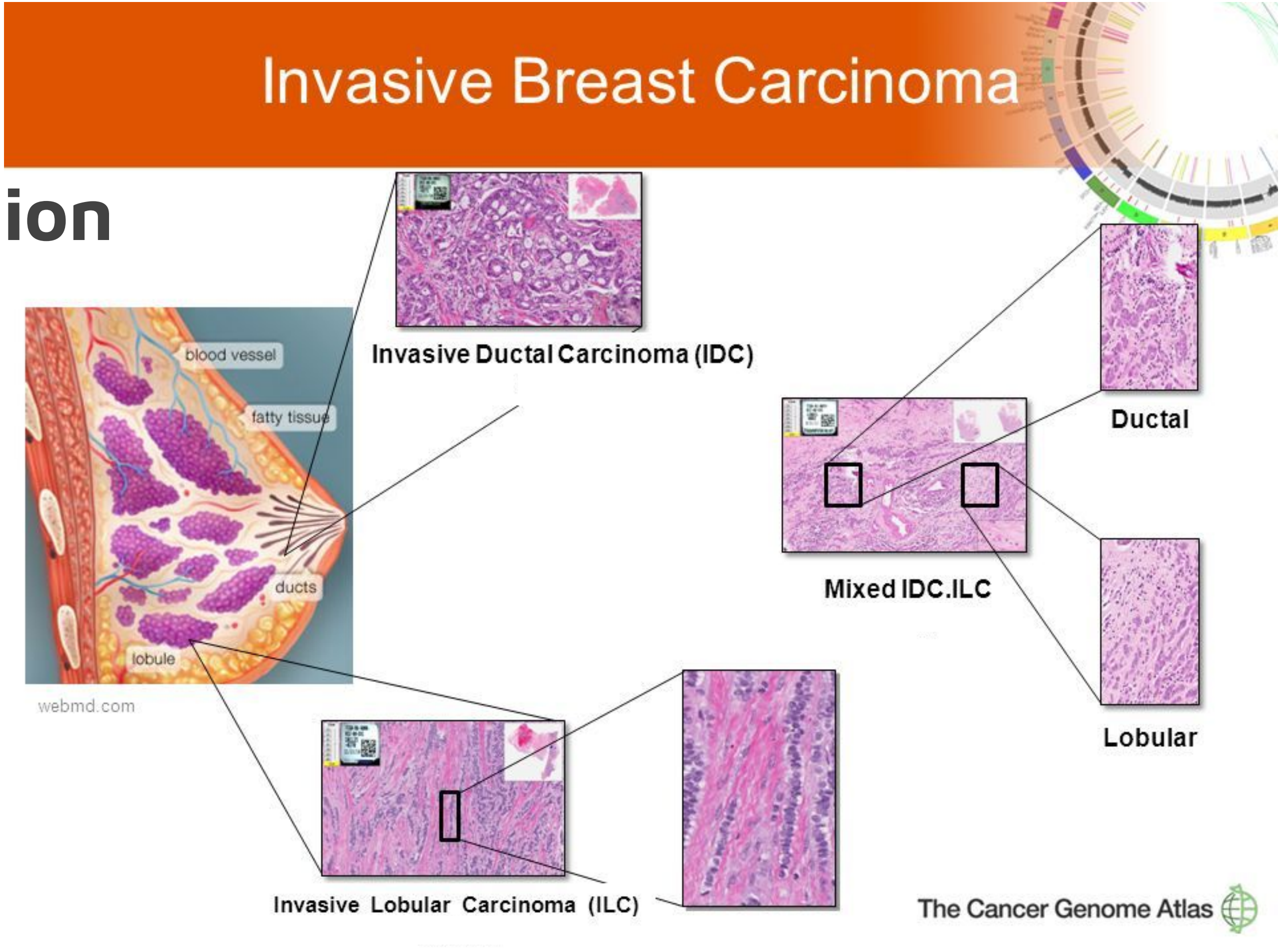
Deciphering Invasive Ductal and Lobular Breast Carcinomas: A Multi-Omics Approach to Clinical Correlations and Molecular Insights

Mehlam Saifudeen, Qilin Zhu, Tesh Pierre

Introduction

- Most breast cancers are invasive
    - Invasive Ductal Carcinoma
    - Invasive Lobular Carcinoma
- Detection method:
    - IDC
        - Mammogram
    - ILC
        - Loss of E-Cadherin
- Limited insights from Genomic studies in the biological underpinnings



https://www.cancer.org/cancer/types/breast-cancer/about/types-of-breast-cancer/invasive-breast-cancer.html

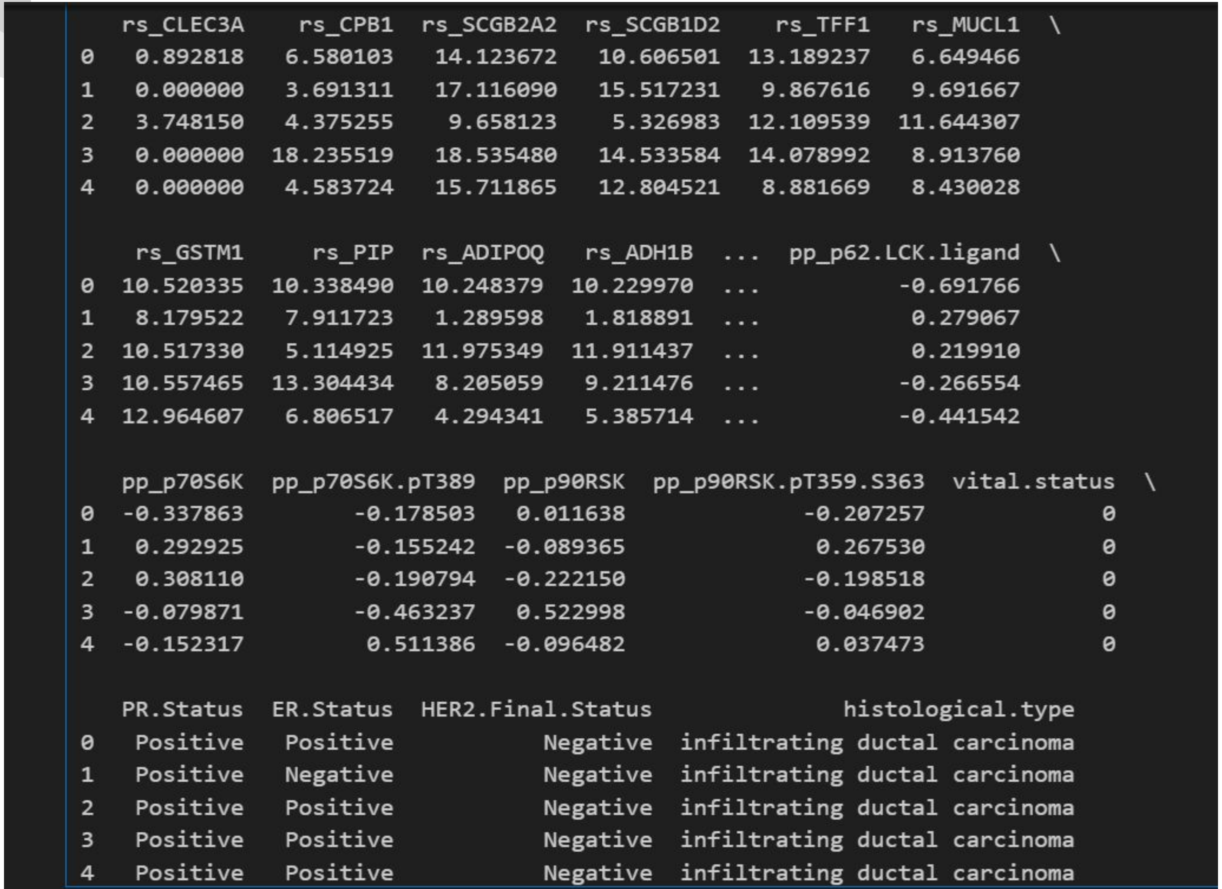
Study Goal

Identify key biomarkers for invasive breast carcinomas to predict vital status and provide a more accurate molecular description of breast cancer

Strategy

- Predict Vital Status Using ML Models
- Gene Enrichment Analysis
- Individual Cluster Analysis

