



**JOURNÉE R**  
**Vendredi 24 mai 2013**

# Le package XCMS pour l'analyse de profils de chromatographie liquide – spectrométrie de masse

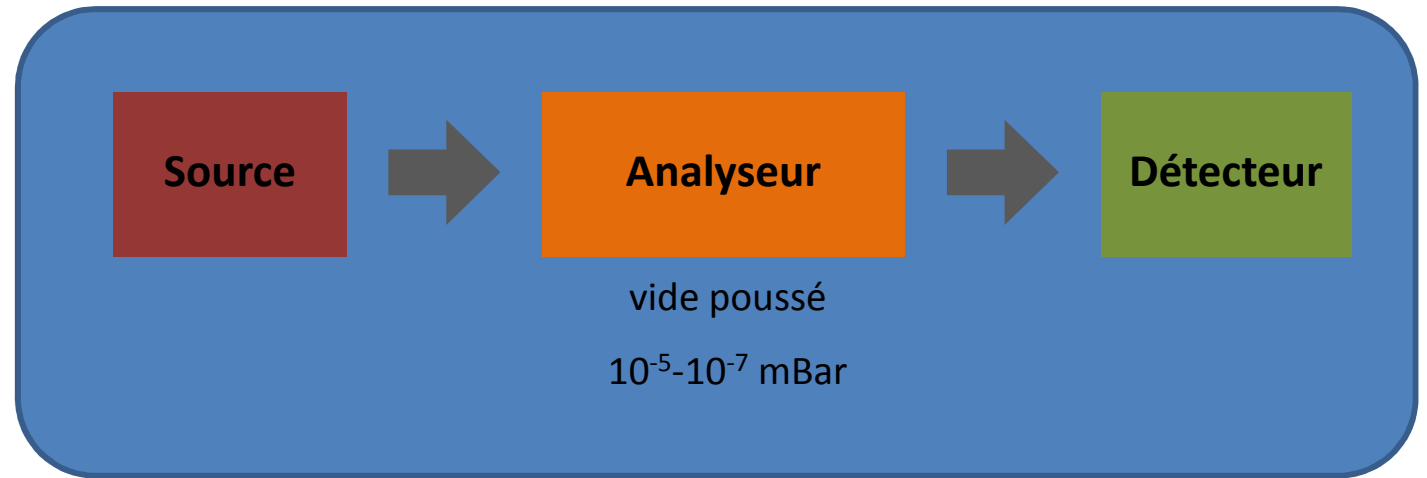
**Séverine ZIRAH ([szirah@mnhn.fr](mailto:szirah@mnhn.fr))**

Muséum National d'Histoire Naturelle, Département RDDM

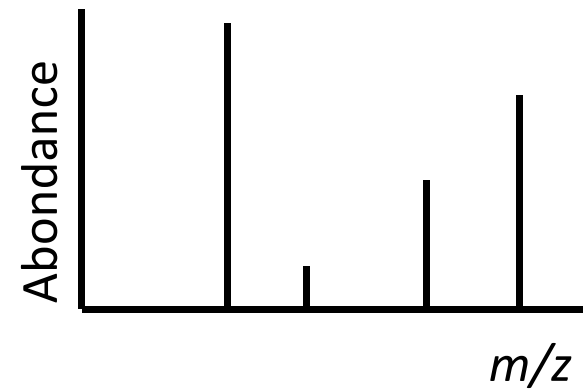
UMR 7245 CNRS - MNHN

Molécules de communication et adaptation des micro-organismes (MCAM)

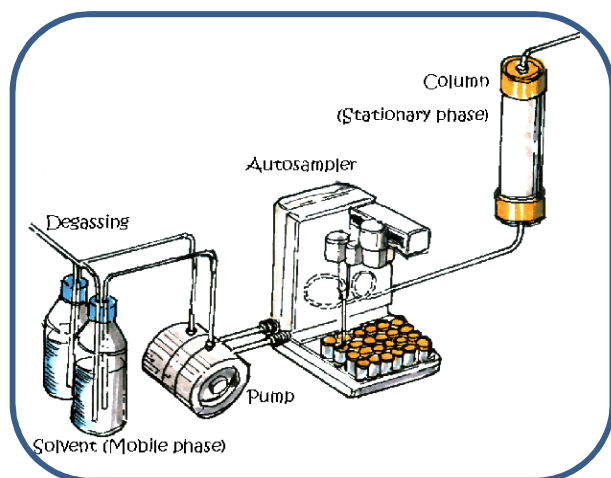
# Couplage chromatographie liquide – spectrométrie de masse (LC-MS)



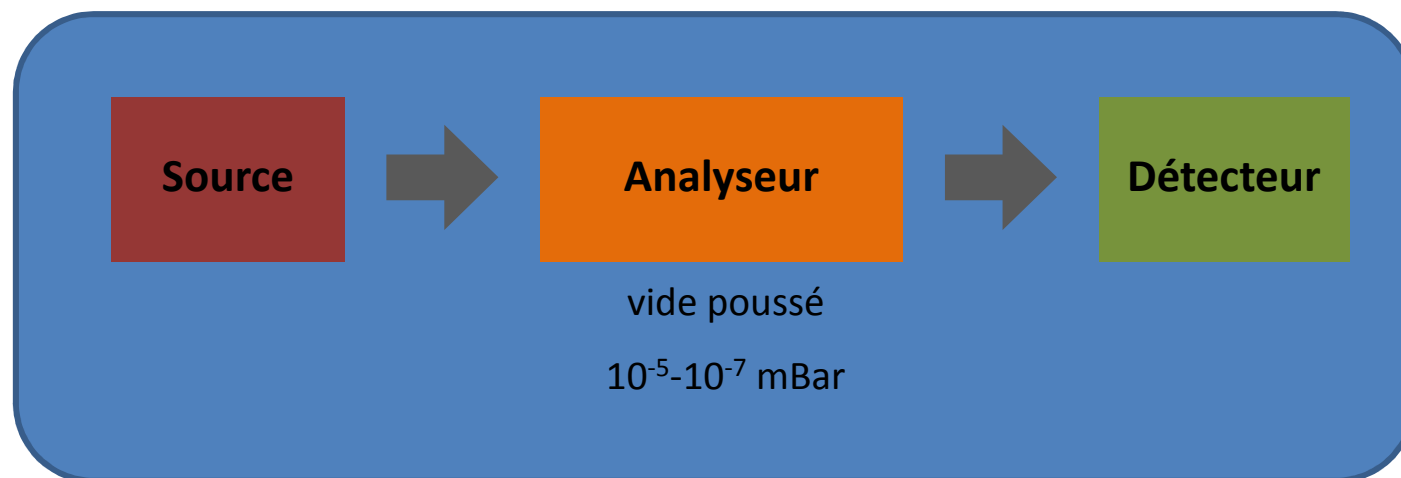
Spectromètre de masse



# Couplage chromatographie liquide – spectrométrie de masse (LC-MS)

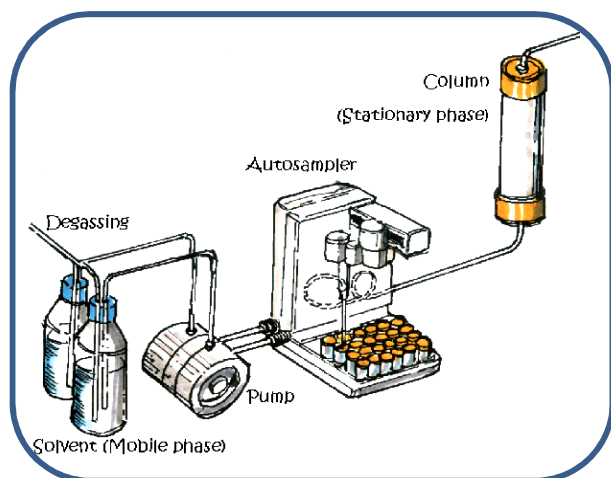


Chromatographie liquide  
haute performance (HPLC)

