Computational methods for data integration

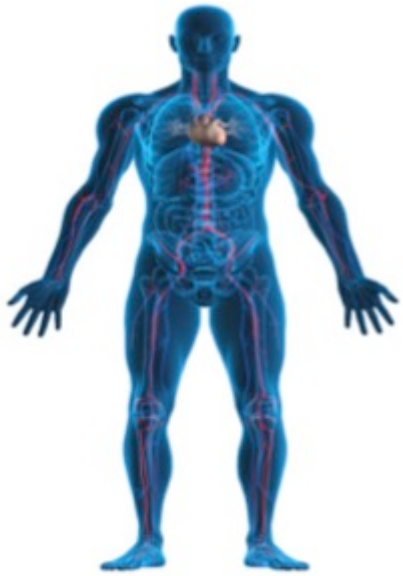
July 21, 2017
Karan Uppal, PhD
Assistant Professor of Medicine
Director of Computational Metabolomics & Integrative Omics,
Clinical Biomarkers Laboratory
Emory University

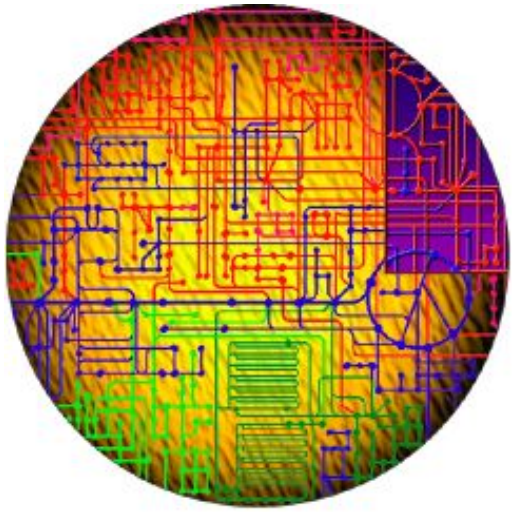
Learning Objectives

- Understanding of different data integration 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

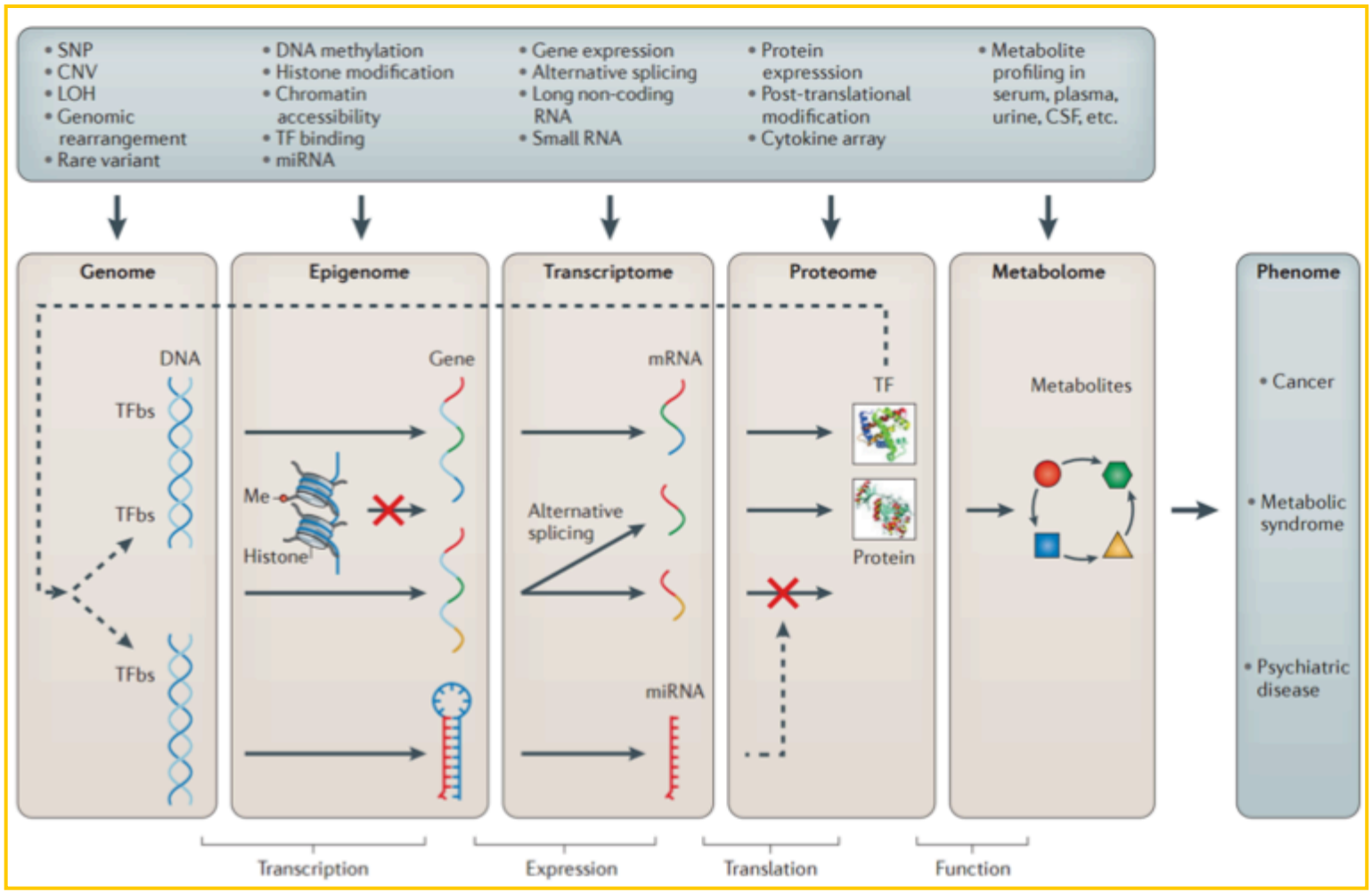




Exposures
Internal measurements
Disease states

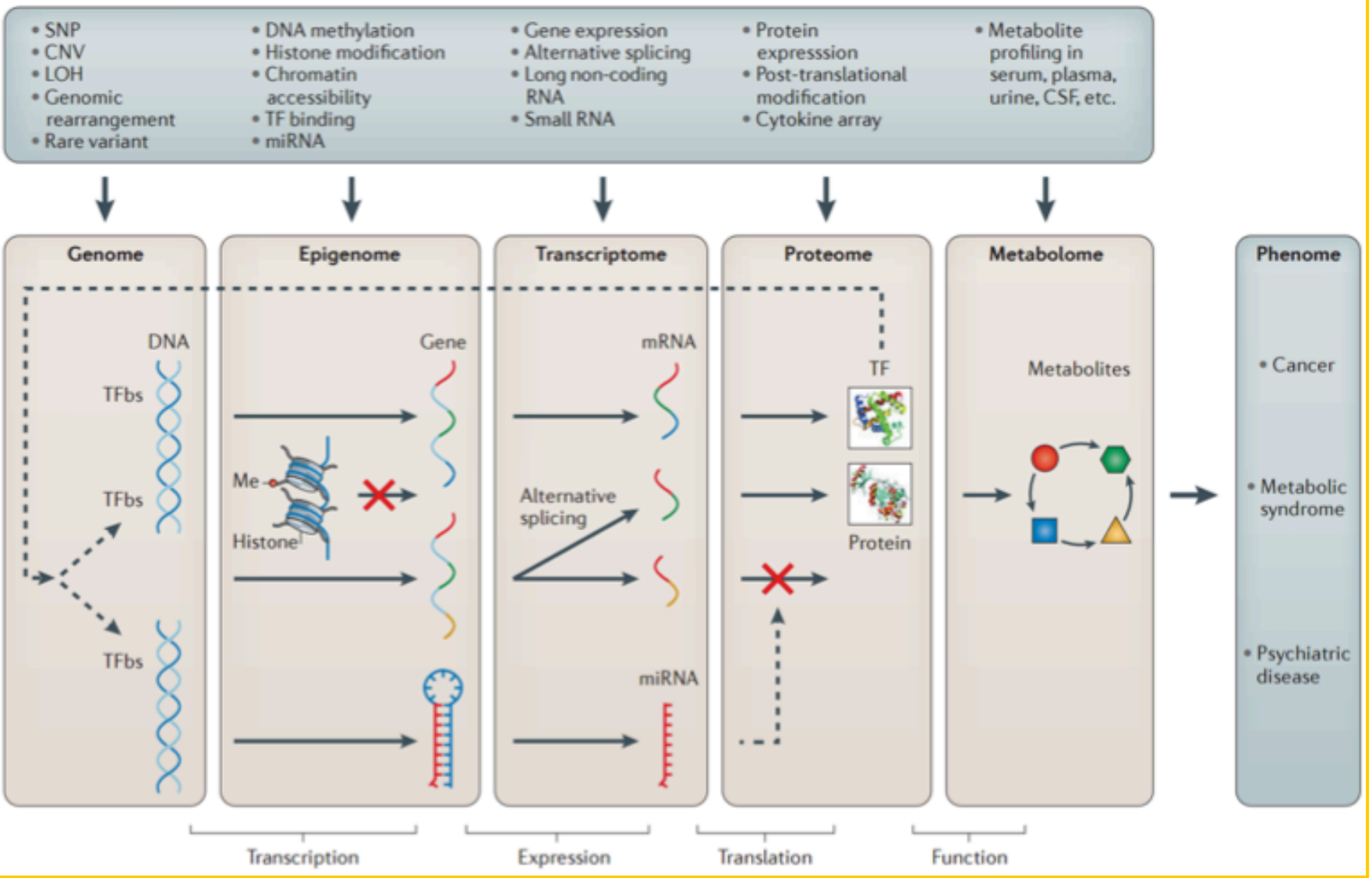
Systems Biology
"*Integrative approach in which scientists study pathways and networks will touch all areas of biology, including drug discovery*"
C. Henry and C. Washington

