

**New
Uzbekistan
University**

OCTOBER 2-4, 2026 · TASHKENT, UZBEKISTAN
2026 New Uzbekistan University – Soongsil University,
G-LAMP, International Workshop
Recent Research in AI: Mathematics, Statistics and
Chemistry

Deep-Learning Pathway Analysis using Hierarchical Structured Component Models

Taesung Park

Department of Statistics, Seoul National University, Korea

Interdisciplinary Program in Bioinformatics, Seoul National University, Korea

BIBS = Bioinformatics + Biostatistics



“The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information.”



– Frank Tekaiia



Bioinformatics is the field of science in which *biology*, *computer science*, and *information technology* merge into a single discipline

– National Center for Biotechnology Information 2001

What is Bioinformatics?

1. Bioinformatics is an interdisciplinary field that integrates biology, statistics, mathematics, and computer science
2. Its goal is to extract biological meaning from large-scale data
3. Central question:
How do we translate massive biological data into knowledge, insight, and impact?

Why bioinformatics?

1. Find an answer quickly

Most *In silico* biology is faster than *in vitro*

2. Massive amounts of data to analyze

- a. Need to make use of all information
- b. Not possible to do analysis by hand
- c. Automation is the key!

Why bioinformatics matters?

1. Modern biology produces data beyond human intuition
2. High-throughput technologies generate **millions to billions of measurements**
3. Without computational analysis, data remains **silent and unused**
4. Bioinformatics turns data into:
 - Biological understanding
 - Clinical decisions
 - Societal benefit

Major Data Types

- 1. Genomics:** DNA variants, whole genome / exome sequencing
- 2. Transcriptomics:** RNA-seq, single-cell RNA-seq
- 3. Proteomics:** protein expression, structure, interactions
- 4. Metabolomics:** metabolic profiles
- 5. Clinical & Epidemiological Data:** phenotypes, outcomes, longitudinal data

Typical Bioinformatics Workflow

1. Data Generation

- NGS, mass spectrometry, imaging

2. Preprocessing & Quality Control

- Filtering, normalization, alignment

3. Statistical Modeling & Algorithms

- High-dimensional inference, machine learning

4. Biological Interpretation

- Pathways, mechanisms, validation

Statistical & Computational Foundations

1. High-dimensional statistics

- Regularization (Lasso, Ridge)
- Mixed and hierarchical models

2. Machine Learning / AI

- Classification, prediction, representation learning

3. Network & systems modeling

- Gene–gene and protein–protein interactions

Data complexity demands methodological humility and rigor.

Machine Learning & AI in Bioinformatics

1. Supervised learning

- a. Classification, risk prediction

2. Unsupervised learning

- a. Clustering, latent structure discovery

3. Deep learning

- a. Sequence modeling

- b. Protein structure and interaction prediction

4. Emphasis on **interpretability** and **generalizability**

Key Application Areas

1. Precision Medicine

- Personalized diagnosis and treatment

2. Cancer Genomics

- Driver mutations, molecular subtypes

3. Infectious Disease Modeling

- Transmission dynamics, variant evolution

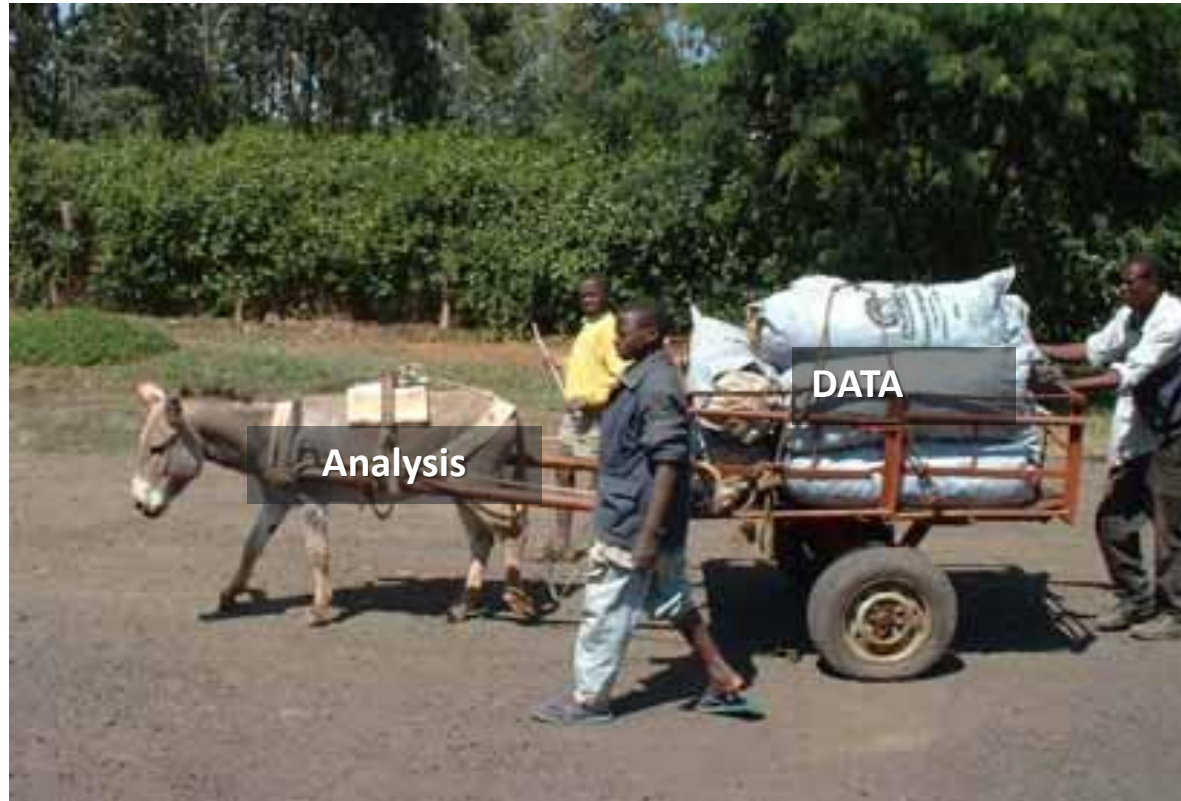
4. Drug Discovery

- Target identification, structure prediction

5. Systems Biology

- Multi-omics integration

Bioinformatics

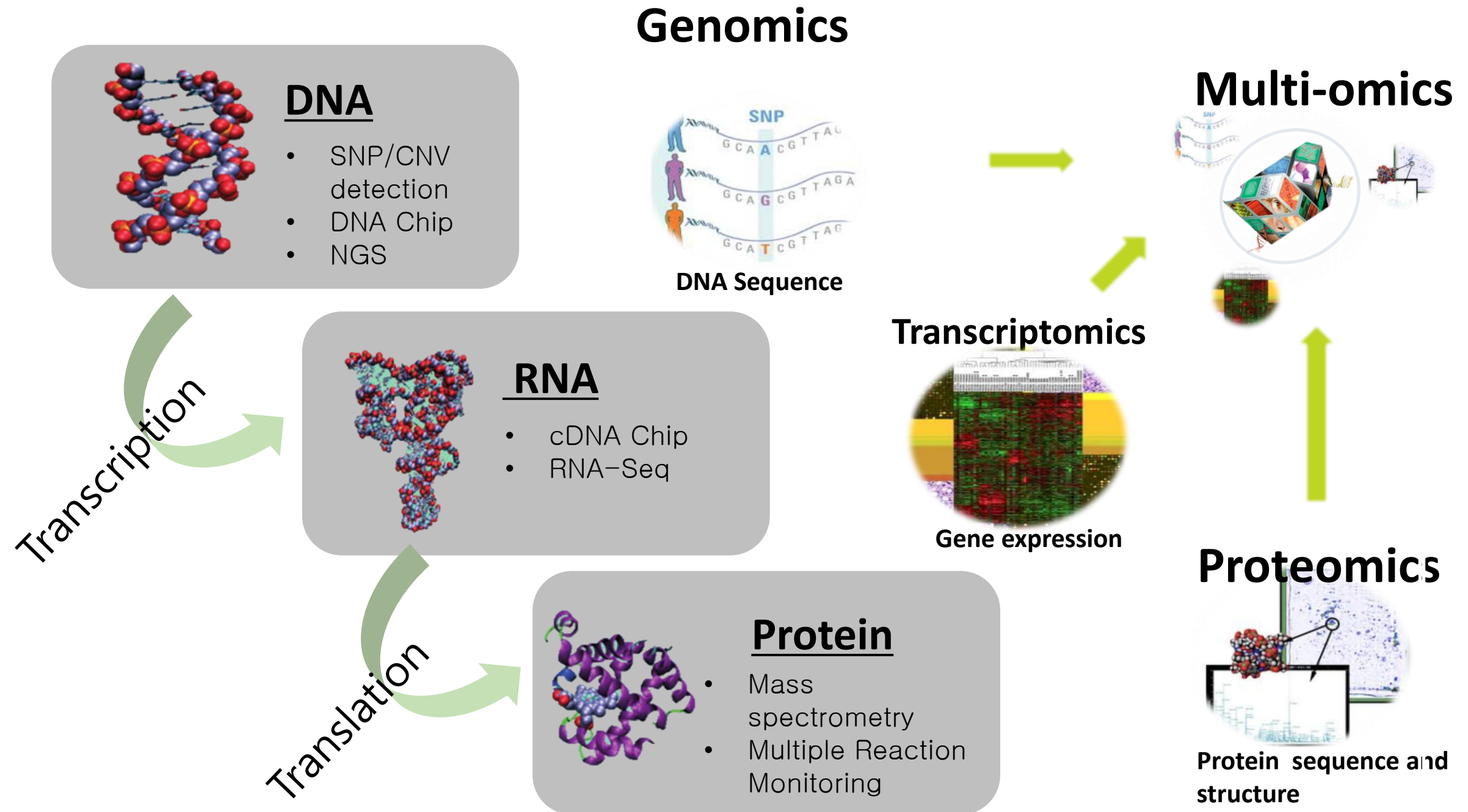


- High-Dimensional Omics Data

Data Generation < **Analysis** < **Interpretation**

\$1,000 < **\$100,000** < **\$1,000,000**

Bioinformatics



Omics

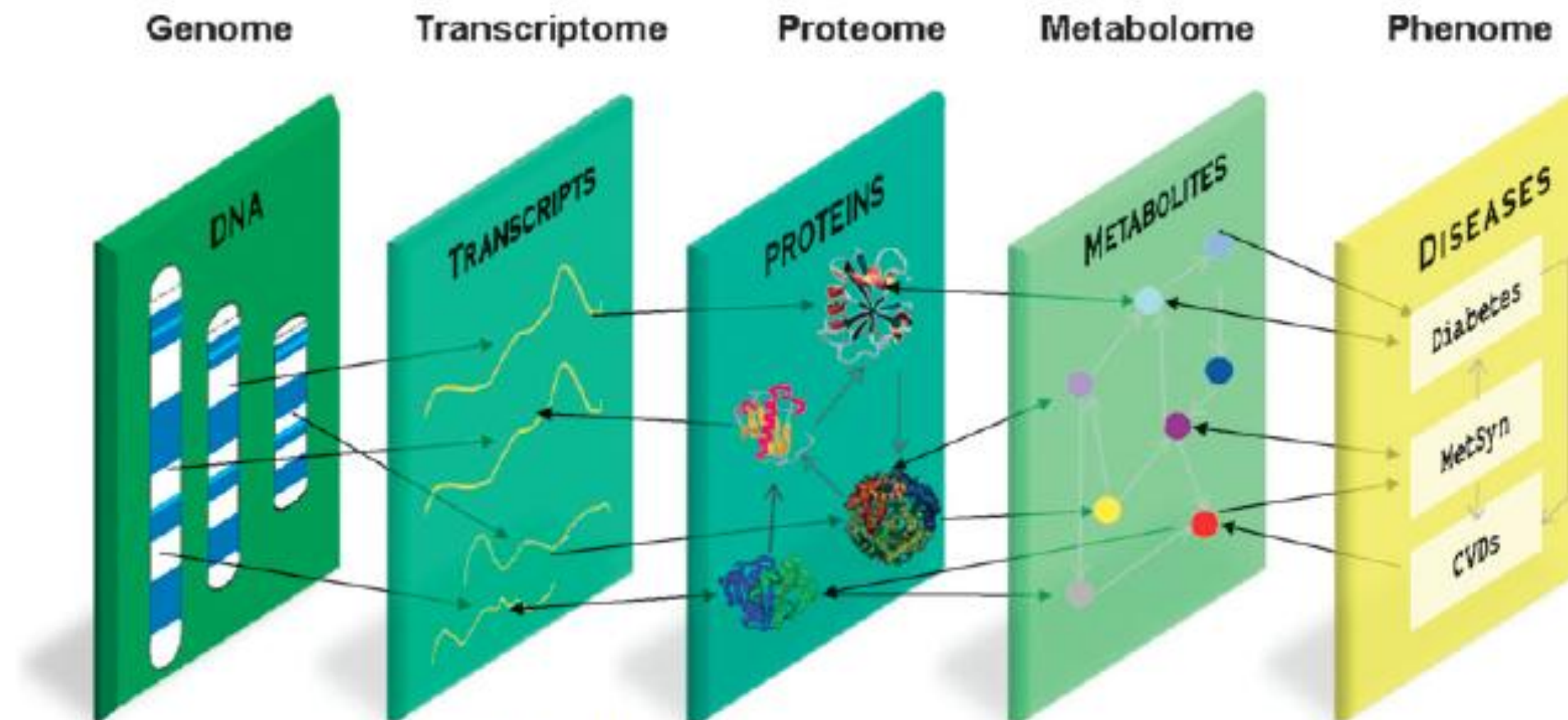
1. OMICS – “Totality” of specific fields

- a. Genomics : “the study of genes and their function” (Human Genome Project (HGP) 2003)
- b. Proteomics : the study of proteins
- c. Transcriptomics : the study of mRNA

2. Bioinformatics has become an essential part of omics research and requires unique practical and analytical skills for appropriate results interpretation

Bioinformatics

A systems perspective of complex disease



Contents

- 1 · **Introduction**
- 2 · **Method**
- 3 · **Real Data Analysis**
- 4 · **Simulation Study**
- 5 · **Discussion**

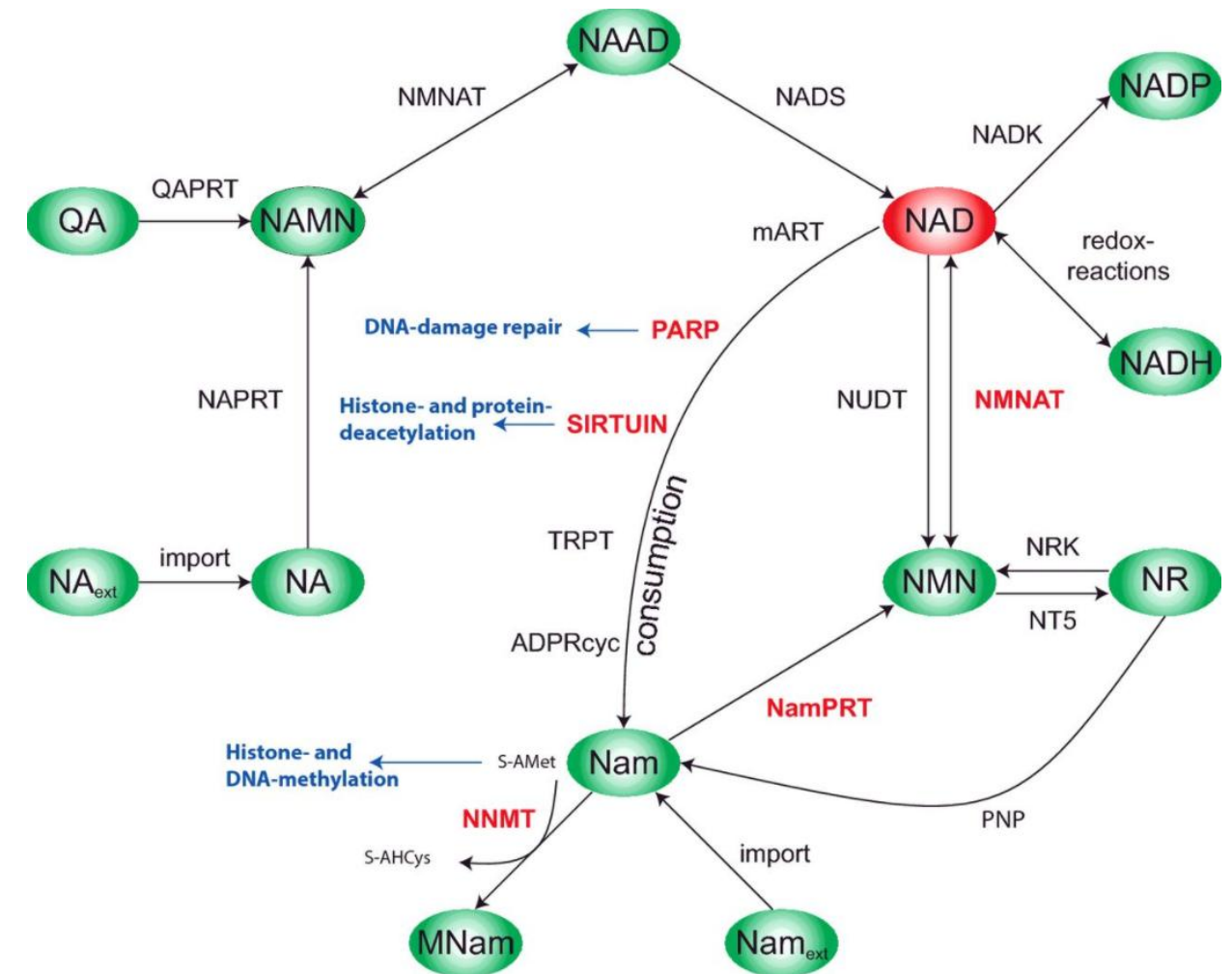


Introduction

Statistical Significance vs Biological Interpretation

Biological pathways

- Genes do not work alone, but in an intricate network of interactions
- Series of actions among molecules in a cell that leads to a certain product or a change in the cell
- Trigger the assembly of new molecules, such as a fat or protein



Biological pathways

- Well annotated databases available
 - **KEGG (Kanehisa *et al.*, 2004)**
 - Biocarta (www.biocarta.com)
 - MSigDB (Tamayo *et al.*, 2005)
 - WikiPathways (Pico *et al.*)
 - and etc.

KEGG:
Kyoto Encyclopedia of
Genes and Genomes



Pathway-based approach to disease



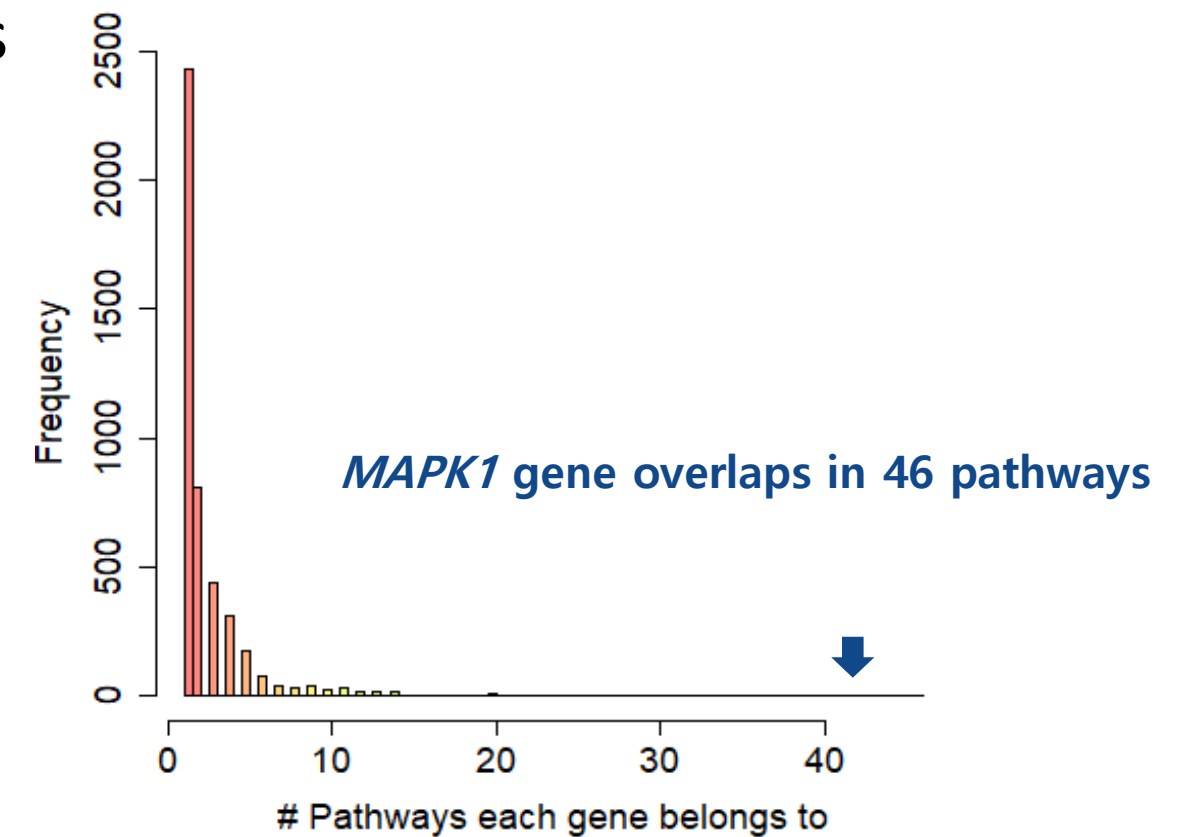
- Effectively explore human diseases by studying biological pathways; providing intuition and clues about relationship between biological factors and disease
- In particular, pathway analysis plays an important role in understanding the mechanism of complex diseases
- For example, cancer, which is complex disease that is usually caused by multiple disorder gene mutations or pathways
- Because of these advantages, pathway-based methods for different types of omics data analysis have been actively developed.

Pathway-based approach to disease

Omics data	Methods
Transcriptomic (i.e. gene expression)	• Gene Set Enrichment Analysis (GSEA) • sparse Linear Discriminant Analysis (sLDA) • Logistic kernel machine test • GlobalTest, GlobalAncova, etc • Pathome
Genomic (i.e. SNP)	• SNP Set Enrichment Analysis (SSEA) • Association List Go AnnoTatOR (ALIGATOR) • SNP ratio test (SRT), etc • SNP-PRAGE
Metabolomic	• Metabolite Set Enrichment Analysis (MSEA) • Network-based Metabolite Pathway Identification (MPINet), etc

Pathway-based approach to disease

- Most of pathway-based methods consider a single pathway at a time.
- Multiple biomarkers are shared between pathways making pathways to be highly correlated.
 - Example showing that multiple genes are shared between pathways.
- A biological process is a result of complex entangled interactions among pathways.
- The lack of method to consider all pathways simultaneously and to provide a systematic insight into the biological processes.



