Differentially expressed genes in metabolic disorders

By: Number Ninjas (Kevin Rouse, Anton Hung, Bofan Chen)

Background

- Using genomic data as a diagnostic tool is a growing field
- Cost and time of sequencing the genome has gone down drastically
- The demand for data scientists to analyze this data has also increased

Objectives

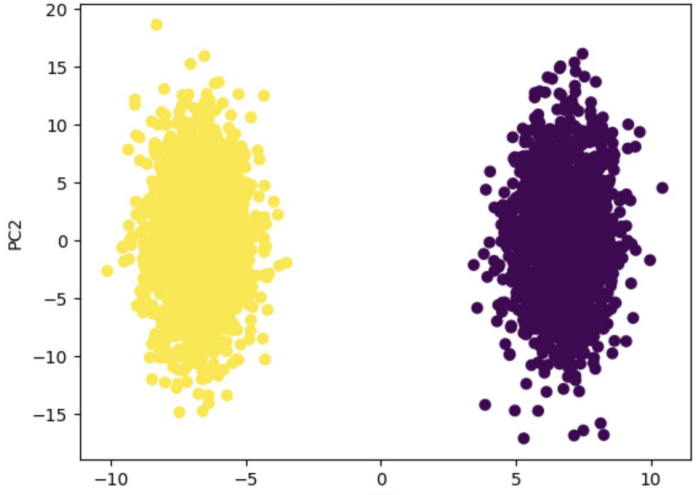
- Data is high dimensional, must reduce dimensions in order to analyze
- Separate new dimensions into two clusters
- Determine differences in gene expression between the two groups
- Use differences in gene expression as a diagnostic tool

Methods - Overview

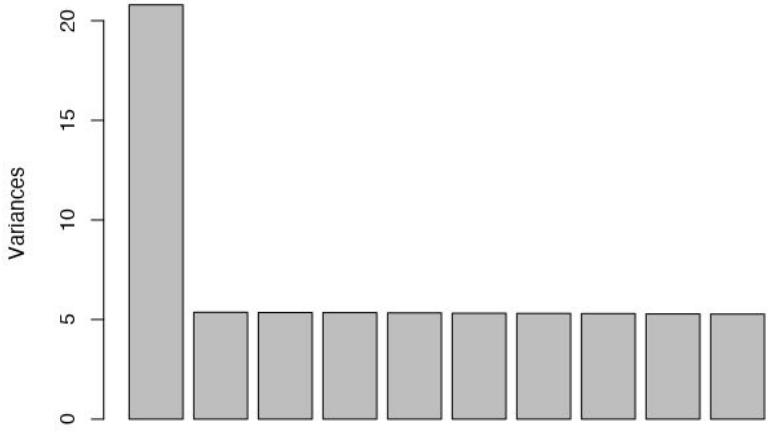
1. Clustering our samples into two distinct groups
    a. PCA (Principal Component Analysis)
    b. K-means Clustering

2. Determining which genes most contributed to driving the clustering pattern.
    a. Statistical Tests
    b. Machine Learning Models

Methods - PCA and K-means Clustering



Plot of first two principal components



Scree plot illustrates that PC 1 is explaining most of the variance

Methods - Statistical Tests

- Determining which genes most contributed to driving the clustering pattern
- Statistical analysis to determine greatest differential expression
  - Normality test: shapiro wilk test
  - Wilcoxon rank sum test
  - DIfferentially expressed genes and p-values
    - 30 genes with p-values equal to 0, and 180 genes with p-value less than 0.01

