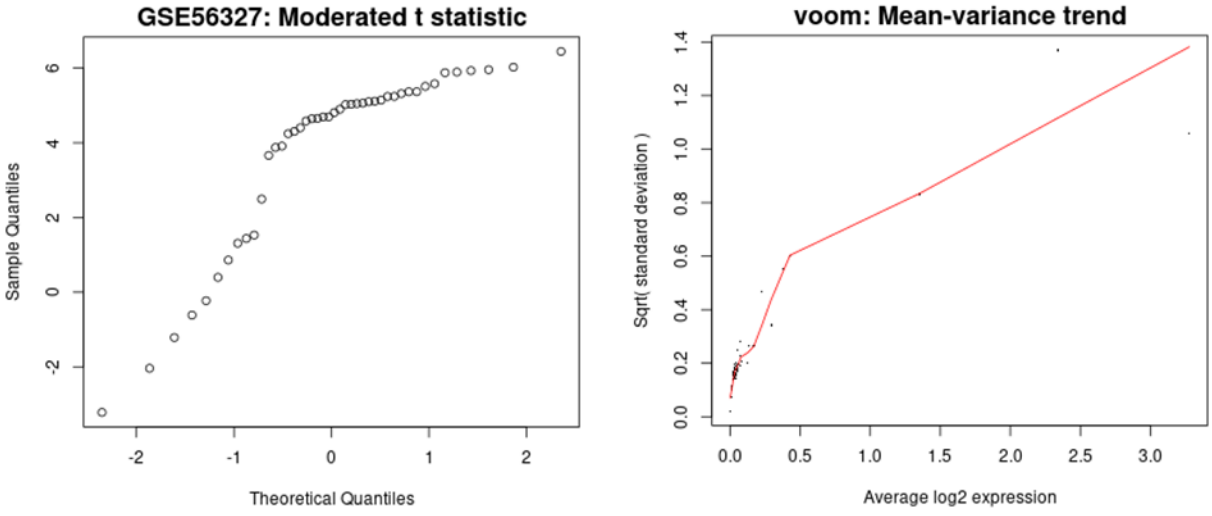

Figure 5: Moderated t-statistic quantile-quantile (q-q) plot for assessing the quality of the limma test results and Mean-variance trend for checking the mean-variance relationship of the expression data

For the second part of our analysis, we decided to apply force normalization to reanalyze the top differentially expressed genes after analyzing the results from the first part and we found the below results:

ID	adj.P.Val	P.Value	t	UniGene_ID	Description
NES	0.00000324	7.37e-08	-6.42	Hs.527971	Nestin
TEK	0.00004088	1.86e-06	-5.47	Hs.89640	TEK tyrosine kinase, end…
ALPL	0.00005045	3.44e-06	5.29	Hs.75431	Alkaline phosphatase, liv…
ST3GAL2	0.00099603	9.05e-05	4.30	Hs.368611	ST3 beta-galactoside alp…
CAV3	0.00162202	1.84e-04	-4.07	Hs.98303	Caveolin 3
CD3E	0.0235802	4.82e-03	-2.96	Hs.3003	CD3E molecule, epsilon (…
MAP2	0.00183301	2.50e-04	-3.98	Hs.368281	Microtubule-associated p…
ALB	0.00568988	1.03e-03	-3.51	Hs.418167	Albumin
KDR	0.0048416	7.70e-04	-3.61	Hs.479756	Kinase insert domain rec…
PTPRC	0.09433014	3.00e-02	2.24	Hs.654514	Protein tyrosine phosphat…
CXCR4	0.17386841	6.72e-02	-1.88	Hs.593413	Chemokine (C-X-C motif)…
ITGAM	0.26637111	1.42e-01	1.50	Hs.172631	Integrin, alpha M (comple…
CX3CR1	0.42116114	2.97e-01	1.06	Hs.78913	Chemokine (C-X3-C moti…
NT5E	0.05377335	1.47e-02	2.54	Hs.153952	5'-nucleotidase, ecto (CD…
CD14	0.94379913	8.90e-01	1.40e-01	Hs.163867	CD14 molecule
ENO2	0.19802727	9.00e-02	-1.73	Hs.511915	Enolase 2 (gamma, neur…
KRT14	0.13441749	4.58e-02	-2.05	Hs.654380	Keratin 14
NKX2-5	0.02928872	6.66e-03	-2.84	Hs.54473	NK2 transcription factor r…
NOS3	0.0473261	1.18e-02	-2.62	Hs.707978	Nitric oxide synthase 3 (e…
POU5F1	0.19250706	8.31e-02	-1.77	Hs.249184	POU class 5 homeobox 1…
CD68	0.93204405	8.47e-01	-1.94e-01	Hs.647419	CD68 molecule
ALDH1	0.79222291	6.30e-01	4.85e-01	Hs.76392	Aldehyde dehydrogenase…
SFTBP	0.21127673	1.01e-01	-1.67	Hs.512690	Surfactant protein B
VWF	0.19250706	7.98e-02	1.79	Hs.440848	Von Willebrand factor
PECAM1	0.26637111	1.51e-01	1.46	Hs.514412	Platelet/endothelial cell a…

