Zapping through software for the statistical analysis of spectroscopic data

Edoardo Gaude

Workshop dedicato alla Chemometria Applicata alla Spettroscopia NMR

06/13/2012

Contents

- Fee-paying software
- Open-source web-based tools
- Command line tools
- muma

Contents

- Fee-paying software
- Open-source web-based tools
- Command line tools
- muma

Fee-paying software (I)

- PLS-Toolbox
- CAMO
- AMIX
- SIMCA-P

- User-friendly graphic interface
- Prompt maintenance
  (Updates, custom service, etc)
- Cost

Shared features

- Spectra/Data pre-processing
- Data exploration
- Multivariate analysis/modeling
- Graphical outputs/interpretation

PLS-Toolbox

- Fee: Academic - $695/$395
  Industrial - $2195/1395
- With MatLab *or* Solo

PLS_Toolbox
Advanced Chemometrics Software for use with **MATLAB®**

- Possibility to build your own script
- Friendly Pipe-line
- Interactive plots

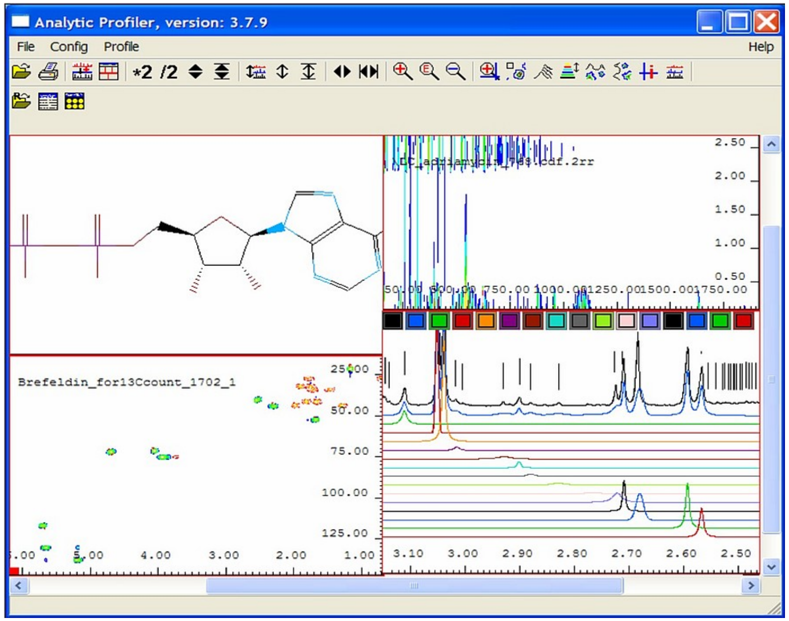
PLS-Toolbox

- Pre-processing
  - Noise, Baseline
  - Filtering (OSC)
  - Normalization
  - Centering/Scaling
- Model building

  Exploration: PCA, Multiway PCA
  Classification: SIMCA, KNN, PLS-DA, SVM
  Regression: PLSR, PCR, MLR, …
- Model validation
  - Cross-validation
- Design of experiment (DoE)

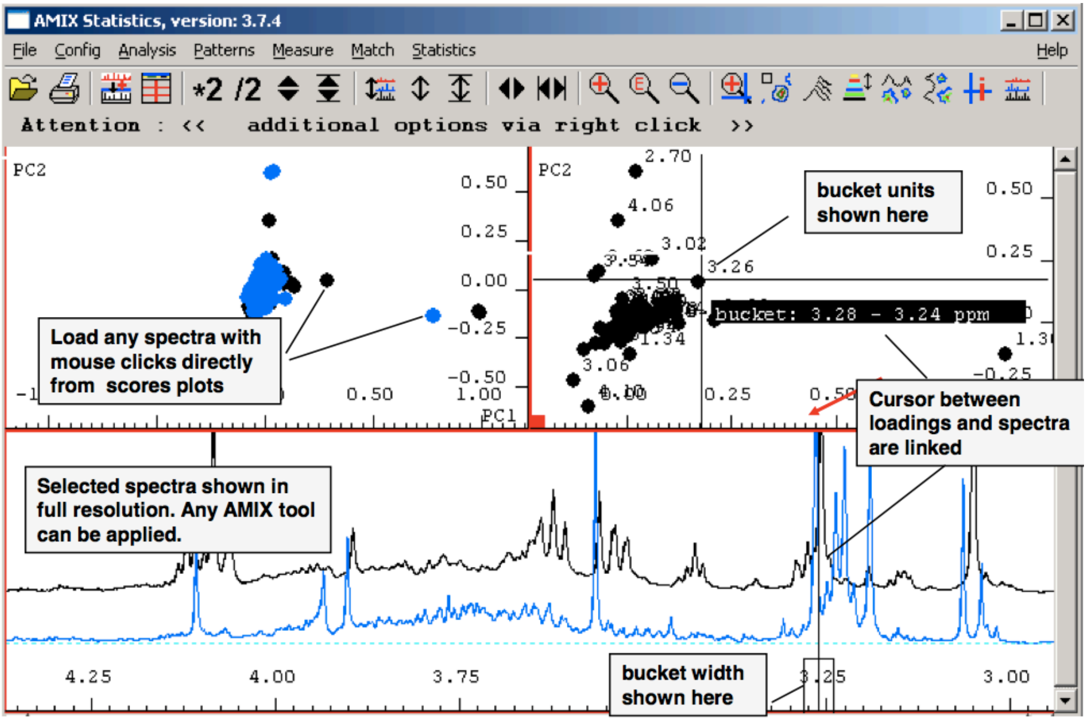
AMIX

- Spectra management
  - Color, Bucketing
  - Metabolite concentration via deconvolution
  - Baseline correction
  - Line shaping
  - Reference DB (1D,2D,J-res)
- Statistics
  - Pareto/Auto scaling
  - Exploration – PCA
  - Classification – PLS, SIMCA
  - Covariance Analysis (STOCSY)
  - Boxplots of selected variables



AMIX

- Interaction between statistics and spectra



Load any spectra with mouse clicks directly from scores plots

bucket units shown here

Cursor between loadings and spectra are linked

Selected spectra shown in full resolution. Any AMIX tool can be applied.

bucket width shown here

- Routine analysis automation

- Link with DBs
    - HMDB: - Import spectra
      - Query
    - KEGG, BMRB, ChEBI, PubChem

CAMO

- Fee: ???
- OS: Windows 7, Vista, XP
- Data pre-treatment
- Exploratory data analysis
  - Descriptive (Mean, SD, ..)
  - Univariate (T, F, Contingency)
  - K-means/Hierarchical clustering
  - PCA (SVD or NIPALS) + ROTATION METHODS
  - Multivariate Curve Resolution (MCR)

CAMO

- Regression
    - PCR, PLSR, OPLS, SVMR
    - L-PLSR
        - "Z" matrix
        - Reduced false positive rate
        - Accuracy
- Classification
    - SIMCA, LDA, SVM
- No extensive model validation

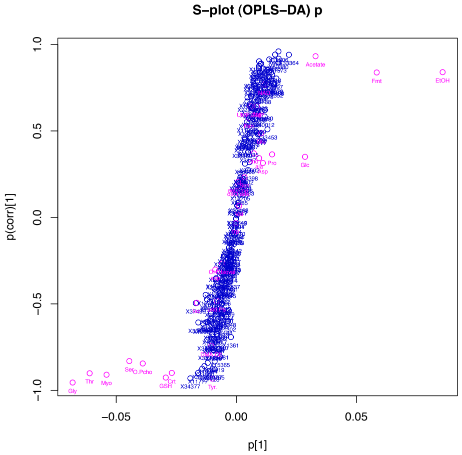
- Design of experiment
- On-line implementation
    - Automation
    - Industrial

