

# Computational methods for data integration

**Karan Uppal, PhD**

Assistant Professor of Medicine  
Director of Computational Systems Medicine &  
Metabolomics Lab  
Emory University

# Learning Objectives

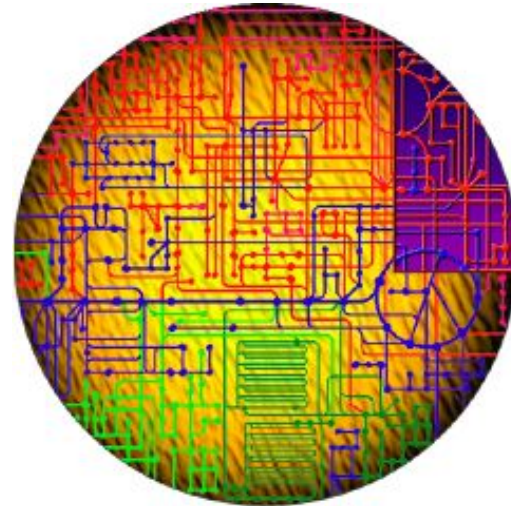
- Understanding of different integrative network analysis approaches
- Familiarity with tools for data integration and network visualization

# Introduction: A Systems Biology Framework

- The goal of **Systems Biology**:
  - Systems-level understanding of biological systems
  - Analyze not only individual components, but their interactions as well and emergent behavior

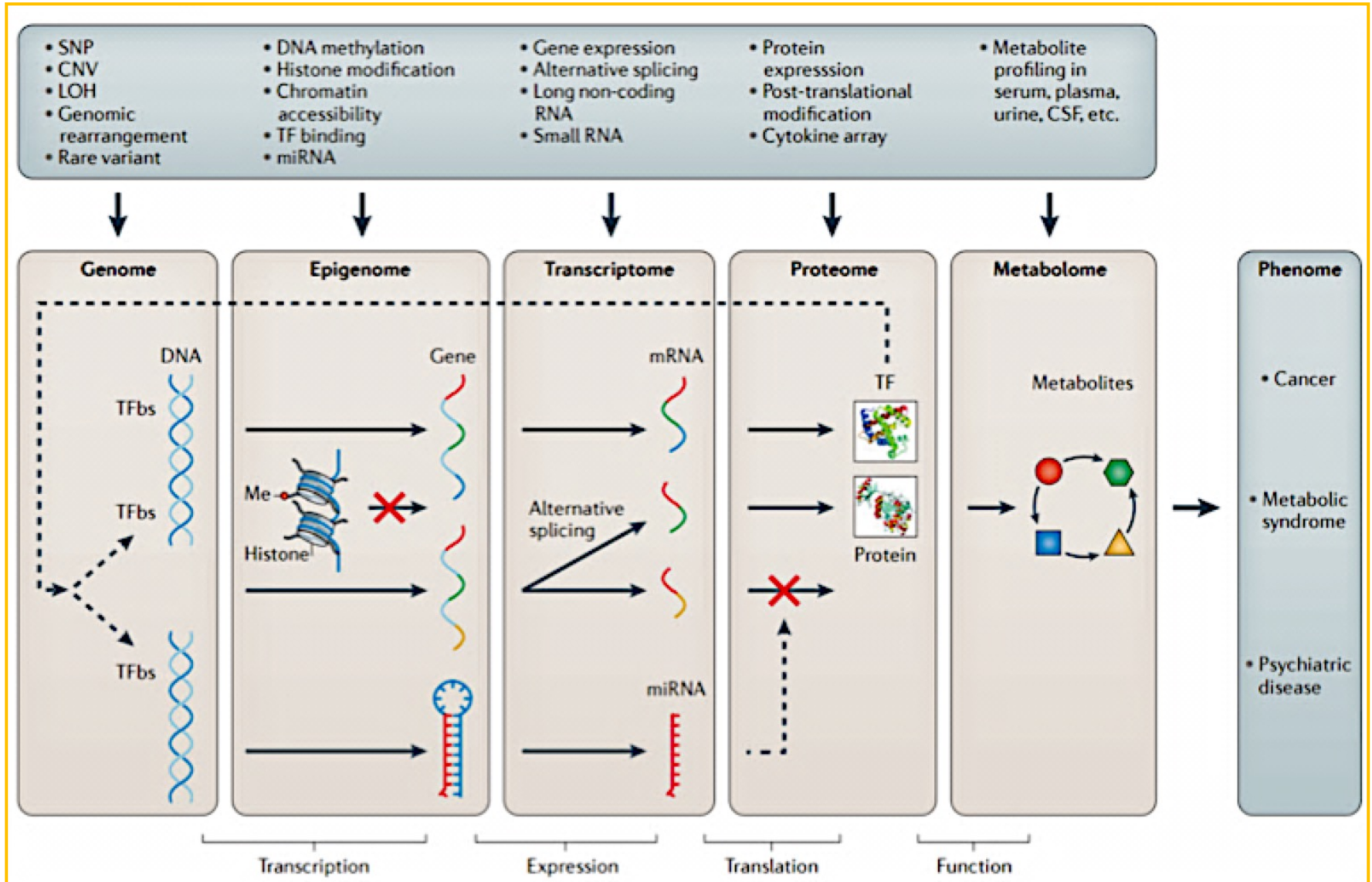


Exposures  
Internal measurements  
Disease states



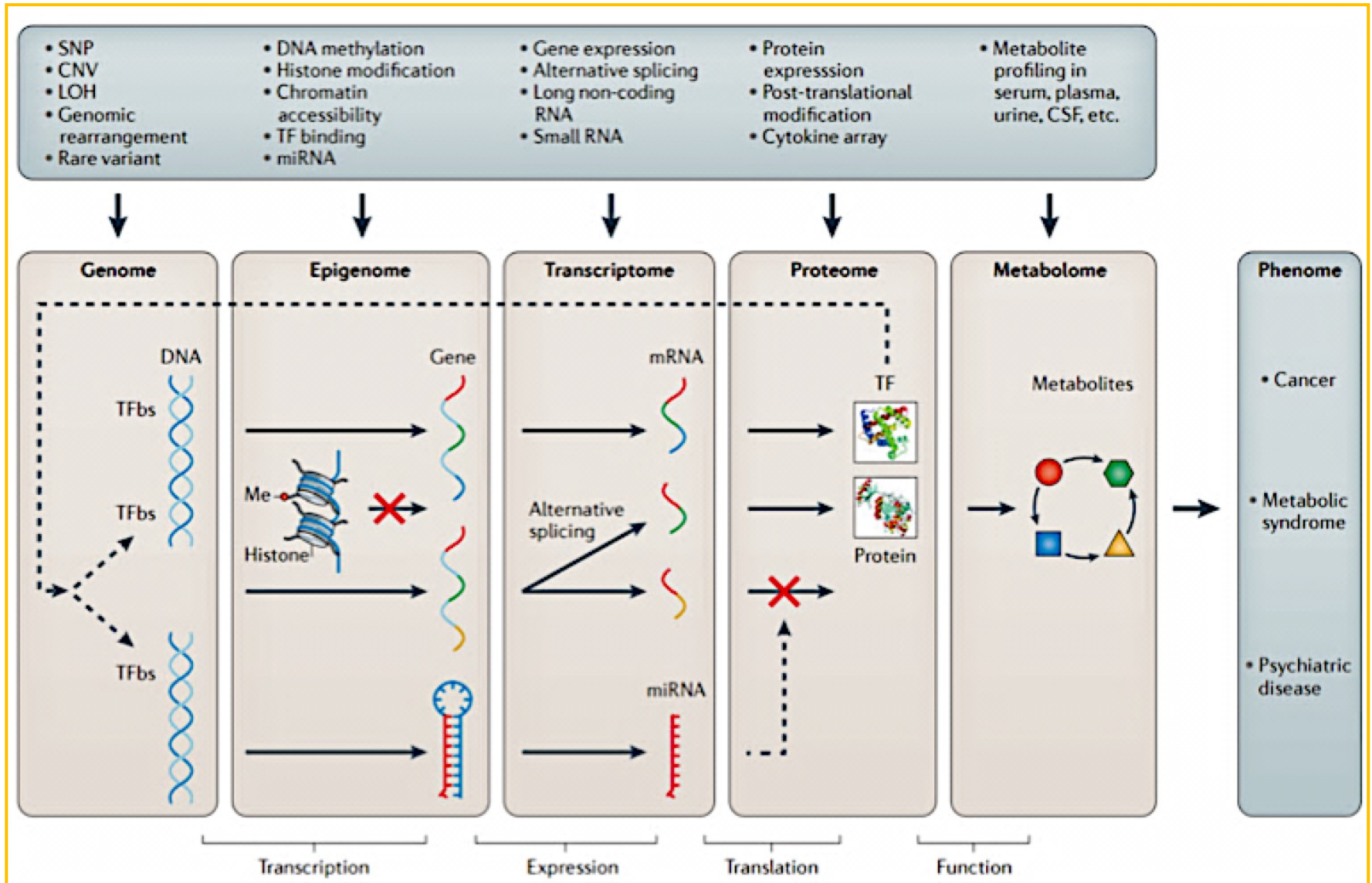
**Systems Biology**  
*“Integrative approach in which scientists study pathways and networks will touch all areas of biology, including drug discovery”*

