# Computational methods for data integration

**Karan Uppal, PhD**

Assistant Professor of Medicine

Director of Computational Systems Medicine & Metabolomics Lab

Emory University

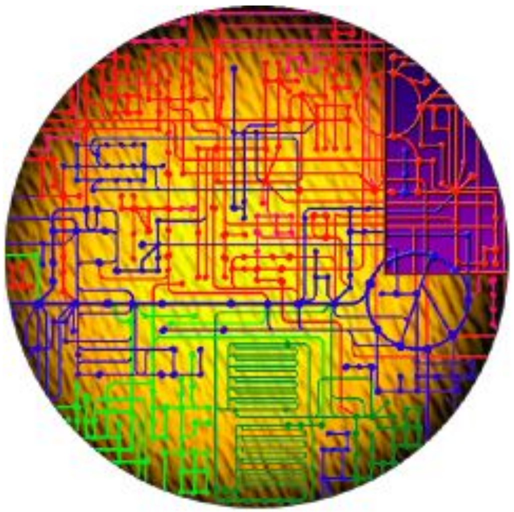
# Learning Objectives

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

# Introduction: A Systems Biology Framework

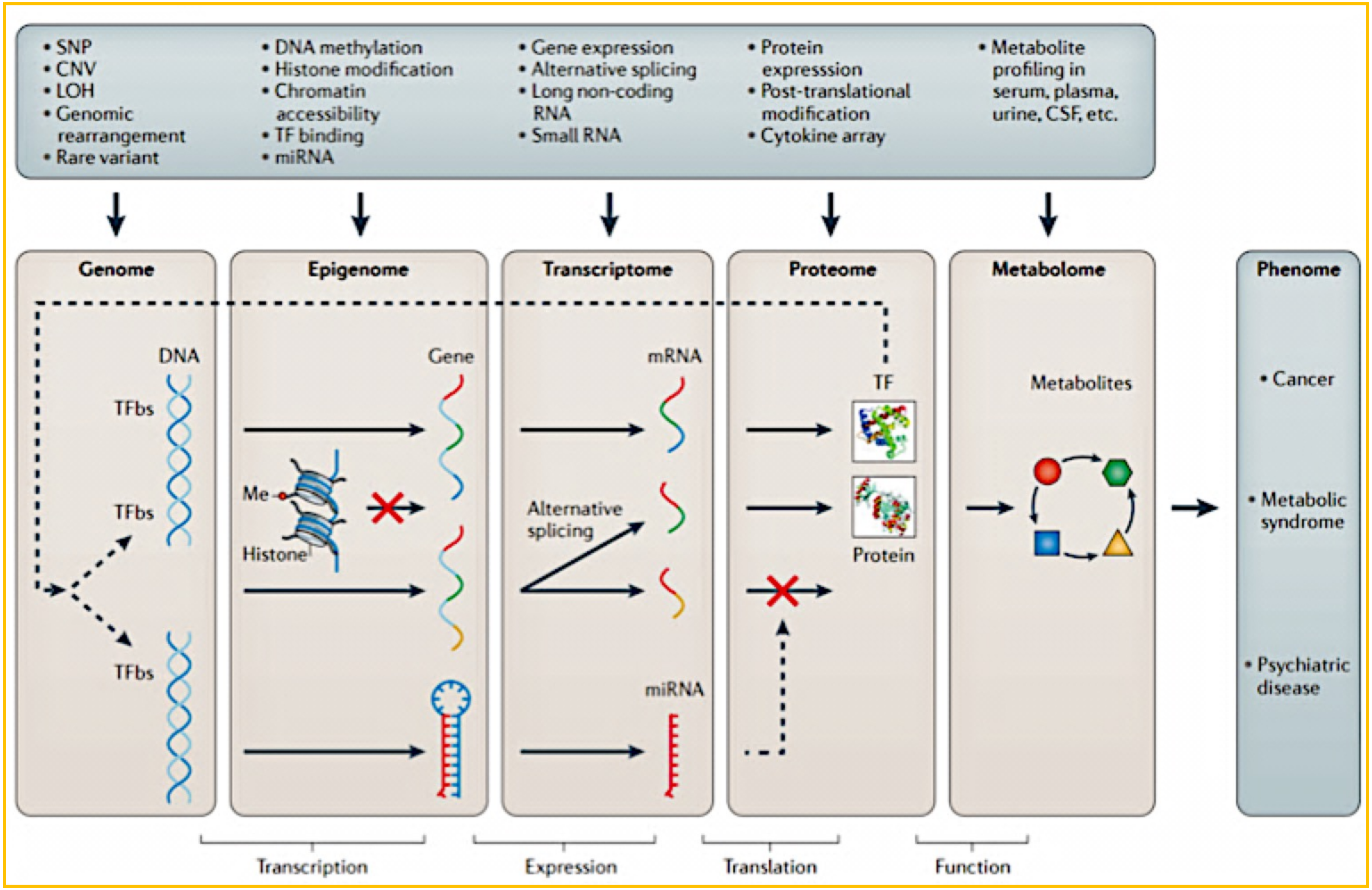
- The goal of **Systems Biology**:
  - Systems-level understanding of biological systems
  - Analyze not only individual components, but their interactions as well and emergent behavior

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*"
C. Henry and C. Washington

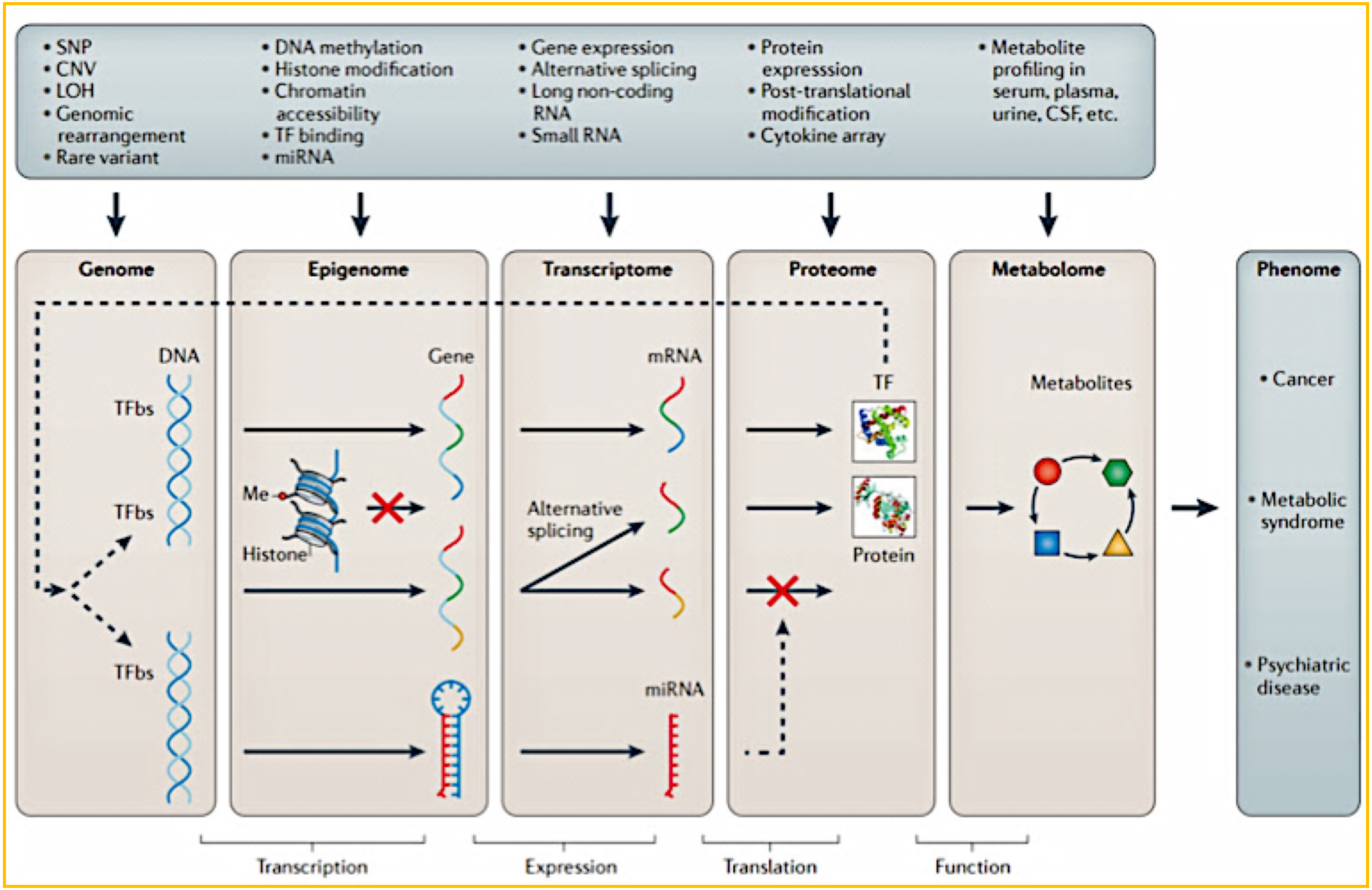
# **Integrative omics:** dissecting the biological system via -omics






Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. 2015. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet *16:85-97.*

