Informatics links between histological features and genetics in cancer

Kun Huang, PhD

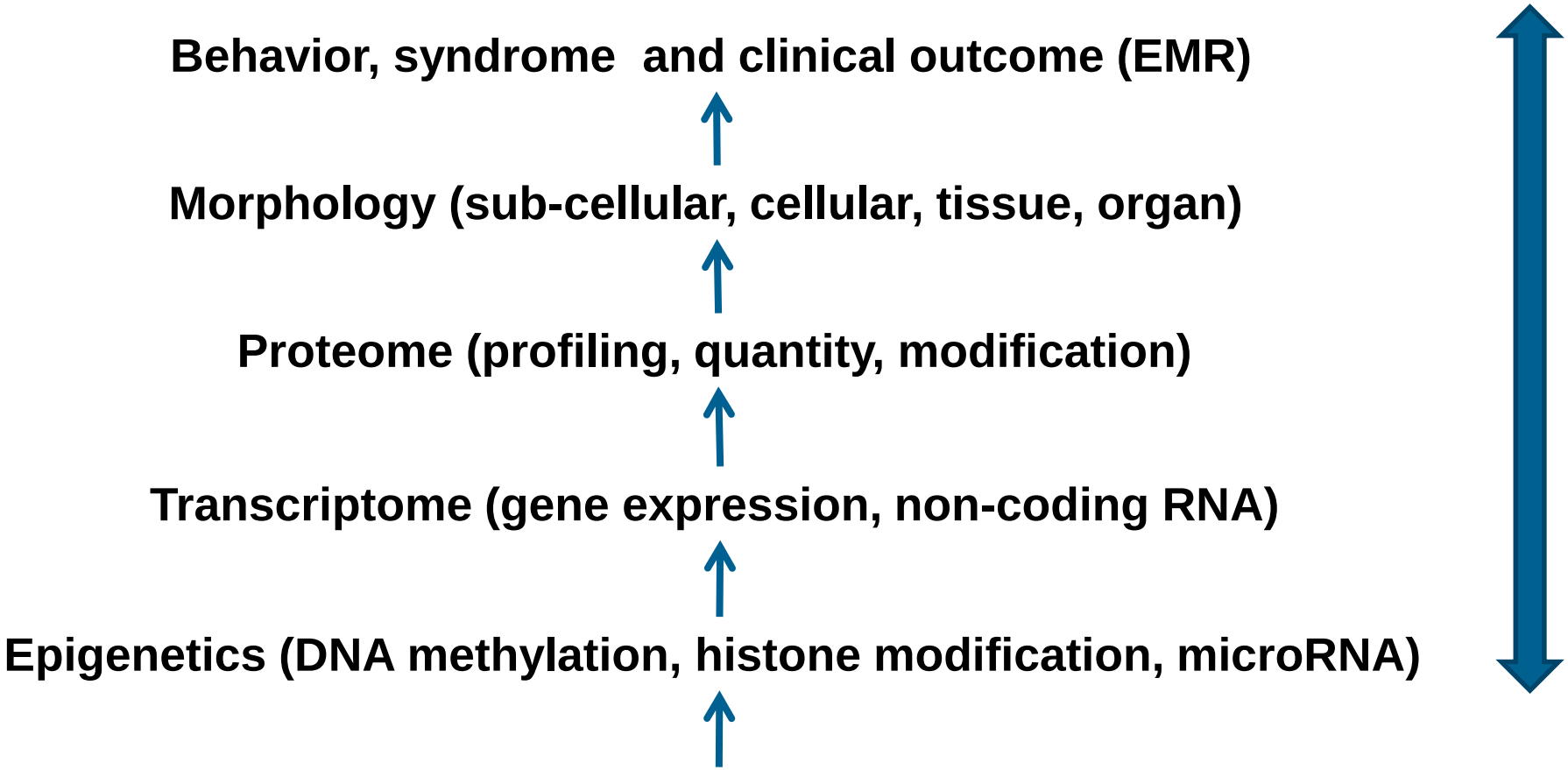
Department of Biomedical Informatics
The Ohio State University Wexner Medical Center

June 13, 2016

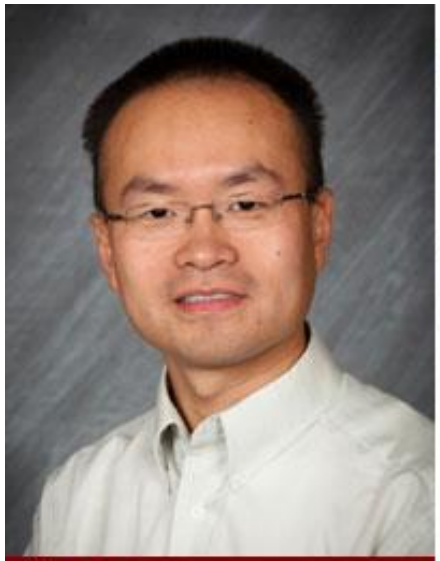
Wexner Medical Center

Data Integration

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



Team



Huang

Machiraju

Jeff Baumes, Kitware

Jie Zhang

David Manthey, Kitware

dical Center

Leveraging

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

Catalyurek

HPC

Mallick

Proteomics

NATIONAL CANCER INSTITUTE

CLINICAL PROTEOMIC TUMOR ANALYSIS CONSORTIUM

OHIO STATE Wexner Medical Center

4

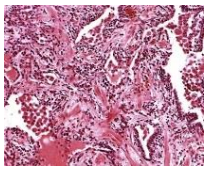
Tasks

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***.

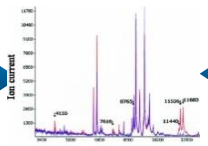
Aim 2 – Develop an integrative and expandable open source platform for *managing, analyzing, and integrating multiple data types* in integrative genomics for cancer with ***visual analytic capabilities*** for cancer biomarker discovery.

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.

