Computational methods for integrative omics analysis

July 21, 2016

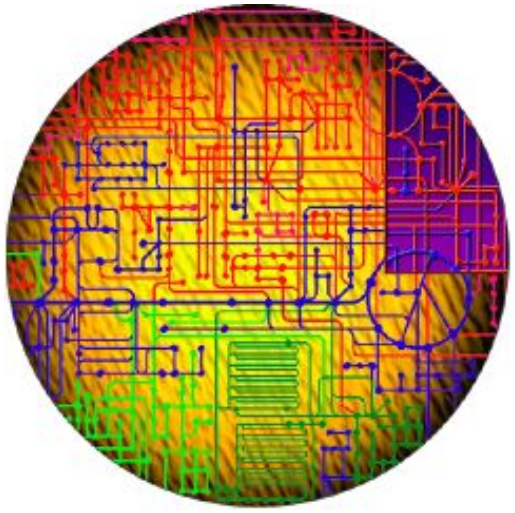
Karan Uppal, PhD
Assistant Professor
Emory University School of Medicine

Learning Objectives

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

Introduction: A Systems Biology Framework

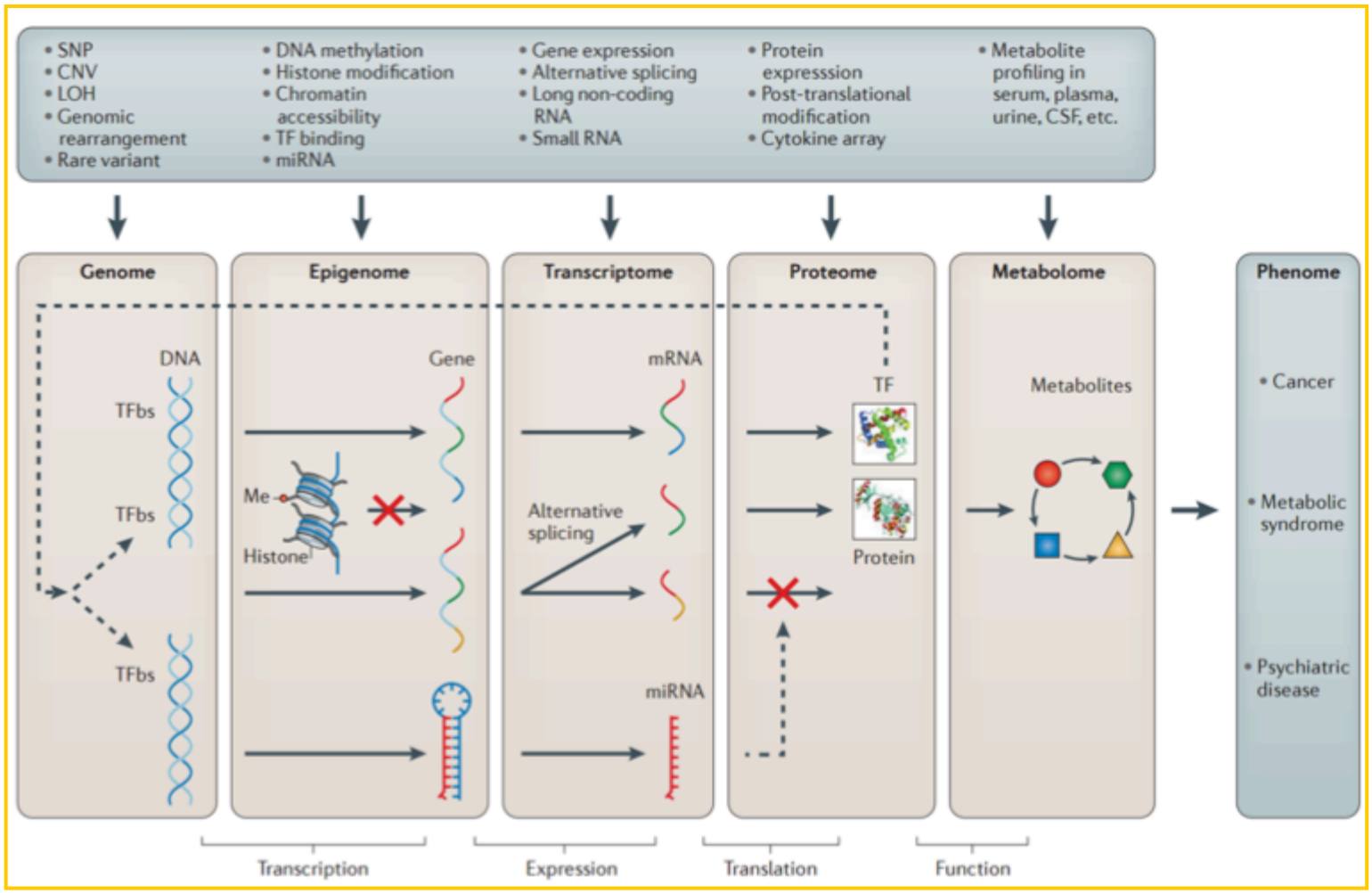
- The goal of **Systems Biology**:
  - Systems-level understanding of biological systems
  - Analyze not only individual components, but their interactions as well and emergent behavior



Exposures
Internal measurements
Disease states

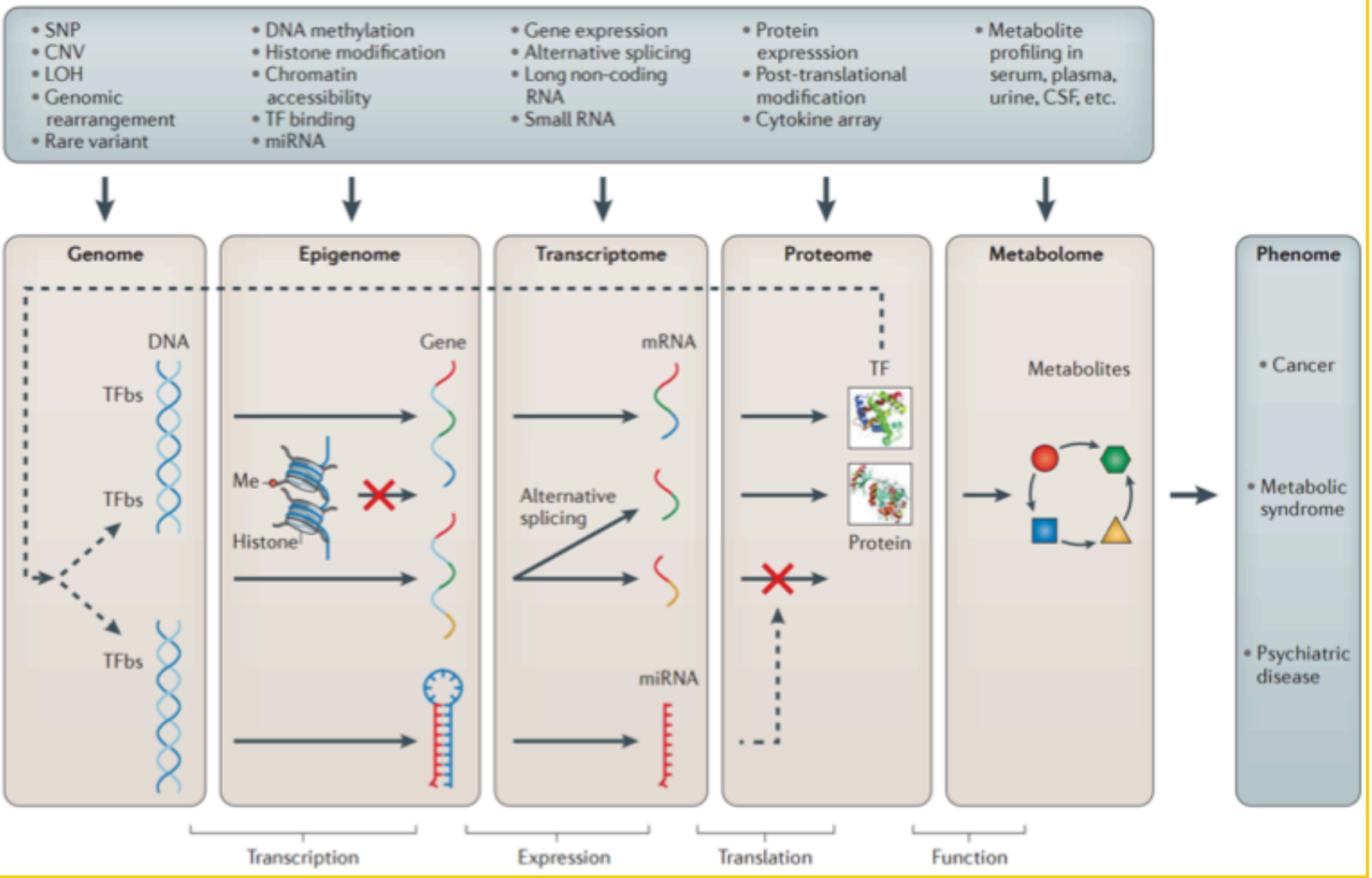
Systems Biology
"*Integrative approach in which scientists study pathways and networks will touch all areas of biology, including drug discovery*"
C. Henry and C. Washington