Data Table



```
   rs_CLEC3A    rs_CPB1  rs_SCGB2A2  rs_SCGB1D2    rs_TFF1   rs_MUCL1  \
0   0.892818   6.580103   14.123672   10.606501  13.189237   6.649466
1   0.000000   3.691311   17.116090   15.517231   9.867616   9.691667
2   3.748150   4.375255    9.658123    5.326983  12.109539  11.644307
3   0.000000  18.235519   18.535480   14.533584  14.078992   8.913760
4   0.000000   4.583724   15.711865   12.804521   8.881669   8.430028

    rs_GSTM1     rs_PIP  rs_ADIPOQ   rs_ADH1B  ...  pp_p62.LCK.ligand  \
0  10.520335  10.338490  10.248379  10.229970  ...          -0.691766
1   8.179522   7.911723   1.289598   1.818891  ...           0.279067
2  10.517330   5.114925  11.975349  11.911437  ...           0.219910
3  10.557465  13.304434   8.205059   9.211476  ...          -0.266554
4  12.964607   6.806517   4.294341   5.385714  ...          -0.441542

   pp_p70S6K  pp_p70S6K.pT389  pp_p90RSK  pp_p90RSK.pT359.S363  vital.status  \
0  -0.337863        -0.178503   0.011638             -0.207257             0
1   0.292925        -0.155242  -0.089365              0.267530             0
2   0.308110        -0.190794  -0.222150             -0.198518             0
3  -0.079871        -0.463237   0.522998             -0.046902             0
4  -0.152317         0.511386  -0.096482              0.037473             0

   PR.Status  ER.Status  HER2.Final.Status             histological.type
0   Positive   Positive           Negative  infiltrating ductal carcinoma
1   Positive   Negative           Negative  infiltrating ductal carcinoma
2   Positive   Positive           Negative  infiltrating ductal carcinoma
3   Positive   Positive           Negative  infiltrating ductal carcinoma
4   Positive   Positive           Negative  infiltrating ductal carcinoma
```

705 Breast Tumor Patient Cancer Samples

611 patients survived
94 patients died

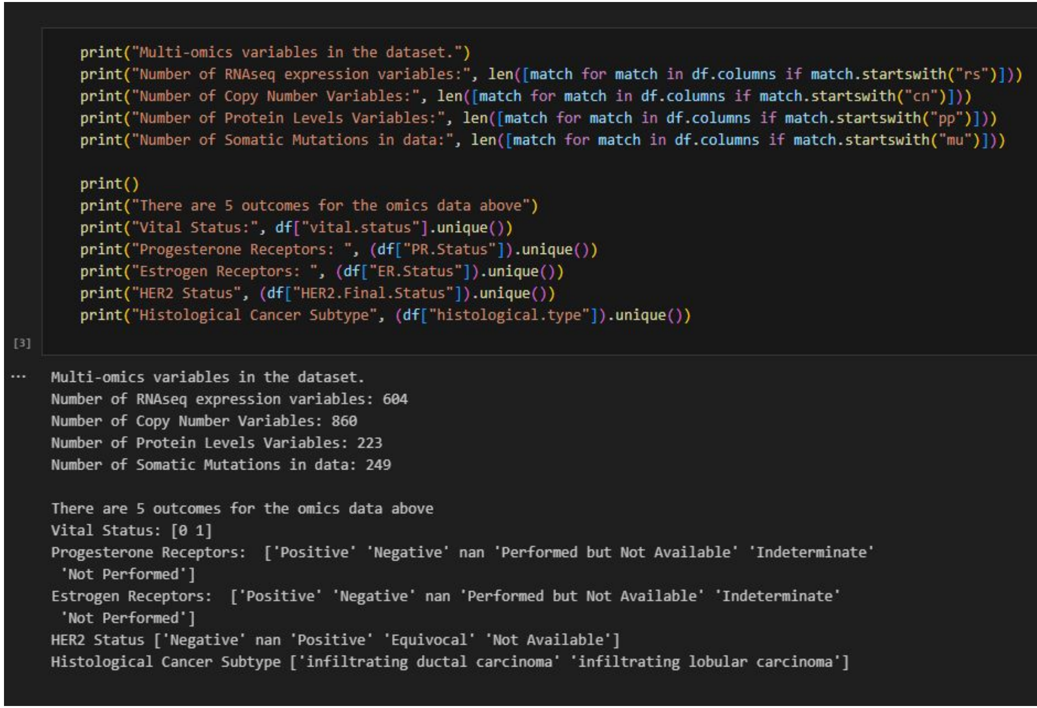
631 BRCA genes

7 BRCA genes with all 4 dimensions of multi-omics data

RS - RNA Seq
PP - Protein Expression
CN - Copy number var
MU - Somatic Mutations

Data Preprocessing and Exploration

- Obtained a dataset of wide range of multi-omics data such as :
  - RNA sequencing expression data
  - Copy Number Variables
  - Protein Expression Levels
  - Somatic Mutations in Data
- Dataset also contains patient outcomes:
  - Vital Status
  - Progesterone Receptors Status (PR Status)
  - Estrogen Receptors Status (ER Status)
  - Human Epidermal Growth Factor Receptor 2 (HER2 Status)
  - Histological Cancer Subtype



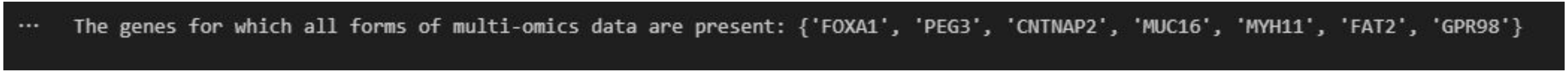
```
print("Multi-omics variables in the dataset.")
print("Number of RNAseq expression variables:", len([match for match in df.columns if match.startswith("rs")]))
print("Number of Copy Number Variables:", len([match for match in df.columns if match.startswith("cn")]))
print("Number of Protein Levels Variables:", len([match for match in df.columns if match.startswith("pp")]))
print("Number of Somatic Mutations in data:", len([match for match in df.columns if match.startswith("mu")]))

print()
print("There are 5 outcomes for the omics data above")
print("Vital Status:", df["vital.status"].unique())
print("Progesterone Receptors: ", (df["PR.Status"]).unique())
print("Estrogen Receptors: ", (df["ER.Status"]).unique())
print("HER2 Status", (df["HER2.Final.Status"]).unique())
print("Histological Cancer Subtype", (df["histological.type"]).unique())
```

[3]

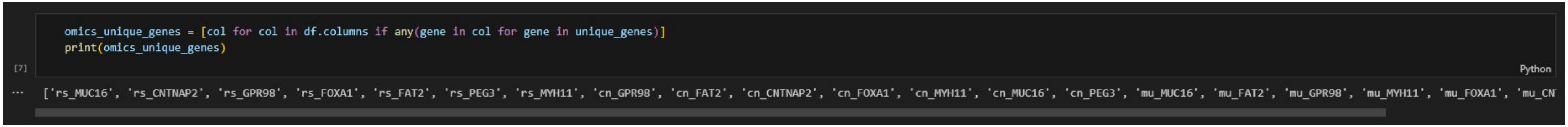
```
Multi-omics variables in the dataset.
Number of RNAseq expression variables: 604
Number of Copy Number Variables: 860
Number of Protein Levels Variables: 223
Number of Somatic Mutations in data: 249

There are 5 outcomes for the omics data above
Vital Status: [0 1]
Progesterone Receptors:  ['Positive' 'Negative' nan 'Performed but Not Available' 'Indeterminate'
 'Not Performed']
Estrogen Receptors:  ['Positive' 'Negative' nan 'Performed but Not Available' 'Indeterminate'
 'Not Performed']
HER2 Status ['Negative' nan 'Positive' 'Equivocal' 'Not Available']
Histological Cancer Subtype ['infiltrating ductal carcinoma' 'infiltrating lobular carcinoma']
```

Obtain set of unique genes with all omics data

- Out of all the genes for BRCA in the dataset, all omics variables were only available for the below genes:



```
The genes for which all forms of multi-omics data are present: {'FOXA1', 'PEG3', 'CNTNAP2', 'MUC16', 'MYH11', 'FAT2', 'GPR98'}
```

