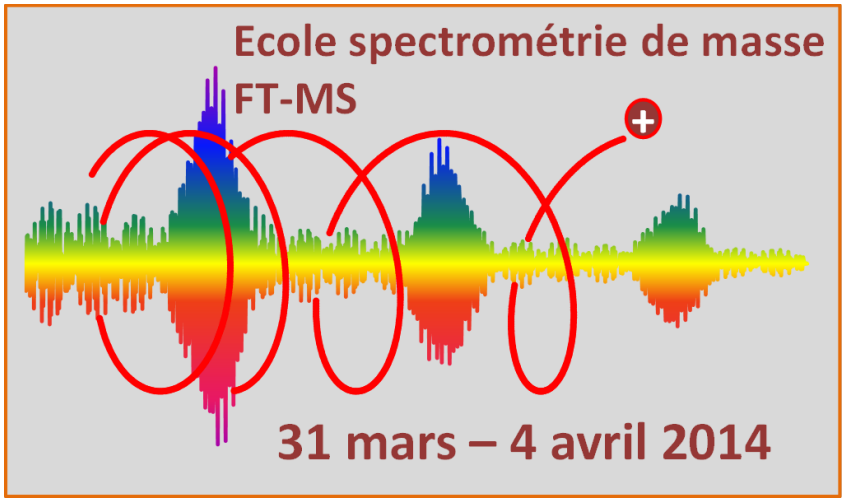
Partie II

De l'identification à la quantification

Thomas Burger

--

Etude de la Dynamique
des Protéomes



Plan

1. Lien avec l'identification
2. Overview du pipe-line (données XIC)
3. Statistiques descriptives
4. Pre-processing
5. Analyse différentielle
6. Conclusion

- Comparaison et quantification relative
- Méthodes, outils et choix d'EDyP
- Formalisation du problème

LIEN AVEC L'IDENTIFICATION

Nature des conditions à comparer

- Recherche de biomarqueurs, AP, etc. :

Sain ou référence **Malade ou pêche**

- Localisation subcellulaire / subcompartimentale :

Comp. 1 **Comp. 2** **Comp. 3**

- Comparaisons de plusieurs quanti :

Quanti. 1

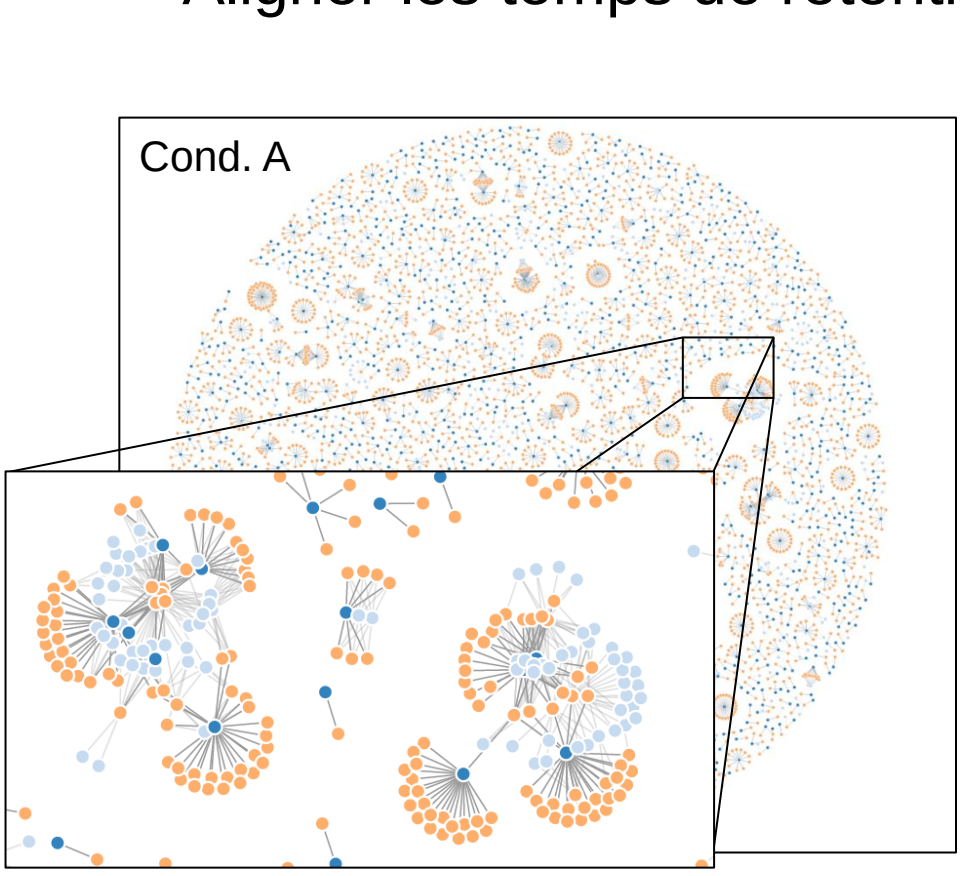
Protéines D.A.

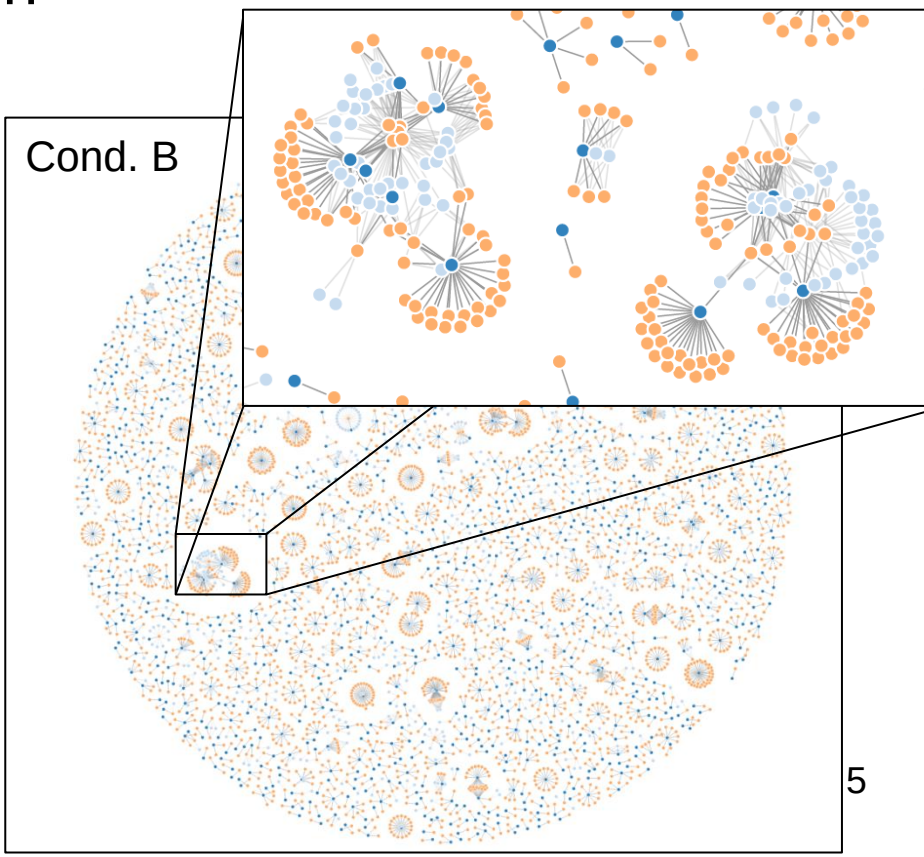
Quanti. 2

Protéines D.A.

