Analyse statistique de données de protéomique et de métabolomique quantitative

Thomas Burger

Journée Intégrative de Protéomique et Métabolomique

~~26 mars 2020 à Lyon~~ pfff….
~~08 Octobre 2020 à Lyon~~ ☹
08 Octobre 2020 en visioconférence !

Le classique: « Moi, ma vie, mon œuvre… »

- **Chercheur CNRS en science des données**
  - Machine learning
  - Statistics
  - Signal processing
  - Data fusion
  - Etc.

- Affecté au laboratoire **EDyP** (Etude de la Dynamique des Protéomes)

  - **Recherche méthodologique**
    Amélioration des méthodes de traitement de données LC/MS

  - **Développement de Prostar**
    Outil logiciel permettant l'analyse statistique de données de protéomiques de découvertes

EDyP et la métabolomique

- **ProMétIS**
  - Avec les différentes infrastructures nationales (proteo, metabo, bioinfo, etc.)
  - Projet piloté par Etienne

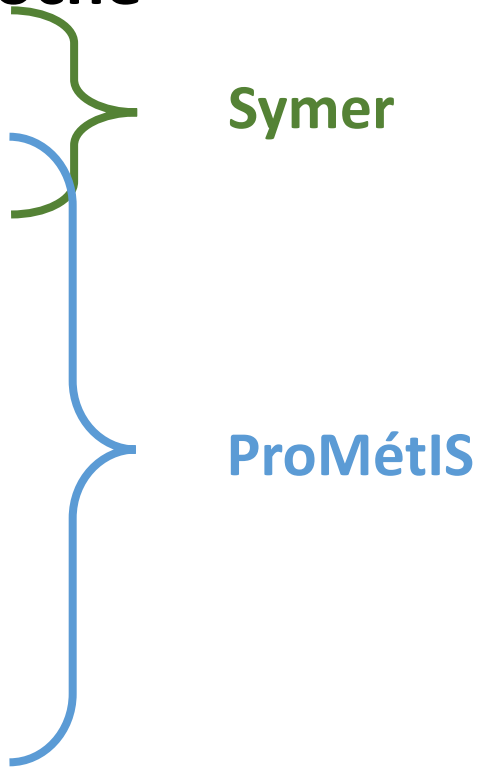
- **Symer**
  - Programme inter-interdisciplinaire grenoblois
  - Accompagner le développement computationnel de la plateforme de métabolomique du CHU Grenoble-Alpes
  - Etendre Prostar à la métabolomique
  - Recrutement d'Enora Fremy (ingénieure en bioinformatique~~, ici présente~~)

- De plus en plus de projets impliquant un volet « **multi-omique** »

Le « pourquoi? » de ces journées

- **Traiter des données omiques dont la production est proche**
  - Tenir compte des spécificités du pipeline LC-MS
  - Echange de bonnes pratiques
  - Identification de workflows communs

- **Réflexion dans un contexte multi-omique**
  - Outils d'analyse compatibles avec les deux types de données
  - Reconstruction d'une histoire biologique cohérente
  - Analyse computationelle commune



Problématique

- Que nous pouvons faire de la même manière ?
  (Pourquoi / Comment ?)

- Existe-t-il des spécificités nécessitant des traitements différenciés ?
  (Pourquoi / Comment ?)

Des workflows +/- similaires

	R_1	...	R_n	...	R_N
Cond.	1	...	...	...	2
B.R.	1	...	...	...	8
T.R.	1	...	...	...	3
A.R.	1	...	...	...	2

