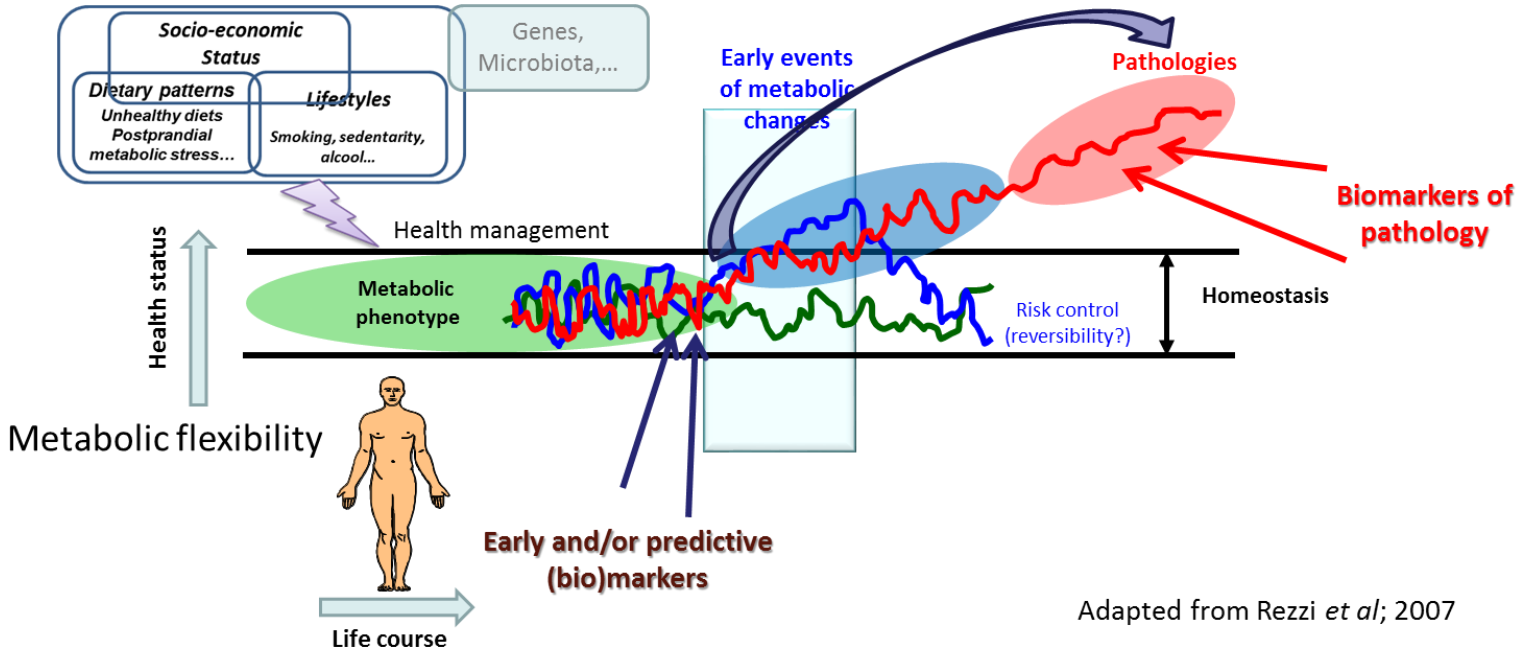
Feature selection methods for early predictive biomarker discovery using untargeted metabolomic data

Dhouha Grissa, Mélanie Pétéra, Marion Brandolini, Amedeo Napoli, Blandine Comte and Estelle Pujos-Guillot

4th Workshop
on Holistic Analytical Methods
for Systems Biology Studies
17-19 April 2016

INRA
SCIENCE & IMPACT

DYNAMICS OF METABOLIC PHENOTYPE AND EARLY CHANGES

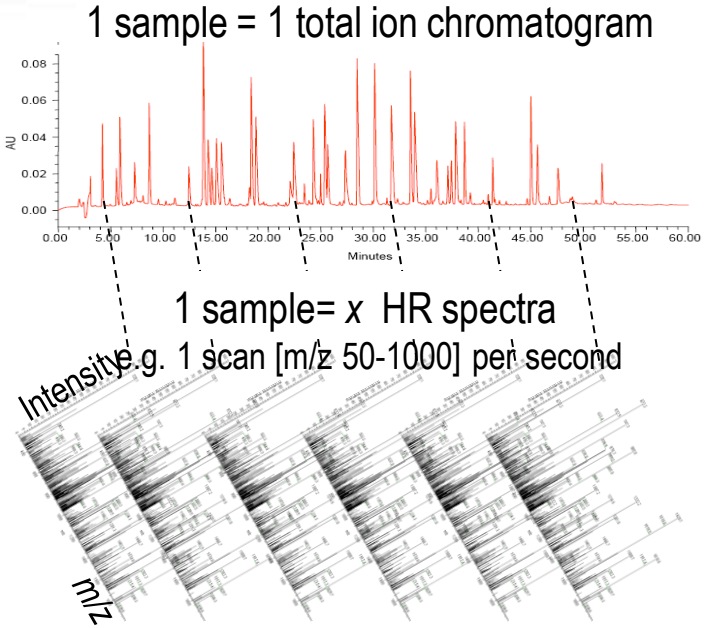


Adapted from Rezzi *et al*; 2007

➢ METABOLOMICS: A POWERFUL PHENOTYPING TOOL

comprehensive and integrative vision of biological systems

UNTARGETED MS-BASED METABOLOMICS



➤ **Need to optimize** two parameters:
(1) the biomarker performance
(2) the number of metabolites used in the predictive model.

METABOLIC PROFILES: MULTIPLE BIOMARKERS

TOWARDS THE DISCOVERY OF PREDICTIVE BIOMARKERS

1) Can the clinician measure the biomarker?

a) **Accurate and reproducible analytical method(s)**
b) **Pre-analytical issues (including stability) evaluated and manageable**
c) **Assay is accessible**
d) Available assays provide high through-put and rapid turn-around
e) Reasonable cost

2) Does the biomarker add new information?

