BENG 183 Final Project

Application of Network Based Co-Expression Expression Patterns to Differential Expression of Genes in Tumorigenic Prostate Epithelial Cells

Network Based Co-Expression Patterns

A tool for integrating time-course expression data into differential expression analysis

Network-based comparison of temporal gene expression patterns

Wei Huang[1,2,†], Xiaoyi Cao[3,†] and Sheng Zhong[1,3,*]

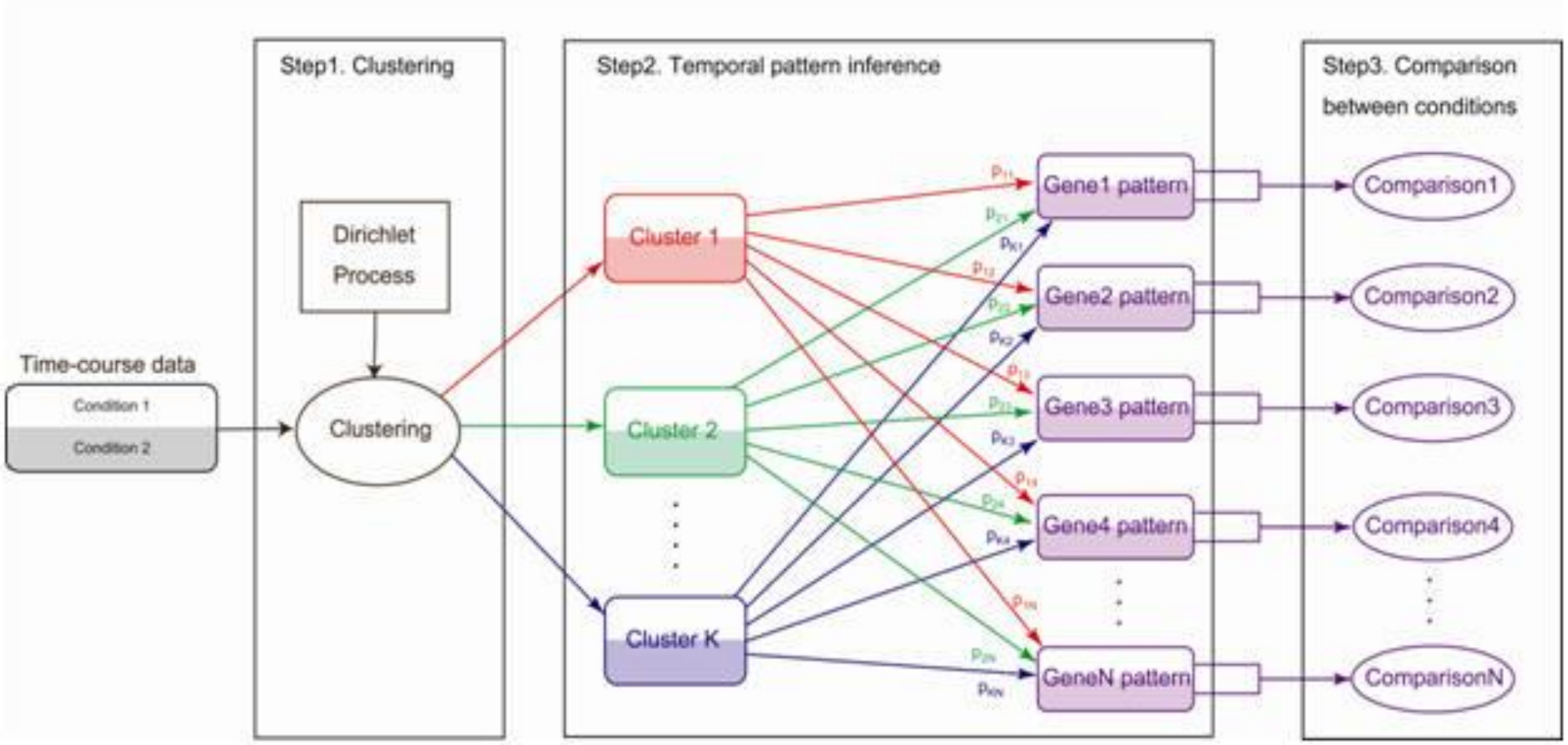
[1]Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA, [2]Key Laboratory for Applied Statistics of Ministry of Education, Northeast Normal University, Changchun, Jilin, China and [3]Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Associate Editor: Trey Ideker

