Using Statistic and ML to Find Biomarkers Suitable For Ovarian Cancer Prognosis

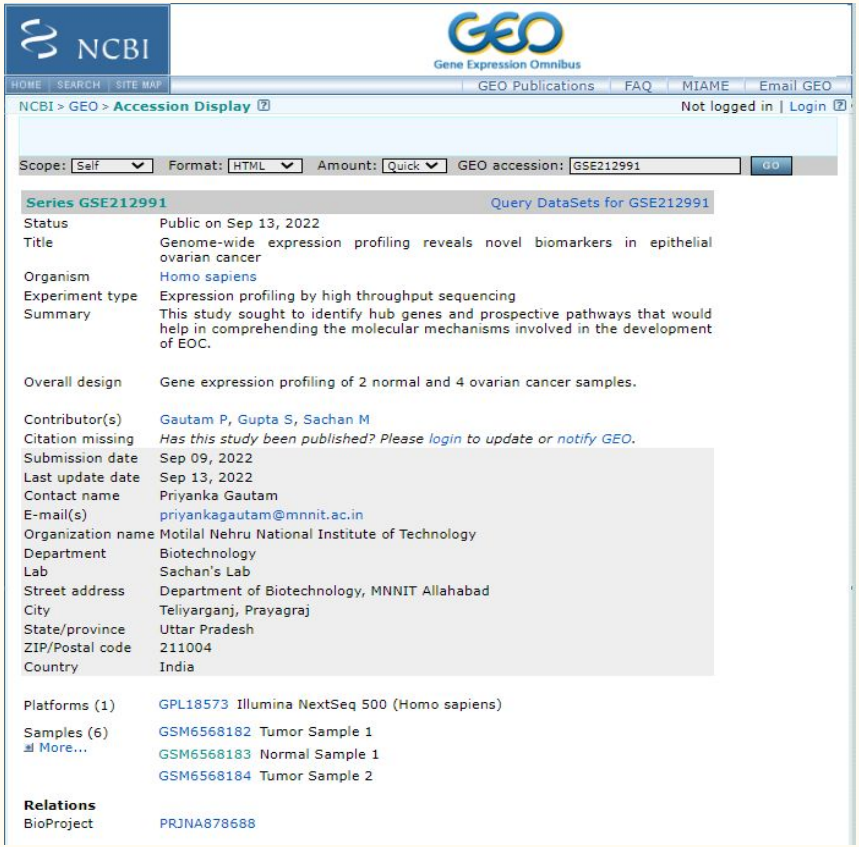
魏燿良 4108029016
曾德旭 4108029019

Contents

Dataset Introduction

Dataset Introduction

- Dataset name : Genome-wide expression profiling reveals novel biomarkers in epithelial ovarian cancer
- Source: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE212991



NCBI

GEO Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > Accession Display

Not logged in | Login

Scope: Self Format: HTML Amount: Quick GEO accession: GSE212991 GO

Series GSE212991 Query DataSets for GSE212991

Status	Public on Sep 13, 2022
Title	Genome-wide expression profiling reveals novel biomarkers in epithelial ovarian cancer
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	This study sought to identify hub genes and prospective pathways that would help in comprehending the molecular mechanisms involved in the development of EOC.
Overall design	Gene expression profiling of 2 normal and 4 ovarian cancer samples.
Contributor(s)	Gautam P, Gupta S, Sachan M
Citation missing	*Has this study been published? Please login to update or notify GEO.*
Submission date	Sep 09, 2022
Last update date	Sep 13, 2022
Contact name	Priyanka Gautam
E-mail(s)	priyankagautam@mnnit.ac.in
Organization name	Motilal Nehru National Institute of Technology
Department	Biotechnology
Lab	Sachan's Lab
Street address	Department of Biotechnology, MNNIT Allahabad
City	Teliyarganj, Prayagraj
State/province	Uttar Pradesh
ZIP/Postal code	211004
Country	India
Platforms (1)	GPL18573 Illumina NextSeq 500 (Homo sapiens)
Samples (6) More...	GSM6568182 Tumor Sample 1 GSM6568183 Normal Sample 1 GSM6568184 Tumor Sample 2

Relations

BioProject	PRJNA878688

Screenshot of NCBI website

Dataset Introduction

- Gene expression profiling of 2 normal and 4 ovarian cancer samples.
- Columns : 7
- Rows : 28026

GSM6568182	Tumor Sample 1
GSM6568183	Normal Sample 1
GSM6568184	Tumor Sample 2
GSM6568185	Tumor Sample 3
GSM6568186	Normal Sample 2
GSM6568187	Tumor Sample 4

Attribute information

	Geneid	N1	N2	T1	T2	T3	T4
0	A1BG	7	12	35	22	41	46
1	A1BG-AS1	0	0	7	0	4	5
2	A1CF	1	21	43	16	37	32
3	A2M	10	10	43	312	51	56
4	A2M-AS1	0	0	4	0	5	4
...	...	...	...	...	...	...	...
28021	ZYG11A	37	48	39	45	57	61
28022	ZYG11B	1	362	90	416	135	99
28023	ZYX	24	0	18	40	2	40
28024	ZZEF1	72	65	121	68	98	190
28025	ZZZ3	76	31	25	20	46	74

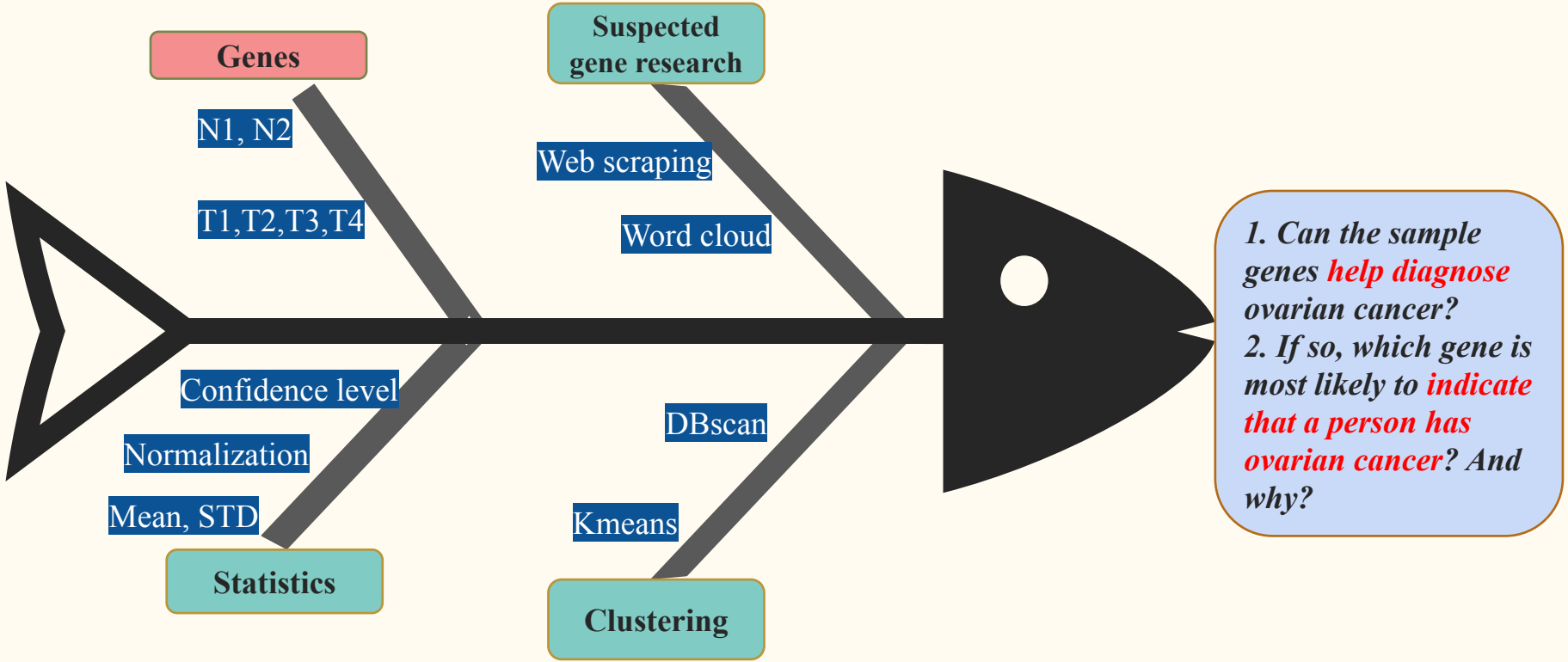
28026 rows × 7 columns

Overview of dataset

Research Purpose and Motivation

- Purpose:
  - the main goal of our research is to find potential biomarkers suitable for prognosis.
- Motivation :
  - learn more about how to do microarray data analysis.
  - Try using our technical ability to contribute and add value to the medical industry.

Fishbone Diagram



Literature Review

1

Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer

John B. Welsh*, Patrick P. Zarrinkar*†, Lisa M. Sapinoso*, Suzanne G. Kern*, Cynthia A. Behling‡, Bradley J. Monk§, David J. Lockhart*¶, Robert A. Burger§, and Garret M. Hampton*‖

*Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, San Diego, CA 92121; ‡Department of Pathology, University of California, San Diego, CA 92103; ¶Affymetrix, Inc., 3380 Central Expressway, Santa Clara, CA 95051; and §Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of California, Irvine–Medical Center, Orange, CA 92868

Communicated by Peter G. Schultz, Genomics Institute of the Novartis Research Foundation, La Jolla, CA, December 1, 2000 (received for review November 2, 2000)

2

Ye et al. *Cell Death Discovery* (2021)7:71
https://doi.org/10.1038/s41420-021-00451-x

Cell Death Discovery

ARTICLE Open Access

A novel defined pyroptosis-related gene signature for predicting the prognosis of ovarian cancer

Ying Ye[1,2], Qinjin Dai[3] and Hongbo Qi[1,2]

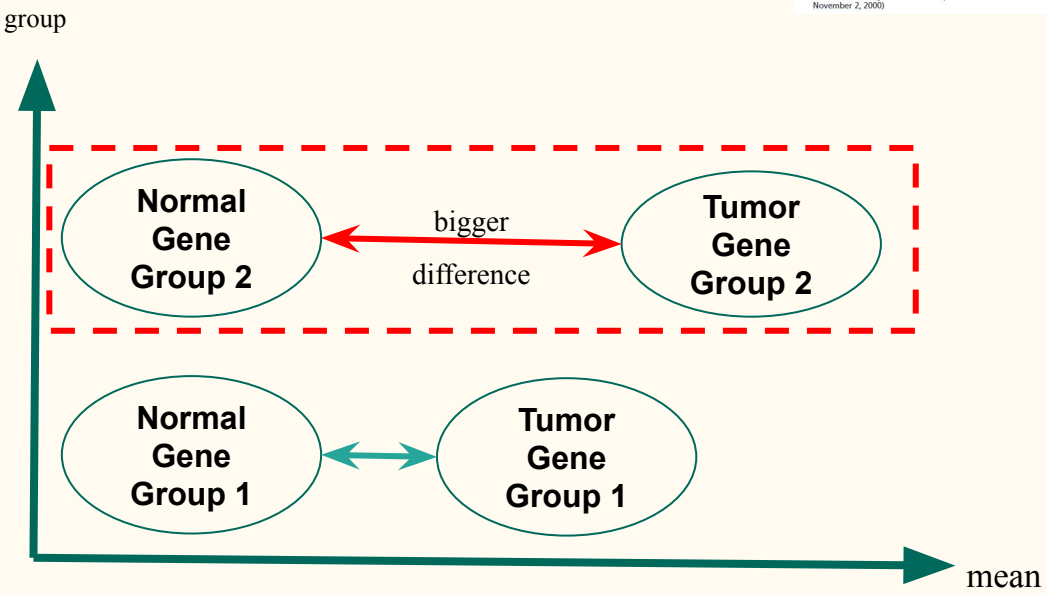
Literature Review 1

Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer

John B. Welsh*, Patrick P. Zarrinkar*†, Lisa M. Sapinoso*, Suzanne G. Kern*, Cynthia A. Behling‡, Bradley J. Monk§, David J. Lockhart*¶, Robert A. Burger§, and Garret M. Hampton*‖

*Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, San Diego, CA 92121; †Department of Pathology, University of California, San Diego, CA 92103; ¶Affymetrix, Inc., 3380 Central Expressway, Santa Clara, CA 95051; and ‡Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of California, Irvine–Medical Center, Orange, CA 92868

Communicated by Peter G. Schultz, Genomics Institute of the Novartis Research Foundation, La Jolla, CA, December 1, 2000 (received for review November 2, 2000)



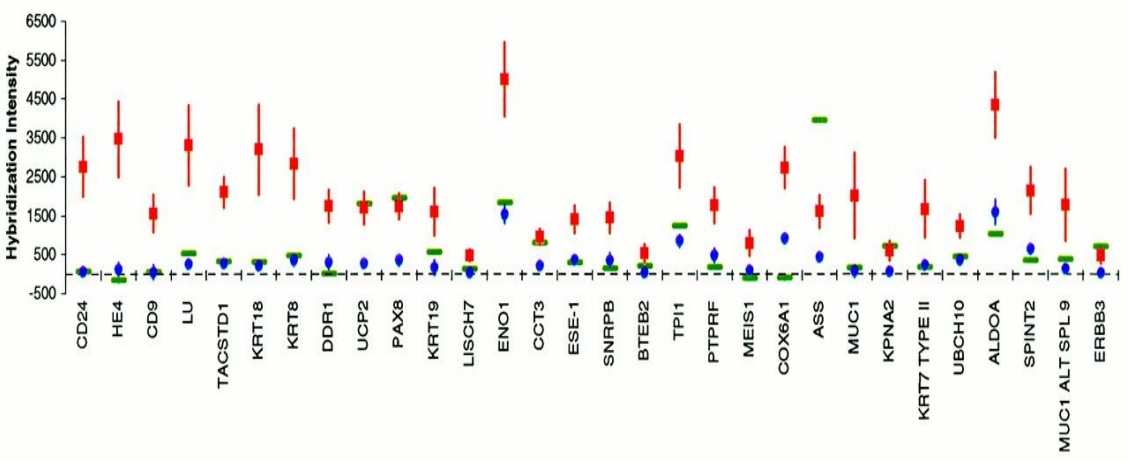
Literature Review

Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer

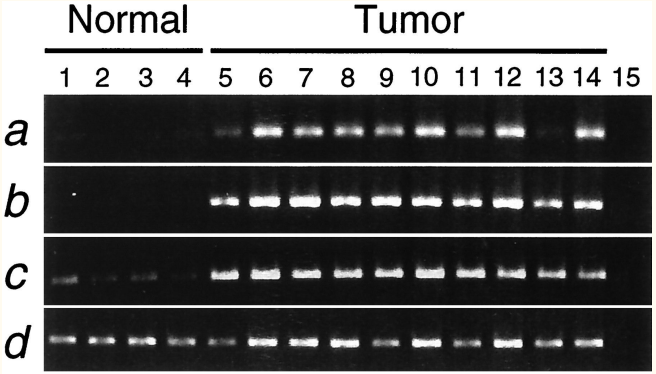
John B. Welsh*, Patrick P. Zarrinkar*†, Lisa M. Sapinoso*, Suzanne G. Kern*, Cynthia A. Behling‡, Bradley J. Monk§, David J. Lockhart*¶, Robert A. Burger§, and Garret M. Hampton*‖

*Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, San Diego, CA 92121; †Department of Pathology, University of California, San Diego, CA 92103; ¶Affymetrix, Inc., 3380 Central Expressway, Santa Clara, CA 95051; and ‡Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of California, Irvine–Medical Center, Orange, CA 92868

Communicated by Peter G. Schultz, Genomics Institute of the Novartis Research Foundation, La Jolla, CA, December 1, 2000 (received for review November 2, 2000)



Genes expression

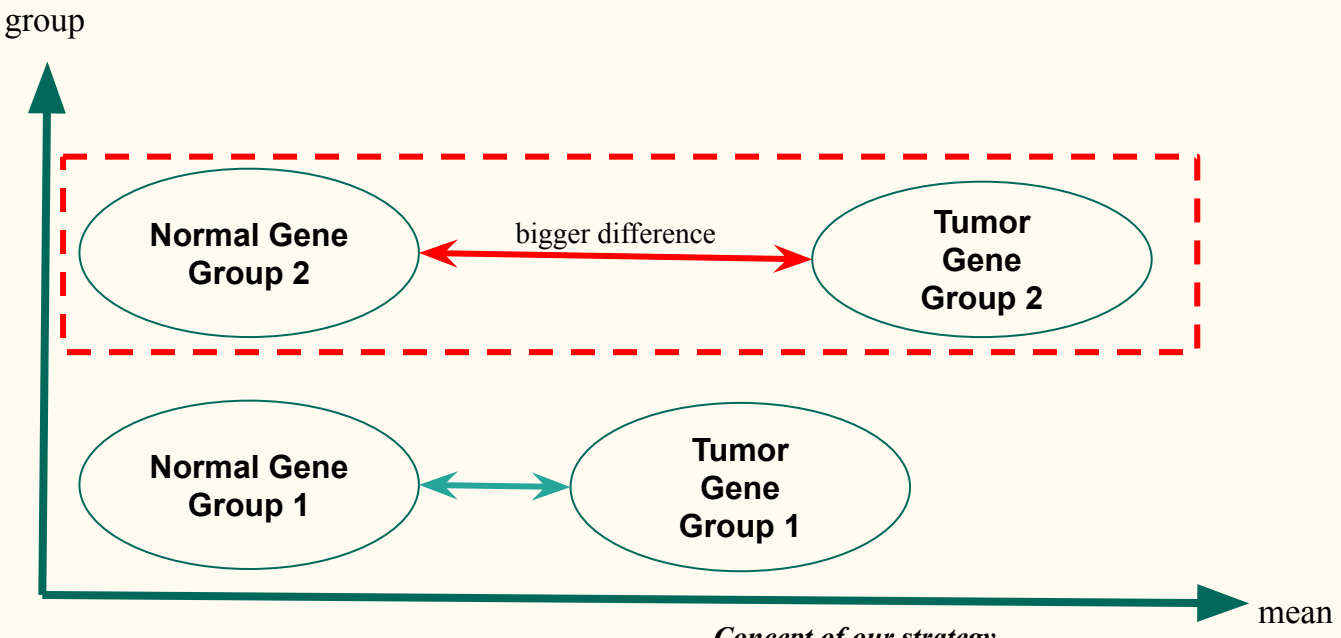


Result verification

source: Welsh, J. B., Zarrinkar, P. P., Sapinoso, L. M., Kern, S. G., Behling, C. A., Monk, B. J., ... & Hampton, G. M. (2001). Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proceedings of the National Academy of Sciences, 98(3), 1176-1181.

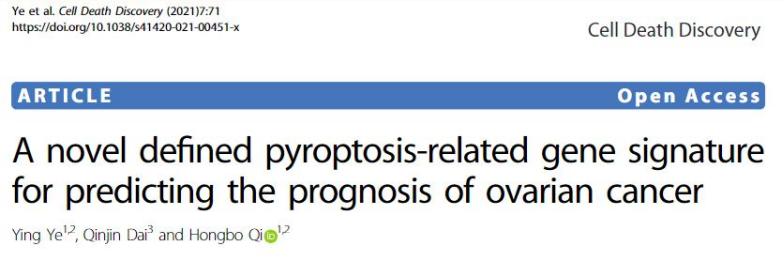
Our research purpose and strategy

- Identify gene groups with high differential expression between normal genes and tumor genes.
- Finding potential biomarkers suitable for prognosis.



Concept of our strategy

Literature Review - 2


Ye et al. *Cell Death Discovery* (2021)7:71
https://doi.org/10.1038/s41420-021-00451-x

Cell Death Discovery

ARTICLE Open Access

A novel defined pyroptosis-related gene signature for predicting the prognosis of ovarian cancer

Ying Ye[1,2], Qinjin Dai[3] and Hongbo Qi[1,2]

- Given the existing findings, knowing that pyroptosis plays an important role in the development of tumours and antitumour processes; however, its specific functions in OC have been less studied.
- Thus, performing a systematic study to determine the expression levels of pyroptosis-related genes between normal ovarian and OC tissues, explore the prognostic value of these genes, and study the correlations between pyroptosis and the tumour immune microenvironment.

Literature Review

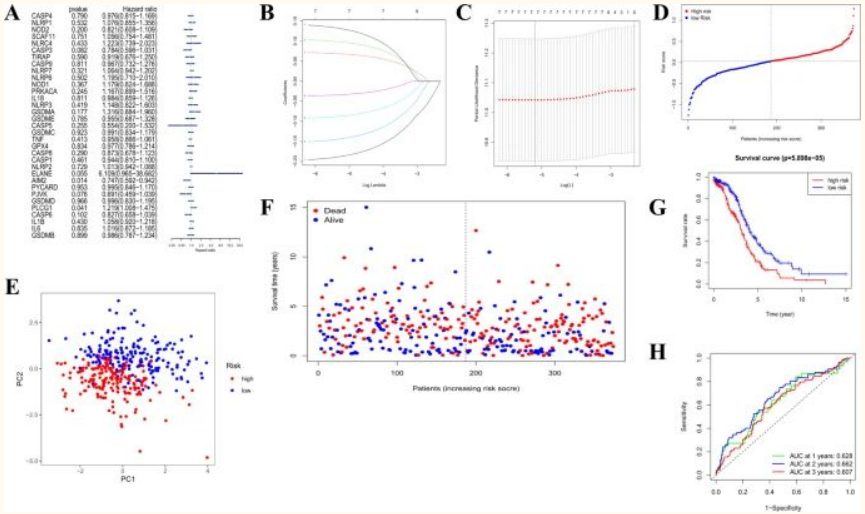
Ye et al. *Cell Death Discovery* (2021)7:71
https://doi.org/10.1038/s41420-021-00451-x

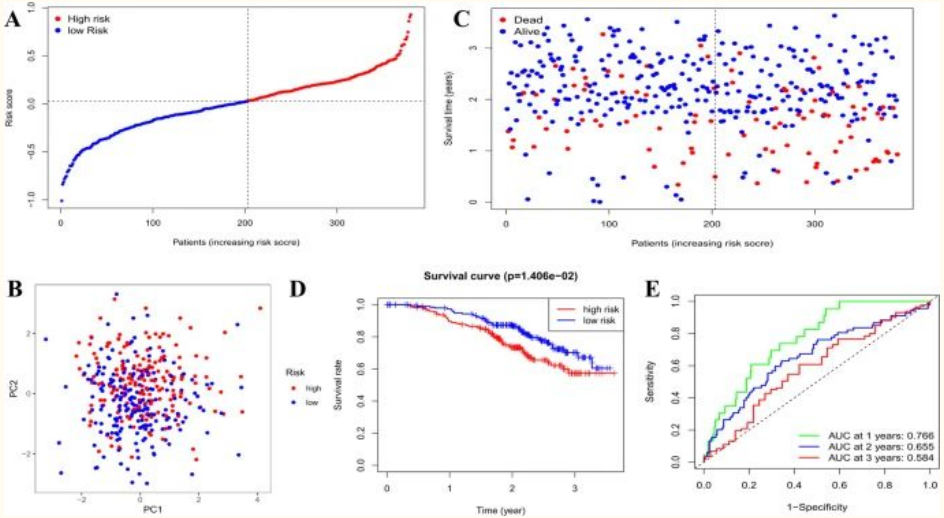
Cell Death Discovery

ARTICLE Open Access

A novel defined pyroptosis-related gene signature for predicting the prognosis of ovarian cancer

Ying Ye[1,2], Qinjin Dai[3] and Hongbo Qi[1,2]

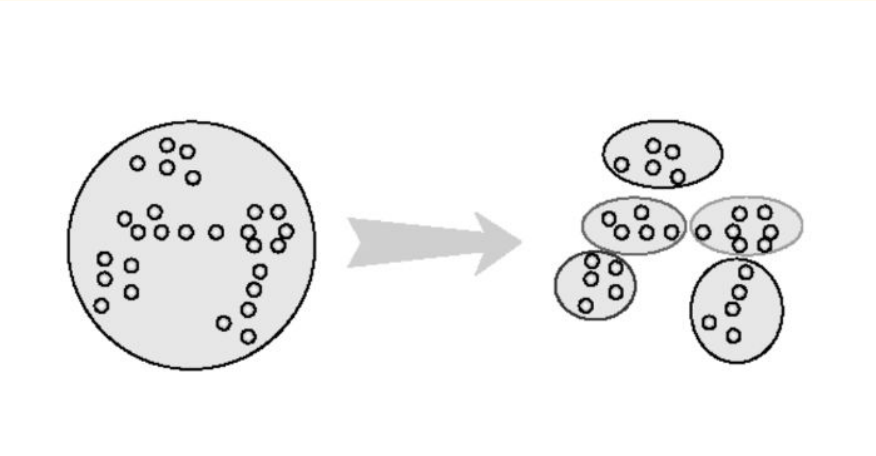


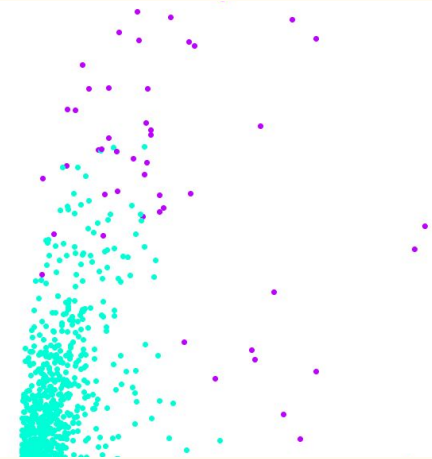


source: Ye, Y., Dai, Q., & Qi, H. (2021). A novel defined pyroptosis-related gene signature for predicting the prognosis of ovarian cancer. Cell death discovery, 7(1), 1-11.

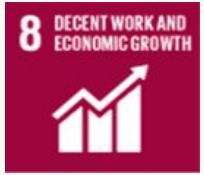
Literature Review--Our ideas

- we decided to try some clustering methods, such as K-means or DBscan, etc with PCA decomposition for cluster analysis to find potential clustering rules.
- Also, we discussed the principle of choosing the groups we want. Same as the above method, it must have strong differential expression and convergent results from normal state genes.





- Implementing the new technologies of today to improve our healthcare industries.

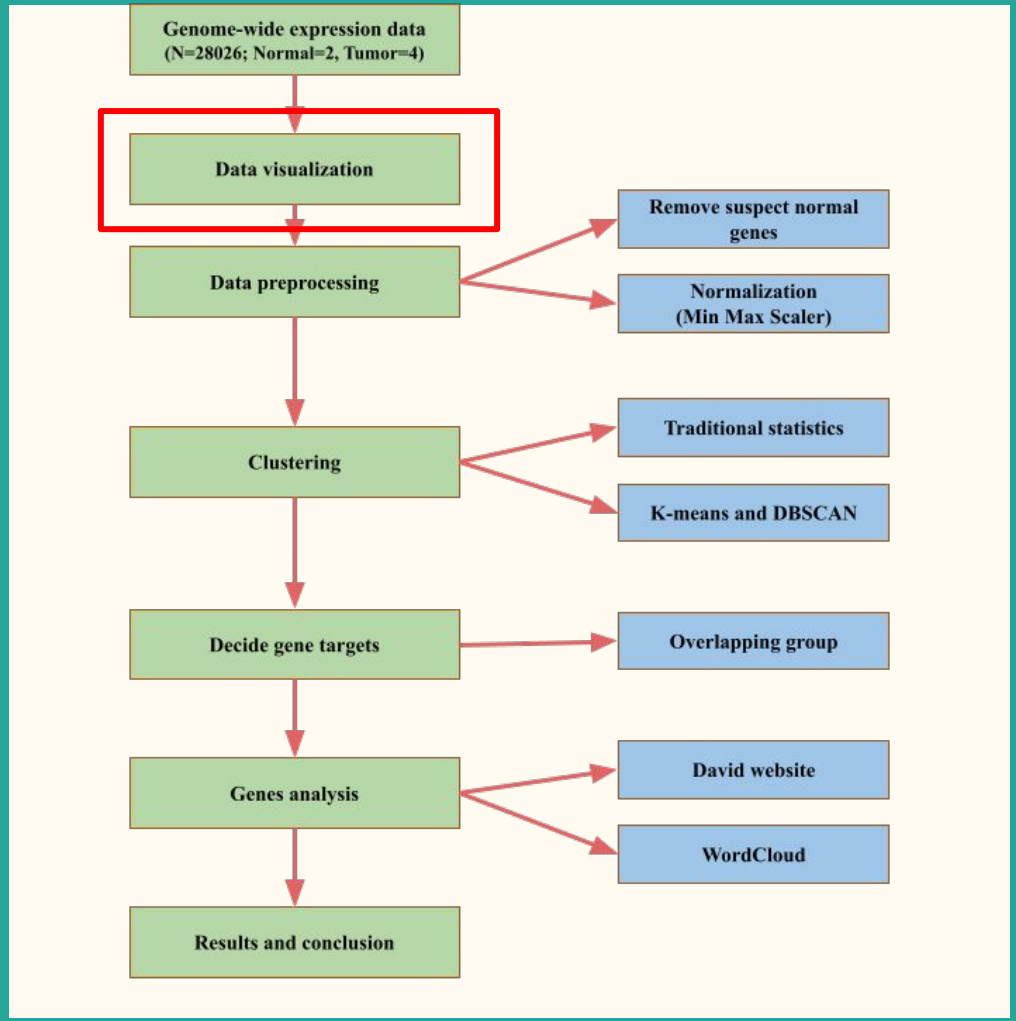


- With successful implementation of new technologies, we can:
  - Improve the quality of healthcare (more competitive)
  - Reduce cost
- A healthy society also means a more productive society

- With the help of modern technologies like AI, automation, etc. we can be more efficient, reduce the cost of healthcare.
- Lower cost means healthcare will be more accessible to everyone, not only for the rich.

