Diagnosing Colorectal Cancer with Gene Expression Analysis

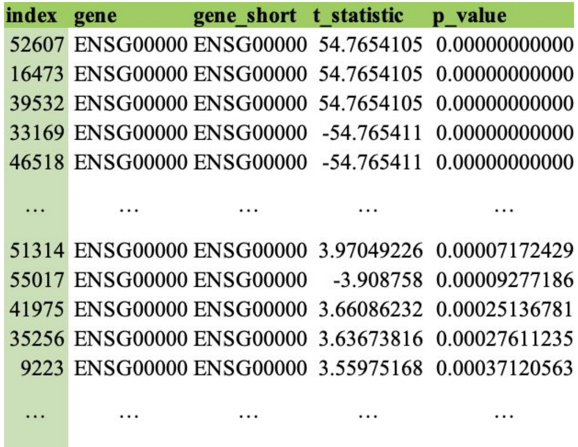
Team Number Ninjas: Bofan Chen, Anton Hung and Kevin Rouse

Genetic Sequencing Background

- Obtaining an individual’s genetic information is requiring increasingly less time and resources
- Traditional diagnosis is difficult to do until patients present with symptoms
- Analyzing a patient’s genetic makeup can reveal disease risk before onset
- Patient’s genome can be sequenced and re-sequenced by their physician over time

Finding the Genes - Statistical Tests

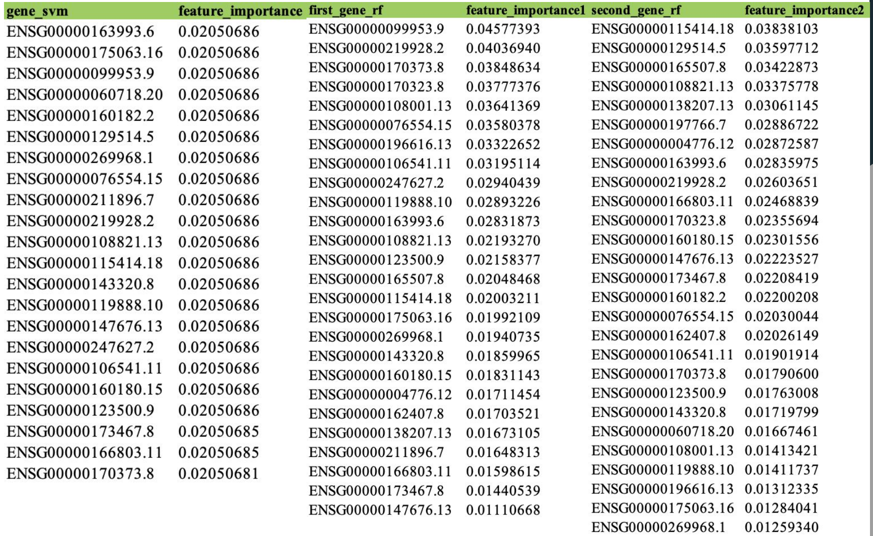
- Identify the differentially expressed genes and p-values
- Data is transposed and labeled for two groups based on PCA
- Statistical analysis to determine greatest differential expression
  - Normality test: shapiro wilk test
  - Wilcoxon rank sum test
  - Bonferroni correction (α = 0.05)
- 30 genes 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	ENSG00000	ENSG00000	54.7654105	0.00000000000
16473	ENSG00000	ENSG00000	54.7654105	0.00000000000
39532	ENSG00000	ENSG00000	54.7654105	0.00000000000
33169	ENSG00000	ENSG00000	-54.765411	0.00000000000
46518	ENSG00000	ENSG00000	-54.765411	0.00000000000
...	...	...	...	...
51314	ENSG00000	ENSG00000	3.97049226	0.00007172429
55017	ENSG00000	ENSG00000	-3.908758	0.00009277186
41975	ENSG00000	ENSG00000	3.66086232	0.00025136781
35256	ENSG00000	ENSG00000	3.63673816	0.00027611235
9223	ENSG00000	ENSG00000	3.55975168	0.00037120563
...	...	...	...	...

