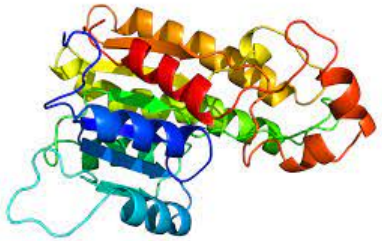
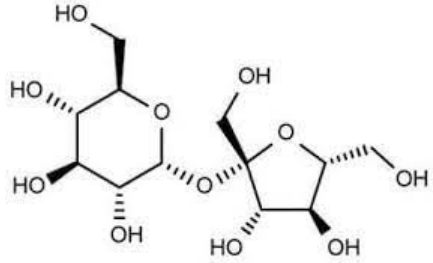
proteome - metabolome

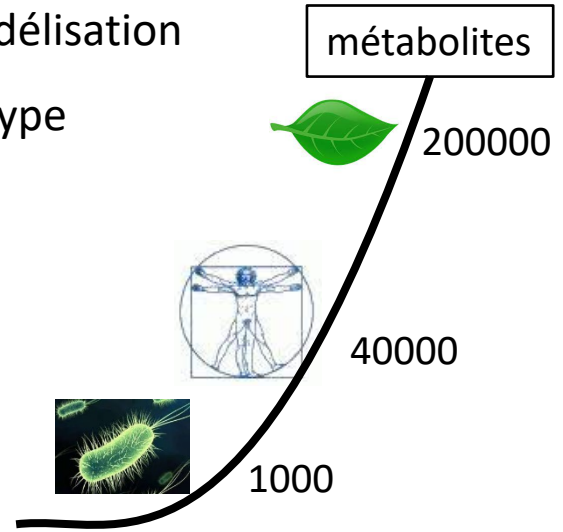
protéines



métabolites

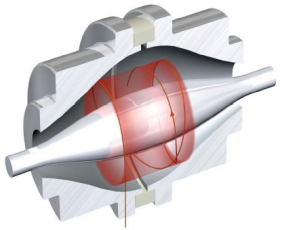
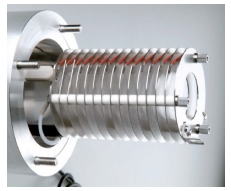


- ce sont les molécules actives des systèmes biologiques, les **acteurs moléculaires**
- niveaux d'informations nécessaires à l'intégration de données et à la modélisation
- pour étude des processus biologiques complexes au plus près du phénotype
- protéines : polymères complexes
- métabolites : petites molécules **< 1,5 kDa**
- **complexité** des échantillons
- **gamme dynamique** d'abondance :
  - metabolome : 7-9 ordres de grandeur (picomoles -> millimoles)
  - proteome : jusque 12 ordres (femtomoles -> millimoles)



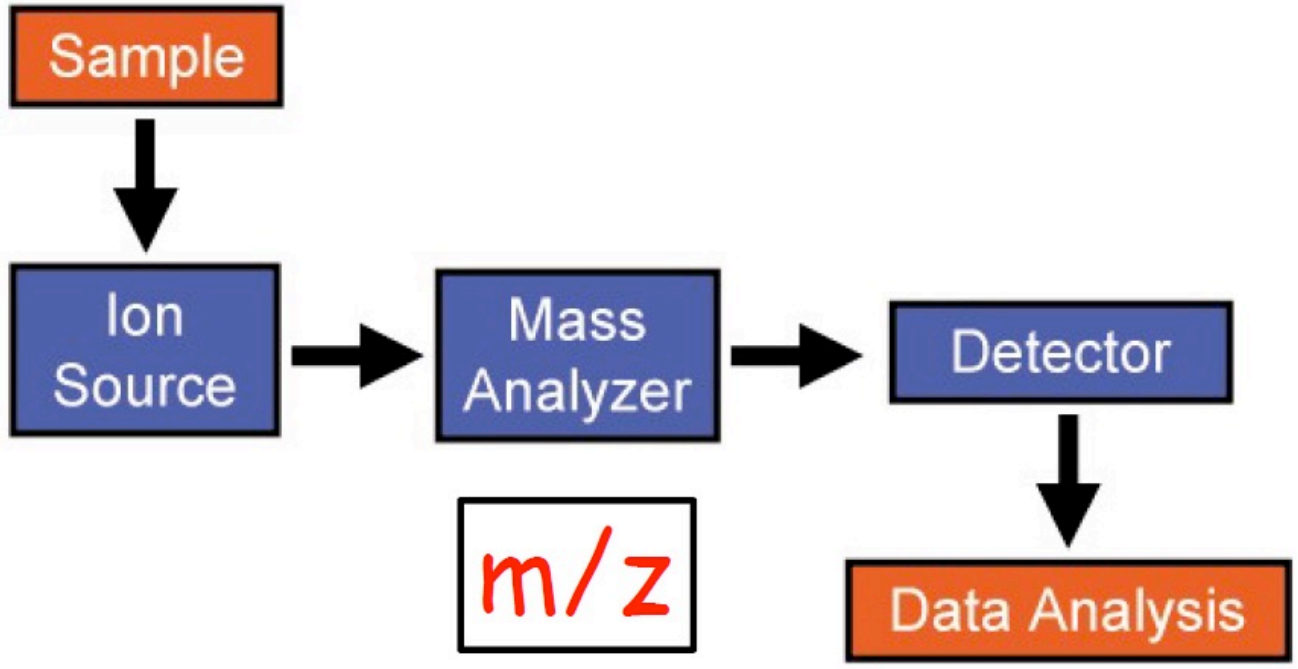
technologies

- approches courantes : globales, non ciblées, sans marquage
- MS : standard actuel
  - quadrupole
  - trappe d'ions
  - orbitrap
  - tof
- séparation préalable :
  - GC
  - LC
- combinaisons classiques :
  - LC + quadrupole
  - LC + orbitrap
  - GC + quadrupole

