# UNIVERSITÉ DE STRASBOURG

*ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES*

**UMR 7178**

# THÈSE présentée par :


## Alvaro Sebastian VACA JACOME


soutenue le : **4 Juillet 2016**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Chimie Analytique

## Progress towards a better proteome characterization by quantitative mass spectrometry method development and proteogenomics

**Vers une meilleure caractérisation du protéome par le développement de méthodes de spectrométrie de masse quantitative et par l'analyse protéogénomique**

**THÈSE dirigée par :**

| | |
|---|---|
| **Dr. Alain VAN DORSSELAER** | Directeur de recherche, CNRS, Université de Strasbourg |
| **Dr. Christine CARAPITO** | Chargée de recherche, CNRS, Université de Strasbourg |
| co-encadrante | |

**RAPPORTEURS :**

| | |
|---|---|
| **Dr. Bruno DOMON** | Professeur des universités, Université de Lille |
| **Dr. Philippe MARIN** | Directeur de recherche, CNRS, Université de Montpellier |

---

**AUTRES MEMBRES DU JURY :**

| | |
|---|---|
| **Dr. Francis BITSCH** | Senior Investigator, Novartis Institutes for Biomedical Research |
| **Dr. Catherine JUSTE** | Chargée de recherche, INRA |

A mi Pa, a mi Ma, al Polo y a la Nata.

"Progress in science depends on new techniques, new discoveries and new ideas, probably in that order"- Sydney Brenner

# Acknowledgment

I would like to thank Alain Van Dorsselaer and Sarah Cianferani for their support and for giving me the opportunity to be part of the LSMBO. Thank you for the perfect environment you have created to enable people to learn, to improve and to let everyone find its place in which they can contribute to the team with all their potential.

I would like to especially thank Christine Carapito to have supervised my thesis and for all her support. Thank you for your advices, for all the discussions, and above all, for your kindness. You are not a boss; you are a leader and a colleague. You know exactly how to get the best of everyone and it was a pleasure to work with you.

I would like to thank Bruno Domon, Philippe Marin, Francis Bitsch and Catherine Juste for accepting to review and evaluate my thesis.

I would like to thank all the people that have collaborated closely with me: Thierry Rabilloud, Catherine Juste, Joël Doré, Raphaël Carapito, Nicodème Paul, Ghada Alsaleh, Seiamak Bahram, Carmella Giglione, Willy Bienvenut, Lydie Lane and Amos Bairoch.

I warmly thank the entire LSMBO for its friendly environment. Thank you for sharing all your knowledge and creating a laboratory truly rooted in solidarity that makes us grow together. And thank you for all the “pot” and BBQs that makes our waistline grow as well.

Thanks to the trio Diego, Sarah Lennon and Marine for their warm welcome to the lab and for teaching me everything that I needed to know. Thanks for always be accessible and answering all my questions.

Thanks to one of the pillars of the LSMBO, Jean-Marc, for your very frequent help on the machines and for your kindness. Thanks to Kevin for his patience when faced to my many administrative problems.

Thanks to the Informatics team: Fabrice V., Aymen, Patrick, Alex 2 and specially to Alex, thanks for all the jokes, the help and all the macros.

Thanks to “Les permanents”: Christine, Laurence, Fabrice B., Helene, Danièle and François. I cannot stress enough how much I learned from you.

Thanks to the members of my bureau. Thanks to Luc. Too bad we couldn’t continue to work together on a very ambitious project. All we needed was the samples… Thanks to Magali for your kindness and always letting us discover something new. Thanks to Charlotte for all your help. You had to do the most ungrateful tasks of my projects: PRIDE and manual validation of TMPPs! Thanks for your courage. Thanks to Georg for your expertise and excellent advices. I still want to see a live concert! Thanks to Alisson, for all your funny stories. I hope you are going to continue to dance. But forget Kizomba and “salsa cubaine” and try Salsa Porto On2!

Nina, Margaux, Marianne and Aurélie thanks for saying “Salut” all the time. I will miss it :)

Thanks to Gauthier for sharing the load of responsibility of an instrument that I will not name. Not the first nor the second or even the third one  It was a pleasure to work with you, especially for all the chocolate. You will do great! I am sure.

Thanks to Guillaume, Benoit and Johann. Thanks for speaking in “belge” all the time. I can now go to Belgium and blend in with the people. Thanks for all the moments we spent together (Izeste, Saint-Louis and New York).

Good luck to the new PHD students: Paola, Thomas and Anthony. Good luck to Maxime. You were coached by the best (and Gauthier), so I am sure you will do just fine in Maastricht

Thanks to “En Bal et Vous?”: Virginie, Tahina, Julie, Anssen, Fanny, Stéphanie et Nicodème. Thanks for letting me be part of the troupe. It was a wonderful experience and I learned a lot.

Thanks to Lucie and Dixie. Thanks for everything. Thanks for being there during my PhD. You made it a lot easier.

I would like to thank my family from the bottom of my heart. Gracias por todo el apoyo y el cariño que me han dado. Sin ustedes no podría estar aquí y hacer todo lo que quiero hacer. Siempre están conmigo y los quiero mucho. Este trabajo es fruto de su ayuda y de la inspiración que me han dado mostrándome siempre el ejemplo. Ustedes me enseñaron a luchar, a no dejarme vencer y trabajar duro para llegar a donde quiera. Es por eso que les dedico completamente este trabajo.

# Table of contents

# Part I Résumé en Français

## Chapter I Introduction Générale

### A. Les avancées dans le domaine de la protéomique quantitative

Des informations quantitatives peuvent être extraites de données shotgun acquises en mode DDA par des approches de comptage de spectres MS/MS mais la nature stochastique de la sélection des peptides fragmentés rend cette méthode sujette au sous échantillonnage, à des valeurs manquantes, à un manque de reproductibilité et donc à une quantification aujourd'hui considérée comme approximative [1].

Les avancées technologiques extraordinaires de ces dernières années en spectrométrie de masse permettent de profiter d'un vaste panel de modes d'acquisition hautement performants pour l'analyse quantitative de protéines. Trois approches basées sur des modes d'acquisition différents peuvent être distinguées: l'extraction de signaux chromatographiques à partir de données acquises en mode DDA, le suivi de peptides prédéfinis par des approches ciblées, et une approche récemment introduite basée sur l'acquisition des données en mode Data-Independent Acquisition (DIA).

L'extraction de pics chromatographiques pour l'ensemble des peptides identifiés, voire détectés, à partir de données DDA permet d'obtenir des résultats quantitatifs plus fiables et reproductibles que des approches par comptage de spectres mais est plus complexe à mettre en œuvre et les outils logiciels existants doivent encore être améliorés.

Les approches ciblées permettent de doser de manière précise et sensible des dizaines, voire centaines de protéines prédéfinies dans un mélange contenant plusieurs milliers de protéines [2]. Le choix des protéines ciblées est fondé sur la vérification d'hypothèses et/ou la découverte préalable de cibles par des approches globales. La méthode de référence en analyse ciblée est l'analyse en mode Selected Reaction Monitoring (SRM) couplée à l'utilisation de la dilution isotopique. Cette méthode utilise des appareils de type triple quadripôle. La détection de plusieurs signaux d'un même ion parent permet une quantification très précise grâce à la mesure simultanée de peptides de référence marqués aux isotopes stables, identiques aux peptides ciblés et introduits en quantité connue. Des instruments à haute-résolution peuvent aussi être utilisés en mode d'analyse ciblée pour obtenir une sélectivité plus grande. La méthode est alors communément appelée Parallel Reaction Monitoring (PRM) [3, 4]. Les approches ciblées sont souvent utilisées dans des étapes de vérification de candidats biomarqueurs, notamment dans des fluides biologiques, et de très nombreux paramètres relatifs à la préparation d'échantillon et à l'instrumentation sont à optimiser pour atteindre les sélectivités et les sensibilités nécessaires pour la quantification des protéines ciblées dans des mélanges aussi complexes que du plasma ou de l'urine. L'optimisation de méthodes ciblées est longue mais des méthodes bien optimisées ont montré les meilleures sensibilités et les limites de détection atteintes sont encore inégalées par des approches globales aujourd'hui [5]. Cependant, la vulnérabilité de cette méthode est le faible nombre de protéines quantifiées relativement au grand nombre de protéines généralement présentes dans un protéome complexe.

Le mode d'acquisition DIA récemment introduit promet de combiner les avantages d'une analyse sans *a priori* de type DDA à la reproductibilité, la sensibilité et la justesse des méthodes de quantification ciblée [6]. En DIA, le spectromètre de masse génère des spectres MS/MS de tous les peptides isolés dans une grande fenêtre de masse

prédéfinie (25 Da en général). Les spectres MS/MS sont donc des spectres multiplexés de tous les fragments des peptides co-elués et co-isolés à un temps donné. Ce mode d'acquisition permet, en théorie, de cartographier les fragments de tous les peptides d'un échantillon complexe [7].

Devant ce choix, le travail du protéomiste consiste à définir la stratégie la plus appropriée, à développer et optimiser des méthodes de quantification capables de fournir la sensibilité, la justesse, la sélectivité et la couverture nécessaires afin de répondre au mieux au questionnement biologique posé.

### B. L'analyse protéogénomique

La protéogénomique est un nouveau domaine de recherche à l'intersection entre la génomique, la transcriptomique et la protéomique. Ce domaine avait initialement été introduit avec comme objectif d'utiliser des données protéiques pour améliorer les annotations génomiques [8]. Sa définition a depuis été étendue à la découverte de nouveaux peptides liés à des variants de séquence d'acides aminés, des variants d'épissage, d'édition de l'ARN, d'un nouveau cadre de lecture ouvert codant et d'autres évènements biochimiques qui sont finalement traduits jusqu'au niveau protéique. L'identification de ces nouveaux peptides est d'un intérêt majeur puisqu'ils pourraient être directement utilisés comme biomarqueurs diagnostiques ou pronostiques d'une pathologie donnée [9].

Des modifications pré- ou post-traductionnelles peuvent aussi donner lieu à la présence de biomolécules dont la fonction diffère d'autres biomolécules provenant d'un même gène et conduire à un changement de la fonction de la protéine. Dans ce contexte, le début d'une protéine, sa partie N-terminale, est une caractéristique importante qui va définir sa stabilité, sa localisation et sa fonction dans la cellule. De nombreuses méthodes spécifiques existent pour analyser spécifiquement les parties N-terminales des protéines et ces méthodologies entrent parfaitement dans le cadre de la protéogénomique [10].

Enfin, est intégré dans le champ des stratégies de protéogénomique l'ensemble des stratégies qui viseront à tirer avantage de l'utilisation conjointe des données produites par différentes approches omiques sur des échantillons donnés. Ces approches seront optimales dans les cas où des données multi-omiques pourront être acquises sur un même échantillon [11].

Les principaux verrous de ces approches de protéogénomique résident dans les outils bioinformatiques partiellement développés ou encore manquants pour permettre leur mise en œuvre.

## Chapter II Développements méthodologiques en analyse protéomique quantitative

### A. Optimisation du workflow de développement de méthodes de quantification ciblée

Lors de mon travail de thèse j'ai participé à l'optimisation du workflow pour le développement de méthodes ciblées par LC-SRM. Ceci a été fait en déterminant les paramètres clés à optimiser afin d'augmenter la sensibilité et la spécificité des méthodes. La Figure IV-1 montre le workflow pour le développement des méthodes SRM que j'ai optimisé, qui commence par le développement de la méthode de préparation des échantillons qui doit être adaptée et optimisée à chaque nouveau type d'échantillon. En parallèle, le développement d'une méthode spécifique à chaque protéine ciblée est fait. Cela commence par le choix des peptides signatures pour chaque protéine. Ces peptides doivent avoir des propriétés physicochimiques particulières pour être visibles en spectrométrie de masse, ne pas être sujets à des modifications indésirables et être uniques à la protéine. L'utilisation de peptides marqués isotopiquement

permet l'optimisation des paramètres expérimentaux (énergie de collision, temps de rétention). La partie analytique a aussi été améliorée grâce à la mise en place et l'utilisation en routine d'échantillons standards, avant et pendant une série d'analyses, pour vérifier les performances et le bon fonctionnement des appareils du couplage LC-SRM.

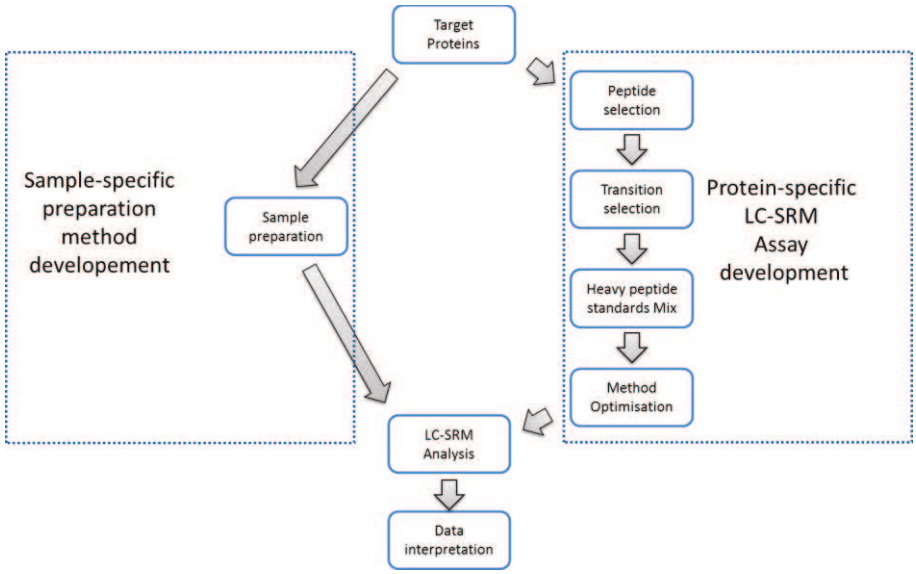



**Figure I-1 : Workflow de développement de méthodes de quantification ciblée.**

## B. Evaluation de la compatibilité d'une méthode de séparation de protéine (SDS-PAGE) avec une quantification ciblée par LC-SRM

Mon travail de thèse a également consisté à optimiser et développer des méthodes de préparation d'échantillons compatibles avec des études quantitatives. Une première étude a eu pour but de mesurer les performances analytiques du couplage d'une méthode de fractionnement de protéines par gel d'électrophorèse (SDS-PAGE) avec une méthode de quantification ciblée par LC-SRM. Cette méthode permet de réduire la complexité d'un échantillon et ainsi de réduire la gamme dynamique de chaque fraction analysée, augmentant ainsi la spécificité de l'analyse. Cependant cette méthode requiert un temps d'analyse conséquent et n'est donc pas compatible avec une quantification à haut débit. De plus, l'interprétation de données d'échantillons fractionnés n'est pas triviale.

Afin d'obtenir une méthode de préparation d'échantillons compatible avec des méthodes de quantification, j'ai évalué la compatibilité de l'utilisation du gel stacking SDS-PAGE en amont d'une stratégie de quantification. Cette stratégie analytique permet de bénéficier de l'excellente capacité du SDS à extraire et solubiliser des protéines. Cette méthode a également l'avantage de ne pas introduire de fractionnement et est donc plus compatible avec des études quantitatives. Après avoir optimisé les paramètres clés pour améliorer le protocole, nos résultats ont montré que cette méthode donne des résultats quantitatifs très satisfaisants (Figure I-2). Les performances analytiques trouvées sont mêmes supérieures à celles obtenues par une méthode de digestion liquide.

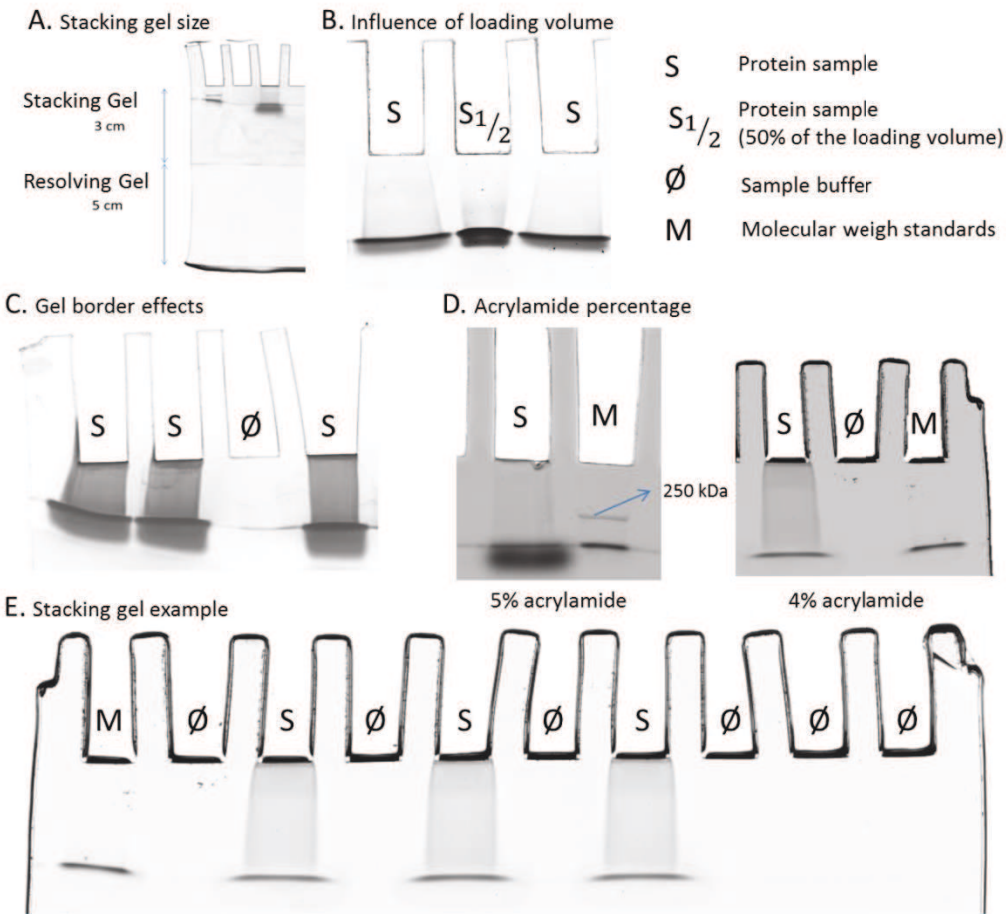



**Figure I-2 : Développement d'un protocole de purification de protéines par Gel Stacking SDS-PAGE.**
Plusieurs paramètres clés ont été optimisés pour améliorer les performances quantitatives de cette approche de préparation d'échantillons : la taille des gels stackings (A), l'influence des volumes d'échantillons déposés (B), l'influence des effets de bord (C) et le pourcentage d'acrylamide (D). Un exemple d'un gel stacking optimisé est montré (E).

## C. Mise en place de stratégies d'évaluation des performances analytiques des couplages LC-MS et des outils bioinformatiques dédiés à la quantification

Devant le nombre important de méthodes d'identification et de quantification du protéome, le travail du protéomiste consiste à définir la stratégie la plus appropriée, à l'adapter et à l'optimiser pour répondre au mieux au questionnement biologique posé. Afin d'obtenir l'analyse la plus robuste et la plus juste, les performances instrumentales doivent être monitorées très fréquemment.

Pour évaluer avec précision et objectivité une étape ou l'ensemble du workflow de l'analyse protéomique, un échantillon standard bien caractérisé doit être utilisé. Cet échantillon doit être conçu pour pouvoir définir une valeur vraie avec laquelle les résultats des évaluations seront comparés.

L'échantillon standard développé est constitué de d'un digestat de levure dans lequel sont rajoutés :

- Des peptides tryptiques des protéines d'un mélange de 48 protéines humaines purifiées (Universal Protein Standard, UPS1, Sigma),
- Leurs peptides homologues lourds isotopiquement marqués,
- 11 peptides standards de temps de rétention (iRT standard peptides, Biognosys).

Cet outil s'est révélé extrêmement utile pour évaluer et améliorer les workflows d'analyse tels que le développement d'un test de performance sensible et objectif pour les plateformes LC-SRM, l'évaluation des pipelines bioinformatiques

pour l'analyse quantitative Label-free et l'amélioration de l'extraction de signal à partir de données acquises en mode DIA.

Afin d'évaluer les performances des couplages LC-SRM du laboratoire, un test de performance sensible et objectif a été développé. Les protéines UPS1 ont été rajoutées à une concentration finale de 2,5 fmol/µl dans 1 µl de digestat de levure à 500 ng/µl. Ce test de performance a été utilisé pour évaluer régulièrement les performances des instruments au cours d'une année, et a été injecté au moins 4 fois par mois. La Figure I-3 montre le résultat d'un test de performance. Des critères individuels sur les caractéristiques chromatographiques (intensité, aire, largeur à mi-hauteur, temps de rétention) de certains ions ont été mis en place. Egalement, des critères globaux ont été mis en place pour déceler des perturbations dans le système LC-SRM. Le chromatogramme a été découpé en trois parties (la zone hydrophile en début du gradient, le milieu du gradient et la zone hydrophobe en fin de gradient). Une déviation des performances dans une des zones permet de diriger la stratégie la plus adapté pour le dépannage.

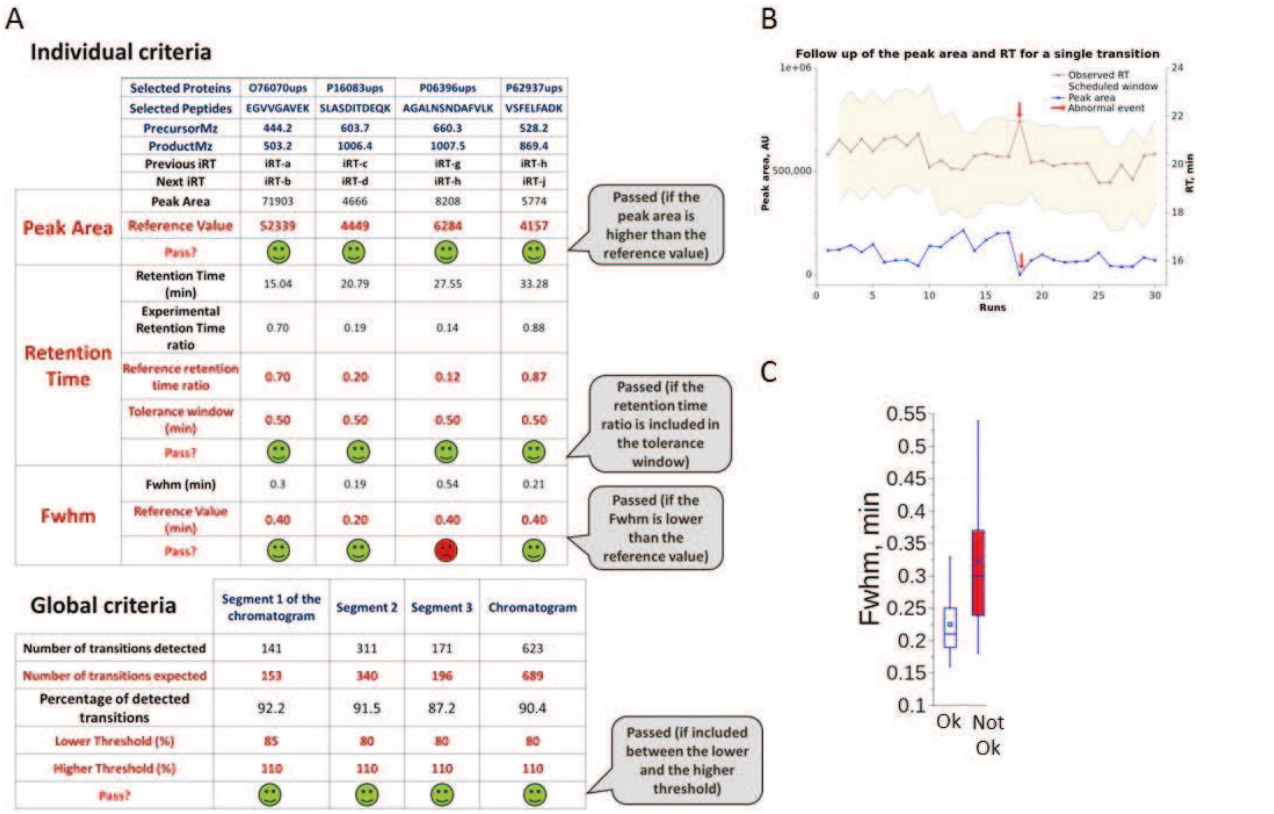



**Figure I-3 : Test de performance appliqué à l'évaluation de performances de plateformes LC-SRM.**
A. Les critères individuels et globaux avec leurs critères d'acceptation correspondants sont listés en rouge dans le tableau. Les valeurs extraites d'une analyse sont en noir. Ce test de performances est sensible, objectif et permet de détecter des perturbations dans le système LC-SRM et diriger la stratégie la plus adapté pour le dépannage. B. Suivi des temps de rétention et des aires des pics d'un peptide cible pendant une année. C. Comparaison des largeurs à mi-hauteurs d'un système performant et d'un système non-performant.

Une autre stratégie pour l'évaluation des plateformes instrumentales a été développée. Elle est basée sur l'utilisation de peptides isotopologues, c'est-à-dire de peptides de même séquence primaire mais constitués d'acides aminés isotopiquement marqués qui leur confèrent des masses différentes (Figure I-4). Cette stratégie repose sur le fait que ces peptides ont les mêmes caractéristiques physico-chimiques donc répondent de la même manière en chromatographie et en spectrométrie de masse. Ces peptides ont été mélangés à des concentrations différentes afin de créer des droites de calibration couvrant une large gamme de concentration (5,3 log). Cette stratégie permet

d'évaluer rapidement et de manière fiable la gamme dynamique, la sensibilité et les limites de quantification de différentes plateformes instrumentales. Des travaux ont montré la preuve de concept pour évaluer des plateformes de LC-MS par l'utilisation de peptides isotopologues [12, 13]. Nous avons étendu ce concept et nous l'avons appliqué à différentes plateformes de LC-MS. Nous avons aussi évalué les effets de différents paramétrages d'une méthode en mode DIA sur la sensibilité de l'appareil lors de l'analyse de mélanges simples et complexes. Enfin, nous avons utilisé cette stratégie pour confirmer que des systèmes chromatographiques en débit capillaire(5 µl/min) permettent d'atteindre une sensibilité similaire voire meilleure que celle obtenue en débit nano (450 nl/min) tout en apportant plus de robustesse et confort à l'utilisateur [14]. Une publication présentant ces résultats est en cours de préparation.

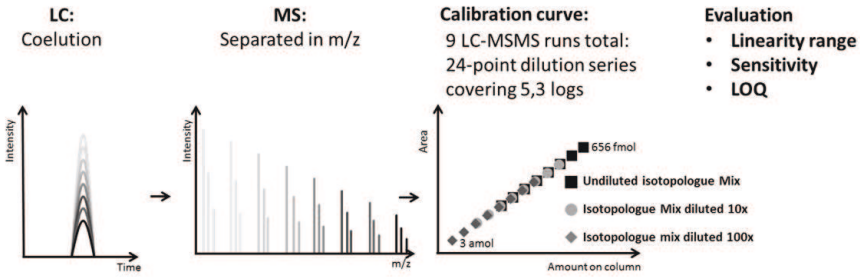



**Figure I-4 : Utilisation de peptides isotopologues pour l'évaluation de performances de plateformes de LC-MS/MS.**
Un mélange de peptides isotopologues ayant le même comportement chromatographique mais séparés en spectrométrie de masse ont été utilisés pour créer des droites de calibration. Ces droites, constituées de 24 points allant de 3 amol à 656 fmol et couvrant une gamme de 5,3 log, ont été utilisées pour évaluer la gamme dynamique, la sensibilité et les limites de quantification de différents instruments.

## D. Optimisation et développement de méthodes pour l'analyse en mode DIA

La DIA est souvent présentée comme une méthode standardisée de type « plug-and-play» utilisant un unique set de paramètres pour quantifier des protéines indépendamment du type d'échantillon. Cependant, pour les méthodes de DIA basées sur l'utilisation de fenêtres d'isolement successives (type SWATH [6]), il est important de paramétrer correctement l'instrument afin d'obtenir la meilleure sélectivité, sensibilité, précision de quantification et la plus grande couverture du protéome.

Gillet *et al.* a proposé une approche pour l'analyse de données DIA, initialement appliquée pour les données de type SWATH, nommée extraction ciblée de données [6]. L'identification et la quantification d'un peptide sont effectuées en utilisant des informations obtenues au préalable par des approches DDA (stockées dans une librairie spectrale). Les traces des ions fragments sont ensuite extraites pour des peptides d'intérêt dans les données DIA. La qualité des données est évaluée et un score est attribué afin de pouvoir valider l'identification du peptide. Cette approche centrée sur les peptides (par opposition aux approches centrées sur les spectres qui utilisent des algorithmes de recherche de banque de données pour identifier des protéines) utilise les caractéristiques chromatographiques des signaux extraits afin de vérifier l'identification des peptides ciblés. Comme pour la SRM, les paramètres de validation de l'identification du peptide sont la co-élution d'ions fragments, la forme des pics, les intensités relatives et le temps de rétention. D'autres mesures supplémentaires peuvent également être utilisées lors de l'analyse de données DIA comme la précision de la masse, la co-élution des ions précurseurs et des ions fragments et la co-élution des différents états de charge d'un même peptide. Cette approche confère aux données DIA une structure ressemblant à celle obtenue en SRM. Comme pour la SRM, les données sont complètes pour les peptides ciblés et aucune valeur manquante n'est présente dans les données. Cependant les données DIA sont très bruiteuses, ce qui rend très difficile l'identification des pics à intégrer.

Deux outils logiciels ont été évalués en utilisant un échantillon standard bien caractérisé composé d'un mélange de 48 protéines humaines purifiées (Universal Protein Standard, UPS1, Sigma) rajouté à un digestat de levure. Deux points de concentration ont été utilisés pour l'évaluation : 5 et 25 fmol d'UPS1 dans 1 µg de lysat de levure. Les deux outils logiciels évalués sont PeakView (AB Sciex) et Skyline [15]. Les résultats sont décrits dans la Figure I-5. La liste de peptides pour lesquels un signal a été extrait est la même entre les deux outils logiciels. Pour Skyline, le choix et l'identification des pics chromatographiques à intégrer ont été dirigés par l'algorithme mProphet [16]. Le taux de faux positifs (False Discovery Proportion, FDP), le taux de vrais positifs (True Positive Rate, TPR) et la justesse des rapports de l'abondance des protéines entre les deux échantillons ont été utilisés comme métriques pour évaluer ces logiciels.

Les deux logiciels arrivent à discriminer les protéines d'UPS1 et ceux de la levure. Toutefois, les protéines UPS1 semblent être plus proches du facteur de variation attendu (Fold change = 5) en utilisant PeakView. Cependant PeakView est un logiciel propriétaire et n'est compatible qu'avec des données obtenues par un instrument de la marque AB Sciex. Même si Skyline a de moins bonnes performances, il reste un logiciel open-source compatible avec des données brutes provenant de différents instruments et dont les principes de fonctionnement sont connus et donc peuvent être optimisés.

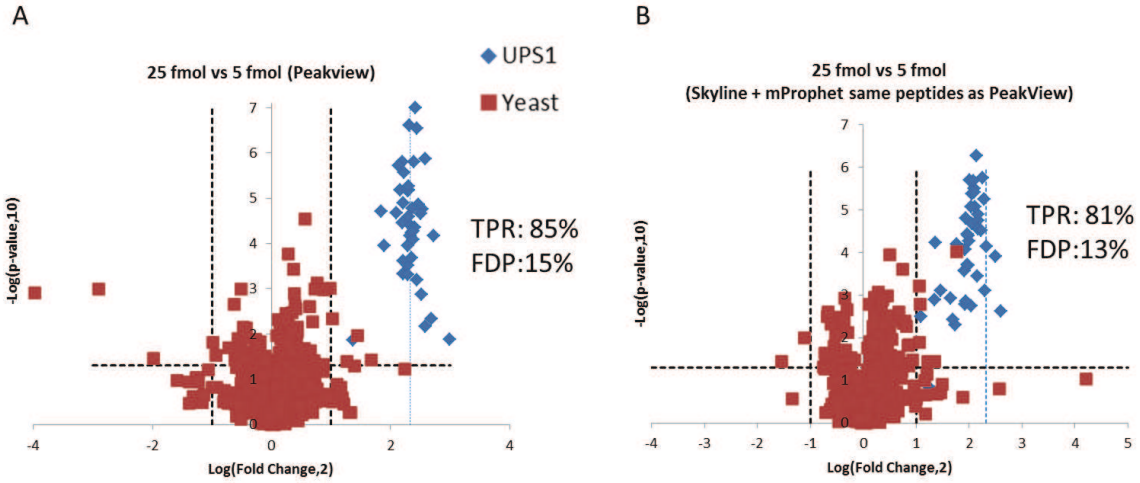



**Figure I-5 : Evaluation de deux outils logiciels pour l'analyse des données DIA.**

En raison de leur nature, l'analyse de données acquises en mode DIA est compliquée. La complexité des spectres MS/MS et le nombre élevé de peptides à quantifier (dizaines de milliers) rendent la validation visuelle et manuelle des données très difficile et peu pratique. Les pipelines de validation automatique (mProphet [16]) peuvent être utilisés, mais pour le moment il y a encore une marge de progression importante pour le traitement des données DIA. Les problèmes majeurs restent le choix automatisé des pics chromatographiques à intégrer et l'alignement des temps de rétention entre analyses.

## E. Application à la validation de biomarqueurs de la maladie de Crohn

Ce projet a été réalisé en collaboration avec l'unité MICALIS de l'Institut National de la Recherche Agronomique (INRA), et en particulier avec Catherine Juste et Joël Doré.

La maladie de Crohn est un type de maladie inflammatoire chronique de l'intestin (MICI) caractérisée par une inflammation chronique et récurrente des segments intestinaux et peut potentiellement être accompagnée par des

manifestations extra-intestinales [17]. Cette maladie touche environ 0,32% de la population en Europe et en Amérique du Nord. Elle est plus fréquente dans les pays développés et moins fréquente en Asie et en Afrique [18].

Il n'existe pas de cure pour la maladie de Crohn. Des médicaments et la chirurgie sont utilisés pour soulager les symptômes, maintenir la rémission, et prévenir les rechutes [19]. Le diagnostic de cette maladie est très difficile car il n'y a pas de symptômes spécifiques de la maladie et ses manifestations sont communes avec d'autres pathologies telles que la gastro-entérite, la rectocolite hémorragique et le syndrome de l'intestin irritable. Il est très important de pouvoir distinguer l'ensemble de ces pathologies car chacune nécessite un traitement thérapeutique particulier. Le diagnostic de la maladie de Crohn est basé sur un ensemble d'arguments cliniques qui prend du temps pour être rassemblé. Le temps moyen pour poser le bon diagnostic pour la maladie de Crohn est d'environ 2,6 ans. Pour l'instant, aucun biomarqueur moléculaire spécifique à la maladie de Crohn n'a atteint le stade de l'usage clinique.

Dans ce contexte, une méthode LC-SRM a été développée pour la validation de treize protéines du microbiote intestinal marqueurs de la maladie de Crohn.

Le développement de cette méthode LC-SRM pour des protéines microbiennes a été compliqué à cause de la nature de l'échantillon. En effet, le microbiome intestinal humain est un échantillon d'une extrême complexité, l'existence de plus de 9,8 millions de gènes différents a récemment été confirmée [20]. Ceci représente 445 fois plus de gènes que dans le génome humain complet. En outre, il y a une grande diversité dans la composition microbienne entre individus. La Figure I-6 résume le workflow analytique mis en place pour ce projet. Une cohorte de patients sains et malades a été analysée. La fraction microbienne a été extraite à partir des échantillons de selles fraiches grâce à un gradient de densité sous conditions inertes. Le protocole de purification de protéines par Gel Stacking SDS-PAGE a été employé. Cette étude a permis de confirmer grâce à une méthode ciblée les tendances de sous- et surexpressions de protéines microbiennes préalablement identifiées lors d'une étude de découverte par 2D-DIGE (Figure I-7).

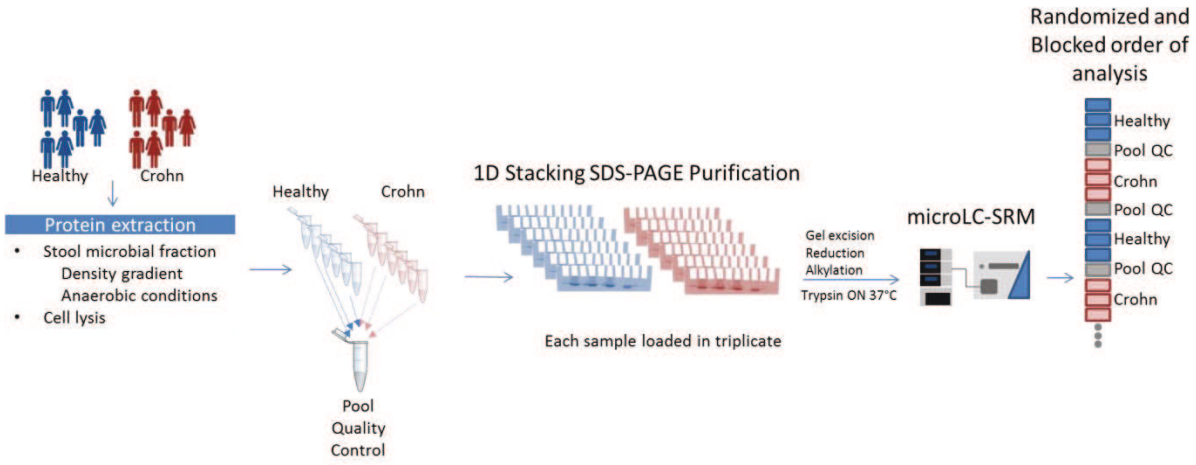



**Figure I-6 : Workflow analytique pour la validation de protéines microbiennes par LC-SRM.**

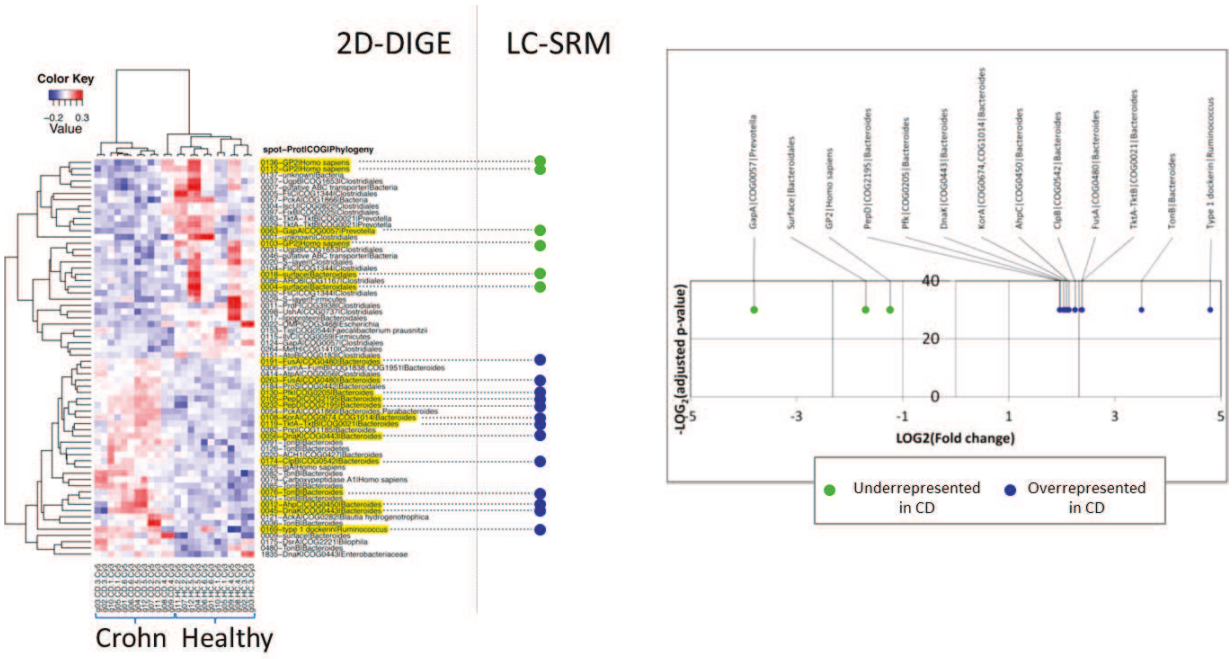



**Figure I-7 : Résultats de l'analyse par LC-SRM de 13 protéines microbiennes liées à la maladie de Crohn.**
Heat map des résultats de l'étude de 2D-DIGE montrant les spots de gels avec un changement significatif de l'abondance de protéines entre des patients sains et malades. Les protéines en jaune ont été choisies pour être validées par LC-SRM. Les tendances observées lors de l'étude de quantification par 2D-DIGE ont été validées par LC-SRM.

Cette étude apporte la première preuve que des protéines bactériennes du microbiome intestinal humain peuvent être liées à la maladie de Crohn [21]. La quantification de ces protéines cibles a été faite sans fractionnement dans un milieu d'une extrême complexité. Ce travail a été valorisé par un dépôt de brevet et une publication.

## F. Applications à la quantification relative et absolue de méthionines aminopeptidases

Ce projet a été réalisé en collaboration avec l'Institut de biologie intégrative de la cellule (I2BC) à Gif-Sur-Yvette, et en particulier avec Frédéric Frottin, Willy Bienvenut, Thierry Meinnel et Carmela Giglione.

Les méthionines aminopeptidases (MetAP) sont en charge de l'excision de la méthionine N-terminale. Ce processus biologique a une grande importance dans la cellule illustrée par le fait qu'il est un processus hautement conservé entre les organismes. La plupart des protéines sont synthétisées avec une méthionine sur le premier résidu. Toutefois, pour deux tiers des protéines cette méthionine est éliminée par la suite. La raison exacte de ce processus est mal connue. Ce processus est supposé contrôler la stabilité de la protéine et sa demi-vie [22]. Dans les cellules eucaryotes, il existe deux classes de méthionine aminopeptidases, MetAP1 et MetAP2. Frottin *et al.* a montré que ces deux protéines ont une spécificité de substrat *in vitro* très similaires et qu'elles sont interchangeables dans les plantes [23, 24]. Ces enzymes sont très régulées à différents stades de la vie de la cellule ou lorsque celle-ci est soumise à des conditions de stress [22, 25]. Toutefois, il n'est pas encore bien compris comment la régulation des MetAPs affecte le protéome.

Une façon d'étudier les rôles respectifs des MetAP1 et MetAP2, et d'en connaitre plus sur le rôle de l'excision de la méthionine N-terminale dans la cellule, consiste à utiliser des médicaments ciblant spécifiquement la MetAP2. La fumagilline est un médicament qui se lie et inhibe MetAP2 mais n'a aucune incidence sur l'activité de MetAP1 [23, 26]. La fumagilline et ses dérivés, provoque un arrêt du cycle cellulaire dans les cellules endothéliales et dans plusieurs lignées cellulaires cancéreuses. Ceci suggère que MetAP2 pourrait être une cible pour la thérapie contre certains

cancers. Cependant, les dérivés de ce médicament ont été montrés comme provoquant une neurotoxicité dans des essais cliniques de phase III [27].

Nos collaborateurs ont identifié des lignées cellulaires avec des sensibilités différentes à la fumagilline:

- des lignées cellulaires hautement sensibles, à savoir ayant une prolifération faible lorsqu'elles sont exposées à la drogue : HUVEC, U87, U937 et A549.
- des lignées cellulaires insensible : Jurkat, HCT116 et K562.

Avec cette information, ils ont caractérisé les protéomes et les profils N-terminomiques de plusieurs lignées de cellules pour identifier les variations possibles au niveau de la protéine qui pourraient expliquer la différence de sensibilité à la fumagilline. De cette étude, ils ont conclu que les variations spécifiques des protéomes n'expliquent pas la sélectivité dans le phénotype. Par ailleurs, les MetAP1 et MetAP2 n'ont pas pu être identifiées. Par conséquent, pour vérifier si la différence d'abondance de ces deux protéines peut expliquer la différence de sensibilité de différentes lignées cellulaires, des méthodes de quantification par immunodétection ont été développées en utilisant plusieurs anticorps commerciaux. Cependant aucune des MetAPs n'a pu être détectée ce qui suggère que ces protéines sont présentes à des niveaux très faibles dans les lignées cellulaires.

Nous avons donc développé une méthode de quantification ciblée par LC-SRM pour la quantification relative des deux protéines MetAP1 et MetAP2 dans un premier temps, et la quantification absolue pour la protéine MetAP2 dans un deuxième temps. Trois lignées cellulaires ont été choisies pour être analysées : la lignée cellulaire la plus sensible parmi les tissus cancéreux (U87), une lignée sensible parmi les cellules endothéliales (HUVEC) et une lignée de cellules insensibles à la fumagilline (K562). Dans cette étude nous avons montré que la sensibilité des cellules à cette drogue est bien corrélée à l'abondance de MetAP2 dans la cellule (Figure I-8). Ce résultat a été obtenu par le développement d'une méthode de quantification relative utilisant des peptides isotopiquement marqués, et confirmé par une méthode de quantification absolue s'appuyant sur des courbes de calibration pour un peptide signature de la protéine de MetAP2. Une publication résumant ces résultats est en cours de révision.

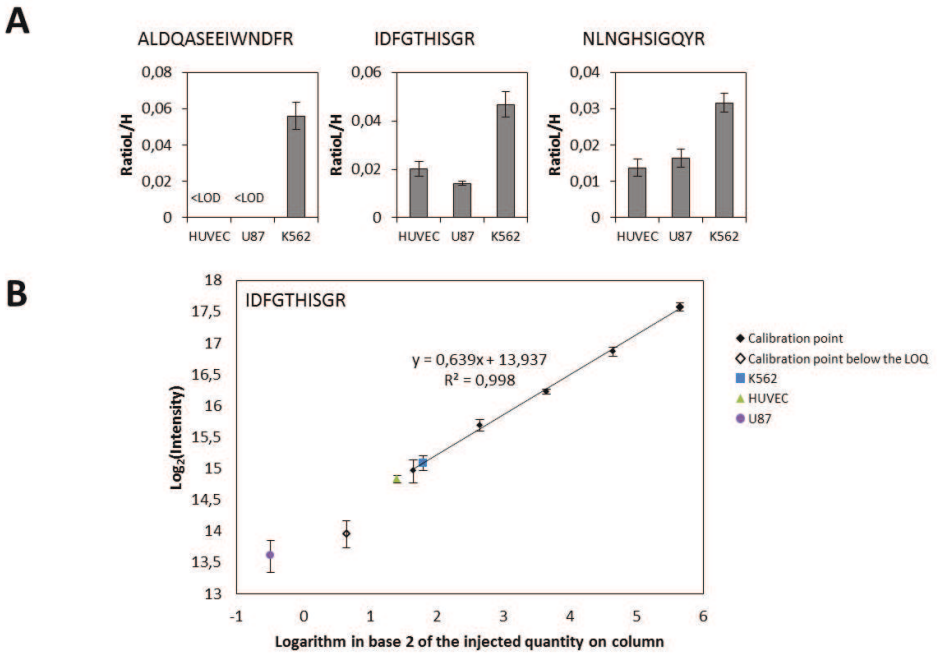



**Figure I-8 : Résultats de la quantification relative et absolue pour la protéine MetAP2 dans trois lignées cellulaires.**
A. Ratios de la somme des aires de toutes les transitions des versions légères et lourdes des peptides ciblées (L/H) pour les peptides de MetAP2 dans trois lignées cellulaires. Chaque analyse a été répliquée 4 fois. B. Droite de calibration pour la quantification absolue de METAP2 en suivant le peptide IDFGTHISGR. La droite de calibration a été faite en utilisant les points avec des CVs inférieurs à 15% et une exactitude entre 80% et 120% (losanges remplis), le dernier point respectant ces critères est la limite de quantification (LOQ). Les points ne respectant pas ces limites n'ont pas été utilisés pour générer la droite de calibration (losanges vides).

# Chapter III Développements méthodologiques en analyse protéogénomique

Dans le vaste champ couvert par les approches protéogénomiques, mes travaux de thèse ont porté dans deux directions :

## A. Développement d'une approche de caractérisation du N-terminome

Ce projet a été réalisé en collaboration avec l'Institut de Biosciences et Biotechnologies de Grenoble et en particulier avec Thierry Rabilloud.

Ce projet a aussi été réalisé en collaboration avec le groupe CALIPHO (Computer Analysis and Laboratory Investigation of Proteins of Human Origin) Group, de l'institut Suisse de Bioinformatique (SIB), et en particulier avec Lydie Lane et Amos Bairoch, pour l'intégration des résultats dans la banque de données UniProtKB/Swissprot.

Dans le contexte d'améliorer la caractérisation du protéome, j'ai participé au développement d'une approche pour l'identification des parties N-terminale des protéines. Cette méthode est basée sur la derivatisation spécifique des parties N-terminales par le triméthoxyphényl phosphonium (TMPP) (Figure I-9) [28].

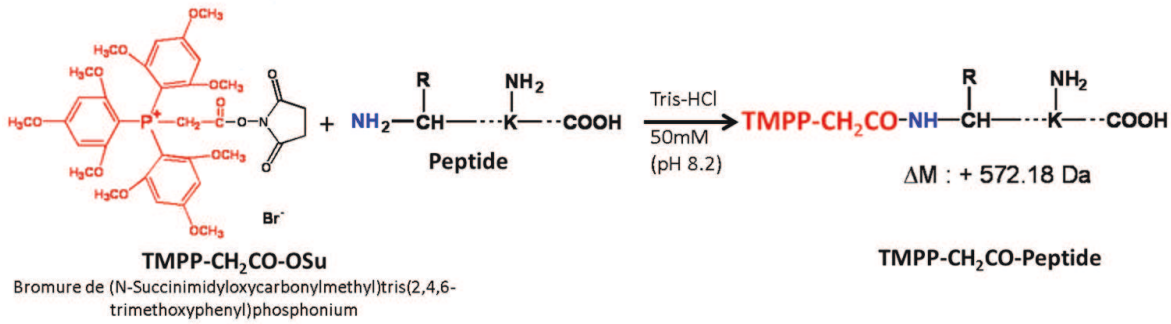



**Figure I-9 : Marquage spécifique des parties N-terminales des protéines par le TMPP.**

Dans ce projet j'ai développé un workflow analytique et bioinformatique pour la validation automatique et fiable des peptides marqués au TMPP. Cette étape finale de validation des peptides marqués était jusqu'ici réalisée manuellement et constituait le frein à l'application de cette méthode car trop compliquée et sujette à des erreurs subjectives d'interprétation. A présent ce workflow permet l'analyse haut-débit des parties N-terminales de protéines dans des mélanges complexes et l'étape de validation peut être faite en quelques minutes.

J'ai appliqué cette stratégie pour caractériser le N-terminome mitochondrial humain, et des fractions enrichies en mitochondries de cellules humaines en culture ont été utilisées. Les protéines sont extraites et marquées sur leur partie N-terminale libre au TMPP léger ($^{12}$C-TMPP-Ac-OSu) et lourd ($^{13}C_9$-TMPP-Ac-OSu). Les échantillons sont ensuite purifiés par Gel Stacking SDS-PAGE ou fractionnés par gel SDS-PAGE (Figure I-10). Les bandes de gel sont coupées, réduites, alkylées et digérées enzymatiquement pendant une nuit. Les peptides résultants sont extraits et analysés par nanoLC-MS / MS.

Une stratégie utilisant une série de deux recherches dans des banques de données a été mise en œuvre (Figure I-10). La première recherche est une recherche avec les paramètres classique de l'analyse protéomique pour identifier des peptides internes. Les spectres de bonne qualité sont extraits de l'ensemble des spectres non attribués lors de la première recherche. Les peptides TMPP marqués sont ensuite recherchés sur ce plus petit jeu de données et validés si et seulement si l'identification d'une séquence peptidique marquée au TMPP léger et lourd est confirmée, si la paire de peptides porte les mêmes modifications et si l'identification de cette paire a été réalisée avec des temps de rétention proche (< 30 s). Les positions N-terminales ne sont validées que si le spectre utilisé permet d'obtenir une identification non-ambiguë de la séquence peptidique. Egalement, la séquence doit être unique dans la banque de données recherchée pour que la position exacte du marquage en position N-terminale soit bien confirmée (Figure I-10).

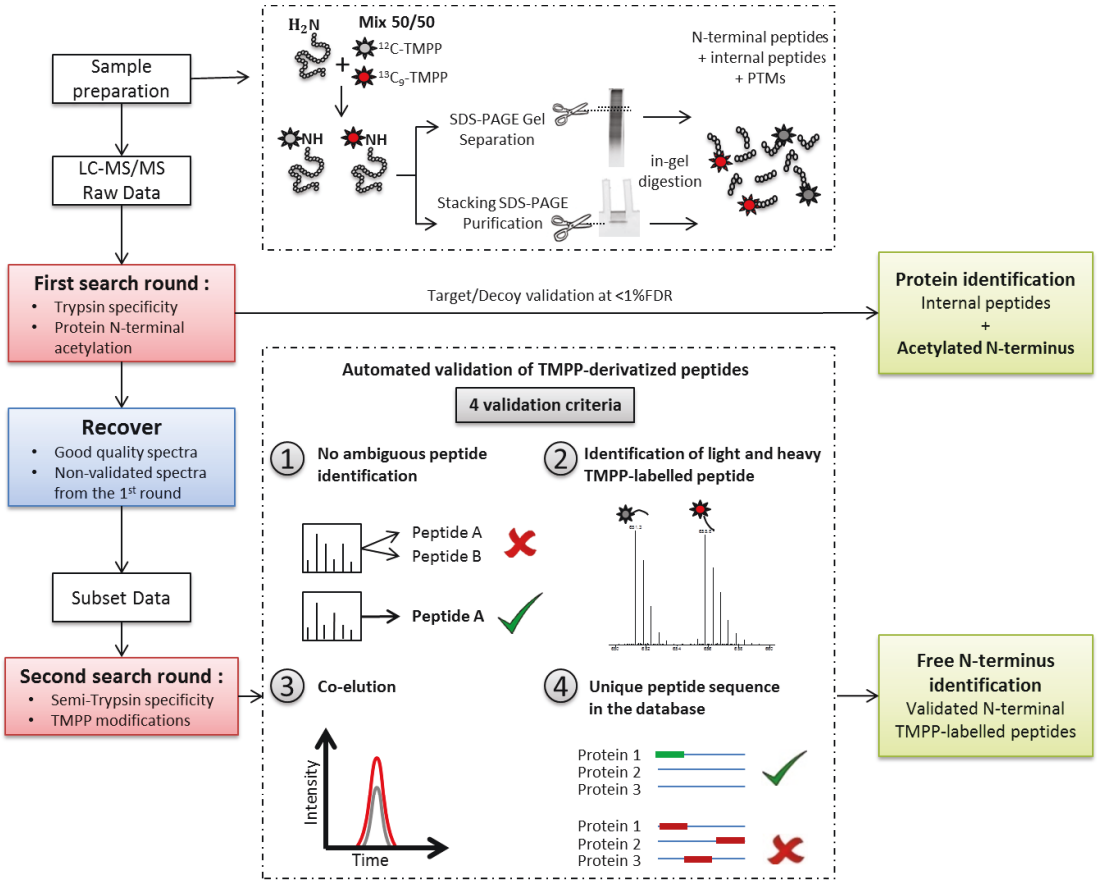



**Figure I-10 : Vue d'ensemble de la stratégie analytique pour la préparation d'échantillons et pour la validation des données.**

J'ai appliqué cette stratégie comme preuve de principe à la correction des annotations des codons d'initiation d'un génome bactérien *Herminiimonas arsenicoxydans* [28]. J'ai ensuite appliqué ces développements à la caractérisation du protéome mitochondrial humain. Nous avons ainsi pu établir un catalogue de plus de 4600 protéines dont 963 sont mitochondriales et obtenir l'identification de la partie N-terminale pour 35% de l'ensemble des protéines identifiées [29]. Ce haut pourcentage a pu être atteint grâce au marquage TMPP qui permet d'augmenter la sélectivité et la sensibilité des peptides marqués. Les résultats de ce projet ont servi à améliorer les annotations protéiques (acétylation en position N-terminales, sites de clivages de peptide signaux et transit…) dans la banque UniProtKB/SwissProt et ont fait l'objet de trois publications (dont une est en cours de soumission).

## B. Développement d'une stratégie de recherche en banques personnalisées grâce à l'utilisation conjointe de données de génomique, transcriptomique et protéomique acquises sur les mêmes échantillons

Ce projet a été réalisé en collaboration avec la Plateforme GENOMAX du Laboratoire d'ImmunoRhumatologie Moléculaire de l'Université de Strasbourg, et en particulier Raphaël Carapito, Nicodème Paul, Ghada Alsaleh, Louise Ott et Seiamak Bahram.

J'ai développé une stratégie permettant de construire des banques de séquences protéiques personnalisées, stratégie qui s'inscrit parfaitement dans les objectifs fixés par la protéogénomique. J'ai pu réaliser cela dans le contexte d'un projet biologique dans lequel des études du génome (par séquençage d'exome) et du transcriptome (par RNASeq) ont été menées sur les mêmes échantillons en parallèle des analyses protéomiques (Figure I-11). Cette étude multi-

omique avait pour but d'étudier des membres d'une famille atteinte de fièvre récurrente avec hyper-IgD. Nous avons ainsi étendu la banque de séquences protéiques de référence (UniProtKB/Swissprot) avec des informations de variants de séquence et d'épissage alternatif propres à chaque individu étudié. Ceci a permis d'avoir une banque plus complète sans compromettre la spécificité et la sensibilité de la recherche des peptides.

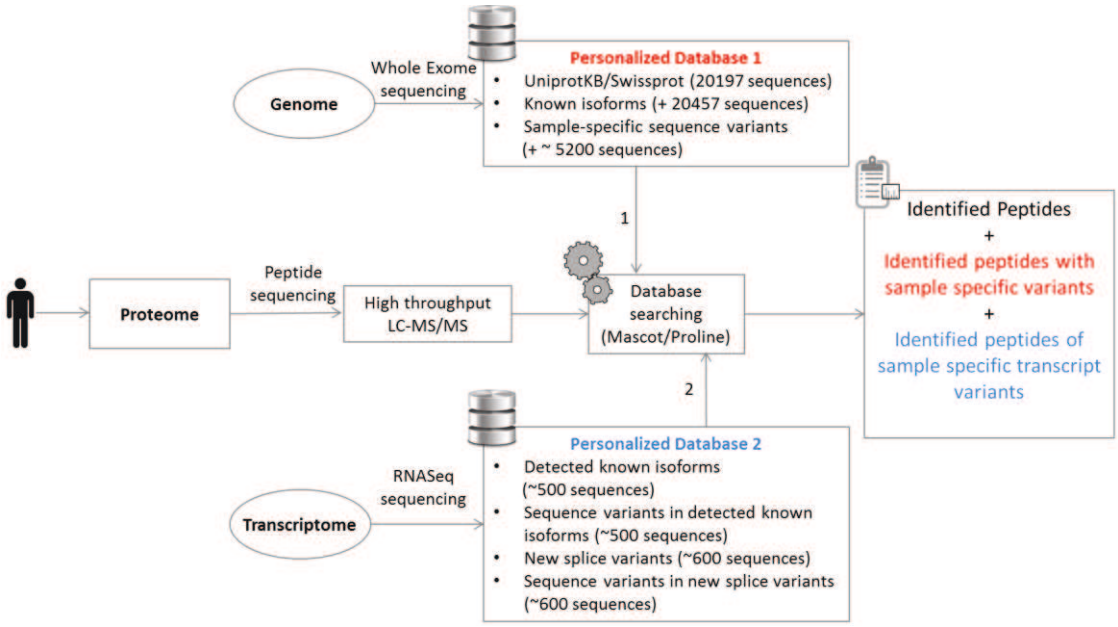



**Figure I-11 : Approche protéogénomique permettant l'amélioration de la caractérisation du protéome par l'utilisation de données de séquençage du génome et du transcriptome pour générer des banques personnalisées.**

Notre approche protéogénomique a permis d'améliorer l'identification des protéines et d'augmenter la couverture des séquences protéiques. Un exemple peut être vu dans la Figure I-12 où les avantages de l'utilisation d'une banque personnalisée sont illustrés. Pour la protéine présentée dans la figure, la recherche utilisant une banque personnalisée a permis l'identification de deux peptides supplémentaires contenant des variants de séquences génomiques spécifiques à chaque individu. Un acide aspartique et une lysine ont été remplacés par deux acides glutamiques dans la séquence peptidique VLWLDEIQQAVDDANVDKDR. Une leucine a été remplacée par une valine dans la séquence peptidique QTFIDNTDSIVK. SI le protéome de référence avait été utilisé cette information aurait été perdue.

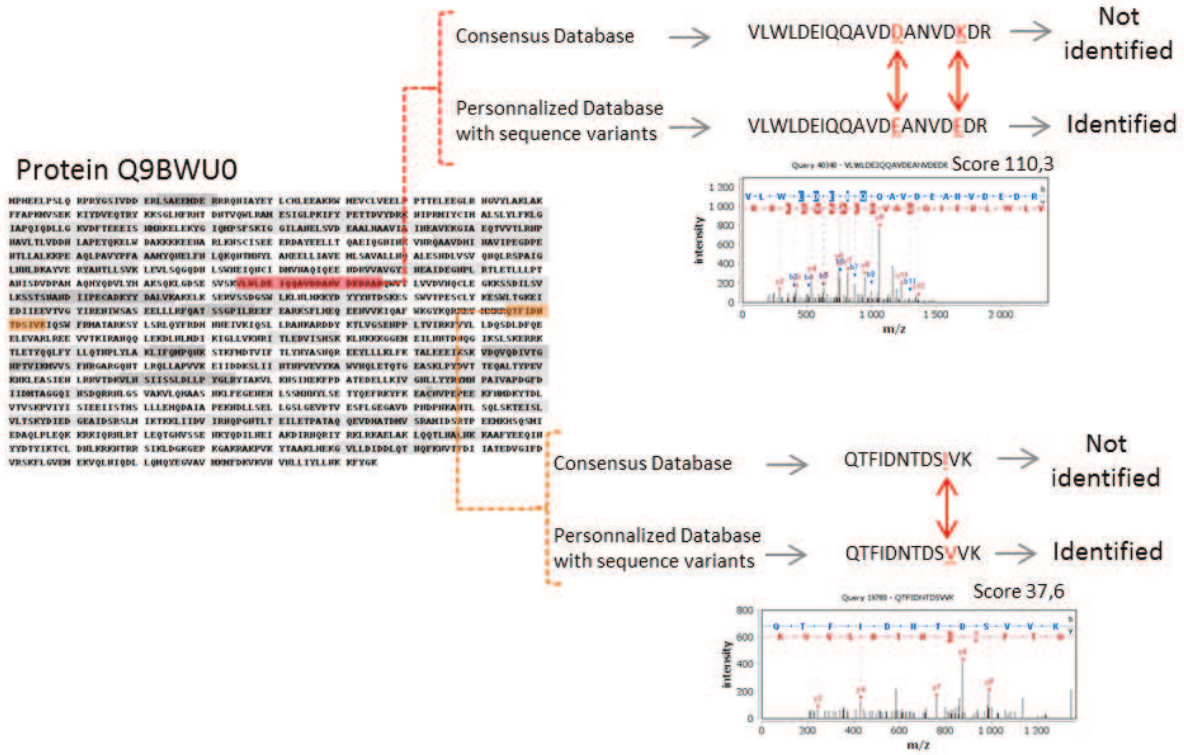



**Figure I-12 : Avantages de la protéogénomique pour l'identification de protéine et la couverture des séquences.**
La recherche utilisant une banque personnalisée a permis l'identification de deux peptides supplémentaires, pour la protéine présentée dans la figure, contenant des variant de séquences génomiques spécifiques à chaque individu. Un acide aspartique et une lysine ont été remplacés par deux acides glutamiques dans la séquence peptidique VLWLDEIQQAVDDANVDKDR. Une leucine a été remplacée par une valine dans la séquence peptidique QTFIDNTDSIVK.

De la même manière nous avons montré que cette approche permet d'identifier de nouveaux peptides provenant de mutations spécifiques dans le génome de chaque individu étudié. Egalement nous avons pu identifier des paires de produits de gènes hétérozygotes. L'utilisation d'informations de séquençage du l'ARN a également permis l'identification de variants d'épissage spécifiques à chaque individu.

L'amélioration de l'identification de protéines par l'utilisation de banques de données personnalisées implique également l'amélioration de la quantification des protéines. La Figure I-13.A. montre que l'augmentation de la couverture de séquence améliore la quantification de la protéine et la rend plus juste. La Figure I-13.B. montre que la protéine canonique P32455 a été vue sous deux protéoformes provenant de deux allèles différents d'un même gène. Deux peptides, l'un avec une thréonine et l'autre avec une serine en position 349, ont été identifiés. Les résultats du comptage spectral ont montré que l'une des formes hétérozygotes est surexprimée dans une condition analysée, tandis que l'autre forme ne l'est pas. Cet exemple démontre que la quantification de produits d'allèles spécifiques est possible au niveau de la protéine.

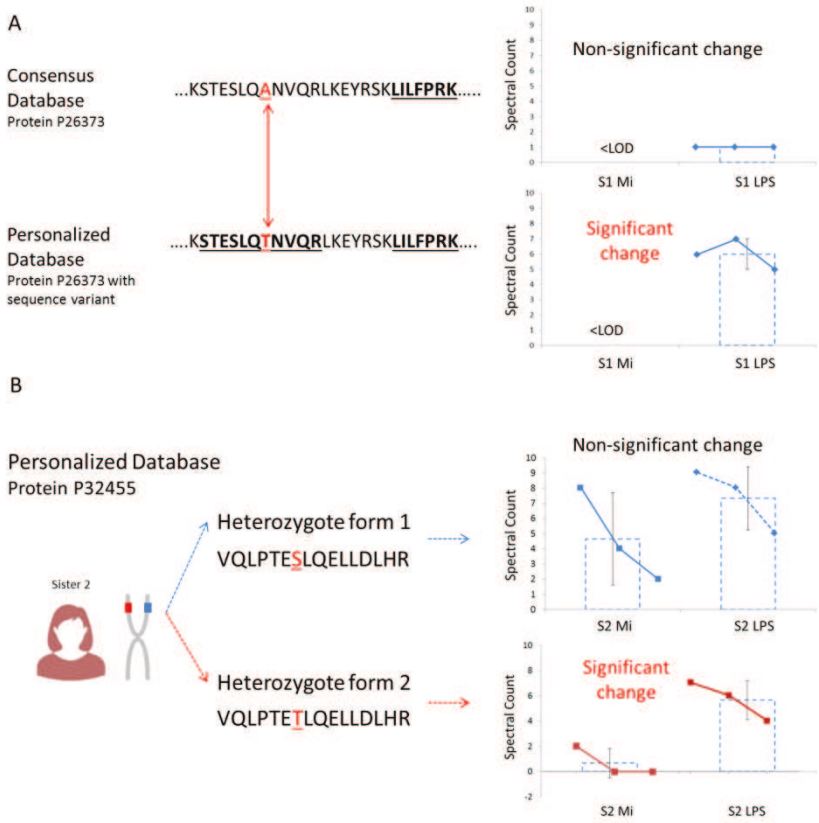



**Figure I-13 : L'amélioration de l'identification de protéines par l'utilisation de banques de données personnalisées implique l'amélioration de la quantification des protéines.**

A. En utilisant une approche d'analyse protéomique classique avec une banque de données consensus un seul peptide a été identifié (peptide souligné). Les résultats de la quantification par comptage de spectres ont montré un changement non significatif. Cependant, lorsqu'on utilise une banque de données personnalisée un peptide supplémentaire est trouvé et la quantification relative montre une surexpression de cette protéine. B. En utilisant une banque de données personnalisée la quantification des produits spécifiques d'allèles différents d'un même gène est possible au niveau de la protéine. Dans cet exemple la forme hétérozygote 1 a un changement non significatif alors que la forme hétérozygote 2 est surexprimée dans une des conditions étudiée.

Nous avons pu ainsi obtenir des informations qui n'auraient pas pu être obtenues par une approche classique. Nous avons pu identifier 106 variants de séquence appartenant à 96 protéines en utilisant les banques de séquence personnalisées à partir du séquençage de l'exome et 2 nouveaux variants d'épissage à partir du séquençage de l'ARN. Cette méthode a permis d'identifier de nouveaux peptides provenant de mutations spécifiques dans le génome de chaque individu, la preuve d'expression de gènes hétérozygotes et l'identification de variants d'épissage spécifiques à chaque individu. Cette approche a également permis d'obtenir une quantification relative plus juste entre les individus étudiés.

# Chapter IV Conclusion Générale

En conclusion, ce travail de thèse m'a permis d'acquérir des compétences en analyse protéomique et spectrométrie de masse. Les développements méthodologiques que j'ai réalisés ont permis d'une part d'améliorer et d'organiser la stratégie de développement de méthodes quantitatives ciblées par LC-SRM et de mettre en place les outils pour la quantification globale par DIA. D'autre part, j'ai développé des échantillons standards permettant des contrôles internes et externes pour améliorer la fiabilité des analyses et obtenir un suivi juste des performances instrumentales. Les développements que j'ai réalisés pour améliorer les approches de protéomique quantitative m'ont permis de résoudre avec succès une série de questionnements biologiques : j'ai pu montrer que la différence de la sensibilité de différentes lignées cellulaires face à la fumagilline (un traitement contre certains types de cancer) est bien corrélée à l'abondance de la protéine MetAP2 dans la cellule. Un résultat utile pour comprendre comment cette drogue induit l'arrêt de la croissance cellulaire. Mon travail de thèse a également ouvert des perspectives intéressantes pour la mise au point d'un nouveau diagnostic de la maladie de Crohn.

Le second volet de mon travail de thèse a consisté à développer de nouvelles méthodologies en protéogénomique. Dans ce contexte, mes travaux ont permis de créer des outils innovants de protéogénomique pour améliorer l'annotation des génomes et des protéomes et corriger les banques de séquences. D'une part, j'ai mis au point et optimisé un workflow complet, robuste, rapide et fiable pour l'analyse N-terminomique. J'ai appliqué cette stratégie pour corriger les annotations des codons d'initiation d'un génome bactérien (*Herminiimonas arsenicoxydans*) et pour caractériser finement le N-terminome du protéome mitochondrial humain. D'autre part, j'ai développé les outils nécessaires à l'amélioration de l'identification des protéines grâce à l'utilisation de données multi-omiques (séquençage d'exome et du RNA) pour créer des banques de séquences protéiques personnalisées. J'ai appliqué cette stratégie dans le contexte d'une recherche de biomarqueurs d'une maladie rare, la fièvre récurrente avec hyper-IgD.

# General introduction

Proteomics is the field of science that focuses in identifying, characterizing and quantifying all the proteins of a sample in a given moment and in a given condition[30]. The aim of Proteomics is to provide information on protein expression, cellular localization, post-translational modifications (PTMs), protein interactions, and protein turnover [31]. This is of major importance as proteins have a direct function in living organisms, and thus the comprehensive study of the proteome can offer the understanding of complex biological processes.

Contrary to the Genome which is static, the proteome is dynamic. The dimension of time, space, and nature of the samples, make the comprehensive analysis of a proteome more challenging than that of a genome. Additionally the full study of the proteome is hampered by its extreme complexity due to the multiplicity of chemically different versions of a same protein, called proteoforms [32]. These originate from genomic sequence variants, alternative splicing events, proteolytic events or post-translational modifications. Moreover, added to this complexity, the depth of analysis that can be achieved is challenged by the protein abundance dynamic range [33].

Proteomic analysis by liquid chromatography coupled to tandem mass spectrometry (LC-MS-MS/MS) is a technology-driven science that has rapidly developed by the advances in techniques for protein extraction, peptide and protein separation and mass spectrometry analysis. In recent years rapid advances in mass spectrometry enable now large-scale protein analysis due to advances in terms of resolution, mass accuracy, sensitivity and scan rates. In parallel, the exponential growth of protein sequence databases, their curation, the quality of the annotations and the improvement of bioinformatic resources have also played a key role in the development of Proteomics.

Global proteomic strategies are used for discovery studies that try to identify the maximum number of proteins present in a sample, and if possible, characterize their modifications. This approach uses mass spectrometers with high sensitivity, high resolution and high acquisition rates which enable the acquisition of data in data-Dependent Acquisition mode (DDA). In this mode, the most intense precursor ions of a MS spectrum are successively isolated and fragmented in a collision cell. The resulting fragments are recorded in a MS/MS spectrum. The identification of a protein is done by comparing the experimental mass list of parents and fragment ions to a mass list obtained in silico from a protein sequence database. This method has proved to be a powerful method to identify thousands of proteins in complex biological samples, and is today the standard method for the characterization of proteomes [34].

However, while the coverage of the proteome is becoming greater and greater - and the lists of identifications become longer and longer - without the quantitative information associated to each protein the response to biological questions remains only partial and insufficient in most cases.

It is also important to keep in mind that this strategy is not a *de novo* peptide sequencing method but is based on the matching of experimental and expected lists of masses, calculated from a reference protein sequence database. The inherent limitation of this approach is thus that it relies heavily on the protein sequence database used and on its completeness.

My doctoral work falls within this context and was intended to improve the proteome characterization by quantitative mass spectrometry and Proteogenomic method development. The methodological developments were optimized for and applied to several biological projects.

**The first part** of this manuscript provides a summary of the state of the art of bottom-up proteomics. It describes the steps and tools used to analyze and identify proteins in complex sample matrix by LC-MS-MS/MS. It then describes quantification strategies, both global and targeted. Lastly it will introduce the state of the art of recent methodologies promising the comprehensive proteome analysis, termed Data-independent Acquisition mode.

**The Second part** of this manuscript will focus on the state of the art of Proteogenomics, a field of research at the intersection between proteomics, transcriptomics and genomics. In this approach, personalized protein sequence databases are generated using genomic and transcriptomic information. This approach aims at eliminating the inherent dependency problem to protein sequence databases of Proteomics. In this part I will describe the current state of proteogenomic methods and software tools. Additionally I will summarize the state of the art of methodological strategies aiming to analyze the ensemble of proteins N-termini.

**In the third part** the results of the methodological developments for quantitative proteomics and its applications are presented.

- **The first chapter** will describe in a general way the optimizations made for the development of quantitative proteomics. The SRM assay method development workflow will be thoroughly described. Then the results of the evaluation of sample preparation protocols compatible with quantification methods will be discussed. During my thesis I also set up performance standard samples for targeted and global quantification platforms. These will be presented in this part as well. Additionally, an alternative targeted quantification method will be presented: Parallel Reaction Monitoring (PRM). The optimization and method development of Data-independent Acquisition methods will also be presented.
- **The second chapter** will describe the optimization steps and the results of the quantification of microbial proteins in the human gut microbiome, a sample of extreme complexity, without fractionation. This project aimed at validating biomarkers of Crohn's disease.
- **The third chapter** will present the development of targeted methods for the relative and absolute quantification of Methionine Aminopeptidase Proteins.

**In the Four part** the methodological developments concerning proteogenomic approaches and their applications will be presented.

- **In the first chapter** I will present the optimization of N-terminomic approach based on chemical labelling of proteins' N-termini using light/heavy TMPP reagent. The analytical workflow will be described as well as the steps towards the engineering of an automated workflow for the data interpretation for this approach. This method was applied to deeply characterize the proteome and N-terminome of human mitochondria.
- **In the last chapter** a personalized multi-omics profiling strategy to improve the proteome characterization with the use of personalized databases derived from exome sequencing and RNASeq data will be discussed. This method was applied to the study of a rare disease, Hyperimmunoglobulinemia D and periodic fever syndrome (HIDS).

# Part II State of the art of quantitative proteomics

## Chapter I Bottom-up proteomics

### A. Strategies for protein analysis by mass spectrometry

Three proteomic strategies have been developed in recent years: bottom-up, middle-down and top-down proteomics (Figure II-1). These strategies are described below.

**Bottom-up proteomics:** Bottom-up proteomics is a term referring to the characterization of proteins by the analysis of peptides created after enzymatic digestion, most commonly by trypsin (Figure II-1). The resulting tryptic peptides are easily fractionated, ionized and fragmented, making their analysis possible by liquid chromatography coupled to mass spectrometry. The identification of a protein is done by comparing the experimental mass list of parents and fragments ions to a mass list obtained *in silico* from a protein sequence database. In bottom-up proteomics the measure of protein is indirectly inferred from tryptic peptides. The protein inference is performed by assembling multiple peptide identifications a protein. Since peptides can be either unique to a given protein or shared by multiple proteins the identified proteins may be further grouped. The result of bottom-up proteomics is the smallest list of identified protein groups explaining the maximum number of peptide identification [35]. This approach has become the most popular approach for the large-scale analysis of protein by mass spectrometry[31].

This approach was used for all the studies detailed in this manuscript and will be further detailed in this chapter.

**Top-down proteomics:** in this approach intact proteins are characterized without an enzymatic digestion step (Figure II-1). The identification of proteins is done using the information of the MS and the MS/MS signals. The top-down approach promises to provide several advantages over the bottom-up approach, especially for the study of post translational modifications (PTMs) and proteoforms. Large scale studies have identified more than a thousand proteins using multi-dimensional separations in complex samples [36, 37].

However, the top-down method has major limitations compared with bottom-up proteomics due to the challenges regarding protein solubility, protein fractionation, protein ionization and fragmentation in the gas phase [38]. Due to the high complexity of the signals obtained in top-down proteomics, and the multiple charge states of intact proteins, this approach needs highly purified and fractionated samples. Additionally this method needs high-resolution instruments in order to resolve isotopic envelopes of multiple proteins present in several charge states. Time-of-flight or Fourier-transform based instrument can be used. However the scanning rates for this type of analysis are high. For example the Orbitrap Elite using a resolving power of 120000 at 400 m/z scans at 2,3 Hz handicapping thus the high-throughput capability of this method. Lastly, one of the bottlenecks for top-Down proteomics is the lack of dedicated instrumental software and bioinformatic pipelines[38].

**Middle-down proteomics:** This is a hybrid method between bottom-up and top-down proteomics. For middle-down proteomics larger peptide fragments are generated when compared to bottom-up proteomics. This minimizes the number of shared peptides between proteins. Additionally the larger peptides provide the main advantage of top-down proteomics, i.e. the characterization of PTMs, minus the challenges inherent to the analysis of intact proteins.

Bottom-up proteomics is now the method of choice for the analysis of proteins and their PTMs. The complementarity of top-down and middle-down methods will certainly improve the characterization of the proteome in the future.

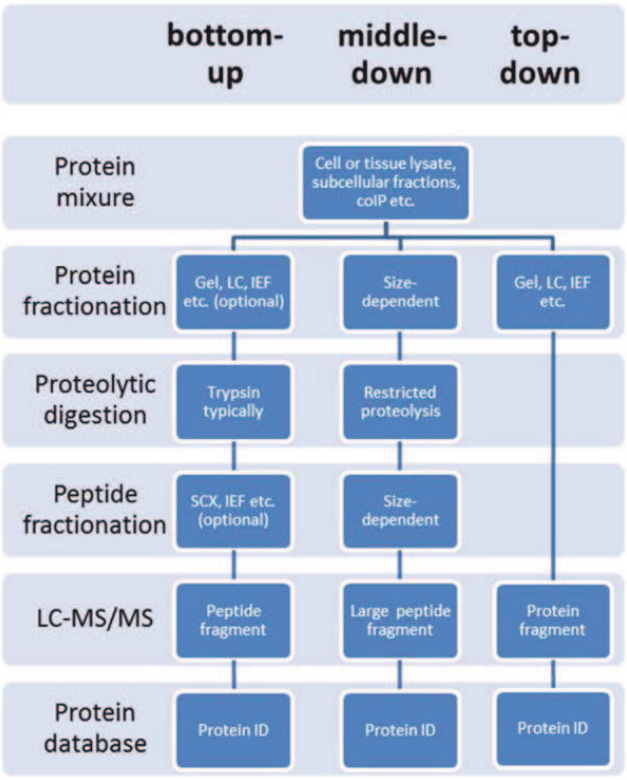



**Figure II-1: Bottom-up, middle-down and top-down workflows (Adapted from [31]).**

## B. Analytical workflow considerations for Bottom-up Proteomics

The proteome is an extremely complex sample, not only due to the number of proteins present in a sample (20250 protein-coding genes for humans), but also for the multiplicity of chemically different versions of a same protein, called proteoforms [32]. These originate from genomic sequence variants, alternative splicing events, proteolytic events or post-translational modifications [9]. Moreover, added to this complexity, the depth of analysis that can be achieved is challenged by dynamic range of protein abundance that can reach more the 6-7 orders of magnitude [5, 33]. The dynamic range of an MS instrument can reach 3 to 5 orders of magnitude depending on the acquisition parameters. To be able to sensitively and selectively analyze proteins of interest it is necessary to decomplexify the samples prior to MS analysis. Depletion of highly abundant proteins, the fractionation of samples at the protein-level and/or the separation of peptides prior to MS analysis can be used to reduce the overall complexity.

### *B.1. Protein separation and purification*

Several methods exist to separate proteins based on their different physico-chemical properties [39]. These are among others the molecular weight (size exclusion chromatography, sodium dodecyl sulfate-Polyacrylamide gel electrophoresis (SDS-PAGE), the charge or the hydrophobicity (hydrophobic interaction chromatography, ion exchange chromatography, affinity chromatography, immunoaffinity chromatography, reversed phase

chromatography…), the isoelectric point (Isoelectric focusing) or a combination of many (Two-dimensional gel electrophoresis).

In the projects described in this manuscript, when a fractionation was needed, the SDS-PAGE approach was used[40]. Sodium dodecyl sulfate (SDS) is an amphipathic detergent. It has an anionic group and a lipophilic tail. It binds uniformly and non-covalently to proteins, with an approximated ratio of 1.4g SDS/1g protein. SDS denatures proteins, disassociates complexes and confers a uniformly distributed negative charge. The proteins' intrinsic charge is thus masked and all proteins have very similar charge-to-mass ratios. During the migration in SDS-PAGE, the protein migration will be determined by the molecular weight. A clear advantage of this sample preparation is the high yield of protein extraction and solubilization.

When a fractionation was not necessary, or not recommended for the type of analysis that was going to be performed, a liquid digestion protocol was most commonly used. In this protocol the proteins were extracted using an urea buffer, and then separated from non-protein contaminants using a protein precipitation step (Acetone precipitation). An unfractionated sample preparation protocol using the SDS-PAGE approach was developed and will be presented in Part IVChapter IB.2 on page 93.

### *B.2. Peptide separation and purification*

The separation of peptides prior to mass spectrometry analysis is crucial in order to increase the sensitivity, the selectivity and the depth of the analysis of a proteome. The most commonly used approach is liquid chromatography (LC) [41].

In the projects presented in this manuscript reverse-phase LC was used to separate peptides prior to MS-MS/MS analysis. Three types of LC systems were used and are described in Table II-1.

**Table II-1: Description of LC-systems present in the laboratory.**

| LC system | MicroLC | nanoLC-Chip | UPLC™ |
|---|---|---|---|
| **Manufacturer** | **Dionex or Agilent** | **Agilent** | **Waters** |
| **Stationnaire phase** | C18 | C18 | C18 |
| **Colum lenght (mm)** | 150 | 150 | 200 |
| **Internal diameter (µm)** | 300 | 75 | 75 |
| **Particles size (µm)** | 3,5 | 5 | 1,7 |
| **Flow rate (µL/min)** | 5 | 0,3 | 0,3 or 0,45 |
| | Capillary-flow | Nano-flow | Nano-flow |

In Proteomics the UPLC systems are now the method of choice for protein identification. The use of small particles provides excellent chromatographic resolution. Nano-flow LC platforms provide also good sensitivity, high peak capacity, high resolution and it enables low sample injection volume. But technical problems common to nano-flow LC systems still remain a challenge (ESI spray instability, not-easily detectable leaks, high back pressure or dead volumes). However the main advantage of nano-flow LC is the detection sensitivity that can be achieved as a result of reduced sample dilution. It is ideally suited for studies in which the sample amount is limited.

Miniaturized systems (nanoLC-Chip) enable the reduction of void volumes and analysis times. However the sample capacity is limited.

Standard-flow provides lower sensitivity compared to nano-flow LC. However in some cases this can be countered by the higher sample capacity. A study showed that standard-flow can provide globally superior sensitivity than nano-flow, with higher retention time reproducibility and increased ease of use [14].

Nano-flow and capillary-flow systems were used in my thesis. Nano-flow LC was used for discovery studies and capillary-flow systems were used for targeted quantification studies in which the robustness was necessary.

## C. Mass spectrometry analysis

### *C.1. Tandem Mass spectrometry*

In bottom-up Proteomics hybrid instruments combining the properties of different mass analyzers are used. These instruments are used to obtain information about the sequence of peptides.

This strategy is called tandem Mass Spectrometry. Two stages of MS are coupled in series, this way a given peptide can be isolated and fragmented in a collision cell. A mass spectrum of the resulting fragments is then generated called MS/MS or MS2 spectrum [30].

In the projects described in this manuscript several types of hybrid instruments were used: Triple quadrupole (QQQ), Quadrupole-Time-of-flight (Q-TOF) and Quadrupole-Orbitrap (Q-Orbitrap).

### *C.2. Peptide fragmentation*

Four fragmentation modes are commonly used in Proteomics: CID (Collision induced Dissociation) [42], HCD (Higher collision Dissociation)[43], ETD (Electron Transfer Dissociation)[44] and ECD (Electron Capture Dissociation) [45]. In proteomics the most common fragmentation mode is CID. The isolated ions are accelerated to induce a high kinetic energy and then they collide with neutral molecules present in the collision cell (helium, nitrogen or argon). As a result of the collision some of the kinetic energy is converted into internal energy which induces the fragmentation of the peptide. In CID the fragmentation follows the mobile proton model [46]. This type of fragmentation is perfectly adapted to tryptic peptides as they usually possess two charges (One in the N-terminal position and one on the side-chain of Lysine or Arginine). The fragmentation in CID provokes the peptide bond breakage. This fragmentation characteristic is why this fragmentation mode made possible the rapid development of Proteomics. The peptide fragmentation rules and nomenclature was extensively studied by Biemann [47] (Figure II-2). The most predominant and informative ions are the fragments resulting from the breakage of the amide bond between amino acids. The resulting ions are called b- and y-ions if the charge is respectively retained by the N-terminal or the C-terminal part of the peptide. When using CID b- and y-ions are the predominant ions. In ETD and ECD, z- and c-ions are most common.

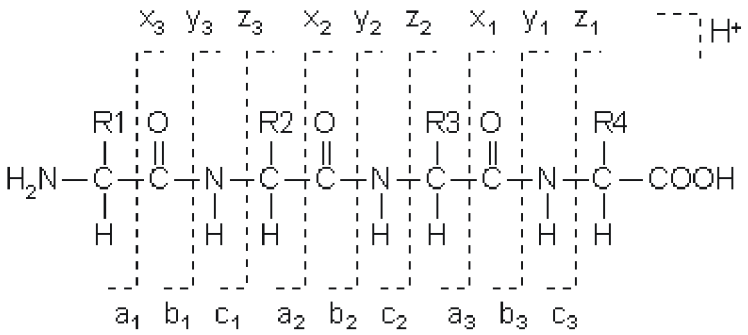



**Figure II-2: Biemann nomenclature for peptide fragmentation.**

Recent instrumental developments have rendered possible the combination of ETD and CID fragmentation mode, termed EThcD [48].

After the fragmentation, all the fragment ions of a peptide are measured simultaneously on a MS/MS spectrum. Using this MS/MS spectrum the sequence of the peptide can be determined (Figure II-3).

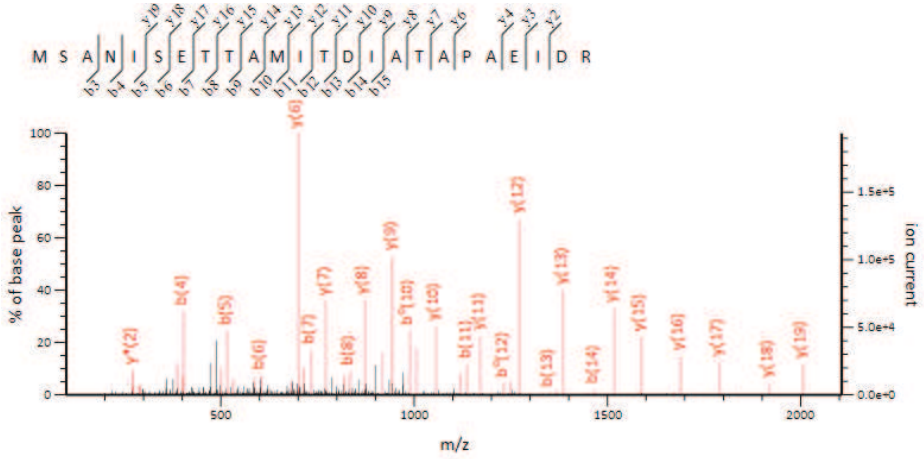



**Figure II-3: Example of an annotated MS/MS spectrum providing the peptide sequence.**

## D. Data-dependent acquisition (DDA)

The most common acquisition mode used in bottom-up Proteomics is Data-Dependent Acquisition (DDA). In this mode the instrument acquires alternative MS and MS/MS cycles (Figure II-4). First a MS survey scan is acquired and a number N of precursor ions are selected to be analysis by a MS/MS scan. The cycle of one MS and N MS/MS repeats throughout the whole duration of the analysis. The N precursor ions which are chosen are the most intense ions in the MS spectrum. The analysis of a given peptide will depend thus on its intensity and on the intensity of the other co-eluting peptides. This is why this method is called Data-Dependent Acquisition.

To improve this acquisition mode, technological efforts have been made to improve instrumental sensitivity and scan rates. The depth of the identifications can be increased by excluding already selected peptides for a given amount of time (frequently the half of the average chromatography peak width) to reduce spectral redundancy.

Even if this method has proved to be a powerful tool for Proteomics, enabling the identification of thousands of proteins per run [34], its major drawback is the stochastic nature of the peptide selection that causes undersampling[1]. The analysis of samples in replicates can help to reduce the undersampling. To improve the depth of analysis exclusion or inclusion list of precursor ions can be used [49, 50].

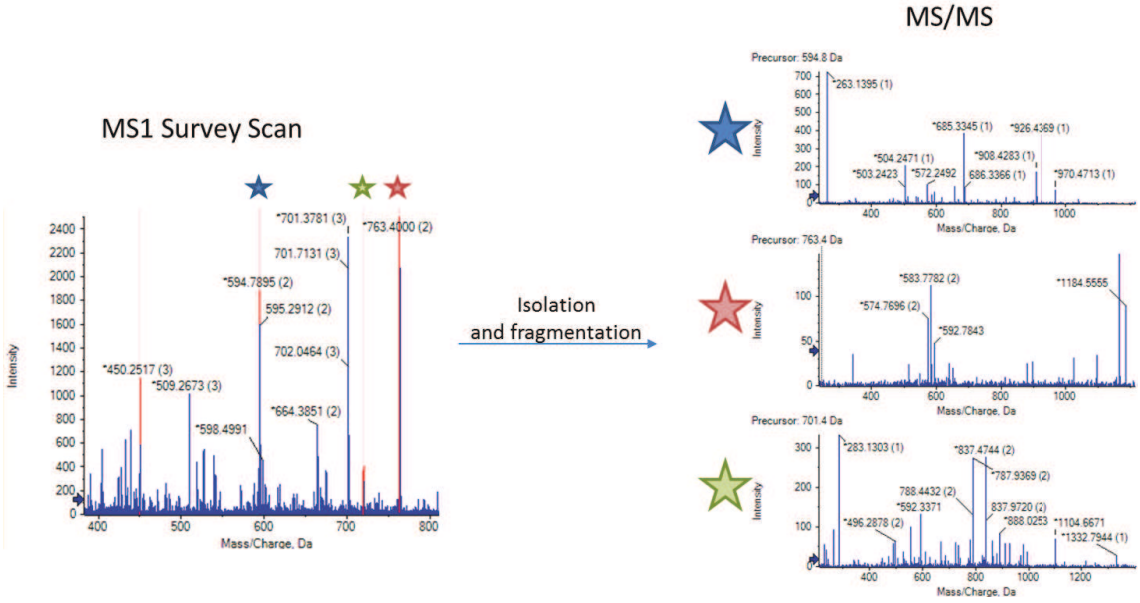


Figure II-4: Scheme of Data-Dependent Acquisition mode.

## E. Protein identification strategy

### *E.1. Search engines*

The method to identify peptides from DDA data is termed peptide fragmentation fingerprinting (PFF) [51]. This strategy consists in transforming the raw DDA data in mass lists composed of the precursor peptide and its corresponding fragments m/z ratios. These experimental mass lists are matched with the mass lists all peptides from a protein sequence database digested and fragmented *in silico.* Many search engines have been developed in the last years, such as Sequest [52], Mascot [53], OMSSA [54], X!Tandem [55] and Andromeda [56].

All these search engines need information about the experimental and instrumental conditions in which the data were acquired. The following information is necessary to perform the search:

- The tolerance of the precursor and the fragment ion m/z ratio.
- The charge of the precursor and the fragment ions.
- The digestion enzyme used and the maximum number of tolerated missed cleavages.
- The protein sequence database.
- The type of fragmentation mode (CID, ETD).

In the work presented in this manuscript the search engines used were Mascot and OMSSA.

Mascot is a proprietary search engine and thus the complete description of the algorithm is not openly accessible. For each MS/MS spectra an ion score is calculated. This corresponds to the probability that the observed match between the experimental mass list and the *in silico* calculated mass list happens by chance. The higher the score the more confident the peptide identification is. This score does not depend on the search space, only on the quality of the spectra. Moreover Mascot also calculates an identity score which is associated to the size of the search space. This value will provide an idea of how well the peptide's identification separates from the distribution of random scores. If it is well separated then this match is not a random event. However, the larger the search space, the higher the identity score will be [53]. Then for each spectrum all possible peptide identifications, termed Peptide Spectrum Matches (PSM), are ranked.

For OMSSA, the score is an expectation value (e-value) which is the probability that the matching of a peptide sequence to an experimental MS/MS spectrum would occur by chance if the trial was repeated several times [57]. For example, an e-value of 1 indicates that there is the same chance of having a true or false positive identification. An e-value of 0.01 indicates that the matching of a peptide sequence to an experimental MS/MS spectrum would occur by chance one time in 100 given many trials. This value directly depends on the search space [54].

Since the algorithms of the search engines are not based on the same principles (scoring, ranking…) they provide complementary results. The combination of search engines provides thus complementary results that increase the total number of protein identifications and give high confidence to spectra for which multiple search engines identified the same peptide sequence [58-60].

## *E.2. Protein sequence databases*

The protein sequence database has a central place in the identification of proteins. It is thus very important to work with high quality and curated databases. However the publically available databases have different degrees of data quality [61].

### E.2.1 NCBI's Entrez Protein database

This database was created by the National Center for Biotechnology Information (NCBI) [62]. This database is an example of a protein sequence repository with high redundancy and no data curation. The Entrez Protein database is composed of protein sequences translated from nucleotide sequences databases (EMBL, DDBJ[63] and GenBank [64]). It also contains sequences from Swiss-Prot, Protein Information Ressource (PIR) [65], RefSeq [66] and the Protein Databank (PDB) [67]. This database does not create new annotations but extracts the information from the databases cited above. Additionally the database is very redundant.

### E.2.2 NCBI's RefSeq Database

The RefSeq database is also a database produced by the NCBI [66]. RefSeq however is a curated and non-redundant database. For each protein the link between the protein and the gene and transcript information is done. The data is periodically updated, curated and stored in a consistent format. However, the data is automatically generated with very little manual curation. In May 2016, the database obtained more than 61 million protein sequences for more than 58 thousand organisms (RefSeq Release 75).

These databases can be prone to hold sequence and annotation errors. Errors in gene annotation can propagate errors in protein sequences, such as incorrectly defined open reading frames or wrong annotated translational start sites, thus leading to errors or impossible MS spectra identifications [68].

### E.2.3 UniProtKB Database

The UniProtKB/SwissProt database has now emerged as the database of reference for Proteomic analysis of model organisms[69]. This database is the manually annotated and reviewed section of the UniProt Knowledgebase. This database was created by the European Bioinformatics Institute (EBI) and the Swiss Institute of Bioinformatics (SIB). This database is non-redundant and integrates information from other databases. This database provides annotations of high-quality, accurate, up-to-date and manually curated from the literature. The database provides, among others, information of function, subcellular location, related pathologies, related pathways, PTMs, Processing events, level of expression and structure.

The other part of this knowledgebase is UniProtKB/TrEMBL which is composed of automatically annotated and classified data. The data from this database is then selected for full manual annotation and integration into UniProtKB/SwissProt.

### E.2.4 neXtProt Database

The neXtProt database is a human protein-centric knowledgebase [70]. It was created by the SIB and focuses exclusively on human proteins. It is a high-quality highly curated database and also includes a focus on Proteomic analysis by mass spectrometry. Very helpful information for proteomics is accessible such as information of all proteins with respect to their existence (protein evidence index), their abundance, their distribution and subcellular localization. Information is also given about the isoforms expressed, post-translational modifications, sequence

variants, peptide unicity and information on previously observed peptides by LC-MS-MS/MS, with links to PeptideAtlas and SRMAtlas.

In this manuscript high quality databases were preferred, when possible. The databases were customized to contain only the proteins corresponding to the taxonomy of the samples being analyzed. This was done to reduce the processing time and more importantly to reduce the number of false positives.

### *E.3. Validation of protein identification*

The protein identifications returned by search engines have to be validated. The score given by a search engine cannot by its own validate the true identification of a protein, as the possibility of having false positive identifications in the data is non-negligible.

The most common approach to validate proteomic data is the target-decoy strategy [71]. In this approach a concatenated protein database containing the target proteins (the real protein sequences) and decoy proteins (reversed protein sequences) is generated. The identifications of reverse peptide sequences, which are by definition false, provide the rate of false positive identifications.

The False Discovery Rate (FDR) [72] can then be calculated as :

$$FDR\% = 2 \times \frac{Number\ of\ decoys}{Number\ of\ decoys + number\ of\ targets} \times 100$$

The calculation of the FDR has to be done at the PSM, peptide and protein level [57]. To obtain high-confidence results a 1% false discovery rate (FDR) at the protein level is often used.

Several guidelines for publication have set stringent validation criteria for proteomic data[73]. In the context of the Human Proteome Project (HPP) guidelines have also been established for the publication and the dissemination of results [74].

# Chapter II Global Quantification approaches

## A. Stable-isotope Label-based quantification

Peptide and protein quantification using stable-isotope labeling relies on the fact that the labeled and unlabeled peptides have the same physicochemical properties. This implies that they will have the same behavior in mass spectrometry (MS signal response, ionization efficiency, fragmentation pattern…) and in chromatography conditions. Different samples can thus be compared using the MS signal response.

### *A.1. Metabolic labeling strategies*

In metabolic labelling strategies, the isotope label is done during cell growth and division. This ensures that every protein is labelled in the sample of interest. It has the major advantage of introducing the labelling step at the earliest possible step in a proteomic workflow. This can correct confounding errors originating from sample preparation steps and thus provide high quantification accuracy and high precision. The most wide-spread metabolic labelling method is the stable isotope labeling with amino acids in cell culture (SILAC) [75]. In this approach isotopically labeled arginine and lysine ($^{13}C$, $^{15}N$) are introduced in the culture medium. This way all tryptic peptides contain at least one labeled amino acid. The samples to be compared are mixed together and the relative quantification is done by comparing the intensities of the labelled and unlabelled peptides.

However this technique is time-consuming as the culture process is slow. Additionally the samples have to be compatible with cell culture processes, which is not always the case. [76].

### *A.2. Chemical protein and peptide labelling*

It this approach a peptide's reactive group (α-amino groups, ε-amino group of lysine or thiol of cysteine) is modified using stable-isotope labels.

**Isotopic labelling**: Among this type of approaches, historically, the first approach was the one termed ICAT (Isotope Coded Affinity Tag)[77]. This technique is based on the use of light and heavy reagent containing biotine group. The reagent reacts with thiol groups of cysteine. The biotine group enables the protein purification by affinity chromatography prior to MS analysis. However with this technique only cysteine-containing peptides can be analyzed. Additionally the deuterated labelling of the heavy reagent induces a shift in retention times during reversed-phase LC compared to the light reagent[78].

**Isobaric tags:** This method targets mostly the peptide or protein N-terminus and the ε-amino group of lysine. Several variations of the same tag having the same mass (isobaric) but producing fragment ions of different masses (reporter ions). The reporter ions are in the lower mass region of MS/MS spectra. The relative quantification of several samples is done by labelling each sample with a different isobaric tag at the peptide level. The intensities of the reporter ions are used for the quantification. The advantages of isobaric labeling approaches are the higher multiplexing capability compared to isotopic labelling approaches (8 or 10-plex) [79, 80]. Also the complexity of LC separations and of the MS analysis is not increased thanks to the co-elution of isobaric labelled-peptides. The most popular approaches are the isobaric Tag for Relative and Absolute Quantification (iTRAQ) [81] and the tandem Mass Tags (TMT) approaches [82].

## B. Label-Free Quantification

Label-based approaches are rather expensive and the total multiplexing capabilities is limited by the number of stable-isotope reagents. In order to overcome these limitations label-free approaches have been developed in recent years [83]. In label-free approaches the number of samples that can be analyzed is, in theory, not limited. Moreover contrary to the SILAC approach which can only be used for samples compatible with cell culture conditions, label-free approach can be applied to all types of samples (tissue, cells, body fluids…). The data used in these types of approaches is acquired in DDA mode.

### *B.1. Spectral Counting*

The spectral count approach is based on the assumption that in DDA mode a protein concentration is correlated to the number of acquired MS/MS spectra for the protein [84].

The major advantage of this quantification strategy is the simplicity of the data analysis. However this method is limited by the undersampling of the DDA acquisition mode which creates incomplete data with missing values. Moreover to be able to carry out a spectral counting quantification the DDA method has to be set up differently than a classical DDA discovery approach where the objective is to identify low-abundant proteins. For spectral count, the exclusion times have to removed or very short, to have a spectral redundancy that represents the protein abundance in the sample. This means that the depth of the analysis (the number of identified/quantified proteins) will be reduced.

The data analysis of spectral count data is different when compared to other quantitative techniques. In spectral count the values attributed to the peptides and proteins are discrete values. Thus the statistical procedures performed on the data (imputation of missing values, normalization, testing the significance of protein abundance variations) have to be adapted to the nature of the data [85]. To be able to confidently quantify a difference in protein abundance the spectral count for a protein has to be high. This is why this method performs better to quantify high-abundant proteins and is not very performant to detect variations on low abundant proteins.

### *B.2. MS1 Filtering - Extracted Ion Chromatograms (XIC)*

This strategy consists in extracting ion chromatograms from MS1 scans of precursor ions of peptides of interest. The area under the curve is then used for the relative quantification between samples [86]. This approach requires instruments with high-resolution, high-accuracy instruments and high scanning speed, in order to be able to reconstruct well-defined chromatogram peaks and reduce the risk of integrating wrong precursor ions with close m/z ratios.

Two approaches have developed for the extraction of ion chromatograms:

**Extraction of all detected ions:** The detection and integration in the MS1 signal is done for all ions having an isotopic pattern resembling that of a peptide. The integrated ions are called features. The identification of the peptide sequence corresponding to each feature is not necessary for this first step. The advantage of this approach is that, for all analyses that have to be quantified, the retention times are aligned and matched to a reference analysis. The XIC for a given feature can be extracted even if the peptide could not be identified. In a second step, each feature will be matched to the corresponding peptide sequence. Several software tools have been developed using this approach. The most popular ones are Progenesis LC-MS (Nonlinear Dynamics), MaxQuant [87], MFPaQ [88].

A software tool implementing such an approach is being developed in the frame of the French Proteomics Infrastructure (ProFI), named Proline (http://proline.profiproteomics.fr/) and I have participated in its evaluation as an alpha tester during my PhD.

**Targeted Extraction of validated peptide precursor ions:** The list of precursor ions for which a XIC will be extracted is defined by the list of identified peptides. Since the data is acquired in DDA mode the same data can be used to identify and quantify the peptides present in the sample. However, this limits the quantification only to the list of identified peptides. A spectral library is often used to guide the extraction of the chromatograms. For a given peptide the extraction and the signal integration is done around the retention time indicated in the MS/MS spectrum which was used to identify the peptide. The most popular software tool using this approach is Skyline [15]. The principles of operation of Skyline are described in Figure II-5.

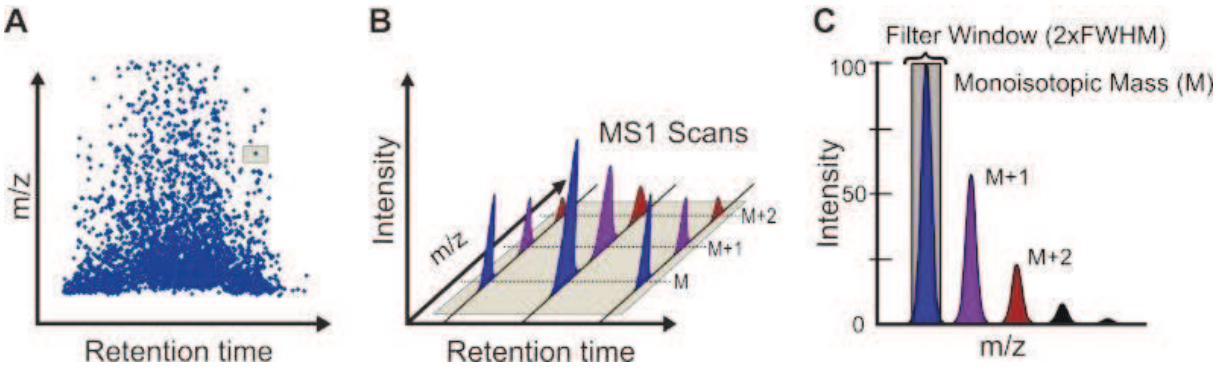



**Figure II-5: Schematic of MS1 filtering (Adapted from [86])**
A. Peptide distribution in the m/z and the retention time dimension. B. Isotopic envelope for the molecular ion of a peptide with peaks at M, M + 1, and M + 2 selected showing changes in MS1 intensity over time. C. High resolution data allow specific filtering of molecular ions and separation of individual peaks within the isotope distribution. Skyline sums intensities within a window of twice the theoretical resolution, predicted full width at half-maximum (2×FWHM). The resolution setting can be selected by the user depending on MS instrument type.

## C. Label-free quantification development Workflow

The steps required for the development of a Label-free quantification method are shown in Figure II-6. The steps required to develop a label-free method start with a hypothesis. Then the LC-MS instrument parameters are optimized according to the sample and the instrument used. The parameters used for a XIC quantification method will be the same as those used for discovery method where the objective is to identify the maximum number of peptides. After the acquisition is done, the peptides and proteins present in the sample are identified and validated (<1% FDR). A spectral library with the information of a peptide's m/z ratio and retention time is created. The use of a spectral library is not necessary for the extraction of ion chromatograms but it is highly recommended as the coordinates of the targeted peptides are inferred. This enables to reduce the processing time and reduces the chances of integrating wrong peaks originating from peptides having close m/z ratios. Finally the confirmation of the presence of a given peptide in the sample is done.