index	gene	gene_short	t_statistic	p_value
52607	ENSG00000123500.9	ENSG00000123500	54.7654105	0.00000000000
16473	ENSG00000269968.1	ENSG00000269968	54.7654105	0.00000000000
39532	ENSG00000175063.16	ENSG00000175063	54.7654105	0.00000000000
33169	ENSG00000004776.12	ENSG00000004776	-54.7654105	0.00000000000
46518	ENSG00000196616.13	ENSG00000196616	-54.7654105	0.00000000000
8973	ENSG00000143320.8	ENSG00000143320	54.7654105	0.00000000000
11565	ENSG00000165507.8	ENSG00000165507	-54.7654105	0.00000000000
31708	ENSG00000197766.7	ENSG00000197766	-54.7654105	0.00000000000
40628	ENSG00000160180.15	ENSG00000160180	54.7654105	0.00000000000
40630	ENSG00000160182.2	ENSG00000160182	54.7654105	0.00000000000
39026	ENSG00000170373.8	ENSG00000170373	54.7654105	0.00000000000
59788	ENSG00000219928.2	ENSG00000219928	54.7654105	0.00000000000
29515	ENSG00000108821.13	ENSG00000108821	54.7654105	0.00000000000
49551	ENSG00000247627.2	ENSG00000247627	54.7654105	0.00000000000
22608	ENSG00000211896.7	ENSG00000211896	54.7654105	0.00000000000
57636	ENSG00000076554.15	ENSG00000076554	54.7654105	0.00000000000
23982	ENSG00000166803.11	ENSG00000166803	54.7654105	0.00000000000
57671	ENSG00000170323.8	ENSG00000170323	-54.7654105	0.00000000000
12376	ENSG00000138207.13	ENSG00000138207	-54.7654105	0.00000000000
58179	ENSG00000147676.13	ENSG00000147676	54.7654105	0.00000000000
45330	ENSG00000163993.6	ENSG00000163993	54.7654105	0.00000000000
41291	ENSG00000099953.9	ENSG00000099953	54.7654105	0.00000000000
7317	ENSG00000162407.8	ENSG00000162407	-54.7654105	0.00000000000
21228	ENSG00000129514.5	ENSG00000129514	54.7654105	0.00000000000
53659	ENSG00000106541.11	ENSG00000106541	54.7654105	0.00000000000
53662	ENSG00000173467.8	ENSG00000173467	54.7654105	0.00000000000
38015	ENSG00000115414.18	ENSG00000115414	54.7654105	0.00000000000
7973	ENSG00000060718.20	ENSG00000060718	54.7654105	0.00000000000
12979	ENSG00000108001.13	ENSG00000108001	-54.7654105	0.00000000000
35367	ENSG00000119888.10	ENSG00000119888	54.7654105	0.00000000000
51314	ENSG00000204482.10	ENSG00000204482	3.970492261	0.00007172429
55017	ENSG00000274709.1	ENSG00000274709	-3.908757952	0.00009277186
41975	ENSG00000220702.1	ENSG00000220702	3.660862322	0.00025136781
35256	ENSG00000207218.1	ENSG00000207218	3.636738158	0.00027611235
9223	ENSG00000143228.12	ENSG00000143228	3.559751683	0.00037120563
1251	ENSG00000231257.8	ENSG00000231257	3.504205865	0.00045797117
58251	ENSG00000251840.1	ENSG00000251840	-3.44818085	0.00056437601
59703	ENSG00000236521.1	ENSG00000236521	3.440171409	0.00058134588
485	ENSG00000225834.1	ENSG00000225834	3.429245709	0.00060526131
19225	ENSG00000256875.2	ENSG00000256875	-3.352040172	0.00080218378

Methods - Machine Learning Models

- Transpose the data and label two groups based on the labels from PCA
- Built machine learning models to predict the groups
  - Support Vector Machine (SVM)
  - Random Forest
- Extract the list of feature importance

gene_svm	feature_importance
ENSG00000163993.6	0.02050686
ENSG00000175063.16	0.02050686
ENSG00000099953.9	0.02050686
ENSG00000060718.20	0.02050686
ENSG00000160182.2	0.02050686
ENSG00000129514.5	0.02050686
ENSG00000269968.1	0.02050686
ENSG00000076554.15	0.02050686
ENSG00000211896.7	0.02050686
ENSG00000219928.2	0.02050686
ENSG00000108821.13	0.02050686
ENSG00000115414.18	0.02050686
ENSG00000143320.8	0.02050686
ENSG00000119888.10	0.02050686
ENSG00000147676.13	0.02050686
ENSG00000247627.2	0.02050686
ENSG00000106541.11	0.02050686
ENSG00000160180.15	0.02050686
ENSG00000123500.9	0.02050686
ENSG00000173467.8	0.02050685
ENSG00000166803.11	0.02050685
ENSG00000170373.8	0.02050681

