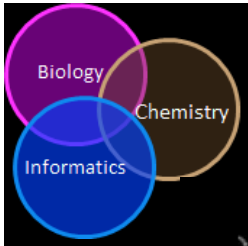
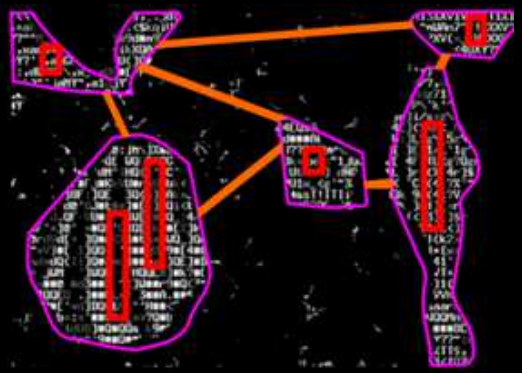
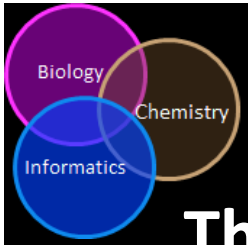




Introduction

2015 Workshop on Statistics and Data Analysis for Metabolomics

Dmitry Grapov, PhD



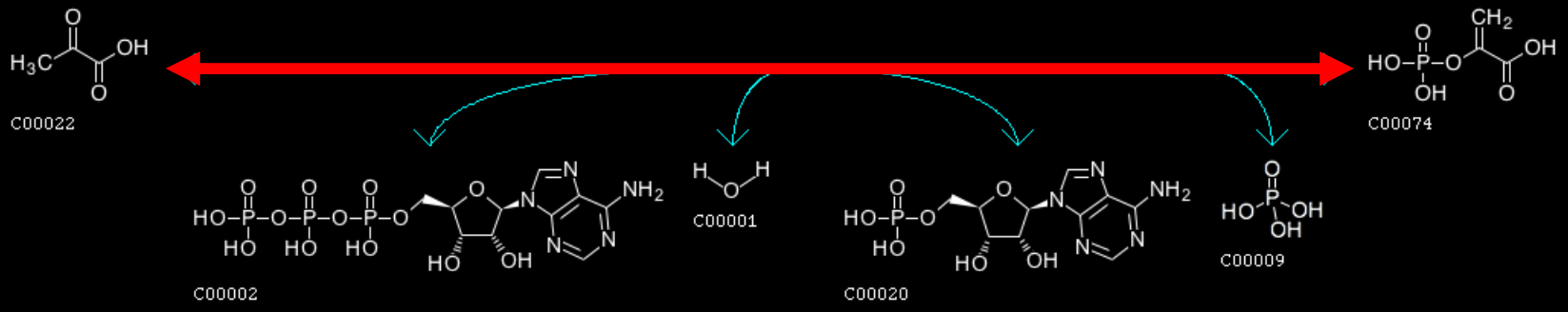
Important

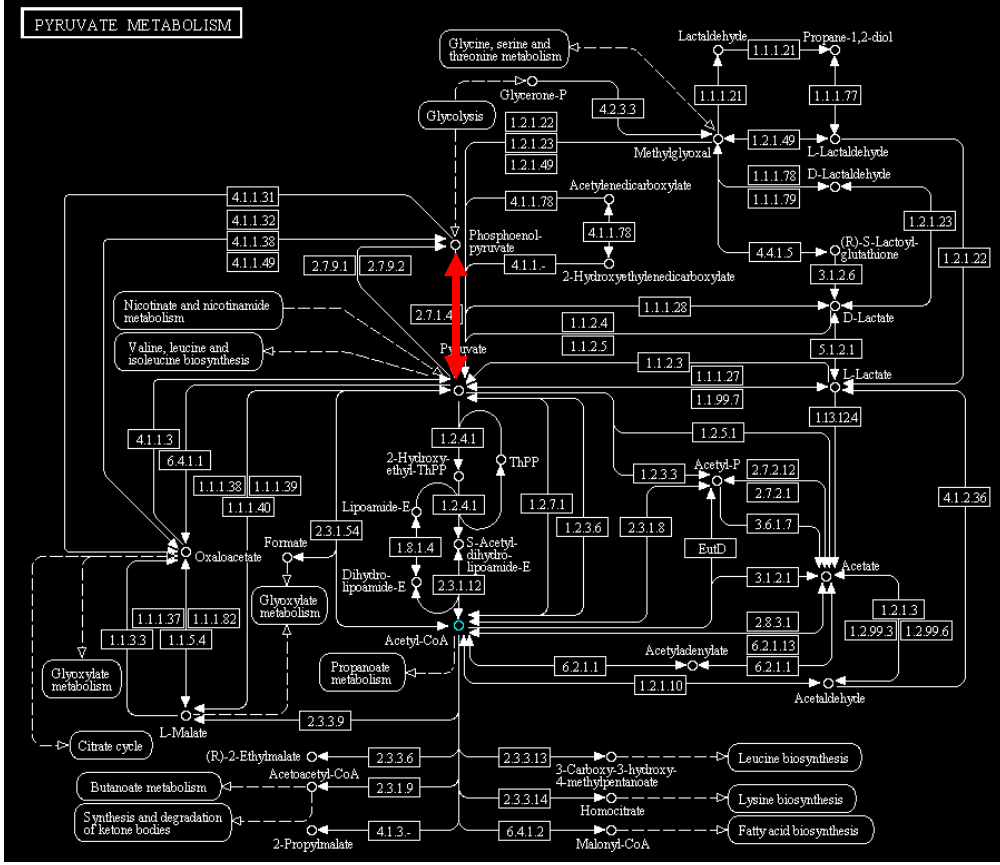
This is an introduction to a series of tutorials for metabolomic data analysis

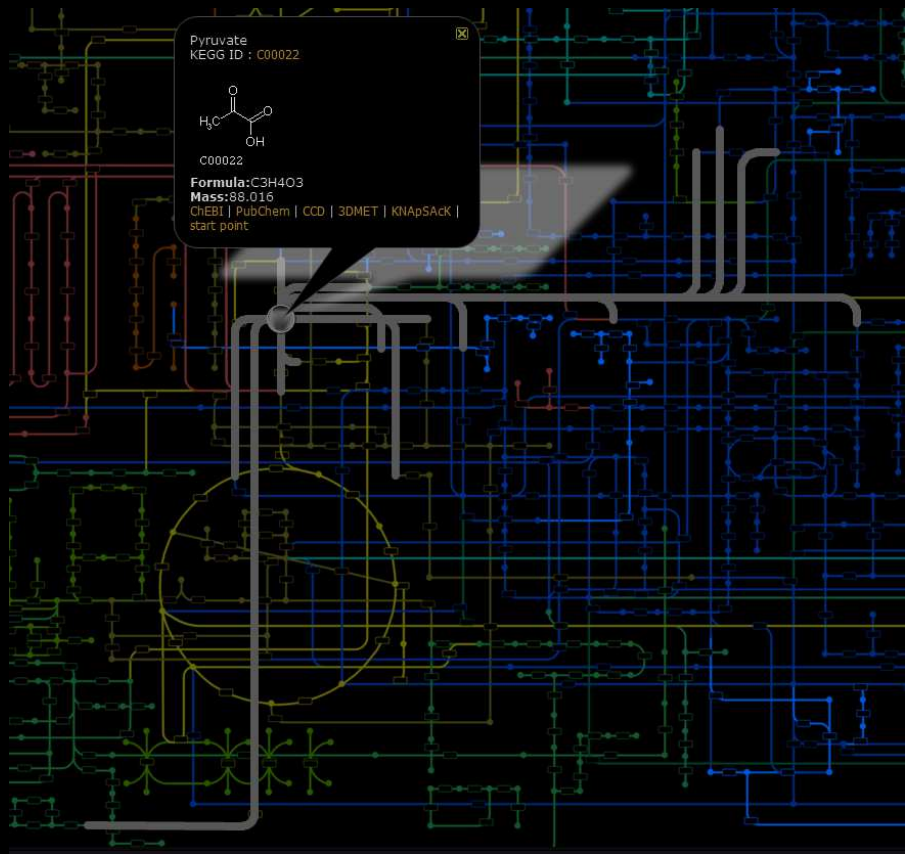
Introduction

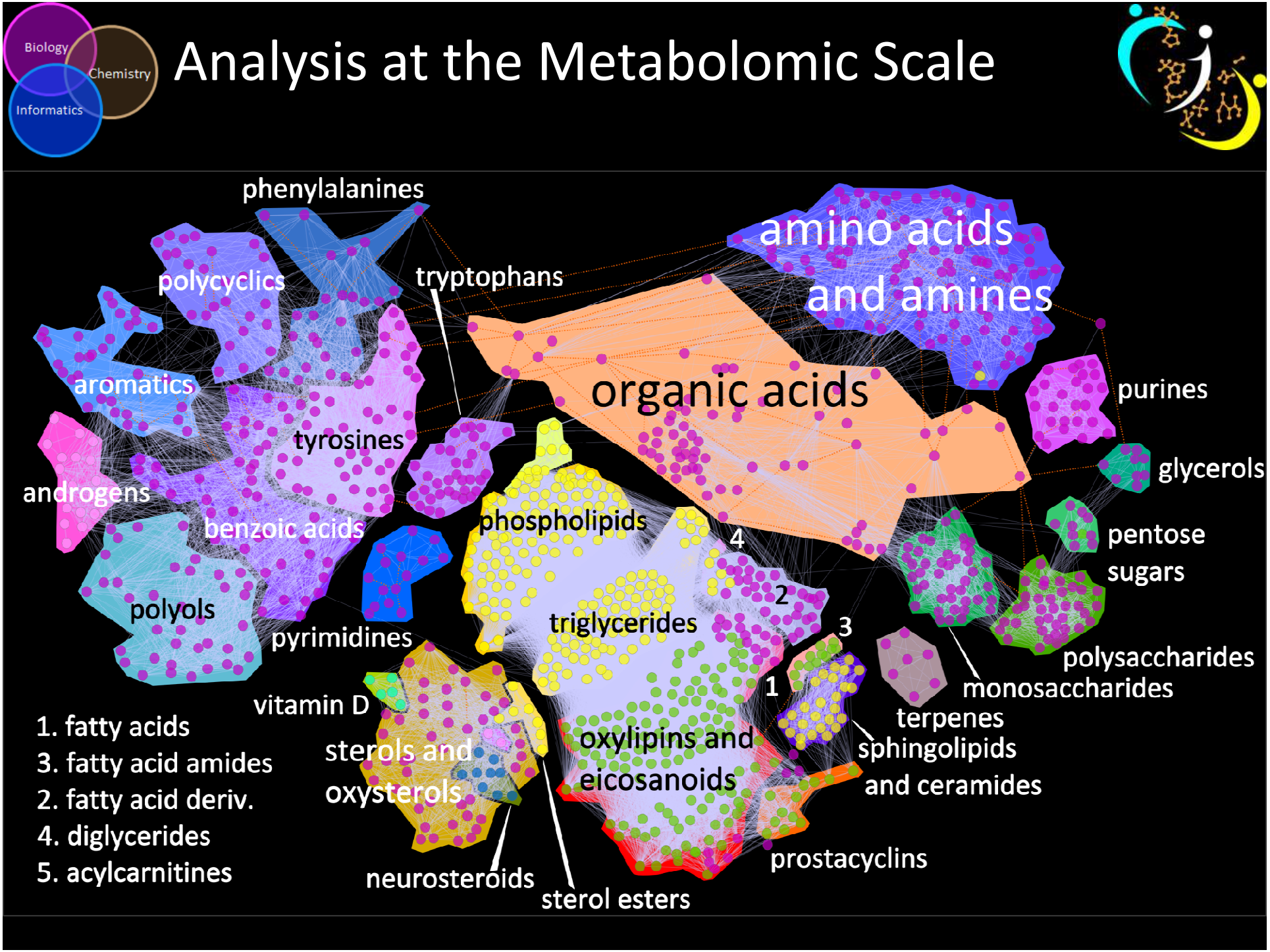
1. Download all the required files and software at: TBD
2. Make sure you have installed R (≥v 3.1.1, http://cran.us.r-project.org/), shiny (v0.10.2.2, http://shiny.rstudio.com/) and a modern browser (e.g., Chrome).
3. Run the code in the folder software the file startup.R to launch all accompanying software; or download most current versions at the links below
   - **DeviumWeb** (https://github.com/dgrapov/DeviumWeb)
   - **MetaMapR** (https://github.com/dgrapov/MetaMapR)
4. Have a great time!

Goals?

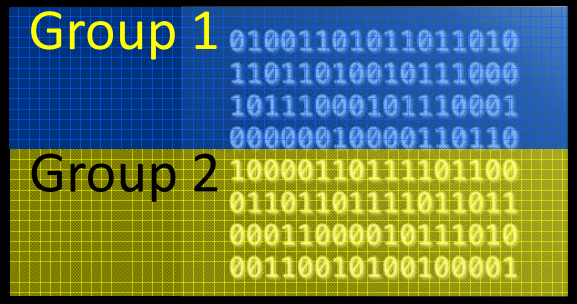

C00022
C00074
C00001
C00002
C00020
C00009


PYRUVATE METABOLISM
Glycine, serine and threonine metabolism
Glycerone-P
Glycolysis
Lactaldehyde
Propane-1,2-diol
Methylglyoxal
L-Lactaldehyde
D-Lactaldehyde
Acetylenedicarboxylate
Phosphoenol-pyruvate
2-Hydroxyethylenedicarboxylate
(R)-S-Lactoyl-glutathione
Nicotinate and nicotinamide metabolism
Valine, leucine and isoleucine biosynthesis
D-Lactate
L-Lactate
2-Hydroxy-ethyl-ThPP
ThPP
Lipoamide-E
S-Acetyl-dihydro-lipoamide-E
Dihydro-lipoamide-E
Acetyl-P
EutD
Acetate
Formate
Oxaloacetate
Glyoxylate metabolism
Acetyl-CoA
Propanoate metabolism
Acetyladenylate
Acetaldehyde
L-Malate
Citrate cycle
(R)-2-Ethylmalate
Acetoacetyl-CoA
Butanoate metabolism
Synthesis and degradation of ketone bodies
2-Propylmalate
3-Carboxy-3-hydroxy-4-methylpentanoate
Homocitrate
Malonyl-CoA
Leucine biosynthesis
Lysine biosynthesis
Fatty acid biosynthesis


Pyruvate
KEGG ID : C00022
C00022
Formula:C3H4O3
Mass:88.016
ChEBI | PubChem | CCD | 3DMET | KNApSAcK |
start point

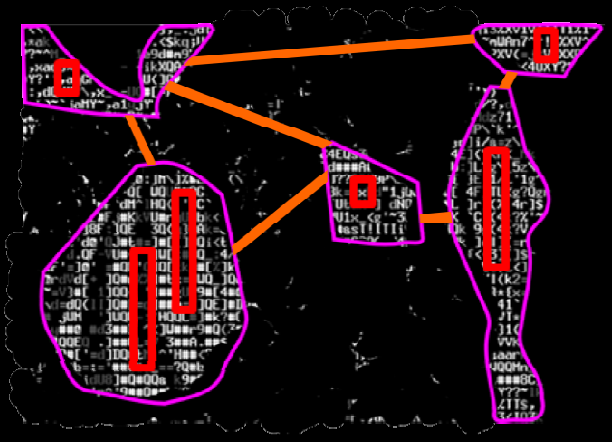

Biology
Chemistry
Informatics
Analysis at the Metabolomic Scale
phenylalanines
polycyclics
tryptophans
amino acids
and amines
aromatics
organic acids
purines
tyrosines
glycerols
androgens
phospholipids
4
pentose
sugars
benzoic acids
2
polyols
triglycerides
3
pyrimidines
polysaccharides
monosaccharides
1
vitamin D
terpenes
sterols and
oxysterols
oxylipins and
eicosanoids
sphingolipids
and ceramides
prostacyclins
neurosteroids
sterol esters
1. fatty acids
3. fatty acid amides
2. fatty acid deriv.
4. diglycerides
5. acylcarnitines

Statistical and Multivariate Analyses

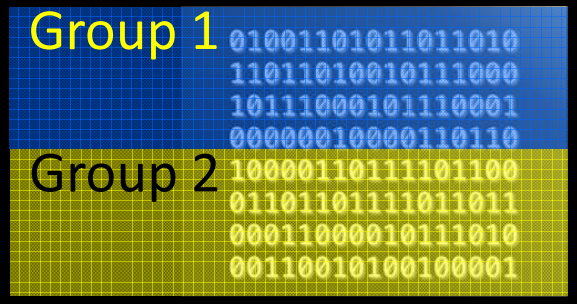


What analytes are different between the two groups of samples?

Statistics
+
Multivariate
+
Context
=

Statistical	Multivariate	Network Mapping
t-Test	O-PLS-DA	
		
significant differences lacking rank and context	ranked differences lacking significance and context	Ranked statistically significant differences within a a biochemical context

Statistical and Multivariate Analyses



What **analytes** are different between the **two groups of samples**?

