# UNIVERSITÉ DE STRASBOURG

*ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES*

**UMR 7178**

# THÈSE présentée par :

**Gauthier HUSSON**

soutenue le : **10 novembre 2017**

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

Discipline / Spécialité : Chimie / Chimie Analytique

# DEVELOPMENT OF HOST CELL PROTEIN IMPURITIES QUANTIFICATION METHODS BY MASS SPECTROMETRY TO CONTROL THE QUALITY OF BIOPHARMACEUTICALS

**THÈSE dirigée par :**

| | |
|---|---|
| **Dr. VAN DORSSELAER Alain** | Directeur de recherche, CNRS, Université de Strasbourg, France |
| **Pr. BRACEWELL Daniel** | Professor, University College London, United Kingdom |

**RAPPORTEURS :**

| | |
|---|---|
| **Dr. FERRO Myriam** | Directeur de recherche, CEA, Grenoble, France |
| **Dr. O'HARA John** | Director, Characterisation, UCB, Slough, United Kingdom |

---

**AUTRES MEMBRES DU JURY :**

| | |
|---|---|
| **Dr. GERVAIS Annick** | Director PCMD I, ASB, UCB, Braine l'Alleud, Belgium |
| **Dr. CARAPITO Christine** | Chargée de recherche, CNRS, Université de Strasbourg, France |

A mes parents,

Mon frère,

Ma famille,

Mes amis,

*« Le doute est le plus sûr instrument pour la découverte de la vérité ; c'est un verre brut qu'on donne à polir au temps et au génie pour aider tous les yeux à voir enfin les objets tels qu'ils sont »*

Joseph Michel Antoine Servan

# Acknowledgments

Tout d'abord, je tiens à remercier Alain Van Dorsselaer et Sarah Cianférani pour m'avoir accueilli au Laboratoire de Spectrométrie de Masse Bio-Organique.

Je voudrais remercier mes encadrants: un énorme merci Christine Carapito, pour tout ce que tu m'as appris, pour ton soutien au quotidien, pour ces discussions toujours enrichissantes. Merci pour ton implication, pour ta gentillesse. Merci de m'avoir fait découvrir et aimer la spectrométrie de masse. Merci Alain Van Dorsselaer pour vos précieux conseils tout au long de ma thèse. Un tout grand merci à Aurélie Delangle et Annick Gervais pour votre confiance, votre encadrement, vos réponses à mes nombreuses questions, votre soutien. Merci de me faire encore confiance aujourd'hui. A big thank you to Dan Bracewell, for your supervision at University College London, and your precious advices. I learned a lot during these five months, including mAb production and purification, but also English!

Thank you to John O'Hara and Myriam Ferro for accepting to review and evaluate my work.

Je voudrais remercier ceux avec qui j'ai collaboré durant ma thèse pour différents projets: Bruno Maucourt, Sabrina Bibi-Triki, Françoise Bringel, Muriel Bonnet et Sylvain Debard.

Thank you to my colleagues from UCL: Hai Yuan, thanks for all the time we spent together in the lab, for the discussions, the brainstorming, your kindness. Sushobhan, or Shubho, I was so lucky to have you in the lab. You were my personal teacher, and I really enjoyed learning so many things from you. Damiano, thanks for having welcoming me in the team, learning me suspension cell culture and for all the after work sessions. Alex, mon soutien francophone parmi tous ces anglophones, merci pour ta bonne humeur, et ton soutien tout au long de mon séjour à l'UCL.

Thanks to my colleagues from UCB: Matthew Hinchliffe, thank you for your strong support and all the advices regarding the cell culture and mAb purification. Thank you John O'Hara for your support, and your implication in the project. Merci Anne-Sophie pour ton aide concernant la recherche bibliographique.

Merci à mes collègues du LSMBO, où j'ai passé la majeure partie de ma thèse, pour votre gentillesse et votre bonne humeur.

Merci aux anciens: Diego, Gilles, Marine, merci pour votre accueil et vos conseils. Johann et Guillaume, merci pour vos réponses à mes nombreuses questions, et pour votre humour fort belge ! Merci Benoît pour ton humour un peu fou (et un peu belge aussi), et pour m'avoir accompagné au RU jusqu'au bout.

Georg, merci pour ta gentillesse, ton humour, et tout ce que tu m'as appris lors de nos discussions scientifiques tardives ! Angela te passe le bonjour depuis l'espace ;) Merci Maurane pour avoir partagé une partie de ma thèse (je me souviendrai longtemps de ce développement de méthode SRM !).

Merci Hélène, Agnès, Véronique, Fabrice Varrier, Fabrice Bertile, Luc, Alfred, Laurence, Christine Schaffer, Martine et Stella pour votre bonne humeur. Merci Danièle pour ta gentillesse cachée derrière cette façade alsacienne ! François, merci pour ta bonne humeur et pour tes blagues, j'espère que tu as commandé un chargeur de téléphone pour Noël ☺ Merci aux informaticiens pour leur aide au quotidien, Alex (vraiment merci pour ton aide précieuse à chaque fois que j'en avais besoin), Alex 2, Aymen, Patrick.

Jean-Marc, merci pour tout ce que tu m'as appris, pour ton aide tout au long de ma thèse, pour ta bienveillance envers tous les p'tits jeunes ! Merci aussi pour ta bonne humeur. Le labo a de la chance de t'avoir ! Prends soin de Nelson ;)

Merci aux supramol : Anthony et Thomas, bon courage pour vos thèses propres maggle ! Oscar, j'espère qu'un jour tu trouveras un sac à dos aussi beau que le mien. Maxime B, j'ai vraiment aimé discuter avec toi de science comme d'autres choses (« mets-toi donc à ma place, Stephen ! »), ces macdo tardifs, ces parties de pétanque, … Bon courage pour la fin de ta thèse, tes problèmes de chaleurs et d'humidité (je parle bien sûr du tiroir du robot). Stéphane, prends soin de ma co-équipière, et Steeve, bonne continuation dans ton nouveau bureau !

Merci Ludivine, Paola, Justine, Pauline, Blandine, Leslie et Nicolas pour votre gentillesse et votre bonne humeur. Paola et Blandine, bon courage pour vos thèses ! Justine, bon courage pour le Triple TOF ☺ Leslie, je me souviens de ce retour de SFEAP plutôt rigolo ! Félicitations pour ta thèse que tu viens de soutenir. Nicolas, bon courage pour ta thèse (écoute bien tout ce que te dit Jojo !).

Merci Maxime E, « mon père », pour ta bonne humeur, et ton aide sur le traitement de données DIA… Ce n'était pas évident mais tu t'en es bien sorti, et je te souhaite le meilleur pour ta thèse à Maastricht.

Merci Sebastian, pour tout ce que tu m'as appris. Je ne connaissais rien à la spectrométrie de masse en arrivant au laboratoire, et j'ai toujours pu compter sur toi lorsque j'avais des questions ou des doutes. On s'est bien pris la tête sur ce traitement de données DIA, la normalisation des temps de rétention etc… Merci aussi pour ton humour aiguisé, c'était du haut niveau ! Je te souhaite le meilleur à Boston, et apparemment ça te plaît, que ça continue !

Merci Joanna, j'ai vraiment aimé travailler avec toi. Tu es intéressée, consciencieuse et rigoureuse, et je suis fier de voir que tu es arrivée aussi vite à ce niveau de maîtrise. Je suis sûr que la suite de ta thèse va très bien se passer. Merci également pour ton humour un peu décalé, et pour tous ces « désolé » qui nous ont valu de très bons gâteaux ☺ Merci pour ton aide et ton soutien notamment pour la fin de ma thèse, pour retraiter les données, terminer la mise en page du manuscrit... Bon courage pour la fin de ta thèse, et pour t'occuper de notre instrument préféré ! :3

Le bureau des gros… Merci infiniment pour ces 3 années et quelques. C'est comme si le destin avait rassemblé les plus gros morphales du labo dans le même bureau ! Merci pour la bonne humeur qui y régnait, toutes ces choses qu'on a mangées ensemble, tous ces moments qu'on a partagés, et tous ces chats ! Merci Nina, ma Chouquette, pour ton humour très décalé, pour ta gentillesse, les apéros, la colocation à San Antonio, etc. Je te souhaite le meilleur pour la suite, ainsi qu'à Martoune et ~~Kevin Jean Bernard Edouard~~ Raphaël. Merci Margaux, Margrosse, nos échanges de mots doux vont me manquer ! Merci pour ton soutien, et toutes les franches rigolades qu'on a eues. Nina et Margaux, j'ai presque cohabité avec vous pendant plus de 3 ans, rarement tôt le matin, mais des fois tard le soir, lorsque l'inspiration était à son paroxysme : c'est là qu'ont été créés le kidnappeur de tractopelle (et son jumeau encore plus maléfique le kidnappeur de kidnappeur de tractopelle), les blagues de photos de vacances, les canulars téléphoniques… Marianne et Aurélie, merci infiniment pour votre gentillesse, pour votre soutien… Vous avez toujours été là pour me faire rire, me nourrir (c'est vous les meilleures pâtissières ☺), m'héberger… Et aussi pour cette dernière soirée, vous étiez impliquées comme si c'était votre propre thèse, et on l'a finie ensemble! Merci pour tout, je n'y serais pas arrivé sans vous. Marianne, ma voisine de bureau, j'ai vraiment aimé travailler à côté de toi, te charrier… Je te souhaite vraiment le meilleur au Canada ainsi qu'à Fadi. Tu passeras le bonjour à Jacky si tu le croises, tabarnak ! ;) Aurélie, tu vas me manquer… Tu étais ma coéquipière de badminton, tu es toujours encore là pour aider (les fameux 100 tube gels entre autres), tu m'as hébergé avant la soutenance et j'en garde un très bon souvenir malgré le stress ! T'es un peu comme une crêpe au Nutella ;) Merci également à David, mon frère de bouclettes et d'humour au ras des pâquerettes ! Bon courage pour ta thèse !

Merci à mes amis : les f**** de Nancy (Diane, Aussama, Thierry et Magali) pour votre soutien et les weekends passés ensemble ; les mosellans (Thierry, Serguei, Mèche et Marc), merci de prouver qu'on peut encore garder contact même en vivant aussi éloignés les uns des autres, merci pour ces soirées poker mémorables ! ; Sandra, merci pour ta gentillesse et ton soutien, ainsi que tes superbes blagues ! ; merci à mes collègues aïkidokas pour les séances de défouloir.

Un grand merci à mes parents, à mon frère Thibault et à Ana Alice de m'avoir supporté et soutenu pendant ces trois années plutôt intenses, pour tous les tupperwares, les déménagements… Merci à mes mamies toujours présentes pour me soutenir et me gâter avec des gâteaux, tartes, beignets, chaussons aux pommes, … Merci à ma filleule préférée Louise et à Pierre pour les sessions de maths, de jeux, de piscine, de ping pong, de télé… Merci à Michel pour les chantiers qui me changeaient les idées le weekend. Merci à tonton Jean-Marc, mieux connu sous le nom de « tonton Hermann » en Alsace, pour tous ses mouths. Enfin, merci à toute ma famille qui a toujours été là pour me soutenir.

Et merci aux chocobons ☺

# Table of contents

List of figures ........................................ 13

List of tables ........................................ 15

List of abbreviations ........................................ 16

Résumé en français ........................................ 21

**Chapter I** Introduction bibliographique ........................................ 21

**I.** Etat de l'art de la protéomique ........................................ 21

**II.** Anticorps monoclonaux ........................................ 23

**Chapter II** Résultats ........................................ 25

**I.** Optimisation de l'analyse protéomique globale ........................................ 25

**II.** Evaluation de différents couplages instrumentaux pour l'analyse protéomique ciblée ...... 27

**III.** Optimisation d'un workflow « data independent acquisition » ........................................ 28

A. Analyse des données ........................................ 29

B. Acquisition des données ........................................ 31

C. Préparation d'échantillon ........................................ 31

D. Comparaison entre DDA et DIA ........................................ 32

E. Conclusion ........................................ 34

**IV.** Développement d'approches de spectrométrie de masse de pointe pour quantifier les protéines de la cellule hôte ........................................ 35

**Chapter III** Conclusion générale ........................................ 36

General introduction ........................................ 41

Part I Bibliographic introduction ........................................ 45

**Chapter I** Bottom-up proteomics ........................................ 47

**I.** Analytical workflow ........................................ 48

A. Sample preparation ........................................ 49

A.1. Total protein quantification ........................................ 49

A.2. Protein purification and separation ........................................ 50

A.3. Enzymatic digestion ........................................ 51

A.4. Peptide purification and fractionation ........................................ 51

B. Mass spectrometry analysis ........................................ 52

B.1. Tandem mass spectrometry ........................................ 53

B.2. Fragmentation modes ........................................ 53

**II.** Data dependent acquisition ........................................ 54

**III.** Protein identification 56
- A. Search engines 56
- B. Protein sequence databases 57
  - B.1. NCBI 57
  - B.2. UniProtKB 58
- C. Validation of protein identification 58

**IV.** Global quantitative proteomics 58
- A. Spectral counting 59
- B. MS1 filtering – extracted ion chromatogram 60
  - B.1. Extraction of all detected features 60
  - B.2. Targeted extraction of identified peptides 61

**V.** Targeted proteomics 61
- A. Selected reaction monitoring 61
  - A.1. Method development 62
    - A.1.1. Selection of the targets 63
    - A.1.2. Time scheduling 64
    - A.1.3. Collision energy optimisation 65
    - A.1.4. Isotope dilution 65
  - A.2. Quantification 67
  - A.3. Linearity range 67
- B. Parallel reaction monitoring 67

**VI.** Data independent acquisition 69
- A. Principle 69
- B. Data acquisition methods 70
- C. Data analysis 71
  - C.1. Peptide-centric analysis 72
  - C.2. Spectrum-centric analysis 73

**Chapter II** Monoclonal antibodies 75

**I.** Expression systems 75

**II.** Manufacturing process 76
- A. Upstream process 76
- B. Downstream process 77

**III.** Host cell protein monitoring 77
- A. Immuno-specific methods 78

A.1. ELISA ..... 79
A.2. Western blot ..... 81
B. Non-specific methods ..... 82
B.1. Gel electrophoresis ..... 83
B.2. Mass spectrometry ..... 84
C. Host cell protein monitoring methods comparison ..... 85
Part II Results ..... 87
**Chapter I** Optimisation of shotgun proteomics analysis ..... 91
**I.** Liquid chromatography ..... 92
**II.** Interface ..... 93
**III.** Mass spectrometry ..... 94
A. Accumulation time ..... 95
B. MS/MS spectra collection ..... 97
B.1. Intensity threshold ..... 97
B.2. Dynamic exclusion ..... 99
C. MS/MS spectra quality ..... 100
C.1. Q1 resolution ..... 101
C.2. Collision energy spread ..... 102
C.3. ToF resolution ..... 103
**IV.** Conclusion ..... 104
**Chapter II** Benchmarking of targeted proteomics configurations ..... 107
**I.** Workflow ..... 107
**II.** Results ..... 109
A. Sensitivity ..... 109
B. Accuracy ..... 110
C. Precision ..... 110
**III.** Conclusion ..... 111
**Chapter III** Optimisation of a data independent acquisition workflow ..... 113
**I.** Data analysis ..... 114
A. Targeted data extraction ..... 115
A.1. Workflow ..... 115
A.2. Retention time tolerance ..... 118
A.3. Extraction window ..... 120
A.4. Number of transitions ..... 121

A.5. FDR threshold ........................................................................................ 122

A.6. Software tool ........................................................................................ 123

A.7. Conclusion ........................................................................................ 125

A.8. Perspectives........................................................................................ 126

B. Spectral library ........................................................................................ 126

**II.** Data acquisition........................................................................................ 131

A. Number and size of the isolation windows ........................................................ 133

B. Variable isolation windows ........................................................................ 135

C. Conclusion ........................................................................................ 137

**III.** Sample preparation........................................................................................ 137

A. Workflow ........................................................................................ 137

B. Load on stacking gel ........................................................................................ 138

C. Injected amount ........................................................................................ 139

D. Conclusion ........................................................................................ 139

**IV.** Comparison between DDA and DIA ........................................................ 140

**V.** Conclusion ........................................................................................ 142

**Chapter IV** Development of cutting edge mass spectrometry approaches to monitor host cell protein impurities during bioprocess development ........................................................ 145

General conclusion ........................................................................................ 205

References........................................................................................ 209

# List of figures

Figure 1 : Comparaison de deux gradients de chromatographie liquide. ........................................... 26
Figure 2 : Workflow pour la comparaison de quatre configurations instrumentales pour l'analyse protéomique ciblée. ........................................................................................................................ 28
Figure 3 : Comparaison des performances entre une librairie homemade et la librairie publique SWATHAtlas humaine........................................................................................................................ 30
Figure 4 : Comparaison des performances d'identification en mode DDA et DIA. ............................. 33
Figure 5 : Comparaison des performances de quantification en mode DDA et DIA. .......................... 34
Figure 6 : Distribution of protein abundances in NIH3T3 mouse fibroblasts (adapted from [2]). ........... 41
Figure 7 : Overview of the three MS-based proteomics approaches (adapted from [73]). .................... 48
Figure 8 : Biemann nomenclature for peptide fragmentation. ........................................................ 54
Figure 9 : Annotated MS/MS spectrum allows the determination of the peptide's amino acid sequence. ........................................................................................................................................ 54
Figure 10 : Principle of data dependent acquisition. ........................................................................ 55
Figure 11 : Principle of peptide fragment fingerprinting (PFF). ........................................................ 56
Figure 12 : Principle of the MS1 filtering (adapted from [14]). ........................................................... 60
Figure 13 : Principle of selected reaction monitoring (SRM) (adapted from [12]). ............................... 62
Figure 14 : SRM method optimisation workflow. .............................................................................. 63
Figure 15 : Principle of scheduled-SRM. ............................................................................................ 64
Figure 16 : Collision energy optimisation for SRM method. ............................................................. 65
Figure 17 : Principle of isotope dilution. ........................................................................................... 66
Figure 18 : Principle of parallel reaction monitoring (adapted from[12]). ............................................ 68
Figure 19 : The growing interest in data independent acquisition. ................................................... 69
Figure 20 : Principle of data independent acquisition (adapted from [12]). ......................................... 70
Figure 21 : Peptide-centric and spectrum-centric analyses (adapted from [159]). ................................. 72
Figure 22 : Overview of a generic mAb manufacturing process using mammalian cells (adapted from [173]). ............................................................................................................................................ 76
Figure 23 : Principle of a sandwich ELISA assay (from www.mybiosource.com). ............................... 79
Figure 24 : Possible outcomes for HCP detection by sandwich ELISA (adapted from [56]). ..................... 80
Figure 25 : Principle of protein detection by Western blot (adapted from www.leinco.com). ............ 81
Figure 26 : Evaluation of the HCP coverage of commercially available anti-HCP antibodies (adapted from [45]). ....................................................................................................................................... 82
Figure 27 : Limitations of 2D-PAGE method for HCP detection. ....................................................... 83
Figure 28 : Overview of my PhD work. ............................................................................................. 90
Figure 29 : Optimisation of the LC gradient. ..................................................................................... 93
Figure 30 : Optimisation of the source gas and heating. .................................................................. 94
Figure 31 : Optimisation of the cycle time sharing. .......................................................................... 96
Figure 32 : Evaluation of dynamic accumulation. ............................................................................ 97
Figure 33 : Optimisation of the precursor ion intensity threshold using a Top 50 x 50 ms method. ... 98
Figure 34 : Optimisation of dynamic exclusion. ................................................................................ 99
Figure 35 : Optimisation of the m/z exclusion window. ................................................................. 100
Figure 36 : Isotopic envelope description. ....................................................................................... 101

Figure 37 : Optimisation of the Q1 resolution. .................................................................................. 102
Figure 38 : Optimisation of precursor ions fragmentation. ............................................................. 103
Figure 39 : Optimisation of the number of time bins to sum........................................................... 103
Figure 40 : Overview of the LC-MS/MS key parameters driving the proteome coverage on the microLC-Triple TOF 6600 coupling. ............................................................................................ 104
Figure 41 : Workflow for the benchmarking of four targeted proteomics configurations. ............... 108
Figure 42 : Evaluation of the sensitivity of the targeted MS configurations..................................... 109
Figure 43 : Evaluation of the accuracy of the targeted MS platforms. ............................................. 110
Figure 44 : Evaluation of the precision of the targeted MS platforms. ............................................. 111
Figure 45 : Overview of the DIA workflow optimisation strategy. .................................................... 114
Figure 46 : Workflow for the optimisation of targeted DIA data extraction..................................... 117
Figure 47 : Retention time alignment performed by Skyline and Peakview. .................................... 119
Figure 48 : Evaluation of Skyline and Peakview software tools for targeted DIA data extraction. .... 124
Figure 49 : Comparison of identification performances between a homemade spectral library and the publicly available human spectral library from SWATHAtlas. .......................................................... 128
Figure 50 : Comparison of identification performances between the homemade and the SWATHAtlas spectral libraries when extracting only common peptides. ............................................................. 129
Figure 51 : Target and decoy score distribution obtained using the mProphet peak scoring model in Skyline.............................................................................................................................................. 130
Figure 52 : Relationship between the key parameters of DIA data acquisition................................ 132
Figure 53 : Overlap between Q1 isolation windows. ........................................................................ 132
Figure 54 : Evaluation of different number and size of isolation windows....................................... 134
Figure 55 : Generation of a variable window SWATH method from DDA data. ............................... 135
Figure 56 : Evaluation of the use of variable SWATH windows. ....................................................... 136
Figure 57 : Effect of the stacking gel load on proteome coverage.................................................... 138
Figure 58 : Effect of the column load on proteome coverage. ......................................................... 139
Figure 59 : Comparison of the identification performances of DDA and DIA modes. ....................... 141
Figure 60 :  Comparison of the quantification performances of DDA and DIA modes. ..................... 142
Figure 61 : Timescale of the Top 3-ID-DIA method transfer to industry. ........................................... 207

# List of tables

Tableau 1 : Paramètres optimisés pour l’analyse en mode DDA sur le couplage microLC-Triple TOF 6600. .......... 27
Tableau 2 : Paramètres optimisés pour l’extraction ciblée des données DIA. .......... 30

Table 1 : Description of the LC systems used. .......... 52
Table 2 : Summary of bottom-up proteomics approaches. .......... 74
Table 3 : Summary of HCP monitoring methods. .......... 85
Table 4 : Optimised parameters for DDA method on the microLC-Triple TOF 6600 coupling. .......... 105
Table 5 : Retention time tolerance optimisation for targeted DIA data extraction. .......... 120
Table 6 : m/z extraction window optimisation for targeted DIA data extraction. .......... 121
Table 7 : Optimisation of the number of extracted transitions per peptide for targeted DIA data extraction. .......... 122
Table 8 : FDR threshold optimisation for targeted DIA data extraction. .......... 123

# List of abbreviations

2D-PAGE: two Dimensional-PolyAcrylamide Gel Electrophoresis
ACN: Acetonitrile
BCA: BiCinchoninic Acid
CCCF: Clarified Cell Culture Fluid
CCF: Cell Culture Fluid
CE: Collision Energy
CES: Collision Energy Spread
CHO: Chinese Hamster Ovary
CID: Collision Induced Dissociation
cps: counts per second
CV: Coefficient of Variation
Da: Dalton
DDA: Data Dependent Acquisition
DDBJ: DNA Data Bank of Japan
DIA: Data Independent Acquisition
DNA: DeoxyriboNucleic Acid
DSP: DownStream Process
ECD: Electron Capture Dissociation
ELISA: Enzyme-Linked ImmunoSorbent Assay
EMBL: European Molecular Biology Laboratory
ESI: ElectroSpray Ionization
ETD: Electron Transfer Dissociation
FASP: Filter Aided Sample Preparation
Fc: Fragment crystallizable
FDP: False Discovery Proportion
FN: False Negative
FP: False Positive
FWHM: Full Width at Half Maximum
g: gram
HCD: Higher Energy Collisional Dissociation
HCP: Host Cell Protein
HPLC: High Performance Liquid Chromatography

HR/AM: High Resolution/Accurate Mass
HRP: Horseradish Peroxidase
ICAT: Isotope Coded Affinity Tag
ID: Isotope Dilution
IEF: IsoElectric Focusing
Ig: Immunoglobulin
iTRAQ: isobaric Tags for Relative and Absolute Quantification
L: liter
LC: Liquid Chromatography
LLOQ: Lower Limit Of Quantification
m/z: Mass over charge ratio
mAb: monoclonal Antibody
MALDI: Matrix-Assisted Laser Desorption Ionisation
min: minute
m: meter
MRM: Multiple Reaction Monitoring
ms: milliseconds
MS: Mass Spectrometry
MS/MS or MS2: tandem Mass Spectrometry
MW: Molecular Weight
NCBI: National Center for Biotechnology Information
PAGE: Polyacrylamide Gel Electrophoresis
PDB: Protein DataBank
PFF: Peptide Fragment Fingerprinting
pI: Isoelectric Point
PIR: Protein Information Ressource
PPA: Post Protein A
ppm: parts per million
PRM: Parallel Reaction Monitoring
FDR: False Discovery Rate
PSM: Peptide Spectrum Match
PTM: Post Translational Modification
Q-ToF: Quadrupole-Time of Flight
RefSeq: Reference Sequence
RF: Radio Frequency

RT: Retention Time

s: second

SDS: Sodium Dodecyl Sulfate

SILAC: Stable Isotope Labelling with Amino Acids in Cell culture

SPE: Solid Phase Extraction

SRM: Selected Reaction Monitoring

SWATH-MS: Sequential Windowed Acquisition of all THeoretical fragment ion Mass Spectra

TMB: 3,3’,5,5’-Tetramethylbenzidine

TMT: Tandem Mass Tags

TP: True Positive

TPR: True Positive Rate

UniProtKB: UniProt Knowledgebase

USD: Ultra Scale Down

USP: Upstream Process

XIC: eXtracted Ion Chromatogram

# Résumé en français

## **Chapter I** Introduction bibliographique

### I. Etat de l'art de la protéomique

Les protéines sont des biomolécules composées d'acides aminées, qui ont un rôle majeur dans une large gamme de procédés biologiques, comme le maintien de la structure des cellules, la signalisation cellulaire ou la réalisation de réactions biochimiques. Le nombre de protéine par cellule est estimé à 10 milliards pour une cellule de mammifère[1], comprenant 10 000 protéines différentes et chaque protéine étant présente de 1 copie à 10 millions de copies[2]. De plus, le contenu en protéines d'une cellule évolue en fonction du temps, de stimuli externes ou du type de cellule. L'ensemble des protéines présentes dans une cellule à un instant donné est appelé le protéome, et l'étude du protéome est appelée la protéomique[3-4].

L'évolution de la protéomique a été guidée par les avancées majeures en techniques de séparation, spectrométrie de masse et bioinformatique[5]. Les avancées de ces 20 dernières années en terme de sensibilité, résolution, précision et rapidité des instruments, ainsi que le séquençage des génomes et la création de banques de données protéiques ont permis à la spectrométrie de masse de s'imposer aujourd'hui comme la technique majeure pour l'analyse protéomique[6]. La protéomique joue aujourd'hui un rôle majeur dans différents secteurs de la recherche, comme la compréhension des procédés biologiques ou la recherche de biomarqueurs[7-8].

Aujourd'hui, ce sont les approches dites « bottom-up » qui sont les plus utilisées en protéomique[4]. Elles sont basées sur une digestion des protéines en morceaux de protéines appelés peptides, et leur petite taille facilite grandement leur analyse par spectrométrie de masse. Après séparation par chromatographie liquide, les peptides sont analysés par spectrométrie de masse, et par déduction les protéines sont ainsi identifiées et quantifiées.

Le mode d'acquisition le plus utilisé en protéomique est le mode « data dependent acquisition » (DDA), qui permet d'identifier et quantifier des milliers de protéines en seulement une heure[9]. Ce mode d'acquisition est basé sur la mesure des masses des peptides, puis les peptides les plus intenses sont fragmentés et les masses de ces fragments sont déterminées. La fragmentation des peptides se produit au niveau des liaisons peptides, c'est-à-dire entre les acides aminés qui les composent, et les masses des fragments détectés permettent de déduire la séquence en acides aminés des peptides[10]. Généralement, des moteurs de recherche sont utilisés pour comparer les masses mesurées des peptides et des fragments à des banques de données contenant les séquences de toutes les protéines d'un organisme, ainsi que les masses théoriques des peptides et fragments correspondants[11].

Cependant, l'identification seule des protéines ne suffit pas souvent pour fournir une réponse biologique claire, et le développement des techniques de quantifications des protéines était nécessaire[12]. Aujourd'hui, ce sont les approches sans marquage qui sont les plus utilisées, dû à leur polyvalence ainsi qu'à leur faible coût et rapidité comparé aux techniques de marquage[13]. La quantification sans marquage permet de quantifier de façon relative des peptides entre plusieurs échantillons. Le plus souvent, les peptides sont quantifiés par extraction des courants d'ions des peptides[14]. Cependant, étant donné que seuls les peptides les plus intenses sont fragmentés, le mode DDA souffre de limitations en terme de sensibilité (les peptides moins abondants ne sont pas fragmentés), de reproductibilité (les peptides les plus intenses ne sont pas toujours les mêmes au même moment de l'analyse) et de gamme dynamique[15].

Lorsqu'un nombre limité de protéines doit être quantifié dans un grand nombre d'échantillons, ce qui est le cas lorsque des candidats biomarqueurs identifiés en DDA doivent être validés, des approches ciblées peuvent être employées[16-23]. Elles permettent de quantifier ≈ 50-100 protéines connues dans des matrices complexes, avec une meilleure sensibilité, spécificité et reproductibilité. En effet, lors d'une analyse de protéomique ciblée, un groupe de peptides ainsi que leurs fragments vont être analysés de manière ciblée, même s'ils ne sont pas les plus intenses. La méthode ciblée la plus utilisée est appelée « selected reaction monitoring » (SRM), réalisée sur un instrument de type triple quadripôle. Le développement et l'optimisation d'une méthode SRM demandent un investissement conséquent en temps et en matériel, pour le développement d'une méthode de type « scheduled » qui permet d'augmenter drastiquement le nombre de peptides analysés, ou l'optimisation des énergies de collision afin d'obtenir une sensibilité optimale. De plus, la méthode SRM est souvent couplée à l'utilisation de peptides standards marqués aux isotopes stables correspondant aux peptides analysés, et qui sont ajoutés dans chaque échantillon en quantité connue. De ce fait, en faisant le rapport entre les signaux des peptides endogènes et des peptides standards marqués, une quantification absolue des peptides endogènes est possible. Récemment, des méthodes ciblées ont été développées sur des instruments de dernière génération dits HR/AM pour « high resolution / accurate mass », comme la « parallel reaction monitoring » (PRM)[24] réalisée sur un instrument de type quadripôle-orbitrap, ou la « multiple reaction monitoring in high resolution » (MRM HR)[25-26] réalisée sur des instruments de type quadripôle-tube-de-vol. L'utilisation de ces instruments HR/AM permet d'augmenter la spécificité de la quantification. De plus, ces instruments permettent également de réaliser des approches globales, et donc la découverte et la validation de biomarqueurs peuvent être réalisées sur le même couplage instrumental, ce qui facilite grandement le transfert de méthodes d'analyse.

Récemment, un nouveau mode d'acquisition a vu le jour, appelé « data independent acquisition » (DIA). Le mode DIA promet de combiner les avantages des approches DDA et des approches ciblées, en permettant (i) une couverture du protéome comparable et même supérieure à celle de l'approche DDA, et (ii) des sensibilité, spécificité et robustesse comparables à celles des approches ciblées. Ce mode est basé sur la fragmentation et l'obtention de données MS/MS de tous les peptides de la gamme de m/z analysée. Diverses techniques existent, et sont soit (i) basées sur l'analyse de tous les fragments en un seul balayage, soit (ii) les fragments sont analysés par fenêtres de m/z. Ces différentes méthodes DIA, en co-isolant un grand nombre de peptides, génèrent des spectres MS/MS très complexes, contenant les fragments de tous les peptides co-isolés, et le traitement des données en devient très difficile et c'est aujourd'hui le goulot d'étranglement des approches DIA. En effet, il n'est pas possible d'identifier les peptides comme dans une analyse classique DDA car chaque spectre MS/MS contient les fragments de plusieurs peptides, et la masse précise des peptides parents n'est pas connue non plus. Il existe deux façons de traiter ces données : (i) l'approche centrée sur les peptides, c'est-à-dire que l'on va extraire les données DIA de manière ciblée en recherchant des peptides d'intérêt à l'aide d'une librairie spectrale[27], et (ii) l'approche centrée sur les spectres, c'est-à-dire que des pseudo-spectres DDA sont générés en regroupant les peptides et leurs fragments qui co-éluent, et ensuite une recherche classique peut être effectuée[28]. L'approche centrée sur les peptides et utilisant une librairie spectrale est aujourd'hui la plus utilisée car elle donne les meilleurs résultats, et l'approche centrée sur les spectres identifie pour le moment un trop grand nombre de faux positifs[29]. Malgré ces défis pour le traitement de données, les approches DIA sont très prometteuses et suscitent beaucoup d'intérêt de la part de la communauté scientifique.

## II. Anticorps monoclonaux

Les anticorps monoclonaux (mAbs) sont des molécules intéressantes pour le traitement de maladies, car ils sont hautement spécifiques et peu toxiques comparés aux traitements classiques. Depuis la commercialisation du premier mAb, la classe des anticorps et molécules dérivées a rapidement évolué et est devenu aujourd'hui la classe dominante au sein du marché biopharmaceutique[30]. Aujourd'hui, plus de 70 mAbs et produits dérivés ont été approuvés par la Food and Drug Administration (FDA) et la European Medicines Agency (EMA), et plus de 50 mAbs sont en cours d'évaluation dans des études cliniques[31]. Ils sont utilisés pour traiter une large gamme de maladies, et principalement pour les maladies auto-immunes et les cancers[32]. Leur mode d'action varie depuis les fonctions naturelles des anticorps à l'adressage de médicaments[33]. Les ventes globales pour tous les mAbs thérapeutiques représentaient 107 milliards de dollars en 2016, et sont estimées à 145 milliards de dollars en 2020[34].

Les mAbs sont produits de manière recombinante dans des systèmes d'expressions, c'est-à-dire qu'après insertion du gène codant pour l'anticorps dans une cellule hôte, celle-ci le produit en continu. Les cellules d'ovaire de hamster chinois (CHO) sont les plus utilisées aujourd'hui pour la production de mAbs[35-36]. Le mAb est sécrété par ces cellules dans le milieu de culture, qui est récupéré et purifié en utilisant différentes techniques de chromatographie et de filtration afin d'éliminer les impuretés comme les acides nucléiques, les lipides ou les protéines de la cellule hôte (HCP). Ces impuretés doivent être quantifiées et leur taux communiqué aux autorités régulatrices[37].

Les HCP présentes dans la forme finale du biomédicament peuvent réduire l'efficacité du mAb, en particulier si ce sont des protéases qui peuvent dégrader le mAb[38-40], ou alors déclencher des effets secondaires chez les patients comme des réactions immunitaires[41-42]. La détection des HCP est particulièrement difficile, car (i) elles sont très peu présentes à côté du mAb, (ii) un grand nombre d'HCP doit être quantifié et (iii) la population d'HCP peut changer pendant le développement du procédé de production[43]. Typiquement, le taux d'HCP doit être réduit à < 100 ppm, c'est-à-dire < 100 ng HCP / mg mAb. Il existe plusieurs méthodes pour quantifier les HCP, qui peuvent être divisées en deux catégories : les méthodes immuno-spécifiques et les méthodes non-spécifiques.

Les méthodes immuno-spécifiques sont basées sur l'utilisation d'anticorps dirigés contre les HCP pour les détecter. Aujourd'hui, la méthode la plus utilisée pour quantifier les HCP est une méthode immuno-spécifique : l'ELISA. Cette méthode permet de quantifier les HCP à haut-débit, haute sensibilité et haute spécificité[43-45]. Cependant, le principal défaut de cette méthode est lié à l'utilisation d'anticorps dirigés contre les HCP. En effet, la population d'HCP détectable par les méthodes immuno-spécifiques est limitée aux HCP ciblées par les anticorps anti-HCP, et il est impossible de créer des anticorps anti-HCP reconnaissant toutes les HCP qui pourraient potentiellement apparaître dans les échantillons.

Le développement de méthodes non-spécifiques est donc nécessaire afin de détecter les HCP non reconnues par les anticorps anti-HCP. Parmi ces méthodes, la spectrométrie de masse est la plus prometteuse, car elle permet d'identifier et de quantifier les HCP individuellement en une seule analyse. La quantification individuelle des HCP est une information capitale pour comprendre comment améliorer le procédé de purification, ou encore prédire la dangerosité de cette HCP pour les patients. De plus, les avancées récentes en spectrométrie de masse, et notamment le développement des approches ciblées et des approches DIA, ont permis un gain en sensibilité d'un facteur 2 à 8[27], ainsi qu'un gain en spécificité et en gamme dynamique, ce qui est crucial pour la détection des HCP qui sont présentes à l'état de traces. Les limitations de ces techniques de spectrométrie de masse sont (i) le manque de banque de séquence protéique de CHO de bonne qualité[46], (ii) le besoin d'un personnel hautement qualifié et (iii) l'accès à un matériel très onéreux.

# Chapter II Résultats

Dans ce contexte, l'objectif de mon travail de thèse a été divisé en deux volets : (i) améliorer la caractérisation des protéomes par spectrométrie de masse en optimisant des méthodes d'analyse et de traitements de données, et en particulier du workflow complet pour l'analyse en mode DIA, et (ii) la production d'une large gamme d'échantillons de mAb, et l'analyse des HCP contenues dans ces échantillons par des techniques de pointe de spectrométrie de masse, et notamment l'approche DIA.

## I. Optimisation de l'analyse protéomique globale

Aujourd'hui, le mode d'acquisition DDA est le plus utilisé pour identifier et quantifier des protéines. Au laboratoire, nous avons acquis un couplage de dernière génération microLC-Triple TOF 6600, c'est-à-dire une chromatographie liquide (LC) opérée en mode micro (1-10 µL/min) couplée à spectromètre de masse de type quadripôle-tube de vol. Au cours de ma thèse, j'étais responsable de ce couplage, et j'ai optimisé une méthode d'analyse DDA pour ce couplage, afin de fournir au laboratoire une méthode optimale permettant l'identification et la quantification d'un maximum de peptides et de protéines.

Tout d'abord, la séparation chromatographique des peptides a été optimisée, et elle s'est avérée être l'un des facteurs majeurs influençant le nombre de peptides identifiés, de l'ordre de 30%. En effet, la séparation chromatographique permet d'injecter dans le spectromètre de masse les peptides de façon graduelle, et doit fournir le mélange de peptides le plus simple possible tout au long de l'analyse (**Figure 1**).