# Integrative omics: dissecting the biological system via -omics





# “Information Overload”: >10,000 variables per –omics experiment

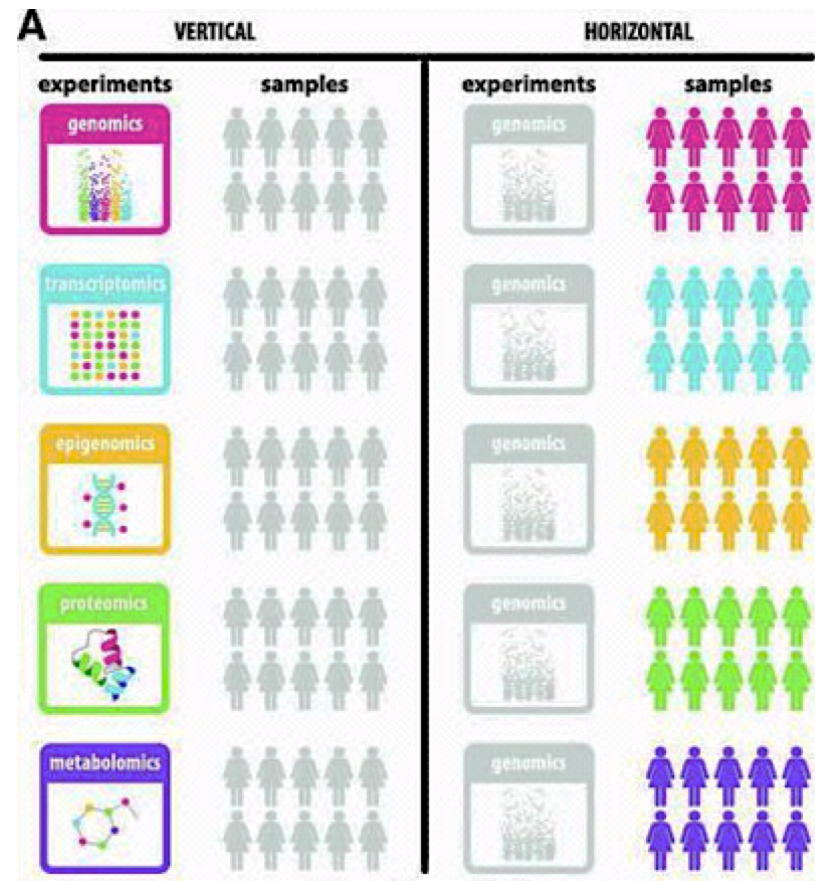


# Why data integration?

- Systems level analysis provides:
  - more detailed overview of underlying mechanisms;
  - exploration of interactions between different biomedical entities (genes, proteins, metabolites, etc.)
- Combining multiple types of data collected on the same subjects compensates for noise or unreliable information in a single data type
- More confidence in results if multiple sources of evidence pointing to the same gene or pathway

# Data integration study designs

- Paired or vertical integrative analysis
  - Integrative analysis of **multiple omics datasets** from the **same N subjects**
  - Discover networks of associations or correlated variables (e.g. genes, proteins, metabolites, microbiome, epigenetic alterations, clinical variables)
    - Univariate or multivariate regression
    - Example: explaining protein abundance with respect to gene expression
- Horizontal integrative analysis
  - Meta-analysis of **multiple studies/cohorts** looking at the same type of data
  - Cross-laboratory or cross-platform comparisons



Eidem 2018, BMC Med Genomics

Ref: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6245874/>

# Main approaches for data integration

- Pathway-based integration
  - Pathway information from KEGG or other databases
  - Datasets are analyzed individually (differentially expressed genes, metabolites, proteins) and integration is performed at the pathway level
  - Examples: MetaboAnalyst, iPEAP, MetScape, MetaCore
- Data-driven integration using meta-dimensional analysis
  - Integration is performed globally such that data from multiple omics layers are combined simultaneously
  - Interpretation using pathway analysis tools
  - Examples: 3Omics, mixOmics, xMWAS
- Using literature-derived associations for integration
  - Using co-occurrence criteria for establishing relationship
  - Examples: Comparative Toxicogenomics Database, CoPub, ArrowSmith, SEACOIN

# Main approaches for data integration

- **Pathway-based integration**
  - Pathway information from KEGG or other databases
  - Datasets are analyzed individually (differentially expressed genes, metabolites, proteins) and integration is performed at the pathway level
  - Examples: MetaboAnalyst, iPEAP, MetScape, MetaCore
- Data-driven integration using meta-dimensional analysis
  - Integration is performed globally such that data from multiple omics layers are combined simultaneously
  - Interpretation using pathway analysis tools
  - Examples: 3Omics, mixOmics, xMWAS
- Using literature-derived associations for integration
  - Using co-occurrence criteria for establishing relationship
  - Examples: HiPub, CoPub, ArrowSmith

# Pathway-based data integration - I

Metabolomics data  
(n subjects X p metabolites)

	M1	M2	-	Mp
Subject1	199	19	-	100
Subject2	10	40		90
-	-	-		-
SubjectN	50	30	-	20



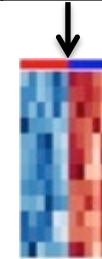
**Differentially  
expressed  
metabolites**

**Over-represented pathways**

Rank	Pathway ID (hsa:)	Pathway Title
1	04974	Protein digestion and absorption
2	02010	ABC transporters
3	00250	Alanine, aspartate and glutamate metabolism
4	00330	Arginine and proline metabolism
5	00480	Glutathione metabolism
6	00260	Glycine, serine and threonine metabolism
7	00910	Nitrogen metabolism
8	00460	Cyanoamino acid metabolism
9	00270	Cysteine and methionine metabolism
10	00770	Pantothenate and CoA biosynthesis

Transcriptomics data  
(n subjects X q genes)

	G1	G2	-	Gq
Subject1	19	19	-	100
Subject2	10	40	-	90
-	-	-	-	-
SubjectN	10	40	-	50



**Differentially  
expressed  
genes**

**Over-represented pathways**

Rank	Pathway ID (hsa:)	Pathway Title
1	00260	Glycine, serine and threonine metabolism
2	00340	Histidine metabolism
3	00480	Glutathione metabolism
4	00450	Selenoamino acid metabolism
5	00360	Phenylalanine metabolism
6	00071	Fatty acid metabolism
7	00330	Arginine and proline metabolism
8	00561	Glycerolipid metabolism
9	00380	Tryptophan metabolism
10	00250	Alanine, aspartate and glutamate metabolism

# Pathway-based data integration - I

Metabolomics data  
(n subjects X p metabolites)

	M1	M2	-	Mp
Subject1	199	19	-	100
Subject2	10	40		90
-	-	-		-
SubjectN	50	30	-	20



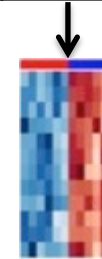
Differentially  
expressed  
metabolites

Over-represented pathways

Rank	Pathway ID (hsa:)	Pathway Title
1	04974	Protein digestion and absorption
2	02010	ABC transporters
3	00250	Alanine, aspartate and glutamate metabolism
4	00330	Arginine and proline metabolism
5	00480	Glutathione metabolism
6	00260	Glycine, serine and threonine metabolism
7	00910	Nitrogen metabolism
8	00460	Cyanoamino acid metabolism
9	00270	Cysteine and methionine metabolism
10	00770	Pantothenate and CoA biosynthesis

Transcriptomics data  
(n subjects X q genes)

	G1	G2	-	Gq
Subject1	19	19	-	100
Subject2	10	40	-	90
-	-	-	-	-
SubjectN	10	40	-	50



Differentially  
expressed  
genes

Over-represented pathways

Rank	Pathway ID (hsa:)	Pathway Title
1	00260	Glycine, serine and threonine metabolism
2	00340	Histidine metabolism
3	00480	Glutathione metabolism
4	00450	Selenoamino acid metabolism
5	00360	Phenylalanine metabolism
6	00071	Fatty acid metabolism
7	00330	Arginine and proline metabolism
8	00561	Glycerolipid metabolism
9	00380	Tryptophan metabolism
10	00250	Alanine, aspartate and glutamate metabolism

Pathway rank aggregation

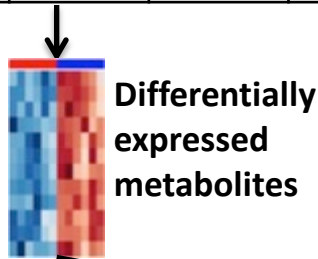
Rank	Pathway ID (hsa:)	Pathway Title
1	00260	Glycine, serine and threonine metabolism
2	00330	Arginine and proline metabolism
3	00480	Glutathione metabolism



# Pathway-based data integration - II

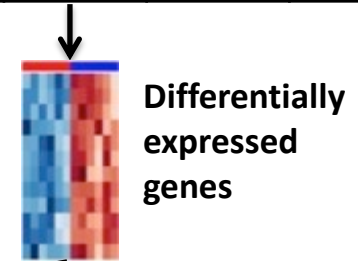
Metabolomics data  
(n subjects X p metabolites)

	M1	M2	-	Mp
Subject1	199	19	-	100
Subject2	10	40		90
-	-	-		-
SubjectN	50	30	-	20



Transcriptomics data  
(n subjects X q genes)

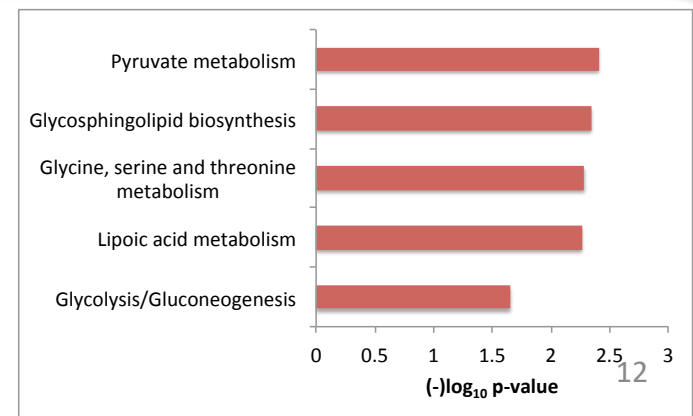
	G1	G2	-	Gq
Subject1	19	19	-	100
Subject2	10	40	-	90
-	-	-	-	-
SubjectN	10	40	-	50



