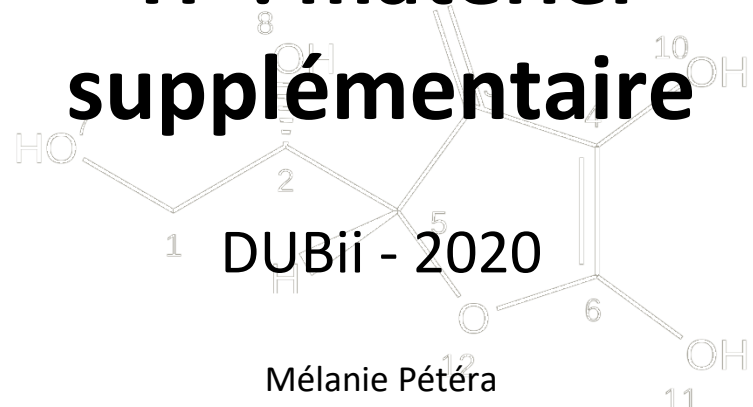


# 4 Wm

Workflow4metabolomics

## TP : matériel supplémentaire



04/06/2020

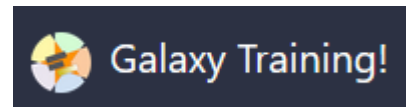
v 1.0.0

Galaxy Training Material :  
*Mass spectrometry: LC-MS analysis*

# TP : EXEMPLE DE TRAITEMENT DE DONNÉES LC-MS

# Principe

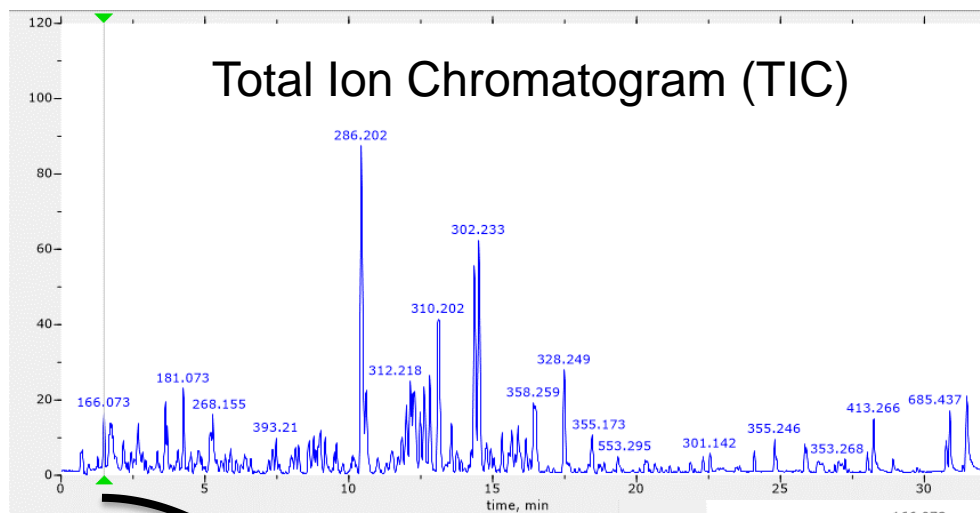
- Appréhender les principales étapes qui constituent un workflow d'analyse de données de métabolomique non-ciblée
- Cas d'étude : échantillons d'urine humaine analysés par LC-MS
- Support de TP : Galaxy Training Material disponible sur le site du GTN (Galaxy Training Network) :
  - <https://galaxyproject.github.io/training-material/>



Première partie

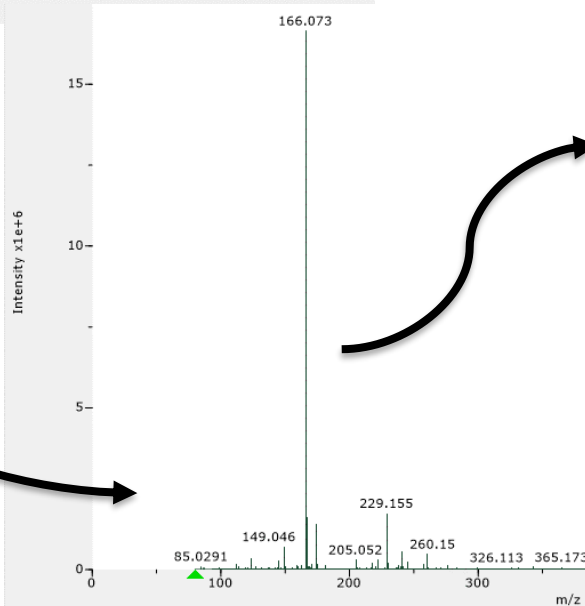
# PREPROCESSING : XCMS

# Ce qu'on veut faire



```

fichier mzXML
<scan num="263"
  scanType="Full"
  centroided="1"
  msLevel="1"
  peaksCount="10453"
  polarity="+"
  retentionTime="PT215.853S"
  basePeakMz="180.089111"
  basePeakIntensity="1.2813312e07"
  totIonCurrent="7.1073584e07"
  msInstrumentID="1">
<peaks compressionType="none"
  compressedLen="0"
  precision="64"
  byteOrder="network"
</scan>
  