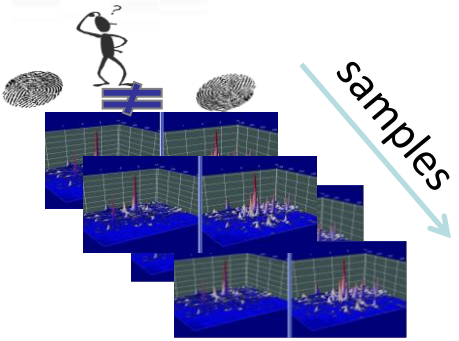
a) **Strong and consistent association between the biomarker and the outcome or disease of interest in multiple studies**
b) **Information adds to or improves upon existing tests**
c) Decision-limits are validated in more than one study
d) Evaluation includes data from community-based populations

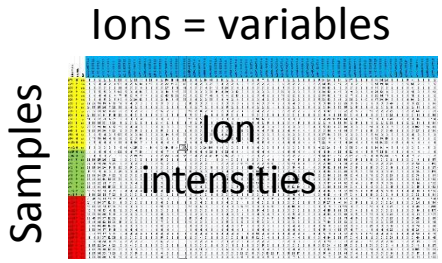
3) Will the biomarker help the clinician to manage patients ?

a) **Superior performance to existing diagnostic tests, or**
b) **Evidence that associated risk is modifiable with specific therapy, *or***
c) **Evidence that biomarker-guided triage or monitoring enhances care**
d) Consider each of multiple potential uses (SEE PANEL B)

Morrow *et al.*, 2014

UNTARGETED METABOLOMICS AND PREDICTION

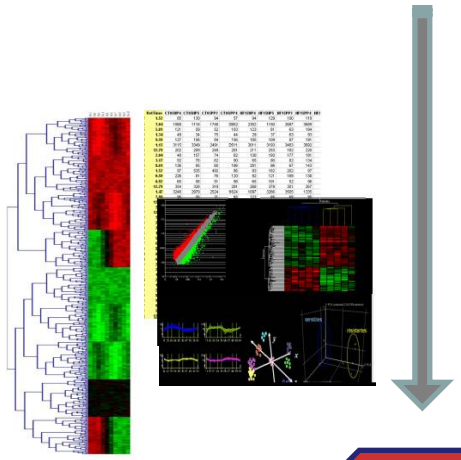




- Data from instrument signal: noisy, variable
- Range of linearity, missing data
- High redundancy / degree of correlation:

- one metabolite gives several ions
- several metabolites are in the same pathway

- High number of variables compared to the number of samples



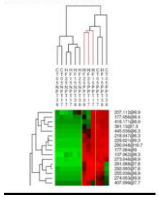
- Need ways to extract information from the data
- Obtain reliable, predictive information
- Ignore random variation (noise)

Knowledge

DISCOVERY OF THE BEST PREDICTIVE FEATURES
EVIDENCE FOR BIOLOGICAL MECHANISMS

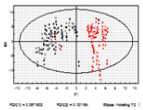
ALTERNATIVE TOOLS AND METHODS

- **Statistical** methods:
  - Univariate analyses ANOVA
  - Clustering methods,

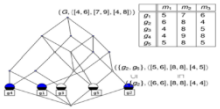


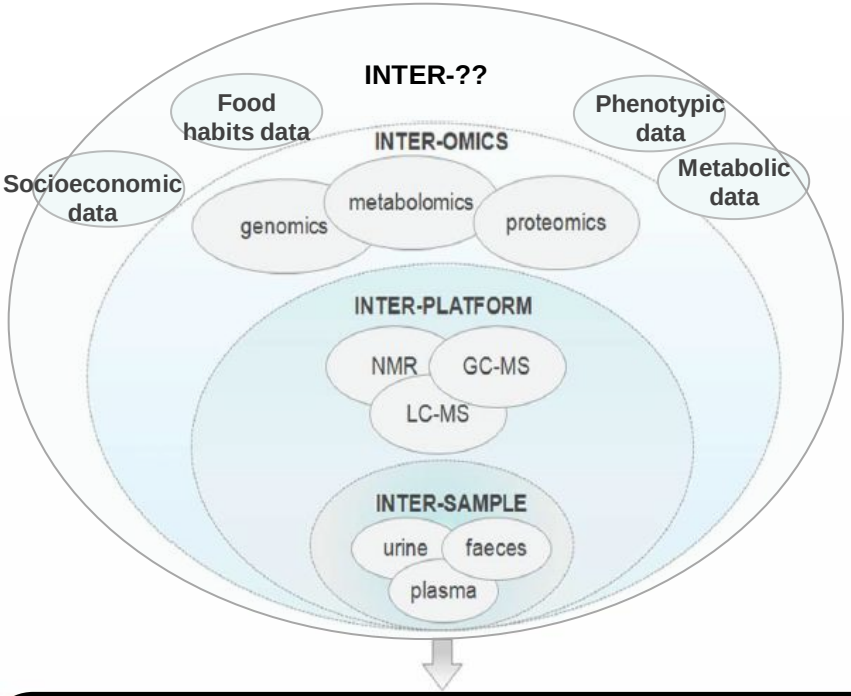
e.g., k-means, HAC;

  - Principal components analysis (PCA), PLS regression, PLS-DA



- **Data mining** methods:
  - Supervised classification: Random Forest, Support Vector Machine (SVM)…
  - Visualization with unsupervised methods: Formal concept analysis (FCA);
  - Association rules;