Data Analysis

Workflow Part 1

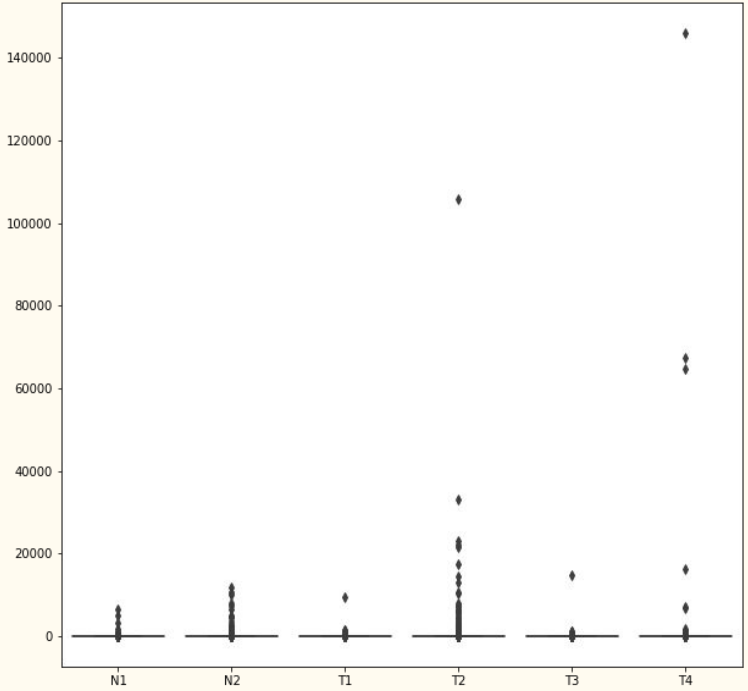


WorkFlow Part 1--Data visualization

- The graph of the boxplot is very small, which is very strange.

	N1	N2	T1	T2	T3	T4
count	28026.000000	28026.000000	28026.000000	28026.000000	28026.000000	28026.000000
mean	18.046849	26.375009	27.834190	53.482516	26.949012	46.020659
std	63.183212	165.972202	67.068825	746.548294	95.228727	1041.508054
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	4.000000	1.000000	2.000000	7.000000
50%	0.000000	4.000000	16.000000	12.000000	13.000000	22.000000
75%	27.000000	25.000000	39.000000	36.000000	37.000000	49.000000
max	6473.000000	11743.000000	9465.000000	105727.000000	14654.000000	145957.000000

Overview and description of dataset

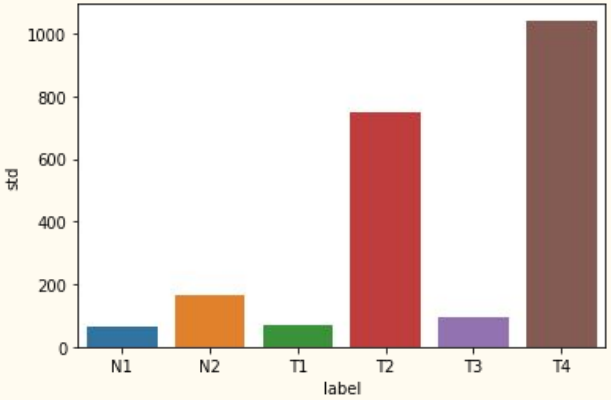


Boxplot of each column

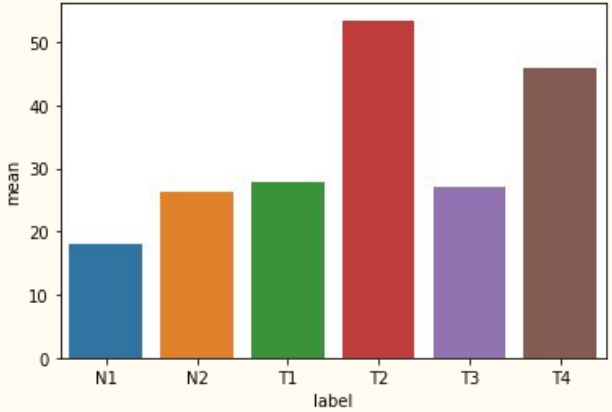
WorkFlow Part 1--Data visualization

- Because of the "Max" value, boxplot is compressed.
- This also shows that there is a large difference between each data, because the data distribution is not uniform. Especially T2, T4.

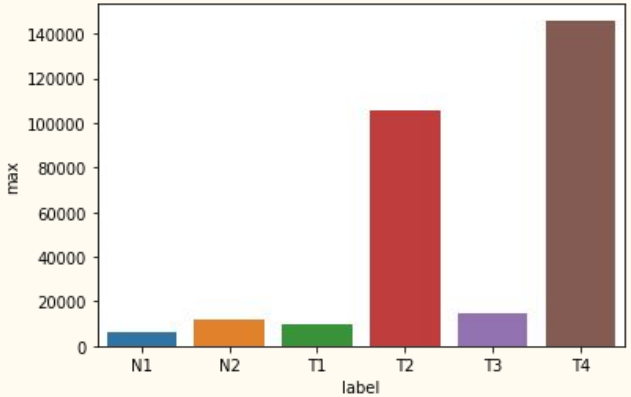
Barplot of SD



Barplot of mean



Barplot of Max

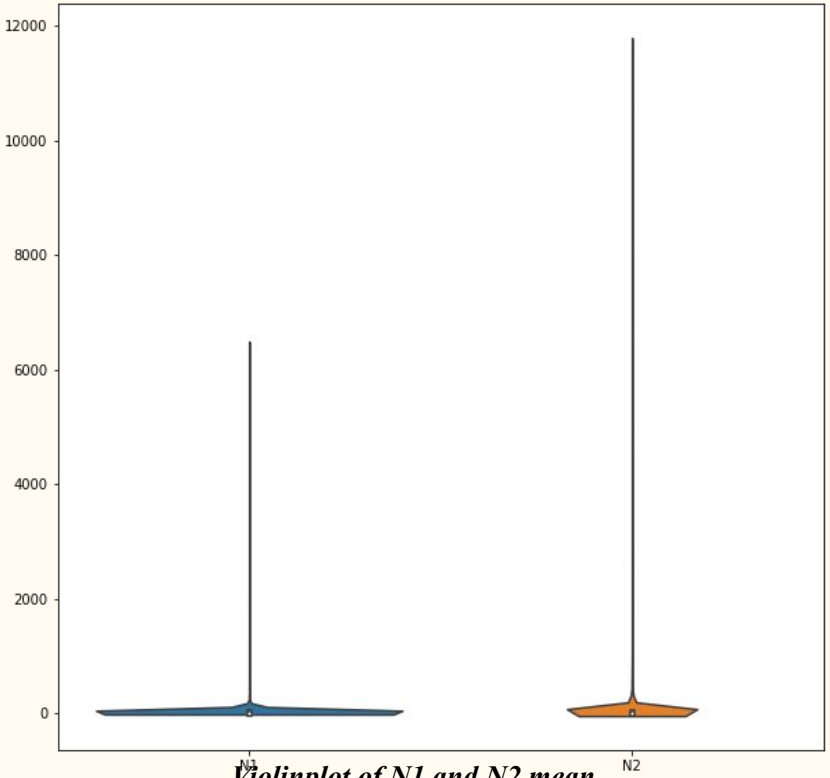


WorkFlow Part 1--Data visualization

- For us, the data distribution of N1 and N2 is important.
- We found that the data distribution of N1 and N2 is not the same.

	N1	N2
count	28026.000000	28026.000000
mean	18.046849	26.375009
std	63.183212	165.972202
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	4.000000
75%	27.000000	25.000000
max	6473.000000	11743.000000

Overview and description of N1 and N2

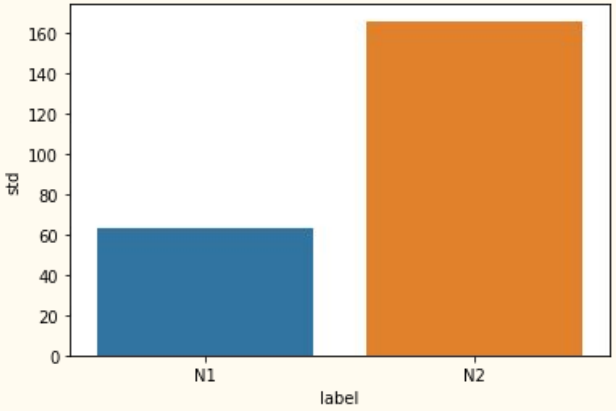


Violinplot of N1 and N2 mean

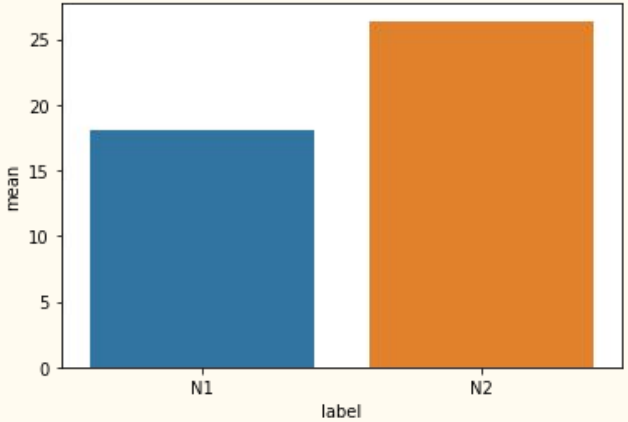
WorkFlow Part 1--Data visualization

- We believe that the factor causing the different distributions is that the mean, sd, and max of N2 are all greater than N1. It means that there must be some problem genes in N1 or N2.
- Therefore, the problematic part needs to be removed before starting the analysis.

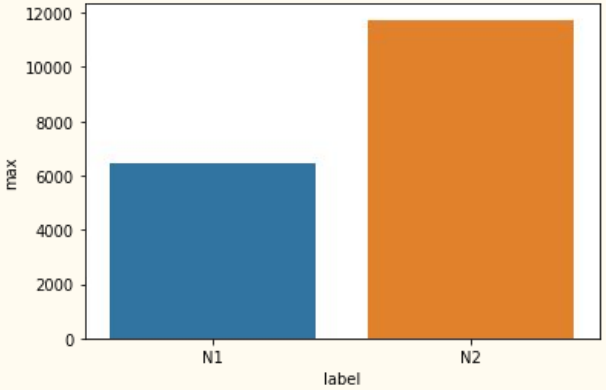
Barplot of SD



Barplot of mean



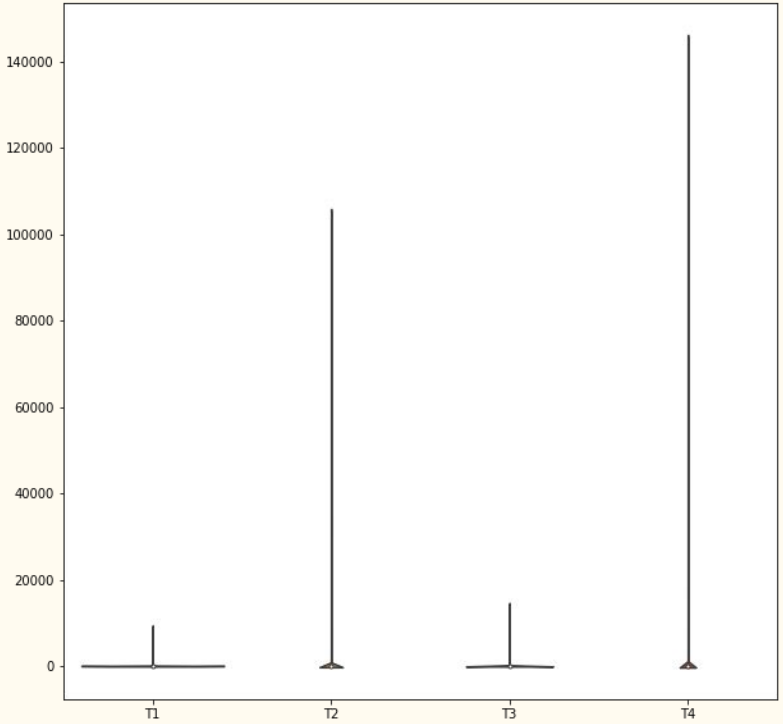
Barplot of Max



WorkFlow Part 1--Data visualization

	T1	T2	T3	T4
count	28026.000000	28026.000000	28026.000000	28026.000000
mean	27.834190	53.482516	26.949012	46.020659
std	67.068825	746.548294	95.228727	1041.508054
min	0.000000	0.000000	0.000000	0.000000
25%	4.000000	1.000000	2.000000	7.000000
50%	16.000000	12.000000	13.000000	22.000000
75%	39.000000	36.000000	37.000000	49.000000
max	9465.000000	105727.000000	14654.000000	145957.000000

Overview and description of T1-T4

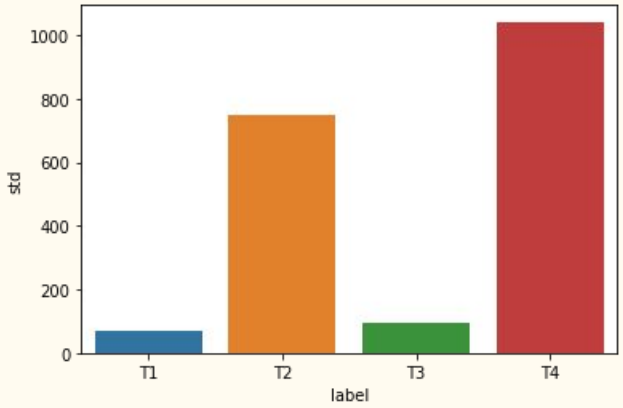


Violinplot of T1-T4 mean

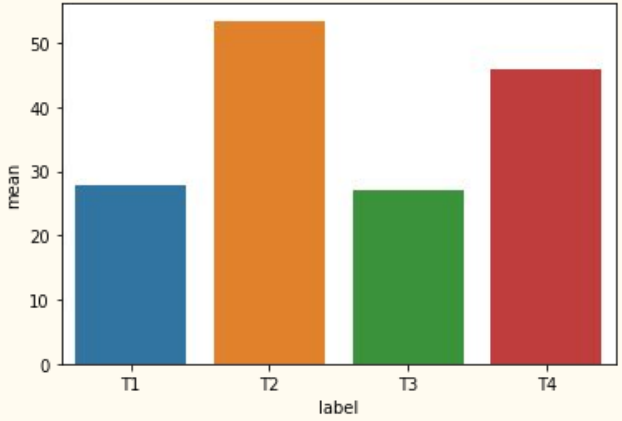
WorkFlow Part 1--Data visualization

- We don't spend time preprocessing the group data because they are unbalanced as expected.