Motivation:

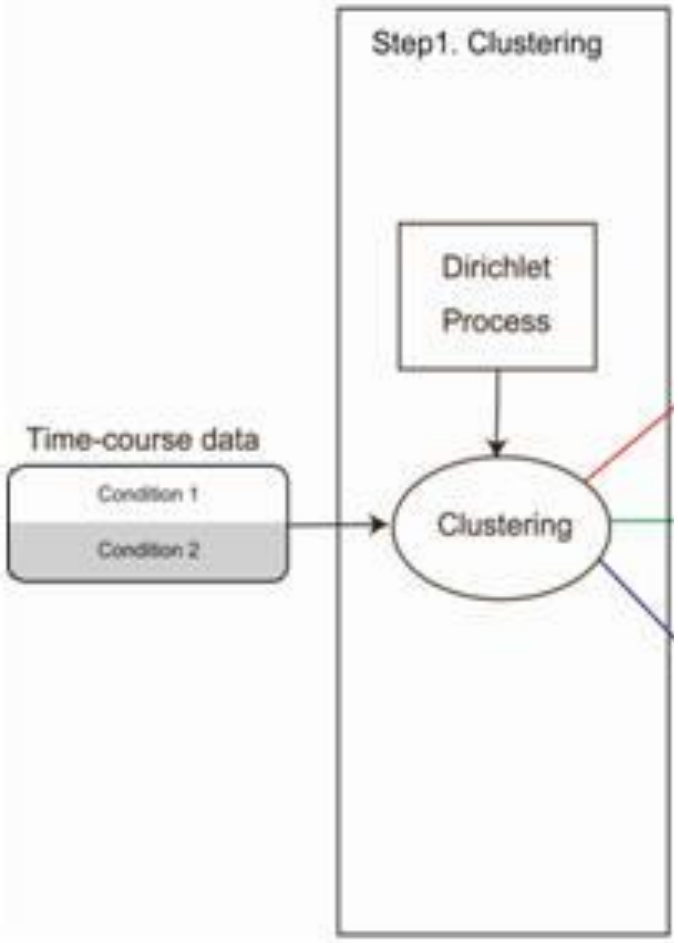
- Incorporate time-course expression data in differential expression analysis
- Prior work was limited to an independent gene-by-gene approach
- NACEP uses expression pattern clustering to solve these issues

Network Based Co-Expression Patterns (NACEP)



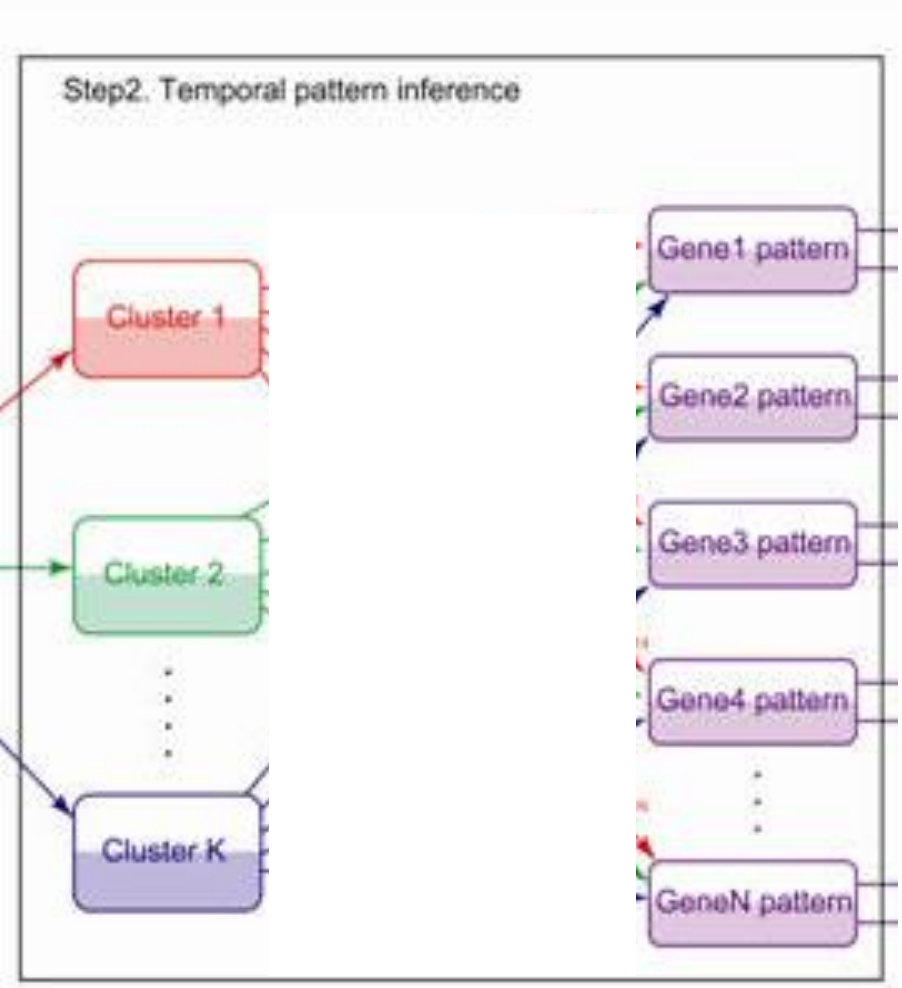
Huang et al., Bioinformatics (2010)

