

Informatics links between histological features and genetics in cancer

Kun Huang, PhD

**Department of Biomedical Informatics
The Ohio State University Wexner Medical Center**

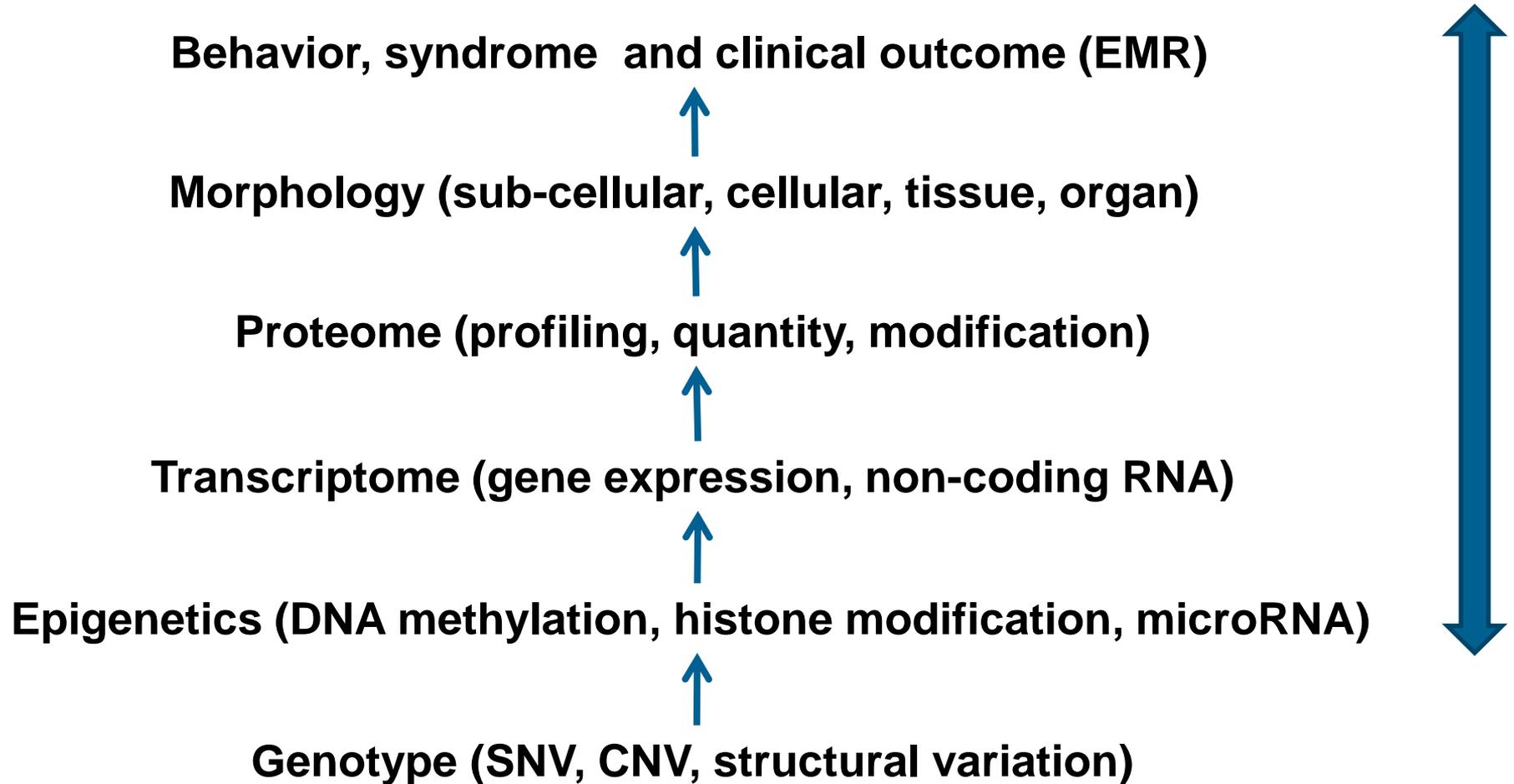
June 13, 2016



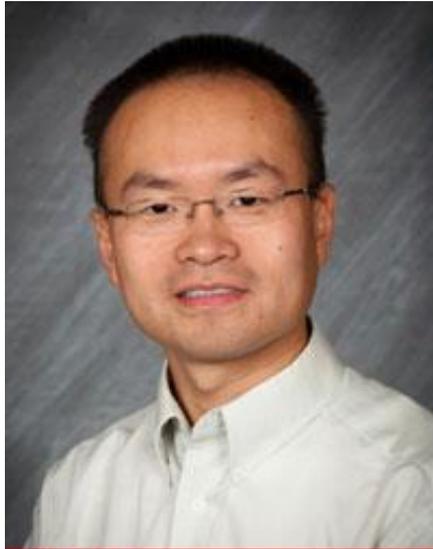
Wexner Medical Center

Data Integration

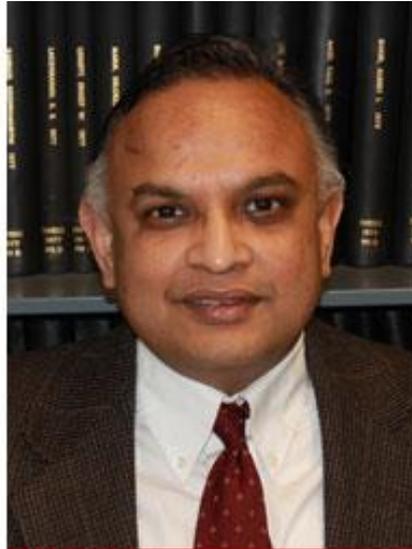
- **Integrative genomics / trans-omics approach**



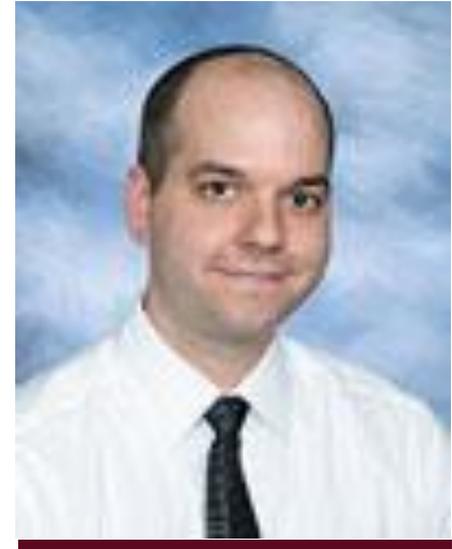
Team



Huang



Machiraju



Jeff Baumes, Kitware



Jie Zhang



*David Manthey,
Kitware*



Leveraging

- NCI CPTAC Contract
 - Integrate proteomics data from CPTAC project
 - High performance computing (GPU and cluster)



HPC



Proteomics



Wexner Medical Center

Tasks

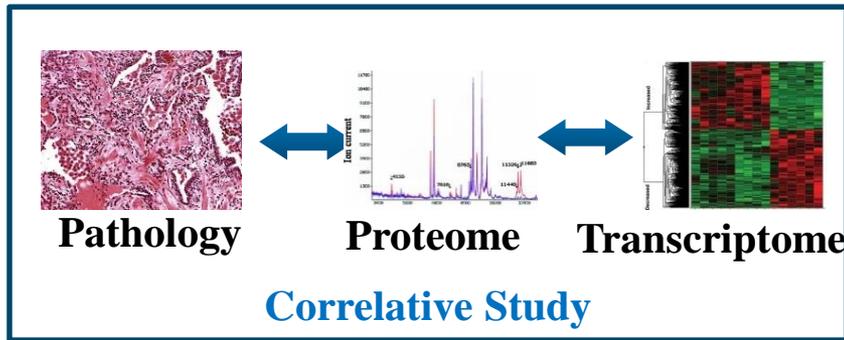
Aim 1 – Develop software libraries for integrative genomics in cancer research, specifically for ***integrating genomic, histological images and clinical data for cancer biomarker discovery and subtyping.***

Aim 2 – Develop an integrative and expandable open source platform for ***managing, analyzing, and integrating multiple data types*** in integrative genomics for cancer with ***visual analytic capabilities*** for cancer biomarker discovery.

