Predicting Cancer Through Differential and Unsupervised Gene Expression Analysis

Nathalie Bonin, Franco Krepel, William Mukuvi, Julien Wakim

Team Number 14

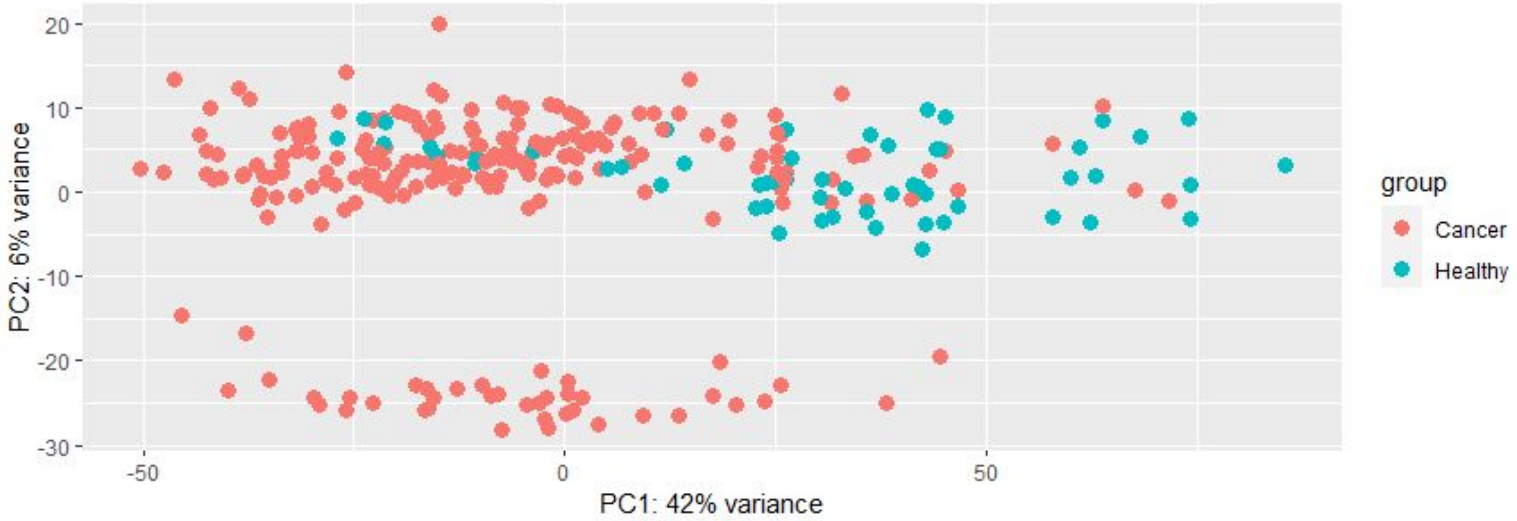
Introduction

- Cancer remains one of the deadliest disease
- Importance of early diagnosis in survival
- Benefits of liquid biopsies
- Efficacy issues