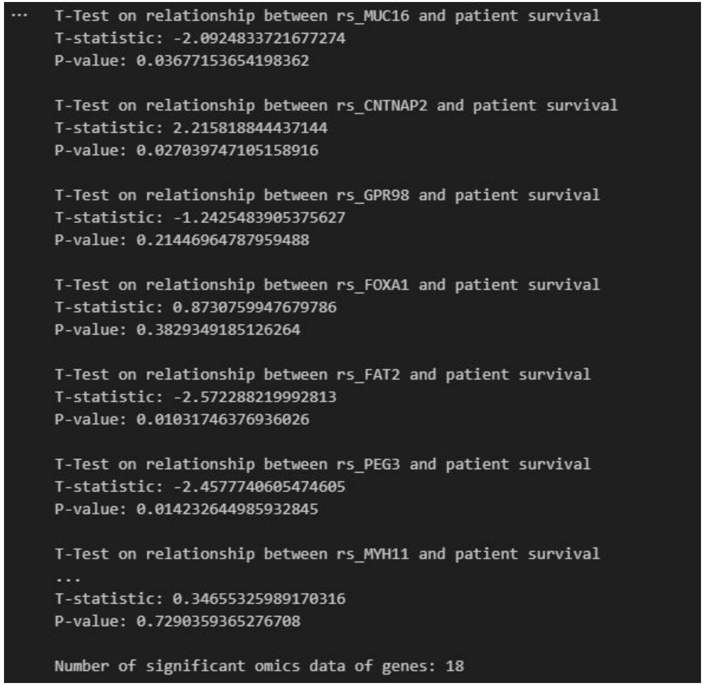


```
omics_unique_genes = [col for col in df.columns if any(gene in col for gene in unique_genes)]
print(omics_unique_genes)
```

```
['rs_MUC16', 'rs_CNTNAP2', 'rs_GPR98', 'rs_FOXA1', 'rs_FAT2', 'rs_PEG3', 'rs_MYH11', 'cn_GPR98', 'cn_FAT2', 'cn_CNTNAP2', 'cn_FOXA1', 'cn_MYH11', 'cn_MUC16', 'cn_PEG3', 'mu_MUC16', 'mu_FAT2', 'mu_GPR98', 'mu_MYH11', 'mu_FOXA1', 'mu_CN
```

Statistical Testing and Analysis

- Conducted statistical analysis in the form of T-test to examine association between omics data and patient survival.
- Obtained p-values from the t-tests to understand the probability of chance playing a role here.
- Out of 28 omics data for 7 genes above, 18 had p-values < 0.05.
- Results indicate that these variables are statistically significant to vital status.
- Can be used to identify biomarkers for prognosis of breast cancer.

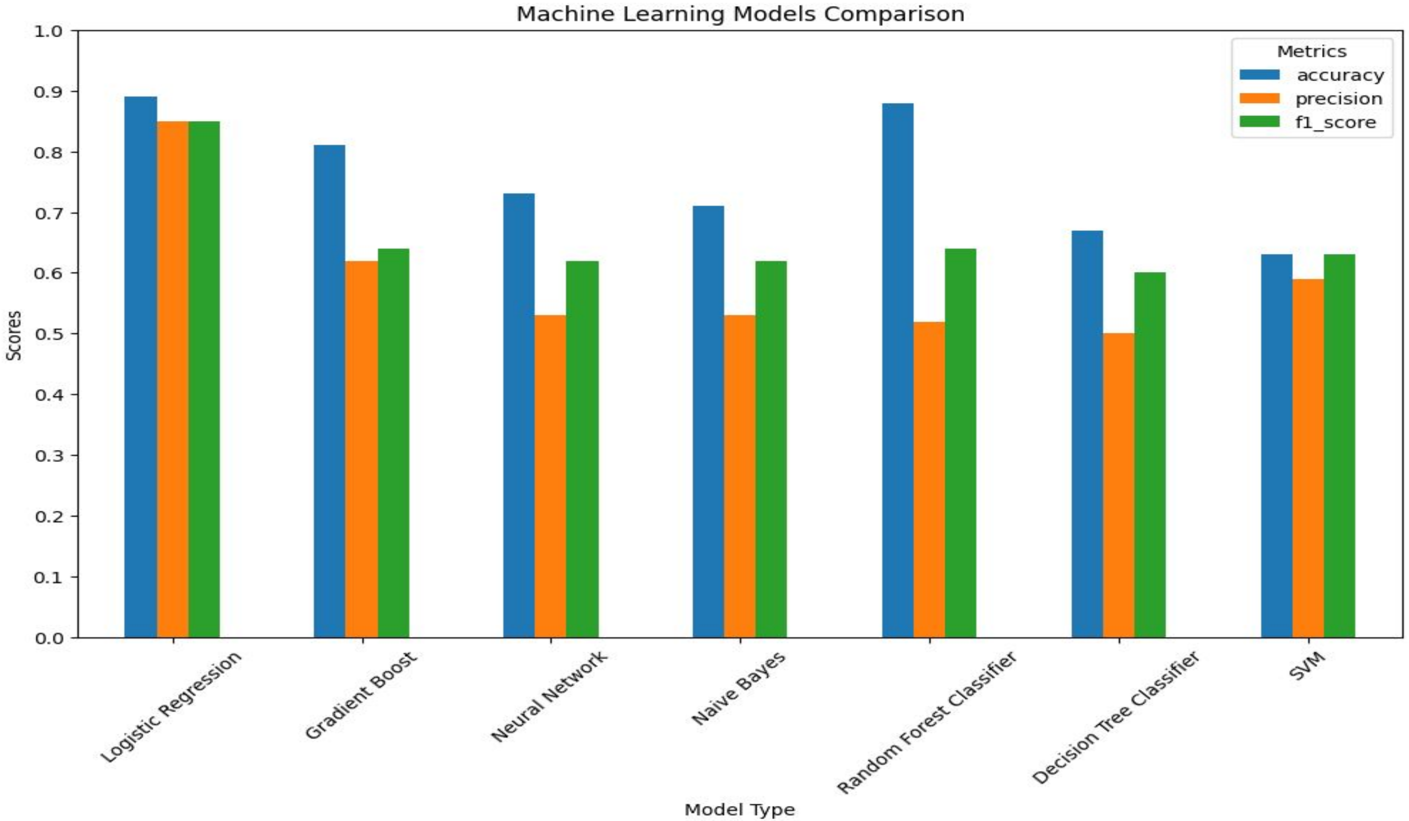


```
...   T-Test on relationship between rs_MUC16 and patient survival
      T-statistic: -2.0924833721677274
      P-value: 0.03677153654198362

      T-Test on relationship between rs_CNTNAP2 and patient survival
      T-statistic: 2.215818844437144
      P-value: 0.027039747105158916

      T-Test on relationship between rs_GPR98 and patient survival
      T-statistic: -1.2425483905375627
      P-value: 0.21446964787959488

      T-Test on relationship between rs_FOXA1 and patient survival
      T-statistic: 0.8730759947679786
      P-value: 0.3829349185126264

      T-Test on relationship between rs_FAT2 and patient survival
      T-statistic: -2.572288219992813
      P-value: 0.01031746376936026

      T-Test on relationship between rs_PEG3 and patient survival
      T-statistic: -2.4577740605474605
      P-value: 0.014232644985932845

      T-Test on relationship between rs_MYH11 and patient survival
      ...
      T-statistic: 0.34655325989170316
      P-value: 0.7290359365276708

      Number of significant omics data of genes: 18
```

Machine Learning Models Outcomes

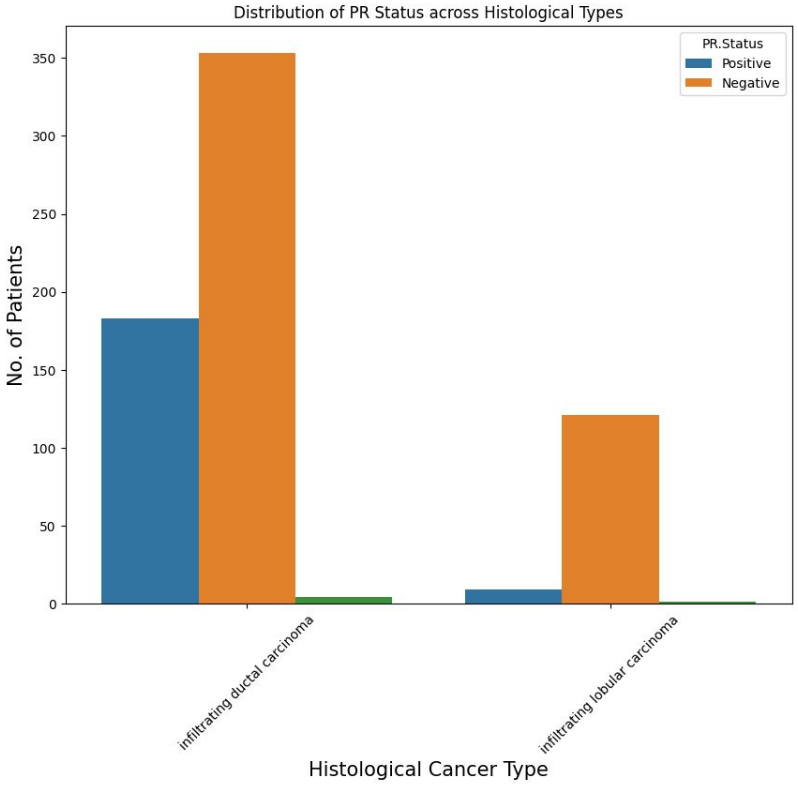
- Deployed 7 different Machine Learning (ML) models to predict vital status of breast cancer patients.
- These vary across ensemble methods (Random Forest and Gradient Boost), SVM, Deep Learning (Neural Networks), linear classifier (Logistic Regression), Naive Bayes and non-linear classifiers (Decision Trees).
- Performance metrics - Accuracy, F1 score and Precision - were used to show how each model compared to the others.
- All models performed well overall.

