FROM RESEARCH TO INDUSTRY

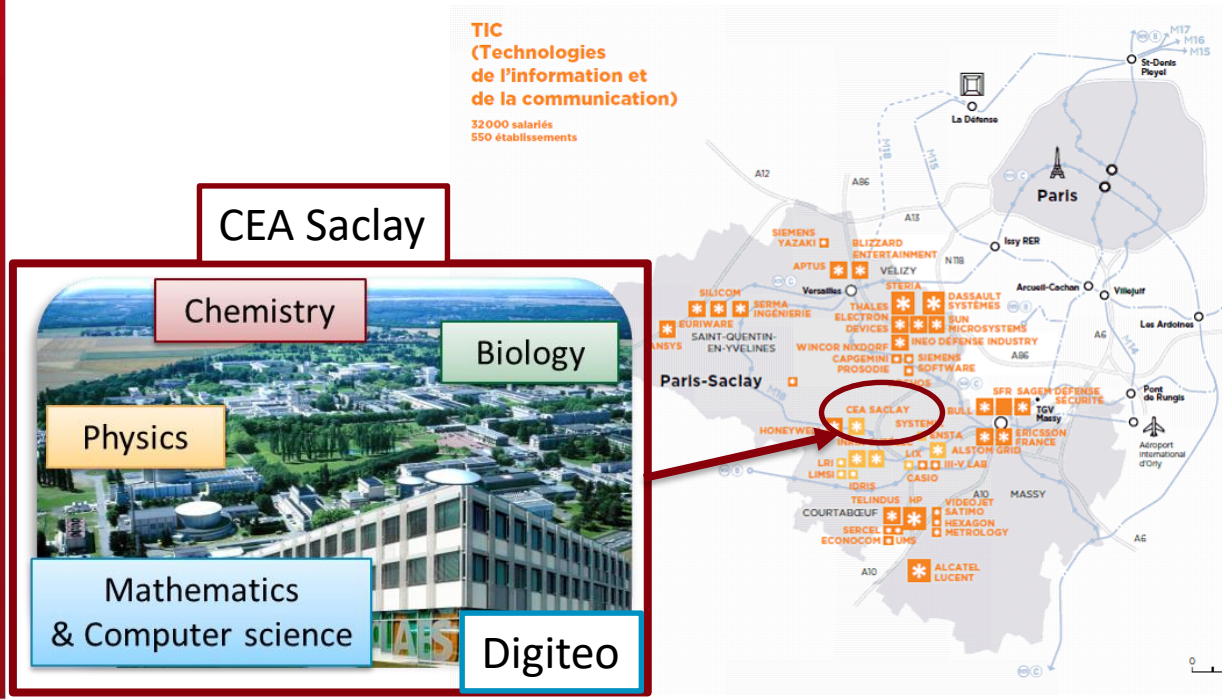
Omics data analysis for high-throughput phenotyping

Etienne Thévenot *et al.*

CEA, LIST (Saclay, France)
Laboratory for Data Analysis and Systems' Intelligence
MetaboHUB

etienne.thevenot@cea.fr

http://etiennethevenot.pagesperso-orange.fr/

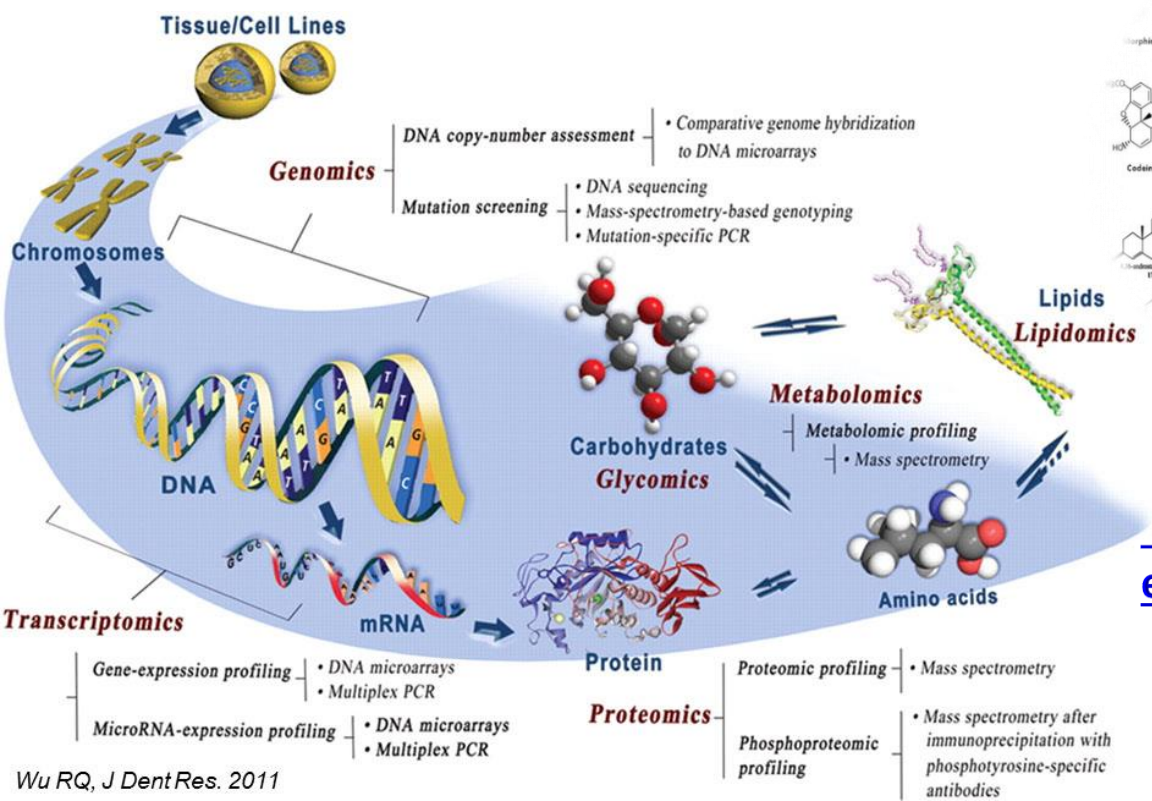


www.cea.fr

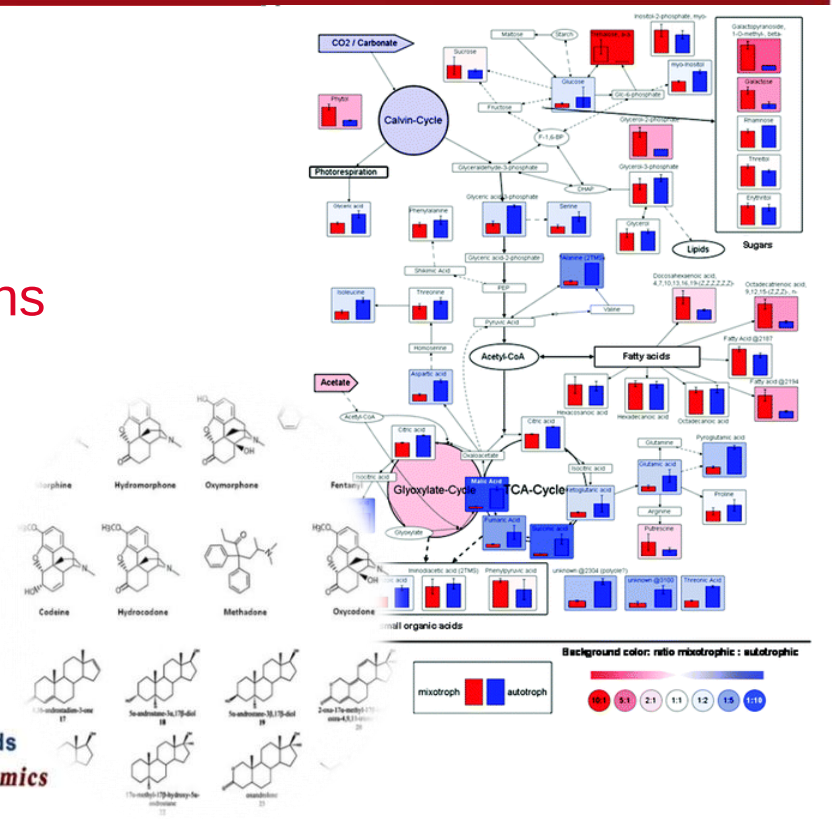
➢ Metabolomics

Metabolomics

- omics science
- dedicated to small molecules (< 1kDa)
- involved in metabolic chemical reactions



Wu RQ, J Dent Res. 2011



FUN MOOC La métabolomique : enjeux technologiques et scientifiques

Metabolomics successes for clinical biomarkers

- **Cancer** and oncometabolites: Yang et al., 2013


Oncometabolites: linking altered metabolism with cancer

Ming Yang,[1] Tomoyoshi Soga,[2,3] and Patrick J. Pollard[1,4]

[1]Cancer Biology and Metabolism Group, Nuffield Department of Medicine, Henry Wellcome Building for Molecular Physiology, University of Oxford, Oxford, United Kingdom. [2]Institute for Advanced Biosciences and [3]Oxford-Keio Metabolomics Consortium, Keio University, Yamagata, Japan. [4]Oxford-Keio Metabolomics Consortium, Oxford, United Kingdom.

- **Memory** impairment and phospholipids: Mapstone et al., 2014

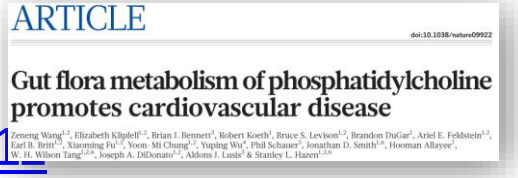

LETTERS

nature medicine

Plasma phospholipids identify antecedent memory impairment in older adults

Mark Mapstone[1], Amrita K Cheema[2,3], Massimo S Fiandaca[4,5], Xiaogang Zhong[6], Timothy R Mhyre[5], Linda H MacArthur[5], William J Hall[7], Susan G Fisher[8,14], Derick R Peterson[9], James M Haley[10], Michael D Nazar[11], Steven A Rich[12], Dan J Berlau[13,14], Carrie B Peltz[13], Ming T Tan[6], Claudia H Kawas[13] & Howard J Federoff[6,5]

- **Cardiovascular** disease and TMAO: Wang et al., 2011


ARTICLE

doi:10.1038/nature09922

Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease

Zeneng Wang[1,2], Elizabeth Klipfell[1,2], Brian J. Bennett[3], Robert Koeth[1], Bruce S. Levison[1,2], Brandon DuGar[1], Ariel E. Feldstein[1,2], Earl B. Britt[1,2], Xiaoming Fu[1,2], Yoon-Mi Chung[1,2], Yuping Wu[4], Phil Schauer[5], Jonathan D. Smith[1,6], Hooman Allayee[7], W. H. Wilson Tang[1,2,8], Joseph A. DiDonato[1,2], Aldons J. Lusis[3] & Stanley L. Hazen[1,2,8]

- **Diabetes** and amino acids: Wang et al., 2011


ARTICLES

nature medicine

Metabolite profiles and the risk of developing diabetes

Thomas J Wang[1–3], Martin G Larson[3,4], Ramachandran S Vasan[3,5], Susan Cheng[2,3,6], Eugene P Rhee[1,7,8], Elizabeth McCabe[2,3], Gregory D Lewis[1,2,8], Caroline S Fox[3,9,10], Paul F Jacques[11], Céline Fernandez[12], Christopher J O'Donnell[2,3,8], Stephen A Carr[8], Vamsi K Mootha[8,13,14], Jose C Florez[8,13], Amanda Souza[8], Olle Melander[15], Clary B Clish[8] & Robert E Gerszten[1,2,8]

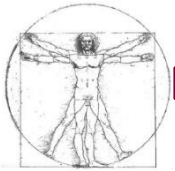
- Review: Wishart et al., 2016


REVIEWS

Emerging applications of metabolomics in drug discovery and precision medicine

David S. Wishart[1–3]

Metabolomics workflow



Biological question → **Design of Experiment**

NMR

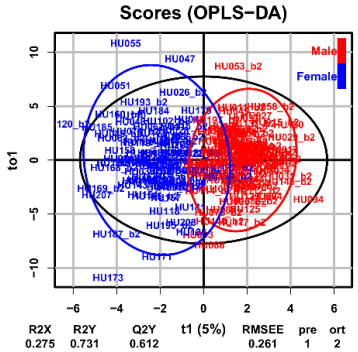
LC-MS
GC-MS

Analytical Chemistry

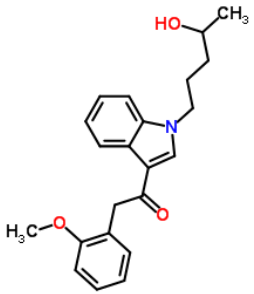
Preprocessing

mz	rt	Db_015	...	Db_068
75.0322	41.28	22162	...	48575
75.0441	174.83	1371	...	820
75.0634	56.23	49111	...	91769
...	...	...	...	...
999.6653	844.61	571	...	636
999.6759	844.61	711	...	665
999.6865	844.61	698	...	612

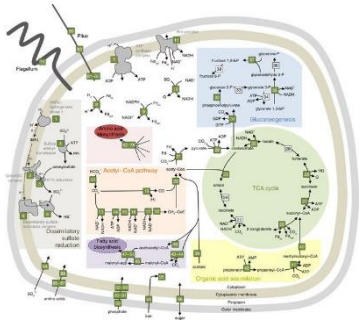
Statistics



Chemical identification



Network analysis

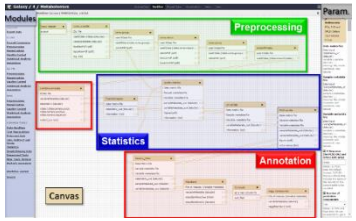


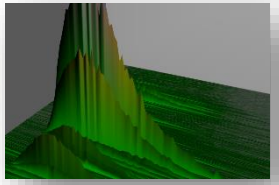
Data integration

Developments & contributions from CEA

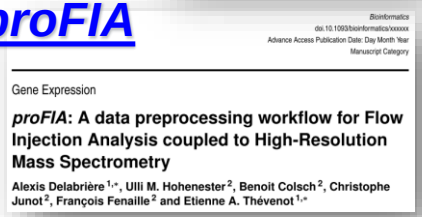
Biological question → Design of Experiment → Analytical Chemistry → Preprocessing → Statistics → Chemical identification → Network analysis → Data integration

Workflow4Metabolomics





proFIA

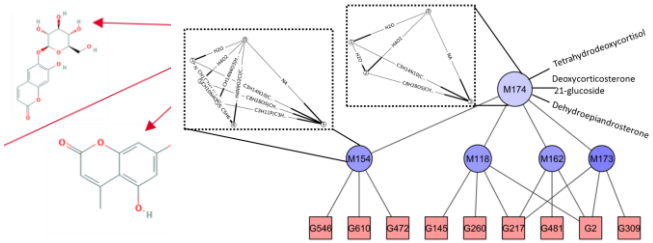

Bioinformatics
Advance Access Publication Date: Day Month Year
Manuscript Category

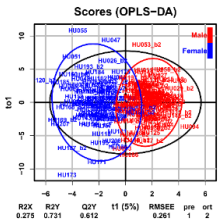
Gene Expression

proFIA: A data preprocessing workflow for Flow Injection Analysis coupled to High-Resolution Mass Spectrometry

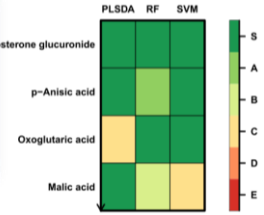
Alexis Delabrière[1,*], Ulli M. Hohenester[2], Benoit Colsch[2], Christophe Junot[2], François Fenaille[2] and Etienne A. Thévenot[1,*]

MineMS2 **rbiodb**

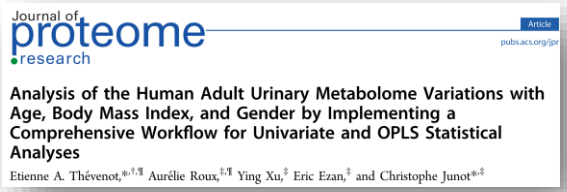




ropls



biosigner


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[∥,‡]

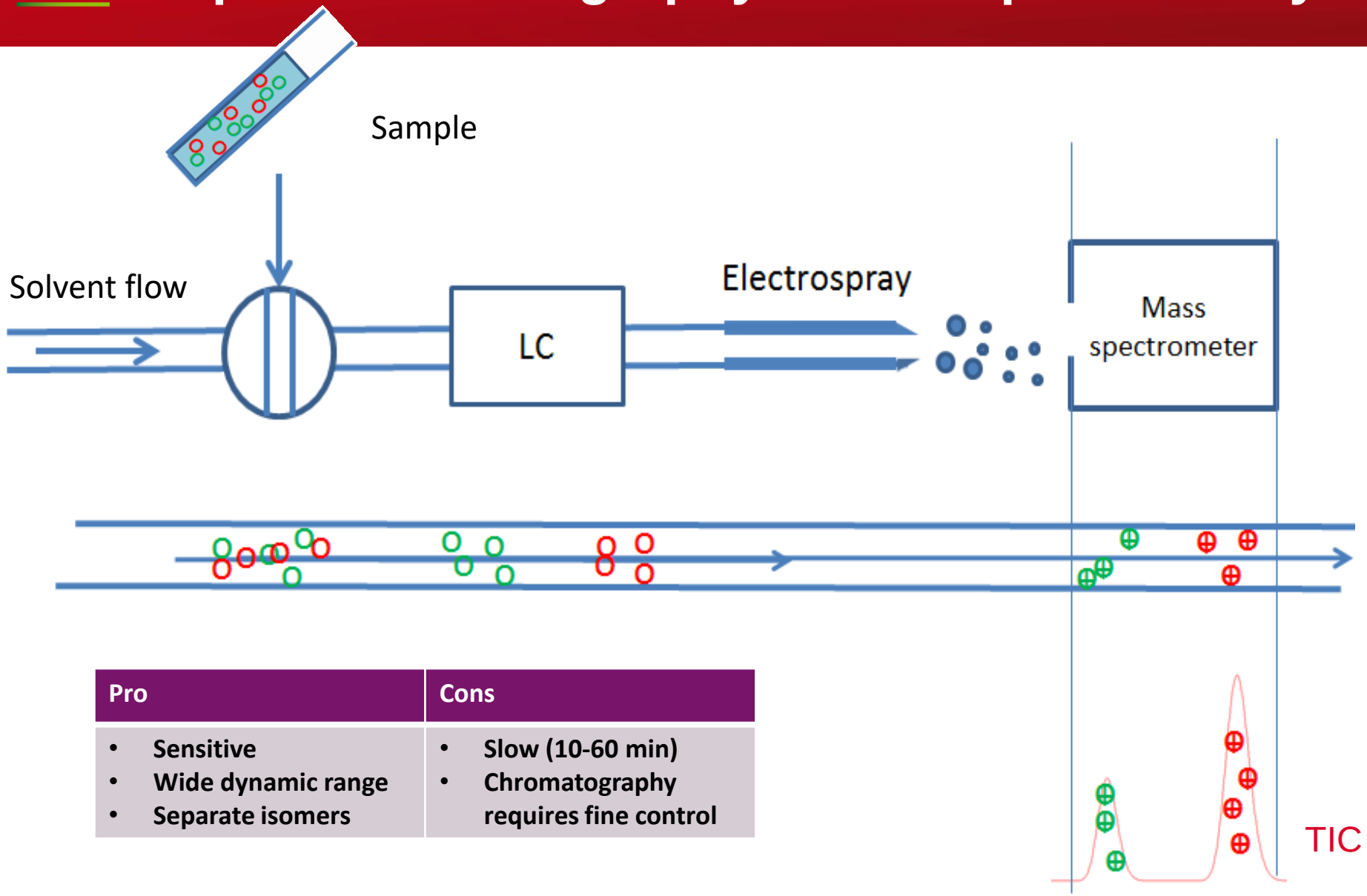

frontiers in Molecular Biosciences

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*]

