Multivariate analysis

Etienne Thévenot
W4M Core Team
etienne.thevenot@cea.fr

10/10/2018

- **The Sacurine study**

Exploratory data analysis

Multivariate modeling

Selection of molecular signatures

Urine diagnosis: From Antiquity to modern metabolomics

I.  explained as *sinatu pizu*, "white or pure urine."

II.  explained as *sinatu zalmi*, "black or dark urine."

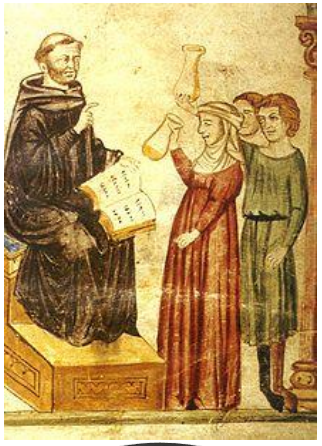
III.  or explained as *urpati sinatu*, "clouds of the urine."

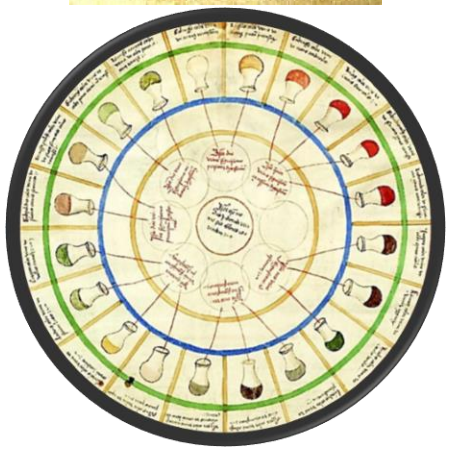
IV.  (lost). Explained as *tidu sa sinatu*, "mud or sediment of the urine."

V.   explained as *sinatu bursi*.

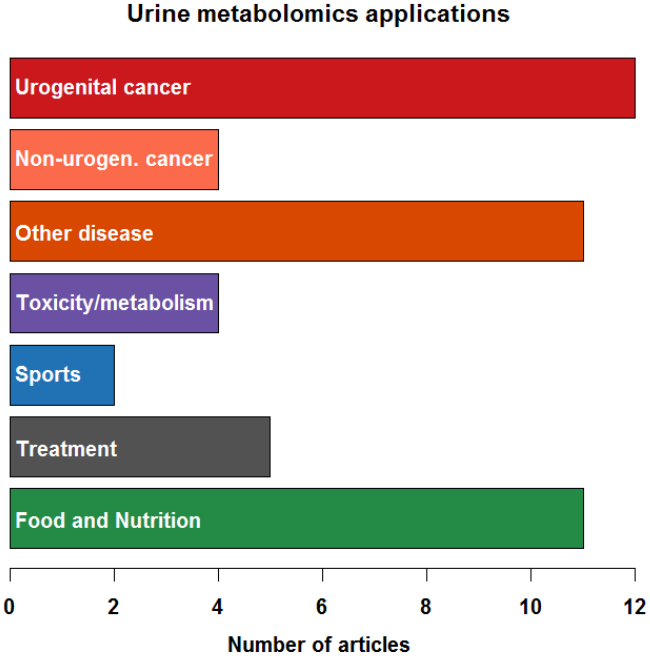
This is a very interesting group, as the second square means "bright, very bright red," and evidently indicates blood-coloured urine.

Summarian and Babylonian physicians (-4000).





Pinder (1506). Epiphanie Medicorum.



adapted from Zhang and Watson (2015). *Analyst,* **140:**2907-2915.

Wellcome H. (1911). The evolution of urine analysis: An historical sketch of the clinical examination of urine. Burroughs Wellcome and Co .

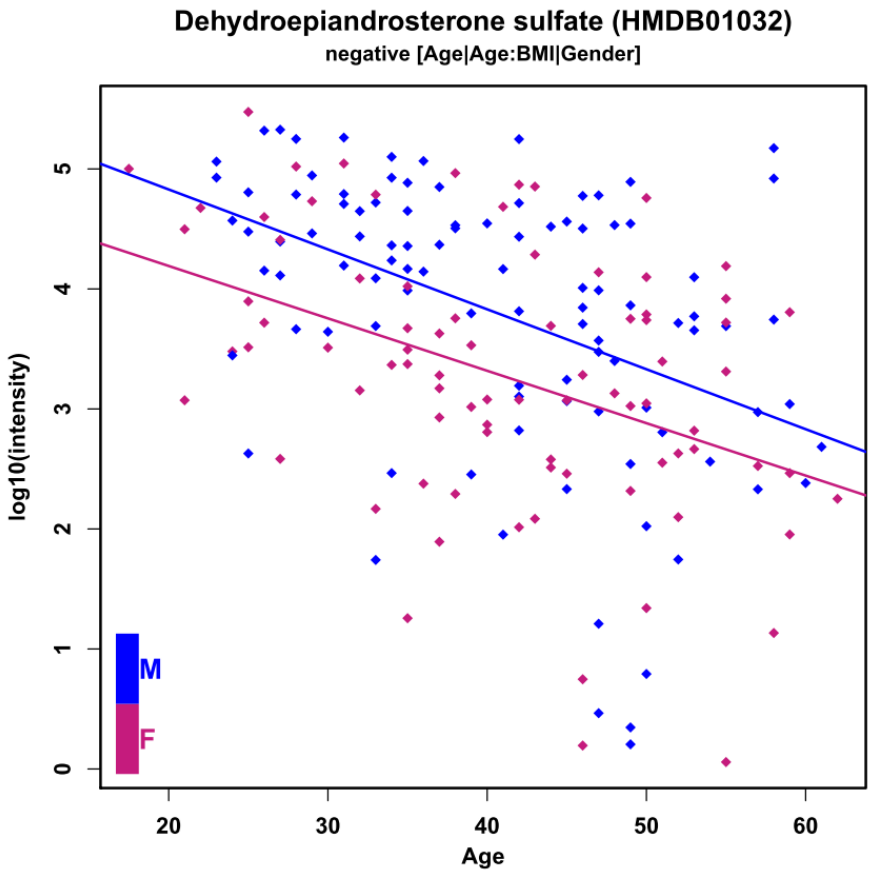
Physiological variations are poorly documented

Identification | Taxonomy | Ontology | Physical properties | Spectra | Biological properties | Concentrations | Links | References | enzymes (2)

Show 2 proteins | XML

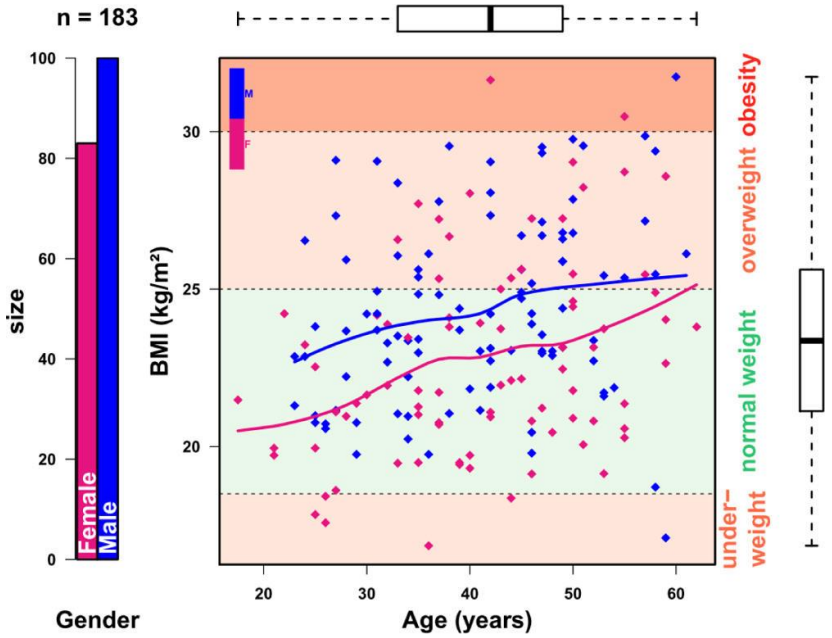
Normal Concentrations

Biofluid	Status	Value	Age	Sex	Condition	Reference	Details
Blood	Detected and Quantified	3.47 +/- 1.20 uM	Adult (>1 old)				
Blood	Detected and Quantified	4.5 (1.4-8.2) uM	Adult (>1 old)				
Urine	Detected and Quantified	1.25 (0.0026-2.52) umol/mmol creatinine	Adult (>1 old)				
Urine	Detected and Quantified	1.34 +/- 0.23 umol/mmol creatinine	Adult (>1 old)				



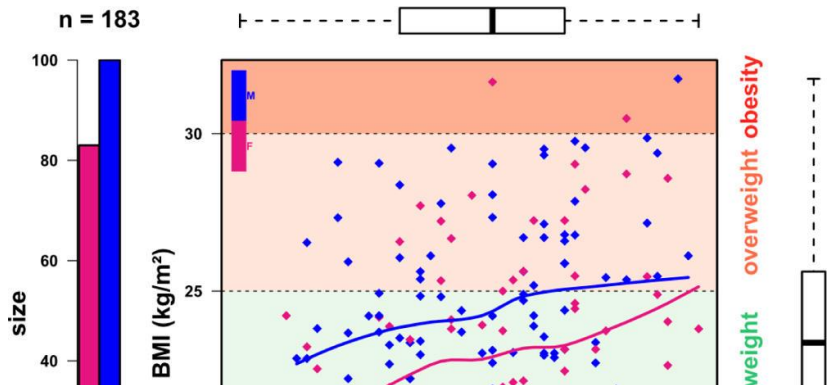
Sacurine dataset (MTBLS404)

- Objective: influence of age, body mass index and gender on metabolite concentrations in urine
- Cohort: 184 employees from the CEA institute
- Analytics: LTQ-Orbitrap (negative ionization mode)
- Annotation: 109 metabolites were identified or annotated at the MSI level 1 or 2.
- Pre-processing:
  - XCMS followed by Quan Browser
  - Signal drift and batch effect correction
  - Normalization to the osmolality
  - log10 transformation



Sacurine dataset (MTBLS404)

- Objective: influence of age, body mass index and gender on metabolite concentrations in urine
- Cohort: 184 employees from the CEA institute
- Analytics: LTQ-Orbitrap (negative ionization mode)
- Annotation: 109 metabolites were identified or annotated at the MSI level 1 or 2.
- Pre-processing:
  - XCMS followed by Quan Browser
  - Signal drift and batch effect correction
  - Normalization to the osmolality
  - log10 transformation



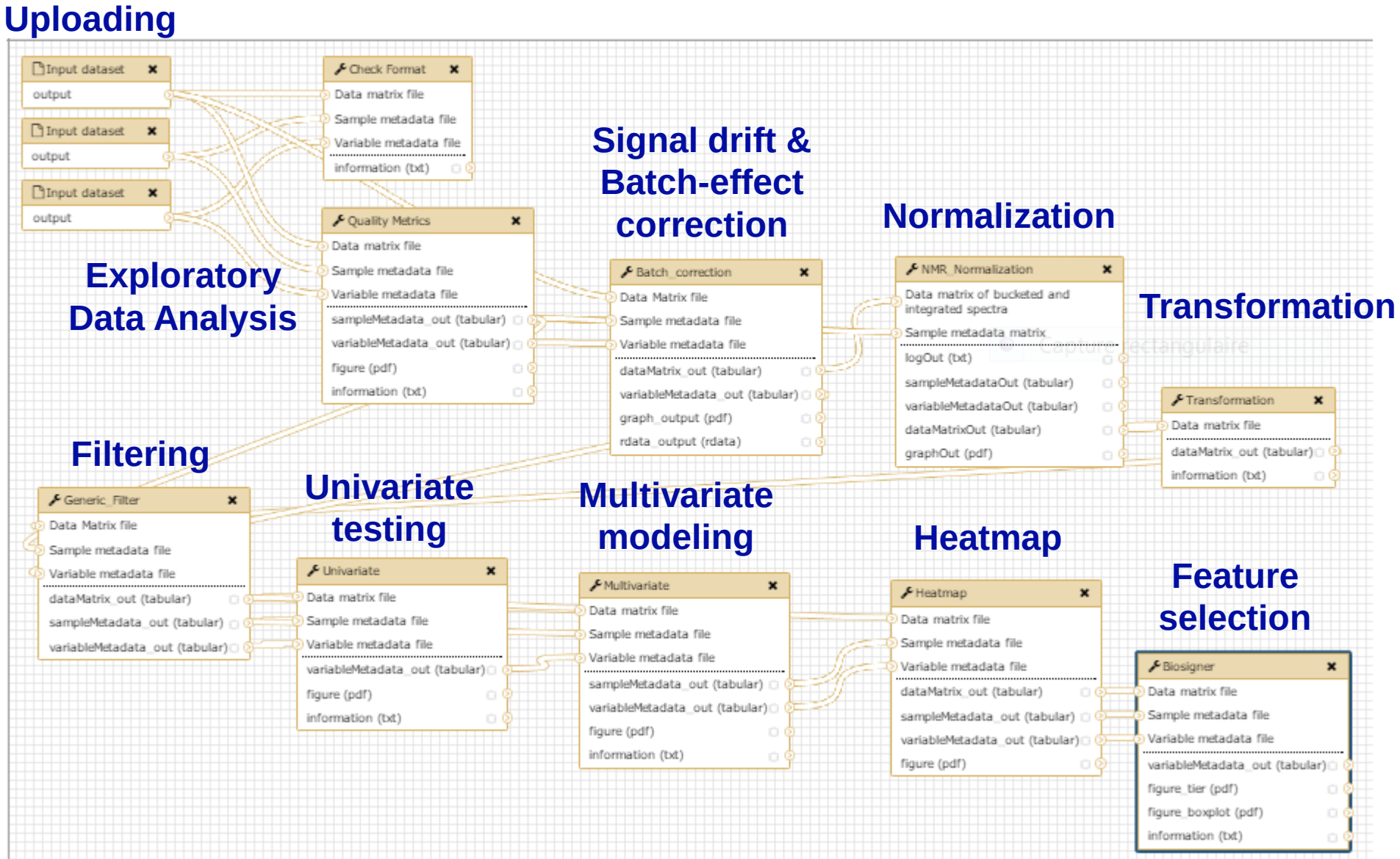
Journal of proteome research
Article
pubs.acs.org/jpr

Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive Workflow for Univariate and OPLS Statistical Analyses

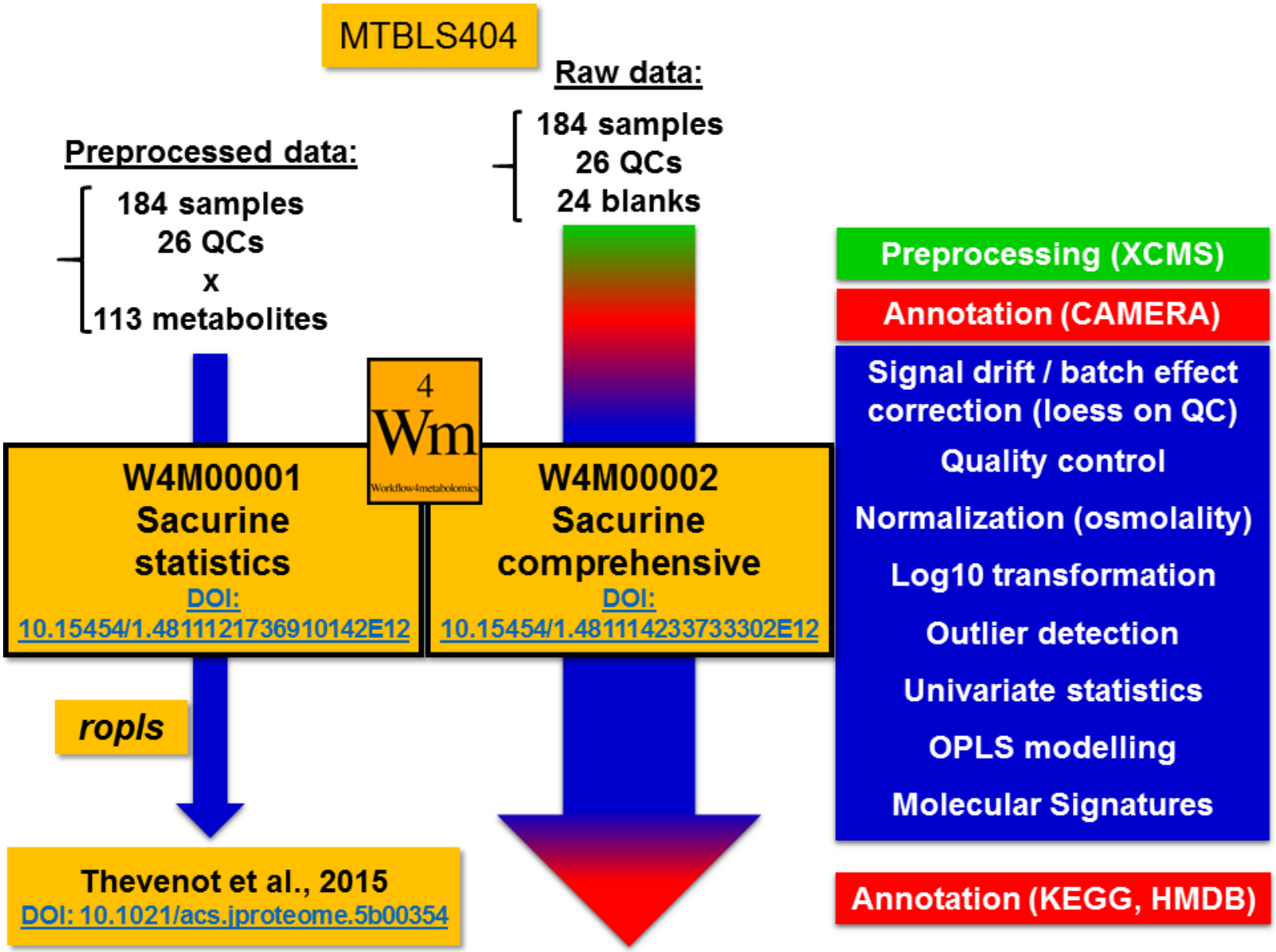
Etienne A. Thévenot,*[,†,¶] Aurélie Roux,[‡,¶] Ying Xu,[‡] Eric Ezan,[‡] and Christophe Junot*[,‡]

[†]CEA, LIST, Laboratory for Data Analysis and Smart Systems, MetaboHUB Paris, F-91191 Gif-sur-Yvette, France
[‡]Laboratoire d'Etude du Métabolisme des Médicaments, DSV/iBiTec-S/SPI, MetaboHUB Paris, CEA-Saclay, Gif-Sur-Yvette, France

W4M00001_Sacurine-statistics



W4M00001 and 2: Human physiology



Developing & implementing new methods

Methods




FIA processing (O)PLS(-DA) Feature selection

Packages

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

proFIA *ropls* *biosigner*

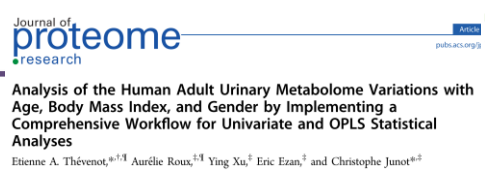
W4M Modules

Galaxy

proFIA Multivariate Biosigner

References

Delabriere et al, *under review*


Journal of proteome research
Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive Workflow for Univariate and OPLS Statistical Analyses
Etienne A. Thévenot, Aurélie Roux, Ying Xu, Eric Ezan, and Christophe Junot

Thevenot et al, 2015 *J. Prot. Res.*


frontiers in Molecular Biosciences
biosigner: A New Method for the Discovery of Significant Molecular Signatures from Omics Data
Philippe Rinaudo, Samia Boudah, Christophe Junot and Etienne A. Thévenot

Rinaudo et al, 2016, *Front. Mol. Biosciences*

The Sacurine study

➢ **Exploratory Data Analysis**

Multivariate modeling

Selection of molecular signatures

Objectives

- Visualize your data
- Detect potential clusters of samples
- Detect potential sample outliers

=> To be performed before any statistical analysis

The "Quality Metrics" tool

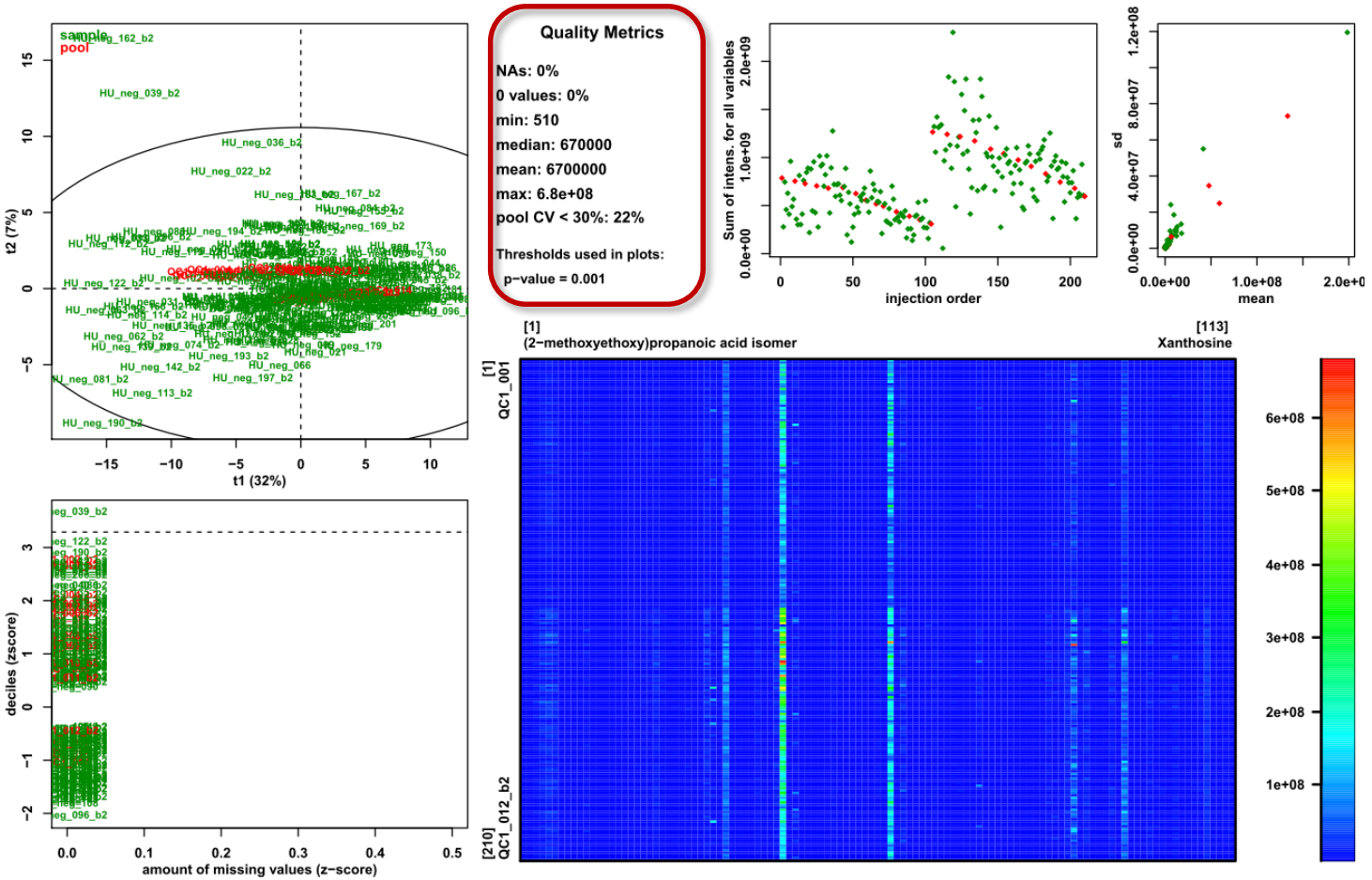
Quality Metrics

- Includes:
  - graphics for data visualization
  - metrics for outlier detection and quality control
    - *p*-values for samples (columns added in sampleMetadata)
      - Hotelling's T2 (Mason et al, 1997)
      - intensity distribution (Alonso et al, 2011)
      - proportion of missing values (Alonso et al, 2011)
    - metrics for variables (columns added in variableMetadata)
      - coefficient of variation
- Depending on the results:
  - intensities can be log transformed — Transformation
  - outliers can be discarded — Generic Filter

The "Quality Metrics" tool

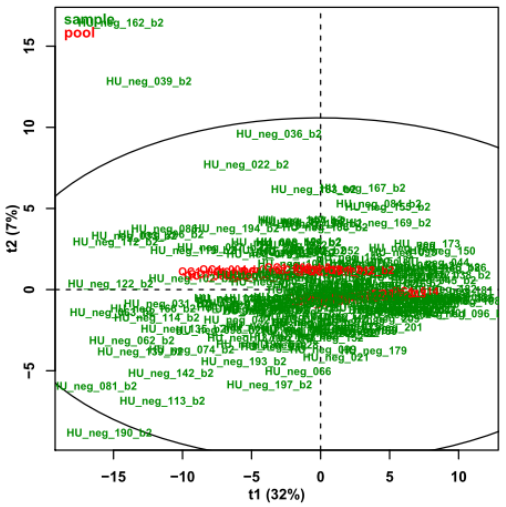
Quality Metrics

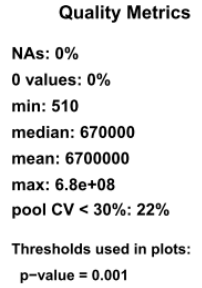
Summary of the intensities in the dataMatrix