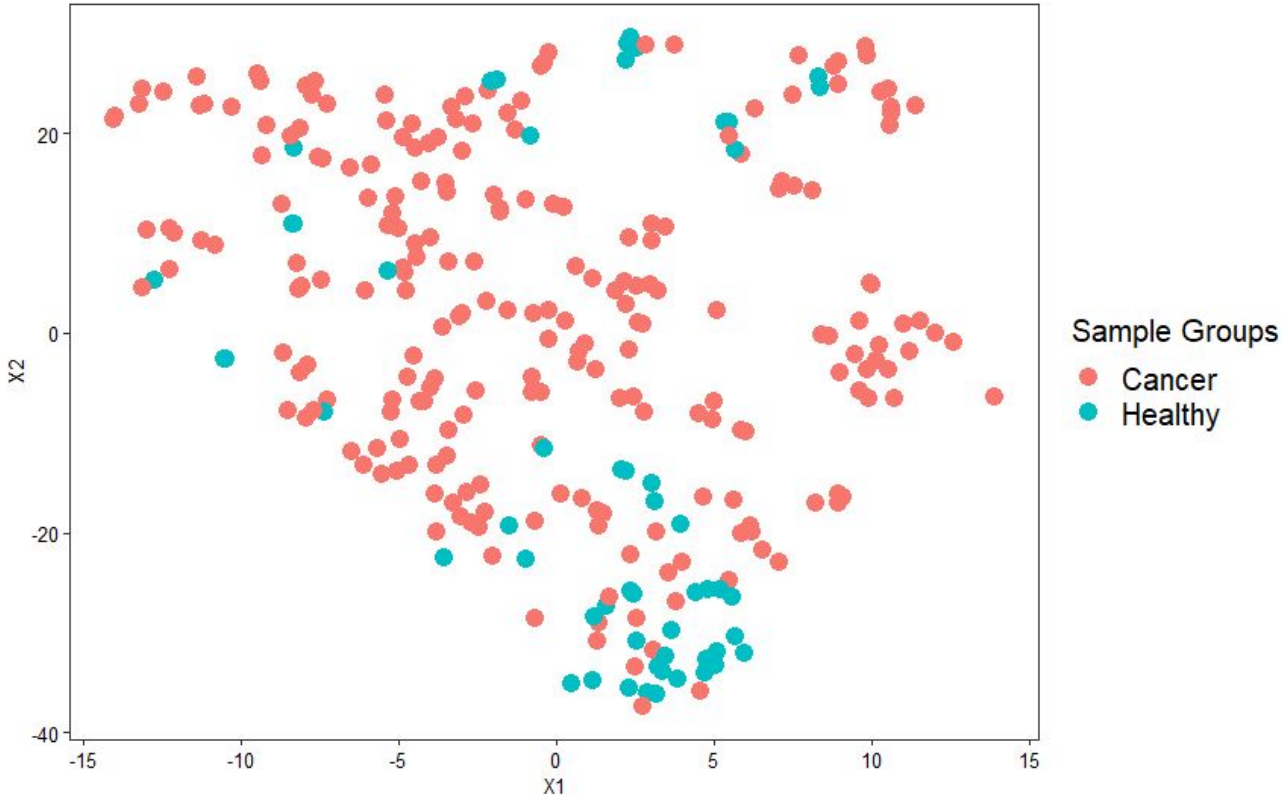
Introduction

- RNA-sequencing data of 285 blood platelet samples
- Samples collected from six different malignant tumors
  - (lung cancer, colorectal cancer, pancreatic cancer, glioblastoma, breast cancer and hepatobiliary carcinomas)
- 230 “Cancer” and 55 “Healthy” samples

PCA Plot


20
10
0
-10
-20
-30
PC2: 6% variance
-50
0
50
PC1: 42% variance
group
Cancer
Healthy