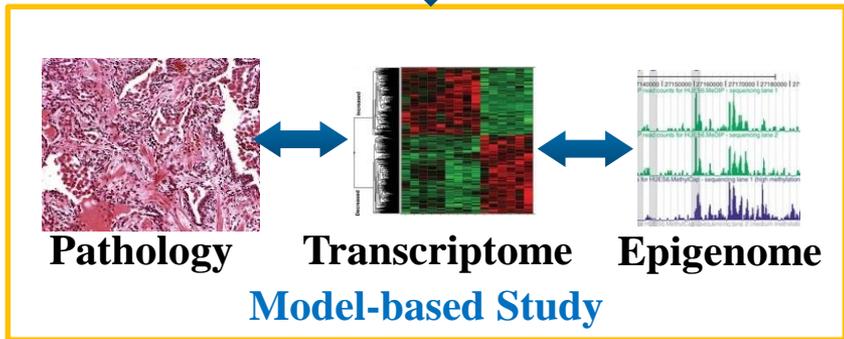
Aim 3 – Test the completed ***software platform*** with cancer systems biology studies and build ***an ecosystem based on the open source framework for integrative genomics*** and in particular for imaging genomics in cancer.



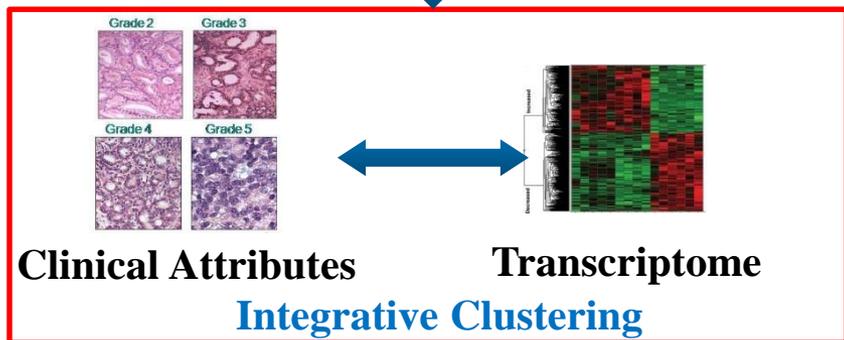
From Correlation to Integration



- Applications**
- Triple Negative Breast Cancer
 - Breast Cancer Proteomics

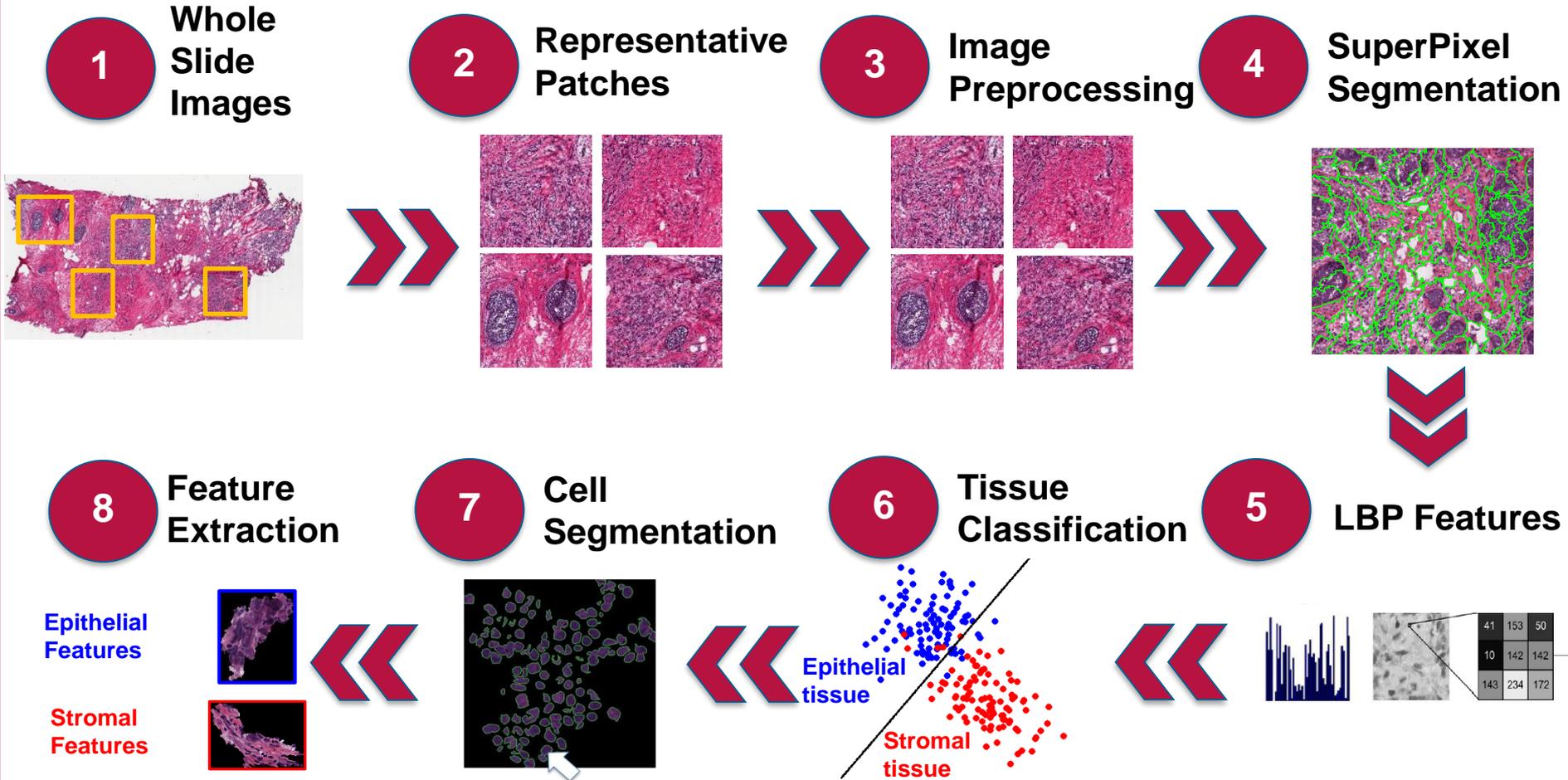


- Applications**
- Lung Adenocarcinoma

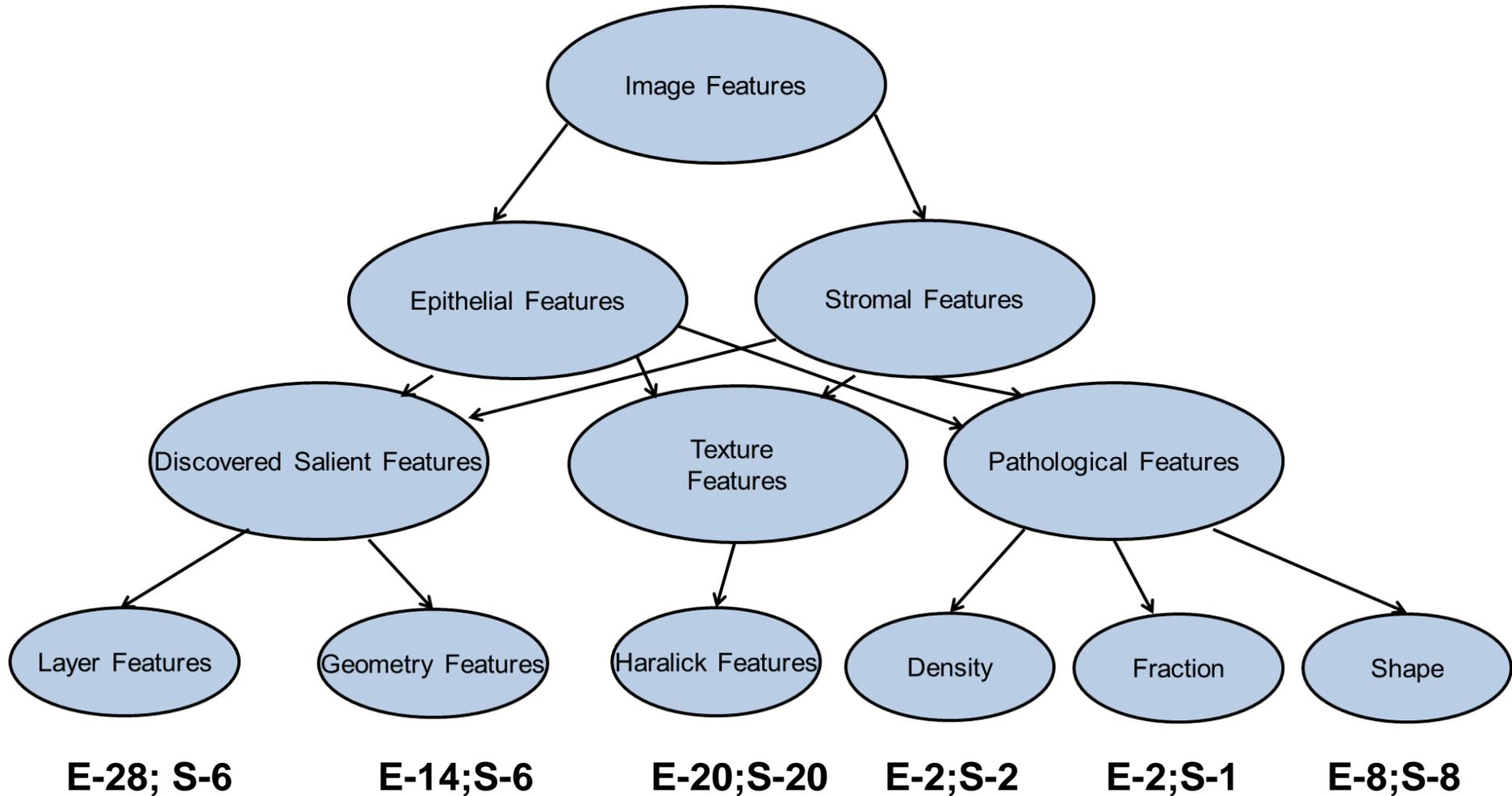


- Applications**
- Breast Cancer Patient Stratification
- A two-step algorithm
 - Molecular regularized algorithm