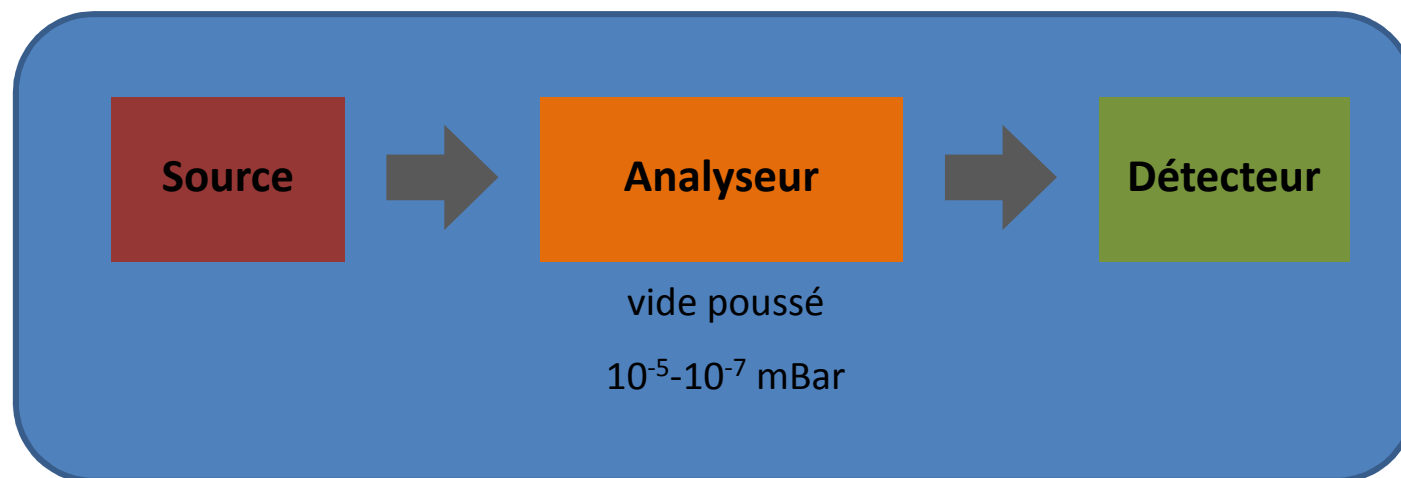


Spectromètre de masse

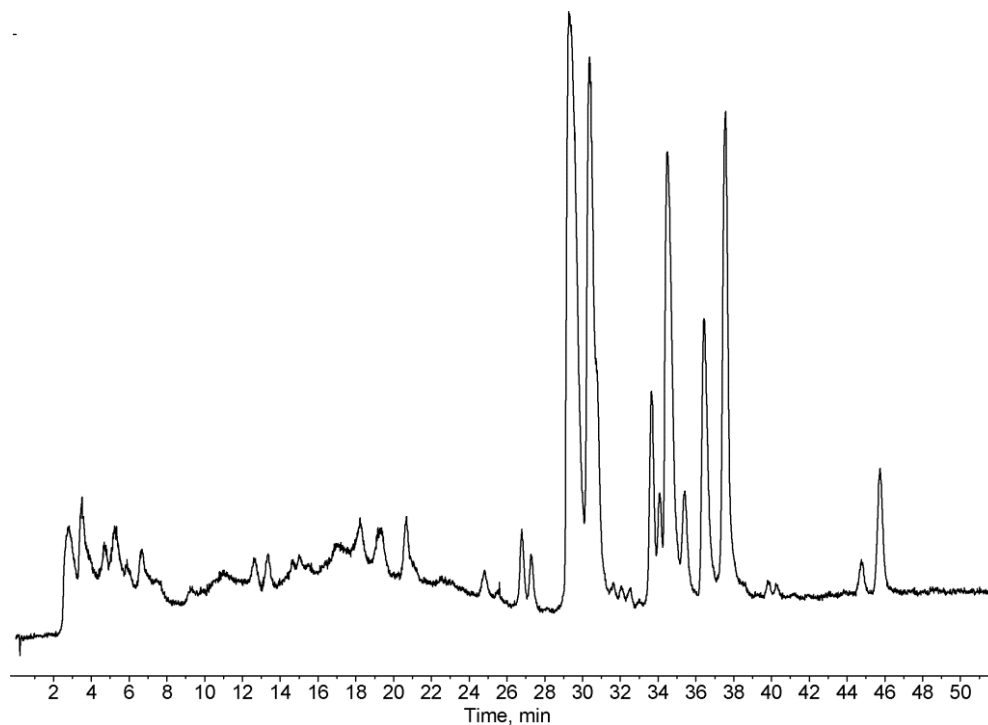
# Couplage chromatographie liquide – spectrométrie de masse (LC-MS)



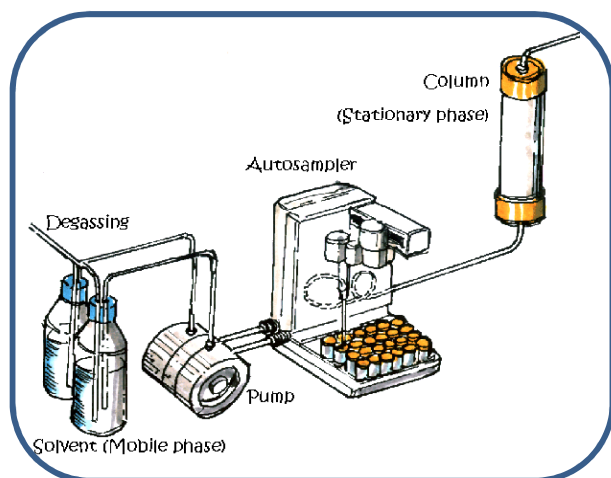
Chromatographie liquide  
haute performance (HPLC)



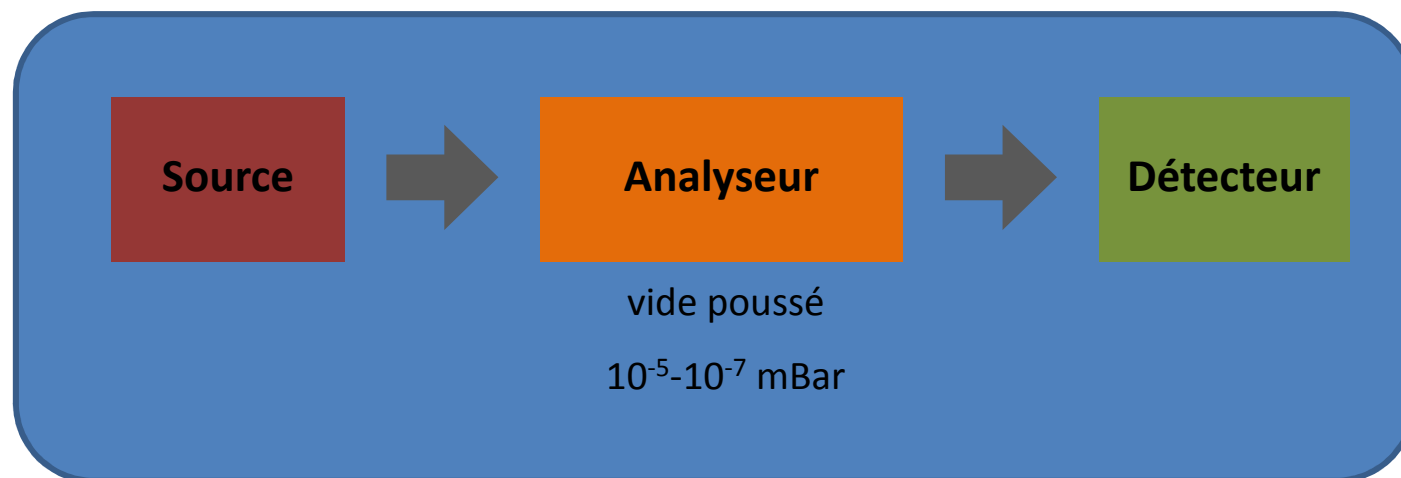
Spectromètre de masse



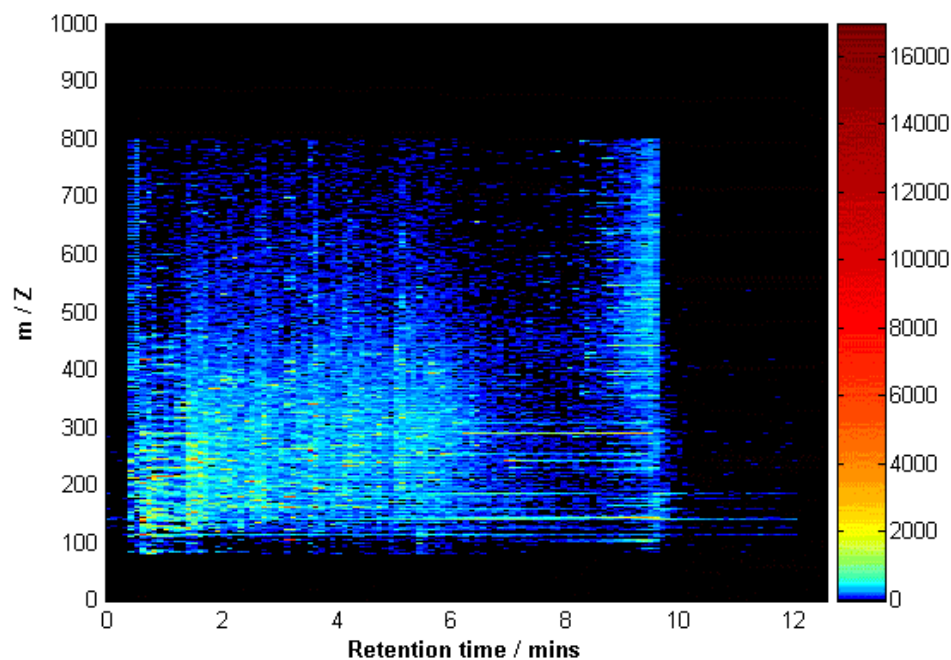
# Couplage chromatographie liquide – spectrométrie de masse (LC-MS)



Chromatographie liquide haute performance (HPLC)

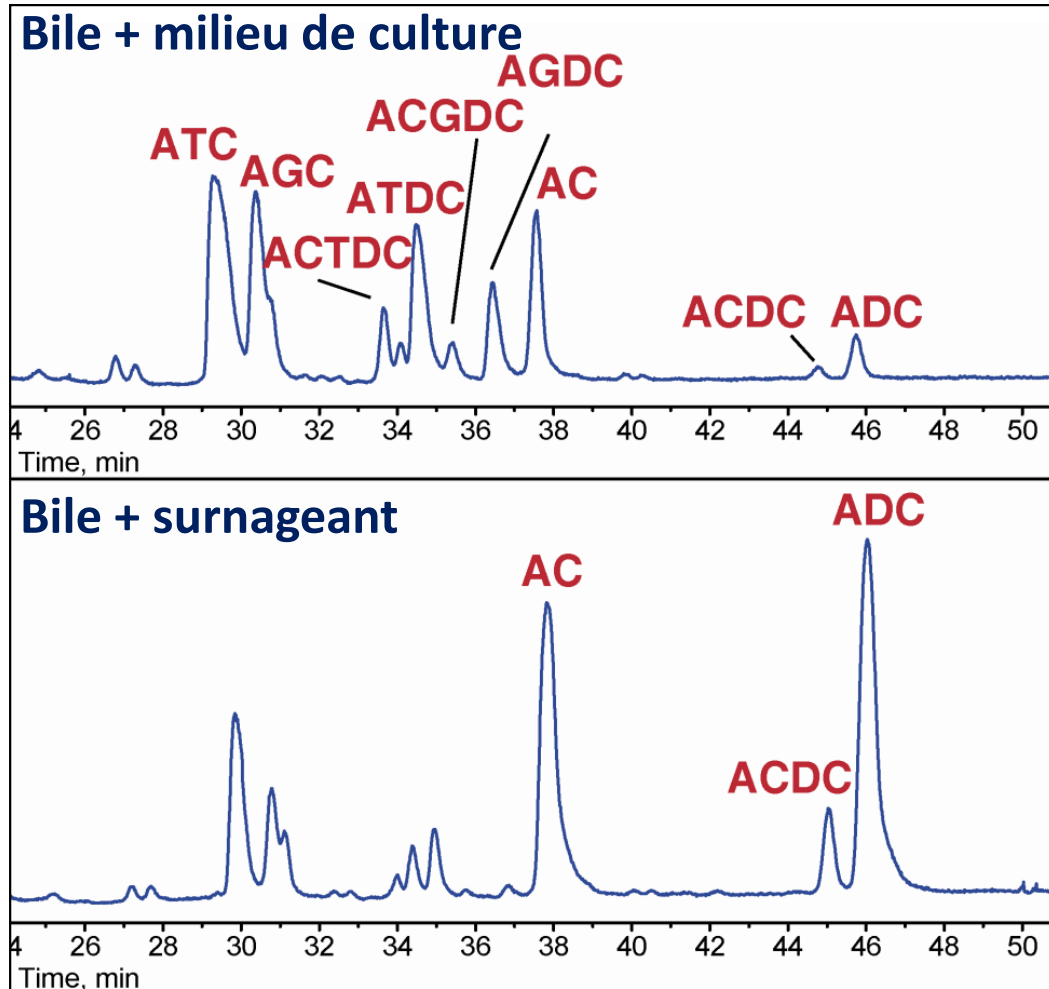


Spectromètre de masse



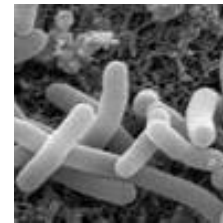
# Comparaison de profils LC-MS

## Profil des sels biliaires après incubation avec un surnageant bactérien



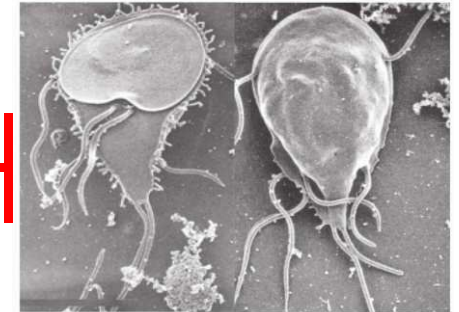
## Interaction entre *Giardia lamblia* et *Lactobacillus johnsonii*

