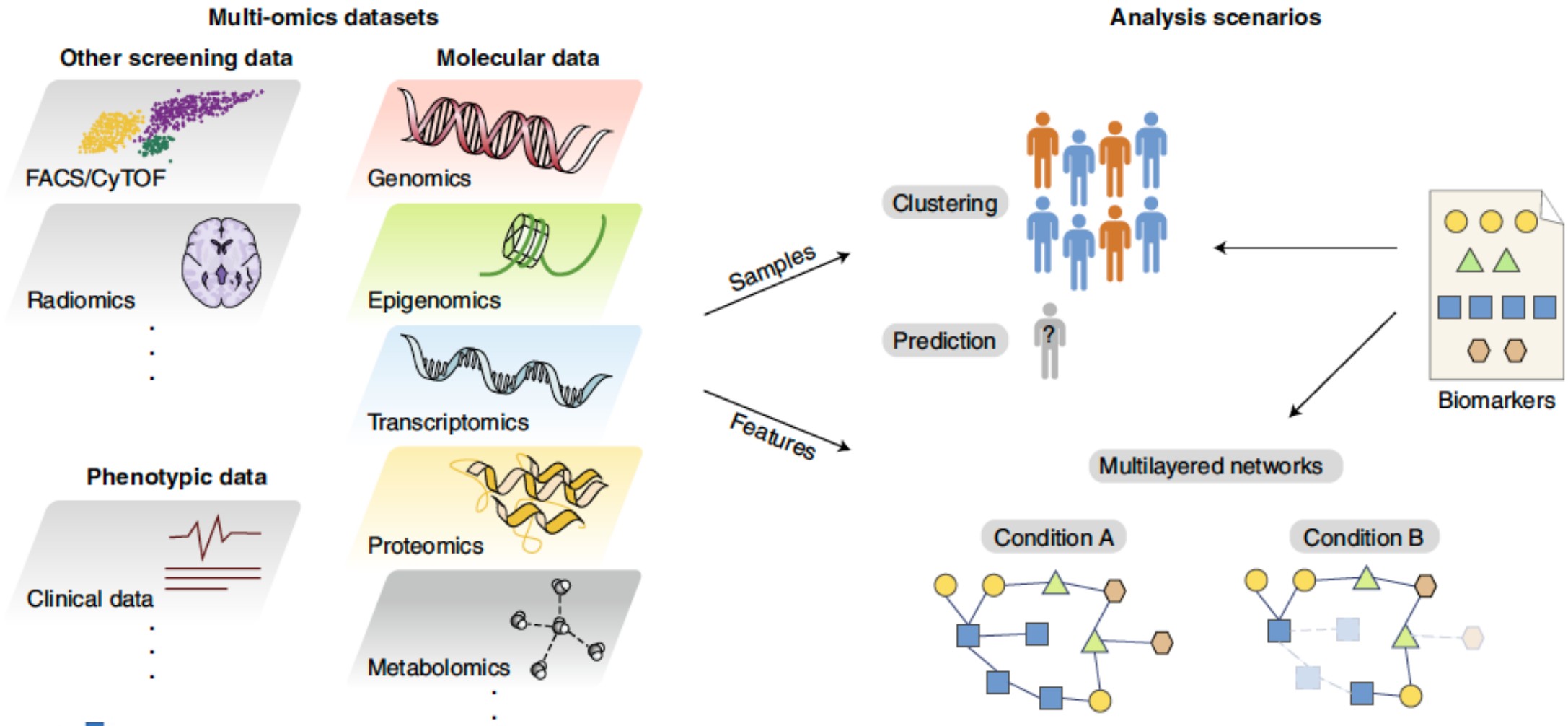


Figure from Tarazona et al. (2021) *Nature Computational Science*

Selective Introduction of Multi-Omics Analysis

American Statistical Association (ASA):

Section on Statistics in Genomics and Genetics (SSGG)

Directors:

George Tseng (University of Pittsburgh)

Katerina Kechris (University of Colorado Denver)

Instructors:

Session 1: Wenjia Wang (University of Pittsburgh)

Session 2: Sierra Niemiec (University of Colorado Denver)

Session 3: Jack Pattee (University of Colorado Denver)

Session 4: Rick Chang (University of Pittsburgh)

Acknowledgement of support from SSGG:

Nancy Zhang (Section Chair 2023)

Michael C. Wu (Past Chair 2022)

Yijuan Hu (Treasurer)

Yuchao Jiang (Communications Officer)

Housekeeping

- Questions
  - Ask questions in chat
  - One lecturer will teach, other lecturers will monitor the chat
  - You can also send questions privately to a particular lecturer if you prefer
- Video recording
  - Videos recording will be shared through **Section on Statistics in genomics and genetics**
- GitHub site: https://github.com/KechrisLab/ASAShortCourse-MultiOmics
- Break
  - We will have a 5 minutes break for each lecture.

Short Course Overview

- Multi-omics studies now common in small groups + large consortium studies
- Many statistical and computational challenges

Learning objectives

1. Learn about the landscape of multi-omics problems and analysis methods.
2. Understand some of the theoretical justification behind selected multi-omics methods.
3. Gain hands-on experience with multi-omics analysis software and data applications.

Short Course Overview

- Prerequisites: basic knowledge of statistics, molecular biology, genetics, and familiarity with R programming
- Each session has 2 parts
  - 1st part: motivation, basic principles, representative tools, and examples
  - 2nd part: real-time lab session - provide hands-on experience and lasting take-home messages
- Lab sessions:
  - Reproducible examples, relevant data and annotated code through GitHub (https://github.com/KechrisLab/ASAShortCourse-MultiOmics)
  - Can practice within class or review at their own pace after the short course.

Outline

- Lecture 1 (Wenjia Wang)
  - Brief introduction
  - Example method: **MetaOmics**
- Lecture 2 (Sierra Niemiec)
  - Unsupervised clustering of multi-omics data
  - Example method: **MOVICS**
- Lecture 3 (Jack Pattee)
  - Dimension reduction for multi-omics data
  - Example method: **JIVE**
- Lecture 4 (Rick Chang)
  - Multi-omics causal mediation analysis
  - Single cell multi-omics analysis
  - Example methods: **HIMA, Seurat**

AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

SECTION ON STATISTICS IN GENOMICS AND GENETICS

https://bit.ly/3lK4Ja5

https://bit.ly/3DKXyos

https://bit.ly/3AOZBWy

https://bit.ly/3p2aK49

https://bit.ly/30jXvBp

Lecture 1
Overview of Multi-Omics Data Analysis and Horizontal Data Integration

April 11, 2023

Instructor: **George Tseng**
Wenjia Wang

Outline

I. Background of Multi-Omics Data Integration

a) Why integrate omics data?

b) Multi-omics data source

c) Common analysis themes and examples

d) Overview of omics data integration

II. Methods for Vertical Multi-Omics Data Integration

a) Parallel integration approaches

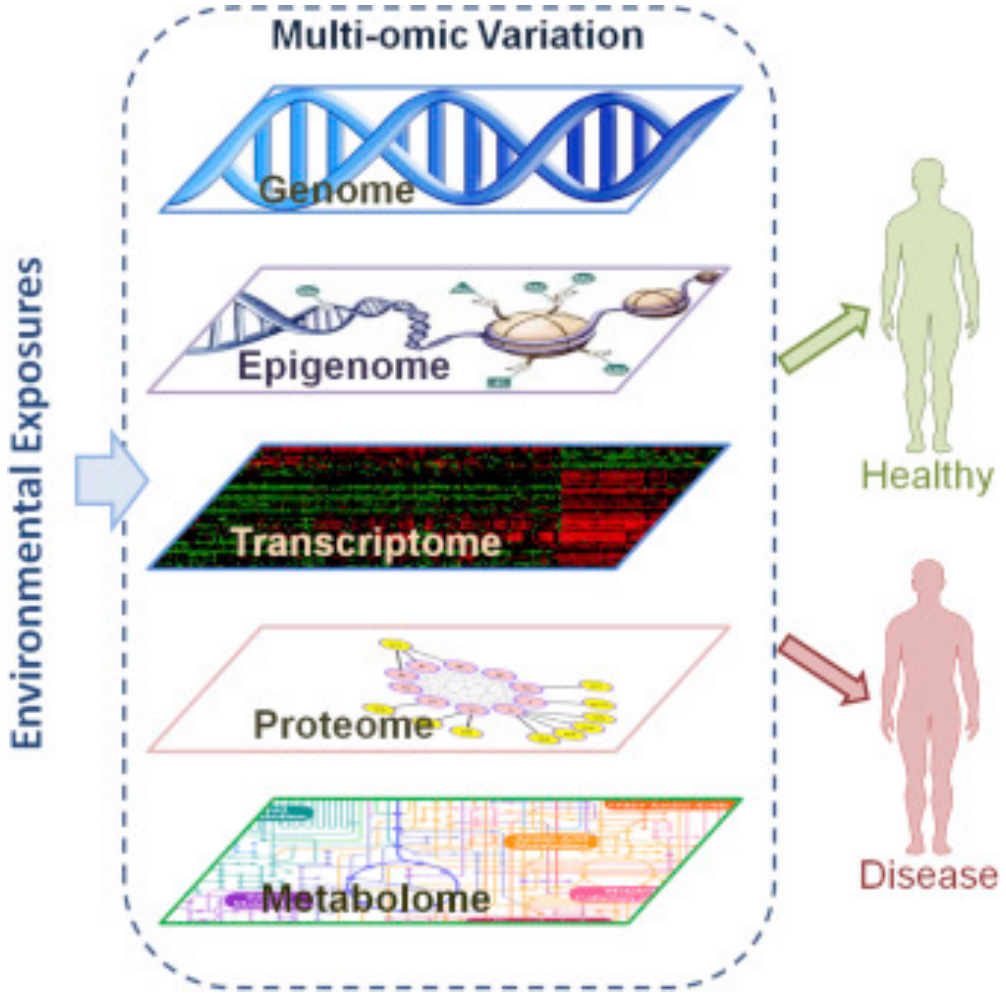
b) Hierarchical integration approaches

III. Horizontal Omics Data Integration

IV. Lab Session: MetaOmics

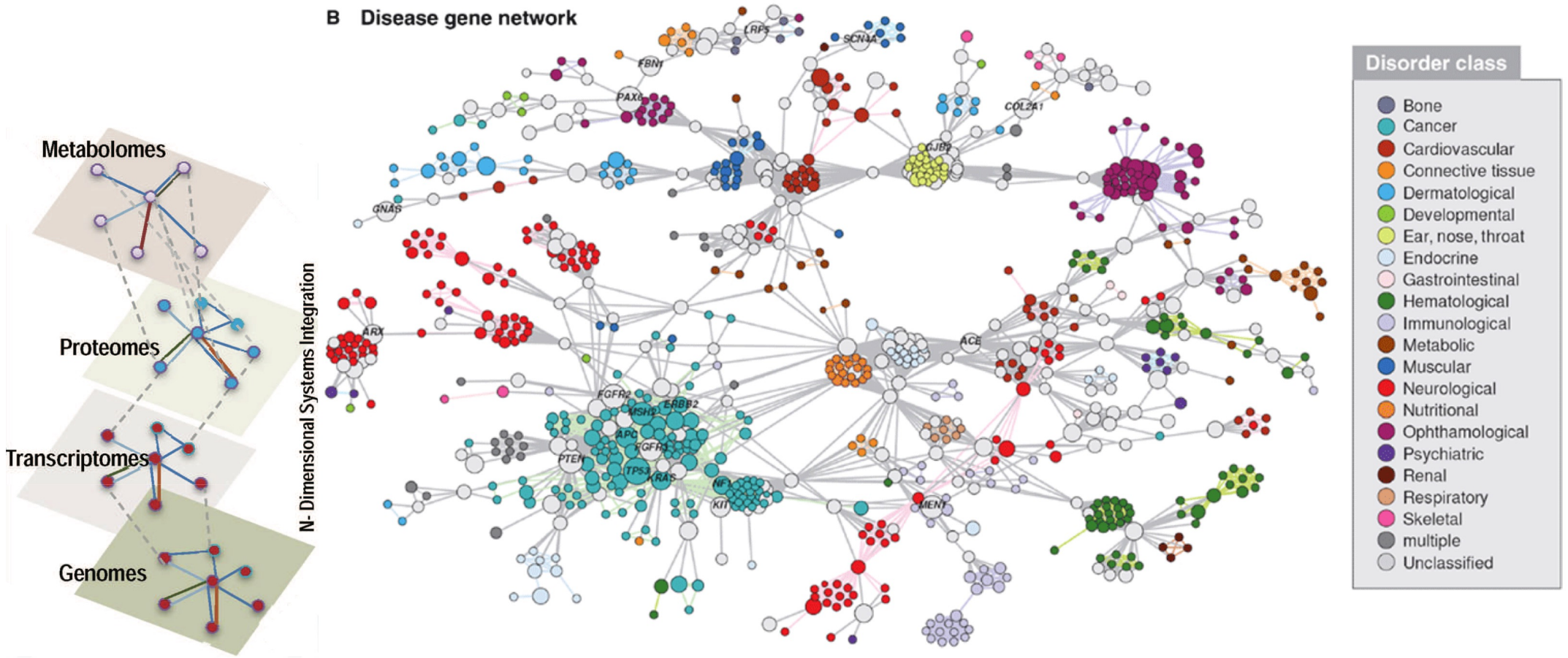
Why Integrate Omics Data?