Network Based Co-Expression Patterns (NACEP)



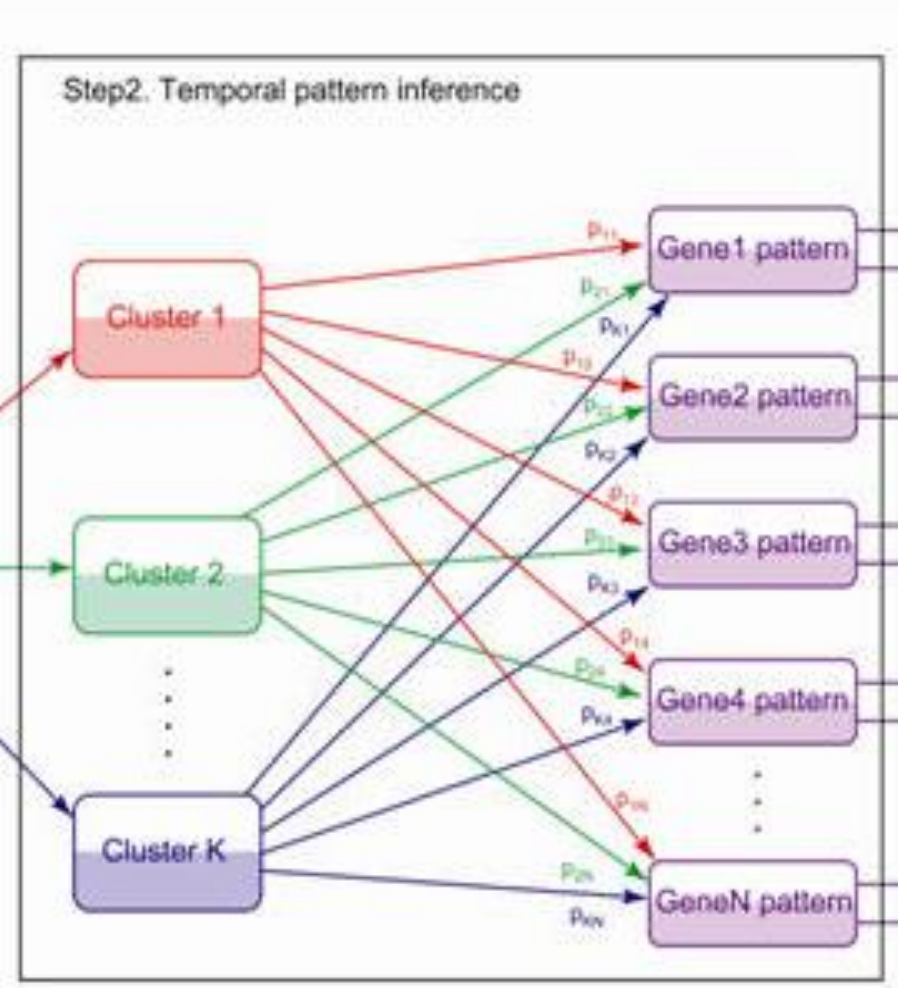
- Infinite-mixture model for clustering time-course data
- Number of clusters is determined by Dirichlet Process
- Cluster memberships are missing data, generated by Chinese Restaurant Process