Collaboration Marie-Agnès Travers, Cissé Sow, Isabelle Florent, Philippe Grellier



*Lactobacillus johnsonii* LA1

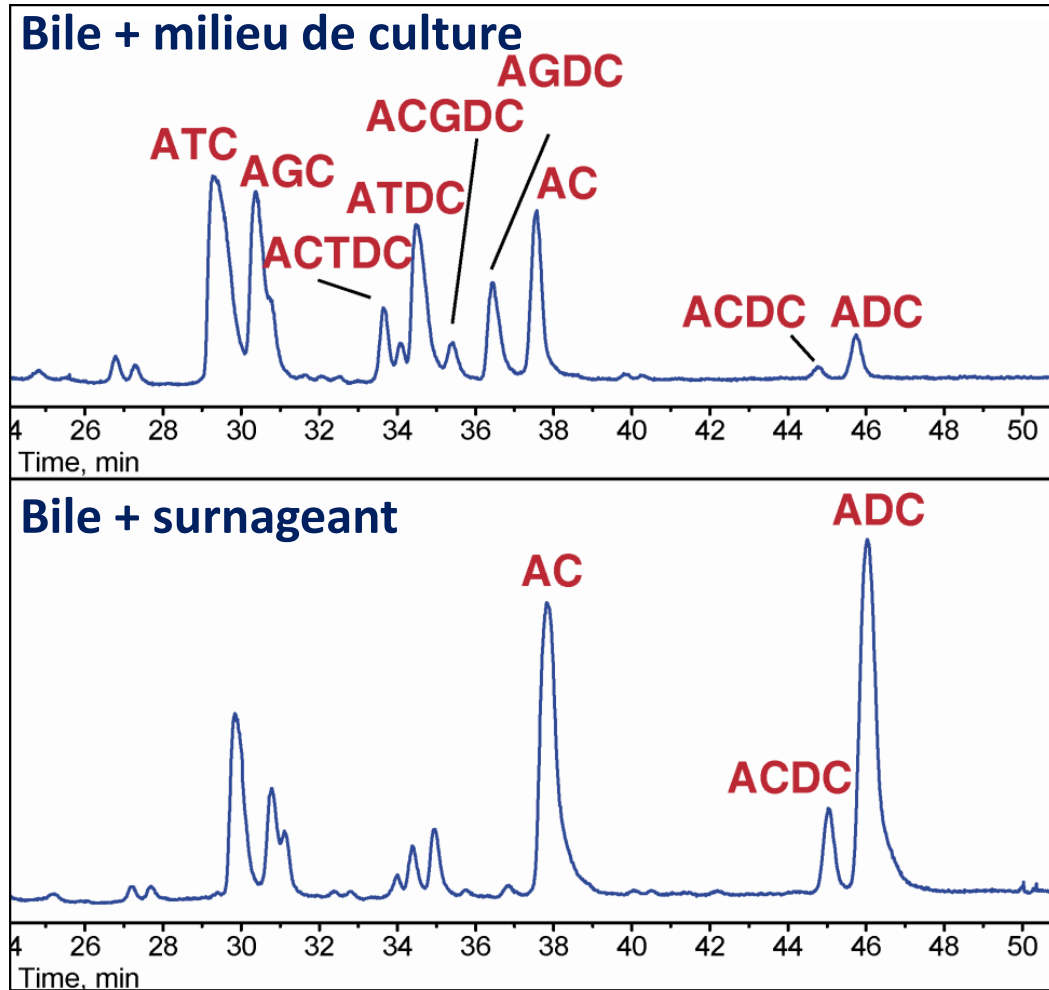
surnageant



*Giardia lamblia*

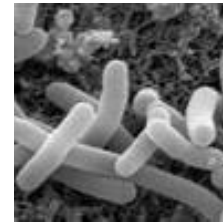
# Comparaison de profils LC-MS

## Profil des sels biliaires après incubation avec un surnageant bactérien

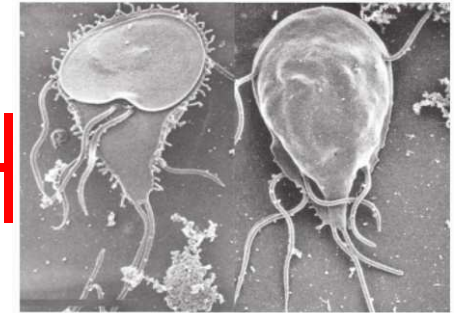


## Interaction entre *Giardia lamblia* et *Lactobacillus johnsonii*

Collaboration Marie-Agnès Travers, Cissé Sow, Isabelle Florent, Philippe Grellier



*Lactobacillus johnsonii* LA1



*Giardia lamblia*

surnageant

**ADC :** Acide déoxycholique

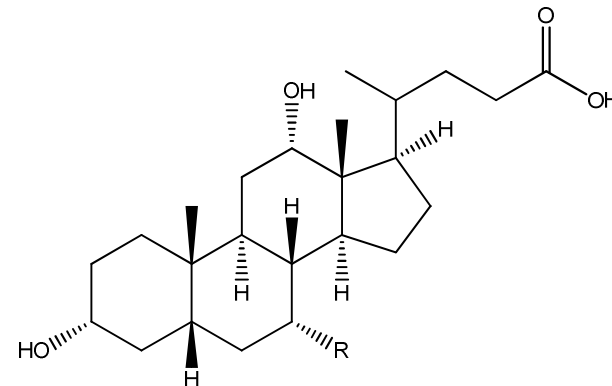
R=H

[M-H]<sup>-</sup> : m/z 391,2

**AC :** Acide cholique

R=OH

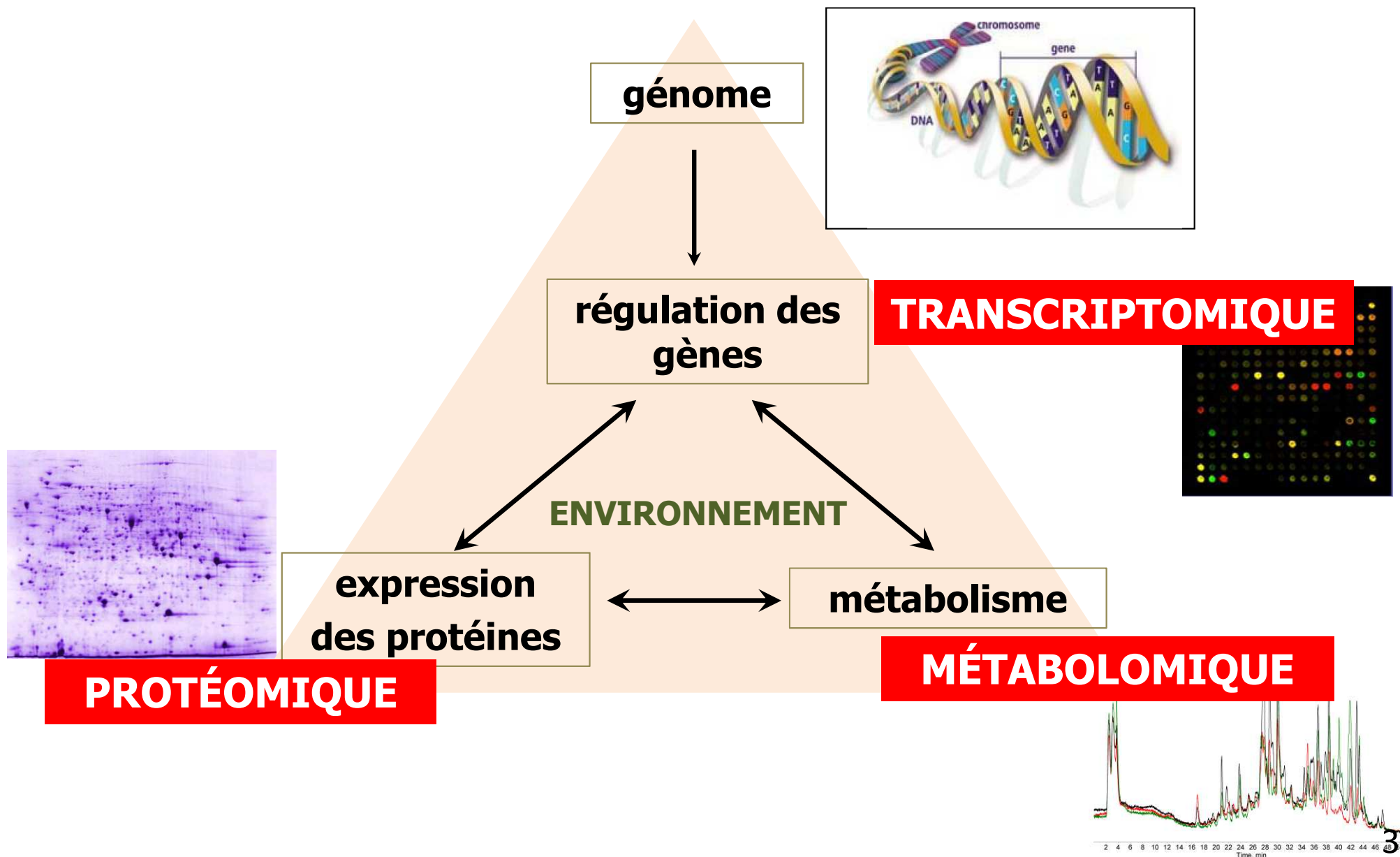
[M-H]<sup>-</sup> : m/z 407,2



# Comparaison de profils LC-MS pour la métabolomique

## METABOLOMIQUE

Mesure quantitative de l'ensemble des métabolites dans un système cellulaire sous des conditions données... ceux qui sont détectables avec les méthodes analytiques utilisées.