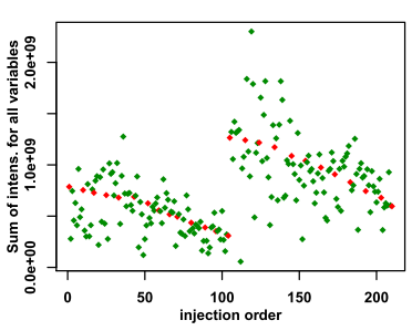


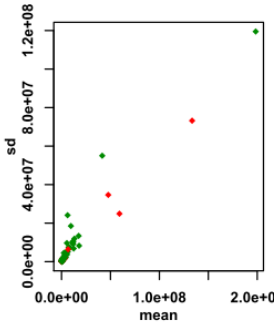
The "Quality Metrics" tool

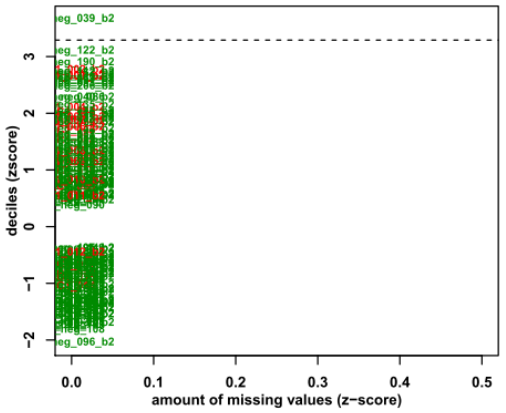
Quality Metrics

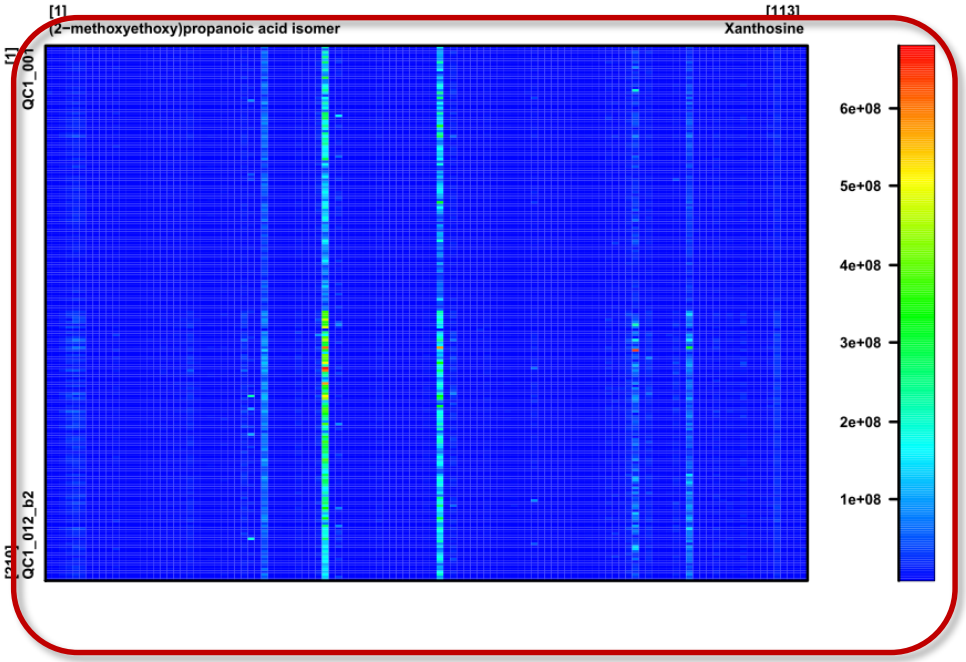










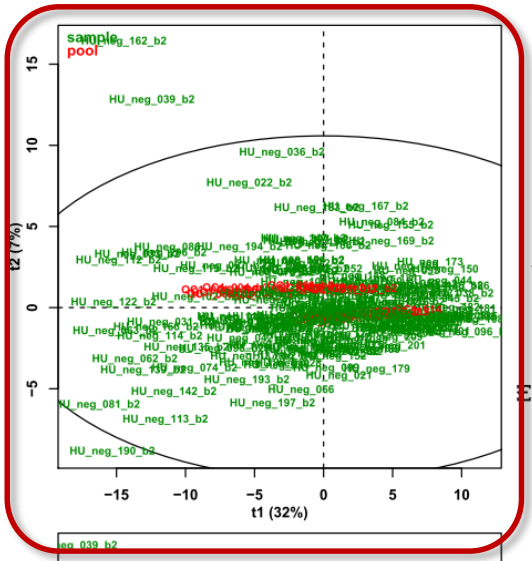


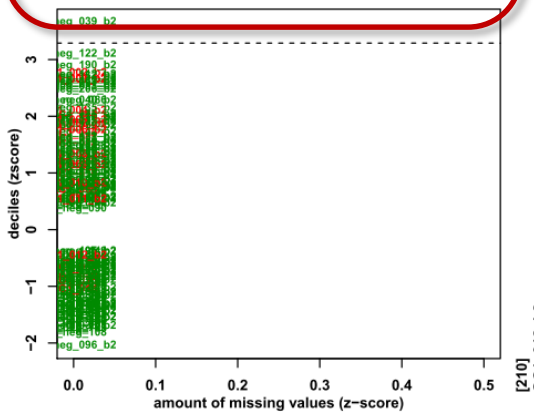
View of the intensities in the dataMatrix

The "Quality Metrics" tool

Quality Metrics

View the PCA scores (check for clusters, outliers)

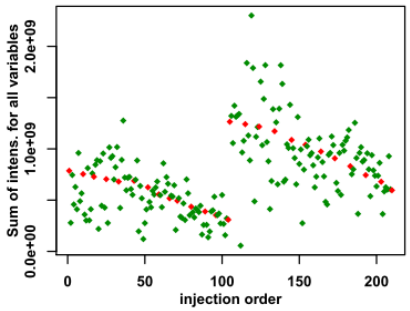


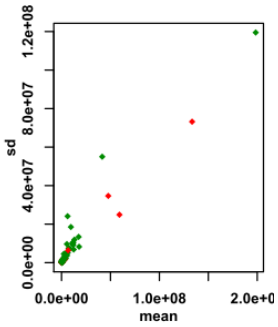


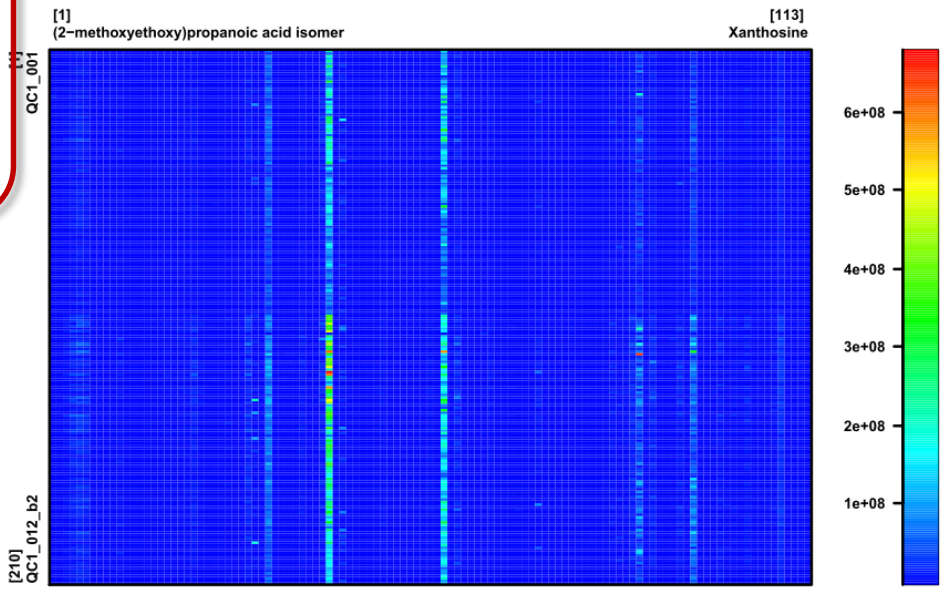
Quality Metrics

NAs: 0%
0 values: 0%
min: 510
median: 670000
mean: 6700000
max: 6.8e+08
pool CV < 30%: 22%

Thresholds used in plots:
p−value = 0.001



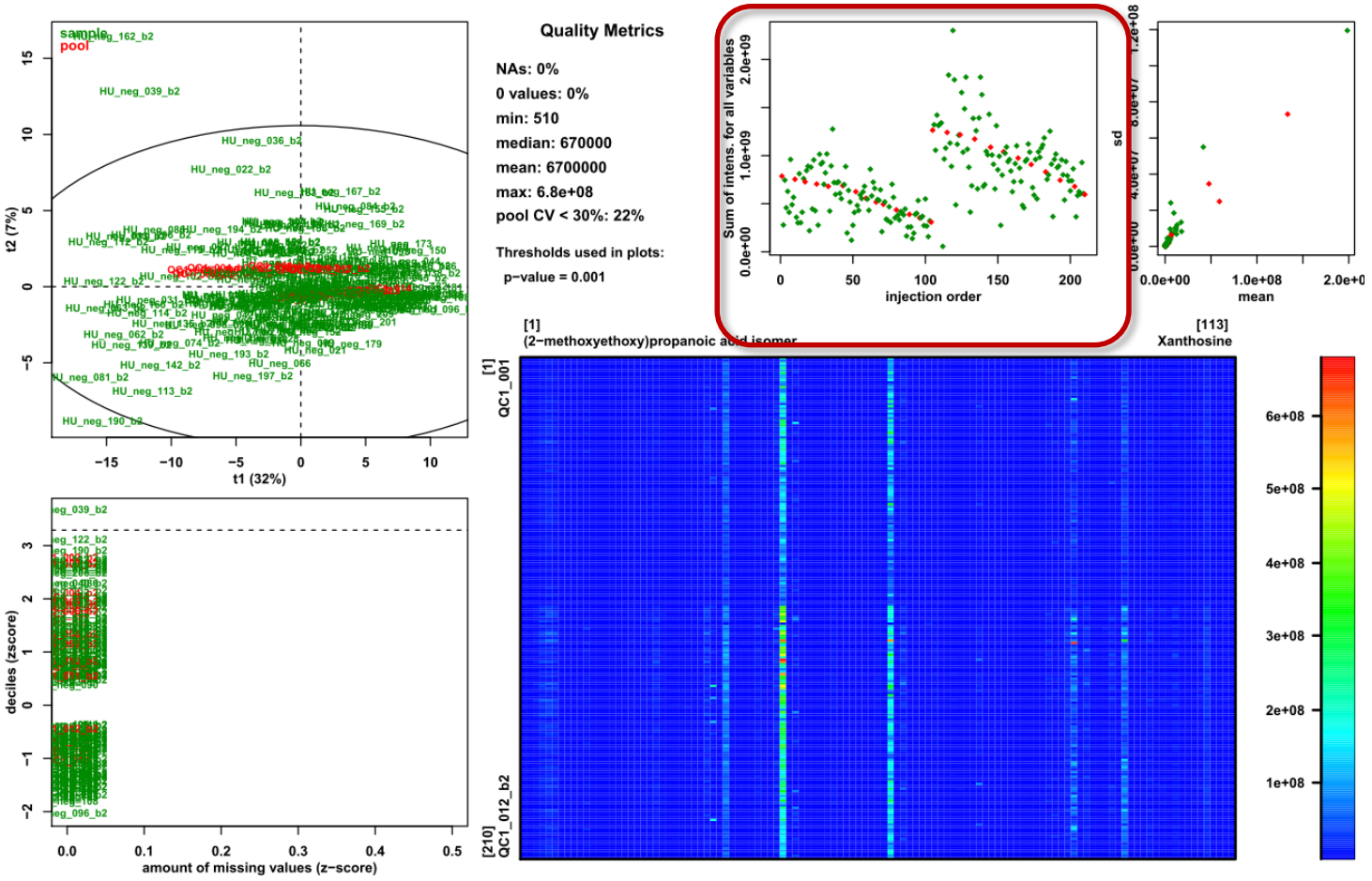




The "Quality Metrics" tool

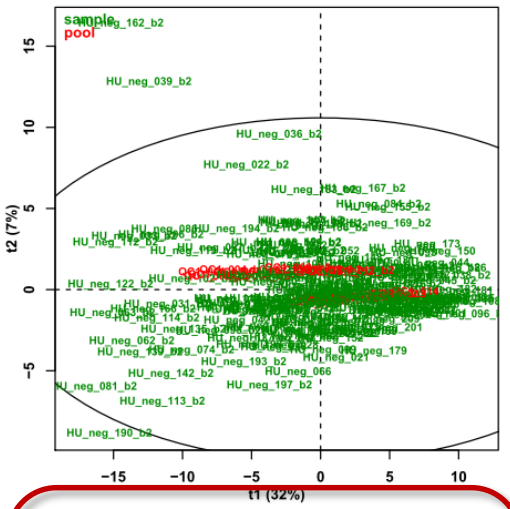
Quality Metrics

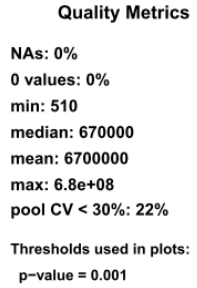
Check the absence of signal drift (LC-MS)

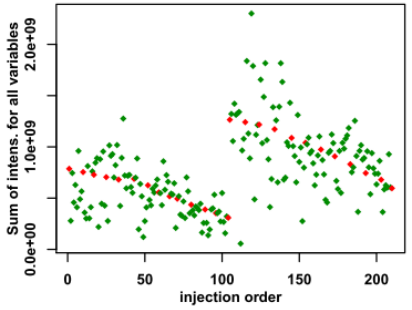


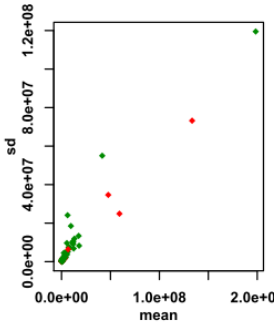
The "Quality Metrics" tool

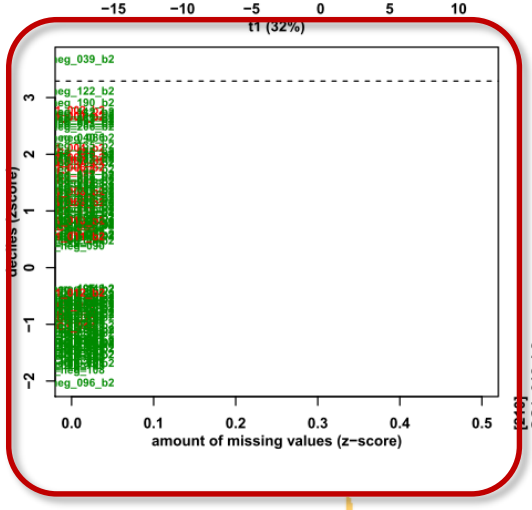
Quality Metrics

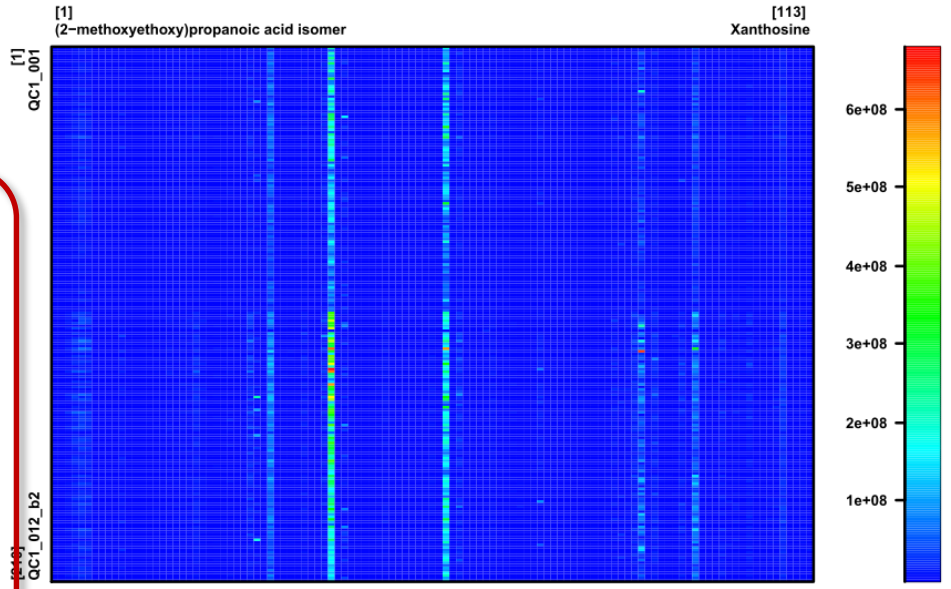










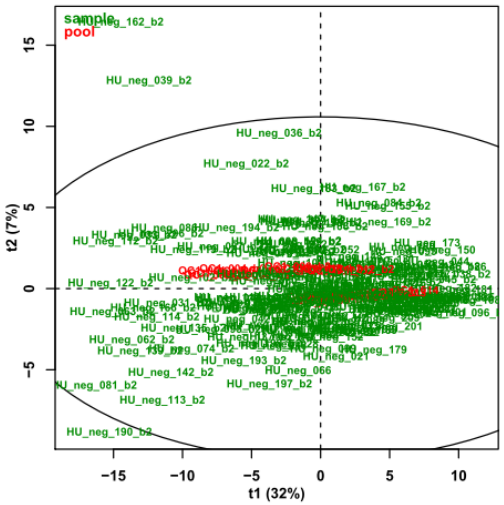


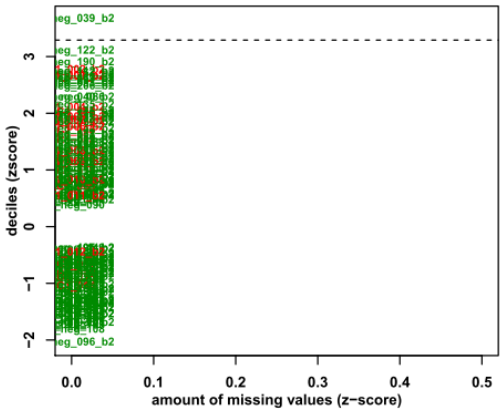
Check the absence of samples with outlier intensity distribution or outlier proportion of missing values

The "Quality Metrics" tool

Quality Metrics

Check the absence of correlation between mean and standard deviation





Quality Metrics

NAs: 0%

0 values: 0%

min: 510

median: 670000

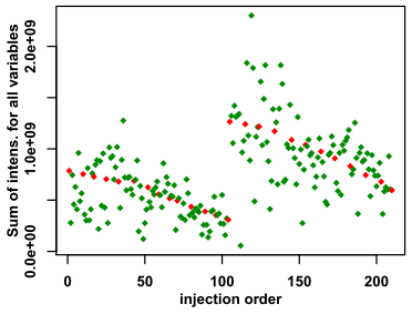
mean: 6700000

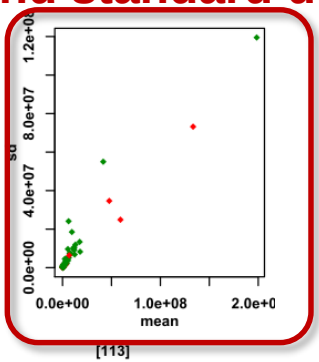
max: 6.8e+08

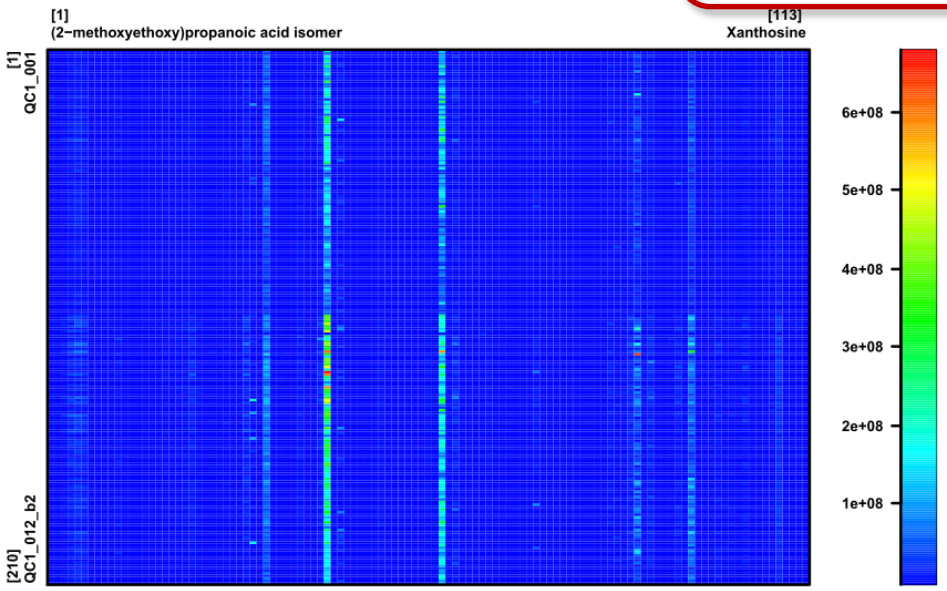
pool CV < 30%: 22%

Thresholds used in plots:

p−value = 0.001







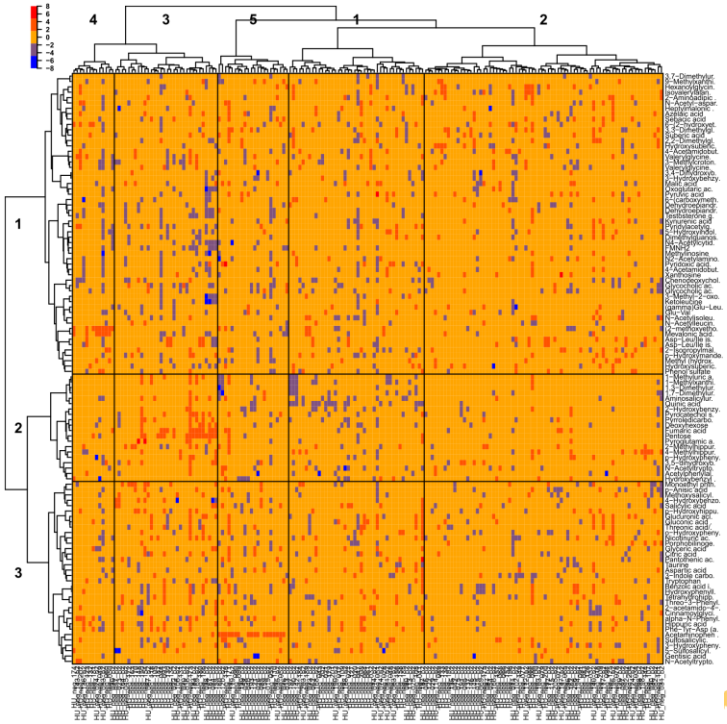
Other tools for unsupervised analysis

- Principal Component Analysis

ACP **Multivariate**

- Clustering

Hierarchical Clustering **Heatmap**



The Sacurine study

Exploratory Data Analysis

➢ **Multivariate modeling**

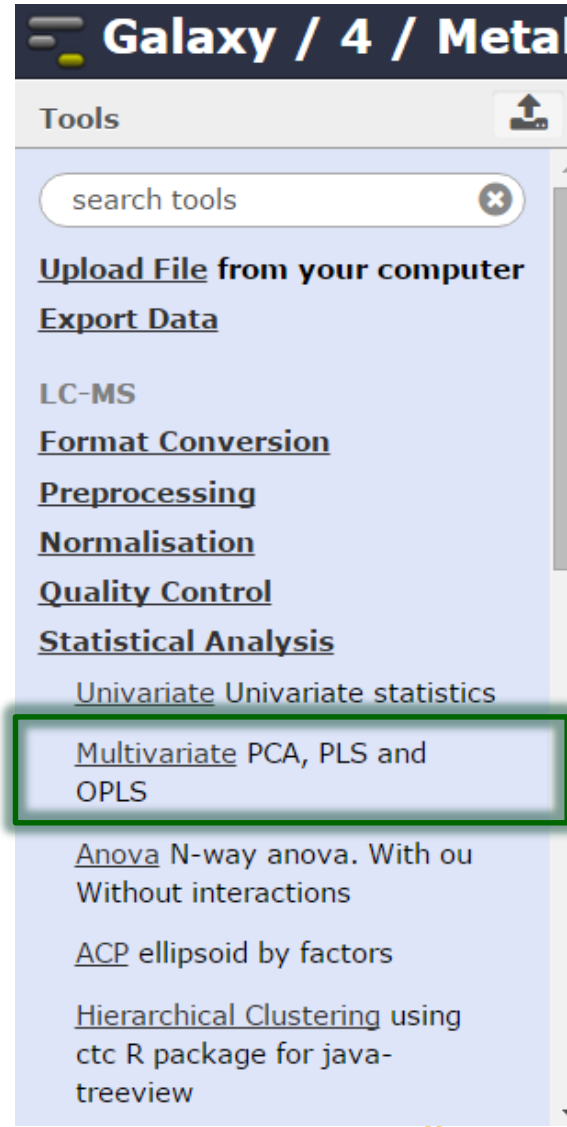
Selection of molecular signatures

The "Multivariate" module

Multivariate

- The "**Multivariate**" module allows you to perform:
  - Principal Component Analysis (**PCA**)
  - Partial Least-Squares regression (**PLS**) and discriminant analysis (**PLS-DA**)
  - Orthogonal Partial Least-Squares regression (**OPLS**) and discriminant analysis (**OPLS-DA**)

- It is available in the "Statistical Analysis" sections of LC-MS, GC-MS, and NMR



The "Multivariate" module

Multivariate

- The Multivariate module uses internally the *ropls* R module from bioconductor

http://bioconductor.org/packages/ropls

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

  - implements the original, NIPALS based, algorithms for PCA, PLS and OPLS
  - diagnostics to detect outliers, overfitting
  - graphics (scores, loadings, predictions)
  - feature selection (VIP, regression coefficients)

Thévenot E.A., Roux A., Xu Y., Ezan E. and Junot C. (2015). Analysis of the human adult urinary metabolome variations with age, body mass index and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. *Journal of Proteome Research,* **14:**3322-3335. http://dx.doi.org/10.1021/acs.jproteome.5b00354

Objectives

- Multivariate analysis:
  1. PCA [unsupervised]: Visualize the structure of the dataMatrix: $\mathbf{X}$
  2. (O)PLS(-DA) [supervised]: How can a factor of interest (response; column of sampleMetadata) be explained as a linear combination of **all** the variables (predictors) from dataMatrix: $\mathbf{y} = f(\mathbf{X})$
     a. when the response $\mathbf{y}$ is quantitative: (O)PLS regression
     b. when $\mathbf{y}$ is qualitative: (O)PLS(-DA) classification

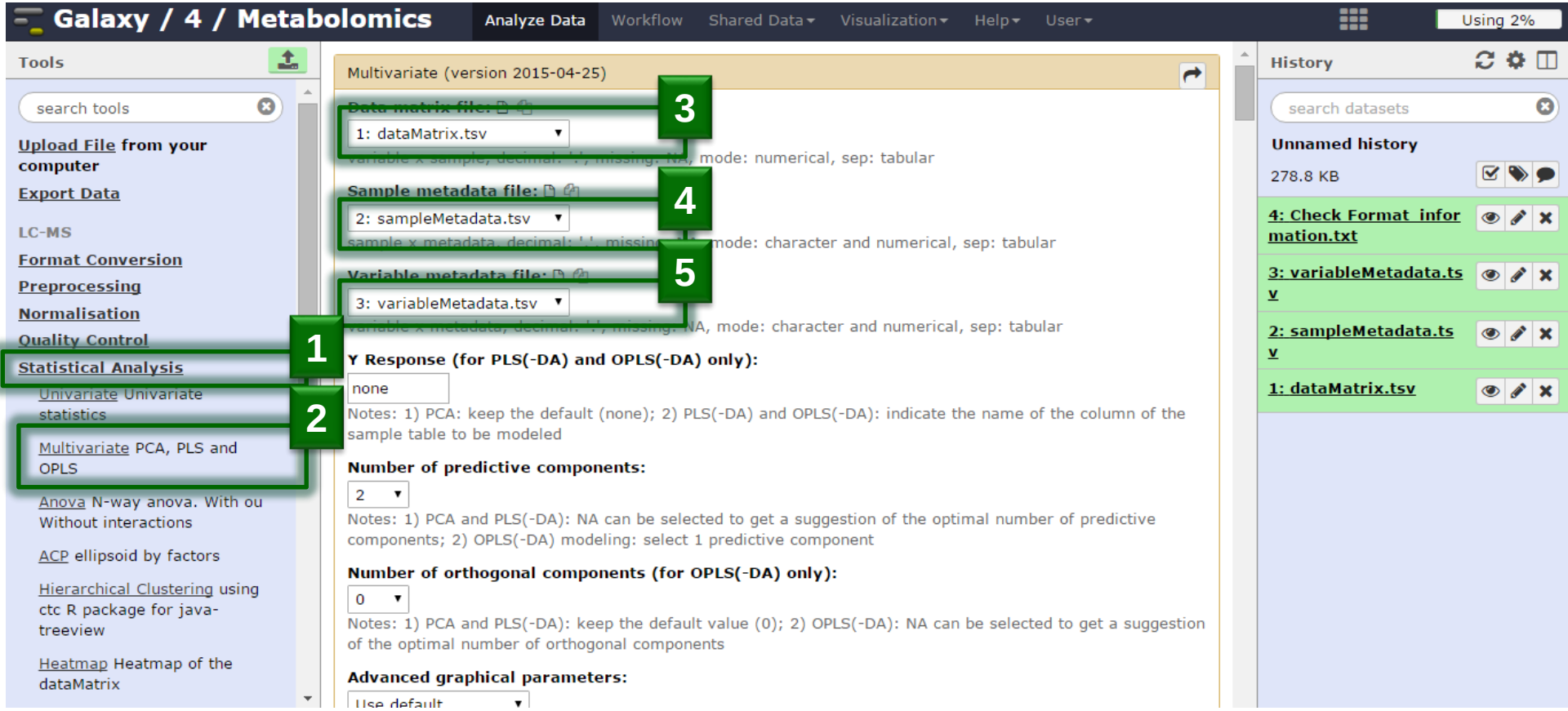
     Complementary to univariate analysis (where variables are tested independently)

Latent variable methods

- PCA and (O)PLS(-DA) are **latent variable** methods: new components are computed as linear combinations of the original variables
- The assumption is that a few components can efficiently represent the whole dataset (PCA) or model the factor of interest (O)PLS(-DA)
- **Other powerful multivariate methods** exist for regression and classification (Support Vector Machine, Random Forest, etc.) => biosigner module

Open the "Multivariate" module

- and select your 3 files of interest:



- you are now ready to start your multivariate analyzes!

