Informatics links between histological features and genetics in cancer

Kun Huang, PhD

Department of Biomedical Informatics
The Ohio State University Wexner Medical Center

Department of Medicine
Precision Health Initiative – Data Science Program
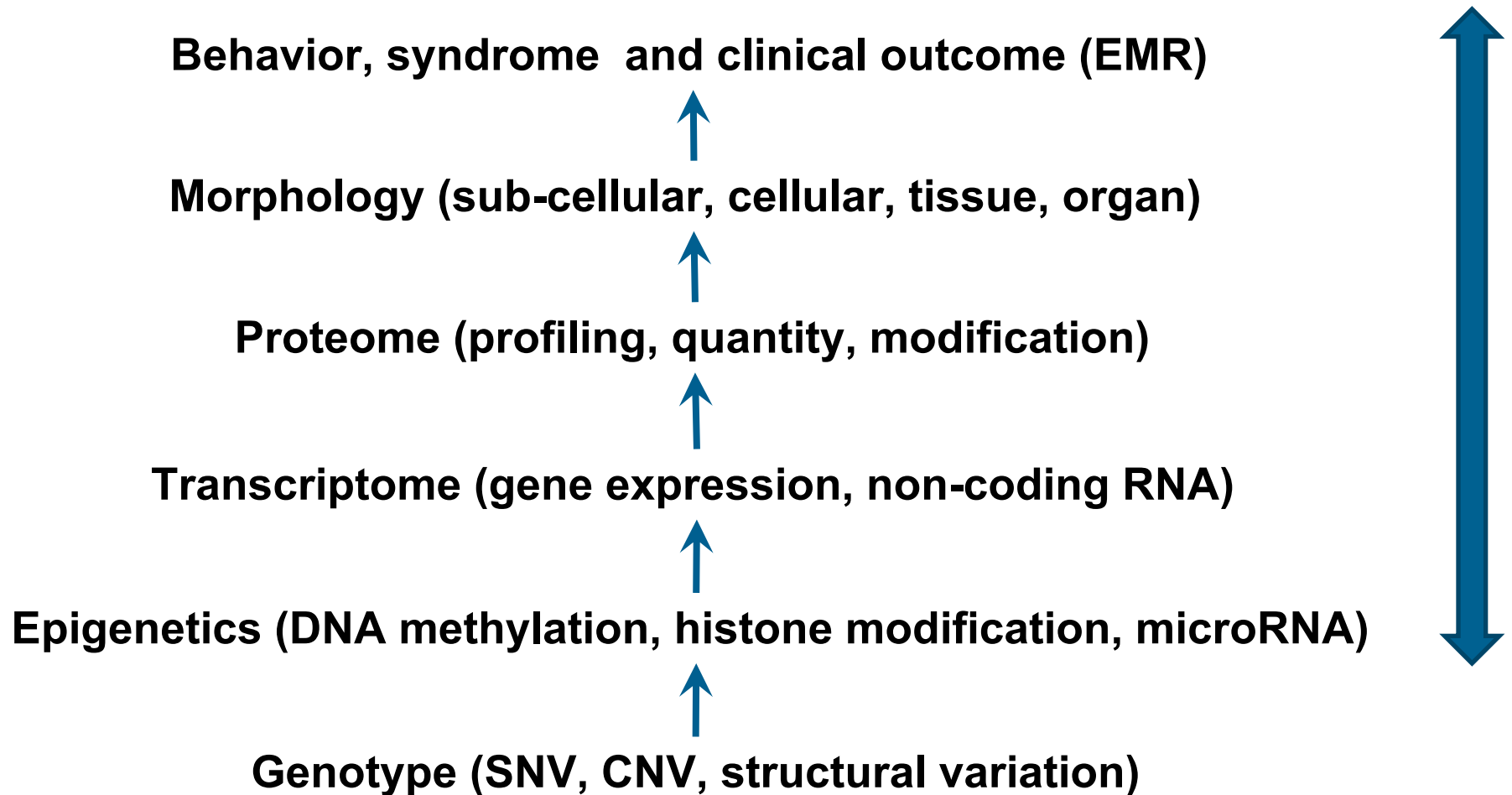**Indiana University School of Medicine**

Data Integration

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

Behavior, syndrome and clinical outcome (EMR)

↑

Morphology (sub-cellular, cellular, tissue, organ)

↑

Proteome (profiling, quantity, modification)

↑

Transcriptome (gene expression, non-coding RNA)

↑

Epigenetics (DNA methylation, histone modification, microRNA)

↑

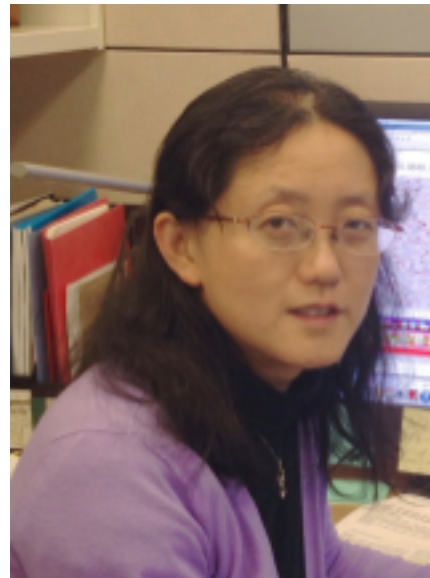**Genotype (SNV, CNV, structural variation)**

Team
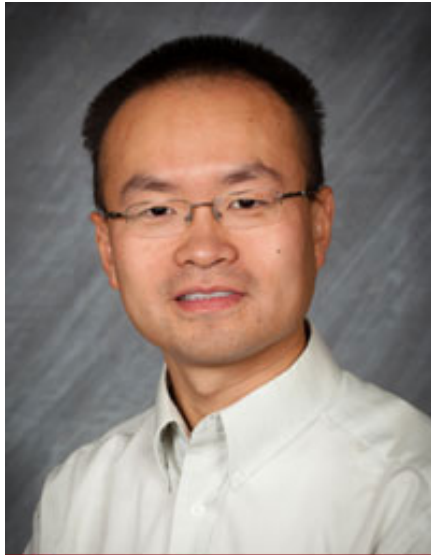
Huang

Machiraju

Jeff Baumes, Kitware

Jie Zhang

Omar Padron, Kitware

dical Center

Leveraging

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

Catalyurek

Mallick

HPC

Proteomics

Trainees

- Chaitanya Kulkarni (OSU CSE)
- Arunima Srivastava (OSU CSE)
- Xusheng Wang (OSU ECE)
- Jun Cheng (SMU, China)
- Yatong Han (HEU, China)
- Tongxin Wang (OSU ECE)
- Dr. Hao Ding (OSU CSE, Walmart Lab)
- Dr. Chao Wang (OSU ECE, Thermo Fisher)
- Dr. Xing Tang (OSUCCC, St. Jude Children's Hospital)
- Sanaya Shroff (Cornell U)
- Kelly Pan (Brown U)
- Duong Au (OSUMC)

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.
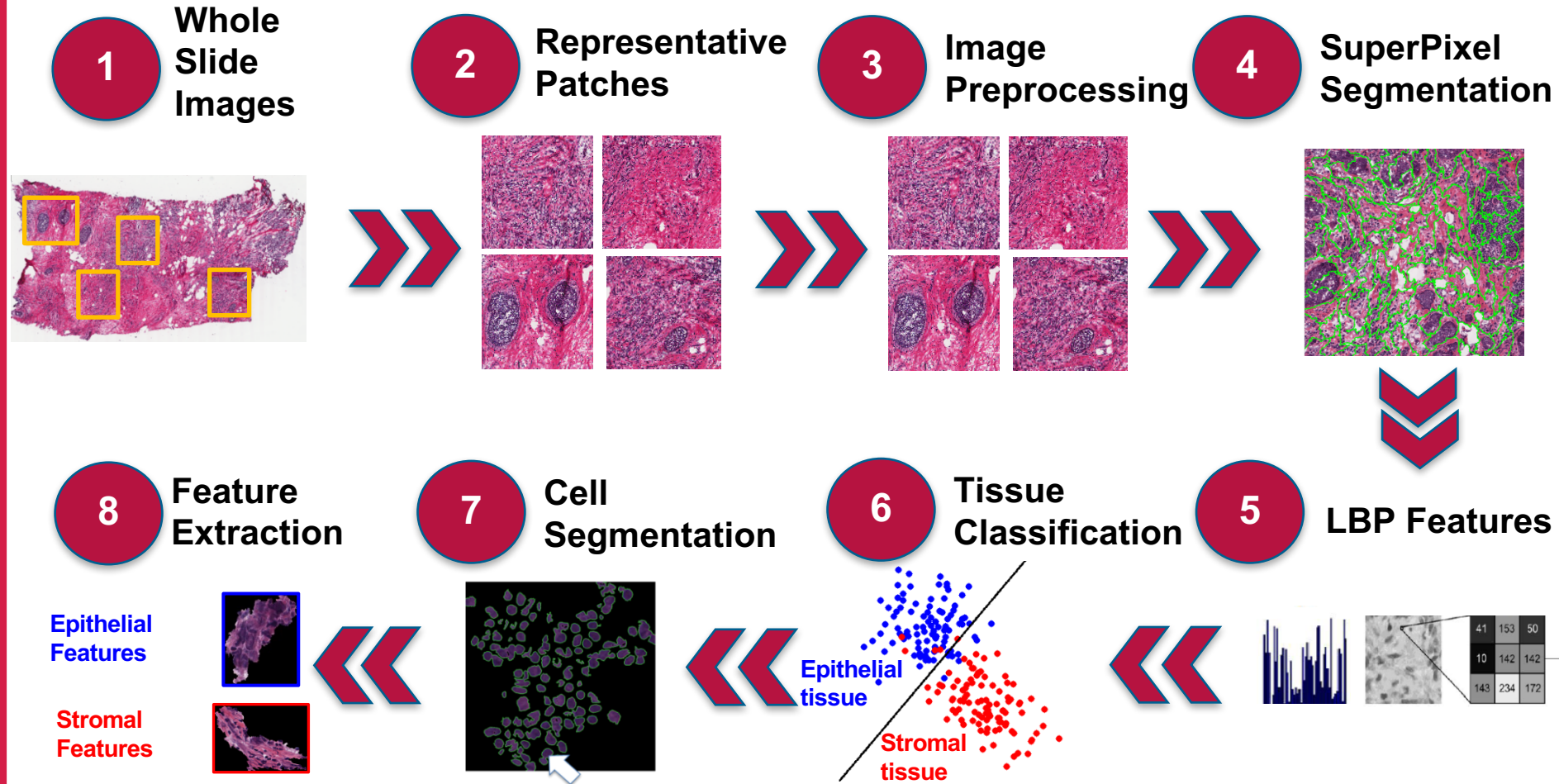
Wexner Medical Center

