Computational methods for integrative omics and relation discovery between biomedical entities

Feb 11, 2016

Karan Uppal (PhD) and Sophia A. Banton (MS, PhD candidate)

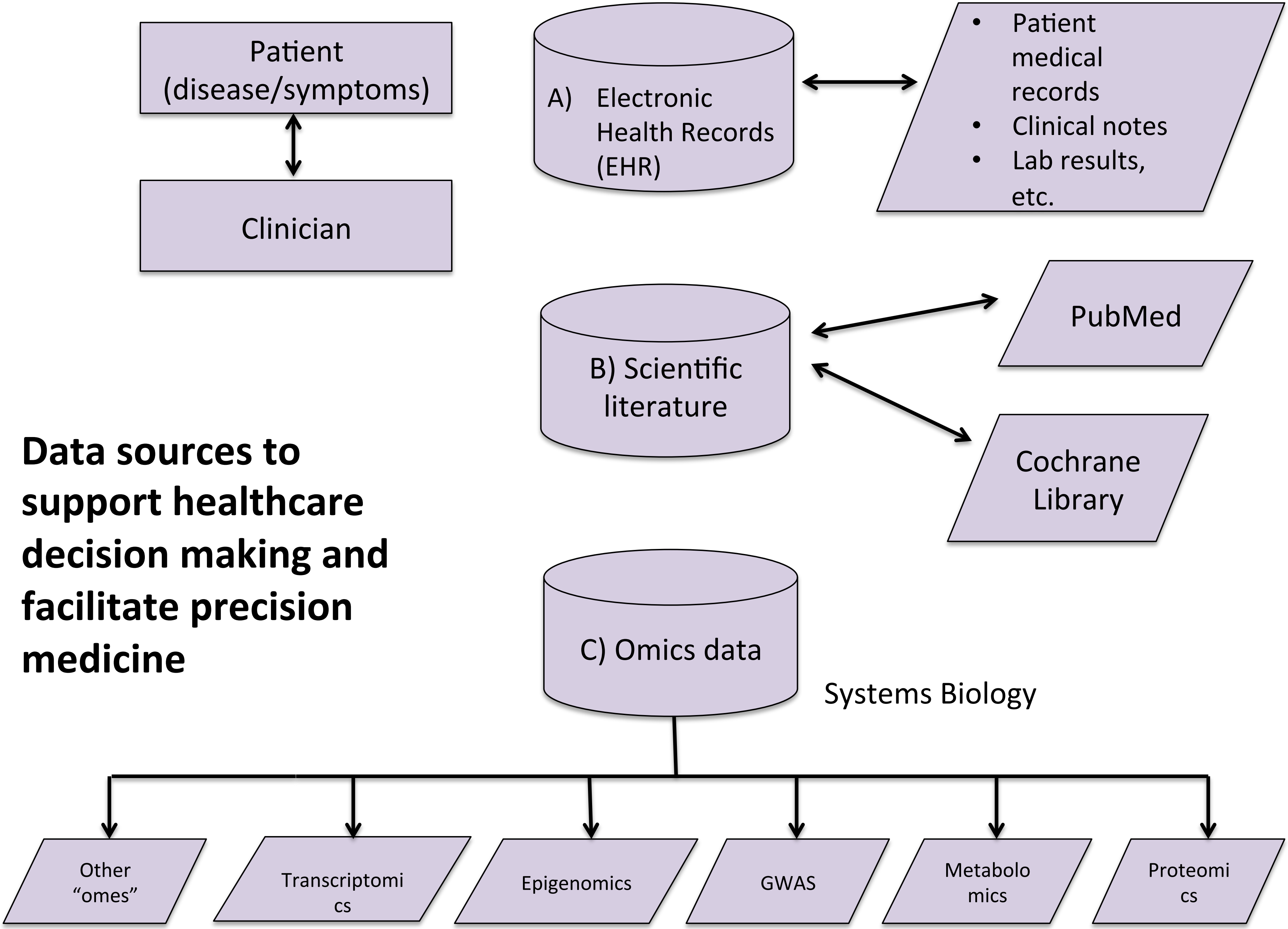
Clinical Biomarkers Laboratory

Emory University School of Medicine

Learning Objectives

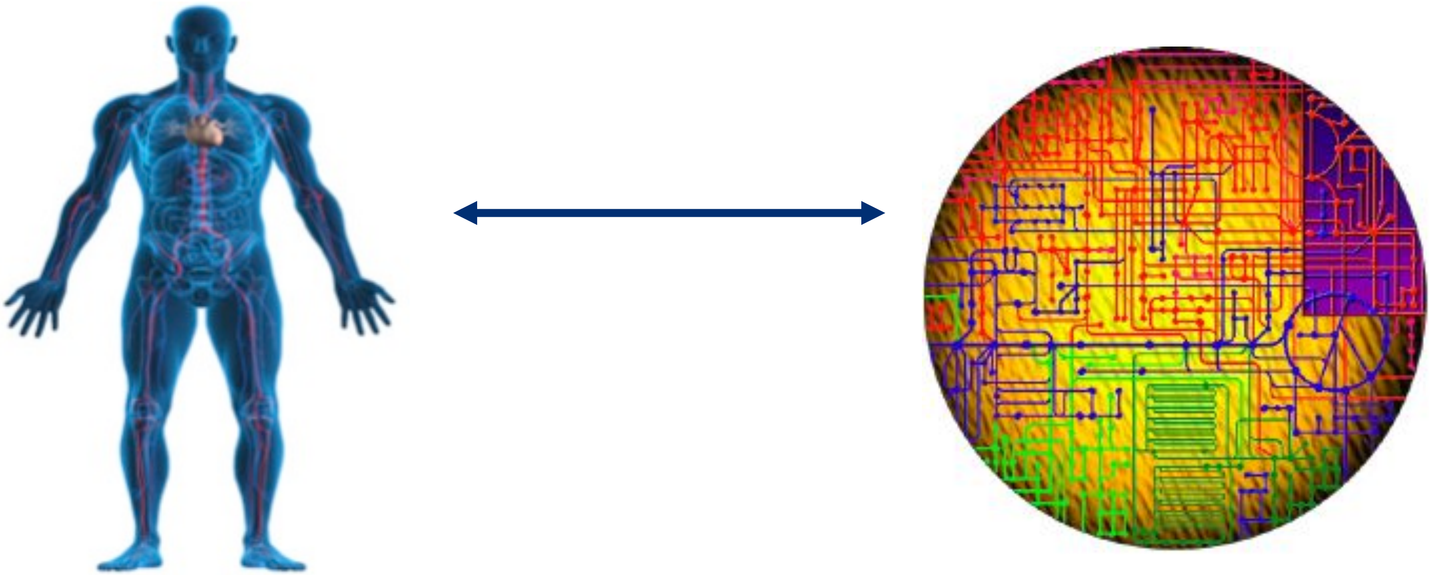
- Data-driven methods for integrating paired –omics data and visualizing associations (Karan Uppal)
- Knowledge-driven methods for integrating paired –omics data (Sophia Banton)

Data sources to support healthcare decision making and facilitate precision medicine



Introduction: A Systems Biology Framework

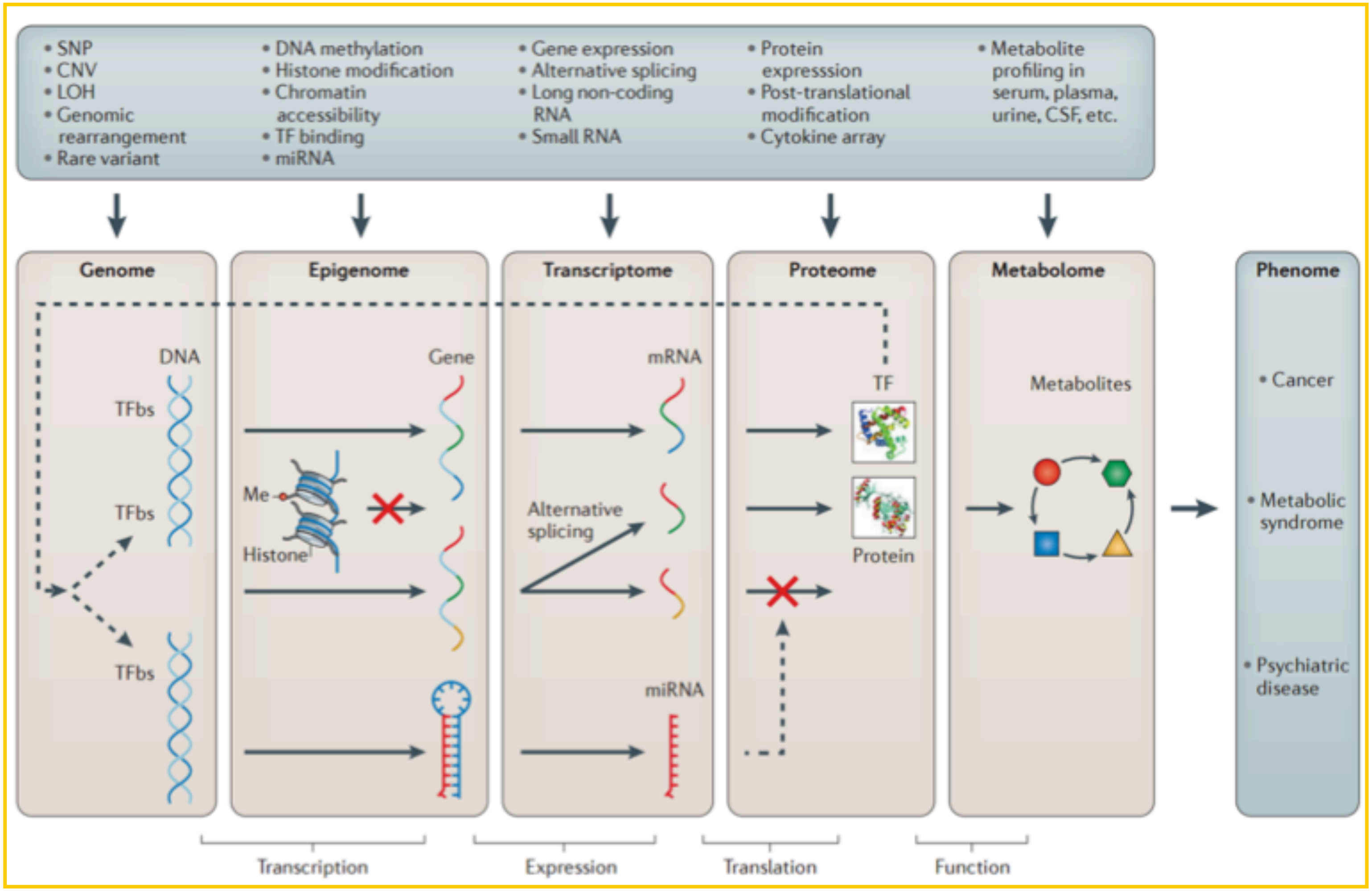
- The goal of **Systems Biology**:
  - Systems-level understanding of biological systems
  - Analyze not only individual components, but their interactions as well and emergent behavior



Exposures
Internal measurements
Disease states

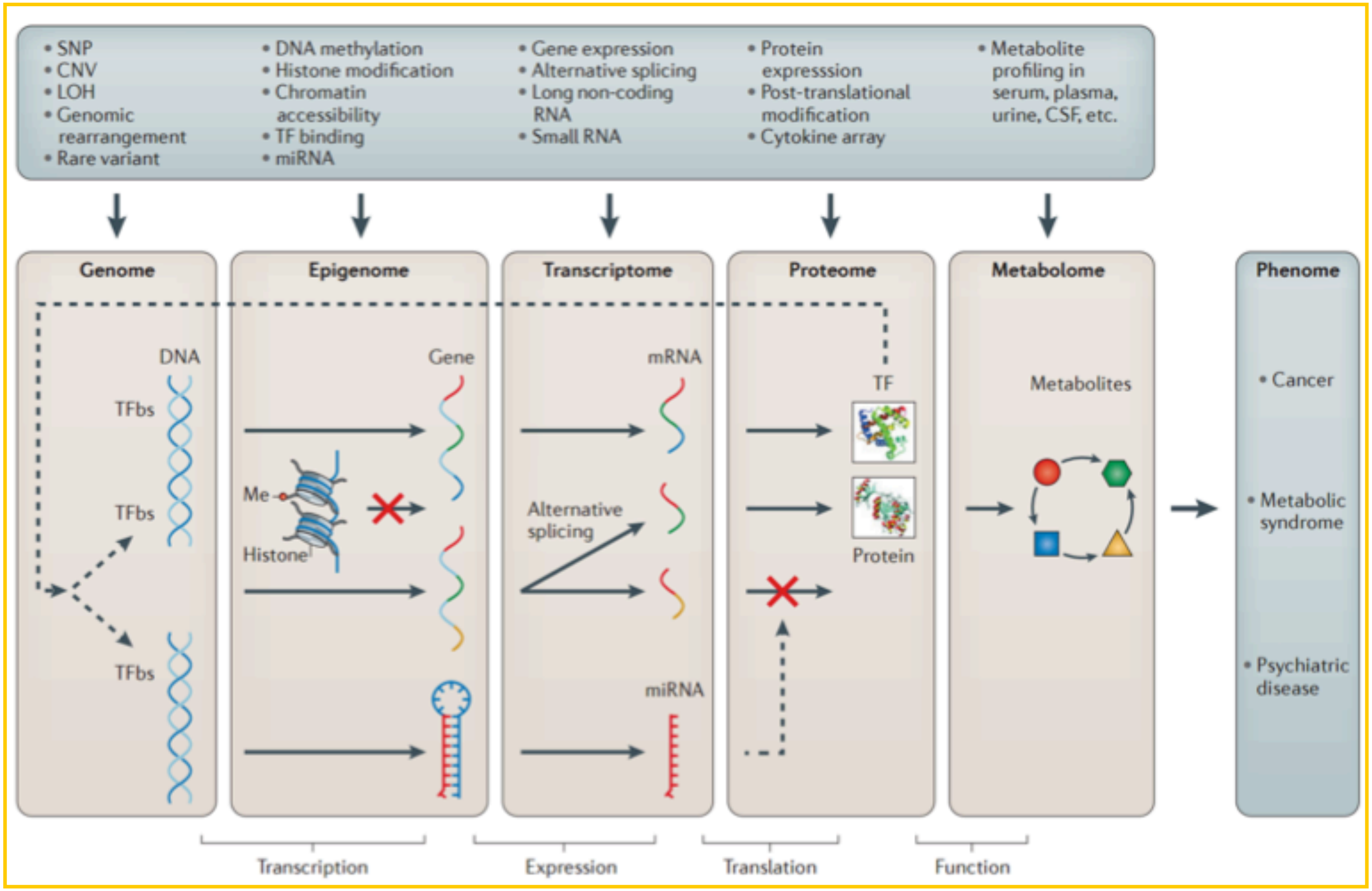
Systems Biology
“*Integrative approach in which scientists study pathways and networks will touch all areas of biology, including drug discovery*”
C. Henry and C. Washington

