重慶醫科大學
CHONGQING MEDICAL UNIVERSITY

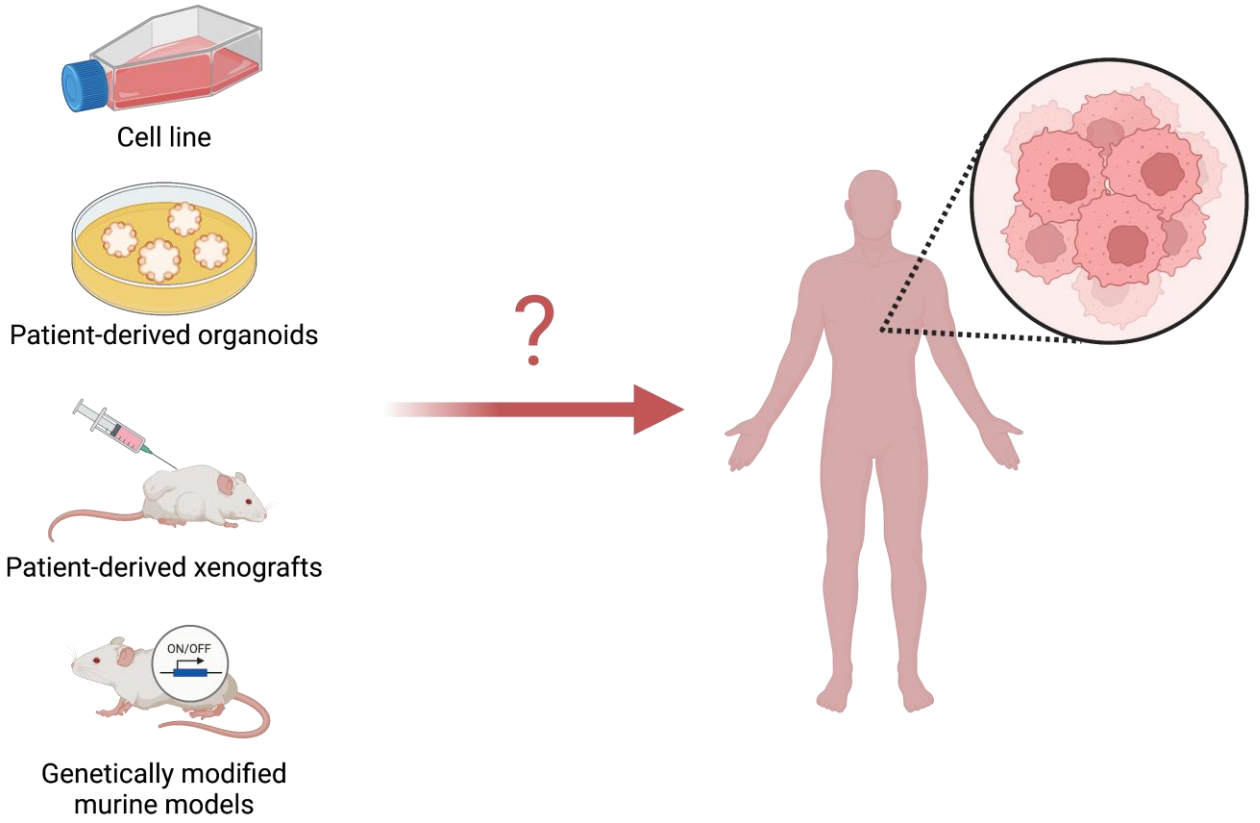
Systems approach for congruence and selection of cancer models towards precision medicine

Jian Zou, Ph.D.
Department of Statistics, School of Public Health
Chongqing Medical University
jianzou@cqmu.edu.cn