- Use of two or more omics data sets (e.g., epigenomics, transcriptomics, metabolomics)
- Confirm or gain new insights that may not be possible using single-omics data
- Attain systems perspective for biological processes and disease mechanisms



Sun & Hu (2016) *Advances in Genetics*

Why Integrate Omics Data?



Public Data Sources

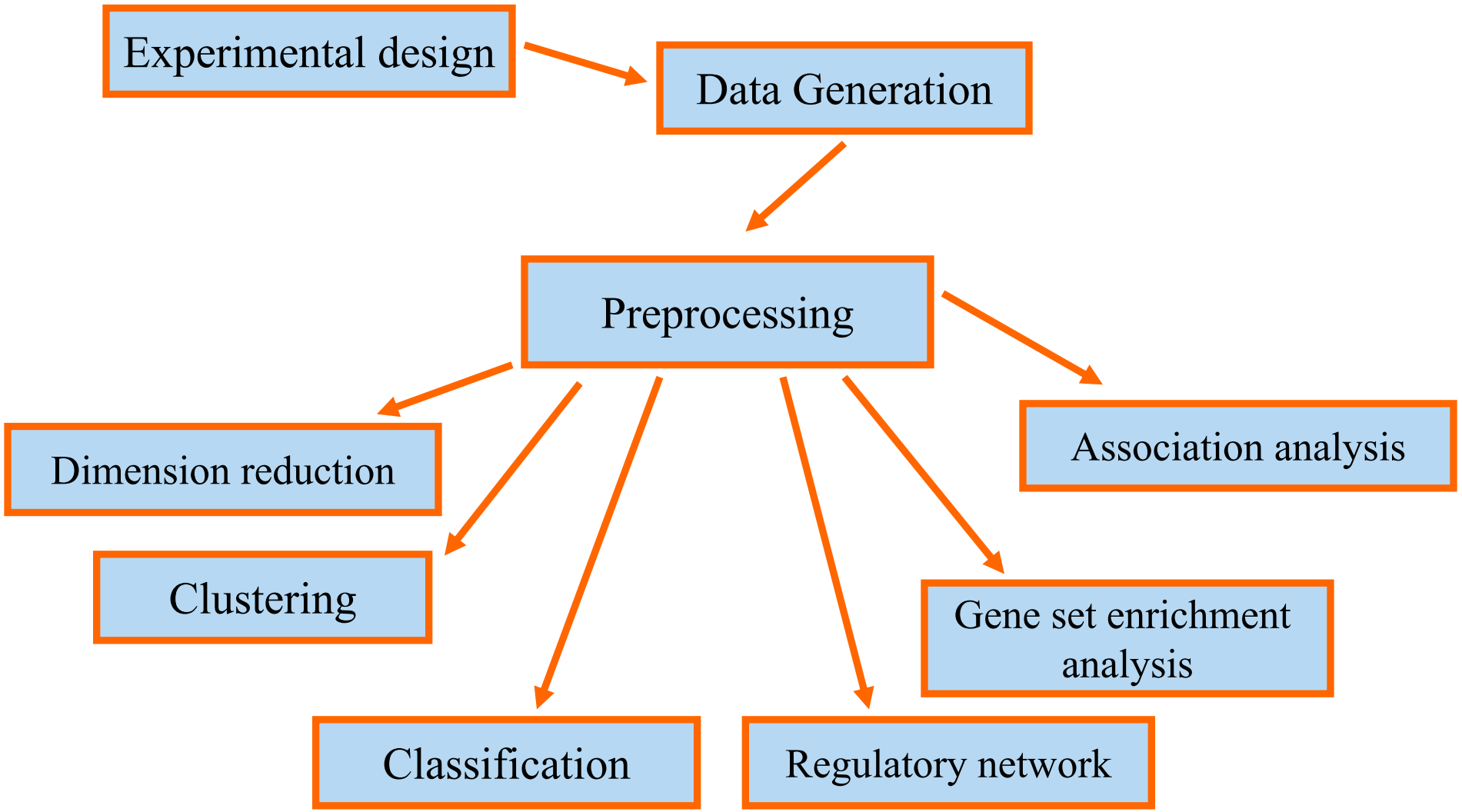
Table 1. List of multi-omics data repositories.

DATA REPOSITORY	WEB LINK	DISEASE	TYPES OF MULTI-OMICS DATA AVAILABLE
The Cancer Genome Atlas (TCGA)	https://cancergenome.nih.gov/	Cancer	RNA-Seq, DNA-Seq, miRNA-Seq, SNV, CNV, DNA methylation, and RPPA
Clinical Proteomic Tumor Analysis Consortium (CPTAC)	https://cptac-data-portal.georgetown.edu/cptacPublic/	Cancer	Proteomics data corresponding to TCGA cohorts
International Cancer Genomics Consortium (ICGC)	https://icgc.org/	Cancer	Whole genome sequencing, genomic variations data (somatic and germline mutation)
Cancer Cell Line Encyclopedia (CCLE)	https://portals.broadinstitute.org/ccle	Cancer cell line	Gene expression, copy number, and sequencing data; pharmacological profiles of 24 anticancer drugs
Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)	http://molonc.bccrc.ca/aparicio-lab/research/metabric/	Breast cancer	Clinical traits, gene expression, SNP, and CNV
TARGET	https://ocg.cancer.gov/programs/target	Pediatric cancers	Gene expression, miRNA expression, copy number, and sequencing data
Omics Discovery Index	https://www.omicsdi.org	Consolidated data sets from 11 repositories in a uniform framework	Genomics, transcriptomics, proteomics, and metabolomics

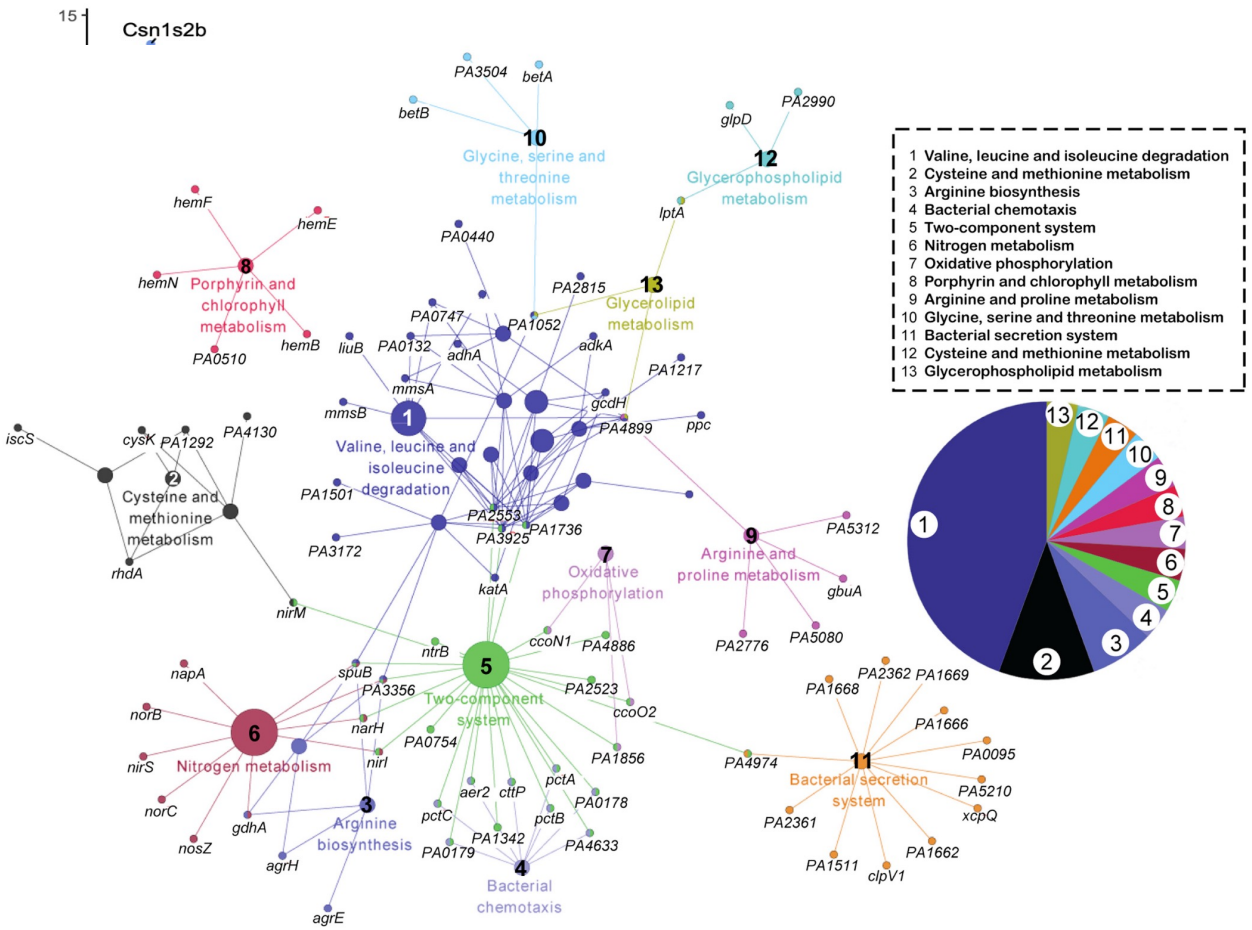
Abbreviations: CNV, copy number variation; miRNA, microRNA; RPPA, reverse phase protein array; SNP, single-nucleotide polymorphism; SNV, single-nucleotide variant.

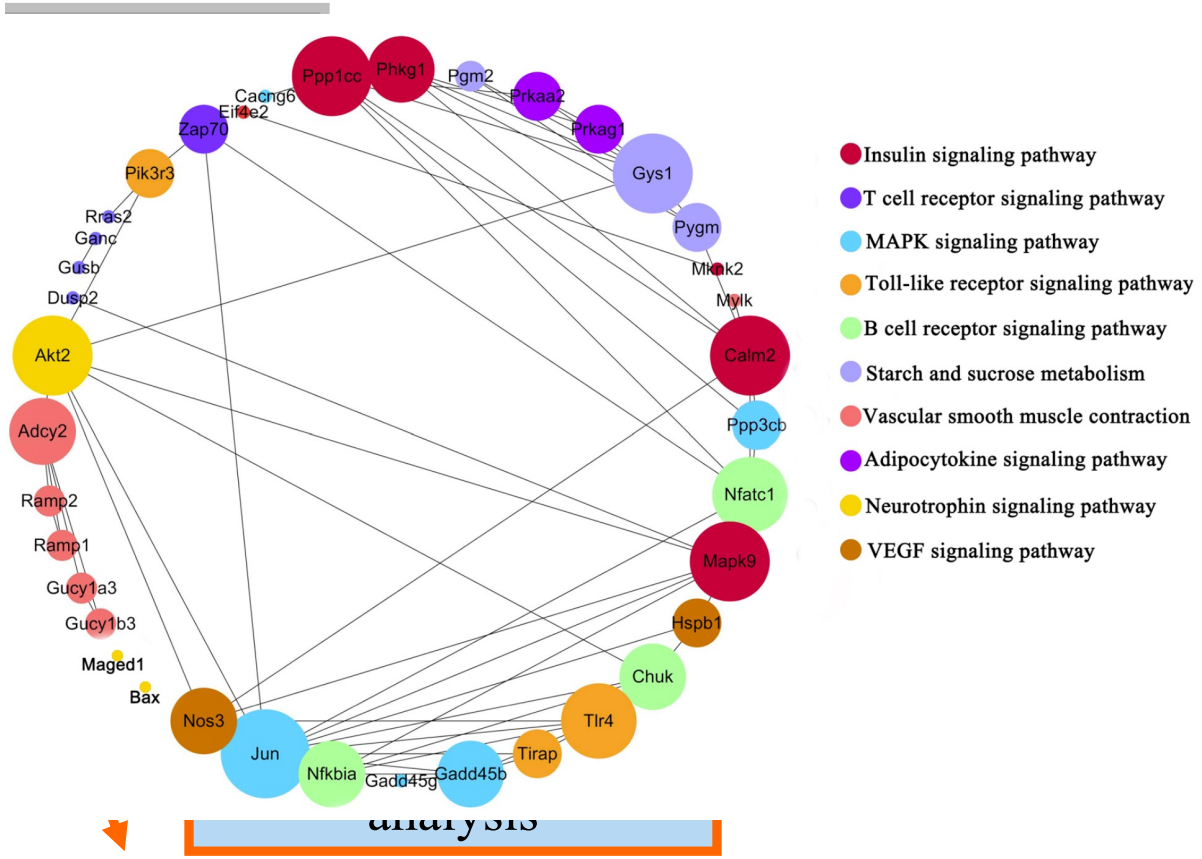
Subramanian et al. (2020) *Bioinformatics and Biology insights*