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





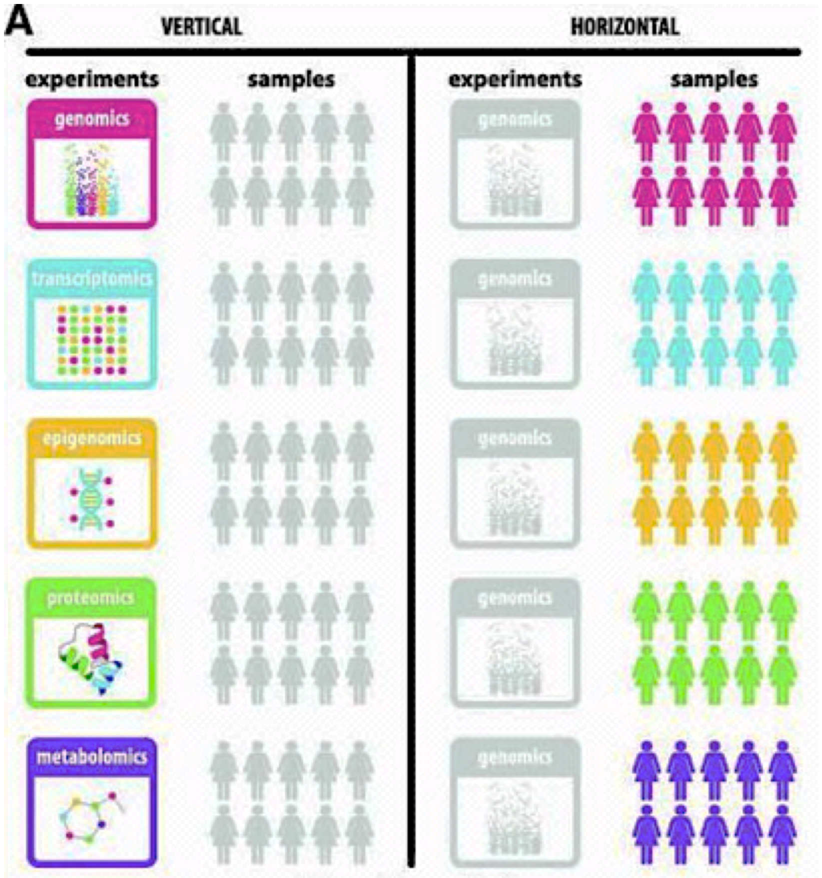

Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. 2015. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet *16:85-97.*

# 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

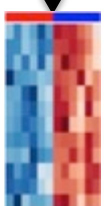
# Main approaches for data integration

- **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**

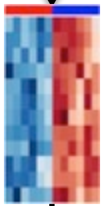
**Over-represented pathways**

| 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 |



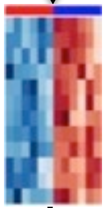
**Differentially expressed metabolites**

**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 |

iPEAP (Sun 2013)

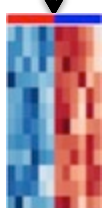
# Pathway-based data integration - II

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 |

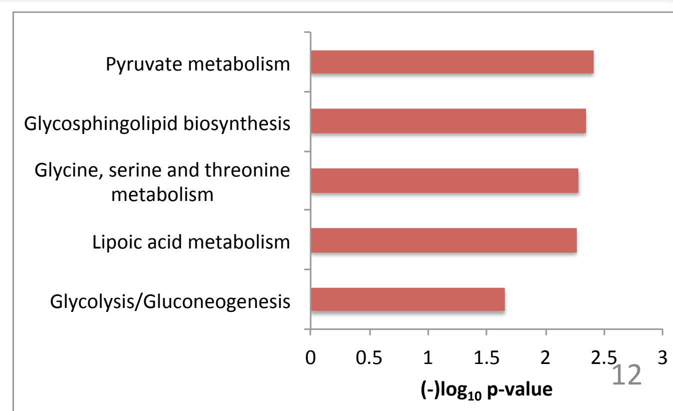


**Differentially expressed metabolites**

**Differentially expressed genes**

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

Over-representation analysis in KEGG using gene and metabolite IDs

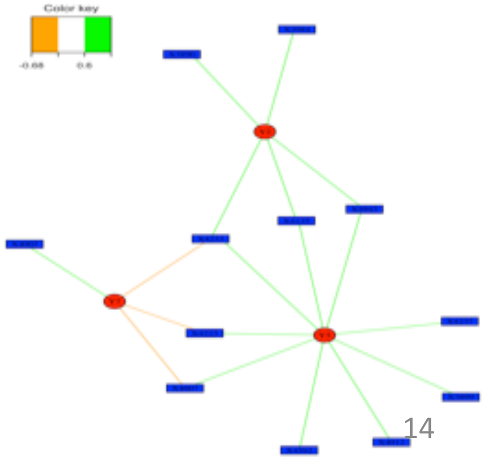


# Main approaches for data integration

- 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 a 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/

Transcriptomics

| Gene | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|
| akap9 | -0.24 | -0.6 | -0.47 |
| macf1 | -0.3 | -0.3 | 0.48 |
| RNPEP | 0.24 | 0.85 | 0.15 |
| SDHA | 0.1 | 0.37 | 0.18 |

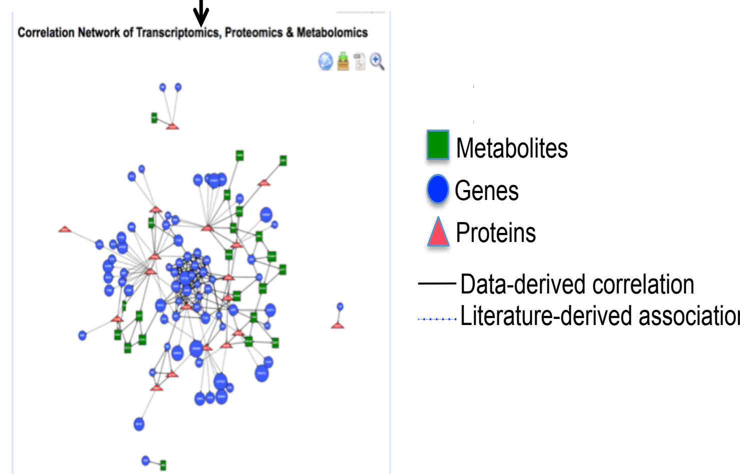
Metabolomics

| Metabolite | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|
| 4277439 | -0.3109937 | -0.2792995 | -0.2548517 |
| 441 | -0.2967872 | -0.2895908 | -0.2674823 |
| 69362 | -0.3183692 | -0.2828533 | -0.272917 |
| 10258 | -0.0614116 | -0.1180467 | -0.1231662 |

Proteomics

| Protein | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|
| P14060 | 2.06 | 1.2 | 1.61 |
| P26439 | 1.8 | 3.57 | 2.04 |
| P29372 | -0.64 | -0.71 | -0.21 |
| 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

| Gene | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|
| akap9 | -0.24 | -0.6 | -0.47 |
| macf1 | -0.3 | -0.3 | 0.48 |
| RNPEP | 0.24 | 0.85 | 0.15 |
| SDHA | 0.1 | 0.37 | 0.18 |

### Metabolomics

| Metabolite | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|
| 4277439 | -0.3109937 | -0.2792995 | -0.2548517 |
| 441 | -0.2967872 | -0.2895908 | -0.2674823 |
| 69362 | -0.3183692 | -0.2828533 | -0.272917 |
| 10258 | -0.0614116 | -0.1180467 | -0.1231662 |

### Proteomics

| Protein | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|
| P14060 | 2.06 | 1.2 | 1.61 |
| P26439 | 1.8 | 3.57 | 2.04 |
| P29372 | -0.64 | -0.71 | -0.21 |
| Q96J02 | -0.52 | -1.34 | -0.15 |

↓

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

↓

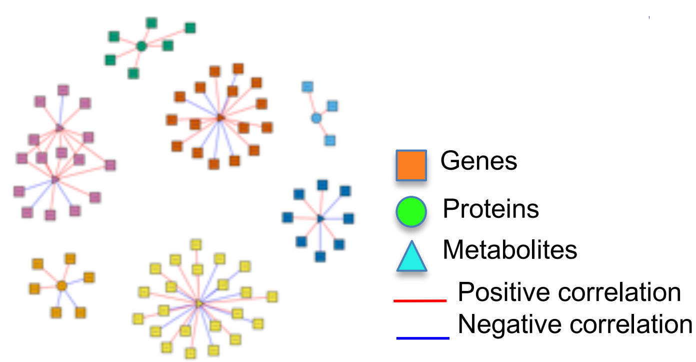
Approximation of Pearson correlation using PLS components

↓

Filtering based on |r|>threshold and p-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 $\mathrm{cov}(X_{u,} Y_v)$
where
$u_1, u_2 \ldots u_H$ and $v_1, v_2 \ldots v_H$ 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 |