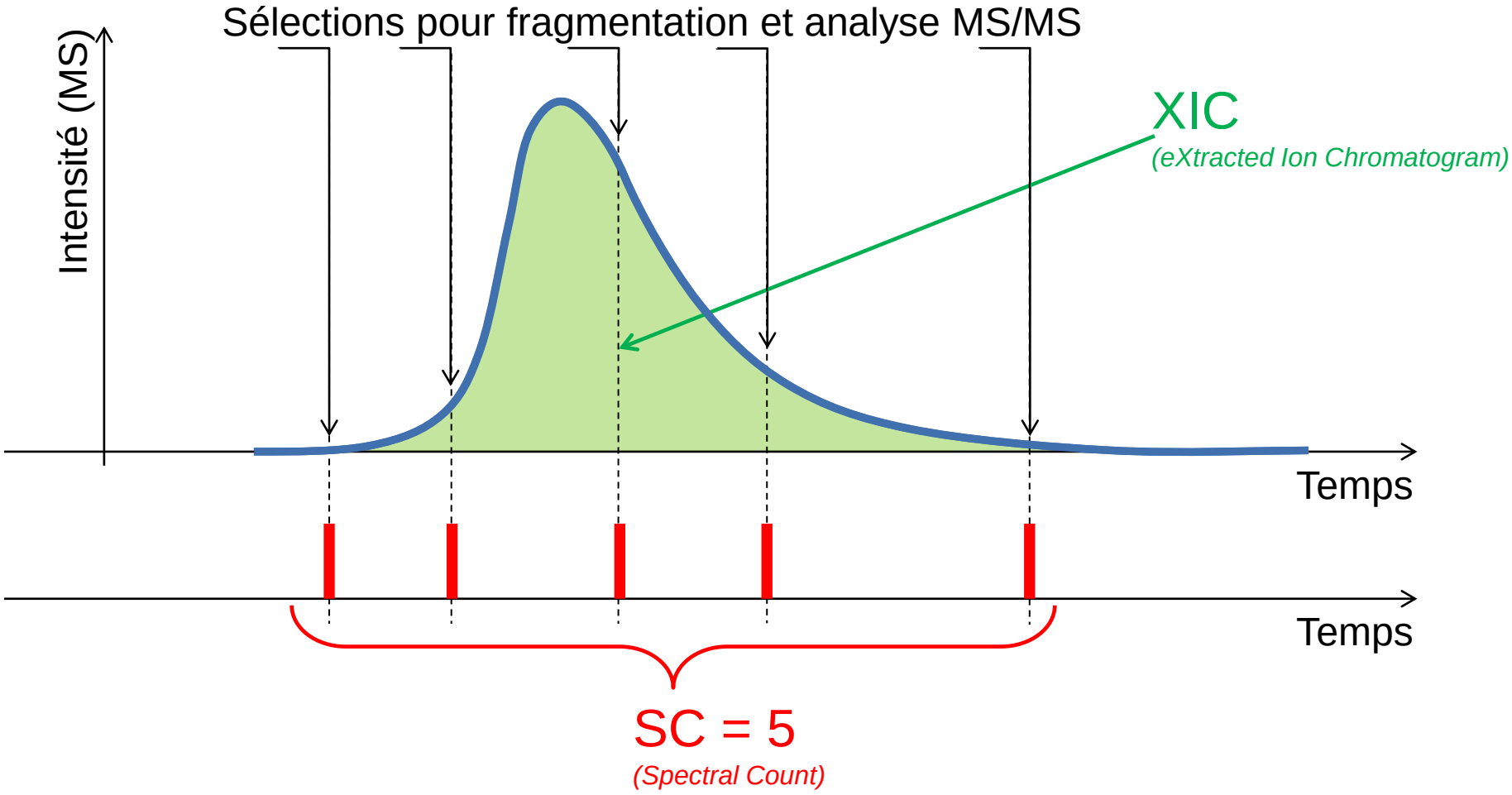
Comparaison des identifications

- Comparer les identifications de peptides
- Garantir la cohérence de l'inférence de protéines
- Aligner les temps de rétention





Méthodes de quanti. (label-free)



Outils disponibles et choix d'EDyP

MSn Extract
MaxQuant
MFPaQ
Mascot Distiller
Mascot
hEIDI
LC-Progenesis
Viper
MassCroQ
IRMa
DeconTools
OpenMS
Andromeda
Scaffold
Census
Sieve
Skyline

- **Spectral Count** :
  Mascot / IRMa / hEIDI

- **eXtracted Ion Chromatogram** :
  Andromeda / MaxQuant

- **Data Independent Analysis** :
  Effort de recherche actuel

Format des données

	1	2	3	4	5	6
Cond.	1	1	1	2	2	2
B.R.	1	2	3	1	2	3
T.R.	1	1	1	1	1	1
A.R.	1	1	1	1	1	1

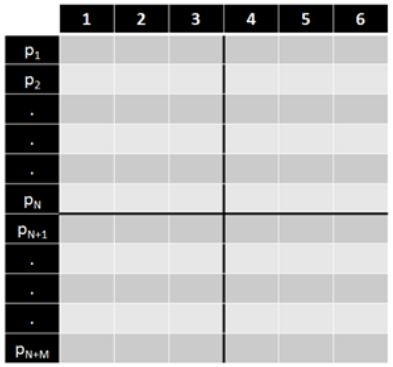
	1	2	3	4	5	6
p_1						
p_2						
.						
.						
.						
p_N						
p_{N+1}						
.						
.						
.						
p_{N+M}						

	Comments	DA
p_1		0
p_2		0
.		.
.		.
.		.
p_N		0
p_{N+1}		1
.		.
.		.
.		.
p_{N+M}		1

- Etapes de traitement des données XIC
- Roline / Shiny
- Quelques mots sur le SC

OVERVIEW DU PIPE-LINE

Les étapes du traitement



Statistiques descriptives

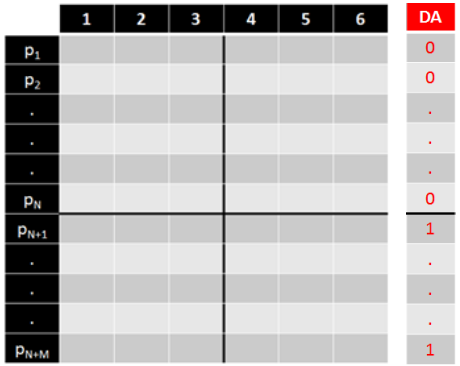
- Log-transformation
- Filtrage
- Quality control