Types of Multi-Omics Data Analysis



Types of Multi-Omics Data Analysis



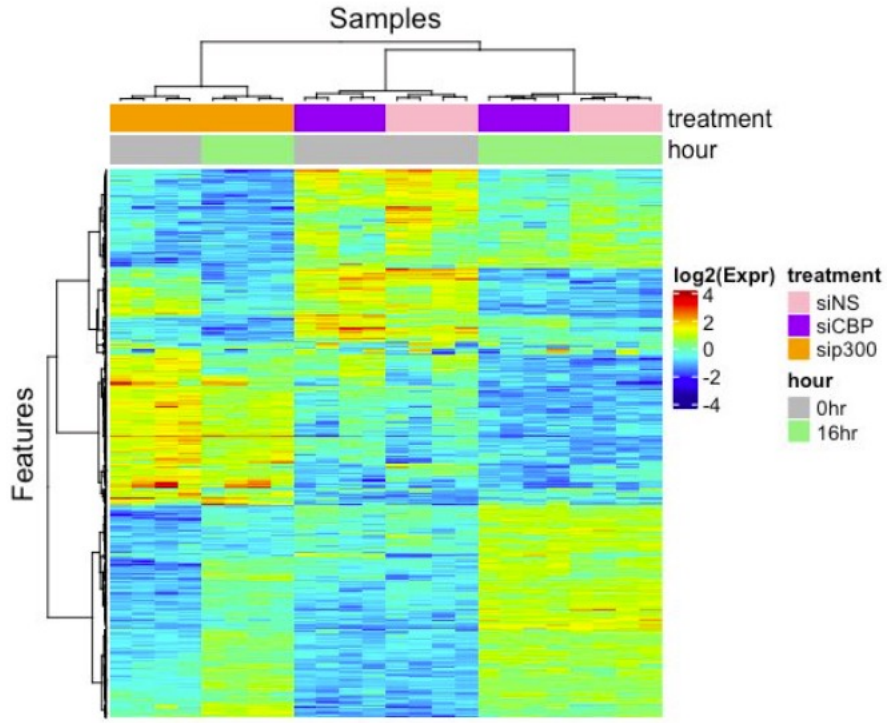


Classification

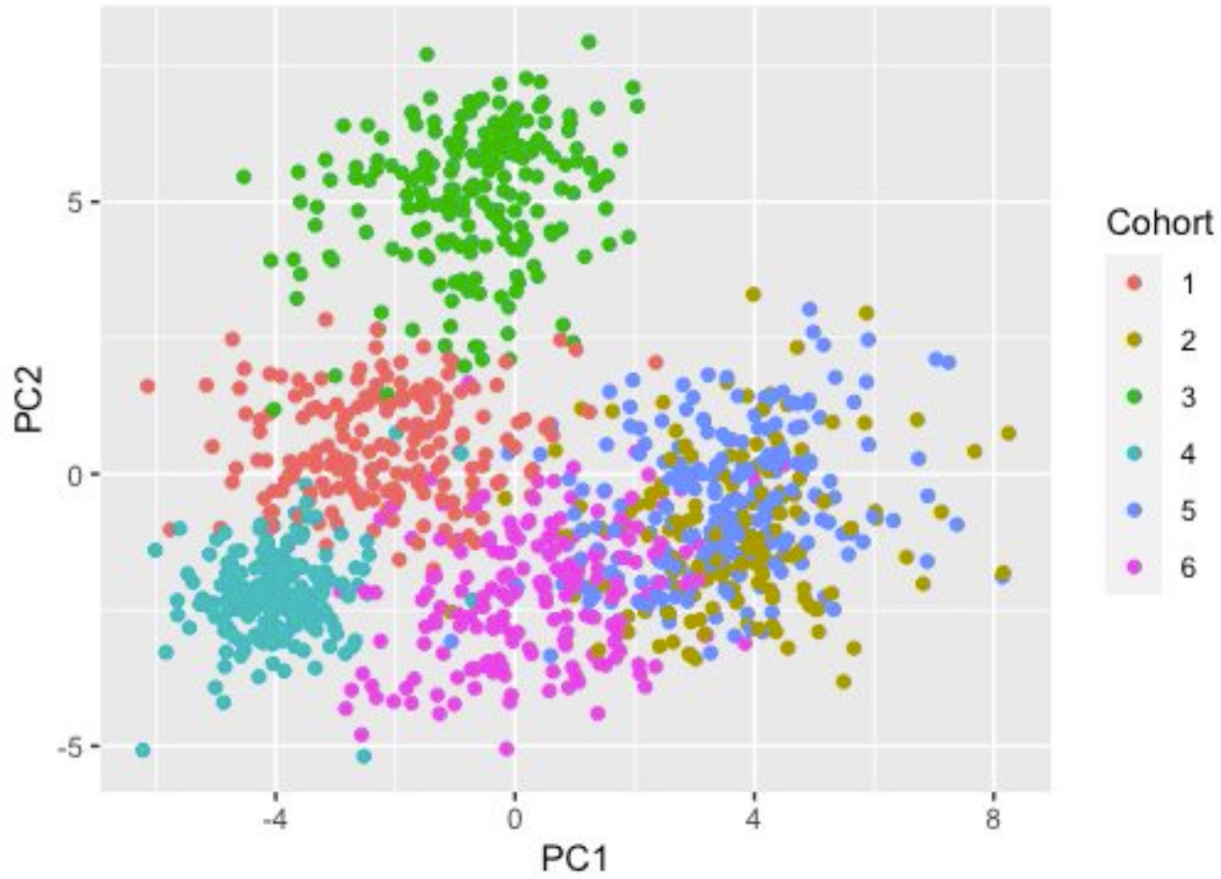
Regulatory network

Gu et al. (2021) *Sci Rep*; Zhao et al. (2017) *Front Microbiol*.

Types of Multi-Omics Data Analysis



Clustering

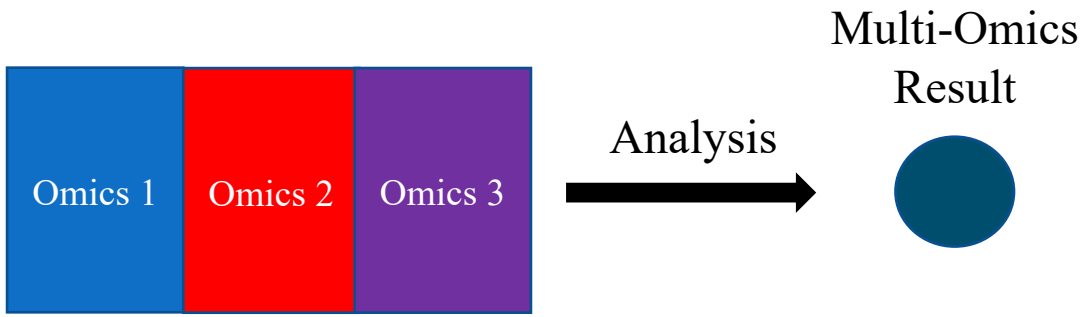


Classification

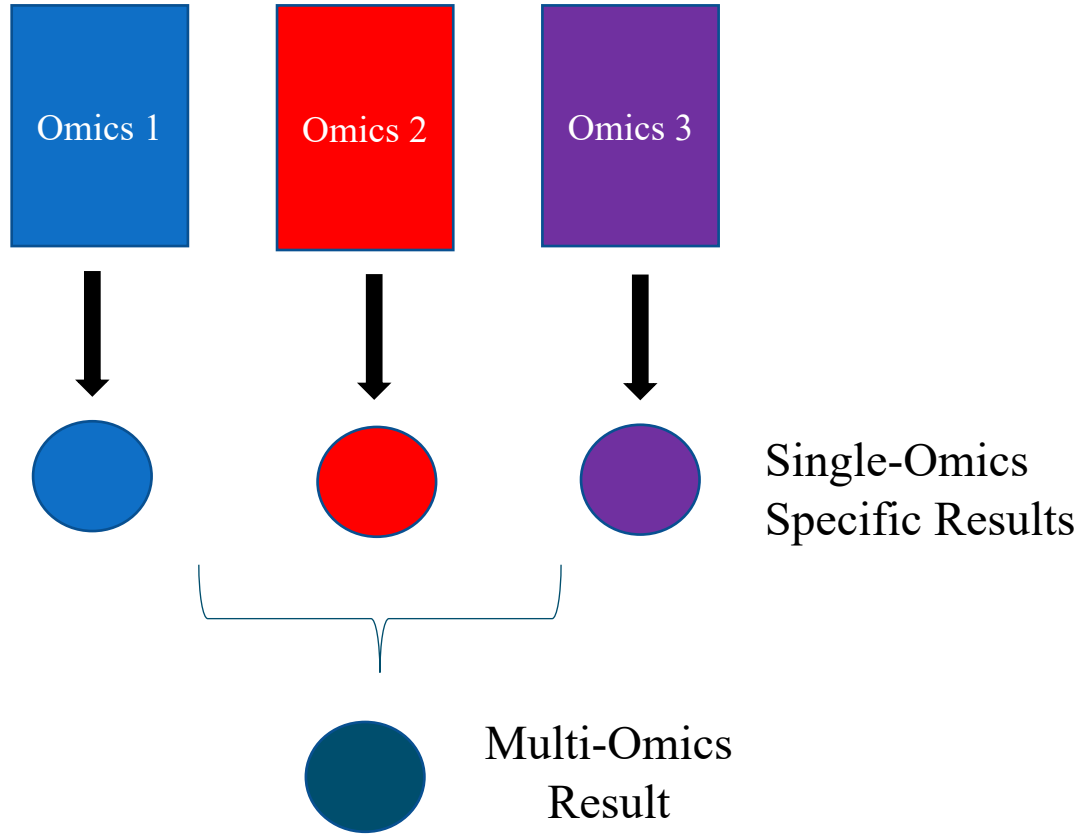
Regulatory network

When to Integrate?

1. Early integration
(concatenated or separated)

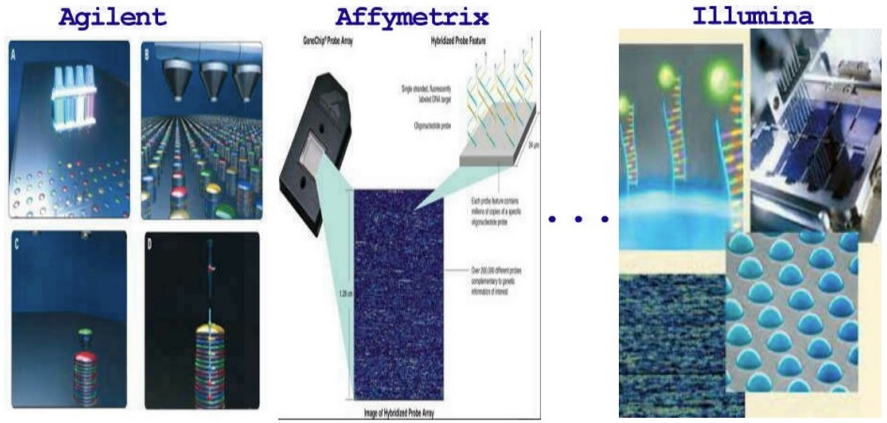


2. Late integration
(analyze separately, then integrate)



Horizontal Multi-Omics Integration

Combine microarray studies { **Cross-lab**, **Cross-sample cohort**, **Cross-platform**