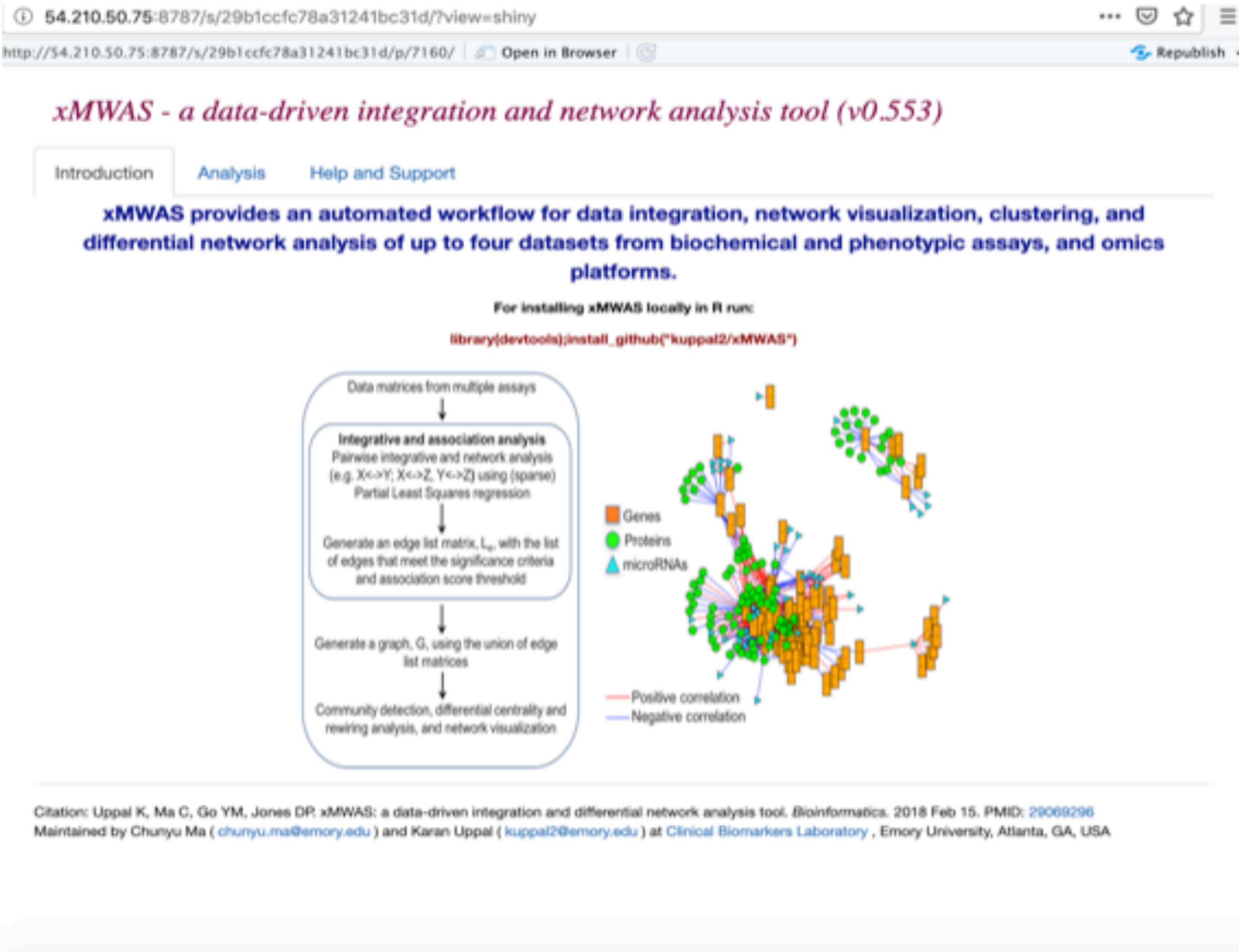
# Community detection

- 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

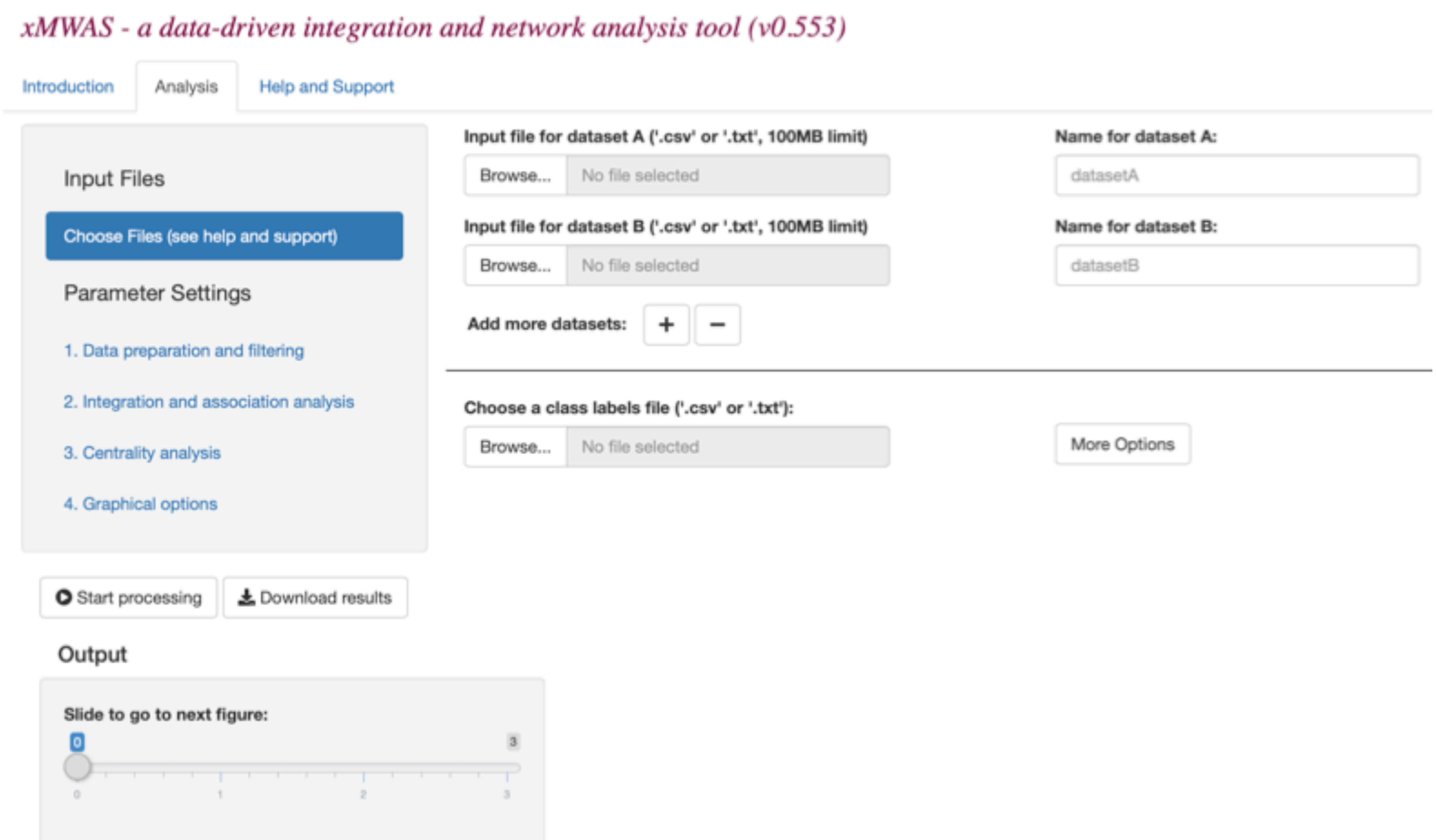
# Centrality analysis

- 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. $|centrality_{exposed} – centrality_{control}|$

