The Fifth International Conference of Metabolomic Society

Metabolomic Data Processing & Statistical Analysis

Jianguo (Jeff) Xia
Dr. David Wishart Lab
University of Alberta, Canada

Outline

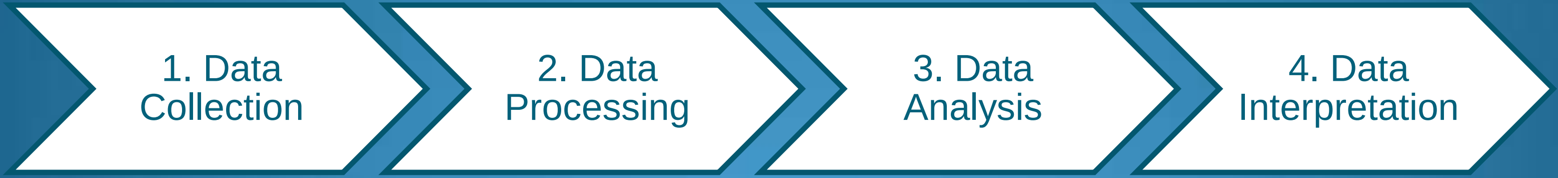
I. Overview of procedures for metabolomic studies

II. Introduction to different data processing & statistical methods

III. MetaboAnalyst – a web service for metabolomic data processing, analysis and annotation

IV. Conclusions & future directions

A data-centric overview of metabolomic studies



1. Data Collection

2. Data Processing

3. Data Analysis

4. Data Interpretation

Data collection

❖ Biological Samples → Spectra

Separation Techniques

- Gas Chromatography (GC)
- Liquid Chromatography (LC)
- Capillary Electrophoresis (CE)

Detection Techniques

- Nuclear Magnetic Resonance Spectroscopy (NMR)
- Mass Spectrometry (MS)

Hyphenated Techniques

- Gas Chromatography - Mass Spectrometry (GC-MS)
- Liquid Chromatography -Mass Spectrometry (LC-MS)
- Liquid Chromatography - Nuclear Magnetic Resonance (LC-NMR)

Data processing

❖ Raw Spectra → Data Matrix

Quantitative

- Compound concentration data;
- Involving compound identification & quantification;
- Currently labor intensive with a lot of manual efforts

Chemometric

- Spectral bins (NMR, Direct injection–MS)
- Peak lists (LC/GC – MS)
- Largely automated process

Data analysis

❖ Extract important features/patterns

Exploratory Analysis

- Data overview
- Outlier detection
- Grouping patterns

Biomarker discovery

- To identify metabolites that are significantly different between groups

Classification

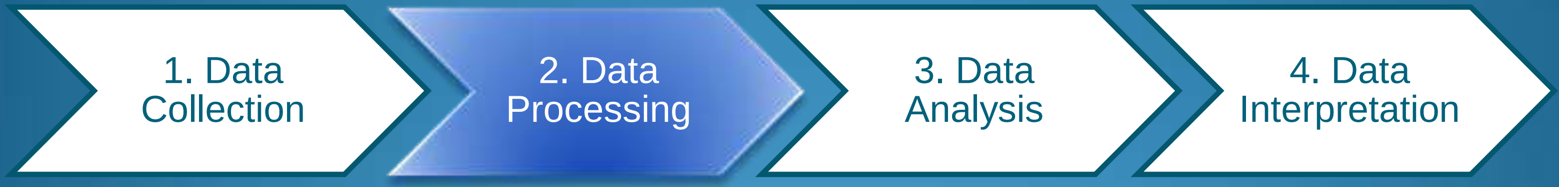
- To build a model for the prediction of unlabeled new samples

Data interpretation

❖ Features/patterns → biological knowledge

- Mainly a manual process
- Require domain expert knowledge
- Tools are coming:
  - Comprehensive metabolite databases
  - Network visualization
  - Pathway analysis

Data processing & normalization



Data processing (I)

- Purposes:
  - To convert different metabolomic data into data matrices suitable for varieties of statistical analysis
  - Quality control
    - To check for inconsistencies
    - To deal with missing values
    - To remove noises

Data processing (II)

Compound concentrations

- Nothing to do

A data matrix with **rows represent samples** and **columns represents features** (concentrations/intensities/ areas)

GC/LC-MS spectra

- Peak picking
- Peak alignment

Data normalization

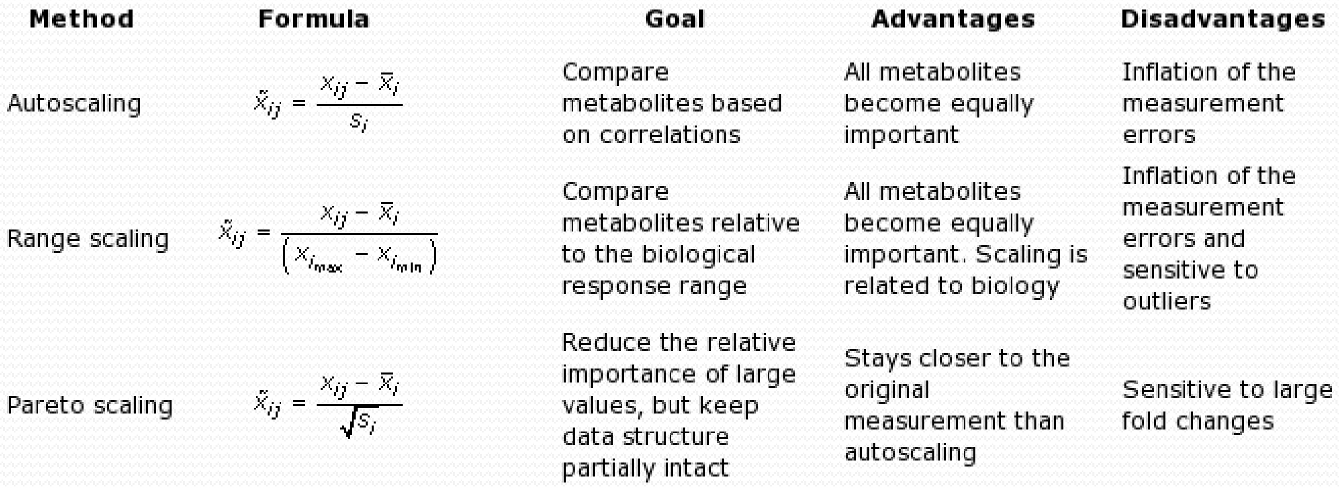
- Purposes:
  - To remove systematic variation between experimental conditions unrelated to the biological differences (i.e. dilutions, mass)
    - Sample normalization (row-wise)
  - To bring variances of all features close to equal
    - Feature normalization (column-wise)

Sample normalization

- By sum or total peak area
- By a reference compound (i.e. creatinine, internal standard)
- By a reference sample
  - a.k.a "probabilistic quotient normalization" *(Dieterle F, et al. Anal. Chem. 2006)*
- By dry mass, volume, *etc*

Feature normalization

- Log transformation
- Scaling



Method	Formula	Goal	Advantages	Disadvantages
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	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$	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