t-SNE Plot



Hypothesis

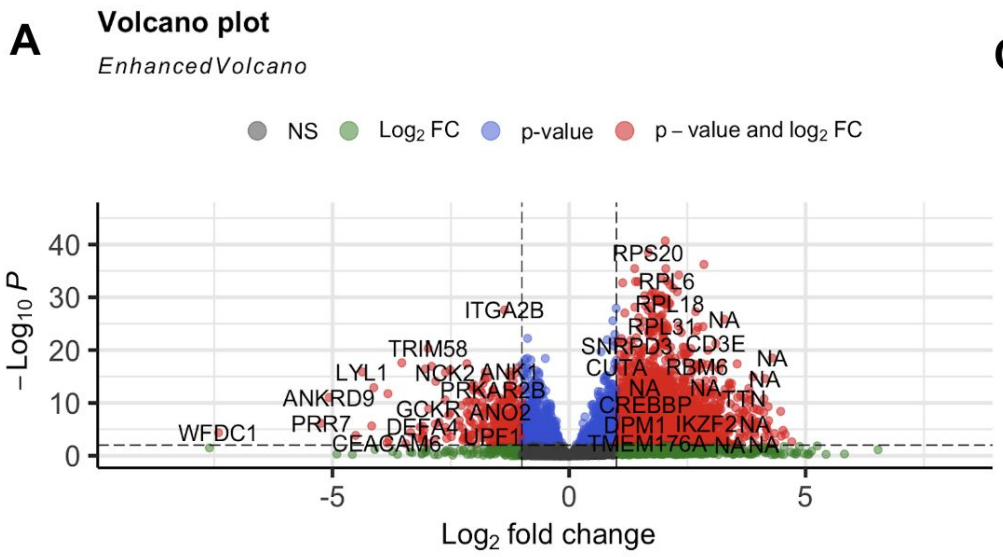
- Can the presence of cancer be determined based on the gene expression?
- Implications of differences in gene expression
- Potential benefits to diagnosis
- Potential identification of key genes

Differential Expression and Enrichment Analysis

- Used DeSeq with our DESeqDataSet object
- Volcano Plot created with differentially expressed genes
- Heatmap constructed with the overall significantly differentially expressed genes

Differential Expression and Enrichment Analysis

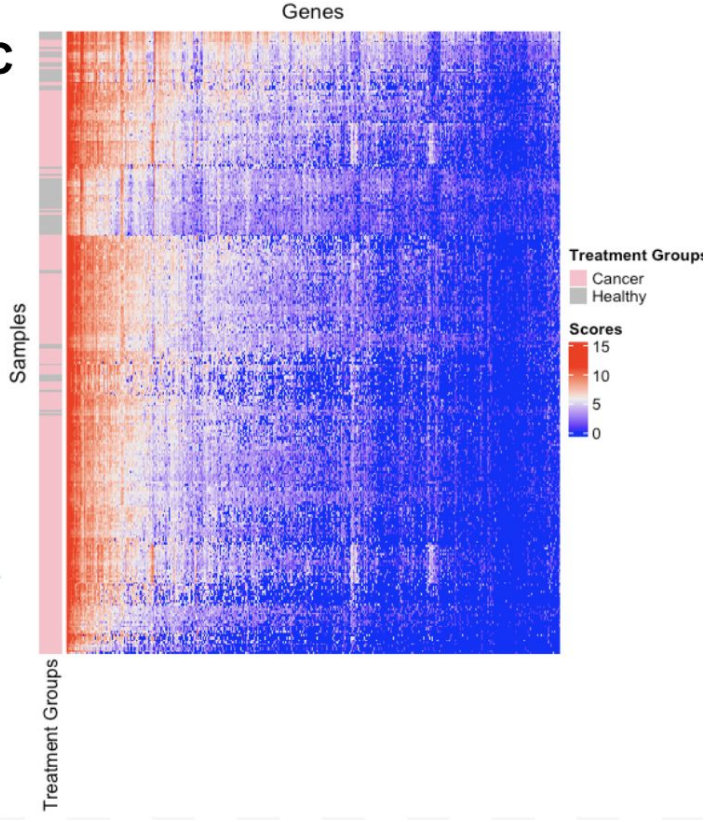
A



B

	genes	baseMean	log2FoldChange	lfcSE	pvalue	padj	threshold
1	DPM1	2.560248e+01	1.4058481	0.27961392	1.601275e-07	1.238843e-06	TRUE
2	SCYL3	6.299382e+00	1.6538750	0.42915992	3.094044e-05	1.350748e-04	TRUE
3	FGR	9.218809e+01	0.9811150	0.24821756	3.680709e-05	1.573083e-04	TRUE
4	FUCA2	3.592357e+01	0.4022840	0.14344232	4.471876e-03	1.117666e-02	TRUE
5	NFYA	4.896060e+00	1.0473509	0.41540838	4.516917e-03	1.127041e-02	TRUE
6	LAS1L	1.337220e+01	1.2037124	0.37070961	4.194998e-04	1.374315e-03	TRUE