first_gene_rf	feature_impor
ENSG00000099953.9	0.04577393
ENSG00000219928.2	0.04036940
ENSG00000170373.8	0.03848634
ENSG00000170323.8	0.03777376
ENSG00000108001.13	0.03641369
ENSG00000076554.15	0.03580378
ENSG00000196616.13	0.03322652
ENSG00000106541.11	0.03195114
ENSG00000247627.2	0.02940439
ENSG00000119888.10	0.02893226
ENSG00000163993.6	0.02831873
ENSG00000108821.13	0.02193270
ENSG00000123500.9	0.02158377
ENSG00000165507.8	0.02048468
ENSG00000115414.18	0.02003211
ENSG00000175063.16	0.01992109
ENSG00000269968.1	0.01940735
ENSG00000143320.8	0.01859965
ENSG00000160180.15	0.01831143
ENSG00000004776.12	0.01711454
ENSG00000162407.8	0.01703521
ENSG00000138207.13	0.01673105
ENSG00000211896.7	0.01648313
ENSG00000166803.11	0.01598615
ENSG00000173467.8	0.01440539
ENSG00000147676.13	0.01110668

second_gene_rf	feature_importance2
ENSG00000115414.18	0.03838103
ENSG00000129514.5	0.03597712
ENSG00000165507.8	0.03422873
ENSG00000108821.13	0.03375778
ENSG00000138207.13	0.03061145
ENSG00000197766.7	0.02886722
ENSG00000004776.12	0.02872587
ENSG00000163993.6	0.02835975
ENSG00000219928.2	0.02603651
ENSG00000166803.11	0.02468839
ENSG00000170323.8	0.02355694
ENSG00000160180.15	0.02301556
ENSG00000147676.13	0.02223527
ENSG00000173467.8	0.02208419
ENSG00000160182.2	0.02200208
ENSG00000076554.15	0.02030044
ENSG00000162407.8	0.02026149
ENSG00000106541.11	0.01901914
ENSG00000170373.8	0.01790600
ENSG00000123500.9	0.01763008
ENSG00000143320.8	0.01719799
ENSG00000060718.20	0.01667461
ENSG00000108001.13	0.01413421
ENSG00000119888.10	0.01411737
ENSG00000196616.13	0.01312335
ENSG00000175063.16	0.01284041
ENSG00000269968.1	0.01259340

SVM's Confusion Matrix

Predicted/ Actual	0	1
0	407	0
1	0	393

Random Forest's Confusion Matrix

Predicted/ Actual	0	1
0	410	0
1	0	390

Results

Genes with p-value of 0

[Statistical Test]

Overlaps

Genes with feature importance >= 0.01

[Machine Learning Models]