-- van den Berg RA, *et al*. BMC Genomics (2006) 7:142

Statistical Analysis

1. Data Collection

2. Data Processing

3. Data Analysis

4. Data Interpretation

Data Analysis

Univariate

- Fold change analysis,
- T-tests
- ✓ Volcano plots

Chemometrics

- Principal component analysis (PCA)
- ✓ Partial least squares - discriminant analysis (PLS-DA)

High-dimensional feature selection

- ✓ Significance analysis of microarrays (and metabolites) (SAM)
- Empirical Bayesian analysis of microarrays (and metabolites) (EBAM)

Clustering

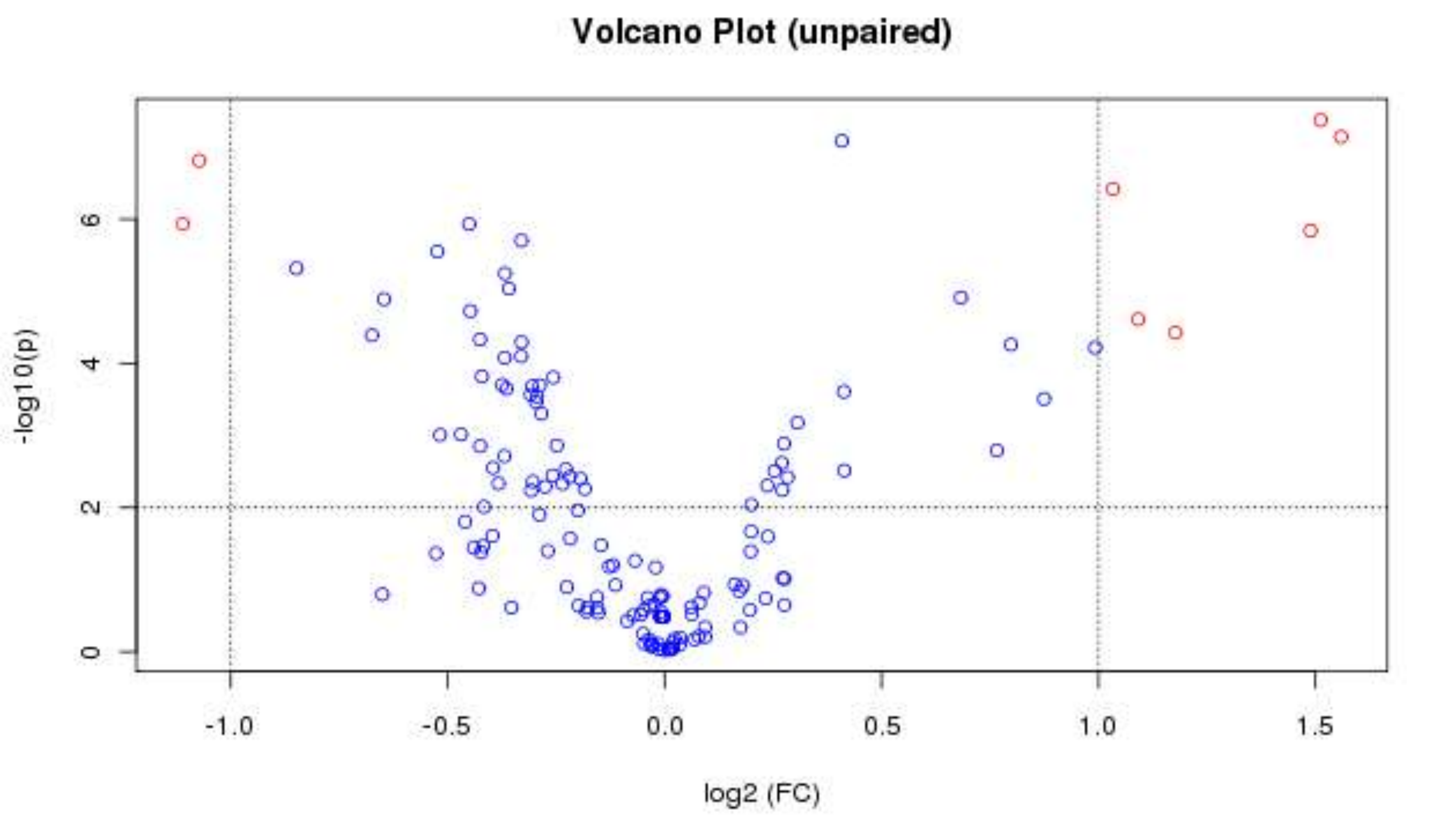
- Dendrogram & Heatmap
- K-means, Self Organizing Map (SOM)

Classification

- Random Forests
- Support Vector Machine (SVM)