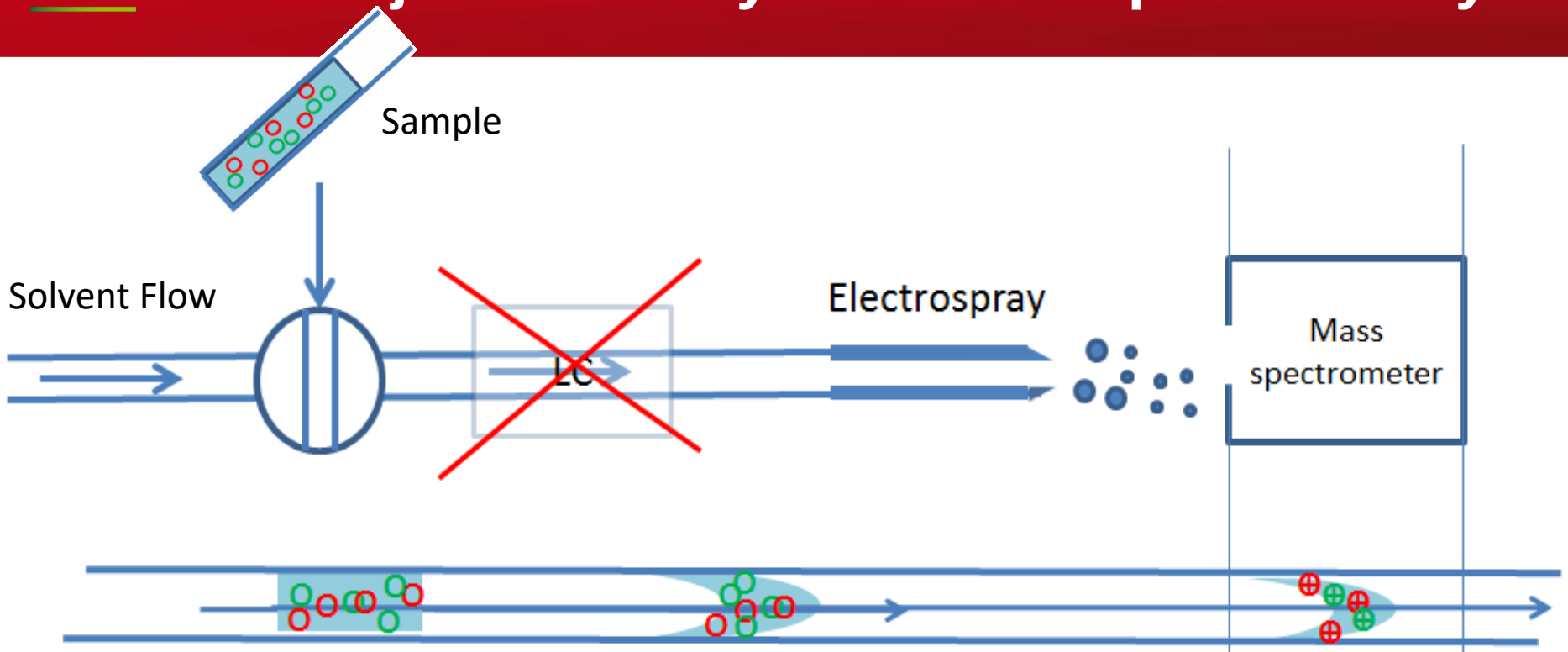
- **Data preprocessing**
  - **Flow injection analysis**

Liquid chromatography - mass spectrometry



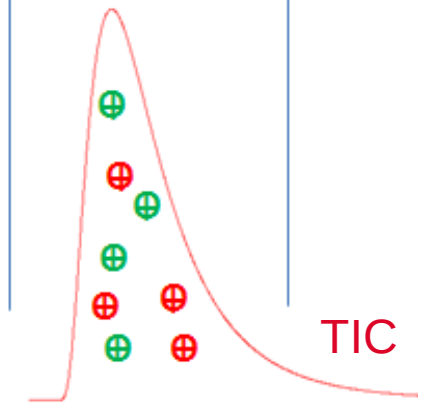
Pro	Cons
• Sensitive • Wide dynamic range • Separate isomers	• Slow (10-60 min) • Chromatography requires fine control

Flow injection analysis - mass spectrometry



	Pro	Cons
LC-MS	• Sensitive • Wide dynamic range	• Slow (10-60 min) • Expensive and difficult chromatography
FIA-MS	• Fast (1-3 min)	• Matrix effect • Isomers not separated

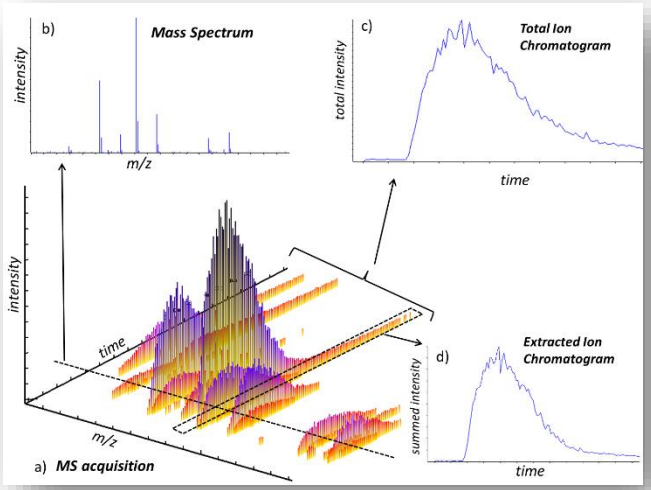
FIA is adapted to high-throughput screening



The *proFIA* workflow

Alexis Delabrière

Raw files



Detection of m/z bands → Noise estimation → Peak model estimation → Signal filtering → Peak alignment → Missing values imputation

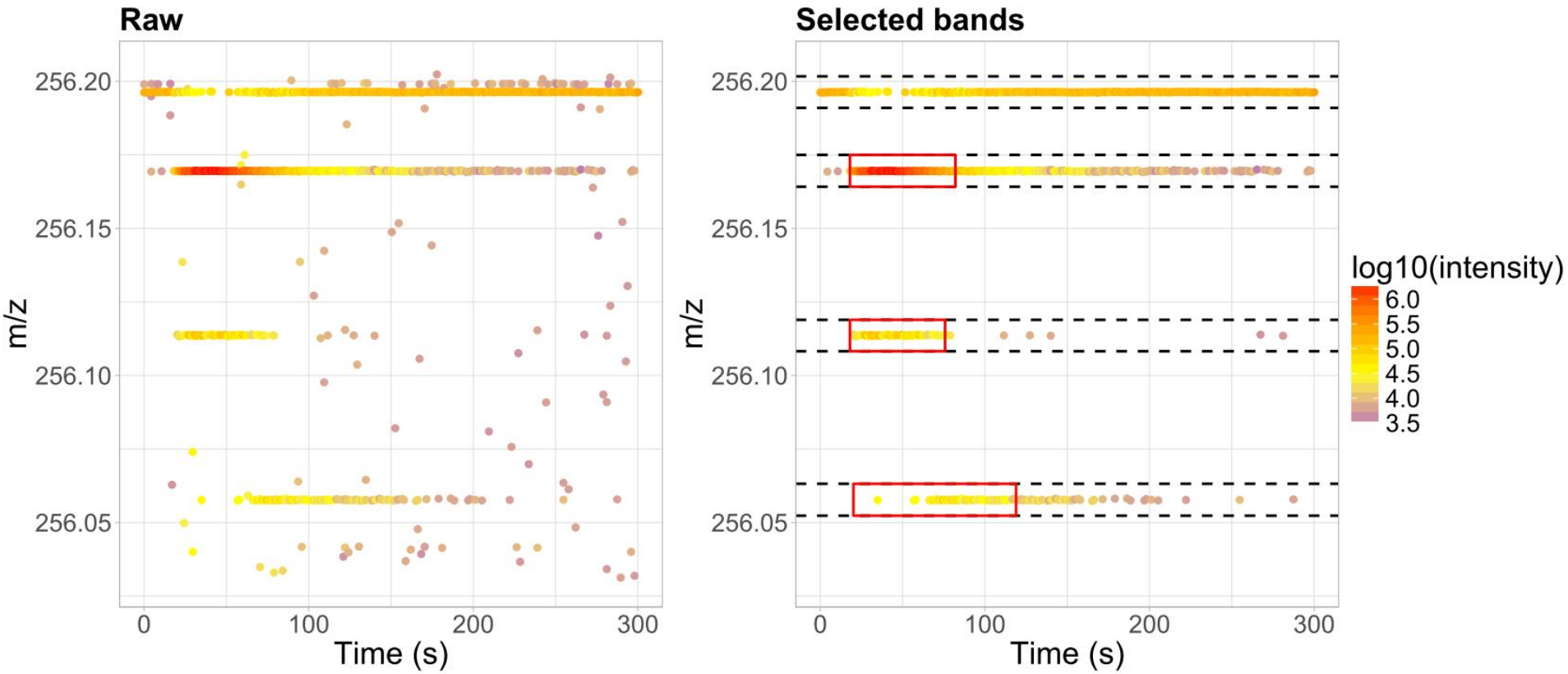
Original work
Major modification of existing algorithm

Variable by sample table of peak intensities

	1	2	3	4	5	6
1	mzMed	mzMin	mzMax	meanSolvent	corMean	UlH344_12_A
2	163.02777	163.02767	163.02782	845186.744	0.42770334	2105187
3	163.03893	163.03889	163.039	0	0.71292516	10484
4	163.11577	163.11562	163.11586	0	0.54386442	183310
5	164.02923	164.02919	164.0294	0	0.6646757	119575

Delabriere *et al.* (2017). proFIA: A data preprocessing workflow for Flow Injection Analysis coupled to High-Resolution Mass Spectrometry. *Bioinformatics*. 33:3767-3775.

Detection of m/z bands

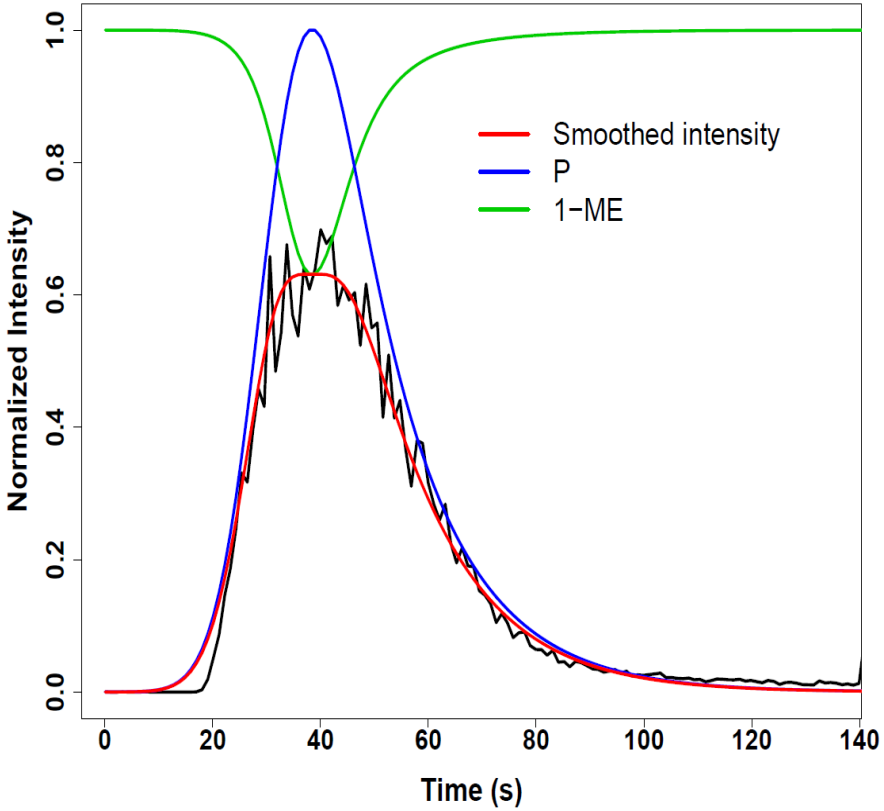


Peak model estimation

➢ We proposed a model based on Kolev (1994) and Nanita (2012)

With :

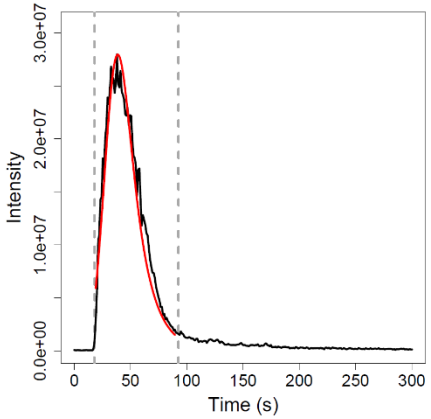
- I_A the observed intensity
- k_A a constant specific to the molecule
- P is exponentially modified gaussian
- ME_A is a second order exponential
- B_A is the baseline constant for analyte
- ϵ is the heteroscedastic noise



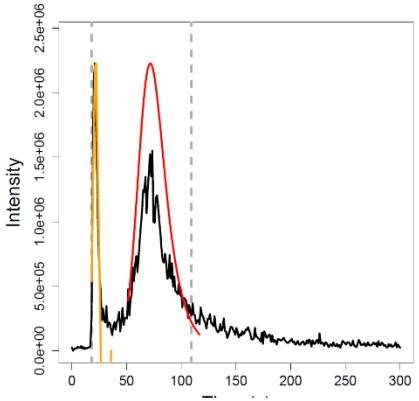
➢ Intense peaks without baseline are selected and a regression is performed leading to a peak model P

Signal filtering

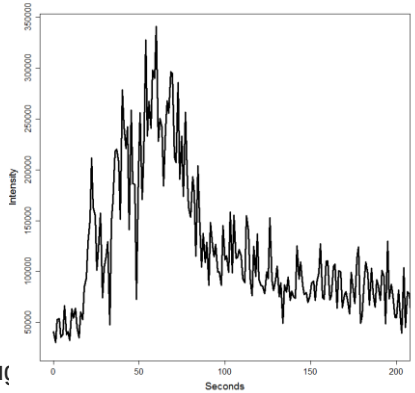
- This peak model is used to perform matching filtration on the signal



- The match can be extended if a second maximal is found on the filter. If not, a triangular filter is used for coarser grain

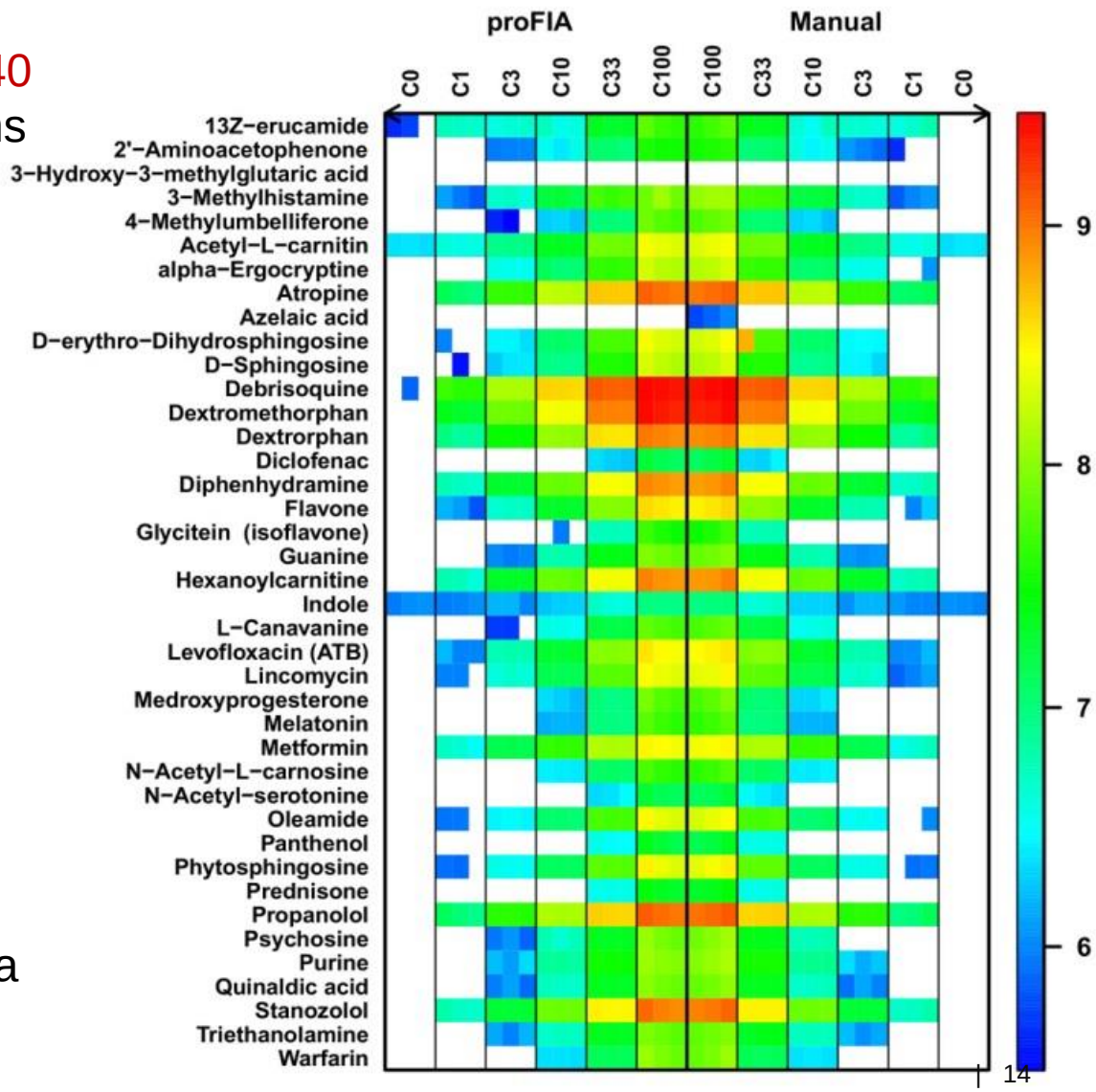


- A statistical test has been developed to discard signals too close to the baseline



Application to metabolomics data

- Dataset:
  - plasma sample spiked with 40 molecules at 6 concentrations
- Running time:
  - < 15 s per file
- Comparison with manual integration:
  - precision of 0.96
  - recall of 0.98
  - mean intensity error < 5%
- Annotation:
  - 211 signals out of 1082 had a unique match on HMDB



The *proFIA* software

- R package: Bioconductor (DOI:10.18129/B9.bioc.proFIA)

- Galaxy tool: Toolshed, Workflow4Metabolomics, PhenoMeNal

- Publication: *Bioinformatics* (DOI:10.1093/bioinformatics/btx458)

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Gene Expression

proFIA: A data preprocessing workflow for Flow Injection Analysis coupled to High-Resolution Mass Spectrometry

Alexis Delabrière [1,*], Ulli M. Hohenester [2], Benoit Colsch [2], Christophe Junot [2], François Fenaille [2] and Etienne A. Thévenot [1,*]

LC-MS

Preprocessing

xcms.xcmsSet Filtration and Peak Identification using xcmsSet function from xcms R package to preprocess LC/MS data for relative quantification and statistical analysis

xcms.xcmsSet Merger Merge xcms.xcmsSet xset in one to be used by group

xcms.group Group peaks together across samples using overlapping m/z bins and calculation of smoothed peak distributions in chromatographic time.

xcms.retcor Retention Time Correction using retcor function from xcms R package

xcms.fillPeaks Integrate a sample's signal in regions where peak groups are not represented to create new peaks in missing areas

xcms.summary Create a summary of XCMS analysis

CAMERA.annotate CAMERA annotate function. Returns annotation results (isotope peaks, adducts and fragments) and a diffreport if more than one condition.