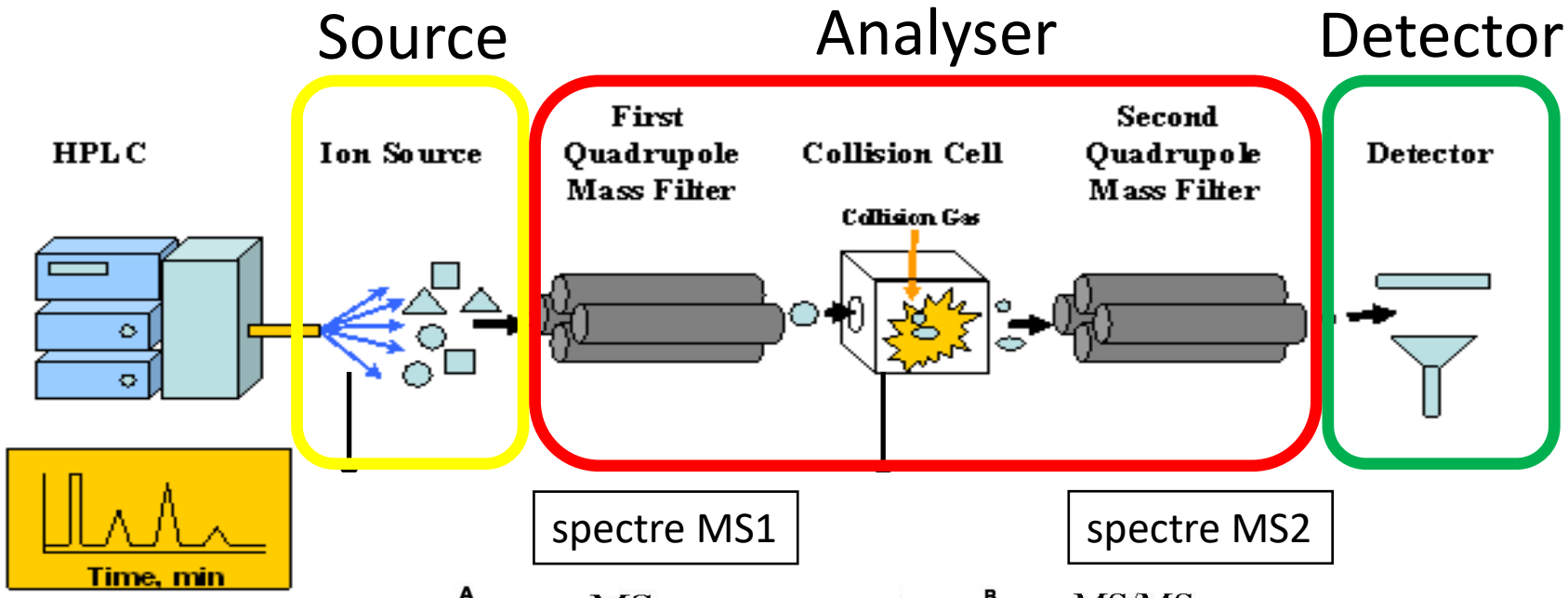





technologies

- structure **SAD** des spectromètres de masse : Source-Analyseur-Détecteur

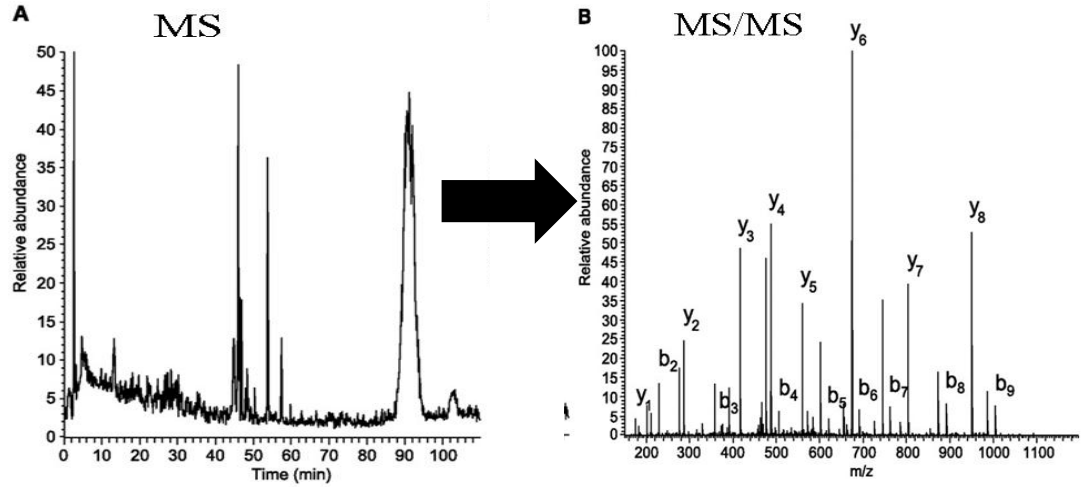


technologies



spectre MS1

spectre MS2



technologies

proteome

- plateforme PAPPSO (michel zivy)
- **LC-MS/MS** : Sciex nanoUPLC + Orbitrap Thermo Q Exactive
- génère des spectres MS1 et MS2

metabolome

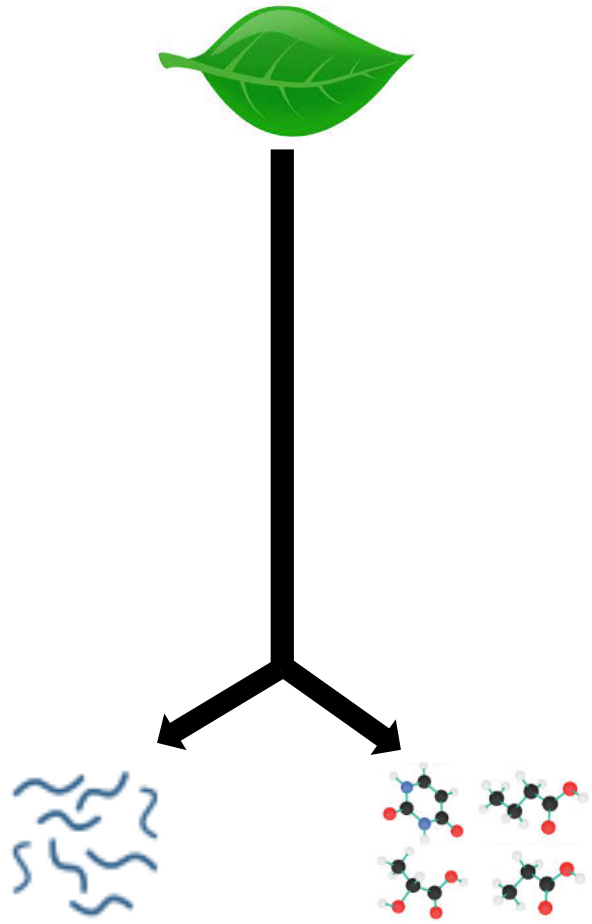
- plateforme OV-Chimie (gregory mouille)
- **GC-MS** : Agilent 7890 + Quadrupole Agilent 5975
- génère des données à 2 dimensions = indices de rétention + spectres

protocoles

proteome

LC-MS/MS

- matériel bio
- extraction
- reprise
- digestion
- injection peptides



metabolome

GC-MS

- matériel bio
- extraction
- standard interne
- dérivatisation
- injection composés