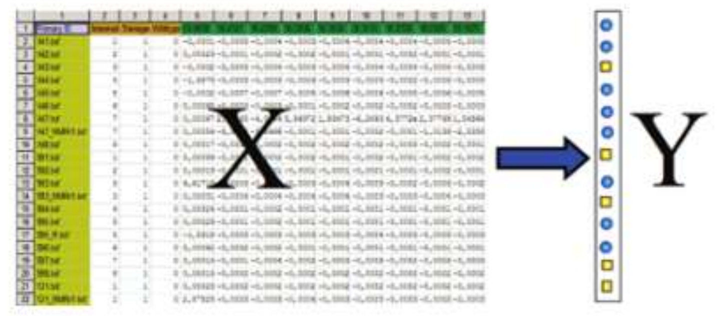
Volcano-plot

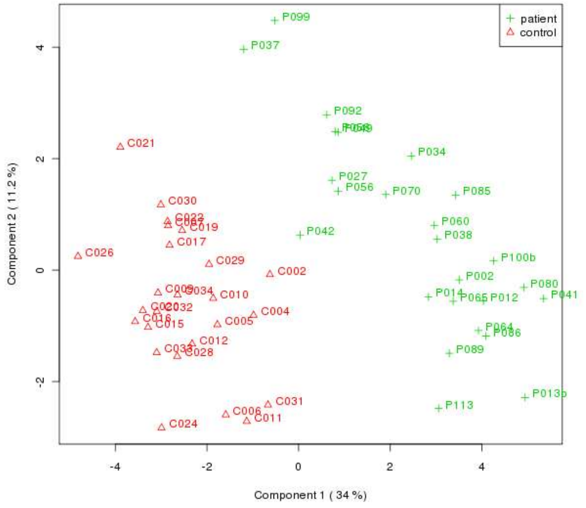
-
-
-



PLS-DA

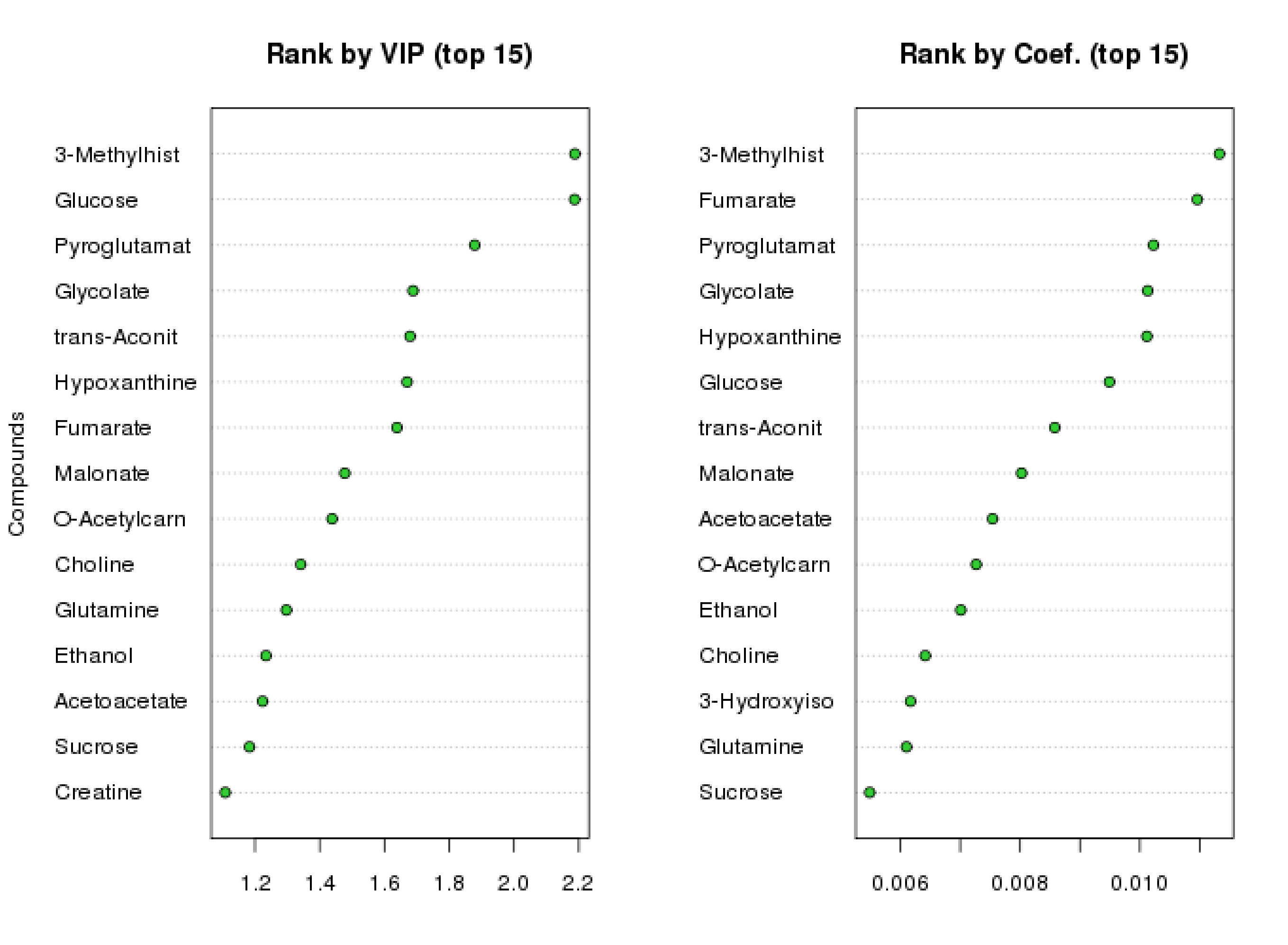
- *De facto* standard for chemometric analysis
- A supervised method that uses multiple linear regression technique to find the direction of maximum covariance between a data set (*X*) and the class membership (*Y*)
- Extracted features are in the form of latent variables (LV)





PLS

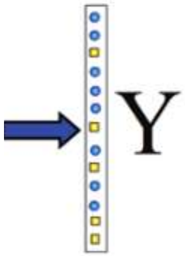
- Var
  - A ights
    a h
    c
  - E
    c
  - Tl f
    sc



Over fitting problem

- PLS-DA tend to over fit data
  - It will try to separate classes even there is no real difference between them!
    - Westerhuis, C.A., *et al*. (2007) Assessment of PLSDA cross validation. *Metabolomics*, 4, 81-89.
- Require more rigorous validation
  - For example, to use permutations to test the significance of class separations