Pipeline Overview



Feature Extraction



E:epithelial features;S:stromal features



Image Analysis

Whole-slide image

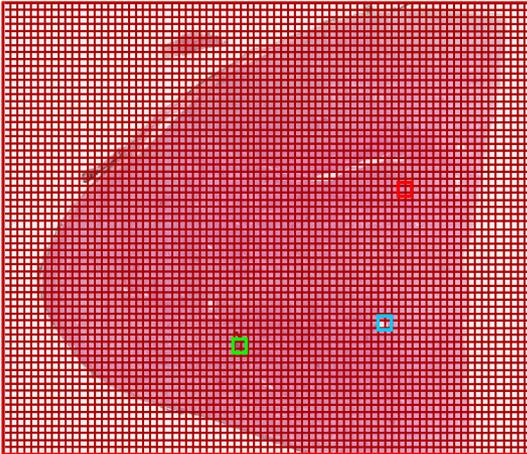
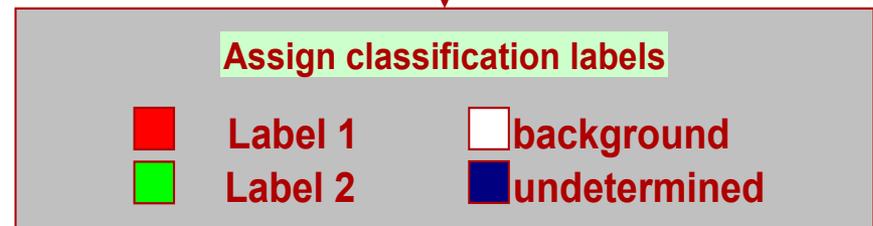
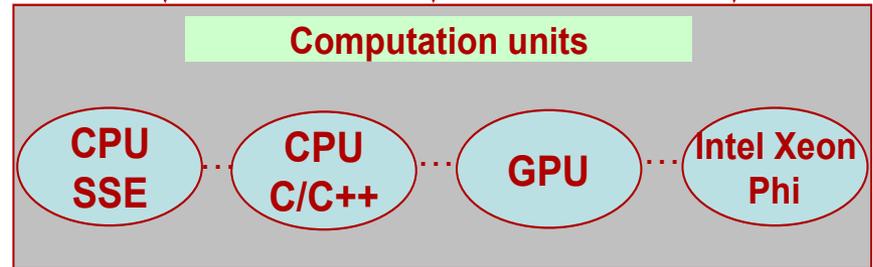
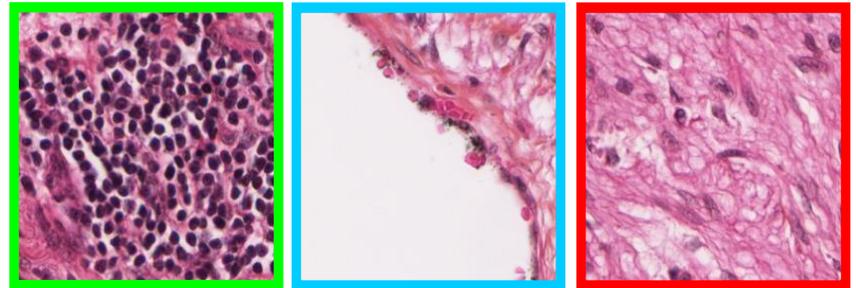
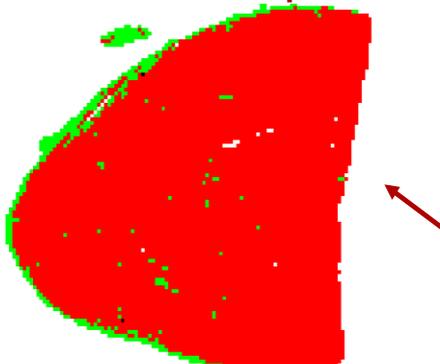


Image tiles (40X magnification)



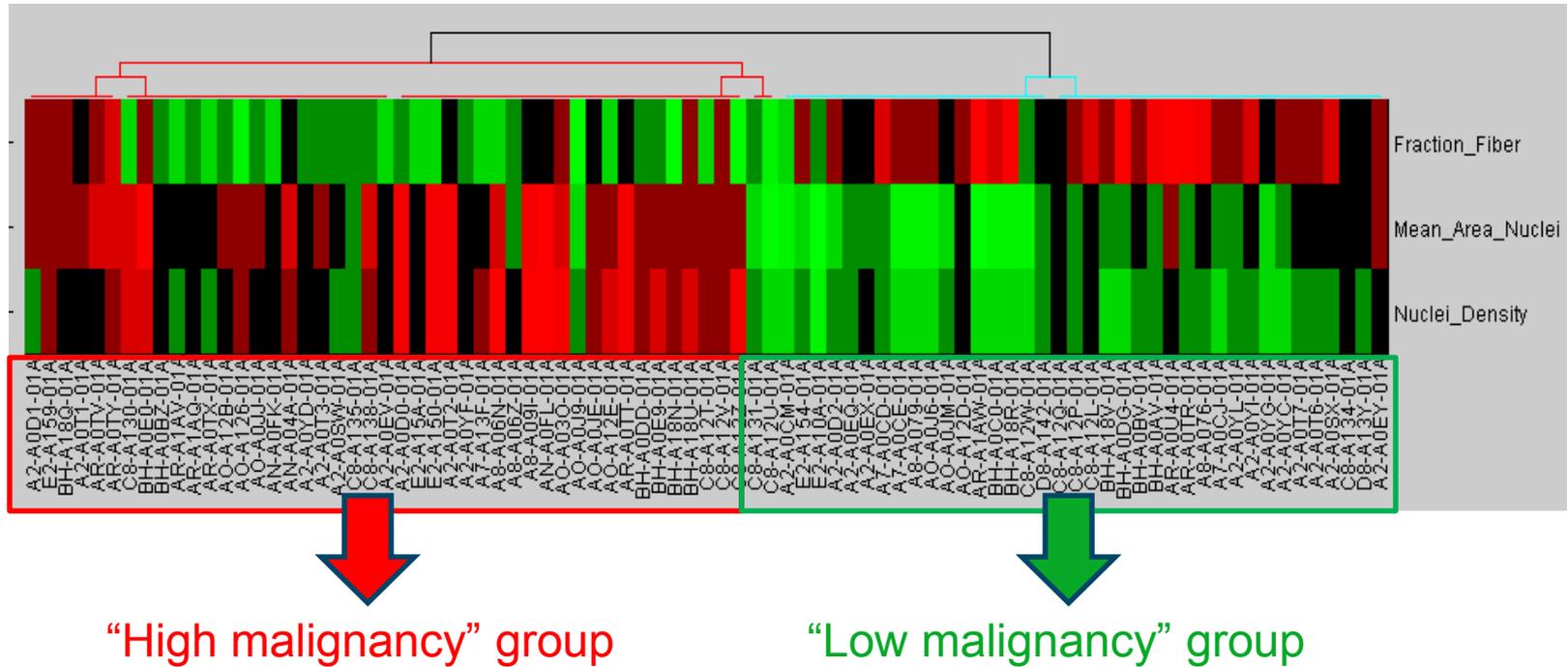
Classification map



Catalyurek



Patient Stratification on Morphology



Hierarchical clustering of breast cancer patients based on imaging features.