Statistics
+
Multivariate
+
Context
=

Statistical	Multivariate	Network Mapping
t-Test	O-PLS-DA	

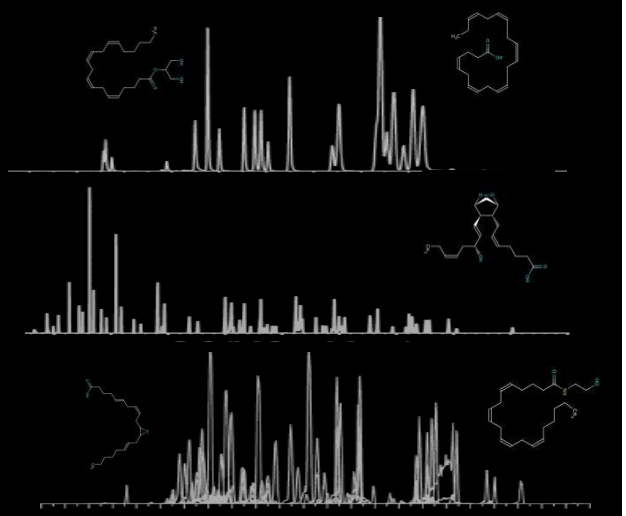
To see the big picture it is necessary too view the data from multiple different angles

Cycle of Scientific Discovery

Hypothesis

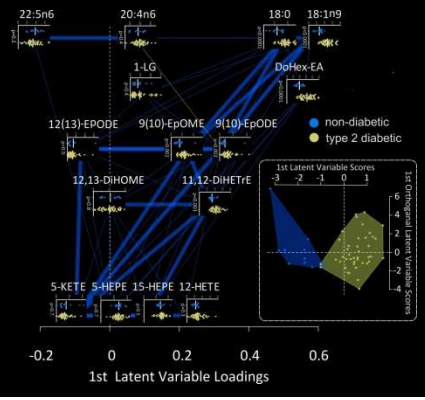
Data Acquisition

Data Processing

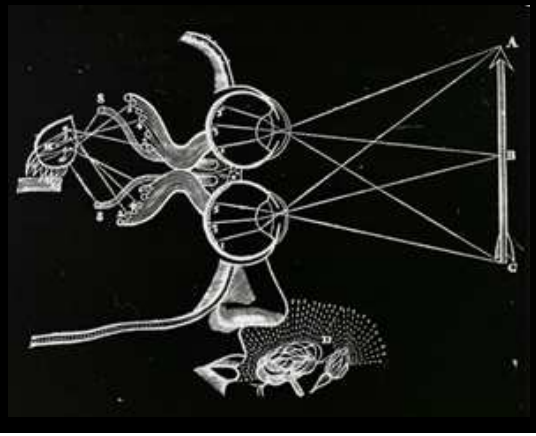


Data

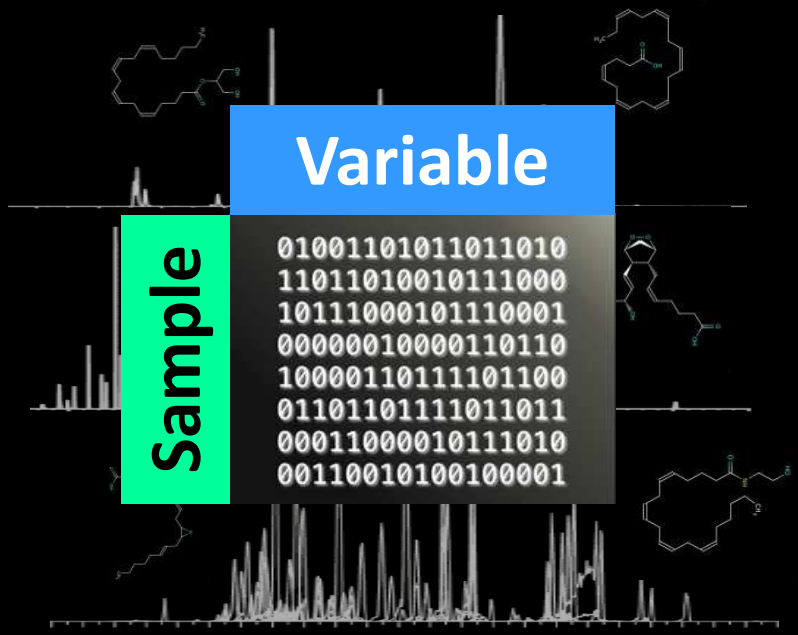
Data Analysis



Hypothesis Generation



Data



Sample | Variable

experimental design

- organism, sex, age etc.

analyte description and metadata

- biochemical class, mass spectra, etc.

Analysis and Visualization

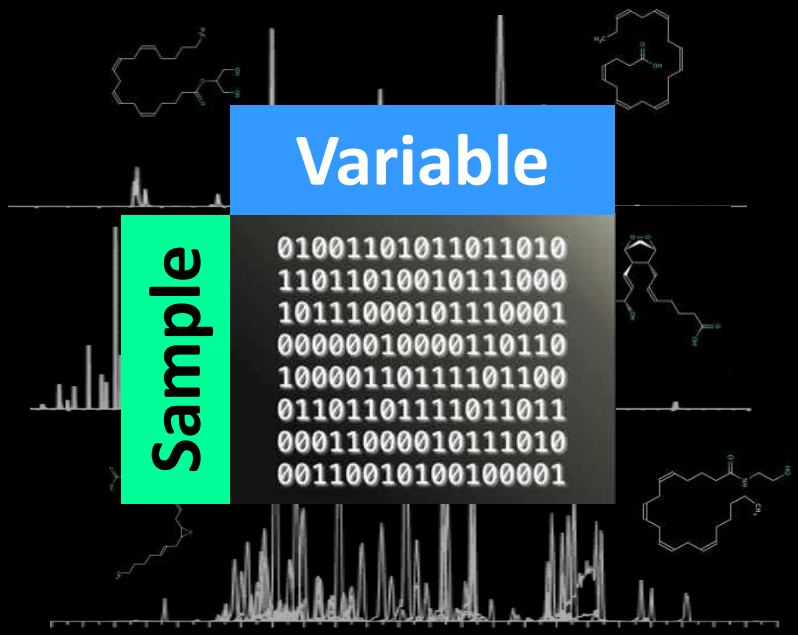
Quality Assessment

Statistical and Multivariate

Functional

Network

Data



Sample | Variable

experimental design

- organism, sex, age etc.

analyte description and metadata

- biochemical class, mass spectra, etc.

Analysis and Visualization

Quality Assessment

- use replicated mesurements and/or internal standards to estimate analytical variance

Statistical and Multivariate

- use the experimental design to test hypotheses and/or identify trends in analytes

Functional

- use statistical and multivariate results to identify impacted biochemical domains

Network

- integrate statistical and multivariate results with the experimental design and analyte metadata

Network Mapping

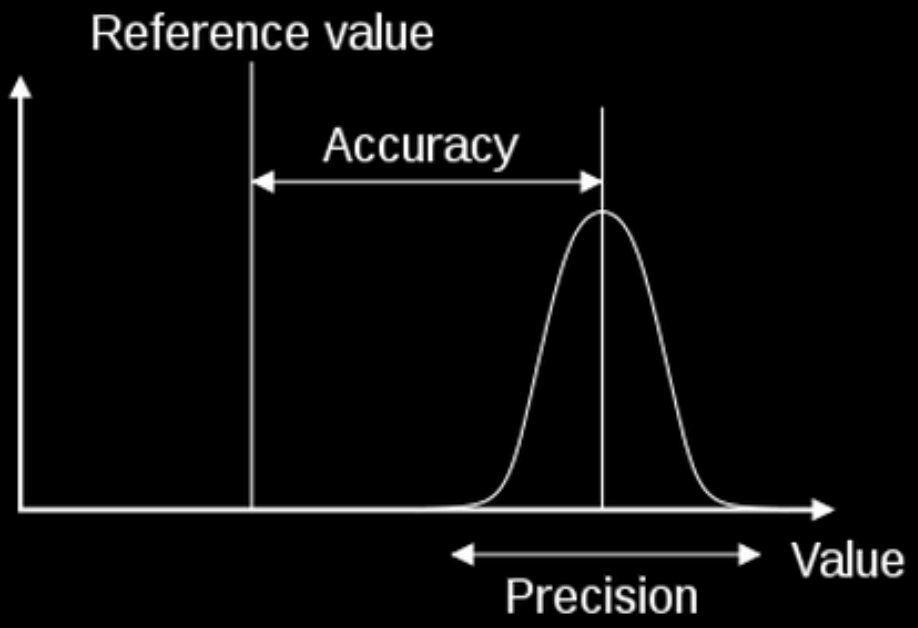
Data Quality Assessment

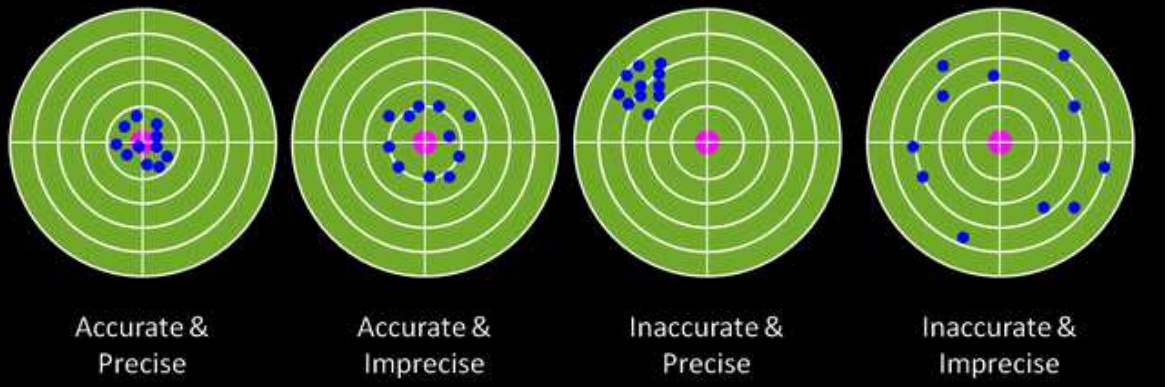
Quality metrics

- Precision (replicated measurements)
- Accuracy (reference samples)

Common tasks

- normalization
- outlier detection
- missing values imputation

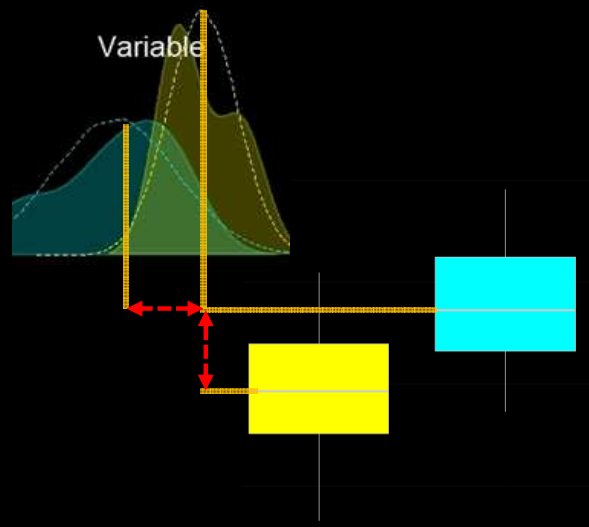


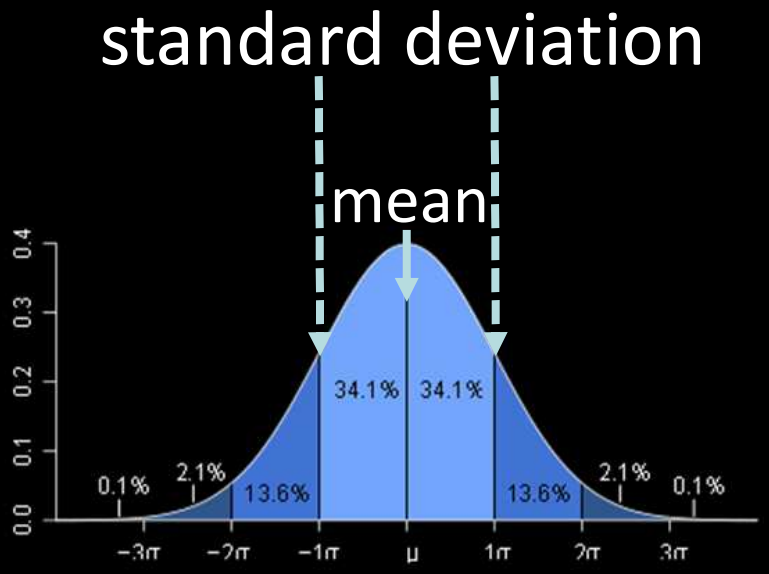


*Finish lab: 1-Data quality

Univariate Qualities

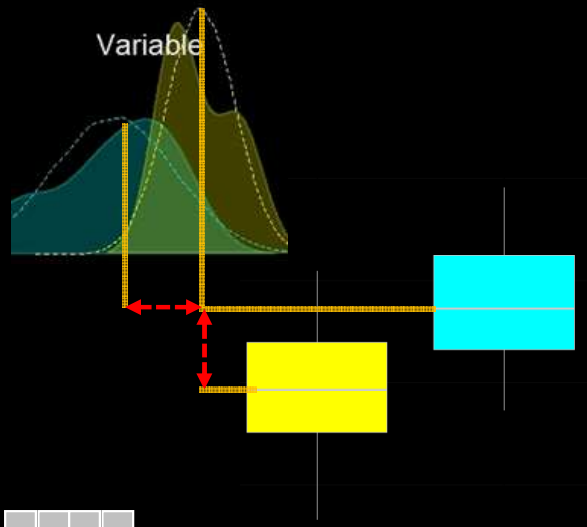
- length (sample size)
- center (mean, median, geometric mean)
- dispersion (variance, standard deviation)
- range (min / max),
- quantiles
- shape (skewness, kurtosis, normality, etc.)

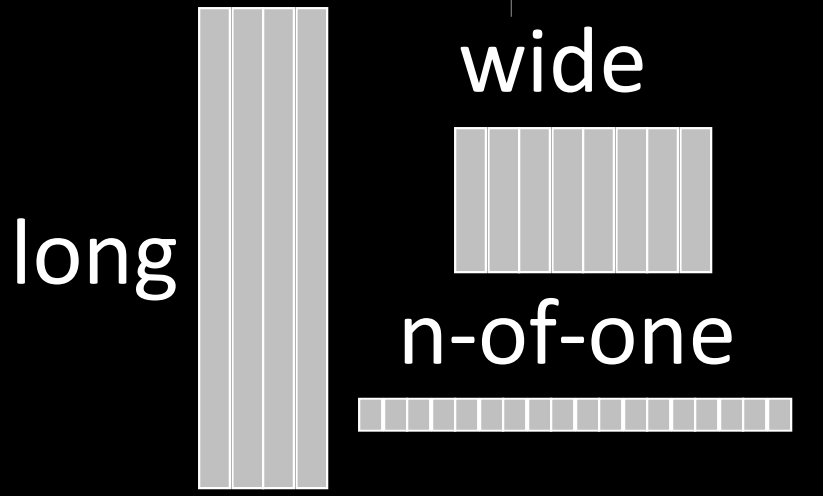




Univariate Analyses

- **Identify differences in sample population means**
- **sensitive to distribution shape**
  - **parametric = assumes normality**
- **error in Y, not in X (Y = mX + error)**
- **optimal for long data**
- **assumed independence**
- **false discovery rate (FDR)**



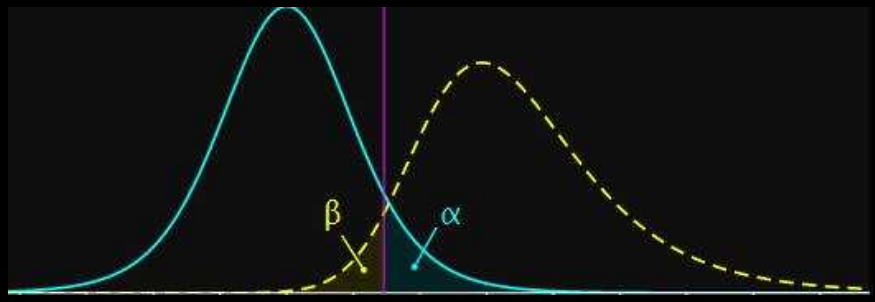


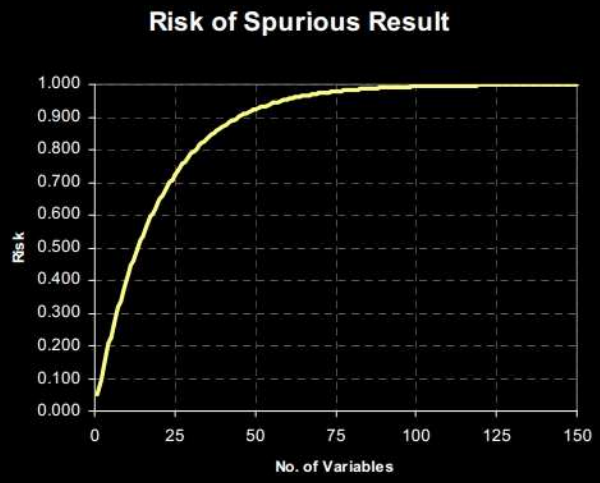
False Discovery Rate (FDR)

Type I Error: False Positives

- **Type II Error: False Negatives**
- **Type I risk =**
  - **$1-(1-p.value)^m$**

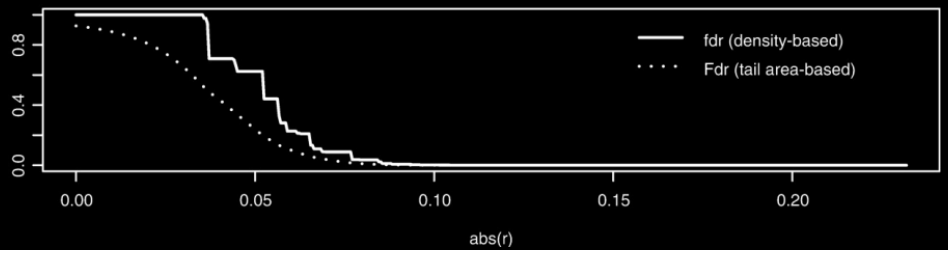
m = number of variables tested





FDR correction

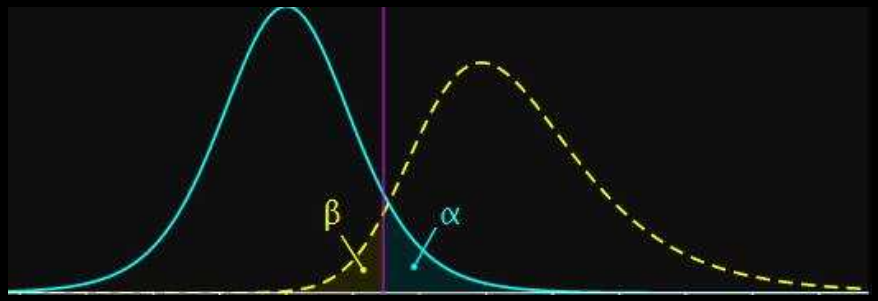
- **p-value adjustment or estimate of FDR (Fdr, q-value)**