```

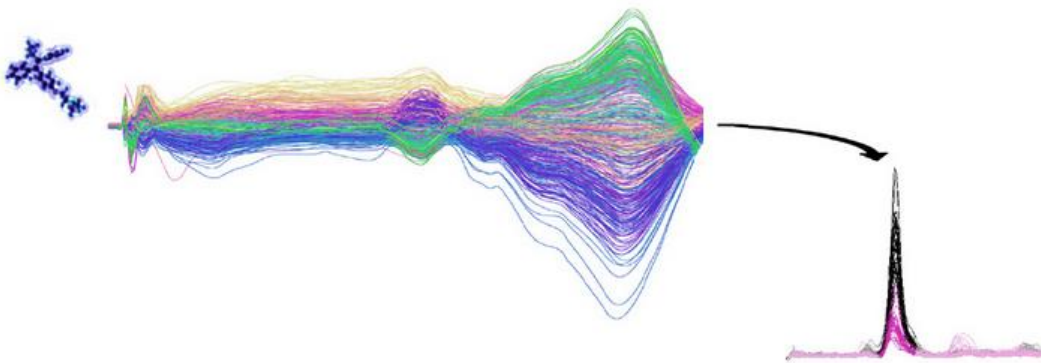


## Data matrix

ions	RetTime	Mass	T38CT05N	T38CT05N
114.067T1.5	1.5	114.067	9206.7362	4014.3652
137.072T1.5	1.5	137.072	2083.1412	3437.6839
212.853T1.5	1.5	212.853	0	2095.7974
196.88T1.5	1.5	196.88	0	1531.1653
162.114T1.5	1.5	162.114	1985.5564	267.3418
201.937T1.5	1.5	201.937	1934.2631	2295.2461
141.067T1.5	1.5	141.067	1656.8438	1182.8188
229.119T1.5	1.5	229.119	676.5843	688.6075
152.026T1.5	1.5	152.026	1002.5317	372.6582
407.186T6.1	6.1	407.186	183.2912	588.2105
359.059T6.1	6.1	359.059	36.4557	0
211.11T6.1	6.1	211.11	117.1308	175.5949

# Une solution libre et gratuite : XCMS

- ✓ R based software,
- ✓ Free
- ✓ A lot of parameters to tune,
- ✓ No graphical interface
- ✓ Need to write a R script



Web for documentation:

<https://bioconductor.org/packages/release/bioc/html/xcms.html>

Forums :

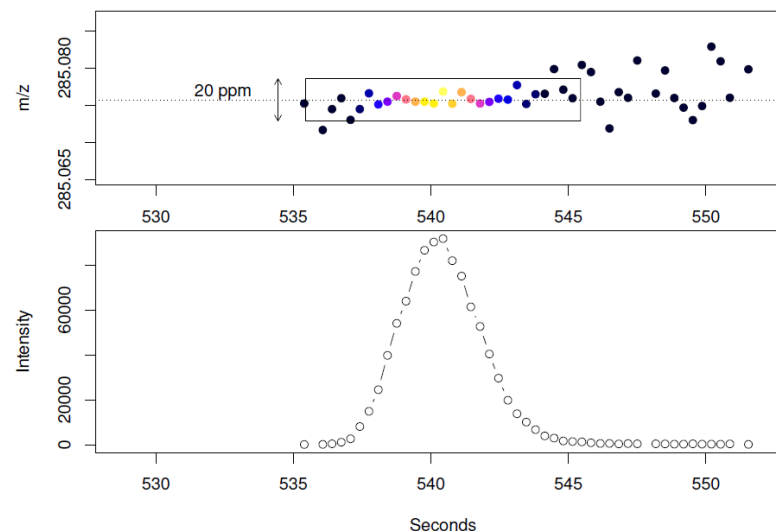
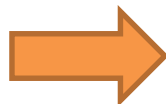
<http://metabolomics-forum.com>

# Première étape : extraire les pics

pour un échantillon

Données scan par scan :

```
fichier mzXML
<scan num="263"
  scanType="Full"
  centroided="1"
  msLevel="1"
  peaksCount="10453"
  polarity="+"
  retentionTime="PT215.853S"
  basePeakMz="180.089111"
  basePeakIntensity="1.2813312e07"
  totIonCurrent="7.1073584e07"
  msInstrumentID="1">
  <peaks compressionType="none"
    compressedLen="0"
    precision="64"
    byteOrder="network"
  </scan>
```



**Figure 1**  
Mass trace and chromatographic peak of Biochanin A  $[M + H]^+$  mass signal. The upper panel shows the mass of the biochanin A  $[M + H]^+$  mass signal across 10 seconds with colour-coded intensities. The corresponding chromatographic peak is shown below.

Tautenhahn R. *BMC Bioinformatics* 2008

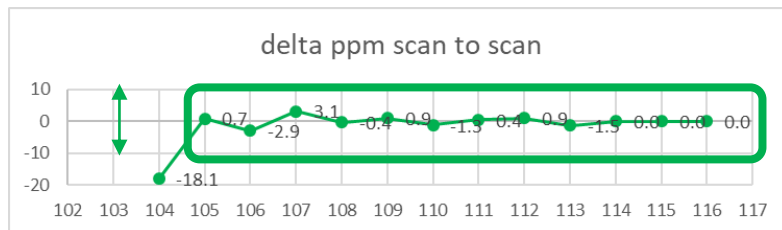
Où commence le pic ? Où s'arrête-t-il ? Qu'en est-il du bruit ?  
Si on considère qu'il s'agit d'un pic, comment synthétiser les informations suivantes ?