C



Differential Expression and Enrichment Analysis

- 4 different enrichment analysis methods
  - topGO with the biological process ontology
  - clustProfiler with the biological process ontology
  - gProfiler2 with the molecular function gene ontology
  - gProfiler2 with the human phenotype ontology

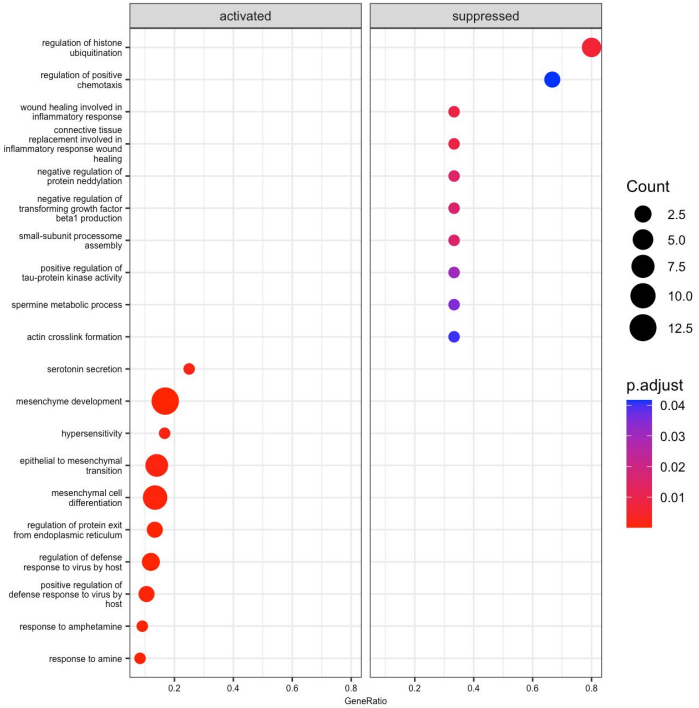
topGo: Biological Process Ontology

```
Description: Simple session
Ontology: BP
'classic' algorithm with the 'fisher' test
4151 GO terms scored: 126 terms with p < 0.01
Annotation data:
    Annotated genes: 6423
    Significant genes: 5261
    Min. no. of genes annotated to a GO: 10
    Nontrivial nodes: 4151
```

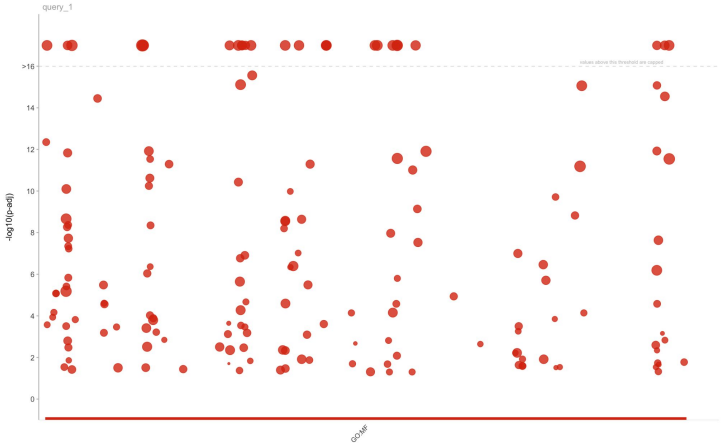
```
Description: Simple session
Ontology: BP
'elim' algorithm with the 'fisher : 0.01' test
4151 GO terms scored: 38 terms with p < 0.01
Annotation data:
    Annotated genes: 6423
    Significant genes: 5261
    Min. no. of genes annotated to a GO: 10
    Nontrivial nodes: 4151
```