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.*

Dissecting the Biological system via -omics



"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 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

Paired integrative –omics analysis

- Discover networks of associations or correlated variables (genes, proteins, metabolites, microbiome, epigenetic alterations, clinical variables, etc.) from paired –omics data measured across same samples
  - Univariate or multivariate regression
  - Example: explaining protein abundance with respect to gene expression
- Determine if different –omics data point to same disease mechanism
- Generate novel hypotheses for further investigation

Main approaches for data integration

- Pathway or knowledge-based integration
  - Datasets are analyzed individually (differentially expressed genes, metabolites, proteins) and integration is performed at the pathway level
  - Examples: MetaboAnalyst, iPEAP, MetScape
- Data-driven integration using meta-dimensional analysis
  - Integration is performed globally such that data from multiple omics layers are combined simultaneously
  - Examples: 3Omics, mixOmics, xMWAS
- Using literature-derived associations for integration
  - Using co-occurrence criteria for establishing relationship
  - Examples: CoPub, ArrowSmith, SEACOIN2.0

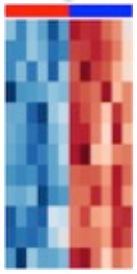
Main approaches for data integration

- **Pathway or knowledge-based integration**
  - **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
  - Examples: 3Omics, mixOmics, xMWAS
- Using literature-derived associations for integration
  - Using co-occurrence criteria for establishing relationship
  - Examples: CoPub, ArrowSmith, SEACOIN2.0

Pathway or knowledge-based integration

Metabolomics data
(n subjects X p metabolites)

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

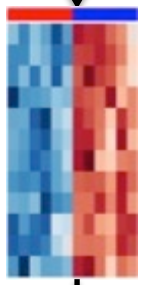


Differentially expressed metabolites

Pathway analysis for metabolites

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

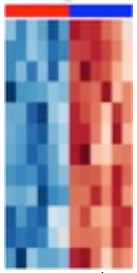
Pathway analysis for genes

Common pathways or pathway rank aggregation

Pathway or knowledge-based integration

Metabolomics data
(n subjects X p metabolites)

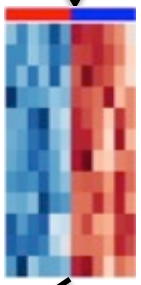
	M1	M2	-	Mp
Subject1	199	19	-	100
Subject2	10	40		90
-	-	-		-
SubjectN	50	30	-	20



Differentially expressed metabolites

Transcriptomics data
(n subjects X q genes)

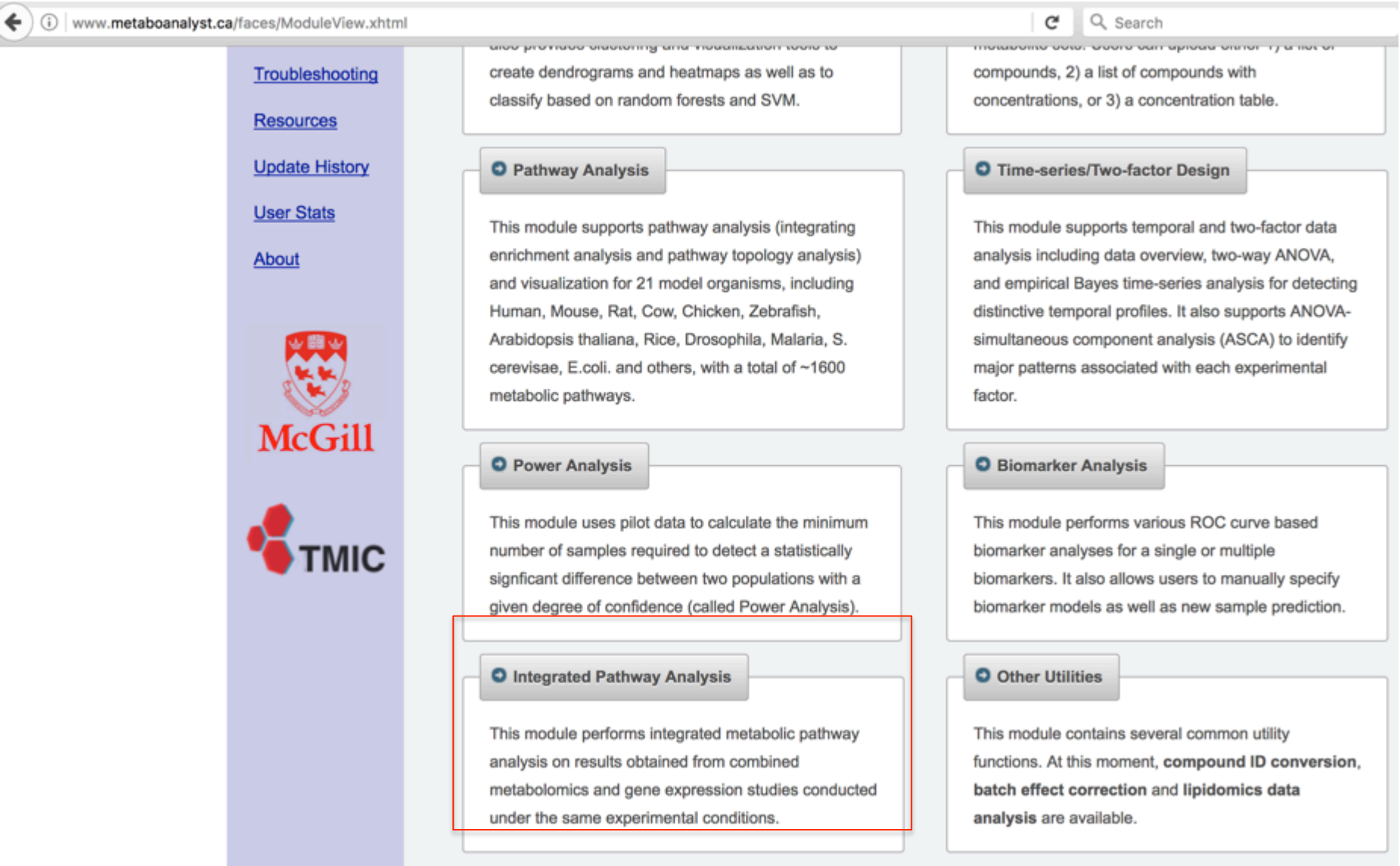
	G1	G2	-	Gq
Subject1	19	19	-	100
Subject2	10	40	-	90
-	-	-	-	-
SubjectN	10	40	-	50



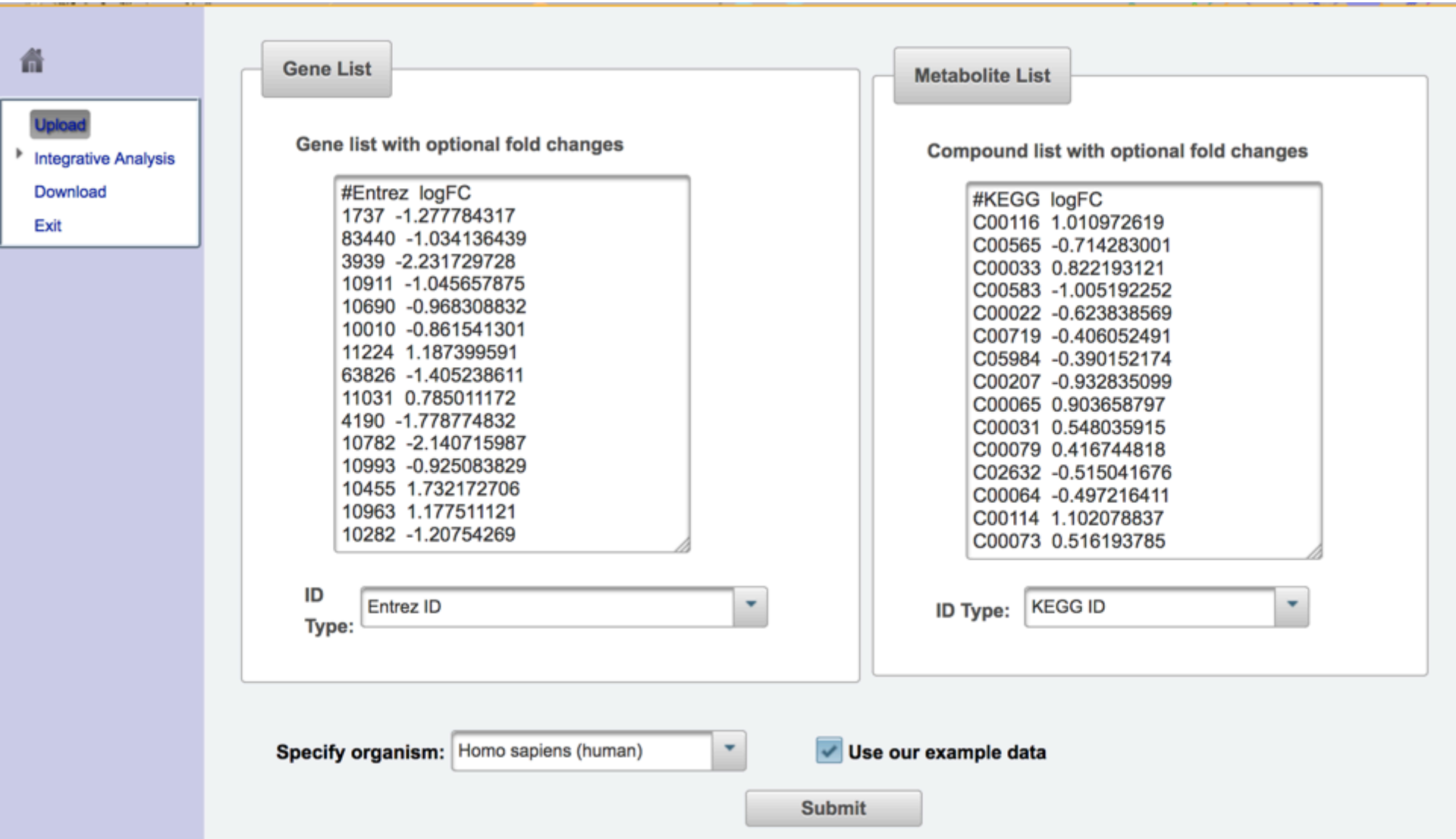
Differentially expressed genes

Pathway analysis using genes and metabolites (joint)