2023-10-16

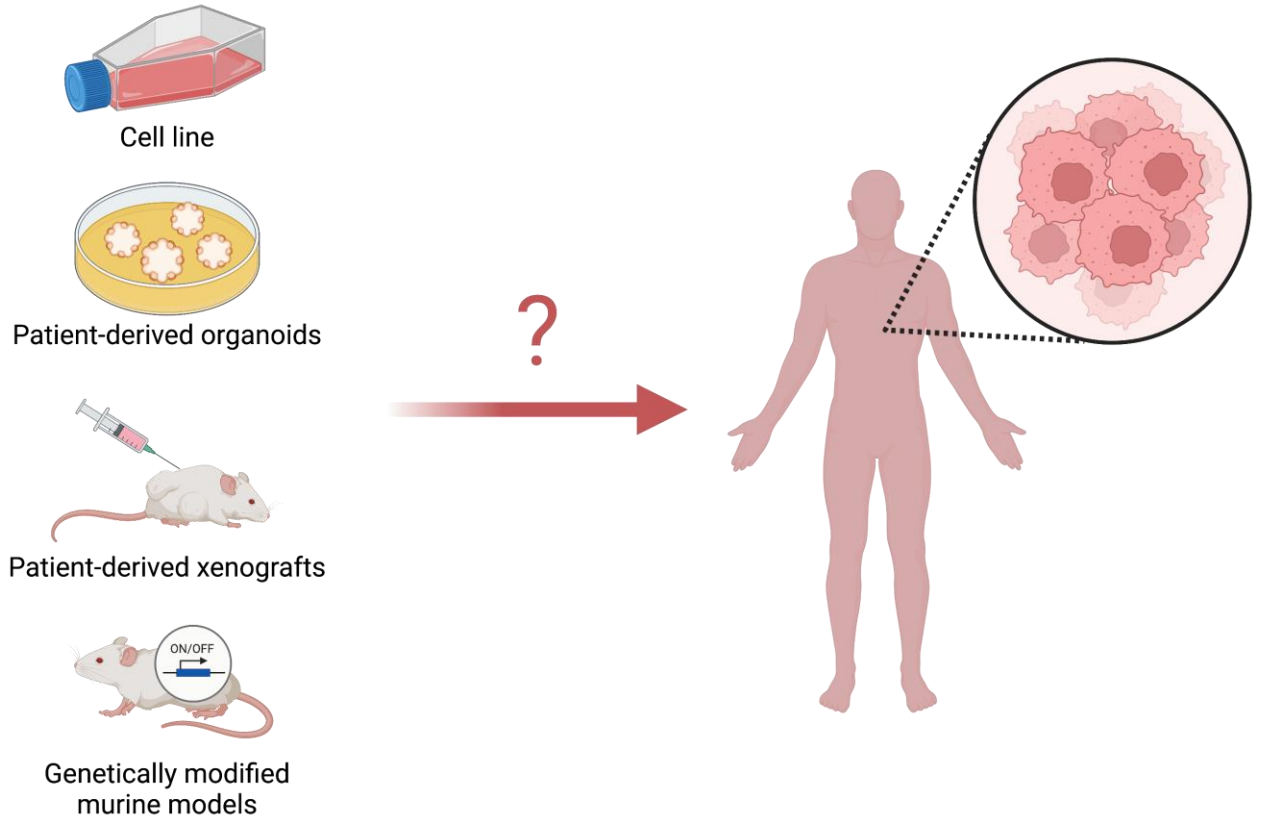
Background

- **Cancer models** are essential tools in cancer research for exploring carcinogenesis and developing drugs in translational and clinical studies.
- Evaluation and comparison of **cancer models** with human tumors have drawn increasing attention in recent years.
- Existing approaches:
  - Congruence (correlation) analysis
  - Authentication (machine learning) analysis

