# Biomarker Discovery: Computational Approaches

Seungchan Kim

Center for Computational Systems Biology

Prairie View A&M University

CCSB@PVAMU [https://ccsb.pvamu.edu]

BCAN
Bladder Cancer Advocacy Network

https://ccsb.pvamu.edu/

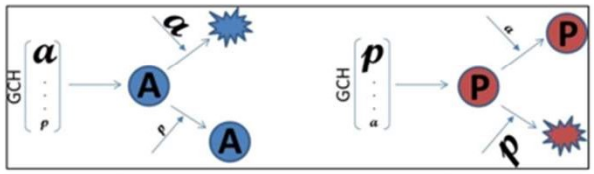


# Welcome to CCSB@PVAMU

The Center for Computational Systems Biology at the Prairie View A&M University (CCSB@PVAMU) was established in 2017, by a funding from the Texas A&M University Systems Chancellor's Research Initiative (CRI) to become a nationally recognized computational systems biology research center.

As a newly established center, we are rapidly developing research programs to study 1) biomedical questions of high translational significance and impact and 2) challenging questions in plant and agricultural science, utilizing various genomics and computational tools. The center aims to achieve:

- Recruitment of high-quality faculty and research scientists
- Promoting interdisciplinary research across Roy G. Perry College of Engineering, College of Agriculture and Human Sciences, College of Arts and Sciences, and other research centers/labs within PVAMU campus
- Development of active collaboration with leading biomedical research institutes as well as bioinformatics research centers
- Development of training programs for bioinformatics and computational systems biology at both the undergraduate and the graduate levels.




Iacobas et al. 25 July 2019. Genes 10(8):560.

## Recent posts

2019 IEEE BIBM Workshop - Single Cell -Omics: Challenges and

[DOD-CDMRP Translational Team Science Award] Development of

[CCSB Seminar Series] Radiation Dosimetry Measurements Onboard Air

[CCSB Seminar Series] Single Cell Genomics at the Center for

# CCSB Collaboration: SU2C

## Transatlantic Pancreatic Cancer Dream Team


Atlanta 2019

# What is a Biomarker?

# Biomarkers can be

- Pulse
- Blood pressure
- Basic chemistries or more complex laboratory tests of blood and other tissues

# Molecular Biomarkers

- Gene mutations
- Single Nucleotide Polymorphisms
- Gene expressions
- Small molecules such as miRNAs

- Emerging from analysis of high throughput molecular measurements such as DNA and RNA sequencing data

# What is a Biomarker?

- NIH [1998] – a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention
- WHO [2001] – any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease, or
- WHO [2001] – almost any measurement reflecting an interaction between a biological system and a potential hazard, which may be chemical, physical, or biological.

*Curr Opin HIV AIDS.* 2010 November ; 5(6): 463–466

# What is a Biomarker?

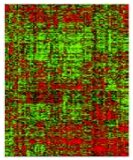
- NIH [1998] – a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention
- WHO [2001] – any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease, or
- WHO [2001] – almost any measurement reflecting an interaction between a biological system and a potential hazard, which may be chemical, physical, or biological. The measured response may be functional and physiological, biochemical at the cellular level, or a molecular interaction.

*Curr Opin HIV AIDS.* 2010 November ; 5(6): 463–466

# Biomarkers

- **classification and prediction,**
- as surrogate outcomes in clinical trials,
- as measures of toxic or preventive exposures, or
- as a guide to individual treatment choice

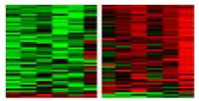
# Biomarker Discovery: Workflow



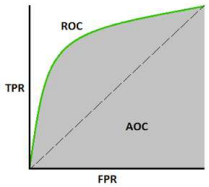
Raw data

Preprocessing

ctrl vs. case

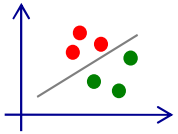


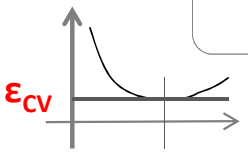
Feature selection




Internal validation

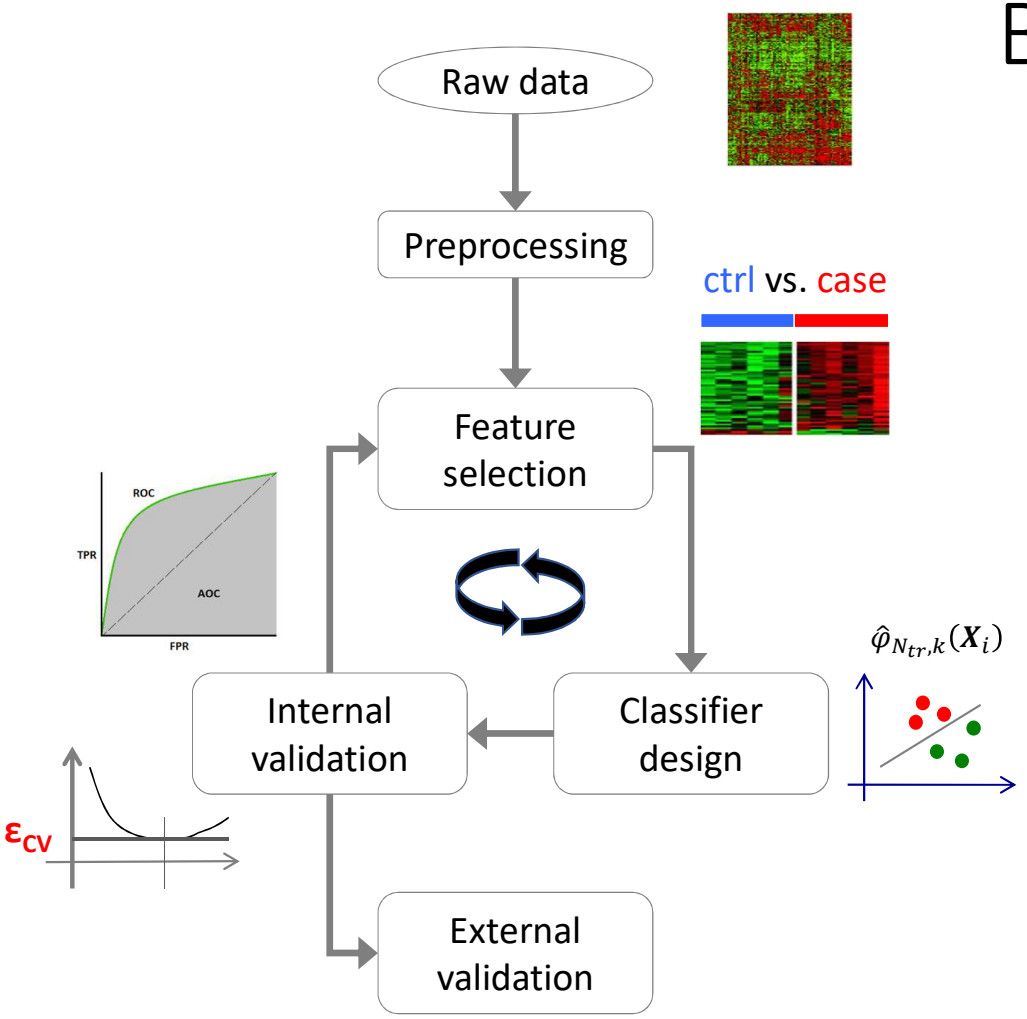
Classifier design






NCI EDRN Guideline
*JNCI*, Vol. 93, No. 14, July 18, 2001

External validation → Clinical Assay & Validation → Retrospective longitudinal → Prospective screening → Cancer Control



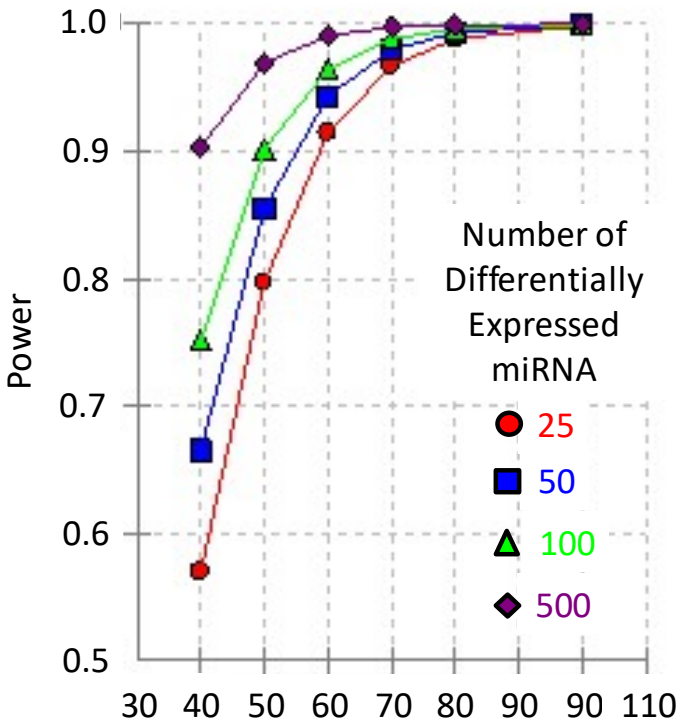

# Bottle necks/Challenges

- Sample size
  - How to estimate statistical power
- Preprocessing
  - Normalization
  - Batch correction
- Internal validation
  - Statistical validation
  - Error estimation