Associated Protein Co-Expression Network

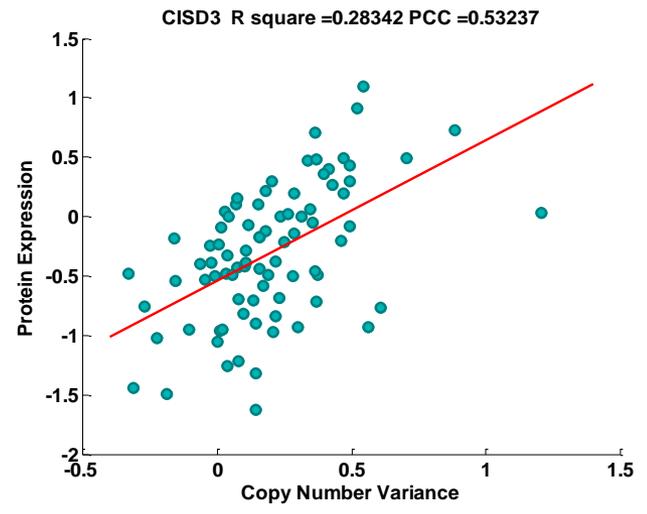
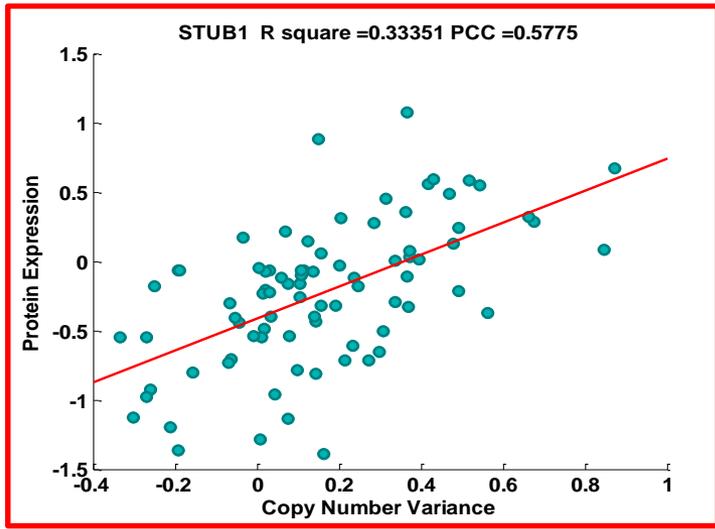
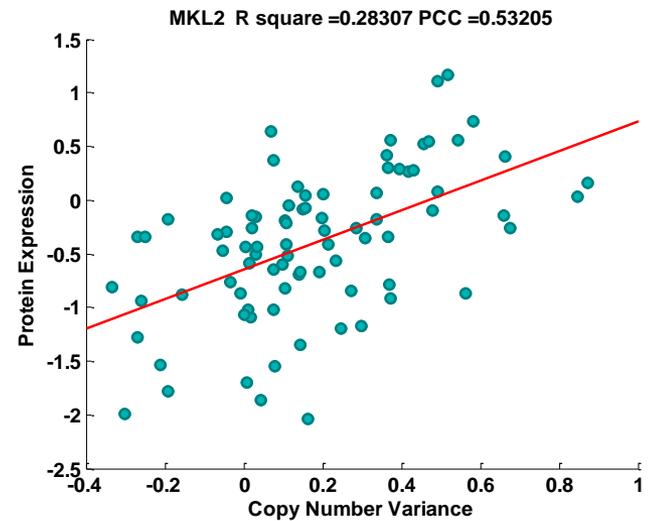
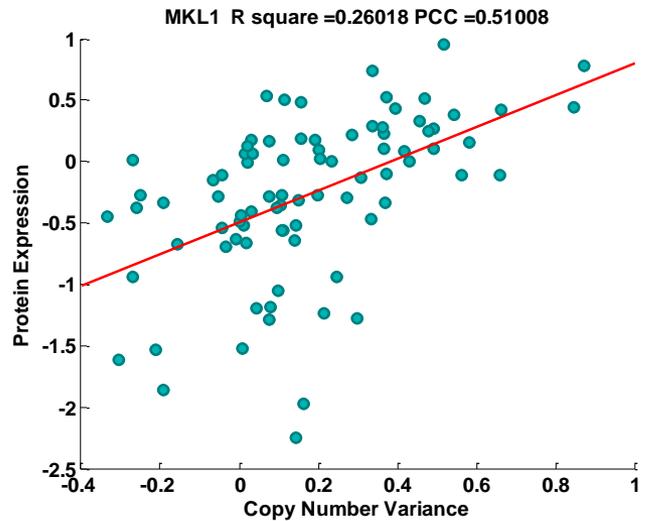
- Identified 124 protein co-expression networks
- 4 protein networks differentially expressed

Protein Networks	Proteins in the Networks	Cytoband	pValue	Enriched Cytobands	Top GO Molecular Functions (pValue)	Note
1	MYOCD	17p11.2	3.880E-2	16p13 P – value = 2.170E-5	smooth muscle cell differentiation(8.184 E-7); muscle cell differentiation(1.889 E-4); regulation of transforming growth factor beta receptor signaling pathway.	Enriched in 16q13
	MKL1	22q13	5.986E-3			
	MKL2	16p13.12	2.306E-3			
	FLYWCH2	16p13.3	2.170E-5			
	CISD3	17q12	1.439E-1			
	HN1L	16p13.3	2.170E-5			
	STUB1	16p13.3	2.170E-5			
	SUCLG1	2p11.2	1.536E-2			

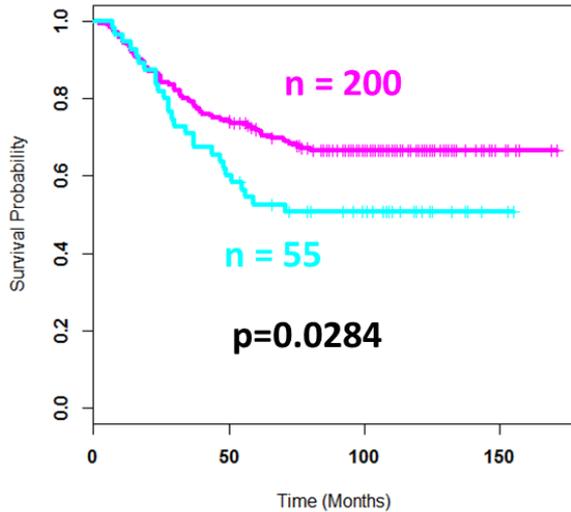


DNA Copy Number Variance

Protein Networks	Proteins in the Networks
1	MYOCD
	MKL1
	MKL2
	FLYWCH2
	CISD3
	HN1L
	STUB1
	SUCLG1