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.*

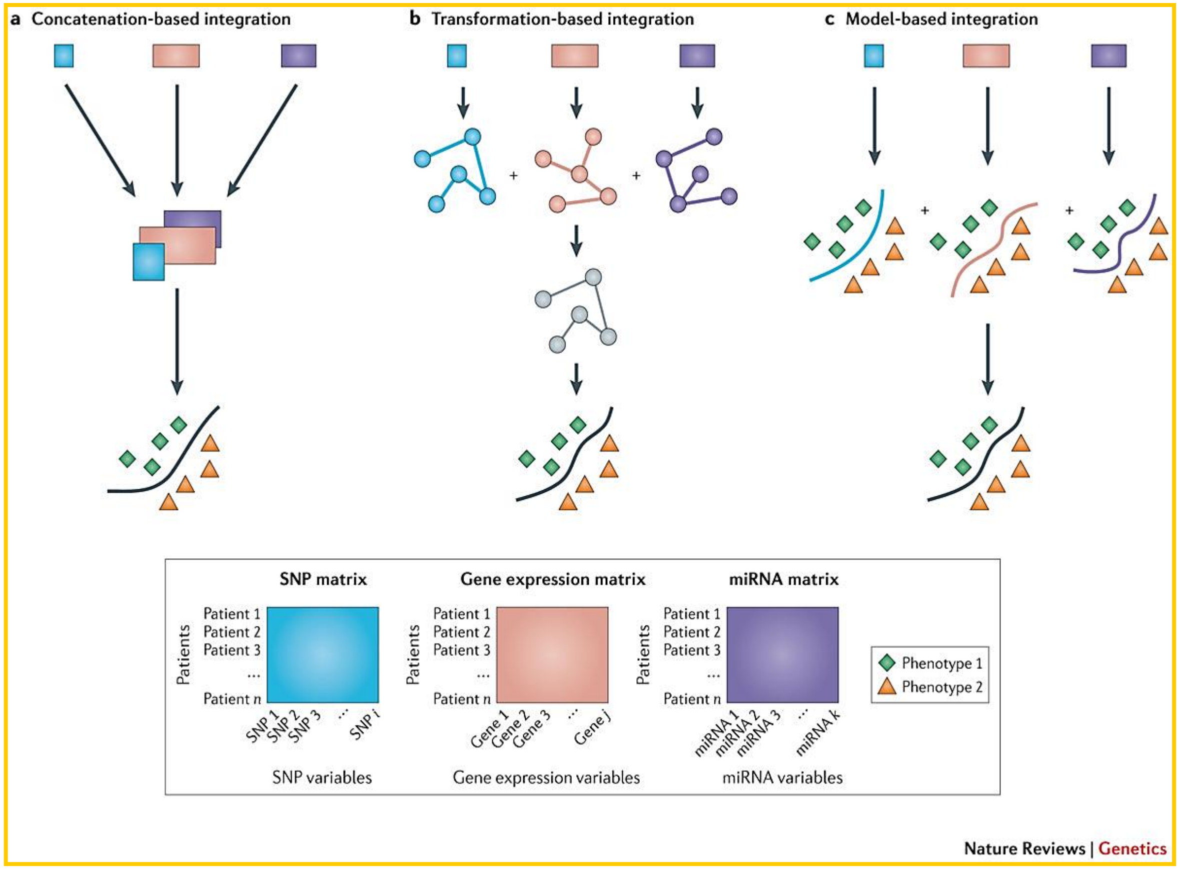
Dissecting the Biological system via -omics



"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.*

Methods of 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.*

Data-driven methods for integration

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

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

Workflow

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

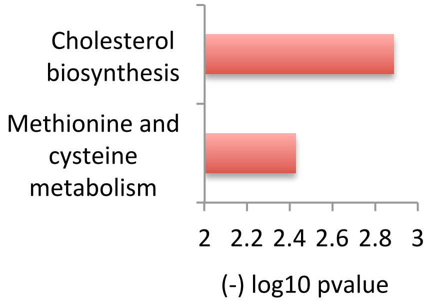
Univariate

- Pearson Correlation
- MetabNet (Uppal2015)

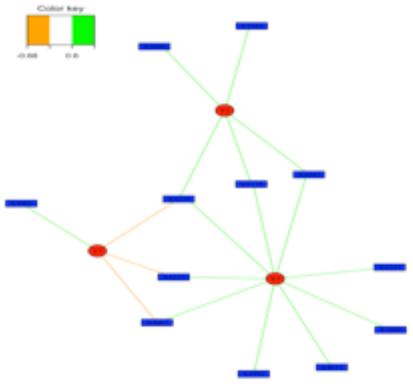
Multivariate

- PLS, CCA, sparse PLS
- mixOmics (Cao 2009)

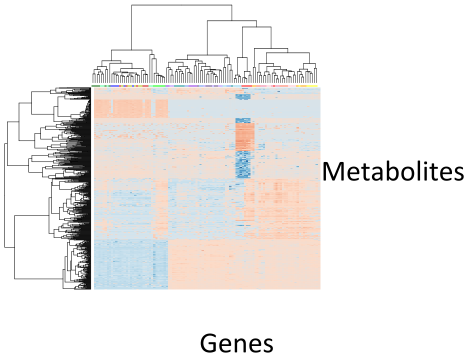
Pathway enrichment



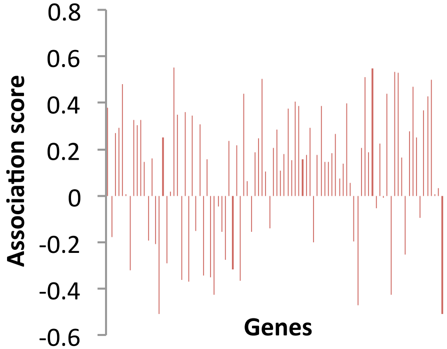
Relevance networks



Clustering



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



Metabolomics data
(n subjects X p metabolites)

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

Microbiome data
(n subjects X q bacterial species)

	B1	B2	-	Bn
Subject1	19	19	-	100
Subject2	10	40	-	90
-	-	-	-	-
SubjectN	10	40	-	50

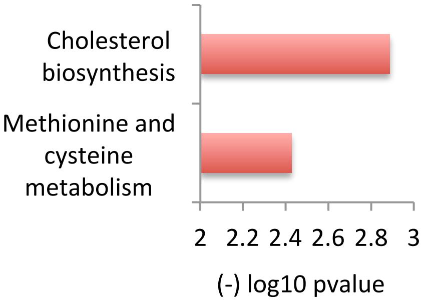
Association matrix

	B1	B2	-	Bn
M1	0.4	0.9	-	0.3
M2	0.7	0.1	-	0.5
M3	0.1	0.6		0.8

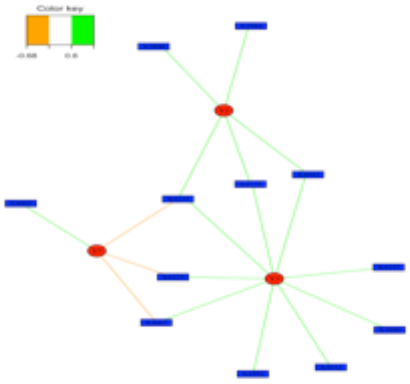
Univariate
- Pearson Correlation
- MetabNet (Uppal2015)

Multivariate
- PLS, CCA, sparse PLS
- mixOmics (Cao 2009)

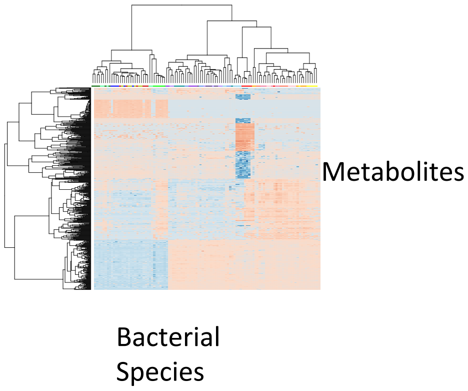
Pathway enrichment



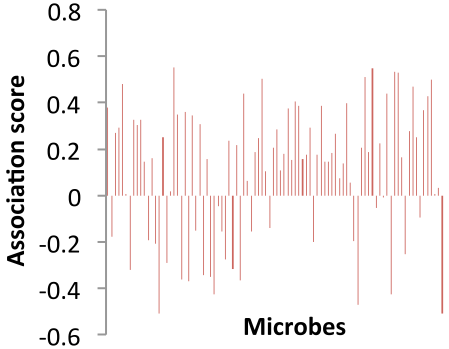
Relevance networks



Clustering

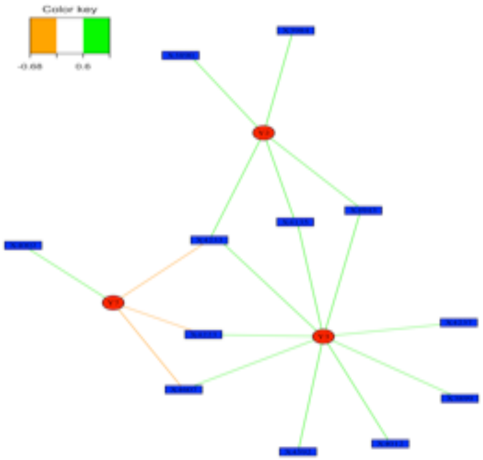


Targeted investigation



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 2001; 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

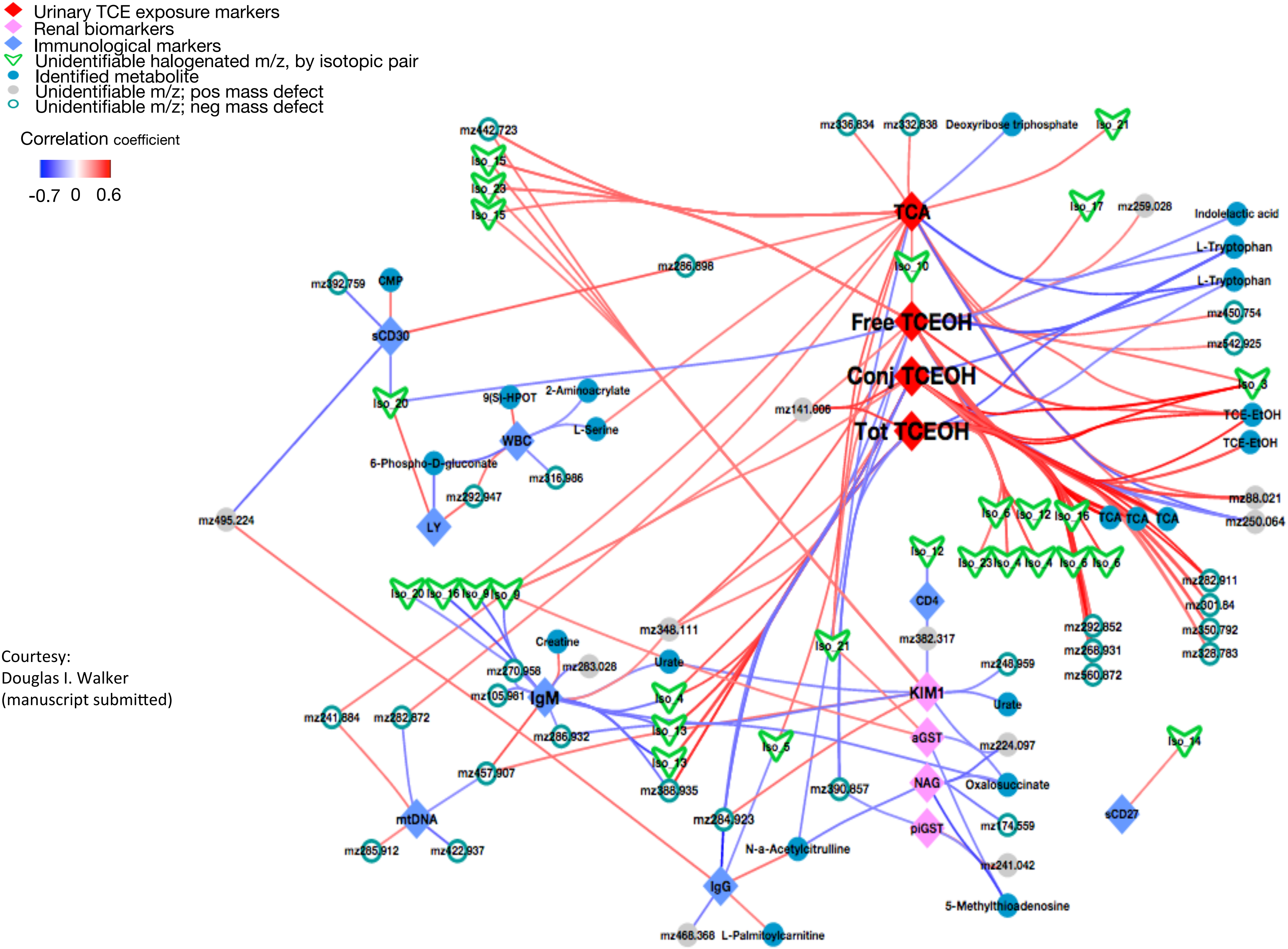
- Univariate
  - Pairwise Pearson or Spearman correlation between data from different biomedical/clinical technologies (Butte et al. 2009, Uppal et al. 2015)
  - Software:
    - MetabNet (Uppal 2015; R package for performing pairwise correlation analysis and generating relevance networks)
  - Application: 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 (R package; Uppal 2015)

- Performs pairwise correlation (Pearson or Spearman) or partial correlation analysis to generate association matrix (p x q) and relevance network using the data measure on same N
- 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
- More suitable when number of variables in at least one layer (p or q) is small
- Availability: Software and tutorial available on sourceforge (https://sourceforge.net/projects/metabnet/)

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: 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

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:

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

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

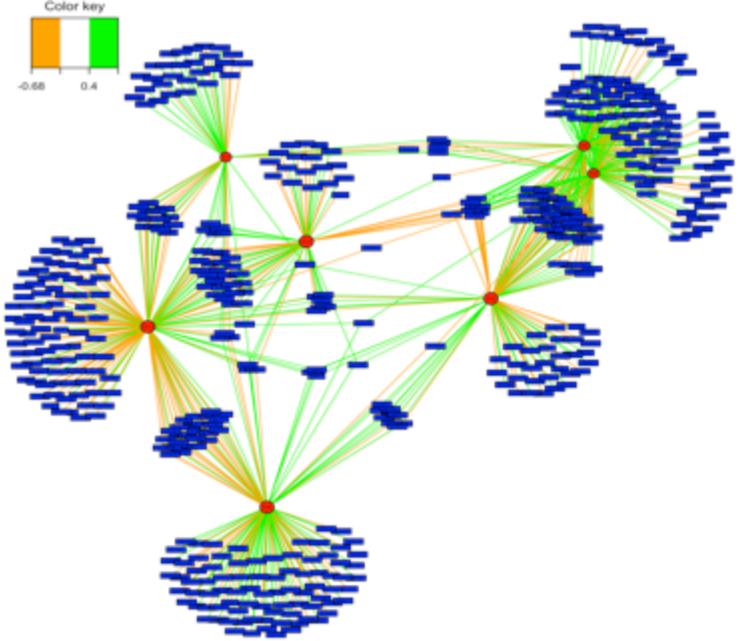
u_1, $u_2 \ldots u_H$ are the loading vectors