Aim 1a - Histopathology Image Features

- **Patch-based features**

- **Whole-slide cell morphology statistics**

- **Topological features**

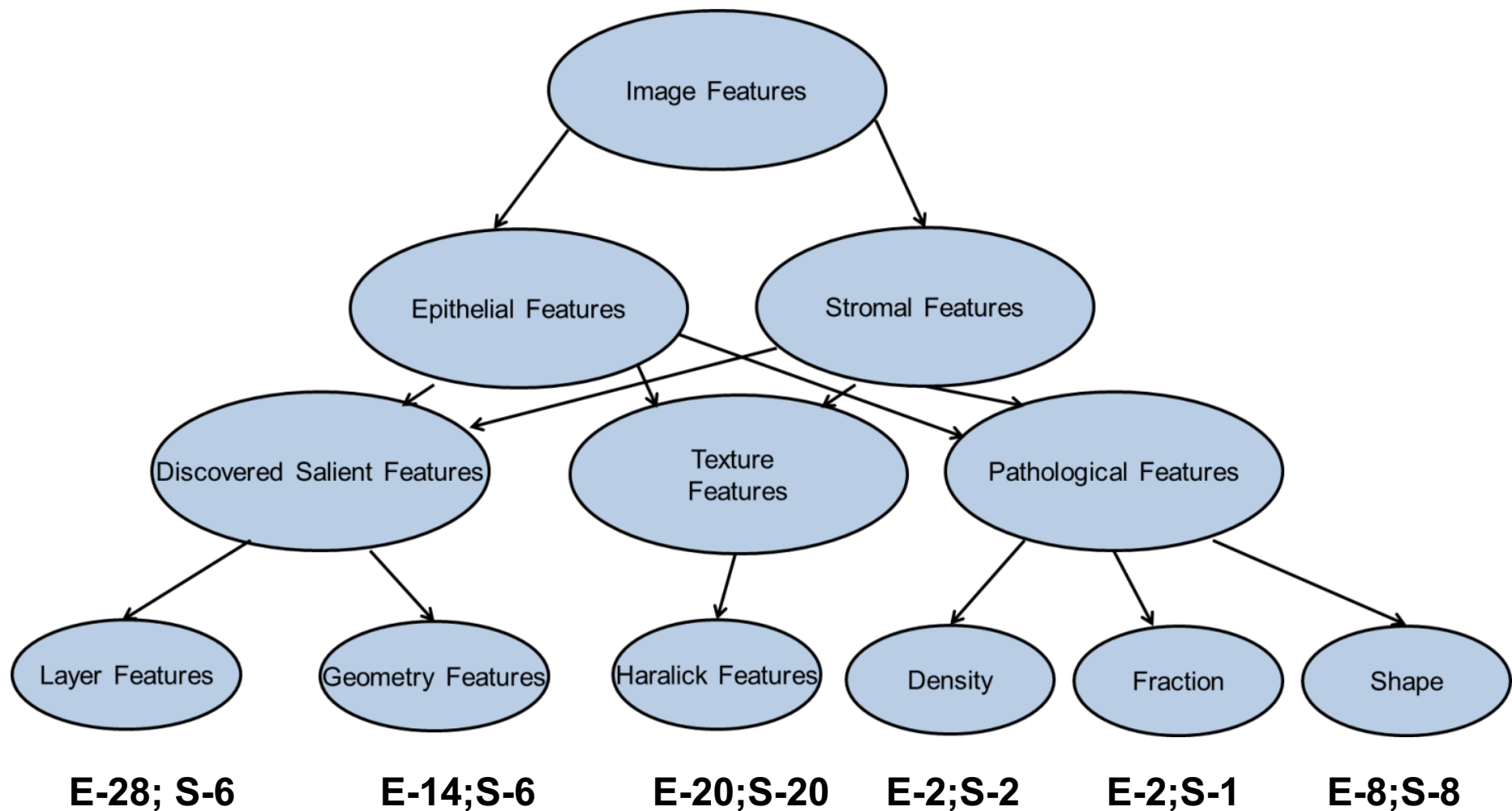- **Features from other tools (e.g., CellProfiler) or collaborative groups**

Wexner Medical Center

Pipeline Overview
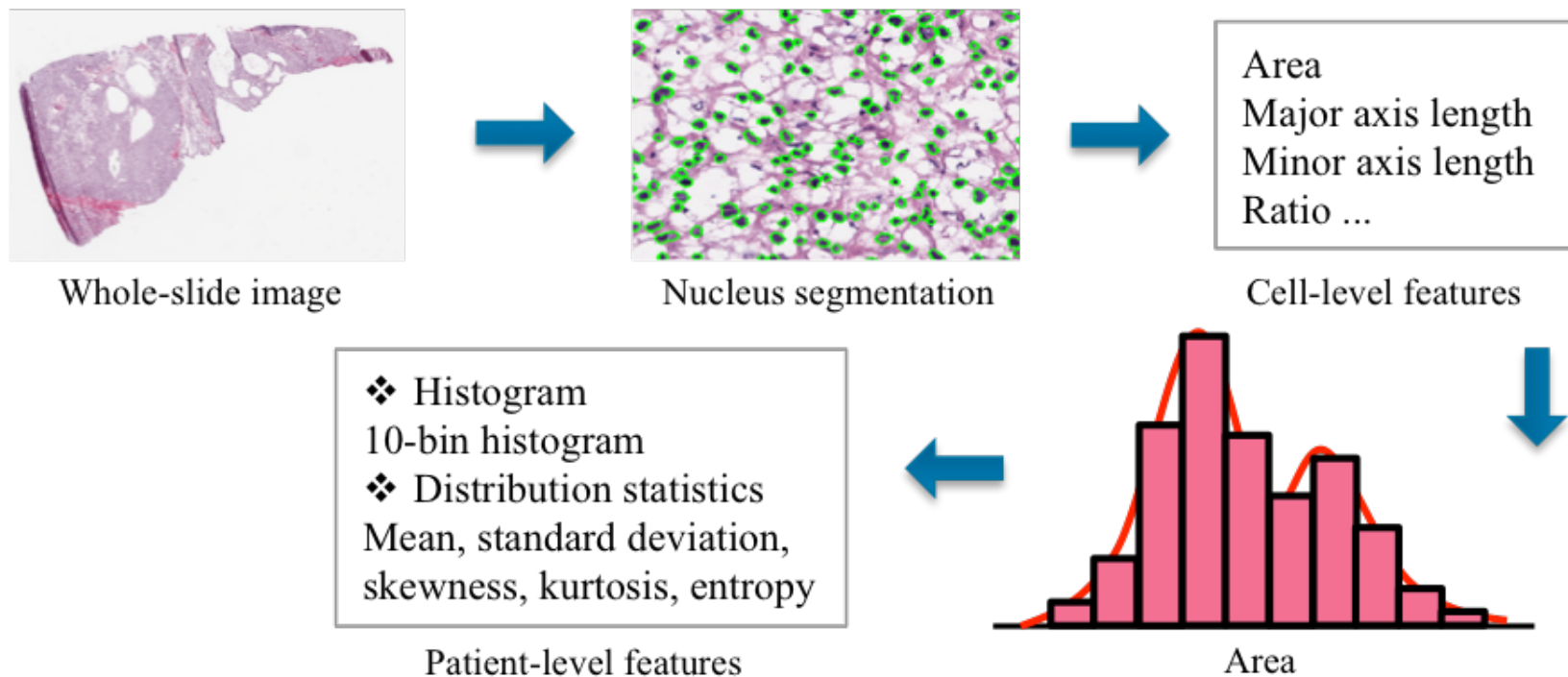

1 Whole Slide Images

2 Representative Patches

3 Image Preprocessing

4 SuperPixel Segmentation

5 LBP Features

6 Tissue Classification

Epithelial tissue

Stromal tissue

7 Cell Segmentation

8 Feature Extraction

Epithelial Features

Stromal Features

Wexner Medical Center

Feature Extraction

E-28; S-6   E-14;S-6   E-20;S-20   E-2;S-2   E-2;S-1   E-8;S-8

E:epithelial features;S:stromal features

Wexner Medical Center

Cell-Level Distribution Features

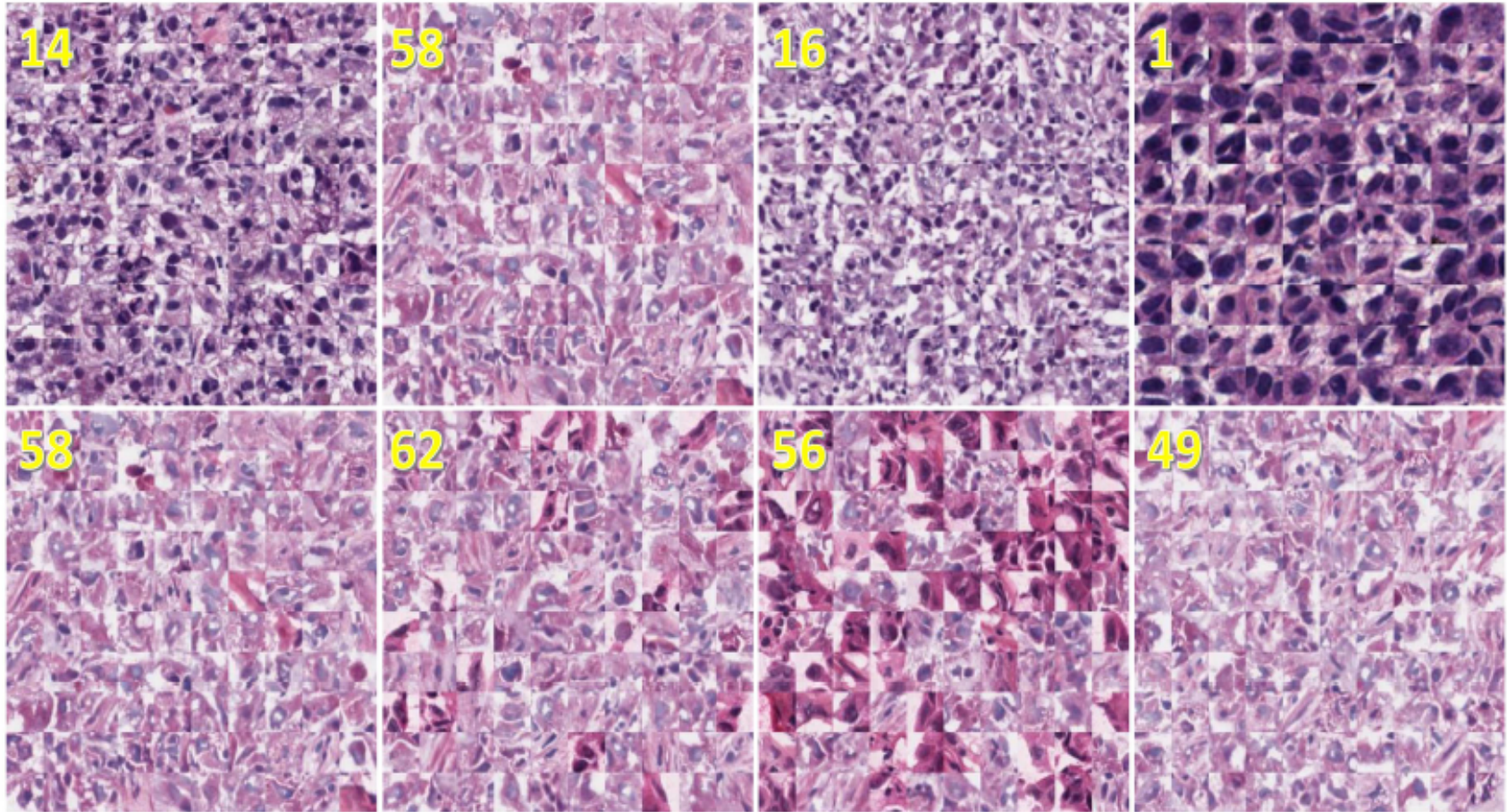Whole-slide cell based analysis with distribution as features.