Pre-processing

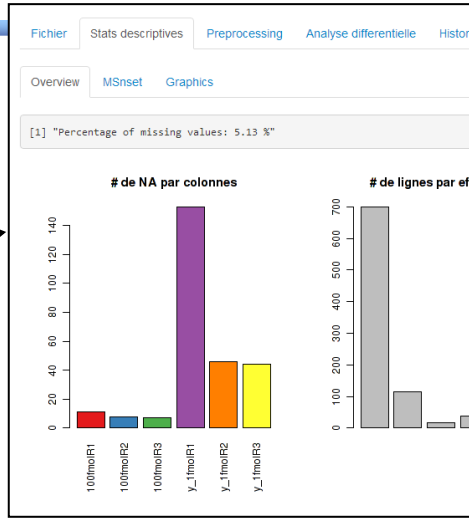
- Normalisation
- Imputation

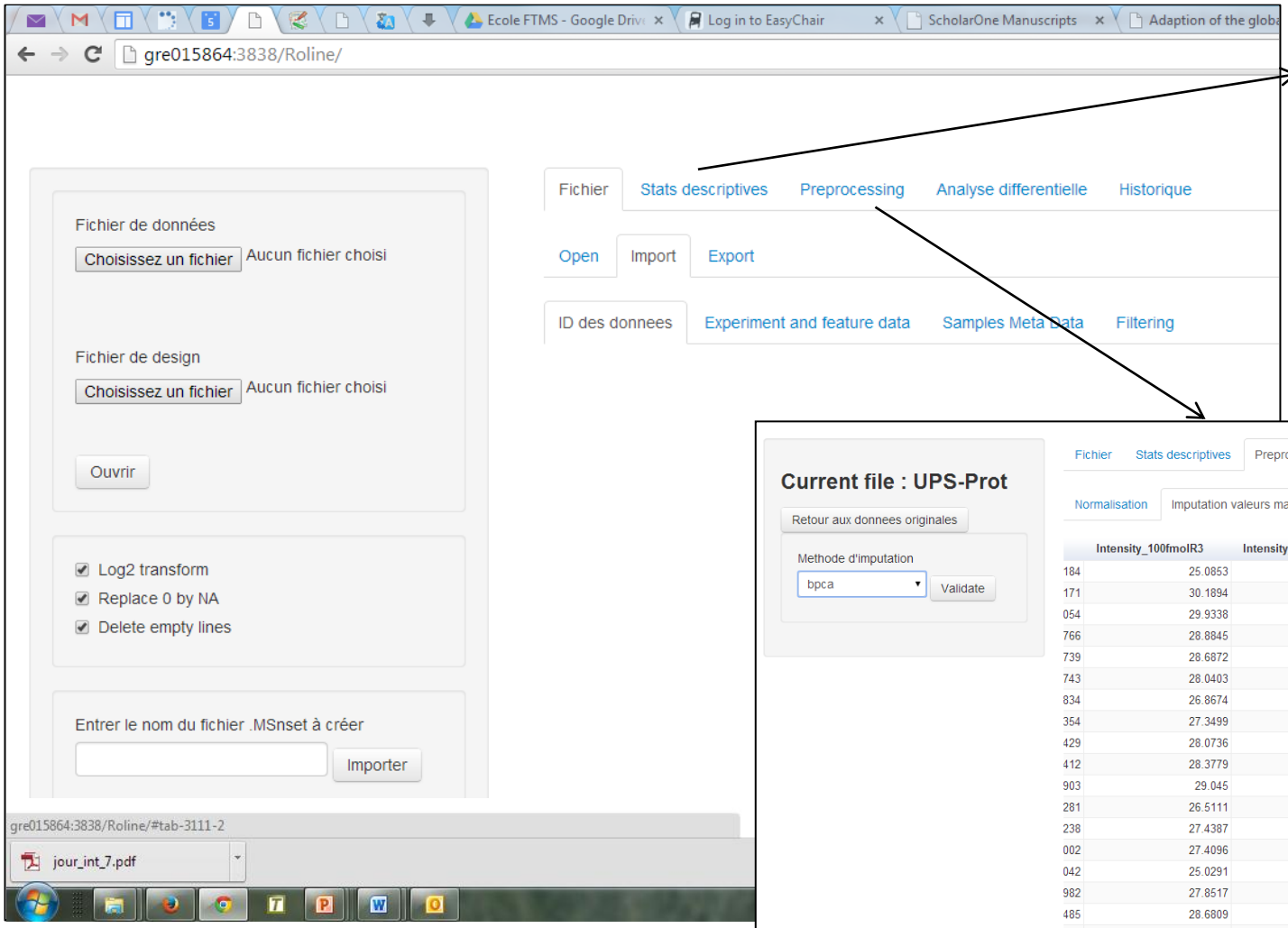
Analyse différentielle

- Tests d'hypothèse
- Correction tests multiples



Packages R et interface Shiny





Quelques mots sur le SC

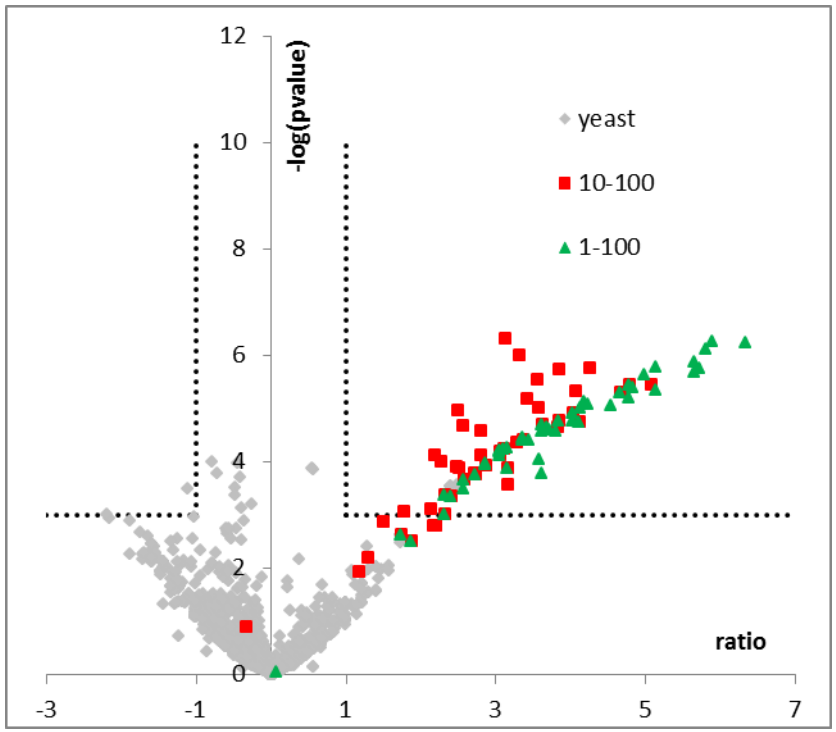
SC data