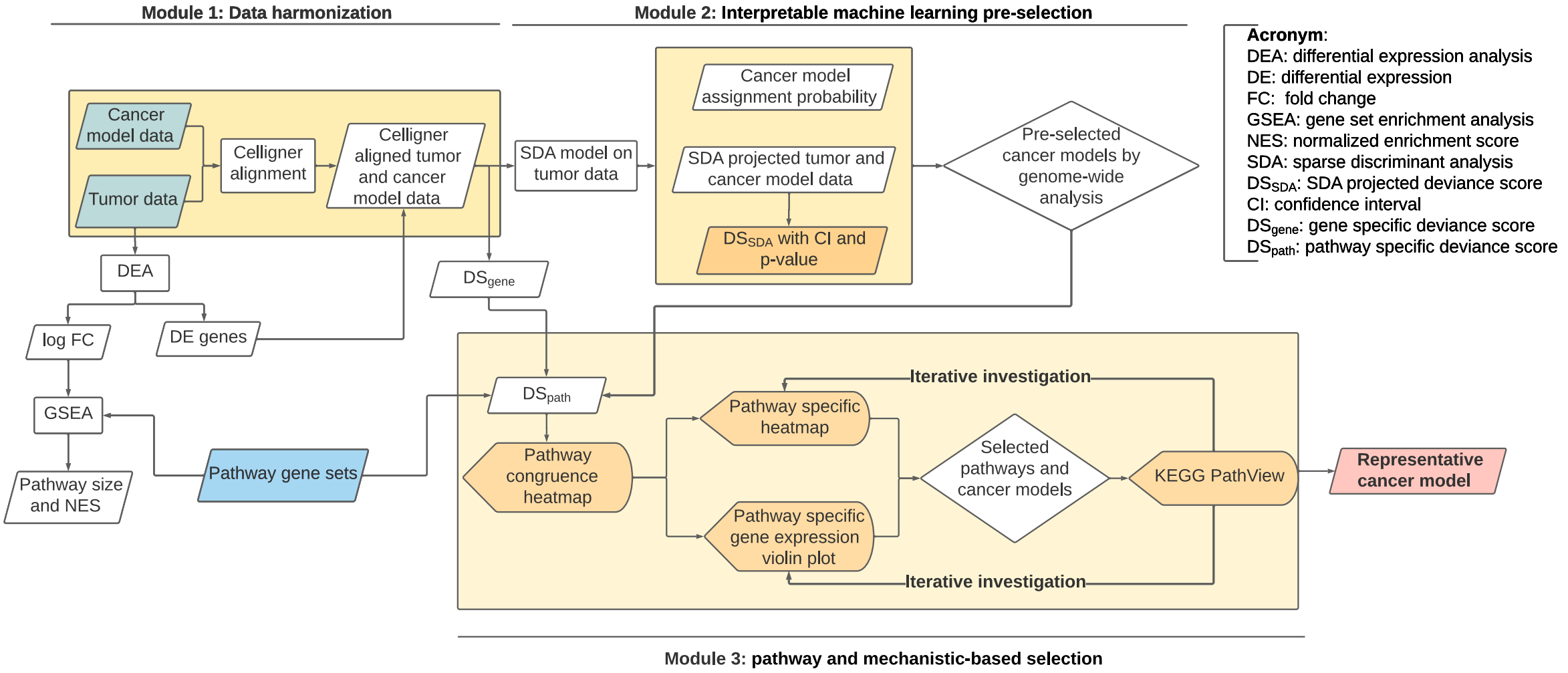


Challenges

- **Congruence** analysis provides low prediction accuracy.
- **Authentication** analysis cannot prioritize the cancer models.
- Data harmonization between human tumors and cancer models are seldomly considered.
- Current studies are limited to the genome-wide analysis without any pathway-based evaluations.



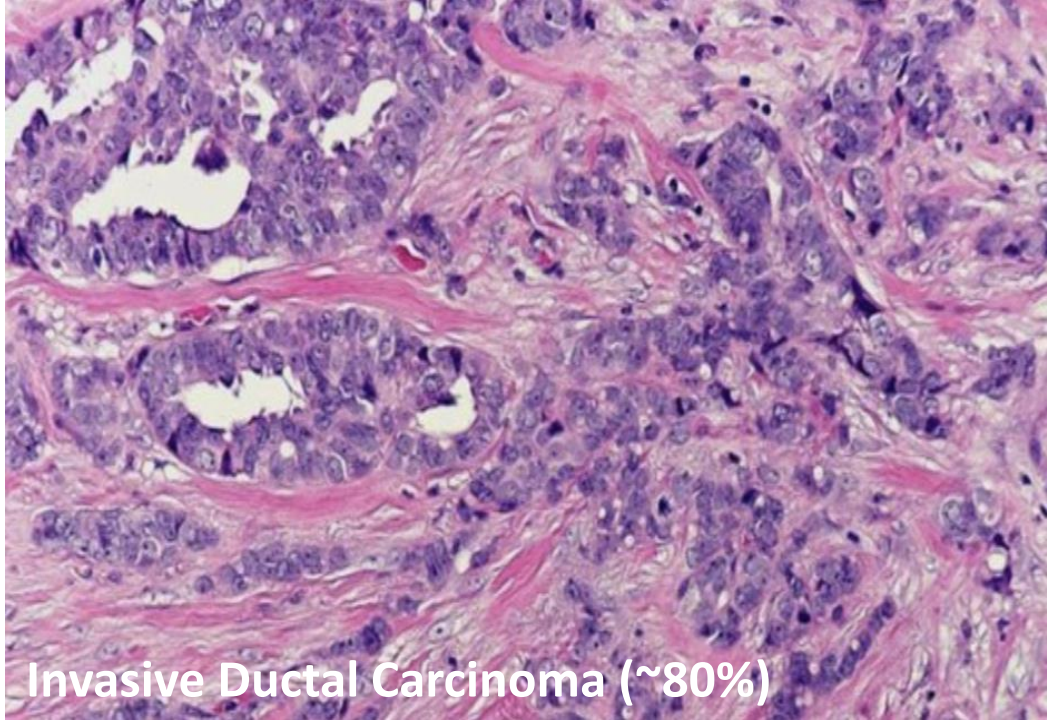
Congruence Analysis and Selector of CAncer Models (CASCAM)