Whole-slide image

Nucleus segmentation

Area
Major axis length
Minor axis length
Ratio ...

Cell-level features

❖ Histogram
10-bin histogram
❖ Distribution statistics
Mean, standard deviation,
skewness, kurtosis, entropy

Patient-level features

Area

Wexner Medical Center

Topological Features

Nucleus segmentation

Stacked sparse autoencoder to learn compact features

Input   Encoder   Encoder

K-means clustering in feature space

(a)

Delaunay triangulation graph

Bag of edge histogram

frequency

AA AB AC BB BC CC

Survival-related image features

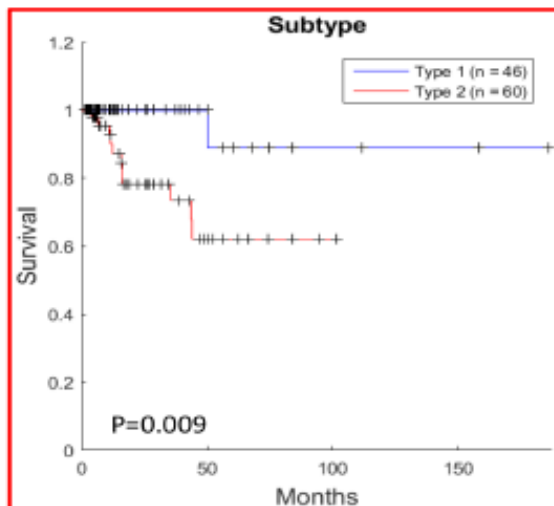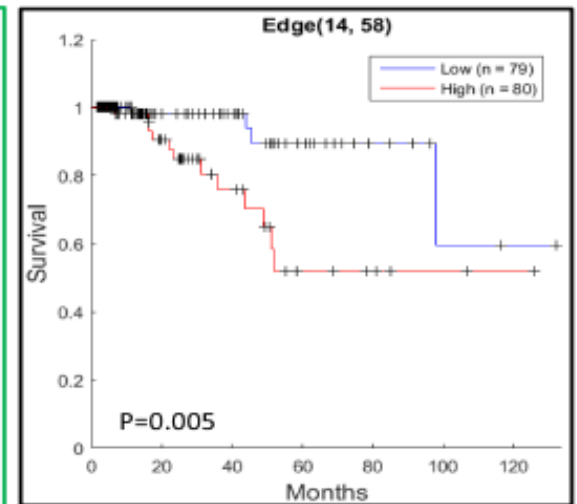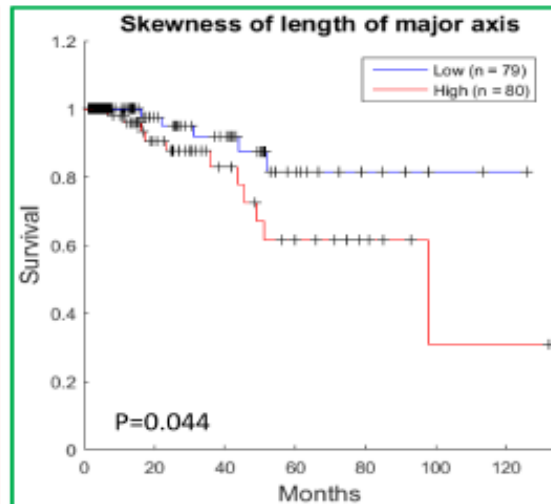survival

time

(b)

cal Center

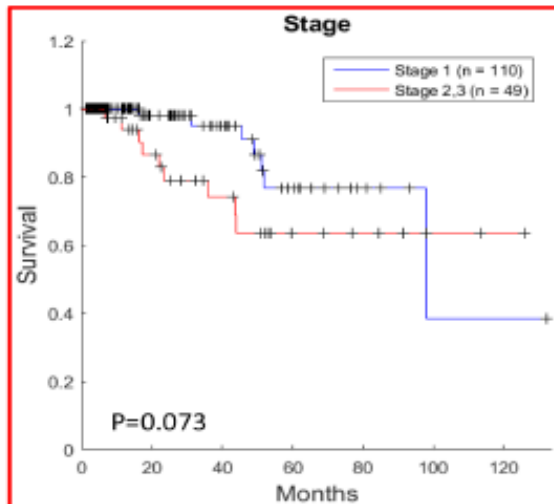# Papillary Cell Renal Cell Carcinoma

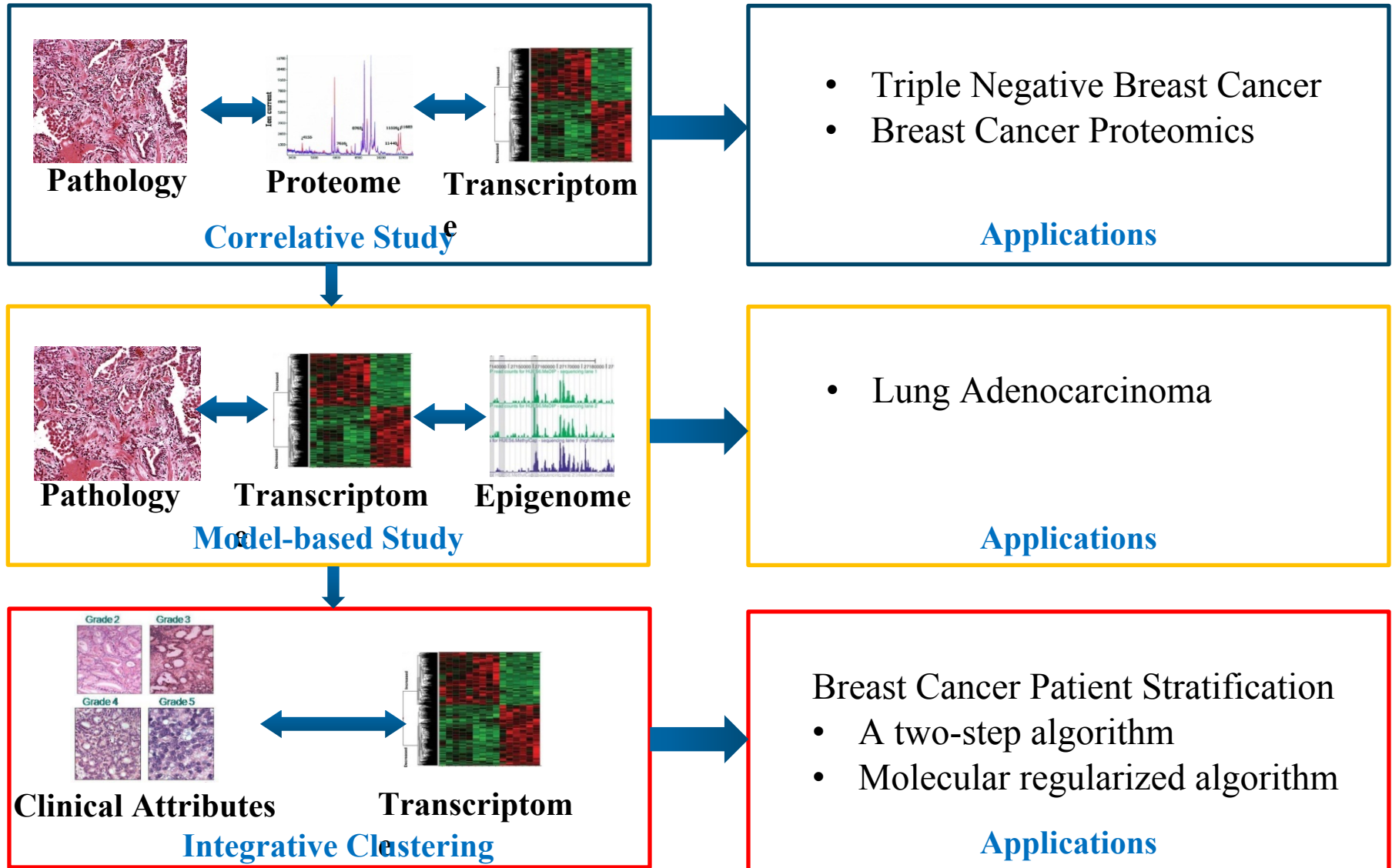Edge(14, 58)   Edge(58, 62)   Edge(16, 56)   Edge(1, 49)