Network Based Co-Expression Patterns (NACEP)



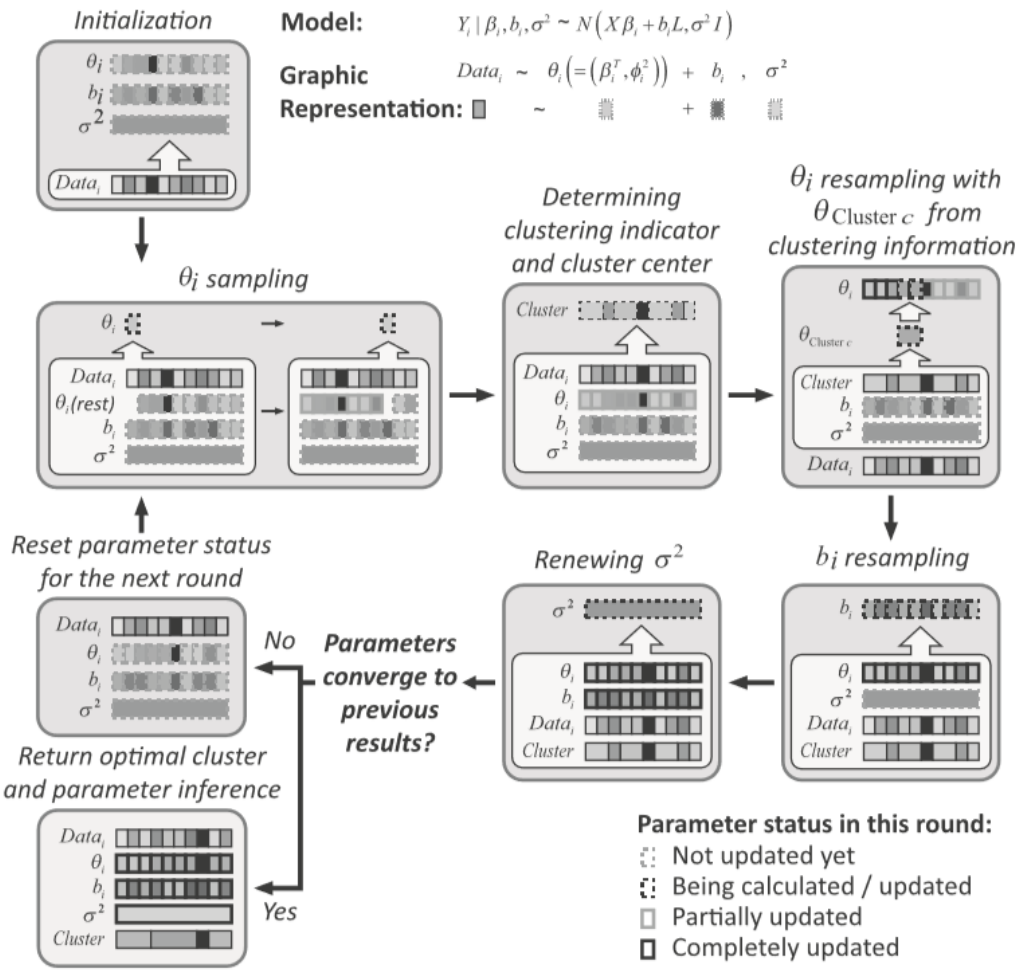
- Mixed-effects model of temporal gene expression patterns
  - $Y_{ijkl} = f_{cj}(t_k) + b_i + \varepsilon_{ijkl}$
  - Y_{ijkl} = Gene$_i$'s expression under condition$_j$ at time$_k$ under replicate$_l$
  - b_i = random gene effect
  - ε_{ijkl} = noise parameter
- Cluster mean profile $f_{cj}(t_k)$ is modeled as a B-spline (smooth function that passes through control points)
- **Generates expression levels under generated clusters**