PRINCIPAL COMPONENT ANALYSIS (PCA)

Objectives

- Visualize the dataMatrix
  - by selecting a few components which capture most of the spread (variance) of the cloud of samples
- Detect outliers
  - which may bias the computation of the component
- Detect clusters of samples
  - which may suggest an internal structuration of the data

Unsupervised analysis

***p* = 110 (quantitative) variables**

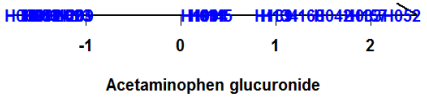
***n* = 183 samples**

	1,7-Dimethyluric acid	Dehydroepiandrosterone sulfate	Acetaminophen glucuronide	...
H011	2114	29025	44	
H023	43274	639	2	
H033	22386	325	1933	
H042	8185	13938	933	
H052	22385	357	5004	
H062	6380	292	1	
H073	10012	22781	1	
H083	30414	105	1	
H092	6637	35156	1	
H103	12100	2	1	
H114	33362	149041	46	
H124	11197	84536	1	
H134	18698	34053	254	
H145	14435	212398	52	
H157	31732	19317	2200	
H168	10221	78	475	
H180	22936	463	1	
H189	14423	1039	220	
H199	2888	12272	37	
H209	12563	100236	2	

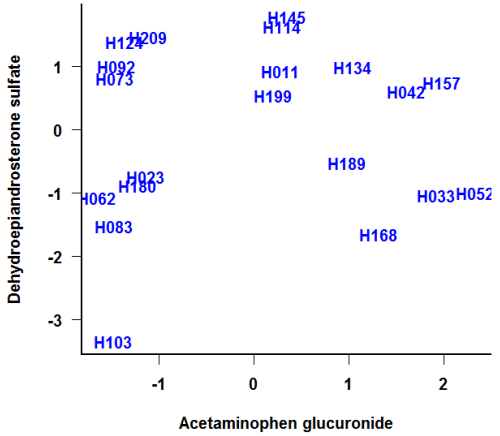
X

How to visualize multivariate observations?

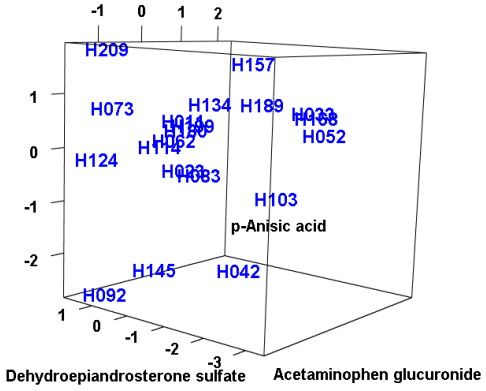
1 variable



2 variables



3 variables

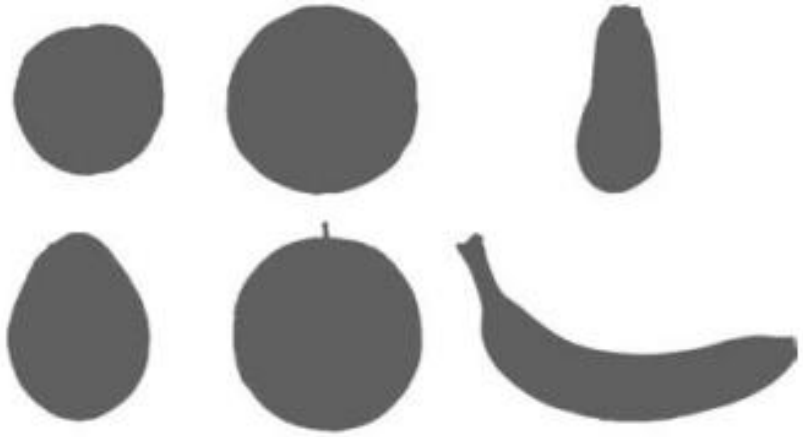


***p* variables**

=> Dimension reduction

Projection

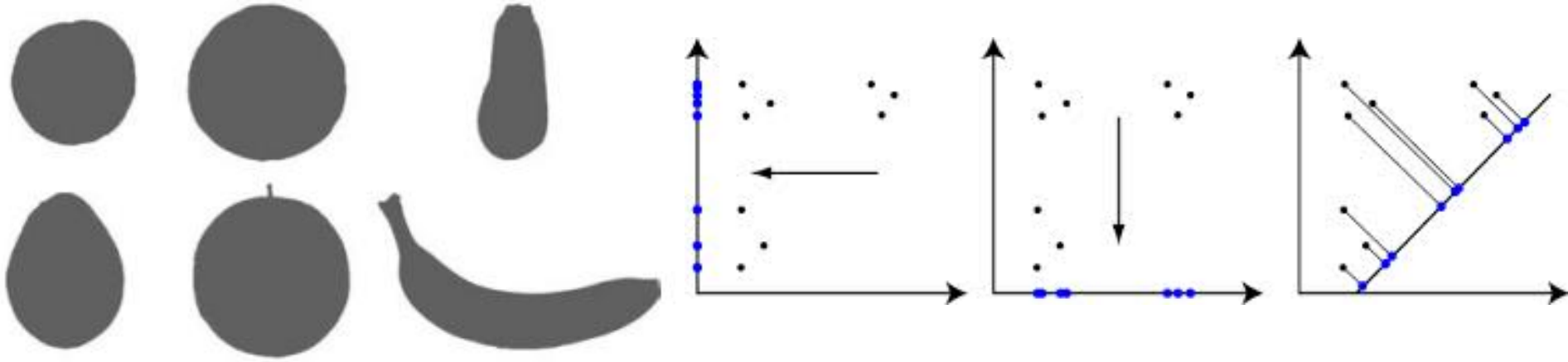
- Projected distances as high as possible



Husson and Pages (2011). Exploratory multivariate analysis by example using R. Chapman & Hall/CRC

Projection on latent variables

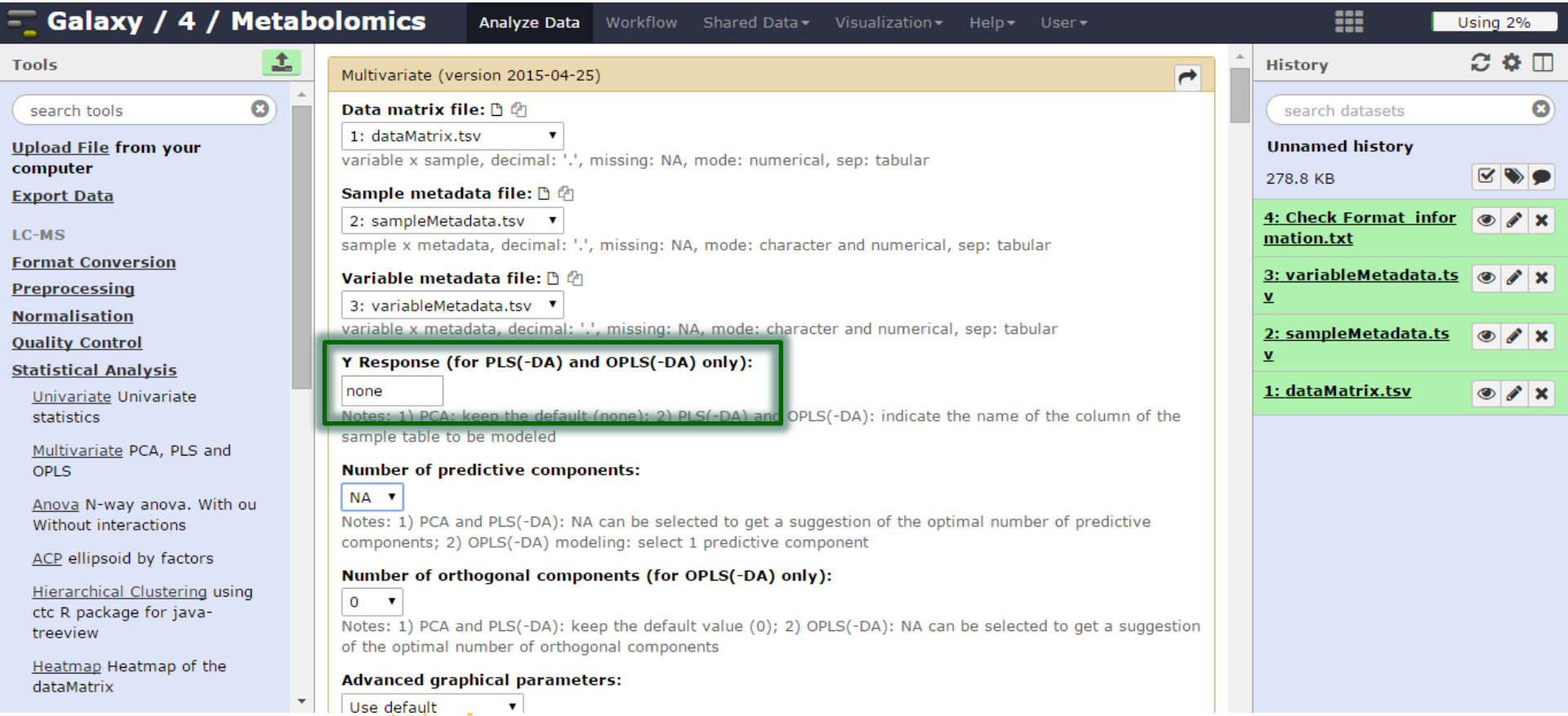
- Projected distances as high as possible
- Define new variables as linear combination of original ones



Husson and Pages (2011). Exploratory multivariate analysis by example using R. Chapman & Hall/CRC

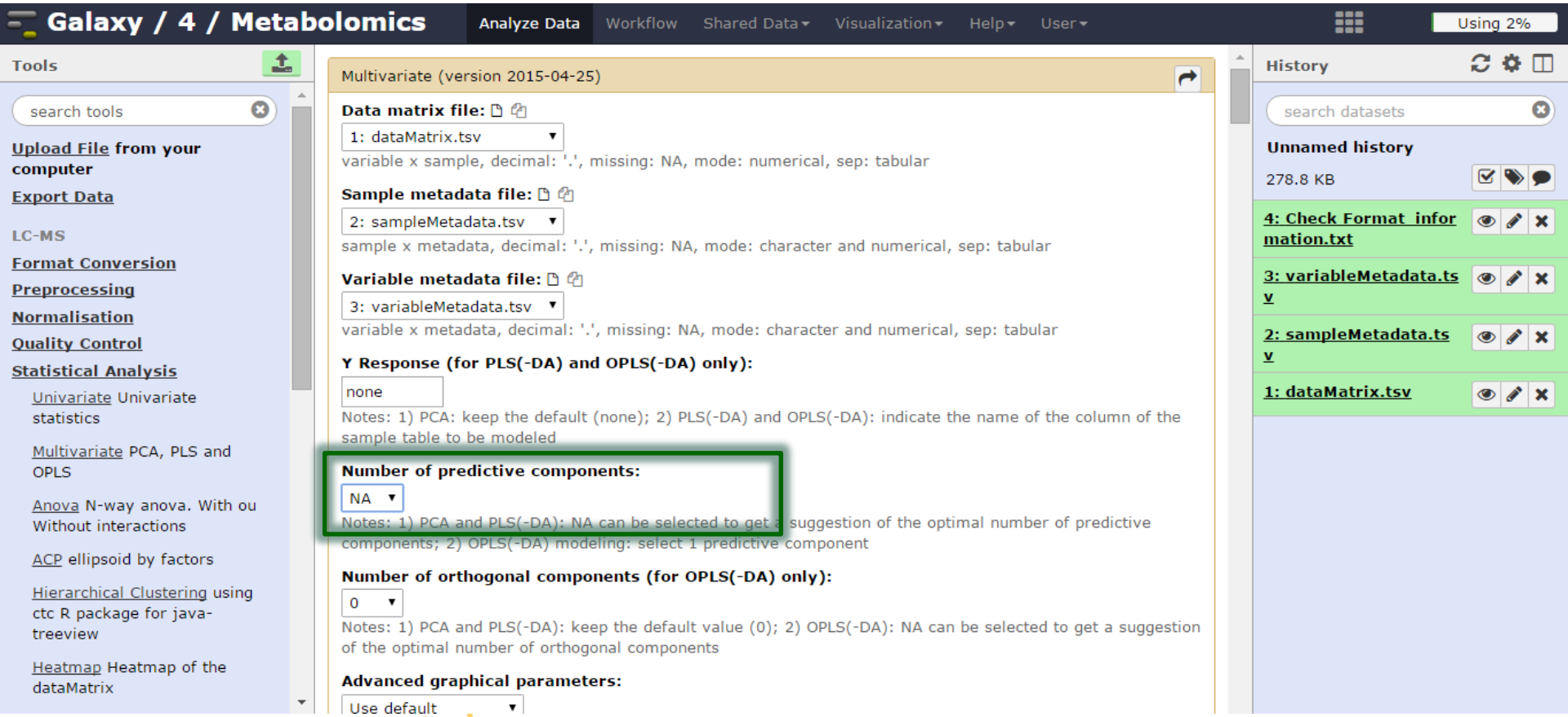
Selection of PCA as the type of analysis

- Keep the "Y response" to 'none' for PCA (unsupervised analysis)



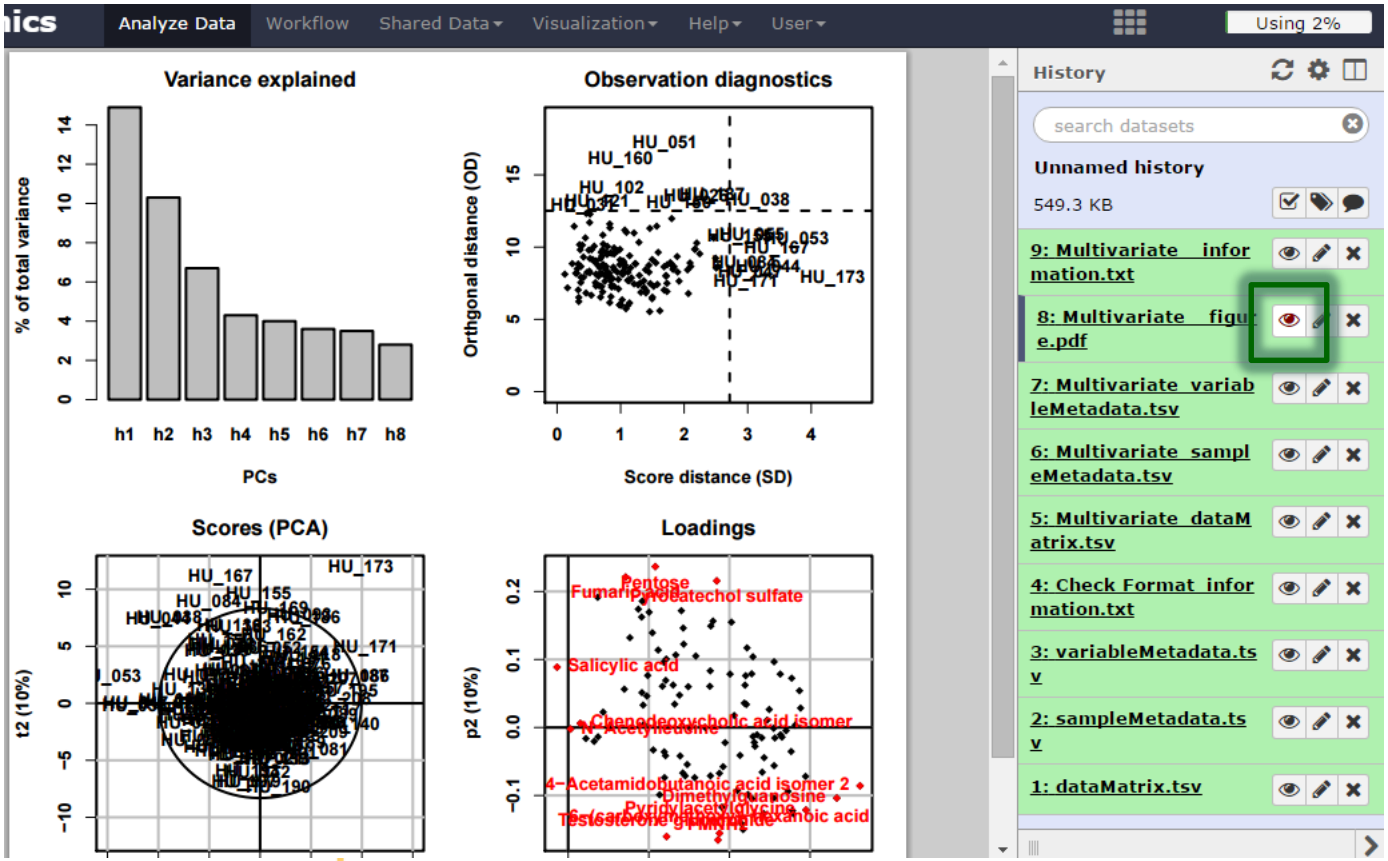
Automatic selection of the number of components

- Until the variance is less than the mean variance of all components



Graphical results

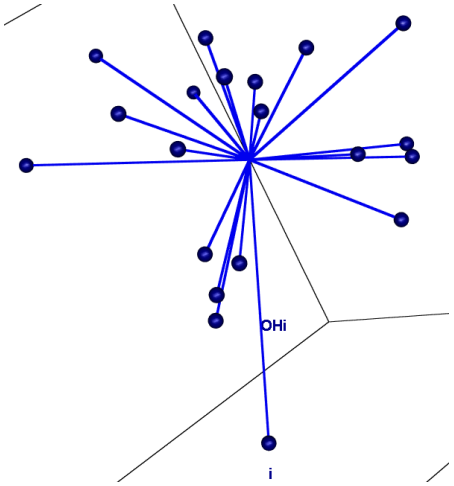
- scree plot, outliers, and the loading and score plots



Diagnostics R2X: How much of the original inertia is still reflected by the model?

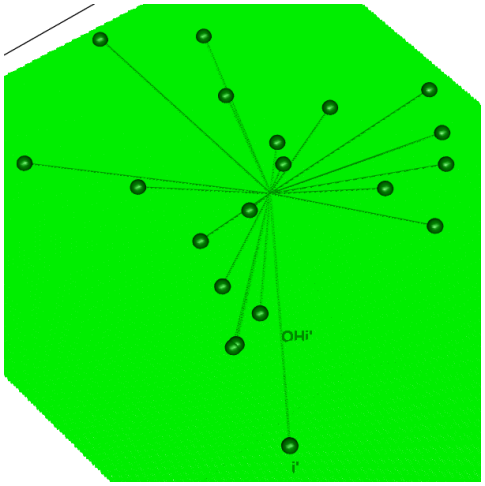
Total

$$TSS = \sum_{i=1}^{n} OH_i{}^2$$



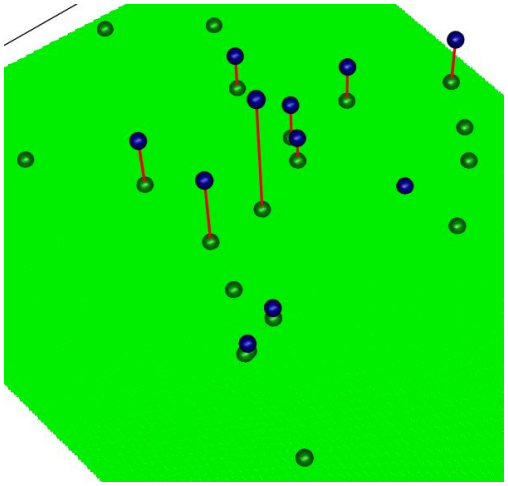
Explained

$$ESS = \sum_{i=1}^{n} OH'_i{}^2$$



Residual

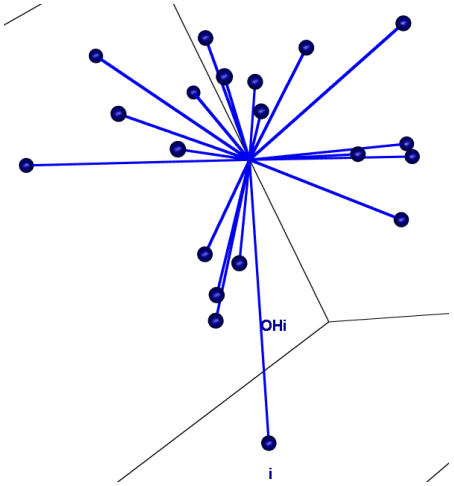
$$RSS = \sum_{i=1}^{n} HH'_i{}^2 = TSS - ESS$$



Diagnostics R2X: How much of the original inertia is still reflected by the model?

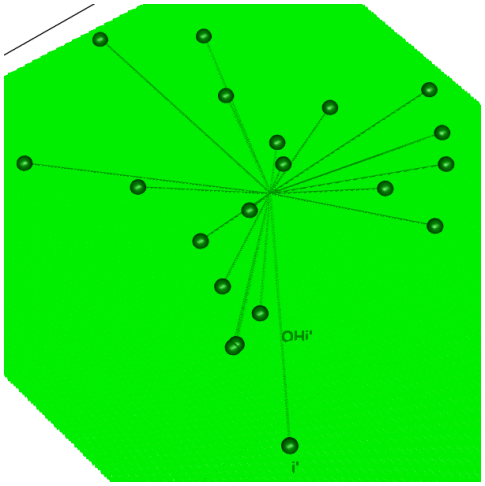
Total

$$TSS = \sum_{i=1}^{n} OH_i{}^2$$



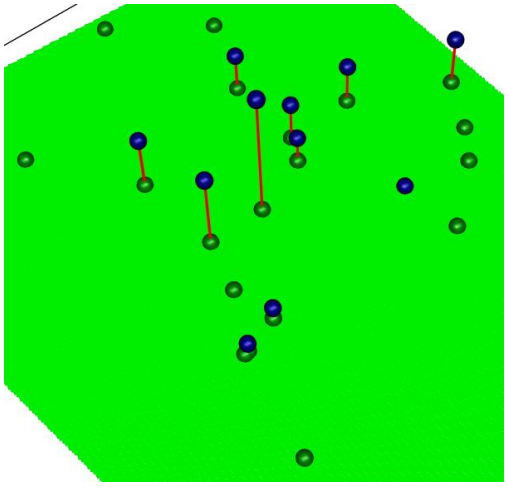
Explained

$$ESS = \sum_{i=1}^{n} OH'_i{}^2$$



Residual

$$RSS = \sum_{i=1}^{n} HH'_i{}^2 = TSS - ESS$$

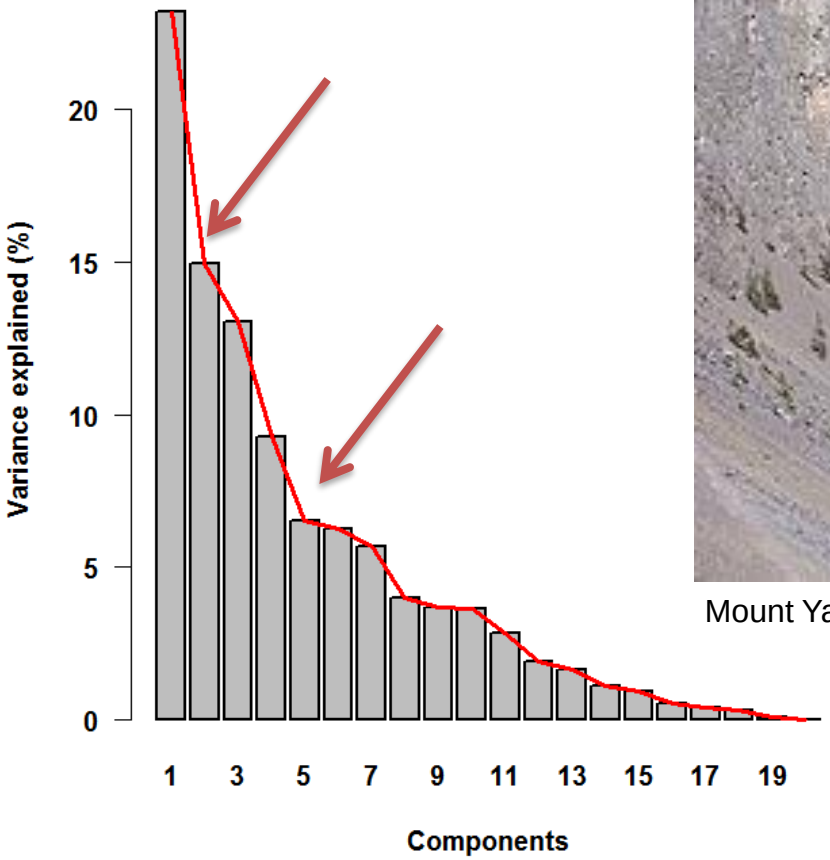


$$R2X = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \qquad 0 \leq R2X \leq 1$$

- R2X increases with the number of components in the model
- For a given number of components, the higher the R2X, the more inertia is captured by the model (projection)