Finding the Genes - Machine Learning

- Transpose the data
- Build machine learning models to predict the groups
    - Decision Tree
    - Support Vector Machine (SVM)
    - Random Forest
- Extract the list of feature importance



gene_svm	feature_importance	first_gene_rf	feature_importance1	second_gene_rf	feature_importance2
ENSG00000163993.6	0.02050686	ENSG00000099953.9	0.04577393	ENSG00000115414.18	0.03838103
ENSG00000175063.16	0.02050686	ENSG00000219928.2	0.04036940	ENSG00000129514.5	0.03597712
ENSG00000099953.9	0.02050686	ENSG00000170373.8	0.03848634	ENSG00000165507.8	0.03422873
ENSG00000060718.20	0.02050686	ENSG00000170323.8	0.03777376	ENSG00000108821.13	0.03375778
ENSG00000160182.2	0.02050686	ENSG00000108001.13	0.03641369	ENSG00000138207.13	0.03061145
ENSG00000129514.5	0.02050686	ENSG00000076554.15	0.03580378	ENSG00000197766.7	0.02886722
ENSG00000269968.1	0.02050686	ENSG00000196616.13	0.03322652	ENSG00000004776.12	0.02872587
ENSG00000076554.15	0.02050686	ENSG00000106541.11	0.03195114	ENSG00000163993.6	0.02835975
ENSG00000211896.7	0.02050686	ENSG00000247627.2	0.02940439	ENSG00000219928.2	0.02603651
ENSG00000219928.2	0.02050686	ENSG00000119888.10	0.02893226	ENSG00000166803.11	0.02468839
ENSG00000108821.13	0.02050686	ENSG00000163993.6	0.02831873	ENSG00000170323.8	0.02355694
ENSG00000115414.18	0.02050686	ENSG00000108821.13	0.02193270	ENSG00000160180.15	0.02301556
ENSG00000143320.8	0.02050686	ENSG00000123500.9	0.02158377	ENSG00000147676.13	0.02223527
ENSG00000119888.10	0.02050686	ENSG00000165507.8	0.02048468	ENSG00000173467.8	0.02208419
ENSG00000147676.13	0.02050686	ENSG00000115414.18	0.02003211	ENSG00000160182.2	0.02200208
ENSG00000247627.2	0.02050686	ENSG00000175063.16	0.01992109	ENSG00000076554.15	0.02030044
ENSG00000106541.11	0.02050686	ENSG00000269968.1	0.01940735	ENSG00000162407.8	0.02026149
ENSG00000160180.15	0.02050686	ENSG00000143320.8	0.01859965	ENSG00000106541.11	0.01901914
ENSG00000123500.9	0.02050686	ENSG00000160180.15	0.01831143	ENSG00000170373.8	0.01790600
ENSG00000173467.8	0.02050685	ENSG00000004776.12	0.01711454	ENSG00000123500.9	0.01763008
ENSG00000166803.11	0.02050685	ENSG00000162407.8	0.01703521	ENSG00000143320.8	0.01719799
ENSG00000170373.8	0.02050681	ENSG00000138207.13	0.01673105	ENSG00000060718.20	0.01667461
		ENSG00000211896.7	0.01648313	ENSG00000108001.13	0.01413421
		ENSG00000166803.11	0.01598615	ENSG00000119888.10	0.01411737
		ENSG00000173467.8	0.01440539	ENSG00000196616.13	0.01312335
		ENSG00000147676.13	0.01110668	ENSG00000175063.16	0.01284041
				ENSG00000269968.1	0.01259340