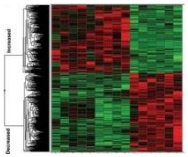
From Correlation to Integration



Pathology



Proteome

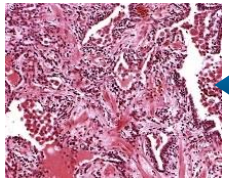


Transcriptome

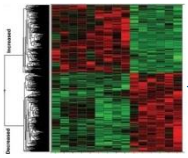
Correlative Study

- Triple Negative Breast Cancer
- Breast Cancer Proteomics

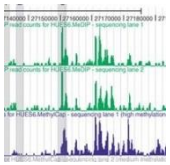
Applications



Pathology



Transcriptome

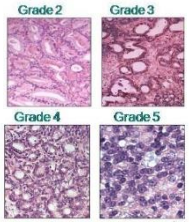


Epigenome

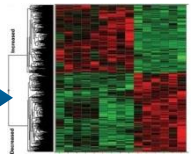
Model-based Study

- Lung Adenocarcinoma

Applications



Clinical Attributes



Transcriptome

Integrative Clustering

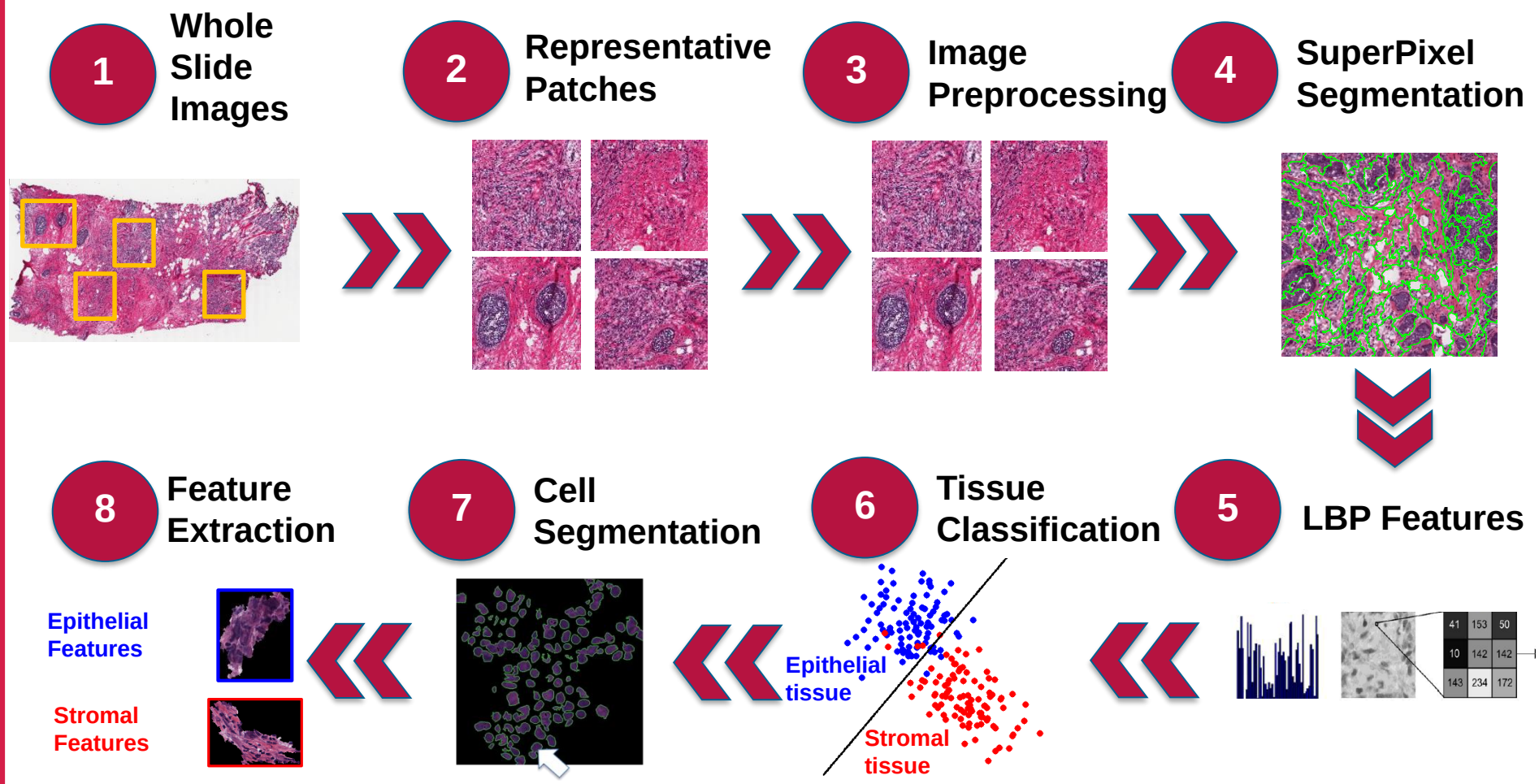
Breast Cancer Patient Stratification

- A two-step algorithm
- Molecular regularized algorithm

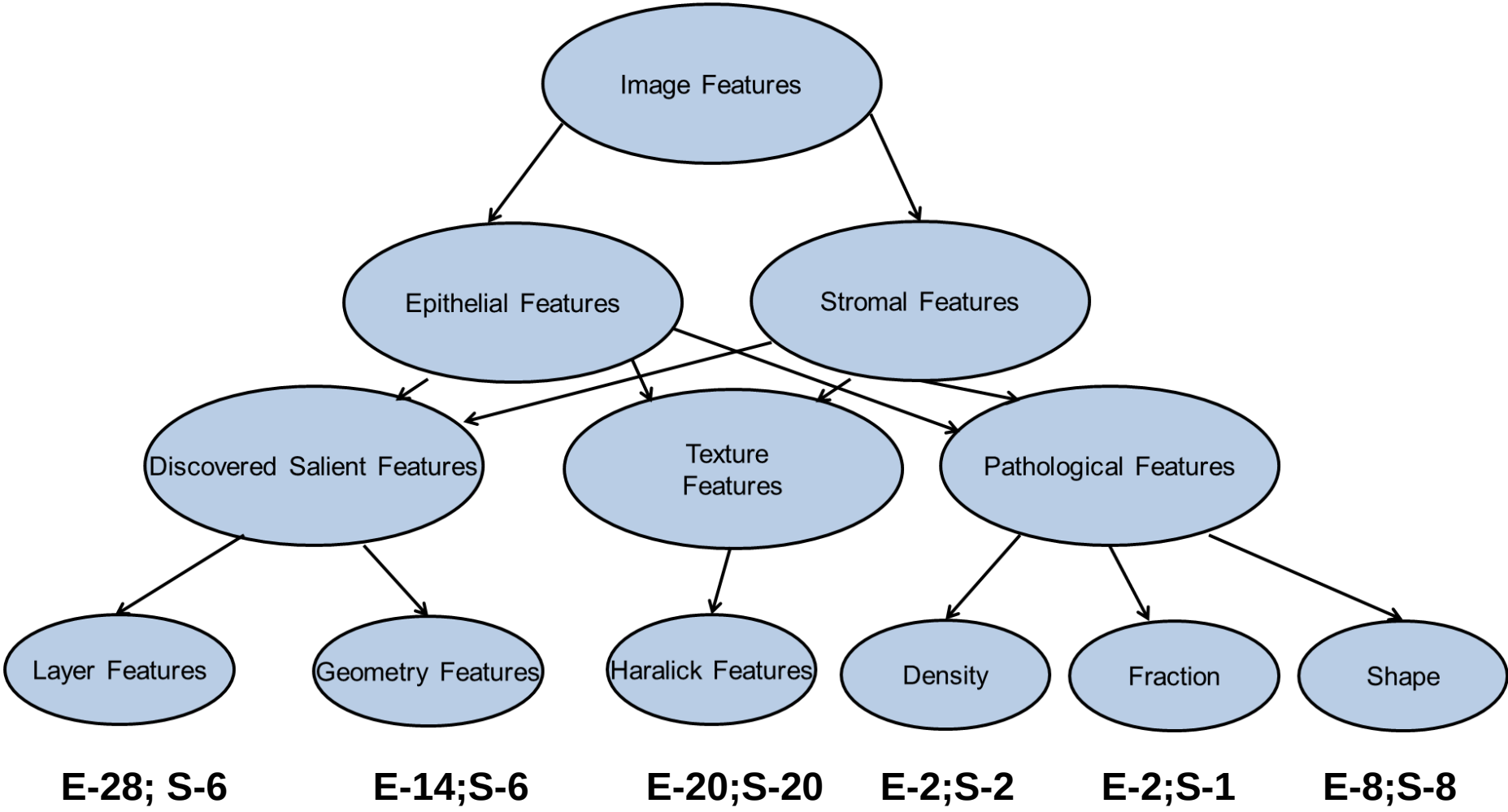
Applications

Wexner Medical Center

Pipeline Overview



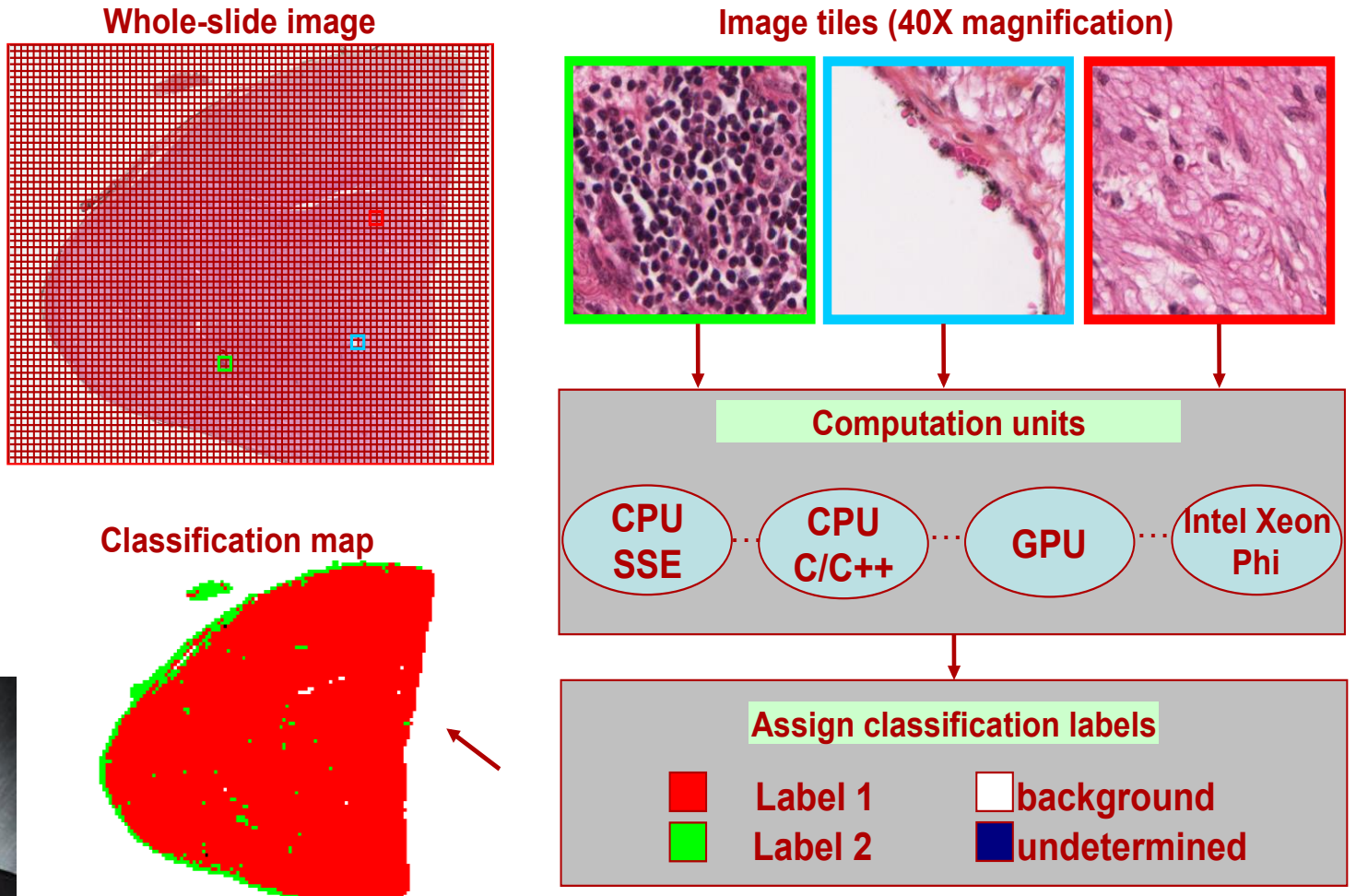
Feature Extraction



E:epithelial features;S:stromal features