Network Based Co-Expression Patterns (NACEP)



- Cluster assignment probabilities calculated from Bayesian model
  - NACEP model is rewritten in Bayesian form with Dirichlet Process prior
- Bayesian posterior probabilities = probability for any gene belonging to any cluster for all genes and all clusters
- Gibbs sampling algorithm makes model inferences

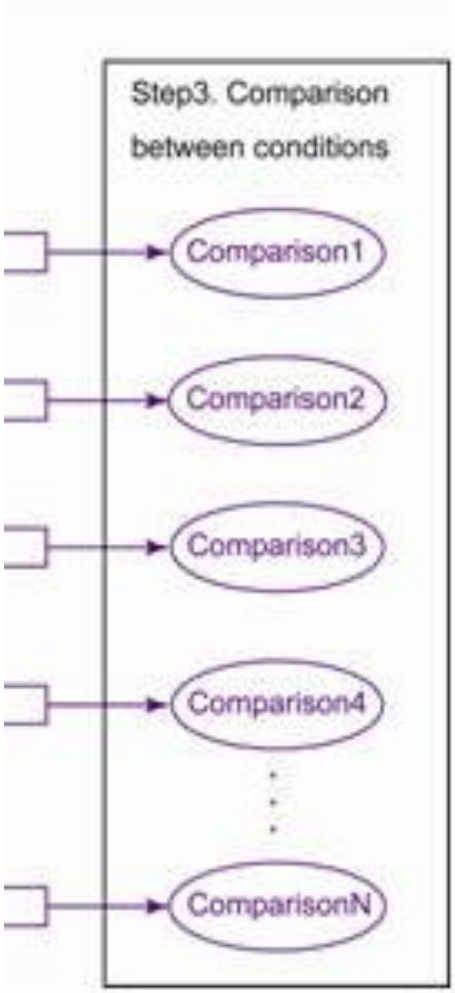
Network Based Co-Expression Patterns (NACEP)



- Gibbs Sampling Algorithm
  - Θ_i = collection of all parameters
- Estimates parameters for Bayesian form of NACEP model
- Runs until parameters converge OR specified number of iterations

Huang et al., Bioinformatics (2010)

Network Based Co-Expression Patterns (NACEP)



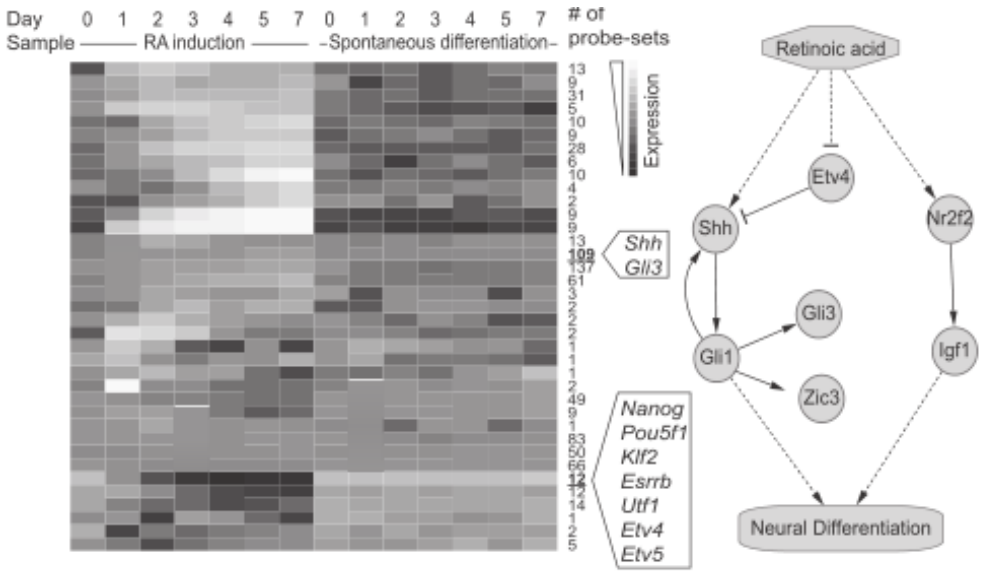
- Differences between experimental conditions are calculated as **distances**
- Distance is calculated as weighted average of differences between temporal patterns per cluster
  - Weights = posterior probability of gene-cluster membership
- Differentially expressed genes are ranked by statistical significance (FDR-corrected for multiple hypothesis testing)