**MetaboAnalyst4.0 – Joint Pathway Analysis module**



Over-representation analysis in KEGG using gene and metabolite IDs

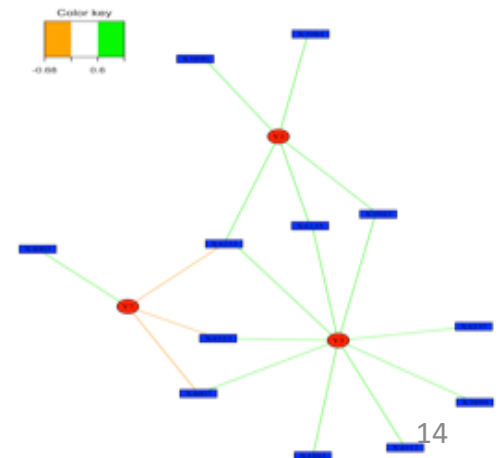


# Main approaches for data integration

- Pathway-based integration
  - Pathway information from KEGG or other databases
  - Datasets are analyzed individually (differentially expressed genes, metabolites, proteins) and integration is performed at the pathway level
  - Examples: MetaboAnalyst, iPEAP, MetScape, MetaCore
- **Data-driven integration using meta-dimensional analysis**
  - **Integration is performed globally such that data from multiple omics layers are combined simultaneously**
  - **Interpretation using pathway analysis tools**
  - **Examples: 3Omics, mixOmics, xMWAS**
- Using literature-derived associations for integration
  - Using co-occurrence criteria for establishing relationship
  - Examples: Comparative Toxicogenomics Database, CoPub, ArrowSmith, SEACOIN

# Relevance networks

- What is a network (or graph)?
  - A set of nodes (vertices) and edges (links)
  - Edges describe a relationship (e.g. correlation) between the nodes
- What is a relevance network?
  - Networks of highly-correlated biomedical/clinical entities (Butte 2000; PNAS)
  - Metabolomics x Proteomics, Transcriptomics x Proteomics, Metabolomics x Microbiome, Metabolomics x Clinical variables/phenotypes, etc.
  - Generate a bipartite graph network using an association threshold (e.g. 0.5) to visualize positive or negative associations



Circles: genes  
Rectangles: metabolites

# Methods for generating relevance networks

- Univariate
  - 3Omics (Kuo 2013; a web-based tool for analysis, integration and visualization of human transcriptome, proteome and metabolome data)
  - MetabNet (Uppal 2015; R package for performing pairwise correlation analysis and generating relevance networks)
- Multivariate
  - Multivariate regression techniques such as partial least squares (PLS), sparse partial least squares regression (sPLS), multilevel sparse partial least squares (msPLS) regression, etc.
  - mixOmics (Cao et al. 2009, Liquet et al. 2012; R package for integration and variable selection using multivariate regression)
  - xMWAS (Uppal 2018): a data-driven integration and differential network analysis
    - Availability: <https://kuppal.shinyapps.io/xmwas> (Online) and <https://github.com/kuppal2/xMWAS/> (R)

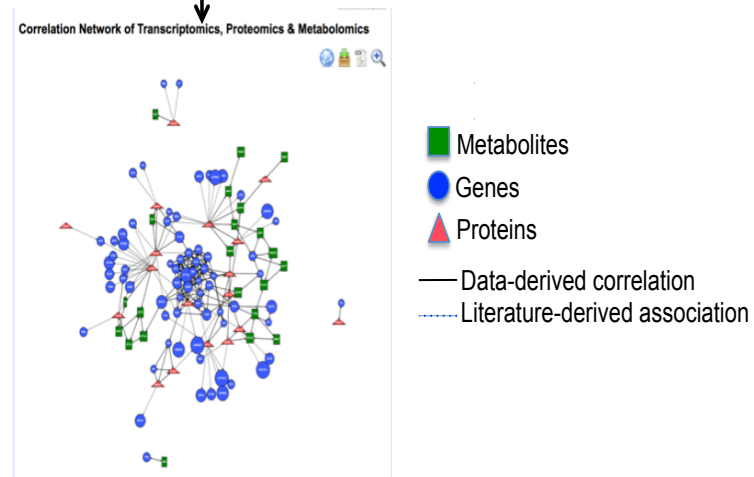
# 3Omics: a web-based tool for analysis, integration and visualization of human transcriptome, proteome and metabolome data (Kuo 2013, BMC Systems Biology)

- Web-based tool
- Correlation analysis and network visualization
- Additional features:
  - Metabolic pathway analysis
  - Gene ontology enrichment analysis
  - Hierarchical clustering analysis

URL: <http://3omics.cmdm.tw/>

Gene	Transcriptomics			Metabolite	Metabolomics			Protein	Proteomics		
	Sample 1	Sample 2	Sample 3		Sample 1	Sample 2	Sample 3		Sample 1	Sample 2	Sample 3
akap9	-0.24	-0.6	-0.47	4277439	-0.3109937	-0.2792995	-0.2548517	P14060	2.06	1.2	1.61
macf1	-0.3	-0.3	0.48	441	-0.2967872	-0.2895908	-0.2674823	P26439	1.8	3.57	2.04
RNPEP	0.24	0.85	0.15	69362	-0.3183692	-0.2828533	-0.272917	P29372	-0.64	-0.71	-0.21
SDHA	0.1	0.37	0.18	10258	-0.0614116	-0.1180467	-0.1231662	Q96J02	-0.52	-1.34	-0.15

Pairwise Pearson correlation analysis



# xMWAS: a data-driven integration and differential network analysis (Uppal 2018, Bioinformatics)

URL: <https://kuppal.shinyapps.io/xmwas/>

R package: <https://github.com/kuppal2/xMWAS>

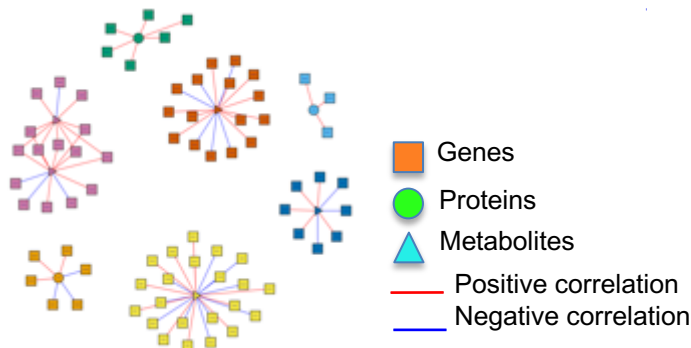
Transcriptomics				Metabolomics				Proteomics			
Gene	Sample 1	Sample 2	Sample 3	Metabolite	Sample 1	Sample 2	Sample 3	Protein	Sample 1	Sample 2	Sample 3
akap9	-0.24	-0.6	-0.47	4277439	-0.3109937	-0.2792995	-0.2548517	P14060	2.06	1.2	1.61
macf1	-0.3	-0.3	0.48	441	-0.2967872	-0.2895908	-0.2674823	P26439	1.8	3.57	2.04
RNPEP	0.24	0.85	0.15	69362	-0.3183692	-0.2828533	-0.272917	P29372	-0.64	-0.71	-0.21
SDHA	0.1	0.37	0.18	10258	-0.0614116	-0.1180467	-0.1231662	Q96J02	-0.52	-1.34	-0.15

Pairwise (sparse) Partial Least Squares regression for data integration (Cao 2009)

Approximation of Pearson correlation using PLS components

Filtering based on  $|r| > \text{threshold}$  and  $p\text{-value} < \alpha$  criteria

Community (clusters) detection and centrality (importance) analysis



# Sparse Partial Least Squares (PLS) regression method (Cao 2009, Liquet 2012)

- sPLS is a variable selection and dimensionality reduction method that allows integration of heterogeneous omics data from same set of samples
- Robust approximation of Pearson correlation using regression and latent (principal) variates
- Multilevel sparse PLS – accounts for repeated measures
- Eg: transcriptome (matrix X) and metabolome (matrix Y) data  
where,  
matrix X is an  $n \times p$  matrix that includes  $n$  samples and  $p$  metabolites  
matrix Y is an  $n \times q$  matrix that includes  $n$  samples and  $q$  genes

Objective function

$\max \text{cov}(X_u, Y_v)$


where

$u_1, u_2 \dots u_H$  and  $v_1, v_2 \dots v_H$  are the loading vectors

$H$  is the number of PLS-DA dimensions

A Lasso based optimization is used to select most relevant variables

Association matrix using the PLS components



	X1	X2	-	Xn
Y1	0.4	0.9	-	0.3
Y2	0.7	0.1	-	0.5
Y3	0.1	0.6		0.8



# Community detection

- Community: set of densely connected nodes that have more connections with the nodes in the same community as compared to nodes in other communities
- Multilevel community detection: a multi-step procedure
  - 1) each node is assigned to a different community
  - 2) each node is moved to a community with which it achieves the highest positive contribution to modularity
  - 3) Step 2 is repeated for all nodes until no improvement can be achieved
  - 4) Each community after step 3 is now considered a node and step 2 is repeated until there is a single node left or the modularity can no longer be improved

# Centrality analysis

- Centrality: measure of importance of a node in the network
- Common centrality measures
  - Eigenvector: based on the number and quality of connections
  - Betweenness: based on the extent to which a node lies on the path between other nodes
  - Degree.count: based on the number of connections
  - Degree.weight: based on the magnitude of edges (association scores)
  - Closeness: based on the closeness of a node to all other nodes
- Differential centrality analysis: delta centrality between two conditions (e.g.  $|\text{centrality}_{\text{exposed}} - \text{centrality}_{\text{control}}|$ )

# xMWAS: ShinyApp interface

54.210.50.75:8787/s/29b1ccfc78a31241bc31d/?view=shiny  
http://54.210.50.75:8787/s/29b1ccfc78a31241bc31d/p/7160/ Open in Browser Republish