Comparison of ML Models Metrics



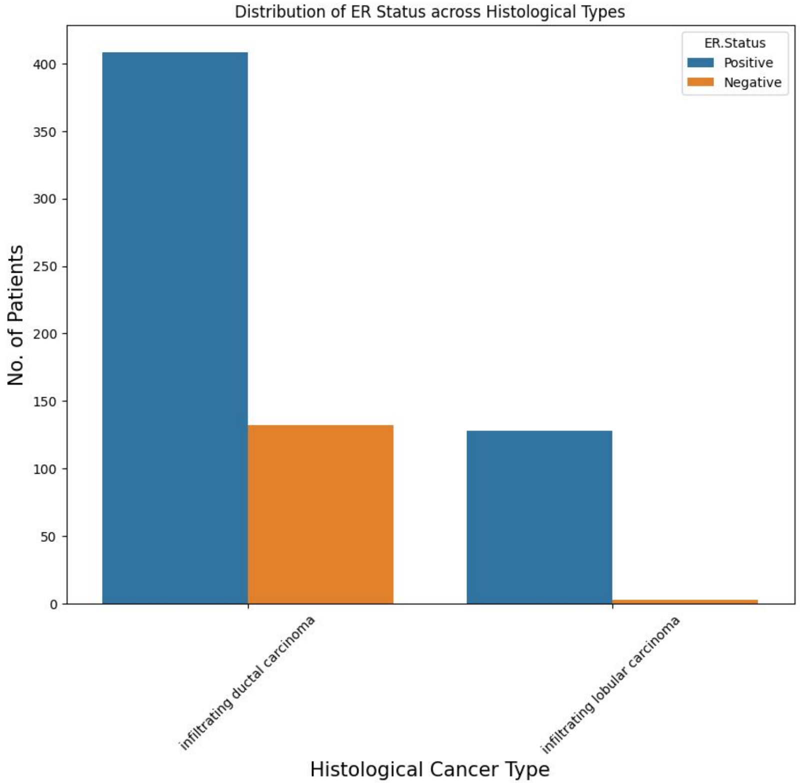
Hormonal Clinical Biomarker Identification Across Histological Types (PR Status)

- Analyze PR Status in patients across histological types.
- Can be used to personalize treatment for patients.
- Low prevalence of PR Negative status for both cancer subtypes suggests that hormone therapy may not be an effective treatment option.
- **Chi-square test for PR.Status and histological type, p-value**: 6.274956314187463e-09



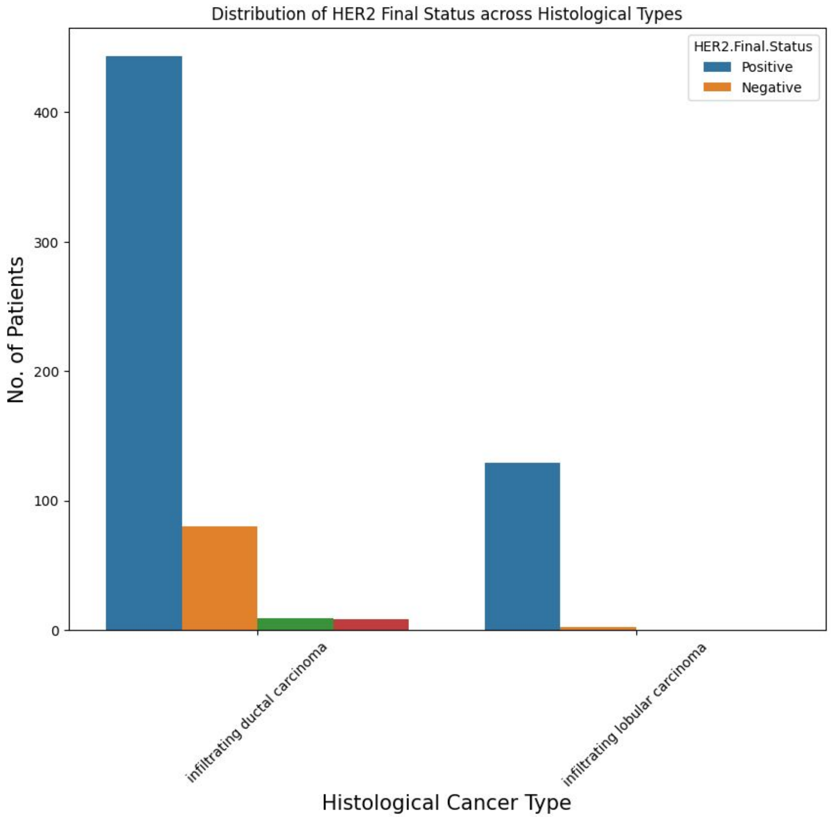
(ER Status)

- Analyze ER Status in patients across histological types.
- High prevalence of ER Positive status for both cancer subtypes suggests that hormone therapy can be an effective treatment option for a lot of patients.
- **Chi-square test for ER.Status and histological type, p-value**: 2.8125967930730625e-08

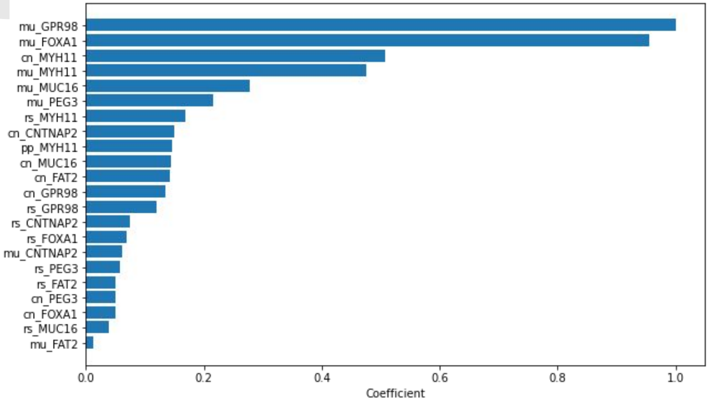


(HER2 Status)

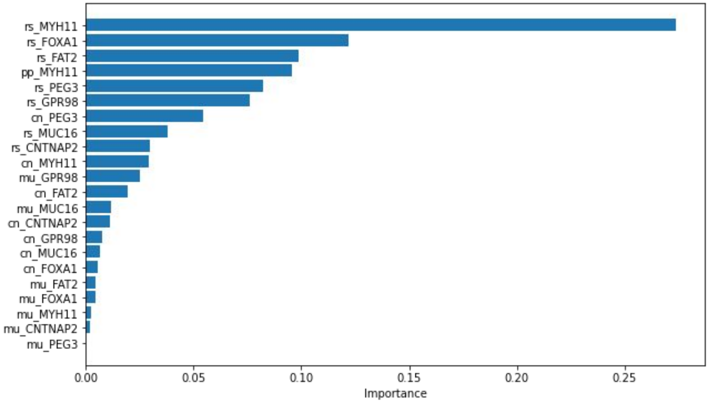
- Analyze HER2 Status in patients across histological types.
- High prevalence of ER Positive status for both cancer subtypes suggests that hormone therapy can be an effective treatment option for a lot of patients.
- **Chi-square test for HER2.Final.Status and histological type, p-value:** 4.668546114461214e-05