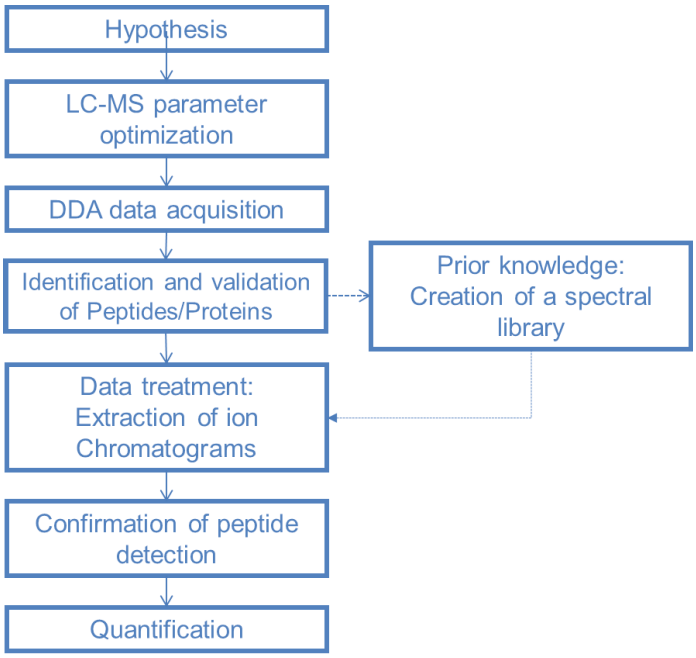



**Figure II-6: Label-free quantification development Workflow.**

# Chapter III Targeted Quantification

## A. Selected-Reaction Monitoring

### *A.1. Principle of Selected Reaction Monitoring MS for targeted quantification*

The SRM approach is a targeted quantification strategy that allows the quantification of only a predefined set of peptides of interest, such as biomarker candidates. It is used when a reproducible and accurate quantification is needed across large number of samples.

It is mostly run on triple quadrupole type instruments composed of two quadrupole mass analyzers with a collision-induced dissociation (CID) cell between them (Figure II-7). In Selected Reaction Monitoring (SRM) or Multiple Reaction Monitoring (MRM) mode the first quadrupole (Q1) acts as a mass filter for a predefined mass. It will isolate a precursor ion that will then be fragmented in the collision cell. The third quadrupole (Q3) also acts as a mass filter for a predefined mass of a fragment ion. The couple of precursor and fragment ion is called a transition.

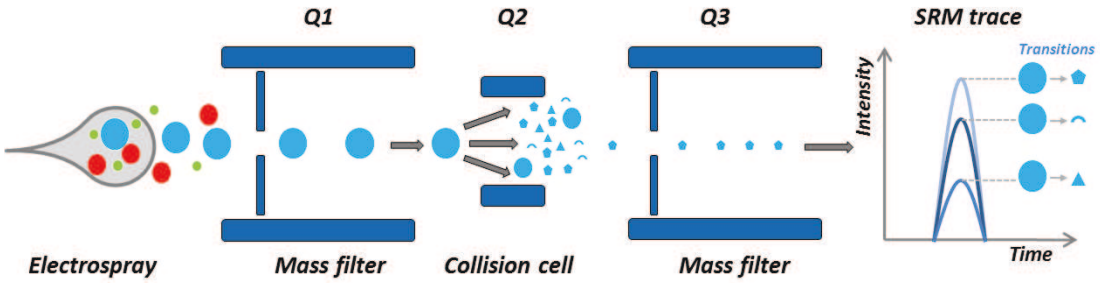



**Figure II-7: Principle of Selected Reaction Monitoring (SRM) or Multiple Reaction Monitoring (MRM) scanning mode.**

The SRM acquisition mode is an iterative process: during an analysis, the mass spectrometer scans all the predefined transitions sequentially. The double selection of a precursor and a specific fragment ion provides high specificity and sensitivity. The higher the number of monitored transitions per peptide is the higher the specificity will be (Figure II-8).

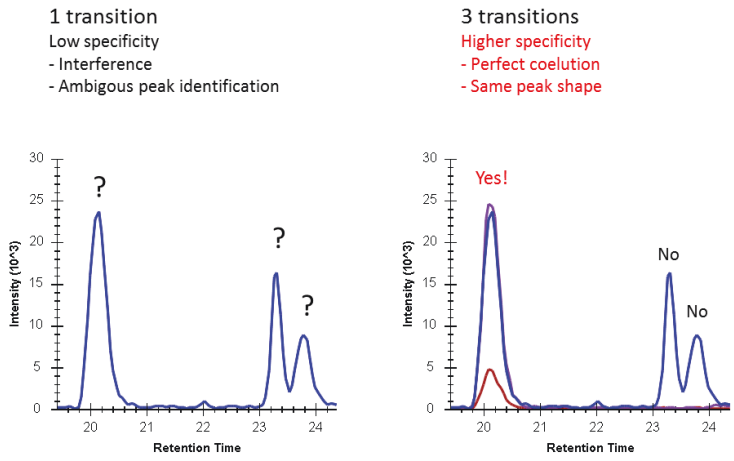



**Figure II-8: Monitoring multiple transitions per peptide provides higher specificity.**

The perfect coelution of transitions with similar peak shape proves the presence of the peptide in the analyzed sample without any interference and without any ambiguous peak identification.

Additional characteristics of targeted proteomics are that it provides comprehensive data without any missing value conversely to global label free MS1 quantification strategies. Since it is a deterministic approach it is able to monitor peptides across their whole chromatographic elution peak. The precise and accurate quantification is done by integrating the area under the curve for each trace. Even when a peptide is not present, a background signal is

measured thus proving that the peptide is absent or in a concentration lower than the instrument's detection limit. Targeted proteomics is therefore often used when reproducible and accurate quantification is required across a large number of samples.

This method has become the gold-standard method for precise quantitative proteomics and has won the title of Method of the year by the journal Nature Methods in 2012 [89].

During my PhD, I have enhanced the LC-SRM assays development workflow by optimizing key parameters to attain highest sensitivity and specificity (Part IVChapter IA. on page 69).

### *A.2. SRM assay development Workflow*

The steps required for the development of an SRM assay are shown in Figure II-9. SRM is a hypothesis-driven technique and the quantification is done on a predefined set of proteins of interest. The proteins will be quantified by a set of proteotypic peptides, i.e. peptides unique to the protein and visible in mass spectrometry. For each one, the best-responding transitions will be chosen. Then the LC-SRM platform's parameters need to be optimized in order to achieve the best sensitivity. All these steps and the methodological developments towards improving this workflow will be discussed in detail in Part IVChapter I.

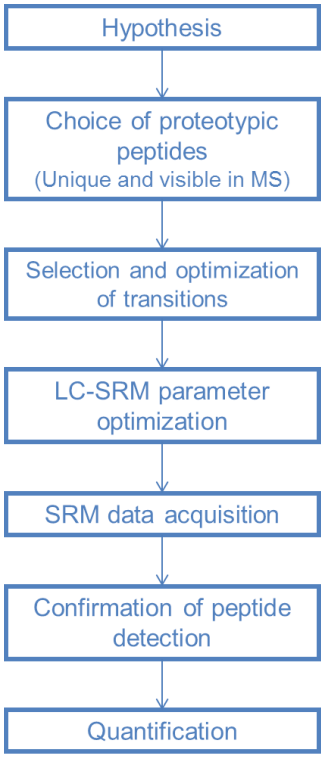



**Figure II-9: SRM assay development workflow.**

### *A.3. Balancing instrument time and multiplexing capabilities*

During an LC-SRM analysis hundreds of transitions can be analyzed in a single run. The time spent to measure a given transition is called the dwell time. It often ranges between 5ms and 100ms. This parameter is related to the sensitivity of the analysis. The higher the dwell time the higher the signal-to-noise ratio and thus the higher the sensitivity. The cycle time corresponds to the time necessary to monitor the complete list of transitions. This parameter is related to the accuracy of the quantification. The lower the cycle time, the more points will be acquired to reconstruct the chromatographic peak for a given transition. Typical chromatographic peak widths range between 15 to 30 s. The cycle time needs to be adjusted to obtain at least 8 to 10 points across the chromatographic peak. The cycle time

often ranges between 1 and 3 seconds. The inter-scan time is the time necessary to change the voltages to monitor another transition. It has a fixed value around 1 ms (depending on the instrument used).

Optimizing an LC-SRM method consists in finding a balance between the dwell time, the cycle time and the number of transitions to be analyzed. All these parameters are related by the following equation:

$$\textbf{\textit{Cycle time}} = \textbf{\textit{Number of transitions}} \times (\textbf{\textit{Dwell time}} + \textit{interscan time})$$

| User-defined (has to be as short as possible) | Is fixed for each MS analysis | User-defined (Has to be as long as possible) | Is fixed |
|---|---|---|---|

During a SRM experiment a typical set of parameters will be a cycle time of 3 seconds, a dwell time of 20ms and 3 transitions monitored per peptide. This means that 150 transitions corresponding to 50 peptides can be analyzed in a given run.

In order to extend the number of analyzed peptides in a single run, the transitions of a given peptide can be analyzed only during the time the peptide is eluted out of the chromatographic column. This method is known as Scheduled-SRM [2] (Figure II-10). The transitions are only monitored during a time window often between 2 to 5 min. This method optimizes the instrument time and increases its multiplexing capabilities. In this method, the cycle time is kept constant but the number of transitions changes during the analysis. The dwell time is thus optimized to reach the best sensitivity. The instrument time parameters are now related by the following equation:

$$\textbf{\textit{Cycle time}} = \textbf{\textit{Number of transitions}} \times (\textbf{\textit{Dwell time}} + \textit{interscan time})$$

| User-defined (fixed-value) | Changes throughout the gradient | Changes throughout the gradient (Longer when the number of transitions is small and vice versa) | Is fixed |
|---|---|---|---|

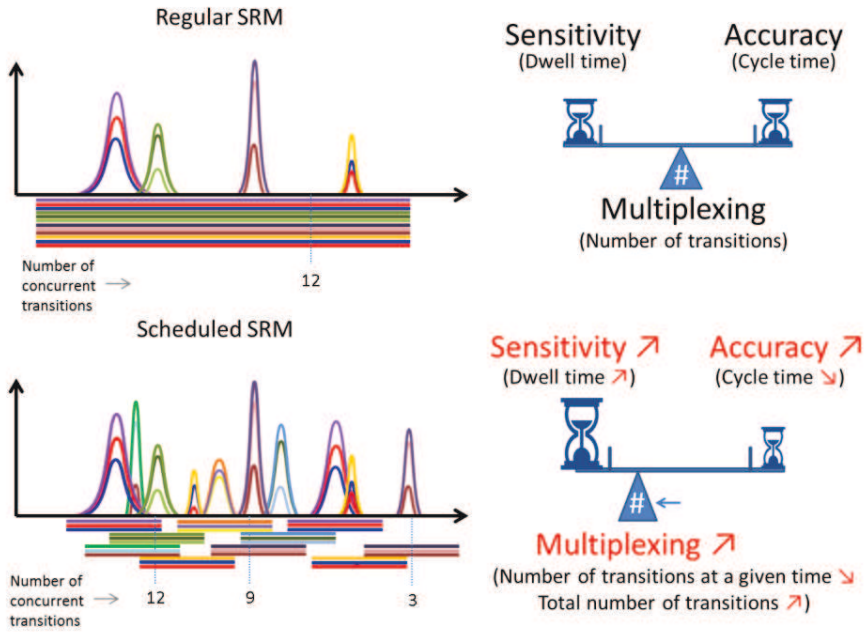



**Figure II-10: Scheduled SRM optimizes instrument time and increases multiplexing capabilities.**

Figure II-11 illustrates the multiplexing capabilities of Scheduled SRM for the analysis of highly complex samples. We analyzed a set of almost 300 heavy-labelled standard peptides by monitoring almost 1000 transitions. To analyze this sample using a regular SRM method with a 1 hour run time and with the typical set of parameters described above, it

would take 7 independent runs to analyze all the targeted peptides whereas with a time scheduled SRM method all the peptides could be analyzed in a single run. This is possible using 2,5 minutes time windows but this requires a highly reproducible liquid chromatography system.

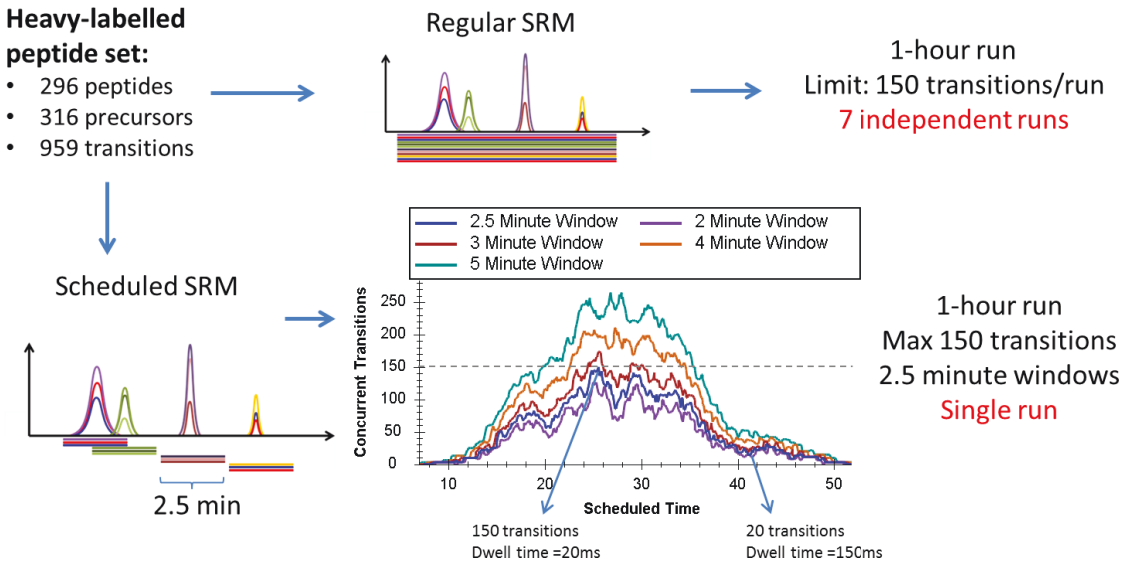



**Figure II-11: Illustration of the multiplexing capabilities of the Scheduled-SRM approach.**

When developing an LC-SRM assay to answer a given biological question, it is of crucial importance to balance the instrument scanning times to reach best quantification performances. Developing a SRM method consists in finding a trade-off between the accuracy, the sensitivity and the multiplexing (Figure II-12). These parameters depend on the instrument's user-defined scan speeds. They have to be adapted to the purpose of the quantification. SRM can be used as a discovery tool to screen several proteins. In this case the sensitivity and quantification accuracy are reduced in favor of the multiplexing capabilities. In a context of the need of precise quantification, then the multiplexing capabilities are reduced in favor of a higher sensitivity and accuracy.

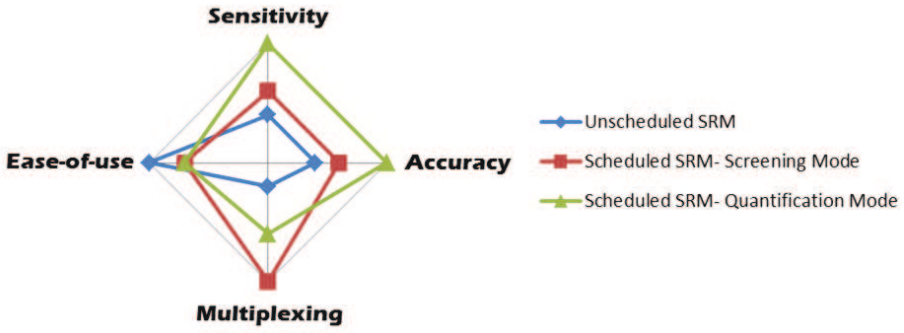



**Figure II-12: Trade-off between sensitivity, accuracy, multiplexing and easiness of use.**
These parameters depend of the instrument's user-defined scan speeds. They have to be adapted to the purpose of the quantification.

### *A.4. Use of isotope dilution for precise quantification*

The use of synthetic isotopically-stable internal standards enables the precise quantification of proteins. The most common approach is the use of the Stable Isotope Dilution (SID) method that consists in adding to all analyzed samples isotopically-labelled standards in the same known amount. The concentration of the targeted peptides can be measured by comparing the signals from the synthetic heavy-labelled and endogenous unlabeled species (Figure II-13). The endogenous and heavy-labelled forms of a peptide have the same physicochemical properties and differ

only by their mass. The standard samples are also used to correct for LC-MS signal fluctuations and sample preparation biases. The light-to-heavy ratio is then used to inform of the peptide's concentration.

Since Trypsin is the most commonly used enzyme in proteomics workflows, the targeted peptides are often tryptic peptides. The heavy-labelled versions of these peptides are thus commonly labelled on C-terminal Lysine or Arginine residues ($^{13}C$ and $^{15}N$).

**Low-quality Crude synthetic peptides:** Several quality ranges exist for synthetic standards. The quality of the internal standard depends on the objective of the quantification. A screening quantification for putative biomarkers between several different samples each representing a particular condition (healthy vs. diseased, time-course variations, different chemical perturbations…) does not necessarily need accurately quantified internal standards. Crude synthetic standards of low-purity are frequently used in these cases as they are cheap to produce. They are also very useful to optimize the LC-SRM methods. We mostly used crude PEPotec peptides from Thermo Fischer.

**High-quality synthetic peptides:** When a project needs a precise relative quantification (absolute quantification) highly purified and accurately quantified internal standards are required (AQUA peptides) [90]. However the high cost of these standards remains a drawback and this often limits their use in highly multiplexed assays.

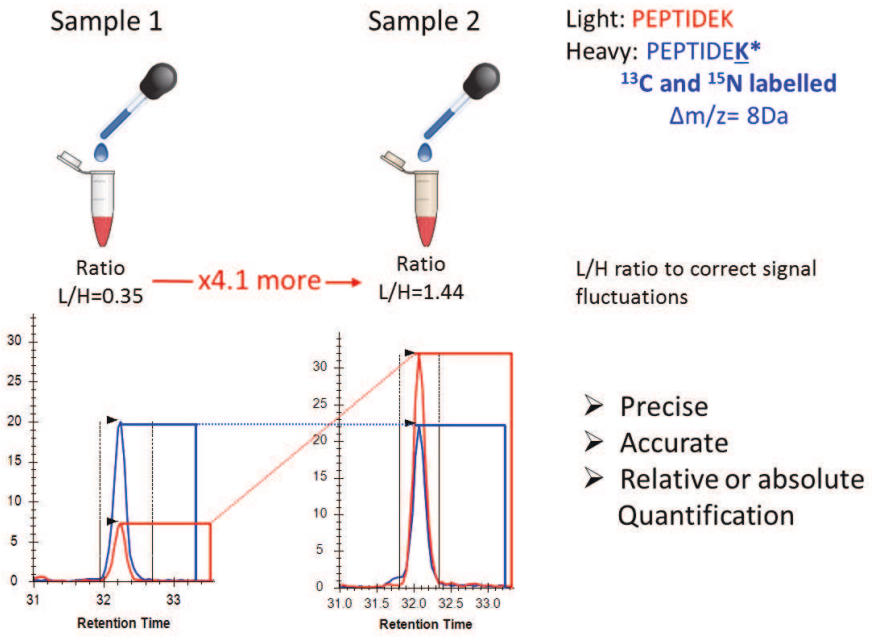



**Figure II-13: Use of isotope dilution for precise quantification.**

If heavy-labelled standard peptides are used it is important to spike the peptides as early as possible to correct eventual biases introduced during sample preparation. However, when using peptide level standards, some confounding errors cannot be corrected, namely incomplete or nonspecific enzymatic digestion [91, 92], artifactual chemical modifications of the targeted peptides or incomplete solubilization of the standard peptide.

**Concatemer proteins:** An alternative approach to AQUA peptides consists in using QconCAT (Quantification conCATamer)[93]. In this strategy, artificial proteins that are concatamers of tryptic peptides originating from several proteins are built. This method consists of the design, the expression in a media enriched in isotopically-labelled amino acids of a concatemer protein combining all targeted peptides and a purification step. The advantage of this approach is that this heavy labelled protein can be spiked at an early step of the sample preparation protocol thus allowing correcting for biases introduced at these steps. However a drawback of this approach is that the efficiency of

the tryptic digestion step highly depends on the protein sequence and might thus differ between the QconCAT protein and the endogenous proteins. Mirzaei et al. showed that the efficiency of expression of QconCAT proteins depends on the arrangement order of concatenated peptides and changing this order can help to increase the yield [94]. This work also showed that peptides that are difficult to synthetize chemically can alternatively be generated by the QconCAT method. This shows that there is no fit-for-all method and that it is the proteomist's task to choose the method that better fits to the objective of the quantification.

**Full-length stable-Isotopically labelled protein:** Alternatively full-length stable-Isotopically labelled proteins can be used. The PSAQ (Protein Standard Absolute Quantification) method was introduced by Brun et al. in 2007 [95]. All Lysine and Arginine in the protein are labelled with $^{13}C$ and $^{15}N$ stable isotopes. This method has the advantage to recreate the same sequence between the internal standard and the endogenous protein. All tryptic peptides can thus be used for the quantification of a protein. However, this strategy remains very expensive and requires a long development time for the synthesis, the purification and the precise quantification of these proteins. Additionally PTMs are not taken into account as they are not present in the internal standard protein which can also result in a different behavior between light and heavy proteins.

All in all, targeted MS-based quantitative assays using internal standards are characterized by high specificity, high sensitivity, high multiplexing capability, and high precision which make this approach ideal for biomarker verification studies for instance.

## B. Parallel-Reaction Monitoring

### *B.1. Principle of PRM*

Parallel Reaction Monitoring (PRM) is performed using high-resolution high mass-accuracy instruments (Figure II-14). The most common instruments used for PRM are hybrid quadrupole Orbitrap (Q-Orbitrap) or quadrupole time-of-flight (Q-TOF) mass spectrometers [3, 96]. The approach resembles SRM on a triple quadrupole instrument as the targeted precursor ion is first selected at unit resolution by a quadrupole and is then fragmented in a collision cell. A full scan spectrum of all resulting fragments is acquired. Then the PRM traces for each transition (couple of precursor/fragment ions) are extracted in a post-acquisition step. The simultaneous measurement of all fragments of a peptide facilitates the method development steps as only the precursors' m/z ratios and chromatographic coordinates are needed. This is an advantage over SRM which needs to previously determine the best transitions for a given peptide. Also if interferences from the background matrix are found, in PRM it is possible to easily refine the transitions list.

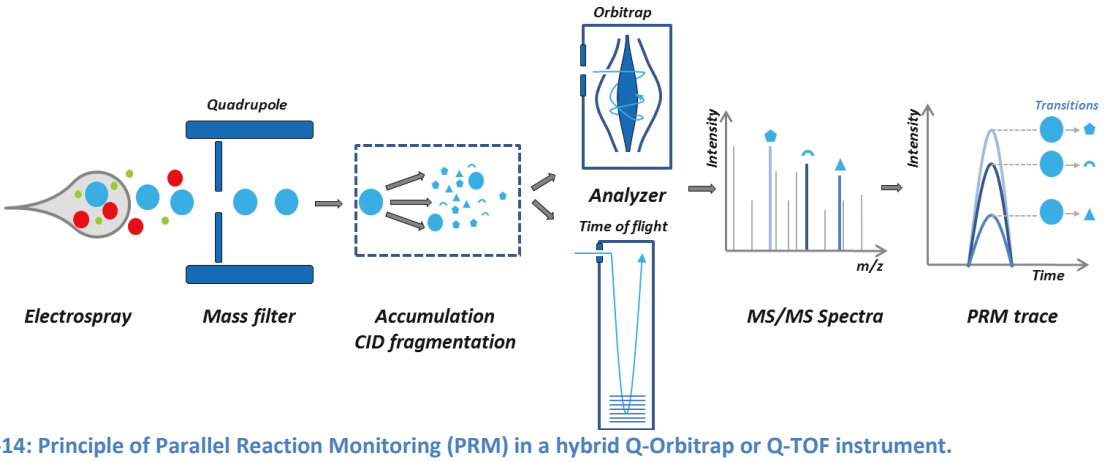



Figure II-14: Principle of Parallel Reaction Monitoring (PRM) in a hybrid Q-Orbitrap or Q-TOF instrument.

One additional advantage of PRM over SRM is the higher resolution of Q-Orbitrap and Q-TOF instruments compared to QqQ instruments. For example the AB Sciex Triple-TOF 6600 instrument can reach a resolution of 15k (in high sensitivity mode) or 30k (in high resolution mode) and hybrid Q-Orbitrap instrument such as the Thermo Scientific Q-Exactive Plus instrument can work at resolutions of 18k, 35k, 70k or 120k (@200 m/z). This high-resolution added to the high-accuracy measurements drastically increase the selectivity of the analysis. This higher selectivity can also mean a higher sensitivity as well, because the targeted signal is characterized by higher signal-to-noise ratio and is less likely to be interfered by biochemical background noise.

The development of a LC-PRM method follows the same steps as the ones for an LC-SRM method. However these are simplified due to the fact that there is no need to pick the best-responding transitions for each peptide.

### *B.2. Instrumental parameters for a hybrid Q-Orbitrap instrument*

The development of a targeted PRM method needs fewer parameters to be optimized. The necessary information to analyze a peptide is only its m/z ratio and, optionally, its elution time. This makes PRM significantly user-friendly and amenable to rapidly perform targeted quantitative assays on a limited list of targets. The optimization of individual transitions is not required. Additionally as for SRM, time-scheduled acquisition can also be performed in PRM methods.

The parameters to be set on a Q-Exactive plus instrument are the resolving power, the maximum injection time, the AGC target, the normalized collision energy and the scheduling parameters. The isolation width of the quadrupole is normally set at 2Da. The resolving power value is of critical importance as it will have an influence on the selectivity of the analysis but also on the sensitivity and on the multiplexing capabilities. Indeed, on an Orbitrap there is a trade-off between the resolving power and the acquisition time. There are four working resolution settings on the Q-Exactive plus: 18k, 35k, 70k and 140k at 200 m/z. The respective transient times are: 64, 128, 256 and 512ms. Thus, it is important to balance the instrument parameters to obtain the best sensitivity, selectivity and accuracy for the quantification.

An important feature of the Q-Exactive Plus is its capacity of parallel acquisition, i.e the accumulation and preparation of ions for injection into the Orbitrap during the analysis of the previous ion package in the Orbitrap. This is illustrated in Figure II-15. A. This feature results in faster scan cycles. Moreover, the transient time of the Orbitrap becomes the limiting factor of the analysis scanning rate.

To obtain a reliable quantification at least 8 to 10 data points are needed across a chromatographic peak. Therefore, for an average FWHM of 30 s, the cycle time must be shorter than 3 s. To fully optimize the instruments duty cycle, the maximum fill time of the C-Trap must be shorter than the transient time of the Orbitrap. Otherwise there is down time during which the instrument does not acquire any data. If this condition is met then the cycle time can be calculated with the following equation in the case of a scheduled PRM method:

$$if\ Max.Fill\ Time < Transient\ time$$

$$Cycle\ time = Number\ of\ precursors \times (Transient\ time)$$

Changes throughout the gradient — Changes throughout the gradient — Fixed value (Defined by the resolving power used)

An important characteristic of the Q-Exactive Plus is its multiplexing capabilities. The multiplex mode is illustrated in Figure II-15. B. In this mode 2 to 10 precursors are sequentially isolated and fragmented in the HCD cell and all the

resulting fragments are trapped and accumulated in the HCD cell. Then all the fragments are analyzed simultaneously in the Orbitrap. In this case the maximum fill time for each precursor has to be set shorter than the transient time divided by the number of co-analyzed precursors (multiplexing degree). The cycle time can be calculated with the following equation:

$$if\ Max.Fill\ Time < \frac{Transient\ time}{Multiplexing\ degree}$$

$$Cycle\ time = Number\ of\ precursors \times \left(\frac{Transient\ time}{Multiplexing\ degree}\right)$$

This mode can increase the total number of analyzed precursors in a single run. However it can impair the sensitivity and the accuracy of the quantification as the fill time of the c-trap is greatly reduced and the resulting data are convoluted spectra. This is discussed in further detail on section Part IVChapter IC. on page 100.

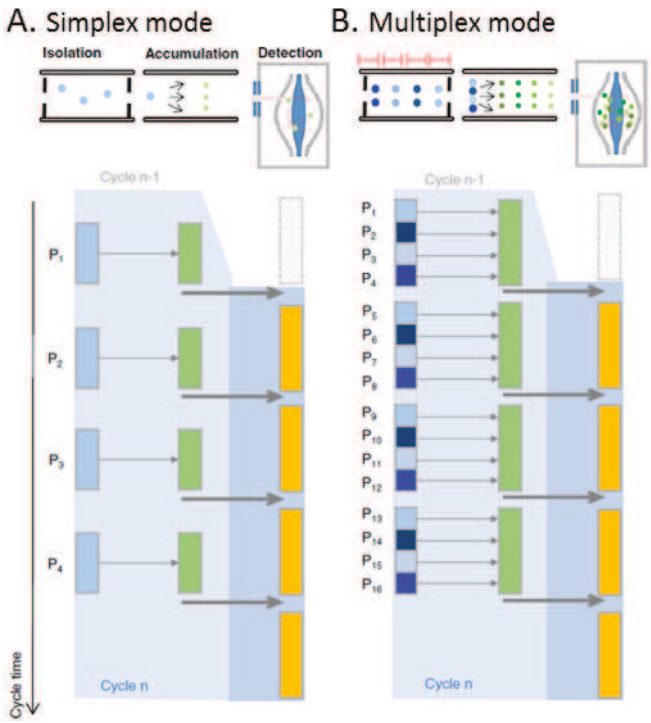



Figure II-15: Parallel Reaction Monitoring on a Thermo Scientific Q-Exactive Plus (Adapted from [97])

# Chapter IV Data-Independent Acquisition (DIA)

## A. Principles of DIA

Data-Independent Acquisition (DIA) is a group of newly developed acquisition methods that combine the benefits of the unbiased characteristics of DDA with the reproducibility, sensitivity and accuracy of targeted methods.

In DIA, the acquisition of MS/MS spectra is completely independent from the information of MS survey scans. In fact the MS scan is not required for some DIA methods. The acquisition of MS/MS spectra is not made to target a given precursor ion, making DIA an untargeted and unbiased acquisition mode. The instrument generates MS/MS spectra of all precursor ions isolated in a given predefined large isolation window, or eventually the entire mass range (Figure II-16). The MS/MS spectra are thus multiplexed data composed of fragments coming from all peptides coeluted and coisolated at a given time.

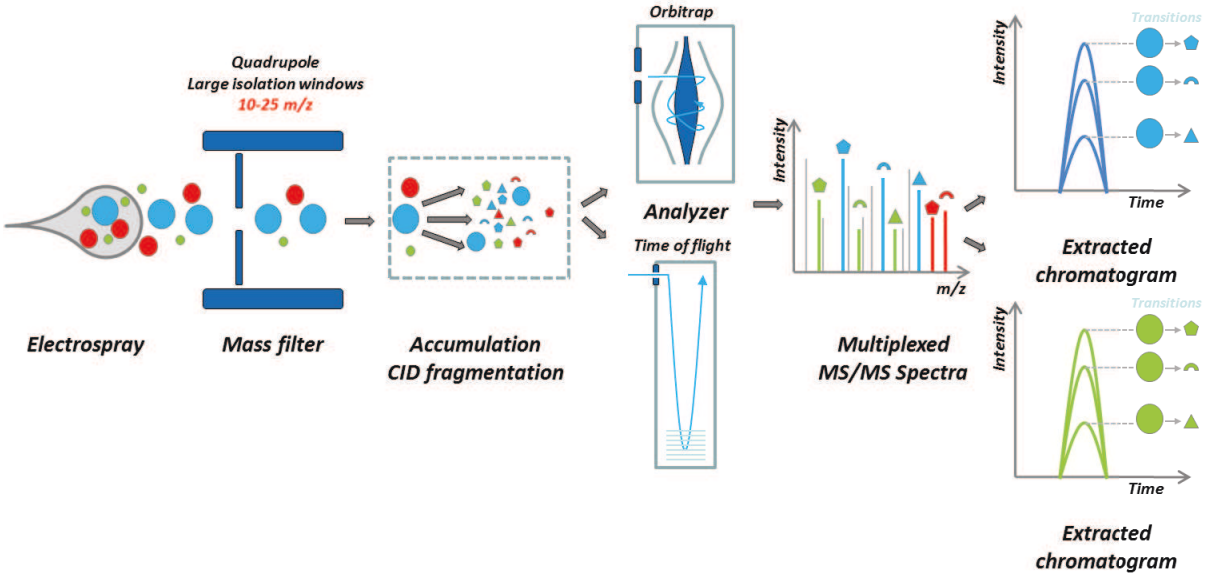



**Figure II-16: Principle of Data-Independent Acquisition**

This mode promises the comprehensive MS/MS sampling of all fragment ions of all peptides in a complex sample, with the only limitation of them being above the instrument's limit of detection.

DIA has emerged in recent years due to the extraordinary technological developments enabling faster scan rates, high resolution and reproducible LC conditions. DIA is executed using high-resolution, high mass-accuracy and high scan-rate instruments. The most common instruments used for DIA are hybrid quadrupole Orbitrap (Q-Orbitrap) or quadrupole time-of-flight (Q-TOF) mass spectrometers [7, 98].

## B. The development of DIA

The first proof-of-principle study concerning DIA was done by Purvine *et al.* in 2003 [99]. In this study two analyses of the same sample were done using low and high nozzle-skimmer voltages, the first produced mainly MS/MS spectra of precursor ions and the second of fragment ions of multiple precursor ions. They also proposed the idea of matching extracted chromatographic peak shape of precursor ions to those of suspected fragment ions. This method was termed shotgun-CID. In 2005, a refinement of this approach was proposed and commercialized by Waters under the name $MS^E$ [100]. This approach was executed in a hybrid Q-TOF instrument which scans the full m/z range with alternating low and high collision energies (Figure II-17.A).

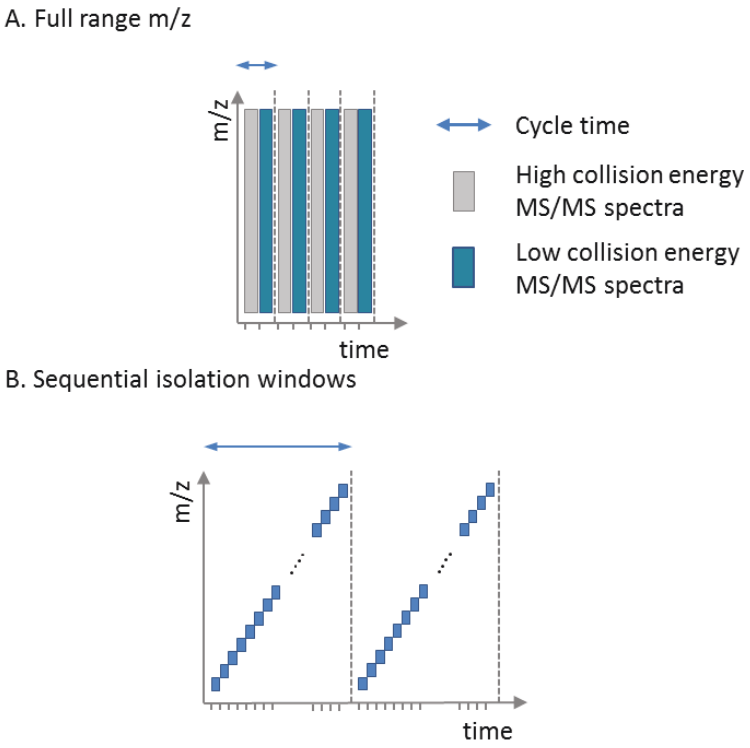



**Figure II-17: Signal acquisition scheme of data-independent acquisition**

Another school of thought developed in parallel as in 2004 Venable et al. proposed a method based on the sequential analysis of small isolation windows (Figure II-17.B) on a linear ion trap mass spectrometer and introduced the term data-independent acquisition [101]. In this study sequential twenty 10 m/z isolation windows were analyzed to cover only the 900– 1100 m/z range. The gain of signal-to-noise ratio was established when comparing MS2 to MS1 extracted chromatograms. In 2009 Panchaud *et al.* proposed a method called precursor acquisition independent from ion count, PAcIFIC, which consisted of using 2,5 Da isolation windows to reduce the complexity of MS/MS spectra. However multiple runs were needed to analyze a single sample (67 injections during 5 days of analysis) [102]. In 2012, the Fourier-transform all reaction monitoring (FT-ARM) was proposed by Weisbrod *et al.* which consisted in using broad 12 or 100m/z isolation windows and profit from the ultrahigh-resolution of FT-instruments to match fragment ions to *in silico* calculated peptide fragment ions [103].

Even if these strategies offered quantification advantage, the use of sequential isolation windows was not commonly adopted. This was due to the low scan rate of the mass spectrometers at the time and due to the complexity of MS/MS spectra which made the identification of peptides difficult.

With the commercialization of a faster scanning, more accurate and high-resolution instrument, Gillet et al. proposed the Sequential Window Acquisition of all Theoretical Mass Spectra (SWATH MS) method. This was executed on a Q-TOF instrument that scanned sequential 32x25Da isolation windows to cover the 400-1200 m/z range [6]. The identification and quantification of a peptide of interest is done by targeted signal extraction approach. The signal of a given peptide is looked for at its corresponding SWATH window and around the retention time window informed by a spectral library. Since then new method developments and DIA strategies have been developed and will be presented and discussed in Part IVChapter IE. .

Even if several new studies have used DIA strategies for quantitative proteomic studies with high sensitivity and high precision [104, 105], several challenges still remain. Since DIA generates very complex convoluted MS/MS spectra, the data analysis is non-trivial. Indeed, even if high resolution and high accuracy instruments are used, due to the wide

isolation windows of DIA experiments this approach is still vulnerable of co-eluting interferences that handicap the quantification. Moreover several software tools exist but a direct side-by-side assessment of these tools have not been done yet. The best practices for protein quantification using DIA data are still to be defined. Finally, the real gain in terms of sensitivity, dynamic range and selectivity are still to be proved. This will be discussed in further details in Part IVChapter IE. .

### C. DIA assay workflow

The steps required for the development of a DIA assay are shown in Figure II-18. The steps required to develop a DIA are different from those for targeted proteomics. The quantification analysis starts with a hypothesis, generally less specific than those for targeted quantification workflows. Then the LC-MS instrument parameters are optimized according to the sample and the instrument used. After the acquisition is done, the data is analyzed using prior knowledge concerning the analyzed sample. A targeted signal extraction approach is most commonly used. The coordinates of the targeted peptides are inferred by the use of spectral libraries, obtained by previous discovery analyses or using publically available databases. Then the confirmation of the presence a given peptide is necessary to be able to reliably quantify it. All these steps and the methodological developments towards improving this workflow will be discussed in detail in Part IVChapter IE. .

One of the major advantages of DIA is the possibility of reprocessing the same dataset to test new hypothesis without the need to reacquire the sample. For example, if a protein of interest was found to be up-regulated after a given stimulus, all proteins related to the pathway where the protein interacts can be quantified without the need to develop a targeted SRM method or time-consuming ELISA assays.

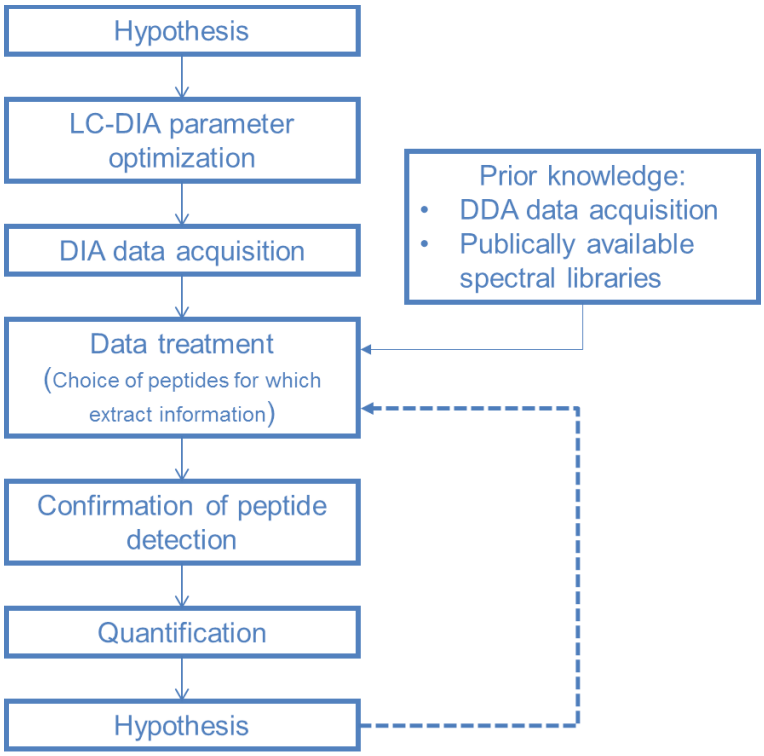



**Figure II-18: DIA assay development workflow.**

# Chapter V Fit-for purpose strategy for discovery or quantitative Proteomics

## A. Considerations for the choice of the analytical strategy

In order to choose the best fitting strategy to respond to a biological question, the following information is needed:

- The nature of the sample and the nature of the proteins of interest.
- The proteome coverage that needs to be analyzed (global or targeted).
- If a targeted method must be used: How many proteins are targeted and if the quantitative response needs to be relative or absolute.
- The MS instruments accessible for the analysis.
- How many samples need to be analyzed in a short-term and in a long-term.
- The budget accessible for the project.

Given the high number of quantification approaches now accessible to the proteomics community, it becomes the Proteomist's task to choose the strategy fitting the best to the purpose of the quantification. The proteomist's work consists in developing and optimizing the quantification method by finely tuning the LC-MS parameters in order to correctly balance between the needed sensitivity, the highest accuracy, the highest selectivity and the broadest protein coverage. But also keep in mind to reduce the cost of the analysis, the throughput needed in shot- and long-term.

## B. Mass spectrometers used during my doctoral work

Table II-2 presents the description of the MS instruments used in the projects described in this manuscript. For each instrument its performance is described for common operation conditions. A description of the application possibilities of each platform is provided.

Table II-2: Description of Mass Spectrometry instruments present in the laboratory

| Name | AmaZon ETD | Synapt G1 | Maxis4G | Impact HD | TripleTOF 5600 | TripleTOF 6600 | TSQ Vantage | Q-Exactive Plus |
|---|---|---|---|---|---|---|---|---|
| **Manufacturer** | Bruker Daltonics | Waters | Bruker Daltonics | Bruker Daltonics | AB Sciex | AB Sciex | Thermo Fisher Scientific | Thermo Fisher Scientific |
| **Analyzer type** | IT (3D) | Q-TOF | Q-TOF | Q-TOF | Q-TOF | Q-TOF | QQQ | Q-Orbitrap |
| **Resolution** | 3K | 9K | 40K | 40K | 30K (MS)<br>15K (MS/MS) | 30K (MS)<br>15K (MS/MS) | Unit | 17,5K to 140K at 200 m/z<br>70K (MS) and 17,5K (MS/MS)<br>70K (MS) and 17,5K (MS/MS) |
| **Mass accuracy** | 0,1 - 0,5 Da | 15 ppm | 10 ppm | 10 ppm | 15ppm | 15ppm | 0,7 Da | 5 ppm |
| **Mass range** | 3 000 m/z | 20 000 m/z | 20 000 m/z | 20 000 m/z | 40 000 m/z | 40 000 m/z | 3 000 m/z | 6000 m/z |
| **Q1 selection range** | - | 4000 m/z | 4000 m/z | 4000 m/z | 5–1250 m/z | 5–2250 m/z | 3 000 m/z | 6000 m/z |
| **Acquisition speed** | 4Hz | 1Hz | 10 Hz | 17 Hz | 20 Hz | 20 Hz | 50Hz | 13 Hz (17,5K @200 m/z) 7Hz (35K @200 m/z) 3 Hz (70K @200 m/z) |
| **Year of installation** | 2009 | 2009 | 2011 | 2013 | 2014 | 2015 | 2011 | 2014 |
| **Applications** | Discovery Proteomics (Highly fractionated samples) | Discovery Proteomics (Highly fractionated samples) | Discovery Proteomics (Highly fractionated samples) | Discovery and Quantitative Proteomics | Discovery and Quantitative Proteomics | Discovery and Quantitative Proteomics | Quantitative Proteomics | Discovery and Quantitative Proteomics |
| **Quantification possibilities** | Spectral Count | Spectral Count | Spectral Count | MS1 Filtering | PRM, MS1 Filtering and DIA | PRM, MS1 Filtering and DIA | SRM | PRM, MS1 Filtering and DIA |

## C. Challenges for Quantification

As stated before the proteome is an extremely complex sample due the high number of proteoforms that can be present for a single canonical protein [3], and also due to the broad dynamic range of protein abundance [4, 11]. Since the dynamic range of an MS instrument can reach 3 to 5 orders of magnitude depending on the acquisition mode, complex sample preparation workflows have to be carried out in order to reduce the dynamic range or to decomplexify the sample. However this can introduce biases in overall protein recovery.

For quantification the data reproducibility and accuracy is very important. The LC-MS systems must be regularly monitored to avoid confounding errors due to instrumental perturbations. As well, the sample preparation protocols must be validated and optimized to minimize sample losses and the inherent variability that is associated with complex and multi-step sample preparation protocols.

In bottom-up Proteomics the inference of protein quantitative information from peptide information is still a challenge and a “best-practice” method is still not defined [35]. Since the most common approach for protein identification highly depends on the protein database, this must be as complete and as adapted to the sample

analyzed as possible. However due to the enormous number of proteoforms that can be expected, a consensus database will never completely represent the protein content of a sample. This has consequences in the quantification of a protein as the analysis of a given protein could only be partial. Likewise, a challenge of proteomics is to determine post-translational modifications and the added difficulty of the quantification of PTMs comes from the fact that they are dynamic.

The quantification of a proteome can only benefit from extending the reach of bottom-up proteomics to comprehensively analyze the proteome. Understanding the limitations of the methodologies is absolutely necessary to propose solutions to them. These limitations can originate from analytics or informatics.

In this context Proteogenomics can have a significant beneficial impact on the quantification of a proteome. This field of science will be presented in the next part.

# Part III State of the art of proteogenomics

## Chapter I Proteogenomic analysis

### A. Introduction

Protein identification is carried out by search algorithms that compare the experimental lists of precursor and fragment masses obtained by LC-MS-MS/MS to a list of masses produced *in silico* from a protein database. This procedure relies on the completeness of the protein database used by the search algorithm. If a peptide sequence differs, even slightly, from the peptide sequence present in the database, it will not be identified. Yet databases used in bottom-up proteomics are often incomplete.

Additionally it is necessary to control the database and the search space size. The more candidates there are to explain a given spectrum, the lower the specificity of the identification and the higher the risk to have false positive interpretations. The reliability of protein identification relies on the correct balance between the database size and its completeness in regard with the proteome being analyzed.

For non-model organisms, protein databases are populated by homologous protein sequences from related organisms. This allows the identification of highly conserved peptide sequences but this approach does not allow the identification of organism-specific variant peptides.

Furthermore not all protein-coding genes are known and thorough gene annotation is far from complete. Novel protein coding genes are still being identified, even for the human genome [106]. Errors in gene annotation can propagate errors in protein sequences, such as incorrectly defining open reading frames and badly identifications of translational start sites, thus making identification of MS spectra impossible [68]. Furthermore, a single gene can give rise to many different biomolecules, for example splice variants [107], and these biomolecules might not be present in the reference database. Furthermore even if an extremely important amount of genomic, transcriptomic and proteomic knowledge is already available, the challenge remains to integrate all this big-data towards the improvement of peptide identification without compromising sensitivity and specificity.

To identify peptides that are not present in the database several bioinformatic methodologies have been developed, such as variant-tolerant sequence-tag based algorithms [108, 109] and *de novo* sequencing [110]. Variant-tolerant sequence-tag based algorithms search to identify small sequence tags, of 3 to 5 amino acids, in a given spectra and find candidate peptides in the protein database that possess the sequence-tag allowing a mutation in the peptide sequence. These techniques enable the identification of protein sequence variants or splice isoforms. However, they require large computational power and are prone to error when used in large-scale studies. Additionally these approaches need a reference protein database for peptide identification. Even if they allow a small change in the peptide sequence compared to the one present in the database, they still depend on the completeness of the database. As a response to this problem the field of Proteogenomics has developed.

## B. Proteogenomics

### *B.1. Definition*

Proteogenomics is a field at the interface between Genomics and Proteomics. At the early stages of this science, proteogenomics aimed at refining genome annotations from protein analysis [8]. Now the meaning of Proteogenomics was extended to all applications integrating multi-omics data in order to unify genome information and measures of abundance and characterization of proteins. This enables a better understanding of a biological system and provides information that could not be obtained otherwise by the isolated analysis of DNA, RNA or proteins [9, 111, 112].

Several factors have helped the fast development of this field:

- The expansion of Knowledgebases and data repositories.
- The remarkable technical development in sequencing techniques (next generation sequencing (NGS), exome sequencing (WES), RNA sequencing (RNA-seq), ribosome profilling).
- The appearance of proteogenomics bioinformatics tools.
- The developments of mass spectrometers with high scanning rates, high resolution, high sensitivity and high mass accuracy.

### *B.2. Types of novel peptides identified*

Proteogenomics allows the identification of novel peptides that would otherwise be missed by a classical proteomics approach. Several studies have shown the capability of proteogenomics to identify sequence variants and polymorphisms [11], small open reading frames [113], and proteins arising from annotated genes as novel protein-coding genes, pseudogenes [114], long non coding RNA (encoding polypeptides) [115] and gene fusions (Figure III-1.A). When using RNA data it is possible to identify peptides mapped to alternative splicing, RNA edits, and regions annotated as introns, exon boundaries and untranslated regions (3'-UTR and 5'-UTR). Also there is the possibility to identify out- of-frame peptides arising from alternative open reading frames.

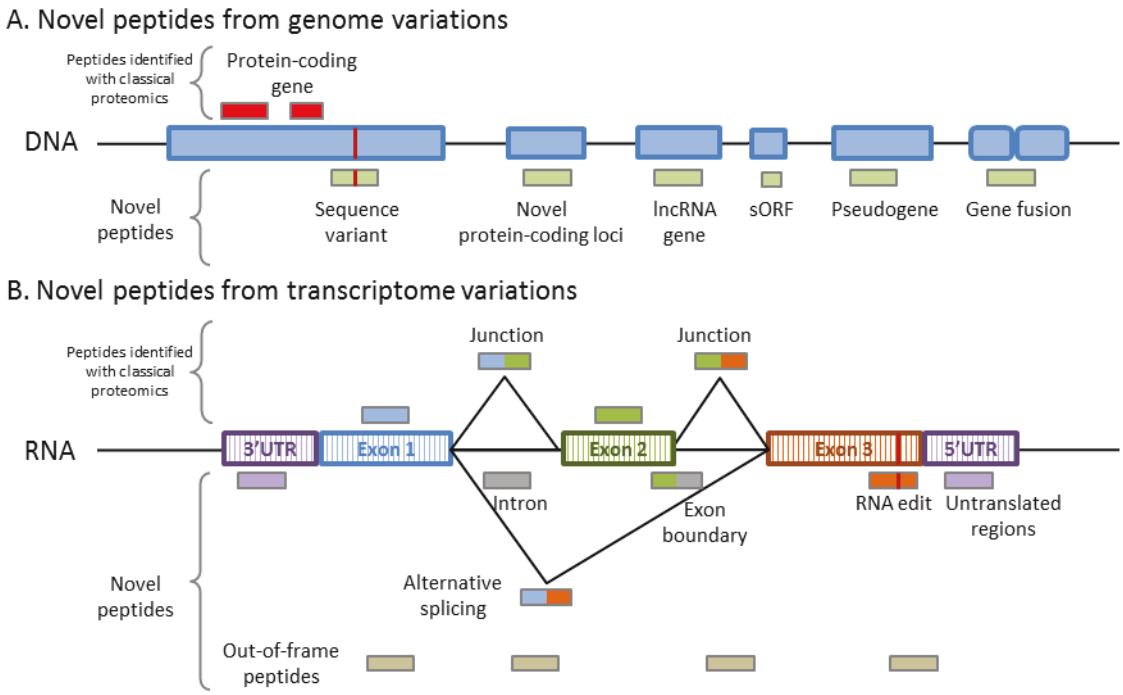



**Figure III-1 : Overview of variant peptides that can be identified in proteogenomics.**

## C. Tools for proteogenomics

### *C.1. Database customization using public resources*

Proteogenomics can improve protein identification by creating customized protein sequence databases using genomic and transcriptomic information. Different methodologies exist to use the information already provided by the scientific community. Table III-1 lists the most common resources for genomic, transcriptomic and proteomic annotations.

The most direct approach is to use a six-frame translation of the whole genome. This approach does not depend on gene models and contains all possible proteins with the exception of peptides mapping the exon junction regions [116]. The major drawback of this approach is the large size of the database and the large amount of background noise. For example, the translation of the whole human genome is 70 times larger than a curated reference database. However, only 2% of the genome is protein-coding. This approach drastically increases the search space. Consequently there's a high risk of false positive and false negatives. Also it requires a great amount of computational power.

Some approaches have been developed to circumvent this problem. The use of multi-stage searches with multiple databases, fractionated by chromosomes for example [117], has been proposed. Other methods use a prior step of peptide fractionation and several customized databases based on predicted physico-chemical characteristics, such as pI [118]. Approaches using *ab initio* prediction algorithms to identify protein-coding regions have been used to reduce the database size [119]. The use of exon prediction has the advantage to account for splicing events. A database containing theoretical exon-exon junctions accounting for all possible combinations can be created in order to identify novel slicing variants [120]. Even if when using these methods the computational efficiency can be increased, large amount of background noise are still generated, as there is no proof of the transcriptional evidence of these predicted biomolecules.

Using data from annotated RNA transcripts - using GENCODE [121] or RefSeq [66] databases for example – to make a three-frame translation it is possible to generate a protein sequence database. This enables the identification of out-of-frame peptides and alternative translation initiation sites. The database can also contain information of RNA transcripts annotated as long non-coding RNA [122] or as pseudogenes [123]. However, the problem of the size expansion of the protein database remains.

Single nucleotide variants or polymorphisms can also be included in a customized database. Databases listing these genomic variations exist (COSMIC [124], dbSNP [125]) and protein-centric knowledgebases also include annotations of sequence variants (Uniprot [69], Nextprot[70]). However, the number of reported SNPs is very large and the majority of them are rare. In the latest NeXtProt release (January 2016) 2,481,976 variants have been annotated for the entire human proteome. To reduce the number of SNPs, it is possible to work in a specialized subset of variants. For example, Li et al. created a database for single amino acid variations found to be related to human cancer [126]. Finally, when looking for SNPs one must keep in mind to make sure that the small mass changes found are not originated by a chemical modification that generates the same mass change.

| Name | Description | URL | Ref. |
|---|---|---|---|
| COSMIC | Catalogue of Somatic Mutations in Cancer. | http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/ | [124] |
| dbSNP | The NCBI dbSNP database of genome variation complements GenBank by providing the resources to build comprehensive catalogs of common genomic variations in humans and other organisms. | http://www.ncbi.nlm.nih.gov/SNP/ | [125] |
| ECgene | Genome annotation for alternative splicing. | http://genome.ewha.ac.kr/ECgene/ | [127] |
| ENSEMBL | Publicly available software system which produces and maintains automatic genomic annotation and integration of this annotation with other available biological data. | http://www.ensembl.org/ | [128] |
| GenBank | GenBank is a comprehensive database that contains publicly available nucleotide sequences for more than 300 000 organisms named at the genus level or lower, obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects. | http://www.ncbi.nlm.nih.gov | [64] |
| GENCODE | High quality reference gene annotation and experimental validation for human and mouse genomes. | http://www.gencodegenes.org/ | [121] |
| neXtProt | The human protein-centric knowledgebase provides a state of the art resource for the representation of human biology by capturing a wide range of data, precise annotations and fully traceable data provenance. | http://www.nextprot.org | [70] |
| NONCODE | A database of noncoding RNAs. | http://noncode.org/ | [122] |
| OMIA | Online Mendelian Inheritance in Animals - Catalogue/compendium of inherited disorders, other (single-locus) traits, and genes in 219 animal species. | http://omia.angis.org.au/ | [129] |
| Pseudogene.org | Comprehensive database of identified pseudogenes, utilities used to find pseudogenes, various publication data sets and a pseudogene knowledgebase. | http://pseudogene.org/ | [123] |
| RefSeq | The NCBI Reference Sequence (RefSeq) database provides curated non-redundant sequence standards for genomic regions, transcripts (including splice variants), and proteins. | http://www.ncbi.nlm.nih.gov/RefSeq/ | [66] |
| UniProt | The Universal Protein Resource (UniProt) is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot and TrEMBL. | http://www.uniprot.org/ | [69] |

**Table III-1 : Most common databases used in proteogenomics studies**

# Chapter II N-terminomics analysis

## A. The importance of N-terminomics

As seen above, pre- and post-translational modifications give rise to a plethora of proteoforms originating from a single gene. These proteoforms can differ by amino acid composition, chemical modifications and size. The start of a protein, the N-terminus, is a highly critical protein characteristic that has a great impact in protein function, protein stability and its localization in the cell. The protein N-terminus can contain signal peptides that direct protein translocation or activation. For example, in mitochondria and chloroplasts transit peptides guide proteins specifically to these subcellular locations. Once the proteins reach their destination the transit peptides are often cleaved. Furthermore, some proteins need proteolysis to reach maturation. For example, trypsin is synthetized as its precursor inactive form Trypsinogen. The proteolysis causes a slight rearrangement of the protein structure that activates the protease site. Additionally proteolysis can also regulate biological processes such as removal of N-terminal methionine and protein degradation. Protein truncation is an irreversible modification, and as such, it can have a direct effect on the phenotype. Proteases exist in all domains of life and constitute one of the largest enzyme families in humans [130]. 2% of the human genome codes for proteases but their respective substrates and functions have not been fully characterized [131].

N-terminal post-translational modifications also play a role in biological processes [132]. N-α-acetylation is a very common modification. It is present in more than 50% and 80% of yeast and human cytosolic proteins [22, 133]. It has been shown that this modification is essential for cell viability and survival but the reasons for this are yet to be found. Some studies link N-terminal acetylation to protein protection or specific signaling for degradation, protein delivery and localization and complex formation [134].

In the proteogenomics context, studying the real start positions of proteins has an interest in the refinement of database annotation. The majority of protein databases are obtained from prediction and translation of DNA sequences. For example, the database of all species in UniprotKB/Swissprot (version 2016_02) contains only 27, 1% of proteins with evidence at protein or transcript level.

## B. State of the art of N-terminomics

Multiple strategies exist for the analysis of proteins N-termini In a bottom-up proteomics experiment highly complex samples are analyzed and N-terminal peptides are greatly outnumbered and overshadowed by internal peptides. In a common DDA experiment only the information of the canonical protein N-terminus present in the database can be obtained. To be able to study specifically, with high confidence and sensitivity, the exact position of the protein N-terminus after processing, enrichment and tagging protocols are necessary. I listed in the following sections recent developments in N-terminomics separated by positive (i.e. concentration of N-terminal peptides) and negative selection strategies (i.e. depletion of internal peptides). N-terminomics approaches are reviewed in the following references: [134-136].

### *B.1. Positive selection of protein N-termini*

Positive selection strategies for the analysis of protein N-termini often consist in a chemical or enzymatic derivatization of the protein N-terminus followed by trypsination and ending with a concentration and purification step. For this to be possible the α-amino group of the protein must be unmodified. An illustration of the methods presented here is available on Figure III-2 and a summary of their characteristics is presented in Table III-2.

A method called N-terminalomics by Chemical Labeling of the α-Amine of Proteins (N-CLAP) uses a protection step by phenylisothiocynate of all the primary amines in the protein (α-amines of protein N-termini and ε-amines of lysine side chain) followed by a single cycle of the Edman degradation reaction [137]. The new N-terminus is linked to a sulfo-NHS-SS-biotin group which can be cleaved by a reduction step. The protein is typsynized and then captured by immobilized avidin. A similar strategy uses a modified subtiligase enzyme to enzymatically biotinylate the α-amino group of protein N-termini [138]. In this approach the Biotin is bonded to a tobacco etch virus (TEV) cleavage motif. After digestion and immobilized avidin capture, the n-terminal peptide can be released by a highly specific TEV proteolysis step. N-terminal peptides keep a SerTyr tag in the N-terminal position. Possible drawbacks of these methods are the low enzymatic biotin reaction efficiency [139] needing thus high quantities of samples (typically 50–100 mg of a complex mixture per experiment) and the selection bias of the subtiligase specificity to the N-terminal sequence of the protein [140]. A recent development of the enzymatic biotynilation by the modified subtiligase enzyme method, enabled the targeted quantification of N-terminal peptides by LC-SRM [141].

Two other methods are based on a first step consisting of guanidination protection of ε-amino groups. Then the n-termini is labelled by chemical biotylination with sulfo-NHS-SS-biotin previous to digestion and capture by immobilized avidin [142]. N-succinimidyl S-acetylthioacetate (SATA) can also be used to label free protein N-termini and add a thiol group. Then after digestion, N-terminal peptides are covalently bound to a thiol-reactive resin. Internal peptides are washed away and N-terminal peptides are released by a reduction reaction [143].

Bland et al. introduced a recently developed N-terminomics technique[144]. It consists of a TMPP-derivatization step ((N-Succinimidyloxycarbonylmethyl)tris(2,4,6-trimethoxyphenyl)phosphonium) at the protein level followed by a digestion step. Reverse phase chromatography is used to eliminate the excess of TMPP and degradation by-products. Then TMPP-derivatized N-terminal peptides are captured onto magnetic beads having an anti-TMPP antibody.

All methods mentioned above enable the study of free protein N-termini. In order to study natural occurring acetylated N-termini different strategies must be employed. An example is the used of dimethyl labelling of the α-amino groups of internal peptides after protein digestion. Then acetylated N-terminal peptides are purified by SCX-SPE. Dimethylation does not change the overall charge of the peptide but it increases its basicity enabling thus the separation from acetylated peptides[145].

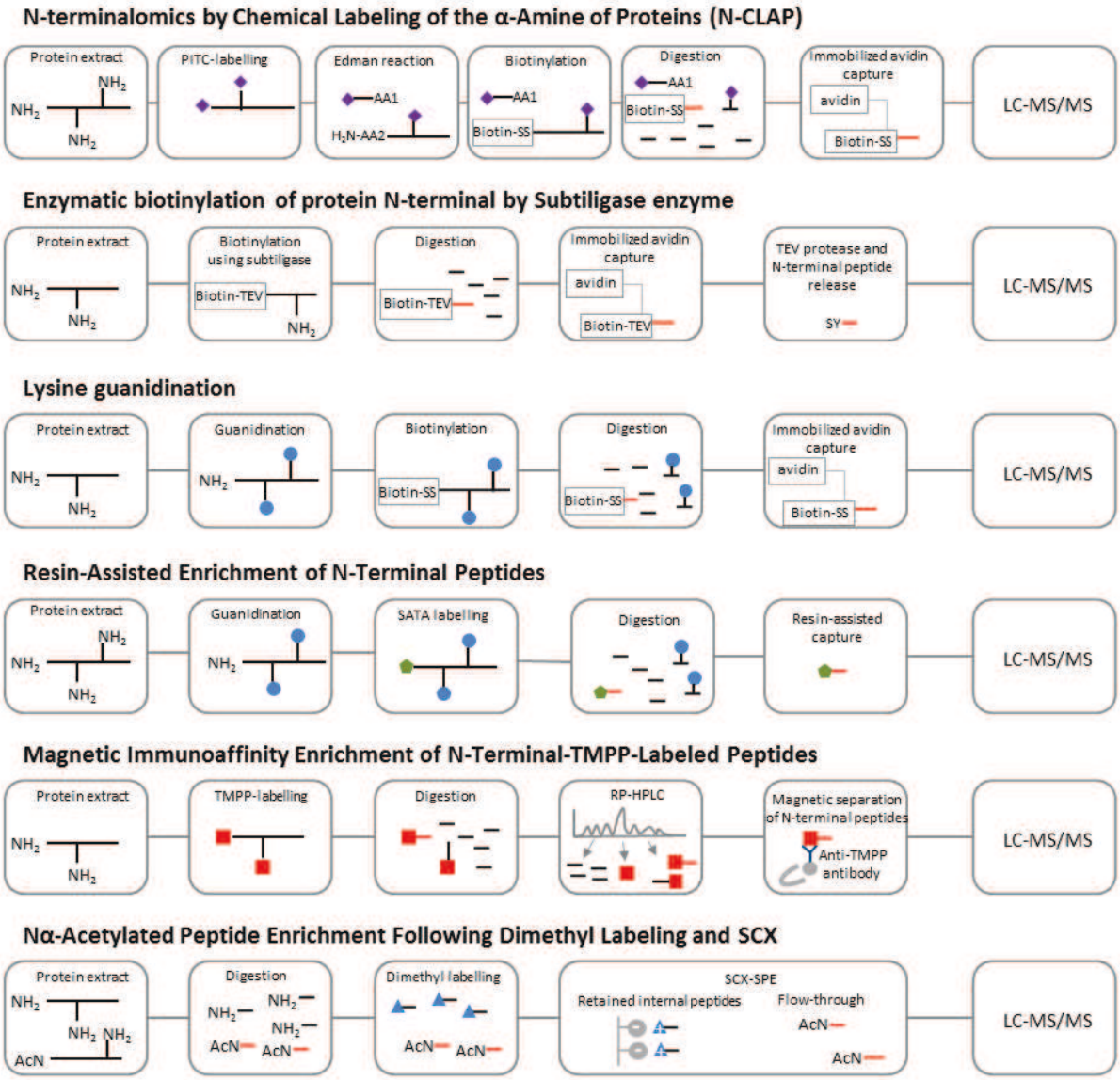



Figure III-2: N-terminomics strategies by positive selection of protein N-termini

### *B.2. Negative selection of protein N-termini*

Negative selection strategies consist in selectively depleting internal peptides. They enrich both free and modified protein N-termini thus providing a more comprehensive analysis of the N-terminome. These strategies consist in a chemical derivatization of the protein N-terminus followed by digestion. The internal peptides have now a free α-amino group and these will be used to selectively remove them. An illustration of the methods presented here is available on Figure III-3 and a summary of their characteristics is presented in Table III-2.

One of the most mentioned method is the combined fractional diagonal chromatography (COFRADIC). In it α-amino and ε-amino groups are acetylated, then proteins are digested and the resulting peptides are fractionated by reverse phase chromatography. For each fraction the α-amino group of internal peptides is derivatized by 2,4,6-Trinitrobenzenesulfonic acid (TNBS). Trinitrophenylated internal peptides are more hydrophobic than prior to the derivatization. During a second reverse phase chromatography step, internal peptides retention time is shifted to the right and separated from N-terminal peptides whose retention time did not change. Although this method enables a comprehensive analysis, the extensive fractionation can lead to sample loss and it requires significant instrument time (>100 fractions per run) [146]. The COFRADIC method also enables N-terminome relative quantification by introducing stable heavy isotopically labeled reagents during the workflow (such as the trideutero-acetylation of primary free

amines in the first stage to determine the degree of naturally occurring α-N-acetylation) or the metabolic labeling by SILAC approach [147].

Variations of this protocol exist where the chemical label can change. A method using TMPP to label the protein N-termini was recently reported [148]. After the acetylation of ε-amino groups of lysine the protocol is the same as COFRADIC. Another derived method is the charged-based fractional diagonal chromatography (chaFRADIC) where the protein N-termini are dimetylated, then proteins are digested and fractionated using SCX-HPLC [149]. Internal peptides are then trideutero-acetylated, decreasing thus their charge by one state of charge. A second SCX-HPLC step is carried out, N-terminal peptides having not change their charge state will elute at the same elution time.

Another widely reported method is the terminal amine isotope labeling substrate (TAILS) [150]. In this method protein α-amino groups are dimetylated before digestion with trypsin. Internal peptides are then captured by an amine-reactive polymer (hyperbranched polyglycerol-aldehydes, HPG-ALD). Internal peptides are removed by ultrafiltration and N-terminal peptides collected. The TAILS method also enables N-terminome relative quantification by using light or heavy stable-isotope dimethyl labelling to compare two samples or iTRAQ labelling to achieve an 8-plex comparison [151].

As the TAILS strategy, two other methods use a demethylation step prior to digestion and use the reaction between an aldehyde and a primary amine to remove internal peptides. The dimethyl Isotope-Coded Affinity Selection (DICAS) method also resembles this approach but utilizes an aldehyde matrix to deplete internal peptides (POROS-AL) [152]. The phospho tagging method (PTAG) uses glyceraldehyde-3-phoshate (GAP3) reagent to derivatize internal peptides which are subsequently depleted through binding to $TiO_2$ [153].

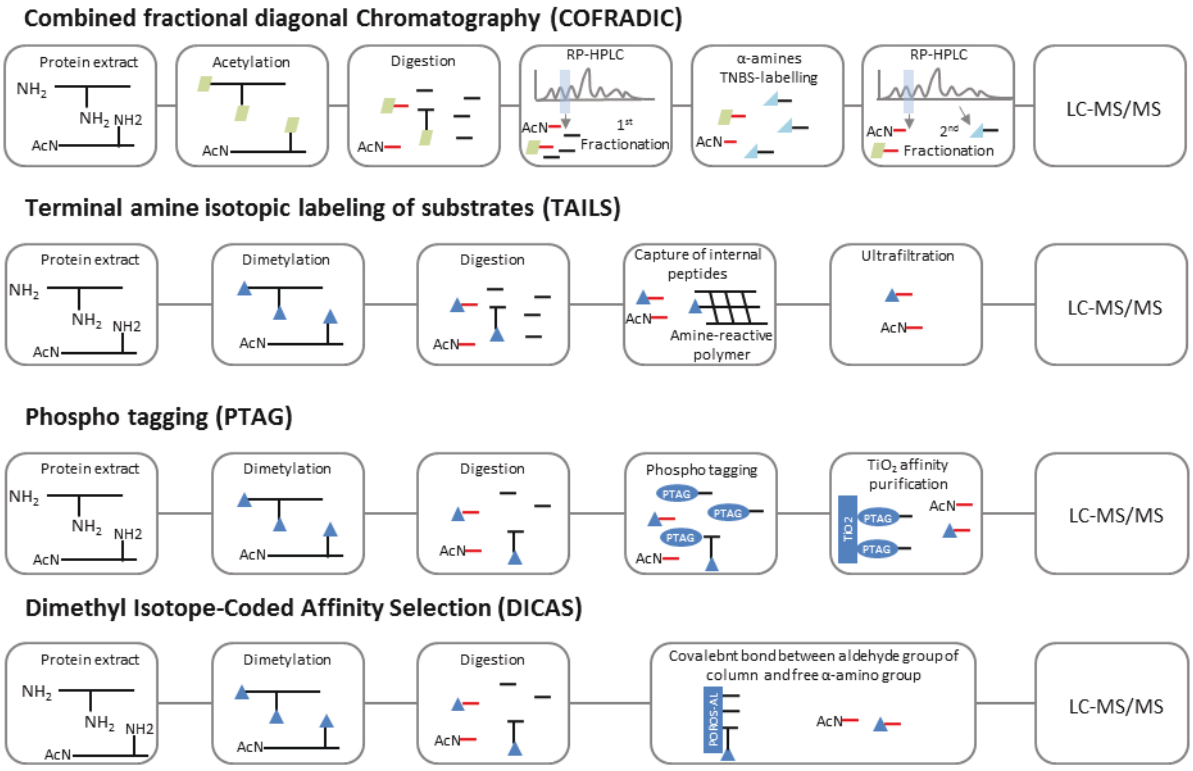



Figure III-3 : N-terminomics strategies by negative selection of protein N-termini

### B.3. Databases dedicated to the study of proteolysis

Several databases dedicated to the study of proteolysis exist. The MEROPS database classifies proteases from all species and provides information about their substrates, cleavage sites and inhibitors [154]. However access to original data is not available.

Other databases are specific to a certain N-terminomic approach. The data coming from the Gevaert laboratory that developed the COFRADIC approach is stored in the online protein processing resource (TOPPR) [155]. It provides information of proteolytic sites coming from studies on human and mouse samples. The Wells laboratory, which developed the enzymatic biotynilation by the modified subtiligase enzyme method, stores its data on the DegraBase [156]. This database contains proteolysis information coming from normal and apoptotic human cells. The Overall laboratory, which developed the TAILS method, created the Termini-oriented protein Inferred Database (TopFIND) database [157]. Contrary to the two previous databases this one is open to contribution from the scientific community but access to the original data is not possible.

### B.4. Current limitations

N-terminomic analysis is not an easy task, several approaches exist and each one is accompanied by its set of limitations. These will be discussed here and more details can be found on this subject in Table III-2.

As for proteomic analysis, one of the most difficult limitations to overcome in N-terminomic analysis is the large dynamic range of complex proteomes that hinders proteome coverage. To reduce the impact that highly abundant peptides have on lower abundant peptides, strategies using extensive fractionation can be used. Also protocols enriching subproteomes, such as organelle enrichment, can be used to increase specific proteome coverage.

Another shortcoming of some N-terminic strategies is the non-specific labelling of ε-amino groups on lysines. Even if α-amines and ε-amines have close reactivity, close attention must be put on finding experimental conditions in which derivatization specificity for α-amines is attained. Especially when using positive selection strategies. Furthermore positive selection strategies are also very dependent on the completeness of the α-amines derivatization. If incomplete this can impede the identification of N-terminal positions present in the sample.

The goal of enrichment is to separate internal peptides from N-terminal peptides. However enrichment is sometimes incomplete and can result in sample loss. For example, when using a non-covalent interaction the internal peptides are not fully depleted. For example when using a SCX-based protocols to enrich acetylated peptides more than 50% of identified peptides are internal peptides [145]. These number is reduced when using covalent interaction as in the TAILS protocol where 6% of all identification are internal peptides [132].

Distinguishing between naturally occurring proteolysis and protein degradation during sample preparation is also a challenge. To tackle this problem protease inhibitors should be used and the number of sample preparation steps at the protein level should be reduced.

The enzyme used for digesting proteins into peptides has a direct effect on the proteome coverage. Depending on the physicochemical properties of the N-terminal peptide (charge, hydrophobicity, size…) it can maybe not be detected even if it is present in the sample. Complementary proteases can be used to create different sequences for the n-terminal peptides and enlarge the coverage [158].

Finally the most challenging problem in N-terminomics is to understand the biological origin of a given N-terminal position. Fortelny et al. illustrated the problem of the current ignorance of the genesis of truncated protein N-termini [157]. In this work they took the N-terminal positions measured experimentally from the TopFIND database and tried

to match them to predicted N-terminal positions. These predictions came from predicted protease cleavage sites and predicted N-termini coming from alternative translation and alternative splicing. They found out that only 6% of all experimentally observed N-terminal positions can be explained.

This gap in knowledge illustrates the need for more enriched and curated databases. It also shows that the identification of N-terminal positions alone is no longer sufficient to answer biological questions. Quantification strategies should be developed in order to find specific cleavage sites induced by specific experimental conditions.

**Table III-2 : List and characteristics of some N-terminomic approaches**

| | Strategy | Quantity needed per experiment (in µg) | Unmodified N-terminus | Modified N-terminus | Relative quantification | Comments | Ref. |
|---|---|---|---|---|---|---|---|
| Positive enrichment of N-termini | N-terminalomics by Chemical Labeling of the α-Amine of Proteins (N-CLAP) | 100 | yes | | | - Peptides loose first amino acid after Edman reaction. The true N-terminal amino acid is not measured.<br>- Incomplete PITC-labelling reaction could lead to erroneous identification of the protein N-terminus. | [137] |
| | Enzymatic biotinylation of protein N-terminal by Subtiligase enzyme | 20000-100000 | yes | | yes | - Expensive patent-protected enzyme<br>- Selection bias of the subtiligase enzyme<br>- Large quantities of sample are needed | [138] |
| | Lysine guanidination | 2000 | yes | | | - labelling reagent NHS-SS-biotin can react with side chains of serine, threonine, histidine and unreacted lysine residues. | [142] |
| | Resin-Assisted Enrichment of N-Terminal Peptides | 100 | yes | | | - Very dependent of the efficiency of the guanidination reaction. Must not have overguanidination of α-amino groups or low efficiency at ε-amino groups.<br>- Needs thiol-reactive resin | [143] |
| | Magnetic Immunoaffinity Enrichment of N-Terminal-TMPP-Labeled Peptides | 3000 | yes | | | - Multiple preparation steps prone to sample loss | [144] |
| | Nα-Acetylated Peptide Enrichment Following Dimethyl Labeling and SCX | 50 | | yes | | - Loss of unmodified N-terminal peptides | [145] |
| Negative enrichment of N-termini | Combined fractional diagonal Chromatography (COFRADIC) | 1000 | yes | yes | yes | - Extensive fractionation prone to sample loss<br>- Instrument time consuming (>50 fractions/sample) | [146] |
| | Terminal amine isotopic labeling of substrates (TAILS) | 100 | yes | yes | yes | - losses due to non-specific binding to polymer | [150] |
| | Phospho tagging (PTAG) | 100 | yes | yes (except phosphorylated) | | - Losses due to non-specific binding to $TIO_2$<br>- Time-consuming preparation protocol (4-days) | [153] |
| | Dimethyl Isotope-Coded Affinity Selection (DICAS) | 100 | yes | yes | | - Needs one home-made packed cartridge per sample<br>- 8 to 10h of loading time | [152] |
| doublet N-terminal Oriented Proteomics (dN-TOP) | | 50-100 | yes | yes | yes | - Modified N-termini are not lost but they are not enriched. | [159] |

# Part IV Results: Quantitative proteomics developments and applications

## Chapter I Methodological developments for quantitative proteomics

### A. Optimization of targeted proteomics method development workflow

I developed a fast and efficient method for SRM assay development. The general workflow can be seen in Figure IV-1.

The development of a LC-SRM assay starts with defining the list of targeted proteins. This choice can be hypothesis-driven, based on previous knowledge or result from previous discovery proteomics studies.

The development of the assay is then followed by the development of, in one side, the sample-specific preparation method, and in the other side the development of the protein-specific LC-SRM assay. The detailed explanation of each step is presented in the following paragraphs (paragraphs A.1 to A.5) and a step-by-step walkthrough for SRM assay development is described in paragraph A.6.

#### *A.1. Sample preparation*

The sample preparation method needs to be optimized every time and adapted for each analyzed sample. It is important to keep in mind that in order to obtain a more precise and accurate quantification, the sample preparation method has to be a simple and non-fractionated protocol.

The development of a rapid and non-fractionated sample using Stacking SDS-PAGE protocol was discussed in Chapter I part B.2above.

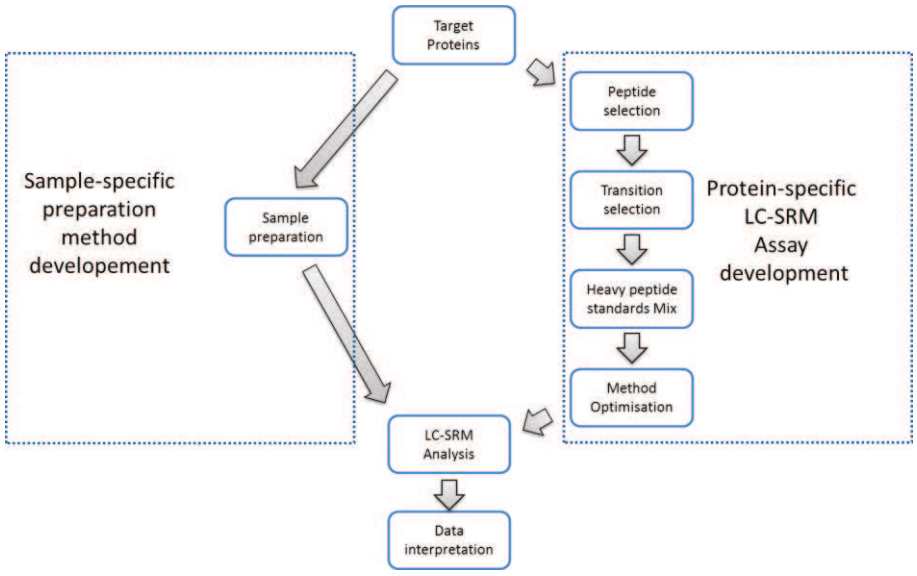



**Figure IV-1: LC-SRM method development workflow.**

### A.2. Peptide selection

To choose the best signature peptides to quantify the targeted proteins all possible tryptic peptides are filtered based on the peptide backbone sequence, on its unicity and on their physico-chemical properties that make them visible in LC-MS conditions and make them accurate proxys for the proteins of interest. This is illustrated in Figure IV-2.

First all possible peptides are filtered based on their backbone sequence. Tryptic peptides between 7 to 25 amino acids are chosen to fit within the m/z range analyzable with triple-quadrupole instruments (50-2000 range). Peptides below 7 amino acids are not well retained on a reverse phase column and long hydrophobic peptides are not well eluted and should be avoided. Peptides without any missed cleavages are preferred and possibly without ragged ends (two enzymatic cleavage sites next to each other) which produces irreproducible digestion yields. An advantage of SRM is that it enables the possibility to monitor interesting PTMs (acetylation, phosphorylation…). However it is important to verify that the surrogate peptides are not prone to unwanted modifications, such as Oxidation (Met, Trp), N-terminal pyroglutamic acid (N-terminal Glu under acidic conditions) and deamidation (Asn to Asp, Gln to Glu). This information can be found in Protein-centric Knowledgebases as Uniprot and NeXtprot. Also it is important to verify that the chosen peptides are not subject to other types of modifications (proteolysis, signal/transit cleavage sites…) or have possible sequence variants (SNPs).

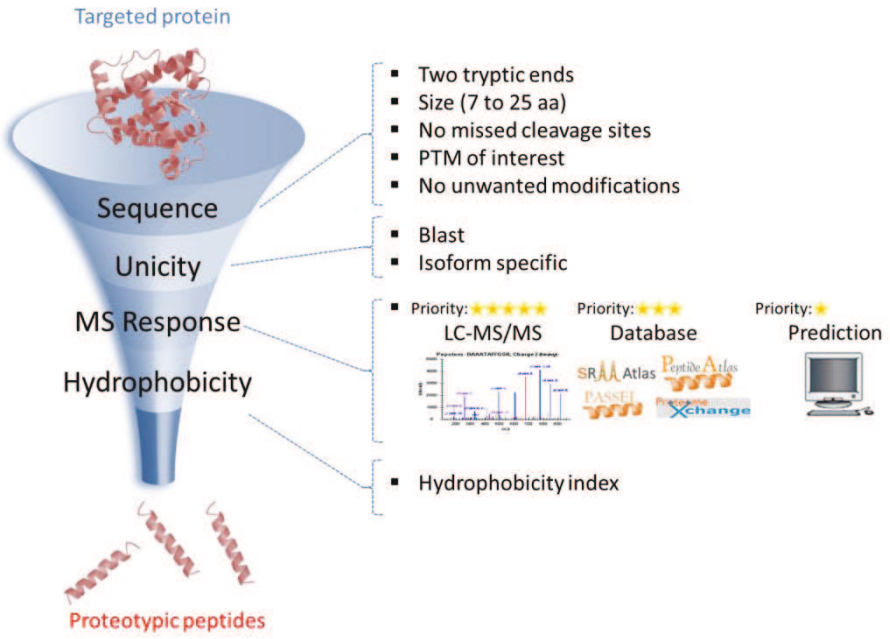



**Figure IV-2: Choosing proteotypic peptides.**
All possible tryptic peptides are filtered based on the peptide backbone sequence, on its unicity and on their physico-chemical properties that make them visible in LC-MS conditions.

If the sample preparation conditions are well controlled to avoid oxidation, Tryptophan-containing peptides can be chosen preferentially since the tryptophan residue is quite rare which provides a higher degree of specificity. Moreover proline-containing peptides can also be chosen preferentially as their fragmentation at the N-terminal side of the proline residue generates an intense fragment ion which can result in high sensitivity. It is also important to choose signature peptides well distributed across the protein sequence to obtain a good coverage of the protein.

Once a set of possible candidates is found the unicity of each peptide is checked by performing a BLAST against the whole proteome of interest. At this point the method can target specific isoforms or sequence variants.

An important step is to verify the peptide's response to LC-MS conditions. To choose the best peptide to quantify a protein a priority-based system was developed. A high priority is given to peptides that have already been seen and

respond well to LC-MS conditions in previous experiments. Data provided by the scientific community in databases or repositories (SRMAtlas [160], PeptideAtlas [161], ProteomeXchange [162], ProteomicsDB [163]). If no LC-MS data has been acquired for the surrogate peptides prediction algorithms exist to predict LC-MS behavior (PeptideSieve [164], ESP predictor [165], Detectability predictor [166], STEPP [167], PeptidePicker [168]).

If enough peptides pass the filters described above, an additional hydrophobicity filter can be used to select the best responding peptide. For example, the SSRCalc algorithm uses a sequence-specific model to predict peptide behavior in reverse phase liquid chromatography[169]. It serves in this case to eliminate possible too much hydrophilic or too much hydrophobic peptides.

The peptides that pass all the filters mentioned above are called proteotypic peptides. These peptides will be signature peptides to accurately represent the level of the targeted protein.

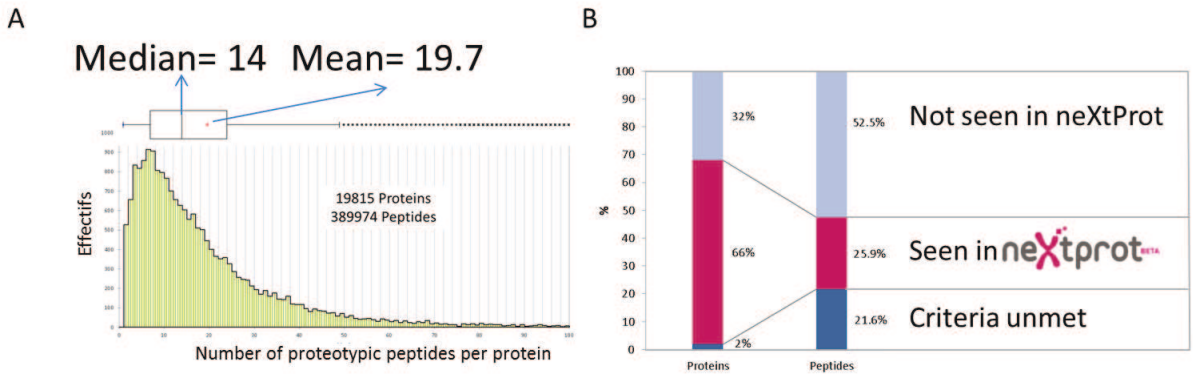



**Figure IV-3: Number of proteotypic peptides per protein in the entire human proteome.**
The entire human proteome was digested in silico and filtered in order to obtain all possible proteotypic peptides. A. The distribution of the number of proteotypic peptides per protein is shown. B. For all tryptic peptides between 7 to 25 amino acids in length the percentage of peptides that have been observed in LC-Ms experiments is given. The small percentage of observed peptides illustrates that not all of the potential proteotypic peptides are good candidates for quantification.

To illustrate the importance of these selection criteria we applied these filters to all the possible tryptic peptides between 7 to 25 amino acids of the entire human proteome (UniportKB/Swissprot version 2014_04). No shared peptide and no methionine-containing peptides were kept. We obtained with these filters 389974 tryptic peptides belonging to 19815 different proteins. For 98% of all proteins in the database a potential proteotypic peptide was found. The results showed that the mean number of potential proteotypic peptides per protein is around 19 and the median is at 14 (Figure IV-3). Seeing the distribution of proteotypic peptides per protein we could therefore conclude that there are enough peptides to quantify all proteins in the human proteome. However, to evaluate the peptides' response to LC-MS conditions, we looked for the number peptides that have been detected. We extracted this information from the high-quality and highly-curated database neXtProt [70]. As of April 2014, only 26% of all possible proteotypic peptides have been detected using LC-MS systems covering 66% of all proteins. This shows that not all of the potential proteotypic peptides are good candidates for quantification. That's why the highest priority has to be given to peptides already seen in LC-MS experiments. This ensures that the developed LC-SRM assay will give the best results.

### *A.3. Transition selection*

For each signature peptide, it is important to make the choice of the transitions that will provide the best sensitivity and specificity, i.e. the highest signal response and the lowest interfering signals.

Using CID fragmentation singly charged y- and b- fragment ions are chosen. Small m/z ratios are not specific enough since multiple amino acid combinations can result in the same mass. Fragment ions with high m/z ratios are preferred

over small fragments, since this gives more specificity to the measure, preferably those with m/z ratios higher that the precursor m/z ratio.

To choose the best transitions it is possible to calculate *in-silico* all the fragments m/z ratios for a given peptide and filter them with the criteria listed above. But this does not take into account how a peptide will respond to mass spectrometry and CID fragmentation. Public repositories and databases can be used to export spectra that can be used to choose the best-responding fragments (SRMAtlas [1], PeptideAtlas [2], ProteomeXchange [3], Proteomics DB [163]).

To facilitate the choice of transitions crude heavy-labelled peptides can be synthetized and use them to create spectral libraries from which the most intense fragment ions for each peptide can be picked. The library can come from different instrument configurations (IT, QqQ, Q-TOF or Q-Orbitrap). This method has the advantage of knowing how the peptide responds chromatographically (peak shapes, retention times) and in mass spectrometry (response factor, fragmentation pattern, charge states). Some small differences in the fragmentation pattern and the peptide charge distribution can exist between different instruments [170], however choosing the most intense transitions from a spectral library enables to fasten the SRM assay development step. The most intense transitions identified in Q-TOF, IT or Q-Orbitrap using CID fragmentation will be the same in triple-quadrupole instruments. The instrument-specific changes in relative intensities/fragmentation pattern will be discussed below.

Finally it is important to keep in mind to look for interferences using the real sample matrix, especially in complex digest samples where hundreds of peptides can coelute and have close m/z ratios which can impair the quantification.

### *A.4. Concentration-balanced mixture of synthetic heavy –labelled peptides*

To be able to correctly quantify the targeted proteins, the heavy-labelled and the light peptides have to be close in intensity. This enables a more accurate estimation of the endogenous protein abundance.

In our approach we use crude heavy-labelled peptides (Thermo Fischer PEPotec peptides) that are not fully purified and the exact concentration is not known. These peptides are used to optimize LC and MS parameters. First a mixture of all targeted peptides is prepared by empirically finding the dilution factors needed to obtain a mixture where all peptides are in detectable levels. Each peptide has to be in a sufficient amount to be easily detectable but not in an excessive amount to overshadow other analytes or have a bad behavior in LC-MS conditions (column saturation, ion suppression effects...).

A first estimation of the correct dilution factor for each peptide can be obtained when creating the spectral library. To do this, a series of LC-MS/MS runs with different dilutions factor are performed. If a peptide can be detected than it goes into the spectral library, if it does not a less diluted solution is prepared and analyzed until the peptide is detected correctly. This way the targeted peptides can be divided into categories of dilution factors to generate a concentration balanced mixture. In a latter step of the SRM assay optimization the levels of each heavy-labelled peptide are adjusted to match the levels of the endogenous peptides. Kuzyk et al. showed that the use of concentration-balanced mixtures improve the precision of the quantification [171].

### *A.5. LC-MS parameter optimization*

#### A.5.1 Retention time prediction

In a classical SRM experiment, all transitions are monitored throughout the whole run. Using this scanning mode we are limited in the multiplexing and the scanning time parameters. A way to enhance the instruments capabilities is to

use the Scheduled-SRM mode, where the transitions of a peptide are monitored only around its retention time. This allows optimizing the instrument's scanning time and also increasing the total number of transitions and peptides that can be monitored in single run.

Thousands of SRM signals can be monitored in a single run using scheduled SRM mode with small time windows [172]. But this is only possible if the retention times of all targeted analytes are precisely known and the chromatography system is precisely controlled and reproducible.

Several methods exist to determine the retention time of a peptide. We evaluated four of these methods by listing all the characteristics of what the ideal method for retention time determination should have. It should be straightforward and accurate, based on experimental data, not sample- and time-consuming, take PTM into account and easily allow method transferability between platforms. To decide which one would be implemented for our SRM method development workflow, we graded how each method performed according to the different criteria listed above (Figure IV-5). The most straightforward method was the direct injection method, which consists of making an LC-MS run using the same instrumentation and LC-MS parameters as for the real analysis conditions. However this method is sample- and time-consuming and has to be performed each time the LC conditions change. Repositories and databases can inform about the targeted peptide's chromatographic behavior in reverse-phase chromatography. However the stored values have to be adapted to the LC-MS system. This method is not as straightforward as the previous one, it also has to be adapted each time there is a change in LC conditions and the accuracy of the retention time estimation is not very good. Algorithms exist to predict *in silico* retention times based on peptide backbone sequences. An example of this is the SSRcalc algorithm that uses a sequence-specific model to calculate hydrophobicity indexes for each peptide based on their peptide sequences. Then peptides of known retention time are used to create a calibration curve from which one can predict a peptide's behavior in reverse phase liquid chromatography[169]. This method is fast and very useful to obtain a first idea of a peptide's retention time. Also this method does not consume a lot of sample since it only requires the analysis of a few peptides used to create the calibration curve. However, the accuracy is not very good thus long time windows are needed for scheduled SRM if this method is used. Also the algorithm does not take PTMs into account which impairs the accuracy even more.

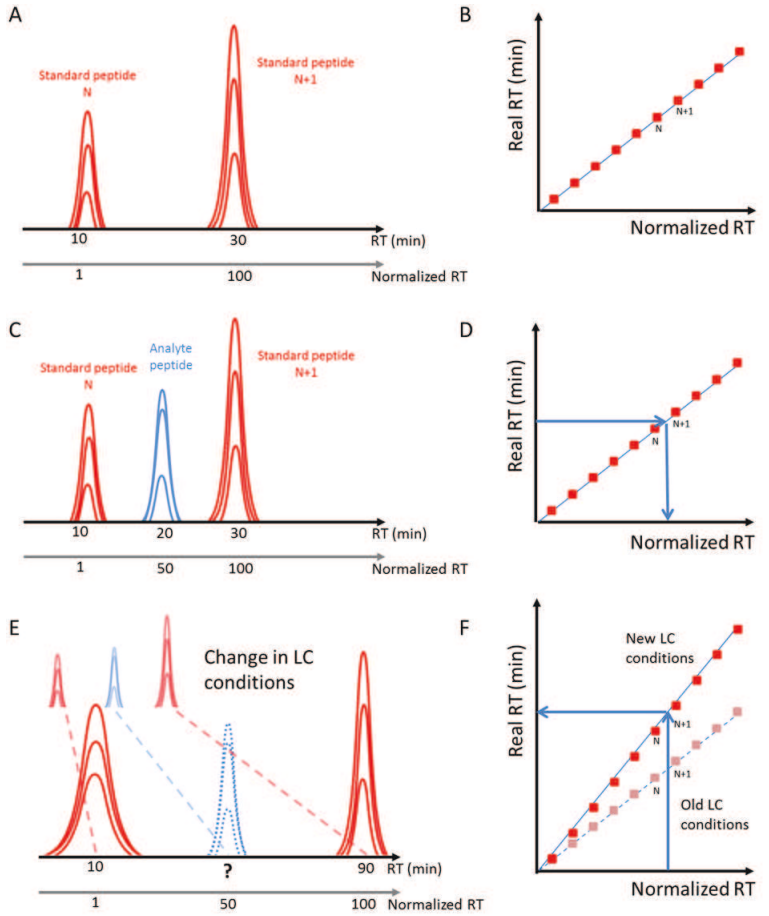



**Figure IV-4: Standard retention time peptides to predict retention times.**
Standard peptides that elute throughout the whole LC gradient are used (A). Each standard peptide is associated to a dimensionless normalized RT value to calibrate the linear equation between the two (B). Based on experimental data (C) normalized retention time values for the targeted peptides are calculated (D). These are relative to standard peptides and dimensionless values that are independent of the LC conditions. A database of these normalized values is created and once the LC conditions change (E), the retention times of all peptides can be easily predicted by recalibrating the linear equation in the new LC conditions (D).

The use of retention time standard peptides is a method that combines both experimental measurements and *in silico* prediction. For this method, standard peptides that elute throughout the whole gradient are used. In a given chromatographic condition a calibration curve is created to transform the real RT values into normalized RT values (Figure IV-4. A and B). Then each targeted peptide is analyzed in the same LC conditions and a normalized RT value is calculated for each one (Figure IV-4. C and D). These normalized values are relative to the standard RT peptides and are dimensionless values independent of the LC conditions. This step has to be performed only once for each peptide and the calculated values can be stored in a database. So when the LC conditions change the retention times of all peptides can easily be predicted by analyzing only the standard RT peptides in the new LC conditions to calibrate the calibration curve and, from it, predict the new retention times of each targeted peptide (Figure IV-4. E and F). We decided to implement this approach for determining RT values as it meets most of our expectations as illustrated on Figure IV-5. This method is rather straightforward. It does not require a lot of sample and it is not time-consuming. It uses experimental data and takes PTMs into account. It also facilitates method transfer between platforms and it is an accurate RT determination method.

Several standard retention time peptides exist and they can be present in various forms to evaluate not only retention time but also other steps as the enzymatic digestion yields, overall LC-MS performance or be used for mass recalibration. They can be in the form of a peptide standard kit( iRT-Kit [173] (Biognosys AG,Zurich, Switzerland), Peptide Retention Time Calibration Mixture (Pierce, Rockford, IL), MS RT Calibration Mix (Sigma-Aldrich, Poole, U.K.)),

in the form of recombinant proteins (DIGESTIF [174], Reversed-phase liquid chromatography calibrant (RePLiCal) [175]) or highly conserved peptides across many samples (Common internal Retention Time standards (CiRT) [176]).

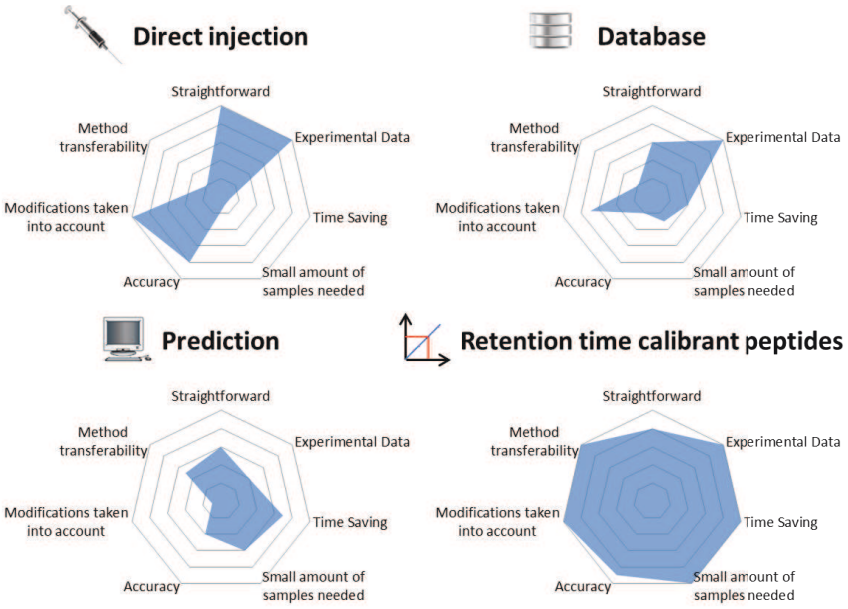



**Figure IV-5: Evaluation of retention time determination methods.**
These four radar charts show the results of the evaluation of different methods for RT determination according to seven criteria. Each criterion was empirically graded from 1 to 10.

### A.5.2 Collision energy

The optimal collision energy (CE) for a given peptide can be estimated using a linear equation between the collision energy and the mass over charge ratio of the precursor ion [177, 178]. This equation is dependent of the instrument used and the charge state. Using this approach provides a good estimation of the optimal CE. However to fully achieve the best possible sensitivity it is necessary to optimize the CE of each targeted peptide. The use of these linear equations provides a good starting point for collision energy optimization.

To optimize a peptide's CE in a fast and reliable way, Sherwood et al. developed a method consisting in making a subtle alteration of the precursor and product m/z targets of a given transition. This enables the monitoring of the same transition within a single SRM run but having a different CE value [178]. This approach has also the advantage to be very reproducible as all the measurements for a given peptide are performed within a single run, thus reducing the variance due to the LC-MS system. Additionally this approach is fully implemented into Skyline [177].

Figure IV-6.A shows the calibration curve of the CE values against the precursor m/z values for doubly charge peptide for the Thermo Scientific TSQ Vantage instrument. For a peptide of 960 m/z the red square shows the estimated optimal CE value that is equal to 32 Volts. To rapidly optimize the CE the transitions for this peptide are monitored with several CE values centered on the predicted value. In this example 9 different collision energies were followed, 4 steps on each side of the predicted value, each being separated by 2 volts, covering the range of 24 to 40 Volts. The optimization can be done by precursor m/z, i.e. using the same CE value for all the transitions and finding which value gives the best sensitivity for the summed area for the targeted precursor (Figure IV-6. B.). In this example, recreated chromatograms and the summed intensities of all transitions for a given CE value are shown. The triply charged peptide EGAEQIISEIQNQLQNLK gains more than 135% of signal intensity. The optimal CE value for this peptide is 6V lower than the predicted value. The optimization can also be done by each transition, i.e. looking for the best CE value for each transition of a given peptide. This method is preferred as it enables reaching the maximum possible

sensitivity. In Figure IV-6. C, the CE optimization results are shown for three transitions for the doubly charged peptide GLTLAPADGPTTDEVTLQVSGER. The CE optimization for the y5 and y8 transition gave a very small increase in the measured intensity. The calibration curve gives a good estimate of the optimal CE value in this case. However, the intensity of the y10 transition had an increase of +60% of the intensity with an 8V-increase of the CE value.

This method has the disadvantage of changing the fragmentation pattern and can possibly affect the calculation of dot-products using a spectral library (see Part IVChapter IA.7.2 on page 81). However this effect was shown to be minimal [177].

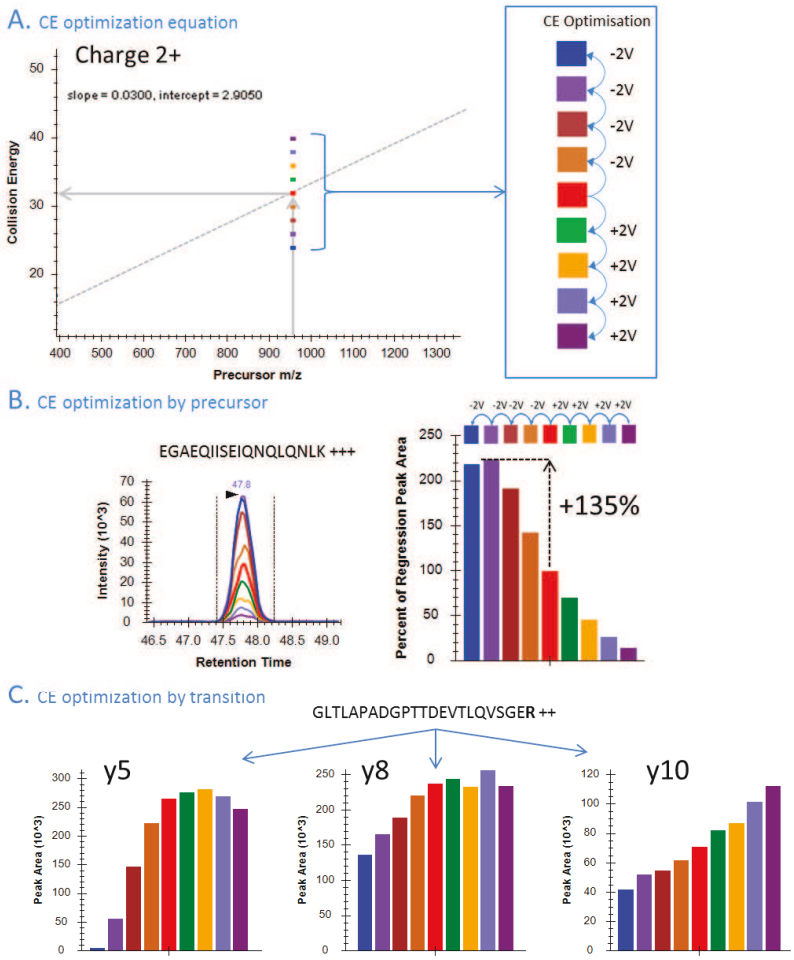



**Figure IV-6: Collision energy optimization using linear equations as a starting point.**

I used this approach to optimize the CE values of a set of 270 peptides and 867 transitions in order to determine the correct linear equation to predict CE values for our instrument. Figure IV-7. A. and B. show the results of this study. The linear equations for our instrument could be improved. The old linear equation for the doubly charge peptides was not adapted to m/z ratios higher than 1000. For the triply charge peptides, the old linear equation was too energetic and did not correlate with the data. These new equations will enable to have a better estimation of the optimal collision energy and provide a more correct starting point for the CE optimization. We believe that the collision energy optimization has to be part of the development workflow of a targeted method in order to achieve the best sensitivity.

Figure IV-7. C. shows the gain in intensity for all peptides in the data set using the same CE value for all transitions for a given peptide (lower boxplot) or using the best CE value for each transition (upper boxplot). Using the same CE value for a given precursor resulted in an increase of intensity of more than 13% for a quarter of all peptides with a maximum of 140%. Optimizing the collision energies for each transition gave even better results as seen in the upper

boxplot where a quarter of all peptides have a gain in intensity of more than 23%. One in 10 peptides had an intensity gain of more than 45% and 1 in 20 peptides had an increase of more than 72%. This significant increase in sensitivity is peptide-dependent and difficult to predict. The tools in Skyline to create and manage SRM methods for CE optimization make this step very easy and fast. Furthermore the optimized values can be stored in a database and thus be easily accessible. This step is not time-consuming and has to be done only once for each peptide and instrument setup.

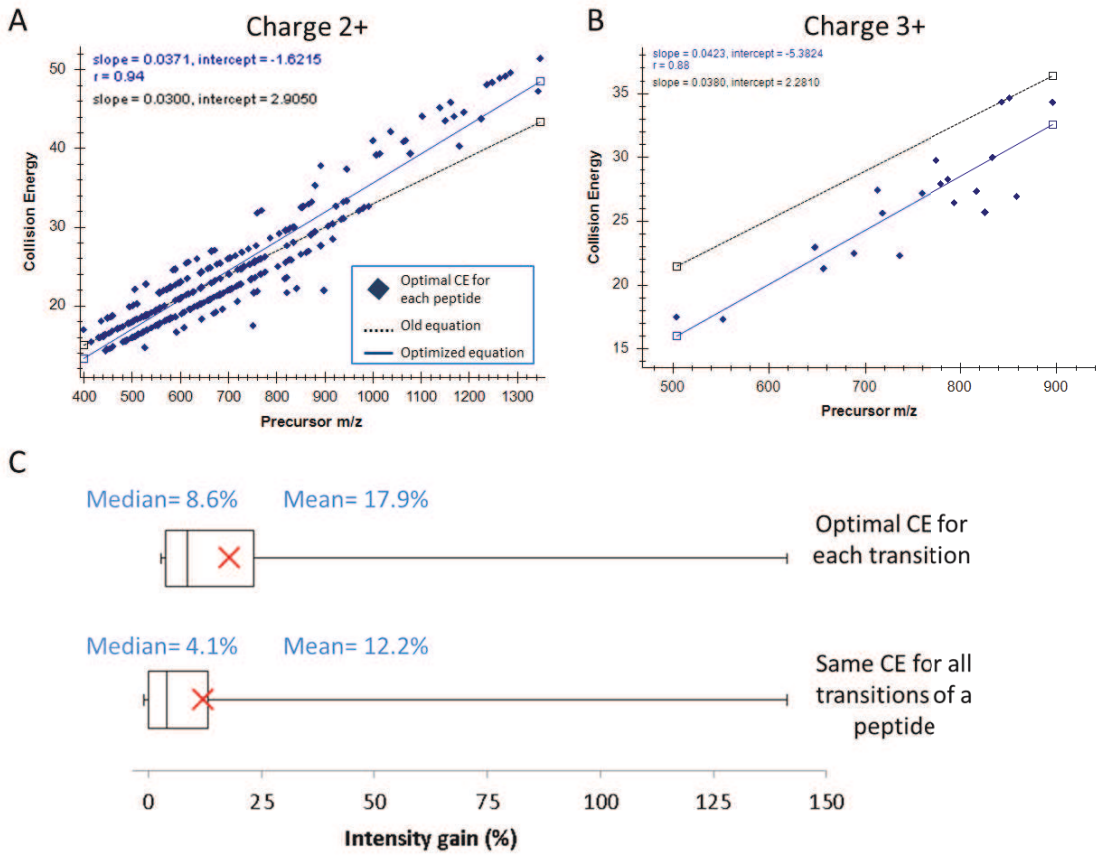



**Figure IV-7: Results of the collision energy optimization of 270 heavy-labelled standard peptides on a Thermo Scientific TSQ Vantage.**

### *A.6. Step-by-step walkthrough for the method development of targeted quantitative proteomics*

The development of a targeted method requires taking into account many different parameters. In order to obtain the best selectivity and sensitivity, multiple parameters must be optimized. The difficulty for a non-experienced researcher wanting to develop a targeted quantification method is to know where to start, how to do it and in what order to do it. In this section, I present a step-by-step walkthrough to guide proteomist in the development of targeted proteomics method development. The workflow is illustrated in Figure IV-8.