MetaboAnalyst
(http://www.metaboanalyst.ca)



Data upload


Upload
Integrative Analysis
Download
Exit
Gene List
Gene list with optional fold changes
#Entrez logFC
1737 -1.277784317
83440 -1.034136439
3939 -2.231729728
10911 -1.045657875
10690 -0.968308832
10010 -0.861541301
11224 1.187399591
63826 -1.405238611
11031 0.785011172
4190 -1.778774832
10782 -2.140715987
10993 -0.925083829
10455 1.732172706
10963 1.177511121
10282 -1.20754269
ID Type:
Entrez ID
Metabolite List
Compound list with optional fold changes
#KEGG logFC
C00116 1.010972619
C00565 -0.714283001
C00033 0.822193121
C00583 -1.005192252
C00022 -0.623838569
C00719 -0.406052491
C05984 -0.390152174
C00207 -0.932835099
C00065 0.903658797
C00031 0.548035915
C00079 0.416744818
C02632 -0.515041676
C00064 -0.497216411
C00114 1.102078837
C00073 0.516193785
ID Type:
KEGG ID
Specify organism:
Homo sapiens (human)
Use our example data
Submit

Enrichment Analysis

Enrichment analysis aims to evaluate whether the observed genes and metabolites in a particular pathway are significantly enriched (appeatr more than expected by random chance) within the dataset. You can choose over-representation analysis (ORA) based on either hypergenometrics analysis or Fisher's exact method.

- (•) Hypergeometric Test
- ( ) Fisher's Exact Test

Topology Analysis

The topology analysis aims to evaluate whether a given gene or metabolite plays an important role in a biological response based on its **position** within a pathway. Degree Centrality measures the number of links that connect to a node (representing either a gene or metabolite) within a pathway; Closeness Centrality measures the overall distance from a given node to all other nodes in a pathway; Betweenness Centrality measures the number of shortest paths from all nodes to all the others that pass through a given node within a pathway.

- (•) Degree Centrality
- ( ) Betweenness Centrality
- ( ) Closeness Centrality

Pathway Databases

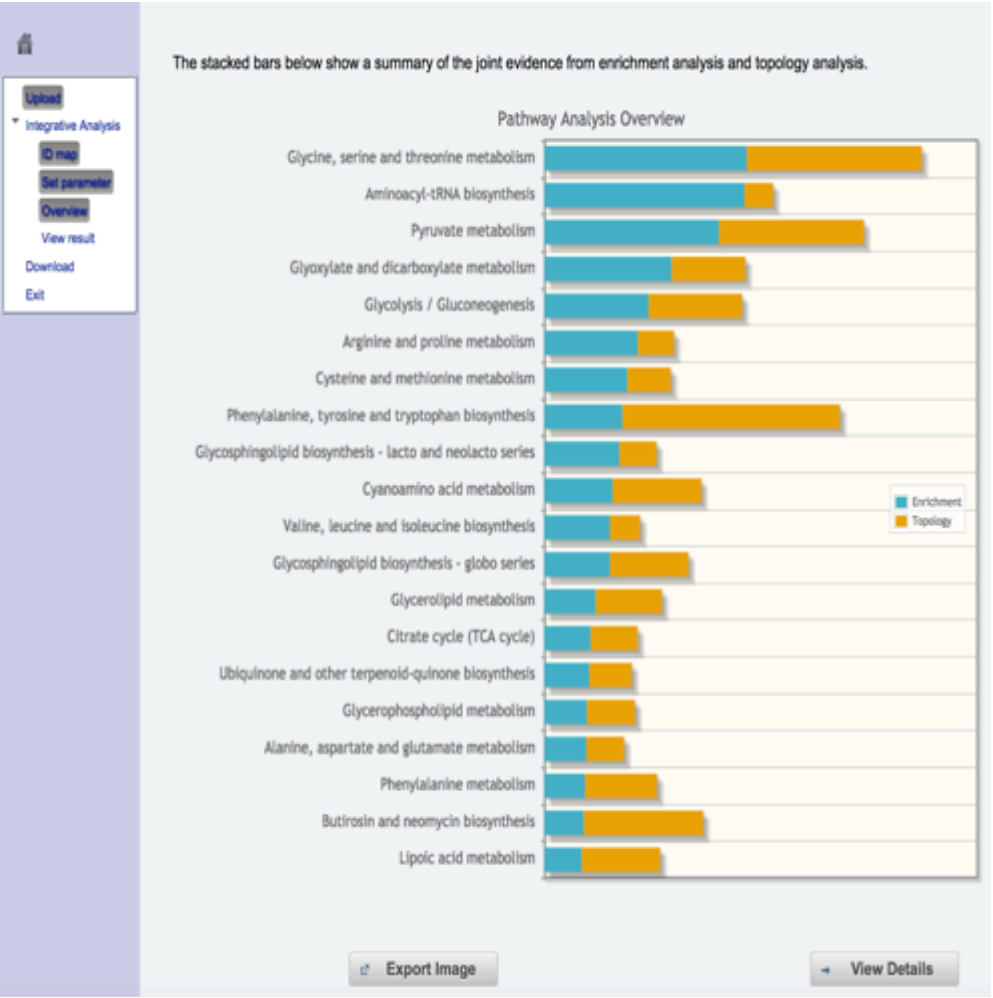
Users can choose one of three different modes of pathways: - the gene-metabolite mode (default) allows joint-analysis and visualization of both significant genes and metabolites; while the gene-centric or metabolite-centric mode allows users to identify enriched pathways driven by significant genes or metabolites, respectively.

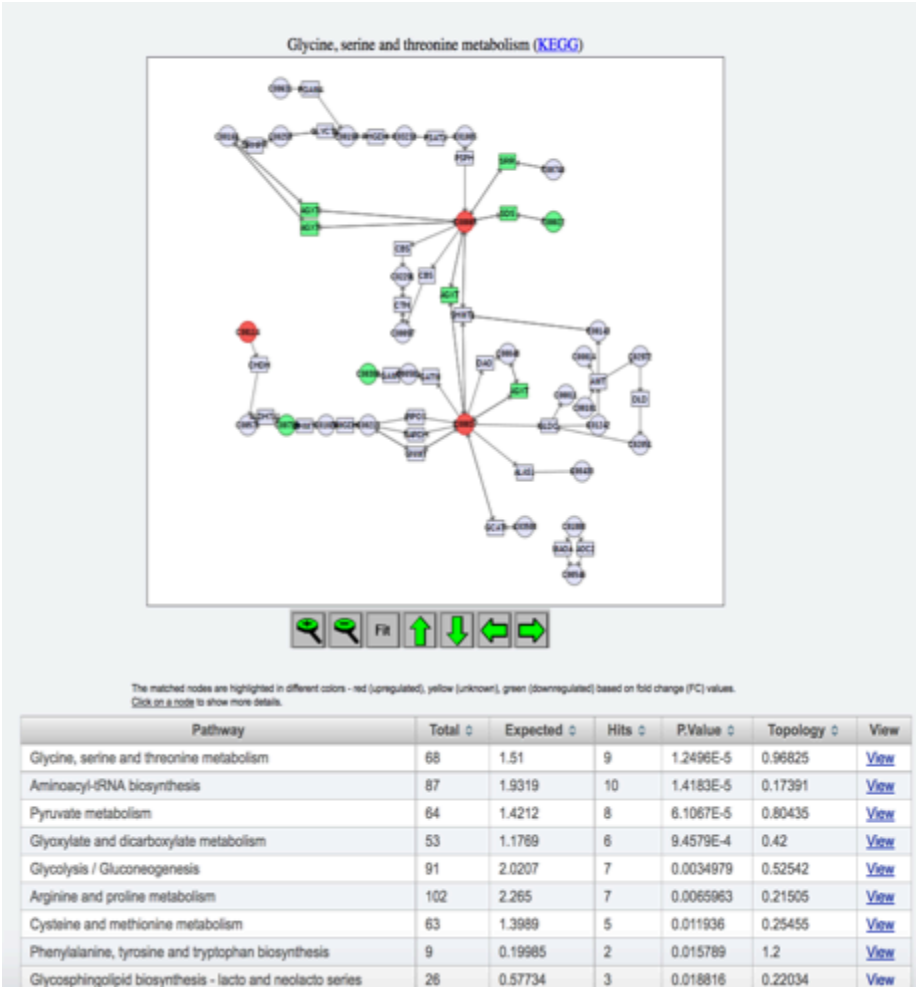
- (•) Gene-metabolite pathways
- ( ) Gene-centric pathways
- ( ) Metabolite-centric pathways

→ Submit

Set Parameters

Results





Pathway	Total	Expected	Hits	P.Value	Topology	View
Glycine, serine and threonine metabolism	68	1.51	9	1.2496E-5	0.96825	View
Aminoacyl-tRNA biosynthesis	87	1.9319	10	1.4183E-5	0.17391	View
Pyruvate metabolism	64	1.4212	8	6.1067E-5	0.80435	View
Glyoxylate and dicarboxylate metabolism	53	1.1769	6	9.4579E-4	0.42	View
Glycolysis / Gluconeogenesis	91	2.0207	7	0.0034979	0.52542	View
Arginine and proline metabolism	102	2.265	7	0.0065963	0.21505	View
Cysteine and methionine metabolism	63	1.3989	5	0.011936	0.25455	View
Phenylalanine, tyrosine and tryptophan biosynthesis	9	0.19985	2	0.015789	1.2	View
Glycosphingolipid biosynthesis - lacto and neolacto series	26	0.57734	3	0.018816	0.22034	View

Main approaches for data integration

- Pathway or knowledge-based integration
  - Datasets are analyzed individually (differentially expressed genes, metabolites, proteins) and integration is performed at the pathway level
  - Examples: MetaboAnalyst, iPEAP, MetScape
- **Data-driven integration using meta-dimensional analysis**
  - **Integration is performed globally such that data from multiple omics layers are combined simultaneously**
  - **Examples: 3Omics, mixOmics, xMWAS**
- Using literature-derived associations for integration
  - Using co-occurrence criteria for establishing relationship
  - Examples: CoPub, ArrowSmith, SEACOIN2.0

Workflow

Metabolomics data
(n subjects X p metabolites)

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

Transcriptomics data
(n subjects X q genes)

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

Association matrix

	G1	G2	-	Gn
M1	0.4	0.9	-	0.3
M2	0.7	0.1	-	0.5
M3	0.1	0.6		0.8

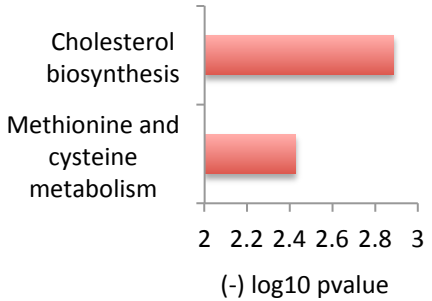
Univariate

- Pearson, Spearman, Partial Correlation
- Tools: 3Omics, MetabNet, etc.