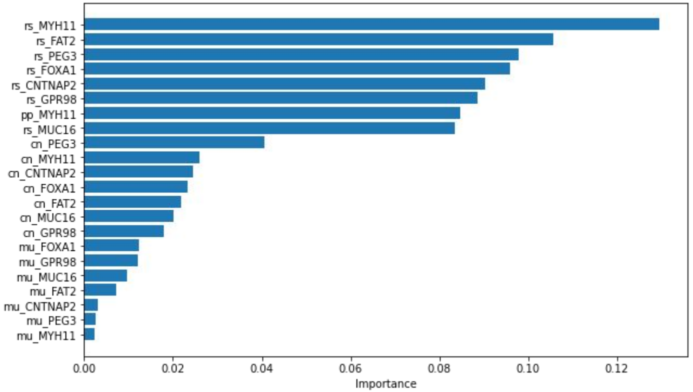
Feature Importance



Feature Coefficient from Logistic Regresion

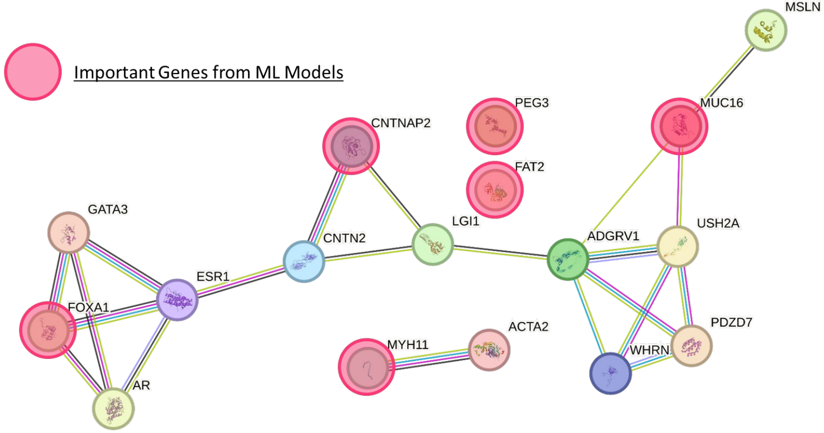


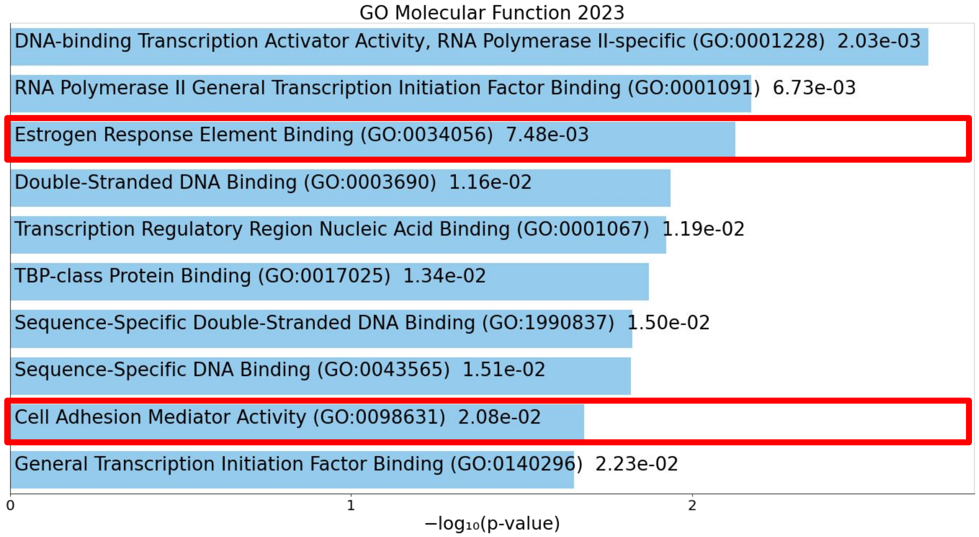
Feature Importance from Gradient Boost



Feature Importance from RandomForest

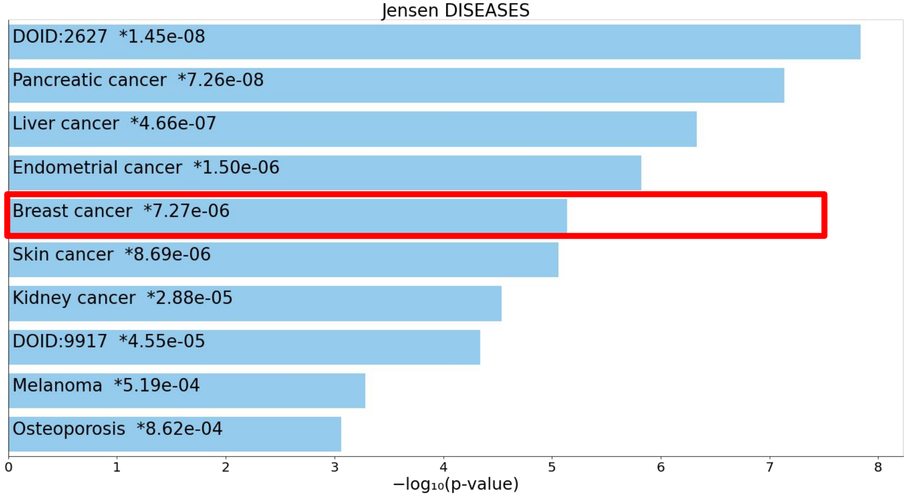
Gene Enrichment Based on Top Genes

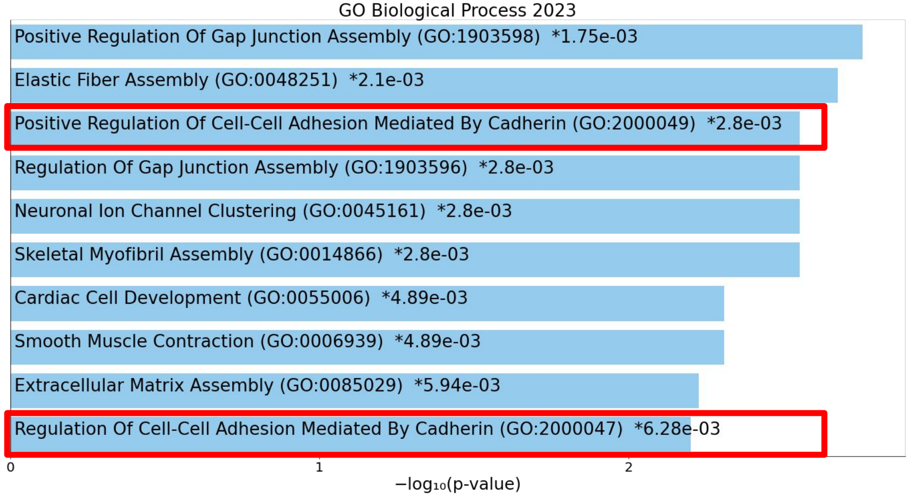




https://string-db.org/; https://maayanlab.cloud/Enrichr/

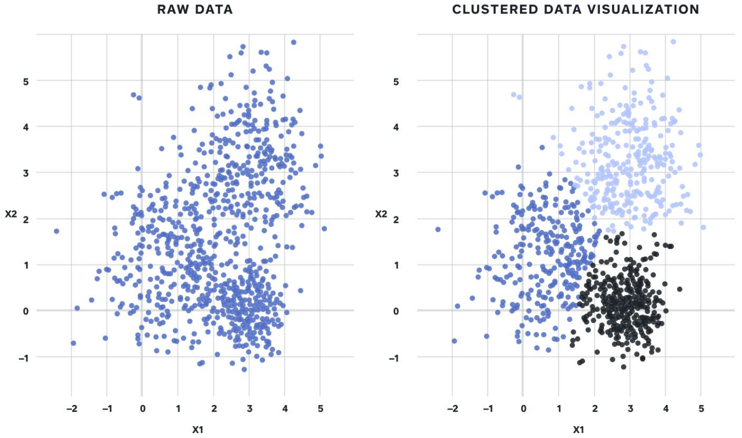
Gene Enrichment Based on Top Genes





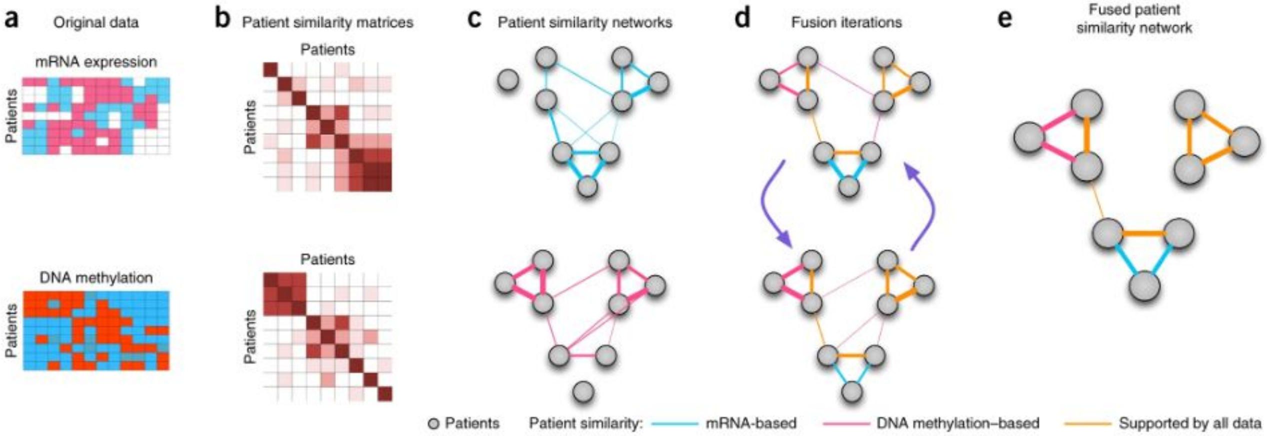
Clustering

- Importance and Necessity
  - Make groups which can carry biological or pathological meaning
  - Explore the potential complex biological relationships and patterns
- Similarity Network Fusion(SNF)