Perr

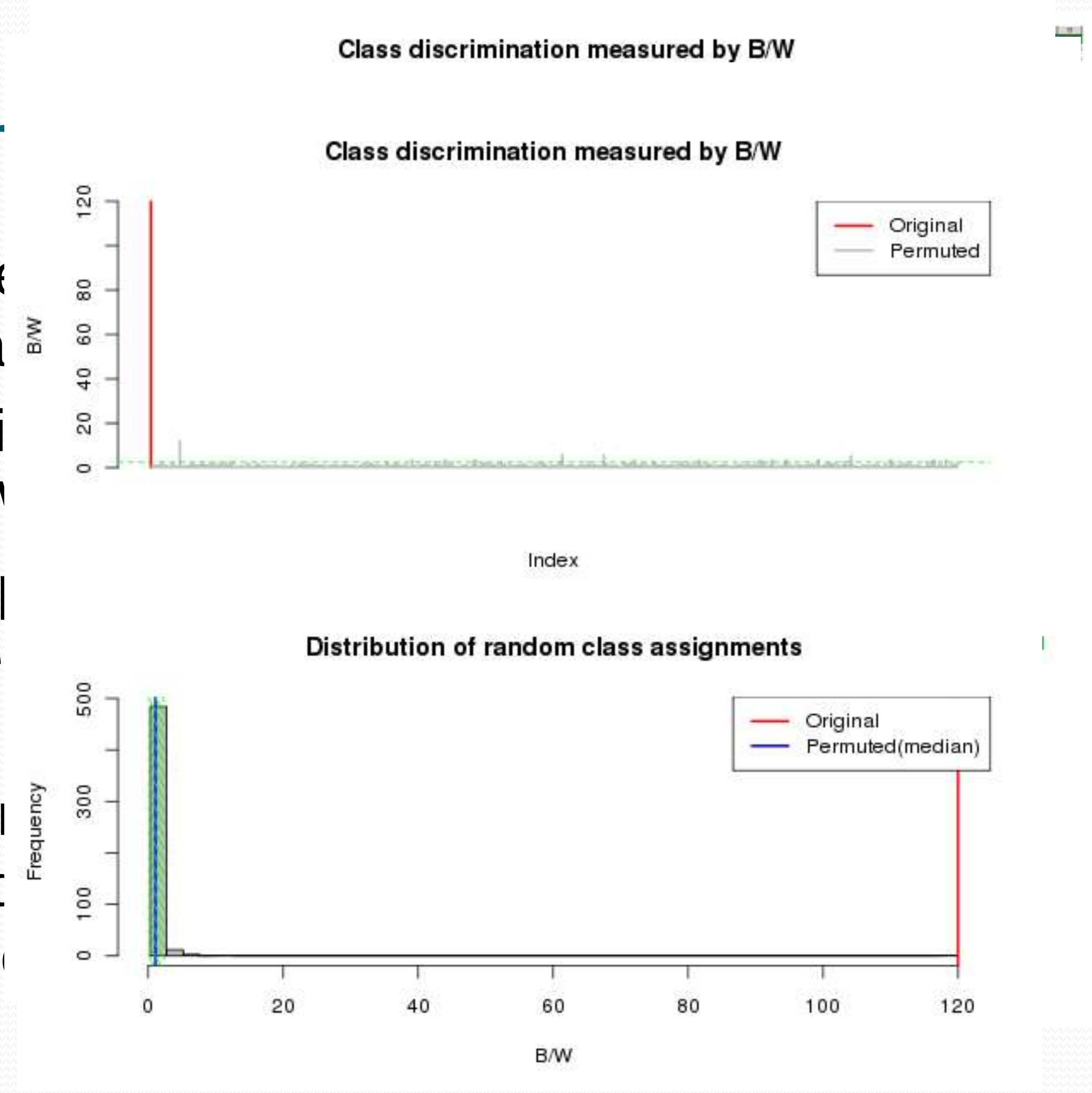


1) Use
rea

2) Bui ... nce
(B/W

3) Rep ... on of
the ... llows
a n

4) Co ... label
and
lab

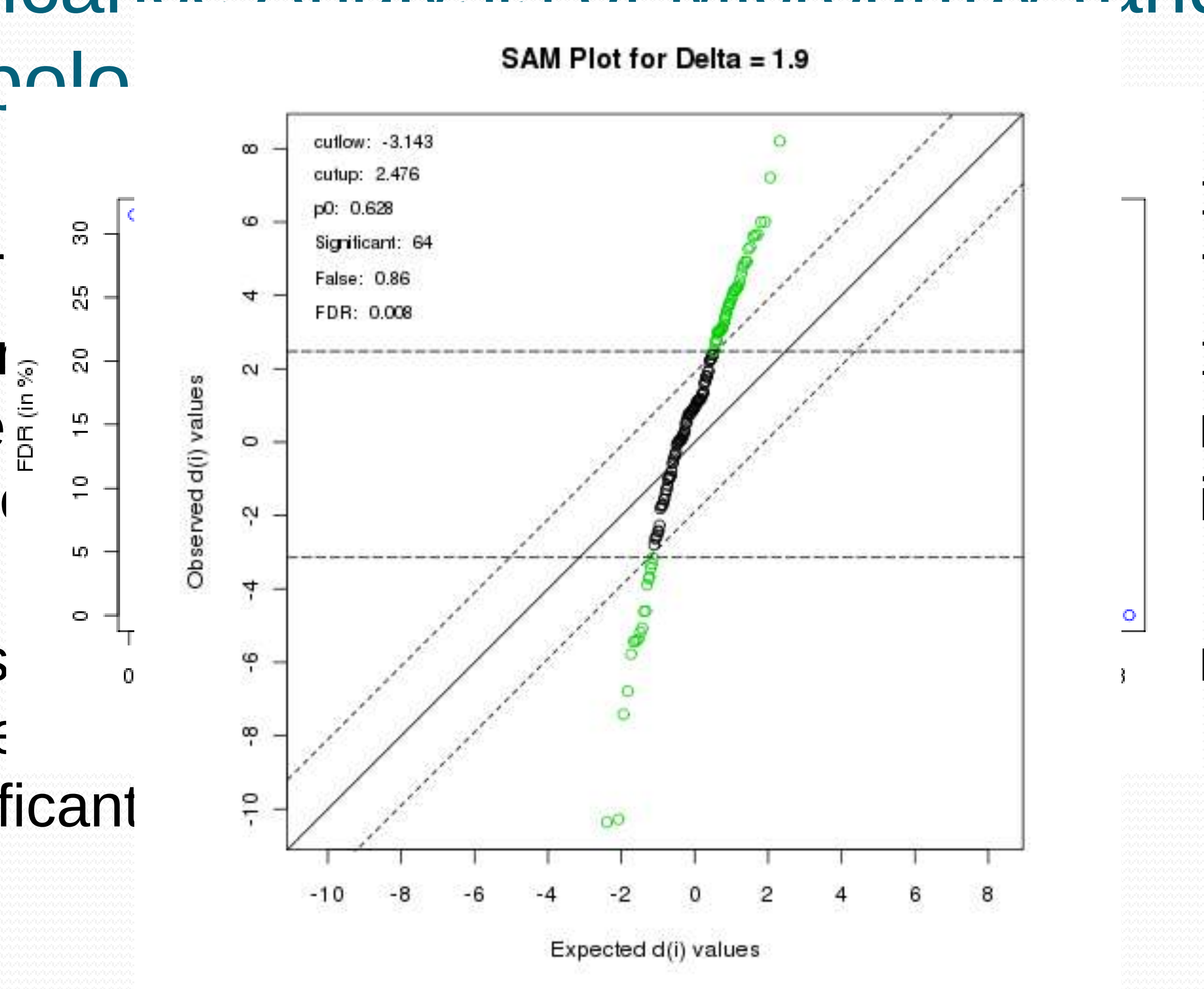


Multi-testing problem

- P-value appropriate to a single test situation is inappropriate to presenting evidence for a set of changed features.
  - Adjusting p-values
    - Bonferroni correction
    - Holm step-down procedure
  - Using false discovery rate (FDR)
    - A percentage indicating the expected false positives among all features predicted to be significant
    - More powerful, suitable for multiple testing

Significance Analysis of Microarray (and Metabolo

- A we cation of differ iments
- Use r or each gene nship betw iable (Y).
- Uses itations of the ene is significant



Clustering

- Unsupervised learning
- Good for data overview
- Use some sort of distance measures to group samples
  - PCA
  - Heatmap & dendrogram
  - SOM & K-means