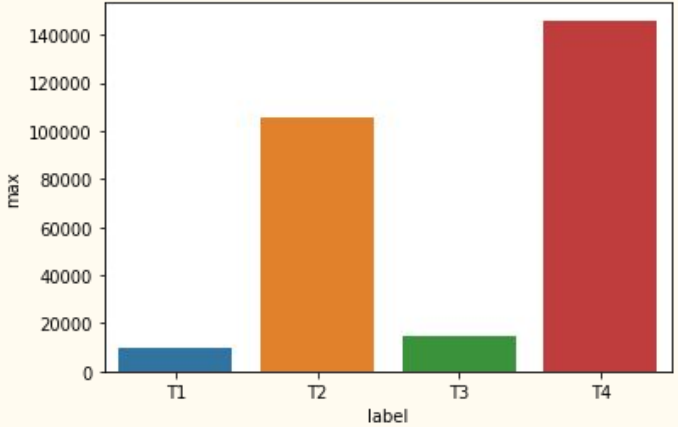
Barplot of SD



Barplot of mean

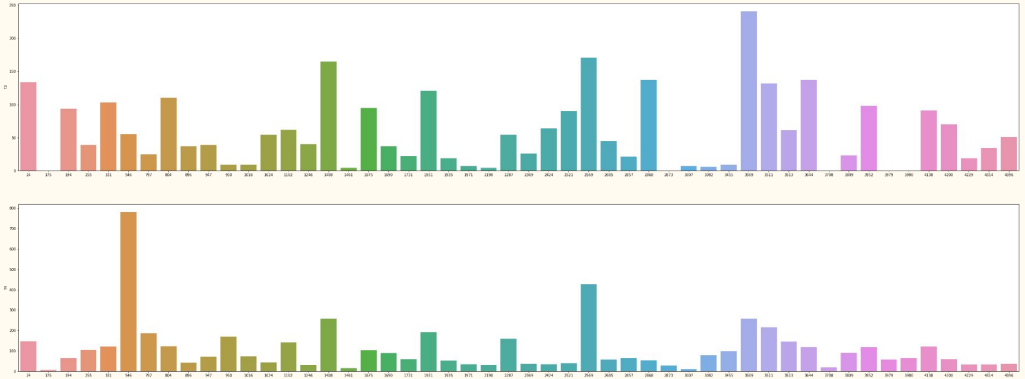


Barplot of Max



WorkFlow Part 1--Data visualization

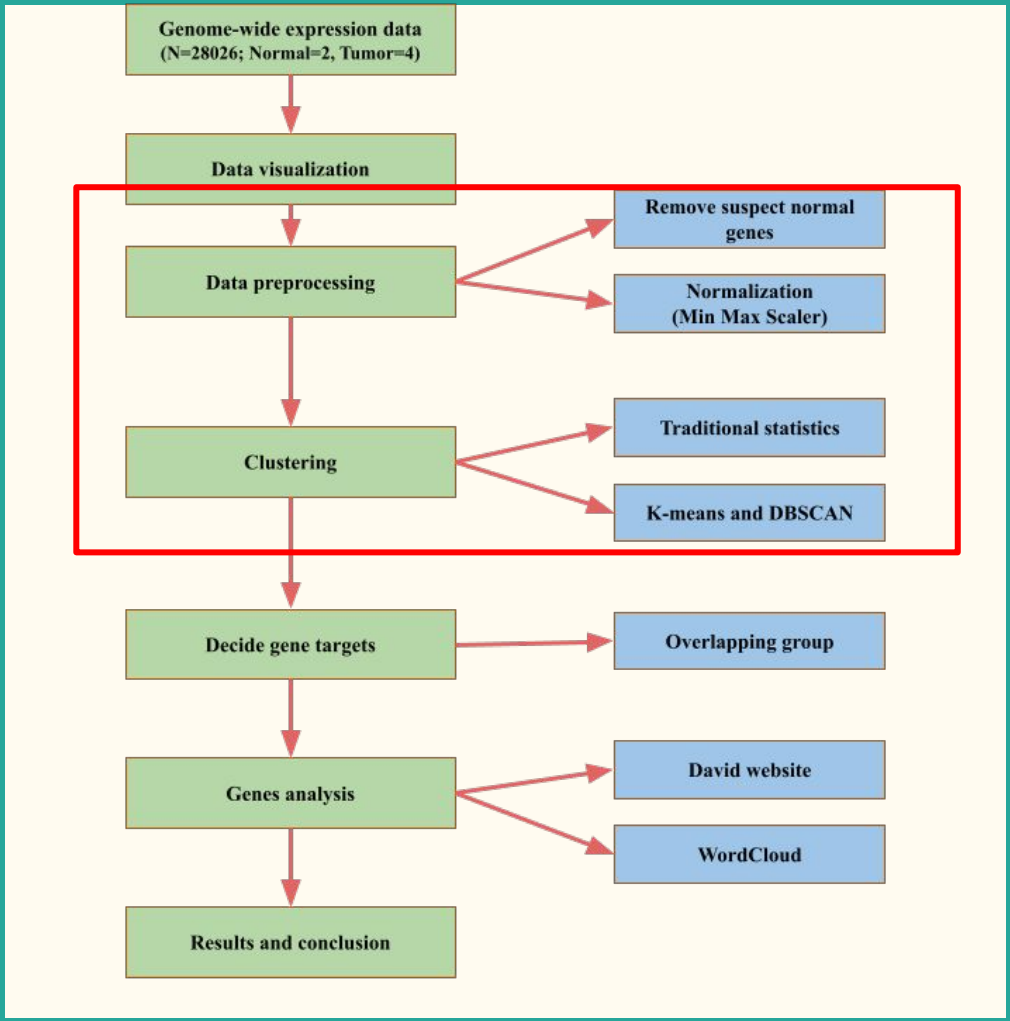
- Using bar graphs to show genes volume and patterns.
- Using heat-map to show the correlation between genes.



Expression levels and patterns of 50 genes in T3 and T4

Correlation heatmap of dataset

Workflow Part 2



WorkFlow Part 2

- Checking data for null values before processing.
- There are no missing or null values.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28026 entries, 0 to 28025
Data columns (total 7 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Geneid  28026 non-null  object
 1   N1      28026 non-null  int64
 2   N2      28026 non-null  int64
 3   T1      28026 non-null  int64
 4   T2      28026 non-null  int64
 5   T3      28026 non-null  int64
 6   T4      28026 non-null  int64
dtypes: int64(6), object(1)
memory usage: 1.5+ MB
```

Checking null value