bioinfo

proteome

- identification : Mascot, Sequest, Xtandem
- identification + quantification : Progenesis QI, Proteome Discoverer, MaxQuant
- **pipeline PAPPSO** :
  - Xtandem pipeline + Masschroq + MasschroqR : java, c++, xml, R
  - approche bottom-up : des peptides vers les protéines

bioinfo

proteome

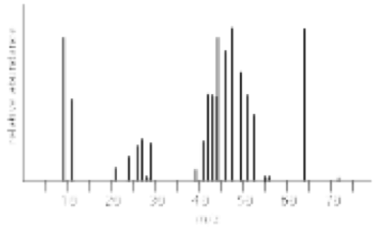
- **Xtandem pipeline** : identification des protéines
  - processus en 2 étapes
  - **mass matching** :
    - on compare les spectres expérimentaux à des spectres de référence générés in silico
    - On en déduit la séquence peptidique la plus probable associée au spectre observé
  - **assignation protéique**
    - étape la plus complexe, la plus cruciale, notamment en proteome quantitative
    - forme d'alignement
    - écarter les faux positifs

bioinfo

Mass Matching

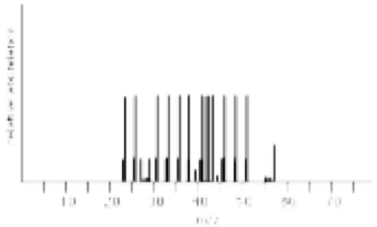
MS/MS

Experimental spectrum



compare

Theoretical spectrum



Protein sequence database

Output :
Ranked list of peptides matches

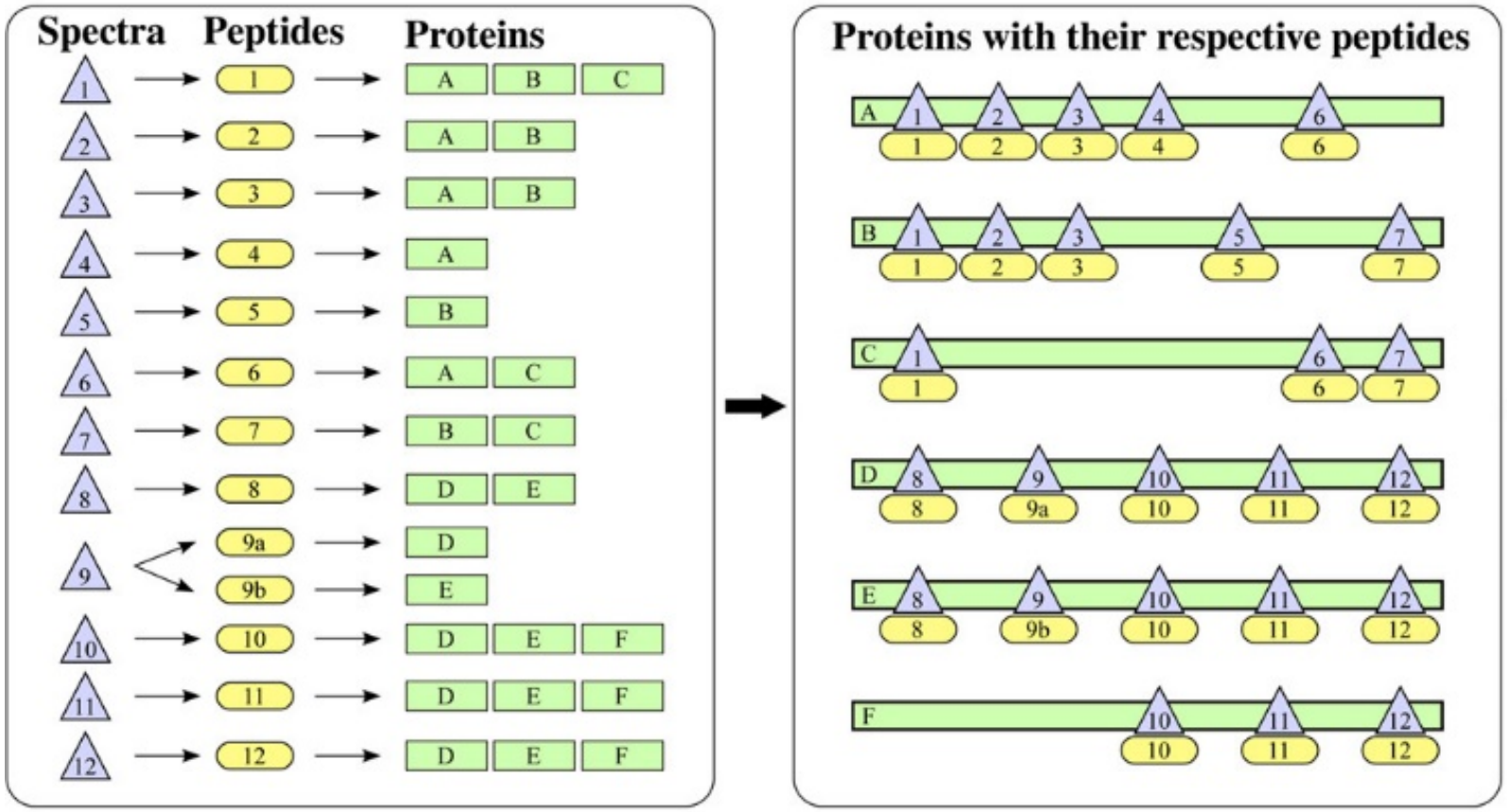
Peptide	score
KVVGTAWWLPQIP	3.5
ISLAVBCAQENFQE	2.1
AEKISIVVPENGKYY	1.8
SYLFGNTWCQIPVV	0.4