- ✓ Genes in TCGA-Pancreatic adenocarcinoma
- ✓ 186 pathways based KEGG database

- **Hierarchical structured CoMponent analysis (HisCoM)**
 - Biologically hierarchical structure
 - Multiple pathway analysis
 - All pathways are analyzed simultaneously in an augmented model
 - Model-based analysis
 - Assume a **linear relationship** between latent variables (pathways) and biological factors
 - However, biological processes may not be well modelled using linear models

Objective

- Propose **deep-learning** pathway analysis based on hierarchical structured component models assuming **non-linear relationship** between pathways and biological factors

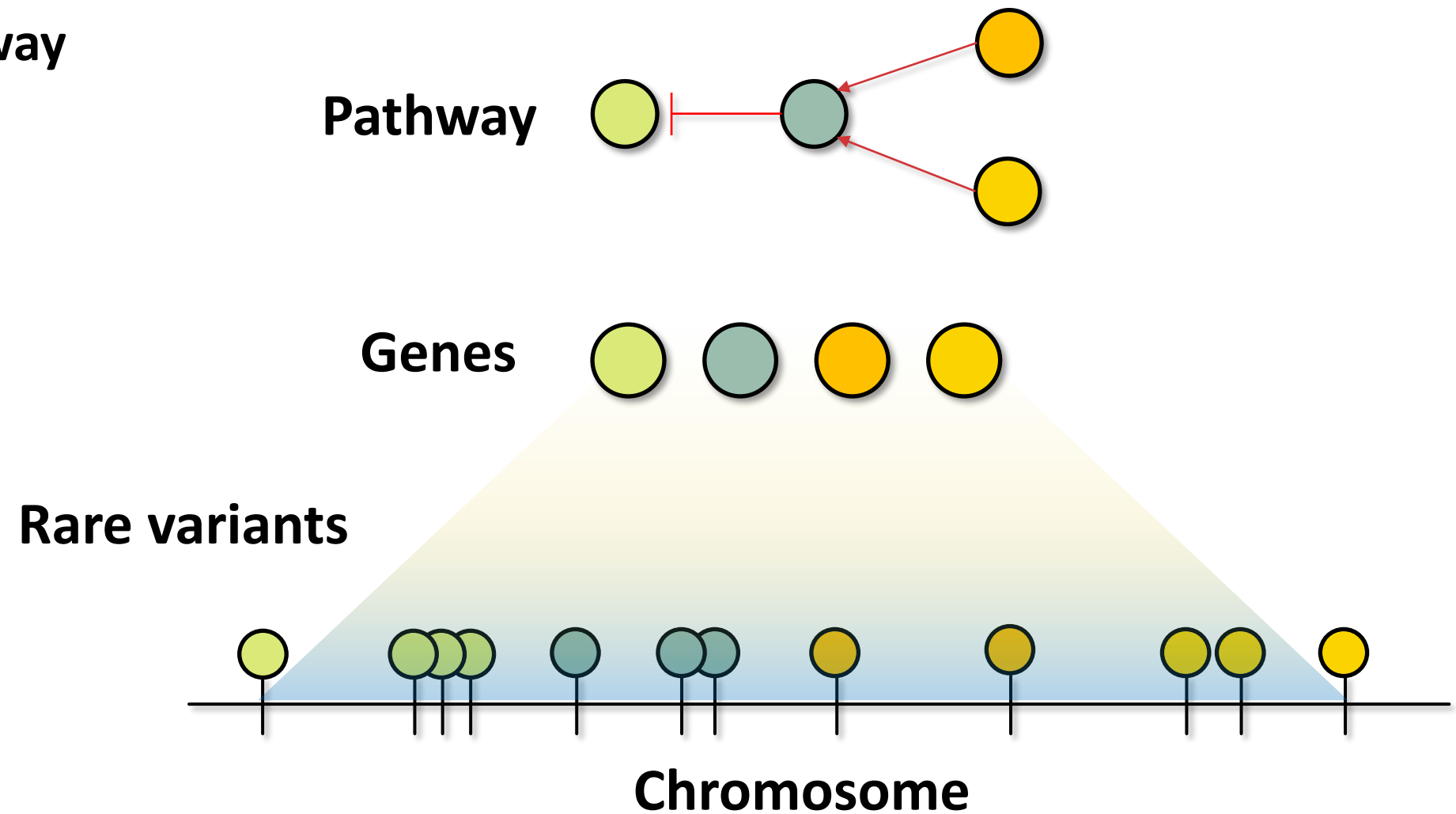
Deep - Hierarchical structured Component Model (Deep-HisCoM)

- **Deep-HisCoM** can identify non-linear trend between pathway and biological factors based on the state of the art machine learning method



Methods

- **Earlier work: HisCoM (PARAOH)**
 - ✓ Hierarchical Structures
 - ✓ Rare variants to a pathway



- **Earlier work I: HisCoM (PARAOH)**

✓ Rare variants to a pathway

G

SNP 1

SNP 2

SNP 3

SNP 4

SNP 5

SNP 6

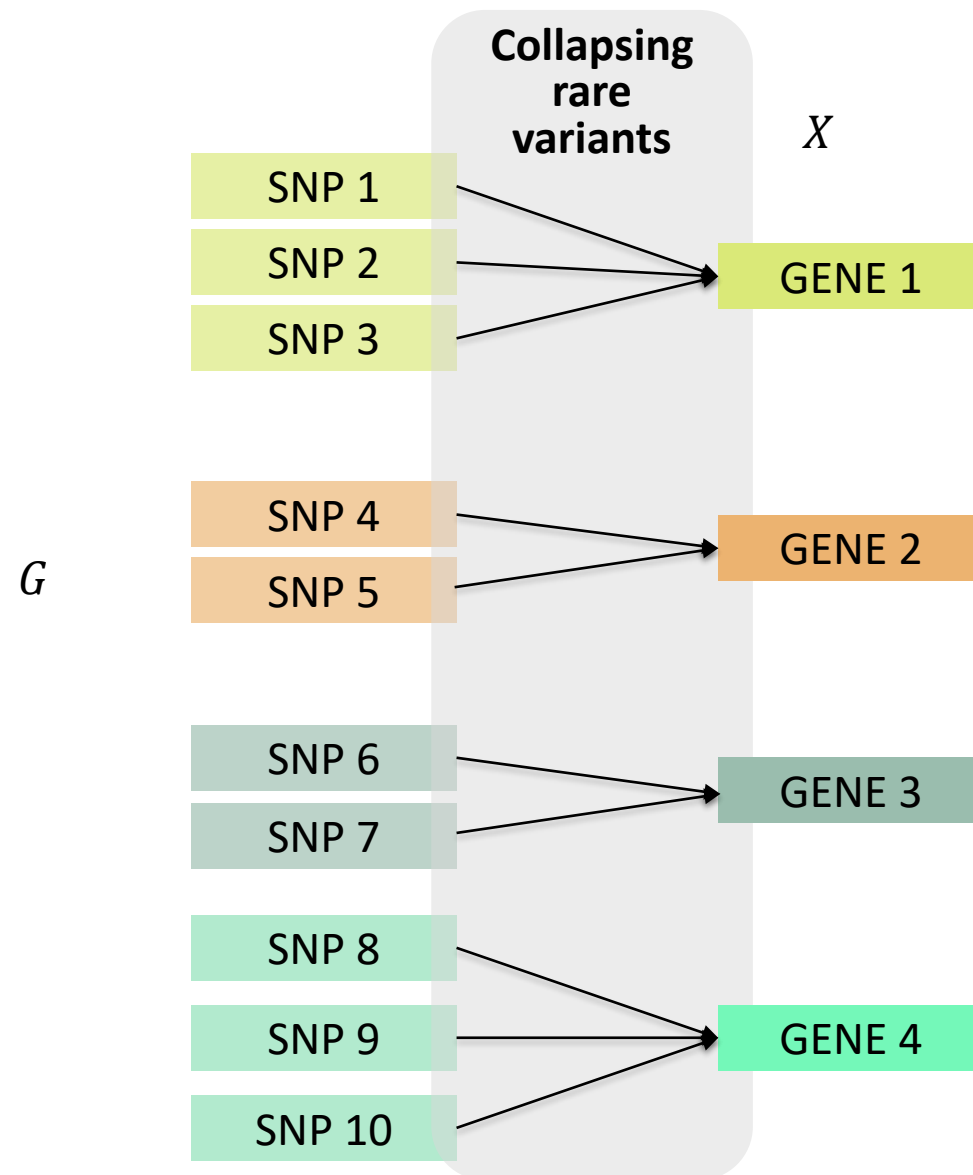
SNP 7

SNP 8

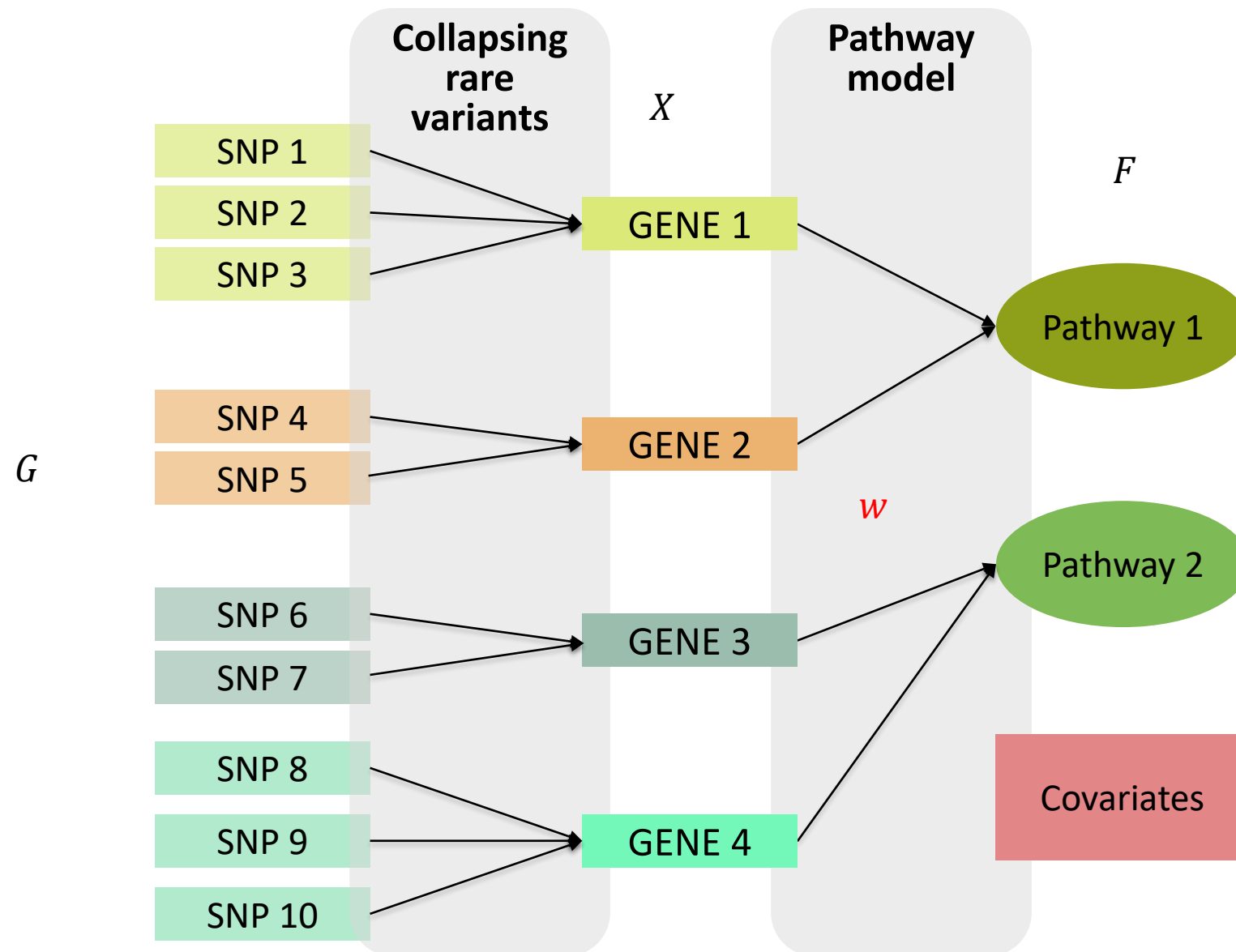
SNP 9

SNP 10

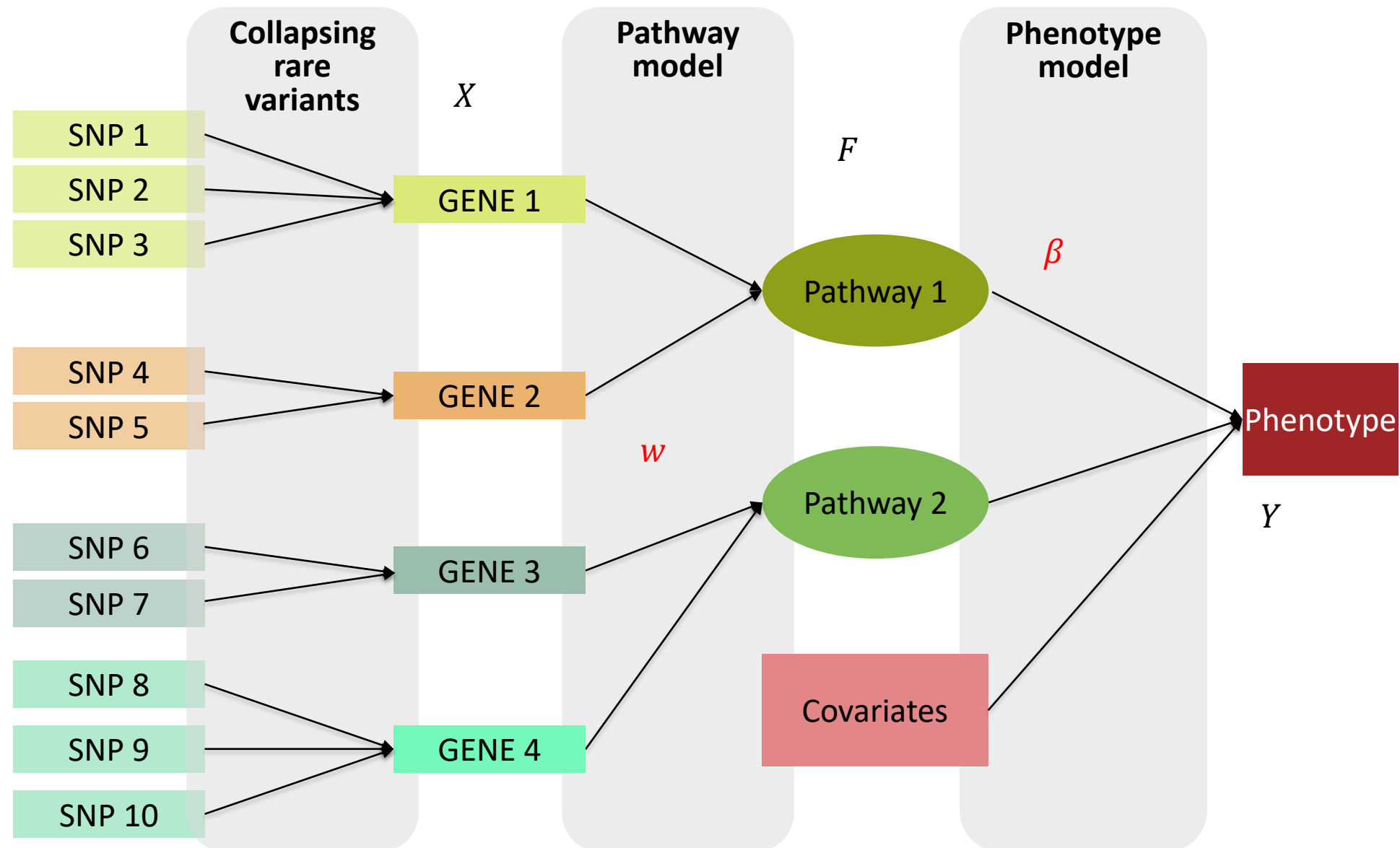
- Earlier work I: HisCoM (PARAOH)



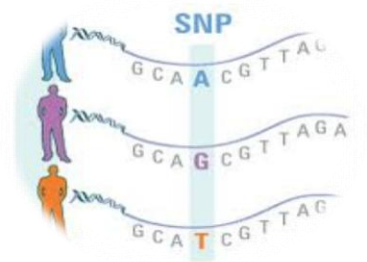
- Earlier work I: HisCoM (PARAOH)



- Earlier work I: HisCoM (PARAOH)



Genomics



- **Earlier work I: HisCoM (PARAOH)**

- ✓ The relationship between a linear predictor and a link function

$$\begin{aligned}\eta_j &= \beta_0 + \sum_{k=1}^K \left[\sum_{t=1}^{T_k} x_{jkt} w_{kt} \right] \beta_k \\ &= \beta_0 + \sum_{k=1}^K f_{jk} \beta_k \\ &= \sum_{k=0}^K f_{jk} \beta_k = g(\mu_j)\end{aligned}$$

- f_{jk} : j^{th} observation's score of the k^{th} pathway
- Using a canonical link function
 - Identity $\Rightarrow$ For the normal distribution
 - Logit $\Rightarrow$ For the binomial distribution
 - Log $\Rightarrow$ For the Poisson distribution
 - etc.





- **Earlier work I: HisCoM (PARAOH)**

- To estimate parameters, we seek to maximize a penalized log-likelihood function with ridge parameters

$$\phi_1 = \sum_{j=1}^N \log p(y_j; \gamma_j, \delta) - \frac{1}{2} \lambda_G \sum_{k=1}^K \sum_{t=1}^{T_k} w_{kt}^2 - \frac{1}{2} \lambda_P \sum_{k=1}^K \sum_{t=1}^{T_k} \beta_k^2$$

- where $p(y_j; \gamma_j, \delta)$: the probability distribution for y_j ,

$$p(y_j; \gamma_j, \delta) = \exp \left(\left(y_j \gamma_j - \xi(\gamma_j) \right) / \zeta(\delta) + v(y_j, \delta) \right)$$
- respect to w_{kt} and β_k , subject to the conventional scaling constraint $\sum_{i=1}^N f_{jk}^2 = N$ (Takane and Hwang, 2005)
- λ_G & λ_P : ridge parameters for genes and pathways



- **Earlier work I: HisCoM (PARAOH)**

- Pathway-based analysis using **Hier**archical component of collapsed **RA**re variants **Of** **H**igh-throughput sequencing data

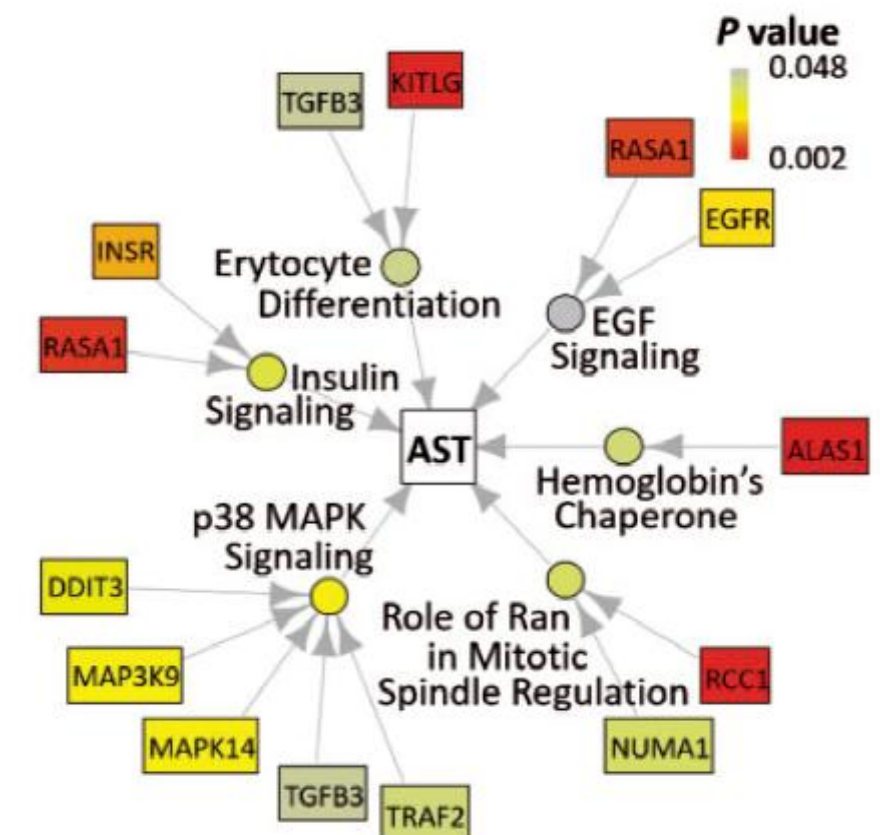
- ✓ Collapsing step

- ✓ One hierarchical model

- ✓ Analyzes entire genes and pathways simultaneously

- Inference

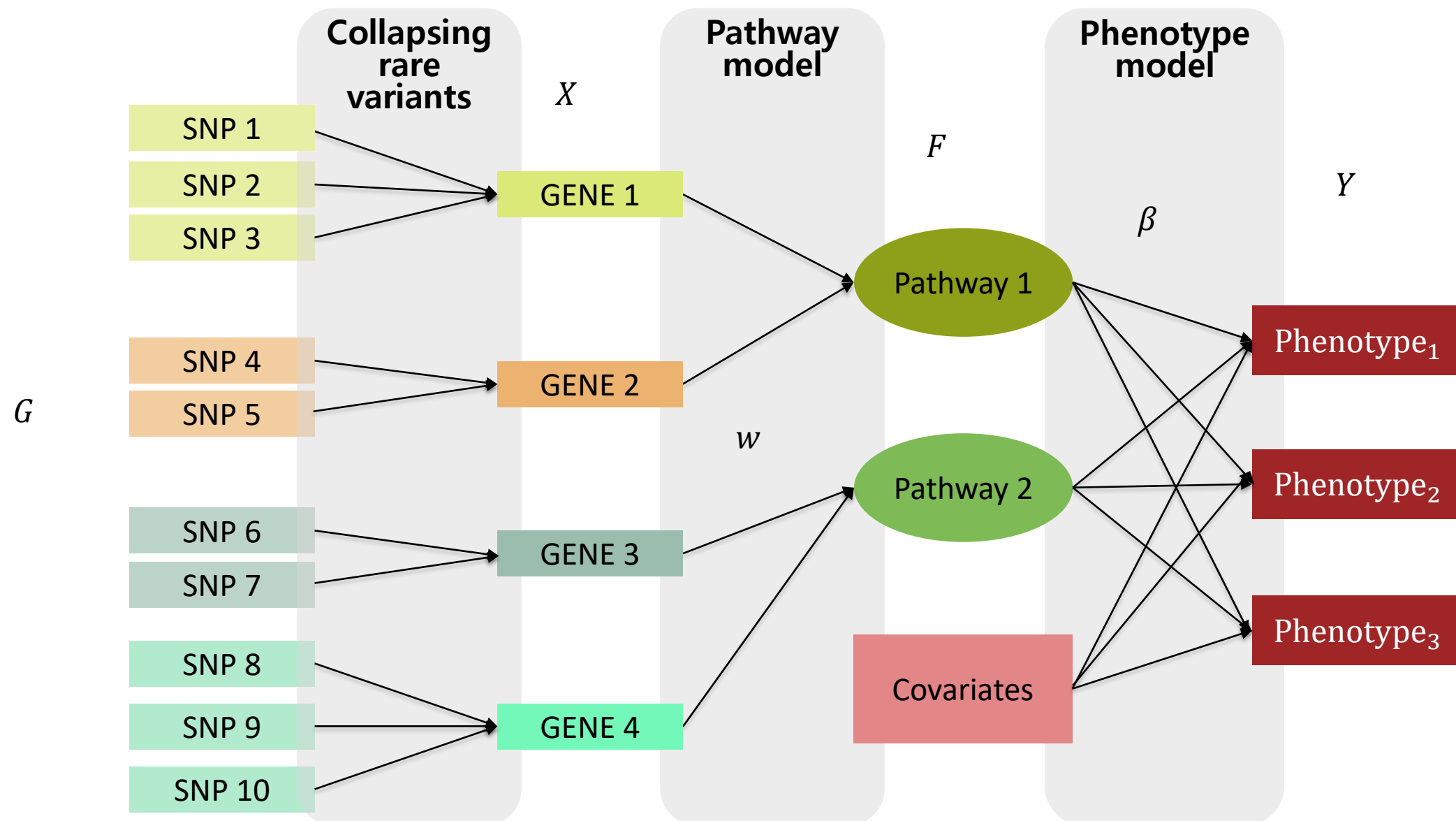
- Generalized Structuralized Component Analysis
 - Hwang and Takane (2004)
 - Alternating weighted least square estimation method
 - Significance test via permutation



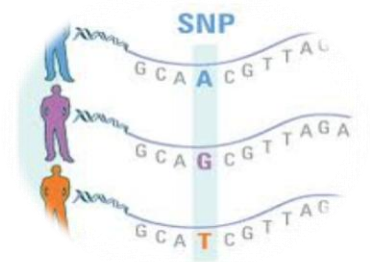
Lee *et al.*, Bioinformatics (2016)

- Earlier works II, III

- PHARAOH-multi, PHARAOH-GEE

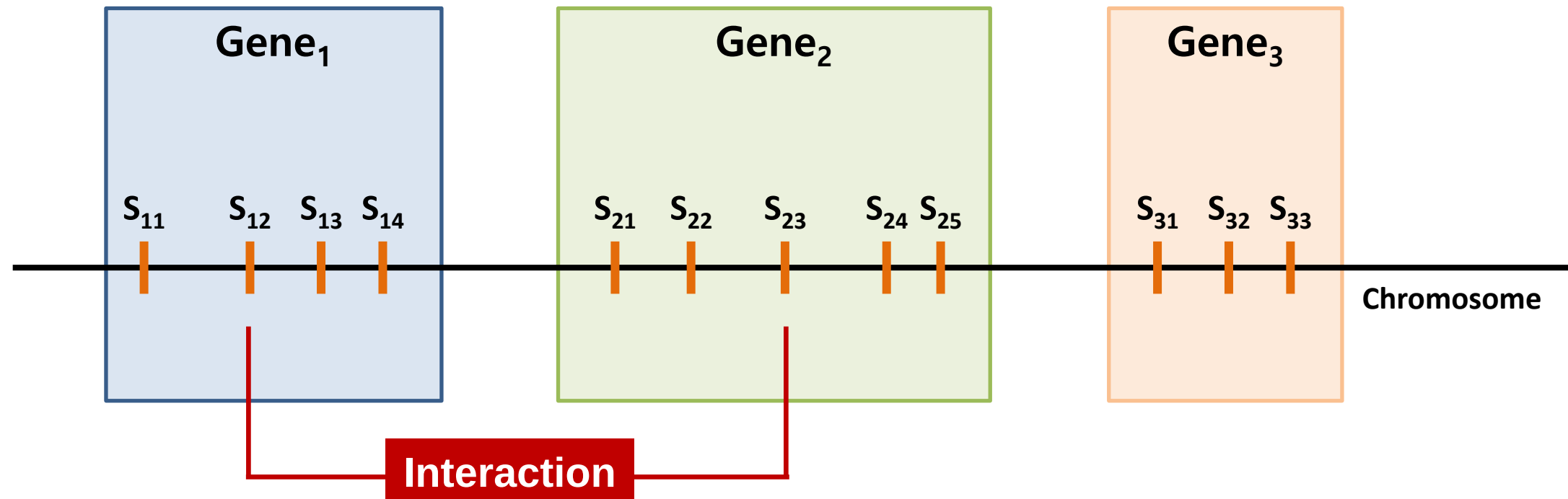


Genomics



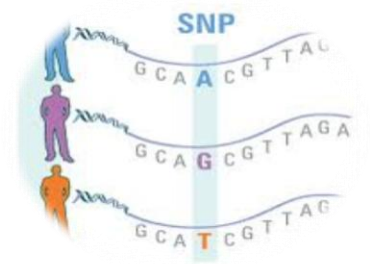
- Earlier works IV

- Gene-gene interaction



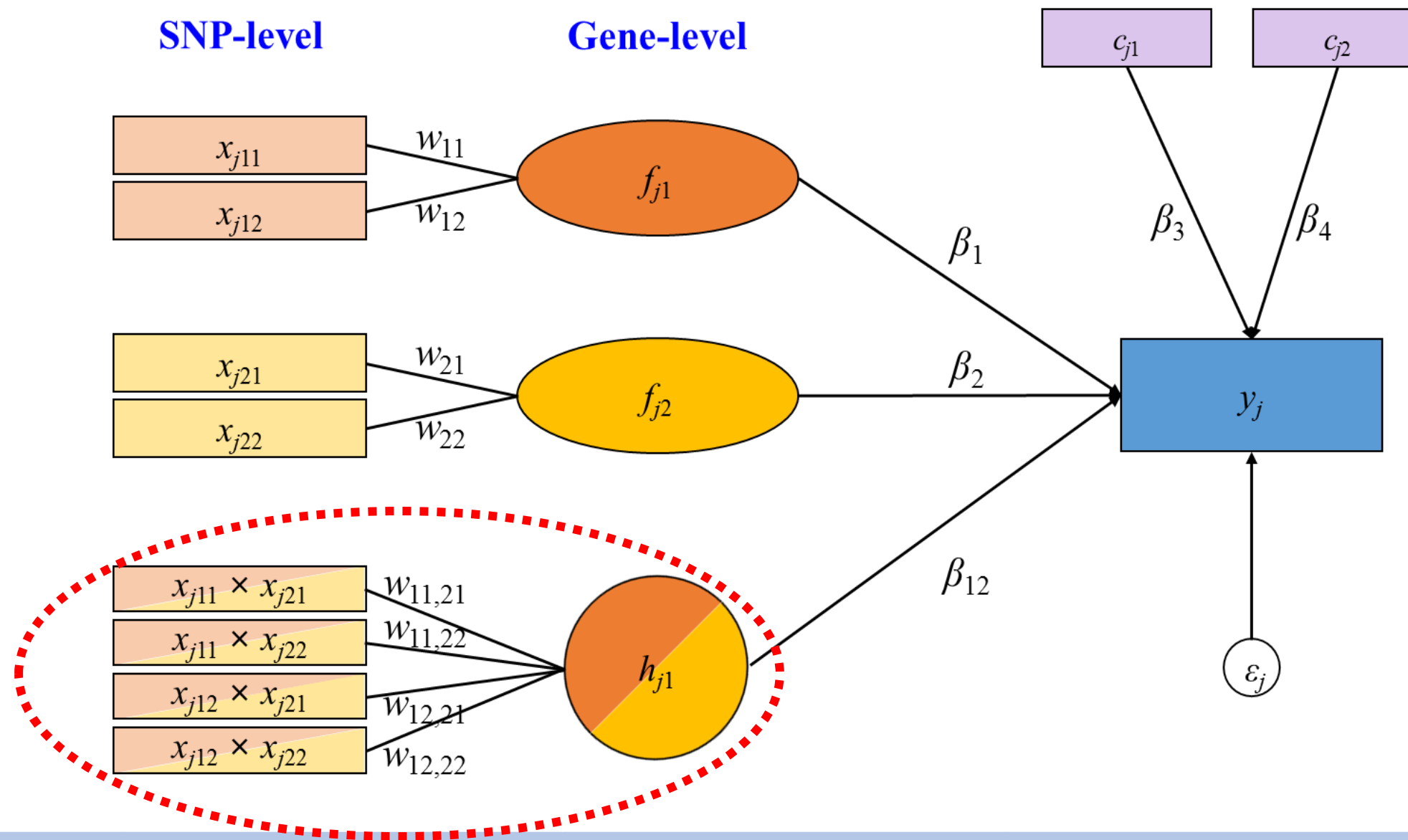
- SNP based SSI analysis: $_{12}C_2 = 66 \rightarrow {}_{1M}C_2 = 5.00 \times 10^{11}$
- Gene based GGI analysis: $_3C_2 = 3 \rightarrow {}_{25K}C_2 = 3.12 \times 10^8$