	GO.ID	Term	Annotated	Significant	Expected	Rank in classicFisher	classicFisher	elimFisher
1	GO:0002181	cytoplasmic translation	125	123	102.39	8	4.8e-09	4.8e-09
2	GO:0000398	mRNA splicing, via spliceosome	211	198	172.83	16	2.8e-07	2.8e-07
3	GO:0006364	rRNA processing	187	175	153.17	25	2.4e-06	7.4e-05
4	GO:0022618	ribonucleoprotein complex assembly	134	125	109.76	33	0.00011	0.00011
5	GO:0042274	ribosomal small subunit biogenesis	63	61	51.60	39	0.00036	0.00036
6	GO:0045727	positive regulation of translation	82	78	67.17	40	0.00039	0.00039
7	GO:0010467	gene expression	2649	2278	2169.76	1	3.6e-13	0.00062
8	GO:0061077	chaperone-mediated protein folding	45	44	36.86	54	0.00134	0.00134
9	GO:0051098	regulation of binding	163	147	133.51	61	0.00223	0.00223
10	GO:0072655	establishment of protein localization to...	69	65	56.52	66	0.00277	0.00277

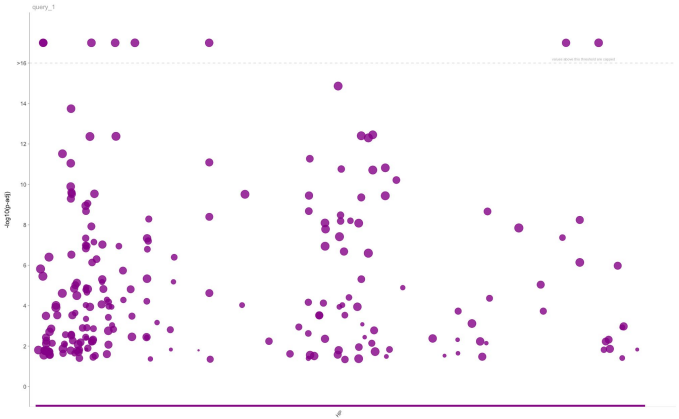
clustProfiler and gProfiler2



clustProfiler (Biological Process)



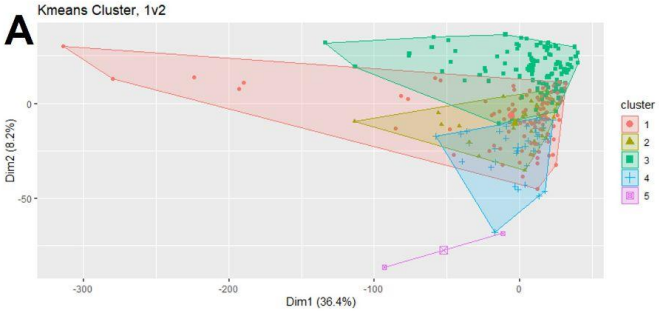
gProfiler2 (Molecular Function)

