Integrating metabolomics and transcriptomics to study metabolic reprogramming

Janet Li, Weiruo Zhang

Plevritis Lab, Summer 2018

Background

Metabolic Reprogramming

- **Metabolic reprogramming is a cancer hallmark**
  - Promotes tumor invasion, progression and resistance to treatment
  - Oncogenic metabolites can alter cell signaling and block cellular differentiation
- **Challenges**
  - Due to technology limitations, only a small number of metabolites can be reliably measured at once
  - Most papers only analyze 1-2 metabolites, instead of looking at global metabolic pathways

L.M. Phan et al., *Cancer biology & medicine*, 2014
P.S. Ward et al., *Cancer cell*, 2012

The Problem

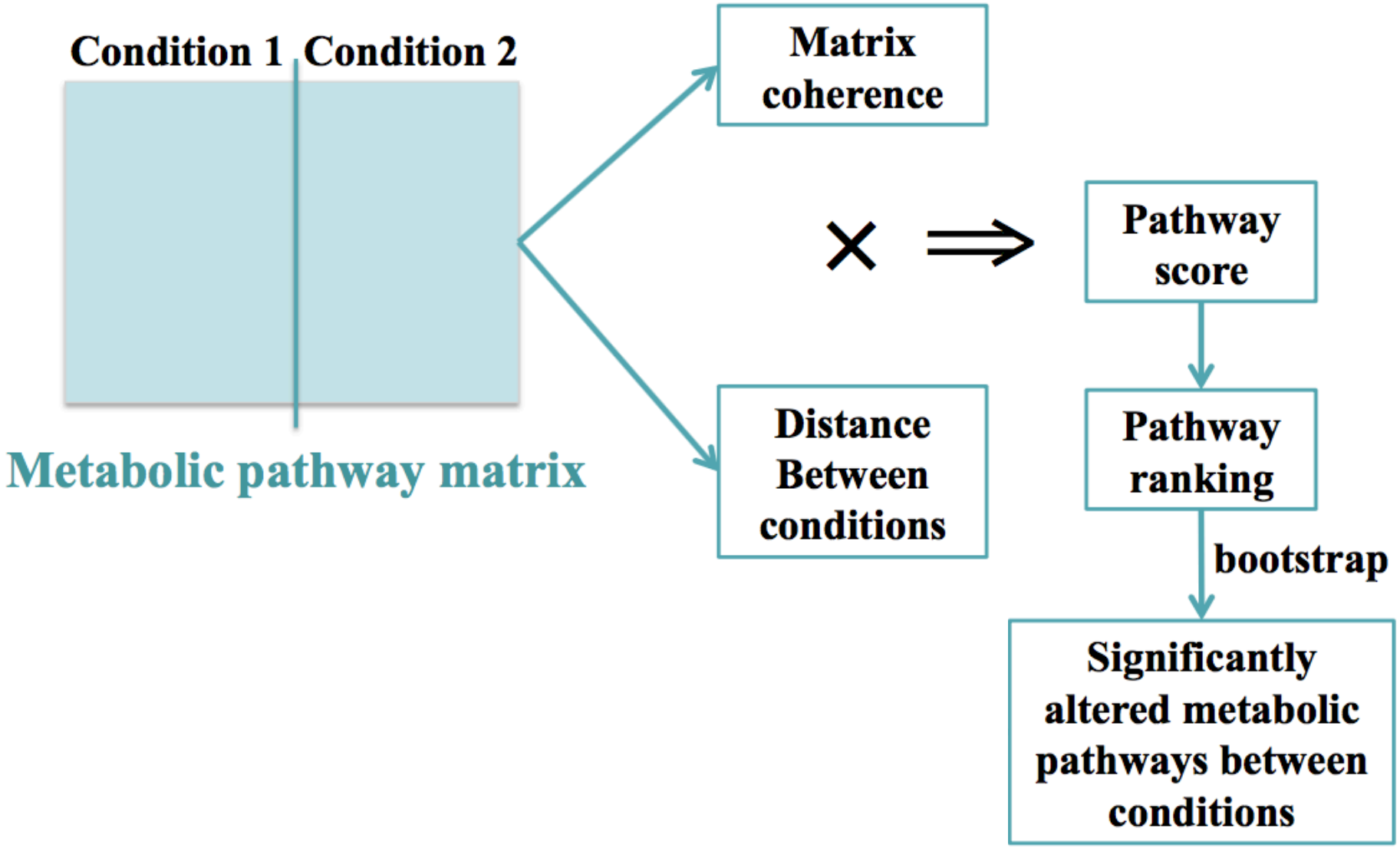
- **Input data**
  - Metabolomics
  - Transcriptomics (gene expression or RNAseq)
  - Metabolic pathway (KEGG)
- **Goal: to study metabolic pathway changes with respect to two conditions (e.g. Normal vs. Tumor)**