H is the number of PLS-DA dimensions

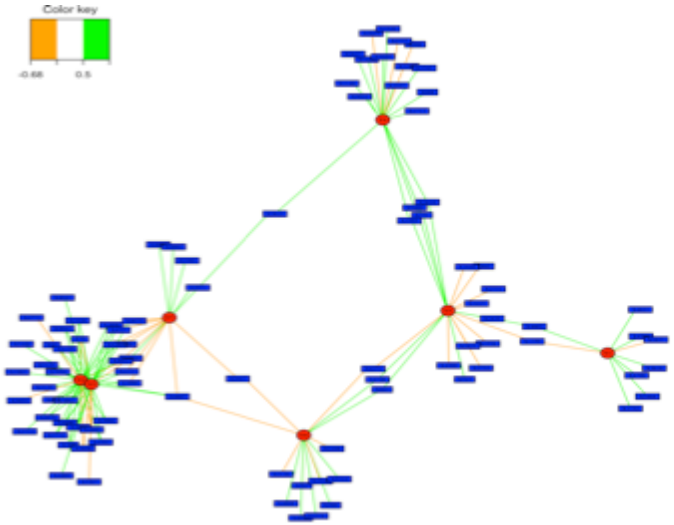
A Lasso based optimization is used to select most relevant variables

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

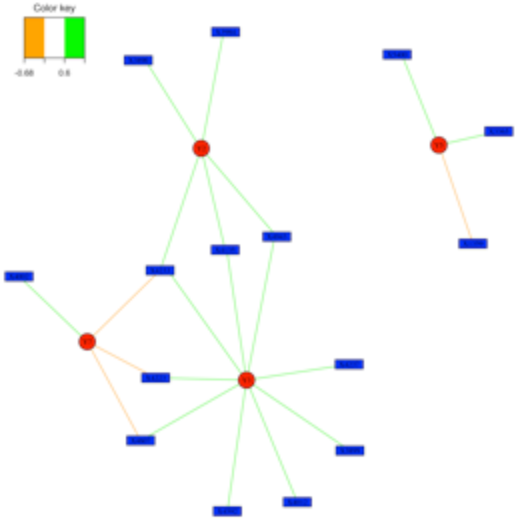
Legend
Circles: targeted bile acid
Rectangles: metabolome features



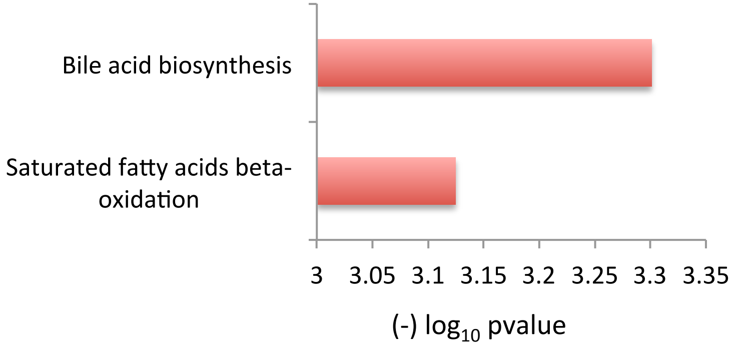
A. Association threshold: 0.4



B. Association threshold: 0.5



C. Association threshold: 0.6



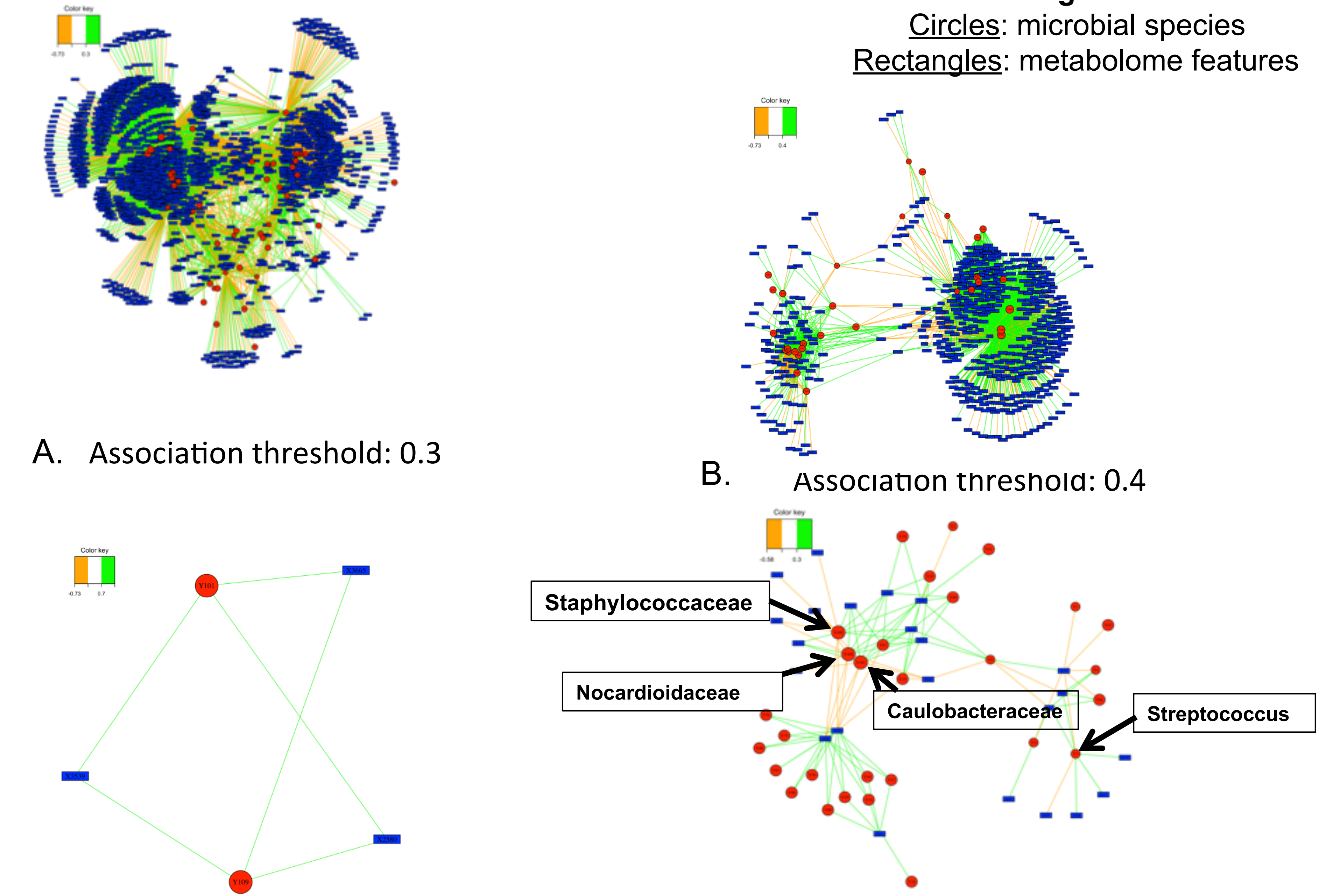
D. Pathway analysis (only top two displayed)

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

Legend

Circles: microbial species

Rectangles: metabolome features



A. Association threshold: 0.3

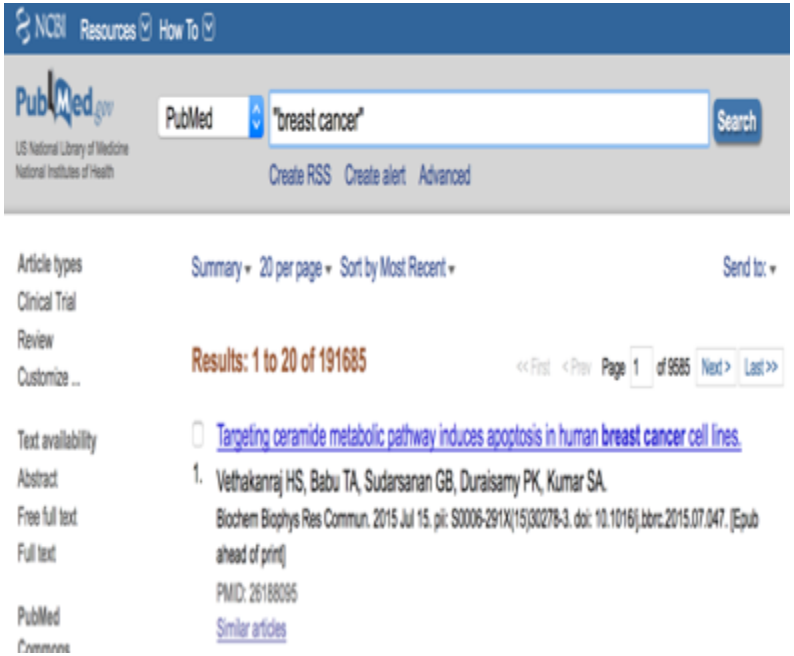
B. Association threshold: 0.4

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

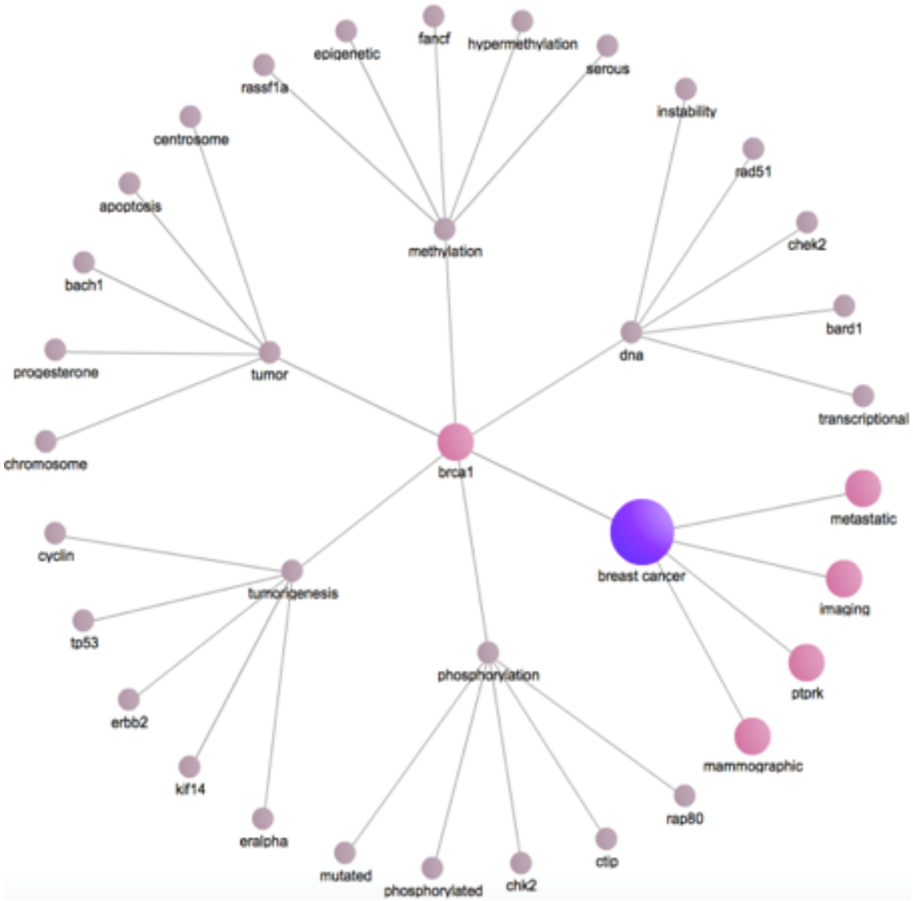


"breast cancer"

Results: 1 to 20 of 191685

□ Targeting ceramide metabolic pathway induces apoptosis in human **breast cancer** cell lines.

1. Vethakanraj HS, Babu TA, Sudarsanan GB, Duraisamy PK, Kumar SA.
Biochem Biophys Res Commun. 2015 Jul 15. pii: S0006-291X(15)30278-3. doi: 10.1016/j.bbrc.2015.07.047. [Epub ahead of print]
PMID: 26188095
Similar articles



Association mining based on co-occurrence

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

Association mining algorithm for constructing relation trees

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

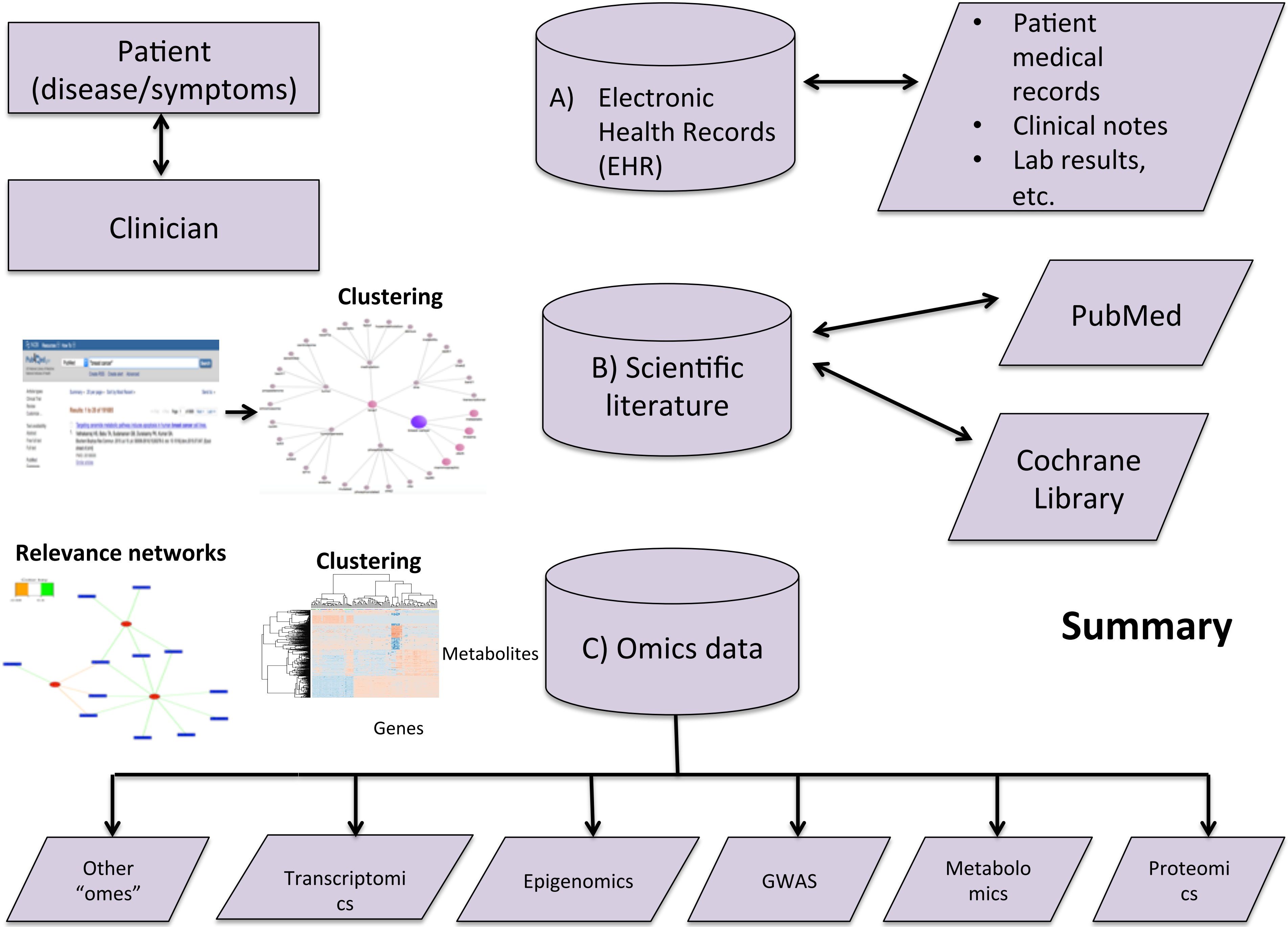
v_i is 1 if term t_2 is present in the controlled vocabulary, 0 otherwise;

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

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

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

Summary



Patient (disease/symptoms)

Clinician

A) Electronic Health Records (EHR)

- Patient medical records
- Clinical notes
- Lab results, etc.