# Sample Size

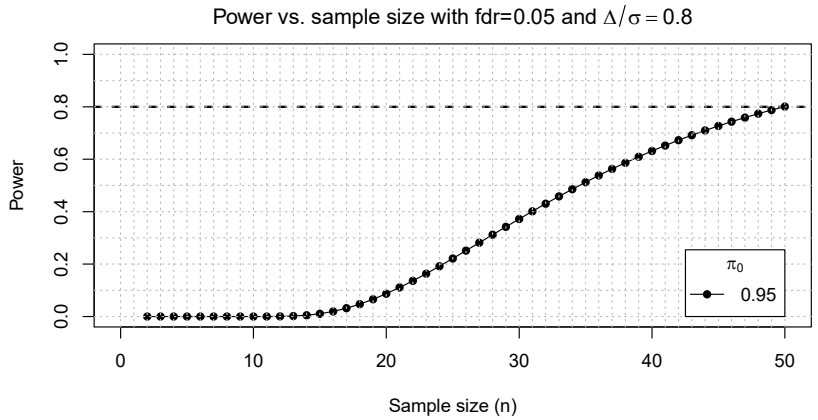
- Is the study sufficiently statistically powered to test hypothesis?
- Traditional method to estimate sample size is often based on a simple, irrelevant hypothesis test
  - Effect size – often arbitrarily chosen
- Simulation-based approach is more desirable, where
  - Synthetic data are generated from a plausible data model,
  - Classifier design and evaluation should performed with varying size of cases and controls to determine sample size

# Sample Size

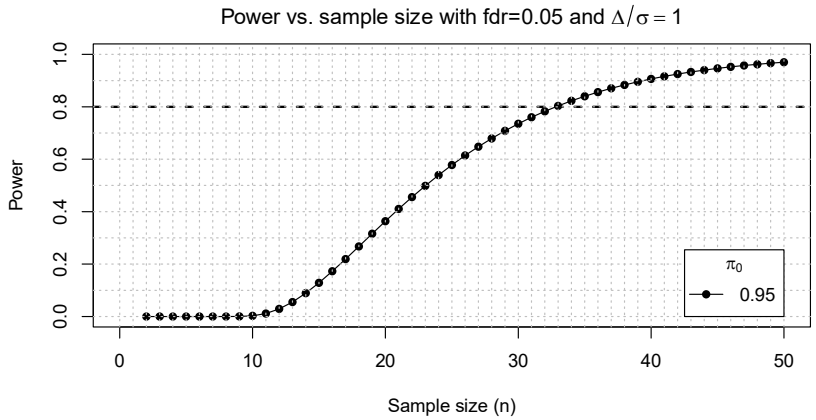



- The study is **adequately powered (>0.9)** to detect a **2 fold difference** in as few as **25 miRNAs out of 7,000** while adjusting for multiple tests using an **FDR correction of 0.05**