M. Kanehisa et al., *Nucleic Acids Research*, 2000

W. Zhang et al., AACR 2018

Algorithm

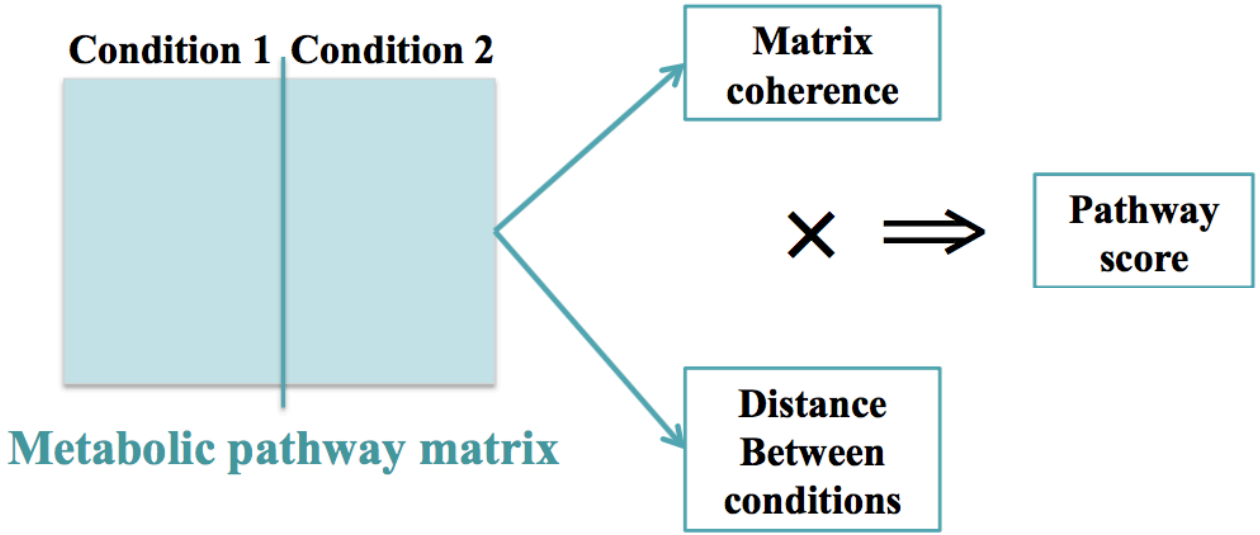
Algorithm Overview



W. Zhang et al., AACR 2018

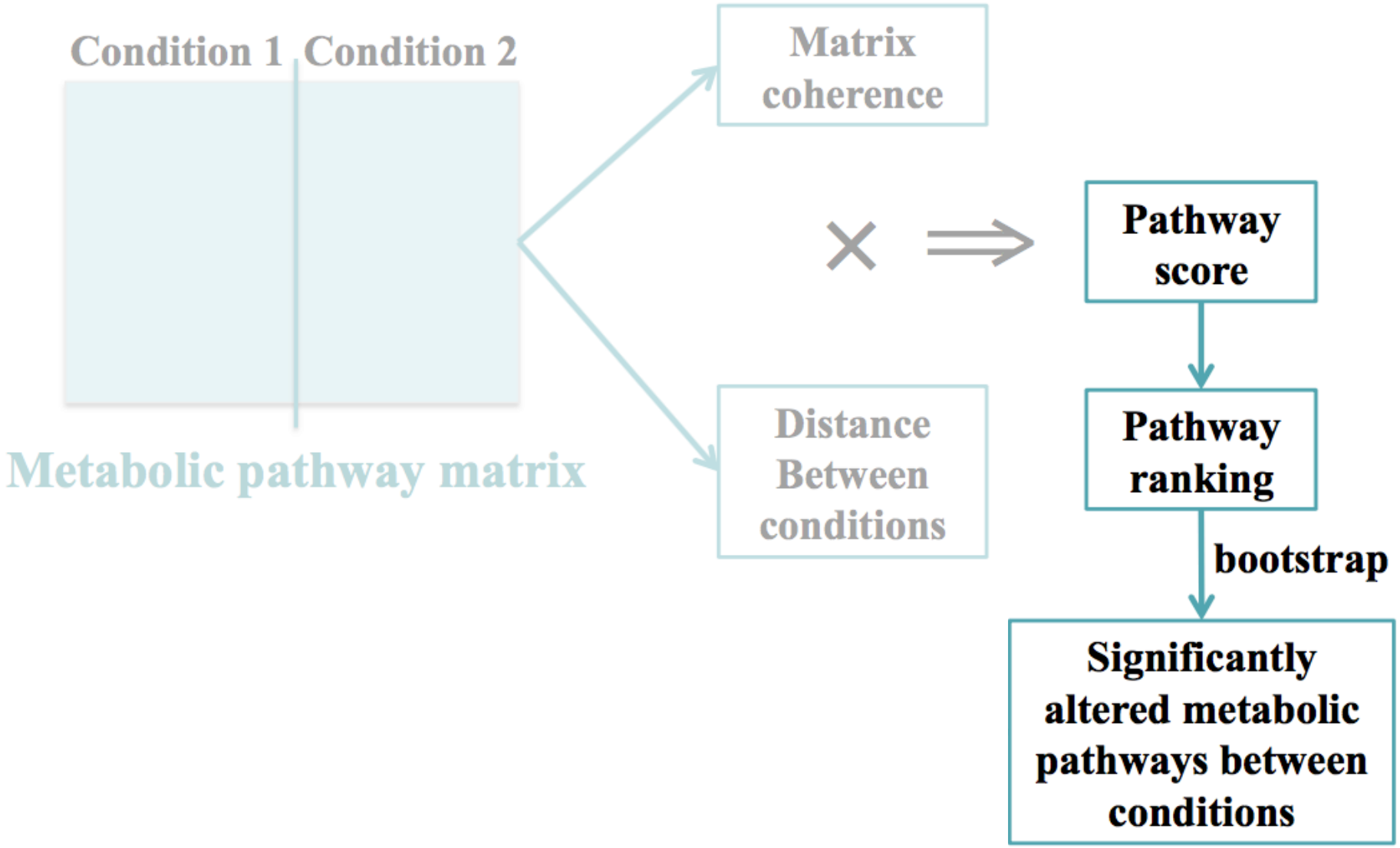
Pathway Score

- **Matrix coherence**
  - Mean of absolute value of correlation between the rows (genes and metabolites) of the pathway matrix
- **Distance between conditions**
  - Probabilistic principal component analysis: variation of PCA
    - Calculate optimal number of principal components through maximum likelihood estimation
  - Calculate distance between centroids of Condition 1 and Condition 2 in reduced dimensions space