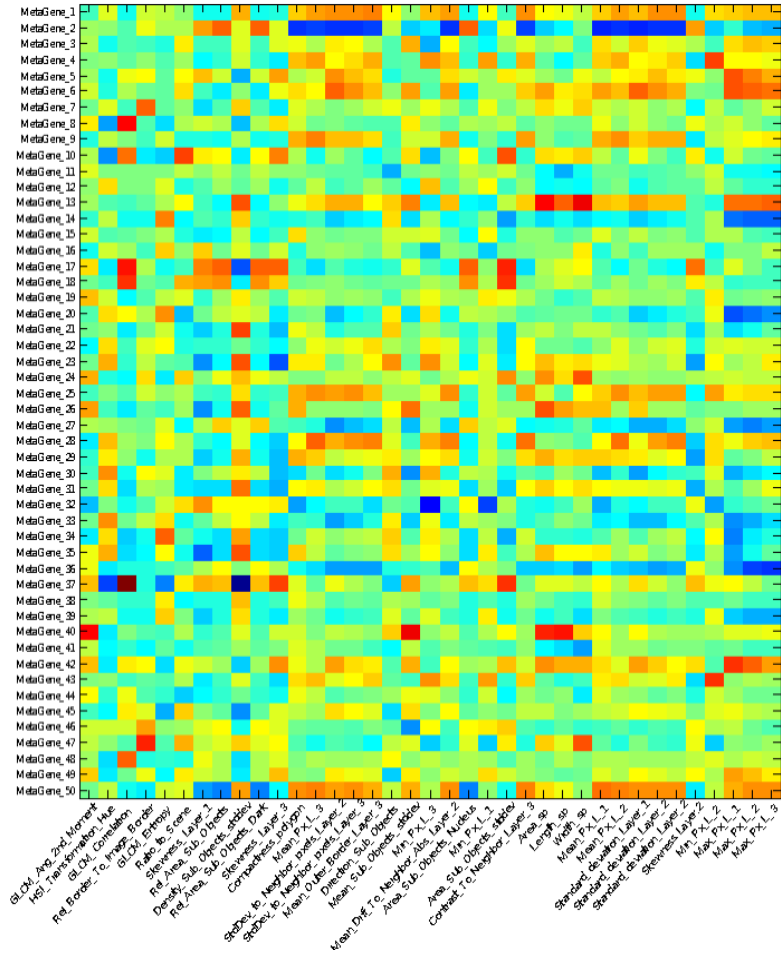# Papillary Cell Renal Cell Carcinoma
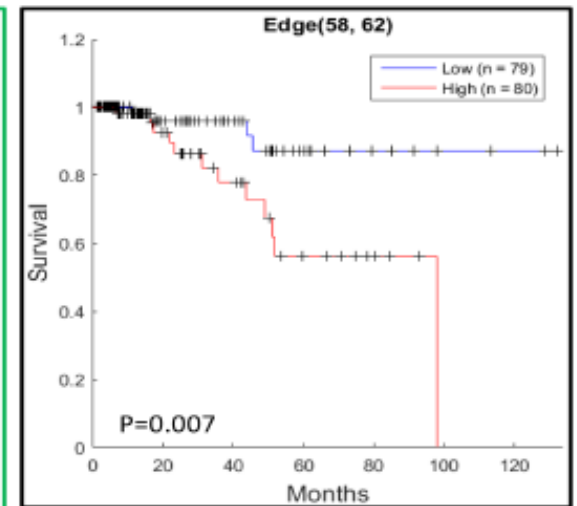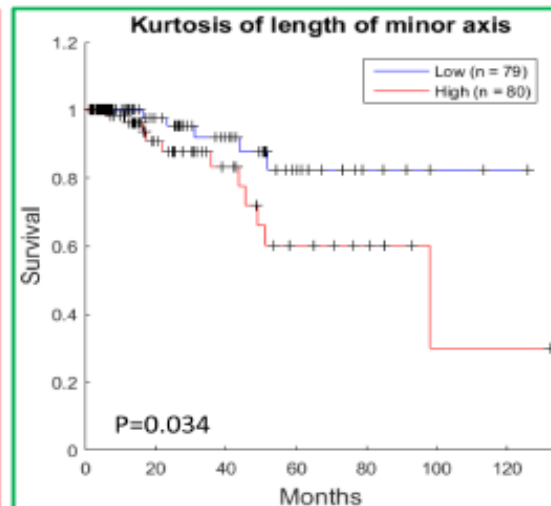
Aim 1b - Correlation to Integration

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

Eigengene – Imaging Feature Correlation

Wexner Medical Center
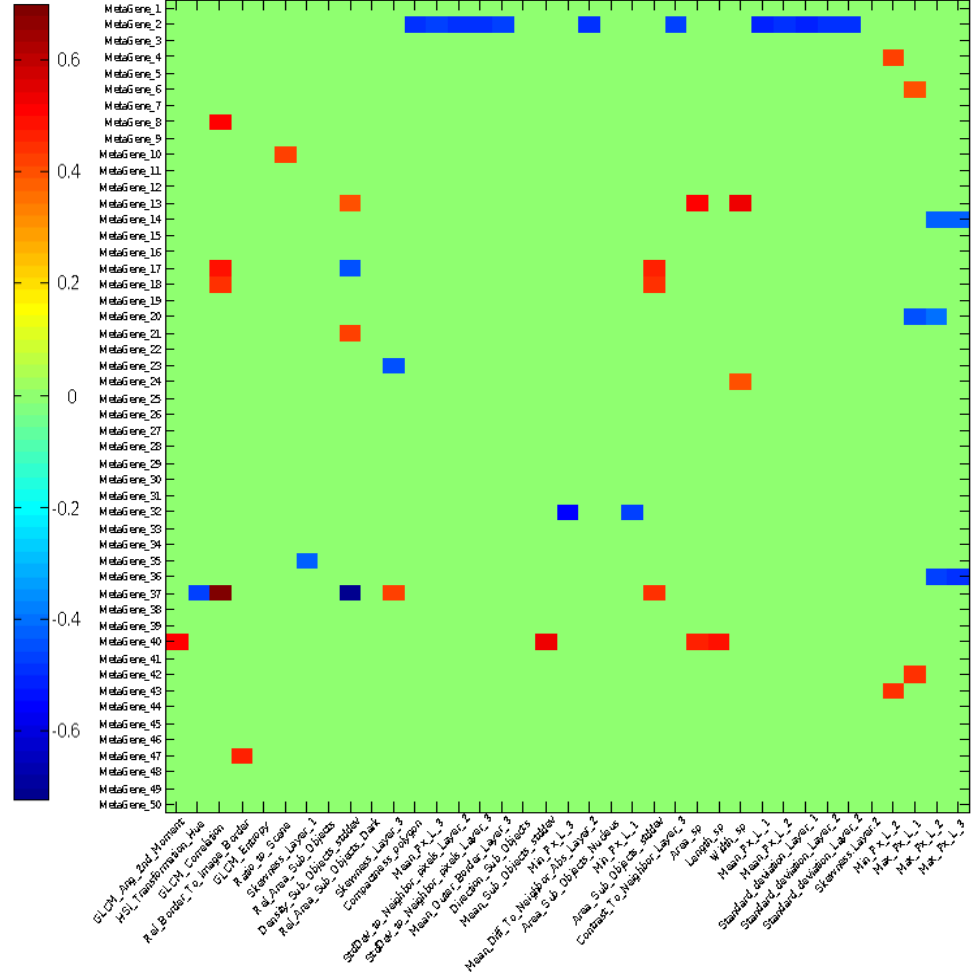
Breast Cancer

Patient Stratification on Morphology

"High malignancy" group   "Low malignancy" group
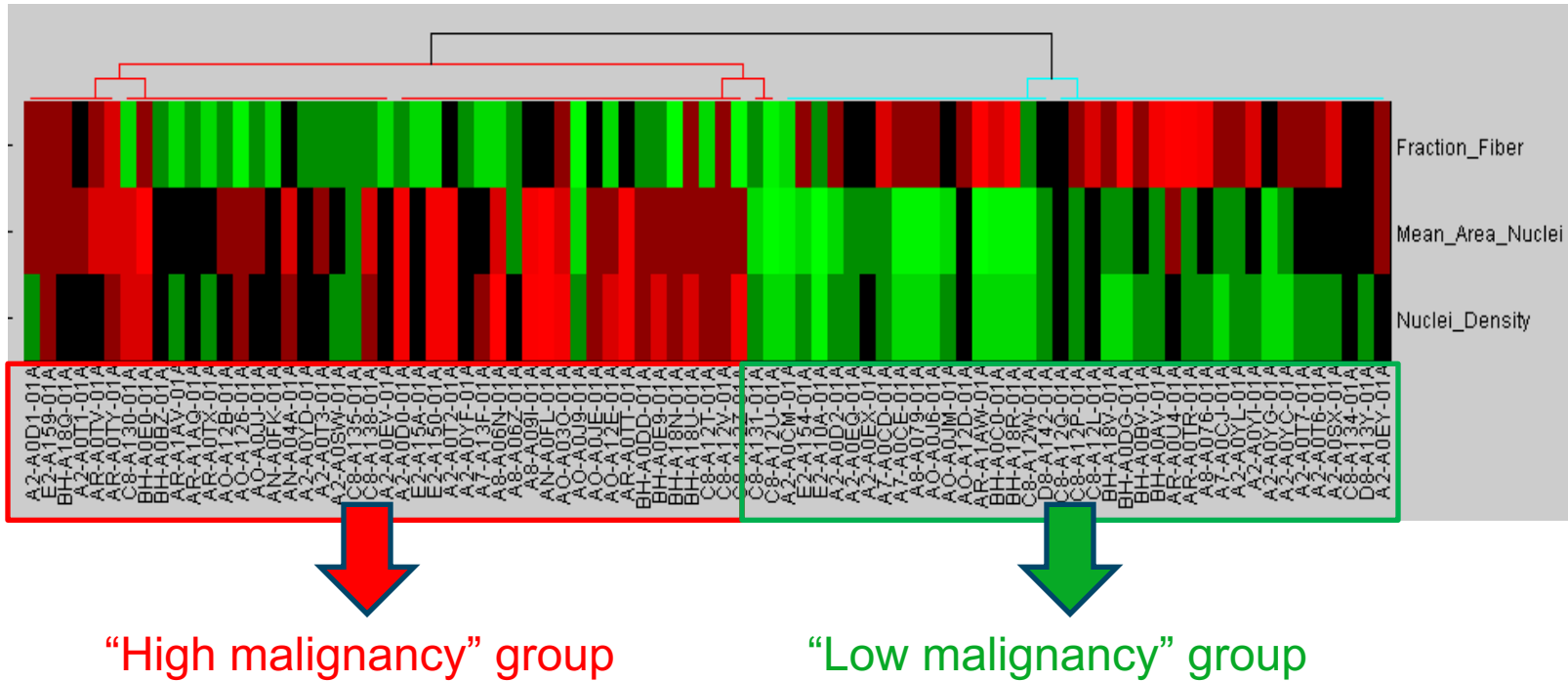
Hierarchical clustering of breast cancer patients based on imaging features.