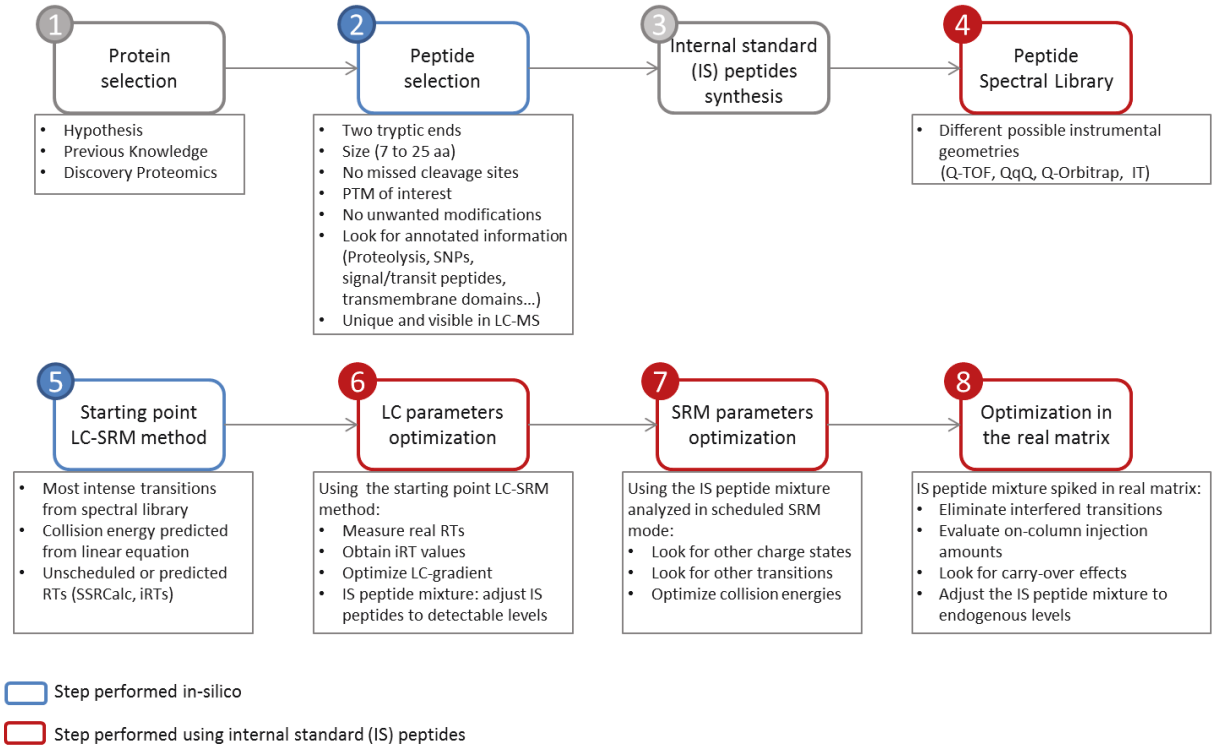



**Figure IV-8: Step-by-step LC-SRM method development workflow.**

**Step 1 and 2:** These steps have been discussed above and are done *in silico*.

**Step 3:** Once the final choice of the signature peptides is done, the heavy-labelled standard peptides can be synthetized. This step takes at least two months.

**Step 4:** Once the peptides have been synthetized. A spectral library is created and in the process it is possible to determine roughly the dilution factors necessary to create a concentration-balanced mixture. The spectral library can be done on different instrument geometries (IT, QqQ, Q-orbitrap or Q-TOF). For crude synthetic heavy-labelled peptides (Thermo Fischer PEPotec peptides) that are of low-purity and for which the concentration is not exactly known, a mixture of all targeted peptides can be done where all peptides are diluted by a factor of 1000. Then proceed to analyze the mixture. Do another mixture at a lower dilution factor solely of peptides that could not be observed in the first analysis. Do this until all peptides can be detected with good quality spectra. The peptides can now be divided into categories of dilution factors. Using this determined dilution factors, proceed to create a mixture where all peptides are present in detectable amounts. This mixture can be done on a pure solvent (5% Acetonitrile + 0.1% formic acid) or using a simple background matrix (BSA digest) to reduce peptide loss due to the coating on the vial walls.

**Step 5:** Create a method that will be the starting point for the optimization. In this method 3-5 transitions are chosen for each peptide using the spectral library. They correspond to the most intense transitions in the MS/MS spectra of each peptide. The collision energy is predicted using a linear equation (CE=a*m/z+b). The easiest way at this point is to create unscheduled methods but if the number of targeted peptides is too high the retention times can be predicted based on a RT prediction algorithm (SSRcalc), and large time windows are used.

**Step 6:** The starting point method can now be used to analyze the mixture where all peptide are present in detectable amounts spiked with retention time standard peptides (iRTs). This mixture will be used to optimize the chromatography parameters. The exact retention time can be measured and normalized RT values can be determined. These values will be saved on a database. The chromatographic conditions can now be easily optimized.

The amounts of each peptide in the mixture can now be adjusted so that all peptides reach a correct level of detectability. The intensity should not be too low (close of the limit of detection (LOD)) as the signal will not be reproducible. It also should not be too high to avoid oversaturation of the chromatographic column or suppression effects.

**Step 7:** Knowing now the exact retention time of each peptide scheduled-SRM methods can now be created to analyze the mixture of heavy-labelled peptides. Since the multiplexing capabilities are increased when using scheduled SRM, all peptides can be monitored by following a higher number of transitions to look for other possible precursor charge states that had not been seen when creating the spectral library and also look for other well-responding transitions. An example to illustrate this can be seen in Figure IV-9. In this example two triple-quadrupole instruments were used to analyze the peptide GSYTAQFEHTILLRPTCK, a Thermo Scientific TSQ Vantage and an Agilent 6490. This peptide was analyzed in the 3+ and 4+ state of charge. The chromatograms and the relative intensities of the transitions for the 3+ and 4+ charge states are shown. The relative intensities are not exactly the same in the two instruments since the collision cells are not the same. So it is important to look for transitions that can respond well in the instrument used for the quantification. Another important result is the difference in the distribution of charge states between the two instruments. On the Thermo Scientific TSQ Vantage the 3+ and 4+ charge states have almost the same intensity. However on the Agilent 6490 the 4+ charge state is more than three times more intense than the 3+ charge state. These differences can be explained by the different source configurations (geometry, voltage, temperature) of the two instruments that might favor high charge states on the Agilent 6490.

This shows that it is important to adapt the method and do a fast screening using the same instrument used for the quantification to account for specific changes in fragmentation patterns and charge distributions.

Once the best transitions have been chosen then collision energies can be optimized using scheduled SRM to reduce the number of runs necessary. This step can be done in triplicate but a duplicate is enough as the approach to optimize collision energies for a given peptide within a single run makes it very reproducible.

**Step8:** Until this step all previous optimization steps have been done using heavy-labelled standards in a pure solvent or a simple background matrix. This final step is done using the real sample matrix. The transition choice is refined to eliminate interfered transitions. This can be done by comparing dot-products calculated using the spectral library (Dotp) or calculated using the light and heavy SRM traces for a given peptide (rDotp) (see Part IVChapter IA.7.2 on page 81).

The heavy-labelled standard peptides are finally adjusted to match the levels of the endogenous peptides to create a concentration-balanced mixture.

If a sufficient amount of sample is available the optimal loading amount on column can be determined. This is a parameter that can depend on each peptide [14]. We have seen that sometimes loading a lower amount of sample can result in a better signal-to-noise ratio and thus a better sensitivity. This can also be explained by a lower response due to ion suppression when higher amounts are loaded into the column. Another possible explanation is that the total amount of protein is estimated using colorimetric assays that can be biased by interfering compounds present in the sample. We have determined that in our capillary-flow system 10 µg of total protein digest is the maximum loading capacity. But depending on the sample matrix, the real protein amount can be under- or overestimated. Thus, if possible, the optimal loading capacity of each different sample matrix must be evaluated to reach the best sensitivity.

Finally to obtain an unbiased quantification carry-over effects must be sought by analyzing a blank sample after a real sample analysis, and one must look for residual compounds from the previous injection. To illustrate this, an example is given in Figure IV-56 on page 154.

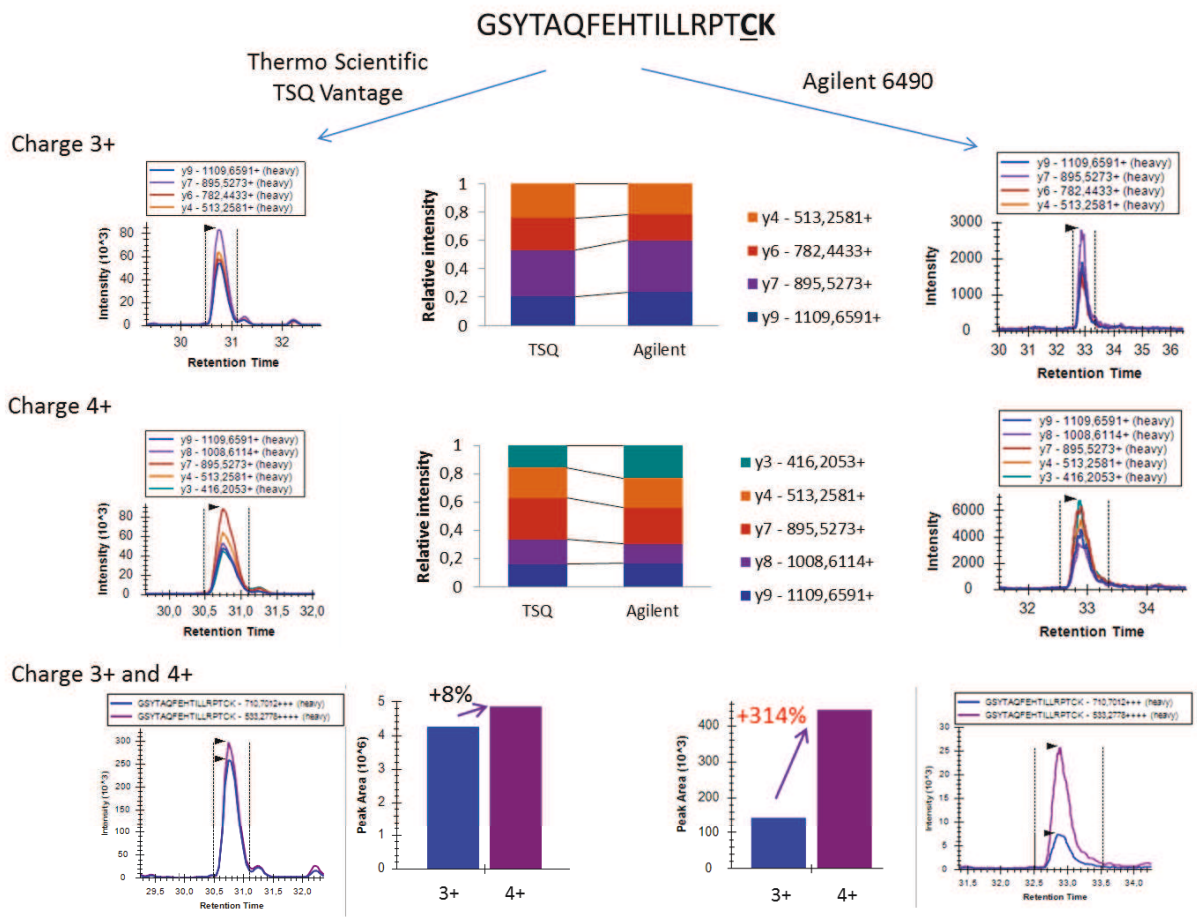



**Figure IV-9: Instrument-specific changes in fragmentation patterns and charge distribution.**
Two triple-quadrupole instruments were used to analyze the peptide GSYTAQFEHTILLRPTCK, a Thermo Scientific TSQ Vantage and an Agilent 6490. This peptide was analyzed in the 3+ and 4+ state of charge. Instrument-specific changes in fragmentation patterns and charge distribution can be observed. It is thus important to look for well-responding precursor m/z and transitions for the instrument used for the quantification.

## *A.7. Data analysis*

### A.7.1 The use of Skyline for SRM and PRM data analysis

Skyline is an open-source software [15], capable of importing raw files from different vendors (Agilent, AB Sciex, Thermo Scientific, Bruker Daltonics, Waters) and acquired using different acquisition modes (DDA, SRM, PRM and DIA). It enables the user to visualize SRM/PRM data, perform peak picking and integration of transition peak areas.

The user-friendly interface facilitates the development of targeted quantitative methods. Additionally, Skyline supports the use of MS/MS spectral libraries to aid the creation of SRM/PRM assays and to verify the correct peak group identification. During my PhD this is the software that I used for all targeted quantification studies.

### A.7.2 Metrics for peak identification and validation

When analyzing complex samples a validation set of criteria must be defined to validate the signals used for the quantification. Even in targeted quantification there is the possibility of having interfered signals. Especially in SRM assays due to the fact that the acquisition is performed on low-resolution instruments. Another challenge is that in complex samples different peptides can have the same or close precursor or fragment ion masses. This means that several peak groups will be extracted for a same precursor/fragment ion mass. The challenge is thus to pick the correct peak group corresponding to the peptide of interest (Figure IV-10).

Validating a peak group for the quantification thus means two things:

- Verifying the correct identification of the peak group
- Verifying that the peak group does not contain bad-quality signals due to low intensity or interferences by other chemical compounds.

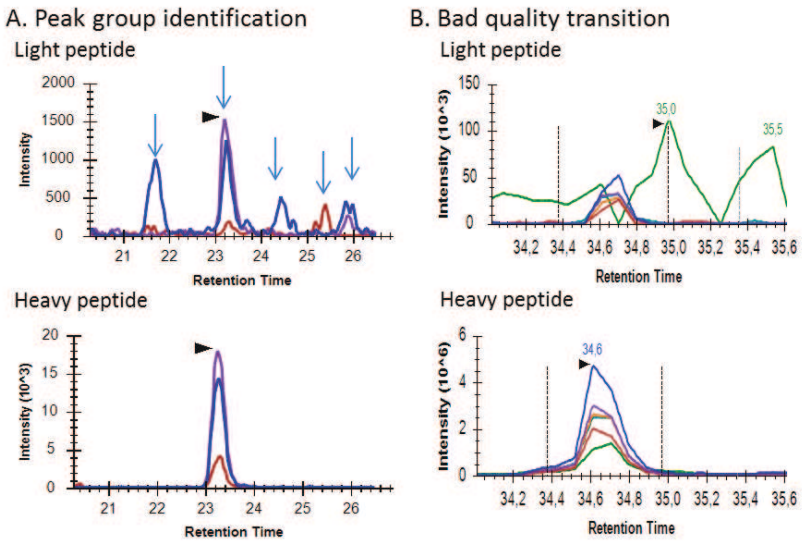



**Figure IV-10: Validating a peak group for quantification means verifying the correct peak identification and checking the quality of the signals.**
A. Several peak groups having an m/z ratio close to the one of the peptide of interest can be seen. The use of Heavy-labelled peptides enables to determine the correct peak group identification and is the method that enables the highest selectivity. B. A bad quality transition can be seen for the light peptide.

In order to validate a peak group a set of metrics are used and will be presented here.

**Co-elution and peak shape of light and heavy-labelled peptides:** First, the strategy that provides the highest selectivity is the use of heavy isotopically-labelled peptide standards. Since the light and heavy-labelled peptides have the same physico-chemical properties, they will behave in the same way in liquid-chromatography and in mass spectrometry. In Figure IV-10.A. several peak groups can be seen in the upper panel. The identification of the correct peak group is facilitated by verifying the coelution of the light and heavy peptides. Another metric to identify the correct peak group is the fact that the peak shape of the two peptides has to be the same. This also helps to find interfered transitions.

**Fragment ion relative intensities:** A Spectral library not only facilitates the choice of transitions but can also be used to verify the correct peak group identification and also identify bad quality transitions. An example can be seen in Figure IV-11. The relative fragment ion intensities are show in the histograms. The relative intensities from the MS/MS spectra of the chosen fragment ions are annotated as "library". The relative intensities of the SRM trace of the light peptide in sample S1 are the same as the ones from the library, suggesting that this is the correct peak. For the light peptide of sample S2, the relative ion intensities do not match the ones on the spectral library. Since the heavy-labelled peptide is present, we know that this is the correct peak integration but the signal is of low-intensity and it is likely interfered by signals of other compounds.

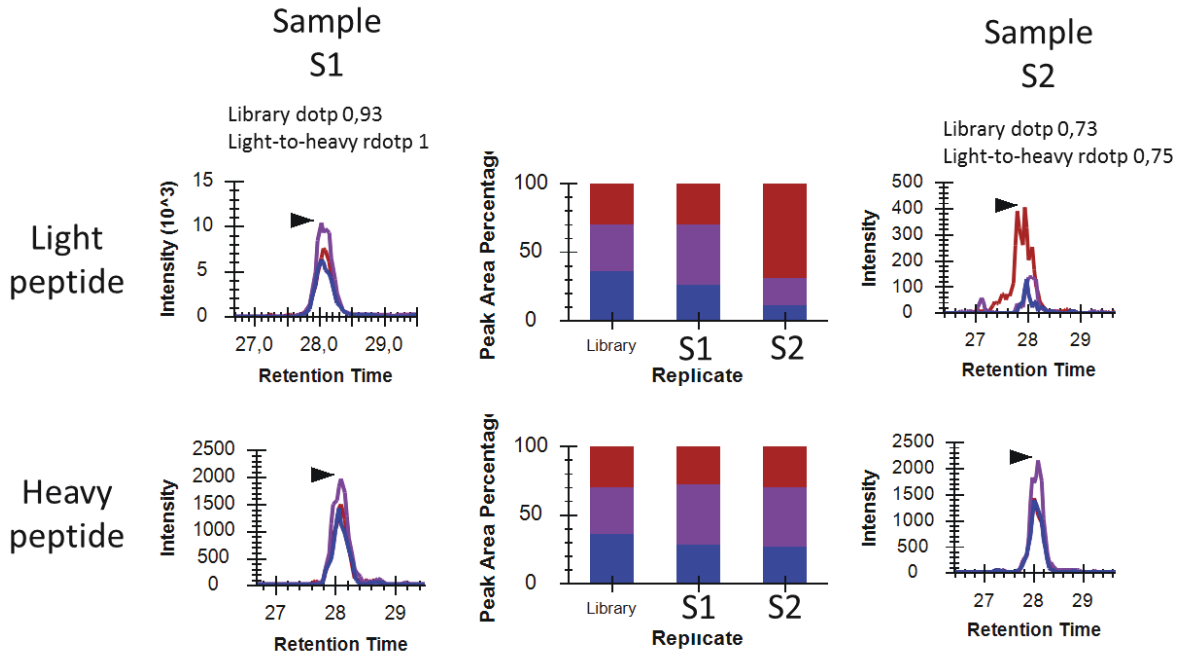



Figure IV-11: Comparing relative fragment ion intensities to assess the quality of the data.

**Library and Light-to-heavy Dot-products:** To quantify the similarity of the relative ion intensities a very useful metric is the calculation of dot-products. The calculation of dot-products was originally used to assess the similarity between MS/MS spectra [179]. The output of this calculation is a value between 0 and 1. The closer the dot-product is to 1 the more similar the two spectra are. In the case of SRM or PRM signals this tool is very useful to assess the quality of the signals used for the quantification. Skyline provides two types of dot-product calculations. The first one is the calculation of a dot-product between the full scan MS/MS spectrum fragment ion intensities and the SRM traces. In Figure IV-11 the calculated library dot-product are shown for the two samples. The dotp for sample S1 is 0.93 which shows a high similarity between the SRM trace and the MS/MS spectrum, on the contrary the sample S2 shows low similarity (dotp=0.73). If heavy-labelled peptides are present a more accurate metric is the dot-product calculation between the light and the heavy-labelled peptides' SRM traces. This has the advantage that both peptides are analyzed in exactly the same conditions and this eliminates the difference of fragmentation patterns that can exist between the SRM trace and the MS/MS spectrum if this has been obtained with another instrument.

**Automatic validation pipelines:** To validate large-scale and highly multiplexed assays, software solutions have been developed (AuDIT [180], mProphet [16], Ariadne [181]). The output information of these algorithms is a score reflecting the quality of the integrated signals based on a series of metrics (peak shape, co-elution of transitions, co-elution of light/heavy peptides, difference between measured and predicted retention times, relative fragment ion

intensities...). Some of these algorithms have implemented a false discovery rate assessment. As for shotgun peptide identification approaches, target/decoy strategies have been developed for targeted proteomics [16].

These automated pipelines work very well to discriminate signals of very good-quality from signals of very bad quality. It also helps to point out the problematic signals. However, several issues still remain and we have found that a visual verification step and manual peak picking and integration are still needed. Automatic tools can be used to reduce the number of peptides that the user has to manually inspect.

Additionally, to assess the FDR of the quantification with these strategies, decoy transitions have to be measured in other to determine the distribution of decoy targets. In SRM this is very constraining as this significantly reduces the multiplexing capabilities of the SRM assay.

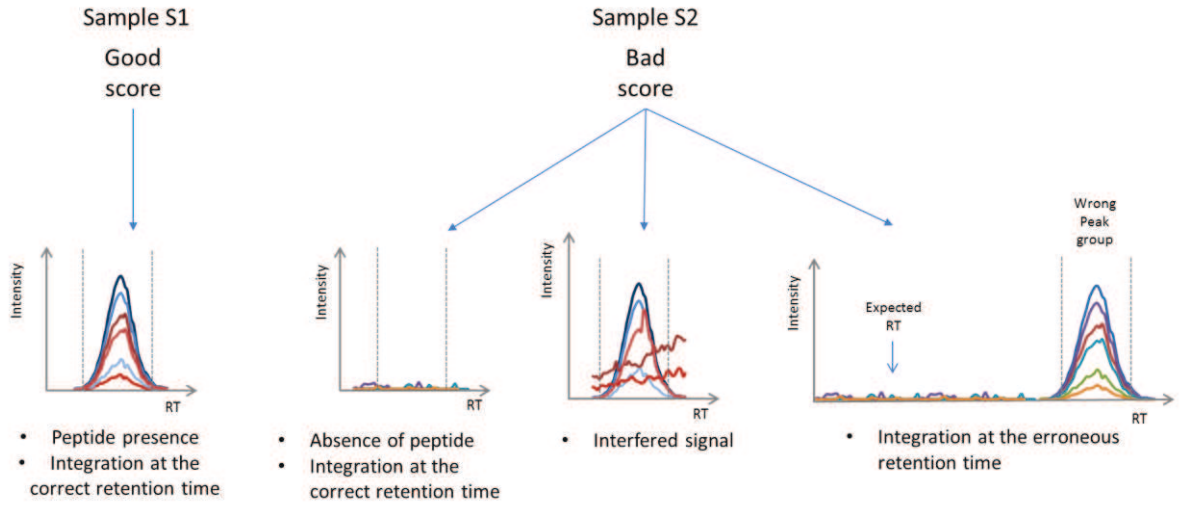



**Figure IV-12: The difficulty of assessing the meaning of bad-scoring SRM traces.**

Figure IV-12 shows the case of the comparison of two samples using an automated algorithm to score the quality of the peaks. For sample S1 the signal has a good score. The peptide was integrated in the right time window and the peptide can be correctly quantified in this sample. However in sample S2 the signal obtained a bad score. In this situation three cases are possible:

- The targeted peptide is not present in sample S2, or it is below the limit of detection. This is the ideal case in which a bad score implies the detection of the non-presence of the peptide. In this case the comparison of the two samples can be done. The overexpression of the peptide in sample S1 thus can be measured.
- The targeted peptide is present in sample S2 but the signal is of bad quality or interfered. In this case the comparison of the two samples is biased. The visual inspection of the chromatograms and manual validation would be the best option in this case. However, due to the large number of targeted peptides this is very constraining. Figure IV-13 shows the choices that the proteomist has to make in this type of situation. If an automatized workflow is chosen, in order to compare the two conditions the proteomist has to either consider the measured signal for the sample 2 as a missing value and use an imputed value to make the comparison [182]. He can also eliminate the peptide completely since the value for this peptide in a single condition cannot be used as it is interfered and including this peptide in the quantification would bias the quantification. Another option is to visual inspect the chromatograms and manually validate the peak group to obtain an accurate quantification. The automatic validation software tools can in this case guide the user to the signals that need a manual verification.

- The targeted peptide is present (or not) in sample S2 but the integration is done at the wrong retention time. We have seen that one of the major causes of errors in quantification is the erroneous peak picking, i.e. an integration of the signal at the wrong retention time. Even if software tools now have integrated the use of scoring and FDR assessment strategies for the peak picking and integration steps (Skyline[15], Spectronaut [16]), one of the most challenging steps in data analysis of chromatogram signals is the correct peak group identification. This is often the case when the targeted peak is not present (or below the LOD) and thus the peak picking algorithm tries to find the best peak in the extracted signal.

In conclusion, these automatic validation pipelines can help to accelerate the data analysis of large-scale targeted quantitative proteomic studies but the user has to be conscious of the problems that underlie behind them. We recommend to visually inspect the signals of the peptides of interest. And use these automated workflows to reduce the number of signals that the user has to inspect.

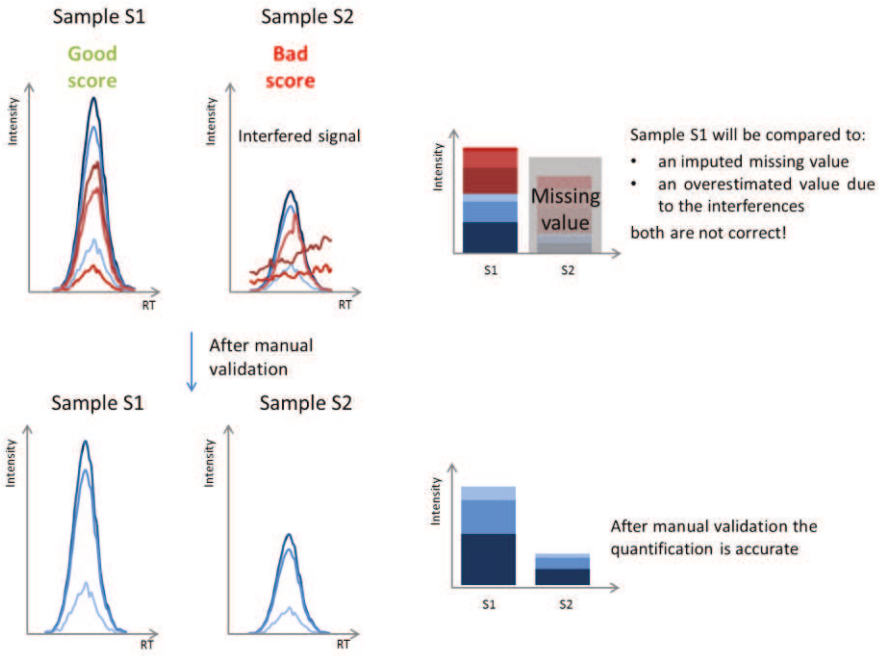



**Figure IV-13: Visual inspection and manual peak picking are still needed to obtain an accurate and reliable quantification.**

### A.7.3 Relative quantification

For all relative quantification studies I carried out, I used crude heavy-labelled standard peptides (PEPotec peptides, Thermo Scientific). The overall reproducibility of the experiment was verified by calculating light/heavy area ratios for each transition, and verifying that coefficients of variation were lower than 20% for triplicate injections. To compare different samples, the light over heavy area ratios of the sum of all the transitions was used. Evidently, all transitions with interfered signals were eliminated.

The protein relative quantification and the testing for differential protein expression were performed using the R package MSstats [183, 184].

The acceptance criteria for statistically different protein abundance changes between two conditions were set at a p-value lower than 0.05 and a fold change higher than 2. The use of these two criteria is necessary as the use of the p-value alone is not recommended. A statistical test shows if, in a pairwise comparison, the differential expression is different from zero, but it does not show if the difference observed is biologically meaningful.

### A.7.4 Absolute quantification

For all absolute quantification studies I carried out, I used high-quality, highly purified and accurately quantified internal standard peptides (AQUA peptides) [90].

For absolute quantification experiments it is important to determine the range of linearity and the limits of detection (LOD) and limit of quantification (LOQ). The LOQ is the lowest amount of an analyte in a sample that can be quantitatively determined with acceptable precision and accuracy. To define the LOQ, a signal-to-noise ratio higher than 10 is commonly used [185]. However, due to the nature of SRM and PRM, the background noise is extremely low. Accurately measuring the co-eluting noise of a SRM/PRM transition is challenging. In 2007, Keshishian *et al.* quantified low-abundant plasma proteins using the stable isotope dilution method. In this study they calculated the signal to noise ratio by dividing the peak intensity at the apex by a fixed intensity. This value was obtained based upon visual inspection of preceding and following regions around the targeted peptide's chromatographic peak [185]. Linnet and Kondratovitch proposed a procedure to determine the LOD using measurements of blank samples (processed matrix sample without analyte and without internal standards). In 2009, Keshishian *et al.* modified this equation and calculated the LOD as:

$$LOD = mean_b + \frac{t_{0.95} \times (\sigma_b + \sigma_S)}{\sqrt{n}}$$

Where $mean_b$ is the limit of blank; $\sigma_b$ and $\sigma_S$ are the standard deviations of the blank samples and the lowest level spike-in sample; n is the number of replicates [186]. Then, the LOQ was defined as three times the LOD.

However, it is very difficult to assess the LOQ this way as there is no practical way to obtain blank samples. Hence a more adapted way to assess the LOQ in proteomic studies is to define the LOQ as the lowest analyte concentration that can be measured with <20% CV [171].

The definition used for the linearity range and the LOQ for the experiments described in this thesis are the following:

- Calibration points in standard curves must show an average CV precision below 20% among triplicate injections.
- The coefficient of determination $R^2$ should be higher than 0,99 between the area under the peaks (sum of all transitions) and the injected amount on column (Figure IV-14.A).
- The coefficient of determination $R^2$ should be higher than 0,99 between the back-calculated and the real injected amount on column (Figure IV-14.B).
- Calibration points must show a 80–120% accuracy range by back-calculating expected injected amounts using regression equations after logarithmic transformation. (Figure IV-14.C).
- The limit of quantitation (LOQ) is the lowest point satisfying all the criteria reported above.

All signals were visually evaluated and validated to ensure high-quality results. Only the points satisfying all these criteria were used to calculate the linear regression equation and coefficient of determination.

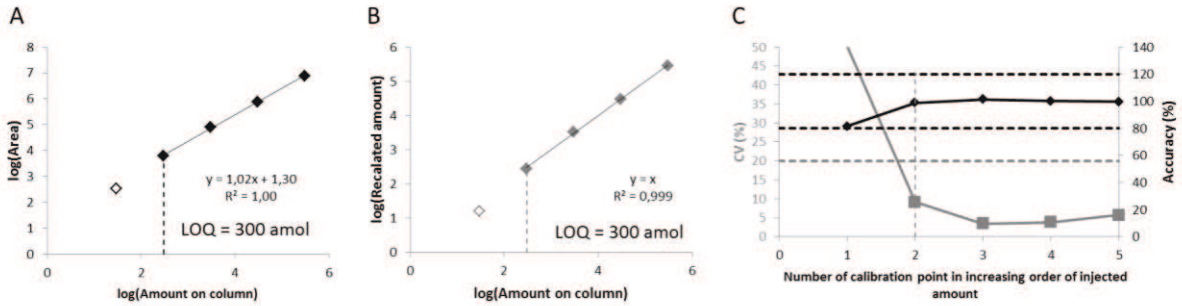



**Figure IV-14: Determination of the limit of quantification.**

A. Logarithmic area under curve against the logarithm of the amount of peptide injected on column. The full boxes are validated calibration points (CV<20%; accuracy between 80 and 120%, $R^2$>0.98), the limit of quantitation (LOQ) is shown by the dashed line and empty boxes are calibration points below the LOQ. B. Logarithm of the back-calculated peptide amount against the real injected amount on column. C. CV (%) and the accuracy (%) against the number of calibration points arranged in increasing amount of injected peptide.

# B. Evaluation of the compatibility of SDS-PAGE with targeted quantification by LC-SRM

## *B.1. SDS-PAGE separation prior to quantification*

### B.1.1 Context of the project

As stated before, one of the major challenges in quantitative proteomics is to overcome the extremely large dynamic range of protein abundance in biological fluids. In order to quantify proteins in complex samples in a reproducible and accurate manner, sample preparation strategies must be developed. One possibility to overcome the problem of the protein abundance dynamic-range is to fractionate the samples. This will decomplexify the samples, increase the proteome coverage and make the detection of low-abundant proteins possible. Even if any type of fractionation is not ideal for further quantification, it is sometimes the only way to be able to reach the dynamics and sensitivity required to quantify specific targets of interest.

In this section I will present the results of an evaluation of the compatibility of SDS-PAGE separation with targeted quantification by LC-SRM. Indeed, numerous sources of variability can compromise the quantification of several samples, namely the electrophoretic migration, the gel cutting and the in-gel migration [187].

### B.1.2 Experimental design

The analytical workflow of this experiment is illustrated in Figure IV-15. A whole yeast digest was used as a background matrix to mimic a complex biological sample. Then, the Universal Protein Standard (UPS1, Sigma) consisting of 48 human proteins was spiked-in at three different known amounts. Then, the samples were loaded on a monodimensional SDS-PAGE system. The equivalent of 100µg of yeast lysate was loaded with UPS1 spiked-in at 250 fmol, 500 fmol and 1 pmol. After migration, the gels were washed with water and fixed using 3% phosphoric acid in 50:50 methanol:water (v:v). Gels were stained by a colloidal coomassie blue method (G250, Fluka, Buchs, Switzerland). To avoid introducing errors due to the gel cutting step, gel grids commonly used in gel-based workflows were not used in this experiment. Instead, the gel was cut in two steps. First the gels were cut horizontally using a ruler and a bistoury in order to minimize the variations of this step. Then the gel was cut vertically to excise the gel bands (Figure IV-15). Bands were divided into three pieces and washed to get rid of the coomassie blue dye (25mM ammonium bicarbonate followed by acetonitrile, 3 times) using the MassPrep Station (Waters, Milford, MA, USA), proteins were in-gel reduced with Dithiothreitol (10 mM DTT in a 25mM ammonium bicarbonate solution, 30 min) and alkylated with Iodoacetamide (55mM IAA in a 25mM ammonium bicarbonate solution, 20 min). Finally, proteins were in-gel digested overnight at 37°C using modified porcine trypsin (Promega, Madison, WI). Resulting tryptic peptides were extracted using 60% ACN in 0.1% formic acid for 1h at room temperature. The volume was reduced in a vacuum centrifuge and resuspended using 0.1% formic acid in water before capillaryLC-SRM analysis.

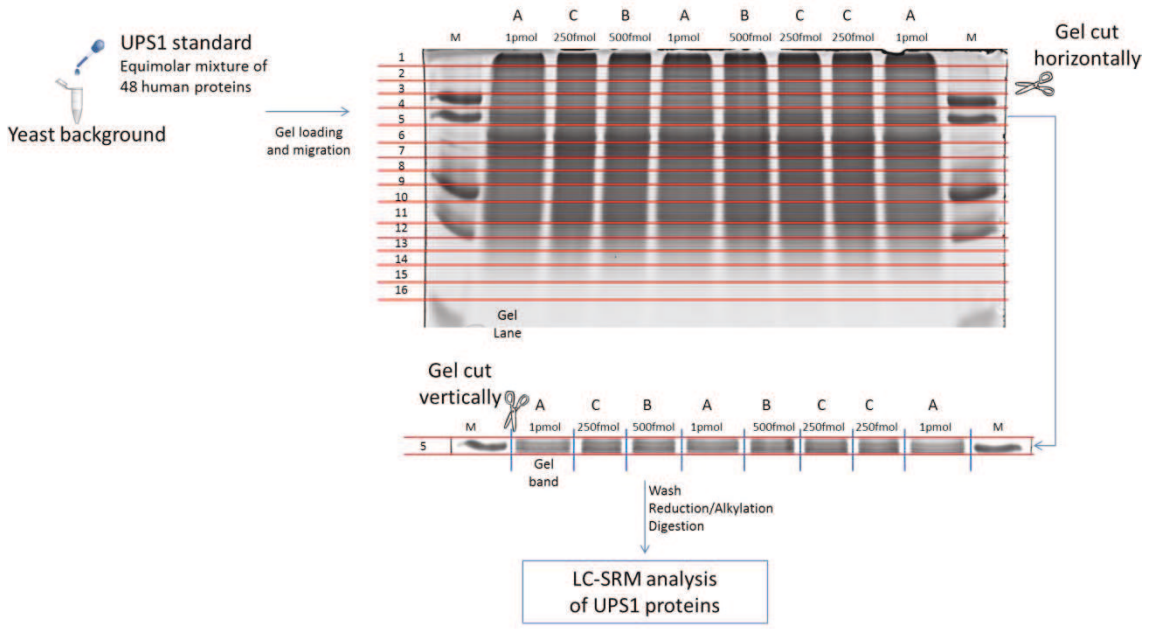



**Figure IV-15: Overview of the analytical workflow for the evaluation of the compatibility of SDS-PAGE separation with targeted quantification by LC-SRM.**

### B.1.3 Protein-specific migration profiles

Figure IV-16 shows the migration profiles of four peptides signatures of four different proteins P16083. First, one important result is the different types of migration profiles observed for each protein. These migration profiles are dependent on the protein. For example proteins P99999 and P16083 have a Gaussian profile that spans more than 8 gel bands, whereas proteins P63165 and P08263 migrated in only a few number of gel bands. This difference in migration profiles can be a problem for the quantification that will be discussed later on. I will now first list the advantages of this approach.

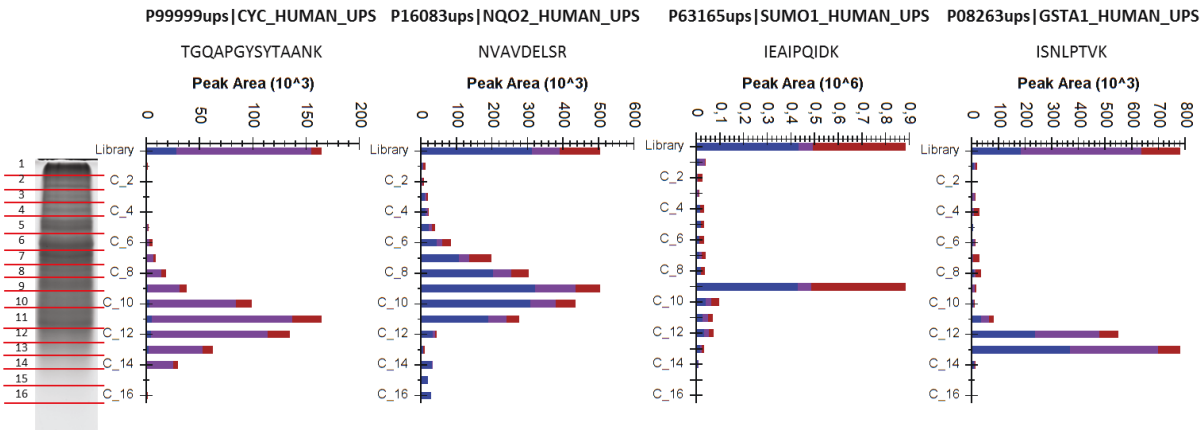



**Figure IV-16: Protein-specific migration profiles in monodimensional SDS-PAGE.**

### B.1.4 Advantages of the SDS-PAGE separation technique prior to LC-SRM analyses

Figure IV-17.A. shows the migration profile of protein P16083 monitored by three peptides covering different parts of the protein sequence. These three peptides have the same migration profiles confirming that the protein is only present in a single form. As for chromatographic peak shapes, the migration profile can provide an added level to the selectivity of the protein quantification. Quantifying a protein only using the peptides that have the same migration profile can increase the accuracy and selectivity for a given proteoform, knowing that there is no other isoform that can bias the quantification. The corollary of this statement is also true. Coupling the SDS-PAGE separation to targeted quantification by LC-SRM can enable the quantification of isoforms. In Figure IV-17.B two forms of protein P02788 can

be discriminated as two migration profiles of a long and a short form of the protein can be observed with an apex at bands 2 and 4 respectively. This migration profile was observed in all 8 samples and with three different peptides confirming that the migration profile is due to the presence of two protein forms and cannot be attributed to an instrumental artifact. The protein P02788 is obtained from milk, HPLC purified and quantified by AAA. Two isoforms exist for this protein with one missing 44 amino acids in the N-terminal part of the protein. Unfortunately the three signature peptides chosen to monitor this protein are after the position 285 in the protein sequence. However the two migration profiles are likely to be these two isoforms. Thus, the use of 1D-SDS-PAGE coupled to mass spectrometry can inform about the presence of different isoforms and proteoforms like proteolytic events. In 2008 Dix *et al.* introduced the Protein Topography and Migration Analysis Platform (PROTOMAP) approach to analyze proteolytic events using SDS-PAGE coupled to quantification by mass spectrometry using spectral count [188].

Furthermore, this approach enables to increase the specificity of the quantification as possible biomolecular interferences are separated from the proteins of interest. An example can be seen in Figure IV-17.C. in which the peptide VLDALQAIK was monitored using three transitions. The protein was identified in the gel band number 10 and the peptide was eluted at minute 24. Another molecule having a precursor and a fragment m/z ratio close to the y4 transition of peptide VLDALQAIK and eluting at exactly the same retention time was found in the gel band number 7. Fortunately, the added separation dimension of the SDS-PAGE enabled to completely discriminate these two molecules. Using an unfractionated protocol, the y4 transition would be interfered and could not be used for the quantification.

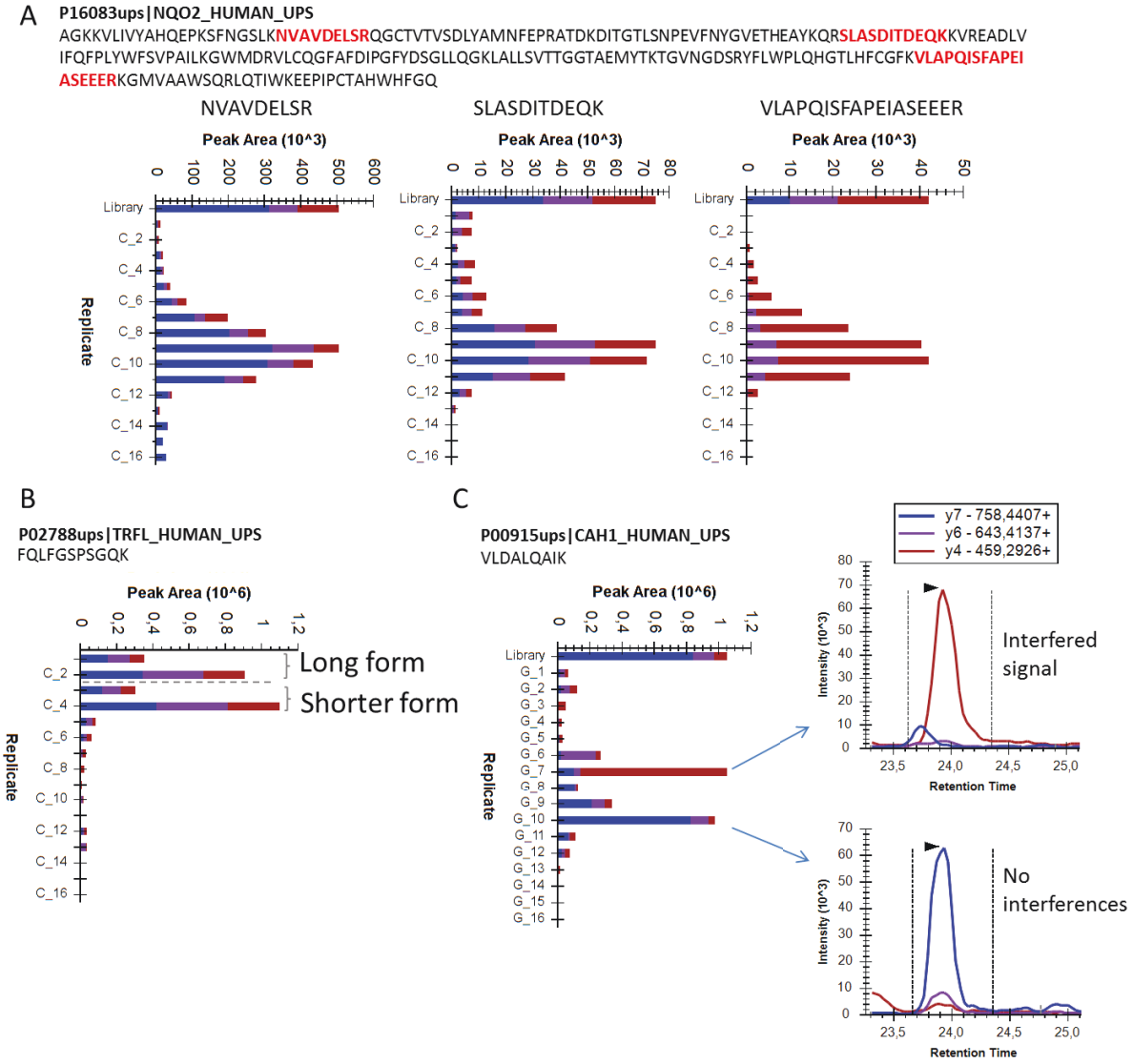



**Figure IV-17: Advantages of SDS-PAGE-LC-SRM approach.**
The added separation dimension of the 1D-SDS-PAGE approach increases the specificity of the quantification (A), enables the quantification of different proteoforms of a given protein (proteolysis products, isoforms) (B) and reduces the sample complexity and lowers the presence of interferences (C).

### B.1.5 Drawbacks of the approach

Even if this approach promises several advantages it also has several drawbacks. First, this approach requires longer instrument time when compared to unfractionated protocols. In this study 16 gel bands were analyzed for each sample with a 1-hour gradient. This very long analysis time constraints the overall number of samples that can be analyzed, especially if the samples are analyzed in replicates. Additionally, studies with long analysis time are more prone to quantification biases. Indeed, the instrumental performances can decrease over time (or even fail to perform) and produce confounding errors.

Additionally, as seen in Figure IV-16 the migration of several proteins can span across multiple gel bands. This means that the overall sensitivity for these proteins is lowered as the proteins are diluted in multiple gel bands. This is due to the low resolving power of the SDS-PAGE system.

Moreover the quantification when using this approach can be biased as each gel band contains a different background matrix and this can change the ionization efficiency due to ion suppression effects. In order to limit these effects, internal standards can be used. In this study however internal standards were not used and the quantification was done with a label-free approach. Another way to reduce the impact of the changing instrumental performance over

time is to analyze the equivalent gel bands of all samples close to each other (Horizontal analysis) and avoid analyzing the entire lane for each sample before analyzing another sample (Vertical analysis) as this will increase the time between the analyses of two equivalent gels bands.

### B.1.6 Data analysis of fractionated samples by SDS-PAGE

All these factors have to be taken into account in the strategy used for the data analysis. We wanted to evaluate which approach enabled the correct quantification of our targeted proteins. First, we wanted to evaluate the reproducibility when using only the most intense transition or three transitions per peptide. We also wanted to compare the reproducibility of the quantification when using only the most intense gel band or when summing the intensities of three gel bands centered on the most intense band (the most intense band and the adjacent gel bands on either side). Two factors can influence the quantification when summing several gel bands: the ion suppression effects due to different background matrices in each gel band and the migration of proteins across multiple gel bands.

The Skyline open-source software package was used to visualize the SRM data. A manual inspection of the peak picking and the integration of transition peak areas were done. In the study of fractionated samples, the manual inspection of all signals is very important as the proteins are only present in a few fractions. It is important to confidently determine in which fractions the proteins are present. And in fractions in which the protein is not present, Skyline erroneously picks and integrates another peak group. This is a problem as it can result in errors of peak group identification. It also compromises the use of automated workflows. We have seen that a manual inspection is required. This step however is time consuming. For example, in this experiment 126 peptides were monitored with three transitions in 8 samples and each sample consisted of 16 fractions resulting in more than 48300 SRM traces to be inspected manually.

### B.1.7 Results of the evaluation

The results of the evaluation are presented in Figure IV-18. We have seen that the migration profile was reproducible in different gel lanes. An example can be seen in Figure IV-18.A in which protein P02787 is monitored in three gel lanes. The migration profile is the same and the most intense band is the 4th gel band for all lanes. Overall, 65 out of 126 peptides were quantified for 34 of the 48 proteins. We also evaluated the quantification reproducibility when using 1 or 3 transitions per peptide and when using only the most intense gel band or when summing the intensities of three gel bands centered on the most intense band. The coefficient of variation was calculated for each quantified peptide. The frequency and the cumulative frequency are shown in Figure IV-18 B. and C. where we can clearly see that the quantification is not significantly affected when quantifying with 1 (solid lines) or 3 transitions (dashed lines) per peptide. However, the quantification is undoubtedly more precise when summing the three adjacent gel bands (purple lines) when compared to using only the most intense gel band (green lines).

Since this experiment did not use internal standard peptides a global normalization procedure was carried out. This consisted in aligning for all replicates of the same sample, the median value of all the intensities measured. This will correct for different loading amounts on the gel and overall MS performance but will not have an impact on further variations such as ion suppression effects. The normalization did improve the precision of the quantification. Overall the coefficient of variations showed a mean and median value of 17% when not normalized and a mean value of 13% and a median of 10% when the global normalization procedure was applied (Figure IV-18.D and E).

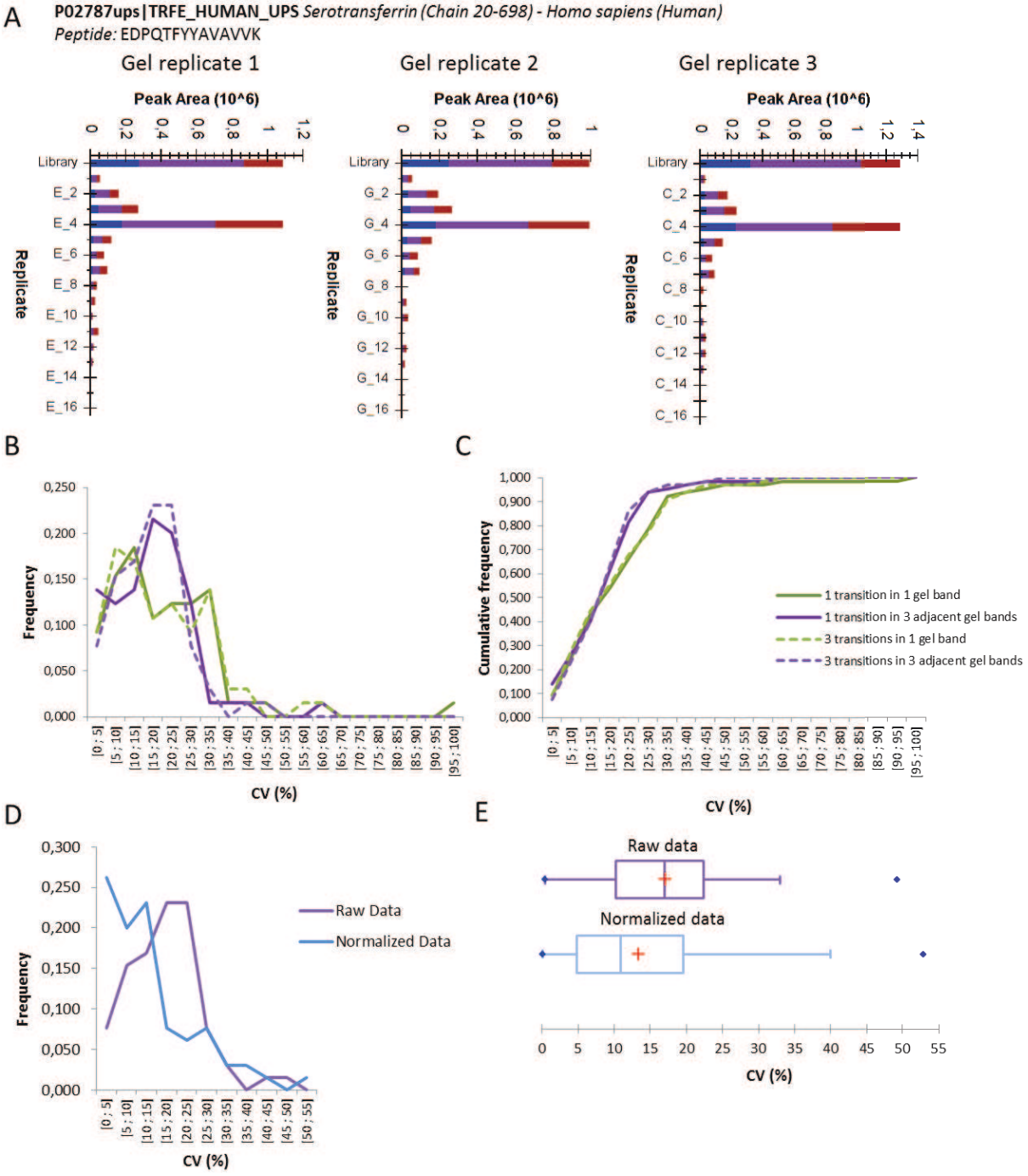



**Figure IV-18: Results of the evaluation of the SDS-PAGE-LC-SRM workflow.**
We evaluated the reproducibility of the quantification of the technique. An example of the migration profile reproducibility of the same protein in three different gel lanes is shown (A). The quantification was evaluated when using 1 (solid lines) or 3 transitions (dashed lines) per peptide and when using only the three adjacent gel bands (purple lines) when compared to using only the most intense gel band (green lines). The distribution (B) and the cumulative frequency (C) of coefficient of variations are shown. The need for a normalization step was also evaluated. The distribution (D) and the boxplots (E) of the coefficients of variations of raw and the normalized data are shown.

### B.1.8 Evaluation of a high-resolution SDS-PAGE system

The low resolution and the diffusion of protein peaks in the gel can be explained by the Joule heating caused by the electric current passing through the gel. Joule heating not only increases the fluid temperature, but also produces temperature gradients that can broaden the protein peaks and the widths of the migration lanes. In order to improve the coupling of SDS-PAGE with LC-SRM a high performance electrophoresis system was evaluated. The HPE flattop tower (The Gel Company, San Francisco, CA, USA) system uses very thin gels with 420 µm of thickness covalently polymerized to a thin film support. The system uses a horizontal system with a cooling plate constantly cooled by

water delivered by a pump. This enables an efficient heat dissipation to obtain rapid and straight electrophoretic migration. As a result of the controlled temperature the resolution of the SDS-PAGE system should be improved.

However the very small thickness of the gel implies two problems: low loading volume capacity and low protein amount loading capacity. The loading volume capacity is maximum 15µl. We tried to use the system by loading 100µg per sample. However the gels burned due to high tension possibly due to the presence of salts.

When loading lower amounts (<50µg), the migration of the samples could be done. Nevertheless, since the gel is covalently bonded to the plastic support is was not possible to separate them without losing some sample. All in all, this system did not meet our needs since the low amount of protein loading was not compatible with the amounts needed for quantification by capillaryLC-SRM. The very low gel thickness and the fact that it was covalently bond to the support made this item very impractical, very time-consuming and prone to sample losses.

#### B.1.9 Conclusion

In conclusion, the SDS-PAGE separation coupled to a targeted quantification by LC-SRM can be an alternative when a limited number of proteins have to be quantified in a small number of samples. This strategy can be used to efficiently reduce the sample complexity and analyze low-abundant proteins. However, it requires a significant amount of instrument time and a dedicated data analysis workflow.

If only a few proteins are to be quantified, then this workflow can be used to purify them and analyze only a few bands, or even pool several bands together in order to reduce the complexity while keeping a reduced number of samples to be analyzed.

### *B.2. Development of an unfractionated stacking Gel SDS-PAGE protein purification protocol*

#### B.2.1 The principle of stacking gels

In order to benefit from the advantages of SDS to solubilize and extract proteins, we wanted to evaluate an unfractionated SDS-PAGE protocol to obtain a more precise, accurate and simple quantification workflow. To do this we wanted to take advantage of the fact that the stacking gel focusses all proteins in sharp bands before they enter the resolving gel.

The principle behind the staking gel is illustrated in Figure IV-19. The electrophoresis buffer contains glycine which is a zwitterion that at low pH is protonated and thus uncharged. At pH 6,8 of the stacking gel the glycine migrates very slowly (trailing ion). The stacking gel contains chloride ions (leading ions) which have a very high mobility. This creates a narrow zone of very low conductance (very high electrical resistance). The protein molecules are trapped in a sharp band between the leading ions and the trailing ions. The negatively charged proteins move forward due to the very high field strength but they can never go faster than the chloride ions. If somehow they would outrun the leading ions they would be in a region of very low field strength, due to the high conductance (low resistance) and would immediately slow down. As a result, all the proteins move in the stacking gel in a sharp band. When the proteins reach the resolving gel, the glycine becomes deprotonated and negatively charged as the pH is now of 9. Its mobility increases and the mobility of the proteins decrease due to the sieving properties of the gel (smaller pore size of the resolving gel). The proteins are no longer in a narrow zone of very high electric field. They are now in a uniform electric field where they are separated based on their size.

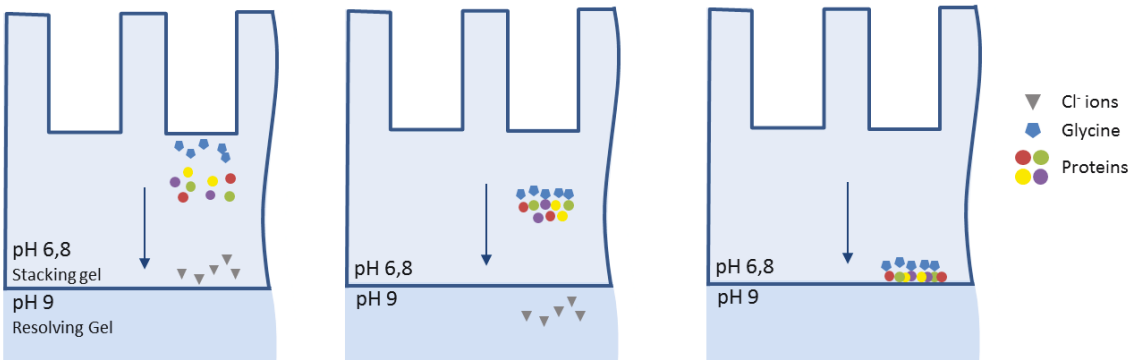



**Figure IV-19: Principle of a SDS-PAGE stacking gel.**

### B.2.2 The optimization of key parameters to enhance the quantification performances of the stacking gel protocol

We developed and optimized a protocol to use stacking gel as an unfractionated sample-preparation method compatible with further quantification by mass spectrometry. Figure IV-20 shows some practical results obtained during the development of the optimized protocol. The upper right part of the figure shows how the gel should be polymerized. The percentage of a stacking gel is normally 5% of acrylamide. It should not be polymerized by its own as it is very fragile. A resolving gel should be used to hold stacking gels in place and to easily manipulate the gels without breaking them.

We also found out that the width of the migration lane depends on the loading volume on the gel (Figure IV-20.B). To increase the reproducibility of the migration when comparing several samples, all samples must be loaded with the same volume. The samples should not be loaded on neighboring loading wells as the migration front could be very close and difficult to cut or possibly mixed between two neighboring samples. To avoid this, a loading well should be left empty between each sample. But it should contain the same volume of sample buffer with the exact same composition as the one used for the samples. If not, the migration front will not be straight. Additionally it is best to avoid the loading wells on the borders of the gel. We have seen that the migration front in these positions is irregular (Figure IV-20.C). Finally, the acrylamide percentage of the stacking gel was also studied. A 5% staking gel is commonly used in proteomics workflows. We evaluated this acrylamide percentage and found out that large proteins (>250 kDa) might not migrate at the same speed than smaller proteins (Figure IV-20.D), and will thus not be correctly stacked in the migration front. Using a 4% acrylamide stacking gel, this phenomenon was reduced. However, lower acrylamide percentage could not be used as the gel lost its consistency and could not be easily handled. An example of the stacking gel outline is shown in Figure IV-20.E. Using a 10 loading wells gel, only four samples can be loaded per gel.

Empirically, we found that the best results in terms of sensitivity and reproducibility were found when more than 50µg were loaded into the stacking gels. Below this amount we believe that peptides are most likely partially lost due to adsorption on the vial walls in subsequent preparation steps.

Since all proteins are stacked in a single sharp band, the coloration of the gel should not exceed 15 minutes. The high concentration of proteins in a tight band is rapidly colored and if longer times are used then the coomassie blue cannot be removed and can affect chromatographic MS conditions.

The optimized protocol is described in the Experimental Part on page 216.

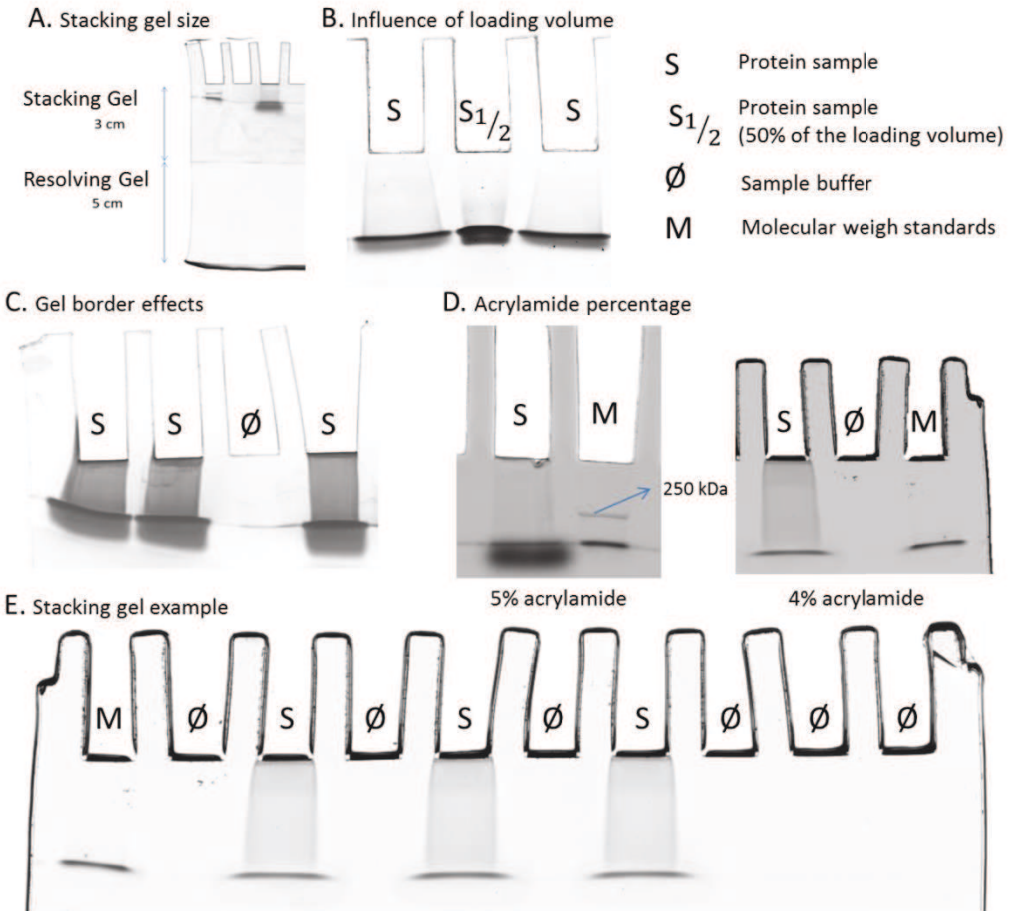



**Figure IV-20: Development of the SDS-PAGE stacking gel purification protocol.**
Several key parameters were optimized to enhance the quantitative performances of the sample preparation protocol these were the stacking gel size (A), the influence of loading volumes (B), the effects of loading samples in the border loading well (C), The acrylamide percentage (D). An example of the optimized stacking gel is also shown (E).

### B.2.3 Evaluation of the reproducibility: experimental design

We tested three ways to cut the gels. The Figure IV-21 shows in red the template to cut the gels. The first one consisted in migrating the samples 2 cm into the stacking gel and then cutting only the migration front, where the proteins should be stacked in a sharp band (Protocol 1). The second consisted in migrating the samples 1 cm into the stacking gel and cutting all the gel above and including the migration front (Protocol 2). And the third one consisted in stopping the migration after the samples entered the resolving gel (Protocol 3).

To evaluate this parameter, a whole lysate of human cells (HepaRG cell line) was used. 50µg of total lysate was loaded on stacking gels in triplicates on a single gel (intra-gel replicates) and on three different gels (inter-gel replicates). Additionally, the stacking gels were also compared to a liquid digestion protocol (Protocol 4). The nanoLC-MS/MS analyses were done using a Bruker Daltonics MaXis 4G Q-TOF instrument.

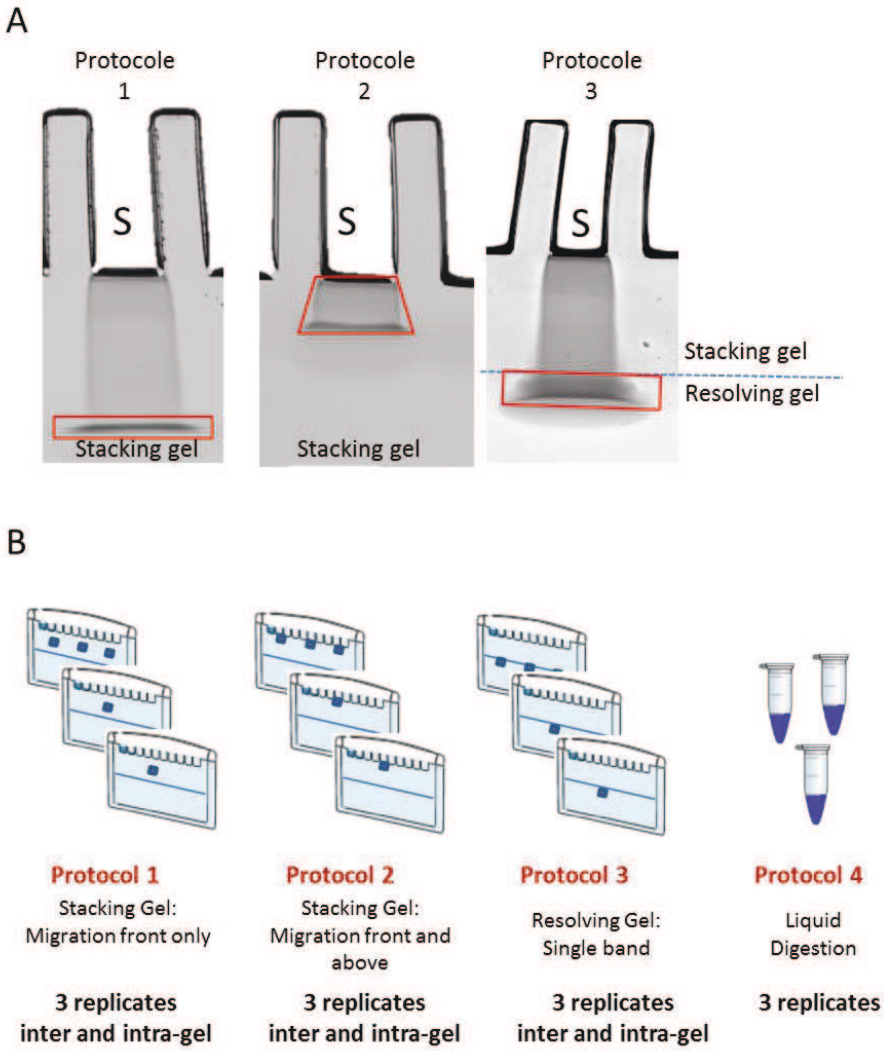



**Figure IV-21: Evaluating the cutting template of stacking gels.**
Three ways of cutting the gels were evaluated. The cutting pattern is shown by the red rectangle (A). The samples were prepared in inter- and intra-gel triplicates. And these were compared to a liquid digestion/C18-SPE protocol (B).

### B.2.4 Evaluation of the reproducibility: results

The number of protein and peptide identifications obtained with each sample preparation protocol is shown in Figure IV-22. The average number of proteins and peptides identified in each sample is shown in part A and the cumulated number of identifications is shown in part B of the figure. One interesting result is the significant difference between the average number of peptides identified in a single run, around 3400, and the total cumulated identifications, around 6000. This important difference shows the important undersampling effect of which DDA analysis still suffer on this Q-TOF platform.

Of note is the fact that gel-based sample preparation protocols showed overall higher performances than the liquid digestion protocol. This can be explained by two factors. The SDS used in the Laemmli buffer helps to extract and solubilize the proteins when compared to the urea buffer used for the liquid digestion protocol. Also, in gel digestion has been proven to be more effective than in liquid digestion [189]. And, in the case of stacking gels, the lattice is much lower than the one of resolving gels, 4-5% and 10-12% respectively. This facilitates the access of the digestion enzyme increasing thus the yield of digestion. Among the gel-based protocols, it is the protocol 1 gave the highest number of identifications.

Parts C and D of the figure show the complementarity of the protocols. This can partly be explained by different protein migration profiles that were not present or partially present in the region of the gel that was cut and analyzed. It can also be explained by the undersampling of the instrument. This is illustrated in part E of the figure, that shows the cumulative number of identifications according to the number of replicates analyzed. In this case the data from the 5 replicate analyses using the protocol 2 are shown. This clearly shows the undersampling of the instrument as each new replicate adds a significant number of new identifications.

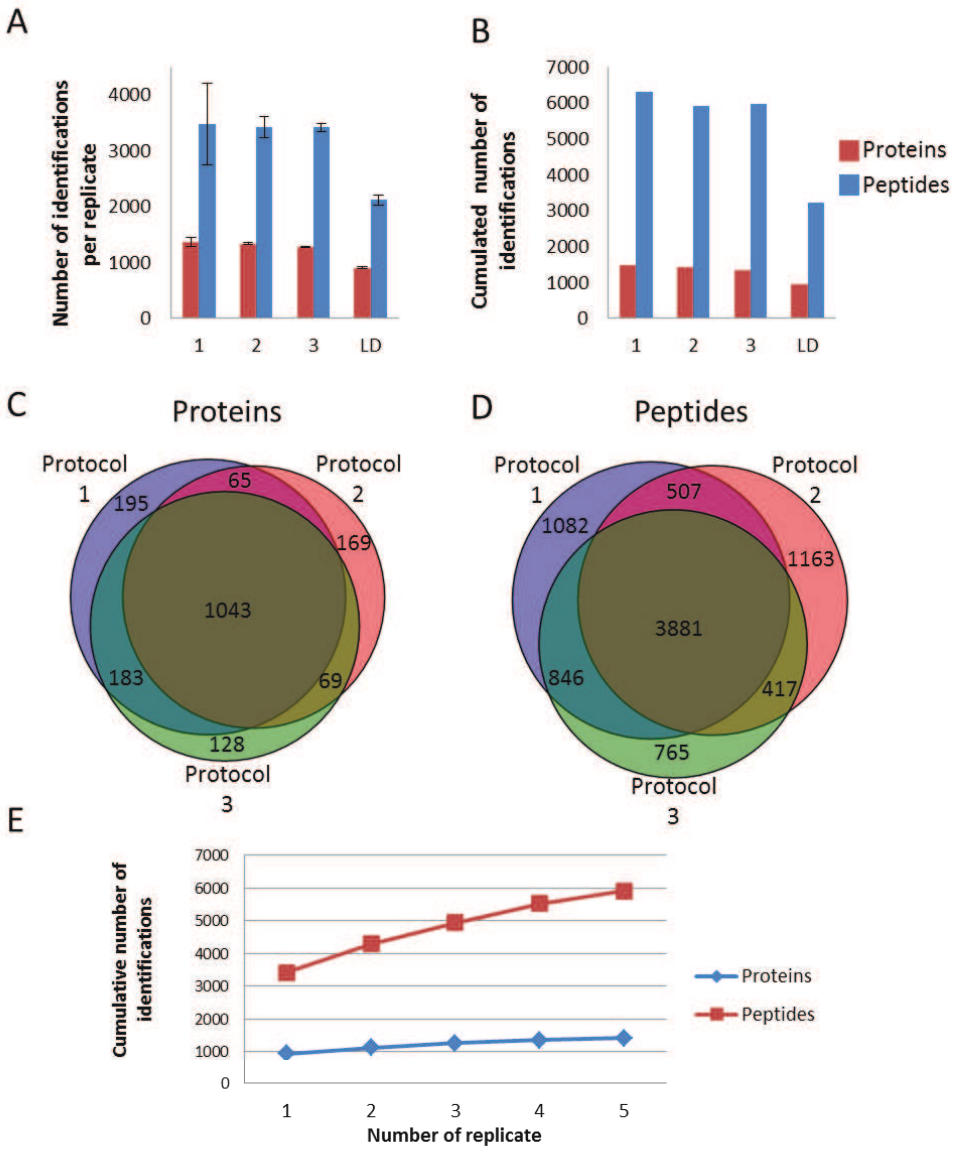



**Figure IV-22: Number of identifications of proteins and peptides according to the sample preparation protocol.**
The average number of proteins and peptides identified in each sample (A) and cumulated number of identifications of peptides and proteins (B) according to the sample preparation protocol are shown. The Venn diagrams showing the complementarity of the methods in terms of identifications of proteins (C) and peptides (D) are also shown. The cumulative number of identifications according to the number of replicates analyzed (E) illustrates the undersampling of the instrument.

We wanted next to assess the analytical performances of intra-gel and inter-gel replicates using each sample preparation protocol. To do this we used a Label-free MS1 filtering strategy.

One of the major problems in MS1-filtering is the correct peak identification and integration (see Part IVChapter ID.4 on page 109), which can increase the number of false positives and false negatives. To minimize this problem and obtain a value of the variability that originates solely from the sample preparation steps, we used the following data analysis strategy.

To minimize the errors of peak picking and peak integration due to errors in matching peptides across different runs, only the set of peptides that were identified in all four sample preparation protocols were used for the quantification. We used Skyline to extract the 3 isotopes (P, P+1 and P+2) for each peptide. We used a spectral library to provide the retention time coordinates of each peptide in each LC-MS run. The quantification method extracted information for 1407 peptides and 1623 precursor ions and monitored 4869 MS1 signals. For each peptide, the summed area under the 3 isotope peaks was used. Additionally, to minimize the variability originating from the LC-MS instrument (sample dilution, sample injection, MS performance…) a normalization step was used. This way we ensure that the remaining variability is a product only of the sample preparation protocol. To perform the normalization step we used the Normalyzer tool [190], an open-source tool in R language [184]. We used a Log2 transformation and each sample in a replicate group was normalized to the median of all the samples in the replicate group.

The results can be seen in Figure IV-23. The results show the analytical performance of inter- and intra-gel replicates. The boxplots show the distribution of log2-transformed intensities in each sample after the normalization. All three gel-based protocols have the same levels of intensities, and have overall higher intensities than the liquid digestion protocol. The Relative Log Expression (RLE) plots show the ratio between the intensity of a peptide and the median intensity of the peptide across all samples. Since the assumption that the majority of peptides are unchanged across all samples is made, the samples should be aligned around zero. Any deviation would indicate discrepancies in the data. What is interesting here is to see that the peptide intensities from the liquid digestion protocol are quite different than those obtained by the gel-based protocols. From this plot we can see that at least 25% of the peptides have intensities 2 times lower than the corresponding median intensity of the peptide across all samples. This is in agreement with the previous result. What it interesting is to see that there are also peptides that are more intense in the liquid digestion protocol than with the gel-based protocol. Further data mining must be performed to determine whether this result comes from protein or peptide physico-chemical characteristics that make a peptide respond better in one condition compared to the other. However, only 37 peptides were found to have a higher intensity when using the protocol 4 when compared to the gel-based protocols.

Moreover it is important to note that the gel-based protocols are very similar to each other. This can also be seen in the dendograms where all the gel-based protocols are grouped together. And among the gel-based protocols the two stacking gel protocols (protocol 1 and 2) are also grouped together and cannot be distinguished. Finally, the distribution of coefficients of variations for all quantified peptides is shown. All protocols show similar performances. Though, one intra-gel replicate of protocol 1 had a problem of the injected volume. This affected the coefficient of variations in the intra-gel conditions. However the plot showing the inter-gel replicates show that the protocol 1 and specially the protocol 2 have overall lower CVs.

Finally in terms of practicability the protocol 2 is the one that requires the less amount of expertise from the researcher performing the experiment.

All in all, the protocol 2 is the one providing the best analytical performances in terms of precision and proteome coverage. A parameter not discussed here is the accuracy that can be obtained by this method. Further studies can be done to evaluate this important parameter using a well-characterized sample in which known amounts of variant proteins are spiked into a complex biological sample. This strategy will be discussed in the next section.

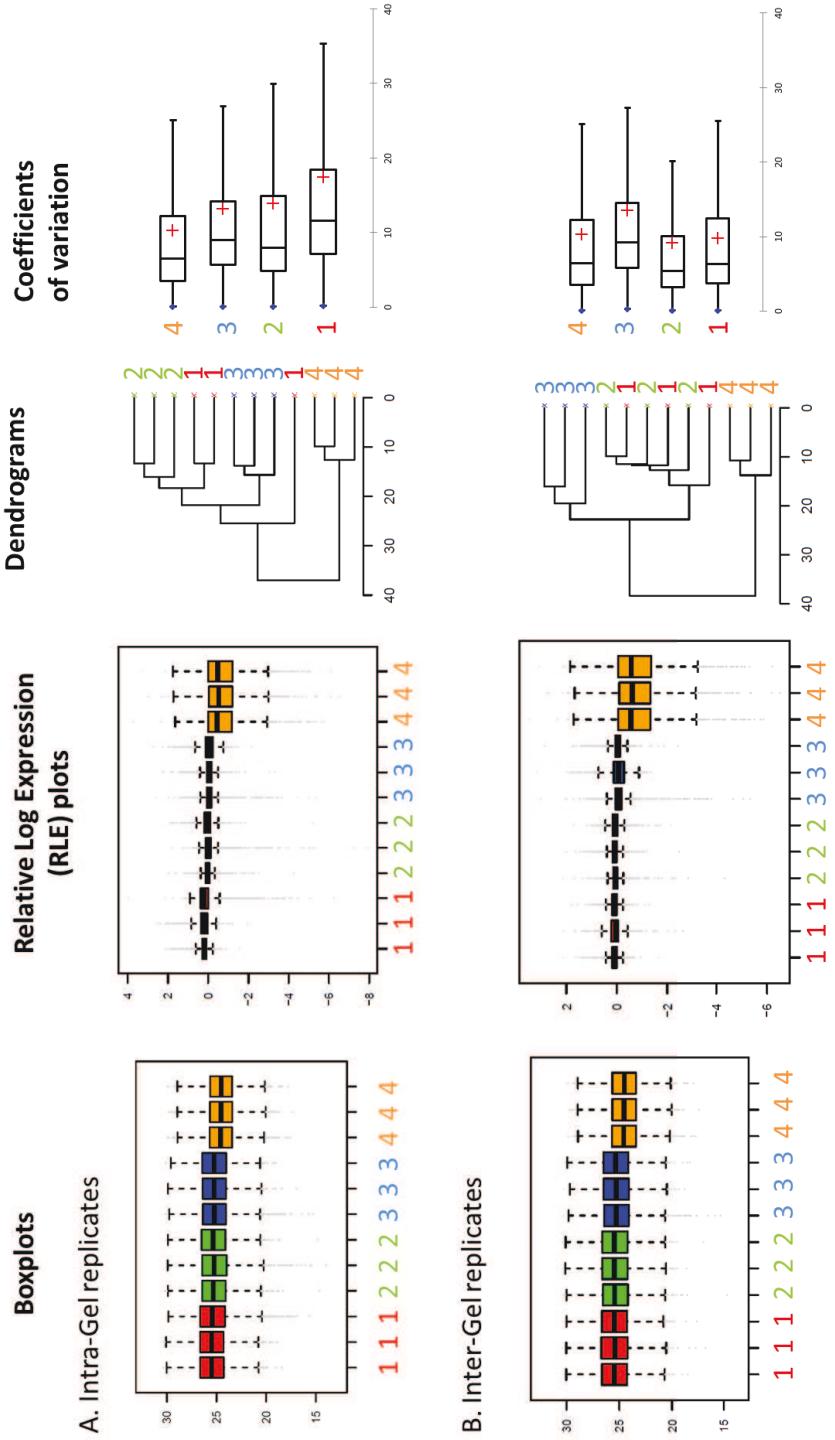


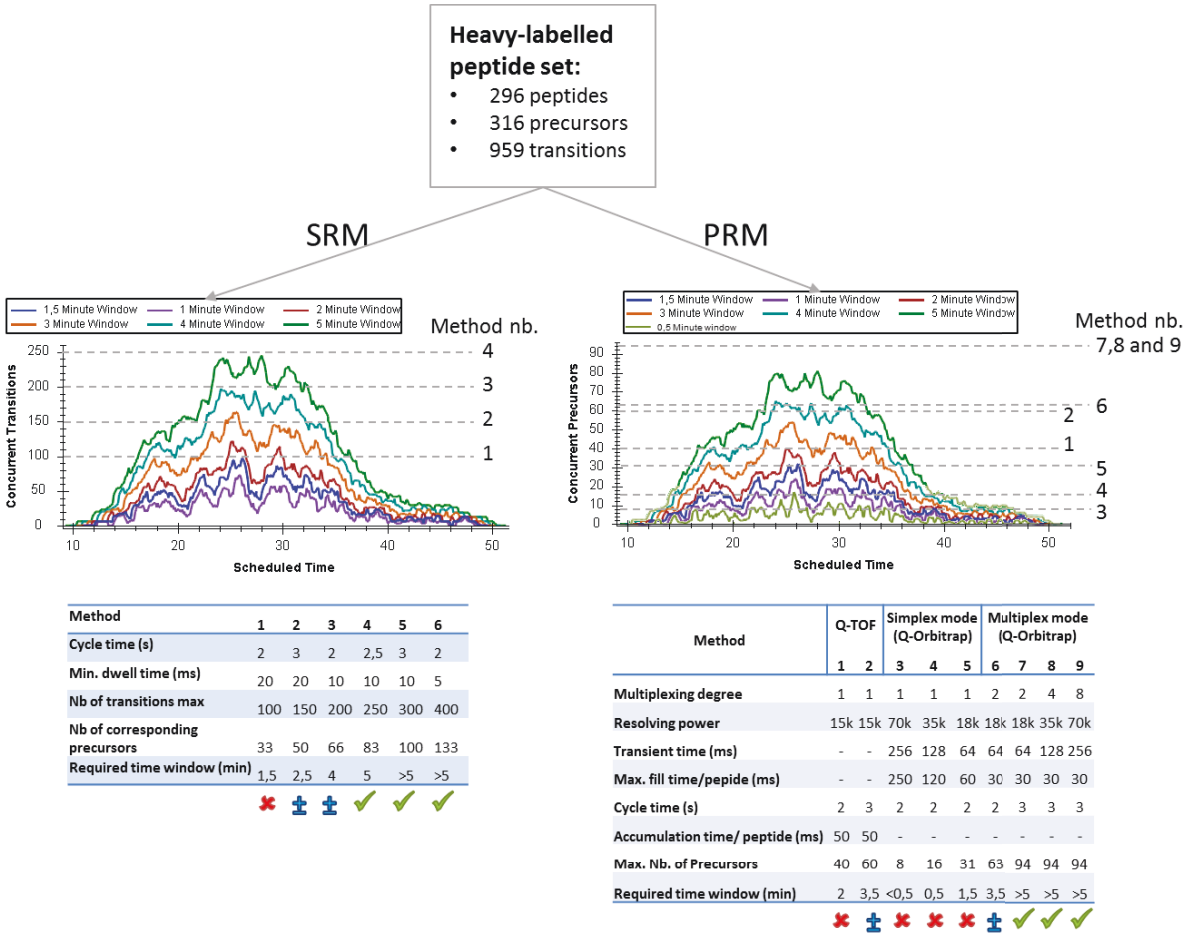

**Figure IV-23: Intra-gel and inter-gel replicate performance assessment of each sample preparation protocol.**
The results of the evaluation of the different protocols by label-free quantification are summarized here for the intra-gel (A) and the inter-gel replicates (B). The boxplots show the distribution of log2-transformed intensities in each sample after a normalization step. The Relative Log Expression (RLE) plots show the ratio between the intensity of a peptide and the median intensity of the peptide across all samples. The similarity of quantification results is summarized by the dendograms and the distribution of the CVs is also shown.

## C. Setup of an alternative targeted quantification method: Parallel Reaction Monitoring (PRM)

Figure IV-24 shows a comparison of the multiplexing capabilities of SRM versus PRM. For SRM and PRM different acquisition methods were evaluated. The instrumental parameters in Figure IV-24 are adapted to the instruments present in the laboratory. The SRM platform is a Thermo Scientific TSQ Vantage; the Q-TOF platform is an AB Sciex Triple-TOF 6600; and the Q-Orbitrap is a Thermo Scientific Q-Exactive Plus. The data presented here shows the scheduling of a heavy-labelled peptide set containing 316 targeted precursor ions and 959 transitions. This is a non-biased standard sample to evaluate the multiplexing capabilities as it mimics the elution profile of multiple peptides in a biological complex digest sample. In this section, the optimization of the chromatographic conditions to better separate the peptides and thus increase the multiplexing will not be discussed. This dataset will be used to evaluate whether or not the analysis of all the peptides can be done in a single run. It is important to keep in mind that scheduled windows higher than 5 minutes are commonly used in order to take account for common retention time shifts due to small differences in mobile phase composition, stationary phase, flow rate, temperature and matrix effects. Using shorter time windows lower than 2-minutes represents a high risk of missing the analytes. Indeed if the retention times shifts then a targeted peptide could elute before or after the scheduled time-window or it can be truncated rendering its quantification impossible. The scheduled time windows were set to be at least 5-minutes long in order to account for possible retention time variations. This value will be used as a metric for the comparison of SRM and PRM.

For SRM, at least 3 transitions per peptide were monitored. For the two first methods (Figure IV-24), the dwell times and cycle times were adjusted to enable an accurate and sensitive quantification. Using these methods all the transitions can be analyzed in a single run with Scheduled-SRM. However, the time windows need to be respectively of 1,5 min and 2,5 min. Using such small time windows size represents a high risk. It is thus not reasonable to use these methods to analyze this set of peptides. To be able to confidently quantify all the transitions in this peptide set, it is possible to reduce the minimum dwell time to 9-14 ms, knowing that the TSQ Vantage can go as low as 5ms. And keeping in mind that the average FWHM for a chromatographic peak in our LC conditions is 30 seconds, the cycle time can be increased to 2,5-3 seconds. Respectively 12 and 10 data points can be measured per peak in these conditions which is enough for an accurate quantification. With these parameters (SRM methods 3 and 4), all the transitions can be monitored within a single run with 4 and 5 min time windows.

It is important to understand that in a scheduled SRM method the cycle time is fixed and the dwell time is optimized according to the number of concurrent transitions. That means that dwell times are short only when a high number of concurrent transitions have to be measured. Otherwise the dwell time is longer, as the cycle time is divided by a smaller number of transitions, which allows increasing the sensitivity.




| Method | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Cycle time (s) | 2 | 3 | 2 | 2,5 | 3 | 2 |
| Min. dwell time (ms) | 20 | 20 | 10 | 10 | 10 | 5 |
| Nb of transitions max | 100 | 150 | 200 | 250 | 300 | 400 |
| Nb of corresponding precursors | 33 | 50 | 66 | 83 | 100 | 133 |
| Required time window (min) | 1,5 | 2,5 | 4 | 5 | >5 | >5 |

| Method | Q-TOF | | Simplex mode (Q-Orbitrap) | | | Multiplex mode (Q-Orbitrap) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Multiplexing degree | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 4 | 8 |
| Resolving power | 15k | 15k | 70k | 35k | 18k | 18k | 18k | 35k | 70k |
| Transient time (ms) | - | - | 256 | 128 | 64 | 64 | 64 | 128 | 256 |
| Max. fill time/pepide (ms) | - | - | 250 | 120 | 60 | 30 | 30 | 30 | 30 |
| Cycle time (s) | 2 | 3 | 2 | 2 | 2 | 2 | 3 | 3 | 3 |
| Accumulation time/ peptide (ms) | 50 | 50 | - | - | - | - | - | - | - |
| Max. Nb. of Precursors | 40 | 60 | 8 | 16 | 31 | 63 | 94 | 94 | 94 |
| Required time window (min) | 2 | 3,5 | <0,5 | 0,5 | 1,5 | 3,5 | >5 | >5 | >5 |

**Figure IV-24: Multiplexing comparison between SRM and PRM.**

These results show that the SRM multiplexing capability allows reliably measuring all peptides in the dataset within a single run. Additionally, the SRM method can still be tuned to achieve higher multiplexing (SRM methods 5 and 6). However using these screening methods the sensitivity and quantification accuracy are reduced.

Using a PRM method on an AB Sciex Triple-TOF 6600 Q-TOF instrument with a 50ms accumulation time for each MS/MS scan and a 2 second cycle time, the analysis using a scheduled-PRM method of all heavy-labelled peptides set requires the use of 2-min scheduling time windows (Method 1). For this Q-TOF instrument, scanning at a higher rate using smaller accumulation time is not recommended. Thus it would not be recommended to monitor the full set of peptides in a single run because of possible LC variations that can compromise the quantification. To be able to increase the multiplexing a cycle time of 3s can be used. In this case a 3,5-minutes time window is required (Method 2). In a recent study, Schilling et al. showed that scheduled-PRM on a Triple-TOF 5600 reached the same levels of sensitivity (dynamic range and LOD/LOQ) as SRM [191]. It achieved the quantification of around 500 peptides in a single run in different complex samples by using accumulation times of 50-60ms and 1-2 minutes time windows to obtain a maximum of 50 concurrent precursor ions at a given time. However, the number of analyzed samples is small (10 runs to evaluate RT shifts using a standard protein and a triplicate injection per sample matrix) and it is not possible to assess the risk of using such small scheduling time windows in large-scale analyses during several weeks.

Using a PRM method on a Q-Exactive Plus instrument results in the same problem. The limiting factor in this case is the transient time of the Orbitrap. To be able to quantify in a simplex mode (non-multiplexing mode) all precursors in our set, very small time-windows are required. Less than 30s time windows are required using a resolving power of 70000 or 35000 (Methods 3 and 4) which is not realistic. The same can be said for the method 3 using a resolving power of 17500. Using this method the time windows needed are smaller than 1,5 minutes. An advantage of the Q-

Exactive Plus is its ability to analyze multiple precursors in a multiplexed way. In the multiplex mode, several precursors are sequentially isolated and sequentially fragmented in the HCD cell. All fragments are trapped and accumulated, and then sent into the Orbitrap where they are analyzed together. This approach enables to increase the total multiplexing. As seen in Figure IV-24 the methods 6, 7, 8 and 9 use the multiplexing mode. These methods enable the analysis of all the precursors in our peptides set with correct scheduling time-window size. However, it is important to note that these methods have several downsides, the first being the fact that the fill times are considerably reduced (30 ms) and this highly reduces the sensitivity. Gallien et al. showed the negative impact of reducing the fill times to this extend on low-abundant peptides and suggest not using low fill times (30-60 ms) when co-analyzing wide m/z ranges (4x2m/z, 8x2m/z) [97]. Additionally, it is important to keep in mind that in a PRM analysis the intensity values observed in the MS/MS spectra are normalized values corresponding to the number of charges of the precursor and its corresponding fill time. In this context, in a multiplexed analysis the intensity values observed are wrongly normalized by the total fill time. A supplemental data processing step is necessary to correct for this. Finally the multiplexing approach can lead to the co-analysis of peptides with very different abundances. This can decrease the dynamic range and thus the overall sensitivity.

Additional methodological developments have been reported to increase the multiplexing capabilities notably by enabling the reduction of time scheduled window sizes. One of these approaches is “on-the-fly” retention time correction with the analysis of retention time standard peptides evenly distributed throughout the gradient [172]. Another approach termed IS-PRM uses internal standards to trigger PRM events and thus optimize the instrument analysis time and increase the multiplexing [4]. This approach promises to highly increase the data quality (as higher resolutions can be used) and the analysis throughput.

All in all, Parallel Reaction Monitoring enables the targeted analysis of peptides of interest with a higher selectivity due to the significantly higher resolving power (15k-70k) and higher accuracy (5-10ppm) when compared to SRM. It also provides greater assay flexibility as all fragments for a given peptide are simultaneously measured, enabling the post-acquisition refinement of signals by eliminating weak, noisy or interfered transitions. Several studies have compared PRM and SRM methods [4, 191-193]. These studies found that PRM reaches similar performances as SRM in selectivity, accuracy and sensitivity. And in certain conditions (IS-PRM) it can outperform it. However for the moment PRM is limited by its low multiplexing capabilities. Recent technological advances can alleviate this problem as faster scanning instruments have been developed [194]. Moreover, of SRM and PRM the latter is the one with the most potential for progress.

Though, triple quadrupole instrumentation is significantly low-cost when compared to Q-TOF or Q-Orbitrap instruments that could achieve the same analytical performances. The low-price and the ease-of-use of triple quadrupole instruments make them the most promising technology for routine analyses if mass spectrometry assays become commonly used in hospitals.

## D. Setup of performance standard samples for targeted and global quantification platforms

As seen above several new acquisition modes allowing large-scale protein quantification by mass spectrometry have recently emerged due to the remarkable technical progress. These can be grouped in three categories:

- Data dependent acquisition Label-free quantification based on MS signal.
- Targeted data acquisition (SRM/MRM and PRM).
- Data-independent acquisition (DIA) modes (Swath, MSX, $MS^E$…).

All these approaches have been shown to be suitable for protein quantification, and in this context it becomes the proteomist's task to choose the strategy fitting the best to the purpose of the quantification. The proteomist's work consists in developing and optimizing the quantification method by finely tuning the LC-MS parameters in order to correctly balance between the needed sensitivity, the highest accuracy, the highest selectivity and the broadest protein coverage (Figure IV-25). In order to assess the performances and the effects of the optimization, simple and standardized tools are needed.

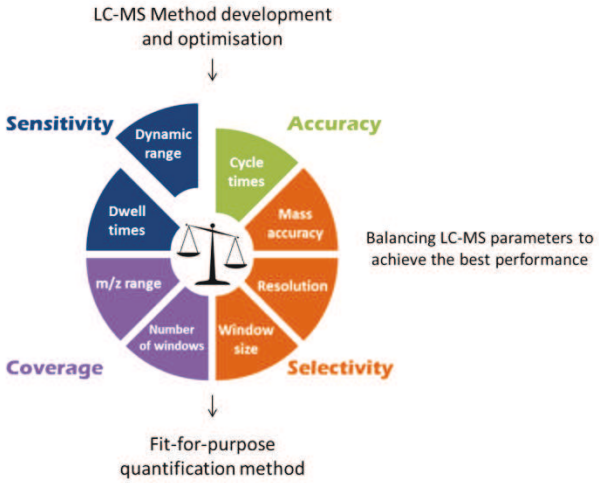



**Figure IV-25: Balancing LC-MS parameters to optimize quantification methods.**
The figure shows a list of LC-MS parameters impacting the accuracy, the selectivity, the proteome coverage and the sensitivity of a quantification method.

Additionally in order to obtain robust and reliable quantification data it is important to evaluate each step of the proteomics workflow, form sample preparation, LC-MS analysis to the bioinformatic data treatment. Indeed, each step is characterized by an inherent technical variance that is added to the total analysis variance. This is illustrated in Figure IV-26 which shows a fishbone diagram describing the technical variability that must be evaluated and controlled in a proteomic experiment [195]. The causes of the variability observed are separated in six categories. In an analytical study, the majority of the variance is due to the sample preparation step. But Piehowski et al. showed that the variance due to the LC-MS instrumentation can in some cases contribute to 25% of the overall variability in studies requiring long periods of time [196]. It is thus imperative to systematically verify the instrumental performances.

Undeniably acquiring data on an instrument with suboptimal performances will culminate in lower peptide identifications and irreproducible quantitative data that can even end in loss of precious biological samples. And more importantly it will be accompanied with a significant loss of the investment of personnel time, instrument time and, of course, financial resources. Moreover, if the data is not of good quality there will be no data transformation or statistical tool capable of correcting it.

In order to evaluate LC-MS performances several types of performance tests have been developed based on either peptide identifications or chromatogram signal extraction [195] using simple [197] or complex protein digests that can also be spiked with standard references [85, 198]. This evaluation must also identify the sources of the disturbances that handicap the instrumental performances and guide the troubleshooting corrective actions. To obtain robust quantification data the systematic assessment of the system performance must be planned from early stages of the experimental design.

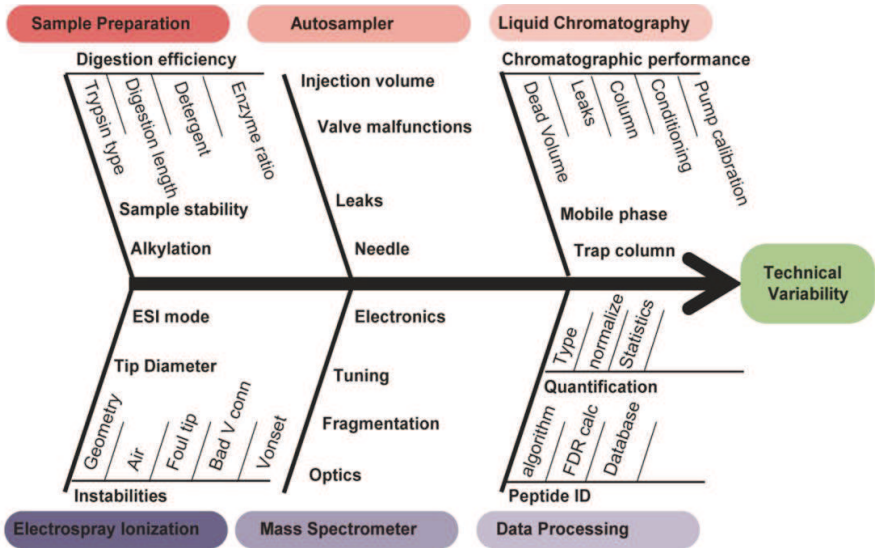



**Figure IV-26: A fishbone diagram describing the technical variation that must be evaluated and controlled in a proteomic experiment (Adapted from [195]).**

The choice of the bioinformatic workflow used for the protein identification and quantification is also of high importance. In quantitative proteomics the data processing consists in multiple sequential steps, such as pick picking, pick area integration, retention time alignment, data normalization, inference of protein abundances from peptide abundances, and application of statistical tests to find proteins with differential abundances. Several software tools have been developed in the last years for protein identification and protein quantification, such as MaxQuant [87], MFPaQ [88], Scaffold [199] or Skyline [15]. However, each software tool uses a different strategy to tackle each of these steps. The same tool can produce good or bad results according to the parameters used and the user's level of expertise. In order to evaluate these workflows a well-characterized dataset can be used to evaluate and find the best practices for protein quantification and determine the specificities of each bioinformatic tool.

### D.1.1 Well-characterized standard samples and datasets to evaluate proteomics workflows

In order to accurately and objectively evaluate a step or an entire proteomic workflow, a well-characterized and standardized sample must be used. This sample must be designed to set ground truth characteristics with which the results of the evaluation will be compared.

The LSMBO is one of the three nodes of the French Proteomics Infrastructure ProFI (together with IPBS from Toulouse and EdyP from Grenoble, http://www.profiproteomics.fr). This consortium aims at improving the fields of computing/bioinformatics, and method development for high throughput targeted and global quantitative proteomics. These methodological developments are applied to the dynamic analysis of biological systems and to biomarker discovery. In this context, one of ProFI's initial goals was to develop standardized samples and metrics for proteomics workflows performance evaluation and I have actively participated in this task.

To set up a standard sample, we have chosen the Universal Proteomics Standard (UPS1, Sigma) consisting of 48 purified human proteins that we spiked in a whole yeast proteome. To mimic relative quantitative changes produced by biological up-or down-regulations of a set of proteins, different spike concentrations were defined and prepared [85].

This tool revealed to be extremely useful to evaluate and improve analytical workflows such as the development of a sensitive and objective performance test for LC-SRM platforms, the side-by-side evaluation of label-free bioinformatic pipelines and the improvement of signal extraction of Data-Independent Acquisition data as described in the following sections.

## *D.2. Engineering of a sensitive and objective performance test for Targeted data acquisition (LC-SRM)*

### D.2.1 Designing a sensitive and highly multiplexed standard sample for the evaluation of LC-SRM platforms

In the context of the ProFI consortium it was necessary to assess the intra- and inter-laboratory transferability of SRM assays across the three laboratories. A standard sample and a standard targeted method needed to be developed in order to standardize the way that targeted quantification was done in each platform. Additionally, the LC-MS platforms present in each laboratory were not of the same vendors (AB Sciex Q-Trap5500 and 6500 and Thermo Scientific TSQ Vantage) and were not set to use the same flow-rates (nanoLC and capillaryLC, respectively). As part of the method transferability across LC-SRM platforms, it was important to develop a standardize performance test. The yeast+UPS1 sample was chosen to routinely verify the instrumental platforms performances. The key steps of the development of the performance test are described below.

First, each laboratory chose and identified the set of well-responding peptides for their corresponding LC-SRM platform. Since different LC and QqQ instruments were used, each platform found a different set of peptides. However a common list of peptides could be defined. 117 peptides were chosen for the 48 human proteins of the UPS1 mixture. Heavy isotopically stable labelled peptides were spiked in the sample to facilitate the identification of the peaks and correct LC-MS signal fluctuations. Additionally, retention time standard peptides (iRT standard peptides, Biognosys) were used to predict retention times. Each peptide was monitored using at least 3 transitions, totaling 669 transitions. The developed assay was a time-scheduled SRM method with 6-minutes time windows.

In order to obtain a reliable and accurate performance test that shows the state of an instrument, it should be sensitive to small changes in performances. For the TSQ Vantage platform, the only performance test recommended by the vendor is an infusion of a polytyrosine solution. This only enables to check for mass calibration, mass resolution and roughly for the sensitivity of the MS instrument. To check the coupling of the LC and the MS, we had introduced in the lab the injection of a BSA digest using a fast gradient to determine some chromatography and ESI problems (spray instabilities, poorly made connections, void volumes…). However this simple test does not give a full picture of the state of the instrument's performance and a complex standard mimicking a real biological sample analyzed with a long gradient is by far more relevant. It is also important to evaluate the instrument using a highly multiplexed method in order to detect the global state of the instrument and more importantly detect peptide-dependent variations by individually monitoring chosen peptides. That way the performance test will be sensitive to small changes in a subset group of peptides, indicating for example a bad ionization efficiency of hydrophilic peptides at the hydrophilic region of the chromatogram, spray instabilities or retention time shifts at the hydrophobic regions of the chromatogram.

Taking all these aspects into account, a 4-points dilution series of UPS1 (0.1, 1.0, 5.0, 10.0 fmol) spiked in 2µg of yeast was created to evaluate the performances of each platform in terms of sensitivity and linearity,. From this linearity experiment, the breaking point of the instruments performance was found to be around 1 fmol of UPS1 injected into the column. At this quantity the majority of peptides were below the limit of quantification. For the performance test, the amount of UPS1 to be spiked was thus determined to be 5fmol spiked in 1µg of yeast background matrix. The amount of yeast background matrix was reduced to 1µg of total protein injected into the column as 2µg seemed to damage the Nano-flow columns and a carry-over effect of highly hydrophobic peptides was observed. This quantity of background matrix was thus not compatible with a routine performance test evaluation.

The amount of UPS1 was chosen as it is close to the breaking point of the instrument's performances, i.e. a small deviation from the acceptable instrumental performances would induce very detectable consequential changes in the SRM traces of UPS1 peptides. At 5 fmol of injected UPS1 all the peptides were observed but the signals were close to the limits of detection and would thus enable easily detecting a loss in sensitivity.

The performance test, consisting of tryptic peptides of UPS1 proteins, their heavy-labeled peptide counterparts and 11 retention time standard peptides (iRT standard peptides, Biognosys) spiked-in a yeast lysate background matrix, was named the Performance Evaluation Standard (PES). The UPS1 proteins were spiked at 2,5fmol/µl in a 500ng/µl yeast lysate background matrix. Two microliters of the PES were injected into the LC-SRM system. The PES was used to routinely evaluate the performances of the instruments during a year and it was injected at least 4 times per month.

### D.2.2 Automated and rapid performance evaluation of LC-SRM platforms

To facilitate the data treatment and fasten the decision making of whether or not the instrument platform is suitable for accurate and precise analyses, an automated workflow was developed.

The skyline open-source software was chosen for the data treatment steps [15]. Heavy-labelled peptides were spiked in for every targeted UPS1 peptide to facilitate the peak picking and integration. The heavy-labelled peptides were spiked in a sufficient amount to always be detected but not suppress the signal of the targeted peptides. The use of these standard peptides eliminated the need to manually verify all SRM traces fastening the data analysis step. Additionally Retention time peptides were used to correct for shifts in retention time. The PES was designed to be used as a routine performance test throughout long periods of time. In this time period the retention times of the peptides can change as the chromatographic conditions can change (column changes, solvent changes, different tubing lengths …). To account for this, the nominal retention times were not used instead normalized retention times and relative to the standard RT peptides were used. The description of the set of global and individual criteria that were established is given below.

To obtain a general view of the instrument's performance the global criteria were the following: first the chromatogram was divided in three regions: the hydrophilic part, the intermediate part and the hydrophobic part. The boundaries of each region were set by the elution times of RT standard peptides (Figure IV-27). That way the retention time shifts were taken into account and the three regions will always be composed of the same set of eluting peptides. Then, for each region the number of transitions observed with S/N ratio > 3 were counted and compared to an expected value with a tolerance value. This makes the performance test sensitive to small changes in a subset group of peptides and can thus guide to the appropriate troubleshooting procedure.

Furthermore, the performance test also follows a restricted set of peptides to determine peptide-dependent variations. The so called individual criteria are the following: six peptides well-distributed across the gradient were

chosen. For each peptide the best transitions were chosen and the following information was extracted: the peak area, the peak height (intensity) and the full width at half maximum were compared to a reference value. The reference values were set by analyzing the PES multiple times and determining the values that enabled to confidently discriminate between a good and a bad instrument state.

As stated before, the nominal retention times could not be used. This is why we developed a mean to use normalized retention times to RT standard peptides. To do this, for each peptide we calculated the difference in time between the peptide's RT and the RT of the previous RT standard peptide (value $a_{mesured}$ in Figure IV-27.B) and divided it by the difference in time of the two RT standard peptides surrounding the peptide (value $b_{mesured}$ in Figure IV-27.B). The difference of the measured a/b ratio and the a/b reference ratio was used to calculate the difference in time between the RT of the measured peptide and the predicted RT for that same peptide (Δ in Figure IV-27.B). The difference has to be smaller than half of the time tolerance defined by the user (c Figure IV-27.B). This means that the retention time of the peptide falls within the time tolerance window defined by the user.