Dissecting the Biological system via -omics



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

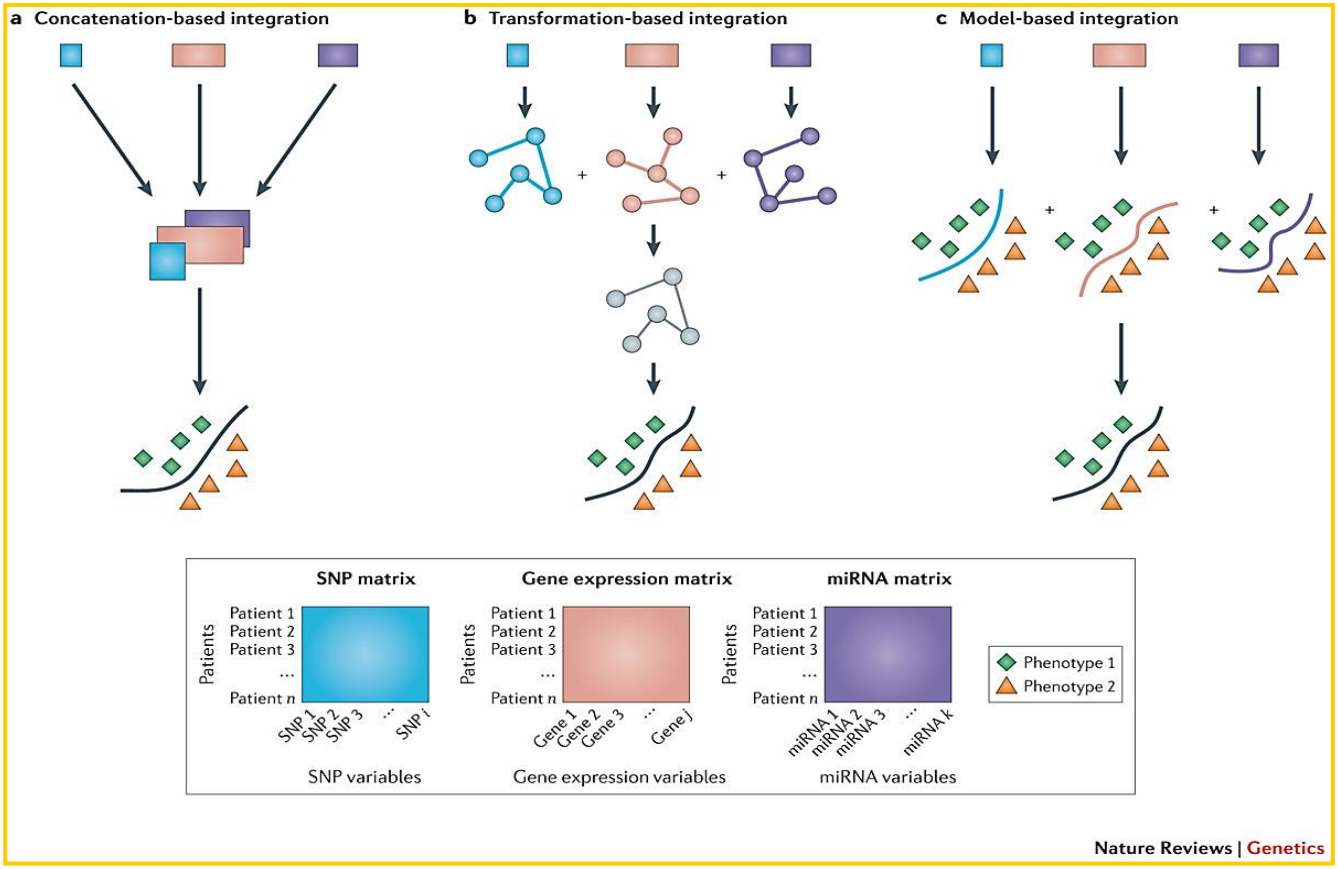
Paired integrative –omics analysis

- 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

- 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.

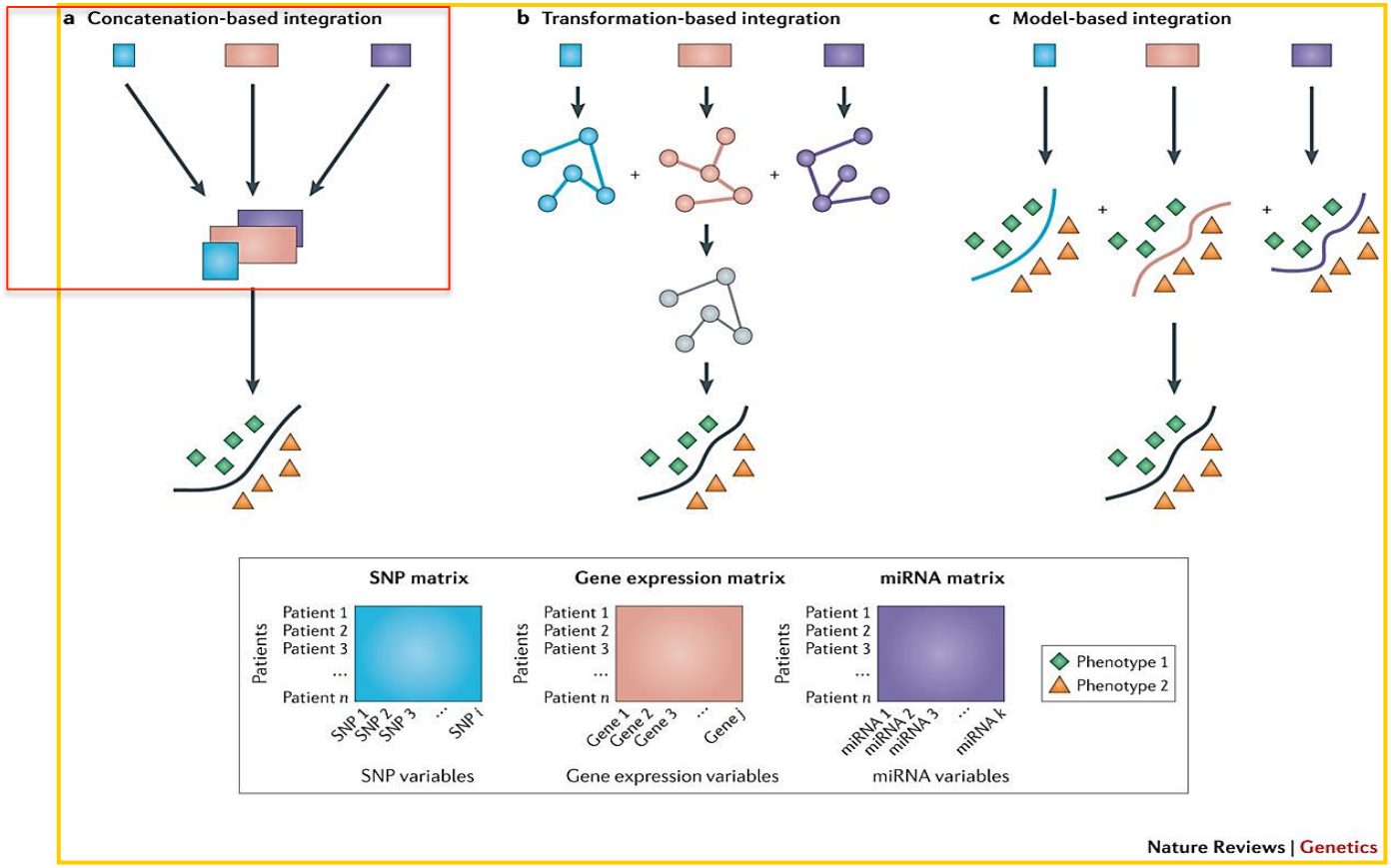
Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. 2015. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet *16:85-97.*

Categories of meta-dimensional omics integration

1) Concatenation-based integration:

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

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



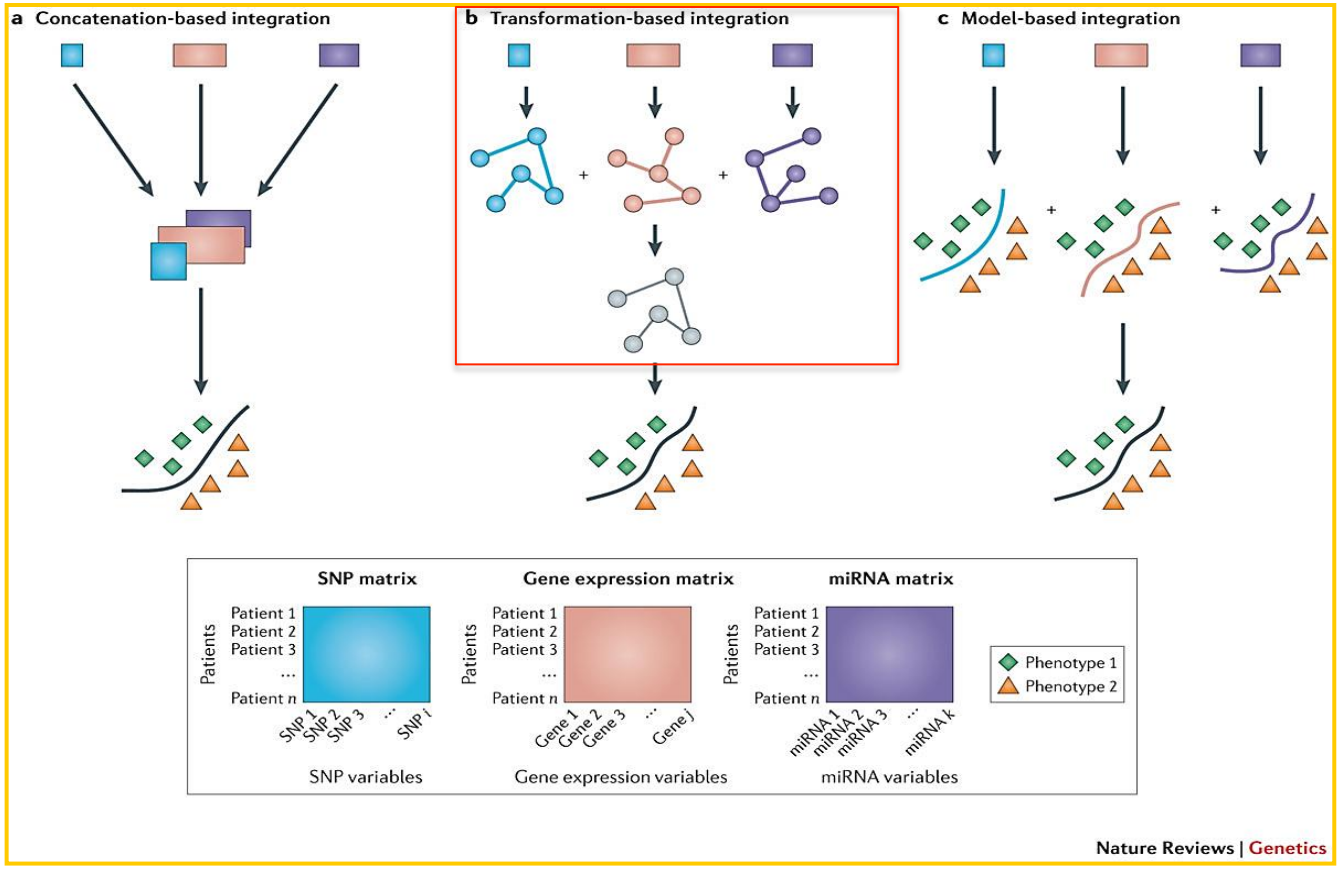
Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. 2015. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet *16:85-97.*

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



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.*

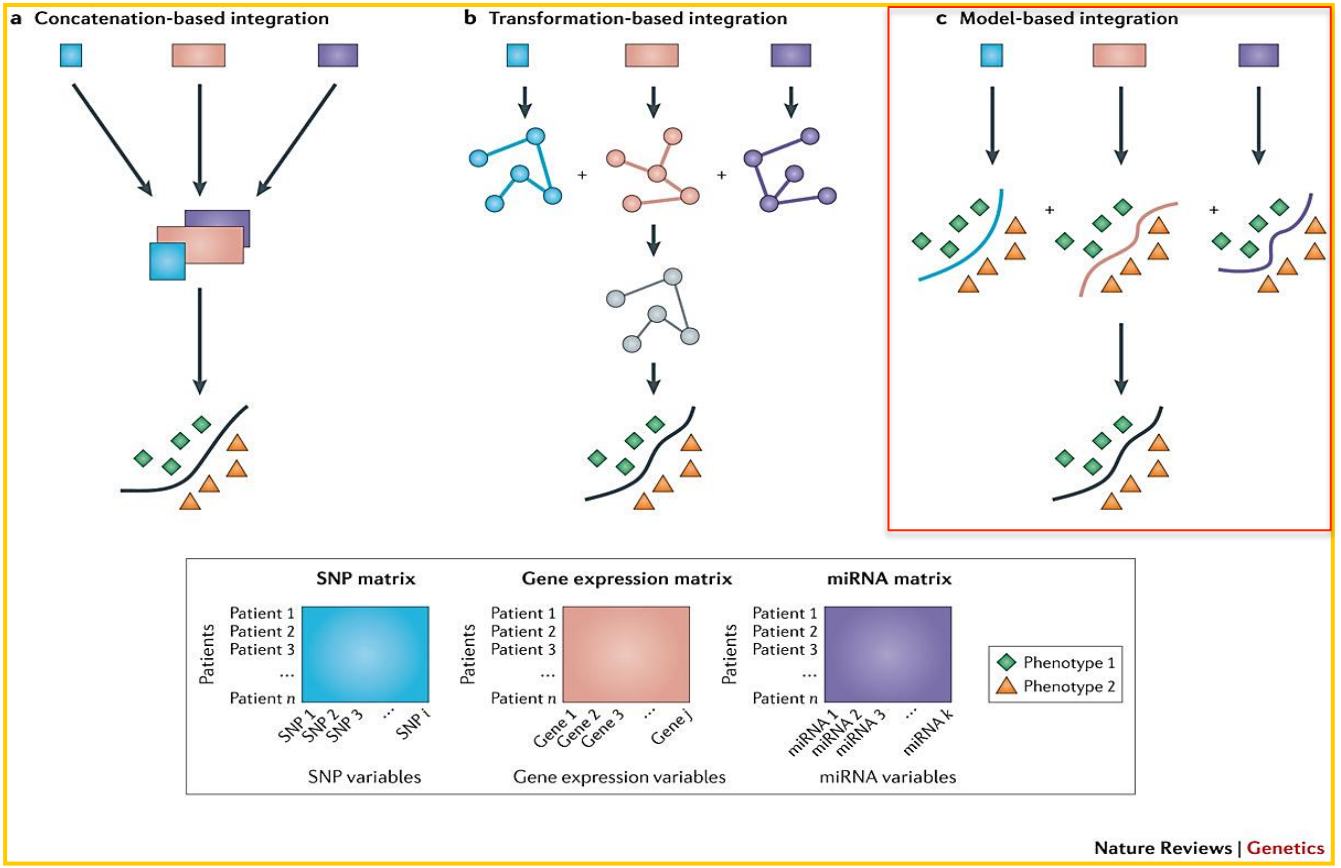
Categories of meta-dimensional omics integration

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 << number of variables)

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



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.*

Tools and techniques for multi-omics integration and relationship visualization

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

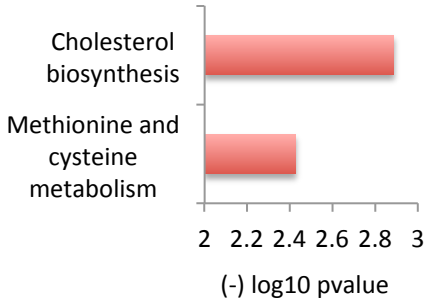
Univariate

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

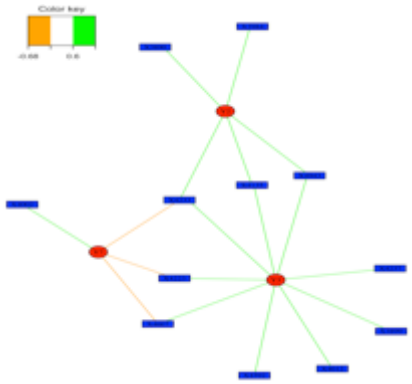
Multivariate

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

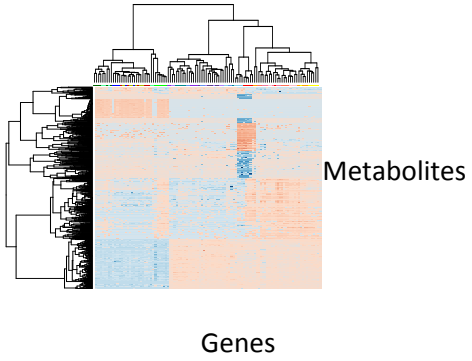
Pathway enrichment



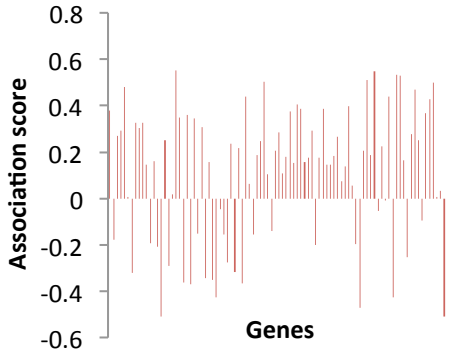
Relevance networks



Clustering

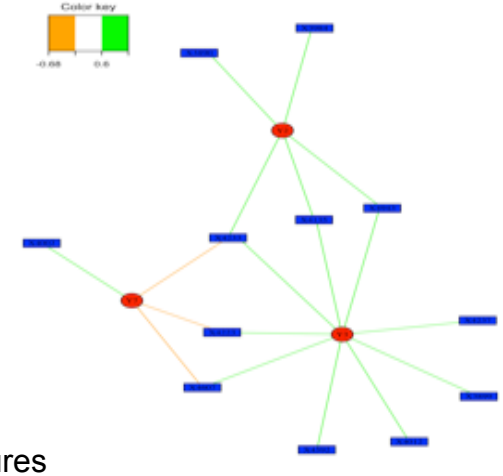


Targeted investigation
(e.g.: Arginine x Transcriptome)