B) Scientific literature

PubMed

Cochrane Library

Clustering

C) Omics data

Relevance networks

Clustering

Metabolites

Genes

Other “omes”

Transcriptomics

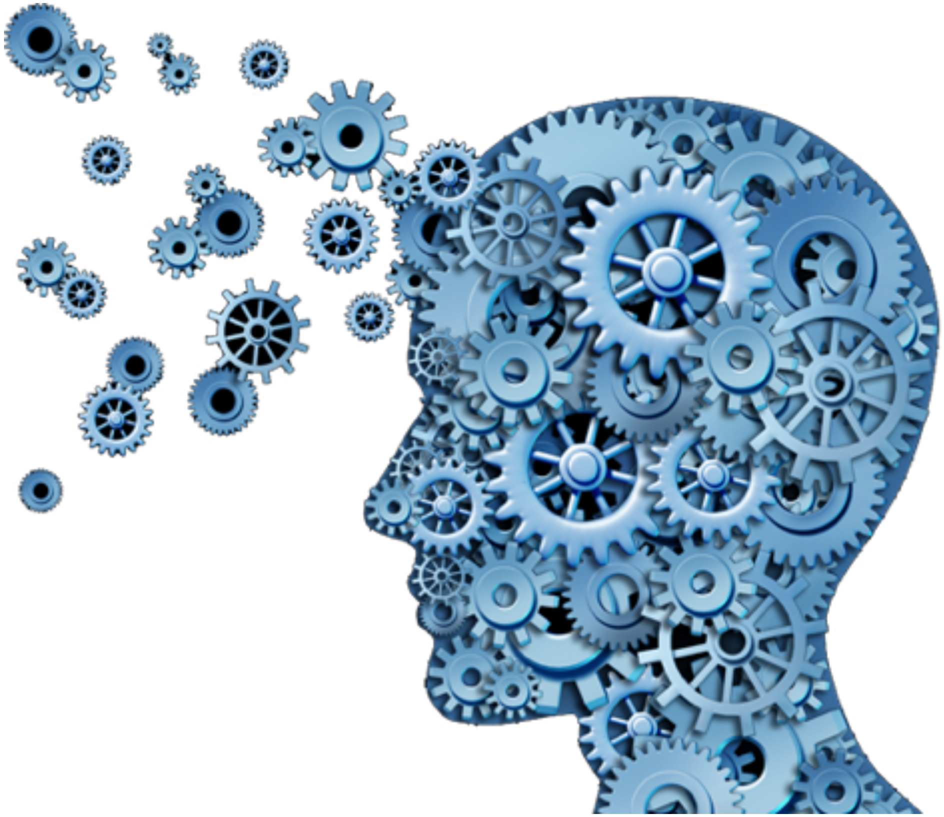
Epigenomics

GWAS

Metabolomics

Proteomics

Knowledge-Based Approaches of Data Integration



Sophia A. Banton, Shuzhao Li
Clinical Biomarkers Laboratory
Emory University School of Medicine

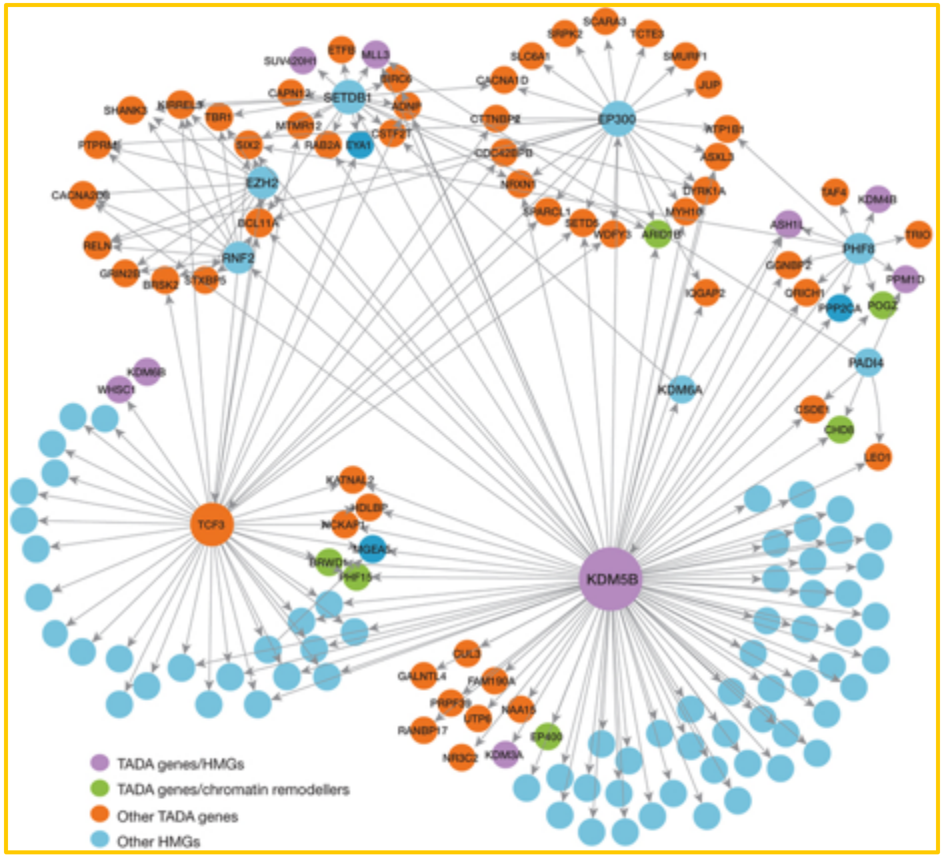
Introduction: biological networks

- Types of biological networks:
  - Intra-cellular networks
    1. **Transcriptional regulatory networks**
    2. **Metabolic networks**
    3. RNA networks
    4. Protein-protein interaction (PPI) networks
    5. Cell signaling networks

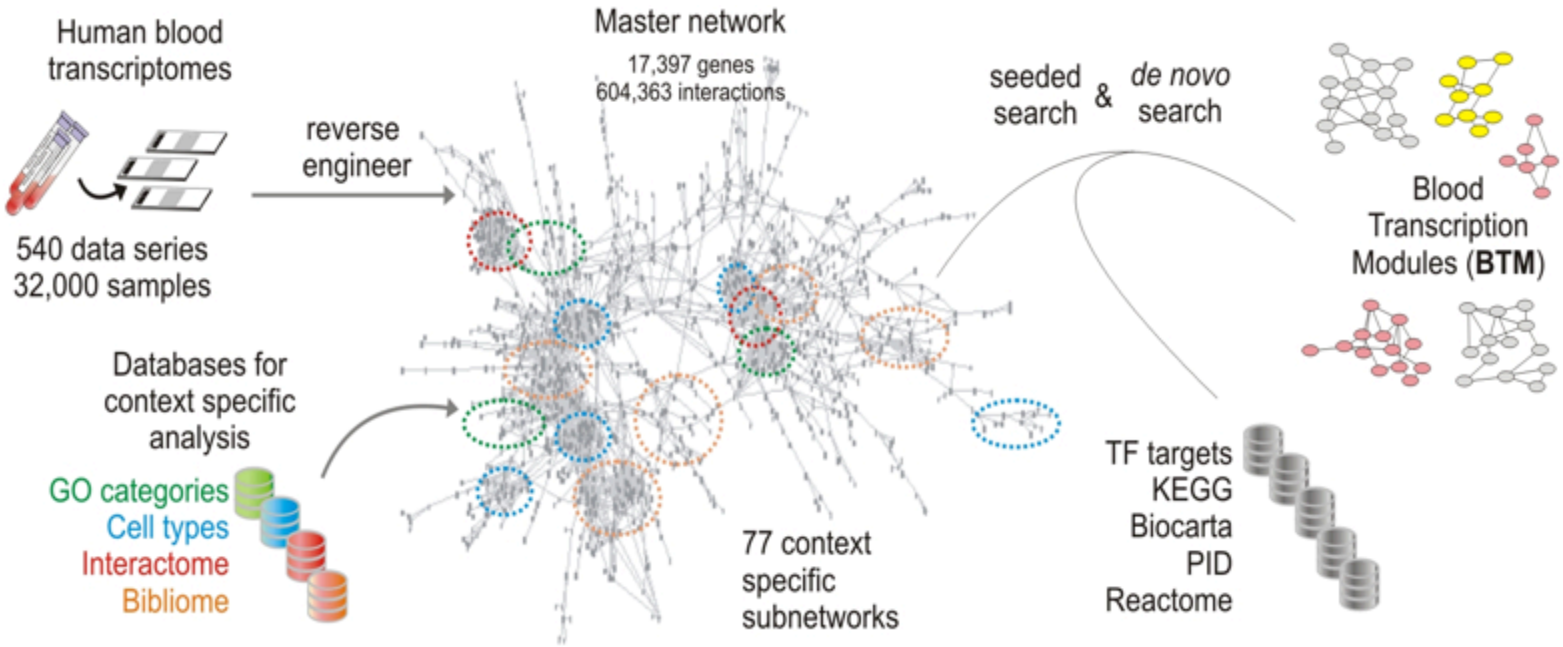
  - Other biological networks
    - Neuronal synaptic connection networks
    - Brain functional networks
    - Ecological food webs
    - Phylogenetic networks
    - Correlation networks (e.g., gene co-expression)
    - Disease – "disease gene" association networks
    - Drug – "drug target" networks

Transcriptional regulation networks and modules

- Model regulation of **gene expression**
  - Gene → mRNA → protein
- Nodes correspond to genes
- Directed edges correspond to interactions through which the products of one gene affect those of another



Blood Transcription Modules (BTMs)

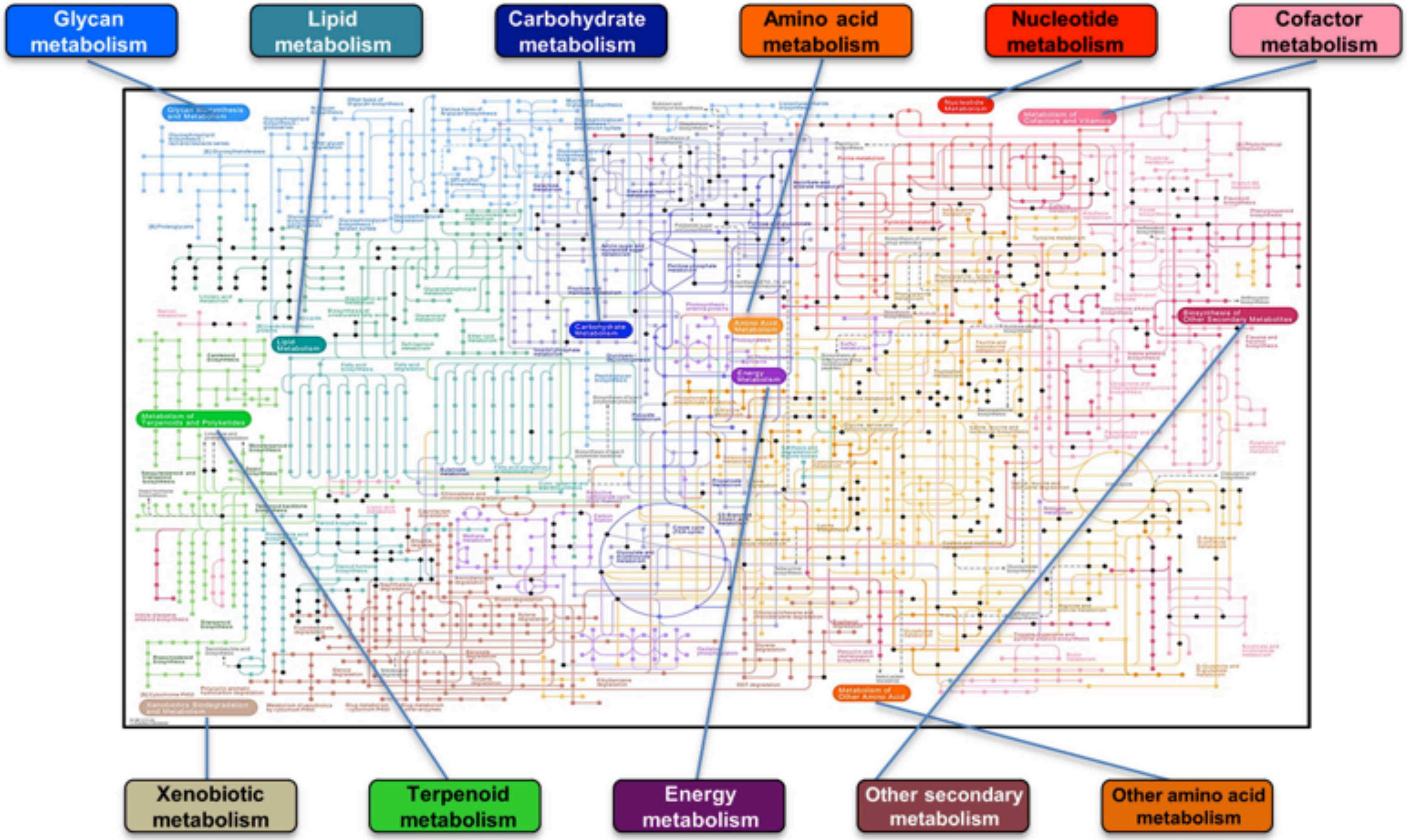


Blood Transcription Modules (BTM) as a powerful new tool. High quality gene network was first inferred from public transcriptomic data. Context specific subnetworks were derived by intersecting GO, cell types, interactome and bibliome. Gene modules were extracted from these subnetworks by search algorithms that take into account connection density and underlying conditions. KEGG, BioCarta, PID, Reactome and TF targets were integrated as search seeds. These BTM modules can be used as alternative to pathways, and often offer better sensitivities.