Figure 6: Top Differentially Expressed Genes after applying force normalization

We observed from figure 6 that it ranked the genes according to their P-values as evidence of differential expression. And in this part, we considered force normalization along with FDR and limma precision weights to get those top differentially expressed genes. Considering the same, we noticed from the Venn diagram and Histogram (Figure 7) below that the number of significant genes decreased. More specifically, there were 11 significant genes which were differentially expressed, and 33 genes were not significant. We used 0.05 as significance level to determine a cutoff threshold from the adjusted P-values.

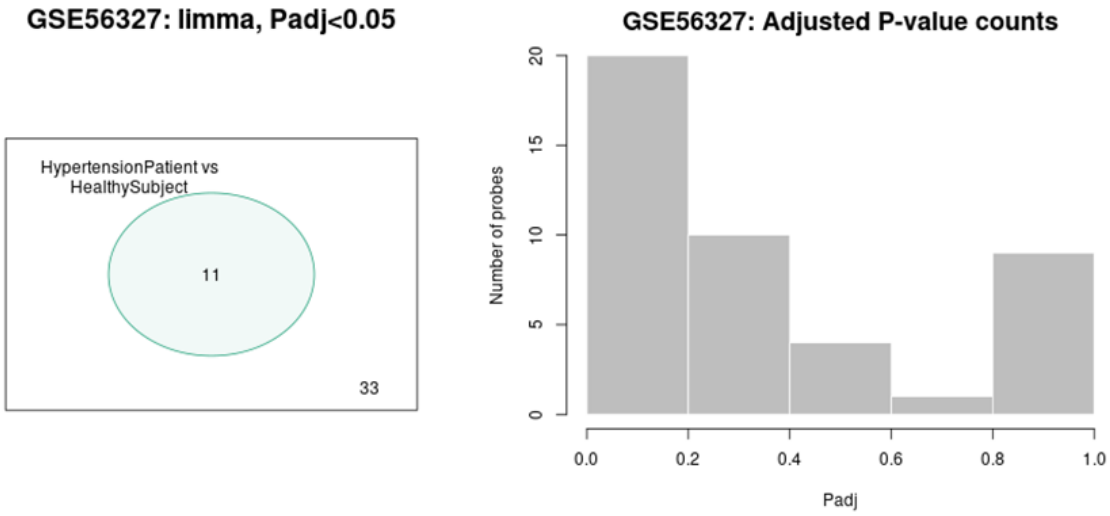

Figure 7: Venn diagram and Histogram for Adjusted P-value counts after force normalization