W. Zhang et al., AACR 2018

Pathway Ranking

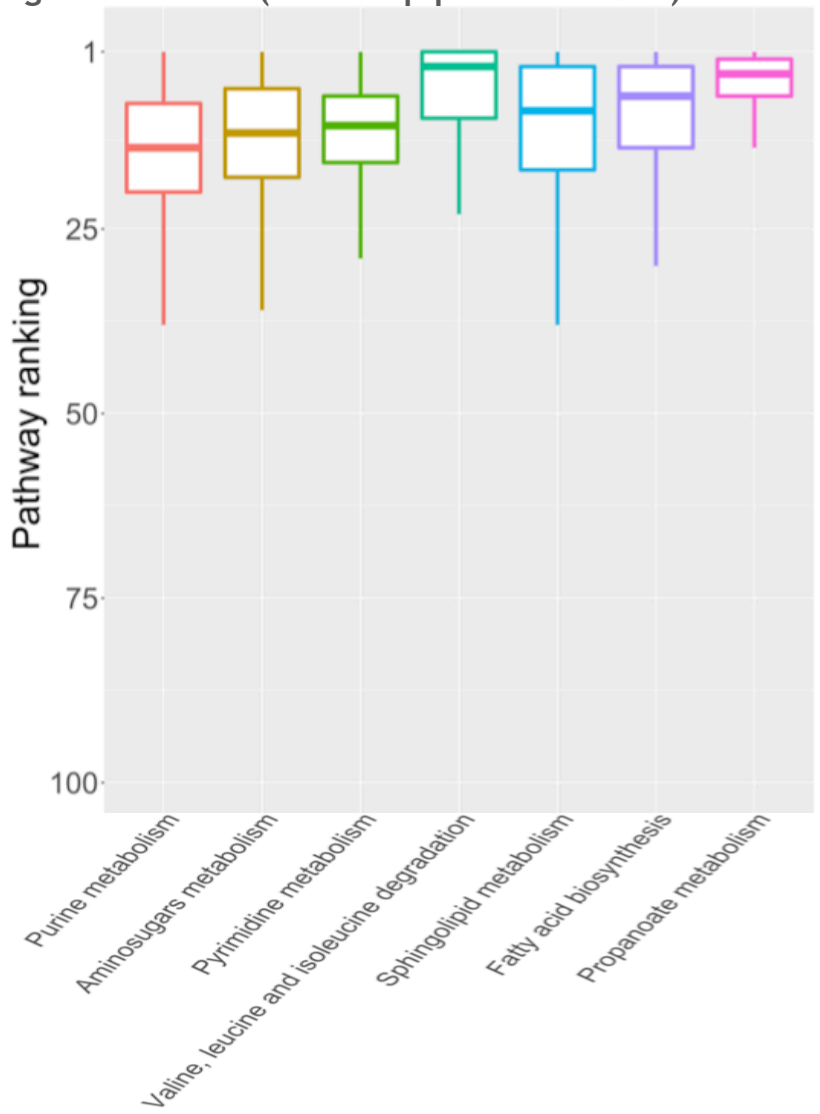


W. Zhang et al., AACR 2018

Results on cell lines

- Condition 1: Control of NSCLC cells
  Condition 2: NSCLC cells treated with TGF-β to induce EMT
- Metabolic pathway database: KEGG (104 metabolic pathways included)

Top-ranked metabolic pathways that were significantly changed after EMT (bootstrap p-value < 0.04)