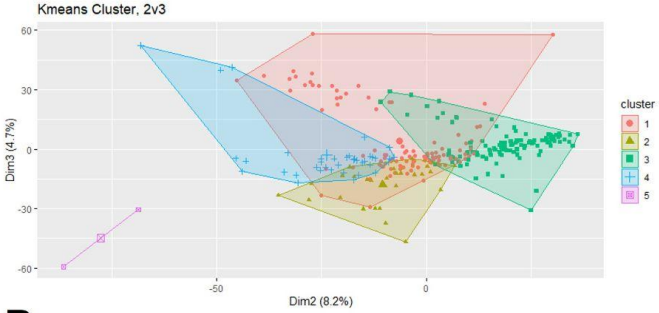


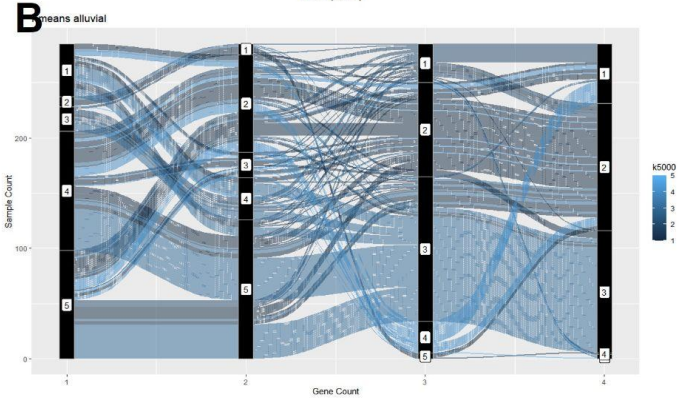
gProfiler2 (Human Phenotype)

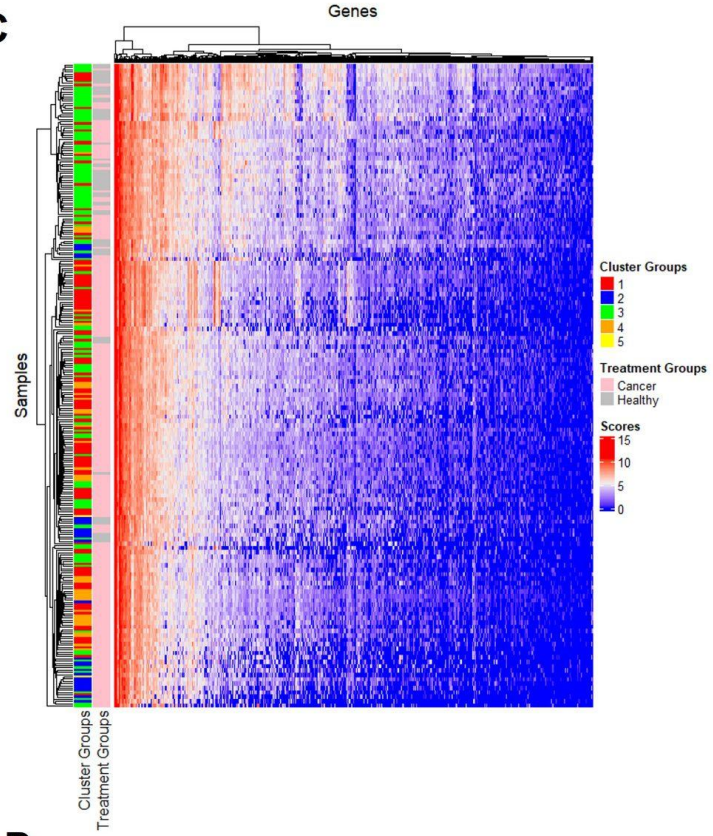
Clustering & Enrichment Analysis

- 4 different clustering algorithms
  - K-means clustering
  - ConsensusClusterPlus
  - PAM clustering
  - Gaussian clustering