Similarity Network Fusion(SNF)

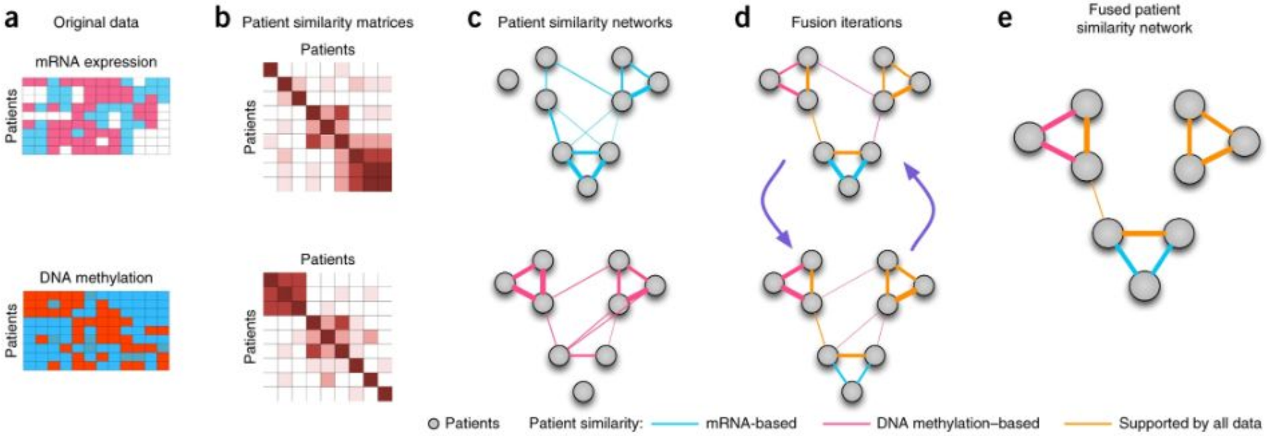
- Receive raw original multi-omics data
- Create similarity matrices/networks
- Fusion iterations
- Fused similarity network



Similarity Network Fusion(SNF)

- Data pre-processing
  - Outlier removal
  - Missing-data imputation
  - Data normalization

$$\tilde{f} = \frac{f - E(f)}{\sqrt{Var(f)}}$$



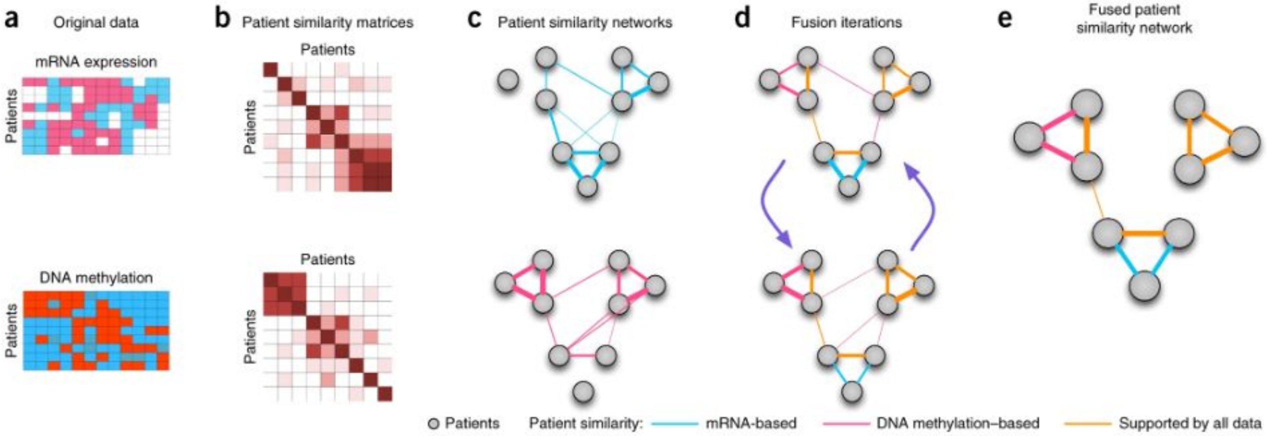
Similarity Network Fusion(SNF)

- Create similarity matrices/network

$$G = (V, E)$$

$$W(i,j) = \exp\left(-\frac{\rho^2(x_i, x_j)}{\mu\epsilon_{i,j}}\right)$$

$$\epsilon_{i,j} = \frac{\mathrm{mean}(\rho(x_i, N_i)) + \mathrm{mean}(\rho(x_j, N_j)) + \rho(x_i, x_j)}{3}$$



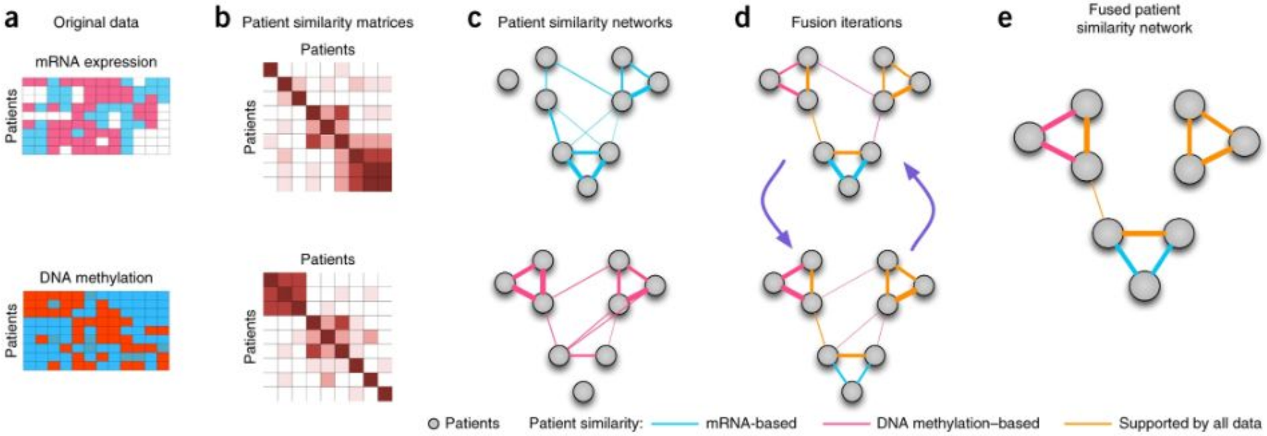
Similarity Network Fusion(SNF)

- Create similarity matrices/network

$$P(i,j) = \begin{cases} \frac{W(i,j)}{\sum_{k \neq i} W(i,k)} & \text{if } j \neq i \\ \frac{1}{2} & \text{if } j = i \end{cases}$$