Genomics

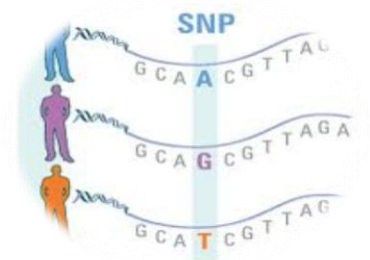


- Earlier works IV

- Gene-gene interaction



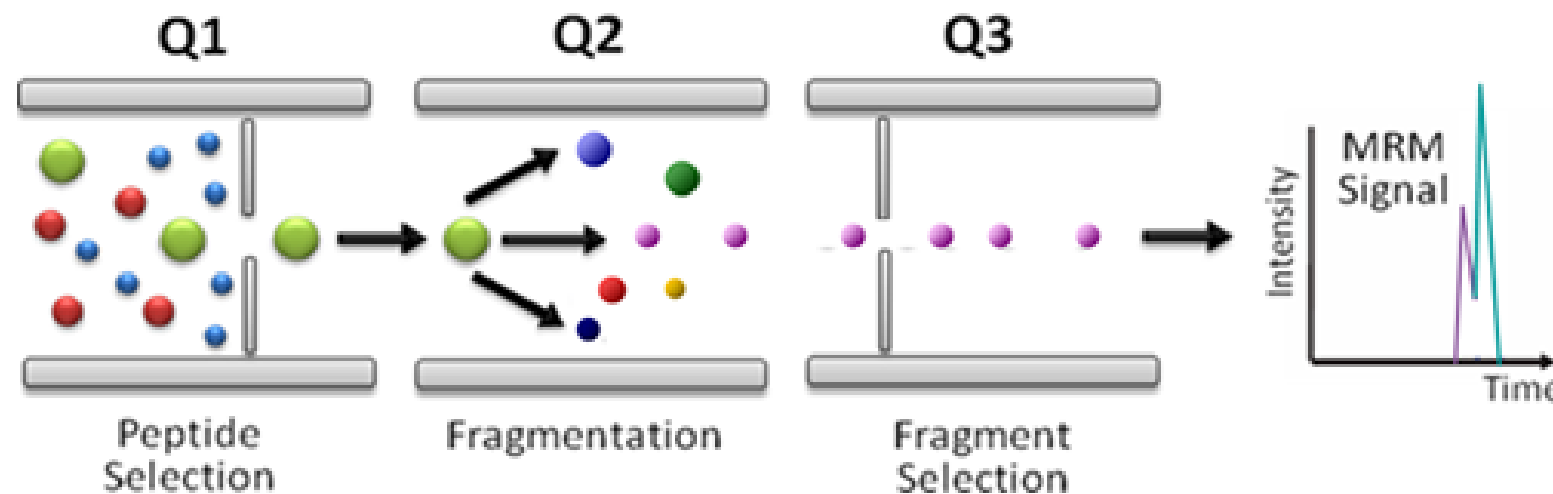
Genomics



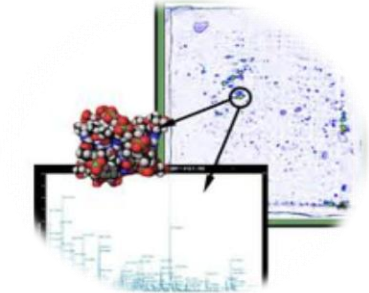
- **Earlier works V**

- **HisCoM-PIP**

- **Multiple Reaction Monitoring (MRM)** quantifies protein or peptides in biological samples
 - Peptides are detected by mass spectrometry analysis and the exact concentration can be determined
 - It is a highly sensitive and selective method.
 - How to take into the biological structures into the prediction model
 - **One protein with multiple peptides**

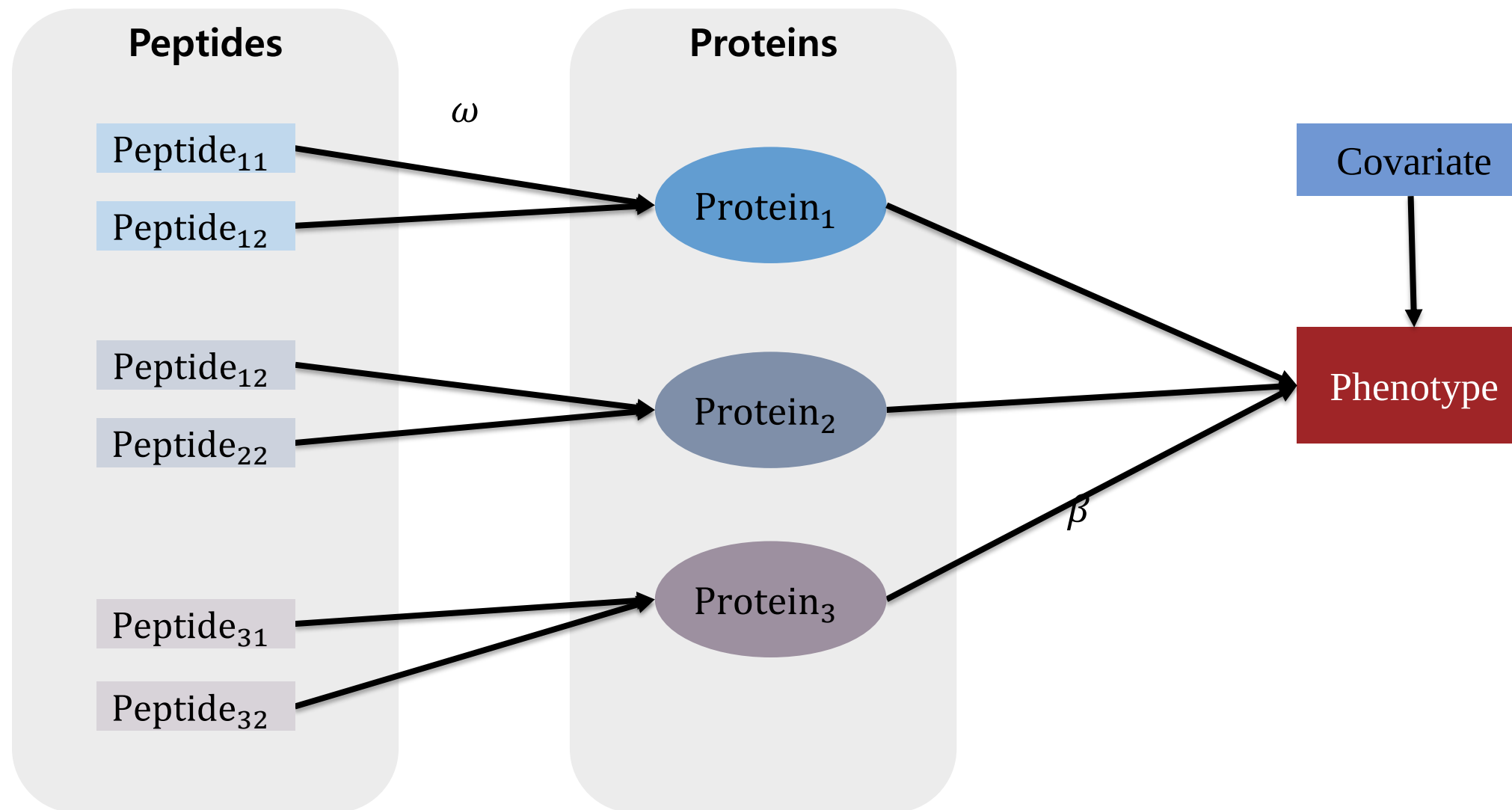


Proteomics

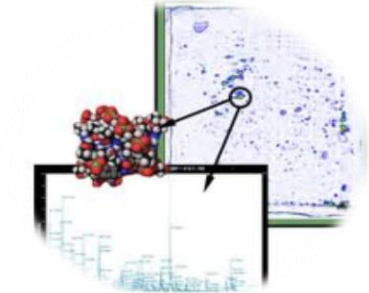


- Earlier works V

- HisCoM-PIP

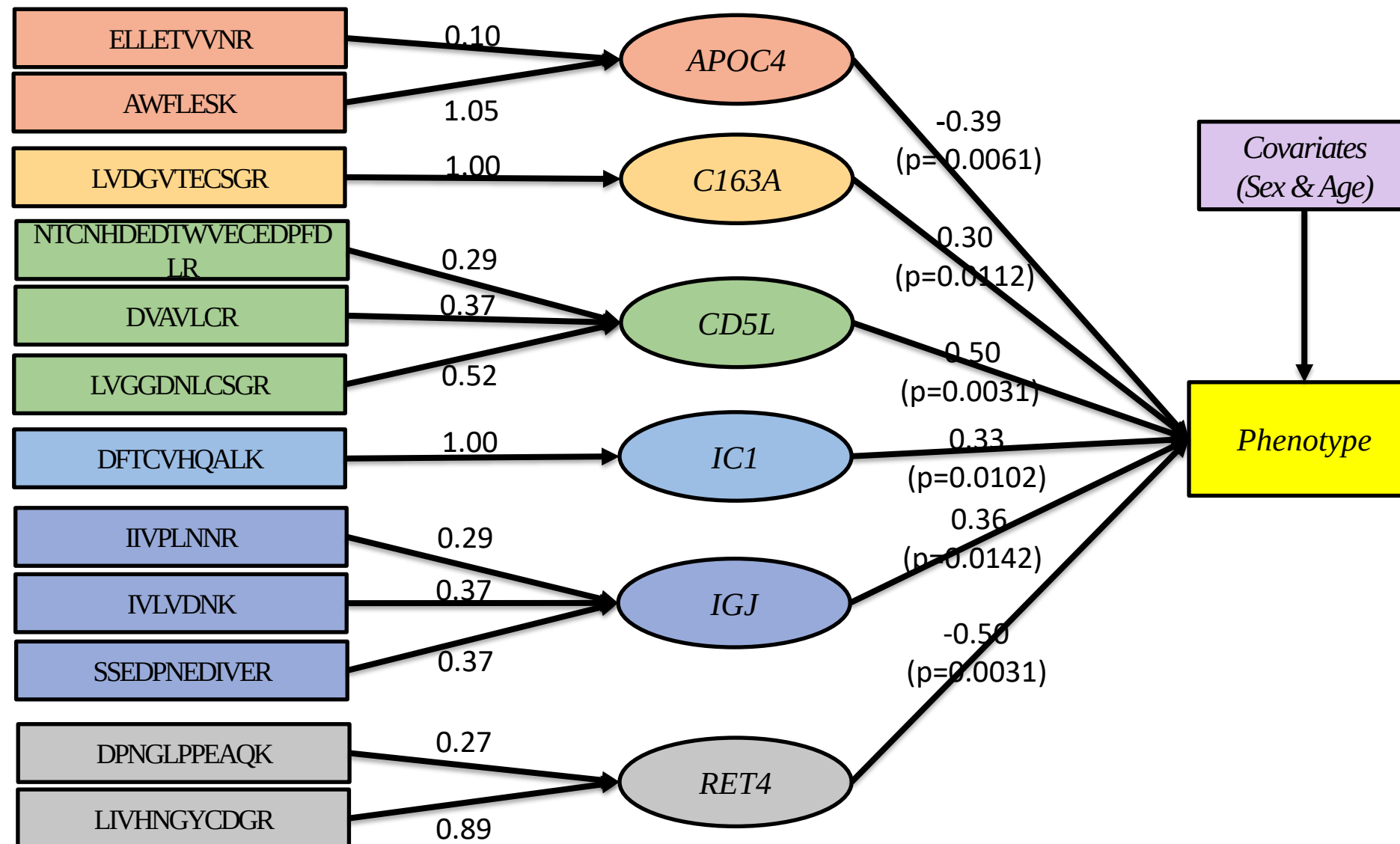


Proteomics

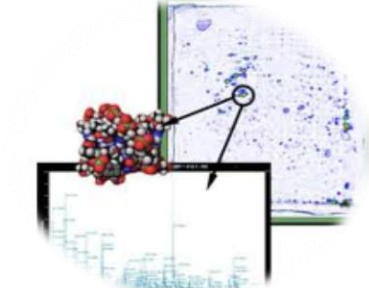


- Earlier works V

- HisCoM-PIP



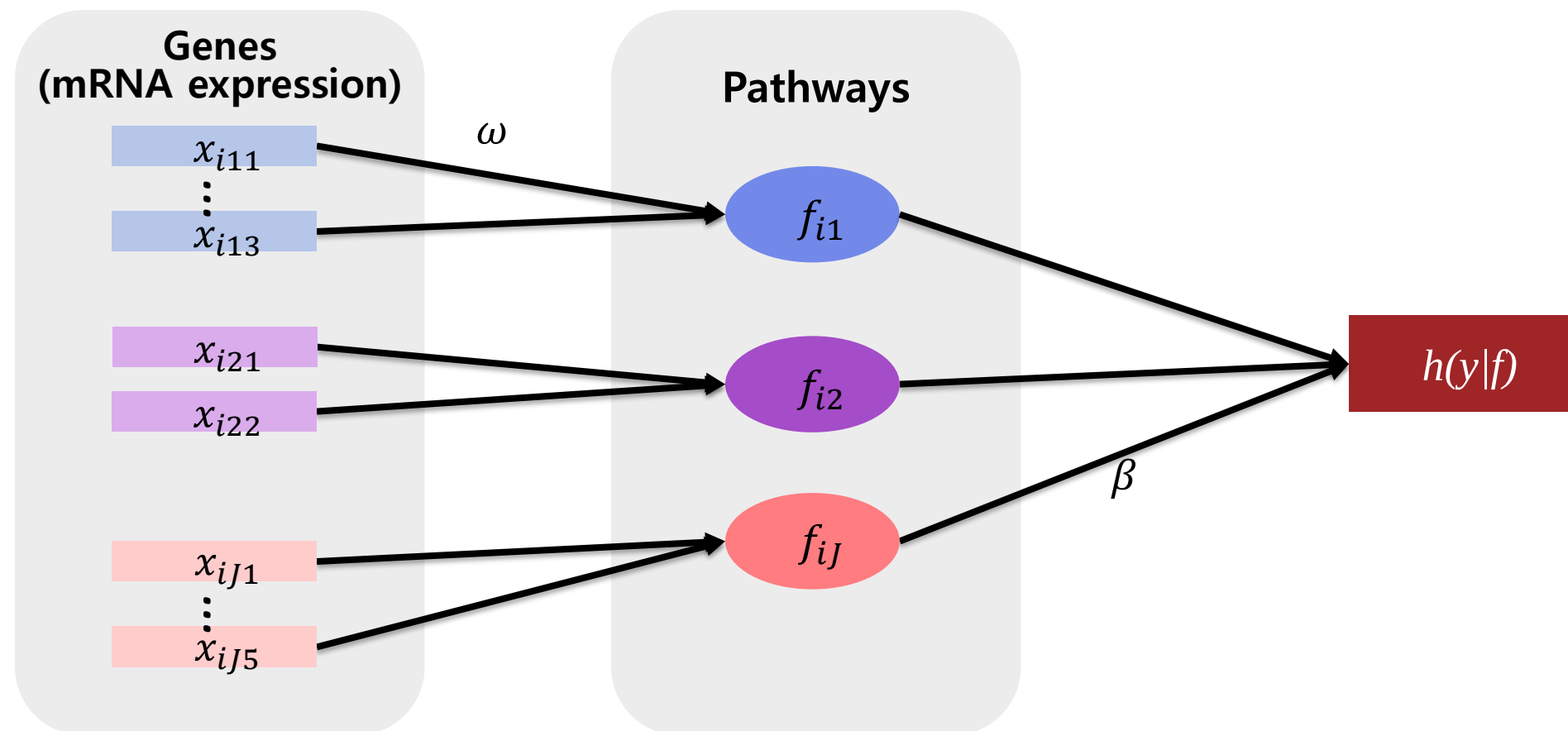
Proteomics



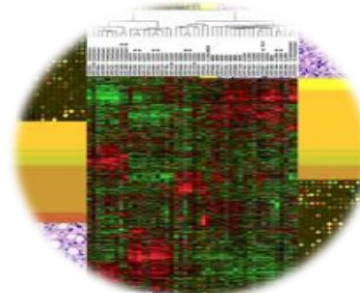
• Earlier works VI

➤ HisCoM-PAGE

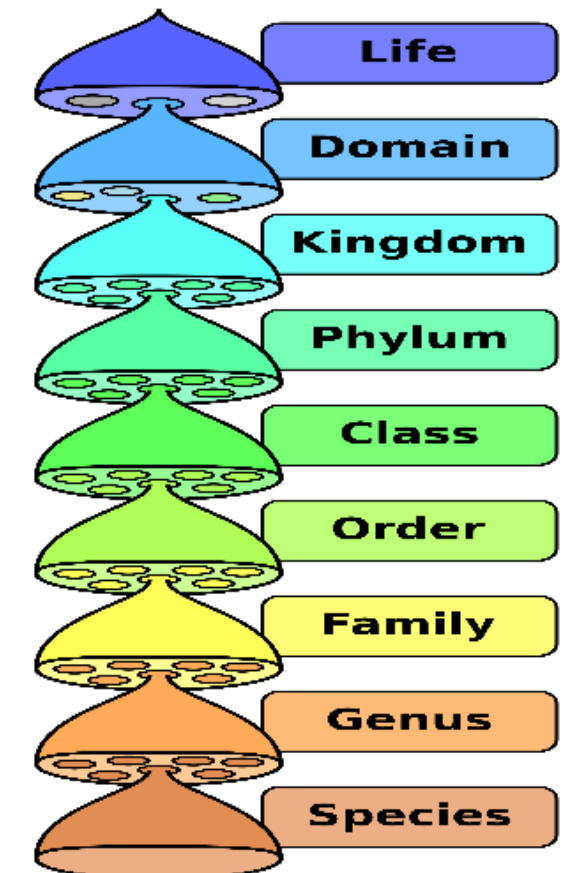
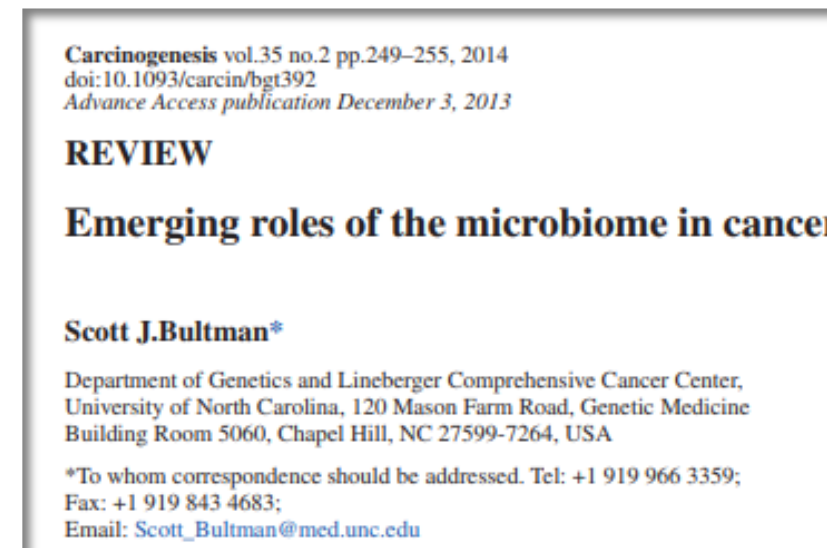
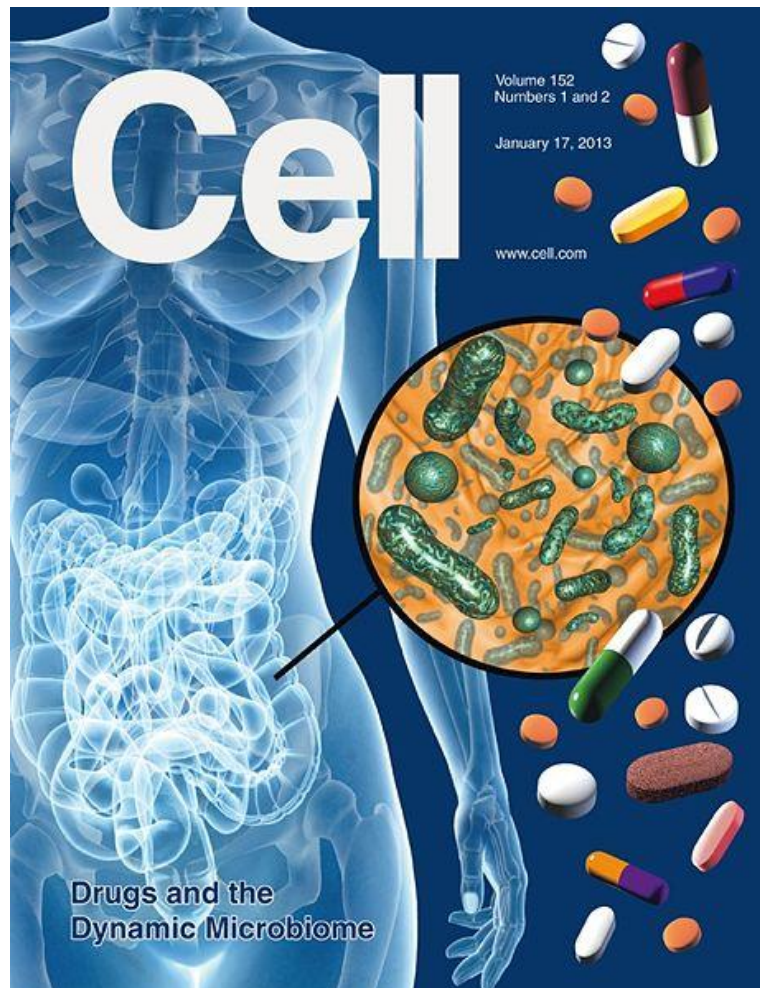
- Pathway-based analysis for *gene expression* for disease.
- Hierarchical component structures
- Identify *genes* and *pathways* associated with a specific trait using a statistical approach.