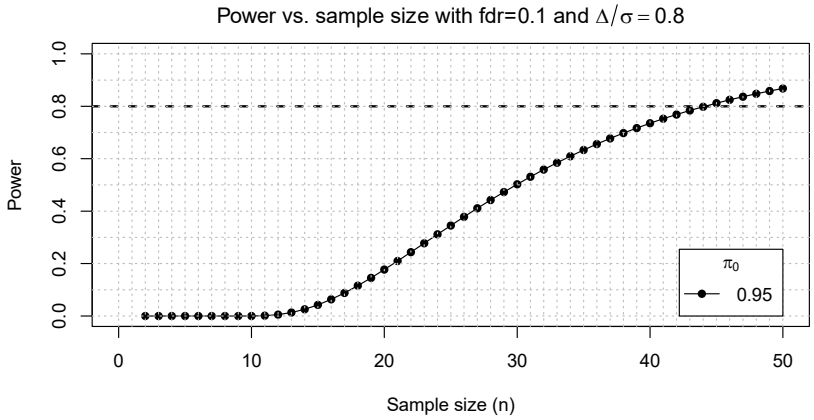
# Sample Size: effect size vs. FDR



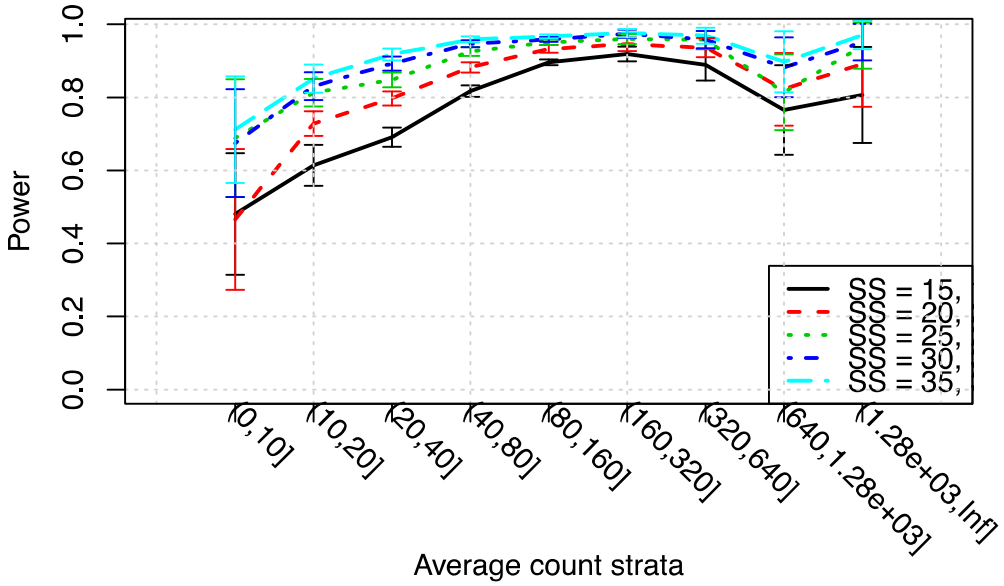

effect size

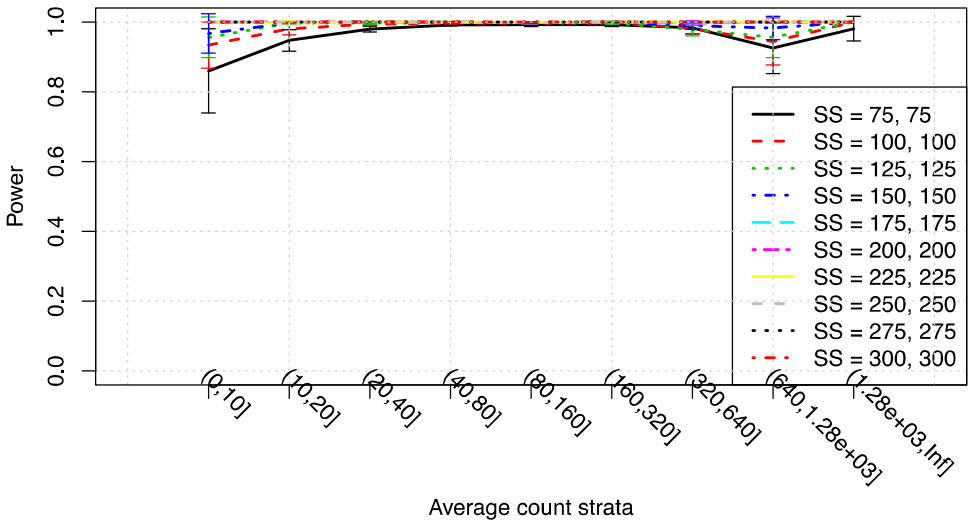



FDR



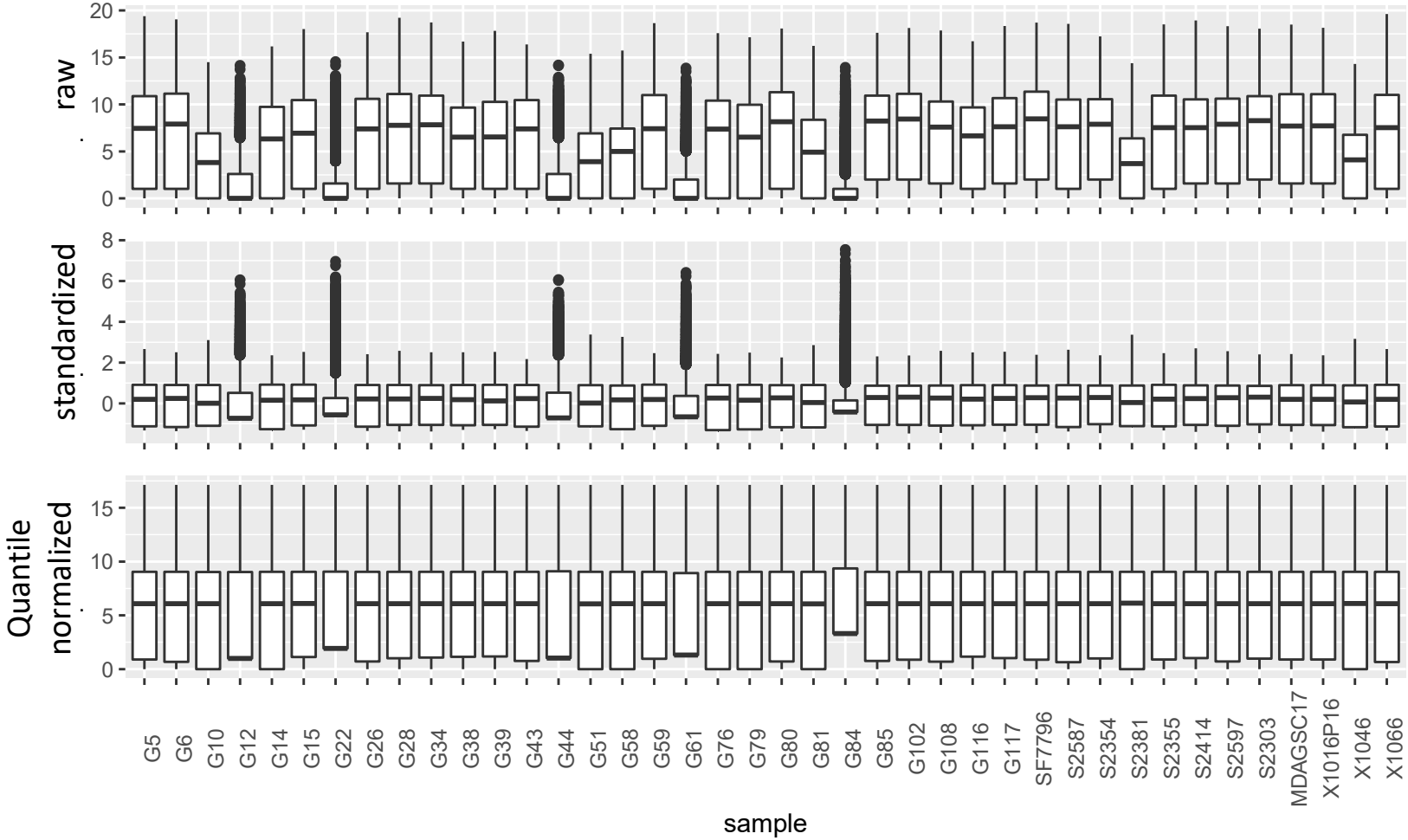
# Sample Size: simulation-based



# Normalization

- Compensate for technical and/or biological covariates such as (in RNAseq):
  - Sequencing depth
  - Transcript length
- Also, done to transform the measurements so that they fit into a mathematical model – often called standardization
- Aggregate vs single-sample normalization


raw
standardized
Quantile normalized
20
15
10
5
0
8
6
4
2
0
15
10
5
0
G5
G6
G10
G12
G14
G15
G22
G26
G28
G34
G38
G39
G43
G44
G51
G58
G59
G61
G76
G79
G80
G81
G84
G85
G102
G108
G116
G117
SF7796
S2587
S2354
S2381
S2355
S2414
S2597
S2303
MDAGSC17
X1016P16
X1046
X1066
sample

# Batch correction

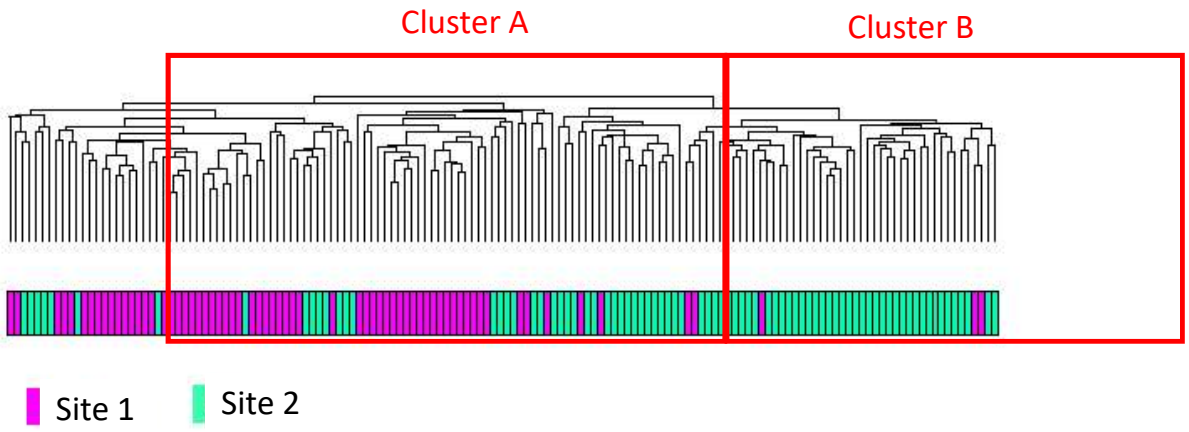
- Minimize systematic, undesired difference in measurements between batches of samples
  - Difference between different sites where the samples are collected
  - Difference between different times when the samples are collected

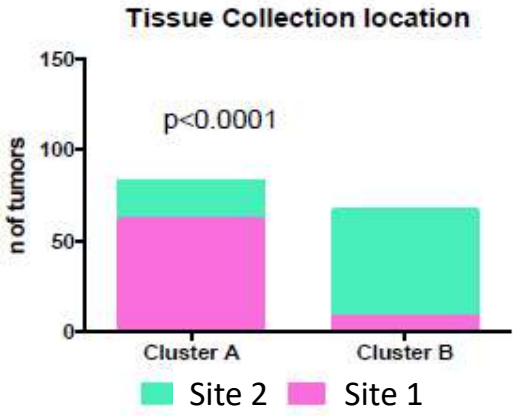
  - Standardized protocols to collect samples should mitigate this issue, but …

- Some of the conditions are not controllable.