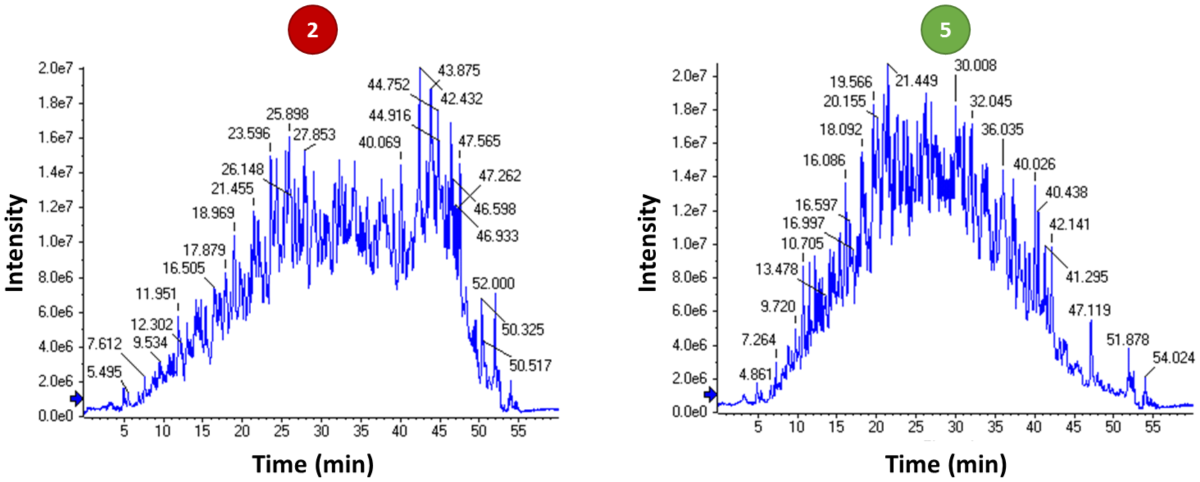


***Figure 1 : Comparaison de deux gradients de chromatographie liquide.***
*Le gradient 2 a permis de mieux étaler les peptides tout au long de l'analyse (maximum d'intensité de 25 à 45 min) alors qu'en utilisant le gradient 5, la majorité des peptides élue entre 20 et 30 min.*

Après la séparation chromatographique, les peptides doivent être ionisés et transférés en phase gazeuse pour pouvoir être analysés par spectrométrie de masse : c'est le rôle de l'interface. La position de l'aiguille à la sortie de la chromatographie liquide a été optimisée afin de permettre à un maximum d'ions de pénétrer dans le spectromètre de masse. L'interface utilisée permet également d'utiliser des gaz ainsi qu'un chauffage pour aider à la désolvatation des peptides et leur transfert en phase gazeuse. Les paramètres de la source ont été optimisés, permettant de gagner environ 10% de peptides identifiés en plus.

Enfin, l'acquisition des données par le spectromètre de masse en mode DDA a été optimisée. Tout d'abord, l'utilisation de l'accumulation dynamique, permettant au spectromètre de masse de gérer lui-même le temps qu'il passe pour analyser un peptide en fonction de son intensité, a permis d'augmenter le nombre d'identifications de ≈ 10% comparé à une méthode classique. De plus, l'optimisation de la collecte des spectres MS/MS en utilisant une exclusion dynamique, a permis également d'augmenter le nombre d'identification de ≈ 10%. La qualité des spectres acquis a également été optimisée, grâce principalement au réglage de la résolution du quadripôle qui a permis de gagner ≈ 10% d'identifications.

En conclusion, ces optimisations m'ont permis de mieux comprendre le fonctionnement de ce couplage instrumental, et de dégager des paramètres optimaux pour l'analyse des échantillons en mode DDA (**Tableau 1**).

*Tableau 1 : Paramètres optimisés pour l'analyse en mode DDA sur le couplage microLC-Triple TOF 6600.*

| | Paramètre | Optimum |
|---|---|---|
| Chromatographie liquide | Gradient | Gradient 2 |
| Interface | Gaz coaxial | 18 psi |
| | Gaz de chauffage | 20 psi |
| | Chauffage | 100 °C |
| Spectrométrie de masse | Temps d'accumulation | Accumulation dynamique |
| | Seuil d'intensité | 10 cps |
| | Exclusion dynamique | ½ pic chromatographique |
| | Résolution du Q1 | 1 Da |
| | Dispersion de l'énergie de collision | 0 V |
| | Résolution du tube de vol | 8 bins |

## II. Evaluation de différents couplages instrumentaux pour l'analyse protéomique ciblée

Aujourd'hui, la méthode SRM couplée à la dilution d'isotopes stables et effectuée sur un instrument de type triple quadripôle est la méthode de référence pour les approches ciblées. Elle permet de quantifier de façon absolue des protéines d'intérêt dans des échantillons complexes de façon robuste et avec une grande sensibilité[47-48]. La sensibilité et la robustesse sont les paramètres les plus importants pour les approches ciblées, car des protéines souvent faiblement abondantes doivent être analysées dans des centaines d'échantillons de façon reproductible. Récemment, des approches ciblées ont été développées sur des instruments de dernière génération de type HR/AM, comme la PRM[24] ou la MRM HR[25-26], permettant d'accroître la spécificité de la quantification et faciliter le développement et le transfert de méthode.

Au laboratoire, nous disposons de différentes configurations instrumentales permettant d'effectuer des approches de protéomique ciblée. Nous avons comparé quatre de ces couplages en utilisant un échantillon modèle dans lequel nous avons quantifié de manière précise 39 peptides, grâce à des peptides marqués aux isotopes stables (**Figure 2**).

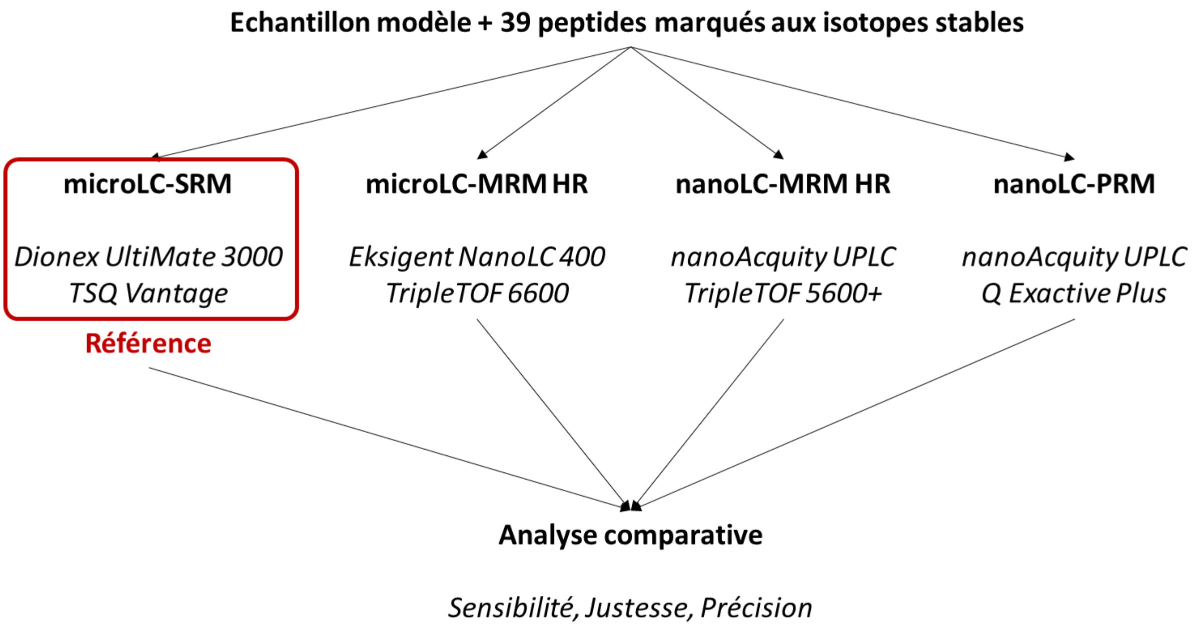

Echantillon modèle + 39 peptides marqués aux isotopes stables
microLC-SRM
Dionex UltiMate 3000
TSQ Vantage
Référence
microLC-MRM HR
Eksigent NanoLC 400
TripleTOF 6600
nanoLC-MRM HR
nanoAcquity UPLC
TripleTOF 5600+
nanoLC-PRM
nanoAcquity UPLC
Q Exactive Plus
Analyse comparative
Sensibilité, Justesse, Précision

*Figure 2 : Workflow pour la comparaison de quatre configurations instrumentales pour l'analyse protéomique ciblée.*

La sensibilité de chaque couplage a été estimée en comparant les ratios signal / bruit des courants d'ions extraits des fragments suivis, la justesse de quantification a été estimée en considérant le couplage SRM comme référence, et la précision a été estimée en calculant des coefficients de variation entre des triplicatas techniques.

Globalement, les quatre couplages testés ont montré des performances équivalentes en terme de sensibilité, justesse et précision, ce qui concorde avec d'autres études[25, 49-50]. Le choix du couplage à utiliser pour des analyses ciblées doit donc se faire selon d'autres critères. Par exemple, la robustesse est un point clé des approches ciblées, et un couplage fonctionnant en mode microLC est bien plus robuste qu'un couplage nanoLC[51]. La disponibilité des instruments doit aussi être prise en compte, les instruments de type triple quadripôle étant généralement dédiés aux analyses SRM, alors que les instruments HR/AM peuvent également réaliser des analyses DDA ou DIA.

En conclusion, ceci nous conforte dans notre choix d'utiliser préférentiellement des couplages de type microLC-triple quadripôle pour nos analyses ciblées.

## III. Optimisation d'un workflow « data independent acquisition »

Le mode d'acquisition « data independent acquisition » (DIA) a été introduit récemment pour les instruments de dernière génération de type HR/AM. Le mode DIA promet de combiner les avantages des modes DDA et ciblé, en permettant la quantification de tous les peptides au-dessus du seuil de détection avec de grandes sensibilité, spécificité et reproductibilité.

Cependant, de par la nouveauté et la complexité de l'analyse des données du mode DIA, son utilisation n'est pas commune. De plus, la DIA doit encore faire ses preuves en terme de sensibilité et reproductibilité. Nous avons donc optimisé un workflow complet pour l'analyse en mode DIA de type « sequential windowed acquisition of all theoretical fragment ion mass spectra » (SWATH), de la préparation d'échantillons jusqu'au traitement des données, en passant par l'acquisition des données.

### A. *Analyse des données*

Nous avons tout d'abord optimisé l'analyse des données DIA, qui est aujourd'hui la partie la plus difficile du workflow. Aujourd'hui, c'est la stratégie centrée sur les peptides qui donne les meilleurs résultats[29]. Dans cette approche, une librairie spectrale est utilisée pour extraire les données DIA de façon ciblée. La librairie spectrale contient les informations de m/z des peptides et fragments à rechercher, leur temps de rétention ainsi que l'intensité relative des fragments. Afin d'optimiser l'extraction des données, nous avons utilisé un échantillon bien défini, constitué d'un extrait protéique de levure dans lequel a été ajouté soit 25 fmol soit 5 fmol d'un mélange équimolaire de 48 protéines standards UPS1 (Universal Proteomics Standard, UPS1, Merck). Après avoir construit une librairie spectrale à partir de données que nous avons acquises en mode DDA, ces deux échantillons ont été analysés en mode DIA-SWATH. Les données ont été extraites grâce à la librairie spectrale en utilisant deux logiciels, Skyline qui est un logiciel libre[52], et Peakview qui est le logiciel propriétaire de SCIEX. De plus, différents réglages ont été comparés pour la tolérance en temps de rétention (RT), la fenêtre d'extraction, le nombre de transitions par peptide, ainsi que le seuil de faux positifs (false discovery rate ou FDR). Chaque paramètre a été optimisé indépendamment en utilisant le fait que nous savons précisément ce qu'il y a dans nos échantillons et ce que nous sommes censés observer : les protéines de levure sont en concentration égale dans les deux échantillons et les protéines UPS1 sont cinq fois plus abondantes dans l'échantillon avec 25 fmol comparé à l'échantillon avec 5 fmol. Nous avons donc calculé des taux de vrais positifs (TPR) et de faux positifs (FDP), les vrais positifs étant les peptides UPS1 détectés comme différentiels entre les deux échantillons, et les faux positifs étant les protéines de levure détectées comme différentielles entre les deux échantillons. Nous avons pu ainsi dégager des paramètres optimisés pour l'extraction des données (**Tableau 2**).

*Tableau 2 : Paramètres optimisés pour l'extraction ciblée des données DIA.*

| Paramètre | Optimum |
|---|---|
| Logiciel | Skyline |
| Tolérance RT | Déterminée empiriquement |
| Fenêtre d'extraction | Egale à la résolution du ToF |
| # transitions | 6 |
| Seuil FDR | 1 % |

Ensuite, nous avons utilisé un échantillon d'hépatocytes humains afin d'évaluer l'utilisation d'une librairie spectrale publique (SWATHAtlas humaine[53]) comparée à une librairie que nous avons construite par analyse DDA de 27 bandes de gel SDS-PAGE, dite « homemade ». L'utilisation d'une librairie spectrale publique, en plus d'être plus complète, pourrait faire gagner un temps précieux. Les données DIA ont été extraites en utilisant notre librairie homemade contenant 30 982 peptides correspondant à 3 644 protéines, et la librairie SWATHAtlas contenant 139 449 peptides correspondants à 10 316 protéines (**Figure 3**).

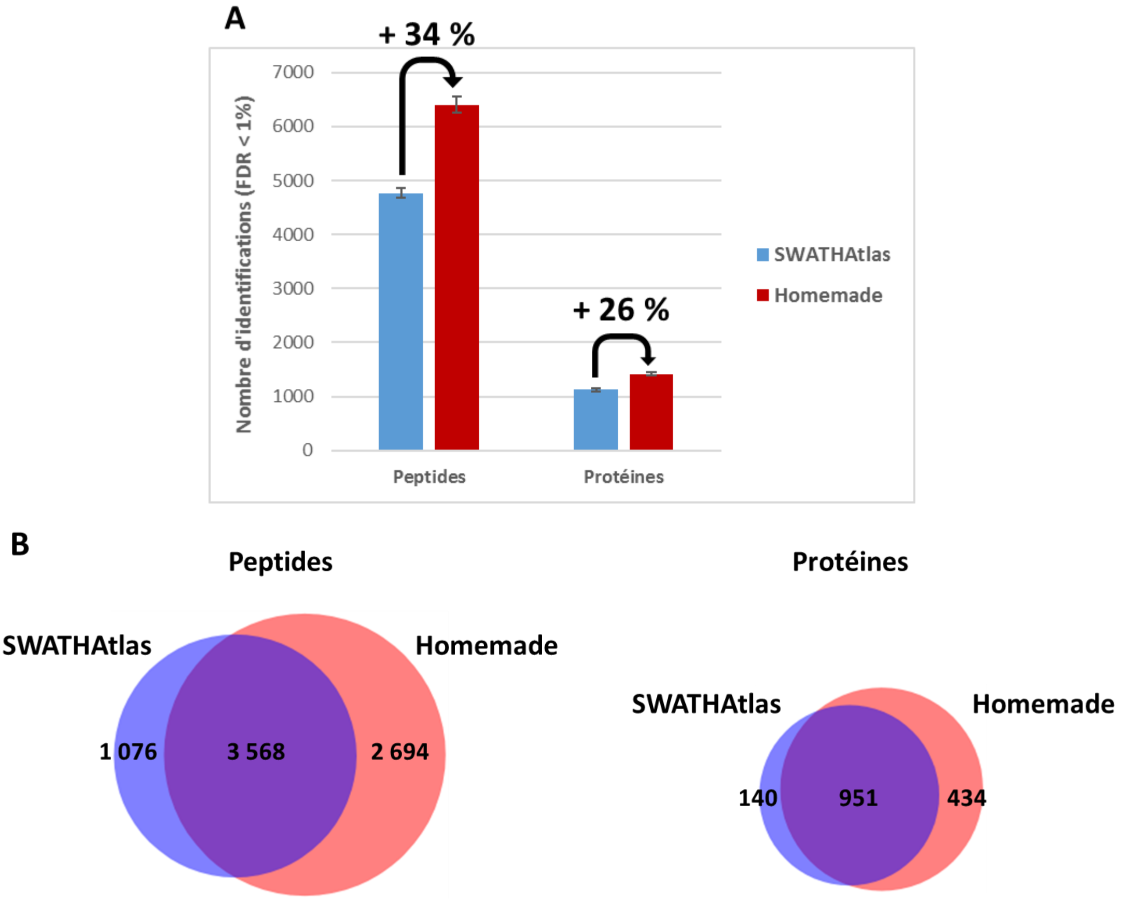


***Figure 3 : Comparaison des performances entre une librairie homemade et la librairie publique SWATHAtlas humaine.*** *A. Le nombre de peptides et protéines identifiés est présenté. B. Des diagrammes de Venn ont été réalisés pour les peptides et les protéines identifiés dans au moins deux réplicas techniques sur trois.*

En utilisant la librairie homemade, 34% de peptides et 26% de protéines supplémentaires ont été identifiés comparé à l'utilisation de la librairie publique SWATHAtlas humaine. Nous avons ensuite montré que cette différence n'était pas due à la qualité des informations présentes dans la librairie spectrale SWATHAtlas, mais plutôt au nombre trop important de peptides dans la librairie. En effet, ce nombre trop important, avec une majorité de peptides qui sont en réalité absents de nos échantillons (la librairie SWATHAtlas humaine a été construite à partir de 331 analyses DDA de différents tissus et lignées cellulaires humaines), empêche la différenciation entre les peptides ciblés et les peptides leurres utilisés dans les approches « target decoy » afin de définir un seuil de faux positifs. En effet, la grande majorité des peptides ciblés, tout comme les peptides leurres, sont absents de notre échantillon.

### B. *Acquisition des données*

En mode DIA-SWATH, les peptides sont isolés par fenêtres le long de la gamme de m/z analysée. L'acquisition des données peut être optimisée principalement en modifiant le schéma des fenêtres d'isolement, notamment en changeant le nombre de fenêtre ou en utilisant des fenêtres variables.

Nous avons montré que l'utilisation de fenêtres plus petites permettait d'augmenter le nombre de peptides identifiés en réduisant les interférences. En effet, une méthode avec 68 fenêtres de 12.5 Da nous a permis d'identifier 42% de peptides et 31% de protéines supplémentaires comparé à une méthode avec 34 fenêtres de 25 Da.

L'optimisation d'une méthode SWATH utilisant des fenêtres variables se fait en fonction de la densité des peptides le long de la gamme de m/z : plus les régions seront denses, plus les fenêtres seront petites, et ceci dans le but d'égaliser la densité des peptides dans les fenêtres et ainsi réduire les interférences dans les zones denses. Nous avons comparé deux méthodes utilisant 100 fenêtres d'isolement, l'une avec des fenêtres variables, l'autre avec des fenêtres fixes, mais nous n'avons pas pu observer d'amélioration significative en nombre d'identifications en utilisant des fenêtres variables. Ceci est probablement dû au grand nombre de fenêtres utilisées, rendant la différence en taille des fenêtres trop faible pour qu'elle ait un effet visible. Ceci reste donc à confirmer.

### C. *Préparation d'échantillon*

La préparation d'échantillon est la première étape de tout workflow de protéomique. Son importance est souvent négligée, mais cette étape va conditionner la sensibilité et la reproductibilité des analyses. La préparation d'échantillon doit être aussi simple et rapide que possible, car chaque étape peut introduire de la variabilité. Les échantillons étant souvent disponibles en quantités limitées, nous avons évalué l'impact de différentes quantités d'échantillon chargées sur un gel de concentration, ainsi que différentes quantités injectées sur un couplage nanoLC-Triple TOF 5600.

Nous avons montré qu'en déposant moins de 50 µg sur un gel, la sensibilité de l'analyse était légèrement réduite, sans doute à cause de la fixation des peptides après la digestion sur les parois en plastique des tubes ou des plaques 96 puits. En chargeant entre 50 et 100 µg, la sensibilité restait équivalente.

Nous avons ensuite montré que plus on injecte d'échantillon sur le couplage, plus on identifie de peptides et protéines (+ 95% de peptides et +84% de protéines identifiées en injectant de 100 ng à 1 µg), jusqu'à atteindre un plateau pour 1 µg d'échantillon injecté. Il est inutile d'injecter plus car la sensibilité n'en sera pas améliorée, et ce sera délétère pour la stabilité de l'instrument. Cependant, cette observation compte pour un couplage incluant une nanoLC, et pour ce type d'échantillon (digest de levure). Pour un échantillon moins complexe, il faudra injecter moins d'échantillon, et pour un échantillon plus complexe, on pourra injecter plus d'échantillon, tout en faisant attention à ne pas saturer le spectromètre de masse. En effet, ce qui est important pour l'identification des peptides, c'est la quantité des peptides individuels qui entre dans le spectromètre de masse au cours du temps, et la quantité d'échantillon à analyser doit donc être adaptée au nombre de peptides qu'il contient.

### D. *Comparaison entre DDA et DIA*

Enfin, nous avons souhaité conclure ces optimisations par la comparaison des performances d'une analyse DIA avec celles d'une analyse DDA. Nous avons donc comparé les performances en terme de recouvrement de protéome entre (i) une analyse DIA effectuée sur un échantillon de levure, en extrayant les données comme optimisé précédemment, et en validant les identifications avec le modèle de notation des pics mProphet[54] à 1% FDR, et (ii) une analyse DDA effectuée sur le même échantillon, en effectuant une recherche Mascot et en validant les identifications avec le logiciel Proline (http://proline.profiproteomics.fr/) à 1% FDR (**Figure 4**).

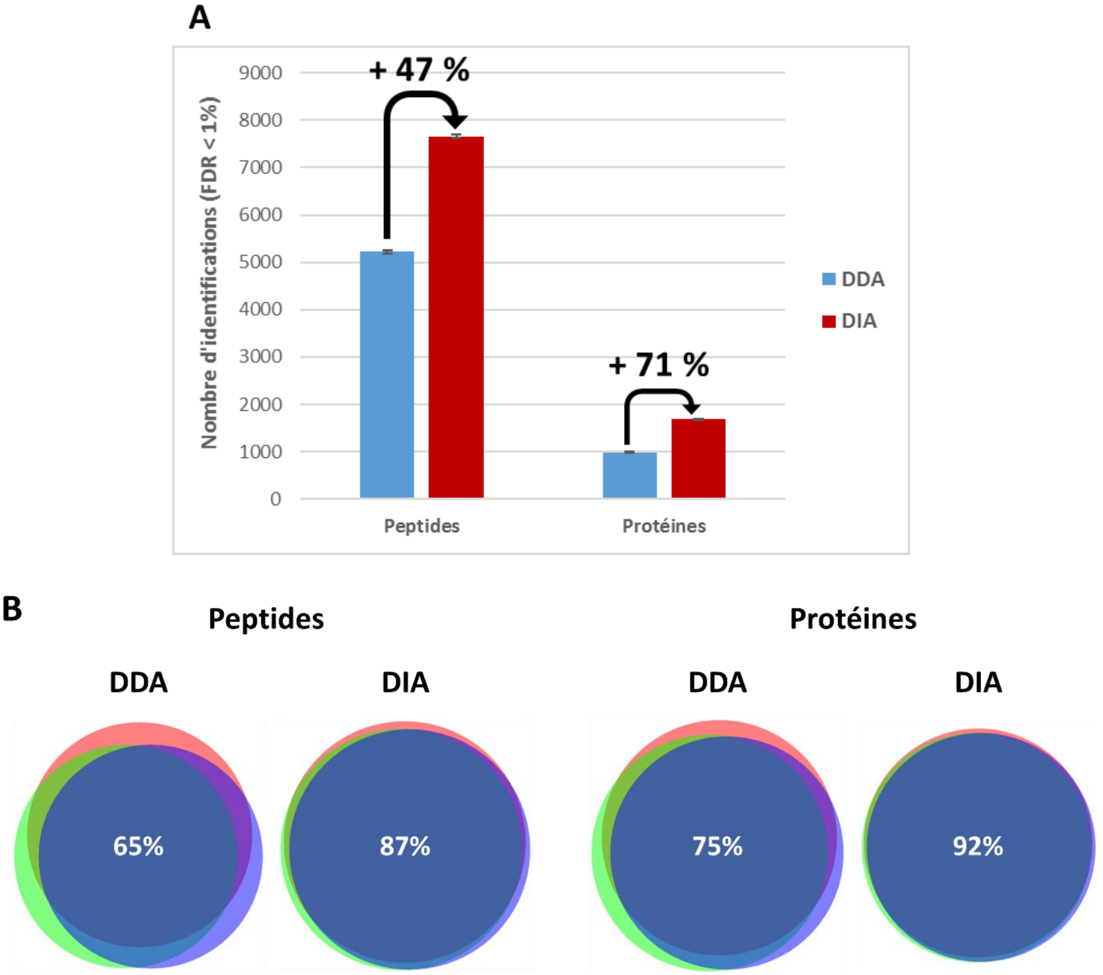


***Figure 4 : Comparaison des performances d'identification en mode DDA et DIA.***
*A. Le nombre de peptides et protéines identifiés en DDA et DIA sont présentés. B. La reproductibilité en terme d'identifications a été évaluée en réalisant des diagrammes de Venn pour les peptides et protéines identifiés dans chaque réplica technique. La valeur indiquée dans chaque diagramme est le pourcentage d'identifications communes entre les trois réplicas.*

Nous avons pu montrer que sur cet échantillon, les performances d'identification du mode DIA sont nettement meilleures que celles du mode DDA. Le mode DIA nous a permis d'identifier 42% de peptides et 31% de protéines supplémentaires comparé au mode DDA. De, plus la reproductibilité des identifications en DIA est également nettement meilleure, avec des recouvrements au sein de triplicatas techniques de 87% pour les peptides et 92% pour les protéines, contre seulement 65% pour les peptides et 75% pour les protéines en mode DDA, ce qui montre que le sous-échantillonnage est bien réel en mode DDA. Néanmoins, des identifications croisées sont possible en mode DDA, et même si un peptide n'a pas été identifié directement dans une analyse mais l'a été dans une autre, en alignant les temps de rétention de ces échantillons le peptide peut être identifié si son précurseur est détecté, ce qui permet de réduire le sous-échantillonnage du mode DDA[55].

Nous avons ensuite comparé le nombre de peptides quantifiés de manière reproductible en mode DDA et DIA, en appliquant un filtre de coefficient de variation de 20% sur les valeurs d'aires sous les pics entre les triplicatas techniques. Pour le mode DDA, nous avons effectué une quantification par extraction des courants d'ions des ions précurseurs (XIC MS1), et en DIA nous avons utilisé les paramètres optimisés précédemment (**Figure 5**).

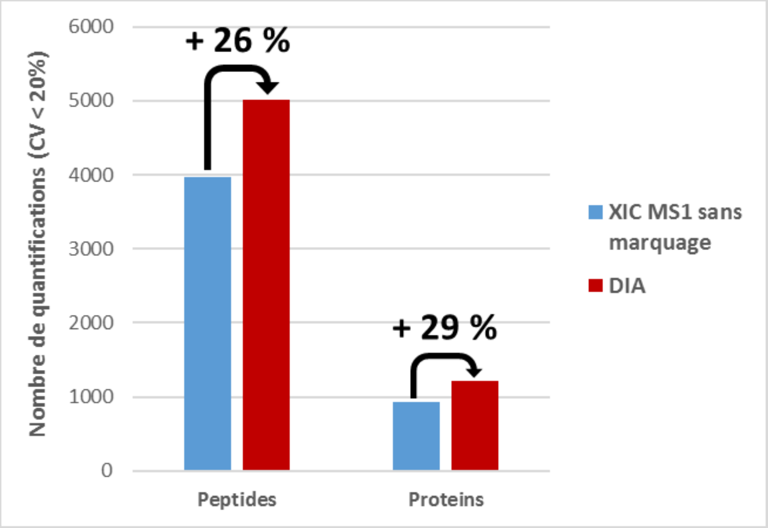


***Figure 5 : Comparaison des performances de quantification en mode DDA et DIA.***
*Le nombre de peptides et protéines quantifiés avec un coefficient de variation entre les triplicatas techniques inférieur à 20% (appliqué sur les valeurs d'aires sous les pics) par XIC MS1 sans marquage (à partir des données DDA) et DIA-SWATH sont présentés.*

Le mode DIA nous a permis de quantifier 26% de peptides et 29% de protéines supplémentaires par rapport au mode DDA couplé à la quantification XIC MS1 sans marquage. Cette comparaison entre les modes d'acquisition DDA et DIA doit cependant être approfondie, et des analyses sont en cours au laboratoire afin d'évaluer la justesse et la gamme dynamique de l'approche DIA.

### E. *Conclusion*

Dans cette partie, nous avons optimisé le workflow complet de l'analyse DIA. Premièrement, le point le plus critique était l'analyse des données DIA, que nous avons optimisée à l'aide d'un échantillon bien défini. Puis nous avons montré que pour le moment, les librairies dites « homemade » donnent de bien meilleurs résultats que les librairies publiques comme SWATHAtlas. Ensuite, nous avons montré que l'utilisation d'un grand nombre de fenêtres permettait de réduire le taux d'interférences et ainsi de gagner en sensibilité. Nous avons également montré que la quantité de protéines chargées sur un gel pouvait légèrement influer sur la sensibilité, et qu'elle devient optimale à partir de 50 µg de protéines chargées. La quantité de peptides injectés est d'une importance cruciale, et sur un système nanoLC, injecter 1 µg d'échantillon semble optimal.

Enfin, nous avons comparé les performances en terme de recouvrement du protéome entre l'analyse en mode DDA classique et l'analyse en mode DIA, et nous avons montré qu'en termes de sensibilité et de reproductibilité, la DIA surpasse largement la DDA. Donc déjà aujourd'hui, alors que la DIA est encore en plein développement, elle offre de meilleures performances que la DDA, avec cependant le désavantage de nécessiter la création d'une librairie spectrale. Cependant, des approches de traitement de données centrées sur les spectres permettront peut-être dans le futur de se passer de cette étape.

## IV. Développement d'approches de spectrométrie de masse de pointe pour quantifier les protéines de la cellule hôte

Aujourd'hui, les protéines de la cellule hôte (HCP) sont généralement quantifiées par ELISA[43-45]. Cependant, l'ELISA souffre d'inconvénients majeurs, comme (i) un recouvrement des HCP incomplet, car les anticorps anti-HCP utilisés ne peuvent pas reconnaître toutes les HCP présentes, et (ii) l'ELISA ne fournit aucune information quant à l'identité des HCP détectées et ne fournit qu'une valeur de concentration d'HCP totale. De plus, un nombre croissant d'études montre que l'ELISA n'est pas capable de fournir une quantification complète des HCP[43-45, 56-58]. Il y a donc un besoin urgent de développer des méthodes alternatives, parmi lesquelles la spectrométrie de masse est la plus prometteuse, car elle permet d'identifier et de quantifier individuellement chaque HCP, sans les biais inhérents à l'ELISA[44-45].

Dans ce contexte, nous avions pour objectif de développer des approches de spectrométrie de masse de pointe pour identifier et quantifier les HCP dans des échantillons de mAbs. Pour ce projet, j'ai réalisé toutes les étapes, depuis la préparation d'échantillon à l'University College London, jusqu'à la quantification des HCP par spectrométrie de masse à l'Université de Strasbourg.

La préparation d'échantillons a été réalisée au sein du Département d'ingénierie biochimique à l'University College London, où j'ai pu suivre de nombreuses formations pour la culture de cellules, la production et la purification de mAbs. Pour ce projet, UCB Pharma nous a fourni une lignée de cellules CHO produisant un mAb que nous avons utilisée comme modèle. J'ai cultivé ces cellules à petite échelle dans des flasques sous agitation, récupéré le surnageant de culture contenant le mAb sécrété, et réalisé la première étape de purification classique d'un mAb, à savoir une purification par chromatographie d'affinité en utilisant une colonne de protéine A, qui permet d'éliminer la grande majorité des HCP[59]. Ainsi, j'ai généré une gamme d'échantillons à partir de différentes étapes du procédé de purification, c'est-à-dire du surnageant de culture cellulaire clarifié (CCCF) et des échantillons post-protéine A (PPA), obtenus en utilisant différentes conditions de production, comme différentes durées de culture, différents stress de cisaillement, et différents protocoles de purification protéine A. Au total, 4 fractions CCCF et 8 PPA ont été générées, pour un total de plus de 600 aliquotes.

De retour à l'Université de Strasbourg au sein du laboratoire de spectrométrie de masse bioorganique (LSMBO), j'ai développé une gamme de méthodes analytiques basées sur la spectrométrie de masse afin de quantifier les HCP présentes dans les échantillons produits. Après avoir construit une librairie spectrale complète des HCP, nous avons développé une méthode originale basée sur le mode d'acquisition DIA, combinant un profilage global des HCP avec une quantification absolue d'HCP clés en une seule analyse. Le profilage global des HCP a été réalisé par des estimations dites « Top 3 », c'est-à-dire qu'on admet que le signal des trois peptides les plus intenses par mole de protéine est constant, avec un coefficient de variation de moins de 10%[60]. La quantification absolue des HCP clés a été réalisée grâce à la dilution isotopique (ID). Globalement, les HCP ont été quantifiées dans une gamme couvrant 5 ordres de magnitude, et jusqu'à moins de 1 ppm. Cette méthode, appelée Top 3-ID-DIA, a été comparée aux méthodes de référence ELISA pour la quantification des HCP, et SRM couplée à la dilution isotopique (ID-SRM) pour la quantification absolue par spectrométrie de masse. La méthode Top 3-ID-DIA a montré des sensibilité, justesse et précisions comparables à celles de la méthode ID-SRM.

En conclusion, la méthode Top 3-ID-DIA développée pourrait fournir une aide conséquente pour le développement de procédés de production ainsi que pour s'assurer de la pureté d'un biomédicament.

Ce travail a été soumis au journal Analytical Chemistry de l'American Chemical Society.

# **Chapter III** Conclusion générale

La première partie de ce manuscrit est une introduction bibliographique présentant les états de l'art de l'analyse protéomique et du domaine des anticorps monoclonaux. La seconde partie du manuscrit décrit les principaux résultats obtenus, pour le développement et l'optimisation d'approches MS, et pour la quantification des HCP dans des échantillons d'anticorps monoclonaux.

Tout d'abord, j'ai été responsable d'un couplage de dernière génération microLC-Triple TOF 6600, pour lequel j'ai optimisé une méthode DDA pour des analyses de type « shotgun », permettant à la fois d'identifier des protéines et de les quantifier par l'extraction des courants d'ions.

Nous avons également comparé les performances de différentes configurations d'analyse protéomique ciblée afin de nous guider dans notre choix instrumental lorsque de telles approches sont envisagées. Nous avons montré que des systèmes microLC étaient aussi performants que des systèmes nanoLC, si la quantité d'échantillons disponible est suffisante. De plus, la SRM et la PRM ou la MRM HR ont montré des sensibilité, justesse et précision équivalentes. Donc le choix de l'approche à utiliser pour des analyses de protéomique ciblée doit plutôt se faire sur des critères de (i) quantité d'échantillon, en préférant le mode microLC par rapport au nanoLC lorsque c'est possible car il offre une meilleure robustesse, (ii) instrument disponible, sachant que les instruments de type triple quadripôle sont souvent dédiés aux analyses SRM alors que les instruments HR/AM peuvent réaliser d'autres types d'approches, i.e. « shotgun » ou DIA, (iii) le développement de méthode est facilité si l'approche ciblée est réalisée sur le même instrument que l'approche globale.

Nous avons également optimisé le workflow complet DIA, incluant la partie préparation d'échantillons, acquisition des données et analyse des données. L'analyse des données est aujourd'hui le goulot d'étranglement de cette technique, c'est pourquoi nous l'avons profondément optimisé. Le workflow optimisé DIA nous a permis d'atteindre des recouvrements de protéome bien meilleurs que l'analyse « shotgun » classique DDA, et des sensibilité, spécificité et reproductibilité équivalentes à celles des approches ciblées. En conclusion, l'approche DIA semble tenir ses promesses, et les futures avancées instrumentales en terme de vitesse, sensibilité et résolution, ainsi que les améliorations de l'analyse des données DIA, feront sans doute de cette méthode la référence pour l'analyse protéomique dans les années à venir.

Enfin, j'ai produit une large gamme d'échantillons de mAb provenant de différentes étapes et conditions du procédé de production d'un mAb, et ai développé une approche DIA innovante, appelée Top 3-ID-DIA, permettant à la fois un profilage complet de la population des HCP et une quantification absolue d'HCP clés. Nous avons pu quantifier les HCP avec une gamme dynamique de 5 ordres de magnitude, avec une sensibilité inférieure à 1 ppm. Cette méthode a été comparée aux méthode de référence ELISA pour la quantification des HCP, et SRM pour la quantification absolue par spectrométrie de masse. L'approche Top 3-ID-DIA nous a permis d'atteindre des sensibilité, justesse et précisions comparables à la SRM, tout en permettant une quantification non biaisée et plus complète comparé à l'ELISA.

Cette méthode peut être transférée en industrie, et peut être appliquée après seulement deux mois de développement, là où il faudrait plus d'un an pour développer un test ELISA. Bien que cette méthode soit en apparence plus chère qu'un ELISA (un couplage LC-MS vaut ≈ 600 k€, et le développement d'un ELISA ≈ 200 k€), il ne faut pas oublier qu'un nouveau test ELISA doit être développé lorsqu'il n'y a plus de réactif disponible, et les résultats entre le nouveau et l'ancien kit ELISA sont rarement concordants. De plus, la qualité et la quantité des informations obtenues en MS sont nettement supérieures à ce qui est obtenu en ELISA : par exemple, nous avons quantifié plus de 3 000 HCP en MS, alors qu'on estime que l'ELISA quantifie ≈ 1 000 HCP. La MS nous donne également une quantification individuelle des HCP, alors que l'ELISA ne fournit qu'une quantité totale d'HCP, sans information concernant le nombre ou l'identité des HCP détectées.

En conclusion, l'approche Top 3-ID-DIA pourrait offrir un support important pour le développement de procédé de production de mAb et la vérification de la pureté des mAbs, permettant au final la production de biomédicaments plus purs et plus sûrs. A court terme, la méthode Top 3-ID-DIA pourrait fournir des résultats complémentaires à ceux obtenus en ELISA, et à long terme le remplacer totalement.

# General introduction

Proteins constitute a key class of biomolecules, which perform a range of essential functions involved in various biological processes like biochemical reactions, cell structure maintenance, intracellular trafficking or cell signalling, protein regulation or signalisation. The protein composition in amino acids is coded by the genome, and post translational maturations lead to their active three-dimensional conformation. The total number of proteins per cell is estimated at 3-4 million for the bacteria *Escherichia coli*, 100-150 million for the yeast *Saccharomyces cerevisiae*, and 10 billion for a mammalian cell[1]. Contrary to the genome, which is constant within an organism, the proteome is a dynamic entity evolving according to time, external stimuli or cell type. For instance, mouse fibroblasts express about 10 000 different proteins, covering a dynamic range of seven orders of magnitude (from one to ten million copies per cell) following a bell-shape distribution (**Figure 6**).