Best match	score
KVVGTAWWLPQIP	3.5

Peptide assignment to spectrum

bioinfo

Assignation protéique

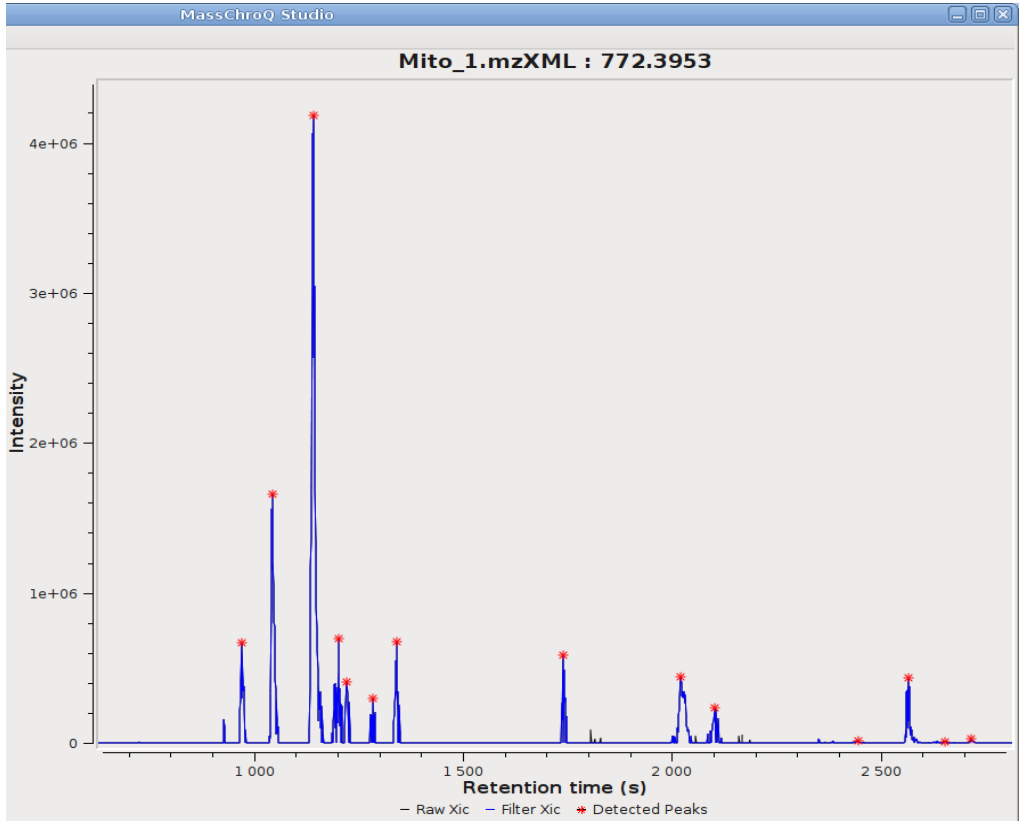


Langella et al., 2016

bioinfo

proteome

Masschroq : quantification des peptides
XIC : Extracted Ion Chromatogram



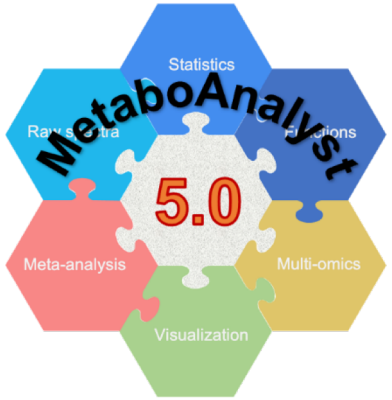
MasschroqR : quantification des protéines

- **mcq.drop.variable.rt**(XICRAW.SDS, cutoff=200)
- **mcq.drop.wide.peaks**(XIC.SDS, cutoff=200)
- **mcq.compute.normalization**(XIC.BY.TRACK, refSample="msruna1_P2", method="median")
- **mcq.compute.peptide.imputation**(XIC)
- **mcq.compute.protein.abundances**(XIC)

bioinfo

metabolome

- complexité de l'identification des métabolites :
  - grande diversité de structures chimiques
  - diversité accrue par des modifications chimiques
- **MetaboAnalyst** : axé LC-MS
- **Workflow4Metabolomics**
- fonctionnalités :
  - traitement des donnée de chromato et spectrales
  - analyses statistiques classiques et spécifiques
  - analyses fonctionnelles de pathway



bioinfo

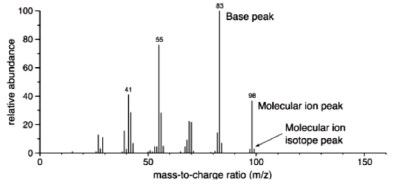
metabolome

- **Amdys** : identification
  - matching sur banques RI + spectres
  - banque ijpb de plus de 300 composés des plantes
  - banque NIST générale : 72361 composés

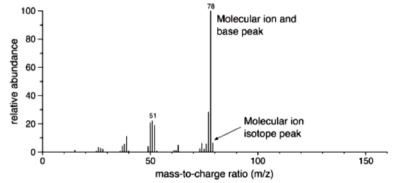
- **Quanlynx** : quantification
  - intégration des intensités MS
  - normalisation