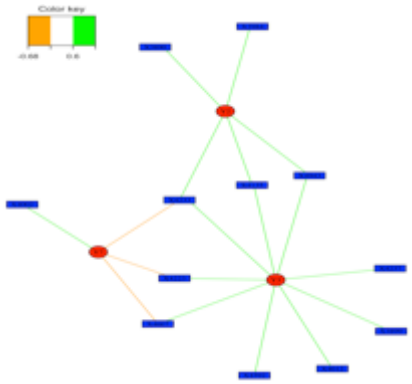
Multivariate

- PLS, CCA, sparse PLS
- Tools: mixOmics (Cao 2009), etc.

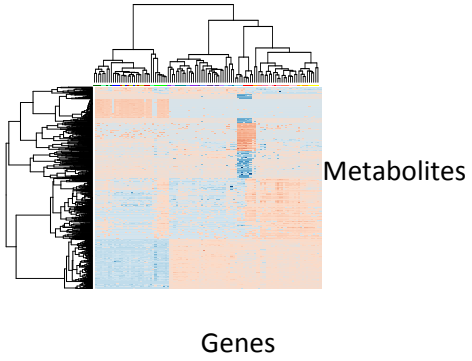
Pathway enrichment



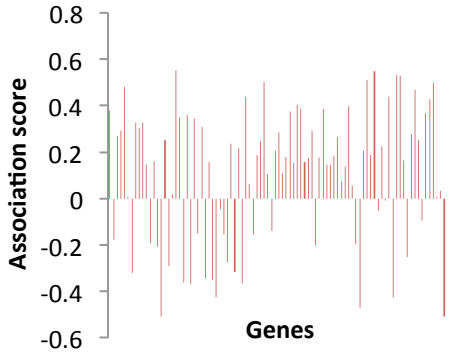
Relevance networks



Clustering

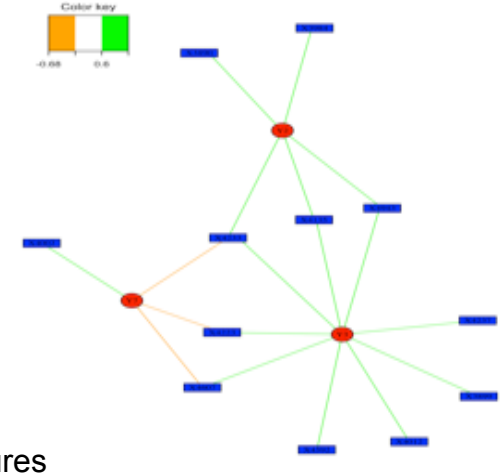


Targeted investigation
(e.g.: Arginine x Transcriptome)



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: microbial species
Rectangles: metabolome features

Methods for generating relevance networks

- Univariate
  - Pairwise Pearson or Spearman correlation between data from different biomedical/clinical technologies (Butte et al. 2000, Uppal et al. 2015)
  - 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 et al. Submitted): R package for data-driven integration and differential network analysis

Univariate methods

3Omics (Kuo et al. BMC Systems Biology 2013)

- A web-based tool for analyzing, integrating and visualizing transcriptomic, proteomic and metabolomic data
- http://3omics.cmdm.tw/

3Omics - homepage

Features

- Correlation analysis and network visualization
  - Pairwise Pearson correlation analysis
- Database-derived relationships in correlation analysis
  - Uses an internal database based on NCBI Entrez gene, Uniprot proteins, and KEGG metabolites to determine gene-protein-metabolite relationship
- Coexpression analysis
  - Two-way hierarchical clustering analysis
  - Rows: variables (Genes + proteins + metabolites, genes+metabolites, etc.)
  - Columns: samples
- Phenotype analysis
  - Uses OMIM databases to link genes with phenotypes
- Pathway and Gene Ontology Enrichment analysis
  - Using KEGG, HumanCyc, and DAVID

Data upload

Please select the desired analysis.

a. Transcriptomics-Proteomics-Metabolomics
b. Transcriptomics-Proteomics
c. Proteomics-Metabolomics
d. Transcriptomics-Metabolomics
e. Transcriptomics only
f. Proteomics only
g. Metabolomics only

Please refer to the help page for more details about each integrating method.

← Back

User may upload three kinds of -omic expression data. All analyses will be performed.

Use example data

Transcriptomics

Browse... No file selected.

GenBank ID: e.g. NAT1, ABL1

Proteomics

Browse... No file selected.

Uniprot Accession: e.g. P31946, P62258

Metabolomics

Browse... No file selected.

Data format

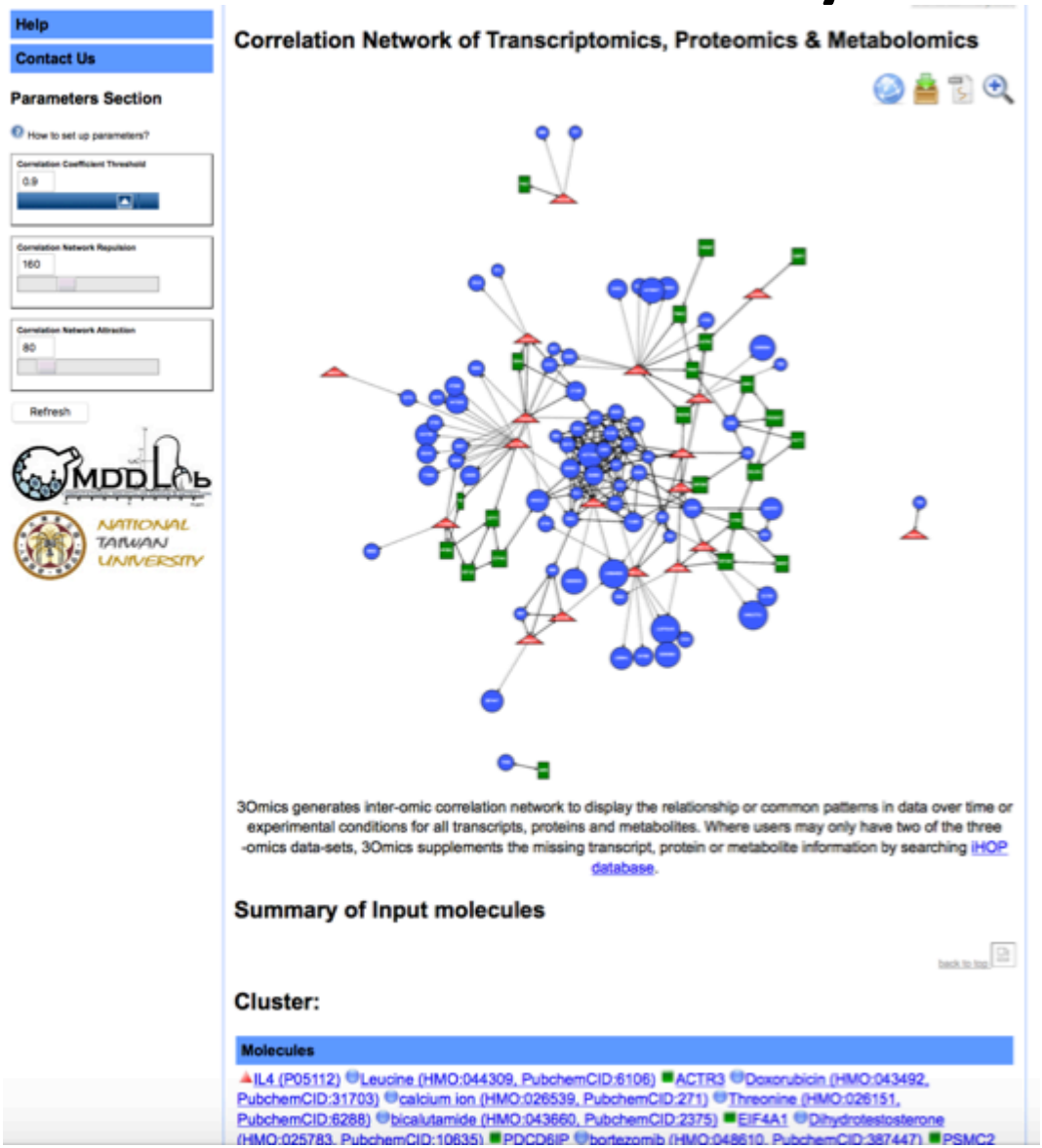
(http://3omics.cmdm.tw/help.php#examples)