Validation of NACEP Methodology

- Authors used NACEP on gene expression data for Embryonic Stem Cells (ES) (Data obtained from Ivanova *et. al*)
- Tested Retinoic Acid (RA) induced differentiation vs. spontaneous differentiation
- Tested with data collected over time period of 0 to 7 days
- Results were compared to Ivanova *et. al (2007)*

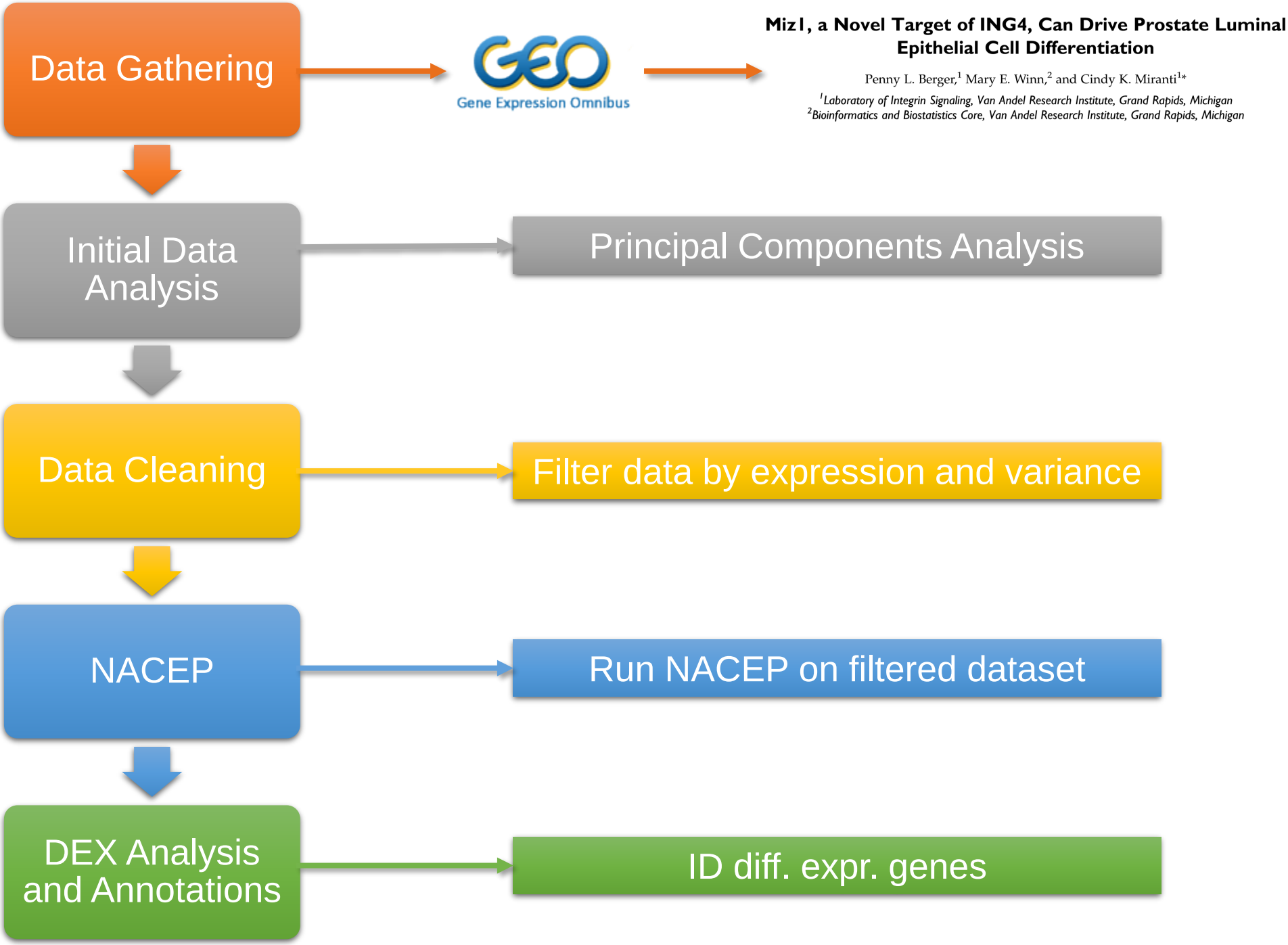
RA induced vs. Spontaneous

- NACEP ranked Gli3, Zic3, and Shh highly in the RA sample
- Gli and Shh were clustered together – consistent with hypothesis
- Consistent with results from Ivanova *et. al*