Transcriptomics

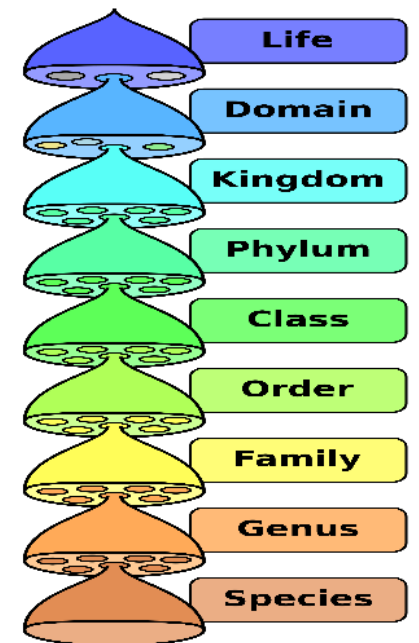
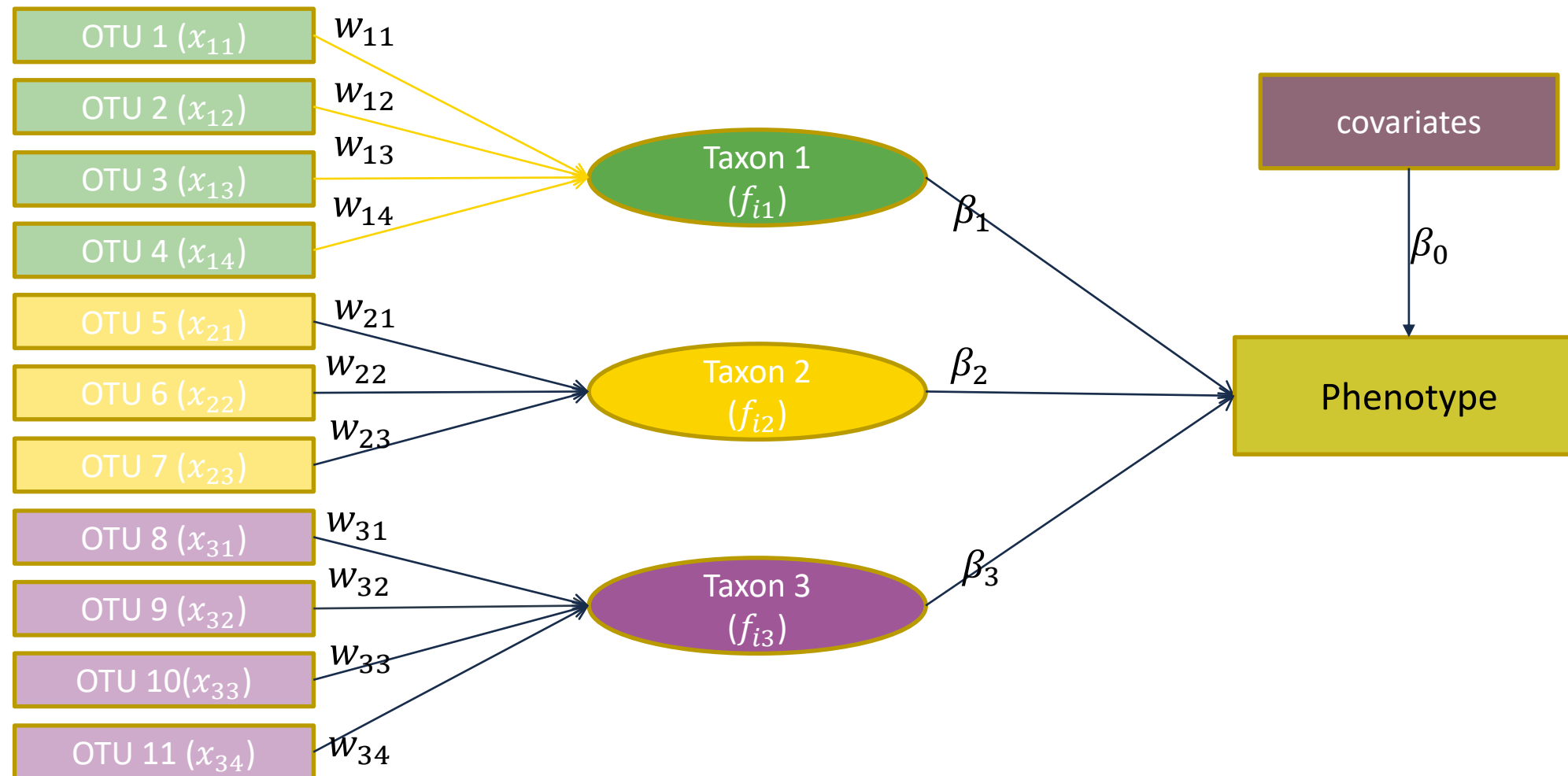


- Earlier works VII
 - Microbiome



- Earlier works VII

- Microbiome: HisCoM-microb



Metagenomics



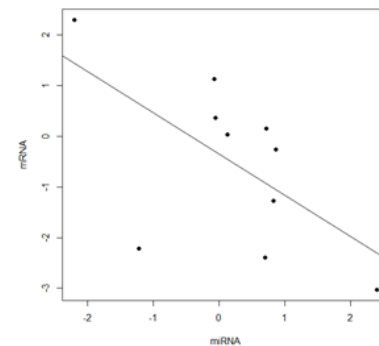
- Earlier works VIII

- HisCoM-mimi

Construct
miRNA-mRNA
Network

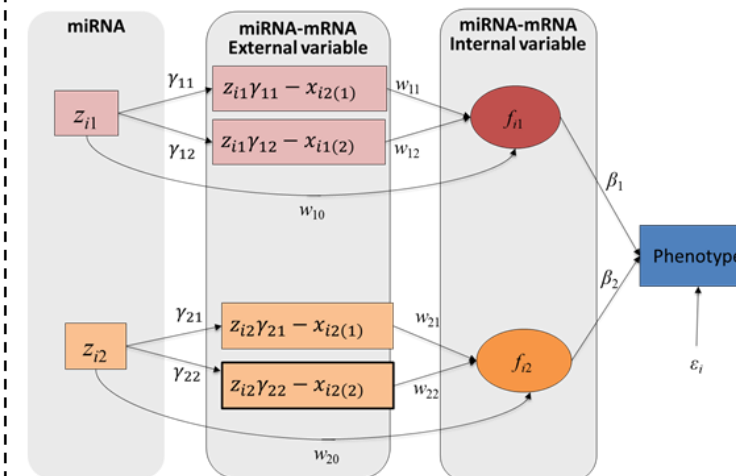


Collect predicted mRNA targets
of miRNAs by TargetScan



Find significant negative linear
regression coefficients between
miRNA and target mRNAs.

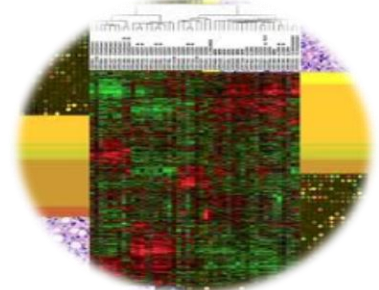
Find
significant
Networks with
phenotype



Find significant path and
weight coefficients by mimi
algorithm

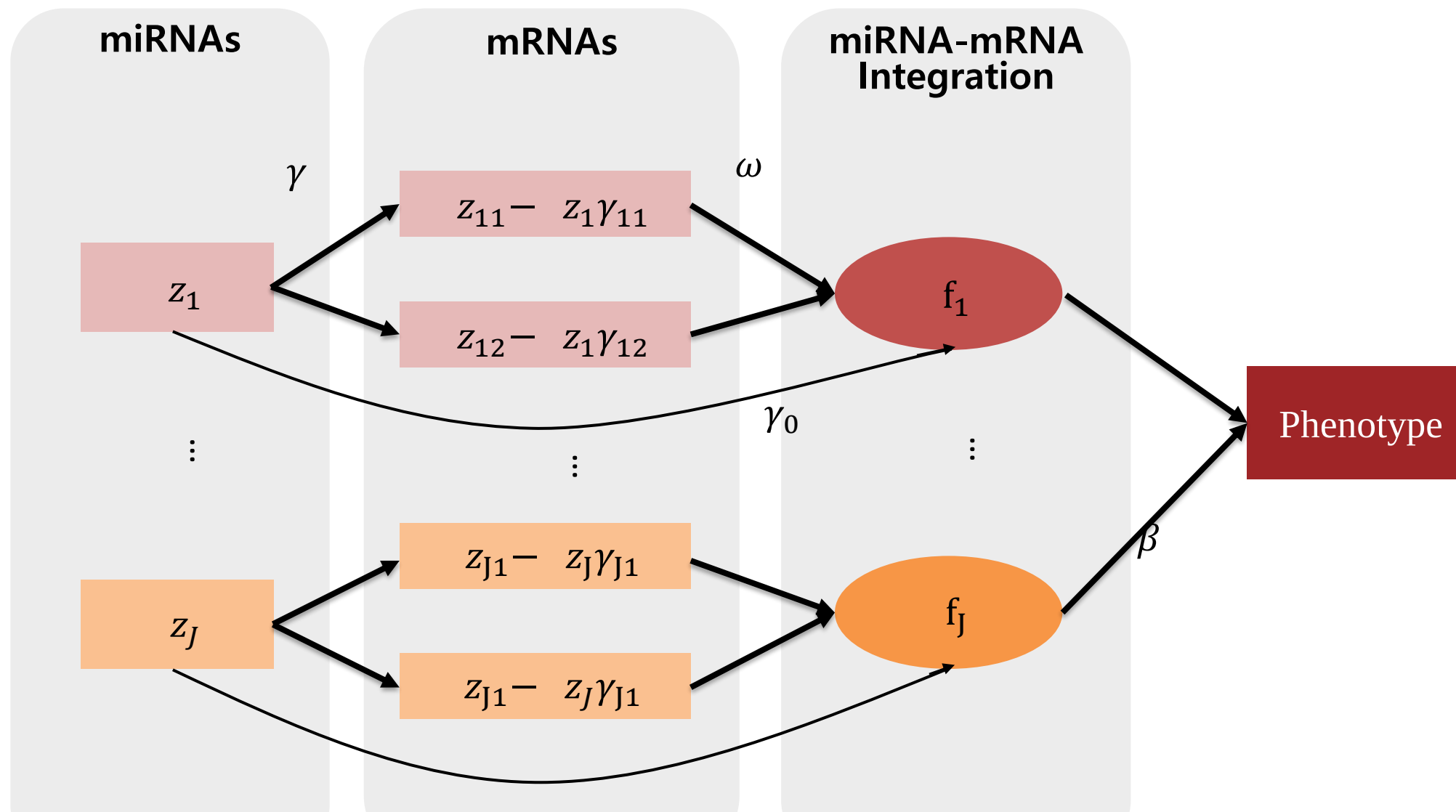


Transcriptomics



- Earlier works VIII

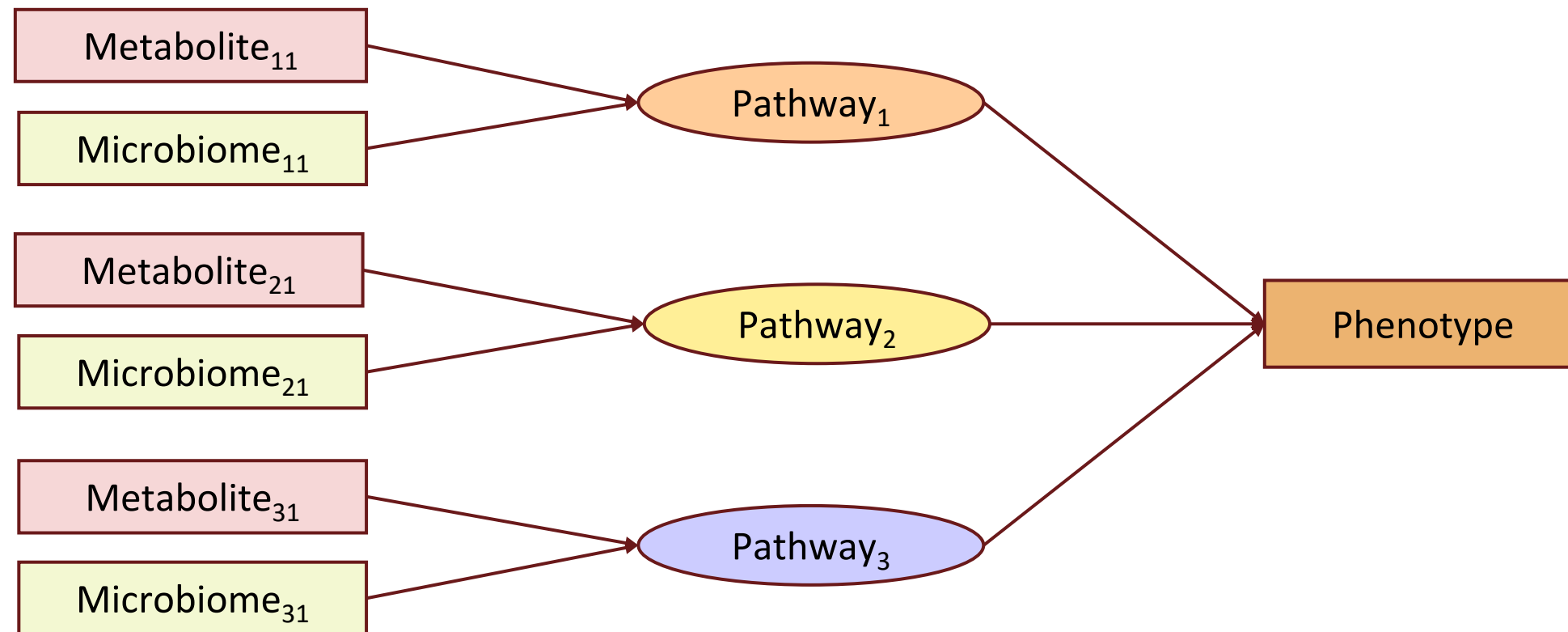
- HisCoM-mimi



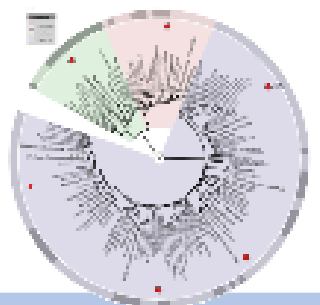


• Earlier works IX

- **HisCoM-MnM = metabolite+microbiome**
- Integration analysis of metabolite and microbiome data
- Pathway analysis for gene and microbiome data



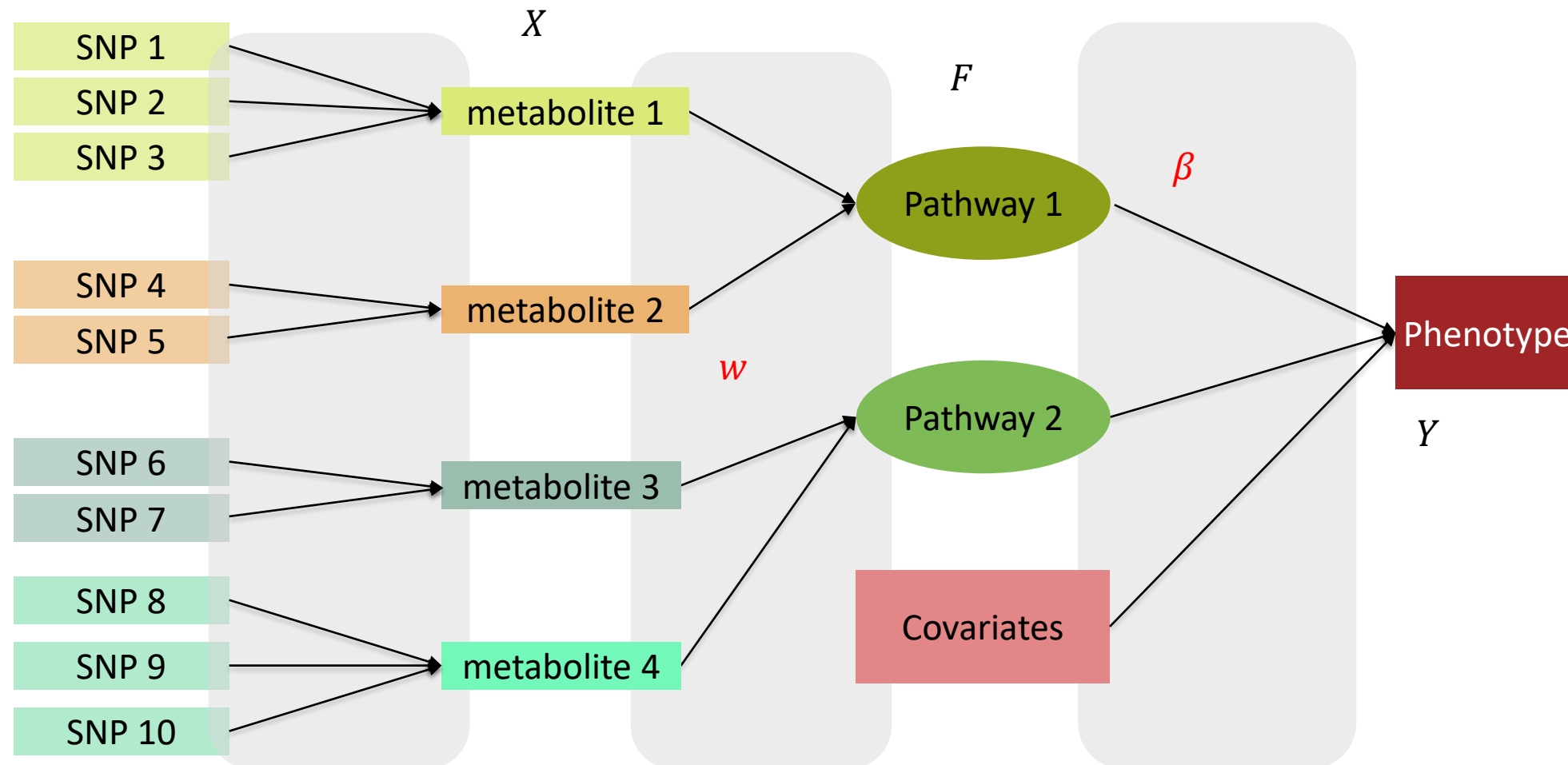
Metagenomics



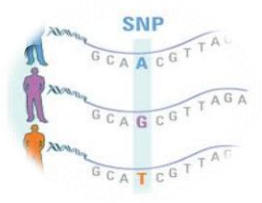


- **Earlier works X**

- **HisCoM-S&M = SNP+metabolite**
- Integration analysis of SNP and metabolite and microbiome
- Pathway analysis for SNP and metabolite data



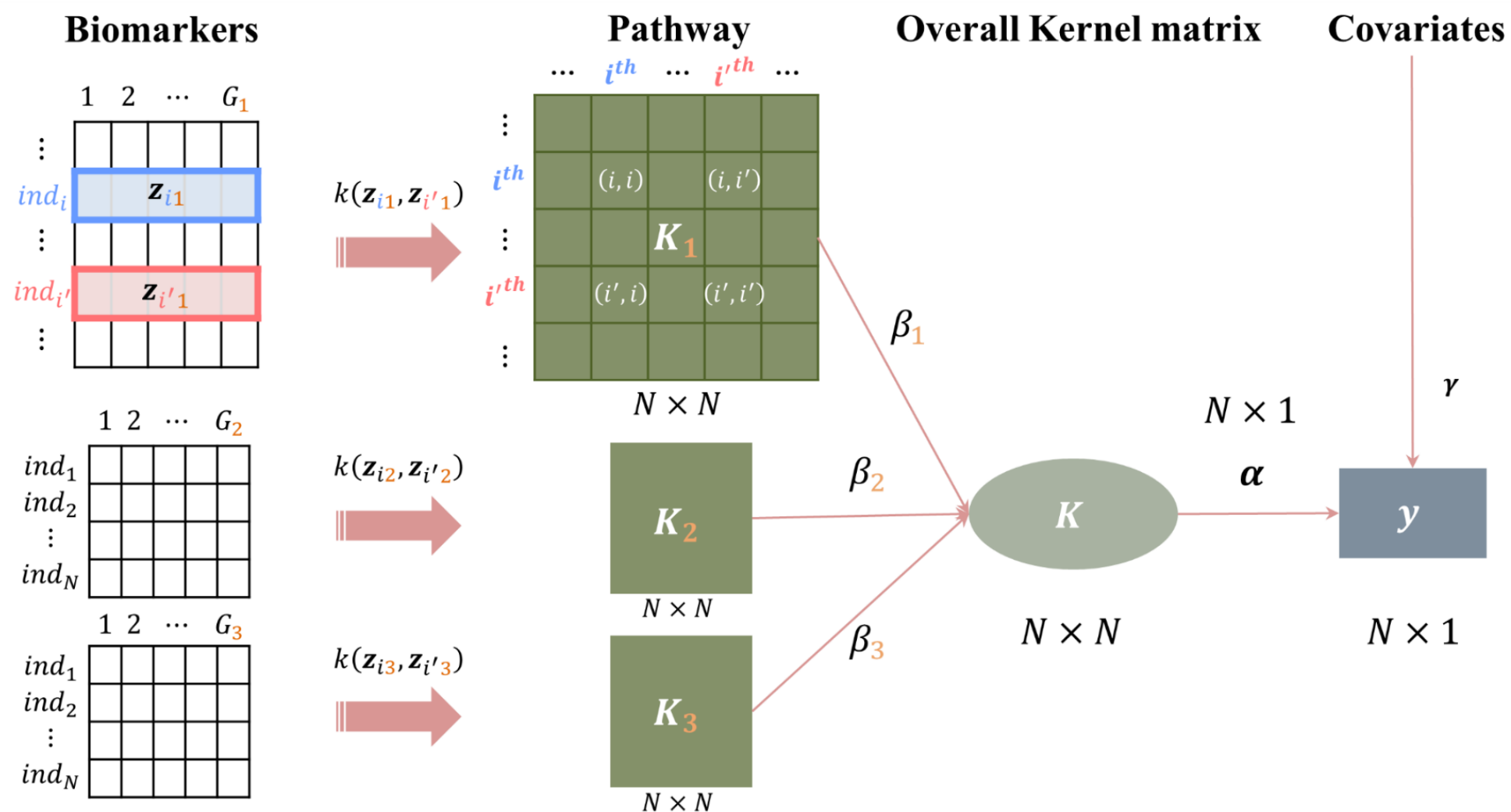
Genomic



- Earlier works XI

- HisCoM-Kernel

- Linear model => Non-linear model



- **Earlier works X**

- HisCoM-Kernel

- Linear model => Non-linear model

- Various kernel functions $k(\mathbf{z}_{ij}, \mathbf{z}_{i'j})$ are available for HisCoM-Kernel.
- Widely used kernel functions
 - Gaussian kernel: $k(\mathbf{z}_{ij}, \mathbf{z}_{i'j}; \rho) = \exp \left[-\sum_{g=1}^{G_j} (z_{ijg} - z_{i'jg})^2 / (G_j \cdot \rho) \right]$
 - ρ is Kernel parameter. It is also tuning parameter using cross-validation.
 - Second degree polynomial kernel: $k(\mathbf{z}_{ij}, \mathbf{z}_{i'j}) = \left(\mathbf{z}_{ij} \mathbf{z}_{i'j}^T / G_j + 1 \right)^2$
 - Linear kernel: $k(\mathbf{z}_{ij}, \mathbf{z}_{i'j}) = \mathbf{z}_{ij} \mathbf{z}_{i'j}^T$



Previous works on HisCoM

Omics data	Model
Rare variant data	• PHARAOH • PHARAOH-GEE • PHARAOH-Multi
Common genetic variant data	• HisCoM-GGI • HisCoM-GxE • HisCoM-PCA
Gene expression data Metabolomics data	• HisCoM-PAGE • HisCoM-Kernel • DeepHisCoM
Protein data	• HisCoM-PPI
Metagenome data	• HisCoM-microb
Multi-omics integration	• HisCoM-mimi, mimiSurv (miRNA & mRNA) • HisCoM-MnM (Metabolite & Microbiome) • HisCoM-S&M (SNP & Metabolite)

Previous works on HisCoM

Pathway-based approach using hierarchical components of collapsed rare variants

Sungyoung Lee^{1,†}, Sungkyoung Choi^{1,†}, Young Jin Kim², Bong-Jo Kim², T2d-Genes Consortium, Heungsun Hwang^{3,*} and Taesung Park^{1,4,*}

Bioinformatics

Pathway analysis of rare variants for the clustered phenotypes by using hierarchical structured components analysis

Sungyoung Lee^{1†}, Sunmee Kim^{2†}, Yongkang Kim³, Bermseok Oh⁴, Heungsun Hwang² and Taesung Park^{3,5*}

BMC Med Genomics

Pathway-based approach using hierarchical components of rare variants to analyze multiple phenotypes

Sungyoung Lee¹, Yongkang Kim², Sungkyoung Choi¹, Heungsun Hwang³ and Taesung Park^{1,2*}

BMC Bioinformatics

HisCoM-GGI: Hierarchical structural component analysis of gene–gene interactions

Sungkyoung Choi^{*,||}, Sungyoung Lee^{†,**}, Yongkang Kim[‡], Heungsun Hwang^{§,††} and Taesung Park^{‡,||,§§}

J Bioinform Comput Biol.

Hierarchical structural component modeling of microRNA-mRNA integration analysis

Yongkang Kim¹, Sungyoung Lee², Sungkyoung Choi², Jin-Young Jang³ and Taesung Park^{1,2*}

BMC Bioinformatics

Hierarchical structural component model for pathway analysis of common variants

Nan Jiang¹, Sungyoung Lee² and Taesung Park^{1,3*}

BMC Med Genomics

Drug response prediction model using a hierarchical structural component modeling method

Sungtae Kim¹, Sungkyoung Choi¹, Jung-Hwan Yoon², Youngsoo Kim³, Seungyeoun Lee⁴, Taesung Park^{5,6}

BMC Bioinformatics

Hierarchical Structured Component Analysis for Microbiome Data Using Taxonomy Assignments

Sun Ah Kim, Naveon Kang, Taesung Park

IEEE/ACM TCBB

HisCoM-PAGE: Hierarchical Structural Component Models for Pathway Analysis of Gene Expression Data

Lydia Mok¹, Yongkang Kim², Sungyoung Lee³, Sungkyoung Choi⁴, Seungyeoun Lee⁵, Jin-Young Jang⁶ and Taesung Park^{1,2,*}

Genes

HisCoM-G×E: Hierarchical Structural Component Analysis of Gene-Based Gene–Environment Interactions

Sungkyoung Choi¹, Sungyoung Lee², Iksoo Huh³, Heungsun Hwang⁴ and Taesung Park^{5,6,*}

Int J Mol Sci.

Pathway-Based Integrative Analysis of Metabolome and Microbiome Data from Hepatocellular Carcinoma and Liver Cirrhosis Patients

Boram Kim¹, Eun Ju Cho², Jung-Hwan Yoon², Soon Sun Kim³, Jae Youn Cheon⁴, Sung Won Cho³ and Taesung Park^{1,4,*}

Cancers

Identifying miRNA-mRNA Integration Set Associated With Survival Time

Yongkang Kim^{1†}, Sungyoung Lee^{2,3†}, Jin-Young Jang⁴, Seungyeoun Lee⁵ and Taesung Park^{1,6*}

Front Genet.

Integrative Pathway Analysis of SNP and Metabolite Data Using a Hierarchical Structural Component Model

Taeyeong Jung¹, Youngae Jung², Min Kyong Moon³, Oran Kwon⁴, Geum-Sook Hwang², Taesung Park^{1,5}

Front Genet.