SIMCA

- Fee: Academic: €1500
  Industrial: €8000
- OS: Windows 7, Vista, XP
- Data visualization
  - Friendly GUI
  - Comprehensive data import and management
  - Spectrum plot from dataset
- Pre-processing
  - Derivatives, MSC/SNV, de-noising, etc
  - Wavelet denoising/compression
  - Variable and Block scaling

SIMCA

- Modeling
  - Overview: PCA
  - Regression and Discriminant: PLS, OPLS, O2PLS
  - Cluster analysis with PLS tree
- Model validation
  - Cross validation (random/custom)
  - Permutation test
  - CV ANOVA, CV scores
- Graphics
  - Model summary/diagnostics (Hotelling's T2, DModX, ..)
  - VIP
  - Calibration diagnostics
  - Contribution plots
  - Observation/Variable plots
  - Y-related profiles for OPLS and O2PLS



- 2D, 3D scatter, line, column, time series
- Wavelet structure
- Auto/Cross correlation
- Including/Removing data

INTERACTIVE

Contents

- Fee-paying software
- Open-source web-based tools
- Command line tools
- muma

MetaboAnalyst

- First release: 2009

- Update 2.0: March 2012

- Metabolomic-specific data processing and statistical analysis
- Potentiated server

- FAQs section and tutorials

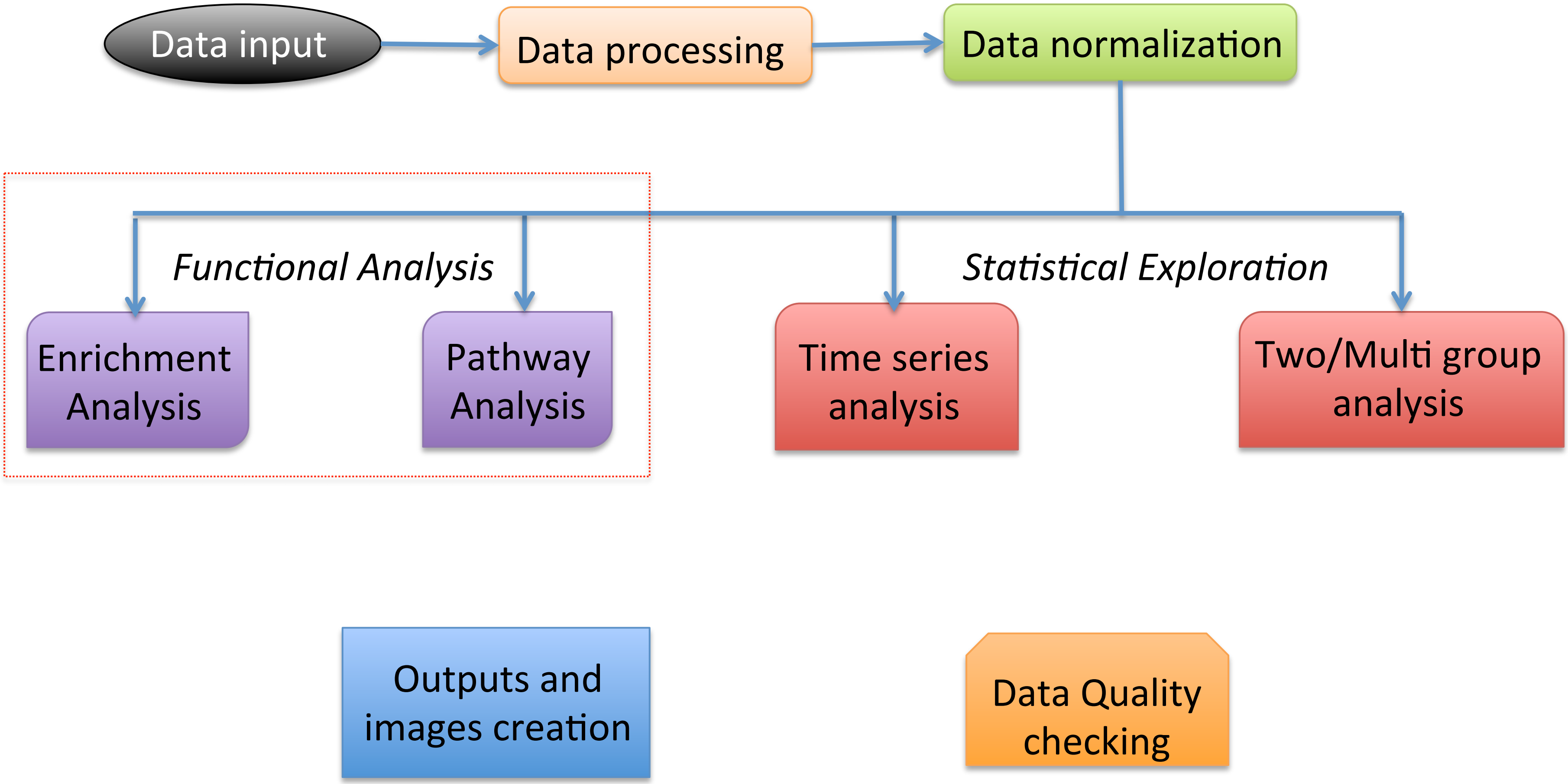
- Downloadable for local installation

FREE

Xia J et al. Nucl Ac Res 2012. 1:7

MetaboAnalyst - Overview

MetaboAnalyst 2.0
-- a comprehensive



Data input

Data processing

Data normalization

Functional Analysis

Enrichment Analysis

Pathway Analysis