W. Zhang et al., AACR 2018

Patient Tissue Data

Prostate Cancer and Breast Cancer

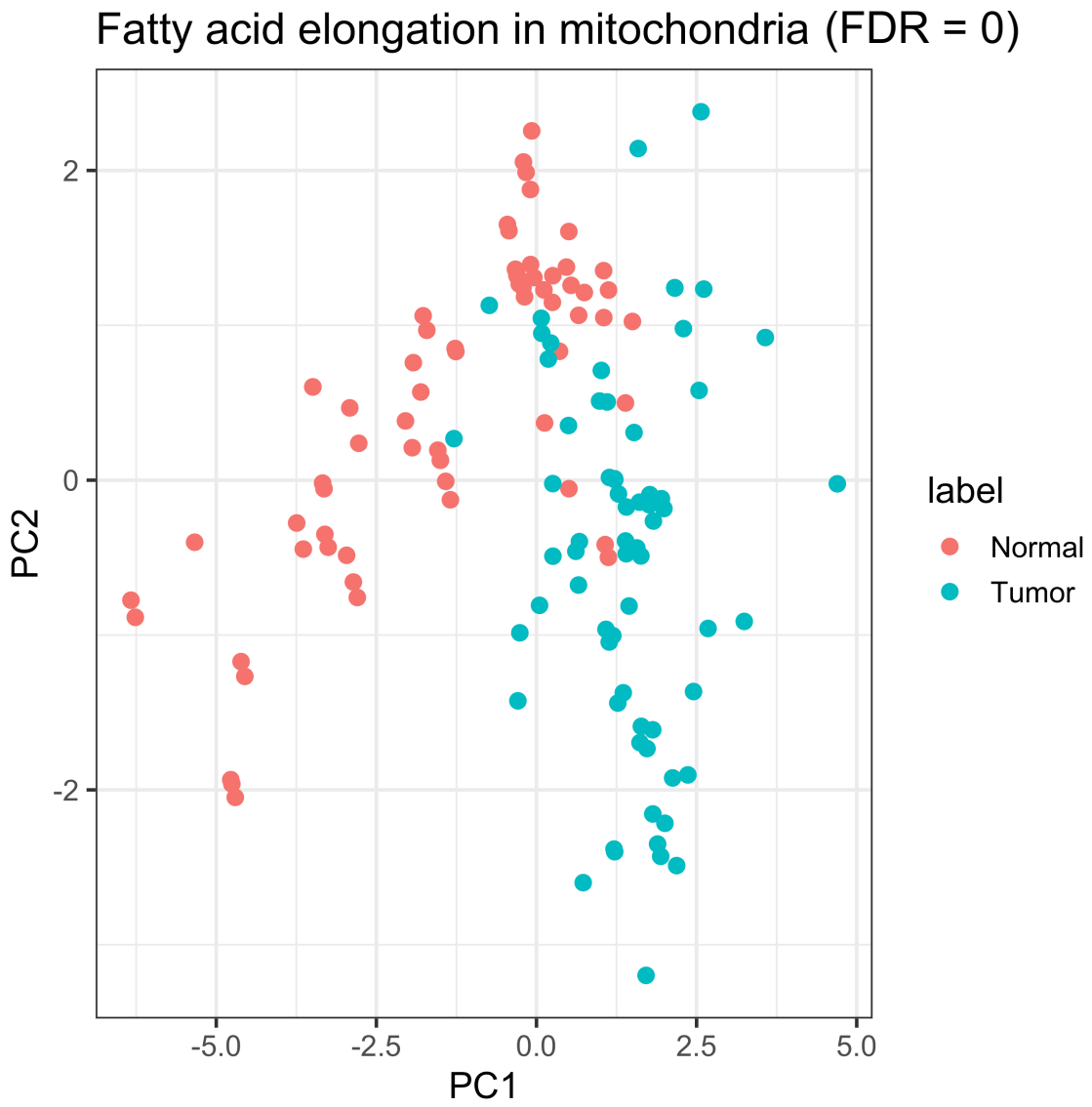
Breast Cancer Data and Results

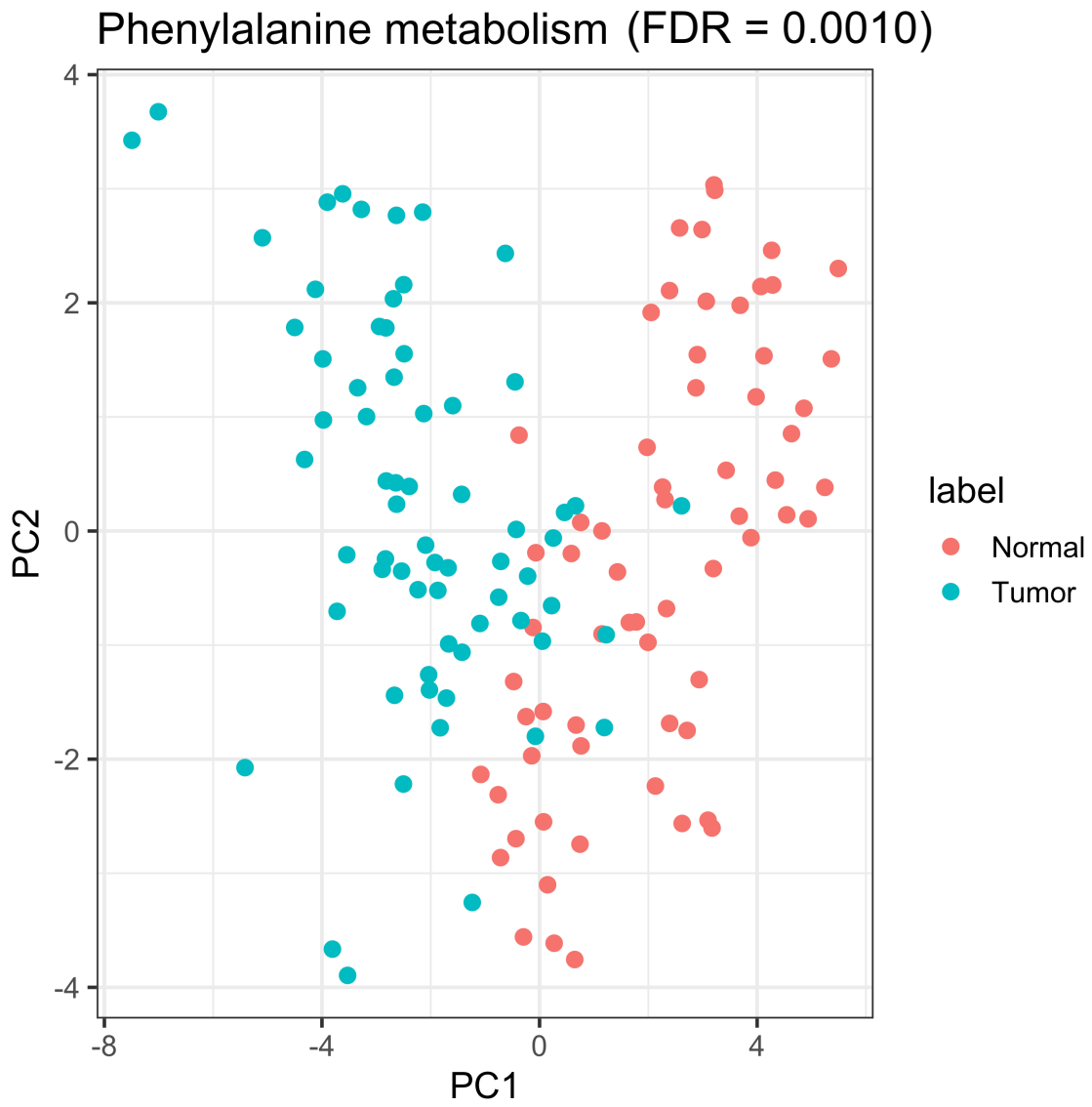
- 25,759 genes
  61 human breast tumors
  47 normal
- 352 identified metabolites
  67 human breast tumors
  65 tumor-adjacent noncancerous tissues
- Focused on 2-hydroxyglutarate (2HG)