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)
  - Software:
    - MetabNet (Uppal 2015; R package for performing pairwise correlation analysis and generating relevance networks)
    - 3Omics (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:
    - mixOmics (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

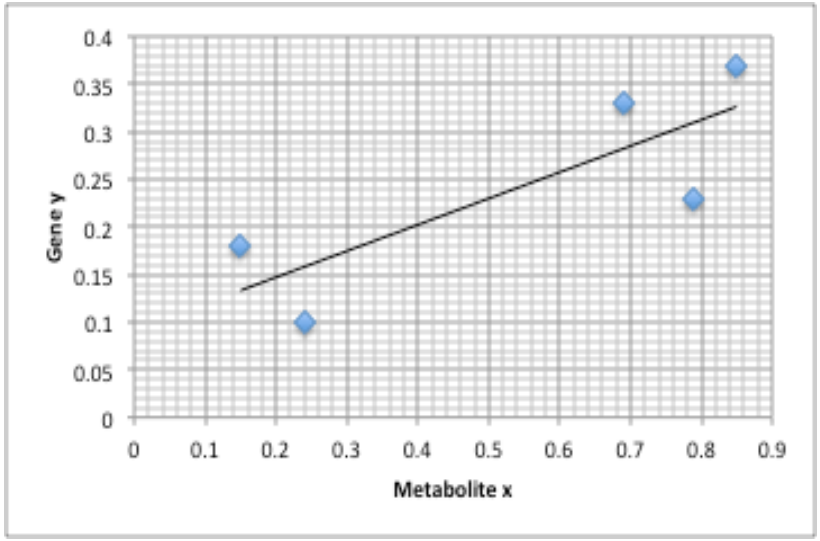
MetabNet (Uppal 2015)

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

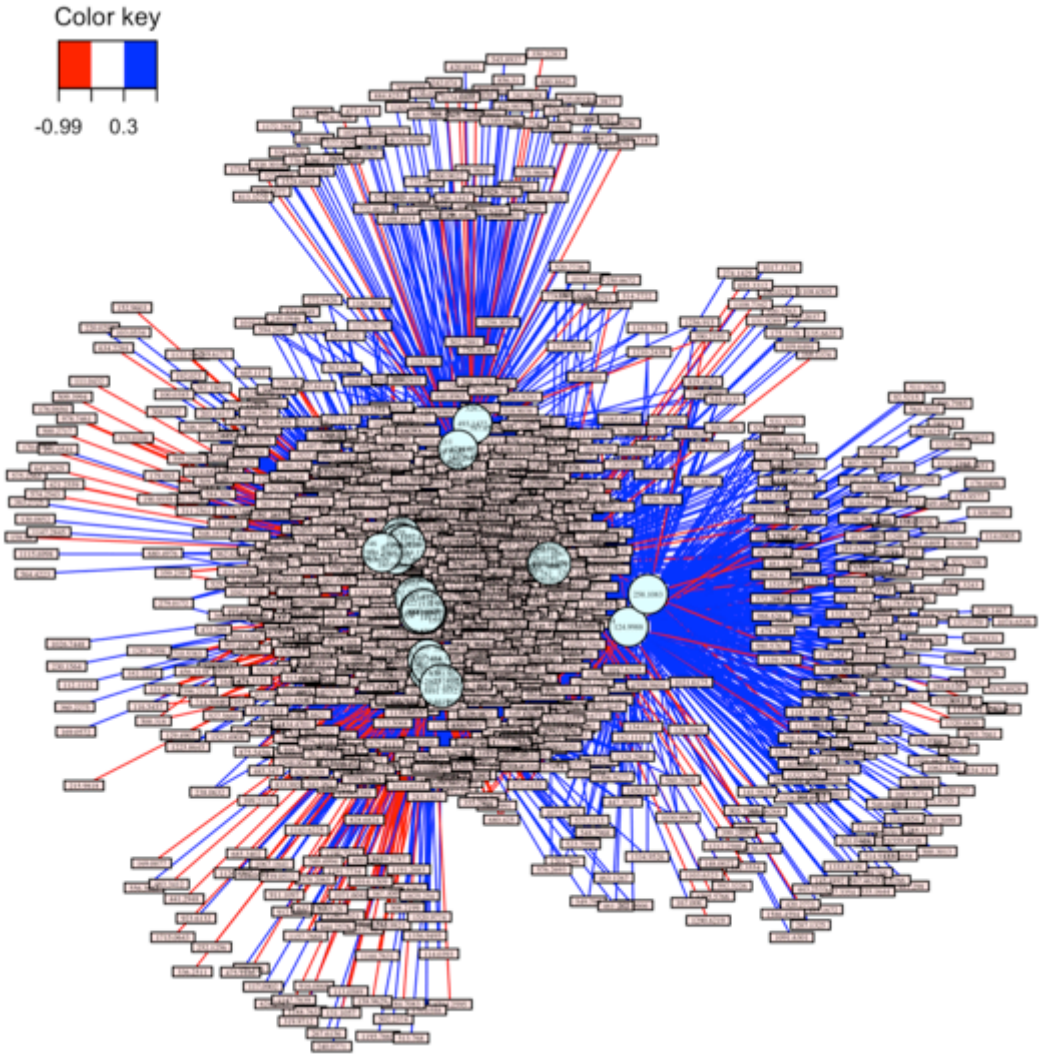
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (1)$$

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

where,
x,y -> different omics data
n -> number of samples
r -> Pearson Correlation
t -> t-statistic



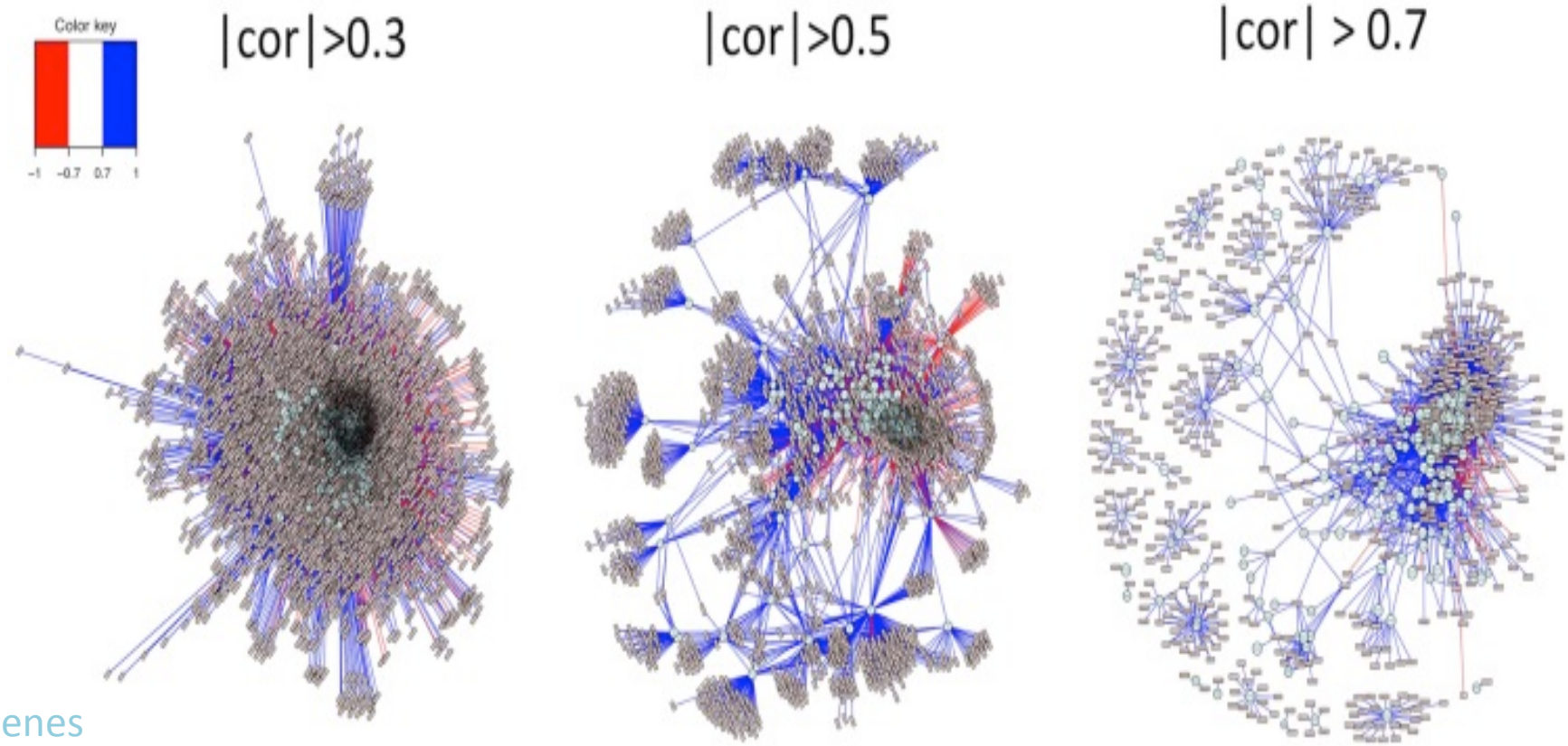
MetabNet output



Genes
Metabolites

MetabNet output at different correlation thresholds

Increasing stringency



MetabNet R package

- Availability: Software and tutorial available on sourceforge (https://sourceforge.net/projects/metabnet/)
- Caveats:
  - Large number of possible associations (p x q)
    - E.g.: 2x10^8 possible associations for 20,000 genes x 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