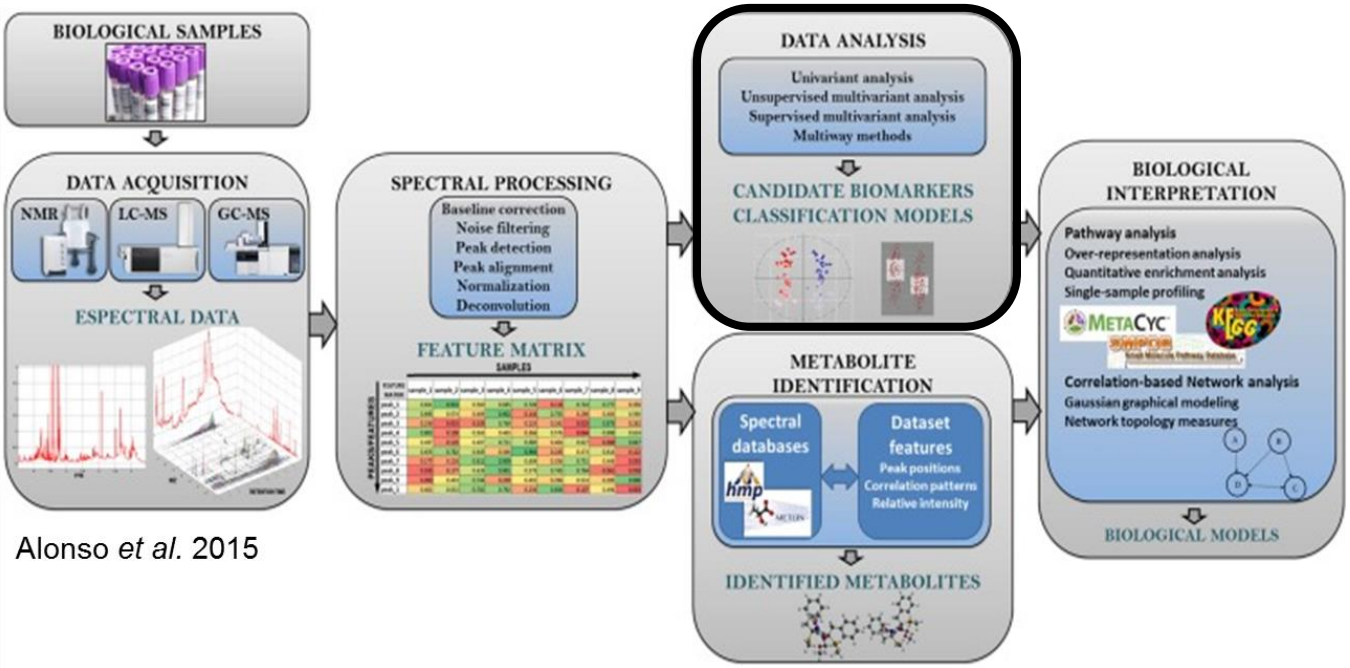


Management of extracted datasets

I- unsupervised
II- supervised
III- explanatory/inductive methods

Adapted from Vernocchi *et al.*, 2012

OBJECTIVE

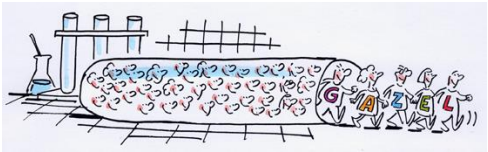


Alonso *et al.* 2015

Studying a workflow describing the **general feature selection process**, using knowledge discovery and data mining methodologies to propose advanced solutions **for predictive biomarker discovery from untargeted metabolomic data**

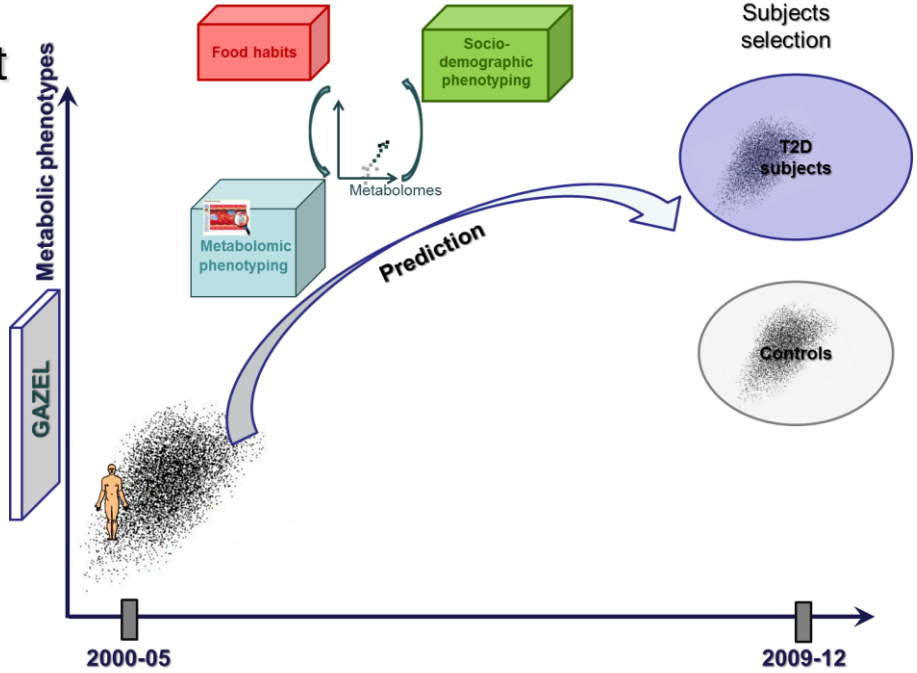
DATA COLLECTION

Case / Control study within the GAZEL cohort



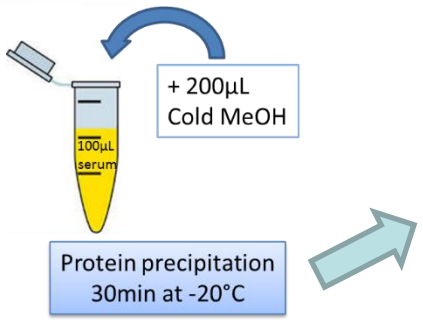
n=112 men
52-64 y.o, overweight 25≤BMI<30

Cases: T2D in 2009, free of T2D in 2004
Controles : matched for age and BMI classes

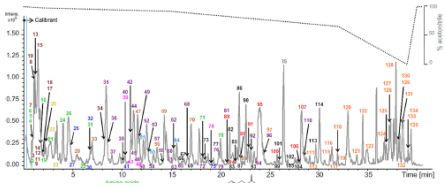


Untargeted metabolomics

DATA COLLECTION



UPLC-(ESI)QToF

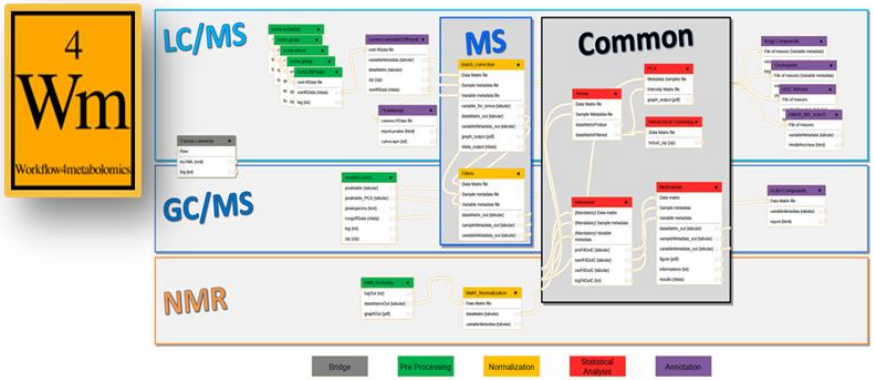


UPLC QToF Bruker Impact II

HSS T3 150 x 2.1mm 1.8µm
A : water + 0.1% FA
B : ACN + 0.1% FA
0.4mL/min

Pereira H. et al, Metabolomics 2010

Workflow4Metabolomics.org Galaxy



ifb INSTITUT FRANÇAIS DE BIOINFORMATIQUE

METABOHUB

Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics

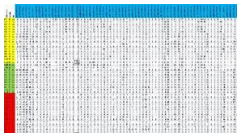
Franck Giacomoni[1,†], Gildas Le Corguillé[2,†], Misharl Monsoor[2], Marion Landi[1], Pierre Pericard[2], Mélanie Pétéra[1], Christophe Duperier[1], Marie Tremblay-Franco[3], Jean-François Martin[3], Daniel Jacob[4], Sophie Goulitquer[2], Etienne A. Thévenot[5,*] and Christophe Caron[2,*]