Retention Index Mass Spectrum

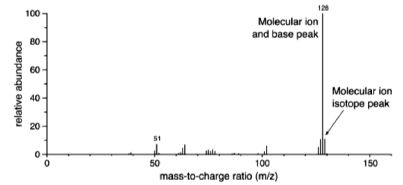
10 ⟷



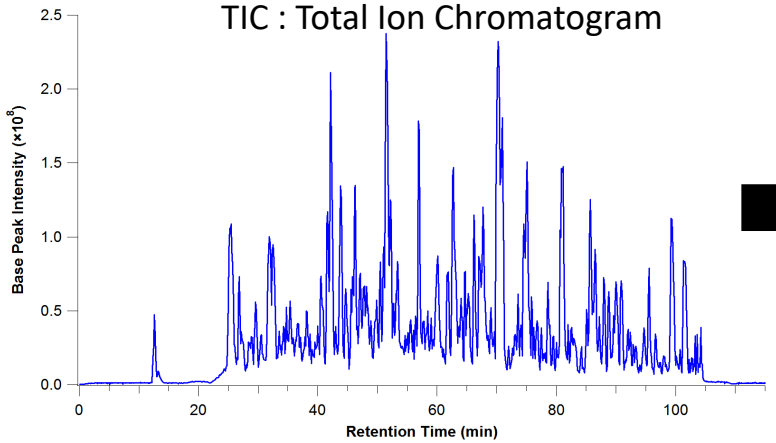
45 ⟷



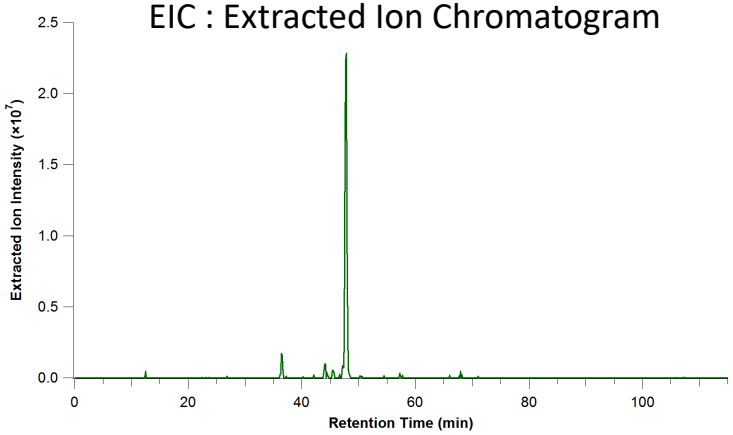
50 ⟷



TIC : Total Ion Chromatogram

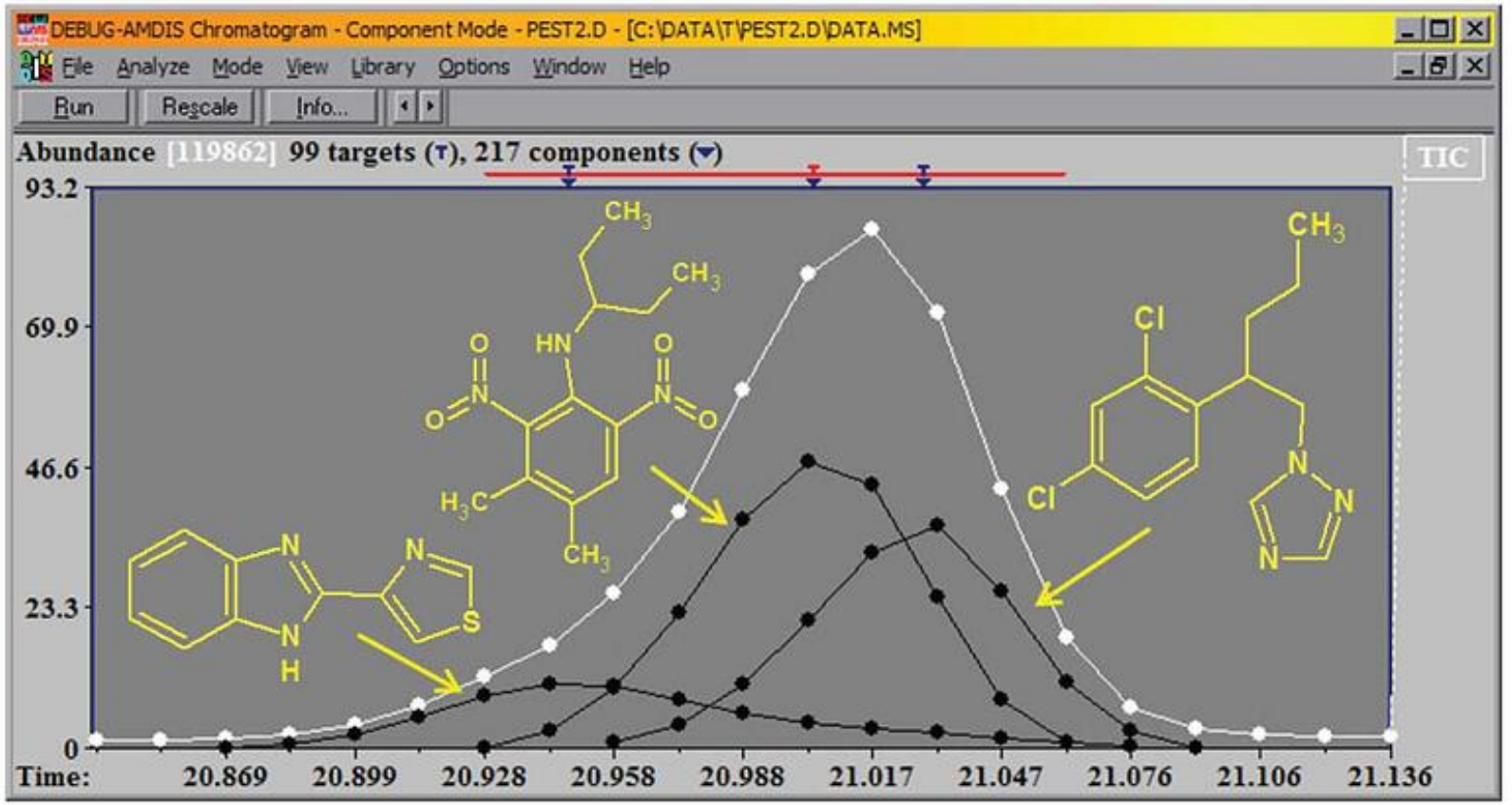


EIC : Extracted Ion Chromatogram



bioinfo

metabolome



données

proteome

- types de données :
  - SC : Spectral Counting
  - PC : Peak Counting
  - **XIC : Extracted Ion Chromatogram**
- données d'intensités MS
- matrices complètes sans NA
- richesse des protéomes:
  - de 1000 à 4000 protéines quantifiées sur graines
  - protéomes les plus complet chez les plantes : **10000** protéines

	p_AT1G01900	p_AT1G02305	p_AT1G02700	p_AT1G02780
Elevated_DS_1	8.284	6.403	7.395	7.501
Elevated_DS_2	8.315	6.862	7.422	7.548
Elevated_DS_3	8.303	6.716	7.357	7.520
Elevated_EI_1	8.421	6.722	6.965	7.648
Elevated_EI_2	8.552	6.629	6.845	7.498
Elevated_EI_3	8.560	6.626	7.075	7.704
Elevated_LI_1	8.349	6.545	6.922	7.826
Elevated_LI_2	8.527	6.906	6.799	7.744
Elevated_LI_3	8.606	6.654	6.755	7.638