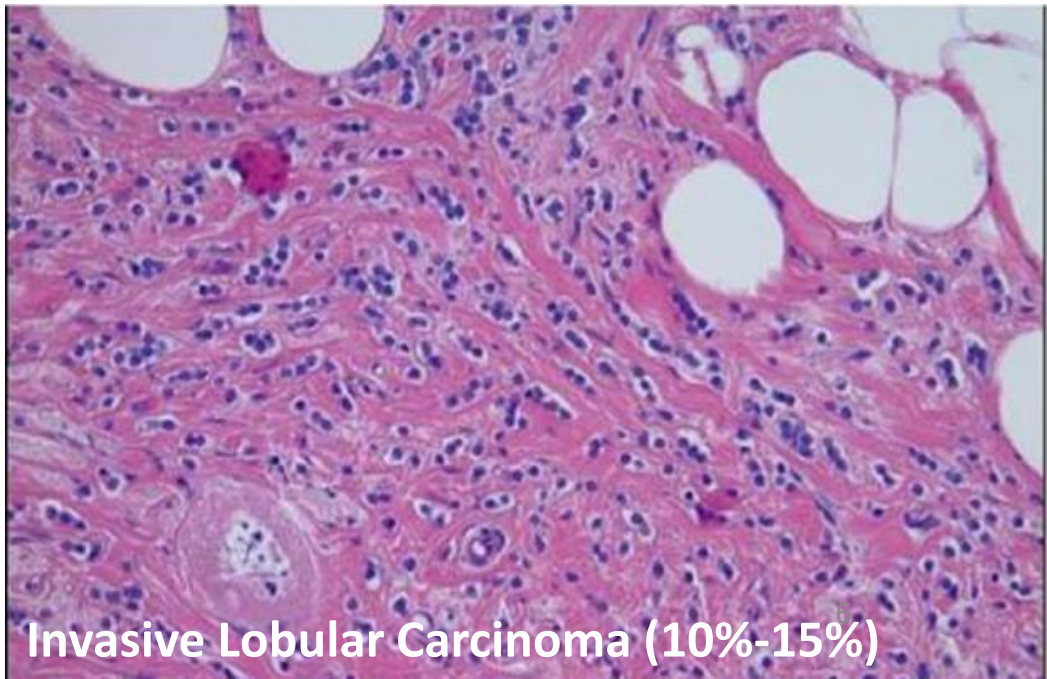


Case Study: Identify one representative cell line for the specific histological subtype in breast cancer

Data source

- We focus on two histological subtypes in breast cancer (BC).
- 960 BC patient samples from TCGA and 65 BC cell lines from CCLE are recruited for analysis.