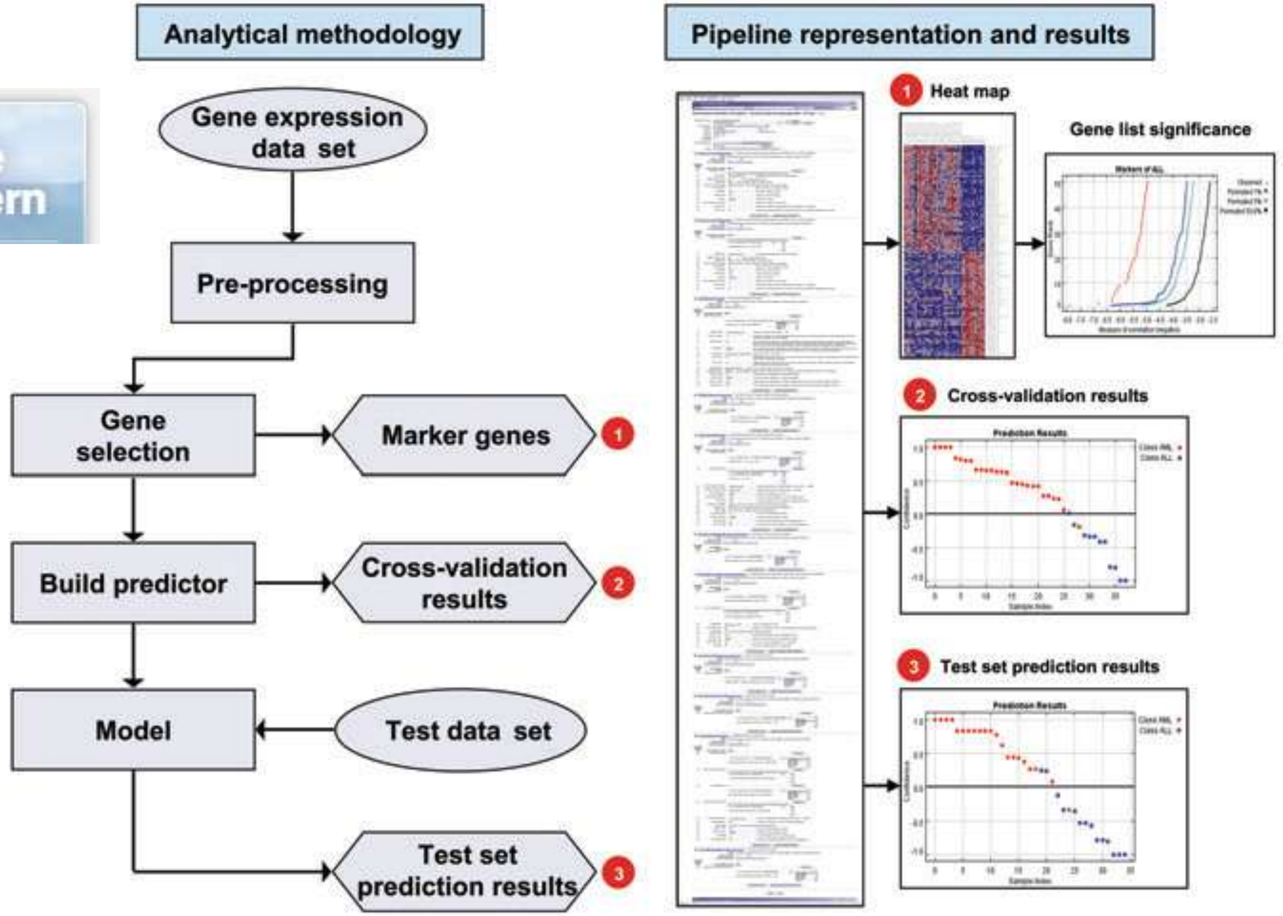
Classification

- Supervised learning
- Many traditional multivariate statistical methods are not suitable for high-dimensional data, particularly small sample size with large feature numbers
- New or improved methods, developed in the past decades for microarray data analysis
  - Support vector machine (SVM)
  - Random Forests

To develop a pipeline service for metabolomic studies

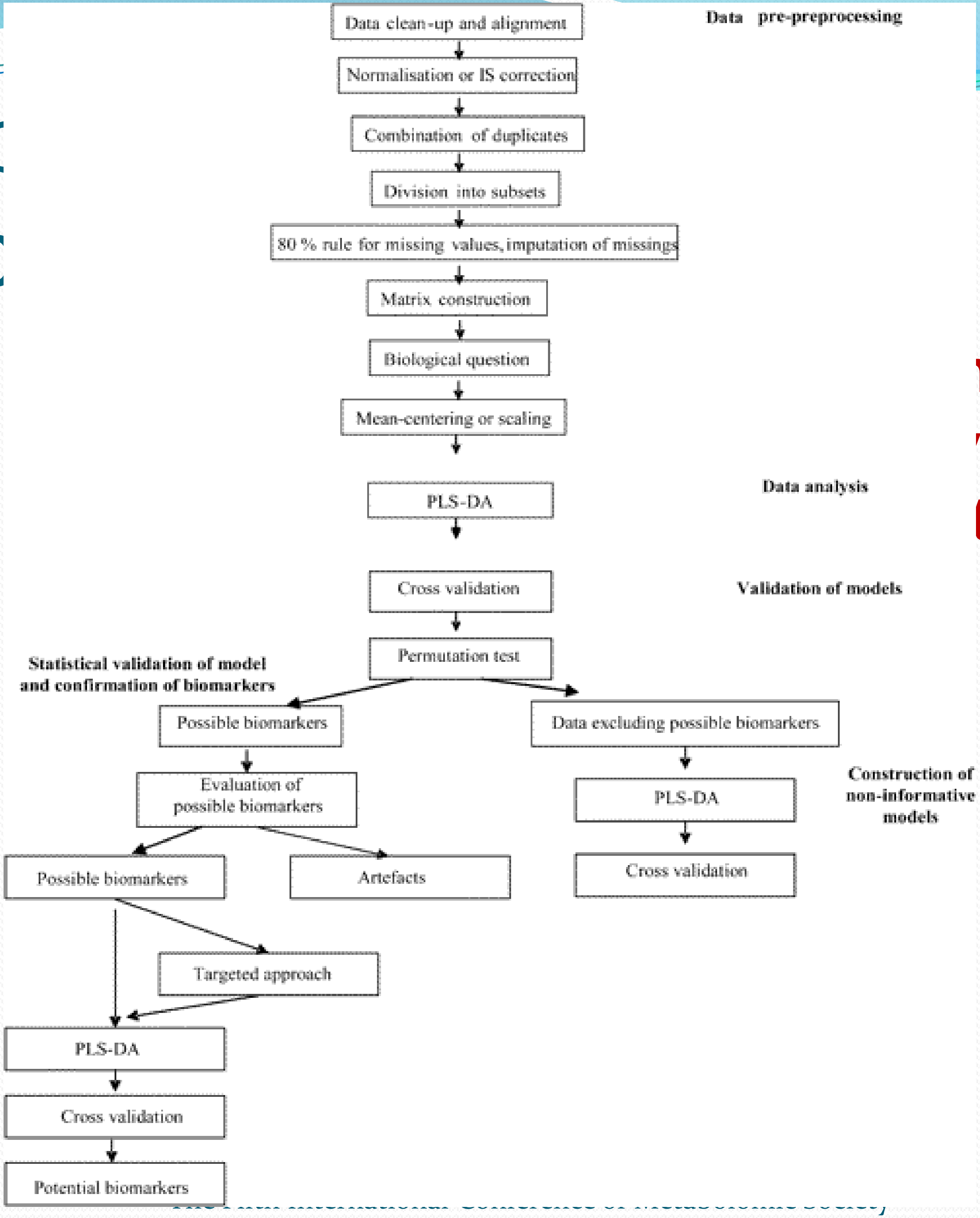
Microarray data analysis pipeline

Gene Pattern



A prop metab

- n ... for ... tion.