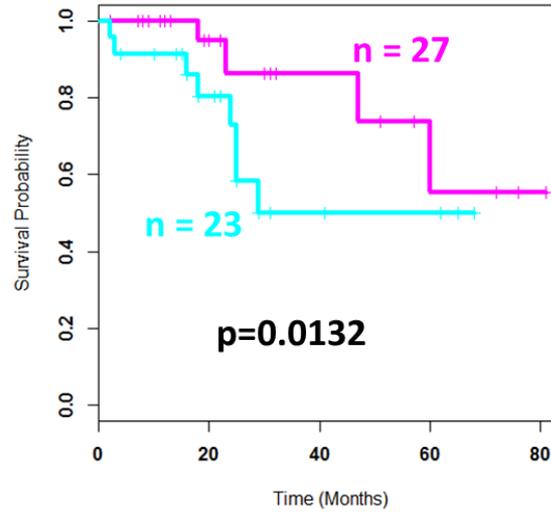


Prognosis Validation



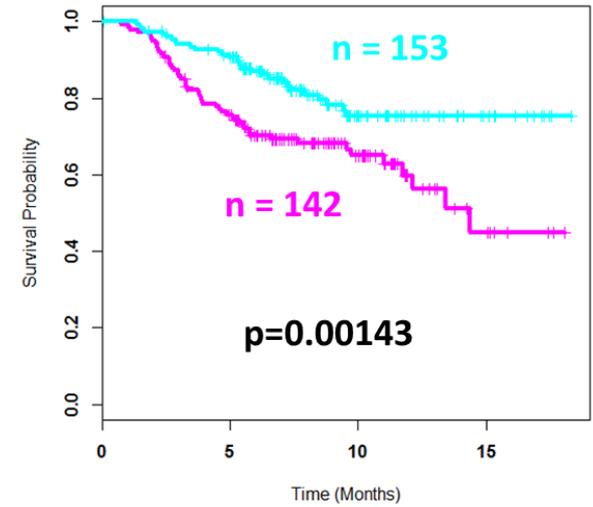
A

Wang



B

Perou



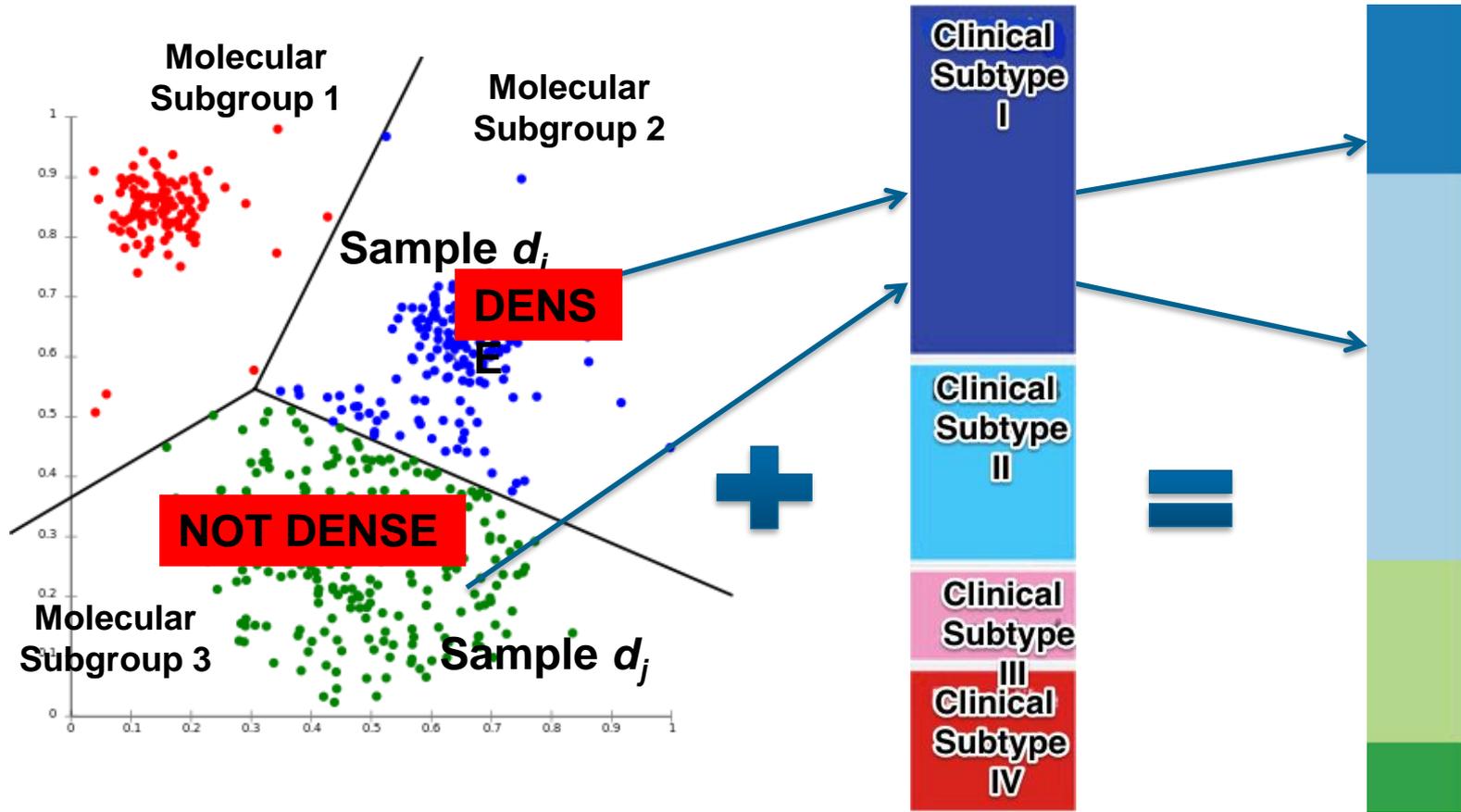
C

NKI

Kaplan–Meier survival curves of prognostic model on multiple public breast cancer datasets. From left to right: Wang, Perou and NKI data respectively.



Concept



Wang et al, Methods, 2013



Wexner Medical Center

The Algorithm

Algorithm 1: Molecular Regularized Consensus Patient Stratification

Data: Similarity Matrix \tilde{S} , Molecular Density Weight Matrix W , the number of clusters in final consensus k , $MaxIter$, precision ϵ

Result: Cluster indicator matrix U .

initialize $\tilde{U}^{(1)} > 0$, $t = 1$, $\Delta = +\infty$;

while $t < MaxIter$ and $\Delta > \epsilon$ **do**

$$\text{Update } \tilde{U}_{ij}^{(t+1)} \leftarrow \tilde{U}_{ij}^{(t)} \sqrt{\frac{[(W \circ S)\tilde{U}D]_{ij}}{[(W \circ \tilde{U}D\tilde{U}^T)\tilde{U}D]_{ij}}};$$

$$\text{Update } D_{ij}^{(t+1)} \leftarrow D_{ij}^{(t)} \sqrt{\frac{[\tilde{U}^T(S \circ W)\tilde{U}]_{ij}}{[\tilde{U}^T(\tilde{U}D\tilde{U}^T \circ W)\tilde{U}]_{ij}}};$$

$$\text{Compute } \Delta = \|\tilde{S} - W \circ (\tilde{U}D\tilde{U}^T)\|_F^2;$$

$t = t + 1$;

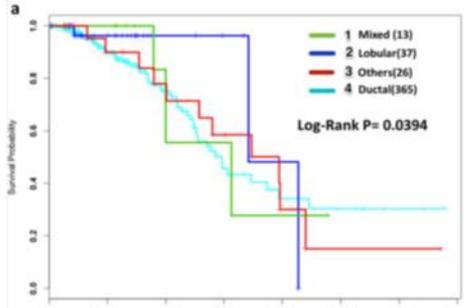
end

Discretize \tilde{U} to binary membership matrix.

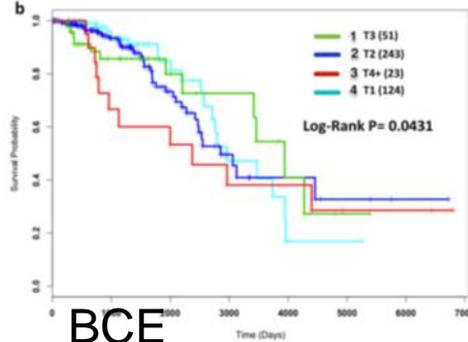
Algorithm 1: Molecular Regularized Consensus Patient Stratification