Scree plot

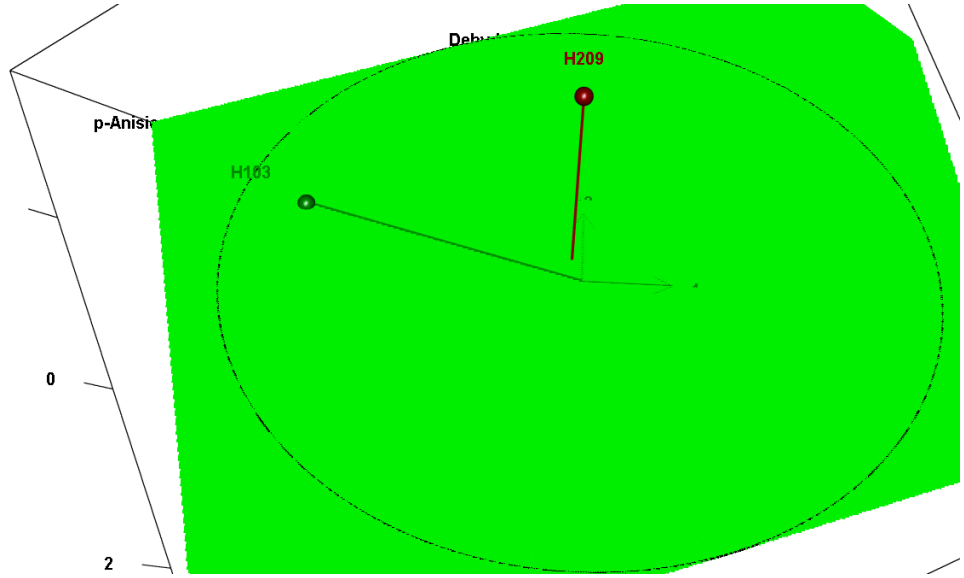
- Check that the first components capture most of the variance

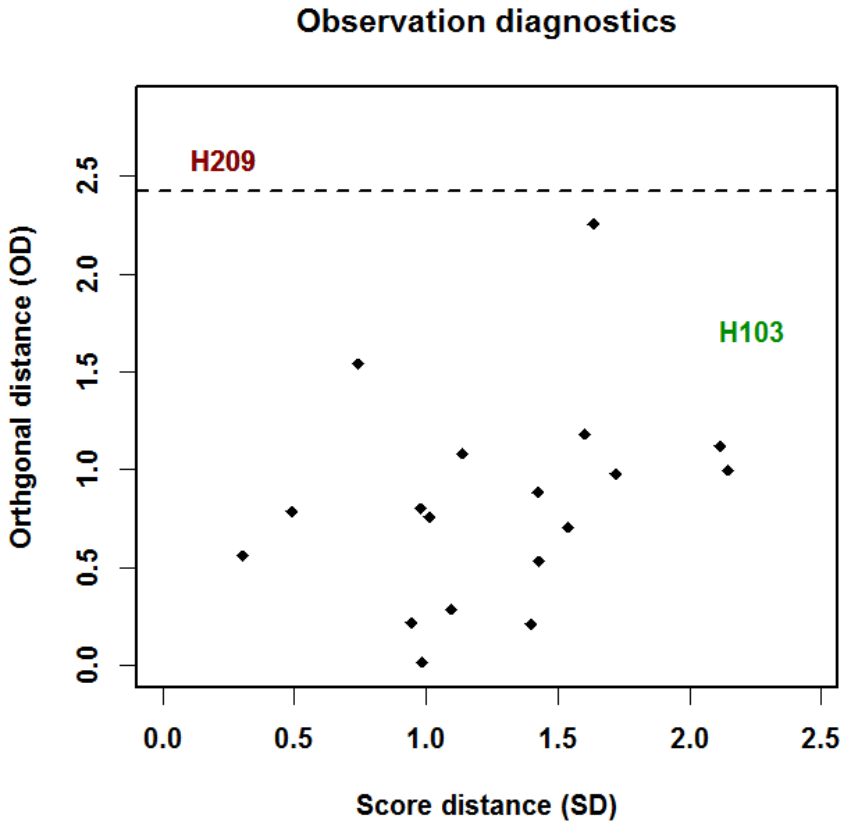


Mount Yamnuska, Alberta. Wikipedia

Observation diagnostics

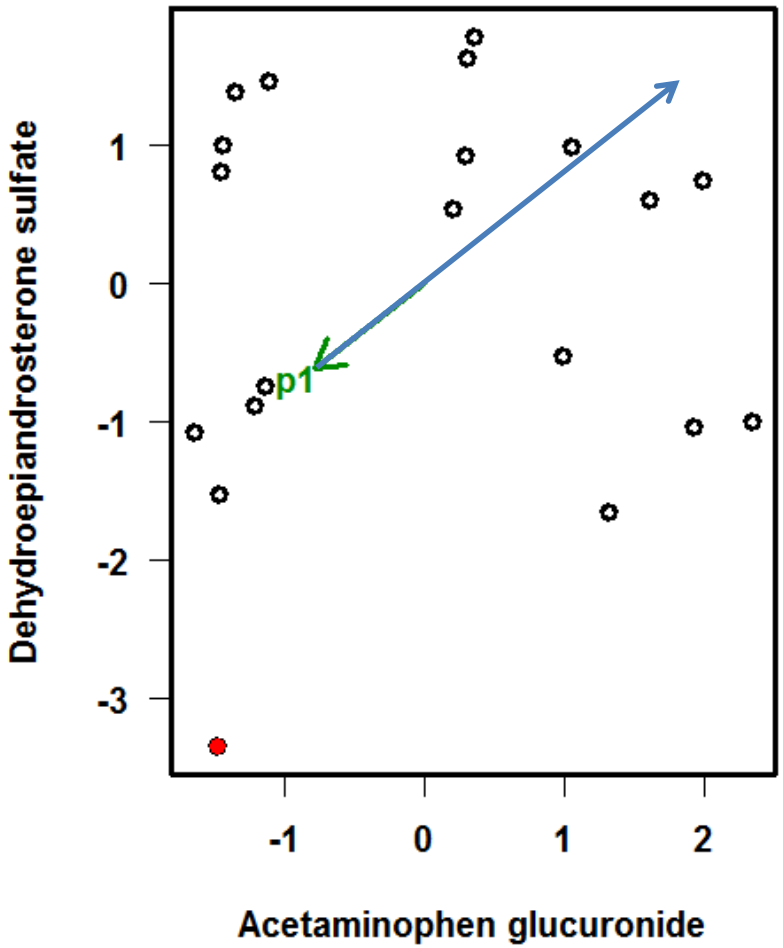
- Samples which may bias the PCA computation and/or may not be faithfully visualized by the score plot

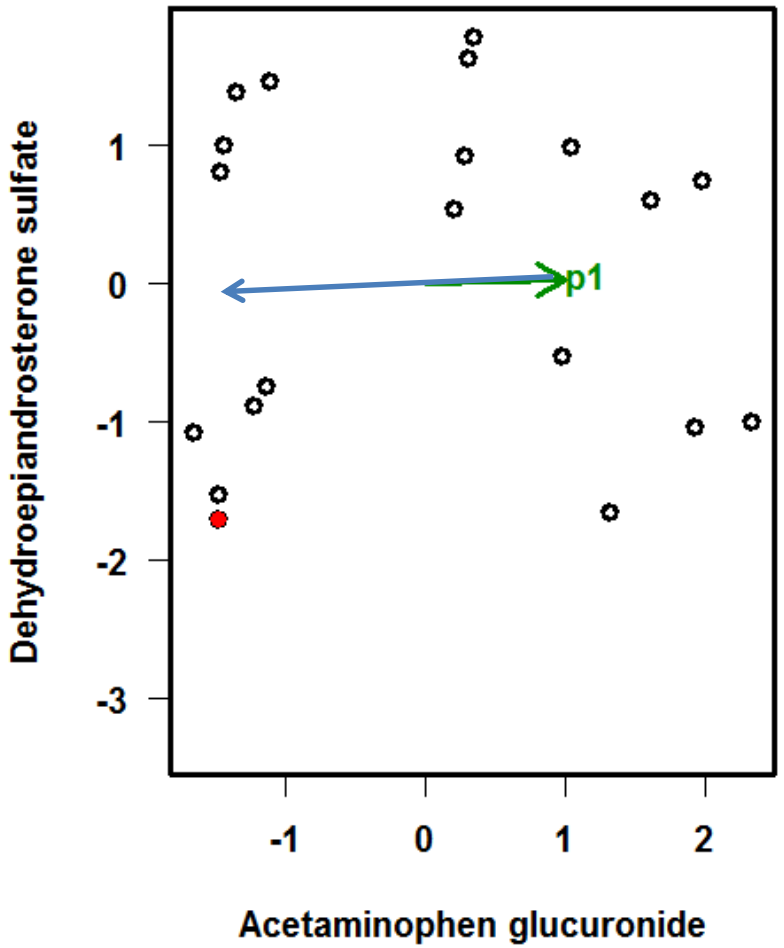




Hubert M., Rousseeuw P. and Vanden Branden K. (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics,* **47:**64-79. DOI: 10.1198/004017004000000563

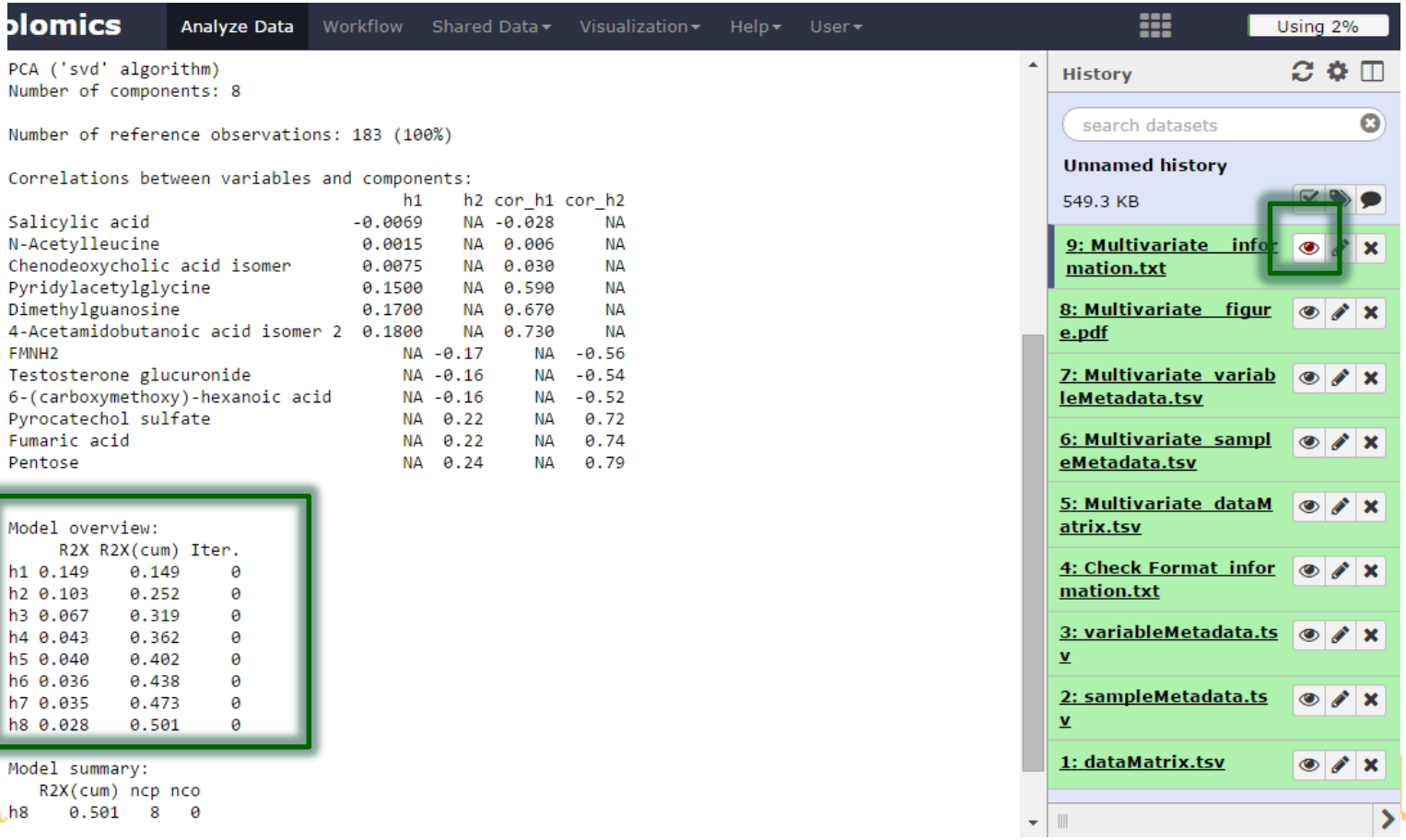
Sensitivity to outliers





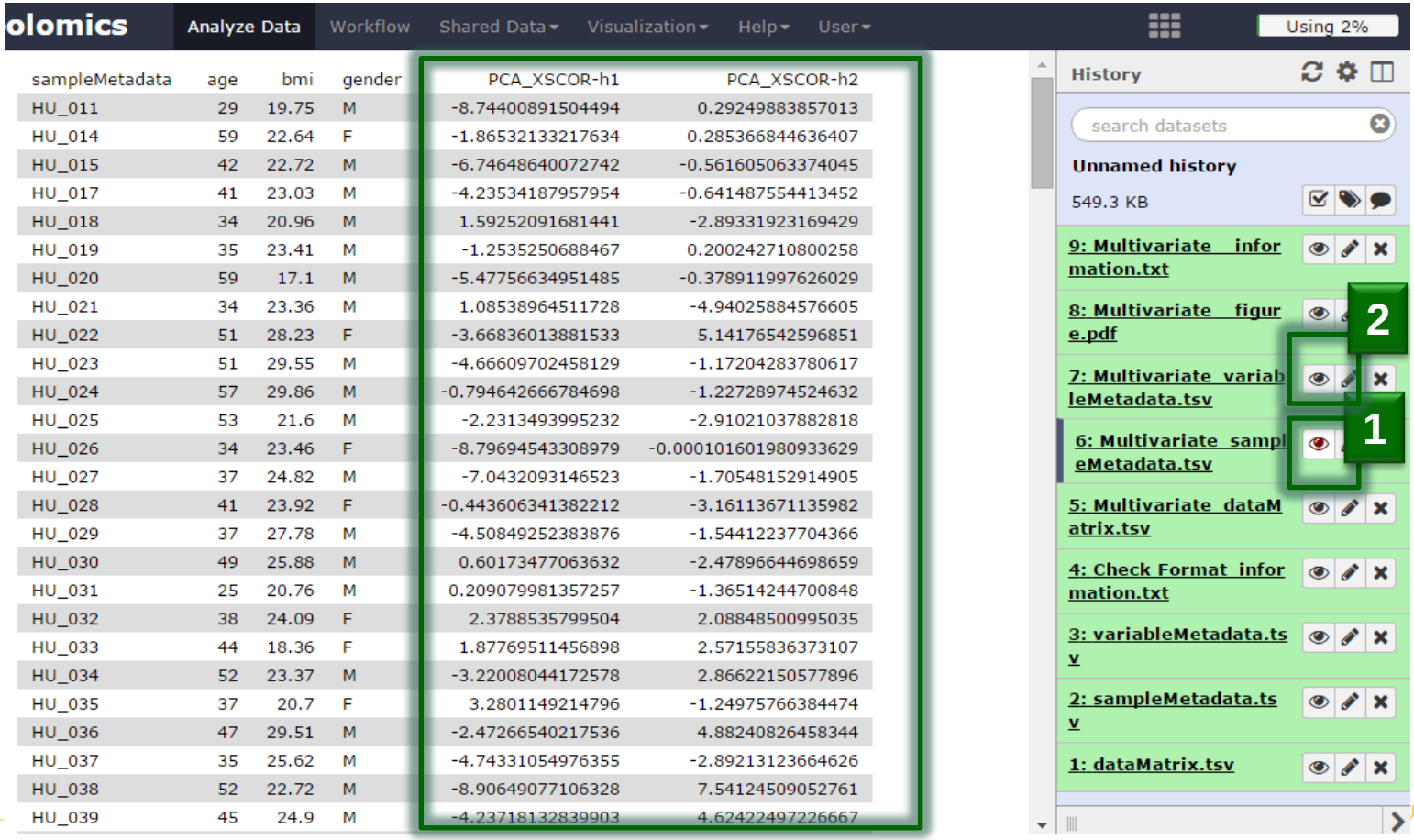
Numerical results

- Numerical results (including the percentage of explained inertia) can be viewed in the "information.txt" file



Score and loading values

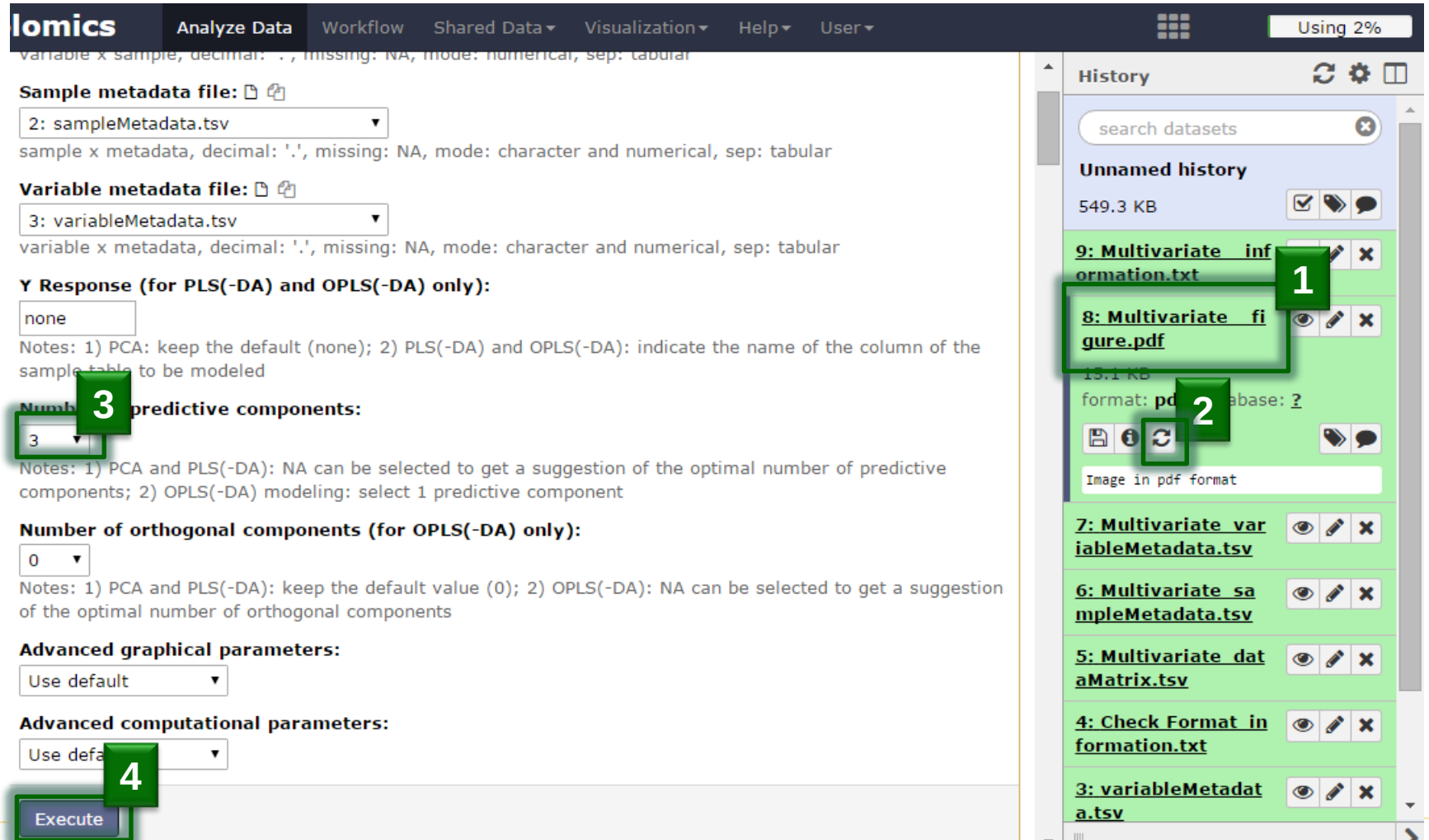
- The score (resp. loading) values of the selected components have been added as columns in the sampleMetadata (resp. variableMetadata) files



sampleMetadata	age	bmi	gender	PCA_XSCOR-h1	PCA_XSCOR-h2
HU_011	29	19.75	M	-8.74400891504494	0.29249883857013
HU_014	59	22.64	F	-1.86532133217634	0.285366844636407
HU_015	42	22.72	M	-6.74648640072742	-0.561605063374045
HU_017	41	23.03	M	-4.23534187957954	-0.641487554413452
HU_018	34	20.96	M	1.59252091681441	-2.89331923169429
HU_019	35	23.41	M	-1.2535250688467	0.200242710800258
HU_020	59	17.1	M	-5.47756634951485	-0.378911997626029
HU_021	34	23.36	M	1.08538964511728	-4.94025884576605
HU_022	51	28.23	F	-3.66836013881533	5.14176542596851
HU_023	51	29.55	M	-4.66609702458129	-1.17204283780617
HU_024	57	29.86	M	-0.794642666784698	-1.22728974524632
HU_025	53	21.6	M	-2.2313493995232	-2.91021037882818
HU_026	34	23.46	F	-8.79694543308979	-0.000101601980933629
HU_027	37	24.82	M	-7.0432093146523	-1.70548152914905
HU_028	41	23.92	F	-0.443606341382212	-3.16113671135982
HU_029	37	27.78	M	-4.50849252383876	-1.54412237704366
HU_030	49	25.88	M	0.60173477063632	-2.47896644698659
HU_031	25	20.76	M	0.209079981357257	-1.36514244700848
HU_032	38	24.09	F	2.3788535799504	2.08848500995035
HU_033	44	18.36	F	1.87769511456898	2.57155836373107
HU_034	52	23.37	M	-3.22008044172578	2.86622150577896
HU_035	37	20.7	F	3.2801149214796	-1.24975766384474
HU_036	47	29.51	M	-2.47266540217536	4.88240826458344
HU_037	35	25.62	M	-4.74331054976355	-2.89213123664626
HU_038	52	22.72	M	-8.90649077106328	7.54124509052761
HU_039	45	24.9	M	-4.23718132839903	4.62422497226667

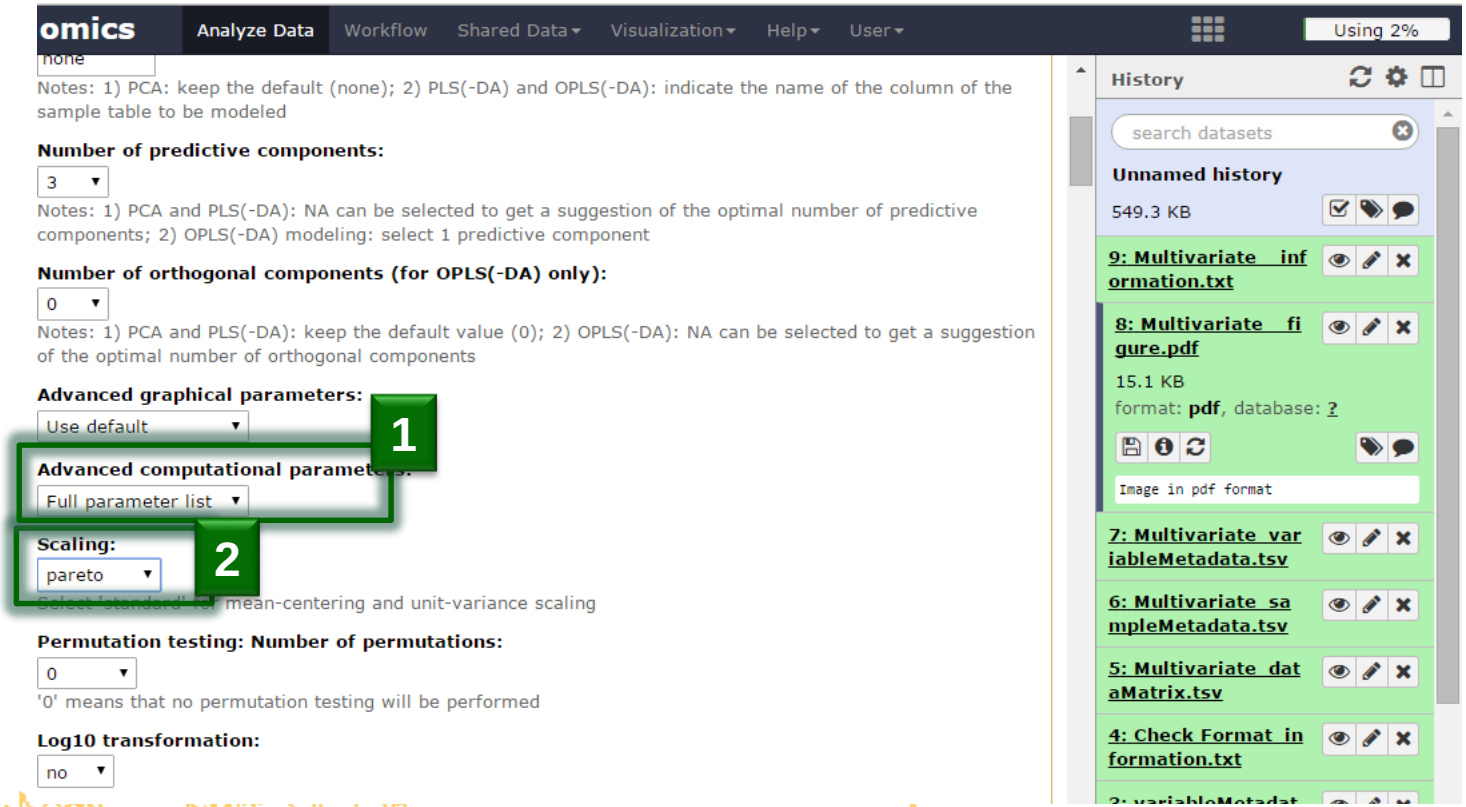
Tuning the parameters

- You can recall the page with your parameters, modify them, and restart the analysis



Advanced parameters: Scaling

- Variables are mean-centered for PCA
- By default, they are also unit-variance scaled
  - absence of variance scaling or changing to Pareto scaling can be selected in the advanced computational parameters



Advanced parameters: Ellipses

- Indicate the column name of sampleMetadata to be used

olomics | Analyze Data | Workflow | Shared Data | Visualization | Help | User

Variable metadata file:

3: variableMetadata.tsv

variable x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular

Y Response (for PLS(-DA) and OPLS(-DA) only):

none

Notes: 1) PCA: keep the default (none); 2) PLS(-DA) and OPLS(-DA): indicate the name of the column of the sample table to be modeled

Number of predictive components:

3

Notes: 1) PCA and PLS(-DA): NA can be selected to get a suggestion of the optimal number of predictive components; 2) OPLS(-DA) modeling: select 1 predictive component

Number of orthogonal components (for OPLS(-DA) only):

0

Notes: 1) PCA and PLS(-DA): keep th lt value (0); 2) OPLS(-DA): NA can be selected to get a suggestion of the optimal number of orthogonal ents