Statistical Exploration

Time series analysis

Two/Multi group analysis

Outputs and images creation

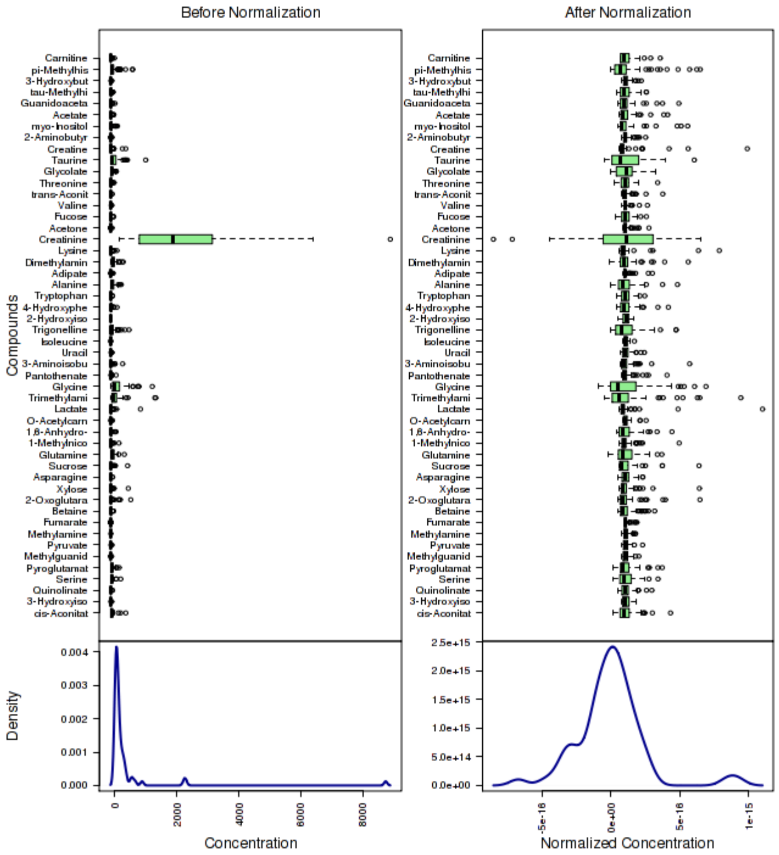
Data Quality checking

Xia J et al. Nucl Ac Res 2012. 1:7

Processing and normalization

MetaboAnalyst 2.0
-- a comprehensive

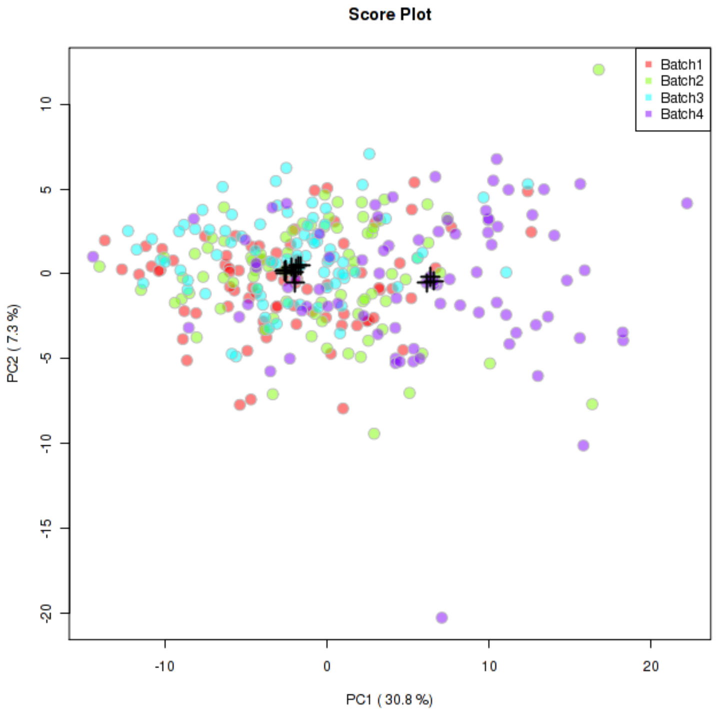
- Data Filtering
  - Remove low-quality data points
  - Low-value threshold
  - Low-variance (noise)
- Data Editing
  - Exclude/Include
- Use standard names of compounds
  - Link to web databases
- Normalization
  - 11 different procedures
  - Diagnostic plots aiding the choice..



Data quality checking

MetaboAnalyst 2.0
-- a comprehensive

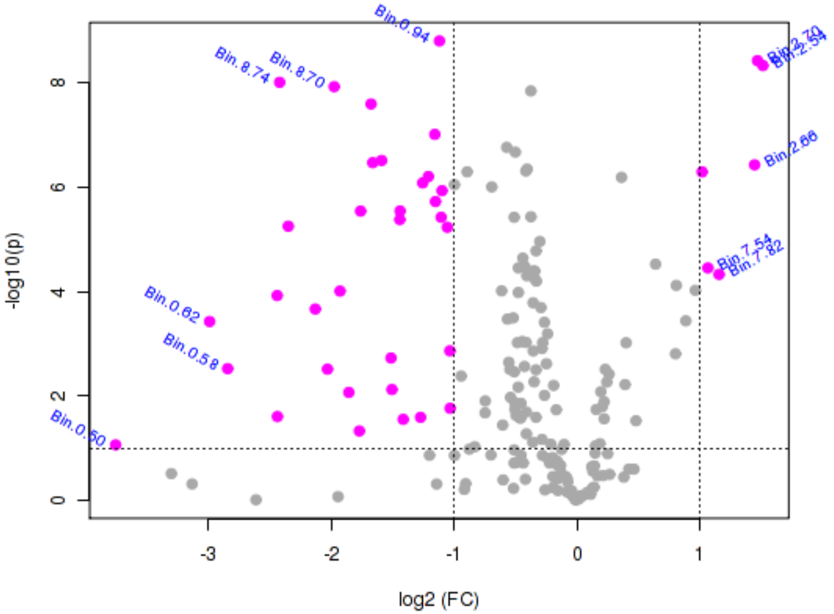
- Pair-wise comparison of 2 measurements
  - – Consistency of 2 protocols/instruments/platforms …
- Temporal drift
  - – Variations related to long-period analyses
- Batch effect
  - – Batch systemic variation
  - – Uni/Multi-variate methods
- Reference concentration ranges
  - – Link with databases
  - – Only for human



Important compounds identification

MetaboAnalyst 2.0
-- a comprehensive

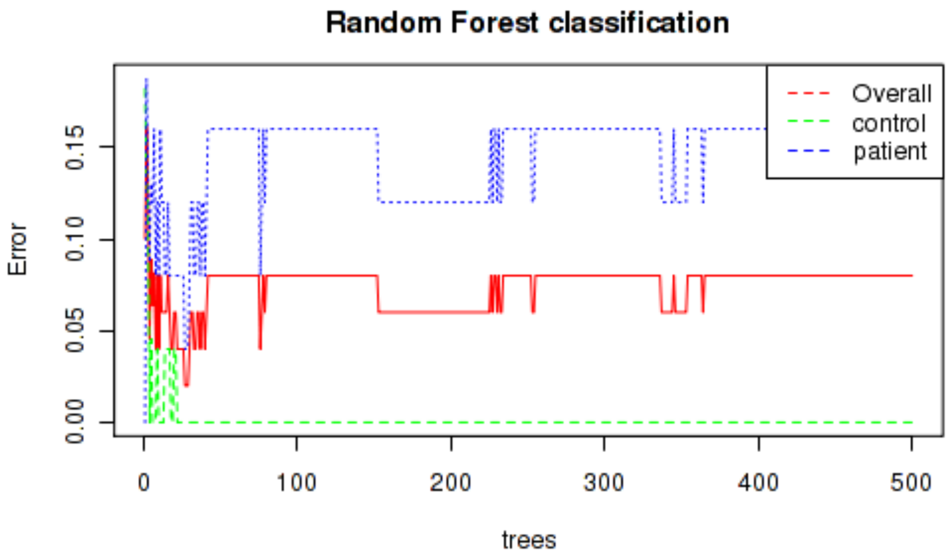
- Differential expression
  - Univariate methods
  - T test/ANOVA
  - Multiple test correction (FDR or Bonferroni)
- Co-expression analysis
  - Correlation heatmaps (STOCSY)
  - 1D and 2D
  - Euclidean distances, Pearson/Spearman correlations
- Pattern search
  - Template matching method
  - (Supervised-like)