Wexner Medical Center

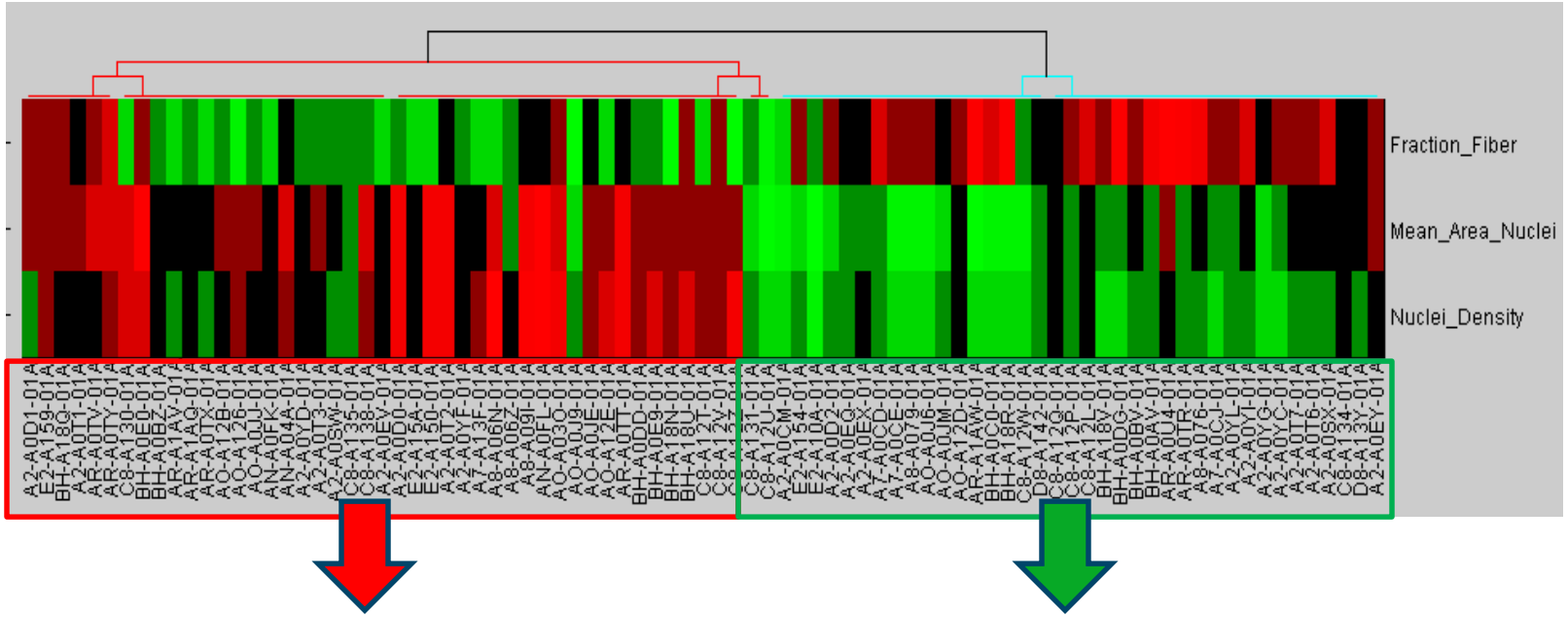
Image Analysis



Catalyurek

Wexner Medical Center

Patient Stratification on Morphology



Hierarchical clustering of breast cancer patients based on imaging features.

Wexner Medical Center

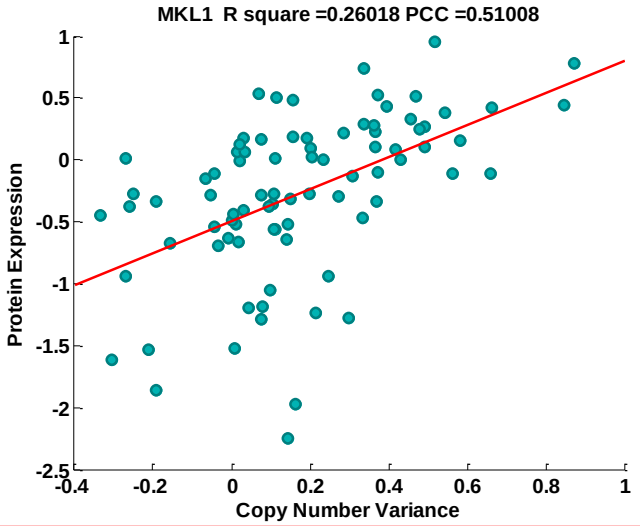
Associated Protein Co-Expression Network

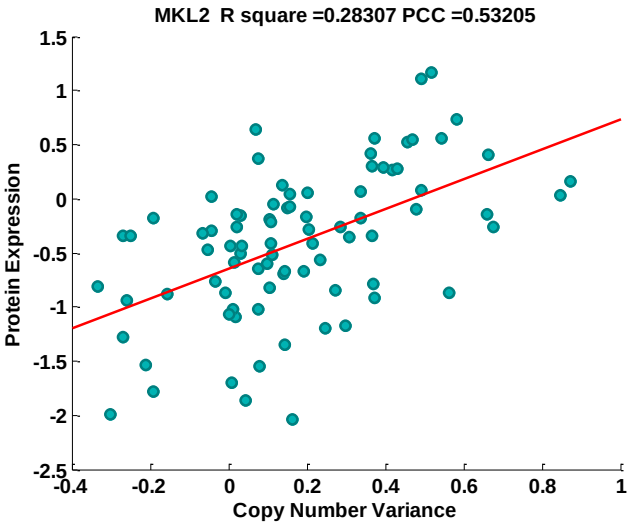
- Identified 124 protein co-expression networks
- 4 protein networks differentially expressed

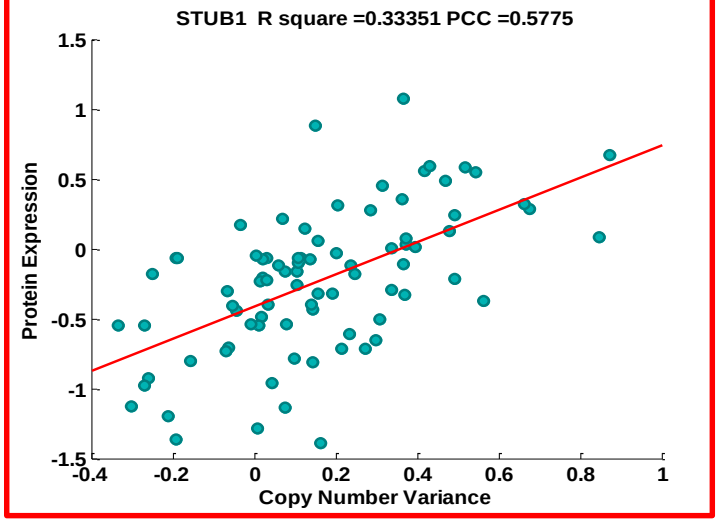
Protein Networks	Proteins in the Networks	Cytoband	pValue	Enriched Cytobamds	Top GO Molecular Functions (pValue)	Note
1	MYOCD	17p11.2	3.880E-2	16p13 P – value = 2.170E-5	smooth muscle cell differentiation(8.184 E-7); muscle cell differentiation(1.889 E-4); regulation of transforming growth factor beta 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			