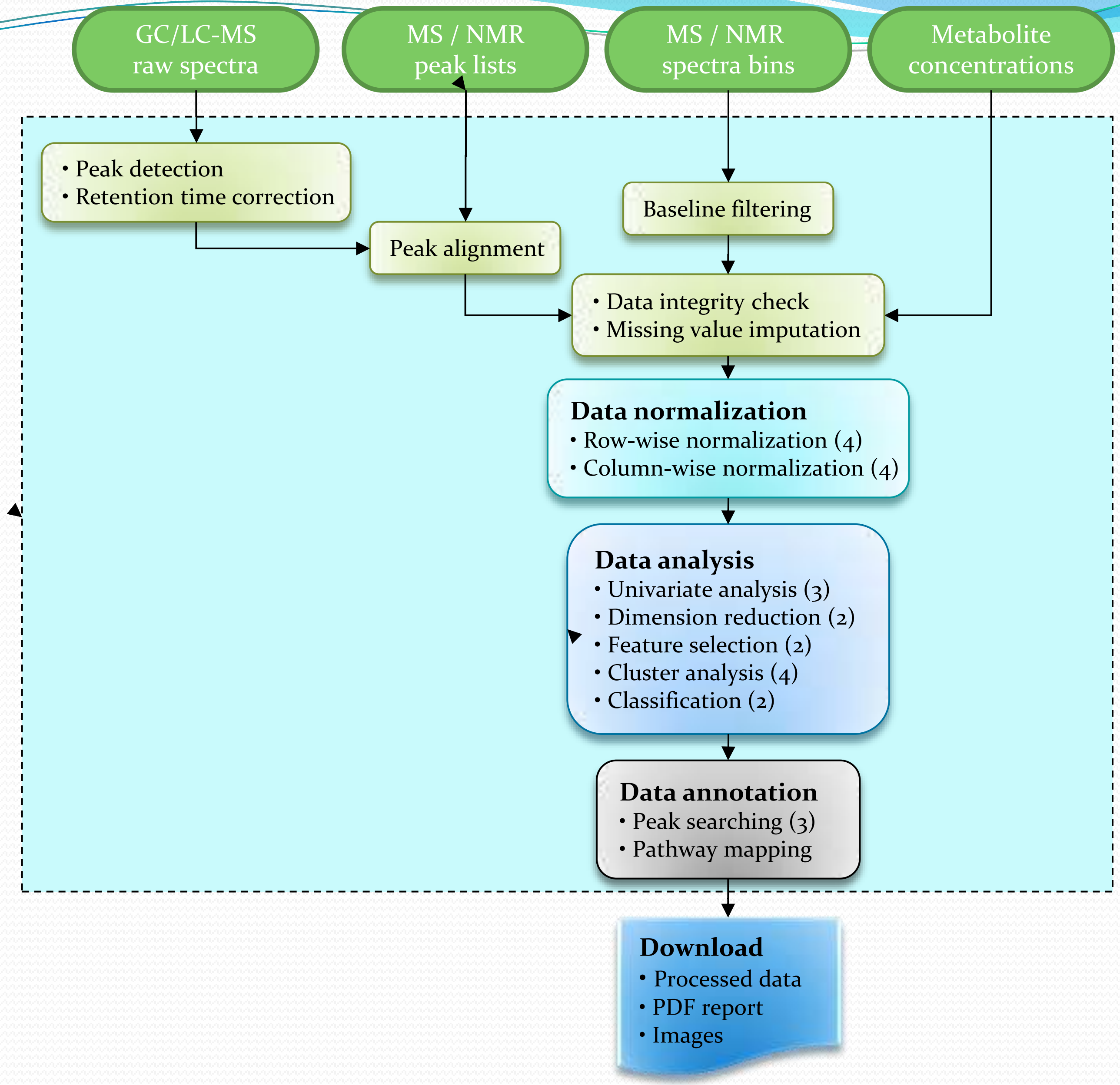

Data pre-preprocessing
Data clean-up and alignment
Normalisation or IS correction
Combination of duplicates
Division into subsets
80 % rule for missing values, imputation of missings
Matrix construction
Biological question
Mean-centering or scaling
Data analysis
PLS-DA
Validation of models
Cross validation
Permutation test
Statistical validation of model and confirmation of biomarkers
Possible biomarkers
Data excluding possible biomarkers
Evaluation of possible biomarkers
PLS-DA
Construction of non-informative models
Possible biomarkers
Artefacts
Cross validation
Targeted approach
PLS-DA
Cross validation
Potential biomarkers

MetaboAnalyst

-- A web service for high-throughput metabolomic data processing, analysis and annotation

-- Implementation of all the methods mentioned in the form of user-friendly web interfaces

-- www.metaboanalyst.ca



GC/LC-MS raw spectra

MS / NMR peak lists

MS / NMR spectra bins

Metabolite concentrations

- Peak detection
- Retention time correction

Peak alignment

Baseline filtering

- Data integrity check
- Missing value imputation

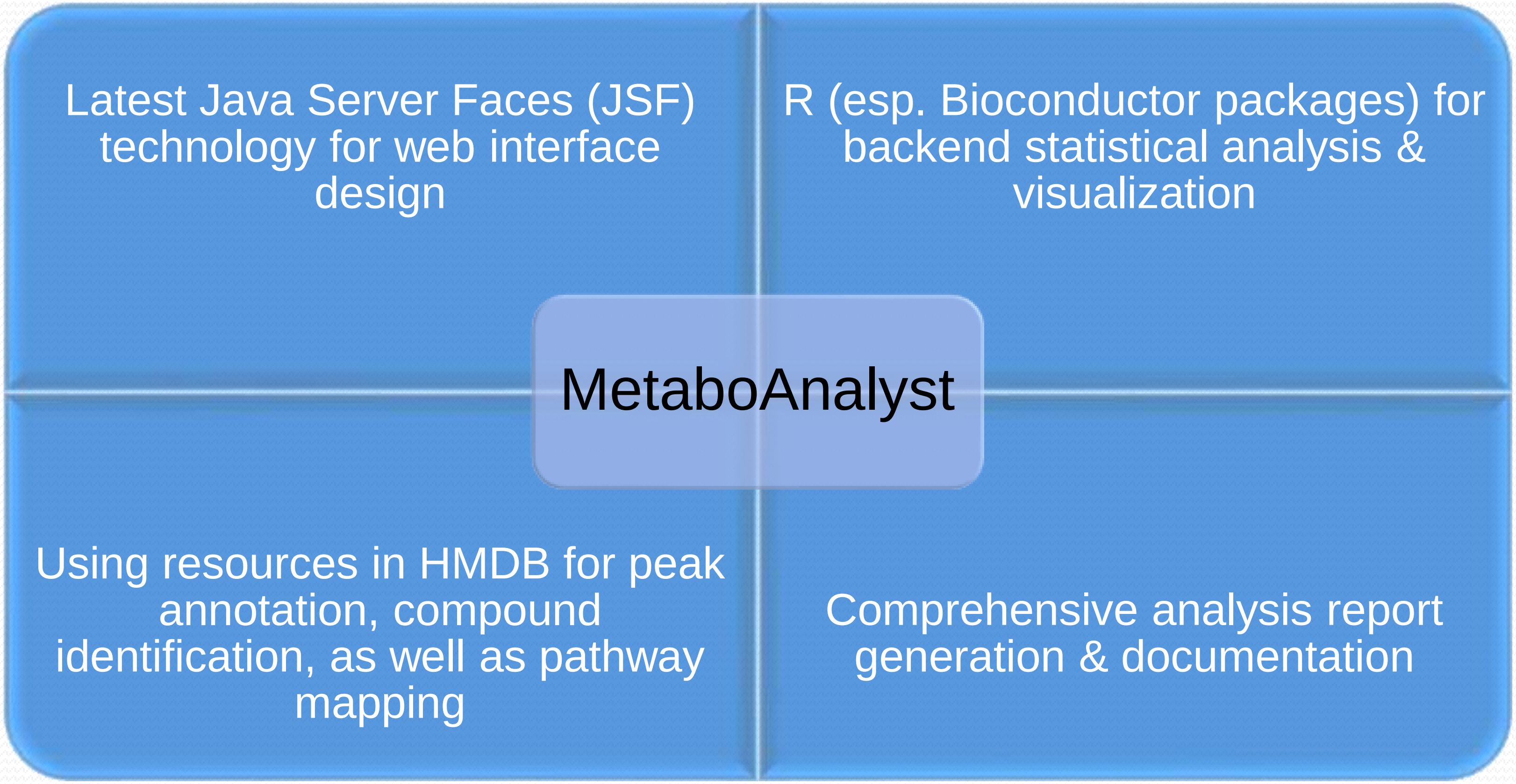
Data normalization
- Row-wise normalization (4)
- Column-wise normalization (4)