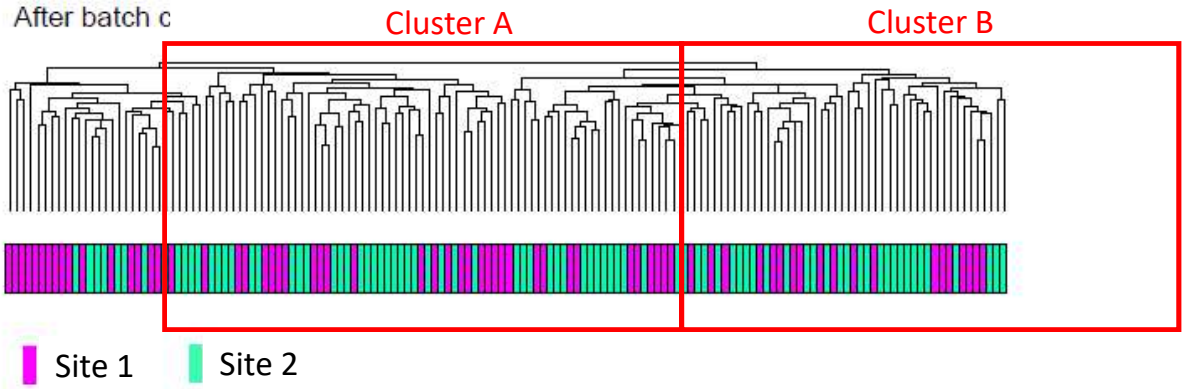
# Batch: Multiple sites

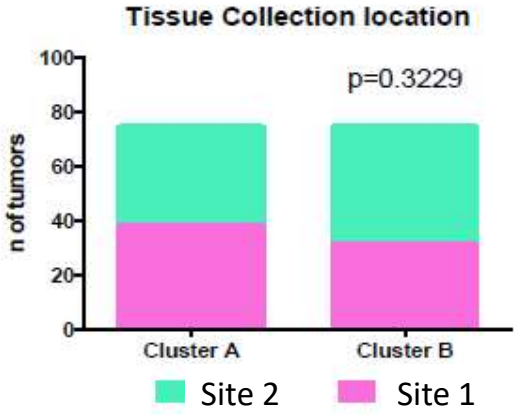
Before Batch Correction



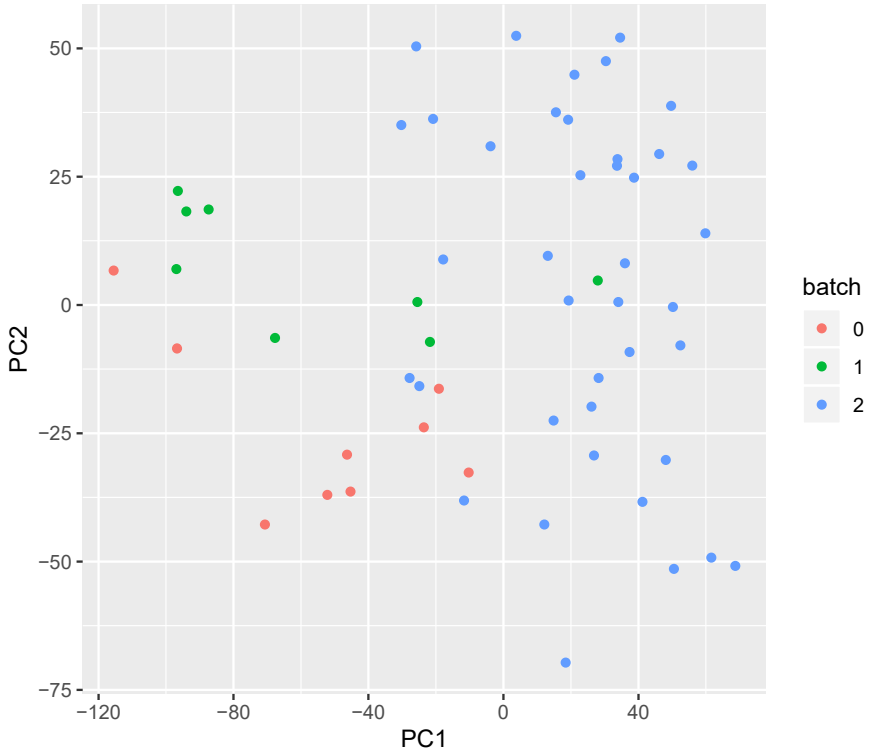

After Batch Correction



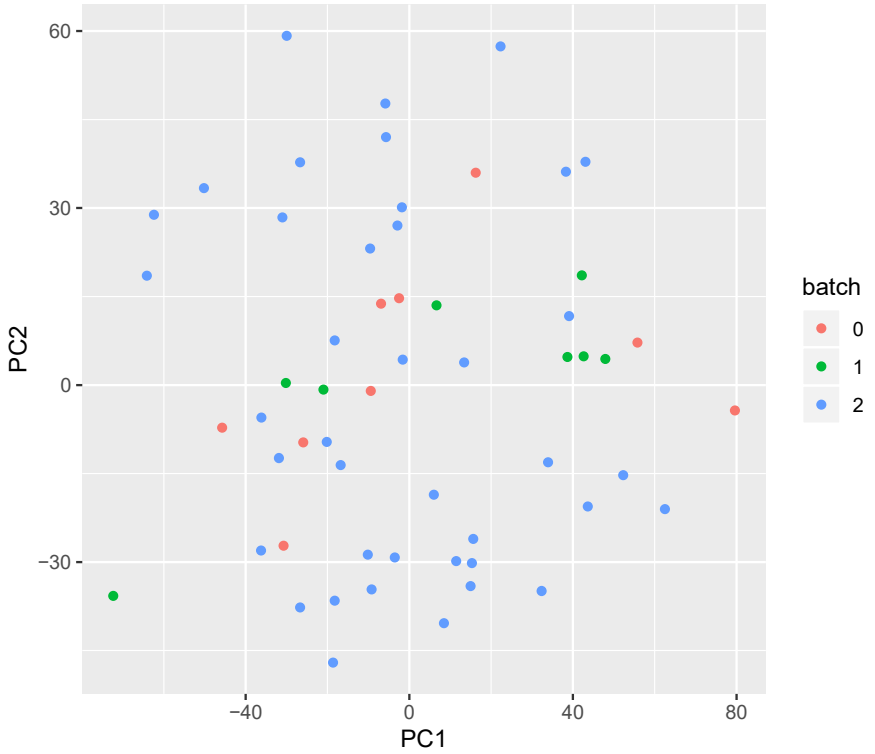



# Another Example

## Before

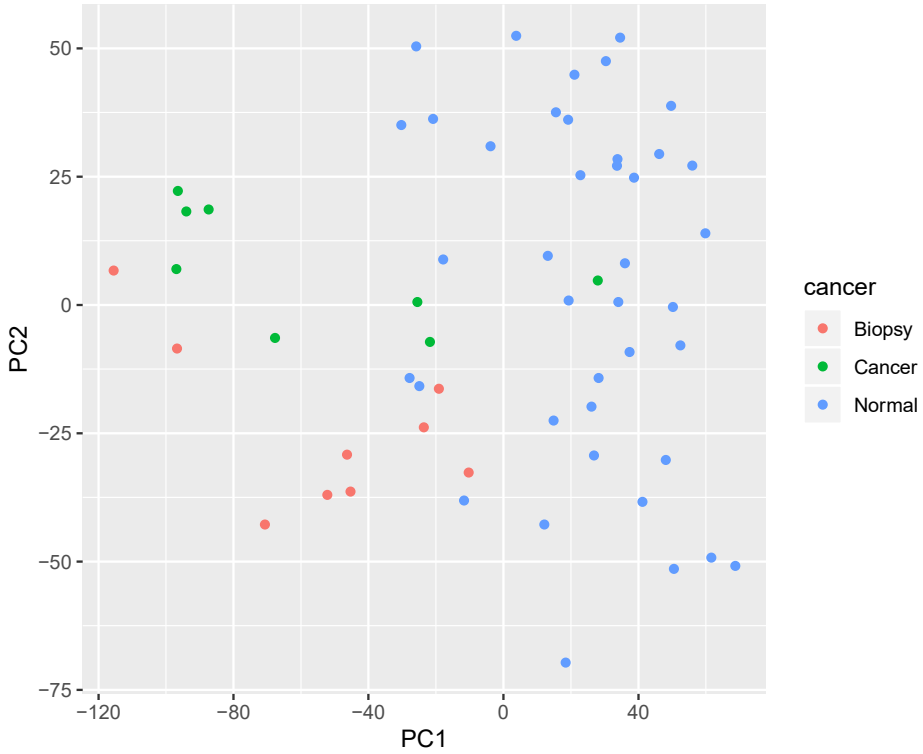



## After

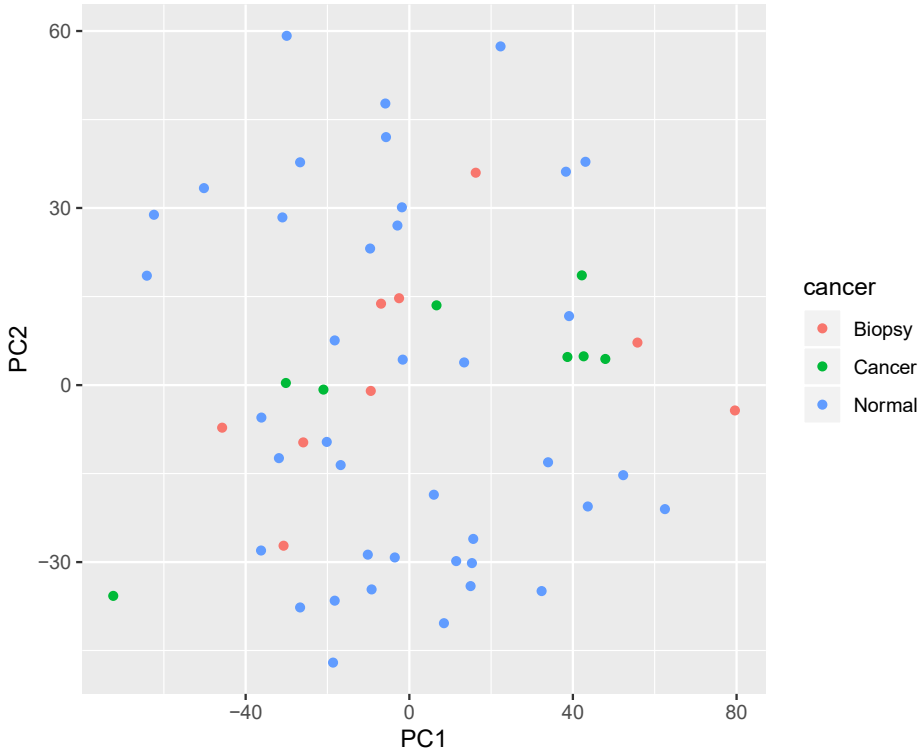


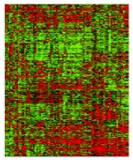
# Another Example

## Before




## After

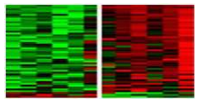




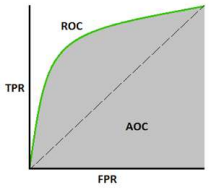
Raw data

Preprocessing

ctrl vs. case

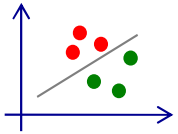


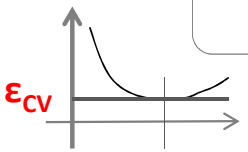
Feature selection



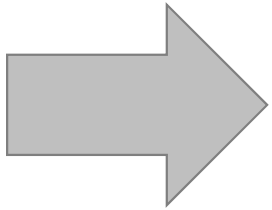

Internal validation

Classifier design






External validation



