

Computational methods for integrative omics analysis

July 21, 2016

Karan Uppal, PhD

Assistant Professor

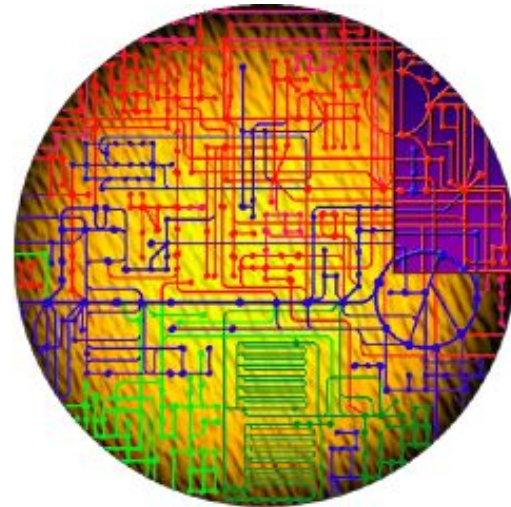
Emory University School of Medicine

Learning Objectives

- Data-driven methods for integrating paired – omics data and visualizing associations

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



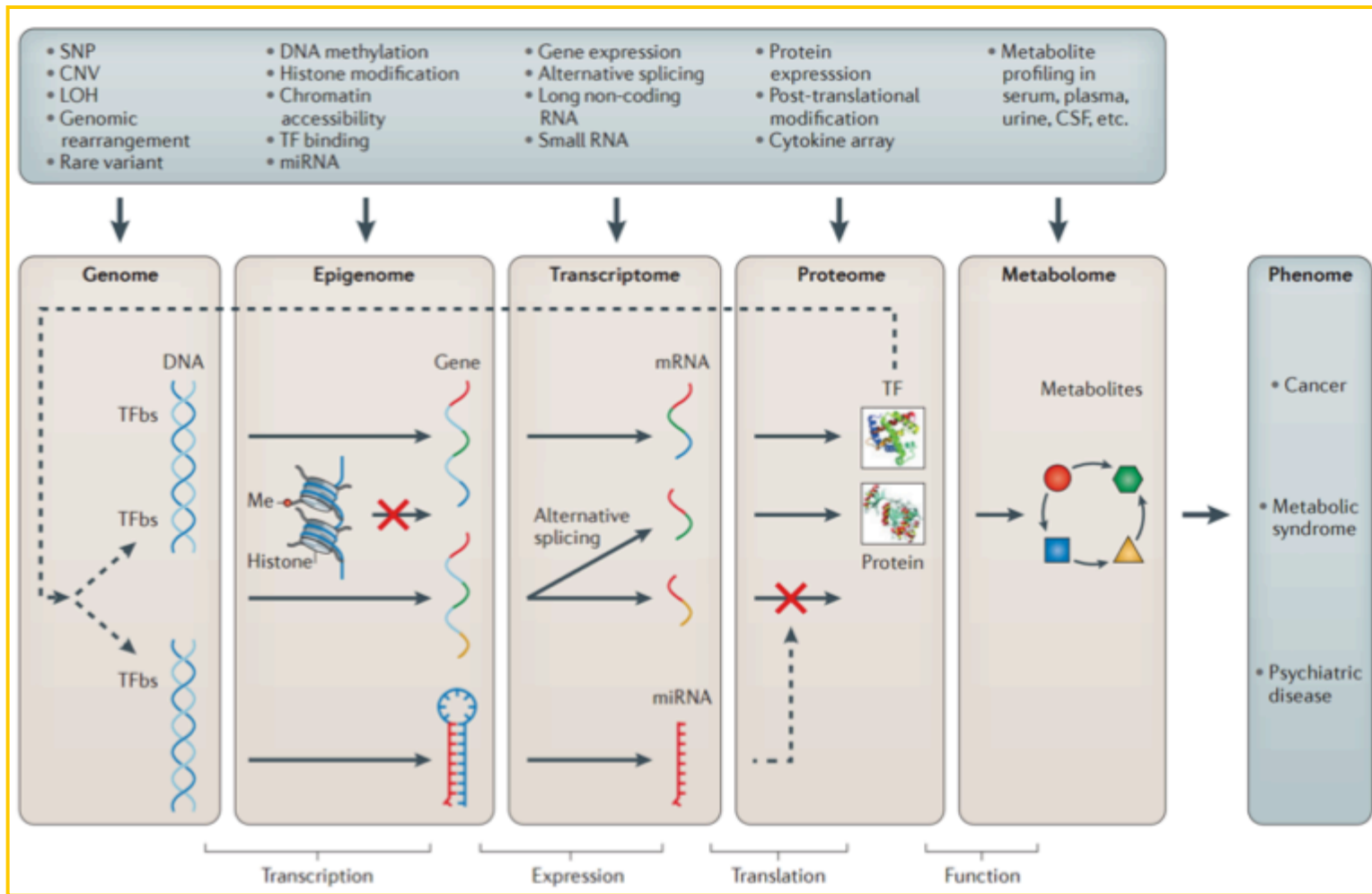
Exposures
Internal measurements
Disease states

Systems Biology

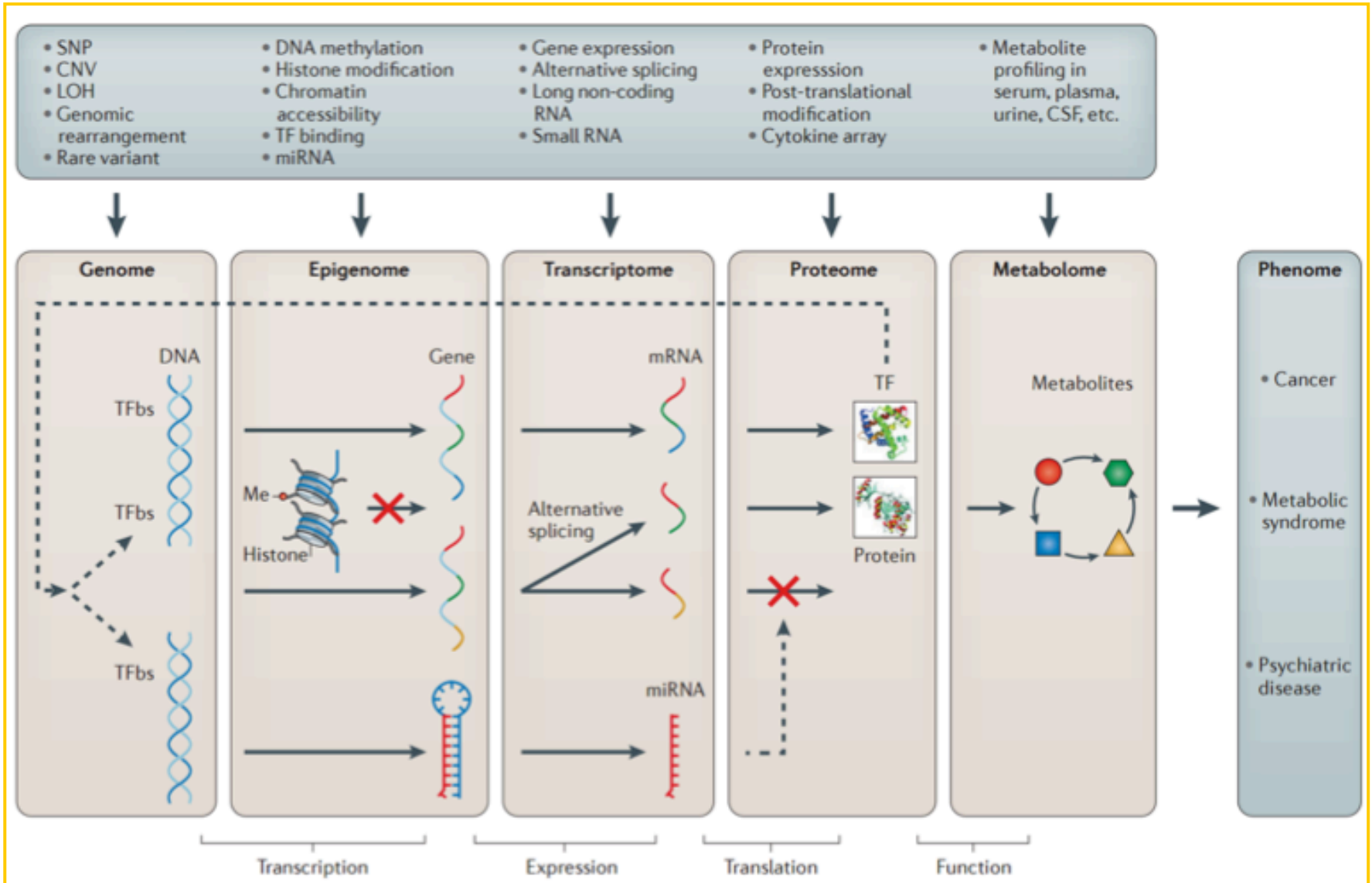
“Integrative approach in which scientists study pathways and networks will touch all areas of biology, including drug discovery”