Data analysis
- Univariate analysis (3)
- Dimension reduction (2)
- Feature selection (2)
- Cluster analysis (4)
- Classification (2)

Data annotation
- Peak searching (3)
- Pathway mapping

Download
- Processed data
- PDF report
- Images

Implementation features



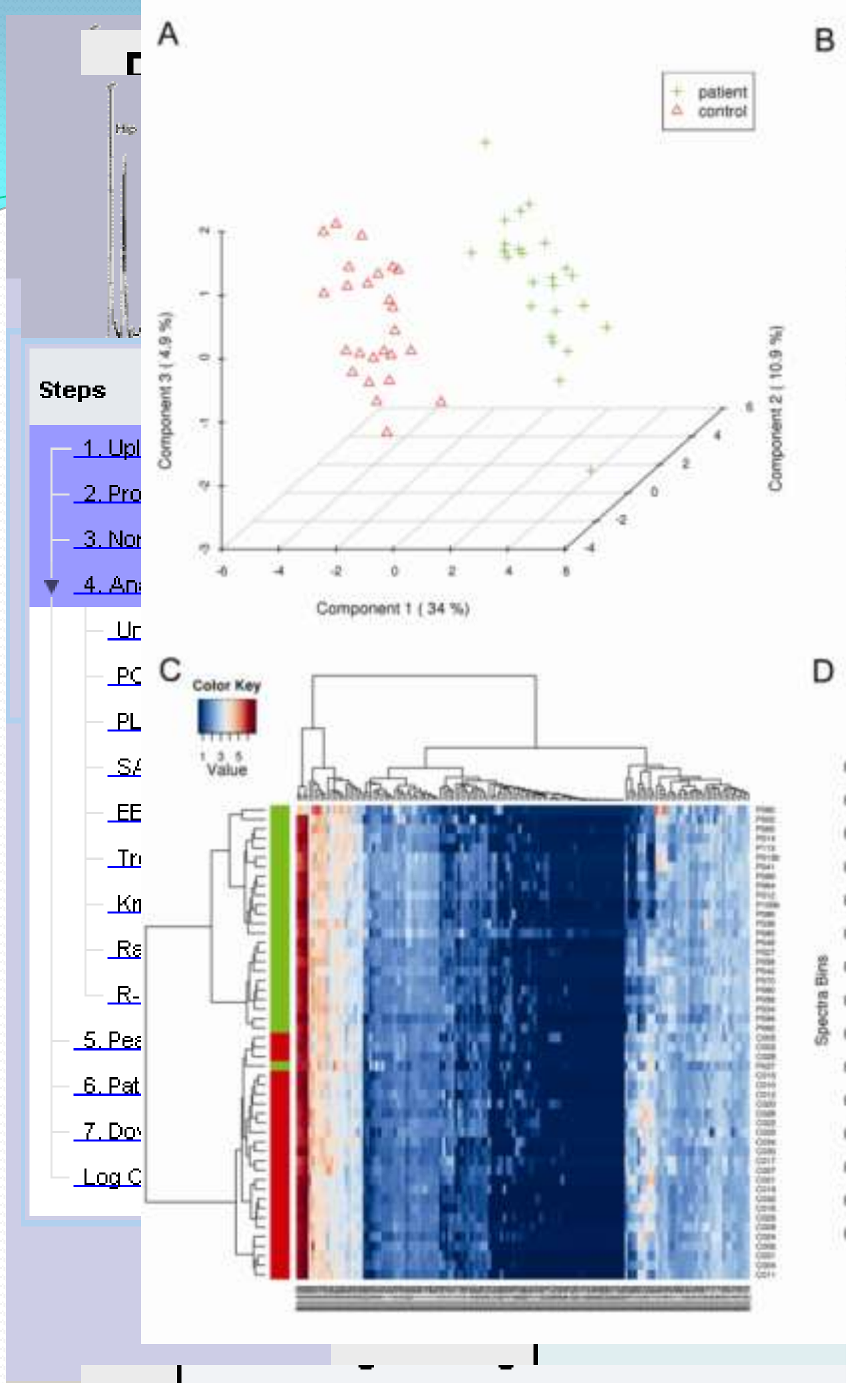
Latest Java Server Faces (JSF) technology for web interface design

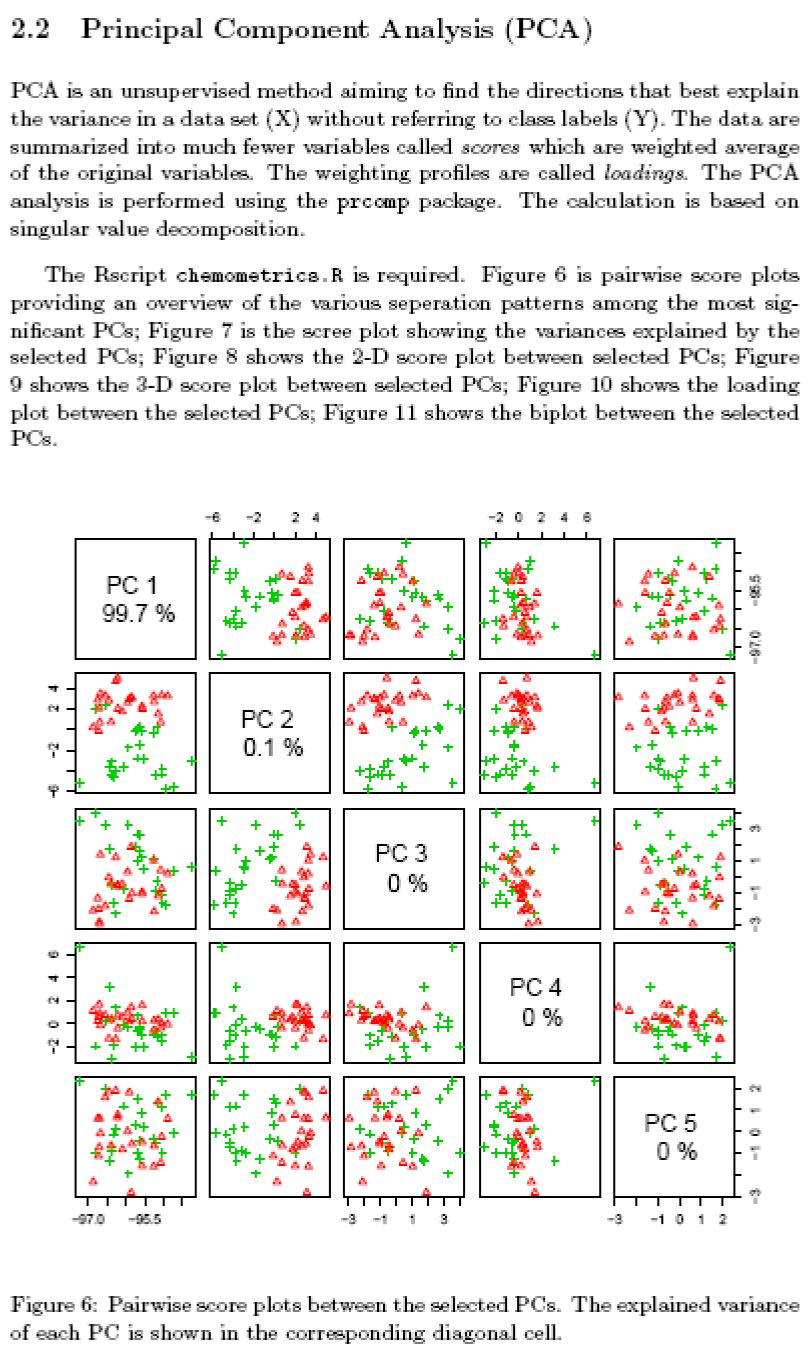
R (esp. Bioconductor packages) for backend statistical analysis & visualization