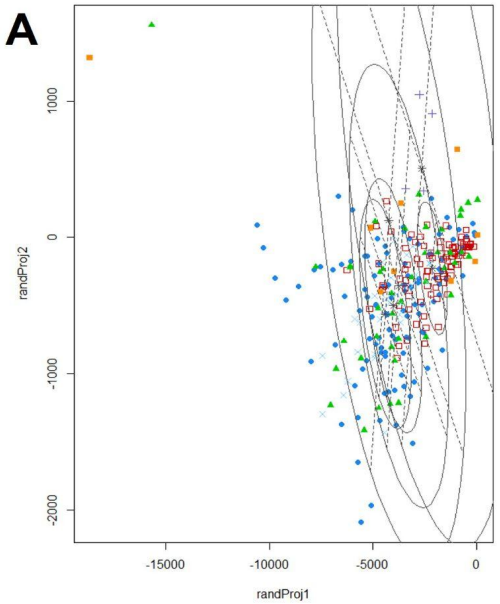


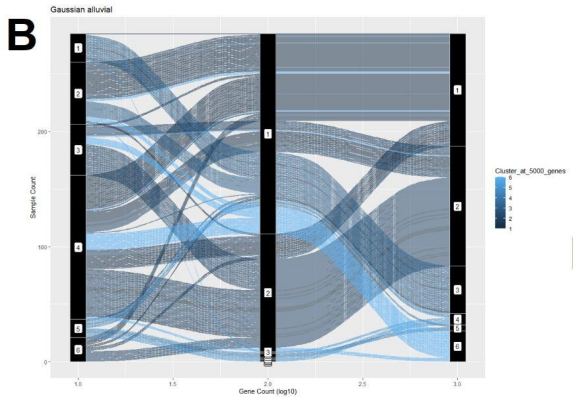
D

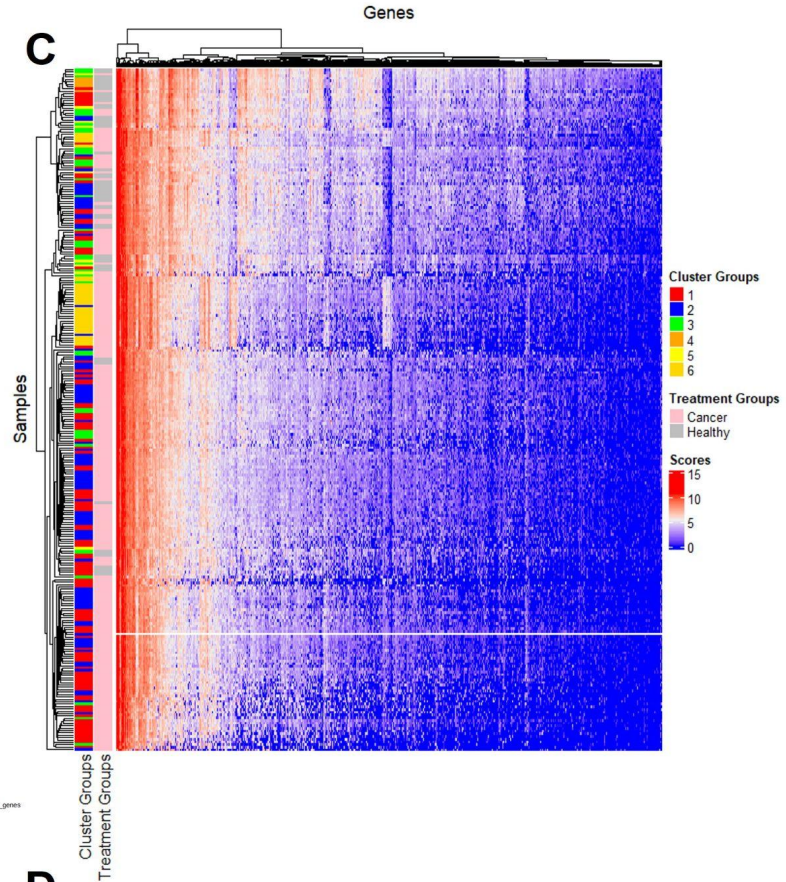
```
        Pearson's Chi-squared test

data:  table(kmeans_df)
X-squared = 38.481, df = 4, p-value = 8.914e-08
```

Clustering & Enrichment Analysis - K-Means







D

```
Pearson's Chi-squared test

data:  table(gaussian_df)
X-squared = 29.027, df = 5, p-value = 2.291e-05
```

Clustering & Enrichment Analysis - Other Methods

Conclusions & Future Work

- Did we answer the question: moderate to strong correlation
- Bioethical implications:
    - privacy/ownership, insurance companies
    - Obligations to disclose possible future diagnosis?
        - Children, high-risk, Effect on remainder of life
- Future of non-invasive diagnosis, early treatments