Pathway Name	FDR
Fatty acid elongation in mitochondria	0
1- and 2-Methylnaphthalene degradation	0
Glyoxylate and dicarboxylate metabolism	0
Methane metabolism	0
Reductive carboxylate cycle (CO2 fixation)	0
Phenylalanine metabolism	0.001
Cyanoamino acid metabolism	0.001
Bile acid biosynthesis	0.002
Propanoate metabolism	0.002
Beta-Alanine metabolism	0.005
Valine, leucine and isoleucine degradation	0.006
Glycine, serine and threonine metabolism	0.007
Urea cycle and metabolism of amino groups	0.010
Lysine degradation	0.022
Histidine metabolism	0.034
Fatty acid metabolism	0.037
Glycolysis/Gluconeogenesis	0.044
Phenylalanine, tyrosine and tryptophan biosynthesis	0.049

A. Terunuma et al., *The Journal of Clinical Investigation*, 2014

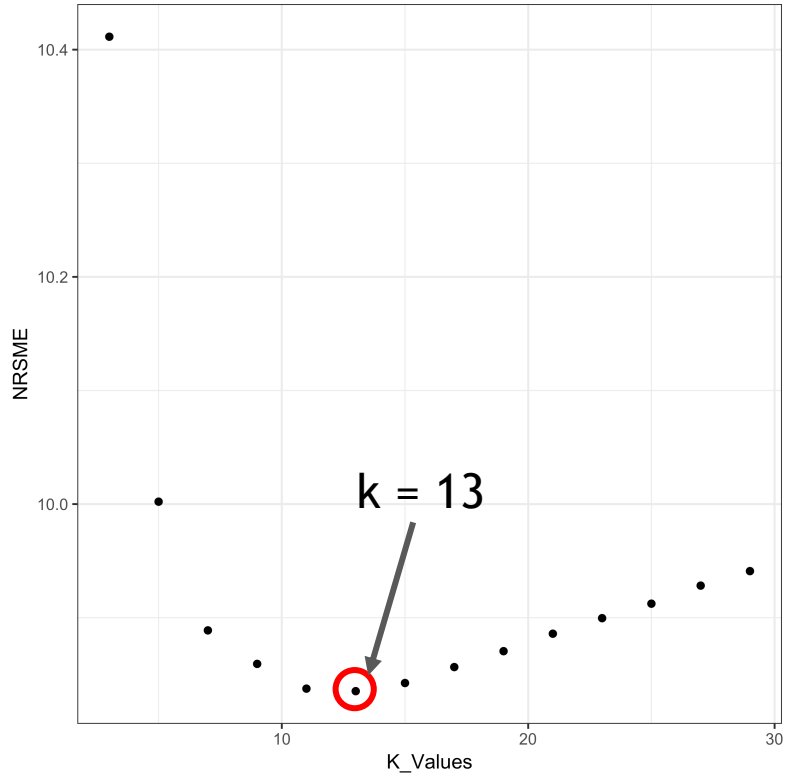
PCA plots of top pathways





Prostate Cancer Data

- 12 prostate cancer samples, 16 normal samples
- 31,035 genes, 186 metabolites
- About 20% of the gene expression data is missing, so we use k-nearest neighbors (kNN) imputation to estimate the missing data
  - Find k-nearest genes with similar expression based on Euclidean distance, then uses weighted average to estimate the missing values.
- Finding the optimal value for k
  - Take the subset of genes with no missing gene expression information (the complete matrix)
  - Randomly delete 20% of the data from the complete matrix
  - Use kNN imputation to estimate the values for varying k
  - Calculate the error between the imputed matrix and the complete matrix



A.K. Kaushik et al., *Nature Communications*, 2016
O. Troyanskaya et al., *Bioinformatics*, 2001

Results

Pathways identified by Kaushik et al. (2016)

Pathway Name	P-value	
Riboflavin metabolism	<0.0001	(0.139)
Biotin metabolism	0.0148	(0.999, only 4 genes)
Amino sugar metabolism	0.0269	
Valine, leucine and isoleucine biosynthesis	0.0285	