- m/z
- temps de rétention (RT)
- intensité

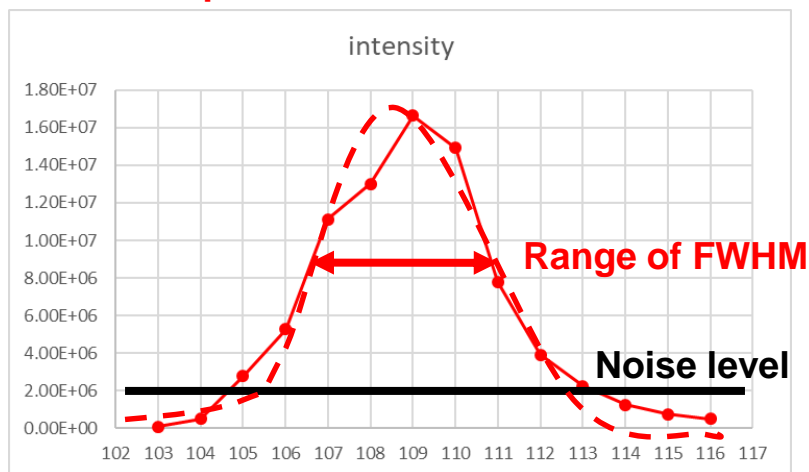
# Exemple de l'algorithme CentWave

Détecter et délimiter des régions d'intérêt (ROI : « region of interest »)

Exemple du paramètre « ppm »



Exemple des paramètres  
« peakwidth » et « noise »



Quelle valeur d'intensité ?

- Hauteur du pic ?
- Intégration du pic (aire sous la courbe) ?
- Doit-on soustraire le bruit ?



# Deuxième étape : grouper les pics

Ce qu'on a au départ	Listes de pics par échantillons (listes indépendantes)	<table><tr><th colspan="3">pool1B1</th><th colspan="3">pool1B2</th><th colspan="3">pool1B3</th></tr><tr><th>mz</th><th>rt</th><th>int</th><th>mz</th><th>rt</th><th>int</th><th>mz</th><th>rt</th><th>int</th></tr><tr><td>196.0905</td><td>66.6</td><td>7810936</td><td>196.0910</td><td>66.7</td><td>11733921</td><td>196.0902</td><td>66.6</td><td>7933325</td></tr><tr><td>158.1180</td><td>67.4</td><td>71736</td><td>342.0310</td><td>69.0</td><td>74594</td><td>158.1173</td><td>67.4</td><td>82969</td></tr><tr><td>342.0308</td><td>67.6</td><td>202268</td><td>267.0581</td><td>65.5</td><td>260877</td><td>342.0308</td><td>21.3</td><td>2581</td></tr><tr><td>267.0581</td><td>65.5</td><td>282039</td><td>283.0318</td><td>65.2</td><td>424631</td><td>283.0320</td><td>65.3</td><td>357448</td></tr></table>	pool1B1			pool1B2			pool1B3			mz	rt	int	mz	rt	int	mz	rt	int	196.0905	66.6	7810936	196.0910	66.7	11733921	196.0902	66.6	7933325	158.1180	67.4	71736	342.0310	69.0	74594	158.1173	67.4	82969	342.0308	67.6	202268	267.0581	65.5	260877	342.0308	21.3	2581	267.0581	65.5	282039	283.0318	65.2	424631	283.0320	65.3	357448									
pool1B1			pool1B2			pool1B3																																																											
mz	rt	int	mz	rt	int	mz	rt	int																																																									
196.0905	66.6	7810936	196.0910	66.7	11733921	196.0902	66.6	7933325																																																									
158.1180	67.4	71736	342.0310	69.0	74594	158.1173	67.4	82969																																																									
342.0308	67.6	202268	267.0581	65.5	260877	342.0308	21.3	2581																																																									
267.0581	65.5	282039	283.0318	65.2	424631	283.0320	65.3	357448																																																									
Ce qu'on fait pour grouper les pics des différents échantillons	Grouper les pics par m/z	<table><tr><th>mz</th><th>rt</th><th>int</th><th>mz</th><th>rt</th><th>int</th><th>mz</th><th>rt</th><th>int</th></tr><tr><td>196.0905</td><td>66.6</td><td>7810936</td><td>196.0910</td><td>66.7</td><td>11733921</td><td>196.0902</td><td>66.6</td><td>7933325</td></tr><tr><td>158.1180</td><td>67.4</td><td>71736</td><td>342.0310</td><td>69.0</td><td>74594</td><td>158.1173</td><td>67.4</td><td>82969</td></tr><tr><td>342.0308</td><td>67.6</td><td>202268</td><td>267.0581</td><td>65.5</td><td>260877</td><td>342.0308</td><td>21.3</td><td>2581</td></tr><tr><td>267.0581</td><td>65.5</td><td>282039</td><td>283.0318</td><td>65.2</td><td>424631</td><td>283.0320</td><td>65.3</td><td>357448</td></tr></table>	mz	rt	int	mz	rt	int	mz	rt	int	196.0905	66.6	7810936	196.0910	66.7	11733921	196.0902	66.6	7933325	158.1180	67.4	71736	342.0310	69.0	74594	158.1173	67.4	82969	342.0308	67.6	202268	267.0581	65.5	260877	342.0308	21.3	2581	267.0581	65.5	282039	283.0318	65.2	424631	283.0320	65.3	357448																		
	mz	rt	int	mz	rt	int	mz	rt	int																																																								
196.0905	66.6	7810936	196.0910	66.7	11733921	196.0902	66.6	7933325																																																									
158.1180	67.4	71736	342.0310	69.0	74594	158.1173	67.4	82969																																																									
342.0308	67.6	202268	267.0581	65.5	260877	342.0308	21.3	2581																																																									
267.0581	65.5	282039	283.0318	65.2	424631	283.0320	65.3	357448																																																									
	Pour chaque groupe de m/z, séparer par RT	<table><tr><th>mz</th><th>rt</th><th>int</th><th>mz</th><th>rt</th><th>int</th><th>mz</th><th>rt</th><th>int</th></tr><tr><td>196.0905</td><td>66.6</td><td>7810936</td><td>196.0910</td><td>66.7</td><td>11733921</td><td>196.0902</td><td>66.6</td><td>7933325</td></tr><tr><td>158.1180</td><td>67.4</td><td>71736</td><td></td><td></td><td></td><td>158.1173</td><td>67.4</td><td>82969</td></tr><tr><td></td><td></td><td></td><td></td><td></td><td></td><td>342.0308</td><td>21.3</td><td>2581</td></tr><tr><td>342.0308</td><td>67.6</td><td>202268</td><td>342.0310</td><td>69.0</td><td>74594</td><td></td><td></td><td></td></tr><tr><td>267.0581</td><td>65.5</td><td>282039</td><td>267.0581</td><td>65.5</td><td>260877</td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td>283.0318</td><td>65.2</td><td>424631</td><td>283.0320</td><td>65.3</td><td>357448</td></tr></table>	mz	rt	int	mz	rt	int	mz	rt	int	196.0905	66.6	7810936	196.0910	66.7	11733921	196.0902	66.6	7933325	158.1180	67.4	71736				158.1173	67.4	82969							342.0308	21.3	2581	342.0308	67.6	202268	342.0310	69.0	74594				267.0581	65.5	282039	267.0581	65.5	260877							283.0318	65.2	424631	283.0320	65.3	357448
mz	rt	int	mz	rt	int	mz	rt	int																																																									
196.0905	66.6	7810936	196.0910	66.7	11733921	196.0902	66.6	7933325																																																									
158.1180	67.4	71736				158.1173	67.4	82969																																																									
						342.0308	21.3	2581																																																									
342.0308	67.6	202268	342.0310	69.0	74594																																																												
267.0581	65.5	282039	267.0581	65.5	260877																																																												
			283.0318	65.2	424631	283.0320	65.3	357448																																																									
Ce qu'on génère en fin d'étape	Attribuer un m/z et un RT de référence	<table><tr><th>mz</th><th>rt</th><th>pool1B1</th><th>pool1B2</th><th>pool1B3</th></tr><tr><td>196.0905</td><td>66.6</td><td>7810936</td><td>11733921</td><td>7933325</td></tr><tr><td>158.1176</td><td>67.4</td><td>71736</td><td></td><td>82969</td></tr><tr><td>342.0308</td><td>21.3</td><td></td><td></td><td>2581</td></tr><tr><td>342.0309</td><td>68.3</td><td>202268</td><td>74594</td><td></td></tr><tr><td>267.0581</td><td>65.5</td><td>282039</td><td>260877</td><td></td></tr><tr><td>283.0319</td><td>65.2</td><td></td><td>424631</td><td>357448</td></tr></table>	mz	rt	pool1B1	pool1B2	pool1B3	196.0905	66.6	7810936	11733921	7933325	158.1176	67.4	71736		82969	342.0308	21.3			2581	342.0309	68.3	202268	74594		267.0581	65.5	282039	260877		283.0319	65.2		424631	357448																												
mz	rt	pool1B1	pool1B2	pool1B3																																																													
196.0905	66.6	7810936	11733921	7933325																																																													
158.1176	67.4	71736		82969																																																													
342.0308	21.3			2581																																																													
342.0309	68.3	202268	74594																																																														
267.0581	65.5	282039	260877																																																														
283.0319	65.2		424631	357448																																																													