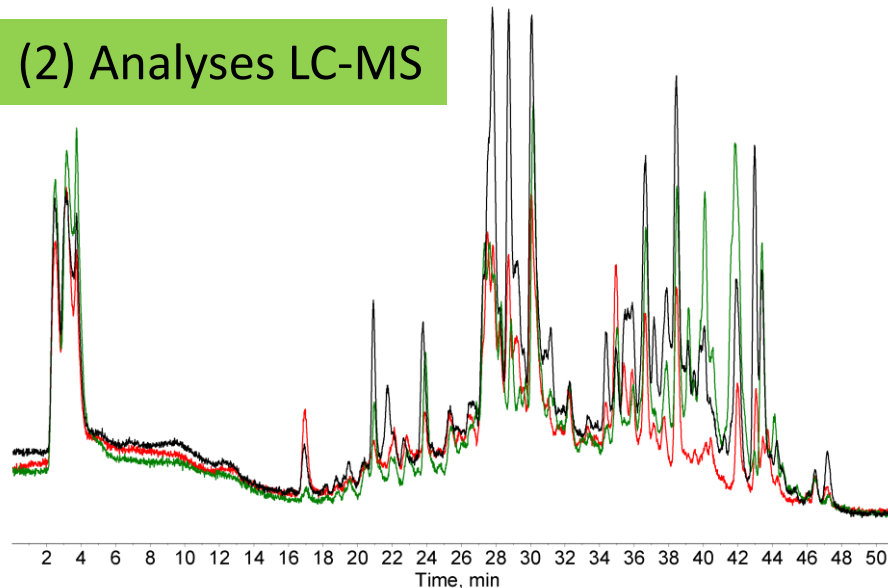


# Comparaison de profils LC-MS pour la métabolomique

(1) Préparation des échantillons



(2) Analyses LC-MS



(3) Traitement des données



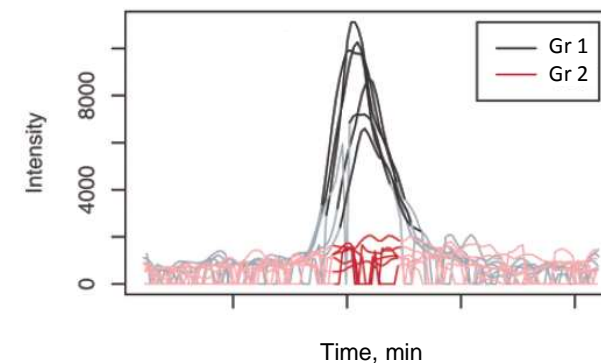
(4) Analyse statistique

name	fold	tstat	pvalue	mzmed	rtmed	npeaks	Groupe1	Groupe2	Ech1	Ech2	Ech3
1 M268T234	3.424489552	4.834454003	2.04E-05	268.10	233.95	20	1	19	152.63	160.65	549.13
2 M532T1458	1.908698484	-3.976496454	0.001110454	532.16	1457.87	24	12	12	452.10	195.04	481.12
3 M284T1624	1.968937295	-4.016948562	0.001119564	284.21	1624.32	37	12	24	1261.31	841.17	1460.25
4 M641T2099	2.492096758	-3.854338971	0.001872463	641.19	2098.51	45	12	33	2665.65	736.15	3005.83

(5) Identification des composés d'intérêt  
→ biomarqueurs

Recherche dans des bases de données  
Comparaison avec un standard  
Expériences MS/MS

Isolement et caractérisation par d'autres méthodes analytiques (RMN, IR...)

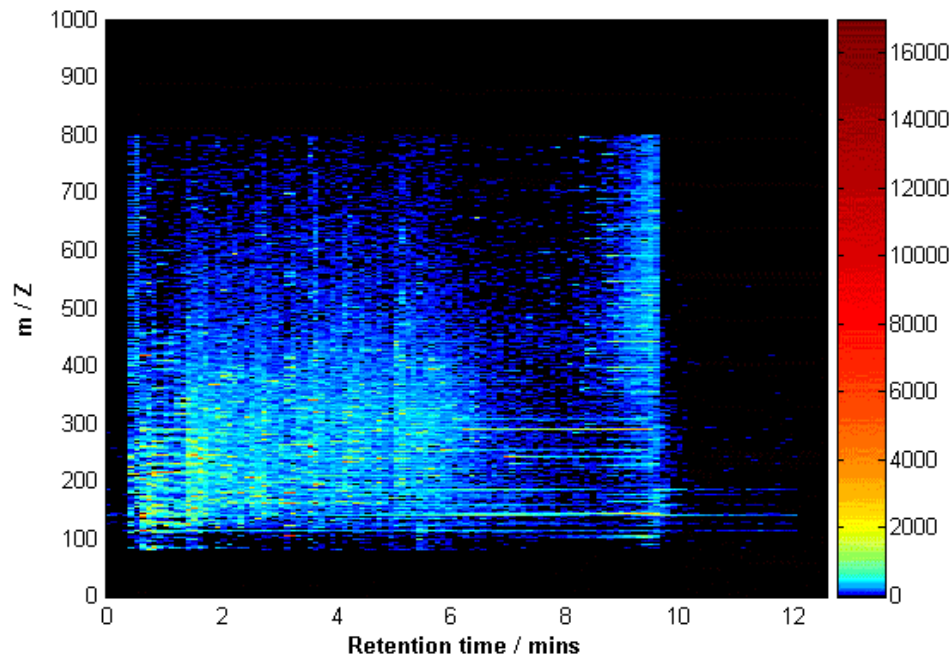


# Exigences pour l'analyses de données LC-MS

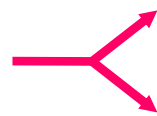
**n analyses LC-MS**



**Matrice de données  
 $n \times k$**



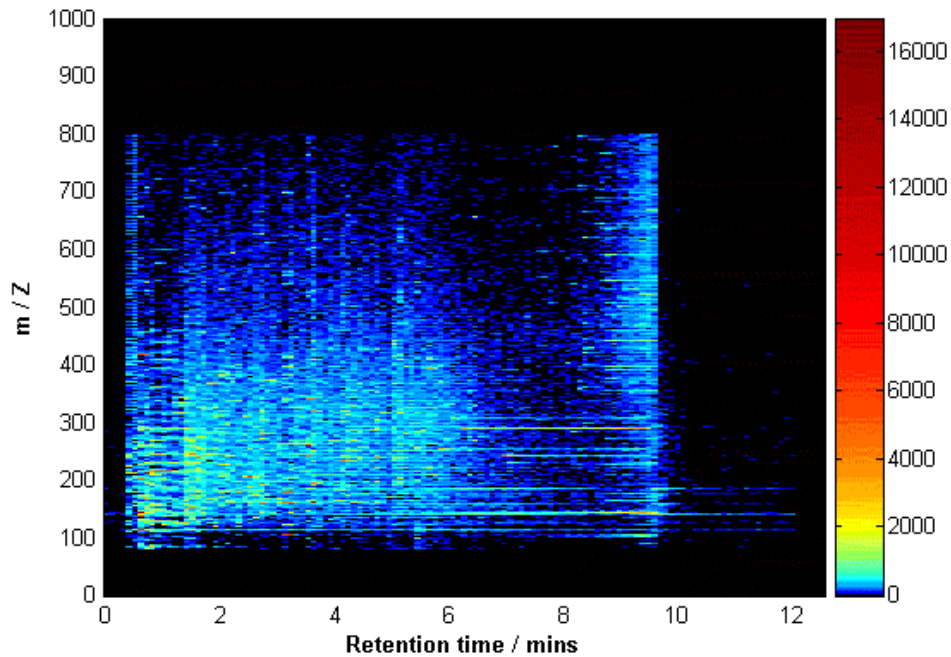
- Chaque échantillon est caractérisé par le **même nombre de variables**
- Chacune de ces variables est représentée sur **l'ensemble des individus**
- Une variable d'un échantillon a la **même signification biologique** (représente le même métabolite) pour tous les autres échantillons
- Capacité de prédiction