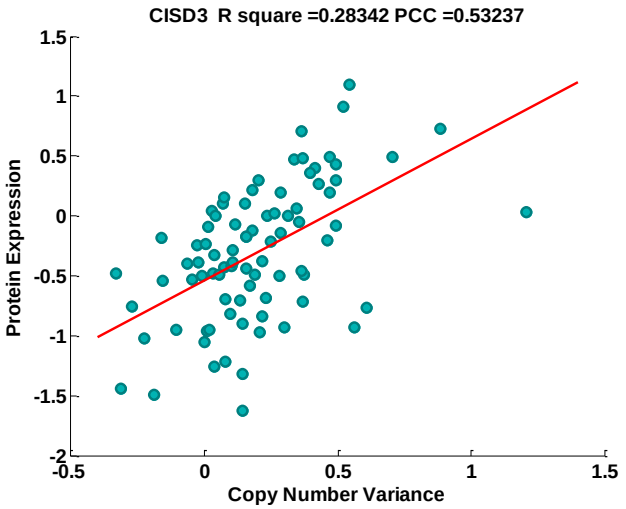
DNA Copy Number Variance

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



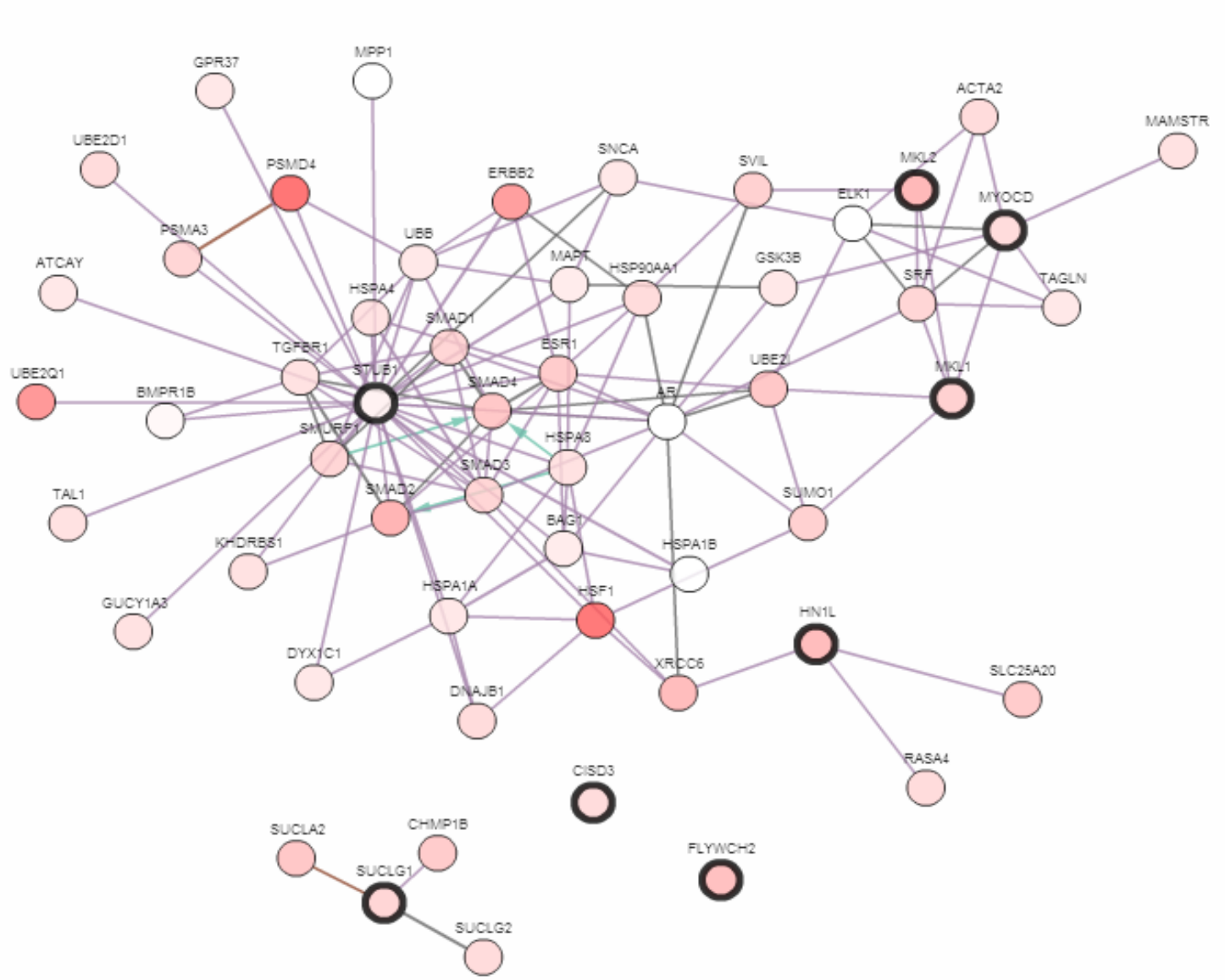




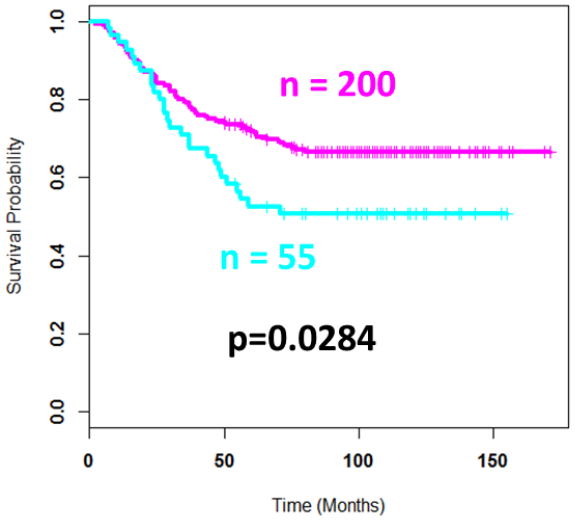


Wexner Medical Center

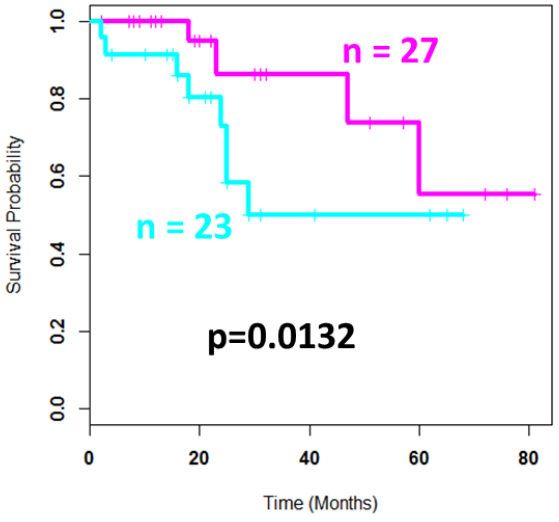
STUB1 gene as a hub gene



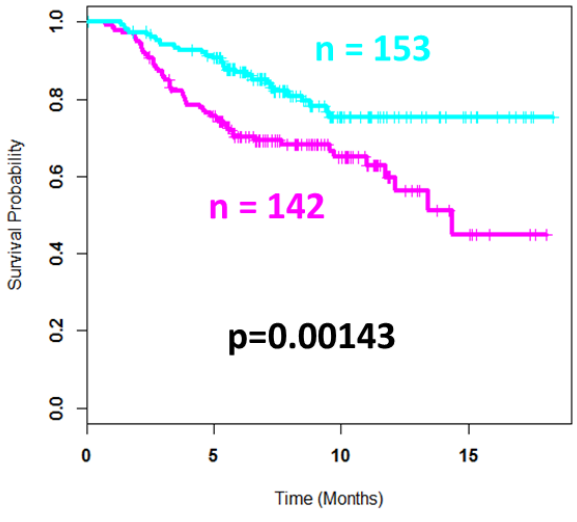
Prognosis Validation



A



B



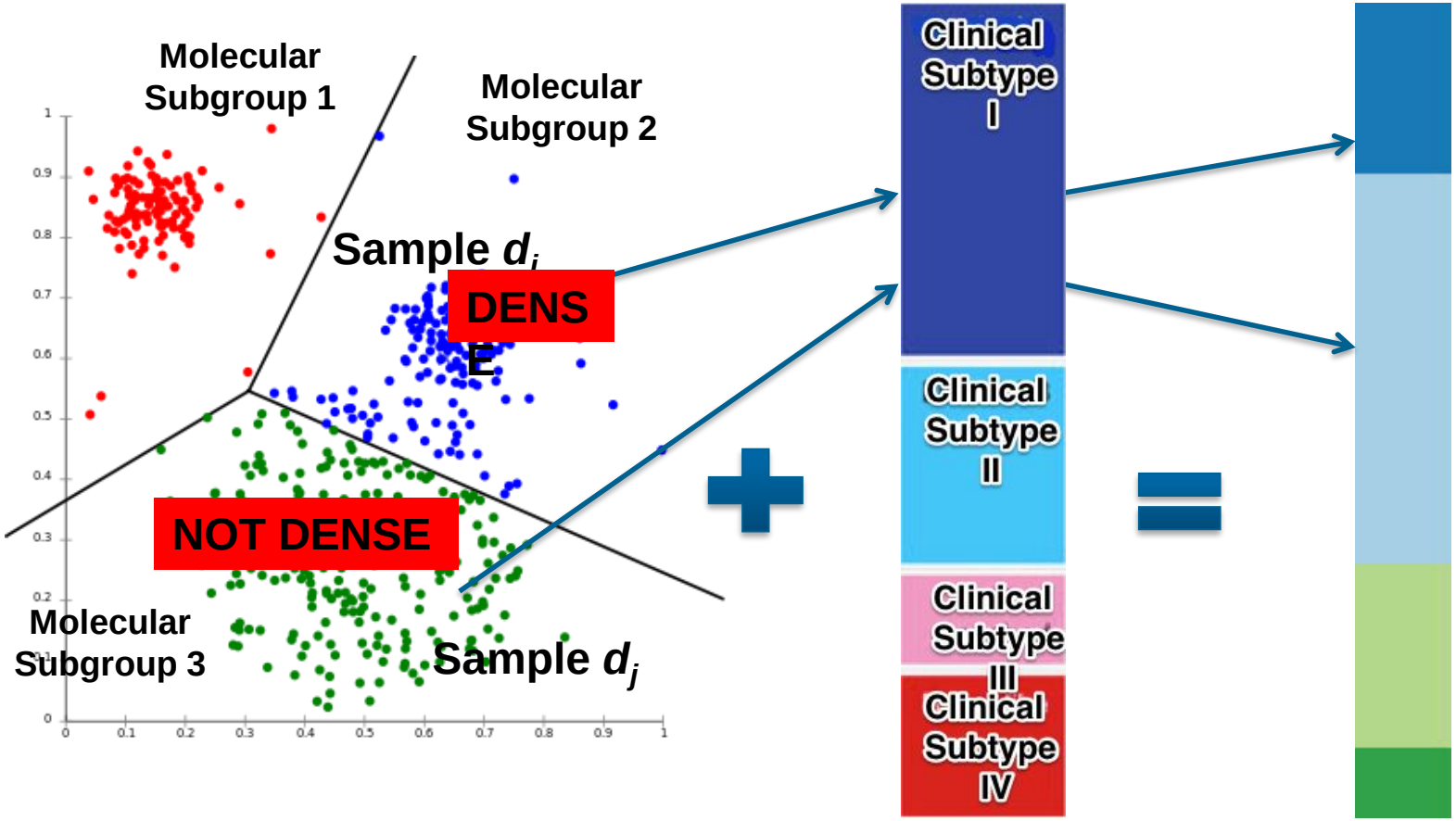
C

Wang Perou NKI

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

Wexner Medical Center

Concept



Wang et al, Methods, 2013

The Algorithm

Algorithm 1: Molecular Regularized Consensus Patient Stratification

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

Result: Cluster indicator matrix U.

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

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

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

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

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

$t = t + 1$;

end

Discretize $\widetilde{U}$ to binary membership matrix.

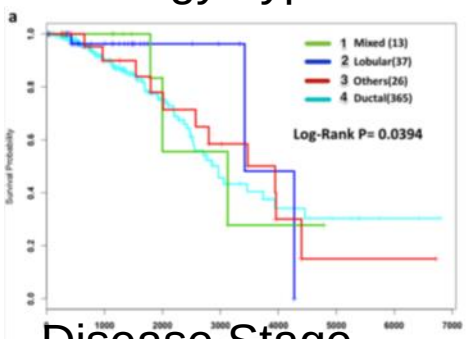
Algorithm 1: Molecular Regularized Consensus Patient Stratification