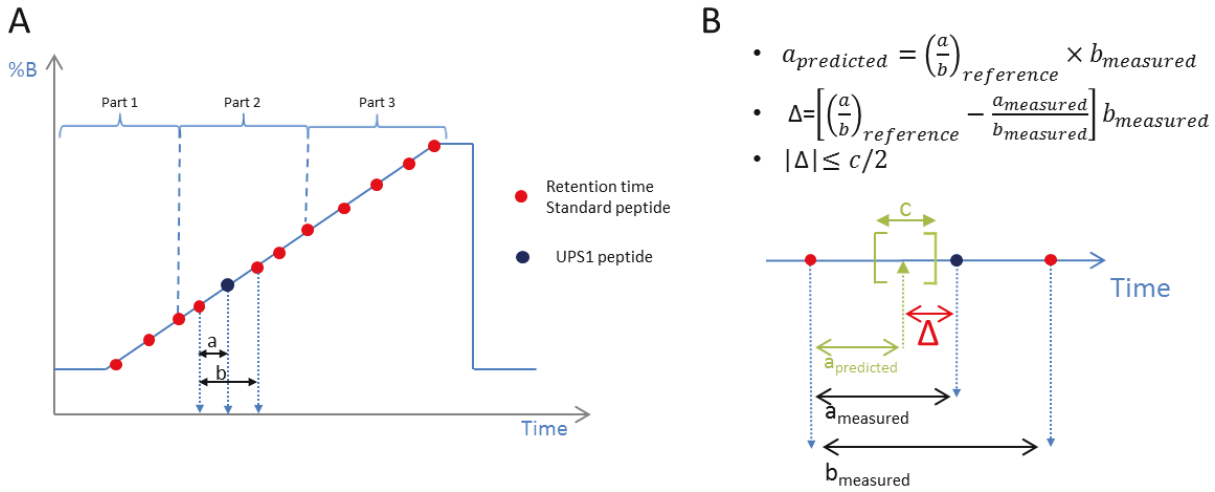



**Figure IV-27: Engineering a performance test using standard retention time peptides to correct for RT shifts in long periods of time.**

A. Normalized retention times and relative to the standard retention time peptides were used. A ratio was calculated using the RT value of the target peptide and the two Retention time standard peptides surrounding the peptide (a/b ratio). B. The difference in time between the RT of the measured peptide and the predicted RT (Δ) is calculated. In order to classify a peptide as having the correct retention time, the observed RT should fall within the time tolerance window defined by the user (c).

The choice of the reference values is a challenge as they have to be sensitive enough to detect perturbations in the system's performance, but they don't have to become an obstacle to the use of the instrument. For the intensity and the peak areas, the defined threshold value was chosen after running several PES analyses and it was defined as 70% of the average value. For the retention time, the reference value is a/b ratio and a user-defined tolerance time window. For the FWHM, it was the maximum FWHM observed after several analyses of the PES.

An in-house Excel tool was developed to fasten the data analysis, and rapidly and confidently decide whether or not the instrument platform is suitable for accurate and precise analyses, or if cleaning and maintenance are required. This macro only needs the peptide sequences for which the information is going to be extracted.

The output of this tool is a table giving a pass/fail summary (Figure IV-28.A).

### *D.3. Results of a year-long routine evaluation of LC-SRM platforms*

The PES was used to routinely assess the LC-SRM instrument performances for more than a year. The monitoring of the retention times, scheduling window and peak areas of a targeted peptide over a year can be seen in Figure

IV-28.B. An abnormal event is shown by a red arrow. In this case it was the chromatography that was in cause as the peptide was eluted outside of the scheduling window. Figure IV-28.C. shows the results using another metric. The comparison of FWHM shows the clear difference of a well-performing system and a suboptimal system.

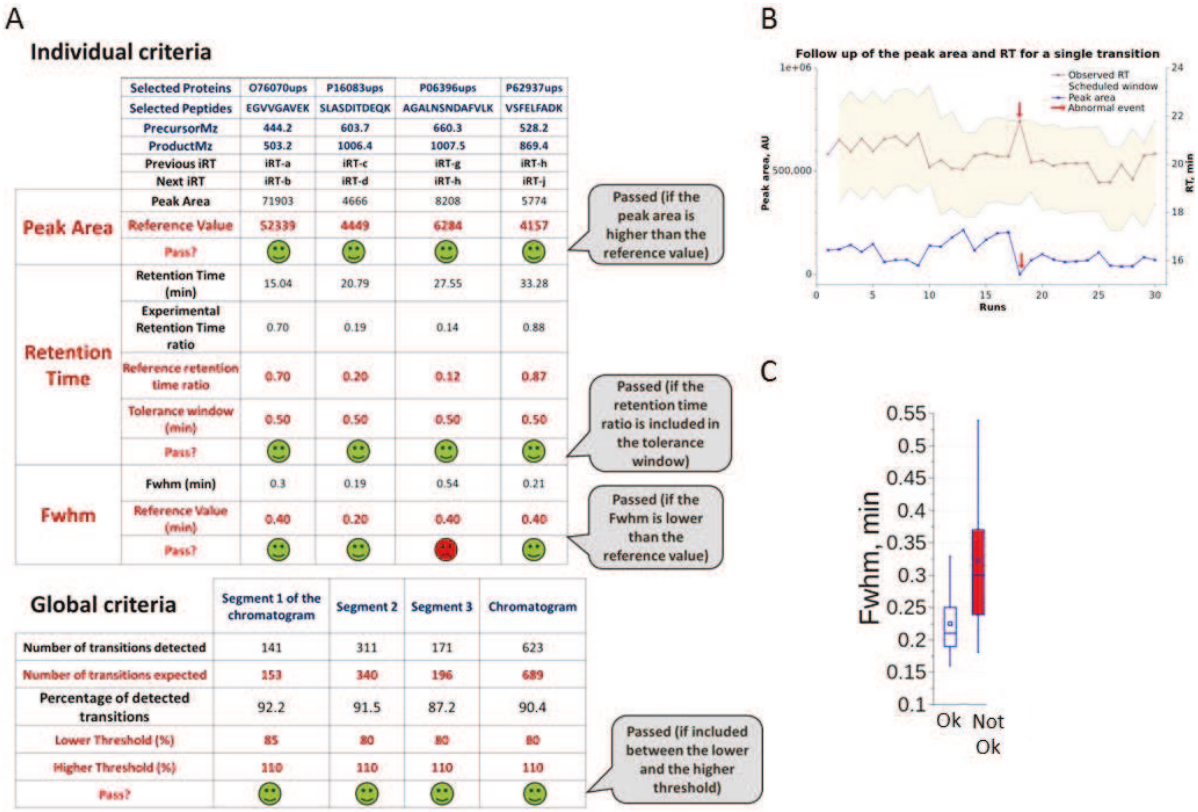



**Figure IV-28: Pass/Fail performance test applied to a year-long routine evaluation of LC-SRM platforms.**
A. The individual and global criteria and the thresholds are listed in this table in red. The values extracted from a given LC-SRM analysis are in black. This is a pass/fail performance test sensitive enough to detect any perturbations and direct the most appropriate troubleshooting. B. Monitoring of the retention times, scheduling window and peak areas of a targeted peptide over a year. C. Comparison of FWHM of a well-performing and a suboptimal system.

## D.4. PES applied to the benchmarking of Label-free LC-MS data processing workflows

The results presented here were published in 2016 in the *Journal of Proteomics* and the dataset was published in the journal *Data in brief*. The first paper is available on page 115.

### D.4.1 Context of the project

Label-free quantification is based on high-throughput peptide sequencing by LC-MS/MS. This approach has emerged in recent years due to significant technological improvements of mass spectrometers in terms of sequencing-speed, resolution, and dynamic range. This technique is a powerful tool to deeply characterize and quantify whole proteomes. However, it is important to objectively assess label-free methods and bioinformatic pipelines.

This project is also inscribed in the context of the ProFI consortium which objectives are, in part, to optimize all the analytical steps involved in a workflow for global protein quantification and to define robust analytical methods.

The accuracy of the quantification by label-free methods is highly dependent on the bioinformatic pipeline used to process the data. Using the yeast+UPS1 standardized sample, the aim of this study was to make a side-by-side evaluation of several bioinformatic tools, namely MaxQuant [87], MFPaQ [88] and Skyline [15]. Having a sample providing a ground truth, the different bioinformatic pipelines were dissected in order to determine the best parameter sets to be used for each one.

### D.4.2 Experimental design

Nine samples were prepared by spiking different quantities of UPS1 (Sigma) standard into a yeast background matrix. The concentrations spanned from 50 amol to 50 fmol of UPS1 proteins in 1 µg of yeast lysate. Protein samples were digested with trypsin, and resulting peptides were analyzed by nanoLC–MS/MS on a LTQ Velos-Orbitrap instrument using a Top 20 data-dependent acquisition. Each sample was analyzed in triplicate resulting in 27 .raw data files (Figure IV-29).

This dataset was then used to evaluate different quantitative workflows. In total, the protein identification was done using two different software (Mascot and Andromeda) and the protein quantification was done using 5 different tools (Scaffold, MFPaQ and IRMa/hEIDI for the spectral count quantification; MaxQuant, MFPaQ and Skyline for the XIC MS1 Label-free quantification). Finally, 8 different quantitative datasets were obtained.

Three pairwise comparisons were chosen to evaluate the different bioinformatic pipelines. These were the comparison A (500 amol/µg versus 50 fmol/µg; Fold change =100), comparison B (5 fmol/µg versus 50 fmol/µg; Fold change =10) and comparison C (12.5 fmol/µg versus 25 fmol/µg; Fold Change=2). They respectively mimicked a condition where in one sample the proteins are under the detection level of the instrument, a high fold change up-regulation and a low fold change up-regulation, thus covering the spectrum of what can be observed in a real complex biological sample.

The same statistical processing method was performed on all results output files of each pipeline. The criteria used to compare the software were the true positive rate (the number of UPS1 successfully classified as variant) and the false discovery proportion (the number of yeast proteins erroneously classified as variant).

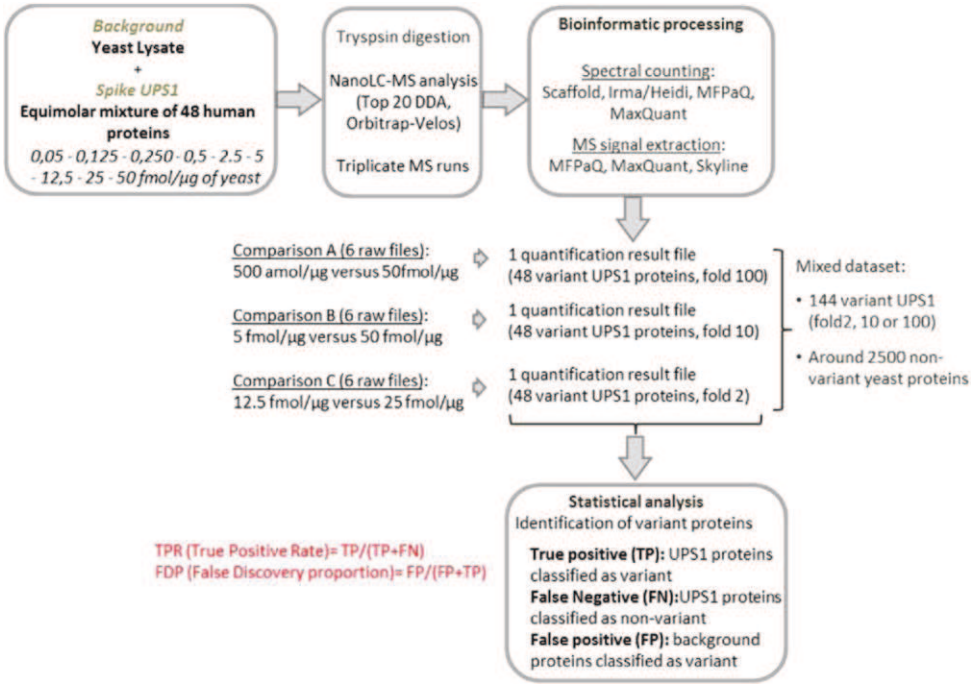



**Figure IV-29: Experimental design (adapted from [85])**
A series of 9 yeast lysate samples spiked with growing concentrations of the Sigma UPS1 standard was analyzed in triplicate by nanoLC–MS/MS mass spectrometry on a LTQ Velos-Orbitrap instrument. Different computational workflows were used to identify, validate, and quantify proteins based on spectral counting or MS signal analysis. In the present study, 3 different pairwise quantitative comparisons (A, B, and C) were performed between samples spiked with different amounts of UPS1, involving in each case the quantification of 6 raw files (2 conditions × 3 replicates), trying to mimic distinct biochemical situations. The 3 individual quantitative datasets containing protein abundance values were then gathered. This global quantitative dataset was generated for each data processing workflow, and identical downstream statistical processing methods were then applied for classification of variant proteins.

### D.4.3 Characteristics of the Skyline Software

During this project, I was in charge of evaluating the Label-free XIC MS1 quantification using Skyline. Skyline is an open-source software capable of importing .raw files from different vendors (Agilent, AB Sciex, Thermo Scientific, Bruker Daltonics, Waters) and acquired using different acquisition modes (DDA, SRM, PRM and DIA). It has a large and growing user community. This is why we chose to evaluate this software.

Skyline is a very flexible software. It does not have a default parameters setting or a single-path workflow. This is a very useful feature that enables to use Skyline for many applications. However, this can also be a problem for non-experienced users, as there are many parameters to be set correctly to obtain an accurate quantification.

Moreover, Skyline does not have an implemented algorithm for protein identification, protein validation nor protein grouping. All these steps have to be performed previously in another software tool. For this project, the Mascot searches followed by Scaffold validation [199] was chosen as it was the commonly used workflow in the laboratory at that time. It is important to understand that Skyline can extract signals of any defined target, whether it is a validated or non-validated peptide. This can highly increase the number of false positives. We have found that to obtain an accurate quantification it is best to perform the protein quantification using only validated peptides.

Additionally, Skyline has two drawbacks that handicap its performance: (i) it does not perform a recalibration of the m/z dimension and (ii) it does not align the retention times of the quantified peptides.

(i) Skyline extracts for a given peptide the signal of its corresponding precursor m/z ratio with a mass tolerance related to the resolving power of the instrument. If the mass calibration shifted considerably during the analysis this could not be corrected and the quantification will be handicapped (Figure IV-30). This highlights the importance of working with

well-performing instruments and systematically controlling the instrument's state with quality control samples. In the case of a shift of the mass calibration a higher tolerance of the extraction mass can be set, but this reduces the selectivity of the method.

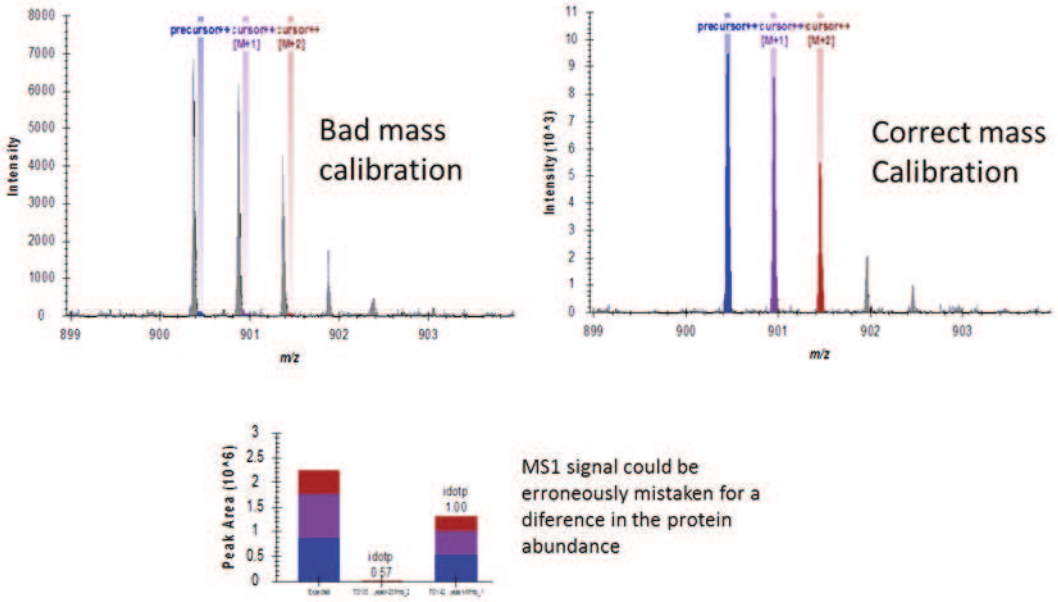



**Figure IV-30: Example of the negative consequences of a badly calibrated instrument.**
Skyline extracts the signal of the precursor calculated m/z ratio with a mass tolerance related to the resolving power of the instrument. The color squares represent the signal extraction window. If the mass calibration is not done correctly the extraction will be wrong and the quantification will be handicapped.

(ii)The lack of retention time alignment considerably hinders the quantification with Skyline. Skylines uses an alignment algorithm to direct the chromatogram extraction in the time region where a peptide should elute according to a spectral database or using a prediction based on retention time standard peptides. However, once the signal is extracted, another algorithm will try to detect the best peak in these, eventually multiple, extracted signals. These two steps are independent so that the peak picking does not benefit from a retention time alignment step and this has an impact on the selection of the peak group to be quantified.

### D.4.4 Results of the evaluation of Skyline for MS1 Label free quantification

Figure IV-31 shows the results of the evaluation. The volcano plots, showing the negative logarithmic of the t-test p-values plotted against the base-2 logarithmic fold changes, are given for the raw skyline output results and for the output after manual validation. The criteria to consider a protein as having a statistically significant changing abundance between two comparisons (variant protein) are in this case a t-test p-value lower than 0.05 and a fold change higher than 2. The combination of these two criteria is very important to discriminate false from true positives. The use of the p-value alone is not recommended as the statistical test shows if, in a pairwise comparison, the differential expression is different from zero, but it does not show if the difference observed is biologically meaningful. This can clearly be seen in the volcano plots (Figure IV-31. left panels) where the yeast proteins are represented by grey diamonds. If only the p-value is used to validate the significance of a differential expression at the commonly employed value of 0.05 (represented by the horizontal dashed line), the quantification results would have hundreds of proteins wrongly declared as variants.

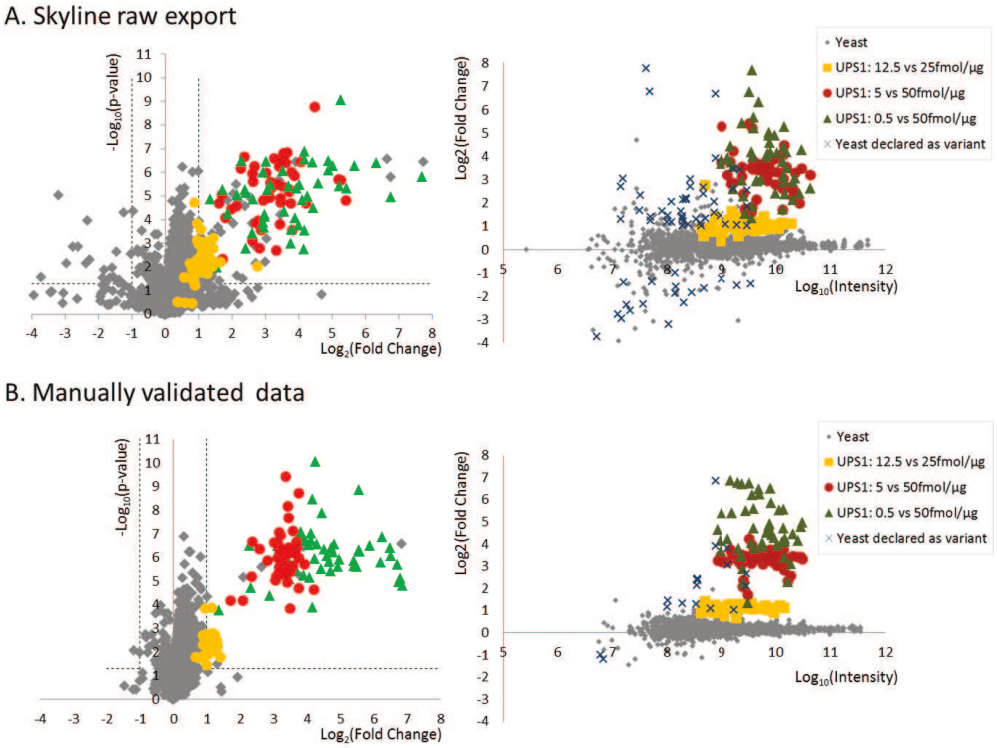



**Figure IV-31: Manual peak picking and integration correction and validation.**
Volcano plots showing the negative logarithmic of the t-test p-values plotted against the base-2 logarithmic fold changes and base-2 logarithmic fold changes plotted against the logarithmic intensities. The results are shown for the Skyline raw output (A) and after manually validating each MS1 signal (B).

We can see that the volcano plot made using the Skyline .raw output files shows many yeast proteins considered as variant and many UPS1 proteins are falsely considered as non-variant proteins in the comparison C (yellow points) (Figure IV-31.A). The overall true positive rate and the false discovery proportion, defined in Figure IV-29, were used as the criteria to assess the quality of the quantification. In this case, they were respectively 88% and 22%. When compared to the other bioinformatic pipelines, Skyline showed the poorer results. In the left panel of Figure IV-31.A the fold changes are plotted against the intensity in log scale. The blue crosses show false positives proteins (yeast proteins erroneously classified as variant) and it can clearly be seen that the majority of false positives are of low intensity, suggesting that the errors in the quantification are mostly done on low-abundant peptides close to the LOQ and where irreproducible signals are common. The peak picking is thus more challenging for these peptides.

In order to have an accurate and reproducible quantification, each signal has to be manually validated. This is a very time-consuming step and subjective to the user interpretation. At the time of this evaluation the mProphet [16] automatic validation was not implemented into Skyline and was thus not evaluated. However, it was later tested and did not significantly improve the results.

### D.4.5 Improving the results of MS1 Label free quantification by manually validating each MS1 signal

Contrary to other software, Skyline shows the peak integration boundaries and they can be manually corrected. The manual correction and validation of every MS1 signal was carried out. The results can be seen in Figure IV-31.B. After the manual validation the results clearly improved. The overall true positive rate and the false discovery proportion were respectively 97% and 7% and these results were better than any other quantification workflow. Another important result is that the false positive proteins (blue crosses) distributed in the intensity axis (left panel Figure IV-31.B). This means that there is no longer a correlation between the intensity of the proteins and the fact that they are erroneously considered as variant.

Each false positive protein was examined and we found that the reason of their wrong classification was that they had an isobaric peptide that had exactly the same mass as an UPS1 peptide and they were eluted at close (or at the same) retention time. This means that UPS1 peptides contaminated the yeast signals and thus the corresponding proteins were classified as variants. This can be seen in the volcano plot in Figure IV-31.B. as the majority of false positives appeared in the same reagion as UPS1 proteins. An example is illustrated in Figure IV-32, the peptide FVGTAVNFEDNLR belonging to a yeast protein is an isobaric peptide with the exact same mass as peptide AFYNVLNEEQR belonging to an UPS1 protein. The MS/MS spectra that served to identify each peptide are shown. This peptide was only observed when the UPS1 protein was present in a low concentration. The XICs for these two peptides are thus exactly the same. In the bottom panels, the histograms show the total intensities in 5 samples analyzed in triplicate; these also describe the same trend for the two peptides. This clearly shows a problem that is inherent to MS1 XIC label-free quantification. In this case the two pepetides are isobaric but this problem can also exist when there are coeluting peptides with prescursor masses that are close to each other.

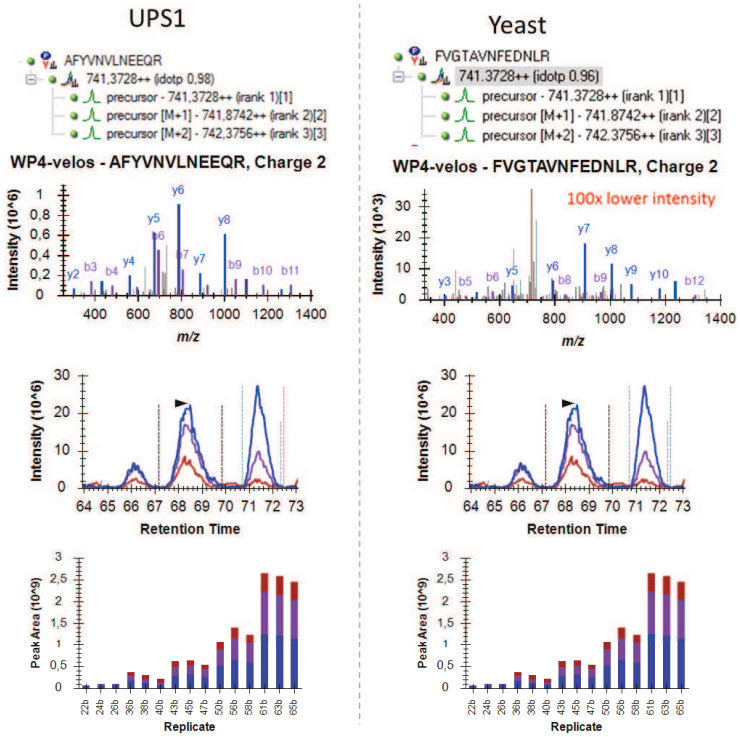



**Figure IV-32: Example of isobaric peptides with the same RTs distorting the quantification results at the MS1 level.**
Precursor masses of peptide FVGTAVNFEDNLR belonging to a yeast protein and peptide AFYNVLNEEQR belonging to an UPS1 protein. The MS/MS spectra that served to identify each peptide and the chromatograms are shown for each peptide. In the bottom panels the histograms show the total intensities in 5 samples analyzed in triplicate.

### D.4.6 Conclusion and perspectives

In conclusion, the percentage of false positives in label-free quantification remains still significantly high (8-22% of false positives). A way to reduce the number of false positives is to manually verify the quantitative information extracted from the raw data, verify and correct the peak picking and the integration boundaries. This is a unique feature of Skyline and most alternative software do not allow this manual correction of peak integrations.

There are still challenges that are inherent to the nature of Label-free quantification using MS1 signals like the coelution of peptides with close prescursor masses. A way to overcome these limitations requires the increase of the resolving power of the instrument or the use of MS2 information as it is the case when using Data-Independent Acquisition. However, we will see in the Part IVChapter IE.3 on page 141 that other challenges arise with the use of DIA.

Journal of Proteomics 132 (2016) 51–62

Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com/locate/jprot

# Benchmarking quantitative label-free LC–MS data processing workflows using a complex spiked proteomic standard dataset

Claire Ramus [a,d,e,f,1], Agnès Hovasse [a,g,1], Marlène Marcellin [a,b,c,1], Anne-Marie Hesse [a,d,e,f,1], Emmanuelle Mouton-Barbosa [a,b,c], David Bouyssié [a,b,c], Sebastian Vaca [a,g], Christine Carapito [a,g], Karima Chaoui [a,b,c], Christophe Bruley [a,d,e,f], Jérôme Garin [a,d,e,f], Sarah Cianférani [a,g], Myriam Ferro [a,d,e,f], Alain Van Dorssaeler [a,g], Odile Burlet-Schiltz [a,b,c], Christine Schaeffer [a,g], Yohann Couté [a,d,e,f], Anne Gonzalez de Peredo [a,b,c,*]

[a] ProFi, Proteomic French Infrastructure, France
[b] CNRS UMR5089 Institut de Pharmacologie et de Biologie Structurale, 205 route de Narbonne, 31077 Toulouse, France
[c] Université de Toulouse, 118 route de Narbonne, 31062 Toulouse, France
[d] CEA, DSV, iRTSV, Laboratoire de Biologie à Grande Echelle, Grenoble F-38054, France
[e] INSERM U1038, Grenoble F-38054, France
[f] Université Grenoble, F-38054, France
[g] Laboratoire de Spectrométrie de Masse BioOrganique (LSMBO), IPHC, Université de Strasbourg, CNRS, UMR7178, 25 Rue Becquerel, 67087 Strasbourg, France

ARTICLE INFO

*Article history:*
Received 4 September 2015
Received in revised form 4 November 2015
Accepted 8 November 2015
Available online 14 November 2015

*Keywords:*
Proteomic standard
NanoLC–MS/MS
Label-free quantification
Computational proteomics
Spectral counting
MS signal analysis

ABSTRACT

Proteomic workflows based on nanoLC–MS/MS data-dependent-acquisition analysis have progressed tremendously in recent years. High-resolution and fast sequencing instruments have enabled the use of label-free quantitative methods, based either on spectral counting or on MS signal analysis, which appear as an attractive way to analyze differential protein expression in complex biological samples. However, the computational processing of the data for label-free quantification still remains a challenge. Here, we used a proteomic standard composed of an equimolar mixture of 48 human proteins (Sigma UPS1) spiked at different concentrations into a background of yeast cell lysate to benchmark several label-free quantitative workflows, involving different software packages developed in recent years. This experimental design allowed to finely assess their performances in terms of sensitivity and false discovery rate, by measuring the number of true and false-positive (respectively UPS1 or yeast background proteins found as differential). The spiked standard dataset has been deposited to the ProteomeXchange repository with the identifier PXD001819 and can be used to benchmark other label-free workflows, adjust software parameter settings, improve algorithms for extraction of the quantitative metrics from raw MS data, or evaluate downstream statistical methods.

*Biological significance:* Bioinformatic pipelines for label-free quantitative analysis must be objectively evaluated in their ability to detect variant proteins with good sensitivity and low false discovery rate in large-scale proteomic studies. This can be done through the use of complex spiked samples, for which the "ground truth" of variant proteins is known, allowing a statistical evaluation of the performances of the data processing workflow. We provide here such a controlled standard dataset and used it to evaluate the performances of several label-free bioinformatics tools (including MaxQuant, Skyline, MFPaQ, IRMa-hEIDI and Scaffold) in different workflows, for detection of variant proteins with different absolute expression levels and fold change values. The dataset presented here can be useful for tuning software tool parameters, and also testing new algorithms for label-free quantitative analysis, or for evaluation of downstream statistical methods.



* Corresponding author at: CNRS UMR5089 Institut de Pharmacologie et de Biologie Structurale, 205 route de Narbonne, 31077 Toulouse, France.
*E-mail addresses:* claire.ramus@cea.fr (C. Ramus), ahovasse@unistra.fr (A. Hovasse), marlene.marcellin@ipbs.fr (M. Marcellin), anne-marie.hesse@cea.fr (A.-M. Hesse), emmanuelle.mouton@ipbs.fr (E. Mouton-Barbosa), bouyssie@ipbs.fr (D. Bouyssié), sebastian.vaca@etu.unistra.fr (S. Vaca), ccarapito@unistra.fr (C. Carapito), karima.chaoui@ipbs.fr (K. Chaoui), christophe.bruley@cea.fr (C. Bruley), jerome.garin@cea.fr (J. Garin), sarah.cianferani@unistra.fr (S. Cianférani), myriam.ferro@cea.fr (M. Ferro), vandors@unistra.fr (A. Van Dorssaeler), schiltz@ipbs.fr (O. Burlet-Schiltz), christine.schaeffer@unistra.fr (C. Schaeffer), yohann.coute@cea.fr (Y. Couté), gonzalez@ipbs.fr (A. Gonzalez de Peredo).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jprot.2015.11.011
1874-3919/

## 1. Introduction

Label-free quantitative methods based on LC–MS/MS have become increasingly popular in proteomic studies, as an attractive and powerful way to analyze differential protein expression in complex biological samples [1–3]. They can be based either on the measurement of the MS/MS sampling rate for a particular protein (spectral counting), or on the MS chromatographic peak area of its corresponding peptides in the survey MS scan (MS trace analysis), both values being directly related to protein abundance. Both approaches have benefited from tremendous improvements in instrumentation, namely increased sequencing speed for spectral counting approaches (up to 15–20 Hz in recent orbitrap or Q-TOF mass spectrometers) and higher resolution allowing more accurate MS signal analysis and improved matching of complex LC–MS maps. These methods have concomitantly gained in analytical depth, and can now routinely be used to profile the expression of thousands of proteins from biological systems submitted to different conditions. An important point is however to be able to assess, minimize, and eventually correct the variability associated to the LC–MS/MS analytical workflow, to ensure sufficient repeatability of the measurements and provide robust relative quantification of proteins across samples. To this respect, the development of proteomic standards has proved to be essential to assess the performances of LC–MS platforms, provide a quality control of the system and identify potential sources of variability. Importantly, they are also needed to evaluate the downstream elements of the analytical pipeline, i.e. bioinformatics processing and statistical analysis, which represent critical steps to generate the final comparative results.

The yeast *Saccharomyces cerevisiae* proteome has been used in many studies as a test sample to illustrate the benefits of various technological optimizations in the LC–MS/MS workflow. Due to its wide availability and relatively high complexity and dynamic range, it can be considered as a good surrogate to many real biological samples, both for method development and quality control. In previous studies, yeast samples have been used to establish and demonstrate the efficiency of a wide range of metrics to evaluate the LC–MS/MS performances [4,5]. These metrics were directly related to the LC system, the MS instrument (electrospray source, MS1 and MS2 intensities), the dynamic sampling, and also the first steps of data processing, i.e. peptide identification results. They were applied by Paulovich et al. for LC–MS benchmarking of several instrumental systems operated in different laboratories [6]. Instead of focusing on specific proteins or peptides, the monitoring proposed by the authors allowed them to give a global and very exhaustive view of the quality of the analysis through general metrics reflecting for example the median peak FWHM on the whole peptide population, the number of MS1 or MS2 scans triggered over various portions of the chromatogram, the level of TIC, the median MS1 signal for the population of identified precursors, or the number of peptides and proteins identified.

However, the final objective of most label-free studies is to measure quantitative levels, and detect variation of some proteins across samples. To evaluate the performances of a workflow in this respect, it is relevant to use a standard spiked with known amounts of some peptides or proteins, which can then be specifically monitored to assess the ability of the analysis in detecting relative quantitative changes. Controlled datasets based on spike-in experiments thus represent a useful tool to objectively assess the performances of quantitative methods for differential analysis. Paired comparison between spiked versus non-spiked samples can be performed to benchmark analytical and computational pipelines for biomarker discovery. Such controlled datasets with known "ground truth" have been for example generated in the past in the field of microarray analysis, by spiking at different concentrations a panel of 100–200 specific RNAs into a well-defined, constant background of RNA species [7], and was then widely used as a gold standard to evaluate various data processing methods [8–12]. In the proteomics field, spiked samples are also often used to evaluate MS methods or data processing tools, although generally the number of spiked proteins or peptides is relatively low [13–16]. Interestingly, as exemplified in the report from Paulovich et al. [6], the use of a more complex spiked material, such as the UPS1 standard containing 48 well-characterized purified proteins, allowed the authors to compute more extensively the exact proportion of false discoveries (number of yeast false positives relative to the total number of proteins declared as variant) and of true discoveries (number of true positives out of the 48 real variant UPS1 proteins). As a proof of concept of the kind of benchmarking that can be done with this spiked standard, they showed the performances of a spectral count approach (the SASPECT method) for detection of biomarkers when comparing the spiked sample (simulating a case sample) and the pure yeast reference sample (control sample).

In the present study, we wanted to extend this concept and use the yeast + UPS1 standard to benchmark several tools developed in recent years for relative quantification, including widely used software such as MaxQuant and Skyline. Indeed, while numerous software tools have been developed and are more and more routinely used for label-free quantitation, stringent and side-by-side evaluations have to be performed to prove the efficiency of the quantification. In addition, proper tuning and parameter settings in each of these software tools are also important for optimal downstream analysis. We thus generated a dataset from yeast samples spiked with 9 different concentrations of UPS1, analyzed in triplicate on an Orbitrap-Velos mass spectrometer. Starting from this dataset, different data processing workflows were implemented to perform relative quantification of proteins. Common statistical tests and fold-change criteria were used to identify differential peptides and proteins, for several theoretical fold variations of the spiked UPS1 standard. This experimental design allowed us to assess the performances of several workflows (4 based on spectral-count analysis and 4 based on MS signal analysis) in discovering true positive (UPS1 proteins successfully classified as variant) and avoiding false positive (yeast proteins erroneously detected as variant). Overall, this study allowed to objectively evaluate label-free quantitative methods and concretely illustrate what one can expect from these approaches in terms of false discovery proportion and sensitivity for the detection of variant proteins.

## 2. Experimental procedures

### 2.1. Sample preparation

A yeast cell lysate was prepared in 8 M urea/0.1 M ammonium bicarbonate buffer, protein concentration was adjusted at 8 μg/μL after Bradford assay, and this lysate was used to resuspend and perform a serial dilution of the UPS1 standard mixture (Sigma). Twenty microliters of each of the resulting samples, corresponding to 9 different spiked levels of UPS1 (respectively 0.05–0.125–0.250–0.5–2.5–5–12.5–25–50 fmol of UPS1/μg of yeast lysate), was reduced with DTT and alkylated with iodoacetamide. The urea concentration was lowered to 1 M by dilution, and proteins were digested in solution by the addition of 2% of trypsin overnight. Enzymatic digestion was stopped by the addition of TFA (0.5% final concentration).

### 2.2. NanoLC–MS/MS analysis

Samples (2 μg of yeast cell lysate + different spiked levels of UPS1) were analyzed in triplicate by nanoLC–MS/MS using a nanoRS UHPLC system (Dionex, Amsterdam, The Netherlands) coupled to an LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). 2 μL of each sample was loaded on a C-18 precolumn (300 μm ID × 5 mm, Dionex) at 20 μL/min in 5% acetonitrile, 0.05% TFA. After 5 min desalting, the precolumn was switched online with the analytical C-18 column (75 μm ID × 15 cm, in-house packed with C18 Reprosil) equilibrated in 95% solvent A (5% acetonitrile, 0.2% formic

acid) and 5% solvent B (80% acetonitrile, 0.2% formic acid). Peptides were eluted using the following gradient of solvent B at 300 nL/min flow rate: 5 to 25% gradient during 75 min; 25 to 50% during 30 min; 50 to 100% during 10 min. The LTQ-Orbitrap Velos was operated in data-dependent acquisition mode with the XCalibur software. Survey scan MS were acquired in the Orbitrap on the 300–2000 m/z range with the resolution set to a value of 60,000. The 20 most intense ions per survey scan were selected for CID fragmentation and the resulting fragments were analyzed in the linear trap (LTQ). Dynamic exclusion was employed within 60 s to prevent repetitive selection of the same peptide.

### 2.3. MS data processing

The dataset was processed according to different workflows listed in Table 1, consisting in the following steps: peaklist generation, database search, validation of the identified proteins and extraction of quantitative metric (spectral count or MS signal). According to the different tools used for each step, eight distinct workflows were evaluated. The same databases were used for peptide identifications: yeast database from UniprotKB (S_cerevisiae_ 20121108.fasta, 7798 sequences) and a compiled database containing the UPS1 human sequences (48 sequences).

#### 2.3.1. Workflow 1: ExtractMSn/Mascot/MFPaQ/Spectral counting

The Mascot Daemon software (version 2.4; Matrix Science, London, UK) was used to perform database searches, using the Extract_msn.exe macro provided with Xcalibur (version 2.0 SR2; Thermo Fisher Scientific) to generate peaklists. Parameters used for creation of the peaklists were: parent ions in the mass range 400–4500, no grouping of MS/MS scans, and threshold at 1000. Peaklists were submitted to Mascot database searches (version 2.4.2). ESI-TRAP was chosen as the instrument, trypsin/P as the enzyme and 2 missed cleavages were allowed. Precursor and fragment mass error tolerances were set at 5 ppm and 0.8 Da, respectively. Peptide variable modifications allowed during the search were: acetyl (Protein N-ter), oxidation (M), whereas carbamidomethyl (C) was set as fixed modification. To calculate the false discovery rate (FDR), the search was performed using the "decoy" option in Mascot. Validation was performed with an in-house developed module associated to MFPaQ [17] (http://mfpaq.sourceforge.net/), based on the target-decoy strategy, as described before [18]. Briefly, FDR at peptide level was calculated as described in [19] and set at 5% by adjusting peptide p-value threshold. Validated peptides were assembled into protein groups following the principle of parsimony (Occam's razor) [20]. Protein groups were then validated to obtain a FDR of 1% at the protein level, by adjusting the threshold on a protein group score defined as the sum of peptide score offsets (difference between each peptide Mascot score and its homology or identity threshold). The total spectral count metric was extracted for each protein group by MFPaQ in each analytical run.

#### 2.3.2. Workflow 2: Andromeda/MaxQuant/Spectral counting

Acquired MS data were processed using MaxQuant version 1.3.0.5 [21]. Derived peak lists were submitted to the Andromeda search engine [22]) (www.maxquant.org). For database searches, the precursor mass tolerance was set to 20 ppm for first searches and 6 ppm for main Andromeda database searches. The fragment ion mass tolerance was set to 0.5 Da. Trypsin/P was chosen as the enzyme and 2 missed cleavages were allowed. Oxidation of methionine and protein N-terminal acetylation were defined as variable modifications, and carbamidomethylation of cysteine was defined as a fixed modification. Minimum peptide length was set to six amino acids. Minimum number of unique peptides was set to one. Maximum FDR – calculated by employing a reverse database strategy – was set to 1% for peptides and proteins. Proteins identified as "reverse" and "only identified by site" were discarded from the list of identified proteins. In this particular workflow, total spectral count for each validated protein group was computed from msms.txt table.

#### 2.3.3. Workflow 3: Mascot Distiller/Mascot/IRMa-hEIDI/Spectral counting

Data were processed automatically using Mascot Distiller software (version 2.4.3.0, Matrix Science). ESI-TRAP was chosen as the instrument, trypsin/P as the enzyme and 2 missed cleavages were allowed. Precursor and fragment mass error tolerances were set at 5 ppm and 0.8 Da, respectively. Peptide variable modifications allowed during the search were: acetylation (Protein N-ter), oxidation (M), whereas carbamidomethyl (C) was set as fixed modification. The IRMa software v1.31 [23] was used to filter the results. Filters used were: (1) peptides whose score ≥ query homology threshold ($p < 0.5$) and rank ≤1 are marked as significant; (2) single match per query filter was: Move to ambiguous all peptides which aren't assigned to best protein for this query (best is higher protein score); (3) FDR seeker filter : Seek a 1% FDR based on score filtering; (4) Accession filter : Delete proteins coming from reverse database; (5) Specific peptide filter : Accept only protein hits whose specific peptides count ≥1. The filtered results were then compiled and structured within dedicated relational Databases and a homemade tool (hEIDI) was used for the compilation, grouping and comparison of the proteins from the different samples, analytical replicates and conditions to compare (Hesse et al., in preparation). In such workflow, total spectral count values calculated for each protein groups are used for quantification.

#### 2.3.4. Workflow 4: ExtractMSn/Mascot/Scaffold/Spectral counting

Peaklists generation and protein identifications were made as detailed in workflow 1. Mascot results were loaded into the Scaffold software (Version 3.6.5, Proteome Software, Portland, USA). To minimize false positive identifications, results were subjected to very stringent filtering criteria as follows. For the identification of proteins, a Mascot ion score had to be minimum 30 and above the 95% Mascot significance threshold ("Identity score"). The target-decoy database search allowed us to control and estimate the false positive identification rate of our study, and the final catalog of proteins presented an

**Table 1**
LC–MS quantification workflows evaluated. Combinations of tools were used for peaklist creation, database search, validation and quantification, resulting in 8 different workflows based on either spectral counting or MS signal extraction procedures, as described in details in the Experimental procedures section. The software tools used for spectral count quantification were Scaffold, IRMa/hEIDI, MaxQuant and MFPaQ. In the case of MS intensity-based quantification, protein intensity metrics were obtained from MFPaQ, MaxQuant or Skyline.

| Workflow number | Peaklist creation device | Database search engine | Validation of identified proteins/ spectral counting device | MS signal extraction device | Quantification method |
|---|---|---|---|---|---|
| 1 | ExtractMSn | Mascot | MFPaQ | | Spectral counting |
| 2 | Andromeda | Andromeda | MaxQuant | | Spectral counting |
| 3 | Mascot Distiller | Mascot | IRMa/hEIDI | | Spectral counting |
| 4 | ExtractMSn | Mascot | Scaffold | | Spectral counting |
| 5 | ExtractMSn | Mascot | MFPaQ | MFPaQ | MS signal analysis |
| 6 | Andromeda | Andromeda | MaxQuant | MaxQuant (Intensity) | MS signal analysis |
| 7 | Andromeda | Andromeda | MaxQuant | MaxQuant (LFQ) | MS signal analysis |
| 8 | Mascot Distiller | Mascot | Scaffold | Skyline | MS signal analysis |

estimated false discovery rate (FDR) below 5%. The spectral count metric used for quantitation corresponds to the Unweighted Spectrum Count values in Scaffold.

*2.3.5. Workflow 5: ExtractMSn/Mascot/MFPaQ/MS Signal analysis*

The first steps (peaklist creation, database search, validation) were the same than in workflow 1. Quantification of proteins was then performed using the label-free module implemented in the MFPaQ v4.0.0 software, as previously described [18,24]. Briefly, the software uses the validated identification results and retrieves the XIC of the identified peptide ions in the corresponding raw nanoLC–MS files, based on their experimentally measured RT and monoisotopic *m/z* values. Peptide ions identified in all the samples to be compared are used to build a retention time matrix and re-align in time LC–MS runs. For peptides not identified by MS/MS in a particular run, this re-alignment matrix is used to perform cross-assignment and extract their XIC signal starting from a predicted RT. Normalization across conditions is performed based on the median of XIC area ratios for all the extracted peptide ions. Protein quantification is based on a protein abundance index calculated as the average of XIC area values for at most three intense reference tryptic peptides per protein.

*2.3.6. Workflow 6 and 7: Andromeda/MaxQuant/MS Signal analysis*

The first steps (database search with Andromeda and validation) were the same as in workflow 2. For quantification purposes, either Intensities (workflow 6) or LFQ [25] (workflow 7) calculated by MaxQuant were used. The LFQ metric, as described in [25], is derived from the raw intensities by the MaxLFQ algorithm, which uses a specific normalization procedure, as well as a particular aggregation method to calculate protein intensities, by taking into account, for each protein, all the peptide ratios measured in all pair-wise comparisons of the different quantified samples. "Match between run" time window was set to 2 min. For LFQ quantification, only protein ratios calculated from at least two unique peptides ratios (min LFQ ratio count = 2) were considered for calculation of the LFQ protein intensity.

*2.3.7. Workflow 8: Mascot Distiller/Mascot/Skyline/MS Signal analysis*

Peaklist creation was performed with Mascot Distiller as described in workflow 3, then database searches were performed with Mascot and validated with Scaffold as described for workflow 4. XIC signal corresponding to all validated peptides were extracted using the Skyline software [26] (Skyline version v2.5, daily updates of April 2014, https://skyline.gs.washington.edu). This method was well described by Schilling et al. (Schilling et al., MCP, 2012). Total areas, corresponding to the sum of the 3 extracted isotopes areas, were used for statistical analysis.

*2.4. Statistical analysis*

For pairwise comparisons of samples spiked at different concentrations of UPS1, same statistical tests and fold-change criteria were applied to the quantitative data obtained from each workflow, as follows:

When working on spectral count metrics (workflows 1–2–3–4), a beta-binomial test was performed based on triplicate MS/MS analyses. p-values were calculated with the software package BetaBinomial_1.2 [27] implemented in R. Fold change was calculated as ratio of average spectral counts from both conditions. For proteins absent in all replicates of one specific condition, their spectral count values were modified by adding 1 spectrum to all 6 samples in order to be able to calculate a fold change for these particular proteins. To classify proteins as variant and non-variant and plot ROC curves, different combinations of criteria were tested ($|\log_2$ fold change$| > x$, from 0.8 to 3; p-value $< y$, from 0.05 to 0.0001).

When working on MS signal intensity-based metrics (workflows 5–6–7–8), proteins were filtered out if they were not quantified in at least all replicates from one condition. Missing protein intensity values were replaced by a constant value calculated independently for each sample as the 5-percentile value of the total population. A Welch t-test (two-tailed t-test, unequal variances) based on triplicate MS analyses was then performed on $\log_2$ transformed values using the Perseus toolbox (version 1.4.0.11; http://141.61.102.17/perseus_doku). Criteria used to classify the proteins were the Welch t-test difference calculated by Perseus (difference between the two compared conditions of the mean $\log_2$ transformed value for triplicate MS/MS analyses), and the Welch t-test p-value. Results were filtered using different combinations of these criteria: |Welch t-test difference| > x (from 0 to 7) and p-value < y (from 0.3 to 0.0001). z-score was also calculated as z-score = {(Welch t-test difference) − Median [(Welch t-test difference) for all quantified proteins]} / Standard deviation [(Welch t-test difference) for all quantified proteins].

## 3. Results

*3.1. Experimental design, sample preparation and analysis*

In order to evaluate different quantitative workflows in their ability to correctly detect known variant proteins in complex samples, we prepared a series of 9 yeast lysate samples spiked with growing concentrations of the Sigma UPS1 standard composed of an equimolar mixture of 48 human proteins. To that aim, UPS1 lyophilized proteins were directly resuspended using the yeast lysate prepared in urea buffer, and a serial dilution of this initial mixture was then performed using the same yeast lysate, resulting in spiked UPS1 concentrations ranging from 50 amol/μg up to 50 fmol/μg of yeast lysate. Protein samples were digested with trypsin, and resulting peptides were analyzed by nanoLC–MS/MS on a LTQ Velos-Orbitrap instrument, using routine chromatographic conditions (15 cm C18 reverse-phase column, 2 h gradient) and data-dependent acquisition MS parameters (resolution 60,000 for MS survey scan, top 20 CID sequencing in the ion trap). Triplicate MS analysis was performed for each sample, resulting in 27 raw data files that were subsequently processed in different ways, using several computational workflows (Table 1). Two different software were used for protein identification (Mascot and Andromeda) and 5 solutions were employed for protein quantification (Scaffold, IRMa/hEIDI (Hesse et al., in preparation), MaxQuant [21,22,28], MFPaQ [17,24] and Skyline [26,29]), some of them generating a unique quantitative output, either spectral counting or MS signal extraction data, and some of them generating both types of quantitative data. Finally, 8 different quantitative datasets were obtained, as indicated in Table 1 and described in details in the Experimental procedures section.

We first evaluated the identification datasets in a qualitative way by simply reporting the number of identified and validated proteins for both the background (yeast proteins) and the spiked standard (UPS1 proteins) in each sample. Sup data 1 shows the number of proteins identified by MS/MS sequencing and validated by various bioinformatics workflows. As expected, the total number of proteins, reflecting mainly the constant yeast background proteome, was fairly reproducible across triplicate MS analysis and across the series of spiked samples, whereas the number of identified UPS1 proteins increased with the spiked amount. While no UPS1 protein was correctly identified at a concentration of 500 amol/μg (as none of the peptide sequence matches could be validated at this concentration), all 48 human proteins were sequenced and correctly identified at 50 fmol/μg. From these results, we decided to select different concentration levels of UPS1 to perform pairwise quantitative comparisons of samples, trying to mimic distinct biochemical situations, as illustrated in Fig. 1. Comparison A (500 amol/μg versus 50 fmol/μg) should reflect a case were a protein is typically under the detection level of the instrument in one condition, and strongly expressed in the other condition with a fold change of 100. In comparison B (5 fmol/μg versus 50 fmol/μg), the protein may be in turn detectable in both conditions, and strongly up-regulated with a fold change of 10.

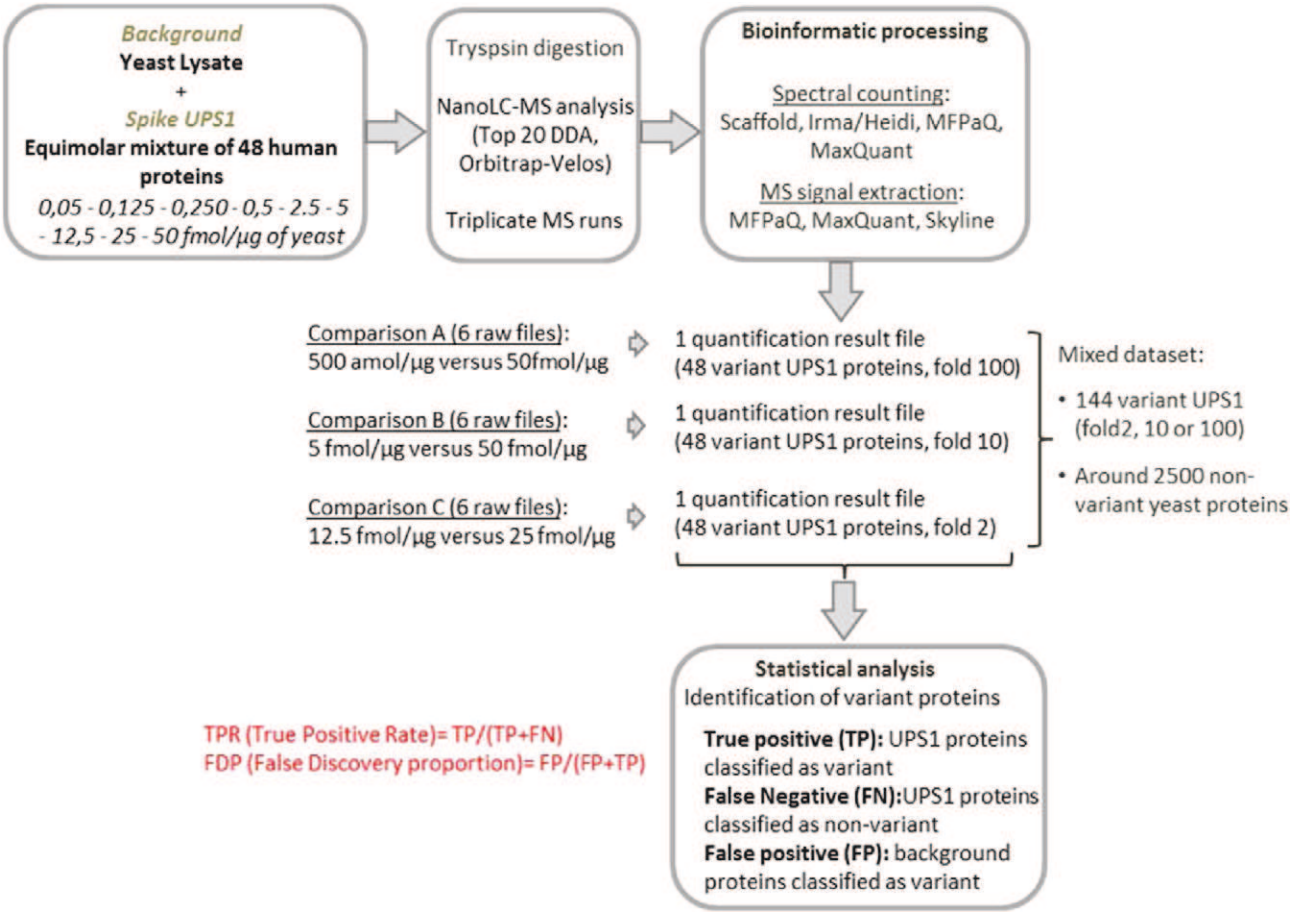



**Fig. 1.** Experimental design. A series of 9 yeast lysate samples spiked with growing concentrations of the Sigma UPS1 standard was analyzed in triplicate by nanoLC–MS/MS mass spectrometry on a LTQ Velos-Orbitrap instrument. Different computational workflows were used to identify, validate, and quantify proteins based on spectral counting or MS signal analysis. In the present study, 3 different pairwise quantitative comparisons (A, B, and C) were performed between samples spiked with different amounts of UPS1, involving in each case the quantification of 6 raw files (2 conditions × 3 replicates), trying to mimic distinct biochemical situations. The 3 individual quantitative datasets containing protein abundance values were then gathered. This global quantitative dataset was generated for each data processing workflow, and identical downstream statistical processing methods were then applied for classification of variant proteins.

Finally, comparison C (12.5 fmol/µg versus 25 fmol/µg) should simulate a situation where the protein is detectable in both conditions, but only slightly up-regulated with a fold change of 2. Because "real-life" biological samples usually contain many proteins with a differential abundance, encompassing a wide range of absolute expression levels and fold change values, we tried to approximate such a situation by gathering together the quantitative data obtained for each binary comparison, after computational processing. Using this post-processing assembly of the 3 individual datasets, we composed a global quantitative dataset containing theoretically 144 variant proteins (UPS1 proteins issued from the 3 relative quantitative analyses, and thus expected to vary with a fold change of 100, 10 or 2), and a background of around 2500 non-variant yeast proteins (measured and quantified in the different pairwise comparisons) (see ref [30], Sup Table 1). The generation of this synthetic dataset allowed us to illustrate, in a single representation, the performance of quantitative proteomic tools and methods, challenged with different situations.

The final aim of relative quantitative analysis in discovery proteomics is usually to identify differentially expressed proteins. Therefore, the tested informatics workflows were mainly evaluated in their ability to correctly detect the expected variants, rather than in the accuracy of the measured fold change. The experimental design and the spiked standard used here allowed us to unambiguously assess such performances by counting the number of true-positives (TP) and false-positives (FP), respectively UPS1 or yeast background proteins found to be differentially expressed. Clearly, the classification of proteins as variant (positive hits) or non-variant (negative hits) both relies on the one hand, on the accuracy of the quantitative metrics generated by the bioinformatics software, and on the other hand, on the performance of the statistical test and criteria used to discriminate the positive and negative populations. Here, we mainly tried to benchmark the former step of the workflow (extraction of quantitative metrics by informatics tools), and we didn't aim to evaluate statistical methods. We thus used a common, simple statistical test for protein classification, based either on the beta-binomial method for spectral count datasets [27], or on a modified t-test for datasets containing peptide intensity-based values (see Experimental procedures section and below). Proteins were classified as variant or non-variant by a combined filtering on the p-value of this statistical test and on the fold change value, as very often performed in "real life" biological studies [31–34]. Following such classification, the sensitivity of the workflows for the detection of variant proteins (number of true positive hits relative to the real total number of variant proteins, i.e. TP/144), and false discovery proportion (FDP, defined as the number of false positive hits relative to the total number of proteins found as variant, i.e. FP/(TP + FP)) could easily be computed.

### 3.2. Performances of spectral counting for discrimination of variant proteins

Fig. 2A shows the volcano plots obtained by applying spectral counting quantification methods, in which the $\log_{10}$(p-value) (calculated from the results of the BetaBinomial R package) is plotted against the calculated protein $\log_2$(fold change). As illustrated on these graphs,

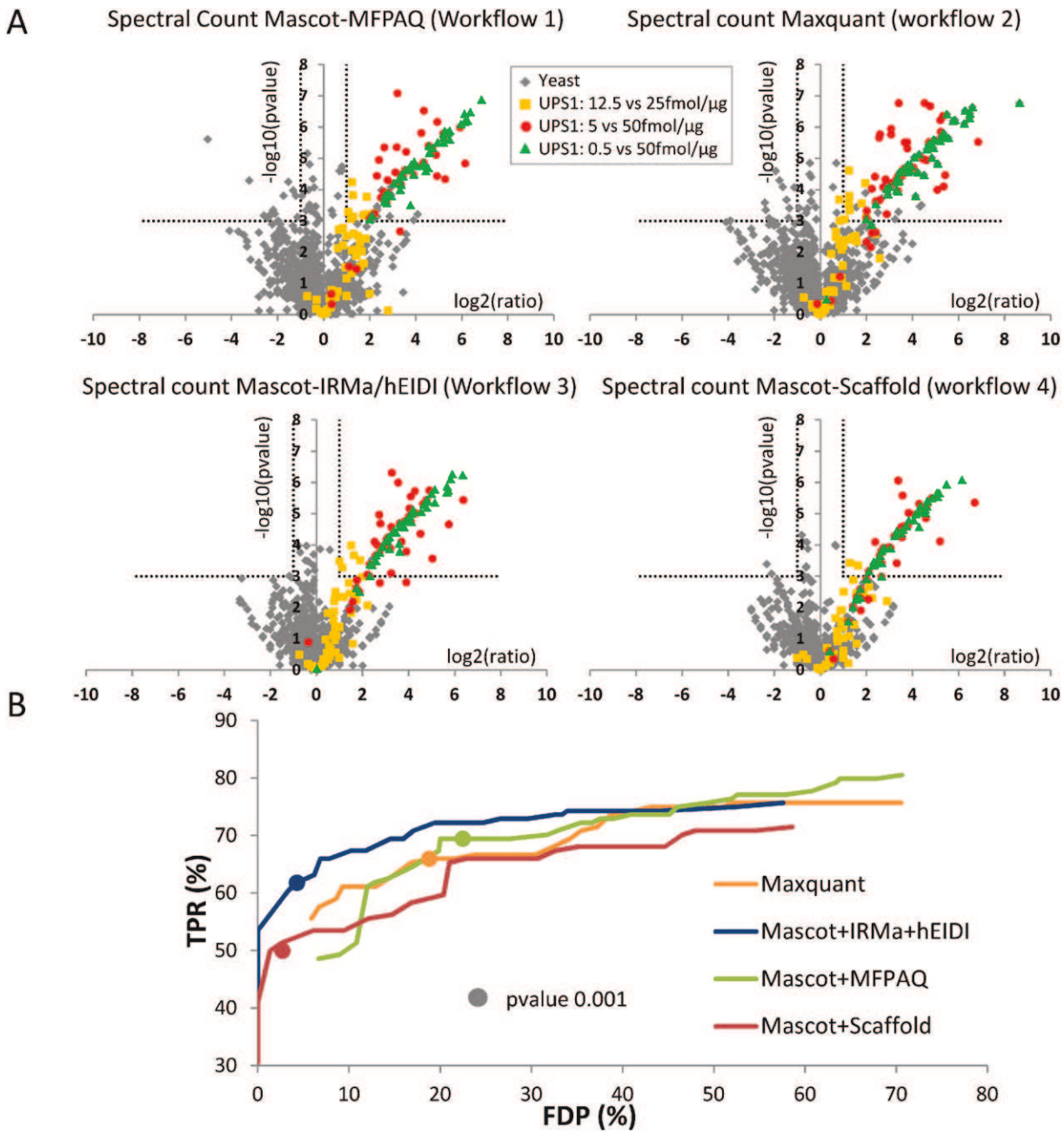



**Fig. 2.** Quantitative results obtained with spectral counting workflows. **A.** Volcano plots ($-\log_{10}$(p-value) of the beta-binomial test versus protein $\log_2$(fold change)) are shown for the different software tools tested. The graphs illustrate the quantitative results for the UPS1 proteins quantified in each binary comparison (green: comparison A, 0,5 fmol/µg versus 50 fmol/µg, theoretical fold change 100; red: comparison B, 5 fmol/µg versus 50 fmol/µg, theoretical fold change 10; yellow: comparison C, 12.5 fmol/µg versus 25 fmol/µg, theoretical fold change 2). Gray dots correspond to yeast proteins quantified in all of these comparisons. Dotted lines represent a fixed p-value threshold of 0001 and a fixed $|\log_2$(fold change)| threshold of 1. B. For each spectral count workflow, proteins of the mixed dataset (comparison A + B + C) were classified as variant after application of different p-value thresholds combined to a fixed $\log_2$(fold change) threshold of 1. The number of true positives (TP) and false positives (FP) was retrieved, and true positive rate (TPR or sensitivity = TP/144) was plotted as a function of false-discovery proportion (FDP = FP / (TP + FP)).

the majority of UPS1 proteins from comparisons A and B (green and red populations, theoretical fold changes of respectively 100 and 10) were easily discriminated from the background of yeast proteins (gray), by both their p-values and fold changes. This was particularly the case with software tools such as IRMa/hEIDI and Scaffold. These results indicated the ability of the spectral count-based quantitative approaches to confidently detect protein variations of high to medium amplitude while minimizing the level of false discoveries. However, it can be noted that the UPS1 proteins quantified in the comparison C (12.5 fmol/µg versus 25 fmol/µg, yellow dots) were not well segregated from the background independently of the software used. Overall, these observations pointed out some limitations of quantification with spectral count data when dealing with low fold change variations or weakly concentrated proteins.

From these data, we tried to determine which criterion was best suited to retrieve significantly variant proteins. Sensitivity-FDP curves were plotted for the data obtained from the different workflows by using either the fold change or the p-value as a unique criterion to classify the proteins, and we further wanted to apply combinations of these filters to improve the classification. Resulting curves (Sup data 2A) show that the beta-binomial test was per se more efficient than a simple fold change to discriminate the TP from the TN. However,

applying an additional fixed fold change cutoff improved significantly the results, as could be anticipated already from the volcano plots. On the dataset presented here, the best classification was obtained for all the workflows by applying this double-filtering approach with a threshold of 2 (or 1/2) on the fold change. Therefore sensitivity-FDP curves were plotted this way (variation of the p-value combined with a fixed threshold of 1 on the absolute $\log_2$(fold change)) for the different spectral count workflows as shown in Fig. 2B. Globally, the best results were obtained with workflow 3 (Mascot/IRMa-hEIDI) which allowed for example to obtain a reasonable sensitivity (62%) with a very low FDP (4%) when setting a stringent p-value threshold of 0.001. Leveraging the p-value threshold at 0.0025 led to a slightly better sensitivity (67%) at the cost of a FDP increase to about 10%. Interestingly, in the case of workflows 3 (Mascot/IRMa-hEIDI) and 4 (Mascot/Scaffold), it was possible to reach really low FDP values by increasing the stringency on the p-value, showing the efficiency of these data processing tools for the exclusion of FP. Altogether, it turns out that spectral count approaches were very efficient for detecting high levels of variations on relatively abundant proteins, but tends to fail to reach high sensitivity on the present dataset which includes a population with moderate fold change variations. Markedly, very low levels of FDP can be reached with appropriate filtering.

### 3.3. Performances of MS intensity-based methods

Fig. 3A shows the volcano plots from data obtained using different MS feature extraction tools ($-\log_{10}$(p-value) – calculated with the two-samples Welch t-test from Perseus – plotted against the $\log_2$(fold

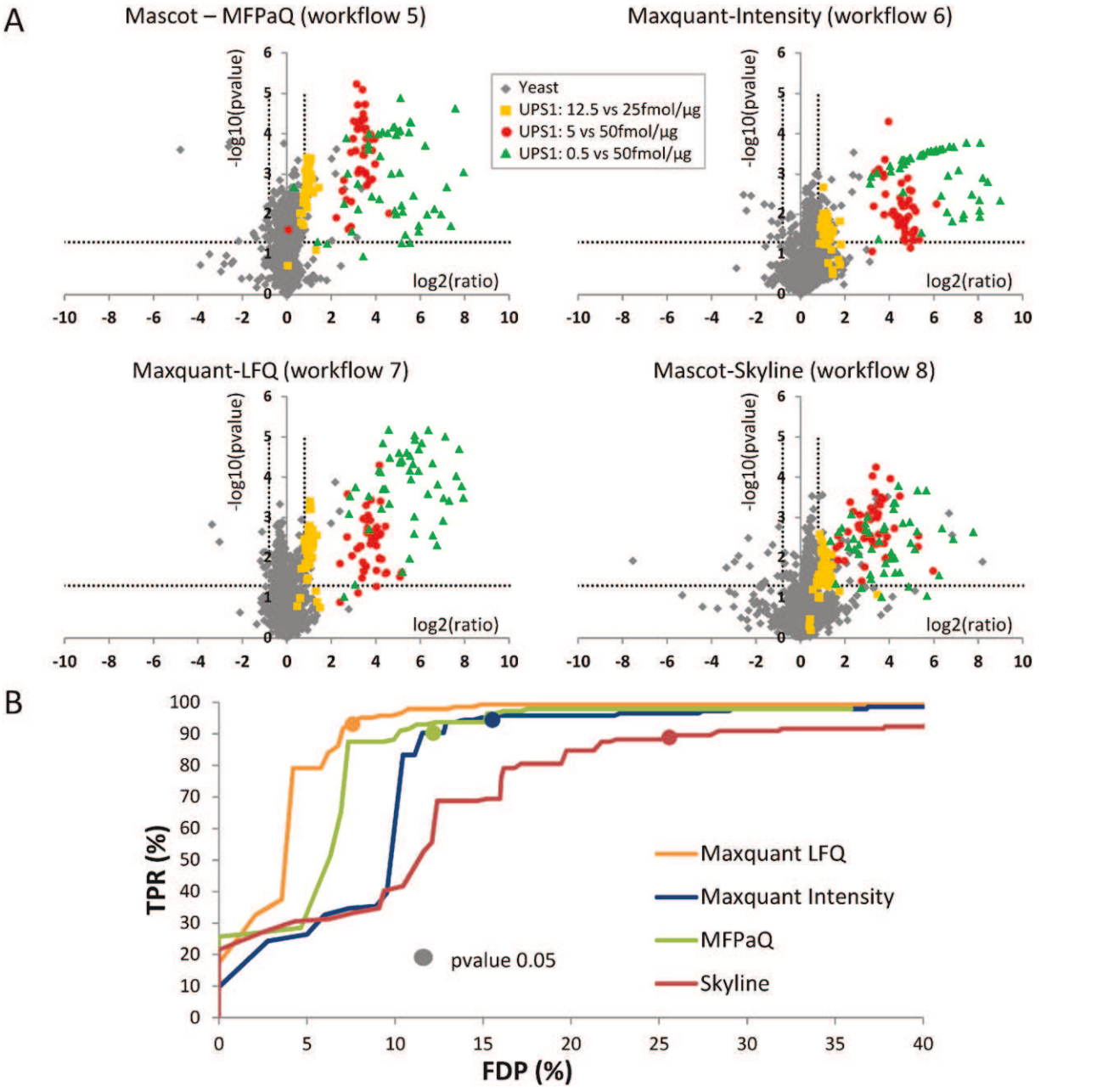



**Fig. 3.** Quantitative results obtained with MS feature extraction workflows **A.** Volcano plots ($-\log_{10}$(p-value) of the Welch t-test versus protein Welch t-test difference) are shown for the different software tools tested. As in Fig. 2, the graphs illustrate the quantitative results for the UPS1 proteins quantified in the different binary comparisons A, B and C. Gray dots correspond to yeast proteins quantified in all of these comparisons. Dotted lines represent a fixed p-value threshold of 0.05 and a fixed |Welch t-test difference| threshold of 1. B. For each MS signal analysis workflow, proteins of the mixed dataset (comparison A + B + C) were classified as variant after application of different p-value thresholds combined to a fixed |z-score| threshold of 1. TPR (sensitivity) = TP/144) was plotted as a function of false-discovery proportion (FDP = FP / (TP + FP)).

change)). Conversely to what we observed with spectral-counting, the plots obtained with MS intensity-based techniques show that a large majority of UPS1 proteins quantified in the different pair-wise comparisons (green, red, and yellow populations) can be visually discriminated from the background of yeast proteins. While proteins with high signal levels and high theoretical fold changes were most often easily classified as variant (good p-values and high calculated fold changes), it can be noticed that even the UPS1 proteins quantified in the comparison C can be segregated from background, although with a partial overlap.

Here again we plotted different sensitivity-FDP curves by classifying the proteins either on their absolute fold change, on their Welch t-test p-value, or by a combination of these criteria (setting up a fixed threshold for one of them and varying the other) (Sup data 2B). In the case of MS intensity values obtained in our dataset, the fold change appeared to be generally a more efficient filter to discriminate TP from background than a simple statistical test based on the variance of the protein intensities. Indeed, the modified Welch t-test may produce a high number of FP hits on this particular dataset containing only three analytical replicates, finally leading to a high FDP after multiple testing. For example, on the MaxQuant LFQ dataset (workflow 7), filtering the proteins at a 0.05 cutoff only on the Welch p-value allowed to efficiently retrieve almost all UPS1 variant proteins (134 out of 144, e.g. 94% sensitivity), but with as many as 387 FP yeast proteins declared as variant (i.e. a final FDP of 74%). On the other hand, correction of the p-values for multiple-testing with methods such as the Benjamini–Hochberg (BH) procedure can be used to limit the number of FP and control the final FDR, but at the cost of a much lower sensitivity. For example, applying this correction on the same dataset and filtering afterwards with a BH adjusted p-value cutoff of 0.05 led to only 3 FP yeast proteins, but the number of TP UPS1 proteins also dropped to 50 (i.e. a calculated final FDP of 6%, close to the desired theoretical value, but a sensitivity of only 35%, see ref [30], Sup Table 1). Finally, combining fold change and Welch t-test p-value criteria emerged as the most discriminant approach, and allowed to reach good sensitivity with relatively low FDP. It has to be noticed that, unlike with the statistical t-test, setting a fold change threshold was quite sensitive to any shift in the population fold change distribution and to the optional normalization procedure applied in the workflows. Since some of the used methods contained a normalization step (e.g. MFPaQ or MaxQuant with the LFQ metric) and others not (e.g. Skyline or MaxQuant based on summed peptide intensity values), we used a z-score to avoid possible discrepancies between quantitative data depending on their origin. This z-score reflects, for each protein, the distance between the protein fold change and the mean of the population fold changes, relative to the standard deviation of this population (see Experimental procedures for calculation of the z-score). The combination of z-score and p-value criteria gave efficient discrimination results, as shown in Sup data 2B. For example, in the case of the MaxQuant LFQ workflow, we obtained a sensitivity of 94% and a calculated FDP of 8% when combining a |z-score| threshold of 1 and a Welch t-test p-value threshold of 0.05.

Fig. 3B shows the sensitivity-FDP curves obtained for the MS intensity based workflows by varying the Welch t-test p-value filter, with a fixed |z-score| cut-off of 1. Altogether, it appeared that the tested label-free tools based on MS signal analysis have the potential to be globally very sensitive (detect a large proportion of the true variant UPS1 proteins), with sensitivity values up to 94% when setting a p-value of 0.05. Comparative results for the different software are shown in Table 2 with sensitivity and FDP for this specific p-value. It has to be noticed however that all workflows produced still relatively elevated FDP values, that may be related to signal extraction errors by the software. The best compromise between sensitivity and FDP was obtained using the LFQ metric from MaxQuant [25] and the Top 3 metric from MFPaQ [24].

### *3.4. Use of the spiked standard dataset to highlight data processing problems and optimize the workflows.*

We next wanted to take advantage of this model dataset to identify quantification errors associated to the generation of false-negative (FN) and false-positive (FP) proteins, and illustrate a number of possible mistakes introduced by the different MS intensity based workflows. Protein quantification is a multi-step process, and possible errors associated to each of these steps may influence the final result. Obviously, processing steps based on peptide validation, grouping, and peptide-to-protein inference are important for final protein quantification. Sup data 3A illustrates a case where quantification based on non-specific peptides, shared between a stable yeast protein and a UPS1 variant protein (Ubiquitin-40S ribosomal protein S27a), compromised the result and led to classification of the spiked protein as a FN. Most of the time however, errors seem to take place at the signal extraction step itself. Sup data 3B shows a situation with overlapping isotopic patterns from several coeluting species, in which the MFPaQ software wrongly picked, in addition to the monoisotopic peak of the correct peptide, the third and second isotope peaks from other species, as well as the monoisotopic peak of a closely eluting isobaric peptide. Such errors could be avoided through a better recognition by the algorithms of peptide isotopic patterns. In addition, in the cases illustrated here, 16 peptides were correctly quantified for the protein, while signal extraction error occurred occasionally on a single peptide. Enabling the detection and elimination of outlier peptides with adequate testing procedures (option not enabled in that case) would alleviate such problems. Good alignment of LC–MS runs in retention time is also important for correct peak picking when cross assignment between runs is implemented. Some errors in Skyline could be attributed to wrong selection of a particular peptide in one of the runs in which the peptide was not sequenced by MS/MS, and in which XIC extraction was thus performed based on the RT of the peptide in another run (not shown). It must be noticed that tracking and eventually correcting these signal extraction errors is quite dependent on the software interface. To this respect, a software like MFPaQ offers a visualization interface that enables a rapid inspection of the XICs extracted for each peptide in the different conditions, and possibly unselects some of them to eliminate these peptides from the final quantification of the protein. However, it does not allow going back to raw MS data and correct for example the selection of the integration area directly on the chromatogram. This in turn is possible in Skyline, which really offers an interactive interface to efficiently review the results and manually correct possible mistakes. We thus wanted to take advantage of this feature and evaluate whether

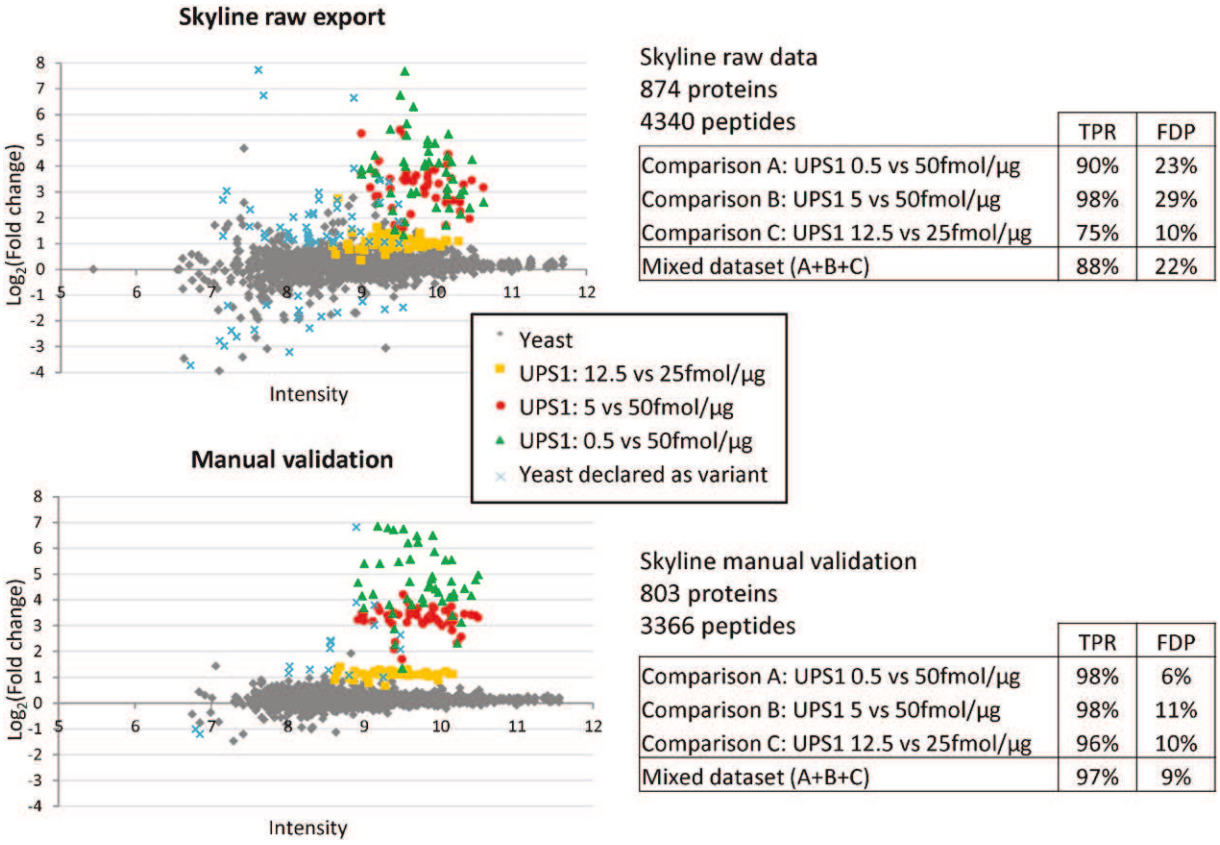
**Table 2**
FDP and TPR obtained on the spiked dataset for different quantitative workflows. Similar criteria were used for all workflows to classify proteins as variant (positive hits), i.e. |z-score| > 1 and Welch t-test p-value < 0.05. Human UPS1 proteins and yeast proteins verifying these criteria were counted respectively as True Positive and False Positive. False Discovery Proportion and True Positive Rate (sensitivity) were computed as described in the table.

| | MFPaQ (workflow 5) | Maxquant intensity (workflow 6) | Maxquant LFQ (workflow 7) | Skyline (workflow 8) |
|---|---|---|---|---|
| True positive | 135 | 130 | 134 | 126 |
| False positive | 25 | 18 | 11 | 36 |
| FDP = FP / (FP + TP) * 100 | 16% | 12% | 8% | 22% |
| TPR = TP / (TP + FN) * 100 | 94% | 90% | 93% | 88% |

manual validation of the entire dataset was practically possible and how efficient it could be to improve the quantitative results. It took around 15 h to manually check all the peptide ions from the dataset and either validate or correct the integration of the corresponding XIC. Fig. 4 shows the result of this exhaustive reviewing of the data on the accuracy of the quantitative result. While relatively time consuming, the manual correction clearly reduces the number of both false positive and false negative. The sensitivity was thus improved (from 88% using raw data to 97% after manual correction) and the FDP was significantly reduced (from 22% to around 9%) (sensitivity and FDP values calculated by filtering proteins based on a Welch t-test p-value < 0.05 calculated with the two-samples test from Perseus and |z-score| > 1). In addition, the calculated fold changes were closer to the expected theoretical values. It appeared that most of the extraction errors generating false positive hits were related to low intensity signals, as illustrated in Fig. 4. Finally, after manual correction, no more than 8 yeast proteins were classified as variant. Out of these 8 false positive hits, 3 contained peptides that were clearly "contaminated" with UPS1 peptides, 4 had very low intensity signals, and one of them was detected as variant while the expression profile of the related peptides did not follow that of UPS1 peptides. Altogether, these residual mistakes remaining after in-depth manual validation may reflect the minimal margin of error of the label-free, MS intensity-based quantification process, which may be difficult to reduce even by improving the automatic signal extraction algorithms of the software.

## 4. Discussion

In this study, we generated a complex, spiked proteomic standard dataset, in which the ground truth is well characterized, and showed its utility for benchmarking label-free relative quantification computational workflows. Different protein standards have been used in the past to measure the performances of such software and data processing methods, ranging from simple mixtures of recombinant proteins, to complex cellular extracts spiked with a known amount of exogenous proteins. In the design of such a standard, it is important to be able to easily differentiate the spiked proteins from the background after the database search and identification process, in order to perform a correct classification of spiked (TP) and background (TN) molecules. The most straightforward approaches are either to apply some isotopic labeling on the background or the spiked samples, or to use sets of proteins from different species. Ideally, the number of spiked molecules should be large enough to provide a relevant statistical estimation of the sensitivity and FDP of the quantitative methods. Typically, samples can be spiked with recombinant purified proteins added in known quantities to the background, or with a much more complex sample, such as a biological extract from another species. In recent studies aiming at benchmarking software tools, such "double-proteome" samples have been used. For example, a mixture of lysates from human cells and from the *Streptococcus pyrogenes* bacterium at different ratios was used in a comparative study to show the performances of the OpenMS




| | TPR | FDP |
|---|---|---|
| Comparison A: UPS1 0.5 vs 50fmol/µg | 90% | 23% |
| Comparison B: UPS1 5 vs 50fmol/µg | 98% | 29% |
| Comparison C: UPS1 12.5 vs 25fmol/µg | 75% | 10% |
| Mixed dataset (A+B+C) | 88% | 22% |

| | TPR | FDP |
|---|---|---|
| Comparison A: UPS1 0.5 vs 50fmol/µg | 98% | 6% |
| Comparison B: UPS1 5 vs 50fmol/µg | 98% | 11% |
| Comparison C: UPS1 12.5 vs 25fmol/µg | 96% | 10% |
| Mixed dataset (A+B+C) | 97% | 9% |

**Fig. 4.** Manual feature-extraction correction in Skyline. The graphs illustrate the $\log_2$(fold change) calculated from protein intensity values in each binary comparison (A, B, and C) as a function of protein intensity. Protein intensity values were calculated as the sum of all peptide area values extracted by Skyline for each protein, and fold changes were computed from the mean of triplicate protein intensity values for each spiked concentration point. Results were plotted either from the raw Skyline output, or after an extensive manual check of all the peptide ions from the dataset (leading to either validate or correct the integration of the corresponding XIC, or eliminate the peptide from quantification). UPS1 proteins quantified in each binary comparison are represented as indicated in the legend, and yeast proteins are represented either as gray dots (non-variant, true negatives) or blue crosses (variant, false-positives). Tables on the right indicate the number of proteins and peptides actually quantified in each case. Proteins were classified as variant after application of a p-value thresholds of 0,05 combined to a fixed $\log_2$(fold change) threshold of 1. TPR (TP/144) and FDP(FP / (TP + FP)) are indicated after classification of the proteins individually for each binary comparison (A, B or C), or on the mixed dataset (comparison A + B + C).

software [35]. Similarly, Cox et al. used a complex digest of Hela cells, spiked with an *Escherichia coli* digested cellular extract at two different amounts, creating a 3 fold variation of the *E. coli* proteins in the quantitative comparison [36]. In that later case, the spiked population represents a significant portion of the total sample (about one third of the identified proteins). Such a dataset may simulate particular biological experiments where a stimulation could for example induce a massive variation of the proteome, or some interaction proteomics experiments where a control is compared to an affinity purified sample containing many up-regulated proteins. However, normalization of such datasets may be difficult, because the usual hypothesis underlying normalization procedures is that the major part of the protein population remains stable, and the median of the fold change distribution should be 1. On the other hand, spiking a proteome background with a calibrated set of recombinant purified proteins is statistically less representative, as the number of TP decreases, but allows to simulate easily a classical expression proteomics experiment, in which a very minor part of the proteome will undergo a fold change. The UPS1 commercial standard, containing an equimolar mixture of 48 purified human proteins, represents a convenient sample for a spiking scheme experiment, and offers already a significant number of TP that allows to get an estimation of the sensitivity and FDP of the computational methods.

As software tools are expected to perform unequally depending on the fold change and amount of the spiked proteins, producing signals that will be more or less difficult to extract from the raw data according to their intensities, it is important to challenge them with different simulated variations. In a previous study, Cox et al. spiked UPS1 in combination with the UPS2 standard, which contains the same proteins than UPS1, but distributed into 6 groups of decreasing concentration, spanning 5 orders of dynamic range [36]. By adding respectively these two standards into a background *E. coli* proteome, the authors simulated a situation where groups of proteins vary with different ratios, in a single pairwise comparison (6 analytical runs corresponding to 2 conditions with 3 technical replicates). However, in that case, only a small number of proteins are representative of each ratio, and many highly diluted UPS2 proteins are hardly detectable, creating a significant set of proteins which are differentially expressed but not really quantifiable.

In the present study, we chose to spike the UPS1 mixture at 9 different concentrations in a background yeast proteome, as described previously in Paulovich et al. [6], and analyzed these samples in triplicate, resulting in a dataset of 27 runs. In order to artificially recreate a simulated dataset containing TP with different intensities and fold change values, we performed several pairwise comparisons by label-free quantification, and then combined the quantitative outputs. This approach has the benefit to illustrate the performances of the computational and statistical methods in a more comprehensive way. As a proof of principle, we show here the results obtained by simulating 3 kinds of variations (comparisons A, B and C: detection in only one condition; high fold change; moderate fold change). In principle, more comparisons could be performed and gathered to better approximate the inherent complexity of the variations that take place in a real biological experiment. For example, we didn't challenge here the software tools with comparisons involving only the more diluted spikes of the UPS1 concentration range, which would simulate variations of lower abundance proteins. Nevertheless, the different UPS1 spikes considered here could represent different types of biological samples, notably affinity purifications for large fold change analyses, or more classical proteome-wide analyses including moderate but significant expression fold change for some regulated proteins.

While label-free methods are more and more used for quantification of complex protein mixtures in biological studies, they are sometimes still considered as less accurate and reliable than label-based approaches. In addition, while many software tools for label-free quantification have been developed and are available, it may be difficult for an unexperienced user to choose a particular workflow. Finally, the quality of the results may be influenced by the parameter settings and the user's expertise with the programs. Consequently, test datasets are really needed to assess the performances of a given label-free workflow, adjust the parameters of a particular algorithm, and optimize post-processing methods such as missing value imputation, normalization, and statistical tests. The dataset presented here offers such possibilities, as illustrated on 8 different label-free pipelines which were objectively evaluated, and for which the number of FN and FP could be easily measured. The results obtained here show that label-free approaches are indeed efficient to detect variant proteins on the standard dataset. Globally, compared to signal extraction procedures, spectral counting workflows exhibited limited sensitivity (see Sup data 4A, showing overlaid ROC curves for both type of approaches). Even with lenient p-value cutoff, spectral count methods could only reach sensitivity levels up to 70–80%, mainly due to inefficiency to classify low abundance proteins with moderate fold change (comparison C). However, it must be noticed that they are easier to implement (shorter data processing time), and work quite well to sort out proteins with medium to high fold change (comparisons A and B). Noticeably, they also proved to be quite specific, with the possibility to reach low level of FDP. Indeed, with data from such workflows, it was possible to set stringent filtering criteria and to almost completely avoid the detection of false positive yeast proteins, whereas this was much more difficult with MS intensity based methods (see below). Thus, as illustrated in Sup data 4B, at a given FDP level of e.g. 5%, spectral count approaches globally provided better sensitivity levels than MS intensity based approaches. In other words, if one is interested in the generation of a very “clean” and reduced list of differentially expressed proteins, the analysis of spectral count data with stringent filtering may represent a safe way to sort out very confident hits — probably with some compromise on sensitivity. Among spectral count workflows, coupling Mascot peptide identification with IRMa validation and hEIDI grouping and comparison ended up with the best compromise between sensitivity and FDP (Fig 2B). Indeed, even if retrieving the spectral count metric could per se be seen as a basic process which is not error-prone, depending on the workflow used, some differences in FDP were observed at the same sensitivity levels. In fact, spectral count approaches are still dependent on the quality of peptide validation, selection and grouping, which may directly influence the performances of the different software tools tested here.

On the other hand, our results indicated that workflows based on signal extraction clearly have the potential to be globally very sensitive, and are effective in detecting large variations as well as accurately measuring moderate fold changes. Sensitivity levels up to 90–100% could be reached by relaxing filtering criteria. Thus, when admitting FDP levels higher than 10%, such workflows outperformed spectral count methods for the classification of differentially expressed proteins in the dataset (Sup data 4B). They represent promising approaches to detect variations even on minor proteins expressed at low level in the sample, and/or showing subtle changes. However, it has to be noticed that at present, software tools based on MS intensity analysis still generate a significant number of FN and FP. The presence of false positive hits (type I error) associated to statistical tests in multiple comparisons is a well documented problem when using high-throughput analytical methods which enable the quantification of hundreds or thousands of species. When a large number of statistical tests are performed, the final proportion of false discoveries (FDP) is actually larger than the user-specified p-value cutoff used for each individual test. Multiple testing correction procedures are classically used to adjust the individual p-values of each gene or protein, and to control the final FDR, such as the Benjamini-Hochberg method. Interestingly, spiked datasets, such as the yeast-UPS1 dataset provided here, allow to experimentally measure this FDP rate as well as the associated sensitivity, and could represent a useful tool for optimization of statistical processing steps for proteomic data. The Benjamini–Hochberg adjustment, while very effective for controlling the final FDR of the process, appeared to be very conservative

and reduced strongly the sensitivity of the workflows. In our hands, empiric filtering based on the combination of p-value and fold change (or z-score) cutoffs offered a more efficient compromise to obtain good sensitivity with relatively low levels of experimentally measured FDP, although this FDP was not formally controlled through the statistical process. Clearly, further studies will be needed to implement statistical methods allowing to control the FDR rate when looking for differentially expressed proteins in proteomic experiments. For example, while we used here arbitrary, fixed fold-change and p-value cutoffs, other approaches have been described in which the fold-change cutoff can be modulated as a function of the t-test P-value, to increase sensitivity for a given FDR after Benjamini-Hochberg correction [37]. Additionally, pre-filtering can also be implemented to eliminate lowly abundant proteins which tend to give artificially high fold change values after spectral count quantification, and create false positives [37]. Finally, other statistical methods have bee proposed previously for microarray data in order to take into account a fold-change threshold of interest in a formal hypothesis test with FDR control [38–40].

The occurrence of FP and FN hits is also a problem that has to be tackled upstream of statistical processing, at the level of quantitative analysis and raw data processing tools, as these false hits are very often associated to signal extraction or matching problems (Sup data 3). Indeed, extraction of peptide intensity values is a complex process based on MS peak picking, isotope pattern and chromatographic peak recognition, and association of peptide features with MS/MS identification results, which can be complicated by the frequent occurrence of overlapping peptides in the LC–MS space. In our comparison, the MaxQuant software performed the best when using the LFQ metric (Fig. 3B). In MaxQuant, the data analysis starts from the detection of features in the LC–MS map, based on recognition of elution peaks and peptide isotope profiles. In contrast, the processing in MFPaQ and Skyline is based on direct XICs extraction, using as a starting point m/z and RT coordinates derived from MS/MS identification results. Our study indicates that the later approach can however also produce good results, as illustrated by the good sensitivity and FDP obtained from MFPaQ quantification. A higher number of false positive hits were obtained with Skyline, which could be attributed in most instances to the absence of realignment procedure in the version of Skyline used for this study, and incorrect retrieving of peptide signals at deviated RT in some of the conditions. On the other hand, the interactive interface of Skyline allowed to efficiently check the signal extraction, and enabled an in-depth manual verification which clearly improved the final quantitative results, and particularly allowed to reduce the number of false-positive. The reduction of false-positive is an important challenge in label-free based discovery proteomic approaches, as it will directly influence the success of further validation steps, based on the selection of protein candidates from the first quantitative analysis. Although manual validation of the whole population of peptide ions, as performed in this study, is certainly overly long and impracticable in "real-life" biological studies, the ability to go back to the raw data for manual inspection of some specific proteins is probably an important feature for a label-free quantitative software. Indeed, the user can in this way really check the evidence for the differential expression of a protein, directly on the XIC and MS spectra of the different peptides. This manual verification can be performed on specific proteins that make biological sense (e.g. on some expected markers which would not be found as variants, due to signal extraction errors by the software, but also on new candidate proteins that will be subsequently selected for further validation studies, to ensure that these are not false positive).

In summary, our study on the presented standard dataset indicates that 1/the number of false-positive hits from label-free quantitative analysis is still significant, even with the best performing workflows, 2/that manual verification by the expert allows to reduce it, illustrating that there is still some margin of improvement for the automatic signal extraction step by label-free software, and 3/that a residual number of errors remain inherently difficult to avoid, independently of the quality of the signal extraction procedure, particularly in the case of co-elution and overlapping peptide features, which would in turn require better resolution of both chromatographic and MS instruments. Ideally, label-free software should offer good performances in order to keep this number of FP relatively low, but also offer a user-friendly interface allowing to efficiently going back to the raw data and check the MS signal extraction on all the peptides of a particular candidate protein.

## 5. Conclusion

As outlined in previous reports, benchmark datasets are really needed to evaluate software algorithms in mass spectrometry-based protein analysis, and should be made freely available [41]. All raw MS data generated from the spiked standard presented here have been deposited to the ProteomeXchange Consortium [42] via the PRIDE partner repository with the dataset identifier PXD001819, and quantitative outputs from the different workflows tested are given in ref [30], Sup Table 1. It must be noticed that all these results are dependent on the parameter settings used for each computational workflow, and to this respect, one main utility of this model dataset may be to help the users in optimizing the tuning and finding the best parameters for a particular tool. Additionally, we hope that such spiked datasets could be useful for developers in order to efficiently test algorithms and improve the extraction of intensity metrics for protein quantitation. Finally, post-processing steps such as possible normalization, imputation of missing values, and downstream statistical analysis will also strongly influence the results. The use of spiked datasets could be beneficial to objectively evaluate their performances and their ability to reduce the level of FP and correctly classify variant proteins in large-scale studies.

This material is available free of charge via the Internet at http://pubs.acs.org. Supplementary data associated with this article can be found in the online version, at http://dx.doi.org/10.1016/j.jprot.2015.11.011.

## Notes

The authors declare no competing financial interest.

## Transparency document

The transparency document associated with this article can be found, in online version

## Acknowledgments

This work was funded through the French National Agency for Research (ANR) (grant ANR-10-INBS-08; ProFI project, "Infrastructures Nationales en Biologie et Santé"; "Investissements d'Avenir" call).

## References

[1] S. Nahnsen, C. Bielow, K. Reinert, O. Kohlbacher, Tools for label-free peptide quantification, Mol. Cell. Proteomics: MCP. 12 (2013) 549–556.
[2] K.A. Neilson, N.A. Ali, S. Muralidharan, M. Mirzaei, M. Mariani, G. Assadourian, et al., Less label, more free: approaches in label-free quantitative mass spectrometry, Proteomics 11 (2011) 535–553.
[3] M. Sandin, J. Teleman, J. Malmstrom, F. Levander, Data processing methods and quality control strategies for label-free LC–MS protein quantification, Biochim. Biophys. Acta 2014 (1844) 29–41.
[4] A. Beasley-Green, D. Bunk, P. Rudnick, L. Kilpatrick, K. Phinney, A proteomics performance standard to support measurement quality in proteomics, Proteomics 12 (2012) 923–931.
[5] P.A. Rudnick, K.R. Clauser, L.E. Kilpatrick, D.V. Tchekhovskoi, P. Neta, N. Blonder, et al., Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses, Mol. Cell. Proteomics : MCP. 9 (2010) 225–241.
[6] A.G. Paulovich, D. Billheimer, A.J. Ham, L. Vega-Montoto, P.A. Rudnick, D.L. Tabb, et al., Interlaboratory study characterizing a yeast performance standard for benchmarking LC–MS platform performance, Mol. Cell. Proteomics : MCP. 9 (2010) 242–254.

[7] S.E. Choe, M. Boutros, A.M. Michelson, G.M. Church, M.S. Halfon, Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset, Genome Biol. 6 (2005) R16.
[8] A.R. Dabney, J.D. Storey, A reanalysis of a published Affymetrix GeneChip control dataset, Genome Biol. 7 (2006) 401.
[9] B. De Hertogh, B. De Meulder, F. Berger, M. Pierre, E. Bareke, A. Gaigneaux, et al., A benchmark for statistical microarray data analysis that preserves actual biological and technical variance, BMC Bioinf. 11 (2010) 17.
[10] R.A. Irizarry, L.M. Cope, Z. Wu, Feature-level exploration of a published Affymetrix GeneChip control dataset, Genome Biol. 7 (2006) 404.
[11] R.A. Irizarry, Z. Wu, H.A. Jaffee, Comparison of Affymetrix GeneChip expression measures, Bioinformatics 22 (2006) 789–794.
[12] R.D. Pearson, A comprehensive re-analysis of the Golden Spike data: towards a benchmark for differential expression methods, BMC bioinf. 9 (2008) 164.
[13] B. Hoekman, R. Breitling, F. Suits, R. Bischoff, P. Horvatovich, msCompare: A framework for quantitative analysis of label-free LC–MS data for comparative candidate biomarker studies, Molecular & Cellular Proteomics (. 2012) 11.
[14] C.C. Tsou, C.F. Tsai, Y.H. Tsui, P.R. Sudhir, Y.T. Wang, Y.J. Chen, et al., IDEAL-Q, an automated tool for label-free quantitation analysis using an efficient peptide alignment approach and spectral data validation, Mol. Cell.: MCP. 9 (2010) 131–144.
[15] R.B.A. Zhang, J. Brittenden, J.T. Huang, D. Crowther, Evaluation for computational platforms of LC–MS based label-free quantitative proteomics: a global view, J. Proteomics Bioinf. 3 (2010) 260–265.
[16] C. Christin, H.C. Hoefsloot, A.K. Smilde, B. Hoekman, F. Suits, R. Bischoff, et al., A critical assessment of feature selection methods for biomarker discovery in clinical proteomics, Mol. Cell. Proteomics : MCP. 12 (2013) 263–276.
[17] D. Bouyssie, A. Gonzalez de Peredo, E. Mouton, R. Albigot, L. Roussel, N. Ortega, et al., Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: application to the proteomics study of membrane proteins from primary human endothelial cells, Mol. Cell Proteom.: MCP 6 (2007) 1621–1637.
[18] V. Gautier, E. Mouton-Barbosa, D. Bouyssie, N. Delcourt, M. Beau, J.P. Girard, et al., Label-free quantification and shotgun analysis of complex proteomes by one-dimensional SDS-PAGE/NanoLC–MS: evaluation for the large scale analysis of inflammatory human endothelial cells, Mol. Cell. Proteomics: MCP. 11 (2012) 527–539.
[19] P. Navarro, J. Vazquez, A refined method to calculate false discovery rates for peptide identification using decoy databases, J. Proteome Res. 8 (2009) 1792–1796.
[20] B. Zhang, M.C. Chambers, D.L. Tabb, Proteomic parsimony through bipartite graph analysis improves accuracy and transparency, J. Proteome Res. 6 (2007) 3549–3557.
[21] J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification, Nat. Biotechnol. 26 (2008) 1367–1372.
[22] J. Cox, N. Neuhauser, A. Michalski, R.A. Scheltema, J.V. Olsen, M. Mann, Andromeda: a peptide search engine integrated into the MaxQuant environment, J. Proteome Res. 10 (2011) 1794–1805.
[23] V. Dupierris, C. Masselon, M. Court, S. Kieffer-Jaquinod, C. Bruley, A toolbox for validation of mass spectrometry peptides identification and generation of database: IRMa, Bioinformatics 25 (2009) 1980–1981.
[24] E. Mouton-Barbosa, F. Roux-Dalvai, D. Bouyssie, F. Berger, E. Schmidt, P.G. Righetti, et al., In-depth exploration of cerebrospinal fluid by combining peptide ligand library treatment and label-free protein quantification, Mol. Cell. Proteomics : MCP. 9 (2010) 1006–1021.
[25] J. Cox, M.Y. Hein, C.A. Luber, I. Paron, N. Nagaraj, M. Mann, Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ, Mol. Cell. proteomics : MCP 13 (2014) 2513–2526.
[26] B. MacLean, D.M. Tomazela, N. Shulman, M. Chambers, G.L. Finney, B. Frewen, et al., Skyline: an open source document editor for creating and analyzing targeted proteomics experiments, Bioinformatics 26 (2010) 966–968.
[27] T.V. Pham, S.R. Piersma, M. Warmoes, C.R. Jimenez, On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics, Bioinformatics 26 (2010) 363–369.
[28] J. Cox, I. Matic, M. Hilger, N. Nagaraj, M. Selbach, J.V. Olsen, et al., A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics, Nat. Protoc. 4 (2009) 698–705.
[29] B. Schilling, M.J. Rardin, B.X. MacLean, A.M. Zawadzka, B.E. Frewen, M.P. Cusack, et al., Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation, Mole. Cell. Proteomics : MCP. 11 (2012) 202–214.
[30] C. Ramus, A. Hovasse, M. Marcellin, A.M. Hesse, E. Mouton-Barbosa, D. Bouyssié, et al., Spiked Proteomic Standard Dataset for Testing Label-free Quantitative Software and Statistical Methods, 2015 (Data in Brief, submitted for publication).
[31] J. Albrethsen, J. Agner, S.R. Piersma, P. Hojrup, T.V. Pham, K. Weldingh, et al., Proteomic profiling of Mycobacterium tuberculosis identifies nutrient-starvation-responsive toxin–antitoxin systems, Mol. Cell. Proteomics : MCP. 12 (2013) 1180–1191.
[32] C. Bell, L. English, J. Boulais, M. Chemali, O. Caron-Lizotte, M. Desjardins, et al., Quantitative proteomics reveals the induction of mitophagy in tumor necrosis factor-alpha-activated (TNFalpha) macrophages, Mol. Cell. proteomics : MCP. 12 (2013) 2394–2407.
[33] H. Ichikawa, A. Yoshida, T. Kanda, S. Kosugi, T. Ishikawa, T. Hanyu, et al., Prognostic significance of promyelocytic leukemia expression in gastrointestinal stromal tumor; integrated proteomic and transcriptomic analysis, Cancer Sci. 106 (2015) 115–124.
[34] L. Zhang, Z. Wang, Y. Chen, C. Zhang, S. Xie, Y. Cui, et al., Label-free proteomic analysis of PBMCs reveals gender differences in response to long-term antiretroviral therapy of HIV, J. Proteome 126 (2015) 46–53.
[35] H. Weisser, S. Nahnsen, J. Grossmann, L. Nilse, A. Quandt, H. Brauer, et al., An automated pipeline for high-throughput label-free quantitative proteomics, J. Proteome Res. 12 (2013) 1628–1644.
[36] J. Cox, M.Y. Hein, C.A. Luber, I. Paron, N. Nagaraj, M. Mann, MaxLFQ allows accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, Mol. Cell. Proteomics : MCP (2014).
[37] P.C. Carvalho, J.R. Yates, V.C. Barbosa, Improving the TFold test for differential shotgun proteomics, Bioinformatics 28 (2012) 1652–1654.
[38] D. Dembele, P. Kastner, Fold change rank ordering statistics: a new method for detecting differentially expressed genes, BMC bioinf. 15 (2014) 14.
[39] D.J. McCarthy, G.K. Smyth, Testing significance relative to a fold-change threshold is a TREAT, Bioinformatics 25 (2009) 765–771.
[40] E. Vaes, M. Khan, P. Mombaerts, Statistical analysis of differential gene expression relative to a fold change threshold on NanoString data of mouse odorant receptor genes, BMC Bioinf. 15 (2014) 39.
[41] J.R. Yates 3rd, S.K. Park, C.M. Delahunty, T. Xu, J.N. Savas, D. Cociorva, et al., Toward objective evaluation of proteomic algorithms, Nat. Methods 9 (2012) 455–456.
[42] J.A. Vizcaino, E.W. Deutsch, R. Wang, A. Csordas, F. Reisinger, D. Rios, et al., ProteomeXchange provides globally coordinated proteomics data submission and dissemination, Nat. Biotechnol. 32 (2014) 223–226.

## D.1. Evaluation of instruments and acquisition methods performance using isotopologue peptides

### D.1.1 Context of the project

An essential parameter for quantification, the dynamic range, is not often assessed as its evaluation needs a high number of runs and instrument time (Figure IV-25). The routine evaluation of dynamic range remains a challenge. Previous work presented the proof of principle concept of using isotopologues peptides for evaluating LC-MS/MS performance [12, 13].

Isotopologues only differ by the number of isotopically stable heavy-labeled amino acids included in their peptide sequence. In chromatography, all isotopologues of the same peptide have the same behavior, i.e. the same retention time and elution peak shape. In mass spectrometry, all isotopologues can be separated by their m/z ratio in the MS signal while they have the same response factor and fragmentation patterns. These characteristics make isotopologues ideal to perform absolute targeted protein quantification by SRM [2]. In this context heavy-labeled peptides are used to correct sample preparation biases and MS signal fluctuations. Recent studies have used isotopologues to correct for differences in response factors [13] , as a mean to trigger in real-time PRM experiments in order to optimize the instrument's scanning time [4] or for the routine quality control of both sample preparation and LC–MS platforms[200].

In the present study, we wanted to extend this concept and develop a performance assessment workflow using dilution series of isotopologue standards by deriving calibration curves with an increased number of calibration points and a large concentration range. The use of isotopologue peptides facilitates the creation of calibration curves with large concentration ranges. In this study we wanted to take advantage of this approach's capability to rapidly and confidently evaluate an instrument's dynamic range and sensitivity. We applied it to the comparison of the performance of different LC-MS platforms. We benchmarked different acquisition modes within a single instrument. Using PRM and DIA methods, we wanted to answer whether or not the isolation windows widths in DIA mode have an effect on the dynamic range, and evaluated the quantification using MS1 or MS2 signals in simple and complex samples. We compared quantification performance of nano-flow and capillary-flow platforms. And finally we developed a workflow for routine evaluation of the dynamic range as a performance test. Overall, this study was carried out on five different LC-MS platforms and 17 different experiments were carried out, resulting in 52 different LOQ and dynamic range determinations.

The detailed description of the experimental parameters can be seen on the Experimental Section on page 216.