Classification and 2 way

MetaboAnalyst 2.0
-- a comprehensive

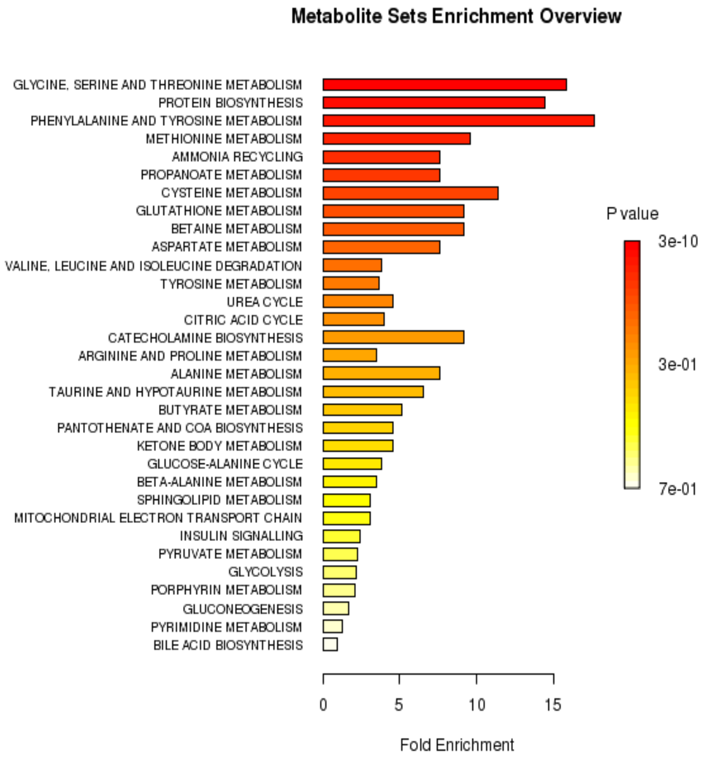
- Chemometrics
  - PCA, PLS-DA
- Machine learning
  - Hierarchical/ K-means clustering
  - Self Organizing Maps (SOMs)
  - SUPERVISED: SVM, Random Forest



- Time course/2 factors Analysis
  - Clustering for 2-way data
  - Within/Between subject ANOVA
  - Multivariate time course with Bayes approach

Functional Interpretation

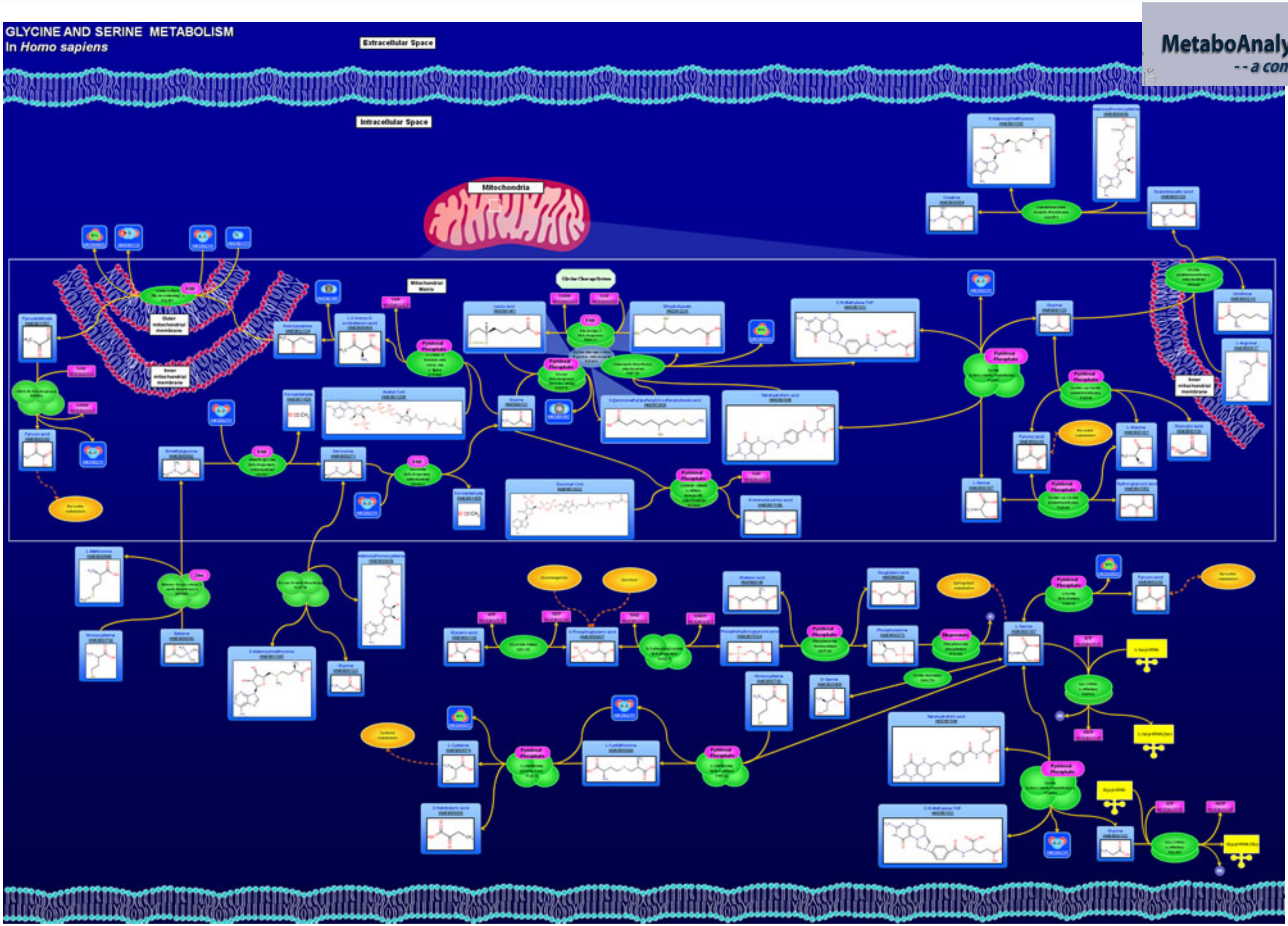
- Over-representation Analysis (ORA)
- Single sample profiling (SSP)
  - Reference concentrations of human biofluid
- Quantitative enrichment Analysis (QEA)
  - Quantification data sets
- Metabolic Pathway Analysis
  - Enrichment analysis + Network topology
  - Metabolic pathways identification
  - Link to DBs