# xMWAS: ShinyApp interface

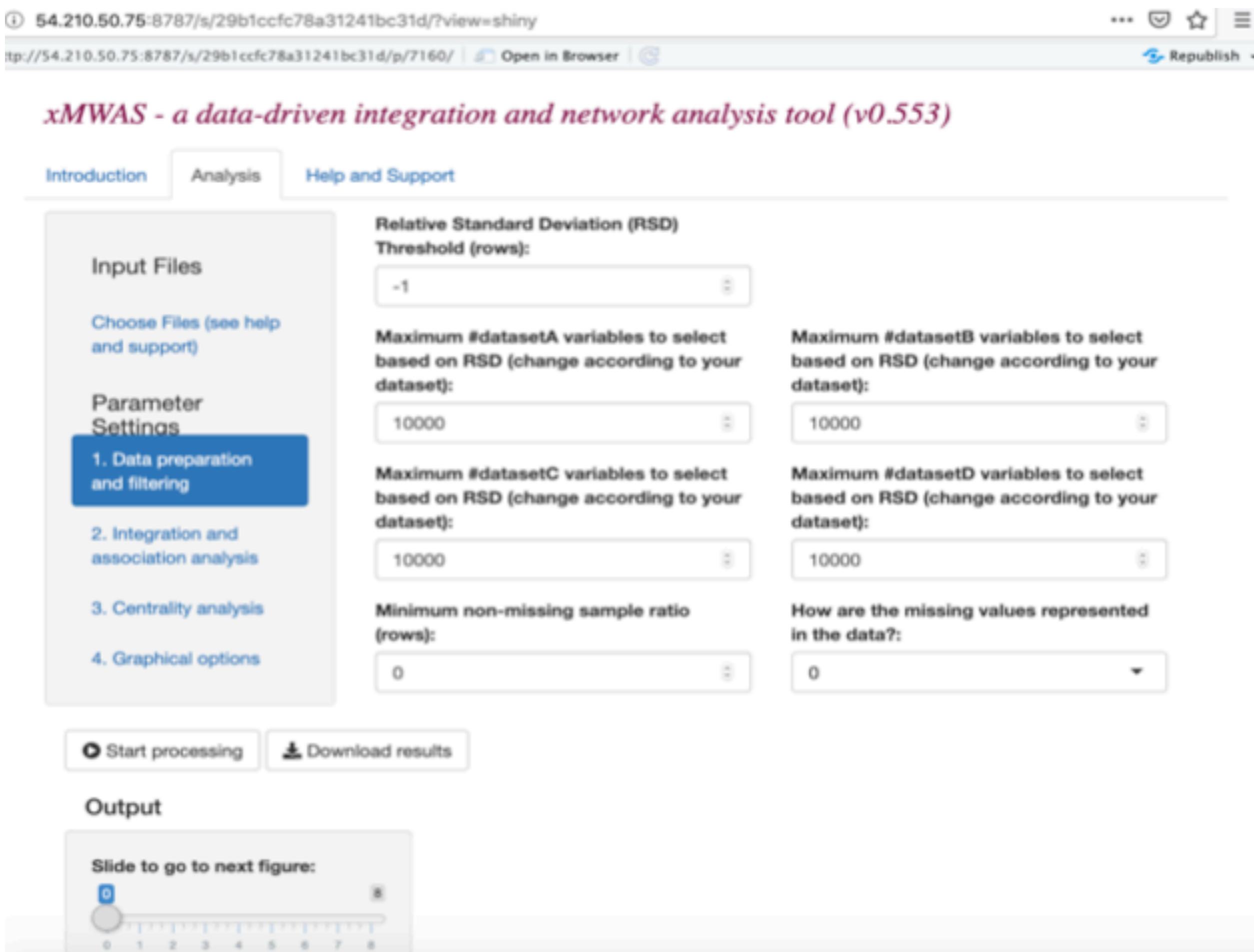


# Step 1. Upload data files



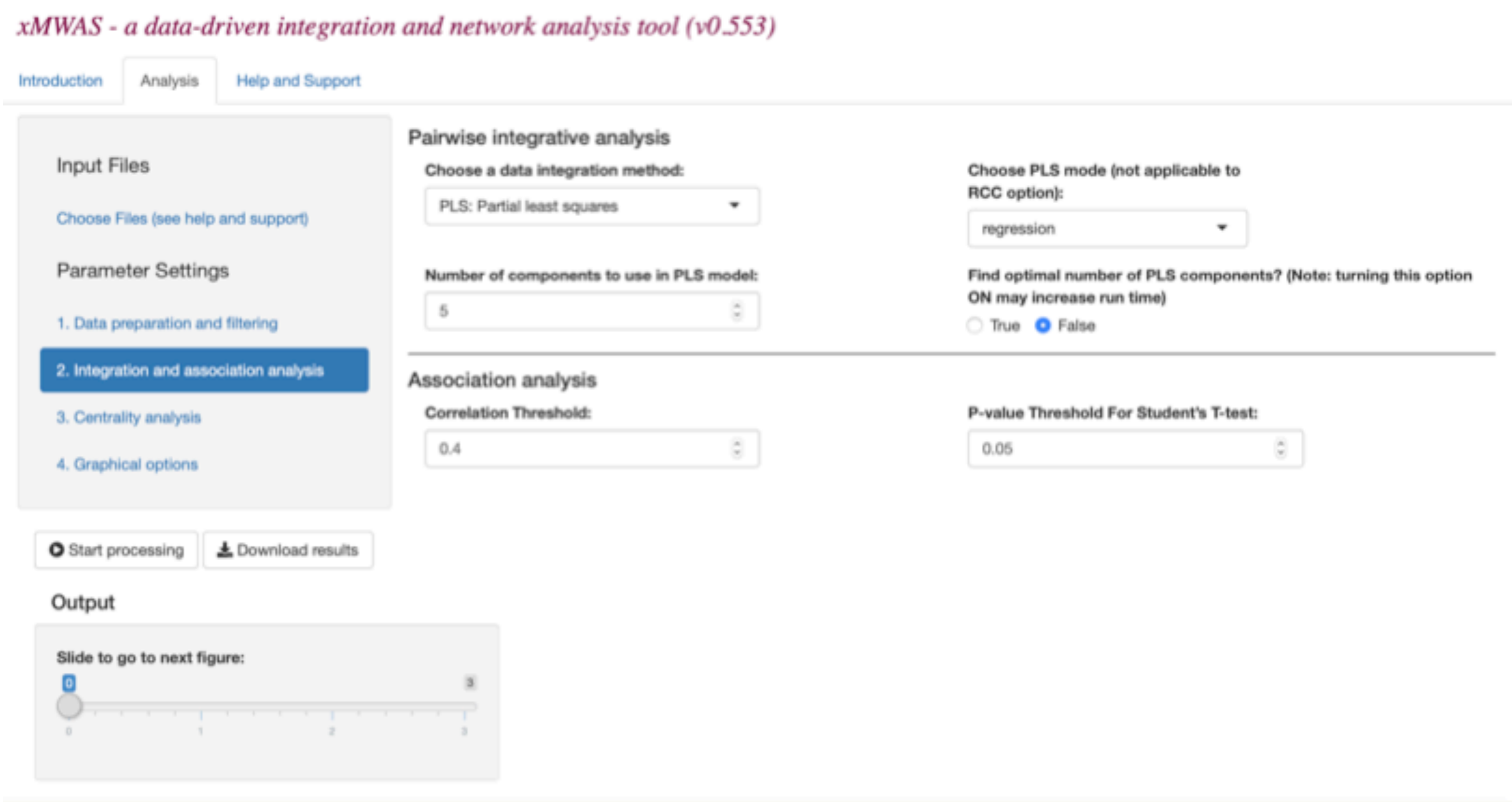

https://kuppal.shinyapps.io/xmwas/
(See: Help & Support)

# Step 2. Data preprocessing and filtering



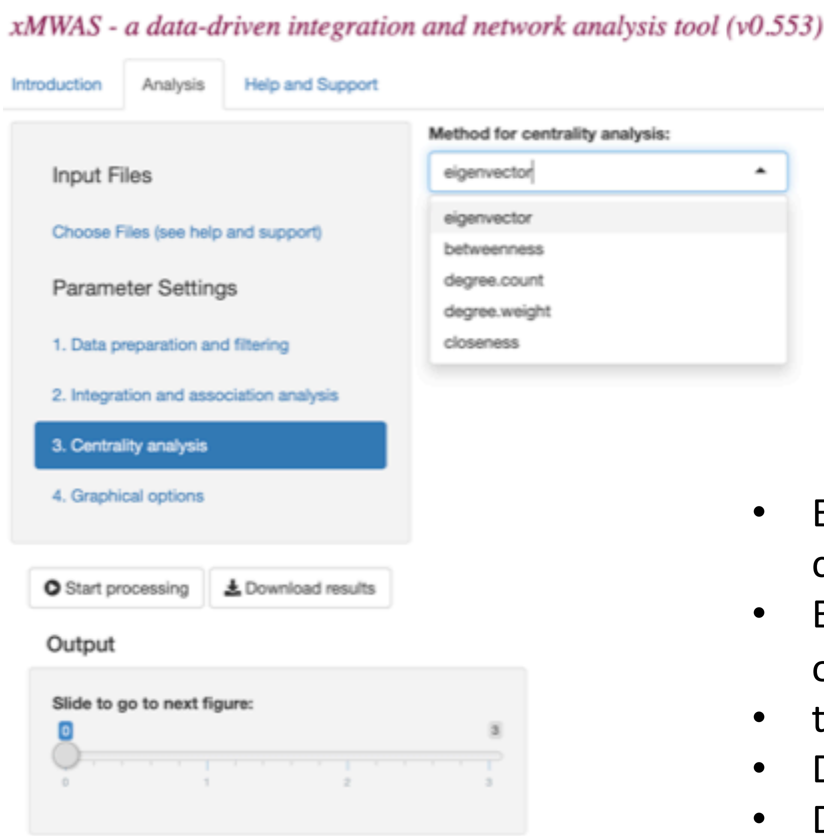

https://github.com/kuppal2/xMWAS/blob/master/example_manual_tutorial/xMWAS-manual.pdf