## *xMWAS - a data-driven integration and network analysis tool (v0.553)*

Introduction

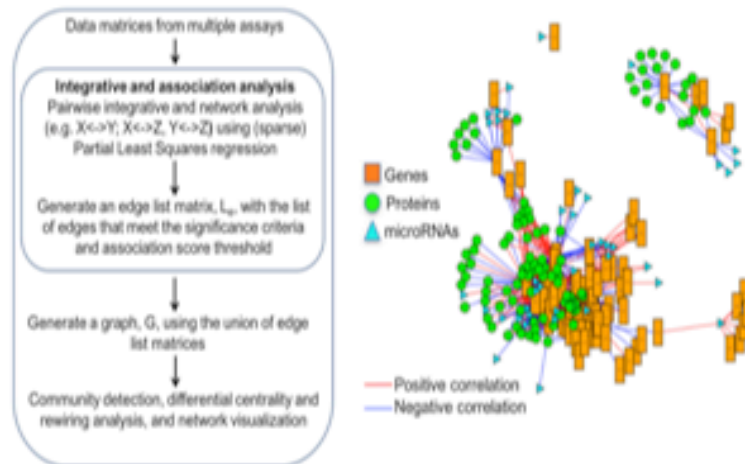
Analysis

Help and Support

**xMWAS provides an automated workflow for data integration, network visualization, clustering, and differential network analysis of up to four datasets from biochemical and phenotypic assays, and omics platforms.**

For installing xMWAS locally in R run:

```
library(devtools);install_github("kuppal2/xMWAS")
```



Citation: Uppal K, Ma C, Go YM, Jones DP: xMWAS: a data-driven integration and differential network analysis tool. *Bioinformatics*. 2018 Feb 15. PMID: 29069296  
Maintained by Chunyu Ma ( [chunyu.ma@emory.edu](mailto:chunyu.ma@emory.edu) ) and Karan Uppal ( [kuppal2@emory.edu](mailto:kuppal2@emory.edu) ) at Clinical Biomarkers Laboratory , Emory University, Atlanta, GA, USA

# Step 1. Upload data files

*xMWAS - a data-driven integration and network analysis tool (v0.553)*

Introduction Analysis Help and Support

### Input Files

Choose Files (see help and support)

### Parameter Settings

1. Data preparation and filtering
2. Integration and association analysis
3. Centrality analysis
4. Graphical options

**Input file for dataset A ('.csv' or '.txt', 100MB limit)**  
Browse... No file selected

**Input file for dataset B ('.csv' or '.txt', 100MB limit)**  
Browse... No file selected

Add more datasets: + -

**Name for dataset A:**  
datasetA

**Name for dataset B:**  
datasetB

---

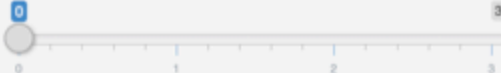
**Choose a class labels file ('.csv' or '.txt'):**  
Browse... No file selected

More Options

Start processing Download results

### Output

Slide to go to next figure:



<https://kuppal.shinyapps.io/xmwas/>

(See: Help & Support)

# Step 2. Data preprocessing and filtering

The screenshot displays the xMWAS web application interface. At the top, the browser address bar shows the URL: `http://54.210.50.75:8787/s/29b1ccfc78a31241bc31d/p/7160/`. The page title is *xMWAS - a data-driven integration and network analysis tool (v0.553)*. The navigation menu includes **Introduction**, **Analysis** (selected), and **Help and Support**.

The **Analysis** tab is active, showing a sidebar with **Input Files** and **Parameter Settings**. Under **Parameter Settings**, the first option, **1. Data preparation and filtering**, is highlighted in blue. Other options include **2. Integration and association analysis**, **3. Centrality analysis**, and **4. Graphical options**.

The main content area contains several input fields for parameter settings:

- Relative Standard Deviation (RSD) Threshold (rows):** A text input field containing the value `-1`.
- Maximum #datasetA variables to select based on RSD (change according to your dataset):** A text input field containing the value `10000`.
- Maximum #datasetB variables to select based on RSD (change according to your dataset):** A text input field containing the value `10000`.
- Maximum #datasetC variables to select based on RSD (change according to your dataset):** A text input field containing the value `10000`.
- Maximum #datasetD variables to select based on RSD (change according to your dataset):** A text input field containing the value `10000`.
- Minimum non-missing sample ratio (rows):** A text input field containing the value `0`.
- How are the missing values represented in the data?:** A dropdown menu with the value `0` selected.

At the bottom of the parameter settings, there are two buttons: **Start processing** and **Download results**.

Below the buttons is the **Output** section, which includes a slider control labeled **Slide to go to next figure:**. The slider is currently positioned at `0` on a scale from `0` to `8`.

# Step 3. Set parameters for integration and association analysis

*xMWAS - a data-driven integration and network analysis tool (v0.553)*

Introduction Analysis Help and Support

Input Files

Choose Files (see help and support)

Parameter Settings

1. Data preparation and filtering
2. Integration and association analysis
3. Centrality analysis
4. Graphical options

### Pairwise integrative analysis

Choose a data integration method:  
PLS: Partial least squares

Number of components to use in PLS model:  
5

Choose PLS mode (not applicable to RCC option):  
regression

Find optimal number of PLS components? (Note: turning this option ON may increase run time)  
 True  False

---

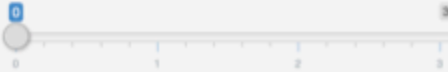
### Association analysis

Correlation Threshold:  
0.4

P-value Threshold For Student's T-test:  
0.05

### Output

Slide to go to next figure:



[https://github.com/kuppal2/xMWAS/blob/master/example\\_manual\\_tutorial/xMWAS-manual.pdf](https://github.com/kuppal2/xMWAS/blob/master/example_manual_tutorial/xMWAS-manual.pdf)

# Step 4. Select method for centrality analysis

*xMWAS - a data-driven integration and network analysis tool (v0.553)*

The screenshot shows the xMWAS web interface. At the top, there are navigation tabs: 'Introduction', 'Analysis', and 'Help and Support'. Below the tabs, there are three main sections: 'Input Files', 'Parameter Settings', and 'Output'. The 'Input Files' section has a 'Choose Files (see help and support)' link. The 'Parameter Settings' section has four steps: '1. Data preparation and filtering', '2. Integration and association analysis', '3. Centrality analysis' (highlighted in blue), and '4. Graphical options'. Below the settings, there are two buttons: 'Start processing' and 'Download results'. The 'Output' section has a slider labeled 'Slide to go to next figure:' with a range from 0 to 3. The 'Method for centrality analysis:' dropdown menu is open, showing the following options: 'eigenvector', 'betweenness', 'degree.count', 'degree.weight', and 'closeness'. The 'eigenvector' option is currently selected.

- Eigenvector: based on the number and quality of connections
- Betweenness: based on the extent to which a node lies on the path between other nodes
- Degree.count: based on the number of connections
- Degree.weight: based on the magnitude of edges (association scores)
- Closeness: based on the closeness of a node to all other nodes



# Step 5. Click on “Start processing” – sit back and relax for a bit

*xMWAS - a data-driven integration and network analysis tool (v0.553)*

Introduction Analysis Help and Support

### Input Files

Choose Files (see help and support)

### Parameter Settings

1. Data preparation and filtering
2. Integration and association analysis
3. Centrality analysis
- 4. Graphical options**

Size of the Labels:

Size of the Nodes:

Maximum number of associations to include in the network (any numeric value >0 or -1 to use all):

Use dataset A as reference?  
 True  False

Node shape for dataset A:

Node shape for dataset B:

Node shape for dataset C:

Node shape for dataset D:

Seed for Random Number Generator:

**Output**

Slide to go to next figure:  3

# Step 6. Download the results

## 4. Graphical options

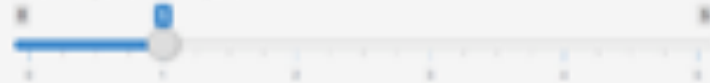
Processing complete. Please click on Download to save the results.

Start processing

Download results

### Output

Slide to go to next figure:



# Additional methods

- Additional methods
  - MINT and Diablo in mixOmics R packages for horizontal and vertical data integration
  - Recent review article by Meng et al. in Bioinformatics reviewed over 20 dimensionality reduction methods for data integration

Comparison of 5 methods for assessing **pairwise associations** between variables implemented in MetabNet, 3Omics, xMWAS, and mixOmics

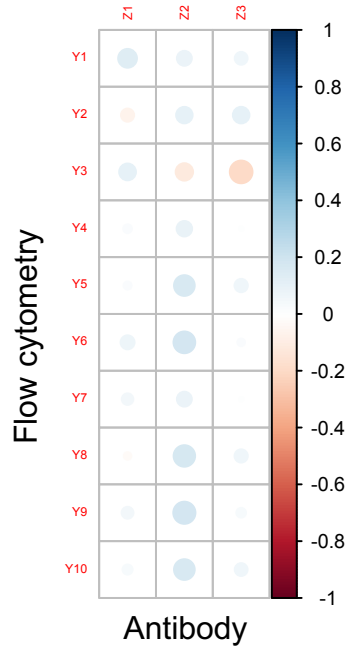
	3Omics	MetabNet	xMWAS	mixOmics
Pearson correlation	x	x	x	x
PLS (regression)			x	x
PLS (canonical)			x	x
Regularized canonical correlation analysis (RCC)				X
sparse generalized canonical correlation analysis (sGCCA)				x

# T cell responses to H1N1v and a longitudinal study of seasonal influenza vaccination (TIV) - 2011

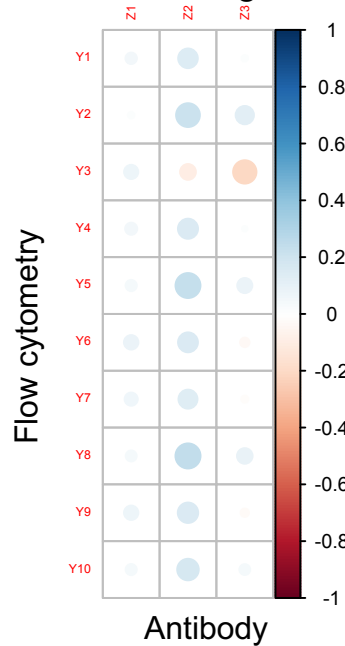
(<https://www.immport.org/shared/study/SDY112>)