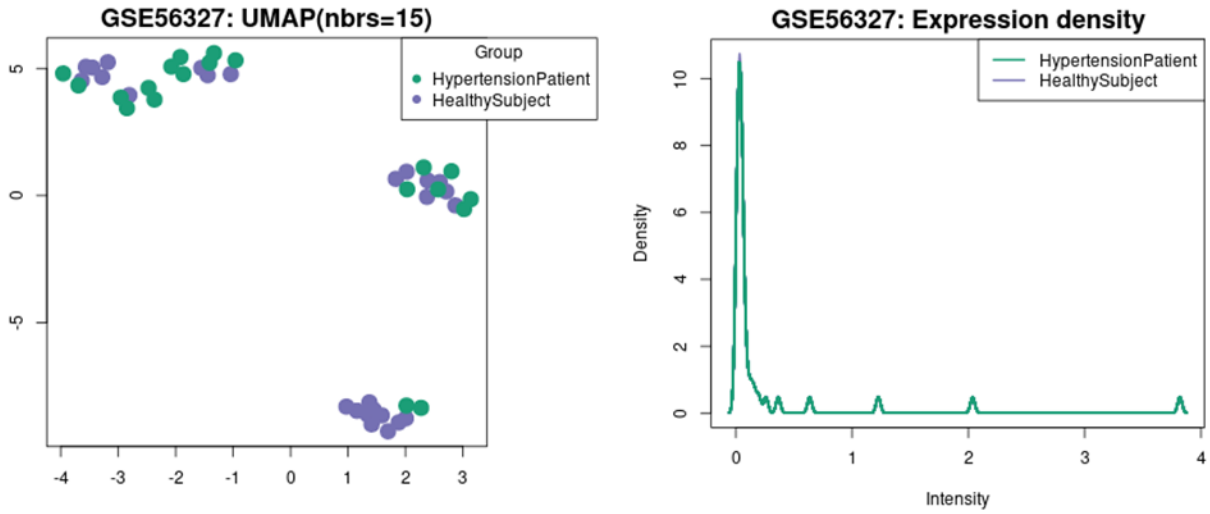

Figure 8: Uniform Manifold Approximation and Projection (UMAP) and Expression Density plot after force normalization

After applying force normalization, we examined from the above UMAP plot (Figure 8) that the clusters were clearly classified and by observing the Expression Density plot (Figure 8), we analyzed the distribution of the values for the selected samples. In addition, we considered the boxplot (figure 9) again for two different groups to check the distribution whether it fixed the variations of medians or not. And finally, we found median-centered values which indicate that the data are now normalized and cross-comparable.

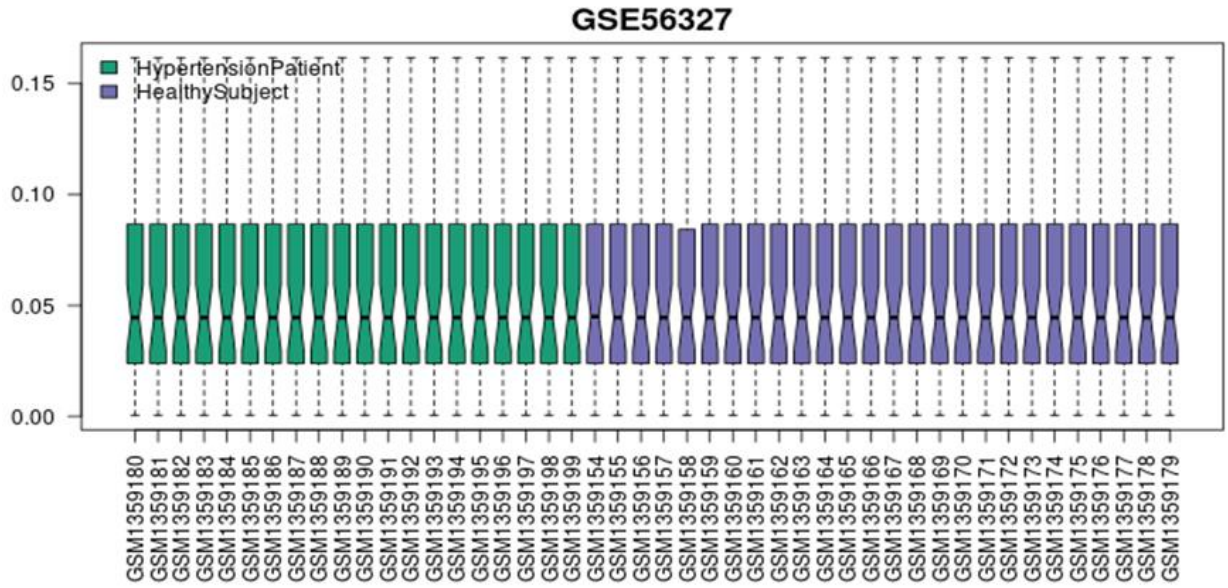

Figure 9: Boxplot for viewing the distribution of the values of the selected Samples after force normalization

We ran both q-q plot and Mean-variance trend plot (Figure 10) again after considering force normalization to observe the quality of the limma test results and the mean-variance relationship of the expression data after fitting a linear model. And from the q-q plot, we noticed that the points did lie along a straight line as per expectation. Hence, we can say that the values for moderated t-statistics computed during the test followed their theoretically predicted distribution. And from the Mean-variance trend plot, we also noticed that the expression data were located close to the red line which is the mean-variance trend approximation that can be considered during differential gene expression analysis.