Analysis of Tumorigenic Prostate Epithelial Cells

Incorporating time-course data in differential expression analysis with NACEP


Data Gathering
Gene Expression Omnibus
Miz1, a Novel Target of ING4, Can Drive Prostate Luminal Epithelial Cell Differentiation
Penny L. Berger,[1] Mary E. Winn,[2] and Cindy K. Miranti[1]*
[1]Laboratory of Integrin Signaling, Van Andel Research Institute, Grand Rapids, Michigan
[2]Bioinformatics and Biostatistics Core, Van Andel Research Institute, Grand Rapids, Michigan
Initial Data Analysis
Principal Components Analysis
Data Cleaning
Filter data by expression and variance
NACEP
Run NACEP on filtered dataset
DEX Analysis and Annotations
ID diff. expr. genes

Miz1, a Novel Target of ING4, Can Drive Prostate Luminal Epithelial Cell Differentiation

Penny L. Berger,[1] Mary E. Winn,[2] and Cindy K. Miranti[1]*

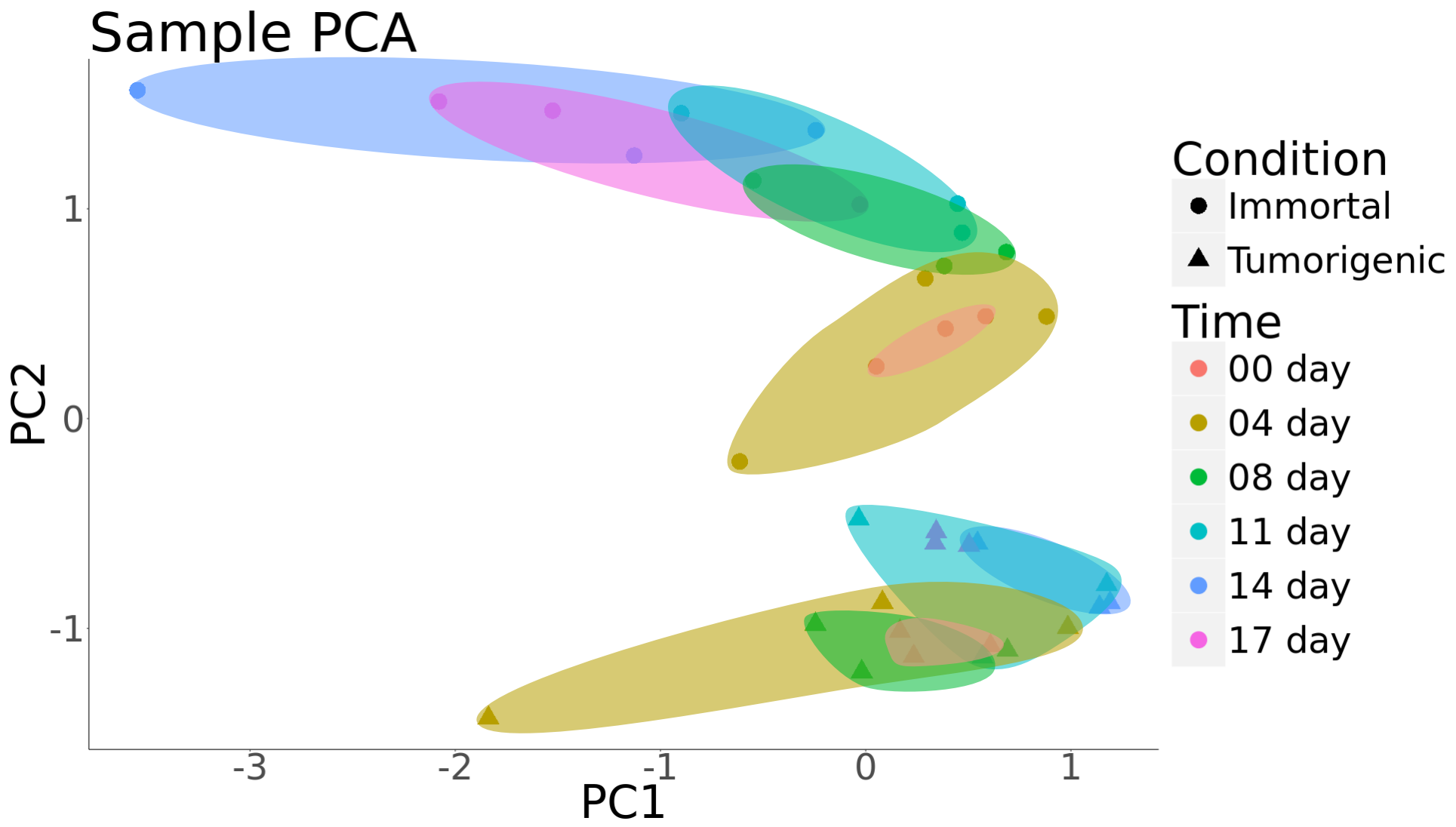
[1]*Laboratory of Integrin Signaling, Van Andel Research Institute, Grand Rapids, Michigan*
[2]*Bioinformatics and Biostatistics Core, Van Andel Research Institute, Grand Rapids, Michigan*