MetaboAnalyst

Using resources in HMDB for peak annotation, compound identification, as well as pathway mapping

Comprehensive analysis report generation & documentation





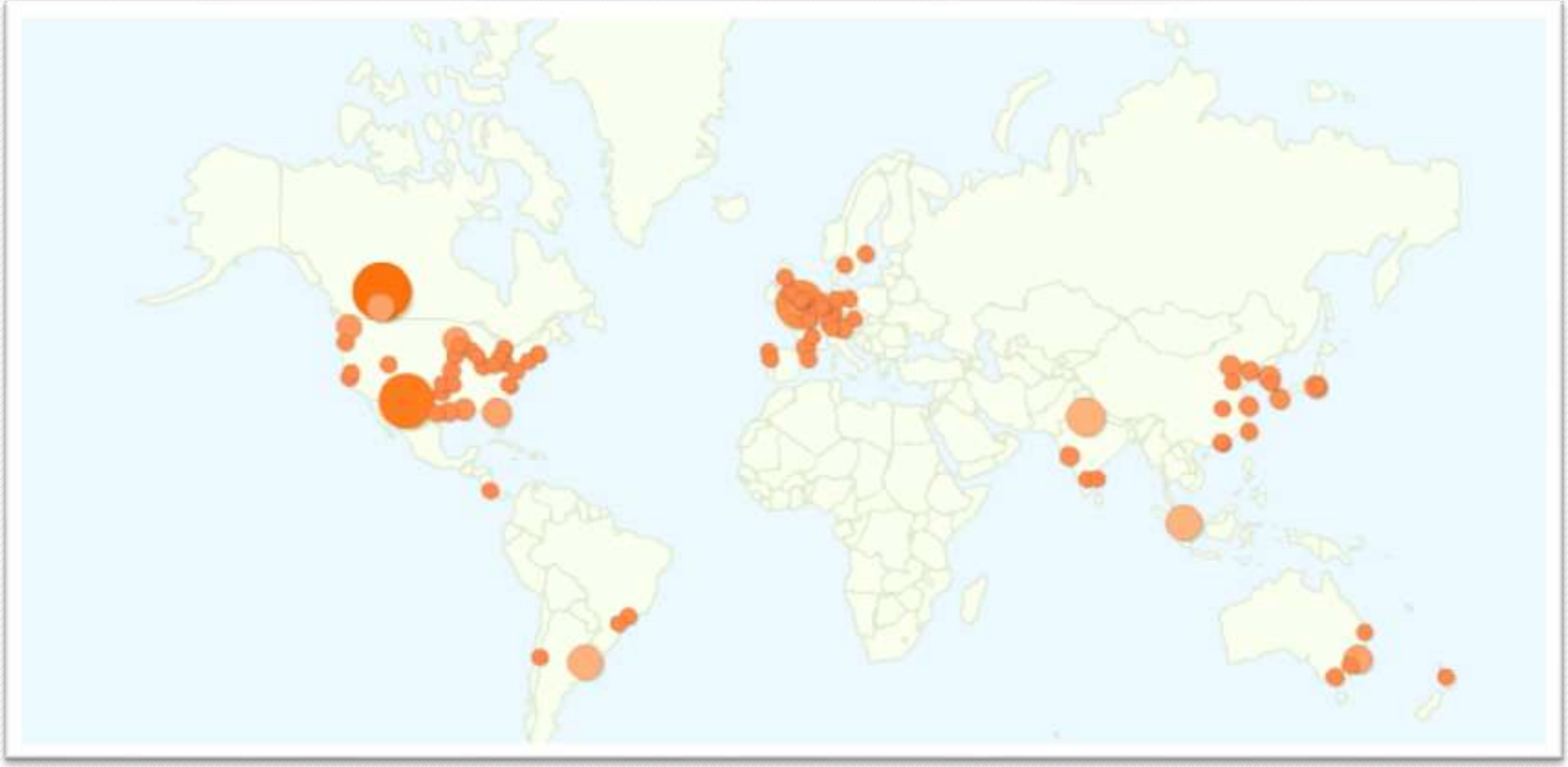
2.2 Principal Component Analysis (PCA)

PCA is an unsupervised method aiming to find the directions that best explain the variance in a data set (X) without referring to class labels (Y). The data are summarized into much fewer variables called *scores* which are weighted average of the original variables. The weighting profiles are called *loadings*. The PCA analysis is performed using the `prcomp` package. The calculation is based on singular value decomposition.

The Rscript `chemometrics.R` is required. Figure 6 is pairwise score plots providing an overview of the various seperation patterns among the most significant PCs; Figure 7 is the scree plot showing the variances explained by the selected PCs; Figure 8 shows the 2-D score plot between selected PCs; Figure 9 shows the 3-D score plot between selected PCs; Figure 10 shows the loading plot between the selected PCs; Figure 11 shows the biplot between the selected PCs.

Figure 6: Pairwise score plots between the selected PCs. The explained variance of each PC is shown in the corresponding diagonal cell.

Some usage statistics



Over 1,200 visits since publication (~15 / day)

Current status

- ✓ Differential Analysis (Biomarker Identification)
- ✓ Class Prediction (Supervised learning)
- ✓ Class Discovery (Clustering)
- UNDER CONSTRUCTION Pathway Analysis

Challenges & future directions

- Unbiased and comprehensive survey of metabolome
  - NMR only able to detect more abundant compound species (> 1 µmol)
  - MS are usually optimized to detect compounds of certain classes
- Systematic classification of compounds (ontology)
- More efficient pathway analysis & visualization

Acknowledgement

- Dr. David Wishart
- Dr. Nick Psychogios
- Nelson Young

- Alberta Ingenuity Fund (AIF)
- The Human Metabolome Project (HMP)
- University of Alberta