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

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

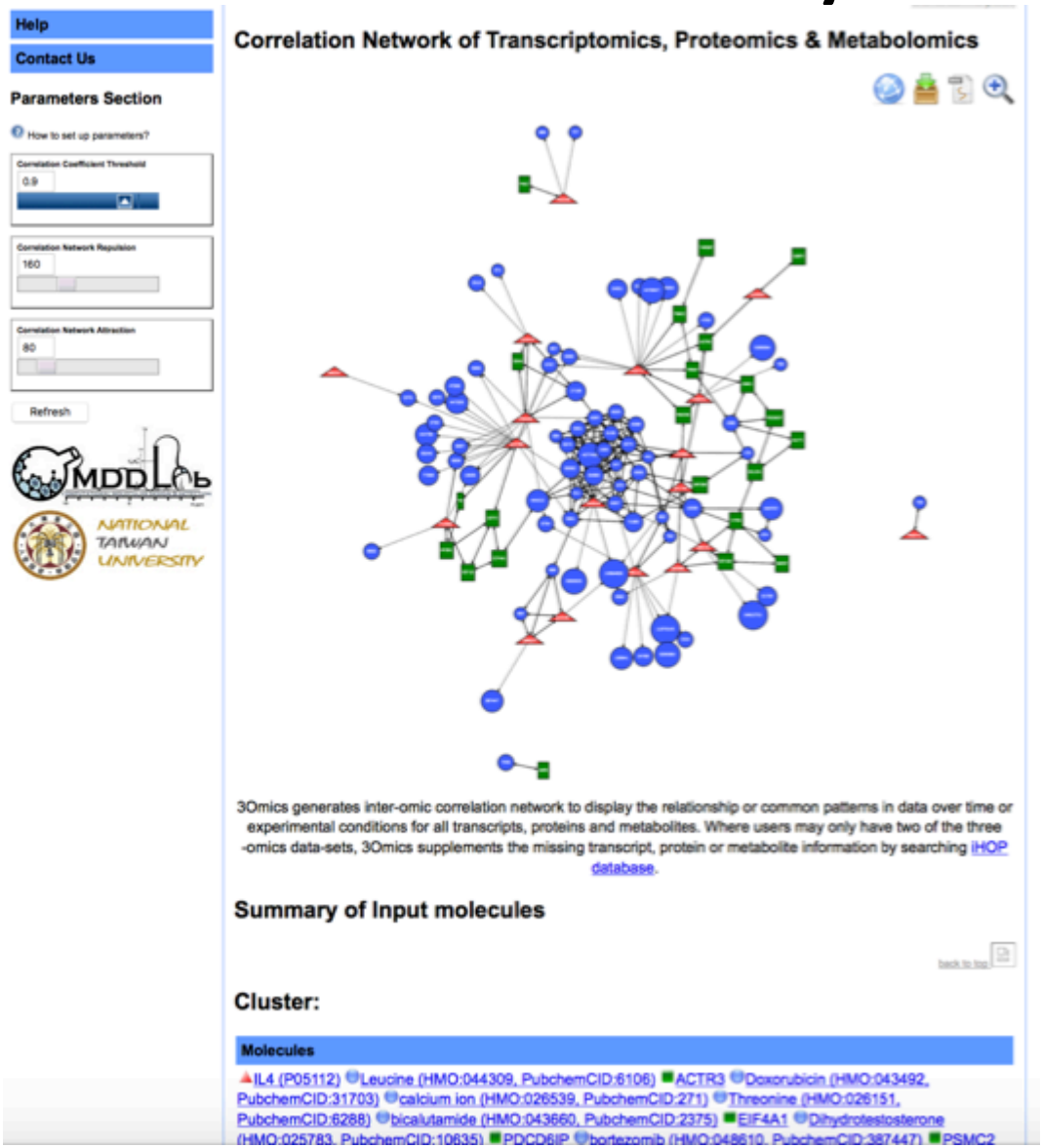
(http://3omics.cmdm.tw/help.php#examples)