D_1 D_2 .... D_K

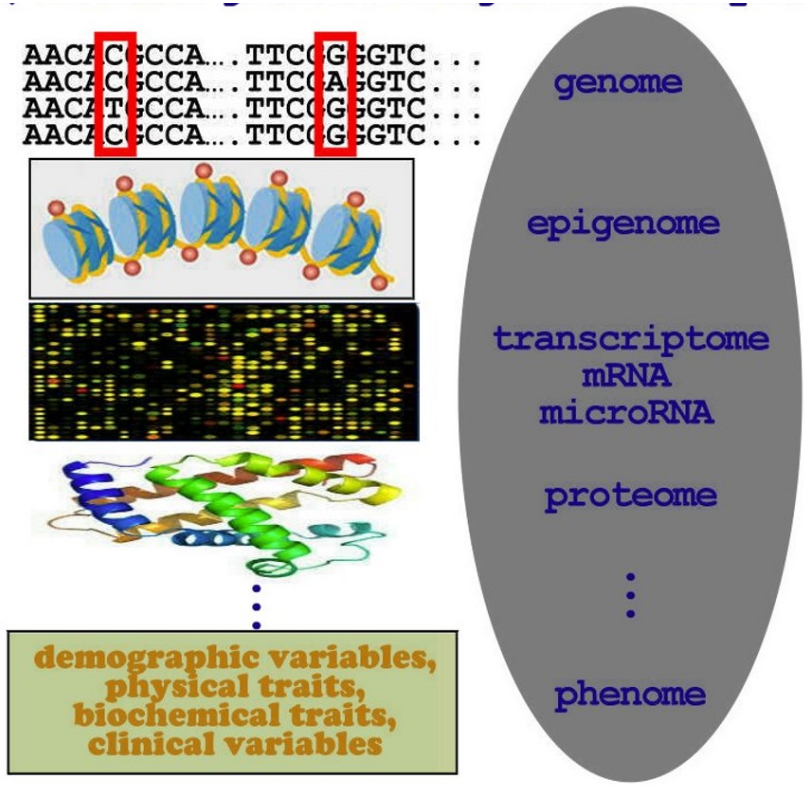
G genes× S_1 samples **G genes× S_2 samples** **G genes× S_K samples**

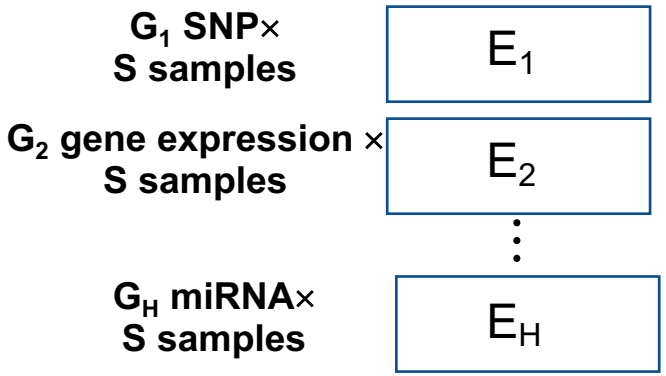
- **Same kind of omics data on K different samples or subject cohorts** (GWAS, gene expression, methylation, eQTL…)
- **Increase statistical power and generate robust discoveries**

LECTURE 1: Combining multiple genomic studies for horizontal meta-analysis and data integration (MetaOmics)

Tseng et al. (2012) *Nucleic Acids Res.*

Vertical Multi-Omics Integration





- Same samples or subject cohort analyzed using different omics technology
- Understand the complex biology and diseases systematically and holistically

Tseng et al. (2012) *Nucleic Acids Res.*

Outline

I. Background of Multi-Omics Data Integration

a) Why integrate omics data?

b) Multi-omics data source

c) Common analysis themes and examples

d) Overview of omics data integration

II. Methods for Vertical Multi-Omics Data Integration

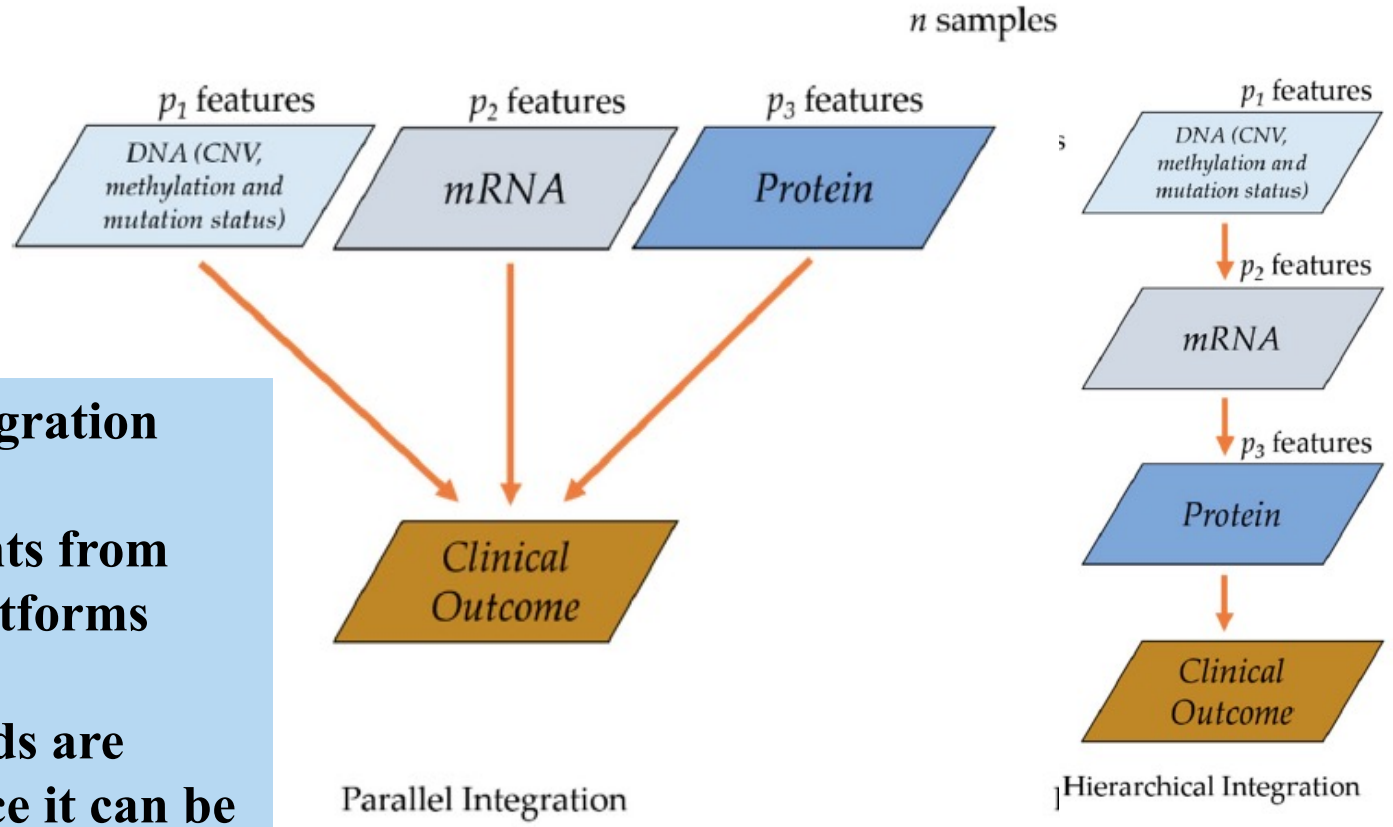
a) Parallel integration approaches

b) Hierarchical integration approaches

III. Horizontal Omics Data Integration

IV. Lab Session: MetaOmics

Vertical Multi-Omics Integration



- **<u>Parallel</u> integration treat omics measurements from different platforms equally**
- **Most methods are parallel, since it can be easily generalized to arbitrary number & types of omics data.**

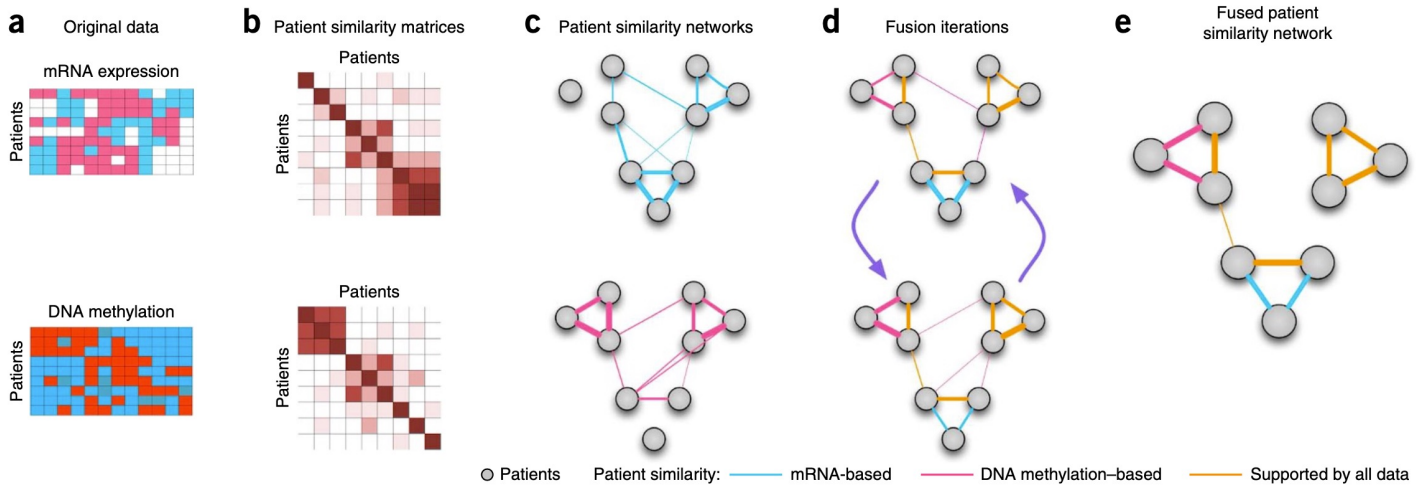
Vertical Integration Scheme: Parallel vs Hierarchical Integration

- **<u>Hierarchical</u> integration incorporates prior knowledge of regulatory relationship among different omics data**
- **Hierarchical integration can more closely reflect the biological nature of multidimensional data, but it lacks generalizability.**

Wu et al. (2019). *High Throughput;* Hasin et al. (2017). *Genome Biology.*

Parallel Multi-Omics Integration

- **Similarity Network Fusion method (SNF)** uses networks of samples as a basis for integration to cluster the samples.



LECTURE 2: Clustering of Samples (MOVICS)

Instructor: **Sierra Niemiec**

- **Joint and Individual Variation Explained (JIVE)** decomposes the concatenated data into a sum of three terms: (1) **joint structure** between data types; **(2) structure individual** to each data type; **(3) residual noise**.

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix} : p \times n, \xrightarrow{\text{scale}} X^{\text{scaled}} = \begin{bmatrix} X_1^{\text{scaled}} \\ \vdots \\ X_k^{\text{scaled}} \end{bmatrix} \xrightarrow{\text{decompose}}$$

$$JA_i^T = 0_{p \times p_i} \text{ for } i = 1, \ldots, k$$

$$\begin{aligned} X_1 &= J_1 + A_1 + \varepsilon_1 \\ &\vdots \\ X_k &= J_k + A_k + \varepsilon_k, \end{aligned}$$

J: joint structure matrix of rank r

A_i: individual structure of rank r_i

$$p = p_1 + p_2 + \ldots + p_k.$$

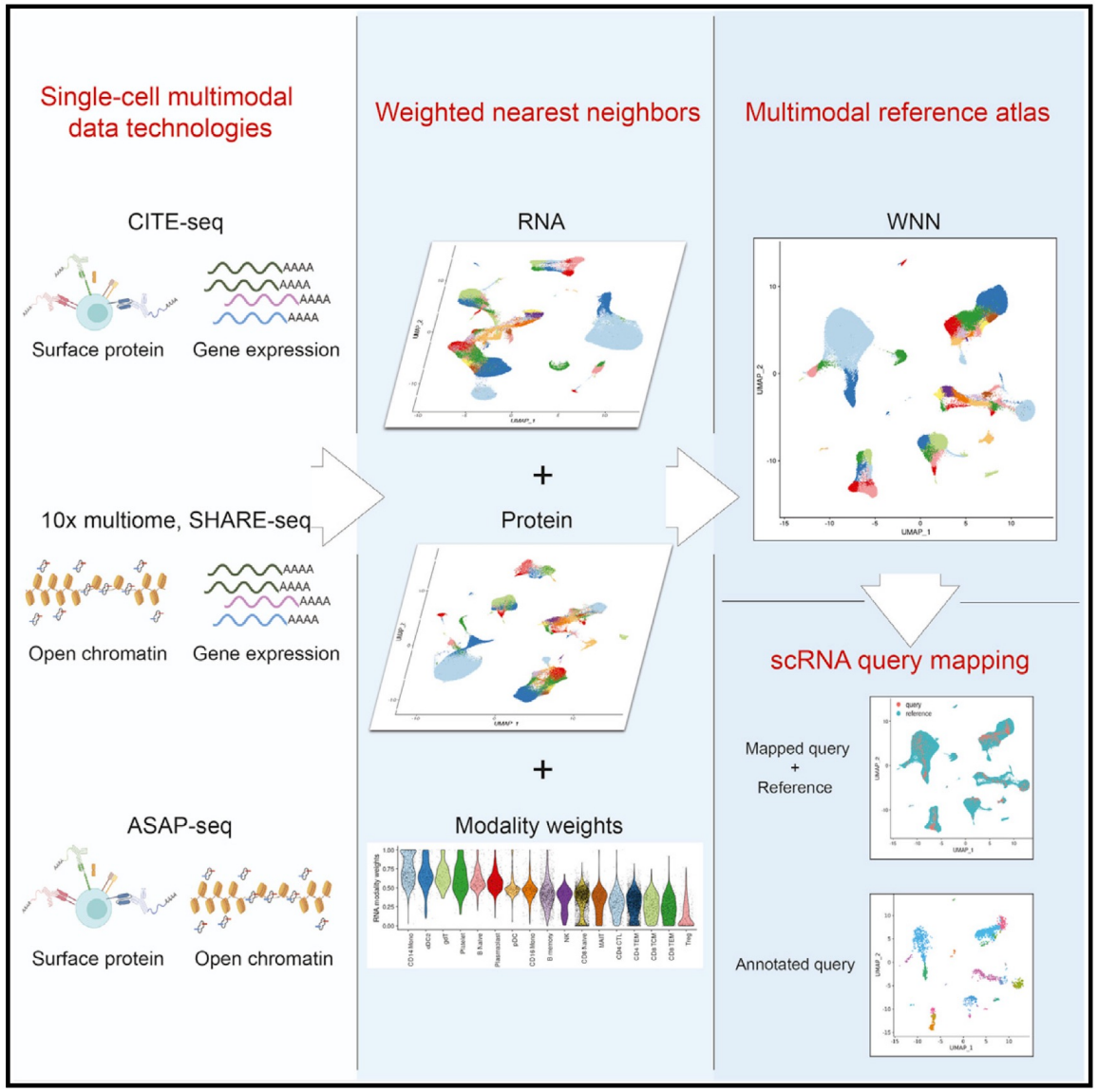
Concatenate data matrices from K platform