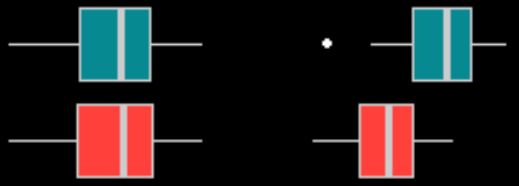


Bioinformatics (2008) 24 (12):1461-1462

Statistical Analysis: achieving ‘significance’

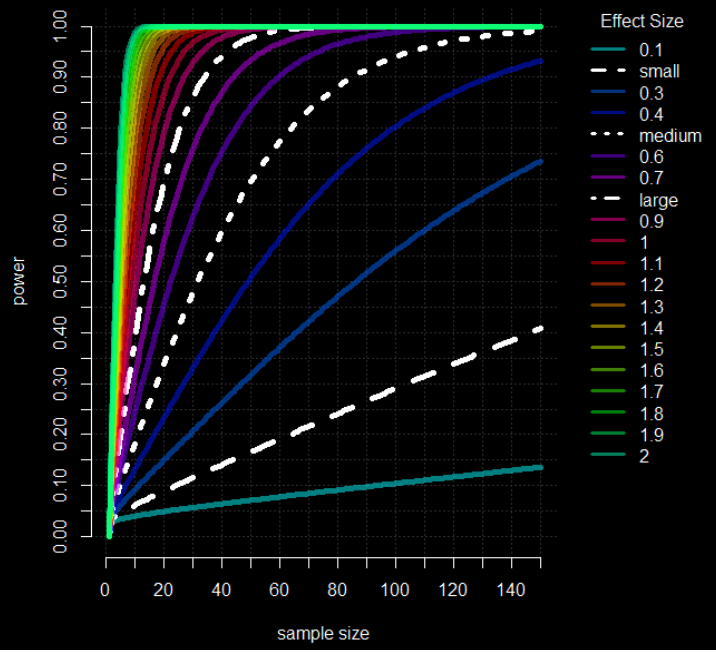
significance level (α) and power (1-β)





effect size (standardized difference in means)

sample size (n)



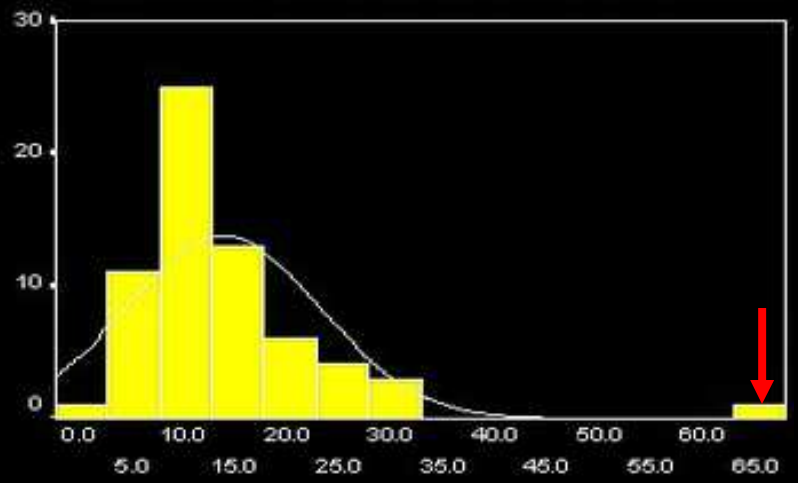
*finish lab 2-statistical analysis

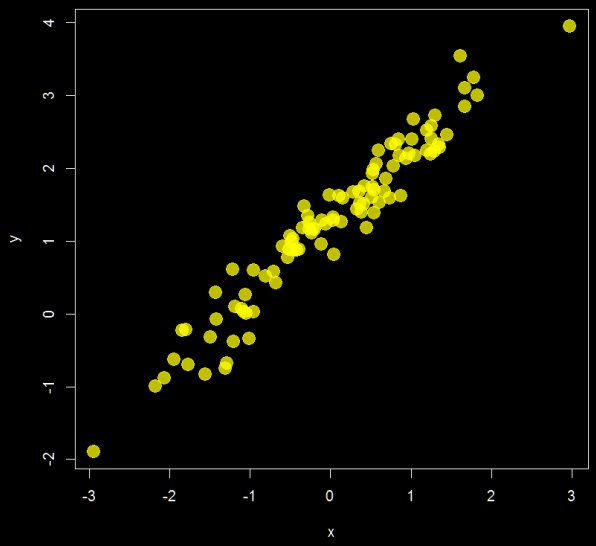
Power analyses can be used to optimize future experiments given preliminary data

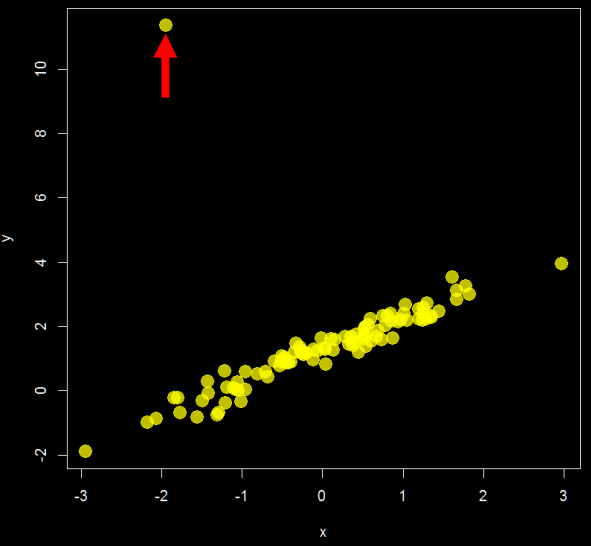
Example: use experimentally derived (or literature estimated) effect sizes, desired p-value (alpha) and power (beta) to calculate the optimal number of samples per group

Outlier Detection

- 1 variable (univariate)
- 2 variables (bivariate)
- >2 variables (multivariate)

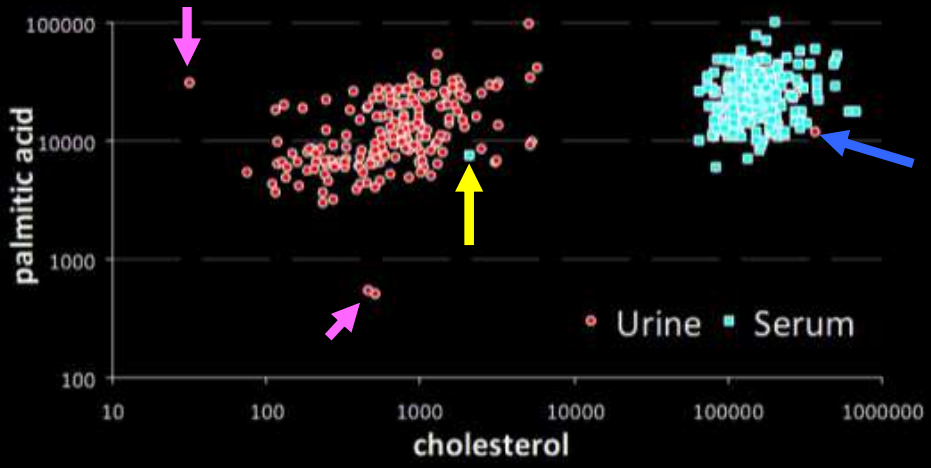






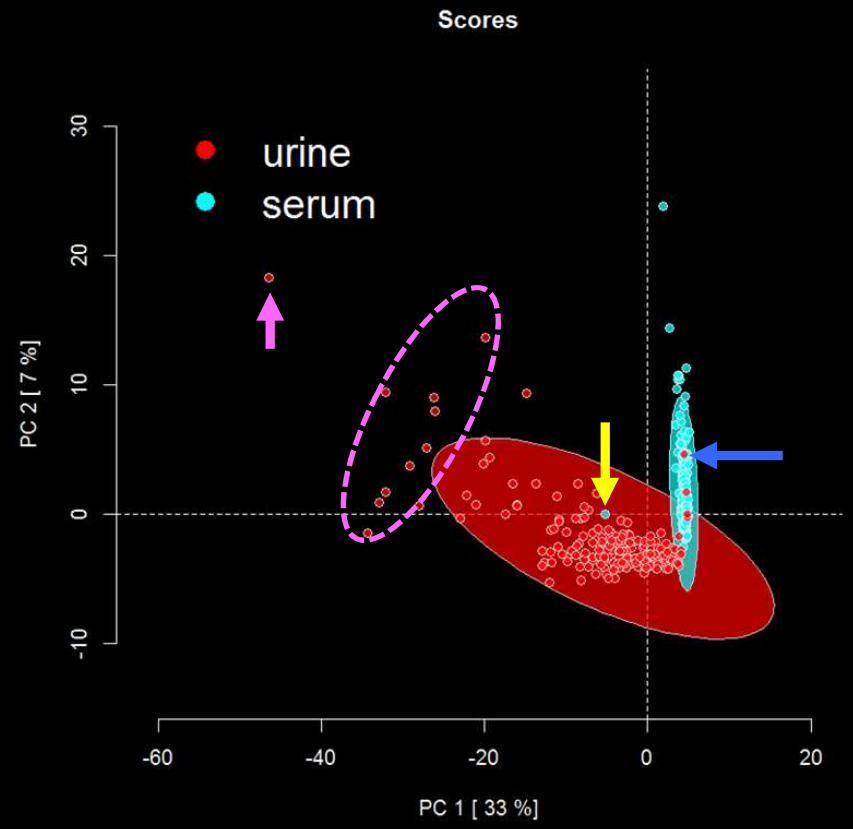
Outlier Detection

bivariate
(scatter plot)



vs.

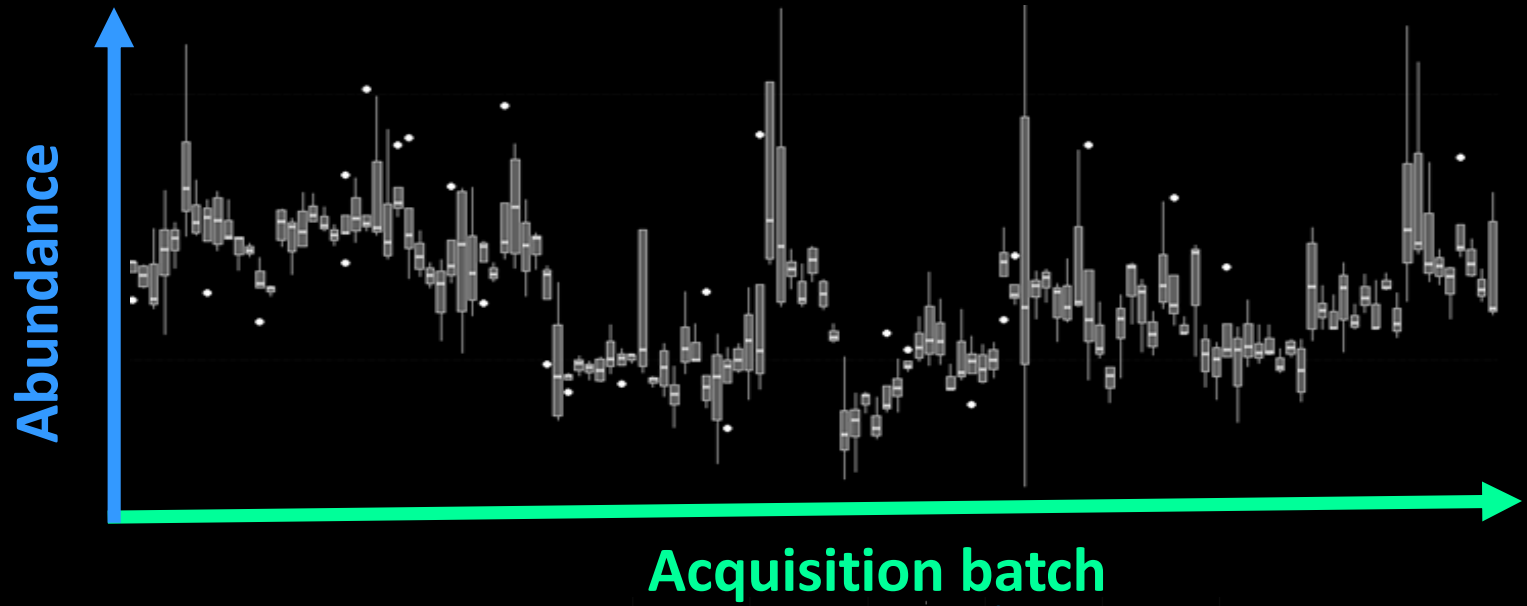
multivariate
(PCA scores plot)



→outliers?
⇉mixed up samples

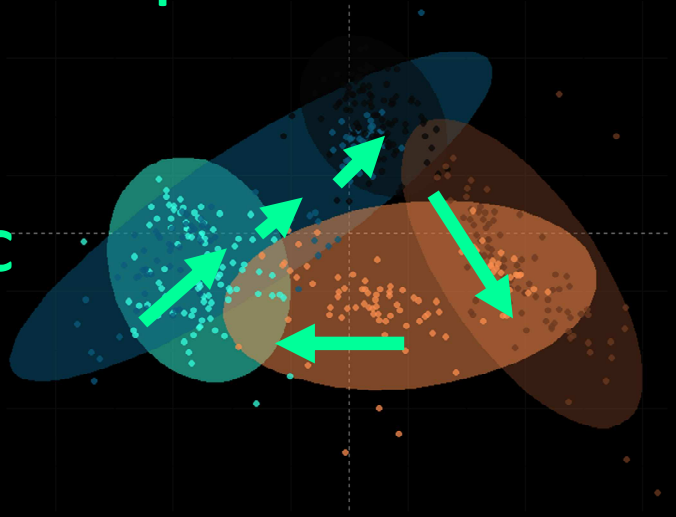
Batch Effects

Drift in >400 replicated measurements across >100 analytical batches for a single analyte



QCs embedded among >5,5000 samples (1:10) collected over 1.5 yrs

Principal Component Analysis (PCA) of all analytes, showing QC sample scores

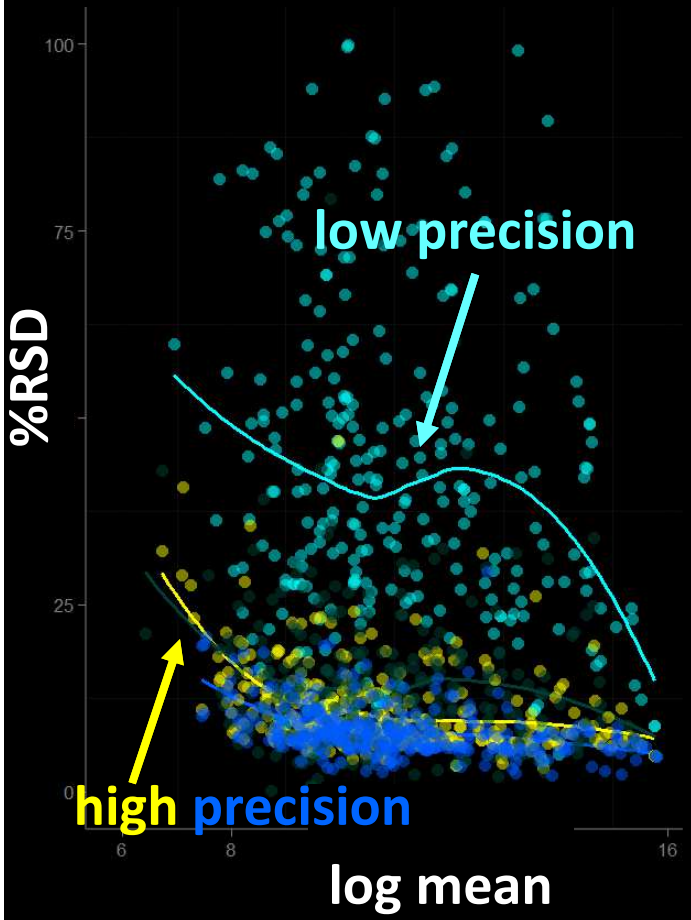


If the biological effect size is less than the analytical variance then the experiment will incorrectly yield insignificant results

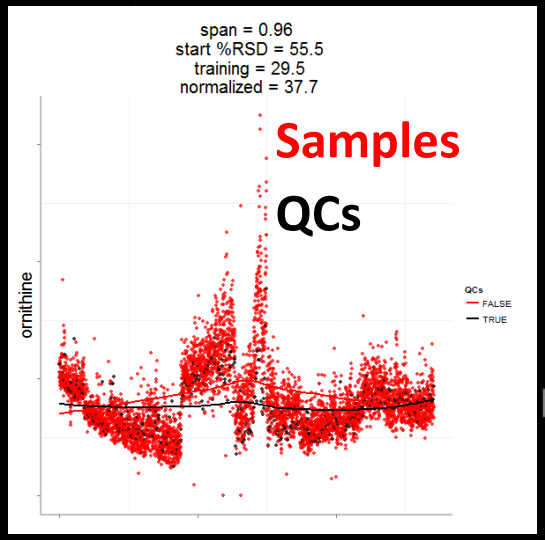
Batch Effects

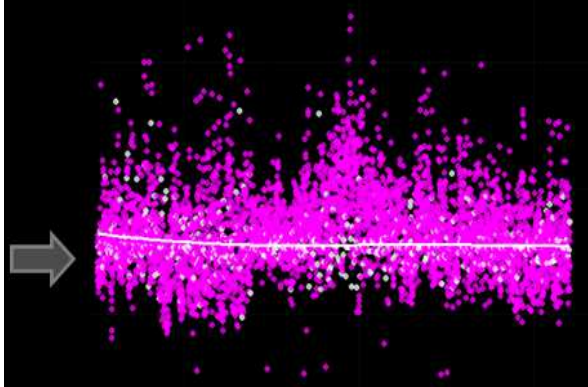
Analyte specific data quality overview

Sample specific normalization can be used to estimate and remove **analytical variance**