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

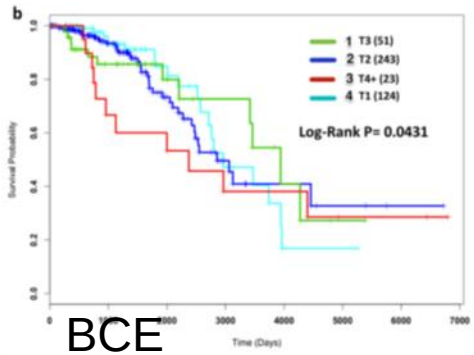
Wexner Medical Center

Prognosis

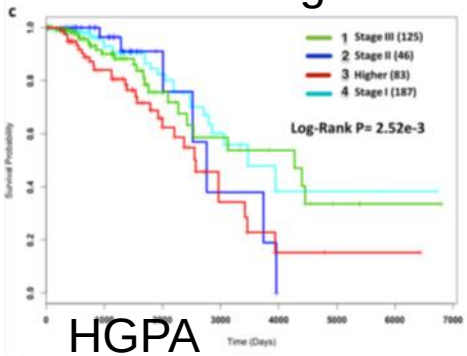
Histology Type



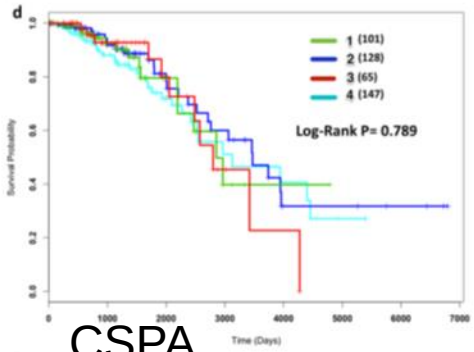
Tumor Grade



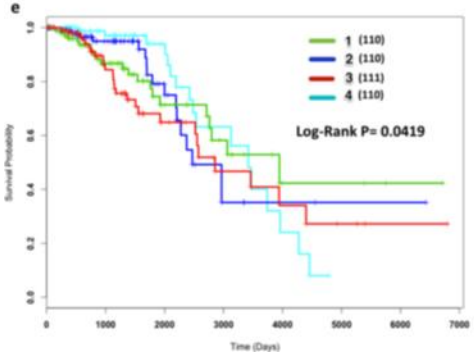
Disease Stage



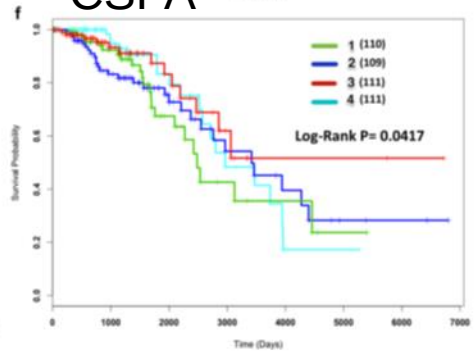
BCE



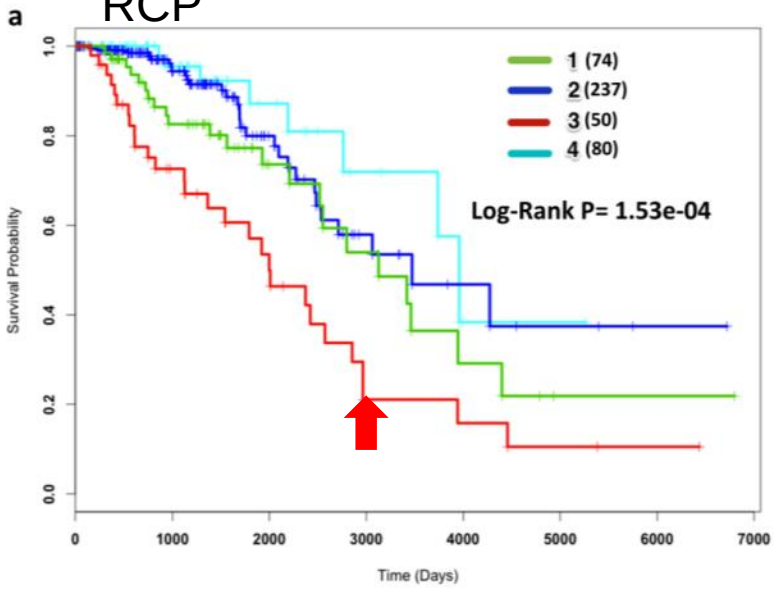
HGPA



CSPA



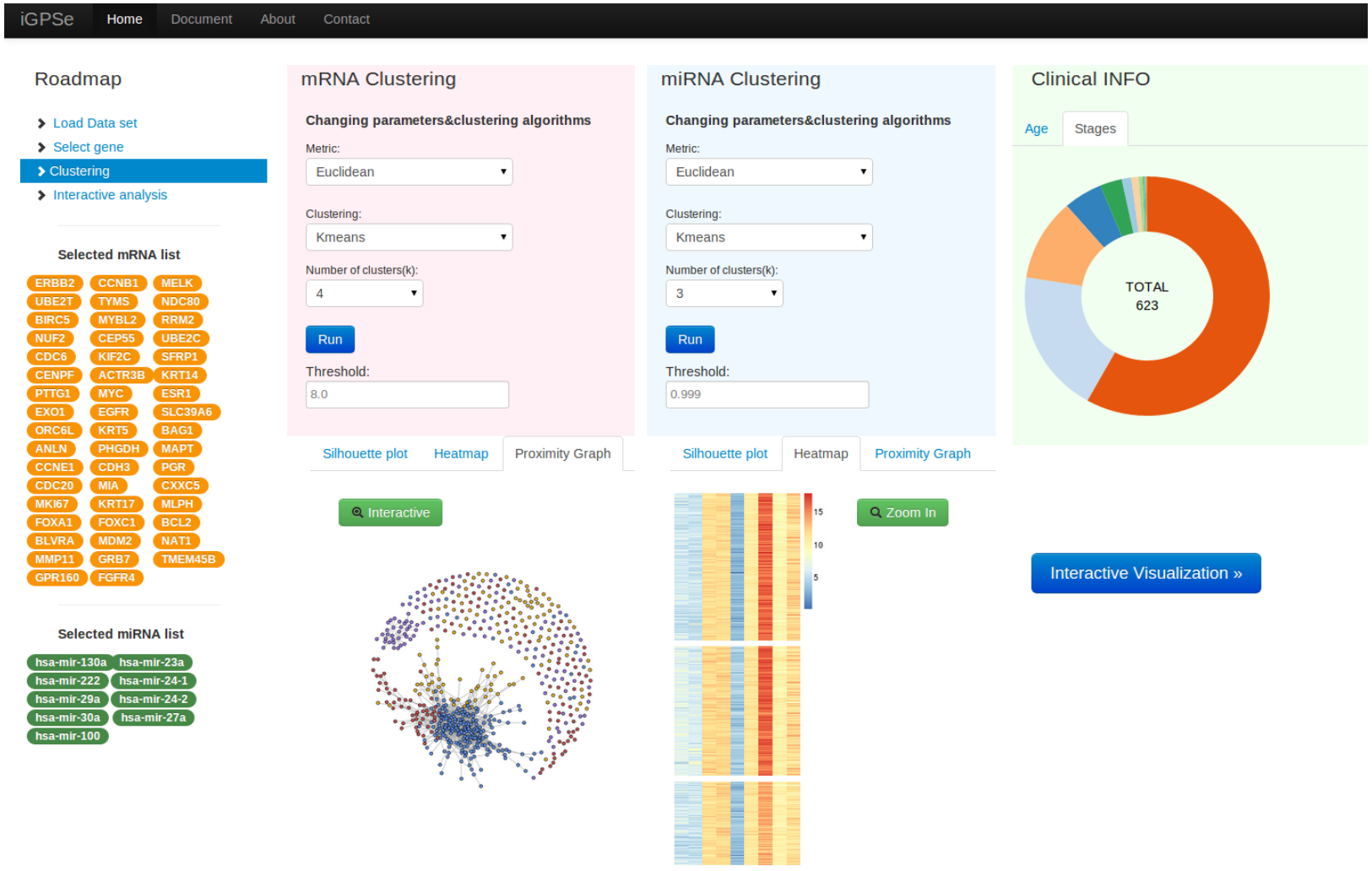
RCP



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

Wexner Medical Center

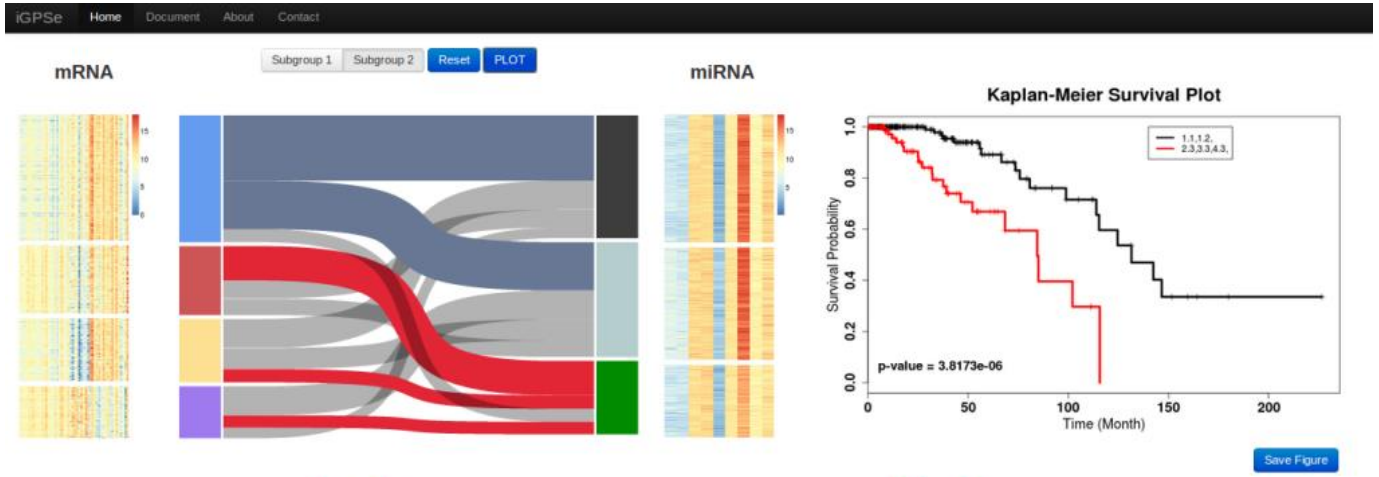