Preprocess XIC

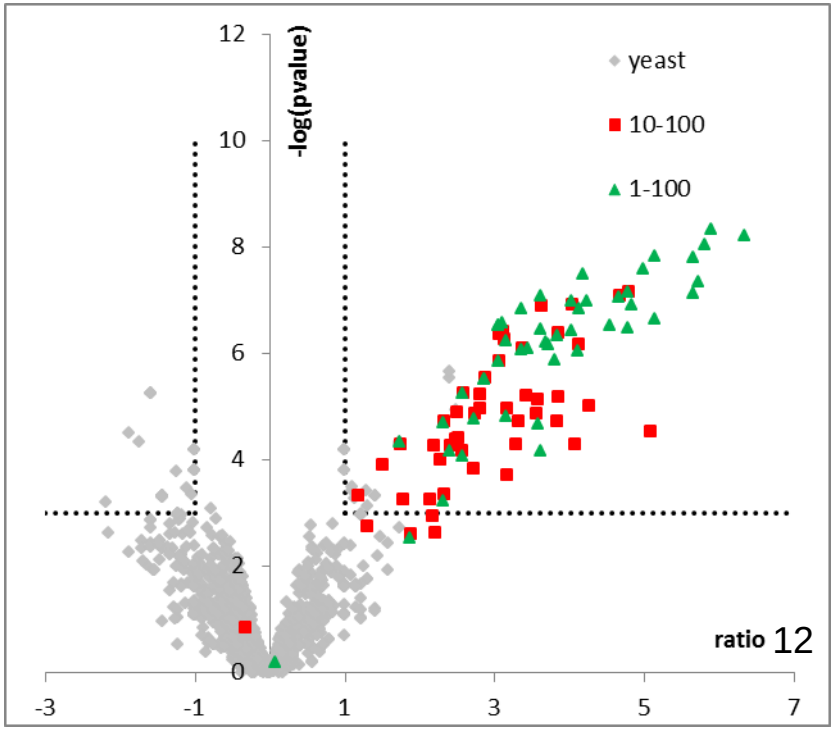
Test Beta Binomial

Test LIMMA

TDP 88%
FDP 6%



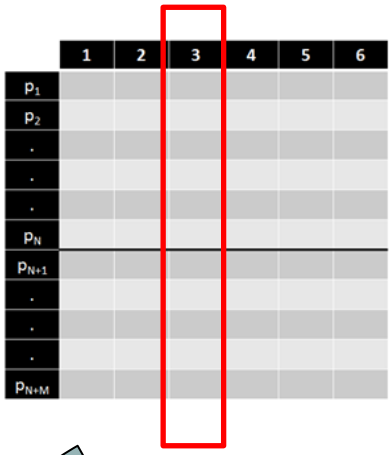
TDP 93%
FDP 24%



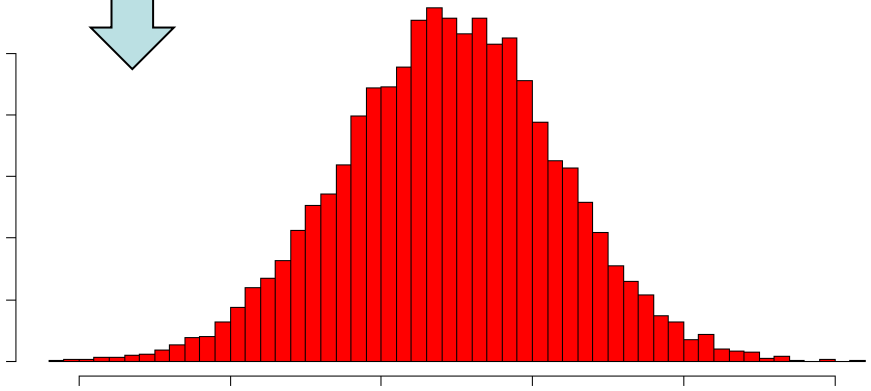
…

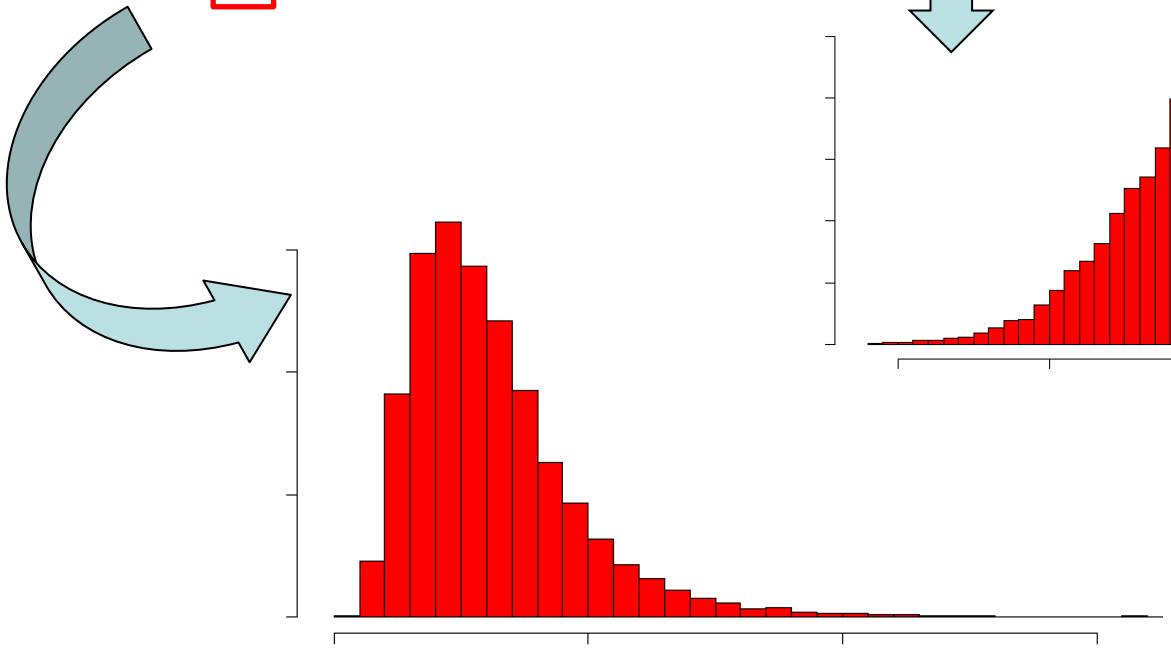
ETAPE 1:
STATISTIQUES DESCRIPTIVES

Log2-transformation



Log2()