Li et al., 2014. Nature Immunology, 15:195

Metabolic networks

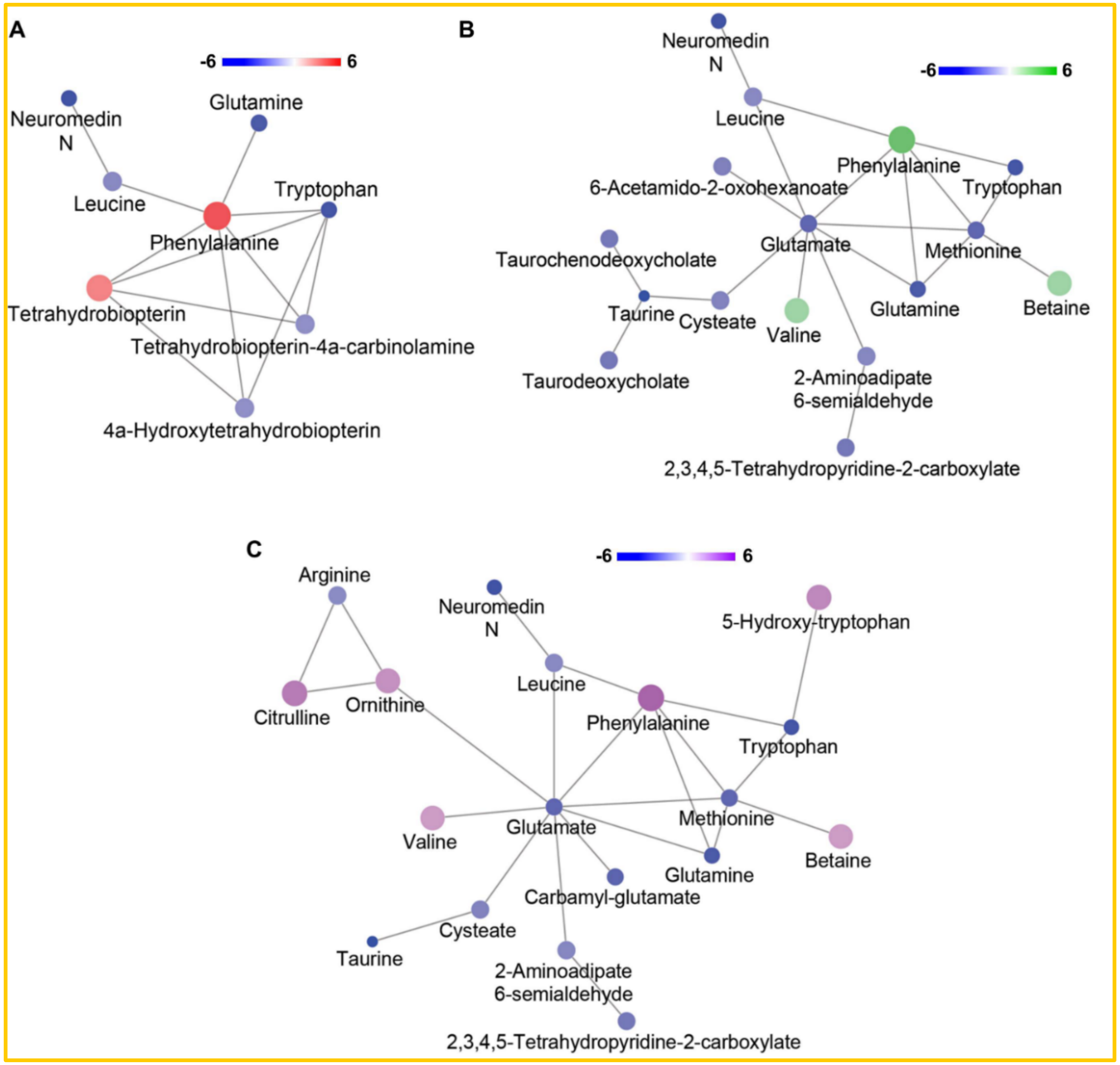
- Used for studying and modeling **metabolism**
  - Biochemical reactions in cells that allow an organism to carry out essential life functions



Banton et al 2016. *JAALAS* -KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway mapping of marmoset plasma metabolites associated with a change from the NE diet to the purified diet. The black dots represent metabolites in the pathways that were identified by using Mummichog

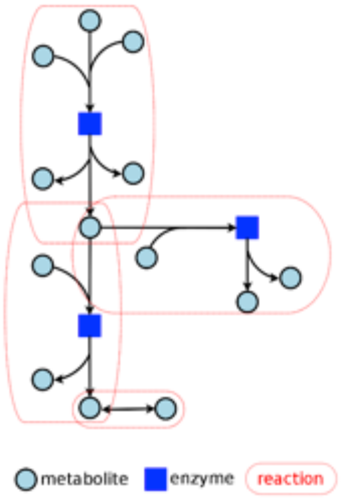
Metabolic networks

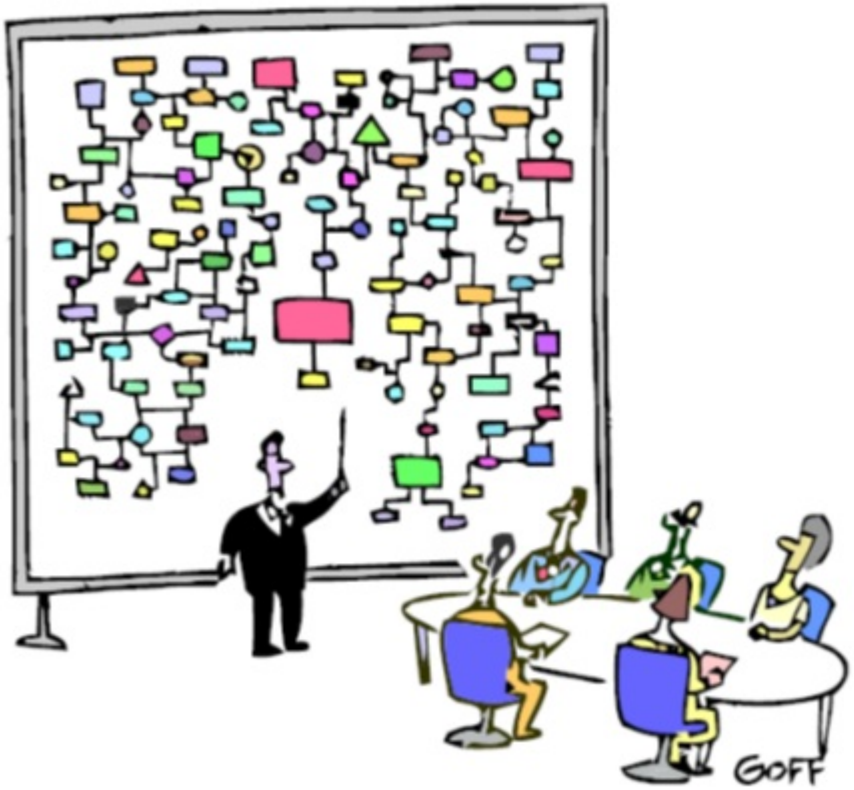
- **Metabolites**
  - Small molecules
  - Macromolecules
- **Metabolic pathways**
  - Series of successive biochemical reactions for a specific metabolic function, e.g., glycolysis, or penicillin synthesis, that convert one metabolite into another



Banton et al 2016. *JAALAS.* Metabolic network activity of plasma amino acid concentrations affected by changes between the baseline, standard, and synthetic diets fed to the common marmoset (*Callithrix jacchus*).

Metabolic model at Genome-scale: need for bioinformatics tools

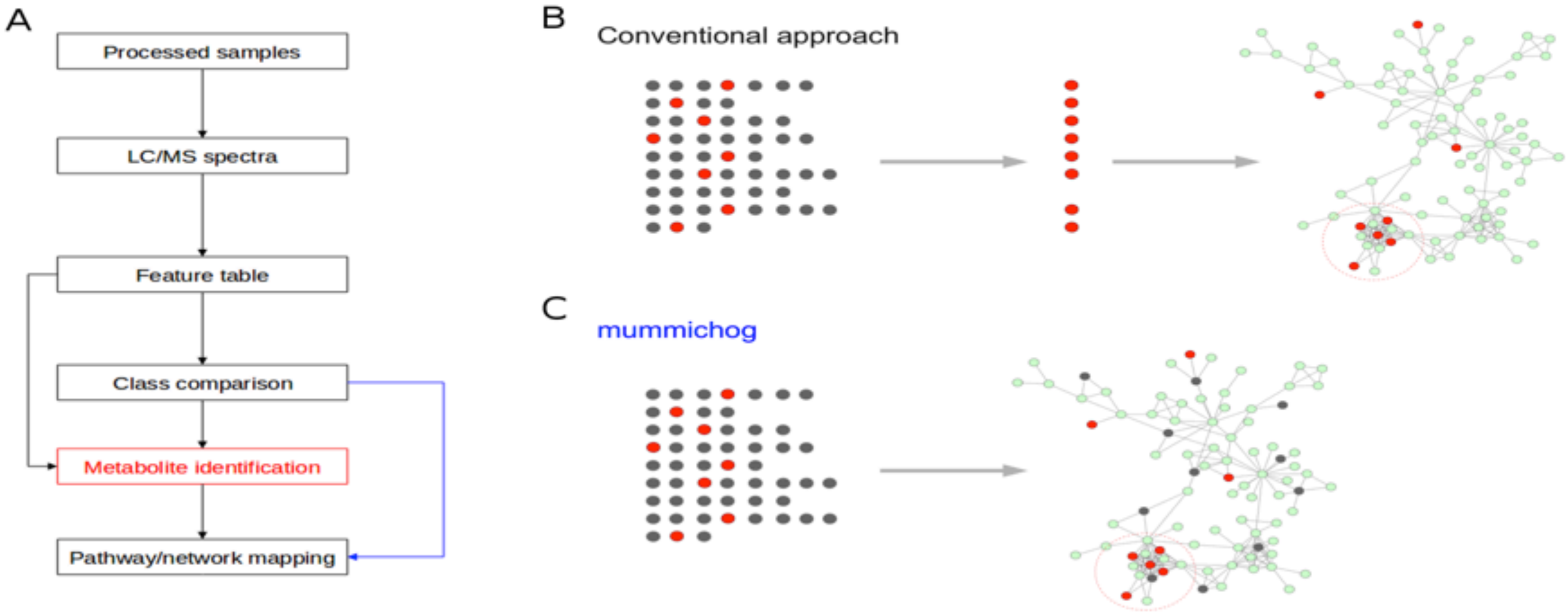




"And that's why we need a computer."

— Courtesy: Keck Graduate Institute

Mummichog: rewriting metabolomics workflow



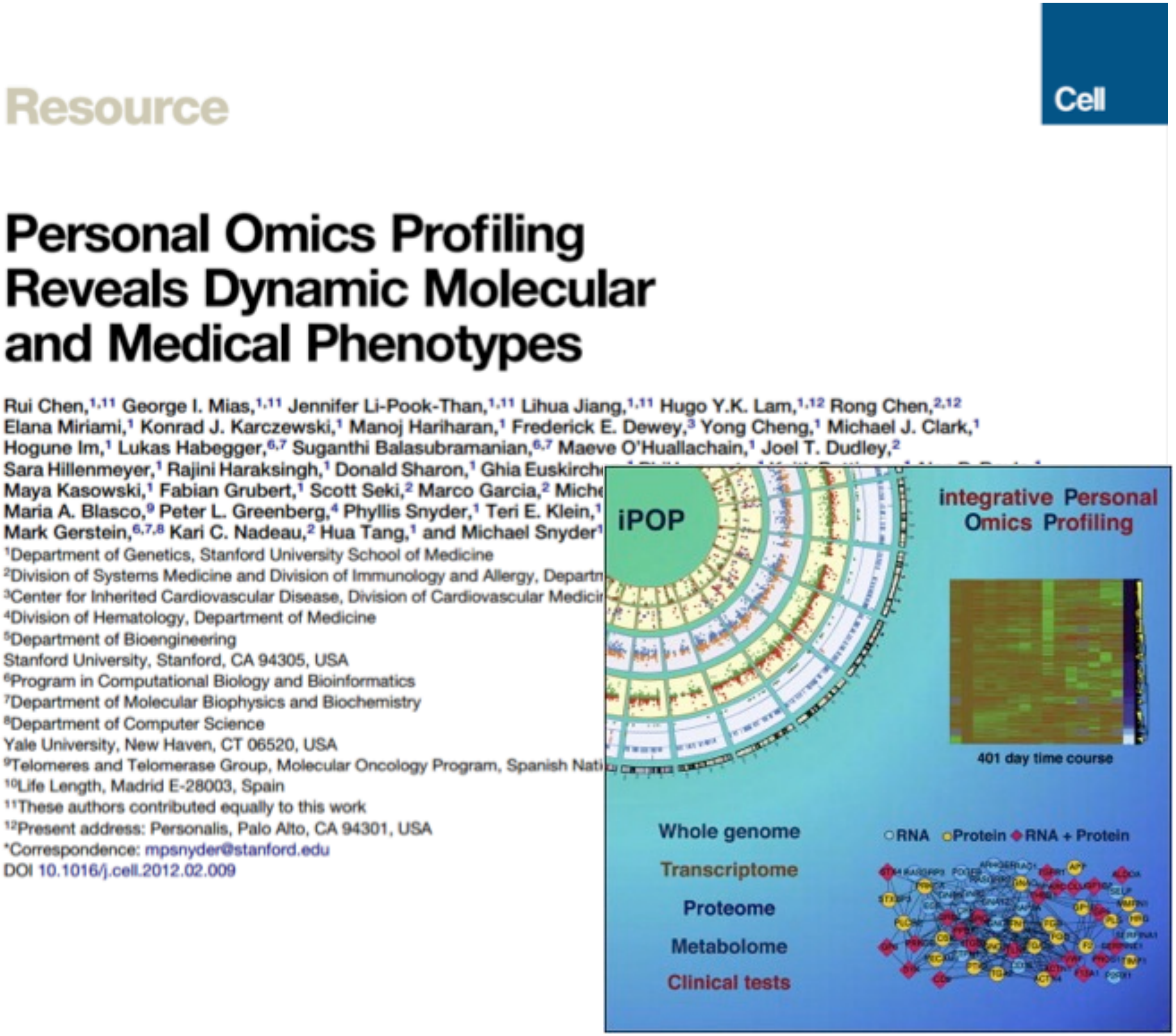
A) In the work flow of untargeted metabolomics, the conventional approach requires the metabolites to be identified before pathway/network analysis, while mummichog (blue arrow) predicts functional activity bypassing metabolite identification. **B)** Each row of dots represent possible matches of metabolites from one *m/z* feature, red the true metabolite, gray the false matches. The conventional approach first requires the identification of metabolites before mapping them to the metabolic network. **C)***mummichog* maps all possible metabolite matches to the network and looks for local enrichment, which reflects the true activity because the false matches will distribute randomly.