C. Henry and C. Washington

Dissecting the Biological system via -omics



Dissecting the Biological system via -omics



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

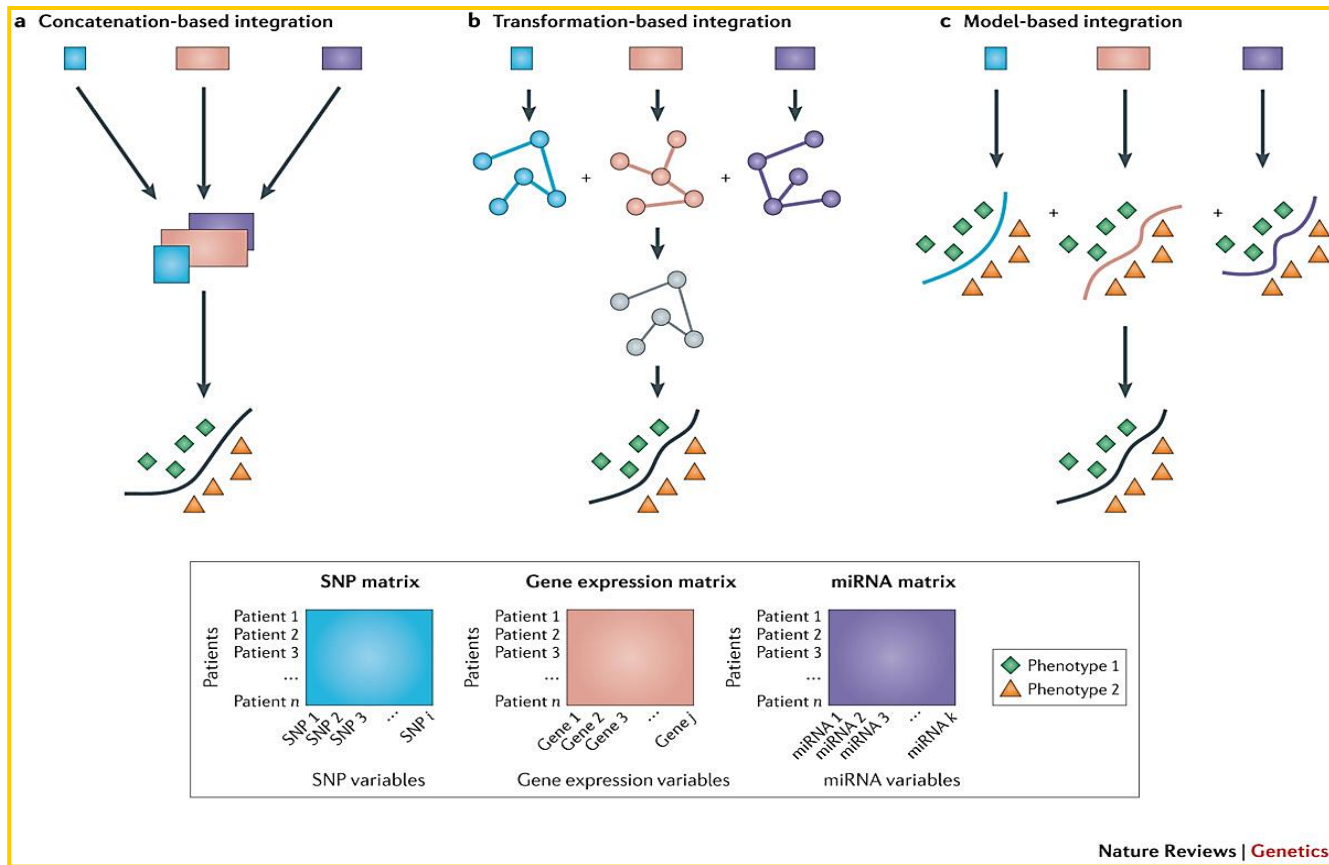
Paired integrative –omics analysis

- 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

- Multi-stage analysis:
 - A step-wise procedure where first associations are found between different data types, and then between the data types or phenotype of interest
 - Example:
 - 1) SNPS -> Phenotype
 - 2) SNPS selected from 1) are associated with other -omic data
 - 3) Omic data from 2) are then associated with Phenotype
- Meta-dimensional analysis:
 - Integration is performed globally such that data from multiple omics layers are combined simultaneously

Categories of meta-dimensional omics integration



Meta-dimensional analysis can be divided into three categories. **a** | Concatenation-based integration involves combining data sets from different data types at the raw or processed data level before modelling and analysis. **b** | Transformation-based integration involves performing mapping or data transformation of the underlying data sets before analysis, and the modelling approach is applied at the level of transformed matrices. **c** | Model-based integration is the process of performing analysis on each data type independently, followed by integration of the resultant models to generate knowledge about the trait of interest. miRNA, microRNA; SNP, single-nucleotide polymorphism.

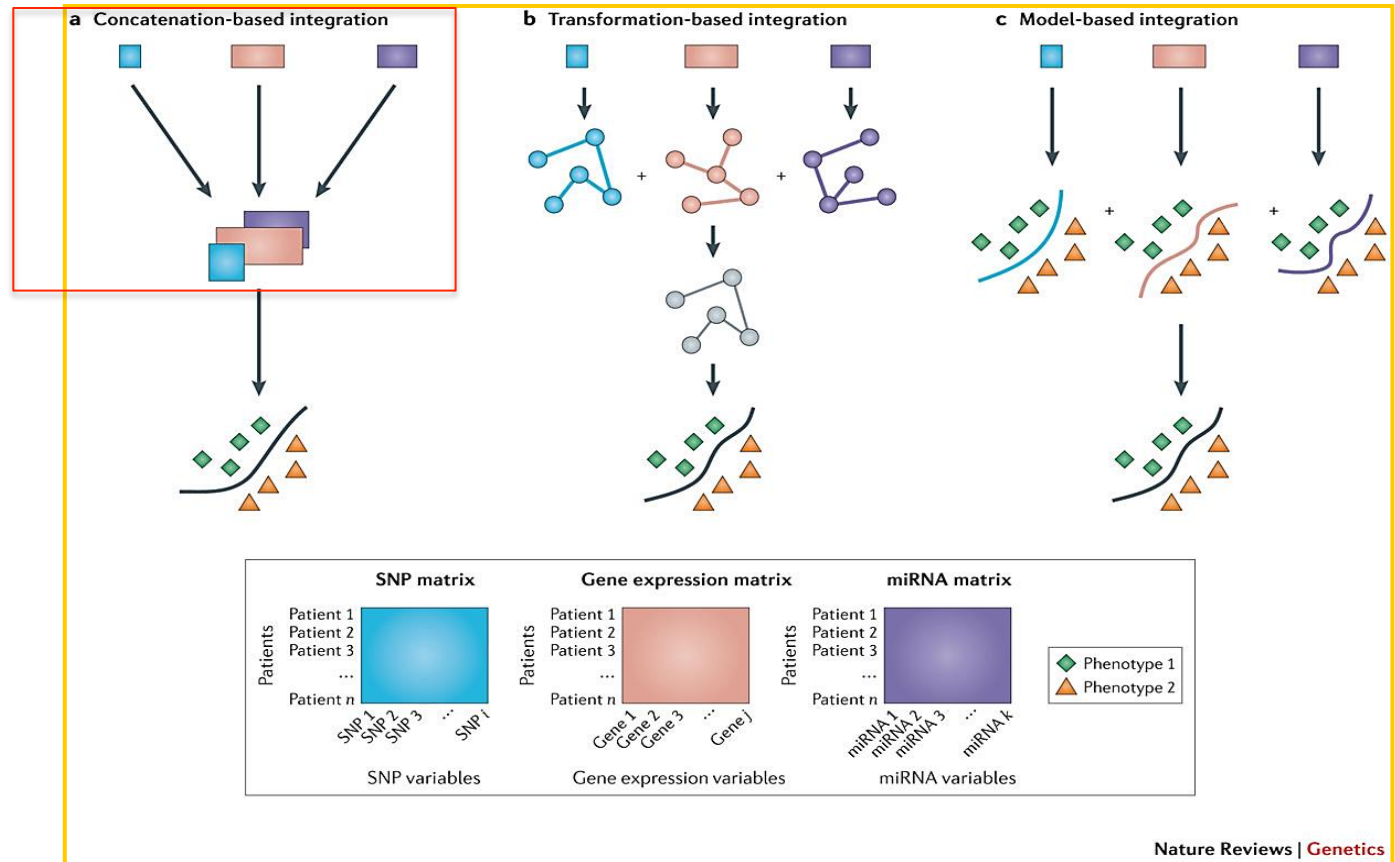
Categories of meta-dimensional omics integration

1) Concatenation-based integration:

combining data sets from different data types at the raw or processed data level before modelling and analysis

Caveats:

-Different data types should be at the same scale (discrete vs continuous)



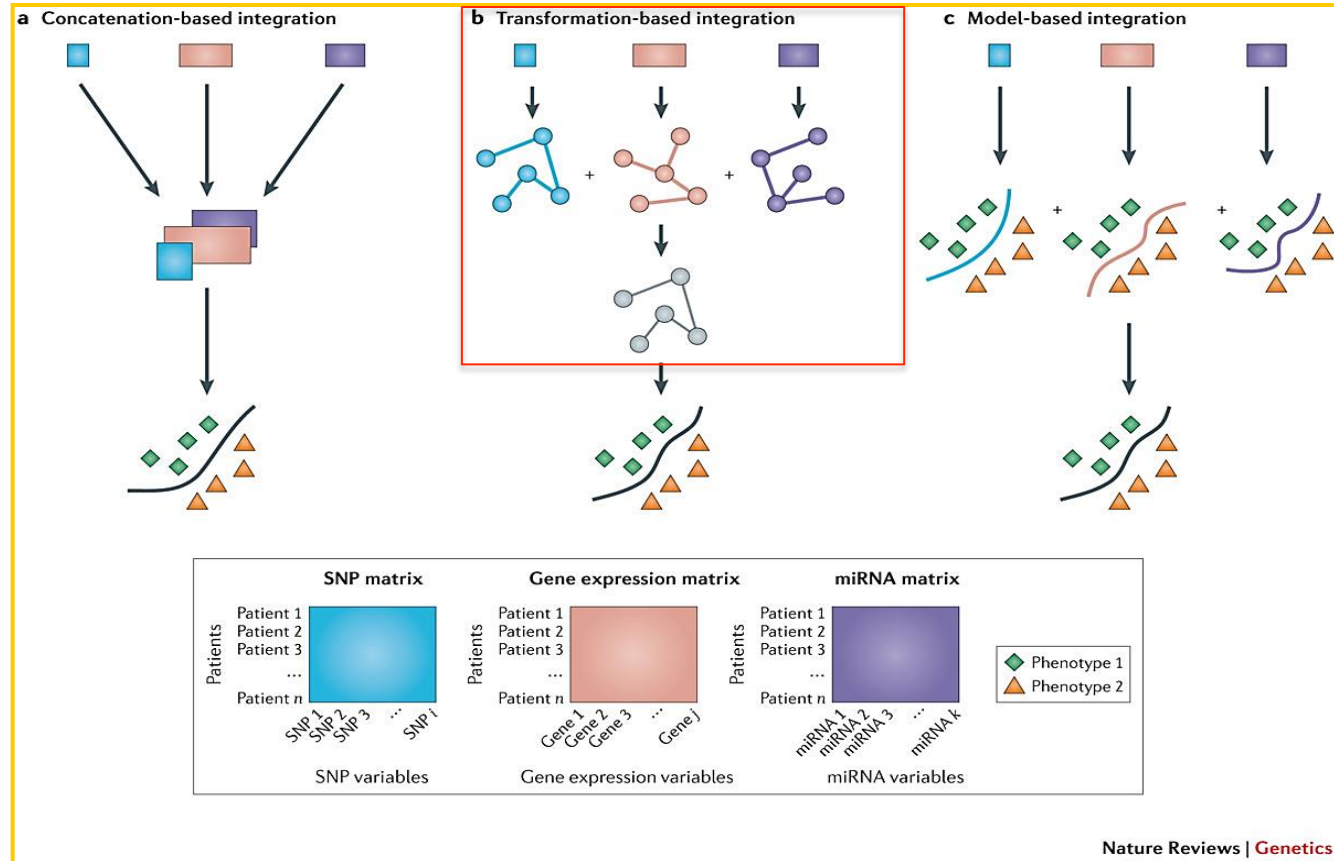
Categories of meta-dimensional omics integration