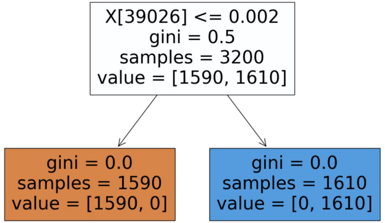
SVM's Confusion Matrix

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

Visualization of Decision Tree



The Genes

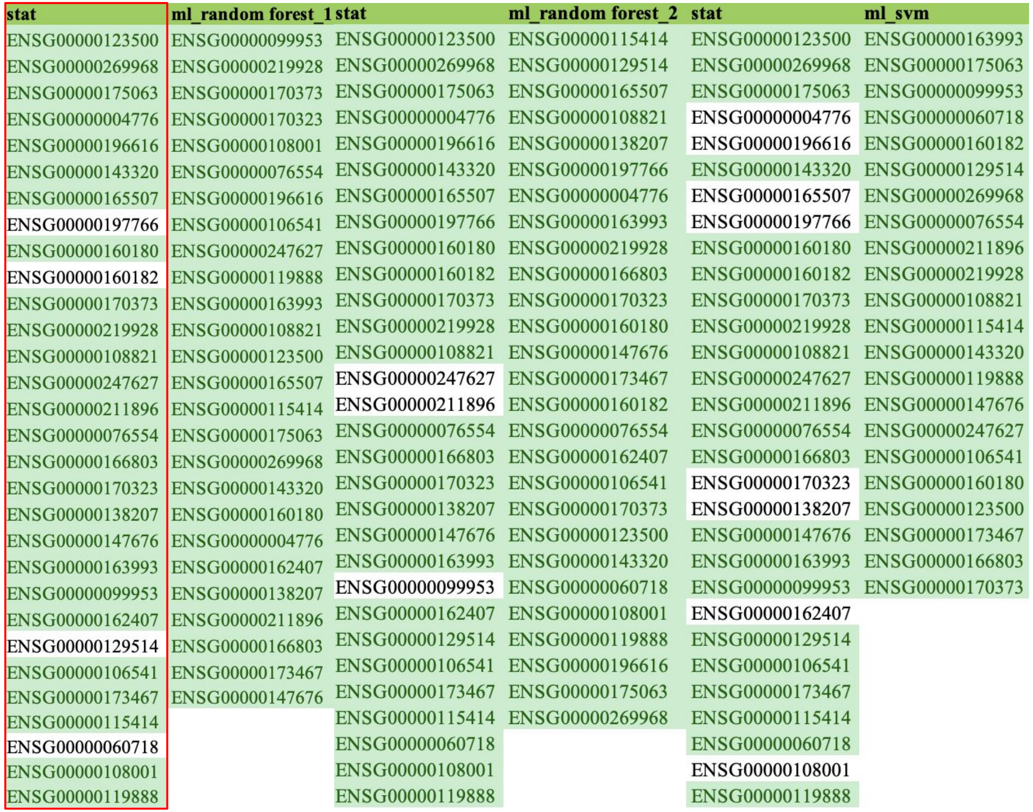
Genes with p-value of 0

[Statistical Test]

Matches with

Genes with feature importance >= 0.01

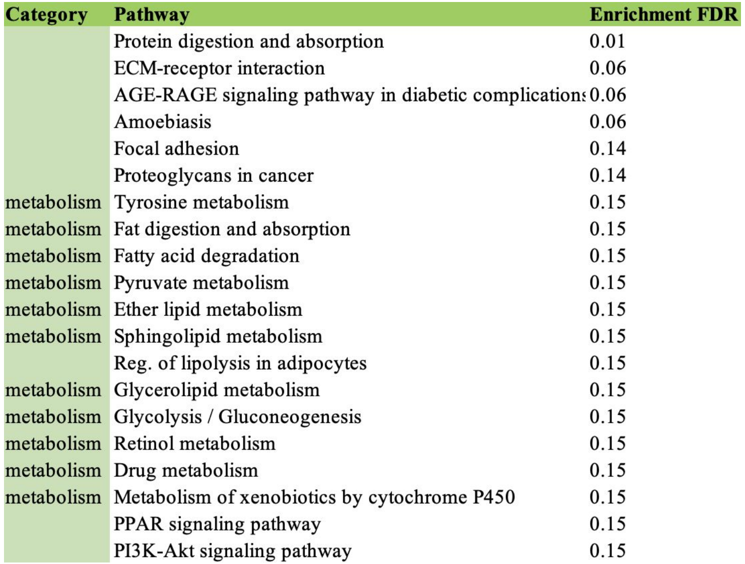
[Machine Learning Models]



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	

Diagnosis - Colorectal Cancer

- Compared highly differentiated genes between the two groups
- First, looked for similar pathways amongst genes using KEGG
- Many pathways related to metabolism, but an ailment could not be directly identified from these pathways
- Then began to compare up and down-regulation of gene expression in our groups to trends found in the literature
- Discovered that the trends in our groups were similar to those of colorectal cancer



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

(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)

Diagnosis - Sick Group Identification

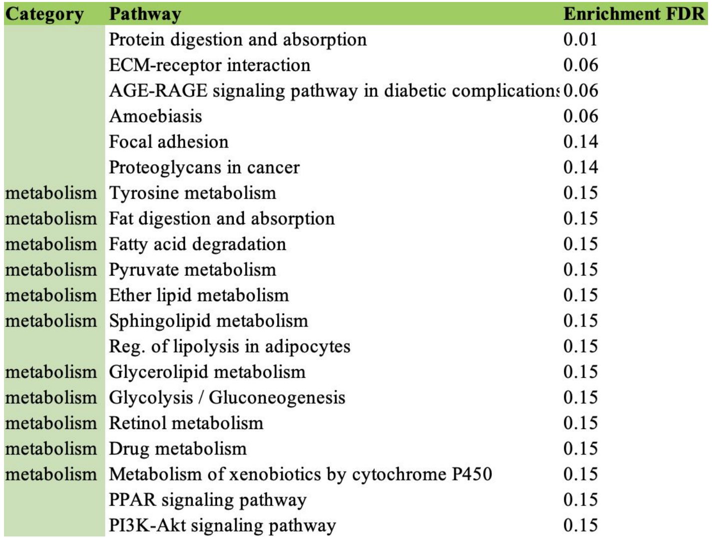
Gene	Relationship between level of gene expression and colorectal cancer (Literature)	Higher Gene Expression (Group 1 or Group 2)
ENSG00000163993(S100P)	Direct[1]	Group 2
ENSG00000004776 (HSPB6)	Indirect[2]	Group 1
ENSG00000123500(COL10A1)	Direct[3]	Group 2
ENSG00000170373(CST1)	Direct[4]	Group 2