Delta de m/z ? De RT ? Garde-t-on tous les pics même s'ils sont peu présents ? Etc.

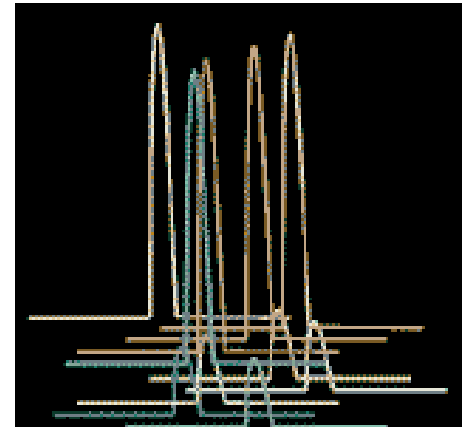
# Etape facultative : aligner les RT des échantillons

Déviation de temps de rétention d'un échantillon à l'autre

=> difficulté à regrouper les ions entre eux, en particulier de façon automatisée

Exemple de stratégie : « **peakgroups** »

- Identifier des « beaux » pics, qui vont servir de référence (« well-behaved peaks »)
- Aligner les RT en fonction des déviations observées sur ces beaux pics



Quelles caractéristiques pour définir un « beau » pic ? Quelle méthode de régression pour corriger ?

*Une étape d'alignement de temps de rétention doit être suivie, de nouveau, par une étape de groupement de pics.*

# Dernière étape : combler les NA

mz	rt	pool1B1	pool1B2	pool1B3
196.0905	66.6	7810936	117921	7933325
158.1176	61.3	71736		82969
342.0308	21.3			2581
342.0309	68.3	202268	74594	
267.0581	65.5	282039	260877	
283.0319	65.2		424631	35747

Plusieurs raisons possibles à l'absence de valeur

- ⇒ Pas de composé à l'origine dans l'échantillon
- ⇒ Incapacité à détecter correctement le pic
- ⇒ Pic non retenu lors de la première étape (trop faible intensité, forme du pic mauvaise...)