Pathway link

MetaboAnalyst 2.0

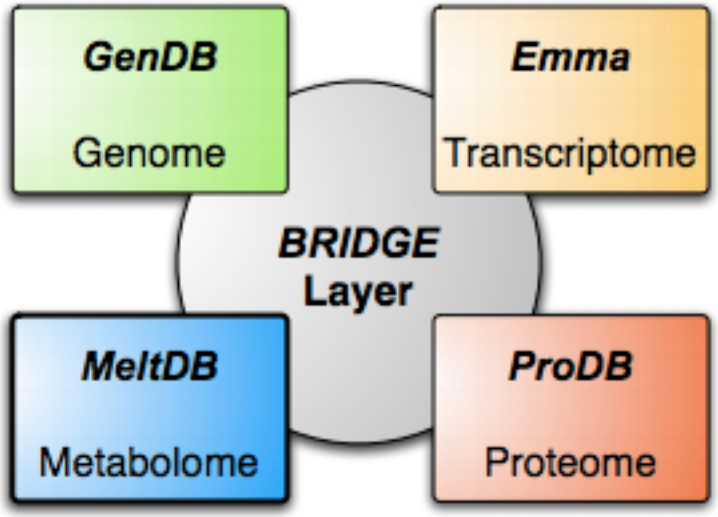
-- a comprehensive



MeltDB

- Raw GC- or LC-MS data sets
- Spectra processing

- R functions integrated
  - T-test, ANOVA
  - PCA
  - Hierarchical Clustering



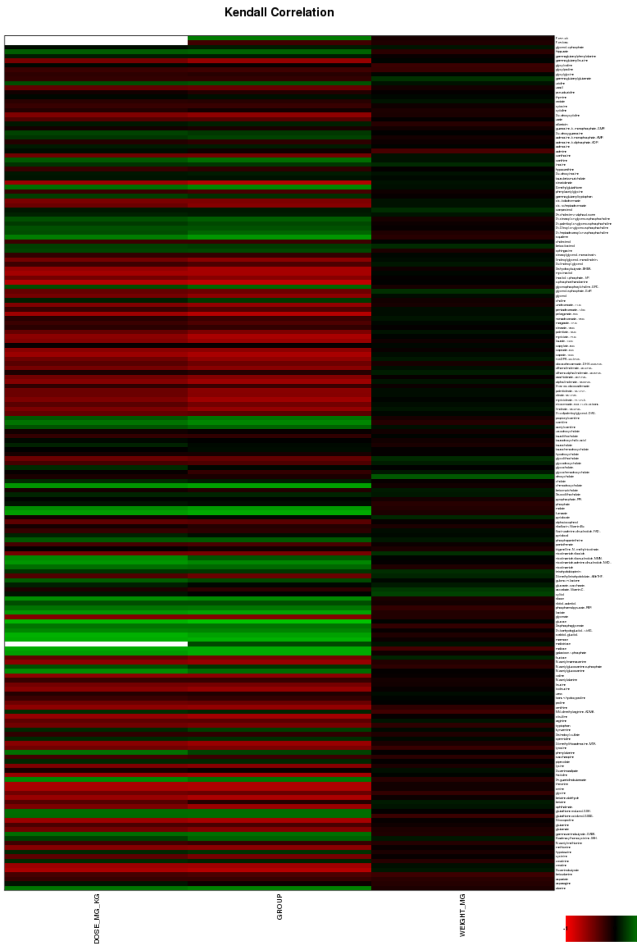
- Link with omics DBs

- On-line forum
  - Real-time discussion

MetaP server

metaP

- No pre-processing
- Common statistical tools
  - PCA
  - Correlation
  - Hypothesis tests
- Ready-to-use PDF reports
  - Box plots, PCA plots, etc…



Contents

- Fee-paying software
- Open-source web-based tools
- Command line tools
- muma

Contents

- Fee-paying software
- Open-source web-based tools
- Command line tools
- muma

R packages

- MS-specific

- CRMN: Cross-contribution compensating multiple standard normalization for quantification of MS metabolomics data sets

- Metab: GC-MS specific for spectra deconvolution and compound identification

R packages - Bioconductor

- Flagme
- Target Search

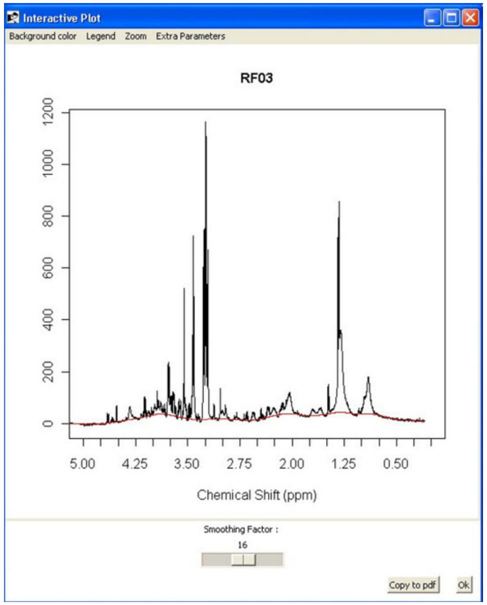
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

- GC-MS spectra processing
  - Peak detection, alignment, retention time shifts, plot spectra, de-noising, etc..
- No statistical analysis

Metabonomic R package

- R-Tcl/Tk Graphical User Interface (GUI)
- R functions
- Windows OS

- Import NMR spectra
  - Processed (text file)
  - Raw (FID, Bruker)

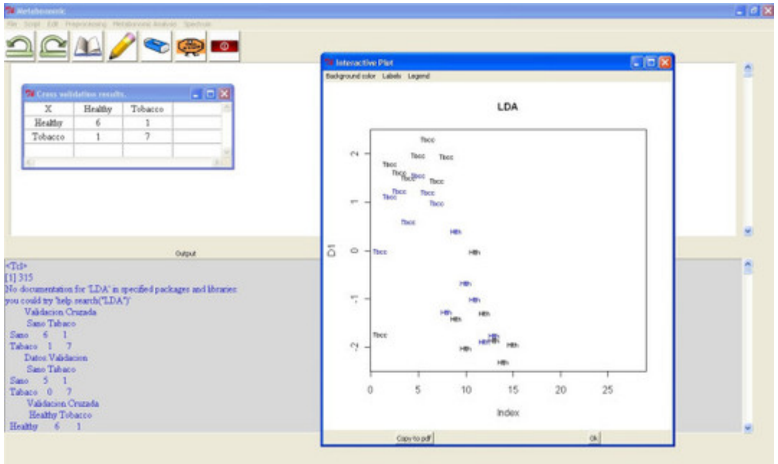


- *Not excellent for spectra management*

Izquierdo-Garcia, JL. *BMC Bioinf*. 2009. 10:363

Metabonomic R package

- Spectra pre-processing
    - Region exclusions
    - Baseline correction (LOESS, FTICRMS)
    - Binning
    - Peak detection
    - Alignment
- Statistics
    - – PCA
    - – LDA
    - – PLS-DA
    - – KNN classification


Metabonomic
File Script Edit Preprocessing Metabonomic Analysis Spectrum
Interactive Plot
Background color Labels Legend
LDA
D1
2
1
0
-1
-2
0
5
10
15
20
25
Index
Tbcc
Hlth
Copy to pdf
Ok
Output

Cross validation results.

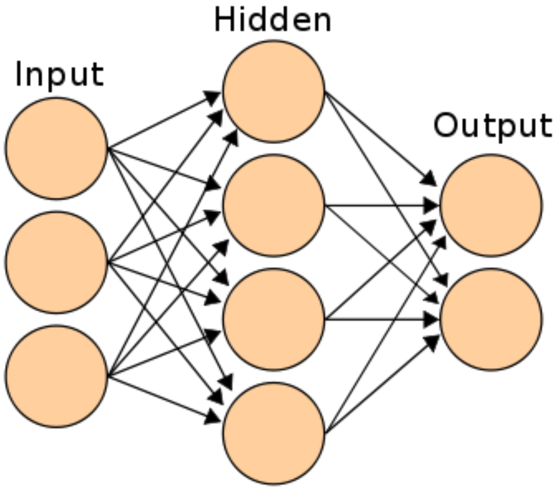
X	Healthy	Tobacco	
Healthy	6	1	
Tobacco	1	7	

```
<Tcl>
[1] 315
No documentation for 'LDA' in specified packages and libraries:
you could try 'help.search("LDA")'
       Validacion Cruzada
        Sano Tabaco
 Sano      6      1
 Tabaco    1      7
       Datos Validacion
        Sano Tabaco
 Sano      5      1
 Tabaco    0      7
        Validacion Cruzada
         Healthy Tobacco
 Healthy      6      1
```

Izquierdo-Garcia, JL. *BMC Bioinf*. 2009. 10:363

Metabonomic R package

- Artificial Neural Networks (ANN)
  - Custom/Random training set
  - Single/Multiple layers networks
- Group-specific spectral differences
  - Plot means of specific ppms