2) Transformation-based integration:

Data transformation is performed before analysis or modelling

Caveats:

-transformation should preserve original properties of the data to avoid loss of information



Categories of meta-dimensional omics integration

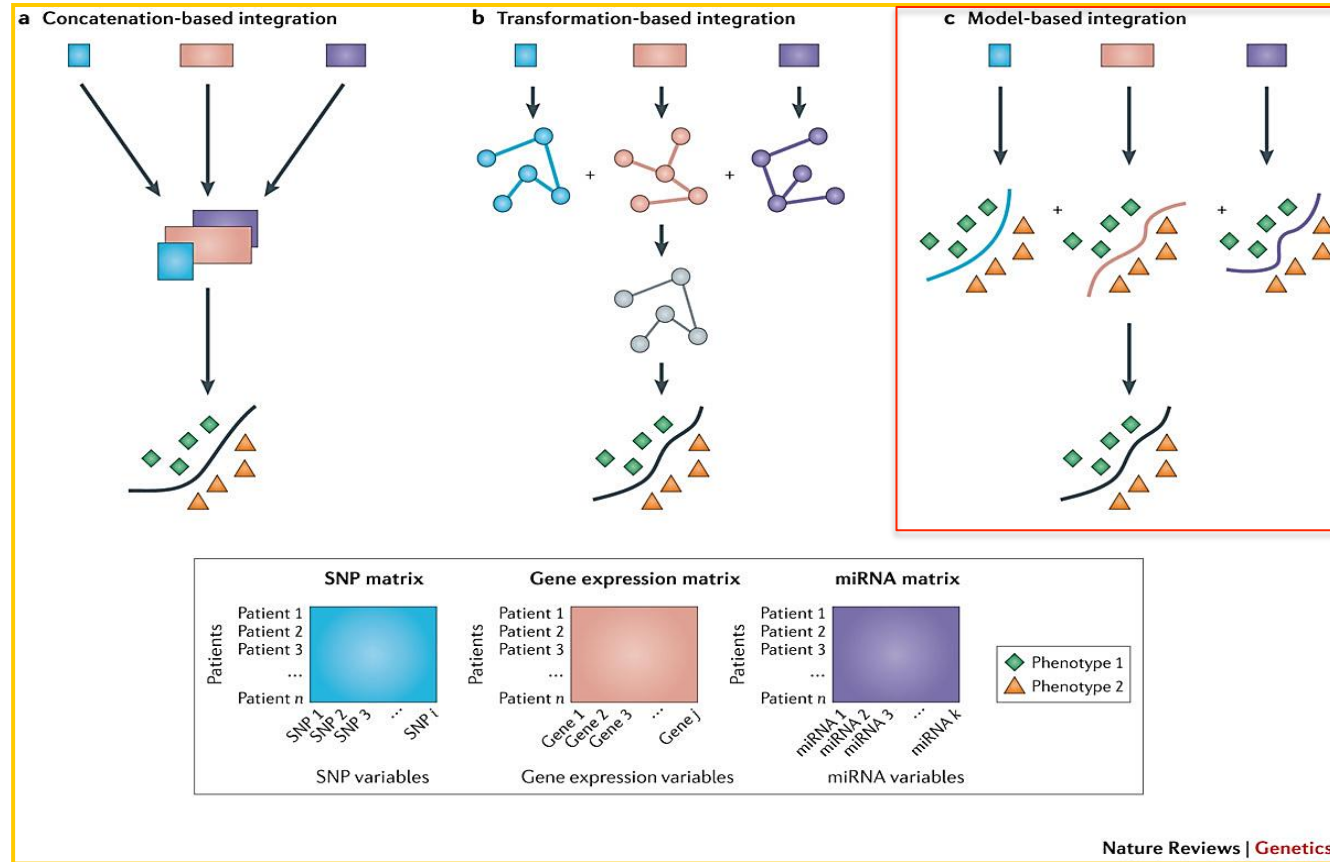
3) Model-based integration:

Variables are first selected from each omics layer in stage 1 based on the dependent variable (e.g. variables that discriminate cancer patients from controls) and integration is performed in stage 2

Caveats:

-model overfitting (number of samples \ll number of variables)

-chances of missing relevant interactions if variables are associated with the outcome through their interaction only



Tools and techniques for multi-omics integration and relationship visualization

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.

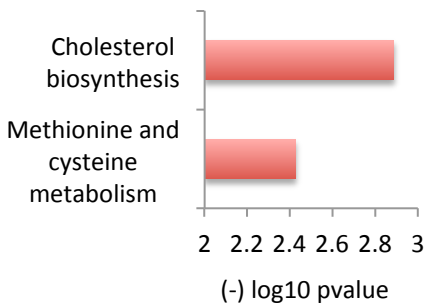
Multivariate

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

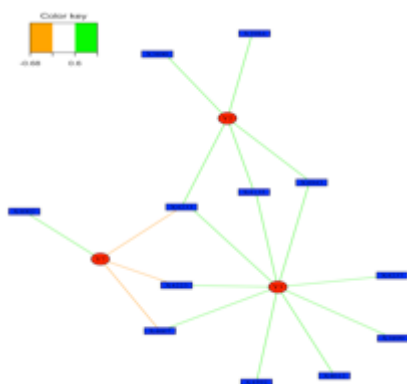
Workflow



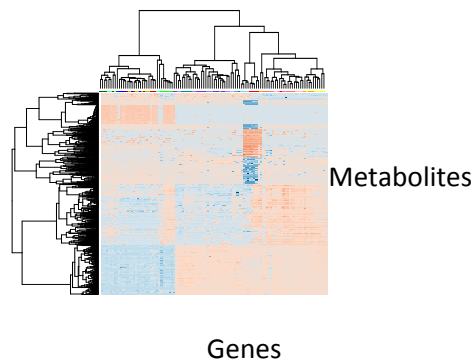
Pathway enrichment



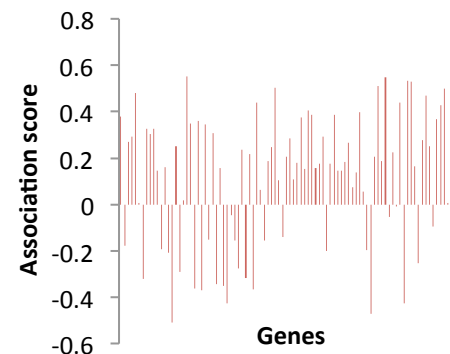
Relevance networks



Clustering

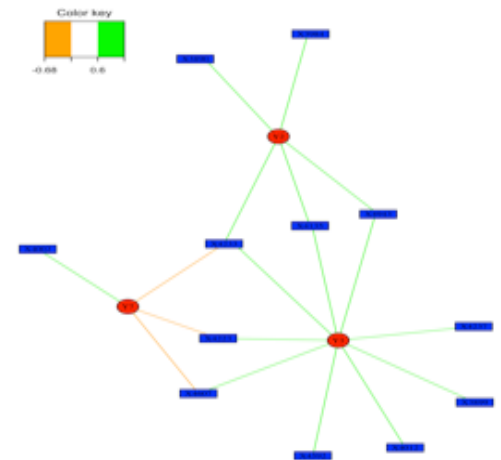


Targeted investigation (e.g.: Arginine x Transcriptome)



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 an 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)
 - Software:
 - MetabNet (Uppal 2015; R package for performing pairwise correlation analysis and generating relevance networks)
 - 3Omic (Kuo 2013; a web-based tool for analysis, integration and visualization of human transcriptome, proteome and metabolome data)
 - **Application:** Integration of TCE exposure data and physiological markers with metabolomics (Douglas I. Walker et al. submitted)
- **Multivariate**
 - Multivariate regression techniques such as partial least squares (PLS), sparse partial least squares regression (sPLS), multilevel sparse partial least squares (msPLS) regression, etc.
 - Software:
 - mixOmic (Cao et al. 2009, Liquet et al. 2012; R package for integration and variable selection using multivariate regression)
 - **Applications:**
 - Transcriptome x Metabolome (Roede, Uppal et al. 2013)
 - Microbiome x Metabolome (Cribbs, Uppal et al. 2016 in press)

Univariate methods

MetabNet (Uppal 2015)

- Performs pairwise correlation analysis to generate association matrix, M, e.g. (p metabolites x q genes)