Samples

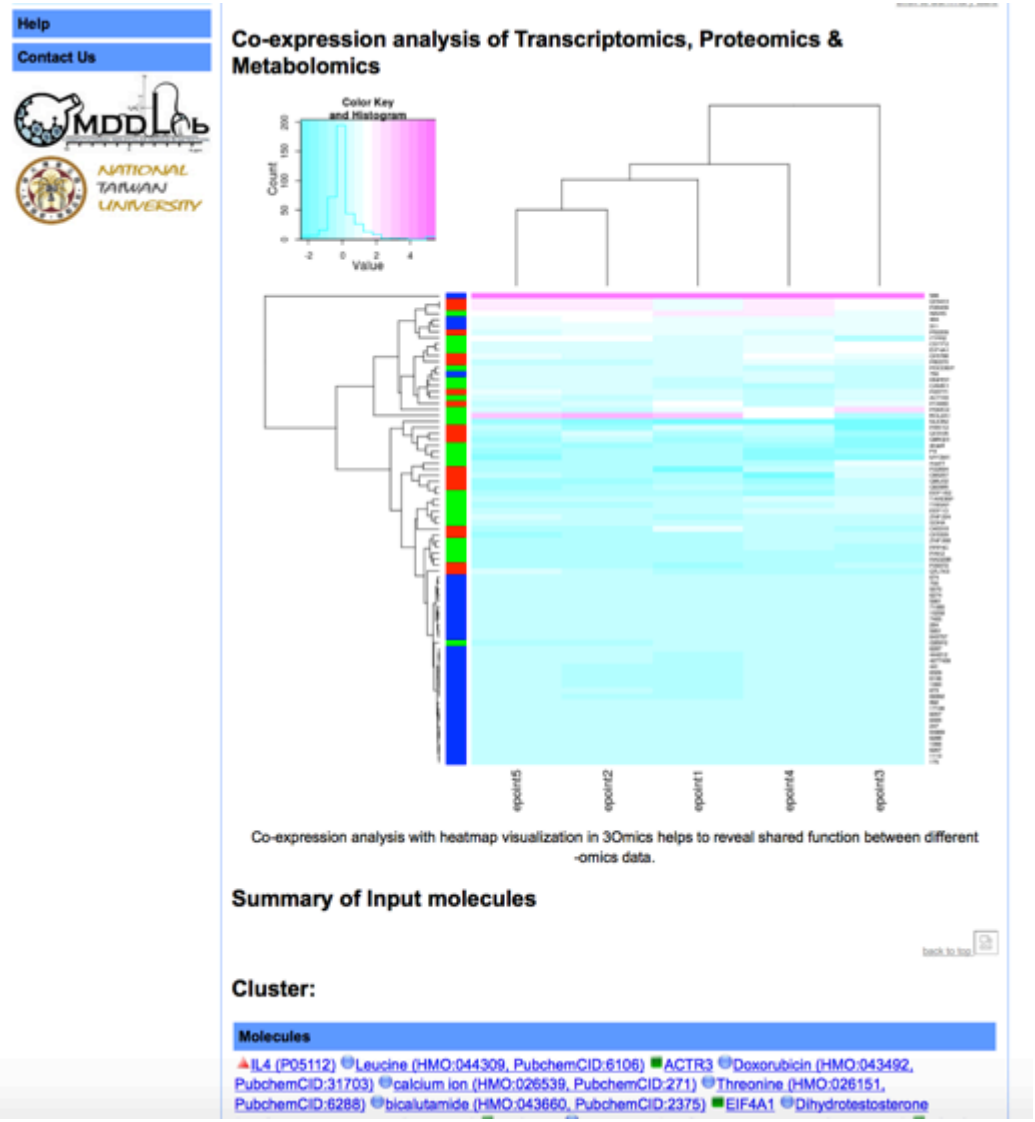
Variables

	timepoint1	timepoint2	timepoint3	timepoint4	timepoint5
akap9	-0.24	-0.6	-0.47	-0.38	-0.31
macf1	-0.3	-0.3	0.48	0.07	-0.36
RNPEP	0.24	0.85	0.15	0.79	0.69
SDHA	0.1	0.37	0.18	0.23	0.33
EEF1B2	-0.04	-0.31	0.06	-0.39	-0.46
EEF1D	0.07	0.29	0.22	0.75	0.47
EIF4A1	0.42	0.65	0.66	0.97	0.78
WARS	1.47	1.72	0.58	1.79	1.69
G3BP2	0.15	0.09	0.1	0.2	-0.22
PAK2	-0.21	-0.14	-0.15	-0.31	-0.4
PPP4C	-0.13	0.05	-0.09	0.21	-0.12
ZNF224	-0.06	0.31	0.17	0.27	0.61
ZNF268	-0.23	0.08	0.01	0.1	-0.1
TRRAP	0.07	-0.12	0.41	0.45	-0.09
RAD23B	-0.07	-0.32	-0.02	-0.02	-0.44
TARDBP	0.23	0.18	0.39	0.63	0.23
CSTF2	0.51	0.65	0.71	1.18	0.89
PSMC2	0.82	0.57	1.15	1.75	0.58
F8	-0.19	-0.02	-0.35	-0.82	-0.81
MYOM1	-0.28	-0.29	-0.54	-1.06	-1.03
ACTR3	0.57	0.48	0.39	0.32	0.72
ITPR2	0.62574	1.771	-0.057392	1.2612	1.7769
NUCB2	-1.1943	-0.96016	-0.71549	-1.1877	-0.70604
CAMK1	0.33342	0.87499	0.059355	0.062122	0.53605
BCL2A1	2.2913	3.8479	-0.12343	1.6604	3.3933
PDCD6IP	0.46362	0.88049	0.20539	0.36177	0.62012

Correlation analysis



Co-expression analysis



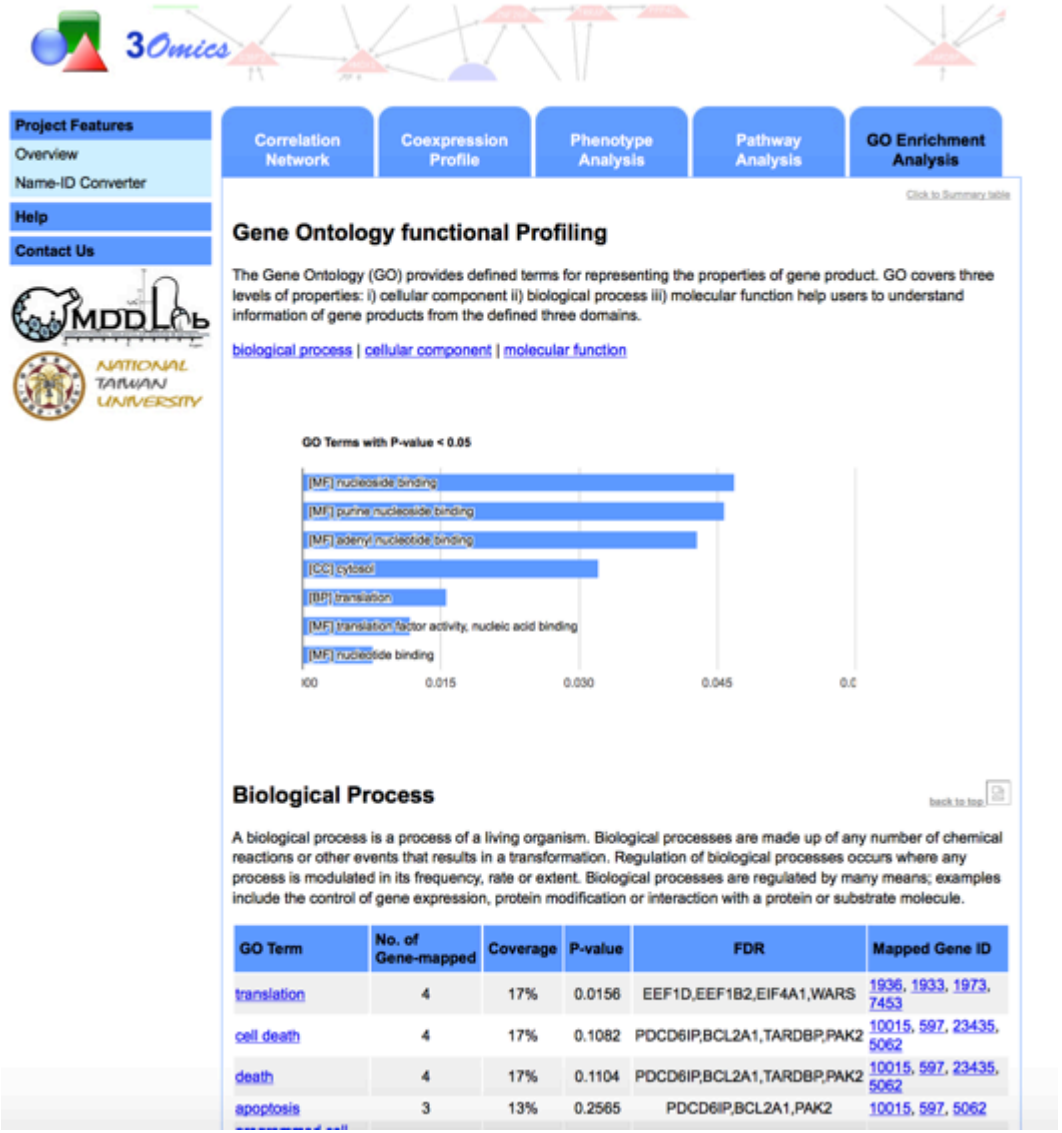
Rows: Variables

Columns: Samples

Phenotype analysis

Pathway analysis

GO Enrichment Analysis



Multivariate methods

Generating relevance network using sPLS or msPLS techniques (Cao 2009, Liquet 2012)