# Step 3. Set parameters for integration and association analysis



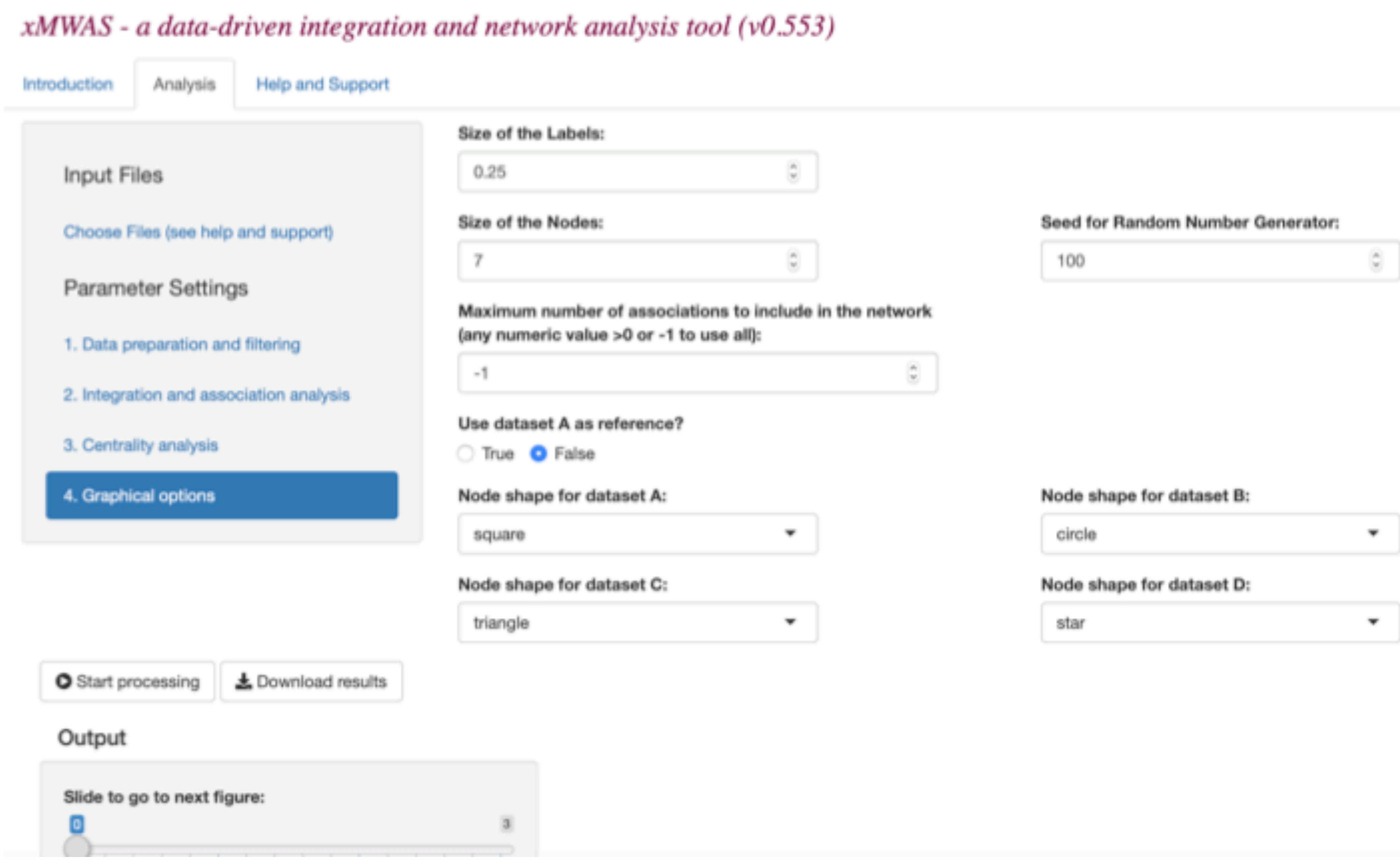

https://github.com/kuppal2/xMWAS/blob/master/example_manual_tutorial/xMWAS-manual.pdf

# Step 4. Select method for centrality analysis




- 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



# Step 6. Download the results

# 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

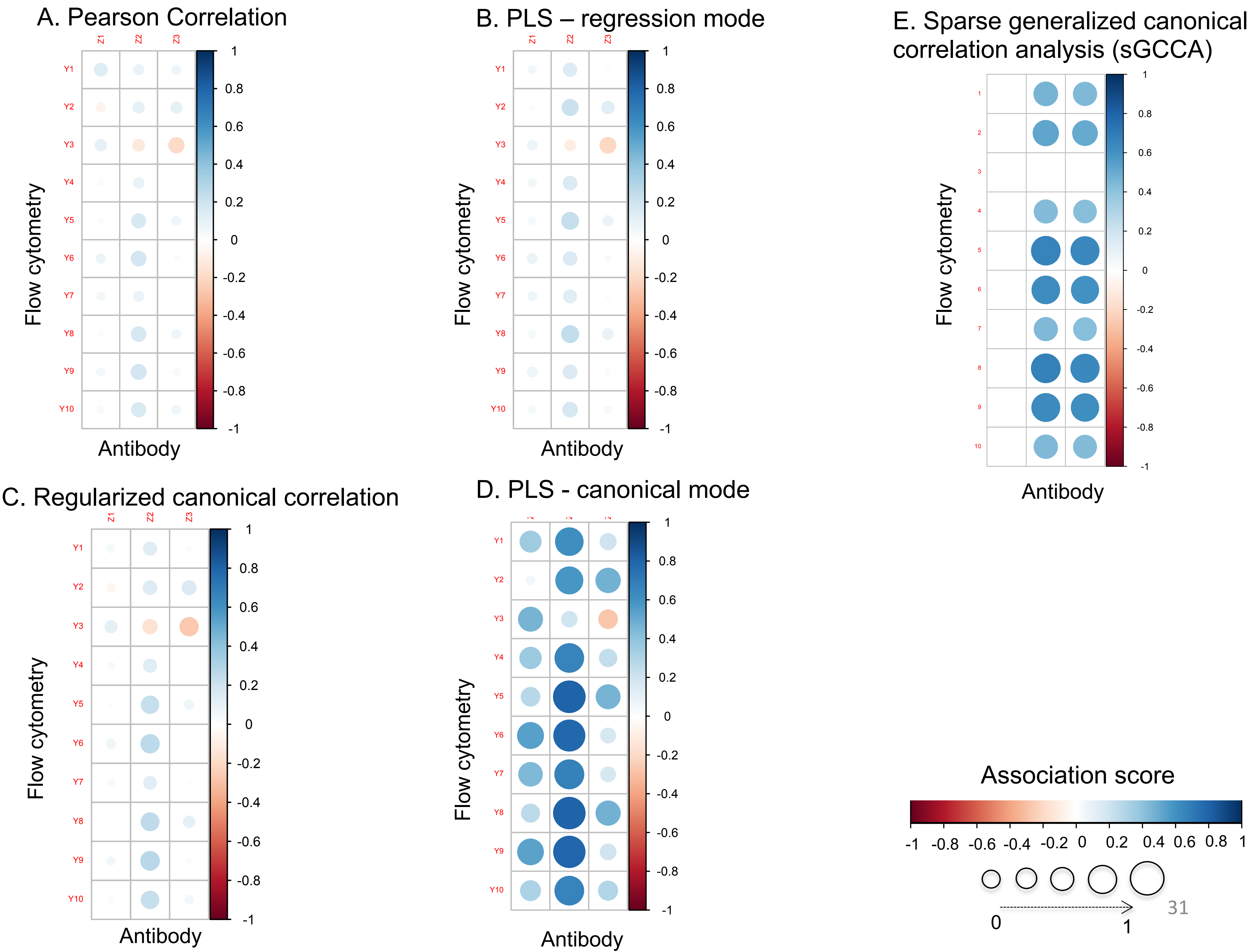
# Comparison of 5 methods for assessing **pairwise associations** between variables implemented in MetabNet, 3Omics, xMWAS, and mixOmics

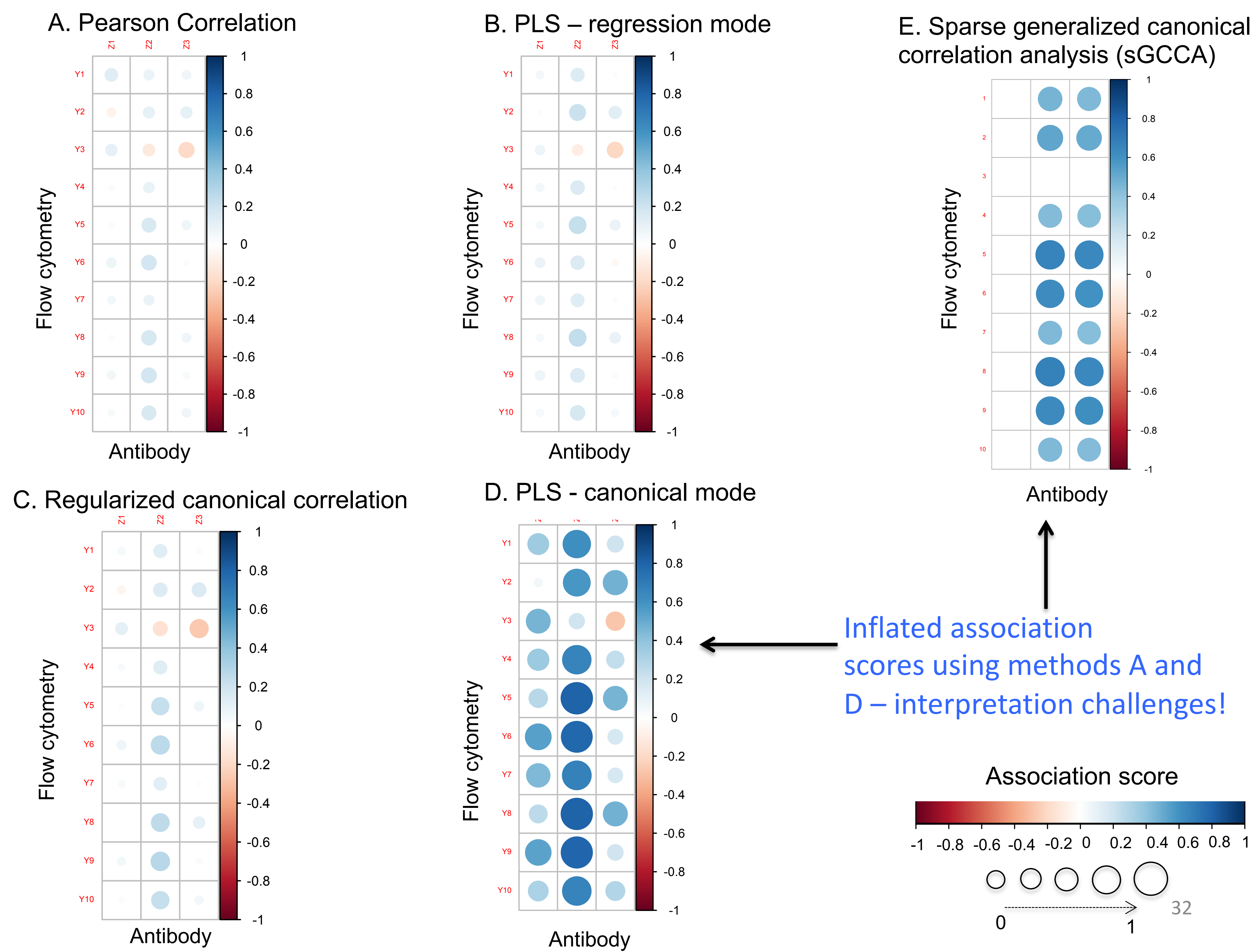
| | 3Omics | MetabNet | xMWAS | mixOmics |
|---|---|---|---|---|
| Pearson correlation | x | x | x | x |
| PLS (regression) | | | x | x |
| PLS (canonical) | | | x | x |
| Regularized canonical correlation analysis (RCC) | | | | X |
| sparse generalized canonical correlation analysis (sGCCA) | | | | 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
Flow cytometry
Antibody
Association score
Z1
Z2
Z3
Y1
Y2
Y3
Y4
Y5
Y6
Y7
Y8
Y9
Y10
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