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

$$t = r \sqrt{\frac{n-2}{1-r^2}} \quad (2)$$

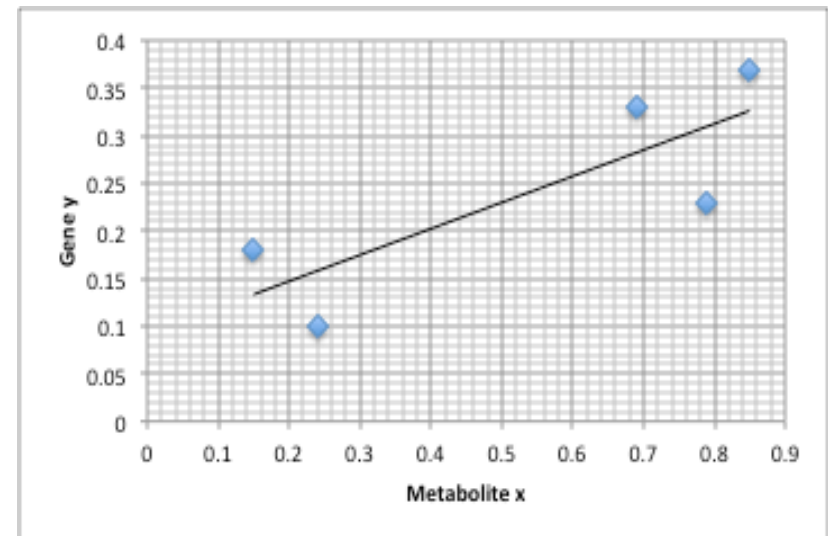
where,

x,y -> different omics data

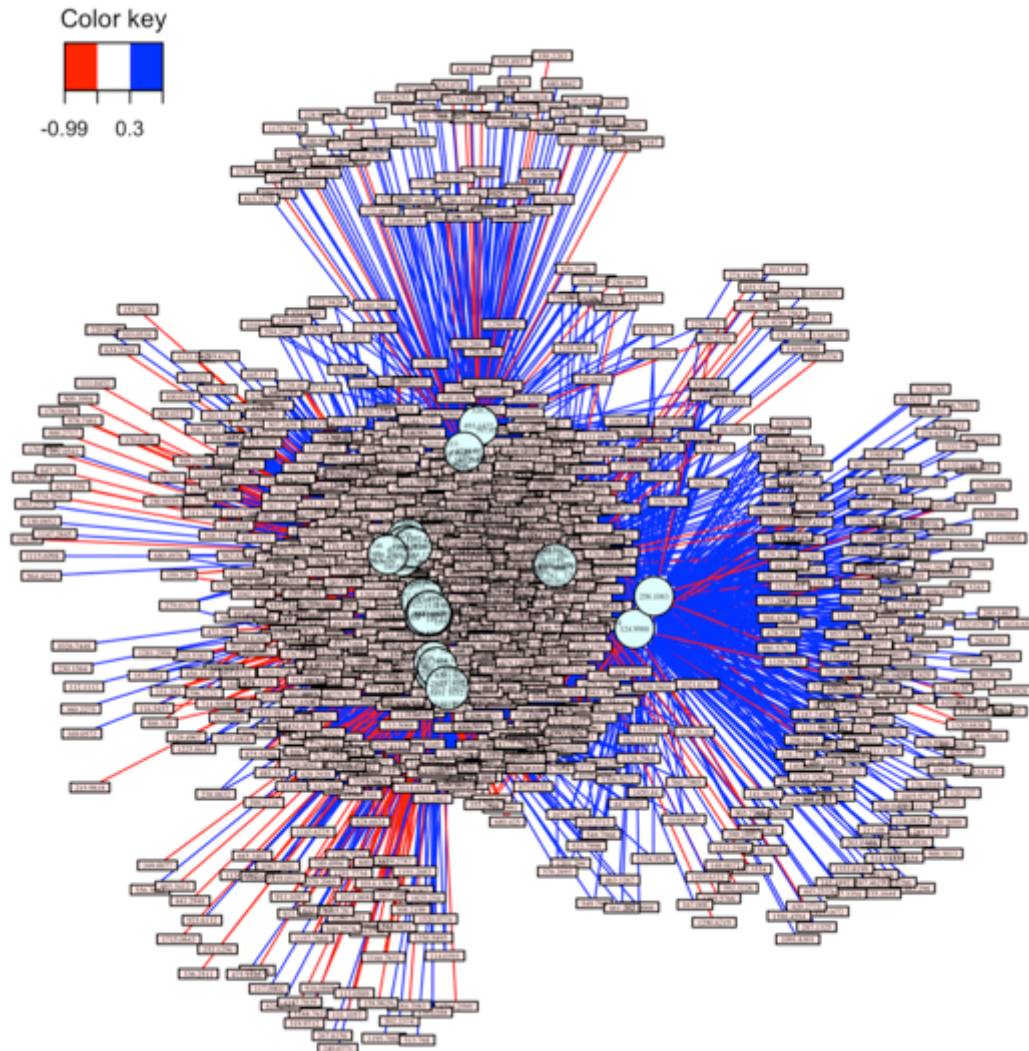
n -> number of samples

r -> Pearson Correlation

t -> t-statistic

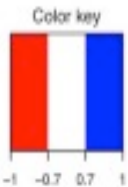


MetabNet output

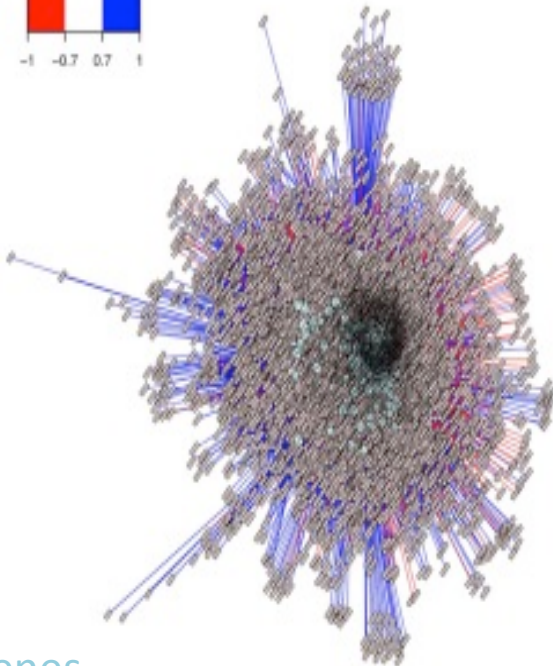


MetabNet output at different correlation thresholds

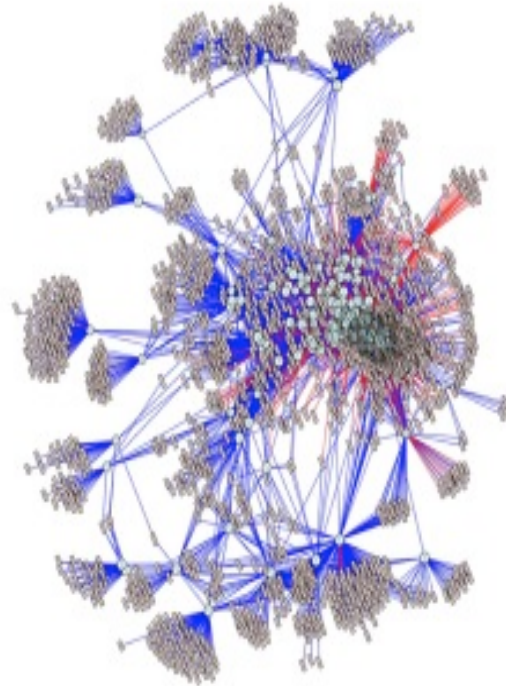
Increasing stringency



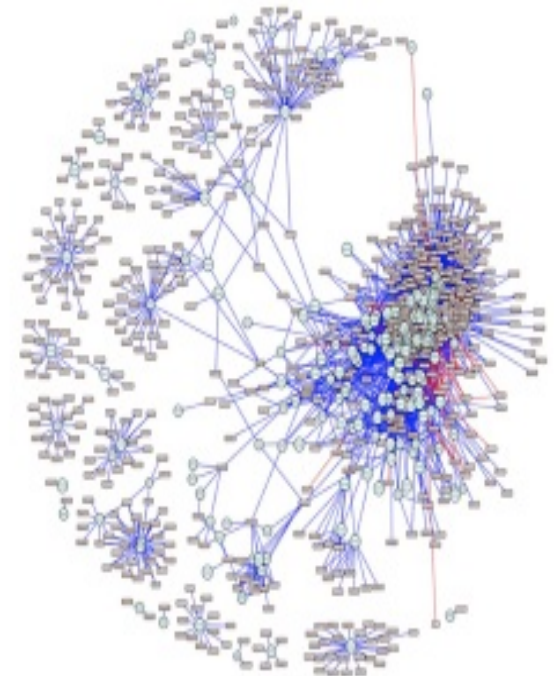
$|\text{cor}| > 0.3$



$|\text{cor}| > 0.5$



$|\text{cor}| > 0.7$



Genes
Metabolites

MetabNet R package

- Availability: Software and tutorial available on sourceforge (<https://sourceforge.net/projects/metabnet/>)
- Caveats:
 - Large number of possible associations ($p \times q$)
 - E.g.: 2×10^8 possible associations for 20,000 genes \times 10,000 metabolic features
 - Computationally intensive and hard to interpret results for large number of variables; Some pre-filtering based on missing values or other quality measures might be required
- More suitable when number of variables in at least one layer (p or q) is small

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

3omics.cmdm.tw

3Omics

Project Features

- Overview
- Name-ID Converter
- Help
- Contact Us

Overview

3Omics: A web based systems biology visualization tool for integrating human transcriptomic, proteomic and metabolomic data

3Omics is a one-click web tool for visualizing and rapidly integrating multiple inter- or intra-transcriptomic, proteomic, and metabolomic human data. It covers and connects cascades from transcripts, proteins, and metabolites and provides five commonly used analyses including correlation network, co-expression, phenotype generation, KEGG/HumanCyc pathway enrichment, and GO enrichment.

Please select the desired analysis:

- Transcriptomics-Proteomics-Metabolomics
- Proteomics-Metabolomics
- Transcriptomics only
- Transcriptomics-Proteomics
- Transcriptomics-Metabolomics
- Proteomics only
- Metabolomics only

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

Legend:

- Transcripts
- Proteins
- Metabolites
- Correlation
- Literature-derived relationship

Copyright (c) 2006-2012 Computational Molecular Design & Metabolomics Laboratory | BMRI | NTU | Contact us
We recommend using the latest version of Google Chrome or Mozilla Firefox to get the best experience using 3Omics services.