Clinical assay

Feature selection

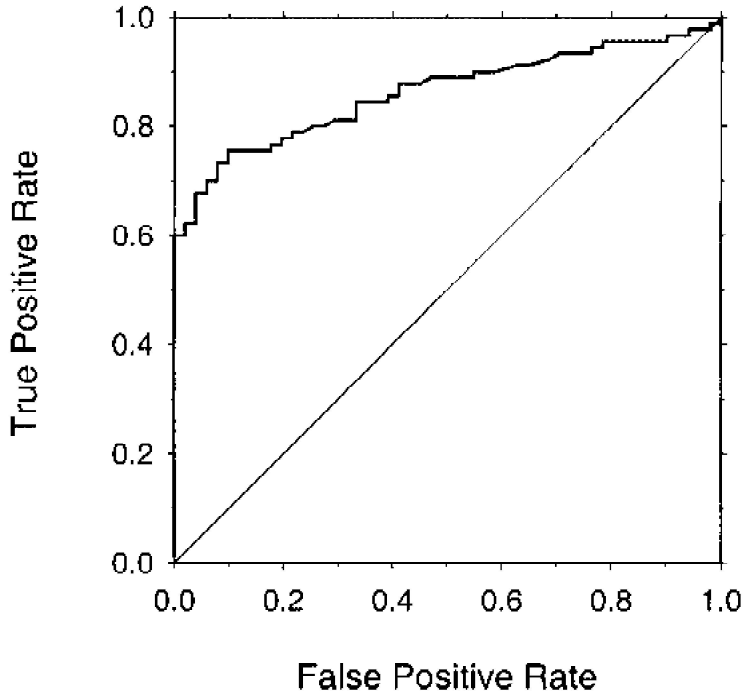
Classifier design

Internal validation

External validation

# Error Estimation

- TPR or sensitivity
- FPR or 1 - specificity
- ROC
- AUC




- Error or accuracy

- ***How well (accurately) can we estimate these?***

# Error Estimation

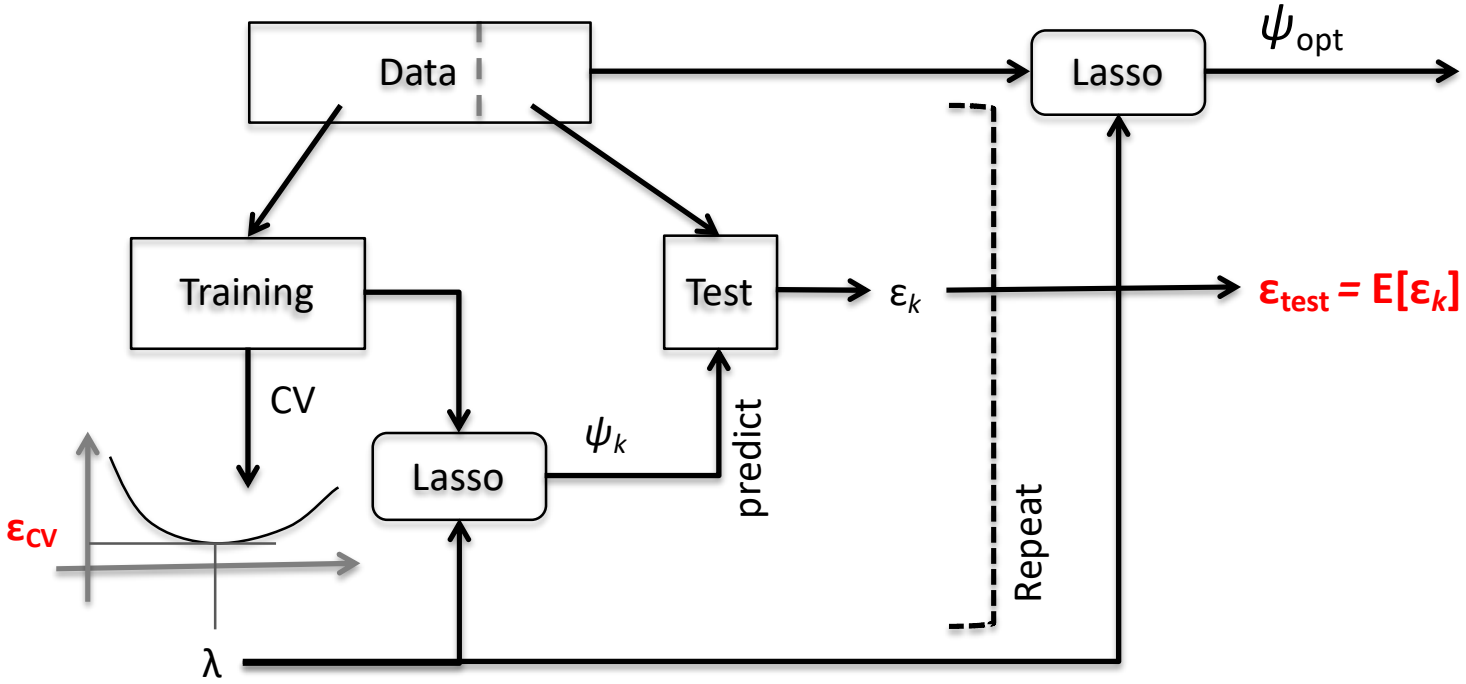
$$\varepsilon = E[(\varphi(\boldsymbol{X}_i) - Y_i)^2]$$ ← True Error

$$\hat{\varepsilon} = E\left[\left(\hat{\varphi}_{N,k}(\boldsymbol{X}_i) - Y_i\right)^2\right]$$ ← Estimated Error

$$= E\left[\left(\varphi(\boldsymbol{X}_i) + \left(\hat{\varphi}_k(\boldsymbol{X}_i) - \varphi(\boldsymbol{X}_i)\right) + \left(\hat{\varphi}_{N,k}(\boldsymbol{X}_i) - \hat{\varphi}_k(\boldsymbol{X}_i)\right) - Y_i\right)^2\right]$$