CAMERA.combinexsAnnos Wrapper function for the combinexsAnnos CAMERA function. Returns a dataframe with recalculated annotations.

proFIA Preprocessing of FIA-HRMS data

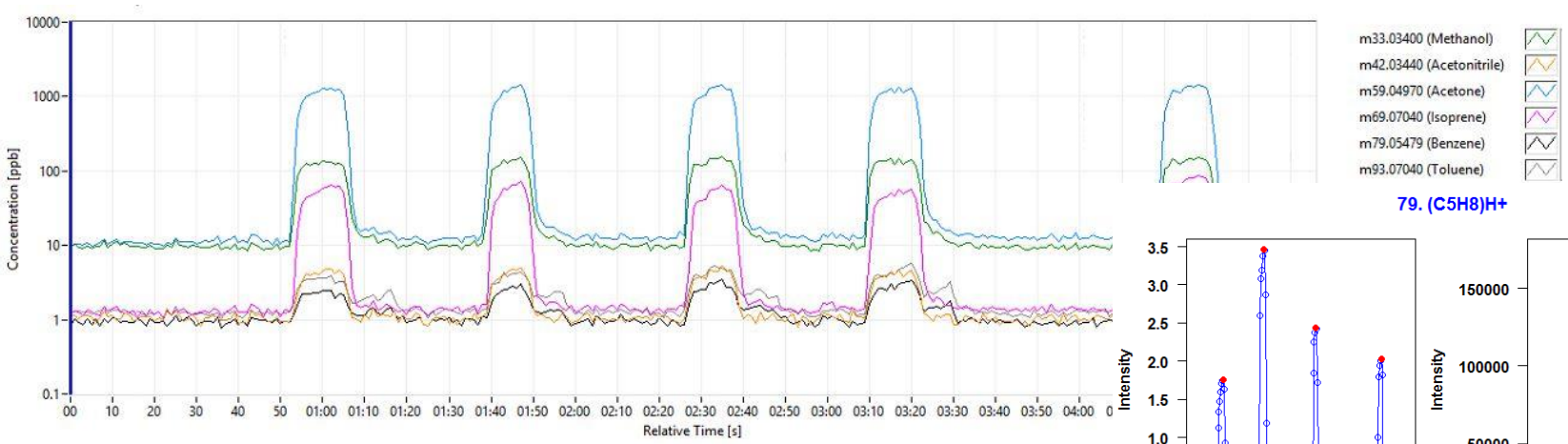
metaMS.runGC GC-MS data preprocessing using metaMS package

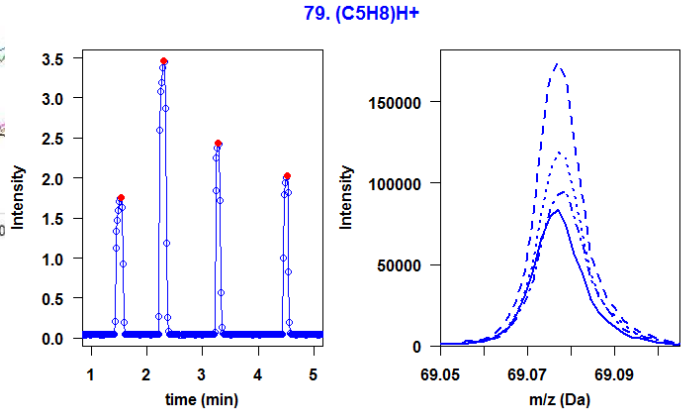
Breathomics: real time analysis of exhaled air in disease and response to treatments

- Objective: comprehensive analysis of metabolism-derived Volatile Organic Compounds (VOCs)
- Technology: PTR-TOF-MS at the patient bedside (Foch Hospital)
- Project: develop innovative algorithms and software environment for the processing of real-time analysis of VOCs in exhaled air

HOPITAL FOCH

S. Grassin-Delyle





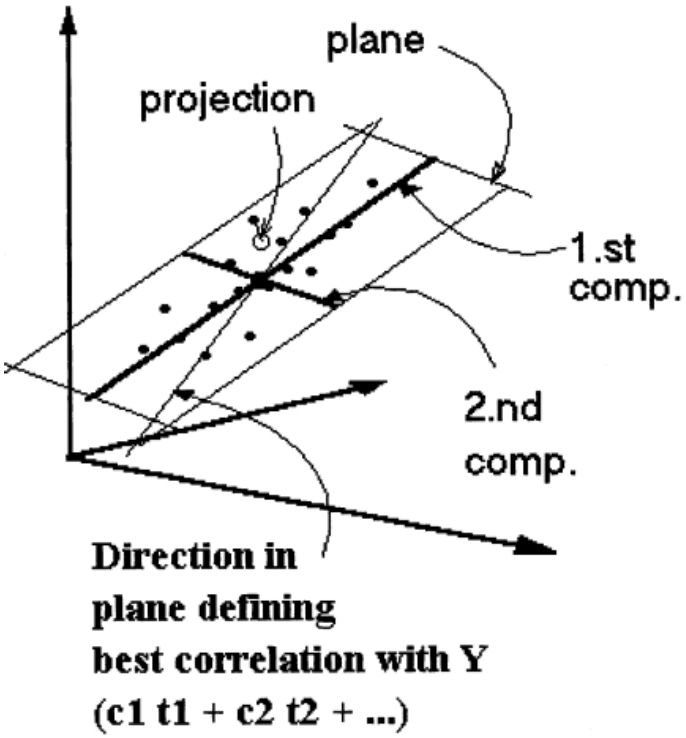
Omics data analysis

➢ Statistical analysis

- Orthogonal Partial Least Squares

Partial Least Squares (PLS)

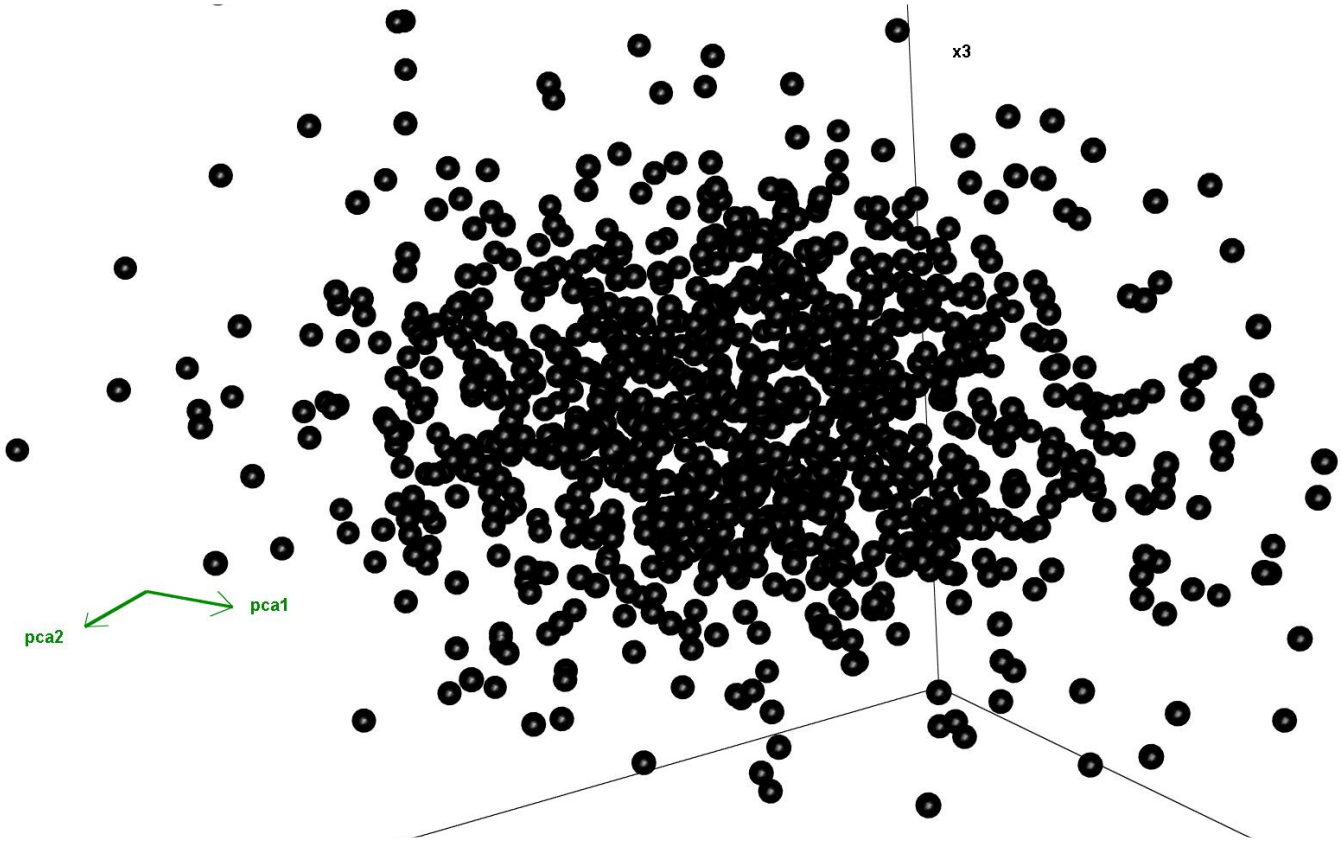
- Multivariate regression approach
- Handle data sets
  - of high dimension (n < p)
  - correlated variables
  - including missing values
- Based on latent variables
  - maximizing covariance with the response Y
- Developed by Wold H. and S.
- Can be used for classification (PLS-DA)



Wold *et al.* (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* 58, 109–130.

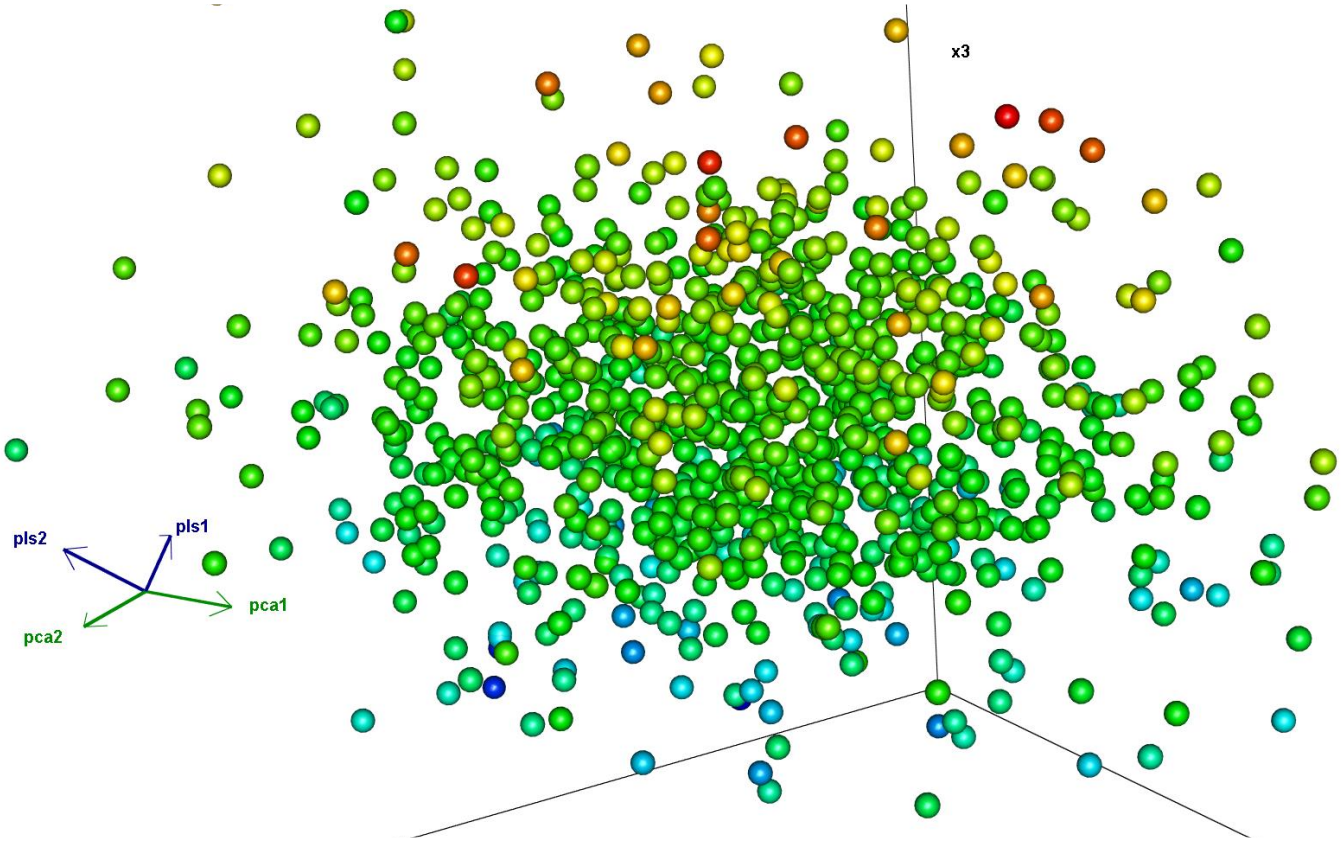
PLS vs PCA

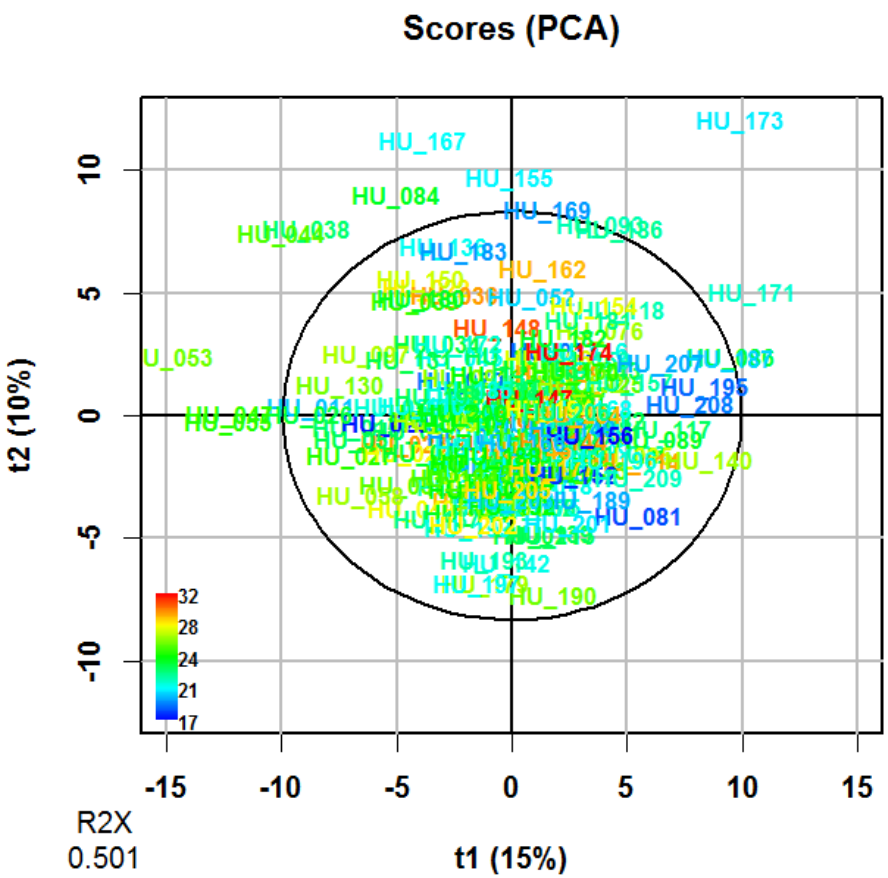
- PCA finds the directions of maximum variance

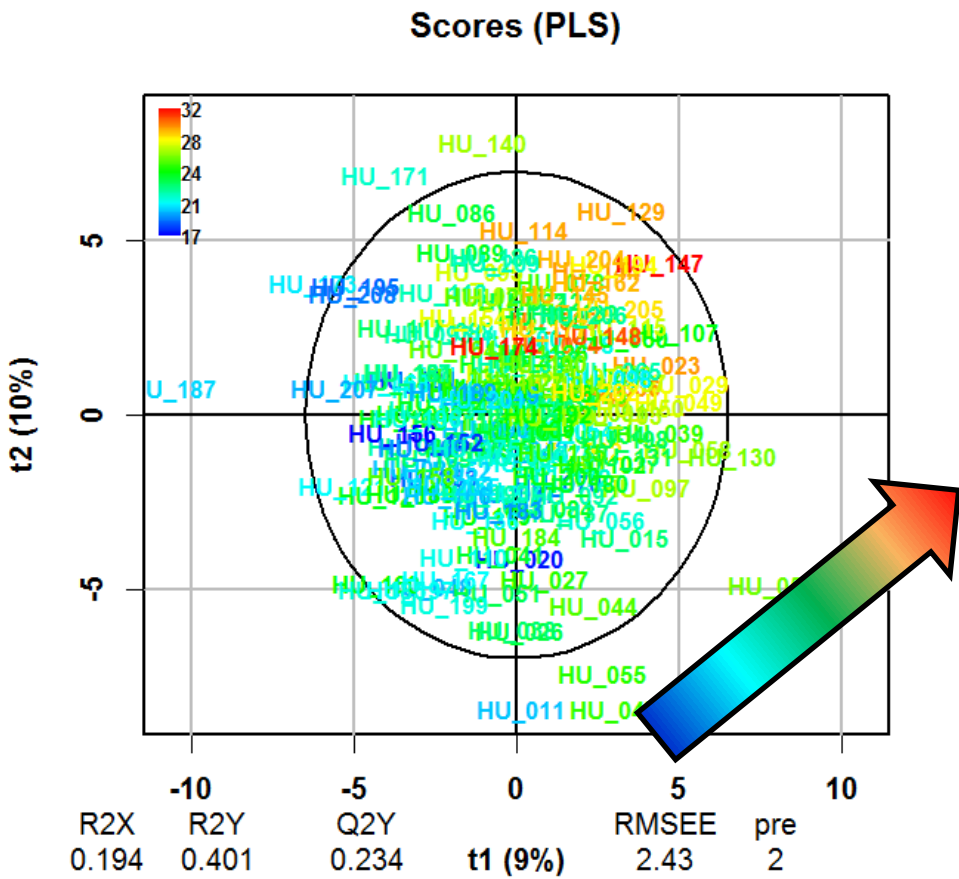


PLS vs PCA: score plots

- PLS includes the labels into the model

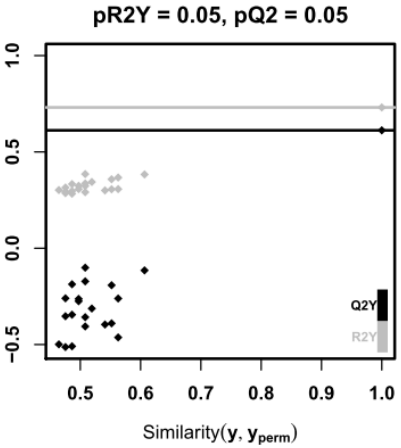


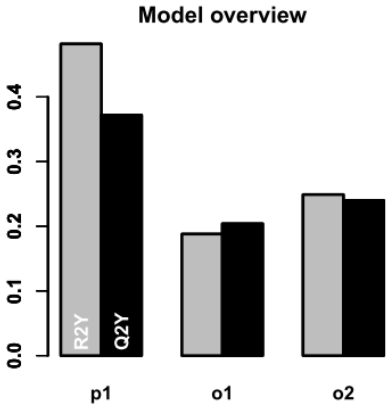

Scores (PCA)
t2 (10%)
t1 (15%)
R2X
0.501
32
28
24
21
17


Scores (PLS)
t2 (10%)
t1 (9%)
R2X
0.194
R2Y
0.401
Q2Y
0.234
RMSEE
2.43
pre
2
32
28
24
21
17

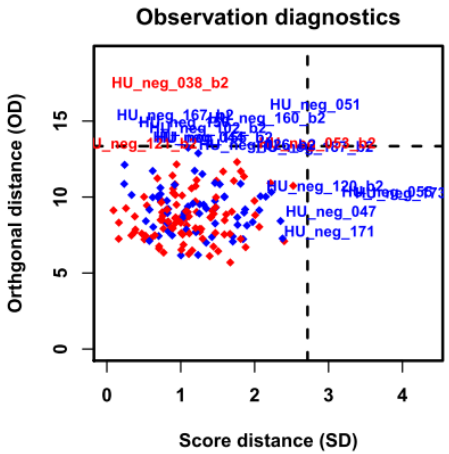
ropls package: R implementation of the (O)PLS(-DA) modeling algorithms

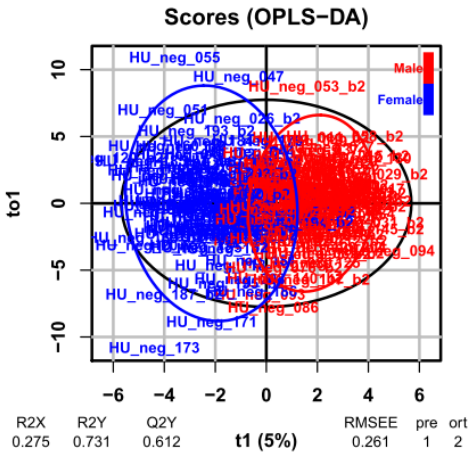
- Full diagnostics
  - outliers
  - permutation testing





- Full numerical and graphical results
  - R2X, R2Y, Q2Y
  - VIPs

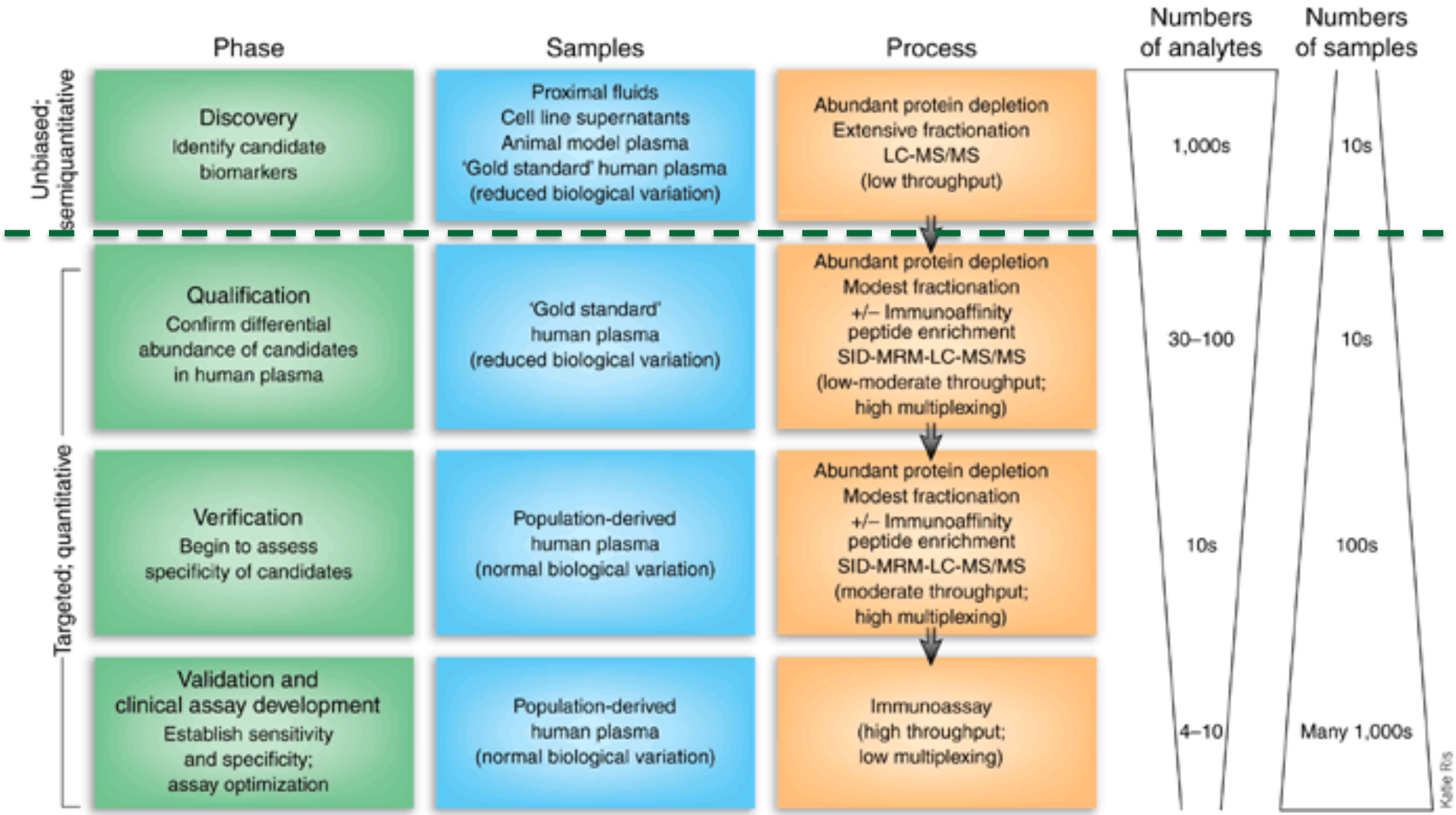




Thévenot *et al.* (2017). Analysis of the human adult urinary metabolome variations with age, body mass index and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. *J. Proteome Res.* 14:3322-3335.