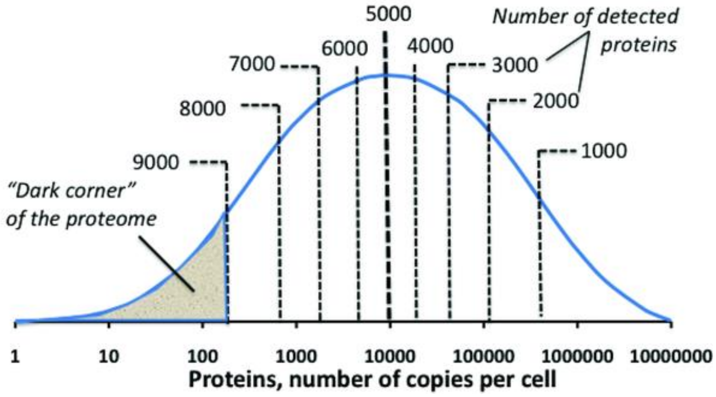


***Figure 6 : Distribution of protein abundances in NIH3T3 mouse fibroblasts (adapted from [2]).***
*The "dark corner" of the proteome is the most challenging part for detection and represents about 1 000 proteins.*

Moreover, like alternative splicing adds a level of complexity from the genome to the transcriptome, post translational modifications add a supplemental level of complexity from the transcriptome to the proteome. Different versions of a protein are called proteoforms[61]. In human cells, it is estimated that 90% of the ≈ 20 000 genes undergo alternative splicing[62], and approximately 100 000 proteoforms can be expressed[63]. The total protein content of a cell at a given time is called the proteome.

Proteomics is the large scale and comprehensive study of the proteome: it aims to identify, quantify and characterise all the proteins of a proteome[3-4]. The expansion of proteomics was driven by technological advances in separation techniques, mass spectrometry (MS) and bioinformatics[5]. In the past two decades, improvements of the sensitivity, resolution, mass accuracy and scan rate of mass spectrometers, as well as the growth of curated and annotated protein sequence databases have made MS the most important and popular tool for high throughput and large scale proteomics[6]. Today, proteomics plays an essential role in major research areas, including systems biology and biomarker discovery, and strongly contributes to the understanding of biological processes[7-8].

My doctoral work was intended to improve proteome analysis by MS by setting up and evaluating a new method using Data Independent Acquisition (DIA), and to demonstrate the major interest and potential of MS methodologies for the study of host cell protein impurities in monoclonal antibody solutions.

**The Part I** of this manuscript is a bibliographic introduction. **The Chapter I** summarises the state of the art of proteomics, including shotgun analyses, targeted approaches and the very recent and promising data independent acquisition mode. **The Chapter II** presents the field of monoclonal antibodies and a state of the art of host cell protein impurities detection methods.

**The Part II** of this manuscript presents the main analytical and methodological developments and evaluations that were realised during this PhD, and their application to the study of host cell protein impurities in monoclonal antibody samples. **The Chapter I** presents the key parameters of a data dependent acquisition method and their optimisation to improve the proteome coverage of shotgun proteomics analyses. **The Chapter II** describes a benchmarking of four targeted proteomics platforms, including the gold standard triple quadrupole for selected reaction monitoring (SRM) and new generation mass spectrometers performing high resolution SRM-like experiments. **The Chapter III** focuses on the optimisation of the data independent acquisition workflow, from the sample preparation to data acquisition and analysis. **The Chapter IV** presents the application of these analytical developments to the study of host cell proteins.

# Part I Bibliographic introduction

# **Chapter I** Bottom-up proteomics

Proteomics can be divided into three approaches: top-down, bottom-up, and middle-down.

In the **top-down approach**, proteins are directly analysed intact by MS and following fragmentation by MS/MS, providing information on the intact protein mass and amino acid sequence. The objective of the top-down approach is to provide high sequence coverage and a comprehensive characterisation (e.g. post translational modifications, proteoforms) of a targeted protein. The top-down approach allowed the analysis of proteins > 200 kDa[64-65], and identification of more than 1 000 proteins and thousands of proteoforms[66-68]. However, this approach suffers from a sensitivity limitation (> 100 fmol) linked to difficulties with protein solubility, separation, ionisation and fragmentation[69-70]. Due to the complexity of the signals and the multiple charge states of intact proteins, the top-down approach is still best suited for the analysis of highly purified samples. High resolution instruments, like time-of-flight or Fourier transform-based instruments, are also needed to resolve isotopic envelopes of the proteins. Finally, dedicated instrumental software and bioinformatic pipelines still need to be improved[70].

The **bottom-up approach** is based on the digestion of proteins into peptides of ≈ 500-3 000 Da prior to MS analysis. The analysis of peptides rather than proteins offers an increased sensitivity due to a better separation by liquid chromatography (LC), a lower molecular weight and fewer charge states[71]. Peptides are identified by comparing the measured masses to theoretical masses obtained *in silico* using a protein sequence database. Since peptides can be either assigned to a single protein or shared among several proteins, the identified proteins are scored and grouped based on their identified peptides. The result of bottom-up proteomics is the smallest list of identified protein groups explaining the maximum number of peptide identification[11].

The **middle-down approach** aims to combine the best of top-down and bottom-up approaches[72]. The proteins are digested into large peptides of ≈ 3 000-20 000 Da to minimise peptides redundancy between proteins compared to the bottom-up approach. These large peptides allow an improved characterisation regarding PTMs without the challenges of analysing intact proteins.

A summary of these approaches is presented in **Figure 7**.

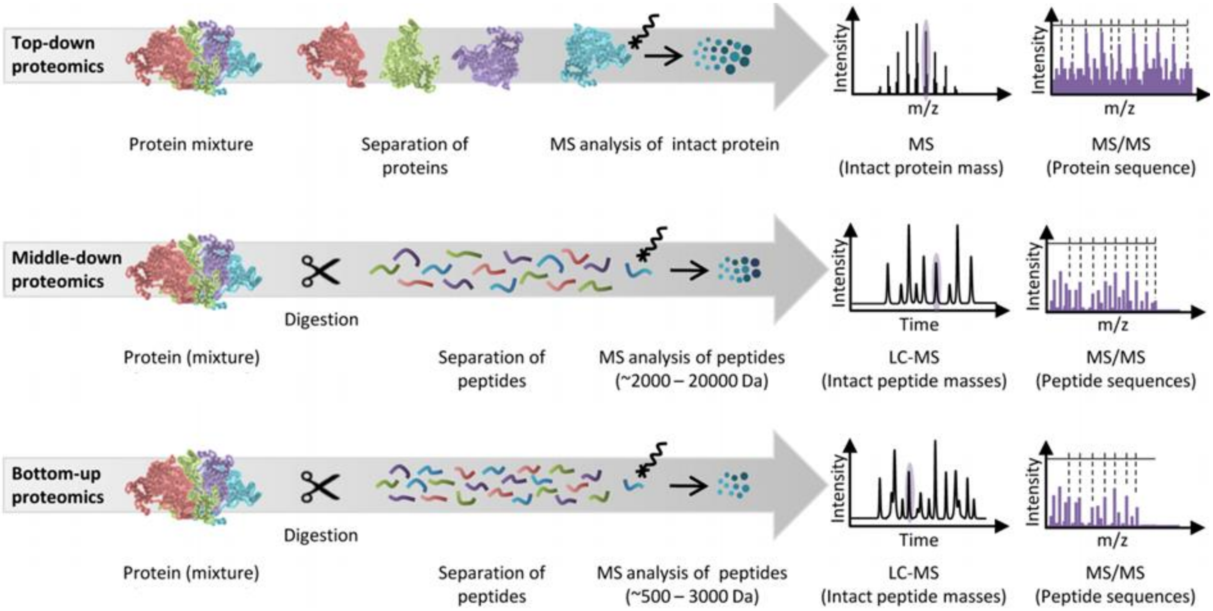

Top-down proteomics
Protein mixture
Separation of proteins
MS analysis of intact protein
Intensity
m/z
MS
(Intact protein mass)
MS/MS
(Protein sequence)
Middle-down proteomics
Digestion
Protein (mixture)
Separation of peptides
MS analysis of peptides (~2000 – 20000 Da)
Time
LC-MS
(Intact peptide masses)
MS/MS
(Peptide sequences)
Bottom-up proteomics
MS analysis of peptides (~500 – 3000 Da)

*Figure 7 : Overview of the three MS-based proteomics approaches (adapted from [73]).*

While top-down and middle-down methods will likely provide complementary information in the future, today, the bottom-up approach is still the major workhorse method for the large scale analysis of proteins and their characterisation. It is widely used in many research fields like disease biomarker discovery or systems biology[4].

In this work, we have exclusively used bottom-up approaches, which are detailed below.

## I. Analytical workflow

The proteome is an extremely challenging sample for analytical sciences, not only due to the number of different proteins with up to 100 000 proteoforms in human cells[63], but also to its dynamic range up to seven orders of magnitude[2, 74]. On the other hand, mass spectrometers can reach a dynamic range of 3 to 5 orders of magnitude depending on the acquisition parameters. Therefore, to reach an optimal sensitivity, specificity and proteome coverage, it is necessary to reduce the sample complexity prior to MS analysis. The complexity of a sample can be reduced by depletion of highly abundant proteins, or fractionation at the protein level or the peptide level.

## A. *Sample preparation*

The sample preparation is the first step of any analytical workflow. Since each operation can bring variability, it must ideally be as easy and as fast as possible for an optimal reproducibility. Prior to MS analysis, the sample preparation usually starts with a quantification of total proteins, as it is important to inject in the end an adapted quantity of sample into the mass spectrometers to avoid dirtying them and to preserve their performances. To perform relative quantification between samples, it is also important to compare the same amount of total proteins of each sample. After quantification of their total protein amounts, the samples can be purified and their complexity reduced at the protein level and, after digestion, at the peptide level to improve the sensitivity and the proteome coverage of the assay[75].

### A.1.Total protein quantification

Total protein quantification is one of the most frequently performed assay in biological research. It is often an underestimated step, but it will actually condition the quantification results, as for any quantification experiment, the same amount of total proteins must be analysed. When working with highly complex samples, the most used methods for total protein quantification of complex samples are the Bradford assay and the modified Lowry assay, which are compared below.

The Bradford assay[76] is based on the binding of the dye Coomassie Brilliant blue G-250 in acidic conditions to arginine, histidine, phenylalanine, tryptophan and tyrosine residues, which induce a metachromatic shift from 465 to 595 nm[77]. The advantages of the Bradford assay include its ease of use, sensitivity and low cost. However, it is interfered with detergents, and the majority of the observed signal is due to the binding of the dye to arginine residues, resulting in wide variations between proteins according to their arginine content.

The Lowry assay[78] is based on the Biuret reaction, involving the reduction of copper $Cu^{2+}$ to $Cu^{+}$ by proteins in alkaline solution, followed by the reduction of the Folin-Ciocalteu reagent. A blue color with absorbance maximum at 750 nm is produced by $Cu^{+}$-peptide bond complex, but also tyrosine, tryptophane, and to a lesser extent cystine, cysteine and histidine residues[77]. Since peptide bonds are the major actors that produce the dye, less variations between proteins are observed using the Lowry assay. The Lowry assay has been modified to improve its tolerance to interfering agents, speed, dynamic range and stability[79]. However, it is still interfered with reducing agents.

Because of its reduced variability between proteins, I preferentially used a Lowry-based kit to quantify total protein amounts in my samples.

### A.2.Protein purification and separation

Proteins can be purified and/or separated according to their physico-chemical properties, including their molecular weight using size exclusion chromatography or sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), their charge using ion exchange chromatography, their hydrophobicity using reversed phase chromatography, their identity using affinity chromatography, their isoelectric point using isoelectric focusing, or a combination of properties using two dimensional-polyacrylamide gel electrophoresis (2D-PAGE)[6, 80].

During my PhD, I exclusively used SDS-PAGE methods for protein purification and fractionation[81]. The major advantage of these approaches is the use of SDS to solubilise the proteins. SDS is a strong detergent with high solubilising power, composed of an anionic head group and a lipophilic tail. It binds uniformly and non-covalently the proteins, denaturing them at high temperature and providing them negative charges, whatever their original charge state. However, it interferes with trypsin digestion and must therefore be removed after protein solubilisation. To purify the proteins and remove SDS, we used gel-based methods, which allow the washing of SDS while proteins are immobilised into the gel matrix.

When protein fractionation was needed, proteins were separated by SDS-PAGE. Since all proteins were negatively charged due to SDS, they were separated only according to their molecular weight.

When protein fractionation was not needed, a stacking gel approach was used to purify the proteins: like for standard SDS-PAGE approach, the proteins were focused in a sharp band into the stacking gel, but their migration was stopped prior to their separation into the resolving gel.

More recently, tube-gel approaches were proposed, in which the proteins are directly incorporated into a polyacrylamide gel matrix without electrophoresis[82-83]. However, this approach is not compatible with all samples, e.g. samples containing thiols.

Alternatively, the filter aided sample preparation (FASP) approach was developed. It aims to combine the advantages of in-gel (impurities removal for an optimal digestion) and in-solution digestion (digestion enzyme accessibility, less variability, automation possibility)[84-86]. In this procedure, the proteins can be solubilised in a strong detergent like SDS, which is subsequently removed using a filter to exchange the buffer for a protease compatible one. After digestion on the filter surface, the peptides are retrieved by an additional filtering step. However, the FASP approach seems to suffer from protein loss due to adsorption on the filter during buffer exchange[87].

### A.3. Enzymatic digestion

Protein digestion can be performed in solution, in-gel or on a filter (FASP, described above).

When proteins were extracted without SDS, e.g. with a urea buffer, and when no protein fractionation is necessary, they can be digested in-solution by adding directly the digestion enzyme into the sample.

When using gel-based approaches, proteins are digested inside the gel[88]: briefly, gel bands of interest are cut, the dye is washed out, the immobilised proteins are reduced and alkylated to provide optimal accession to the digestion enzyme, and the gel bands are dehydrated. The gel bands are then re-swelled in the protease solution for an optimal digestion enzyme penetration into the gel[89]. After digestion, the peptides are extracted.

### A.4. Peptide purification and fractionation

If the proteins were digested in-gel or using FASP, subsequent peptide purification is not necessary since the impurities were already removed. However, if an in-solution digestion was performed, peptides should be purified prior to MS analysis to remove contaminants like urea, for instance using reversed phase solid phase extraction (SPE) or an enrichment column. Both peptide purification techniques are based on the hydrophobic binding of peptides onto reversed phase (i.e. nonpolar, typically C18) allowing impurities removal, followed by peptides elution in a nonpolar solvent like acetonitrile, which can then be removed using a vacuum drier.

Peptide separation prior to MS analysis is crucial to reduce the sample complexity and thus increase the ionisation efficiency, sensitivity and specificity resulting in a better proteome coverage. It is usually performed by reversed phase high performance liquid chromatography (HPLC)[75], which is based on the hydrophobic binding of peptides onto reversed phase (C18) and their progressive elution according to their hydrophobicity using a gradient of nonpolar solvent, typically acetonitrile.

Several parameters can influence the fractionation performances: (i) a long column improves the fractionation capacity but also the required analysis time, (ii) a reduced internal diameter increases the sensitivity by reducing the required solvent volumes and therefore sample dilution, (iii) small particle and pore sizes improve the chromatographic resolution. During my PhD, I used two types of reversed phase HPLC systems which are described in **Table 1**.

.

*Table 1 : Description of the LC systems used.*

| HPLC system | nanoLC | microLC |
|---|---|---|
| Manufacturer | Waters | Eksigent or Agilent |
| Stationary phase | C18 | C18 |
| Column length (mm) | 200 | 150 |
| Internal diameter (µm) | 75 | 300 |
| Particles size (µm) | 1.7 | 3.5 |
| Pore size (Å) | 130 | 300 |
| Flow rate (µL/min) | 0.3<br>Nano-flow | 5<br>Capillary-flow |

In the laboratory, the available sample quantity is often limited. In this context, nanoLC systems using column with reduced internal diameter are the best option for an optimal sensitivity[90], requiring small sample amount (typically from 100 ng to 1 µg of complex digest). Moreover, the use of small particles provides excellent chromatographic resolution and peak capacity. However, at such low flow rate, the interface between the LC and the MS becomes very delicate, and any undetectable leak, dead volume or sprayer issue can result in electrospray instability.

On the other hand, higher flow rates like capillary-flow provide more robust LC systems when higher sample amount is available, typically from 1 to 10 µg for capillary-flow. However, higher flow rates provide generally a lower sensitivity compared to nano-flow systems, but in some cases the decrease in sensitivity can be countered by the increased sample capacity[51].

### B. *Mass spectrometry analysis*

Mass spectrometry is an analytical technique which measures the mass over charge ratio of ions within a sample. A mass spectrometer is composed of an ionisation source, one or two analysers and a detector.

Prior to MS analysis, the peptides must be ionised and transferred to gas phase. In proteomics, soft ionisations techniques are used like matrix-assisted laser desorption ionisation (MALDI)[91], or more commonly for complex samples electrospray ionisation (ESI)[92] which can be directly coupled to LC[93]. In ESI, peptides in liquid phase will be ionised at the tip of a needle: under high voltage, droplets will take the form of a cone (Taylor cone), and the peptides will be transferred to gas phase after solvent evaporation. Once ionised, peptides are called precursor ions.

In this work, only LC-ESI-MS couplings were used.

### B.1.Tandem mass spectrometry

In bottom-up proteomics, mass analysers are mostly used in tandem, combining their properties to obtain information about the peptides sequences. The most common tandem mass spectrometer includes (i) a quadrupole used either in radio frequency (RF)-only mode to serve as an ion guide for all ions, or in analyser mode to isolate a given peptide or m/z range, (ii) a second quadrupole which is used as a collision cell to fragment the peptides (see B.2), and (iii) the analyser which can be a third quadrupole, a time-of-flight (ToF) or an Orbitrap. Therefore, by using the first quadrupole alternatively in RF-only and analyser modes, tandem mass spectrometry allows the collection of the m/z of peptides within MS spectra, and the m/z of peptide fragments within MS/MS spectra[3].

During my PhD, I used triple quadrupole, quadrupole-time-of-flight (Q-ToF) and quadrupole-orbitrap (Q-Orbitrap) type instruments.

### B.2.Fragmentation modes

Three fragmentation modes are commonly used in bottom-up proteomics : collision induced dissociation (CID)[94], electron transfer dissociation (ETD)[95] and electron capture dissociation (ECD)[96]. The most used fragmentation mode in bottom-up proteomics is CID, based on the mobile proton model[97]. A high kinetic energy is provided to the isolated ions, and their collision with neutral molecules present in the collision cell (e.g. helium, nitrogen or argon) induces the conversion of this kinetic energy into internal energy resulting in the peptide bond breakage by the mobile proton. CID fragmentation is particularly well suited for fragmentation of tryptic peptides since they usually possess at least two positive charges, one in N-terminal ($NH_3^+$) and one in the side chain of lysine (K) or arginine (R) (trypsin specifically cleaves in C-terminal position of K and R).

The higher energy collisional dissociation (HCD)[98] refers to the fragmentation mode used in Orbitrap (ThermoFisher Scientific). It follows the same principle as CID, but requires a higher energy amount because peptides must be trapped before their entry into the collision cell.

Peptide fragmentation follows well-established rules that were described in 1990 by Biemann[99] (**Figure 8**).

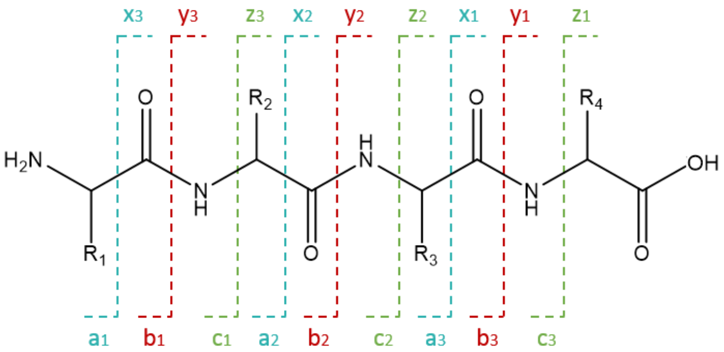


*Figure 8 : Biemann nomenclature for peptide fragmentation.*

While CID preferentially induces the production of b- and y-ions, ETD and ECD produce essentially z- and c-ions. A recently developed fragmentation mode called EThcD was developed to improve the peptide sequence coverage by combining ETD and HCD, which produce c-, z-, b- and y-ions[100].

The collection of an MS/MS spectrum of a given peptide, displaying the m/z of its fragments, allows the determination of its amino acid sequence (**Figure 9**).

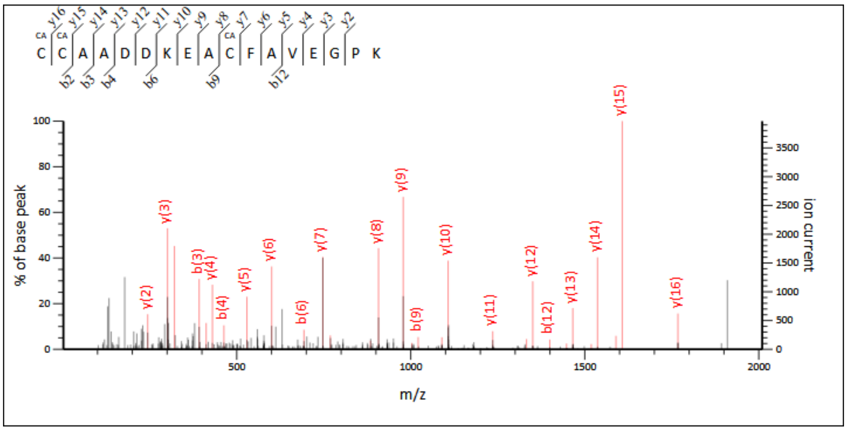


*Figure 9 : Annotated MS/MS spectrum allows the determination of the peptide's amino acid sequence.*

## II. Data dependent acquisition

In bottom-up proteomics, the most used acquisition mode is data dependent acquisition (DDA). In this mode, the mass spectrometer first acquires an MS spectrum, and the N most intense precursor ions of this MS spectrum are sequentially isolated and fragmented to collect their MS/MS spectra.

Collection of an MS spectrum followed by N dependent MS/MS spectra is called a cycle, and cycles are repeated throughout the analysis time (**Figure 10**).

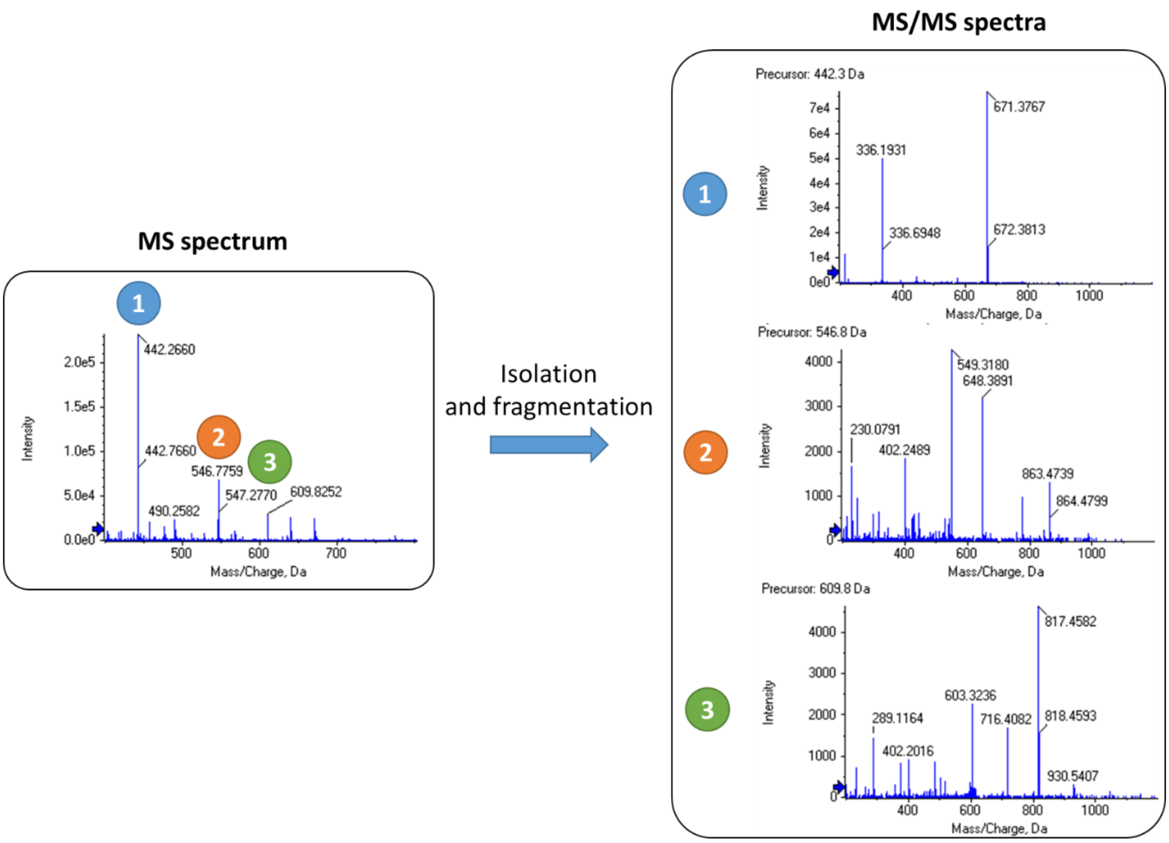


*Figure 10 : Principle of data dependent acquisition.*
*In this example, the three most intense precursor ions were sequentially isolated and fragmented.*

Technological efforts have been made to improve the sensitivity of DDA, for instance with higher scan rates or the use of dynamic exclusion to avoid selecting again and again the same most intense precursor ion and reduce the MS/MS spectra redundancy. DDA has proven to be a powerful tool for proteomics, enabling today the identification of thousands of proteins in only one hour[9].

However, DDA still suffers from an inherent limited sensitivity, because only the most intense peptides are fragmented, and only when they are selected (which is often not the moment when the peptide is the most intense), leading to a stochastic undersampling effect[101]. This also leads to a lack of reproducibility.

A thorough description of a DDA workflow that I have set up and optimised on a late generation Q-ToF instrument (Triple TOF 6600, SCIEX) will be presented in Part IIChapter I.

## III. Protein identification

In bottom-up proteomics, proteins are identified by inference from their corresponding peptides identification. Today, peptides are identified by peptide fragment fingerprinting (PFF)[10], consisting in comparing measured experimental masses of peptides and their corresponding fragments to theoretical masses from a protein sequence database using a search engine, as illustrated in **Figure 11**. Identifications can be further validated using statistical filters.

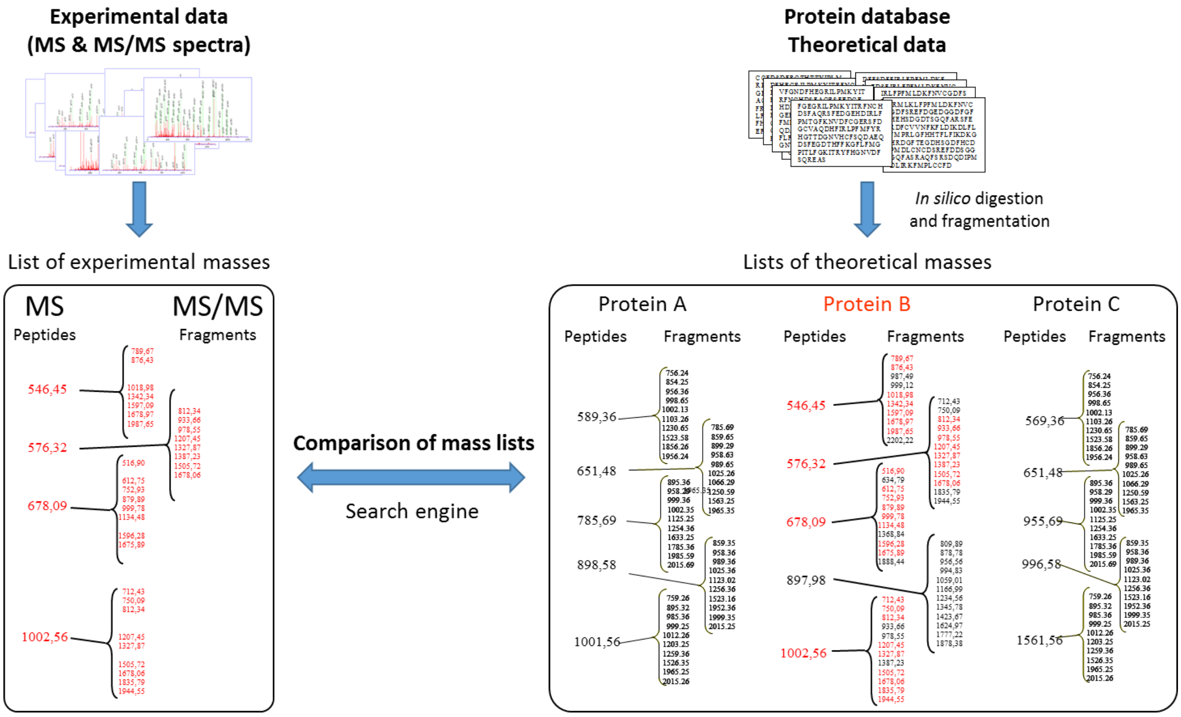

Experimental data
(MS & MS/MS spectra)
Protein database
Theoretical data
*In silico* digestion
and fragmentation
List of experimental masses
Lists of theoretical masses
MS
MS/MS
Peptides
Fragments
546,45
576,32
678,09
1002,56
Comparison of mass lists
Search engine
Protein A
Protein B
Protein C
Peptides
Fragments
589,36
651,48
785,69
898,58
1001,56
546,45
576,32
678,09
897,98
1002,56
569,36
651,48
955,69
996,58
1561,56

***Figure 11 : Principle of peptide fragment fingerprinting (PFF).***
*The experimental m/z are compared to theoretical m/z obtained from in silico digestion of the protein sequence database. In this example, experimental data allowed identification of protein B.*

### A. *Search engines*

Once acquired, the raw data are converted into mass lists including the m/z of the precursor ions and their corresponding fragments. These experimental m/z lists are compared with theoretical m/z lists of all theoretical peptides from a protein sequence database that has been digested and fragmented *in silico*. Identified peptides, which can be either assigned to a single protein or shared among multiple proteins, are used to score and group proteins to provide the smallest list of identified protein groups explaining the maximum number of peptide identifications[11].

Today, many search engines are available such as Sequest[102], Mascot[103], OMSSA[104], X!Tandem[105] and Andromeda[106]. The search engine needs the following information to perform the search: the m/z tolerance for the precursor and fragment ion; the charge states of the precursor and fragment ions; the digestion enzyme used and a maximum number of allowed missed cleavages; the fragmentation mode (e.g. CID); the protein sequence database.

In the work presented in this manuscript, I exclusively used the Mascot search engine. It is a proprietary search engine, and therefore a full description of its search algorithm is not publicly available. Roughly, an expectation value and an ion score are attributed to each peptide spectrum match (PSM), which are linked to the probability that the observed match did not occur by chance. For each MS/MS spectrum, all possible PSM are ranked and the best PSM is used for peptide identification.

Since search engines rely on different algorithms, they provide different but complementary results, giving more confidence in overlapped identifications[107]. However, it has been shown that if consistent validation criteria are used, little difference is observed between search engines[108].

### B. *Protein sequence databases*

The protein sequence database is of crucial importance for protein identification by MS, because it will condition which protein can or cannot be identified. Therefore, the protein sequence database must be adapted to the analysed sample, containing all possibly present proteins but not containing too many entries because this can lead to wrong matching by chance and false positive identifications. Moreover, the quality of the database, i.e. its curation or annotation, will condition the quality of the identifications. Several reference protein sequence databases are publicly available and the most used are briefly described below[109].

#### B.1.NCBI

The National Center for Biotechnology Information (NCBI) proposes two protein sequence databases, namely the Reference Sequence (RefSeq)[110] database and the Entrez Protein database[111]. RefSeq contains protein sequences from multiple sources with variable levels of manual curation and annotation. Entrez Protein database is a larger database with high redundancy and no data curation, containing protein sequences from publicly available databases, including RefSeq, UniProtKB/Swiss-Prot, Protein Information Ressource (PIR)[112] and the Protein Databank (PDB)[113], but also automatic translations from European Molecular Biology Laboratory (EMBL)[114], DNA Data Bank of Japan (DDBJ)[115] and GenBank[116].

### B.2.UniProtKB

The UniProt Knowledgebase (UniProtKB) aims to combine all available data for proteins and provide rich annotations, like structural information, function, localisation and cross references[117]. The database is divided into UniProtKB/TrEMBL and UniProtKB/Swiss-Prot. While UniProtKB/TrEMBL contains automatically translated and annotated protein sequences awaiting manual curation, UniProtKB/Swiss-Prot contains non redundant, manually curated and well annotated entries. Each protein entry is also scored according to the degree of evidence, e.g. if the protein has ever been detected. Today, the UniProtKB/Swiss-Prot database is the reference database for proteomics analysis of model organisms.

### C. *Validation of protein identification*

As explained previously, each peptide and protein identification is scored, and this score will be used to validate the identification. The most common approach to validate proteomics identifications is the target decoy approach[118-119]. In this approach, target proteins, i.e. the real protein sequences, are searched together with decoy proteins, which can be reversed or shuffled protein sequences. Decoy identifications are used to calculate a false discovery rate (FDR)[120], which can be calculated as follows:

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

Since decoy identifications present lower scores compared to target identifications, a score threshold is usually employed to reduce the FDR, most commonly down to 1-5%. The calculation of the FDR can be done at the PSM, peptide and protein levels, and combining FDR thresholds at different levels can improve the confidence in identifications[121].

In this work, the identifications were validated using the Proline software developed by the French proteomics infrastructure ProFI our laboratory belongs to (http://proline.profiproteomics.fr/), with the following criteria: Mascot ion score above 25 and a false discovery rate below 1% at both the peptide and protein levels.

## **IV.** Global quantitative proteomics

Protein identifications alone are most often not sufficient to answer a biological question, and quantitative information is necessary. Global quantitative proteomics aims to provide quantitative information for all detected peptides and proteins. Global quantification approaches include stable isotope label-based approaches, and label-free approaches[12].

Stable isotope label-based approaches are based on the fact that the stable isotope labelled and the unlabelled peptides have the same physico-chemical properties (same elution profile and ionisation efficiency) but a slightly different mass. Therefore, labelled and unlabelled samples can be mixed and analysed together, and a relative quantification can be performed by comparing the intensities of the labelled and unlabelled peptides. Labelling techniques are divided into (i) *in vivo* labelling like stable isotope labelling with amino acids in cell culture (SILAC)[122], and (ii) *in vitro* labelling which relies on enzymatic ($^{18}O$-labelling[123]) or chemical (isotope coded affinity tag (ICAT)[124], isobaric tags for relative and absolute quantification (iTRAQ)[125], tandem mass tags (TMT)[126]) labelling. It is of note that the *in vivo* labelling is not suited for all type of samples, and the preferred approaches today for *in vitro* labelling are iTRAQ and TMT approaches.

However, mass spectrometers are sensitive instruments which cannot handle unlimited number of ions, and if too many ions are analysed they will dirty the mass spectrometer, which will ultimately lead to a loss in sensitivity and a cleaning will be required. Therefore, the injected sample amount must be limited, and thus analysing multiple samples together logically induce a reduction of individual sample amounts that are analysed compared to a dedicated analysis of each individual samples. For example, a protein that is present in one out of three samples will be diluted by the sample mixing, and may not be detected. Moreover, label-based approaches are rather expensive, and they are limited in multiplexing by the number of stable-isotope reagents. In this context, with the improvement of mass spectrometers reproducibility, label-free approaches have emerged with satisfying performances[13]. They require less complex sample preparation, are suited to all types of samples, and they are not limited in multiplexing. Typically, label free quantification is performed using DDA data, allowing both identification and quantification within a single analysis. Label free quantification can be performed either by spectral counting or by MS1 filtering coupled to the extraction of ion chromatograms (XIC)[127].

### A. *Spectral counting*

The spectral count approach is based on the assumption that the number of collected MS/MS spectra for a protein is proportional to its abundance[128].

The major advantage of this method is the data analysis simplicity. The number of collected MS/MS spectra is attributed to each peptide and protein, which will be used to perform relative quantification between samples. However, only peptides that were selected for fragmentation will be considered, and this method thus suffers from the undersampling of DDA, which creates missing values. Moreover, dynamic exclusion must not be used or very limited for spectral counting to allow the MS/MS spectra redundancy used for quantification, but it prevents the identification and quantification of low abundance proteins, and spectral counting is therefore very limited in sensitivity. In addition, to be able to confidently quantify a difference in protein amount, the number of collected MS/MS spectra should be high, and therefore spectral counting performs better for high abundance proteins[129].

## B. *MS1 filtering – extracted ion chromatogram*

The MS1 filtering – extracted ion chromatogram (XIC) approach (MS1 XIC), is based on the extraction of precursor ion chromatograms from MS spectra. The area under the curve is then used to attribute a quantification value to each peptide and protein[14]. This method requires high resolution / accurate mass (HR/AM) instruments to specifically extract isotopes of the precursor ions of interest (**Figure 12**).

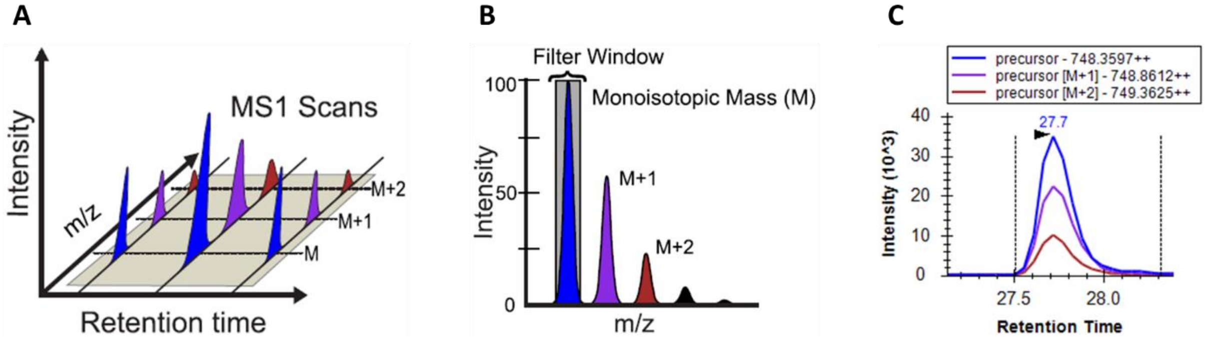

A
B
C
Intensity
MS1 Scans
m/z
M+2
M+1
M
Retention time
Filter Window
100
Monoisotopic Mass (M)
Intensity
50
M+1
M+2
0
m/z
precursor - 748.3597++
precursor [M+1] - 748.8612++
precursor [M+2] - 749.3625++
40
30
20
10
0
Intensity (10^3)
27.7
27.5
28.0
Retention Time

***Figure 12 : Principle of the MS1 filtering (adapted from [14]).***
*A. In DDA mode, an MS spectrum is collected at the beginning of each cycle. The isotopic envelope presents the three first isotopes at M, M+1 and M+2. B. The use of high resolution / accurate mass instruments allows the specific filtering of isotopes from MS spectra. C. The MS peak areas are used to build extracted ion chromatograms, usually for the three first isotopes of targeted precursor ions.*

Two approaches can be used for MS1 XIC approach: (i) extraction of all detected features, or (ii) targeted extraction of identified peptides.