Features

- Correlation analysis and network visualization
 - Pairwise Pearson correlation analysis
- Literature-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

No file selected. [?](#)

GenBank ID: e.g. [NAT1](#), [ABL1](#)

Proteomics

No file selected. [?](#)

Uniprot Accession: e.g. [P31946](#), [P62258](#)

Metabolomics

No file selected. [?](#)

Data format

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

Samples

	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

[Help](#)

[Contact Us](#)

Parameters Section



How to set up parameters?

Correlation Coefficient Threshold
0.9

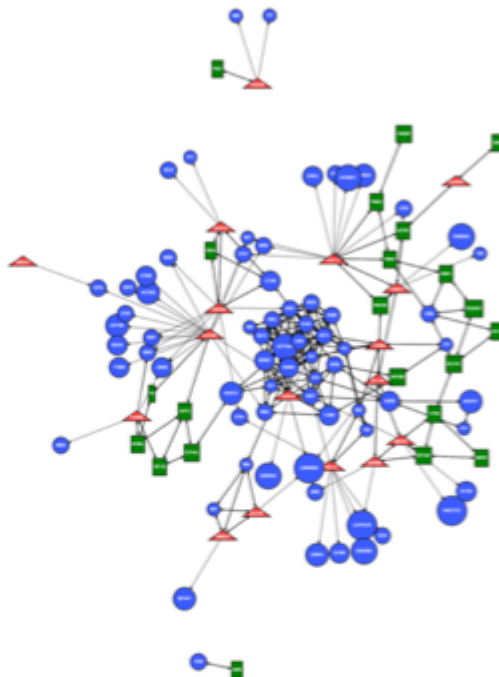
Correlation Network Repulsion
160

Correlation Network Attraction
80

[Refresh](#)

 **MDDL**
 **NATIONAL TAIWAN UNIVERSITY**

Correlation Network of Transcriptomics, Proteomics & Metabolomics



3Omics generates inter-omic correlation network to display the relationship or common patterns in data over time or experimental conditions for all transcripts, proteins and metabolites. Where users may only have two of the three -omics data-sets, 3Omics supplements the missing transcript, protein or metabolite information by searching [iHOP database](#).

Summary of Input molecules

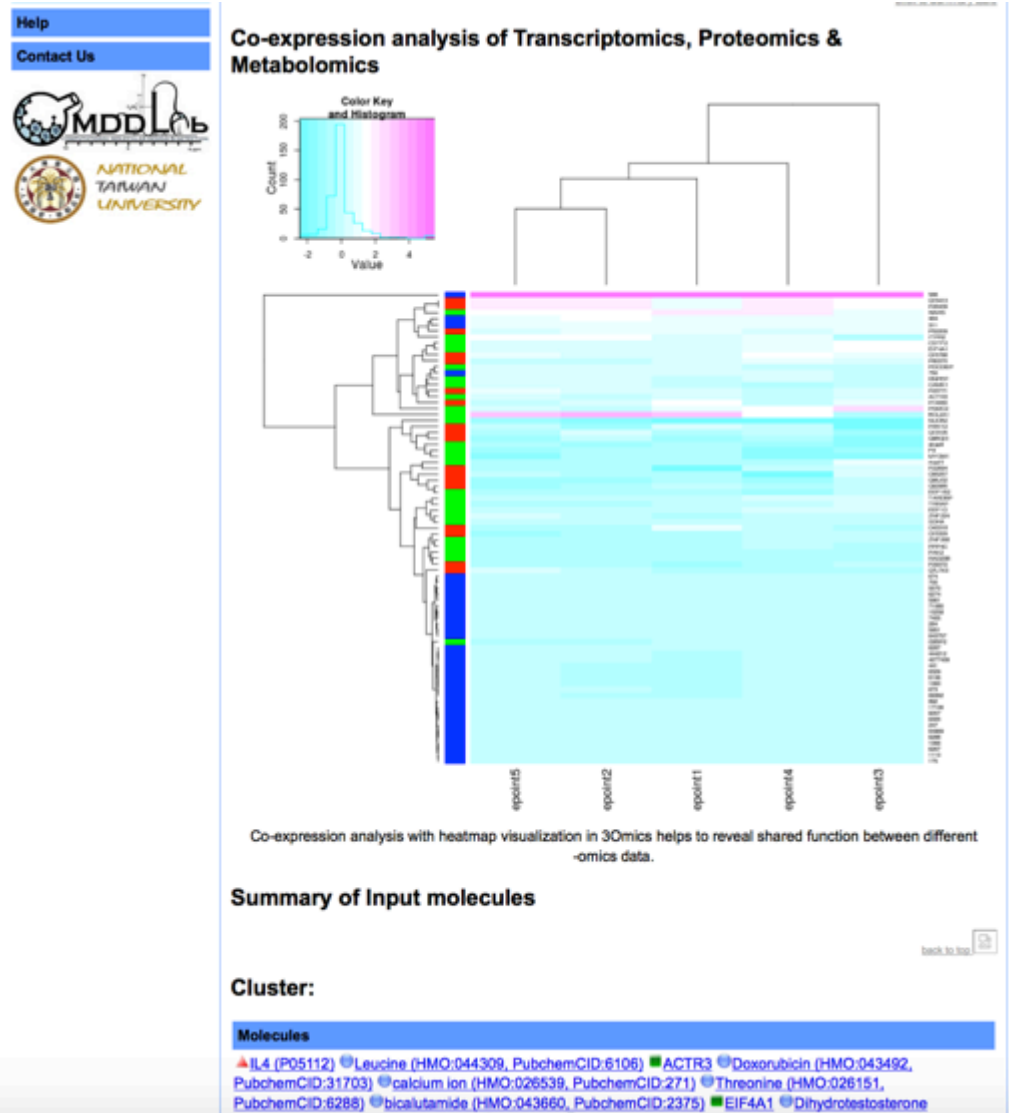
[back to top](#)

Cluster:

Molecules

- ▲ IL4 (P05112) ● Leucine (HMO:044309, PubchemCID:6106) ■ ACTR3 ● Doxorubicin (HMO:043492, PubchemCID:31703) ● calcium ion (HMO:026539, PubchemCID:271) ● Threonine (HMO:026151, PubchemCID:6288) ● bicalutamide (HMO:043660, PubchemCID:2375) ■ EIF4A1 ● Dihydrotestosterone (HMO:025783, PubchemCID:10635) ■ PDCC6IP ● bortezomib (HMO:048610, PubchemCID:387447) ■ PSMC2

Co-expression analysis



Rows: Variables
Columns: Samples

Phenotype analysis



Project Features

Overview

Name-ID Converter

Help

Contact Us



Correlation Network

Coexpression Profile

Phenotype Analysis

Pathway Analysis

GO Enrichment Analysis

[Click to Summary table](#)

Phenotype Analysis

A phenotype is defined as any observable characteristic or trait of an organism arising from gene expression, the influence of environmental factors, and the interactions between them. With phenotype-gene association from OMIM, genes and genetic disorders containing information to relate genes in the human genome with specific phenotypes can be identified.

The Transcriptomics data you've input have been used to search through the OMIM database, and the related phenotype and genes can be listed as below:

Please click the link for description and molecular genetic information on OMIM website.

Human-related Phenotype	Related-Gene
[OMIM: 611820] LONG QT SYNDROME 11	akap9
[OMIM: 256000] LEIGH SYNDROME	SDHA
[OMIM: 612069] AMYOTROPHIC LATERAL SCLEROSIS 10, WITH OR WITHOUT FRONTOTEMPORAL DEMENTIA WITH TDP43 INCLUSIONS	TARDBP
[OMIM: 306700] HEMOPHILIA A COAGULATION FACTOR VIII, INCLUDED	F8

Summary of Input molecules

[back to top](#)

Cluster:

Molecules

▲ IL4 (P05112) [Leucine \(HMO:044309, PubchemCID:6106\)](#) [ACTR3 \(HMO:043492, PubchemCID:31703\)](#) [calcium Ion \(HMO:026539, PubchemCID:271\)](#) [Threonine \(HMO:026151, PubchemCID:6288\)](#) [bicalutamide \(HMO:043660, PubchemCID:2375\)](#) [EIF4A1 \(HMO:048610, PubchemCID:387447\)](#) [PSMC2 \(HMO:025783, PubchemCID:10635\)](#) [PDCD6IP \(HMO:048610, PubchemCID:387447\)](#) [PSMC2](#) [trigonelline \(HMO:033252, PubchemCID:5570\)](#) [TARDBP](#) [RYR3 HBRR \(Q15413\)](#) [dimethylamine \(HMO: PubchemCID:674\)](#) [HSD3B1 3BH HSDB3A \(P14060\)](#) [Hydrocortisone \(HMO:043177, PubchemCID:5754\)](#) [Tyrosine \(HMO:026152, PubchemCID:6057\)](#) [Methotrexate \(HMO:042925, PubchemCID:126941\)](#) [formic acid \(HMO:044577, PubchemCID:284\)](#) [hippuric acid \(HMO:033093, PubchemCID:464\)](#) [Testosterone Propionate \(HMO:043961, PubchemCID:5995\)](#) [Androsterone \(HMO:027989, PubchemCID:5879\)](#) [MYOM1](#) [Leucine \(HMO:042148, PubchemCID:857\)](#) [zinc fluoride \(HMO:040479, PubchemCID:24551\)](#) [3d0b \(HMO:049721, PubchemCID:24812721\)](#) [Mifepristone \(HMO:043298, PubchemCID:55245\)](#) [MAP3K7 TAK1 \(Q43318\)](#) [Aconitic Acid \(HMO:033434, PubchemCID:444212\)](#) [Indican \(HMO:049137, PubchemCID:10258\)](#) [Estradiol \(HMO:026665, PubchemCID:5757\)](#) [NTH \(HMO:049464, PubchemCID:5289054\)](#) [Inositol \(HMO:036496,](#)

Pathway analysis



Project Features

- Overview
- Name-ID Converter
- Help
- Contact Us

Correlation Network

Coexpression Profile

Phenotype Analysis

Pathway Analysis

GO Enrichment Analysis

[Click to Summary table](#)

Pathway analysis - Normal, non-enrichment

[KEGG section](#) | [HumanCyc section](#)

KEGG Pathway analysis [back to top](#)

KEGG pathway enrichment analysis operates upon metabolomic data to reveal enriched pathways in a KEGG Pathway database by ranking the biological pathways commonly shared by metabolites.