- **Data extraction:** XCMS Centwave. Prefilter (3,500), S/N=3
- **Data cleaning:** batch correction, noise removal, normalization, transformation

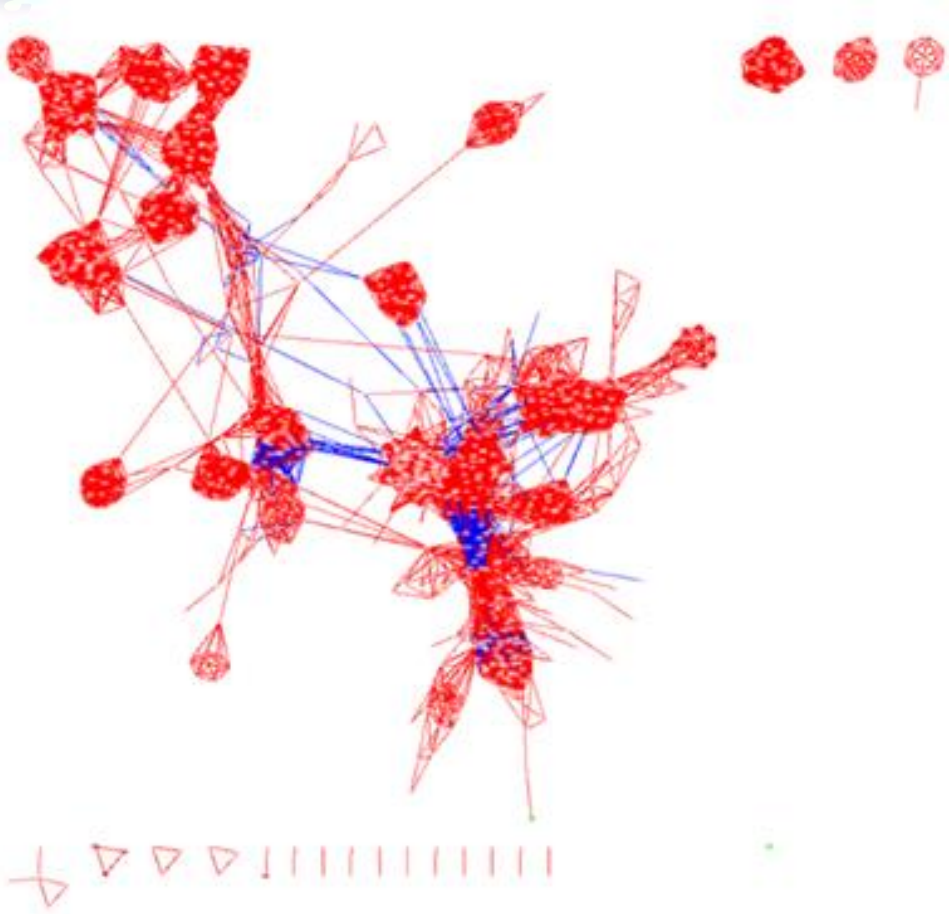
Signals > 2 blanks,
CVpool < 1.25CVsamples, CV<30%,
deisotope data

INRA SCIENCE & IMPACT

DATA CHARACTERISTICS



1,195 m/z variables
111 individuals

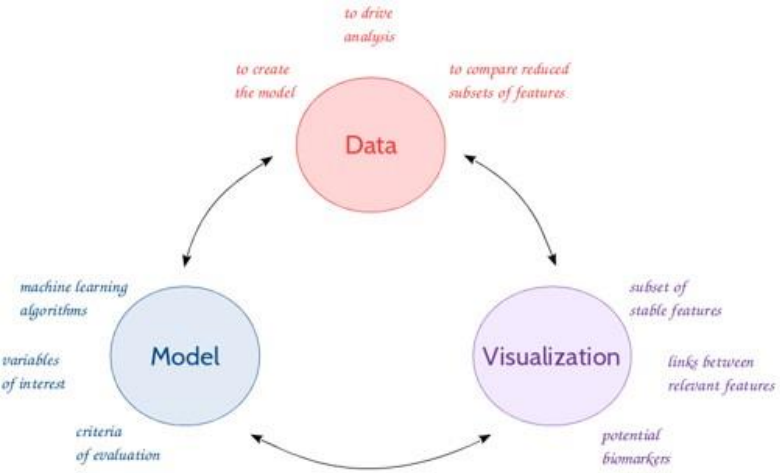


- ANOVA: 52 significant ions (4.3%) (p-value <0.5 after BH correction)

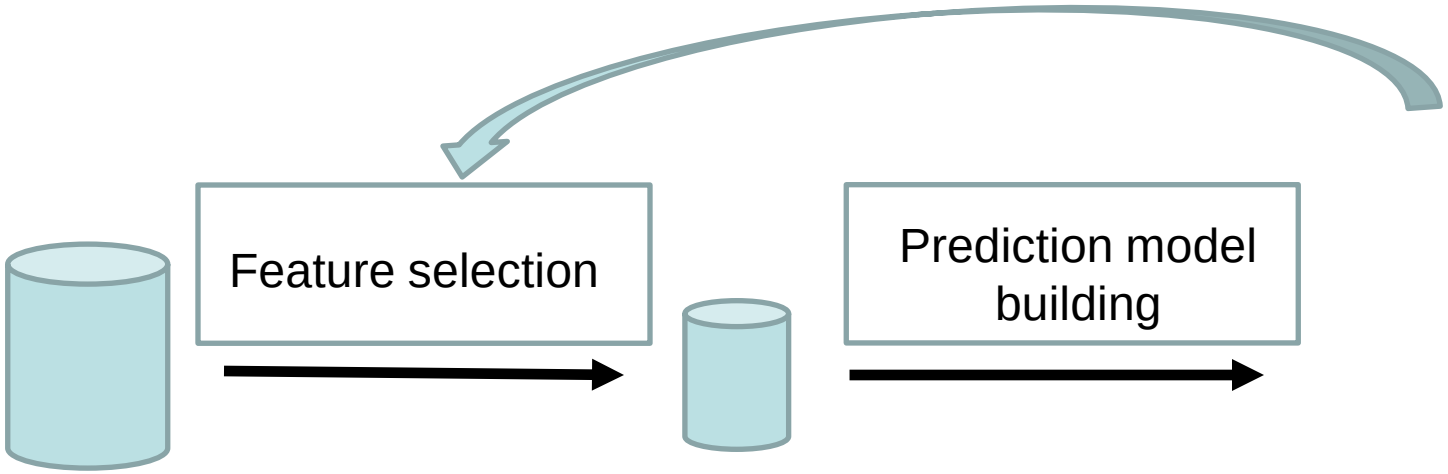
- 2.4% ions with correlation coefficient > 0.5, with 576 ions with a least one correlation >0.8.