Advanced graphical parameters: 1

Full parameter list

Graphic:

x-score

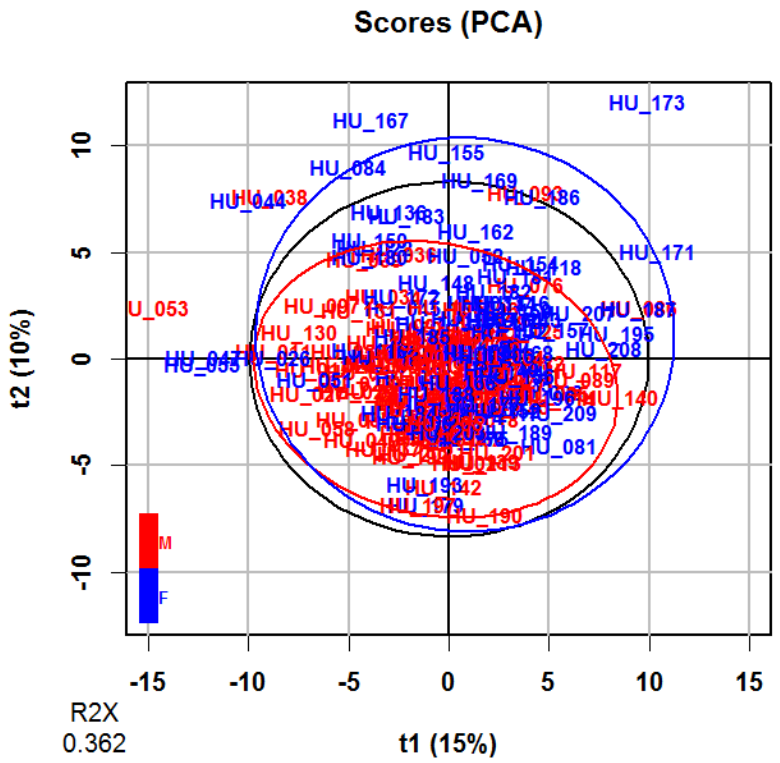
Mahalanobis ellipses: 2

gender

data column with the classes to be used for drawing ellipses (for (O)PLS-DA: indicate the same name as the 'Response' argument above); If you do not want ellipses, keep the default, none

Sample colors:

none



References

- Husson F., Le S. and Pages J. (2011). Exploratory multivariate analysis by example using R. *Chapman & Hall/CRC*

- Ringner M. (2008). What is principal component analysis? *Nature Biotechnology*, **26**:303-304. http://dx.doi.org/10.1038/nbt0308-303

- Baccini A. (2010). Statistique descriptive multidimensionnelle (pour les nuls). www.math.univ-toulouse.fr/~baccini/zpedago/asdm.pdf

PARTIAL LEAST SQUARES REGRESSION (PLS) AND DISCRIMINANT ANALYSIS (PLS-DA)

PLS(-DA) modelling

- Powerful regression method when

$$n_{samples} < p_{variables}$$

- **Complementary to univariate hypothesis testing** (where variables are tested independantly)

- **Risk of overfitting**: i.e., building a model whose (apparently) good performances result from chance only

Supervised analysis (i.e. with labels)

1 response

p = 110 (quantitative) variables

n = 183 samples

	bmi
H011	19.8
H023	29.6
H033	18.4
H042	19.8
H052	20.1
H062	22.2
H073	25.4
H083	29.8
H092	21.8
H103	26.8
H114	29.4
H124	22.2
H134	22.9
H145	29.1
H157	22.0
H168	20.8
H180	23.7
H189	19.4
H199	21.0
H209	21.5

y

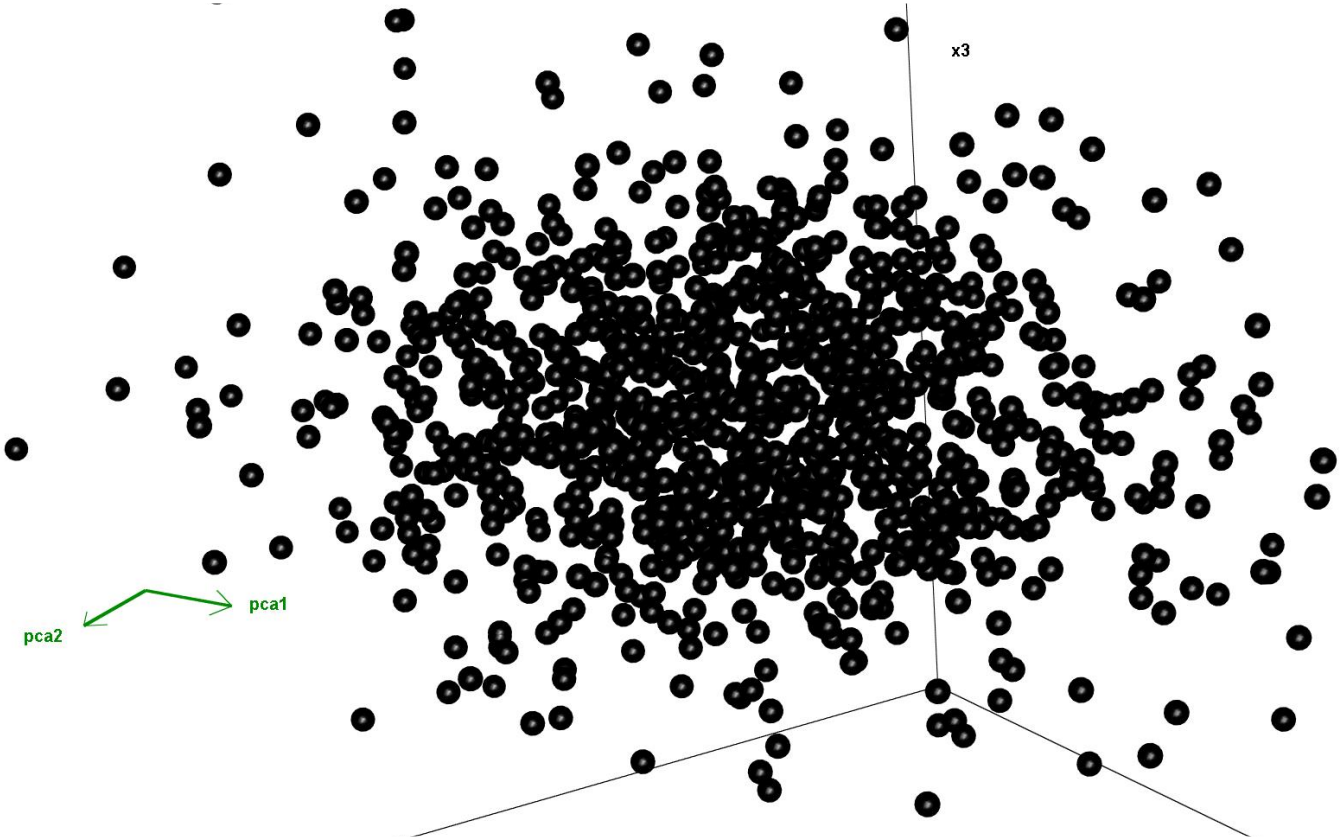
	1,7-Dimethyluric acid	Dehydroepiandrosterone sulfate
H011	3.33	4.46
H023	4.64	2.81
H033	4.35	2.51
H042	3.91	4.14
H052	4.35	2.55
H062	3.80	2.47
H073	4.00	4.36
H083	4.48	2.02
H092	3.82	4.55
H103	4.08	0.21
H114	4.52	5.17
H124	4.05	4.93
H134	4.27	4.53
H145	4.16	5.33
H157	4.50	4.29
H168	4.01	1.89
H180	4.36	2.67
H189	4.16	3.02
H199	3.46	4.09
H209	4.10	5.00

...

X

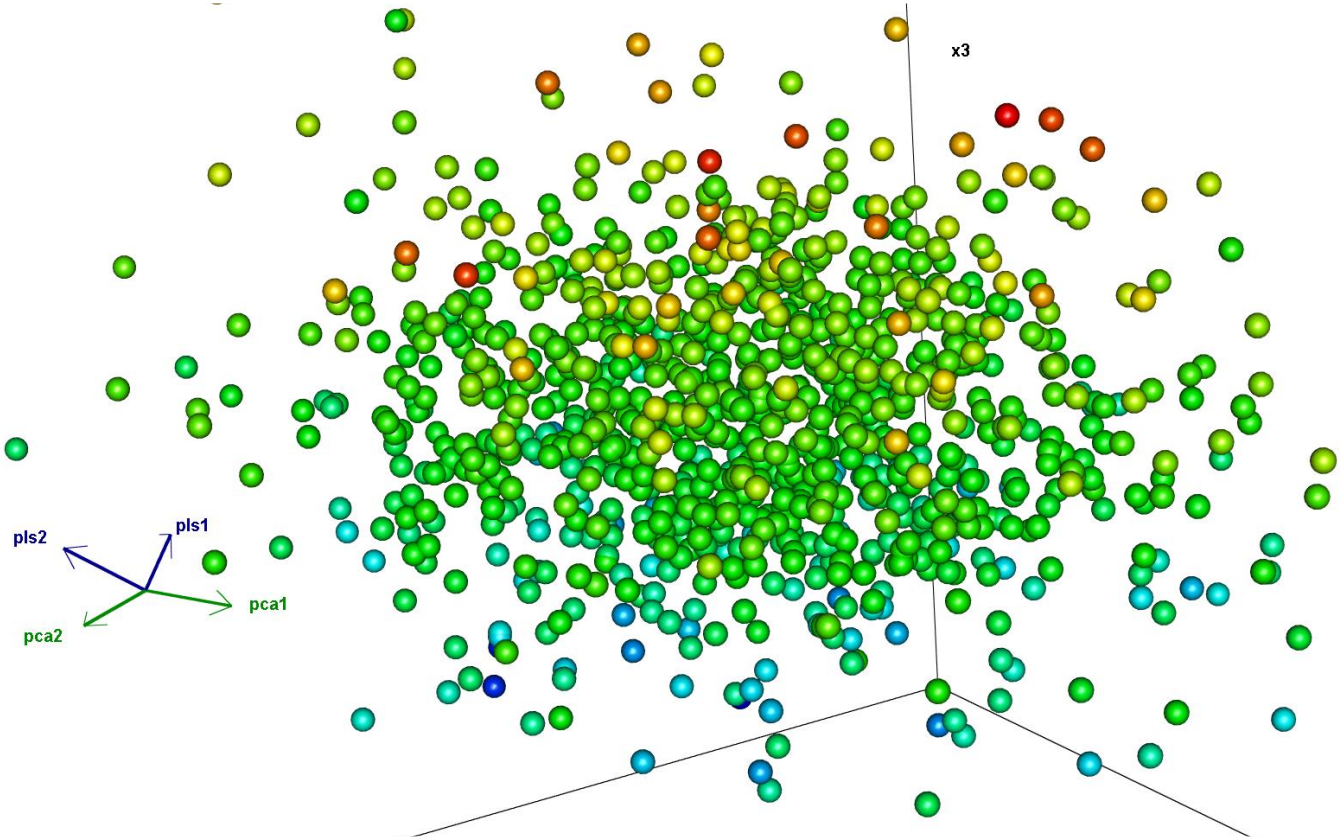
PLS vs PCA

- PCA finds the directions of maximum variance



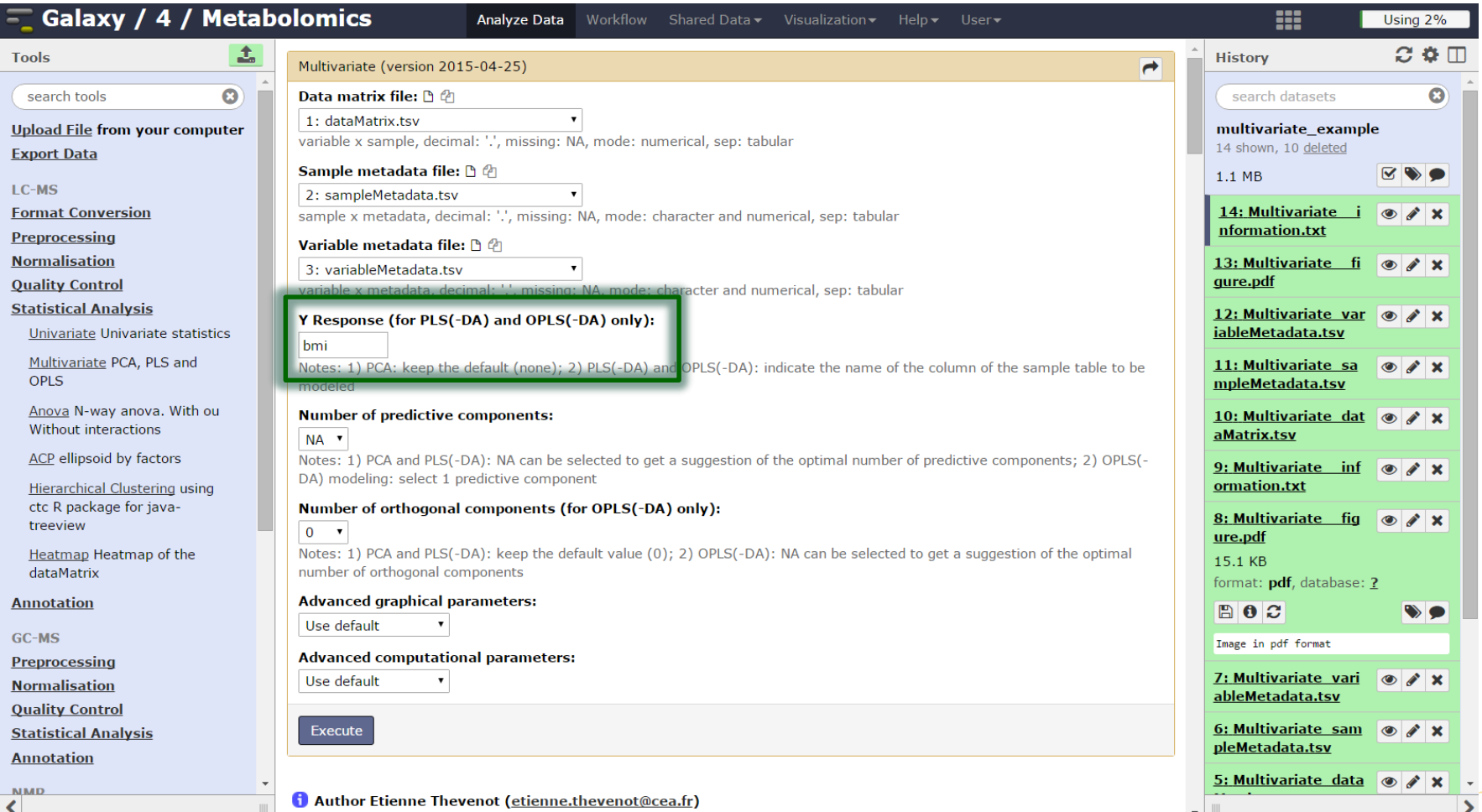
PLS vs PCA

- PLS includes the labels into the model



Selection of PLS(-DA) as the type of analysis

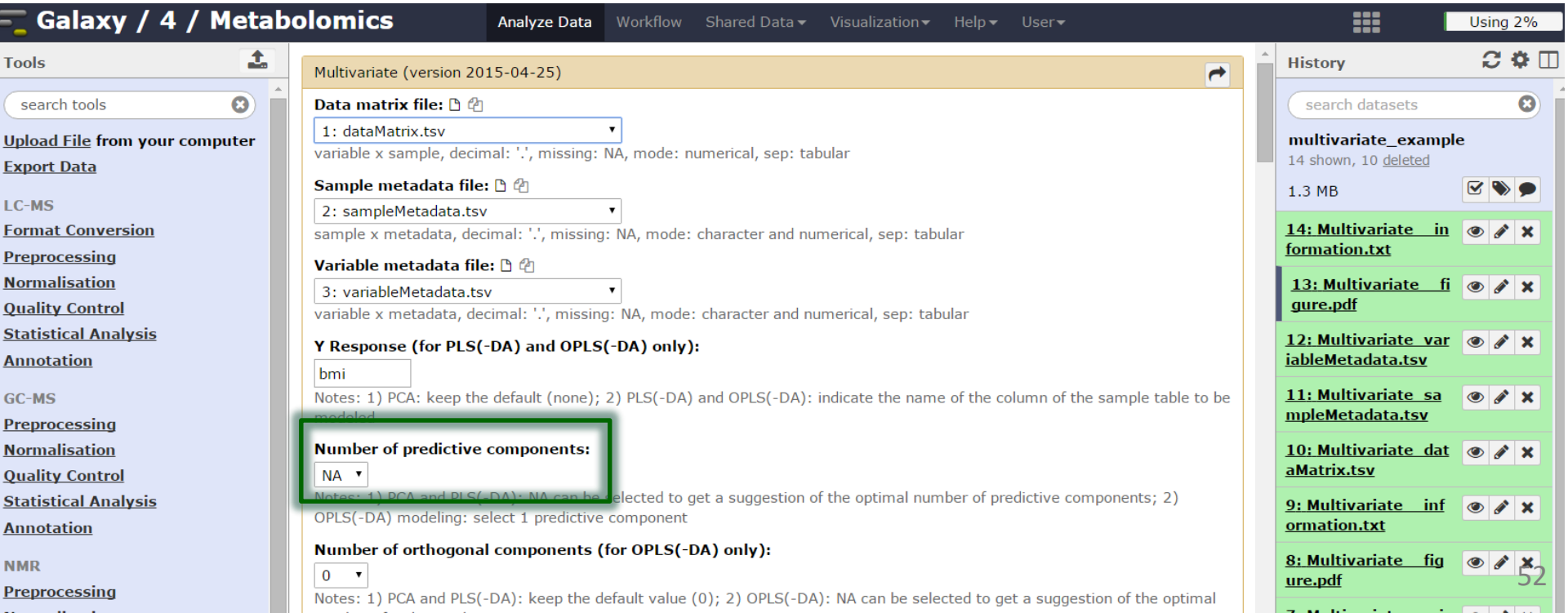
- Select the "Y response" to be modelled (column of sampleMetadata):
  - column of numbers (age, bmi): **PLS** regression
  - column of characters ('M'/'F', 'patient'/'control'): **PLS-DA** classification



Automatic selection of the number of components

- A new component *h* is added to the model if:
  - $R2Y_h \geq 1\%$
  - $Q2Y_h \geq 0$ $\left(\text{or } 5\% \text{ if } n_{samples} \leq 100\right)$

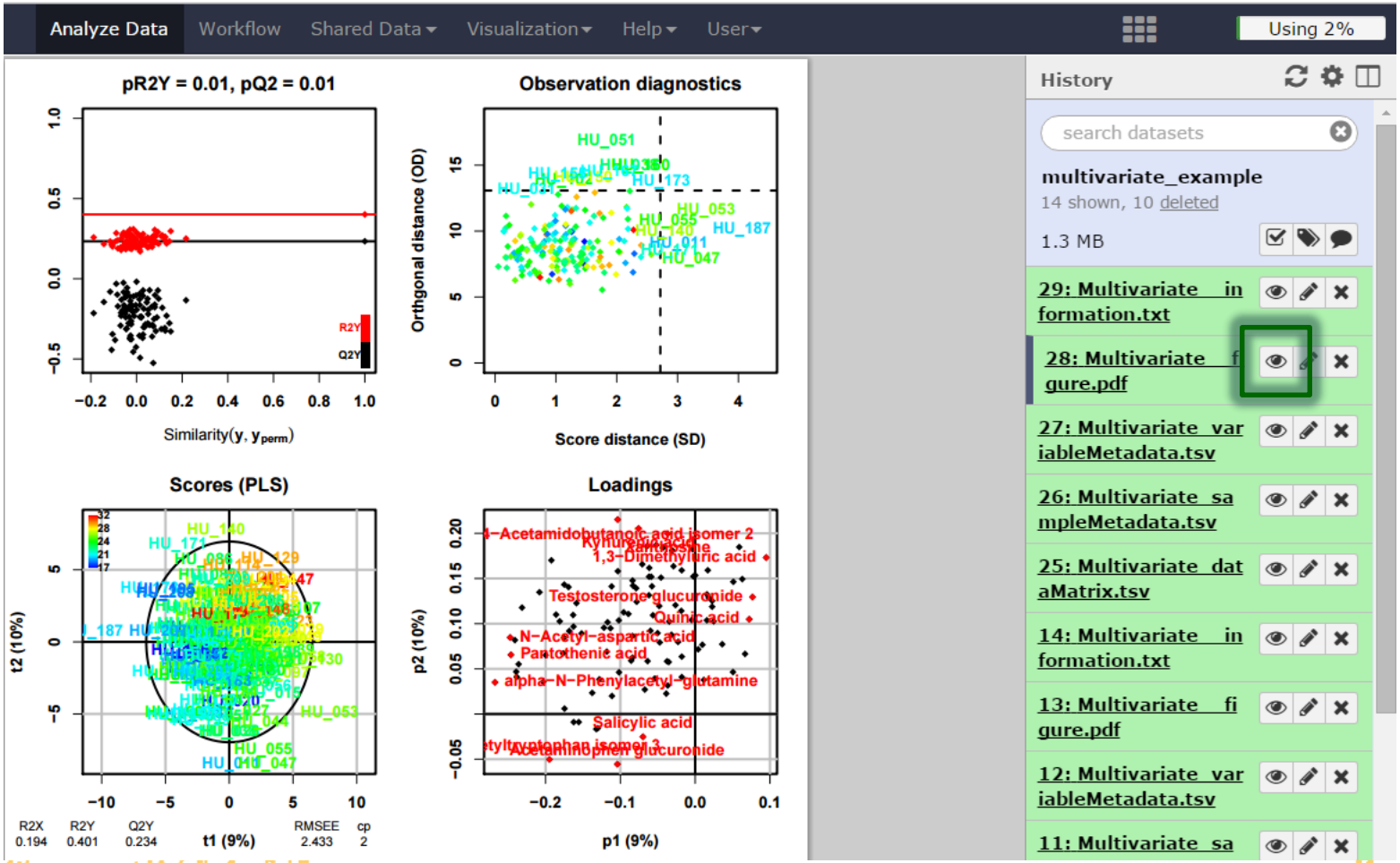
Note: $Q2Y_h = 1 - \frac{PRESS_h}{RSS_{h-1}}$ where $PRESS_h$ is estimated by cross-validation



Graphical results

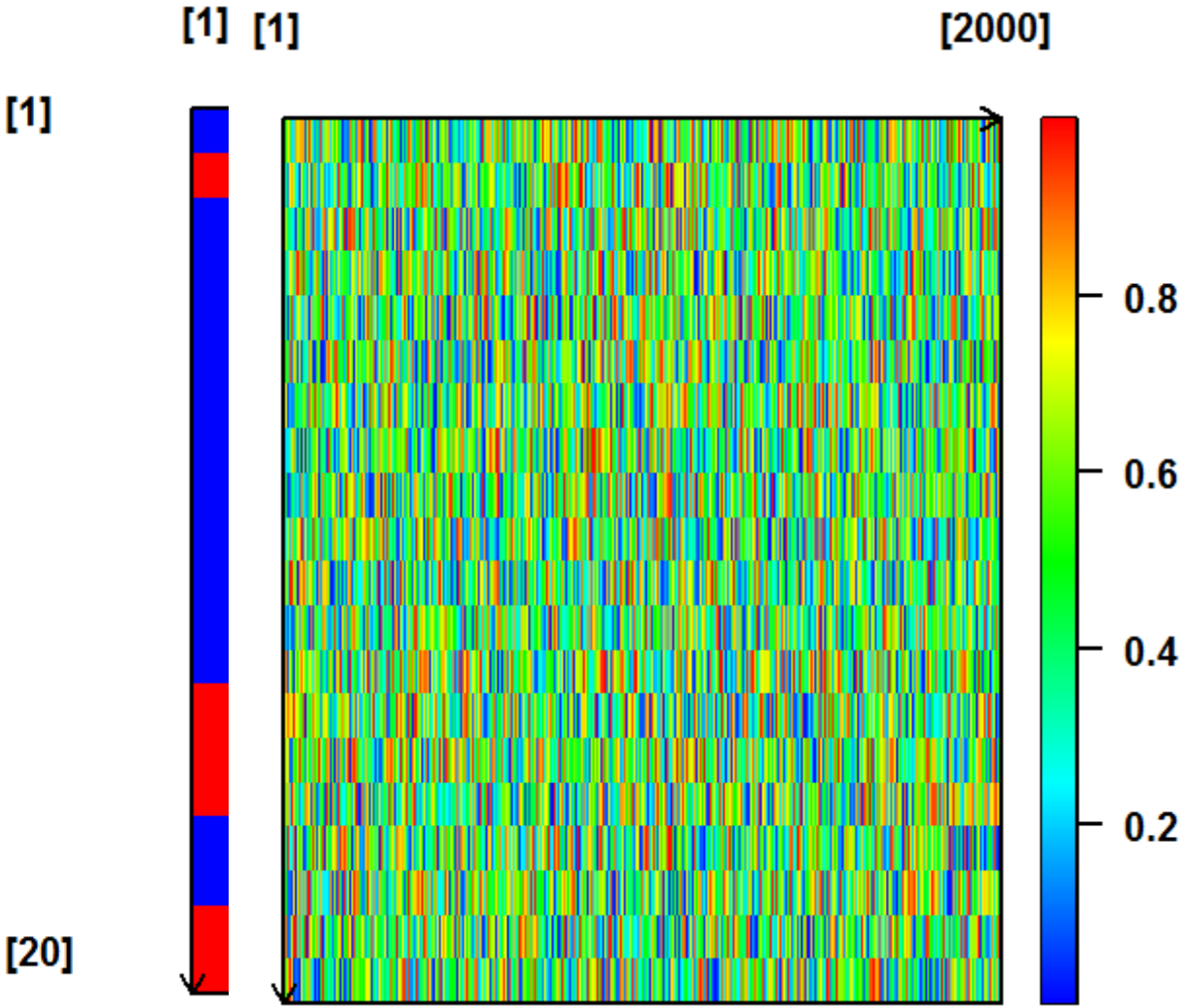
- permutation, overview, outlier, and score plots displayed as the default ('summary')

'summary'



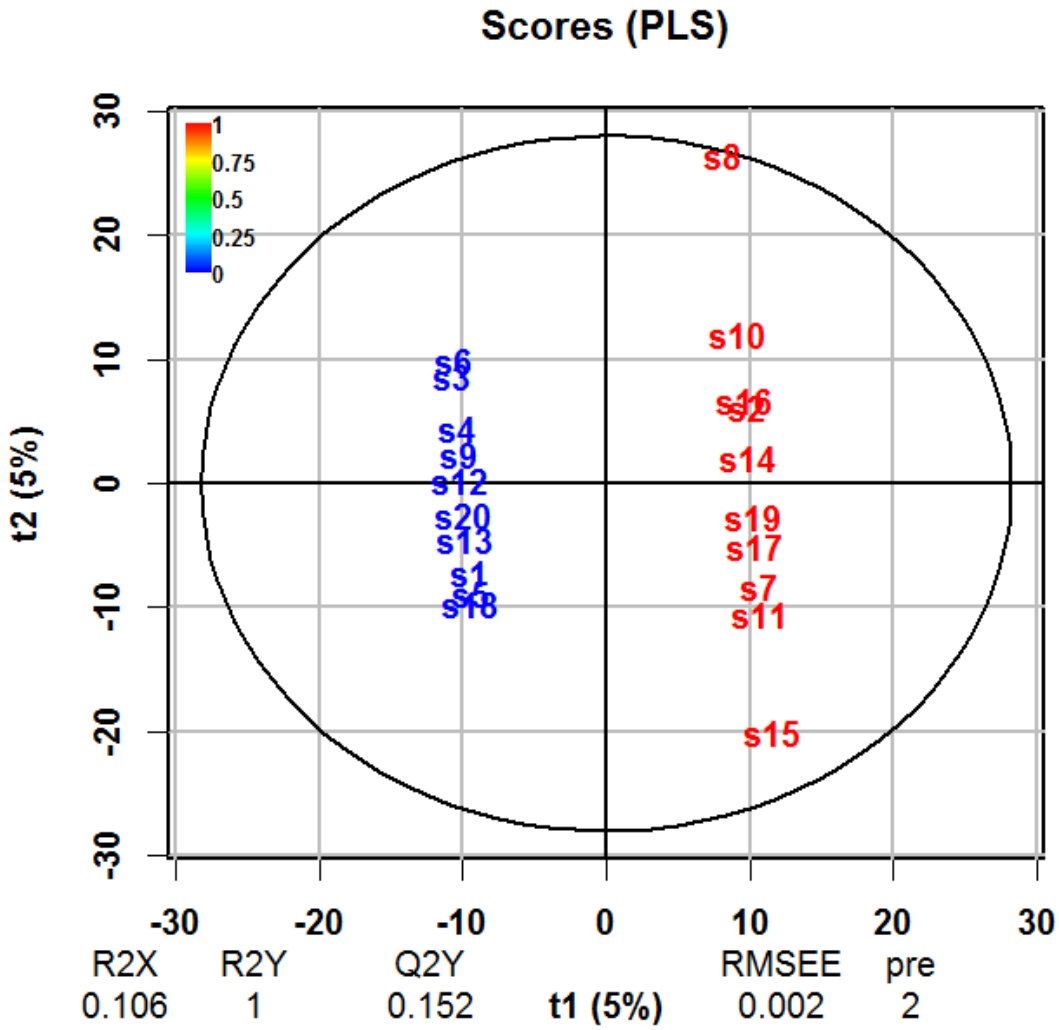
Overfitting

- X: 20 × 2,000 matrix of **random** numbers
  - Uniform distribution between 0 and 1
- Y: 20 x 1 matrix of **random** labels
  - 0 or 1 values



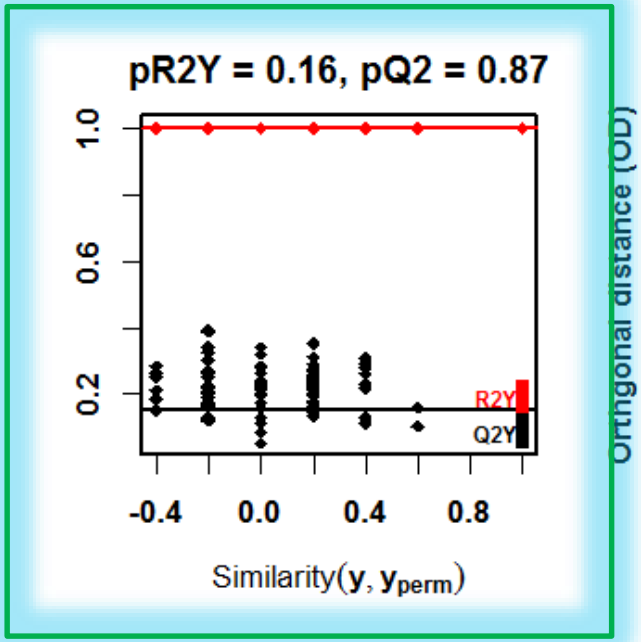
adpated from Wehrens (2011). Chemometrics with R. Springer.