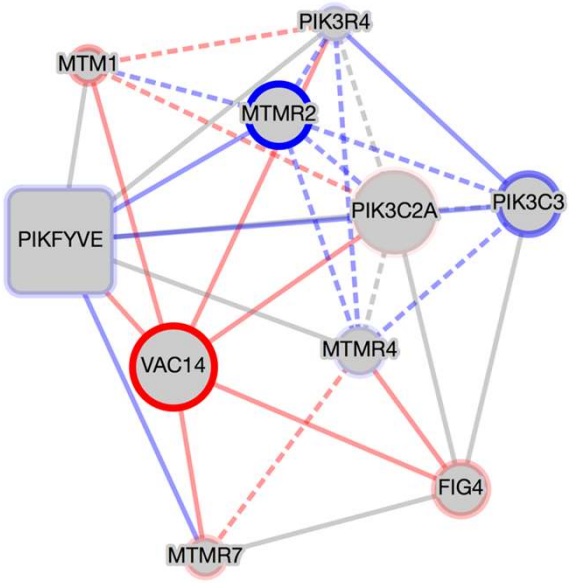
$$= E\left[\left((\varphi(\boldsymbol{X}_i) - Y_i) + \left(\hat{\varphi}_k(\boldsymbol{X}_i) - \varphi(\boldsymbol{X}_i)\right) + \left(\hat{\varphi}_{N,k}(\boldsymbol{X}_i) - \hat{\varphi}_k(\boldsymbol{X}_i)\right)\right)^2\right]$$

# Interval Validation

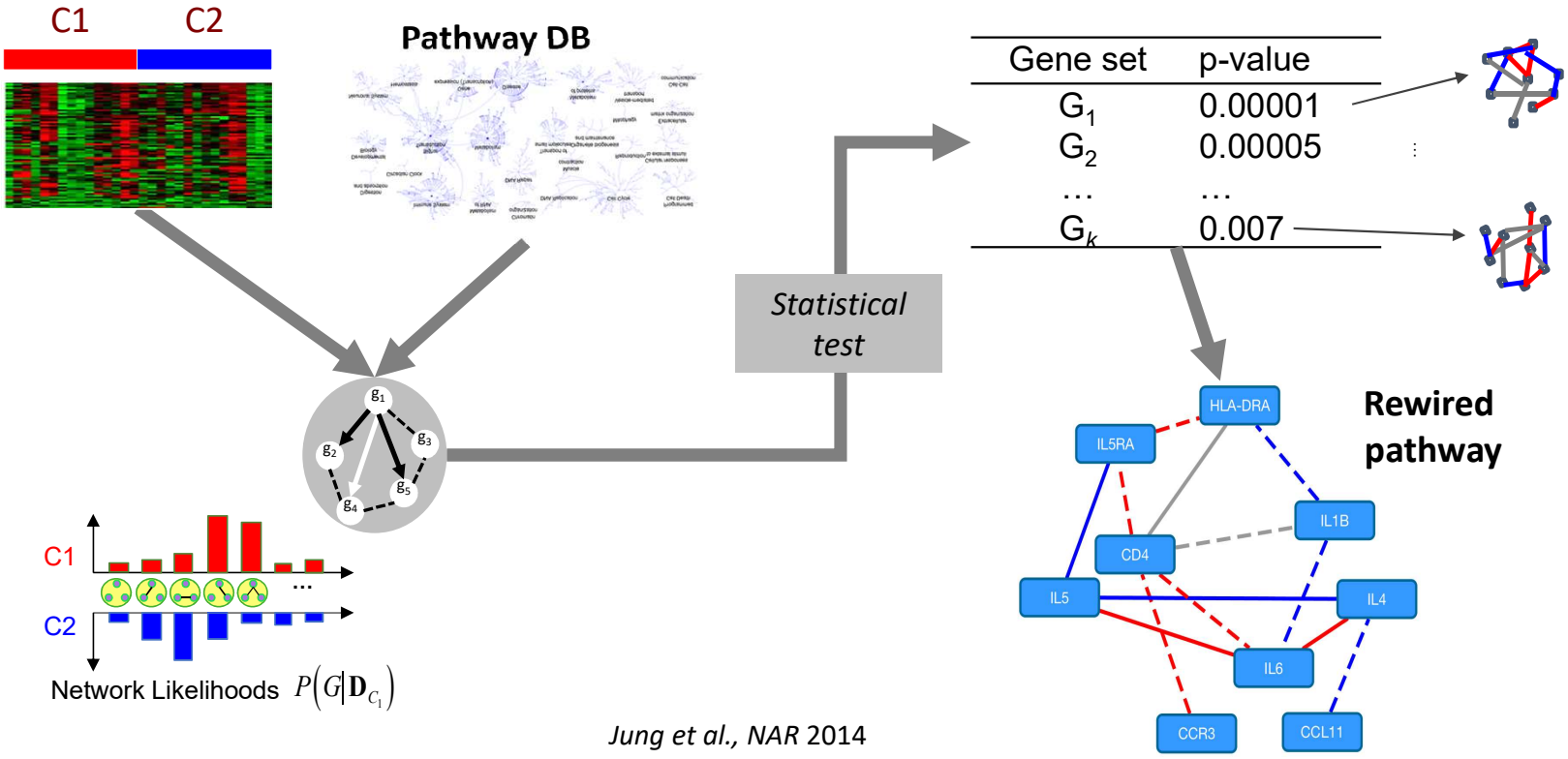



CV: cross-validation – find λ (lambda) for Lasso

# Network-based Approach



# EDDY: Evaluation of Differential DependencY




*Jung et al., NAR* 2014
*Speyer et al., PSB* 2016

# EDDY + CTRP-CCLE

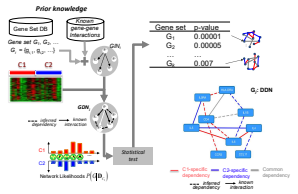
CCLE Cancer Cell Line Encyclopedia

RNAseq of 935 cancer cell lines
**>75TB in size**

CENTER FOR THE SCIENCE OF THERAPEUTICS

Cell lines chemical sensitivity for 481 small compounds

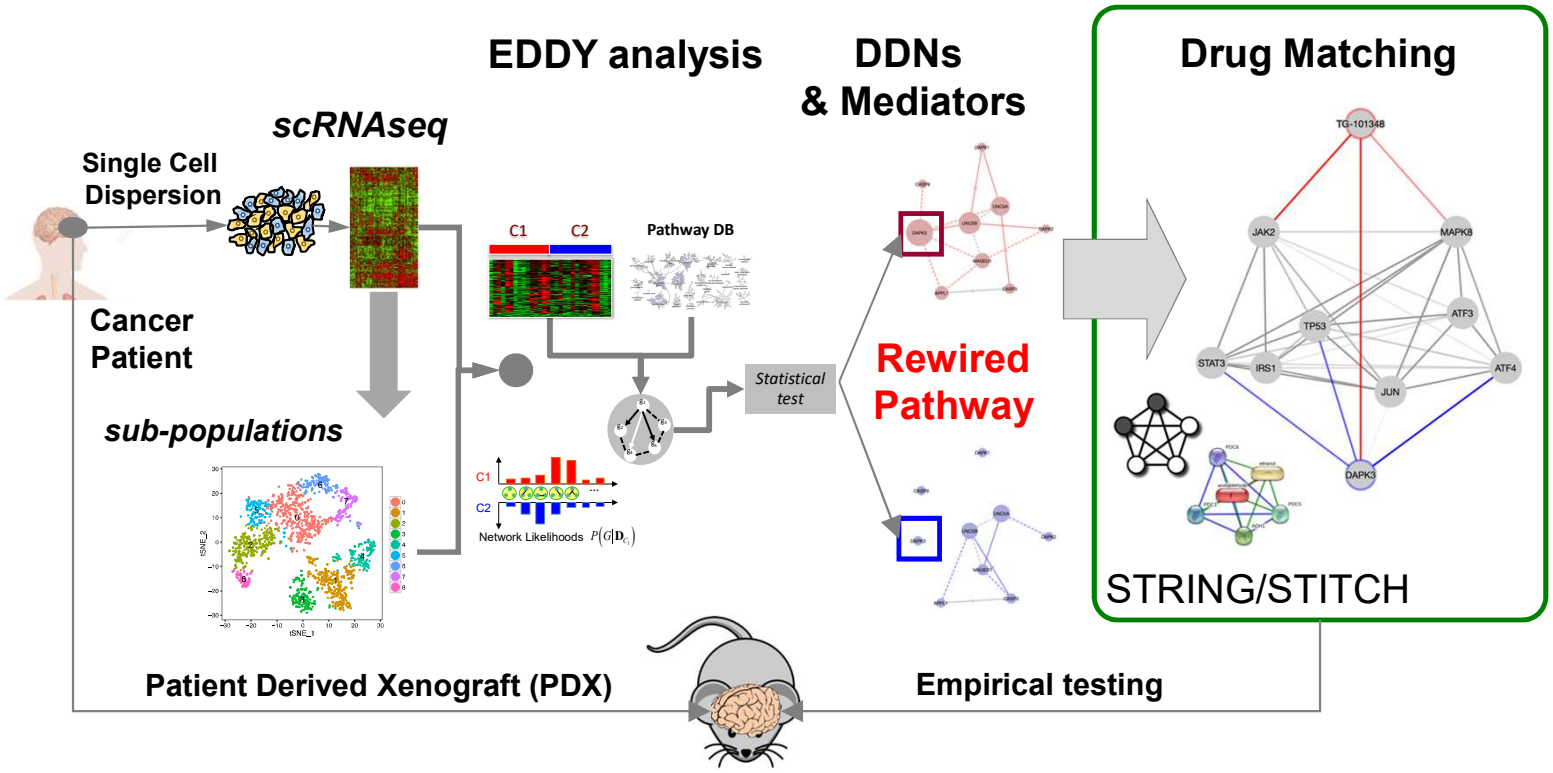
CTD² Cancer Target Discovery and Development