Li et al. 2013. PLoS Computational Biology. 9:e10031323

Case Studies

1. Reanalysis of Snyderome using new tools
2. Galactosemia: GALT transcriptomics and metabolomics integration
3. Formation of memory CD8+ T cells
4. Integration of Metabolomics and Transcriptomics to evaluate the effect pyrimethamine on plasma hemoglobin using Group LASSO

Case Study 1: **Reanalysis of Snyderome using new tools**



Resource

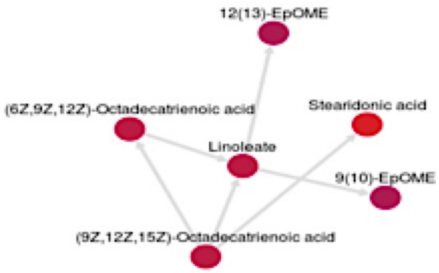
Cell

Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes

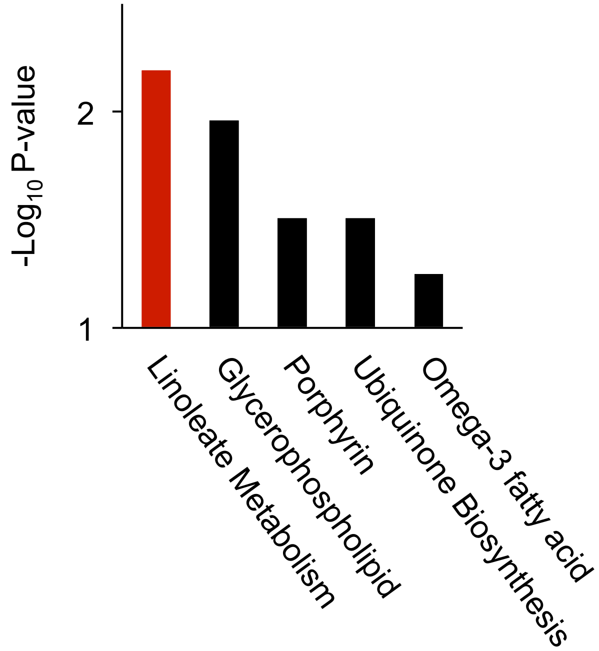
Rui Chen,[1,11] George I. Mias,[1,11] Jennifer Li-Pook-Than,[1,11] Lihua Jiang,[1,11] Hugo Y.K. Lam,[1,12] Rong Chen,[2,12]
Elana Miriami,[1] Konrad J. Karczewski,[1] Manoj Hariharan,[1] Frederick E. Dewey,[3] Yong Cheng,[1] Michael J. Clark,[1]
Hogune Im,[1] Lukas Habegger,[6,7] Suganthi Balasubramanian,[6,7] Maeve O'Huallachain,[1] Joel T. Dudley,[2]
Sara Hillenmeyer,[1] Rajini Haraksingh,[1] Donald Sharon,[1] Ghia Euskirch
Maya Kasowski,[1] Fabian Grubert,[1] Scott Seki,[2] Marco Garcia,[2] Miche
Maria A. Blasco,[9] Peter L. Greenberg,[4] Phyllis Snyder,[1] Teri E. Klein,[1]
Mark Gerstein,[6,7,8] Kari C. Nadeau,[2] Hua Tang,[1] and Michael Snyder[1]

[1]Department of Genetics, Stanford University School of Medicine
[2]Division of Systems Medicine and Division of Immunology and Allergy, Departn
[3]Center for Inherited Cardiovascular Disease, Division of Cardiovascular Medicir
[4]Division of Hematology, Department of Medicine
[5]Department of Bioengineering
Stanford University, Stanford, CA 94305, USA
[6]Program in Computational Biology and Bioinformatics
[7]Department of Molecular Biophysics and Biochemistry
[8]Department of Computer Science
Yale University, New Haven, CT 06520, USA
[9]Telomeres and Telomerase Group, Molecular Oncology Program, Spanish Nati
[10]Life Length, Madrid E-28003, Spain
[11]These authors contributed equally to this work
[12]Present address: Personalis, Palo Alto, CA 94301, USA
*Correspondence: mpsnyder@stanford.edu
DOI 10.1016/j.cell.2012.02.009

Case Study 1: *Mummichog* **interpretation of Snyder metabolome**

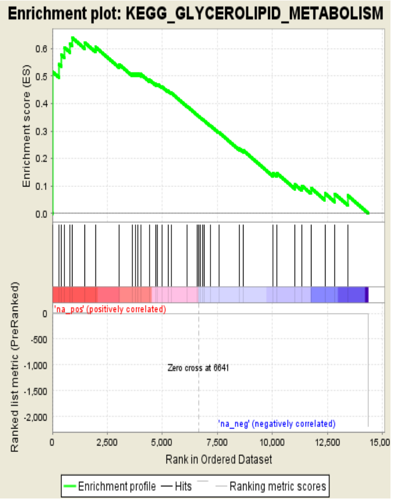

12(13)-EpOME
(6Z,9Z,12Z)-Octadecatrienoic acid
Stearidonic acid
Linoleate
9(10)-EpOME
(9Z,12Z,15Z)-Octadecatrienoic acid

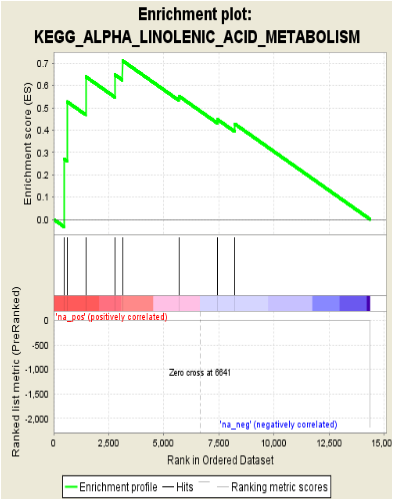
Metabolic Network module corresponding to Linoleate pathway.

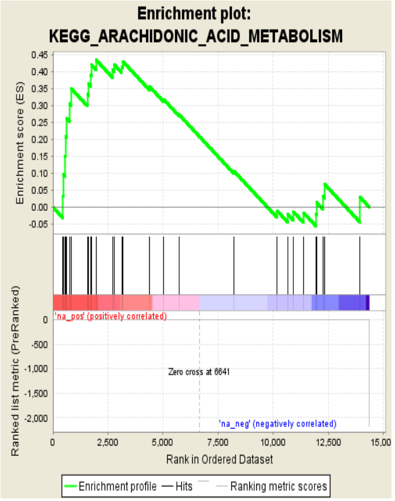


Significant Metabolic Pathways during RSV infection (P < 0.05). Porphyrin pathway pinpoints the clinical phenotype of anemia.

Gene Pathways from Snyder transcriptome are Consistent with Metabolite Pathways

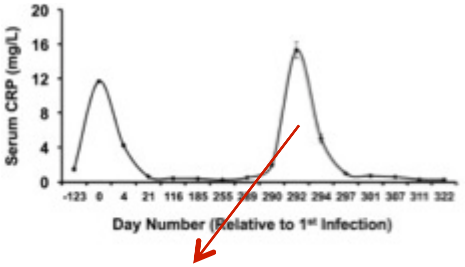






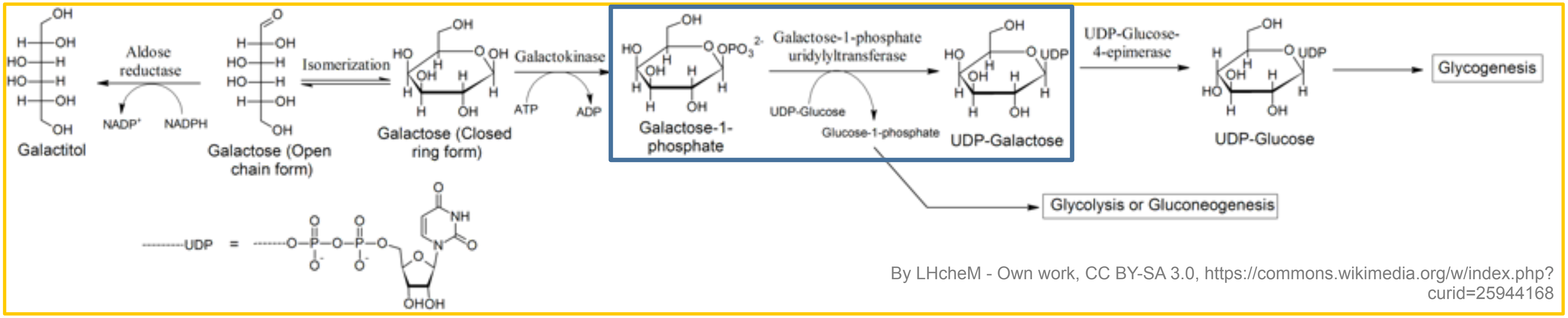
Case Study 1: **Reanalysis of Snyderome using new tools**

The transcriptomics signature during RSV infection, using BTMs. Enrichment test was performed within GSEA software, using BTMs as custom gene sets.



Name	NES	NOM p-val
Enriched in monocytes (II) (M11.0)	2.386807	0
Enriched in monocytes (I) (M4.15)	2.08188	0.005102
RIG-1 like receptor signaling (M68)	1.867297	0.00978
Complement Activation (I) (M112.0)	1.948812	0.011086
Enriched in monocytes (III) (M73)	1.932553	0.011905
Cell Activation (IL15, IL23, TNF) (M24)	1.89332	0.012136
Cell Cycle and Growth Arrest (M31)	1.892822	0.016908
Formyl peptide receptor mediated neutrophil response (M11.2)	1.920023	0.021028
RA, WNT, CSF receptors network (Monocyte) (M23)	1.842038	0.022321
Extracellular Matrix, Collagen (M210)	1.76067	0.025316
Signaling in T Cells (I) (M35.0)	1.743513	0.027972
Myeloid cell enriched receptors and transporters (M4.3)	1.942556	0.02849
Inflammatory response (M33)	1.636564	0.029056
Enriched in activated dendritic cells (II) (M165)	1.918629	0.02973
Viral sensing & immunity; irf2 targets network (I) (M111.0)	1.815972	0.03038
Blood Coagulation (M11.1)	1.855351	0.030691
TLR and Inflammatory Signaling (M16)	2.084829	0.031884
Lysosome (M209)	1.678483	0.032407
Innate Antiviral Response (M150)	1.679814	0.033413

Case Study 2: **Background**



By LHcheM - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=25944168

- **Galactosemia** is an autosomal recessive condition that affects an individual's ability to metabolize galactose

- In *Drosophila melanogaster* **dGALT** is the presumed ortholog of the human GALT gene which converts Galactose-1-phosphate to UDP-Galactose

- Genotypes:
  - **Ap2** is the imprecise excision of a p element in the dGALT gene and results in loss of dGaLT activity (Knockout).
  - **C2** is the precise excision of the same p element and results in normal GALT activity (Wild-type).

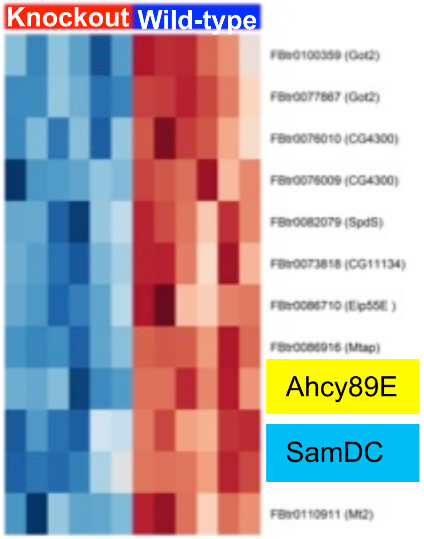
Transcriptome x Metabolome

- Transcriptomics
- 15-20 ug of Larva were used for each RNA extraction and subsequent cDNA synthesis.
- Dye hybridization and microarray were performed using Nimblegen technology.

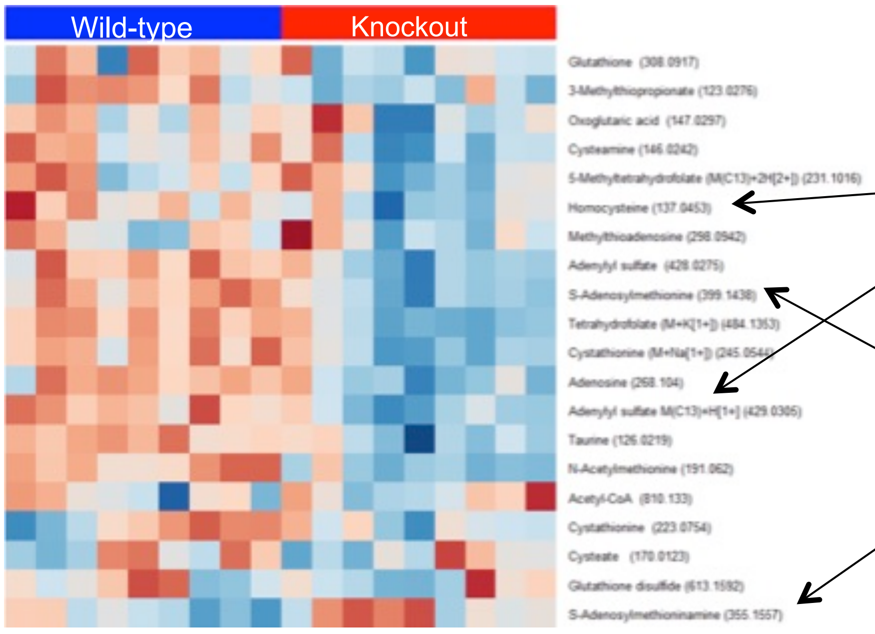
- Metabolomics
- Human Data
  - 3 Cases and 6 controls
  - Using 19,505 metabolite features
- Fly Data
  - 9 Knockout and 9 Wild-Type
  - Using 9,767 metabolite features

Case Study 2: Transcriptome x Metabolome

Fly Genes



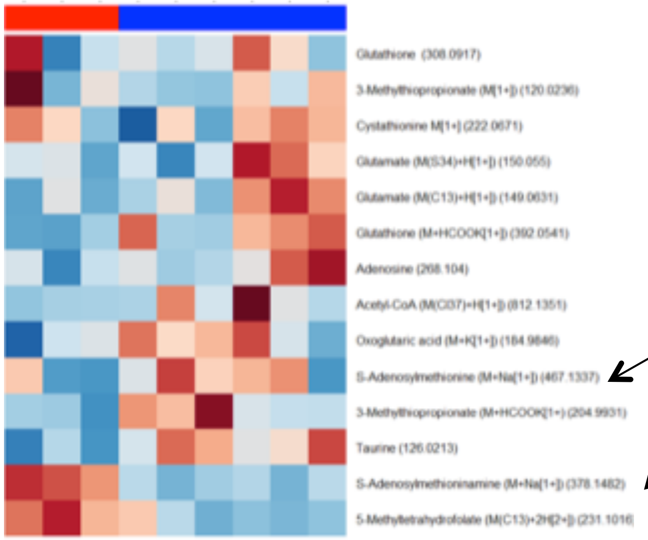
Fly Metabolites



Products of Ahcy89E

Substrate and Product of SamDC

Human Metabolites



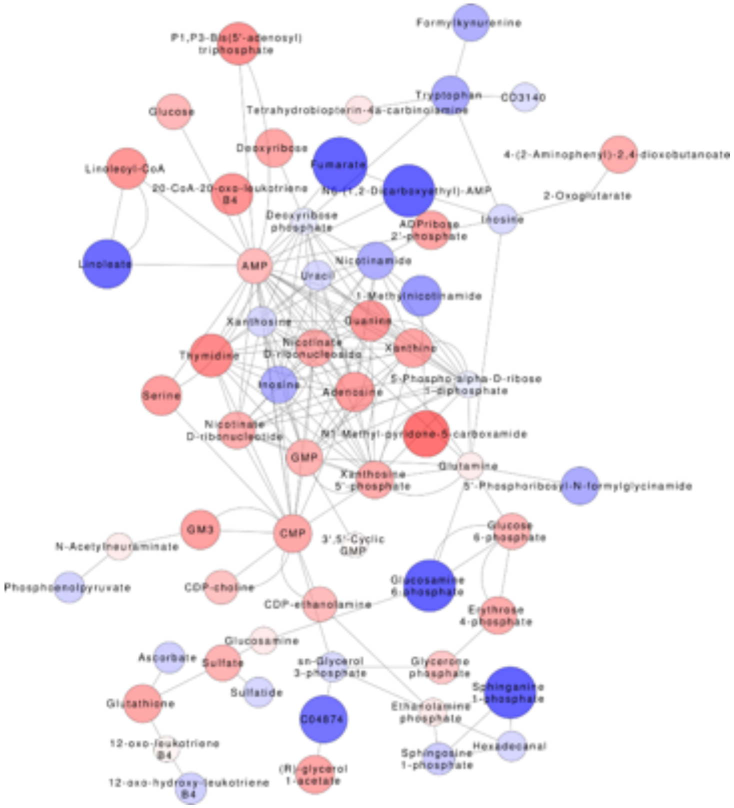
Substrate and Product of SamDC

Case Study 3: Transcriptome x Metabolome

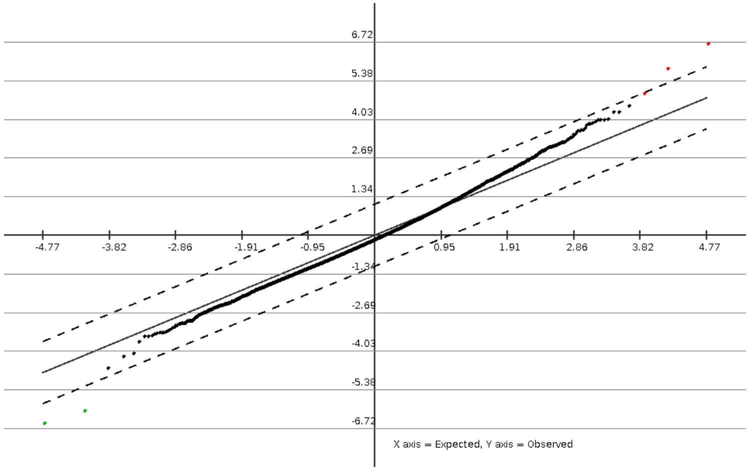
nature
immunology

Autophagy is essential for effector CD8+ T cell survival and memory formation

Xiaojin Xu[1,5], Koichi Araki[1,5], Shuzhao Li[2], Jin-Hwan Han[1], Lilin Ye[1], Wendy G Tan[1], Bogumila T Konieczny[1], Monique W Bruinsma[3], Jennifer Martinez[4], Erika L Pearce[3], Douglas R Green[4], Dean P Jones[2], Herbert W Virgin[3] & Rafi Ahmed[1]