Raw Data → Normalized Data





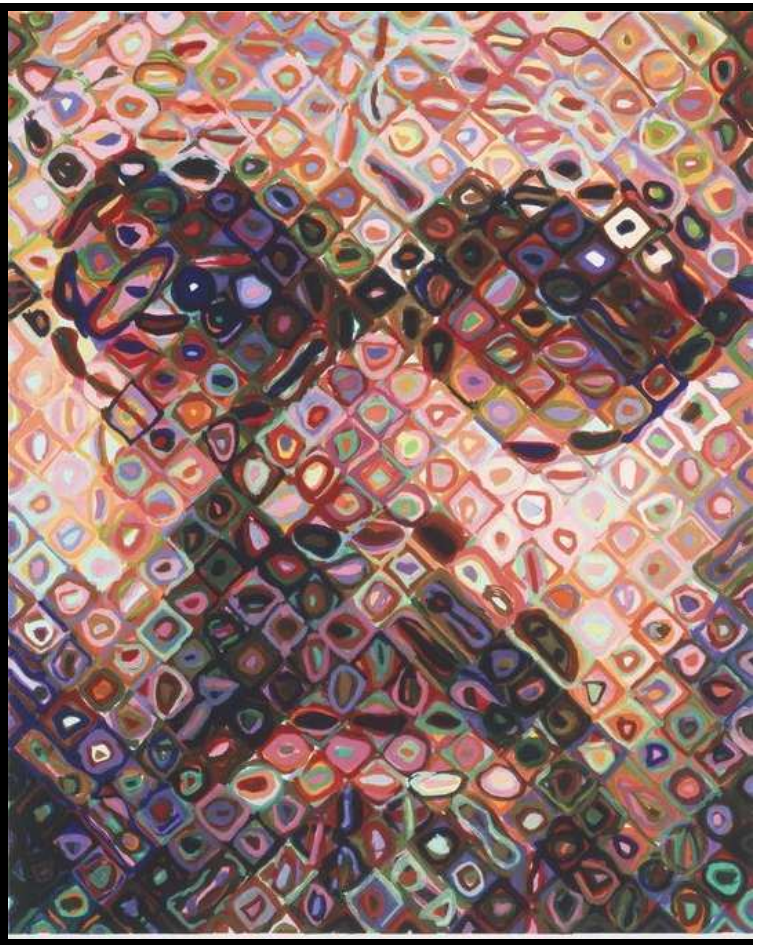
*finish lab 3-Data Normalization

Clustering

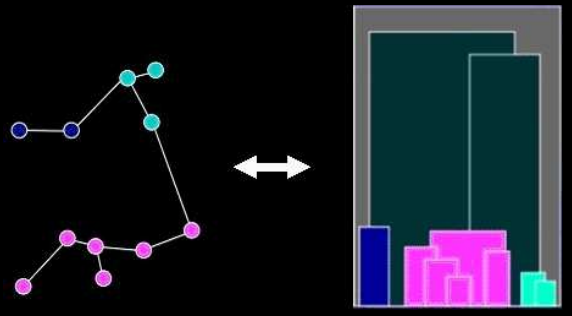
Identify

- **patterns**
- **group structure**
- **relationships**

- **Evaluate/refine hypothesis**
- **Reduce complexity**



Artist: Chuck Close



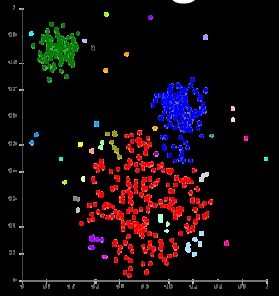
Cluster Analysis

Use the concept similarity/dissimilarity to group a collection of samples or variables

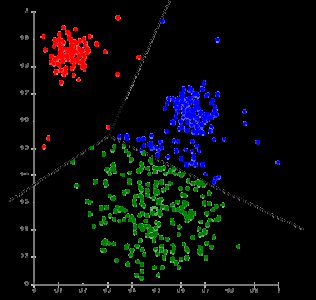
Approaches

- **hierarchical (HCA)**
- **non-hierarchical (k-NN, k-means)**
- **distribution (mixtures models)**
- **density (DBSCAN)**
- **self organizing maps (SOM)**

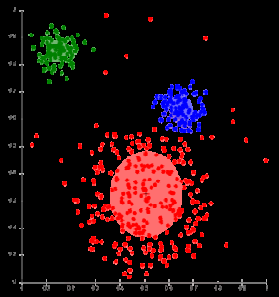
Linkage



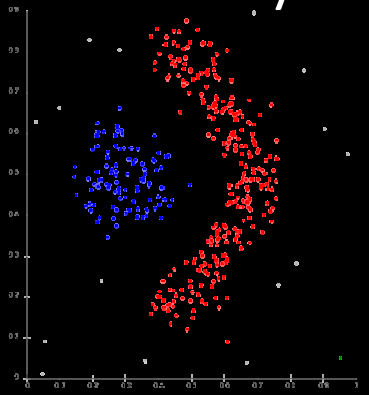
k-means



Distribution

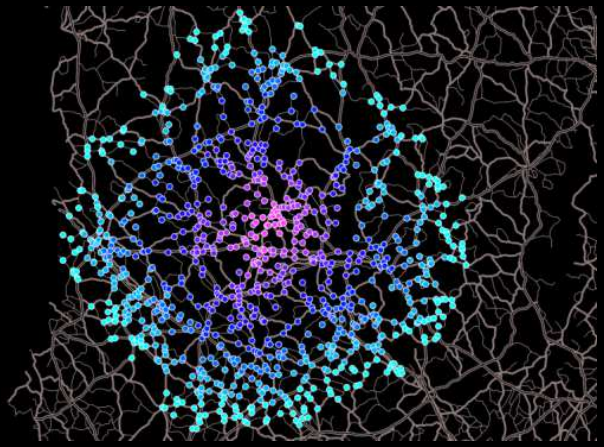


Density

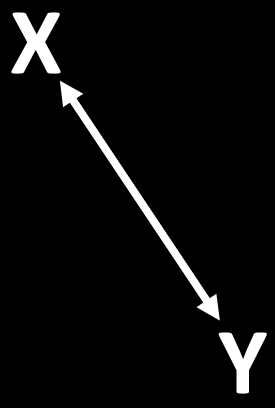


Hierarchical Cluster Analysis

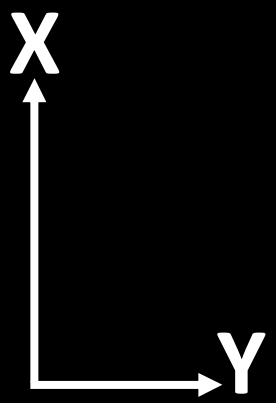
- similarity/dissimilarity defines "nearness" or distance



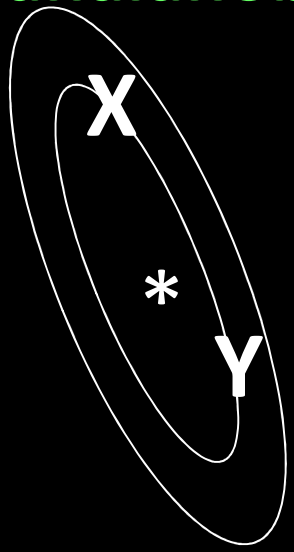
euclidean



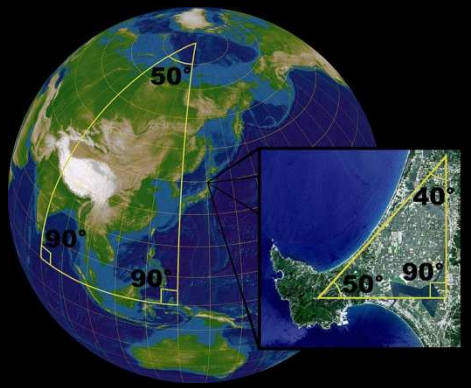
manhattan



Mahalanobis

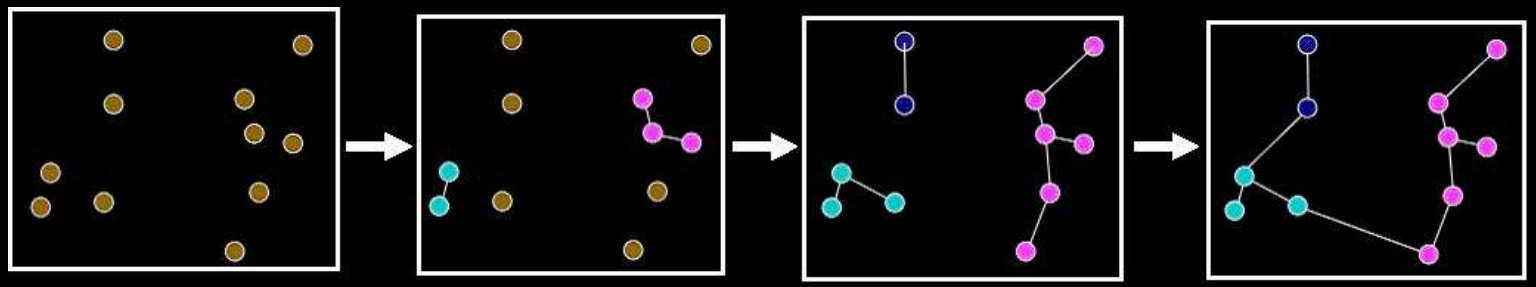


non-euclidean

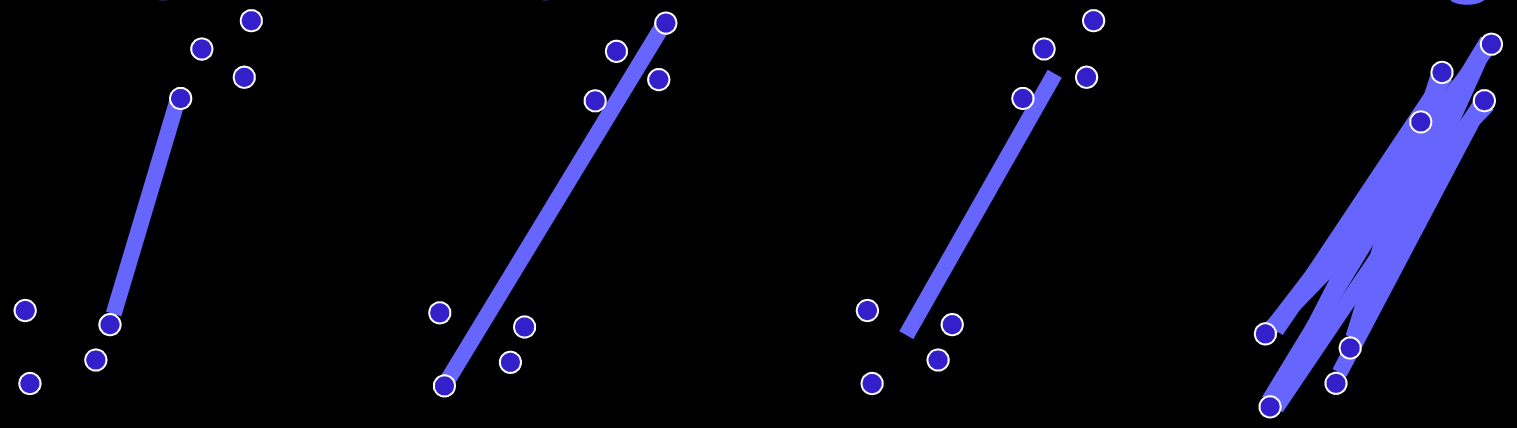


Hierarchical Cluster Analysis

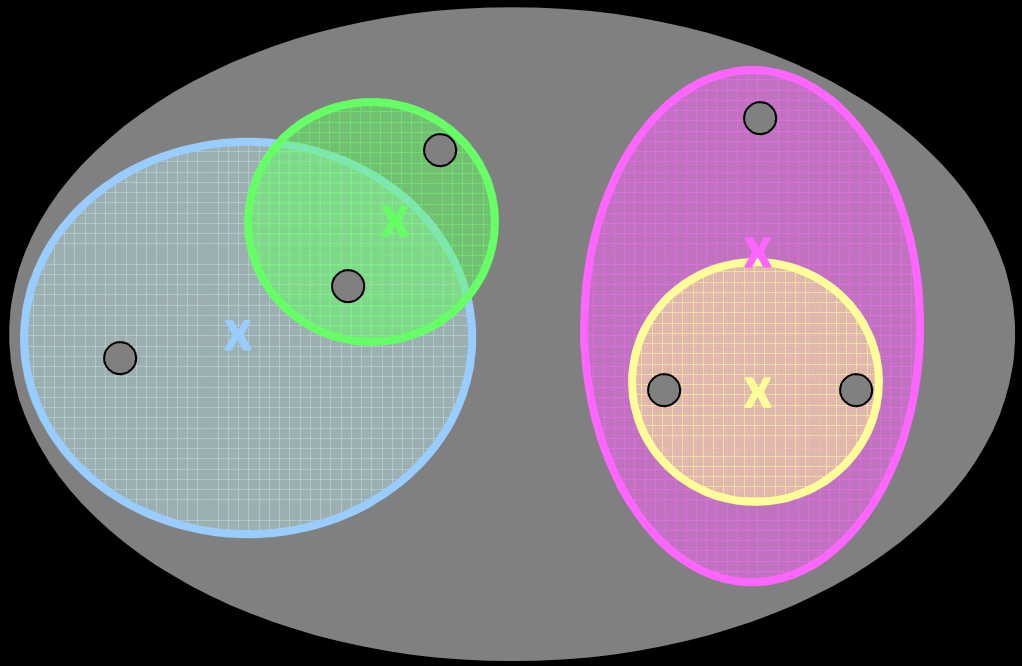
Agglomerative/linkage algorithm defines how points are grouped

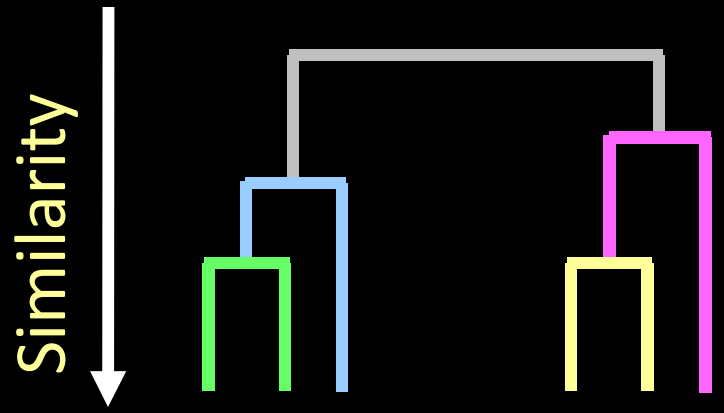


single complete centroid average



Dendrograms

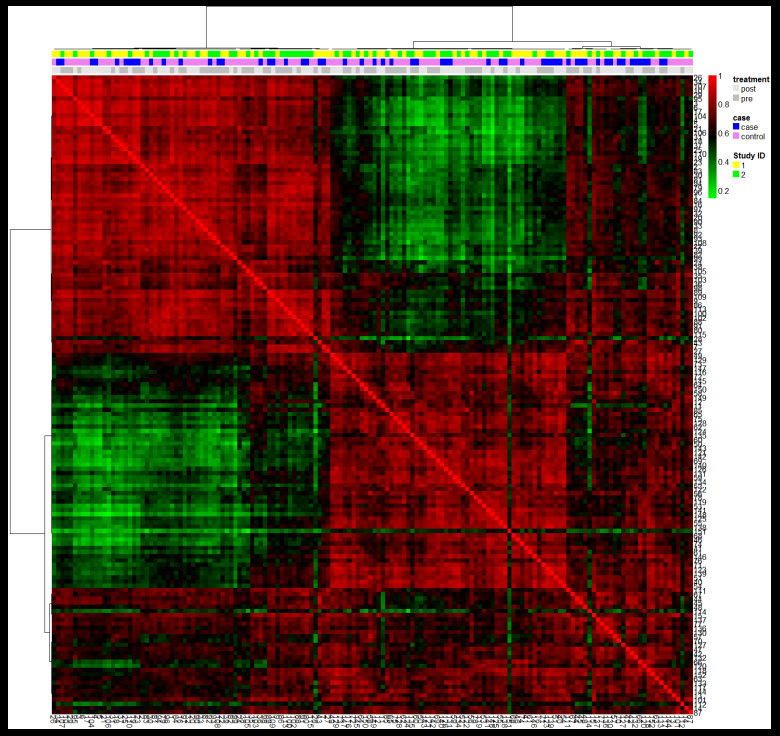




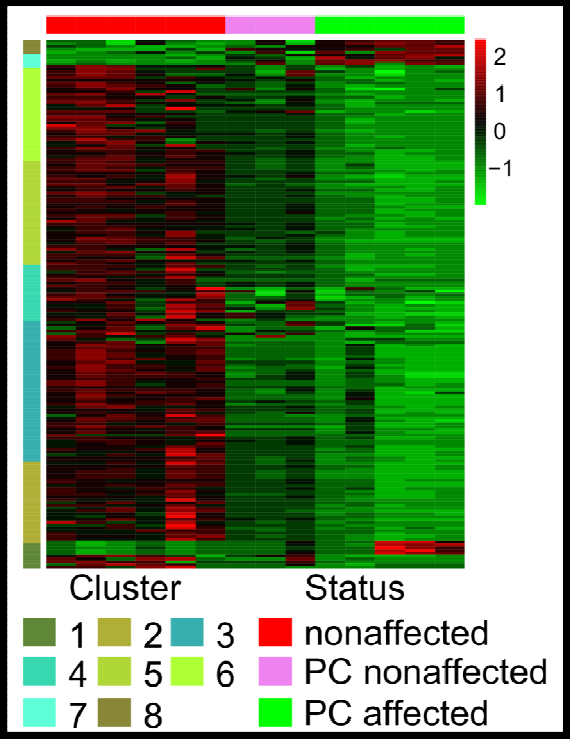
Hierarchical Cluster Analysis

How does my metadata match my data structure?