- sparse partial least squares (sPLS) regression or multilevel partial least squares (msPLS) method
- One-step procedure for variable selection as well as integration
- Comparison of different multivariate integration techniques showed that sPLS generates (Cao 2009)
- msPLS – for repeated measures
- Implemented in the R package mixOmics
- Generates association matrix and allows visualization of associations using bipartite relevance networks (Liquet 2012)

sPLS method

- 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
- 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 $cov(X_u, Y_v)$

  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

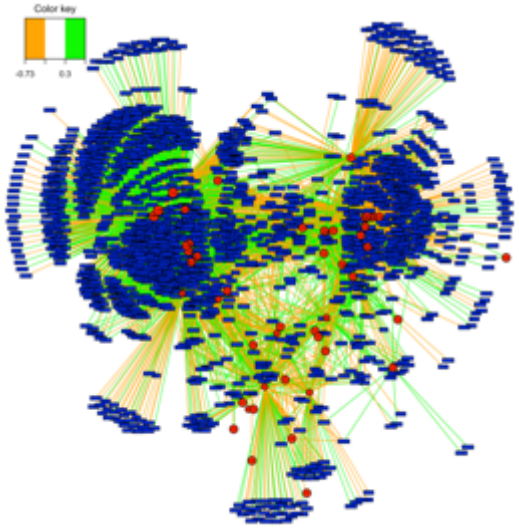
  A Lasso based optimization is used to select most relevant variables

Case Study 2: Application of sPLS technique for integrative –omics. Microbiome-Metabolome Wide Association Study of Lung BAL: Global integration of 5930 m/z features with 153 microbial species using sparse Partial Least Squares regression (Cribbs et al. Microbiome 2016)

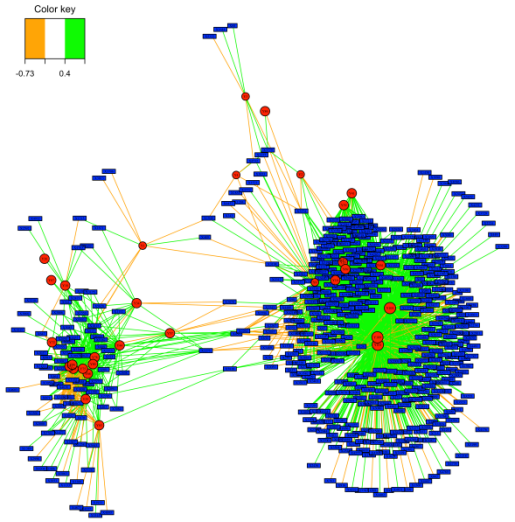
Legend

Circles: microbial species

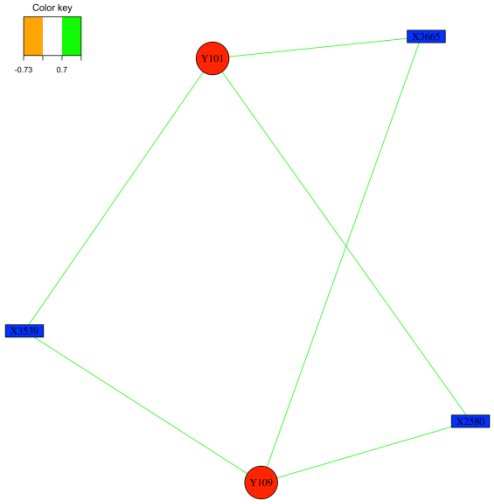
Rectangles: metabolome features



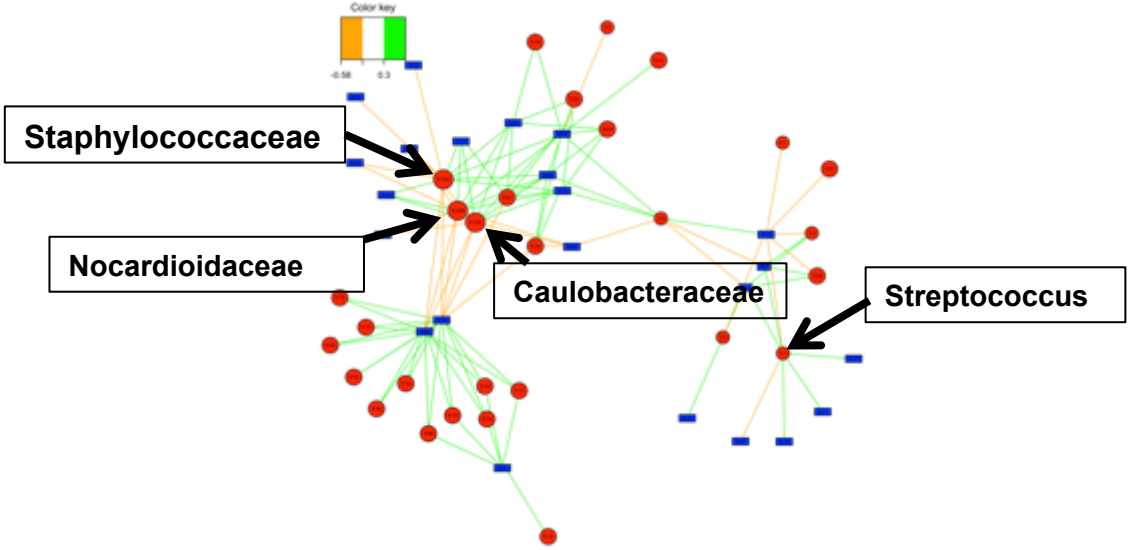
A. Association threshold: 0.3



B. Association threshold: 0.4



C. Association threshold: 0.7



D. Using only subset of metabolic features also associated with HIV status (+ve or –ve)

Integrating more than two datasets

Proteomics data
(n subjects X s proteins)

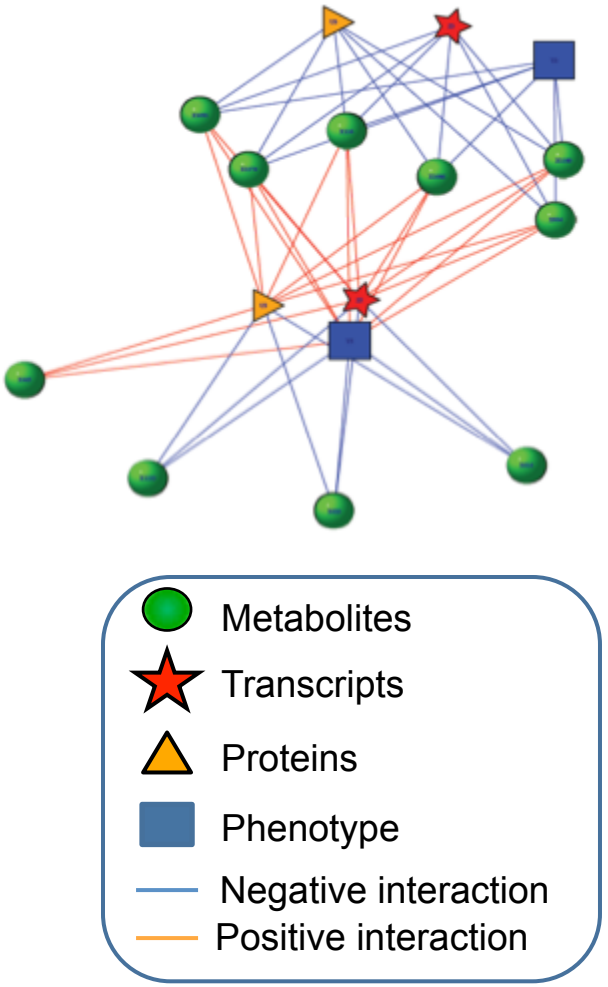
	E1	E2	-	Es
Subject1	199	19	-	100
Subject2	10	40		90
-	-	-		-
SubjectN	50	30	-	20

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



xMWAS: R package for data integration and differential network analysis

(Uppal et al. Submitted to Bioinformatics)

URL: https://sourceforge.net/projects/xmwas/

A.

Data matrices from multiple assays

↓

Integrative and association analysis

Pairwise integrative and network analysis (e.g. X<->Y; X<->Z, Y<->Z) using (sparse) Partial Least Squares regression

↓

Generate an edge list matrix, L_e, with the list of edges that meet the significance criteria and association score threshold