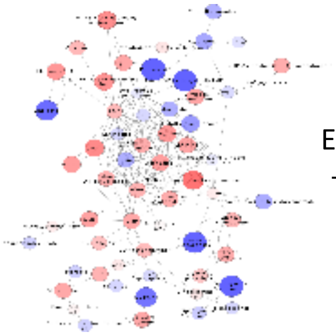


metabolomics



transcriptomics

Enzymes associated with significant metabolites



Metabolites

Enzymes →

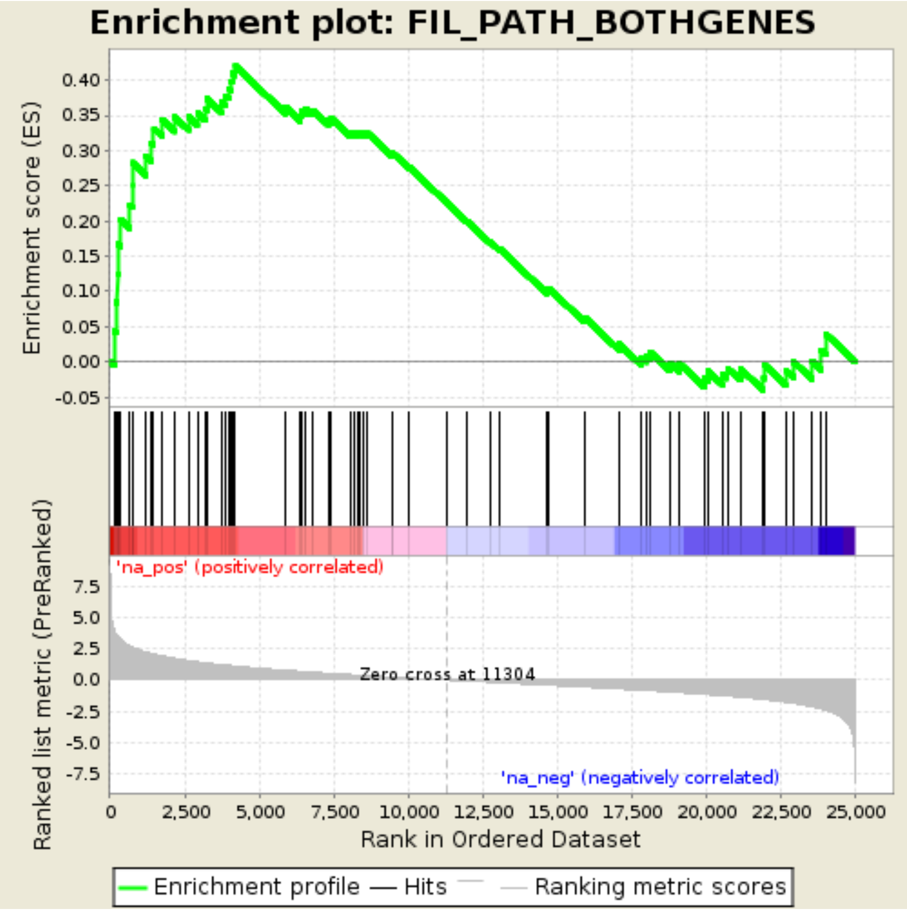
2.5.1.56, 2.7.1.91, 2.4.2.9, 2.4.2.8, 1.14.16.4, 1.14.16.5, 3.6.1.22, 2.4.2.1, 2.4.1.80, 3.1.4.35, 2.4.2.4, 2.4.2.7, 2.4.2.14, 2.4.2.11, 2.4.2.12, 3.5.4.17, 2.4.2.19, 1.1.1.94, 3.1.6.8, 3.1.6.1, 4.3.2.2, 1.14.14.1, 3.1.3.4, 3.1.3.5, 3.1.4.46, 2.4.1.141, 1.3.99.13, 3.6.1.5, 3.6.1.6, 3.6.1.9, 3.6.1.8, 2.1.1.1, 3.5.1.9, 2.7.1.1, 2.7.1.8, 3.1.1.4, 2.7.8.-, 3.2.1.18, 2.7.8.2, 2.7.8.5, 2.7.8.8, 1.1.99.4, 1.1.99.5, 2.7.1.74, 2.7.7.14, 3.6.1.29, 3.6.1.19, 3.6.1.17, 2.7.1.138, 2.4.1.47, 6.3.5.1, 6.3.5.3, 6.3.5.2, 6.2.1.3, 1.1.1.102, 4.1.3.3, 1.14.13.30, 3.2.2.1, 2.5.1.18, 3.5.1.23, 1.13.11.11, 2.6.1.7, 2.7.1.59, 4.1.2.13, 2.4.99.8, 2.4.99.9, 1.3.3.6, 3.1.3.10, 3.2.1.46, 3.2.1.45, 6.3.4.4, 2.2.1.1, 2.2.1.2, 6.3.4.1, 2.7.8.1, 2.7.1.20, 1.7.1.7, 2.4.2.22, 2.3.1.24, 2.7.8.11, 2.7.8.15, 3.5.4.3, 3.1.4.2, 3.5.4.6, 2.7.6.1, 2.6.1.16, 3.1.4.12, 3.1.4.17, 2.4.1.117, 1.2.3.1, 3.5.4.4, 1.4.3.2, 4.1.2.27, 3.1.4.3, 6.1.1.2, 4.2.1.17, 3.2.2.2, 3.1.2.2, 3.2.2.6, 3.2.2.5, 3.5.99.6, 3.2.2.8, 1.1.1.8, 3.7.1.3, 1.13.11.34

Genes →

Gpd1l, Kdsr, Ado, Acox1, Gmpr2, Tkt, Alg5, Alg13, Hprt, Nampt, Gsta4, Gstk1, Gstm1, Gstm4, Gsto1, Gstp1, Gstp2, Gstt2, Hpgds, Gfpt1, Adk, Nagk, Dck, Sphk1, Sphk2, Prps1, Prps2, Cept1, Ept1, Cept1, Cdipt, Plb1, Acot2, Lpin1, Lpin2, Pde1b, Pde2a, Pde3b, Pde4a, Pde4d, Pde7a, Pde8a, Pde5a, Arsa, Gba2, Galc, Bst1, Cd38, Asah1, Asah2, Ada, Ampd1, Ampd2, Ampd3, Cant1, Enpp1, Itpa, Enpp4, Aldoa, Aldoc, Sgpl1, Npl, Acsl1, Acsl3, Acsl4, Acsl5

Gpda, Acox3, Oxla, Gmpr1, T23o, Lox5, Cp4f3, Cp4fe, Cp19a, Cp1a1, Cp1a2, Cp1b1, Cp237, Cp238, Cp239, Cp240, Cp254, Cp255, Cp270, Cp2a4, Cp2a5, Cp2ac, Cp2b9, Cp2ba, Cp2bj, Cp2ct, Cp2d9, Cp2da, Cp2db, Cp2dq, Cp2j5, Cp2j6, Cp2s1, Cp2u1, Cp341, Cp3ab, Cp3ad, Cp3ag, Cp3ap, Cp4b1, Cp4ca, Cp4x1, Cy250, Tph1, Tph2, Alkmo, Nnmt, Tktl1, Tktl2, Taldo, Cegt, Pnph, Typh, Apt, Nadc, Sia8a, Siat9, Gsta1, Gsta2, Gsta3, Gstm2, Gstm5, Gstm6, Gstm7, Gsto2, Gstt1, Gstt4, Maai, Mgst1, Mgst3, Aadat, Aatm, Kat1, Kat3, Gfpt2, Hkdc1, Hxk1, Hxk2, Hxk3, Cerk1, Pcy2, Chpt1, Pgps1, Gpt, Hrsl3, Pa21b, Pa24a, Pa24b, Pa24d, Pa24e, Pa24f, Pa2g5, Pa2ga, Pa2gc, Pa2gd, Pa2ge, Pa2gf, Pa2gx, Pg12a, Plpl9, Aco15, Acot1, Acot3, Acot4, Acot5, Baat, Bach, Them4, Lpin3, Lpp1, Lpp2, Lpp3, Lppr2, Lppr3, Lppr4, Ppc1a, Ppc1b, 5nt1a, 5nt1b, 5nt3a, 5nt3b, 5ntc, 5ntd, Ppap, Gpcp1, Asm, Nsma2, Nsma3, Nsma, Pde10, Pde11, Pde1a, Pde1c, Pde3a, Pde4c, Pde7b, Pde8b, Cncg, Cnrg, Pde10, Pde11, Pde6a, Pde6b, Pde6c, Pde9a, Neur1, Neur2, Neur3, Neur4, Glcm, Kfa, Acer1, Acer2, Guad, Gnpi1, Gnpi2, Entp1, Entp8, Entp4, Entp5, Enpp3, Ap4a, Nud12, Fhit, Kynu, Aldob, Echa, Echm, Echp, Pur8, Sywc, Sywm, Acbg1, Acbg2, Acsl6, S27a2, Pura1, Pura2, Nade, Guaa, Pur4

Enzyme genes significantly enriched towards KO



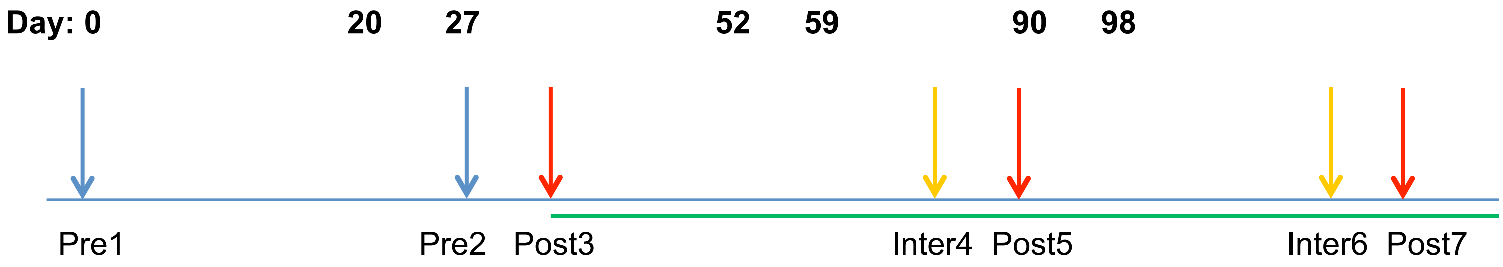
Expression of genes corresponding to related enzymes are enriched for KO cells, DNA microarray data, GSEA (Gene Set Enrichment Analysis).
Nominal p = 0,
FWER p = 0.024.

Case Study 4: Integration of Metabolomics and Transcriptomics to evaluate the effect of subcurative doses of pyrimethamine on plasma hemoglobin in Rhesus macaques using Group LASSO

Experimental Design

- Five macaques were each delivered a sub-curative dose of pyrimethamine at Day 21, and 3-day curative doses commencing at Days 52 and 90, in each case immediately following peripheral blood sampling. This results in two **pre-drug**, three **post-drug**, and two **inter-drug** treatments as indicated.

- Plasma samples were collected over the course of 100 days.



- These correspond to:
  - Time Point 1 = Day 0 (Baseline Sampling Point)
  - Time Point 2 = Day 21
  - Time Point 3 = Day 27
  - Time Point 4 = Day 52
  - Time Point 5 = Day 59
  - Time Point 6 = Day 90
  - Time Point 7 = Day 98

Analysis Summary

- **Question 1**: Are there genes and metabolites that are associated with hemoglobin levels?
- Correlation Analysis
  - 305 metabolites are significantly correlated with Hgb levels
  - 1074 genes are significantly correlated with Hgb levels

- **Question 2**: Are there features that separate subjects based on drug exposure (pre *vs*. inter *vs*. post)
- Differential Expression Analysis
  - 1660 metabolites can separate the subjects into *inter* and *post* drug exposure groups
  - 925 genes can separate the subjects into *pre*, *inter* and *post* drug exposure groups

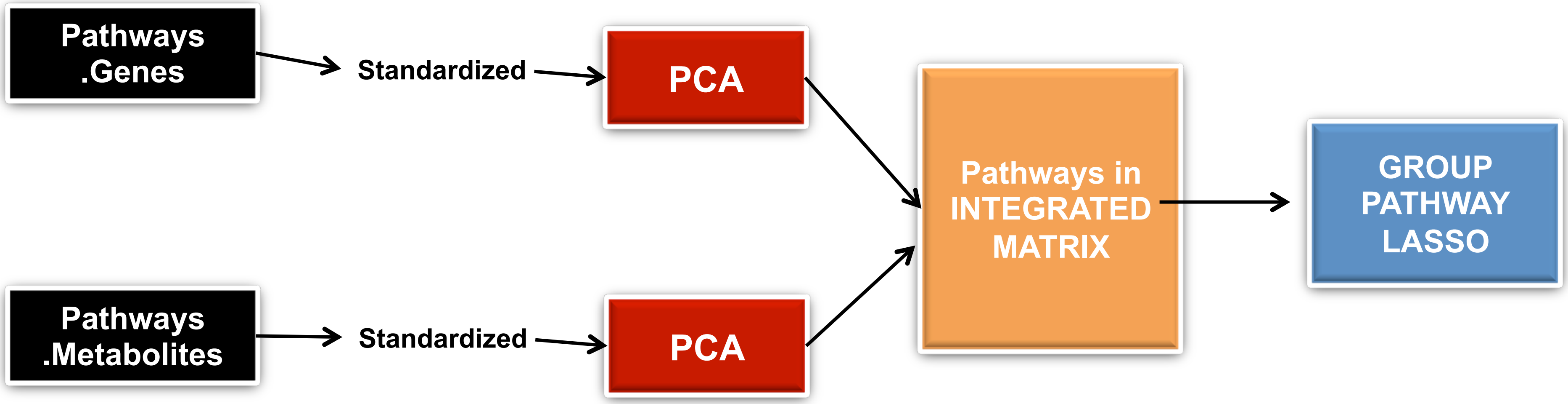
- **Conclusion**
  - The list of potential targets is cumbersome (information overload)

Can a single test answer two questions?