Somme sur le temps de rétention

Alignement // Décomposition en fenêtres de temps

## n analyses LC-MS



## n individus répartis en 2 groupes

→ Matrice des k pics ( $t_r$ ,  $m/z$ ) avec les intensités pour chaque individu

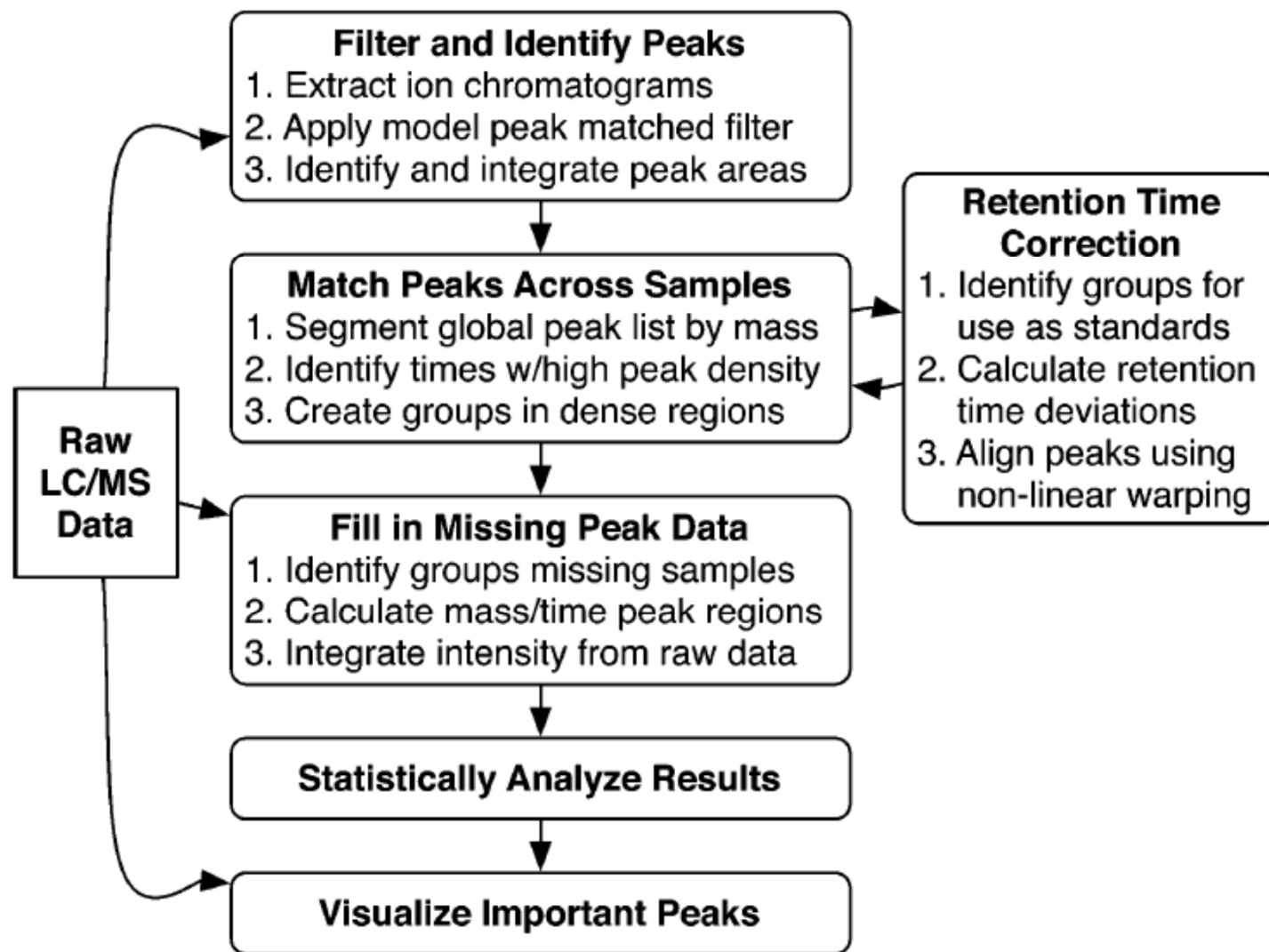


Analyse statistique

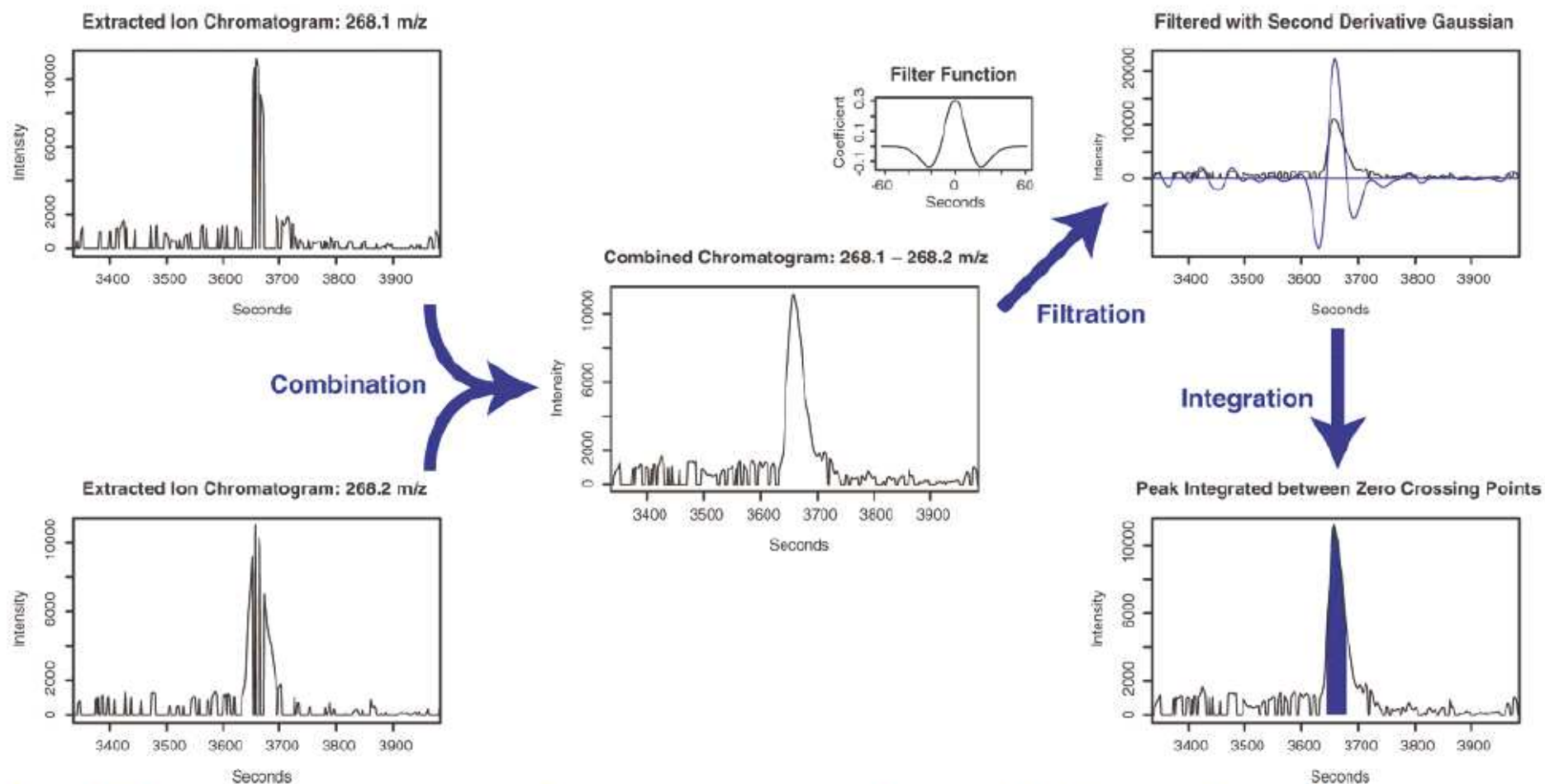


Proposition de « biomarqueurs »