Exploration

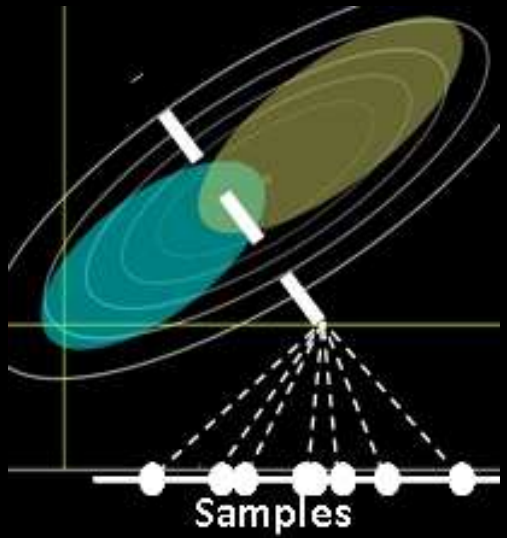


Confirmation



***finish lab 4-Cluster Analysis**

Projection of Data



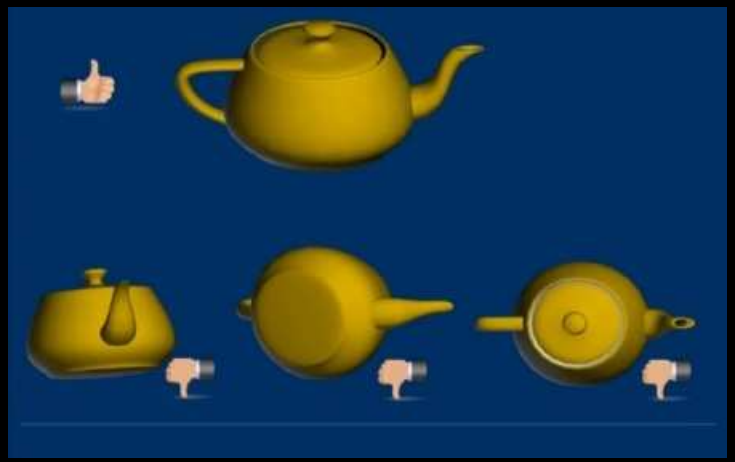
The algorithm defines the position of the light source

Principal Components Analysis (PCA)

- unsupervised
- maximize variance (X)

Partial Least Squares Projection to Latent Structures (PLS)

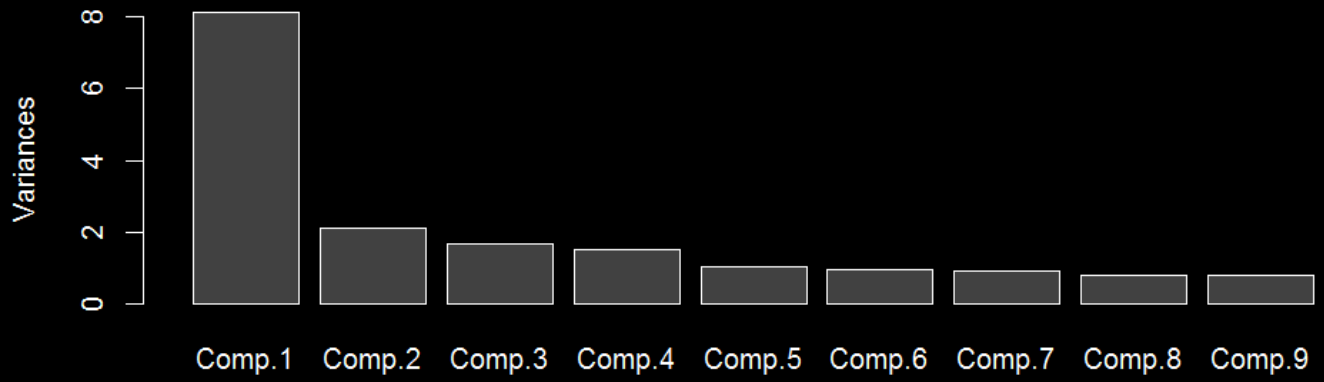
- supervised
- maximize covariance (Y ~ X)



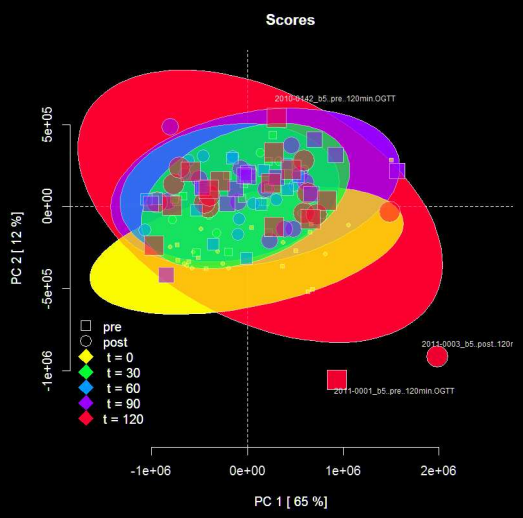
James X. Li, 2009, VisuMap Tech.

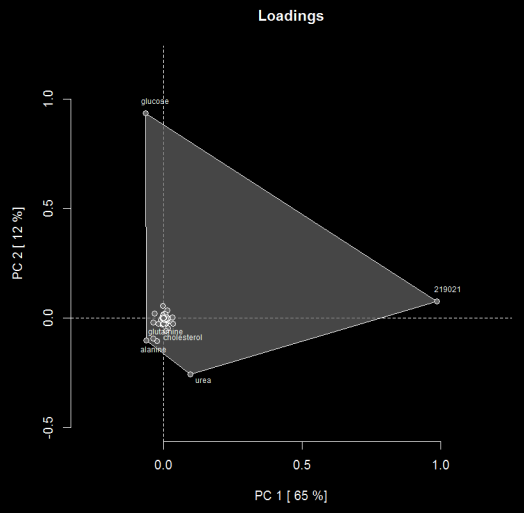
Interpreting PCA Results

Variance explained (eigenvalues)

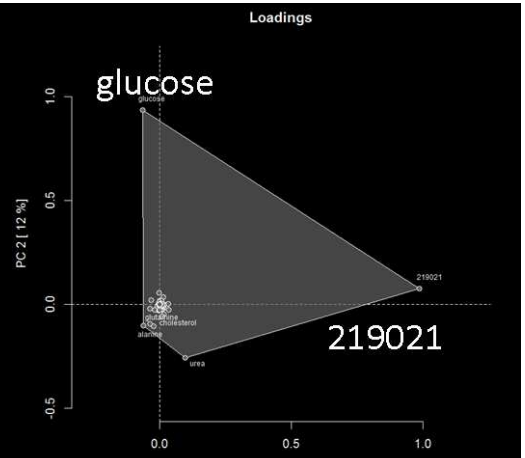


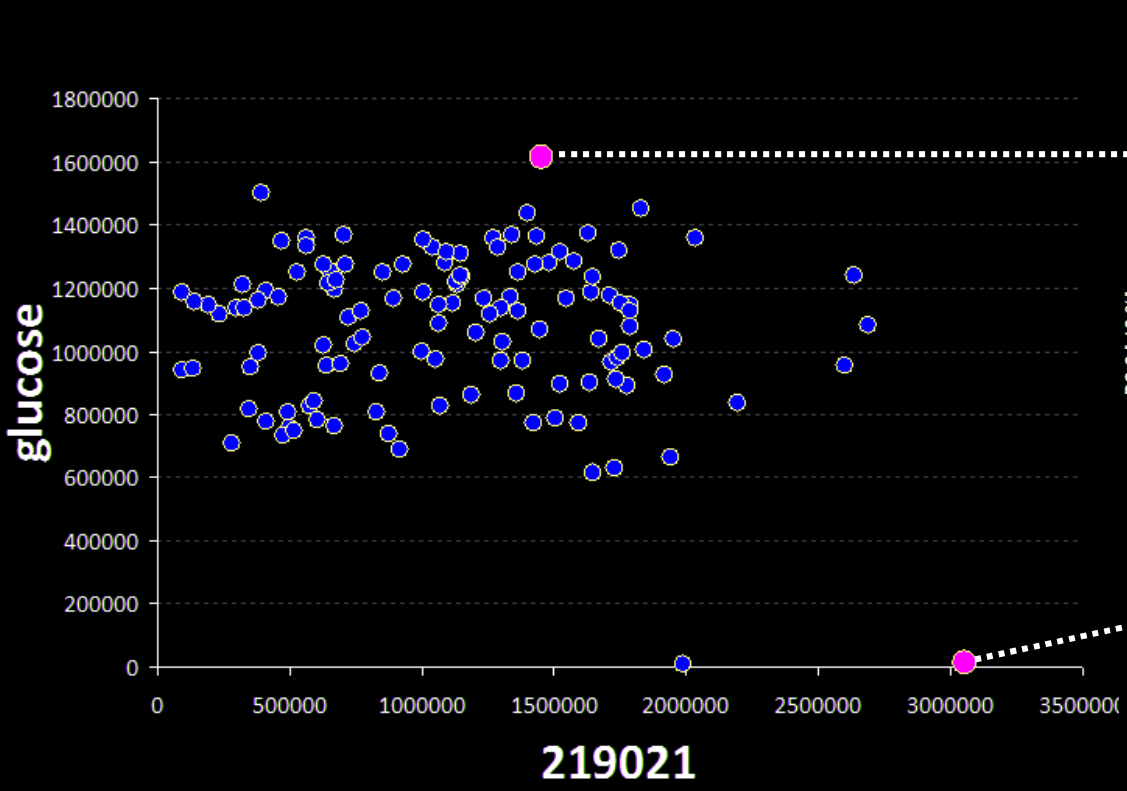
Row (sample) scores and column (variable) loadings

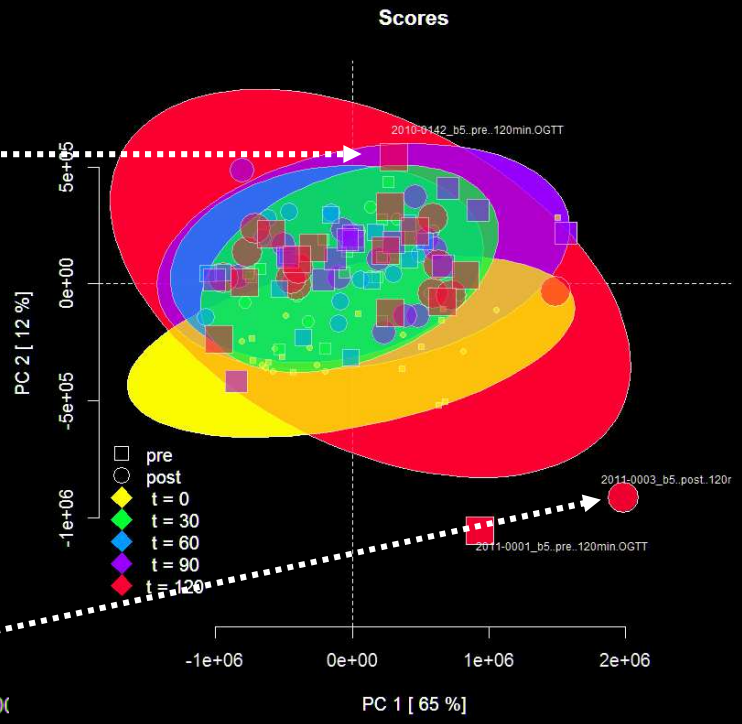




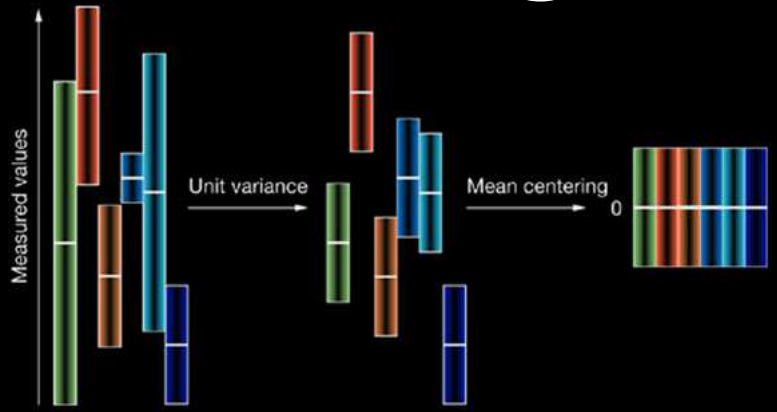
How are scores and loadings related?



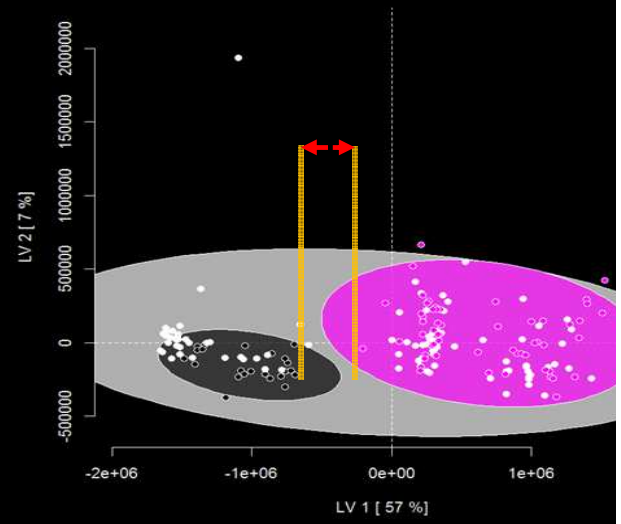


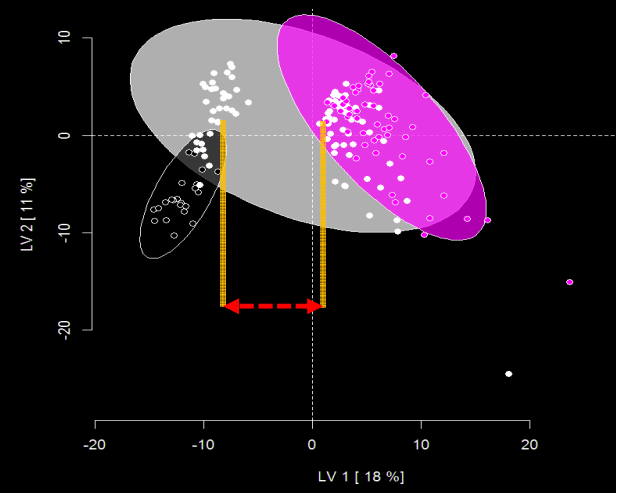


Centering and Scaling



Method	Formula	Unit	Goal	Advantages	Disadvantages
Centering	$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$	0	Focus on the differences and not the similarities in the data	Remove the offset from the data	When data is heteroscedastic, the effect of this pretreatment method is not always sufficient
Autoscaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$	(-)	Compare metabolites based on correlations	All metabolites become equally important	Inflation of the measurement errors
Range scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\left(x_{i_{max}} - x_{i_{min}}\right)}$	(-)	Compare metabolites relative to the biological response range	All metabolites become equally important. Scaling is related to biology	Inflation of the measurement errors and sensitive to outliers
Pareto scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$	0	Reduce the relative importance of large values, but keep data structure partially intact	Stays closer to the original measurement than autoscaling	Sensitive to large fold changes
Vast scaling	$\tilde{x}_{ij} = \frac{\left(x_{ij} - \bar{x}_i\right)}{s_i} \cdot \frac{\bar{x}_i}{s_i}$	(-)	Focus on the metabolites that show small fluctuations	Aims for robustness, can use prior group knowledge	Not suited for large induced variation without group structure
Level scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$	(-)	Focus on relative response	Suited for identification of e.g. biomarkers	Inflation of the measurement errors
Log transformation	$\tilde{x}_{ij} = {}^{10}\log\left(x_{ij}\right)$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	Log 0	Correct for heteroscedasticity, pseudo scaling. Make multiplicative models additive	Reduce heteroscedasticity, multiplicative effects become additive	Difficulties with values with large relative standard deviation and zeros
Power transformation	$\tilde{x}_{ij} = \sqrt{\left(x_{ij}\right)}$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	$\sqrt{0}$	Correct for heteroscedasticity, pseudo scaling	Reduce heteroscedasticity, no problems with small values	Choice for square root is arbitrary.