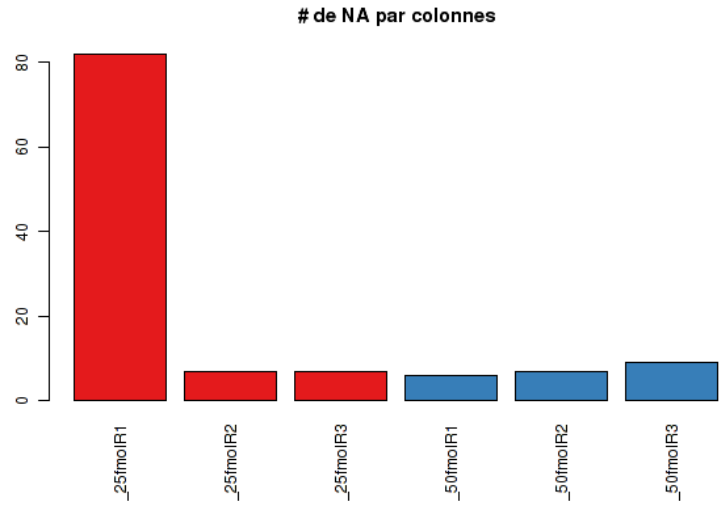


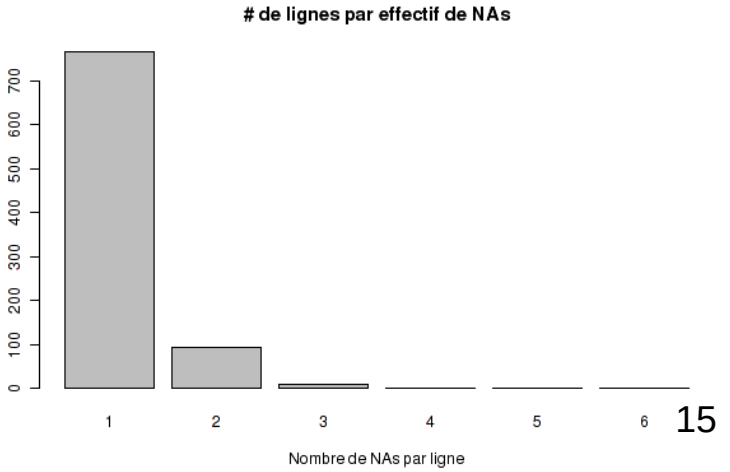


Filtrage – comptage des MV

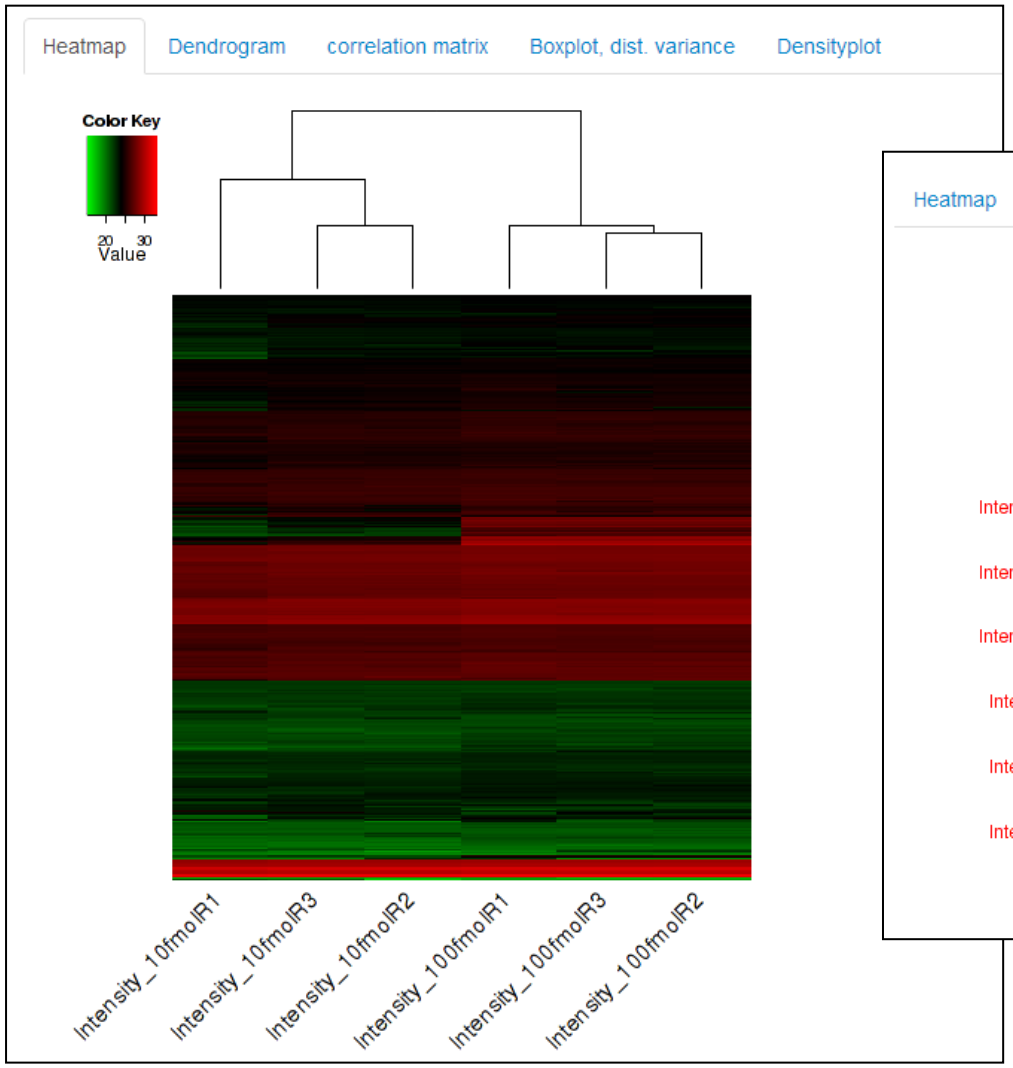
- Comptage des valeurs manquantes
  - Par ligne (protéine)
  - Par colonne (réplicat)
- Suppression des protéines non quantifiées
- Suppression des contaminants
- Etc.