### B.1.Extraction of all detected features

In this approach, all ions presenting peptide-like isotopic pattern are detected and called features. The chromatograms of the main isotopes are extracted for each feature, and their area under the curve are summed and used for quantification. Then, the features will be linked to their corresponding peptides, if they have been identified. The advantage of this approach is that it allows quantification of peptides that were not identified, for instance because they were not fragmented or they are not present in the protein sequence database. Several software tools allow features quantification, like Progenesis LC-MS (Nonlinear Dynamics), MaxQuant[55], or MFPaQ[130]. The Proline software tool, developed by the French proteomics infrastructure ProFI, is being developed to allow such feature detection and quantification.

### B.2. Targeted extraction of identified peptides

In this approach, before quantification, peptides must be identified. Then, a spectral library is built, containing the m/z and retention time at identification of each precursor ion. The spectral library is used to extract the precursor ions MS1 signals in a targeted manner. A drawback of this approach is that it is limited to the quantification of identified peptides, and then suffers from the limitations of DDA for peptide identification (e.g. undersampling, limited dynamic range, reproducibility). The most used software tool allowing targeted extraction of identified peptides is Skyline[52].

## V. Targeted proteomics

While shotgun proteomics allows the identification and quantification of a large number of proteins, data dependent acquisition (DDA) still suffers from limited sensitivity, reproducibility and dynamic range[15]. When a limited number of known proteins have to be detected in a large cohort of samples, targeted approaches are the best candidate. They allow the quantification of a predefined set of ≈ 50-100 known proteins in complex matrices with high sensitivity, specificity, dynamic range and reproducibility. Prior knowledge on the targeted peptides is necessary to build the acquisition method, as data will be collected in a targeted manner. Targeted approaches are often used for biomarker validation[16-23]. The gold standard approach for targeted proteomics is selected reaction monitoring (SRM) performed on a triple quadrupole mass spectrometer[48]. Recently, targeted approaches have also been developed for HR/AM instruments, e.g. parallel reaction monitoring (PRM) performed on a quadrupole-orbitrap mass spectrometer[24].

### A. *Selected reaction monitoring*

Selected reaction monitoring (SRM), also called multiple reaction monitoring (MRM) is the gold standard method for targeted proteomics[47]. It is performed on a triple quadrupole mass spectrometer, which is composed of a first quadrupole used for peptide ion selection, a second quadrupole used as a collision cell, and a third quadrupole used to sequentially isolate the fragment ions of interest. The targeted peptide-fragment couples, called transitions, must be defined in the acquisition method and each transition will be analysed sequentially (**Figure 13**).

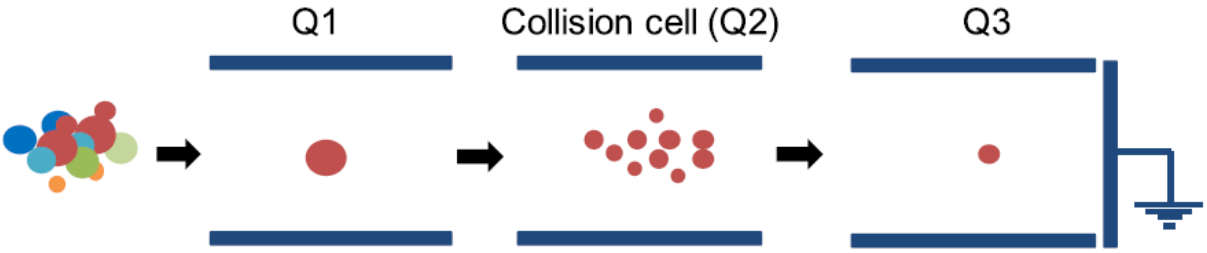


***Figure 13 : Principle of selected reaction monitoring (SRM) (adapted from [12]).***
*The first quadrupole (Q1) isolates a targeted peptide from a complex mixture of peptides. The isolated peptide is then fragmented by the collision cell (Q2) and the targeted fragments are isolated by the third quadrupole (Q3) and detected. This process is repeated for each targeted transition.*

In SRM, the use of fragment ion signals provides high sensitivity and specificity to the quantification, due to the double selectivity at both the peptide and the fragment levels. Moreover, the specificity of the quantification is further improved by the repeated detection of multiple transitions per peptide. The systematic detection of targeted transitions also provides high reproducibility to the assay[48].

### A.1.Method development

The development of a SRM method is a time and labor intensive work, and is divided into selection of targets using previous knowledge, and empiric optimisations to improve the multiplexing capacity and sensitivity of the method (**Figure 14**).

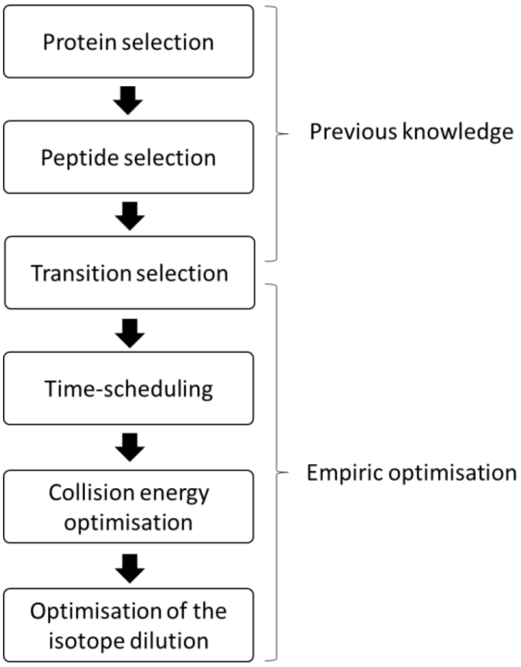

Protein selection
Peptide selection
Transition selection
Time-scheduling
Collision energy optimisation
Optimisation of the isotope dilution
Previous knowledge
Empiric optimisation

***Figure 14 : SRM method optimisation workflow.***
*Using previous knowledge, e.g. DDA data, the proteins, peptides and transitions to target are selected. The best responding transitions are then empirically validated, and a time-scheduled acquisition method is developed. The collision energies are optimised, and usually heavy peptides are used for quantification, for which the injected quantity should be adapted to the quantity of each light endogen peptide.*

### A.1.1. Selection of the targets

Targeted approaches are hypothesis-driven, as the targeted peptides must be chosen prior to the analysis. They should be specific to the targeted protein and not shared among multiple proteins to provide a quantification that is specific to the protein of interest. A peptide that is specific to a unique protein and detectable by MS is called a proteotypic peptide[131]. In the laboratory, we use to choose only fully tryptic peptides that were previously detected in DDA mode, from 7 to 25 amino acid residues, without amino acids prone to modifications like methionine which can be oxidised, without missed cleavages, and if possible peptides should be distributed in the whole protein sequence. After the peptide selection, the best responding transitions of each peptide are selected, if possible based on previous DDA data. The use of multiple transitions per peptide increases the specificity of the quantification and prevents issues if some of them are interfered in the samples.

The next steps of the SRM method development are usually performed empirically by analysing a representative sample and using crude stable isotope labelled peptides corresponding to the peptides of interest.

### A.1.2. Time scheduling

A major optimisation is the development of a time scheduled acquisition method, which greatly improves the multiplexing capacity and the sensitivity of the assay. Indeed, during an SRM analysis, hundreds of transitions are analysed, and the time spent to analyse a transition is called the dwell time. The longer the dwell time, the better signal / noise ratio, and the better sensitivity. Since all transitions are sequentially analysed within a cycle, the cycle time is defined as follows:

$$Cycle\ time = Number\ of\ transitions\ x\ (Dwell\ time + interscan\ time)$$

The interscan time is the time that the mass spectrometer needs to change the voltages to analyse another transition, which is ≈ 1 ms. Therefore, the more transitions, the longer the cycle time. However, 8 to 10 data points should be obtained across the chromatographic peak for a well-defined peak allowing a precise quantification. With average chromatographic peaks of 20 to 30 sec, a cycle should be ≈ 3 sec. The dwell time can vary from 5 to 100 ms, but in the laboratory we use to define a minimum of 20 ms dwell time (on a TSQ Vantage) for an acceptable sensitivity, resulting in a maximum of 150 transitions that can be monitored. In order to increase this number of transitions, they can be analysed only when their corresponding peptide elute out of the chromatographic column, using a time-scheduled acquisition method, also called scheduled SRM[48] (**Figure 15**).

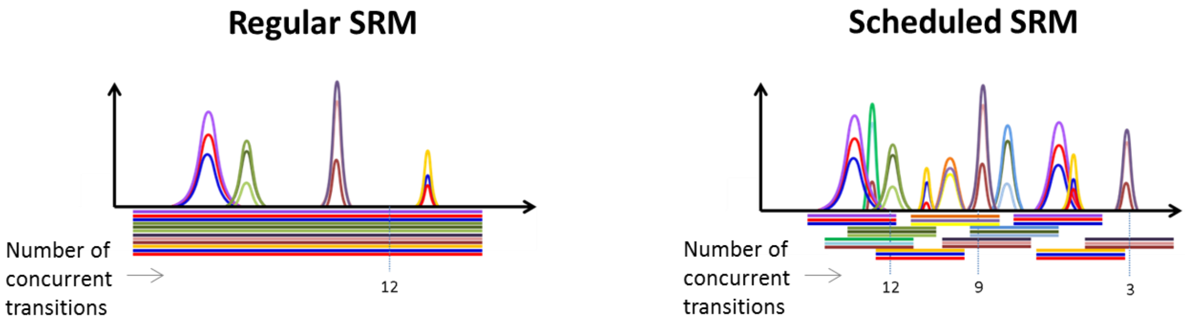

Regular SRM
Scheduled SRM
Number of concurrent transitions
12
Number of concurrent transitions
12
9
3

***Figure 15 : Principle of scheduled-SRM.***
*Using regular SRM, all transitions are monitored during the whole analysis, while using scheduled SRM they are monitored only when the peptides are eluting out from the column.*

The time-scheduling decreases the number of concurrent transitions over the analysis at a given time. According to the equation presented above and given that the cycle time is fixed, the time scheduling allows both an increase in multiplexing (more targeted transitions) and sensitivity (longer dwell time). The predicted retention time and the time window during which the transitions will be monitored must be predefined for each transition in the acquisition method. Typically, I used a time window of 4 min, but it can be reduced when a higher number of transitions must be monitored, or if a longer dwell time is desired, but then special attention should be allocated to liquid chromatography reproducibility and stability.

### A.1.3. Collision energy optimisation

The optimal collision energy (CE) to fragment a given peptide can be estimated using equations that allow the calculation of a CE using the m/z of the precursor ion and its charge state. However, these calculated CE are not optimal for all peptides, and since a limited number of peptides are targeted in SRM, their CE can be empirically optimised to enhance peptide fragmentation[132-133]. Moreover, the CE can be optimised for each transition, i.e. the CE leading to the most intense transition signal, what I did for SRM experiments presented in this manuscript (**Figure 16**).

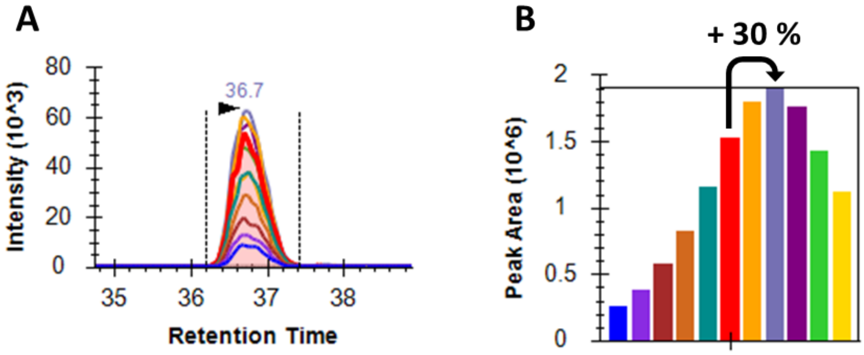


***Figure 16 : Collision energy optimisation for SRM method.***
*A range of collision energy are tested around the calculated value. A. The extracted ion chromatograms of a given transition obtained using different collision energies are presented. B. The transition peak area obtained using each collision energy is plotted. In this example, the collision energy optimisation allowed a gain of ≈ 30% sensitivity compared to the calculated collision energy (in red).*

### A.1.4. Isotope dilution

The accuracy and precision of the quantification can be improved by the use of isotope dilution. The most common approach consists in the addition of the same known amount of stable isotope labelled peptides into the samples, which corresponds to the targeted peptides. The so called heavy peptides will be analysed together with the endogen, or light peptides. The heavy peptides will be used as internal standards to normalise the quantification, reducing signal fluctuation due to technical biases. The ratio between the signals of the light and the heavy peptides will be used for the quantification (**Figure 17**).

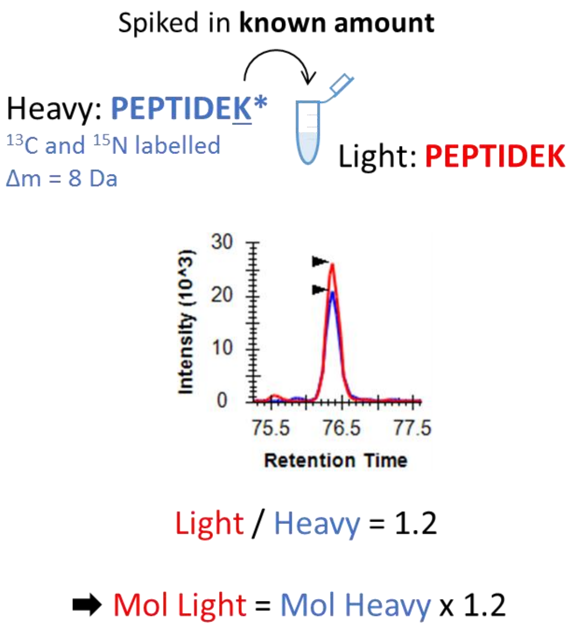


***Figure 17 : Principle of isotope dilution.***
*The heavy peptides are added into the samples in the same known amount, and the ratio between the light and heavy peptides signals are used for quantification.*

For bottom-up proteomics, the most used digestion enzyme is trypsin, and therefore a $^{13}C$ and $^{15}N$ labelling is usually performed on C-terminal Lysine or Arginine residues. The light peptide and its corresponding heavy version share the same physicochemical properties, i.e. same retention time, ionisation efficiency, fragmentation pattern, and differ only by their mass, allowing accurate quantification without bias. For an accurate quantification, each heavy peptide should be added at a ≈ 1/1 ratio compared to its light version. This also avoids too much ionisation competition between both versions of each peptide and maximise the sensitivity of the assay.

During my PhD, I used two types of stable isotope labelled peptides: (i) low quality crude synthetic peptides, which are not accurately quantified and inexpensive, were used for accurate relative quantification between samples, (ii) high quality synthetic peptides, which are highly purified and accurately quantified, were used for accurate and absolute quantification, like AQUA peptides[134], but they are expensive. Low quality crude synthetic peptides were used at the first steps of the targeted proteomics projects to screen for a large number of peptides, and AQUA peptides were bought for the best responding peptides.

The moment when the stable isotope labelled peptides are added will condition which steps of the workflow will be normalised: if they are added just before LC-MS/MS analysis, they will normalise only the LC-MS/MS fluctuations. Alternatively, several approaches use heavy labelled proteins that can be spiked earlier during the sample preparation, which are either (i) heavy labelled concatemer of the targeted peptides (quantification concatemer, QconCAT)[135], or (ii) full length stable isotopically labelled proteins (protein standard absolute quantification, PSAQ)[136]. However, these approaches are very expensive and require a long development time.

### A.2.Quantification

The most widely used software tool for SRM data analysis is Skyline[52], which extracts transitions chromatograms and allows exporting area under the curves for quantification. A thorough signal inspection can be performed using Skyline to detect interfered signals, LC issues or signal instability. The co-elution of all transitions of a given peptide in both heavy and light versions should be checked, as well as the relative intensities of the transitions which should be equivalent. Interfered transitions can be removed, and wrong peak picking can be manually curated.

The area under the curve of the transitions are summed for each peptide, and the ratio between the signals of the light and the heavy versions of the peptides are used for accurate quantification. If highly purified and accurately quantified heavy peptides were used, the absolute light peptide quantity can be calculated.

### A.3.Linearity range

When absolute quantification is performed, the linearity range and the lower limit of quantification (LLOQ) must be determined. The LLOQ is the lowest amount at which an analyte can be accurately and precisely quantified[137-138]. Usually, the linearity range and LLOQ are determined by the realisation of calibration curves using the stable isotope labelled peptides.

## B. *Parallel reaction monitoring*

Parallel reaction monitoring (PRM)[24] has been recently developed to (i) improve the selectivity of the assay, and (ii) ease the method development. It is a SRM-like method performed on a quadrupole-orbitrap mass spectrometer. The first quadrupole sequentially isolates the targeted peptides, which are fragmented in the collision cell and all fragments are analysed by the orbitrap (**Figure 18**).

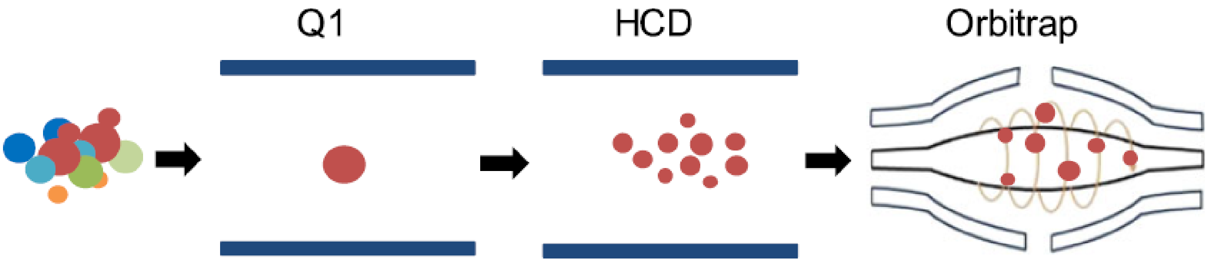


***Figure 18 : Principle of parallel reaction monitoring (adapted from[12]).***
*The first quadrupole (Q1) sequentially isolates the peptides of interest, which will be fragmented in the collision cell (HCD cell), and all fragments will be analysed by the orbitrap.*

When compared to SRM, the use of a HR/AM analyser in PRM improves the specificity of the assay because fragment ions are no more isolated by a quadrupole with a typical resolution of 0.7 Da, but they are extracted during data analysis from MS/MS spectra with high selectivity, commonly ≈ 50 ppm. This better specificity, by allowing interference removal, could also lead to an increased sensitivity. Moreover, the acquisition of complete MS/MS spectra for targeted peptides allows targets refinement during data analysis, which is a clear advantage over SRM for which a limited number of transitions are targeted for each peptide, if several transitions are interfered. Moreover, not needing to define a list of targeted transitions for each peptide also leads to an easier method development, because the best responding transitions do not need to be chosen prior to data acquisition.

The method development workflow is similar to the one presented for SRM (see A.1), except that the best responding transitions do not need to be selected for PRM as they are all analysed. However, several additional parameters should be considered when using a Q Exactive Plus: (i) using an orbitrap, the more acquisition time, the more resolving power, and the best compromise between resolving power and acquisition speed should be found, and (ii) the multiplexing mode should be chosen: the *simplex* mode is used to analyse precursor ions one at a time, the *broadband* mode is used to analyse all isotopes of precursor ions, the *duplex* mode is used to analyse 2 co-eluting precursor ions, and the *multiplex* mode is used to analyse up to 10 co-eluting ions[139].

Equivalent approaches were developed on quadrupole-time of flight mass spectrometers, like multiple reaction monitoring in high resolution (MRM HR)[25-26, 140]. An extensive comparison between several targeted platforms, including SRM and PRM operated on different LC and MS systems, will be presented in Chapter I.

## **VI.** Data independent acquisition

On the one hand, shotgun approaches using data dependent acquisition (DDA) allow large scale protein quantification, but suffer from low sensitivity, specificity, dynamic range and reproducibility. On the other hand, targeted approaches using selected reaction monitoring (SRM) or parallel reaction monitoring (PRM) offer high sensitivity, specificity, dynamic range and reproducibility, but are limited in the number of targeted proteins, and the method development is labor intensive. Recently, improvements in scan rates and high resolution allowed the emergence of data independent acquisition (DIA) approaches, which promises to combine the advantages of shotgun and targeted approaches, allowing global quantification with sensitivity, specificity, dynamic range and reproducibility that are comparable to those of targeted approaches. However, the major bottleneck of DIA approaches is the data analysis[141]. Though DIA is still in development, it starts to be used for a variety of applications like biomarker discovery and validation[142-143], and attracts a growing interest from the scientific community (**Figure 19**).

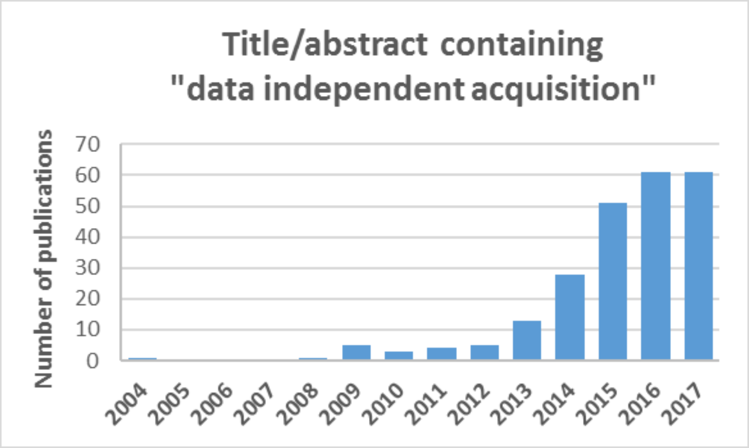


***Figure 19 : The growing interest in data independent acquisition.***
*The number of publications containing "data independent acquisition" in their title or abstract was extracted from PubMed on the 19th of September 2017.*

### A. *Principle*

In data independent acquisition (DIA) mode, MS/MS data are collected all along the analysis for the whole analysed m/z range, independently of any MS data. DIA is performed on quadrupole-time-of-flight or quadrupole-orbitrap mass spectrometers. Precursor ions are sequentially isolated in predefined m/z windows, fragmented together, and all fragments are analysed by the HR/AM analyser. During a cycle, highly multiplexed MS/MS spectra, containing fragments of all co-eluted and co-isolated precursor ions, are collected for the whole m/z range.

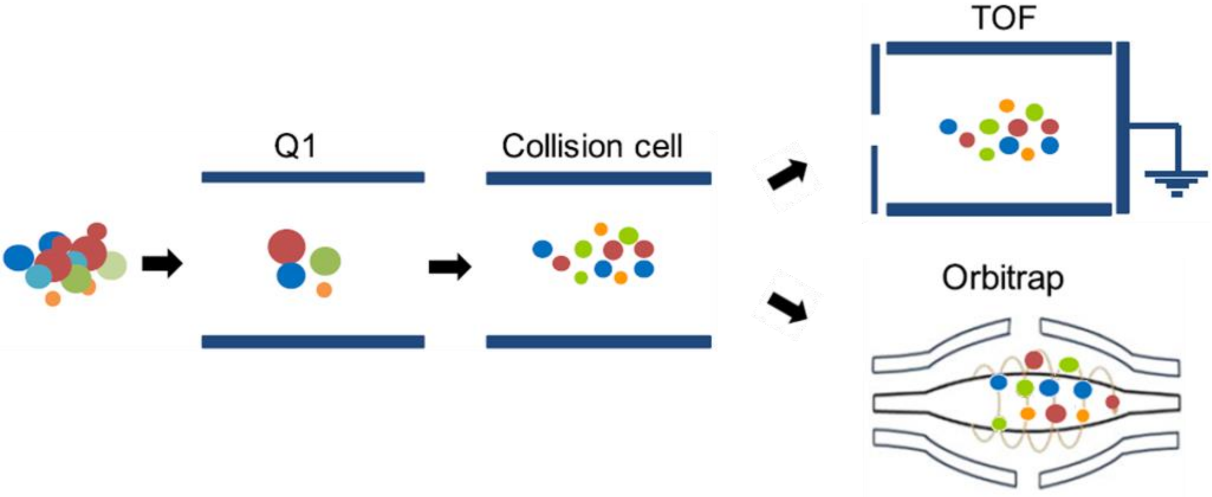


*Figure 20 : Principle of data independent acquisition (adapted from [12]).*
*Precursor ions are sequentially isolated by the first quadrupole (Q1) within predefined large m/z windows, co-isolated precursor ions are fragmented together in the collision cell and all fragments are analysed by the time-of-flight (TOF) or orbitrap analyser. This process is repeated to cover the whole m/z range within a cycle.*

In DIA approaches, MS/MS data are collected for all peptides and during the whole analysis time, and thus they do not suffer from the undersampling of DDA nor the limited number of targets of targeted approaches. The coverage of DIA is only limited by the limit of detection of the instrument[144]. Moreover, the use of MS/MS signals provide sensitivity, specificity, dynamic range equivalent to those of targeted approaches[145]. Finally, the systematic fragmentation of all peptides throughout the analysis provide an ideal reproducibility.

### B. *Data acquisition methods*

In 2003, Purvine *et al.* published the proof-of-principle of DIA. Using a liquid chromatography-time-of-flight coupling, they performed in-source co-fragmentation of multiple peptides, then called shotgun-CID, and used the extracted ion chromatograms information to reconstruct peptide-fragment lineage and identify the peptides[146]. From then, two types of DIA methods were developed, consisting in (i) the fragmentation of the whole m/z range (broadband DIA), or (ii) the sequential fragmentation of the m/z range within predefined m/z windows[141].

The development of DIA approaches collecting MS and MS/MS spectra for the whole m/z range started in 2005, with the introduction of the $MS^E$ methodology by Waters, which is performed on a quadrupole-time-of-flight mass spectrometer. Alternatively to ion-source fragmentation, the peptide ions are fragmented in the second quadrupole which acts as a collision cell. In $MS^E$, low and high collision energy are alternated to collect MS and MS/MS data for the whole m/z range[147]. In 2010, Thermo Scientific developed a similar approach for the analysis of small molecules, called all-ion fragmentation performed on a linear ion trap-orbitrap[148].

Rather than fragmenting peptides from the whole m/z range together, alternative methodologies were developed in which peptides are sequentially fragmented within predefined m/z windows to reduce interferences and improve the sensitivity[141]. In 2004, Venable *et al.* employed for the first time the term data independent acquisition to describe a method based on the sequential isolation and fragmentation of precursor ions within small windows of 10 m/z from 900 to 1 100 m/z, performed on a linear ion trap[149]. In 2009, Panchaud *et al.* introduced the precursor acquisition independent from ion count (PAcIFIC) approach, using 2.5 m/z isolation windows to further reduce interferences, but 67 analyses during 5 days were necessary to cover the same m/z range[150]. In 2011, the same team proposed an improved version of PAcIFIC, reducing the analysis time to ≈ 2 days using a faster ion trap[151]. In 2010, Carvalho *et al.* developed the extended data-independent acquisition (XDIA), which included a high resolution MS scan at the beginning of each cycle, and a combination of ETD and CID for peptide fragmentation. In 2012, Weisbrod *et al.* developed the Fourier-transform all reaction monitoring (FT-ARM) using isolation windows of 12 or 100 m/z[152]. Also in 2012, Gillet *et al.* presented a similar method called sequential windowed acquisition of all theoretical fragment ion mass spectra (SWATH), which uses 26 m/z isolation windows and is performed on quadrupole-time-of-flight mass spectrometers[27]. This methodology is marketed by SCIEX. In 2013, Egertson *et al.* introduced the MSX strategy, consisting in dividing the 500 to 900 m/z range in 100 windows, which will be sequentially analysed by groups of 5 random windows[153].

Recently, four types of improvements were performed for DIA methods: (i) combination of $MS^E$ with ion mobility to improve precursor and fragment ion assignment (high definition $MS^E$ $HDMS^E$ [154], and ultra-definition $MS^E$ $UDMS^E$ [155]), (ii) use of isolation windows of variable sizes over the m/z range to reduce the number of co-isolated precursor ions (SWATH 2.0), (iii) use of parallelisation capacity of the Orbitrap Fusion (Thermo Scientific) to allow both quantification with MS data analysed by the orbitrap, and identification with MS/MS data analysed by the linear ion trap (wide selected-ion monitoring WiSIM-DIA[156], and pSMART-DIA[157]), and (iv) use of a quadrupole that continuously scans a 10 to 35 m/z window moving over the m/z range (SONAR[158]).

### C. *Data analysis*

The data analysis is today the major bottleneck of DIA approaches. Indeed, DIA generates highly multiplexed MS/MS spectra composed of fragments of multiple co-isolated peptides, rendering the classical protein database searching inefficient[159]. Alternative approaches have been developed, namely (i) the peptide-centric analysis, which searches for specific peptides into the DIA data using a spectral library[27] and (ii) the spectrum-centric analysis, which creates pseudo-DDA spectra prior to classic protein database search[28] (**Figure 21**).

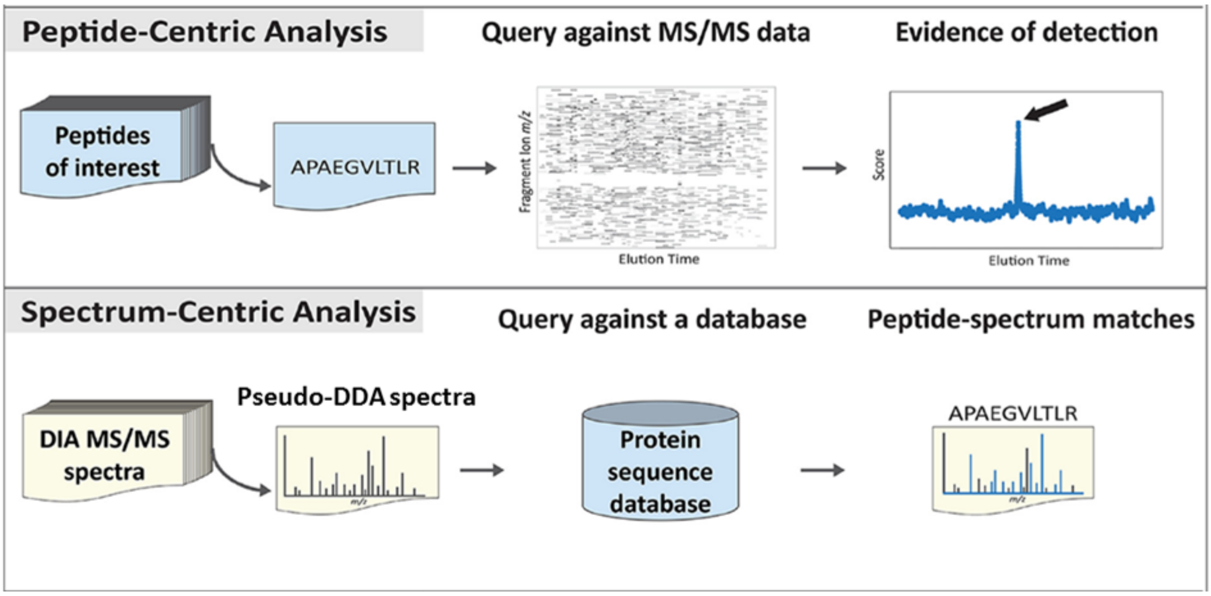

Peptide-Centric Analysis
Query against MS/MS data
Evidence of detection
Peptides of interest
APAEGVLTLR
Fragment ion m/z
Elution Time
Score
Elution Time
Spectrum-Centric Analysis
Query against a database
Peptide-spectrum matches
DIA MS/MS spectra
Pseudo-DDA spectra
m/z
Protein sequence database
APAEGVLTLR
m/z

***Figure 21 : Peptide-centric and spectrum-centric analyses (adapted from [159]).***
*In the peptide-centric analysis, the peptides of interest are queried against the MS/MS data, and extraction of fragment ions chromatograms allows the peptide identification. In the spectrum-centric analysis, pseudo-DDA spectra are queried against a protein sequence database using a classic protein database search.*

### C.1. Peptide-centric analysis

This approach was proposed by Gillet *et al.* in 2012, and was initially applied for SWATH data analysis and named targeted data extraction[27]. It is today the most used approach for DIA data analysis. Several software tools allow peptide-centric data analysis like Peakview (SCIEX), Skyline[160], OpenSWATH[161], or Spectronaut[162].

The peptide-centric approach relies on the use of a spectral library to look for peptides of interest into DIA data. Built from DDA data, the spectral library contains a list of previously identified peptides, and the information that is necessary to extract the fragment ion chromatograms corresponding to these peptides: peptides and fragments m/z, detected retention time, and relative intensity between the fragments. After extraction, each peak is scored according to several quality attributes, including its retention time, relative fragment ion intensities, fragments co-elution or m/z accuracy. Then, a target decoy approach is usually used to validate peptide identification[54].

The spectral library is usually built from previous DDA data, and therefore DIA indirectly suffers from DDA undersampling when using the peptide-centric approach. However, an extensive fractionation of the samples used to build the spectral library can increase the coverage of the spectral library[163]. In addition, retention time standards should be used to allow retention time alignment between the spectral library and the DIA data[164]. Alternatively to a homemade spectral library, several free-of-access spectral libraries are available for some reference taxonomies like human[53, 165] or yeast[166].

### C.2. Spectrum-centric analysis

This approach is based on the generation of pseudo-DDA spectra from co-eluting precursor and fragment ions. These spectra are then queried against a protein sequence database using the classical approach (described in III).

The spectrum-centric approach was first used for DIA data in 2003, when Purvine *et al.* used the co-elution characteristic of peptide and fragment ions to manually create pseudo-DDA spectra[146]. From then, several software tools were developed like DIA-Umpire[28, 167], which allows direct identification from DIA data. However, the number of false positives is still higher when compared to several library-based tools[29].

A comprehensive description and optimisation of a whole DIA workflow using SWATH acquisition and peptide-centric data analysis will be presented in Part IIChapter III.

A summary of the advantages and drawbacks of the different bottom-up proteomics approaches is presented in **Table 2**.

*Table 2 : Summary of bottom-up proteomics approaches.*

| | Method | Advantages | Drawbacks |
|---|---|---|---|
| **Shotgun proteomics** | DDA | High coverage | Low sensitivity<br>Low specificity<br>Low dynamic range<br>Low reproducibility |
| **Targeted proteomics** | SRM, PRM | High sensitivity<br>High specificity<br>High dynamic range<br>High reproducibility | Low coverage (50-100 proteins)<br>Labor intensive method development |
| **Data independent acquisition** | SWATH, $MS^E$, … | Very high coverage<br>High sensitivity<br>High specificity<br>High dynamic range<br>High reproducibility | Data analysis |

# **Chapter II** Monoclonal antibodies

Monoclonal antibodies (mAbs) are attractive for human therapy because they are highly specific and less toxic compared to conventional small molecules. Since the commercialisation of the first therapeutic mAb in 1986, Orthoclone OKT3, for prevention of kidney transplant rejection, the mAb and derived products class has significantly grown to become the dominant product class within the biopharmaceutical market[30]. Today, more than 70 mAbs and related products have been approved by the Food and Drug Administration (FDA) and the European Medicines Agency (EMA), and more than 50 mAbs are under evaluation in late-stage clinical studies[31]. While they are used for the treatment of a wide variety of diseases, the majority of approved mAbs are indicated for autoimmune disorders and cancers[32]. Their mode of action ranges from various natural functions of antibodies (neutralisation, antibody-dependent cell-mediated cytotoxicity (ADCC) or complement-dependent cytotoxicity (CDC)) to drug delivering[33]. Global sales for all therapeutic mAbs represented $107 billion in 2016, and are estimated at $145 billion in 2020[34].

## I. Expression systems

In 1975, Georges Köhler and César Milstein developed the hybridoma technology to continuously produce monoclonal antibodies (mAbs) specific to an antigen of interest, for which they obtained the Nobel Prize in Physiology or Medicine in 1984. The method relies on the injection into a mouse of an antigen to induce a specific immune response. The B lymphocytes of interest are isolated and fused with myeloma cells to produce a hybrid cell called a hybridoma, which combines the ability to secrete a specific antibody from the B cells with immortality from the myeloma cells[168]. However, hybridoma cells are genetically unstable and produce low mAb amounts, and more importantly the produced mAbs originating from the immunised animal may induce immune response in humans[169-170]. Advances in molecular biology and genetic manipulation techniques allowed the production of chimeric and then humanised or even human mAbs to reduce the immune response of patients against the mAb product. To improve their yield, they are produced in a variety of expression systems, ranging from bacteria, yeast, fungi, insect, mammalian cell lines to transgenic plants and animals[171]. However, the wide majority of currently licensed mAbs are produced in mammalian host cells due to their ability to introduce post translational modifications similar to those in humans[172]. Today, the Chinese hamster ovary (CHO) cell line is the most widely used for recombinant mAb production, because (i) CHO cells are robust and versatile cells which can be easily adapted to growth in serum free suspension conditions for large scale culture in bioreactors, (ii) powerful gene amplification systems are available for CHO cells to increase their productivity, and (iii) they have been demonstrated as a safe host, thus facilitating approval from regulatory agencies[35-36].

## II. Manufacturing process

The mAb manufacturing process is divided into the upstream process (USP) consisting in the production and the harvest of the mAb, and the downstream process (DSP) during which the mAb is purified and formulated (**Figure 22**).

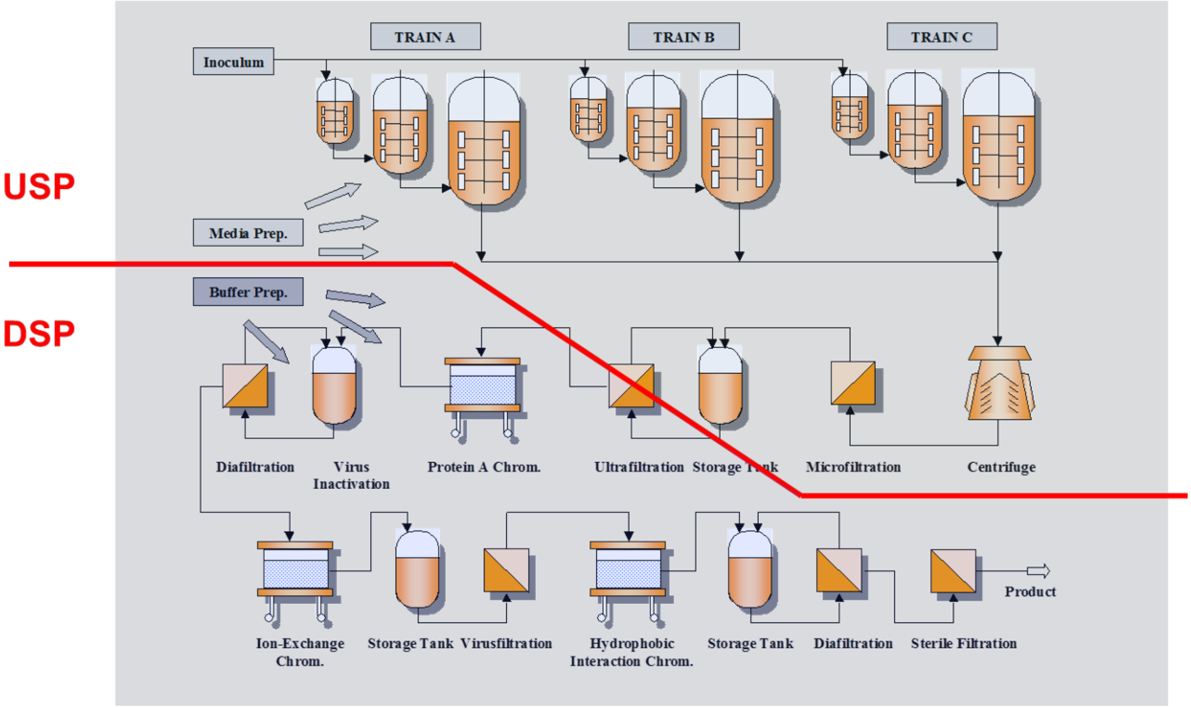

TRAIN A
TRAIN B
TRAIN C
Inoculum
USP
Media Prep.
DSP
Buffer Prep.
Diafiltration
Virus Inactivation
Protein A Chrom.
Ultrafiltration
Storage Tank
Microfiltration
Centrifuge
Ion-Exchange Chrom.
Storage Tank
Virusfiltration
Hydrophobic Interaction Chrom.
Storage Tank
Diafiltration
Sterile Filtration
Product

*Figure 22 : Overview of a generic mAb manufacturing process using mammalian cells (adapted from [173]).*

### A. Upstream process

Once the cell clone is selected (stable, robust and with high yield), it is expanded and several hundreds of cell vials are stored at -180°C in several locations to constitute a cell stock to be used only if necessary, named the Master Cell Bank (MCB). The MCB is extensively characterised regarding its identity, purity and stability. Cells from one vial of the MCB will be grown for several passages and again several hundreds of aliquots will be stored at -180°C to constitute the Working Cell Bank (WCB). Cells from the WCB are used for production.