- Transcriptomics at day 0: 18,867 genes
- Flow cytometry at day 0: 24 cells
- Antibody at day 0 and 28: 3 antibodies (California, Perth, and Brisbane)
- Number of subjects:
  - Total: 89
  - 85 had matching gene expression, flow cytometry, and antibody data at day 0

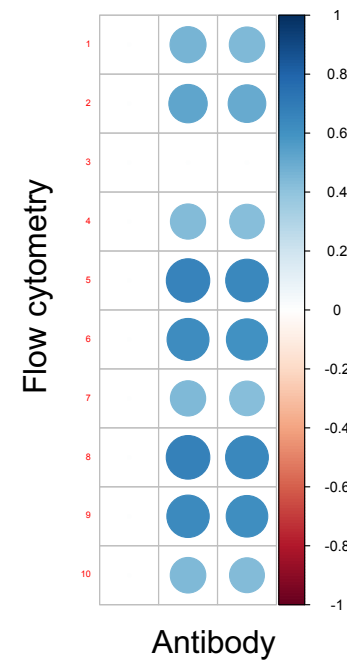
A. Pearson Correlation



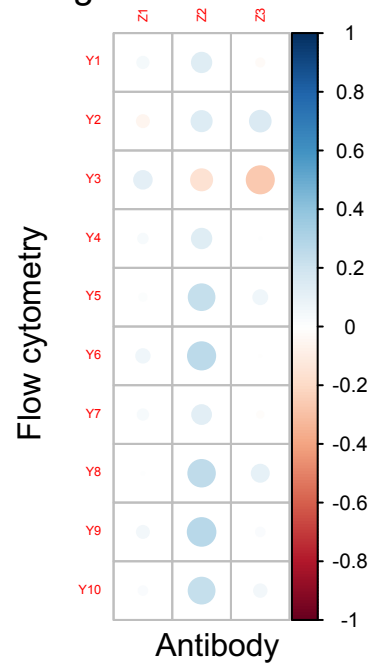
B. PLS – regression mode



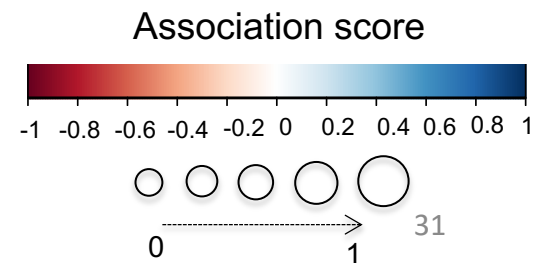
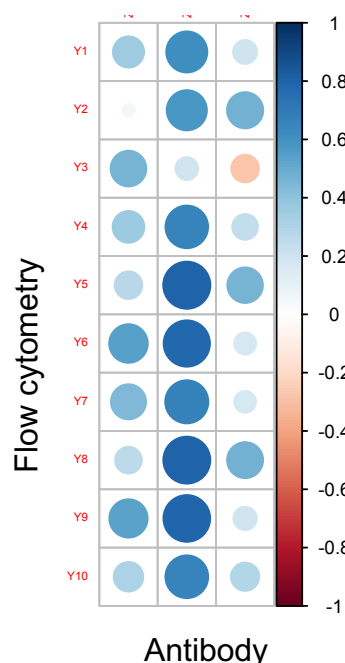
E. Sparse generalized canonical correlation analysis (sGCCA)



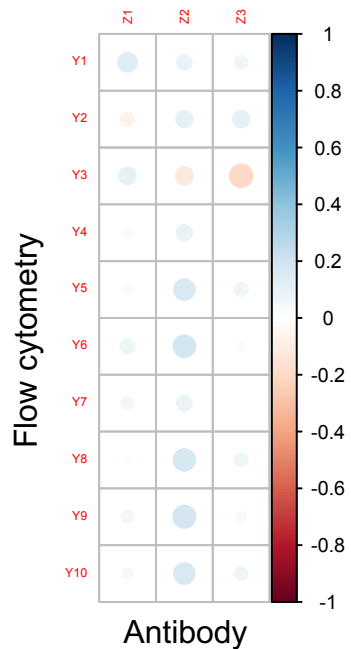
C. Regularized canonical correlation



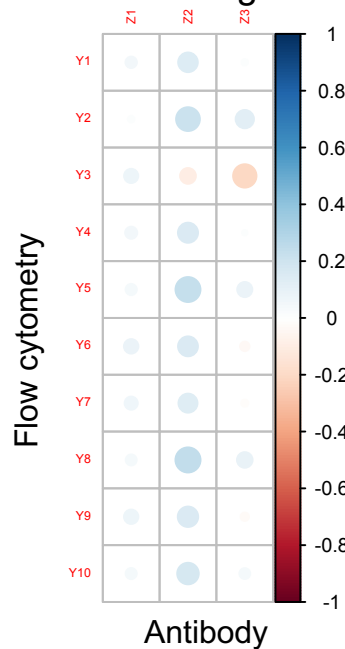
D. PLS - canonical mode



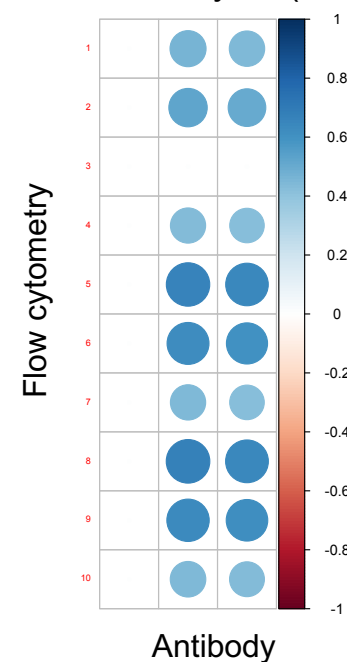
A. Pearson Correlation



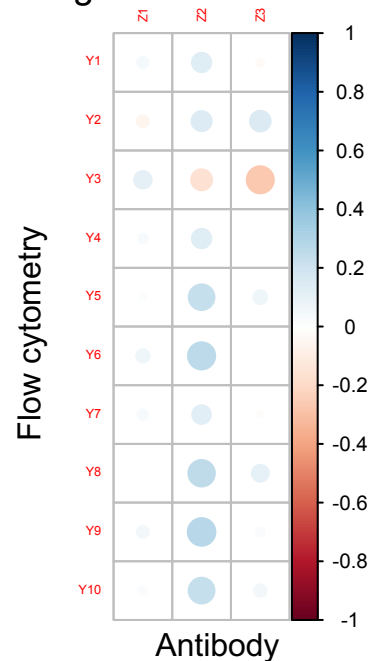
B. PLS – regression mode



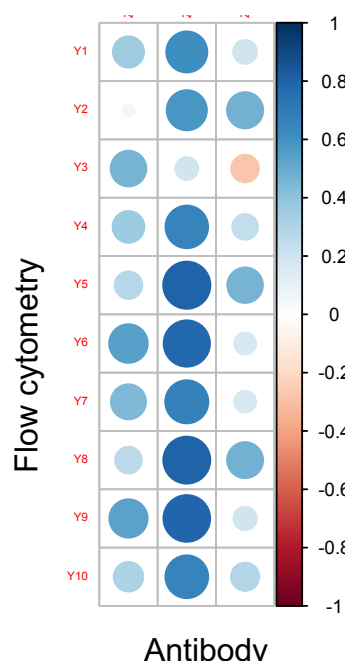
E. Sparse generalized canonical correlation analysis (sGCCA)



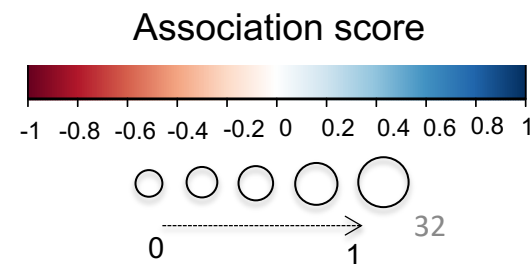
C. Regularized canonical correlation



D. PLS - canonical mode



Inflated association scores using methods A and D – interpretation challenges!





# Main approaches for data integration

- Pathway-based integration
  - Pathway information from KEGG or other databases
  - Datasets are analyzed individually (differentially expressed genes, metabolites, proteins) and integration is performed at the pathway level
  - Examples: MetaboAnalyst, iPEAP, MetScape, MetaCore
- Data-driven integration using meta-dimensional analysis
  - Integration is performed globally such that data from multiple omics layers are combined simultaneously
  - Interpretation using pathway analysis tools
  - Examples: 3Omics, mixOmics, xMWAS
- **Using literature-derived associations for integration**
  - **Using co-occurrence criteria for establishing relationship**
  - **Examples: Comparative Toxicogenomics Database, CoPub, ArrowSmith, SEACOIN**



# Comparative Toxicogenomics Database: Chemical x Disease associations

## Trichloroethylene

Basics Gene Interactions Genes **Diseases** Phenotypes Comps Pathways GO Exposure Studies Exposure Details Refe

These diseases are associated with *Trichloroethylene* or its descendants. Each association is *curated* (M **mark**tion).