Samples

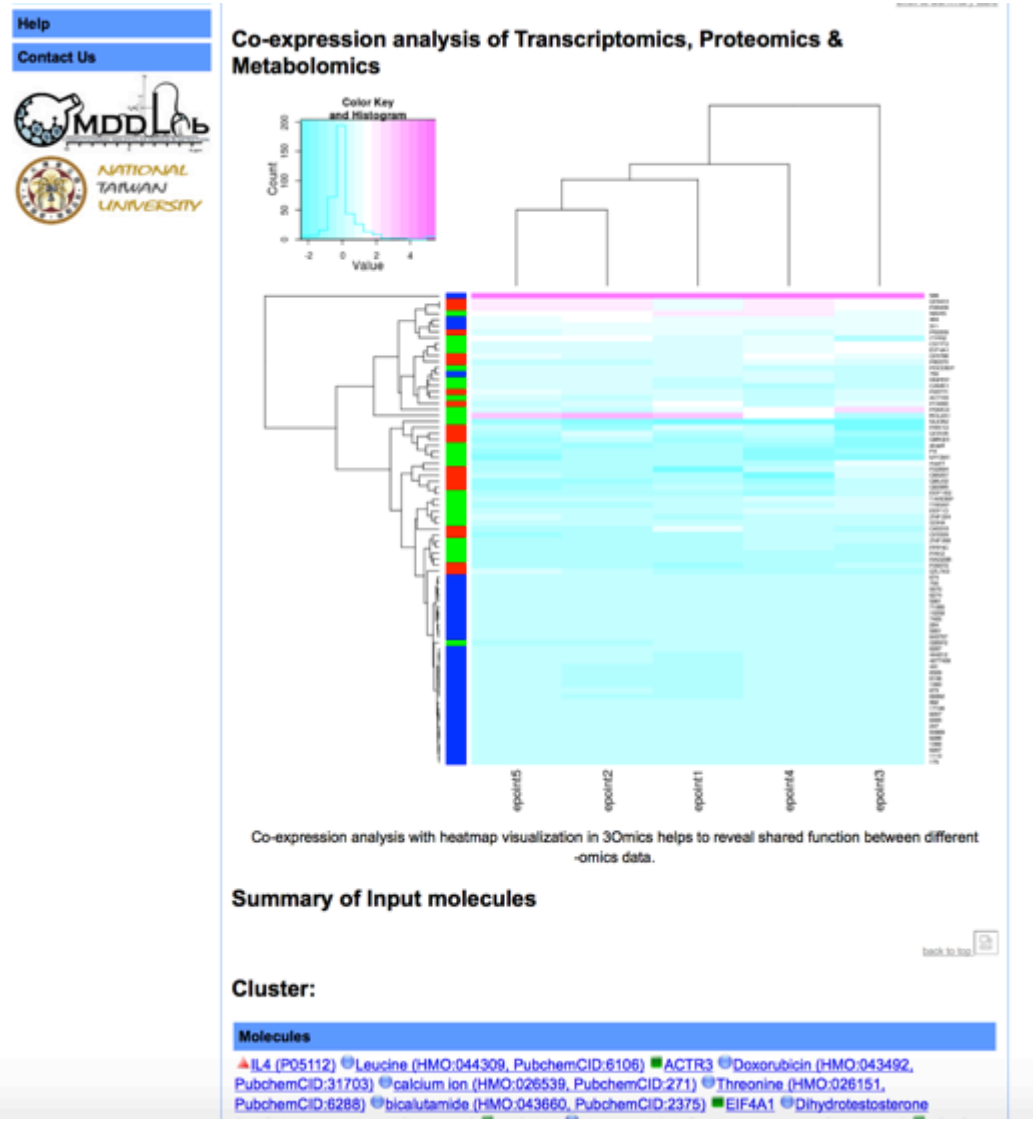
Variables

	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



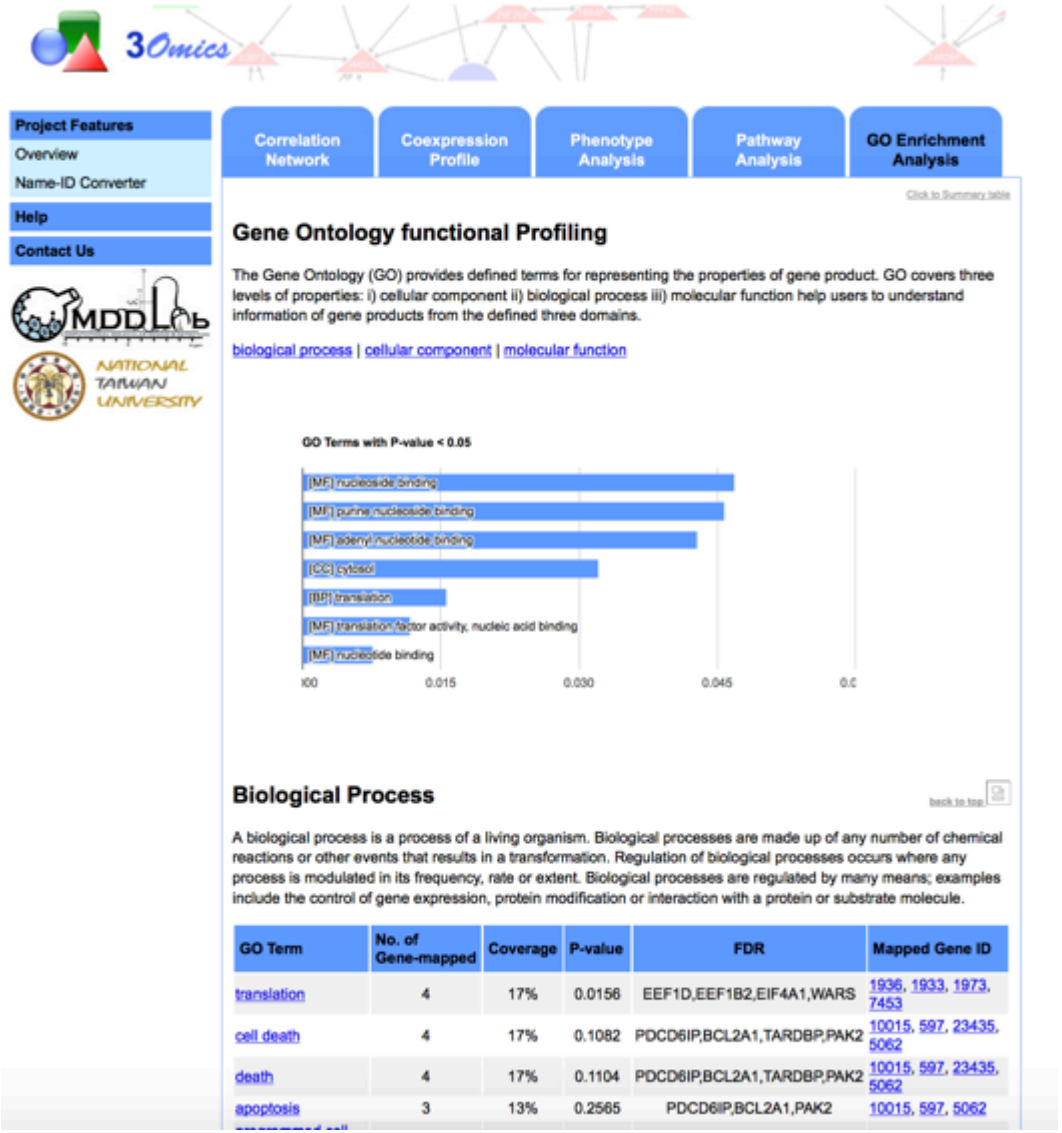
Rows: Variables

Columns: Samples

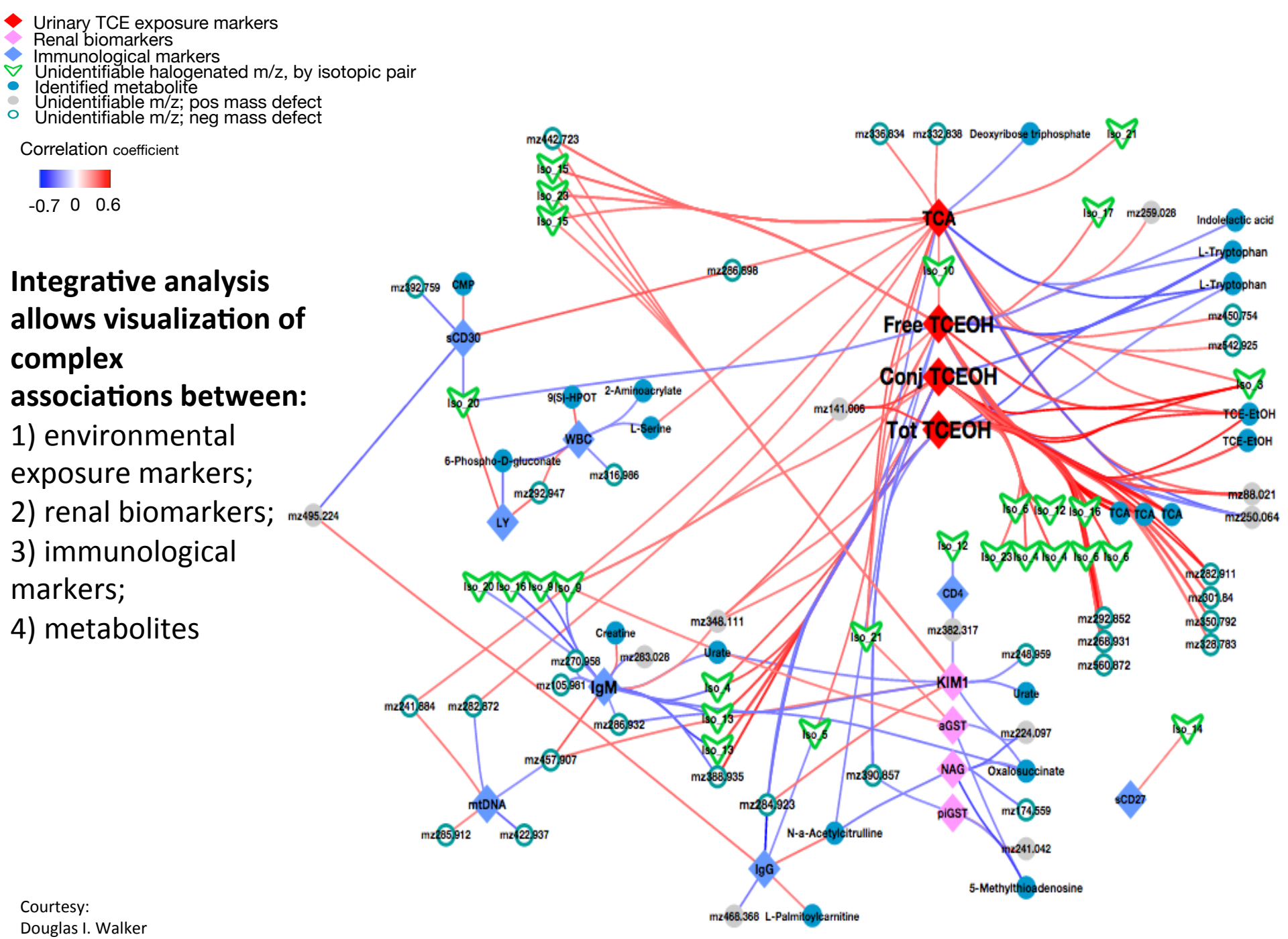
Phenotype analysis

Pathway analysis

GO Enrichment Analysis



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)



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 $\mathrm{cov}(X_u, Y_v)$

  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

multilevel sPLS method for experiments with repeated measurements

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

1) Split-up variation:

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

2) sparse PLS objective function:

$\max cor(Y, X_u) var(X_u)$

where

Y is the matrix indicating group of each sample

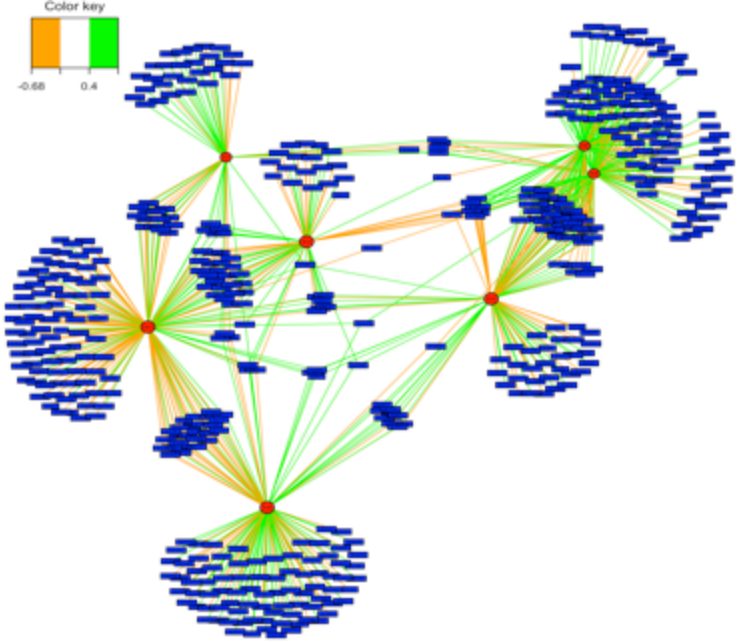
X is the split-up variation

$u_1, u_2 \ldots u_H$ 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

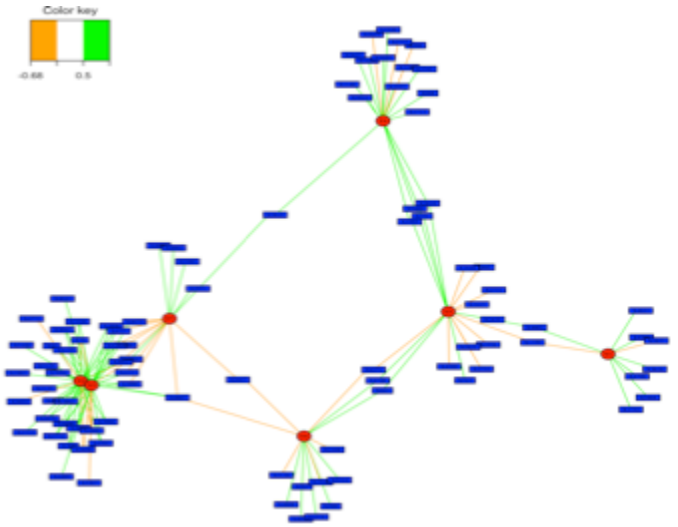


A. Association threshold: 0.4

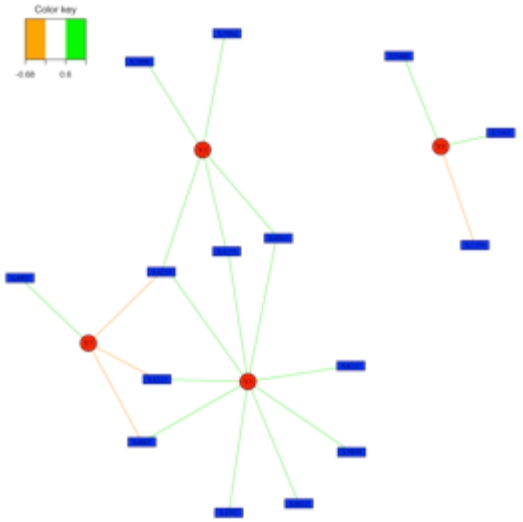
Legend

Circles: targeted bile acid

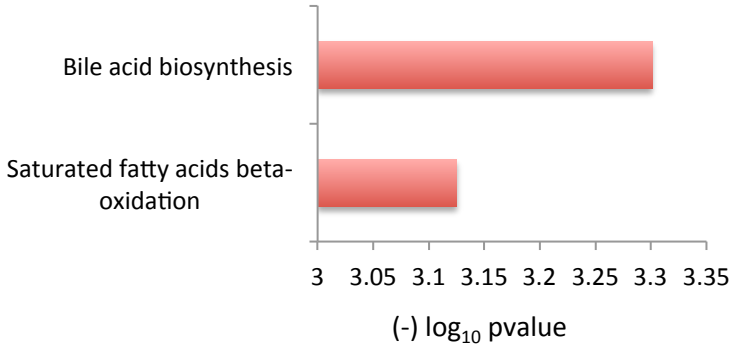
Rectangles: metabolome features



B. Association threshold: 0.5



C. Association threshold: 0.6



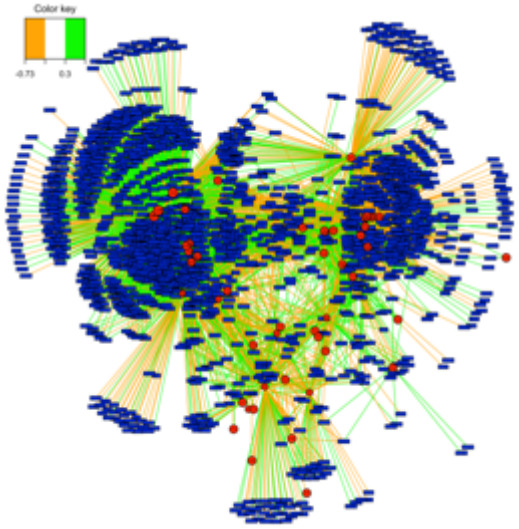
D. Pathway analysis (only top two displayed)

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)

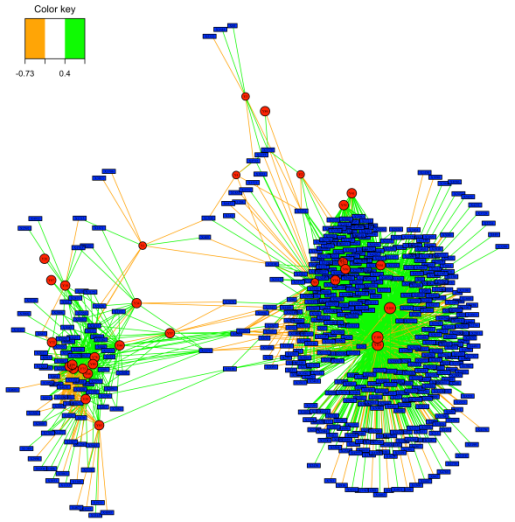
Legend

Circles: microbial species

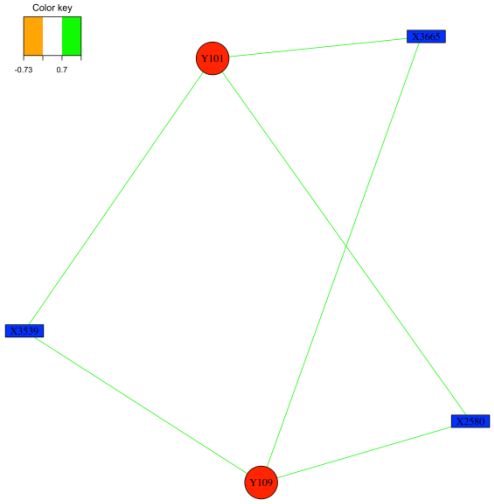
Rectangles: metabolome features



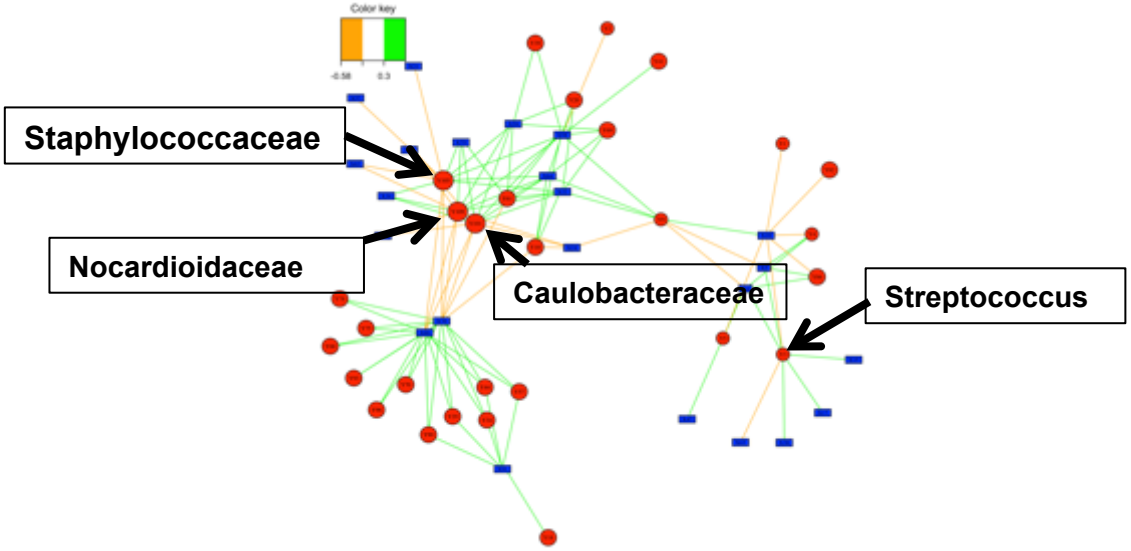
A. Association threshold: 0.3



B. Association threshold: 0.4



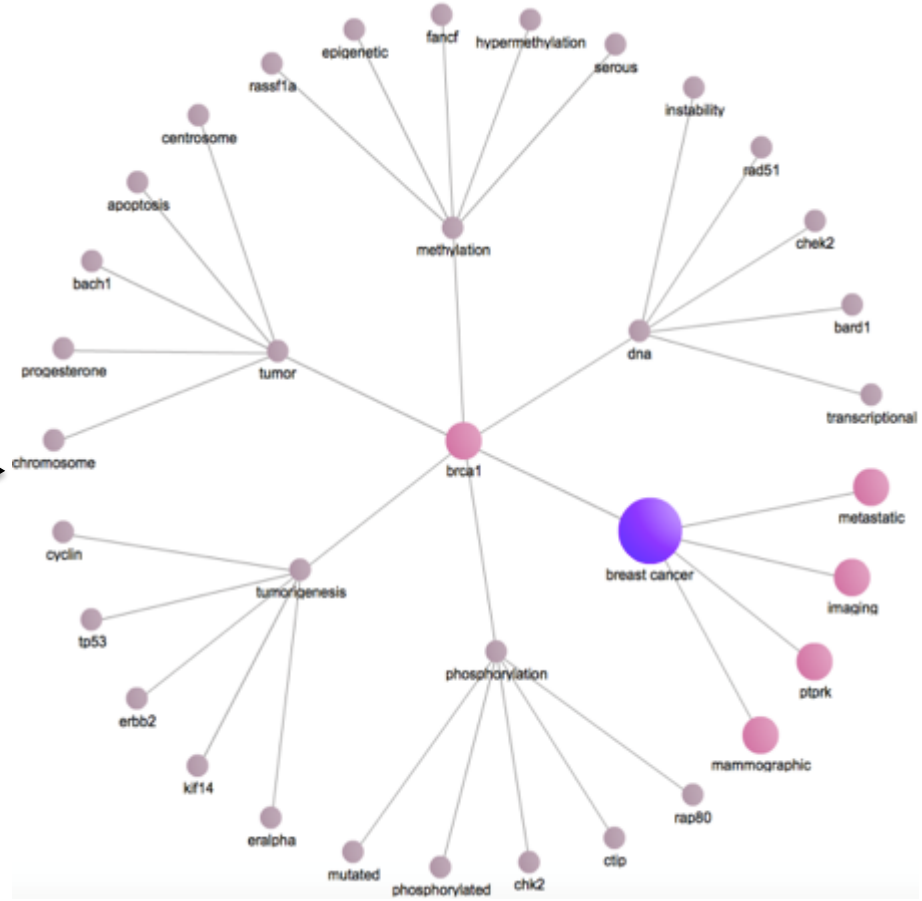
C. Association threshold: 0.7



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

Integrating data from other sources (e.g. PubMed)