Lung Adenocarcinoma

TCGA Slide Images → Image Features → Survival-related features

The Lasso Model

$$\sum_{i=1}^{N} \left(y_i - \sum_j \beta_j x_{ij} \right)^2 + \underline{\lambda} \sum_j \beta_j^2 \qquad \text{penalty}$$
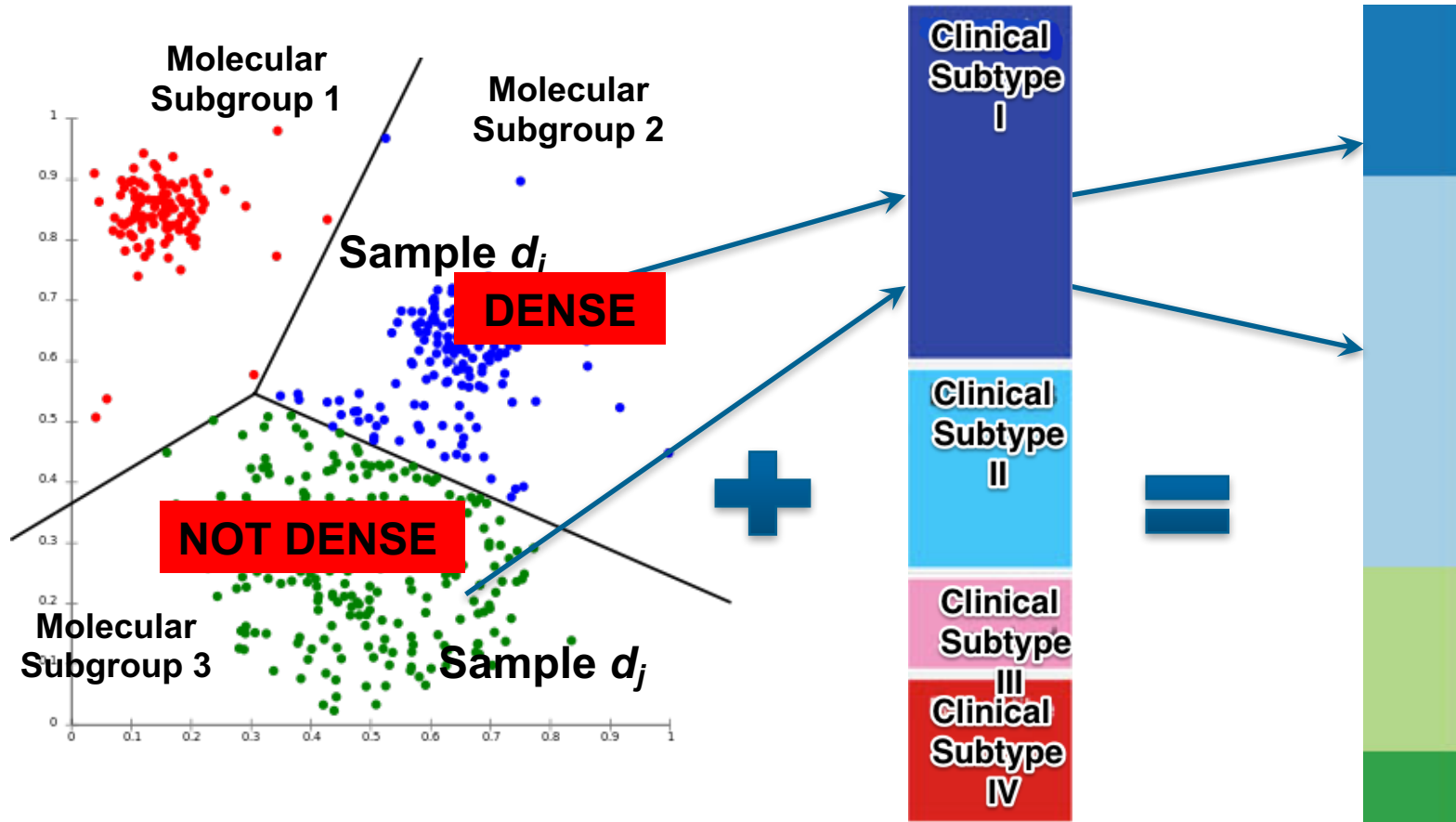
mRNA expressions → Co-expressed Gene Network → Co-expressed Gene Clusters
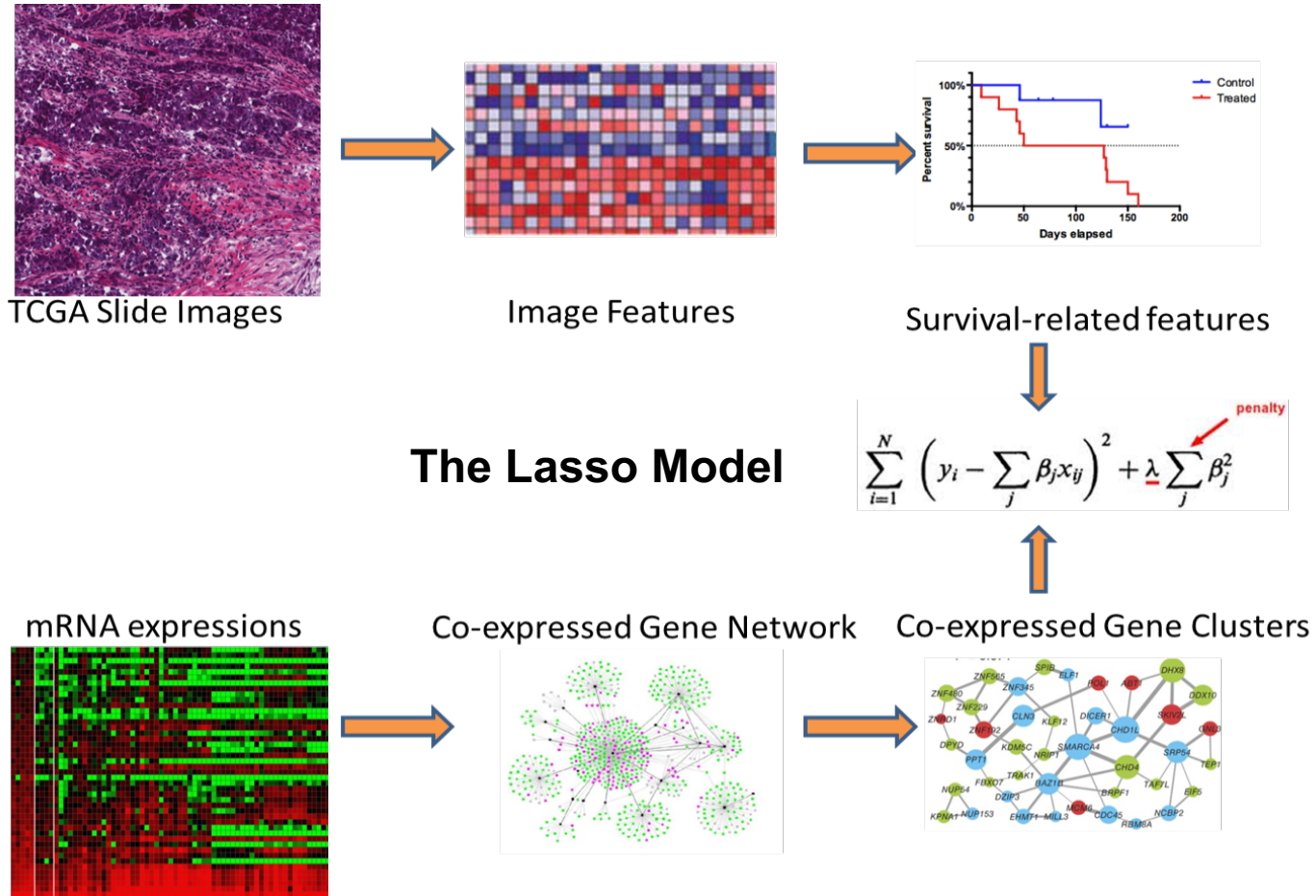
Schematic overview of constructing gene networks and analyzing correlations between gene networks and morphological features

Wexner Medical Center

Morphology and genetic variations

Gene Cluster (size)	beta	GO Biological Process/p-values	Cytobands/p-values	Notes:
4 (59)	-1.1558	GO:0006614 SRP-dependent cotranslational protein targeting to membrane / 9.105E-98		
31 (10)	1.3894			Genes down-regulated in nsopharyngeal carcinoma relative to the normal tissue (p = 5.074e-19, all 10 genes)
33 (10)	-1.7213		19q13.42/5.525e-6	All 10 genes on 19q13.3-4
40 (8)	1.2977		8q24.13/3.263e-5	Seven genes on 8q21-24, one on 8q13
59 (7)	-1.5729		16p11.2/1.364e-10	All seven genes on 16p11
60 (7)	1.2103		Xq28/1.982e-13	All seven genes on Xq27-28
61 (7)	-1.6639		6p21.1/4.436e-7	Six genes on 6p21-22, one on 6p12
70 (6)	2.1783		17q21.31/5.532e-10	All six genes are on 17q21
74 (6)	-2.0544		8p11.2/1.048e-9	All six genes are on 8p11.2
75 (6)	-1.0093	GO:0050680 negative regulation of epithelial cell proliferation / 3.290E-6	17q11.2/6.880e-7	All six genes are on 17q11-12
87 (5)	1.9569		19q13.2/1.131e-6	All five genes on 19q13
95 (5)	2.5783		18q12.1/1.175e-9	All five genes on 18q12

