Project Data Analysis (Part II)

John and Sayed

- Paper reference: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3997360/

- Link to the dataset: https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE56327

- What part of the paper do you attempt to reanalyze?

  We want to analyze 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 genes that are differentially expressed across experimental conditions.

- If only part of the analysis in the paper: Reason why, for example, no data available

  We choose this part of analysis because we have 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 is missing so it was very difficult to conduct further analysis within a very short period of time.

- Describe the experiment and statistical model

  We have analyzed the whole task with GEO2R and RStudio in two different ways. Such as, at first, we have 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 have applied force normalization and reanalyzed. By comparing these two procedures, we have 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.

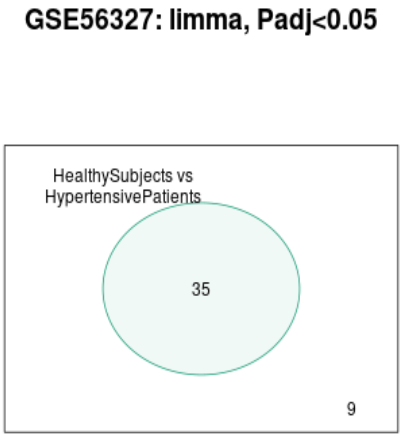
- List all Tests to be done

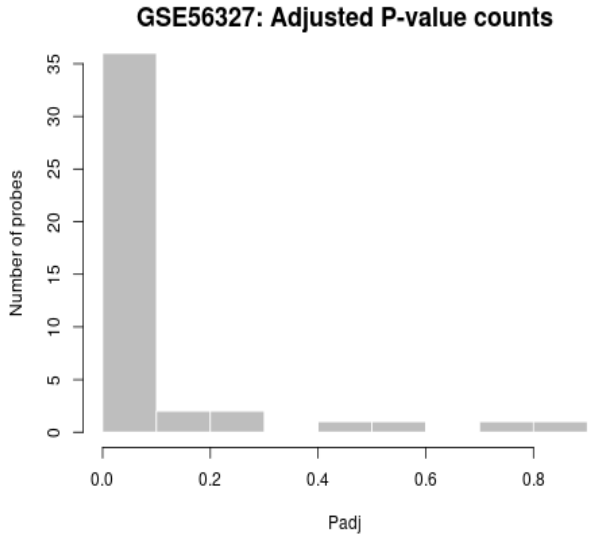
  1. The limma (Linear Models for Microarray Analysis) R package

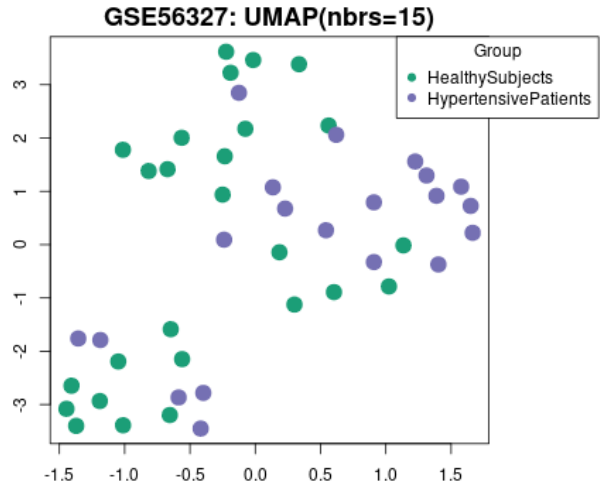
- List the Test Results in tables (not a printout of the program).
  [With Benjamini & Hochberg (False discovery rate) and limma precision weights]

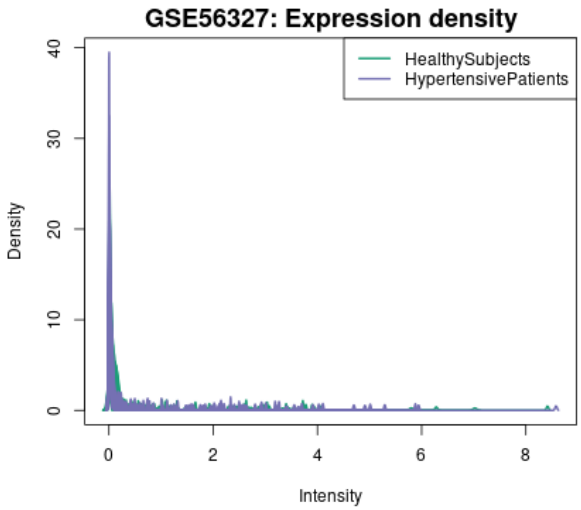
Top Differentially Expressed Genes with Benjamini & Hochberg (False discovery rate) and limma precision weights:

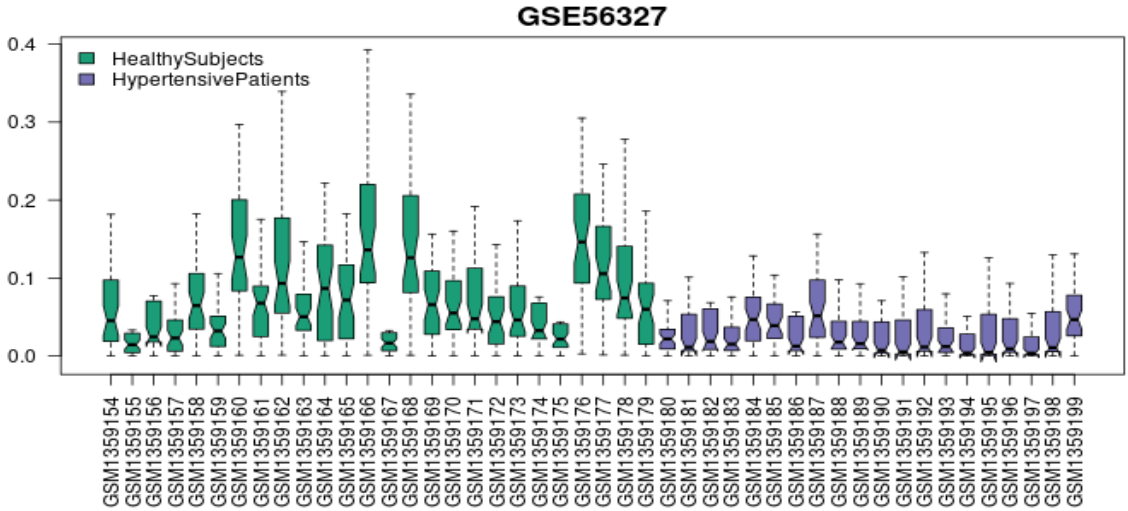
ID	adj.P.Val	P.Value	t	UniGene_ID	Description
CNN1	0.00000235	5.35e-08	6.445	Hs.465929	Calponin 1, basic, smoot…
NOTCH4	0.00000289	2.35e-07	6.024	Hs.436100	Notch 4
BGLAP	0.00000289	2.97e-07	5.957	Hs.654541	Bone gamma-carboxyglu…
SFTBP	0.00000289	3.26e-07	5.931	Hs.512690	Surfactant protein B
ALB	0.00000289	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

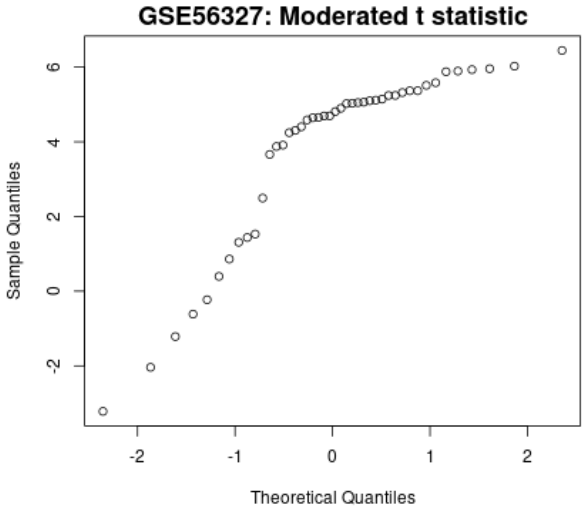


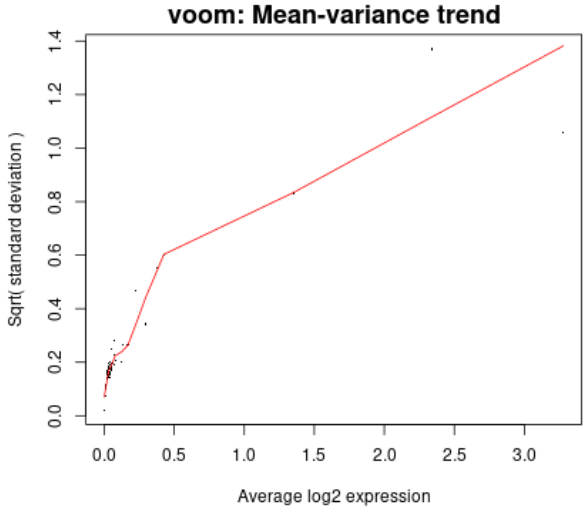



GSE56327: UMAP(nbrs=15)
Group
HealthySubjects
HypertensivePatients


GSE56327: Expression density
HealthySubjects
HypertensivePatients
Density
Intensity