	R_1	...	R_n	...	R_N
E_1					
E_2					
.					
.					
.					
E_p					
.					
.					
E_{P-1}					
E_P					

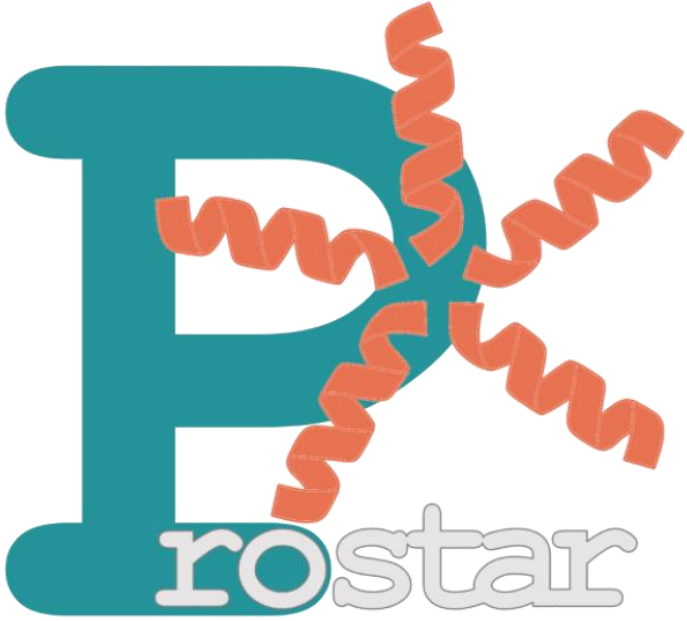
	info1	info2
E_1		
E_2		
.		
.		
.		
E_p		
.		
.		
E_{P-1}		
E_P		

Pipeline standard d'analyse statistique

1. **Filtrage**
   - Des échantillons (si en nombres suffisants)
   - Identifications
     - Qualité (en fonction du score, de la nature target/decoy)
     - Organismes, contaminants, etc.
   - Valeurs d'intensité non observées
     - Réellement manquantes
     - Récupérées par « Match between runs »
2. **Normalisation**
   - Effets de lots (« batch effect »)
   - Meilleure prise en compte de la variabilité de mesure
3. **Imputation des valeurs manquantes**
   - En fonction de leur nature
     - MNAR (sous le seuil de quantification de la MS)
     - MCAR (non-exhaustivité de la chaîne de mesure)
   - En protéomique, à plusieurs niveaux
     - Peptidique
     - protéique
4. **Tests statistiques**
   - Classiquement, comparaisons de paires (analyse differentielle)
   - Beaucoup d'autres plans d'expériences possibles
     - Plusieurs conditions (ANOVA)
     - IP /coIP
     - Prise en compte de facteurs confondants
     - Études longitudinales
     - Etc.
5. **Correction de la multiplicité des tests**
   - Benjamini-Hochberg et contrôle du FDR
   - D'autres corrections peuvent être nécessaires:
     - Ex: Différents tests post-hoc suivent l'ANOVA
     - Interférence forte avec le contrôle du FDR
6. **Traitement post-analyse différentielle**

(Très variable en fonction de la question biologique)