### D.1.2 Description of the strategy

Figure IV-33 presents the overall strategy. Briefly, a mixture of isotopologue peptides was spiked into a simple Bovine Serum Albumin (BSA) or a complex yeast lysate background matrix. Two sample sets were used in this study:

- **Sample set 1:** Eight synthetic stable-isotope $^{15}N$- and $^{13}C$-labeled isotopologue peptides based on the peptide sequence AALPAAFK (provided from Thermo Fisher) were mixed in the different concentrations each. The mixture of isotopologue peptides was diluted by a factor of two in a background matrix (either 5 fmol/µl of BSA digest or 50ng/µl of total yeast digest). Then two dilutions by a factor of 10 and 100 were done by cascade dilution using the background matrix. The background matrix is used to mimic a proteomic sample and also to avoid peptide adsorption to the walls of vials so it should always be added first. Two microliters were analyzed by LC-MS/MS and each solution was analyzed in triplicate. This resulted in nine injections of

three solutions of isotopologues peptides covering a range of 5,3 logs, i.e. 3 amol to 656 fmol injected on column.

- **Sample set 2:** The 6 × 5 LC-MS/MS Peptide Reference Mix which contains six sets of isotopologue peptides (with each isotopologue peptide in a different concentration) was diluted by a factor of two in a background matrix (either 5 fmol/µl of BSA digest or 50ng/µl of total yeast digest). Two microliters were analyzed by LC-MS/MS and each solution was analyzed in triplicate. For each peptide a calibration curve could be created with the following calibration points: 30 amol, 300 amol, 3fmol, 30fmol and 300 fmol of injected amount on column.

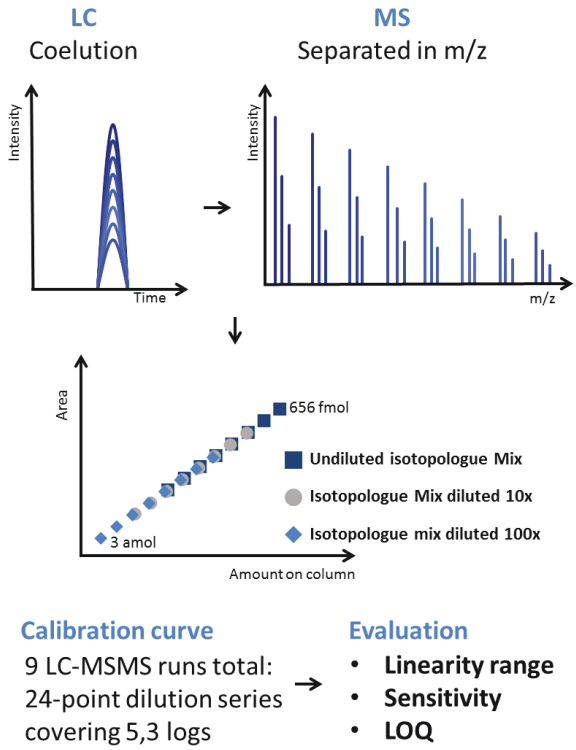



**Figure IV-33: Overview of the analytical workflow.**
A mixture of isotopologue peptides, having the same chromatographic behavior but resolved in MS, was used to create a 24-points calibration curve spanning a 5,3-log range from 3amol to 656 fmol in order to assess the dynamic range, sensitivity and limits of quantification.

To obtain high-quality data, the extracted signal was visually examined to verify the correct peak group identification and integration of peak areas by checking the exact coelution of isotopologue peptides. The limit of quantitation (LOQ) was defined as the last point having a coefficient of variation lower than 20% among triplicate injections, showing an accuracy between 80 and 120% and giving a coefficient of determination $R^2$ higher than 0,98 between the area under the peaks and the injected amount on column, and between the recalculated and the real injected amount on column.

### D.1.3 LC-MS/MS platform comparison

This approach was used to compare the performances of different LC-MS platforms. As an example three nanoLC-MS platforms were compared: AB Sciex TripleTOF 6600, Thermo Q-Exactive plus and a Waters Synapt HDMS all coupled to Waters nanoAcquity systems (Figure IV-34). The isotopologue peptides were spiked in a simple BSA background matrix. The Synapt HDMS, being an 8 years old instrument, was set to monitor only MS scans since its scan rate is lower than the other two instruments. It showed a 2-log dynamic range and below the LOQ (7,3 fmol of injected amount on column), the accuracy and CVs were rapidly off the tolerated values.

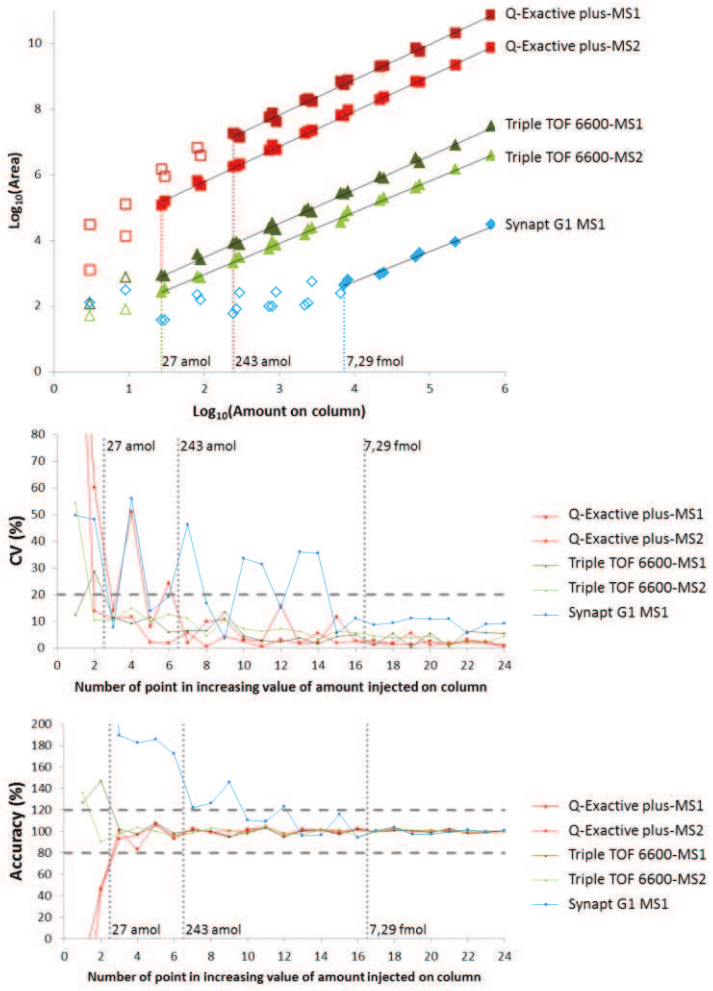



**Figure IV-34: Rapid and accurate performance comparison in terms of sensitivity of three LC-MS platforms.**
Targeted methods were used for each instrument and the MS1 and MS2 signals were used. Five calibration curves are shown for three different LC-MS platforms. Empty calibration points are not validated calibration points. The CV and the accuracy are plotted against the number of calibration points in increasing order of injected amount. All signals were visually evaluated and validated to ensure high-quality results. The vertical dashed lines indicate the LOQ.

The other two instruments were set to perform PRM targeted analysis of the isotopologue peptides. The MS1 and MS2 signal was used to determine the dynamic range. Figure IV-35 shows the results obtained on a Thermo Q-Exactive plus instrument using a PRM method targeting the isotopologue peptides. The curves were created using the MS2 signals. The LOQ was found to be 27 amol of injected peptide into the column. Below this injected amount the accuracy falls out of the 80-120% range. The LC-MS system showed a 4,4-log dynamic range with excellent accuracy and reproducibility. The TripleTOF 6600 proved to have the same performances in targeted analysis when using the PRM MS2 signal with a LOQ equal to 27 amol and a 4,4-log dynamic range (Figure IV-34). For both instruments no saturation effects were detected even at the highest injected amount, suggesting that higher quantities could be used, and that the dynamic range could be larger. Furthermore, the Q-Exactive plus showed a lower LOQ in MS2 signal than with the MS1 signal, respectively 27 and 243 amol of injected peptide on column. This illustrates the gain of sensitivity coming from the increase of the selectivity by quantifying MS2 signals.

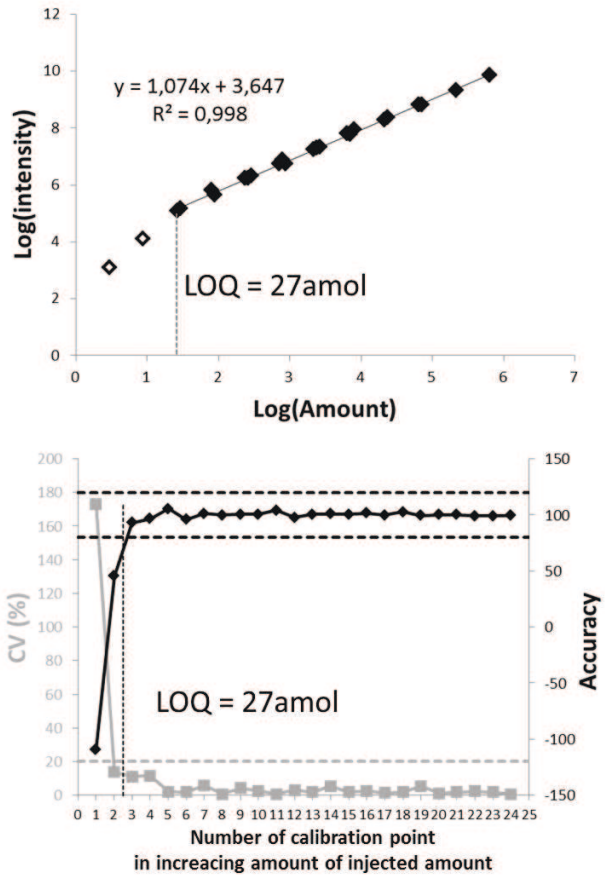



**Figure IV-35: Validation criteria for the LOQ and the dynamic range determination.**
The example here shows data from a PRM experiment where the MS2 signal was used. The bottom graphic shows the CV (%) and the accuracy (%) against the number of calibration points arranged in increasing amount of injected peptide. The upper chart shows the logarithm of the area under curve against the logarithm of the amount of peptide injected on column. The full boxes are validated calibration points (CV<20%; accuracy between 80 and 120%, $R^2$>0.98), the limit of quantitation (LOQ) is shown by the dashed line and empty boxes are calibration points below the LOQ.

The use of isotopologue peptides to determine the dynamic range and limits of quantification is a novel, fast and accurate method to assess instrumental performances. By multiplying the number of occurrences of the same peptide sequence analyzed in a single run, the number of injections needed to create a calibration curve with a satisfying number of points is lowered. Therefore dedicated instrument time is also drastically reduced. Using only 9 LC-MS runs, a calibration curve with 24 points analyzed in triplicate could be constructed. Performing the same experiment without the use of isotopologues would require 72 LC-MS runs and, using a 30-minute method per run, 36 hours would be needed. With the approach described in this paper only 4,5 hours are needed to create a 24-points calibration curve covering a 5,3-log concentration range and analyzed in triplicate.

### D.1.4 Comparison of different acquisition modes in the same instrument

By determining the LOQ in different experimental conditions, we wanted to answer the question of whether or not the isolation windows widths in DIA mode have an effect on the dynamic range. To evaluate this, seven different experimental conditions were tested on an AB Sciex TripleTOF 6600 (Table IV-1). A PRM method was compared the different DIA methods and the analyses were done on a simple and a complex background matrix.

For PRM, an MS survey scan was acquired followed by a set of 8 sequential Q1 isolation windows targeting the isotopologue peptides. The first type of DIA method was a SWATH method covering the 350-1200 m/z range corresponding to the mass range of the majority of tryptic peptides in a bottom-up proteomic experiment. An MS survey scan was acquired followed by a set of 34 sequential Q1 windows with a fixed width of 25 Da. The second type

of DIA method used only 8 isolation windows in order to be able to obtain comparable cycle and dwell times as with the PRM experiments.

**Table IV-1: Changes in sensitivity by the effects of the acquisition mode and the sample complexity.**
These experiments were carried out on an AB Sciex Triple TOF 6600. For each experimental condition, the LOQ is given in attomoles of injected peptide on column and the corresponding dynamic range in logarithmic scale is given in brackets.

| Experiment | Acquisition mode | Number and size of isolation window | Covered range (m/z) | Cycle time (s) | Background matrix | LOQ (amol); Dynamic range Signal: MS1:P, P+1 and P+2 | LOQ (amol); Dynamic range Signal: MS1:P | LOQ (amol); Dynamic range Signal: MS2: y4-y5-y6-y7 | LOQ (amol); Dynamic range Signal: MS2: y6-y7 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | PRM | 8 x 2Da | - | 1,1 | BSA | 27 [4,4] | 27 [4,4] | 27 [4,4] | 243 [3,4] |
| 2 | DIA | 8 x 4Da | 390-422 | 1,1 | BSA | 27 [4,4] | 27 [4,4] | 30 [4,3] | 30 [4,3] |
| 3 | DIA | 8 x 25Da | 369-569 | 1,1 | BSA | 30 [4,3] | 27 [4,4] | N/A* | 81 [3,9] |
| 4 | DIA | 8 x 50Da | 375-775 | 1,1 | BSA | 27 [4,4] | 27 [4,4] | N/A* | 243 [3,4] |
| 5 | SWATH | 34 x 25Da | 350-1200 | 3,1 | BSA | 30 [4,3] | 27 [4,4] | N/A* | 270 [3,4] |
| 6 | PRM | 8 x 2Da | - | 1,1 | Yeast | N/A** | 810 [2,9]*** | 30 [4,3] | 81 [3,9] |
| 7 | SWATH | 34 x 25Da | 350-1200 | 3,1 | Yeast | N/A** | 900 [2,9]*** | N/A* | 300 [3,3] |

*shared fragments making the LOQ assessment impossible; **interfered signals making the LOQ assessment impossible; ***signal detected at lower amounts but is interfered

### *D.1.4.1 Effects of the background matrix on MS1 and MS2 quantification*

A first result that can be noted is that, even when using a high-resolution instrument, the complexity of the yeast background generates interferences in MS1 signals making the determination of the LOQ impossible (Table IV-1). For the PRM experiment (Exp. 6), by eliminating the interfered MS1 signals (P+1 and P+2), the LOQ was determined to be 810 amol of peptides injected on column. This value is significantly higher than the LOQ obtained in a simple BSA matrix (Exp. 1), i.e 27 amol. These results suggest that the MS1 signal is not specific enough and the sensitivity is thus impaired.

For the PRM experiments in both a simple and a complex background matrix (Exp 1 and 6) the LOQ is approximately the same, respectively 27 and 30 amol. The gain in sensitivity when quantifying MS2 signals thanks to the selection of the peptide and the reduction of interfered signals is clearly demonstrated here. The MS2 signal is less affected by the complexity of the sample and can thus be used to achieve a higher sensitivity.

In contrast, when the sample complexity is lower (BSA background matrix, Exp. 1) there is no difference when quantifying MS1 or MS2 signals. For low complexity matrices, quantification on the MS1 signal alone seems to be enough to achieve high sensitivity.

### *D.1.4.2 The problem of shared fragments in DIA experiments*

It is important to note that in DIA multiple peptides are coisolated and fragmented together. The selectivity of this method is thus reduced, leading to the increase of the number of interferences present in the MS2 signals. This is the case in our study where it should be noted that the isotopologue peptides have shared fragment ions. Figure IV-36 shows an example of shared and non-shared fragment ions for two sets of isotopologue peptides used in this study.

A

| AALPAAFK | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Peptide Sequence | AALPAAFK | AALPAAFK | AALPAAFK | AALPAAFK | AALPAAFK | AALPAAFK | AALPAAFK | AALPAAFK |
| precursor | 394,7369 | 402,7511 | 405,2626 | 407,7647 | 410,2731 | 412,7752 | 415,2804 | 418,2873 |
| precursor [M+1] | 395,2384 | 403,2526 | 405,7641 | 408,2662 | 410,7746 | 413,2767 | 415,7821 | 418,7897 |
| precursor [M+2] | 395,7397 | 403,7539 | 406,2652 | 408,7673 | 411,2755 | 413,7776 | 416,2827 | 419,2894 |
| b1 | 72,0444 | 72,0444 | 72,0444 | 72,0444 | 72,0444 | 72,0444 | 76,0515 | 76,0515 |
| b2 | 143,0815 | 143,0815 | 143,0815 | 143,0815 | 143,0815 | 147,0886 | 151,0957 | 151,0957 |
| b3 | 256,1656 | 256,1656 | 263,1827 | 256,1656 | 263,1827 | 260,1727 | 271,1969 | 271,1969 |
| b4 | 353,2183 | 353,2183 | 360,2355 | 353,2183 | 366,2493 | 363,2392 | 368,2497 | 374,2635 |
| b5 | 424,2554 | 428,2625 | 431,2726 | 428,2625 | 441,2935 | 438,2835 | 443,2939 | 449,3077 |
| b6 | 495,2926 | 503,3068 | 506,3168 | 503,3068 | 516,3377 | 513,3277 | 518,3381 | 524,3519 |
| b7 | 642,361 | 650,3752 | 663,4125 | 660,4024 | 673,4334 | 670,4233 | 675,4338 | 681,4476 |
| y1 | 147,1128 | 155,127 | 147,1128 | 155,127 | 147,1128 | 155,127 | 155,127 | 155,127 |
| y2 | 294,1812 | 302,1954 | 304,2084 | 312,2226 | 304,2084 | 312,2226 | 312,2226 | 312,2226 |
| y3 | 365,2183 | 377,2396 | 379,2527 | 387,2669 | 379,2527 | 387,2669 | 387,2669 | 387,2669 |
| y4 | 436,2554 | 452,2838 | 450,2898 | 462,3111 | 454,2969 | 462,3111 | 462,3111 | 462,3111 |
| y5 | 533,3082 | 549,3366 | 547,3425 | 559,3638 | 557,3634 | 565,3776 | 559,3638 | 565,3776 |
| y6 | 646,3923 | 662,4207 | 667,4438 | 672,4479 | 677,4647 | 678,4617 | 679,4651 | 685,4789 |
| y7 | 717,4294 | 733,4578 | 738,4809 | 743,485 | 748,5018 | 753,5059 | 754,5093 | 760,5231 |

B

| YVYVADVAAK | | | | | |
|---|---|---|---|---|---|
| Peptide Sequence | YVYVADVAAK | YVYVADVAAK | YVYVADVAAK | YVYVADVAAK | YVYVADVAAK |
| precursor | 553,8022 | 556,8091 | 559,816 | 562,8229 | 566,83 |
| precursor [M+1] | 554,3038 | 557,3107 | 560,3176 | 563,3245 | 567,3316 |
| precursor [M+2] | 554,8051 | 557,812 | 560,8189 | 563,8257 | 567,8328 |
| y9 | 943,5339 | 949,5477 | 955,5615 | 961,5753 | 969,5895 |
| y8 | 844,4654 | 850,4793 | 856,4931 | 856,4931 | 864,5073 |
| y7 | 681,4021 | 687,4159 | 693,4297 | 693,4297 | 701,4439 |
| y6 | 582,3337 | 588,3475 | 588,3475 | 588,3475 | 596,3617 |
| y5 | 511,2966 | 517,3104 | 517,3104 | 517,3104 | 525,3246 |
| y4 | 396,2696 | 402,2835 | 402,2835 | 402,2835 | 410,2977 |
| y3 | 297,2012 | 297,2012 | 297,2012 | 297,2012 | 305,2154 |
| y2 | 226,1641 | 226,1641 | 226,1641 | 226,1641 | 230,1712 |
| y1 | 155,127 | 155,127 | 155,127 | 155,127 | 155,127 |
| b1 | 164,0706 | 164,0706 | 164,0706 | 164,0706 | 164,0706 |
| b2 | 263,139 | 263,139 | 263,139 | 269,1528 | 269,1528 |
| b3 | 426,2023 | 426,2023 | 426,2023 | 432,2162 | 432,2162 |
| b4 | 525,2708 | 525,2708 | 531,2846 | 537,2984 | 537,2984 |
| b5 | 596,3079 | 596,3079 | 602,3217 | 608,3355 | 608,3355 |
| b6 | 711,3348 | 711,3348 | 717,3486 | 723,3624 | 723,3624 |
| b7 | 810,4032 | 816,417 | 822,4308 | 828,4447 | 828,4447 |
| b8 | 881,4403 | 887,4542 | 893,468 | 899,4818 | 903,4889 |
| b9 | 952,4775 | 958,4913 | 964,5051 | 970,5189 | 978,5331 |

**Figure IV-36: Example of shared and non-shared fragment ions in two isotopologue peptides.**
Two sets of isotopologue peptides are shown. The heavy-labelled amino acid is shown in red. The m/z ratio of precursor and fragment ion for each isotopologue peptide is given. The cells in gray are shared transitions between peptides. A. For peptide AALPAAFK only the b7, y6 and y7 transitions are not shared. B. For peptide YVYVADVAAK only the y9, b8 and b9 are not shared but this transitions are not well-responding. This peptide cannot be used to evaluate DIA approaches with large isolation windows.

When using an isolation window large enough to co-isolate two or more isotopologues the determination of the dynamic range becomes impossible, as the signal from one isotopologue peptide interferes with the signal of another isotopologue peptide. The PROMEGA 6x5 LC-MS/MS peptide mix (sample set 2) could not be used to evaluate DIA strategies as the majority of fragment ions were shared. Also for the peptide from the sample set 1, its two most intense fragment ions are shared between isotopologue peptides. These two transitions can thus not be used to evaluate DIA analyses. However the y-6 and y-7 fragment ions can be used as they are specific to each of the isotopologue peptides. These transitions are only the 3$^{rd}$ and 4$^{th}$ most intense fragment ions for this peptide, thus the sensitivity will be impaired.

### *D.1.4.3 Effects of DIA isolation windows size on the dynamic range*

The isotopologue peptide from the sample set 1 had specific y5- and y6-ions for each of the eight isotopologue peptide. These two transitions are only the third and fourth most intense transitions for this peptide, increasing thus the value of the LOQ for this peptide. However they could still be used to evaluate DIA approaches.

Using the y5- and y6-ions the performances of DIA were found to be similar to PRM as the LOQ was found to be in the range of hundreds of attomoles. Additionally the isolation window size in DIA seems to have no significant effect on the sensitivity of the instrument (Exp 1 to 4).

Another important result is that even if the SWATH experiment (34x25Da) maps a larger m/z range and has a cycle time three times longer than the other methods; the LOQ is not significantly different of that obtained by PRM (Exp. 5 and 7). Overall, the instruments performance does not seem to be massively affected by the different acquisition modes. The most influent parameter on the performance is the loss of selectivity when changing the isolation-window size; this has certainly a higher effect on impairing the sensitivity. However these results should be viewed cautiously as in this comparison the background matrix was not complex and the results reflect the behavior of a small set of peptides.

### D.1.5 Evaluation of chromatography scale: Comparison of Nano-LC-PRM vs. capillaryLC-SRM systems

Nano-flow LC platforms are most widespread in proteomic research laboratories. NanoLC provides good sensitivity, high peak capacity, high resolution and it enables low sample injection volume. But technical problems common to nanoLC still remain a challenge, such as nanoESI spray instability, not-easily detectable leaks, high back pressure or dead volumes. Standard-flow was found to provide globally superior sensitivity than nano-flow, with higher retention time reproducibility and increased ease of use [14]. For projects where sample amount is not an issue, this alternative platform can be ideal to enhance analysis throughput and reduce instrument downtime. Capillary-flow LC is another option that lays in-between nano-flow and standard-flow.

We applied the method described in this paper to compare the performance of our nanoLC-PRM system versus our capillaryLC-SRM system. The sample set 1 and 2 described above with a yeast background matrix were used. The LOQ and dynamic range determination can be seen in Table IV-2.

**Table IV-2: Comparison of NanoLC-PRM to CapillaryLC-SRM.**
For each experimental condition the LOQ values are given in amol of peptide injected on column, the corresponding dynamic range is the value in brackets. For the third experiment, 9,6 times more sample was injected into the column to take advantage of the higher sample capacity of the capillary-flow platform, the equivalent amounts in nano-flow conditions are shown in parenthesis.

| Instrument | Acquisition mode | LC conditions | Relative injected amount | LASVSVSR (y4-y5-y6-y7) | YVYVADVAAK (y4-y6-y7-y8) | VVGGLVALR (y4-y6-y7-y8) | LLSLGAGEFK (y5-y6-y7-y8) | AALPAAFK (y4-y5-y6-y7) |
|---|---|---|---|---|---|---|---|---|
| Q-Exactive plus | PRM | Nano-flow | x1 | 300 [3-log] | 300 [3-log] | 30 [4-log] | 300 [3-log] | 27 [4,4-log] |
| TSQ Vantage | SRM | Capillary-flow | x1 | 300 [3-log] | 300 [3-log] | 300 [3-log] | 300 [3-log] | x |
| TSQ Vantage | SRM | Capillary-flow | x9,6 | 288 (30) [4-log] | 288 (30) [4-log] | 288 (30) [4-log] | 288 (30) [4-log] | 2592 (270) [3,4-log] |

First, a direct comparison was made between the two platforms, i.e. the same solution and exactly the same amount was injected on both systems. Surprisingly, the nano-flow and the capillary-flow systems showed globally the same performances, giving both a 3-log dynamic range and a LOQ of 300 amol. Only the peptide VVGGLVALR was found to have a larger dynamic range (4-logs) and a lower LOQ (30 amol) in nano-low compared to capillary-flow.

A second comparison was made by taking into account the higher sample loading capacity of the capillary-flow platform. The total amount of sample loaded on column was 9,6 higher to account for the up-scaling from nano-flow to capillary-flow column dimensions. The calibration curve spanned now from 2,88 pmol to 288 amol of injected peptide and 9,6µg of yeas background matrix injected on column. From this comparison, the sensitivity of the capillary-flow LC-SRM platform was found to be 10-fold higher when the up-scaling was taked into account. In these conditions, 4 out of 5 peptides attained at least a 4-log dynamic range and a LOQ equal to 288 amol which is equivalent to 30 amol in nano-flow conditions. What is important to note is that for these peptides the last measured

calibration point had coefficients of variation and accuracy values respecting the tolerated criteria. This suggests that the LOQ could be even lower and the dynamic range larger for these peptides.

However, peptide AALPAAFK had the opposite behavior. It showed a better sensitivity when using the nano-flow platform, suggesting that, even if globally the capillary-flow system is more sensitive than the nano-flow system, this trend is peptide-dependent.

As a conclusion, when access to sample is not an issue, capillary-flow seems to be a robust alternative to nanoflow allowing high throughput analysis and lower instrument downtime without losing quantification performances. However it is important to keep in mind that these results only reflect the behavior of a small set of peptides.

### D.1.6 Routine dynamic range evaluation as an instrument performance test

Several system suitability experiments have been developed specifically for quantitative proteomics by LC-MS/MS. These experiments use metrics based on chromatogram signal extraction (such as retention time variability, FWHM, chromatographic peak shape, peak capacity, etc..). In this experiments sensitivity is often determined by analyzing standards at known quantities and comparing the measured response, i.e peak areas or heights, to a reference value [195, 197, 201]. Quality control (QC) standards are often used to check system performance. These samples are regularly injected all over the sample sequence and analyzed at fixed times.

In this context, a mixture of isotopologue peptides would constitute a perfect standard to monitor LC-MS instruments over time, to benchmark instrument performance and routinely assess sensitivity, as it gives a full view of the instrument's dynamic range.

To assess the feasibility of this idea, a targeted PRM experiment was setup on a Bruker Daltonics Impact II system. The same sample set 1 solution was analyzed before and after the instrument cleaning and maintenance (Figure IV-37.A). Before maintenance, the LOQ was 2,4 fmol of injected amount on column and the dynamic range spanned 2,4-logs. The high LOQ value suggested a problem in the instruments sensitivity. The instrument was cleaned and the collision cell received maintenance. After this, the sensitivity was increased by a factor of 8 to reach a 3,3-log dynamic range and a LOQ of 300 amol.

What is important to note is that, since the equation of the calibration curve is directly related to the instrument's sensitivity, the slope and the intercept of the calibration curve changed in a significant manner after the maintenance. Here we propose an idea of how to use these mixtures to make a simple, cost-efficient and rapid performance test (Figure IV-37.B). First, using the method described in this section, a signal intensity corresponding to a level close to the LOQ should be defined for each peptide ($y_{ref}$ in Figure IV-37.B). The equation of the calibration curve can then routinely be easily derived using a single injection of an isotopologue mixture. The LOQ can rapidly be approximated by looking for the intersection of the calibration curve with the horizontal line corresponding to the intensity under which the signal is not of good quality ($y_{ref}$ in Figure IV-37.B). The x-value of this interception ($x_{LOQ}$ in Figure IV-37.B) will be an approximation of the LOQ. This value can be compared to a reference value ($x_{ref}$ in Figure IV-37.B) to obtain a pass/fail type of performance test.

In Figure IV-37.C the calibration curve of the instrument before and after maintenance was plotted using only one sample injection. The $x1_{LOQ}$ and $x2_{LOQ}$ shows the value of the approximated LOQ values for each injection. It can clearly be seen that the sensitivity improved after the maintenance. The estimated value of the LOQ after maintenance was 6,5 times lower than before the maintenance. The equation of the calibration curve changes in a significant manner. If only a single data point was used to assess the sensitivity, the instrument's bad performance could have gone

unnoticed as the two curves are close to each other at high concentrations. Since this performance test provides a full view of the instruments response and dynamic range the sensitivity of the instrument can be more accurately evaluated.

The isotopologue mix could be used as a QC sample with a short LC gradient to limit the dedicated instrument time, but by doing so they may not be giving the best representation of the instrument when using a long gradient. The isotopologue mixture could be used spiked directly into the samples to measure dynamic ranges directly in the sample of interest, and increase sample throughput.

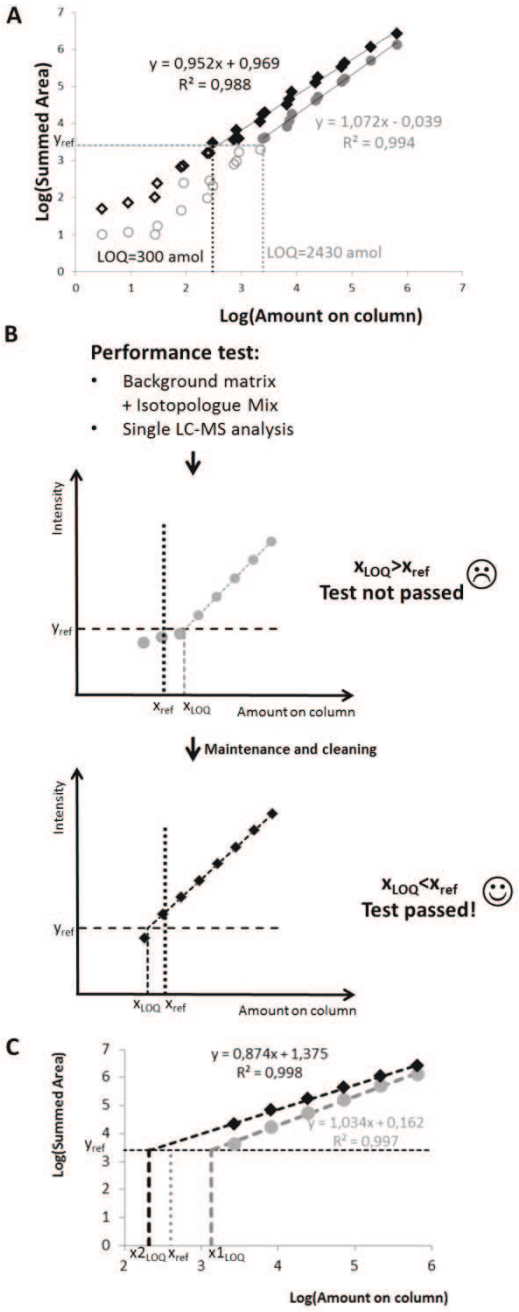



**Figure IV-37: Routine dynamic-range evaluation using isotopologue peptides.**
A. Dynamic range and sensitivity assessment before (gray) and after (black) instrument maintenance and cleaning. The dashed lines correspond to the LOQ values. $y_{ref}$ indicates an intensity under which the signal is no longer of good quality. B. The strategy of a performance test using isotopologue peptides is described. A single LC-MS is sufficient to generate a calibration curve that can be used to evaluate the instrument sensitivity. C. Application of the strategy to an instrument before (gray) and after (black) maintenance. The approximation of the LOQ obtained with the extrapolation of the calibration curve is a good mean to assess the instrument sensitivity.

### D.1.7 Conclusion and perspectives

In the present study we applied and extended the concept of using isotopologue peptide standards for instrument performance evaluation. We developed a performance assessment workflow using dilution series of isotopologue standards by deriving calibration curves with an increased number of calibration points and a larger concentration range. The accurate performance assessment of five different nanoLC-MS platforms with different acquisition methods was conducted. This would not have been possible without the use of isotopologue peptides as more than a thousand injections would have been necessary to reproduce the results presented here. The comparison of nano-flow and capillary-flow platforms showed that capillary-flow systems constitute an appropriate alternative to nanoflow systems for protein quantification as they provide good sensitivity, higher robustness and achieve equal or better performance in terms of sensitivity. Using PRM and DIA methods on the same instrument, we showed that the isolation-window width in DIA analysis does not have a significant effect on the instrument's sensitivity but the most significant impact on the performance comes from the loss of selectivity. However these results should be viewed cautiously as in this comparison the background matrix was not complex and the results reflect the behavior of a small set of peptides.

We also introduced a workflow aimed at routine evaluation of instruments' dynamic range: a performance test that allows a fast, accurate and complete view of the instrument's sensitivity.

Finally, the use of isotopologue peptides is a powerful tool that will be useful in the proteomic community. However the development of peptide mixtures, without shared fragment ions, is still to be achieved to obtain standards suitable for the evaluation for new technologies such as DIA.

A publication resuming the results of this project is in preparation.

## E. Optimization and method development of Data-independent Acquisition methods

### *E.1. Data-Independent method optimization*

DIA is often presented as a "plug-and-play" method using a unique set of parameters to quantify proteins independently of the sample type. However, for DIA methods based on sequential isolation windows it is important to correctly parameter the instrument in order to achieve the best selectivity, sensitivity, quantification accuracy and proteome coverage. During my last 2 PhD years I was responsible for the maintenance, the method development, the training of new users and the day-to-day operations of a new LC-MS system in the laboratory (Waters nanoAcquity; AB Sciex TripleTOF 6600). My task was thus to evaluate the different acquisition methods available, prepare default acquisition methods and train new users to setup their methods in the laboratory. In this context, I gained expertise in DIA methods setup and will thus present and discuss the key parameters to be optimized to setup DIA methods in the following section.

**Scanning times, selectivity and proteome coverage:** The time spent to measure a given isolation window is called the accumulation time. It is often in the range of 40ms and 100ms. This parameter is related to the sensitivity of the analysis. The higher the accumulation time the higher the signal-to-noise ratio and thus the higher the sensitivity. The cycle time corresponds to the time necessary to monitor the complete m/z range. This parameter depends on the average chromatographic peak widths, to obtain at least 8-10 points per peak. The lower the cycle time the more points will be acquired to reconstruct the chromatographic peak for a given transition and thus the better the accuracy of the quantification will be. Typical chromatographic peak widths range between 15 to 30 s. Thus the cycle

time often ranges around 3 seconds. The proteome coverage depends on the total m/z range analyzed. This depends on the peptide distribution on the m/z dimension of the sample to be analyzed. Commonly tryptic peptides distribute in the range of 350 to 1200 m/z with a high density in the range of 500 to 800 m/z. The range that can be covered is also dependent on the scanning rate of the instrument. In order to cover a larger m/z range, DIA methods use large isolation windows (10-50m/z). However even if high-resolution and high-accuracy instruments are used this approach is still vulnerable of co-eluting interferences that handicap the quantification. A balance between the number and the width of isolation windows, and the instrument's scanning times is necessary.

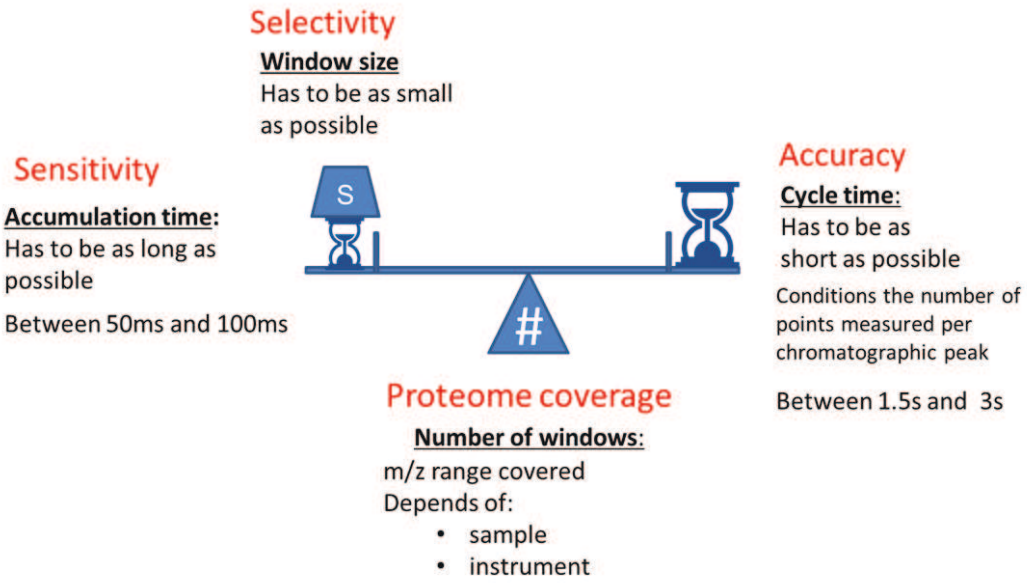



Figure IV-38: Balancing LC-MS parameters to optimize DIA methods

For example, the SWATH method presented by Gilet *et al.* used 32x25Da isolation windows each one acquired with an accumulation time of 100ms to cover the 400-1200m/z range. It also used a MS survey scan with an accumulation time of 100ms and thus totaled a cycle time of 3.3s [6]. In order to increase the selectivity, smaller windows could be used but this could affect the sensitivity as the number of windows required to scan the same total m/z range must be increased, thusly reducing the accumulation time of each isolation window.

**Q1 isolation window overlaps:** The quadrupole ion transmission in Triple-TOf instruments was shown to be almost squared shape. However in the border of the isolation windows the transmission is not optimal. To overcome this, two consecutive SWATH windows need to have an overlap of at least 1Da to ensure that a peptide in the borders of the window will be analyzed correctly in at least one of the overlapping windows. Then 0.5Da margins will be set and no data will be used in these regions (Figure IV-39).

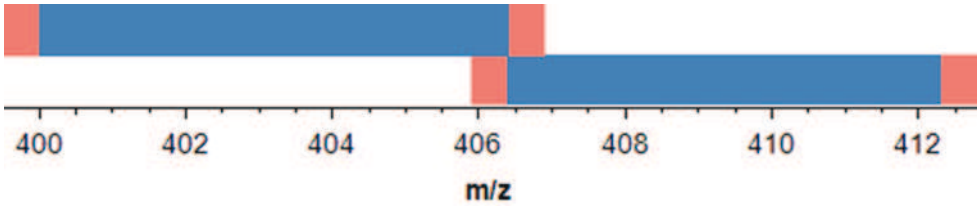



Figure IV-39: Q1 isolation window overlaps and margins.

**Collision energies:** Contrary to targeted quantitative proteomics where all instrument parameters can be optimized for a small set of peptides, in DIA the parameters must be averaged to fit a majority of peptides. The collision energy is an example of this. It cannot be optimized for a given peptide in an isolation window. Commonly, the collision

energies for doubly charged peptides are used. This means that peptides with a charge higher than two can be non-optimally fragmented and it will not be possible to quantify these peptides.

Additionally since a spectral library will be used to target the signal extraction, this must preferably be acquired with the same collision energy values as the ones used for the DIA experiment [202]. This will provide a more accurate comparison between the extracted signals and the corresponding MS/MS spectra in the library.

**MS full scan:** Even if the MS scan is not required to create a DIA method the selectivity can be increased if the MS1 signal is present[203]. It can also help to discriminate a posteriori ambiguous peak group identifications. This will be discussed further in detail below.

**Mass accuracy:** The mass accuracy is a very important parameter for DIA, especially because the signal extraction is done using the exact theoretical mass of the targeted peptides. And contrary to DDA peptide identification where a post-acquisition recalibration can be done, in DIA this step is challenging. When using hybrid Q-Orbitrap instruments this is not a problem as the mass calibration is very stable over time and a lock-mass recalibration is done constantly. However for the Triple-TOF 6600 instrument, the software does not integrate a lock-mass recalibration. Instead the strategy used to correct for mass drifts, is to finely control the temperature inside the instrument and use an external calibration to periodically correct the mass calibration shifts. The external calibration is an injection of a simple mixture of tryptic peptides (Bovine Serum Albumine, BSA in our case), and the instrument extracts the precursors and fragments masses of selected peptides to perform an external calibration of the instrument in MS and MS/MS mode. Figure IV-40 shows the measurement of 50 consecutive injections of a BSA digest using a 30-minutes gradient. The average mass accuracy error of six monitored BSA peptides is shown over time. This experiment was done after observing important mass shifts in the measurements over time. The blue trace shows significant mass accuracy errors over time and important mass deviations from one injection to the other. The reason for this problem was a badly tuned temperature inside the instrument. The temperature regulation system could not correct for temperature changes in the room even if the difference was lower than 2°C during the study. An electronic problem could also be implicated but this hypothesis could not be verified. The red trace shows the results of the same experiment once the instrument received maintenance. The mass errors are within the 5 ppm tolerance and the deviations from one injection to the other are much lower. In this case the deviations can be corrected by the use of periodically injected external calibration samples.

In conclusion, it is important to often check the mass accuracy in this type of instruments. An error in the mass calibration can drastically bias the quantification results (as discussed in Part IVChapter ID.4.3on page 110).

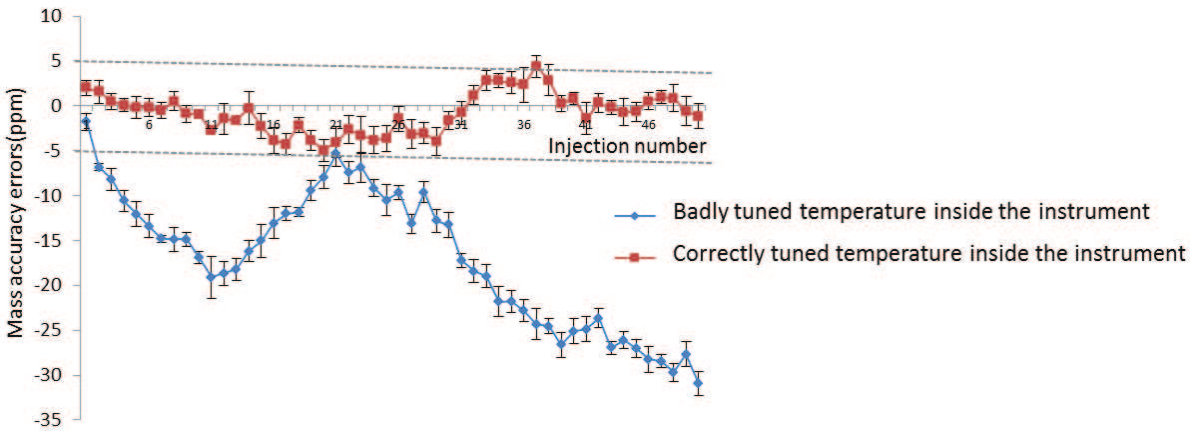



**Figure IV-40: Monitoring mass errors in a Q-TOF instrument over time.**

**Mass resolution:** For Q-TOF instruments there is a trade-off between the resolution and the sensitivity. For the Triple TOF 6600 two resolutions can be used in MS/MS mode: 20k and 35k. However the sensitivity is divided by at least a factor of two when choosing the latter. For the Q-Exactive instrument the resolution is directly proportional to the transient Length. In order to obtain correct scan rates necessary to cover a large m/z range and obtain correct cycle times for quantification, the lowest resolution value must be used (17,5K at 200 m/z). However, this resolution is not constant throughout the m/z dimension and is more than two times lower at 1000m/z (7,8K) than the one of the Triple-TOF 6600 which can be considered constant [204]. However recent technological advances of hybrid Q-orbitrap instruments have doubled the resolving power attained with the same transient times. This can directly benefit DIA methods as a higher resolution can be used without affecting the cycle time [194].

Figure IV-41 shows the proportion of fragment ions showing interference in PRM analyses of 122 isotopically labeled peptides spiked into a urine matrix. When the Orbitrap resolving power is increased the number of interfered transitions decreases. When the isolation window decreases the proportion also decreases considerably.

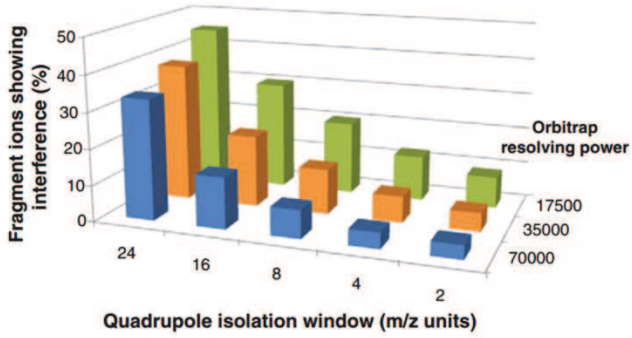



**Figure IV-41: Comparison of the proportion of fragment ions showing interference in PRM analyses of 122 isotopically labeled peptides spiked into a urine matrix as a function of various combinations of experimental conditions (quadrupole isolation window and Orbitrap resolving power) (Adapted from [205]).**

### *E.2. Recent developments in DIA acquisition modes*

Recent remarkable technical progress made in mass spectrometry in terms of resolution, mass accuracy, scan speed and sensitivity have made available to the scientific community a panel of high performing acquisition modes, allowing large-scale protein analysis. For Data-Independent acquisition several new acquisition strategies have emerged in the last years.

**Stepwise sequential predefined isolation windows:** Figure IV-42.A. shows the peptide distribution during an 80-minute long DDA analysis of a whole yeast lysate sample. It is clear to see that the tryptic peptides observed here are not well distributed along the m/z ratio or the time dimension. Small hydrophilic peptides are eluted in early parts of the gradient and large hydrophilic peptides are eluted at late parts of the gradient. For this sample the distribution of precursor ions shows that peptides between 400 to 700 m/z ratio are more common (Figure IV-42.B.). Undoubtedly the use of a common SWATH method for this sample would result in inefficient instrument scanning time, and very complex MS/MS spectra for the isolation windows of low mass range. The use of varying isolation windows according to the ion density would help to evenly distribute the ion population and thus increase the selectivity and the overall quantification performances. The use of variable windows in SWATH method was termed SWATH 2.0 [98].

I have applied this method to the analysis of the whole yeast lysate sample. Figure IV-42.C. shows the common Swath method (32x25Da isolation windows) and a customized variable window Swath method. For this method 67 isolation windows were used with an accumulation time of 45ms to cover the 400-1250 m/z range. A MS1 full scan of 150ms

was also acquired in each cycle. The total cycle time was 3.2s. The size of the isolation windows is show in Figure IV-42.D. Small windows of 7Da were used to cover the m/z range with the most dense ion population and larger windows were used to cover the less populated regions with isolation windows larger than 30 Da and up to 120Da.

Zhang et al. introduced an open-source tool, swathTUNER, to create variable isolation windows based on different information to further optimize the methods [206]. The windows can be set to contain the same number of peptides ions or the same total ion intensity. The latter can be useful to distribute high abundant protein in different windows than low-abundant proteins. This can thus help to avoid loss of intra-spectrum dynamic range.

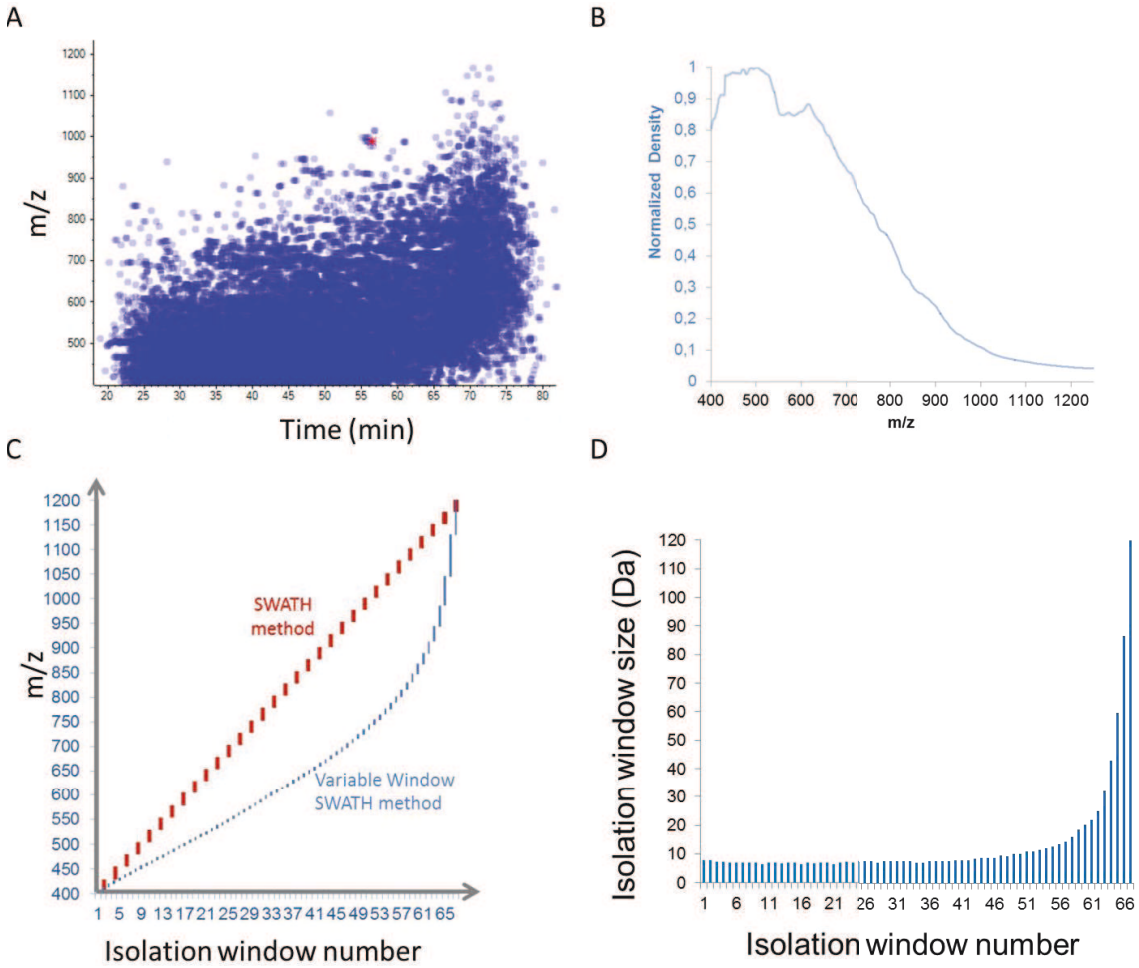



**Figure IV-42: Variable isolation window SWATH method.**
A. Distribution of peptides over the m/z and the time dimension. B. Density of ions according to the m/z ratio. C. Comparison of a typical 32x25Da SWATH method and a customized variable window SWATH method. Smaller isolation windows are used to analyze the regions of high ion density. D. Isolation window sizes. For high ion density regions 7Da windows are used to improve selectivity and sensitivity.

**Multiplexed DIA data Random windows:** Egertson *et al.* presented in 2013 a new acquisition method termed MSX [207, 208]. This method uses the possibility of multiplex capabilities of Q-Exactive instruments (see the description of the multiplex mode in Part IIChapter IIIB.2. on page 51). The method consists in dividing the 500-900m/z range in 100 non-overlapping 4Da isolation windows. For each MS/MS spectrum, five isolation windows are randomly chosen and are sequentially isolated, fragmented and then analyzed simultaneously (Figure IV-43.A). The process is repeated until all 100 isolation windows are analyzed. For each cycle the isolation windows are randomly chosen. This method is equivalent to analyzing the same m/z range with a method of 20x20m/z consecutive isolation windows. However the information contained in the randomly chosen isolation windows and analyzed together will enable the demultiplexing of the spectra to generate demultiplexed pseudo MS/MS spectra corresponding to 4Da isolation windows. This method was shown to improve the selectivity and the signal-to-noise ratio. However, the main

drawback of this method is that the fill times needed to obtain a correct cycle time handicaps the performances of this method. Indeed the ion counts showed that the instrument does not perform at the maximum of its possibilities as fill times do not let the trap to completely fill [207].

To avoid this problem an improved method was proposed consisting of a method resembling a 20x20m/z consecutive isolation windows method. But even-numbered cycles cover the 500-900m/z range and odd-numbered cycles cover have a 10m/z offset to cover the 510-910m/z range (Figure IV-43.B) [209]. The information of overlapping MS/MS spectra can be used to generate demultiplexed MS/MS spectra equivalent to 10m/z isolation windows (Figure IV-43.C). This is done by looking for common ion fragments present in overlapping isolation windows belonging to two consecutive cycles. Using the common fragment ions a new MS/MS spectra will be generate corresponding to an pseudo isolation window of 10Da.

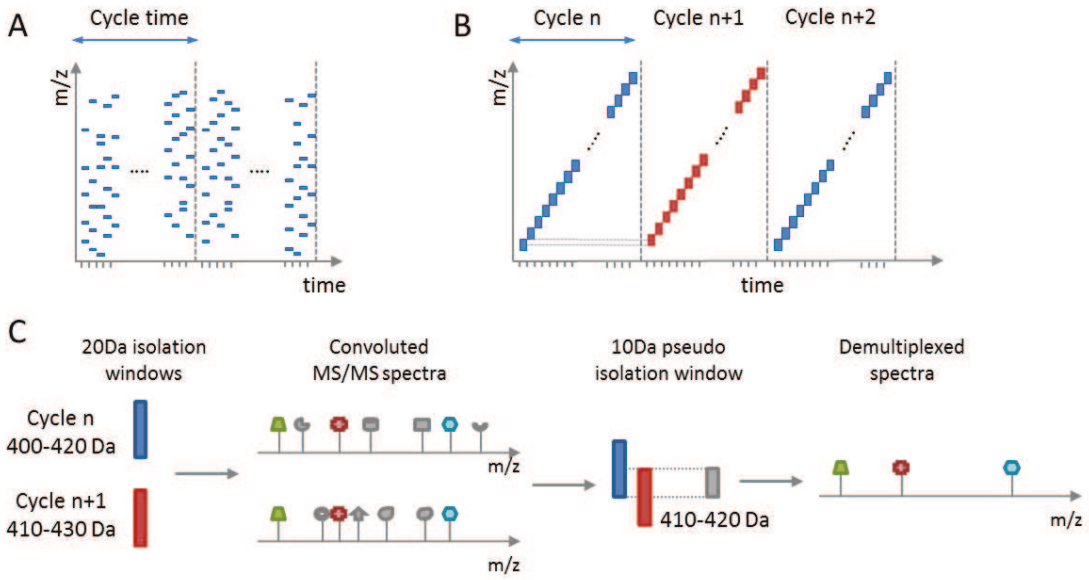



**Figure IV-43: Multiplexed DIA approaches.**

## *E.3. DIA Data analysis*

### E.3.1 Targeted data extraction – Peptide-centric approaches

DIA data by its definition is very complex. As wide isolation windows are used several precursor ions are fragmented together. The MS/MS spectra are highly convoluted data. Contrary to DDA data where one MS/MS spectrum corresponds to only one precursor ion, in DIA this changes. New data analysis methods had to be developed in order to overcome the complexity of DIA data. This approach has been implemented by software like PeakView (AB Sciex), Skyline [15], Spectronaut [16] and OpenSWATH [210].

Gillet *et al*. proposed an approach for DIA data analysis, initially applied for SWATH data, named targeted data extraction [6]. A peptide's identification and quantification is performed by using the peptide's prior information obtained by DDA approaches and stored in a spectral library. Then fragment ions traces are extracted for peptides of interest in DIA data and then the quality of the data is assessed to validate the peptide's identification. This peptide-centric approach (as opposed to spectrum-centric that use database search algorithms to identify the peptides) uses chromatographic characteristics of the extracted signals to verify the identification of the targeted peptides. Like SRM the metrics to validate the peptide's identification are the co-elution of fragment ions, the peak shape, the fragment-ions relative intensities and the retention time. Other additional metrics can be used in DIA due to the fact that the analysis is done using high-accuracy high-resolution instruments. These are the mass accuracy of the signal, the co-

elution of precursor and fragment ions and the co-elution of different charge states. This approach gives to DIA data SRM-like characteristics in the sense of data completeness. Like SRM, the data is comprehensive for targeted peptides and no missing values are present in the data. However DIA has the advantage over SRM of being able to provide data for any peptide of interest by specifically extracting their signals form the data without the need to reacquire new data.

DIA is very noisy and this makes peptide picking very challenging. Common challenges concerning DIA Data analysis will be discussed below.

**Spectral library generation:** Generating a spectral library for DIA experiments has to respect certain considerations.

First, it is important to consider that DDA undersampling is still a problem in modern day instruments. To be able to increase the number of identifications of low-abundant proteins and the quality of their spectra it is recommended to build the spectral library using fractionated samples [202]. It is preferred to use the same instrument to build the spectral library and to perform the DIA analysis, as the fragmentation pattern (and the charge state distribution) can change between instruments. Additionally, contrary to DDA analysis where a single spectrum can be used to identify a peptide, for a spectral library all identified and validated peptide must have high quality spectra with representative fragmentation pattern. To this aim, the instruments parameters must be set to obtain high quality spectra instead of favoring deepness of analysis. Accumulation times must be longer and exclusion times must be short to enable a spectral redundancy to ensure the acquisition of high quality data. To counter the effect of the loss of depth of the analysis, fractionation can be used. Finally retention time standard peptides must be used to correct retention time shifts between the analysis used to build the spectral library, but also to enable the prediction of retention times for the DIA targeted data extraction.

**Interferences**: Event if high-resolution and high-accuracy instruments are used in DIA analysis, due to the use of broad isolation windows, the signal can suffer from interferences. Figure IV-44 shows two examples of peptides with interfered signals. The first one shows a peptide (TREIHNEAESQLR) with interfered MS1 signal. However the MS2 signal is clean. This shows the undeniable advantage of quantification using the MS2 signal. The second example shows a peptide with interfered MS2 signal (GALATYGLTIDDLGVASFHGTSTK). However the advantage of DIA is that all fragment ions are recorded. In this case the interfered could be deleted and this peptide can be correctly quantified. Interferences handicap the use of automatic validation pipelines as these try to assess the identification of the peptides based on the similarity of the chromatogram traces with the corresponding spectra in the spectral library. If interferences are present then the peak picking algorithm can erroneously chose another peptide and bias the quantification. Moreover it can assign a bad score to a peptide that will eventually not be quantified even if it is present in the sample and could be quantified with another set of transitions (see Part IVChapter IA.7.2. on page 81).

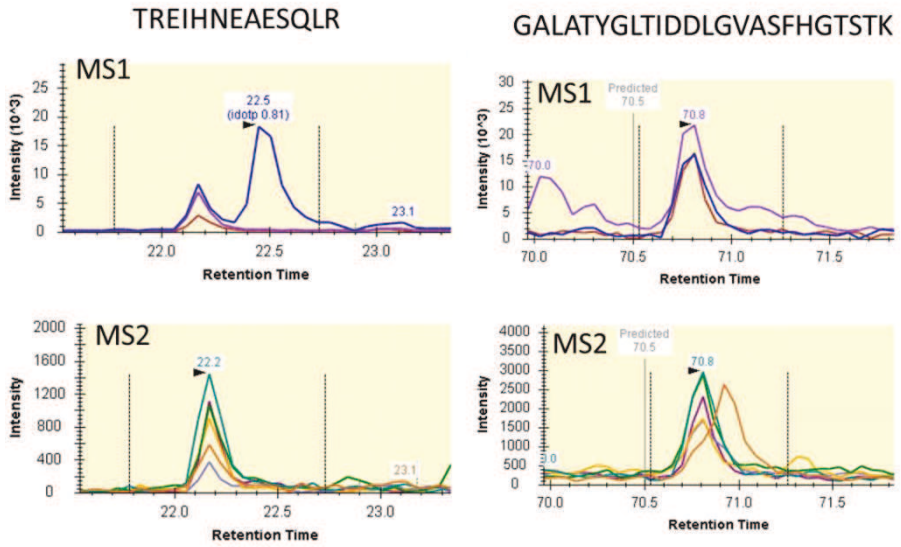



Figure IV-44: Interferences in MS1 and MS2 DIA signals.

**Automatic vs. manual validation**: Figure IV-45 shows the gain of manual data refinement. By eliminating two interfered transitions the dot-product jumps from 0.81 to 0.97 showing a high similarity between the DIA trace and the corresponding spectra in the library.

However in typical DIA studies thousands of proteins and tens of thousands of peptides are analyzed. Additionally when large-scale studies are performed manual validation is no longer feasible. Automatic validation tools like mProphet have emerged as a solution to this problem[16]. However there is still room for progress to automate the data analysis. Two important points need to be addressed: the peak piking and the elimination of interfered signals. An example of this will be shown below for the complicated case of PTMs.

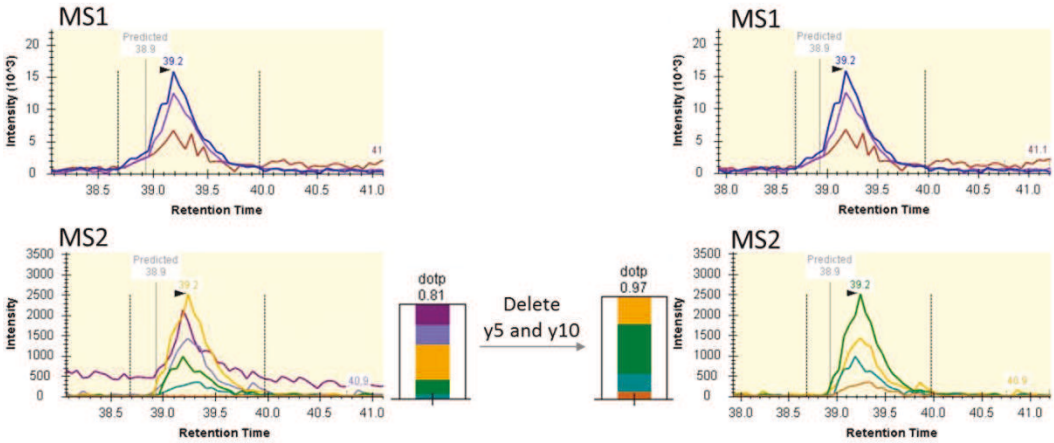



Figure IV-45: Manual refinement of DIA data.

**Post-translational modifications:** A challenge of DIA data is that since broad isolation windows are used, modified and unmodified peptides can be isolated and fragmented together if the modification does not produce a sufficient retention time shift to discriminate them. In this case the presence of the MS1 signal could be useful for the quantification. However the MS1 signal has a lower limit of detection than the MS2 signal, so this could not always be used.

If the modified and unmodified peptides are separated chromatographically, two peak groups having very similar fragmentation patterns are extracted and can produce errors of peak picking and identification, and thus bias the quantification.

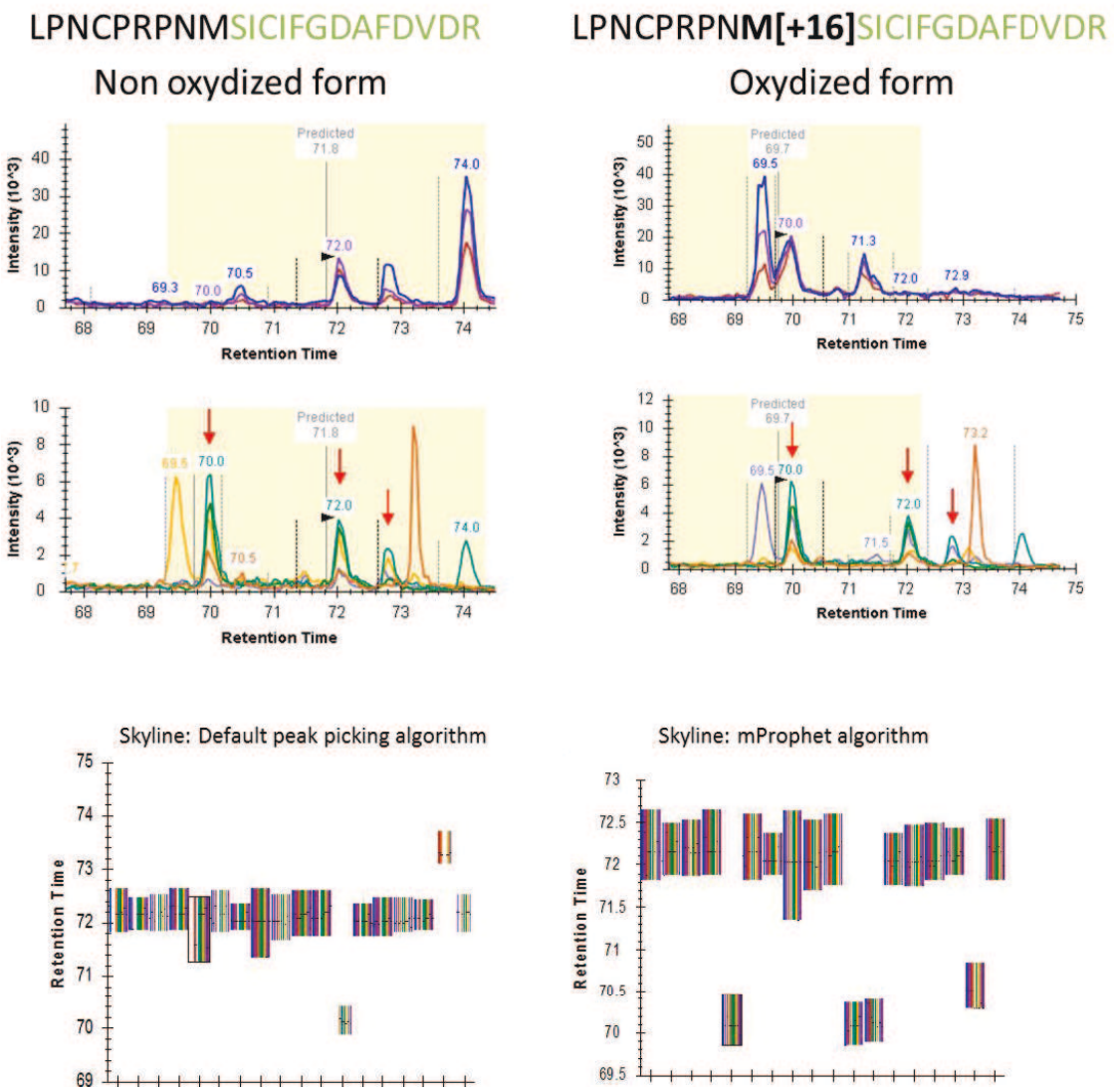



**Figure IV-46: Chemically modified peptides can be isolated in the same isolation window producing similar peak groups thus handicapping the peak picking.**

In Figure IV-46, the oxidized and un-oxidized versions of the LPNCPRPNMSICIFGDAFDVDR peptide are shown. The two versions of this peptide fall in the same isolation window since the triply charge precursors are separated by 5,3Da. In this case three peak groups are present in the MS2 signal. The peak at 70 minutes and the one at 72 minutes have exactly the same fragmentation pattern. In this case the two peaks can be discriminated using the MS1 signal. The peak the un-oxidized peptide and the oxidized peptides are respectively the peaks at 72 and 70 minutes. In this case, the prediction of the retention time can also help to discriminate each peptide. Other useful information to discriminate these two peptides is to use the signal of other charge states that could be analyzed in two distinctive isolation windows instead of one. This was not used here as the doubly charged precursors were too big to be selected by the instrument (>1250Da) and any other charge could be identified. More importantly, since in DIA all fragment ions are recorded, specific fragment ions for a given version of the peptide can be used to discriminate a modified peptide and even find a specific modification site. This has been widely used to quantify phosphopeptides [203, 211, 212].

After a visual inspection of these peaks the correct identification of both peptides can be made. The presence of the chemically modification producing similar peak groups handicaps the peak picking algorithm. In Figure IV-47, the integration boundaries for the un-oxidized version of the peptide are shown over 18 analyses. Two algorithms were used, the Skyline default peak picking algorithm and the implemented version of mProphet into Skyline. It can clearly

be seen that both algorithms still make mistakes, certainly due to the fact that retention times are predicted but they are not strictly aligned. And the identification of a peptide in a sample is not translated to the other samples but each analysis is independently treated. There is still room for progress to automate the analysis of DIA data.

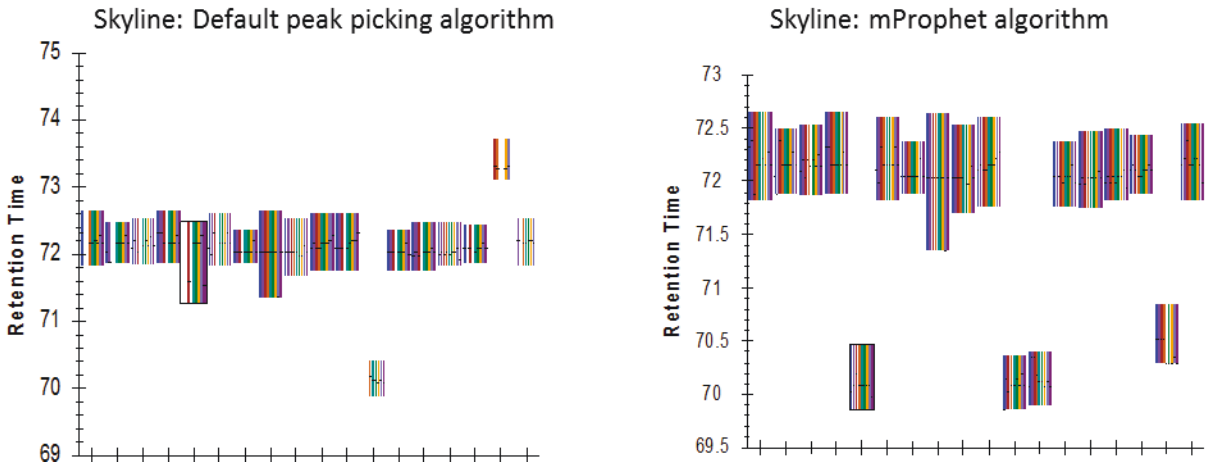



Figure IV-47: Peak boundaries chosen by two algorithms for a modified peptide.

### E.3.2 Evaluation of two peptide-centric software tools for DIA analysis

Two software tools were evaluated using the well-characterized standard sample described in Part IVChapter ID.1.1 on page 104. The dataset used here was the analysis of two samples composed of Universal Protein Standard (UPS1, Sigma) consisting of 48 purified human proteins spiked in a whole yeast proteome. Two concentration points were used for the evaluation: 5 and 25 fmol of UPS1 in 1µg of whole yeast lysate. Retention time standard peptides were also spiked in the samples.

To evaluate the different software tools the False Discovery Proportion (FDP) and the True Positive Rate (TPR) were used. The definition of these two metrics can be seen in Figure IV-29 on page 110. In the spectral library only 43 of the 48 UPS1 proteins could be identified. This means that the maximum value that can be obtained for the TPR is 89.6% (43/48). Additionally, a fold change of 5 is expected for UPS1 proteins. The deviation to this value was also used to evaluate the software tools.

A DDA analysis was performed in the same instrument to generate the spectral library. Then these analyses were validated at 1%FDR using the Proline software (Proline Studio, ProFI, Proteomics French Infrastructure) and a non-redundant spectral library was exported.

The first software tool to be evaluated was PeakView (AB Sciex). This is proprietary software from AB Sciex. PeakView extracts the target peptides present in the spectral library. To be able to control for which peptides a signal is extracted, the only possibility is to create a spectral library only containing validated and high quality spectra. This was done using Proline software. Then the user sets several parameters for the extraction: number of transitions, MS/MS tolerance, number of peptides per protein, extraction window and false discovery rate threshold. The software uses an algorithm close to mProphet [16]. The user then can define a list of peptides to be used as retention time calibration peptides. Then PeakView aligns the retention times of all the analyses to a chosen analysis that is used as a reference. Finally Peakview calculates a false discovery rate for each peptide and only uses peptides for which the FDR is lower than 1% for the protein quantification.

The results of the PeakView evaluation can be seen in Figure IV-48.A. A good discrimination of yeast proteins and UPS1 proteins was done. The UPS1 nicely align in the expected fold change of 5. The TPR was 85% and the FDP was

15%. These values were the reference to which Skyline will be compared. These were very satisfying results. However since PeakView is a proprietary software it cannot be used to analyze data from other vendors. Additionally the manual verification of the chromatographic traces is not easy and the integration boundaries are not shown and cannot be corrected if necessary.

The open-source software Skyline [15] was then evaluated. The workflow for this software consists in using a redundant spectral library in which each peptide's retention time will be normalized using a linear regression curve calibrated with RT standard peptides. The list of validated peptides from the DDA analysis by the Proline software was used as the list of targeted peptides. This list will be used to target the signal extraction.

For each DIA analysis, Skyline uses the retention times of the RT standard peptides to predict where the targeted peptides will elute (see the description of this approach in Part IVChapter IA.5.1 on page 72). The results of the evaluation for Skyline using its default peak picking algorithm can be seen in Figure III 46.B. The discrimination of UPS1 and yeast proteins could be done. However, compared to PeakView the UPS1 proteins do not align correctly around the expected fold change. The number of UPS1 where a variation in protein abundance that could be detected is lower (TPR= 81%) and the FDP is higher (FDP=19%).

To try to improve these results the implemented version of mProphet into Skyline was evaluated [16] (Figure III 46.C.). In this case, for each targeted peptide and for each analysis, all the peak groups found on the chromatogram are scored and the best one is chosen and integrated. This enabled to increase the TPR to 83% and reduce the FDP to 16%. However the UPS1 proteins were still not centered on the expected fold change.

For the last two previous evaluations, all peptides were used to quantify the proteins without eliminating any peptide. This is different than what Peakview does, as it eliminates all peptides below an FDR of 1%. To be certain that the quantification is not biased, the same list of peptides was chosen between Skyline and PeakView. Figure III 46.D. and E. shows the results of the evaluation using the same list of targeted peptides between the two software tools. This step did not drastically improve the TPR or FDP. However the UPS1 proteins seem to be closer to the expected fold change.

Figure IV-49 shows the results of PeakView and Skyline (using the Prophet algorithm) and both software were evaluated using the same peptides. The base-2 logarithm of the fold changes are plotted against the logarithm of the summed area under the peaks. It can clearly be seen that PeakView performs better at lower intensities as the UPS1 and the yeast proteins are close to the axis representing their expected fold change, respectively 2,3 and 0. It is only at lower protein abundances that the points start to spread out. In fact the majority of false positives (green triangles) are at low abundances. However, for Skyline the proteins start to spread out at higher protein abundances compared to PeakView, showing that the peak picking algorithm loses its performances at low concentrations.

As stated above the peak picking algorithms still make mistakes, certainly due to the fact that retention times are predicted but they are not strictly aligned. And the identification of a peptide in a sample is done independently from the other analysis. This phenomenon is more frequent in Skyline.

To illustrate that the major problem of the quantification is the lack of retention time alignment, the CVs and the spread of retention times (maximum minus minimum value) was calculated for all peptides in common between the two software tools. It is important to keep in mind that the signals are extracted only in a small window of time (6 minutes), the CVs and spreads will thus not have very large values but these have to be compared to the elution time

of a chromatographic peak which is between 15 to 30s. The retention time CVs and spread show larger values for Skyline demonstrating that the Peak picking is different between the two software tools.

In conclusion there is still room for progress to automate the analysis of DIA data. A strict retention time alignment seems to be the best option to avoid errors in the quantification. The matching of peptides between different runs can be an option to improve this. Future developments to improve automatic DIA analysis are ongoing.

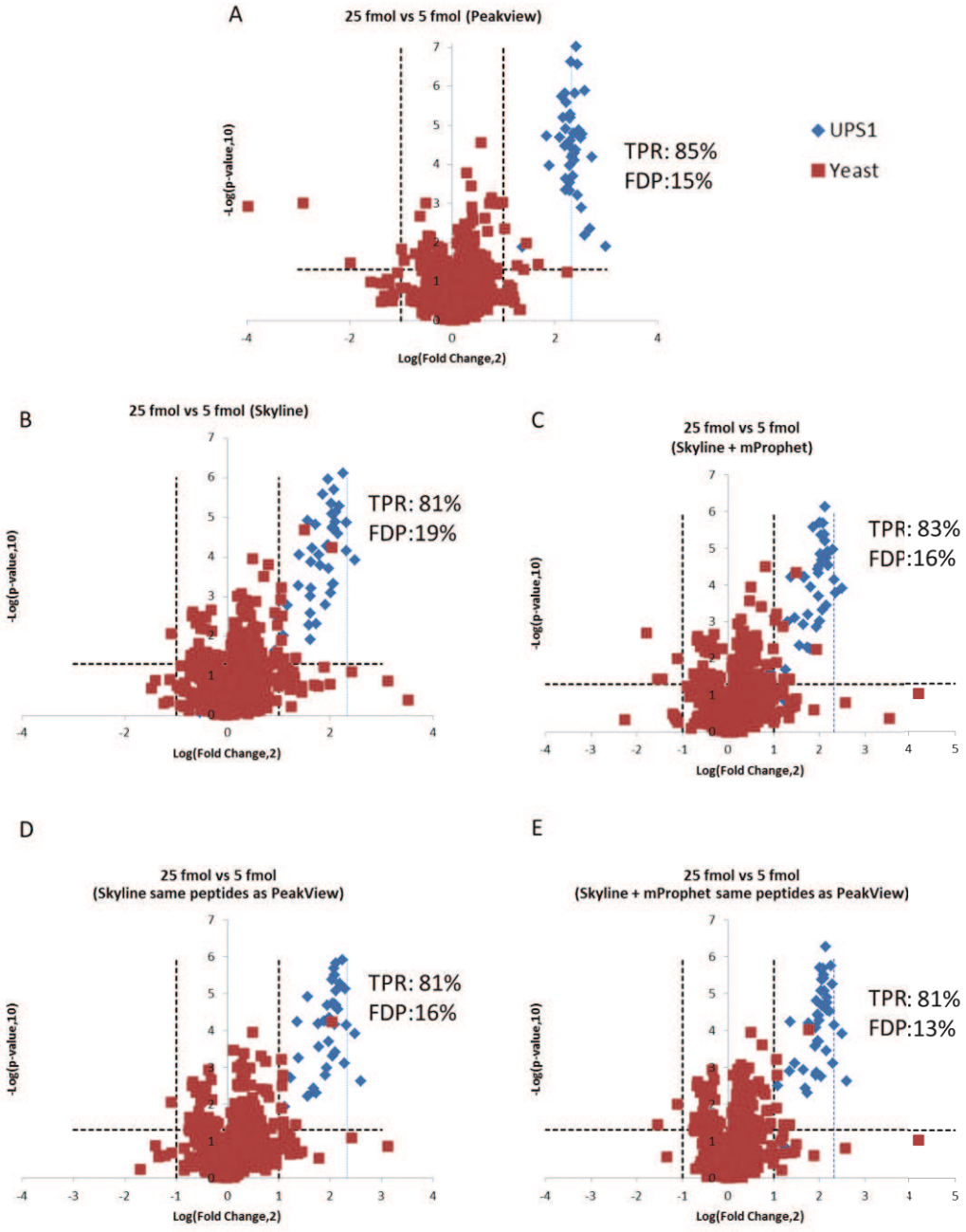



**Figure IV-48: Evaluation of two peptide-centric DIA data analysis software tools.**

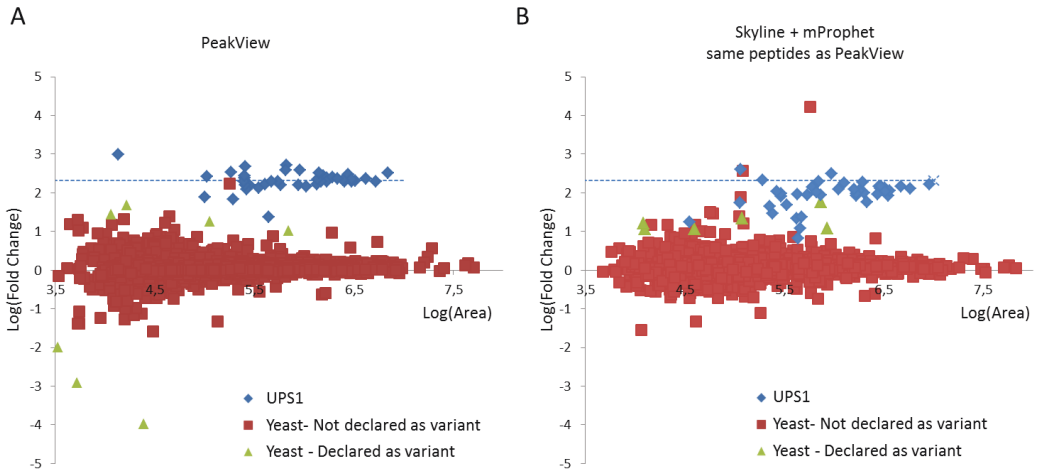



Figure IV-49: Comparison of PeakView and Skyline

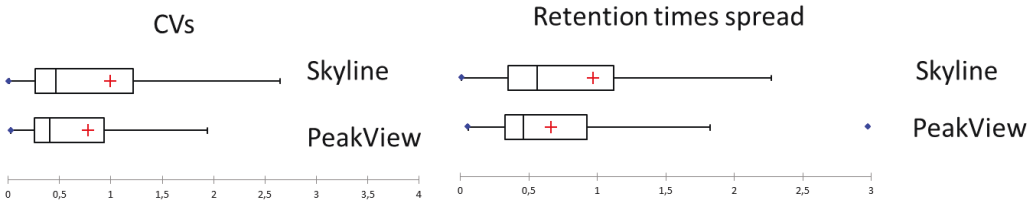



Figure IV-50: Retention time CVs and spread for the same peptides quantified by Skyline and PeakView.

### E.3.3 Spectrum-centric approaches

Spectrum-centric approaches have developed with the appearance of DIA approaches. In these approaches co-eluting fragment and precursor ions are clustered together to generate pseudo DDA spectra. These spectra are then used to perform peptide identification by standard database search algorithms. This approach was first introduced by Purvine *et al*. to reconstruct DDA-like spectra from low and high voltage analysis by manually identifying precursors and fragment ions with similar chromatographic characteristics [99]. This technique was refined and commercialized by Waters under the name $MS^E$ [15]. In theory the $MS^E$ approach could be used in any type of high-resolution instrument. However the software tool capable of treating this type of data (ProteinLynx Global SERVER™, PLGS™) is a proprietary software tool. This added to the complexity of the MS/MS spectra have slowed the progress of this technology.

With the emergence of new DIA methods several bioinformatic research groups have developed new spectrum-centric data analysis algorithms. An example is DIA Umpire, which generates DDA-like spectra from SWATH-like DIA data [213]. This tool enables the identification and quantification of proteins directly from DIA data. However for the moment, it generates an excessive amount of spectra which have a negative impact on the total protein identification and on the time necessary to analyze a single run.

These approaches are still new in the field but they are very promising as they have the advantage of profiting from the discoveries and advances of standard workflows that have been developed in the last 20 years in proteomic analysis.

## *E.4. Conclusion and perspectives*

In conclusion, DIA is a very promising acquisition mode for protein quantification. The method has taken advantage of the recent technological advances that provided the scan rate, sensibility and mass-accuracy necessary for this type of approach.

Due the nature of DIA, data analysis is challenging. The complexity of the MS/SM spectra and the high number of peptides to be quantified (tens of thousands) made the visual and manual validation of the data very difficult and impractical. Automatic validation pipelines must be used but at the moment there is still room for progress.

This is why to ensure an accurate and reliable quantification, the best strategy is to have a defined hypothesis before analyzing the data. This way the list of peptides that have to be analyzed can be reduced. Once this first set of peptides has been quantified, the results can guide the user to reestablish his hypothesis and expand the list of peptides accordingly, and so on. A reduction of protein targets that represent biological processes has been proposed by several groups [212, 214, 215].

Finally, both peptide-centric and spectrum-centric approaches strongly rely on protein databases. A challenge in DIA will certainly be in the future the quantification of protein sequences not present in the consensus databases (sequence variant peptides, isoforms…). In the context of personalized medicine and the developments of Proteogenomic approaches, DIA is likely a very promising type of quantification as novel peptides could be queried in the already acquired data without the need of developing targeted approaches.

# Chapter II Application of targeted proteomics to validate Crohn's disease biomarkers

## *A.1. Crohn's disease*

Crohn's disease (CD) is a type of inflammatory bowel disease (IBD) characterized by chronic and relapsing inflammation of intestinal segments and potentially be accompanied by extra-intestinal manifestations [17]. Crohn's disease affects about 0,32% of people in Europe and North America. It is more common in the developed countries and less common in Asia and Africa [18].

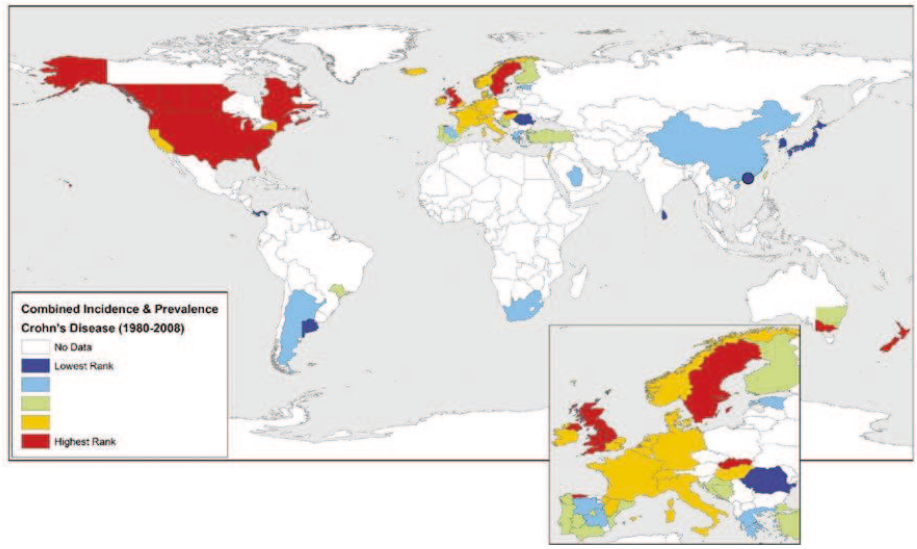



**Figure IV-51: Worldwide CD incidence rates and prevalence for countries reporting data after 1980 (adapted from [18])**
Incidence and prevalence values were ranked into quintiles representing low (dark and light blue) to intermediate (green) to high (yellow and red) occurrence of disease.

CD is caused by a combination of several contributing factors, including heritable traits, environmental cues, abnormalities in intestinal mucosal barrier integrity and function, immune regulation, and gut microbiota. Exaggerated immune responses are presumably directed against normal commensal enteric bacteria in genetically susceptible hosts. Host genetic susceptibility may be related to defective mucosal barrier function and/or bacterial killing, leading to an overexposure to luminal antigens and to inadequate immunoregulation, resulting in abnormal responses and tissue damage [17].

There is no cure for Crohn's disease. Medications and surgery are used to ease symptoms, maintain remission, and prevent relapse [19]. Diagnosing CD is very difficult as there is no specific symptom for the disease and its manifestations are common with others pathologies such as gastroenteritis, ulcerative colitis and irritable bowel syndrome. It is very important to differentiate all these pathologies as they require different therapeutic handling. The diagnosis of CD is based on a body of clinical arguments that takes time to generate. The average time to make the correct diagnosis for CD is 2,6 years. For the moment no molecular maker specific to Crohn's disease has reached clinical use.

## *A.2. The Human Gut Microbiome*

The human gut microbiota is the community of microorganisms present in the human gut. The latest estimation of the total amount of microbial cells in the human gut is 1 to 2 kg and the number of microbial cells is about the same as the total number of human cells in the human body [216]. These microbes play a crucial role in human life for example in nutrition, immune system and protection against pathogens. They thus have a direct or indirect effect on human

health. Several studies have shown that imbalances in the composition of the microbiota are linked to diseases such as obesity, IDB, anorexia, cancer, obesity and autism [217].

The analysis of the gut microbiota is challenging. Early approaches aimed at cultivating bacteria but it is estimated that up to 80% of the bacterial species cannot be cultivated by conventional techniques. This is due to the fact that bacteria in the gut need an anaerobic environment, specific nutrients and they have a codependency on one another. This is why these techniques have now been taken over by molecular techniques such as targeted sequencing of the 16S Ribosomal rRNA-encoding gene and metagenomic sequencing of the whole microbial DNA [218].

Technological advancements in DNA sequencing have enabled the expansion of knowledge through large-scale metagenomic studies of the human gut microbiome, such as the European consortium Metagenomics of the Human Intestinal Tract (MetaHit) [219] and the Human Microbiome Project [220]. In 2014, an integrated and expanded gene catalog of this metagenomic data was published [20]. The human gut microbiome is an organ of high complexity as this study showed the existence of more than 9,8 million different genes. This is more than 445 times more genes than the human genome. However, this number is very large and mostly composed of rare genes (Figure IV-52).

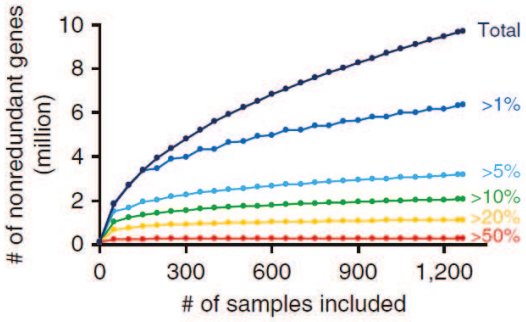



**Figure IV-52: Number of non-redundant genes against the number of samples analyzed (Adapted from [20]).**
Rare genes, those present in less than 1% of samples, constitute the majority of sequenced genes. The most common genes, those present in more than 50% of samples, are around 300,000.

The gut microbiome was found to show a significant diversity between healthy individuals. Only around 300,000 genes have been found to be common bacterial genes present in 50% of all sequenced individuals [20, 219] (Figure IV-52). Figure IV-53 shows the results of the analysis of the microbiome extracted from stool samples. The microbial taxa varies significantly from individual to individual and is mostly constituted by Bacteroidetes and Firmicutes [220]. However the metabolic function of the microbial communities is very stable from individual to individual.

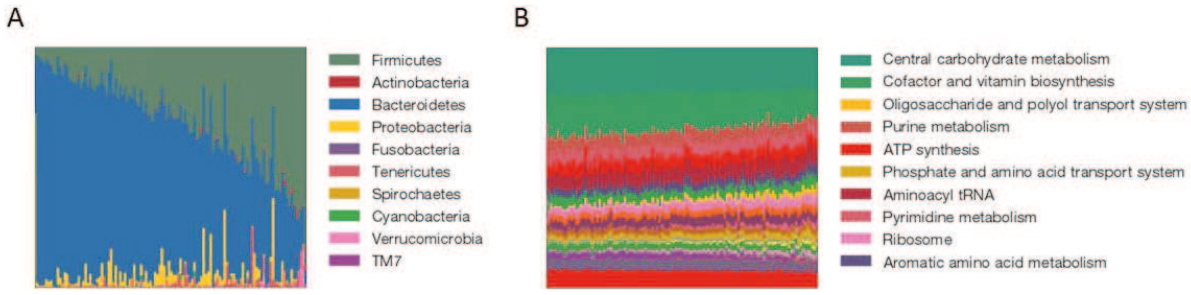



**Figure IV-53 : Microbial taxa composition and metabolic pathways on different stool samples (Adapted from [220]).**
Vertical bars represent microbiome extracted from stool samples. The microbial phyla (A) and the metabolic pathways (B) are shown. The human gut microbiome is highly diverse from individual to individual and is mostly constituted by Bacteroidetes and Firmicutes. The metabolic pathways are very stable from individual to individual.

### *A.3. Context of the project*

This project was carried out in collaboration with the Micalis laboratory of the National Institute of Agricultural Research (INRA), and particularly with Drs. Catherine Juste and Joël Doré.

Beyond functional metagenomic analysis, our collaborators were interested in studying the microbial proteomes of patients with Crohn's disease. They have therefore developed a workflow for microbial extraction and discovery metaproteomic analysis using 2D-DIGE and LC-MS/MS.

The study included six patients with Crohn's disease (CD) and six healthy controls (HC). A conscious selection of the patients was made to avoid confounding errors. The CD and HC patients matched for sex, age and tobacco use. The HC patients did not have any symptoms or family history of gastrointestinal disease and were not medicated. The participants provided fresh stool samples that were collected in anaerobic conditions. Bacterial fractions were separated from fecal matrix using a density gradient (nicodenz based) at low temperature and in anaerobic conditions.

The discovery metaproteomic analysis using 2D-DIGE and LC-MS/MS conducted by our collaborators allowed the identification of 59 gel spots found to have a significant expression change between CD and HC proteomes (30 increased and 29 decreased) (Figure IV-54). From this list of biomarker candidates, we selected a subset of proteins to be further validated using an LC-SRM approach (candidate proteins highlighted in yellow in Figure IV-54).

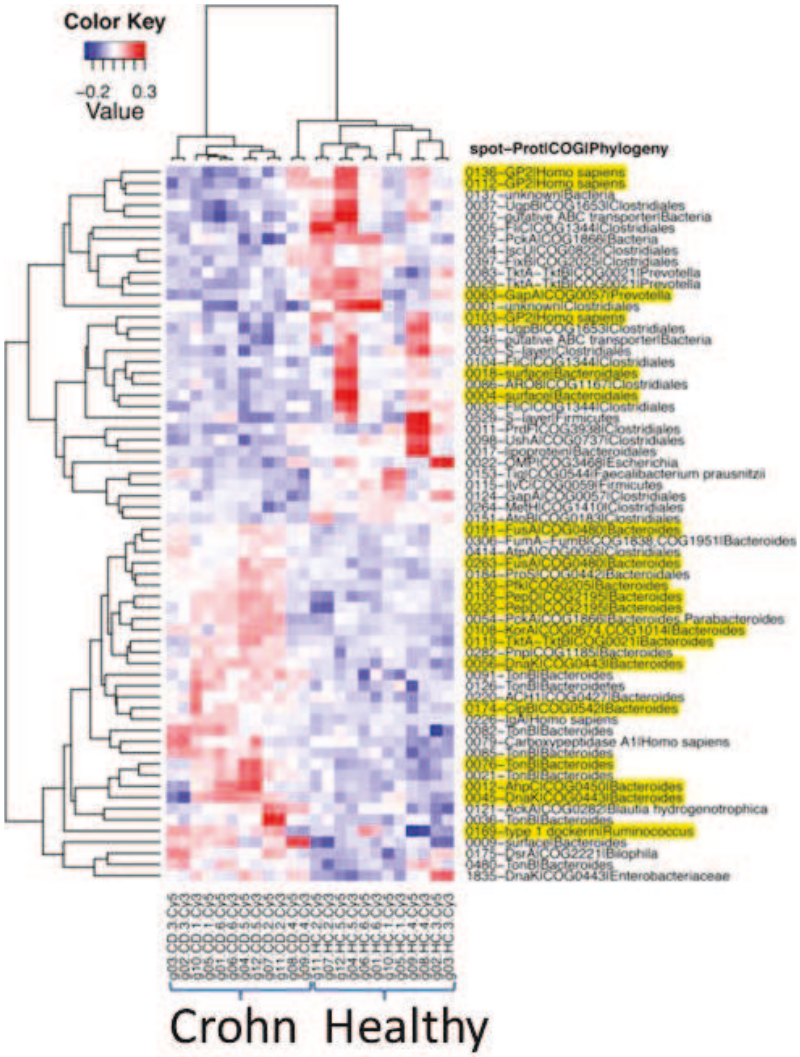



**Figure IV-54: Cluster heat map of 2D-DIGE gel spots with significantly different intensities between HC and CD patients.**
Proteins highlighted in yellow were chosen for further validation by LC-SRM.

### *A.4. LC-SRM method development*

The development of the LC-SRM method for microbial proteins was challenging. As detailed above, the microbiome is a sample of extreme complexity. Furthermore there is large microbial composition diversity between individuals.

These factors complicated the task of creating an LC-SRM assay targeting microbial proteins. Figure IV-55 summarizes the analytical workflow we have setup for this project.

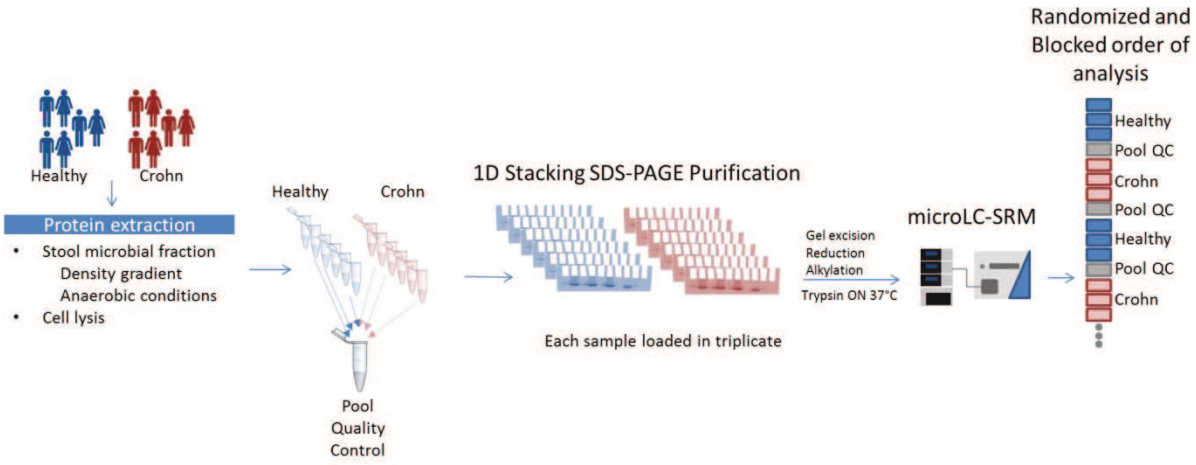



**Figure IV-55: LC-SRM analytical workflow for the validation of microbial proteins.**

Here is a description of most crucial steps of the method development:

***Choice of targeted proteins and signature peptides:*** With the close expertise of our collaborator, a list of 21 microbial proteins was chosen to be further validated by LC-SRM. Among all candidate proteins, using the METAHIT database [219] we chose those having specific and unique peptides to a protein or a group of proteins with the same function from phylogenetically close bacterial strains. In this extremely complex sample it is very difficult (almost impossible) to find truly proteotypic peptides. However, the importance here is to relatively quantify proteins or a group of proteins having the same biological function. To choose the peptides, priority was given to peptides that had already been identified in previous shotgun experiments acquired on equivalent samples, preferentially without fractionation, and showing high-quality MS/MS spectra. Chosen peptides were 7 to 25 amino acids long, contained no miscleavage and no methionine in their sequences.

***Choice of sample preparation protocol:*** The LC-SRM validation was performed on aliquots of microbial proteins prepared in Laemmli buffer for the 2D-DIGE experiment. This restrained the choice of the methodology to be employed and we therefore chose a SDS-PAGE Stacking gel preparation protocol. This is compatible with the Laemmli buffer and allows further quantification without sample fractionation. The development of this unfractionated sample preparation protocol was described above in Part IVChapter IB.2 on page 93.

***Choice of the best transitions and heavy labelled standard peptide mixture:*** In order to determine the best transitions for each peptide, four randomly chosen protein samples were prepared and pooled together, heavy labelled standard peptides were spiked in the samples and injected using the same microLC-SRM system used for sample analysis (See Experimental Section C.1 on page 219). At least 6 transitions (including y- and b-type ions) were monitored for each peptide in an unscheduled method. This allowed determining the retention times of all targeted peptides, verifying endogenous and isotopically-labelled peptides co-elution, eliminating interfered transitions and adjusting the isotopically-labelled peptides concentrations. A concentration-balanced mixture of crude heavy-labelled peptides was prepared in order to obtain comparable signal-intensities to the endogeneous peptides (the peptides were split into 4 groups defined by signal intensities and diluted 2400, 1200, 600 or 300 times from the purchased stock solutions).

***MicroLC-SRM Analyses:*** A Dionex Ultimate 3000 system coupled to a TSQ Vantage Triple Quadrupole instrument (Thermo Fischer Scientific, San Jose, CA, USA) was used for the quantification. A carry over effect was detected at

early stages of the development. A two-step LC gradient was developed. The first step was for peptide separation and LC-SRM analysis and the second step was a rapid washing step to elute potential remaining peptides on the column. Using this two-step gradient we eliminated the carry-over effect and we made sure this phenomenon would not compromise the quantification (Figure IV-56). This also reduced the total amount of instrument time needed as fewer blank runs were needed. The reproducibility of the retention times was good and showed that this two-step LC-gradient did not affect the reproducibility of the elution. A detailed description of the LC-SRM parameters is provided in page 219.

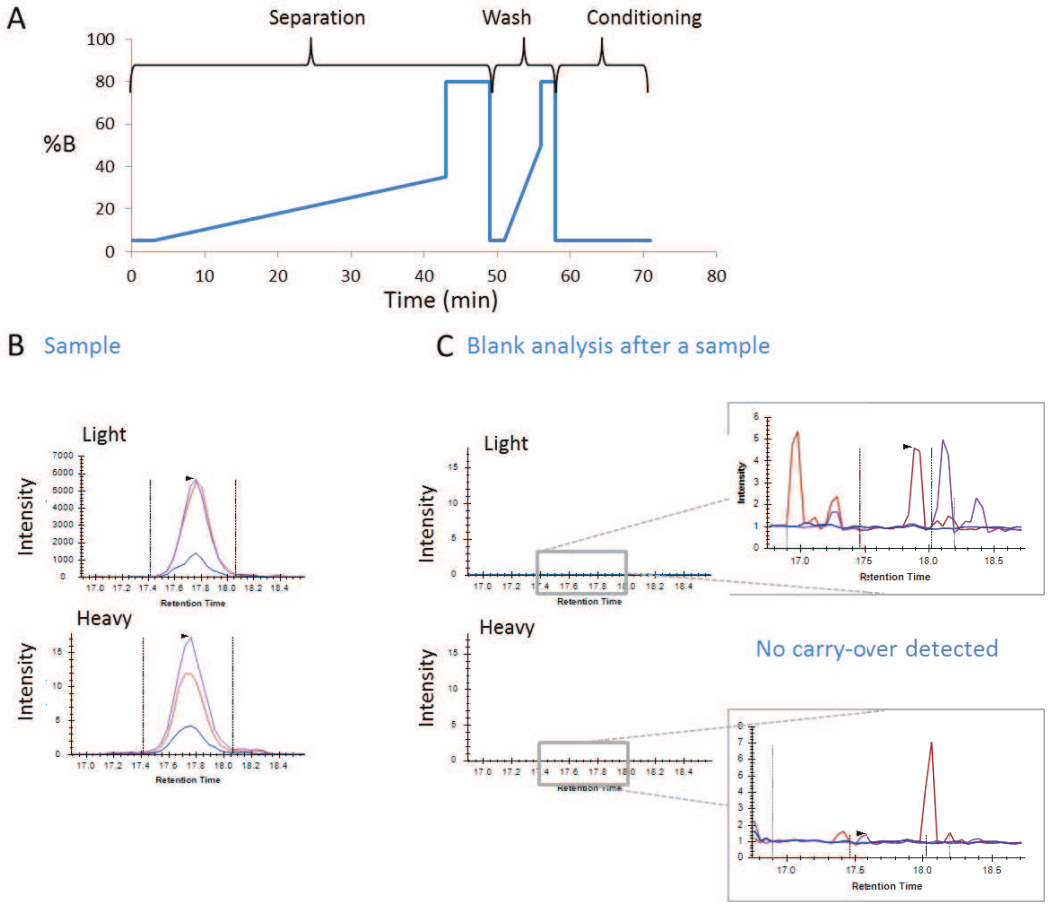



**Figure IV-56: Elimination of the carry-over effect.**
A. A two-step LC gradient was developed. The first step was for peptide separation and LC-SRM analysis and the second step was a rapid washing step to elute potential remaining peptides on the column. B. Example of SRM traces of the endogenous (Light) and the isotopically stable heavy-labelled standard (Heavy) counterpart of the same peptide. C. SRM traces of the blank injection just after the analysis of a sample. Using the two-step LC gradient allowed eliminating the carry-over effect.

***Quality control samples and metrics:*** To assess the LC-SRM system stability and performance over the course of the experiment a quality control sample was created. This quality control was created by pooling together randomly chosen samples and preparing them in the same way as the samples (Figure IV-55). The quality control pool was analyzed ten times over the whole course of the experiment. For each transition, coefficients of variation were calculated for light/heavy area ratios obtained during the ten repeated injections, and we set the acceptance level for coefficients of variation below 20%.

In order to limit the impact of confounding errors we used the blocking and randomization method [183]. The order of sample injection was randomized within the CD and the HC conditions and then blocks by groups of three samples of

the same condition. In between each group a quality control pool sample was introduced. The injection scheme can be seen in Figure IV-55.

***Data Analysis:*** The Skyline open-source software package [15] was used to visualize the SRM data, perform peak picking and integration of transition peak areas. Protein relative quantification and testing for differential protein expression were performed using the R package MSstats [184, 221]. The acceptance criteria for statistically different protein abundance changes between controls and CD patients were set at a p-value lower than 0.05 and a fold change higher than 2.

### *A.5. Results*

Using the quality control pool sample we assessed the overall performance of the LC-SRM system by monitoring the overall behavior of our targeted peptides in the real analysis conditions and within the real analysis matrix. Because the quality control pool sample was made by randomly pooling CD and HC samples, some endogenous peptides were not present in sufficient amount to be detected in the pool. However, this quality control sample was very useful to assess the stability of the overall LC-SRM system. Over the course of the analysis some peptides were found to be very stable overtime (Figure IV-57). Others were found to be unstable maybe due to degradation or coating to the vial walls. The peptides having this latter behavior were not considered for further quantification. We set the acceptance level for coefficients of variation below 20%. All transitions met this criterion, except for those present in low amounts in the quality control pool, thus proving that the LC-SRM system was stable and performing well over the course of the experiment.

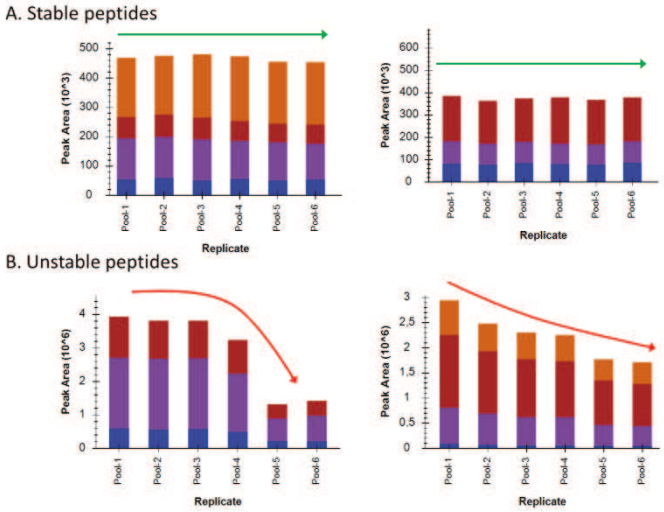



**Figure IV-57: Stable and unstable peptides over the course of the experiment.**
The measured intensity of the heavy-labelled standard peptide is shown for each peptide. Over the course of 6 analyses of the quality control sample, some peptides were found to be stable (A) and others were unstable (B) maybe due to degradation or coating to the vial walls.

The overall reproducibility of the experiment was verified by calculating light/heavy area ratios for each transition, and verifying that coefficients of variation were lower than 20% for triplicate injections. An example of this can be seen in Figure IV-58. In general, all transitions used for quantification in the sample cohort showed coefficients of variation lower than 20% with a mean value of 9.4% and a median value of 8.5%.