GSE56327
HealthySubjects
HypertensivePatients

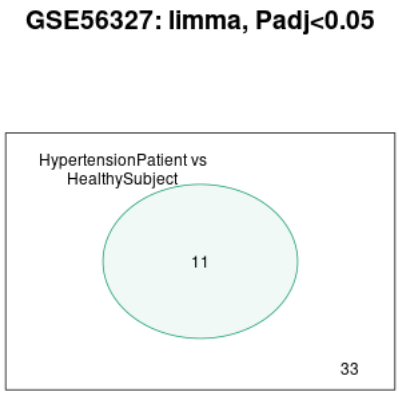

GSE56327: Moderated t statistic
Sample Quantiles
Theoretical Quantiles

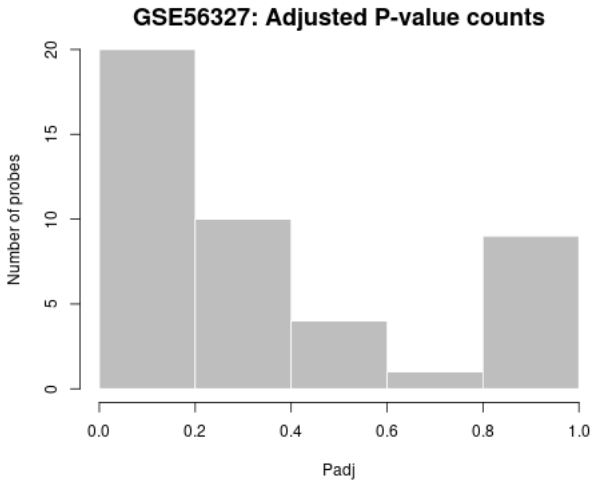

voom: Mean-variance trend
Sqrt(standard deviation)
Average log2 expression

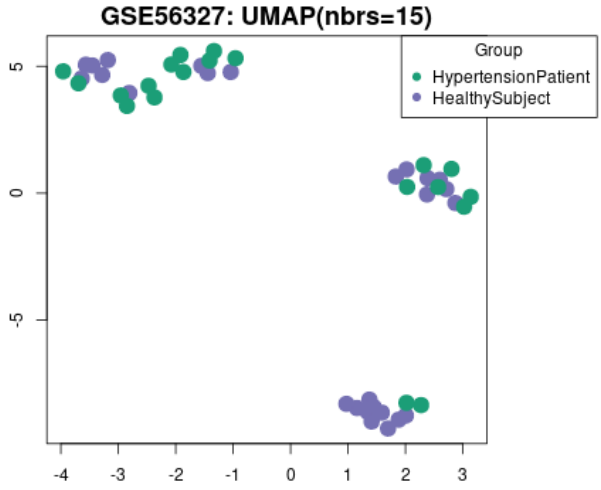
- If not identical to the results of the paper, what else did you try to match the results?

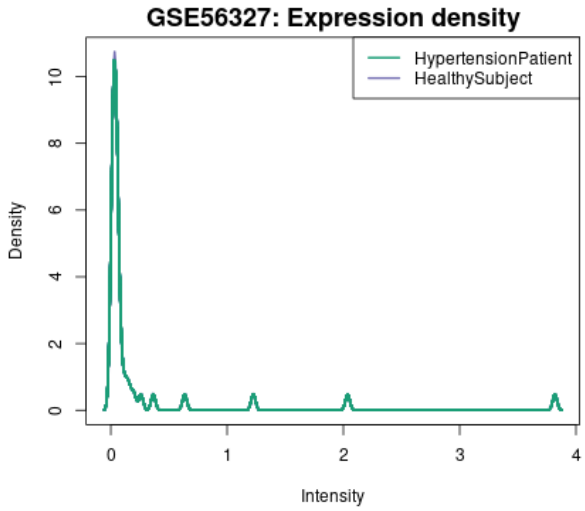
We have applied force normalization to reanalyze the top differentially expressed genes. Results are listed below:

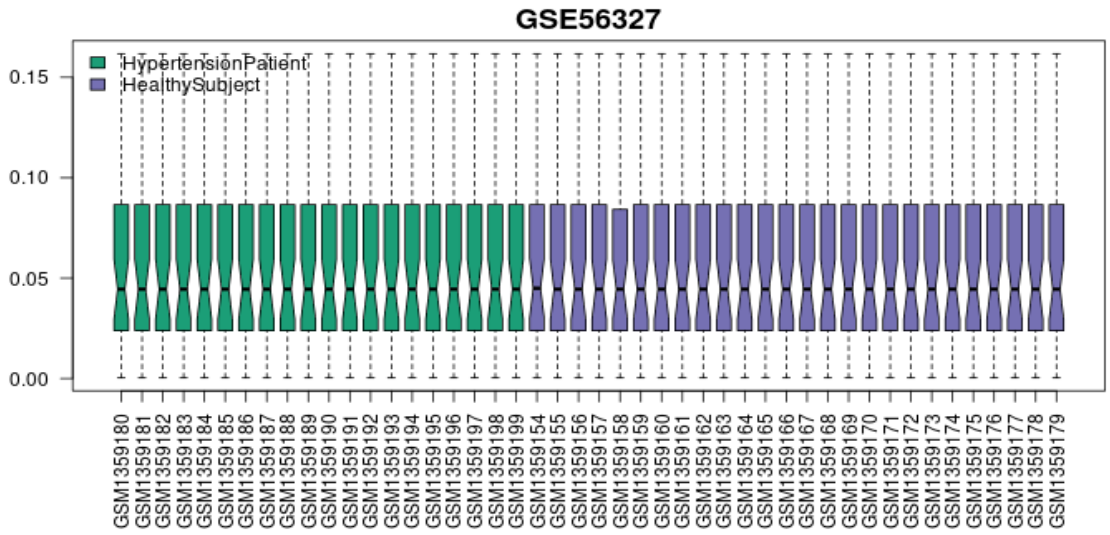
ID	adj.P.Val	P.Value	t	UniGene_ID	Description
NES	0.00000324	7.37e-08	-6.42	Hs.527971	Nestin
TEK	0.00004088	1.86e-06	-5.47	Hs.89640	TEK tyrosine kinase, end…
ALPL	0.00005045	3.44e-06	5.29	Hs.75431	Alkaline phosphatase, liv…
ST3GAL2	0.00099603	9.05e-05	4.30	Hs.368611	ST3 beta-galactoside alp…
CAV3	0.00162202	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…

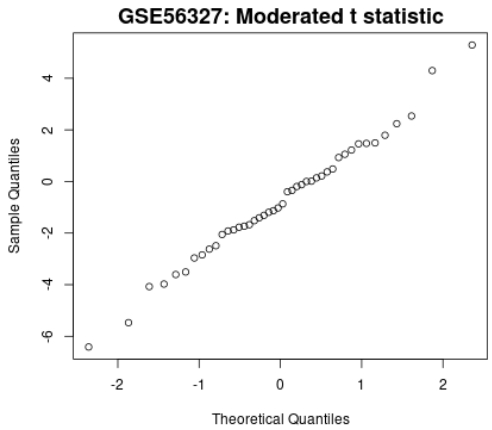


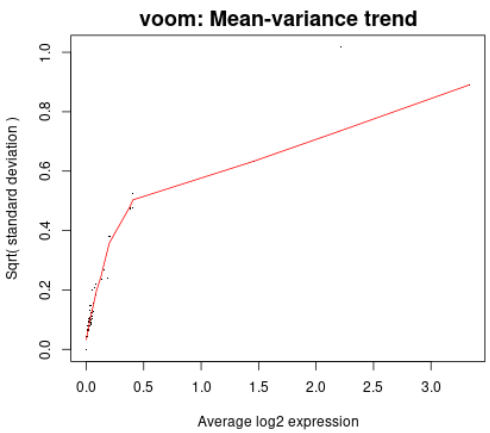



GSE56327: UMAP(nbrs=15)
Group
HypertensionPatient
HealthySubject


GSE56327: Expression density
HypertensionPatient
HealthySubject
Density
Intensity


GSE56327
HypertensionPatient
HealthySubject


GSE56327: Moderated t statistic
Sample Quantiles
Theoretical Quantiles


voom: Mean-variance trend
Sqrt(standard deviation)
Average log2 expression

- Attach the Program code or lab report (summary description of steps to arrive at results).

  The program code has been attached separately. There will be two separate R scripts as we have conducted our analysis in two different ways as we mentioned earlier. After force normalization the number of significant differentially expressed genes has declined.

- Who of the project team did what part of the analysis (load should be equally distributed).

Sayed has drafted the document thoroughly, planned virtual meetings, prompt with communication, Produced the Top Expressed Gene graphs, as well as producing the 2 R scripts.

John has revised the document, laid out the processes, expanded the description, planned virtual meetings, reviewed and sanity checked all graphs produced as well as R codes.