Izquierdo-Garcia, JL. *BMC Bioinf*. 2009. 10:363

Contents

- Fee-paying software
- Open-source web-based tools
- Command line tools
- muma

muma

http://www.biomolnmr.org/software.html

- Metabolomic Univariate & Multivariate Analysis
- R package
- Pipeline for statistical analysis
- Literature-based/well-established methods
- Contents
  - Data pre-processing
  - Dataset exploration
  - Univariate analysis
  - Supervised multivariate analysis
  - NMR molecular assignment
  - Biochemical interpretation
  - Data reporting
  - Graphics for publications

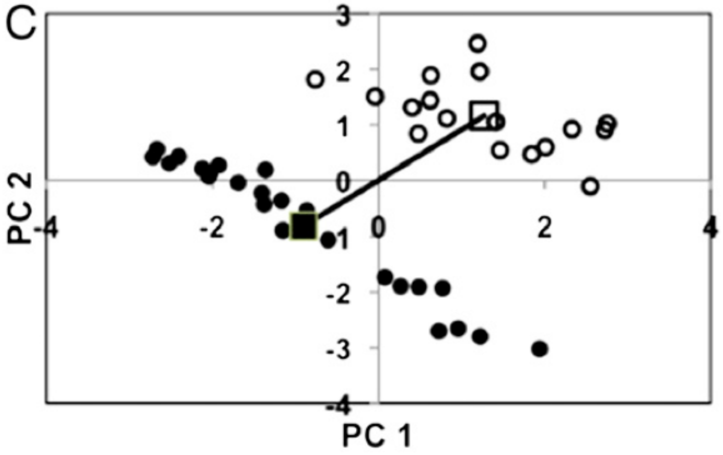
muma – data pre-processing

- Missing values imputation
    - Mean
    - Minimum
    - Half minimum
    - Zero
- Normalization on total spectral area

- Scaling
    - Pareto
    - Auto (Unit Variance)
    - Vast
    - Range

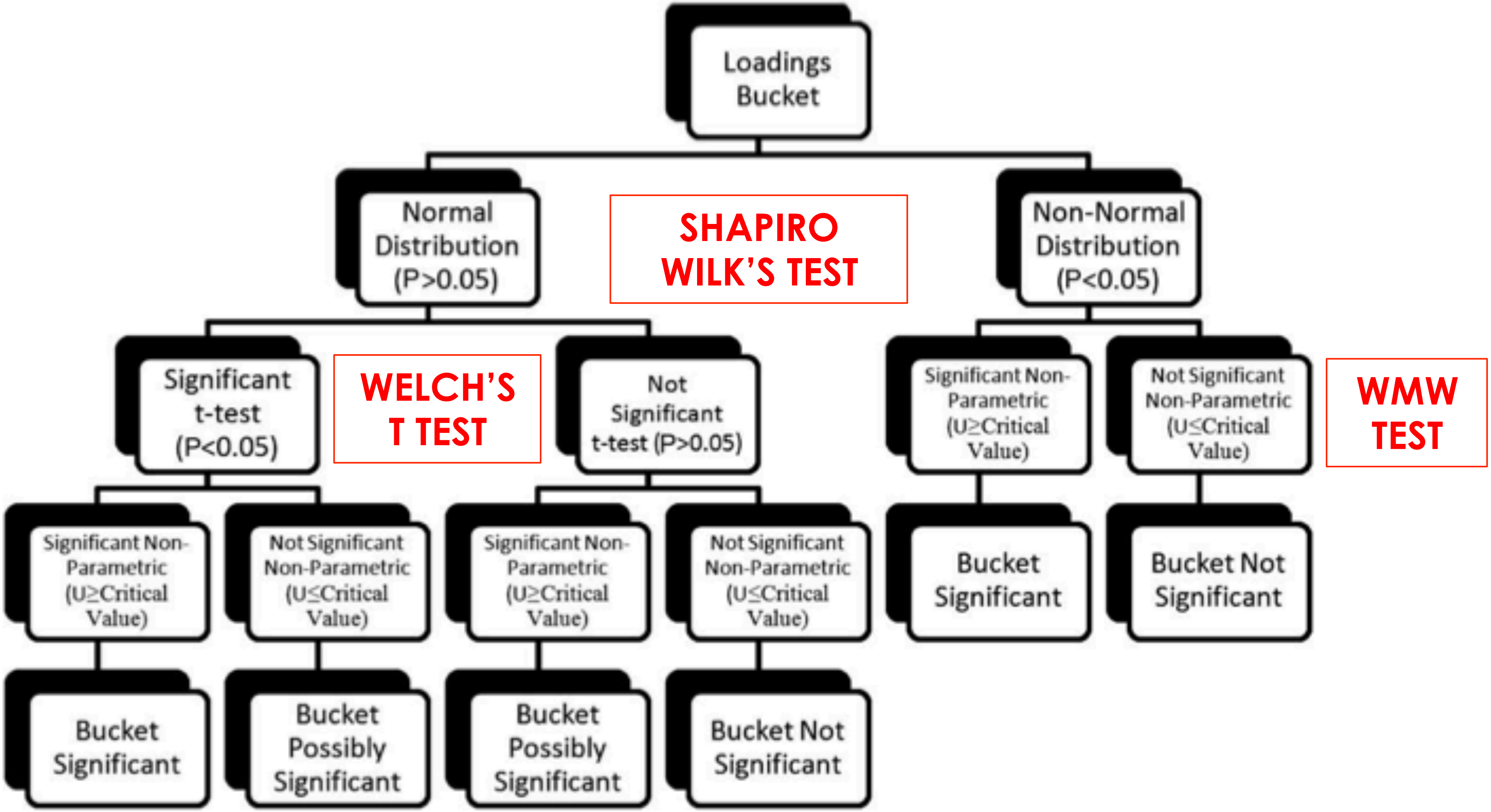
muma – exploration

- PCA
    - Graphical overview of first 10 PCs
    - Automatic ranking of best-separating pairs of PCs
        - Hotelling’s T^2 test
        - F statistics
    - Identify PCs maximizing group separation
- Geometric outlier test
- Score and Loading plots



Goodpaster et al. *Chemometr Intel Lab Sys.* 109 (2011) 162-170

muma – univariate



Loadings Bucket

Normal Distribution (P>0.05)

SHAPIRO WILK'S TEST

Non-Normal Distribution (P<0.05)

Significant t-test (P<0.05)

WELCH'S T TEST

Not Significant t-test (P>0.05)

Significant Non-Parametric (U≥Critical Value)

Not Significant Non-Parametric (U≤Critical Value)

WMW TEST

Significant Non-Parametric (U≥Critical Value)

Not Significant Non-Parametric (U≤Critical Value)

Significant Non-Parametric (U≥Critical Value)

Not Significant Non-Parametric (U≤Critical Value)