données

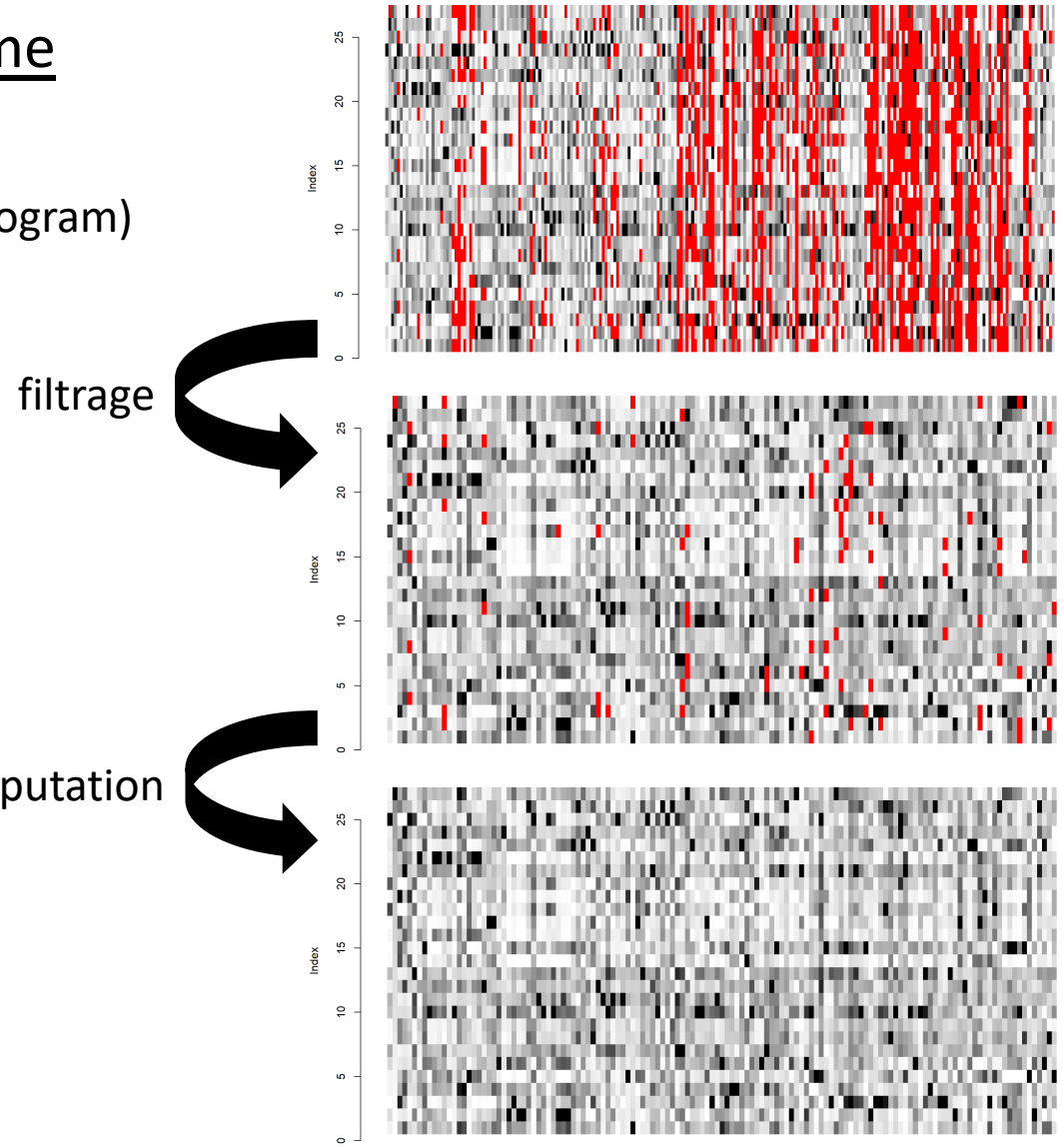
proteome

- transformation non linéaire usuelle : **log10**
- données **relatives** pour du profilage biochimique
- impossibilité de comparer, au sein d'un même echantillon, l'abondance respective des protéines entre elles
- évolutions :
  - quantification **absolue** qui nécessite des protocoles de préparation spécifiques et une utilisation particulière des machines
  - quantification indépendante des isoformes de **PTM** : protéoformes

données

metabolome

- **données d'intensités : EIC** (Extracted Ion Chromatogram)
- données incomplètes, grand nombre de NA
- méthodes d'imputation : ACP, Random Forest…
- gamme dynamique d'abondance très large
- transformation non linéaire usuelle : **log1p**

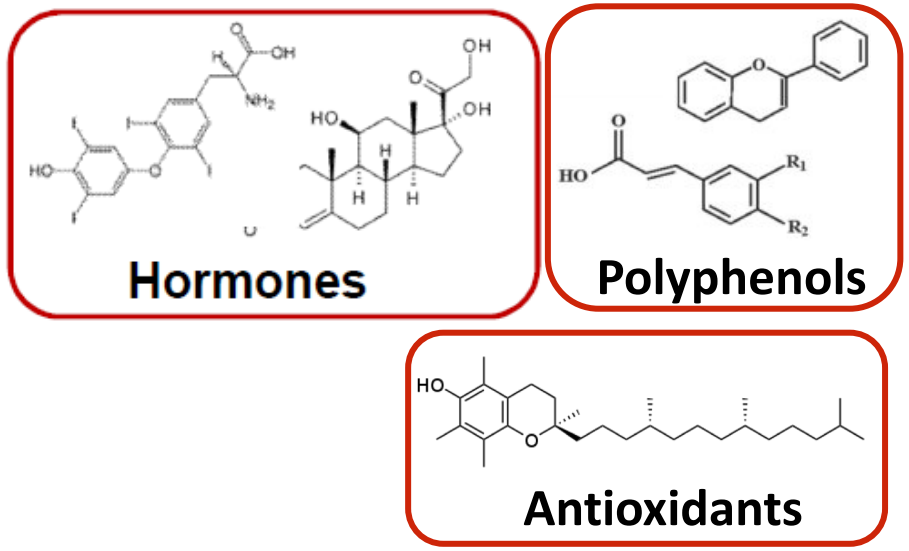


données

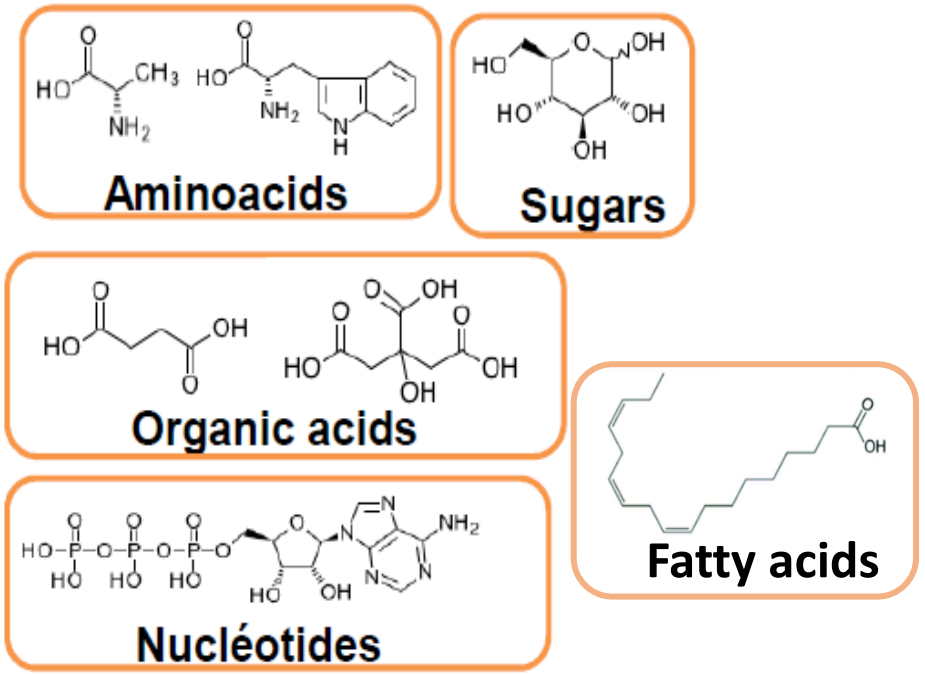
metabolome

- profondeur du métabolome : **300** métabolites
- <u>évolutions</u> :
  - LC-MS pour élargir la couverture
  - quantification absolue