LECTURE 3: Dimension Reduction (JIVE)

Instructor: Dr. **Jack Pattee**

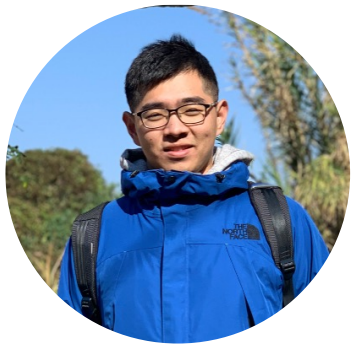
Wang et al. (2014) *Nat Methods*; Lock et al. (2013) *The Annals of Applied Statistics*

Parallel Multi-Omics Integration



Hao et al. (2021) *Cell*

LECTURE 4b: Single-cell Multi-Omics Analysis (Seurat)

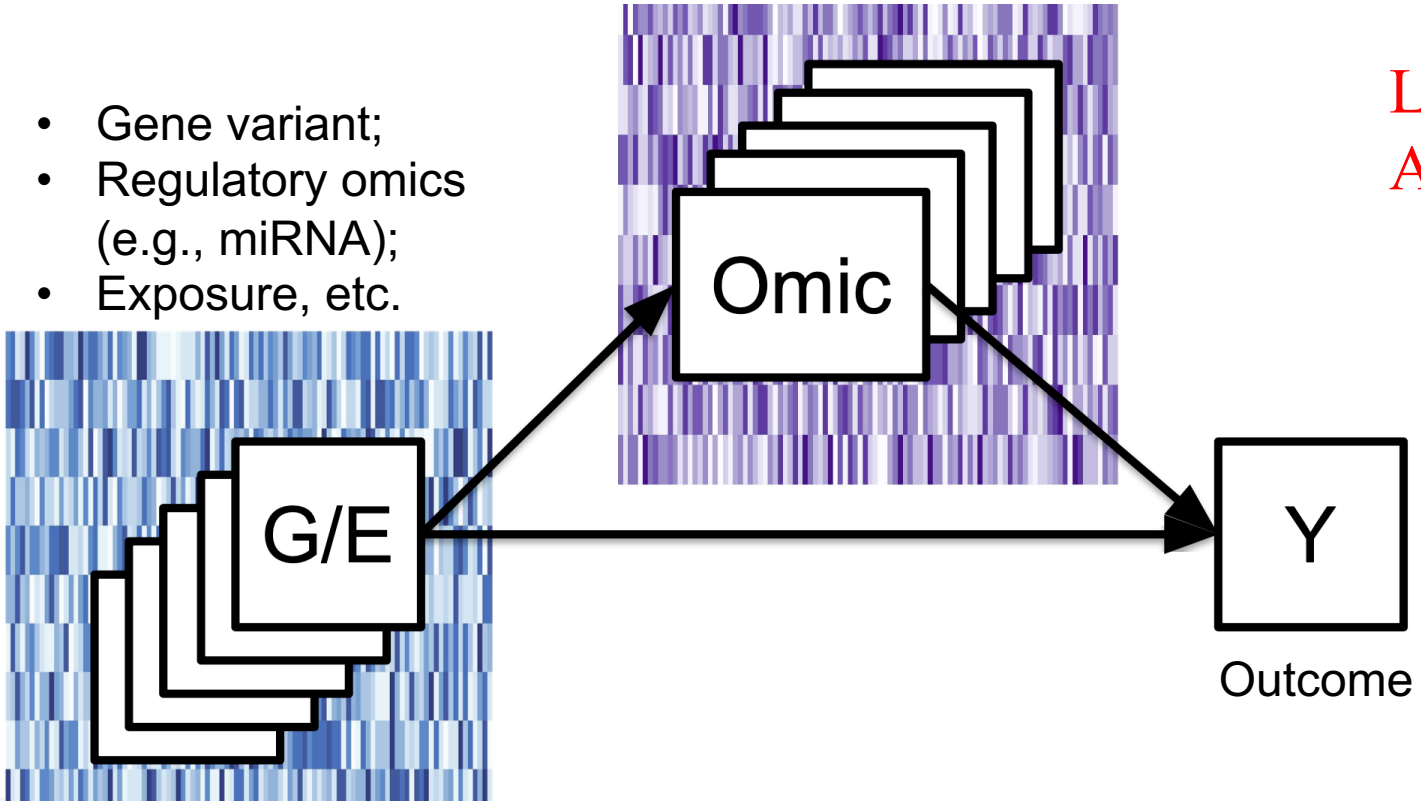


Instructor: **Rick Chang**

Hierarchical Multi-Omics Integration

- Transcriptomics;
- Proteomics;
- Metabolomics, etc.

- Gene variant;
- Regulatory omics (e.g., miRNA);
- Exposure, etc.



LECTURE 4a: Causal Mediation Analysis for Multi-Omics Analysis

Zhang et al. (2016) *Bioinformatics*; Song et al. (2020). BAMA. *Biometrics.*

Outline

I. Background of Multi-Omics Data Integration

a) Why integrate omics data?

b) Multi-omics data source

c) Common analysis themes and examples

d) Overview of omics data integration

II. Methods for Vertical Multi-Omics Data Integration

a) Parallel integration approaches

b) Hierarchical integration approaches

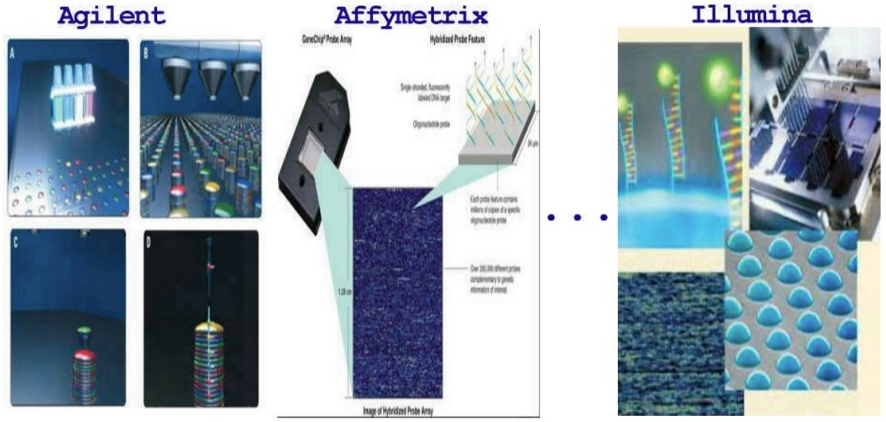
III. Horizontal Omics Data Integration

IV. Lab Session: MetaOmics

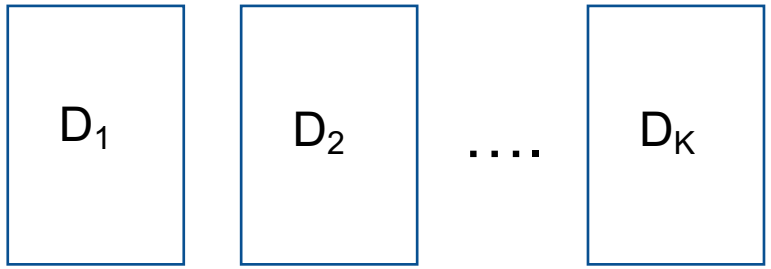
Transcriptomic meta-analysis pipeline and browser-based software suite: MetaOmics

(A) Horizontal genomic meta-analysis

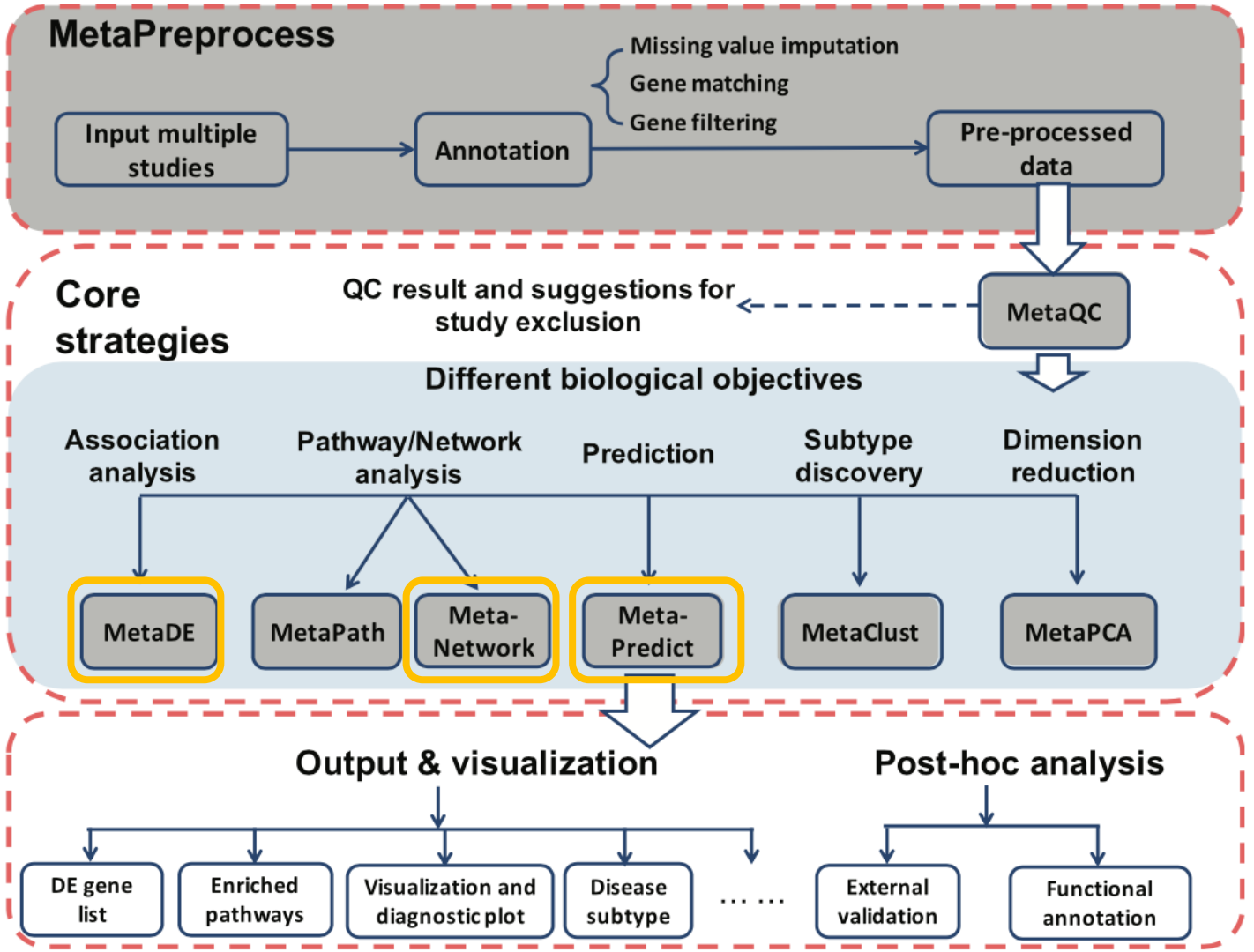
Combine microarray studies { Cross-lab; Cross-sample cohort; Cross-platform



K transcriptomic datasets



G genes× S_1 samples | G genes× S_2 samples | G genes× S_K samples



Ma et al. (2019) *Bioinformatics*.