```
#check missing values
df_ori.isnull().sum()
```

```
Geneid    0
N1        0
N2        0
T1        0
T2        0
T3        0
T4        0
dtype: int64
```

Checking missing value

Traditional Statistics Method

WorkFlow Part 2--Traditional Statistics Method

- Normalize each row by using Min Max Scaler

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

	Geneid	N1	N2	T1	T2	T3	T4
0	A1BG	7	12	35	22	41	46
1	A1BG-AS1	0	0	7	0	4	5
2	A1CF	1	21	43	16	37	32
3	A2M	10	10	43	312	51	56
4	A2M-AS1	0	0	4	0	5	4
...	...	...	...	...	...	...	...
28021	ZYG11A	37	48	39	45	57	61
28022	ZYG11B	1	362	90	416	135	99
28023	ZYX	24	0	18	40	2	40
28024	ZZEF1	72	65	121	68	98	190
28025	ZZZ3	76	31	25	20	46	74

28026 rows × 7 columns

Before normalization

Min Max Scaler

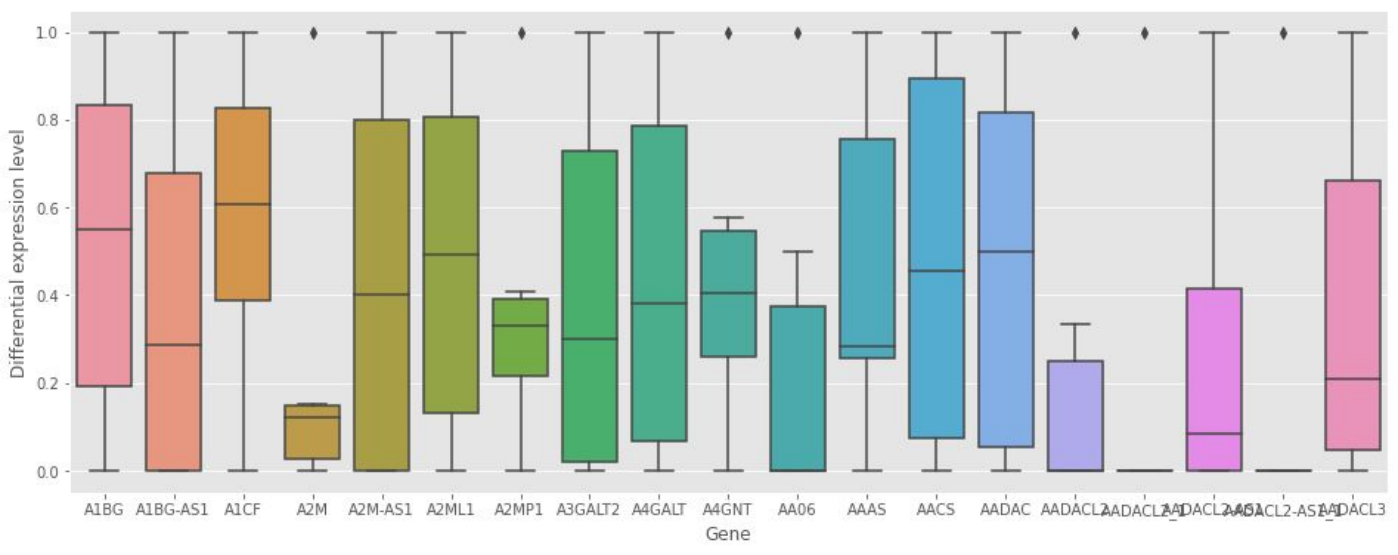
Geneid	N1	N2	T1	T2	T3	T4
A1BG	0.000	0.128205	0.717949	0.384615	0.871795	1.000000
A1BG-AS1	0.000	0.000000	1.000000	0.000000	0.571429	0.714286
A1CF	0.000	0.476190	1.000000	0.357143	0.857143	0.738095
A2M	0.000	0.000000	0.109272	1.000000	0.135762	0.152318
A2M-AS1	0.000	0.000000	0.800000	0.000000	1.000000	0.800000
...	...	...	...	...	...	...
ZYG11A	0.000	0.458333	0.083333	0.333333	0.833333	1.000000
ZYG11B	0.000	0.869880	0.214458	1.000000	0.322892	0.236145
ZYX	0.600	0.000000	0.450000	1.000000	0.050000	1.000000
ZZEF1	0.056	0.000000	0.448000	0.024000	0.264000	1.000000
ZZZ3	1.000	0.196429	0.089286	0.000000	0.464286	0.964286

28026 rows × 6 columns

After normalization

WorkFlow Part 2--Traditional Statistics Method

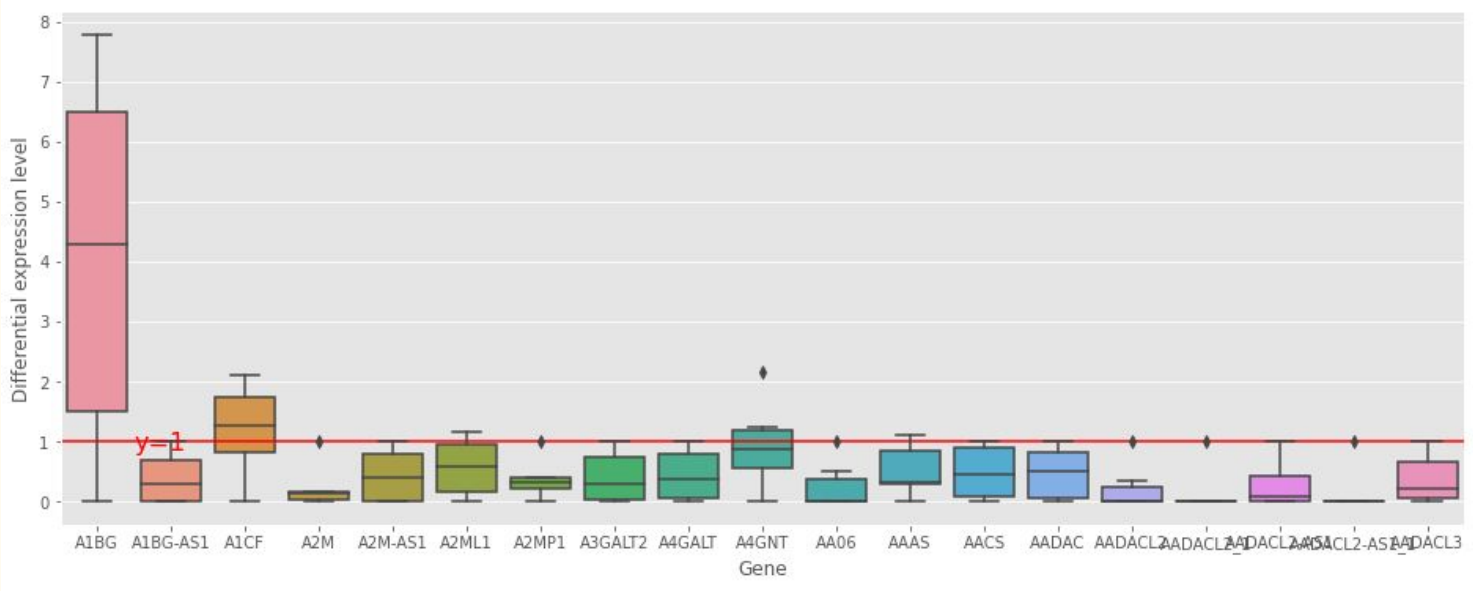
- Scale each row by its own MinMaxScaler, the purpose is to set a universal trade off value.
- Different genes have the same difference ratio now, so different genes can be compared together to determine the universal value of the differential expression of ovarian cancer genes.



Genes value after normalization

WorkFlow Part 2--Traditional Statistics Method

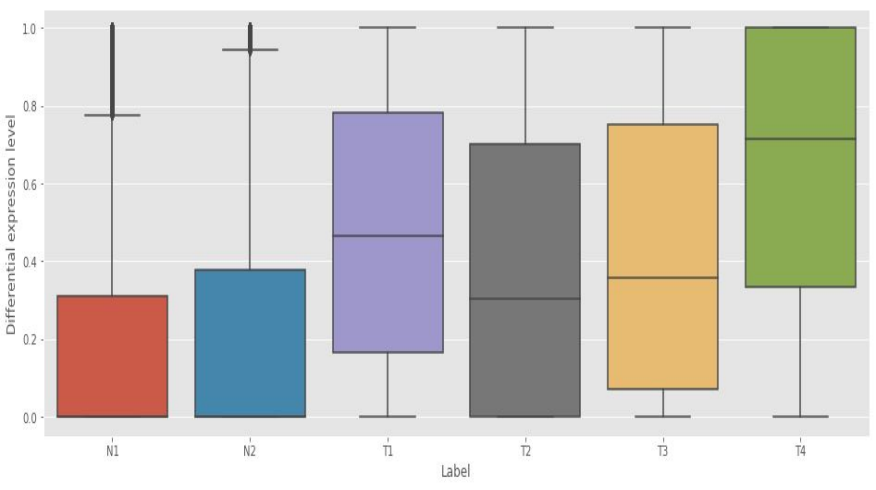
- We select the maximum and minimum values from the N1-N2 data to determine the universal value of the differential expression of ovarian cancer genes.
- Exceeding the red line of y=1 means ovarian cancer gene expression exceeding the normal level.



Genes value after normalization

WorkFlow Part 2--Traditional Statistics Method

- Overview of datasets now.



Boxplot of overview data now

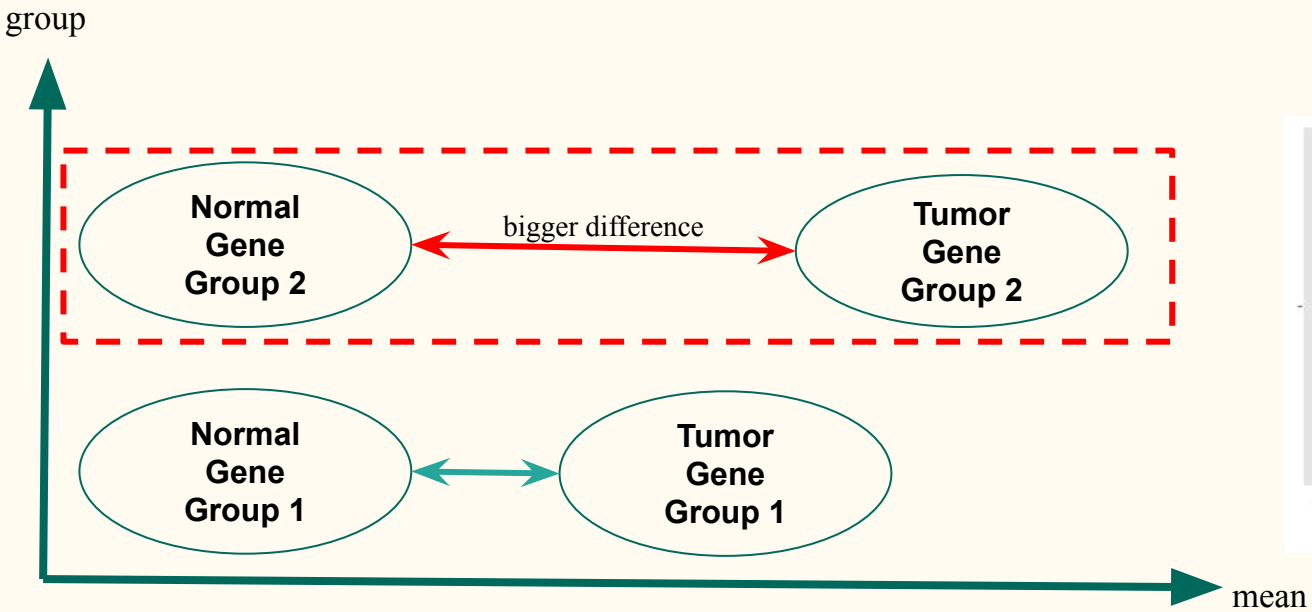
	N1	N2	T1	T2	T3	T4
count	28026.000000	28026.000000	28026.000000	28026.000000	28026.000000	28026.000000
mean	0.213459	0.226588	0.481389	0.393441	0.427960	0.638165
std	0.360583	0.337611	0.346661	0.371784	0.366148	0.360089
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.166667	0.000000	0.070044	0.333333
50%	0.000000	0.000000	0.465116	0.305085	0.357724	0.714286
75%	0.310345	0.377551	0.781250	0.702414	0.752124	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

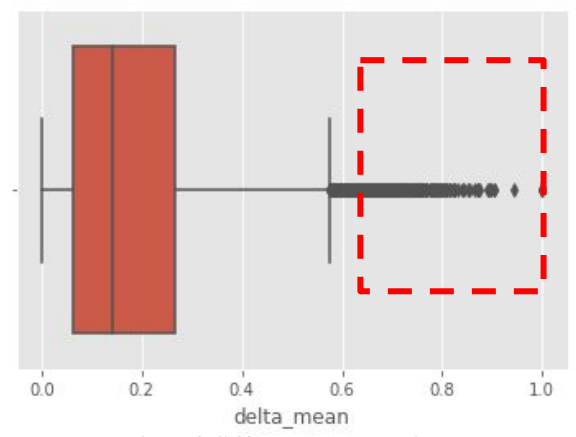
Overview and description of N1-T4

WorkFlow Part 2--Traditional Statistics Method

- Trade off value = mean of difference ± 3 SD / Top 0.01%
- Find the gene groups which difference between normal genes and tumor genes are exceeds the trade-off value.

	N_mean	T_mean	delta_mean
count	28026.000000	28026.000000	28026.000000
mean	0.220023	0.485239	0.182117
std	0.231254	0.194555	0.157499
min	0.000000	0.000000	0.000000
25%	0.000000	0.333333	0.062500
50%	0.142857	0.500000	0.140625
75%	0.450000	0.633333	0.267746
max	1.000000	1.000000	1.000000





Boxplot of difference mean between N-group and T-group

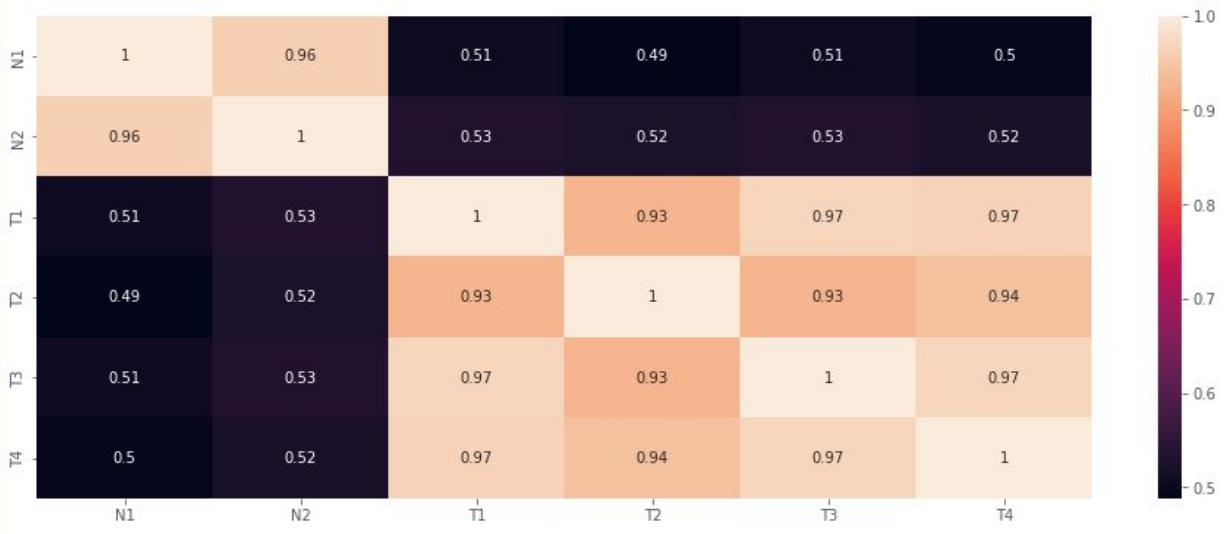
WorkFlow Part 2--Traditional Statistics Method