Visualization

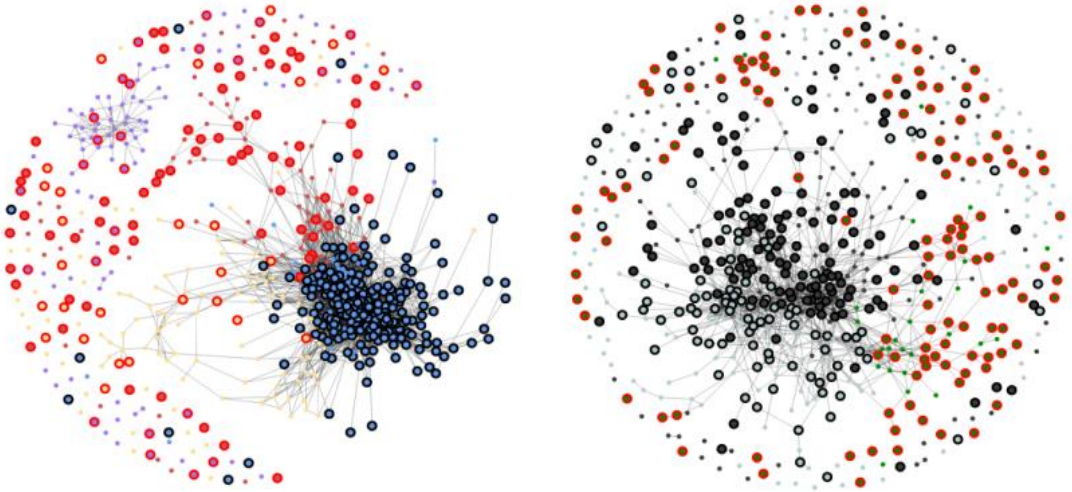


Ding et al, BMC Bioinformatics, 2014

Wexner Medical Center

Interactive Patient Stratification

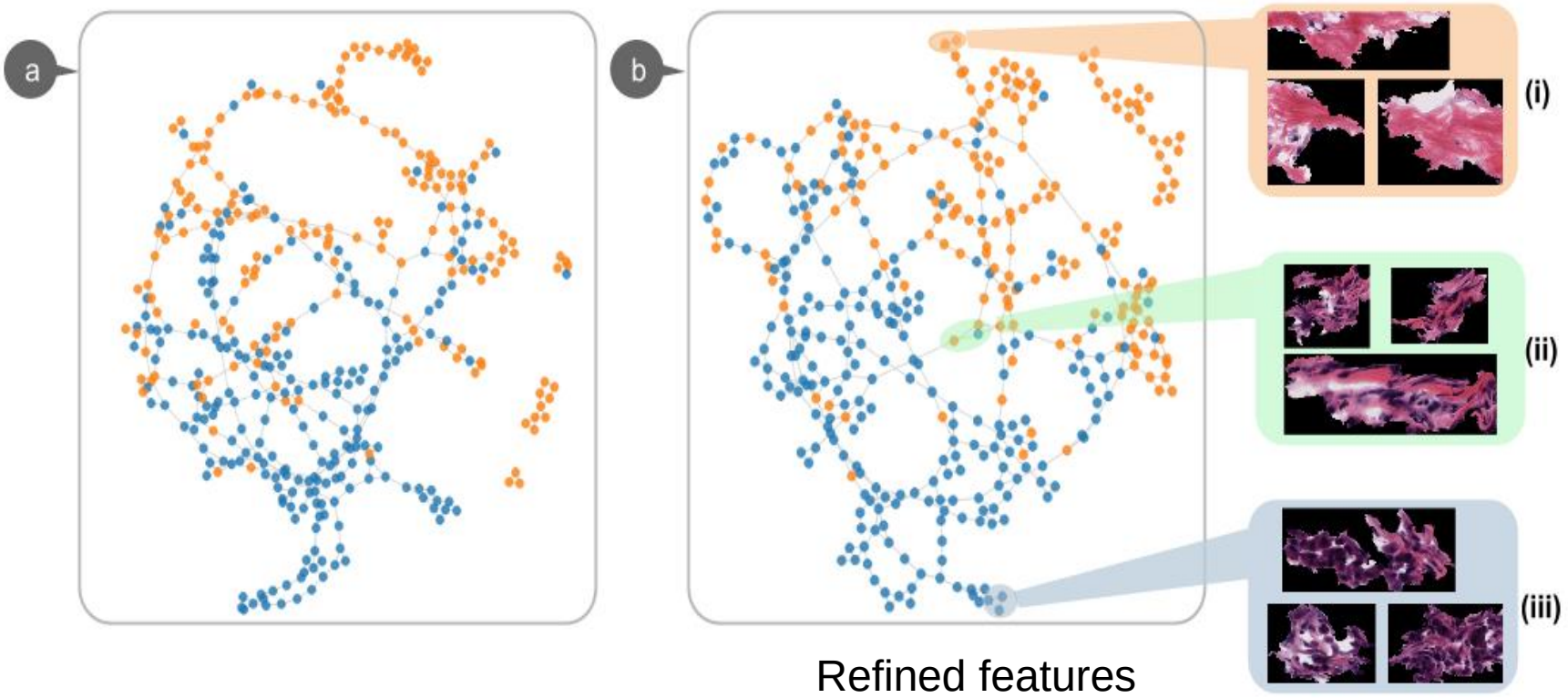




Ding et al, BMC Bioinformatics, 2014

Wexner Medical Center

Graphie – Visual Analytics of Imaging Features



Refined features

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

BMC Bioinformatics

RESEARCH Open Access

GRAPHIE: graph based histology image explorer

Hao Ding[1], Chao Wang[3], Kun Huang[2*], Raghu Machiraju[1*]