Suite logicielle Prostar



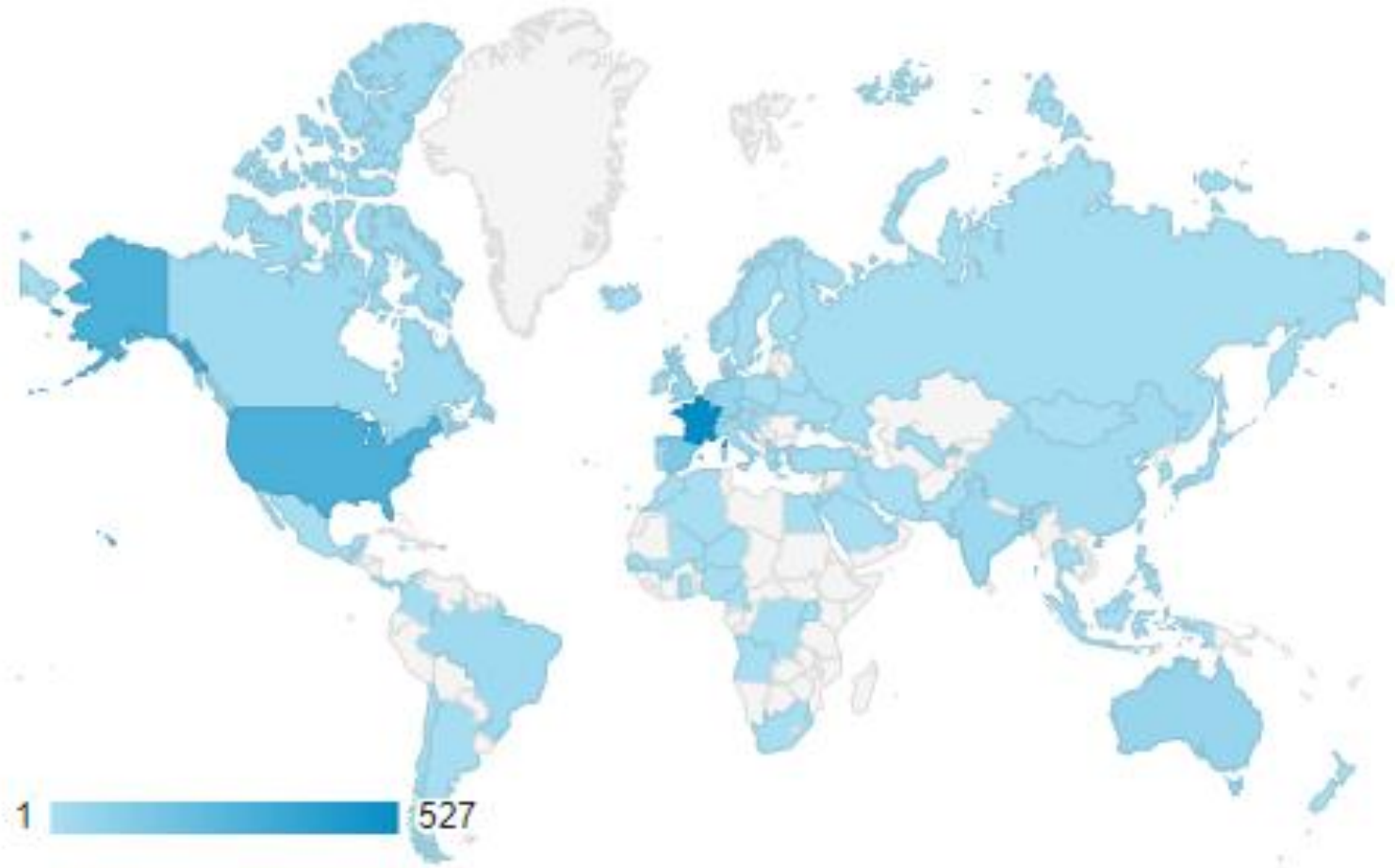
- **Développeurs principaux**:
  Samuel Wiezcorek, Thomas Burger

- **Autres contributeurs**:
  Florence Combes; Quentin Giai-Gianetto (Institut Pasteur, France); Laurent Gatto (Université catholique de Louvain, Belgique); Cosmin Lazar; Helene Borges; Yohann Couté; Christophe Bruley; Anne-Marie Hesse; Alexia Dorffer; **Enora Fremy**

- Ensemble de packages **R** + interfaces graphiques **Shiny**

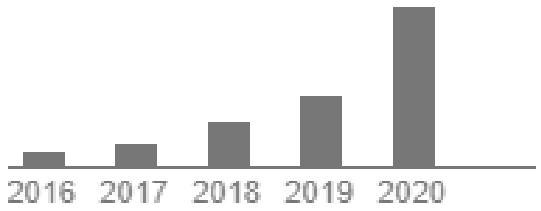
- Implémente le précédent workflow; d'autres sont à venir (métabolique; peptidomique; peptide-level proteomics, etc.)