The first step of the mAb manufacturing process is the thawing of a vial of cells from the WCB, which are first grown in a small volume (50 mL) and then expanded to reach the volume of a production bioreactor (500 to 20 000 L). This expansion phase is followed by a production phase during which the mAb is secreted by mammalian cells, and it accumulates in the culture medium until typical titers of ≈ 1 g/L in batch and 1-10 g/L in fed-batch processes[174]. Then, the cell culture fluid (CCF) is centrifuged and filtered (microfiltration) to remove the cells and cellular debris[175]. The resulting clarified cell culture fluid (CCCF) constitutes the last step of the USP.

### B. *Downstream process*

The DSP aims at releasing a pure mAb solution, concentrated into a solvent ensuring its stability, safety and therapeutic efficacy. The purification is realised by successive chromatography and filtration steps to remove process-related impurities like nucleic acids, lipids, host cell proteins (HCP) and product-related impurities[176].

The majority of mAbs DSP starts with a protein A affinity chromatography step (also called capture step), which removes the majority of impurities from the crude harvest material in a single step[177]. Protein A was originally found in the cell wall of the bacteria *Staphylococcus aureus*, and its natural high affinity to the Fc region of immunoglobulin G (IgG) from various species was first described in 1958[178]. Since then, engineered versions of the protein A present increased stability and binding capacity[179].

Then, the mAb undergoes up to three chromatography steps (also called polishing steps), virus clearance (inactivation and filtration) and filtration steps to concentrate the product (ultrafiltration) or remove buffer components (diafiltration)[173, 180]. The final filtration step aims to concentrate the mAb product into a buffer to allow its formulation (i.e. the addition of excipient) and conditioning (e.g. lyophilisation).

Impurities, in particular HCPs and DNA, must be monitored throughout the process[37], and typical purity targets are < 100 ppm for HCP (< 100 ng HCP / mg mAb), and < 10 ng/dose for DNA[176].

## **III.** Host cell protein monitoring

HCP constitute a major class of impurities that must be monitored and efficiently removed by the purification process. Remaining HCP in the final drug product can reduce the drug efficacy[38-40] or induce immune reactions when injected into patients[41-42]. HCP detection is particularly challenging due to (i) trace levels of HCP present in large excess of mAb product, (ii) large number of HCP that must be quantified and (iii) HCP population may change during process development[43]. A range of methods for the detection and characterisation of HCP are available, which can be classified as either immuno-specific methods like Western blot and ELISA, or non-specific methods like electrophoresis and mass spectrometry[44-45].

### A. *Immuno-specific methods*

These methods detect HCP using polyclonal anti-HCP antibodies, which are usually raised in goats or rabbits by repeated injections of HCP mixtures. The choice of these HCP mixtures is crucial as it will determine the spectrum of HCP that will be detected by the anti-HCP antibodies. To avoid the generation of anti-mAb product antibodies, HCP mixtures are typically generated using a null version of the host cell line, i.e. a mock transfected cell line. The assumption in this approach is that the HCP profiles of both the null cell line and the mAb producing version of this cell line are similar[43-45, 58].

Commercially available anti-HCP antibodies are usually raised using cell lysates or culture supernatant of several null cell strains. Assays employing these antibodies are called generic assays, because they are able to detect a broad spectrum of HCP from various cell strains and process conditions. They are easy and fast to implement, but their low specificity becomes problematic when few HCP must be detected in purified samples, and actually their coverage remains very low even in crude samples (≈ 30%)[45, 181]. To overcome this issue, process-specific anti-HCP antibodies can be generated in-house using material that is specific to the cell line used (upstream process specific), or to the manufacturing process (downstream process specific). Process-specific antibodies are usually raised using partially purified material, leading to an increased sensitivity for remaining HCP throughout the DSP, but such antibodies may be blind to changes in the HCP profile caused by a modification of the manufacturing process.

In conclusion, the major limitation of immuno-specific methods is that no anti-HCP antibody reagent can cover the entire spectrum of HCP that may be present, and it will only detect HCP which elicited immune reaction in animals that were used to generate the anti-HCP antibodies. Furthermore, developing a process-specific immunoassay is costly and time consuming[56], and anti-HCP antibodies are a limited reagent that will need to be reproduced.

### A.1.ELISA

Enzyme-linked immunosorbent assay (ELISA) is often used as a diagnostic tool in the fields of medicine and biotechnology[182-184]. It combines the specificity of antibodies with the sensitivity of assay enzymes to provide a measurement of the targeted protein concentration. It is today the gold standard method for HCP monitoring during process development, manufacturing and in final product formulations due to its high throughput, sensitivity and specificity[43-45].

Several types of ELISA exist, but the most used for HCP detection is the sandwich ELISA, so called because the targeted antigens are detected between two antibodies (**Figure 23**).

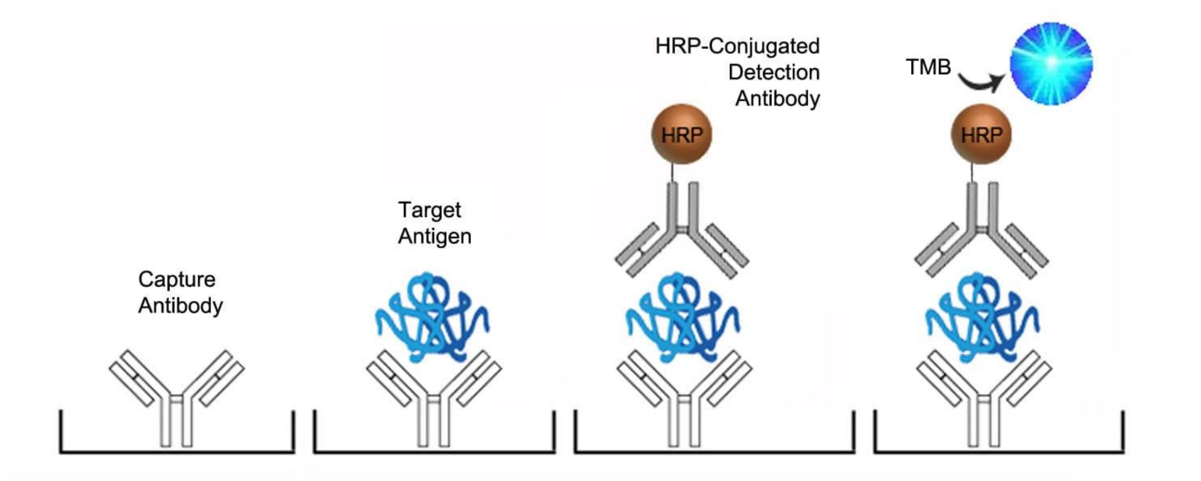


*Figure 23 : Principle of a sandwich ELISA assay (from www.mybiosource.com).*

Briefly, the first step is to coat capture antibodies which are specific to the targeted protein to a multiwell plate, and after incubation, wash out the unbound antibodies. The remaining protein binding sites are then blocked by incubation with for instance bovine serum albumin (BSA) or non-fat dry milk, to prevent subsequent nonspecific binding onto the wells. After washing, the samples are added and the targeted protein will bind the immobilised capture antibodies. After incubation, unbound target proteins are washed out. Detection antibodies are conjugated to an enzyme (e.g. horseradish peroxidase (HRP)), and are specific to another epitope of the targeted protein to allow simultaneous binding of both the capture and detection antibodies to the targeted protein, which is necessary to detect the protein of interest. After addition and incubation with detection antibodies, unbound antibodies are washed out. Finally, the substrate (e.g. TMB) is added and converted by the enzyme, and the product is quantified by measuring its absorbance using a spectrophotometer. The concentration of the antigen is calculated using a standard curve realised with standard samples of known concentration.

Contrary to conventional ELISA which quantifies a single antigen, ELISA for HCP aims to quantify a large number of proteins. Generally, the same anti-HCP antibodies are used for both the binding and the detection of HCP. However, the binding antibodies are either directly coated to the well, or conjugated to biotin to enhance their binding onto plates coated with streptavidin and improve the sensitivity of the assay[185-186]. On the other hand, the detection antibodies are conjugated to an enzyme, most commonly HRP[187-188]. For an effective detection of a given HCP by sandwich ELISA, at least two antibodies must be raised against this HCP, and a simultaneous binding of these two antibodies to the HCP must be sterically possible. Otherwise, the HCP will not be detected (**Figure 24**).

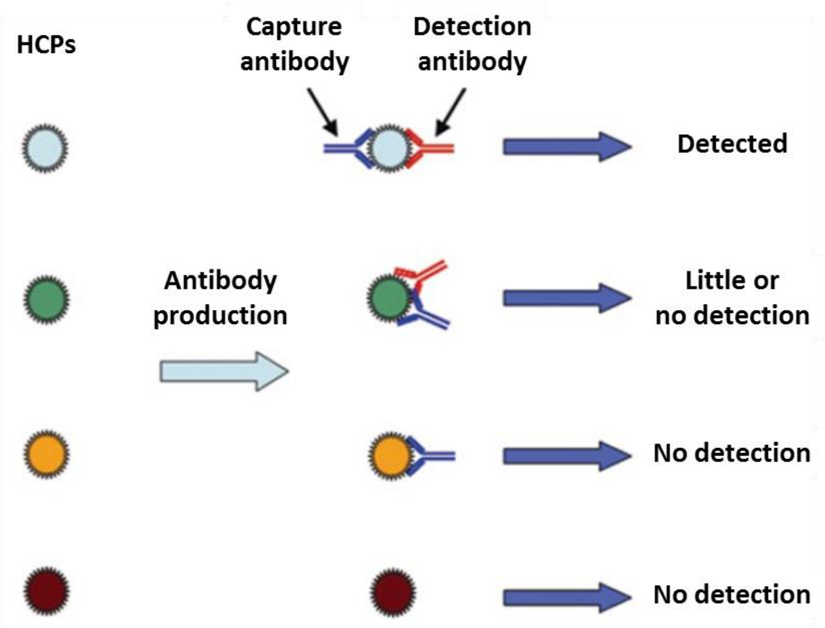


*Figure 24 : Possible outcomes for HCP detection by sandwich ELISA (adapted from [56]).*

Moreover, an increasing number of evidences show that ELISA does not provide comprehensive HCP quantification due to the use of anti-HCP antibodies[43-45, 56-58, 181]. Moreover, HCP quantification by ELISA produces only a total HCP amount without any information about the identity of the detected HCP, rendering a risk-based assessment of HCP very challenging.

### A.2. Western blot

Western blot is used in routine in many fields of scientific research, such as biology and biomedical sciences, to detect specific proteins from a complex sample[189-190]. The technique is divided into (i) separation by size using gel electrophoresis, (ii) transfer to a membrane (the proper western blot step), and (iii) detection of the targeted proteins using specific antibodies (**Figure 25**).

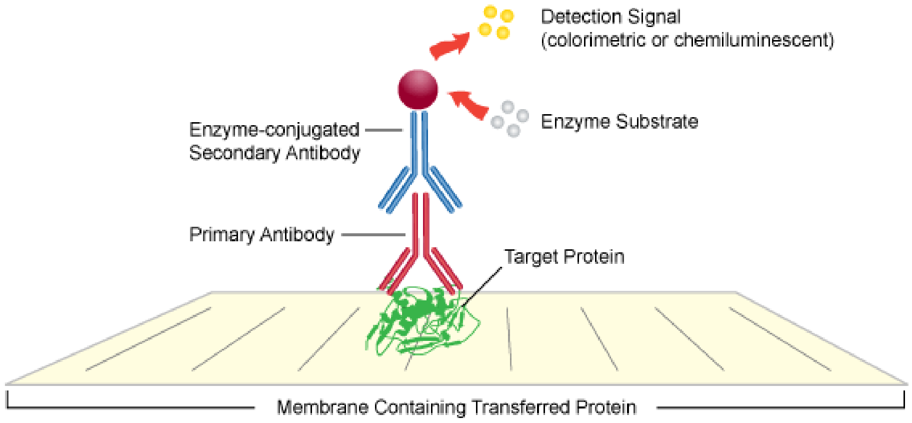
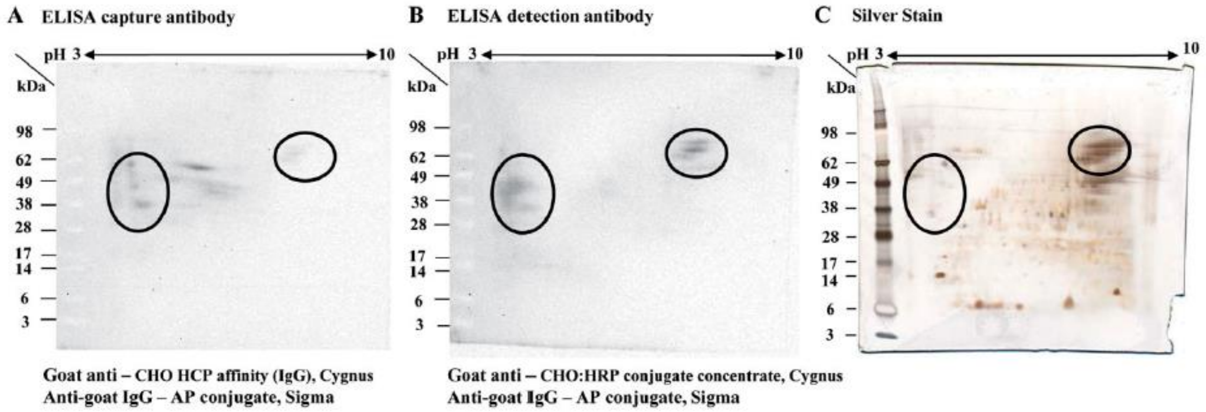


*Figure 25 : Principle of protein detection by Western blot (adapted from www.leinco.com).*

Briefly, the membrane is incubated with a blocking solution (e.g. BSA or non-fat dry milk) to prevent nonspecific antibody binding onto the membrane. After washing, the membrane is incubated with primary antibodies that are specific to the targeted protein. After incubation, unbound primary antibodies are washed. Secondary antibodies, which are conjugated to an enzyme, specifically target the primary antibodies. After incubation and washing, the substrate is added and the enzyme produces a compound which is detected using a spectrophotometer.

In the field of HCP, similarly to ELISA, western blot is used to detect a large number of HCP. However, while ELISA aims to quantify HCP, western blot is usually used to support ELISA development by evaluating the coverage of the anti-HCP antibodies. Typically, throughout the immunisations of animals with HCP mixtures to generate anti-HCP antibodies, the sera of the animals are collected at different steps and characterised to control the immunisation process and follow the anti-HCP antibodies production by the animals. This allows adaptation of the immunisation protocol, for instance an increase in the injected amount of HCP into animals to boost their immune response, or the injection of partially purified samples to enhance immune reactions against a subset of process challenging HCP. Typically, the method used to follow the generation of anti-HCP antibodies is based on the comparison between the global HCP profile, visualised using 2D-gels followed by a global staining like silver staining, and the fraction of the HCP population that is detected by the anti-HCP antibodies after western blot. This comparison can be performed for samples from different steps of the manufacturing process to evaluate the relevance of the anti-HCP antibodies for purified samples. An evaluation of capture and detection anti-HCP antibodies provided in a commercial ELISA kit is presented in **Figure 26**.



Goat anti – CHO HCP affinity (IgG), Cygnus
Anti-goat IgG – AP conjugate, Sigma

Goat anti – CHO:HRP conjugate concentrate, Cygnus
Anti-goat IgG – AP conjugate, Sigma

***Figure 26 : Evaluation of the HCP coverage of commercially available anti-HCP antibodies (adapted from [45]).***
*2D-gels were realised using CHO cell culture supernatant, followed by western blot and HCP detection using (A) capture and (B) detection anti-HCP antibody reagents from a commercial ELISA kit. The western blots were compared to (C) the global HCP profile detected by silver staining. Black circles highlight HCP that are recognised by both capture and detection antibody reagents, and which can therefore be effectively detected by the commercial ELISA kit.*

This example shows that the conjugation process affects the affinity of certain antibodies, as the detected HCP population differs slightly between the capture and detection antibodies. Moreover, the HCP coverage of this generic ELISA kit was found very limited, because only the HCP detected by both the capture and the detection antibodies can be detected using this ELISA kit, which represents a limited fraction of the total HCP population revealed by silver staining. However, it should be noted that if an HCP is recognised by the anti-HCP antibodies in Western blot, this does not guarantee that it will be recognised in ELISA (see **Figure 24**). Inversely, anti-HCP antibodies that do not recognise a denatured HCP in Western blot could recognise its native form in ELISA.

### B. *Non-specific methods*

Orthogonal non-specific methods should be employed to detect non-immunogenic HCP and complement immuno-specific methods for a rigorous HCP monitoring.

### B.1. Gel electrophoresis

Polyacrylamide gel electrophoresis (PAGE) techniques are used to separate proteins according to their size. In denaturing conditions, proteins are linearised and therefore separated according to their molecular weight. Using two dimensional-PAGE (2D-PAGE), a first step called isoelectric focusing (IEF) is performed before separation of proteins according to their molecular weight: IEF consists in the migration of proteins according to their isoelectric point in a polyacrylamide gel strip containing an immobilised pH gradient[191]. After migration, proteins are fixed into the gel and can be stained by global dyes like Coomassie blue[192] or silver staining[193] to allow global protein profiling.

Today, 2D-PAGE and differential gel electrophoresis (2D-DIGE) are widely used for the monitoring of HCPs during process development[58, 194]. They have the advantages of being robust, and offering a visual mapping of the global HCP population along with their molecular weight and isoelectric point. They also allow a direct visualisation of several post translational modifications (PTMs)[191], like phosphorylation which is expected to affect one third of an eukaryotic proteome[195]. Moreover, 2D-PAGE techniques can be used in combination with mass spectrometry to identify proteins in specific gel spots[191, 195].

However, the most important drawback of 2D-PAGE is the limited dynamic range, and the displayed proteins represent only the most abundant portion of the proteome. It is also difficult to analyse very small or very large proteins, extremely acidic or basic proteins, or hydrophobic proteins[5, 191]. The limited dynamic range is particularly problematic for the study of HCP with overwhelming mAb heavy and light chains, which can also hide low abundance HCPs (**Figure 27**).

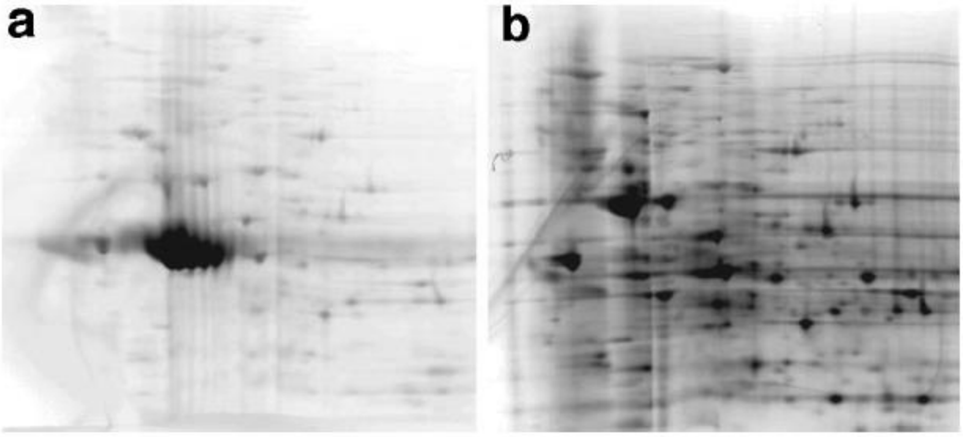


***Figure 27 : Limitations of 2D-PAGE method for HCP detection.***
*2D gels were realised for (a) a CCCF fraction containing the recombinant protein product and (b) for a CCCF fraction without the recombinant protein product which has been removed by affinity chromatography (from [58]).*

Another drawback of 2D-PAGE is its low throughput and labor-intensiveness, combined with its lack of reproducibility and the need for multiple gels to obtain reliable data[45, 191].

This issue can be overcome by 2D-DIGE in which up to five samples can be analysed on the same gel using different fluorescent dyes (cyanine dyes). However, the disadvantage of 2D-DIGE is the labelling which slightly alters the physical properties of the proteins such as their solubility, hydrophobicity and size[45].

### B.2. Mass spectrometry

Liquid chromatography coupled to mass spectrometry (LC-MS/MS) methods are the most promising orthogonal methods to the gold standard ELISA. They allow unbiased and individual HCP identification and quantification within a single analysis, enabling a more comprehensive risk assessment when taking into account the nature of the HCP, e.g. its proteolytic activity or immunogenicity[44, 196].

Moreover, recent advances in the MS field, notably the use of MS/MS signals for quantification by targeted or DIA methods allowed a 2- to 8-fold gain in sensitivity[27], and a significant gain in specificity and dynamic range when compared to the use of MS1 signals. These features are particularly crucial in the HCP field in which very low abundant proteins have to be quantified besides a highly abundant predominant protein.

The targeted approach using SRM coupled to isotope dilution has, for long, been the gold standard MS-based quantification technique offering highest sensitivity, accuracy and robustness[48]. However, targeted approaches are still limited in multiplexing to a few tens of proteins. Besides, DIA allows the collection of MS/MS information for all detectable species in order to extract valuable quantitative information from whole complex proteome maps[27]. For instance, two-dimensional liquid chromatography coupled to DIA-$MS^E$ has been used in a few studies to quantify HCP in mAb solutions[197-202]. However, a run time of more than ten hours is necessary for this type of analysis, which is incompatible with real time process support. Very recently, a 1D-LC DIA-SWATH method was shown to achieve equivalent sensitivity in only one hour[203].

The major limitations of this methodology are (i) the lack of high quality and publicly available CHO protein sequence database[46], and (ii) the requirement of a highly skilled operator and (iii) access to expensive equipment.

### C. *Host cell protein monitoring methods comparison*

Today, ELISA is the workhorse method for HCP monitoring during bioprocess development, manufacturing, and for product purity assessment because it is a highly sensitive and specific detection method. Its high throughput allows multiple samples to be analysed simultaneously in several hours, and it is accepted by the regulatory authorities. However, ELISA suffers from important drawbacks among which the most important is the limited coverage of the HCP population by the anti-HCP antibodies[45]. Ultimately, it can lead to the undetected presence of dangerous HCP in the final drug product, which can degrade the mAb product[38-40] or induce adverse immune reactions when injected into patients[41-42].

Therefore, there is a need for orthogonal methods to detect HCP without the bias linked to the use of antibodies, allowing a more comprehensive HCP coverage and a better characterisation of the detected HCP for a risk-based assessment of HCP. A summary of the available approaches for HCP monitoring with their detection limits and pros and cons is presented in **Table 3**.

***Table 3 : Summary of HCP monitoring methods.***

| | Method | Sensitivity | Advantages | Drawbacks |
|---|---|---|---|---|
| Immuno-specific | ELISA | Total HCP<br>1-100 ppm[45] | High throughput, sensitivity, specificity | Detects only immunogenic HCP<br>≥ 2 antibodies / HCP<br>Total HCP amount<br>No information about the HCP<br>Development costly & time-consuming |
| | Western blot | Individual HCP<br>20-200 ppm[45] | MW and pI<br>Visible PTM | Detects only immunogenic HCP<br>Development costly & time-consuming<br>Labor intensive |
| Non-specific | 2D-PAGE (cyanine dye) | Individual HCP<br>8 ppm[45] | MW and pI<br>Visible PTM<br>MS-compatible | Low dynamic range<br>HCP hidden by the mAb product<br>Labor intensive |
| | MS | Individual HCP<br>1-10 ppm[197, 203] | HCP identification and quantification<br>High sensitivity, specificity | No high quality CHO protein database<br>Highly skilled operator<br>Expensive equipment<br>Labor intensive |

The main objective of my PhD was to develop MS-based HCP monitoring approaches to support process development, manufacturing and final purity assessment. This work will be presented in Part IIChapter IV.

# Part II Results

My PhD work can be divided in two main parts: the first part consisted in developing and optimising analytical workflows for bottom-up proteomics, and the second, major part, was dedicated to the production of a wide range of monoclonal antibody (mAb) samples and the application of the acquired knowledge and optimised workflows to the comprehensive characterisation of host cell protein (HCP) impurities in these samples.

Analytical developments for bottom-up proteomics were performed for the three approaches of bottom-up proteomics, namely (i) shotgun proteomics, which is used for discovery projects, to characterise global protein contents, (ii) targeted proteomics, which is used to quantify a specific set of known proteins with optimal sensitivity and robustness, and (iii) data independent acquisition, a recent methodology which promises to combine the advantages of both shotgun and targeted approaches, but is today still not mastered due to its challenging data analysis.

**Chapter I:** A significant part of my work consisted in setting up a last generation microLC-Q-ToF coupling (microLC-Triple TOF 6600) for proteomics analysis. In this chapter, I describe the extensive optimisations that were performed to design high performing data dependent acquisition (DDA) methods for this new coupling.

**Chapter II:** Here I present the benchmarking work that I performed to compare four targeted proteomics workflows, including the gold standard selected reaction monitoring (SRM) performed on a triple quadrupole instrument (TSQ Vantage), and the more recently developed parallel reaction monitoring (PRM) performed on a Q-orbitrap instrument (Q-Exactive Plus) and equivalent multiple reaction monitoring in high resolution (MRM HR) performed on two Q-ToF instruments (Triple TOF 5600+ and Triple TOF 6600).

**Chapter III:** The main methodological focus of my PhD concerned the thorough evaluation of data independent acquisition (DIA) approaches on the Triple TOF instrument. DIA is a recent methodology which promises to combine the strengths of shotgun and targeted approaches, or even surpass them: in DIA mode, all peptides are fragmented during the whole analysis time. However, DIA is still not a common methodology mainly because of its major bottleneck which is DIA data analysis. In this chapter, I describe the deep optimisations that were performed during my PhD for each step of a DIA workflow, and more precisely DIA-SWATH (sequential windowed acquisition of all theoretical fragment ion mass spectra) methodology, including sample preparation, data acquisition and data analysis. Finally, a comparison between DIA and DDA for identification and quantification is presented.

The second part of my PhD, presented in **Chapter IV**, consisted in the application of upper optimised workflows to quantify host cell proteins (HCP) in monoclonal antibody (mAb) samples. This was the main concern of my PhD, for which I realised all steps, from the sample preparation performed during my 5 months stay at University College London to the quantification of HCP by MS at University of Strasbourg.

An overview of my PhD work is presented in **Figure 28**.

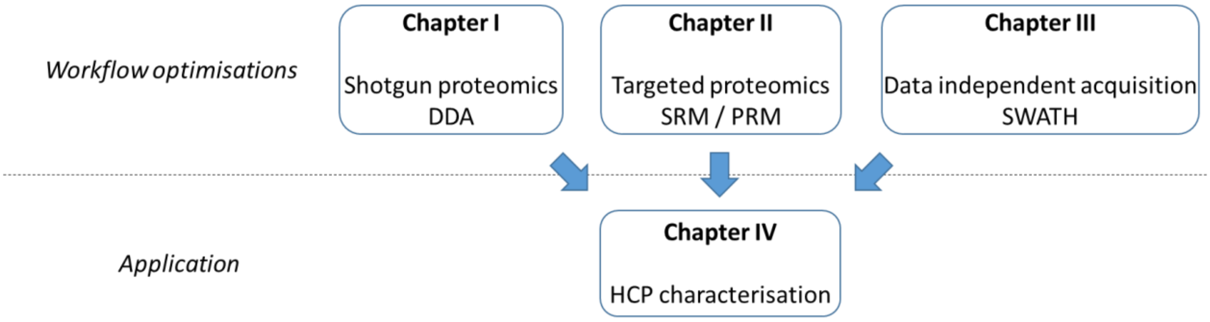

Workflow optimisations
Chapter I
Shotgun proteomics
DDA
Chapter II
Targeted proteomics
SRM / PRM
Chapter III
Data independent acquisition
SWATH
Application
Chapter IV
HCP characterisation

*Figure 28 : Overview of my PhD work.*

# **Chapter I** Optimisation of shotgun proteomics analysis

Today, data dependent acquisition (DDA) is the most used acquisition mode for discovery proteomics as it allows identification and quantification of peptides and proteins in a sample requiring only the protein sequences information[144].

In the laboratory, we acquired a microLC-Triple TOF 6600 coupling. A significant part of my PhD work consisted in setting up this coupling, and optimise DDA methods for both peptide identification and XIC MS1 quantification. To this end, we used a yeast digest as a representative sample which was analysed using a range of DDA methods. Peptides and proteins were identified using Mascot search engine and validated using the Proline software (1% FDR at both peptides and proteins levels). The objective was to develop the most sensitive DDA method for the best proteome coverage that could be used as the standard shotgun proteomics acquisition method on this coupling.

During a standard LC-MS/MS analysis on our microLC-Triple TOF 6600 coupling, the peptides are first fractionated by reverse phase chromatography using a **gradient** of organic solvents in acidic conditions, usually acetonitrile (ACN) with 0.1% formic acid. Eluted peptides reach the interface between LC and MS where the liquid phase nebulises at the end of a high voltage needle and the peptides are transferred to gas phase. An optimised **needle position** allows more charged peptides to enter the mass spectrometer while avoiding contamination by uncharged molecules. The source can be **heated** and several **gas** are used to help peptide desolvation. In DDA mode, the mass spectrometer performs a series of cycles. Each cycle starts with a survey MS scan followed by MS/MS scans. The survey MS scan is acquired when the Q1 is in RF-only mode, allowing all peptides, or precursor ions, to go through the TOF and reach the detector. Each ion that enters the detector creates a current that is converted into a voltage pulse which are summed among time bins, defining the **ToF resolution**, and during a defined **accumulation time** to build an MS spectrum. The **N most intense precursor ions (top N)** of this MS spectra above a defined **intensity threshold** are sequentially isolated within a defined m/z window by the Q1, or **Q1 resolution**, fragmented in the collision cell using an adapted collision energy and a defined **collision energy spread (CES)**, and MS/MS spectra of fragment ions are acquired. In order to increase the coverage of the assay, peptides for which MS/MS spectra were already collected can be **dynamically excluded**, based on their m/z (± **tolerance**) and for a defined **exclusion time**.

The parameters in bold will be detailed within this chapter, and their optimisation will be presented. They were optimised towards the best proteome coverage, which was probed with the number of peptides that were identified using each acquisition method. Indeed, what are rigorously identified by bottom-up proteomics are the peptides and not the proteins, and by using the peptides number we avoid any bias intrinsic to protein inference.

## I. Liquid chromatography

The liquid chromatography (LC) gradient has a direct and huge impact on the sensitivity of the assay, as it determines the sample complexity (number of co-eluting peptides) that will be analysed by the mass spectrometer over time. The longer the LC gradient, the more peptides will be identified, but the throughput of the method will also be reduced. The choice of the LC gradient duration must therefore be a compromise between the desired sensitivity and throughput. For these optimisations, we decided to use short LC gradients, from 5 to 40% ACN in 47 min for a 60 min total analysis time.

Beyond its duration, the LC gradient design will determine the peptides elution profile which should be equalised throughout the analysis to provide the simplest peptides solution to the MS over time. For this purpose, we evaluated five LC gradient designs: a linear gradient from 5 to 40% ACN in 47 min, and four gradients including a step at 36 min at 20%, 25%, 30% or 35% ACN. The gradients were evaluated by comparing the number of peptides that were identified (**Figure 29**).

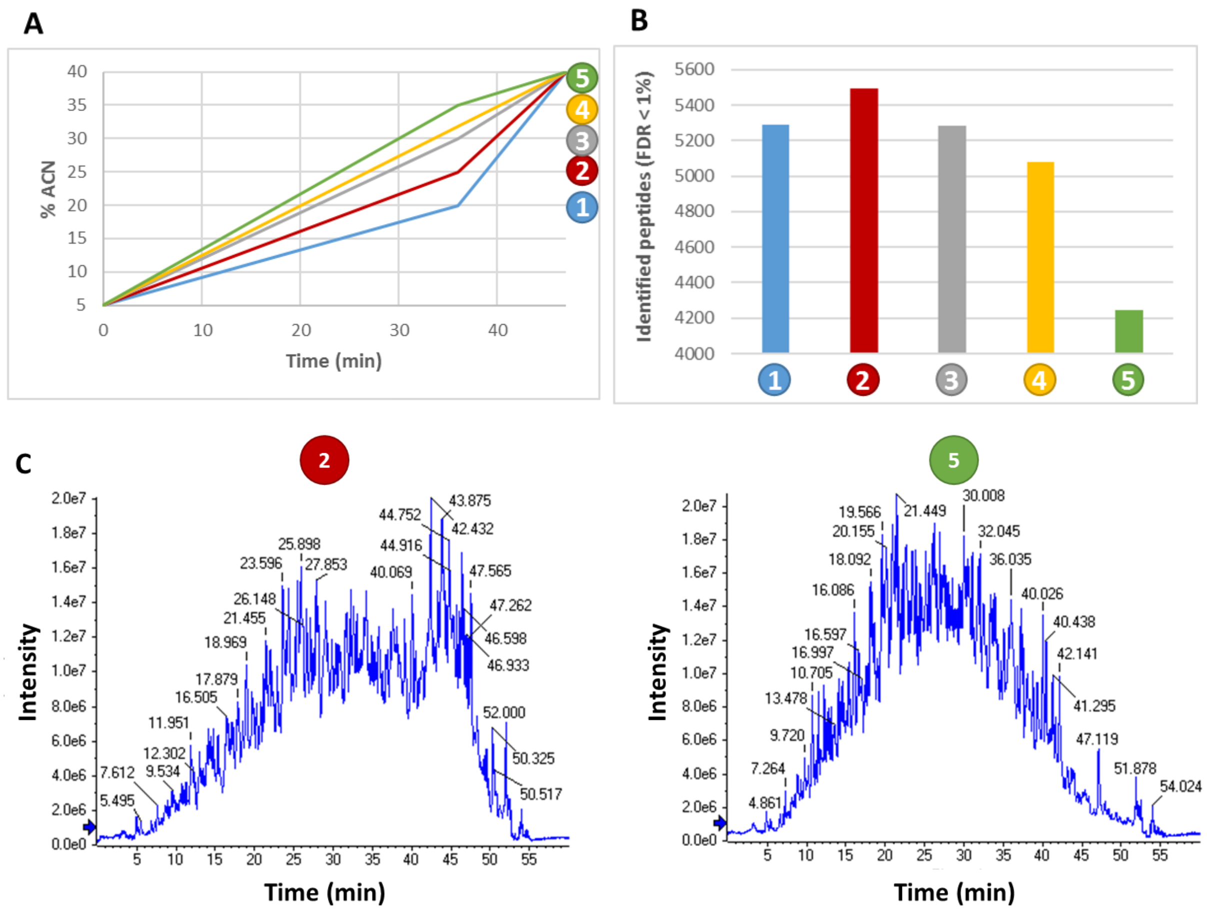


***Figure 29 : Optimisation of the LC gradient.***
*A. Gradients 1 to 5 are represented. B. The number of peptides identified using each LC gradient is displayed. C. A comparison of the total ion current of gradients 2 and 5 was realised.*

As expected, the LC gradient design is a major parameter to optimise since the tested gradients impacted the number of identifications by ≈ 29%. We found that the **gradient 2**, which includes a step at 36 min at 25% ACN, allowed the identification of 5 483 peptides, while other gradients allowed the identification of less peptides with a minimum for the gradient 5. Indeed, if we compare the global elution profile (i.e. the total ion current) between the gradients 2 and 5, we can see that the peptides are better split using the gradient 2. Using gradient 5, the maximum number of MS/MS is reached from ≈ 20 to 35 min and the instrument cannot collect MS/MS spectra for all candidate peptides.

## II. Interface

After LC separation, the role of the interface is to ionise and transfer the peptides from liquid to gas phase. Only ionised peptides will be analysed by the mass spectrometer, and therefore the ionisation efficiency directly impacts the sensitivity of the assay.

At the end of the LC part, the electrospray is formed at the tip of a needle, close to the mass spectrometer orifice. The needle position is crucial: if it is too far from the mass spectrometer orifice, less ions will enter the mass spectrometer, and if it is too close, a lot of uncharged species will dirty the first section of the mass spectrometer and a cleaning will be required. The needle position was optimised for the highest signal intensity without placing it too close from the orifice, and an optimal position was determined with the needle out about 2 mm and both callipers set at 5 mm.

On our microLC-Triple TOF 6600 coupling, we use the DuoSpray ion source. This source helps peptides desolvation with a coaxial gas (GS1), and a heater (TEM) coupled to the heater gas (GS2). We analysed the yeast digest using a range of gas supply pressure and source temperatures, from 10 to 20 psi for GS1 and GS2, and a source heating of 50 or 100°C (**Figure 30**).

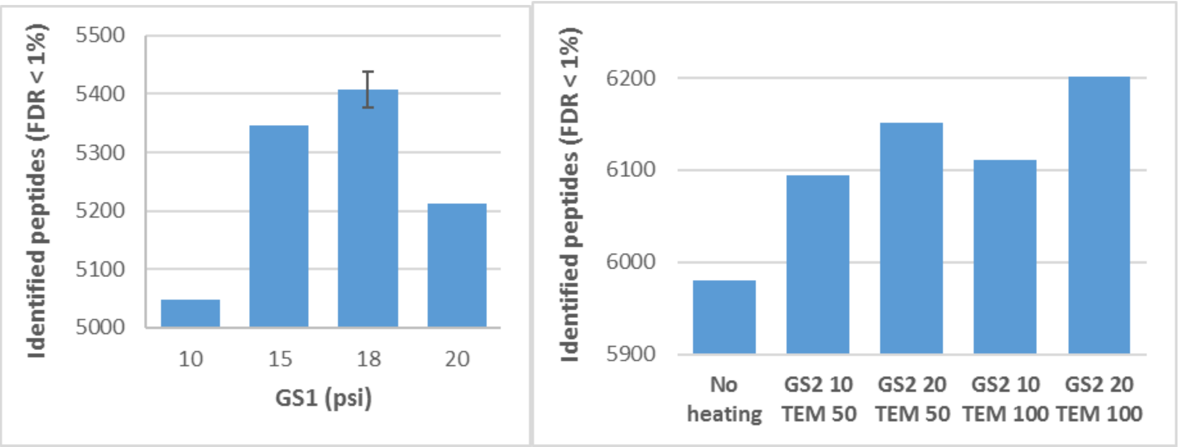


***Figure 30 : Optimisation of the source gas and heating.***
*A range of coaxial gas (GS1), heating (TEM) and heater gas (GS2) were evaluated.*

The tested GS1 supply pressures modulated the number of identification by ≈ 10%, and the GS2 and TEM by ≈ 4%. Optimal source parameters were determined at **18 psi for GS1**, **20 psi for GS2** and **heating at 100 degrees**.