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

Association mining algorithm for constructing relation trees

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

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

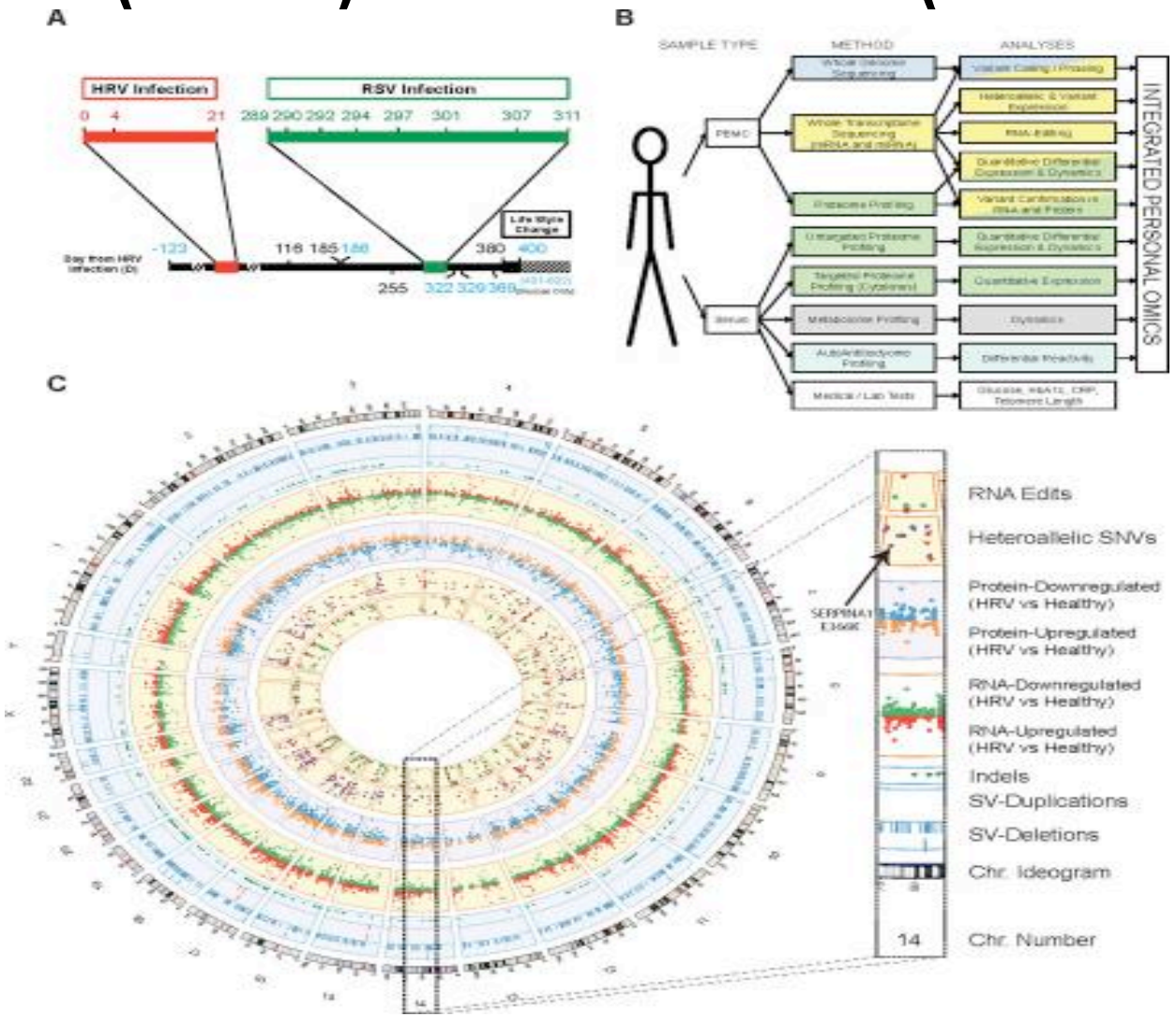
$p(t_1 \text{ and } t_2)$ is the probability of co-occurrence of terms 1 and 2 in the corpus

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

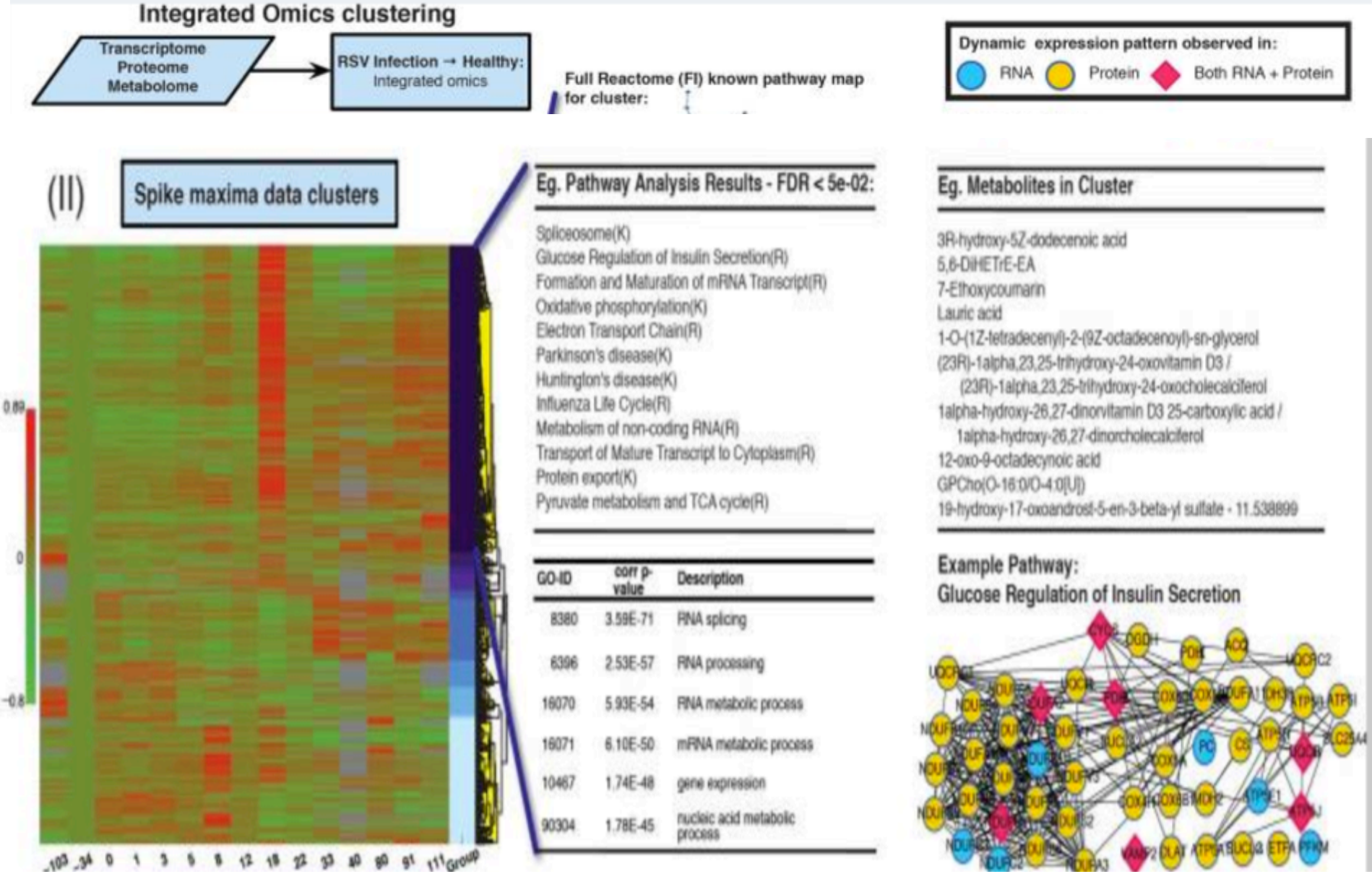
Integrative Personal Omics Profiling (iPOP) – Chen et al. (Cell 2012)



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)



Questions?