Team-building & renforcement d'égo



	1,692 % of Total: 100.00% (1,692)
1. France	**527** (30.98%)
2. United States	**296** (17.40%)
3. Spain	**137** (8.05%)
4. Germany	**68** (4.00%)
5. United Kingdom	**58** (3.41%)

Nombre total de citations — Cité 71 fois



Analyse quantitative en métabo et protéo

- De l'importance de l'identification
  - P: On cherche ce qui bouge parmi ce qu'on a identifié
  - M: On cherche à identifier ce qui bouge

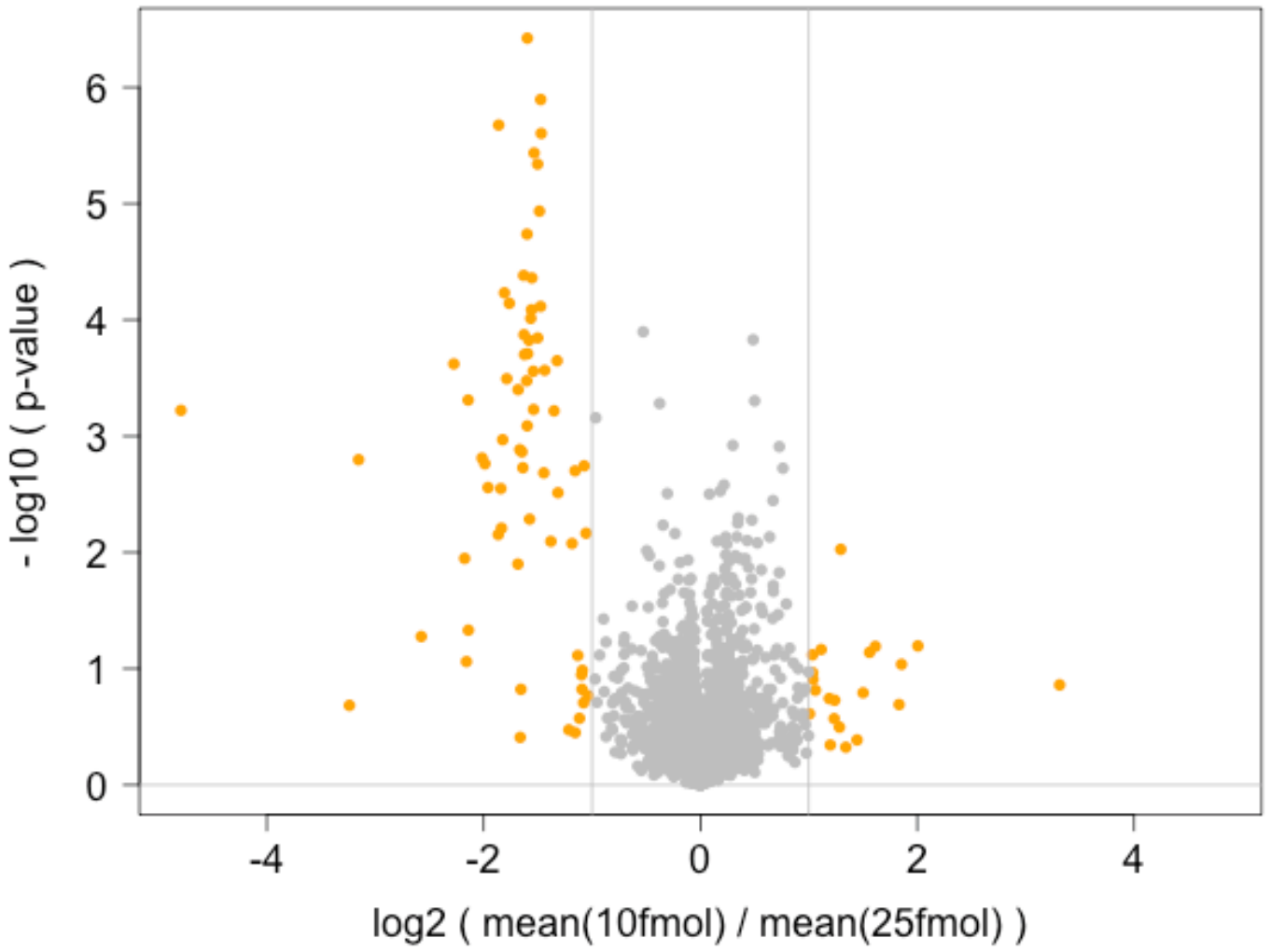
- De la nature multi-échelle de la protéomique

  Peptides (observations) -> Protéines (intérêt) -> Protéoformes (???)