Secondary metabolites

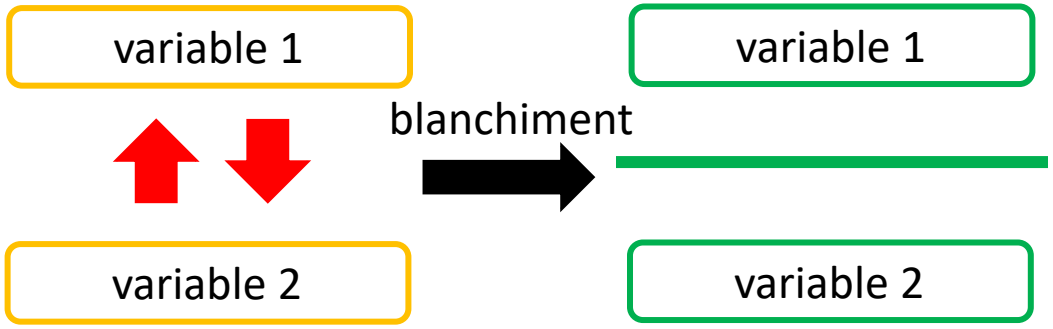


Primary metabolites



analyses statistiques

- choix d'une transformation appropriée : **log**, **scale**
- gestion des NA
  - origines expérimentales ou analytiques?
  - méthode d'imputation, de remplacement
  - imputer ou utiliser des méthodes stat tolérantes aux NAs?
- données quantitatives continues : approches statistiques classiques
- notion de dépendance entre variables :
  - absente en proteome
  - forte en metabolome
  - procédures de blanchiment pour la lever



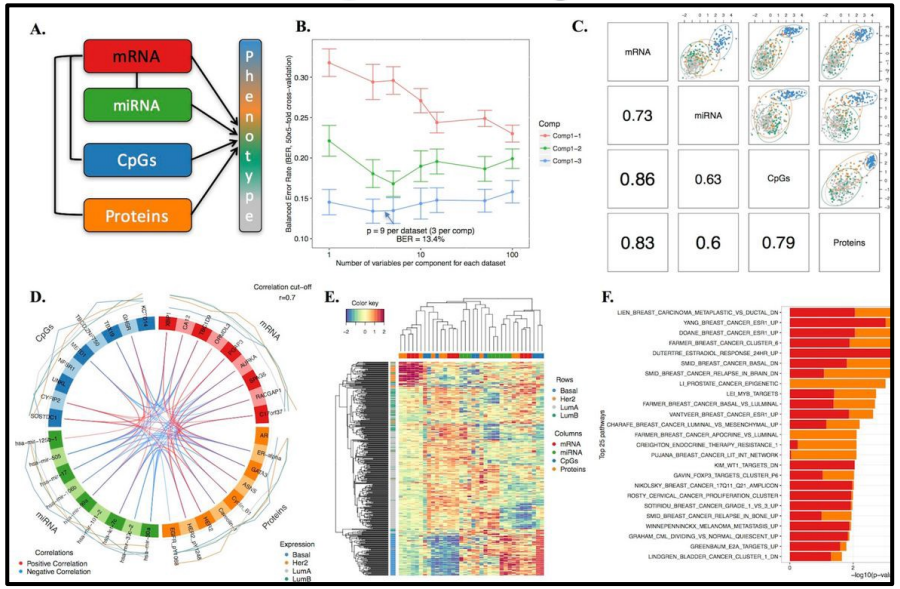
analyses statistiques

- outils plus avancés :
  - **glm lasso** pour sélection de marqueurs forts
  - **limma/DeqMS** pour analyse différentielle plus robuste : moderated Anova
  - **Coseq** sur données continues
  - Enrichissements fonctionnels sur « custom background » : **clusterProfilerR**, **gprofileR**
  - Enrichissements de pathway sur métabolomes : **MetaboAnalyst**

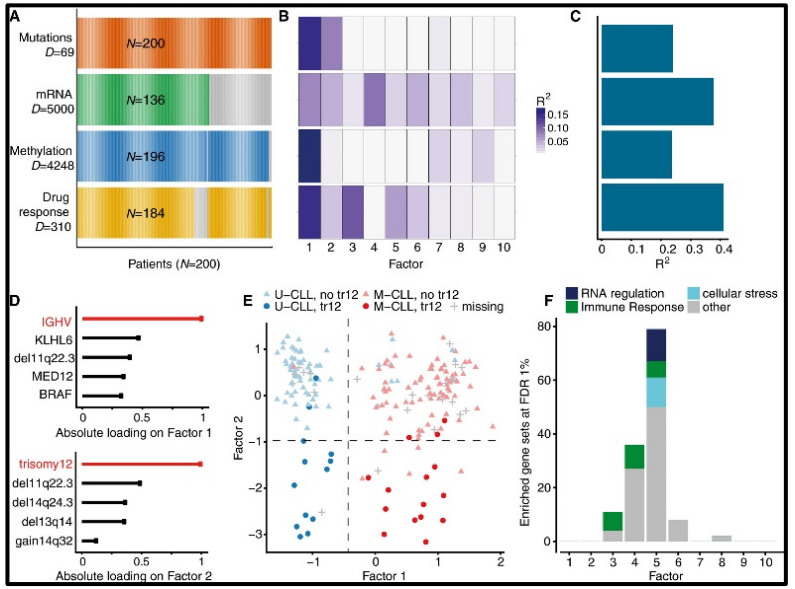
intégrations multi-niveaux

- intégration RNASeq discret avec protéomes et métabolomes continus
- linéarisation RNASeq: log(cpm) ou limma voom?
- <u>approches</u> :
  - **factorisation de matrices**
  - **réseaux**