- **Statistical analysis**
  - **Feature selection**

Feature selection: from biomarker discovery to clinical diagnostics



Rifai *et al.* (2006). Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* 24:971-983.

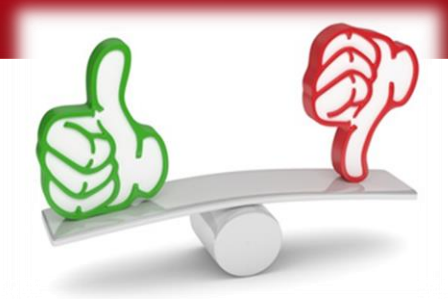
Feature selection: objectives

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

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



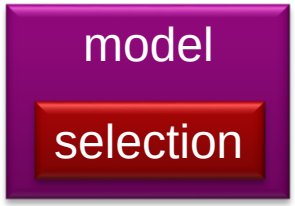
- filter (threshold criterion)
  - e.g., t-test

fast **threshold?**

- wrapper (iterative selection)
  - e.g., SVM RFE, Genetic Algorithm

interaction with classifier **computation intensive**

- embedded (penalization constraint)
  - e.g., Lasso, Elastic Net



fast **stability**

- **Statistical analysis**
  - **The *biosigner* approach**

The *biosigner* wrapper algorithm

Philippe Rinaudo

- Objective: select only features which significantly contribute to the performance of the classifier
- Method: features are significant if the prediction accuracy decreases after random permutation of their values for in test samples
- Algorithm:

1. generate k train/test subset by resampling
2. build the models and rank the variables
3. find the largest non-significant feature subset (half-interval search)
4. repeat steps (1-3) on the dataset restricted to the significant features until the selection is stable (all features are significant)

Rinaudo *et al.* (2016). *biosigner*: a new method for the discovery of significant molecular signatures from omics data. *Front. Mol. Biosci.* 3.

1.1 Generate k subsets (bootstrap resampling)

1.1 Generate k subsets (bootstrap resampling)

1.2 Train F_k models (e.g. PLS-DA)

$$y_k = F_k (X_k)$$

$train_k$

1.3 Find the largest subset of non-significant features

a) Rank the features (e.g. VIP)

1.3 Find the largest subset of non-significant features

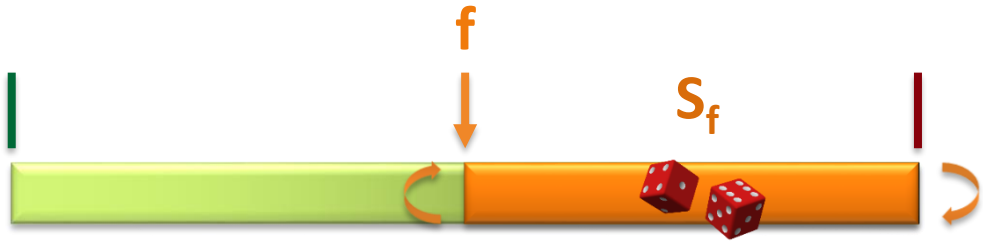
b) Find the feature f of lowest rank such that the subset of all features of higher ranks is not significant:

i) set f to the feature of mean rank

1.3 Find the largest subset of non-significant features

b) Find the feature f of lowest rank such that the subset of all features of higher ranks is not significant:

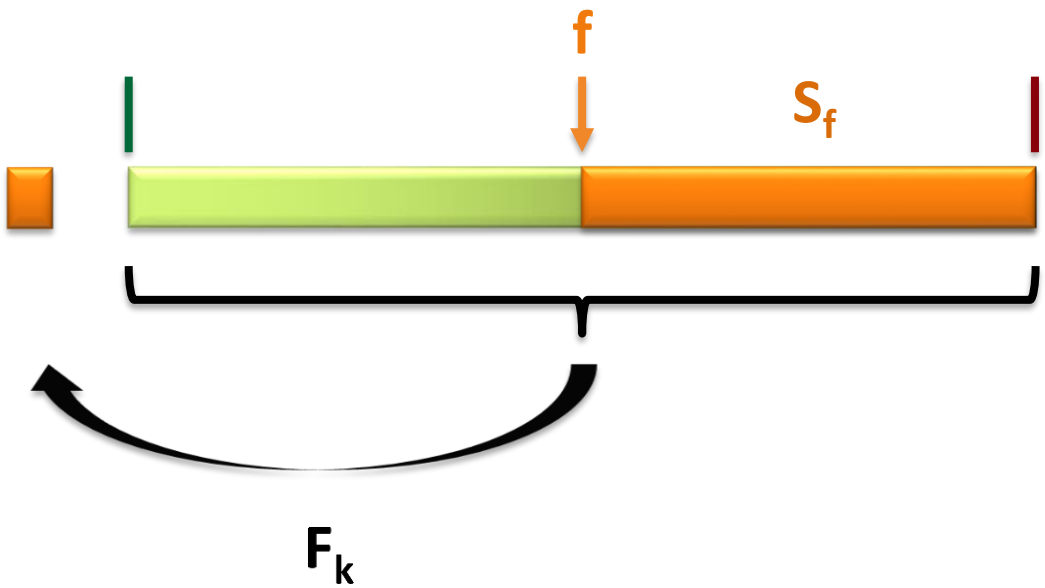
ii) permute in the test set all features of higher rank (ie features in S_f)



1.3 Find the largest subset of non-significant features

b) Find the feature f of lowest rank such that the subset of all features of higher ranks is not significant:

iii) compare the accuracies of the predictions after permutation



1.3 Find the largest subset of non-significant features

b) Find the feature f of lowest rank such that the subset of all features of higher ranks is not significant:

iii-a) accuracy → or ↗

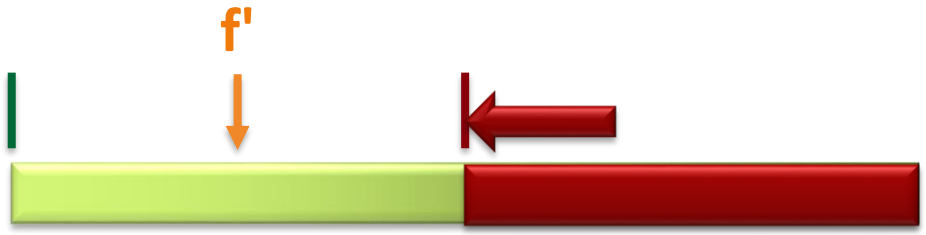
=> S_f does not contain significant features

1.3 Find the largest subset of non-significant features

b) Find the feature f of lowest rank such that the subset of all features of higher ranks is not significant:

iii-a) accuracy → or ↗

=> shift f upward to the mean rank of significant features

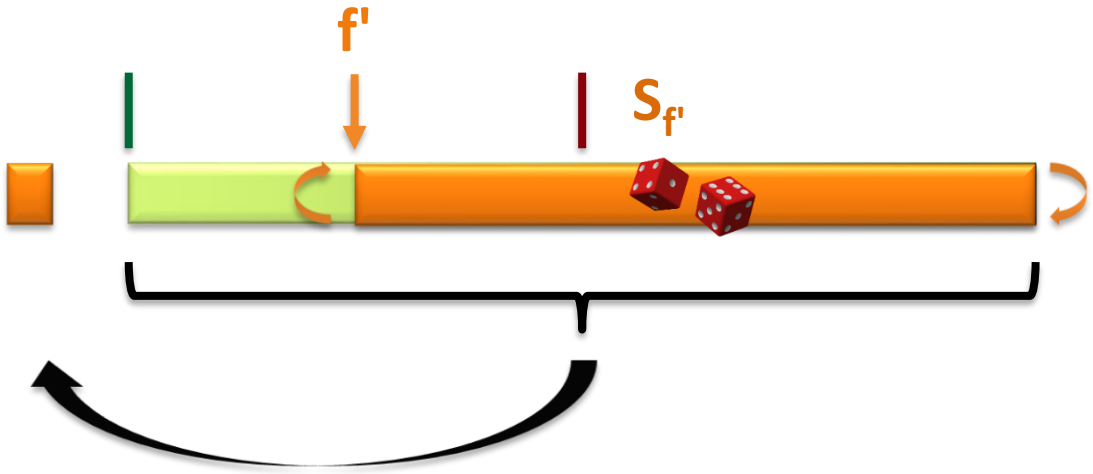


1.3 Find the largest subset of non-significant features

b) Find the feature f of lowest rank such that the subset of all features of higher ranks is not significant:

iii-a) accuracy → or ↗

=> evaluate the performance after permutation of the features in $S_{f'}$

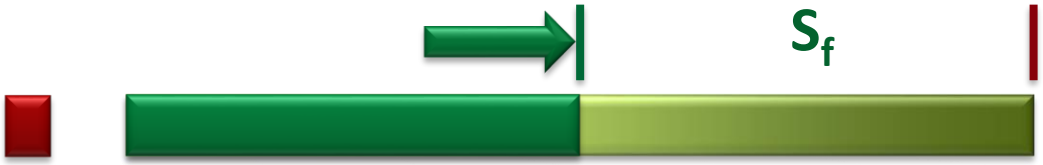


1.3 Find the largest subset of non-significant features

b) Find the feature f of lowest rank such that the subset of all features of higher ranks is not significant:

iii-b) accuracy ↘

=> S_f contains significant features

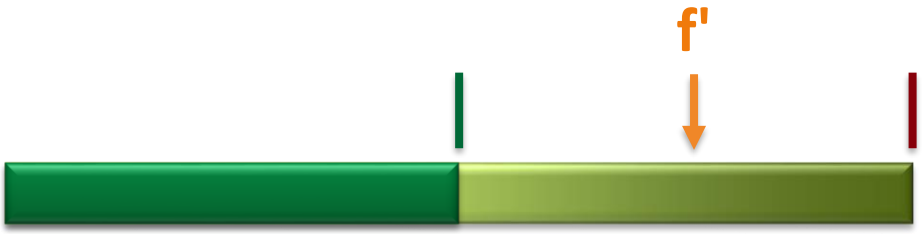


1.3 Find the largest subset of non-significant features

b) Find the feature f of lowest rank such that the subset of all features of higher ranks is not significant:

iii-b) accuracy ↘

=> shift f downward to the mean rank of non-significant features

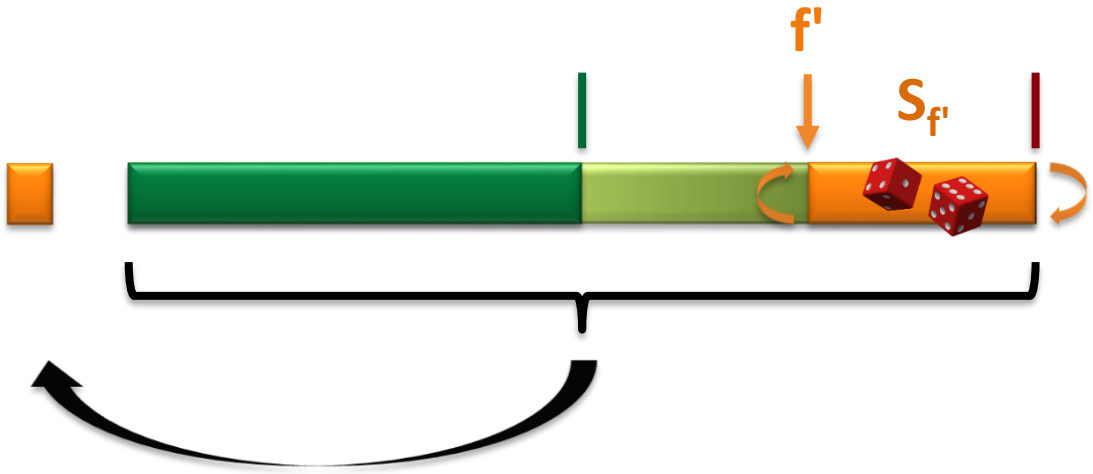


1.3 Find the largest subset of non-significant features

b) Find the feature f of lowest rank such that the subset of all features of higher ranks is not significant:

iii-b) accuracy ↘

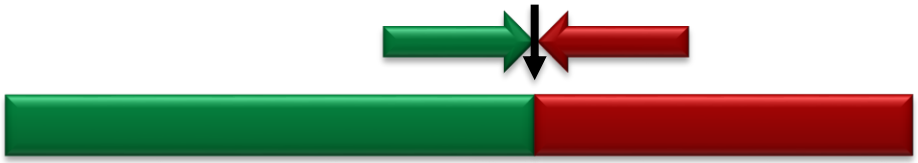
=> evaluate the performance after permutation of the features in $S_{f'}$



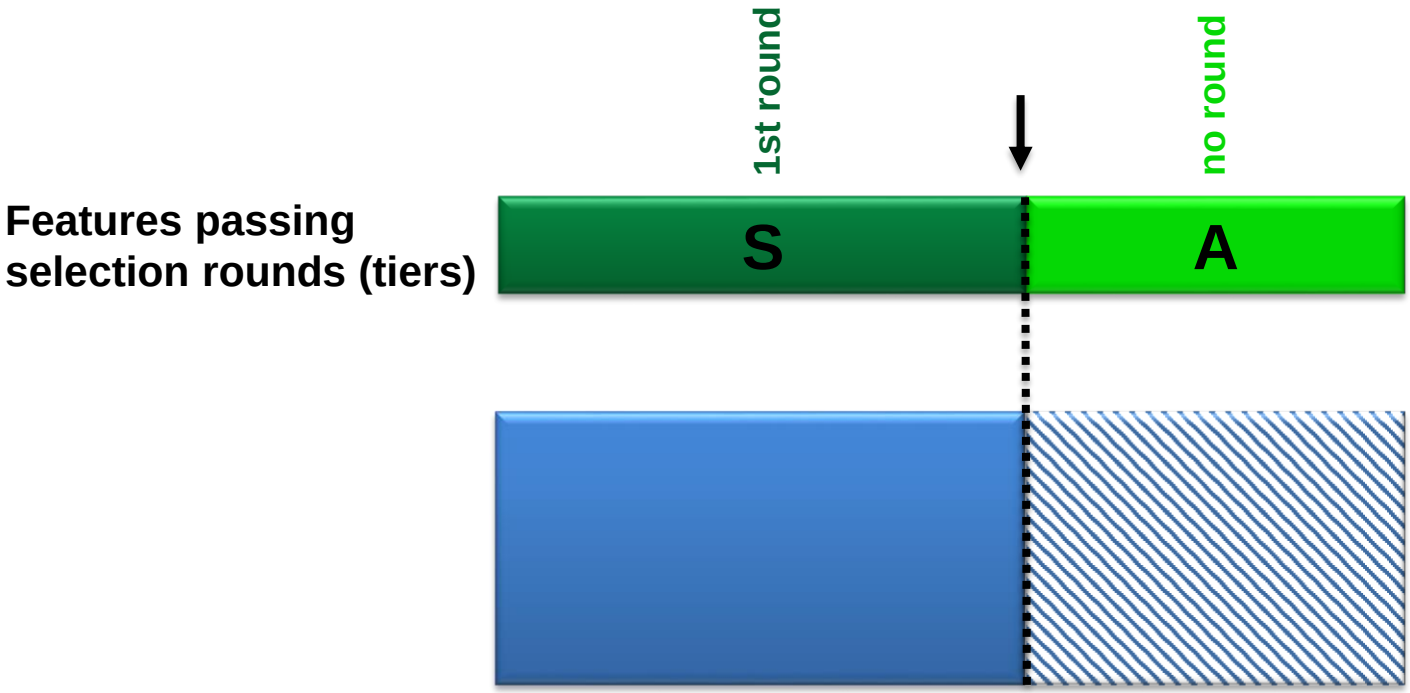
1.3 Find the largest subset of non-significant features

b) Find the feature f of lowest rank such that the subset of all features of higher ranks is not significant:

iv) stop when the upper and lower limits for f converge



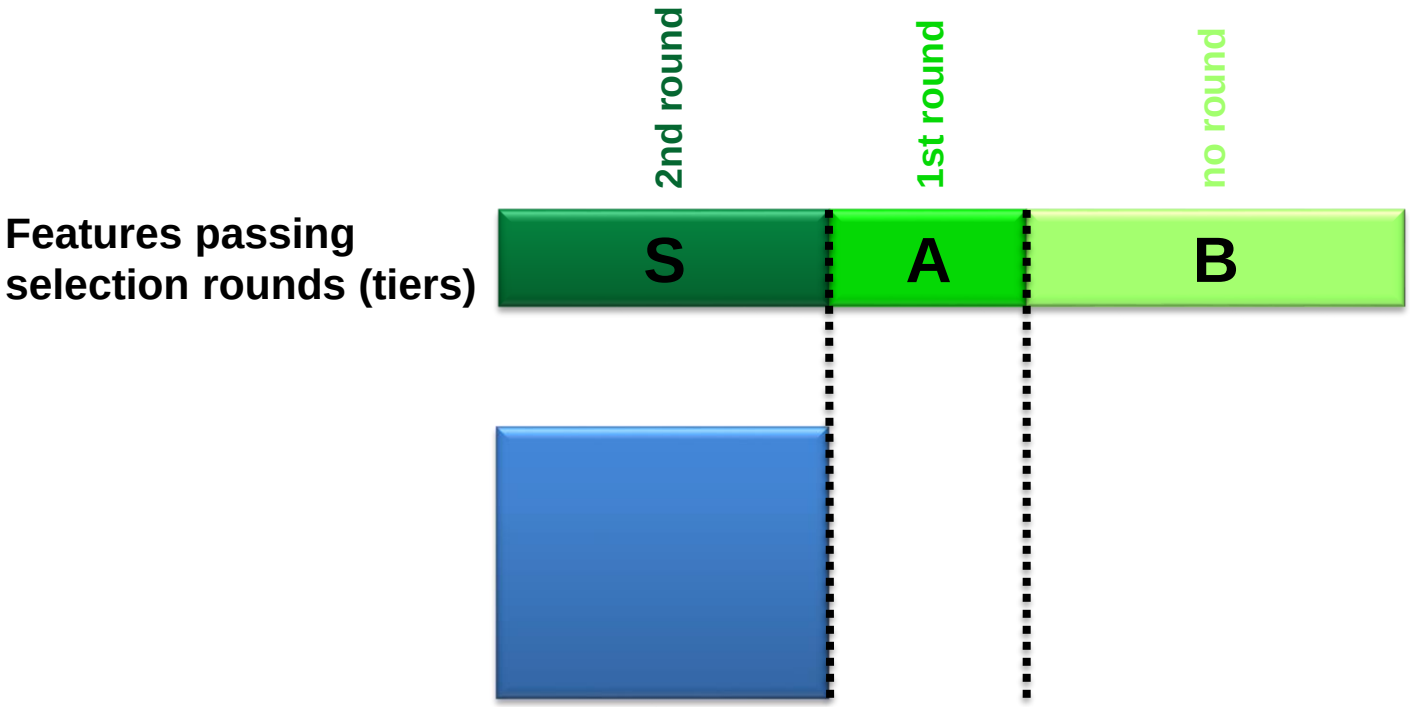
1.4 Restrict the dataset to the significant subset