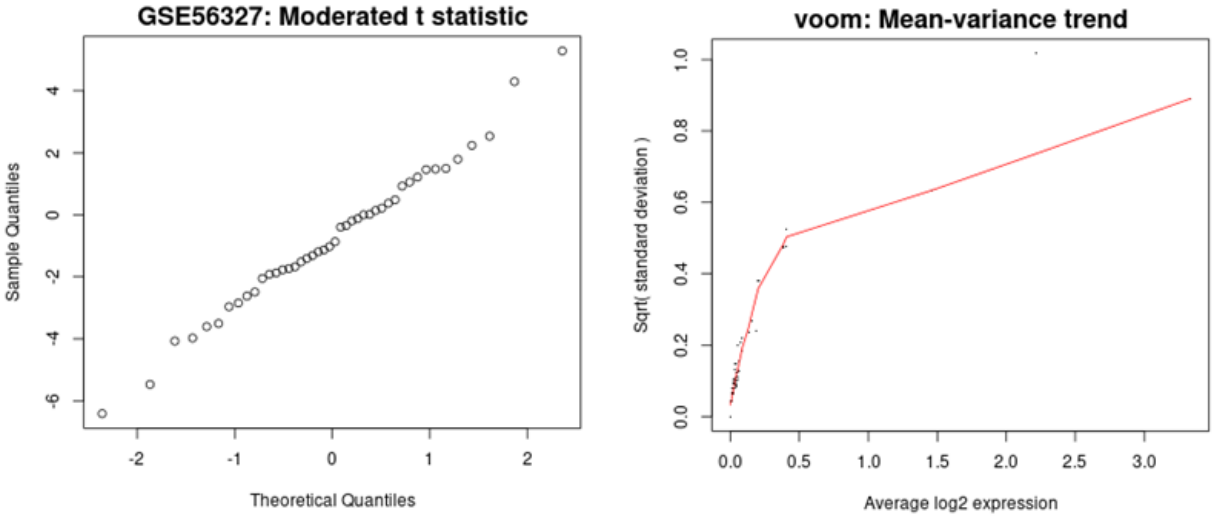

Figure 10: Moderated t-statistic quantile-quantile (q-q) plot and Mean-variance trend after force normalization

RESULTS

We split our projects into two parts. Such as- we conducted our first analysis without the force normalization to see the readiness of the gene expression data for the analysis and we found several inconsistencies by plotting the several plots named Venn diagram, UMAP plot, q-q plot, boxplot etc. We tried to bring the differences or changes in one frame (Figure 11) to compare the two experimental outcomes for differential gene expression analysis. After comparing the two different situations in terms of force normalization, we found the appropriate approach that we must apply force normalization on the given dataset for getting the best differentially expressed genes.

DISCUSSION

After analyzing the genes from 26 healthy subjects and 20 hypertensive patients using GEO2R, we came to a discussion that the samples were not suitable for differential expression analysis by considering the Benjamini & Hochberg (False discovery rate) and limma precision weights only. But we found the same