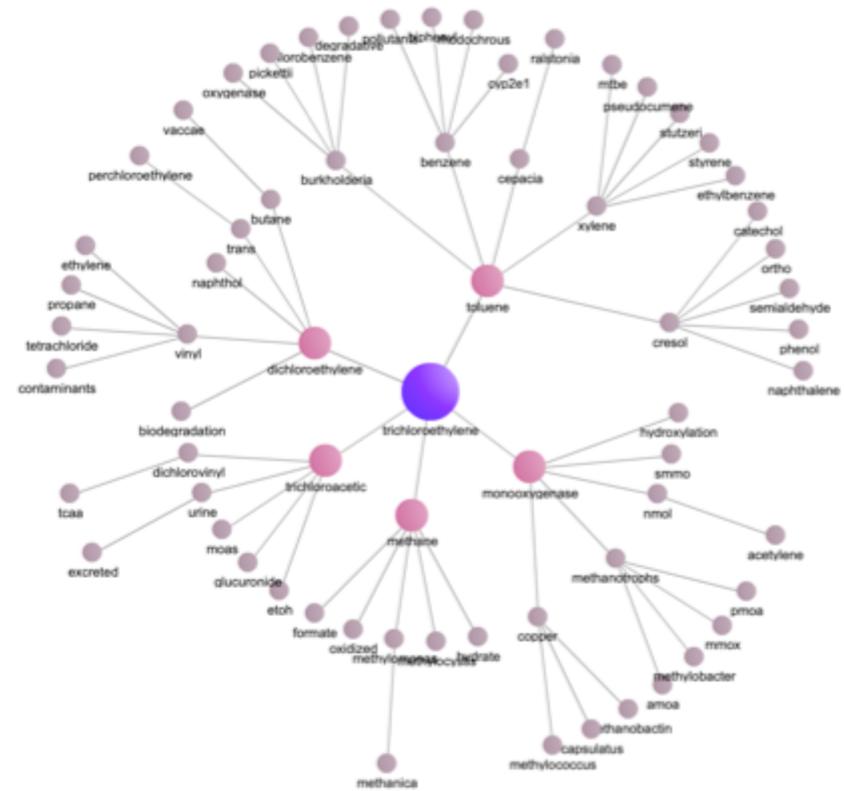

A. Pearson Correlation
B. PLS – regression mode
E. Sparse generalized canonical correlation analysis (sGCCA)
C. Regularized canonical correlation
D. PLS - canonical mode
Flow cytometry
Antibody
Inflated association scores using methods A and D – interpretation challenges!
Association score
-1 -0.8 -0.6 -0.4 -0.2 0 0.2 0.4 0.6 0.8 1
0
1

# 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**

# Text mining tools for literature-based relation discovery biomedical text

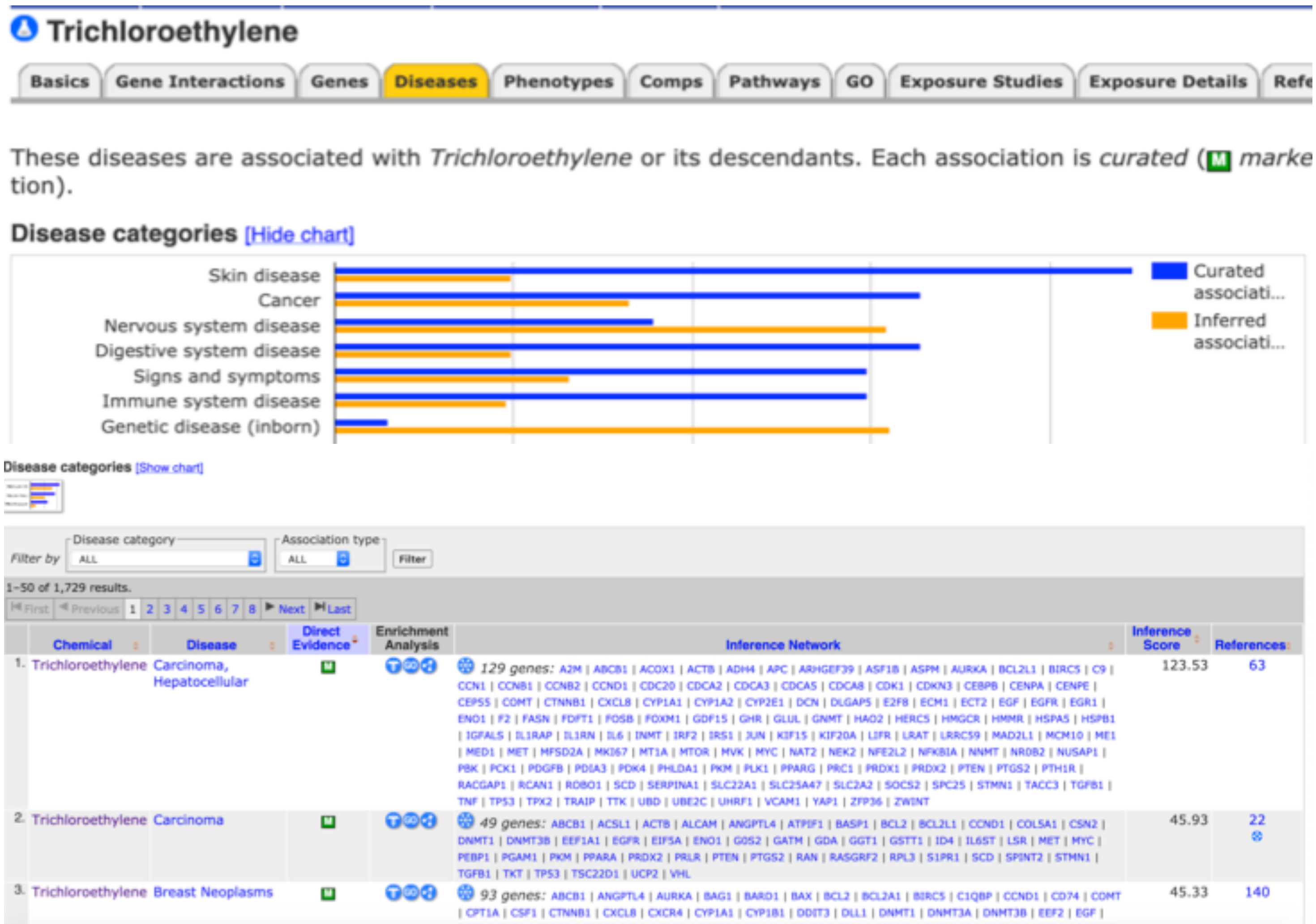

Association mining based on co-occurrence




SEACOIN2.0 Uppal, Lee, et al.

Zhiyong Lu, “PubMed and beyond: a survey of web tools for searching biomedical literature”, Database Vol. 2011