**Idée** : aller récupérer dans la donnée brute l'information contenue à l'emplacement du pic manquant

# TP Time !

Galaxy Training Material utilisé :

- <https://galaxyproject.github.io/training-material/topics/metabolomics/tutorials/lcms/tutorial.html>

**TODO : Parties 1.3 à 1.9, puis 2. Ne pas faire les parties optionnelles.**

Des problèmes lors des exercices de pré-requis : vous n'avez pas pu lancer avec succès l'étape 1.2 ?

Pas de panique ! Vous pouvez importer l'historique au lien qui suit qui contient les premières étapes qu'il vous manque :

<https://workflow4metabolomics.usegalaxy.fr/u/m.petera/h/dubiibackup1>

Deuxième partie

# **DATA PROCESSING : VÉRIFIER, CORRIGER, FILTRER**

# Vérifier les données

Contraintes d'un jeu métabolomique parfois telles qu'on aimerait « **vérifier** » si globalement les données sont fortement **impactées par certains aspects** ou pas.

Étape intéressante **en amont** des analyses statistiques pour **avoir une idée** de ce à quoi on s'attaque.



- Calcul d'indicateurs (e.g. % de NA)
- Visualisation des données (e.g. Analyse en Composantes principales)



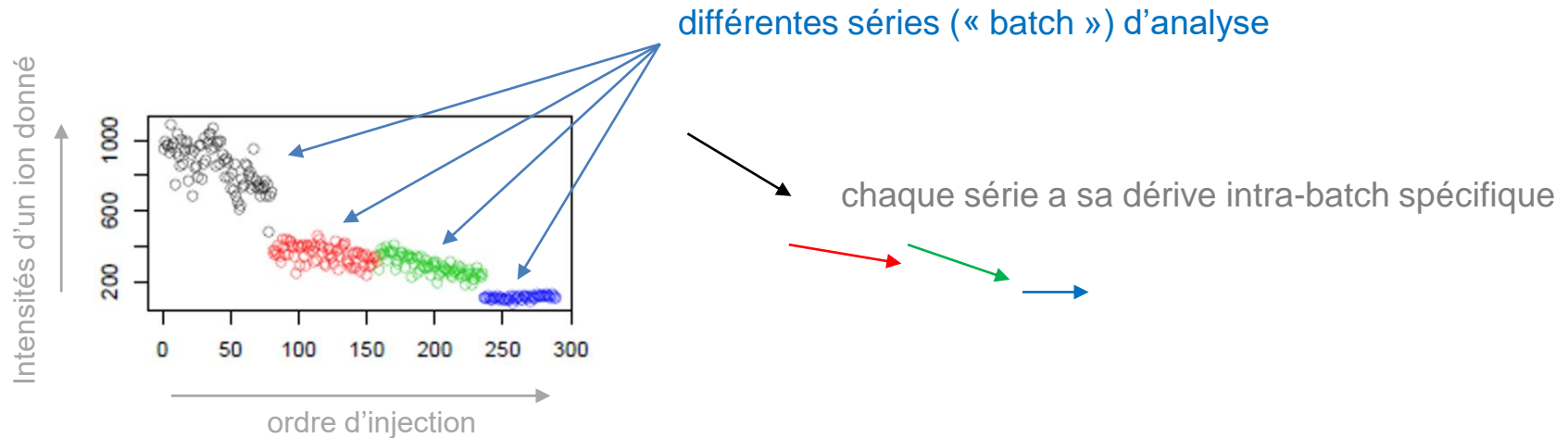
- Détection d'outliers
- Détection d'effets analytiques

Cette étape de vérification débouche communément sur des étapes de correction ou de filtre des données

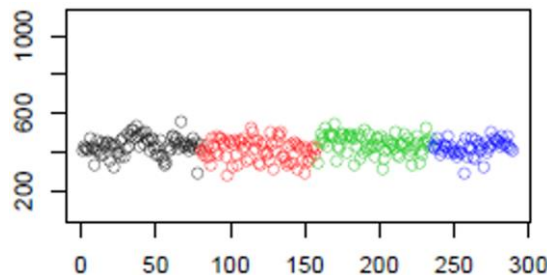
# Corriger les données : effets analytiques

## LC-MS : dérive analytique et effet batch

Variation de la mesure d'un signal du fait de l'encrassement de la machine



Ce qu'il nous faut pour être en mesure de réaliser des analyses statistiques



Intensités comparables

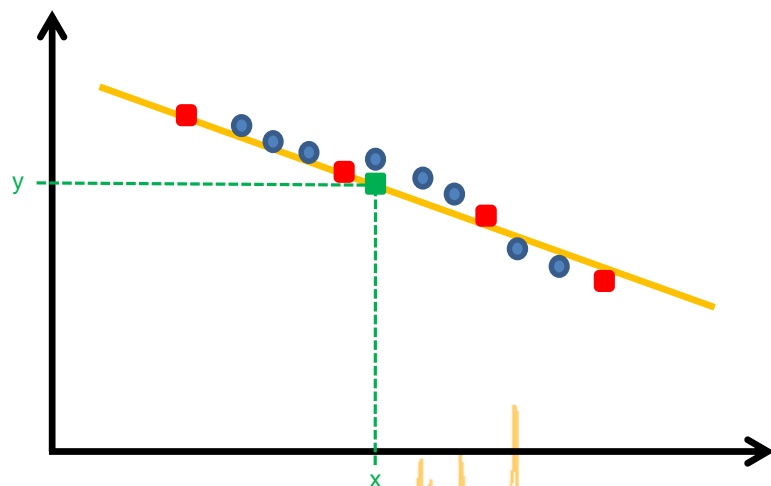
# Corriger les données : effets analytiques