Prognosis

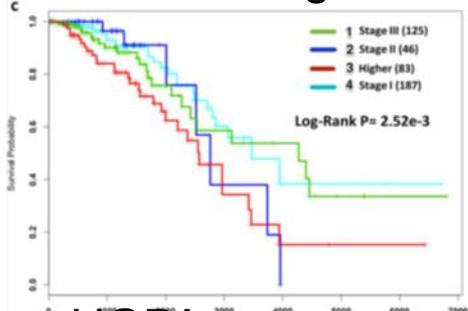
Histology Type



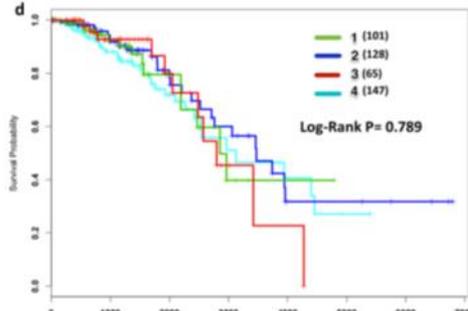
Tumor Grade



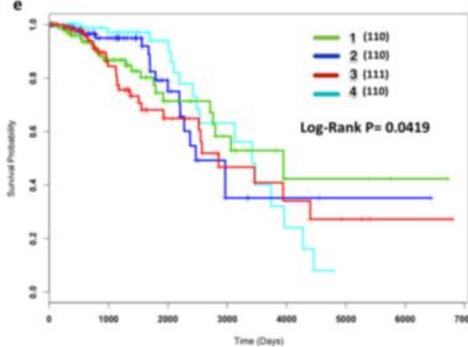
Disease Stage



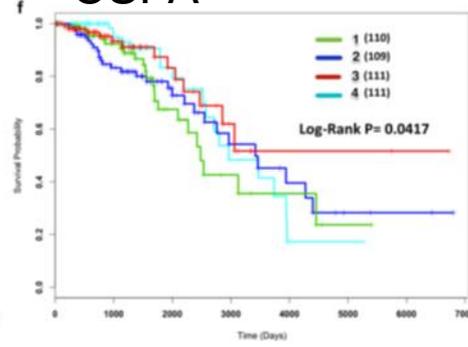
BCE



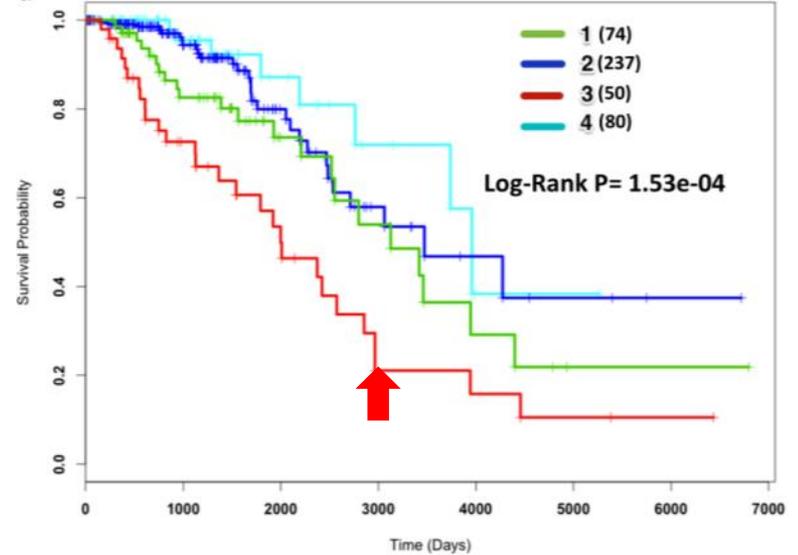
HGPA



CSPA



RCP



Wang C, Machiraju R, Huang K, *Methods*, 2014



Wexner Medical Center

Visualization

iGPSe

Home

Document

About

Contact

Roadmap

> Load Data set

> Select gene

> Clustering

> Interactive analysis

Selected mRNA list

ERBB2 CCNB1 MELK
UBE2T TYMS NDC80
BIRC5 MYBL2 RRM2
NUF2 CEP55 UBE2C
CDC6 KIF2C SFRP1
CENPF ACTR3B KRT14
PTTG1 MYC ESR1
EXO1 EGFR SLC39A6
ORC6L KRT5 BAG1
ANLN PHGDH MAPT
CCNE1 CDH3 PGR
CDC20 MIA CXXC5
MKI67 KRT17 MLPH
FOXA1 FOXC1 BCL2
BLVRA MDM2 NAT1
MMP11 GRB7 TMEM45B
GPR160 FGFR4

Selected miRNA list

hsa-mir-130a hsa-mir-23a
hsa-mir-222 hsa-mir-24-1
hsa-mir-29a hsa-mir-24-2
hsa-mir-30a hsa-mir-27a
hsa-mir-100

mRNA Clustering

Changing parameters&clustering algorithms

Metric:

Euclidean

Clustering:

Kmeans

Number of clusters(k):

4

Run

Threshold:

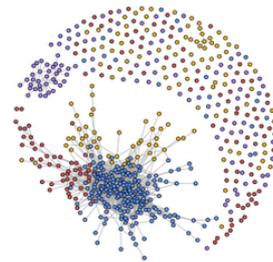
8.0

Silhouette plot

Heatmap

Proximity Graph

Interactive



miRNA Clustering

Changing parameters&clustering algorithms

Metric:

Euclidean

Clustering:

Kmeans

Number of clusters(k):

3

Run

Threshold:

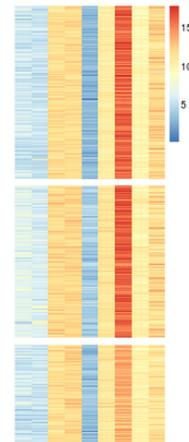
0.999

Silhouette plot

Heatmap

Proximity Graph

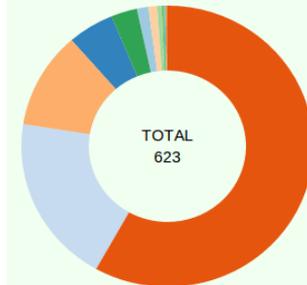
Zoom In



Clinical INFO

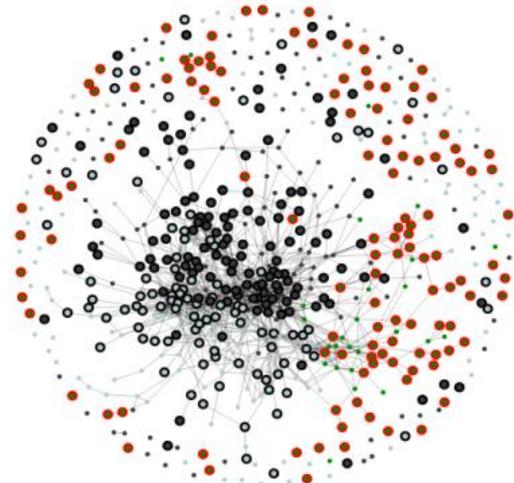
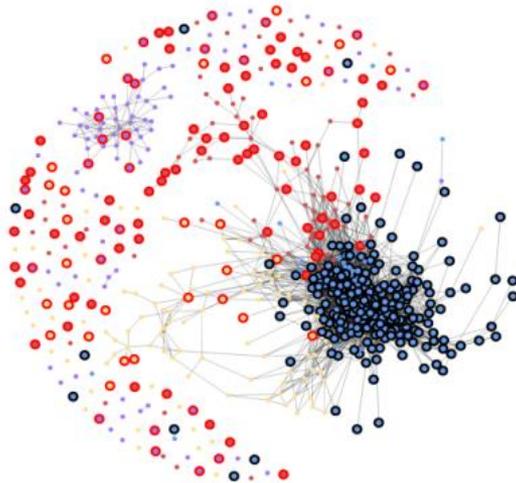
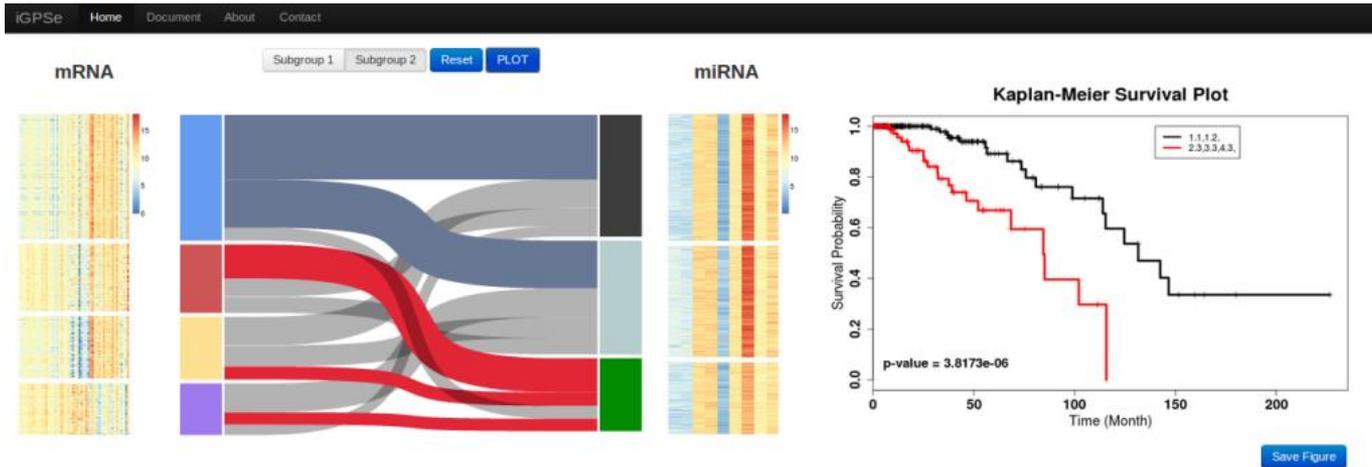
Age