# Le package xcms pour l'analyse de données LC-MS et GC-MS



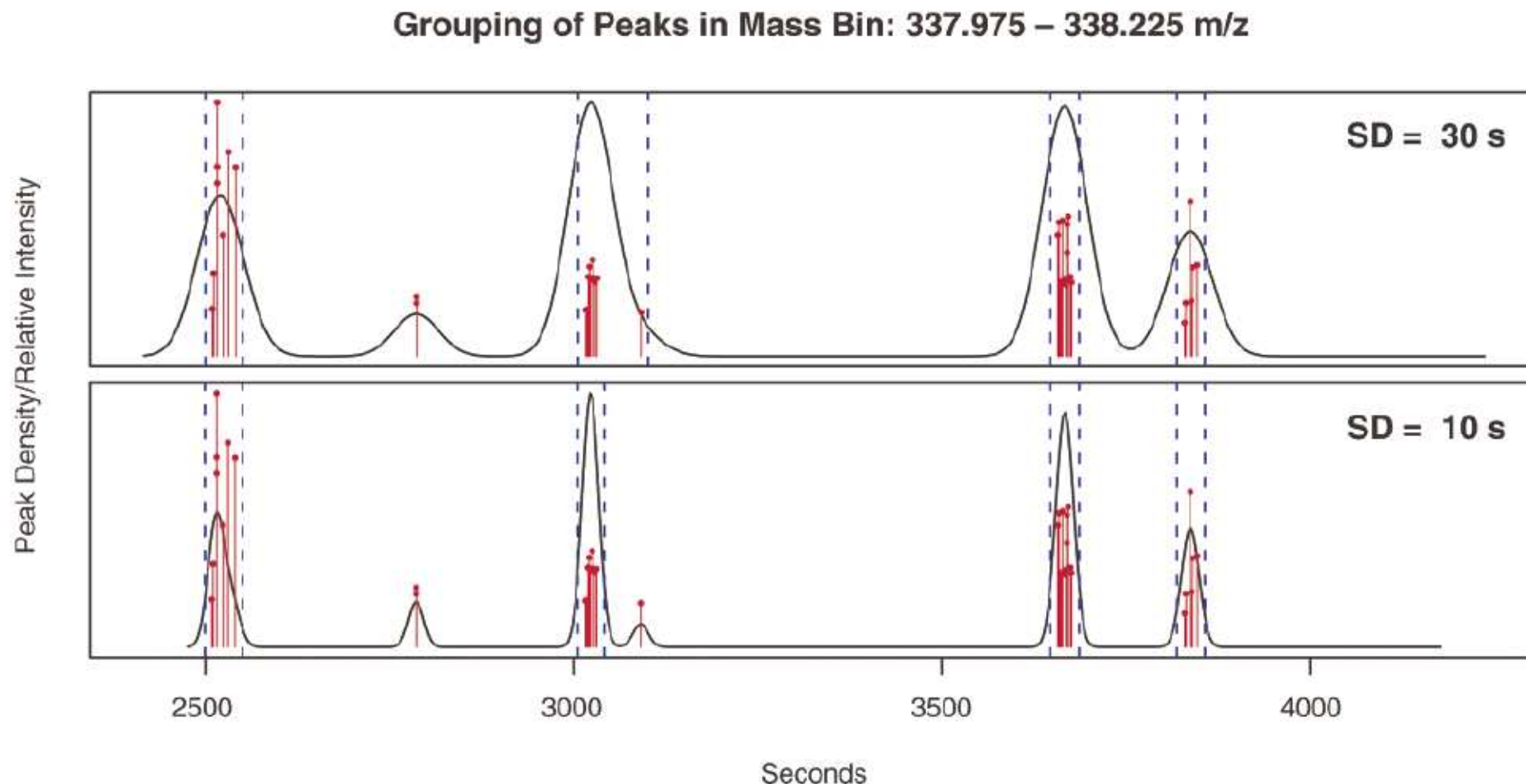
# xcms : (1) détection des pics



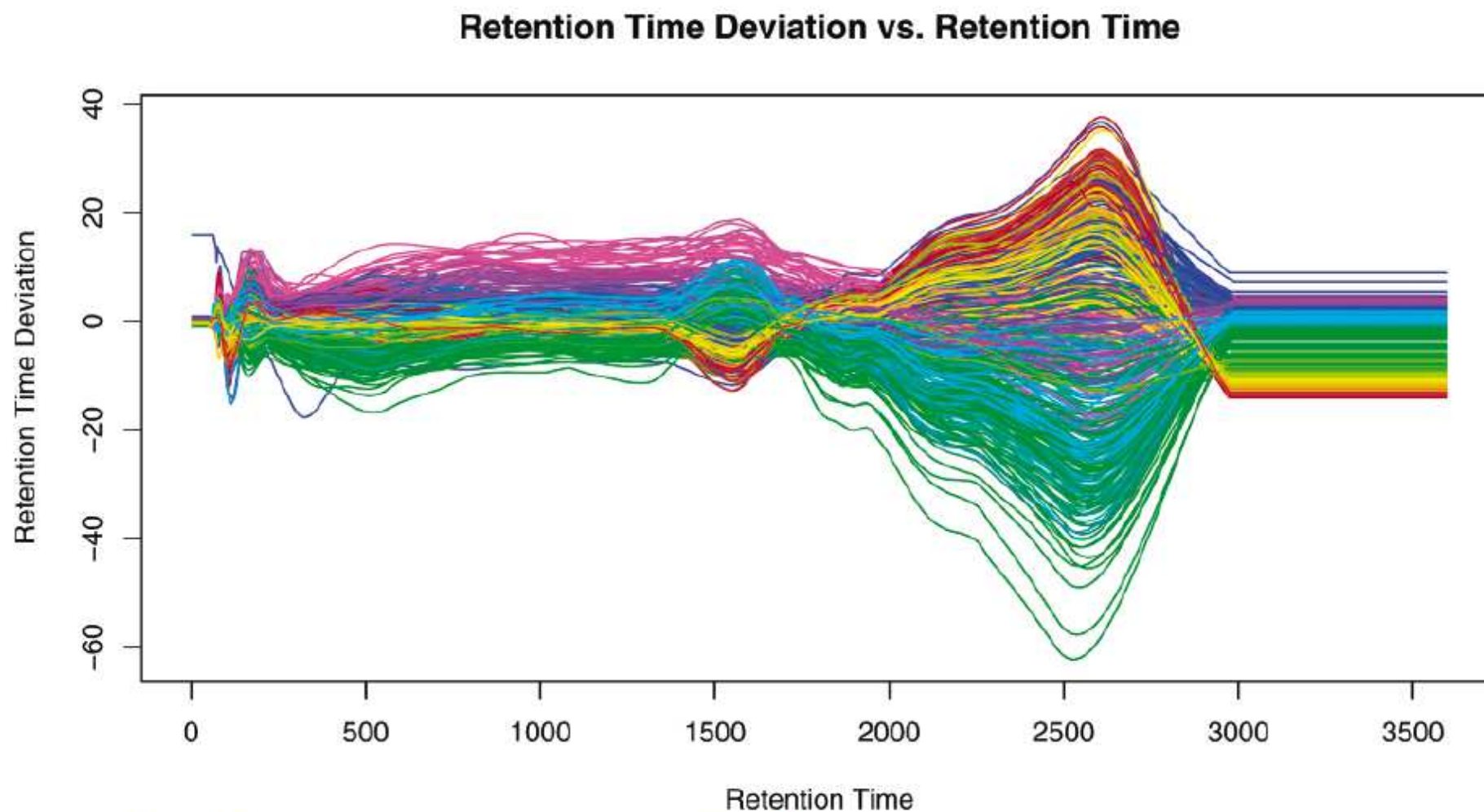
**Figure 2.** Illustration of the peak detection method using a single peak from the centroided FAAH knockout data. The peak shown here is contained in two adjacent chromatographic slices at 268.1 and 268.2  $m/z$ . The algorithm creates overlapping combined chromatograms (i.e.,  $m/z$  268.0/268.1, 268.1/268.2, 268.2/268.3, etc.) with only the  $m/z$  268.1/268.2 chromatogram showing a clean signal. The data are then processed with a matched filter whose coefficients are equal to a second-derivative Gaussian function. The second-derivative transformation causes the filtered chromatogram to cross the x-axis roughly at the peak inflection points. Those zero-crossing points define the area of peak integration. The peak tails, which typically contain only a small fraction of the signal and may overlap with other peaks, are not integrated.



# xcms : (2) correspondance des pics entre échantillons

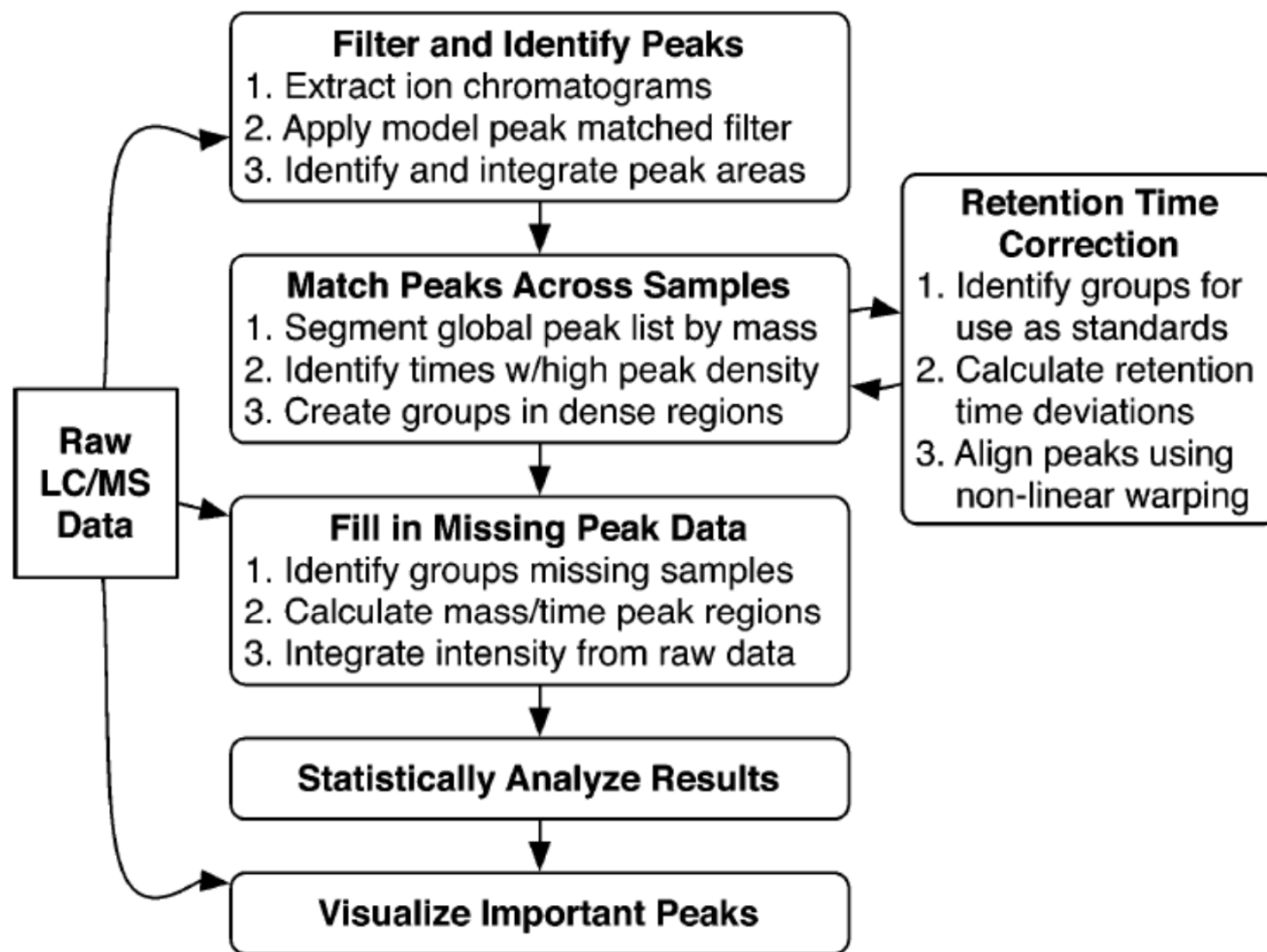


**Figure 3.** Example of cross-sample peak matching from the FAAH data involving 12 samples. Individual peaks are shown as sticks indicating relative intensity. The smoothed peak density profile is shown as a solid continuous line. Identified groups are flanked by dashed lines. The peak density profiles were smoothed with Gaussian functions of SD 30 and 10 s, respectively. Note how decreased smoothing eliminates a peak from the second group.



**Figure 4.** Retention time deviation profiles used for aligning 476 LC/MS analyses from the serum samples. All times are in seconds. The deviation profiles were created after two iterative steps of retention time alignment. A positive deviation indicates that the sample was eluting after the median retention time, and a negative deviation indicates that the sample was eluting before the median retention time. Sample profiles are colored in a rainbow by the order in which they were run, with red being the first samples and violet being the last samples run.

# xcms : (4) remplissage des données de pics manquantes

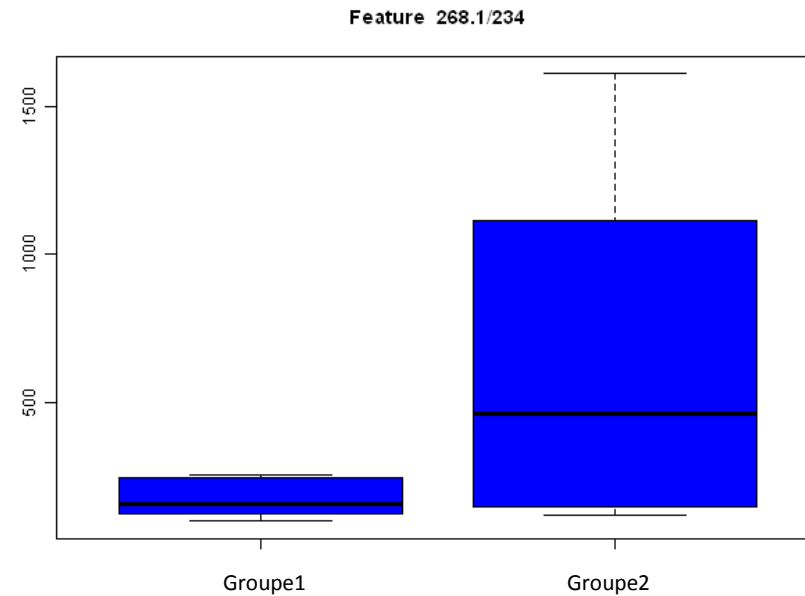
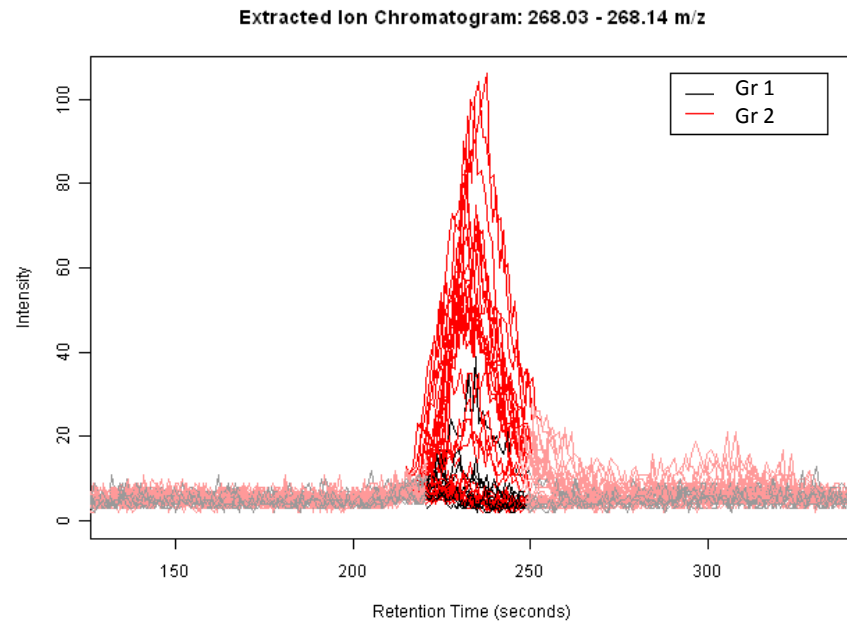