- Correlation networks of the ions with correlations higher than 0.5 showed highly correlated clusters due to both analytical and biological origins

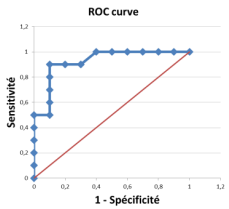
WORKFLOW



Biomarker discovery process:
(1) data pre-processing, (2) biomarker selection, (3) performance evaluation, and (4) final model creation



Performance evaluation



Original dataset

Reduced dataset

Potential biomarkers

FEATURE SELECTION

WHY ?

- To reduce the computational cost
- To improve the identification of specific markers

HOW ?

- Non-informative metabolites filtering:

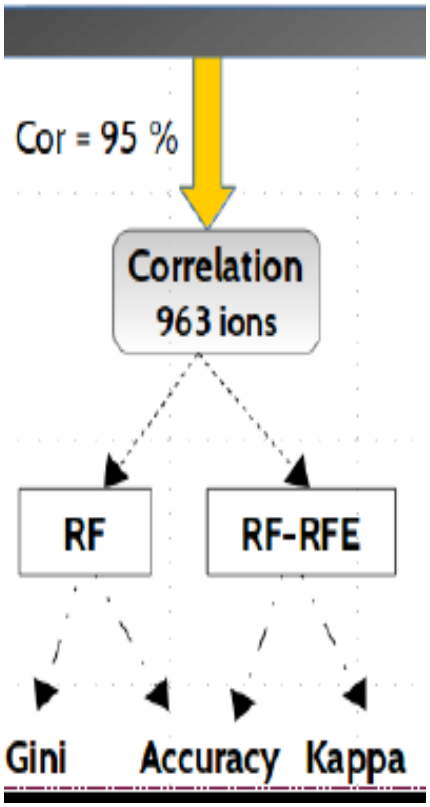
(1) those with very small intensities close to the limit of detection;
(2) those only detected in very few individuals;
(3) those that are near-constant irrespective of the difference in clinical outcome

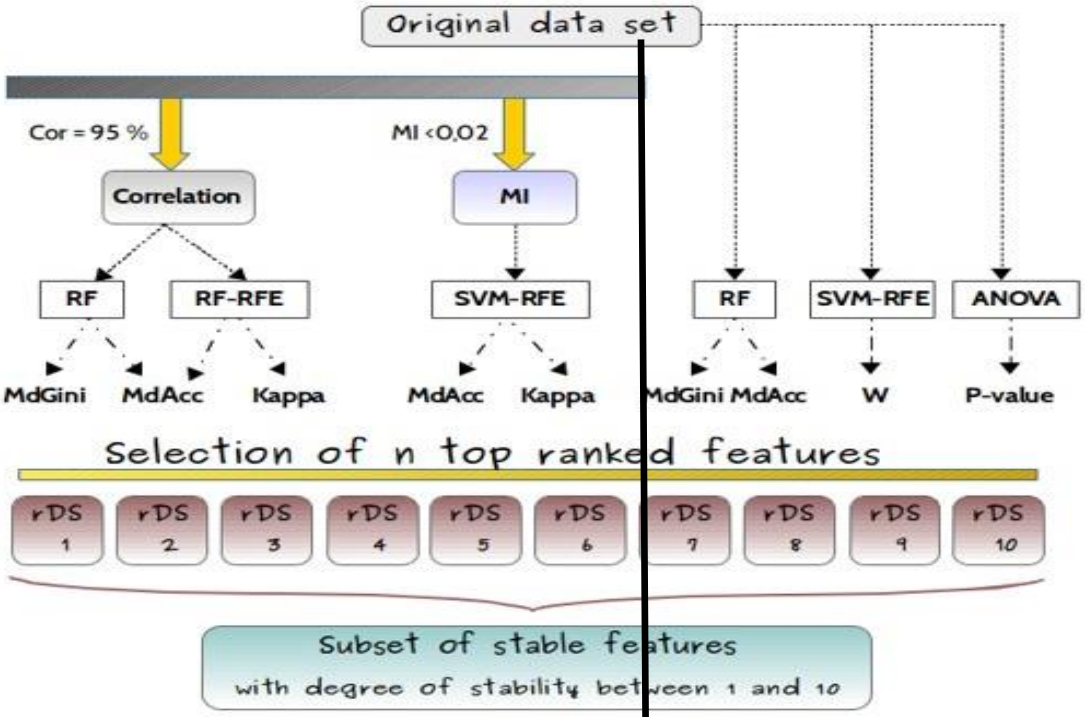
ALTERNATIVE ALGORITHMS:

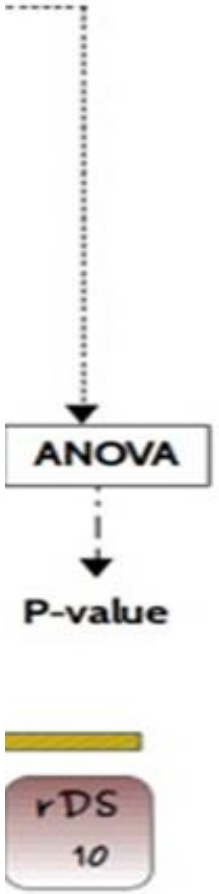
As a pre processing step : use of a statistical filter (t-test)

As a learning step : link the feature ranking to the classification task (wraper methods…)