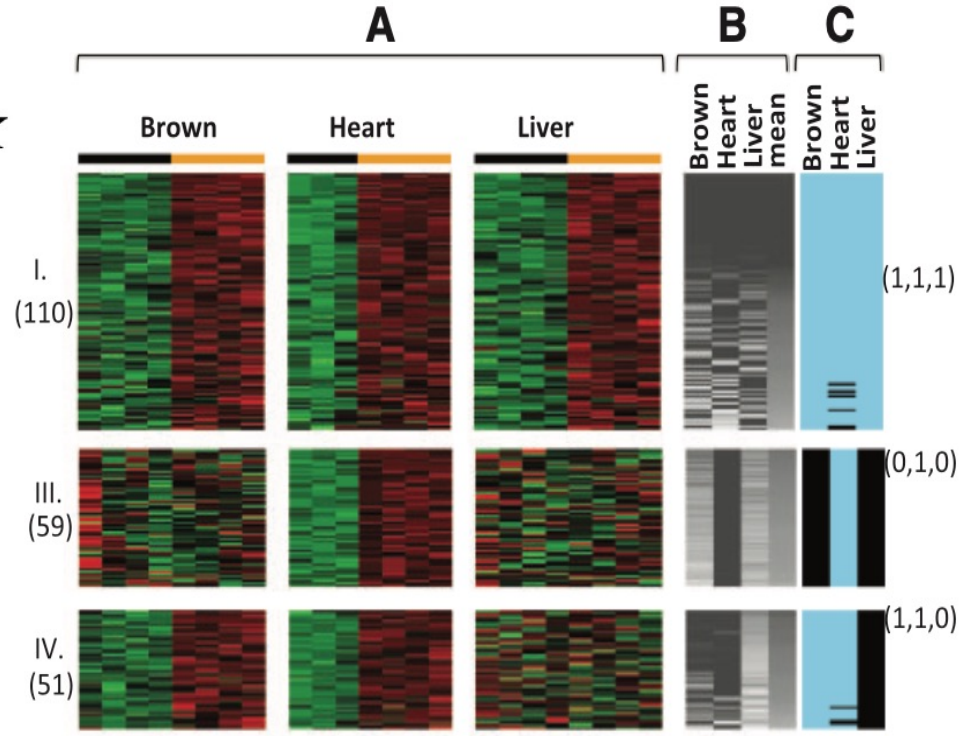
MetaOmics Pipeline: MetaDE

- **Adaptively Weighted Fisher (AWFisher) method** is one of the p-value combination methods that can additionally characterizes which study contributes to the meta-analysis result.

Hypothesis Setting $H_0: \theta_{g1} = \cdots = \theta_{gK} = 0, \quad H_B: \text{at least one } \theta_{gk} \neq 0, 1 \leq k \leq K$ (effect size)

Weighted statistic $U_g(w_g) = -\sum_{k=1}^{K} w_{gk} \log(p_{gk})$, ($p_{gk}$: p-value of gene g in study k)

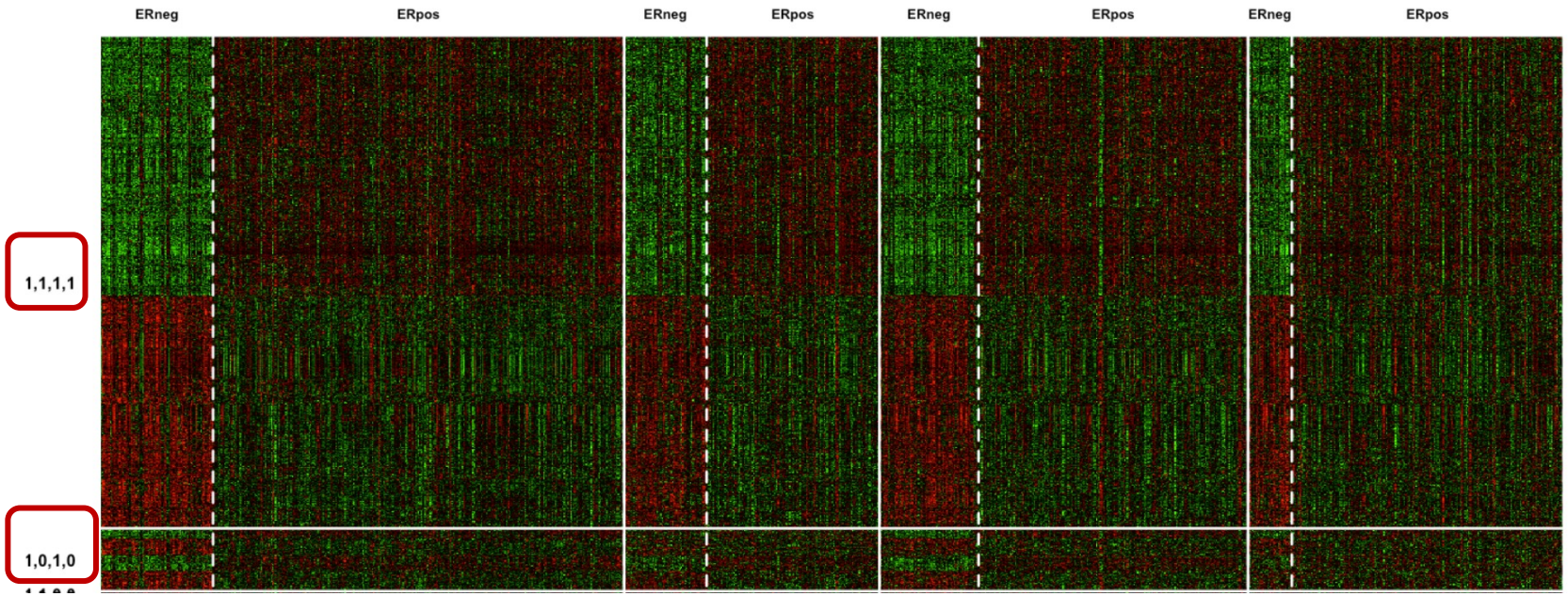
AWFisher statistic $V_g^{AW} = \min_{w_g \in W} p_U(u_g(w_g)), \quad W = \{w \mid w_i \in \{0, 1\}\}$, (Under H_0, p-value of the observed $u_g(w_g)$)



- The resulting weight reflects whether a study contributes to the statistical significance of a gene.
- The AWFisher p-values are calculated for each gene, followed by FDR control.

Huo et al. (2020) *Bioinformatics*.

MetaOmics Pipeline: MetaDE



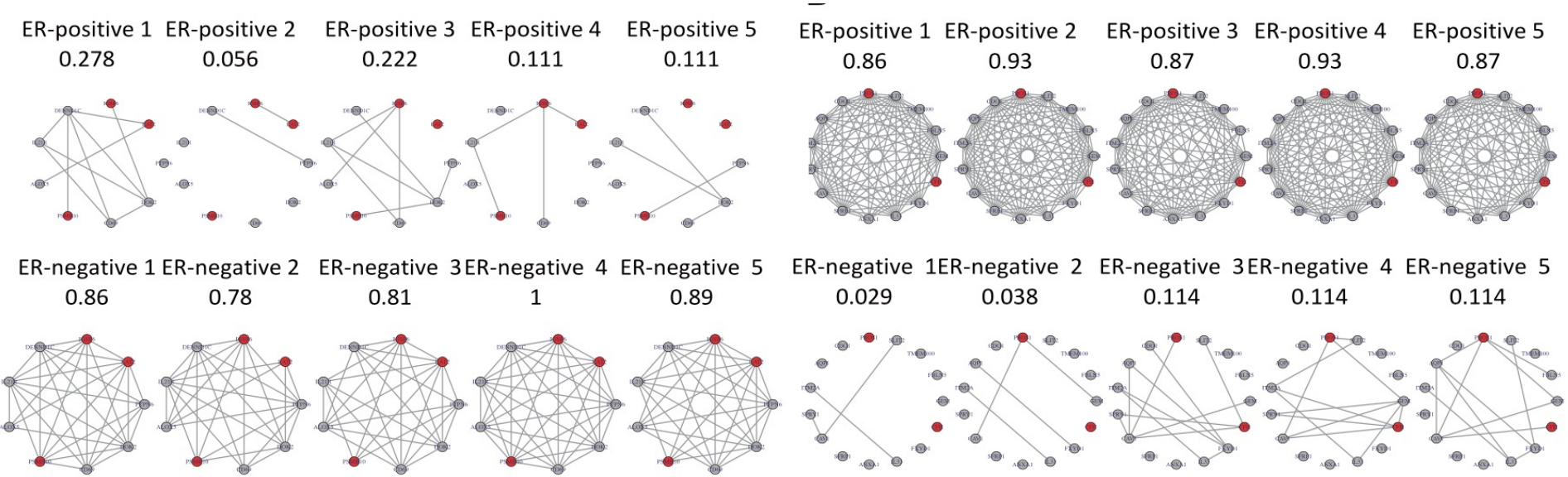
	pvalue	qvalue
KEGG Endocytosis	0.0003202	0.5491
GO:BP cell cycle	0.0007314	0.5491
KEGG DNA replication	0.001104	0.5491
GO:BP second-messenger-mediated signaling	0.001251	0.5491
Reactome DNA strand elongation	0.001471	0.5491
GO:BP regulation of mitotic cell cycle	0.002103	0.654
GO:BP cellular aromatic compound metabolic process	0.00255	0.6797
Reactome Class A/1 (Rhodopsin-like receptors)	0.004122	0.8208
GO:BP phagocytosis	0.004213	0.8208
GO:BP regulation of cell cycle	0.004765	0.8208

Showing 1 to 10 of 1,901 entries Previous 1 2 3 4 5 … 191 Next

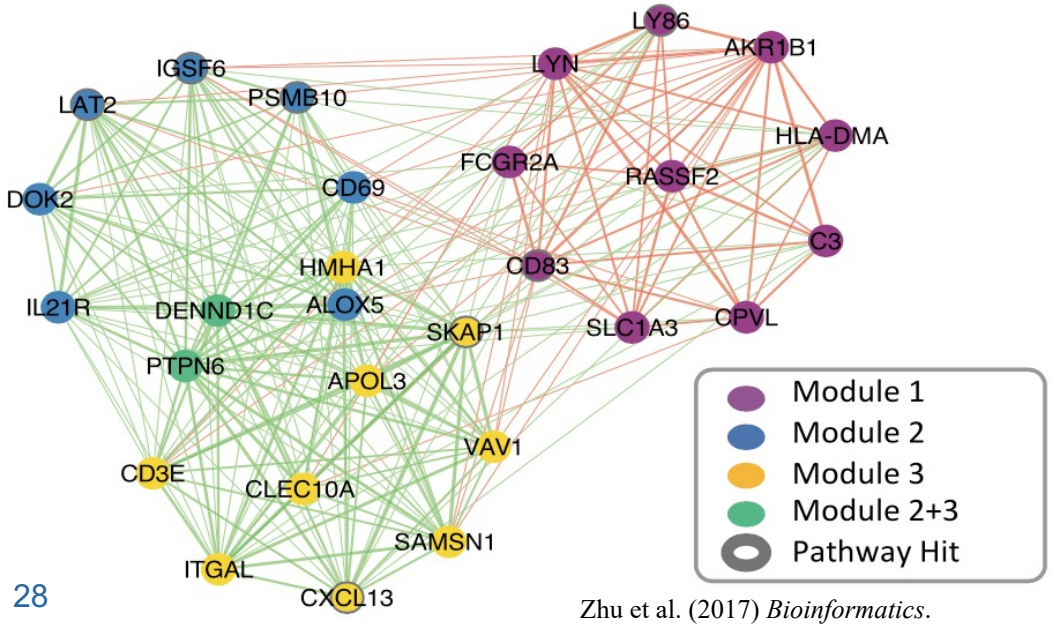
- **Input**: merged individual transcriptomic datasets (microarray or RNAseq) after preprocessing module.
- **Options of 12 major meta-analysis methods (e.g. AWFisher) with 22 variations** for detecting DE genes.
- Also implement a post hoc pathway enrichment analysis to functionally annotate detected DE genes.

Huo et al. (2020) *Bioinformatics*.