- Est-ce qu'un analyte « D.A. » a la même signification ?

  Ex: Caractérisation des différences vs. recherche de biomarqueurs

Fold-Change filtering



A quoi ca sert ?

- En théorie, à éviter de selectionner des analytes :
  - Dont le FC est trop faible pour induire un intérêt biologique
  - Dont la variabilité biologique rendra le FC invisible

- En pratique:
  - A limiter l'anti-conservatisme des p-values (test statistique trop optimiste)
  - A « choisir à la main » (un peu) les analytes selectionnées

- En conclusion:
  C'est un outil utile, mais dont l'usage peut vite se transformer en « *tripatouillage de données pour leur faire dire ce qu'on veut* »

Quand le faire ?
(résumé d'un échange avec Etienne)

- **Etienne**: Forcement APRES le contrôle du FDR, car sinon, le filtrage du FC peut biaiser le calcul du FDR *(ce n'est pas faux)*

- **Thomas**: Forcement AVANT le contrôle du FDR, car si filtrage post-FDR, alors le FDR n'est plus garantie *(ce n'est pas faux non plus)*

- **Synthèse:** il vaut mieux ne pas en faire du tout… ☺

Malgré tout…

- Avec CP4P, il est possible de vérifier que le contrôle du FDR est correct, même si précédé d'un filtrage sur le FC.

- Dans certains cas, la liste différentielle ne sera pas entièrement utilisée de sorte qu'on peut privilégier les analytes avec FC important

- Il est possible de travailler proprement avec un filtre sur le FC
  - Mais le tripatouillage n'est jamais loin !
  - Pour l'éviter: **dans le toute mettre un seuil plus faible !**

Une méthode pour fixer le « FC cut-off »

- **Ne pas chercher à:**
  - Filtrer les fausses découvertes
  - Sélectionner les protéines d'intérêt

- **Mais plutôt à:**
  - Éliminer les protéines qu'on est prêt à perdre malgré une excellente p-value
  - Ne pas chercher à éliminer beaucoup de découvertes

Méthodes d’analyse algébrique

Définition

1. Représentation des données sous la forme d'une matrice (ex: matrice de variance-covariance)

2. Transformation de cette matrice par des outils algébriques
   - Diagonalisation (recherche d'espace propre, de valeurs singulières, etc.)
   - Transformation géométrique des vecteurs-colonnes (projection, rotation, etc.)
   - Minimisation d'un critère numérique sur cette matrice
   - Etc.

3. Usage de la matrice transformée pour mieux comprendre les données

La plus connue…

… au monde:

- Analyse en composantes principales (ACP)
- Cas particulier de PLS
- Usages
  - Visualisation 2D minimisant la variance
  - Réduction de dimensionnalité
  - Filtrage des porteuses du bruit

… en métabolomique:

- Orthogonal PLS – Discriminant Analysis: OPLS-DA
- Cas particulier de PLS
- Usages:
  - Analyse discriminante
  - Recherche de séparabilité optimale
  - Sélection des variables les plus discriminantes

Pour quoi faire ?

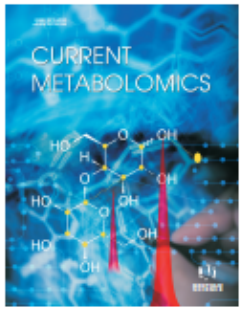
- Pour une analyse différencielle, OPLS-DA est beaucoup plus puissante
- Presque trop… très gros risques d'over-fitting
- Très peu utilisable en pratique en protéomique, mais justifiée dans certains cas en métabolomique
- Mais dans certains cas seulement…

Current Metabolomics, 2016, *4*, 97-103

RESEARCH ARTICLE

BENTHAM SCIENCE

PCA as a Practical Indicator of OPLS-DA Model Reliability



Bradley Worley and Robert Powers*

Department of Chemistry, University of Nebraska-Lincoln, Lincoln, NE 68588-0304, USA

Abstract: ***Background:*** Principal Component Analysis (PCA) and Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA) are powerful statistical modeling tools that provide insights into separations between experimental groups based on high-dimensional spectral measurements from NMR, MS or other analytical instrumentation. However, when used without validation, these tools may lead investigators to statistically unreliable conclusions. This danger is especially real for Partial Least Squares (PLS) and OPLS, which aggressively force separations between experimental groups. As a result, OPLS-DA is often used as an alternative method when PCA fails to expose group separation, but this practice is highly dangerous. Without rigorous validation, OPLS-DA can easily yield statistically unreliable group separation.