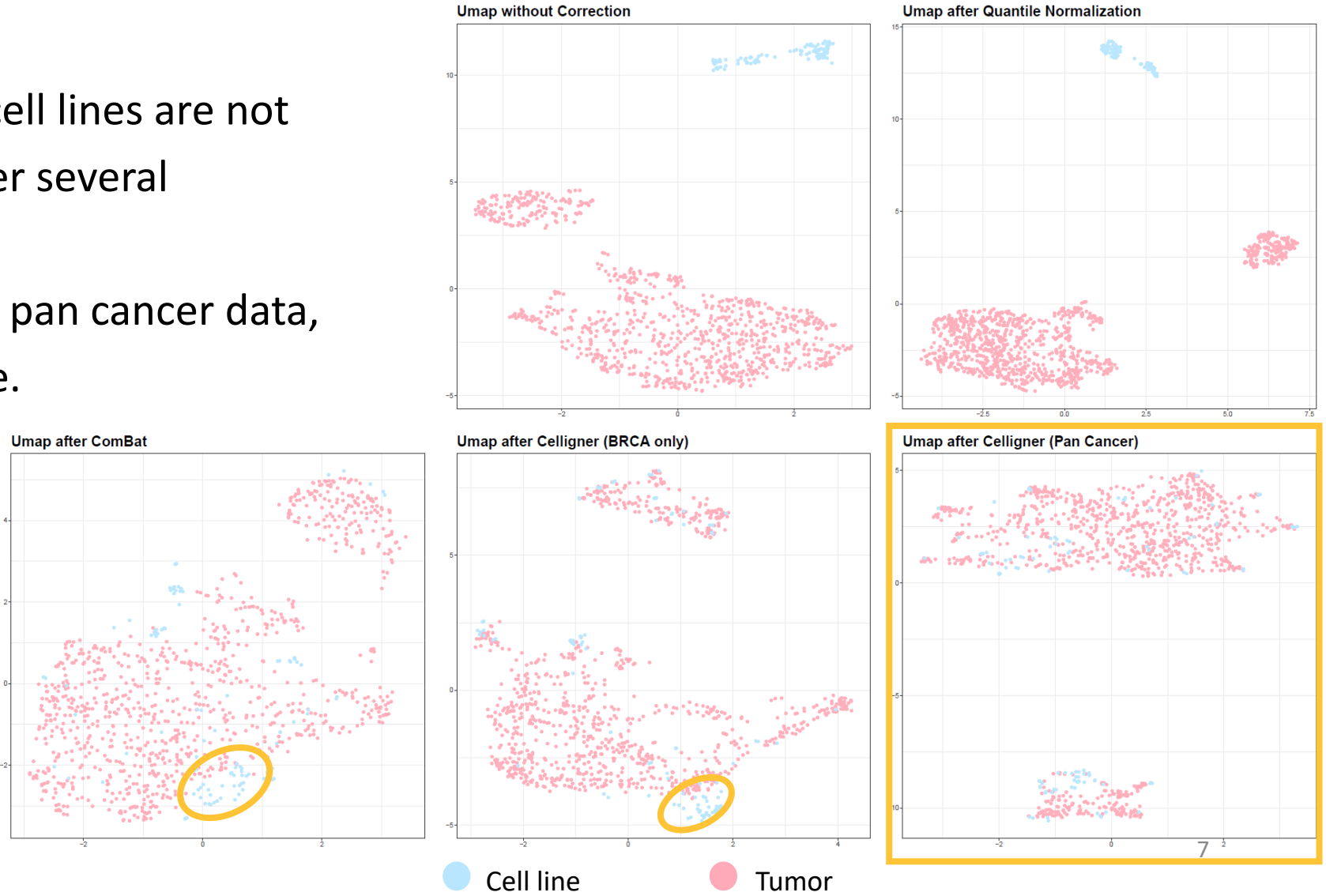




Seminars in Oncology 46 (2019) 121–132
https://pathology.jhu.edu/breast/types-of-breast-cancer

Module 1: Data harmonization

- The RNA-Seq of tumors and cell lines are not directly comparable even after several normalizations.
- After applying *Celligner* using pan cancer data, we can find them comparable.



Module 1: Data harmonization

- The cells from the same origin are gathered, and the basal group is separate from the others.
- The non-basal tumors and cells for the downstream analysis.

Module 2: Interpretable machine learning pre-selection

	Machine learning evaluation			Machine learning relevant properties		
	ILC vs IDC	ER+ vs ER-	BRCA vs other cancers	Gene selection	Assignment probability	Deviance score
	TCGA; 5-fold CV	Training data: TCGA; Test data: CCLE	Training data: TCGA; Test data: CCLE			
SDA	0.91 (0.02)	0.91	0.86	Yes	Yes	Yes
ElasticNet	0.90 (0.03)	0.93	0.85	Yes	Yes	No
2D-Hybrid-CNN	0.87 (0.03)	0.93	0.86	No	No	No
RidgeRegress*	0.88 (0.02)	0.91	0.84	Yes	Yes	No
Pearson25*	0.86 (0.01)	0.86	0.90	No	No	No
KNN	0.85 (0.03)	0.86	0.91	No	Yes	No
2D-Vanilla-CNN	0.86 (0.04)	0.88	0.85	No	No	No
1D-CNN	0.86 (0.03)	0.86	0.86	No	No	No
RandomForest*	0.85 (0.01)	0.91	0.82	Yes	Yes	No
RSLDA	0.81 (0.11)	0.77	0.86	Yes	Yes	Yes
CancerCellNet*	0.79 (0.03)	0.82	0.79	Yes	Yes	No
LDA	0.80 (0.03)	0.68	0.82	No	Yes	Yes
NTP	0.61 (0.03)	0.86	0.82	No	No	Yes
SpearmanMed*	0.40 (0.03)	0.84	0.61	No	No	Yes
PearsonMed*	0.38 (0.04)	0.84	0.62	No	No	Yes
Logistic	0.52 (0.04)	0.43	0.65	No	Yes	No

Module 2: Interpretable machine learning pre-selection

- On the genome-wide, each cell line and tumor sample is projected to the same space through SDA.
- The ***SDA-based deviance score***, $DS_{SDA}^{i,k}$ for cancer model *i* in subtype *k* is defined as