- The distribution of N1 and N2 is very close, and it is also clearly distinguished from T1-T4.

	Geneid	N1	N2	T1	T2	T3	T4
114	ABHD4	4	0	25	21	19	25
132	ABRACL	0	0	5	4	3	4
181	ACMSD	3	0	19	15	21	19
269	ACTRT1	0	0	8	8	4	6
307	ADAM3A	0	0	37	14	37	40
...	...	...	...	...	...	...	...
27624	ZNF503-AS2	0	0	11	13	10	12
27716	ZNF597	0	0	23	12	17	22
27792	ZNF688	0	0	49	48	26	51
27883	ZNF790	0	1	15	25	19	23
27914	ZNF835	0	0	34	18	39	37

370 rows × 7 columns

370 gene groups are top0.01% groups



Correlation heatmap between N1-T4

WorkFlow Part 2--Traditional Statistics Method

- Find genes with the same trait "H"; 465 genes are selected.
- Try to find genes that each patient has in common.

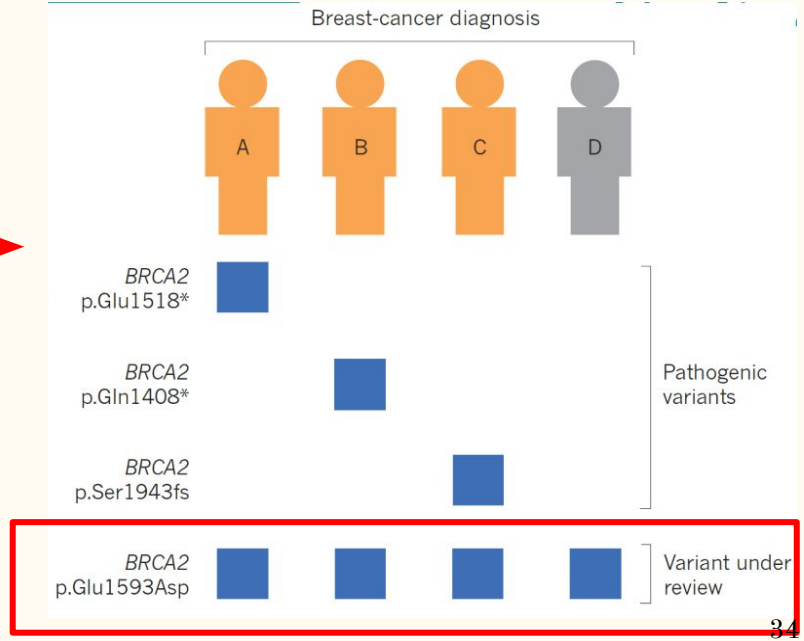
	Geneid	N1	N2	T1	T2	T3	T4
103	ABHD13	0.0	0.0	H	H	H	H
206	ACP6_1	0.0	0.0	H	H	H	H
234	ACTBL2	0.0	0.0	H	H	H	H
473	AFF2_1	0.0	0.0	H	H	H	H
556	AIF1_2	0.0	0.0	H	H	H	H
...	...	...	...	...	...	...	...
27918	ZNF84	0.0	1.0	H	H	H	H
27925	ZNF84-DT	0.0	1.0	H	H	H	H
27942	ZNF90	0.0	1.0	H	H	H	H
27945	ZNF93	0.0	0.0	H	H	H	H
27955	ZNRD1ASP_4	0.0	0.0	H	H	H	H

465 rows × 11 columns

465 rows

465 gene groups have 4 "H" values

Concept of our target



WorkFlow Part 2--Traditional Statistics Method

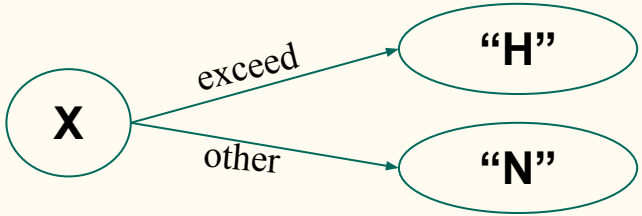
- Scale each row by its own MinMaxScaler.
- Trade off value = mean ± 0.86 SD
- If x exceeds the trade-off value, it is set to “H”, indicating a high risk level. Vice versa, if x is between compromise values, it is set to “N”, which means normal level.

	Geneid	N1	N2	T1	T2	T3	T4	mean_T1toT4	SD_T1toT4	trade_off_lower	trade_off_upper
0	A1BG	0.0	1.0	N	H	N	H	5.800000	2.072036	4.018049	7.581951
1	A1BG-AS1	0.0	0.0	H	H	N	N	4.000000	2.943920	1.468229	6.531771
2	A1CF	0.0	1.0	H	H	N	N	1.550000	0.578792	1.052239	2.047761
3	A2M	0.0	0.0	N	H	N	N	105.500000	131.109369	-7.254057	218.254057
4	A2M-AS1	0.0	0.0	N	H	N	N	3.250000	2.217356	1.343074	5.156926

Genes value after label transformation

	T1	T2	T3	T4
H	7284	14201	9588	13062
N	20742	13825	18438	14964

T1 to T4 value count



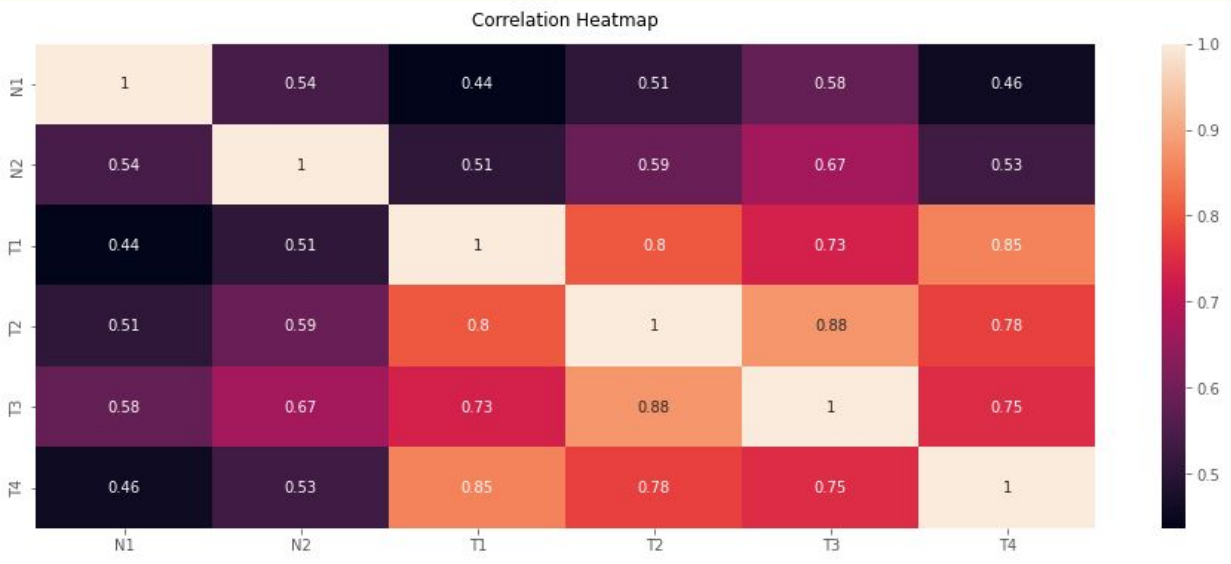
WorkFlow Part 2--Traditional Statistics Method

- The distribution of N1 and N2 is very close, and it is also clearly distinguished from T1-T4.
- More specifically, it means that we have found genes that meet our expectations, and these genes only appear in T1-T4.

	Geneid	N1	N2	T1	T2	T3	T4
103	ABHD13	0.0	0.0	H	H	H	H
206	ACP6_1	0.0	0.0	H	H	H	H
234	ACTBL2	0.0	0.0	H	H	H	H
473	AFF2_1	0.0	0.0	H	H	H	H
556	AIF1_2	0.0	0.0	H	H	H	H
...	...	...	...	...	...	...	...
27918	ZNF84	0.0	1.0	H	H	H	H
27925	ZNF84-DT	0.0	1.0	H	H	H	H
27942	ZNF90	0.0	1.0	H	H	H	H
27945	ZNF93	0.0	0.0	H	H	H	H
27955	ZNRD1ASP_4	0.0	0.0	H	H	H	H

465 rows × 11 columns

465 gene groups have 4 "H" values



Correlation heatmap between N1-T4

WorkFlow Part 2--Traditional Statistics Method

- Final outcome after doing traditional statistical analysis. We get 828 suspect genes.

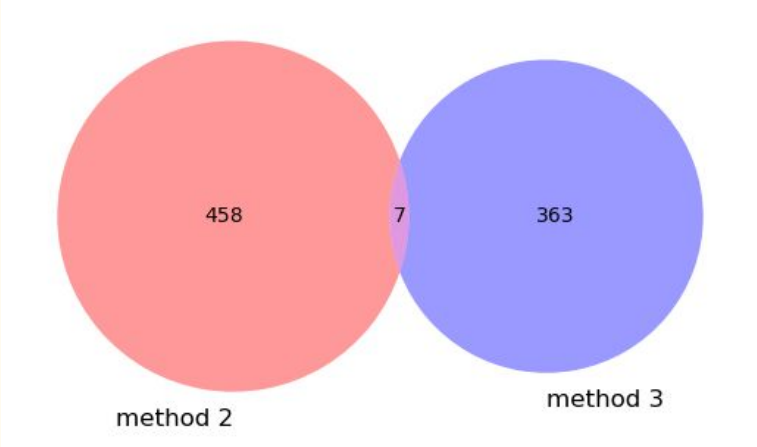
	Geneid	N1	N2	T1	T2	T3	T4
114	ABHD4	4	0	25	21	19	25
132	ABRACL	0	0	5	4	3	4
181	ACMSD	3	0	19	15	21	19
269	ACTRT1	0	0	8	8	4	6
307	ADAM3A	0	0	37	14	37	40
...	...	...	...	...	...	...	...
27624	ZNF503-AS2	0	0	11	13	10	12
27716	ZNF597	0	0	23	12	17	22
27792	ZNF688	0	0	49	48	26	51
27883	ZNF790	0	1	15	25	19	23
27914	ZNF835	0	0	34	18	39	37

370 rows × 7 columns

	Geneid	N1	N2	T1	T2	T3	T4
103	ABHD13	0.0	0.0	H	H	H	H
206	ACP6_1	0.0	0.0	H	H	H	H
234	ACTBL2	0.0	0.0	H	H	H	H
473	AFF2_1	0.0	0.0	H	H	H	H
556	AIF1_2	0.0	0.0	H	H	H	H
...	...	...	...	...	...	...	...
27918	ZNF84	0.0	1.0	H	H	H	H
27925	ZNF84-DT	0.0	1.0	H	H	H	H
27942	ZNF90	0.0	1.0	H	H	H	H
27945	ZNF93	0.0	0.0	H	H	H	H