Robert Powers

ARTICLE HISTORY

Received: April 11, 2016
Revised: June 1, 2016
Accepted: June 6, 2016

Analyse multi-échelle

Ions, peptides, protéines, proteoformes, etc.

Une difficulté supplémentaire

Historiquement, en protéomique, il s'agit d'une réelle difficulté:

- **Computationelle**: Il faut « reconstruire le puzzle »…
  - On ne sait pas forcement comment faire
  - Cela prend du temps

- **Biologique**: Lien indirect entre élément mesuré et élément d'intérêt

Mais un réel avantage !

- Le point de vue statistique gagnant de l'importance, le poids de cette difficulté commence à s'effacer
- On n'observe plus une protéine, mais plusieurs peptides
- La moyenne des mesures sur les peptides est probablement plus fiable qu'une mesure sur une protéine.
- **Nécessite un changement de paradigme:**
  - **Réaliser le plus de traitements de données au niveau peptidique**
  - **Ne considérer les entités protéiques qu'au moment de l'interprétation finale**

Et l’identification ?

Les approches Target-Decoy en métabo

nature methods

Published: 14 November 2016

FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry

Andrew Palmer, Prasad Phapale, Ilya Chernyavsky, Regis Lavigne, Dominik Fay, Artem Tarasov, Vitaly Kovalev, Jens Fuchser, Sergey Nikolenko, Charles Pineau, Michael Becker & Theodore Alexandrov

Nature Methods **14**, 57–60(2017) | Cite this article

nature communications

Article | Open Access | Published: 14 November 2017

Significance estimation for large scale metabolomics annotations by spectral matching

Kerstin Scheubert, Franziska Hufsky, Daniel Petras, Mingxun Wang, Louis-Félix Nothias, Kai Dührkop, Nuno Bandeira, Pieter C. Dorrestein & Sebastian Böcker

Nature Communications **8**, Article number: 1494 (2017) | Cite this article

RETURN TO ISSUE | < PREV **ARTICLE** NEXT >

Target-Decoy-Based False Discovery Rate Estimation for Large-Scale Metabolite Identification

Xusheng Wang*, Drew R. Jones, Timothy I. Shaw, Ji-Hoon Cho, Yuanyuan Wang, Haiyan Tan, Boer Xie, Suiping Zhou, Yuxin Li, and Junmin Peng*

Cite this: *J. Proteome Res.* 2018, 17, 7, 2328–2334
Publication Date: May 23, 2018
https://doi.org/10.1021/acs.jproteome.8b00019
Copyright © 2018 American Chemical Society
RIGHTS & PERMISSIONS Subscribed