$$DS_{SDA}^{(i,k)} = |c_i - \hat{\mu}_k| / \hat{\sigma}$$

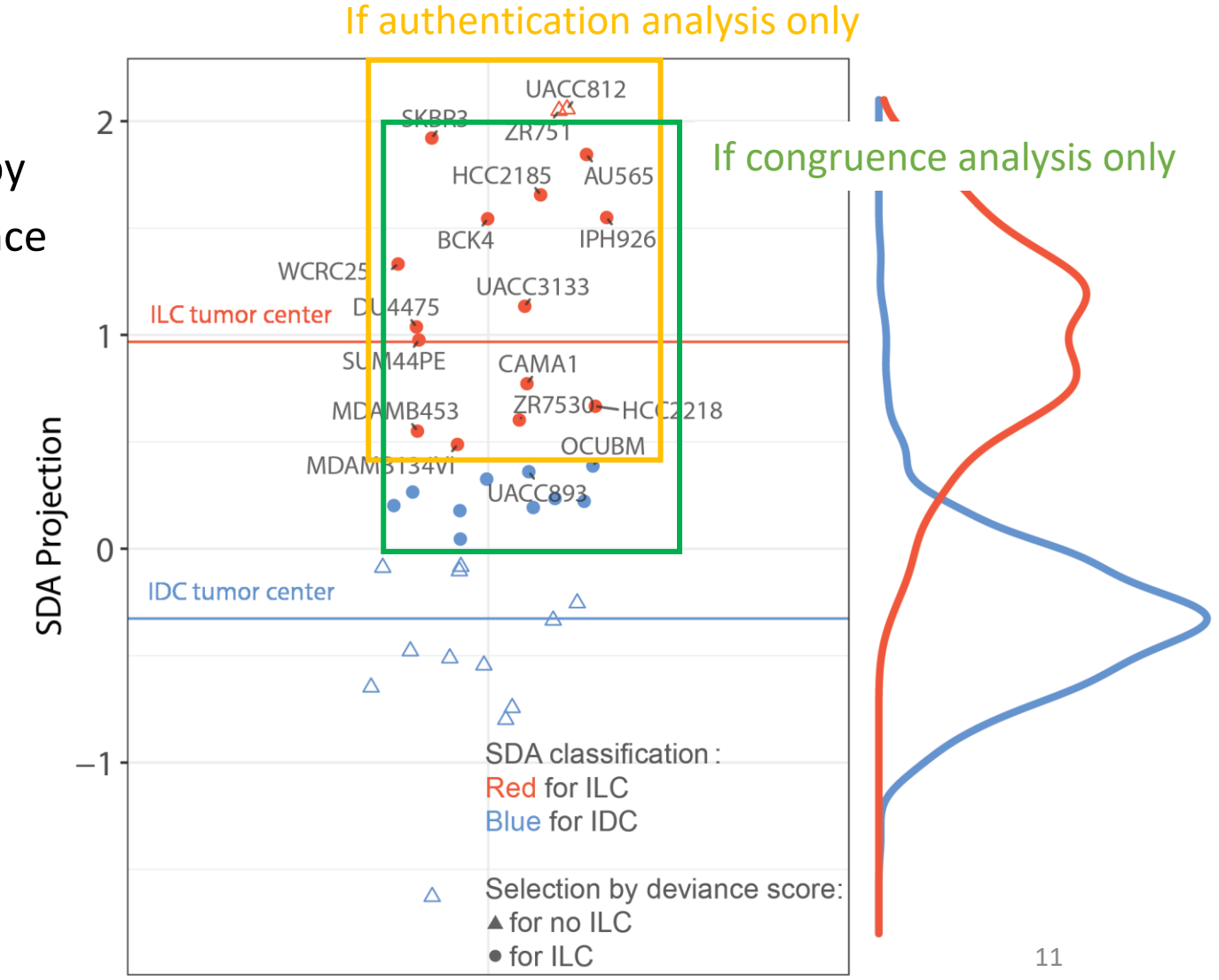
  where $\hat{\mu}_k$ and $\hat{\sigma}$ are the estimated robustized tumor subtype center and standard deviation.
- $pval(DS_{SDA})$ is obtained from the null distribution constructed by tumor samples.
- Assignment probability is denoted as $P_{SDA}^{(i,k)}$.

- DS_{SDA} is for congruence (correlation) analysis.
- P_{SDA} is for authentication (machine learning) analysis.

Module 2: Interpretable machine learning pre-selection

- Red circles are the one classified as ILC cell line by the combination of SDA classification and deviance score.
- $0.025 < pval(DS_{SDA}) < 0.975$ and $P_{SDA} > 0.5$ is used as ILC criteria.

- 14 cell lines are selected for downstream investigation.



Module 3: Pathway and mechanistic-based selection

- The ***gene specific deviance score*** (DS_{gene}) for cell *i* for class *k* in gene *g* is defined as