MetaOmics Pipeline: MetaNetwork



MetaDCN method: constructs the **D**ifferential **C**o-expression **N**etworks **(DCN)**



1. **Generate co-expression network**
2. **Search for basic DCN modules**
3. **Assemble the basic DCN modules into super-modules**

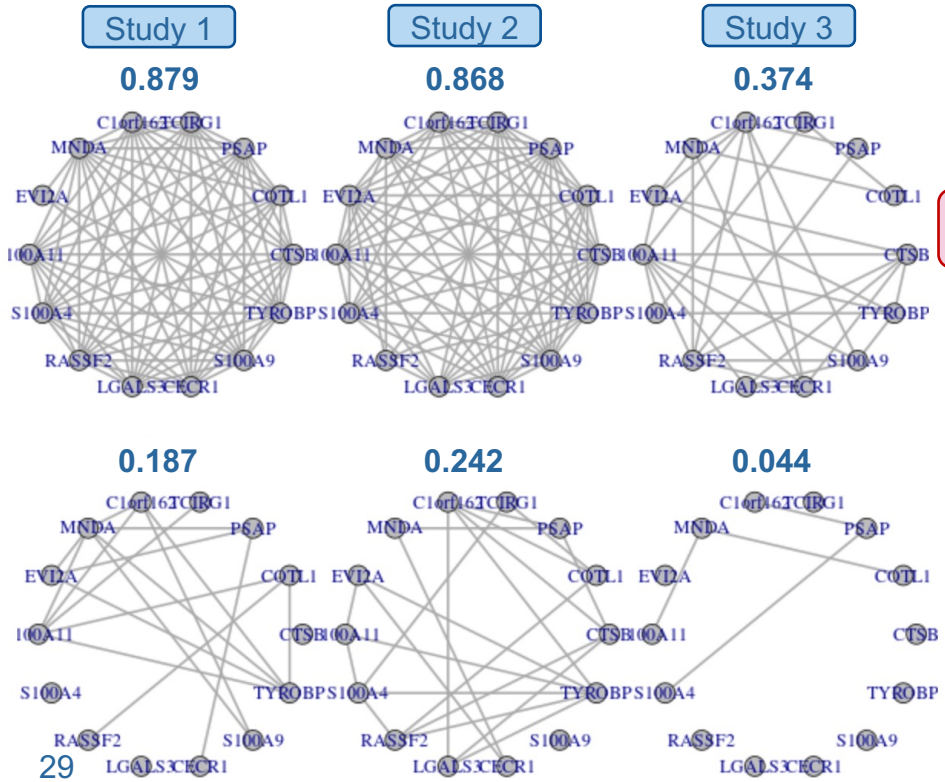
$$E_{tot} = w_1 E_{diff\_mean} + w_2 E_{size} + w_3 E_{diff\_var}$$

- E_{diff_mean}: Mean network density difference between outcome groups across all studies
- E_{size}: Size of the module
- E_{diff_var}: Consistency of the density difference between outcome groups across studies

Zhu et al. (2017) *Bioinformatics*.

MetaOmics Pipeline: MetaNetwork

- **Input**: merged individual transcriptomic datasets after preprocessing module.
- **GOAL**: infer whether the gene-gene correlations change between outcome groups.
- Implement **MetaDCN method** to construct the basic differential co-expression networks (DCN) and assemble the significant basic DCN modules into pathway-guided super-modules.
- The result of super-modules can be uploaded to a Cytoscape App “MetaDCNExplorer” for interactive visualization.



Group 1: Inversion on Chr16

Group 2: Translocation between Chr15 and Chr17

MetaDCN pathway-guided supermodules

Show 10 entries Search:

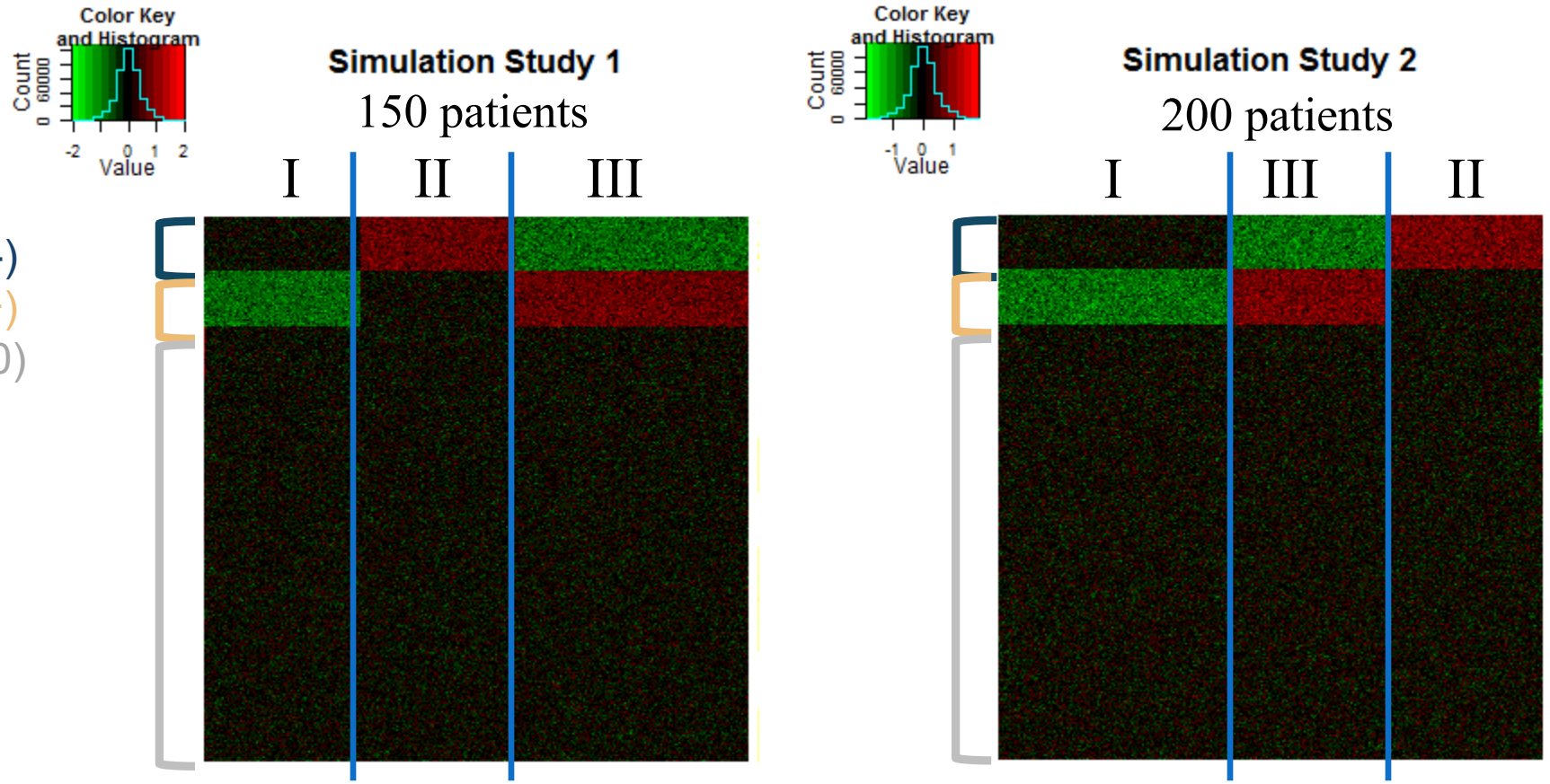
pathway_name	pathway_size	p_value	q_value	size	num_gene_in_set	module_num	module
GO_EXTRINSIC_TO_MEMBRANE	25	0.00907	0.0915	12	2	2	L3,L7
GO_ACTIN_FILAMENT	18	0.00725	0.0915	18	2	2	L7,L8
GO_MONOSACCHARIDE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	10	0.0583	0.0915	12	1	2	L3,L7
GO_SUGAR_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	11	0.0583	0.0915	12	1	2	L3,L7
GO_RUFFLE	31	0.012	0.0915	23	2	2	H6,L7
BIOCARTA_MCALPAIN_PATHWAY	25	0.0206	0.0915	18	2	2	L7,L8
REACTOME_FACILITATIVE_NA_INDEPENDENT_GLUCOSE_TRANSPORTERS	12	0.0583	0.0915	12	1	2	L3,L7
GO_CARBOHYDRATE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	16	0.0583	0.0915	12	1	2	L3,L7
GO_CORTICAL_CYTOSKELETON	20	0.00725	0.0915	18	2	2	L1,L7
GO_CARBOHYDRATE_TRANSPORT	19	0.0583	0.0915	12	1	2	L3,L7

Showing 1 to 10 of 55 entries Previous 1 2 3 4 5 6 Next

MetaOmics Pipeline: MetaClust

MetaSparseKmeans

Gene set 1: (0,+,-)
Gene set 2: (-,0,+)
Gene set 3: (0,0,0)



Three parameters to estimate:

1. Gene selection: genes that participate in the clustering
2. Sample clustering: Sample assignment to clusters
3. Pattern matching: Match cluster patterns across studies

MetaOmics Pipeline: MetaClust

- **MetaSparseKmeans** method extends the sparse K-means method towards a meta-analytic framework.

Sparse K-means for single study

$$\max_{C,\mathbf{w}} \sum_{j=1}^{p} w_j \quad BCSS_j(C) \text{ between-cluster sum of squares}$$

$$\text{subject to} \quad \|\mathbf{w}\|_2 \le 1, \|\mathbf{w}\|_1 \le \mu, w_j \ge 0, \forall j,$$

lasso regularization on gene-specific weights

Meta Sparse K means for s studies

Gene selection | Sample separation | Pattern matching

$$\max_{C^{(s)},\mathbf{w},M} \sum_{j=1}^{p} w_j \times \left[\frac{1}{S} \sum_{s=1}^{S} \frac{BCSS_j^{(s)}\left(C^{(s)}(K)\right)}{TSS_j^{(s)}} + \lambda \times f_j^{match}(M) \right]$$

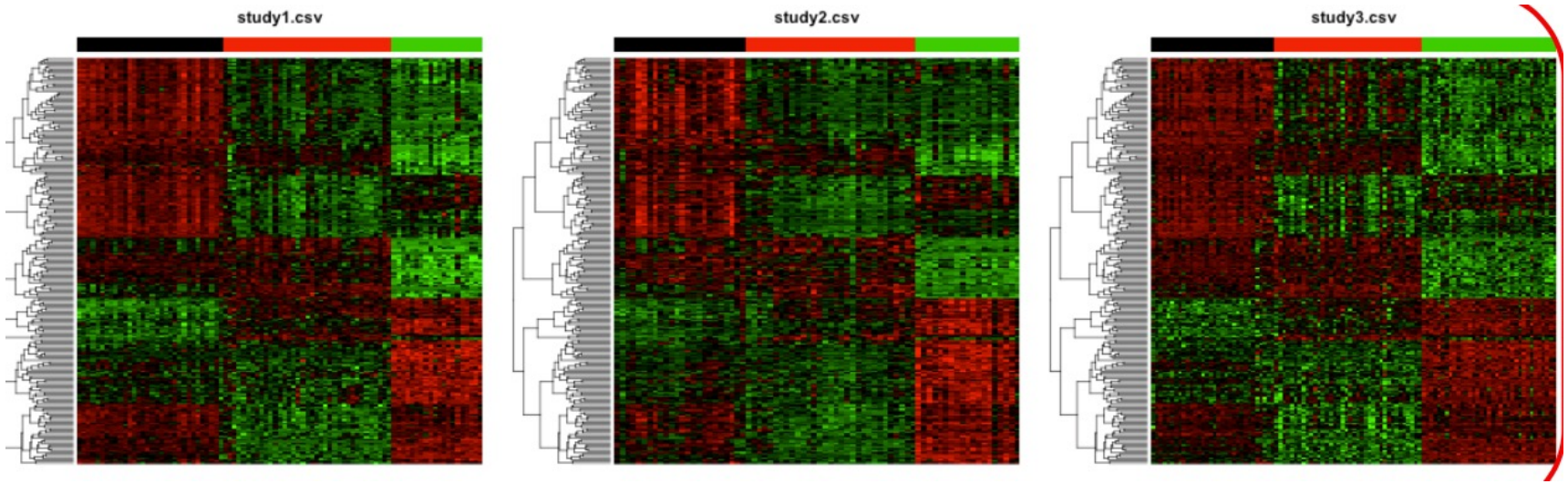
$$\text{subject to} \quad \|\mathbf{w}\|_2 \le 1, \|\mathbf{w}\|_1 \le \mu, w_j \ge 0, \forall j,$$