Thirteen microbial proteins could be quantified and the trends of under- and overrepresentation in CD patients observed in the discovery 2D-DIGE experiment were confirmed (Figure IV-59).

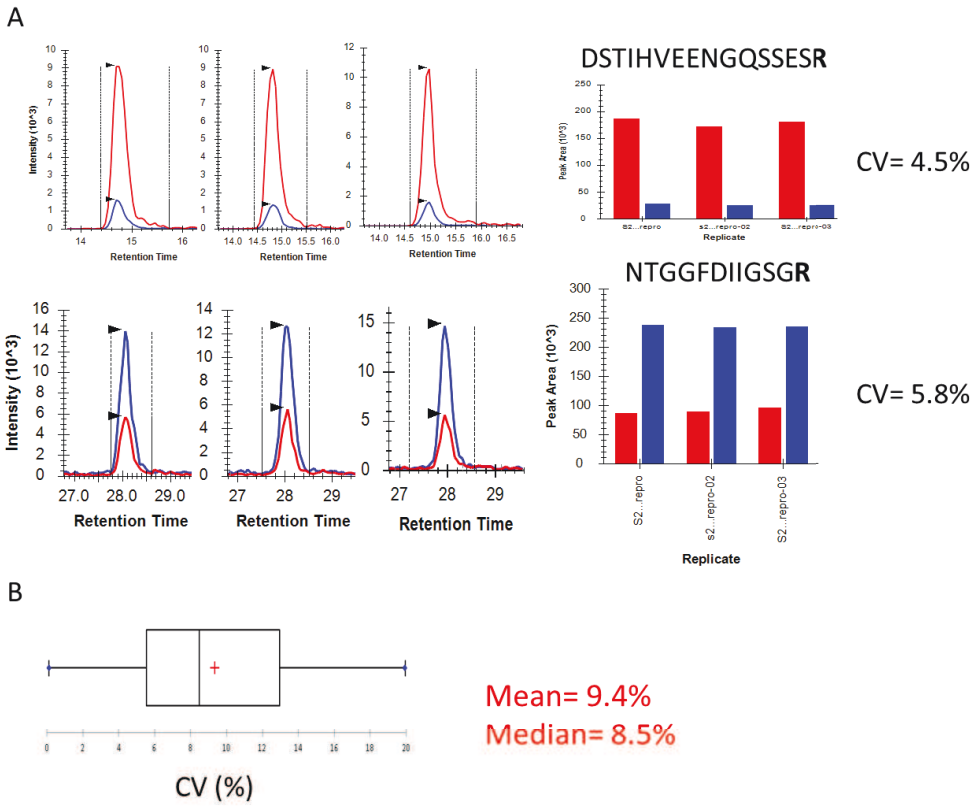



**Figure IV-58: Assessment of the precision of the developed LC-SRM method.**
A. Example of reproducible LC-SRM traces of two peptides. The signals from endogenous and heavy-labelled peptides are respectively in red and in blue. B. Boxplot representing the coefficient of variation of the whole sample cohort.

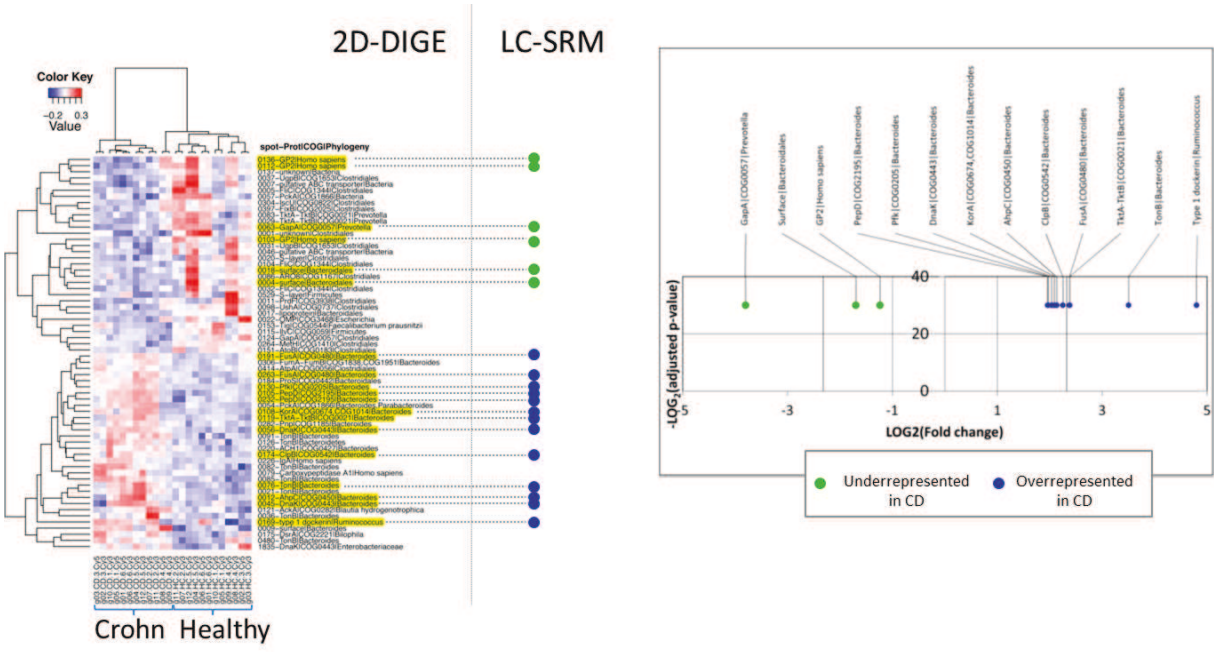



**Figure IV-59: Results of the validation of 13 microbial proteins related to Crohn's disease by LC-SRM.**
The trends observed in the 2D-DIGE experiment were validated by the LC-SRM experiment.

### *A.6. Conclusion and Perspectives*

The results presented here were published in 2014 in the journal *Gut* (available on page 157) and a patent for diagnostic biomarkers for Crohn's disease was filed in December 2014 and extended internationally (PCT) in December 2015 (N°FR1462867 «MARQUEURS DIAGNOSTIQUES DE LA MALADIE DE CROHN»). These biomarkers could help to reliably diagnose Crohn's disease, discriminate between similar intestinal pathologies and assess the therapeutic efficiency of treatments directed against CD.

The biomarkers presented in this work were analyzed in fresh stool samples. In order to carry out a large-scale study of these candidates, the sample preparation step should ideally be simplified. Further studies are on their way to determine whether or not the biomarker candidates can be detected in frozen stool samples. This would alleviate the organization of large–scale studies. Similarly, we are evaluating the feasibility of simplifying the sample preparation protocol by detecting our biomarkers in the fecal water (the liquid extracted after centrifugation of stool samples). This idea originated from the fact that some of the targeted proteins are secreted by the microbial cells. This would considerably simplify the sample preparation protocol. Though, the main drawback of this approach is that the targeted proteins are no longer enriched, as the microbial secreted proteins/peptides are no longer isolated from other components of the stool. Human or food-derived proteins can now become possible interferences. The impact of this will be evaluated in future experiments.

### A.7. Publication

Gut Online First, published on January 16, 2014 as 10.1136/gutjnl-2012-303786


OPEN ACCESS

ORIGINAL ARTICLE

# Bacterial protein signals are associated with Crohn's disease

Catherine Juste,[1] David P Kreil,[2,3] Christian Beauvallet,[4] Alain Guillot,[5] Sebastian Vaca,[6] Christine Carapito,[6] Stanislas Mondot,[1] Peter Sykacek,[2] Harry Sokol,[1,7] Florence Blon,[1] Pascale Lepercq,[1] Florence Levenez,[1] Benoît Valot,[5] Wilfrid Carré,[8] Valentin Loux,[8] Nicolas Pons,[1] Olivier David,[9] Brigitte Schaeffer,[9] Patricia Lepage,[1] Patrice Martin,[4] Véronique Monnet,[1] Philippe Seksik,[7] Laurent Beaugerie,[7] S Dusko Ehrlich,[1] Jean-François Gibrat,[8] Alain Van Dorsselaer,[6] Joël Doré[1]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2012-303786).

For numbered affiliations see end of article.

**Correspondence to**
Dr Catherine Juste, Bâtiment 405, INRA Domaine de Vilvert, Jouy-en-Josas 78350, France; catherine.juste@jouy.inra.fr

Received 19 September 2012
Revised 16 December 2013
Accepted 17 December 2013

## ABSTRACT

**Objective** No Crohn's disease (CD) molecular maker has advanced to clinical use, and independent lines of evidence support a central role of the gut microbial community in CD. Here we explore the feasibility of extracting bacterial protein signals relevant to CD, by interrogating myriads of intestinal bacterial proteomes from a small number of patients and healthy controls.

**Design** We first developed and validated a workflow—including extraction of microbial communities, two-dimensional difference gel electrophoresis (2D-DIGE), and LC-MS/MS—to discover protein signals from CD-associated gut microbial communities. Then we used selected reaction monitoring (SRM) to confirm a set of candidates. In parallel, we used 16S rRNA gene sequencing for an integrated analysis of gut ecosystem structure and functions.

**Results** Our 2D-DIGE-based discovery approach revealed an imbalance of intestinal bacterial functions in CD. Many proteins, largely derived from *Bacteroides* species, were over-represented, while under-represented proteins were mostly from Firmicutes and some *Prevotella* members. Most overabundant proteins could be confirmed using SRM. They correspond to functions allowing opportunistic pathogens to colonise the mucus layers, breach the host barriers and invade the mucosae, which could still be aggravated by decreased host-derived pancreatic zymogen granule membrane protein GP2 in CD patients. Moreover, although the abundance of most protein groups reflected that of related bacterial populations, we found a specific independent regulation of bacteria-derived cell envelope proteins.

**Conclusions** This study provides the first evidence that quantifiable bacterial protein signals are associated with CD, which can have a profound impact on future molecular diagnosis.

### Significance of this study

**What is already known on this subject?**

- There are unmet needs for diagnosis, treatment and patient monitoring in Crohn's disease (CD).
- No molecular marker has yet advanced to clinical use in CD.
- The intestinal microbiota is recognised as an essential contributor to disease initiation and perpetuation and, therefore, represents an enormous reservoir for the discovery of novel signatures that could be used as biomarkers and predictors for different disease phenotypes or stages.

**What are the new findings?**

- The feasibility of extracting bacterial protein signals relevant to CD by interrogating myriads of intestinal bacteria, even from a small number of subjects.
- Twelve bacterial protein signals and one human protein signal (glycoprotein 2 of zymogen granule membranes, GP2) were robustly quantified by targeted MS-based proteomics, without the need for antibodies and ELISA testing. All of them make sense in the context of our understanding of CD.
- Increased IgA at the surface of microbial cells of CD patients coincides with the over-representation of various bacterial proteins with a high immunogenic potential in CD patients.
- Decreased GP2 at the surface of microbial cells of CD patients may favour adhesion of bacteria to the mucosa and then promote inflammation.

**How might it impact on clinical practice in the foreseeable future?**

- Using meta-proteome-wide association studies, we point out new potential biomarkers in CD.

## INTRODUCTION

Independent lines of evidence converge to suggest a central role of the gut microbial community in Crohn's disease (CD): microbiota is required for the development of inflammation in genetically predisposed colitis animal models,[1] reinfusion of luminal contents after ileal resection rapidly produces recurrent disease in CD patients,[2] antibiotics delay postoperative recurrence of CD,[3] and the hitherto identified susceptibility polymorphisms contribute or relate to bacterial sensing through innate and adaptive immune pathways.[4–6] Finally, a vicious

**To cite:** Juste C, Kreil DP, Beauvallet C, *et al*. *Gut* Published Online First: [*please include* Day Month Year] doi:10.1136/gutjnl-2012-303786

circle of 'mutualism breakdown' has been postulated, where the host does not tolerate its own microbiota any longer, and inflammation can favour the selection of aggressive symbionts.[7] However, the micro-organisms, or microbial products that signal the disruption of gut homeostasis and may have a critical role in CD, are unknown. Their identification remains a significant challenge due to the high complexity of the intestinal microbiota, forming the most densely populated microbial community in the body. It is composed of $10^{13}$–$10^{14}$ micro-organisms belonging to about a thousand different species, most of them anaerobic,[8] and is hitherto largely uncultivable (only 20–30% of enteric bacterial species have been propagated in pure culture). Our group[9–11] and others[12–15] have used diverse culture-independent approaches based on molecular profiling of 16S rRNA genes to compare the microbial diversity of intestinal or faecal samples from CD patients and healthy people. Despite the biases inherent to different methodological approaches, varying sampling sites (faeces or mucosa along the intestine), heterogeneity in clinical phenotypes, and variable statistical power, the overall consensus is that the diversity of dominant bacterial species is reduced in CD, notably among members of the Firmicutes phylum.

Largescale metagenomic sequencing (as in MetaHit and the Human Microbiome Project), which analyses whole genomic DNA directly extracted from human intestinal communities, offers a new dimension for the characterisation of these communities from a functional point of view, and represents an enormous reservoir for the discovery of novel signatures that could be used as biomarkers and predictors for different disease phenotypes or stages.[16–18] Beyond functional metagenomics, metaproteomic studies will reveal the true expression of metabolic and cellular functions that govern physiology, become disrupted in disease, and can have a profound impact on molecular diagnosis. Therefore, while there are unmet needs for diagnosis, treatment and patient monitoring in CD, especially while no molecular maker has advanced to clinical use in CD,[19] and considering that the intestinal microbiota is recognised as an essential contributor to disease initiation and perpetuation, we here demonstrate the feasibility of discovering and validating a range of CD-associated bacterial proteins by using a proteomic approach from discovery to validation. With protein profiling providing assays closer to activated functions, such metaproteome-wide association studies have the potential to become an important tool in modern medicine, and could answer major yet unmet clinical needs.

## METHODS

### Subjects and samples

We conducted a cross-sectional study including six patients with CD (four women and two men, aged 26 through 41 years) and six healthy controls (HC) matched for age, sex and tobacco use (table 1). Patients were followed and selected in the gastroenterology unit of the Saint-Antoine Hospital (Paris). We made a conscious selection of different phenotypes to avoid an unnaturally uniform patient population for this pilot study. Exclusion criteria, however, were active disease with a Harvey–Bradshaw score >5, and any use of antibiotics within the preceding 2 months. The control group comprised healthy volunteers with neither symptoms nor a family history of gastrointestinal disease, and with no use of medication. All participants gave informed consent to the protocol that was approved by the ethics committee of the hospital.

Every participant was asked to provide a fresh stool sample collected at home in a Stomacher 400 plastic bag (Seward Medical), which was left open in a one-litre hermetic plastic box containing a catalyst (Anaerocult, Merck, Darmstadt, Germany) to generate anaerobic conditions. This faecal material was maintained in a coolbox and transferred within 2 h into an anaerobic chamber (90% $N_2$, 5% $H_2$ and 5% $CO_2$) for processing. We had verified in preliminary assays, that measurements at a single time point gave a reliable picture of individual metaproteomes, which showed little variation over time (see online supplementary figure S1).

### Preparation of bacterial fractions and diversity profiling

Given the high complexity of faecal samples that contain bacterial, dietary and host proteins, we first extracted bacterial communities, to focus on the collected bacterial proteomes. Bacterial fractions were extracted in duplicate, at low temperature and in an anaerobic atmosphere (see online supplementary method 1, supplementary figure S2), from freshly collected stool specimens. The final bacterial pellets, as well as 150 mg stool aliquots, were kept at −80°C until further analyses. Diversity profiling was performed by 16S rRNA gene pyrosequencing (see online supplementary method 2).

**Table 1** Gender and age of matched participants and clinical characteristics of Crohn's disease patients at the time of stool collection

| Controls | Patients | | | | | | Common to controls and patients |
|---|---|---|---|---|---|---|---|
| Designation Gender/age | Designation Gender/age | Disease location | Disease activity | Diagnosis year | Medication | Surgery | Smoking |
| HC.1 F/39 years | CD.1 F/41 years | L1 | 4 | 1994 | Azathioprine | Small bowel | No |
| HC.2 M/36 years | CD.2 M/37 years | L2 | 2 | 1987 | Azathioprine+Prednisone | Small bowel | Yes |
| HC.3 F/26 years | CD.3 F/29 years | L1+L4 | 5 | 1998 | Azathioprine | No | Yes |
| HC.4 M/27 years | CD.4 M/26 years | L1 | 5 | 2006 | Budesonide | No | No |
| HC.5 F/41 years | CD.5 F/41 years | L3 | 2 | 1990 | Azathioprine | Subtotal colectomy | No |
| HC.6 F/36 years | CD.6 F/38 years | L1 | 5 | 2005 | Mesalazine | No | No |

Disease location according to the Montreal classification: L1 ileum; L2 colon; L3 ileocolon; L4 upper gastrointestinal tract.
Disease activity according to the Harvey–Bradshaw index.[20]
CD, Crohn's disease; HC, healthy controls.

### Discovery of CD-associated gut microbial proteins using 2D-DIGE/LC-MSMS

We used two-dimensional differential gel electrophoresis (2D-DIGE), coupled with tandem mass spectrometry (MS/MS) and searches against metagenomic databases (MetaHit) as a non-targeted comprehensive approach to discovering CD-associated proteins. Briefly, microbial fractions were extracted in duplicate for the 12 participants leading to 24 samples, which were analysed in a dye-swap design comprising 12 gels in total (see online supplementary table S1 and supplementary method 3). For differential expression analysis, we applied two complementary methods, both established and commonly used in microarray gene expression analysis: a hierarchical analysis of variance (ANOVA) (false discovery rate (FDR) <10%) and an empirical Bayes moderated single-group t test *per* gene (FDR <10%). They represent different approaches to the challenge of comparing small sets of samples for thousands of variables (see online supplementary method 4).[21 22] The complementary candidate lists were combined to yield a set of protein spots identified by at least one of the methods as showing significant differences between CD and HC.

For protein identification, nine gels (see online supplementary table S1) were poststained with SYPRO Ruby (BioRad), and spots of interest were robotically excised under computer-assisted visual control. In-gel trypsin digestion and LC-MS/MS analyses are detailed in online supplementary method 5. Finally, we used the X!TandemPipeline to identify and group the differentially expressed proteins (see online supplementary method 6).

### Validation of CD-associated candidate proteins using selected reaction monitoring (SRM)-based targeted proteomics

A targeted LC-SRM assay was developed to validate a subset of 13 candidate proteins discovered in the original 2D-DIGE non-targeted comprehensive survey. The subset of proteins was defined by choosing the ones containing at least two specific peptides for a protein or a group of proteins with identical function in phylogenetically close bacterial strains, and that at the same time, had already been identified in previous label-free shotgun experiments run on equivalent samples, preferentially without prefractionation. Details on sample preparation, the SRM-assay development, the list of transitions, chromatographic and acquisition conditions, data processing and statistical analysis of the quantitative datasets using MSstats,[23] are given in online supplementary method 7. The 284 optimised transitions measured for the 13/46 targeted proteins/peptides are detailed in online supplementary table S2.

Other general statistical analyses are detailed in online supplementary method 8.

## RESULTS

### Pyrosequencing and quality of the microbial extracts

As illustrated by the dendrogram produced by hierarchical clustering of the 16S rRNA pyrosequencing data at the genus and operational taxonomic unit (OTU) levels (figure 1), microbial extracts were closely related to the corresponding stool total 16S rRNA. This illustrates the ability of our extraction method to preserve the microbial diversity observed in the raw sample material. On the other hand, samples did not cluster by clinical diagnosis, CD, or HC, based on their 16S rRNA gene profile alone. This highlights the need for a complementary proteomics viewpoint.

### Discovery of protein signatures of CD-associated gut microbiota by 2D-DIGE

The electrophoretic profile was well conserved across individual samples, and the internal standard (see online supplementary figure S3 and magnified regions thereof in online supplementary figure S4), making it possible to accurately compare spot volumes across the entire experiment. After image alignment and spot co-detection, 2007 protein spots were validated and simultaneously quantified in all 36 images derived from measuring three image channels for each of the 12 gels. There were no pronounced systematic differences between biological replicates

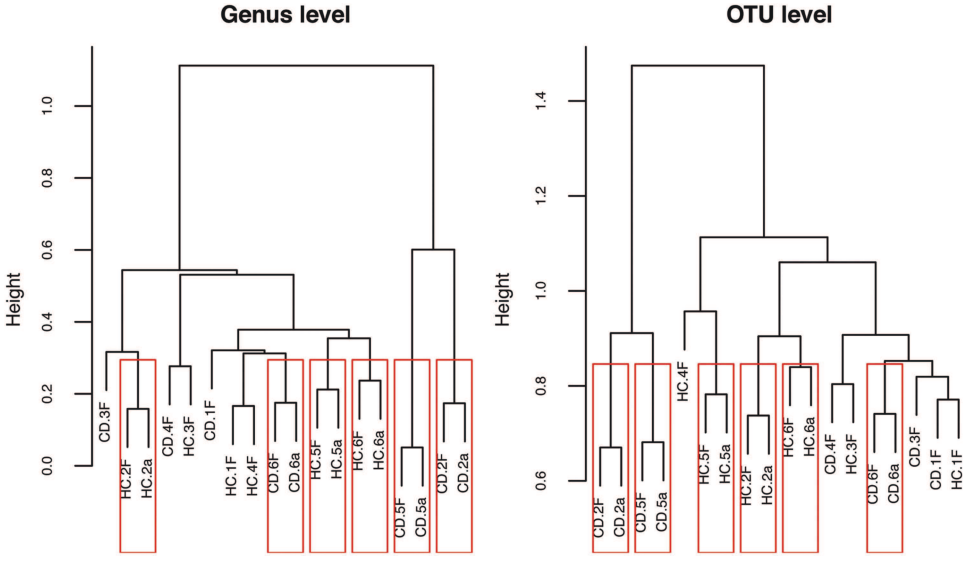



**Figure 1** Structure of all crude samples and half the corresponding extracted microbial pellets profiled by 16S rRNA gene pyrosequencing. Hierarchical clustering of the 16S rRNA gene pyrosequencing dataset (at genus and OTU level) showed a high similarity between population structure of crude samples and those of the corresponding microbial extracts, but did not allow distinguishing between clinical status, CD or HC. HC.1 to HC.6 and CD.1 to CD.6 denote HC and CD patients, respectively; suffix letters, F and a, denote native faeces and bacterial cell extracts, respectively. CD, Crohn's disease; HC, healthy controls.

(samples from a clinical group), and most protein spots (93%) were unchanged between patients and controls allowing reliable normalisation of the data (see online supplementary figure S5). A cluster tree based on the pairwise distances between 2D-DIGE profiles is shown in figure 2. Microbial fractions prepared in duplicate from the same stool specimen always clustered together, reflecting good technical reproducibility, and pairs of duplicates tended to cluster by clinical status, indicating a clinically relevant strong signal. A list of 141 candidate spots (7% of total, 53 increased and 88 decreased in CD patients) was obtained by hierarchical ANOVA or empirical Bayes moderated single-group t test, and all visible selected spots were repeatedly excised from nine SYPRO Ruby poststained gels for LC-MS/MS-based identification (see online supplementary table S1). Eighty-nine spots were found to contain bacterial proteins from a single functional category, and which could be attributed to a defined bacterial subpopulation, with most proteins being from *Bacteroides/Parabacteroides* species, or *Prevotella* species, or members of the order *Clostridiales* (see online supplementary table S3). For robust reporting, however, only a subset of 59 spots were retained (30 increased and 29 decreased in CD patients), which could be identified independently in several gels containing different patient-control pairs (see online supplementary figure S6 for the sequential spot selection process, and see online supplementary table S3 for lists of proteins and peptides). Human proteins were identified in five additional spots with differential signal. Results are summarised in the heat map of figure 3, showing the normalised volumes of those 59 bacterial and 5 human protein spots.

Of the 30 bacterial protein spots which were increased in CD patients, 25 were from the phylum Bacteroidetes, essentially *Bacteroides* species, three were from the phylum Firmicutes, order *Clostridiales*, and two were from the phylum Proteobacteria (see the lower half of figure 3). Human proteins IgA immunoglobulins and carboxypeptidase A1 were identified in two additional spots that we found to be over-represented in CD patients (lower half of figure 3). Proteins that were identified in these spots are reported in table 2, where they are organised according to their lineage and function. Of the 29 bacterial protein spots which were decreased in CD patients, 18 were from the phylum Firmicutes, invariably *Clostridiales* whenever order or lower phylogenic affiliation could be determined, three were from *Prevotella* species, and three others from undefined *Bacteroidales* members, one was from *Escherichia coli*, and four from unknown bacteria (see the upper half of figure 3). Another interesting result was the presence of fragments of human GP2 (pancreatic glycoprotein 2 of zymogen granule membranes) in three under-represented protein spots (upper half of figure 3). Proteins that were identified in these spots are listed in table 3 with their lineage and function.

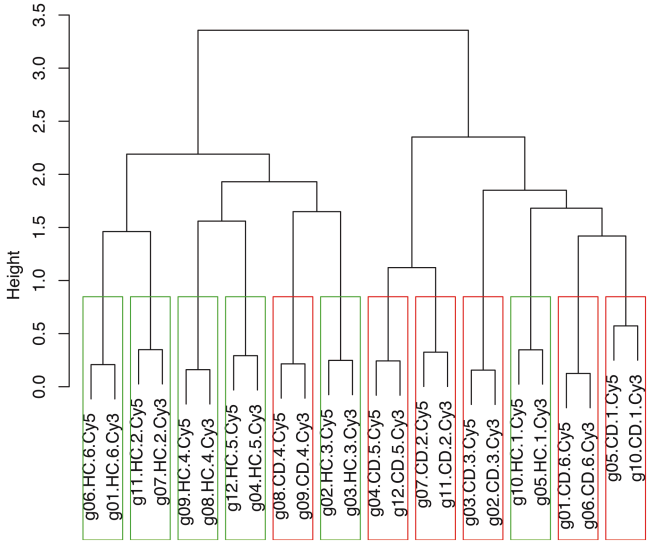



**Figure 2** Cluster tree based on the pairwise distances between 2D-DIGE profiles. Similarities between patterns (normalised volumes of 2007 spots) were assessed by unsupervised hierarchical clustering. HC.1 to HC.6 and CD.1 to CD.6 denote HC and CD patients, respectively; g01–12 denote gel numbers. Microbial fractions prepared in duplicate from the same stool specimen always clustered together, reflecting good technical reproducibility of our method, and pairs of duplicates tended to cluster by clinical status, CD or HC, indicating a clear clinically relevant signal in the proteomics data. CD, Crohn's disease; HC, healthy controls; 2D-DIGE, two-dimensional difference gel electrophoresis.

## Validation of protein signatures of CD-associated gut microbiota by SRM

A subset of 13 proteins (highlighted in yellow on figure 3) found to be differentially abundant between CD and HC on the basis of 2D-DIGE were selected to be validated using a targeted LC-SRM assay. Totally, 46 peptides were chosen and 284 transitions were finely optimised using heavy isotope-labelled synthetic peptides spiked into a sample pool in order to allow the precise relative quantification of the 13 candidate proteins in the sample cohort without further sample fractionation other than a stacking gel (see online supplementary table S2). Thus, all 13 proteins could be unambiguously detected with 2–6 specific peptides in single injections of the total bacterial protein extracts. Results are summarised in figure 4 representing the fold change value (differential expression) and the adjusted p value for each targeted protein from the triplicate analyses of the six CD versus six HC individual samples. Details on individual peptide quantification are given in online supplementary table S4. The differential expression of all candidates detected in the discovery experiment was validated and fold changes spanning 14–28 were detected with a very high confidence.[23]

Clearly, we could validate in CD patients a significant elevation of *Bacteroides* proteins that participate in the protection against oxidative stress (AhpC), in protein synthesis, folding and repair (FusA, DnaK and ClpB), in energy saving, and the maintenance of a high carbon flux within both glycolysis and pentose phosphate pathways (PPi-dependent PfK and TktA-TktB), in the biosynthesis of precursors through the reductive branch of the tricarboxylic acid cycle (KorA), in nutrient acquisition and sensing of the environment (TonB), and in adhesion and colonisation (PepD), while some of these proteins (DnaK, AhpC and TonB-dependent receptors) are recognised for their strong immunogenic properties. We also confirmed elevation of type 1 dockerin from members of the family *Ruminococcaceae*, in CD patients. Other confirmed proteins included the glycolytic enzyme GapA of *Prevotella*, a cell surface protein of undefined *Bacteroidales* members, and the human protein GP2, which were all depleted in CD patients (figure 4).

## Correlating functional shifts with the structure of the bacterial community

We then investigated the question whether the imbalance in bacterial protein abundance that we observed in CD corresponded to changes in gene expression in a stable bacterial community,

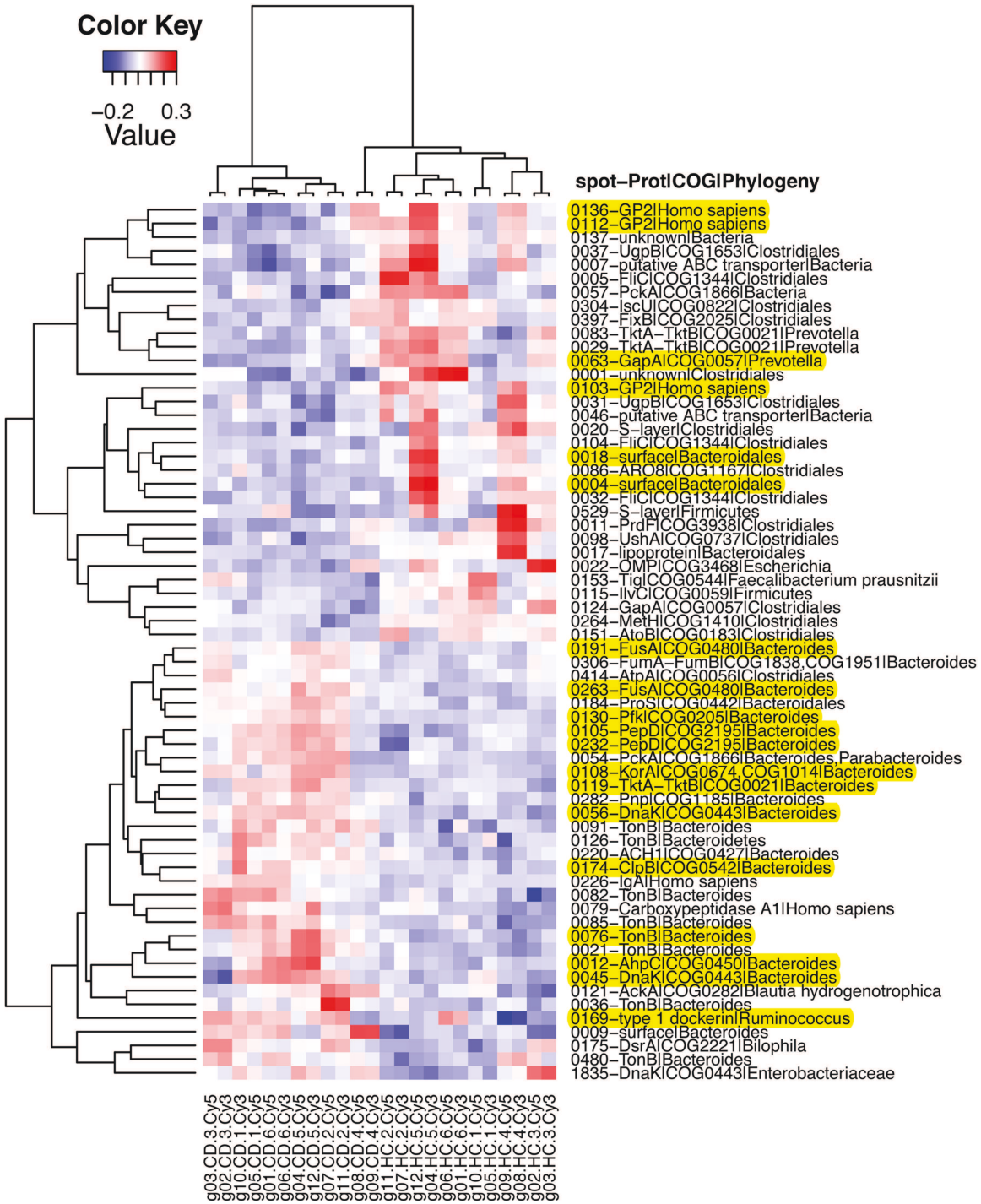



**Figure 3** Cluster heat map constructed from the normalised volumes of spots with significant different intensities between HC and CD patients, and that could be robustly identified. Spot numbers and meaningful names for the associated functions are in the right margin. Similarities between patterns are visualised by unsupervised hierarchical clustering. HC.1 to HC.6 and CD.1 to CD.6 denote HC and CD patients, respectively; g01–12 denote gel numbers. Since all spot variables were centred at the mean (the mean has been subtracted to each value), the new mean for each spot variable is now at 0, making half the values negative as indicated in the colour key. Blue and red tones therefore signify under-represented and over-represented, respectively. Proteins highlighted in yellow in the right margin are those that were chosen for SRM validation. As different forms of the same protein (typically TonB-dependent receptors of *Bacteroides*) may occur in different spots, the number of highlighted spots exceeds 13. Proteins annoted 'surface', 'TonB', 'OMP' and 'lipoprotein' in the right margin, may be grouped into 'cell envelope proteins' in the text when several categories are concerned, including proteins of unknown function that have specific features known to be characteristic of cell envelope localisation. CD, Crohn's disease; HC, healthy controls; SRM, selected reaction monitoring.

or whether they reflected a remodelling of the population structure, or whether both events could have occurred. OTU richness estimated by the bias-corrected Chao 1 richness estimator, was significantly lower ($p=0.015$) in the CD group (840 OTUs, SD 259) compared with the HC group (1371 OTUs, SD 286), but the $\alpha$ diversity Simpson index did not significantly differ between the two groups. This means that a lower number of species was present, which was more evenly distributed in the

**Table 2** List of proteins that were discovered to be over-represented in CD, using the without a priori 2D-DIGE approach

| Lineage | Protein name | Cellular function/pathway | 2D-DIGE | SRM (fold change) |
|---|---|---|---|---|
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides | AhpC | Protection against oxidative stress | + | + (4.4) |
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides | Pnp | Protection against oxidative stress | + | |
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; NA; NA | ProS | Protein synthesis, folding, and repair | + | |
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides | FusA | Protein synthesis, folding, and repair | + | + (5.2) |
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides | DnaK | Protein synthesis, folding, and repair | + | + (4.1) |
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides | ClpB | Protein synthesis, folding, and repair | + | + (4.8) |
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides | PPi-dependent PfK | Energy saving and maintenance of a high flux of carbon within both glycolysis and pentose phosphate pathways | + | + (3.9) |
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides | TktA-TktB | Energy saving and maintenance of a high flux of carbon within both glycolysis and pentose phosphate pathways | + | + (5.2) |
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides | FumA-FumB | Biosynthesis of precursors through the reductive branch of the tricarboxylic acid cycle | + | |
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides | KorA | Biosynthesis of precursors through the reductive branch of the tricarboxylic acid cycle | + | + (4.3) |
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides | TonB-dependent receptors | Nutrient acquisition and sensing of the environment | + | + (11.3) |
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides | Other cell envelope proteins | Nutrient acquisition and sensing of the environment | + | |
| Bacteria; Bacteroidetes; NA; NA; NA; NA | TonB-dependent receptors | Nutrient acquisition and sensing of the environment | + | |
| Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides | PepD | Adhesion and colonisation | + | + (3.9) |
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides | ACH1 | Pyruvate metabolism | + | |
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides/Porphyromonadaceae; Parabacteroides | PckA | Energy saving and maintenance of a high flux of carbon within both glycolysis and pentose phosphate pathways | + | |
| Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcaceae; Ruminococcus | type 1 dockerins | Cellulosome assembly | + | + (27.8) |
| Bacteria; Firmicutes; Clostridia; Clostridiales; NA; NA | AtpA | ATP production | + | |
| Bacteria; Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; Blautia | AckA | ATP production | + | |
| Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; NA | Dnak | Protein synthesis, folding, and repair | + | |
| Bacteria; Proteobacteria; delta/epsilon subdivisions; Deltaproteobacteria; Desulfovibrionales; Desulfovibrionaceae; Bilophila | DsrA | Energy conservation by reducing sulfite | + | |
| Human | IgA immunoglobulins | Coating bacterial cells | + | |
| Human | Carboxypeptidase A1 | Coating bacterial cells | + | |

Those that were confirmed by SRM are indicated in the last column on the right, with their fold change.
CD, Crohn's disease; SRM, selected reaction monitoring; 2D-DIGE, two-dimensional difference gel electrophoresis; NA, not assigned.

CD group. Specific traits of CD microbiota are illustrated by figure 5. Thirty-four OTUs varied or tended to vary in abundance (see online supplementary table S5). Those that were increased in CD were related to *Bacteroides vulgatus* and *Ruminococcus obeum* (genus *Blautia*) and interestingly included one OTU similar to the potentially anti-inflammatory butyrate-producing bacterium SR1/1, while those that were decreased in CD were related to the butyrate-producing bacterium L2-21, to *Roseburia faecis* (T) M72/1, *Faecalibacterium prausnitzii* A2-165, or other clostridial members, and also included one OTU most similar to *Prevotella oralis*. The heat map of figure 6 shows a positive correlation between the abundance of most of the varying protein groups and the abundance of the related varying OTUs. There were, however, a number of interesting exceptions suggesting additional effects at work, for instance a subset of TonB proteins attributed to *Bacteroides* members that were increased in CD patients independently of a modulation of the corresponding bacterial populations (see the green-yellow bands on the right middle part of figure 6).

## DISCUSSION

The present work is a clear demonstration that environmental proteomics of gut microbes can provide molecular signatures of IBDs, becoming a powerful complementary tool for their study and, ultimately, their diagnosis and treatment. So far, long lists of candidate biomarker proteins, notably in oncology,

**Table 3** List of proteins that were discovered to be under-represented in CD, using the without a priori 2D-DIGE approach

| Lineage | Protein name | Cellular function/pathway | 2D-DIGE | SRM (fold change) |
|---|---|---|---|---|
| Bacteria; Firmicutes; Clostridia; Clostridiales; NA; NA | flagellins FliC | Nutrient acquisition and sensing of the environment | – | |
| Bacteria; Firmicutes; Clostridia; Clostridiales; NA; NA | UgpB | Nutrient acquisition and sensing of the environment | – | |
| Bacteria; Firmicutes; Clostridia; Clostridiales; NA; NA | S-layer proteins | Nutrient acquisition and sensing of the environment | | |
| Bacteria; Firmicutes; NA; NA; NA; NA | S-layer proteins | Nutrient acquisition and sensing of the environment | – | |
| Bacteria; Firmicutes; Clostridia; Clostridiales; NA; NA | UshA | Nucleotide transport and metabolism | – | |
| Bacteria; Firmicutes; Clostridia; Clostridiales; NA; NA | ARO8 | Transcription, amino acid transport and metabolism | – | |
| Bacteria, Firmicutes, Clostridia, Clostridiales, Ruminococcaceae, Faecalibacterium, Faecalibacterium prausnitzii | Tig | Folding of newly synthesised proteins | – | |
| Bacteria; Firmicutes; NA; NA; NA; NA | IlvC | Amino acid transport and metabolism | – | |
| Bacteria; Firmicutes; Clostridia; Clostridiales; NA; NA | MetH | Amino acid transport and metabolism | – | |
| Bacteria; Firmicutes; Clostridia; Clostridiales; NA; NA | PrdF | Amino acid transport and metabolism | – | |
| Bacteria; Firmicutes; Clostridia; Clostridiales; NA; NA | IscU | General metabolism | – | |
| Bacteria; Firmicutes; Clostridia; Clostridiales; NA; NA | FixB | General metabolism | – | |
| Bacteria; Firmicutes; Clostridia; Clostridiales; NA; NA | AtoB | General metabolism | – | |
| Bacteria; Firmicutes; Clostridia; Clostridiales; NA; NA | GapA | Glycolysis | – | |
| Bacteria; Firmicutes; Clostridia; Clostridiales; NA; NA | Unknown function | Unknown | – | |
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; NA; NA | Surface proteins | Nutrient acquisition and sensing of the environment | – | –(3.2) |
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; NA; NA | Lipoprotein | Nutrient acquisition and sensing of the environment | – | |
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Prevotellaceae; Prevotella | TktA-TktB | Energy saving and maintenance of a high flux of carbon within both glycolysis and pentose phosphate pathways | – | |
| Bacteria; Bacteroidetes; Bacteroidia; Bacteroidales; Prevotellaceae; Prevotella | GapA | Glycolysis | – | –(14) |
| *Escherichia* | Outer membrane proteins | Nutrient acquisition and sensing of the environment | – | |
| Bacteria; NA; NA; NA; NA; NA | ABC transporters | Nutrient acquisition and sensing of the environment | – | |
| Bacteria; NA; NA; NA; NA; NA | PckA | Energy saving and maintenance of a high flux of carbon within both glycolysis and pentose phosphate pathways | – | |
| Bacteria; NA; NA; NA; NA; NA | Unknown function | Unknown | – | |
| HUMAN | Fragments of GP2 | Coating bacterial cells | – | –(2.4) |

Those that were confirmed by SRM are indicated in the last column on the right, with their fold change.
CD, Crohn's disease; SRM, selected reaction monitoring; 2D-DIGE, two-dimensional difference gel electrophoresis.

remarkably never progressed from research discovery to clinical application because de novo development of multiple ELISA would have been prohibitive in time and money. SRM, by contrast, offers a valuable alternative to antibody-based validations, as it allows the accurate and robust multiplexed quantification of a range of proteins in complex samples without the need for antibodies and ELISA testing. This technology has, for instance, recently been used successfully for the verification of diagnostic and prognostic cancer biomarkers.[24 25] In the present study, we first developed and validated a workflow for the discovery of protein signals specific to CD-associated gut microbial communities without any a priori assumption of the metabolic and/or cellular functions that can accompany CD. Then we developed a SRM assay to confirm a set of candidates, as de novo development of ELISA assays would not have been feasible in a reasonable timeframe and poorly adapted to multiple verifications.

Simple unsupervised hierarchical clustering of samples based on their 2D-DIGE profiles showed that five pairs of replicates out of six within each group, CD and HC, already clustered together, indicating a clear signal, whereas, hierarchical clustering based on 16S rRNA gene pyrosequencing (at the genus and OTU levels) did not allow a distinction between clinical status. This illustrates well that the metaproteomic approach is a powerful tool for highlighting functional imbalances even in the absence of clear major shifts in the dominant bacterial groups or species. The 2D-DIGE strategy further solved the specific question of detecting differences between groups, a problem that faces even greater challenges in label-free shotgun metaproteomics.[26–31] A set of proteins from members of the Bacteroidetes phylum, largely *Bacteroides* species, were over-represented in CD microbiota. By contrast, under-represented proteins were mostly from *Clostridiales* (Firmicutes phylum), and more rarely from *Prevotella* and undefined *Bacteroidales* members. Functions that we found to be increased in the *Bacteroides*/*Bacteroidales* population from CD microbiota included proteins corresponding, in general, to functions related to strategic adaptation for survival in challenging environments. For instance, DnaKs and other chaperones, such as ClpB homologues, have a defensive role against oxidative, nitrosative, nutritional, osmotic and pH stresses that are likely to occur in the gut of CD patients.[32–35] AhpCs represent another important defence mechanism to cope with reactive nitrogen intermediates and reactive oxygen species.[36 37] An over-representation of proteins involved in the binding and import of nutrients (TonB-dependent receptors and other cell envelope proteins) could be related to an increased need for carbon substrates and/or for micronutrients to fuel increased central metabolism in *Bacteroidales*.[38 39] Consistent with this, a set of key enzymes that maximise the energy yield from monosaccharide catabolism in *Bacteroides* (PPi-dependent Pfk, PEP-carboxykinase PckA and

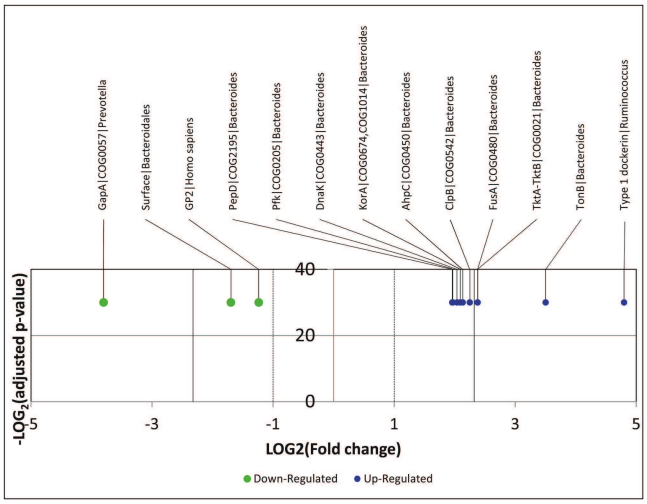



**Figure 4** Volcano plot representing results of the LC-SRM assays on the 13 targeted proteins. The logarithmic fold changes (CD vs HC) are plotted against negative logarithmic adjusted p values calculated with the R package MSstats and performed from triplicate injections.[23] All targeted proteins were found to be either upregulated or downregulated in CD patients compared with controls, and the results validated all candidates identified in the discovery experiments. CD, Crohn's disease; HC, healthy controls; SRM, selected reaction monitoring.

fumarate hydratase FumA-FumB) were also over-represented in CD-associated microbial proteomes.[40] Finally, a higher abundance of PepD could contribute to increased surface colonisation by *Bacteroides* members observed in CD patients.[41–44] Remarkably, the overexpression of 10 of these proteins was reliably confirmed by quantitative SRM measurements in unfractionated bacterial protein extracts. Therefore, we reach an attractive hypothesis, that a number of *Bacteroidales* members, essentially *Bacteroides* species, might be adapted or promoted under environmental conditions specific to the gut of CD patients, and that a set of overabundant bacterial proteins that can be quantified by SRM, could be regarded as bacterial signatures for CD. Moreover, all of them make sense in the context of our understanding of CD. Indeed, a number of these proteins have been proposed as major traits for allowing opportunistic pathogens, including *Bacteroides fragilis*, to colonise mucus layers, breach host barriers, and invade the mucosae. For instance, DnaK and AhpC, which are usually intracellular proteins, may also be found in the outer membrane where they bind human plasminogen and enhance its conversion into plasmin by host activators, a scenario which might promote colonisation and host invasion.[45] DnaK proteins exhibit strong immunostimulatory properties both at the level of the innate and adaptative immune system, and can, moreover, promote the processing and presentation of other bacterial or food antigens by chaperoning them.[46] AphC of several bacteria also demonstrate immunogenic properties as assessed by high titres of seric antibodies against purified/recombinant targets or whole proteome maps,[47] and the β-barrel domains of TonB-dependent receptors are suspected to play an important role in the virulence of Gram-negative bacteria, exposing epitopes on the bacterial surface.[48] Consistent with this and with a previous report,[49] we found an over-representation of microbiota-associated secretory IgA in CD patients. These results encourage inspection of the immunome of the intestinal microbiota to capture those strongly IgA-coated bacteria as well as a set of derived antigenic peptides that could be used to map and distinguish specific antibody profiles in subgroups of IBD patients, and thus facilitate appropriate therapeutic options, evaluation of treatment efficacy, and long-term follow-up.

Conversely, many members of the order *Clostridiales* and some of the genus *Prevotella* appeared unable to meet the ecological challenge imposed in CD patients, as judged from the decrease in abundance of key proteins involved in diverse cellular and biochemical functions within this population. For instance, 25 different flagellin FliC proteins attributed to *Clostridiales* members were identified in three under-represented spots, while common enteric flagellins are proposed as major targets of the CD-associated aberrant immune response.[50] The under-representation of trigger factor Tig attributed to *Faecalibacterium prausnitzii* further suggests that this numerically dominant and potentially anti-inflammatory subgroup,[51] might fail to sustain efficient protein synthesis in CD patients. Finally, under-representation of key enzymes, notably GapA and TktA-TktB, attributed to *Prevotella* members could reflect the inability of this subpopulation in maintaining the flux of carbon within both glycolysis and pentose phosphate pathways. Interestingly, the CD-associated under-representation of GapA from *Prevotella* was confirmed by SRM. Consistent with this, we found that the relative abundances of members of the lineage *Prevotellaceae-Prevotella* tended to decrease in CD patients, and that one OTU assigned to *Prevotella* was under-represented in CD.

While sequencing of the 16S rRNA-encoding genes revealed many positive correlations between the abundances of the varying protein groups and the abundances of the related varying OTUs, it also highlighted a number of unexpected interesting deviations. In particular, some of the TonB-dependent receptors and other uncharacterised proteins localised in bacterial cell envelopes of *Bacteroides* species were increased in CD patients independently of a modulation of the corresponding bacterial populations. Therefore, findings in the present study clearly point to functional changes beyond what can be explained by a mere shift in populations, and extend earlier observations of possible dissociations between structural and functional changes in obese individuals[29 30] or CD patients.[31] The fact that completely different approaches independently identified *Bacteroides* membrane proteins as over-represented in CD,[31] and moreover, that this was confirmed here by SRM, also gives strong support for regarding these proteins as possible relevant bacterial signatures for CD, and highlights the need for future work focusing on microbial cell envelopes in health and disease, inasmuch as this subcellular fraction constitutes the first line of interaction with the host.

Our study also provides the first and unexpected clue towards under-representation of intestinal microbiota-associated GP2 in CD patients. This was confirmed by SRM, and may favour adhesion of bacteria to the mucosa, and then promote inflammation. Indeed, it has been demonstrated recently that recombinant human GP2 binds *Escherichia coli* Type I fimbriae, a bacterial adhesin commonly expressed by members of the intestinal microbiota. A role in host defence has been proposed in which GP2 may serve as a physical barrier that prevents bacteria from binding to host cell receptors.[52] Consistent with this, a higher bacterial biomass on the mucosa, and an adherent mucosal biofilm enriched with *Bacteroides fragilis* were shown to be prominent features in IBD patients.[43] Finally, the question arises as to whether decreased GP2 binding to bacteria might be related to increased anti-GP2 titres reported in CD patients.[53]

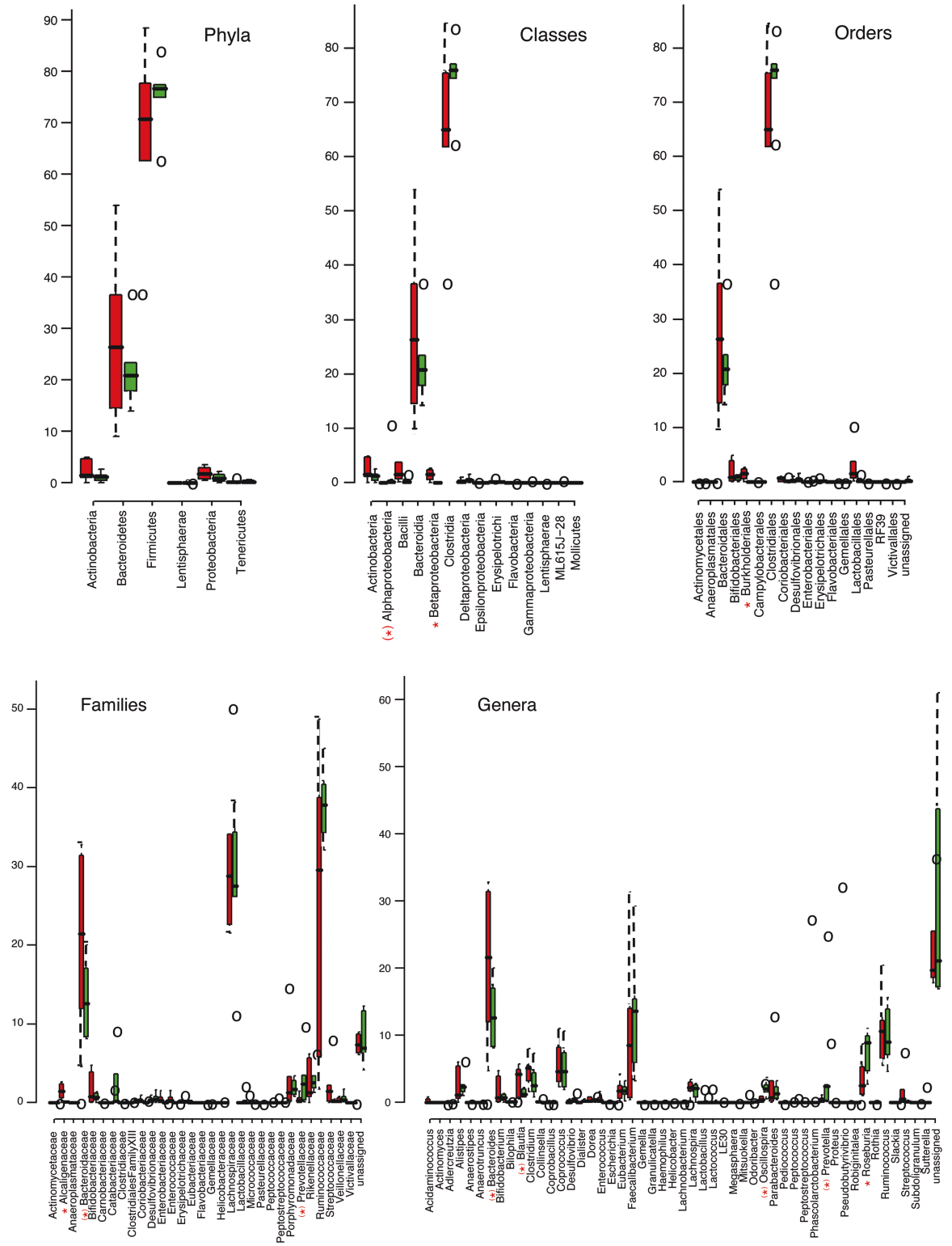



**Figure 5** Box plots of the relative abundances of faecal bacterial populations found by 454 pyrosequencing. Differences between Crohn's patients (□) and HCs (□) at the different phylogenetic levels were considered *significant for p≤0.05, and (*)tendencies were reported up to p≤0.10 ('glm' with the 'quasibinomial family'). Specific traits of CD microbiota were significantly increased abundances of members in the lineage *Betaproteobacteria-Burkhoderiales-Alcaligenaceae*, a tendency towards increased abundances of *Bacteroidaceae-Bacteroides* and *Blautia*, significantly lower numbers of *Roseburia*, and a tendency towards lower numbers of *Alphaproteobacteria*, *Prevotellaceae-Prevotella* and *Oscillospira*. On the other hand, inter-individual variability was higher in CD patients, which is in agreement with heterogeneity of CD. CD, Crohn's disease; HCs, healthy controls.

In conclusion, our metaproteomics approach spanning discovery to confirmation demonstrates, for the first time, the feasibility of extracting bacterial protein signals relevant to CD, by interrogating myriads of intestinal bacteria, even from a small number of patients and HCs. We provide an initial list of CD-associated microbial proteins extracted from a typical group of patients, which could represent major common features in CD patients. The next step should be to validate the specificity

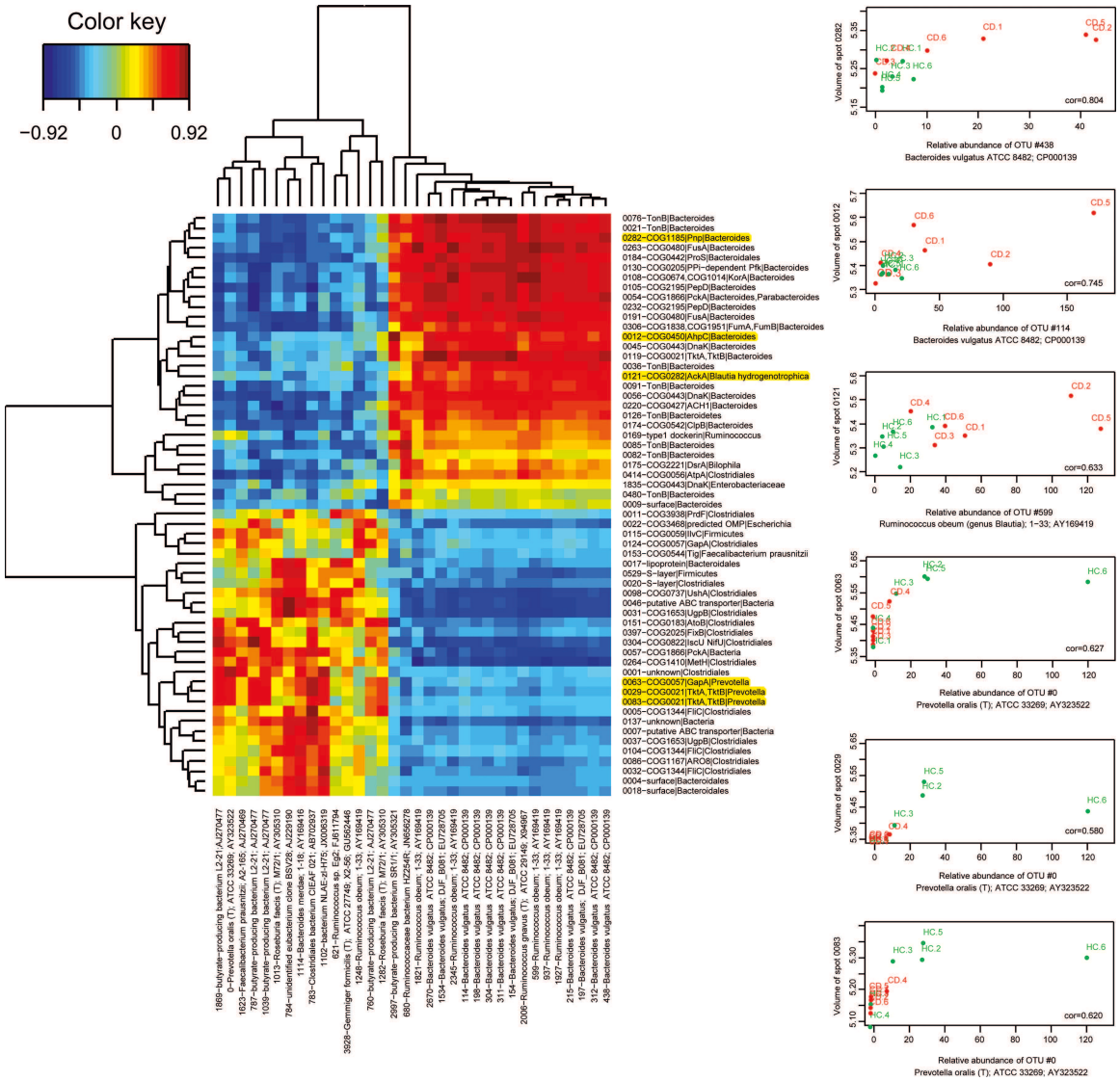



**Figure 6** Heat map of the correlation matrix between abundance of the varying protein groups and the abundance of the related varying OTUs (left panel). Red and hotter orange tones indicate a positive correlation between abundance of a protein group and abundance of the corresponding OTUs; green and yellow tones indicate absence of correlation as for example, TonB proteins and other uncharacterised surface proteins of *Bacteroides* in spots 0082, 0480 and 0009, and DnaK of *Enterobacteriaceae* in spot 1835 (right middle part of the image). Activities that were found to be increased and decreased in CD patients fell in the upper and lower halves of the image, respectively. OTUs that were found to be increased and decreased in abundance in CD patients fell into the right and left halves of the image, respectively. Examples of positive correlations are detailed on right panel and highlighted in yellow on panel A. CD, Crohn's disease.

and sensitivity of bacterial protein signals either in individual clinical trials with well-defined and homogenous CD populations, or in a comprehensive study with a larger heterogeneous patient cohort. Given that effects are harder to detect in smaller samples, one can expect that even more subtle differences could be detected in larger cohorts, and that inclusion and accurate quantification of additional predefined sets of proteins could be used to refine recognition of IBD entities in the very near future. The extraction and robust quantification of bacterial protein signals is also a way to identify disrupted protein networks that drive onset and perpetuation of CD and, therefore, candidate targets for IBD treatment based on gut-ecological intervention strategies.

**Author affiliations**
[1]UMR1319 Micalis, INRA, Jouy-en-Josas, France
[2]Chair of Bioinformatics, Boku University Vienna, Vienna, Austria
[3]Department of Life Sciences, University of Warwick, Warwickshire, UK
[4]UMR1313 GABI, Iso Cell Express (ICE), INRA, Jouy-en-Josas, France
[5]Plate-forme d'Analyse Protéomique de Paris Sud-Ouest (PAPPSO), INRA, Gif-sur-Yvette, France
[6]Laboratoire de Spectrométrie de Masse BioOrganique (LSMBO), IPHC, Université de Strasbourg, Strasbourg, France
[7]Gastroenterology and Nutrition Unit, Hôpital Saint-Antoine, AP-HP, Paris, France
[8]UR1077, Mathématique Informatique et Génome (MIG), INRA, Jouy-en-Josas, France
[9]UR341, Mathématiques et Informatique Appliquées (MIA), INRA, Jouy-en-Josas, France

**Acknowledgements** We are grateful to Bertrand Nicolas for his help in the preparation of figures.

**Contributors** CJ, DPK, CC, HS, PS, LB, JD: conception and design of the study; CB, AG, SV, CC, SM, PL, FB, FL, WC, VL, NP: acquisition of data; CJ, DPK, CB, AG, SV, CC, SM, FB, WC, VL, NP: analysis and interpretation of data; CJ, DPK, CB, SV, CC: drafting of the manuscript; DPK, AG, CC, HS, PS, J-FG, SDE, AVD, JD, BV, PM, VM: critical revision of the manuscript for important intellectual content; CJ, DPK,

SV, CC, OD, BS: statistical analysis; CB, AG: technical or material support; CJ, HS, PS, LB: study supervision.

**Funding** The Boku Chair of Bioinformatics acknowledges funding by the Vienna Science and Technology Fund (WWTF), Baxter AG, Austrian Research Centres Seibersdorf, and the Austrian Centre of Biopharmaceutical Technology. We acknowledge the 'Fondation pour la Recherche Médicale' for funding the triple quadrupole instrument for targeted proteomics experiments.

**Competing interests** None.

**Patient consent** Obtained.

**Ethics approval** This study was conducted with the approval of the Ethics Committee of the St Antoine hospital, Paris.

**Provenance and peer review** Not commissioned; externally peer reviewed.



## REFERENCES

1 Rath HC, Herfarth HH, Ikeda JS, *et al*. Normal luminal bacteria, especially Bacteroides species, mediate chronic colitis, gastritis, and arthritius in HLA-B27/human β2 microglobulin transgenic rats. *J Clin Invest* 1996;98:945–53.
2 D'Haens GR, Geboes K, Peeters M, *et al*. Early lesions of recurrent Crohn's disease caused by infusion of intestinal contents in excluded ileum. *Gastroenterology* 1998;114:262–7.
3 Rutgeerts P, Hiele M, Geboes K, *et al*. Controlled trial of metronidazole treatment for prevention of crohn's recurrence after ileal resection. *Gastroenterology* 1995;108:1617–21.
4 Mathew CG. New links to the pathogenesis of Crohn disease provided by genome-wide association scans. *Nat Rev Genet* 2008;9:9–14.
5 Cho JH. The genetics and immunopathogenesis of inflammatory bowel disease. *Nat Rev Immunol* 2008;8:458–66.
6 Cleynen I, González JR, Figueroa C, *et al*. Genetic factors conferring an increased susceptibility to develop Crohn's disease also influence disease phenotype: results from the IBDchip European Project. *Gut* 2013;62:1556–65.
7 Cerf-Bensussan N, Gaboriau-Routhiau V. The immune system and the gut microbiota: friends or foes? *Nat Rev Immunol* 2010;10:735–44.
8 Tap J, Mondot S, Levenez F, *et al*. Towards the human intestinal microbiota phylogenetic core. *Environ Microbiol* 2009;11:2574–84.
9 Seksik P, Rigottier-Gois L, Gramet G, *et al*. Alterations of the dominant faecal bacterial groups in patients with Crohn's disease of the colon. *Gut* 2003;52:237–42.
10 Mangin I, Bonnet R, Seksik P, *et al*. Molecular inventory of faecal microflora in patients with Crohn's disease. *FEMS Microbiol Ecol* 2004;50:25–36.
11 Manichanh C, Rigottier-Gois L, Bonnaud E, *et al*. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* 2006;55:205–11.
12 Kleessen B, Kroesen AJ, Buhr HJ, *et al*. Mucosal and invading bacteria in patients with inflammatory bowel diseasecompared with controls. *Scand J Gastroenterol* 2002;37:1034–41.
13 Scanlan PD, Shanahan F, O'Mahony C, *et al*. Culture-independent analyses of temporal variation of the dominant fecal microbiota and targeted bacterial subgroups in Crohn's disease. *J Clin Microbiol* 2006;44:3980–8.
14 Gophna U, Sommerfeld K, Gophna S, *et al*. Differences between tissue-associated intestinal microfloras of patients with Crohn's disease and ulcerative colitis. *J Clin Microbiol* 2006;44:4136–41.
15 Frank DN, St Amand AL, Feldman RA, *et al*. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci USA* 2007;104:13780–5.
16 Qin J, Ruiqiang L, Raes J, *et al*. A human gut microbial gene catalog established by deep metagenomic sequencing. *Nature* 2010;464:59–65.
17 Human Microbiome Jumpstart Reference Strains ConsortiumNelson KE, Weinstock GM, Highlander SK, *et al*. A catalog of reference genomes from the human microbiome. *Science* 2010;328:994–9.
18 Le Chatelier E, Nielsen T, Qin J, *et al*. Richness of human gut microbiome correlates with metabolic markers. *Nature* 2013;500:541–6.
19 Vermeire S, Ferrante M, Rutgeerts P. Recent advances: personalised use of current Crohn's disease therapeutic options. *Gut* 2013;62:1511–15.
20 Harvey RF, Bradshaw JM. A simple index of Crohn's disease activity. *The Lancet* 1980;1:514.
21 Kreil DP, Kar NA, Lilley KS. DNA microarray normalization methods can remove bias from differential protein expression analysis of 2D difference gel electrophoresis results. *Bioinformatics* 2004;20:2026–34.
22 Karp NA, Kreil DP, Lilley KS. Determining a significant change in protein expression with DeCyder during a pair-wise comparison using two-dimensional difference gel electrophoresis. *Proteomics* 2004;4:1421–32.
23 Chang CY, Picotti P, Hüttenhain R, *et al*. Protein significance analysis in selected reaction monitoring (SRM)measurements. *Mol Cell Proteomics* 2012;11:M111.014662.
24 Pan S, Chen R, Brand RE, *et al*. Multiplex targeted proteomic assay for biomarker detection in plasma: a pancreatic cancer biomarker case study. *J Proteome Res* 2012;11:1937–48.
25 Hüttenhain R, Soste M, Selevsek N, *et al*. Reproducible quantification of cancer-associated proteins in body fluids using targeted proteomics. *Sci Transl Med* 2012;4:142ra94.
26 Verberkmoes NC, Russell AL, Shah M, *et al*. Shotgun metaproteomics of the human distal gut microbiota. *ISME J* 2009;3:179–89.
27 Cantarel BL, Erickson AR, VerBerkmoes NC, *et al*. Strategies of metagenomic-guided whole-community proteomics of complex microbial environments. *PLoS ONE* 2011;6:e27173.
28 Rooijers K, Kolmeder C, Juste C, *et al*. An iterative workflow for mining the human intestinal metaproteome. *BMC Genomics* 2011;12:6. http://www.biomedcentral.com/1471-2164/12/6
29 Kolmeder CA, de Been M, Nikkilä J, *et al*. Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions. *PLoS ONE* 2012;7:e29913.
30 Ferrer M, Ruiz A, Lanza F, *et al*. Microbiota from the distal guts of lean and obese adolescents exhibit partial functional redundancy besides clear differences in community structure. *Environ Microbiol* 2013;15:211–26.
31 Erickson AR, Cantarel BL, Lamendella R, *et al*. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS ONE* 2012;7:e49138.
32 Keshavarzian A, Banan A, Farhadi A, *et al*. Increases in free radicals and cytoskeletal protein oxidation and nitration in the colon of patients with inflammatory bowel disease. *Gut* 2003;52:720–8.
33 Schill R, Breuer RI, Klein F, *et al*. Comparison of the composition of faecal fluid in Crohn's disease and ulcerative colitis. *Gut* 1982;23:326–32.
34 Nugent SG, Kumar D, Rampton DS, *et al*. Intestinal luminal pH in inflammatory bowel disease: possible determinants and implications for therapy with aminosalicylates and other drugs. *Gut* 2001;48:571–7.
35 Baumgart DC, Sandborn WJ. Inflammatory bowel disease: clinical aspects and established and evolving therapies. *Lancet* 2007;369:1641–57.
36 Chen L, Xie QW, Nathan C. Alkyl hydroperoxide reductase subunit C (AhpC) protects bacterial and human cells against reactive nitrogen intermediates. *Mol Cell* 1998;1:795–805.
37 Rocha ER, Smith CJ. Role of the alkyl hydroperoxide reductase (ahpCF) gene in oxidative stress defense of the obligate anaerobe Bacteroides fragilis. *J Bacteriol* 1999;181:5701–10.
38 Spence C, Wells WG, Smith CJ. Characterization of the primary starch utilization operon in the obligate anaerobe *Bacteroides fragilis*: regulation by carbon source and oxygen. *J Bacteriol* 2006;188:4663–72.
39 Martens EC, Koropatkin NM, Smith TJ, *et al*. Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm. *J Biol Chem* 2009;284:24673–7.
40 Macy JM, Ljungdahl LG, Gottschalk G. Pathway of succinate and propionate formation in *Bacteroides fragilis*. *J Bacteriol* 1978;134:84–91.
41 Van Houdt R, Michiels CW. Role of bacteria cell surface structures in Escherichia coli biofilm formation. *Res Microbiol* 2005;156:626–33.
42 Stoodley P, Sauer K, Davies DG, *et al*. Biofilms as complex differentiated communities. *Annu Rev Microbiol* 2002;56:187–209.
43 Swidsinski A, Weber J, Loening-Baucke V, *et al*. Spatial organization and composition of the mucosal Flora in patients with inflammatory bowel disease. *J Clin Microbiol* 2005;43:3380–9.
44 Kleessen B, Kroesen AJ, Buhr HJ, *et al*. Mucosal and invading bacteria in patients with inflammatory bowel disease compared with controls. *Scand J Gastroenterol* 2002;37:1034–41.
45 Knaust A, Martin VR, Weber MVR, *et al*. Cytosolic proteins contribute to surface plasminogen recruitment of *Neisseria meningitidis*. *J Bacteriol* 2007;189:3246–55.
46 Routsias JG, Tzioufas AG. The role of chaperone proteins in autoimmunity. *Ann NY Acad Sci* 2006;1088:52–64.
47 Chitlaru T, Gat O, Grosfeld H, *et al*. Identification of in vivov-expressed immunogenic proteins by serological proteome analysis of the *Bacillus anthracis* secretome. *Infect Immun* 2007;75:2841–52.

48 Gribun A, Katcoff DJ, Hershkovits G, *et al*. Cloning and characterization of the gene encoding for OMP-PD porin: the major Photobacterium damsela outer membrane protein. *Curr Microbiol* 2004;48:167–74.
49 van der Waaij LA, Kroese FGM, Visser A, *et al*. Immunoglobulin coating of faecal bacteria in inflammatory bowel disease. *Eur J Gastroenterol Hepatol* 2004;16:669–74.
50 Sitaraman SV, Klapproth JM, Moore DA III, *et al*. Elevated flagellin-specific immunoglobulins in Crohn's disease. *Am J Physiol Gastrointest Liver Physiol* 2005;288:G403–6.
51 Sokol H, Pigneur B, Watterlot L, *et al*. *Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patient. *Proc Natl Acad Sci U S A* 2008;105:16731–6.
52 Yu S, Lowe AW. The pancreatic zymogen granule membrane protein, GP2, binds *Escherichia coli* type 1 Fimbriae. *BMC Gastroenterol* 2009;9:58–64.
53 Roggenbuck D, Hausdorf G, Martinez-Gamboa L, *et al*. Identification of GP2, the major zymogen granule membrane glycoprotein, as the autoantigen of pancreatic antibodies in Crohn's disease. *Gut* 2009;58:1620–8.

# Chapter III Application of targeted proteomics for the relative and absolute quantification of Methionine Aminopeptidase Proteins

## *A.1. Context of the project*

This project was carried out in collaboration with the Institute for Integrative Biology of the Cell (I2BC) in Gif-Sur-Yvette, and particularly with Frédéric Frottin, Willy Bienvenut, Thierry Meinnel and Carmela Giglione.

Methionine aminopeptidase proteins are in charge of N-terminal Methionine excision. This biological process has an extreme importance in the cell illustrated by the fact that it is a highly conserved process across organisms. Most proteins are synthetized with a methionine in the first residue. However for two-thirds of all proteins this methionine is removed later on. The exact reason for this process is not well known. It is believed that this process controls protein stability and half-life [22].

In eukaryotic cells there are two classes of methionine aminopeptidase proteins, MetAP1 and MetAP2. Frottin *et al.* showed that these two proteins have highly similar substrate specificity *in vitro* and that they are interchangeable in plants [23, 24]. They also share a similar 3D structure of the active site and a conserved metal binding site.

Enzymes responsible of N-terminal Methionine excision are highly regulated when cells are in different states or when submitted to different stress conditions [22, 25]. However it is not well understood how exactly the regulation of MetAPs affects the proteome.

A way to study the respective roles of MetAP1 and MetAP2 and gain more insight into the crucial role of N-terminal Methionine excision in the cell consists in the use of drugs specifically targeting MetAP2. Fumagillin is a drug that binds and inhibits MetAP2 but does not affect the activity of MetAP1 [23, 26]. Fumagillin and its derivatives cause cell-cycle arrest in endothelial cells and several cancer cell lines. This suggests that MetAP2 could be a target for cancer therapy. However, derivatives of this drug were shown to cause neurotoxicity in phase III clinical trials [27].

To take full advantage of the inhibition of MetAP2 as a cancer therapy, it is crucial to identify and understand the function of the N-terminal Methionine excision processed by MetAPs as well as improving our understanding of the molecular consequences of the MetAP2 inhibition.

Our collaborators identified cell lines with different sensitivities to fumagillin:

- Highly sensitive cell lines, i.e. having a low proliferation when exposed to the drug: HUVEC, U87, U937, A549.
- Insensitive cell lines : Jurkat, HCT116 and K562.

With this information, they characterized the proteomes and N-terminomic profiles of several cell lines to identify possible variations at the protein level that could explain the difference in sensitivity to fumagillin. From this study, they concluded that the specific variations in the proteomes do not explain the selectivity in the phenotype. Also, they established that the MetAP2 inhibition does not significantly affect the protein N-terminal Methionine excision process nor does it affect the downstream N-terminal modifications such as N-myristoylation and N-acetylation.

In this bottom-up proteomic study, the MetAP1 and MetAP2 could not be identified. Therefore, to check if the difference of abundance of these two proteins could explain the difference of sensitivity of the different cell lines, they developed a quantification assay using immunodetection using several commercial antibodies. But none of the MetAPs could be detected even in the highest sensitive cell lines and even using high protein amounts, suggesting that these proteins accumulate in the analyzed cell lines at very low levels.

Thus, we developed a targeted method using Selected Reaction Monitoring to attempt monitoring and quantifying these two proteins. Three cell lines were chosen to be analyzed: the most sensitive cell line from cancer tissues (U87), a sensitive endothelial cell line (HUVEC) and one insensitive cell line (K562).

## *A.2. LC-SRM method development and sample preparation protocol optimization*

### A.2.1 LC-SRM method development

For each of the two methionine aminopeptidase proteins MetAP1 and MetAP2, six signature peptides were chosen. The list of chosen peptides and the reasons why they were chosen can be seen on Figure IV-60. For each protein at least 20 proteotypic peptides between 7 and 25 amino acids are predicted. To find the best peptides, a high priority was given to peptides already seen in LC-MS/MS experiments by our collaborators. Then, we looked for well-responding peptides in LC-MS conditions using SRM Atlas [161] and ProteomicsDB [163]. For each peptide, the hydrophobicity index was calculated using SSRCalc in order to eliminate very hydrophilic or very hydrophobic peptides. The hydrophocity index was calculated using the web server and the following equation HI = -2.6687 + 0.4954xH was used (http://hs2.proteome.ca/SSRCalc/SSRCalcX.html). We empirically determined that with this equation, peptides with a hydrophobicity index between 15 and 45 are good candidates. In the final list of signature peptides, three of them had a potential ragged end in the C-terminal part of the peptide, i.e. the peptide sequence is followed by a cleavage site (KK, RR, KR or RK motif). However these three peptides were found to be very well-responding peptides in experimental data. Furthermore the peptide GSYTAQFEHTILLRPTCK has a missed-cleavage in its sequence. Even if the activity of the trypsin is reduced when the cleavage site is followed by a proline [222] peptides with missed-cleavages should be avoided. In this case this peptide was nevertheless maintained as it was found to be a very well-responding peptide on ProteomicsDB and was chosen as a signature peptide in SRM Atlas. The quantification will not be done solely on the ragged-end and missed-cleavage peptides but they will help to monitor the protein and control the presence or absence of the protein in the samples. Finally, the peptides were also chosen to try to cover different regions of the protein and were thus chosen well distributed over the protein sequences (Figure IV-60).

| Peptide Sequence | LC-MS Data | SRM Atlas | ProteomicsDB | Hydrophobicity Index (SSRCalc) | Ragged End | Missed Cleavages |
|---|---|---|---|---|---|---|
| MetAP 1 | | | | | | |
| LQCPTCIK | | | | 16.41 | | |
| LGIQGSYFCSQECFK | | X | | 35.25 | | |
| SCCTSVNEVICHGIPDR | X | X | | 25.09 | X | |
| ELGNIIQK | | | X | 21.11 | | |
| HAQANGFSVVR | X | | X | 15.30 | | |
| SAQFEHTLLVTDTGCEILTR | X | | X | 40.47 | X | |
| MetAP 2 | | | | | | |
| ALDQASEEIWNDFR | X | X | X | 40.45 | | |
| IDFGTHISGR | | X | X | 17.54 | | |
| NLNGHSIGQYR | X | | X | 16.30 | | |
| NFDVGHVPIR | X | | X | 20.32 | | |
| HLLNVINENFGTLAFCR | X | | | 48.32 | X | |
| GSYTAQFEHTILLRPTCK | | X | X | 32.86 | | X |

**>sp|P53582|MAP11_HUMAN Methionine aminopeptidase 1 OS=Homo sapiens GN=METAP1 PE=1 SV=2**

MAAVETRVCETDGCSSEAKLQCPTCIK LGIQGSYFCSQECFKGSWATHKLLHKKAKDEKA
KREVSSWTVEGDINTDPWAGYRYTGKLRPHYPLMPTRPVPSYIQRPDYADHPLGMSESEQ
ALKGTSQIKLLSSEDIEGMRLVCRLAREVLDVAAGMIKPGVTTEEIDHAVHLACIARNCY
PSPLNYYNFPKSCCTSVNEVICHGIPDRRPLQEGDIVNVDITLYRNGYHGDLNETFFVGE
VDDGARKLVQTTYECLMQAIDAVKPGVRYRELGNIIQK HAQANGFSVVRSYCGHGIHKLF
HTAPNVPHYAKNKAVGVMKSGHVFTIEPMICEGGWQDETWPDGWTAVTRDGKRSAQFEHT
LLVTDTGCEILTRRLDSARPHFMSQF

**>sp|P50579|MAP2_HUMAN Methionine aminopeptidase 2 OS=Homo sapiens GN=METAP2 PE=1 SV=1**

MAGVEEVAASGSHLNGDLDPDDREEGAASTAEEAAKKKRRKKKKSKGPSAAGEQEPDK ES
GASVDEVARQLERSALEDKERDEDDEDGDGDGDGATGKKKKKKKKKRGPKVQTDPPSVPI
CDLYPNGVFPK GQECEYPPTQDGRTAAWRTTSEEKKALDQASEEIWNDFREAAEAHRQVR
KYVMSWIKPGMTMIEICEKLEDCSRKLIKENGLNAGLAFPTGCSLNNCAAHYTPNAGDTT
VLQYDDICKIDFGTHISGRIIDCAFTVTFNPKYDTLLKAVKDATNTGIKCAGIDVRLCDV
GEAIQEVMESYEVEIDGKTYQVKPIRNLNGHSIGQYRIHAGKTVPIVKGGEATRMEEGEV
YAIETFGSTGKGVVHDDMECSHYMKNFDVGHVPIRLPRTKHLLNVINENFGTLAFCRRWL
DRLGESKYLMALKNLCDLGIVDPYPPLCDIKGSYTAQFEHTILLRPTCKEVVSRGDDY

**Figure IV-60: Signature peptides choice and positions in the protein sequence.**

### A.2.2 Sample preparation protocol optimization

Next we tried different sample preparation protocols to asses which one would give the best results in terms of sample extraction and protein coverage. MetAP1 and MetAP2 are cytosolic proteins thus the protocols were chosen

to try to fully extract soluble proteins. Three protocols were tested. The first two were protocols routinely used by our collaborators. The third one was the liquid digestion protocol used at the LSMBO. K562 cell lines were used for this test as it was expected that out of the three cell lines tested, the protein abundance would be highest in this sample.

- Single-band resolving gel: Briefly cells were disrupted using glass beads and the extracted proteins were loaded onto a SDS-PAGE gel. The electrophoretic migration was stopped as soon as the protein sample entered the resolving gel. The proteins were digested using trypsin (Overnight, 37°, enzyme:protein ratio 1:100). A detailed description can be found in Experimental section D.1 on page 220.
- Liquid digestion protocol 1: Briefly cells were disrupted using glass beads and the extracted proteins were precipitated using acetone (4 to 1 volumes of acetone, -20°C, 2 hours). The proteins were resuspended using ammonium acetate buffer. The trypsin digestion was done using a two-step protocol (2x[1h30, 37°, enzyme:protein ratio 1:100]) to avoid non-specific trypsin cleavages. Then, samples were desalted and concentrated using solid phase extraction. A detailed description of this protocol can be found in Experimental section D.2 on page 220.
- Liquid digestion protocol 2: Briefly cells were disrupted using needle sonication and the extracted proteins were precipitated using acetone (9 to 1 volumes of acetone, -20°C, overnight). The proteins were resuspended using urea buffer. The trypsin digestion was performed overnight (1:120 enzyme:protein ratio, 37°C, overnight). Then samples were desalted and concentrated using solid phase extraction. A detailed description of this protocol can be found in D.3on page 221.

The results of this assessment can be found in Figure IV-61.

A.

**Single-band resolving Gel:**
- Glass bead disruption
- 10µg loaded on gel
- Migration until the sample is in a single band in the resolving gel.
- Trypsin Digestion (Overnight, 37°, 1:100)

**Liquid digestion protocol 1:**
- Glass bead disruption
- 2-hour acetone precipitation
- Resuspended in ammonium bicarbonate buffer
- Trypsin Digestion 2x(1h30, 37°, 1:100)

**Liquid digestion protocol 2:**
- Needle sonication
- Overnight acetone precipitation
- Resuspended in urea buffer
- Trypsin Digestion (Overnight, 37°, 1:120)

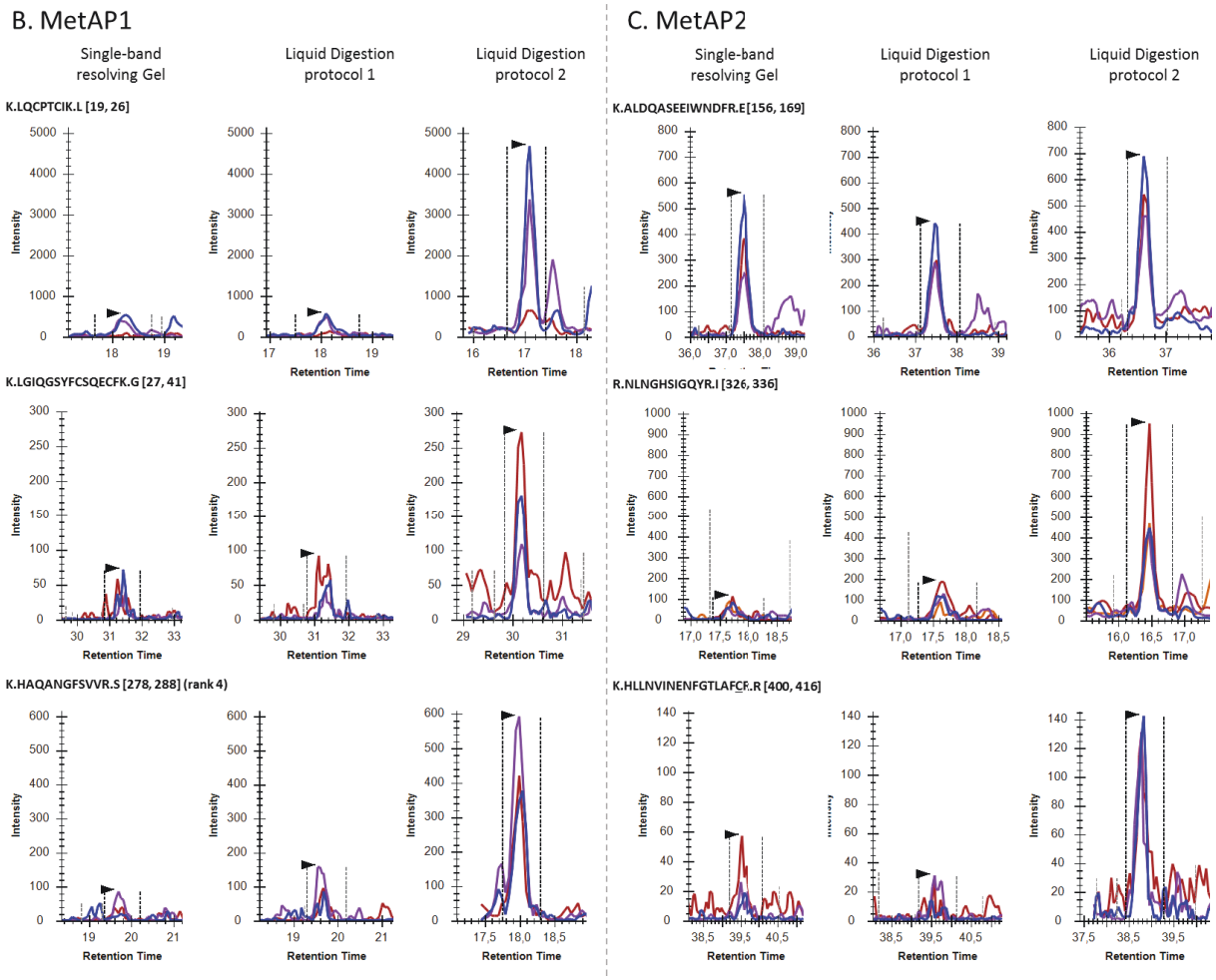



**Figure IV-61: Sample preparation evaluation results.**
A. Overview of the different protocols tested. The chromatograms of endogenous peptides are shown using the three different sample preparation protocols for protein MetAP1 (B) and MetAP2 (C).

The chromatograms of endogenous peptides are shown using the three different sample preparation protocols for protein MetAP1 and MetAP2. For all three protocols the peptide ALDQASEEIWNDFR of protein MetAP2 can be observed with similar intensities. This proves that the protein could be extracted correctly using the three protocols. However, for all the other peptides the signal was clearly most intense when using the liquid digestion protocol 2. This could be explained by an incomplete enzymatic digestion when using short time periods. The idea behind using a two-step trypsin digestion was to avoid non-specific digestion by self-digested trypsin still active, possibly having different specificity. However the digestion did not generate enough peptides to confidently quantify the protein. For the SDS-PAGE the problem was possibly that the quantity of loaded protein (10µg) was not sufficient. We found that the best results are found using a stacking gel protocol when loading more than 50µg. In conclusion, the liquid digestion protocol 2 was chosen for the rest of the project based on these results. Collision energies for each transition were optimized using the heavy labeled synthetic peptides to achieve highest sensitivity.

### A.3. Relative and absolute quantification results

To verify the hypothesis formulated by our collaborators, a first relative quantification experiment was performed using low-purity crude standard peptides. Three cell lines were compared: HUVEC, U87 and K562. The samples were prepared in quadruplicates and injected once. The results can be seen in Figure IV-62.

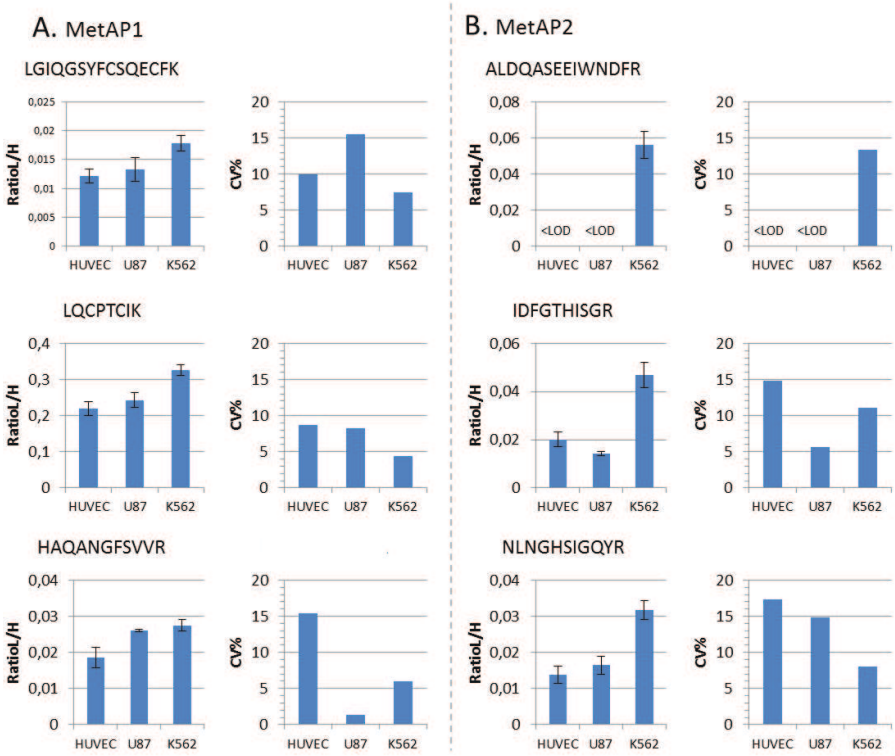



**Figure IV-62: Results of the relative quantification experiment.**
Ratios of the summed areas of all transitions of the light over the heavy-labelled peptide (L/H) and coefficients of variation (CV%) for sample preparation quadruplicates for the 6 targeted peptides of MetAP1 (A) and MetAP2 (B) in three different cell types.

Out of the twelve chosen signature peptides only three peptides per protein were validated. The peak group identification and integration of SRM traces were manually verified, checking the exact coelution and the correct relative fragment-ions intensities between light and heavy-labelled peptides. Overall, all signature peptides showed coefficients of variation lower than 20% with a mean value of 10,2% and a median value of 9,4%. These were calculated using light/heavy area ratios for the sum of all transitions. Since we worked with sample preparation quadruplicates these results prove that the whole sample preparation protocol is very reproducible. The results show that the abundance of MetAP1 and MetAP2 are significantly lower in the sensitive cell lines (U87 and HUVEC) compared to the insensitive cell line (K562).

To validate this trend we carried out an absolute quantification experiment using highly purified and precisely quantified peptides for MetAP2. Two AQUA peptides, ALDQASEEIWNDFR and IDFGTHISGR, were used to monitor MetAP2. However, only peptide IDFGTHISGR could be detected in the samples. A calibration curve was done using a pool of the three cell types. Six calibration points were created: 1.6, 3.1, 6.3, 12.5, 25 and 50 fmol of injected peptide into the column in 10µg of total digest. The results of this study can be seen in Figure IV-63.

The limit of quantitation (LOQ) was defined as the last point having a coefficient of variation lower than 20% among triplicate injections, showing an accuracy between 80 and 120% and giving a coefficient of determination $R^2$ higher than 0,98 between the summed area under the curve and the injected amount on column, and between the recalculated and the real injected amount on column. In this case the LOQ was found to be 3.1 fmol of injected

peptide into the column in 10µg of total digest. All calibration points above had a CV lower than 15%. The coefficient of determination $R^2$ for the two regressions mentioned above was 0.998.

As a conclusion, targeted proteomics allowed confirming the extremely low abundance of MetAP2 in all three cell lines. The limit of quantification was determined to be 310 amol per µg total proteins. For the K562 the abundance of MetAP2 was determined to be 350 amol per µg of total protein. For U87 and HUVEC, a precise estimation could not be done as their abundances were below the limit of quantification.

Furthermore both independent quantification studies showed the same trend: MetAP2 is more abundant in insensitive cell lines than in sensitive cell lines. These results are well correlated with an independent quantification performed by our collaborators of mRNA of the same proteins. These results also show a higher abundance of METAP 1 and 2 in insensitive cell lines.

An important result of this quantification was that mRNA and protein expression are well correlated for the two MetAP proteins. Thusly in future experiments mRNA could be used as proxies for these two proteins.

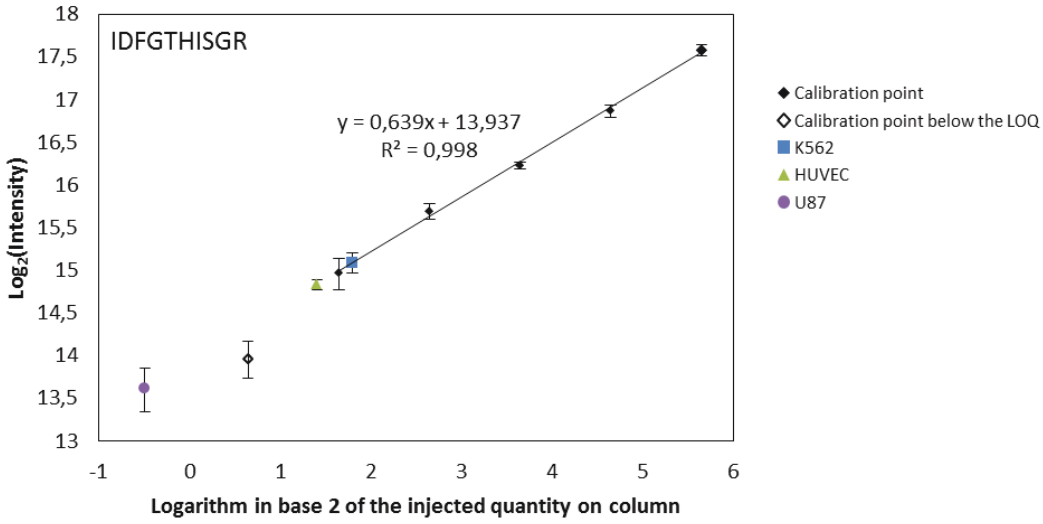



**Figure IV-63: Results of the absolute quantification of MetAP2 in three cell lines.**
The calibration curve is shown for the absolute quantification of METAP2 by monitoring the IDFGTHISGR peptide. The calibration curve was done using the points where CVs were lower than 15% and the accuracy was between 80-120% (full diamonds), the lowest point being the limit of quantification (LOQ). The points not meeting these criteria were discarded (empty diamonds).

### *A.4. Conclusion and perspectives*

The development of an LC-SRM assay enabled the detection and relative quantification of proteins, MetAP1 and MetAP2, that could not be detected in a global shotgun experiment and using an immunodetection assay. The results showed a higher abundance of METAP 1 and 2 in insensitive cell lines, as expected. These results were well correlated with an independent mRNA quantification experiment performed by our collaborators. An absolute quantification of MetAP2 could also be achieved for one of the three cell lines and confirmed the same trend.

In order to be able to quantify the two proteins more accurately, the protein dynamic range should be reduced. A protein depletion protocol could be used to eliminate highly abundant proteins. Since the two proteins have similar molecular weights, 43kDa and 53kDa for MetAP1 and MetAP2 respectively, a SDS-PAGE separation protocol could be used to “roughly isolate” the proteins, reduce the background noise and lower the total protein dynamic range in the sample.

A publication resuming the results of this project is in preparation.

# Part V Results of proteogenomics analysis

## Chapter I Engineering an automated N-terminomics workflow

### A. N-TOP: N-terminal Oriented Proteomics

The N-Terminal oriented proteomics approach (N-TOP) was initially developed at the LSMBO by Sebastien Gallien [223]. It is based on a selective derivatization of α-amine group by TMPP.

TMPP was historically used to increase ionization efficiency in mass spectrometry using fast atom bombardment ionization [224], MALDI [225] and electrospray ionization [226]. The structure of TMPP can be seen in Figure V-1. TMPP improves the ionization efficiency in electrospray because of its hydrophobicity and its permanent positive charge. Hydrophobicity of the reagent increases the ionization efficiency as this means that TMPP-derivatized peptides have an increased affinity for the non-polar electrospray droplet surface [227-229] resulting in more successful competition for excess charge and higher ESI response.

One of the important results from the work of Gallien et al. is the optimized experimental conditions to profit from the difference in pKa of α-amines and the ε-amino group in lysine residues to achieve a good regioselectivity of the derivatization reaction. The regioselectivity is controlled by strictly setting the pH at 8.2 (Figure V-1). This gives selectivity of 78-95% on the free N-terminal amino group (pKa ~7.8) relative to the protonated ε-amino group of the lysine side chain (pKa~11) [225].

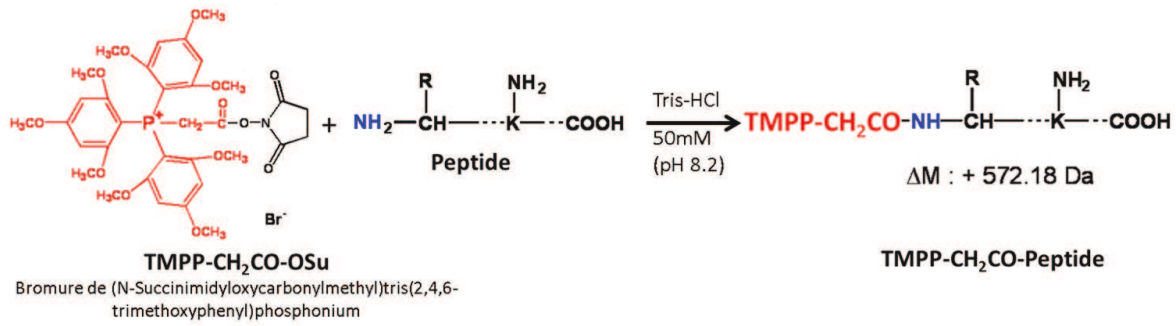



Figure V-1 : TMPP selective derivatization towards N-terminal amines.

The analytical workflow for the N-TOP approach, illustrated in Figure V-2, consisted in a TMPP-derivatization at the protein level after reduction and alkylation, followed by SDS-PAGE separation in order to remove the excess TMPP and decomplexify the samples. Proteins were then in-gel digested and analyzed by LC-MS/MS.

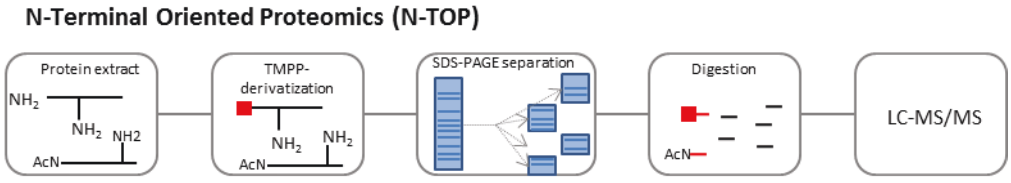



Figure V-2 : Analytical workflow of the N-Terminal Oriented Proteomics (N-TOP) approach

The hydrophobicity of the TMPP reagent induces a shift in retention time towards more hydrophobic regions of the chromatogram. The distribution of retention times of TMPP-labelled and non-labelled peptides can be seen in Figure V-3. The two distributions are well separated and TMPP-labelled peptides are eluted later in the gradient. The separation of TMPP peptides from internal peptides favors their sampling in data dependent acquisition MS/MS experiments and enhances the sensitivity since the complexity in this part of the chromatogram is lower. This behavior added to the efficiency in ESI ionization highly increases the overall sensitivity of N-terminal peptides.

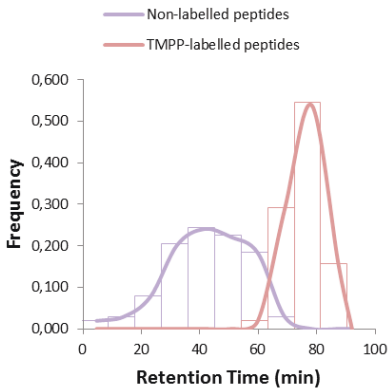



**Figure V-3 : Distribution of retention times for non-labelled and TMPP-labelled peptides.**
The retention times of labelled and un-labelled peptides are shown. The two distributions are well separated and TMPP-labelled peptides are eluted late in the gradient. This helps to increase the sensitivity of N-terminal peptides.

However, TMPP is also known to give a specific fragmentation pattern. TMPP-derivatized peptides are fragmented in CID by a charge-remote fragmentation mechanism [223, 226, 230, 231]. The fixed and permanent positive charge induces a charge-remote fragmentation of derivatized peptides that facilitates peptide fragmentation for MS/MS, constituted predominantly by a- and b-type ions. Compared to a non-derivatized peptide, a spectrum of TMPP-derivatized peptides has few y-ions at the lower mass range. This is expected because only y-ions can be present below 630 m/z (the mass of the a-type fragment corresponding to the TMPP-labeled glycine). An example of a spectrum from a TMPP-derivatized peptide can be seen in Figure V-4.A. This particular pattern of fragmentation is specific to the TMPP-labeled peptides. For these peptides, in spite of high-quality fragmentation spectra, Mascot ion scores are generally lower than for unlabeled peptides because few fragments remain in the lower mass range of the MS/MS spectra and thus creating a bias in the scoring.

As a result of the scoring problem, each spectrum of a TMPP-derivatized peptide had to be individually manually curated. This was a tedious and time-consuming step that was highly prone to errors and subjective to the user interpretation. This remained for some years the bottle-neck of the N-TOP approach.

## B. dN-TOP: doublet N-terminal Oriented Proteomics

The results presented here were published in June 2013 in the *Journal of Proteomic Research* [28] and the newly optimized workflow was published in April 2015 in the Protein N-terminal Biology special issue in *Proteomics* [159] (available at the end of this section).

To overcome the scoring problem of TMPP-derivatized peptides a strategy was developed using a light and heavy isotopically stable version of the TMPP. The approach named doublet N-terminal Oriented Proteomics (dN-TOP) consist in labelling N-terminal peptides with an equimolar mixture of light and heavy TMPP.

Light and heavy TMPP-derivatized peptides had the same behavior in liquid chromatography and mass spectrometry. They only differ their mass corresponding to the isotopically stable tag, i.e. nine $^{13}C$ atoms for the heavy TMPP. The

doublet identification of a peptide derivatized with the light TMPP and the identification of that same peptide but derivatized with the heavy TMPP validates its presence in the sample.

A first try at an automated validation workflow was developed. A two-stage database search pipeline was implemented. The first search round was a classic proteomics search of fully-tryptic peptides validated at 1% FDR. The second search round enabled the identification of TMPP-derivatized peptides. We then developed a macro to search for doublets of N-terminal peptides derivatized with light and heavy TMPP.

This workflow was applied as a proof of concept to the characterization of the N-terminome of a bacterium, *Herminiimonas arsenocoxydans*. 650 unique protein groups could be identified using the internal peptides and 90 N-terminal positions were found. Of these, 77 were correctly annotated in the database and 13 N-terminal positions enabled the correction of the gene annotation or could be explained by proteolysis. However this bacterium is of low complexity: its genome consists of 3,4 mega base pairs almost 950 times smaller than the human genome [232]. To be able to tackle highly complex samples an optimized automated and reliable validation workflow needed to be developed.

### *B.1. Automated validation workflow*

The results presented here were published in April 2015 in the Protein N-terminal Biology special issue in *Proteomics* [159] and available at the end of this section.

To be able to have a reliable and automated workflow allowing high-throughput analysis of samples coming from complex organisms some problems had to be resolved.

- The problem of a possible double interpretation of the same spectrum in the two search rounds.
- The problem of an ambiguous spectral interpretation.
- The problem of an ambiguous TMPP labelling position either in α-amines or ε-amines.
- The problem of shared peptides between different proteins.

These problems were solved by engineering a data analysis workflow that is now completely automated, accurate and user-friendly. The final workflow can be seen in Figure V-5. The important steps of the workflow are described below.

To tackle the problem of a possible double interpretation of the same spectrum in the two search rounds, high-quality spectra were extracted from all non-assigned spectra after the first search round using the in-house developed Recover module of MSDA [233]. This tool extracts spectra using user-defined filters and creates a peaklist made of high-quality spectra. Using this tool spectra were kept if they had at least 4 peaks higher than 3 times the intensity of the background noise and at least one peak above the m/z ratio of the precursor ion. Recover was also used in our workflow to eliminate spectra of singly charged precursor, spectra with unassigned charge states and spectra already used to identify a validated peptide during the first search round .

When analyzing samples of high complexity the problem of a possible ambiguous spectral interpretation arose. Normally for a given spectrum a search engine scores all possible interpretations of that spectrum and ranks them. Then data interpretation software, as Scaffold [199], MaxQuant [87] or Proline (Proline Studio, ProFI, Proteomics French Infrastructure), uses this information and for each spectrum attributes the first ranked peptide sequence to the spectrum. Figure V-4.A. shows a spectrum giving a non-ambiguous identification. In this example the spectrum has enough information to identify one single peptide sequence. An example of an ambiguous interpretation can be seen in Figure V-4.B. In this example the spectrum is not informative enough to discriminate between several possible

peptide sequences present in the database. In this case MASCOT gives the same ranking value to all peptides with identical scores. This ranking value is named the pretty rank. The pretty rank is a value similar to the rank except that equivalent scores get equivalent ranks. In early stages of my PhD, I used the Scaffold software to validate MS-MS/MS results. However, I found out that in the case of an ambiguous identification, Scaffold chose the first peptide sequence in alphabetical order. So in the example the peptide sequence AAAVSPSK would be kept. This biased selection compromised the reliability of the identifications. The Proline software does not make a decision about these peptides and keeps all interpretations with a pretty rank equal to 1. We used this value to detect and eliminate ambiguous spectra. The use of pretty ranks also allows detecting and eliminating spectra for Tyr- or Lys-containing peptides for which the exact TMPP labelling position (α-amino terminal group or ε-amino group of Tyr or Lys) could not be determined (Figure V-4.C). The use of the pretty rank is a criterion of great importance in the reliability of the results in the output of the automation tool.

Finally, the retention times of light and heavy-TMPP derivatized peptides are compared and the difference must be lower than a user-defined value. The information of the unicity in the searched database is also indicated in the output file. This automation tool enabled to reduce the time necessary to confidently validate high-throughput analysis of N-terminomics data using the dN-TOP approach.