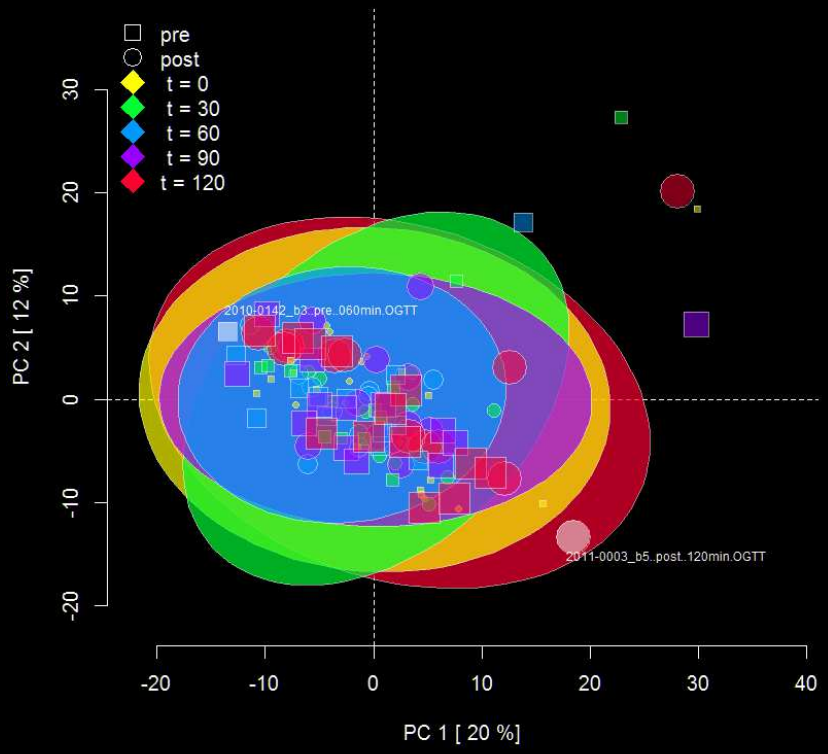
PMID: 16762068

***finish lab 5-Principal Components Analysis**

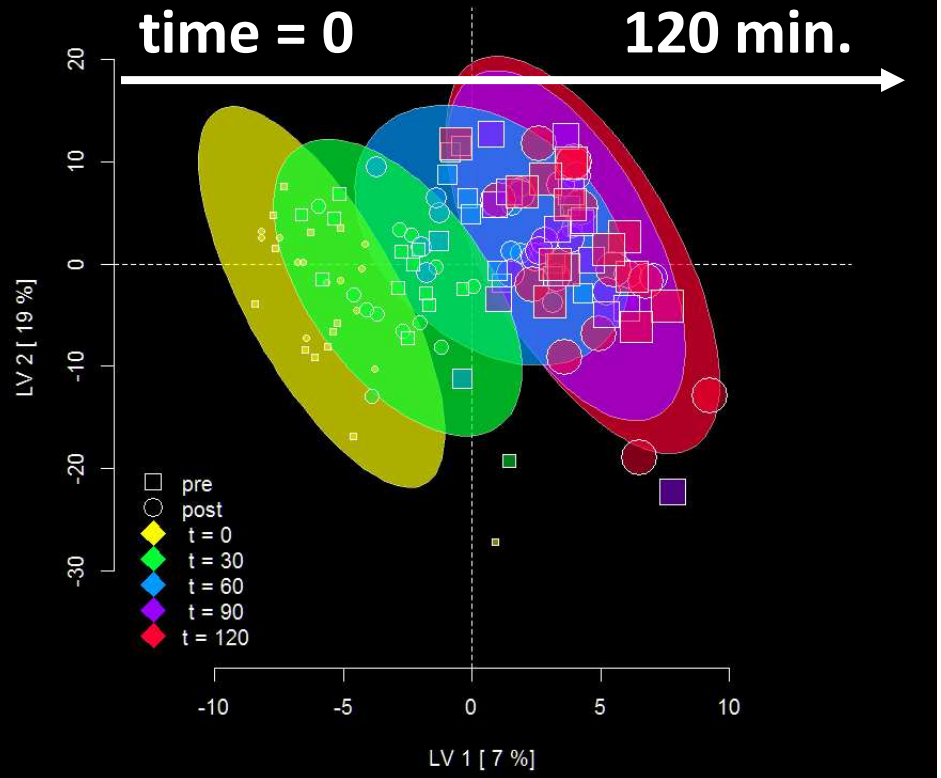
Use PLS to test a hypothesis on a multivariate level

Partial Least Squares (PLS) is used to identify maximum modes of covariance between X measurements and Y (hypotheses)

PCA

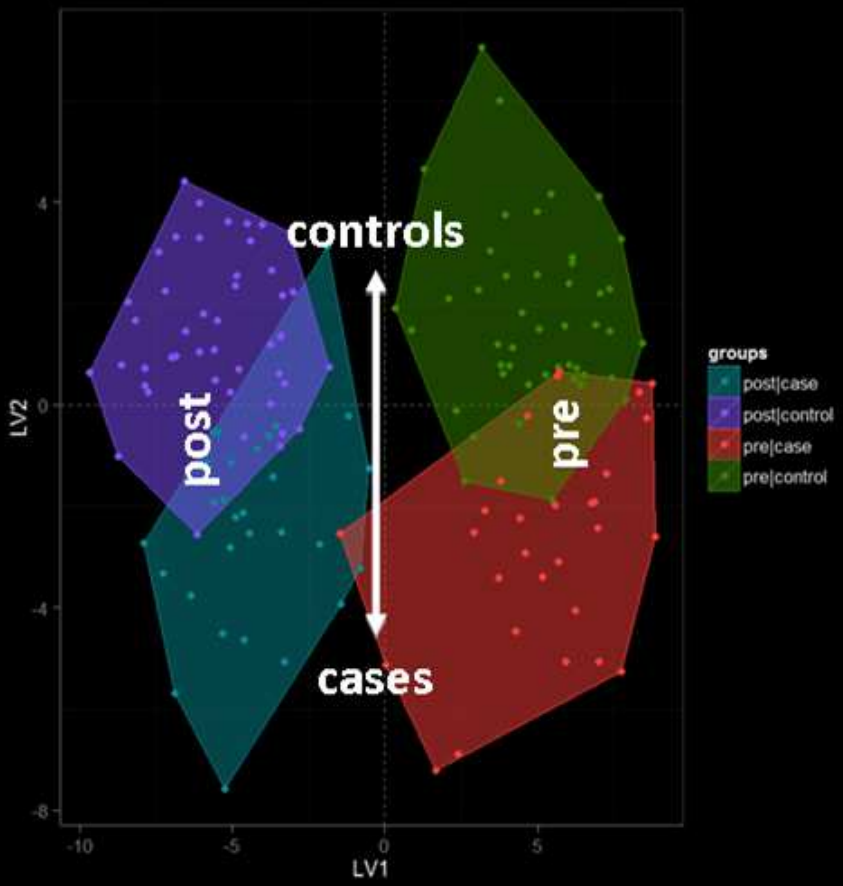


PLS

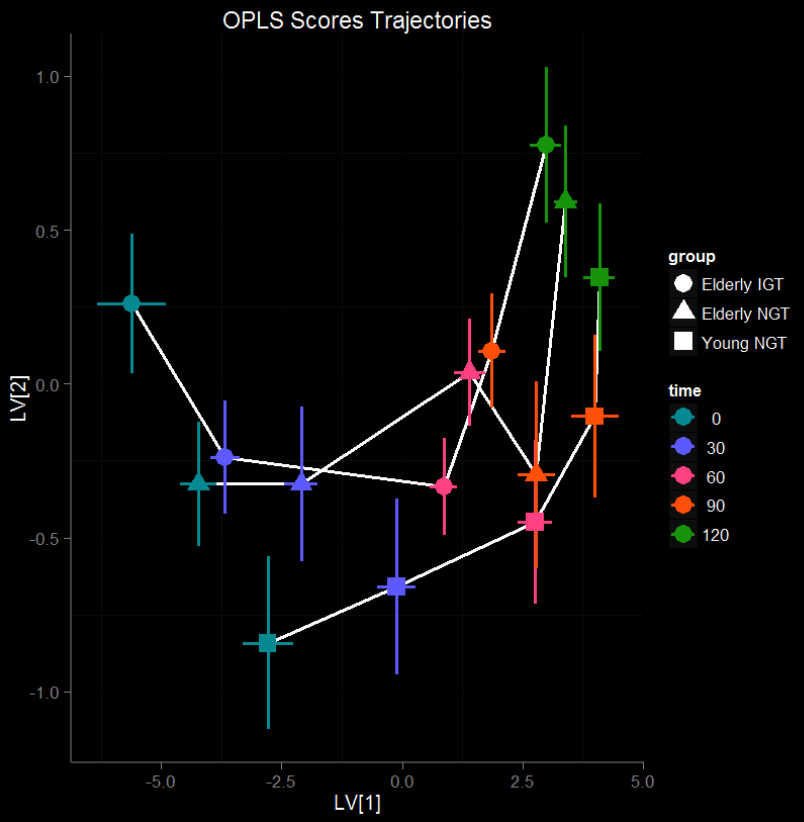


Modeling multifactorial relationships

~two-way ANOVA



dynamic changes among groups



PLS Related Objects

Model

- dimensions, latent variables (LV)
- performance metrics (Q2, RMSEP, etc)
- validation (training/testing, permutation, cross-validation)
- orthogonal correction

Samples

- scores
- predicted values
- residuals

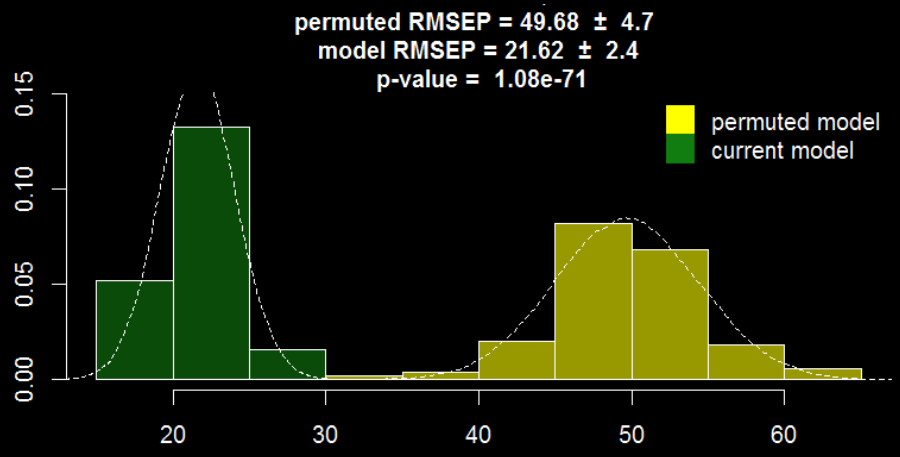
Variables

- Loadings
- Coefficients: summary of loadings based on all LVs
- VIP: variable importance in projection
- Feature selection

"goodness" of the model is all about the perspective

Determine in-sample (Q^2) and out-of-sample error (RMSEP) and compare to a random model

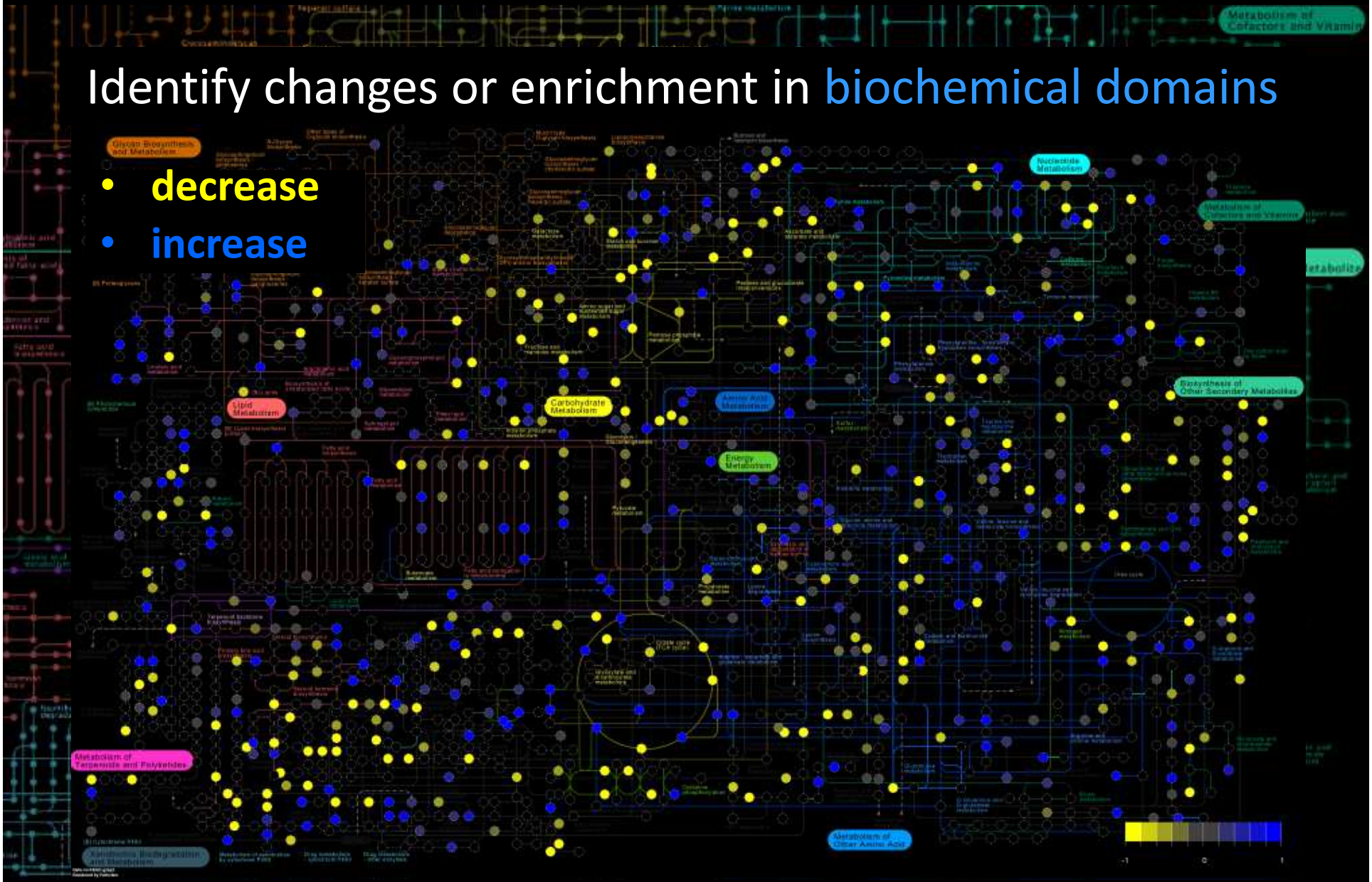
- permutation tests
- training/testing



***finish lab 6-Partial Least Squares and lab 7-Data Analysis Case Study**

Functional Analysis

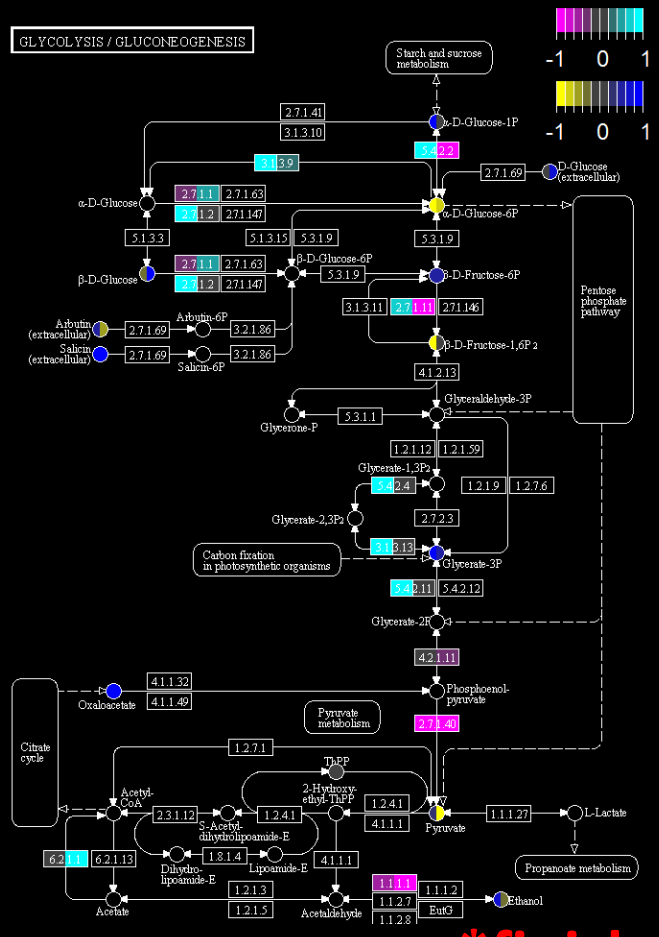
Identify changes or enrichment in biochemical domains



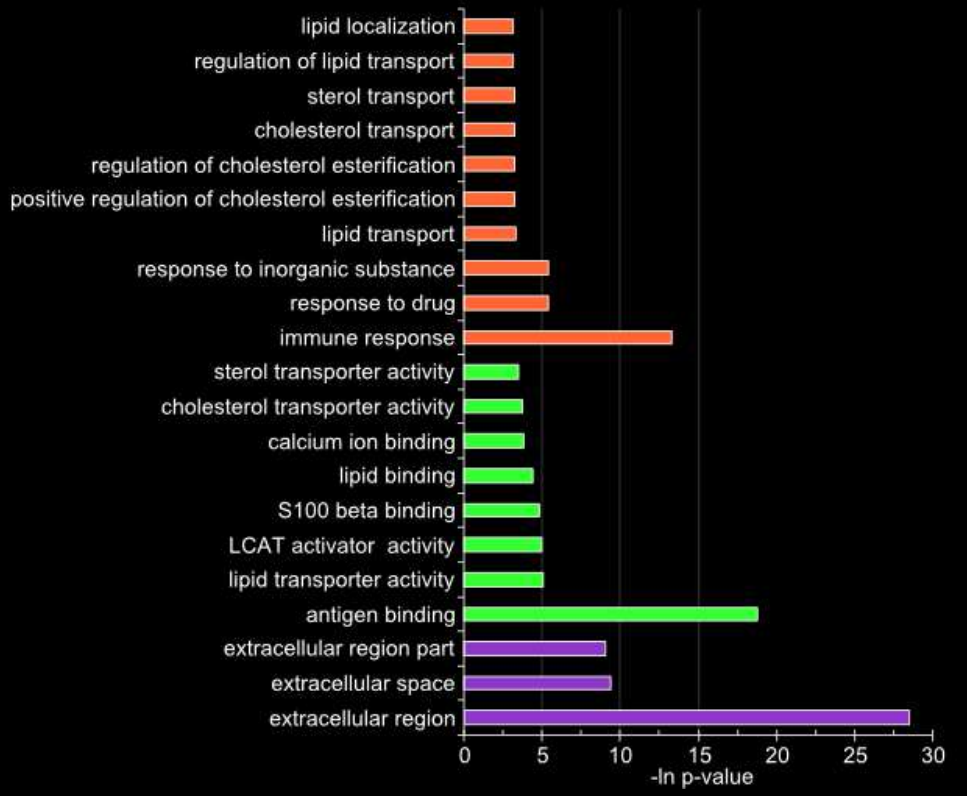
Nucl. Acids Res. (2008) 36 (suppl 2): W423-W426.doi: 10.1093/nar/gkn282