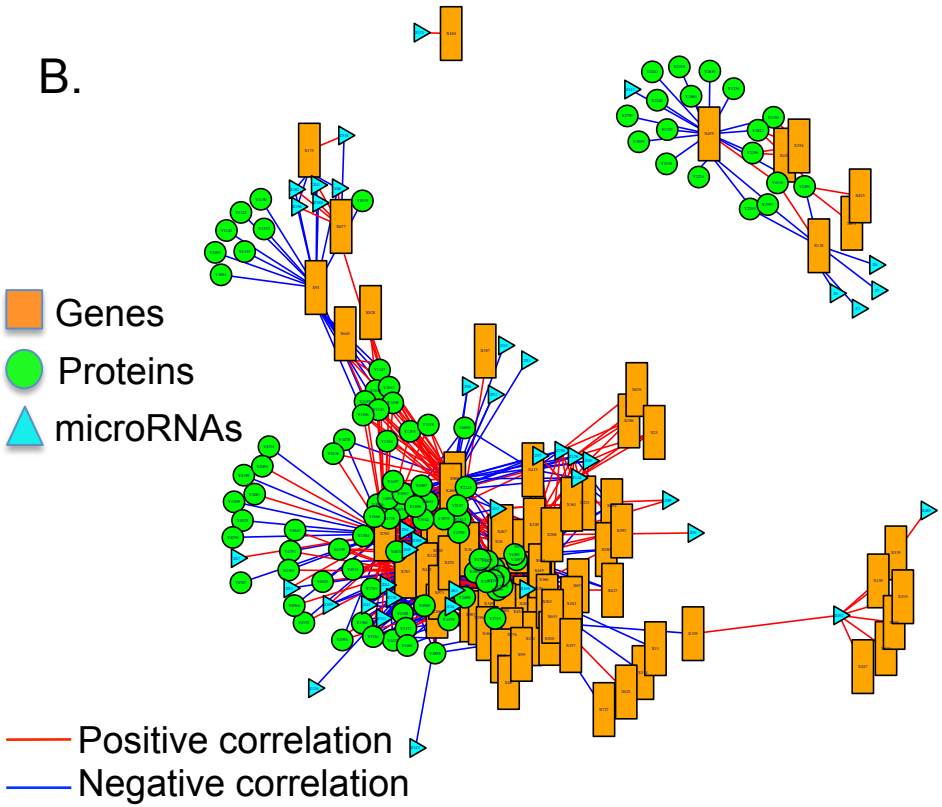
↓

Generate a k-partite graph, G, using the union of edge list matrices

↓

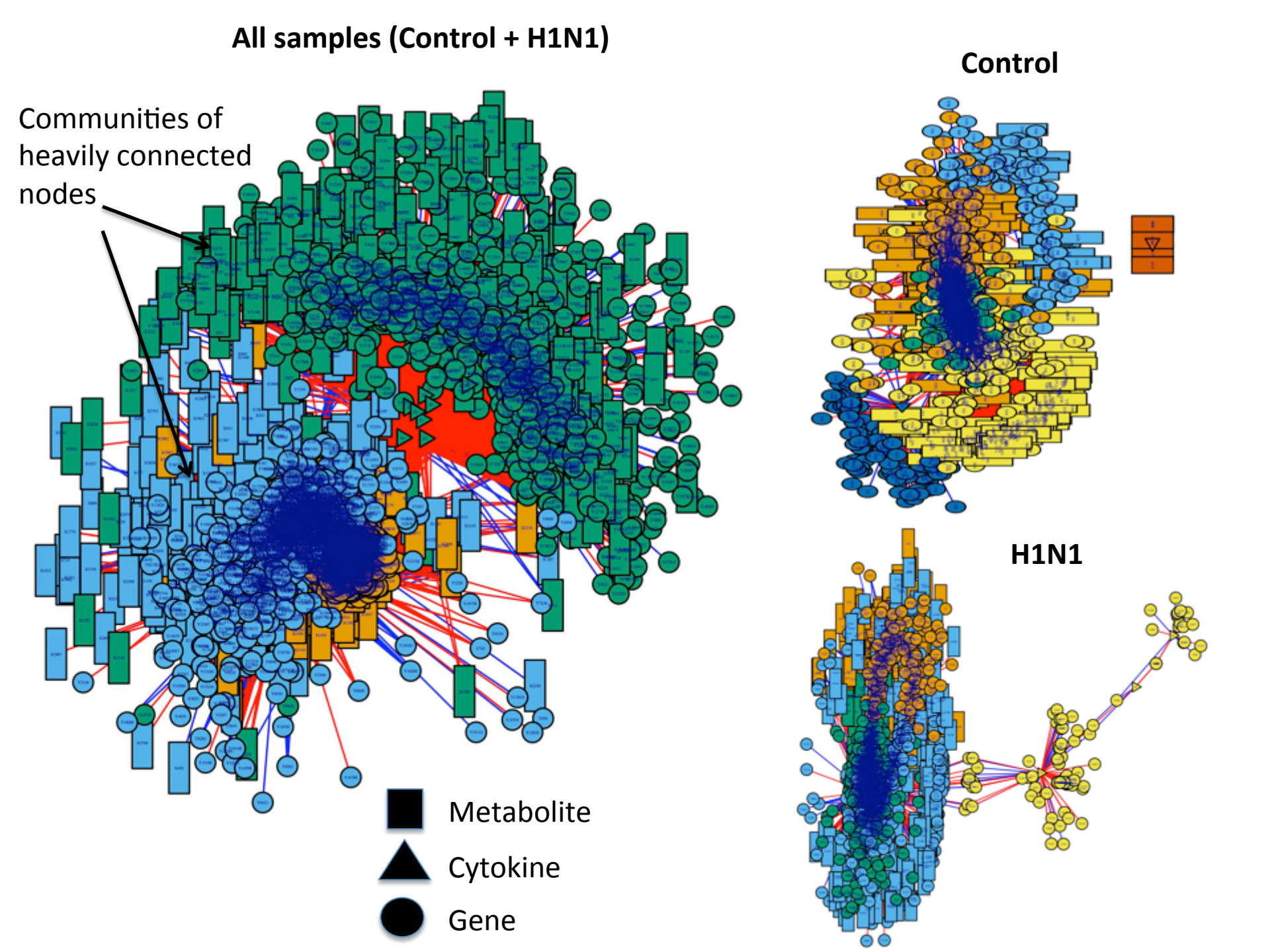
Community detection, differential centrality and rewiring analysis, and network visualization

B.



Case Study 3: Application of xMWAS for integrative –and differential network analysis of more than 2 dataests.

Integrative network analysis of cytokine, metabolome and transcriptome datasets from a study of H1N1 virus infection of mice (Chandler et al. 2016)


All samples (Control + H1N1)
Communities of heavily connected nodes
Control
H1N1
Metabolite
Cytokine
Gene

Pathway Name	IL-1beta	IL-6	IL-10	TNFalpha	IP-1	IFNgamma
Immune response_Alternative complement pathway	x	x	x	x	x	x
GTP metabolism	x	x	x	x	x	
Cytoskeleton remodeling_TGF, WNT and cytoskeletal remodeling	x	x		x	x	x
Cytoskeleton remodeling_Cytoskeleton remodeling	x	x		x	x	x
Alternative complement cascade disruption in age-related macular degeneration	x	x		x	x	x
ATP metabolism	x	x		x	x	
Immune response_Lectin induced complement pathway		x	x		x	x
Regulation of lipid metabolism_Regulation of lipid metabolism by niacin and isoprenaline	x	x		x		
Development_Oligodendrocyte differentiation from adult stem cells		x			x	x
Cell adhesion_Chemokines and adhesion	x			x		
Expression targets of Tissue factor signaling in cancer	x			x		
Cell adhesion_ECM remodeling		x				x
Immune response_Classical complement pathway			x			x

Pathway analysis of genes that were found to be significantly associated with the six cytokines at $p<0.05$ using MetaCore

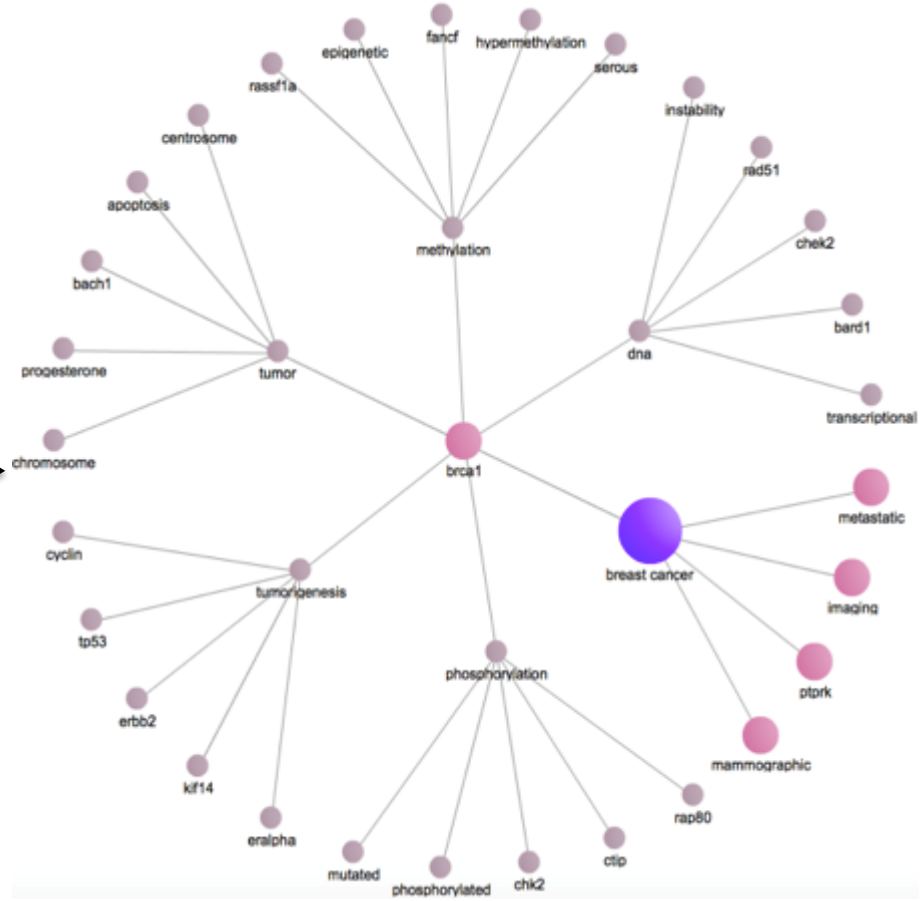
Pathway Name	IL-1beta	IL-6	IL-10	TNFalpha	IP-1	IFNgamma
Carnitine shuttle	x	x	x	x	x	
Glycosphingolipid metabolism	x	x	x	x	x	
Vitamin D_3 (cholecalciferol) metabolism	x	x	x	x	x	
Vitamin B_6 (pyridoxine) metabolism		x		x	x	x
Biopterin metabolism	x	x		x	x	
Saturated fatty acids beta-oxidation		x		x	x	
Vitamin E metabolism	x		x	x		
Nitrogen metabolism		x			x	
Urea cycle/amino group metabolism					x	x
Drug metabolism - cytochrome P450					x	x
Tryptophan metabolism					x	x
Arginine and proline metabolism						x
Bile acid biosynthesis				x		
Fatty acid metabolism			x			
Limonene and pinene degradation			x			
Linoleate metabolism			x			
Tyrosine metabolism		x				

Pathway analysis of metabolic features that were found to be significantly associated with the six cytokines at $p<0.05$ using Mummichog