Bucket Significant

Bucket Not Significant

Bucket Significant

Bucket Possibly Significant

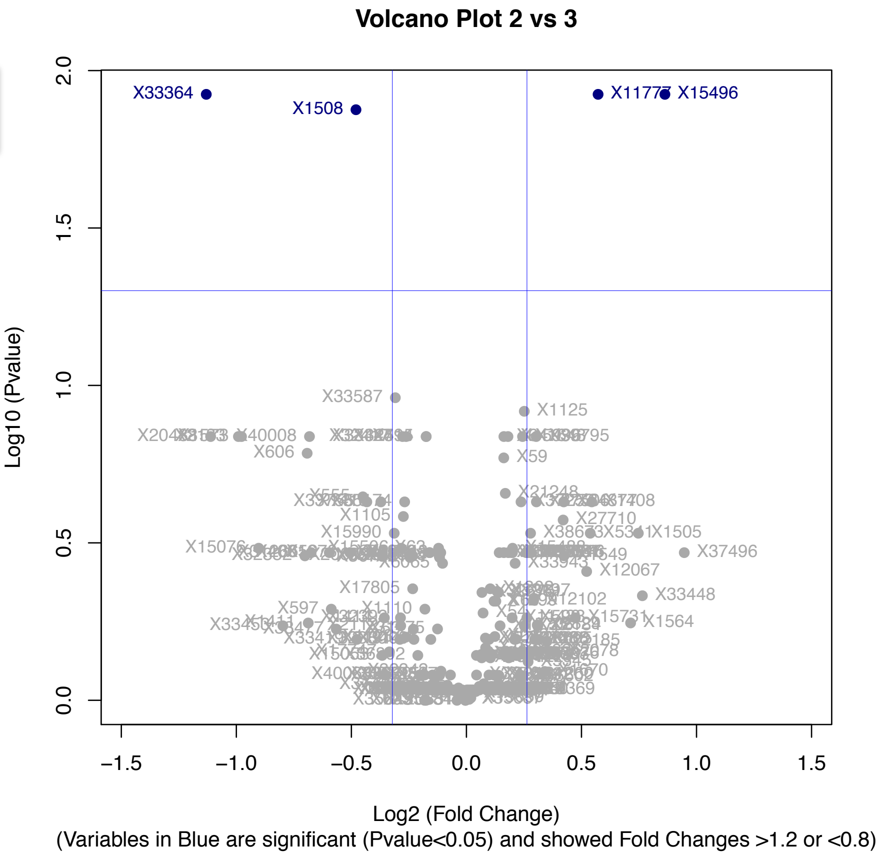
Bucket Possibly Significant

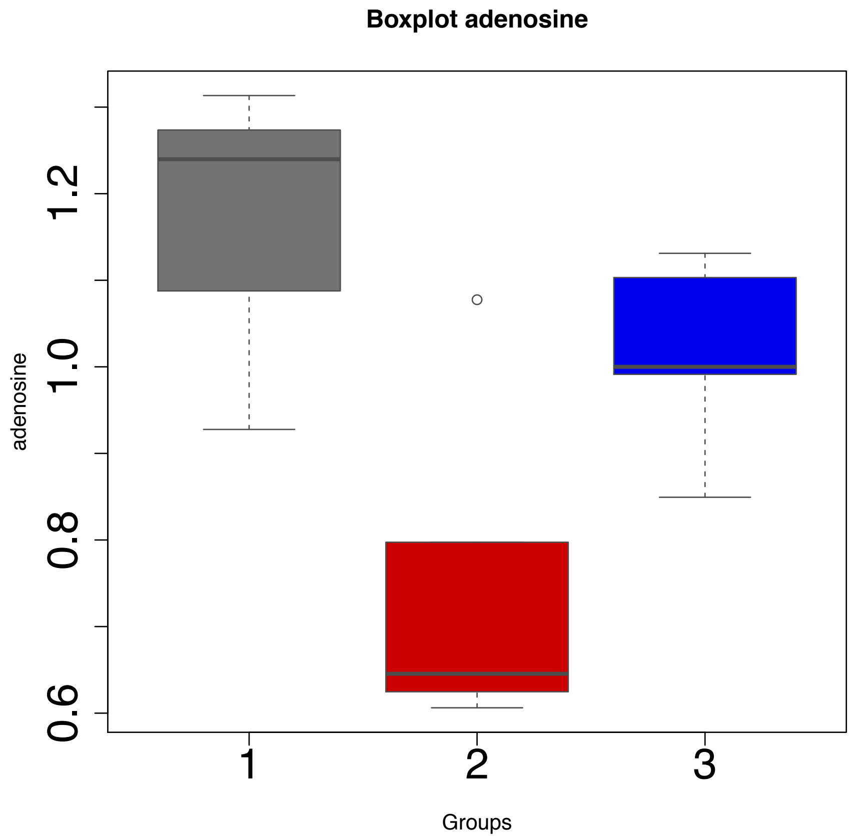
Bucket Not Significant

Goodpaster et al. *Anal Biochem.* 401 (2010) 134-143

muma – univariate graphics

Volcano plots

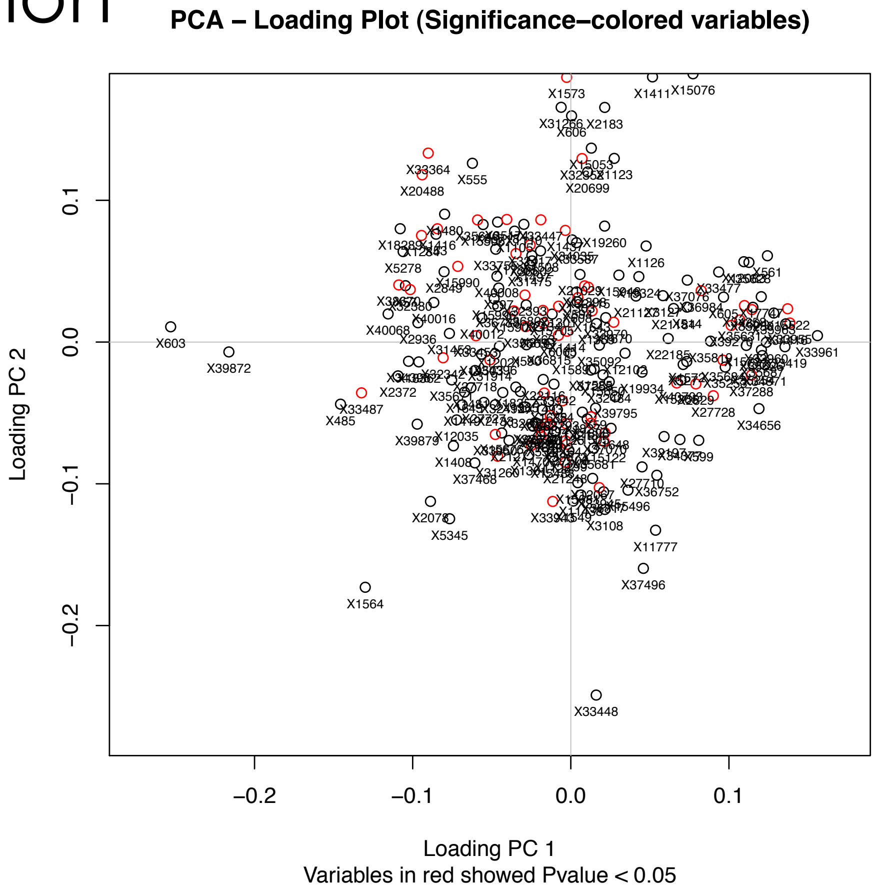




Box plots

muma – multivariate

- Integration of PCA and univariate information

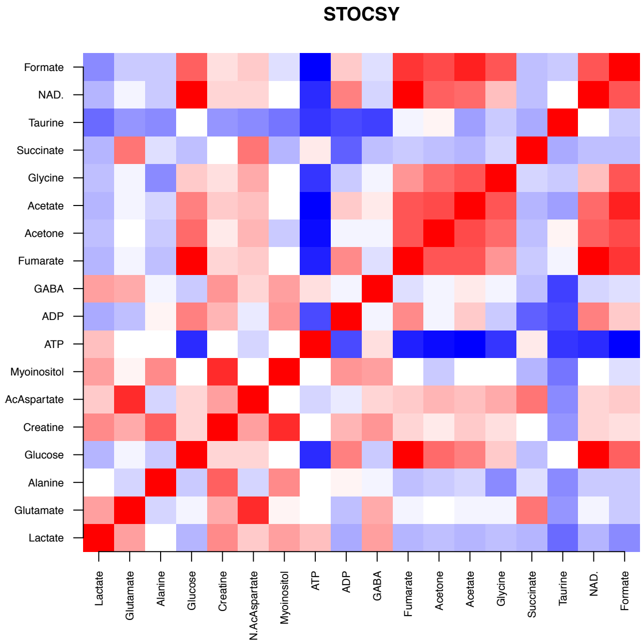


- PLS-DA

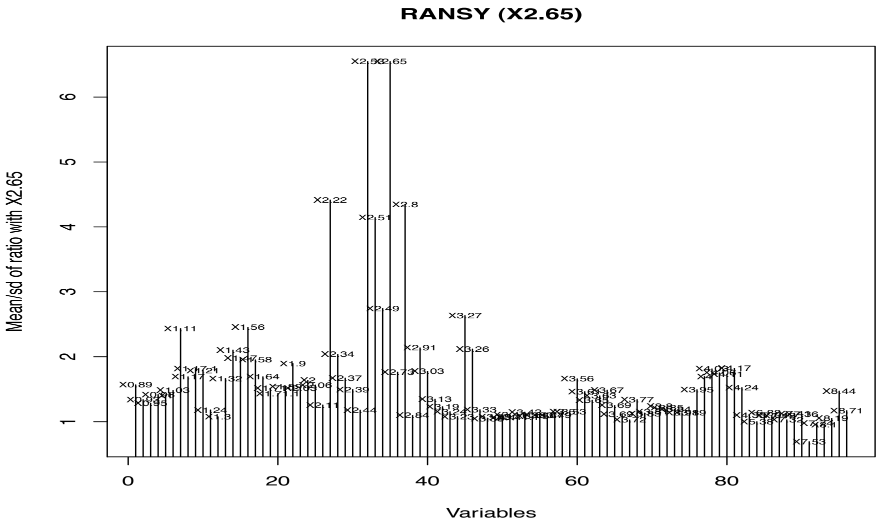
- OPLS-DA (2 classes)

muma – interpretation

- STOCSY
  - Structural correlations (> 0.95)
  - Biochemical correlations (> 0.85 or < -0.85)

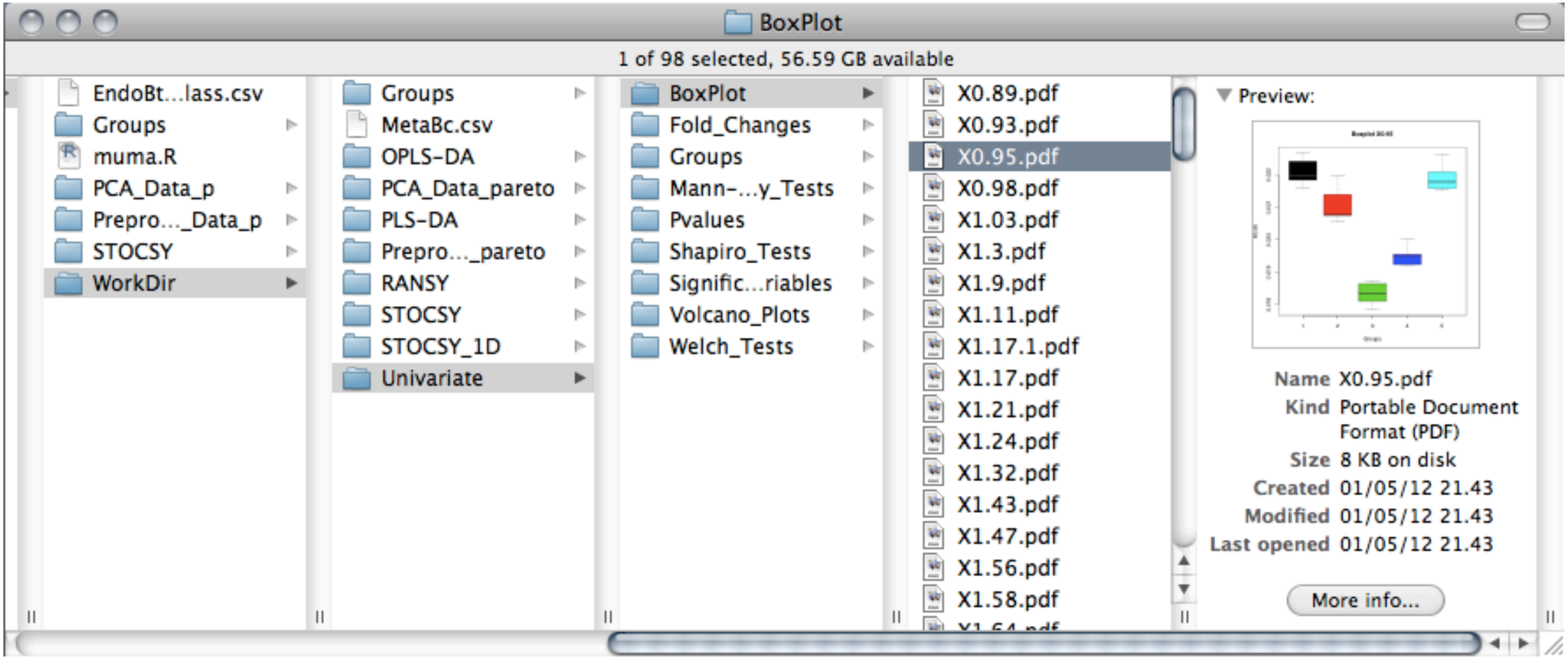


- RANSY
  - Ratio between peaks from the same molecule is conserved