Kernel-based hierarchical structural component models for pathway analysis

Suhyun Hwangbo^{1,2}, Sungyoung Lee², Seungyeoun Lee³, Heungsun Hwang⁴, Inyoung Kim⁵

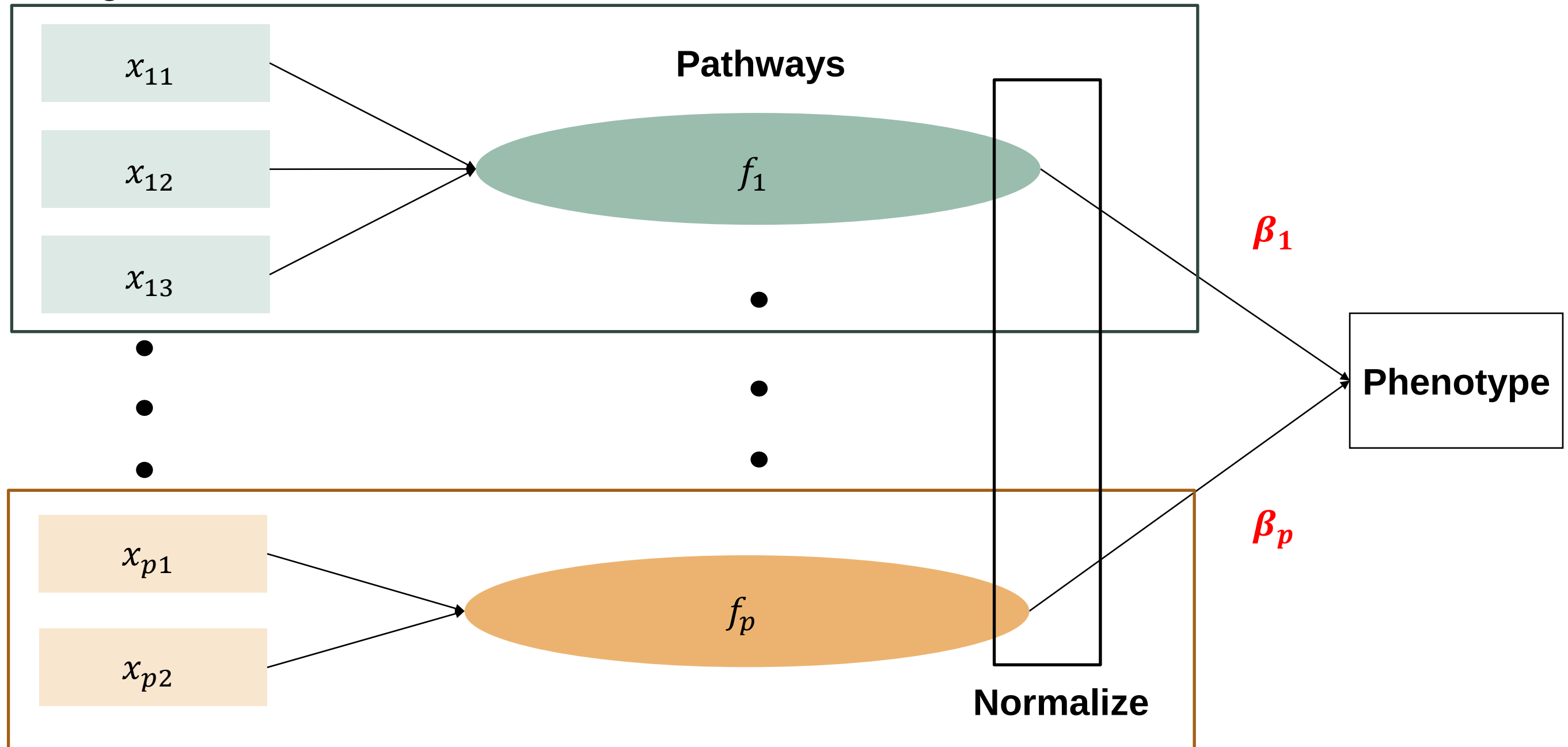
Bioinformatics

DeepHisCoM: deep learning pathway analysis using hierarchical structural component models

Chanwoo Park¹, Boram Kim², Taesung Park^{1,2}

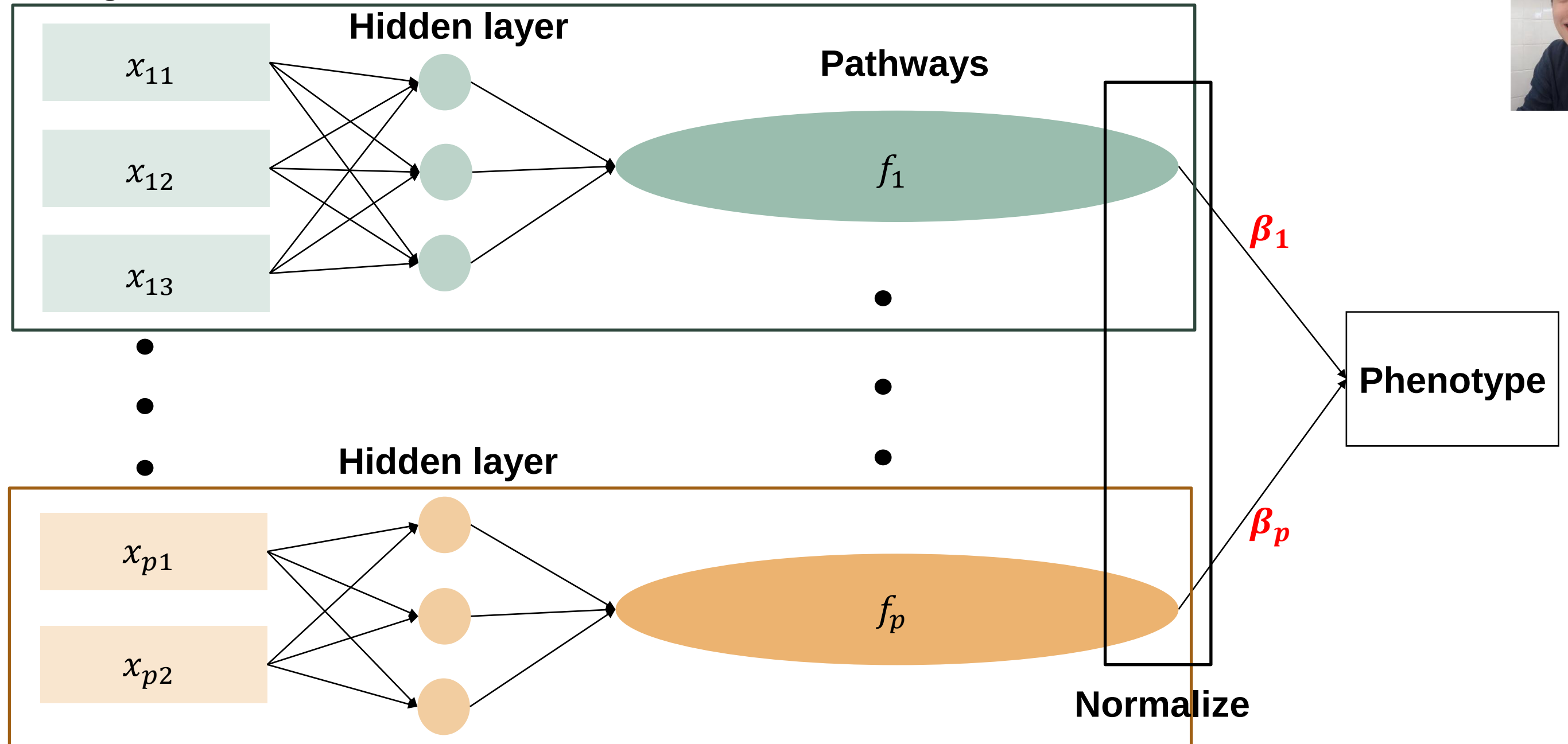
Brief Bioinformatics

Biological factors

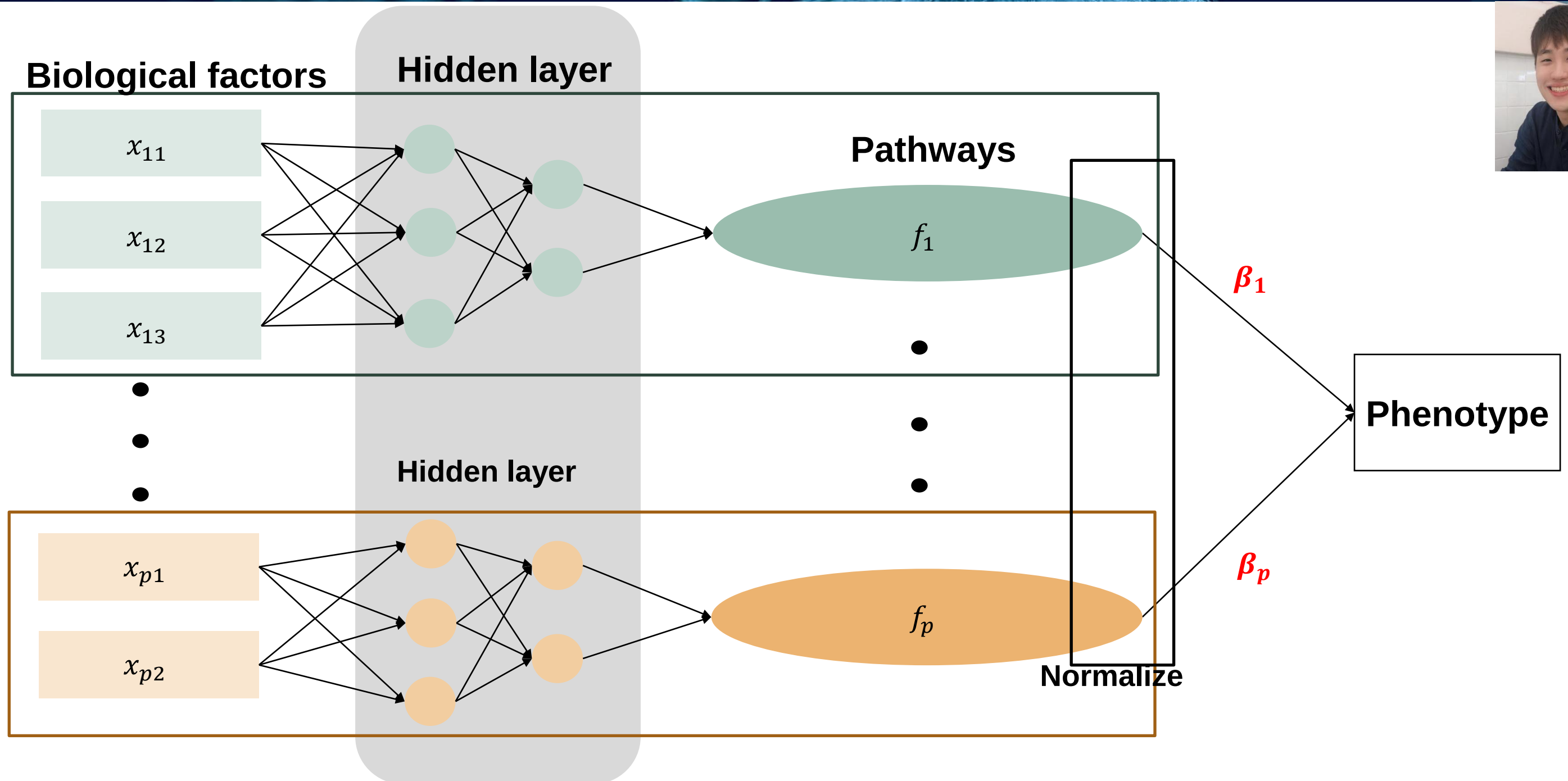


Deep-HisCoM = Deep + HisCoM

Biological factors



Deep2-HisCoM with 2-hidden-layer



- **Deep - Hierarchical structured Component Model (Deep-HisCoM)**

- $x_{ijk} : i^{th}$ pathway, j^{th} biological factor in k^{th} subject
 - $(i = 1, \dots, p(\text{pathway \#}); j = 1, \dots, T_i(\text{biological factor \#}); k = 1, \dots, n(\text{sample \#}))$
- $f_{ik} : \text{latent variable representing the effect of } i^{th} \text{ pathway}$
- $f_{ik} = g_i(x_{i1k}, x_{i2k}, \dots, x_{iT_ik})$ (g_i represents pathway)
- $\text{logit}(\pi_k) = \beta_0 + \sum_{i=1}^p \beta_i f_{ik}$
- $\phi = \text{Loss}(y_1, y_2, \dots, y_n) - \lambda_m \sum_{i,h} w_{ih}^2 - \lambda_p \sum_{i=1}^p \beta_i^2$



- **Deep - Hierarchical structural Component Model (Deep-HisCoM)**
 - Using Hidden layer, we can **approximate continuous functions** between biological factors and pathways (Universal Approximation theorem, Pinkus *et. al*, Acta Numerica, 1999)
 - Algorithm and Inference
 - To optimize ϕ , we use Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (L-BFGS) (Liu, *et. al*, 1989), and repeat iterations until convergence
 - Based on Generalized Structuralized Component Analysis (GSCA) and Deep Neural Network (Schmidhuber, Jürgen *et. al*, 2015, Fukushima *et. al*, 1980)
 - Significant test: permutation test for the coefficients of pathways



Similarity and Difference between HisCoM and Deep HisCoM

	HisCoM	Deep-HisCoM
Local information	Local information (pathway-metabolites)	
Significant pathway detection	p-value via permutation test	
Relation between pathway and metabolites	Linear function	Any continuous function
Optimizer	IWLS	L-BFGS
Loss Function	Loglikelihood = Binary cross entropy	
Hyper-parameter	λ_m, λ_p (Ridge parameters)	Type of Activation Function Hidden layer number L-BFGS learning rate λ_m, λ_p (Ridge parameters)

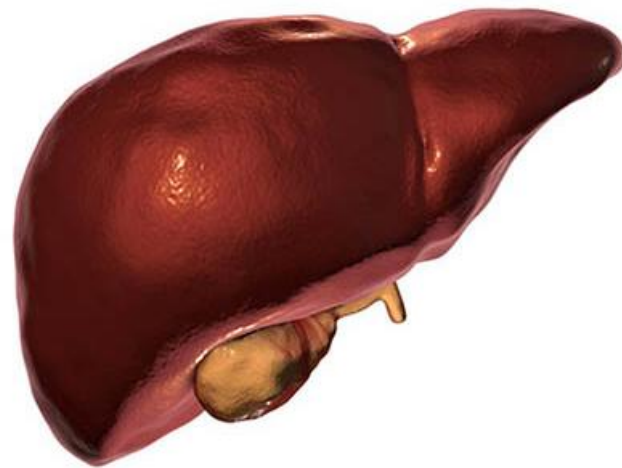


Real Data Analysis

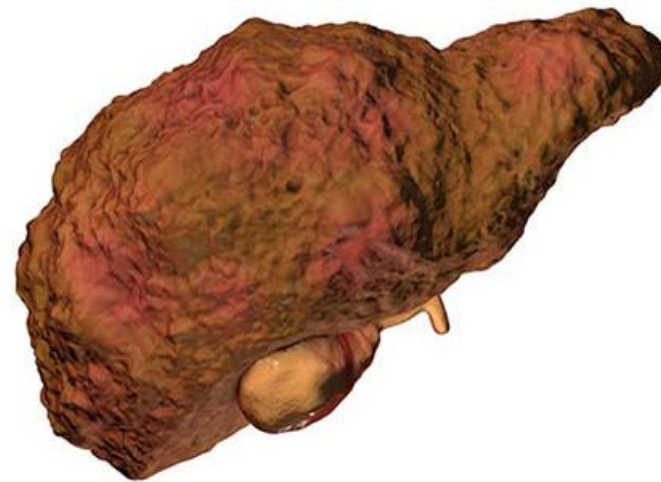
- **HCC Data**
 - Metabolome data
 - Microbiome data
 - RNA-seq data

Hepatocellular Carcinoma (HCC)

- Most common cause of death in a patient with cirrhosis due to hepatitis B, hepatitis C, or alcohol
- Even when diagnosed at the earlier stage, most patients with HCC eventually progresses to the advanced stage



Healthy liver



Cirrhotic liver

HCC Data Description

	Metabolomics
case	103
control	97
#pathway	23
#biomarkers	32

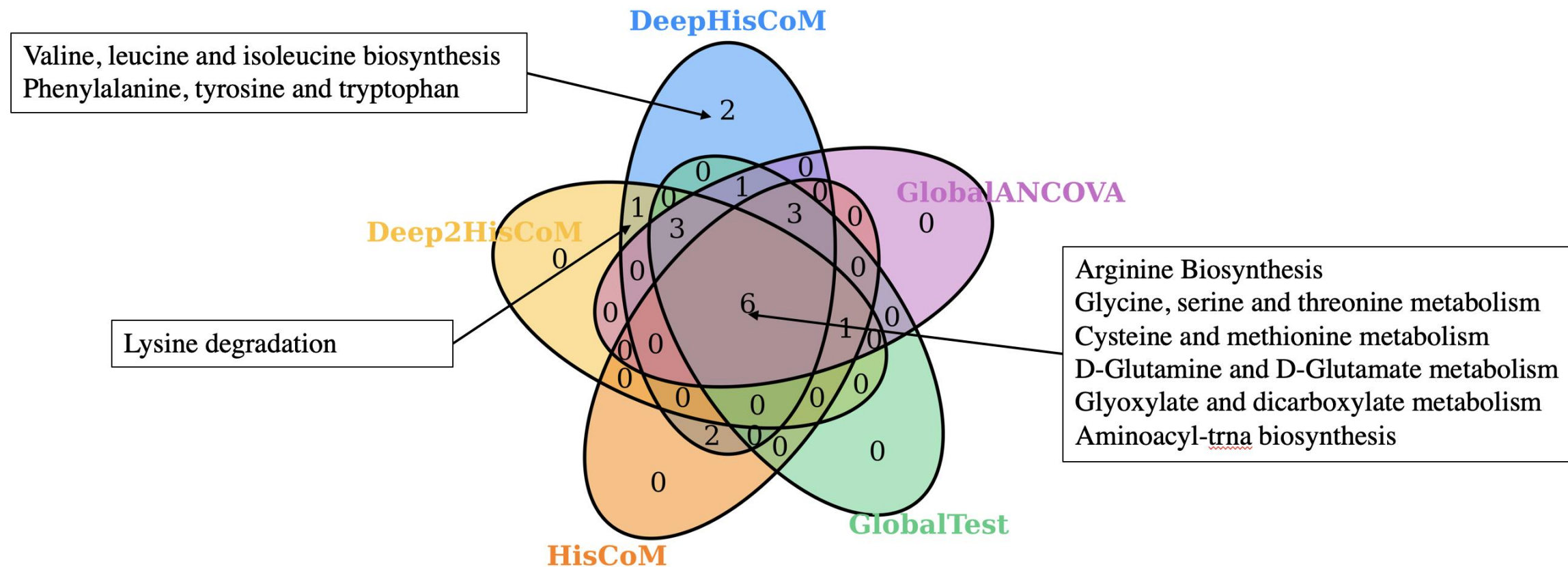
- **HCC metabolome data**
 - 103 HCC, 97 LC
 - Generated from Seoul National University Hospital (Seoul, Korea), and from Ajou University Hospital (Suwon, Korea)
 - Targeted metabolomics analysis using AbsoluteIDQ® p180 kit (BIOCRATES Life Science AG, Innsbruck, Austria), with liquid chromatography and tandem mass spectrometry (LC-MS/MS), was used to quantify metabolites (Kim *et al.*, 2019)
 - **Pathway annotation**
 - Mapping each metabolite to KEGG compound ID using MetaboAnalyst and Human Metabolome Database (HMDB)
 - 32 unique metabolites were mapped to 23 metabolic pathways

- **HCC microbiome data**
 - 93 HCC, 94 LC
 - Generated from Seoul National University Hospital (Seoul, Korea), and from Ajou University Hospital (Suwon, Korea)
 - 16s rRNA based approach using circulating cell-free DNA that was extracted from serum extracellular vesicles (Cho *et al.*, 2019)
 - Closed-reference operational taxonomic unit (OTU)-picking pipeline and Greengenes reference database was used to generate OTU table using QIIME
 - **Pathway annotation**
 - Generating a functional table using PICRUST and KEGG pathway database to predict metagenomic contents, KEGG orthologs (KOs)
 - 752 KOs were mapped to 23 pathways

- **HCC TCGA RNA-seq data**
 - 371 HCC, 50 normal liver tissue
 - RNA-seq and corresponding clinical data for liver hepatocellular carcinoma were obtained from The Cancer Genome Atlas Liver Hepatocellular Carcinoma (TCGA) database
 - The ensemble gene ID was converted into a gene symbol and the missing values were removed
 - Normalization using DESeq2
 - **Pathway annotation**
 - 571 genes were mapped to 18 pathways

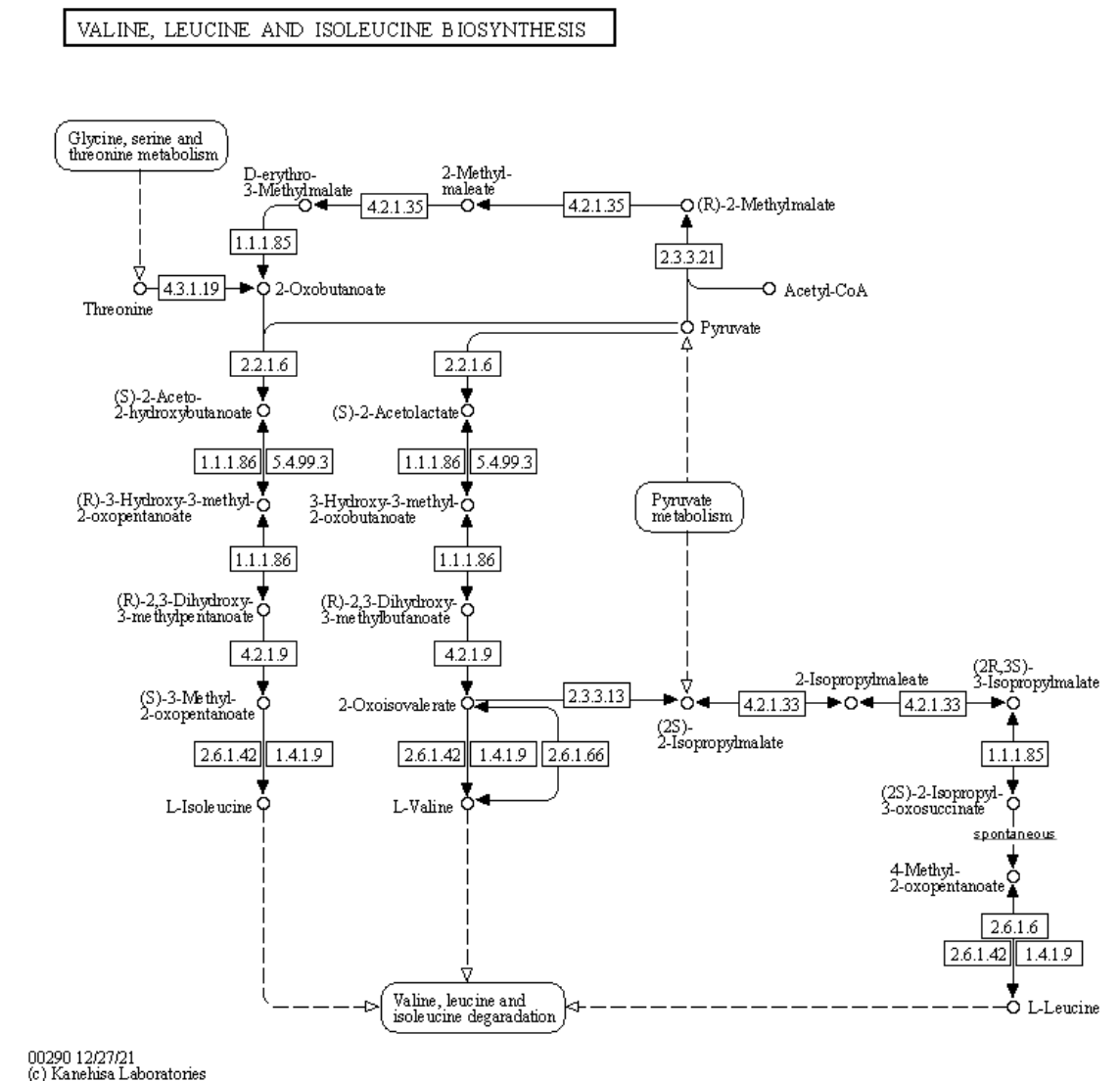
HCC Metabolite Data

- DeepHisCoM and Deep2HisCoM identified 3 significant pathways not found by the traditional pathway analysis (q-value cut = 0.01)



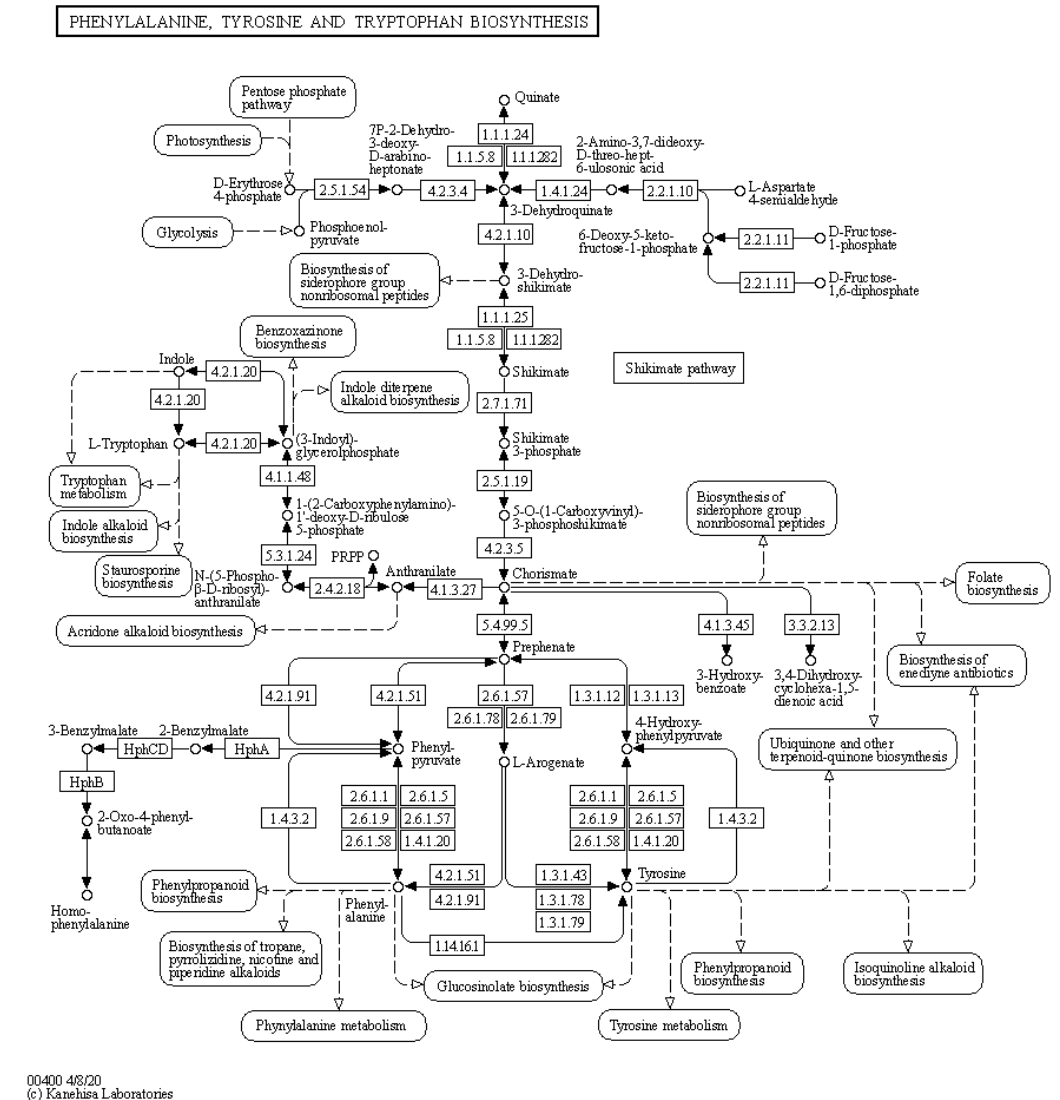
HCC Metabolite Data: Significant pathways