Main approaches for data integration

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

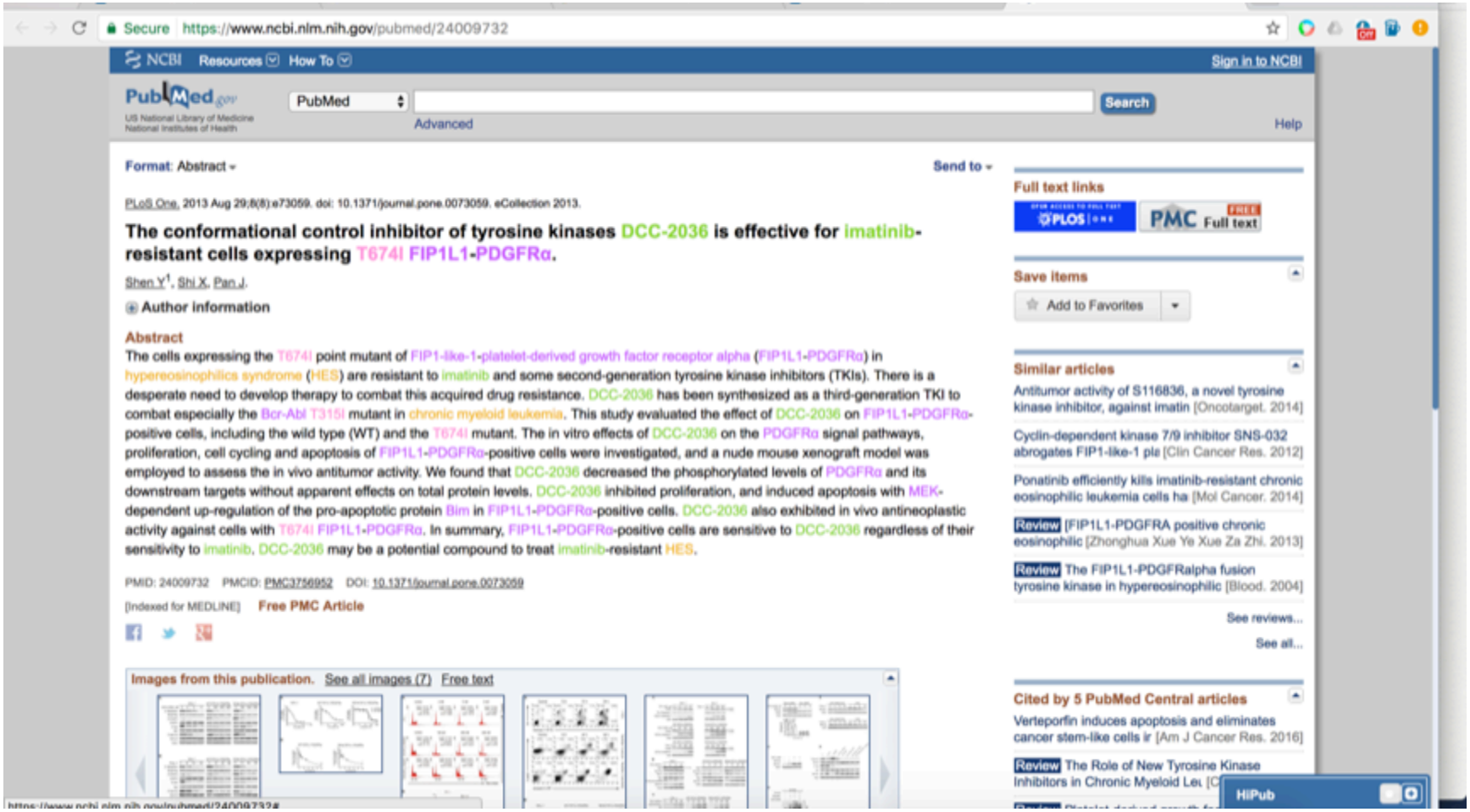
Text mining tools for literature-based relation discovery biomedical text



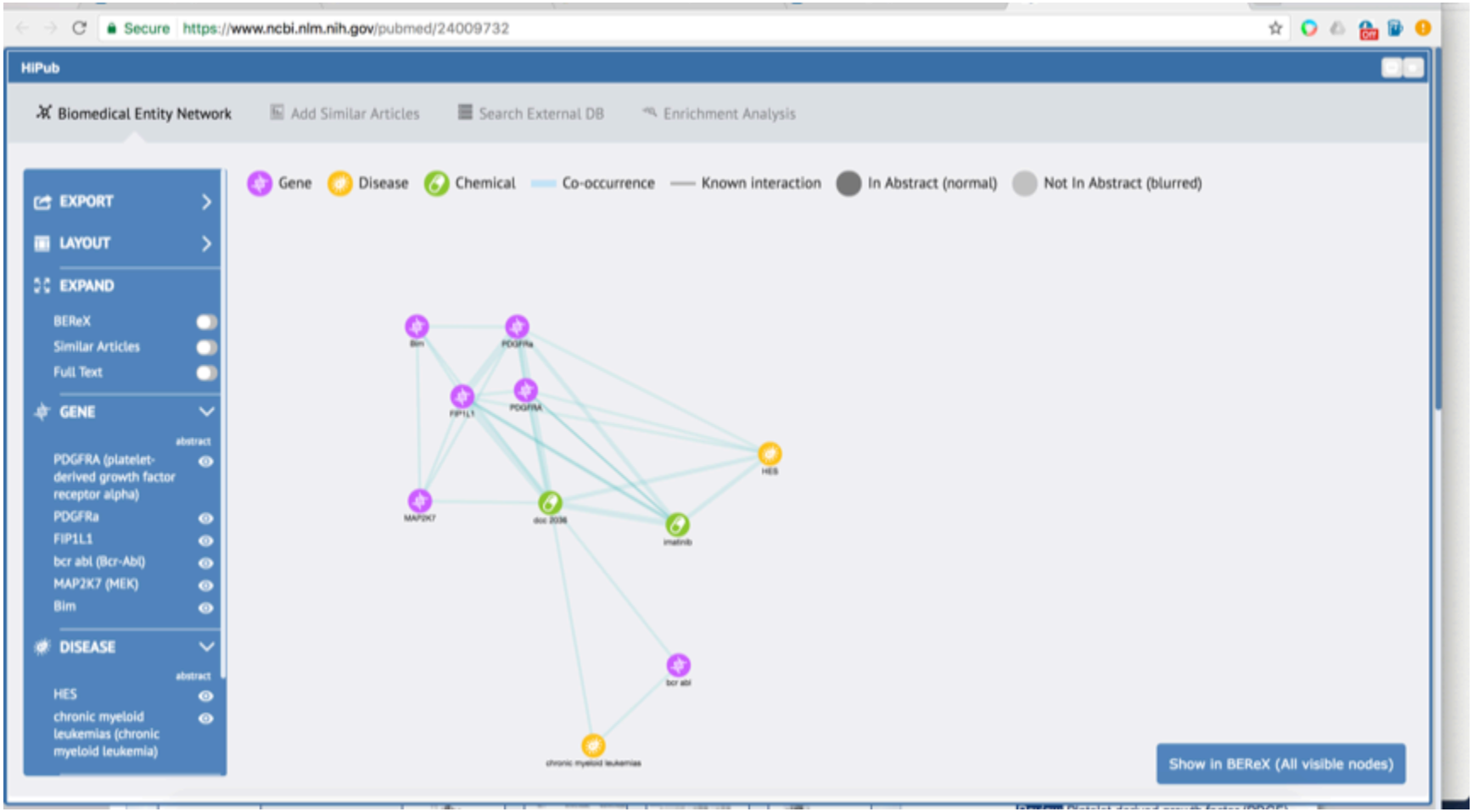
Zhiyong Lu, "PubMed and beyond: a survey of web tools for searching biomedical literature", Database Vol. 2011

Association mining based on co-occurrence

HiPub (Lee 2016): http://hipub.korea.ac.kr/



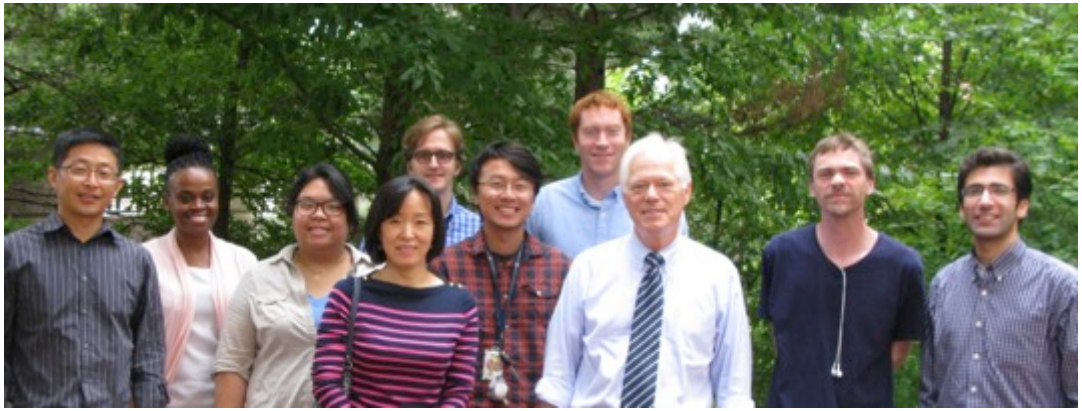
HiPub (Lee 2016):



Summary

- Various tools and techniques are available for integrating and visualization multi –omics data
- Integrative –omics drives systems biology and could play a critical role in personalized medicine

Clinical Biomarkers Laboratory



Dean Jones, Young-Mi Go, Shuzaho Li, Karan Uppal, Douglas Walker, Josh Chandler, Sophia Banton, Ken Liu, Vilinh Tran, Michael Orr, Bill Liang (not shown)

Lab website: http://clinicalmetabolomics.org/

Funding: ES025632, ES023485, ES019776, OD018006, HL095479, EY022618, HHSN272201200031C

Email: kuppal2@emory.edu

Questions?