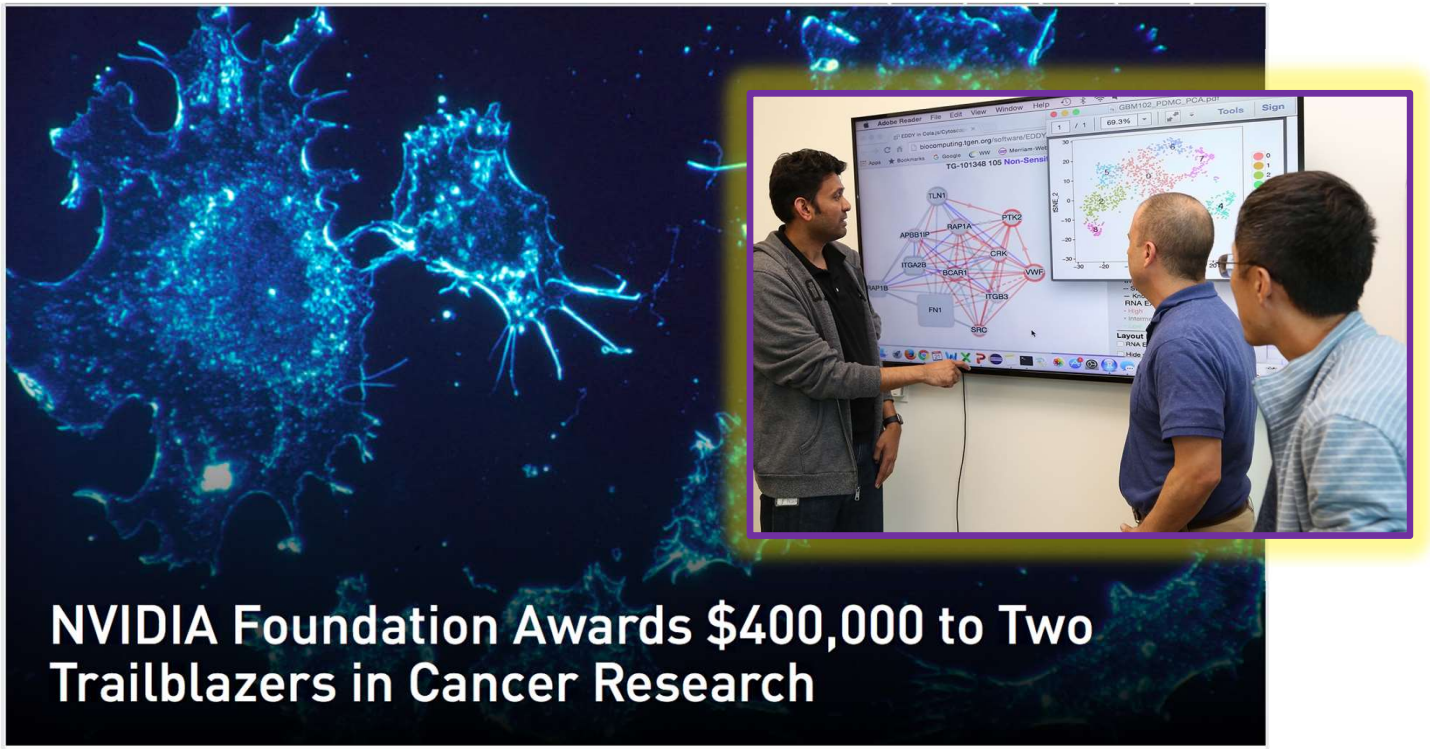
Speyer et al., Pac Symp Biocomput. 2017; 22: 497–508.

- Identifies pathways enriched with differential dependency between sensitive and non-sensitive cancer cell lines, as in DDNs
- Discover mediators of drug sensitivity, i.e. potential targets?

# Fighting Cancer, Cell by Cell



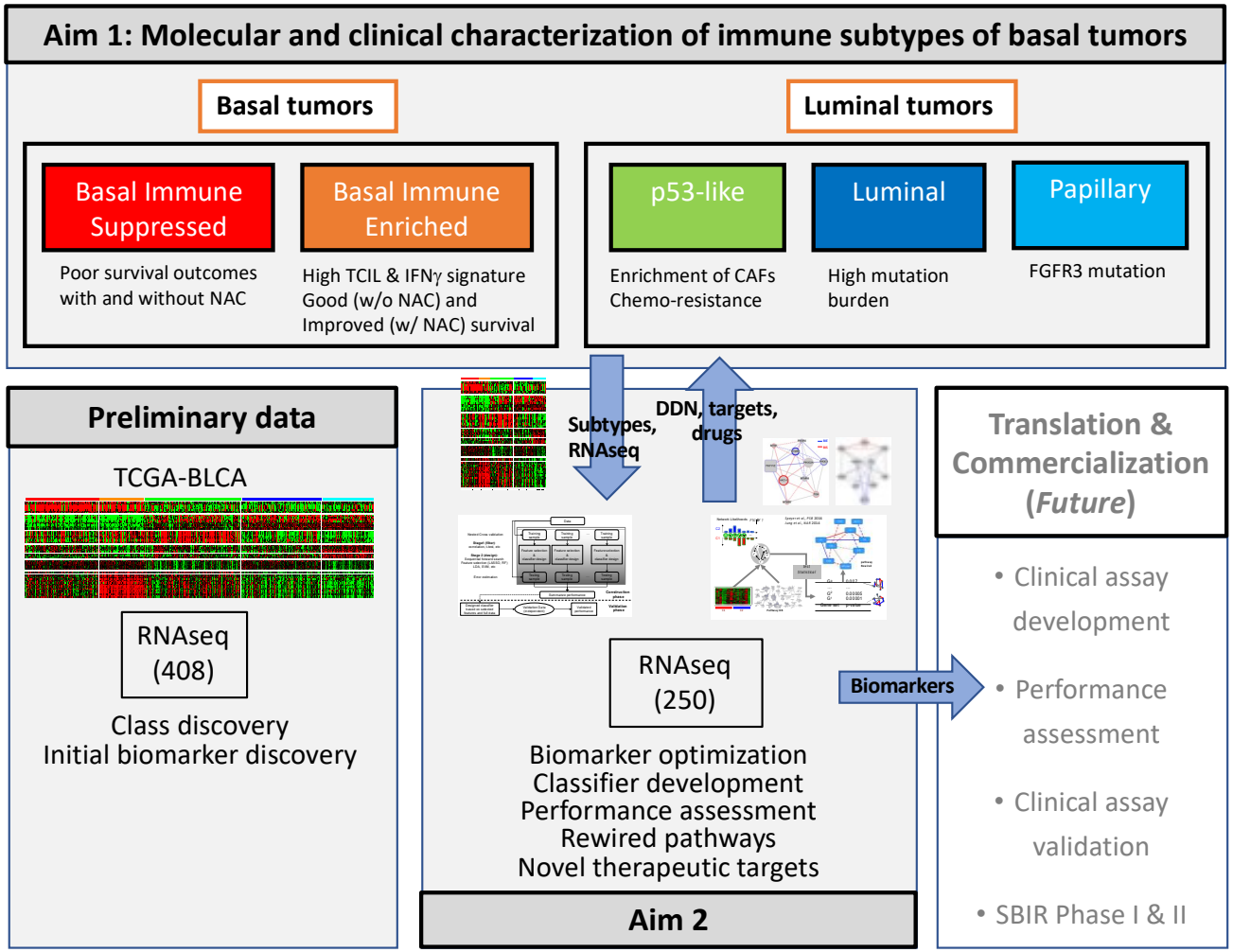
# Fighting Cancer, Cell by Cell



# What about Bladder Cancer!

Translational Team Science Award (DoD-USAMRMC-CDMRP-TTSA),
**"Development of Classifiers for Novel Bladder Cancer Subtypes"**