## Exemple de méthode de correction de dérive analytique et d'effet batch

### Méthode popularisée par Van Der Kloet

- La correction est effectuée pour chaque ion indépendamment
- Pour chaque ion :
  - Une correction intra-batch est faite pour chaque batch indépendamment
    - La dérive analytique est modélisée en utilisant des pools et leur ordre d'injection au sein de la série
    - Chaque intensité d'échantillon est divisée par l'estimation de la dérive analytique correspondant au numéro d'injection de l'échantillon
    - Les valeurs des échantillons sont ensuite multipliées par une valeur de référence (pour conserver l'échelle de valeur originale)
  - L'effet inter-batch est de fait corrigé

Pools = mélanges des échantillons de l'analyse, tous identiques, injectés à intervalles réguliers tout au long des séries



- Intensité du pool observée
- Intensité de l'échantillon observée
- Courbe de régression du modèle de dérive analytique
- Valeur estimée pour le numéro d'injection x

$$\text{valeur normalisée obtenue pour l'échantillon de l'injection numéro x} = \frac{\text{valeur observée pour l'échantillon de l'injection numéro x}}{\text{valeur estimée de l'injection numéro x}} \times \text{valeur de référence}$$



# Filtrer les données

- Les données extraites contiennent souvent **plus de choses que ce qu'on souhaite exploiter**
  - Résidus de bruit
  - Ions non informatifs
  - Ions de maigre fiabilité
  - ...
- La table de données extraite possède des **caractéristiques pénalisantes** pour certains aspects des statistiques qui vont suivre
  - Différents types de redondance dans les données
  - Nombre important de variables par rapport au nombre de sujets
  - ...

➡ Il est profitable de filtrer les données lorsque c'est **possible** et **pertinent**

# Filtrer les données : exemple du CV

- Pourquoi ?

Certains ions peuvent être **trop bruités**, ce qui donne une variabilité artificielle trop forte.

- Comment ?

En exploitant les **pools injectés** et en calculant des **coefficients de variation (CV)** :

Où :

$$CV_i = \frac{\sigma_i}{\mu_i}$$

$CV_i$  = coefficient de variation de l'ion  $i$

$\sigma_i$  = écart – type de l'ion  $i$

$\mu_i$  = moyenne de l'ion  $i$

- En pratique :

Deux indicateurs intéressants :

- **CV des pools** : on s'attend à ce que les pools, qui sont biologiquement identiques, ne varient pas trop en intensité, par exemple qu'ils aient un  $CV < 0,3$  pour que l'ion soit exploitable
- **Rapport CV des pools sur CV des échantillons** : on s'attend à ce que les pools soient moins variables que les échantillons, on peut donc par exemple fixer un rapport maximum « CV pools / CV échantillons » de 1 pour que l'ion soit exploité.

# TP Time !

Rappel du Galaxy Training Material utilisé :

- <https://galaxyproject.github.io/training-material/topics/metabolomics/tutorials/lcms/tutorial.html>

TODO : Parties 3.1 à 3.3.

Des problèmes lors du TP précédent et vous n'avez pas eu le temps d'arriver jusqu'au bout ?

Pas de panique ! Vous pouvez importer l'historique au lien qui suit qui contient les étapes qu'il vous manque pour commencer le prochain TP :

<https://workflow4metabolomics.usegalaxy.fr/u/m.petera/h/dubiibackup2>

Troisième partie

# STATISTIQUES, ANNOTATION

# Statistiques – pourquoi ?

---

## ➤ Grande quantité d'ions récoltés

Trop d'ions pour envisager de tous les identifier

## ➤ Volonté de se focaliser uniquement sur des ions dits d'intérêt

Sélectionner les ions en lien avec une question de recherche donnée pour concentrer les efforts d'identification sur ces seuls ions