Stages

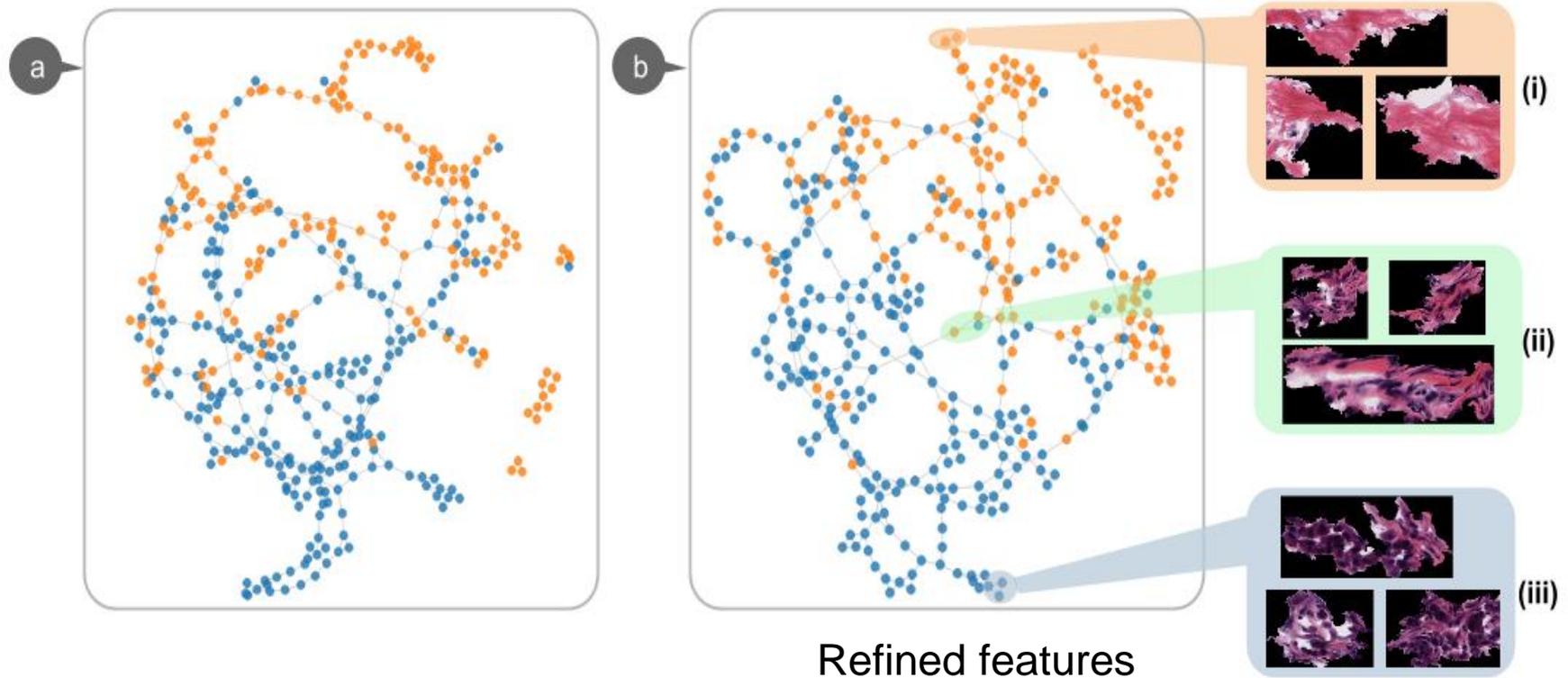


Interactive Visualization >>

Interactive Patient Stratification



Graphie – Visual Analytics of Imaging Features



Ding et al. *BMC Bioinformatics* 2015, **16**(Suppl 11):S10
<http://www.biomedcentral.com/1471-2105/16/S11/S10>



RESEARCH

Open Access

GRAPHIE: graph based histology image explorer

Hao Ding¹, Chao Wang³, Kun Huang^{2*}, Raghu Machiraju^{1*}

From 5th Symposium on Biological Data Visualization
Dublin, Ireland. 10-11 July 2015

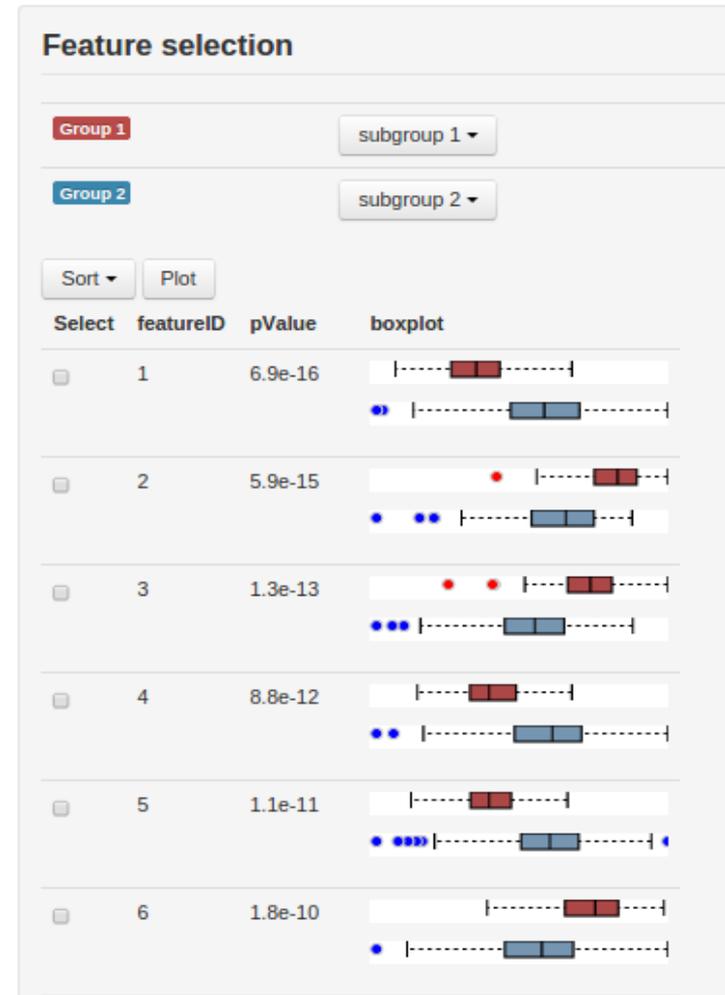


Wexner Medical Center

Graphie –Feature Selection

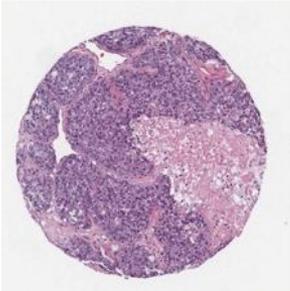
Examine feature distinctiveness for groups of images.

- Student's t-test
- Boxplots
- Re-generate with selected feature subset



Combining Graphie, iGPSE, & SUMO

Histology images



Genome



Epigenome



Proteome



Subtype Analysis

Integrative Biomarkers

Molecular behaviours

...



<http://osumo.org/>



Wexner Medical Center

PIPELINE

DATA SETS



SELECT DATA SET

Breast Invasive Carcinoma (TCGA) 441 samples

DATA SET SUMMARY ?

Title: Breast Invasive Carcinoma (TCGA)

Summary

Data Type(s): mRNA miRNA DNAmethylation

Sample size: 441

FEATURE SET ?

mRNA

miRNA

DNA methylation

User-defined List

User-defined List

User-defined List

Enter Gene Symbols:

Enter Gene Symbols:

Enter Gene Symbols:

State of (O)SUMO

iGPSe: Interactive Genomics Patient Stratification explorer

Upload Dataset

Upload the Gene expression profile

No file chosen

Upload the MicroRNA expression profile

No file chosen

Upload the Clinical profile

No file chosen

If you want a sample .csv file to upload, you can first download the sample [mRNA.sample.csv](#), [miRNA.sample.csv](#) and [time.cancer.csv](#) files, and then uploading them.

Data info [Clustering Analysis](#) [parallelset](#)

Gene expression

null

microRNA

null

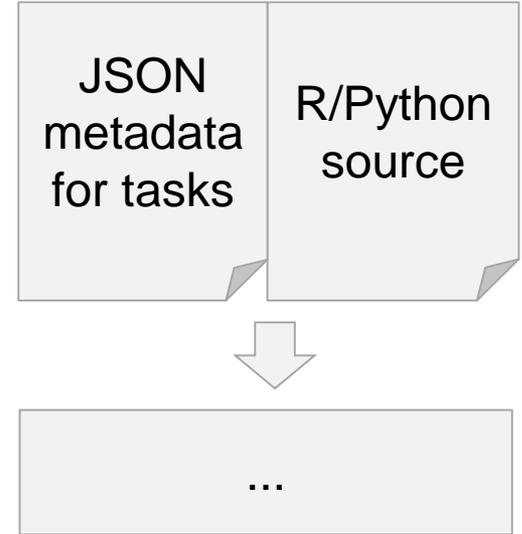
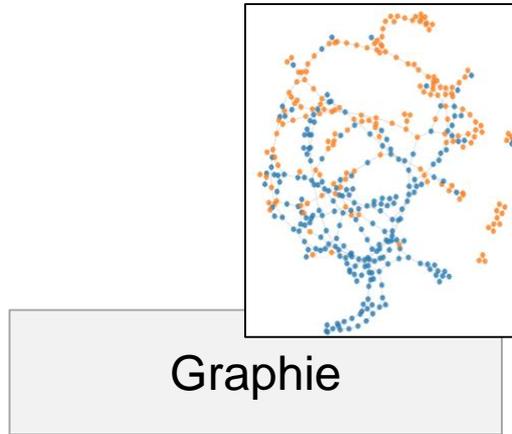
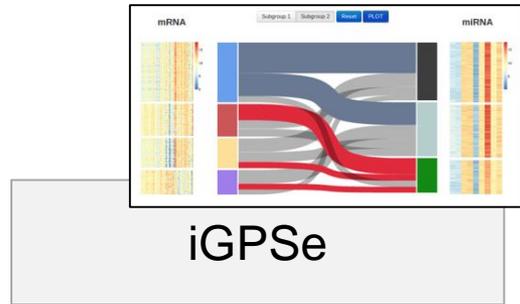
Clinical

null



SUMO Architecture

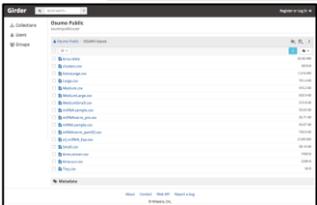
Adding workflows



SUMO (workflow management, UI)