- Calculate local affinity

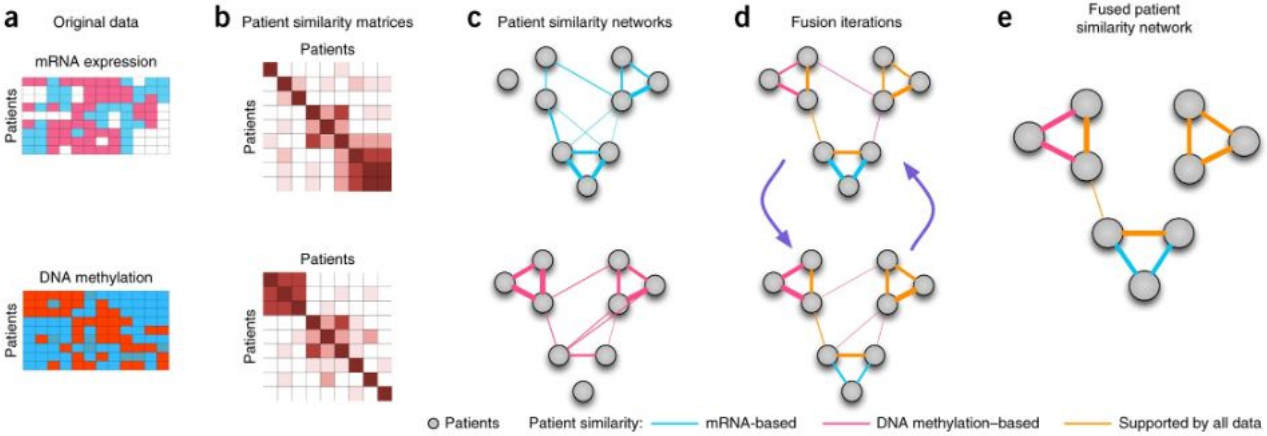
$$S(i,j) = \begin{cases} \frac{W(i,j)}{\sum_{k \in N_i} W(i,k)} & \text{if } j \in N_i \\ 0 & \text{otherwise} \end{cases}$$



Similarity Network Fusion(SNF)

- Initial set up (Assume 2 measurements for easier explanation)

$$P^{(1)}_{t=0} = P^{(1)} \quad P^{(2)}_{t=0} = P^{(2)} \text{ at } t = 0 \quad S^{(1)} \text{ and } S^{(2)}$$



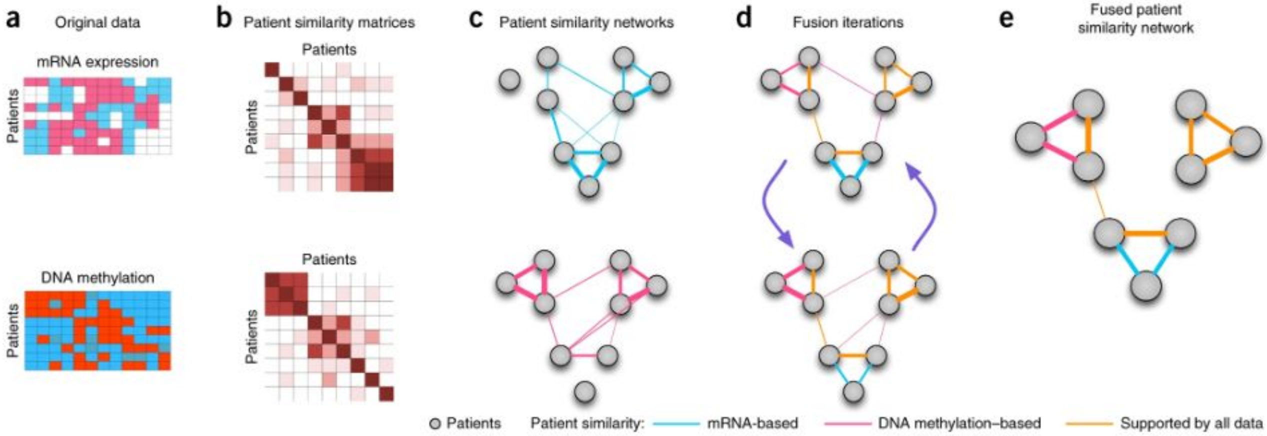
Similarity Network Fusion(SNF)

- Fusion iterations

$$P_{i+1}^{(1)} = S^{(1)} \times P_i^{(2)} \times (S^{(1)})^T$$

$$P_{i+1}^{(2)} = S^{(2)} \times P_i^{(1)} \times (S^{(2)})^T$$

$$P^{(v)} = s^{(v)} \times \left(\frac{\sum_{k \neq v} P^{(k)}}{m-1} \right) \times (s^{(v)})^T, \quad v = 1, 2, \ldots, m$$

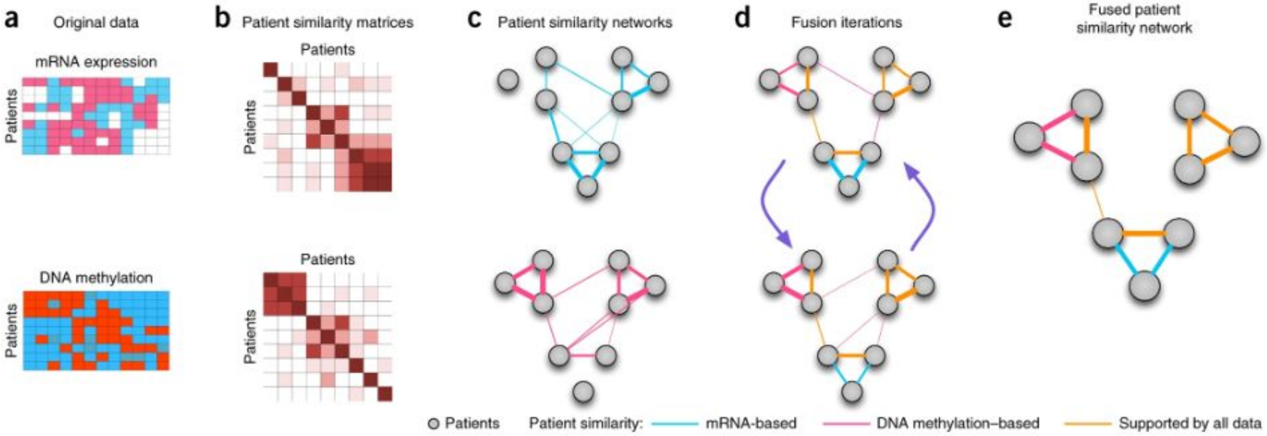


Similarity Network Fusion(SNF)