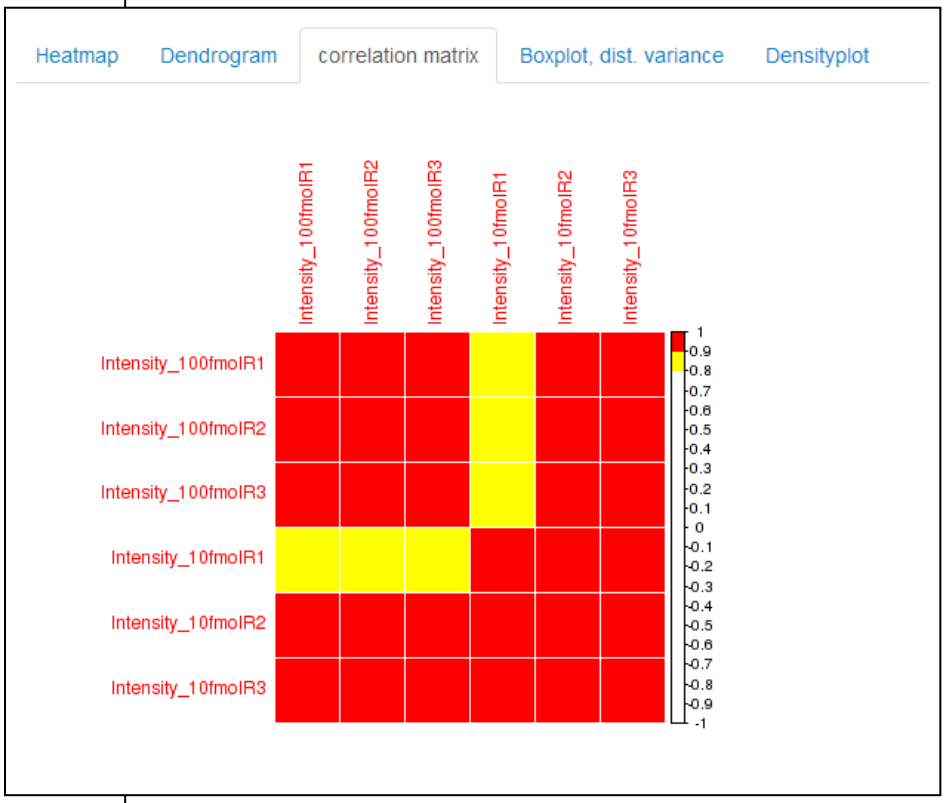
```
[1] "There is 6 samples in your data."
[1] "There is 873 lines in your data."
[1] "Percentage of missing values: 2.25 %"
```



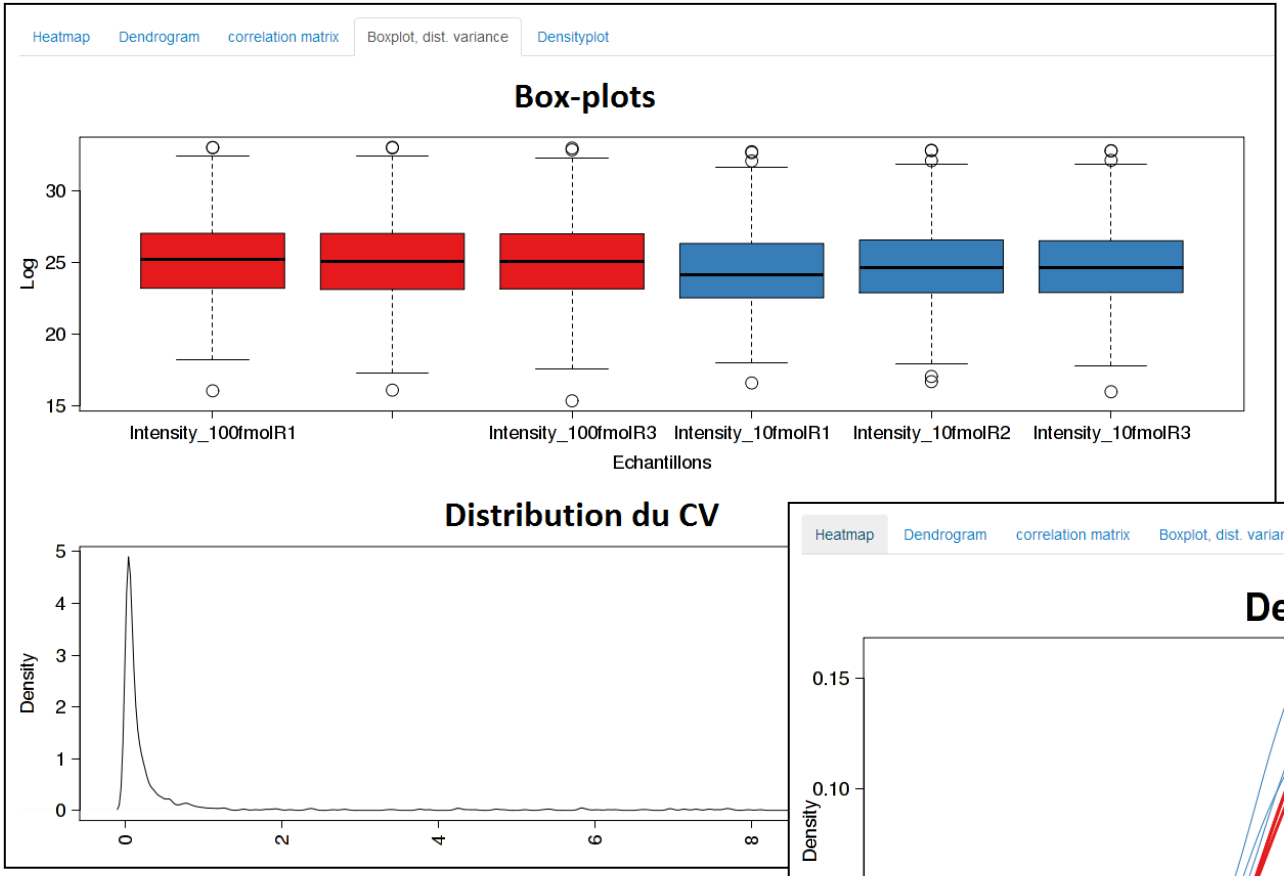


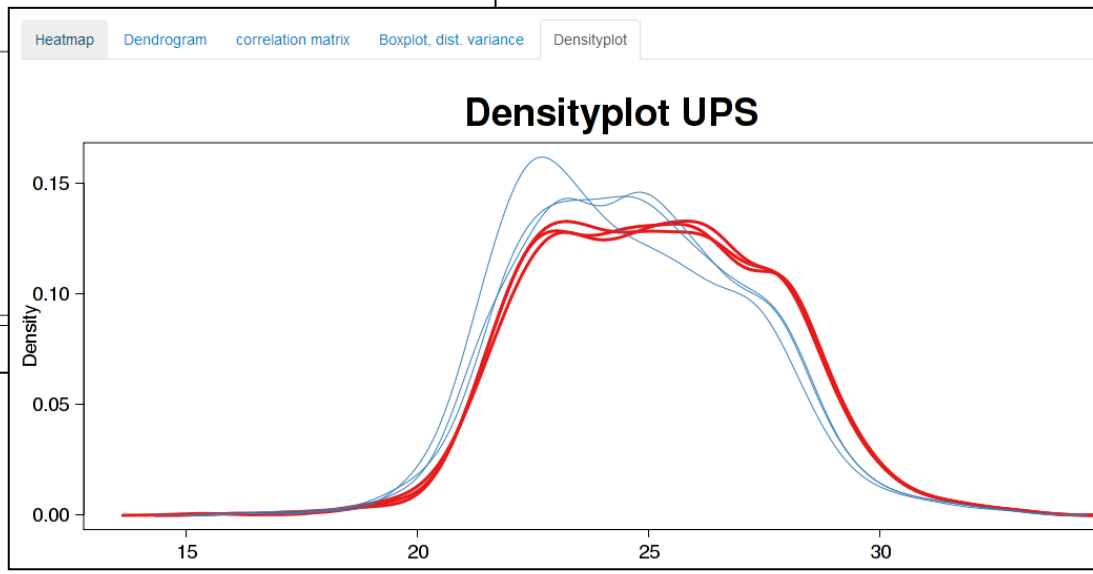
Vérification du design expe.