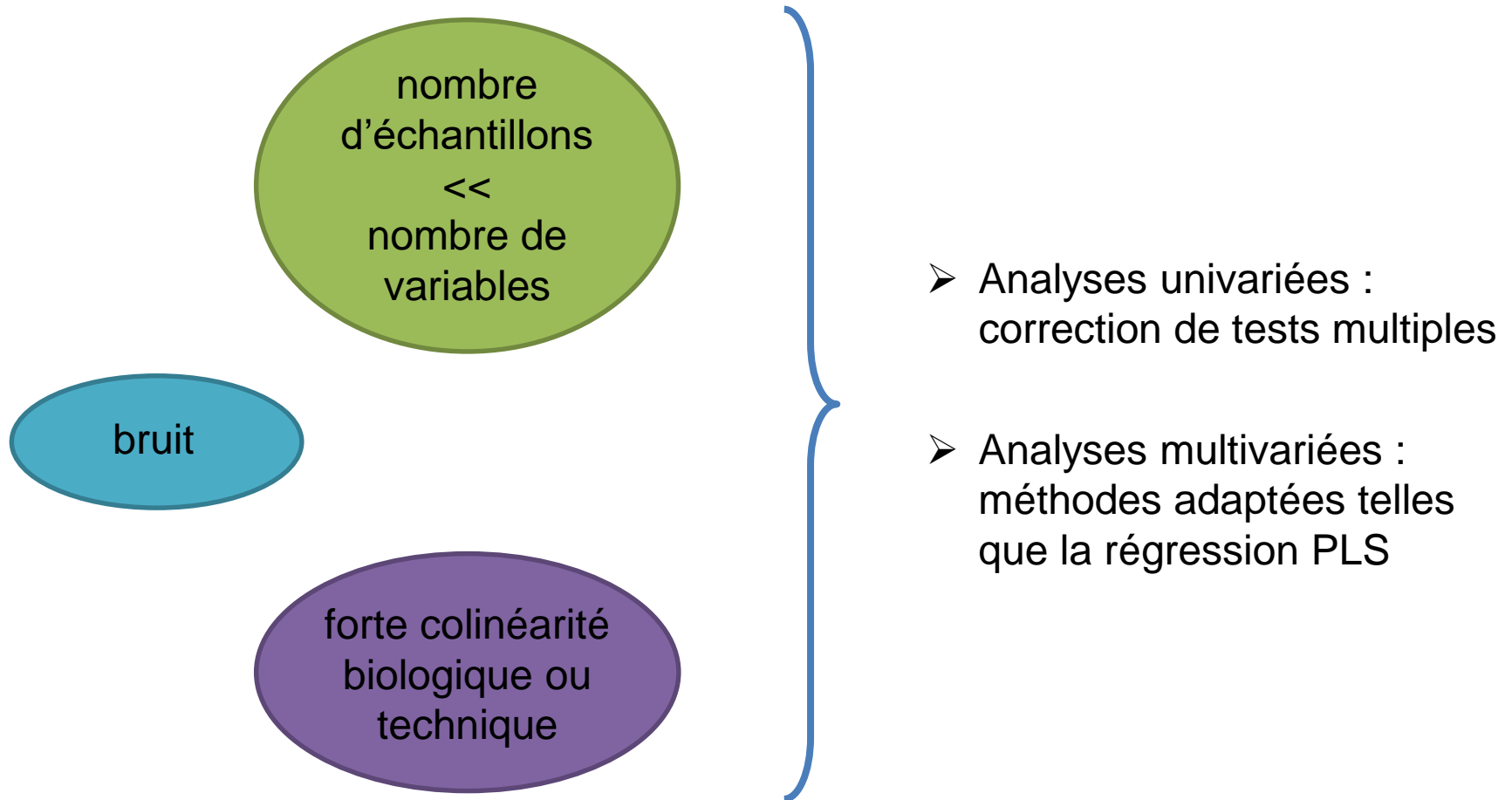
## ➤ Idée

Confronter chaque ion récolté à une ou plusieurs variables d'intérêt

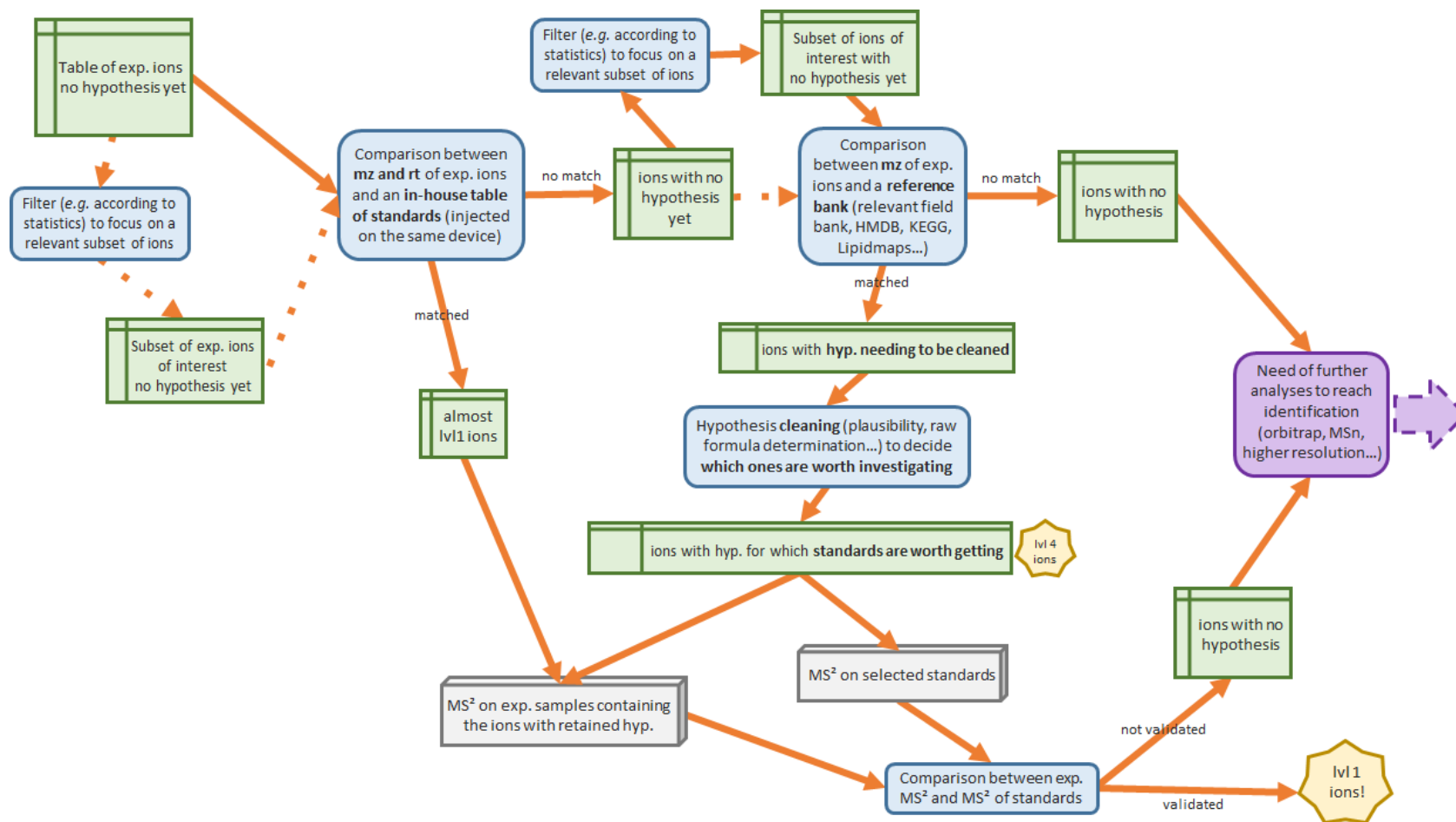
## ➤ Principe de base

Calcul d'**indicateurs** permettant une **sélection de variables** sur la base de **critères** définis en fonction des méthodes et des objectifs

# Statistiques – attentions particulières



# Annotation – un processus complexe



# TP Time !

Rappel du Galaxy Training Material utilisé :

- <https://galaxyproject.github.io/training-material/topics/metabolomics/tutorials/lcms/tutorial.html>

TODO : Parties 4 et 5.

Des problèmes lors du TP précédent et vous n'avez pas eu le temps d'arriver jusqu'au bout ?

Pas de panique ! Vous pouvez importer l'historique au lien qui suit qui contient les étapes qu'il vous manque pour commencer le prochain TP :

<https://workflow4metabolomics.usegalaxy.fr/u/m.petera/h/dubiibackup3>



# Pour aller plus loin



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

Yann Guitton<sup>a,1</sup>, Marie Tremblay-Franco<sup>b,1</sup>, Gildas Le Corguillé<sup>c</sup>, Jean-François Martin<sup>b</sup>,  
Mélanie Pétéra<sup>d</sup>, Pierrick Roger-Mele<sup>e</sup>, Alexis Delabrière<sup>e</sup>, Sophie Goulitquer<sup>f</sup>, Mishari Monsoor<sup>g</sup>,  
Christophe Duperrier<sup>h</sup>, Cécile Canlet<sup>b</sup>, Rémi Servien<sup>b</sup>, Patrick Tardivel<sup>b</sup>, Christophe Caron<sup>g</sup>,  
Franck Giacomoni<sup>h,\*,2</sup>, Etienne A. Thévenot<sup>h,\*,2</sup>

<sup>a</sup> LUNAM Université, Oniris, Laboratoire d'Etude des Résidus et Contaminants dans les Aliments (LABERCA), Nantes, F-44307, France  
<sup>b</sup> Toulouse (Research Centre in Food Toxicology), Université de Toulouse, INRA, ENVT, INP-Purpan, UPS, MetabolHUB, Toulouse, France  
<sup>c</sup> UPMC, CNRS, FRE2426, ARMS, Station Biologique, 29480, Roscoff, France  
<sup>d</sup> INRA, UMR 1019, PFEM, MetabolHUB, 63122, Saint Genes Champanelle, France  
<sup>e</sup> CRA, LIST, Laboratory for Data Analysis and Systems' Intelligence, MetabolHUB, F-91191 Gif-sur-Yvette, France  