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

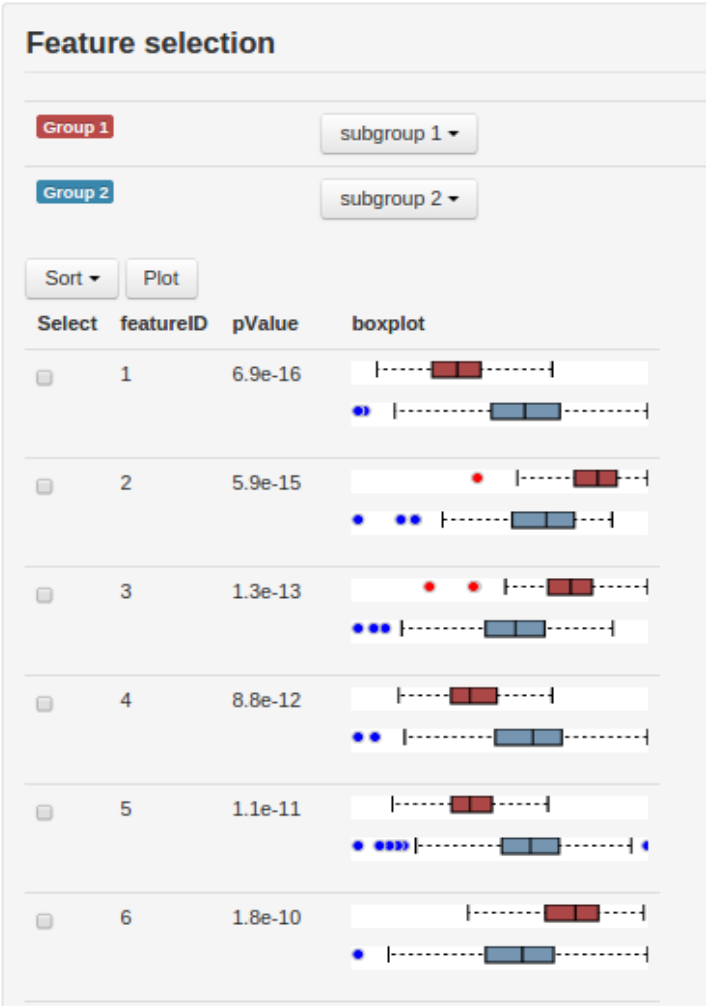
biovis

Wexner Medical Center

Graphie –Feature Selection

Examine feature distinctiveness for groups of images.

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

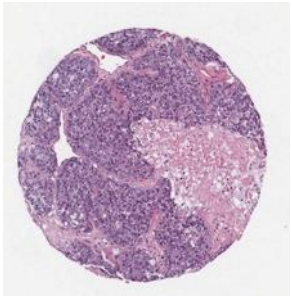


biovis

Wexner Medical Center

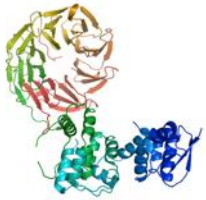
Combining Graphie, iGPSE, & SUMO

Histology images



Genome

Epigenome



Proteome

Subtype Analysis

Integrative Biomarkers

Molecular behaviours

...