The enriched KEGG metabolic pathways are listed on the bottom of the page. Please click to see mapped pathway images on KEGG Pathway.

(Normal Mode) Show Records: 20



Metabolic Pathways	Hits
(hsa01100) Metabolic pathways - Homo sapiens (human) <ul style="list-style-type: none"> • Acetate • D-Alanine • L-Asparagine • Betaine • Citrate • Ethanolamine • Formate • 6-Deoxy-L-galactose • L-Glutamine • Glycine • L-Histidine • N,N-Dimethylglycine • Pyruvate • Pyridine-2,3-dicarboxylate • L-Serine • Succinate • L-Tryptophan • L-Tyrosine • N(p)-Methyl-L-histidine 	19
(hsa00970) Aminoacyl-tRNA biosynthesis - Homo sapiens (human) <ul style="list-style-type: none"> • L-Asparagine • L-Glutamine • Glycine • L-Histidine • L-Serine • L-Threonine • L-Tryptophan • L-Tyrosine 	8
(hsa00250) Alanine, aspartate and glutamate metabolism - Homo sapiens (human) <ul style="list-style-type: none"> • Acetate • L-Asparagine • L-Glutamine • Glycine • Pyruvate • Succinate • L-Tyrosine 	7
(hsa00280) Valine, leucine and isoleucine degradation - Homo sapiens (human) <ul style="list-style-type: none"> • Acetate • Glycine • Pyruvate • Succinate • L-Tryptophan • L-Tyrosine 	6
(hsa00270) Cysteine and methionine metabolism - Homo sapiens (human) <ul style="list-style-type: none"> • Betaine • N,N-Dimethylglycine • Pyruvate • L-Serine • L-Tryptophan • L-Tyrosine 	6
(hsa00330) Arginine and proline metabolism - Homo sapiens (human) <ul style="list-style-type: none"> • Acetate • L-Glutamine • Glycine • Pyruvate • Succinate • L-Tyrosine 	6
(hsa00360) Phenylalanine metabolism - Homo sapiens (human) <ul style="list-style-type: none"> • Acetate • Glycine • L-Histidine • L-Tryptophan • L-Tyrosine 	5
(hsa00340) Histidine metabolism - Homo sapiens (human) <ul style="list-style-type: none"> • Acetate • L-Histidine • L-Tryptophan • L-Tyrosine • N(p)-Methyl-L-histidine 	5
(hsa00260) Glycine, serine and threonine metabolism - Homo sapiens (human) <ul style="list-style-type: none"> • Betaine • Glycine • N,N-Dimethylglycine • Pyruvate • L-Serine 	5
(hsa00520) Amino sugar and nucleotide sugar metabolism - Homo sapiens (human)	4

GO Enrichment Analysis



Project Features

- Overview
- Name-ID Converter
- Help
- Contact Us

Correlation Network
Coexpression Profile
Phenotype Analysis
Pathway Analysis
GO Enrichment Analysis

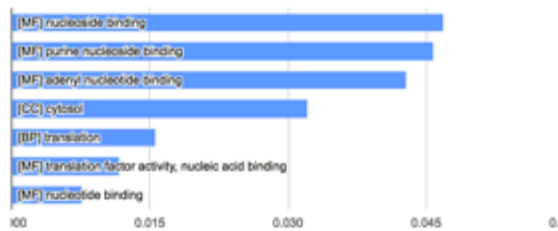
[Click to Summary table](#)

Gene Ontology functional Profiling

The Gene Ontology (GO) provides defined terms for representing the properties of gene product. GO covers three levels of properties: i) cellular component ii) biological process iii) molecular function help users to understand information of gene products from the defined three domains.

[biological process](#) | [cellular component](#) | [molecular function](#)

GO Terms with P-value < 0.05



Biological Process

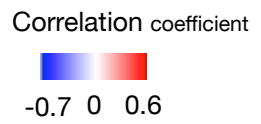
[back to top](#)

A biological process is a process of a living organism. Biological processes are made up of any number of chemical reactions or other events that results in a transformation. Regulation of biological processes occurs where any process is modulated in its frequency, rate or extent. Biological processes are regulated by many means; examples include the control of gene expression, protein modification or interaction with a protein or substrate molecule.

GO Term	No. of Gene-mapped	Coverage	P-value	FDR	Mapped Gene ID
translation	4	17%	0.0156	EEF1D,EEF1B2,EIF4A1,WARS	1936, 1933, 1973, 7453
cell death	4	17%	0.1082	PDCD6IP,BCL2A1,TARDBP,PAK2	10015, 597, 23435, 5062
death	4	17%	0.1104	PDCD6IP,BCL2A1,TARDBP,PAK2	10015, 597, 23435, 5062
apoptosis	3	13%	0.2565	PDCD6IP,BCL2A1,PAK2	10015, 597, 5062

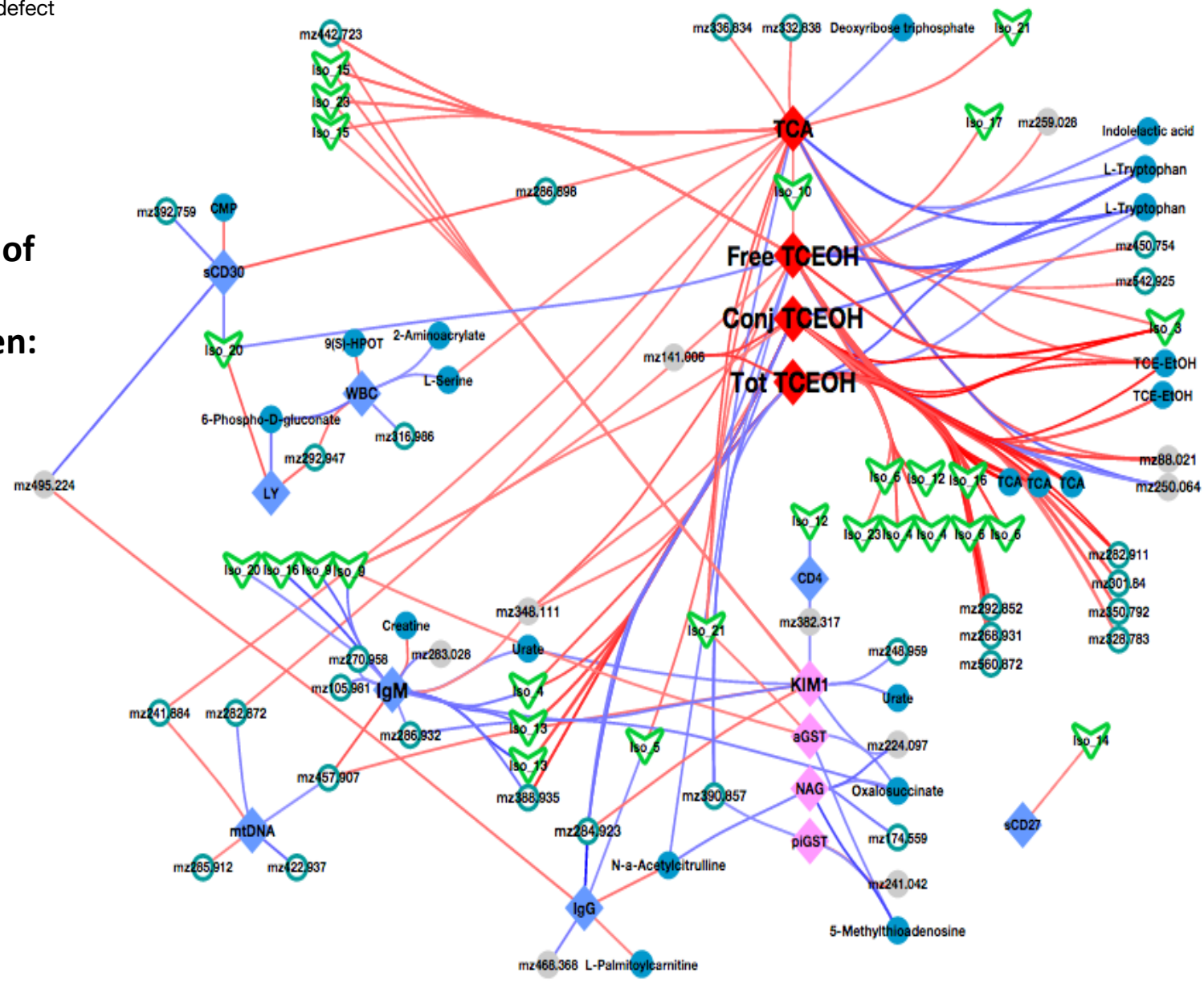
Case Study 1: Using MetabNet for cross-platform paired integrative analysis. Integration of TCE exposure data and physiological markers with metabolomics
(Walker, Uppal et al. manuscript submitted)

- ◆ Urinary TCE exposure markers
- ◆ Renal biomarkers
- ◆ Immunological markers
- ✓ Unidentifiable halogenated m/z, by isotopic pair
- Identified metabolite
- Unidentifiable m/z; pos mass defect
- Unidentifiable m/z; neg mass defect



Integrative analysis allows visualization of complex associations between:

- 1) environmental exposure markers;
- 2) renal biomarkers;
- 3) immunological markers;
- 4) metabolites



Courtesy:
Douglas I. Walker
(manuscript submitted)

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)
- 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: metabolome (matrix X) and transcriptome (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

multilevel sPLS method for experiments with repeated measurements

If X is an $(N \times p)$ intensity matrix, where N is the number of samples and p is the number of m/z features, then

1) Split-up variation:

$$X_w = X_{\text{stimulation}} + X_{\text{time}} + X_{\text{stimulation} \times \text{time}} + X_{\text{residual}} \\ + X_{\text{subject} \times \text{Stimulation}} + X_{\text{subject} \times \text{time}}$$