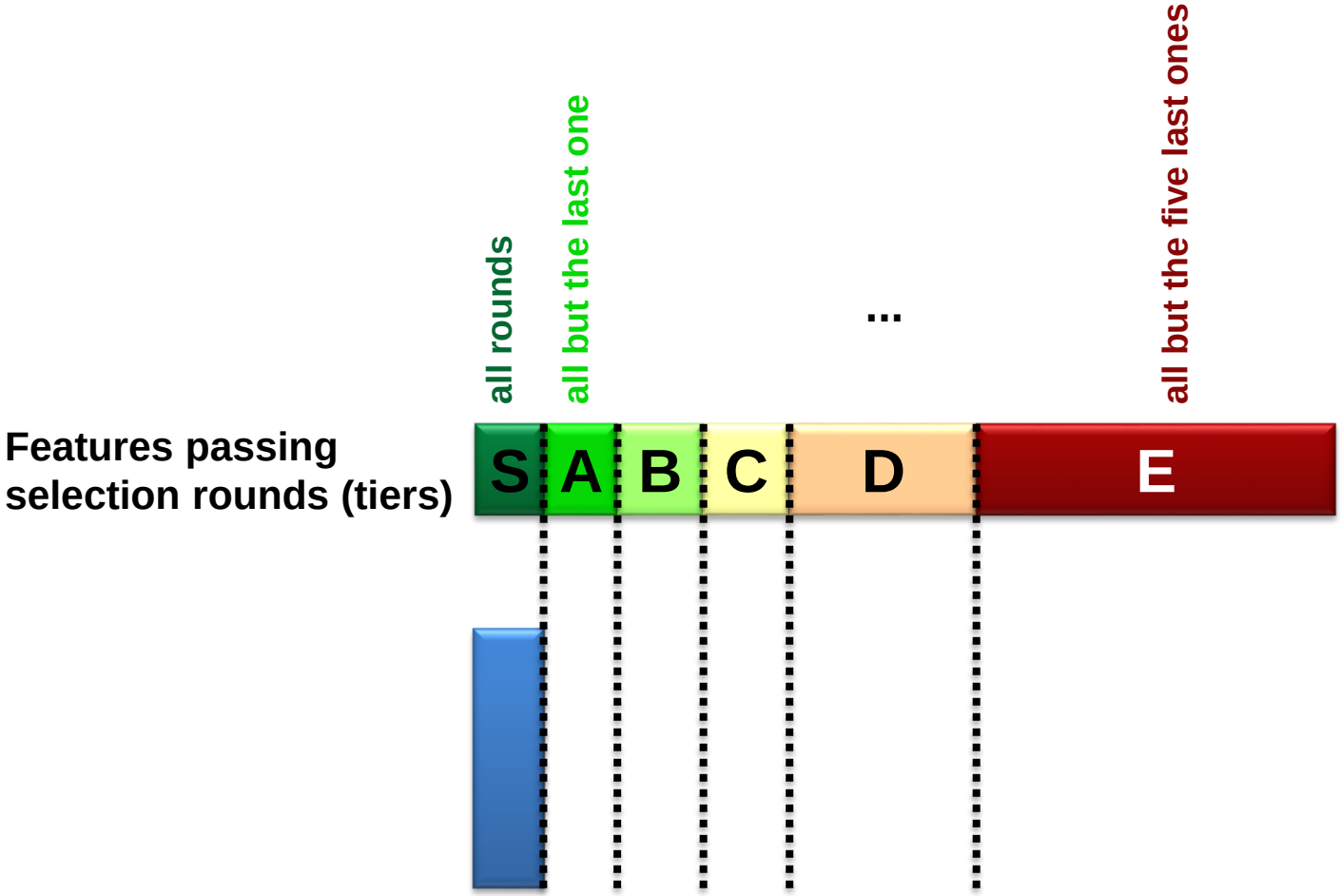
2. Repeat whole feature selection procedure on the restricted dataset

response (y) dataset(X')

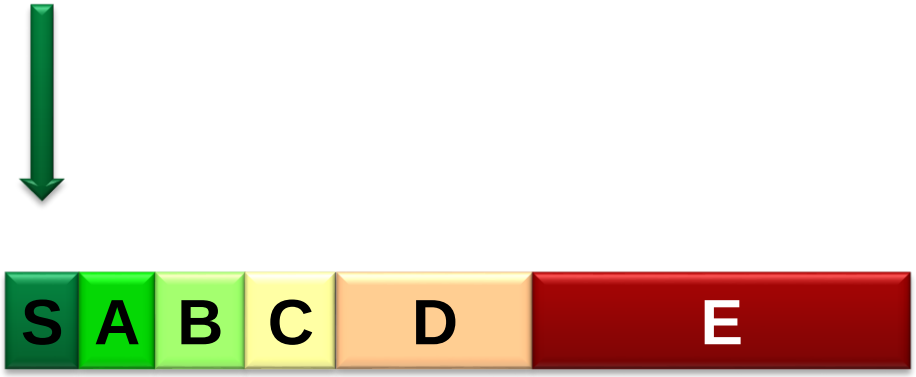
2. Repeat whole feature selection procedure on the restricted dataset



3. Stop when the signature is stable (all features significant)



Selected molecular signature



Final model

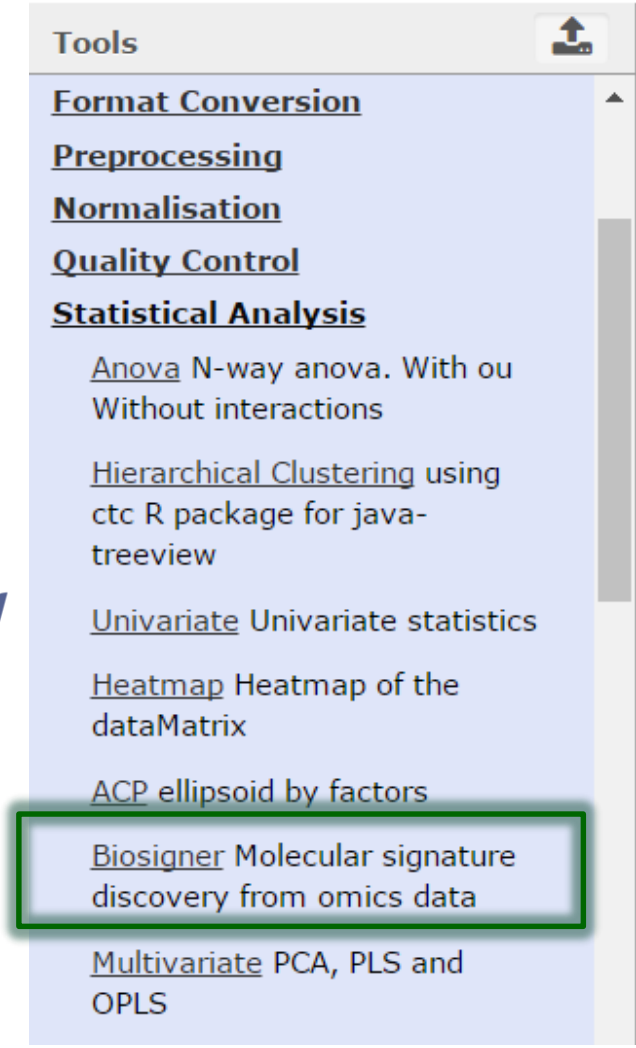
$$y = F_{final}(X_{final})$$

The *biosigner* software

- R package: Bioconductor (DOI:10.18129/B9.bioc.biosigner)

- Galaxy tool: Toolshed, Workflow4Metabolomics, PhenoMeNal

- Publication: Frontiers in Molecular Biosciences (DOI:10.3389/fmolb.2016.00026)




frontiers in Molecular Biosciences

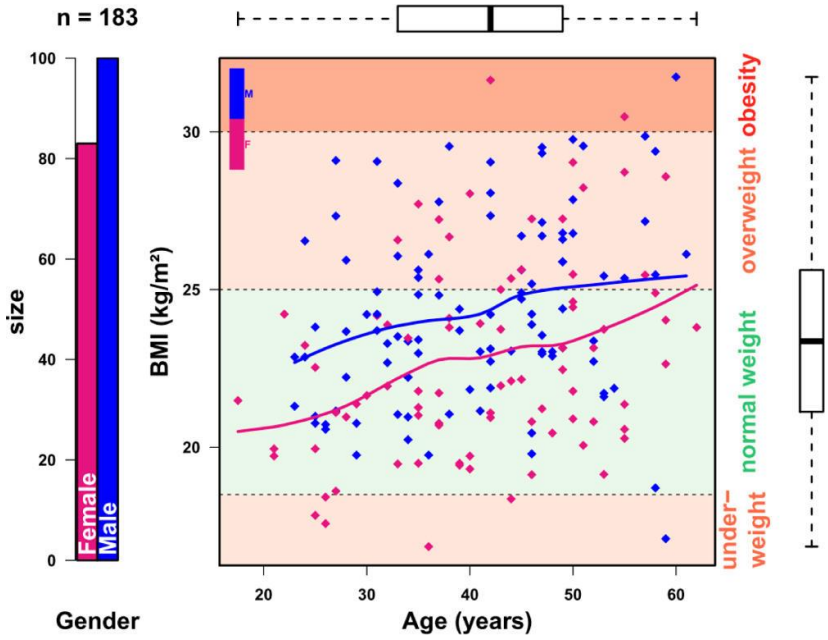
ORIGINAL RESEARCH
published: 21 June 2016
doi: 10.3389/fmolb.2016.00026

***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]**

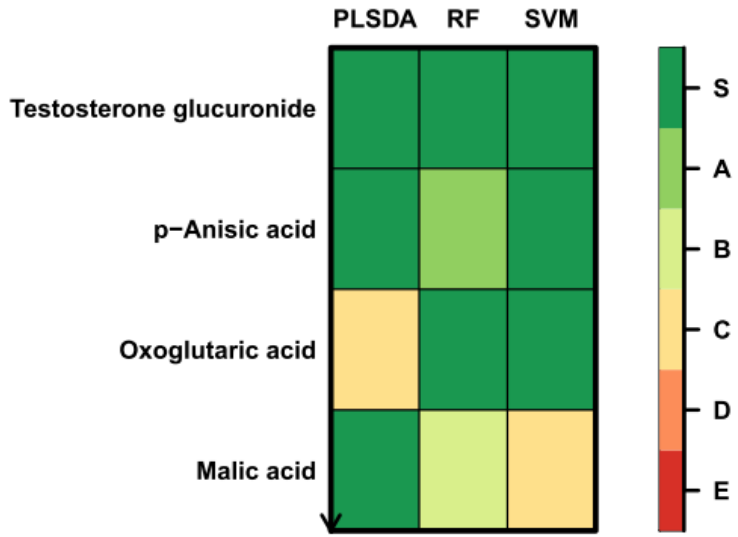
Sacurine dataset (MTBLS404)

- 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



biosigner package: model performances

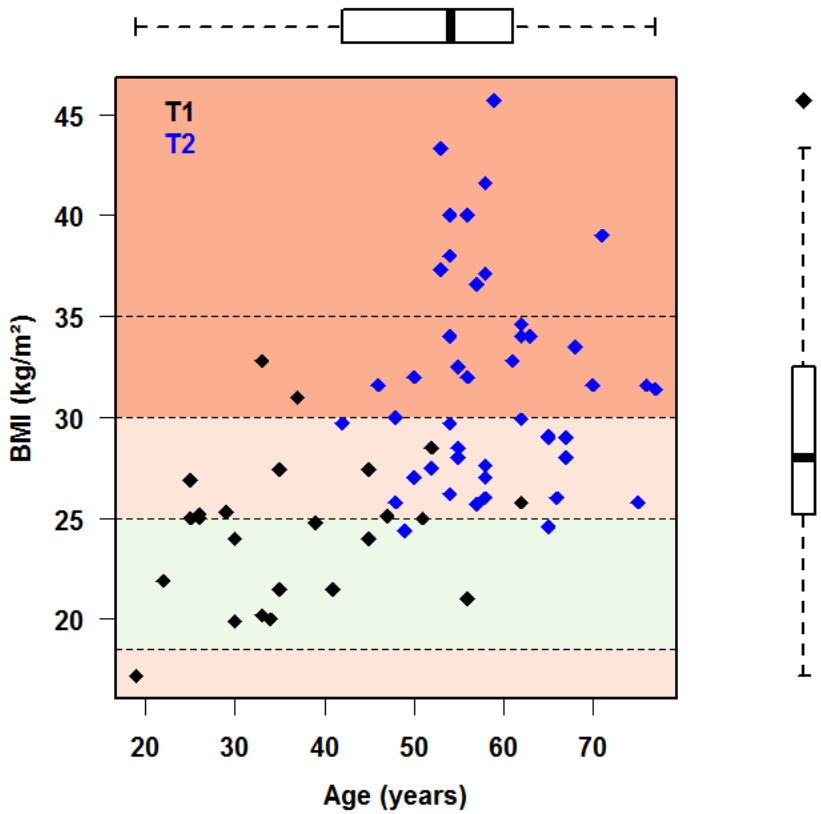
sacurine *(ropls)*



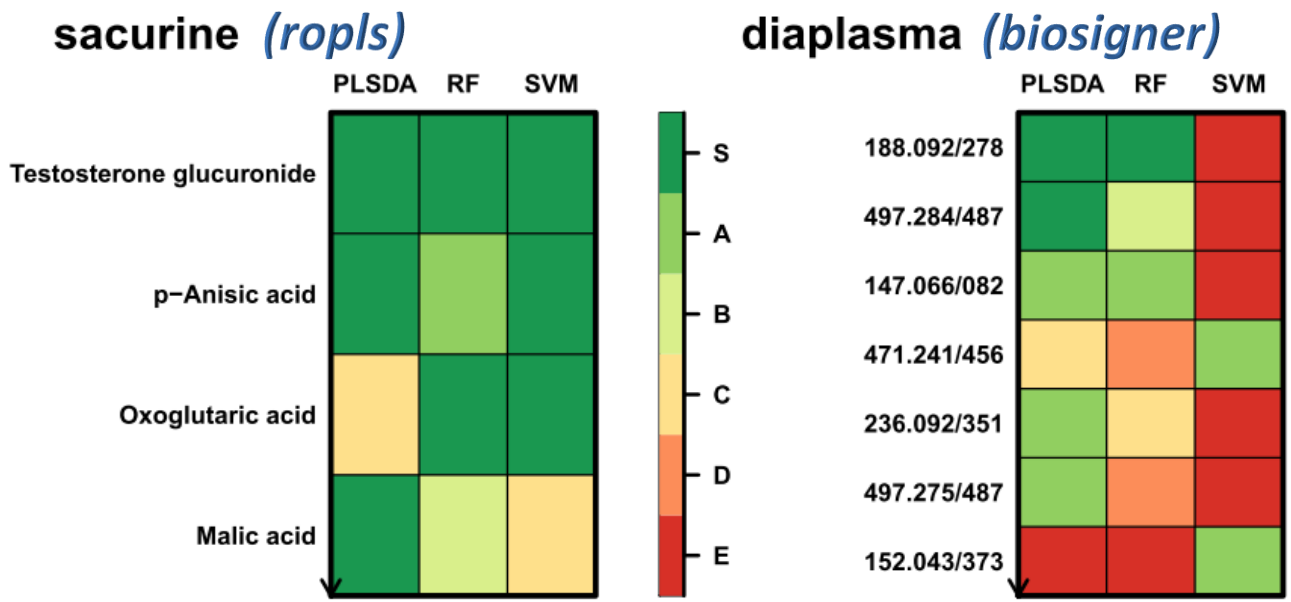
		sacurine
factor		gender
samples		183
features		109
signatures		[2-3]
performances (full -> restricted)	PLS-DA	87% -> **89%**
	Random Forest	86% -> **86%**
	SVM	88% -> **89%**

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

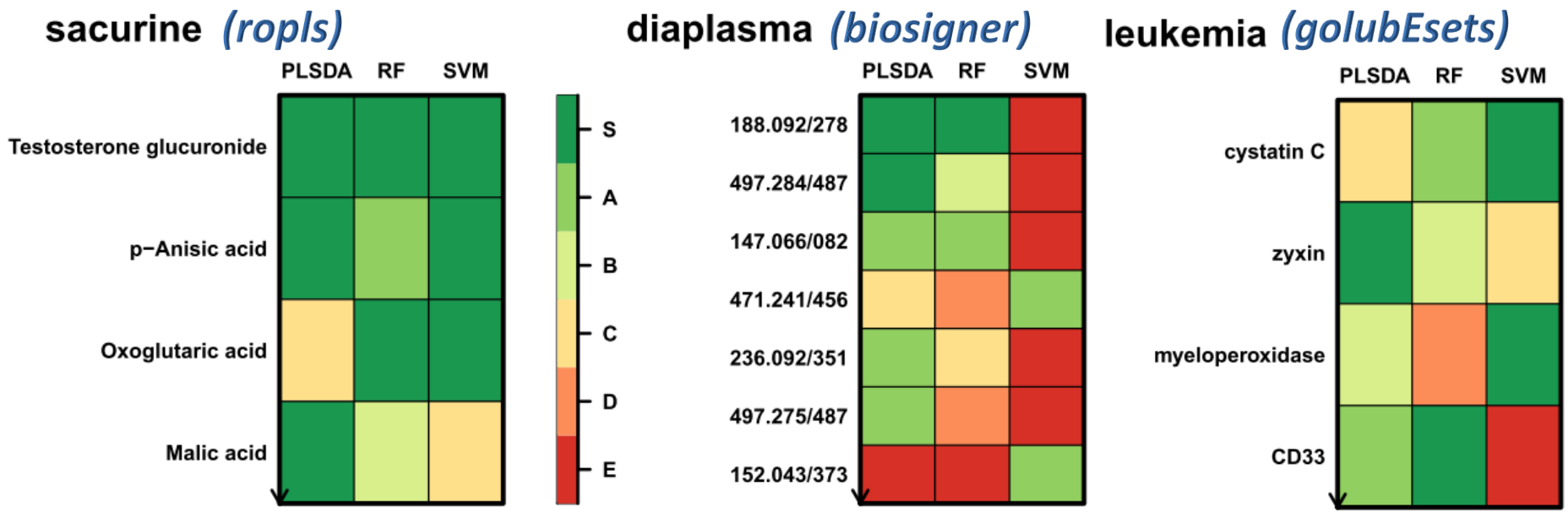


biosigner package: model performances



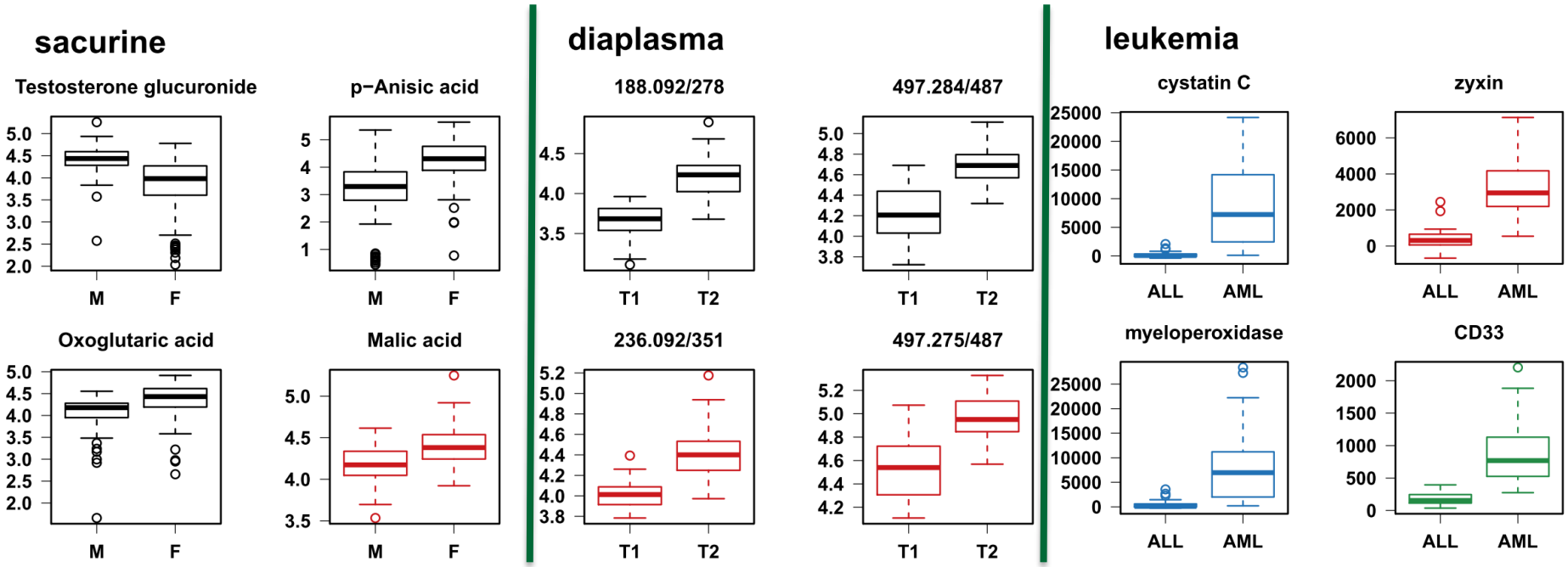
		sacurine	diaplasma
factor		gender	diabetic type
samples		183	69
features		109	5,501
signatures		[2-3]	[0-2]
performances (full -> restricted)	PLS-DA	87% -> **89%**	83% -> **91%**
	Random Forest	86% -> **86%**	81% -> **81%**
	SVM	88% -> **89%**	83% -> **na**