EXPERIMENTAL DESIGN



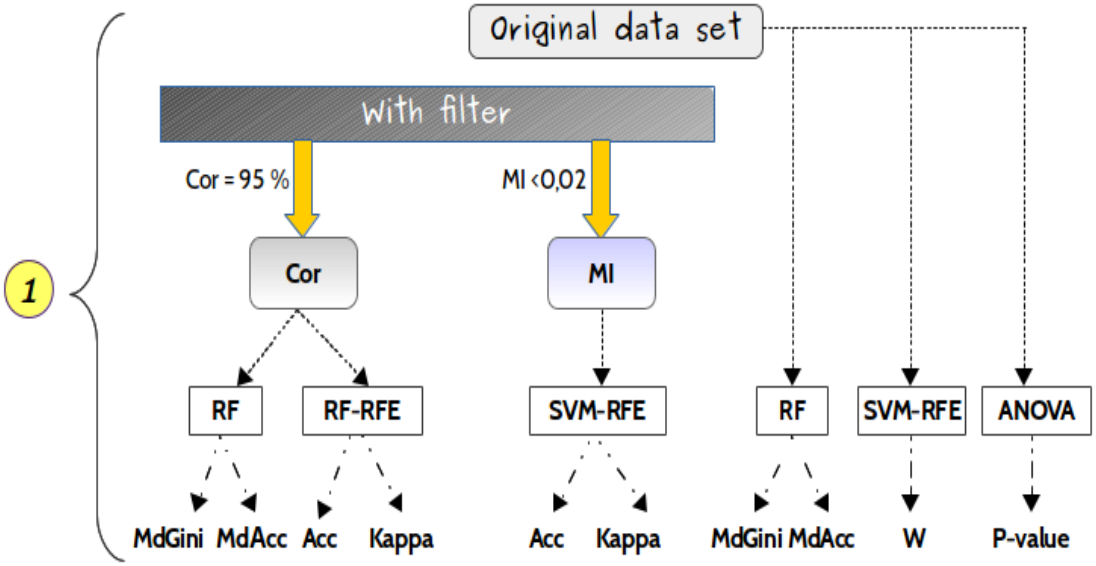




with filters

without filter

RESULTS



- Top 200 ranked features selected
- 107 ions (9%) with p-value <0.1

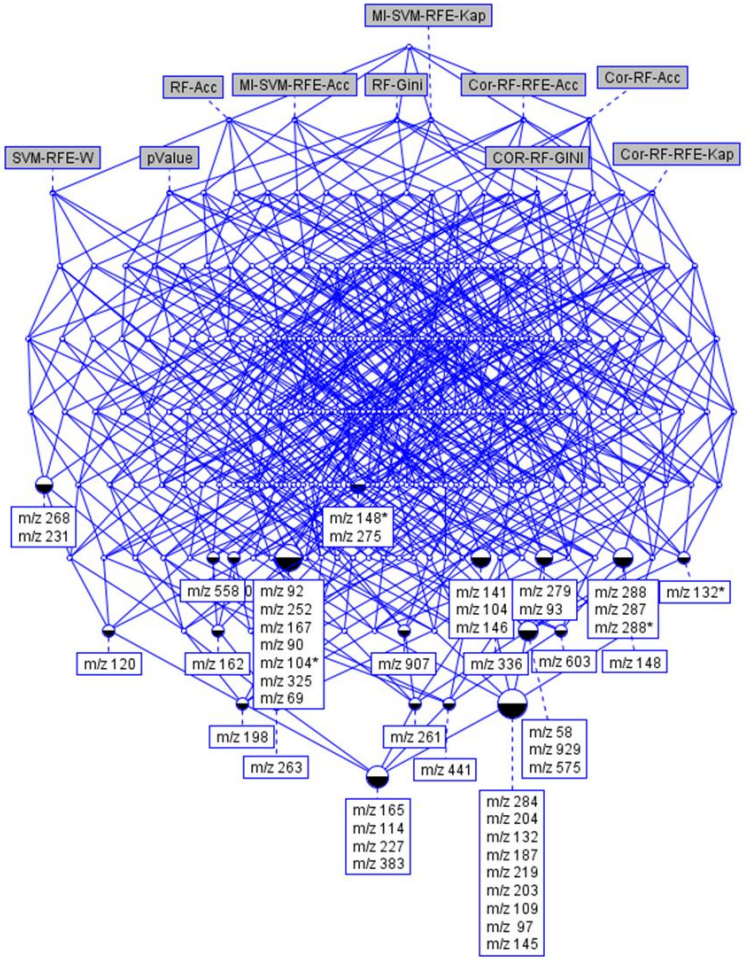
RESULTS AND VISUALIZATION

FCA

Methods

m/z variables

Features / Ions	COR-RF-Gini	Cor-RF-Acc	Cor-RF-RFE-Acc	Cor-RF-RFE-Kap	RF-Gini	RF-Acc	MI-SVM-RFE-Acc	MI-SVM-RFE-Kap	SVM-RFE-W	pValue
m/z 383	1	1	1	1	1	1	1	1	1	1
m/z 227	1	1	1	1	1	1	1	1	1	1
m/z 114	1	1	1	1	1	1	1	1	1	1
m/z 165	1	1	1	1	1	1	1	1	1	1
m/z 145	1	1	1	1	1	1	1	1		1
m/z 97	1	1	1	1	1	1	1	1		1
m/z 441	1	1	1	1	1	1	1		1	1
m/z 109	1	1	1	1	1	1	1	1		1
m/z 203	1	1	1	1	1	1	1	1		1
m/z 219	1	1	1	1	1	1	1	1		1
m/z 198	1		1	1	1	1	1	1	1	1
m/z 263	1	1	1		1	1	1	1	1	1
m/z 187	1	1	1	1	1	1	1	1		1
m/z 132	1	1	1	1	1	1	1	1		1
m/z 204	1	1	1	1	1	1	1	1		1
m/z 261	1	1	1	1	1	1	1	1	1	
m/z 162	1	1			1	1	1	1	1	1
m/z 284	1	1	1	1	1	1	1	1		1
m/z 603	1	1	1	1	1	1			1	1
m/z 148	1	1	1	1	1	1		1		1
m/z 575	1	1	1	1	1		1	1		
m/z 69	1	1			1	1	1	1		1
m/z 325	1	1			1	1	1	1		1
m/z 405	1	1			1	1	1		1	1
m/z 929	1	1	1	1	1	1	1	1		
m/z 58	1	1	1	1	1	1	1	1		
m/z 336	1	1	1	1	1	1	1			1
m/z 146	1		1	1	1		1	1		1
m/z 104	1		1	1	1		1	1		1
m/z 120			1	1	1	1	1	1	1	1
m/z 558	1	1			1	1	1	1	1	
m/z 231					1	1	1	1	1	1
m/z 132*	1	1	1	1	1		1	1		
m/z 93		1	1	1	1	1	1	1		
m/z 907		1	1	1		1	1	1	1	
m/z 279		1	1	1	1	1	1	1		
m/z 104*	1	1			1	1	1	1		1
m/z 90	1	1			1	1	1	1		1
m/z 268					1	1	1	1	1	1
m/z 288*	1	1	1	1	1	1				1
m/z 287	1	1	1	1	1	1				1
m/z 167	1	1			1	1	1	1		1
m/z 288	1	1	1	1	1	1				1
m/z 252	1	1			1	1	1	1		1
m/z 141	1		1	1	1		1	1		1
m/z 279			1	1	1		1	1		1
m/z 148*			1	1	1		1	1		1
m/z 92	1	1			1	1	1	1		1