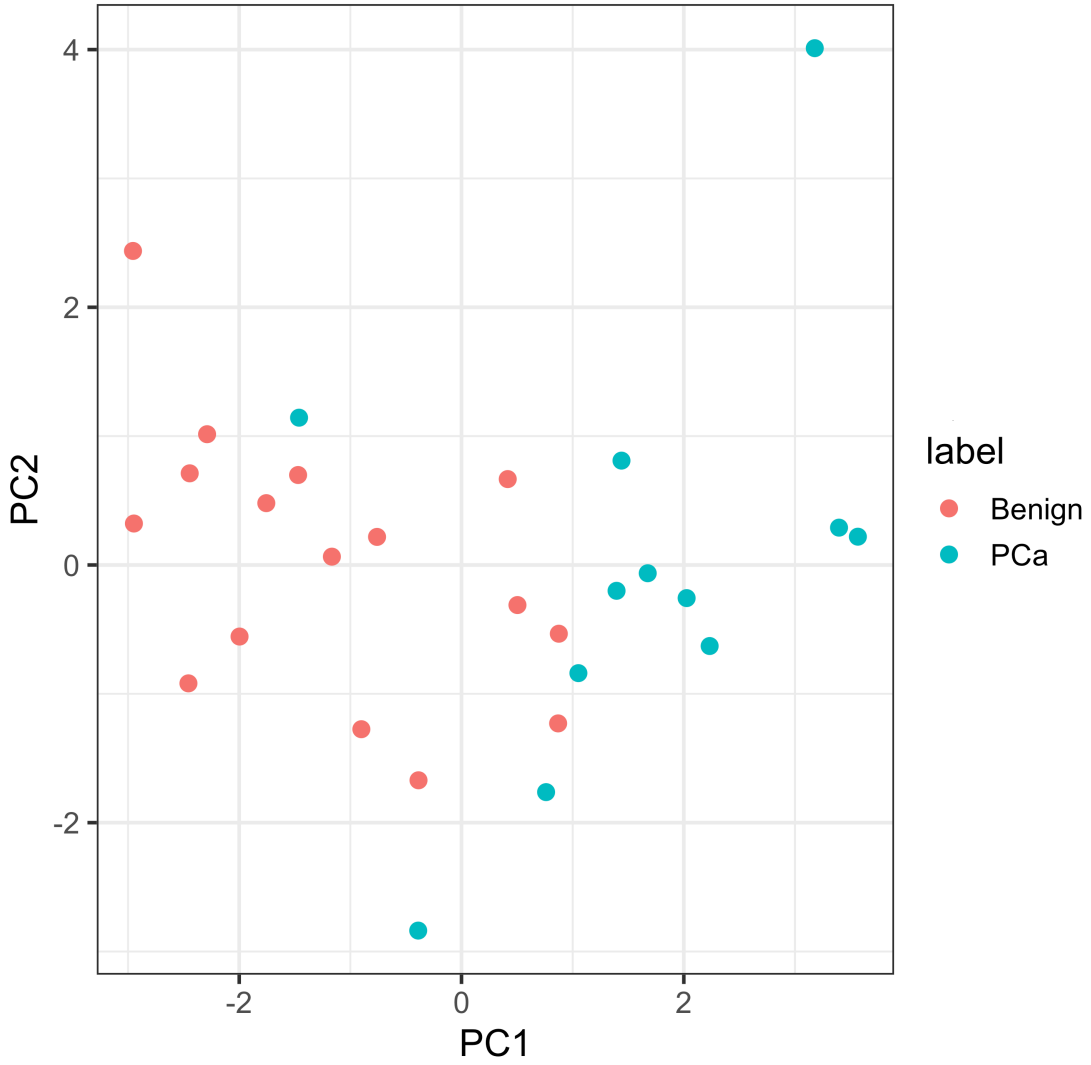
Pathways identified by our algorithm

Pathway Name	FDR
Valine, leucine and isoleucine biosynthesis	0
Glutathione metabolism	0.000999
Urea cycle and metabolism of amino groups	0.001998
Fatty acid metabolism	0.00599
Glycerophospholipid metabolism	0.00999
Oxidative phosphorylation	0.01098
Arginine and proline metabolism	0.01098
Butanoate metabolism	0.01098
Purine metabolism	0.01698
Nitrogen metabolism	0.01998
Fructose and mannose metabolism	0.02797
Aminosugars metabolism	0.03996004
Pyrimidine metabolism	0.04695305