## **III.** Mass spectrometry

After acquisition of a survey MS spectrum, MS/MS spectra are acquired for the most intense peptides. The fragments m/z displayed in the MS/MS spectra and the m/z of the corresponding peptide are compared to a protein sequence database to identify the peptides. Peptide identifications will therefore be mainly affected by the m/z accuracy of the peptide and the quality of the MS/MS spectra. MS/MS spectra quality is mainly determined by the time spent by the instrument to realise the spectra, i.e. their accumulation time. Moreover, MS/MS spectra collection can be optimised to avoid acquiring poor quality or redundant MS/MS spectra which will not result in additional identification, and MS/MS spectra quality can be further optimised by finely tuning peptides isolation, fragmentation and m/z measurement, as detailed below.

### A. *Accumulation time*

In DDA mode, each cycle is composed of one MS survey scan followed by dependent MS/MS scans. For a good definition of peptide chromatographic peaks and perform precise XIC MS1 quantification, about 8-10 MS spectra must be acquired per chromatographic peak. Using our optimised LC gradient, the average peptide chromatographic peak duration was 22 sec, and therefore we used a **cycle time of 2.2 sec**.

The cycle time is defined as follows:

$$Cycle\ time = MS\ Accumulation\ time + Number\ of\ MS/MS \times MS/MS\ Accumulation\ time$$

The accumulation time is the time spent by the mass spectrometer to build a spectrum. Ions are not analysed continuously, but ion groups are periodically accelerated by ToF pulses at the entrance of the ToF, and analysed together. Data from multiple ToF pulses are summed to construct a mass spectrum. Therefore, more ToF pulses are summed with longer accumulation time, leading to an increased signal / noise ratio.

The MS spectra quality is crucial, as it will determine the accuracy of the measure peptides m/z, which will condition both XIC MS1 quantification and identification. To build robust and high quality MS spectra, we used an accumulation time of 150 ms.

Within a cycle, the time management is of main concern, as it will determine the repartition of the accumulation time among MS/MS spectra. Therefore, a compromise between the number of acquired MS/MS spectra and their accumulation time must be found. Several combinations were evaluated, from 20 MS/MS per cycle to 100 MS/MS per cycle, corresponding to 95 to 15 ms accumulation time, respectively.  The number of peptides identified by each method was compared (**Figure 31**).

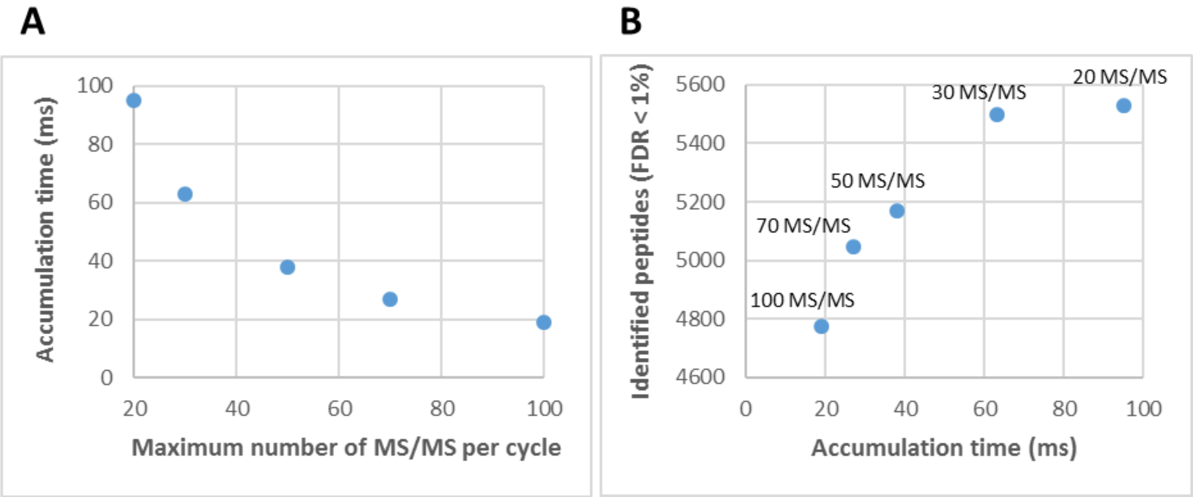


***Figure 31 : Optimisation of the cycle time sharing.***
*A. The MS/MS accumulation times of the five tested methods are plotted against their maximum number of MS/MS per cycle. It is of note that if less candidate precursor ions are detected, longer MS/MS accumulation time will be used to keep constant the cycle time. B. Comparison of the five methods.*

The definition of the maximum number of MS/MS per cycle and the corresponding accumulation time induced changes of ≈ 16% in the number of identified peptides. Using these different methods, we identified from 4 775 to 5 530 peptides. Even if it results in a lower number of acquired MS/MS spectra, longer accumulation time led to more peptide identifications with an optimum for **20 MS/MS per cycle** (also called Top 20 method) and **95 ms accumulation time per MS/MS**. In is of note that with the highest accumulation times we start to see a plateau, and with even higher accumulation times we should see a decrease in identifications because the high MS/MS spectra quality will not compensate the low number of acquired MS/MS spectra anymore.

Alternatively, the dynamic accumulation mode can be used to attribute a variable accumulation time for MS/MS according to their corresponding precursor ion intensity: longer accumulation time will be assigned to less intense ions, and shorter accumulation time to more intense ions, with a minimum of 25 ms for highly intense precursor ions. The number of MS/MS per cycle is therefore variable, and the mass spectrometer will acquire as many MS/MS as possible within the cycle, starting as usually with the most intense precursor ions. Using dynamic accumulation, the intensity threshold must be set very low (10 counts per seconds or cps), because low intensity precursor ions can still produce high quality MS/MS spectra with an extended accumulation time.

First results were very promising, as we were able to identify 5 727 peptides using dynamic accumulation, while a maximum of 5 530 peptides were identified using standard methods with fixed accumulation times. Dynamic accumulation was further evaluated in technical triplicates against a Top 30 and a Top 50 acquisition methods (**Figure 32**).

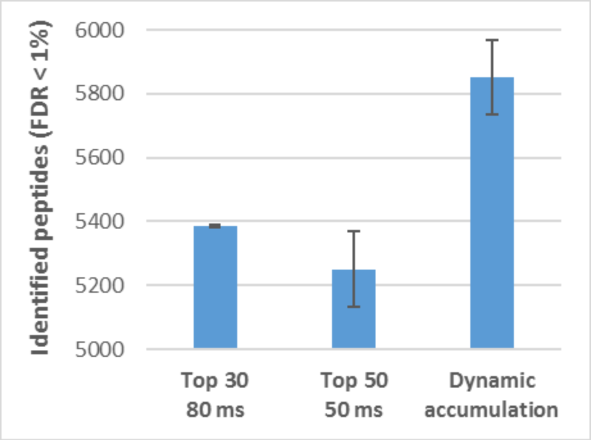


***Figure 32 : Evaluation of dynamic accumulation.***
*A dynamic accumulation method was compared to a Top 30 method with 80 ms accumulation time and a Top 50 method with 50 ms accumulation time.*

These results confirmed that the use of **dynamic accumulation** was beneficial, allowing the identification of ≈ 10% more peptides when compared to standard methods.

## B. *MS/MS spectra collection*

The analysis time is a precious resource that must be used sparingly. The optimisation of MS/MS spectra collection, i.e. avoiding collection of poor quality or redundant MS/MS spectra, can save time which can be better used to analyse informative MS/MS spectra.

### B.1. Intensity threshold

The MS/MS spectra quality, which will condition their identification, depends on their signal / noise ratio, which directly depends on the number of analysed ions. If a precursor ion intensity is weak, it will produce a poor quality MS/MS spectrum which will not lead to an identification. The acquisition of such poor quality MS/MS spectra can be avoided using an intensity threshold defined in the acquisition method, below which the precursor ions will not be selected for fragmentation. Using a Top 50 x 50 ms accumulation time for MS/MS, a range of intensity thresholds were assessed, from 100 to 1 000 counts per second (cps) (**Figure 33**).

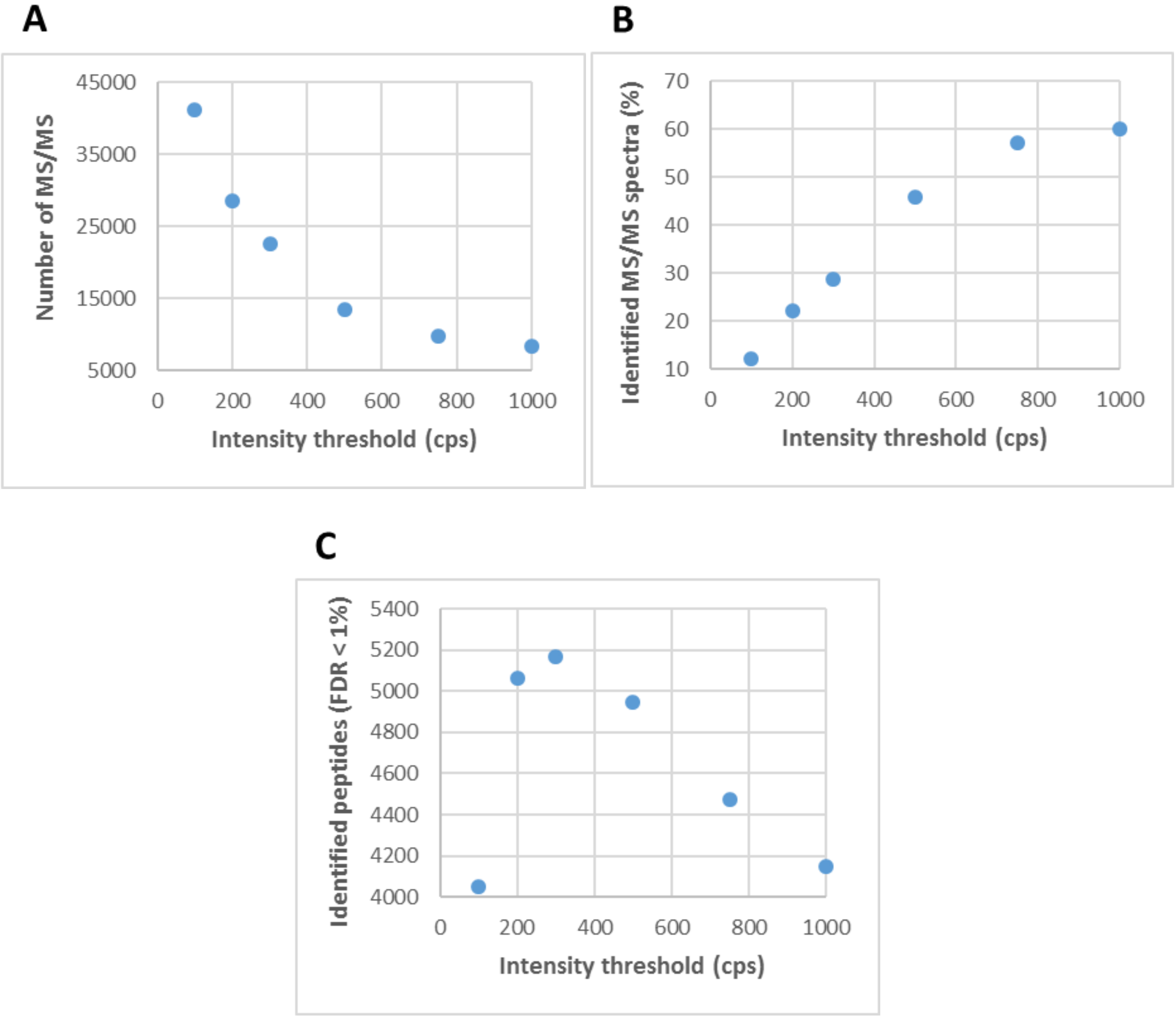


***Figure 33 : Optimisation of the precursor ion intensity threshold using a Top 50 x 50 ms method.***
*A. The total number of MS/MS of the whole analysis was plotted against the tested intensity thresholds. B. The percentage of MS/MS spectra that were identified is plotted against the tested intensity thresholds. C. The number of identified peptides is plotted against the tested intensity thresholds.*

The intensity threshold is of major importance for common Top N acquisition methods, as it modulated the number of identified peptides by ≈ 28%, from 4 052 to 5 168 peptides. While a higher intensity threshold decreases the number of acquired MS/MS spectra, it increases the percentage of identified MS/MS spectra due to their improved quality. The best compromise between the number of acquired MS/MS spectra and their quality was found for a precursor ion **intensity threshold at 300 cps**. At this threshold, one can consider that the wide majority of precursor ions above the lower limit of detection have been fragmented, while no time was wasted to acquire MS/MS spectra on noise peaks.

However, it does not mean that all MS/MS spectra collected from a precursor ion above the threshold will allow peptide identification, and this will be discussed below.

### B.2.Dynamic exclusion

In DDA mode, the most intense precursor ions are fragmented, and to avoid fragmenting again and again the same abundant peptides producing redundant MS/MS spectra, a dynamic exclusion can be used: precursor ions for which an MS/MS spectrum was already acquired will be excluded for a user-defined duration. The dynamic exclusion must be defined in the acquisition method by its duration and m/z exclusion window.

For the exclusion duration, three strategies were evaluated, based on the average chromatographic peak duration of 22 sec: (i) no dynamic exclusion; (ii) dynamic exclusion for 11 sec (i.e. half a chromatographic peak) to acquire an MS/MS spectrum at the top of the chromatographic peak; (iii) dynamic exclusion for 22 sec (i.e. chromatographic peak duration) to acquire only one MS/MS spectrum per precursor ion (**Figure 34**).

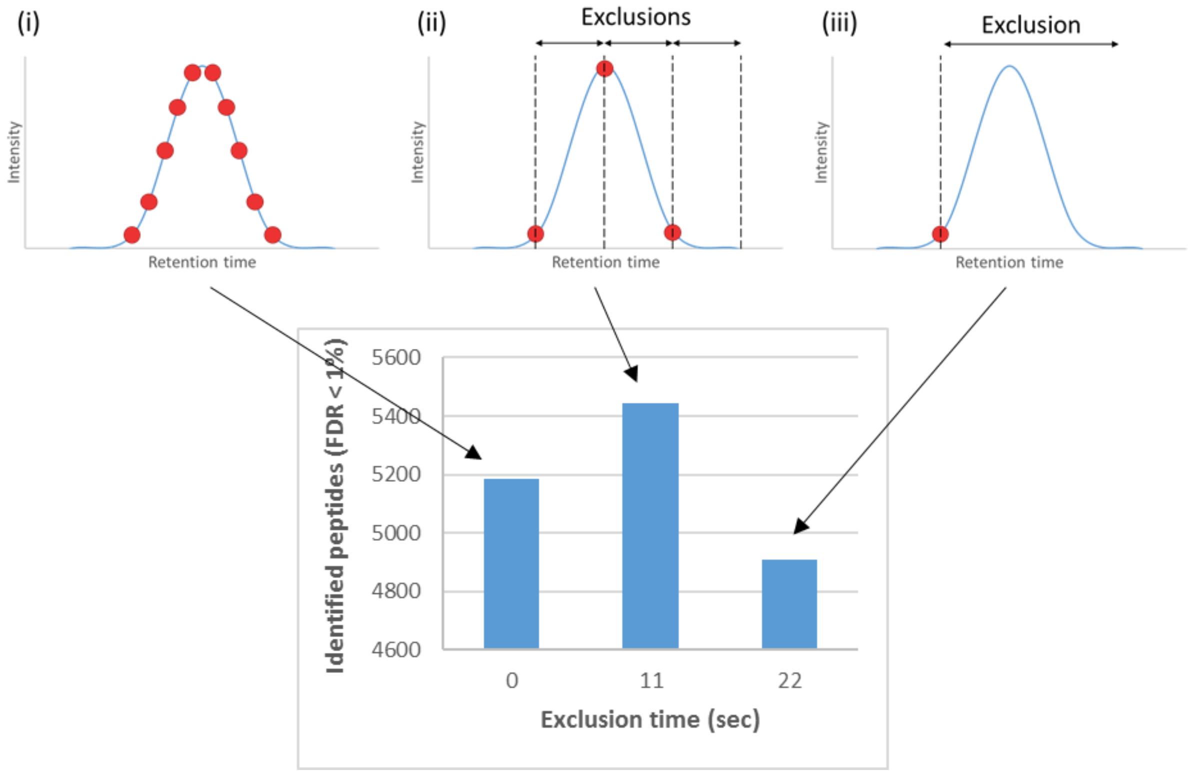


***Figure 34 : Optimisation of dynamic exclusion.***
*Three strategies were evaluated: (i) no exclusion time, (ii) exclusion time of half a chromatographic peak duration and (iii) exclusion time of the whole chromatographic peak duration. Red dots on the precursor ion chromatogram represent acquisition of corresponding MS/MS spectra.*

The tested dynamic exclusion methods accounted for ≈ 11% of the number of identifications, from 4 906 to 5 443 peptides. The best dynamic exclusion duration was found to be **11 sec**, representing half a chromatographic peak duration, aiming to collect an MS/MS spectrum when the precursor ion is the most intense.

The m/z exclusion window must also be set in the acquisition method. This m/z window can be explained as the tolerance with which a newly detected precursor ion will be considered as the same as previously. The m/z exclusion window should be wide enough to take into account the inter scan m/z accuracy variability, but not too wide to avoid excluding new precursor ions with close m/z. We evaluated a range of m/z tolerance, from 10 to 100 ppm (**Figure 35**).

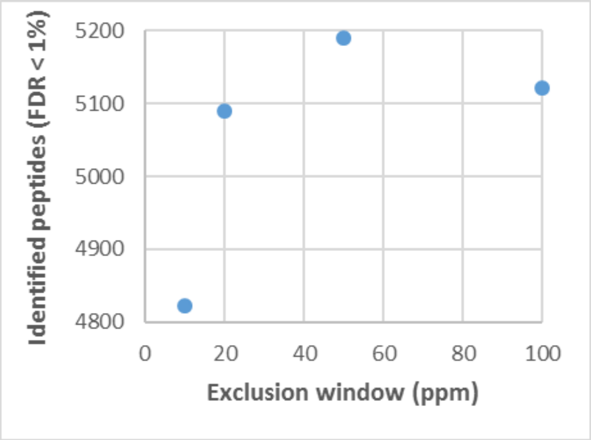


*Figure 35 : Optimisation of the m/z exclusion window.*

The m/z exclusion window had a significant impact of 8% on the number of identifications, from 4 822 to 5 190 peptides. The optimal exclusion window was found at **50 ppm**.

### C. *MS/MS spectra quality*

The spectra quality is a balance between sensitivity and specificity. The sensitivity can be defined as the signal / noise ratio, and the specificity is linked to the instrument selectivity (i.e. resolution) that was used to obtain this signal. The objective here is to find the optimal sensitivity and specificity balance leading to the most identifications.

The Triple TOF system allows the use of high resolution or high sensitivity modes for MS/MS spectra acquisition: using high resolution mode, only the most focalised ions are transmitted to the ToF analyser resulting in highly accurate and resolved signals; using high sensitivity mode, more ions are transmitted to the ToF leading to more intense but less accurate and resolved signals. For better sensitivity, the high sensitivity mode is usually used for MS/MS spectra acquisition.

Spectra quality can be further optimised by finely tuning Q1 resolution, collision energy spread and ToF resolution.

### C.1. Q1 resolution

Due to natural isotopes like mostly $^{13}C$, each peptide presents different masses (+ 1 Da for each $^{13}C$). The peptide form containing only the most abundant isotopes, i.e. no $^{13}C$, is called P, with one $^{13}C$ it is called P+1, with two $^{13}C$ it is called P+2, etc. Using a narrow isolation window, only the first isotope P will be fragmented, but using a wider isolation window allows the fragmentation of other isotopes which can produce identical b- and y-ions compared to P (**Figure 36**).

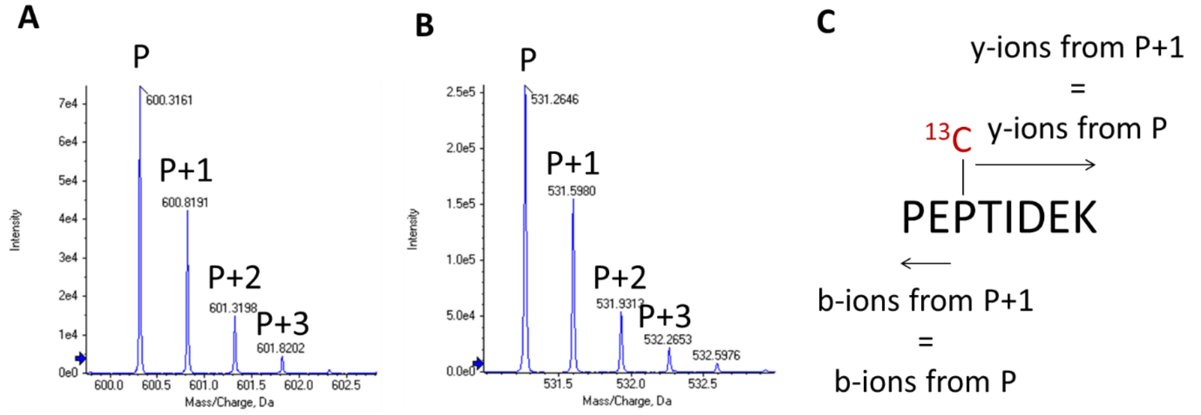


***Figure 36 : Isotopic envelope description.***
*A. Isotopic envelope of a doubly charged precursor ion. B. Isotopic envelope of a triply charged precursor ion. Since isotope masses differ from ≈ 1 Da, their m/z difference allows the determination of the peptide charge state. C. Fragmentation of a P+1 isotope can produce fragments that are identical to the fragments of the P isotope.*

Therefore, isolating isotopes together with the P can lead to the production of more P fragment ions which will lead to increased signal / noise ratio and increased sensitivity. However, a too wide Q1 isolation window can lead to co-isolation of multiple precursor ions with close m/z and generation of chimeric MS/MS spectra with fragments from multiple peptides, which are not easily identified.

The size of the Q1 isolation window can be modified by tuning its resolution, i.e. its full width at half maximum (FWHM). Three Q1 resolutions were evaluated: (i) 0.7 Da to collect only the first isotope P; (ii) 1 Da to collect two isotopes P and P+1; (iii) 2 Da to collect three isotopes P, P+1 and P+2 (**Figure 37**).

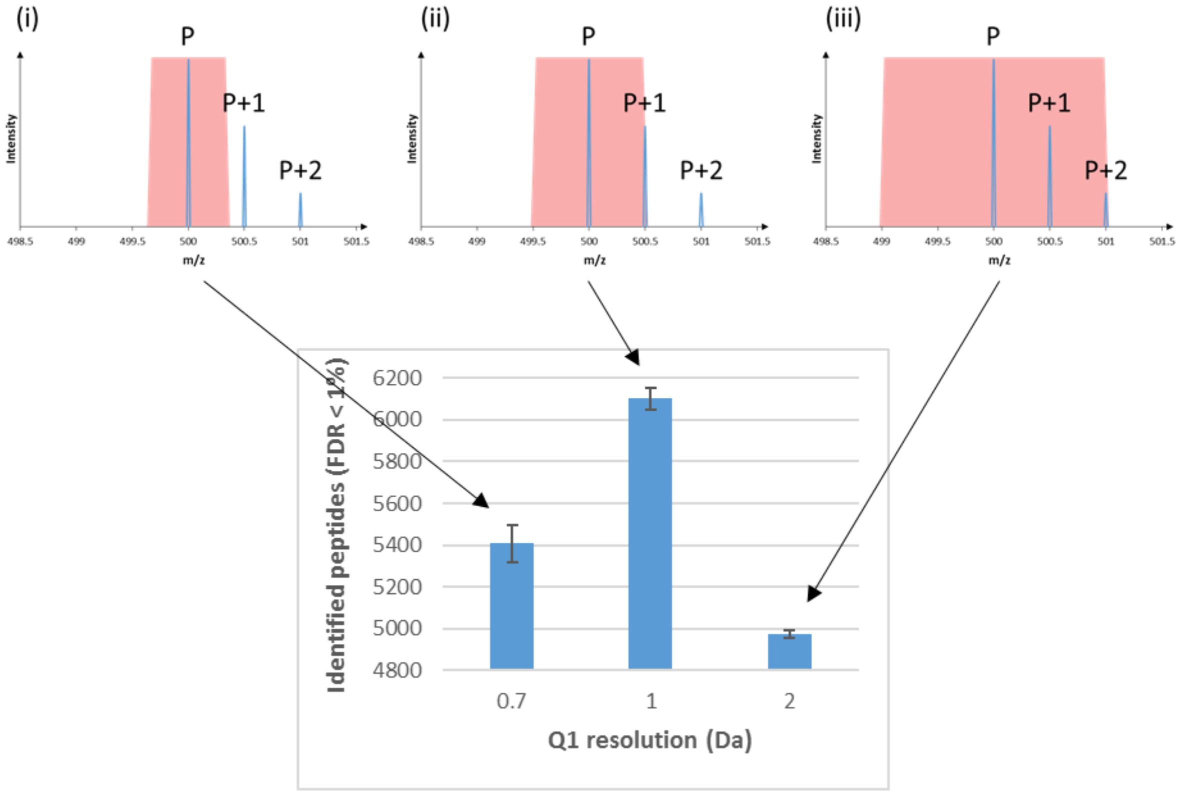
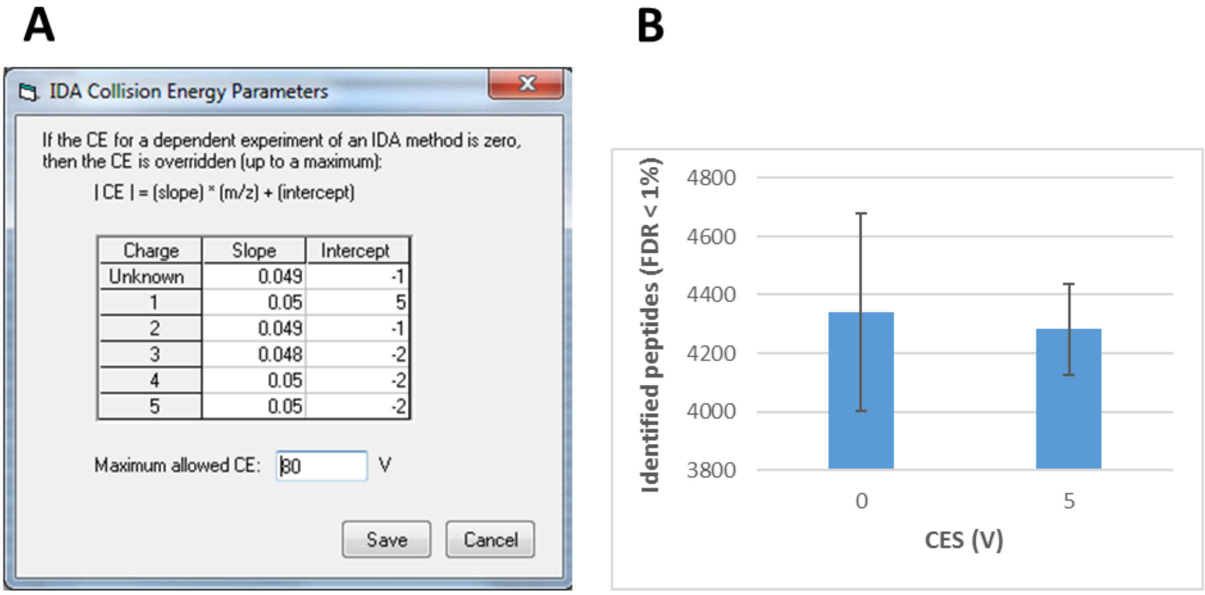


***Figure 37 : Optimisation of the Q1 resolution.***
*Three Q1 resolutions were evaluated: (i) 0.7 Da, (ii) 1 Da and (iii) 2 Da. Red areas represent the Q1 isolation windows.*

The Q1 resolution optimisation allowed a significant gain of 13% more identifications when compared to the standard 0.7 Da Q1 resolution. The optimal Q1 resolution was therefore defined at **1 Da**, which allows isolation of isotopes P and P+1. However, in highly complex samples like for metaproteomics, the Q1 resolution may be better reduced to 0.7 Da due to the increased number of interferences.

## C.2. Collision energy spread

After isolation, the precursor ions are fragmented in the collision cell. The applied collision energy (CE) is calculated for each precursor ion using equations provided by the instrument constructor, which were optimised on a large number of peptides. These equations allow the calculation of a CE using the precursor ion m/z and charge state. However, the optimal collision energy (CE) for each precursor ion depends on its amino acid sequence and its charge state[132]. Therefore, a collision energy spread (CES) can be used to apply a ramping CE from CE – CES to CE + CES, to improve precursor ion fragmentation. We evaluated the effect of using a CES of ± 5 V (**Figure 38**).



IDA Collision Energy Parameters

If the CE for a dependent experiment of an IDA method is zero, then the CE is overridden (up to a maximum):

| CE | = (slope) * (m/z) + (intercept)

| Charge | Slope | Intercept |
|---|---|---|
| Unknown | 0.049 | -1 |
| 1 | 0.05 | 5 |
| 2 | 0.049 | -1 |
| 3 | 0.048 | -2 |
| 4 | 0.05 | -2 |
| 5 | 0.05 | -2 |

Maximum allowed CE: 80 V

***Figure 38 : Optimisation of precursor ions fragmentation.***
*A. The constructor equations used to calculate the applied collision energy (CE) are displayed. B. A collision energy spread (CES) of ± 5 V was evaluated.*

The tested CES of 5 V did not increase the number of identified peptides. The calculated CE seems well estimated, and the use of a CES **not necessary**.

### C.3. ToF resolution

In a mass spectrum, the signal intensity is determined by counting individual ion pulses that reach the detector within time bins, which are summed to construct a mass spectrum. The number of summed time bins determines the ToF resolution, and tips the scales between sensitivity and specificity: while more summed time bins increases the signal / noise ratio, less summed time bins increases the m/z accuracy and resolution. We tested 4 and 8 time bins to sum (**Figure 39**).

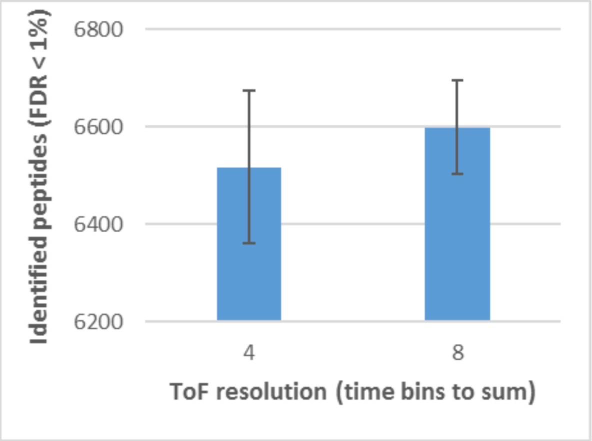


***Figure 39 : Optimisation of the number of time bins to sum.***

Even if no significant difference was observed between 4 and 8 time bins to sum, the results seemed slightly better and MS/MS spectra quality as well using **8 summed time bins**.

## IV. Conclusion

The objective of this work was to further understand the functioning of the microLC-Triple TOF 6600 coupling, and to provide the lab members optimal instrument settings for their shotgun proteomics experiments.

A global view of the key parameters of a DDA method as well as their relationship is presented in **Figure 40**.

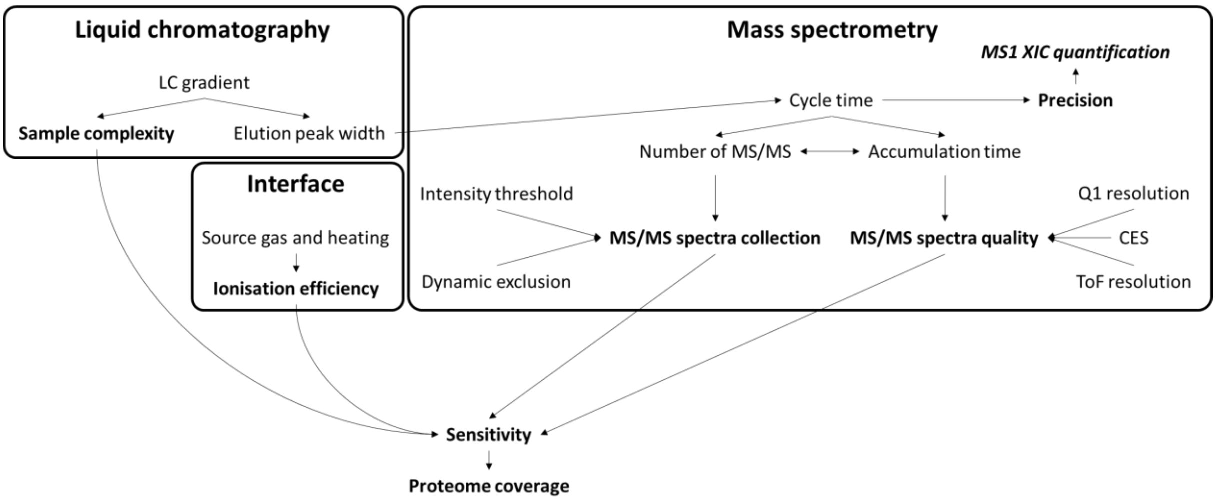

Liquid chromatography
LC gradient
Sample complexity
Elution peak width
Interface
Source gas and heating
Ionisation efficiency
Mass spectrometry
MS1 XIC quantification
Cycle time
Precision
Number of MS/MS
Accumulation time
Intensity threshold
Q1 resolution
MS/MS spectra collection
MS/MS spectra quality
CES
Dynamic exclusion
ToF resolution
Sensitivity
Proteome coverage

***Figure 40 : Overview of the LC-MS/MS key parameters driving the proteome coverage on the microLC-Triple TOF 6600 coupling.***
*The most important parameter was found to be the peptide separation by liquid chromatography. The LC gradient length and design define the sample complexity throughout the analysis and therefore highly affect the sensitivity of the assay. The interface gas supply and heating help peptide ionisation and also directly affect the sensitivity of the assay. Finally, the mass spectrometry part was extensively optimised. First, the optimal cycle time was determined as 1/10 of the average peptide chromatographic peak duration to allow precise XIC MS1 quantification. The optimal cycle time sharing, between number of MS/MS and accumulation time, is automatically determined using the dynamic accumulation mode, which adapt the accumulation time to the precursor ion intensity. The MS/MS spectra collection was optimised to avoid collection of poor quality and redundant MS/MS spectra using an intensity threshold and dynamic exclusion (it is of note that using dynamic exclusion, the intensity threshold should be set very low to let the mass spectrometer freely manage the cycle time sharing). The best MS/MS spectra quality, i.e. the best compromise between sensitivity and specificity, was optimised by finely tuning the Q1 and ToF resolutions.*

Optimised parameters are presented in **Table 4**.

***Table 4 : Optimised parameters for DDA method on the microLC-Triple TOF 6600 coupling.***
*Optimised value for each parameter is provided, as well as an estimation of the contribution of this parameter on the proteome coverage estimated by the range of values obtained during optimisations.*

| | Parameter | Optimum | Contribution |
|---|---|---|---|
| Liquid chromatography | Gradient | Gradient 2 | 29% |
| Interface | Coaxial gas | 18 psi | 10% |
| | Heater gas | 20 psi | 4% |
| | Heater | 100 °C | |
| Mass spectrometry | Accumulation time | Dynamic accumulation | 10% |
| | Intensity threshold | 10 cps | NA |
| | Dynamic exclusion | ½ chromatographic peak | 11% |
| | Q1 resolution | 1 Da | 13% |
| | Collision energy spread | 0 | 1% |
| | ToF resolution | 8 bins | 1% |

This optimised DDA method is now used as the standard shotgun proteomics method on the microLC-Triple TOF 6600 coupling. However, these optimisations were performed on a yeast digest only to provide an optimised "standard" acquisition method. If time and material are available, sample-specific optimisations could still enhance the results quality, for instance because of different sample complexity, or specific protein populations (e.g. hydrophobic peptides).

# Chapter II Benchmarking of targeted proteomics configurations

Today, selected reaction monitoring (SRM) coupled to stable isotope dilution and performed on triple quadrupole instrument is the gold standard approach for targeted proteomics, allowing sensitive, robust and absolute quantification of proteins of interest in complex biological samples[47-48]. Sensitivity and robustness are the most important parameters for such approaches in which low abundant proteins have to be quantified in hundreds of samples. In the laboratory, targeted proteomics is usually performed on microLC-triple quadrupole couplings. MicroLC is preferred to nanoLC because of its increased robustness and reproducibility[51].

Recently, a targeted method was introduced on quadrupole-orbitrap mass spectrometer called parallel reaction monitoring (PRM) which exhibited similar performances when compared to SRM[24]. Equivalent methodologies were developed on quadrupole-time of flight instruments like multiple reaction monitoring in high resolution (MRM HR)[25-26]. These SRM-like methodologies, developed on high resolution/accurate mass (HR/AM) instruments have the advantage of offering an increased specificity linked to the high resolution of the analyser which allows interferences removal. Indeed, stringent m/z extraction windows can be used for data analysis, typically in the range of ≈ 50 ppm, while in SRM the third low resolution quadrupole usually isolates fragment ions in a 0.7 Da window, representing an isolation window of 875 ppm for an ion at 800 m/z. Moreover, the method development is easier when compared to SRM because the best responding transitions do not need to be chosen, and both the discovery and validation steps of a project can be performed with a single instrument using DDA and PRM, respectively: interesting proteins could be identified in DDA mode and further validated on a large cohort of samples by a targeted approach.

In this chapter I will present a benchmarking of four targeted proteomics configurations, including a standard microLC-SRM platform we usually use for targeted proteomics for large cohorts of samples, but also three configurations that are usually used for shotgun approaches but which could also be used for targeted proteomics, including a nanoLC-PRM coupling, a nanoLC-MRM HR coupling and a microLC-MRM HR coupling. The objective of this chapter is to help in the decision making about which instrument to use for targeted proteomics.

## I. Workflow

We used a sample from a targeted proteomics project as a model. This project involved the quantification of 10 biomarker candidates in bovine muscle to predict the meat quality. First shotgun results led to the identification of 10 biomarker candidates. Then, we ordered 39 crude stable isotope labelled peptides corresponding to these 10 biomarker candidates, to validate them on a large cohort of samples by targeted proteomics.