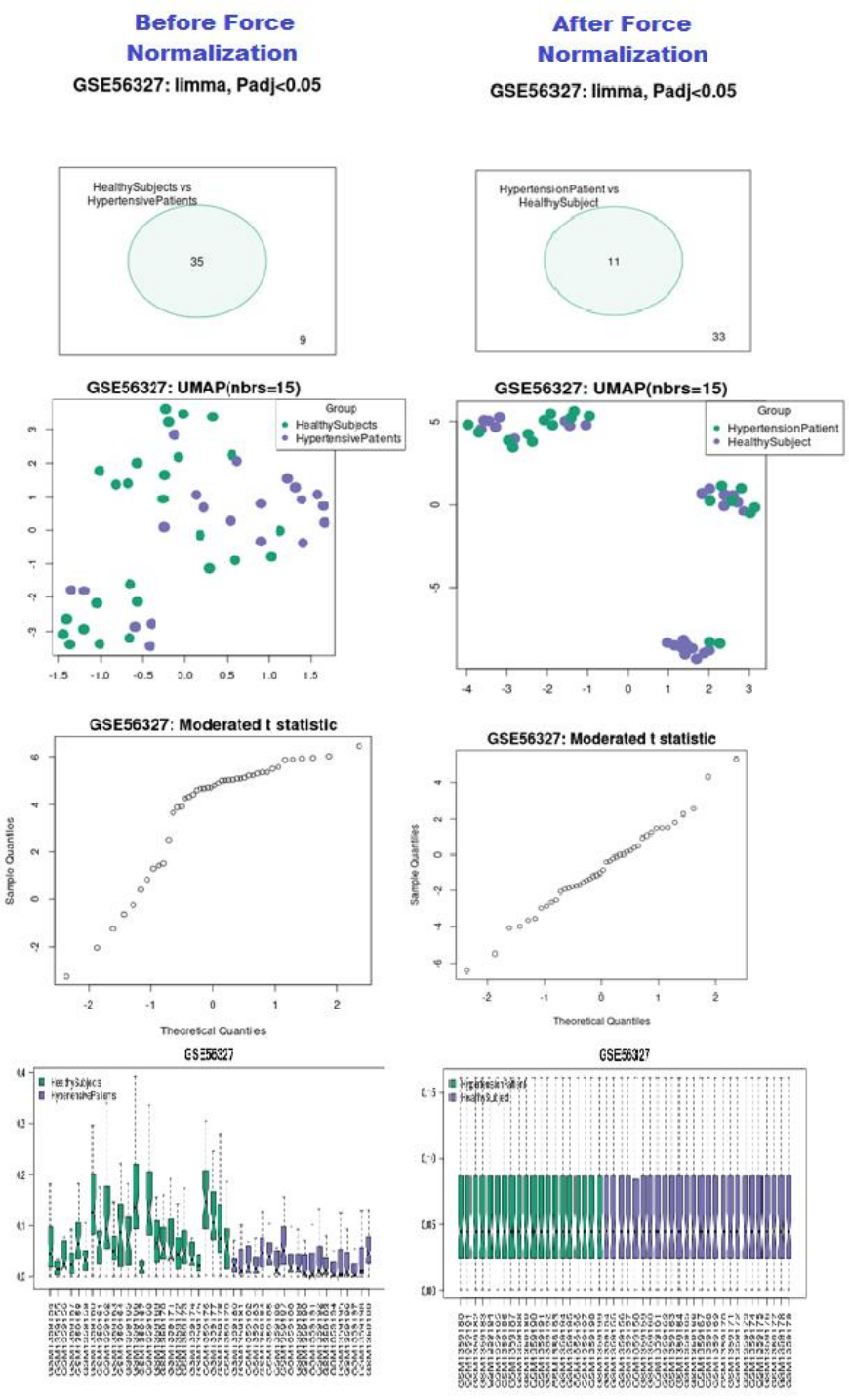

Figure 11: Comparing the results based on applying Force Normalization

samples suitable for differential expression analysis after applying the force normalization which is used to apply quantile normalization to the expression data making identical value distribution for all selected Samples along with considering the Benjamini & Hochberg (False discovery rate) and limma precision weights. We didn't try any partial analysis other than the differential gene expression analysis due to the time constraints along with so many submissions. More importantly, one of our group members was missing so it was very difficult to conduct further analysis within a very short period of time.

CONCLUSION

After this extensive effort conducted while analyzing the data produced from the tests conducted as well as thorough discussions and testing various theories, we decided that the samples were not suitable for either differential expression analysis. This was done by enforcing the False discovery rate (FDR) and limma precision weights. Using FDR, we were able to produce a list of the Top differentially expressed genes, finding that the greater majority of the observed genes were differentially expressed.

Other methods of observation were also conducted, with results both positive (as aforementioned) and negative to our research. We concluded that analyzing the data with a q-q plot and using Mean-variance as a trend plot, the model didn't follow their theoretically predicted distribution without force normalization. Overall, we found that our best success was in applying and considering the Benjamini & Hochberg (False discovery rate) and limma precision weights with force normalization. Our re-analysis of the dataset found few discrepancies with the published article.

References

J J Chen, S.-J. W.-A.-J. (2006). Selection of differentially expressed genes in microarray data analysis. *The Pharmacogenomics Journal*, The Pharmacogenomics Journal. From https://www.nature.com/articles/6500412

Leni Moldovan, M. A. (2014). A module of human peripheral blood mononuclear cell transcriptional network containing primitive and differentiation markers is related to specific cardiovascular health variables. *National Library of Medicine*, PubMed.gov. doi:10.1371/journal.pone.0095124

NCBI. (2014, July 21). *GEO Datasets*. From National Library of Medicine: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56327