### Disease categories [\[Hide chart\]](#)



### Disease categories [\[Show chart\]](#)



Chemical	Disease	Direct Evidence	Enrichment Analysis	Inference Network	Inference Score	References
1. Trichloroethylene	Carcinoma, Hepatocellular			129 genes: A2M   ABCB1   ACOX1   ACTB   ADH4   APC   ARHGAP39   ASF1B   ASPH   AURKA   BCL2L1   BIRC5   C9   CCN1   CCNB1   CCNB2   CCND1   CDC20   CDC42   CDCA3   CDCA5   CDC48   CDK1   CDKN3   CEBPB   CENPA   CENPE   CEP55   COMT   CTNNB1   CXCL8   CYP1A1   CYP1A2   CYP2E1   DCN   DLGAP5   E2F8   ECM1   ECT2   EGF   EGFR   EGR1   ENO1   F2   FASN   FDFP1   FOSB   FOXM1   GDF15   GHR   GLUL   GNMT   HAO2   HERC5   HMGR   HMMR   HSPAS   HSPB1   IGFALS   IL1RAP   IL1RN   IL6   INMT   IRF2   IRS1   JUN   KIF15   KIF20A   LIFR   LRAT   LRRC59   MAD2L1   MCM10   MEI   MED1   MET   MFSD2A   MKI67   MT1A   MTOR   MYK   MYC   NAT2   NEK2   NFE2L2   NFKBIA   NNMT   NR0B2   NUSAP1   PSK   PCK1   PDGFB   PDSA3   PDK4   PHLDA1   PKM   PLK1   PPARC   PRC1   PRDX1   PRDX2   PTEN   PTGS2   PTH1R   RACGAP1   RCAN1   ROBO1   SCD   SERPINA1   SLC22A1   SLC25A47   SLC2A2   SOCS2   SPC25   STMN1   TACC3   TGFBI   TNF   TP53   TPX2   TRAP1   TTK   UBD   UBE2C   UHRF1   VCAM1   YAP1   ZFP36   ZWINT	123.53	63
2. Trichloroethylene	Carcinoma			49 genes: ABCB1   ACSL1   ACTB   ALCAM   ANGPTL4   ATP1F1   BASP1   BCL2   BCL2L1   CCND1   COL5A1   CSN2   DNMT1   DNMT3B   EEF1A1   EGFR   EIF5A   ENO1   G0S2   GATM   GDA   GGT1   GSTT1   ID4   IL6ST   L5R   MET   MYC   PEBP1   PGAM1   PKM   PPARA   PRDX2   PRLR   PTEN   PTGS2   RAN   RASGRF2   RPL3   S1PR1   SCD   SPINT2   STMN1   TGFBI   TXT   TP53   TSC22D1   UCP2   VHL	45.93	22
3. Trichloroethylene	Breast Neoplasms			93 genes: ABCB1   ANGPTL4   AURKA   BAG1   BARD1   BAX   BCL2   BCL2A1   BIRC5   C1QB   CCND1   CD74   COMT   CPT1A   CSF1   CTNNB1   CXCL8   CXCR4   CYP1A1   CYP1B1   ODDT3   DLL1   DNMT1   DNMT3A   DNMT3B   EEF2   EGF   EGFR   ENO1   ERBB2   ERBB3   EVL   FASN   FGFR2   FOXM1   FOXP3   GJA1   GPNMB   GUCY1A2   HADHB   HAPLN4   HMMR	45.33	140

# Comparative Toxicogenomics Database: Chemical x Gene associations

Trichloroethylene

Basics **Gene Interactions** Genes Diseases Phenotypes Comps Pathways GO Exposure Studies Exposure Details References

1-50 of 2,604 results.

First Previous 1 2 3 4 5 6 7 8 Next Last

	Interacting Chemical	Interacting Gene	Interaction	References	Organisms
1.	Trichloroethylene	A2M	Trichloroethylene results in increased methylation of A2M gene	1	1
2.	Trichloroethylene	A430017K17	Trichloroethylene results in decreased expression of A430017K17 mRNA	1	1
3.	Trichloroethylene	ABCA1	Trichloroethylene results in decreased methylation of ABCA1 gene	1	1
4.	Trichloroethylene	ABCA12	ABCA12 affects the susceptibility to Trichloroethylene	1	1
5.	Trichloroethylene	ABCA13	Trichloroethylene results in increased methylation of ABCA13 gene	1	1
6.	Trichloroethylene	ABCA17	Trichloroethylene results in increased methylation of ABCA17 gene	1	1
7.	Trichloroethylene	ABCB1	Trichloroethylene results in increased expression of ABCB1 mRNA	1	1
8.	Trichloroethylene	ABCB10	Trichloroethylene results in increased methylation of ABCB10 gene	1	1
9.	Trichloroethylene	ABCB9	Trichloroethylene results in increased methylation of ABCB9 gene	1	1
10.	Trichloroethylene	ABCC3	Trichloroethylene results in increased methylation of ABCC3 gene	1	1
11.	Trichloroethylene	ABCC4	Trichloroethylene results in increased methylation of ABCC4 gene	1	1
12.	Trichloroethylene	ABCC5	Trichloroethylene results in increased methylation of ABCC5 gene	1	1
13.	Trichloroethylene	ABCD2	Trichloroethylene results in increased expression of ABCD2 mRNA	1	1
14.	Trichloroethylene	ABCF2	Trichloroethylene results in decreased methylation of ABCF2 gene	1	1
15.	Trichloroethylene	ABHD6	Trichloroethylene results in increased expression of ABHD6 mRNA	1	1
16.	Trichloroethylene	ACAA1A	Trichloroethylene results in increased expression of ACAA1A mRNA	2	1
17.	Trichloroethylene	ACAA1B	Trichloroethylene results in increased expression of ACAA1B mRNA	3	2
18.	Trichloroethylene	ACACB	Trichloroethylene results in increased expression of ACACB mRNA	1	1
19.	Trichloroethylene	ACADL	Trichloroethylene results in increased expression of ACADL mRNA	2	1
20.	Trichloroethylene	ACADM	Trichloroethylene results in increased expression of ACADM mRNA	1	1
21.	Trichloroethylene	ACADS	Trichloroethylene results in decreased expression of ACADS mRNA	1	1
22.	Trichloroethylene	ACADVL	Trichloroethylene results in increased expression of ACADVL mRNA	1	1
23.	Trichloroethylene	ACADVL	Trichloroethylene results in increased expression of ACADVL protein	1	1

# Case Study: Integrative analysis of platelet metabolome with mitochondrial bioenergetics

N=13 healthy volunteers

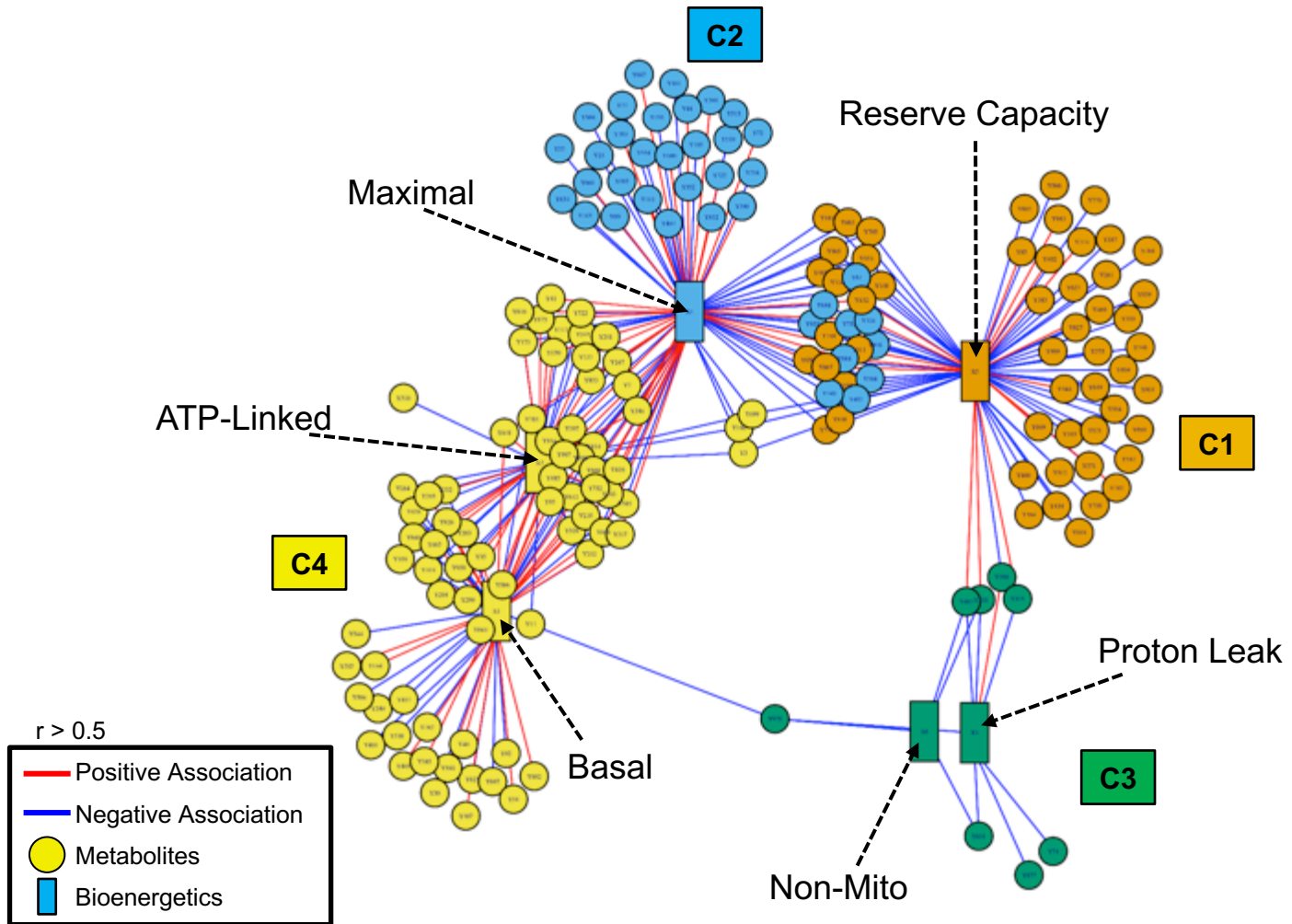
Mitochondrial bioenergetics: 6 energetic parameters (ATP-linked, basal, proton leak, maximal, non-mitochondrial, and reserve capacity OCR)

High-resolution metabolomics: 2,705 metabolic features

Chacko BK, Smith MR, Johnson MS, Benavides G, Culp ML, Pilli J, Shiva S, Uppal K, Go YM, Jones DP, Darley-Usmar VM.