- **Valine, leucine, and isoleucine biosynthesis**
 - The **branched-chain amino acids (BCAAs)** constituting valine, leucine, and isoleucine were reported to show favorable effects on hepatocarcinogenesis
 - Many studies reported that BCAAs suppress angiogenesis, hepatic fibrosis, oxidative stress, and hepatitis C virus replication
 - Moreover, many researchers demonstrated that BCAAs prevented progressive hepatic failure and prolonged survival of cirrhotic patients
 - **Kawaguchi *et al.*, 2012, 2013, Yoshiji *et al.*, 2009, 2012, Ohno *et al.*, 2008, Marchesini *et al.*, 2003**



HCC Metabolite Data: Significant pathways

- **Phenylalanine, tyrosine and tryptophan biosynthesis**
 - Several HCC metabolomic studies have linked aberrant amino acid metabolism to cancer development such as significant upregulation of phenylalanine in HCC versus cirrhosis and upregulation of tyrosine, tryptophan, and phenylalanine in HCC versus healthy controls.
 - Upregulation of such amino acids in HCC may be explained by the reduced aromatic amino acids (phenylalanine, tyrosine, and tryptophan) breakdown in the diseased liver, which reflected the deterioration of liver metabolism and hepatic function.
 - **Fitian *et al.*, 2014, 2017, Chen *et al.*, 2013**

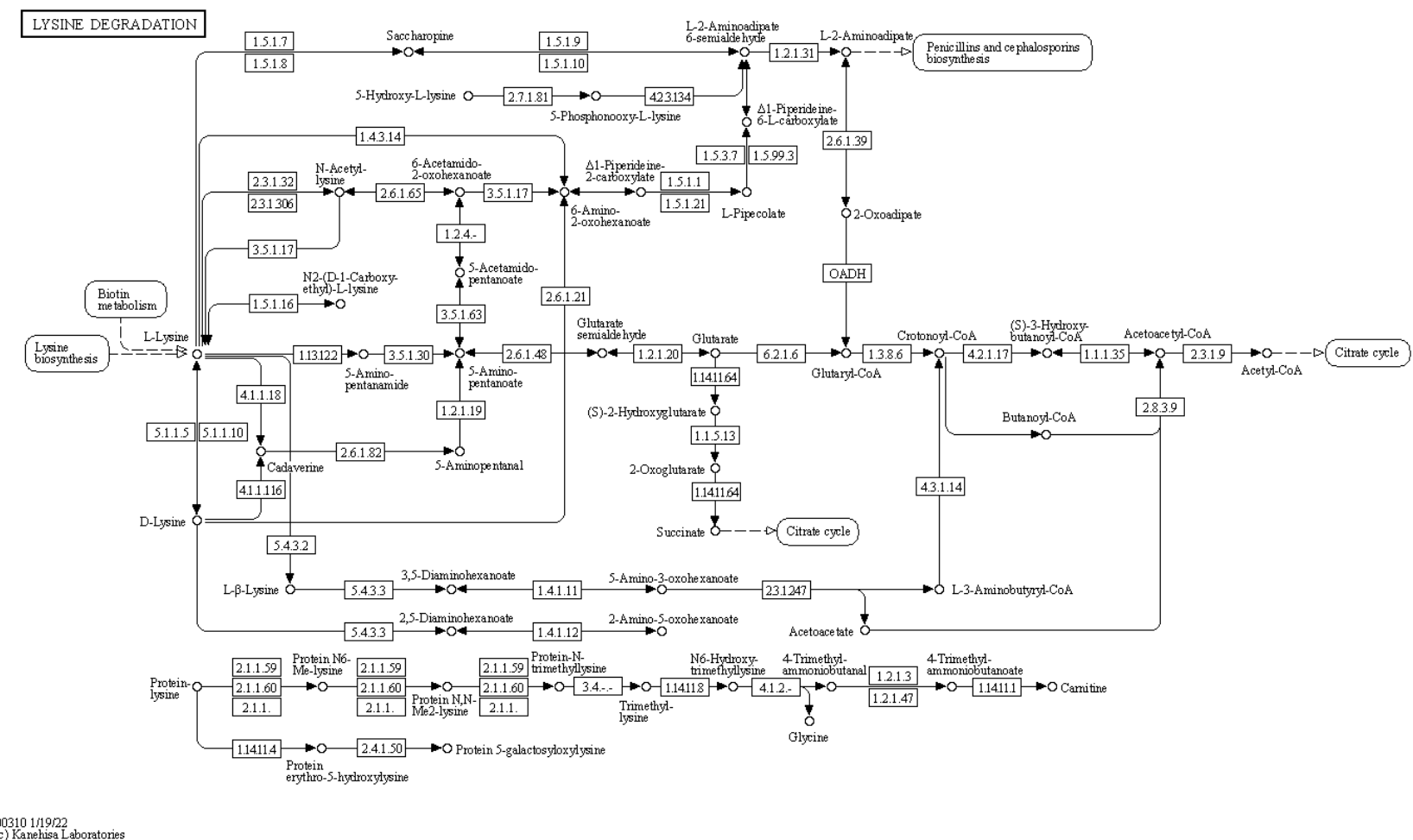


HCC Metabolite Data: Significant pathways

- **Lysine degradation**

- The decrease of lysine in plasma of HCC patients has been reported in previous studies
- These metabolite differences could represent perturbations in specific metabolic pathways that may contribute to the transition to HCC from liver disease

➤ **Chen *et al.*, 2011,**
Clarke *et al.*, 2014



HCC Replication Study: Can pathways be replicated from other datasets?

- In metabolomic dataset analysis, three pathways were detected only by DeepHisCoM but not by other methods at FDR adjusted q-value smaller than 0.01.
 - **lysine degradation**
 - **valine, leucine, and isoleucine biosynthesis**
 - **phenylalanine, tyrosine, and tryptophan**
- To investigate whether these discovered pathways are associated with HCC, we performed additional validation analyses using **metagenomic** and **RNA-seq datasets**.

- **Metagenomic dataset**
 - DeepHisCoM and Deep2HisCoM successfully identified two pathways (lysine degradation, and valine, leucine, and isoleucine biosynthesis) with q-values smaller than 0.01.
- **RNA-Seq dataset**
 - DeepHisCoM and Deep2HisCoM successfully identified two pathways (lysine degradation, and phenylamine, tyrosine and tryptophan) with q-values smaller than 0.01.
- **Summary**
 - Three pathways detected from metabolomic dataset by DeepHisCoM were successfully replicated from metagenomic and RNA-Seq datasets.
 - We conjecture that the metabolites in three pathways **have non-linear relationships** with HCC. Thus, these pathways were not detected by GlobalTest, GlobalANCOVA, and HisCoM in the metabolomic dataset.



Simulation Study

Simulation Description

- Want to compare the performance of methods
- Methods considered
 - HisCoM, DeepHisCoM, Deep2HisCoM, Globaltest, GlobalANOVA
- Measures of performance
 - Empirical type I error rates
 - Empirical power by calculating the proportion of the true causal pathway in the simulation $q < 0.05$ or $q < 0.01$ from 100 replication
- Use original HCC metabolites data
- We simulate model with two pathways (and model with five pathways)
- We used Tryptophan metabolism (map00380) and Primary bile acid biosynthesis (map00120) as true pathways
- HCC phenotype was generated under several pathway model from $\text{Binomial}(\pi_i, 1)$ for each sample, where π_i the success probability.

Simulation model with two pathways

- **Simulation model 1: linear relationship** between metabolites and pathway.

$$\log \frac{\pi_i}{1 - \pi_i} = \beta_{map00400} \left(\sum_i w_i x_i \right) + \beta_{map00860} \left(\sum_i w_i x_i \right)$$

fix $w_i = 1.5$ for all metabolites, and change $\beta = \beta_{map00400} = \beta_{map00860} = 0.1, 0.2, 0.3, 0.4$.

- **Simulation model 2: interaction effects** between metabolites and pathways

$$\log \frac{\pi_i}{1 - \pi_i} = \beta_{map00400} (\sum_{1 \leq i \neq j \leq 3} w_{ij1} x_{i1} x_{j1} + \sum_i w_{i1} x_{i1}) + \beta_{map00860} (\sum_{1 \leq i \neq j \leq 3} w_{ij2} x_{i2} x_{j2} + \sum_i w_{i2} x_{i2})$$

fix $w_{ij1} = w_{ij2} = 1$ for all metabolites and $\beta = \beta_{map00400} = \beta_{map00860} = 0.4$, and change $w_{ij1} = w_{ij2} = 0.1, 0.2, 0.3, 0.4$.

- **Simulation model 3: quadratic relationship** between metabolites and pathway.

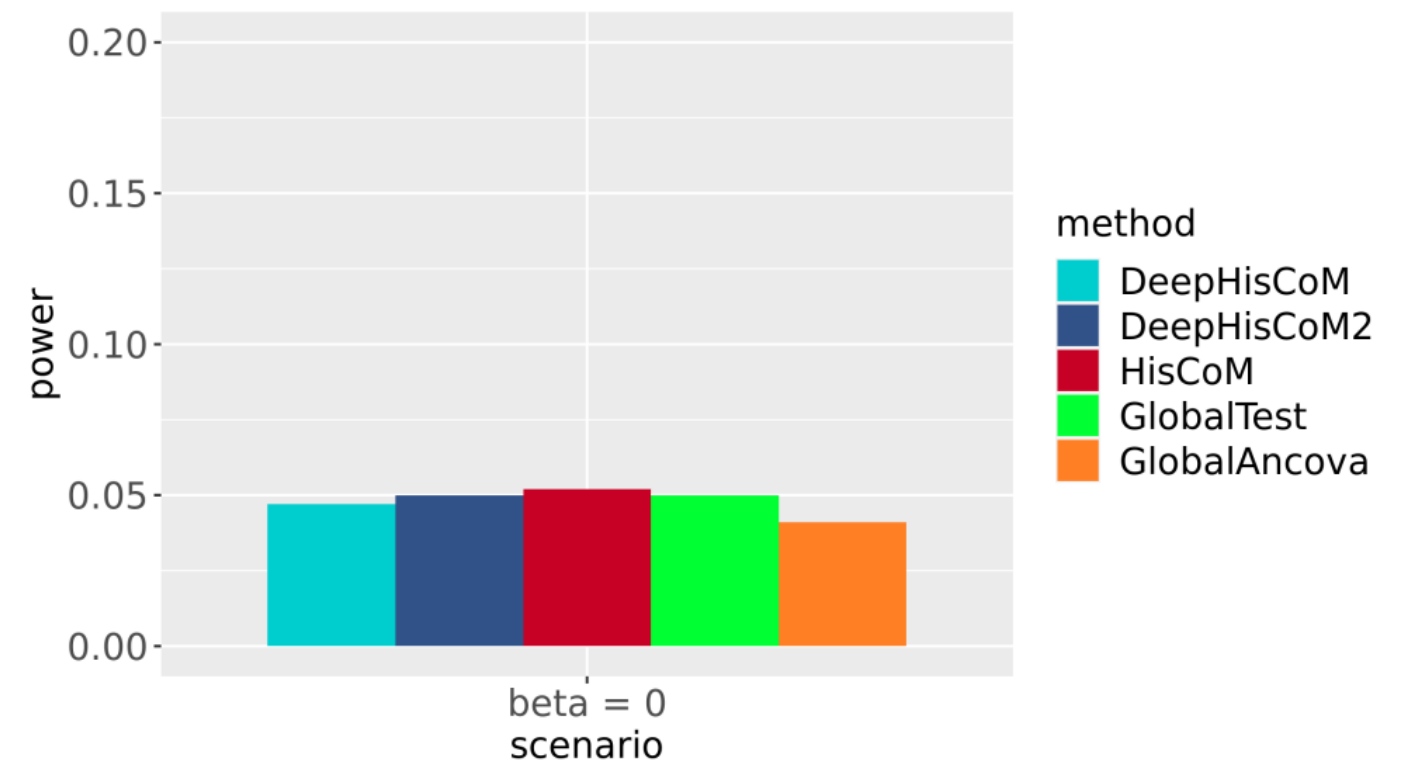
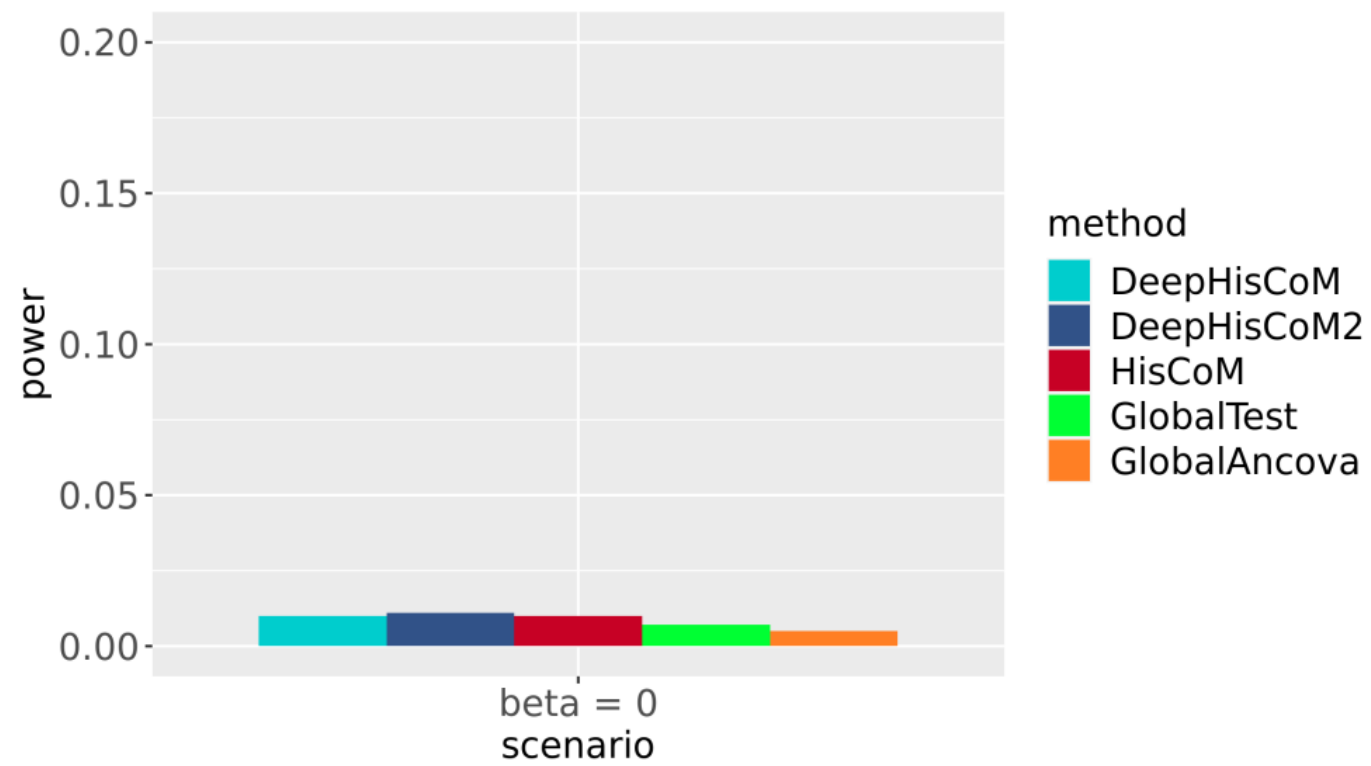
$$\log \frac{\pi_i}{1 - \pi_i} = \beta_{map00400} \left(\sum_i w_i x_i^2 \right) + \beta_{map00860} \left(\sum_i w_i x_i^2 \right)$$

fix $w_i = (-1)^i$ for all metabolites, and change $\beta = \beta_{map00400} = \beta_{map00860} = 0.6, 0.7, 0.8, 0.9$

Simulation model with five pathways

- We randomly selected five pathways that do not overlap and have a similar number of metabolites as much as possible.
- The selected five pathways are map00120, map00220, map00280, map00360, map00380, containing 2, 7, 3, 2, 3 metabolites, respectively.
- Simulation settings are quite similar to those of two pathways.
- For Simulation model 1, we fix $w=0.8$ for all metabolites, and changed $\beta=0.1, 0.2, 0.3, 0.4$.
- The setting of beta and w are the same as the two pathways cases of simulation models 2 and 3.

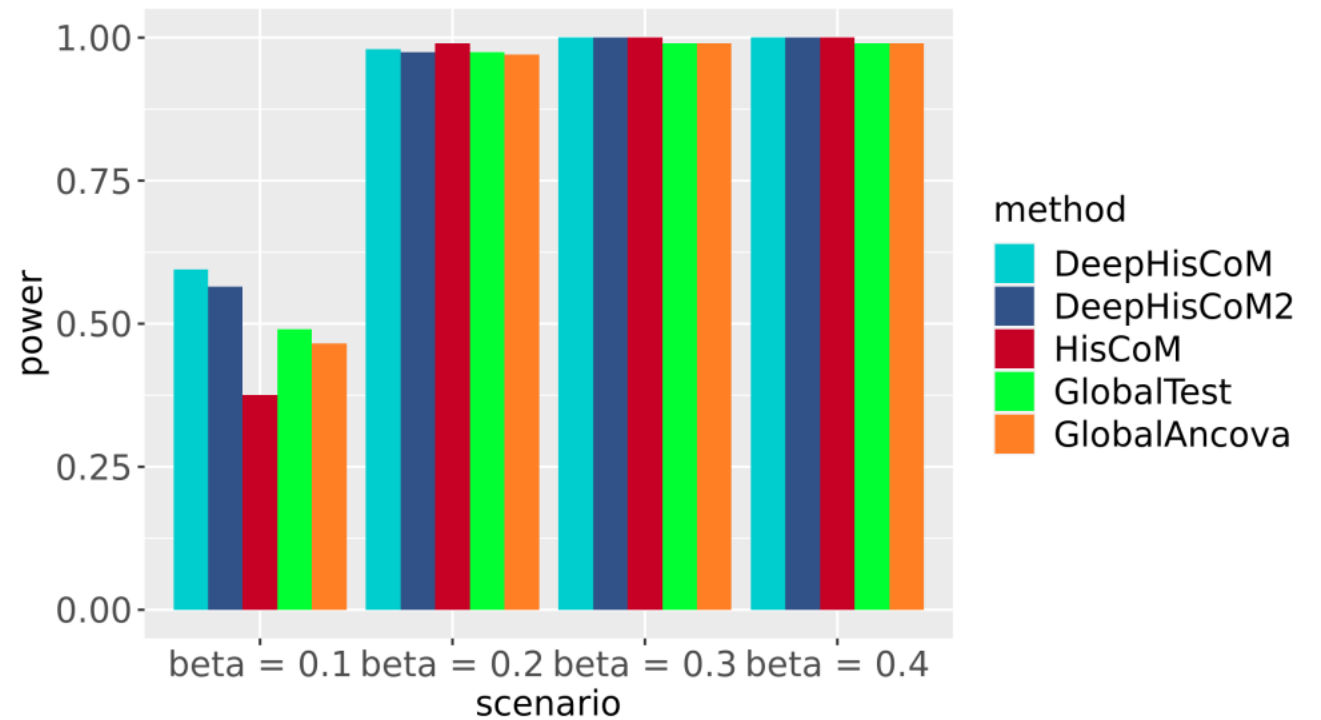
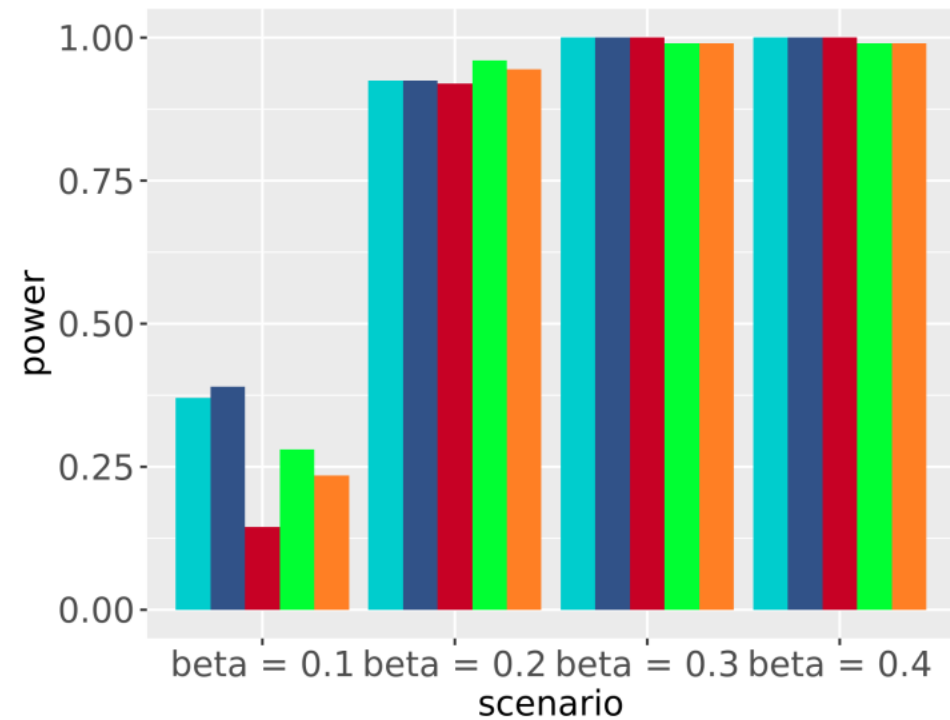
Type 1 error: Two pathways



Type 1 errors are well controlled

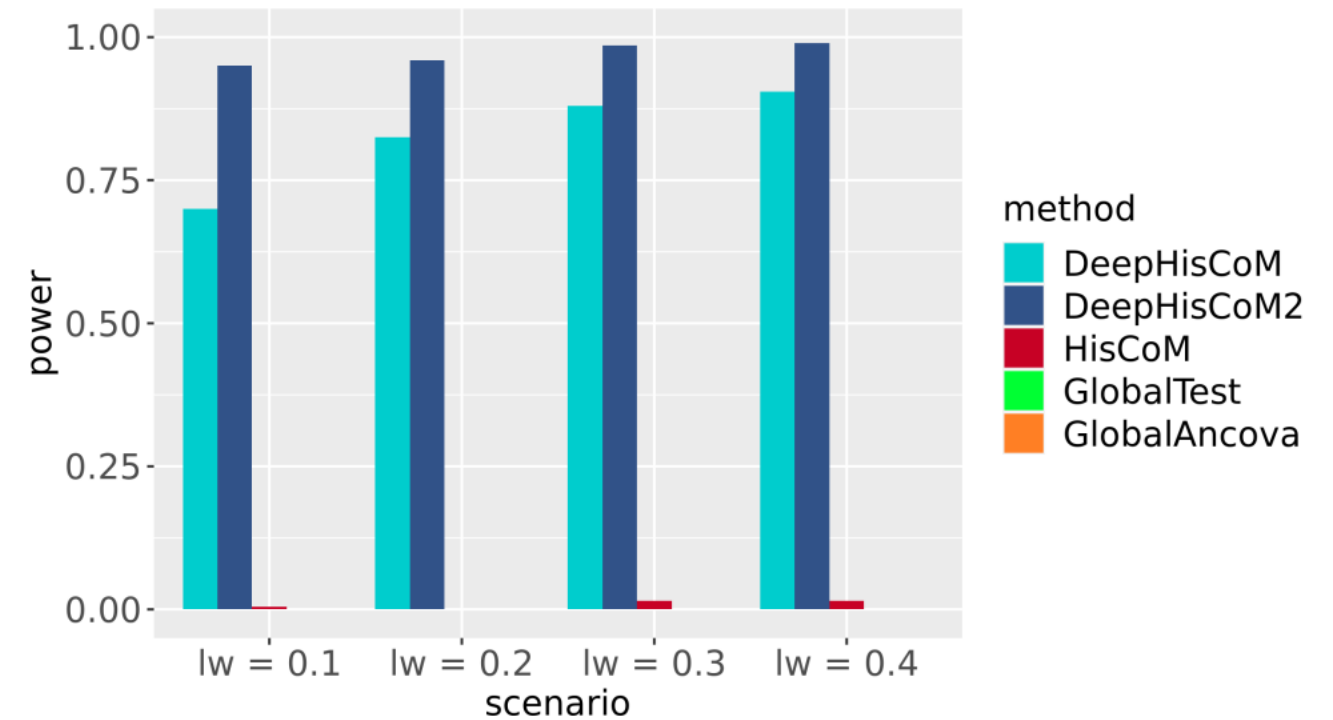
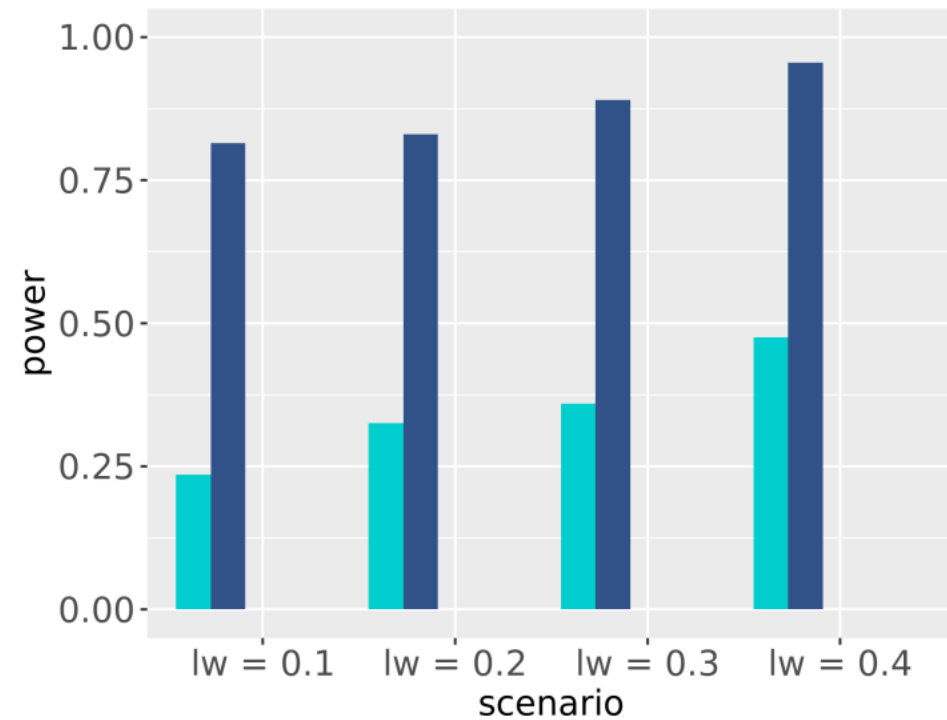
Power for Simulation Model 1: Two pathways