$$DS_{gene}^{(g,i,k)} = |c_{g,i} - \hat{\mu}_{g,k}| / \hat{\sigma}_g$$

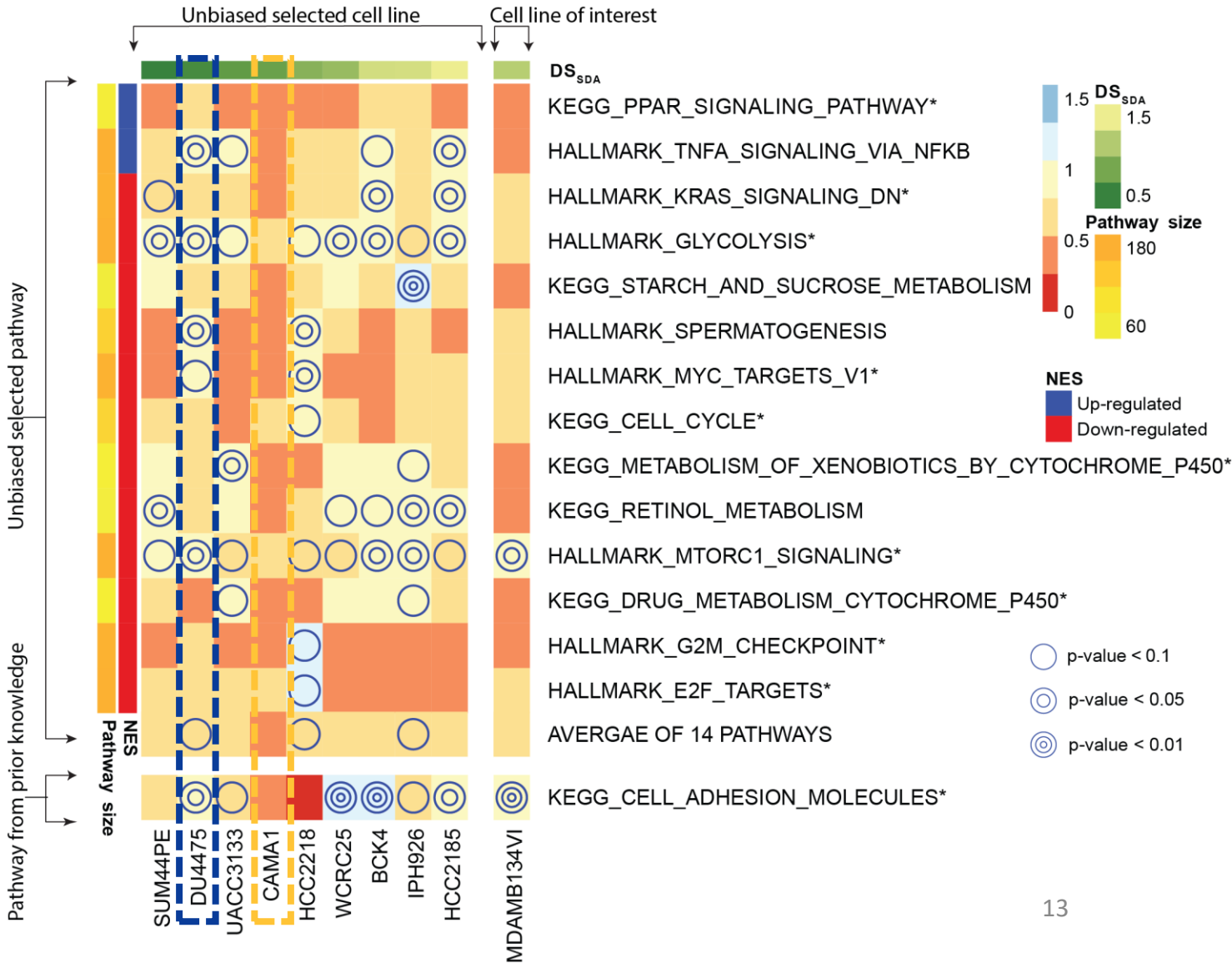
  where $\hat{\mu}_{g,k}$ and $\hat{\sigma}_g$ are the estimated robustized tumor subtype center and pooled standard deviation.

- The ***pathway specific deviance score*** for cell *i* for class *k* in pathway *p* is defined as