Caractérisation des densités



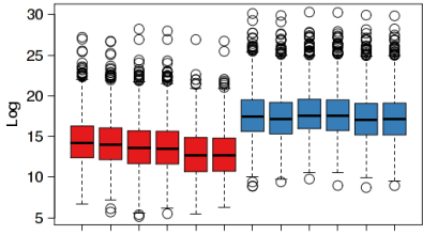


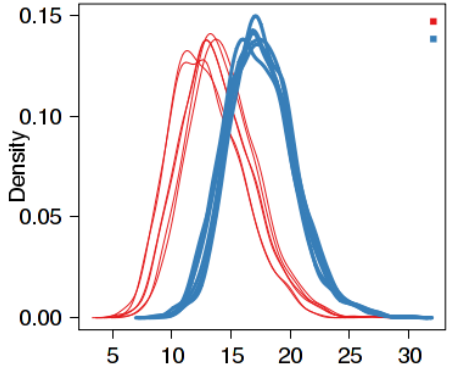
- Normalisation
- Imputation

ETAPE 2: PRE-PROCESSING

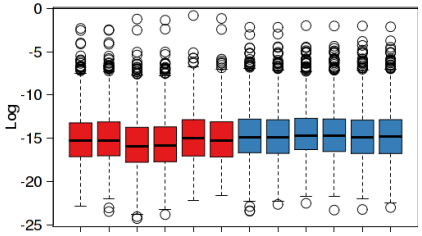
Différentes normalisations existent

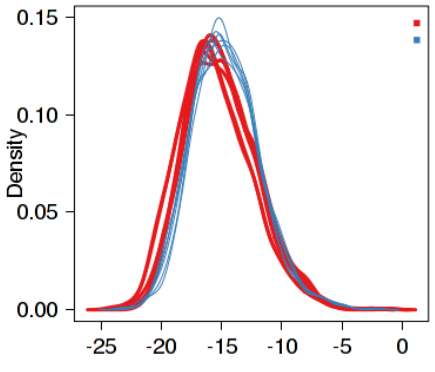
brut



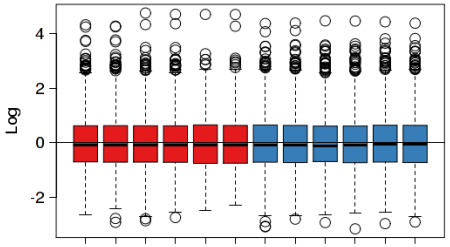


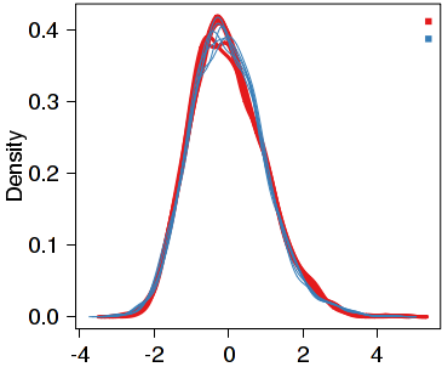
Somme colonnes



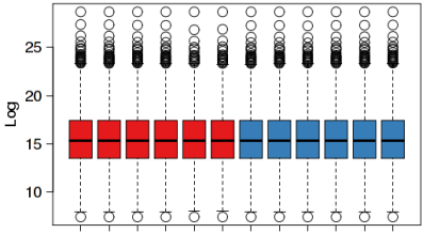


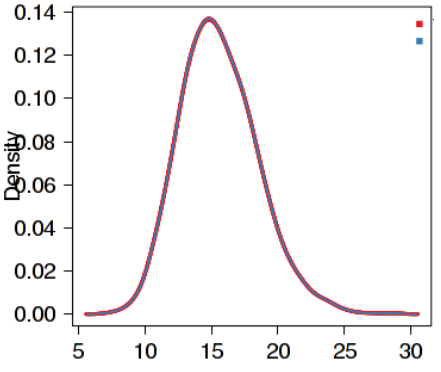
Centrage reduction





Quantile





Comment choisir ?

- Il n'y a pas de meilleure méthode
- Le choix doit dépendre de ce que l'on cherche dans les données
- Eviter malgré tout de prendre la méthode qui donne la liste de protéines la plus longue…

Valeurs manquantes en statistique

- **MAR** et **MCAR** : Missing (Completely) At Random
  - Dues à l'accumulation d'erreur dans le pipe-line
  - Pas liées à l'intensité de la mesure
  - Réparties de manière uniforme dans le dataset

- **MNAR** : Missing Not At Random
  - Dépendent de l'intensité de la mesure

- Les méthodes d'imputation classiques ou issues de la transcriptomique se focalisent sur les MAR

et en protéomique ?

- Principalement des MNAR (lower limit of detection, gamme dynamique, etc.)
- Mais aussi quelques MCAR au milieu… On comprends pourquoi les méthodes classiques ne marchent pas !
- En attendant des méthodes d'imputation originale, pas de solution satisfaisante:
  - Génomique : BPCA (moins pire que les autres)
  - Perseus (MaxQuant) : Tirage aléatoire ad-hoc et mal justifié
  - EDyP (Yohann C.) : valeur déterministe = un faible percentile de chaque colonne

Travaux en cours à EDyP (C. Lazar)

- Mise en place d'un outil de diagnostic MCAR/MNAR
- Développement d'un algorithme d'imputation pour les MNAR
- Benchmark des méthodes MCAR disponibles
- Assemblage dans un logiciel unique d'imputation specifique à la protéomique

Imputation protéique/peptidique

- Il y a moins de valeurs manquantes sur les protéines que sur les peptides.
  - Pourquoi ?
  - Est-ce mieux de travailler sur les protéines ?

- Quand on calcule l'intensité des protéines:
  - Par une somme : les VM des peptides => 0
  - Par une mediane : les VM des peptides => mediane
  - Etc.

- On réalise une imputation implicite de mauvaise qualité. **Il faudrait travailler au niveau peptidique.**

- Utilisation du cadre statistique du test d'hypothèse
- Outils classiques et choix d'EDyP
- Regard sur l'avenir

ETAPE 3: ANALYSE DIFFERENTIELLE

Formalisation du test

- Erreur de Type I ? Erreur de Type II ?
- Hypothèse nulle ? Alternative ?
- Comment interprète-t-on une p-value ?
- Quelles sont les deux décisions consécutives au test d'hypothèse ?