# Comparative Toxicogenomics Database: Chemical x Disease associations



# 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

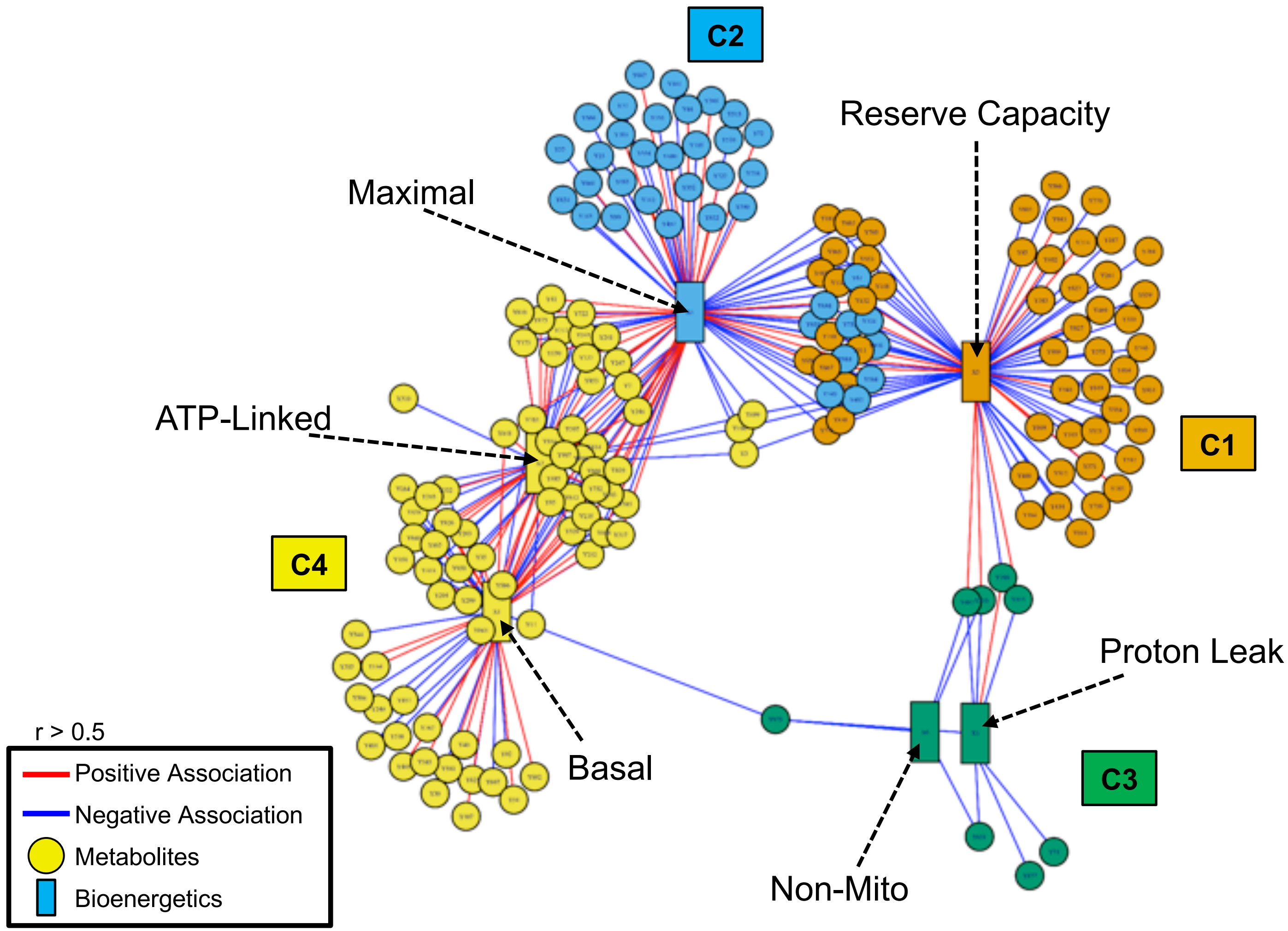
N=13 healthy volunteers

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


C2
Reserve Capacity
Maximal
ATP-Linked
C1
C4
Proton Leak
Basal
C3
Non-Mito
r > 0.5
Positive Association
Negative Association
Metabolites
Bioenergetics

**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

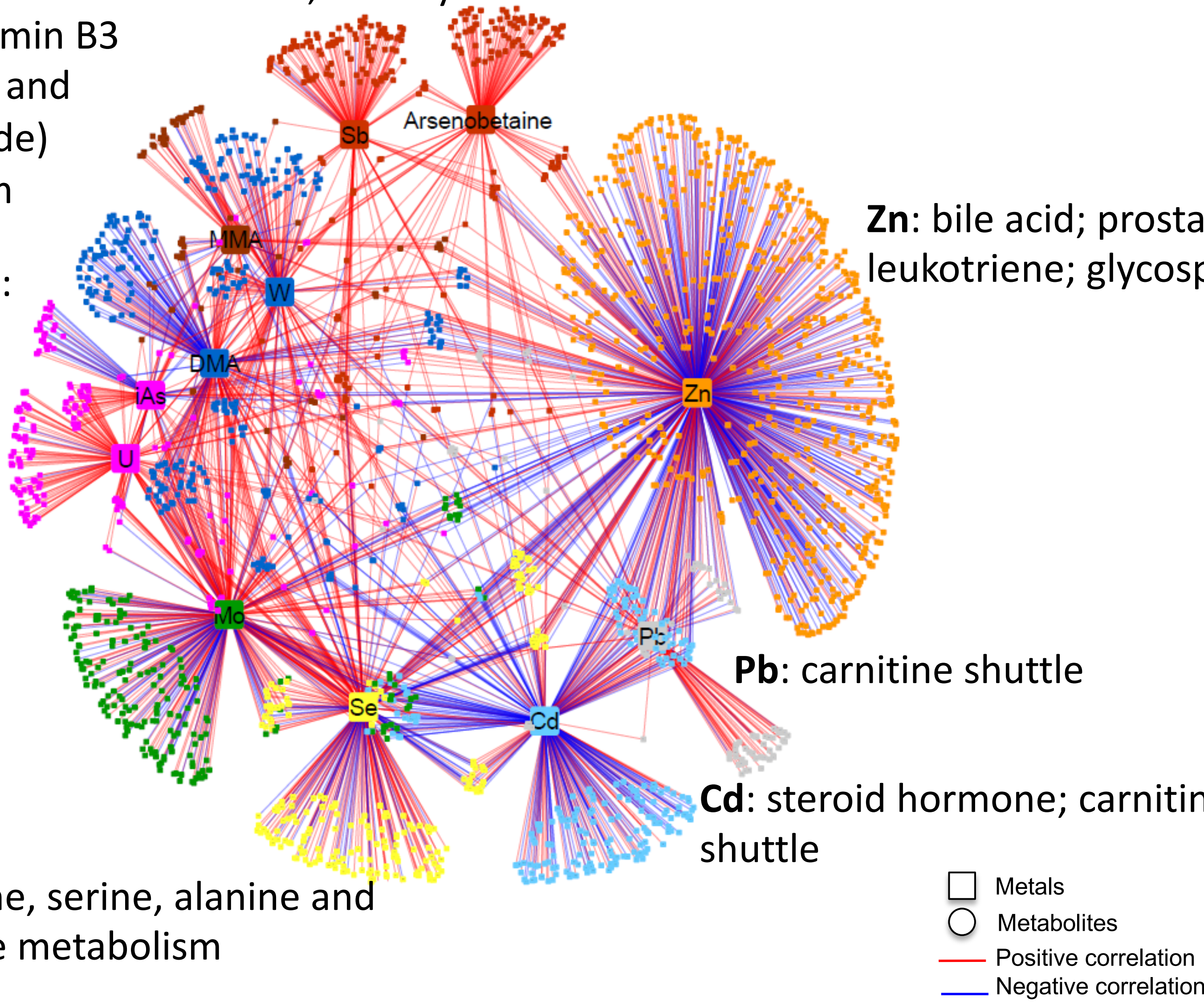
# Metallome-metabolome integrative network using xMWAS

**AsBe and Sb**: Asp and Asn; urea cycle

**MMA**: Vitamin B3 (nicotinate and nicotinamide) metabolism

**Zn**: bile acid; prostaglandin; leukotriene; glycosphingolipid

**DMA and W**: Arg and Pro; linoleate; Ala and Asp




**Mo**: Linoleate metabolism

**Pb**: carnitine shuttle

**Cd**: steroid hormone; carnitine shuttle

**Se**: Glycine, serine, alanine and threonine metabolism

**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 |

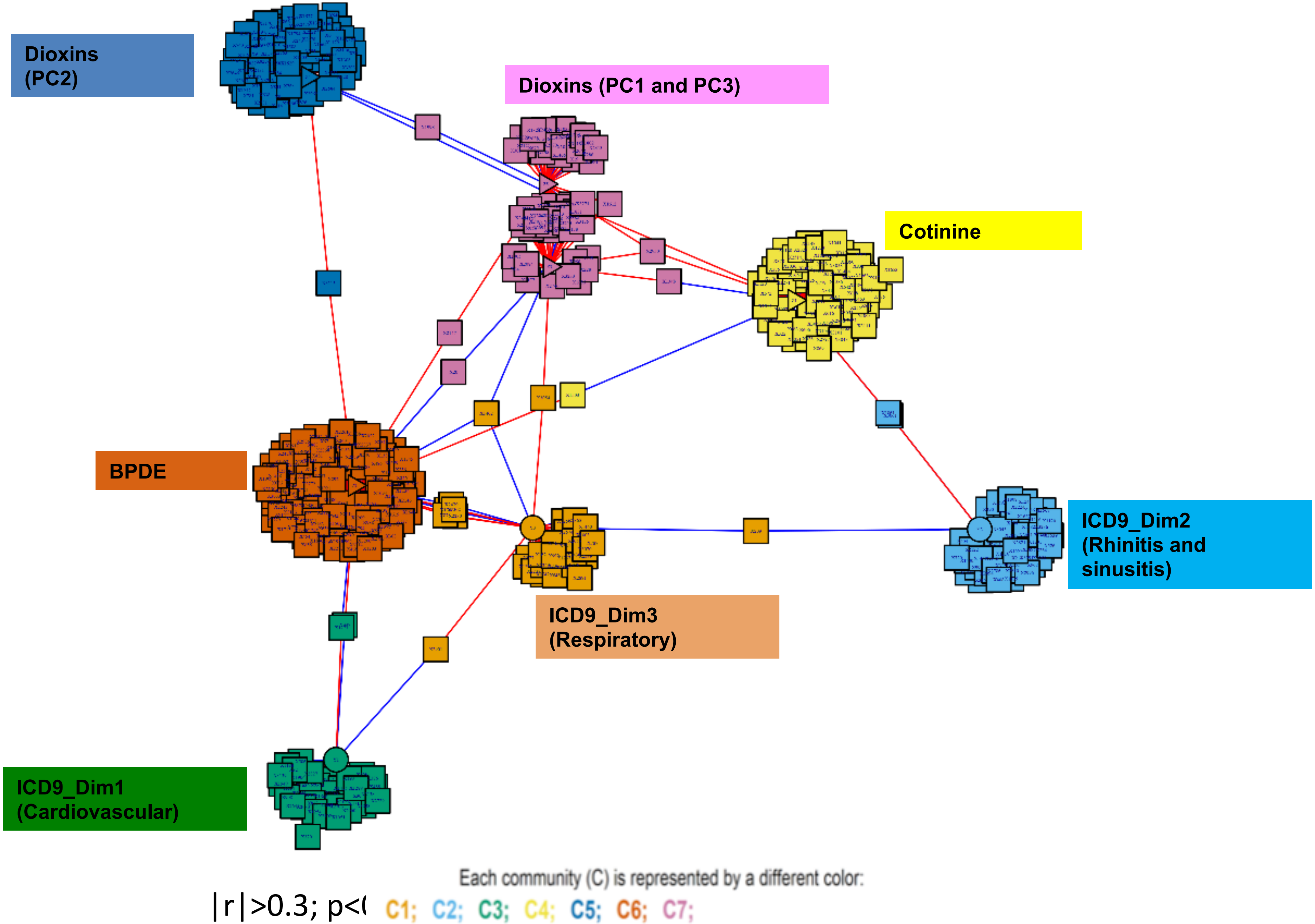
3. **Health outcomes**: ICD-9 codes (49 cardiopulmonary ICD-9 codes x 66 subjects)