Girder (authentication, data management)



Girder Worker (analysis execution)

```
blur_image = {
  'inputs': [
    {'name': 'blur_input', 'type': 'image', 'format': 'pil'},
    {'name': 'blur_radius', 'type': 'number', 'format': 'number'}
  ],
  'outputs': [{'name': 'blur_output', 'type': 'image', 'format': 'pil'}],
  'script': '''
from PIL import ImageFilter
blur_output = blur_input.filter(ImageFilter.GaussianBlur(blur_radius))
...
'''
}
```



Input / output descriptions

```
name: "Silhouette Plot"
description: "Show silhouette plot"
fallbackUser: "osunopublicuser"

task:
  name: silhouette
  mode: r

script: >
  @include(silhouette.r)

inputs:
  - name: sil_input_path
    type: string
    format: text
  - name: silhouette_clusters
    type: number
    format: number

outputs:
  - name: dataplot1
    target: filepath
    type: image
    format: png
  - name: dataplot2
    target: filepath
    type: image
    format: png

inputs:
  - key: sil_input_path
    name: "MicroRNA Expression Profile"
    description: "MicroRNA data."
    type: file
    subtype: rdata
    preferredNames: "'sil_*.csv'"
    notes: "This must be a CSV file with one column per subject, one header row, and one data row per micro RNA"
  - key: silhouette_clusters
    name: "Number of Clusters (k)"
    description: "The number (k) of clusters used in calculations"
    type: integer
    default: 2
    format: number
    notes: "Clustering is performed via k-means."

parameters:
  - key: targetFolderId
    name: "Target Folder"
    type: folder
    input: false

outputs:
  - key: dataplot1
    show: image
    parentType: "folder"
    parent: "parameter:targetFolderId"
    name: "dataplot1.png"
    dataType: "image"
    dataFormat: "png"
    displayName: "Silhouette plot"
  - key: dataplot2
    show: image
    parentType: "folder"
    parent: "parameter:targetFolderId"
    name: "dataplot2.png"
    dataType: "image"
    dataFormat: "png"
    displayName: "Nearest Neighbor Graph"
```

R script

```
## Read data
silData = read.csv(sil_input_path)

# clustering
color = palette()
k = silhouette_clusters
sil = silhouette(silData, k)

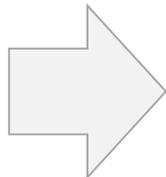
localDir = Sys.getenv("tempdir", unset=tempdir())
save(sil, file=paste0(localDir, "silhouette.png"))
save(sil, file=paste0(localDir, "silhouette.csv"))

plot(sil, col=rep("black", nrow(sil)), border=NA) # with cluster-wise coloring
dev.off()

# plot the graph
res = resnet()
resnet(sil)
M = as.matrix(dist(resnet(sil)))
g = graphLM(M, w=1)
g = as.unweighted()
visNetwork = colorByClustering()

localDir = Sys.getenv("tempdir", unset=tempdir())
save(g, file=paste0(localDir, "nearestNeighbor.png"))
save(g, file=paste0(localDir, "nearestNeighbor.csv"))

plot(g, vertices.size=1, layout=layout_kamada_kamada, vertices.label=NA)
dev.off()
```



Basic Workflow - Silhouette Plot

Task: Silhouette Plot

Show silhouette plot

MicroRNA Expression Profile
This must be a CSV file with one column per subject, one header row, and one data row per micro RNA
sil_miRNA_Exp.csv

Number of Clusters (k)
Clustering is performed via k-means.
2

Process

SUCCESS - 2016-06-13T19:58:24.338000+00:00

Silhouette plot

Silhouette plot of pam(x = t(dmiRNA), k = k)
n = 623

2 clusters C_j
j: n_j | s_{max} C_j

1: 538 | 0.16

2: 85 | 0.08

Average silhouette width: 0.15

Nearest Neighbor Graph

Inputs

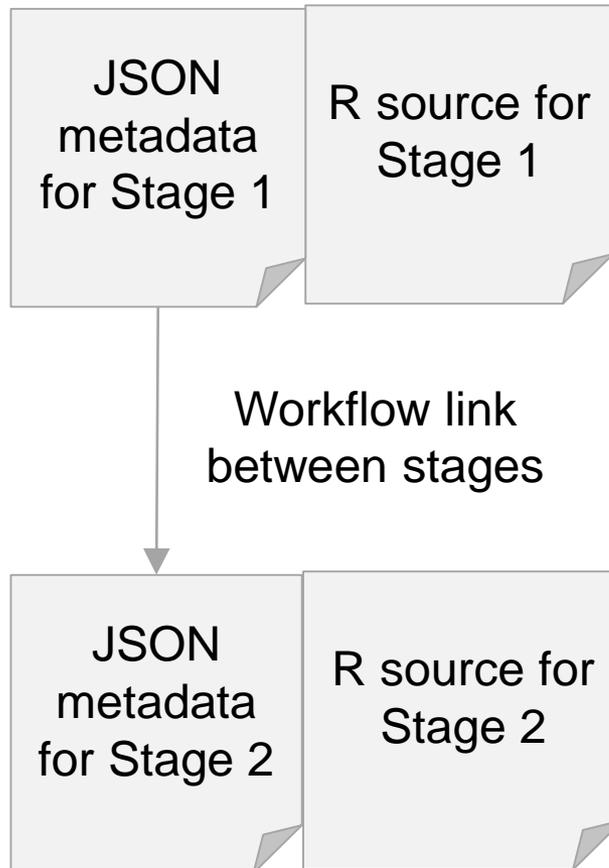
Any Girder-uploaded dataset, access controlled

Outputs

May be static R charts or interactive D3 plots

Outputs are stored in user's Girder space

2-Stage Workflow - iGPSe



Task: iGPSe

iGPSe: Interactive Genomics Patient Stratification explorer
Show a survival plot based on clustering of different data sets.

Gene Expression Profile
This must be a CSV file with one column per subject, one header row, and one data row per gene
mRNAAnorm_pam50.csv
mRNA Number of Clusters (k)
Clustering is performed via k-means.
5

MicroRNA Expression Profile
This must be a CSV file with one column per subject, one header row, and one data row per micro RNA
miRNAAnorm_pre.csv
miRNA Number of Clusters (k)
Clustering is performed via k-means.
5

Clinical Profile
This must be a CSV file with a header row and one row per subject, and columns containing the row description, survival duration, and an indicator field.
time_sur.csv

Process
SUCCESS - 2016-06-13T19:52:09.084000-00:00

mRNA Heatmap
The major groups show how the data was clustered. The columns show different miRNA attributes, and the rows represent each subject.

miRNA Heatmap
The major groups show how the data was clustered. The columns show different miRNA attributes, and the rows represent each subject.

Cluster Selection
Select two groups to compare. The left side shows the clusters generated from mRNA data, and the right side from miRNA data. If a link between the sides is selected, those subjects in the intersection between those two clusters are used. After selecting at least one cluster or link for each of the two groups, select PLOT to generate a comparison survival plot.

Group 1 Group 2 Reset PLOT

SUCCESS - 2016-06-13T19:52:26.598000-00:00

Survival Comparison
This compares the two groups selected from the clustered data.

KM Survival Plot

About Contact Web API Report a bug
© The Ohio State University 2016
All Rights Reserved

Stage 1 Inputs

Genetic and clinical datasets

Stage 1 Outputs

Two static R cluster heatmaps

One interactive D3 set selection that serves as Stage 2 input

Stage 2 Output

Survivability plot for selected sets

SUMO

Initial application available at <http://osumo.org>

Plan for more machine learning algorithms (consensus learning, multiview learning) for data integration on SUMO

Pan-cancer study as driven application

We welcome

- ✧ Beta users
- ✧ Use cases and workflows to deploy
- ✧ Open-source

contributions <https://github.com/osumo/osumo>



Demonstration

- <https://www.youtube.com/watch?v=d96h4EnwRxY>



More Comprehensive Workflow Proposed