A sample pool was used to benchmark the following targeted MS configurations: the standard microLC-SRM configuration (Dionex UltiMate 3000 coupled to a TSQ Vantage triple quadrupole mass spectrometer, both from Thermo Fisher Scientific), a microLC-MRM HR configuration (Eksigent NanoLC 400 system coupled to a TripleTOF 6600 quadrupole-time of flight mass spectrometer both from SCIEX), a nanoLC-MRM HR configuration (nanoAcquity UPLC from Waters coupled to a TripleTOF 5600+ quadrupole-time of flight mass spectrometer from SCIEX) and a nanoLC-PRM configuration (nanoAcquity UPLC from Waters coupled to a Q Exactive Plus quadrupole-orbitrap mass spectrometer from Thermo Fisher Scientific) (**Figure 41**).

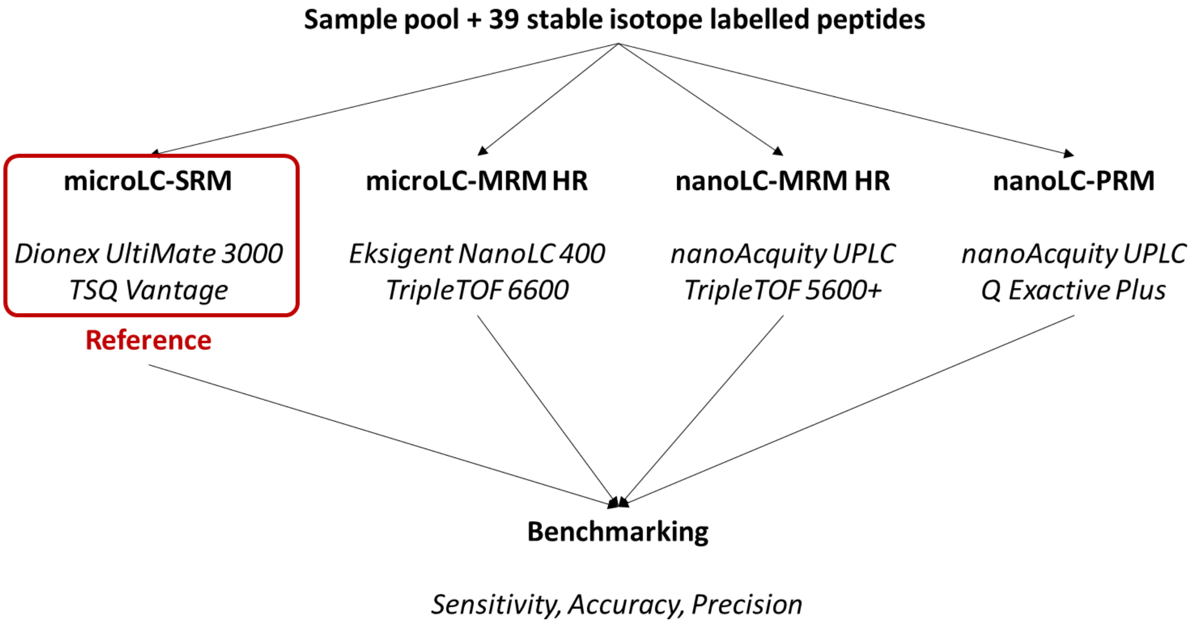

Sample pool + 39 stable isotope labelled peptides
microLC-SRM
Dionex UltiMate 3000
TSQ Vantage
Reference
microLC-MRM HR
Eksigent NanoLC 400
TripleTOF 6600
nanoLC-MRM HR
nanoAcquity UPLC
TripleTOF 5600+
nanoLC-PRM
nanoAcquity UPLC
Q Exactive Plus
Benchmarking
Sensitivity, Accuracy, Precision

*Figure 41 : Workflow for the benchmarking of four targeted proteomics configurations.*

For nanoLC couplings, we used a 250 mm x 75 µm column and injected 800 ng of sample, while for microLC couplings, we used a 150 mm x 300 µm column and injected 6 µg of sample. We used an identical LC gradient for all configurations, from 5 to 25% ACN in 47 min, 25 to 35% ACN in 10 min, 35 to 70% ACN in 2 min, isocratic for 5 min, 70 to 5% ACN in 1 min, and isocratic for 19 min. Time-scheduled methods were developed for each coupling to quantify these 39 peptides with 4 minutes time windows and ≈ 2.5 seconds cycle time. The concentration-balanced mix of heavy peptide was optimised to approach the 1/1 ratio between the heavy and light peptides.

For microLC-SRM configuration, collision energies were optimised for each transition. For nanoLC- and microLC-MRM HR analyses, the accumulation time was set at 150 ms for MS and 100 ms for MS/MS. For nanoLC-PRM analyses, the resolution was set at 35 000 at 200 m/z, with an automatic gain control (AGC) target of $10^6$ and a maximum injection time of 128 ms.

The 39 peptides were monitored in technical triplicates using each configuration, and the data analysis was performed in Skyline. For PRM and MRM HR analyses, the extraction window was adapted to the resolution of the instruments (35 000 at 200 m/z for PRM analyses, and 15 000 for MRM HR analyses). The light / heavy ratios were used to quantify each peptide.

## II. Results

The quantification results of the 39 peptides were used to benchmark the four targeted MS configurations in terms of sensitivity, accuracy and precision.

### A. *Sensitivity*

The sensitivity of a coupling is ideally determined using a calibration curve made with highly purified and precisely quantified heavy labelled peptides. For the early stage of this project, we used crude heavy labelled peptides, thus the accurate quantities of spiked in heavy peptides were not known. However, we estimated the sensitivity of the four targeted MS configurations by visually inspecting the signal / noise ratios (**Figure 42**).

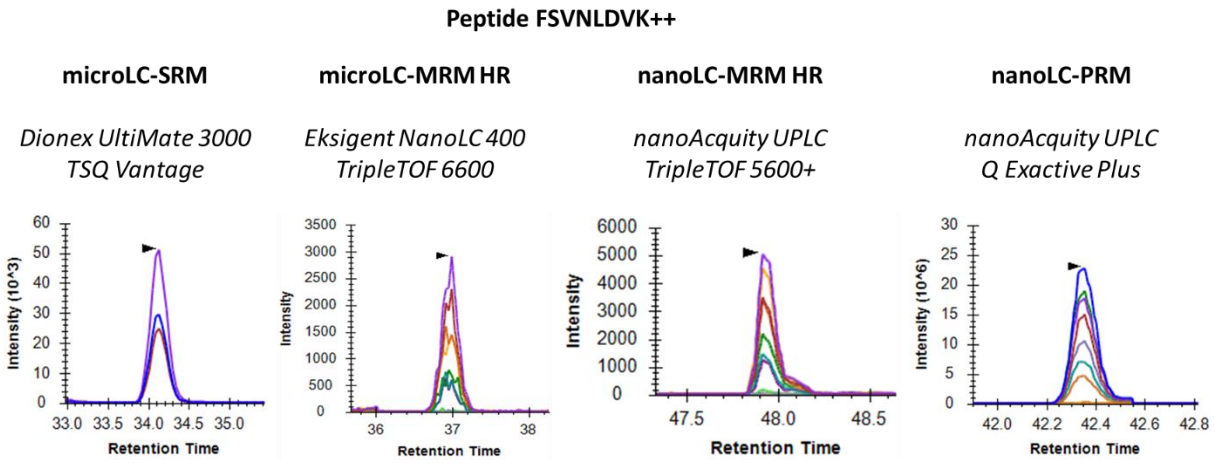


***Figure 42 : Evaluation of the sensitivity of the targeted MS configurations.***
*The sensitivity of the tested couplings was evaluated by comparing the signal / noise ratios. Here is an example of the doubly charged FSVNLDVK peptide. It is of note that using targeted methods on HR/AM instruments, we can extract many more transitions per peptide than the number that can be monitored by SRM.*

In this example, the signal / noise ratio are very good for the four configurations. Even if several differences were observed for certain peptides, globally the tested couplings performed equivalently, and therefore their sensitivity was estimated as equivalent.

### B. *Accuracy*

The accuracy of the targeted MS platforms was evaluated by comparing the endogen / stable isotope labelled peptide ratio obtained using the HR/AM instruments platforms to those obtained using the gold standard microLC-SRM reference (**Figure 43**).

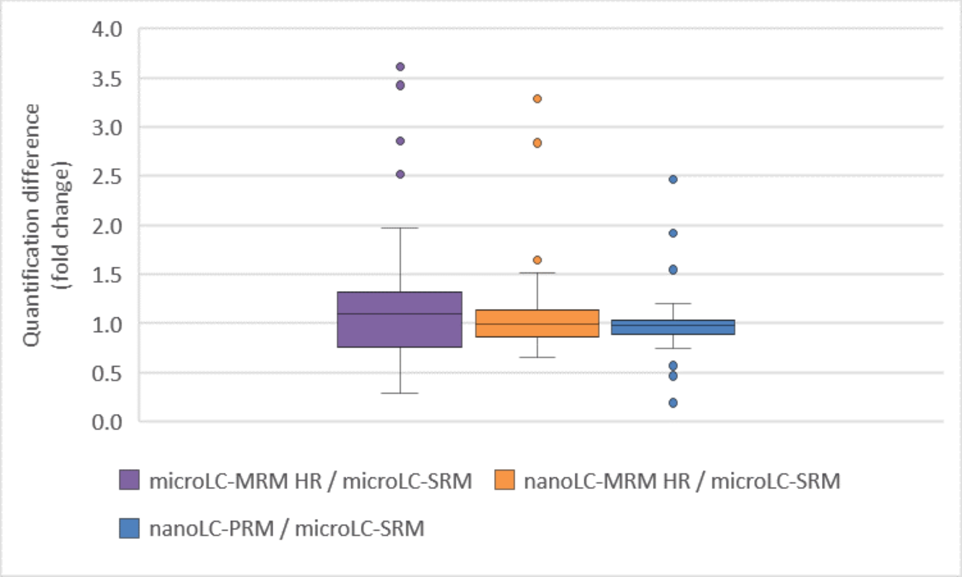


***Figure 43 : Evaluation of the accuracy of the targeted MS platforms.***
*Endogen peptide over stable isotope labelled peptide ratio were compared to the ratio obtained by microLC-SRM.*

Globally, all targeted MS platforms gave consistent results. In more details, 95% of the nanoLC-MRM HR quantifications were accurate within a factor of 2 when compared to microLC-SRM quantifications, and this was the case for 92% of the nanoLC-PRM quantifications and 79% of the microLC-MRM HR quantifications.

### C. *Precision*

The precision of the quantification was probed using coefficients of variation (CV) between technical triplicates (**Figure 44)**.

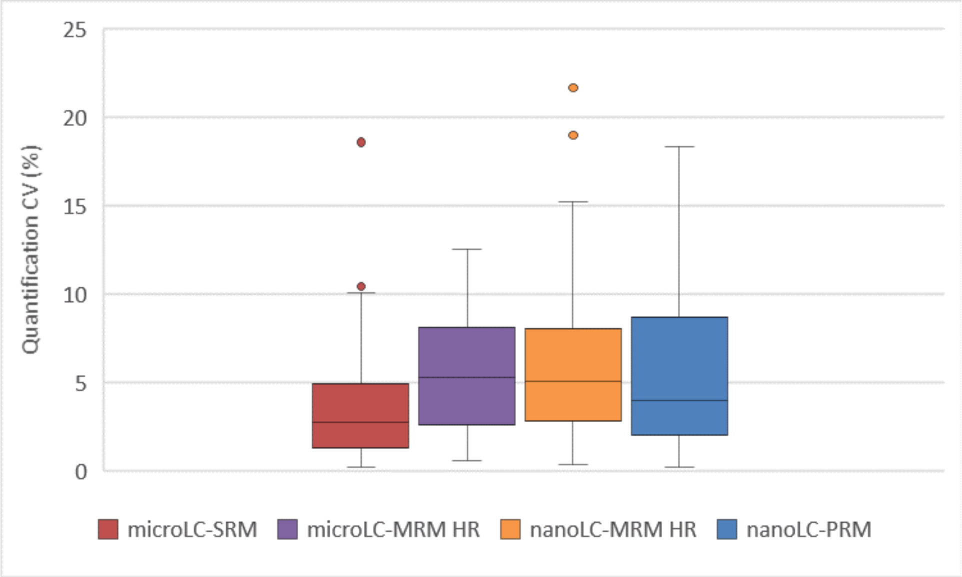


***Figure 44 : Evaluation of the precision of the targeted MS platforms.***
*Coefficients of variation (CV) between technical triplicates were used to compare the precision of the configurations.*

The four targeted MS platforms presented an equivalent good precision, with a majority of CV between technical triplicates around 5%, but with slightly lower CV with microLC-SRM.

### **III.** Conclusion

The objective of this chapter was to present a benchmarking of several targeted proteomics configurations, including the gold standard SRM performed on a triple quadrupole, but also targeted approaches performed on HR/AM last generation instruments, including PRM performed on a quadrupole-orbitrap and MRM HR performed on a quadrupole-time-of-flight. In addition, nanoLC was compared to microLC.

Globally, all targeted configurations gave similar results, in terms of sensitivity (equivalent signal / noise ratio), accuracy (equivalent light / heavy peptide ratio) and precision (equivalent CV between technical triplicates ≈ 5%). These results are consistent with previous studies[25, 49-50]. This means that the targeted quantification performances of HR/AM instruments are equivalent to those of low resolution triple quadrupole instruments, and the decision should rely on other parameters. Since usually targeted proteomics is performed for large cohorts of samples, a key parameter is the robustness of the coupling. In terms of robustness, microLC has proven to be superior to nanoLC and should be preferred for prolonged analyses[51]. Then, the decision making between SRM and PRM could be based on instrument availability, since triple-quadrupoles are usually dedicated to SRM, while

HR/AM instruments can perform shotgun or DIA analyses. If the HR/AM instrument is available, it could be used for targeted proteomics, preferentially operated in microLC mode if enough sample is available. Indeed, PRM or MRM HR offer a better specificity thanks to the use of a high resolution analyser, and allow an easier method development because the best responding transitions have not to be chosen for each peptide. Moreover, if the biomarker discovery was performed on the same coupling, the method transfer is straightforward, as the retention times of the targeted peptides are already known for this LC configuration.

In conclusion, the choice of the targeted configuration, if multiple last generation instruments are available, should be based on the available sample amount and the flowrate in place within the coupling (microLC is more robust than nanoLC), and based on instrument availability.

# **Chapter III** Optimisation of a data independent acquisition workflow

Data Independent Acquisition (DIA) has been recently introduced on high resolution/accurate mass (HR/AM) instruments in order to extract quantitative information from whole complex proteome maps[27]. DIA promises to combine the advantages of both shotgun and targeted proteomics, providing sensitive and reproducible quantification of all detectable peptides.

Shotgun proteomics and data dependent acquisition (DDA) allow today the quantification of thousands of proteins within a single analysis[9], but both dynamic range and reproducibility are still limited[144]. A few years ago, targeted proteomics approaches emerged, like selected reaction monitoring (SRM)[48] or parallel reaction monitoring (PRM)[24]. Due to the use of MS/MS signals for quantification, targeted approaches offer a wider dynamic range, and increased specificity and sensitivity when compared to classical shotgun XIC MS1 quantification. However, they are limited to tens of targeted proteins. Recently, data independent acquisition (DIA) approaches were developed to combine the advantages of both shotgun and targeted proteomics, with an equivalent or even higher proteome coverage than the one of shotgun approaches, and sensitivity, specificity and dynamic range comparable to those of targeted approaches. Indeed, in DIA mode, MS and MS/MS data are collected for the whole m/z range, providing a comprehensive proteome coverage which is only limited by the detection limit of the mass spectrometer, and not by peptide selection issues that are inherent to DDA. Moreover, the use of MS/MS data provide sensitivity, specificity and dynamic range equivalent to those of targeted approaches. Finally, the systematic fragmentation of the whole m/z range provides a reproducibility comparable to targeted approaches. However, the data analysis is today the major bottleneck for DIA approaches due to the complexity of the data. Several DIA data analysis workflow exist, but today the most efficient is the peptide-centric approach which relies on the use of a preliminary built spectral library[29]. A spectral library is made from DDA data, and contains all information necessary to extract DIA data in a targeted way, i.e. the peptides and fragments m/z, their retention time, and the relative intensities of the fragments.

In this chapter, I will deeply describe a whole DIA-SWATH (sequential windowed acquisition of all theoretical fragment ion mass spectra) workflow and the thorough optimisations which were performed during my PhD, including sample preparation, data acquisition and data analysis, with an emphasis on data analysis which is today the major bottleneck of this acquisition mode. Finally, the identification and quantification performances of DIA were compared to those of a classic shotgun approach using DDA (**Figure 45**).

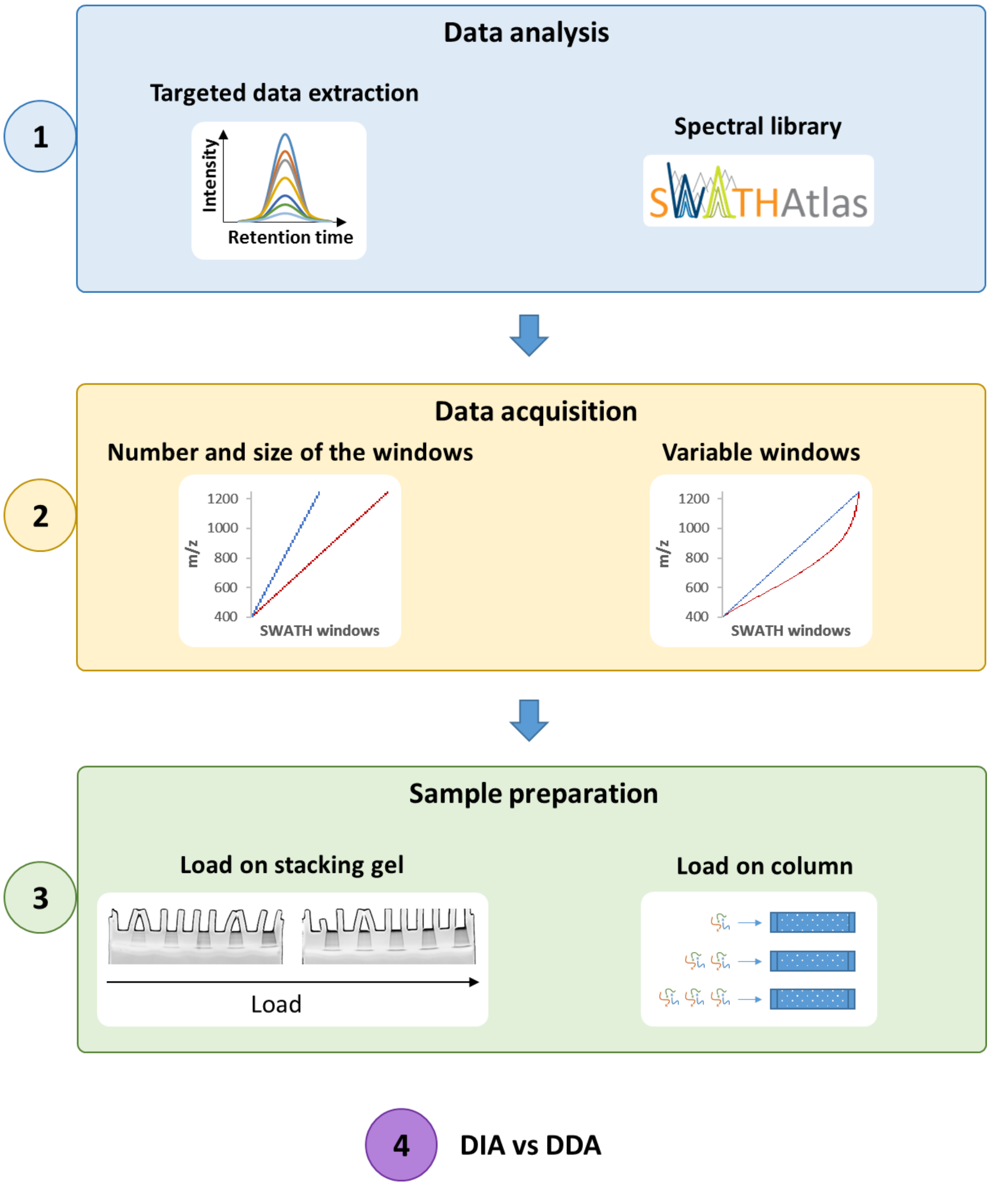

Data analysis
1
Targeted data extraction
Intensity
Retention time
Spectral library
SWATHAtlas
Data acquisition
2
Number and size of the windows
1200
1000
800
600
400
m/z
SWATH windows
Variable windows
1200
1000
800
600
400
m/z
SWATH windows
Sample preparation
3
Load on stacking gel
Load
Load on column
4
DIA vs DDA

***Figure 45 : Overview of the DIA workflow optimisation strategy.***
*First, the DIA data analysis was optimised, i.e. targeted data extraction and a comparison between a homemade and a publicly available spectral library was performed. Then, the Q1 windows settings were optimised for the data acquisition step, including the use of different number and size of windows as well as the use of variable windows. Finally, the sample amount that is loaded on the gel and onto the column was optimised for the sample preparation step.*

## I. Data analysis

The data analysis is today the major bottleneck in the DIA workflow. Various DIA data interpretation strategies are under development, but today the most straightforward way to reliably extract quantitative data from DIA experiments relies on the use of a preliminary built MS/MS spectra library (peptide centric approach)[27]. The main challenge for DIA data analysis is to correctly integrate MS/MS extracted ion chromatograms. In this context, we tested several data extraction workflows to highlight guidelines for DIA (in particular SWATH) data analysis.

## A. *Targeted data extraction*

Targeted data extraction refers to the use of a spectral library to extract information of targeted peptides from DIA data. The spectral library provides the following information to the software used to extract DIA data: protein names, peptides sequences, detected or normalised retention times (RT), precursor ions m/z, fragment ions m/z and the relative intensity of the fragments. The data extraction software will therefore know, for each peptide present in the spectral library, when it should elute, in which Q1 window the precursor ions were isolated, and the m/z of the corresponding fragments to extract. The relative intensity of the fragments helps choosing the most intense ones for the best sensitivity, and provides a quality control (dot product) by correlating their relative intensity in the library to the one found in the DIA data.

### A.1.Workflow

We used a reference sample, consisting in a background of 800 ng of yeast total lysate spiked with either 5 or 25 fmol of an equimolar mix of 48 human proteins (Universal Proteomics Standard, UPS1, Merck). Proteins were purified using tube gels and in-gel digested using trypsin. Peptides were extracted and retention time standards (iRT, Biognosys, Zurich, Switzerland) were spiked for RT alignment.

Samples were analysed on a nanoLC-Triple TOF 5600+ (quadrupole-time of flight mass spectrometer) coupling in DDA mode to build a spectral library. Peptides and proteins were identified using Mascot search engine, and validated using the Proline software (1% FDR at the peptides and proteins levels). In total, we identified 5 336 proteotypic peptides corresponding to 1 422 proteins. Samples were then analysed in DIA mode using a 67 variable windows SWATH method, and data were extracted using different strategies to quantify UPS1 proteins along with all possible yeast proteins. Thereby, we evaluated several parameters to extract DIA data, namely retention time tolerance, m/z window, number of transitions per peptide and FDR threshold, as well as two software solutions.

Each quantification workflow was evaluated in their ability to differentiate varying UPS1 peptides and constant yeast peptides. For this purpose, we evaluated the sensitivity using the True Positive Rate (TPR), and the specificity using the False Discovery Proportion (FDP), which are defined based on previous work[129]:

$$TPR = \frac{TP}{TP + FN}$$

$$FDP = \frac{FP}{TP + FP}$$

True Positive (TP) = variant UPS1 peptide; False Negative (FN) = constant UPS1 peptide; False Positive (FP) = variant yeast peptide.

We considered a peptide as variant between the 25 fmol and the 5 fmol spikes if the fold change was more than 2 and the p-value (Welch’s t-test) less than 0.05.

The whole workflow is presented in **Figure 46**.

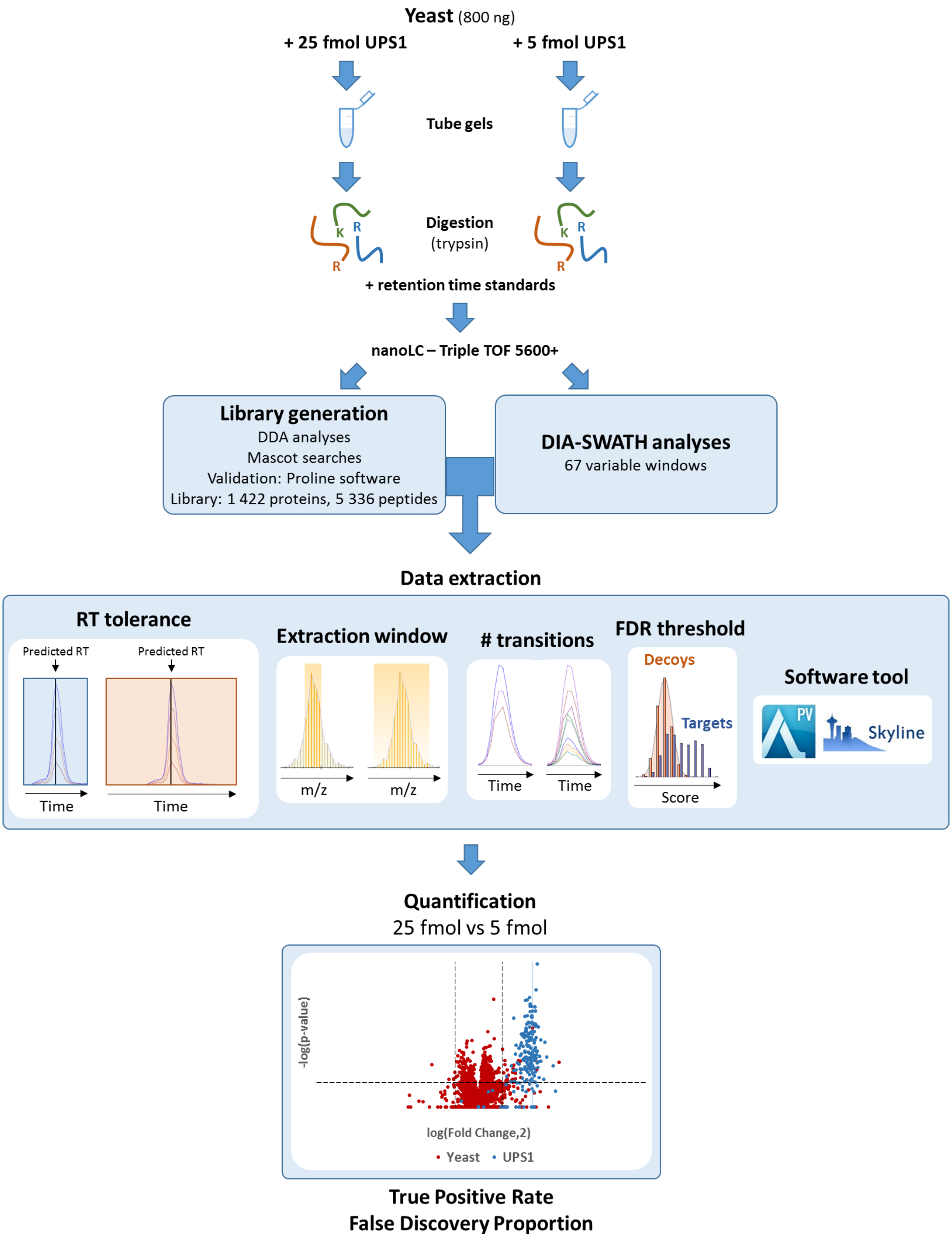

Yeast (800 ng)
+ 25 fmol UPS1
+ 5 fmol UPS1
Tube gels
Digestion
(trypsin)
K
R
R
+ retention time standards
nanoLC – Triple TOF 5600+
Library generation
DDA analyses
Mascot searches
Validation: Proline software
Library: 1 422 proteins, 5 336 peptides
DIA-SWATH analyses
67 variable windows
Data extraction
RT tolerance
Predicted RT
Predicted RT
Time
Time
Extraction window
m/z
m/z
# transitions
Time
Time
FDR threshold
Decoys
Targets
Score
Software tool
PV
Skyline
Quantification
25 fmol vs 5 fmol
-log(p-value)
log(Fold Change,2)
• Yeast • UPS1
True Positive Rate
False Discovery Proportion

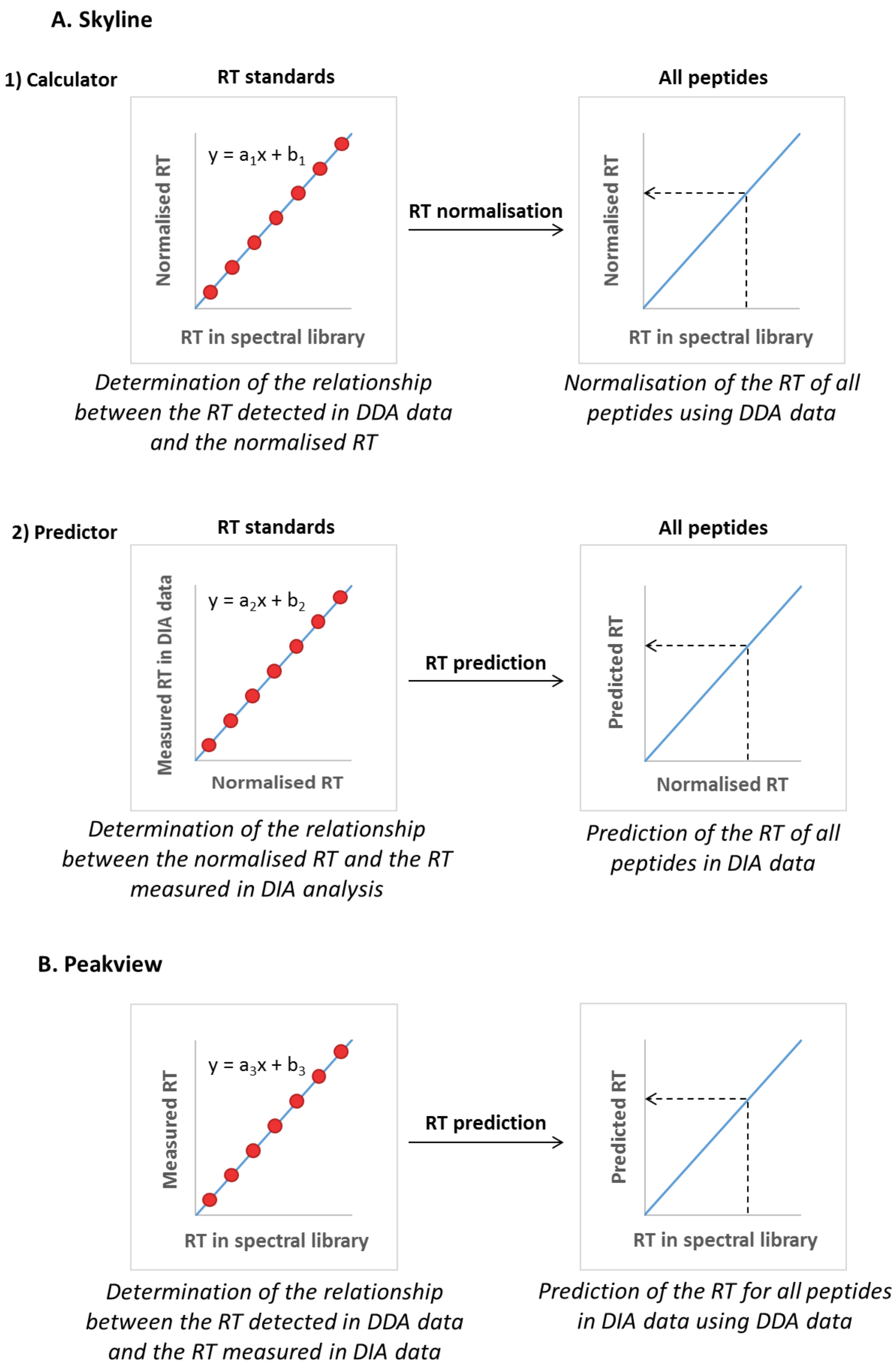
***Figure 46 : Workflow for the optimisation of targeted DIA data extraction.***
*Either 25 or 5 fmol of 48 UPS1 proteins were spiked into a yeast background matrix. After protein purification using tube gels, the proteins were in-gel digested. Peptides were analysed together with retention time standards on a nanoLC-Triple TOF 5600+ coupling, in DDA mode followed by Mascot searches and validation of the identifications by Proline, and then in DIA mode using a 67 variable windows DIA-SWATH method. DIA data were extracted using different sets of parameters, including RT tolerance, m/z extraction window, number of transitions to use, FDR threshold and two software tools (Peakview, SCIEX, and Skyline, MacCoss Lab of Biological Mass Spectrometry, university of Washington). UPS1 peptides were quantified along with yeast peptides to determine a true positive rate and a false discovery proportion which were used to drive the optimisations.*

Starting parameters were 6 min (± 3 min) for the retention time tolerance, 80 ppm for the m/z window, 6 transitions per peptide and 1% FDR for both Skyline and Peakview. From these initial settings, the parameters were optimised one at a time.

### A.2. Retention time tolerance

DIA data are extracted within a defined retention time (RT) tolerance around the predicted RT. RT prediction is performed using a set of peptides that allow RT alignment between the spectral library and the DIA analyses. The peptides used for RT alignment can be either endogen peptides or spiked in standard peptides. The interest of using spiked in RT standards is that they are present in every sample as soon as they are spiked in.

Skyline and Peakview software tools use different strategies to perform RT alignment between the detected RT present in the spectral library and the corresponding predicted RT in the DIA analyses. Skyline uses two steps: (i) a calculator uses the RT standards to perform a linear regression between the RT present in the spectral library and their normalised RT which are provided by the RT standards supplier, and uses the equation of the linear regression to normalise the RT of all peptides present in the spectral library, (ii) a predictor uses the RT standards to perform a linear regression between the normalised RT and the RT detected in the DIA analysis, and uses the equation of the linear regression to predict the RT of all peptides present in the spectral library. On the other hand, Peakview uses RT standards to perform a linear regression between the RT present in the spectral library and the RT detected in DIA analyses, and uses the equation to predict RT for all peptides present in the spectral library (**Figure 47**).


A. Skyline
1) Calculator
RT standards
All peptides
Normalised RT
$y = a_1x + b_1$
RT in spectral library
RT normalisation
Normalised RT
RT in spectral library
2) Predictor
RT standards
All peptides
Measured RT in DIA data
$y = a_2x + b_2$
Normalised RT
RT prediction
Predicted RT
Normalised RT
B. Peakview
Measured RT
$y = a_3x + b_3$
RT in spectral library
RT prediction
Predicted RT
RT in spectral library

*Determination of the relationship between the RT detected in DDA data and the normalised RT*

*Normalisation of the RT of all peptides using DDA data*

*Determination of the relationship between the normalised RT and the RT measured in DIA analysis*

*Prediction of the RT of all peptides in DIA data*

*Determination of the relationship between the RT detected in DDA data and the RT measured in DIA data*

*Prediction of the RT for all peptides in DIA data using DDA data*

***Figure 47 : Retention time alignment performed by Skyline and Peakview.***
*A. With Skyline, a calculator uses RT standards to create a linear regression between the RT present in the spectral library and the normalised RT. The resulting equation is used to normalise RT of all peptides present in the spectral library. Then, a predictor uses RT standards to create a linear regression between the normalised RT and the measured RT in DIA data. The equation is used to predict the RT of all peptides present in the spectral library. B. Peakview uses the RT standards to perform a linear regression between the RT present in the spectral library and the detected RT. The resulting equation is used to predict the RT of all peptides present in the spectral library. Red dots represent the RT standards.*

Different RT tolerances were tested for data extraction, from 3 to 10 min (**Table 5**).

***Table 5 : Retention time tolerance optimisation for targeted DIA data extraction.***
*The true positive rate (TPR) and false discovery proportion (FDP) are presented for each parameter and each software tool. The resulting optimal setting is presented in bold red.*

| Parameter | Value | Skyline | | Peakview | |
|---|---|---|---|---|---|
| | | TPR (%) | FDP (%) | TPR (%) | FDP (%) |
| RT tolerance | 3 min | 63 | 14 | 77 | 39 |
| | **6 min** | 80 | 20 | 82 | 20 |
| | 10 min | 83 | 23 | 79 | 18 |

For these experiments, the optimal RT tolerance was found at **6 min (+/- 3 min)** for both software tools.

More generally, the RT tolerance used for data extraction must be driven by the quality of the RT prediction, which should be evaluated empirically by visually inspecting a range of peptides throughout the gradient. The RT tolerance should allow the extraction of correct chromatographic peaks with a sufficient specificity to avoid as many interfered peaks as possible.

### A.3.Extraction window

In DIA acquisition modes, MS/MS spectra are very complex as they are composed of fragments coming from all peptides co-isolated in the pre-defined Q1 windows. Using the high resolution of the analyser (ToF or orbitrap), the targeted fragments are extracted within a finely tuned m/z extraction window.

During the analyses, the Triple TOF 5600+ reached an average of 23 000 resolving power in MS/MS in high sensitivity mode, corresponding to a resolution or Full Width at Half Maximum (FWHM) of 43 ppm. While we used to extract the MS/MS signals with wide m/z windows to maximise the sensitivity (by default 2 x FWHM ≈ 80 ppm in Skyline, which is recommended by SCIEX as well), narrower m/z extraction windows allowing better specificity were evaluated, down to 15 ppm (**Table 6**).

***Table 6 : m/z extraction window optimisation for targeted DIA data extraction.***
*The true positive rate (TPR) and false discovery proportion (FDP) are presented for each parameter and each software tool.*
*The resulting optimal setting is presented in bold red.*

| Parameter | Value | Skyline | | Peakview | |
|---|---|---|---|---|---|
| | | TPR (%) | FDP (%) | TPR (%) | FDP (%) |
| Extraction window | 15 ppm | 78 | 19 | 71 | 27 |
| | **40 ppm** | 84 | 17 | 84 | 19 |
| | 60 ppm | 86 | 18 | 79 | 21 |
| | 80 ppm | 80 | 20 | 82 | 20 |

We found that an m/z window of **40 ppm** was the best compromise between sensitivity and specificity, resulting in higher TPR and lower FDP. Actually, this 40 ppm extraction window corresponds to the MS/MS **resolution** (**1 x FWHM**). This result is in accordance with Skyline developers finding[29] who implemented a checkbox “Use high-selectivity extraction” in Skyline to reduce the default m/z extraction window from 2 x FWHM to 1 x FWHM. However, contrary to their results, we found that centroiding DIA data did not improve the quantification: using 10 ppm mass accuracy, we obtained 53% TPR and 13% FDP.

It is of note that if using a narrow m/z window improves the quantification performances, the mass accuracy of the instrument becomes an even more crucial parameter. On Triple TOF systems, the mass accuracy is ensured using calibration runs which consist in a LC-MS/MS analysis of a standard sample from which the instrument will be recalibrated in MS and MS/MS using known ions. Between the calibration runs, the mass accuracy is highly affected by temperature changes. For DDA data, the Protein Pilot software (SCIEX) can perform a mass recalibration using high confidence identifications to determine the mass shift and compute a mass correction, but this is not performed for DIA data.