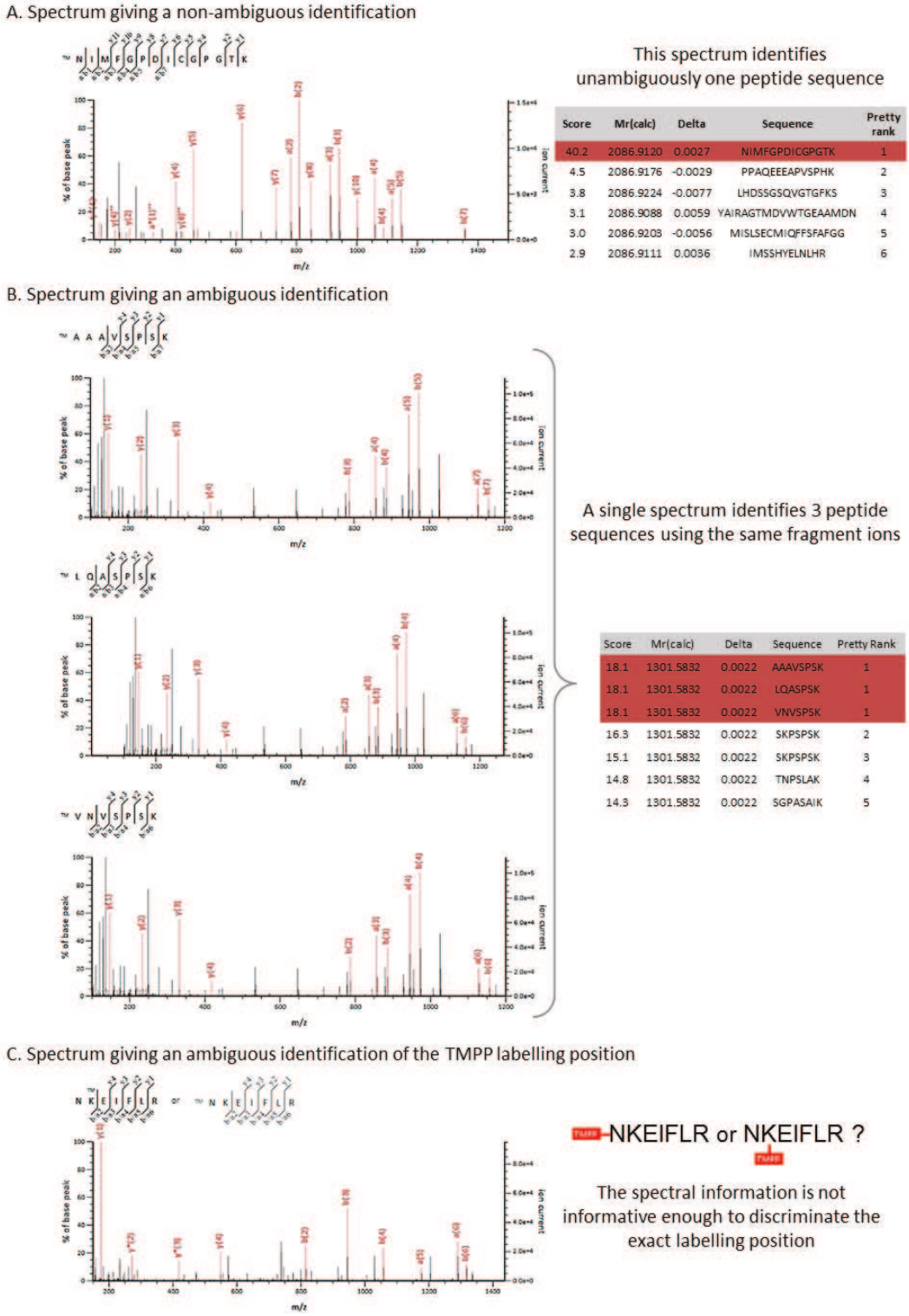



**Figure V-4 : Examples of spectra giving ambiguous and non-ambiguous peptide identifications**

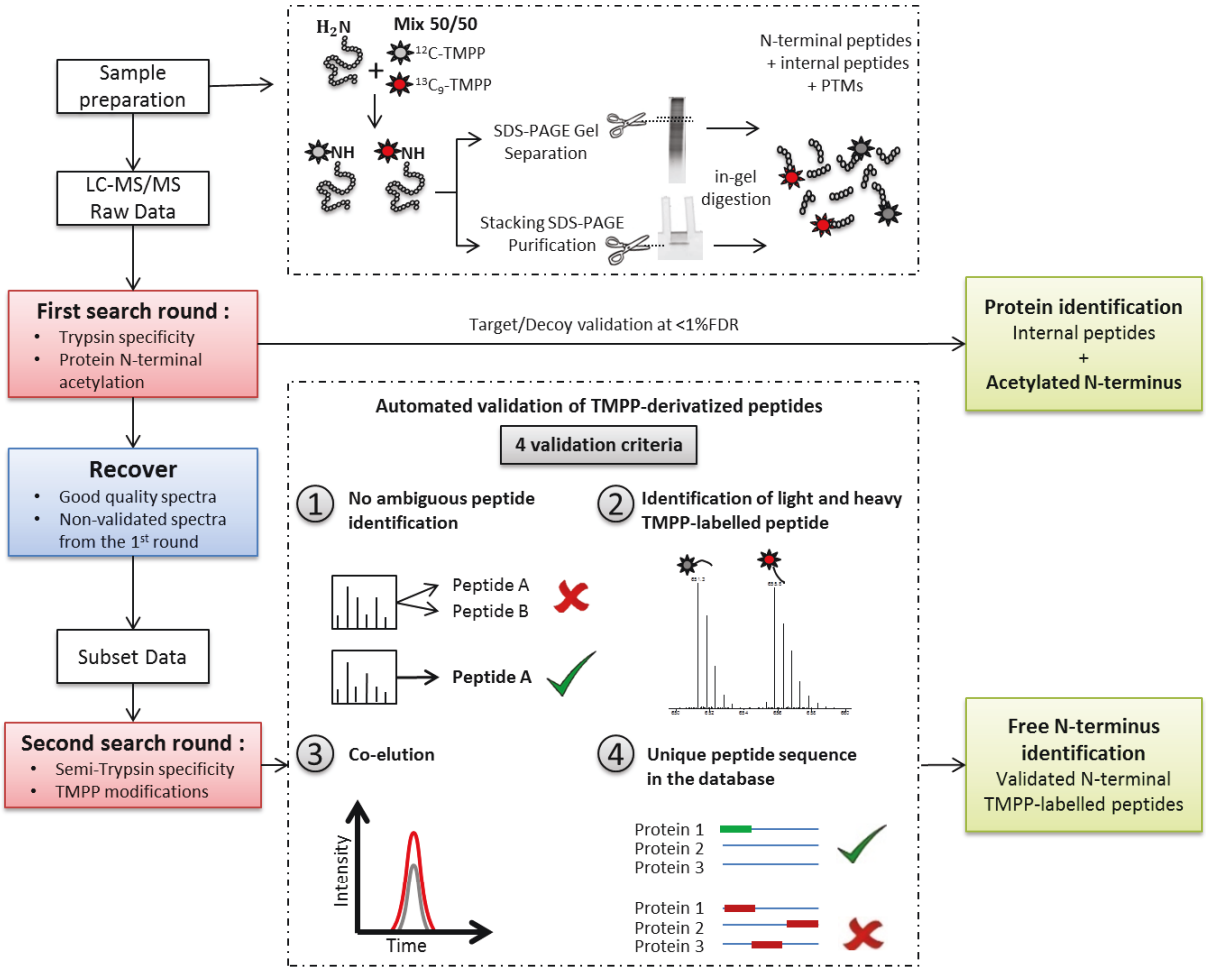



**Figure V-5 : Overview of the analytical workflow for sample preparation and the data validation strategy.**

### *B.2. Advantages of the optimized dN-TOP approach*

Here is a summary of the advantages of using the dN-TOP approach:

- dN-TOP uses a simple SDS-PAGE-LC-MS-MS/MS workflow with just an additional labelling step at the protein level.
- dN-TOP can be used to analyze extensively fractionated samples (SDS-PAGE separation) or unfractionated samples (Stacking Gel SDS-PAGE).
- dN-TOP allows analyzing in a single experiment both N-terminal and internal peptides unlike most alternative N-terminomics approaches focusing only on N-terminal peptides. Among others, it even allows identifying acetylated N-termini.
- dN-TOP does not need dedicated HPLC system and all reagents are commercially available and inexpensive.
- TMPP increases the electrospray ionization efficiency of derivatized peptides.
- TMPP reagent induces a shift in retention time towards more hydrophobic regions of the chromatogram, thus shifts N-terminal peptides to regions usually of low complexity, favoring their sampling in data dependent acquisition MS/MS experiments. This increases the overall sensitivity for derivatized peptides
- The use of a pair of labeling reagents containing the light and an isotopically heavy-labelled form of TMPP allows establishing an unambiguous, reliable and automated identification strategy of N-terminal peptides.
- Double interpretation of a same spectrum is prevented.

- The use of the pretty rank in the automation tool avoids the problem of ambiguous spectral interpretation.
- The use of the pretty rank in the automation tool avoids the problem of ambiguous labelling at α- or ε-amine groups.

## C. N-terminome analysis of the human mitochondrial proteome

This project was carried out in collaboration with the Laboratory of Chemistry and Biology of Metals of the CNRS in Grenoble, and particularly with Thierry Rabilloud.

And in collaboration with the Computer Analysis and Laboratory Investigation of Proteins of Human Origin (CALIPHO) Group, of the Swiss Institute of Bioinformatics, and particularly with Lydie Lane and Amos Bairoch , for the integration of the data into the UniProtKB/Swissprot knowledgebase.

The dN-TOP approach was applied to the characterization of the proteome and N-terminome of human mitochondria.

Mitochondria are very important organelles implicated in vital functions such as bioenergetics, protein folding and degradation, metabolism of amino acids, lipids, heme and iron, signaling and apoptosis [234]. Mitochondria are thus essential for eukaryotic cells.

Mitochondria are believed to have evolved from endosymbiosis of a primitive bacterial cell [235]. From this evolutionary past, 13 human proteins are encoded by mitochondrial DNA. Besides these 13 proteins, it is estimated that around 1200-1500 proteins are located in the mitochondria. These proteins are synthesized as precursor forms in the cytosol and must be imported into mitochondria. As a result of this process, for most mitochondria addressed proteins, a transit peptide, 10 to 100 amino-acids long, is cleaved while passing mitochondrial membranes. However, the exact position of the cleavage sites is not known for most mitochondrial proteins (Figure V-6). The majority of knowledge on the exact position of the cleavage sites comes from predictions and homologies using prediction algorithms [236, 237]. Knowing the exact start position of mature and active proteins inside the mitochondria is important as N-terminal start positions have an impact on protein half-life, protein function and a defect in the process have been shown to be involved in human diseases [238]. Yeast has been used as a model to understand dysfunctions in the mitochondria [239] but human cells contain protease complexes that are not present in yeast [235]. More experimental data is therefore urgently needed using human mitochondria samples.

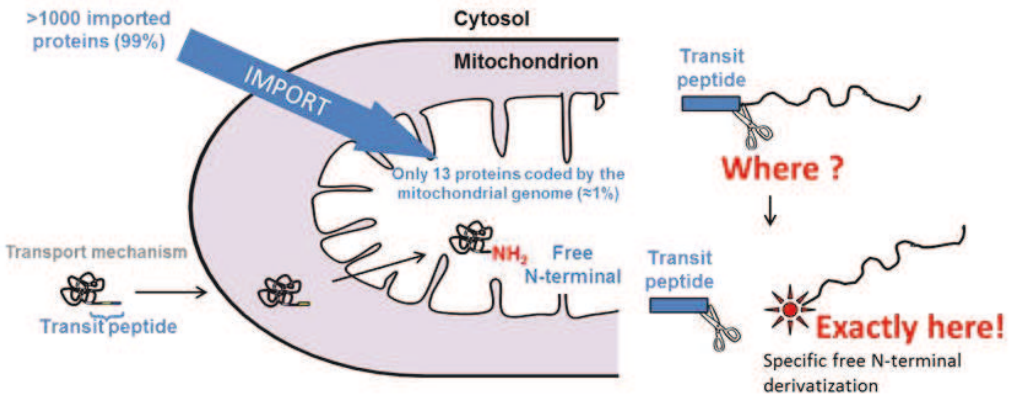



**Figure V-6 : Overview of protein import into mitochondria.**

We applied the dN-TOP approach to the characterization of the proteome and N-terminome of human mitochondria. In the first publication of 2015, I analyzed human mitochondria enriched samples and collected data acquired

between 2010 and 2014. In total, I compiled data from 12 experiments, obtained in low- and high-resolution Q-TOF instruments, with ETD and CID fragmentation and using Asp-N and trypsin digestion. A total of 2714 protein groups and 897 N-terminal peptides were identified (424 N-α-acetylated and 473 TMPP-labelled peptides). Information of the exact position of the N-terminus could be obtained for 26% of all identified proteins (693 unique proteins). 120 already annotated processing cleavage sites were confirmed while 302 new cleavage sites were identified.

In the publication recently compiled in 2016, we have extended these results by using a high-resolution Orbitrap instrument to analyze unfractionated and fractionated mitochondria-enriched samples. Compared to the results of the previous publication, the total number of identified proteins was multiplied by a factor of 2 and the number of unique N-terminal positions identified was multiplied by a factor of 4,4. We identified 4655 protein groups of which 963 are annotated as being mitochondrial proteins. We managed to identify 2740 unique N-terminal positions (687 N-α-acetylated and 2067 TMPP-labelled peptides) which provided N-terminus information for 35% of all identified proteins.

These two datasets provide valuable information as only a small part of the processing cleavage site positions were annotated in Human UniProtKB/SwissProt database, 28% for the first dataset and 12% for the second.

Figure V-7 shows the results for the second dataset, in total only 223 N-terminal positions matched with an annotated processing cleavage with experimental evidence, i.e. 49 canonical protein free N-termini (2,6%), 57 methionine cleavage (3%) and 117 processing cleavages sites (6,2%). Some predicted positions (annotated as “By similarity”, “potential” or “Probable”) were confirmed by our study (6 methionine cleavage (0,3%) and 68 processing cleavage sites (3,6%)). The exact position was corrected from an erroneous annotation for 109 processing cleavage sites (5,8%). Finally, we identified 45 unannotated methionine cleavage sites (2,4%) and 1439 (76,1%) of new processing cleavage site positions.

Among all the identified start positions, 791 were identified in the region where transit and signal peptides are expected (between position 2 and 100 in the protein sequence) (Table 1). 515 of these positions are new non-annotated cleavage sites and 197 of these are annotated either as mitochondrial proteins (138 start positions) or belonging to the endoplasmic reticulum (59 start positions). These are possible new transit/signal peptide cleavage sites due to translocation/processing events.

The first dataset contributed to annotate the Human UniProtKB/Swissprot database : our work was included in the Release 2015_10 (14$^{th}$ October 2015), in which 36 N-terminal acetylations (position 1), 153 N-terminal acetylations (position 2 + methionine initiator), 22 methionine initiator sites, 43 transit peptides, 29 signal peptides, 2 propeptides and 1 sequence variant were included.

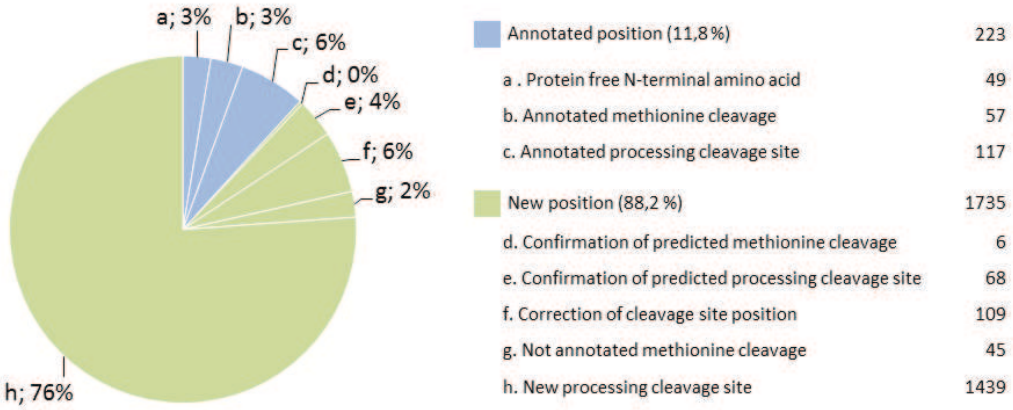



**Figure V-7: Distribution of N-terminal positions identified with the dN-TOP approach.**

## D. Conclusions and perspectives

The results presented here were published in April 2015 in the Protein N-terminal Biology special issue in *Proteomics* [159] (available below) and an expanded publication of the results is in preparation.

The newly engineered dN-TOP approach is now a powerful tool for reliable and accurate high-throughput N-terminomics. It was used to deeply characterize the human mitochondria N-terminome and proteome. This workflow will now be used for a quantitative analysis of N-terminal peptides in different stress conditions in order to expand the understanding of the powerhouse of the cell.

## E. Publications

Journal of proteome research

Technical Note

pubs.acs.org/jpr

# An Improved Stable Isotope N-Terminal Labeling Approach with Light/Heavy TMPP To Automate Proteogenomics Data Validation: dN-TOP

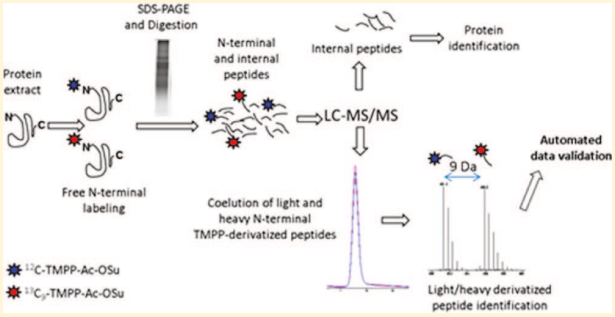
Diego Bertaccini,[†] Sebastian Vaca,[†] Christine Carapito,[†] Florence Arsène-Ploetze,[‡] Alain Van Dorsselaer,[†] and Christine Schaeffer-Reiss[†,*]

[†]Laboratoire de Spectrométrie de Masse BioOrganique, IPHC, Université de Strasbourg, CNRS, UMR7178, Strasbourg, France

[‡]Laboratoire de Génétique Moléculaire, Génomique et Microbiologie, Université de Strasbourg, CNRS UMR7156, Strasbourg, France

Supporting Information

**ABSTRACT:** *In silico* gene prediction has proven to be prone to errors, especially regarding precise localization of start codons that spread in subsequent biological studies. Therefore, the high throughput characterization of protein N-termini is becoming an emerging challenge in the proteomics and especially proteogenomics fields. The trimethoxyphenyl phosphonium (TMPP) labeling approach (N-TOP) is an efficient N-terminomic approach that allows the characterization of both N-terminal and internal peptides in a single experiment. Due to its permanent positive charge, TMPP labeling strongly affects MS/MS fragmentation resulting in unadapted scoring of TMPP-derivatized peptide spectra by classical search engines. This behavior has led to difficulties in validating TMPP-derivatized peptide identifications with usual score filtering and thus to low/underestimated numbers of identified N-termini. We present herein a new strategy (dN-TOP) that overwhelmed the previous limitation allowing a confident and automated N-terminal peptide validation thanks to a combined labeling with light and heavy TMPP reagents. We show how this double labeling allows increasing the number of validated N-terminal peptides. This strategy represents a considerable improvement to the well-established N-TOP method with an enhanced and accelerated data processing making it now fully compatible with high-throughput proteogenomics studies.




**KEYWORDS:** *N-terminome analysis, proteogenomics, TMPP derivatization, automated data validation*

## INTRODUCTION

In a mass spectrometry-based proteomic discovery experiment, protein identifications are achieved by matching the experimentally obtained spectra with the theoretical mass lists obtained by *in silico* digestion and fragmentation of the protein sequences available in databases. This well-known workflow relies on the assumption that the database is an error free, exhaustive list of all proteins coded by a genome. The reality is far from this assumption since protein databases are mainly obtained by *in silico* translation of the genome sequence. An approach that has proven to be prone to errors and to generate incomplete protein data sets.[1−3] The consequence of this can be dramatic as it can affect any biological experiment that has been based on it.

UniProtKB/SwissProt[4] is a curated protein sequence database in which each entry gets thoroughly analyzed and annotated by expert curators ensuring a high standard of annotation and maintaining the quality of the database.[5] When considering the last reported global statistics of UniProtKB/SwissProt, only 14.2% of all entries have evidence at protein level; 70.1% are inferred from homology; 12.7% have evidence at the transcription level; the predicted entries represent 2.7% and the uncertain entries 0.4% (released the 31-Oct-12).

Among the common errors introduced by *in silico* predictions and propagated by ortholog alignments, the incorrect prediction of initiation codon is particularly pervasive as, up to now, any bioinformatics tool is able to properly estimate with high confidence all initiation sites of the mature proteins;[6,7] this is especially the case for prokaryotic genomes with high GC content since they are characterized by many long open reading frames that are not genic.[8] Based on this evidence, the necessity to collect experimental data to assess and refine the quality of the genome annotation becomes obvious and the development of proteogenomic approaches urgent. In this context, proteomics data are unique resources and can improve many of the problematic areas of genome annotation, like the start site assignment.

To maximize the number of identified N-terminal sequences, the classical high-throughput proteomic workflow has been

**Received:** April 3, 2013
**Published:** May 6, 2013

implemented with many complementary approaches able to specifically target the protein N-termini, based on chemical derivatization of the N-terminal function.[9]

The TMPP labeling approach (N-TOP approach) is an efficient N-terminomic approach, that allows the characterization of both N-terminal and internal peptides in a single experiment and has been applied very successfully to various proteomes such as *Mycobacterium smegmatis* and *Sterolibacterium denitrificans* in our laboratory[10,11] and by others.[12] This now well-established N-TOP method is based on a N-terminal protein labeling performed with (*N*-succinimidyloxycarbonylmethyl) tris (2,4,6-trimethoxyphenyl) phosphonium bromide (TMPP-Ac-OSu) on a total biological extract and is fully compatible with all standard detergents, chaotropic agents, and reduction conditions used for protein extraction in proteomics. Two characteristics of this labeling reagent promote the sensitivity of the method: (i) TMPP labeling introduces a permanent positive charge resulting in an enhanced ionization efficiency and thus a better detection of low-abundance proteins; (ii) the hydrophobic TMPP group shifts the retention time of derivatized peptides in reversed phase chromatography toward a less complex part of the chromatogram, therefore, increasing the sensitivity of detection (including the possibility to detect short N-terminal peptides that otherwise would not be retained on the column).

Besides the fact that this approach allows maintaining intact all internal proteolytic peptides, its easy experimental design is the major advantage of this approach: a single chemical derivatization step, performed at the protein level, that can easily be integrated in a classical 1D SDS-PAGE/LC-MSMS proteomics workflow, without requiring any other additional step like immune capture or multidimensional chromatography. Nevertheless, one limitation of the approach resides so far in the validation of labeled peptides, as they present unusual fragmentation patterns. It is well-known that low energy peptide fragmentation (post source decay (PSD) or collision induced dissociation (CID)) is obtained thanks to the delocalization of a proton on the peptidic backbone generating mainly y- and b-type ions.[13]

Alternatively, a peptide labeled with a chemical tag that carries a fixed charge, a permanent positive charge in the case of a TMPP derivatization, behaves in the mass spectrometer in a completely different way; all fragments are generated with a charge remote mechanism that results in a massive production of uncommon ions.[14] The TMPP labeling significantly enhances a- and b-type ions that are usually missing in tryptic peptide MS/MS spectra. Since the search algorithms have been developed and educated for classical fragmentation patterns, a TMPP-derivatized peptide will not be assessed with an optimal score, resulting in a too stringent filtration when operated by the target/decoy approach with 1% FDR and thus in underestimated validation of N-terminal peptides.

To overcome those difficulties, we have developed a new method allowing an easy, reliable and automated TMPP-derivatized peptides' validation based on a stable isotope labeling experiment, a widely applied method in quantitative proteomics.[15−17]

For this purpose, a $^{13}$C-labeled analog of the TMPP reagent was designed and a double labeling was performed (1:1 light and heavy TMPP) allowing to identify doublets of identical N-terminal peptide sequences. We designate this labeling strategy as doublet N-terminal oriented proteomics (dN-TOP).

As proof of concept, we applied this method to a cellular lysate of *Herminiimonas arsenicoxydans*. The 3.4 Mbp single chromosome of this arsenite-oxidizing bacterium has already been sequenced and carefully annotated.[18,19] A previously generated proteome map allowed us to characterize 447 proteins among which 365 proteins are in the soluble fraction, representing 13.6% of the total proteome predicted from the genome sequence for this bacterium. For 5 proteins, proteomic data had allowed correcting 5 start codons, even if no N-terminal labeling strategy was applied.[20] To evaluate the specificity and the labeling kinetics of the new isotopically labeled TMPP compared to the light reagent, we present here the comparison of N-TOP to dN-TOP applied to our model organism *H. arsenicoxydans*.

## EXPERIMENTAL PROCEDURES

Unless otherwise specified, all chemicals were obtained from Sigma Aldrich (St. Louis, MO).

### Growing Conditions and Cell Lysis

*H. arsenicoxydans* was cultivated in a chemically defined medium (CDM) containing 2.66 mM of As(III) ($NaAsO_2$) in the same conditions as previously described.[20] Late exponential phase cells (100 mL) were disrupted as previously described,[20] and the soluble extract was further analyzed.

### Protein Labeling and 1D SDS-PAGE

The protocol used here was carried out according to the original reference paper by Gallien et al. with slight modifications.[10] A batch of heavy labeled (*N*-succinimidyloxycarbonyl-methyl)tris(2,4,6-trimethoxyphenyl)phosphonium bromide ($13C_9$ TMPP-Ac-OSu) was synthetized in collaboration with Alsachim. This $13C_9$-TMPP induces a mass increase of 581.21 Da instead of 572.18 Da for light TMPP on labeled peptides. After reduction and alkylation, an equimolar solution of 0.1 M of $^{12}$C-TMPP-Ac-OSu and $^{13}C_9$-TMPP-Ac-OSu in $CH_3CN$:water (2:8; v/v) was added at a molar ratio of 200:1 to 50 $\mu$g of *H. arsenicoxydans* protein extract solubilized in labeling buffer (50 mM Tris-HCl, 6 M urea, 2 M thiourea, pH 8.2, 1 mM phenylmethylsulfonyl fluoride, 1 mM EDTA, 5 mM TBP (Bio-Rad Laboratories)). Selective N-terminal TMPP derivatization is achieved by a careful control of reaction pH at 8.2, exploiting the weaker basicity of the N-terminal amine relative to the $\epsilon$-amino group of the lysine side chain. After a short mix, the reaction was maintained at room temperature for 1 h. Residual derivatizing reagent was quenched by adding a solution of 0.1 M hydroxylamine at room temperature for 1 h, in order to minimize derivatization of tyrosine residues. N-terminal labeled protein extract was finally supplemented with glycerol at a concentration of 10%. Proteins were then separated on a 12% 1D SDS-PAGE (10.1 cm × 7.3 cm) on a mini PROTEAN (Bio-Rad) apparatus at 10 mA for 20 min and 100 mA until the complete migration of the blue front. After electrophoresis, gels were stained with colloidal Coomassie Blue (BioSafe coomassie stain; Bio-Rad) and whole lanes were systematically cut into 28 bands (5 × 2 mm) using a disposable grid-cutter (The Gel-Company, Tübingen, Germany). Bands were cut into three pieces and in-gel digestion using trypsin (Promega, Madison, WI) was performed overnight at 37 °C after in-gel reduction and alkylation using the MassPrep Station (Waters, Milford, MA). Tryptic peptides were extracted using 60% $CH_3CN$ in 0.1% formic acid for 1 h at room temperature. The volume was reduced in a vacuum centrifuge and adjusted to 10 $\mu$L using 0.1% formic acid in water before nanoLC-MS/


dx.doi.org/10.1021/pr4002993 | J. Proteome Res. 2013, 12, 3063−3070

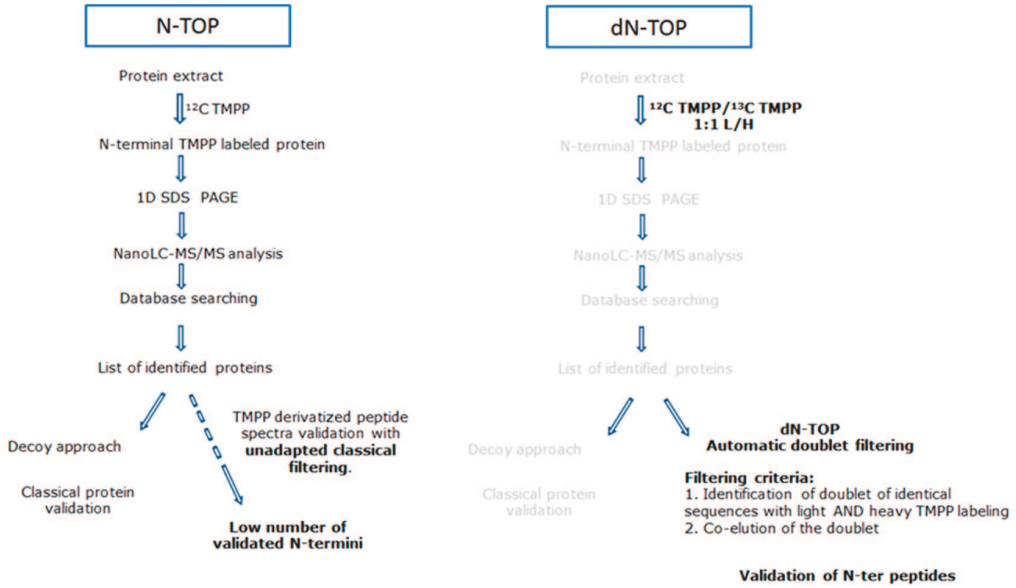



**Figure 1.** Schematic overview of the dN-TOP approach and its improvement steps (in black) when compared to N-TOP, the steps in common are presented in gray in the dN-TOP workflow.

MS (nanoliquid chromatography coupled to tandem mass spectrometry) analysis.

## LC-MS/MS and Data Analysis

NanoLC-MS/MS analyses were performed on a NanoAcquity-LC coupled with a QToF mass spectrometer (maXis 4G, Bruker Daltonics, Bremen, Germany). The UPLC system was equipped with a Symmetry C18 precolumn (0.18 × 20 mm, 5 $\mu$m particle size, Waters, Milford, MA) and an ACQUITY UPLC BEH130 C18 separation column (75 $\mu$m × 200 mm, 1.7 $\mu$m particle size, Waters). The solvent system consisted of 0.1% formic acid in water (solvent A) and 0.1% formic acid in acetonitrile (solvent B). Of each sample 3 $\mu$L was injected. Peptides were trapped during 1 min at 15 $\mu$L/min with 99% A and 1% B. Elution was performed at 60 °C at a flow rate of 450 nL/min, using a linear gradient from 6 to 50% B over 50 min. The mass spectrometer was operating in positive mode, with the following settings: source temperature was set to 160 °C while dry gas flow was at 5 L/min. The nanoelectrospray voltage was optimized to −5000 V. External mass calibration of the TOF was achieved before each set of analyses using Tuning Mix (Agilent Technologies, Paolo Alto, CA) in the mass range of 322−2722 $m/z$. Mass correction was achieved by recalibration of acquired spectra to the applied lock masses (methylstearate ([M + H] + 299.2945 $m/z$) and hexakis-(2,2,3,3,-tetrafluoropropoxy)phosphazine ([M + H] + 922.0098 $m/z$)). For tandem MS experiments, the system was operated with automatic switching between MS and MS/MS modes in the range of 100−2500 $m/z$ (MS acquisition time of 0.4 s), MS/MS acquisition time between 0.05 s (intensity >250 000) and 1.25 s (intensity <5000). The 6 most abundant peptides (absolute intensity threshold of 1500) were selected from each MS spectrum for further isolation and CID fragmentation using nitrogen as collision gas. Ions were excluded after acquisition of one MS/MS spectrum and the exclusion was released after 0.25 min.

Peak lists in mascot generic format (.mgf) were generated using Data Analysis (version 4.0; Bruker Daltonics) and merged for each lane using an in-house developed tool available at https://msda.unistra.fr.

## Internal Peptide Data Processing

MS and the MS/MS data were analyzed using a local Mascot server (version 2.4.1, Matrix Science, London, England). The search were performed against a *H. arsenicoxydans* database composed of all the original entries (created 2013-02-22 and containing 3400 sequences) downloaded from the public available repository (http://www.genoscope.cns.fr/agc/mage/). The reverse sequences of all entries and common contaminants (keratins, trypsin) were added using our in-house toolbox (Mass Spectrometry Data Analysis, MSDA) freely available after registration at https://msda.unistra.fr. Full trypsin enzyme specificity was fixed, carbamidomethylation of Cysteine (+57 Da) and of oxidation of Methionine (+16 Da) were set as variable modifications and mass tolerances on precursor and fragment ions of 10 ppm and 0.02 Da were used, respectively. Mascot results files (.dat files) were uploaded into the Scaffold software (version 3.6.5; Proteome Software Inc., Portland, USA) for identification validation.

The following filtering criteria based on probability-based scoring of the identified peptides were applied in order to obtain a false discovery rate (FDR) <1% based on the number of decoy hits. Peptides having a Mascot Ion scores higher than Mascot's threshold score of identity (95% confidence level) and absolute Mascot Ion scores >25 were validated.

## N-Terminal Labeled Peptide Data Processing

A second Mascot search was performed using semitrypsin enzyme specificity and adding the different TMPP modifications (TMPP N-ter (+572.18 Da), $^{13}$C-TMPP N-ter (+581.21 Da), TMPP derivatization of Tyr and Lys (+572.18 Da) and $^{13}$C-TMPP derivatization of Tyr and Lys (+581.21 Da)) as variable modifications when compared to the first search.

The Mascot results files (.dat files) were uploaded into the Scaffold software and directly exported, without ion score filtration, in an excel file. An automation tool, Validor freely available at https://msda.unistra.fr in the download software


dx.doi.org/10.1021/pr4002993 | *J. Proteome Res.* 2013, 12, 3063−3070

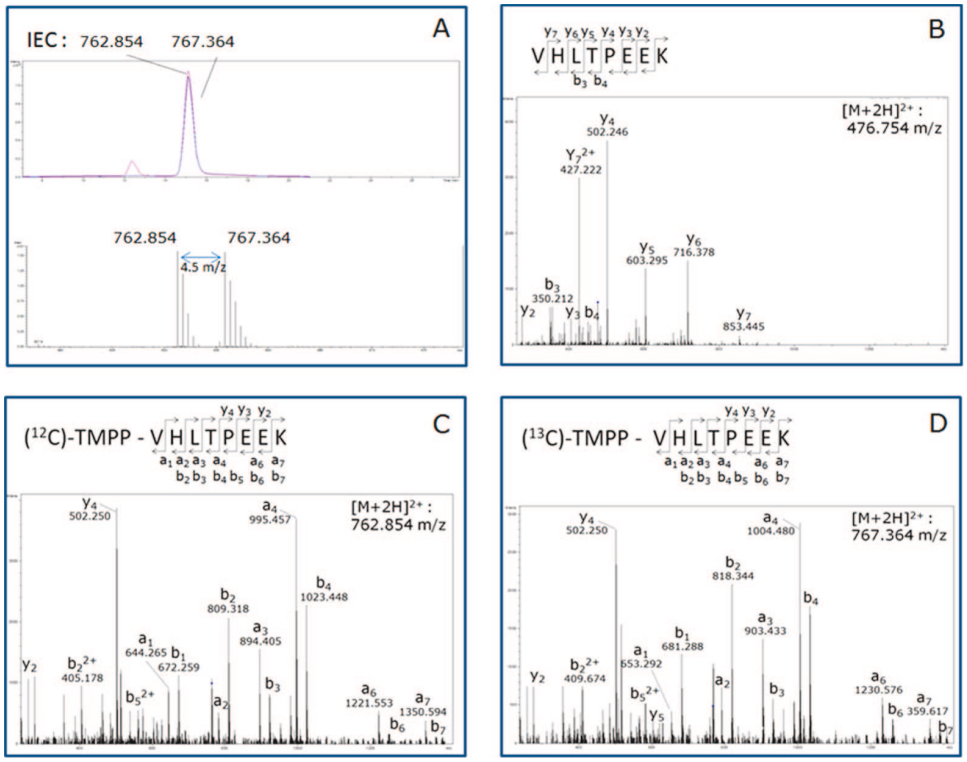



**Figure 2.** Detailed characterization of the tryptic N-terminal peptide VHLTPEEK of the alpha chain of hemoglobin: (A) Ion extracted chromatogram (EIC) from an LC-MS/MS analysis of the peptide derivatized with $^{13}$C-TMPP and $^{12}$C-TMPP that clearly shows the perfect coelution of the two peptides. Below the MS spectrum of the two peptides on which the two isotopic profiles are separated by 9 Da is presented. (B) The underivatized peptide produced the expected fragmentation generating mainly $y_n$- or $b_n$-type ions. (C and D) The comparison of MS/MS spectra of the peptide derivatized with light TMPP and with heavy TMPP, respectively. The derivatized peptides present similar fragmentation patterns with predominant $a_n$- and $b_n$-type ion series when compared to the predominant $y_n$-type ion in the nonderivatized spectrum (B). As expected, the mass difference of 9 Da affects all $a_n$-and $b_n$-type fragments, when comparing the MS/MS spectra of the light and heavy TMPP derivatized peptides.

section, was in-house developed to automate N-terminal peptides validation. In a first step, identical peptide sequences are detected and retained if both the $^{12}$C-TMPP-derivatized and the $^{13}$C-TMPP-derivatized peptides are identified. In a second step, retention times are extracted for every doublet and a user defined tolerance window is applied to ensure coelution of both heavy and light forms. Validor requires an excel file with the following information present in separate columns: accession number, peptide sequence, retention time, peptide modification (see Supporting Information and Method for details).

## RESULTS AND DISCUSSION

A general schematic overview of the N-TOP and dN-TOP strategies is depicted in Figure 1. Both experimental workflows are comparable except for the use of the $^{12}$C-TMPP/$^{13}$C-TMPP mixture for the labeling reaction instead of using only the $^{12}$C-TMPP reagent. This allowed us to significantly improve the so far limiting step of N-terminal peptide validation and to significantly increase the number of the validated protein starts, while maintaining the strength of the approach, i.e. preserving intact all internal peptides.

### dN-TOP Identification and Validation of Internal Peptides

The workflow starts with the denaturation of proteins by reduction and alkylation of cysteine residues to enhance the accessibility of N-termini for chemical derivatization. After treatment of the protein extract with a 1:1 mixture of light and heavy TMPP, a 1D gel separation followed. 1D SDS-PAGE step was shown to be ideal to remove TMPP excess and had the additional advantage not only of being compatible with strong detergents but also reducing the complexity of protein extracts prior to LC-MS/MS analysis. After systematic band cutting, tryptic in gel digestion and nanoLC-MS/MS of each band, all files are merged to generate a global peak list for each lane. This global peak list is then submitted to database searches using Mascot (with full enzyme specificity). Proteins are identified thanks to internal peptides and validated in a usual way (using Scaffold software and score filtering for significant identification at a false discovery rate of 1% with a target/decoy database), since internal peptides are not chemically affected by the TMPP labeling.[10]

### dN-TOP Identification and Validation of N-Terminal Peptides

A second database search is then performed with semitrypsin specificity and TMPP modifications to identify N-terminal peptides. Semitrypsin specificity, which allows a one peptidic termini to be aspecific, is required for this search in order to identify also unpredicted protein starts (downstream of the predicted protein start) that would be missed with full trypsin searches. Indeed, a full trypsin search only allows identifying the N-terminal peptides of the proteins as predicted and present in the database. During this search, both light and heavy forms of


dx.doi.org/10.1021/pr4002993 | J. Proteome Res. 2013, 12, 3063−3070

TMPP are set as variable modifications. The $^{13}C_9$-TMPP modification has been added to the UNIMOD database (http://www.unimod.org). The list of identified peptides contains a series of peptides modified on their N-termini by light or by heavy TMPP. As suspected, those identifications have assigned scores non representative of the spectral quality due to their unusual fragmentation patterns. Figure 2b–d illustrates, in the case of a mixture of model proteins (alpha and beta chains of hemoglobin), the unusual fragmentation pattern of TMPP-derivatized peptides compared to the nonmodified peptides. For TMPP-derivatized peptides, a- and b-type fragmentation ions are dominant due to the permanent charge introduced by the TMPP reagent while the nonmodified peptide produces the expected tryptic peptide fragmentation (mostly y-type ions and a few b-type ions).

Perfect coelution of N-terminal tryptic peptides derivatized by light or heavy TMPP is verified on all doublet peptides. Figure 2a shows that intensities between light and heavy labeled peptides are in close agreement with the initial 1:1 ratio of $^{12}$C- and $^{13}$C-TMPP reagent. The MS spectrum shows that the difference of mass-to-charge values ($m/z$) of the doublet monoisotopic peaks is 4.5 for this doubly charged peptide, corresponding to a mass increase of 9 Da which is adapted to separate the light and heavy peptides' isotope envelopes.

As described in the Experimental Procedures section, Validor allows an automatic validation of the N-terminal peptides based on 2 criteria: the identification of both the $^{12}$C-TMPP-modified and the $^{13}$C-TMPP-modified peptide sequence and a perfect coelution of both forms.

## Application of the Workflow to *H. arsenicoxydans* Proteome

*H. arsenicoxydans* is a $\beta$-proteobacteria which uses organic compounds as an electron donor, oxidizes As(III) and can resist to up to 6 mM As(III) and 200 mM As(V).[18] *H. arsenicoxydans* is the first arsenite-oxidizing bacterium whose genome has been sequenced in 2007 and is rather well annotated.[18,20] However, start site assignment has not yet been validated by experimental proteomics data. Therefore, we have applied our N-TOP and dN-TOP strategies to this organism and we present here a deeper characterization of its proteome, with a special focus on its N-terminome.

## Comparison of N-TOP versus dN-TOP

To verify that the doublet dN-TOP strategy allows identifying a maximum of N-terminal peptides, we have first performed two separate experiments using a single TMPP isotopologue. One protein extract of *H. arsenicoxydans* lysate was treated with $^{12}$C-TMPP while the other one with the $^{13}$C-TMPP, and both derivatized protein mixtures were subjected to the classical N-TOP workflow as described in Figure 1. The two MS/MS data sets were validated using the classical target/decoy approach with a FDR $\leq$1%. These experiments yielded 50 and 74 N-terminal peptides with $^{12}$C-TMPP and $^{13}$C-TMPP labeling, respectively, when using classical validation criteria (Table 1).

Then, the dN-TOP strategy was applied to the same lysate of *H. arsenicoxydans*.

Except for the labeling with a 1:1 mixture of $^{12}$C-TMPP and $^{13}$C-TMPP reagents, the same experimental workflows was applied, i.e., derivatization, separation on 1D SDS-PAGE, in gel digestion and LC-MS/MS analysis of the extracted peptides. In total, Validor allowed the automatic validation of 90 N-terminal peptides thanks to the light and heavy doublet identification (Supporting Information Table S2). This experiment illustrates

**Table 1. Results Obtained with the Classical N-TOP Method Compared to the dN-TOP Approach with Validor**

| | N-TOP $^{12}$C TMPP | N-TOP $^{13}$C TMPP | dN-TOP |
|---|---|---|---|
| Number of N-ter validated with FDR < 1% | 55 | 78 | n.d |
| Number of N-ter validated with Validor | n.d | n.d | 90 |
| (Not expected N-ter) | | | (13) |
| Number of proteins identified with FDR < 1% | 566 | 588 | 504 |

the significantly underestimated validation of N-termini when using the N-TOP strategy and proves the major advantage of the dN-TOP approach as half of the peptides have been discarded in the 2 individual N-TOP experiments (Table 1).

Concerning total protein identifications in these three separate experiments, comparable numbers of proteins have been identified (Table 1 and supplemental Table S1). This proves that the doublet labeling does not affect the global identification rate (even if sample complexity is slightly increased by the labeling with the 2 TMPP forms). This is also due to the fact that TMPP labeling shifts N-terminal peptides' elution times toward a less complex part of the chromatogram, out of the eluting area of internal peptides.

## The *H. arsenicoxydans* N-Terminome with dN-TOP

In total, 504 unique proteins were identified from internal digestion peptides (Table 1). When combining the lists of unique identified proteins over the three experiments ($^{12}$C TMPP labeling N-TOP, $^{13}$C TMPP labeling N-TOP and dN-TOP, Supporting Information Table S1), the total number of proteins raises to 650, increasing the previously published proteome with 384 additional proteins.[20]

From the same data set, 90 unique N-terminal peptides were identified with Validor among which 77 were correctly predicted by the genome annotation (Supporting Information Table S2). The 13 remaining N-terminal peptides did not match to the predicted starts annotated in the *H. arsenicoxydans* database (Table 2). We carefully analyzed these N-terminal peptides in order to highlight possible annotation errors or proteolytic events.

In the case of Flavoprotein HEAR0503, we have identified an N-terminal derivatized peptide presenting a wrongly annotated start site. As illustrated in Figure 3, the identified N-terminal peptide of protein HEAR0503 showed clearly that the start site was experimentally detected 13 amino acids downstream of the annotated translational start site. We checked further if this new start may be in agreement with alternative start prediction algorithms, and if this start would fit to alignments with other known proteins. This identification provides the experimental evidence of remaining incorrectly predicted start sites even after expert manual annotation.[18]

Besides start site annotation errors, we have identified six TMPP-derivatized peptides corresponding to signal peptide cleavage sites (Table 2), allowing to experimentally validate those cleavage sites as predicted by the SignalP 4 algorithm.[21] Interestingly, in one case, identification of the N-terminal TMPP labeled peptide FDFNDVAK supports the predicted cleavage site of protein Glucan Biosynthesis G HEAR3286, and indirectly the prediction of an alternative start codon (Figure 4A). Indeed, in the case of this periplasmic protein involved in the synthesis of membrane-derived oligosaccharides (MDO), two possible starts were predicted according to two different


dx.doi.org/10.1021/pr4002993 | J. Proteome Res. 2013, 12, 3063–3070

**Table 2. List of Identified N-Terminal Peptides That Do Not Match with the Annotated Protein N-Termini**

| protein accession numbers | peptide sequence | peptide start index | SignalP prediction |
|---|---|---|---|
| sp\|HEAR0005 | AIPNDNTPQSPSTLSAAYGASSIQILEGLEAVR | 3 | Between amino acid 27 and 28 |
| sp\|HEAR0225 | TNSIAR | 114 | |
| sp\|HEAR0310 | ATVLK | 28 | |
| sp\|HEAR0348 | AWEPTKPVEFVVPAGTGGGADQMAR | 34 | Between amino acid 33 and 34 |
| sp\|HEAR0415 | DAAYPNK | 23 | Between amino acid 22 and 23 |
| sp\|HEAR0503 | SQNFPDLPNIDPALFTTPTR | 13 | |
| sp\|HEAR1107 | ADITGAGATFPYPIFSK | 26 | Between amino acid 25 and 26 |
| sp\|HEAR1195 | APSAAK | 30 | Between amino acid 29 and 30 |
| sp\|HEAR1337 | TTPAYK | 28 | |
| sp\|HEAR2797 | TMLGFMATDAK | 196 | |
| sp\|HEAR3286 | FDFNDVAK | 31 | Between amino acid 30 and 31 |
| sp\|HEAR3424 | TTTFR | 95 | |
| sp\|HEAR3468 | MLLTR | 97 | |

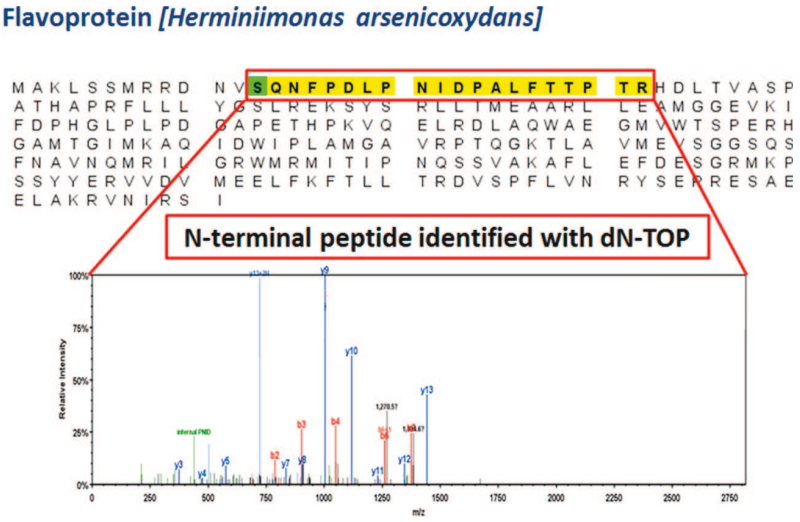



**Figure 3.** Example of a *H. arsenicoxydans* protein, Flavoprotein (HEAR0503), with an experimental start codon correction (13 amino acids after to the currently annotated translation start site).

algorithms (AMIGene and Yuko-Makita), with a good prediction obtained only for the second algorithm. Thus, identification of this TMPP-derivatized peptide allowed to experimentally validate the signal peptide prediction after protein reannotation.

An interesting case of post-translational proteolytic cleavage is illustrated in figure 4B. An N-terminal derivatized peptide was identified starting at position 196 on HEAR 2797, Bifunctional glutamate *N*-acetyltransferase/amino-acid acetyltransferase. A sequence alignment with *C. crenatum* Arginine biosynthesis bifunctional protein showed a high degree of similarity. In this organism, the protein undergoes a proteolytic autolysis between the amino acids 182 and 183, corresponding to residues 195 and 196 in *H. arsenicoxydans*, which generates two chains, the $\alpha$ and the $\beta$ chains.[22] Figure 4 thus shows that the dN-TOP provides a useful tool to identify proteolytic events such as cleavage sites and signal peptide processing.

Five additional TMPP-derivatized peptides were identified with N-termini that could correspond to proteolytic cleavage sites. However, no proteolytic fragments for these proteins identified *in vivo* are yet reported in the literature. Therefore, no biological interpretation can be given to those proteolytic events without additional experiments.

## CONCLUSION AND OUTLOOK

In conclusion, our proof-of-concept experiment on *H. arsenicoxydans* allowed confirming predicted N-termini, correcting wrong start site predictions, and identifying proteolytic events, such as signal peptide cleavages and a proteolytic cleavage sites. We have also demonstrated that dN-TOP presents a significant improvement over the N-TOP approach, for which the labeled peptide validation step was limiting. This improvement makes this methodology compatible with high-throughput and large-scale proteomics studies. This opens also the door to the possibility of performing large-scale experimental validations of predicted genome annotations and dN-TOP reveals to be a powerful proteogenomics tool.

Additionally, the availability of a $^{13}C_9$ TMPP reagent offers the possibility to perform quantitative N-terminomics. It will indeed be possible to compare the N-terminome of two different samples by labeling them with light or heavy TMPP, respectively. The identification and validation method presented here will be useful for a fast detection of the mass spectrum of interest for determining the ratio of the two molecular ions.


dx.doi.org/10.1021/pr4002993 | *J. Proteome Res.* 2013, 12, 3063−3070

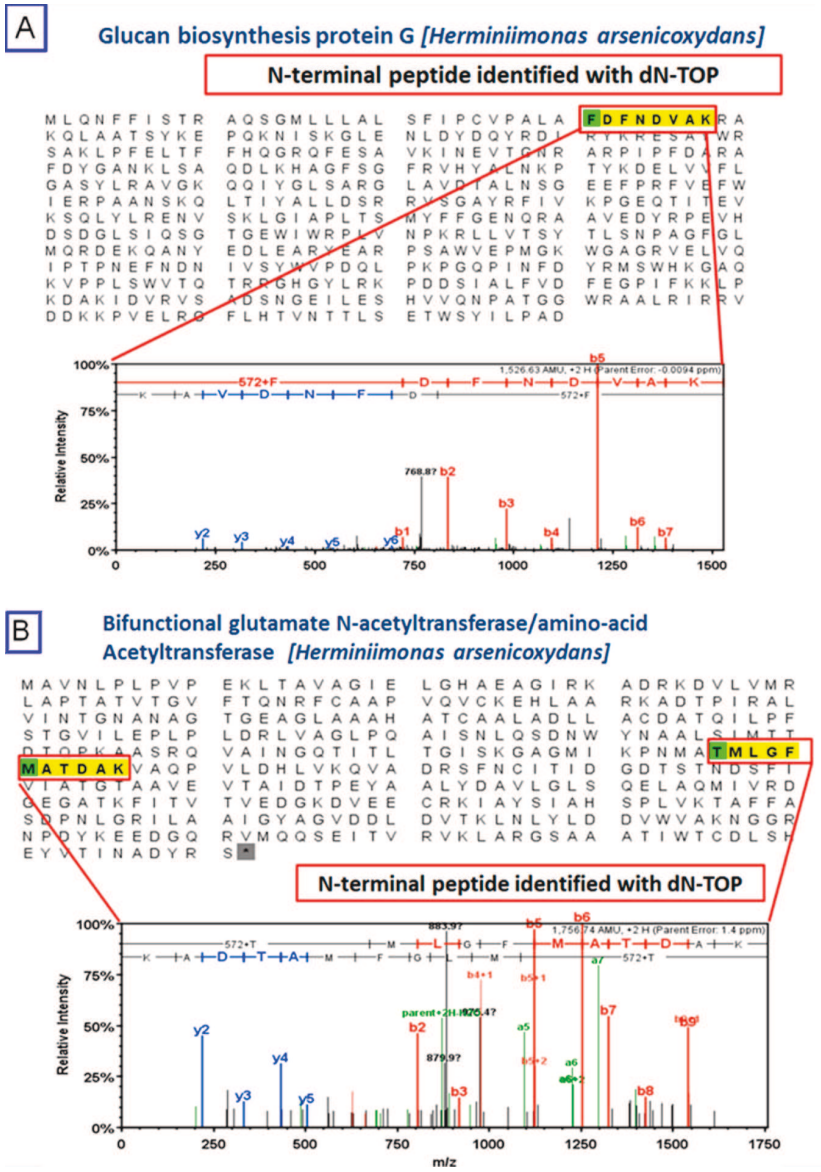



**Figure 4.** (A) Example of an N-terminal processing validated with the dN-TOP approach: the protein Glucan biosynthesis protein G is synthetized with a signal peptide and processed in the ER. The experimental evidence of a mature form (after signal peptide cleavage) with the N-start at position 31 is supported by the *in silico* signal peptide prediction performed with the software SignalP 4.0[21] that predicts a cleavage between amino acid 30 and 31. (B) Example of an endoproteolytic cleavage on protein HEAR 2797. The position 196 has been detected as alternative N-terminal in Bifunctional glutamate N-acetyltransferase/amino-acid acetyltransferase. The closest sequence identity with a *Corynebacterium crenatum* protein which exhibits a proteolytic autolysis confirms a proteolytic cleavage site in this region of the sequence.

## ■ ASSOCIATED CONTENT

### Supporting Information

Additional information as noted in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

### Corresponding Author

*E-mail: christine.schaeffer@unistra.fr. Phone: (+33) 3.68.85.27.79. Fax: (+33) 3.68.85.27.81.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work was supported by the CNRS, the "Agence National de la Recherche" (ANR), the Proteomic French Infrastructure (ProFI; ANR-10-INSB-08-03) and Fondation Pour La Recherche Médicale FRM.

## ■ REFERENCES

(1) Schrimpe-Rutledge, A. C.; Jones, M. B.; Chauhan, S.; Purvine, S. O.; Sanford, J. A.; Monroe, M. E.; Brewer, H. M.; Payne, S. H.; Ansong, C.; Frank, B. C.; Smith, R. D.; Peterson, S. N.; Motin, V. L.; Adkins, J. N. Comparative omics-driven genome annotation refinement: application across Yersiniae. *PLoS One* **2012**, 7 (3), e33903.

(2) Delalande, F.; Carapito, C.; Brizard, J. P.; Brugidou, C.; Van Dorsselaer, A. Multigenic families and proteomics: extended protein

characterization as a tool for paralog gene identification. *Proteomics* **2005**, *5* (2), 450−60.
(3) Blakeley, P.; Overton, I. M.; Hubbard, S. J. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J. Proteome Res.* **2012**, *11* (11), 5221−34.
(4) Bairoch, A.; Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* **1999**, *27* (1), 49−54.
(5) Magrane, M.; Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011**, *2011*, bar009.
(6) Aivaliotis, M.; Gevaert, K.; Falb, M.; Tebbe, A.; Konstantinidis, K.; Bisle, B.; Klein, C.; Martens, L.; Staes, A.; Timmerman, E.; Van Damme, J.; Siedler, F.; Pfeiffer, F.; Vandekerckhove, J.; Oesterhelt, D. Large-scale identification of N-terminal peptides in the halophilic archaea Halobacterium salinarum and Natronomonas pharaonis. *J. Proteome Res.* **2007**, *6* (6), 2195−204.
(7) Bonissone, S.; Gupta, N.; Romine, M.; Bradshaw, R. A.; Pevzner, P. A. N-terminal protein processing: a comparative proteogenomic analysis. *Mol. Cell. Proteomics* **2013**, *12* (1), 14−28.
(8) Venter, E.; Smith, R. D.; Payne, S. H. Proteogenomic analysis of bacteria and archaea: a 46 organism case study. *PLoS One* **2011**, *6* (11), e27587.
(9) Armengaud, J. A perfect genome annotation is within reach with the proteomics and genomics alliance. *Curr. Opin. Microbiol.* **2009**, *12* (3), 292−300.
(10) Gallien, S.; Perrodou, E.; Carapito, C.; Deshayes, C.; Reyrat, J. M.; Van Dorsselaer, A.; Poch, O.; Schaeffer, C.; Lecompte, O. Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. *Genome Res.* **2009**, *19* (1), 128−35.
(11) Chiang, Y. R.; Ismail, W.; Gallien, S.; Heintz, D.; Van Dorsselaer, A.; Fuchs, G. Cholest-4-en-3-one-delta 1-dehydrogenase, a flavoprotein catalyzing the second step in anoxic cholesterol metabolism. *Appl. Environ. Microbiol.* **2008**, *74* (1), 107−13.
(12) Baudet, M.; Ortet, P.; Gaillard, J. C.; Fernandez, B.; Guerin, P.; Enjalbal, C.; Subra, G.; de Groot, A.; Barakat, M.; Dedieu, A.; Armengaud, J. Proteomics-based refinement of Deinococcus deserti genome annotation reveals an unwonted use of non-canonical translation initiation codons. *Mol. Cell. Proteomics* **2010**, *9* (2), 415−26.
(13) Gucinski, A. C.; Dodds, E. D.; Li, W.; Wysocki, V. H. Understanding and exploiting Peptide fragment ion intensities using experimental and informatic approaches. *Methods Mol. Biol.* **2010**, *604*, 73−94.
(14) He, Y.; Parthasarathi, R.; Raghavachari, K.; Reilly, J. P. Photodissociation of charge tagged peptides. *J. Am. Soc. Mass Spectrom.* **2012**, *23* (7), 1182−90.
(15) Altelaar, A. F.; Frese, C. K.; Preisinger, C.; Hennrich, M. L.; Schram, A. W.; Timmers, H. T.; Heck, A. J.; Mohammed, S. Mohammed, S., Benchmarking stable isotope labeling based quantitative proteomics. *J. Proteomics* **2012**, DOI: 10.1016/j.jprot.2012.10.009.
(16) Kline, K. G.; Sussman, M. R. Protein quantitation using isotope-assisted mass spectrometry. *Annu. Rev. Biophys.* **2010**, *39*, 291−308.
(17) Evans, C.; Noirel, J.; Ow, S. Y.; Salim, M.; Pereira-Medrano, A. G.; Couto, N.; Pandhal, J.; Smith, D.; Pham, T. K.; Karunakaran, E.; Zou, X.; Biggs, C. A.; Wright, P. C. An insight into iTRAQ: where do we stand now? *Anal. Bioanal. Chem.* **2012**, *404* (4), 1011−27.
(18) Muller, D.; Medigue, C.; Koechler, S.; Barbe, V.; Barakat, M.; Talla, E.; Bonnefoy, V.; Krin, E.; Arsene-Ploetze, F.; Carapito, C.; Chandler, M.; Cournoyer, B.; Cruveiller, S.; Dossat, C.; Duval, S.; Heymann, M.; Leize, E.; Lieutaud, A.; Lievremont, D.; Makita, Y.; Mangenot, S.; Nitschke, W.; Ortet, P.; Perdrial, N.; Schoepp, B.; Siguier, P.; Simeonova, D. D.; Rouy, Z.; Segurens, B.; Turlin, E.; Vallenet, D.; Van Dorsselaer, A.; Weiss, S.; Weissenbach, J.; Lett, M. C.; Danchin, A.; Bertin, P. N. A tale of two oxidation states: bacterial colonization of arsenic-rich environments. *PLoS Genet.* **2007**, *3* (4), e53.
(19) Muller, D.; Simeonova, D. D.; Riegel, P.; Mangenot, S.; Koechler, S.; Lievremont, D.; Bertin, P. N.; Lett, M. C. Herminiimonas arsenicoxydans sp. nov., a metalloresistant bacterium. *Int. J. Syst. Evol. Microbiol.* **2006**, *56* (Pt 8), 1765−9.
(20) Weiss, S.; Carapito, C.; Cleiss, J.; Koechler, S.; Turlin, E.; Coppee, J. Y.; Heymann, M.; Kugler, V.; Stauffert, M.; Cruveiller, S.; Medigue, C.; Van Dorsselaer, A.; Bertin, P. N.; Arsene-Ploetze, F. Enhanced structural and functional genome elucidation of the arsenite-oxidizing strain Herminiimonas arsenicoxydans by proteomics data. *Biochimie* **2009**, *91* (2), 192−203.
(21) Petersen, T. N.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **2011**, *8* (10), 785−6.
(22) Dou, W.; Xu, M.; Cai, D.; Zhang, X.; Rao, Z.; Xu, Z. Improvement of L-arginine production by overexpression of a bifunctional ornithine acetyltransferase in Corynebacterium crenatum. *Appl. Biochem. Biotechnol.* **2011**, *165* (3−4), 845−55.

DATASET BRIEF

# N-terminome analysis of the human mitochondrial proteome

*Alvaro Sebastian Vaca Jacome[1,2], Thierry Rabilloud[3], Christine Schaeffer-Reiss[1,2], Magali Rompais[1,2], Daniel Ayoub[1,2], Lydie Lane[4,5], Amos Bairoch[4,5], Alain Van Dorsselaer[1,2] and Christine Carapito[1,2]*

[1] BioOrganic Mass Spectrometry Laboratory (LSMBO), Université de Strasbourg, IPHC, Strasbourg, France
[2] IPHC, CNRS, UMR7178, Strasbourg, France
[3] Laboratoire de Chimie et Biologie des Métaux, UMR CNRS-CEA-UGA 5249, iRTSV/LCBM, CEA Grenoble, Grenoble, France
[4] CALIPHO Group, SIB-Swiss Institute of Bioinformatics, Geneva, Switzerland
[5] Department of Human Protein Sciences, Faculty of Medicine, Geneva, Switzerland

Received: December 22, 2014
Revised: March 10, 2015
Accepted: April 30, 2015

The high throughput characterization of protein N-termini is becoming an emerging challenge in the proteomics and proteogenomics fields. The present study describes the free N-terminome analysis of human mitochondria-enriched samples using trimethoxyphenyl phosphonium (TMPP) labelling approaches. Owing to the extent of protein import and cleavage for mitochondrial proteins, determining the new N-termini generated after translocation/processing events for mitochondrial proteins is crucial to understand the transformation of precursors to mature proteins. The doublet N-terminal oriented proteomics (dN-TOP) strategy based on a double light/heavy TMPP labelling has been optimized in order to improve and automate the workflow for efficient, fast and reliable high throughput N-terminome analysis. A total of 2714 proteins were identified and 897 N-terminal peptides were characterized (424 N-α-acetylated and 473 TMPP-labelled peptides). These results allowed the precise identification of the N-terminus of 693 unique proteins corresponding to 26% of all identified proteins. Overall, 120 already annotated processing cleavage sites were confirmed while 302 new cleavage sites were characterized. The accumulation of experimental evidence of mature N-termini should allow increasing the knowledge of processing mechanisms and consequently also enhance cleavage sites prediction algorithms. Complete datasets have been deposited to the ProteomeXchange Consortium with identifiers PXD001521, PXD001522 and PXD001523 (http://proteomecentral.proteomexchange.org/dataset/PXD001521, http://proteomecentral.proteomexchange.org/dataset/PXD0001522 and http://proteomecentral.proteomexchange.org/dataset/PXD001523, respectively).

**Keywords:**
Animal proteomics / dN-TOP approach / Free N-terminome analysis / Human mitochondria / Proteogenomics / Transit peptide

Additional supporting information may be found in the online version of this article at the publisher's web-site

**Correspondence:** Dr. Christine Carapito, BioOrganic Mass Spectrometry Laboratory (LSMBO), Université de Strasbourg, IPHC, CNRS, UMR7178, 25 rue Becquerel, 67087 Strasbourg, France
**E-mail:** ccarapito@unistra.fr
**Fax:** +33368852781

**Abbreviations: dN-TOP**, doublet N-terminal oriented proteomics; **MSDA**, mass spectrometry data analysis; **N-TOP**, N-terminal oriented proteomics; **TMPP**, trimethoxyphenyl phosphonium

Within the frame of large proteome projects and especially the human proteome project, there is an urgent need for large scale determination of human mitochondrial N-termini to achieve a more accurate prediction of mitochondrial proteins and their processing, as most of the existing knowledge and practice is based on predictions and homologies and not on

**Colour Online:** See the article online to view Figs. 1 and 2 in colour.

direct experimental data on human proteins. Their processing is crucial for the normal functioning of mitochondria [1, 2] and deficiency in one of these proteases causes a genetic disease [3]. Such N-terminome data have been generated on yeast [4] and have shed considerable light on the processing events taking place when mitochondrial proteins are imported, e.g. the role of aminopeptidases to generate stable termini.

Although yeast is considered to be a powerful model for the study of mitochondria dysfunction [5], generation of N-terminome data directly on human mitochondria is a more straightforward approach. Furthermore, as it appears more and more obvious that post-transit peptide cleavage processing events are important to produce active and stable mitochondrial proteins, it remains to be determined whether yeast and human mitochondrial proteins are processed in the same way or not. This remains an open question as some of the mitochondrial proteases are shared between yeast and mammals, but mammals also use protease complexes that are not present in yeast [6]. Finally, mitochondrial proteins are degraded by proteases present in the matrix [6, 7] and here again, by determining neo N-termini at downstream positions, N-terminomics will provide the cleavage sites for the endoproteases controlling the half-life of mitochondrial proteins, a feature that has not been fully explored in previous N-terminome approaches [4].

This is why we have undertaken an N-terminomics study on mitochondrial proteins using the recently-developed N-terminomics approach based on a chemical derivatization of the N-terminal alpha amine function with light and heavy (*N*-succinimidyloxycarbonylmethyl)tris(2,4,6-trimethoxyphenyl)phosphonium bromide [8, 9]. Unlike other widely applied N-terminomics strategies requiring specific N-terminal peptide enrichment steps such as the combined fractional diagonal chromatography [10] or the terminal amino isotope selection [11] approaches and others reviewed recently by Hartmann and Armengaud [12], this method enables the characterization, in a single experiment, of both the free protein N-termini and all other internal peptides, and possibly N-terminal α-amino acetylated peptides. The approach also benefits from its easy experimental design, since it requires a single derivatization step at the protein level followed by a classical SDS-PAGE/LC-MS/MS workflow making it fully compatible with standard and efficient proteomics sample preparation protocols (standard buffers and detergents; chaotropic, reducing and alkylating agents; and proteolytic enzymes).

Additionally, trimethoxyphenyl phosphonium (TMPP) derivatization enhances labelled peptides' LC-MS/MS response, both thanks to the added hydrophobicity of the TMPP reagent shifting retention times of derivatized peptides in reverse-phase chromatography to a less complex part of the chromatogram (as illustrated in Supporting Information Fig. 1) and to the introduction of a permanent positive charge that increases their ionization efficiency. Both advantages lead to increased sensitivity to detect low abundant proteins. Furthermore, the dN-TOP [9] strategy recently developed to improve the N-TOP [8] approach is based on the use of a pair of light ($^{12}$C-TMPP) and heavy stable-isotope labelled ($^{13}C_9$-TMPP) TMPP reagents in order to identify doublets of identical N-terminal peptides and automate the validation of labelled peptides. The validation automation is done by looking for this double light/heavy identification of a peptide sequence, with identical modifications, identical elution times and after verifying the peptide's unicity in the searched database as described in Supporting Information. The present work describes an improved workflow of the dN-TOP approach and its direct application to the characterization of the human mitochondrial proteome and N-terminome.

Figure 1A presents the overall sample preparation strategy and a detailed description of the experimental procedures is provided in Supporting Information. Briefly, proteins were precipitated and resuspended in freshly prepared derivatization buffer (50 mM Tris-HCl, pH 8.2, 6 M urea, 2 M thiourea, SDS 1%), reduced (5 mM TBP, 1 h) and alkylated (50 mM Iodoacetamide, 1 h, room temperature). An equimolar solution of light and heavy TMPP (100 mM, 30% Acetonitrile, 170:1 reagent:protein molar ratio, 1 h, room temperature) was added. Light- and heavy-labelled TMPP induce mass shifts on peptides of 572.18 and 581.21 Da, respectively. Derivatized samples were loaded on a 10% mono-dimensional SDS-PAGE. This electrophoresis step is important since it allows separating and decomplexifying the samples and eliminating the excess of unbound TMPP reagent. After colloidal coomassie blue staining, the gels were cut in regular 2 mm bands, proteins were in-gel reduced (10 mM dithiotreitol in 25 mM ammonium bicarbonate), alkylated and digested overnight with trypsin or Asp-N. Extracted peptides were analysed on different nanoLC-MS/MS platforms.

Overall, four different experiments (12 different samples) were carried out as described in Supporting Information Table 1. Among these experiments, different mitochondria enrichment protocols were evaluated, two proteolytic enzymes were used (Trypsin and AspN) and nanoLC-MS/MS analyses were performed on two different instrumental platforms [low-resolution ion-trap with two fragmentation modes (CID and ETD) and high-resolution Q-TOF]. All data were searched with two database search engines (MASCOT, Matrix Science, London, UK and the open-source OMSSA algorithms [13]).

The combination of search engines provides complementary results that increase the total number of protein identifications. The gain of combining OMSSA to MASCOT results goes from 5% for the high-resolution/trypsin digestion/CID fragmentation experiment to 59% for the low-resolution/Asp-N digestion/ETD fragmentation experiment (Supporting Information Fig. 2). Coincidentally, this experiment allowed demonstrating that ETD fragmentation is not adapted for TMPP-labelled peptides.

The new data validation workflow is illustrated in Fig. 1B. In a first round, original peak lists were searched with two search engines against a concatenated target/decoy database including all human entries extracted from UniProtKB-

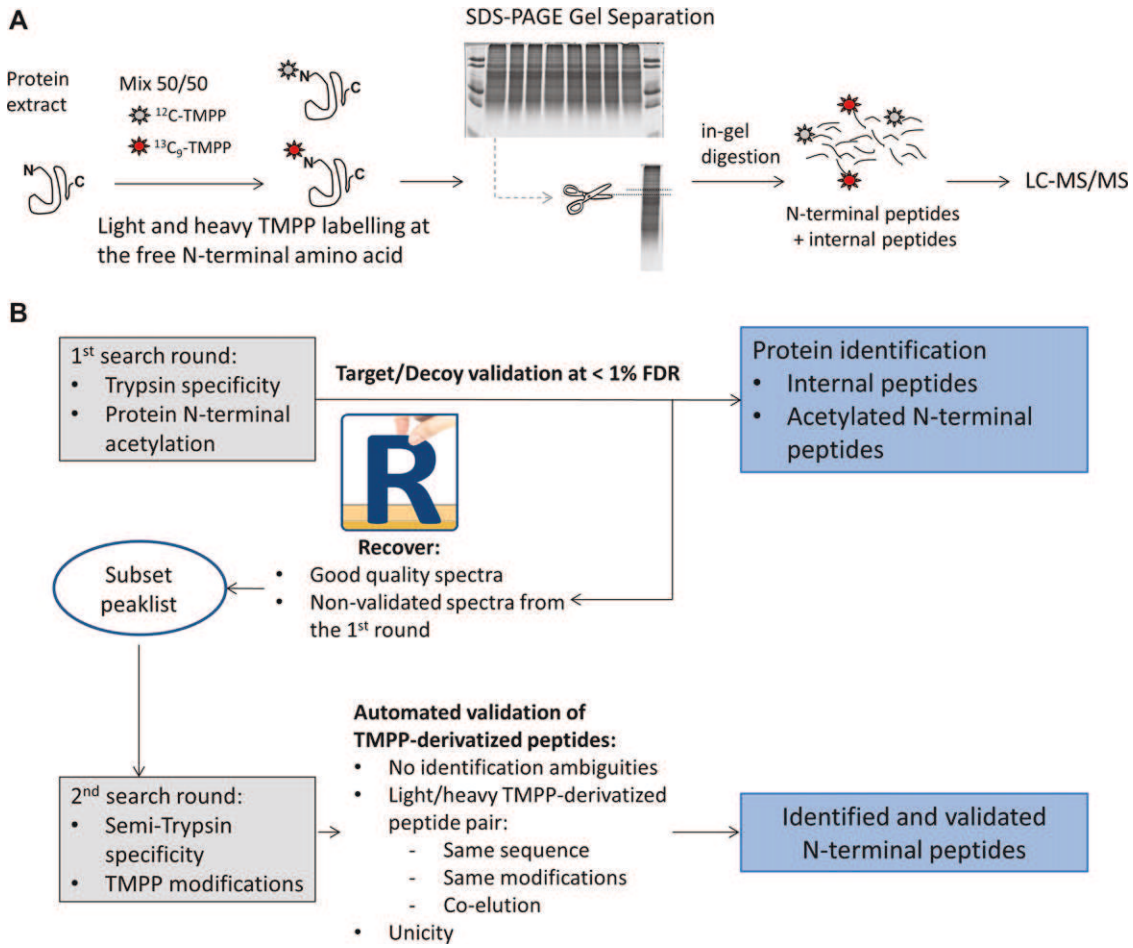



**Figure 1.** Overview of the analytical workflow (A) Proteins and N-termini identification and validation strategy (B).

SwissProt (release-2012_03, 20250 entries) using the database generation toolbox of the MSDA pipeline [14]. MASCOT searches were run on a local server while OMSSA searches were run on a computer grid using the MSDA interface [14]. This first search round was done using full trypsin enzyme specificity, one missed cleavage allowed, carbamidomethylation of cysteine was set as a fixed modification and oxidation of methionine and protein N-terminal acetylation were set as variable modifications. MASCOT and OMSSA results files were loaded into the Scaffold software (version 3.6.5; Proteome Software Inc., Portland, USA). Filtering criteria based on probability-based scoring of the identified peptides were taken into account in order to reach a false discovery rate (FDR) <1% based on the number of decoy hits. This first search round enabled the identification of the protein sets present in the samples thanks to internal peptides (Supporting Information Table 2) and N-terminal α-amino acetylated peptides (Supporting Information Table 3).

A second round of searches was performed to identify free protein N-termini labelled by TMPP. Therefore, the recover module of MSDA [14] was used to create subset peak lists composed of filtered "high-quality" spectra from all non-assigned spectra during the first round of searches. "High-quality" spectra were defined as including at least four peaks higher than 1.5 times the intensity of the background noise and at least one peak above the $m/z$ ratio of the precursor ion and all spectra having led to a successful identification in the first round were filtered out. Using these subset peak lists, a second round of database searches was performed using semi-trypsin enzyme specificity. Light (+572.18 Da) and heavy (+581.21 Da) TMPP derivatization on any peptide N-terminal amino acid, side-chain derivatization of tyrosine and lysine by light and heavy TMPP, and methionine oxidation were set as variable modifications. The searched fragment ions were a-, b- and y-ions. All other parameters were identical to the first search round. Results files were loaded into an in-house developed software Proline Studio and all spectra leading to an identification exceeding a minimum set threshold (MASCOT Ion Score of 13 and OMSSA -log(e-value) of 4) and having a pretty rank (as defined by MASCOT) equal to 1 were kept. The pretty rank is a value similar to the rank except that equivalent scores get equivalent ranks. The use of pretty ranks allows detecting and eliminating ambiguous spectra matching to multiple sequences present

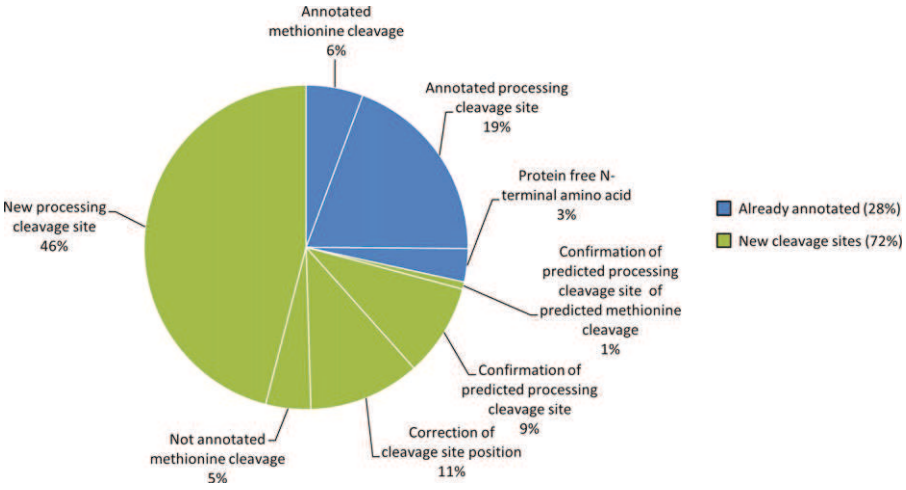



**Figure 2.** Distribution of already annotated and new identified N-terminal positions.

in the database with equivalent scores. Pretty ranks also allow identifying and eliminating spectra for which the exact TMPP labelling position (peptide N-terminus or side chain of tyrosine or lysine) cannot be determined. All ambiguous spectra were discarded. Then, TMPP-labelled peptides were validated only if the identical peptide sequence and modifications were identified for both the light and heavy TMPP-derivatized peptide forms. Also, peptides were validated only when retention times of MS/MS events of both light and heavy forms were within a tolerance window of 30 s to ensure perfect co-elution. A peptide sequence was considered as unique when it belonged to a single protein in the canonical human protein database used for the searches. Further annotation interpretation was exclusively deduced for unique/proteotypic peptides. Validated N-terminal TMPP-labelled peptides meeting all these criteria are listed in Supporting Information Tables 4 and 5.

Taking into account the 12 different experiments, the combined results allowed the identification of 2714 protein groups among which 810 (30%) are annotated to be mitochondrial proteins (Table 1 and Supporting Information Tables 6 and 7). A total of 558 different TMPP-derivatized peptide backbone sequences, of which 473 were unique in the searched database, and correspond to the free N-terminal position of 356 different proteins were validated. 85 (15%) non-unique TMPP-derivatized peptides were validated as their spectra were of high quality. Even if additional internal peptides were identified discriminating a unique protein candidate for some of them, we excluded all these non-unique peptides for any further annotations and comparisons to previous annotation information recorded in neXtProt (designed as "shared peptides" in Table 1).

Among the identified start positions, 245 were identified between amino acids 2 and 100 of the protein sequences, in accordance with typical transit peptide lengths. Beyond the 100th position, we classified the N-termini as further processing cleavage sites, e.g. protein degradation. This hypothesis is substantiated by the fact that these 117 termini are located on 97 proteins, making an average of 1.2 cleavage sites per protein, indicative of potential protein degradation (Table 1).

Overall, 120 (28%) of all identified start positions had already been annotated (14 (3%) free N-termini, 24 (6%) methionine cleavages and 82 (19%) processing cleavage sites) in the highly curated knowledgebase neXtProt. 43 identified positions enabled the experimental validation of so far only predicted cleavage positions (3 (1%) methionine cleavages and 39 (9%) processing cleavage sites). 19 (5%) peptides enabled the identification of new, unpredicted methionine cleavages and 47 (11%) would allow the correction of wrongly annotated processing cleavage sites. Finally, 194 (46%) completely new processing cleavage sites were identified on 160 different proteins (Fig. 2) among which 80 are annotated mitochondrial. Among the 194 new cleavage sites, 88 are located between amino acids 2 and 100 of the protein sequences and 62 have been identified on annotated mitochondrial proteins. They are thus possible new transit peptides cleavage sites due to translocation/processing events. The remaining 106 new cleavage sites, being located beyond the 100th position can be classified as describing protein degradation products.

In addition to the free N-terminome analysis, 424 backbone peptide sequences of N-terminal acetylated peptides corresponding to 357 distinct proteins were identified in our dataset. Among these 357 proteins, 30% are annotated as mitochondrial proteins, justified by the fact that not all mitochondrial proteins undergo processing [15]. Remarkably, for the 2-oxoisovalerate dehydrogenase subunit beta (P21953) we identified the N-terminal acetylated peptide and the TMPP-labelled peptide starting at position 51, already annotated as the transit peptide start site, demonstrating that we have identified the protein both outside, prior to processing, and inside the mitochondria, after transit peptide cleavage [16]. For PGAM5 (Q96HS1), we identified both the N-terminal acetylated peptide at position 2, and a TMPP-labelled peptide at position 25. Cleavage by PARL at this position was shown to have an important physiological role in stress response [17]. For two other mitochondrial proteins (Q02978

**Table 1.** Summary of identification results obtained from all combined experiments

| Type | Number of peptides (unique backbone sequences) | Number of unique N-terminal positions | Number of proteins | Number of mitochondrial proteins | Percentage of mitochondrial proteins |
|---|---|---|---|---|---|
| Total validated free-N-terminal TMPP-labelled peptides | 558 | – | – | – | – |
| Validated N-terminal with unique peptides | 473 | 422 | 356 | 188 | 53% |
| Position 1 | 18 | 14 | 14 | 7 | 50% |
| Position 2 | 54 | 46 | 46 | 20 | 43% |
| 2 < Position ≤ 100 | 279 | 245 | 220 | 142 | 65% |
| Position > 100 | 122 | 117 | 97 | 35 | 36% |
| Shared peptides | 85 | – | – | – | – |
| Total ragged ends | 82 | – | 28 | 24 | 86% |
| Acetylated N-terminus | 424 | 360 | 357 | 106 | 30% |
| Position 1 | 84 | 75 | 75 | 19 | 25% |
| Position 2 | 340 | 285 | 285 | 88 | 31% |
| Combined results of all experiments | – | – | 2714 | 810 | 30% |
| Total N-termini | 897 | 782 | 693 | 283 | 41% |

and P54819), we identified both N-terminal acetylated peptides (at positions 2 and 1, respectively) and TMPP-labelled peptides at positions 6/7 and 2/4, respectively.

Overall, when combining these results with TMPP-based identifications, information about the actual protein N-termini for 26% of all identified proteins could be obtained. This high percentage can be attributed to the unique advantage of the dN-TOP approach offering the possibility to concomitantly identify, in a single run, the proteins' N-terminally acetylated peptides and the free N-termini derivatized by TMPP.

Of note is the identification of multiple successive N-terminal positions for 28 proteins, assimilable to ragged-ends [4, 18], that may indicate that the mitochondrial processing could lead to multiple cleavage sites. As an example, for the isocitrate dehydrogenase [NAD] subunit alpha (P50213), TMPP-derivatized peptides corresponding to positions 27 and 28 were unambiguously identified and confirmed in nine different samples.

Finally, another remarkable note is that the gain in sensitivity provided by TMPP-labelling certainly explains the identification of 60 proteins solely thanks to their N-terminal TMPP-derivatized peptide. For example, the protein Cytochrome c oxidase subunit 8A (P10176) was identified in three different samples exclusively by its free N-terminal peptide. Even more remarkable is that four of these 60 proteins are annotated as "missing proteins" in the context of the chromosome-centric Human Proteome Project cHPP (neXtProt release of 19-09-2014) [19]. This certainly demonstrates the power of our method, combining the gain in sensitivity provided by TMPP-labelling and the mitochondria enrichment preparation protocol, to identify low abundant and so far non-identified proteins (Supporting Information Tables 5 and 6).

*The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium* [20] *with the dataset identifiers PXD001521, PXD001522 and PXD001523. This work was financially supported by the "Agence Nationale de la Recherche" (ANR; ANR eNergiome ANR-13-BSV6-0004-03) and the Proteomic French Infrastructure (ProFI; ANR-10-INSB-08-03). VJAS was supported by a doctoral fellowship from the French Ministry of Research.*

*The authors have declared no conflict of interest.*

## References

[1] Van Dyck, L., Langer, T., ATP-dependent proteases controlling mitochondrial function in the yeast *Saccharomyces cerevisiae*. *Cell. Mol. Life Sci.* 1999, *56*, 825–842.

[2] Gakh, O., Cavadini, P., Isaya, G., Mitochondrial processing peptidases. *Biochim. Biophys. Acta* 2002, *1592*, 63–77.

[3] Casari, G., De Fusco, M., Ciarmatori, S., Zeviani, M. et al., Spastic paraplegia and OXPHOS impairment caused by mutations in paraplegin, a nuclear-encoded mitochondrial metalloprotease. *Cell* 1998, *93*, 973–983.

[4] Vogtle, F. N., Wortelkamp, S., Zahedi, R. P., Becker, D. et al., Global analysis of the mitochondrial N-proteome identifies a processing peptidase critical for protein stability. *Cell* 2009, *139*, 428–439.

[5] Baile, M. G., Claypool, S. M., The power of yeast to model diseases of the powerhouse of the cell. *Front. Biosci.* 2013, *18*, 241–278.

[6] Voos, W., Mitochondrial protein homeostasis: the cooperative roles of chaperones and proteases. *Res. Microbiol.* 2009, *160*, 718–725.

[7] Koppen, M., Langer, T., Protein degradation within mitochondria: versatile activities of AAA proteases and other peptidases. *Crit. Rev. Biochem. Mol. Biol.* 2007, *42*, 221–242.

[8] Gallien, S., Perrodou, E., Carapito, C., Deshayes, C. et al., Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. *Genome Res.* 2009, *19*, 128–135.

[9] Bertaccini, D., Vaca, S., Carapito, C., Arsene-Ploetze, F. et al., An improved stable isotope N-terminal labeling approach with light/heavy TMPP to automate proteogenomics data validation: dN-TOP. *J. Proteome Res.* 2013, *12*, 3063–3070.

[10] Staes, A., Impens, F., Van Damme, P., Ruttens, B. et al., Selecting protein N-terminal peptides by combined fractional diagonal chromatography. *Nat. Protoc.* 2011, *6*, 1130–1141.

[11] Kleifeld, O., Doucet, A., auf dem Keller, U., Prudova, A. et al., Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nat. Biotechnol.* 2010, *28*, 281–288.

[12] Hartmann, E. M., Armengaud, J., N-terminomics and proteogenomics, getting off to a good start. *Proteomics* 2014, *14*, 2637–2646.

[13] Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L. et al., Open mass spectrometry search algorithm. *J. Proteome Res.* 2004, *3*, 958–964.

[14] Carapito, C., Burel, A., Guterl, P., Walter, A. et al., MSDA, a proteomics software suite for in-depth Mass Spectrometry Data Analysis using grid computing. *Proteomics* 2014, *14*, 1014–1019.

[15] Dolezal, P., Likic, V., Tachezy, J., Lithgow, T., Evolution of the molecular machines for protein import into mitochondria. *Science* 2006, *313*, 314–318.

[16] Corral-Debrinski, M., Blugeon, C., Jacq, C., In yeast, the 3′ untranslated region or the presequence of ATM1 is required for the exclusive localization of its mRNA to the vicinity of mitochondria. *Mol. Cell. Biol.* 2000, *20*, 7881–7892.

[17] Sekine, S., Kanamaru, Y., Koike, M., Nishihara, A. et al., Rhomboid protease PARL mediates the mitochondrial membrane potential loss-induced cleavage of PGAM5. *J. Biol. Chem.* 2012, *287*, 34635–34645.

[18] Fortelny, N., Yang, S., Pavlidis, P., Lange, P. F., Overall, C. M., Proteome TopFIND 3.0 with TopFINDer and PathFINDer: database and analysis tools for the association of protein termini to pre- and post-translational events. *Nucleic Acids Res.* 2014, *43*(Database issue), D290–D297.

[19] Lane, L., Bairoch, A., Beavis, R. C., Deutsch, E. W. et al., Metrics for the Human Proteome Project 2013–2014 and strategies for finding missing proteins. *J. Proteome Res.* 2014, *13*, 15–20.

[20] Vizcaino, J. A., Deutsch, E. W., Wang, R., Csordas, A. et al., ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* 2014, *32*, 223–226.

# Chapter II Personalized multi-omics profiling

Proteogenomic analysis is optimized when Genomic, Transcriptomic and Proteomic data are generated from the same sample. When optimally integrated, this enables the creation of a sample-specific protein sequence database. In this approach personalized databases for each individual sample is generated using DNA and/or RNA sequencing. This enables to capture individual-specific differences in the genomic and transcriptomic information onto the protein database without drastically increasing the database size, and thus, keeping a good sensitivity, selectivity and reliability of the protein identification results.

The recent remarkable technical developments in sequencing techniques such as next generation sequencing (NGS), exome sequencing (WES), RNA sequencing (RNA-seq) and ribosome profiling render this possible. The cost to fully sequence a whole human-sized genome has plummeted in recent years, from almost 100 billion dollars in 2001 when the whole genome was first sequenced by the Human Genome Project to almost 1200 dollars in October 2015 [240, 241] (Figure V-8). As well, whole human exome and RNA sequencing are now accessible for less than 500 dollars per sample. This will open the possibility of generating personalized databases for individual samples using DNA and RNA sequencing.

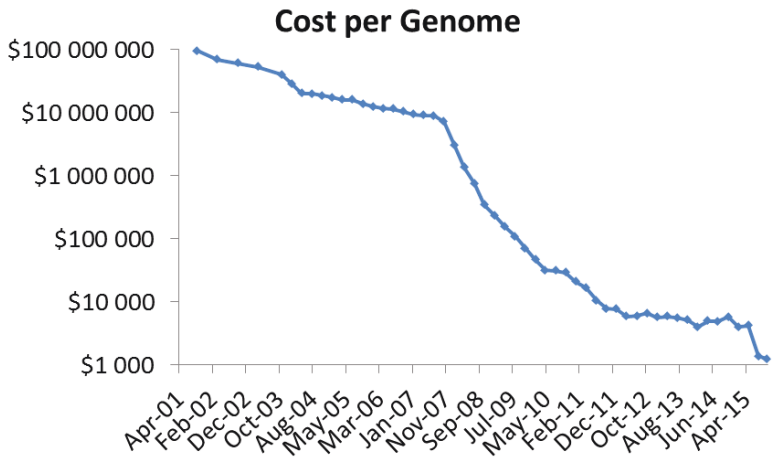



**Figure V-8: Evolution of the cost of sequencing a human-sized genome in dollars (adapted from [240]).**

Recent proof-of concept studies have been proposed:

- Low et al. performed DNA and RNA sequencing to generate a sample-specific database of two rat strains using liver tissue. They analyzed the rat proteomes with an extensive fractionation (36 SCX fractions) and using 5 complementary digestion enzymes. This study validated 1,195 gene predictions, 83 splice events, 126 proteins with nonsynonymous variants, and 20 isoforms with nonsynonymous RNA editing [11].

- By creating a customized database using Ribosome profiling (RIBO-Seq) data [242, 243], Menschaert *et al.* identified new protein products, new protein splice variants, single nucleotide polymorphism variant proteins, and N-terminally extended forms of known proteins in mouse embryonic stem cells [244].

In this context, we have developed a workflow to integrate multi-omics data and build personalized protein databases.

# A. Multi-omics study and generation of personalized databases

## *A.1. Context of the study*

This project was carried out in collaboration with the Human Molecular Immunogenetics laboratory of the University of Strasbourg, and particularly with Raphaël Carapito, Nicodème Paul, Ghada Alsaleh, Louise Ott and Seiamak Bahram.

We were interested to study members of a family in which a member was diagnosed with hyperimmunoglobulinemia D and periodic fever syndrome (HIDS). This disease is characterized by recurrent fever with inflammatory symptoms. A genomic variation in the mevalonate kinase (MK) gene has been identified as being related to this disease [245].

The interesting thing about this family, which pedigree is shown in Figure V-9, is that the parents and the brother are heterozygous for this mutation and do not have the symptoms of the disease. However, the two sisters are homozygous for this mutation but only sister 2 is symptomatic. This suggests that there is another biological process that compensates for the MK deficiency in sister 1. To gain insights into the biomolecular processes differences between the two sisters a multi-omics approach was carried out.

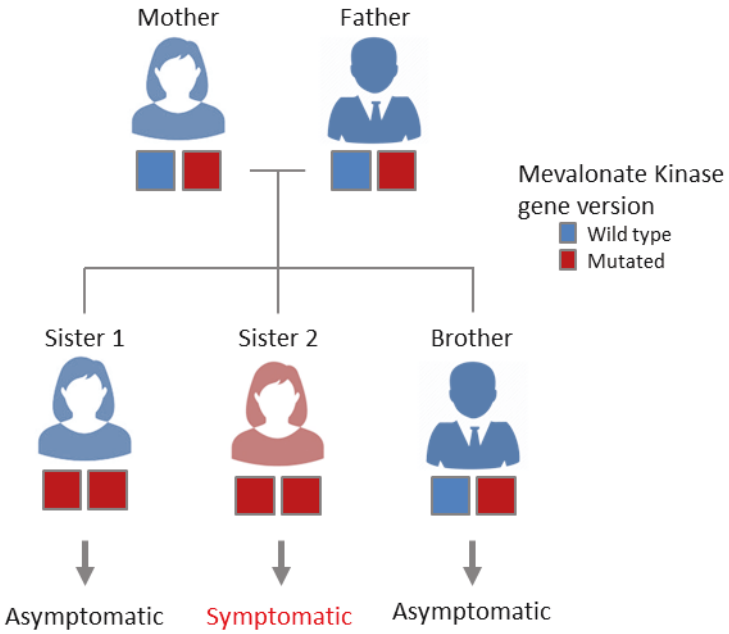



**Figure V-9 : Pedigree of the family showing the allele type of the gene presumably related to HIDS.**

## *A.2. Multi-omics analysis*

We performed a multi-omics analysis of the two sisters with the following technologies (Figure V-10):

- Whole exome sequencing of the two sisters.
- Whole transcriptome sequencing by RNA-seq on total RNA (rRNA were depleted).
- Whole proteome characterization by SDS-PAGE-LC-MS-MS/MS.

This should enable the identification of variants in the genome that have an impact on protein sequences (missense, nonsense, splice-site variants and coding indels). The RNA-Seq analysis should provide information of individual-specific splice-variants. The RNA-Seq and proteomics analyses were performed in triplicate to obtain quantitative information. This integrated personalized multi-omics profiling workflow should give insights to the biomolecular dissimilarities that can explain the two sisters' phenotypic differences. To do this, white blood cells were stimulated

with LPS for 6h (Lipopolysaccharides from Salmonella) to elicit an immune response. The unstimulated and stimulated cells were analyzed for each sister.

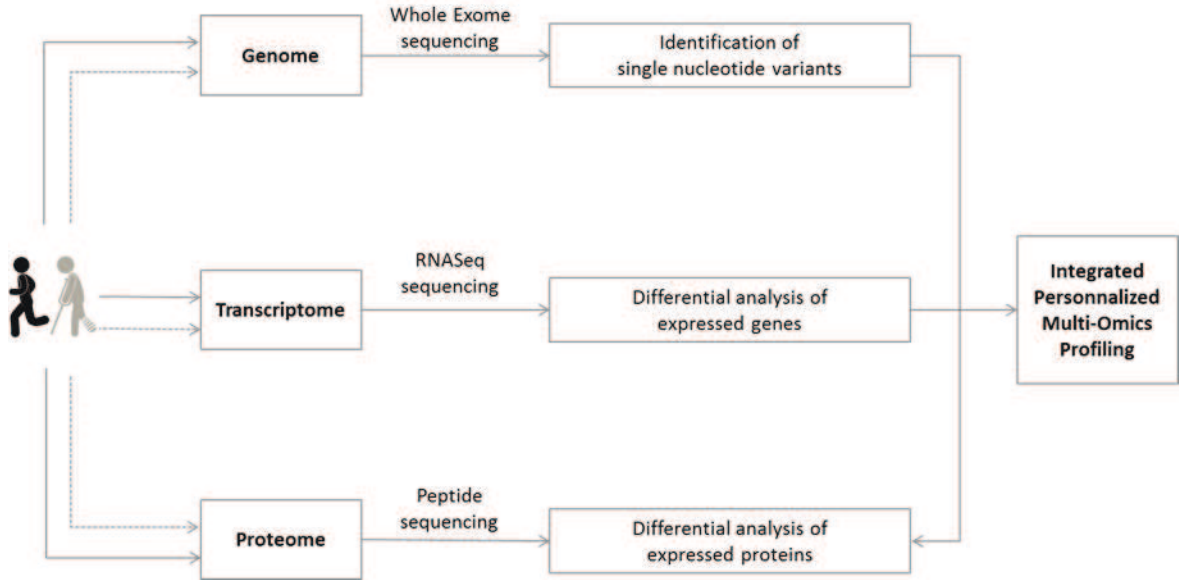



**Figure V-10 : Overview of the personalized multi-omics profiling workflow.**

Additionally, the genomic and transcriptomic data was used to generate personalized databases. These databases were used to improve peptide identification, as peptides originating from genomic variations can be detected. Figure V-11 shows the Proteogenomic approach used to improve peptide identification. More details about the database creation are given below in part A.3.3.

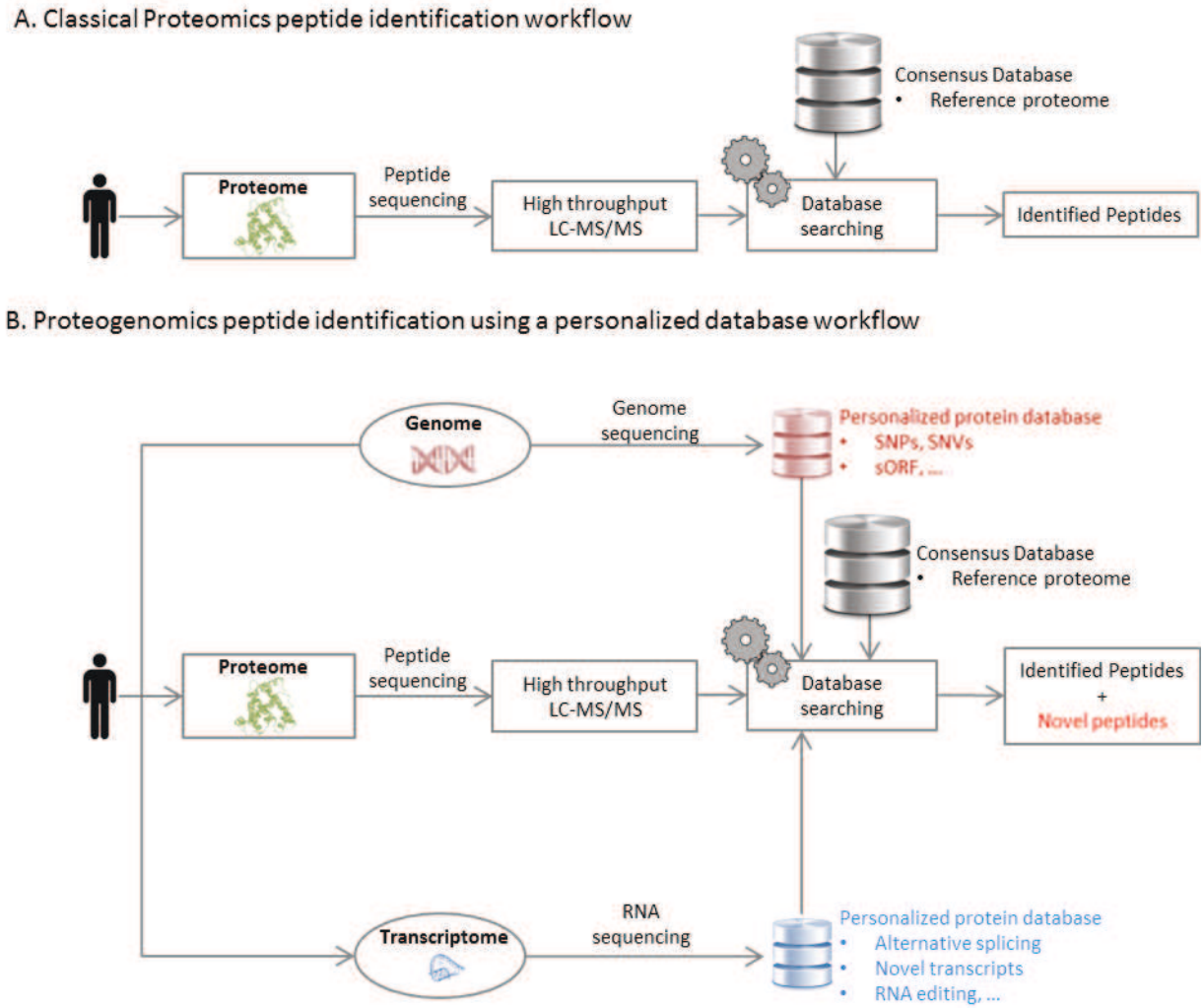



**Figure V-11 : Overview of peptide identification with classical proteomics and omics-based personalized database proteogenomics.**

Two mass spectrometry experiments were performed. The first one was a differential quantification analysis by spectral count aiming at finding differentially expressed proteins in the different conditions analyzed.

The second experiment was designed to reach the maximum proteome coverage in order to get best benefit from personalized database searches. This in-depth analysis of the proteome was done using long LC gradients and a new mass spectrometer whose specifications, at the time, outperformed those of the instruments present at the laboratory.

## *A.3. Experimental Design*

### A.3.1 Sample preparation

An SDS-PAGE separation followed by LC-MS-MS/MS was chosen for this study as a large protein abundance dynamic range was expected. The samples needed to be fractionated to achieve good proteome coverage. A global protein quantification was performed with a Bradford assay prior to the loading into the gel. We made sure that the same protein quantities were used for all samples. Enzymatic digestion using Trypsin was performed overnight at 37°C. The resulting peptides were extracted and analyzed in injection triplicates. More details are given in Experimental Section E.1 on page 221.

### A.3.2 nanoLC-MS analysis

**Relative quantification analysis by spectral count:** The nanoLC-MSMS experiments were carried out on a NanoAcquity LC-system (Waters, Milford, MA, USA) coupled to a maXis 4G QToF mass spectrometer (Bruker Daltonics, Bremen, Germany). The 6 most intense peptides were chosen for MS/MS analysis. The instrument was set to do a spectral count analysis. We used a 12 seconds exclusion time to have a compromise between the spectral redundancy (i.e. how many spectra per peptide were measured) and the proteome coverage (i.e. how many different peptides were measured). Full nanoLC-MSMS parameter details are given in Experimental Section E.2 on page221.

**In-depth proteome characterization experiment:** The relative quantification experiments were carried out on February 2013, and the second experiment was performed in April 2014. In the meantime a new mass spectrometer arrived at the laboratory, a TripleTOF 5600+ (AB Sciex, Framingham, US USA). This instrument has a faster scanning speed, a larger dynamic range and a higher sensitivity than the maXis 4G. To obtain the largest proteome coverage this instrument was thus chosen. Also the nanoLC sytem, a NanoAcquity LC-system (Waters, Milford, MA, USA), was set to provide a 145-minutes long gradient. This enabled an in-depth characterization of the proteome. More details are provided in in Experimental Section E.3 on page 222.

### A.3.3 Personalized database creation

Our collaborators performed whole exome sequencing and RNA-seq experiments on the same samples. Using these data, we created personalized protein sequence databases for each individual including following information:

- Personalized database 1: an exome-derived personalized database
- Personalized database 2: a RNA-seq derived personalized database.

**Personalized database 1:** The first step for the creation of the personalized database was the matching of RefSeq identifiers [66] to UniProt identifiers [69]. Tools for the conversion of identifiers across databases exist, but there are a significant number of identifiers whose correspondence remains unresolved. It was thus necessary to do the annotation using the RefSeq and UniProt databases. Using the RefSeq annotation information and the reference Genome (version hg19) the corresponding RNA sequences for each RefSeq entry were generated. These RNA sequences were then translated into proteins to be compared with the UniProt database. This provided all corresponding protein sequences and their genomic locations. These genomic coordinates are essential for the introduction of the sequence variants observed in the exome into the sample-specific databases.

Using a sequence alignment algorithm (BLAT) [246] the protein sequences were then aligned to the protein sequences in the UniProt database. This way, for every protein in the UniProt database, we obtained its genomic location and its sequence of introns and exons.

The analysis of an individual's exome data enabled the identification of the positions of the variations relative to a reference genome. Using this information, an individual's exome can be fully reconstructed. All exon variations do not necessarily induce a variation in translated proteins. It is thus important to identify exon variations affecting protein sequences. The list of non-synonymous variants thus determined enables making changes to the RNA sequences. The result is the exome-derived personalized protein database.

**Personalized database 2:** Patient-specific alternative splicing was obtained from RNA-Seq data. Having found the correspondence between the RefSeq and the Uniprot database and having obtained the genomic coordinates for each protein, the patient-specific alternative splicing proteins were also introduced in the personalized database.

After taking into account exome variations as well as alternative splicing sites, RNA sequences were then generated and translated into proteins. Knowing the correspondence of each protein to the UniProt protein entries, a new version of the UniProt database was generated. Some proteins were modified taking into account the genetic specificity of the studied individuals.

### A.3.4 Data analysis

The data coming from the samples of each of the two sisters was analyzed using its corresponding personalized database.

The Scaffold software [199] was first used to analyze the data. However, Scaffold failed to make a clear distinction of two protein sequences having a single amino acid variant. These two proteins would be grouped together as a single protein group making the search for amino acid variants difficult.

A two-stage search workflow was engineered using the recently in house developed Proline software to improve peptide identification (Figure V-12). For a given sister, in the first search stage the data was searched against its personalized database 1 containing all human entries including known isoforms extracted from the UniProtKB-SwissProt database (40654 target entries) and the personalized-genome derived database (Database1-S1: 45834 target entries, Database1-S2: 45883 target entries) (See part A.3.3 above for details). The false discovery rate was controlled to reach <1%.

Then a new peaklist file was created using the Recover module of MSDA [233] containing all unassigned spectra during the first search. This subset peaklist files were searched against the personalized databases 2 (RNASeq-derived database). This enabled the identification of patient-specific sequence variants and transcript variants.

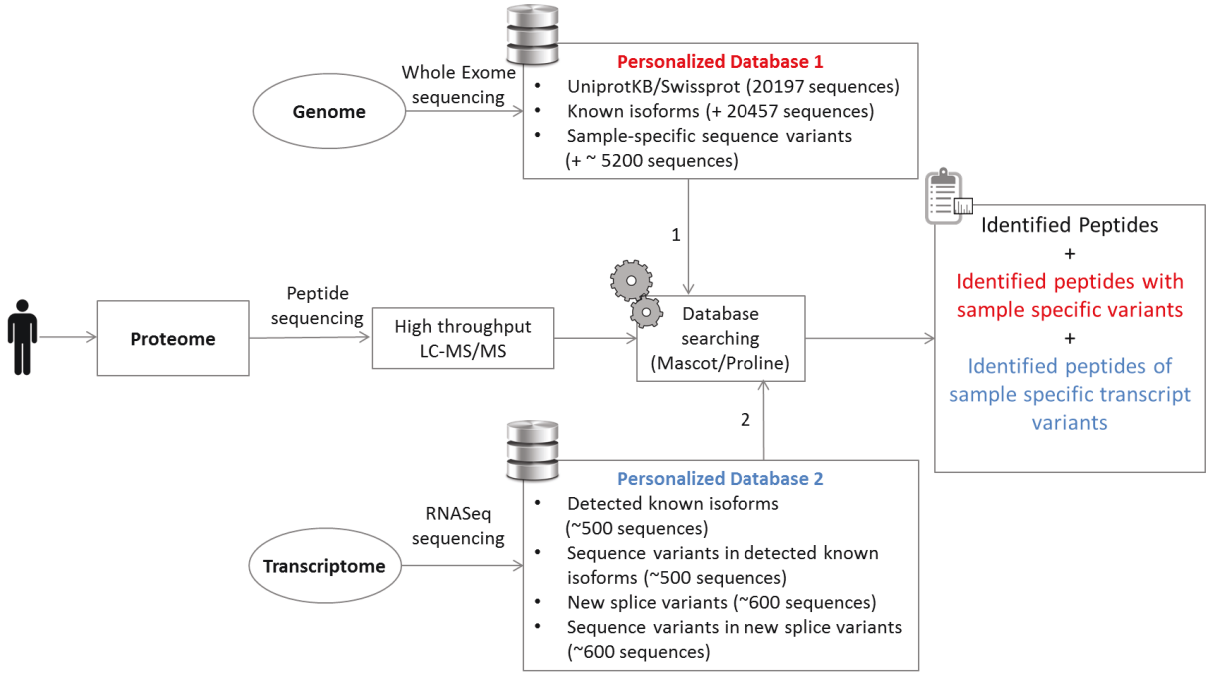



**Figure V-12 : Improving peptide identification with a Proteogenomic approach using Genome and RNA sequencing data to derive personalized protein databases.**

## *A.4. Results of the study*

Overall we identified in the in-depth proteome characterization experiment, more than 4200 protein groups and more than 31200 peptide sequences per sample. We identified 106 nonsynonymous sequence variants on 96 proteins using the exome-seq derived databases 1 and 2 new splice variants using the RNA-seq derived databases 2.