2) sparse PLS objective function:

$$\max \text{cor}(Y, X_u) \text{var}(X_u)$$

where

Y is the matrix indicating group of each sample

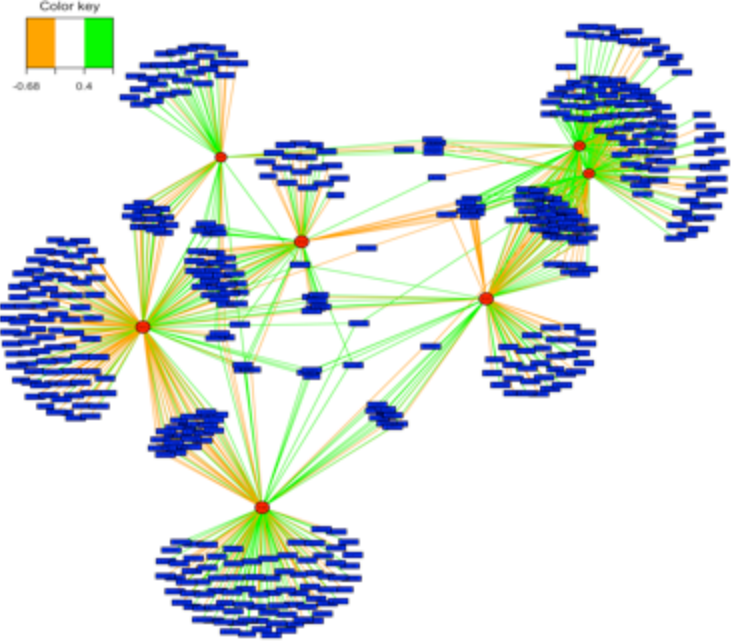
X is the split-up variation

u_1, u_2, \dots, u_H are the loading vectors

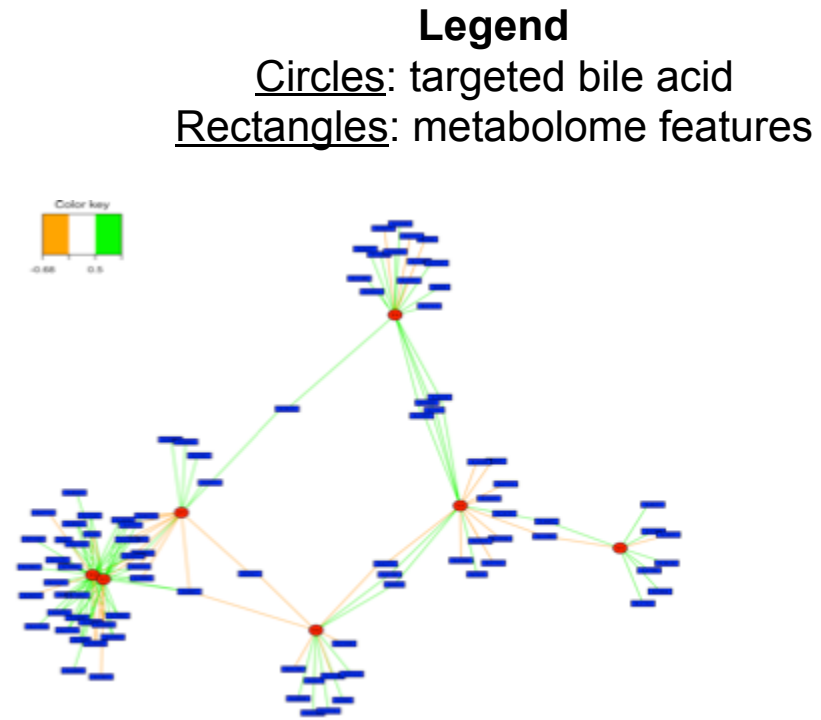
H is the number of PLS-DA dimensions

A Lasso based optimization is used to select most relevant variables

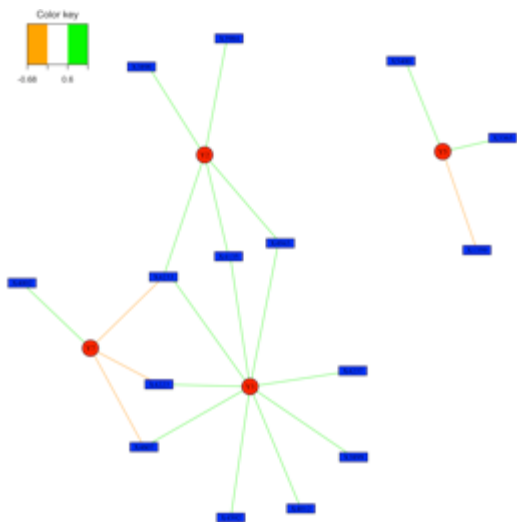
Case Study 2: Application of sPLS technique for cross-platform paired integrative analysis. Integration of targeted bile acids measurements and clinical variables (age, BMI, etc.) with metabolomics



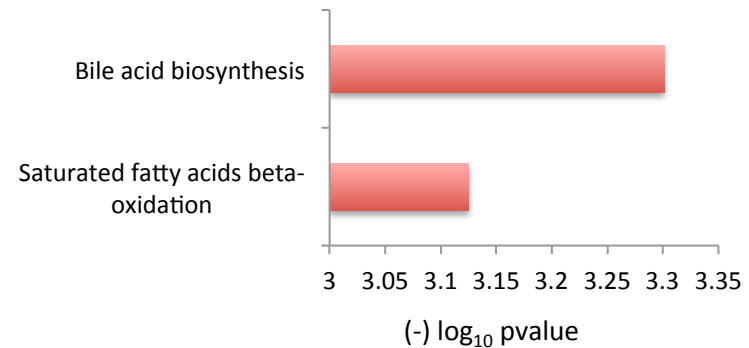
A. Association threshold: 0.4



B. Association threshold: 0.5



C. Association threshold: 0.6

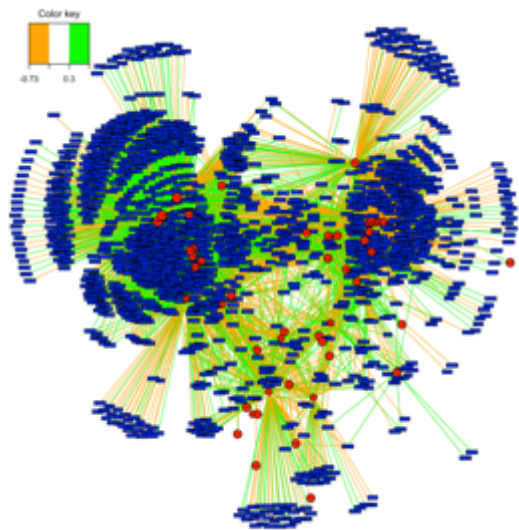


D. Pathway analysis (only top two displayed)

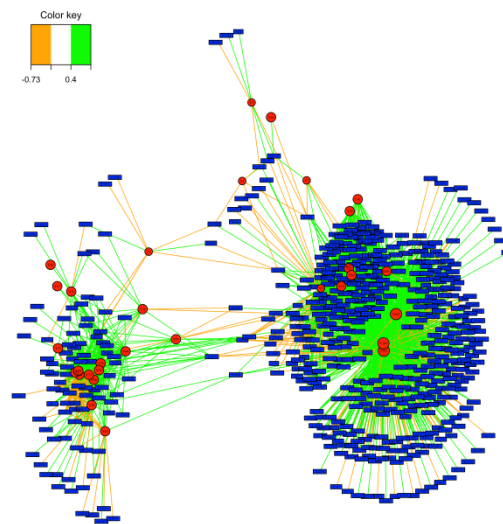
Case Study 3: Application of sPLS technique for integrative –omics. Microbiome-Metabolome Wide Association Study of Lung BAL: Global integration of 5930 m/z features with 153 microbial species using sparse Partial Least Squares regression (Cribbs et al. Microbiome 2015)

Legend

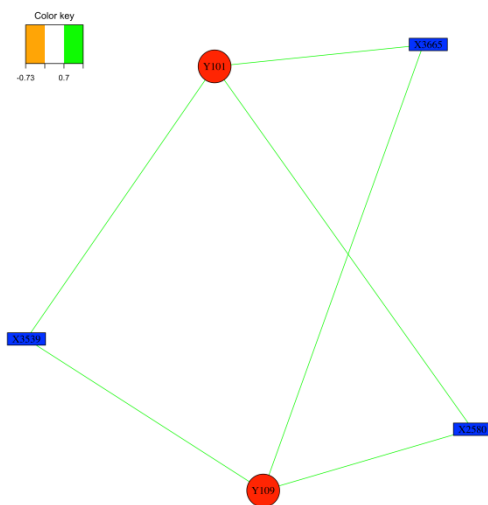
Circles: microbial species
Rectangles: metabolome features



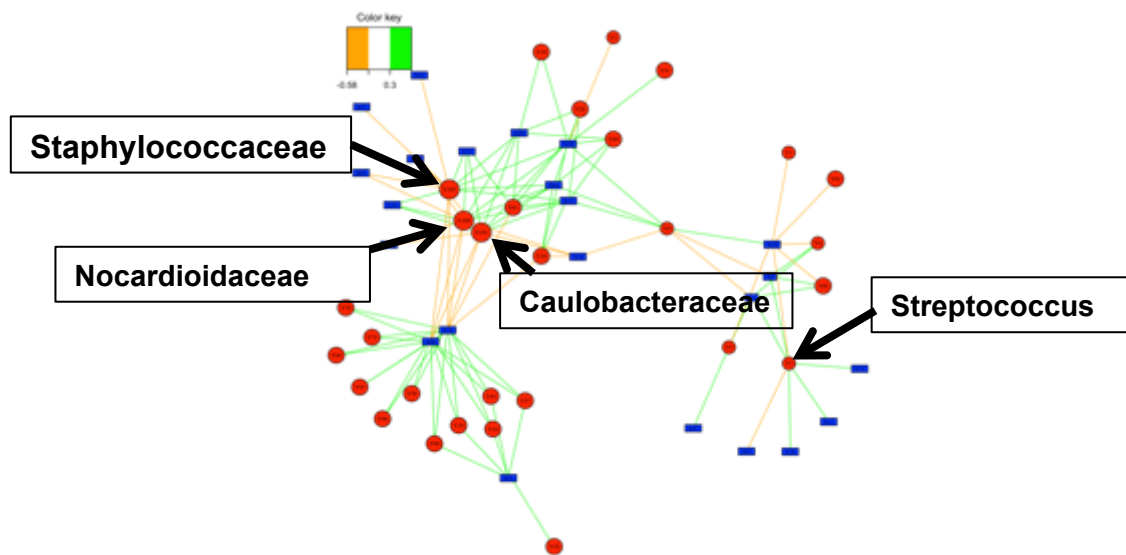
A. Association threshold: 0.3



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 data from other
sources (e.g. PubMed)