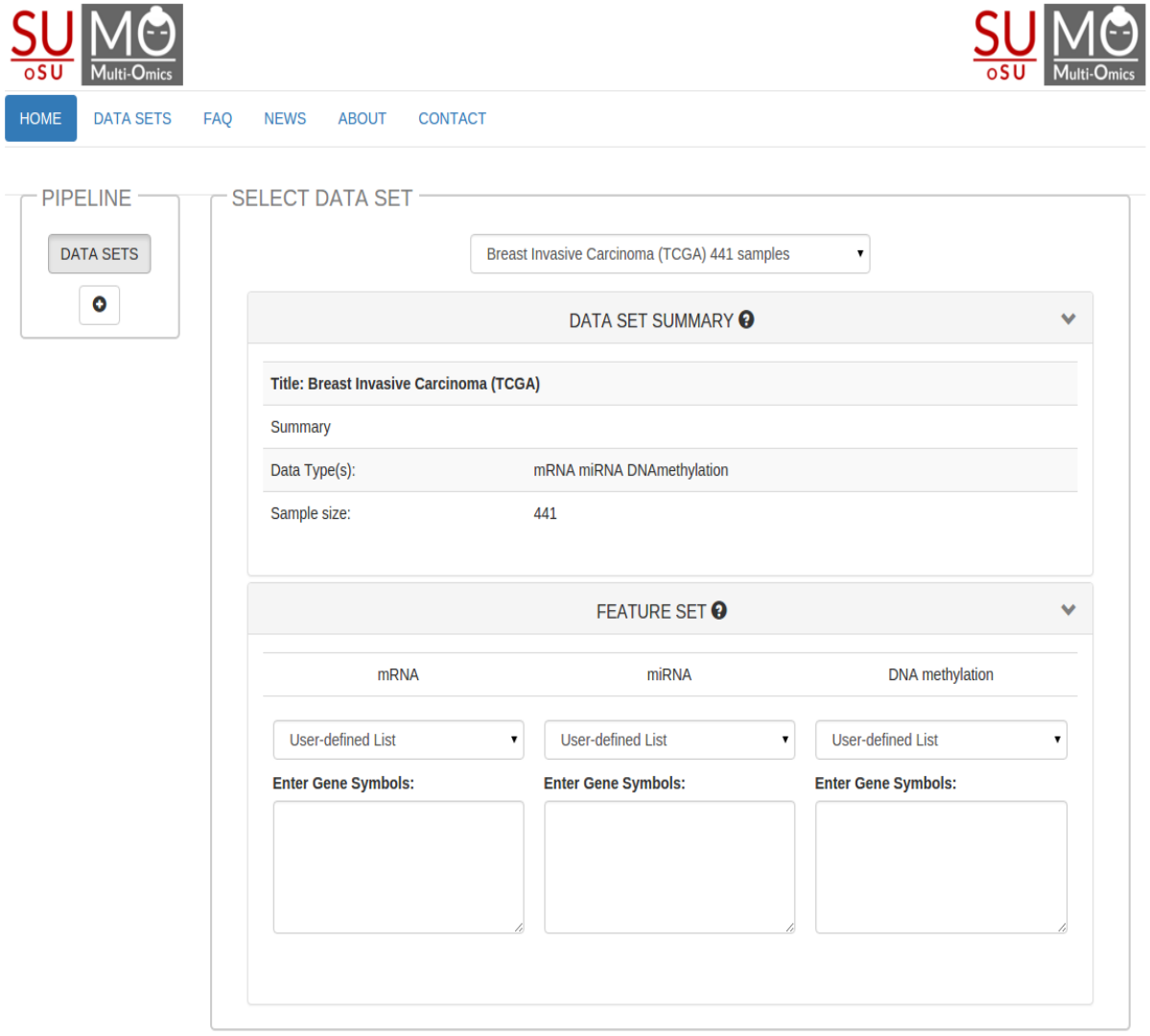
biovis

http://osumo.org/

Kitware

Wexner Medical Center

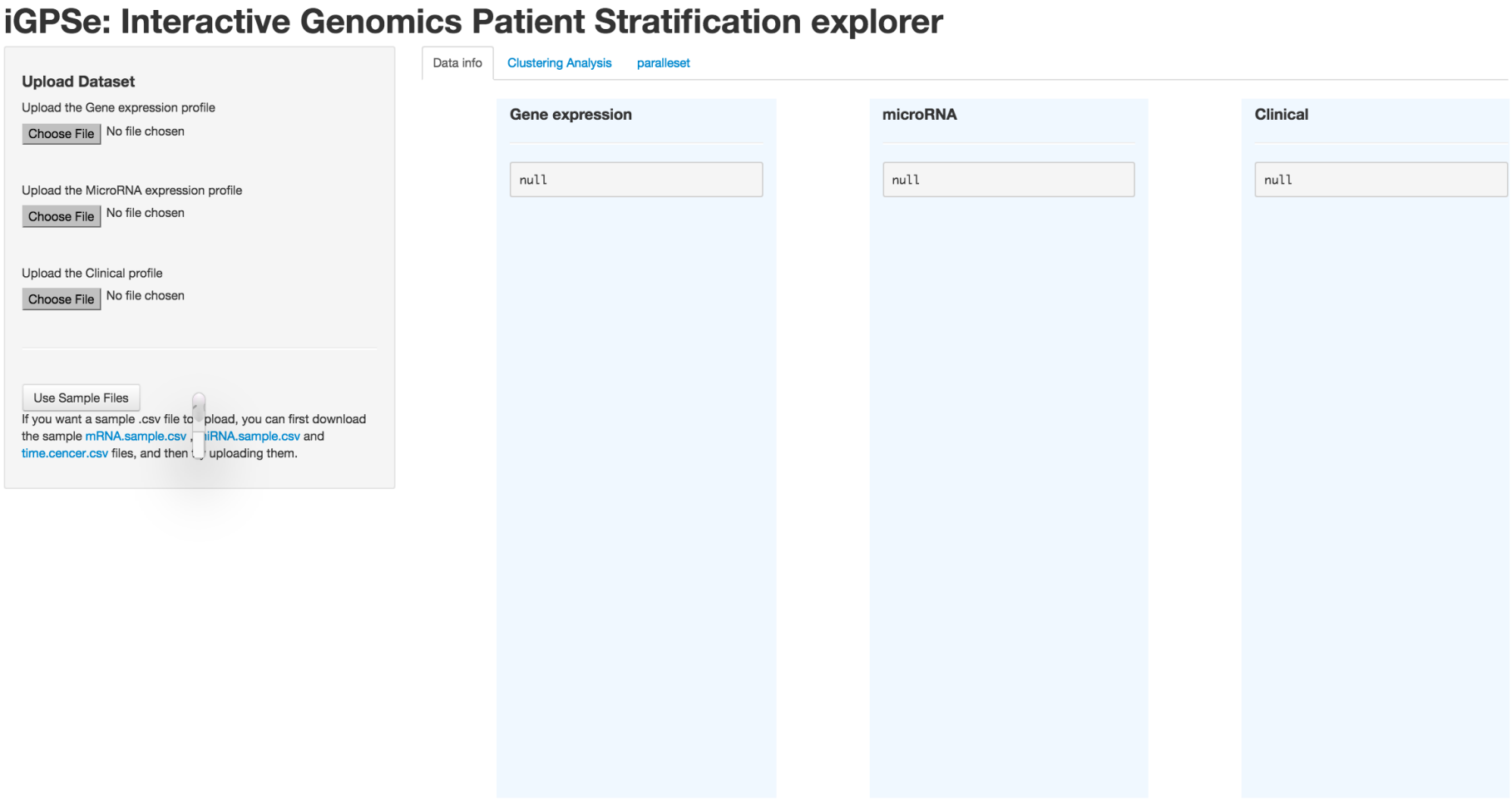
(O)SUMO



Wexner Medical Center

23

State of (O)SUMO



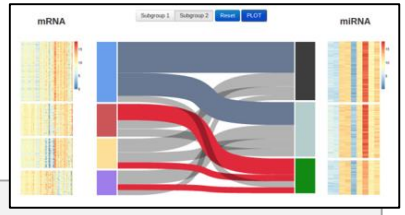
Wexner Medical Center

SUMO Architecture

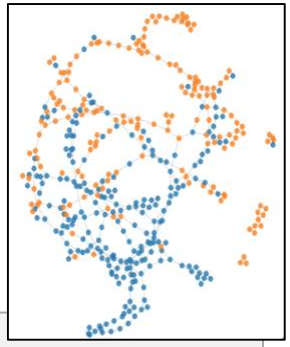
Adding workflows

JSON metadata for tasks

R/Python source



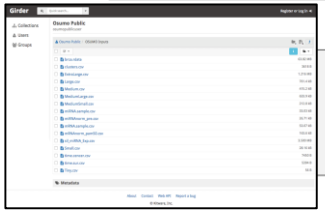
iGPSe



Graphie

...

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))
'''
}
```

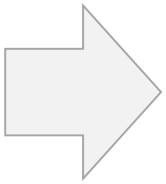
Basic Workflow - Silhouette Plot

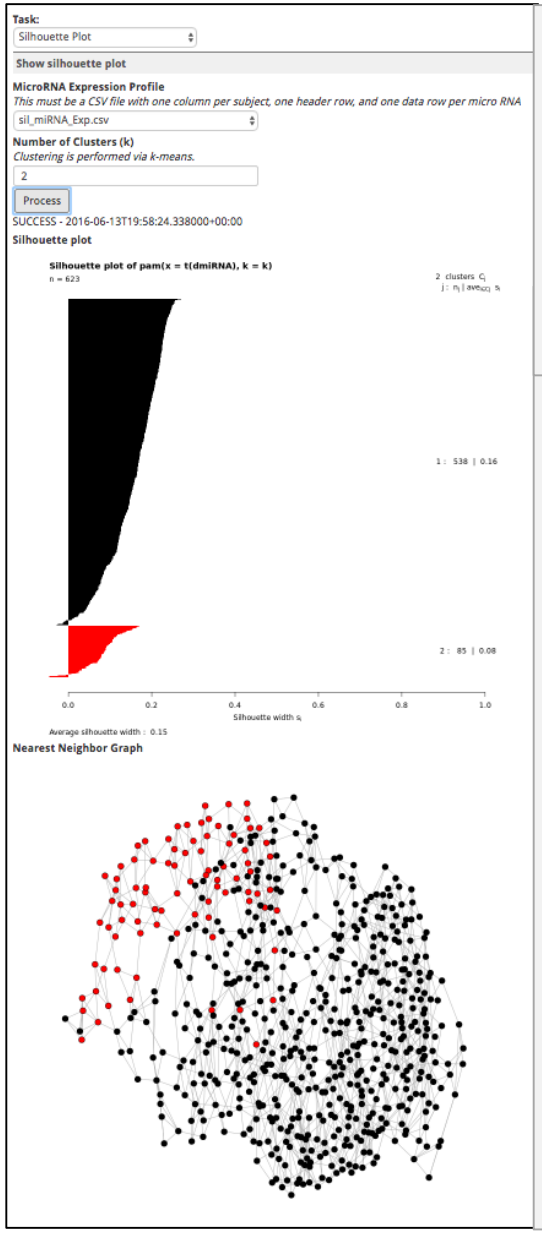
Input / output descriptions

```
---

name: "Silhouette Plot"

description: "Show silhouette plot"

fallbackUser: "osumopublicuser"

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: "miRNA data."
    type: file
    subtype: rdata
    preferredNames: "^sil_.*csv$"
    notes: "This must be a CSV file with one column per subject, one hea
  - 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
dmiRNA = read.csv(sil_input_path)

#clustering
color <- palette()
require(cluster)
k <- silhouette_clusters
cl <- pam(t(dmiRNA),k)
si <- silhouette(cl)

localdir = Sys.getenv('tempdir', unset=tempdir())
dataplot1 = tempfile('plot', tmpdir=tempdir, fileext='.png')
png(filename=dataplot1, width=800, height=800, units="px")

plot(si, col=color[1:k], border=NA) # with cluster-wise coloring

dev.off()


# plot the graph
require(ecod)
require(igraph)
ADJmiRNA <- as.matrix(dist(t(dmiRNA)))
g <- nng(ADJmiRNA, k=3)
g <- as.undirected(g)
V(g)$color <- color[cl$clustering]

localdir = Sys.getenv('tempdir', unset=tempdir())
dataplot2 = tempfile('plot', tmpdir=tempdir, fileext='.png')
png(filename=dataplot2, width=800, height=800, units="px", pointsize=8)

plot(g, vertex.size=3, layout=layout.kamada.kawai, vertex.label=NA)

dev.off()
```





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

al Center

2-Stage Workflow - iGPSe

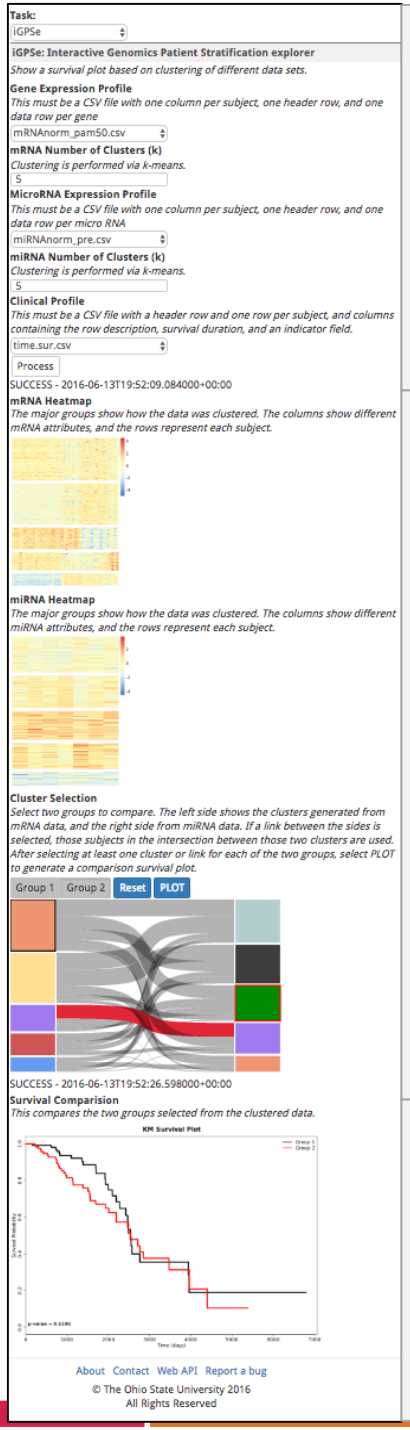
JSON metadata for Stage 1

R source for Stage 1

Workflow link between stages

JSON metadata for Stage 2

R source for Stage 2



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 contributionshttps://github.com/osumo/osumo

Demonstration

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

More Comprehensive Workflow Proposed