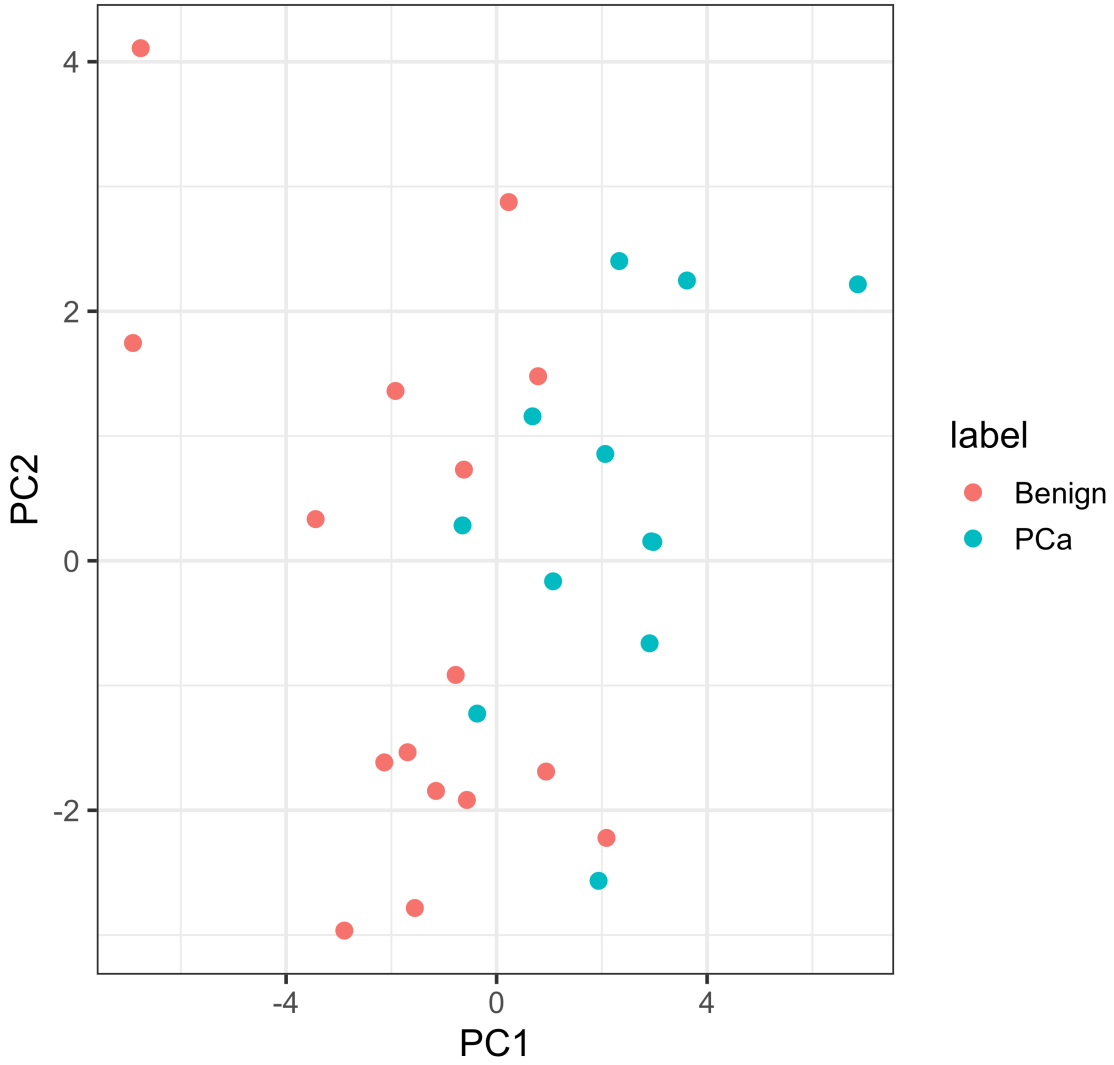
A.K. Kaushik et al., *Nature Communications*, 2016

PCA plots of top pathways from our algorithm

Valine, leucine and isoleucine biosynthesis (FDR = 0)



Aminosugars metabolism (FDR = 0.040)



Acknowledgements

Thanks to

Weiruo Zhang

Sylvia Plevritis

Plevritis Lab

Additional Slides

kNN Imputation

- About 20% of the gene expression data is missing, so we use k-nearest neighbors (kNN) imputation to estimate the missing data
  - Example: Patient 3 is missing data for Gene A
- Compute the Euclidean distance between Gene A and the other genes, using the gene expression data that is not missing for Gene A (i.e. for Patients 1, 2 and 4)
- Find the K closest genes (lowest Euclidean distance) which have a value present for Patient 3 (i.e. Gene C, but not Gene B)
- Estimate the missing gene expression for Gene A by using a weighted average of the gene expression of Patient 3 in the K closest genes
  - The more similar a gene's expression is to Gene A (the lower the Euclidean distance), the more weight the gene has.

	Patient 1	Patient 2	Patient 3	Patient 4
Gene A	-0.2758	0.0151	NA	-0.4138
Gene B	-0.0639	-0.1069	NA	-0.1892
Gene C	0.4853	-0.4678	-0.3163	-0.7069

Finding the Optimal Value for K

- Take the subset of genes with no missing gene expression information (the complete matrix)
- Randomly delete 20% of the data from the complete matrix
  - 20% of the information is missing from the original gene expression data
- Use kNN imputation to estimate the values for varying k
- Calculate the error between the imputed matrix and the complete matrix