Association mining algorithm for constructing relation trees

$$\text{Pointwise Mutual Information}(t_1, t_2) = v_i * \log_2 \frac{p(t_1 \text{ and } t_2)}{p(t_1) p(t_2)}$$

where

v_i is 1 if term t_2 is present in the controlled vocabulary, 0 otherwise;

$p(t_1)$ is the probability of term 1 in the corpus,

$p(t_2)$ is the probability of term 2 in the corpus, and

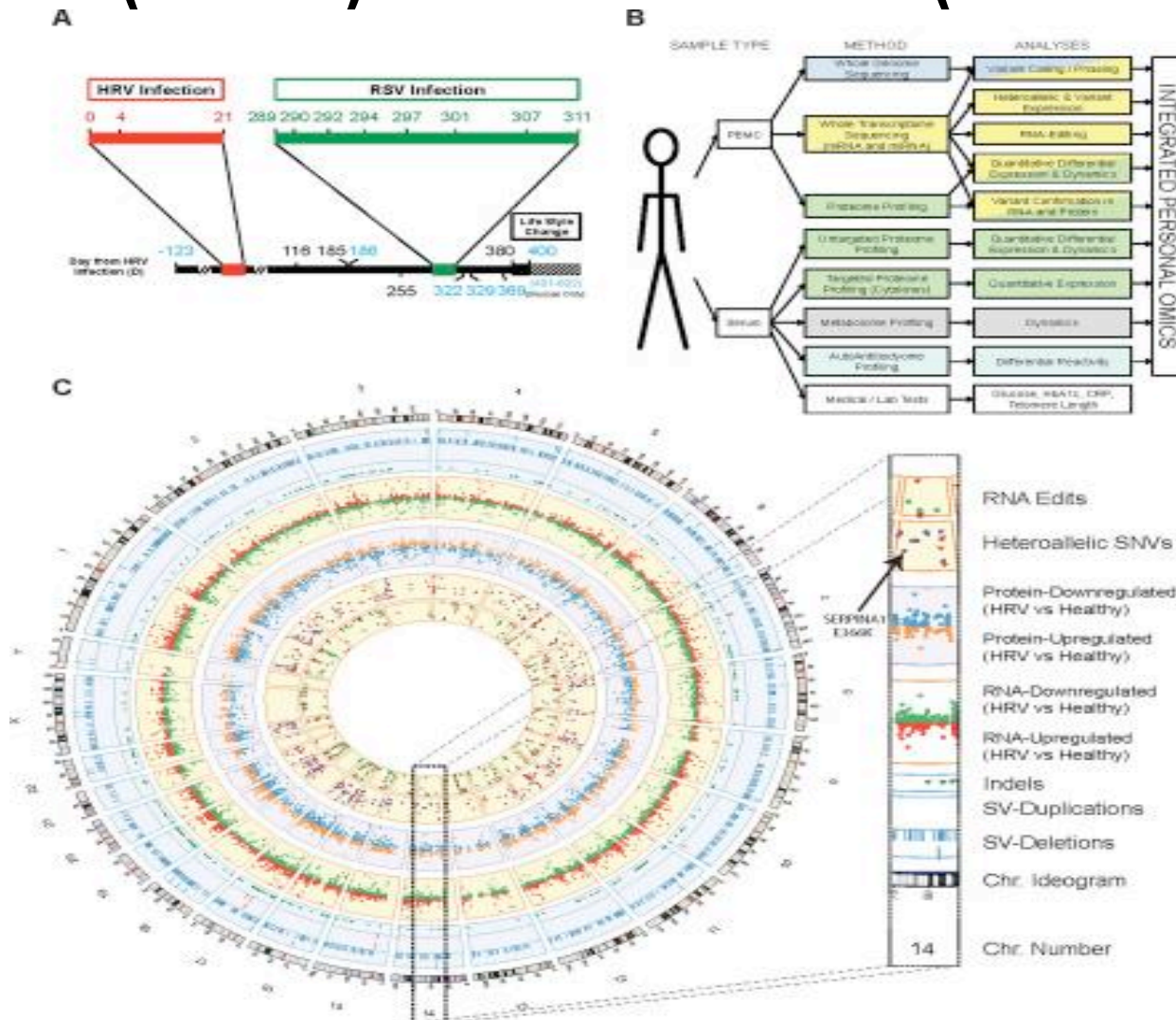
$p(t_1 \text{ and } t_2)$ is the probability of co-occurrence of terms 1 and 2 in the corpus

Dictionaries for biomedical terms: PubTator, MeSH, NCBI databases

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

Integrative Personal Omics Profiling (iPOP) – Chen et al. (Cell 2012)

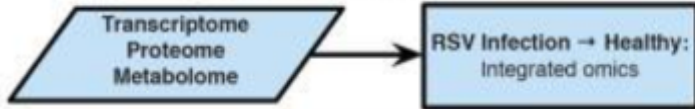


Temporal multi-omic profiling of one individual for 14 months

Figure 1. Study summary Chen et al. (Cell 2012)

Figure 4. Integrative Omics analysis
Chen et al. (Cell 2012)

Integrated Omics clustering



Full Reactome (FI) known pathway map
for cluster:



(II) Spike maxima data clusters

Eg. Pathway Analysis Results - FDR < 5e-02:

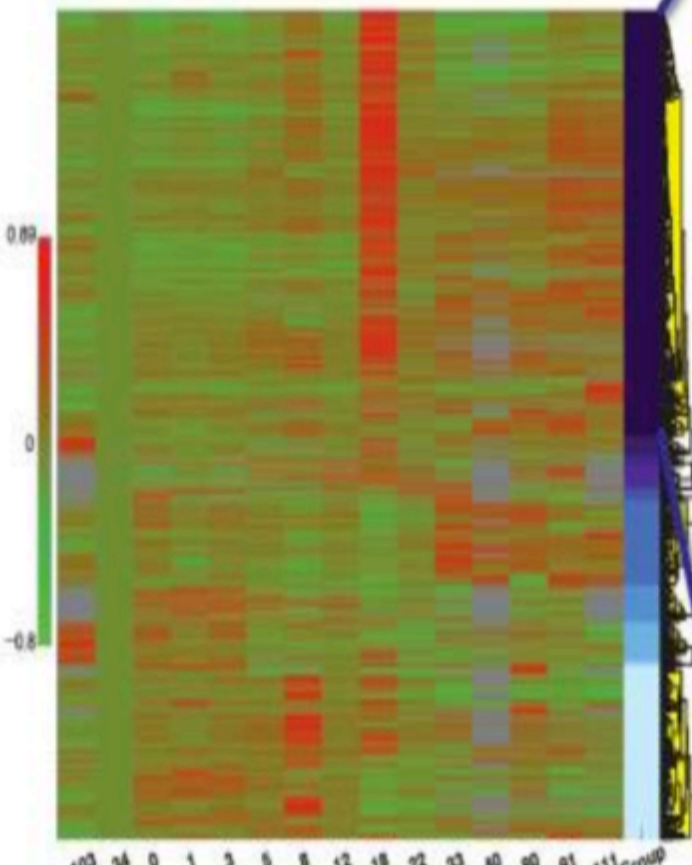
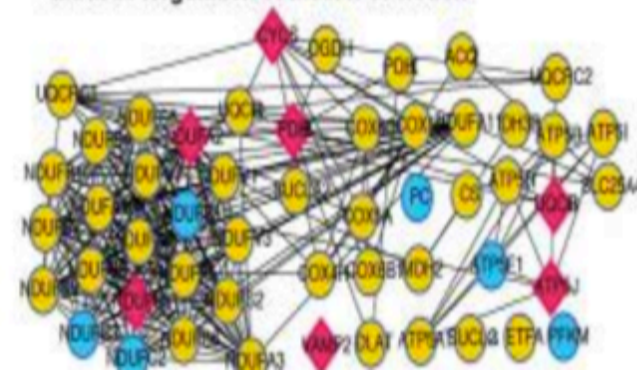
- Spliceosome(K)
- Glucose Regulation of Insulin Secretion(R)
- Formation and Maturation of mRNA Transcript(R)
- Oxidative phosphorylation(K)
- Electron Transport Chain(R)
- Parkinson's disease(K)
- Huntington's disease(K)
- Influenza Life Cycle(R)
- Metabolism of non-coding RNA(R)
- Transport of Mature Transcript to Cytoplasm(R)
- Protein export(K)
- Pyruvate metabolism and TCA cycle(R)

GO-ID	corr p-value	Description
8380	3.59E-71	RNA splicing
6396	2.53E-57	RNA processing
16070	5.93E-54	RNA metabolic process
16071	6.10E-50	mRNA metabolic process
10467	1.74E-48	gene expression
90304	1.78E-45	nucleic acid metabolic process

Eg. Metabolites in Cluster

- 3R-hydroxy-5Z-dodecenoic acid
- 5,6-DIHTe-EA
- 7-Ethoxycoumarin
- Lauric acid
- 1-O-(1Z-tetradecenyl)-2-(9Z-octadecenyl)-sn-glycerol
- (23R)-1alpha,23,25-trihydroxy-24-oxovitamin D3 / (23R)-1alpha,23,25-trihydroxy-24-oxocholecalciferol
- 1alpha-hydroxy-26,27-dinorvitamin D3 25-carboxylic acid / 1alpha-hydroxy-26,27-dinorcholecalciferol
- 12-oxo-9-octadecynoic acid
- GPCho(C-16:0/O-4:0[U])
- 19-hydroxy-17-oxoandrost-5-en-3-beta-yl sulfate · 11.538899

Example Pathway: Glucose Regulation of Insulin Secretion



Questions?