Wexner Medical Center

Concept

Molecular Subgroup 1

Molecular Subgroup 2

Sample d_i

DENSE

NOT DENSE

Molecular Subgroup 3

Sample d_j

Clinical Subtype I

Clinical Subtype II

Clinical Subtype III

Clinical Subtype IV

Wang et al, Methods, 2013

Wexner Medical Center

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**

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

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

$\quad$ Compute $\Delta = \|\widetilde{S} - W \circ (\widetilde{U}D\widetilde{U}^T)\|_F^2$;
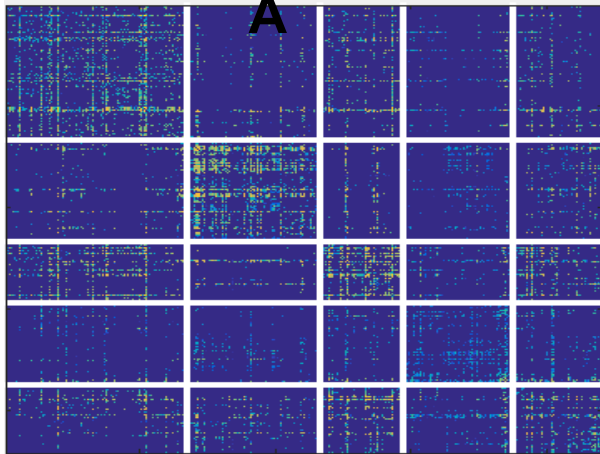
$\quad$ $t = t + 1$;

end

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

Algorithm 1: Molecular Regularized Consensus Patient Stratification

Wang C, Machiraju R, Huang K, *Methods*, 2014
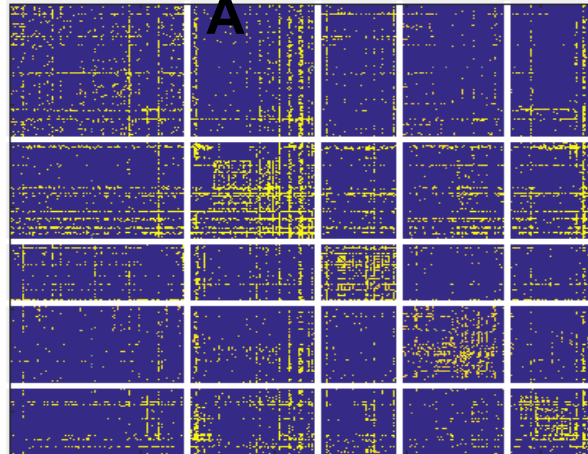
Ongoing – Grassmanian Integration

mRNA

miRNA

DNA methylation

Integration

Wexner Medical Center
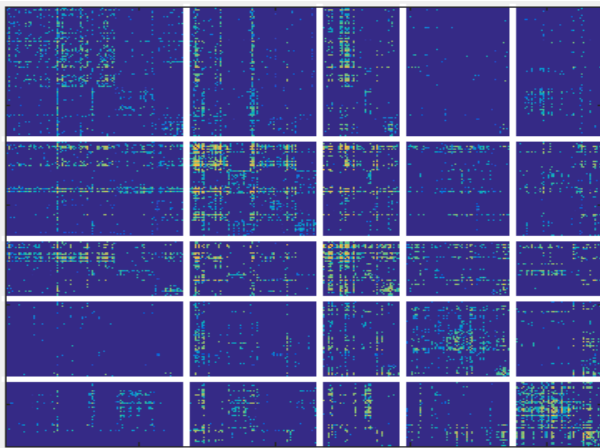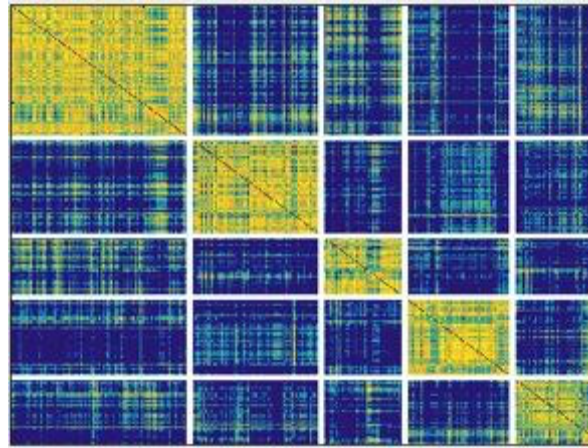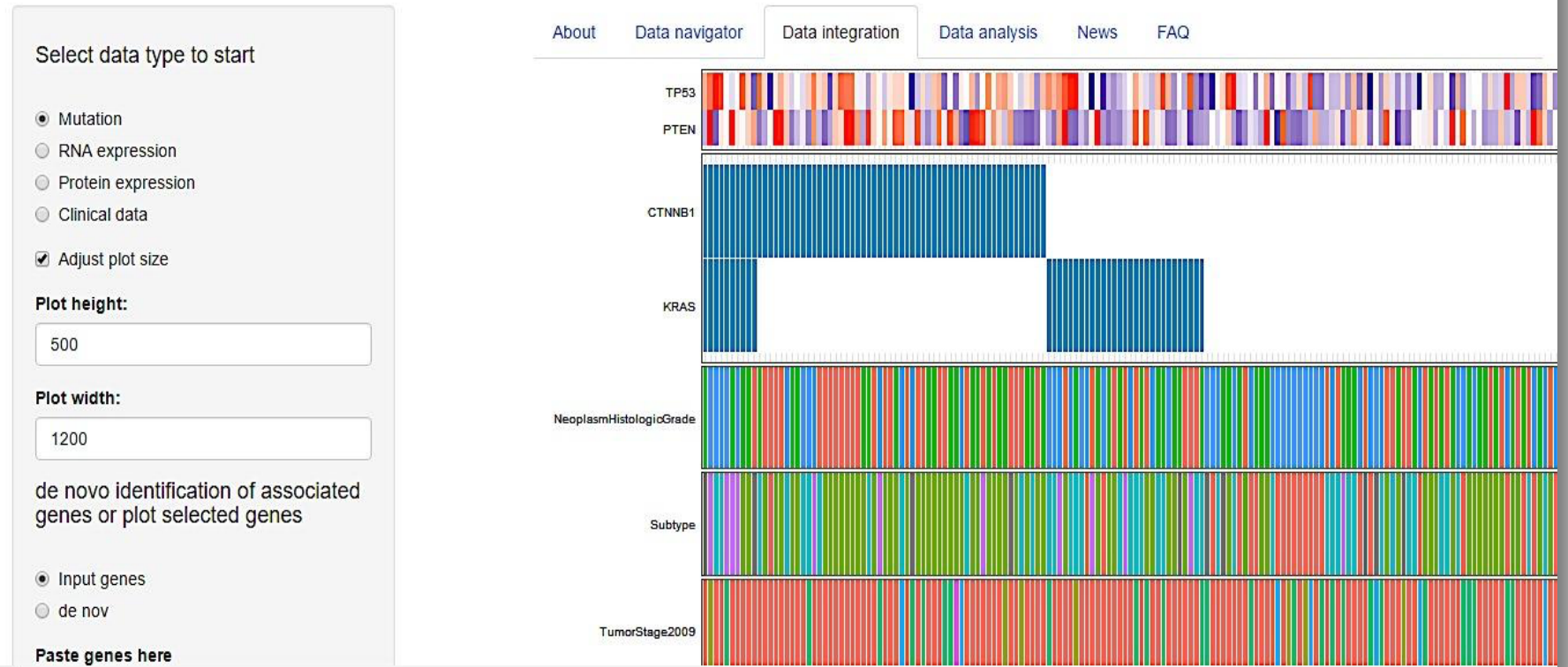
Aim 2 - Software

- **OSUMO (Ohio State University Multi-Omics tool)**

- **iGenomicsR (R/Shiny)**

- **WINGS (in collaboration with Dr. Parag Mallick)**

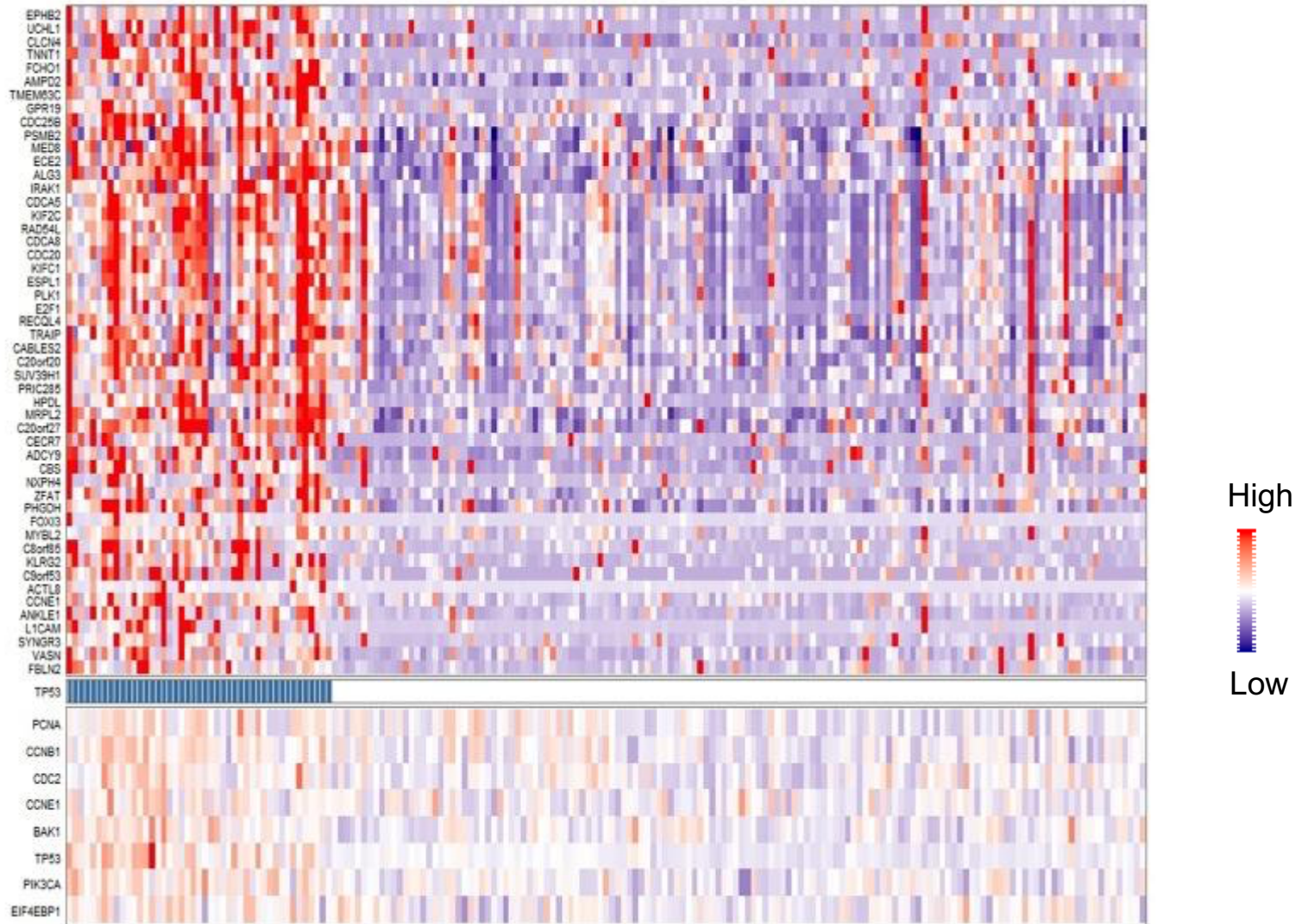- **GRAPHIE**

- **AnnoPeak**

- **ImQCM**

Wexner Medical Center

iGenomicsR

Cell cycle genes is highly expressed in TP53 mutated patients

WINGS

Workflow for Nature Communications publication

Whole slide image
(TCGA/GDC & TMA)

Tiling of image
(bftools – Open Microscopy Environment)

Choose **10** densest tiles
(Calculating % of whitespace – image processing)

Segmented dense tiles
(CellProfiler)

Cell Size and Shape

Pixel Intensity of the Cell
Radial Distribution of Pixel Intensity of the Cell

Cell Texture

Area Occupied by Cells

Image feature (~9K) value extraction
(CellProfiler)

Tiles	Cell Size	Cell Texture	....
T1	.08	.89	1.2
...	...	...	...