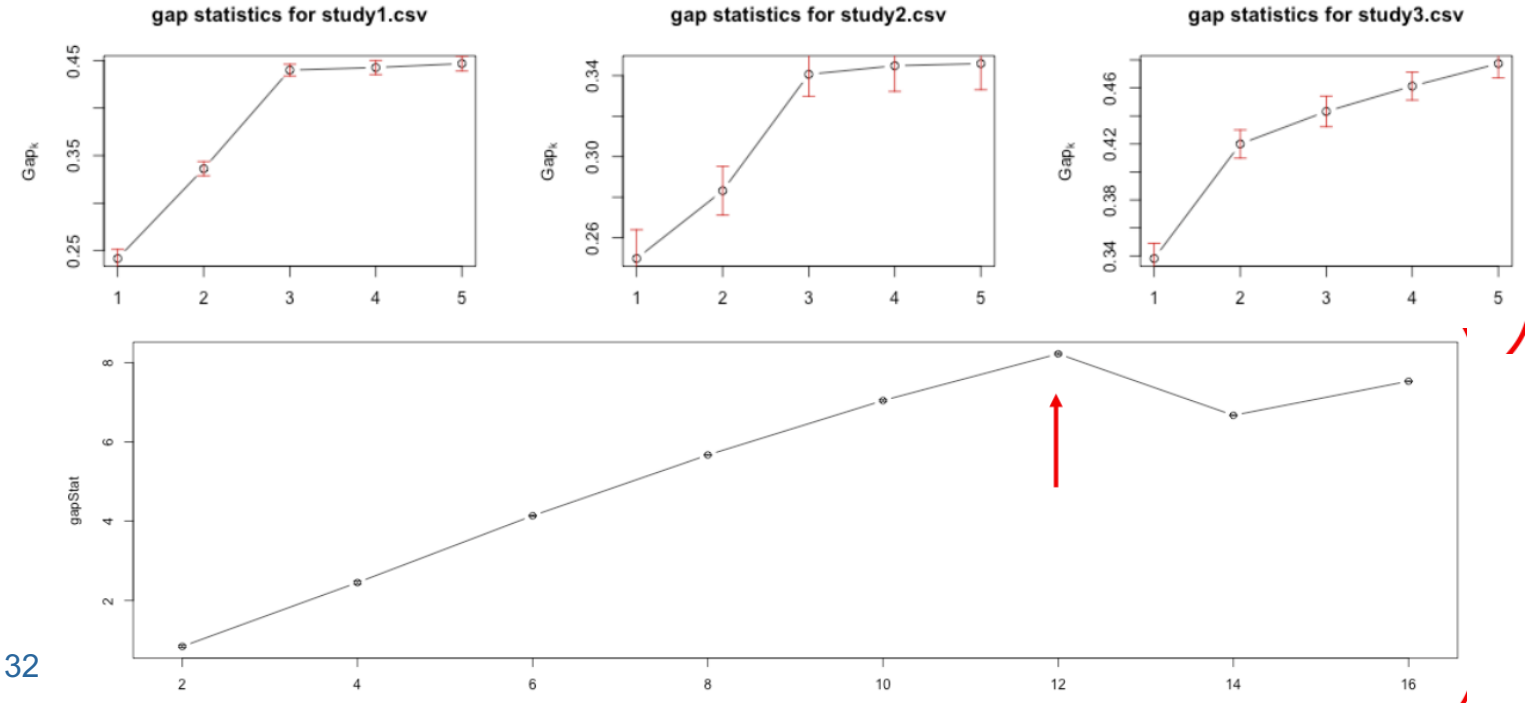
- Maximizing BCSS is equivalent to minimizing WCSS (within-cluster sum of squares).
- Many $\{w_j, 1 \le j \le p\}$ are shrunken to 0. μ controls the amount of non-zero weights.
- Only a small portion of genes have non-zero weights to contribute to the subtype modeling.

- Combine S studies. Estimate common w_j across all studies.
- BCSS may not be comparable across studies (different platform, sample size and intensity scale). Use BCSS/TSS instead.
- Add a penalty function to ensure disease subtypes of similar expression patterns are matched.
- λ balances between separation of clusters in the studies and the goodness of cluster pattern matching across studies.

Huo et al. (2016) *JASA*.

MetaOmics Pipeline: MetaClust

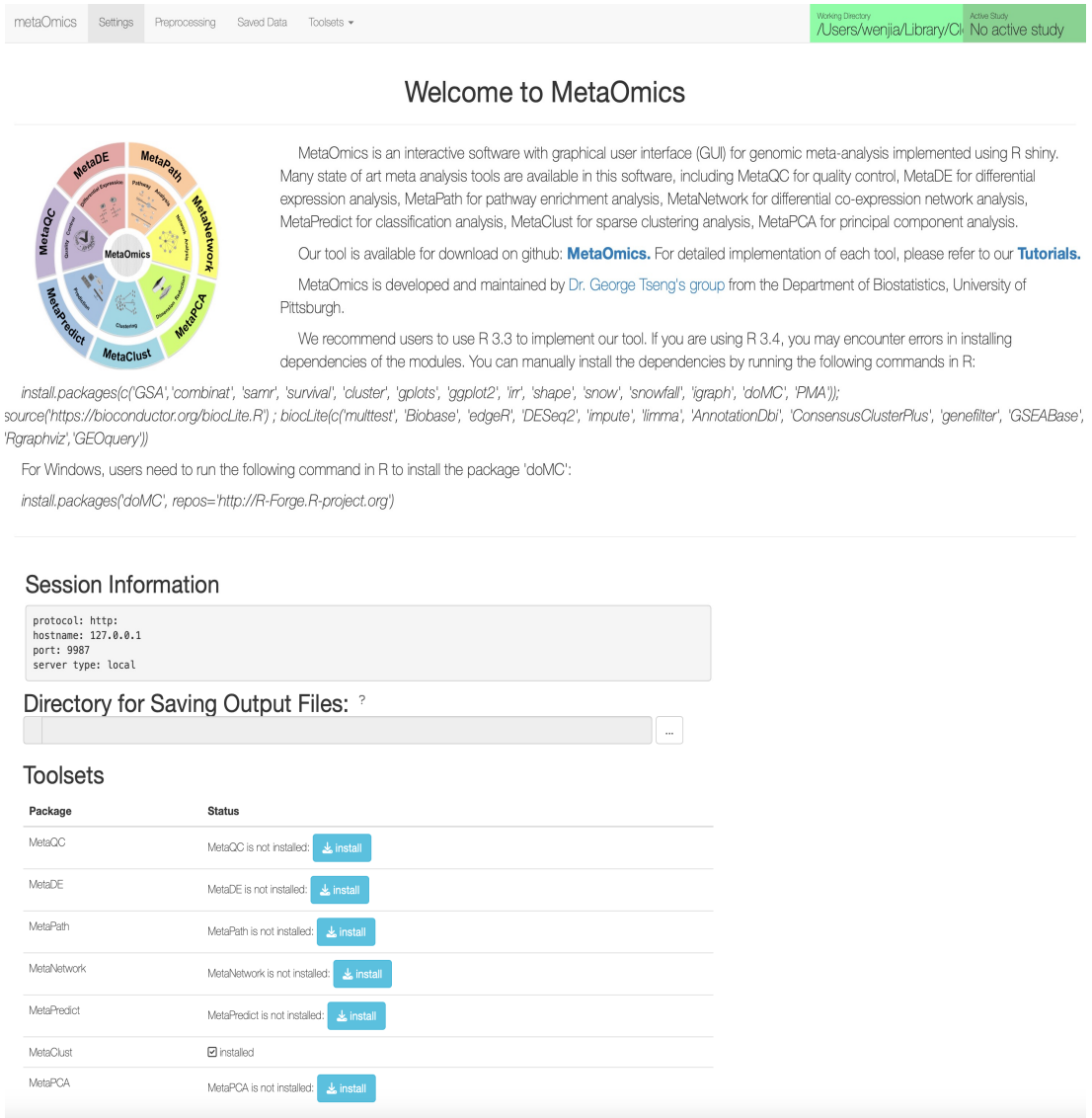




- Implement **MetaSparseKmeans** method to cluster samples and select the "intrinsic gene" set.
- Optionally tune the parameters before clustering by gap statistic: the **number of clusters (K)** and the **regularization parameter (Wbounds).**

Module	Methods	
MetaPreprocess	• Upload individual studies • Four preprocessing steps: gene annotation, missing value imputation (if needed), gene matching and preliminary gene filtering	
MetaQC (quality control)	Utilizes six quantitative QC measures: IQC, EQC, AQCg and CQCg, as well as AQCp and CQCp to make inclusion/exclusion decision	
MetaDE (differential expression analysis)	Combining effect sizes	• fixed effects model (FEM) • six variations of random effects model (REM)
	Combining p-values	Fisher, Stouffer, adaptively weighted Fisher (AW-Fisher), minimum p-value (minP), maximum p-value (maxP) and rth ordered p-value (rOP) and their one-sided correction variations
	Combining ranks	sum of ranks, product of ranks (PR) and RankProd
	Multi-class meta analysis	minimum multi-class correlation (minMCC) method
MetaPath (pathway enrichment analysis)	• Meta-Analysis for Pathway Enrichment (MAPE) (Shen and Tseng, 2010 *Bioinformatics*) • Comparative Pathway Integrator (CPI) (Zeng, 2018 *Genes*)	
MetaNetwork (differential co-expression network analysis)	MetaDCN method (Zhu, et al. 2017 *Bioinformatics*) to integrate multiple transcriptomic studies for differential co-expression networks (DCN) detection.	
MetaPredict (differential co-expression network analysis)	MetaKTSP method (Kim, et al. 2016 *Bioinformatics*)	
MetaClust (clustering analysis)	MetaSparseKmeans algorithm (Huo, et al. 2016 *JASA*)	
MetaPCA (dimension reduction)	MetaPCA method (Kim, 2018 *Bioinformatics*): sum of variance (SV) decomposition and sum of squared cosines (SSC) decomposition	

Lab Session: MetaOmics



General biological insights:

1. Gligorijevic and Przulj (2015) Methods for biological data integration: perspectives and challenges. J R Soc Interface.
2. Hasin et al. (2017) Multi-omics approaches to disease. Genome Biology.
3. Huang et al. (2017) More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Frontiers in Genetics.
4. Dihazi et al. (2018) Integrative omics - from data to biology. Expert Review of Proteomics.
5. Misra et al. (2019) Integrated omics: tools, advances and future approaches. Journal of Molecular Endocrinology.
6. Noor et al. (2019) Biological insights through omics data integration. Current Opinion in Systems Biology.
7. Wang et al. (2019) Toward multiomics-based next-generation diagnostics for precision medicine. Personalized Medicine

Mathematical/statistical challenges:

1. Richardson, Tseng and Sun. (2016) Statistical Methods in Integrative Genomics. Annual Review of Statistics and Its Application.
2. Bersanelli et al. (2016) Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics
3. Wu et al. (2019) A Selective Review of Multi-Level Omics Data Integration Using Variable Selection. High Throughput.
4. Machine learning perspective:
5. Li et al. (2019) A review on machine learning principles for multi-view biological data integration. Briefings in Bioinformatics.
6. Zitnik et al. (2019) Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion.
7. Mirza et al. (2019) Machine Learning and Integrative Analysis of Biomedical Big Data. Genes.
8. Rappoport and Shamir (2018) Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Research.

Specialized disease areas:

1. Zhang et al. (2010) Integrating multiple 'omics' analysis for **microbial biology**: application and methodologies. Microbiology
2. Sathyanarayanan et al. (2019) A comparative study of multi-omics integration tools for **cancer** driver gene identification and tumour subtyping. Briefings in Bioinformatics.
3. Chakraborty et al. (2018) Onco-Multi-OMICS Approach: A New Frontier in **Cancer** Research. BioMed Research International.

   Canzler et al. (2020) Prospects and challenges of multi-omics data integration in **toxicology**. Archives of Toxicology.
4. Beltran et al. (2017) Proteomics and integrative omic approaches for understanding host–pathogen interactions and **infectious diseases**. Molecular Systems Biology.
5. Sun and Hu (2018) Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of **Complex Human Diseases**.
6. Donovan et al. (2019) The current state of omics technologies in the clinical management of **asthma and allergic diseases**. Ann Allergy Asthma Immunol.
7. Higdon et al. (2015) The Promise of Multi-Omics and Clinical Data Integration to Identify and Target Personalized Healthcare Approaches in **Autism** Spectrum Disorders. OMICS: A Journal of Integrative Biology.

Specific community-centered:

1. Pinu et al. (2019) Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. Metabolites.
2. Zhang and Kuster (2019) Proteomics Is Not an Island: Multi-omics Integration Is the Key to Understanding Biological Systems. Molecular & Cellular Proteomics.

Single cell multi-omics:

1. Chappell et al. (2018) Single-Cell (Multi)omics Technologies. Annual Review of Genomics and Human Genetics.
2. Hu et al. (2018) Single Cell Multi-Omics Technology: Methodology and Application. Front. Cell Dev. Biol.

Thank you!