- 48 metabolites selected with at least 6 methods

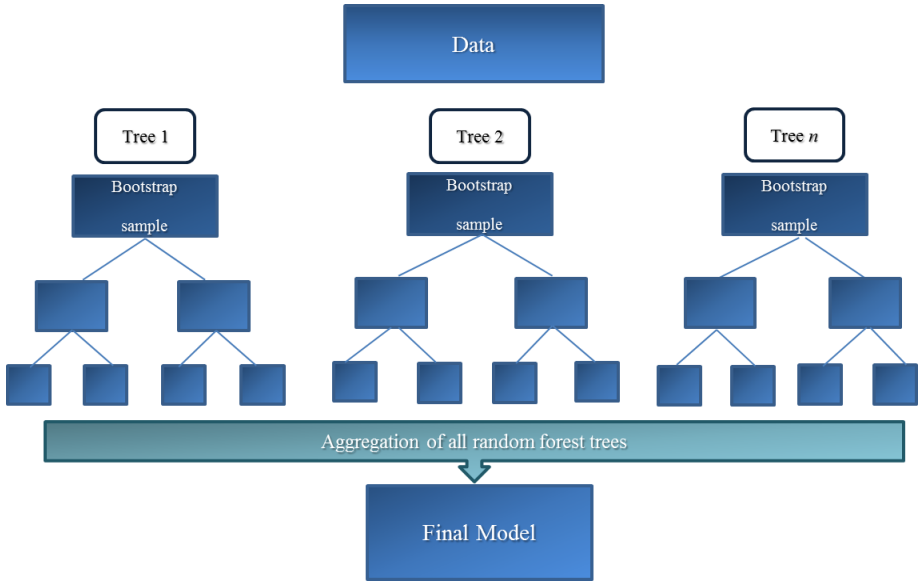
PREDICTIVE MODEL BUILDING

- Logistic regressions $Y = f(X_1, X_2, X_3 \ldots)$

  Prediction equation
  Low number of input variables required
  Elimination of correlated features

- Random Forest

Decision tree
Supports high number of input variables



PERFORMANCE EVALUATION

Indicators

		Measured	
		case	Control
Predicted	case	TP	FP
	Control	FN	TN

TP : true positive
FP : false positive
TN : true negative
FN : false negative

Sensibility = VP/(VP+FN)

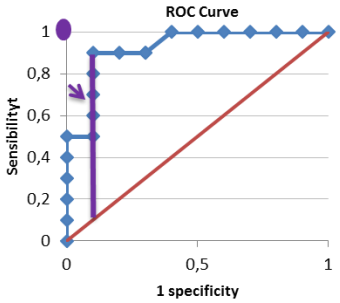
- Ratio of case predicted case

Specificity = VN/(VN+FP)

- Ratio of controls predicted controls

ROC curve (receiver operating characteristic)

- Determine an optimal threshold
- AUC (area under the curve): global model efficacy



Validation

Evaluation on training set

- Calculate indicators by predicting samples from training sets
- Optimistic evaluation (surestimate predictive capacity)

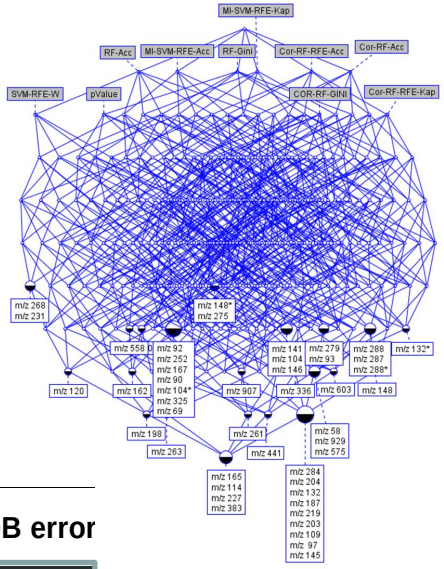
cross-validation

Iterative methods

Evaluation on validation set

- Calculate indicators by predicting samples from an independant validation
- Ideal when the subject number is big enough

FEATURE SELECTION FOR PREDICTIVE MODEL BUILDING



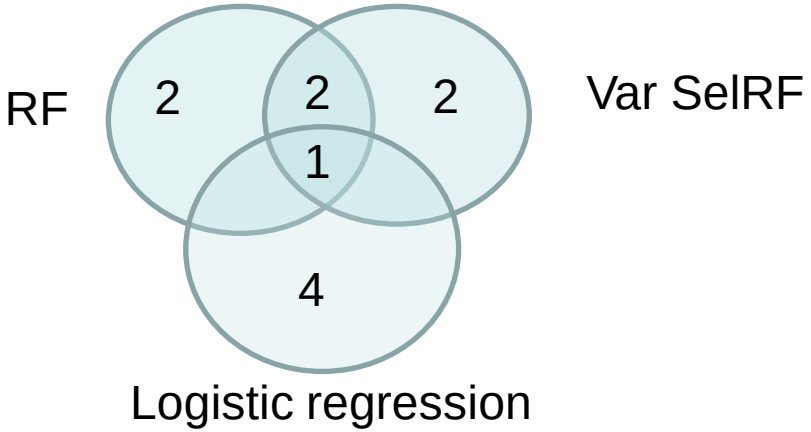
- Interest of working on reduced dataset

Metrics	Sensitivity	Specificity	Accuracy	Precision	Misclassification (%)	OOB error
1.195-Rf-acc	0.81	0.65	0.73	0.71	27	0.261
200-Rf-acc	0.86	0.82	0.84	0.84	16	0.154
48-Rf-acc	0.93	0.80	0.87	0.83	13	0.131
40-Rf-acc	0.85	0.88	0.87	0.87	13	0.131
30-Rf-acc	0.83	0.90	0.87	0.90	13	0.131
20-Rf-acc	0.90	0.85	0.88	0.86	12	0.119
10-Rf-acc	0.85	0.86	0.85	0.85	15	0.142
5-Rf-acc	0.86	0.85	0.85	0.86	14	0.142

PREDICTIVE MODELS

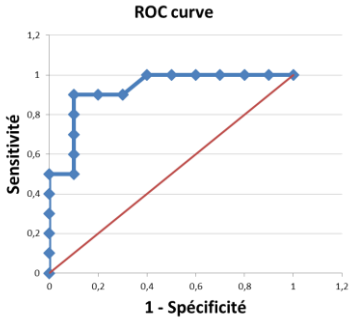
For prediction, the subset of the 48 stable top-ranked features was selected and different alternative techniques were used: RF, VarSelRF and logistic regression

- All final predictive models included 5 variables
- 11 selected variables in total



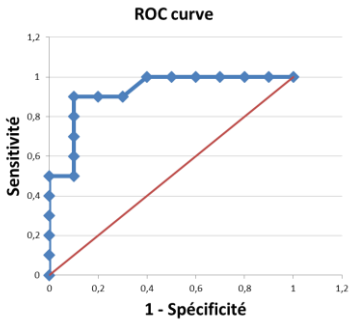
- Predictive capacity of the 11 selected variables