Across Cell Aggregation for tile
(Mean, Median, SD, deciles)

input
Classifier
output

Machine learning classifiers and predictive models
(SVM flavors, Random forest, Bagging, Naïve Bayes)

Evaluation

Testing models
(ROC curve, Cross Validation,)

Wexner Medical Center

Graphie – Visual Analytics of Imaging Features
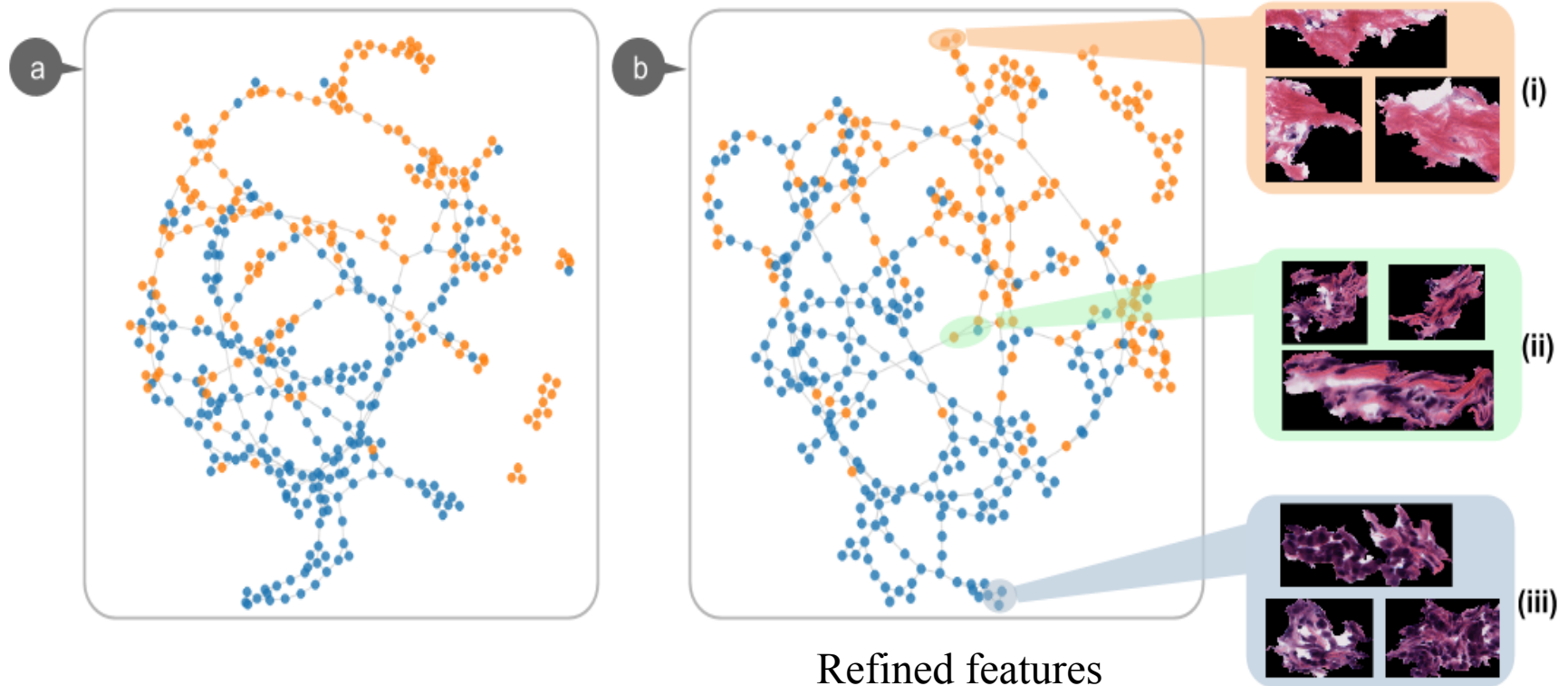
Refined features

BMC Bioinformatics

RESEARCH   **Open Access**

GRAPHIE: graph based histology image explorer

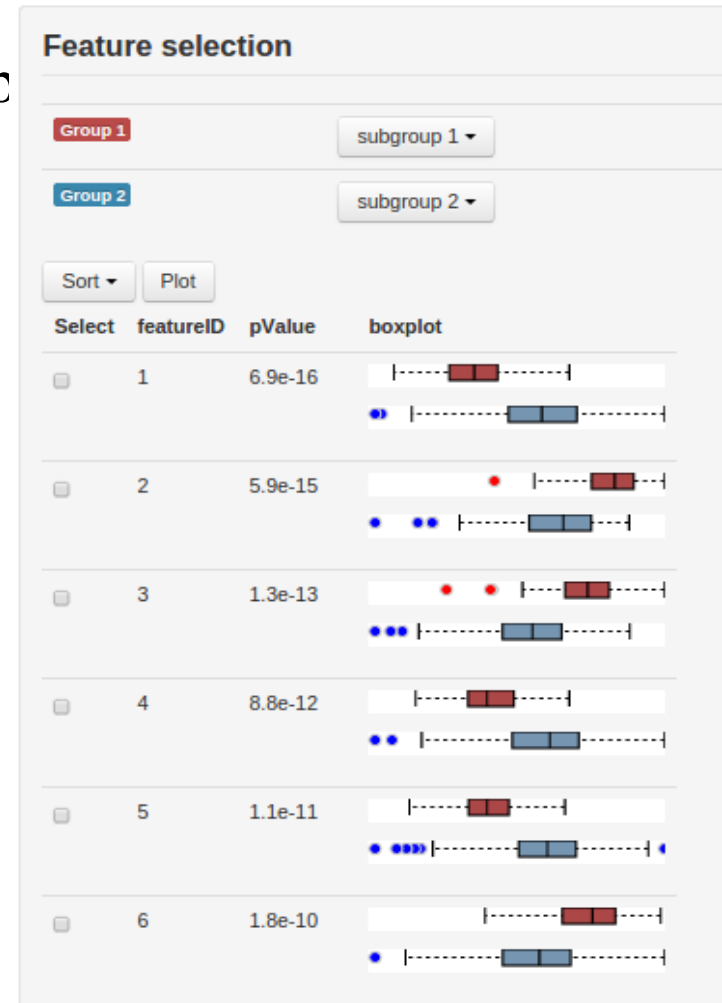Hao Ding[1], Chao Wang[3], Kun Huang[2*], Raghu Machiraju[1*]

Wexner Medical Center

Graphie –Feature Selection

Examine feature distinctiveness for group of images.

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

Aim 3 – Translational Applications

- **Triple negative breast cancer biomarker discovery (Drs. *Charles Shapiro, MSMC*)**

- **Triple negative breast cancer neoadjuvant therapy outcome prediction (Dr. *Zaibo Li - OSU, Yuan Zhang – Harbin Cancer Hospital*)**

- **Lung adenocarcinoma immune therapy response prediction (Drs. *David Carbone - OSU, Yilong Wu – Guangdong Lung Cancer Inst.*)**

- **Lung adenocarcinoma prognosis prediction and subtyping (Dr. Lin Yang – UFL, David Foran – Rutgers U)**

- **Papillary cell renal cell carcinoma (Dr. *Anil Parwane - OSU*)**

- **Clear cell renal cell carcinoma (Dr. *Anil Parwane - OSU*)**

- **Thyroid cancer SPORE (Drs. *Matt Ringel*, Kevin Coombes - *OSU*)**

- **Gastric cancer (Dr. *Jiafu Ji – Beijing Cancer Hospital*)**

- **Liver cancer (Dr. Lei Liu – Fudan University)**

- **Neuroblastoma (CAMDA)**

- **Cytology image classification for cancers (Dr. *Zaibo Li - OSU*)**

Aim 3 – Basic Science Applications

- **Cancer cell cycle pathway characterization (Dr. Gustavo Leone, MUSC)**

- **Cancer genome stability gene mutation (Dr. Jeffrey Parvin, OSU)**

- **Cancer epigenetics / CCSB U54 (Drs. Tim Huang, UTSA)**

- **Cancer drug repurposing (Drs. Philip Payne, WashU)**

- **Cancer co-methylation markers (Dr. Jie Zhang, OSU)**

- **Systems modeling of breast cancer (Dr. Lang Li, IU)**

- **Functions of STUB1 gene (Dr. Yufeng Yang, Fuzhou U)**

Wexner Medical Center

Aim 3 – Informatics Collaborations

- **Integrative analysis of cancer proteogenomics data (Dr. Parag Mallick, Stanford U)**

- **Integrative analysis and visualization of CPTAC data (Dr. Bing Zhang, Baylor College of Medicine)**

- **Integrative visualization tool development (Dr. Xing Tang, St. Jude Children's Hospital)**

- **Integrative tool evaluation (Dr. Simon Lin, Nationwide Children's Hospital)**