Mitochondria in precision medicine; linking bioenergetics and metabolomics in platelets. **Redox Biol.** 2019

Collaboration between Emory and UAB



# Case Study: Application of xMWAS for integrative network analysis of metabolome and metallome datasets from the Strong Heart study

N=145 (12 American Indian communities; free of type-2 diabetes)

Metallome: arsenobetaine (AsBe), monomethylarsonate (MMA), dimethylarsinate (DMA), inorganic arsenic (iAs), cadmium (Cd), lead (Pb), antimony (Sb), tungsten (W), uranium (U), zinc (Zn), selenium (Se), molybdenum (Mo)

Metabolome: High-resolution metabolomics data for 8,810 features

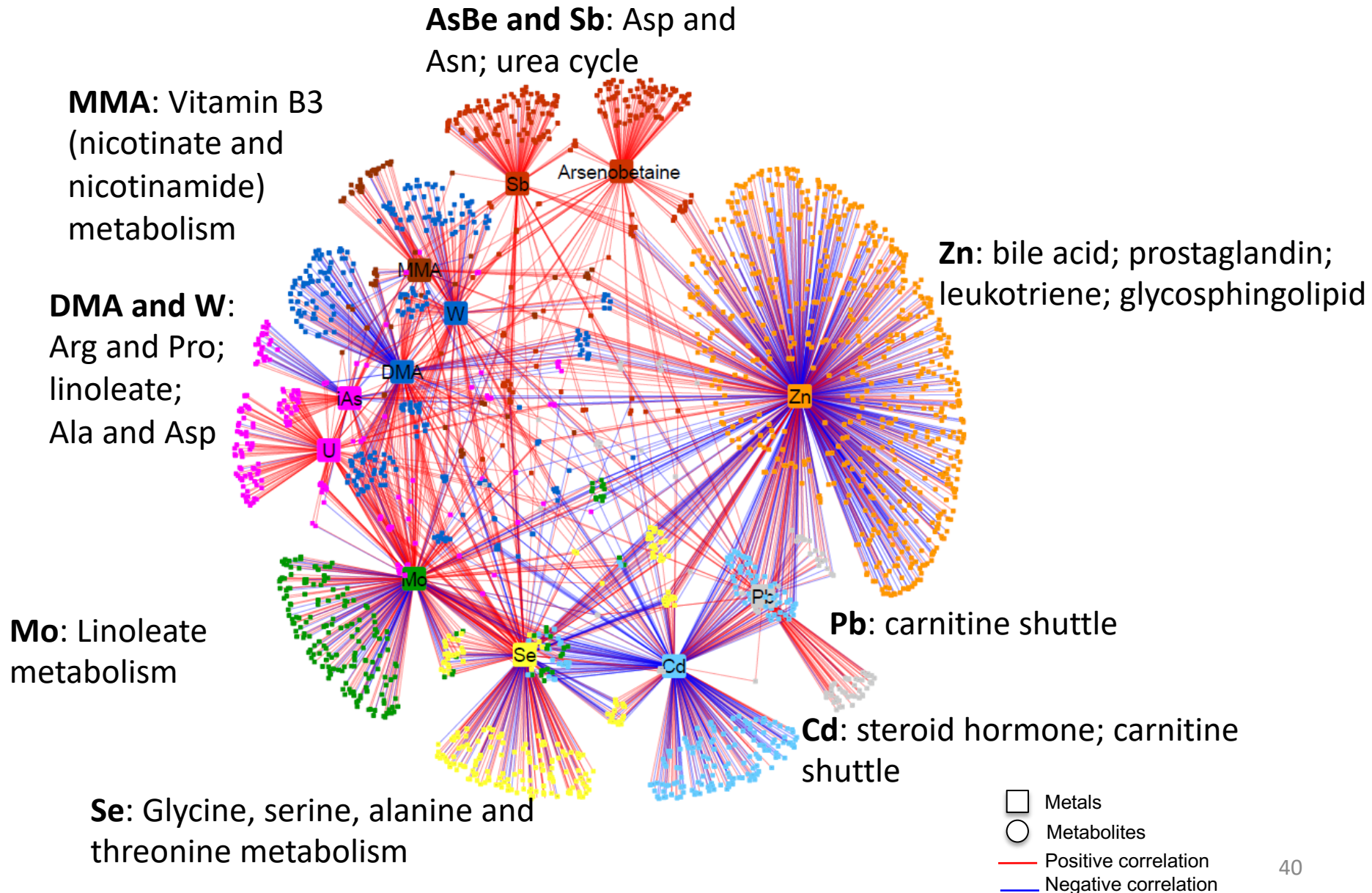
Tiffany R. Sanchez, Xin Hu, Nancy Loiacono, ViLinh Tran, Jinying Zhao, Young-Mi Go, Dean P. Jones, Ana Navas-Acien, Karan Uppal (In preparation)



Collaboration between Emory and Columbia Universities



# Metallome-metabolome integrative network using xMWAS





**Case Study:** Integrative network analysis of clinical, biomolecular (metabolites, microRNAs, plasma protein markers, and cytokines), and environmental exposure data from a dataset of 66 service personnel post-deployment

Juilee Thakar, Thomas H. Thatcher, Matthew Ryan Smith, Collynn F. Woeller, Douglas I. Walker, Mark J. Utell, Philip K. Hopke, Timothy M. Mallon, Pamela L. Krahl, Patricia Rohrbeck, Young-Mi Go, Dean P. Jones, and Karan Uppal  
(under revision)

Collaboration between Emory, Rochester, and Department of Defense

# Input data for xMWAS

1. **Molecular data:** metabolites, miRNAs, cytokines, and proteins  
(3,274 molecular variables x 66 subjects)

	Subject1	Subject2	-	Subject N
Metabolite1	199	19	-	100
-	-	-		-
miRNA1	50	30	-	20
-	-	-		-
Cytokine1	33	12	-	39

2. **Environmental chemicals**  
(5 variables x 66 subjects)

	Subject1	Subject2	-	Subject N
DioxinPC1	3	2.5	-	13
DioxinPC2	5	1.4	-	10
DioxinPC3	2	13	-	5
Cotinine	1	4	-	9
Benzo(a) pyrene diol epoxide	5	3	-	2

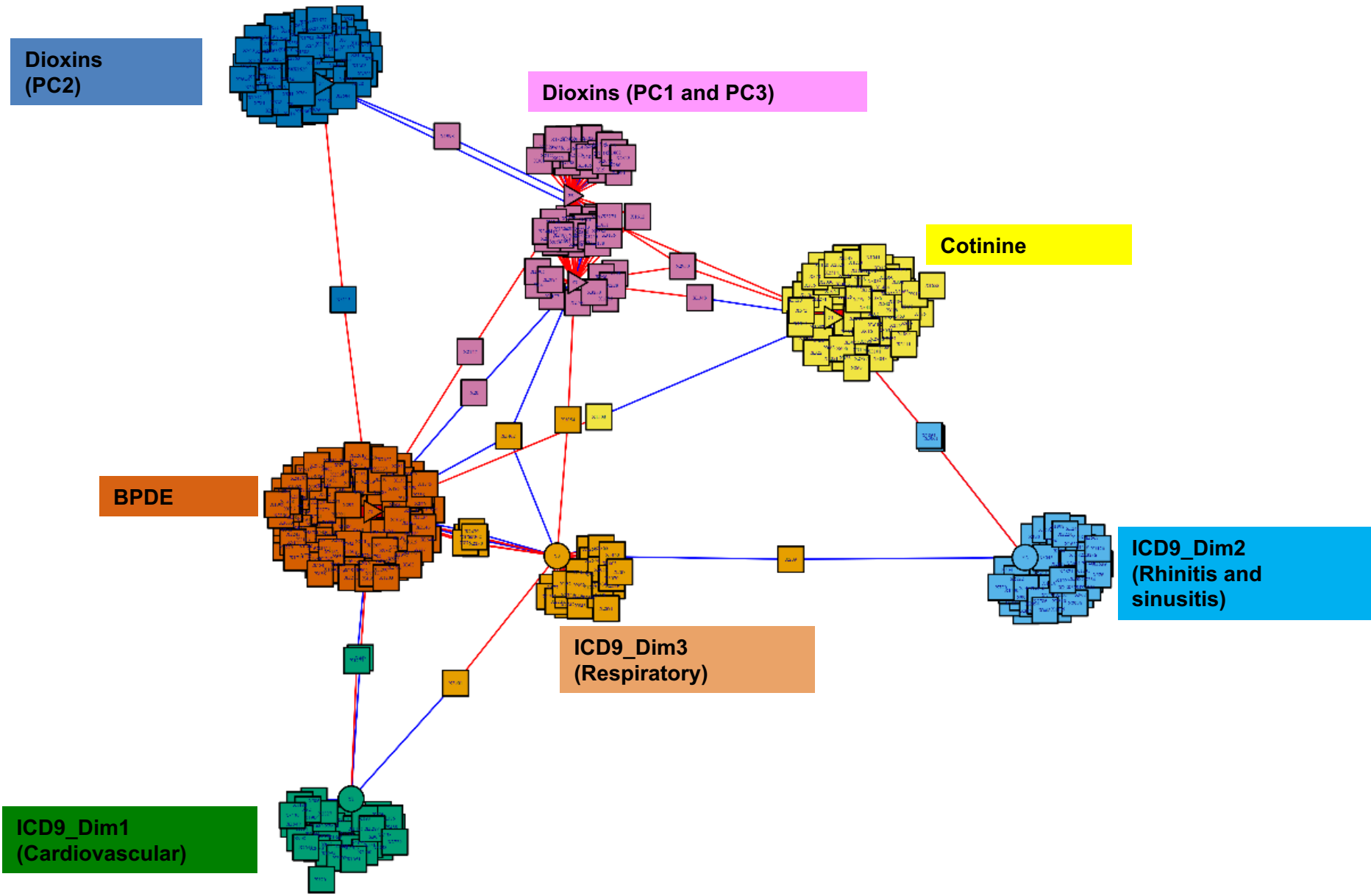
3. **Health outcomes:** ICD-9 codes (49 cardiopulmonary ICD-9 codes x 66 subjects)