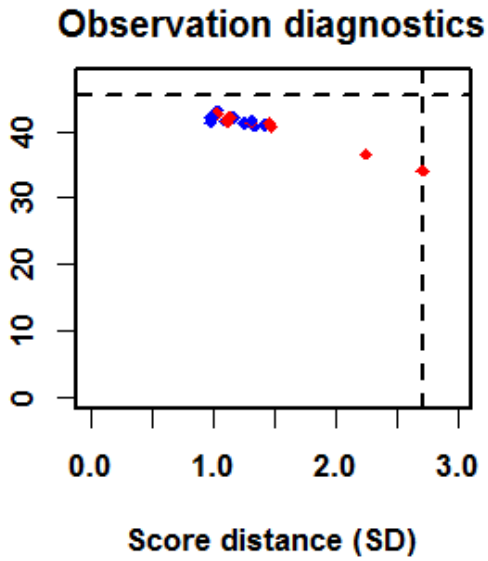
Score plot!

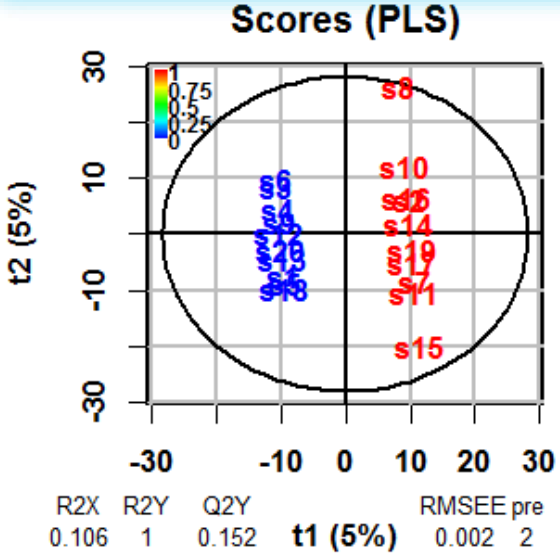

Scores (PLS)
1
0.75
0.5
0.25
0
s8
s10
s6
s3
s16
s2
s4
s9
s14
s12
s20
s19
s13
s17
s1
s7
s18
s11
s15
t2 (5%)
30
20
10
0
-10
-20
-30
-30
-20
-10
0
10
20
30
R2X
0.106
R2Y
1
Q2Y
0.152
t1 (5%)
RMSEE
0.002
pre
2

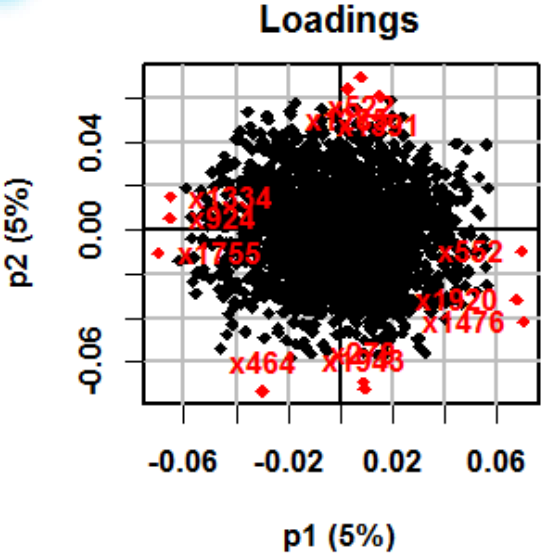
Importance of diagnostics

- **Permutation testing:** comparing the R2Y and Q2Y values of the model built with the true Y labels with n_{perm} **models built with random permutation of Y labels**







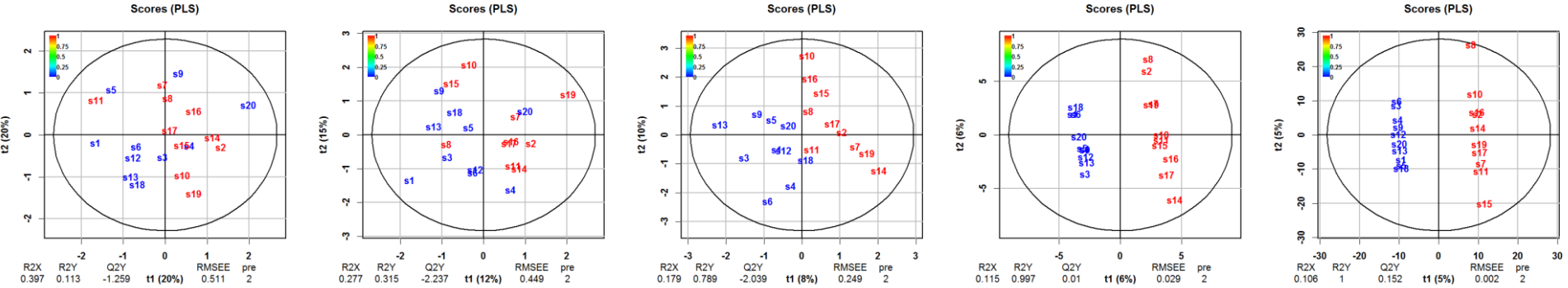


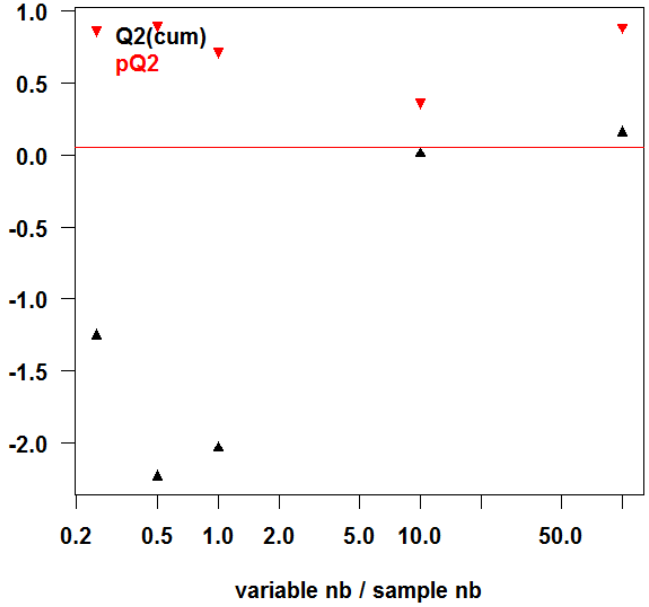
Szymanska E., Saccenti E., Smilde A. and Westerhuis J. (2012). Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics*, **8:**3-16. DOI: 10.1007/s11306-011-0330-3

Risk of overfitting when *n* < *p*

$$\frac{\text{variables}}{\text{samples}} =$$

0.2 | 0.5 | 1 | 10 | 100





Significance of the model

- The algorithm randomly permutates the **y** labels, builds the models and computes the $R2X$, $R2Y$, $Q2Y$

1 response

p = 110 (quantitative) variables

***n* = 183 samples**

	bmi
H011	19.8
H023	29.6
H033	18.4
H042	19.8
H052	20.1
H062	22.2
H073	25.4
H083	29.8
H092	21.8
H103	26.8
H114	29.4
H124	22.2
H134	22.9
H145	29.1
H157	22.0
H168	20.8
H180	23.7
H189	19.4
H199	21.0
H209	21.5

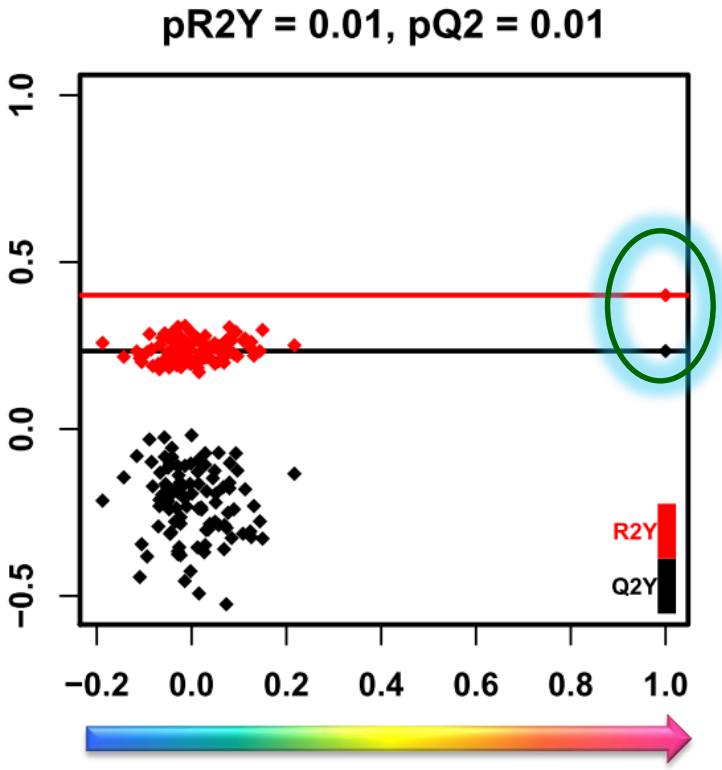
	1,7-Dimethyluric acid	Dehydroepiandrosterone sulfate	...
H011	3.33	4.46	
H023	4.64	2.81	
H033	4.35	2.51	
H042	3.91	4.14	
H052	4.35	2.55	
H062	3.80	2.47	
H073	4.00	4.36	
H083	4.48	2.02	
H092	3.82	4.55	
H103	4.08	0.21	
H114	4.52	5.17	
H124	4.05	4.93	
H134	4.27	4.53	
H145	4.16	5.33	
H157	4.50	4.29	
H168	4.01	1.89	
H180	4.36	2.67	
H189	4.16	3.02	
H199	3.46	4.09	
H209	4.10	5.00	

y_{random}

X

Significance of the model

- Counting the number of $R2Y$ (and $Q2Y$) metrics from random models which are superior to the values of the true model gives an indication of the significance of the PLS modelling



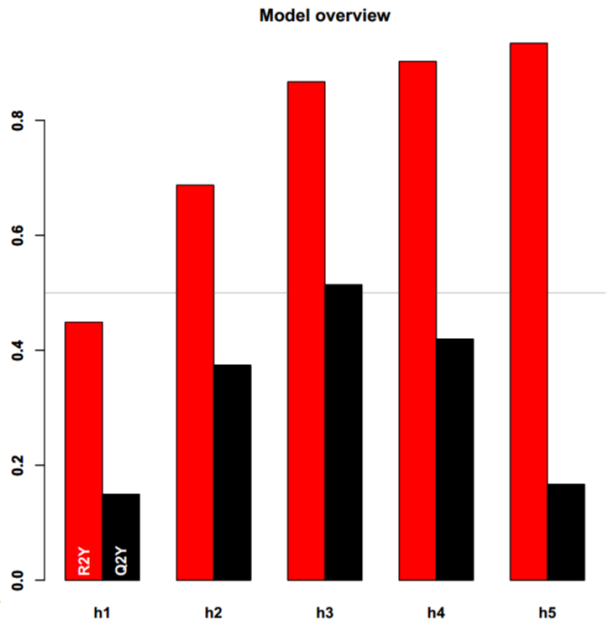
$R2Y$ and $Q2Y$ of the model with the true $\mathbf{y}$ values

'permutation'

Diagnostic metrics

- $0 \leq R2X \leq 1$: percentage of X inertia explained by the model
- $0 \leq R2Y \leq 1$: percentage of Y inertia explained by the model
- $0 \leq Q2Y \leq 1$: estimation of the predictive performance of the model by cross-validation

- $R2X$ and $R2Y$ increase with the number of components while $Q2Y$ reaches a maximum (due to overfitting):



'overview'

Numerical results

- The details of the $R2X$, $R2Y$, and $Q2Y$ values are stored in the "information.txt" file

olomics Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 2%

```
Y: mean-centering and unit-variance scaling

PLS ('nipals' algorithm)
Number of predictive components: 3

Number of reference observations: 183 (100%)

Correlations between variables and components:
                                    h1     h2 cor_h1 cor_h2
alpha-N-Phenylacetyl-glutamine    -0.220     NA  -0.64     NA
Phe-Tyr-Asp (and isomers)         -0.220     NA  -0.63     NA
Glucuronic acid and/or isomers    -0.220     NA  -0.62     NA
Asp-Leu/Ile isomer 1               0.080     NA   0.22     NA
6-(carboxymethoxy)-hexanoic acid   0.097     NA   0.27     NA
Testosterone glucuronide           0.180     NA   0.50     NA
Acetaminophen glucuronide             NA -0.093     NA -0.240
p-Anisic acid                         NA -0.066     NA -0.180
Malic acid                            NA -0.030     NA -0.078
p-Hydroxymandelic acid                NA  0.200     NA  0.520
1-Methyluric acid                     NA  0.200     NA  0.530
Porphobilinogen                       NA  0.200     NA  0.530

Model overview:
      R2X R2X(cum)    R2Y R2Y(cum)    Q2 Q2(cum) Signif. Iter.
h1 0.0984   0.0984 0.4791    0.479 0.401   0.401      R1     1
h2 0.0861   0.1846 0.1892    0.668 0.256   0.555      R1     1
h3 0.0907   0.2752 0.0615    0.730 0.065   0.584      R1     1

Model summary:
   R2X(cum) R2Y(cum) Q2(cum) RMSEE ncp nco
h3    0.275     0.73   0.584 0.262   3   0
```

History

search datasets

multivariate_example
14 shown, 10 deleted

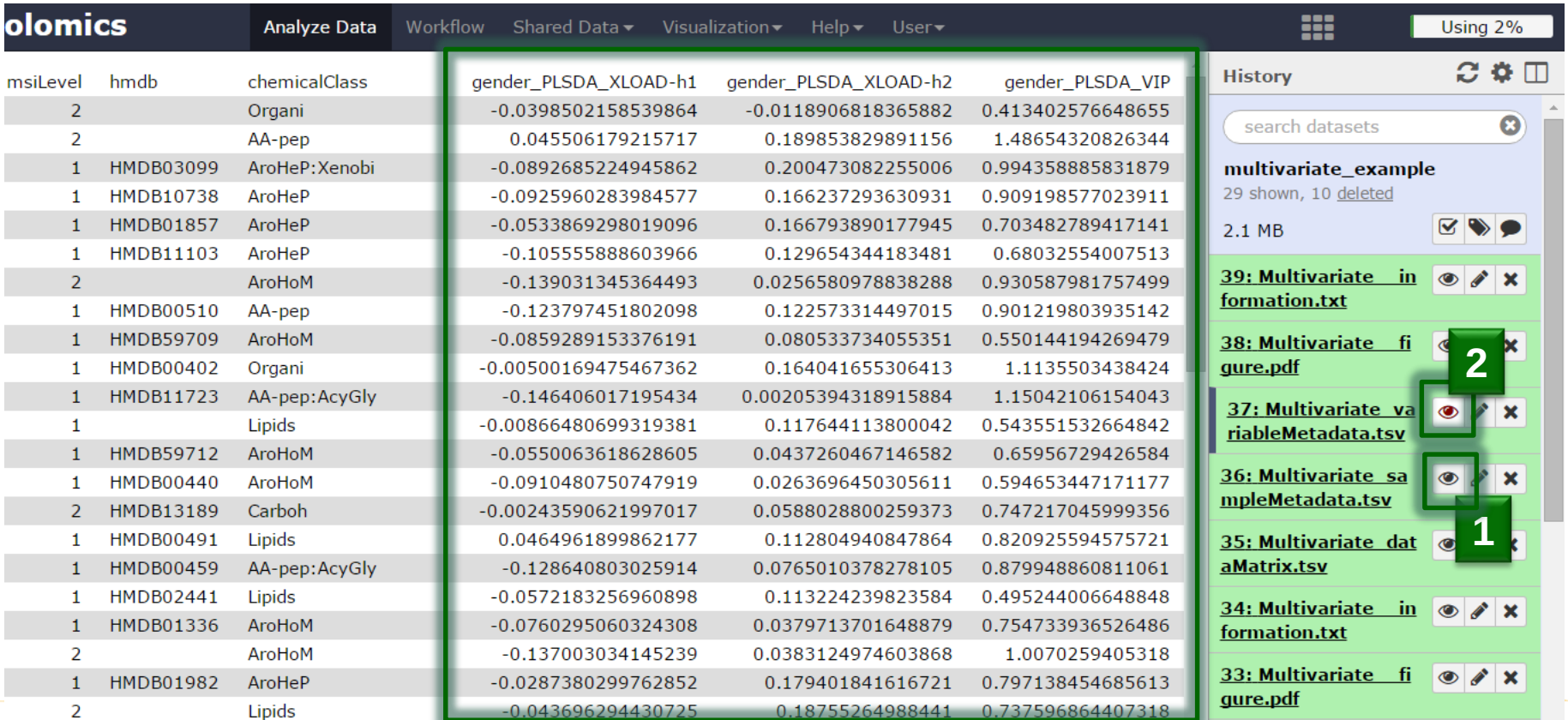
2.1 MB

1

39: Multivariate__information.txt
38: Multivariate__figure.pdf
37: Multivariate_variableMetadata.tsv
36: Multivariate_sampleMetadata.tsv
35: Multivariate_dataMatrix.tsv
34: Multivariate__information.txt
33: Multivariate__figure.pdf
32: Multivariate_variableMetadata.tsv

Scores, loadings and VIPs

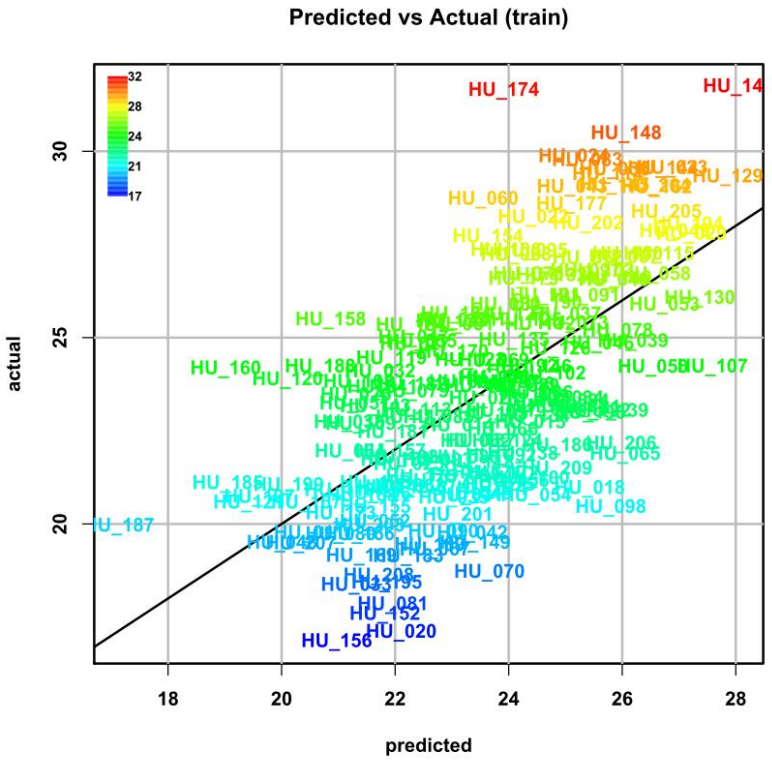
- The score (resp. loading and VIPs) of the selected components have been added as columns in the sampleMetadata (resp. variableMetadata) files



msiLevel	hmdb	chemicalClass	gender_PLSDA_XLOAD-h1	gender_PLSDA_XLOAD-h2	gender_PLSDA_VIP
2		Organi	-0.0398502158539864	-0.0118906818365882	0.413402576648655
2		AA-pep	0.045506179215717	0.189853829891156	1.48654320826344
1	HMDB03099	AroHeP:Xenobi	-0.0892685224945862	0.200473082255006	0.994358885831879
1	HMDB10738	AroHeP	-0.0925960283984577	0.166237293630931	0.909198577023911
1	HMDB01857	AroHeP	-0.0533869298019096	0.166793890177945	0.703482789417141
1	HMDB11103	AroHeP	-0.105555888603966	0.129654344183481	0.68032554007513
2		AroHoM	-0.139031345364493	0.0256580978838288	0.930587981757499
1	HMDB00510	AA-pep	-0.123797451802098	0.122573314497015	0.901219803935142
1	HMDB59709	AroHoM	-0.0859289153376191	0.080533734055351	0.550144194269479
1	HMDB00402	Organi	-0.00500169475467362	0.164041655306413	1.1135503438424
1	HMDB11723	AA-pep:AcyGly	-0.146406017195434	0.00205394318915884	1.15042106154043
1		Lipids	-0.00866480699319381	0.117644113800042	0.543551532664842
1	HMDB59712	AroHoM	-0.0550063618628605	0.0437260467146582	0.65956729426584
1	HMDB00440	AroHoM	-0.0910480750747919	0.0263696450305611	0.594653447171177
2	HMDB13189	Carboh	-0.00243590621997017	0.0588028800259373	0.747217045999356
1	HMDB00491	Lipids	0.0464961899862177	0.112804940847864	0.820925594575721
1	HMDB00459	AA-pep:AcyGly	-0.128640803025914	0.0765010378278105	0.879948860811061
1	HMDB02441	Lipids	-0.0572183256960898	0.113224239823584	0.495244006648848
1	HMDB01336	AroHoM	-0.0760295060324308	0.0379713701648879	0.754733936526486
2		AroHoM	-0.137003034145239	0.0383124974603868	1.0070259405318
1	HMDB01982	AroHeP	-0.0287380299762852	0.179401841616721	0.797138454685613
2		Lipids	-0.043696294430725	0.18755264988441	0.737596864407318

Advanced parameters: Graphics

- Several types of graphics are available:
  - e.g., predict-train and predict-test (the latter being available only if the train/test partition has been selected)



'predict-train'

PLS-DA

- The two response levels are encoded as numbers

n = 183 samples

Qualitative

	gender
H011	M
H023	M
H033	F
H042	M
H052	F
H062	M
H073	M
H083	M
H092	M
H103	M
H114	M
H124	M
H134	M
H145	M
H157	F
H168	F
H180	F
H189	F
H199	M
H209	F

$\mathbf{y}$

Quantitative

	gender
HU_017	0.5
HU_028	0.5
HU_034	-0.5
HU_051	0.5
HU_060	-0.5
HU_078	0.5
HU_091	0.5
HU_093	0.5
HU_099	0.5
HU_110	0.5
HU_130	0.5
HU_134	0.5
HU_138	0.5
HU_149	0.5
HU_152	-0.5
HU_175	-0.5
HU_178	-0.5
HU_185	-0.5
HU_204	0.5
HU_208	-0.5

$\mathbf{y}$

PLS

Quantitative

	gender
H011	0.40
H023	0.10
H033	-0.61
H042	0.39
H052	-0.47
H062	0.46
H073	0.36
H083	0.11
H092	0.47
H103	0.23
H114	0.25
H124	0.56
H134	0.12
H145	0.93
H157	-0.19
H168	-0.49
H180	-0.20
H189	0.00
H199	0.54
H209	0.05