# xcms : (5) analyse statistique : test t univarié pour chaque pic

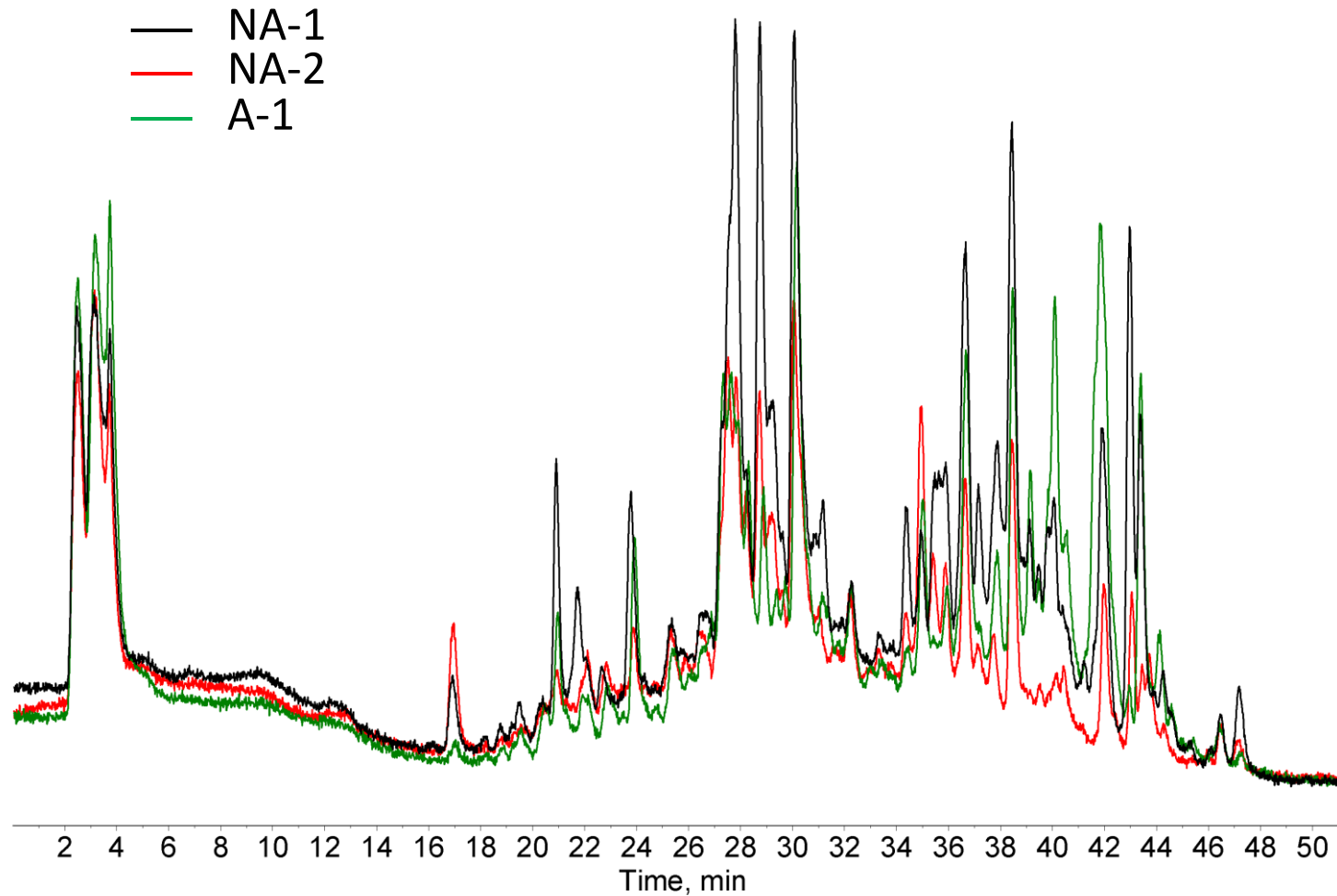
	name	fold	tstat	pvalue	mzmed	rtmed	npeaks	Groupe1	Groupe2	Ech1	Ech2	Ech3
1	M268T234	3.42	4.83	2.04E-05	268.10	233.95	20	1	19	152.63	160.65	549.13
2	M532T1458	1.91	-3.98	0.00111045	532.16	1457.87	24	12	12	452.10	195.04	481.12
3	M284T1624	1.97	-4.02	0.00111956	284.21	1624.32	37	12	24	1261.31	841.17	1460.25
4	M641T2099	2.49	-3.85	0.00187246	641.19	2098.51	45	12	33	2665.65	736.15	3005.83
5	M639T2099	2.56	-3.83	0.00192063	639.18	2098.62	45	12	33	31828.55	7503.64	34338.54
6	M640T2099	2.56	-3.80	0.00205874	640.19	2098.52	45	12	33	10205.49	2502.55	11341.15

## xcms : (6) visualisation des pics d'intérêt



# Exemple : La cécidomyie du buis / Collaboration Michel Baylac


## Profil d'extraits de feuilles buis attaqués ou non par des cécidomyies



→ Conversion en format universel : NetCDF

→ Répartition des fichiers NetCDF en dossiers par groupe

 Attaque

 Non attaque

```
cdffiles<-list.files(myDataDir,recursive=TRUE,full.names=TRUE)
```

```
xset <- xcmsSet(cdffiles)
```

```
xset <- group(xset)
```

```
xset2 <- retcor(xset,family="symmetric",plottype="mdevden")
```

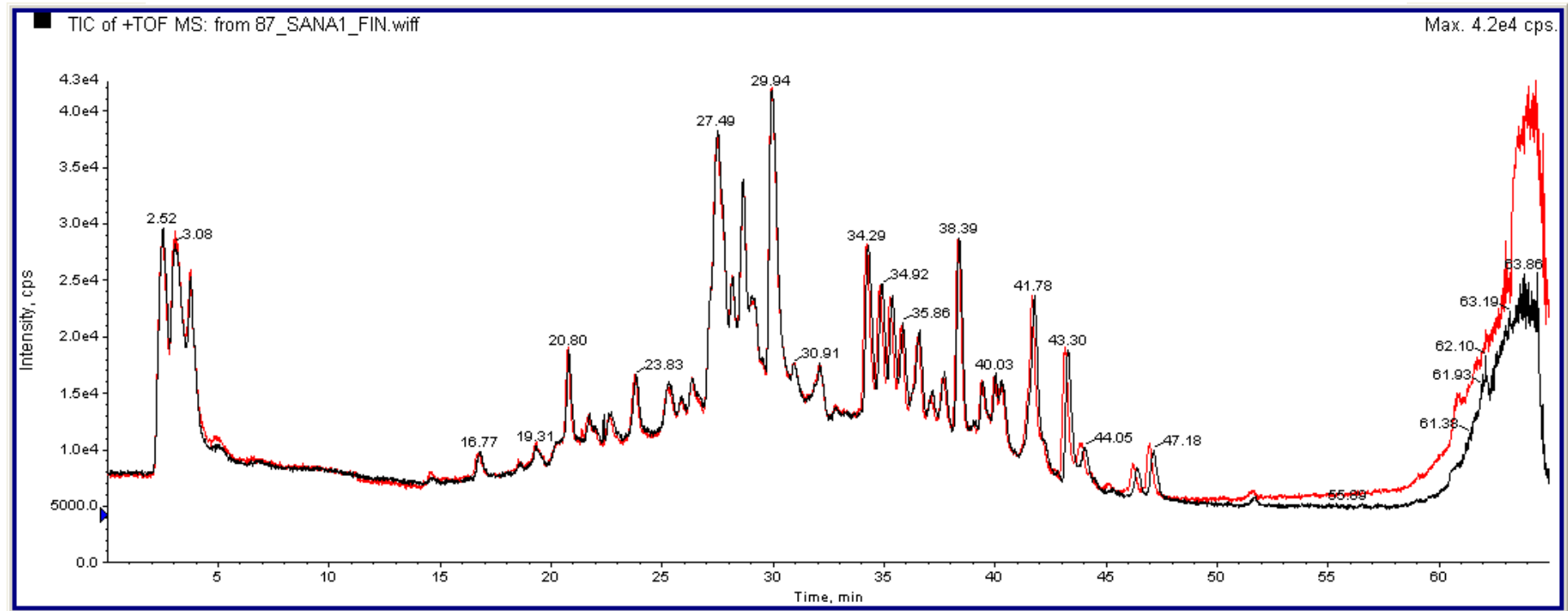
```
xset2 <- group(xset2)
```

```
xset3 <- fillPeaks(xset2)
```

```
reporttab <- diffreport(xset3, "Attaque", "Non attaque", "example", 10, metlin = 0.15)
```

# Choix de la gamme de temps de rétention

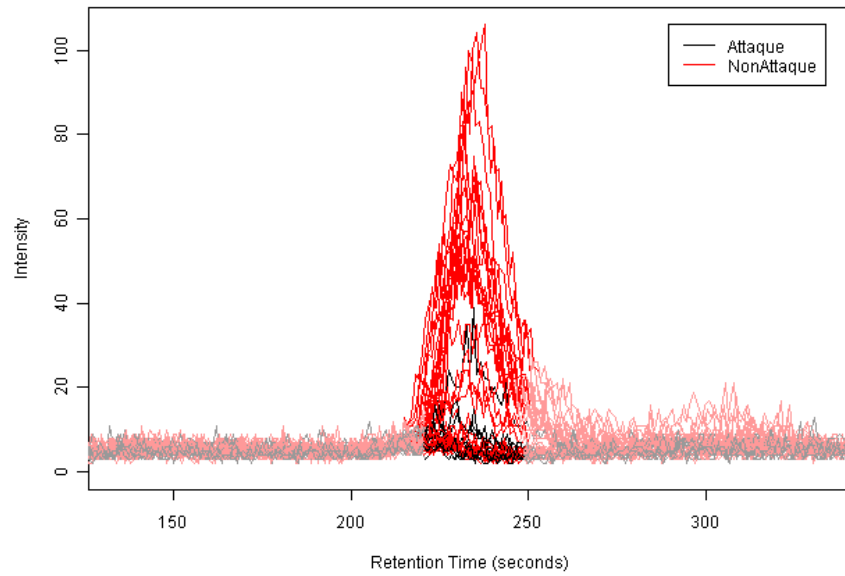
1<sup>er</sup> échantillon de la série 3 (en rouge) et même échantillon lancé après les 86 runs de la série 3 (en noir)