	Subject1	Subject2	-	Subject N
4019	0	1	-	0
4011	1	1	-	0
-	-	-	-	-
49301	1	0	-	0

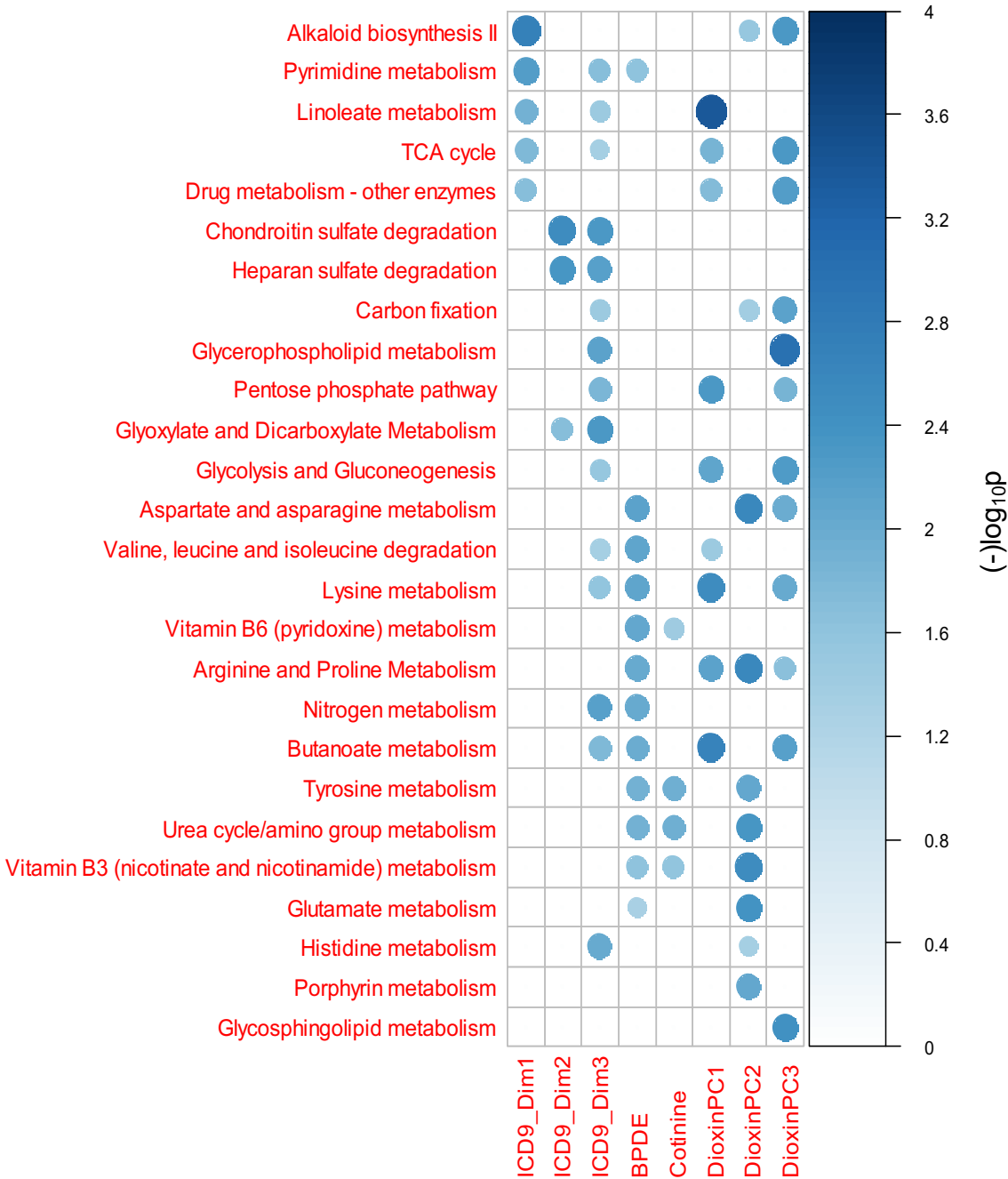


Multiple correspondence analysis  
(8 dimensions x 66 subjects)  
(Only dimensions with >5% variance explained were included)

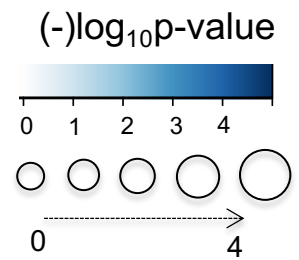
	Subject1	Subject2	-	Subject N
Dim 1	199	19	-	100
Dim 2	10	40		90
-	-	-		-
Dim 8	50	30	-	20



Each community (C) is represented by a different color:  
 $|r| > 0.3; p < (C1; C2; C3; C4; C5; C6; C7;$

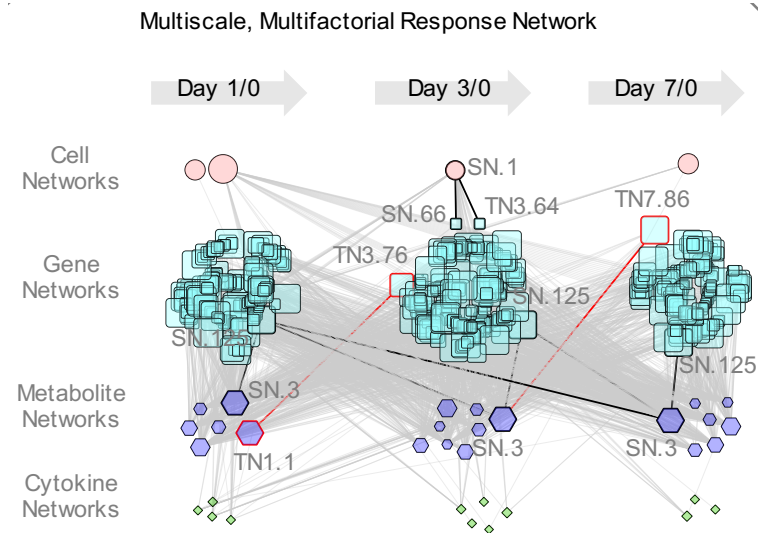


- Bubble plot showing metabolic pathways associated with clinical and environmental exposures data
- Metabolic pathway analysis performed using Mummichog



# Current challenges and future work

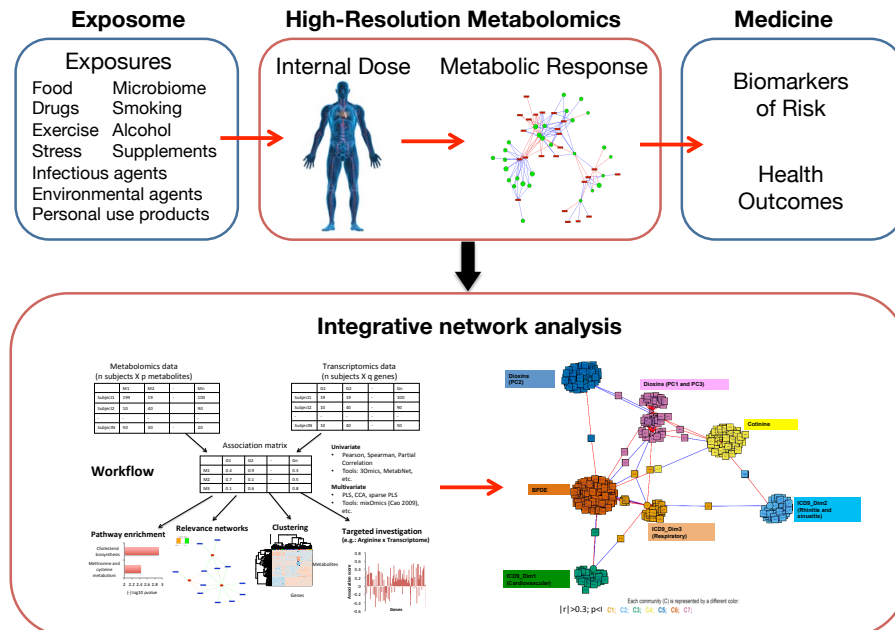
- Development of hybrid methods
  - combined knowledge-based and data-driven approaches
  - incorporation of literature-derived associations in xMWAS
    - Using co-occurrence criteria for establishing relationship (PolySearch2.0)
- Improving scalability
  - Ability to handle >100,000 variables
  - Performing integrative analysis at communities, clusters, or eigenvariables (first PCs) level



Li et al., 2017. *Cell* 169, 862–877

# Summary

- Various tools and techniques are available for integrating and visualizing multi-omics data
- Integrative network analysis approaches can be used to understand the multi-scale interactions between environmental exposures, molecular response, and health outcomes



# Acknowledgements



Emory



Dean Jones, Young-Mi Go, Shuzhao Li, Chunyu Ma, Ken Liu, Kristine Dennis, ViLinh Tran, Michael Orr, Ryan Smith, Xin Hu, Jolyn Fernandes, Bill Liang, Yating Wang, Tiantian Zhang

## **Collaborators:**

Victor M. Darley-Usmar, Ana Navas-Acien, Tiffany R. Sanchez, Nancy Loiacono, Jinying Zhao, Timothy Mallon, Mark Utell, Juilee Thakar, Gary Miller, Douglas Walker, Milam Brantley, Ihab Hajjar, Arshed Quyyumi, John Roback, MoTrPAC, and others

## **Funding**

NIEHS, NIA, NCI, NHLBI, NIDDK, NIAAA, NIAID, Woodruff Foundation, Emory Dept of Medicine, Georgia Research Alliance Exposome Center Integrated Health Science and Facilities Core  
NIEHS P30 ES019776  
DK112341 (MoTrPAC)  
AG057470  
AA026928  
EY022618  
ES026071  
DK117246-01

# Questions?

Email: [kuppal2@emory.edu](mailto:kuppal2@emory.edu)

Website: [www.csm2l.com](http://www.csm2l.com)