Here, the second group would be the group with colorectal cancer given that the gene expression levels match the literature trends

1. Dong, Lei, et al. “Overexpression of S100P Promotes Colorectal Cancer Metastasis and Decreases Chemosensitivity to 5-FU in Vitro.” Molecular and Cellular Biochemistry, vol. 389, no. 1–2, Apr. 2014, pp. 257–64. PubMed, https://doi.org/10.1007/s11010-013-1947-5
2. Ju, Young-Tae, et al. “Decreased Expression of Heat Shock Protein 20 in Colorectal Cancer and Its Implication in Tumorigenesis.” Journal of Cellular Biochemistry, vol. 116, no. 2, Feb. 2015, pp. 277–86. PubMed, https://doi.org/10.1002/jcb.24966
3. Huang, Haipeng, et al. “High Expression of COL10A1 Is Associated with Poor Prognosis in Colorectal Cancer.” OncoTargets and Therapy, vol. 11, 2018, pp. 1571–81. PubMed, https://doi.org/10.2147/OTT.S160196
4. Li, Taiyuan, et al. “Prognostic Significance of Cystatin SN Associated Nomograms in Patients with Colorectal Cancer.” Oncotarget, vol. 8, no. 70, Dec. 2017, pp. 115153–63. PubMed Central, https://doi.org/10.18632/oncotarget.23041

Pathways - Colorectal Cancer Link

- Given that many of the pathways include metabolism, we wanted to relate this to colorectal cancer
- Hypothesized that the samples were collected from cancer cells that metabolize differently than normal cells
- One of the pathways was also related to cancer

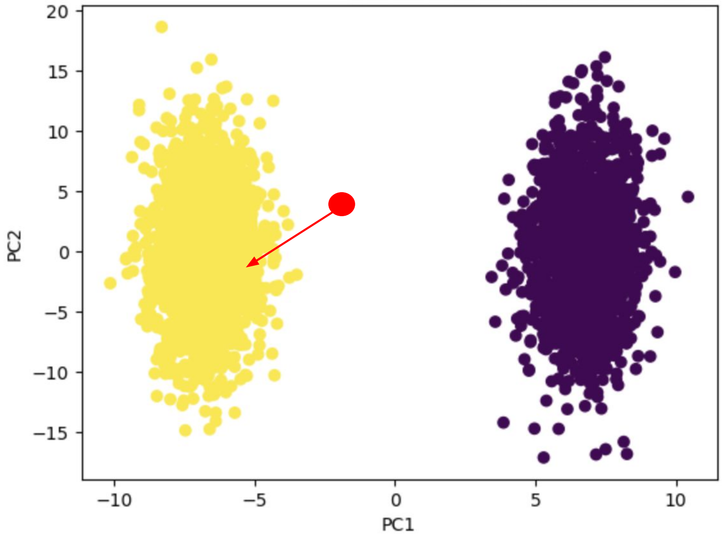


Category	Pathway	Enrichment FDR
	Protein digestion and absorption	0.01
	ECM-receptor interaction	0.06
	AGE-RAGE signaling pathway in diabetic complications	0.06
	Amoebiasis	0.06
	Focal adhesion	0.14
	Proteoglycans in cancer	0.14
metabolism	Tyrosine metabolism	0.15
metabolism	Fat digestion and absorption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

(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)

Diagnostic Application - Machine Learning

- Given a patient’s genetic sequence:
  - Use K-nearest neighbours, the machine learning algorithm
  - Euclidean distance scoring with feature importance to classify the patient to the nearest cluster
  - If they are classified to the affected group, then the physician can continue to follow up on and look out for symptoms



Ethics and Equity

- Cost to the patient
- Discrimination. Underrepresentation of low socioeconomic status
- Confidentiality of personal health information
- The life-long burden of diagnostic results

Limitations and Next Steps

1. Our conclusion about which group is sick or not depended on what we assumed to be the ailment (colorectal cancer).
2. Identification of the ailment using only 4 articles
3. Genes can be implicated in many diseases (potentially a different kind of cancer)
4. Dependent on the quality of literature
5. The data is too perfect

Next step: Robust validation of our findings: Perform RNA-sequencing analysis on patient genomes known to have colorectal cancer, and comparing the similarity of their differentially expressed genes to our differentially expressed genes

Thank you!
Sincerely, Number Ninjas