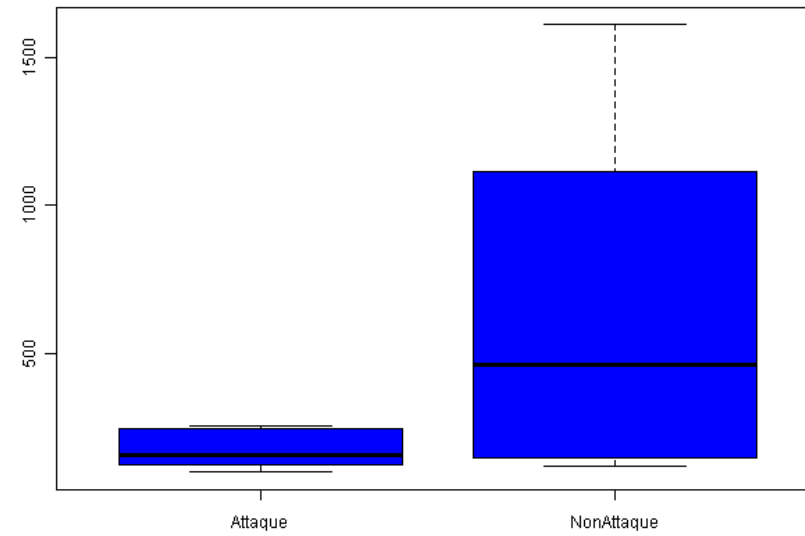
```
idx<-which(xset@peaks[,"rt"] < 3000)  
head(idx)  
xset@peaks<-xset@peaks[idx,]
```

# Contenu de « diffreport »

Extracted Ion Chromatogram: 268.03 - 268.14 m/z

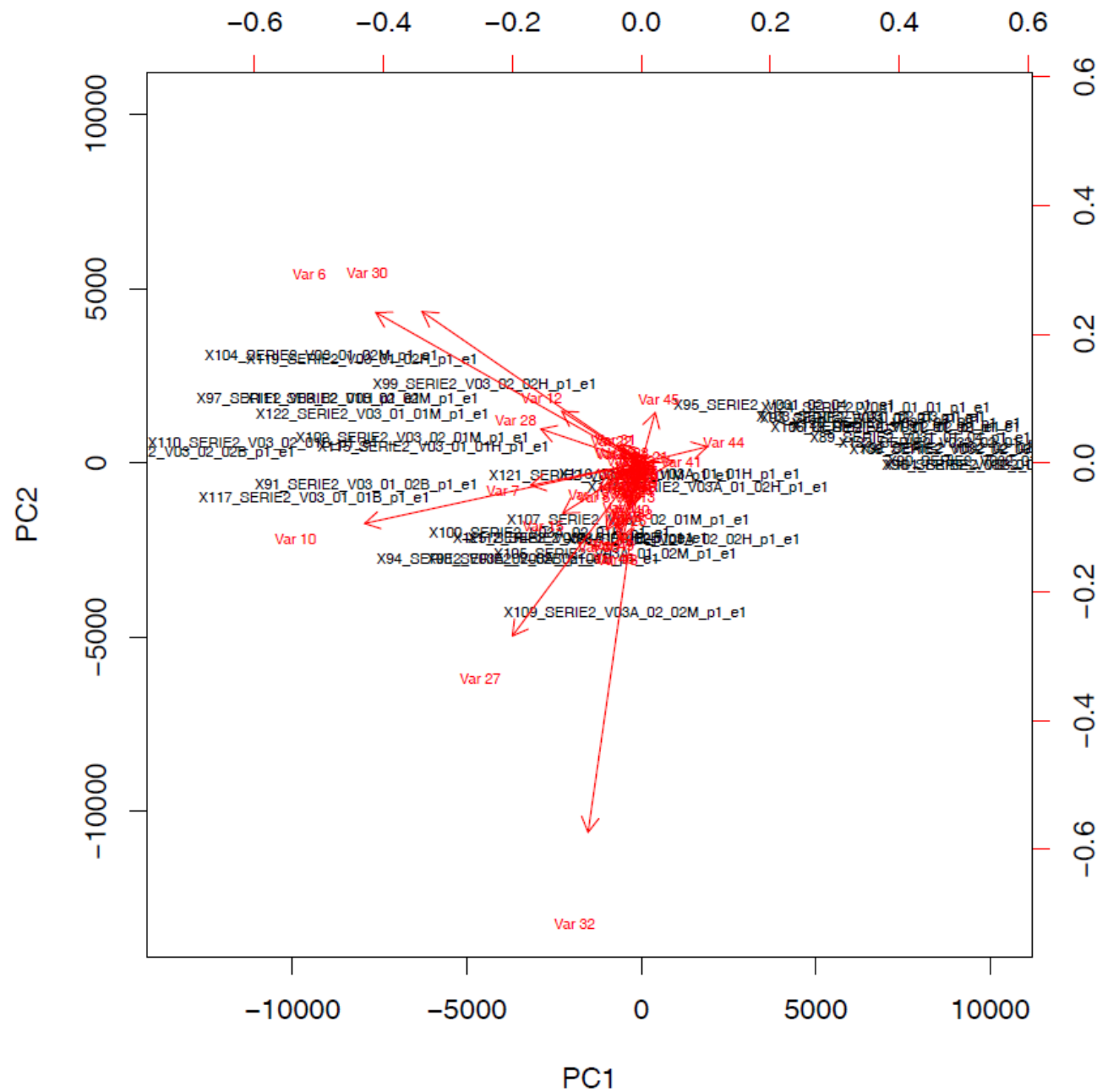


Feature 268.1/234



	name	fold	tstat	pvalue	mzmed	rtmed	npeaks	Attaque	Non attaque	Ech1	Ech2	Ech3
1	M268T234	3.42	4.83	2.04E-05	268.10	233.95	20	1	19	152.63	160.65	549.13
2	M532T1458	1.91	-3.98	0.00111045	532.16	1457.87	24	12	12	452.10	195.04	481.12
3	M284T1624	1.97	-4.02	0.00111956	284.21	1624.32	37	12	24	1261.31	841.17	1460.25
4	M641T2099	2.49	-3.85	0.00187246	641.19	2098.51	45	12	33	2665.65	736.15	3005.83
5	M639T2099	2.56	-3.83	0.00192063	639.18	2098.62	45	12	33	31828.55	7503.64	34338.54
6	M640T2099	2.56	-3.80	0.00205874	640.19	2098.52	45	12	33	10205.49	2502.55	11341.15

# ACP à partir de la matrice des pics





<http://metlin.scripps.edu/>

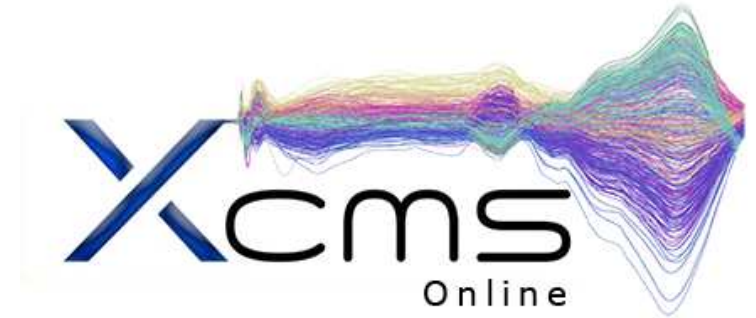
Scripps Center for Metabolomics  
and Mass Spectrometry

Total: 11 Metabolites

METLIN ID	MASS	Appm	NAME	MS/MS	STRUCTURE
67029	[M+H] <sup>+</sup> m/z 268.1027 M 267.0954	10	Miserotoxin Formula: C <sub>9</sub> H <sub>17</sub> NO <sub>8</sub> CAS: 24502-76-9	NO	
3394	[M+H] <sup>+</sup> m/z 268.1027 M 267.0954	10	Neuraminic acid Formula: C <sub>9</sub> H <sub>17</sub> NO <sub>8</sub> CAS: 114-04-5	NO	
64632	[M+H] <sup>+</sup> m/z 268.0968 M 267.0895	11	PD 98059 Formula: C <sub>16</sub> H <sub>13</sub> NO <sub>3</sub> CAS: 167869-21-8	<a href="#">View</a>	
85746	[M+Na] <sup>+</sup> m/z 268.1030 M 245.1137	11	Glutamyl-Valine Formula: C <sub>10</sub> H <sub>17</sub> N <sub>2</sub> O <sub>5</sub> CAS:	NO	
86018	[M+Na] <sup>+</sup> m/z 268.1030 M 245.1137	11	Valyl-Glutamate Formula: C <sub>10</sub> H <sub>17</sub> N <sub>2</sub> O <sub>5</sub> CAS:	NO	
4149	[M+H] <sup>+</sup> m/z 268.1040 M 267.0968	15	3-Deoxyguanosine Formula: C <sub>10</sub> H <sub>13</sub> N <sub>5</sub> O <sub>4</sub> CAS: 3608-58-0	<a href="#">View</a>	
3066	[M+H] <sup>+</sup> m/z 268.1040 M 267.0968	15	Zidovudine Formula: C <sub>10</sub> H <sub>13</sub> N <sub>5</sub> O <sub>4</sub> CAS: 30516-87-1	<a href="#">View</a>	

**xcms2** : analyse de spectres de fragmentation (MS/MS) → interrogation de bases de données.

**metaxcms** : comparaison de plus de deux classes d'individus.



## Autres packages pour l'analyse métabolomique

**CAMERA et Astream** : annotation de données métabolomiques LC-MS.

[http://link.springer.com/content/pdf/10.1007%2F978-3-540-71233-6\\_29.pdf](http://link.springer.com/content/pdf/10.1007%2F978-3-540-71233-6_29.pdf)

Alonso et al, Bioinformatics 27:1339-40 (2011)

**Metabonomic** : traitement des données + analyses multivariées (PCS, LDA, PLS...).

Izquierdo-García et al, BMC Bioinformatics 10: 363 (2009)

**MetabolAnalyze** : analyses multivariées (PPCA, PPCCA, MPPCA).

Nyamundanda, G., Gormley, I.C. and Brennan, L. Probabilistic principal components analysis for metabolomic data. Technical Report. University College Dublin (2010)

**Metabolomics** : fonctions graphiques : dendrogrammes, matrices de contrastes...

<http://cran.r-project.org/web/packages/metabolomics/metabolomics.pdf> (avril 2013)

**Metab** : pré-traitement de données GC-MS.

Agio et al, Bioinformatics 27:2316-8 (2011)



Metlin metabolite Database (Données MS)	<a href="http://metlin.scripps.edu/">http://metlin.scripps.edu/</a>
ChemSpider (Données MS)	<a href="http://www.chemspider.com">http://www.chemspider.com</a>
Human Metabolome Database	<a href="http://www.hmdb.ca/">http://www.hmdb.ca/</a>
KNAPSAcK (Métabolites de plantes)	<a href="http://prime.psc.riken.jp/?action=metabolites_index">http://prime.psc.riken.jp/?action=metabolites_index</a>

BioMeta Database (Voies métaboliques)

<http://biometa.cmbi.ru.nl/>

BiGG (Biochemically, Genetically and Genomically metabolic network reconstructions)

<http://bigg.ucsd.edu/main.pl>

Biocyc (Collection de voies métaboliques/génomiques)

<http://biocyc.org/>

Ecocyc (Gènes et Métabolisme de *Escherichia coli* K-12)

<http://www.ecocyc.org/>

Metacyc (Voies métaboliques)

<http://www.metacyc.org/>



# Profils LC-MS pour la métabolomique

