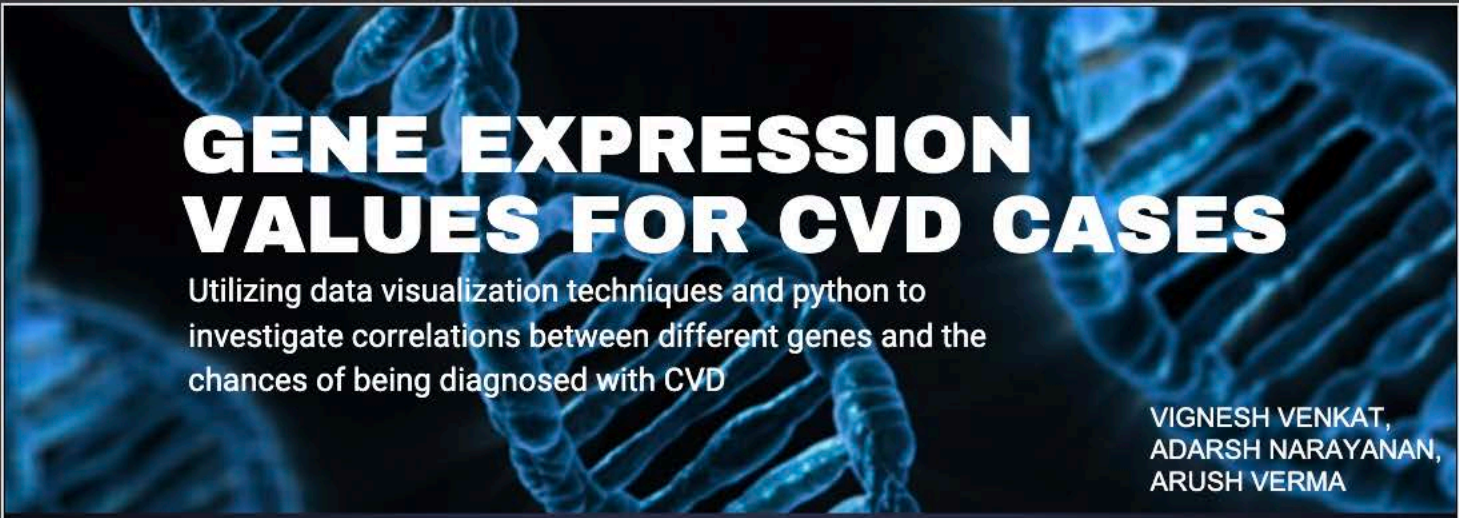


GENE EXPRESSION VALUES FOR CVD CASES

Utilizing data visualization techniques and python to investigate correlations between different genes and the chances of being diagnosed with CVD

VIGNESH VENKAT,
ADARSH NARAYANAN,
ARUSH VERMA

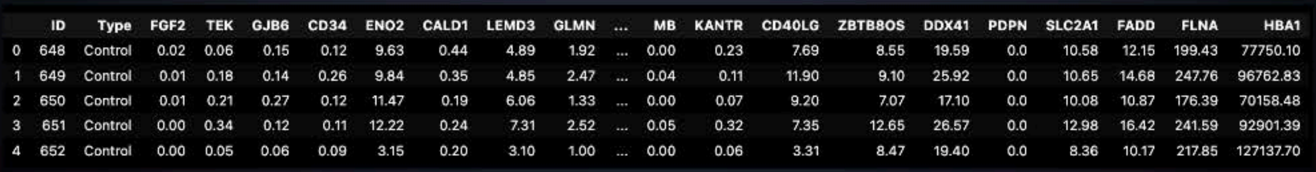
OVERVIEW

According to John Hopkins, more than 250,000 people in the U.S. die every year from medical errors and from other records that number only increases. Around the globe, nearly 73% of global deaths are caused by non-communicable diseases. Focusing on a more common and pressing disease, cardiovascular disease happens to be a leading cause of death in the US.

Using exploratory data analysis and visualization, we tried to find a trend or pattern to show which gene traits in people lead to higher chances in cardiovascular disease. This analysis and visualization showing patterns in gene traits could then be utilized by hospitals and clinics to optimistically aid and assist in diagnosis for those who are at risk of CVD.

Our goal for this datathon was to focus on CVD and identify a pattern linking gene expression values to risk levels of CVD for different types of genes.

THE DATA



	ID	Type	FGF2	TEK	GJB6	CD34	ENO2	CALD1	LEMD3	GLMN	...	MB	KANTR	CD40LG	ZBTB8OS	DDX41	PDPN	SLC2A1	FADD	FLNA	HBA1
0	648	Control	0.02	0.06	0.15	0.12	9.63	0.44	4.89	1.92	...	0.00	0.23	7.69	8.55	19.59	0.0	10.58	12.15	199.43	77750.10
1	649	Control	0.01	0.18	0.14	0.26	9.84	0.35	4.85	2.47	...	0.04	0.11	11.90	9.10	25.92	0.0	10.65	14.68	247.76	96762.83
2	650	Control	0.01	0.21	0.27	0.12	11.47	0.19	6.06	1.33	...	0.00	0.07	9.20	7.07	17.10	0.0	10.08	10.87	176.39	70158.48
3	651	Control	0.00	0.34	0.12	0.11	12.22	0.24	7.31	2.52	...	0.05	0.32	7.35	12.65	26.57	0.0	12.98	16.42	241.59	92901.39
4	652	Control	0.00	0.05	0.06	0.09	3.15	0.20	3.10	1.00	...	0.00	0.06	3.31	8.47	19.40	0.0	8.36	10.17	217.85	127137.70

APPROACH

- Before analysis, researchers at the Robert Wood Johnson Medical School and Institue for Health took the transcriptome of 72 consensual patients for sequencing.
- The sequencing engine used was Illumina.
- The genomes that have had previous literature link them to Cardiovascular disease were used.
- We imported the dataset into a conda environment and pip installed all required packages. Afterwards, we preprocessed our data, using descriptive statistics to understand the data as a whole.
- We then applied a seaborn heatmap visualization to the correlation matrix of the original dataset.
- We noted down the top 5 correlation values for further evaluation.
- We created 21 box plots, based on the 21 genes available in the data cohort, to map the difference in distribution between the case and cohort populations.

SIGNIFICANT RESULTS

LEMD3	0.8
GLMN	0.7
ATP2A2	0.7
FLNA	0.7
GLB6	0.6

CONCLUSION

- Based on our analysis, we were able to plot differences in gene expression based on diagnosis of cardiovascular disease and account for gender differences.
- For a future analysis, we would like to conduct a whole transcriptomic analysis with a larger number of participants.
- A limitation that should be recognized is that large pools of genetic data are not widespread and quite difficult to acquire.
- However, we would like to conduct our own Illumina Sequencing program at a willing hospital to garner more information and potentially be able to conduct Machine Learning and Deep Learning analysis to create a predictive engine for cardiovascular disease.
- The genetic links found at BITS will potentially be the first step in creating a generalizable pipeline available for all heath institutions to assist clinicians with diagnosis.