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

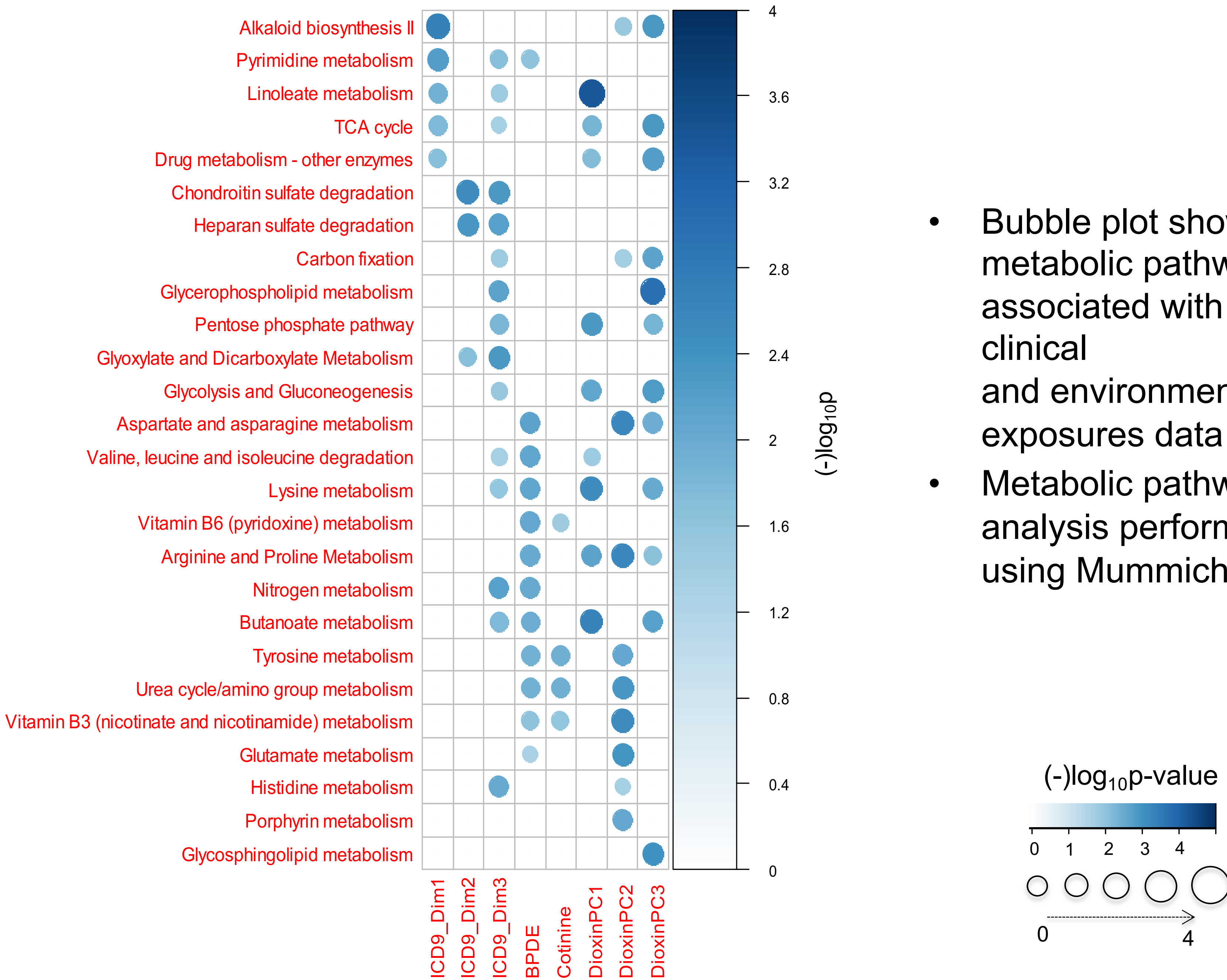
→

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 |


Dioxins (PC2)
Dioxins (PC1 and PC3)
Cotinine
BPDE
ICD9_Dim2 (Rhinitis and sinusitis)
ICD9_Dim3 (Respiratory)
ICD9_Dim1 (Cardiovascular)
Each community (C) is represented by a different color:
C1; C2; C3; C4; C5; C6; C7;

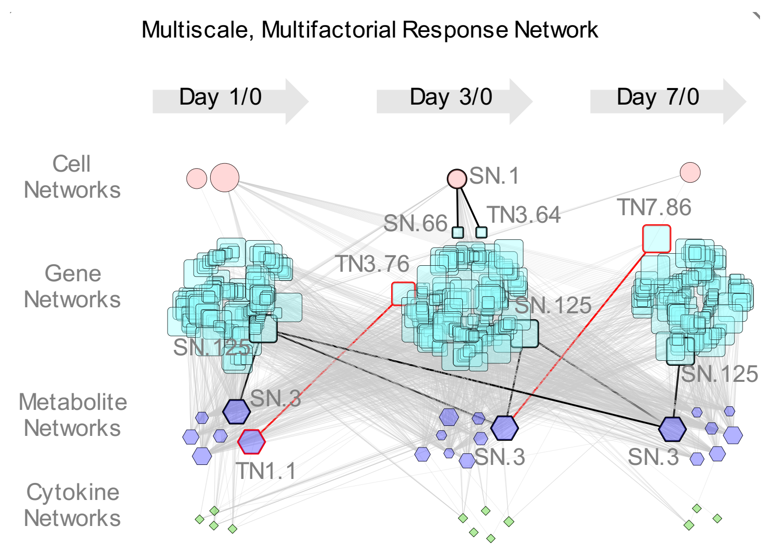
|r|>0.3; p<(




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

# Current challenges and future work

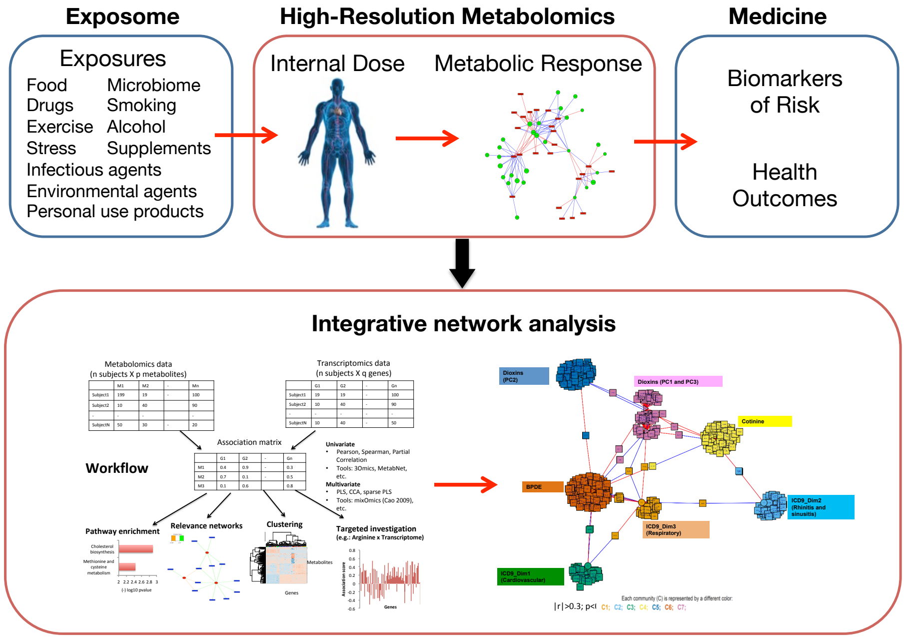
- Development of hybrid methods
  - combined knowledge-based and data-driven approaches
  - incorporation of literature-derived associations in xMWAS
    - Using co-occurrence criteria for establishing relationship (PolySearch2.0)
- Improving scalability
  - Ability to handle >100,000 variables
  - Performing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

EMORY

**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
Website: www.csm2l.com