biosigner package: 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



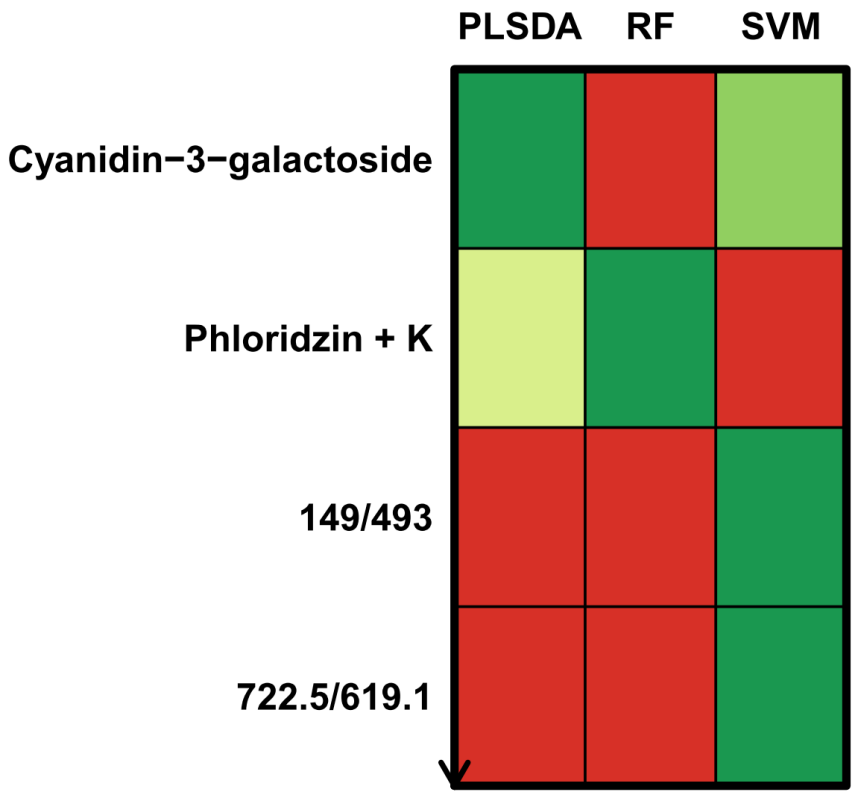
Biomarker in prostate cancer: Zhang et al. (2013). *PLoS ONE*, **8:**e65880.

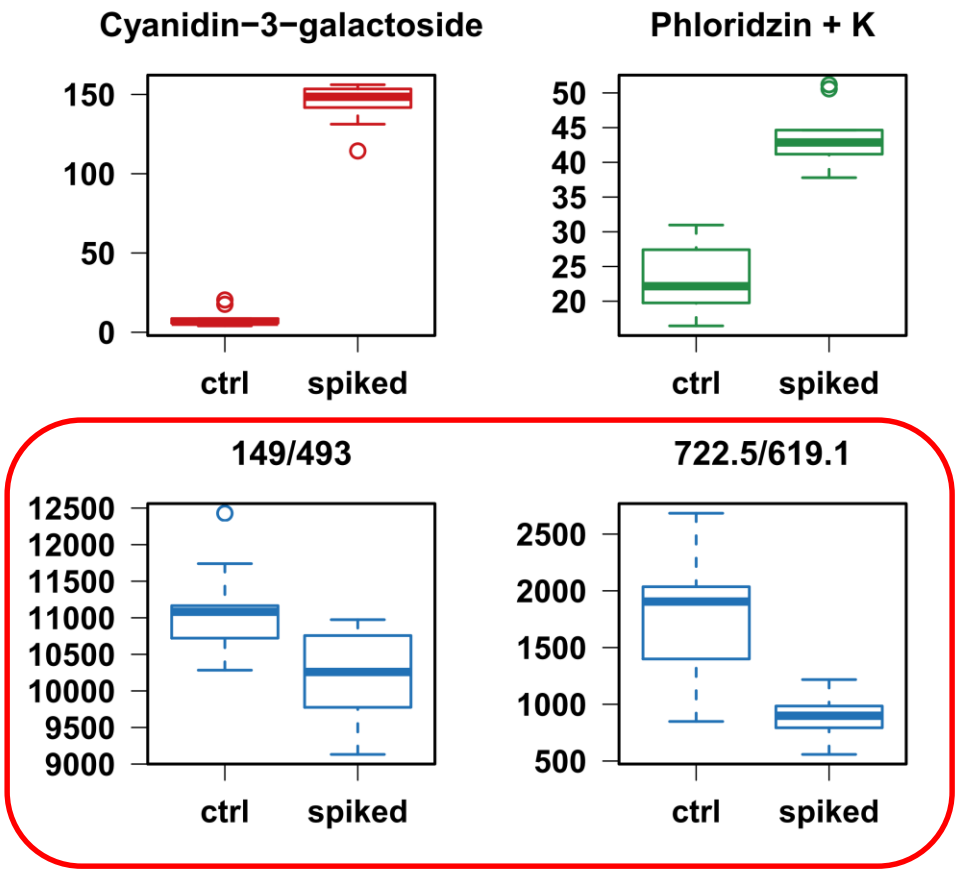
Taurochenodeoxycholic acid: variation in type 2 diabetic patients: Taylor et al. (2014). PLoS ONE, **9**:e93540.

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

Spiked apples

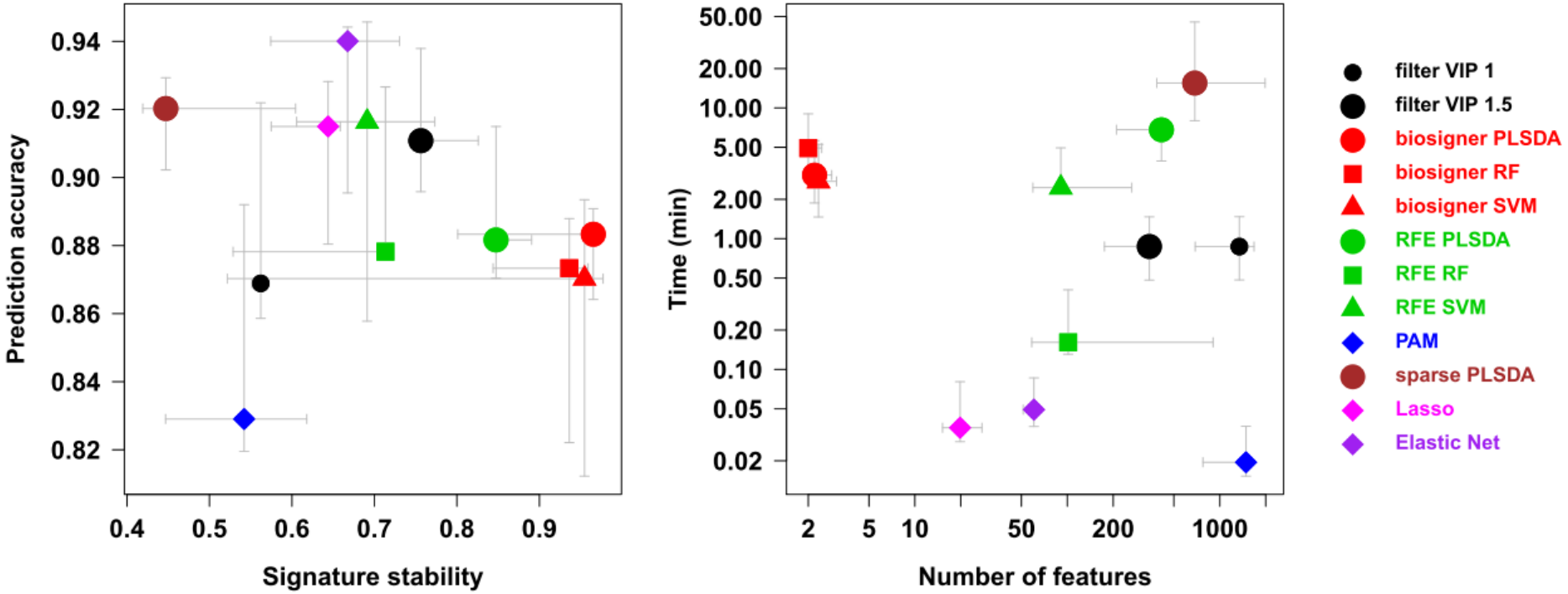
(BioMark)





➢ SVM highlights features with decreased concentrations in spiked samples

Comparison with alternative feature selection methods



- *biosigner* finds mall signatures providing a good compromise between prediction accuracy, signature stability and computation time

Conclusions

- Biosigner:
  - efficient selection of significant signatures for binary classification
  - easy access (R and Galaxy)

- Depends on the structure of the dataset (distribution, limit of detection, correlation)

- Validation on an independant dataset is mandatory

- Importance of public datasets and code to benchmark new algorithms

Statistics

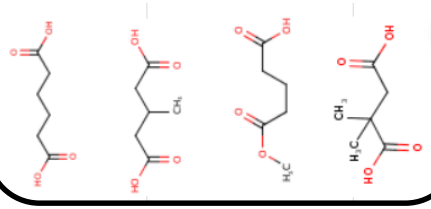
- Data integration

ProMetIS: integration of proteomics and metabolomics data

- New methods and bioinformatics tools
  - statistical integration (multivariate-based approaches)
  - network analysis (pathway-based approaches)
- 2 case studies:
  - high-throughput phenotyping of mouse models
  - systems microbiology
- Large consortium
  - CEA (LIST, IG, BIG), INRA (PFEM, TOXALIM), CNRS (LABGeM)
- 2 year project (« Integrative Bioinformatics » workpackages from IFB)
- Perspectives:
  - application to human phenotyping (France Medecine Genomique 2025)

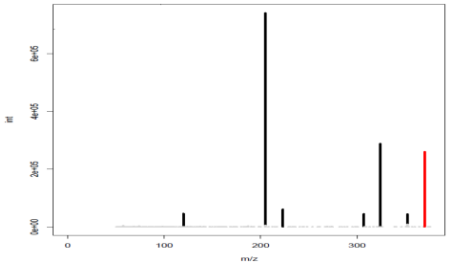
Spectrum identification

Molecular and/or spectral **database**



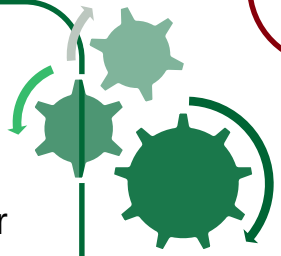
Input:

- A mass spectrum
- A precursor m/z

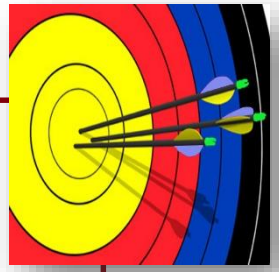


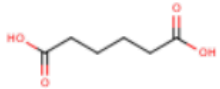
Identification sofware*

Model learned on a database.
CFM-ID: Transition probabilities
CSI-FingerID: Linear transformation between kernel spaces

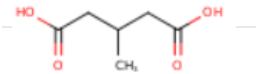


Output :
Scored molecules

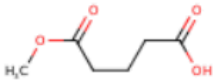




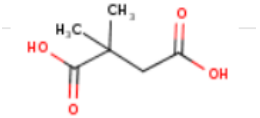
0,98



0,85



0,63



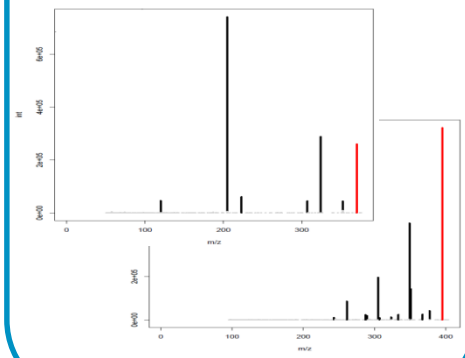
0,82

- CSI-FingerID (Shen, 2014)
- CFM-ID (Allen, 2014)
- MetFrag (Wolf, 2010)
- Etc…

Mining spectral libraries

Input:

- A set of spectra

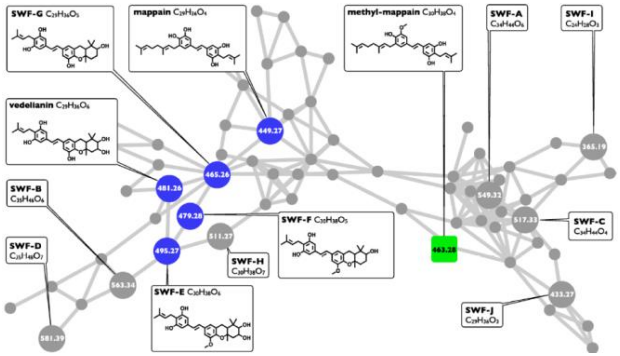


Spectral Mining software

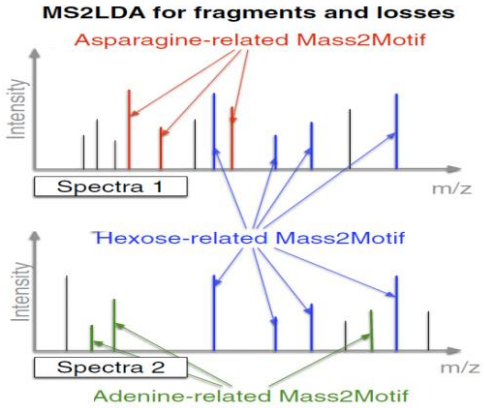
Output:

Information about structural similarities

- **Networks** (GNPS, Wang, 2014)

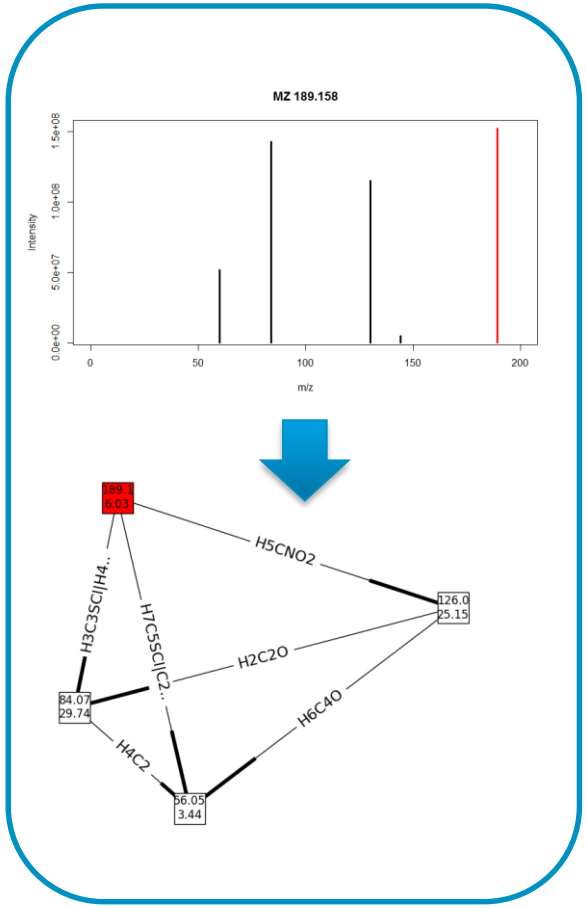


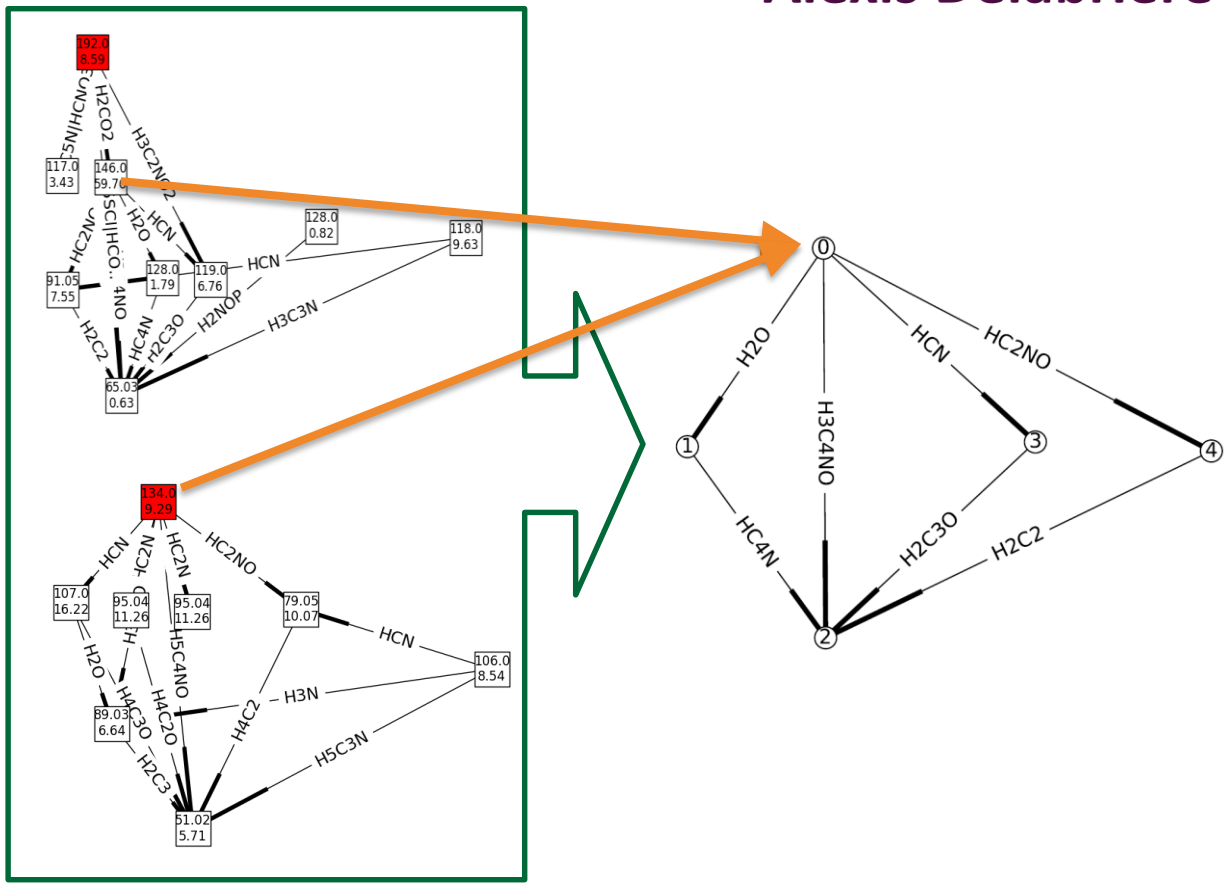
- **Motifs** discovery (MS2LDA, Van der Hooft, 2016)