27955	ZNRD1ASP_4	0.0	0.0	H	H	H	H

465 rows × 11 columns

Final outcome with traditional statistical analysis



Overlapping group of statistics method

Modern ML Method

WorkFlow Part 2--Modern ML Method

- Removing 10,648 suspected normal genes, those genes have high variance. We removed the genes:
  - Which difference between N1 and N2 is greater than the overall difference.
    - |N1-N2| > overall mean
  - Which difference ratio between N1 and N2 is greater than the overall difference.
    - |N1/N2| > overall mean ratio or |N2/N1| > overall mean ratio

Geneid	N1	N2	T1	T2	T3	T4
ZYG11B	1	362	90	416	135	99

This gene is what we don't want

	diff_N1nN2_byScale	diff_N1nN2
count	28026.000000	28026.000000
mean	11.375524	25.954364
std	81.496408	158.408316

Overall mean and SD of difference and difference ratio of N1 and N2

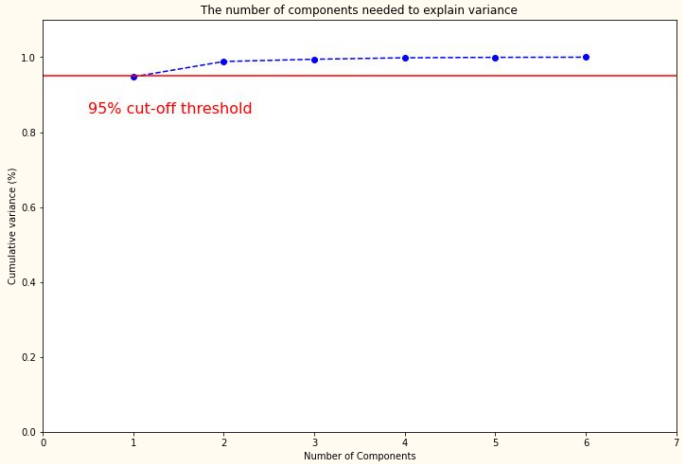
28026 rows × → -10,648 → 17378 rows ×

WorkFlow Part 2--Modern ML Method

DBSCAN (eps=70, min_samples=45, n=17378) → PCA(2 dimension) → Plot()

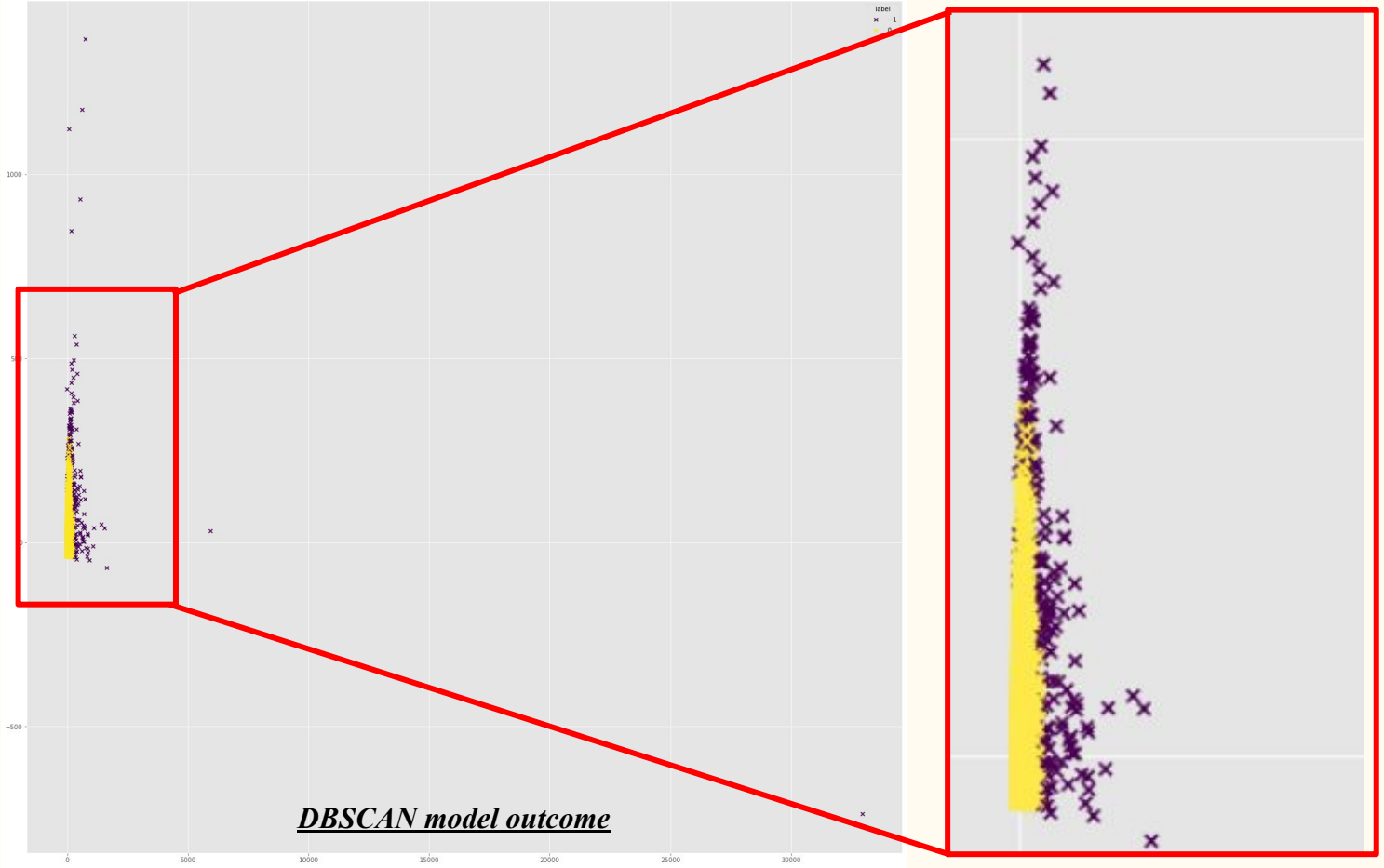
- Fit the DBSCAN model with 17378 genes, parameters are eps=70 and min_samples=45.
- Decompose the 6-dimensional data of 17378 genes into 2 dimensions.
- Plot the graph.

	1	2	3	4	5	6
0	0.947592	0.98843	0.994539	0.998364	0.999606	1.0



<u>The number of PCA components needed to explain variance</u>

WorkFlow Part 2--Modern ML Method



DBSCAN model outcome

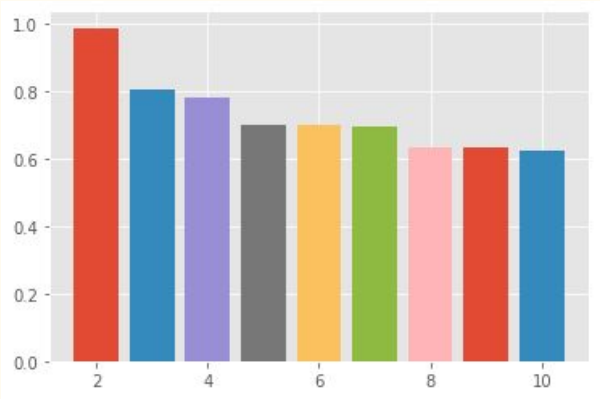
	label
0	17193
-1	185

DBSCAN labels value count

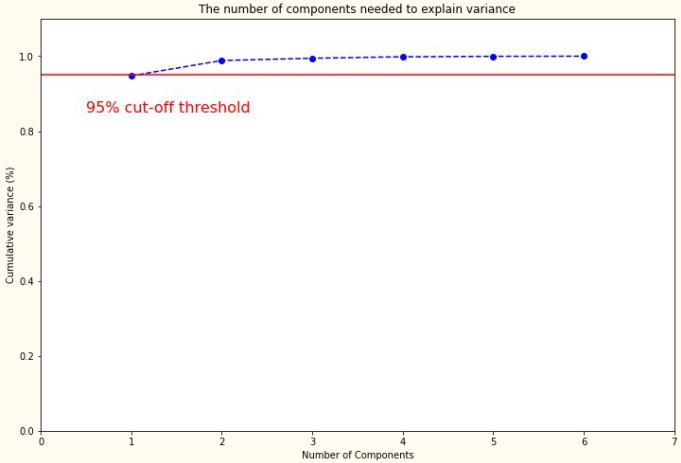
WorkFlow Part 2--Modern ML Method

K-means
(n_clusters=6, n=17378)
→
PCA(2 dimension)
→
Plot()

- Fit the K-means model with 17378 genes, parameters are n_clusters=6 and min_samples=45.
- Decompose the 6-dimensional data of 17378 genes into 2 dimensions.
- Plot the figure.



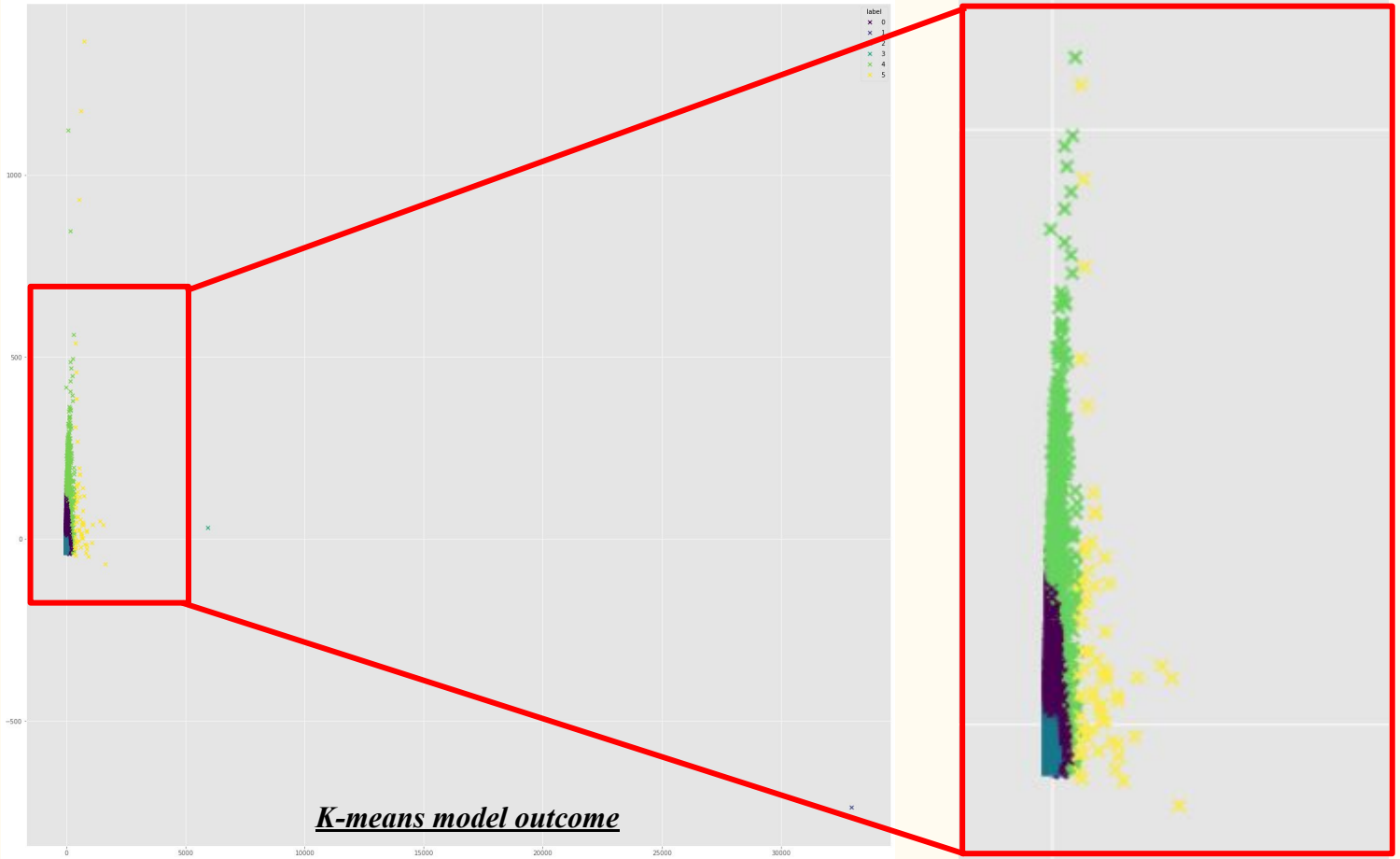
Elbow for K-means clustering



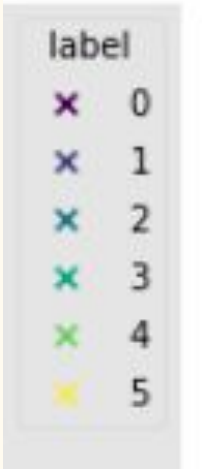
	1	2	3	4	5	6
0	0.947592	0.98843	0.994539	0.998364	0.999606	1.0

The number of PCA components needed to explain variance

WorkFlow Part 2--Modern ML Method



K-means model outcome



2.0	12335
0.0	4041
4.0	931
5.0	69
3.0	1
1.0	1

Name: label, d

K-means labels value count

WorkFlow Part 2--Modern ML Method

- Final outcome after doing ML analysis. We get 259 suspect genes.

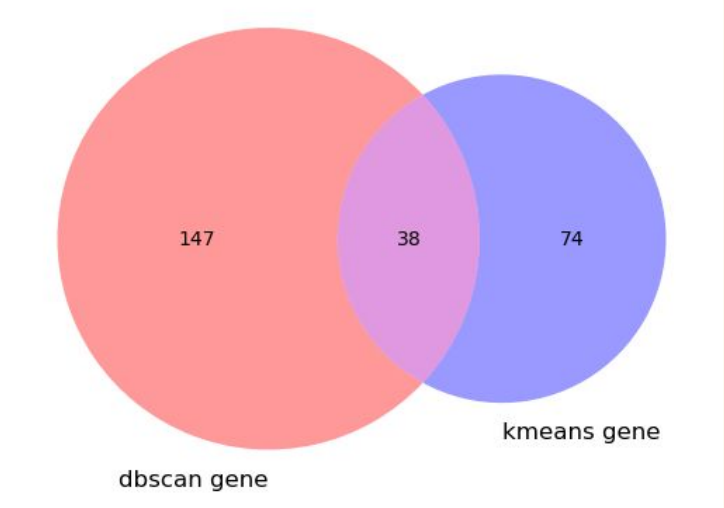
	Geneid	N1	N2	T1	T2	T3	T4
0	C3	1	0	8.833333	495.500000	2.833333	11.666667
1	IFI44L	1	0	23.000000	429.000000	23.500000	10.500000
2	IGHA1	0	0	0.000000	420.000000	0.000000	1.000000
3	NPDC1	0	0	27.000000	421.000000	2.000000	46.000000
4	RAC1	0	1	-19.000000	459.000000	-17.000000	-2.000000
...	...	...	...	...	...	...	...
107	VWA1	0	0	14.000000	142.000000	8.000000	28.000000
108	ZNHIT1	0	1	15.000000	208.000000	7.000000	30.000000
109	ACKR1	0	0	3.000000	1668.000000	0.000000	6.000000
110	HBA1	1	0	-0.090909	1499.727273	-0.454545	0.000000
111	HBA2	0	1	0.075000	1321.587500	0.000000	0.075000

112 rows × 8 columns

	Geneid	N1	N2	T1	T2	T3	T4
0	A2M	10	10	43	312	51	56
1	ACACB	115	105	189	201	150	207
2	ACKR1	0	0	3	1668	0	6
3	ACTB	50	54	41	585	41	253
4	ADGRV1	175	191	220	205	190	155
...	...	...	...	...	...	...	...
180	VPS53	110	115	115	220	128	160
181	WDFY3	109	122	199	148	161	155
182	YWHAB	28	28	47	617	50	69
183	ZEB2	68	93	63	459	76	134
184	ZNF106	116	98	115	358	100	108

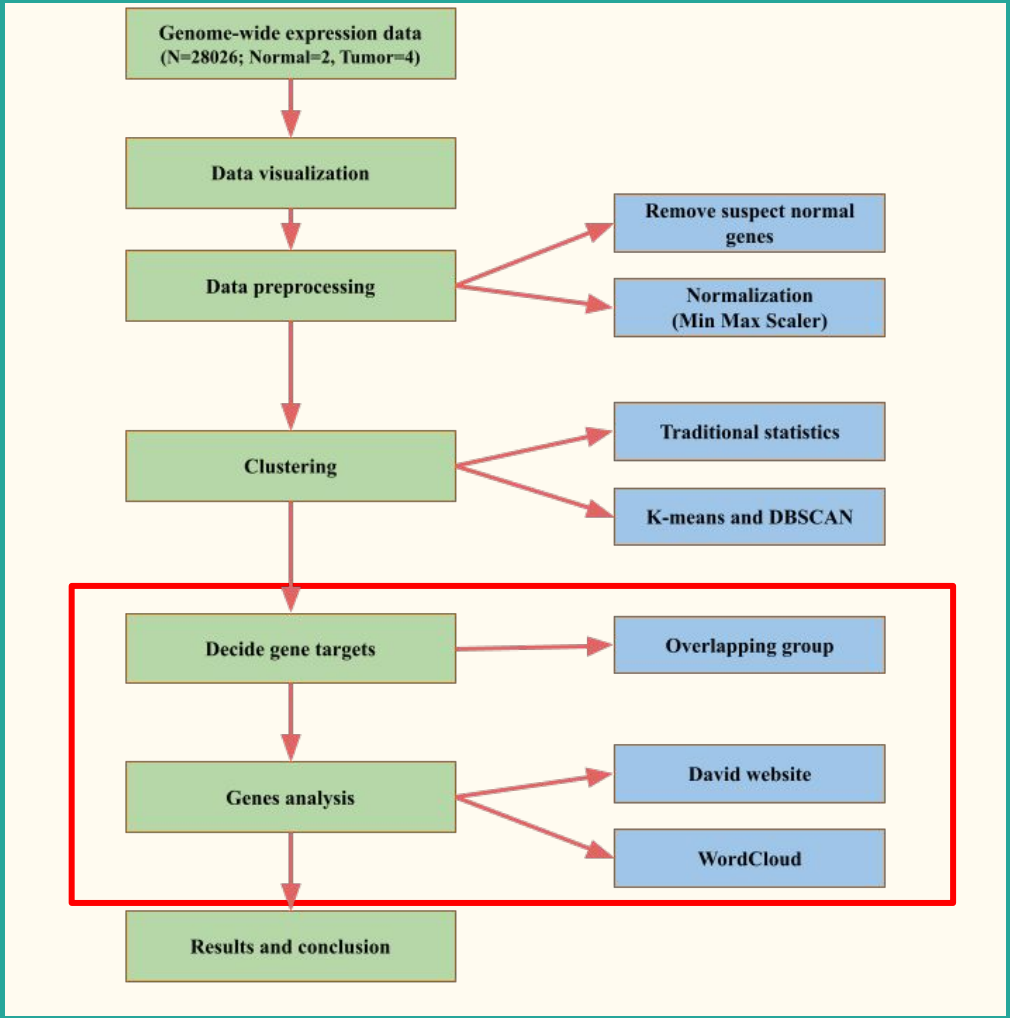
185 rows × 8 columns

Final outcome with ML analysis



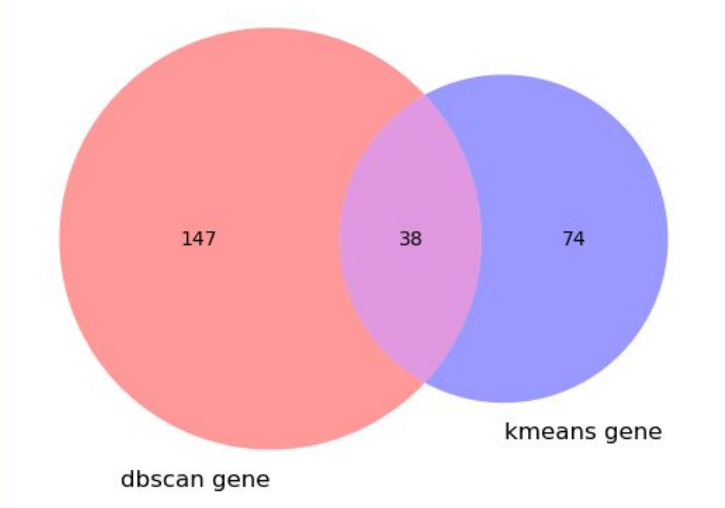
Overlapping group of ML method

Workflow Part 3

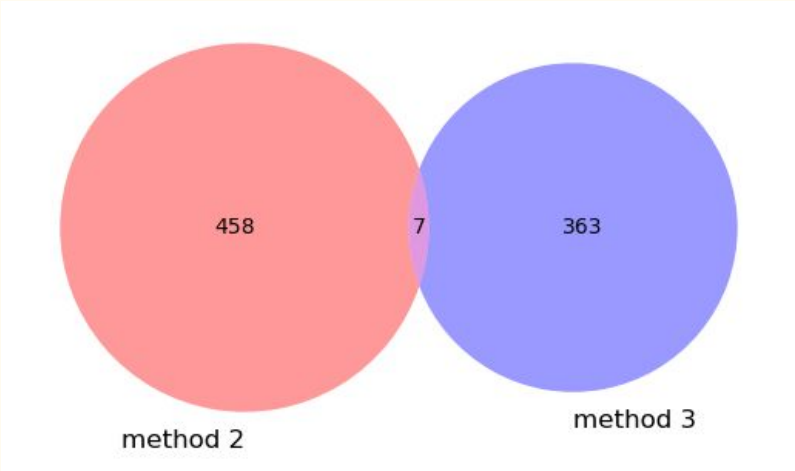


WorkFlow Part 3--Overlapping group