For 77 genomic mutations included in the personalized exome-seq derived databases from both S1 and S2, an evidence of expression at the protein level was found by mass spectrometry. 43 were identified in the samples of both sisters and some were only observable in a sample coming from only one of the sisters, 18 for S1 and 16 for S2 (Figure V-13). 29 genomic mutations specific to only one sister were also observed for sister S1 and S2; 13 and 16 respectively.

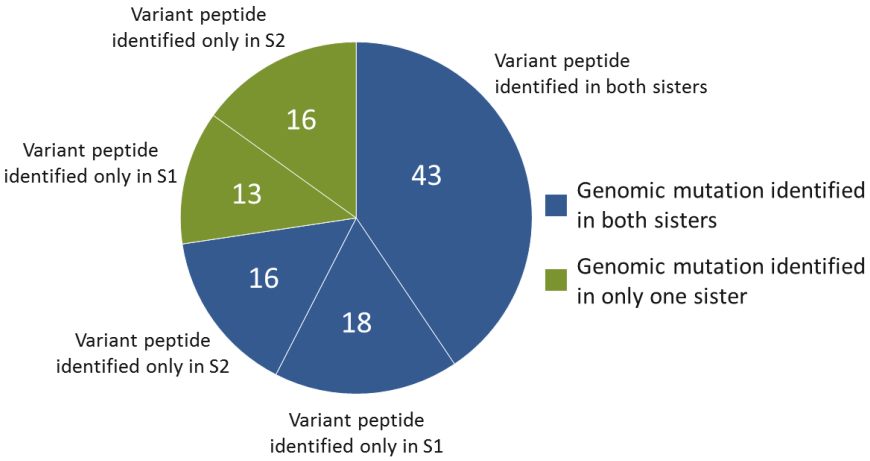



**Figure V-13 : Distribution of the sequence variants showing if the genomic mutation is present in the genome of both sisters and the number of sequence variants observed by mass spectrometry.**

### A.4.1 The use of personalized database enables the identification of sequence variants

Our Proteogenomic approach improved protein identification and increased the protein sequence coverage. An example of this can be seen in Figure V-14 where the benefits of using a personalized database can clearly be seen. The search against the personalized database enabled the identification of two additional peptides for the protein presented in the figure. The two peptides contained individual-specific genomic variants. An aspartic acid and a lysine were replaced by two glutamic acids in the peptide sequence VLWLDEIQQAVDDANVDKDR. A leucine was replaced by a valine in the peptide sequence QTFIDNTDSIVK. Using the consensus reference proteome this information was totally missed.

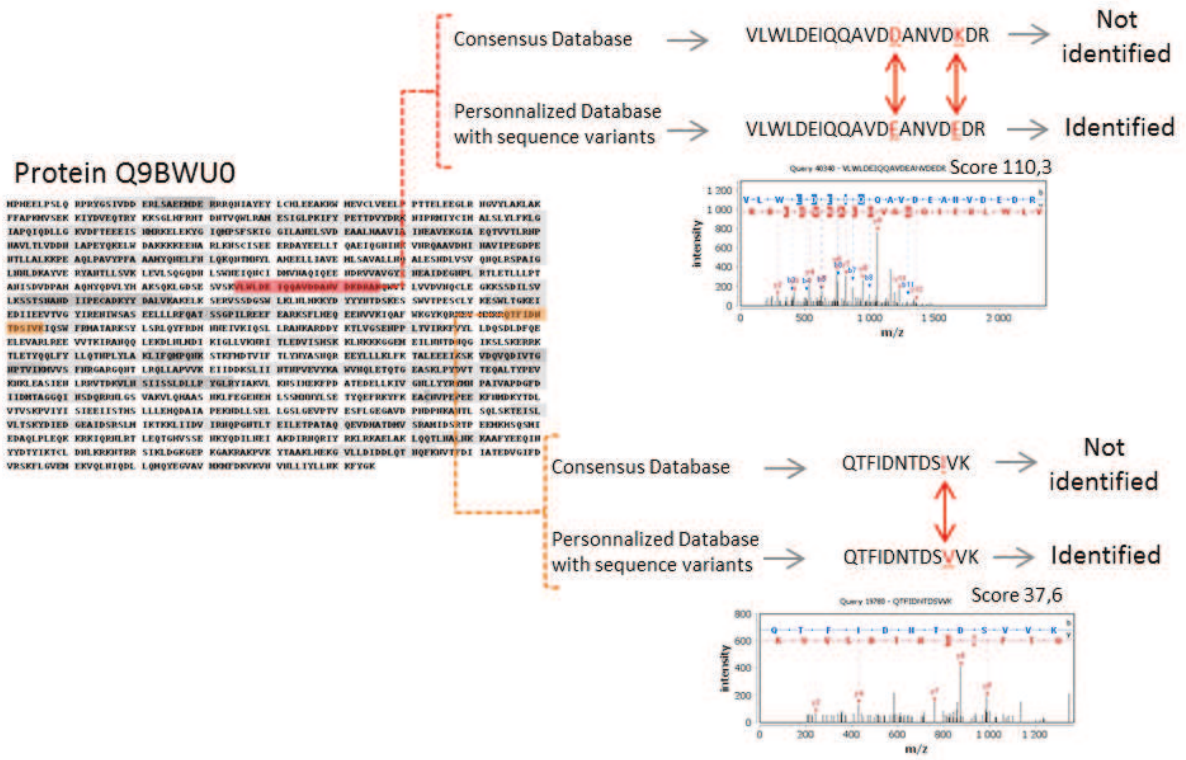



**Figure V-14 : Benefits of proteogenomics for protein identification and sequence coverage.**
The use of personalized databases improves the proteome coverage. The search against the personalized database enabled the identification of two additional peptides for this protein. The two peptides contained individual-specific genomic variants. An aspartic acid and a lysine were replaced by two glutamic acids in the peptide sequence VLWLDEIQQAVDDANVDKDR. A leucine was replaced by a valine in the peptide sequence QTFIDNTDSIVK.

### A.4.2 The use of personalized database enables the identification of new expressed splice variants

The use of a sample-specific databases containing information derived from RNA-seq enabled the identification of two novel splice-variants. Figure V-15 shows an example of a splice variant that is longer than the protein sequence annotated in the consensus reference database. The protein that was translated from RNA-seq data starts 178 amino acids before the annotated position. This splice isoform could be identified in both S1 and S2. The peptide that was identified is highlighted in red. The sequence was blasted against the whole human proteome and it was found to be unique for this proteoform.


**Figure V-15 : Alignment of the consensus UniprotKB/Swissprot sequence and the personalized sequence derived from RNA-sequencing.**
The peptide identified is highlighted in red. The new splice variant sequence for protein O15027 starts before the position annotated in the UniprotKB/Swissprot reference database.

### A.4.3 The use of personalized database enables the identification of patient-specific sequence variants

Another important result that can be obtained when using personalized databases is the identification of protein sequence variants specific to each patient. In Figure V-16 the identification of a proteoform specific to each sister was done. For sister S1 the protein was identified with a peptide having the same sequence as the one present in the reference database. However, for sister S2, the same protein was identified with a peptide having a sequence variant. In this example a Proline is substituted by a Threonine. This result is also important since this protein was identified for sister 2 solely by this peptide sequence.


**Figure V-16 : Identification of protein sequence variants specific to each individual.**
This protein was identified in the samples from the two sisters but they differed by a single amino acid variant. In the personalized database for sister 2 a Proline is substituted by a Threonine.

### A.4.4 The use of personalized database enables the identification of allelic pair products

Our approach does not only improve proteome coverage and the identification of individual-specific variants, but also supports the identification of allelic pair products. An example can be seen in Figure V-17. The peptides shown in the figure belong to the same canonical protein (P13489) and two heterozygote forms of this protein are shown. Sister 2 is homozygote for this gene and only the protein sequence present in the reference database was identified. Two peptides confirm this identification (ELTVSNNDINEAGVR and VLCQGLK). For Sister 1 a SNP was sequenced in her genome. Two proteoforms were identified in mass spectrometry, the protein sequence from the reference database and the personalized protein sequence containing the single amino acid variant. Sister 1 is heterozygote for this gene and evidence at protein level for the expression of both alleles was found.

The identification of the products of allele pairs was only possible in our approach because we included the reference protein database in the personalized database. All heterozygote products we found were constituted of a SNP seen in the genome and the corresponding reference protein sequence. This is due to the fact that when the SNP calling was made, if two mutations were found in the same position when comparing to the reference genome, then both mutations were discarded to minimize false positives. This is done because having two mutations at a unique position is extremely rare. Proteomics can in this case be an extremely useful tool to validate genome sequencing data.

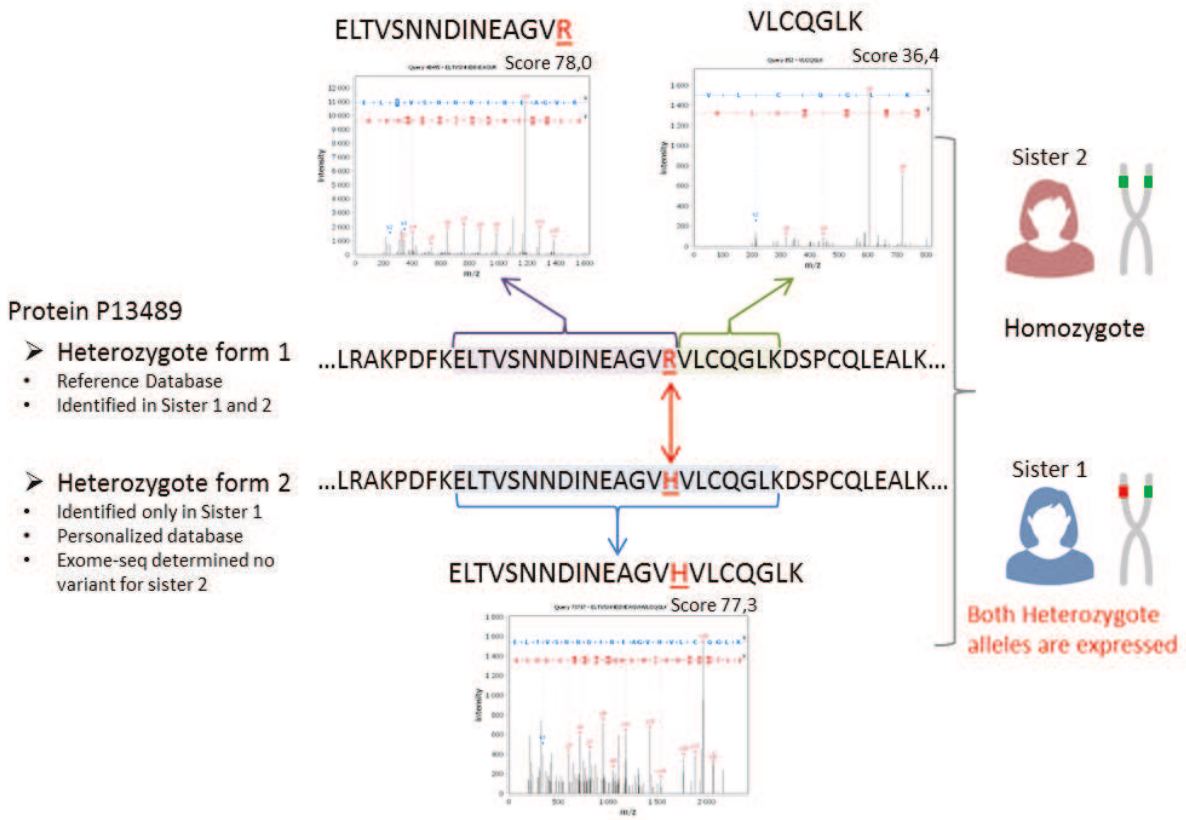



**Figure V-17 : Examples of identified peptides belonging to two proteins from heterozygote genes.**
Two heterozygote forms are shown for protein P13489. The identified peptide and its corresponding spectrum are shown. Sister 2 is homozygote for this gene. However, Sister 1 is heterozygote for this gene and both alleles are expressed.

### A.4.5 The use of personalized database enables to improve protein quantification

A relative quantification analysis by spectral count was done using a Bruker Daltonics MaXis 4G mass spectrometer. In this experiment 1200 protein groups and 7200 different peptides per sample could be identified. This experiment enabled the identification of 757 up- and down-regulated proteins when comparing two conditions: protein extract from cells in a basal media (Mi) and protein extract from cells after a stimuli with LPS (LPS).

The use of personalized databases for protein identification improves protein coverage and characterization which consequently improves protein quantification. As seen above, in a classical proteomics workflow sample-specific information is lost and this could result in an erroneous estimation of protein abundances. To illustrate this in Figure V-18.A. the results of a relative quantification by spectral count are shown. Two conditions are compared: protein extract from cells in a basal media (Mi) and protein extract from cells after a stimuli with LPS (LPS). A stretch of protein P26373 shows the search results using the consensus database (UniProtKB-Swissprot) and the personalized database. The identified peptides are underlined. When using the consensus database only a single peptide was identified and the spectral count values for protein P26373 show a non-significant change in abundance. In a high throughput analysis this protein would have not have been considered for further investigation since its abundance is not affected by the stimuli. However, when using the personalized database, an additional tryptic peptide including a patient-specific variant was identified. In peptide STESLQANVQR the Alanine is substituted by a Threonine at position 112. The relative quantification using spectral count is more accurate as the sequence coverage is higher. A significant overexpression of the protein could be detected. This protein now becomes a target for further examination.

Finally having patient-specific information can allow quantifying allele-specific products. In Figure V-18.B. the canonical protein P32455 was found to be present as two proteoforms originating from two different alleles of the

same gene. Two peptides, one with Threonine and one with Serine at position 349, were identified. The spectral count results showed that one of the heterozygote forms is overexpressed in the LPS condition whereas the other is not. This example demonstrates that allele-specific quantification is possible at protein level.

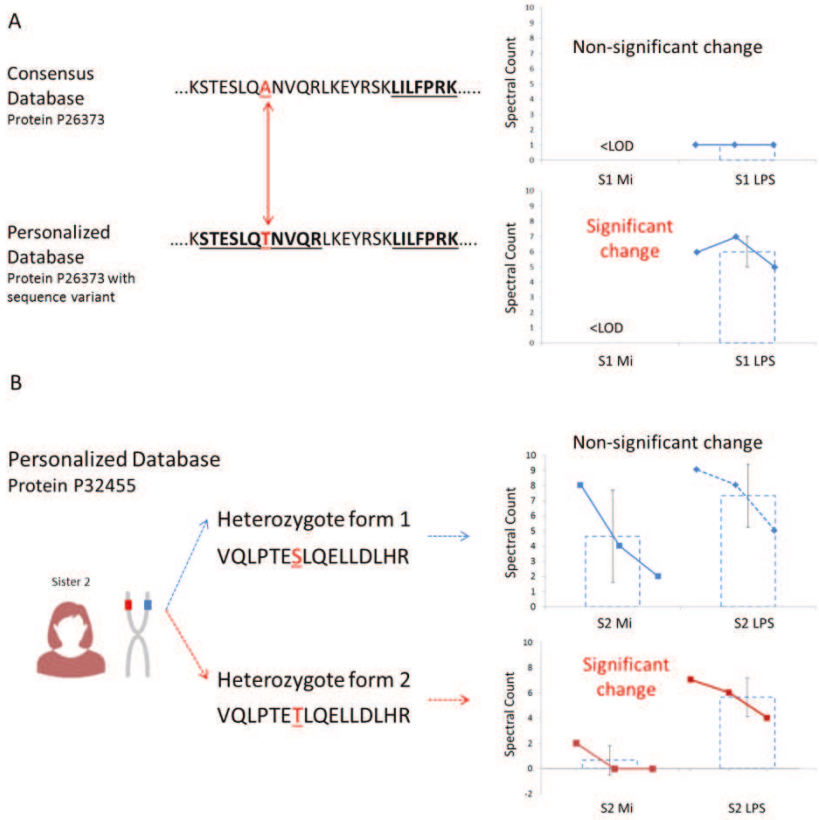



**Figure V-18 : Personalized databases imply more accurate protein quantification.**
A. Using a classical approach with a consensus database only one peptide was identified (peptide underlined) and the spectral count results show a non-significant change. However, when using a personalized database an additional peptide is found and the relative quantification shows an overexpression for this protein. B. Using a personalized database the quantification of allele-specific products is possible at the protein level. Here the heterozygote form 1 has a non-significant change whereas the heterozygote form 2 is overexpressed in the LPS condition.

### *A.5. Conclusion*

In conclusion our approach demonstrated the potential of personalized databases to improve the proteome characterization. We identified 106 nonsynonymous sequence variants on 96 proteins using the exome-seq derived databases and 2 new splice variants using the RNA-seq derived database.

It also enabled the identification of protein sequence variants specific for each patient and the unambiguous identification of allele-specific products. This improvement of protein identification implies the possibility of a more accurate quantification. And it also opens the possibility of quantifying allele-specific products at the protein level.

Finally this information can be extremely useful to understand the phenotype of the two sisters. The genomic and transcriptomic findings can be of higher value if they are propagated to the protein level, as they have a higher possibility of being functionally significant. In turn, proteomic data can be a tool to validate and filter DNA/RNA sequencing data.

The integrative analysis of the exome sequencing and the relative quantification of the transcriptome and the proteome to provide biological insights of the HIDs syndrome are underway.

A publication resuming the results from this project is in preparation.

## B. Challenges and Perspectives of proteogenomics

One of the main challenges in Proteogenomics is the bioinformatics. Several tools have appeared recently to create customized databases from Genomic information [247-249], RNA-Seq [250, 251] and RIBO-Seq [252]. However, these tools are very specific to the application they were developed for. To be able to obtain software that can use any type of genomics, transcriptomic and proteomic data to build custom databases, more standardization of files and annotations is necessary. This would help to build bridges between gene variations and proteoforms. Also a paradigm shift in proteomic software is necessary to move from software highly depending on a single database to multi-omic strategies which can capture and highlight small differences in proteoforms. Also search engines must evolve to let users benefit from meta-information already available in databases or sample-specific DNA/RNA information.

Furthermore when creating customized databases it is important to control false discovery rates. Nesvizhskii et al. proposed in 2014 a series of guidelines for the validation of novel peptides identified by proteogenomics [9]. Peptides identified with custom databases should be queried against all major reference databases for the organism of interest and common sample contaminants. Different FDR estimation should be determined separately for known and novel peptides. Efforts must be done to eliminate the most likely sources of false positives (PTMs, chemical modifications, errors in mass measurements).

Compared to genomic and transcriptomic sequencing, the coverage of the proteome is still lacking behind. To fully characterize a proteome, intensive fractionation is still necessary because of the large dynamic range of protein abundance. One of the reasons that Proteogenomic is still in its infancy is that sensitivity of mass spectrometers and thus global coverage, did not reach a sufficient level until now. Additionally there is the inherent problem that not all tryptic peptides have appropriate physico-chemical properties to be analyzed by LC-MS. However in recent years instrumental progress in sensitivity has enabled Proteomics to reach a significantly improved maturity of sequencing techniques.

Proteogenomics promises to better characterize the proteome. It can provide the direct evidence that a variation in the genome is ultimately translated into a protein. This gives a higher significance to this variation as it can have a direct impact on the phenotype. Proteogenomics could set the first steps towards personalized medicine using patient-specific -omic information that would help to identify more accurately the causes of a certain phenotype.

Finally, proteogenomics is key to correctly quantify a proteome. By performing proteomics analysis using incomplete databases important information is lost.

# General Conclusion

My doctoral work intended to improve the proteome characterization by quantitative mass spectrometry and Proteogenomic method development. The methodological developments were optimized for and applied to several biological projects.

The first and second parts of this manuscript were focused on describing the state of the art of bottom-up proteomics and proteogenomics. It describes the strategies used to identify and characterize proteins in complex sample matrices by LC-MS-MS/MS. The strategies for global or targeted protein quantification are then described, including recent methodologies promising a comprehensive proteome analysis using Data Independent Acquisition mode. The state of the art of proteogenomics and N-terminomics was also described.

In this context of Bottom-up Proteomic analysis the objectives of this thesis towards the improvement of proteome characterization were defined:

- The improvement of the method development workflow for targeted proteomics by SRM/PRM.
- The development of sample preparation protocols compatible with quantitative studies.
- The introduction of standard peptides, both retention time standards and heavy labelled peptides to optimize all peptide-specific parameters.
- The development of standard and well-characterized samples to assess LC-MS platform performances compatible with a routine usage.
- The development of the analytical strategy and the data-treatment workflow for high-throughput N-terminomics analysis by the dN-TOP approach.
- The improvement of proteome characterization with the use of personalized databases derived from exome sequencing and/or RNA-Seq data.

In response to these questions the results presented in this thesis have helped to reach clear conclusions:

**Quantitative Proteomics:** The SRM assay method development workflow was optimized and the key parameters to increase the selectivity and the sensitivity of targeted quantitative methods were determined. The organization of each step enabled to obtain a fast and reliable method development workflow that was routinely used throughout my thesis. The introduction of retention time standard peptides to normalize retention times, and heavy labelled peptides to optimize all peptide-specific parameters (collision energies, choice of the monitored transitions, retention times…) fastened the development of targeted quantification methods. These standards also highly increased the selectivity, multiplexing and the overall throughput capabilities of the method.

The optimization and method development of DDA, SRM, PRM and DIA methods was shown throughout this thesis. As I was responsible for the maintenance, the method development, the training of new users and the day-to-day operations of LC-MS systems in the laboratory, my task was to evaluate the different acquisition methods available, prepare default acquisition methods and train new users to setup their methods in the laboratory. In this context, I gained expertise in developing methods adapted to different types of samples and determined the key parameters to be optimized according to the objectives of each study.

I developed internal and external standard samples for quality control to improve the reliability of the analysis and have an accurate monitoring of instrumental performances. For quantification studies, the data must be reproducible and accurate. To achieve this, the LC-MS system must be regularly monitored to avoid confounding errors due to instrumental perturbations. And in order to assess the performances of a given step in a proteomic workflow (a sample preparation protocol, the LC-MS system performances, a given instrument tuning, a bioinformatic pipeline…) well-characterized standard samples were developed. This tool revealed to be extremely useful to evaluate and improve analytical workflows such as the development of a sensitive and objective performance test for LC-SRM platforms, the side-by-side evaluation of label-free bioinformatic pipelines and the improvement of signal extraction of Data-Independent Acquisition. This method enabled understanding where the limitations of each methodology lie; and this is absolutely necessary to propose innovative solutions. These limitations can originate from analytics or informatics.

Additionally I evaluated sample preparation protocols compatible with quantification methods. I showed, after optimizing all main parameters of a Stacking gel SDS-PAGE protocol, that this strategy not only is fully compatible with targeted proteomics but it provides even better performances than a classical liquid digestion protocol.

All these methodological developments were successfully applied to the quantification of 13 microbial proteins related to Crohn's disease in the human gut microbiome, a sample of extreme complexity, without fractionation. The microbial proteins could be quantified and the trends of under- and overrepresentation in Crohn's disease patients observed in a previous discovery experiment were confirmed. These biomarkers could help to reliably diagnose Crohn's disease, discriminate between similar intestinal pathologies and assess the therapeutic efficiency of treatments directed against the disease.

Moreover the development of an LC-SRM assay enabled the detection and relative quantification of proteins, MetAP1 and MetAP2, that could not be detected in a global shotgun experiment or using an immunodetection assay. The SRM results were well correlated with an independent mRNA quantification experiment performed by our collaborators. Using an absolute quantification method targeting MetAP2 we could verify this results.

Lastly in bottom-up proteomics the most commonly used approach for protein identification highly depends on the protein database, this must be as complete and as adapted as possible to the analyzed sample. However due to the high multiplicity of proteins (proteoforms) that can be present in a sample, a consensus database will never completely represent the protein content. This has negative consequences in the quantification of a protein as the coverage of a given protein could only be partial and thus possibly missing the full state of the protein. The quantification of a proteome can only benefit from extending the reach of bottom-up proteomics to comprehensively analyze the proteome. In this context Proteogenomics can have a significant beneficial impact in the quantification of a proteome.

**Proteogenomics:** An N-terminomic approach based on the specific chemical labelling of proteins' N-termini using the TMPP reagent has been developed by the LSMBO. I presented in this manuscript my work in the optimization of this method and the engineering of an automated workflow for the data treatment. The limitations related to this approach were determined and corrected to be able to obtain a reliable and accurate approach enabling high-throughput analysis of N-terminomes. This method was applied to deeply characterize the proteome and N-terminome of human mitochondria.

Finally a personalized multi-omics profiling strategy was developed to improve the proteome characterization with the use of personalized databases derived from exome sequencing and RNA-Seq data. This method was applied to the study of a rare disease, hyperimmunoglobulinemia D and periodic fever syndrome (HIDS) characterized by recurrent fever with inflammatory symptoms. With this approach, nonsynonymous sequence variants and new splice-isoforms could be identified. Also the identification of protein sequence variants specific for each patient and the unambiguous identification of allele-specific products were possible.

This improvement of protein identification implies the possibility of a more accurate quantification. And it also opens the possibility of quantifying allele-specific products at the protein level.

Finally this extended information revealed to be extremely useful to understand the phenotype of the two sisters. The genomic and transcriptomic findings can be of higher value if they are propagated to the protein level, as they have a higher possibility of being functionally significant. In turn, Proteomics data can be a tool to validate and filter DNA/RNA sequencing data.

However the main challenges in Proteogenomics remains the bioinformatics. To be able to obtain software that can use any type of genomics, transcriptomic and proteomic data to build custom databases, more standardization of files and annotations is still necessary. This will help to build bridges between gene variations and proteoforms. Also a paradigm shift in proteomic softwares is necessary to move from software highly depending on a single database to multi-omic strategies which can capture and highlight small differences in proteoforms.

Finally Proteogenomics promises to better characterize the proteome and this is key to obtain accurate protein quantification data.

# Part VI Experimental section

## A. Unfractionated stacking Gel SDS-PAGE protein purification protocol

Unfractionated stacking Gel SDS-PAGE protein purification protocol was carried out using a 4% polyacrylamide stacking gel and a 10% running gel in a Mini PROTEAN cell (Bio-Rad). Running gels are only used to hold stacking gels in place. 50 µg of proteins, in freshly prepared denaturating buffer were loaded per lane. A loading-well containing only the loading buffer was intercalated between each sample. The same loading volumes were used for all samples. No sample was loaded on the loading wells at the border of the gels. The samples were migrated over 1cm in the stacking gel (50 V for 25min). After migration, the gels were washed with water and fixed using 50:50 methanol:water (v:v)/3% Phosphoric acid. Gels were stained by a colloidal blue method (G250, Fluka, Buchs, Switzerland). Gel bands were excised manually using a ruler and a bistoury.

## B. Evaluation of instruments and acquisition methods performance using isotopologue peptides

### *B.1. Materials*

Modified porcine trypsin and the 6 × 5 LC-MS/MS Peptide Reference Mix was obtained from Promega (Madison, WI, US). Eight synthetic stable-isotope $^{15}$N- and $^{13}$C-labeled peptides were acquired from Thermo Fisher Scientific (Ulm, Germany). Bovine Serum Albumin Digest was purchased from Bruker Daltonics (Bremen, Germany). All other chemicals were obtained from Sigma-Aldrich (St. Louis, MO, US) unless otherwise specified.

### *B.2. Sample Preparation*

**Yeast Digest.** S. cerevisiae protein extracts were prepared using the strain CEN.PK113-7D. Yeast cells were collected by centrifugation (3000g, 10 min, 5°C) and washed with MilliQ water. The cells were resuspended in extraction buffer (50mM NaPO3, pH 7,3, 1mM EDTA, 5% glycerol and protease inhibitors). The cells were disrupted using glass beads (10x30s vortexing steps on ice). After centrifugation (5400g, 5min, 5°C) the supernatant was removed and the beads were washed using extraction buffer. The supernatant was centrifuged (20000g, 30 min, 5°C) to eliminate any cell debris. Protein concentration was determined by Bradford assay and aliquots of 1µg of total protein were created. Acetone precipitation was carried out (6 volumes of acetone to 1 volume of sample; overnight; -20°C). After centrifugation (15 min, 15000g, 4°C), the protein pellet was dried and resuspended in solubilization buffer (8M urea, 0.1M Ammonium bicarbonate, pH8). Proteins were reduced (12mM DTT, 37°C, 30 min) and alkylated (40mM IAA, 25°C, 1 hour, dark). The samples were diluted to reach a 1M urea concentration using a solution of freshly prepared 0.1M ammonium bicarbonate and proteins were digested using Trypsin (1:100 enzyme:substrate ratio, 37°C, overnight). After acidification using formic acid (pH 3), the samples were desalted and concentrated using solid phase extraction (Sep-Pak C18, 1cc, 50mg, Waters). The volume was reduced in a vacuum centrifuge and adjusted in water + 0.1% formic acid to reach a final concentration of 1µg/µl.

**Sample set 1:** Eight synthetic stable-isotope $^{15}$N- and $^{13}$C-labeled isotopologue peptides based on the peptide sequence AALPAAFK were mixed in the following concentrations: AALPAAFK 300 amol, AALPAAFK 900 amol, AALPAAFK 2,7 fmol, AALPAAFK 8,1 fmol, AALPAAFK 24,3 fmol, AALPAAFK 72,9 fmol, AALPAAFK 218,7 fmol and

AALPAAFK 656,1 fmol. The mixture of isotopologue peptides was diluted by a factor of two in a background matrix (either 5 fmol/µl of BSA digest or 50ng/µl of total yeast digest). Then two dilutions by a factor of 10 and 100 were done by cascade dilution using the background matrix. The background matrix is used to mimic a proteomic sample and also to avoid peptide adsorption to the walls of vials so it should always be added first. Two microliters were analyzed by LC-MS/MS and each solution was analyzed in triplicate. This resulted in nine injections of three solutions of isotopologues peptides covering a range of 5,3 logs, i.e. 3 amol to 656 fmol injected on column.

**Sample set 2:** Two picomoles of the lyophilized 6 × 5 LC-MS/MS Peptide Reference Mix was resuspended in 200µl of water+0,1% formic acid and then diluted by a factor of two in a background matrix (either 5 fmol/µl of BSA digest or 50ng/µl of total yeast digest). The background matrix is used to mimic a proteomic sample and also to avoid peptide adsorption to the walls of vials so it should always be added first. Two microliters were analyzed by LC-MS/MS and each solution was analyzed in triplicate. This mixture contains 6 sets of isotopologue peptides. However the most hydrophilic peptide could not be detected as a trapping column was used and the peptide was not retained. For each peptide a calibration curve could be created with the following calibration points: 30 amol, 300 amol, 3fmol, 30fmol and 300 fmol of injected amount on column.

### *B.3. LC-MS/MS*

**Nano Liquid Chromatography.** NanoLC-MS/MS analyses were performed on a NanoAcquity LC-system (Waters, Milford, MA, USA) coupled to a mass spectrometer. For each analysis a volume of 2µl of sample was injected into a Symmetry C18 precolumn (0.18x20mm, 5µm particle size, Waters) and then peptides were separated using an ACQUITY UPLC® BEH130 C18 separation column (75µm×200mm, 1.7µm particle size, Waters). Peptides were eluted using a gradient of water + 0.1% formic acid (solvent A) and acetonitrile +0.1% formic acid (solvent B). Peptide trapping was performed during 3 min at a flow rate of 5µL/min with 99% A and 1% B and elution was performed at 50 °C at a flow rate of 300 nL/min when the system was coupled to a AB Sciex Triple TOF 6600 (AB SCIEX, Framingham, US) and a flow rate of 450 nL/min when coupled to another mass spectrometer. For simple matrix samples (BSA digest) the following gradient was used: from 3 %B to 25 %B over 15 minutes, from 25 %B to 50 %B in 2 minutes, from 50%B to 80%B in 2minute, 80%B for 3 minutes and the column was reconditioned at 3%B. For complex matrix samples (whole yeast digest) the following gradient was used: from 3 %B to 8%B over 2 minutes, from 8%B to 35%B in 77minutes, from 35%B to 90%B in 1minute, 90%B for 5 minutes, from 90%B to 3% in 2minutes and the column was reconditioned at 3%B.

**AB Sciex Triple-TOF 6600.** NanoLC-MS/MS analyses were performed on a NanoAcquity LC-system (Waters, Milford, MA, USA) coupled to a TripleTOF 6600 (AB SCIEX, Framingham, US). The mass spectrometer was operated in positive mode, with the following settings: ionspray voltage floating (ISVF) 2300 V, curtain gas (CUR) 30, interface heater temperature (IHT) 75, ion source gas 1 (GS1) 2, declustering potential (DP) 80 V. For PRM, an MS survey scan was acquired followed by a set of sequential Q1 isolation windows. The MS scan had a dwell time of 250ms in the mass range of 350-1250 m/z and the MS/MS scans 100 ms in the mass range of 100 – 1800 in high sensitivity mode giving a total cycle time of 1.1s. The CE was adapted to each peptide calculated using Skyline using the following equation CE= 0,036m/z + 8,857. For SWATH acquisition, two types of methods were used. The first type of method covered the 350-1200 m/z range corresponding to the mass range of the majority of tryptic peptides in a bottom-up proteomic experiment. A MS survey scan was acquired followed by a set of 34 sequential Q1 windows with a fixed width 25 Da. The accumulation time for each MS and MS/MS experiment was respectively 250ms and 82.4 ms for a total cycle time

of 3.1 s. The CE for each window was determined according to the calculation for a charge 2+ ion centered upon the window with a spread of 15. The second type of SWATH method used only 8 isolation windows in order to be able to obtain comparable cycle and dwell times as with the PRM experiments. For each method, a MS survey scan covering the precursor m/z range of 350–1250 Da was acquired followed by a set of 8 sequential Q1 windows with either a width of 4, 25 or 50 Da covering the range of 390-422 Da, 369-569 Da and 375-775 Da respectively. The accumulation time for each MS and MS/MS experiment was respectively 250ms and 100 ms for a total cycle time of 1.1 s. The complete system was fully controlled by Analyst TF v1.7 software.

**Thermo Fischer Q-Exactive plus.** NanoLC-MS/MS analyses were performed on a NanoAcquity LC-system (Waters, Milford, MA, USA) coupled to a Q-Exactive Plus (Thermo Fisher Scientific, Waltham, MA, USA) mass spectrometer. The mass spectrometer was operated in positive mode. For PRM analyses the instruments was set to the following parameters: a full MS scan at a resolution of 70000 (at m/z 200), AGC target $3 \times 10^6$, and a 50 ms maximum injection time. Full MS scans were followed by 8 PRM scans at 17500 resolution (at m/z 200), AGC target of $1 \times 10^6$, 100 ms maximum injection time, isolation windows of 2Da, normalized collision energy (NCE) of 27. MS/MS scans were acquired with a starting mass range of 100 m/z and acquired as a profile spectrum data type. For Data-independent acquisition analysis the instruments was performed using the following parameters: a full MS scan at a resolution of 17,500 (at m/z 200), AGC target $3 \times 10^6$, and a 50 ms maximum injection time. Full MS scans were followed by a DIA experiment at 17,500 resolution (at m/z 200), AGC target of $1 \times 10^5$, 100 ms maximum injection time, count loop of 10, a default charge of 2, isolation windows of 4Da, normalized collision energy (NCE) of 27. MS/MS scans were acquired with a starting mass range of 100 m/z and acquired as a profile spectrum data type. The complete system was fully controlled by Xcalibur 3.0.63 (Thermo Fisher Scientific).

**Bruker Daltonics Impact II.** NanoLC-MS/MS analyses were performed on a NanoAcquity LC-system (Waters, Milford, MA, USA) coupled to a Q-TOF Impact II (Bruker Daltonics, Bremen, Germany). The mass spectrometer was equipped with a CaptiveSpray source and a nanoBooster operating in positive mode, with the following settings: source temperature was set at 150 °C while drying gas flow was at 3 L/min. The nano-electrospray voltage was optimized to −1300 V. Mass correction was achieved by recalibration of acquired spectra to the applied lock masses hexakis (2,2,3,3,-tetrafluoropropoxy) phosphazine ($[M+H]^+$ = 922.0098 m/z)]. For PRM analyses the instruments was set to scan sequentially 8 PRM (MRM) experiments with the following parameters: isolation windows of 2Da, collision energy of 28, MS/MS scans were acquired with a rolling average of 2, at a scan rate of 10 Hz, with a mass range of 150 to 2200 m/z and acquired as a centroid spectrum data type. The complete system was fully controlled by Hystar 3.2 (Bruker Daltonics, Bremen, Germany).

**Waters Synapt HDMS.** NanoLC-MS/MS analyses were performed on a NanoAcquity LC-system (Waters, Milford, MA, USA) coupled to a Synapt HDMS mass spectrometer (Waters Corp., Milford, USA). The internal parameters of the Synapt HDMS were set as follows: The electrospray capillary voltage was set to 3.2 kV, the cone voltage set to 35 V, and the source temperature set to 90°C. The MS survey scan was m/z 250–800 with a scan time of 0.5 s. Calibration was performed using GFP in 50% acetonitrile + 0.1% formic acid. Data acquisition was piloted by MassLynx software V4.1.

**Thermo Ficher TSQ Vantage.** SRM analysis was done on a capillaryLC-SRM platform: Dionex Ultimate 3000 system coupled to a TSQ Vantage Triple Quadrupole instrument (Thermo Fischer Scientific, San Jose, CA, USA). For each

analysis, a volume of 2µL of sample was injected and trapped on a precolumn (Zorbax C18 stable bond, 5 µm, 1.0 × 17 mm, Agilent Technologies) then separated on a C18 column (Zorbax 300 SB C18, 3.5 µm, 150 × 0.3 mm, Agilent Technologies). Trapping was performed for 3 min at a flow rate of 50 µL.min$^{-1}$ with solvent A. Elution was performed at a flow rate of 5 µL.min$^{-1}$. For complex matrix samples (whole yeast digest) the following gradient was used: from 3 %B to 8%B over 2 minutes, from 8%B to 35%B in 77minutes, from 35%B to 90%B in 1minute, 90%B for 5 minutes, from 90%B to 3% in 2minutes and the column was reconditioned at 3%B. The TSQ vantage mass spectrometer was operated with the following parameters: The system was operated in positive mode, the ion spray voltage was set at 3000V, the capillary temperature at 300°C, the nitrogen collision gas pressure was set to 1.5 mTorr, Q1 and Q3 resolution set to 0.7 Da and the collision energy was individually optimized for each transition. The system was controlled by Chromeleon Xpress software (v. 6.8) for the liquid chromatography system and Xcalibur (v. 2.1.0) software for the mass spectrometry system.

### *B.4. Data Analysis*

The Skyline open-source software package[15] was used to visualize the data, to perform peak picking and integration of transition peak areas, and to manually verify the correct peak group identification by checking the exact coelution of isotopologue peptides. For MS1 signals multiple isotopic precursors were extracted (P, P + 1 and P + 2) and for MS2 signals at least 4 transitions were extracted per peptide to provide confirmation of proper identification of the selected peak and verify the absence of interferences. The linearity range required experimental dots in standard curves to exhibit an average CV precision that was below 20% among triplicate injections. Experimental dots also had to fall within the average 80–120% accuracy range in calculating expected injected amounts using regression equations after logarithmic transformation. The coefficient of determination $R^2$ should be higher than 0,98 between the area under the peaks and the injected amount on column, and between the recalculated and the real injected amount on column. All signals were visually evaluated and validated to ensure high-quality results. The limit of quantitation (LOQ) is the lowest point satisfying all the criteria reported above. Only the points satisfying all these criteria were used to calculate the linear regression equation and coefficient of determination.

## C. Application of targeted proteomics to validate Crohn's disease biomarkers

### *C.1. MicroLC-SRM parameters*

After in-gel reduction and alkylation using a MassPrep Station (Waters, Milford, MA), the protein bands excised from the stacking gel were in-gel digested using a 1:100 trypsin:protein ratio (Promega, Madison, WI) overnight at 37 °C. The resulting tryptic peptides were extracted using 60% acetonitrile in 0.1% formic acid for 1h at room temperature. Equal amounts of the concentration-balanced mixture of stable isotope-labeled crude peptides were spiked in each peptide extract. The total volume was reduced in a vacuum centrifuge and adjusted to 15µl using 0.1% formic acid in water before microLC-SRM analysis. Peptides were analyzed on a Dionex Ultimate 3000 system coupled to a TSQ Vantage Triple Quadrupole instrument (Thermo Fischer Scientific, San Jose, CA, USA). For each analysis, a volume of 1.5µL of sample, i.e. 10µg of protein, was injected and trapped on a precolumn (Zorbax C18 stable bond, 5 µm, 1.0 × 17 mm, Agilent Technologies) then separated on a C18 column (Zorbax 300 SB C18, 3.5 µm, 150 × 0.3 mm, Agilent Technologies). The peptides were eluted with a linear gradient of 2% acetonitrile/98% water/0.1% formic acid (solvent A) and 98% acetonitrile/2% water/0.1% formic acid (solvent B). Trapping was performed for 3 min at a flow rate of 50 µL·min−1 with solvent A. Elution was performed at flow rate of 5 µL.min$^{-1}$ using a two-step optimized gradient : Step One (Elution gradient): 3min at 5% B; from 5% to 35% B in 40 min; 5min at 80% B; 2min at 5% B; Step Two (Column

washing and regeneration gradient): from 5% B to 50% B in 5min; 2min at 80% B; 15min at 5% B. For optimal microLC-SRM, the TSQ vantage mass spectrometer was operated with the following parameters. Triplicate injections of each sample were performed with two distinct methods, each monitoring a subset of all transitions (supplementary Table 4). The system was operated in positive mode, the ion spray voltage was set at 3000V, the capillary temperature at 300°C, the argon collision gas pressure was set to 1.5 mTorr, Q1 and Q3 resolution set to 0.7 Da and the collision energy was calculated using the following equation: CE=0.03 x (Precursor ion m/z) + 2.905. Scheduled SRM was used for data acquisition, each transition was monitored during a 7 minutes time window centered at previously determined peptide retention times, with a cycle time of 2.6 s and minimal dwell times of 22ms and 25ms for the SRM method 1 and 2, respectively. The system was controlled by Chromeleon Xpress software (v. 6.8) for the liquid chromatography system and Xcalibur (v. 2.1.0) software for the mass spectrometry system.

## D. Application of targeted proteomics for the relative and absolute quantification of Methionine Aminopeptidase Proteins

### *D.1. Single-band resolving gel*

Cell pellets were disrupted using glass beads. The extraction buffer was added (50mM Hepes/NaOH pH7,2, 1,5mM $MgCl_2$, 1mM EGTA, 10% Glycerol, 1% Triton, 2mM PMSF, 150mM NaCl and protease inhibitors). The samples were vortexed and centrifuged at 4°C. The supernatant was collected and the total protein amount was determined by the Bradford assay. The samples were loaded on a 10% monodimensional SDS-PAGE (10.1 cmx7.3cm) on a mini PROTEAN (Bio-Rad) apparatus. 10 µg were loaded. The electrophoretic migration was stopped as soon as the protein sample entered the resolving gel. After the migration, the gels were washed with water and fixed using3% Phosphoric acid in 50:50 methanol:water (v:v). Gels were stained by a colloidal coomassie blue method (G250, Fluka, Buchs, Switzerland) and the band containing all the proteins was cut. The gel bands were washed to get rid of the coomassie blue dye. Proteins were reduced, alkylated and digested using trypsin. (1:100 enzyme:substrate ratio, 37°C, overnight). Resulting tryptic peptides were extracted using 60% ACN in 0.1% formic acid for 1h at room temperature. Stable isotope-labeled synthetic peptides were spiked in the samples. The volume was reduced in a vacuum centrifuge and resuspended using 0.1% formic acid in water before nanoLC-MS/MS analysis.

### *D.2. Liquid digestion Protocol 1*

Cell pellets were disrupted using glass beads. The extraction buffer was added (50mM Hepes/NaOH pH7,2, 1,5mM $MgCl_2$, 1mM EGTA, 10% Glycerol, 1% Triton, 2mM PMSF, 150mM NaCl and protease inhibitors). The samples were vortexed and centrifuged at 4°C. The supernatant was collected and the total protein amount was determined by the Bradford assay. The aliquots were stripped of non-protein contaminants using an acetone protein precipitation step (4 to 1 volumes of acetone, -20°C, 2 hours). After centrifugation (10 min, 13000g, 4°C), the protein pellet was dried and resuspended in 50 mM ammo nium bicarbonate. Proteins were reduced, alkylated and digested twice using TPCK modified porcine trypsin (2x[1:100 enzyme:substrate ratio, 37°C, 1,5 hours]). Stable isotope-labeled synthetic peptides were spiked in the samples and after acidification using formic acid (pH 3), the samples were desalted and concentrated using solid phase extraction (Sep-Pak C18, 1cc, 50mg, Waters). The eluate volume was reduced in a vacuum centrifuge and adjusted using water + 0.1% formic acid to reach a final concentration of 4µg/µl.

### D.3. Liquid digestion Protocol 2

Cell pellets were resuspended in 800µl of extraction buffer (8M urea, 2M Thiourea, 1%DTT, protease inhibitors), sonicated with a needle and centrifuged (5 min, 8000g, 4°C). The aliquots were stripped of non-protein contaminants using an acetone protein precipitation step (9 to 1 volumes of acetone, -20°C, overnight). After centrifugation (15 min, 15000g, 4°C), the protein pellet was dried and resuspended in 200µl of solubilization buffer (8M urea, 0.1M Ammonium bicarbonate, pH8). Protein concentration was determined using the RC-DC protein assay (Bio-Rad, Hercules, CA, USA). Proteins were reduced (12mM DTT, 37°C, 30 min) and alkylated (40mM IAA, 25°C, 1 hour, dark). Samples were diluted to reach a 1M urea concentration using a solution of freshly prepared 0.1M ammonium bicarbonate and proteins were digested using Trypsin (1:120 enzyme:substrate ratio, 37°C, overnight). Stable isotope-labeled synthetic peptides were spiked in the samples and after acidification using formic acid (pH 3), the samples were desalted and concentrated using solid phase extraction (Sep-Pak C18, 1cc, 50mg, Waters). The eluate volume was reduced in a vacuum centrifuge and adjusted using water + 0.1% formic acid to reach a final concentration of 4µg/µl.

## E. Personalized multi-omics profiling

### E.1. Sample preparation protocol details

Protein pellets were dissolved in 200µl of 10mM Tris buffer, and protease and phosphatase inhibitors were added. The samples were sonicated for 10s and proteins were quantified using the Bradford assay. The samples were evapored using a vacuum centrifuge and resuspended in laemmli buffer in order to obtain a 10µg/µl concentration, then loaded into a 10% 1D SDS-PAGE (10cmx7.3cmx1mm) and migrated on a mini PROTEAN (Bio-Rad) apparatus at 50V for 20 min and 100V for 1 hour. Two gels were produced, one were samples were 100 µg of total protein amount per sample was loaded and another one with 70µg. Gels were then stained with colloidal Coomassie Blue (BioSafe coomassie stain; Bio-Rad) and whole lanes were systematically cut into bands (5x2 mm) using a disposable grid-cutter. The gel bands were stored at -20°C until further analysis. The gel with 100-µg of deposited protein was used for the relative quantification. The other one was used for the second experiment of in-depth proteome characterization.In-gel digestion using trypsin (Promega, Madison, WI, USA) was performed overnight at 37°C after in-gel reduction and alkylation using a MassPrep Station (Waters, Milford, MA, USA). Tryptic peptides were extracted using 60% acetonitrile in 0.1% formic acid for 1h at room temperature. The volume was reduced in a vacuum centrifuge, resuspended with 0.1% formic acid in water and split into three aliquots to avoid biases due to sample evaporation. Samples were analyzed in triplicate by nanoLC-MS/MS (nanoliquid chromatography coupled to tandem mass spectrometry).

### E.2. nanoLC-MS parameters details for the relative quantification analysis by spectral count

NanoLC-MS/MS analyses were performed on a NanoAcquity LC-system (Waters, Milford, MA, USA) coupled to a maXis 4G QToF mass spectrometer (Bruker Daltonics, Bremen, Germany). For each analysis, a volume of 3µl of sample was injected into a Symmetry C18 precolumn (0.18x20mm, 5µm particle size, Waters) and then peptides were separated using an ACQUITY UPLC® BEH130 C18 separation column (75µm×200mm, 1.7µm particle size, Waters). Peptides were eluted using a linear gradient of water + 0.1% formic acid (solvent A) and acetonitrile +0.1% formic acid (solvent B). Peptide trapping was performed during 1 min at a flow rate of 15µL/min with 99% A and 1% B and elution was performed at 60 °C at a flow rate of 450 nL/min using a linear gradient from 6 to 43.5 % B over 35 minutes. For optimal nanoLC-MS/MS, the mass spectrometer was operated in positive mode, with the following settings: the

source temperature was set to 160°C, the dry gas flow rate was set to 5L/min and the nanoelectrospray voltage was optimized to −5000V. The TOF external mass calibration was achieved before each set of analyses using Tuning Mix (Agilent Technologies, Paolo Alto, USA ) in the mass range of 322-2722 m/z. Mass correction was achieved by recalibration of acquired spectra to the applied lock masses (methylstearate ([M+H]+ 299.2945 m/z) and hexakis(2,2,3,3,-tetrafluoropropoxy)phosphazine ([M+H]+ 922.0098 m/z)). For tandem MS experiments, the system was operated with automatic switching between MS and MS/MS modes in the range of 100-2500 m/z (MS acquisition time of 0.4s), MS/MS acquisition time between 0.05s (intensity > 250000) and 1.25s (intensity <5000). The 6 most abundant peptides (absolute intensity threshold of 1500) were selected from each MS spectrum for further isolation and CID fragmentation using nitrogen as collision gas. Ions were dynamically excluded after acquisition of one MS/MS spectrum and the exclusion was released after 0.2 minutes.

### *E.3. nanoLC-MS parameters details for the in-depth proteome characterization experiment*

NanoLC-MS/MS analyses were performed on a NanoAcquity LC-system (Waters, Milford, MA, USA) coupled to a TripleTOF 5600+ (AB Sciex,, USA). For each analysis a volume of 4µl of sample was injected into a Symmetry C18 precolumn (0.18x20mm, 5µm particle size, Waters) and then peptides were separated using an ACQUITY UPLC® BEH130 C18 separation column (75µm×200mm, 1.7µm particle size, Waters). Peptides were eluted using a linear gradient of water + 0.1% formic acid (solvent A) and acetonitrile +0.1% formic acid (solvent B). Peptide trapping was performed during 3 min at a flow rate of 5µL/min with 99% A and 1% B and elution was performed at 50 °C at a flow rate of 300 nL/min using the following gradient: from 3 %B to 20%B over 110 minutes, from 20% to 40% in 35 minutes, from 40%B to 90%B in 1minute, 90%B for 8 minutes, from 90%B to 3% in 1minute and 3% for 15minutes. For optimal nanoLC-MS/MS, the mass spectrometer was operated in positive mode, with the following settings: ionspray voltage floating (ISVF) 2300 V, curtain gas (CUR) 25, interface heater temperature (IHT) 75, ion source gas 1 (GS1) 0, declustering potential (DP) 100 V. Information-dependent acquisition (IDA) mode was used with Top 40 MS/MS scans. The MS scan had a dwell time of 250ms in the mass range of 400 – 1250 and the MS/MS scans 65 ms in the mass range of 200 – 1600 in high sensitivity mode giving a total cycle time of 2.90s. Switching criteria were set to ions greater than mass to charge ratio (m/z) 350 and smaller than m/z 1250 with charge state of 2–4 and an abundance threshold of more than 75 counts, exclusion time was set at 12 s. IDA rolling collision energy script was used for automatically adapting the CE.

# References


1. Michalski A, Cox J, Mann M: **More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS**. *J Proteome Res* 2011, **10**(4):1785-1793.

2. Picotti P, Aebersold R: **Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions**. *Nat Methods* 2012, **9**(6):555-566.

3. Peterson AC, Russell JD, Bailey DJ, Westphall MS, Coon JJ: **Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics**. *Mol Cell Proteomics* 2012, **11**(11):1475-1488.

4. Gallien S, Kim SY, Domon B: **Large-Scale Targeted Proteomics Using Internal Standard Triggered-Parallel Reaction Monitoring (IS-PRM)**. *Mol Cell Proteomics* 2015, **14**(6):1630-1644.

5. Picotti P, Bodenmiller B, Mueller LN, Domon B, Aebersold R: **Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics**. *Cell* 2009, **138**(4):795-806.

6. Gillet LC, Navarro P, Tate S, Rost H, Selevsek N, Reiter L, Bonner R, Aebersold R: **Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis**. *Mol Cell Proteomics* 2012, **11**(6):O111 016717.

7. Sajic T, Liu Y, Aebersold R: **Using data-independent, high-resolution mass spectrometry in protein biomarker research: perspectives and clinical applications**. *Proteomics Clin Appl* 2015, **9**(3-4):307-321.

8. Jaffe JD, Berg HC, Church GM: **Proteogenomic mapping as a complementary method to perform genome annotation**. *Proteomics* 2004, **4**(1):59-77.

9. Nesvizhskii AI: **Proteogenomics: concepts, applications and computational strategies**. *Nat Methods* 2014, **11**(11):1114-1125.

10. Armengaud J, Trapp J, Pible O, Geffard O, Chaumot A, Hartmann EM: **Non-model organisms, a species endangered by proteogenomics**. *J Proteomics* 2014, **105**:5-18.

11. Low TY, van Heesch S, van den Toorn H, Giansanti P, Cristobal A, Toonen P, Schafer S, Hubner N, van Breukelen B, Mohammed S *et al*: **Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis**. *Cell Rep* 2013, **5**(5):1469-1478.

12. Beri J, Rosenblatt MM, Strauss E, Urh M, Bereman MS: **Reagent for Evaluating Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) Performance in Bottom-Up Proteomic Experiments**. *Anal Chem* 2015, **87**(23):11635-11640.

13. Duriez E, Trevisiol S, Domon B: **Protein quantification using a cleavable reporter peptide**. *J Proteome Res* 2015, **14**(2):728-737.

14. Percy AJ, Chambers AG, Yang J, Domanski D, Borchers CH: **Comparison of standard- and nano-flow liquid chromatography platforms for MRM-based quantitation of putative plasma biomarker proteins**. *Anal Bioanal Chem* 2012, **404**(4):1089-1101.

15. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ: **Skyline: an open source document editor for creating and analyzing targeted proteomics experiments**. *Bioinformatics* 2010, **26**(7):966-968.

16. Reiter L, Rinner O, Picotti P, Huttenhain R, Beck M, Brusniak MY, Hengartner MO, Aebersold R: **mProphet: automated data processing and statistical validation for large-scale SRM experiments**. *Nat Methods* 2011, **8**(5):430-435.

17. Butto LF, Schaubeck M, Haller D: **Mechanisms of Microbe-Host Interaction in Crohn's Disease: Dysbiosis vs. Pathobiont Selection**. *Front Immunol* 2015, **6**:555.

18. Molodecky NA, Soon IS, Rabi DM, Ghali WA, Ferris M, Chernoff G, Benchimol EI, Panaccione R, Ghosh S, Barkema HW *et al*: **Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review**. *Gastroenterology* 2012, **142**(1):46-54 e42; quiz e30.

19. Baumgart DC, Sandborn WJ: **Crohn's disease**. *Lancet* 2012, **380**(9853):1590-1605.

20. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T *et al*: **An integrated catalog of reference genes in the human gut microbiome**. *Nat Biotechnol* 2014, **32**(8):834-841.

21. Juste C, Kreil DP, Beauvallet C, Guillot A, Vaca S, Carapito C, Mondot S, Sykacek P, Sokol H, Blon F *et al*: **Bacterial protein signals are associated with Crohn's disease**. *Gut* 2014, **63**(10):1566-1577.

22. Giglione C, Fieulaine S, Meinnel T: **N-terminal protein modifications: Bringing back into play the ribosome**. *Biochimie* 2015, **114**:134-146.

23. Frottin F, Espagne C, Traverso JA, Mauve C, Valot B, Lelarge-Trouverie C, Zivy M, Noctor G, Meinnel T, Giglione C: **Cotranslational proteolysis dominates glutathione homeostasis to support proper growth and development**. *Plant Cell* 2009, **21**(10):3296-3314.

24. Frottin F, Martinez A, Peynot P, Mitra S, Holz RC, Giglione C, Meinnel T: **The proteomics of N-terminal methionine cleavage**. *Mol Cell Proteomics* 2006, **5**(12):2336-2349.

25. Giglione C, Serero A, Pierre M, Boisson B, Meinnel T: **Identification of eukaryotic peptide deformylases reveals universality of N-terminal protein processing mechanisms**. *EMBO J* 2000, **19**(21):5916-5929.

26. Ross S, Giglione C, Pierre M, Espagne C, Meinnel T: **Functional and developmental impact of cytosolic protein N-terminal methionine excision in Arabidopsis**. *Plant Physiol* 2005, **137**(2):623-637.

27. Satchi-Fainaro R, Puder M, Davies JW, Tran HT, Sampson DA, Greene AK, Corfas G, Folkman J: **Targeting angiogenesis with a conjugate of HPMA copolymer and TNP-470**. *Nat Med* 2004, **10**(3):255-261.

28. Bertaccini D, Vaca S, Carapito C, Arsene-Ploetze F, Van Dorsselaer A, Schaeffer-Reiss C: **An improved stable isotope N-terminal labeling approach with light/heavy TMPP to automate proteogenomics data validation: dN-TOP**. *J Proteome Res* 2013, **12**(6):3063-3070.

29. Vaca Jacome AS, Rabilloud T, Schaeffer-Reiss C, Rompais M, Ayoub D, Lane L, Bairoch A, Van Dorsselaer A, Carapito C: **N-terminome analysis of the human mitochondrial proteome**. In. Saint-Louis, MO, USA: ASMS Congress 2015.

30. Steen H, Mann M: **The ABC's (and XYZ's) of peptide sequencing**. *Nat Rev Mol Cell Biol* 2004, **5**(9):699-711.

31. Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR, 3rd: **Protein analysis by shotgun/bottom-up proteomics**. *Chem Rev* 2013, **113**(4):2343-2394.

32. Smith LM, Kelleher NL: **Proteoform: a single term describing protein complexity**. *Nat Methods* 2013, **10**(3):186-187.

33. Zubarev RA: **The challenge of the proteome dynamic range and its implications for in-depth proteomics**. *Proteomics* 2013, **13**(5):723-726.

34. Richards AL, Hebert AS, Ulbrich A, Bailey DJ, Coughlin EE, Westphall MS, Coon JJ: **One-hour proteome analysis in yeast**. *Nat Protoc* 2015, **10**(5):701-714.

35. Nesvizhskii AI, Aebersold R: **Interpretation of shotgun proteomic data: the protein inference problem**. *Mol Cell Proteomics* 2005, **4**(10):1419-1440.

36. Durbin KR, Fornelli L, Fellers RT, Doubleday PF, Narita M, Kelleher NL: **Quantitation and Identification of Thousands of Human Proteoforms below 30 kDa**. *J Proteome Res* 2016, **15**(3):976-982.

37. Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, Tipton JD, Vellaichamy A, Kellie JF, Li M *et al*: **Mapping intact protein isoforms in discovery mode using top-down proteomics**. *Nature* 2011, **480**(7376):254-258.

38. Gregorich ZR, Ge Y: **Top-down proteomics in health and disease: challenges and opportunities**. *Proteomics* 2014, **14**(10):1195-1210.

39. Mesmin C, van Oostrum J, Domon B: **Complexity reduction of clinical samples for routine mass spectrometric analysis**. *Proteomics Clin Appl* 2016, **10**(4):315-322.

40. Laemmli UK: **Cleavage of structural proteins during the assembly of the head of bacteriophage T4**. *Nature* 1970, **227**(5259):680-685.

41. Xie F, Smith RD, Shen Y: **Advanced proteomic liquid chromatography**. *J Chromatogr A* 2012, **1261**:78-90.

42. Sleno L, Volmer DA: **Ion activation methods for tandem mass spectrometry**. *J Mass Spectrom* 2004, **39**(10):1091-1112.

43. Olsen JV, Macek B, Lange O, Makarov A, Horning S, Mann M: **Higher-energy C-trap dissociation for peptide modification analysis**. *Nat Methods* 2007, **4**(9):709-712.

44. Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF: **Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry**. *Proc Natl Acad Sci U S A* 2004, **101**(26):9528-9533.

45. Zubarev RA, Horn DM, Fridriksson EK, Kelleher NL, Kruger NA, Lewis MA, Carpenter BK, McLafferty FW: **Electron capture dissociation for structural characterization of multiply charged protein cations**. *Anal Chem* 2000, **72**(3):563-573.

46. Wysocki VH, Tsaprailis G, Smith LL, Breci LA: **Mobile and localized protons: a framework for understanding peptide dissociation**. *J Mass Spectrom* 2000, **35**(12):1399-1406.

47. Biemann K: **Appendix 5. Nomenclature for peptide fragment ions (positive ions)**. *Methods Enzymol* 1990, **193**:886-887.

48. Frese CK, Zhou H, Taus T, Altelaar AF, Mechtler K, Heck AJ, Mohammed S: **Unambiguous phosphosite localization using electron-transfer/higher-energy collision dissociation (EThcD)**. *J Proteome Res* 2013, **12**(3):1520-1525.

49. Rudomin EL, Carr SA, Jaffe JD: **Directed sample interrogation utilizing an accurate mass exclusion-based data-dependent acquisition strategy (AMEx)**. *J Proteome Res* 2009, **8**(6):3154-3160.

50. Jaffe JD, Keshishian H, Chang B, Addona TA, Gillette MA, Carr SA: **Accurate inclusion mass screening: a bridge from unbiased discovery to targeted assay development for biomarker verification**. *Mol Cell Proteomics* 2008, **7**(10):1952-1962.

51. Blueggel M, Chamrad D, Meyer HE: **Bioinformatics in proteomics**. *Curr Pharm Biotechnol* 2004, **5**(1):79-88.

52. Eng JK, McCormack AL, Yates JR: **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database**. *J Am Soc Mass Spectrom* 1994, **5**(11):976-989.

53. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data**. *Electrophoresis* 1999, **20**(18):3551-3567.

54. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH: **Open mass spectrometry search algorithm**. *J Proteome Res* 2004, **3**(5):958-964.

55. Craig R, Cortens JP, Beavis RC: **Open source system for analyzing, validating, and storing protein identification data**. *J Proteome Res* 2004, **3**(6):1234-1242.

56. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M: **Andromeda: a peptide search engine**

**integrated into the MaxQuant environment**. *J Proteome Res* 2011, **10**(4):1794-1805.

57. Balgley BM, Laudeman T, Yang L, Song T, Lee CS: **Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy**. *Mol Cell Proteomics* 2007, **6**(9):1599-1608.

58. Searle BC, Turner M, Nesvizhskii AI: **Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies**. *J Proteome Res* 2008, **7**(1):245-253.

59. Jones AR, Siepen JA, Hubbard SJ, Paton NW: **Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines**. *Proteomics* 2009, **9**(5):1220-1229.

60. Shteynberg D, Nesvizhskii AI, Moritz RL, Deutsch EW: **Combining results of multiple search engines in proteomics**. *Mol Cell Proteomics* 2013, **12**(9):2383-2393.

61. Walker JM: **The Proteomics Protocols Handbook**: Humana Press; 2005.

62. McEntyre J: **Linking up with Entrez**. *Trends Genet* 1998, **14**(1):39-40.

63. Mashima J, Kodama Y, Kosuge T, Fujisawa T, Katayama T, Nagasaki H, Okuda Y, Kaminuma E, Ogasawara O, Okubo K *et al*: **DNA data bank of Japan (DDBJ) progress report**. *Nucleic Acids Res* 2016, **44**(D1):D51-57.

64. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic Acids Res* 2015, **43**(Database issue):D30-35.

65. Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE *et al*: **The Protein Information Resource**. *Nucleic Acids Res* 2003, **31**(1):345-347.

66. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D *et al*: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation**. *Nucleic Acids Res* 2016, **44**(D1):D733-745.

67. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.

68. Armengaud J: **A perfect genome annotation is within reach with the proteomics and genomics alliance**. *Curr Opin Microbiol* 2009, **12**(3):292-300.

69. **UniProt: a hub for protein information**. *Nucleic Acids Res* 2015, **43**(Database issue):D204-212.

70. Gaudet P, Michel PA, Zahn-Zabal M, Cusin I, Duek PD, Evalet O, Gateau A, Gleizes A, Pereira M, Teixeira D *et al*: **The neXtProt knowledgebase on human proteins: current status**. *Nucleic Acids Res* 2015, **43**(Database issue):D764-770.

71. Elias JE, Gygi SP: **Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry**. *Nat Methods* 2007, **4**(3):207-214.

72. Navarro P, Vazquez J: **A refined method to calculate false discovery rates for peptide identification using decoy databases**. *J Proteome Res* 2009, **8**(4):1792-1796.

73. Bradshaw RA, Burlingame AL, Carr S, Aebersold R: **Reporting protein identification data: the next generation of guidelines**. *Mol Cell Proteomics* 2006, **5**(5):787-788.

74. Omenn GS, Lane L, Lundberg EK, Beavis RC, Nesvizhskii AI, Deutsch EW: **Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification**. *J Proteome Res* 2015, **14**(9):3452-3460.

75. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M: **Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics**. *Mol Cell Proteomics* 2002, **1**(5):376-386.

76. Bantscheff M, Lemeer S, Savitski MM, Kuster B: **Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present**. *Anal Bioanal Chem* 2012, **404**(4):939-965.

77. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R: **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags**. *Nat Biotechnol* 1999, **17**(10):994-999.

78. Zhang R, Sioma CS, Wang S, Regnier FE: **Fractionation of isotopically labeled peptides in quantitative proteomics**. *Anal Chem* 2001, **73**(21):5142-5149.

79. Latosinska A, Vougas K, Makridakis M, Klein J, Mullen W, Abbas M, Stravodimos K, Katafigiotis I, Merseburger AS, Zoidakis J *et al*: **Comparative Analysis of Label-Free and 8-Plex iTRAQ Approach for Quantitative Tissue Proteomic Analysis**. *PLoS One* 2015, **10**(9):e0137048.

80. Werner T, Sweetman G, Savitski MF, Mathieson T, Bantscheff M, Savitski MM: **Ion coalescence of neutron encoded TMT 10-plex reporter ions**. *Anal Chem* 2014, **86**(7):3594-3601.

81. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S *et al*: **Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents**. *Mol Cell Proteomics* 2004, **3**(12):1154-1169.

82. Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AK, Hamon C: **Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS**. *Anal Chem* 2003, **75**(8):1895-1904.

83. Bakalarski CE, Kirkpatrick DS: **A Biologist's Field Guide to Multiplexed Quantitative Proteomics**. *Mol Cell Proteomics* 2016.

84. Liu H, Sadygov RG, Yates JR, 3rd: **A model for random sampling and estimation of relative protein abundance in shotgun proteomics**. *Anal Chem* 2004, **76**(14):4193-4201.

85. Ramus C, Hovasse A, Marcellin M, Hesse AM, Mouton-Barbosa E, Bouyssie D, Vaca S, Carapito C, Chaoui K, Bruley C *et al*: **Benchmarking quantitative label-free LC-MS data processing workflows using a complex spiked proteomic standard dataset**. *J Proteomics* 2016, **132**:51-62.

86. Schilling B, Rardin MJ, MacLean BX, Zawadzka AM, Frewen BE, Cusack MP, Sorensen DJ, Bereman MS, Jing E, Wu CC *et al*: **Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation**. *Mol Cell Proteomics* 2012, **11**(5):202-214.

87. Cox J, Mann M: **MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification**. *Nat Biotechnol* 2008, **26**(12):1367-1372.

88. Bouyssie D, Gonzalez de Peredo A, Mouton E, Albigot R, Roussel L, Ortega N, Cayrol C, Burlet-Schiltz O, Girard JP, Monsarrat B: **Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: application to the proteomics study of membrane proteins from primary human endothelial cells**. *Mol Cell Proteomics* 2007, **6**(9):1621-1637.

89. Marx V: **Targeted proteomics**. *Nat Methods* 2013, **10**(1):19-22.

90. Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP: **Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS**. *Proc Natl Acad Sci U S A* 2003, **100**(12):6940-6945.

91. Proc JL, Kuzyk MA, Hardie DB, Yang J, Smith DS, Jackson AM, Parker CE, Borchers CH: **A quantitative study of the effects of chaotropic agents, surfactants, and solvents on the digestion efficiency of human plasma proteins by trypsin**. *J Proteome Res* 2010, **9**(10):5422-5437.

92. Fang P, Liu M, Xue Y, Yao J, Zhang Y, Shen H, Yang P: **Controlling nonspecific trypsin cleavages in LC-MS/MS-based shotgun proteomics using optimized experimental conditions**. *Analyst* 2015, **140**(22):7613-7621.

93. Pratt JM, Simpson DM, Doherty MK, Rivers J, Gaskell SJ, Beynon RJ: **Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes**. *Nat Protoc* 2006, **1**(2):1029-1043.

94. Mirzaei H, McBee JK, Watts J, Aebersold R: **Comparative evaluation of current peptide production platforms used in absolute quantification in proteomics**. *Mol Cell Proteomics* 2008, **7**(4):813-823.

95. Brun V, Dupuis A, Adrait A, Marcellin M, Thomas D, Court M, Vandenesch F, Garin J: **Isotope-labeled protein standards: toward absolute quantitative proteomics**. *Mol Cell Proteomics* 2007, **6**(12):2139-2149.

96. Domon B, Gallien S: **Recent advances in targeted proteomics for clinical applications**. *Proteomics Clin Appl* 2015, **9**(3-4):423-431.

97. Gallien S, Bourmaud A, Kim SY, Domon B: **Technical considerations for large-scale parallel reaction monitoring analysis**. *J Proteomics* 2014, **100**:147-159.

98. Bilbao A, Varesio E, Luban J, Strambio-De-Castillia C, Hopfgartner G, Muller M, Lisacek F: **Processing strategies and software solutions for data-independent acquisition in mass spectrometry**. *Proteomics* 2015, **15**(5-6):964-980.

99. Purvine S, Eppel JT, Yi EC, Goodlett DR: **Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer**. *Proteomics* 2003, **3**(6):847-850.

100. Silva JC, Denny R, Dorschel CA, Gorenstein M, Kass IJ, Li GZ, McKenna T, Nold MJ, Richardson K, Young P *et al*: **Quantitative proteomic analysis by accurate mass retention time pairs**. *Anal Chem* 2005, **77**(7):2187-2200.

101. Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR: **Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra**. *Nat Methods* 2004, **1**(1):39-45.

102. Panchaud A, Scherl A, Shaffer SA, von Haller PD, Kulasekara HD, Miller SI, Goodlett DR: **Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean**. *Anal Chem* 2009, **81**(15):6481-6488.

103. Weisbrod CR, Eng JK, Hoopmann MR, Baker T, Bruce JE: **Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification**. *J Proteome Res* 2012, **11**(3):1621-1632.

104. Selevsek N, Chang CY, Gillet LC, Navarro P, Bernhardt OM, Reiter L, Cheng LY, Vitek O, Aebersold R: **Reproducible and consistent quantification of the Saccharomyces cerevisiae proteome by SWATH-mass spectrometry**. *Mol Cell Proteomics* 2015, **14**(3):739-749.

105. Muntel J, Fromion V, Goelzer A, Maabeta S, Mader U, Buttner K, Hecker M, Becher D: **Comprehensive absolute quantification of the cytosolic proteome of Bacillus subtilis by data independent, parallel fragmentation in liquid chromatography/mass spectrometry (LC/MS(E))**. *Mol Cell Proteomics* 2014, **13**(4):1008-1019.

106. Khatun J, Yu Y, Wrobel JA, Risk BA, Gunawardena HP, Secrest A, Spitzer WJ, Xie L, Wang L, Chen X *et al*: **Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions**. *BMC Genomics* 2013, **14**:141.

107. Kornblihtt AR, Schor IE, Allo M, Dujardin G, Petrillo E, Munoz MJ: **Alternative splicing: a pivotal step between eukaryotic transcription and translation**. *Nat Rev Mol Cell Biol* 2013, **14**(3):153-165.

108. Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham AJ, Tabb DL: **TagRecon: high-throughput mutation identification through sequence tagging**. *J Proteome Res* 2010, **9**(4):1716-1726.

109. Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP, Hunter CL, Nuwaysir LM, Schaeffer DA: **The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra**. *Mol Cell Proteomics* 2007, **6**(9):1638-1655.

110. Seidler J, Zinn N, Boehm ME, Lehmann WD: **De novo sequencing of peptides by MS/MS**. *Proteomics* 2010, **10**(4):634-649.

111. Locard-Paulet M, Pible O, Gonzalez de Peredo A, Alpha-Bazin B, Almunia C, Burlet-Schiltz O, Armengaud J: **Clinical implications of recent advances in proteogenomics**. *Expert Rev Proteomics* 2016, **13**(2):185-199.

112. Low TY, Heck AJ: **Reconciling proteomics with next generation sequencing**. *Curr Opin Chem Biol* 2016, **30**:14-20.

113. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL, Saghatelian A: **Peptidomic discovery of short open reading frame-encoded peptides in human cells**. *Nat Chem Biol* 2013, **9**(1):59-64.

114. Brosch M, Saunders GI, Frankish A, Collins MO, Yu L, Wright J, Verstraten R, Adams DJ, Harrow J, Choudhary JS *et al*: **Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome**. *Genome Res* 2011, **21**(5):756-767.

115. Banfai B, Jia H, Khatun J, Wood E, Risk B, Gundling WE, Jr., Kundaje A, Gunawardena HP, Yu Y, Xie L *et al*: **Long noncoding RNAs are rarely translated in two human cell lines**. *Genome Res* 2012, **22**(9):1646-1657.

116. Choudhary JS, Blackstock WP, Creasy DM, Cottrell JS: **Interrogating the human genome using uninterpreted mass spectrometry data**. *Proteomics* 2001, **1**(5):651-667.

117. Bitton DA, Smith DL, Connolly Y, Scutt PJ, Miller CJ: **An integrated mass-spectrometry pipeline identifies novel protein coding-regions in the human genome**. *PLoS One* 2010, **5**(1):e8949.

118. Sevinsky JR, Cargile BJ, Bunger MK, Meng F, Yates NA, Hendrickson RC, Stephenson JL, Jr.: **Whole genome searching with shotgun proteomic data: applications for genome annotation**. *J Proteome Res* 2008, **7**(1):80-88.

119. Tanner S, Shen Z, Ng J, Florea L, Guigo R, Briggs SP, Bafna V: **Improving gene annotation using peptide mass spectrometry**. *Genome Res* 2007, **17**(2):231-239.

120. Mo F, Hong X, Gao F, Du L, Wang J, Omenn GS, Lin B: **A compatible exon-exon junction database for the identification of exon skipping events using tandem mass spectrum data**. *BMC Bioinformatics* 2008, **9**:537.

121. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S *et al*: **GENCODE: the reference human genome annotation for The ENCODE Project**. *Genome Res* 2012, **22**(9):1760-1774.

122. Zhao Y, Li H, Fang S, Kang Y, Wu W, Hao Y, Li Z, Bu D, Sun N, Zhang MQ *et al*: **NONCODE 2016: an informative and valuable data source of long non-coding RNAs**. *Nucleic Acids Res* 2016, **44**(D1):D203-208.

123. Karro JE, Yan Y, Zheng D, Zhang Z, Carriero N, Cayting P, Harrrison P, Gerstein M: **Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation**. *Nucleic Acids Res* 2007, **35**(Database issue):D55-60.

124. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S *et al*: **COSMIC: exploring the world's knowledge of somatic mutations in human cancer**. *Nucleic Acids Res* 2015, **43**(Database issue):D805-811.

125. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic Acids Res* 2001, **29**(1):308-311.

126. Li J, Duncan DT, Zhang B: **CanProVar: a human cancer proteome variation database**. *Hum Mutat* 2010, **31**(3):219-228.

127. Kim P, Kim N, Lee Y, Kim B, Shin Y, Lee S: **ECgene: genome annotation for alternative splicing**. *Nucleic Acids Res* 2005, **33**(Database issue):D75-79.

128. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L *et al*: **Ensembl 2016**. *Nucleic Acids Res* 2016, **44**(D1):D710-716.

129. Lenffer J, Nicholas FW, Castle K, Rao A, Gregory S, Poidinger M, Mailman MD, Ranganathan S: **OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI**. *Nucleic Acids Res* 2006, **34**(Database issue):D599-601.

130. Puente XS, Sanchez LM, Overall CM, Lopez-Otin C: **Human and mouse proteases: a comparative genomic approach**. *Nat Rev Genet* 2003, **4**(7):544-558.

131. Lopez-Otin C, Overall CM: **Protease degradomics: a new challenge for proteomics**. *Nat Rev Mol Cell Biol* 2002, **3**(7):509-519.

132. Marino G, Eckhard U, Overall CM: **Protein Termini and Their Modifications Revealed by Positional Proteomics**. *ACS Chem Biol* 2015, **10**(8):1754-1764.

133. Van Damme P, Arnesen T, Gevaert K: **Protein alpha-N-acetylation studied by N-terminomics**. *FEBS J* 2011, **278**(20):3822-3834.

134. Lai ZW, Petrera A, Schilling O: **Protein amino-terminal modifications and proteomic approaches for N-terminal profiling**. *Curr Opin Chem Biol* 2015, **24**:71-79.

135. Hartmann EM, Armengaud J: **N-terminomics and proteogenomics, getting off to a good start**. *Proteomics* 2014, **14**(23-24):2637-2646.

136. Rogers LD, Overall CM: **Proteolytic post-translational modification of proteins: proteomic tools and methodology**. *Mol Cell Proteomics* 2013, **12**(12):3532-3542.

137. Xu G, Jaffrey SR: **N-CLAP: global profiling of N-termini by chemoselective labeling of the alpha-amine of proteins**. *Cold Spring Harb Protoc* 2010, **2010**(11):pdb prot5528.

138. Mahrus S, Trinidad JC, Barkan DT, Sali A, Burlingame AL, Wells JA: **Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini**. *Cell* 2008, **134**(5):866-876.

139. Agard NJ, Wells JA: **Methods for the proteomic identification of protease substrates**. *Curr Opin Chem Biol* 2009, **13**(5-6):503-509.

140. Huesgen PF, Overall CM: **N- and C-terminal degradomics: new approaches to reveal biological roles for plant proteases from substrate identification**. *Physiol Plant* 2012, **145**(1):5-17.

141. Agard NJ, Mahrus S, Trinidad JC, Lynn A, Burlingame AL, Wells JA: **Global kinetic analysis of proteolysis via quantitative targeted proteomics**. *Proc Natl Acad Sci U S A* 2012, **109**(6):1913-1918.

142. Timmer JC, Enoksson M, Wildfang E, Zhu W, Igarashi Y, Denault JB, Ma Y, Dummitt B, Chang YH, Mast AE *et al*: **Profiling constitutive proteolytic events in vivo**. *Biochem J* 2007, **407**(1):41-48.

143. Kim JS, Dai Z, Aryal UK, Moore RJ, Camp DG, 2nd, Baker SE, Smith RD, Qian WJ: **Resin-assisted enrichment of N-terminal peptides for characterizing proteolytic processing**. *Anal Chem* 2013, **85**(14):6826-6832.

144. Bland C, Bellanger L, Armengaud J: **Magnetic immunoaffinity enrichment for selective capture and MS/MS analysis of N-terminal-TMPP-labeled peptides**. *J Proteome Res* 2014, **13**(2):668-680.

145. Chen SH, Chen CR, Li DT, Hsu JL: **Improved N(alpha)-acetylated peptide enrichment following dimethyl labeling and SCX**. *J Proteome Res* 2013, **12**(7):3277-3287.

146. Staes A, Impens F, Van Damme P, Ruttens B, Goethals M, Demol H, Timmerman E, Vandekerckhove J, Gevaert K: **Selecting protein N-terminal peptides by combined fractional diagonal chromatography**. *Nat Protoc* 2011, **6**(8):1130-1141.

147. Van Damme P, Maurer-Stroh S, Plasman K, Van Durme J, Colaert N, Timmerman E, De Bock PJ, Goethals M, Rousseau F, Schymkowitz J *et al*: **Analysis of protein processing by N-terminal proteomics reveals novel species-specific substrate determinants of granzyme B orthologs**. *Mol Cell Proteomics* 2009, **8**(2):258-272.

148. Bland C, Hartmann EM, Christie-Oleza JA, Fernandez B, Armengaud J: **N-Terminal-oriented proteogenomics of the marine bacterium roseobacter denitrificans Och114 using N-Succinimidyloxycarbonylmethyl)tris(2,4,6-trimethoxyphenyl)phosphonium bromide (TMPP) labeling and diagonal chromatography**. *Mol Cell Proteomics* 2014, **13**(5):1369-1381.

149. Venne AS, Vogtle FN, Meisinger C, Sickmann A, Zahedi RP: **Novel highly sensitive, specific, and straightforward strategy for comprehensive N-terminal proteomics reveals unknown substrates of the mitochondrial peptidase Icp55**. *J Proteome Res* 2013, **12**(9):3823-3830.

150. Kleifeld O, Doucet A, auf dem Keller U, Prudova A, Schilling O, Kainthan RK, Starr AE, Foster LJ, Kizhakkedathu JN, Overall CM: **Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products**. *Nat Biotechnol* 2010, **28**(3):281-288.

151. Prudova A, auf dem Keller U, Butler GS, Overall CM: **Multiplex N-terminome analysis of MMP-2 and MMP-9 substrate degradomes by iTRAQ-TAILS quantitative proteomics**. *Mol Cell Proteomics* 2010, **9**(5):894-911.

152. Shen PT, Hsu JL, Chen SH: **Dimethyl isotope-coded affinity selection for the analysis of free and blocked N-termini of proteins using LC-MS/MS**. *Anal Chem* 2007, **79**(24):9520-9530.

153. Mommen GP, van de Waterbeemd B, Meiring HD, Kersten G, Heck AJ, de Jong AP: **Unbiased selective isolation of protein N-terminal peptides from complex proteome samples using phospho tagging (PTAG) and TiO(2)-based depletion**. *Mol Cell Proteomics* 2012, **11**(9):832-842.

154. Rawlings ND, Barrett AJ, Finn R: **Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors**. *Nucleic Acids Res* 2016, **44**(D1):D343-350.

155. Colaert N, Maddelein D, Impens F, Van Damme P, Plasman K, Helsens K, Hulstaert N, Vandekerckhove J, Gevaert K, Martens L: **The Online Protein Processing Resource (TOPPR): a database and analysis platform for protein processing events**. *Nucleic Acids Res* 2013, **41**(Database issue):D333-337.

156. Crawford ED, Seaman JE, Agard N, Hsu GW, Julien O, Mahrus S, Nguyen H, Shimbo K, Yoshihara HA, Zhuang M *et al*: **The DegraBase: a database of proteolysis in healthy and apoptotic human cells**. *Mol Cell Proteomics* 2013, **12**(3):813-824.

157. Fortelny N, Yang S, Pavlidis P, Lange PF, Overall CM: **Proteome TopFIND 3.0 with TopFINDer and PathFINDer: database and analysis tools for the association of protein termini to pre- and post-translational events**. *Nucleic Acids Res* 2015, **43**(Database issue):D290-297.

158. Tsiatsiani L, Heck AJ: **Proteomics beyond trypsin**. *FEBS J* 2015, **282**(14):2612-2626.

159. Vaca Jacome AS, Rabilloud T, Schaeffer-Reiss C, Rompais M, Ayoub D, Lane L, Bairoch A, Van Dorsselaer A, Carapito C: **N-terminome analysis of the human mitochondrial proteome**. *Proteomics* 2015, **15**(14):2519-2524.

160. Picotti P, Clement-Ziza M, Lam H, Campbell DS, Schmidt A, Deutsch EW, Rost H, Sun Z, Rinner O, Reiter L *et al*: **A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis**. *Nature* 2013, **494**(7436):266-270.

161. Deutsch EW, Sun Z, Campbell D, Kusebauch U, Chu CS, Mendoza L, Shteynberg D, Omenn GS, Moritz RL: **State of the Human Proteome in 2014/2015 As Viewed through PeptideAtlas: Enhancing Accuracy and Coverage through the AtlasProphet**. *J Proteome Res* 2015, **14**(9):3461-3473.

162. Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, Dianes JA, Sun Z, Farrah T, Bandeira N *et al*: **ProteomeXchange provides globally coordinated proteomics data submission and dissemination**. *Nat Biotechnol* 2014, **32**(3):223-226.

163. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H *et al*: **Mass-spectrometry-based draft of the human proteome**. *Nature* 2014, **509**(7502):582-587.

164. Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T *et al*: **Computational prediction of proteotypic peptides for quantitative proteomics**. *Nat Biotechnol* 2007, **25**(1):125-131.

165. Fusaro VA, Mani DR, Mesirov JP, Carr SA: **Prediction of high-responding peptides for targeted protein assays by mass spectrometry**. *Nat Biotechnol* 2009, **27**(2):190-198.

166. Tang H, Arnold RJ, Alves P, Xun Z, Clemmer DE, Novotny MV, Reilly JP, Radivojac P: **A computational approach toward label-free protein quantification using predicted peptide detectability**. *Bioinformatics* 2006, **22**(14):e481-488.

167. Webb-Robertson BJ, Cannon WR, Oehmen CS, Shah AR, Gurumoorthi V, Lipton MS, Waters KM: **A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics**. *Bioinformatics* 2010, **26**(13):1677-1683.

168. Mohammed Y, Domanski D, Jackson AM, Smith DS, Deelder AM, Palmblad M, Borchers CH: **PeptidePicker: a scientific workflow with web interface for selecting appropriate peptides for targeted proteomics experiments**. *J Proteomics* 2014, **106**:151-161.

169. Spicer V, Grigoryan M, Gotfrid A, Standing KG, Krokhin OV: **Predicting retention time shifts associated with variation of the gradient slope in peptide RP-HPLC**. *Anal Chem* 2010, **82**(23):9678-9685.

170. de Graaf EL, Altelaar AF, van Breukelen B, Mohammed S, Heck AJ: **Improving SRM assay development: a global comparison between triple quadrupole, ion trap, and higher energy CID peptide fragmentation spectra**. *J Proteome Res* 2011, **10**(9):4334-4341.

171. Kuzyk MA, Smith D, Yang J, Cross TJ, Jackson AM, Hardie DB, Anderson NL, Borchers CH: **Multiple reaction monitoring-based, multiplexed, absolute quantitation of 45 proteins in human plasma**. *Mol Cell Proteomics* 2009, **8**(8):1860-1877.

172. Gallien S, Peterman S, Kiyonami R, Souady J, Duriez E, Schoen A, Domon B: **Highly multiplexed targeted proteomics using precise control of peptide retention time**. *Proteomics* 2012, **12**(8):1122-1133.

173. Escher C, Reiter L, MacLean B, Ossola R, Herzog F, Chilton J, MacCoss MJ, Rinner O: **Using iRT, a normalized retention time for more targeted measurement of peptides**. *Proteomics* 2012, **12**(8):1111-1121.

174. Lebert D, Louwagie M, Goetze S, Picard G, Ossola R, Duquesne C, Basler K, Ferro M, Rinner O, Aebersold R *et al*: **DIGESTIF: a universal quality standard for the control of bottom-up proteomics experiments**. *J Proteome Res* 2015, **14**(2):787-803.

175. Holman SW, McLean L, Eyers CE: **RePLiCal: A QconCAT Protein for Retention Time Standardization in Proteomics Studies**. *J Proteome Res* 2016, **15**(3):1090-1102.

176. Parker SJ, Rost H, Rosenberger G, Collins BC, Malmstrom L, Amodei D, Venkatraman V, Raedschelders K, Van Eyk JE, Aebersold R: **Identification of a Set of Conserved Eukaryotic Internal Retention Time Standards for Data-independent Acquisition Mass Spectrometry**. *Mol Cell Proteomics* 2015, **14**(10):2800-2813.

177. Maclean B, Tomazela DM, Abbatiello SE, Zhang S, Whiteaker JR, Paulovich AG, Carr SA, Maccoss MJ: **Effect of collision energy optimization on the measurement of peptides by selected reaction monitoring (SRM) mass spectrometry**. *Anal Chem* 2010, **82**(24):10116-10124.

178. Sherwood CA, Eastham A, Lee LW, Risler J, Mirzaei H, Falkner JA, Martin DB: **Rapid optimization of MRM-MS instrument parameters by subtle alteration of precursor and product m/z targets**. *J Proteome Res* 2009, **8**(7):3746-3751.

179. Wan KX, Vidavsky I, Gross ML: **Comparing similar spectra: from similarity index to spectral contrast angle**. *J Am Soc Mass Spectrom* 2002, **13**(1):85-88.

180. Abbatiello SE, Mani DR, Keshishian H, Carr SA: **Automated detection of inaccurate and imprecise transitions in peptide quantification by multiple reaction monitoring mass spectrometry**. *Clin Chem* 2010, **56**(2):291-305.

181. Nasso S, Goetze S, Martens L: **Ariadne's Thread: A Robust Software Solution Leading to Automated Absolute and Relative Quantification of SRM Data**. *J Proteome Res* 2015, **14**(9):3779-3792.

182. Lazar C, Gatto L, Ferro M, Bruley C, Burger T: **Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies**. *J Proteome Res* 2016, **15**(4):1116-1125.

183. Chang CY, Picotti P, Huttenhain R, Heinzelmann-Schwarz V, Jovanovic M, Aebersold R, Vitek O: **Protein significance analysis in selected reaction monitoring (SRM) measurements**. *Mol Cell Proteomics* 2012, **11**(4):M111 014662.

184. R Development Core Team. R: a language and environment for statistical computing RFfSCV, Austria: 2008. ISBN3-900051-07-0. URL http://www.R-project.org.

185. Keshishian H, Addona T, Burgess M, Kuhn E, Carr SA: **Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution**. *Mol Cell Proteomics* 2007, **6**(12):2212-2229.

186. Keshishian H, Addona T, Burgess M, Mani DR, Shi X, Kuhn E, Sabatine MS, Gerszten RE, Carr SA: **Quantification of cardiovascular biomarkers in patient plasma by targeted mass spectrometry and stable isotope dilution**. *Mol Cell Proteomics* 2009, **8**(10):2339-2349.

187. Gautier V, Mouton-Barbosa E, Bouyssie D, Delcourt N, Beau M, Girard JP, Cayrol C, Burlet-Schiltz O, Monsarrat B, Gonzalez de Peredo A: **Label-free quantification and shotgun analysis of complex proteomes by one-dimensional SDS-PAGE/NanoLC-MS: evaluation for the large scale analysis of inflammatory human endothelial cells**. *Mol Cell Proteomics* 2012, **11**(8):527-539.

188. Dix MM, Simon GM, Cravatt BF: **Global mapping of the topography and magnitude of proteolytic events in apoptosis**. *Cell* 2008, **134**(4):679-691.

189. Rosenfeld J, Capdevielle J, Guillemot JC, Ferrara P: **In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis**. *Anal Biochem* 1992, **203**(1):173-179.

190. Chawade A, Alexandersson E, Levander F: **Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets**. *J Proteome Res* 2014, **13**(6):3114-3120.

191. Schilling B, MacLean B, Held JM, Sahu AK, Rardin MJ, Sorensen DJ, Peters T, Wolfe AJ, Hunter CL, MacCoss MJ *et al*: **Multiplexed, Scheduled, High-Resolution Parallel Reaction Monitoring on a Full Scan QqTOF Instrument with Integrated Data-Dependent and Targeted Mass Spectrometric Workflows**. *Anal Chem* 2015, **87**(20):10222-10229.

192. Ronsein GE, Pamir N, von Haller PD, Kim DS, Oda MN, Jarvik GP, Vaisar T, Heinecke JW: **Parallel reaction monitoring (PRM) and selected reaction monitoring (SRM) exhibit comparable linearity, dynamic range and precision for targeted quantitative HDL proteomics**. *J Proteomics* 2015, **113**:388-399.

193. Schiffmann C, Hansen R, Baumann S, Kublik A, Nielsen PH, Adrian L, von Bergen M, Jehmlich N, Seifert J: **Comparison of targeted peptide quantification assays for reductive dehalogenases by selective reaction monitoring (SRM) and precursor reaction monitoring (PRM)**. *Anal Bioanal Chem* 2014, **406**(1):283-291.

194. Scheltema RA, Hauschild JP, Lange O, Hornburg D, Denisov E, Damoc E, Kuehn A, Makarov A, Mann M: **The Q Exactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer**. *Mol Cell Proteomics* 2014, **13**(12):3698-3708.

195. Bereman MS: **Tools for monitoring system suitability in LC MS/MS centric proteomic experiments**. *Proteomics* 2015, **15**(5-6):891-902.

196. Piehowski PD, Petyuk VA, Orton DJ, Xie F, Moore RJ, Ramirez-Restrepo M, Engel A, Lieberman AP, Albin RL, Camp DG *et al*: **Sources of technical variability in quantitative LC-MS proteomics: human brain tissue sample analysis**. *J Proteome Res* 2013, **12**(5):2128-2137.

197. Abbatiello SE, Mani DR, Schilling B, Maclean B, Zimmerman LJ, Feng X, Cusack MP, Sedransk N, Hall SC, Addona T *et al*: **Design, implementation and multisite evaluation of a system suitability protocol for the quantitative assessment of instrument performance in liquid chromatography-multiple reaction monitoring-MS (LC-MRM-MS)**. *Mol Cell Proteomics* 2013, **12**(9):2623-2639.

198. Paulovich AG, Billheimer D, Ham AJ, Vega-Montoto L, Rudnick PA, Tabb DL, Wang P, Blackman RK, Bunk DM, Cardasis HL *et al*: **Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance**. *Mol Cell Proteomics* 2010, **9**(2):242-254.

199. Searle BC: **Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies**. *Proteomics* 2010, **10**(6):1265-1269.

200. Gallien S, Bourmaud A, Domon B: **A simple protocol to routinely assess the uniformity of proteomics analyses**. *J Proteome Res* 2014, **13**(5):2688-2695.

201. Pichler P, Mazanek M, Dusberger F, Weilnbock L, Huber CG, Stingl C, Luider TM, Straube WL, Kocher T, Mechtler K: **SIMPATIQCO: a server-based software suite which facilitates monitoring the time course of LC-MS performance metrics on Orbitrap instruments**. *J Proteome Res* 2012, **11**(11):5540-5547.

202. Schubert OT, Gillet LC, Collins BC, Navarro P, Rosenberger G, Wolski WE, Lam H, Amodei D, Mallick P, MacLean B *et al*: **Building high-quality assay libraries for targeted analysis of SWATH MS data**. *Nat Protoc* 2015, **10**(3):426-441.

203. Rardin MJ, Schilling B, Cheng LY, MacLean BX, Sorensen DJ, Sahu AK, MacCoss MJ, Vitek O, Gibson BW: **MS1 Peptide Ion Intensity Chromatograms in MS2 (SWATH) Data Independent Acquisitions. Improving Post Acquisition Analysis of Proteomic Experiments**. *Mol Cell Proteomics* 2015, **14**(9):2405-2419.

204. Makarov A, Denisov E, Lange O: **Performance evaluation of a high-field Orbitrap mass analyzer**. *J Am Soc Mass Spectrom* 2009, **20**(8):1391-1396.

205. Gallien S, Duriez E, Demeure K, Domon B: **Selectivity of LC-MS/MS analysis: implication for proteomics experiments**. *J Proteomics* 2013, **81**:148-158.

206. Zhang Y, Bilbao A, Bruderer T, Luban J, Strambio-De-Castillia C, Lisacek F, Hopfgartner G, Varesio E: **The Use of Variable Q1 Isolation Windows Improves Selectivity in LC-SWATH-MS Acquisition**. *J Proteome Res* 2015, **14**(10):4359-4371.

207. Egertson JD, Kuehn A, Merrihew GE, Bateman NW, MacLean BX, Ting YS, Canterbury JD, Marsh DM, Kellmann M, Zabrouskov V *et al*: **Multiplexed MS/MS for improved data-independent acquisition**. *Nat Methods* 2013, **10**(8):744-746.

208. Egertson JD, MacLean B, Johnson R, Xuan Y, MacCoss MJ: **Multiplexed peptide analysis using data-independent acquisition and Skyline**. *Nat Protoc* 2015, **10**(6):887-903.

209. Amodei D, Egertson J, McLean B, Johnson R, Vitek O, MacCoss M, Mallick P: **An Instrument-Independent Demultiplexing Method for Computationally Improving the Specificity of Data-Independent Acquisition**. *ASMS 2013 Poster* 2013.

210. Rost HL, Rosenberger G, Navarro P, Gillet L, Miladinovic SM, Schubert OT, Wolski W, Collins BC, Malmstrom J, Malmstrom L *et al*: **OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data**. *Nat Biotechnol* 2014, **32**(3):219-223.

211. Zawadzka AM, Schilling B, Held JM, Sahu AK, Cusack MP, Drake PM, Fisher SJ, Gibson BW: **Variation and quantification among a target set of phosphopeptides in human plasma by multiple reaction monitoring and SWATH-MS2 data-independent acquisition**. *Electrophoresis* 2014, **35**(24):3487-3497.

212. Abelin JG, Patel J, Lu X, Feeney CM, Fagbami L, Creech AL, Hu R, Lam D, Davison D, Pino L *et al*: **Reduced-representation Phosphosignatures Measured by Quantitative Targeted MS Capture Cellular States and Enable Large-scale Comparison of Drug-induced Phenotypes**. *Mol Cell Proteomics* 2016, **15**(5):1622-1641.

213. Tsou CC, Avtonomov D, Larsen B, Tucholska M, Choi H, Gingras AC, Nesvizhskii AI: **DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics**. *Nat Methods* 2015, **12**(3):258-264, 257 p following 264.

214. Soste M, Hrabakova R, Wanka S, Melnik A, Boersema P, Maiolica A, Wernas T, Tognetti M, von Mering C, Picotti P: **A sentinel protein assay for simultaneously quantifying cellular processes**. *Nat Methods* 2014, **11**(10):1045-1048.

215. Creech AL, Taylor JE, Maier VK, Wu X, Feeney CM, Udeshi ND, Peach SE, Boehm JS, Lee JT, Carr SA *et al*: **Building the Connectivity Map of epigenetics: chromatin profiling by quantitative targeted mass spectrometry**. *Methods* 2015, **72**:57-64.

216. Sender R, Fuchs S, Milo R: **Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans**. *Cell* 2016, **164**(3):337-340.

217. Clemente JC, Ursell LK, Parfrey LW, Knight R: **The impact of the gut microbiota on human health: an integrative view**. *Cell* 2012, **148**(6):1258-1270.

218. Shreiner AB, Kao JY, Young VB: **The gut microbiome in health and in disease**. *Curr Opin Gastroenterol* 2015, **31**(1):69-75.

219. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T *et al*: **A human gut microbial gene catalogue established by metagenomic sequencing**. *Nature* 2010, **464**(7285):59-65.

220. **Structure, function and diversity of the healthy human microbiome**. *Nature* 2012, **486**(7402):207-214.

221. Choi M, Chang CY, Clough T, Broudy D, Killeen T, MacLean B, Vitek O: **MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments**. *Bioinformatics* 2014, **30**(17):2524-2526.

222. Rodriguez J, Gupta N, Smith RD, Pevzner PA: **Does trypsin cut before proline?** *J Proteome Res* 2008, **7**(1):300-305.

223. Gallien S, Perrodou E, Carapito C, Deshayes C, Reyrat JM, Van Dorsselaer A, Poch O, Schaeffer C, Lecompte O: **Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol**. *Genome Res* 2009, **19**(1):128-135.

224. Wagner DS, Salari A, Gage DA, Leykam J, Fetter J, Hollingsworth R, Watson JT: **Derivatization of peptides to enhance ionization efficiency and control fragmentation during analysis by fast atom bombardment tandem mass spectrometry**. *Biol Mass Spectrom* 1991, **20**(7):419-425.

225. Huang ZH, Shen T, Wu J, Gage DA, Watson JT: **Protein sequencing by matrix-assisted laser desorption ionization-postsource decay-mass spectrometry analysis of the N-Tris(2,4,6-trimethoxyphenyl)phosphine-acetylated tryptic digests**. *Anal Biochem* 1999, **268**(2):305-317.

226. Sadagopan N, Watson JT: **Investigation of the tris(trimethoxyphenyl)phosphonium acetyl charged derivatives of peptides by electrospray ionization mass spectrometry and tandem mass spectrometry**. *J Am Soc Mass Spectrom* 2000, **11**(2):107-119.

227. Cech NB, Enke CG: **Relating electrospray ionization response to nonpolar character of small peptides**. *Anal Chem* 2000, **72**(13):2717-2723.

228. Cech NB, Krone JR, Enke CG: **Electrospray ionization detection of inherently nonresponsive epoxides by peptide binding**. *Rapid Commun Mass Spectrom* 2001, **15**(13):1040-1044.

229. Zhou S, Cook KD: **A mechanistic study of electrospray mass spectrometry: charge gradients within electrospray droplets and their influence on ion response**. *J Am Soc Mass Spectrom* 2001, **12**(2):206-214.

230. Cheng C, Gross ML: **Applications and mechanisms of charge-remote fragmentation**. *Mass Spectrom Rev* 2000, **19**(6):398-420.

231. He Y, Parthasarathi R, Raghavachari K, Reilly JP: **Photodissociation of charge tagged peptides**. *J Am Soc Mass Spectrom* 2012, **23**(7):1182-1190.

232. Muller D, Medigue C, Koechler S, Barbe V, Barakat M, Talla E, Bonnefoy V, Krin E, Arsene-Ploetze F, Carapito C *et al*: **A tale of two oxidation states: bacterial colonization of arsenic-rich environments**. *PLoS Genet* 2007, **3**(4):e53.

233. Carapito C, Burel A, Guterl P, Walter A, Varrier F, Bertile F, Van Dorsselaer A: **MSDA, a proteomics software suite for in-depth Mass Spectrometry Data Analysis using grid computing**. *Proteomics* 2014, **14**(9):1014-1019.

234. Mossmann D, Meisinger C, Vogtle FN: **Processing of mitochondrial presequences**. *Biochim Biophys Acta* 2012, **1819**(9-10):1098-1106.

235. Voos W: **Chaperone-protease networks in mitochondrial protein homeostasis**. *Biochim Biophys Acta* 2013, **1833**(2):388-399.

236. Emanuelsson O, Brunak S, von Heijne G, Nielsen H: **Locating proteins in the cell using TargetP, SignalP and related tools**. *Nat Protoc* 2007, **2**(4):953-971.

237. Savojardo C, Martelli PL, Fariselli P, Casadio R: **TPpred2: improving the prediction of mitochondrial targeting peptide cleavage sites by exploiting sequence motifs**. *Bioinformatics* 2014, **30**(20):2973-2974.

238. Nolden M, Ehses S, Koppen M, Bernacchia A, Rugarli EI, Langer T: **The m-AAA protease defective in hereditary spastic paraplegia controls ribosome assembly in mitochondria**. *Cell* 2005, **123**(2):277-289.

239. Baile MG, Claypool SM: **The power of yeast to model diseases of the powerhouse of the cell**. *Front Biosci (Landmark Ed)* 2013, **18**:241-278.

240. Wetterstrand KA: **DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts.** . Accessed 03 March 2016.

241. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.

242. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS: **The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments**. *Nat Protoc* 2012, **7**(8):1534-1550.

243. Ingolia NT: **Ribosome profiling: new views of translation, from single codons to genome scale**. *Nat Rev Genet* 2014, **15**(3):205-213.

244. Menschaert G, Van Criekinge W, Notelaers T, Koch A, Crappe J, Gevaert K, Van Damme P: **Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events**. *Mol Cell Proteomics* 2013, **12**(7):1780-1790.

245. Hoffmann F, Lohse P, Stojanov S, Shin YS, Renner ED, Kery A, Zellerer S, Belohradsky BH: **Identification of a novel mevalonate kinase gene mutation in combination with the common MVK V377I substitution and the low-penetrance TNFRSF1A R92Q mutation**. *Eur J Hum Genet* 2005, **13**(4):510-512.

246. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**(4):656-664.

247. Castellana NE, Shen Z, He Y, Walley JW, Cassidy CJ, Briggs SP, Bafna V: **An automated proteogenomic method uses mass spectrometry to reveal novel genes in Zea mays**. *Mol Cell Proteomics* 2014, **13**(1):157-167.

248. Risk BA, Spitzer WJ, Giddings MC: **Peppy: proteogenomic search software**. *J Proteome Res* 2013, **12**(6):3019-3025.

249. Nagaraj SH, Waddell N, Madugundu AK, Wood S, Jones A, Mandyam RA, Nones K, Pearson JV, Grimmond SM: **PGTools: A Software Suite for Proteogenomic Data Analysis and Visualization**. *J Proteome Res* 2015, **14**(5):2255-2266.

250. Wang X, Zhang B: **customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search**. *Bioinformatics* 2013, **29**(24):3235-3237.

251. Woo S, Cha SW, Merrihew G, He Y, Castellana N, Guest C, MacCoss M, Bafna V: **Proteogenomic database construction driven from large scale RNA-seq data**. *J Proteome Res* 2014, **13**(1):21-28.

252. Crappe J, Ndah E, Koch A, Steyaert S, Gawron D, De Keulenaer S, De Meester E, De Meyer T, Van Criekinge W, Van Damme P *et al*: **PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration**. *Nucleic Acids Res* 2015, **43**(5):e29.

# Alvaro Sebastian VACA JACOME

**Progress towards a better proteome characterization by quantitative mass spectrometry method development and proteogenomics**


L'extrême complexité des échantillons biologiques, la variabilité technique et la dépendance de la protéomique envers les banques protéiques empêchent l'analyse complète d'un protéome.

Ce travail de thèse s'est focalisé sur le développement de méthodes pour la protéomique quantitative et la protéogénomique afin d'améliorer la caractérisation du protéome. Premièrement mon travail s'est centré sur le développement de méthodes quantitatives globales et ciblées. La mise en place de standards pour évaluer les performances de tous les niveaux de la stratégie analytique est aussi décrite. Ces méthodes ont été optimisées pour répondre à diverses questions biologiques.

Mon doctorat s'est focalisé aussi autour de la protéogénomique. Une méthode d'analyse N-terminomique à haut débit a été développée et appliquée à l'étude de la mitochondrie humaine. Enfin, ce manuscrit présente une approche multi-omique visant à améliorer l'analyse du protéome avec la création de banques de données personnalisées.



**Mots-clés:** Spectrométrie de masse, Analyse Protéomique Quantitative, Protéogénomique



The high intrinsic complexity of biological samples, the technical variability and the dependency of Bottom-up Proteomics to consensus protein sequence databases handicap the comprehensive analysis of an entire Proteome.

My doctoral work was focused on method development in quantitative Proteomics and Proteogenomics in order to achieve a better proteome characterization. First, I focused on the development of global and targeted quantitative methods. The introduction and development of standard samples to assess the performances at any level of the analytical workflow is also described. These methods were applied to answer different biological questions.

My PhD also focused on Proteogenomic method development. A high throughput N-terminomic analysis approach was developed and applied to the analysis of the human mitochondria. Finally, this manuscript presents a personalized multi-omics profiling strategy to improve the proteome analysis with the use of personalized databases.



**Keywords**: Mass Spectrometry, Quantitative Proteomics, Proteogenomics