MineMS² workflow

Alexis Delabrière





1) Conversion of spectra to graphs

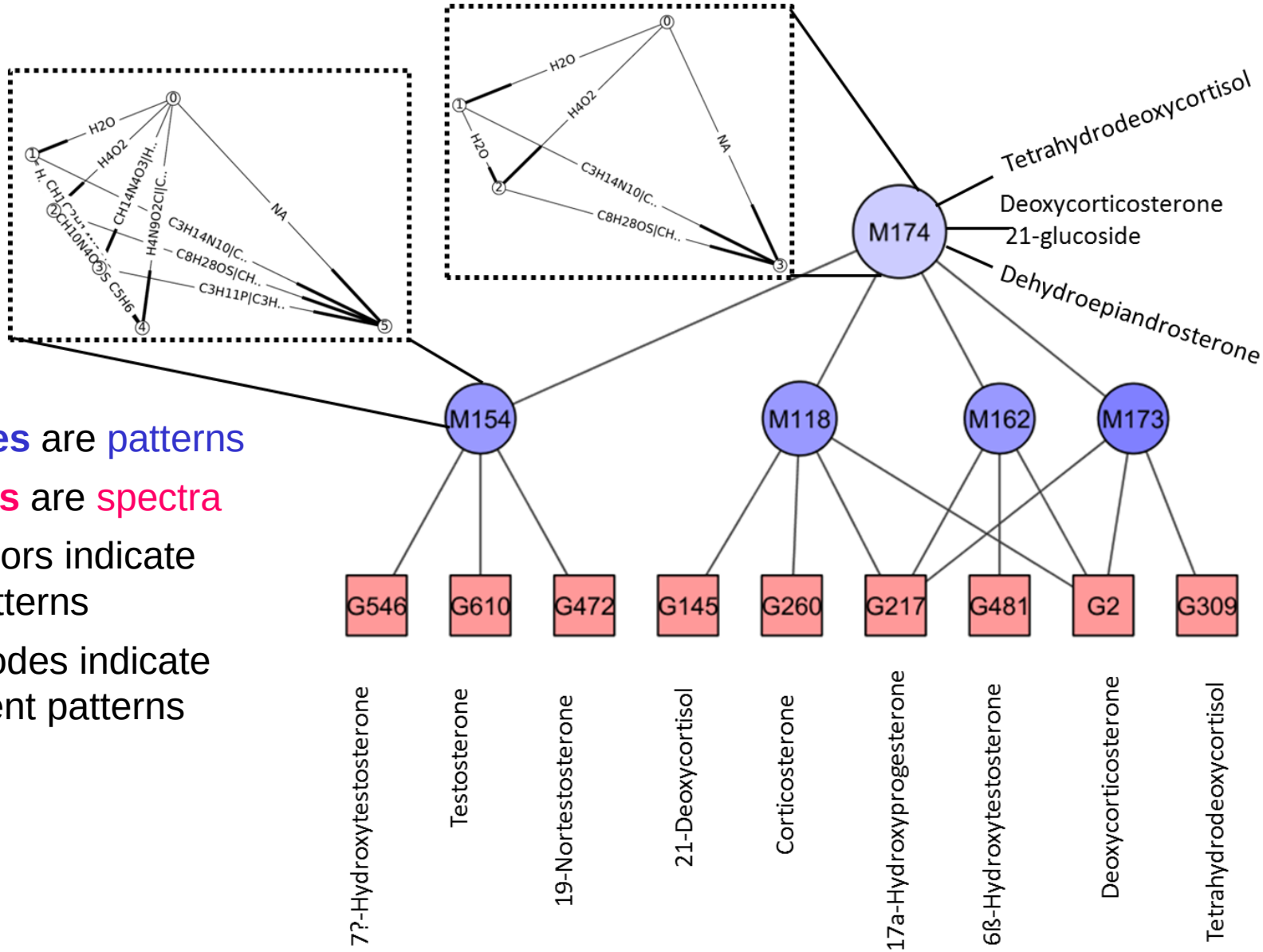
2) Frequent subgraph Mining

Running time and number of patterns

Dataset	*Penicillium verrucosum*	Reference library from pure compounds
Reference	Hautbergue *et al.* 2017 *J Chromatogr B*	Metabolome IDF
Number of spectra	91	834
Graph building	35 s	1 min 45 s
Pattern Mining	10 s	2 min 10 s
Number of patterns	54	832

- Processing time is more dependent on the similarities in the datasets than of the size in the dataset.

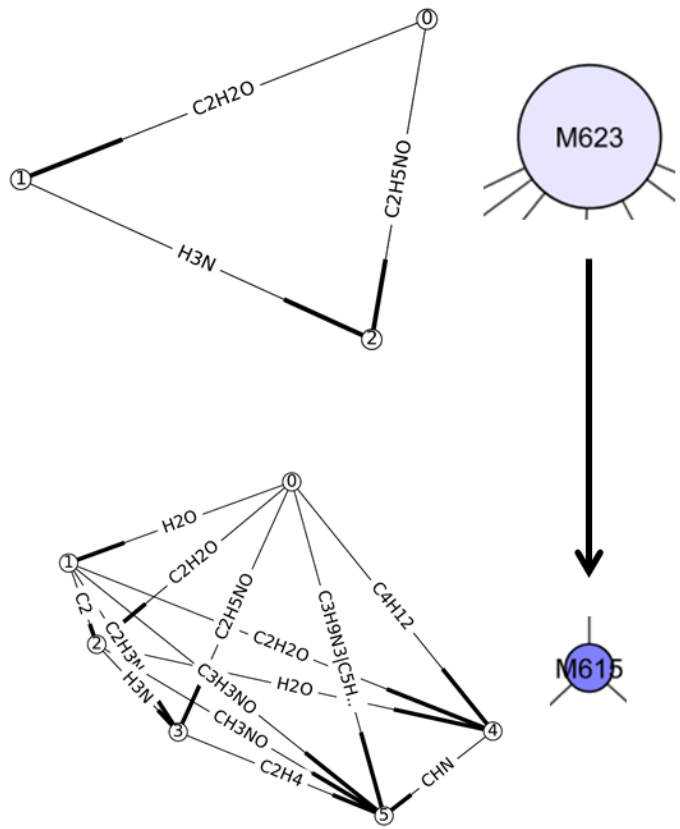
Example: sterone sub-lattice



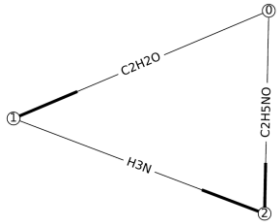
- **Blue nodes** are patterns
- **Red nodes** are spectra
- Lighter colors indicate smaller patterns
- Smaller nodes indicate less frequent patterns

Example: sterone sub-lattice

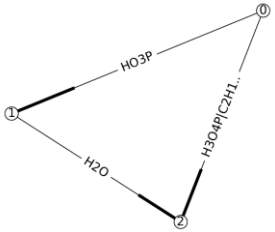
Patterns include coarse and fine grain information



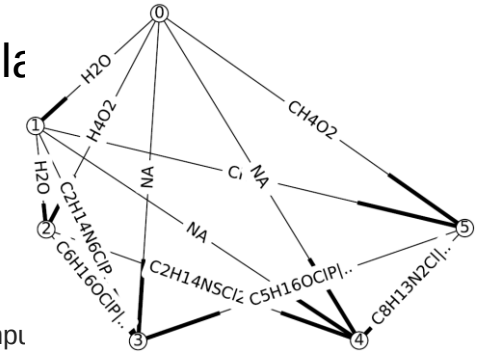
- Exemple of generic patterns:
  - N-Acetyl * compounds



  - Phosphated compounds

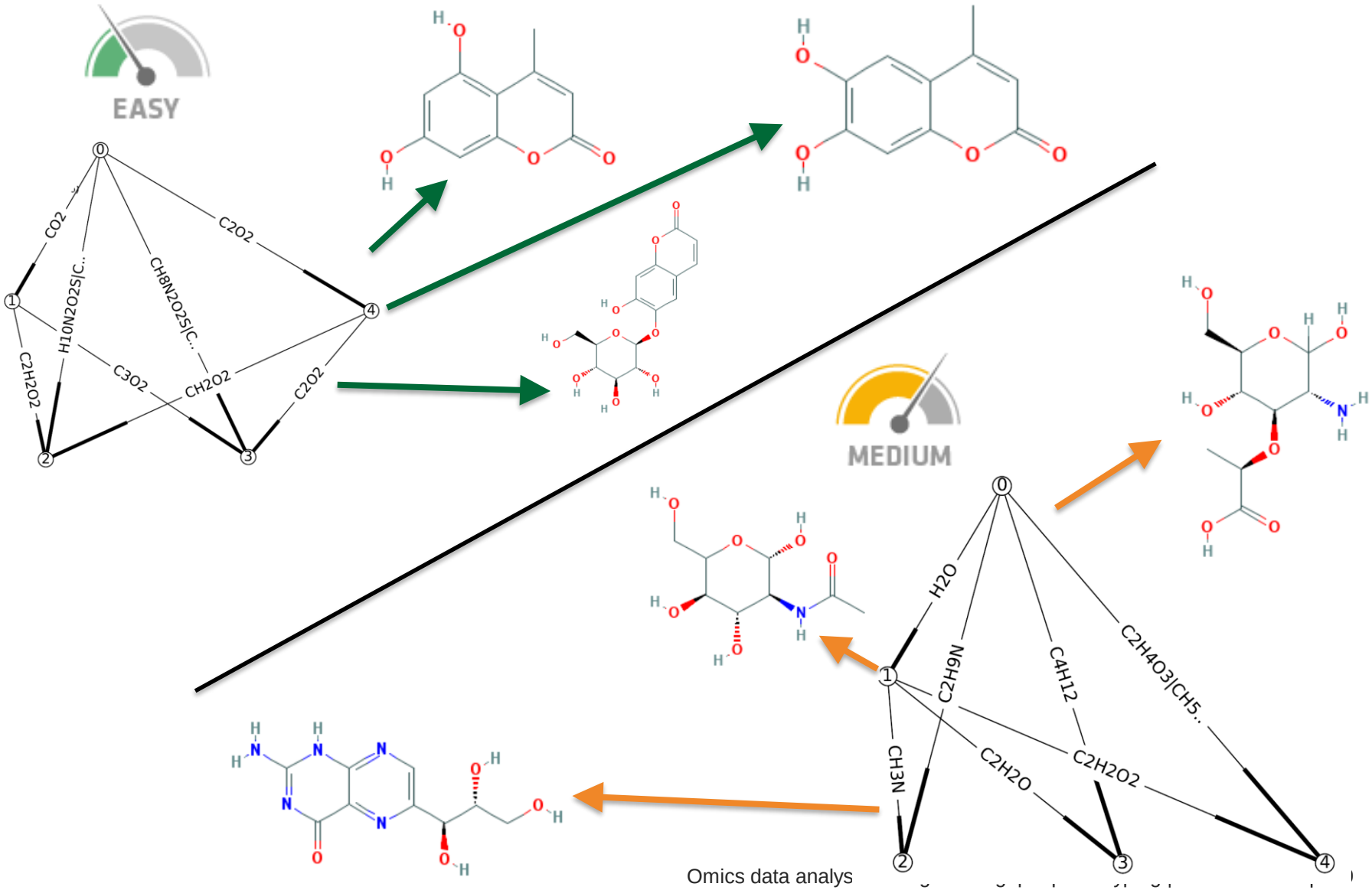


  - ….
- Specific patterns:
  - Sphingosine rela



Major difference with the patterns obtained with Mass2LDA (Van der Hooft *et al.* 2016, *PNAS*)

Relating patterns to chemical (sub-)structures

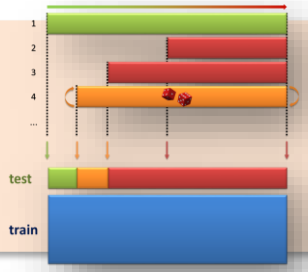


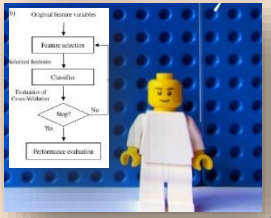
➢ Workflow management

- Workflows

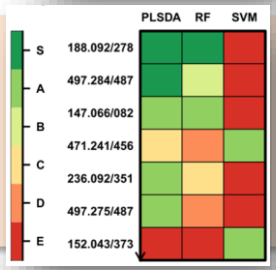
Implementation: From the method to the workflow

Method





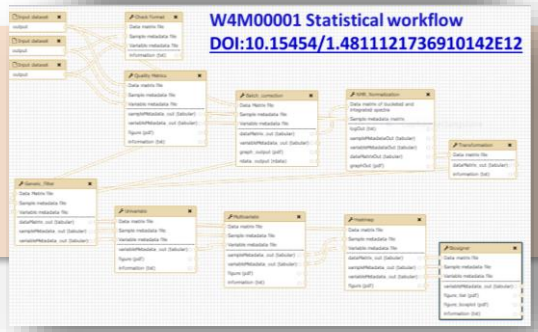
Package



GUI

Workflow

Cloud



http://workflow4metabolomics.org

The workflow challenge: bridging experimenters' and bioinformaticians' talents

Users

- Web-based
- Workflow editor
- User-friendly
- Tutorials

- Shared workflows
- Reproducible science

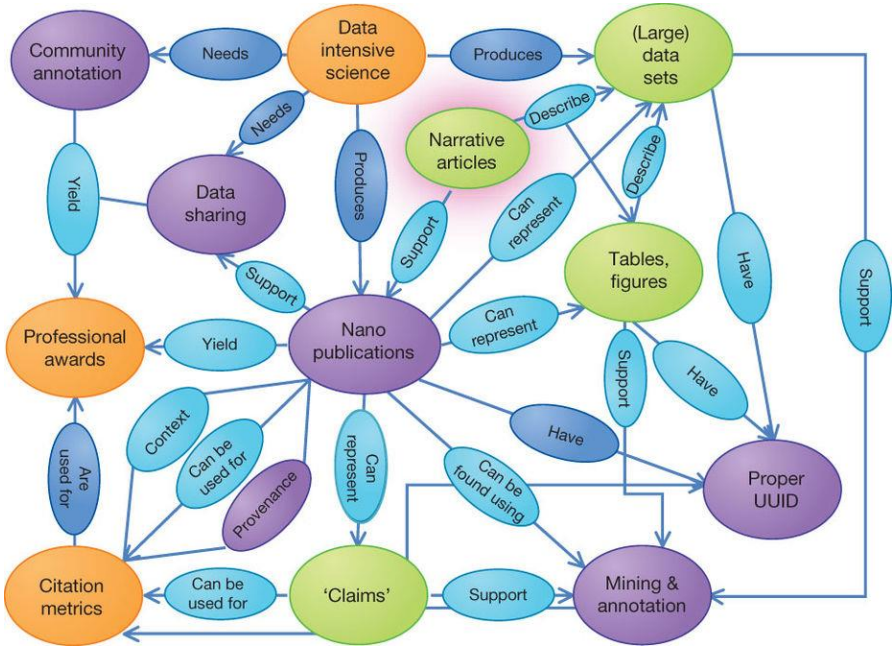
Developers

- Multi-language
- Toolshed
- Open-source

The value of data

- Data acquisition and annotation is costly and time-consuming
  - instruments, reactants, animal lives, your (precious) time
- In the era of data intensive sciences:
  - text (classical publication format) is not an effective way of sharing scientific information
  - computer-readable formatting is needed

Mons *et al.* (2011). The value of data. *Nat. Genet.* 43. 281-283.



- **Workflow management**
  - **Galaxy environment**

The Galaxy project

https://galaxyproject.org

- Workflow management through a classic web browser

Giardine *et al.* (2005). Galaxy: A platform for interactive large-scale genome analysis. *Genome Res.* 15:1451-1455.

- Started in 2005; more than 55,000 users worldwide

Goecks *et al.* (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11:R86.

Afgan *et al.* (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nature Biotechnol.* 29:972-974.

- NGS, transcriptomics, proteomics, metabolomics

Boekel *et al.* (2015). Multi-omic data analysis using Galaxy. *Nature Biotechnol.* 33:137-139.

Guitton *et al.* (2017). Create, run, share, publish, and reference your LC-MS, FIA-MS, GC-MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics. *Int. J. Biochem. Cell. Biol.* 93:89-101.

Tool interface

usegalaxy.org

Galaxy | Analyze Data | Workflow | Shared Data | Lab | Visualization | Admin | Help | User | Using 755.8 GB

Tools

search tools

- Get Data
- Send Data
- Lift-Over
- Text Manipulation
- Datamash
- Convert Formats
- Filter and Sort
- Join, Subtract and Group
- Fetch Alignments/Sequences
- NGS: QC and manipulation
- NGS: Mapping
  - Map with BWA-MEM – map medium and long reads (> 100 bp) against reference genome
  - Map with BWA – map short reads (< 100 bp) against reference genome
  - Bowtie2 – map reads against reference genome
  - Parse blast XML output
  - Megablast compare short reads against htgs, nt, and wgs databases
  - Map with BWA for Illumina
  - Map with Bowtie for Illumina
  - Lastz map short reads against reference sequence

Map with BWA-MEM – map medium and long reads (> 100 bp) against reference genome (Galaxy Version 0.7.12.1)

Versions | Options

Will you select a reference genome from your history or use a built-in index?

Use a built-in genome index

Built-ins were indexed using default options. See `Indexes` section of help below

Using reference genome

C. elegans (WS220): ce10

hg38

- Human (Homo sapiens) (b38): hg38
- Human (Homo sapiens) (b38): hg38 Canonical
- Human (Homo sapiens) (b38): hg38 Canonical Female

Select first set of reads

18: Combine FASTA and QUAL on data 7

Specify dataset with forward reads

Select second set of reads

18: Combine FASTA and QUAL on data 7

Specify dataset with reverse reads

Enter mean, standard deviation, max, and min for insert lengths.

-I; This parameter is only used for paired reads. Only mean is required while sd, max, and min will be inferred. Examples: both "250" and "250,25" will work while "250,,10" will not. See below for details.

Set read groups information?

Do not set

Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

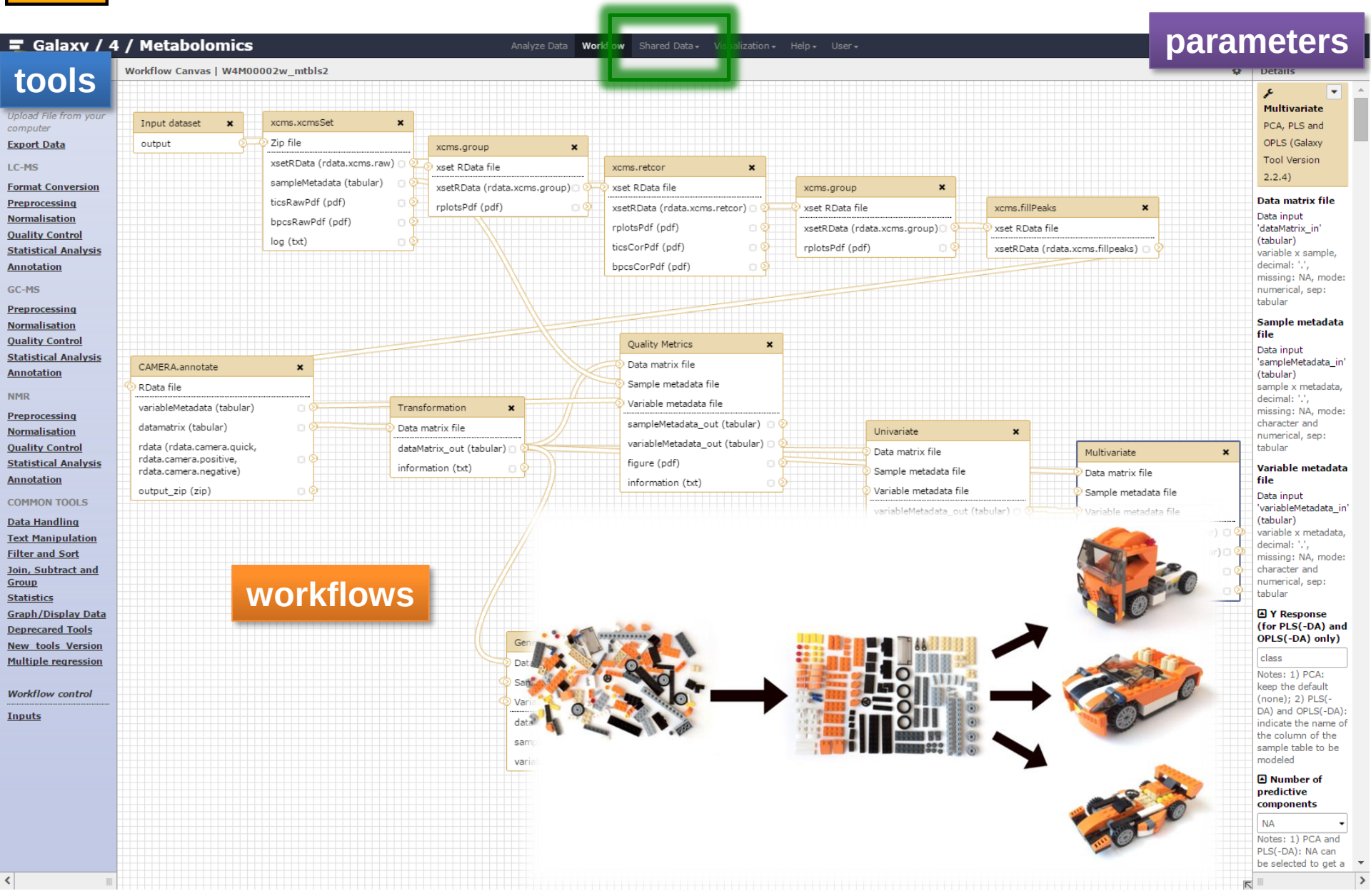
Select analysis mode

History

- 12: Map with BWA-MEM on data 11 and data 10 (mapped reads in BAM format)
- 11: Combine FASTA and QUAL on data 9
- 10: Combine FASTA and QUAL on data 8
- 9: Sequence Content Trimmer on data 6 and data 5
- 8: Sequence Content Trimmer on data 6 and data 5
- 7: Du Novo: Make consensus reads on data 4 (SSCS)
- 6: Du Novo: Make consensus reads on data 4 (mate 2)
- 5: Du Novo: Make consensus reads on data 4 (mate 1)
- 4: Du Novo: Align families on data 3
- 3: Du Novo: Make families on data 2 and data 1
- 2: SRR1799908_reverse
- 1: SRR1799908_forward

Afgan *et al.* (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nat. Biotechnol.* 29:972-974.