- **Integration with TPCA (Dr. Han Liang, MD Anderson)**

- **Morphological feature extraction (Dr. Lee Cooper, Emory U)**

Wexner Medical Center

Clear Cell Renal Cell Carcinoma

**Question –** Can these features do better when combined?

dical Center
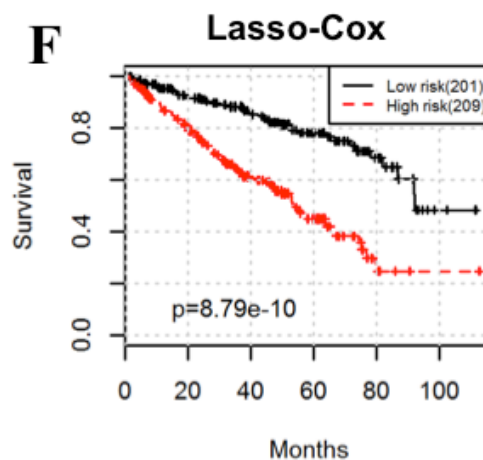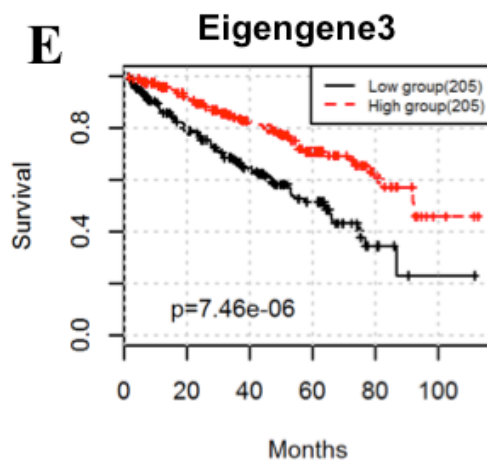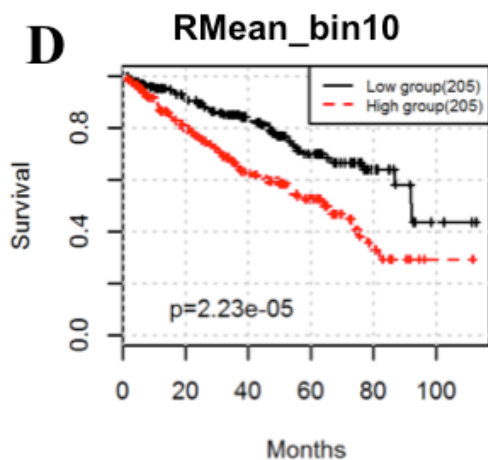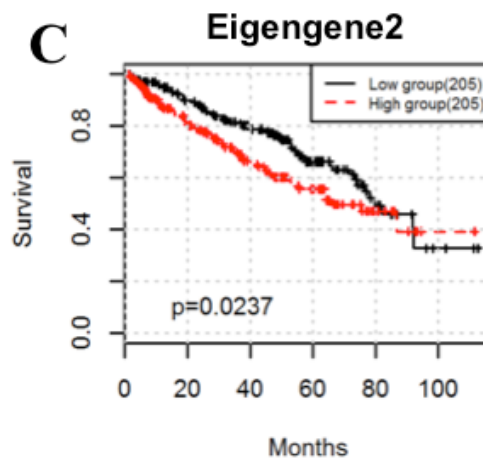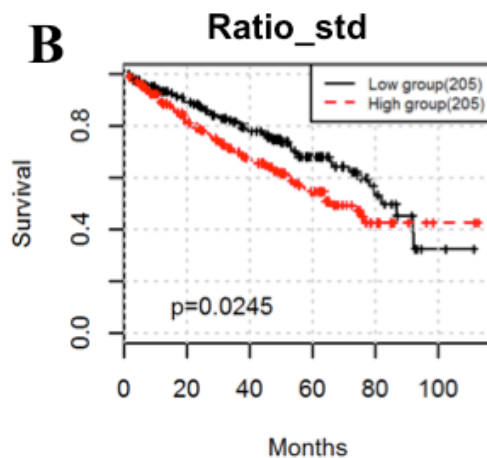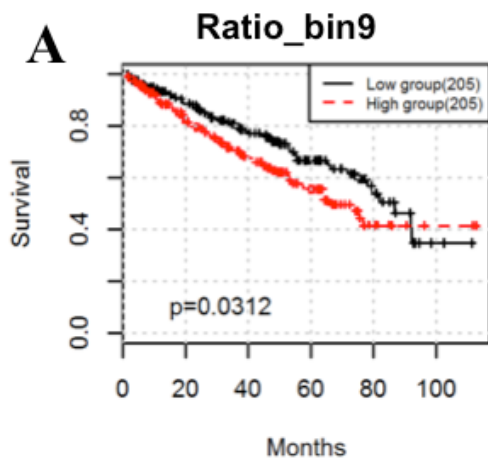
Clear Cell Renal Cell Carcinoma

Question – Can these features do better when combined?

	Univariate Cox regression		Multivariate Cox regression	
Variable	**HR (95% CI)**	**P value**	**HR (95% CI)**	**P value**
Lasso-Cox	3.06 (2.10-4.45)	5.02e-9	2.26 (1.46-3.49)	2.31e-4
Clinical				
Grade	2.38 (1.63-3.5)	8.45e-6	1.46 (0.95-2.23)	8.22e-2
Stage	3.68 (2.57-5.27)	1.12e-12	3.00 (2.00-4.49)	9.23e-8
Gene expression				
CSNK2A1	0.90 (0.64-1.26)	5.34e-1	1.07 (0.74-1.56)	7.11e-1
SPP1	1.15 (0.82-1.61)	4.14e-1	1.10 (0.75-1.63)	6.20e-1
DEFB1	1.41 (1.00-1.98)	4.99e-2	1.36 (0.95-1.95)	9.71e-2
PECAM1	0.77 (0.55-1.09)	1.40e-1	1.04 (0.69-1.58)	8.45e-1
EDNRB	0.50 (0.35-0.71)	9.10e-5	0.96 (0.59-1.57)	8.77e-1
TSPAN7	0.54 (0.38-0.76)	5.12e-4	1.03 (0.64-1.67)	9.07e-1
Somatic mutation				
VHL	0.99 (0.70-1.38)	9.33e-1	1.23 (0.86-1.75)	2.57e-1
PBRM1	0.85 (0.58-1.24)	3.94e-1	1.03 (0.69-1.54)	8.85e-1
BAP1	1.49 (0.78-2.85)	2.22e-1	1.49 (0.74-3.00)	2.60e-1
SETD2	1.29 (0.77-2.14)	3.29e-1	1.03 (0.62-1.74)	9.00e-1
TP53	2.26 (1.00-5.15)	5.13e-2	2.86 (1.19-6.86)	1.85e-2

Medical Center

Clear Cell Renal Cell Carcinoma

Question – What features are correlated with patient outcome in early stage (stages I, II)?
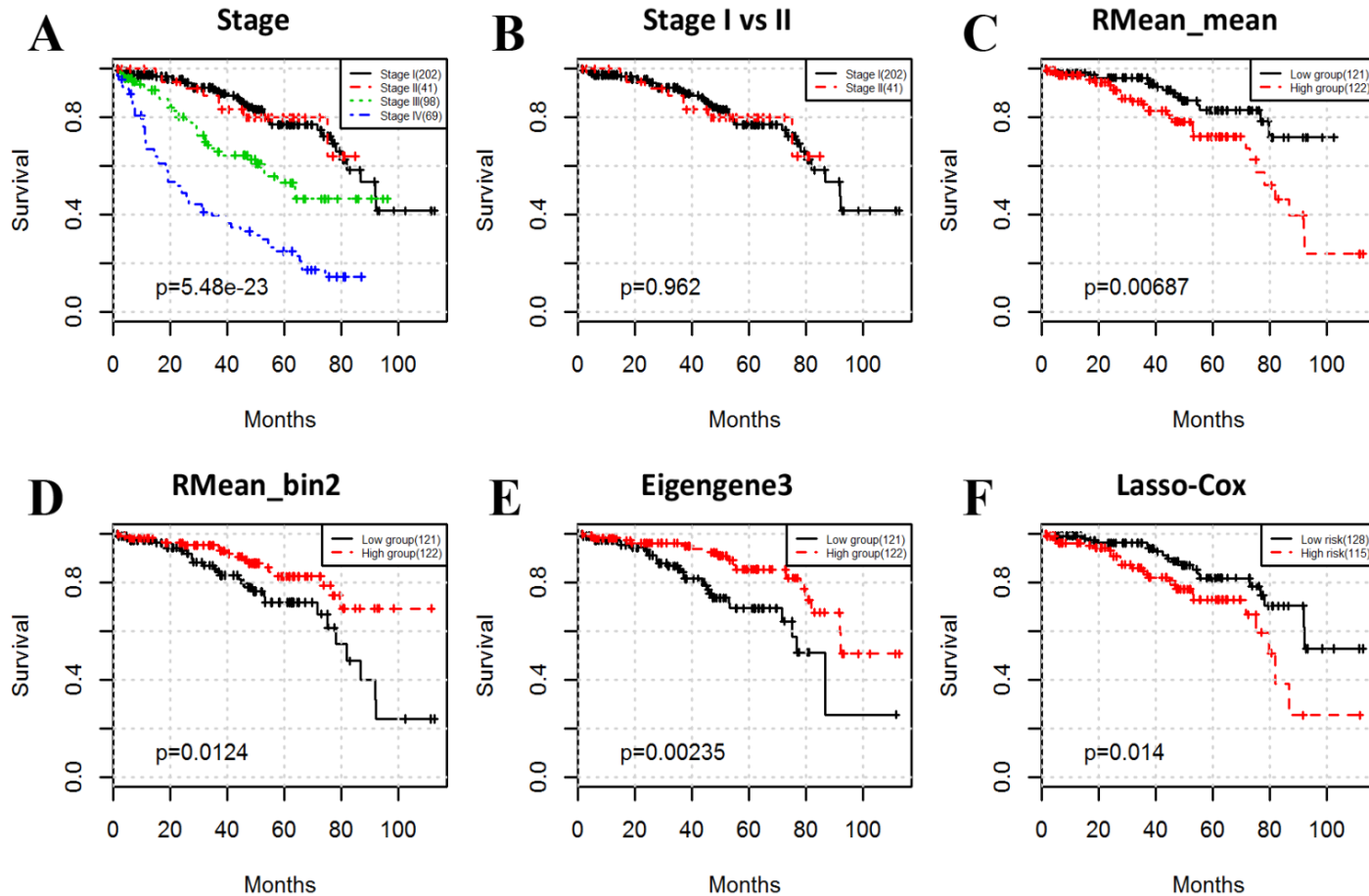
Features and Workflow Released on GitHub

All processed data for clear cell renal cell cancer project (extracted quantitative imaging features, combined gene expression data, etc) and codes with annotations, comments and instructions are available at https://github.com/chengjun583/image-mRNA-prognostic-model.

Wexner Medical Center

Next

- **Multiple releases of web portal and tools**

- **Extensive user evaluations**

- **Outreach - CCSB, ITCR, AMIA, AACR, PSB, ICIBM, …**

- **Interpretation and validation of discoveries from applications**