Woonyoung Choi (JHMI)
Seungchan Kim (PVAMU)

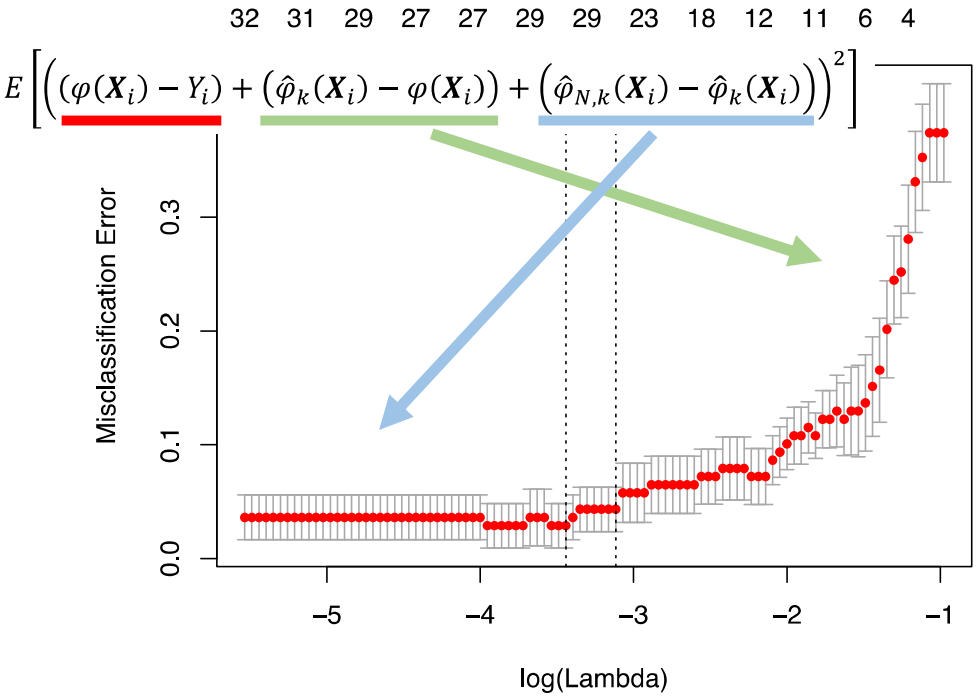


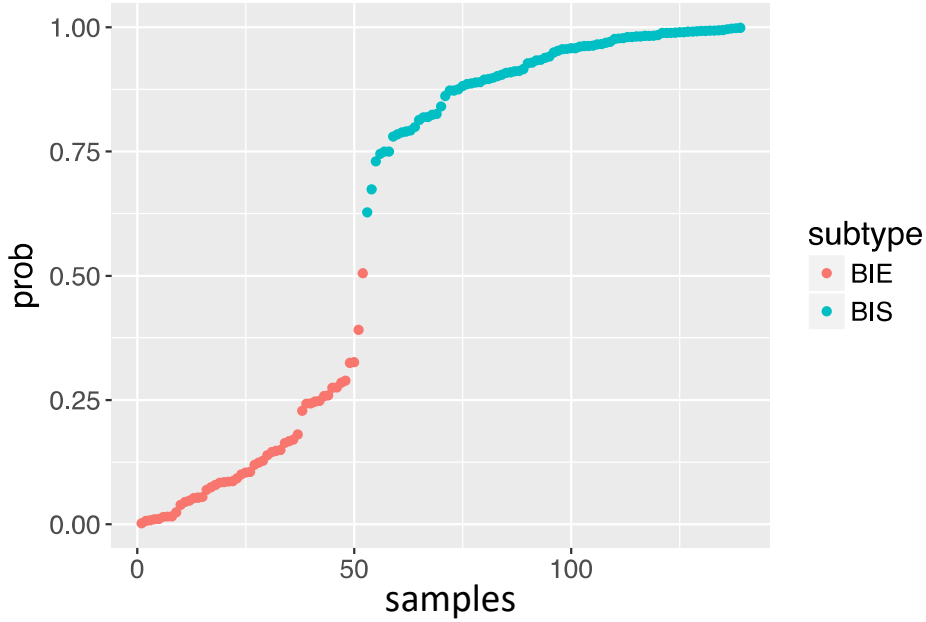
**Basal tumors**

| Basal Immune Suppressed | Basal Immune Enriched |
|---|---|
| Poor survival outcomes with and without NAC | High TCIL & IFNγ signature<br>Good (w/o NAC) and<br>Improved (w/ NAC) survival |

**Luminal tumors**

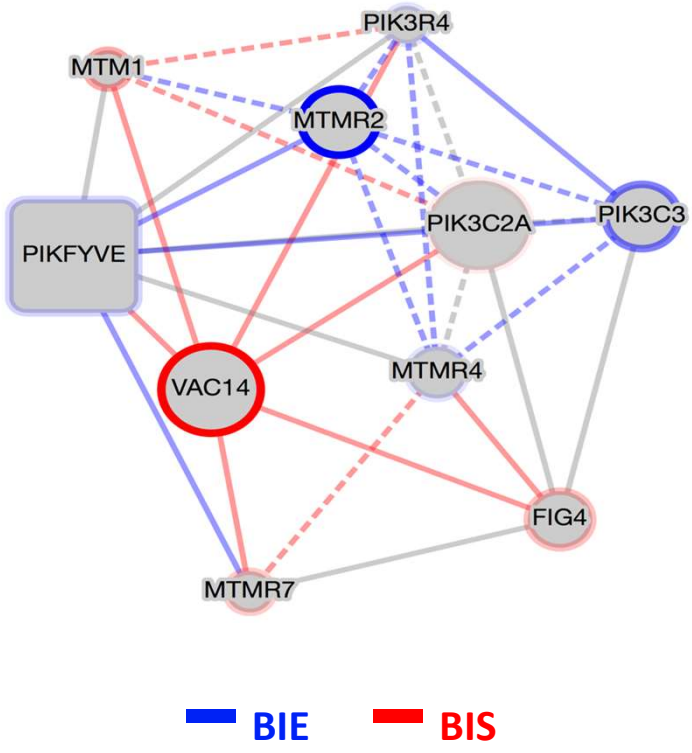
| p53-like | Luminal | Papillary |
|---|---|---|
| Enrichment of CAFs<br>Chemo-resistance | High mutation burden | FGFR3 mutation |



# Pathways enriched with differential dependency between BIE and BIS

| Pathway | # genes | p-val | Rewiring | Mediators |
|---|---|---|---|---|
| SYNTHESIS OF PIPS AT THE LATE ENDOSOME MEMBRANE | 10 | 0.0025 | 0.68 | PIKFYVE |
| MTORC1 MEDIATED SIGNALLING | 11 | 0.0183 | 0.88 | EIF4EBP1, MLST8 |
| PHOSPHORYLATION OF CD3 AND TCR ZETA CHAINS | 16 | 0.0225 | 0.27 | CD4, CD3E, CD3D, CD3G, PAG1, CSK |
| SEMA3A PAK DEPENDENT AXON REPULSION | 15 | 0.0236 | 0.65 | LIMK1 PLXNA1 |
| ELEVATION OF CYTOSOLIC CA2 LEVELS | 10 | 0.0246 | 0.60 | TRPC3 |
| SYNTHESIS OF PIPS AT THE EARLY ENDOSOME MEMBRANE | 12 | 0.0261 | 0.73 | PI4K2A MTMR2 |
| CELL EXTRACELLULAR MATRIX INTERACTIONS | 14 | 0.0327 | 0.55 | PARVA FERMT2 |
| SYNTHESIS SECRETION AND INACTIVATION OF GLP1 | 19 | 0.0390 | 0.87 | CDX2 PAX6 |
| CD28 DEPENDENT VAV1 | 11 | 0.0394 | 0.50 | PAK2 FYN |
| GRB2 SOS PROVIDES LINKAGE TO MAPK SIGNALING FOR INTERGRINS | 15 | 0.0454 | 0.61 | ITGB3 SOS1 TLN1 |
| THE ROLE OF NEF IN HIV1 REPLICATION AND DISEASE PATHOGENESIS | 28 | 0.0479 | 0.69 | AP2S1, CD8B, AP1S2, ELMO1, AP1S1, B2M |




# Center for Computational Systems Biology

Prairie View A&M University

CCSB@PVAMU

https://ccsb.pvamu.edu

# Q&A