$\mathbf{y}_{fitted}$

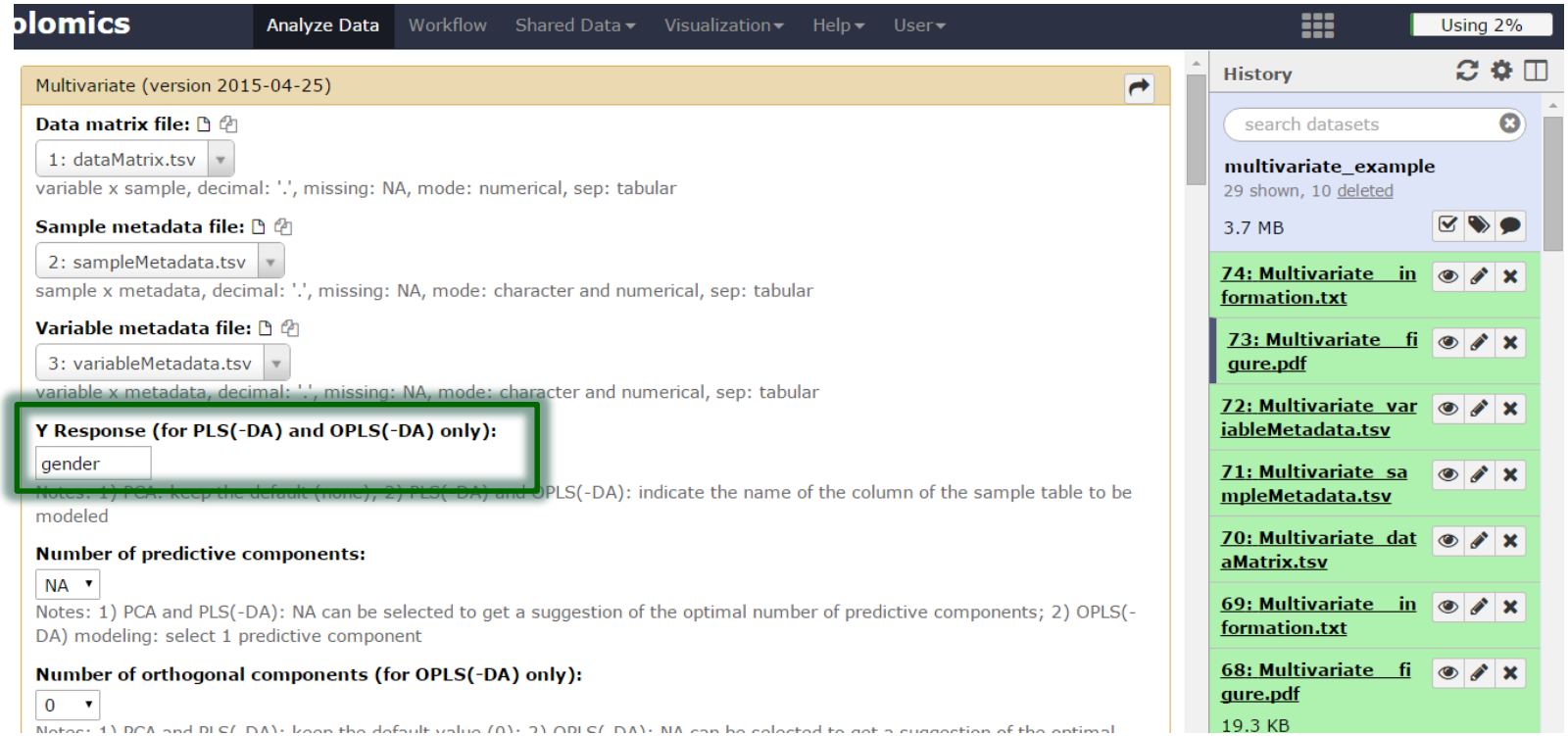
Qualitative

	pred
H011	M
H023	M
H033	F
H042	M
H052	F
H062	M
H073	M
H083	M
H092	M
H103	M
H114	M
H124	M
H134	M
H145	M
H157	F
H168	F
H180	F
H189	M
H199	M
H209	M

$\mathbf{y}_{fitted}$

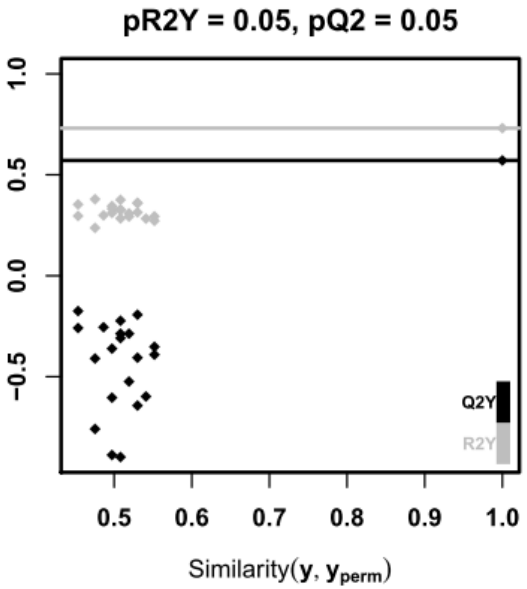
PLS-DA

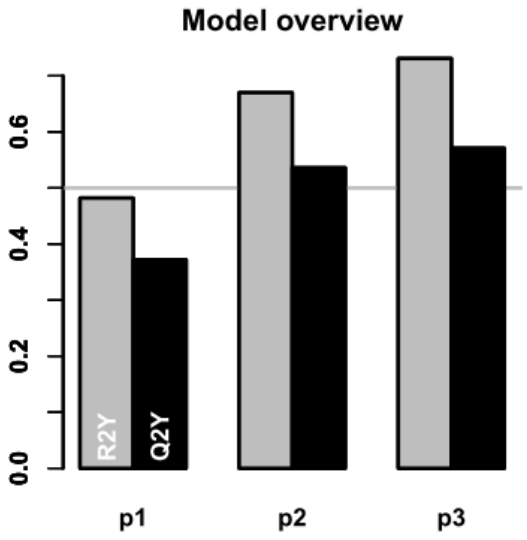
- Automatically selected when the response is qualitative (i.e. the column of sampleMetadata only contains characters)

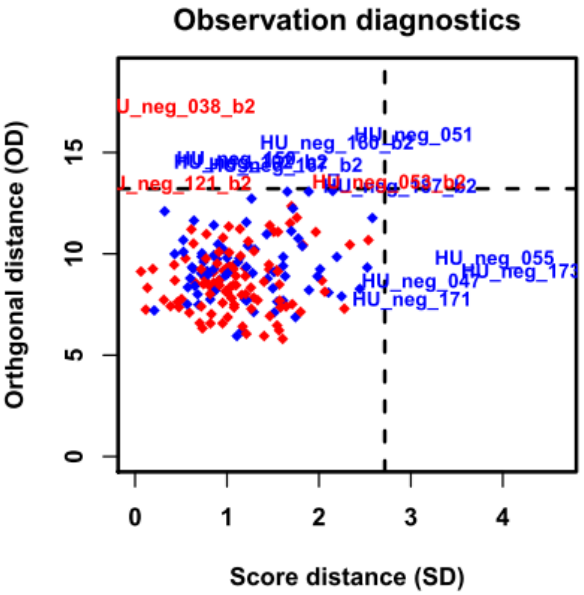


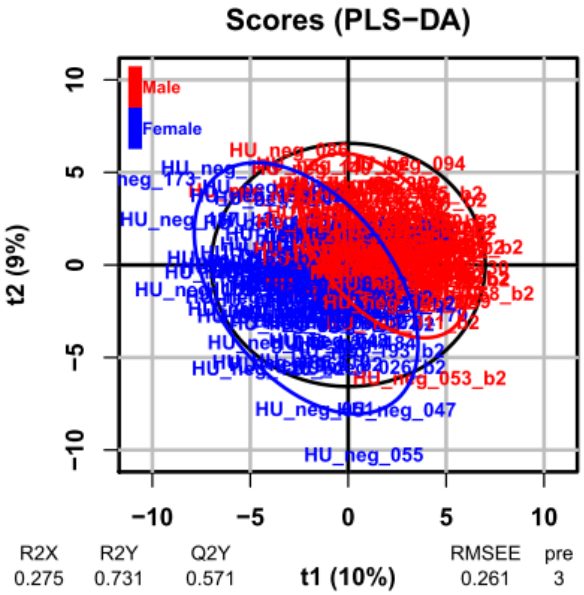
Warning: Use balanced datasets (similar proportions of samples in each of the two classes)

PLS-DA


pR2Y = 0.05, pQ2 = 0.05
1.0
0.5
0.0
−0.5
Q2Y
R2Y
0.5
0.6
0.7
0.8
0.9
1.0
Similarity(y, y_perm)


Model overview
0.6
0.4
0.2
0.0
R2Y
Q2Y
p1
p2
p3


Observation diagnostics
Orthgonal distance (OD)
0
5
10
15
HU_neg_038_b2
HU_neg_051
HU_neg_055
HU_neg_047
HU_neg_171
0
1
2
3
4
Score distance (SD)


Scores (PLS−DA)
Male
Female
t2 (9%)
10
5
0
−5
−10
HU_neg_053_b2
HU_neg_047
HU_neg_055
−10
−5
0
5
10
t1 (10%)
R2X
0.275
R2Y
0.731
Q2Y
0.571
RMSEE
0.261
pre
3

ORTHOGONAL PARTIAL LEAST SQUARES REGRESSION (OPLS) AND DISCRIMINANT ANALYSIS (OPLS-DA)

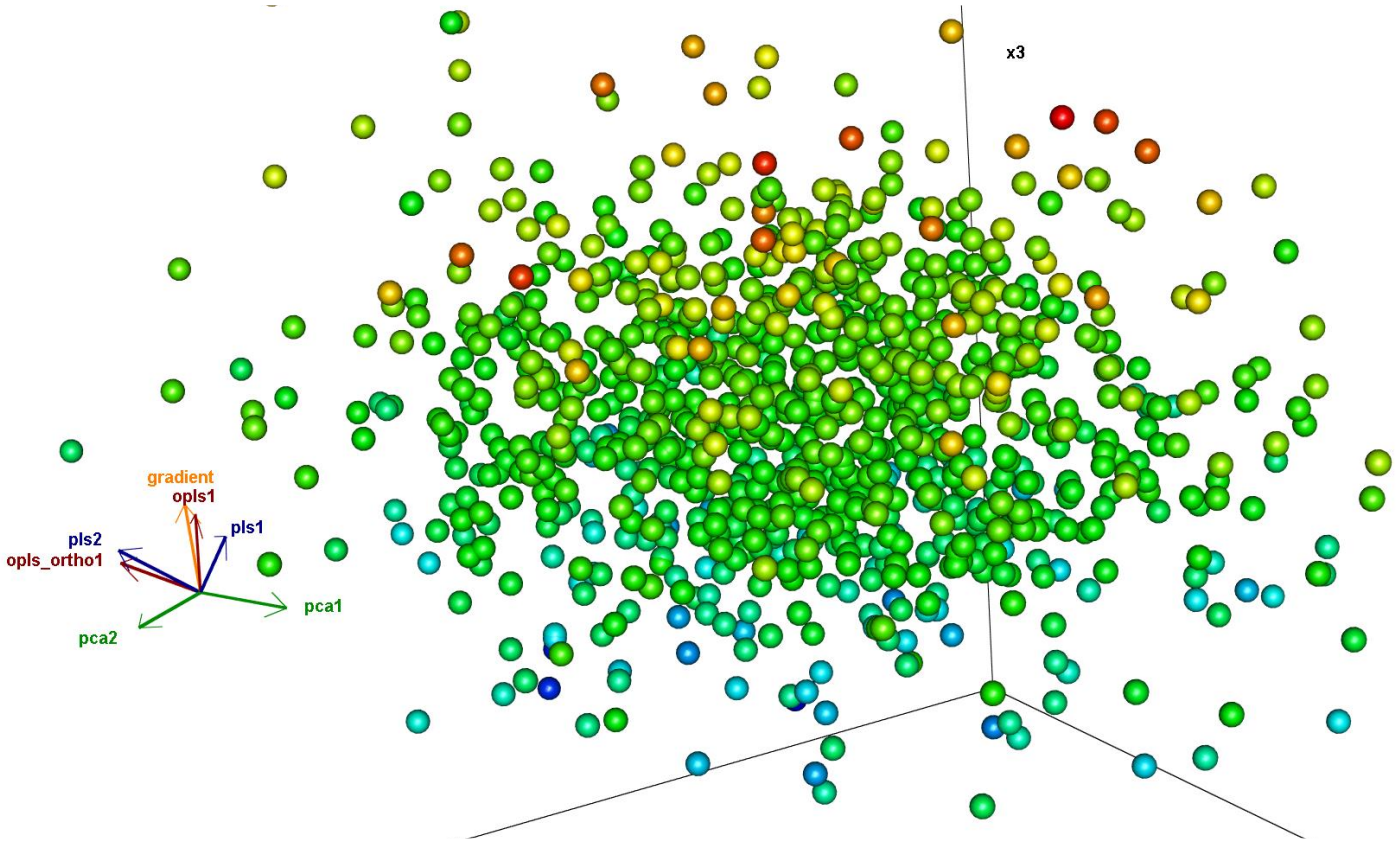
Principles

- Separately models the variations of the predictors correlated and orthogonal to the response
- Improves the interpretation of the components but not the overall predictive performance of the model
- Only one predictive component required for single response models
- Note: As with PLS, care should be taken to avoid too many (orthogonal) components (which would result in overfitting)

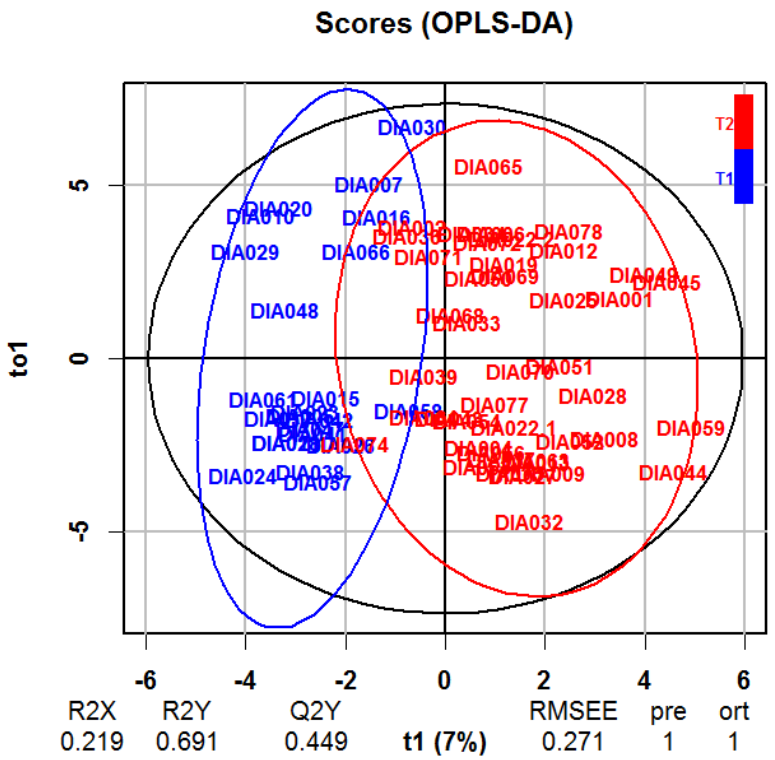
OPLS vs PLS

- Variation not correlated to the response (e.g., technical bias) is modelled separately by the orthogonal component(s)

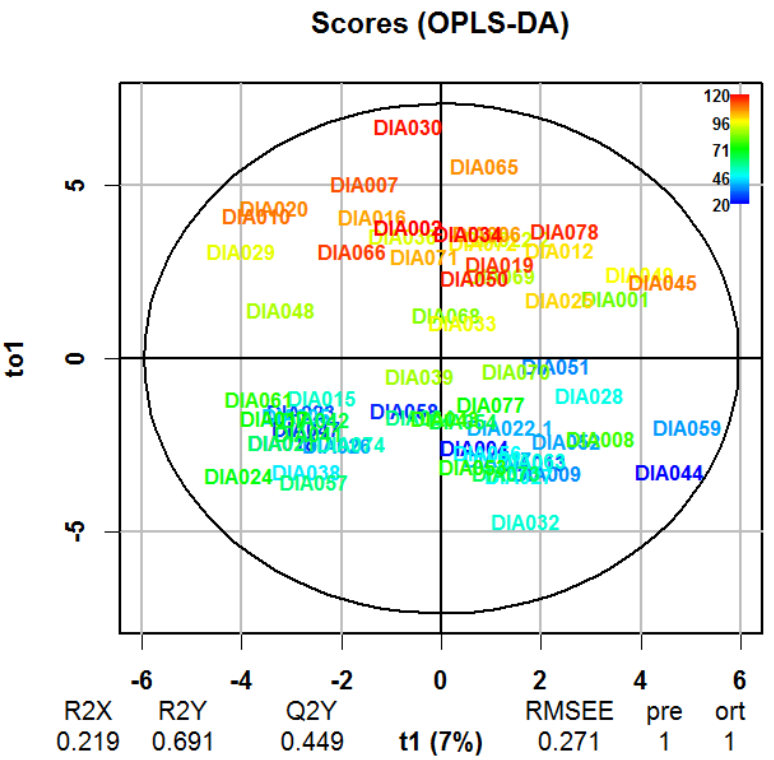
=> The first predictive component is strongly correlated to the response



Predictive and orthogonal variations

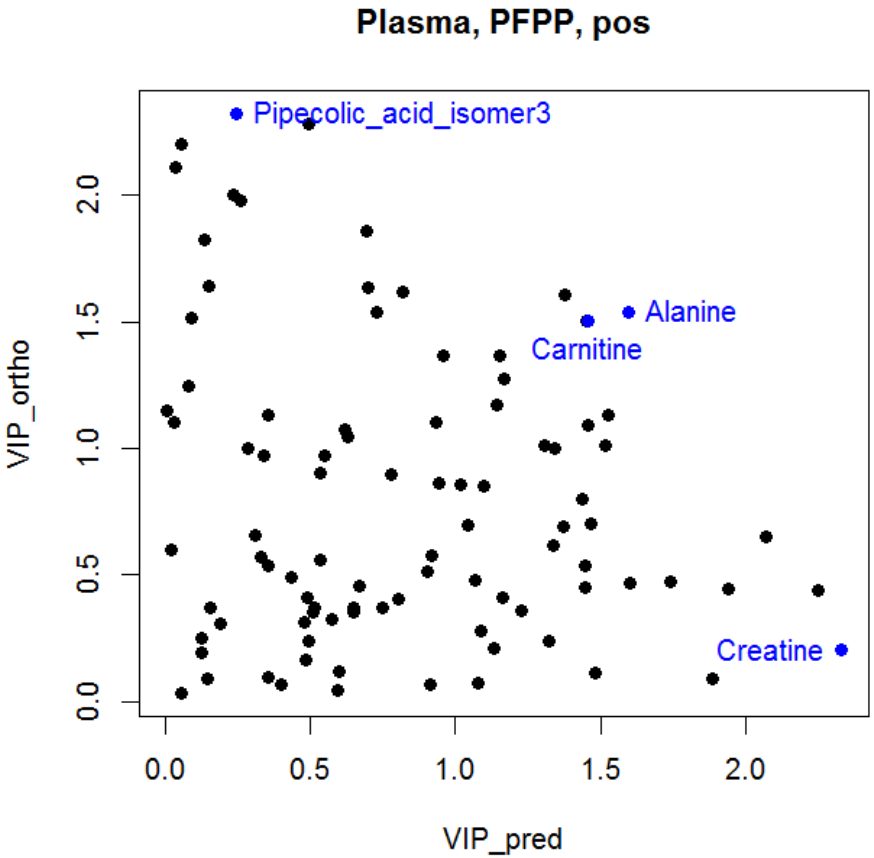


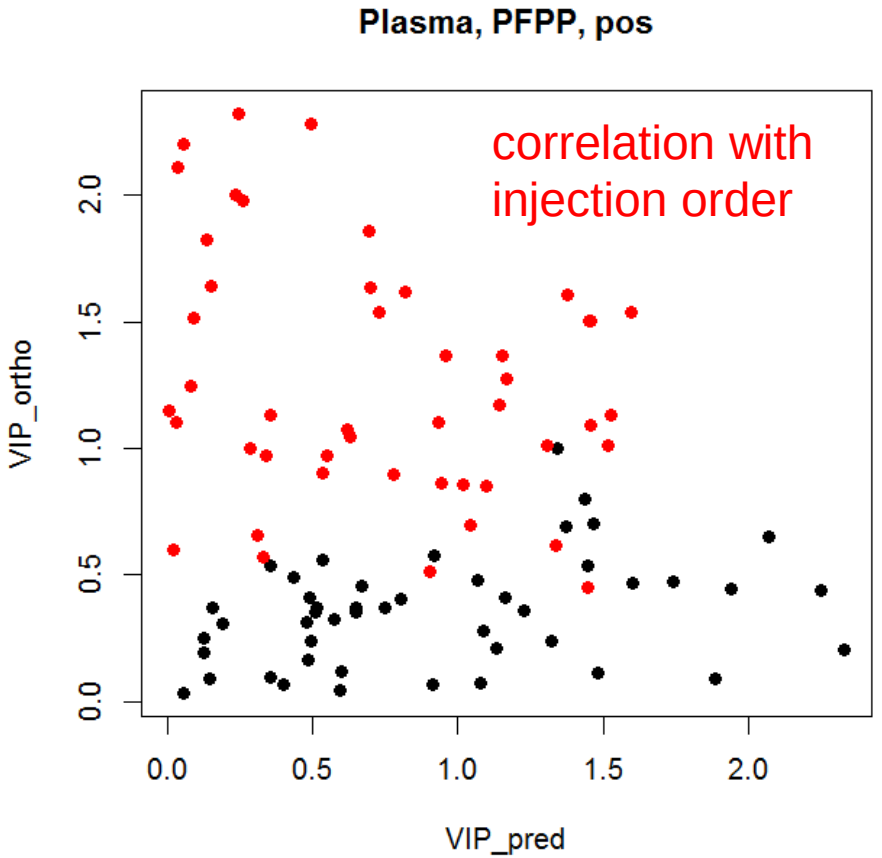
drift



type

Predictive and orthogonal VIP

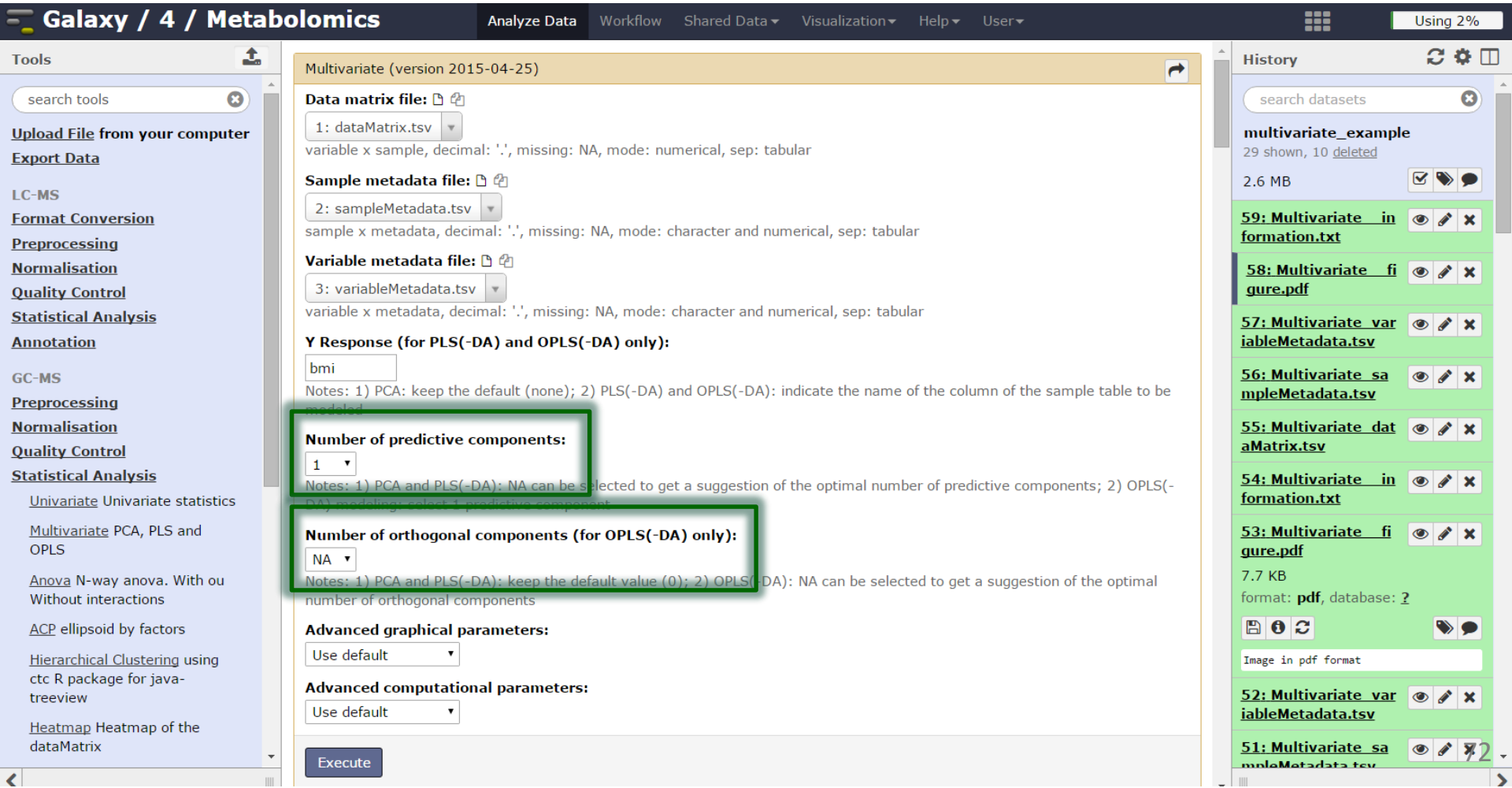




Galindo-Prieto et al (2014). *Journal of Chemometrics,* 28, 623-632.