  - Molecular assignment



muma – reports

- Automatic report of:
  - Graphics
  - Data set transformations
  - Test results
  - P-values



BoxPlot

1 of 98 selected, 56.59 GB available

EndoBt...lass.csv
Groups
muma.R
PCA_Data_p
Prepro..._Data_p
STOCSY
WorkDir

Groups
MetaBc.csv
OPLS-DA
PCA_Data_pareto
PLS-DA
Prepro..._pareto
RANSY
STOCSY
STOCSY_1D
Univariate

BoxPlot
Fold_Changes
Groups
Mann-...y_Tests
Pvalues
Shapiro_Tests
Signific...riables
Volcano_Plots
Welch_Tests

X0.89.pdf
X0.93.pdf
X0.95.pdf
X0.98.pdf
X1.03.pdf
X1.3.pdf
X1.9.pdf
X1.11.pdf
X1.17.1.pdf
X1.17.pdf
X1.21.pdf
X1.24.pdf
X1.32.pdf
X1.43.pdf
X1.47.pdf
X1.56.pdf
X1.58.pdf

Preview:

Name X0.95.pdf
Kind Portable Document Format (PDF)
Size 8 KB on disk
Created 01/05/12 21.43
Modified 01/05/12 21.43
Last opened 01/05/12 21.43