<sup>f</sup> INSERM UMR1078-ECLA, IRISAM, Faculty of Medicine, University of Brest, 29200 Brest, France  
<sup>g</sup> INRA, Ingenuis, Toulouse, France

## ARTICLE INFO

Keywords:  
Metabolomics  
Data analysis  
E-infrastructure  
Workflow  
Galaxy  
Repository

## ABSTRACT

Metabolomics is a key approach in modern functional genomics and systems biology. Due to the complexity of metabolomics data, the variety of experimental designs, and the multiplicity of bioinformatics tools, providing experimenters with a simple and efficient resource to conduct comprehensive and rigorous analysis of their data is of utmost importance. In 2014, we launched the Workflow4Metabolomics (W4M; <http://workflow4metabolomics.org>) online infrastructure for metabolomics built on the Galaxy environment, which offers user-friendly features to build and run data analysis workflows including preprocessing, statistical analysis, and annotation steps. Here we present the new W4M 3.0 release, which contains twice as many tools as the first version, and provides two features which are, to our knowledge, unique among online resources. First, data from the four major metabolomics technologies (i.e., LC-MS, FIA-MS, GC-MS, and NMR) can be analyzed on a single platform. By using three studies in human physiology, alga evolution, and animal toxicology, we demonstrate how the 40 available tools can be easily combined to address biological issues. Second, the full analysis (including the workflow, the parameter values, the input data and output results) can be referenced with a permanent digital object identifier (DOI). Publication of data analyses is of major importance for robust and reproducible science. Furthermore, the publicly shared workflows are of high-value for e-learning and training. The Workflow4Metabolomics 3.0 e-infrastructure thus not only offers a unique online environment for analysis of data from the main metabolomics technologies, but it is also the first reference repository for metabolomics workflows.

Open Course :

<https://usemetabo.org/>

HowTo :

<https://workflow4metabolomics.org/howto>