- Which features from **both platforms** are associated with drug exposure?

- Among those features, which of them are *specifically* associated with hemoglobin levels?

Hybrid concatenation and transformation based integration using Group LASSO (Banton et al.)



Contributions of the method

- Allows integration of multiple omics data types
- The method is not platform specific or dependent
- The method retains functional information provided by pathways
- The method allows prediction of outcomes and thus can be used in the development of clinical biomarkers

Least Absolute Shrinkage and Selection Operator (LASSO)

A popular model selection and shrinkage estimation method (Tibshirani 1995).

The lasso estimator is defined as:

$$\hat{\beta}_{\lambda} = \underset{\beta}{\operatorname{argmin}}(\|\mathbf{Y} - X\beta\|_2^2 + \boxed{\lambda}\sum_{j=1}^{p} |\beta_j|)$$

Where $\boldsymbol{\lambda}$ is the tuning parameter

Extended from the lasso penalty, the group lasso estimator is:

$$\hat{\beta}_{\lambda} = \arg\min_{\beta}(\|Y - X\beta\|_2^2 + \lambda\sum_{g=1}^{G} \boxed{\|\beta_{I_g}\|}_2)$$

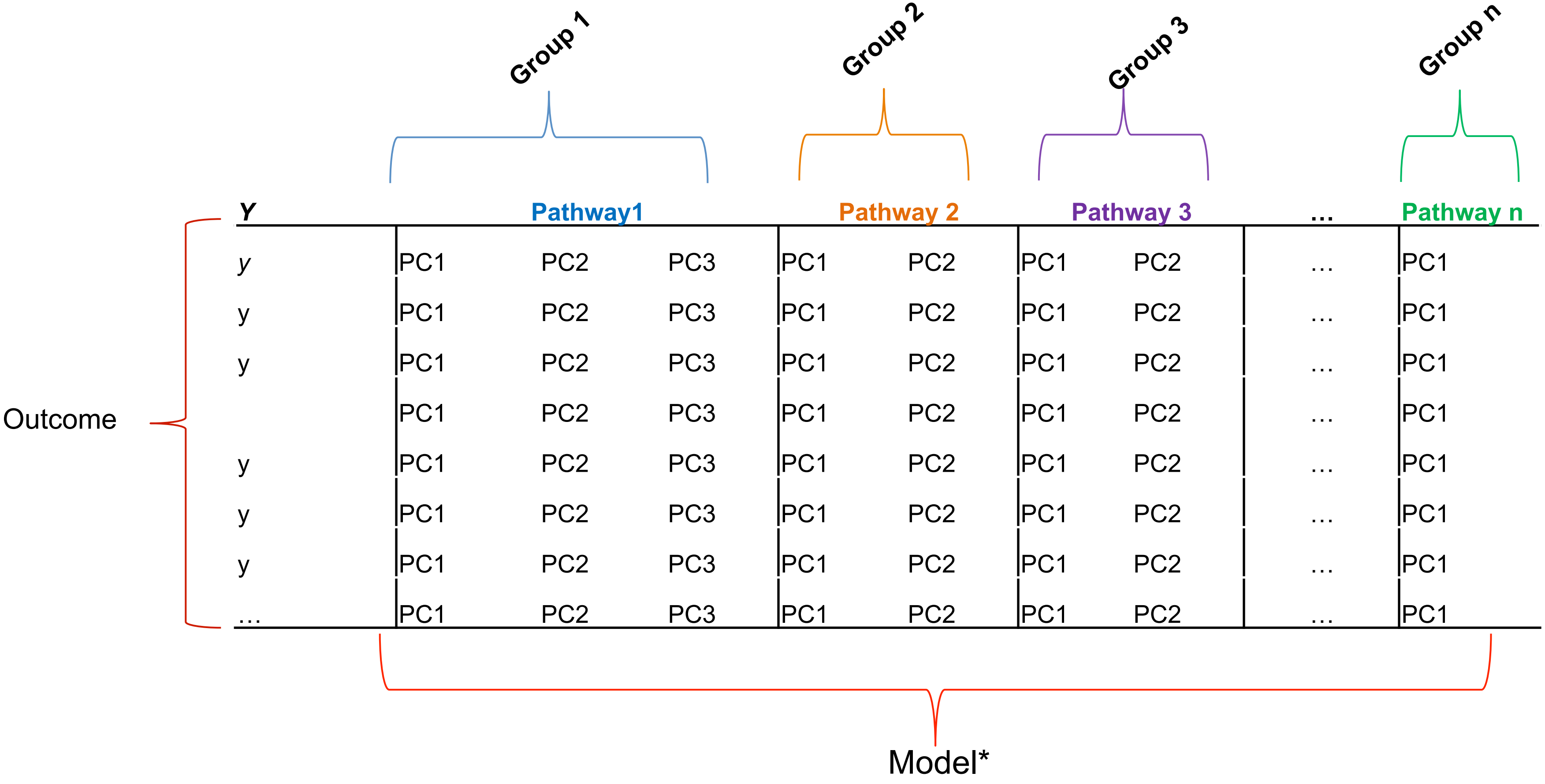
I_g : the index set belonging to the g th group of variables.

The penalty does the variable selection at the group level, belonging to the intermediate between $l1-$ and $l2-$ type penalty.

It encourages that either $\hat{\boldsymbol{\beta}}_g = 0$ or $\hat{\beta}_{g,j} \neq 0$ for all $j \in \{1, \ldots, df_g\}$

Meier L, Geer Svd, Bühlmann P. 2008. The Group Lasso for Logistic Regression. Journal of the Royal Statistical Society Series B (Statistical Methodology) *70:53-71.*

Integration with Group LASSO



Y	**Pathway1**			**Pathway 2**		**Pathway 3**		**…**	**Pathway n**
y	PC1	PC2	PC3	PC1	PC2	PC1	PC2	…	PC1
y	PC1	PC2	PC3	PC1	PC2	PC1	PC2	…	PC1
y	PC1	PC2	PC3	PC1	PC2	PC1	PC2	…	PC1
	PC1	PC2	PC3	PC1	PC2	PC1	PC2	…	PC1
y	PC1	PC2	PC3	PC1	PC2	PC1	PC2	…	PC1
y	PC1	PC2	PC3	PC1	PC2	PC1	PC2	…	PC1
y	PC1	PC2	PC3	PC1	PC2	PC1	PC2	…	PC1
…	PC1	PC2	PC3	PC1	PC2	PC1	PC2	…	PC1

Where ***Y*** is the **hemoglobin** level in each subject

*Model built using cross-validation

Results*: Number of Targets is drastically reduced

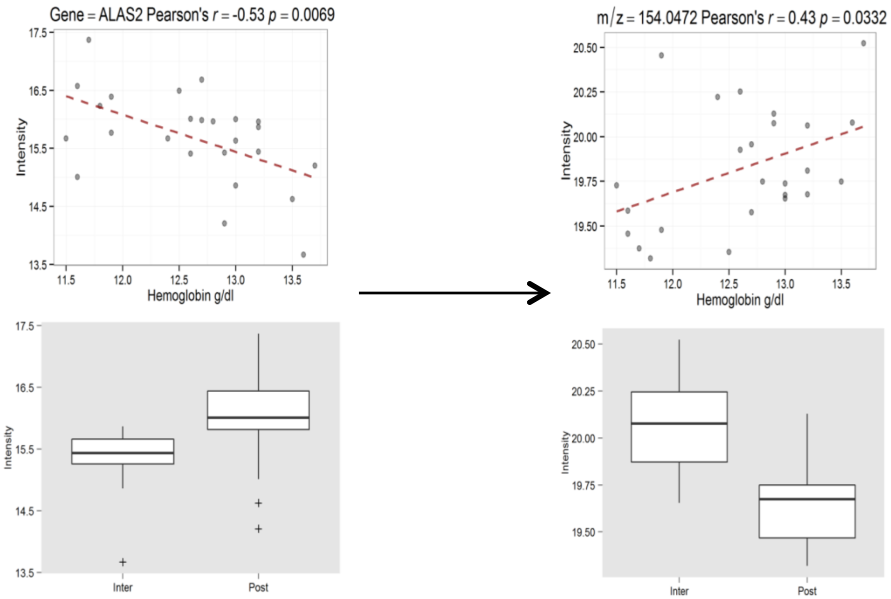
Pathway	Number of Genes	Number of Metabolites
Ascorbate and aldarate metabolism (Vitamin C)	5	4
Glycerophospholipid metabolism	53	7
Linoleic acid metabolism	6	6
Cysteine and methionine metabolism	26	11
Porphyrin and chlorophyll metabolism	**20**	**9**
Retinol metabolism (Vitamin A)	16	6
Valine, leucine and isoleucine degradation	38	6
Nicotinate and nicotinamide metabolism	18	3
Total features	182	52

*Lambda min = 0.009
Lambda se = 0.1872

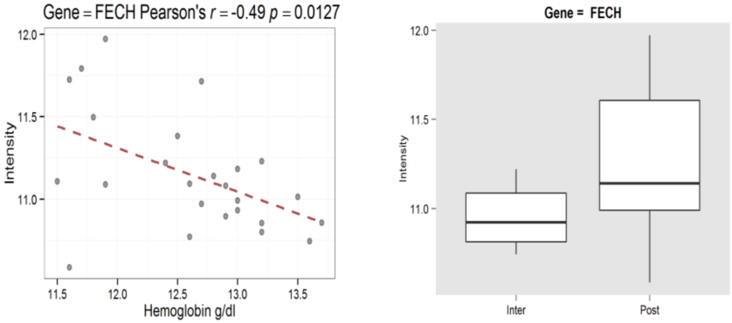
Proof of Concept: Correlations between Hgb and significant genes/metabolites **selected by** Group Pathway Lasso

Pathway: Porphyrin and chlorophyll metabolism (Heme dysregulation in first, second, and eighth steps of biosynthetic pathways)

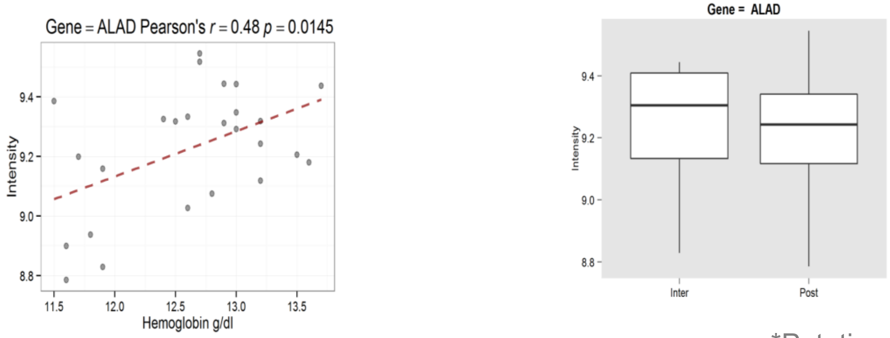
Step 1: Heme biosynthesis (**ALAS2** gene → **aminolevulinic acid** synthase)



Step 8: Terminal step in Heme biosynthesis (**FECH** gene)

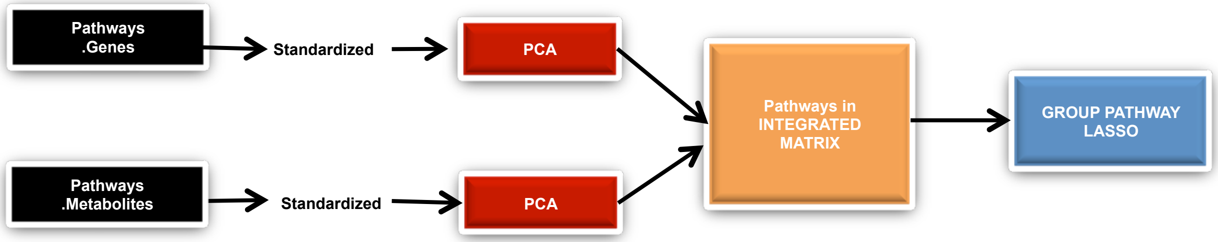


Step 2: Heme biosvnthesis (**ALAD** gene)



*Putative database match

Summary: methods of omics integration



Pathways .Genes

Standardized

PCA

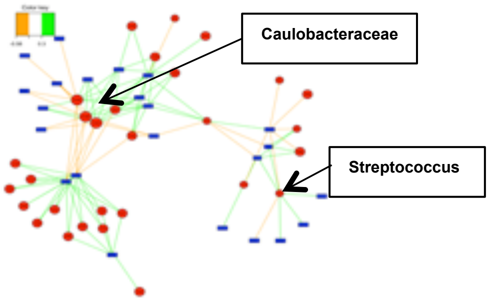
Pathways .Metabolites

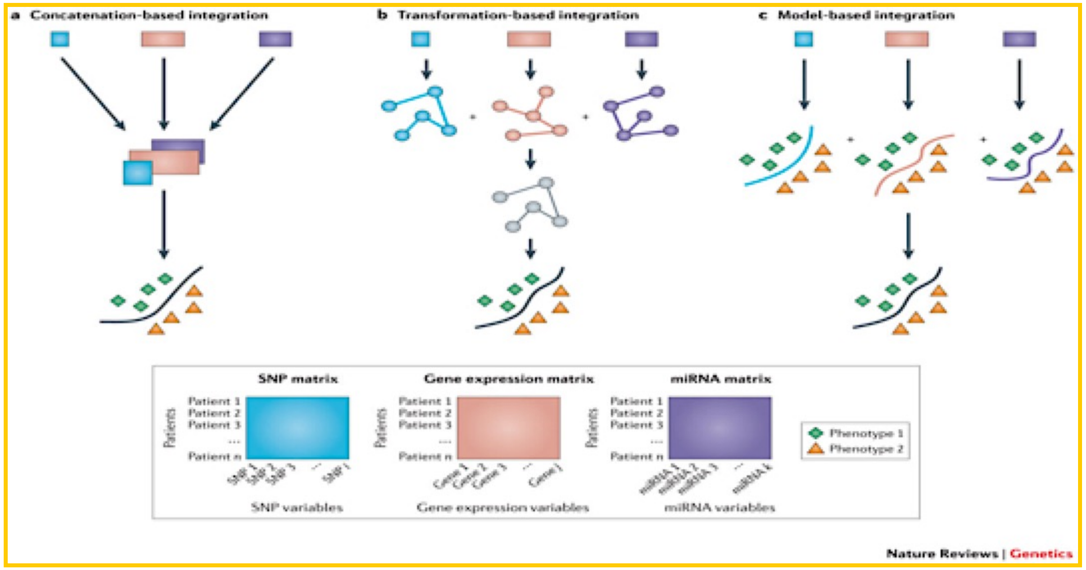
Standardized

PCA

Pathways in INTEGRATED MATRIX

GROUP PATHWAY LASSO





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.*

Acknowledgements

Emory University SOM

Dean P. Jones (Professor, Director of Clinical Biomarkers Laboratory)
Shuzhao Li (Asst. Professor, Director of Immunometabolomics)

Douglas I. Walker (TCE study)
Paul A. Dawson (Bile acids study)
Sushma K. Cribbs (HIV study)
Judy Fridovich-keil (GALT study)
MahPIC (Pyrimethamine study)

Members of Clinical Biomarkers Laboratory

Georgia State University, PHI

Ruiyan Luo

Clinical Biomarkers Laboratory

Questions?