Article Views	Altmetric	Citations
638	4	7

LEARN ABOUT THESE METRICS

Share Add to Export RIS

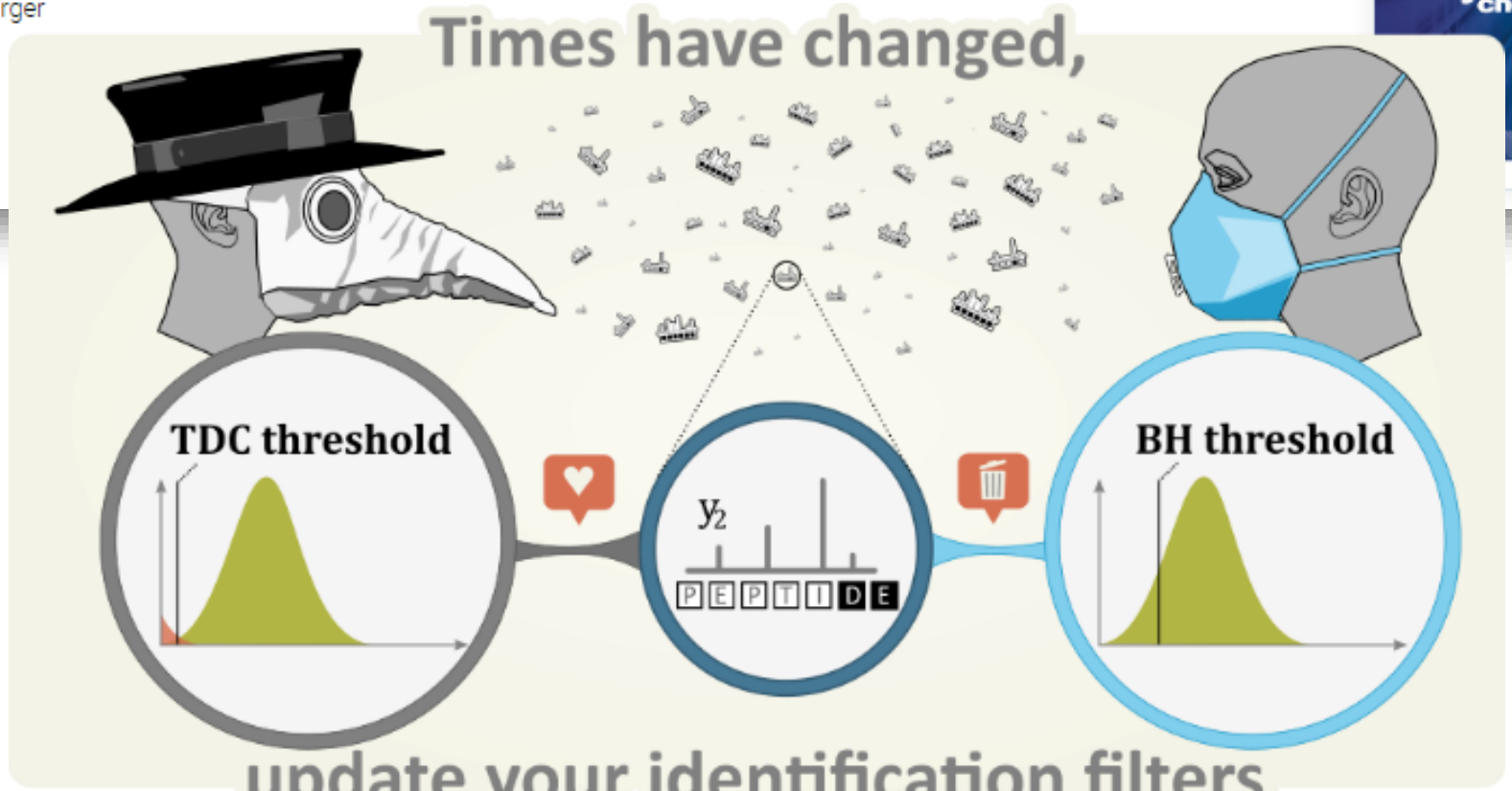
Méfiez-vous des approches Target-Decoys:
Elles ne permettent pas toujours un contrôle du FDR

RETURN TO JUST ACCEPTED MANUSCRIPTS | < PREV **ARTICLE** NEXT >

Beyond target-decoy competition: stable validation of peptide and protein identifications in mass spectrometry-based discovery proteomics

Yohann Couté, Christophe Bruley, and Thomas Burger

Cite this: *Anal. Chem.* 2020, XXXX, XXX, XXX-XXX
Publication Date: September 24, 2020
https://doi.org/10.1021/acs.analchem.0c00328



Remerciements

- Consortium **ProMetIS**
  - Beaucoup de monde…

- Développeurs et testeurs de **Prostar**
  - Les membres du laboratoire EDyP

- Plateforme **Gemeli**, projet **SYMER**
  - Audrey Le Gouellec
  - Bertrand Toussaints
  - Uwe Schlattner

- **Organisateurs** des journées
  pour leur patience et leur capacité de rebond