Outils disponibles

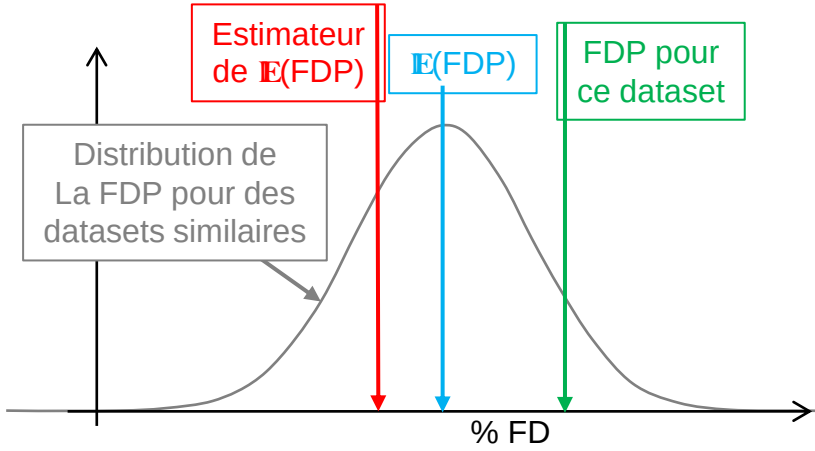
- Les *t*-tests
  - Students (version originale)
  - Welch (nombreuses variations)
  - Fudge factor (S.A.M.)
  - Limma
  - ANOVA, modèles mixtes, etc.

- Likelihood-ratio tests
  Comparer la vraisemblance de 3 modèles dont le ratio suit une loi du χ^2.

- Tests non-paramétriques
  Basés sur la statistique de rang (Mann-Whitney, Log-rank, McNemar, etc.)

Contrôle des fausses découvertes

- Sur des benchmarks:
  - la FDP est souvent > 20%
  - avec un FDR supposé < 5%



- Par ailleurs:
  - **Benjamini-Hochberg** nécessite que les p-values sous H0 suivent une loi uniforme
  - Le **Permutation-Based FDR** ne contrôle pas le FDR
  - La **q-value de Storey** nécessite l'indépendance des tests

- Pas de solution idéale…

Choix d'EDyP

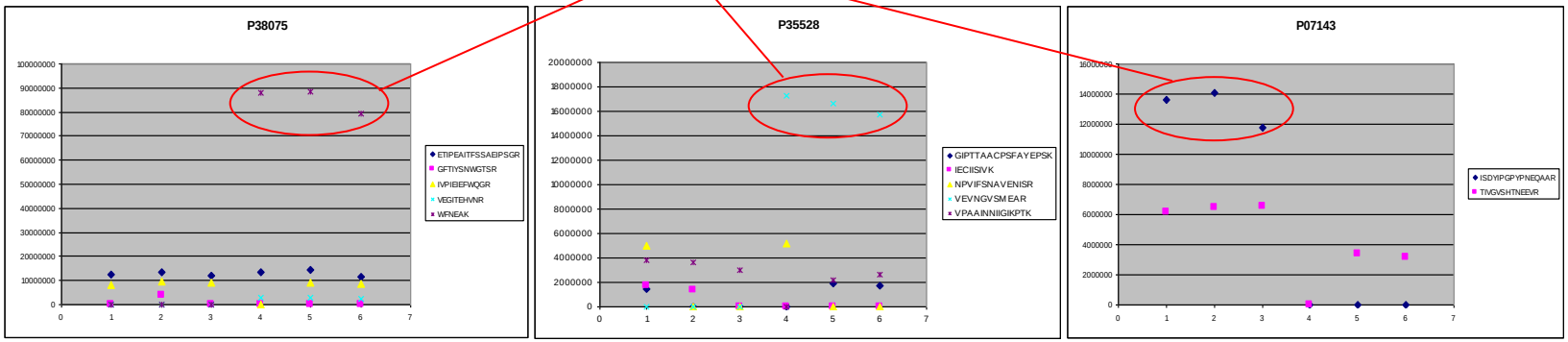
Encore en cours d'affinage…

- LIMMA + seuil sur le FC + correction BH semble préférable (hypothèse uniformité ?)
- SAM test (fudge factor + permutation based FDR)
  - Pb1: le fudge factor est détourné => seuil FC
  - Pb2: la correction par permutation…
- Développement de nouveaux tests spécifiques

Travailler au niveau des peptides

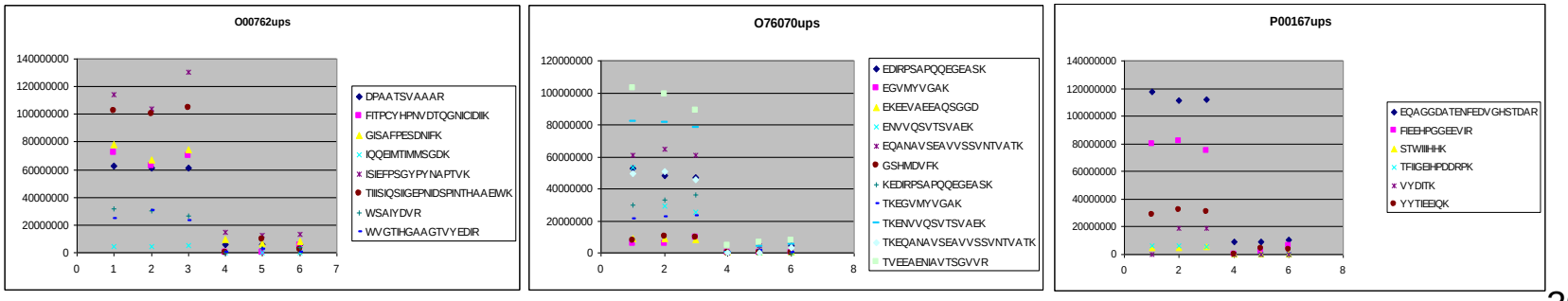
Misidentifications or misalignment ? → coherence of peptide ratios…

Yeast proteins



Protein Intensity = Σ peptide intensities

UPS1 proteins



(Yohann COUTE, Anne-Marie HESSE)

…

CONCLUSIONS GÉNÉRALE

Take Home Message I

- Travailler autant que possible au niveau des peptides !
  - Imputation
  - Test
- Vers la prise en compte des peptides
  - protéo-spécifiques
  - Non observés

THM II

- Il y a beaucoup de questions encore en suspens, impliquant des traitements non-optimaux.

- Il vaut encore mieux des traitements que l'on sait non-optimaux que de ne pas s'être posé la question (=choix au pif)

THM III

- Il faut savoir ce que l'on fait:
  - Questionner la qualité de ces données
  - Avoir conscience des erreurs qui se propagent dans le pipe-line
  - Ne pas utiliser d'outils fermer
  - Etc.

- Accepter que la protéomique est un sujet de recherche pour les statistiques…

  (Et non la dernière étape "pénible et à sous-traiter" avant publication!)