Functional Analysis: Enrichment

Biochemical Pathway

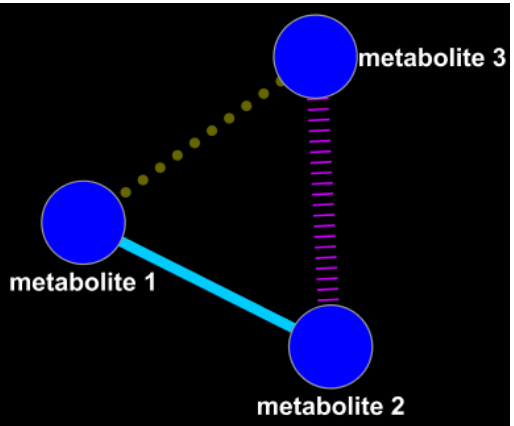


Biochemical Ontology



*finish lab 8-Metabolite Enrichment Analysis

Connections and Contexts



Biochemical (substrate/product)

- Database lookup
- Web query

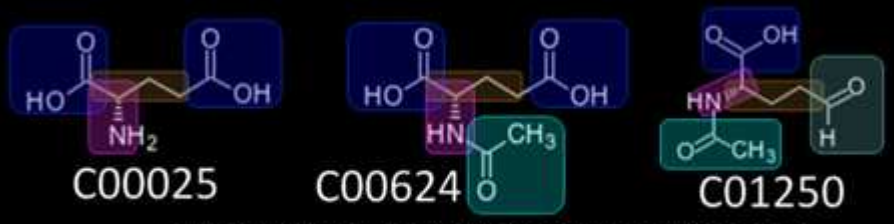
BioCyc Database Collection

REACTOME

KEGG Kyoto Encyclopedia of Genes and Genomes

Chemical (structural or spectral similarity)

- fingerprint generation

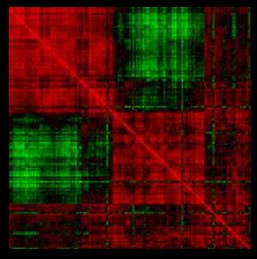


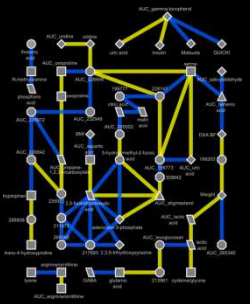
substructure matrix decomposition and Tanimoto chemical similarity calculations

BMC Bioinformatics 2012, **13**:99 doi:10.1186/1471-2105-13-99

Empirical (dependency)

- correlation, partial-correlation



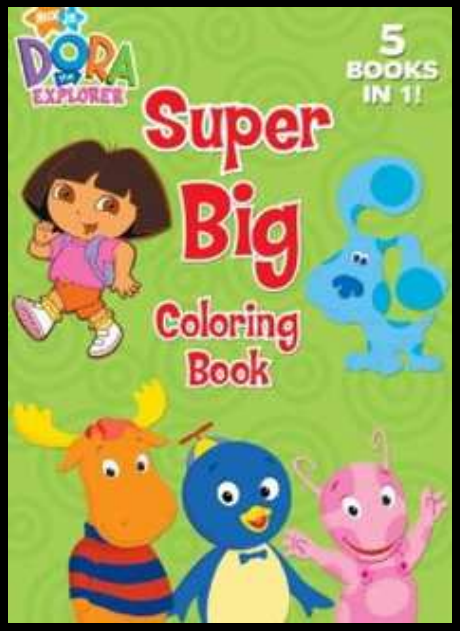


Network Mapping

1. Calculate Connections

2. Calculate Mappings

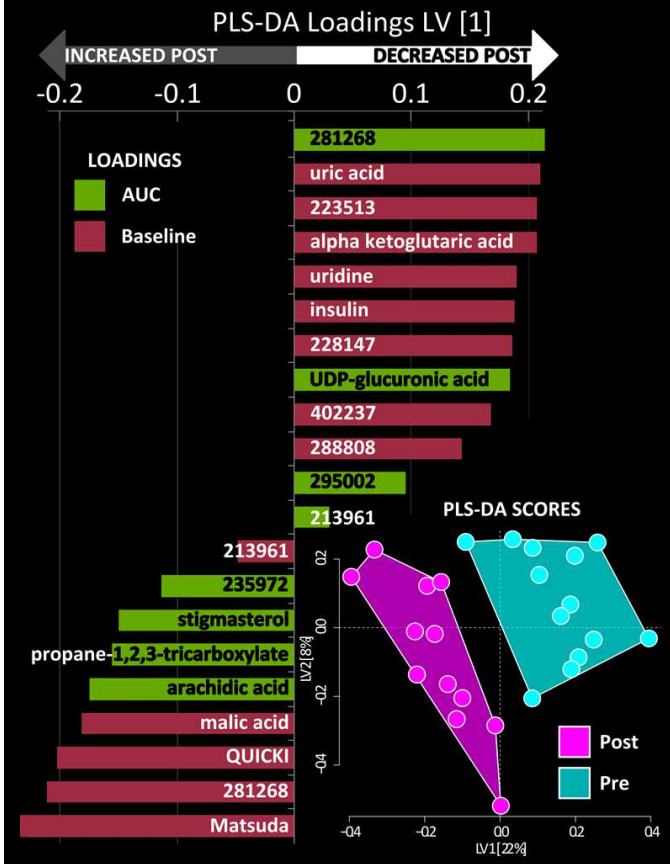
3. Create Mapped Network

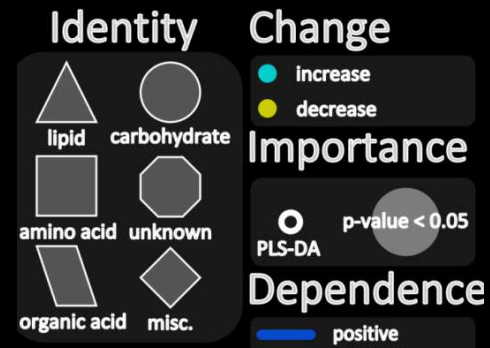


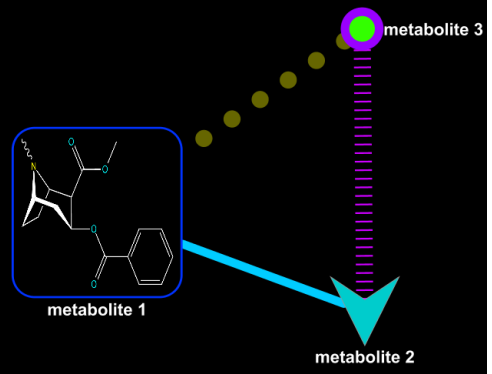
Grapov D., Fiehn O., Multivariate and network tools for analysis and visualization of metabolomic data, ASMS, June 08, 2013, Minneapolis, MN

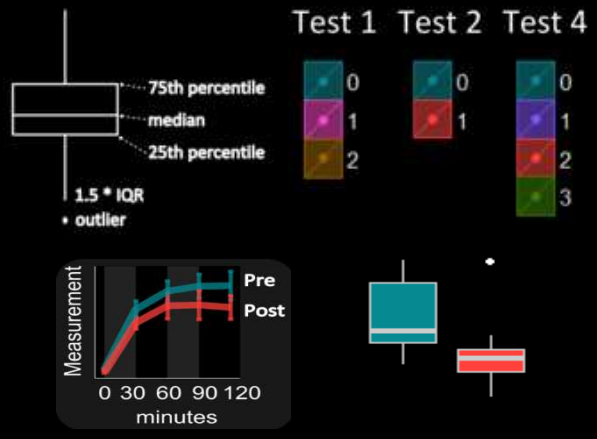
Mapping Analysis Results to Networks

Analysis results → Network Annotation → Mapped Network





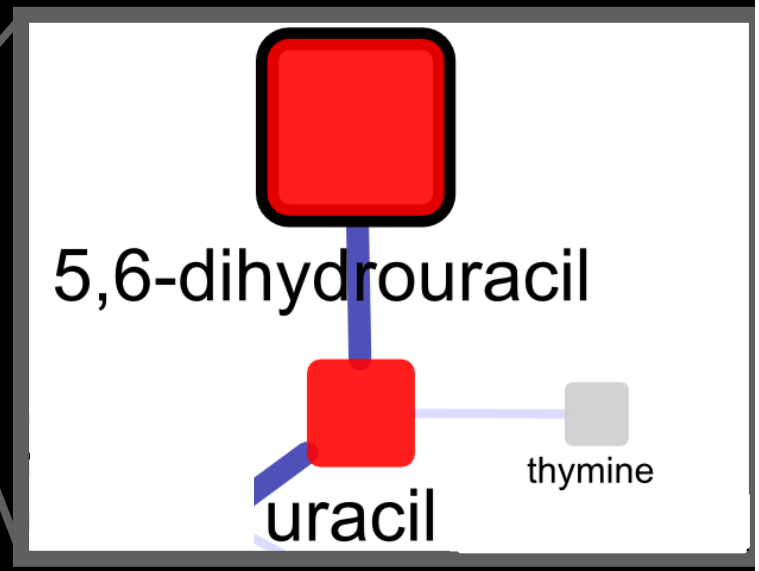


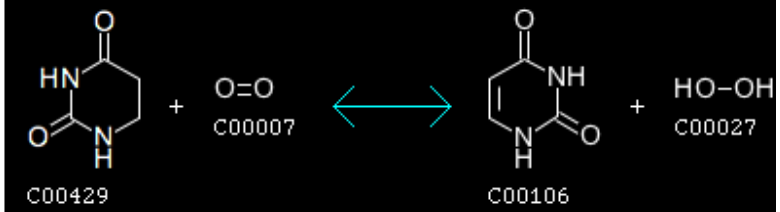


Cytoscape

***finish lab 9-Network Mapping I**

Biochemical Relationships





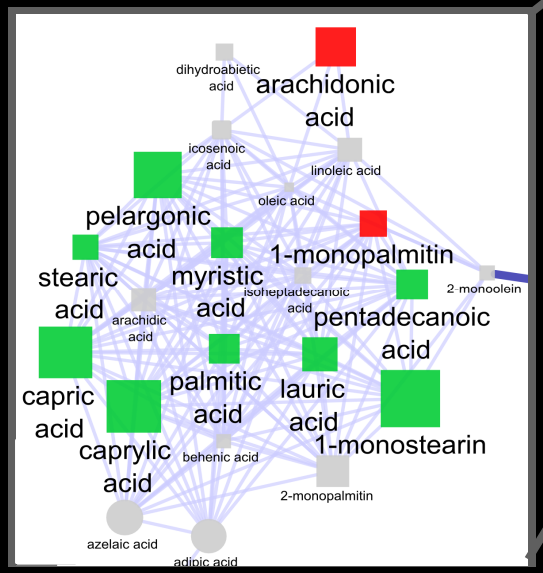
http://www.genome.jp/dbget-bin/www_bget?rn:R00975

arachidonic acid
pelargonic acid
stearic acid
myristic acid
1-monopalmitin
pentadecanoic acid
capric acid
palmitic acid
lauric acid
1-monostearin
caprylic acid
gamma-tocopherol
alpha-tocopherol
dodecanol
arabitol
ribitol
xylitol
arabinose
glycerol
glycerol-3-galactoside
1,5-anhydroglucitol
mannitol
inositol-4-monophosphate
3-phosphoglycerate
tagatose
glucose
5,6-dihydrouracil
UDP-glucuronic acid
uridine
uracil
cytidine-5-monophosphate
UDP GlcNAc
N-acetyl-D-mannosamine
4-hydroxybutyric acid
malic acid
erythronic acid
erythronic acid lactone
glyceric acid
citric acid
fucose
gluconic acid
2-hydroxyglutaric acid
quinic acid
inosine 5'-monophosphate
FAD
xanthine
uric acid
adenosine-5-phosphate
adenosine
adenine
5'-deoxy-5'-methylthioadenosine
glutamic acid
oxoproline
trans-4-hydroxyproline
proline
pipecolic acid
citrulline
lysine
isoleucine
ornithine
spermidine
biuret
allantoic acid
aminomalonic acid
alanine
kynurenine
cysteine
methionine sulfoxide
cysteine-glycine
2-phenylpropanol
benzoic acid

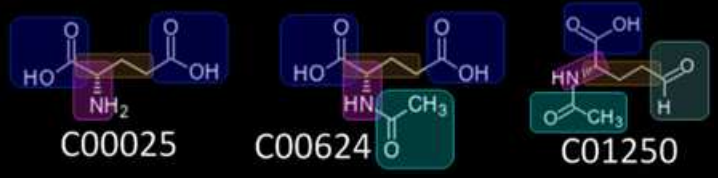
Metabolite Properties
- decrease
- increase
- carbohydrate
- organic acid
- lipid
- amino acid or amine
- purine or pyrimidine
- vitamin
- other