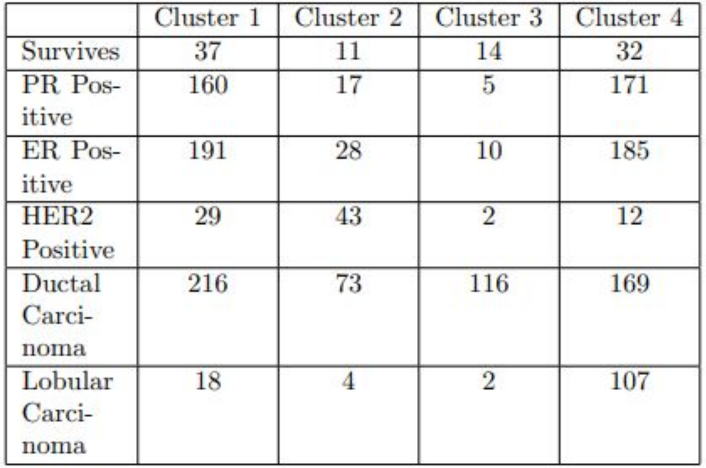
- Fused similarity network

$$P^c = \frac{P_t^1 + P_t^2}{2}$$

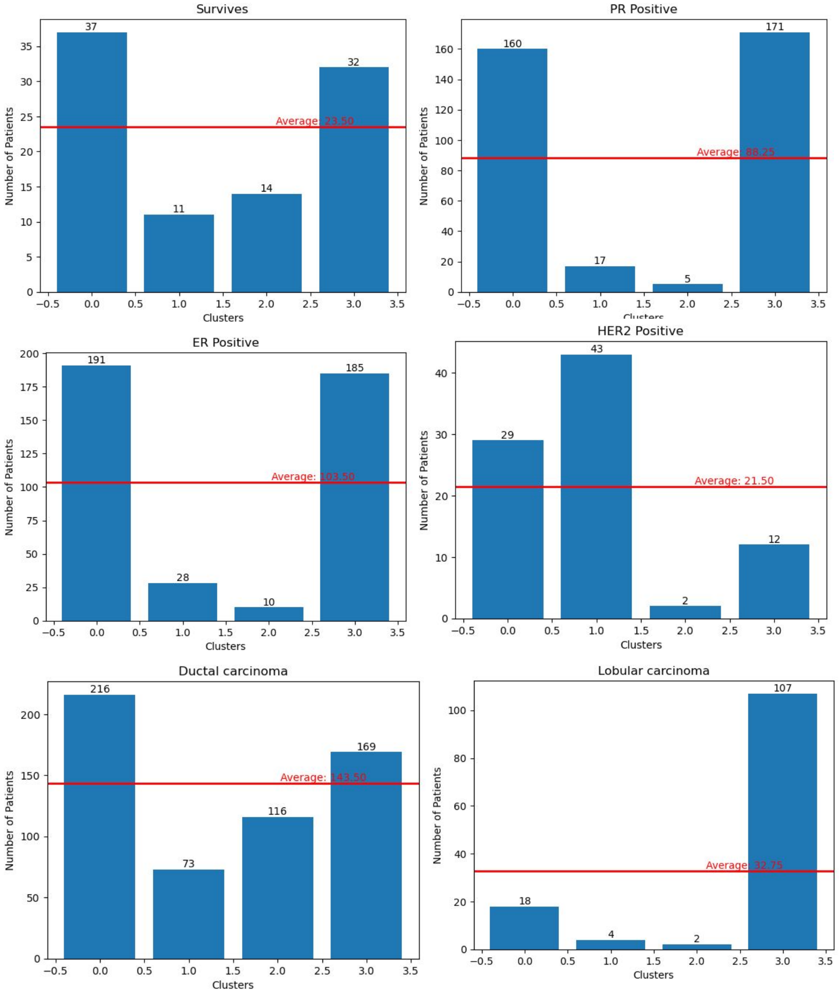
- Spectral clustering



Clustering Result & Analysis



	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Survives	37	11	14	32
PR Positive	160	17	5	171
ER Positive	191	28	10	185
HER2 Positive	29	43	2	12
Ductal Carcinoma	216	73	116	169
Lobular Carcinoma	18	4	2	107



Conclusion

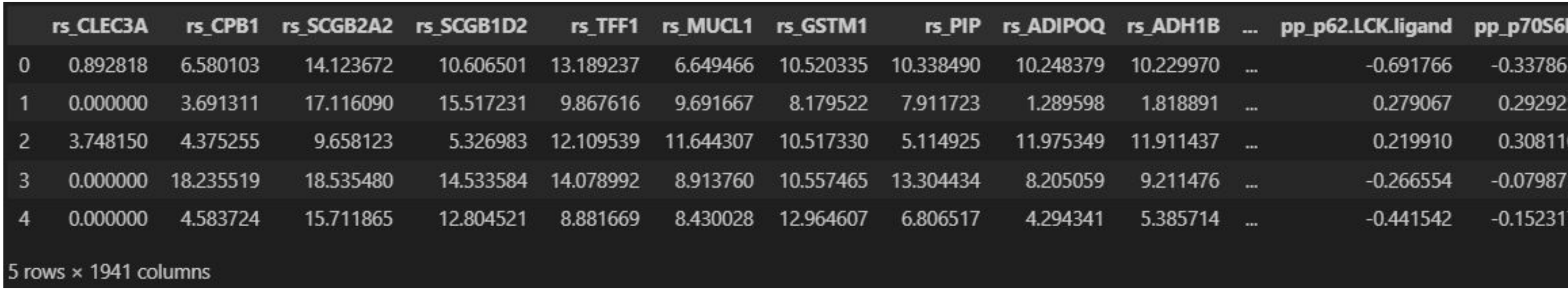
- Rich multi-omics datasets can be used to analyze and generate biomarkers for disease types leading to better treatment options and patient outcomes.
- Random Forest, Logistic Regression, and Gradient Boost had the highest accuracy to predict vital status given the multi-omics dataset.
- SNF clustering method show a potential way to explore the connection between multi-omics data to output statues.
- These models were then used to analyze the feature importance of an individual genes to show that MYH11, FOXA1 and GPR98 have the greatest effect on vital status of a patient
- Receptor statuses are an effective way to analyze the best treatment options for patients with specific cancer types.

References

- Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., Bowlby, R., Shen, H., Hayat, S., Fieldhouse, R., Lester, S. C., Tse, G. M. K., Factor, R. E., Collins, L. C., Allison, K. H., ... Perou, C. M. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell*, *163*(2), 506-519. https://doi.org/10.1016/j.cell.2015.09.033
- Smith, J., Doe, A., & Roe, B. (2020). "Estrogen and Progesterone Receptor Status in Breast Cancer: From Biological Significance to Clinical Implications." Journal of Clinical Oncology, 38(15), 1432-1444.
- Johnson, L., & Patel, S. (2021). "HER2-Positive Breast Cancer: Molecular Pathways and Therapeutic Targets." The New England Journal of Medicine, 384(3), 283-295.
- Lee, C. H., & Kim, Y. J. (2019). "A Comprehensive Review of the Role of Biomarkers in Predicting Disease Progression and Treatment Response in Breast Cancer." Breast Cancer Research and Treatment, 178(2), 259-273.
- https://www.cancer.org/cancer/types/breast-cancer/about/types-of-breast-cancer/invasive-breast-cancer.html
- https://developer.squareup.com/blog/so-you-have-some-clusters-now-what/
- Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhidong Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. Nature Methods, 11(3):333–337, 2014.

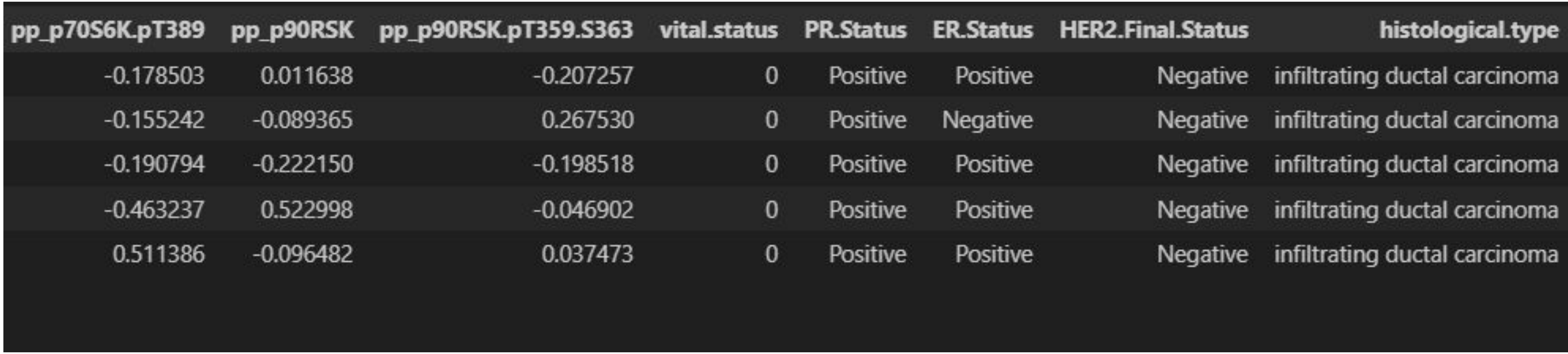
Questions???

Data Sets



	rs_CLEC3A	rs_CPB1	rs_SCGB2A2	rs_SCGB1D2	rs_TFF1	rs_MUCL1	rs_GSTM1	rs_PIP	rs_ADIPOQ	rs_ADH1B	...	pp_p62.LCK.ligand	pp_p70S6
0	0.892818	6.580103	14.123672	10.606501	13.189237	6.649466	10.520335	10.338490	10.248379	10.229970	...	-0.691766	-0.33786
1	0.000000	3.691311	17.116090	15.517231	9.867616	9.691667	8.179522	7.911723	1.289598	1.818891	...	0.279067	0.29292
2	3.748150	4.375255	9.658123	5.326983	12.109539	11.644307	10.517330	5.114925	11.975349	11.911437	...	0.219910	0.30811
3	0.000000	18.235519	18.535480	14.533584	14.078992	8.913760	10.557465	13.304434	8.205059	9.211476	...	-0.266554	-0.07987
4	0.000000	4.583724	15.711865	12.804521	8.881669	8.430028	12.964607	6.806517	4.294341	5.385714	...	-0.441542	-0.15231

5 rows × 1941 columns



pp_p70S6K.pT389	pp_p90RSK	pp_p90RSK.pT359.S363	vital.status	PR.Status	ER.Status	HER2.Final.Status	histological.type
-0.178503	0.011638	-0.207257	0	Positive	Positive	Negative	infiltrating ductal carcinoma
-0.155242	-0.089365	0.267530	0	Positive	Negative	Negative	infiltrating ductal carcinoma
-0.190794	-0.222150	-0.198518	0	Positive	Positive	Negative	infiltrating ductal carcinoma
-0.463237	0.522998	-0.046902	0	Positive	Positive	Negative	infiltrating ductal carcinoma
0.511386	-0.096482	0.037473	0	Positive	Positive	Negative	infiltrating ductal carcinoma