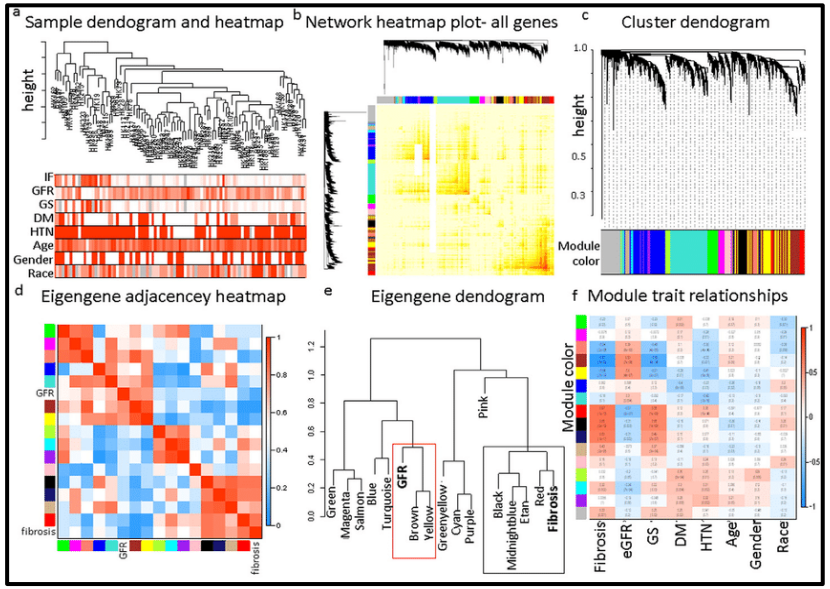
mix Omics



MOFA



WGCNA



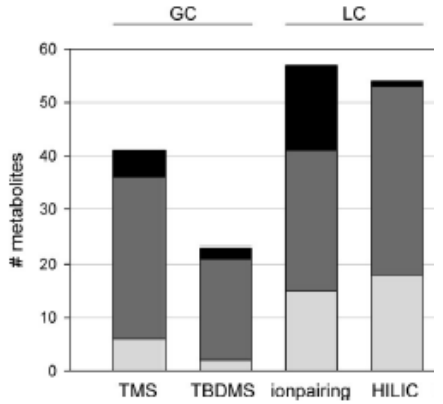
perspectives

couverture

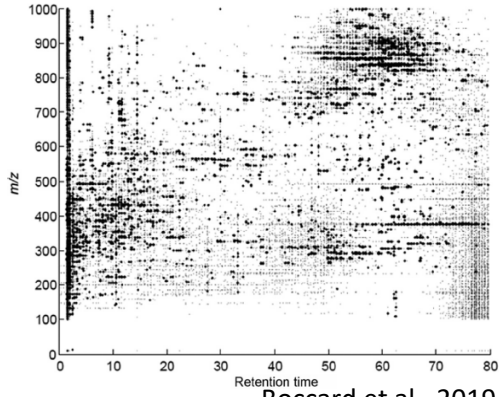
- préparation des échantillons :
  - « **on ne voit que ce qu'on extrait** »
  - apports de la déplétion et de l'enrichissement

- choix des systèmes analytiques :

chromatographe

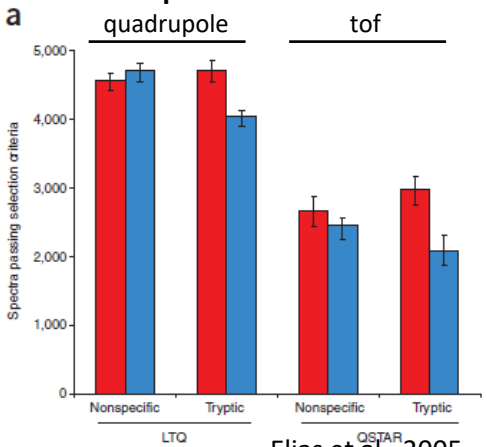


Büscher et al., 2009



Boccard et al., 2019

spectromètre

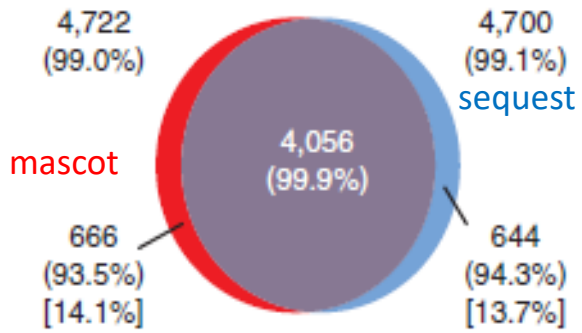


Elias et al., 2005

perspectives

couverture

- favoriser la LC-MS en métabolomique :
  - polyvalence : composés polaires et apolaires, plage de poids moléculaires élargie
  - passer de centaines à des **milliers** de métabolites
- choix des outils bioinfo :



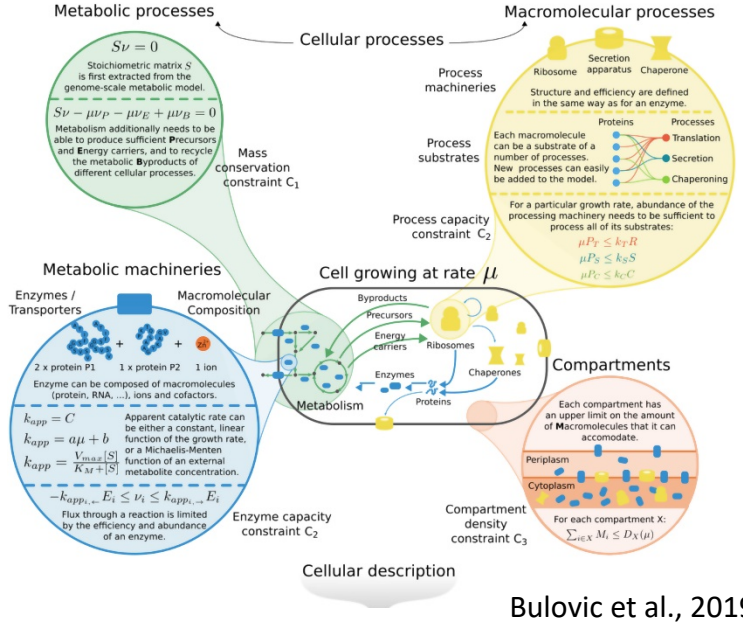
Elias et al., 2005

- problématique de l'assignation et de l'annotation des métabolites :
  - très variable selon les plateformes : pas de standard au niveau dénomination
  - nombre important de composés détectés mais non assignés
  - dépendante des **moteurs de recherche**, de la qualité et de la taille des **bases** interrogées

perspectives

quantification absolue

- nécessite calibration des machines et standards moléculaires diversifiés : protocoles complexes
- rend les données indépendantes des protocoles et des machines
- données comparables entre analyses
- données biologiquement plus parlantes **: concentrations cellulaires**
- importances des données absolues en modélisation quantitative du fonctionnement cellulaire



Bulovic et al., 2019