Linear Assumption



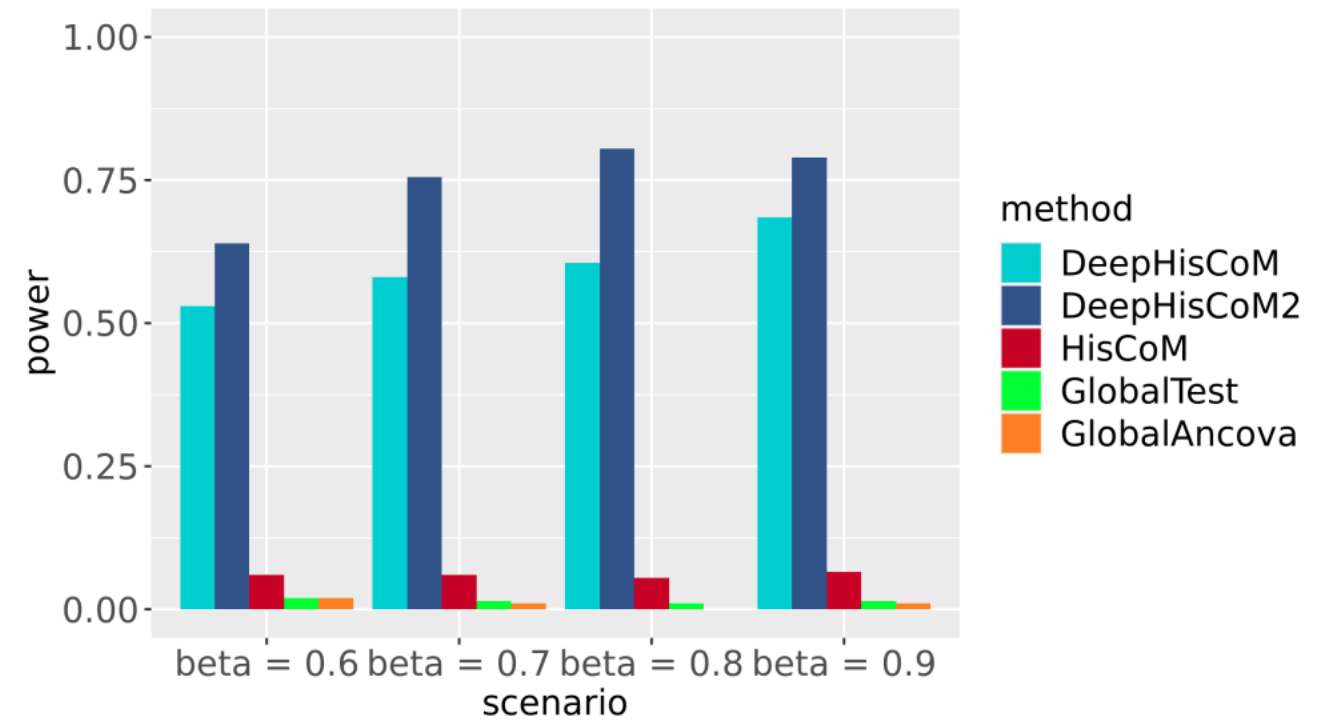
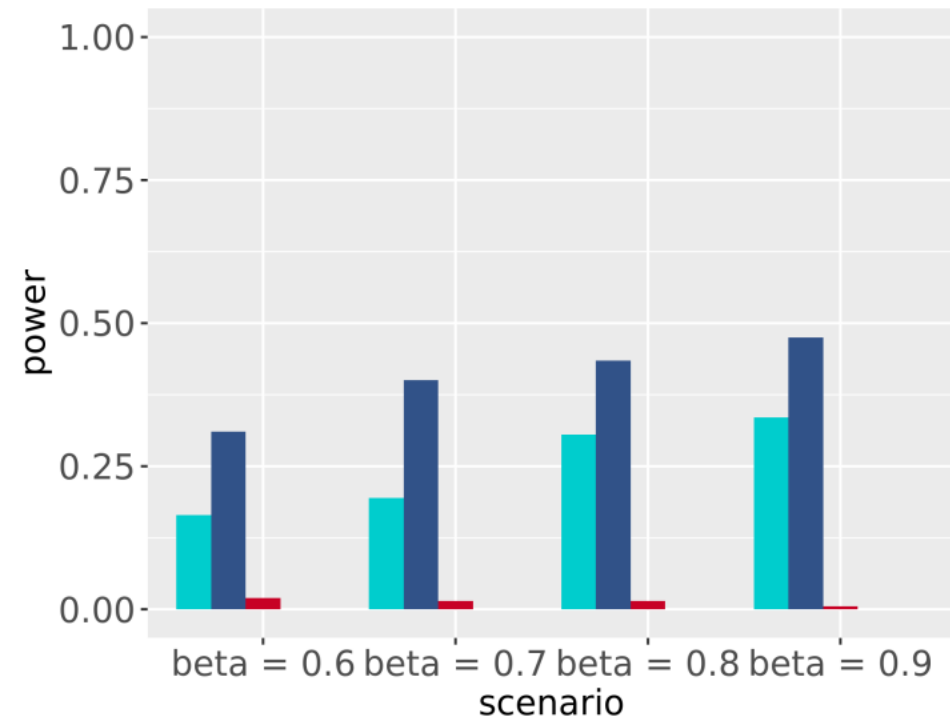
Power for Simulation Model 2: Two pathways

Interaction Assumption

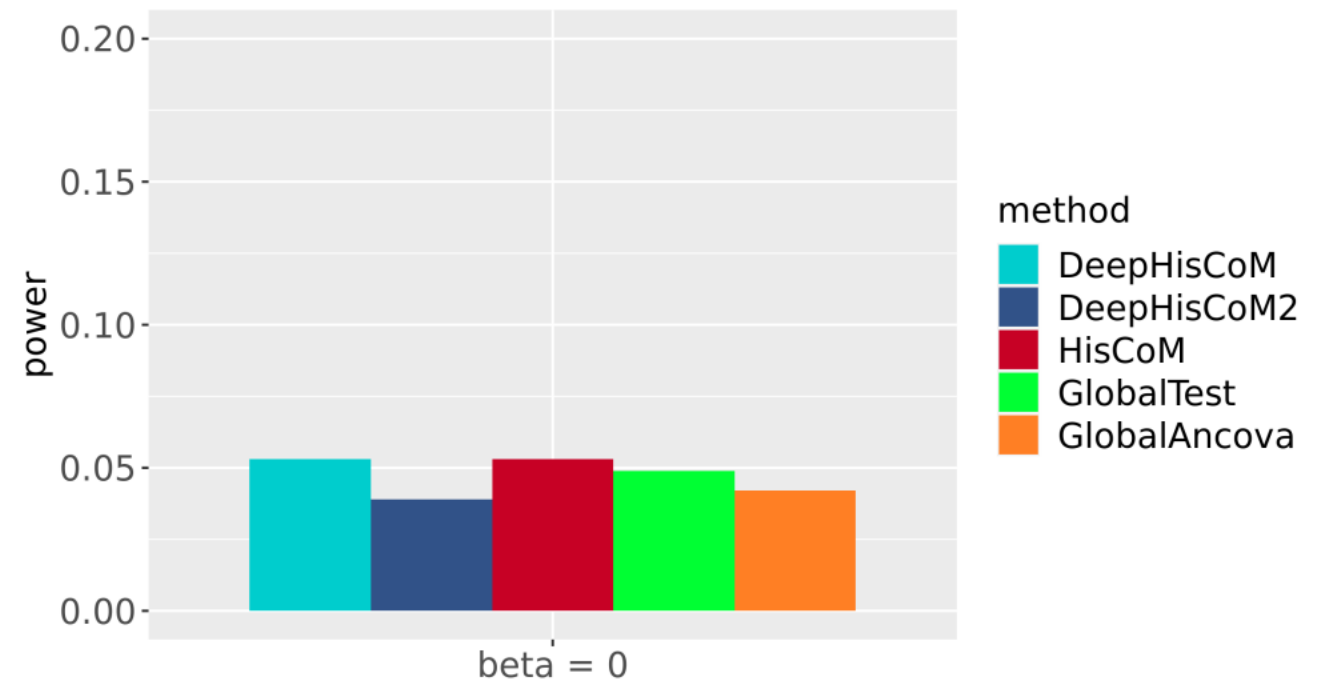
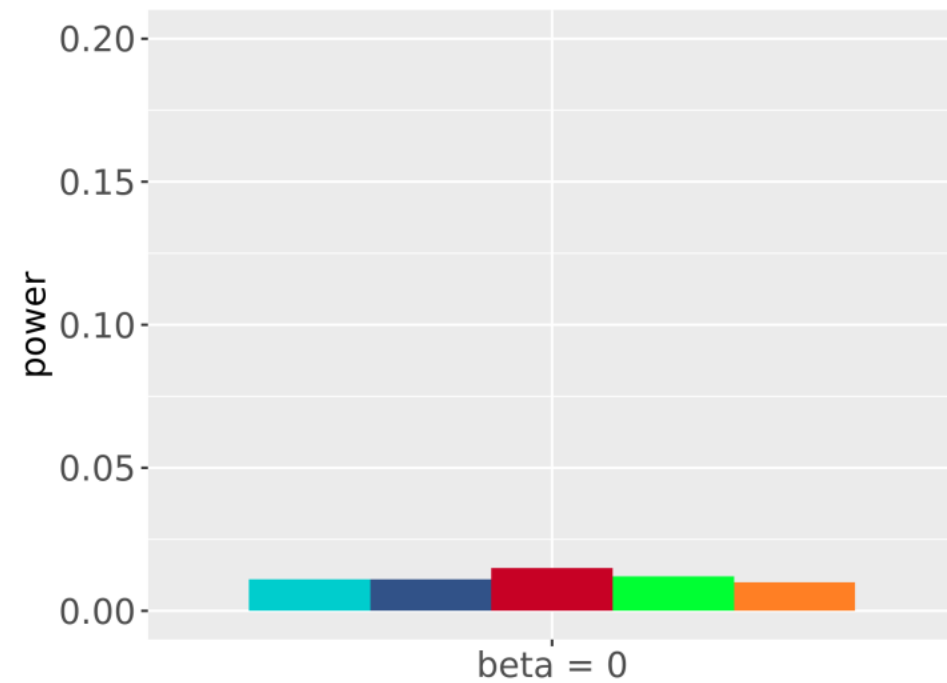


Power for Simulation Model 3: Two pathways

Quadratic Assumption



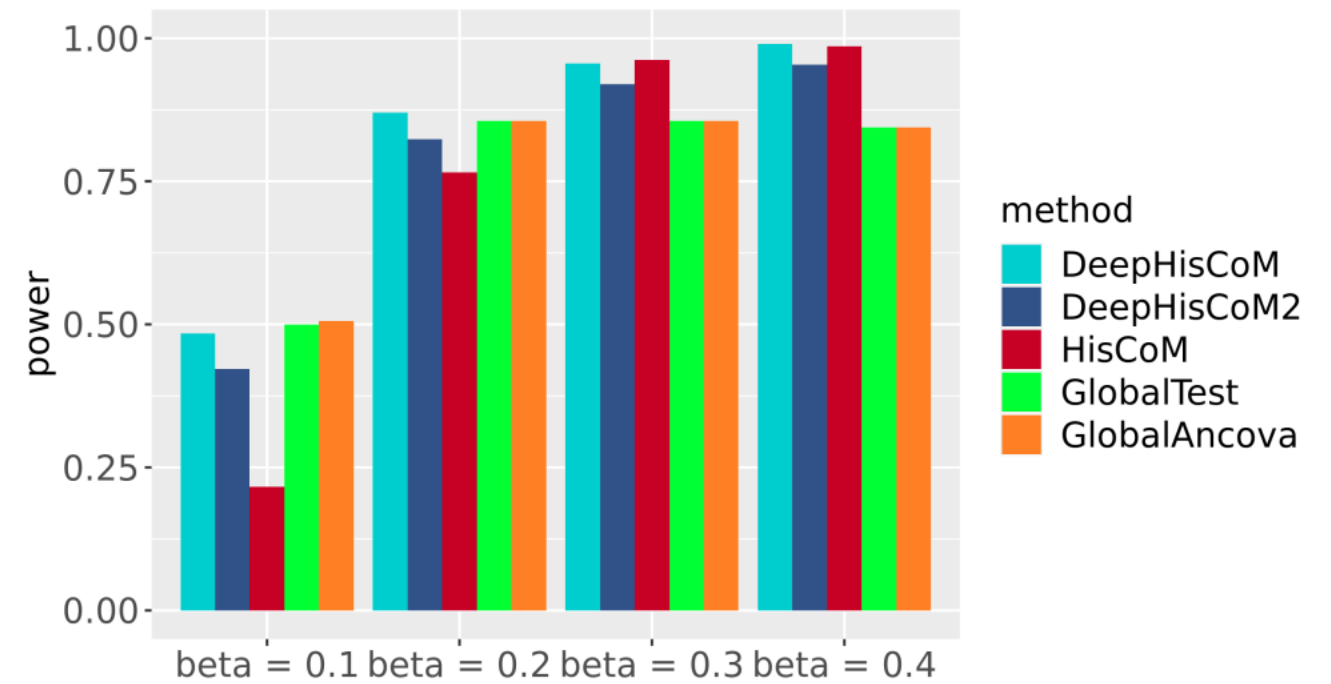
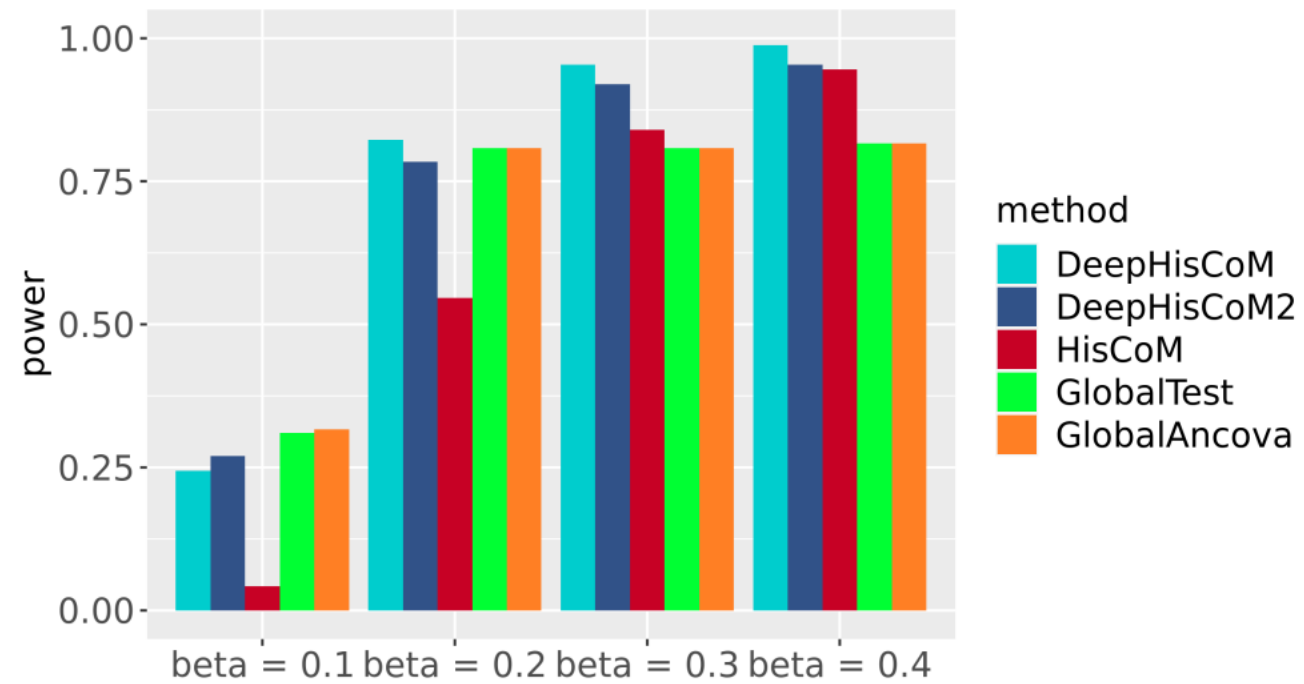
Type 1 error for five pathways



Type 1 errors are well controlled

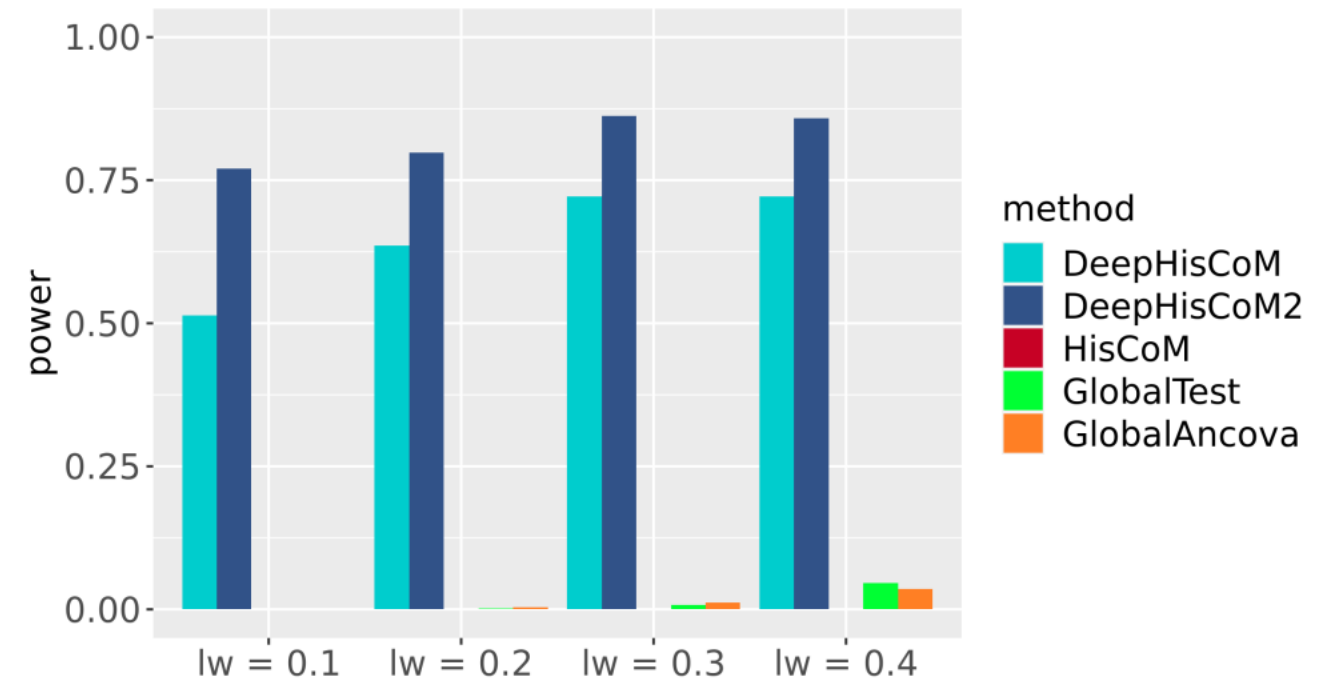
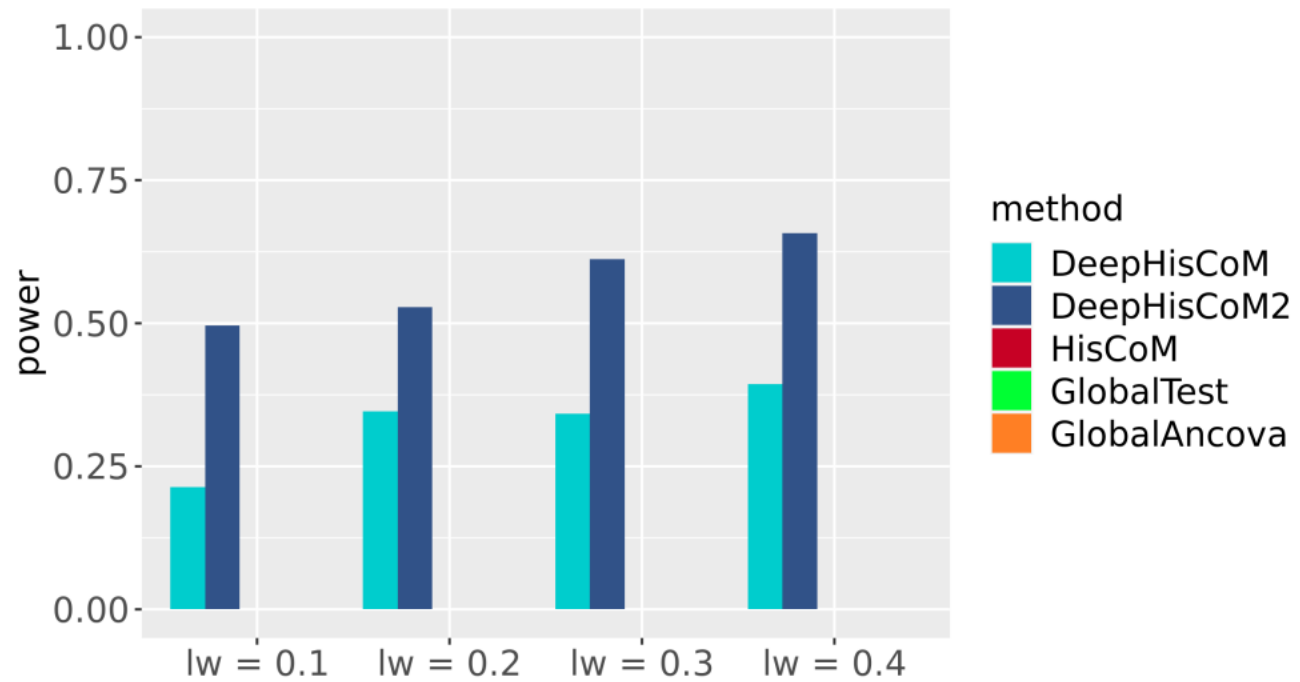
Power for Simulation Model 1: Five pathways

Linear Assumption



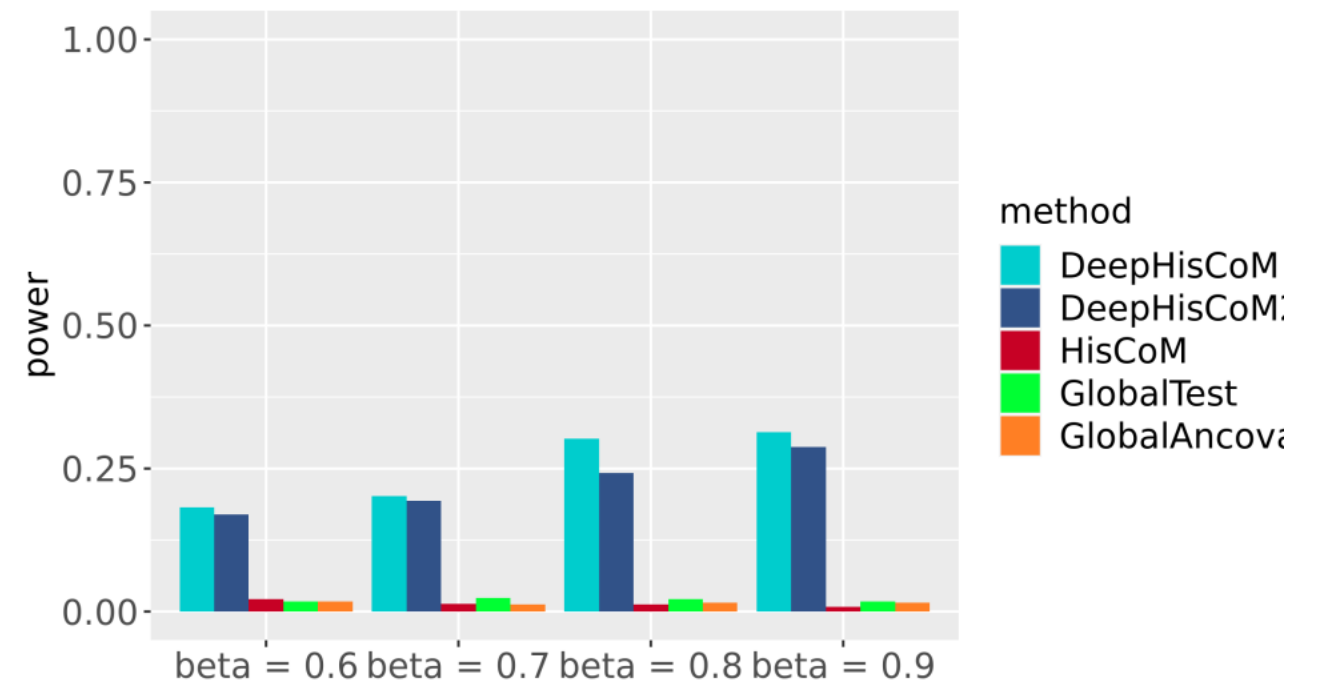
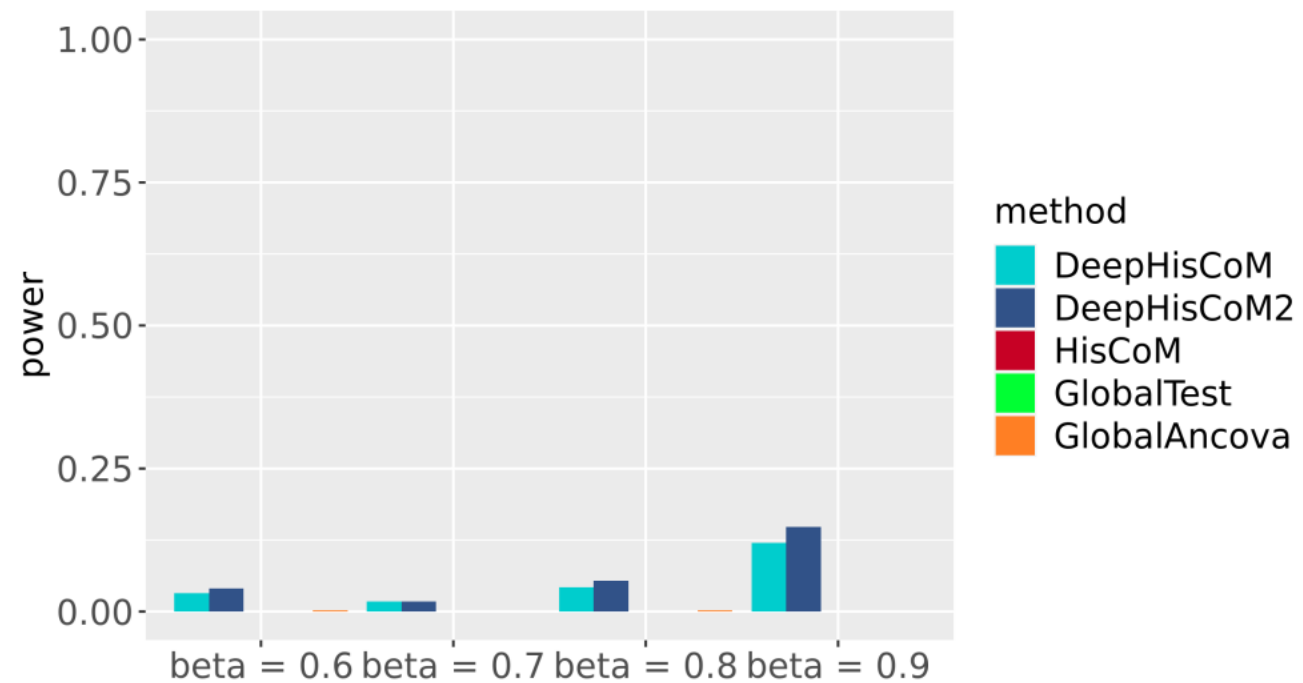
Power for Simulation Model 2: Five pathways