Workflow editor



Management of histories

usegalaxy.org

Galaxy | Analyze Data | Workflow | Shared Data | Lab | Visualization | Admin | Help | User | Using 755.8 GB

Done | search histories | search all datasets | Create new

Current History | Switch to | Switch to | Switch to

Duplex Analysis: ABL1
24 shown
9.93 GB
search datasets
Drag datasets here to copy them to the current history

- 24: Filter on data 23
- 23: Variant Annotator on data 22
- 22: Naive Variant Caller on data 21
- 21: MergeSamFiles on data 20 and data 19: Merged BAM dataset
- 20: Map with BWA on data 18 (mapped reads in BAM format)
- 19: Map with BWA-MEM on data 18 (mapped reads in BAM format)
- 18: Combine FASTA and QUAL on data 7
- 17: Filter on data 16
- 16: Variant Annotator on data 15
- 15: Naive Variant Caller on data 14
- 14: MergeSamFiles on data 13 and data 12: Merged BAM dataset
- 13: Map with BWA on data 11 and data 10 (mapped reads in BAM for

PNAS 2014
2 shown, 47 deleted, 2252 hidden
178.99 GB
Mark

- 1424: MarkDuplicates on collection 470: MarkDuplicate metrics — a list of datasets
- 1423: MarkDuplicates on collection 470: MarkDuplicates BAM output — a list of datasets

Unnamed history
2 shown
5.7 GB
search datasets

- 2: UCSC Main on Human: snp144 (genome)
- 1: UCSC Main on Human: knownGene (genome)

Unnam
17 show
157.19
searc

- 18: ipyt b
- 17: ipyt b
- 16: PNA
- 14: ipyt b
- 13: ipyt b
- 12: ipyt b
- 11: ipyt b
- 10: ipyt b
- 9: ipyth
- 8: ipyth
- 7: ipyth
- 6: Jupyt

Afgan *et al.* (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nat. Biotechnol.* 29:972-974.

Sharing histories

olomics Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 7%

Share or Publish History 'W4M00001_Sacurine-statistics'

Make History Accessible via Link and Publish It

This history is currently **accessible via link and published**.
Anyone can view and import this history by visiting the following URL:

http://galaxy.workflow4metabolomics.org/u/ethevenot/h/w4m00001sacurine-statistics-1

This history is publicly listed and searchable in Galaxy's Published Histories section.
You can:

Unpublish History

Removes this history from Galaxy's Published Histories section so that it is not publicly listed or searchable.

Disable Access to History via Link and Unpublish

Disables this history's link so that it is not accessible and removes history from Galaxy's Published Histories section so that it is not publicly listed or searchable.

Share History with Individual Users

You have not shared this history with any users.

Share with a user

Back to Histories List

History

search datasets

W4M00001_Sacurine-statistics
53 shown
3.99 MB

53: Heatmap_figure.pdf

52: Heatmap_Multivariate_Multivariate_Multivariate_Univariate_Univariate_Univariate_Generic_Filter_Quality_Metrics_Generic_Filter_Quality_Metrics_Batch_correction_all_loess_pool_variableMetadata.tsv

51: Heatmap_Multivariate_Multivariate_Multivariate_Generic_Filter_Quality_Metrics_Generic_Filter_Quality_Metrics_sampleMetadata.tsv

50: Heatmap_Generic_Filter_Transfor

- **Workflow management**
  - **Workflow4Metabolomics online platform**

Workflow4metabolomics

Main menu

- Home
- Events
- History
- Introduction
  - The Galaxy environment
  - The LC-MS workflow
  - The GC-MS workflow
  - The NMR workflow
  - References
- HowTo
- Download
  - Datasets
- Referenced WorkFlows and Histories
- How to contribute?
- Developer resources
  - Source-code
  - Virtual environments

Workflow4Metabolomics 3.0

Welcome to the collaborative portal dedicated to metabolomics data processing, analysis and annotation for Metabolomics community.

" We are happy to announce the next **Workflow4Experimenters (W4E) international course 2018**: *Using Galaxy and the Workflow4metabolomics infrastructure to analyse metabolomics data.* Please save the date: ***8-12 October 2018*** at Pasteur Institute, Paris - France

More news in April ! "

Follow us on Twitter @workflow4metabo

STEP 1 | STEP 2 | STEP 3 | STEP 4

Giacomoni *et al.* (2015). Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. *Bioinformatics*. 31:1493-1495.

French Institute of Bioinformatics

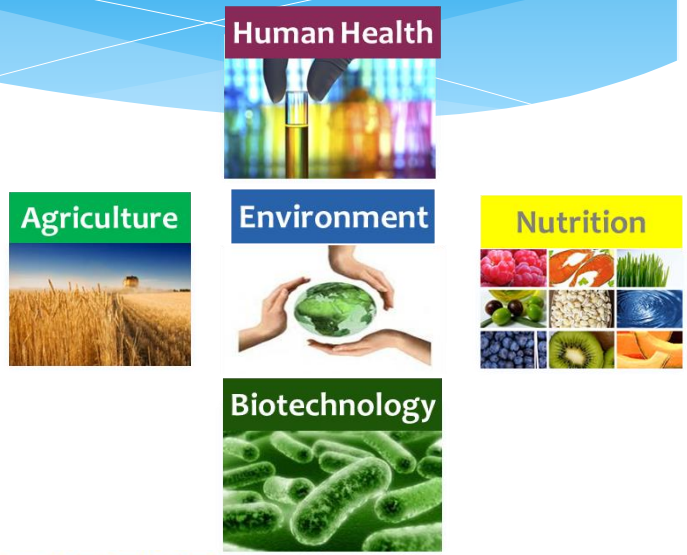
- **Coordinators (C. Médigue & J. van Helden)**
  - **French node from Elixir**
  - **Federation** of 34 national bioinformatics platforms
  - 230 bioinformaticians

- **Missions**
  - **E-infrastructure** components : Storage, Computing (e.g. **cloud**) , Tools, VRE…
  - **Training** (NGS, Galaxy) & community animation
  - Collaboration with national and European infrastructures

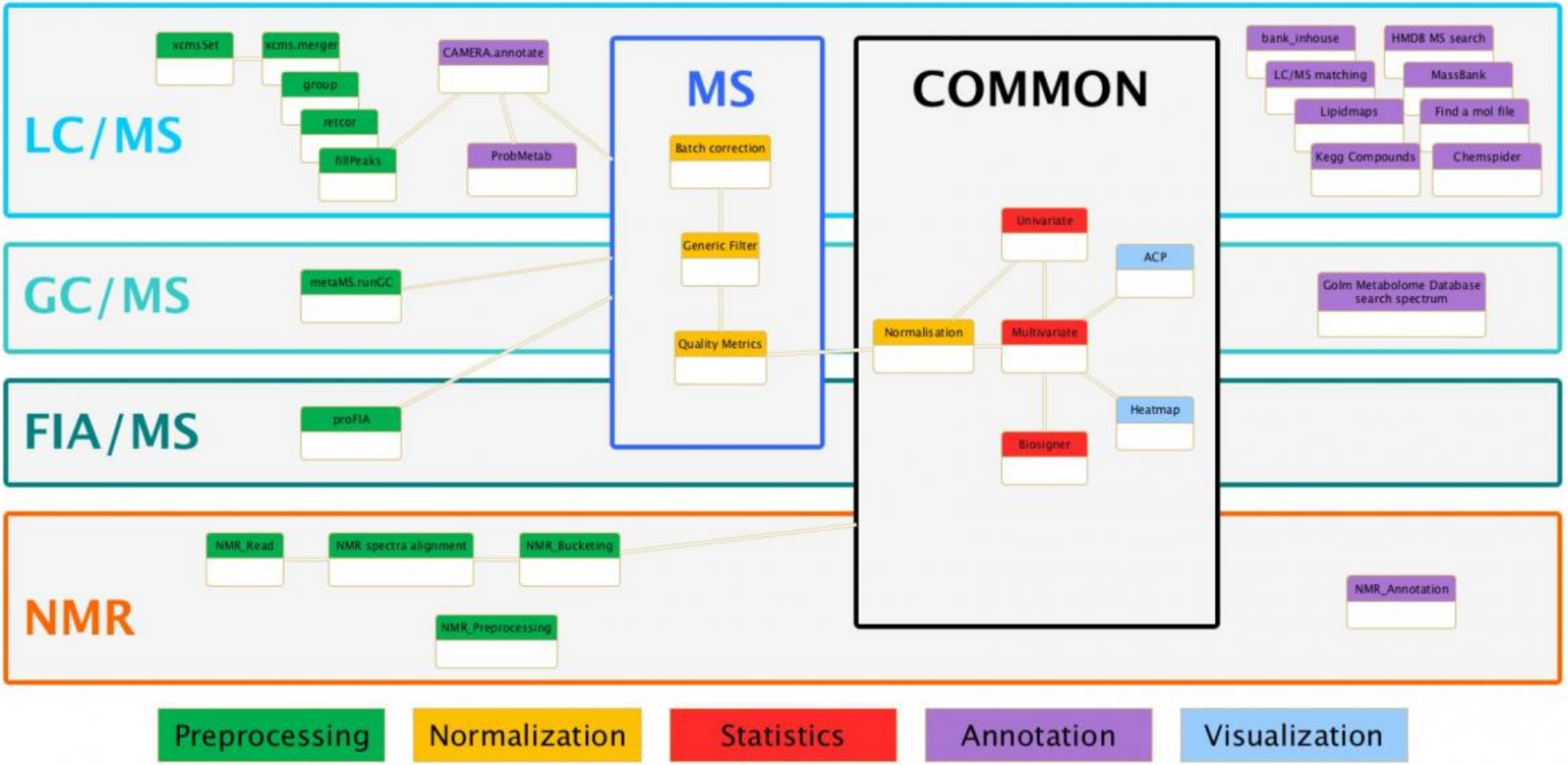
- **National task force**
  - **IFB Galaxy Working Group**
  - **European Galaxy Developer workshop 2017**
  - **Organization of the GCC conference 2017**

MetaboHUB: The French infrastructure for metabolomics and fluxomics

* Coordinator (D. Rolin)
  * 80 permanent scientists
  * Total budget: 45 M€
  * Launched in 2013
* 4 LC-MS, GC-MS and NMR platforms
  * dedicated to Innovation, Service, Technology Transfer and Training
* Built upon the Francophone Network for Metabolomics and Fluxomics (> 300 members)
* 4 online bioinformatics infrastructures
  * workflows, databases, pathways
* Partnerships:
  * bioinformatics, proteomics, cohorts, crops
  * Europe:
* MetaboMOOC (oct. 2018; with IFB)

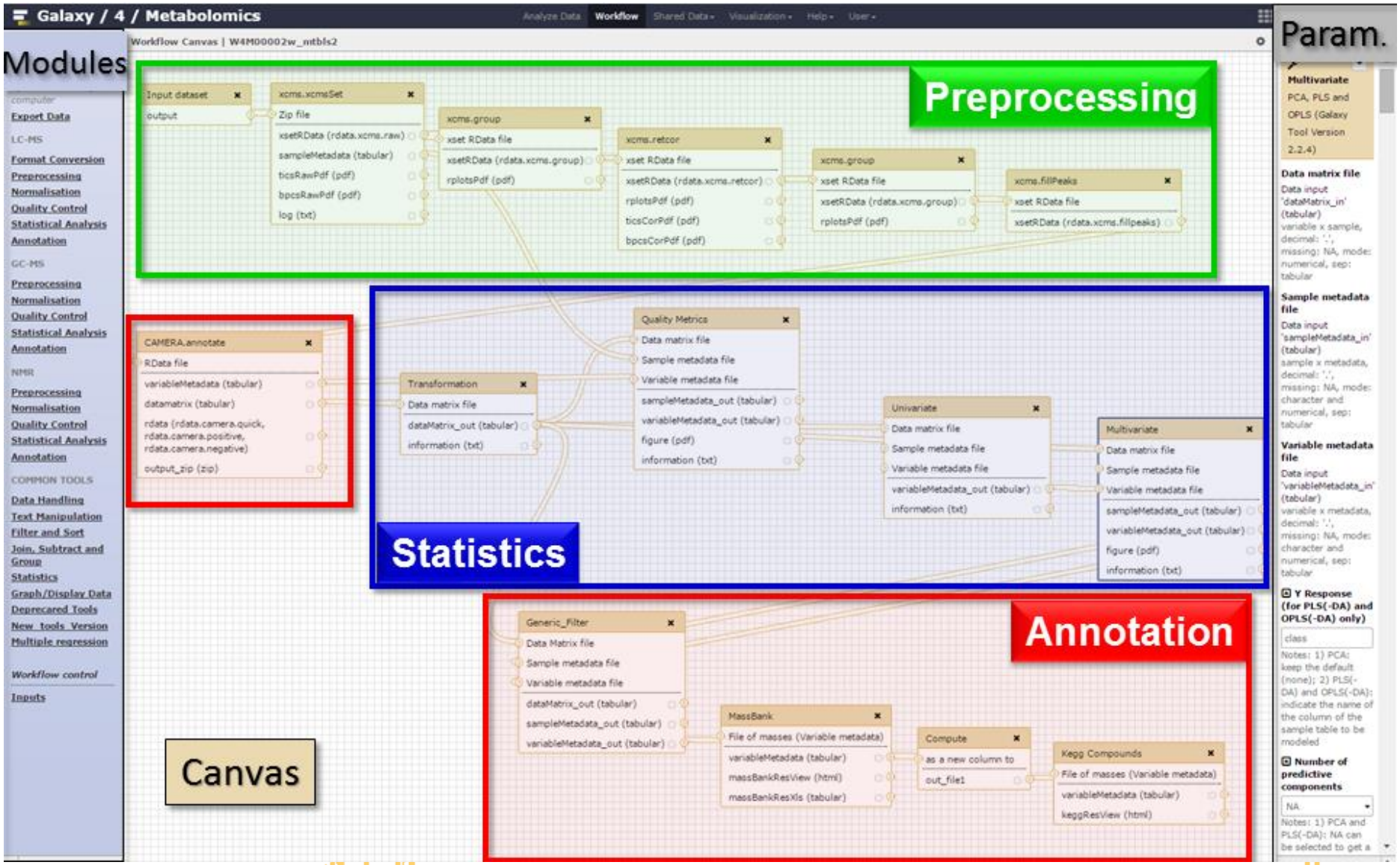


W4M tools



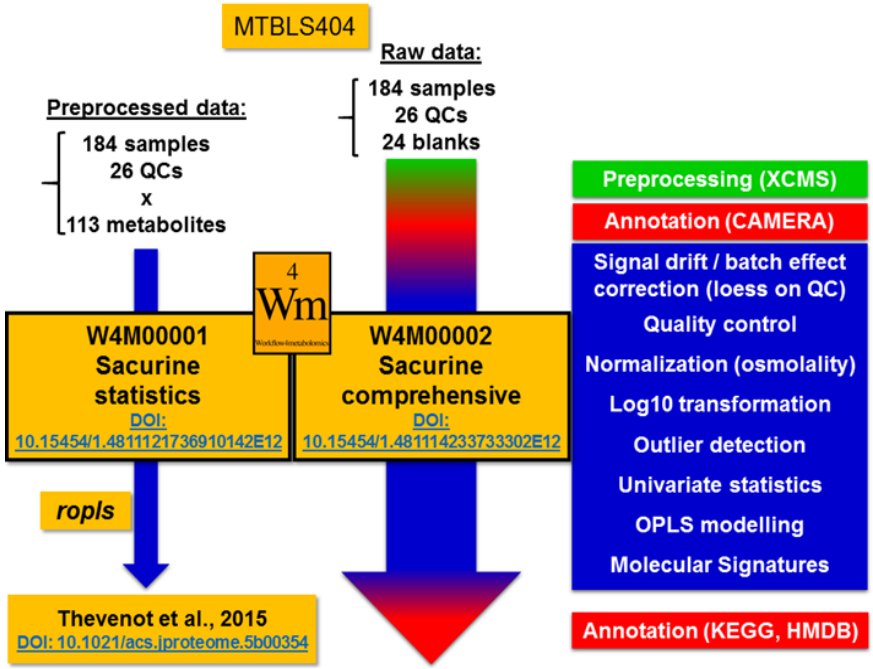
Guitton *et al.* (2017). Create, run, share, publish, and reference your LC-MS, FIA-MS, GC-MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics. *Int. J. Biochem. Cell. Biol.* 93:89-101.

W4M00002_sacurine-comprehensive



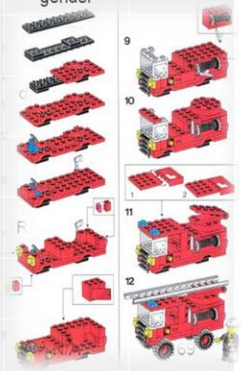
Referencing our analyses

- Demonstrate the value and the reproducibility of your analysis (e.g., to reviewers)
- Receive feedback on your results, get cited, and initiate new collaborations



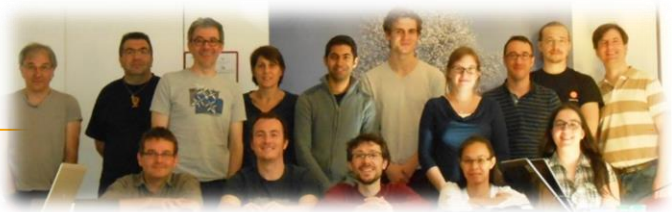
Referenced W4M histories

WOI	Name & DOI	Technology	Species	Matrice	Factor	Samples
W4M00001	*"Sacurine-statistics"* *10.15454/1.4811121736910142E12*	LC-MS	*H. sapiens*	Urine	age, BMI, gender	184
W4M00002	*"Sacurine-comprehensive"* *10.15454/1.481114233733302E12*	LC-MS	*H. sapiens*	Urine		
W4M00003	*"Diaplasma"* *10.15454/1.4811165052113186E12*	LC-MS	*H. sapiens*	Plasma		
W4M00004	*"GCMS Algae"* *10.15454/1.4811272313071519E12*	GC-MS	*E. siliculosus*	Algae		
W4M00005	*"Ractopamine"* *10.15454/1.4811287270056958E12*	LC-MS	*S. scrofa*	Serum		
W4M00006	*"BPA-MMusculus"* *10.15454/1.4821558812795176E12*	NMR	*M. musculus*	Brain		
W4M00007	*"Coffea leaves"* *10.15454/1.4985472277740251E12*	LCMS	Coffea sp.	Leaves		



Guitton *et al.* (2017). Create, run, share, publish, and reference your LC-MS, FIA-MS, GC-MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics. *Int. J. Biochem. Cell. Biol.* 93:89-101.

W4M offer



The W4M Core Team

- Private account
- Computation and storage resources
- Help desk

- Sharing and referencing of histories and workflows (DOI)

- Annual courses (tutoring on your own data)

Save the date: 8-12 October 2018, Pasteur Institute (Paris)

- Installation of local instances

contact@workflow4metabolomics.org

➢ The Team

A team work from talented people

- The CEA team
  - Natacha Lenuzza, Alexis Delabrière, Pierrick Roger-Mele, Bertrand Monfort, Philippe Rinaudo, *et al.*
- The MetaboHUB infrastructure
  - Fabien Jourdan, Franck Giacomoni, Marie Tremblay-Franco, Jean-François Martin, Mélanie Pétéra, Nils Paulhe, Christophe Junot, Estelle Pujos-Guillot, Dominique Rolin, *et al.*
- The IFB infrastucture
  - Christophe Caron, Gildas Le Corguillé, David Vallenet, Claudine Médigue, Jacques van Helden, *et al.*
- The PhenoMeNal consortium
  - Christoph Steinbeck, Steffen Neumann, Namrata Kale, Pablo Moreno, Kenneth Haug, Reza Salek, Philippe Rocca-Serra, Luca Pirredu, *et al.*

Thanks for coming

Questions?