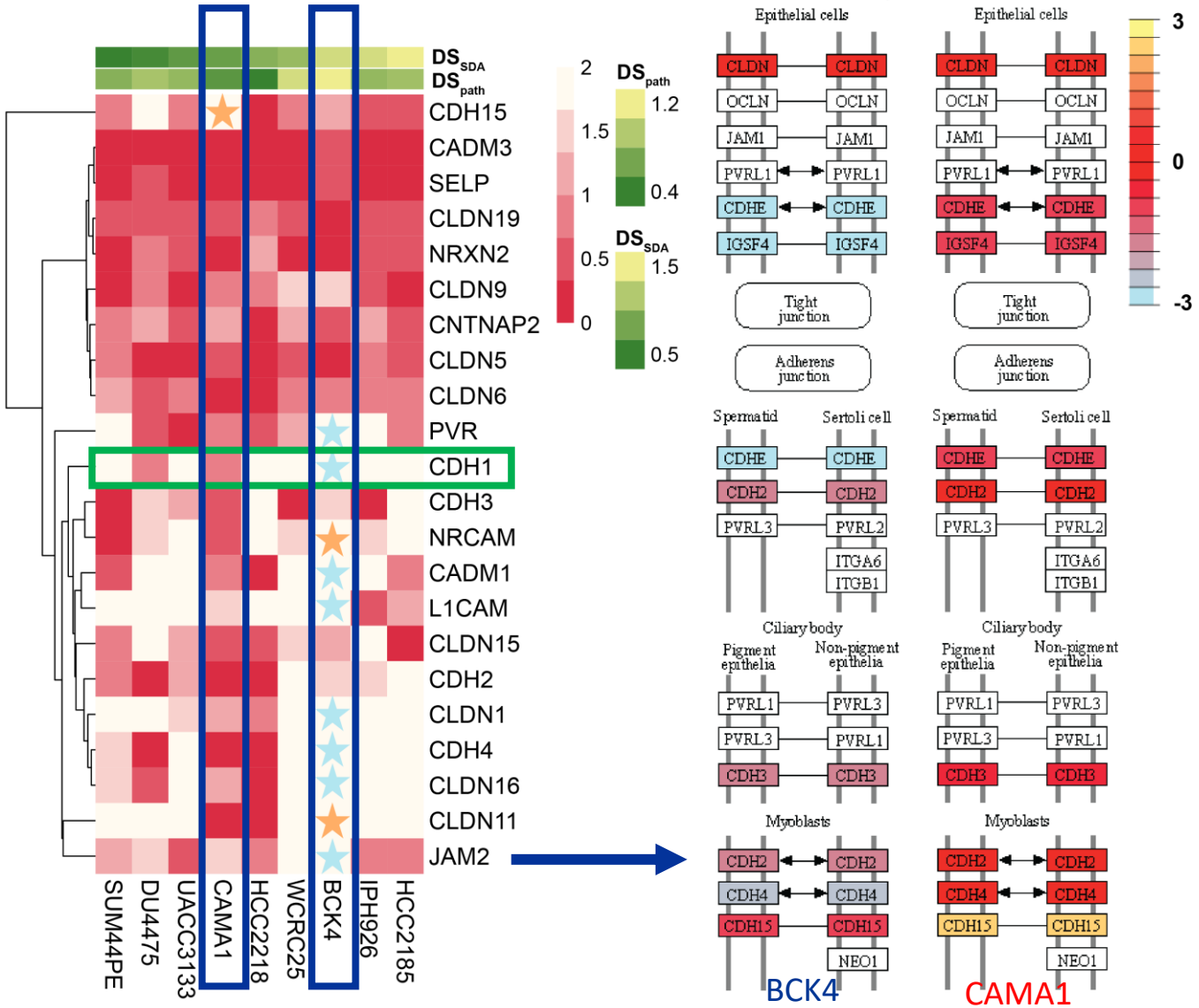
$$DS_{path}^{(p,i,k)} = geometic\ mean_{g \in p}(|DS_{gene}^{(g,i,k)}|)$$

Module 3: Pathway and mechanistic-based selection

- Pathways with # DE > 20, 30 < size < 200, and |NES| > 1.5 are selected.
- CAMA1 has the best averaged DS_{path} though it is not the genome-wide best performer.
- DU4475 has relative worse performance among the genome-wide top 5 models.



Module 3: Pathway and mechanistic-based selection



- *CDH1* is the hallmark of ILC and affects the expression of E-cadherin and dysfunction the cell adhesion.
- We further explore *KEGG Cell Adhesion Molecules* pathway.

CAMA1 is the second-best performer, and BCK4 is the worst.

Conclusion

- CASCAM provide a complete framework for authenticating and selecting the most representative cancer models.
- The heterogeneity exists among different cell lines, even though they are all identified as the same tumor subtype on the genome-wide. (e.g. BCK4 vs. CAMA1)
- CAMA1 is overall the best representative cell line for ILC.

Acknowledgement

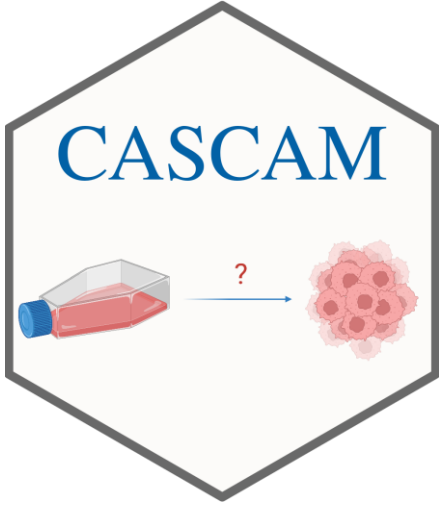
Advisors:

- Dr. George C. Tseng, University of Pittsburgh
- Dr. Steffi Oesterreich (co-advisor), University of Pittsburgh Medical Center
- Dr. Adrian V. Lee (co-advisor), University of Pittsburgh Medical Center

Collaborators:

- Dr. Osama Shah, University of Pittsburgh Medical Center
- Dr. Yu-Chiao Chiu, University of Pittsburgh Medical Center
- Dr. Tianzhou Ma, University of Maryland

Summary



Challenges	Solutions
Data harmonization between human tumors and cancer models are seldomly considered	Celligner is used in this study for data preprocessing
Congruence analysis provides low prediction accuracy	DS_{SDA} is proposed to measure the absolute distance towards the interested tumor subtype center and used for cell line ranking
Authentication analysis cannot prioritize the cancer models	
Current studies are limited to the genome-wide analysis without any pathway-based evaluations	DS_{path} and the related visualization tools are developed for pathway specific cell line selection