stat	ml_random forest_1	stat	ml_random forest_2	stat	ml_svm
ENSG00000123500	ENSG00000099953	ENSG00000123500	ENSG00000115414	ENSG00000123500	ENSG00000163993
ENSG00000269968	ENSG00000219928	ENSG00000269968	ENSG00000129514	ENSG00000269968	ENSG00000175063
ENSG00000175063	ENSG00000170373	ENSG00000175063	ENSG00000165507	ENSG00000175063	ENSG00000099953
ENSG00000004776	ENSG00000170323	ENSG00000004776	ENSG00000108821	ENSG00000004776	ENSG00000060718
ENSG00000196616	ENSG00000108001	ENSG00000196616	ENSG00000138207	ENSG00000196616	ENSG00000160182
ENSG00000143320	ENSG00000076554	ENSG00000143320	ENSG00000197766	ENSG00000143320	ENSG00000129514
ENSG00000165507	ENSG00000196616	ENSG00000165507	ENSG00000004776	ENSG00000165507	ENSG00000269968
ENSG00000197766	ENSG00000106541	ENSG00000197766	ENSG00000163993	ENSG00000197766	ENSG00000076554
ENSG00000160180	ENSG00000247627	ENSG00000160180	ENSG00000219928	ENSG00000160180	ENSG00000211896
ENSG00000160182	ENSG00000119888	ENSG00000160182	ENSG00000166803	ENSG00000160182	ENSG00000219928
ENSG00000170373	ENSG00000163993	ENSG00000170373	ENSG00000170323	ENSG00000170373	ENSG00000108821
ENSG00000219928	ENSG00000108821	ENSG00000219928	ENSG00000160180	ENSG00000219928	ENSG00000115414
ENSG00000108821	ENSG00000123500	ENSG00000108821	ENSG00000147676	ENSG00000108821	ENSG00000143320
ENSG00000247627	ENSG00000165507	ENSG00000247627	ENSG00000173467	ENSG00000247627	ENSG00000119888
ENSG00000211896	ENSG00000115414	ENSG00000211896	ENSG00000160182	ENSG00000211896	ENSG00000147676
ENSG00000076554	ENSG00000175063	ENSG00000076554	ENSG00000076554	ENSG00000076554	ENSG00000247627
ENSG00000166803	ENSG00000269968	ENSG00000166803	ENSG00000162407	ENSG00000166803	ENSG00000106541
ENSG00000170323	ENSG00000143320	ENSG00000170323	ENSG00000106541	ENSG00000170323	ENSG00000160180
ENSG00000138207	ENSG00000160180	ENSG00000138207	ENSG00000170373	ENSG00000138207	ENSG00000123500
ENSG00000147676	ENSG00000004776	ENSG00000147676	ENSG00000123500	ENSG00000147676	ENSG00000173467
ENSG00000163993	ENSG00000162407	ENSG00000163993	ENSG00000143320	ENSG00000163993	ENSG00000166803
ENSG00000099953	ENSG00000138207	ENSG00000099953	ENSG00000060718	ENSG00000099953	ENSG00000170373
ENSG00000162407	ENSG00000211896	ENSG00000162407	ENSG00000108001	ENSG00000162407	
ENSG00000129514	ENSG00000166803	ENSG00000129514	ENSG00000119888	ENSG00000129514	
ENSG00000106541	ENSG00000173467	ENSG00000106541	ENSG00000196616	ENSG00000106541	
ENSG00000173467	ENSG00000147676	ENSG00000173467	ENSG00000175063	ENSG00000173467	
ENSG00000115414		ENSG00000115414	ENSG00000269968	ENSG00000115414	
ENSG00000060718		ENSG00000060718		ENSG00000060718	
ENSG00000108001		ENSG00000108001		ENSG00000108001	
ENSG00000119888		ENSG00000119888		ENSG00000119888	

Results

FDR cutoff of 0.15

Genes related to metabolism

(Ge SX, Jung D & Yao R, Bioinformatics 36:2628–2629, 2020)
(Luo W & Brouwer C, Bioinformatics 29:14:1830–1831, 2013)
(Minoru Kanehisa et al. Nucleic Acids Research, 49:D1:D545–D551, 2021)

Category	Pathway	Enrichment FDR	Genes
	Protein digestion and absorption	0.01	COL11A1 COL1A1 COL10A1
	ECM-receptor interaction	0.06	ADH1B
	AGE-RAGE signaling pathway in diabetic complications	0.06	PLPP3
	Amoebiasis	0.06	ADH1B
	Focal adhesion	0.14	COL1A1 FN1
	Proteoglycans in cancer	0.14	ADH1B
metabolism	Tyrosine metabolism	0.15	PLPP3
metabolism	Fat digestion and absorption	0.15	PLPP3
	Fatty acid degradation	0.15	COL1A1 FN1
metabolism	Pyruvate metabolism	0.15	COL1A1 FN1
metabolism	Ether lipid metabolism	0.15	FABP4
metabolism	Sphingolipid metabolism	0.15	PLPP3
	Reg. of lipolysis in adipocytes	0.15	ADH1B
metabolism	Glycerolipid metabolism	0.15	ADH1B
	Glycolysis / Gluconeogenesis	0.15	ADH1B
metabolism	Retinol metabolism	0.15	ADH1B
metabolism	Drug metabolism	0.15	FABP4
metabolism	Metabolism of xenobiotics by cytochrome P450	0.15	COL1A1 FN1
	PPAR signaling pathway	0.15	COL1A1 FN1
	PI3K-Akt signaling pathway	0.15	COL1A1 FN1

Screening for Future Health Complications

- Early diagnosis enables more effective treatments
- Having doctors adopt genetic sequencing into routine patient check-ups
- Identify differentially expressed genes that could indicate high risk for future health complications

Diagnosis - Distance Scoring

- Sequence a patient’s genetic sequence
- Perform PCA
- Distance scoring to determine nearest cluster
- If they are more similar to the affected group, then the physician can continue to follow up on and look out for symptoms

Limitations and Next Steps

- Only used two principal components
- 180 genes with p-value less than 0.01 in the Wilcoxon Rank Sum Test
- Overfitting of the machine learning models
- Different disease, different genes