- It seems that there is not much overlap items.



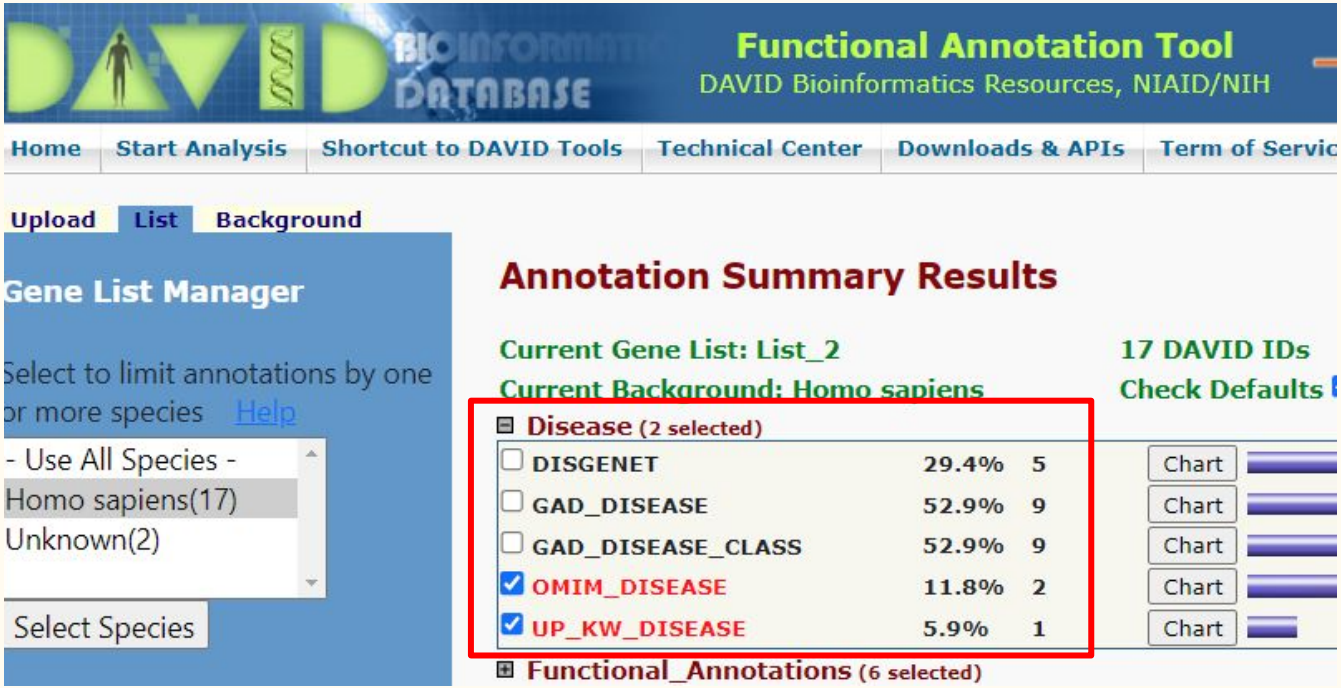
Overlapping group of ML method



Overlapping group of statistics method

WorkFlow Part 3--Web scraping

- Using Python to grab the information that appears on the DAVID website.
- After submitting the gene ID, we only extract the data in "Disease".



Screenshot of DAVID website

WorkFlow Part 3--Web scraping

- The “Term” in the chart is our target. We extract and analyze it to make a word cloud figure.

DAVID Bioinformatics Resources
Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

Functional Annotation Chart

Help and Manual

Current Gene List: List_2
Current Background: Homo sapiens
17 DAVID IDs
Options

Rerun Using Options | Create Sublist

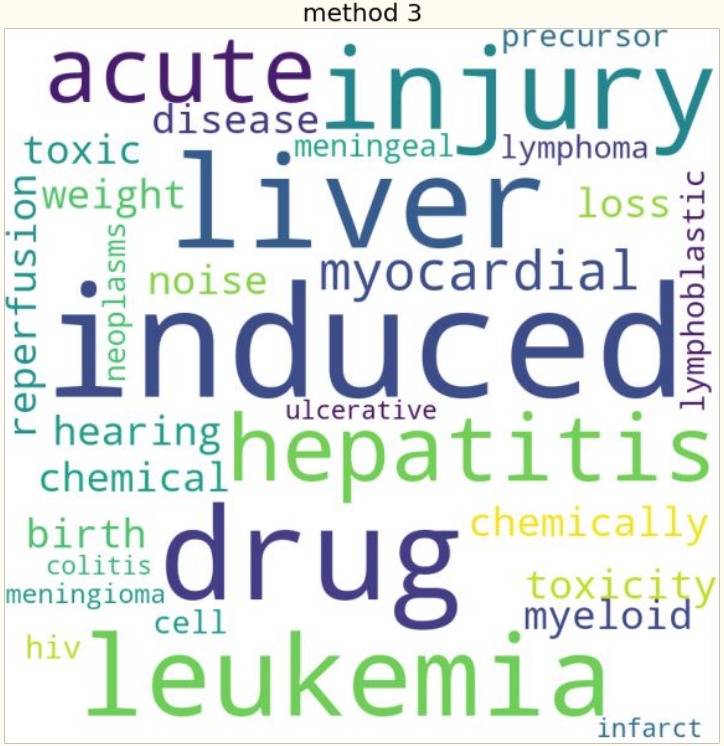
9 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GAD_DISEASE	Asthma|Bronchiolitis, Viral|Respiratory Syncytial Virus Infections	RT		3	17.6	6.5E-3	3.7E-2
	GAD_DISEASE	respiratory syncytial virus bronchiolitis	RT		3	17.6	6.5E-3	3.7E-2
	GAD_DISEASE	Birth Weight|Leukemia|Leukemia, Myeloid, Acute|Precursor Cell Lymphoblastic Leukemia-Lymphoma	RT		3	17.6	6.7E-3	3.7E-2
	GAD_DISEASE	HIV	RT		3	17.6	6.8E-3	3.7E-2
	GAD_DISEASE	Meningeal Neoplasms|meningioma	RT		3	17.6	6.8E-3	3.7E-2
	GAD_DISEASE	Bronchiolitis, Viral|Respiratory Syncytial Virus Infections	RT		3	17.6	7.0E-3	3.7E-2
	GAD_DISEASE	asthma	RT		3	17.6	1.2E-2	5.3E-2
	GAD_DISEASE	ulcerative colitis	RT		2	11.8	5.3E-2	2.1E-1
	GAD_DISEASE	Prostatic Neoplasms	RT		2	11.8	9.4E-2	3.3E-1

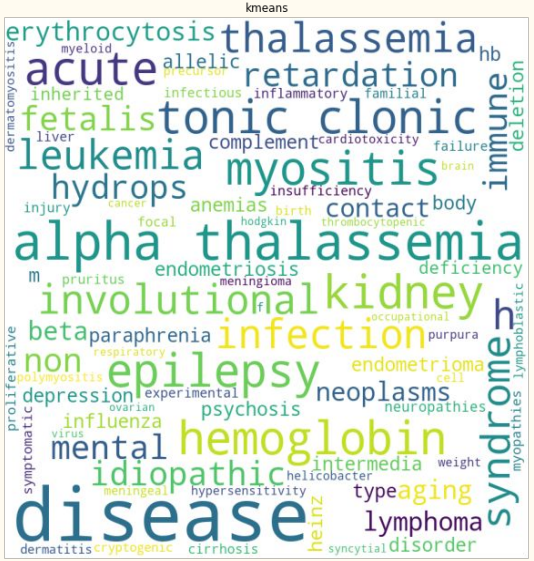
Screenshot of DAVID website -- inside of “Disease”

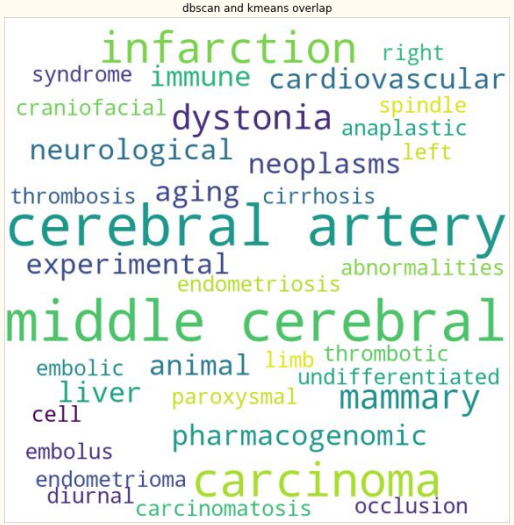
WorkFlow Part 3--Word cloud

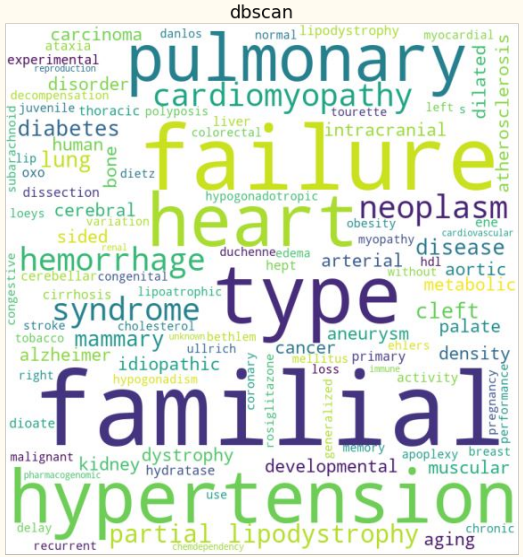


Word cloud results from traditional statistical methods

WorkFlow Part 3--Word cloud

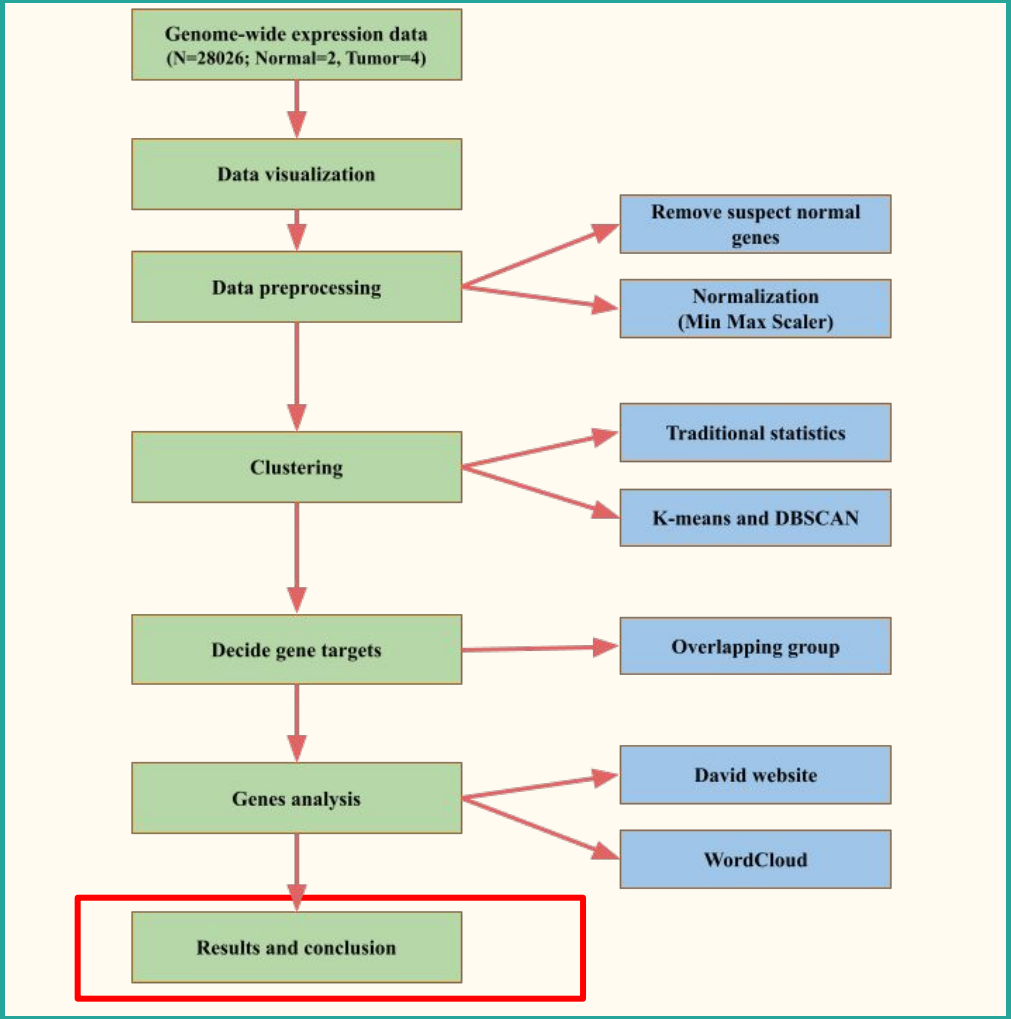




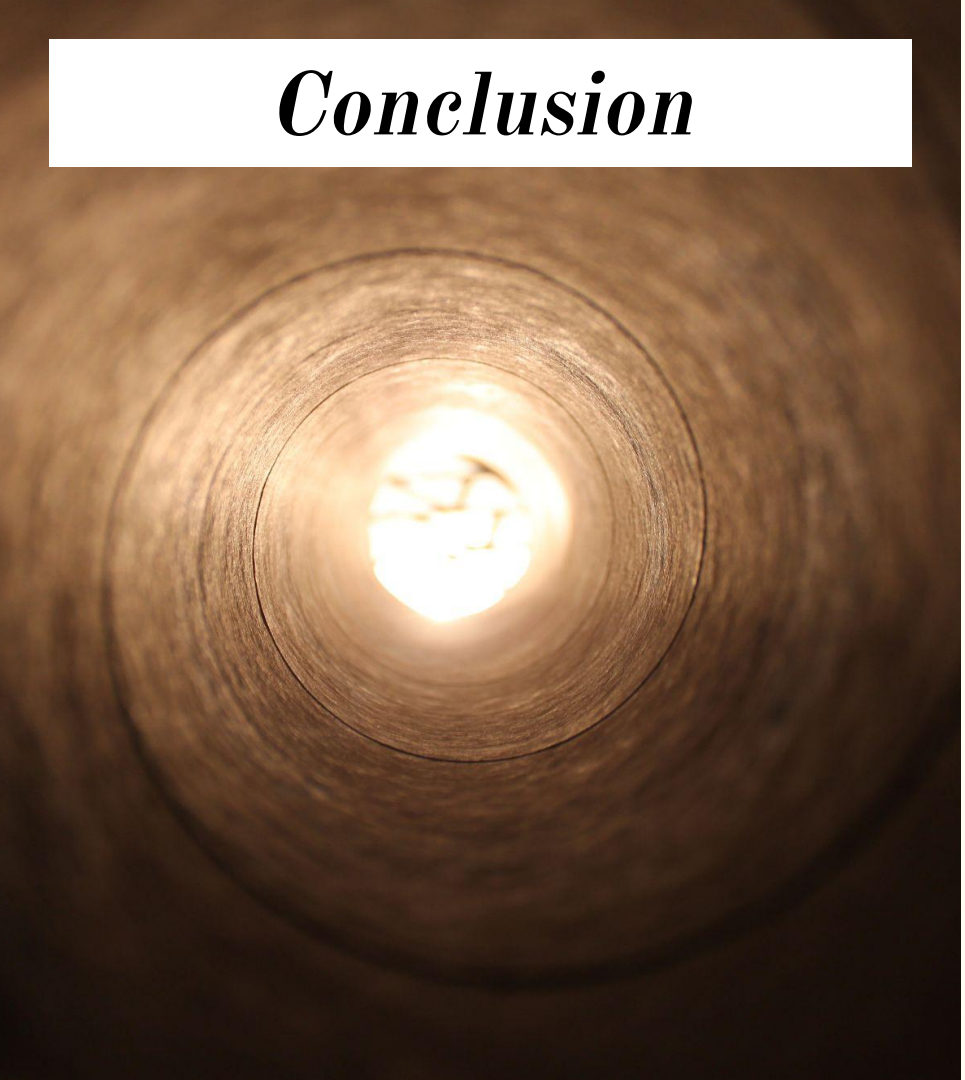


Word cloud results from ML methods

Conclusion


Genome-wide expression data
(N=28026; Normal=2, Tumor=4)
Data visualization
Data preprocessing
Remove suspect normal genes
Normalization
(Min Max Scaler)
Clustering
Traditional statistics
K-means and DBSCAN
Decide gene targets
Overlapping group
Genes analysis
David website
WordCloud
Results and conclusion

Conclusion



1. From the ML method, we believe that the suspected gene may be associated with ovarian cancer. But it requires further research to prove.
   a. one of the reason is that there is not enough research that studies our suspected genes
2. From the statistical method, Some genes that we found have proven to be associated with ovarian cancer.
3. Since our research methods only uses statistics and ML, further research is needed especially with medical analysis.

References

- Quackenbush, J. (2006). Microarray analysis and tumor classification. New England Journal of Medicine, 354(23), 2463-2472.
- Slonim, D. K., & Yanai, I. (2009). Getting started in gene expression microarray analysis. PLoS computational biology, 5(10), e1000543.
- Ye, Y., Dai, Q., & Qi, H. (2021). A novel defined pyroptosis-related gene signature for predicting the prognosis of ovarian cancer. Cell death discovery, 7(1), 1-11.
- Welsh, J. B., Zarrinkar, P. P., Sapinoso, L. M., Kern, S. G., Behling, C. A., Monk, B. J., ... & Hampton, G. M. (2001). Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proceedings of the National Academy of Sciences, 98(3), 1176-1181.
- Yasodha, P., & Ananthanarayanan, N. R. (2018). Detecting the ovarian cancer using big data analysis with effective model. Biomedical Research (0970-938X).