- Control samples: Immortalized Prostate Epithelial Stem Cells
  - iPrEC
  - Generated from primary clinical prostectomies
  - Express only basal epithelial cell markers
- Disease samples: Tumorigenic Prostate Epithelial Stem Cells
  - EMP
  - Overexpression of Erg, Myc, shPten
- 25,702 initial genes with 6 time periods, 3 replicates per time period
- **Datasets downloaded from NCBI Gene Expression Omnibus as read count matrices and converted to TPM**

Sample Principal Components Analysis

- Convert possibly correlated values into linearly uncorrelated **principal components**
  - Resulting vectors are mutually orthogonal
- Summarizes data points by data variance
- Procedure:
  - Perform variance stabilizing transformation on counts matrix
  - Calculate scaled and centered principal components
  - Rotate and plot

Sample Principal Components Analysis



Filtering by Expression and Variance

- NACEP is very computationally expensive
  - Need to drastically reduce dataset size
- Gene Retention Criteria
  - 17% of samples have TPM >= 39 (average TPM)
  - $\log_{10}$(gene variance) >= 2.4
  - Result: 4022 genes retained out of 25702 genes

NACEP Function Call and Parameters

```
source("NACEP.r");
NACEP(filename="dataFilter.txt",     # input file
      spcNum=2,                      # conditions
      Timelength=18,                 # time points
      Knot=15,                       # knots for spline
      loop=300,                      # iterations (default=500)
      compStart=200,                 # begin comparisons
      compInterval=100,              # interval betw. results
      alpha=50                       # cluster strength
      );
```

High Scoring Genes

Entrez ID	Gene Symbol	Description	Previous cancer studies?
3861	KRT14	Epithelial cell cytoskeleton	Yes
4536	MT-ND2	Mitochondrially encoded NADH dehydrogenase 2	Yes
6319	SCD	Fatty acid biosynthesis	Yes
667	DST	Adhesion junction plaque	No
7812	CSDE1	RNA-binding	No
5317	PKP1	Cytoskeleton	Yes
3868	KRT16	Epithelial cell cytoskeleton	No
9168	TMSB10	Actin binding	Yes
6273	S100A2	Cell cycle regulation	Yes

References

Huang, W., Cao, X., & Zhong, S. (2010). Network-based comparison of temporal gene expression patterns. *Bioinformatics*, *26*(23), 2944–2951. http://doi.org/10.1093/bioinformatics/btq561

Berger, P. L., Winn, M. E., Miranti, C. K. (2016). Miz1, a Novel Target of ING4, Can Drive Prostate Luminal Epithelial Cell Differentiation. *The Prostate*. *77(1),* 49-59. http://doi.org/10.1002/pros.23249