Relationship
- biochemical
- structural similarity

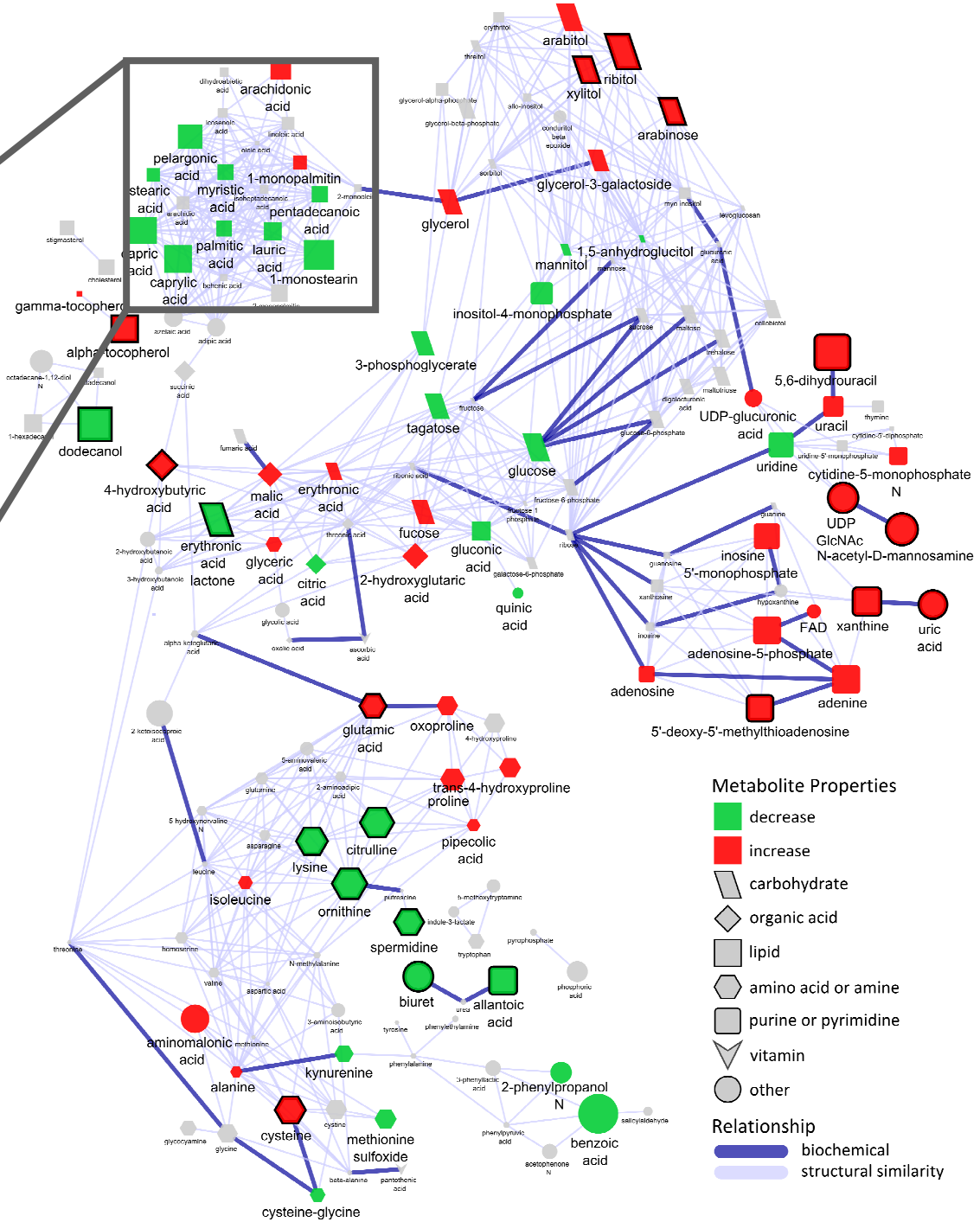
Structural Similarity



PubChem

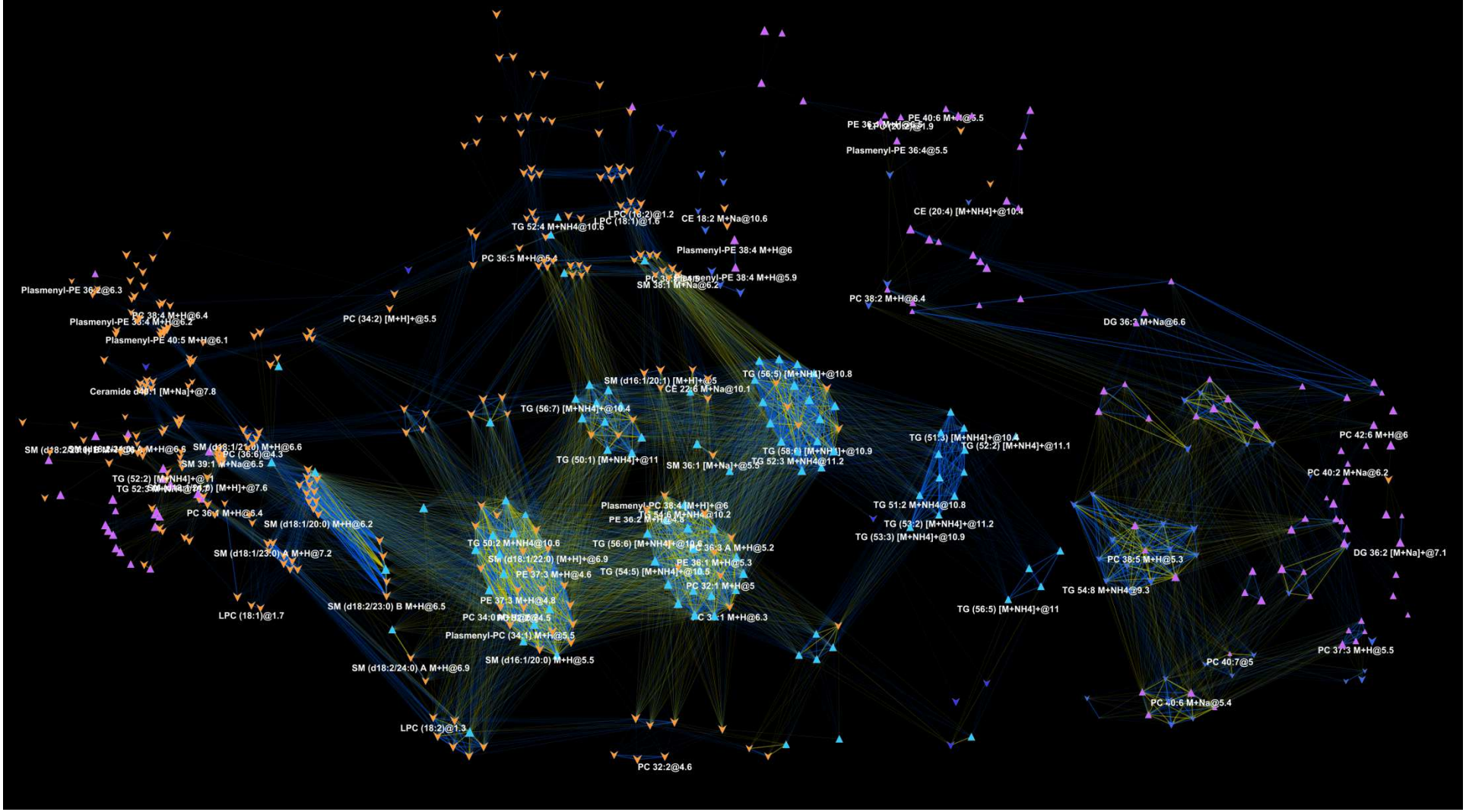


http://pubchem.ncbi.nlm.nih.gov//score_matrix/score_matrix.cgi



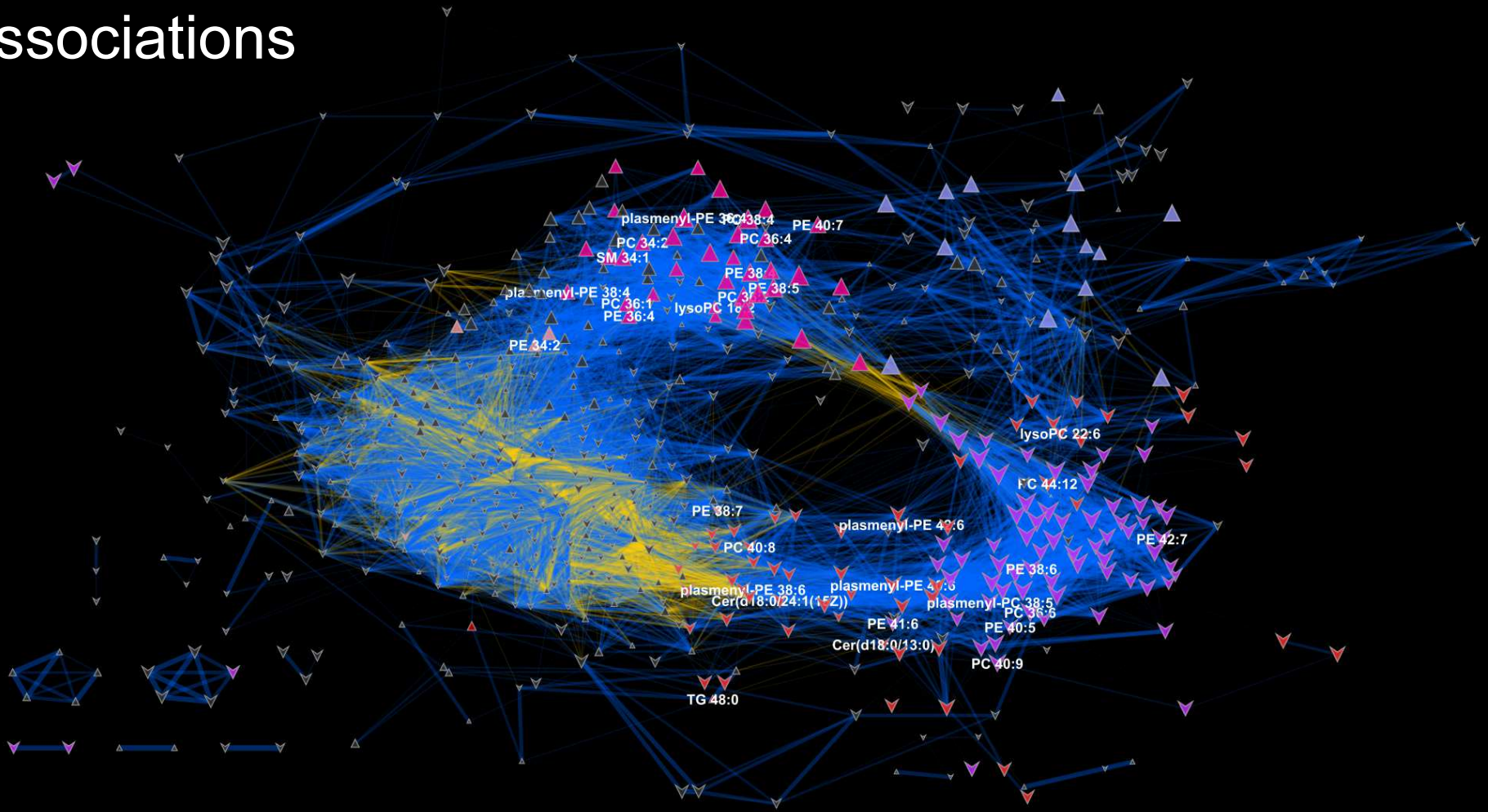
Correlation networks

- simple to calculate

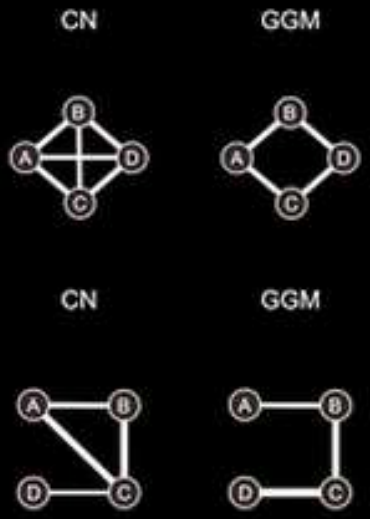


Correlation Networks

- can be difficult to interpret
- poorly discriminate between direct and indirect associations

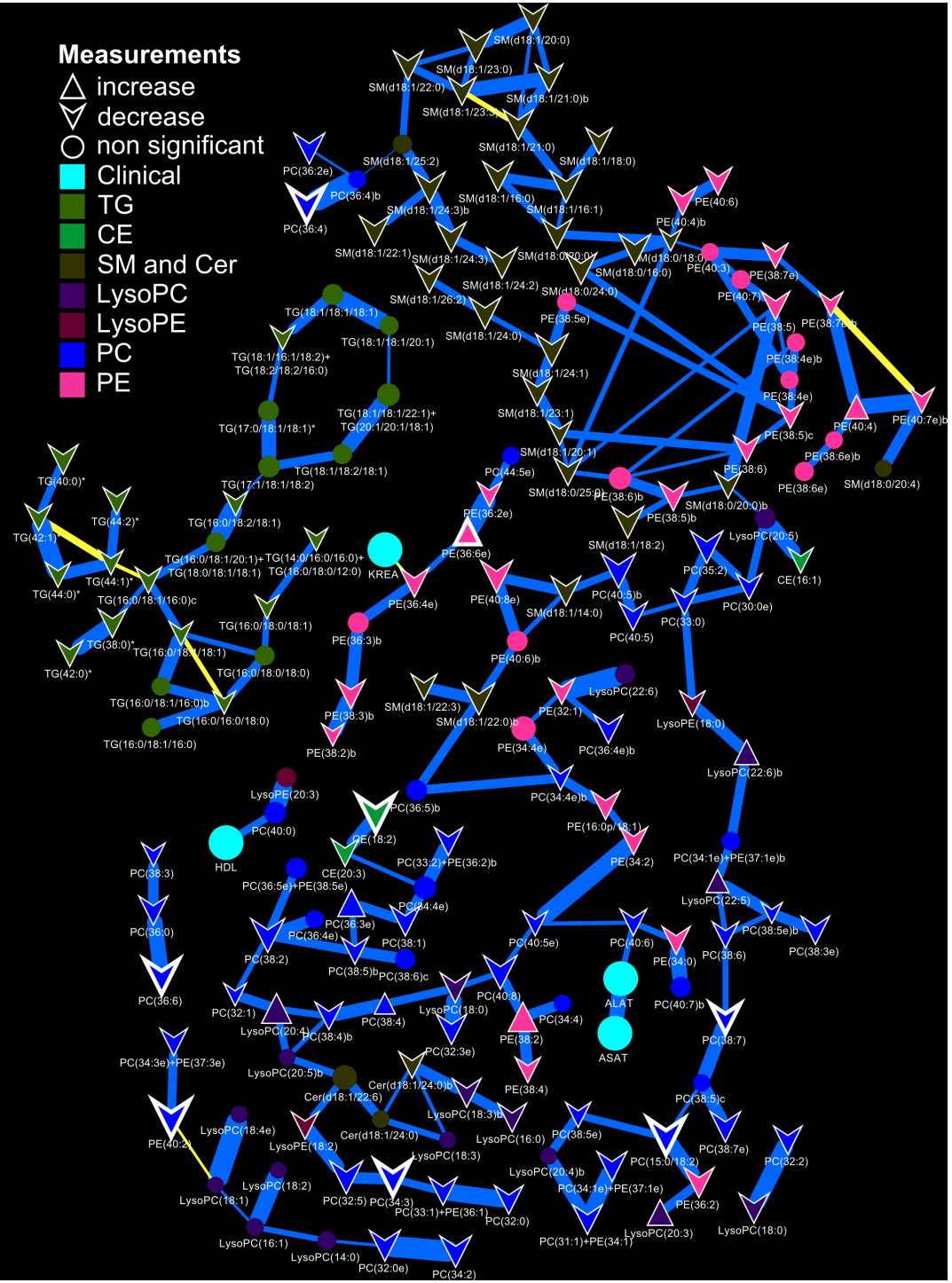


Partial correlations can help simplify networks and preference direct over indirect associations.

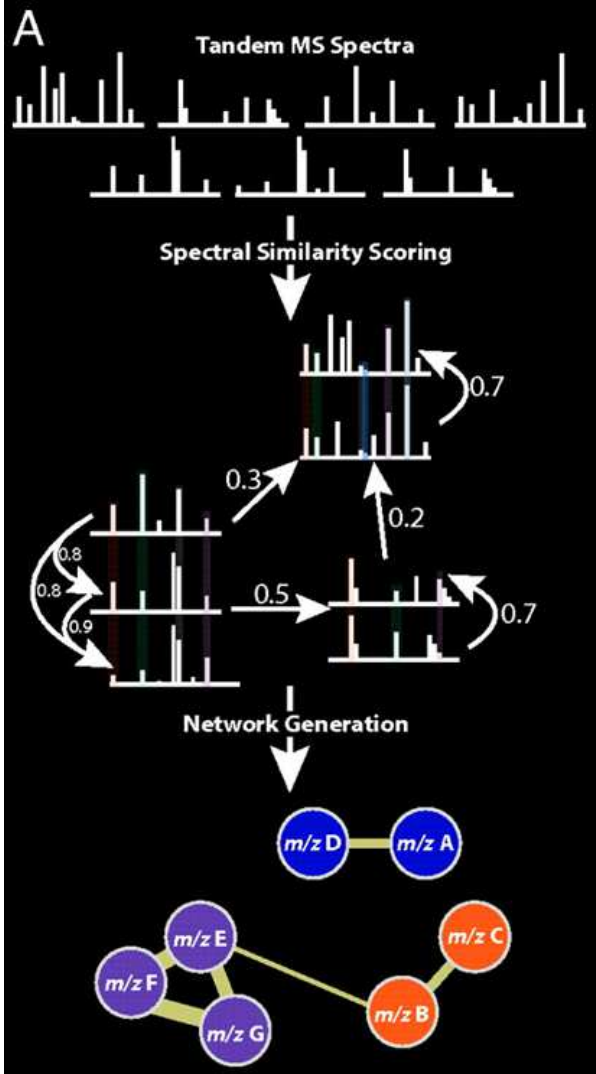


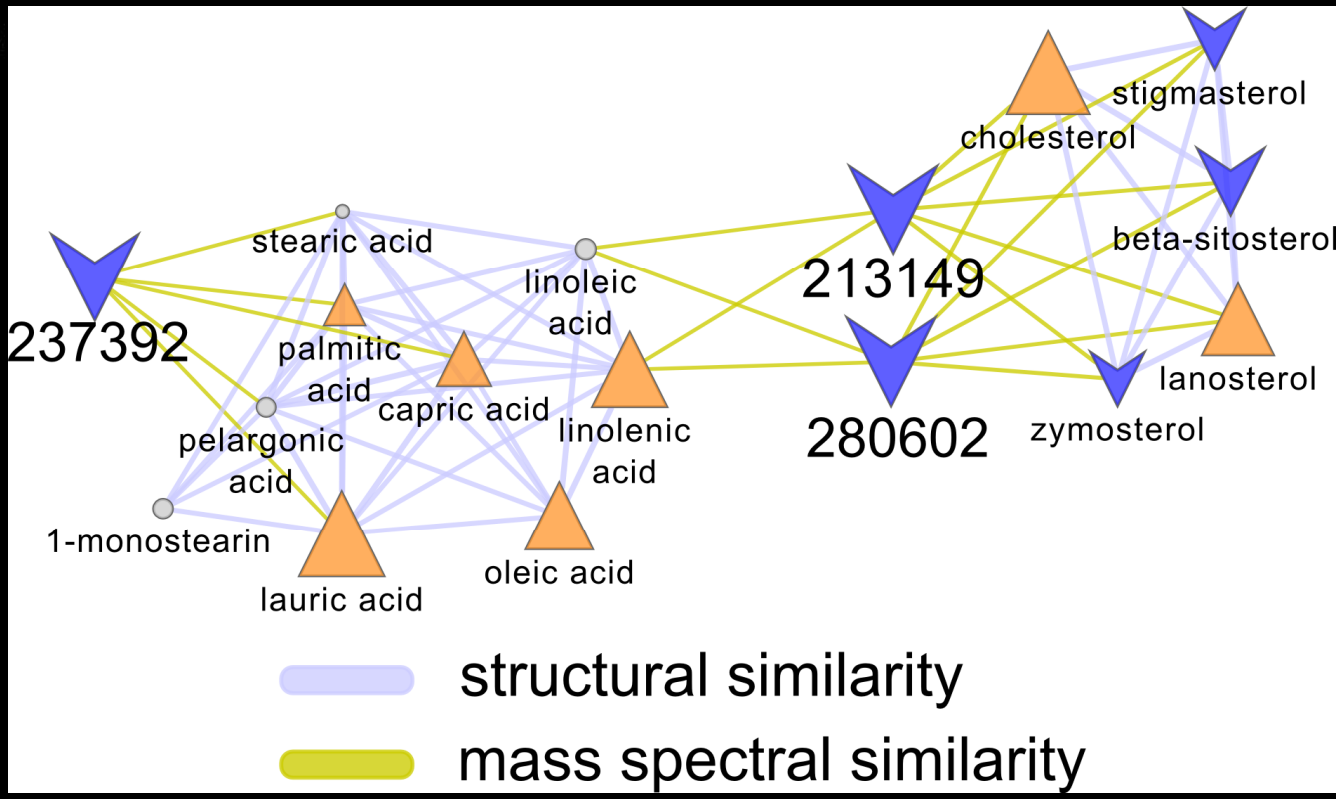
10.1007/978-1-4614-1689-0_17

Complex lipids partial correlation network in human plasma



Mass Spectral Connections





Watrous J et al. PNAS 2012;109:E1743-E1752

***finish lab 10-Network Mapping II**

Software and Resources

- **DeviumWeb- Dynamic multivariate data analysis and visualization platform**
  url: https://github.com/dgrapov/DeviumWeb

- **imDEV- Microsoft Excel add-in for multivariate analysis**
  url: http://sourceforge.net/projects/imdev/

- **MetaMapR- Network analysis tools for metabolomics**
  url: https://github.com/dgrapov/MetaMapR

- **TeachingDemos- Tutorials and demonstrations**
  - **url: http://sourceforge.net/projects/teachingdemos/?source=directory**
  - **url: https://github.com/dgrapov/TeachingDemos**

- **Data analysis case studies and Examples**
  url: http://imdevsoftware.wordpress.com/

NIH: West Coast Metabolomics Center

Questions?

dgrapov@gmail.com