Interaction Assumption



Power for Simulation Model 3: Five pathways

Quadratic Assumption



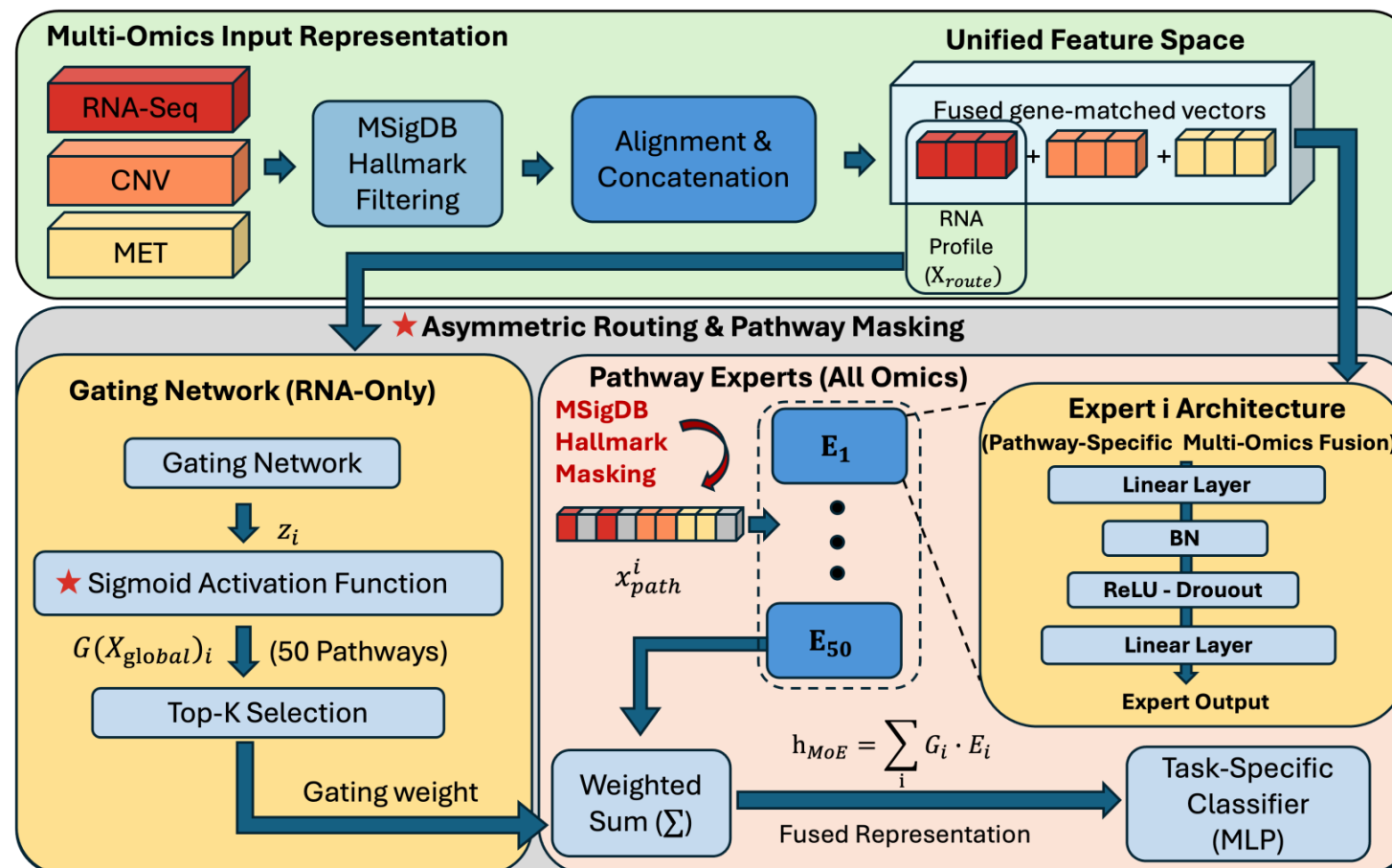
Results from two and five pathways

- DeepHisCoM tends to have slightly higher but negligible type 1 errors than the nominal levels
- For simulation model 1 when the linear effect is true, all models tend to provide similar power.
- For simulation model 2 when there is an interaction effect,
 - Deep2HisCoM >> DeepHisCoM >> HisCoM, Globaltest, GlobalANOVA
- For simulation model 3 when there is a quadratic effect,
 - DeepHisCoM ~ Deep2HisCoM >> HisCoM, Globaltest, GlobalANOVA
- In summary, DeepHisCoM **outperforms most models**

MoPE-MOI | Method

Pathway-guided mixture of experts for interpretable multi-omics integration

FIGURE 1 · MODEL ARCHITECTURE



WORKFLOW

01 Build pathway masks

50 Hallmark masks constrain 4,386 genes; align RNA, CNV and methylation by gene.

02 Route from RNA context

RNA expression provides each patient's global context for pathway routing.

03 Select co-active experts

Independent Sigmoid gates allow pathway co-activation; Top-k selects a sparse expert set.

04 Fuse multi-omics locally

Pathway experts integrate matched omics; a task head predicts subtype or survival.

MoPE-MOI | Results

PAM50 subtypes, patient-level pathway interpretation and a Cox survival extension

TCGA-BRCA · PAM50 · 20-seed hold-out mean ± SD

0.805 ± 0.059

Macro F1

0.811 ± 0.061

Balanced accuracy

Best among comparative models

SUBTYPE-SPECIFIC TOP PATHWAYS

PAM50 subtype	Top routed Hallmark pathway	Selected
Basal-like	Estrogen Response Early	29.7%
HER2-enriched	Estrogen Response Late	25.8%
Luminal A	Estrogen Response Early	19.7%
Luminal B	Myogenesis	17.2%
Normal-like	Coagulation	20.0%

Patient-specific gating weights provide pathway explanations.

COX SURVIVAL EXTENSION · BRCA

Held-out C-index · 20 seeds · 769 patients · 101 deaths

MoPE-MOI 0.633 ± 0.056

Penalized CoxPH 0.621 ± 0.091

Random Survival Forest 0.637 ± 0.060

Comparable to the two survival-specific baselines.



Discussion

Conclusion

- We developed a novel pathway analysis method **DeepHisCoM** that can detect a non-linear relationship between pathway and biological factors
- Simulation studies showed that **DeepHisCoM** tends to have higher power than other methods
- DeepHisCoM could be used for the discovery of new biomarkers

Conclusion

- Application to HCC and LC data successfully discovered HCC-related pathways
 - **lysine degradation**
 - **valine, leucine, and isoleucine biosynthesis**
 - **phenylalanine, tyrosine, and tryptophan metabolism**
- Above pathways were successfully replicated from metagenomic and RNA-Seq datasets.

Future work

- DeepHisCoM can be applicable to any types of omics data
- It can also used in integrative analysis of multi-omics data such as
 - genomics and transcriptomics
 - transcriptomics and metabolomics
 - metabolomics and metagenomics

- Parametric testing approach for HisCoM
 - For testing the significance of the pathway effect, the original HisCoM uses the permutation approach.
 - Sometimes, the permutation test is problematic for high-dimensional data because of its computational burden.
 - To account for this issue, we introduce a parametric test to get the significant value of pathways.
 - ✓ Asymptotic test with adjustments
 - Parametric hypothesis testing for pathway based hierarchical structural component models. Kamruzzaman, M., Park, T. Genes Genom 48, 831–845 (2026).
<https://doi.org/10.1007/s13258-026-01749-9>





- **HisCoM-Categ**

- All the previously developed HisCoM models can handle continuous, binary, or survival phenotypes.
- However, these available methods are not readily applicable for analyzing multinomial (nominal or ordinal) phenotypes.
- HisCoM-Categ, for identifying the association between pathways and multinomial phenotypes.

- **HisCoM-Rcateg**

- an extension of HisCoM-Categ for identifying the association between pathways and longitudinal multinomial phenotypes.
- Hierarchical structural component model for pathway analysis of multinomial phenotypes. Kamruzzaman M, Goo T, Park T. Genes Genomics. 2025 Aug;47(8):923-933. doi: 10.1007/s13258-025-01655-6. Epub 2025 Jul 3. PMID: 40608216.

- **MoPE-MOI: Interpretable multi-omics integration**

- **Objective**

Develop an interpretable multi-omics integration framework that predicts cancer subtypes and survival while identifying pathways relevant to each patient.

- **Motivation**

Existing models often overfit high-dimensional data or lack biological explanations, and fixed pathway models overlook differences between patients.

- **Novel idea**

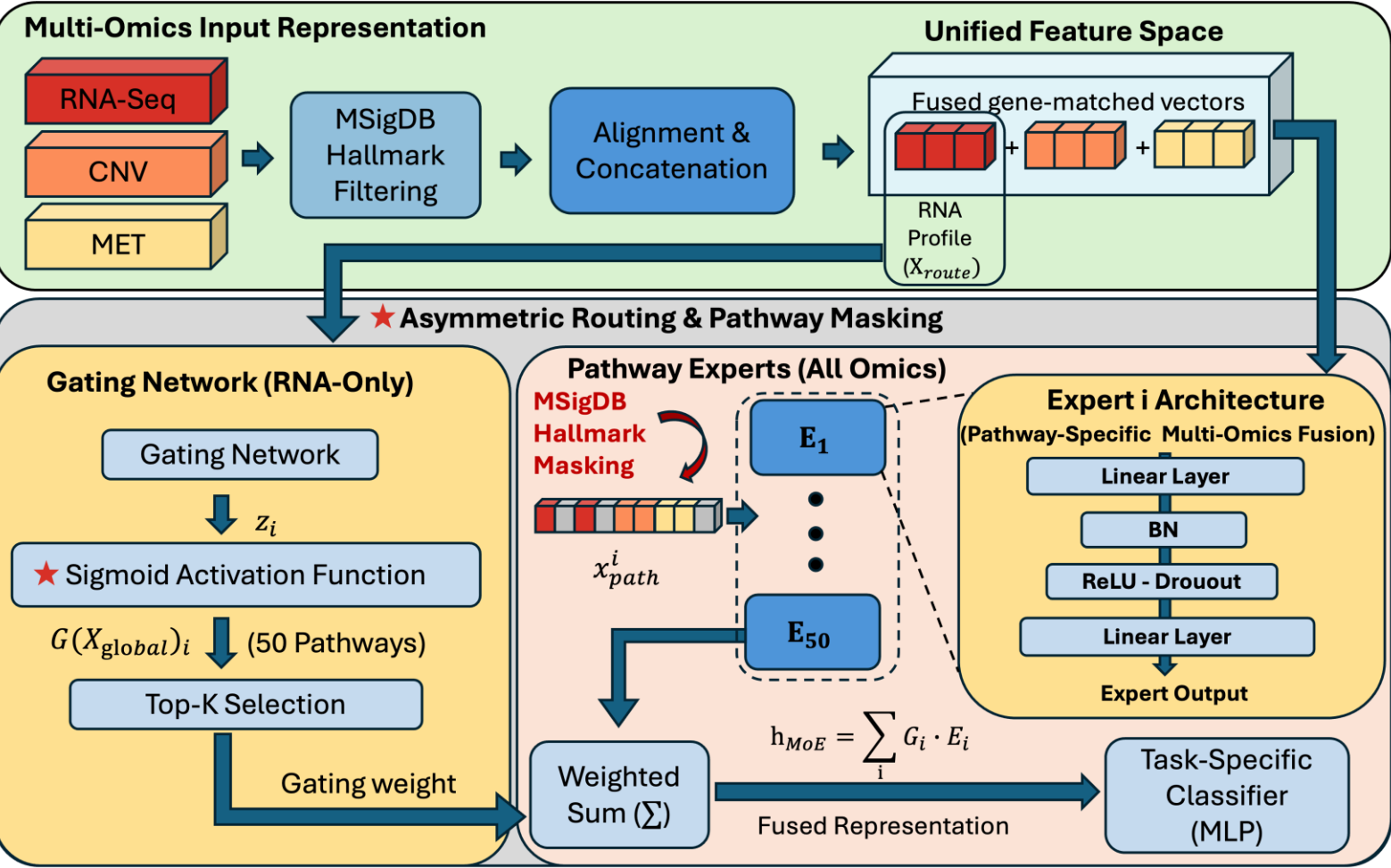
- **Pathway experts:** Restrict each expert to an MSigDB Hallmark pathway.
 - **Patient-specific routing:** Use RNA-guided, noncompetitive sigmoid gating and Top-k selection to activate multiple pathways.
 - **Multi-omics integration:** Fuse RNA, copy number and methylation data within the selected pathway experts.



MoPE-MOI: Interpretable multi-omics integration

Pathway-guided mixture of experts for interpretable multi-omics integration

WORKFLOW



01

Build pathway masks

50 Hallmark masks constrain 4,386 genes; align RNA, CNV and methylation by gene.

02

Route from RNA context

RNA expression provides each patient's global context for pathway routing.

03

Select co-active experts

Independent Sigmoid gates allow pathway co-activation; Top-k selects a sparse expert set.

04

Fuse multi-omics locally

Pathway experts integrate matched omics; a task head predicts subtype or survival.



Deep Score-Guided Gene-Gene Interaction (Epistasis) Discovery Bridging Deep Learning Representation with Rigorous Statistical inference



- **The Core Limitation:** Standard deep learning models for epistasis suffer from biological naivety (ignoring genomic coordinates & LD blocks), severe False Discovery Rate (FDR) inflation (>85%), and rare-variant gradient collapse.
- **The Hybrid Solution (DSG GLM-GGI):** A two-stage framework that projects high-dimensional genotypes into biologically constrained hypergraph modules while enforcing score-guided residual decoupling.
- **Real-World Impact:** Enables scalable, unbiased detection of complex gene-gene interactions across biobank-scale cohorts (e.g., KARE, UK Biobank).

Stage 1 — Biologically Informed Representation Learning

- **Feature Tokenizer:** Encodes SNP dosage, continuous physical base-pair distances, and chromosomal identities while preserving exact resolution.
- **LD-Biased Partitioned Attention:** Restricts attention within gene boundaries and directly modulates logits using local pairwise Linkage Disequilibrium ($LD = \text{logits}(r_{uv}^2)$) matrices to suppress noise.
- **Gene-MLP Layer (Functional Burden Gating):** Executes structured, annotation-aware adaptive pooling—gated by MAF/MAC categories and functional impact classes—to prioritize rare functional variants and prevent gradient collapse
- **HyperGIN Layer:** Combines Graph Isomorphism Networks (GIN) with 4-channel STRING interaction priors (ppi, co-expression, combined, and text) and variational hypergraphs to capture multi-gene module dependencies.



Stage 2: Unbiased Inference & 100% True Positive Recovery

- **Variance Decoupling:** Freezes representation parameters during testing and extracts residuals from GLM null models, eliminating overoptimistic feature selection bias.
- **Freedman-Lane Permutations:** Employs score statistics calibrated via residual permutations for mathematically valid Type-I error control.
- **Empirical Validation:** Will be compared to standard deep learning models (GWAS-NN, GENEPAIR) and Gene-MDR at FDR < 0.05.



Reference

- [Pathway-based approach using hierarchical components of collapsed rare variants.](#) Lee, S., Choi, S., Kim, Y. J., Kim, B. J., T2d-Genes Consortium, Hwang, H., & **Park, T.** *Bioinformatics*. 2016 Sep 1;32(17):i586-i594. doi: 10.1093/bioinformatics/btw425. Online ahead of print. PMID: 27587678
- [Hierarchical structural component modeling of microRNA-mRNA integration analysis.](#) Kim Y, Lee S, Choi S, Jang JY, **Park T.** *BMC Bioinformatics*. 2018 May 8;19(Suppl 4):75. doi: 10.1186/s12859-018-2070-0. Online ahead of print. PMID: 29745843
- [Pathway-based approach using hierarchical components of rare variants to analyze multiple phenotypes.](#) Lee S, Kim Y, Choi S, Hwang H, **Park T.** *BMC Bioinformatics*. 2018 May 8;19(Suppl 4):79. doi: 10.1186/s12859-018-2066-9. Online ahead of print. PMID: 29745849
- [Drug response prediction model using a hierarchical structural component modeling method.](#) Kim S, Choi S, Yoon JH, Kim Y, Lee S, **Park T.** *BMC Bioinformatics*. 2018 Aug 13;19(Suppl 9):288. doi: 10.1186/s12859-018-2270-7. Online ahead of print. PMID: 30367591
- [HisCoM-GGI: Hierarchical structural component analysis of gene-gene interactions.](#) Choi S, Lee S, Kim Y, Hwang H, **Park T.** *J Bioinform Comput Biol*. 2018 Dec;16(6):1840026. doi: 10.1142/S0219720018400267. Epub 2018 Oct 30. Online ahead of print. PMID: 30567476.
- [Pathway analysis of rare variants for the clustered phenotypes by using hierarchical structured components analysis.](#) Lee S, Kim S, Kim Y, Oh B, Hwang H, **Park T.** *BMC Med Genomics*. 2019 Jul 11;12(Suppl 5):100. doi: 10.1186/s12920-019-0517-4. Online ahead of print. PMID: 31296220

Reference

- [HisCoM-PAGE: Hierarchical Structural Component Models for Pathway Analysis of Gene Expression Data.](#) Mok L, Kim Y, Lee S, Choi S, Lee S, Jang JY, **Park T.** *Genes (Basel)*. 2019 Nov 14;10(11):931. doi: 10.3390/genes10110931. Online ahead of print. PMID: 31739607
- [Hierarchical structural component model for pathway analysis of common variants.](#) Jiang N, Lee S, **Park T.** *BMC Med Genomics*. 2020 Feb 24;13(Suppl 3):26. doi: 10.1186/s12920-019-0650-0. Online ahead of print. PMID: 32093692
- [HisCoM-G×E: Hierarchical Structural Component Analysis of Gene-Based Gene-Environment Interactions.](#) Choi S, Lee S, Huh I, Hwang H, **Park T.** *Int J Mol Sci*. 2020 Sep 14;21(18):6724. doi: 10.3390/ijms21186724. Online ahead of print. PMID: 32937825
- [Pathway-Based Integrative Analysis of Metabolome and Microbiome Data from Hepatocellular Carcinoma and Liver Cirrhosis Patients.](#) Kim B, Cho EJ, Yoon JH, Kim SS, Cheong JY, Cho SW, **Park T.** *Cancers (Basel)*. 2020 Sep 21;12(9):2705. doi: 10.3390/cancers12092705. Online ahead of print. PMID: 32967314
- [Identifying miRNA-mRNA Integration Set Associated With Survival Time.](#) Kim Y, Lee S, Jang JY, Lee S, Park T. *Front Genet*. 2021 Jun 29;12:634922. doi: 10.3389/fgene.2021.634922. Online ahead of print. PMID: 34267778
- [Integrative Pathway Analysis of SNP and Metabolite Data Using a Hierarchical Structural Component Model.](#) Jung T, Jung Y, Moon MK, Kwon O, Hwang GS, **Park T.** *Front Genet*. 2022 Mar 24;13:814412. doi: 10.3389/fgene.2022.814412. Online ahead of print. PMID: 35401680
- [Hierarchical Structured Component Analysis for Microbiome Data Using Taxonomy Assignments.](#) Kim SA, Kang N, **Park T.** *IEEE/ACM Trans Comput Biol Bioinform*. 2022 May-Jun;19(3):1302-1312. doi: 10.1109/TCBB.2020.3039326. Epub 2022 Jun 3. Online ahead of print. PMID: 33211665.

Reference

- [Kernel-based hierarchical structural component models for pathway analysis.](#) Hwangbo S, Lee S, Lee S, Hwang H, Kim I, **Park T.** *Bioinformatics*. 2022 Apr 23:btac276. doi: 10.1093/bioinformatics/btac276. Online ahead of print.PMID: 35460238
- [DeepHisCoM: deep learning pathway analysis using hierarchical structural component models.](#) Park C, Kim B, Park T. *Brief Bioinform.* 2022 May 23:bbac171. doi: 10.1093/bib/bbac171. Online ahead of print.PMID: 35598329
- [Kernel-based hierarchical structural component models for pathway analysis on survival phenotype.](#) Hwangbo S, Lee S, Hosain MM, Goo T, Lee S, Kim I, Park T. *Genes Genomics*. 2024 Dec;46(12):1415-1421. doi: 10.1007/s13258-024-01569-9. Epub 2024 Sep 26.
- [Hierarchical structural component model for pathway analysis of multinomial phenotypes.](#) Kamruzzaman M, Goo T, Park T. *Genes Genomics*. 2025 Aug;47(8):923-933. doi: 10.1007/s13258-025-01655-6. Epub 2025 Jul 3. PMID: 40608216.
- [Parametric hypothesis testing for pathway based hierarchical structural component models.](#) Kamruzzaman, M., Park, T. *Genes Genom* 48, 831–845 (2026). <https://doi.org/10.1007/s13258-026-01749-9>

Reference

- [SNP-PRAGE: SNP-based parametric robust analysis of gene set enrichment](#). Lee J, Ahn S, Oh S, Weir B, Park T. *BMC Syst Biol*. 2011;5 Suppl 2(Suppl 2):S11. doi: 10.1186/1752-0509-5-S2-S11. Epub 2011 Dec 14. PMID: 22784568 Free PMC article.
- [Personalized identification of altered pathways in cancer using accumulated normal tissue data](#). Ahn T, Lee E, Huh N, Park T. *Bioinformatics*. 2014 Sep 1;30(17):i422-9. doi: 10.1093/bioinformatics/btu449.
- [Gene-set association tests for next-generation sequencing data](#). Lee J, Kim YJ, Lee J; T2D-Genes Consortium; Kim BJ, Lee S, Park T. *Bioinformatics*. 2016 Sep 1;32(17):i611-i619. doi: 10.1093/bioinformatics/btw429. PMID: 27587681
- [PATHOME-Drug: a subpathway-based poly-pharmacology drug-repositioning method](#). Nam S, Lee S, Park S, Lee J, Park A, Kim YH, Park T. *Bioinformatics*. 2021 Sep 13:btab566. doi: 10.1093/bioinformatics/btab566. Online ahead of print. PMID: 34515762

Acknowledgement

- **Virginia Tech**
 - Inyoung Kim
- **McGill University**
 - Heungsun Hwang
- **Sejong University**
 - Seungyeoun Lee
- **Hanyang University (ERICA)**
 - Sungkyoung Choi
- **Seoul National University Hospital**
 - Sungyoung Lee
 - Suhyun Hwangbo
- **Seoul National University College of Medicine**
 - Joo-Youn Cho
 - Eun Ju Cho
 - Jung-Hwan Yoon
 - Jin-Young Jang
- **Ajou University School of Medicine**
 - Soon Sun Kim
 - Jae Youn Cheong
 - Sung Won Cho

Funding

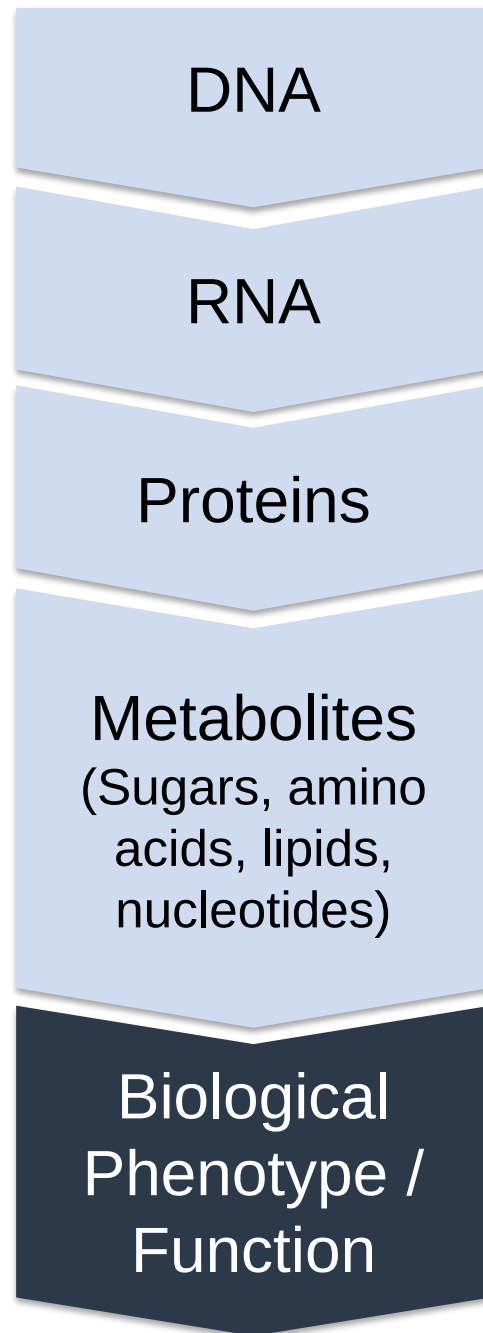


Bio-Synergy Research Center
National Research Foundation of Korea



Korea Post-Genome project
Ministry of Health & Welfare

National Research Foundation of Korea



Genomics

Transcriptomics

Proteomics

Metabolomics

- Metabolites are the substrates and products of metabolism that drive essential cellular functions, such as energy production and storage, signal transduction and apoptosis.
- There is growing motivation to better ascertain their specific functions, as well as to understand their physiological roles
- Identification of metabolites that are associated with particular phenotypes can be done by implementing various metabolomic analyses
- Metabolomics, which is the profiling of metabolites in biofluids, cells and tissues, is routinely applied as a tool for biomarker discovery.

Metabolite and disease