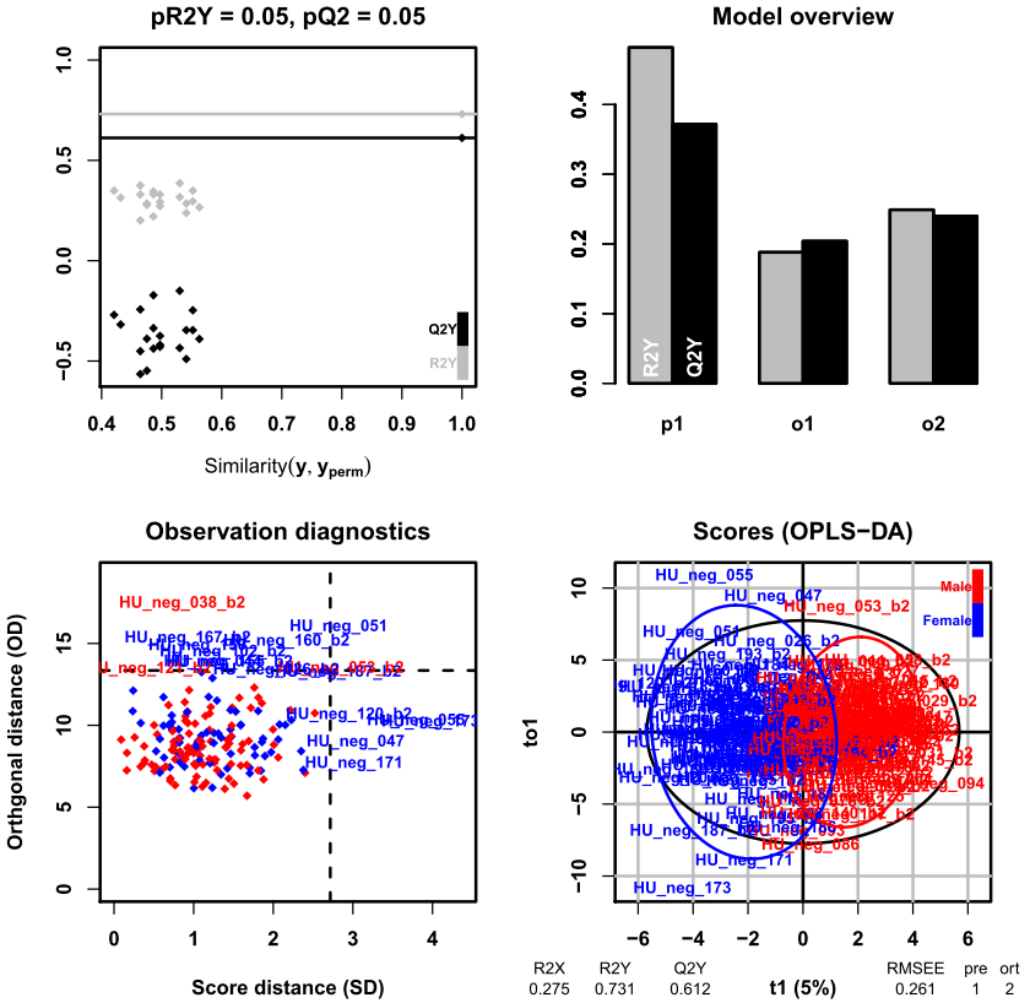
Selection of OPLS(-DA) as the type of analysis

- Set the number of predictive component to 1
- Select the number of orthogonal components (e.g., NA)



Graphical results

- permutation, overview, outlier, and score plots displayed as the default ('summary')



Numerical results

- The details of the $R2X$, $R2Y$, and $Q2Y$ values are stored in the "information.txt" file

```
Start of the 'Multivariate' Galaxy module call: Mon 13 Feb 2017 06:08:23 PM

Warning: OPLS: number of predictive components ('predI' argument) set to 1

OPLS-DA

183 samples x 110 variables and 1 response

standard scaling of dataMatrix and response

       R2X R2X(cum)    R2Y R2Y(cum)     Q2 Q2(cum) Signif.
p1  0.0499   0.0499 0.4820    0.482 0.3720   0.372       R1
o1  0.1250   0.1750 0.1880    0.188 0.2040   0.204       R1
o2  0.0999   0.2750 0.0603    0.249 0.0368   0.240       R1
sum     NA   0.2750     NA    0.731     NA   0.612    <NA>

End of 'Multivariate' Galaxy module call: 2017-02-13 18:08:25
```

References

- Wold S., Sjöström M. and Eriksson L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems,* **58:**109-130. http://dx.doi.org/10.1016/S0169-7439(01)00155-1

- Trygg J., Holmes E. and Lundstedt T. (2007). Chemometrics in Metabonomics. *Journal of Proteome Research*, **6**:469-479. http://dx.doi.org/10.1021/pr060594q

- Brereton R.G. and Lloyd G.R. (2014). Partial least squares discriminant analysis: taking the magic away. *Journal of Chemometrics,* **28:**213-225. http://dx.doi.org/10.1002/cem.2609

The Sacurine study

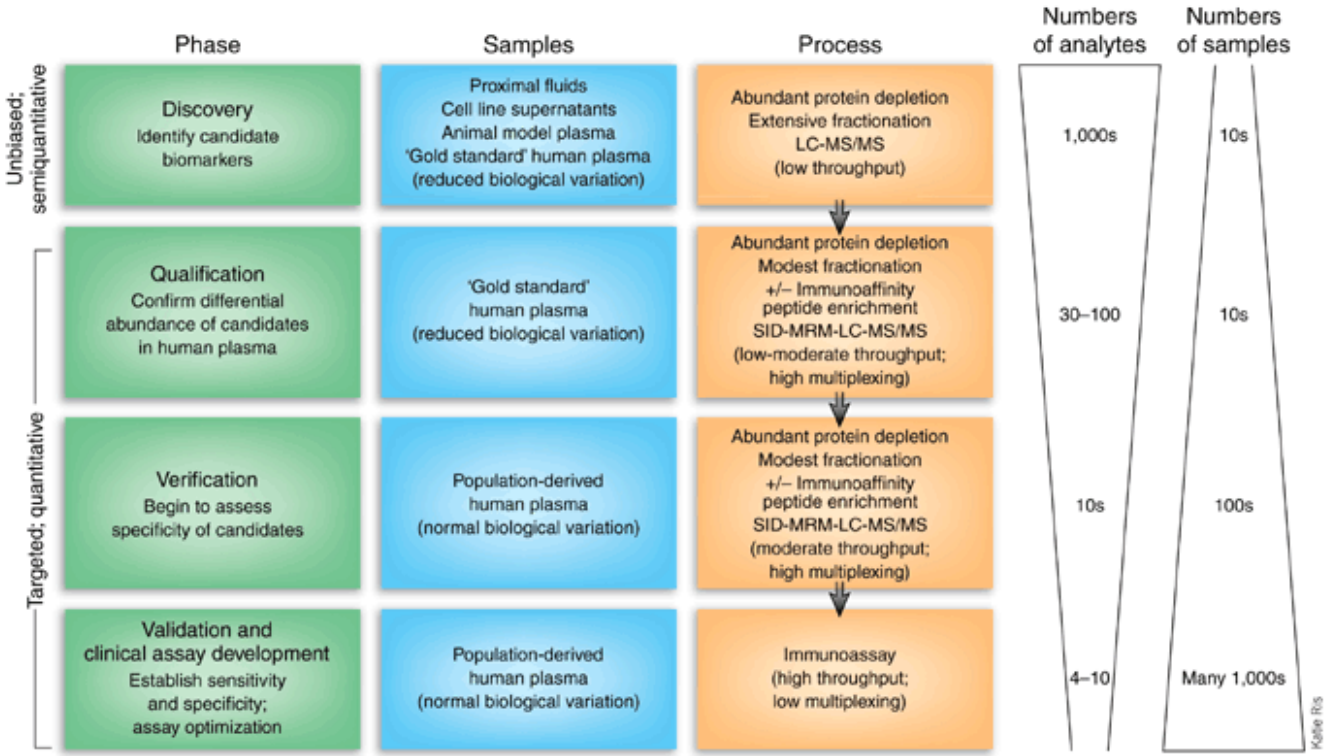
Exploratory Data Analysis

Multivariate modeling

➢ **Selection of molecular signatures**

Feature selection: objectives

- Limit the risk of overfitting
- Stabilize the prediction
- Facilitates interpretation
- Restrict the list of candidates before the subsequent validation phas



Rifai et al (2006).
Nat. Biotechnol., **24**:971-983

Feature selection: challenges

- Testing all combination of features is not computationally tractable
  - efficient search path
- Prediction performance
  - sensitivity, selectivity
- Stability
  - reproducibility
- Relevance
  - selection criterion

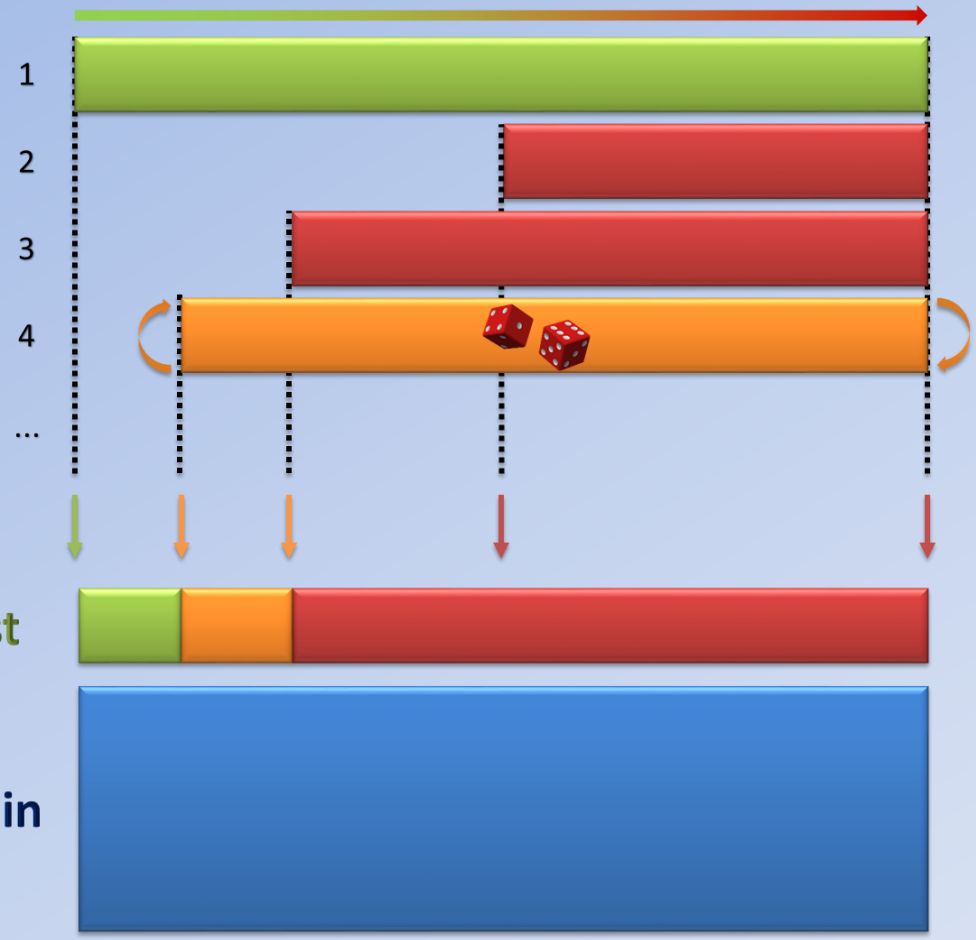
Feature selection: approaches

- filter (threshold criterion)
  - e.g., *t*-test — **fast** — **threshold?**
- wrapper (iterative selection)
  - e.g., SVM RFE — **interaction with classifier** — **computation intensive**
- embedded (penalization constraint)
  - e.g., Lasso, Elastic Net — **fast** — **stability**

The *biosigner* wrapper algorithm

Repeat until selected subset is stable:

2) features ranked by their importance for the classifier on the train subsets



3) half-interval search for the largest irrelevant feature subset

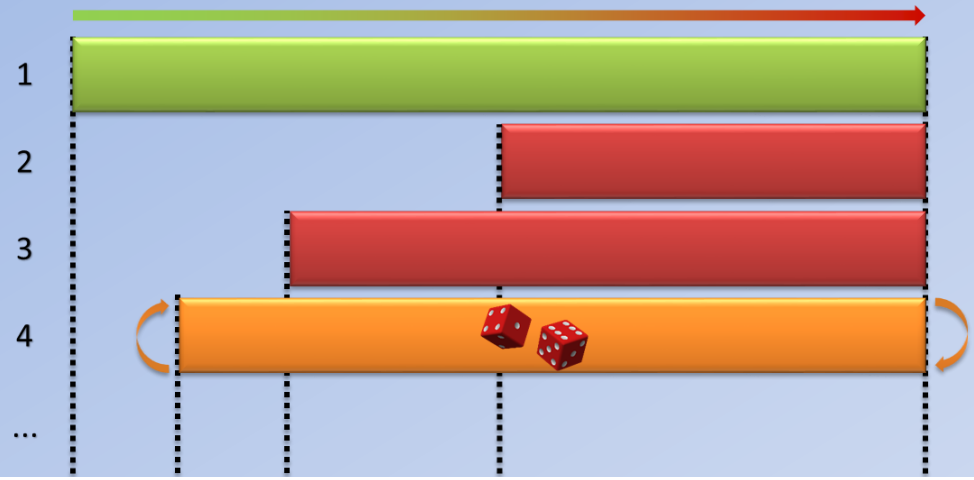
1) resampling (bootstrap)

4) discard irrelevant features

The *biosigner* wrapper algorithm

Repeat until selected subset is stable:

2) features ranked by their importance for the classifier on the train subsets



3) half-interval search for the largest irrelevant feature subset

1) resampling (bootstrap)

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 21 June 2016
doi: 10.3389/fmolb.2016.00026

biosigner: A New Method for the Discovery of Significant Molecular Signatures from Omics Data

*Philippe Rinaudo[1], Samia Boudah[2], Christophe Junot[2] and Etienne A. Thévenot[1]**

[1] CEA, LIST, Laboratory for Data Analysis and Systems' Intelligence, MetaboHUB, Gif-sur-Yvette, France, [2] Laboratoire d'Etude du Métabolisme des Médicaments, DSV/iBiTec-S/SPI, MetaboHUB, CEA-Saclay, Gif-sur-Yvette, France

Feature tiers: number of successful selection rounds

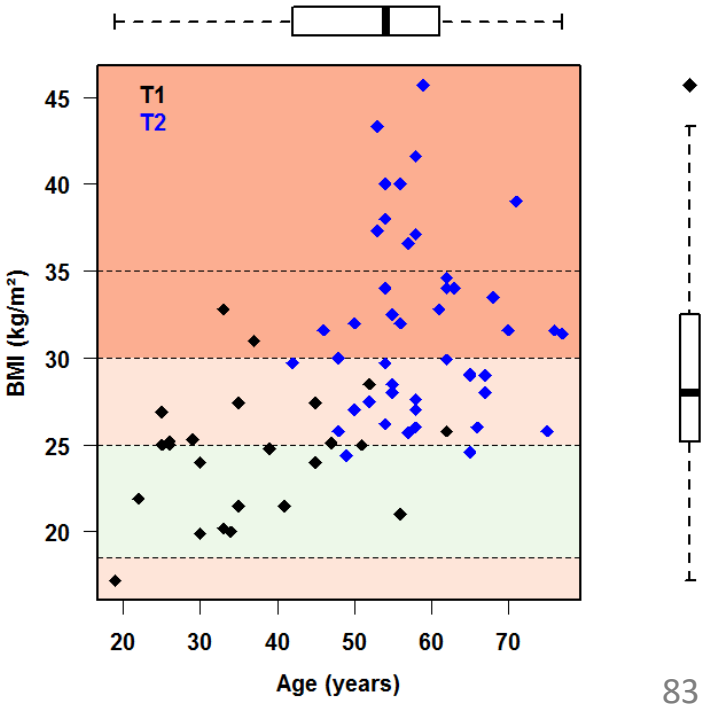
Selection rounds: **all** | **all but the last one** | **...** | **all but the fifth last ones**

molecular signature

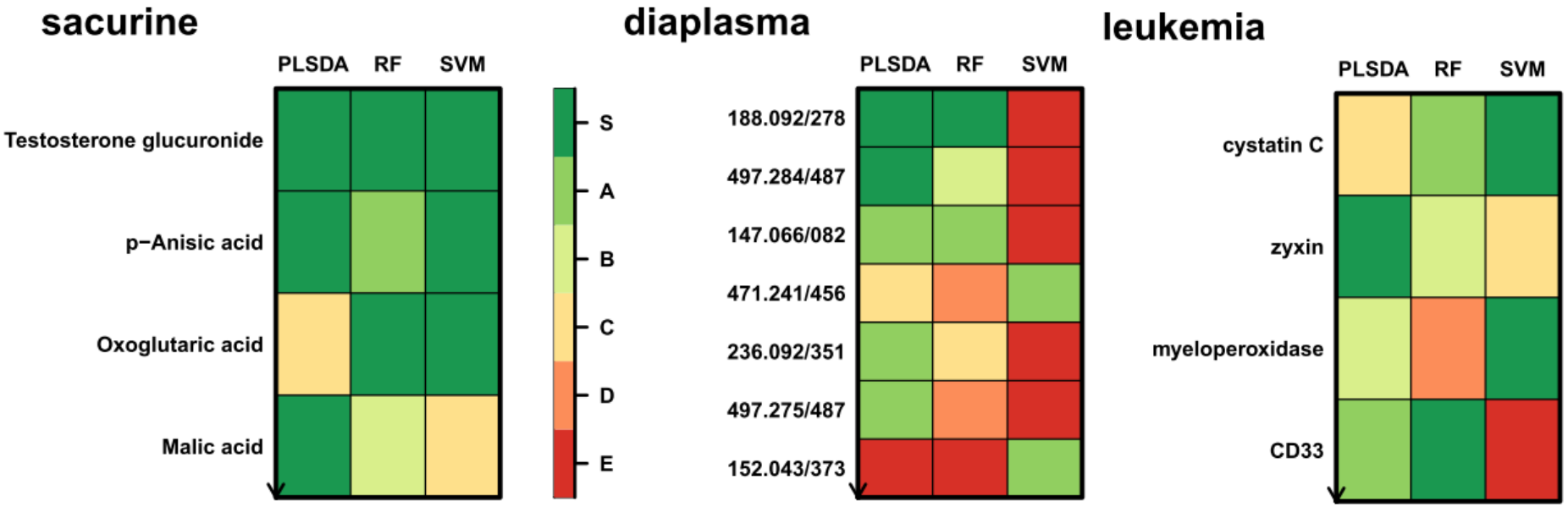
The *biosigner* package

- Feature selection algorithm wrapped around 3 classifiers
  - Partial Least Squares - Discriminant Analysis (PLS-DA)
  - Random Forest
  - Support Vector Machine (SVM)

- *diaplasma* dataset
  - LC-HRMS analysis of plasma
  - from a cohort of 69 diabetic patients
  - type 1 and type 2 patients
  - 5,501 mz/RT features

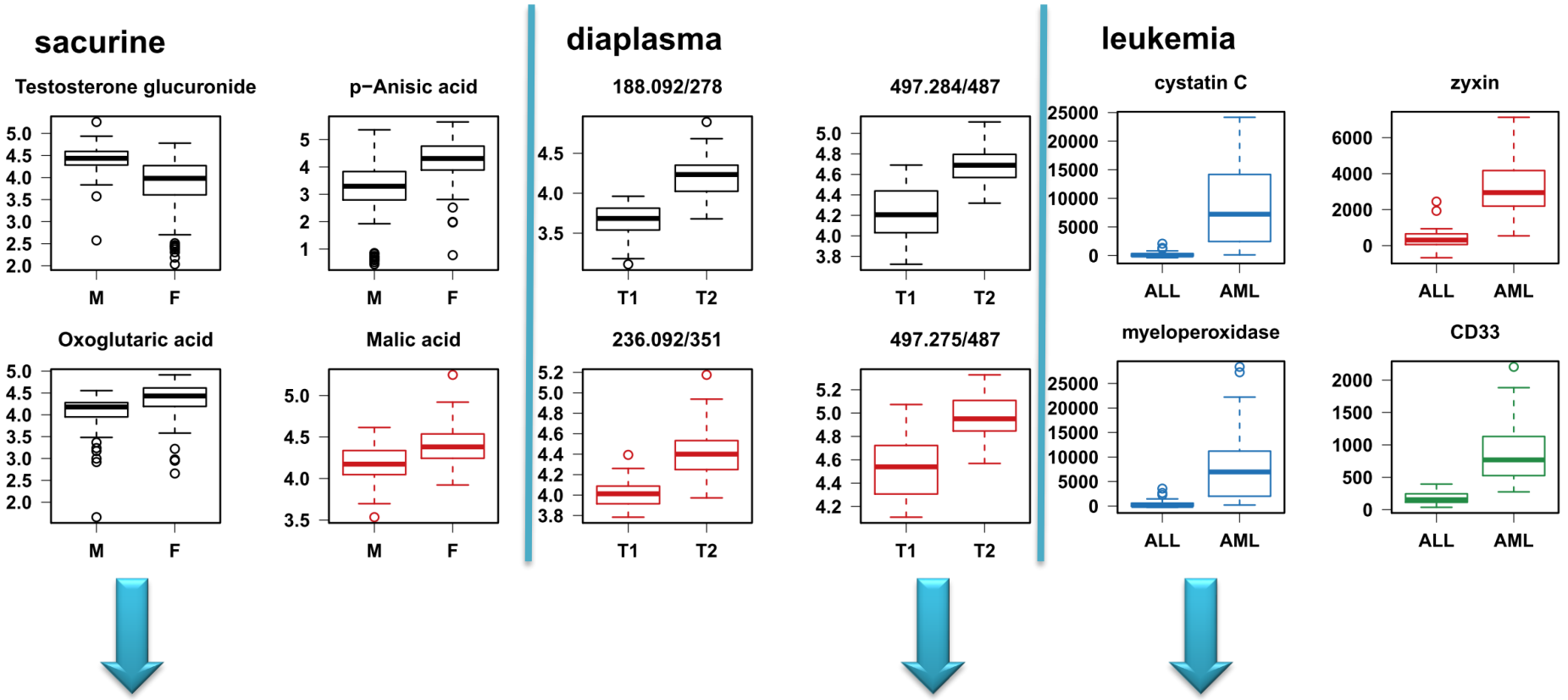


Model performances



		sacurine	diaplasma	leukemia
factor		gender	diabetic type	ALL/AML
samples		183	69	72
features		109	5,501	7,129
signatures		[2-3]	[0-2]	[1-2]
performances (full -> restricted)	PLS-DA	87% -> **89%**	83% -> **91%**	95% -> **87%**
	Random Forest	86% -> **86%**	81% -> **81%**	92% -> **92%**
	SVM	88% -> **89%**	83% -> **na**	93% -> **95%**

Molecular signatures

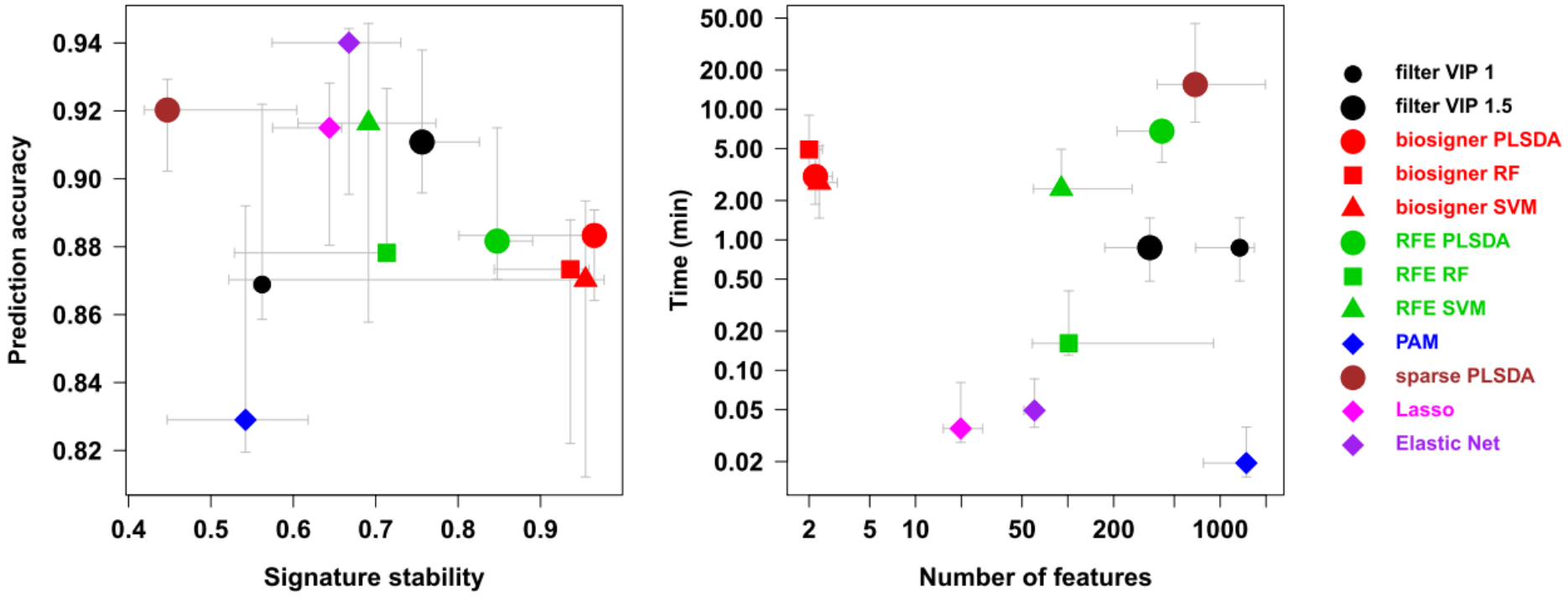


Biomarker in prostate cancer: Zhang et al. (2013). *PLoS ONE,* **8:**e65880.

Taurochenodeoxycholic acid: variation in type 2 diabetic patients: Taylor et al. (2014). PLoS ONE, **9**:e93540.

Cytochemical marker for the diagnosis of AML: Matsuo et al (2003). *Leukemia* **17**:1538-1543.

Comparison with alternative feature selection methods



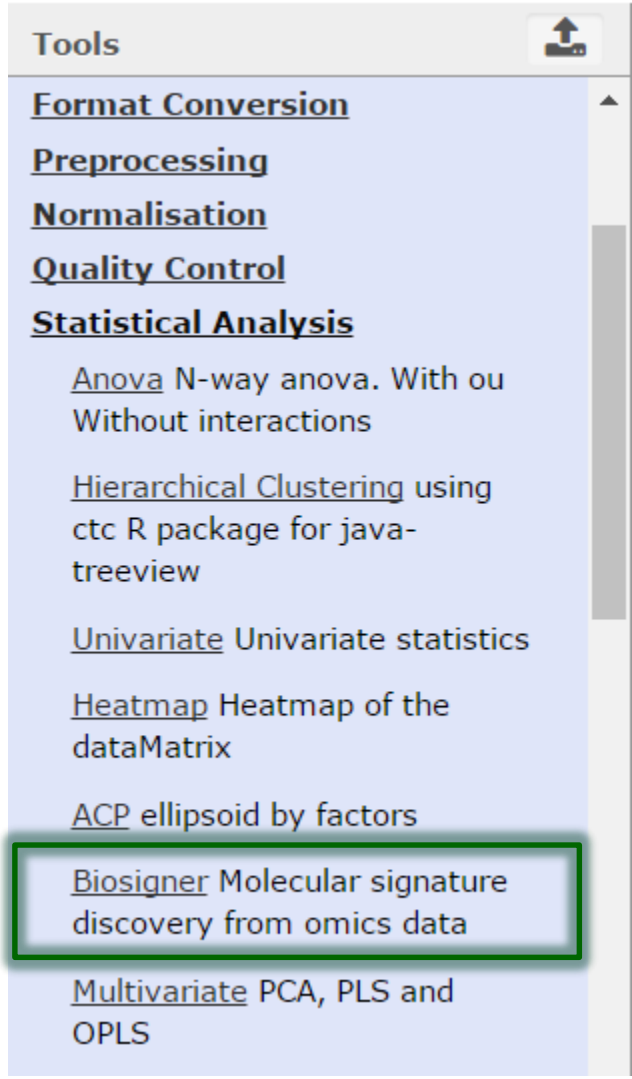
- Small signatures providing a good compromise between prediction accuracy, signature stability and computation time

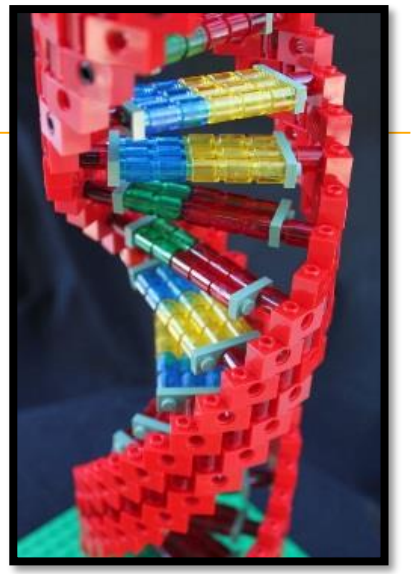
The "Biosigner" module

Biosigner

- selection of the features which are relevant for binary classification with:
  - Partial Least Squares - Discriminant Analysis (**PLS-DA**)
  - Random Forest (**RF**)
  - Support Vector Machine (**SVM**)

- available in the "Statistical Analysis" sections of LC-MS, GC-MS, and NMR





Enjoy your analyses!

Questions?

support@workflow4metabolomics.org

Please cite:

1. Giacomoni *et al.* (2015). *Bioinformatics*, 31:1493-1495
2. Guitton *et al.* (2017). *The International Journal of Biochemistry & Cell Biology*, 93:89-101