Features	AUC	*t*-tests	95 % CI
m/z 145	0.795	1.448E-6	0.657 - 0.896
m/z 97	0.787	1.597E-6	0.657 - 0.898
m/z 325	0.773	2.233E-5	0.627 - 0.896
m/z 268	0.759	4.564E-6	0.614 - 0.866
m/z 263	0.753	5.996E-6	0.642 - 0.874
m/z 219	0.712	1.177E-4	0.162 - 0.798
m/z 162	0.656	0.00195	0.225 - 0.710
m/z 288*	0.634	0.00499	0.252 - 0.708
m/z 148	0.630	0.01778	0.238 - 0.624
m/z 198	0.619	0.01368	0.197 - 0.594
m/z 167	0.541	0.01796	0.190 - 0.715



PERFORMANCE EVALUATION

Features	AUC	*t*-tests	95 % CI
m/z 145	0.795	1.448E-6	0.657 - 0.896
m/z 97	0.787	1.597E-6	0.657 - 0.898
m/z 325	0.773	2.233E-5	0.627 - 0.896
m/z 268	0.759	4.564E-6	0.614 - 0.866
m/z 263	0.753	5.996E-6	0.642 - 0.874
m/z 219	0.712	1.177E-4	0.162 - 0.798
m/z 162	0.656	0.00195	0.225 - 0.710
m/z 288*	0.634	0.00499	0.252 - 0.708
m/z 148	0.630	0.01778	0.238 - 0.624
m/z 198	0.619	0.01368	0.197 - 0.594
m/z 167	0.541	0.01796	0.190 - 0.715



	AUC	95% CI	Misclassification (%)	False positive	False negative
RF	0.830	0.72 - 0.94	19.8	9	13
VarSelRF	0.845	0.76 - 0.94	22.5	14	11
Logistic regression	0.820	0.75 - 0.89	18.0	10	10
Univariate analyses - top 5	0.831	0.73 - 0.93	23.4	12	14
Univariate analyses - top 11	0.869	0.67 - 0.96	18.9	12	9

- using the same number of features (5), univariate and multivariate modeling gave similar predictive results.

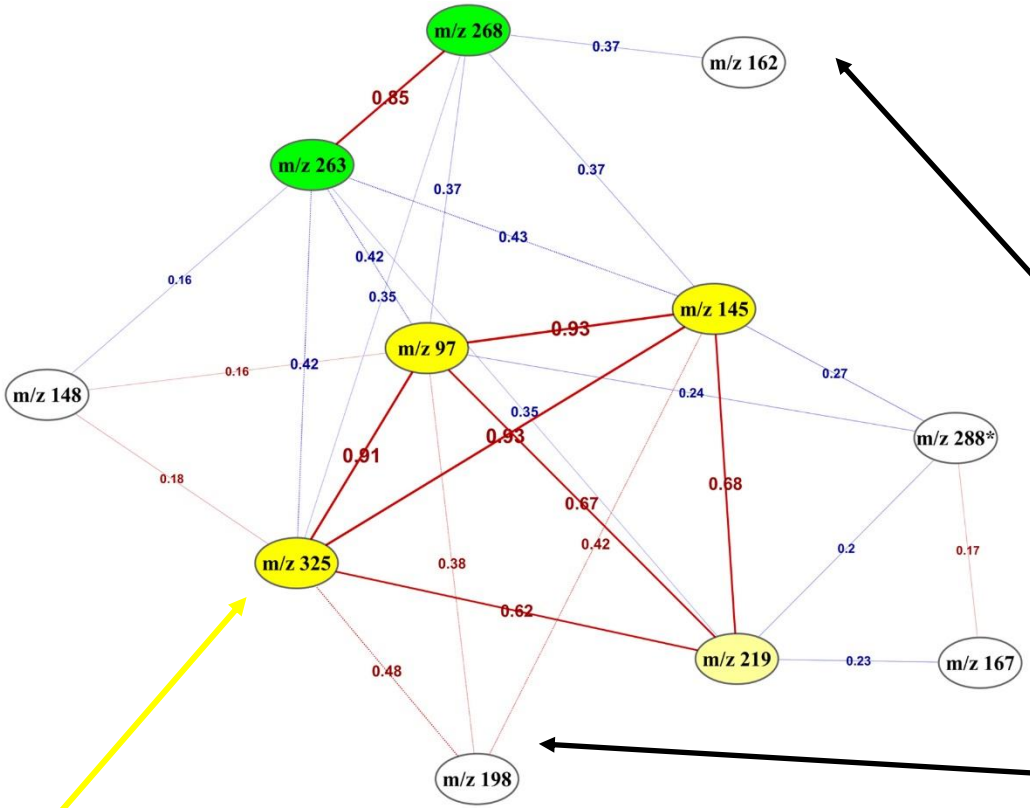
EVALUATION OF THE FEATURE SELECTION METHODS

Ranking of the 11 selected variables:

Features	RF-Acc	RF-Gini	Cor-RF-Gini	Cor-RF-Acc	Cor-RF-RFE-Acc	Cor-RF-RFE-Kap	MI-SVM-RFE-Acc	MI-SVM-RFE-Kap	SVM-RFE-W	Anova-p-value
m/z 145	1	1	1	2	46	53	100	125	323	2
m/z 97	2	2	3	1	142	185	63	67	159	3
m/z 325	5	4	7	5	210	220	38	37	1118	8
m/z 268	13	3	-	-	-	-	168	181	22	4
m/z 263	10	8	5	7	198	249	28	27	166	5
m/z 219	12	15	13	12	84	76	61	65	1022	12
m/z 162	438	31	20	26	211	221	39	38	103	17
m/z 288*	19	53	25	29	140	152	-	-	976	22
m/z 148	384	30	27	86	87	98	66	70	471	38
m/z 198	199	117	150	496	48	36	70	84	167	34
m/z 167	16	70	45	24	505	586	144	154	13	39

- RF combined with ANOVA provided the best feature selection

CORRELATION NETWORKS



Logistic regression

RF

CONCLUSION

- Interest of feature selection methods to identify hidden information in such high dimensional datasets
- Importance of working on reduced datasets to obtain better performances in predictive models
- RF in parallel to ANOVA provided the best feature selection for predictive biomarker discovery

Our recommendation would be to explore these data mining methods !

ACKNOWLEDGMENTS

INRA SCIENCE & IMPACT *UMR1019- Human Nutrition Unit*

MAPPING

METABOHUB

Plate-Forme EXPLORATION du METABOLISME *des gènes aux métabolites*

Inria INVENTEURS DU MONDE NUMÉRIQUE

Loria Laboratoire lorrain de recherche en informatique et ses applications

Inserm

Université de Versailles

From metabolomics to systems biology

"When a thing was new, people said, 'It is not true'. Later, when the truth became obvious, people said, 'Anyway, it is not important.' And when its importance could not be denied, people said, 'Anyway, it is not new." William James (1842–1920).

Goodacre *et al*., 2004

Thanks for your attention