### A.4.Number of transitions

The transitions present in the spectral library that will be extracted for each peptide are chosen by the software tool based on their relative intensity in the spectral library and on several expected quality filters, like a minimal length for the fragment ions to avoid short ions which are not specific and prone to be interfered, or their presence in the Q1 isolation window which also induce potential interference by unfragmented precursor ions.

A range of number of transitions to use per peptide was evaluated, from 3 to 10 transitions per peptide (**Table 7**).

***Table 7 : Optimisation of the number of extracted transitions per peptide for targeted DIA data extraction.***
*The true positive rate (TPR) and false discovery proportion (FDP) are presented for each parameter and each software tool.*
*The resulting optimal setting is presented in bold red.*

| Parameter | Value | Skyline | | Peakview | |
|---|---|---|---|---|---|
| | | TPR (%) | FDP (%) | TPR (%) | FDP (%) |
| # transitions | 3 | 83 | 25 | 61 | 29 |
| | **6** | 80 | 20 | 82 | 20 |
| | 10 | 80 | 19 | 74 | 18 |

The optimal number of transitions to use for each peptide was found to be **6**. Indeed, each extracted transition provides additional data for an optimum peak integration. However, it is rare to find more than 6 informative transitions per peptide, and using too many transitions per peptide can be deleterious because extracting low intensity transitions increases the probability to extract interfered signals.

### A.5.FDR threshold

Using DIA, thousands of peptides and proteins can be analysed, and it is not realistic to check all quantified transitions. Therefore, statistic tools are needed to help in keeping correct peaks while removing wrong peaks. Following data extraction, each peak is scored according to several quality attributes, like the peak intensity, the RT precision, the correlation between the relative intensity of the transitions in DIA data and in the spectral library (dot product), the peak shape, the co-elution of the transitions, the signal to noise ratio, or the m/z accuracy. We used a target decoy approach to determine a false discovery rate (FDR) for each peak based on its score. While the peak scoring model is fixed in Peakview, the mProphet peak scoring model in Skyline can be trained for each dataset[54].

A range of FDR thresholds were compared using both software tools, from 0.5 to 5%, with a control without FDR threshold (**Table 8**).

***Table 8 : FDR threshold optimisation for targeted DIA data extraction.***
*The true positive rate (TPR) and false discovery proportion (FDP) are presented for each parameter and each software tool. The resulting optimal setting is presented in bold red.*

| Parameter | Value | Skyline | | Peakview | |
|---|---|---|---|---|---|
| | | TPR (%) | FDP (%) | TPR (%) | FDP (%) |
| FDR threshold | NA | 78 | 19 | 83 | 26 |
| | 5 % | 80 | 20 | 82 | 22 |
| | **1 %** | 80 | 20 | 82 | 20 |
| | 0.5 % | 78 | 19 | 82 | 19 |

While the use of an FDR threshold showed no major effect using Skyline, it significantly reduced the FDP using Peakview. We therefore decided to comply with the usual **FDR threshold of 1%**.

It is important to note that Peakview FDR application strategy is not common: applying an FDR threshold of 1% in the software settings means that a peptide for which a peak was confidently detected (FDR < 1%) in one analysis of a dataset is automatically considered as correctly detected in all analyses of the dataset. However, even for technical replicates, the peak integration is different between analyses, and therefore the FDR threshold should be applied separately to each individual analysis and this is what we did for Skyline results.

Furthermore, it should be kept in mind that the target decoy strategy is not optimal for all type of samples, e.g. samples with low number of targets like purified samples[118], and additional quality control strategies like a dot product threshold should be used.

### A.6.Software tool

We evaluated the open source and freely available Skyline[52] (MacCoss Lab of Biological Mass Spectrometry, university of Washington) software tool, and the MS/MS$^{ALL}$ with SWATH Acquisition MicroApp 2.0 within the proprietary Peakview (SCIEX) software tool. The parameters optimised above were used to extract DIA data using both software tools, and their quantification performances were compared (**Figure 48**).

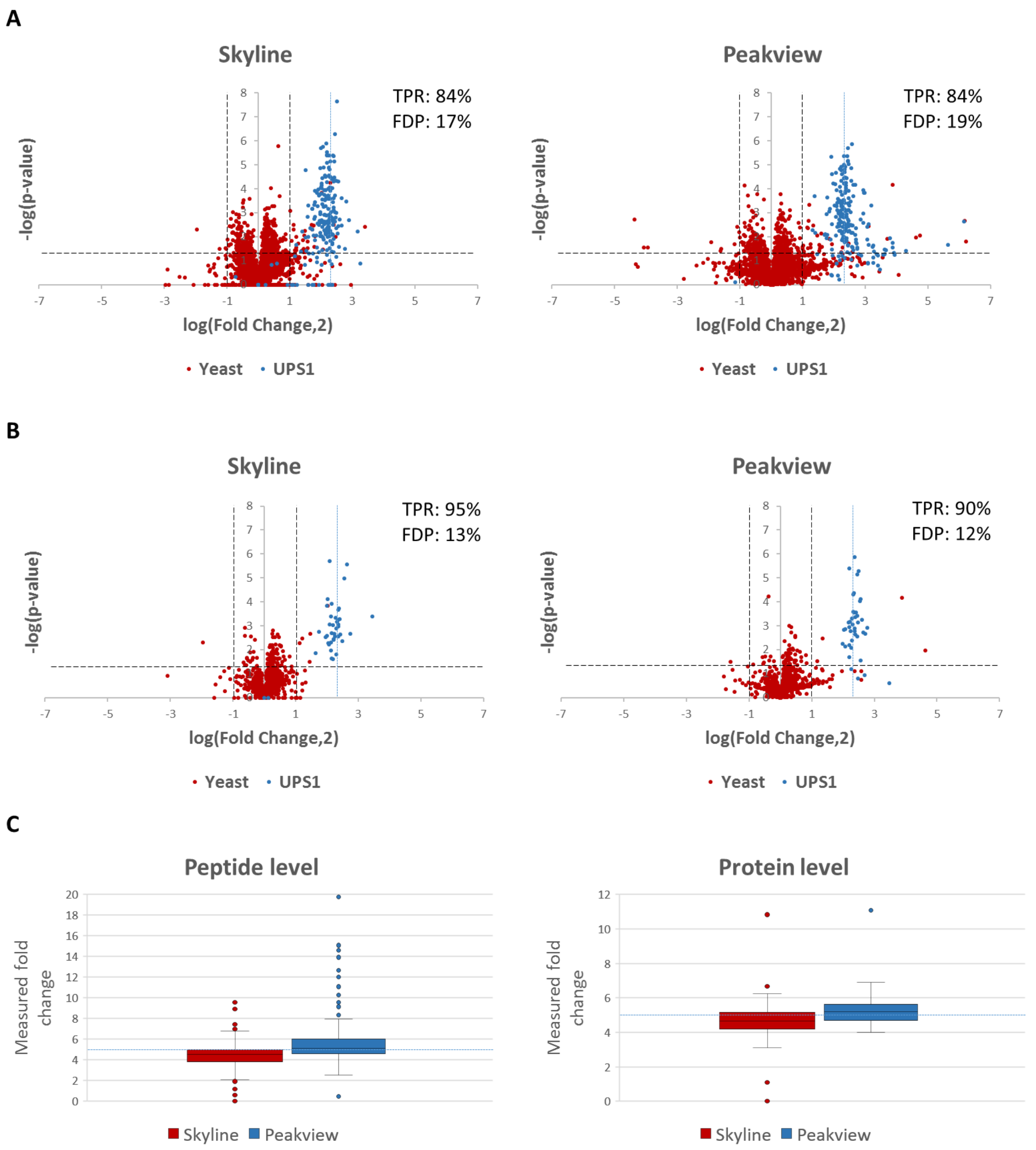


***Figure 48 : Evaluation of Skyline and Peakview software tools for targeted DIA data extraction.***
*A and B. Software tools were operated using the previously optimised parameters, and volcano plots were built for peptides (A) and proteins (B). Fold change and p-value thresholds used for TPR and FDP calculation are presented as black dotted lines, and the expected fold change is presented as a blue dotted line. Red dots represent yeast peptides and proteins, and blue dots represent UPS1 peptides and proteins. C. The accuracy of both software tools was assessed at the peptide and protein levels by comparing the measured fold change of UPS1 peptides and proteins to the expected fold change presented as a blue dotted line. Two outlier values are not represented for Peakview at the peptides level for visual concern: 50 and 71.*

Overall, both software tools perform equivalently in terms of sensitivity (TPR) and specificity (FDP), with TPR at 84% and FDP at ≈ 18% at the peptide level, and TPR at ≈ 93% and FDP ≈ 13% at the protein level. In terms of accuracy, both software tools again perform very similarly, which is consistent with a recent benchmarking of software tools for DIA-SWATH data interpretation[29].

However, a range of differences can help choosing between Skyline and Peakview for DIA data extraction: overall, Skyline allows the user to finely tune each step of the quantification, while Peakview is more straightforward since less control and information are shared with the user. Skyline is compatible with a wide range of search results file formats, and can perform individual retention time normalisation of multiple files to build a unique spectral library, allowing its enrichment with newly acquired DDA data even if different LC gradients were used. On the contrary, Peakview does not allow importing multiple files to build a spectral library. Even if it can be a file merged from multiple analyses, since Peakview does not perform any retention time normalisation, only DDA data obtained with the same LC gradient should be used to build a spectral library. In addition, the way of choosing the targeted transitions and the peak scoring model are fixed in Peakview, while they are adjustable in Skyline. One of the most important difference between both software tools is that Skyline allows manual peak picking curation, which is crucial especially for peptides of interest. Moreover, the Skyline interface allows displaying multiple analyses together, with quality controls like the retention times or the peak areas views to help signals inspection and peak boundaries curation. Using Peakview, only one analysis can be seen at a time and no quality controls are displayed which renders the data reviewing more challenging, and peak integration curation is not permitted. Finally, the Peakview FDR application strategy did not satisfy us (see A.5).

For these reasons, we preferred using **Skyline** to process DIA data.

### A.7.Conclusion

Using a well calibrated standard sample, we benchmarked targeted data extraction parameters using two software solutions. These comparisons allowed us to further understand each parameter in order to better use them and highlight guidelines. Therefore, the following settings will be used for subsequent DIA data analysis: Skyline software tool, empirically determined RT tolerance, an m/z extraction window equal to the MS/MS resolution, 6 transitions per peptide and an FDR threshold of 1%.

Using these optimised settings, we reached a TPR of 95% and an FDP of 13%, which is equivalent to results obtained using a similar sample but quantified by several MS1 XIC label free data analysis tools[129]. However, DIA should prove its superior sensitivity using lower UPS1 spikes amounts, and it is well-known that there is a big room for improvement in automatic peak picking for DIA data.

It should also be kept in mind that these evaluations were performed on a limited number of peptides and proteins, and using a larger number of compared values, e.g. spiking entire proteomes, could improve the statistical significance of such evaluations[29].

### A.8. Perspectives

The data analysis remains today the major bottleneck for DIA approaches, and further developments should focus on (i) an improved retention time (RT) prediction and (ii) a better interference management.

An improved RT prediction could allow a higher RT stringency to help software tools finding the correct peaks. Indeed, RT alignment strategies using a limited number of RT standards and linear regressions, which are used in Skyline and Peakview, were initially developed for SRM assays, for which approximate retention time prediction was sufficient[204]. Linear regressions for RT alignment are optimal only when equivalent LC gradients (different linear gradients or identical nonlinear gradients) are used for both spectral library generation and DIA analyses. Alternatively, nonlinear regression methods like a segmented regression as well as the use of a larger number of anchor peptides can improve RT alignment[164].

Even if DIA is more specific when compared to DDA, it still suffers from interferences when analysing very complex samples. Significant efforts are currently made to detect or even remove interferences[162, 205-207].

## B. *Spectral library*

In the peptide-centric approach, the spectral library is an essential tool which will provide the information that are necessary to extract a list of peptides in a targeted manner into DIA data, including peptides and fragments m/z, their retention time, and the relative intensities of the fragments.

A spectral library is built using previous DDA data, which can have been obtained in-house or from other laboratories, implying the existence of two types of spectral libraries: homemade spectral libraries and publicly available spectral libraries generated by others. A homemade spectral library can be built using the same coupling and LC gradient as the one used for DIA analyses, providing ideal conditions notably for RT alignment. However, a homemade spectral library is limited in proteome coverage to what was previously detected in DDA mode by the analyst. Alternatively, several spectral libraries are available in public repositories for some reference taxonomies such as human[53, 165] or yeast[166]. Such spectral libraries can be extended anytime when new data become available, quickly surpassing the proteome coverage of a homemade library besides saving time and money.

To evaluate the quantification performances of a publicly available spectral library, we used a more complex sample, a HepaRG human cell line protein extract available in the lab. First, we built a comprehensive spectral library using SDS-PAGE protein fractionation in 27 bands. The proteins were in-gel digested using trypsin, and retention time standards were spiked into the peptides extract. The 27 bands were analysed on a nanoLC-Triple TOF 5600 coupling in DDA mode. Peptides and proteins were identified using Mascot search engine, and validated using the Proline software (1% FDR at the peptides and proteins levels), resulting in a spectral library composed of 29 210 proteotypic peptides corresponding to 3 619 proteins.

The HepaRG protein extract was also purified using stacking gels to provide an unfractionated sample for DIA analysis. As previously, the proteins were in-gel digested using trypsin and retention time standards were added to the resulting peptides. The sample was analysed on the same coupling using a 34 x 25 Da windows DIA-SWATH method.

DIA data were extracted as optimised previously in Skyline using either a homemade spectral library built from DDA data of 27 SDS-PAGE bands, containing 30 982 validated peptides corresponding to 3 644 proteins, or the free-of-access human spectral library available in SWATHAtlas[53], combining results from 331 analyses of fractions from different cell lines, tissues and affinity enriched protein samples[53], containing 139 449 peptides corresponding to 10 316 proteins. Retention time alignment was performed using retention time standards (iRT, Biognosys, Zurich, Switzerland), and the proteome coverage allowed by both spectral libraries was compared (**Figure 49**).

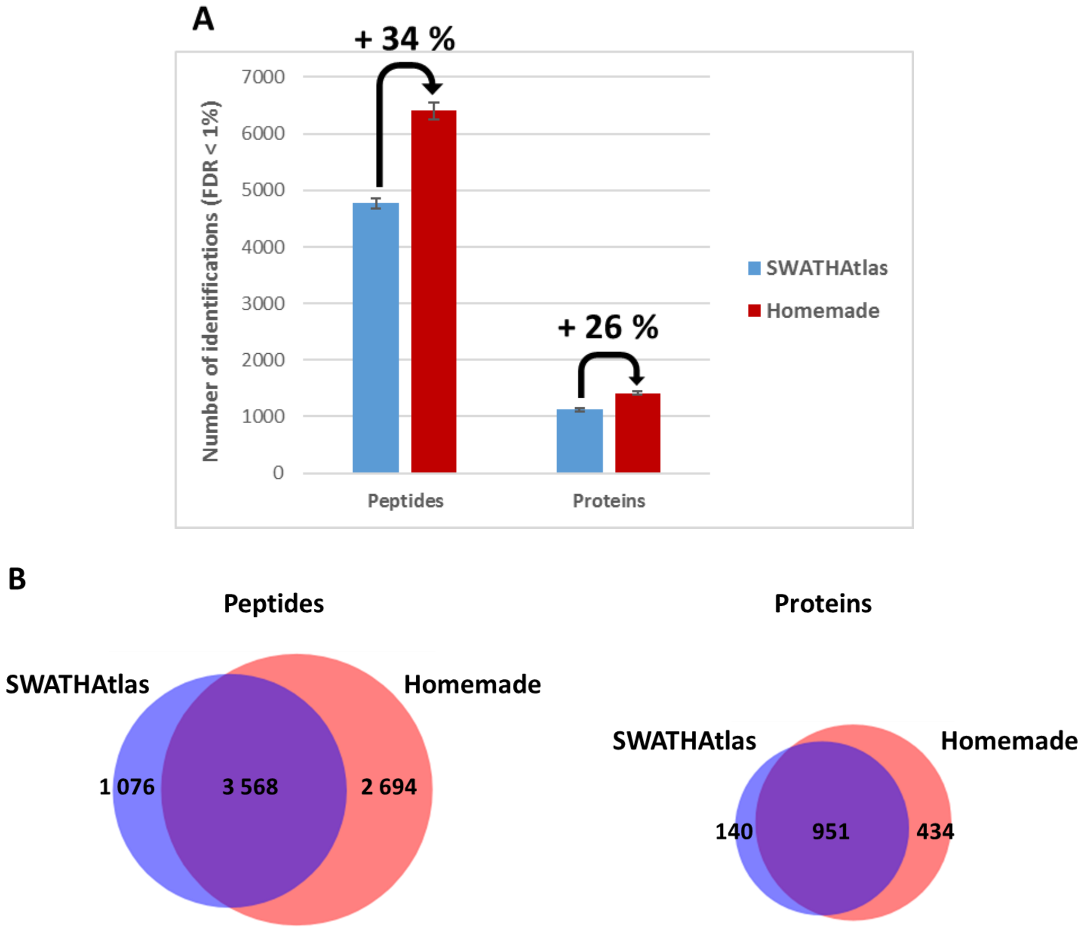


***Figure 49 : Comparison of identification performances between a homemade spectral library and the publicly available human spectral library from SWATHAtlas.***
*A. The number of identified peptides and proteins using both spectral libraries are displayed. B. Venn diagrams were realised for peptides and proteins identified in at least two out of three technical replicates.*

The homemade spectral library allowed the identification of 6 403 peptides and 1 413 proteins, while the SWATHAtlas library allowed the identification of 4 767 peptides and 1 123 proteins, with 49 and 62% overlap at the peptides and proteins levels, respectively. Thereby, the homemade spectral library allowed the identification of 34% more peptides and 26% more proteins when compared to the SWATHAtlas spectral library. The most indicated reason for the lower performances of the public spectral library is linked to the RT alignment. Indeed, the homemade spectral library has the advantage to have been built on the same nanoLC-MS/MS coupling as were performed the DIA analyses with spiked in RT standards, providing ideal conditions for RT prediction in DIA analyses. On the contrary, the public spectral library has been built on different LC-MS/MS couplings, rendering the RT alignment more challenging. This hypothesis was tested by extracting exactly the same peptides using both spectral libraries, i.e. only peptides that were common between the SWATHAtlas and the homemade spectral library (**Figure 50**).

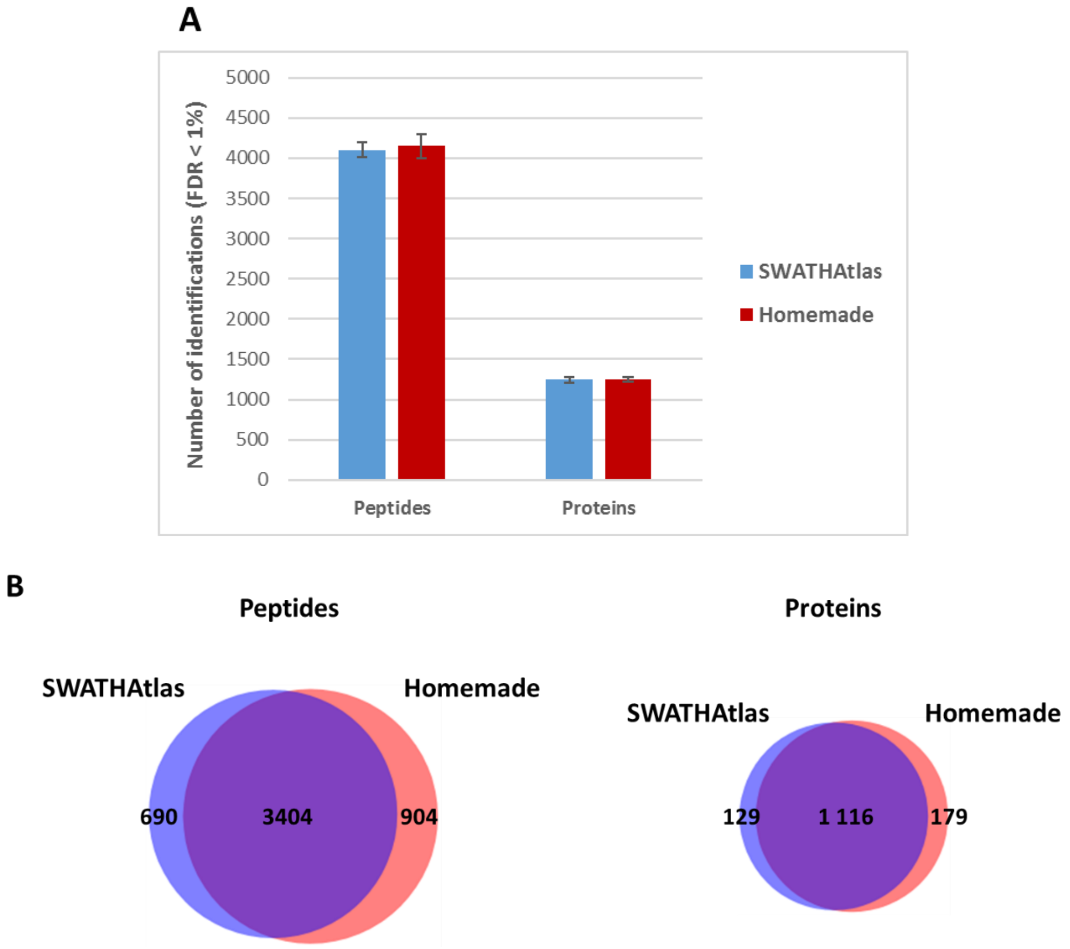


***Figure 50 : Comparison of identification performances between the homemade and the SWATHAtlas spectral libraries when extracting only common peptides.***
*A. The number of identified peptides and proteins using both spectral libraries are presented. B. Venn diagrams were realised for peptides and proteins identified in at least two out of three technical replicates.*

Surprisingly, data extraction of common peptides led to equivalent identification performances, with ≈ 4 100 peptides and ≈ 1 250 proteins identified using both spectral libraries, with a major overlap of 68 and 78% at the peptides and proteins levels, respectively. This result means that the information (normalised retention time and relative intensity of the fragments) provided by the publicly available spectral library are of comparable quality with the homemade spectral library, and when RT standards are used, the RT alignment performs quite well even for a public spectral library. The RT alignment is thus not the reason why the public spectral library performed less well against the homemade spectral library. Another major difference between these libraries is the number of peptides and proteins in each, which could potentially lead to issues for the target decoy validation step at 1% FDR using the public spectral library. Therefore, we compared the target and decoy score distribution obtained using the mProphet peak scoring model in each case, when extracting all peptides or only common ones, for both the homemade and the public SWATHAtlas spectral libraries (**Figure 51**).

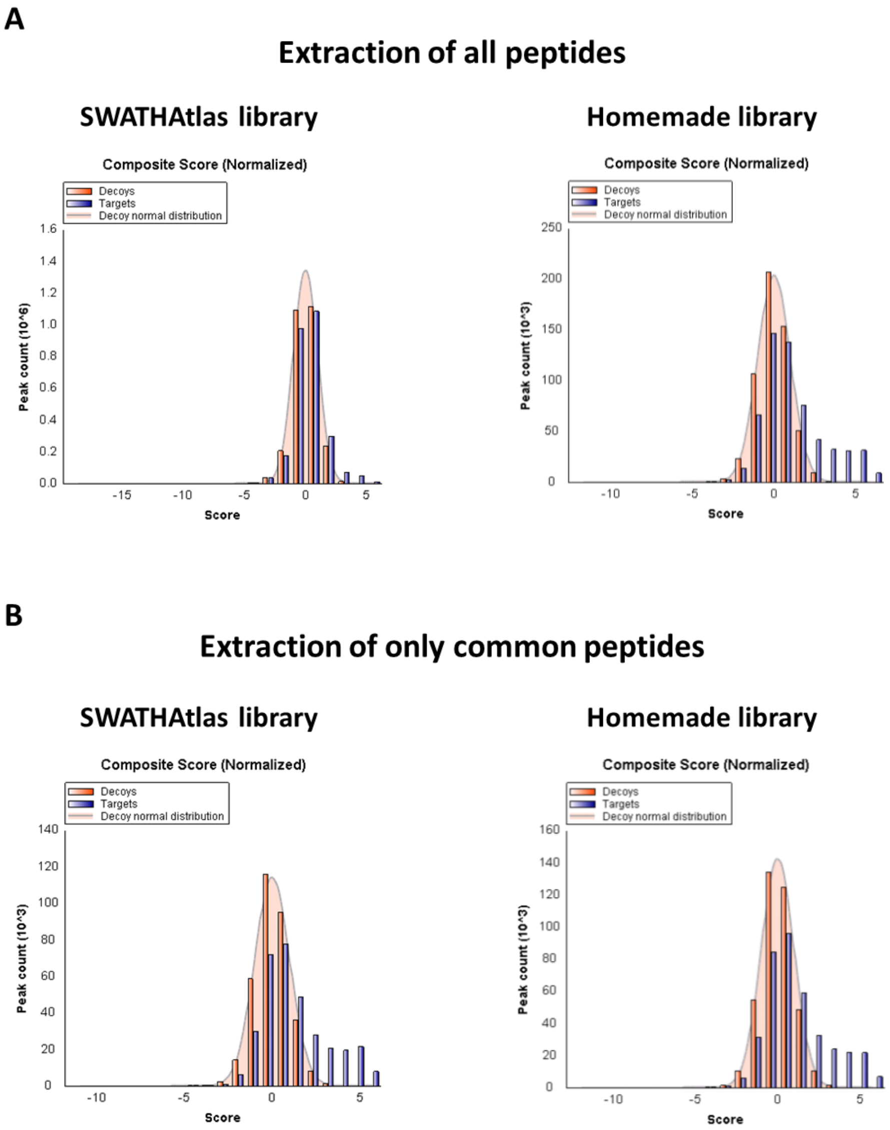


***Figure 51 : Target and decoy score distribution obtained using the mProphet peak scoring model in Skyline.***
*The targets are represented in blue bars, the decoys are represented in orange bars, and the orange curve represents the decoy distribution. A. The targets and decoys score distribution is presented when **all peptides** present in SWATHAtlas or the homemade spectral library were extracted. B. The targets and decoys score distribution is presented when **only common peptides** were extracted.*

When all peptides are extracted, a clear difference can be seen in the target and decoy score distribution between both spectral libraries. While targets can be partially isolated from decoys using the homemade spectral library, target and decoy populations cannot be distinguished when using the SWATHAtlas spectral library. In fact, the target decoy approach used for the peptide-centric approach for DIA data extraction is quite different compared to the target decoy approach used for protein identification validation from DDA data. For DDA data, MS/MS spectra are searched against a database containing both targeted protein sequences and decoy protein sequences. The number of MS/MS spectra matching a decoy peptide is used to estimate a false discovery rate (FDR). However, in our case of analysing DIA data using a peptide-centric approach, data are not searched in a target decoy database, but targets and decoys are searched into the data. Thereby, when using a homemade spectral library, only peptides that were previously identified in a similar sample will be searched for, together with their corresponding decoys. Targets will be generally well picked by the extraction software, and will be attributed a good score, while decoys will not be found and, in majority, background noise will be integrated. This will allow separation of targets and decoys populations by the peak scoring model. However, the majority of the peptides and proteins present in the public SWATHAtlas spectral library (139 449 peptides corresponding to 10 316 proteins) will not be effectively present in the sample. Thereby, the majority of the targeted peptides will be integrated in the background noise, just like the decoy peptides, leading to indistinguishable target and decoy populations, and rendering the FDR estimation inefficient. This is a major issue for public spectral libraries, because their huge size, which was their strength, is also their weakness.

In conclusion, even if the data present in publicly available spectral libraries seems of very good quality, there is an issue with the number of targets in such very large spectral library, and there is a need to adapt the size of the spectral library to what can be present in the analysed sample. It can be performed by extensively characterise the sample in DDA mode prior to DIA analysis but this is the same as building a homemade spectral library. Therefore, today the most reliable workflow for DIA data extraction still relies on the use of a **homemade spectral library** which is representative of the analysed samples. However, the use of subset of public spectral library, e.g. containing only proteins that can be found in the cell type of our sample could be investigated. Alternative scoring and false discovery estimation strategies for the peptide-centric approach could also be investigated.

## II. Data acquisition

Even if DIA method do not need extensive optimisations like DDA or targeted methods, several parameters should not be underestimated, like the definition of the cycle time, the accumulation time, the number and size of the windows. These parameters should not be underestimated as they will determine the sensitivity and specificity of the assay (**Figure 52**).

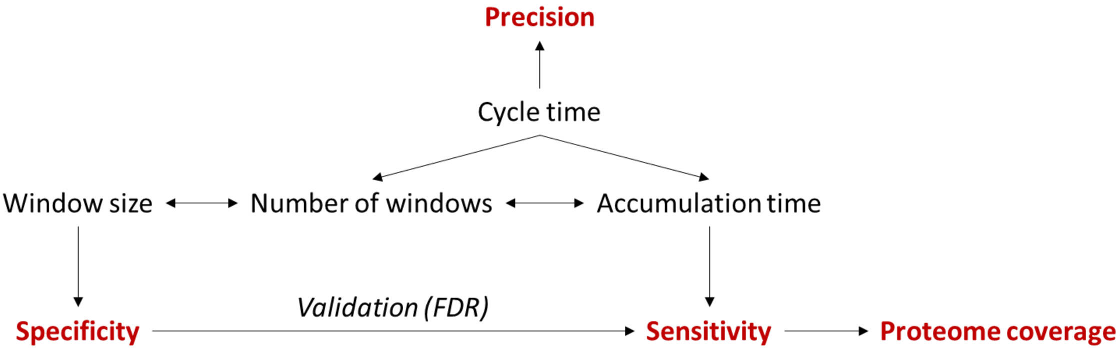


***Figure 52 : Relationship between the key parameters of DIA data acquisition.***
*As described for DDA method, the cycle time conditions the precision of the quantification. It should be defined as ≈ 1/10 of the average chromatographic peak duration to allow precise quantification. The cycle time sharing is a compromise between the number of windows and their respective accumulation time. The accumulation time is directly related to the sensitivity of the assay, as an increased accumulation time leads to an increased signal / noise ratio. The number of windows within a defined m/z range will determine their size, and the window size will condition the number of co-isolated precursor ions and therefore the specificity of the assay. A good specificity will lead to a reduction in interferences and a better signal quality, and due to subsequent scoring and validation (false discovery rate), a better specificity will lead to a better sensitivity and ultimately a better proteome coverage.*

It is of note that even if the quadrupole transmission windows are nearly squared shape, the extremity of each isolation window do not provide optimal ion transmission. Therefore, an overlap between windows should be used, and typically it is set to 1 m/z, and data are not extracted in the 0.5 m/z at the border of each window (**Figure 53**).

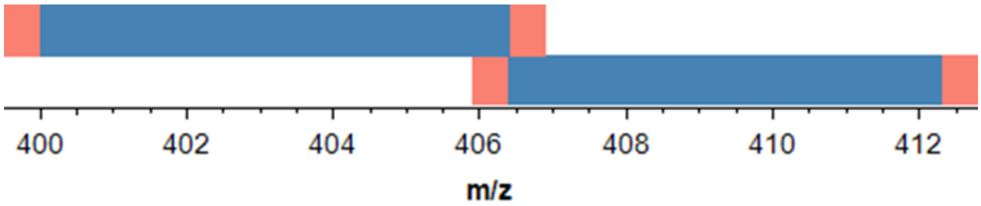


***Figure 53 : Overlap between Q1 isolation windows.***
*In this example, an overlap of 1 m/z was defined, and data were not extracted in the 0.5 m/z at the border of each window (red regions). This way, extracted data are of optimal quality (blue regions).*

Contrary to DDA mode for which the applied collision energy is calculated for each precursor ion, in DIA mode the collision energy is usually calculated for a doubly charge precursor ion at the centre of the isolation window. Therefore, the collision energy spread (CES), which was found to be of limited impact for DDA analysis (see Chapter IIII.C.2), should have more impact on DIA analysis.

In this work, I compared several Q1 isolation windows setups: first, we compared the effect of different windows sizes keeping constant the cycle time, and then we evaluated the use of variable windows aiming to better split the peptides along the m/z range and increase the specificity of the assay.

### A. *Number and size of the isolation windows*

For a defined m/z range, the number and the size of the windows are related: the more windows, the smaller they are. Reducing the size of the isolation windows decreases the number of co-isolated precursor ions and therefore increases the specificity and the sensitivity of the assay. However, to analyse a given m/z range, reducing the windows size induce an increase in the number of windows and therefore either an increased cycle time, which reduces the precision of the quantification because chromatographic peaks are not well defined, or a decreased accumulation time, which reduces the signal / noise ratio and therefore the sensitivity of the assay.

To evaluate the effect of the windows setup on DIA data, the HepaRG human cell line protein extract was prepared as described previously (see I.B), and was analysed on the nanoLC-Triple TOF 5600 coupling using either a 34 x 25 Da windows method with 3 sec cycle time and 90 ms for MS/MS accumulation time, or a 68 x 12.5 windows method with 3.5 sec cycle time and 50 ms for MS/MS accumulation time. The data were extracted in Skyline using the optimised parameters and the homemade spectral library, and the number of identifications was compared between both methods (**Figure 54**).

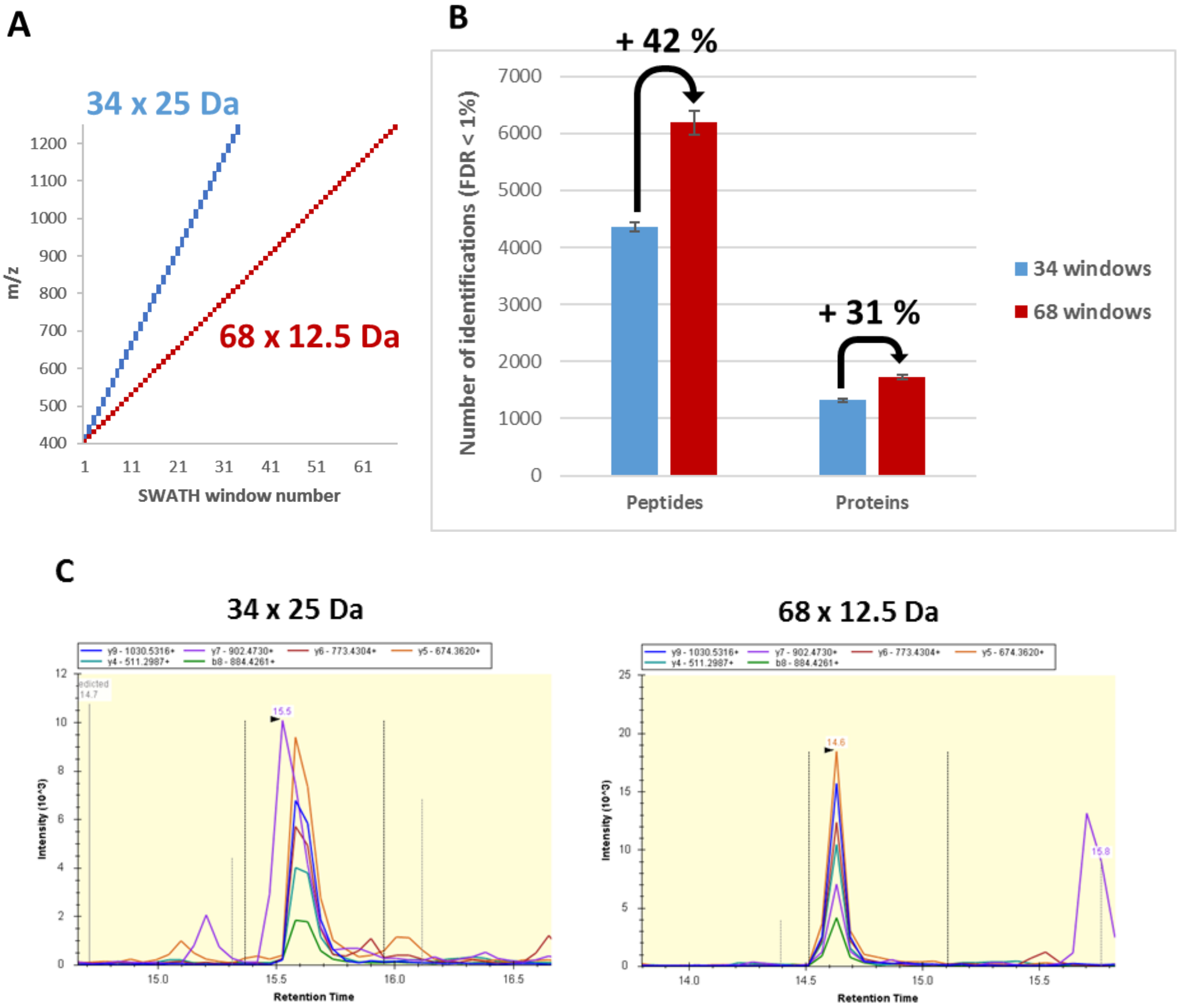


*Figure 54 : Evaluation of different number and size of isolation windows.*

*A. The coverage of the 400-1250 m/z range by the SWATH windows is displayed for the 34 x 25 Da windows and the 68 x 12.5 Da windows methods. B. The number of identified peptides and proteins using both acquisition methods are displayed. C. Example of a peptide interfered using the 34 x 25 Da windows method, which is not interfered using the 68 x 12.5 Da windows methods.*

While the 34 x 25 Da windows method allowed the identification of 4 360 peptides and 1 319 proteins, the **68 x 12.5 Da windows** method identified 6 185 peptides and 1 725 proteins. The 68 x 12.5 Da windows method thus allowed the identification of 42% more peptides and 31% more proteins compared to the 34 x 25 Da windows method. Between both methods, the cycle time was kept approximately constant for a ≈ 3 sec cycle time as the collection of ≈ 8-10 points per peak is necessary for a good chromatographic peak shape and precise quantification[162]. These results show that the reduction of the windows size, leading to a better specificity and less interfered signals, was highly beneficial for peptide identification. The decrease in sensitivity due to the accumulation time reduction from 90 ms to 50 ms was here largely compensated by the gain in specificity. Furthermore, 50 ms accumulation time is still comfortable, and the loss in sensitivity may be more problematic for lower accumulation time, e.g. 20-30 ms.

### B. *Variable isolation windows*

The peptides population m/z is not equally distributed along the m/z range. By using a fixed window size splitting the whole m/z range in equal m/z fractions, several windows will be crowded while others will be nearly empty (especially high m/z regions). Since the more co-isolated peptides, the more interfered signals (see **Figure 54**), it is interesting to adapt the window sizes to the peptides m/z repartition, in order to reduce interferences in crowded m/z regions. The use of variable windows, i.e. different window sizes along the m/z range, can equalise the peptides repartition among isolation windows to achieve better specificity and therefore sensitivity in high density m/z regions. Thereby, smaller windows will be used in high density m/z regions and wider windows will be used in low density m/z regions.

With SWATH 2.0, a variable windows method can be optimised for a sample type using a DDA analysis of a representative sample. DDA data will be used to probe the peptides repartition along the m/z range. Using the SWATH Variable Window Calculator (available in SCIEX website), the m/z range can be divided into windows containing equivalent peptides density, according to user-defined settings for the number of isolation windows, the analysed m/z range and the overlap between windows (**Figure 55**).

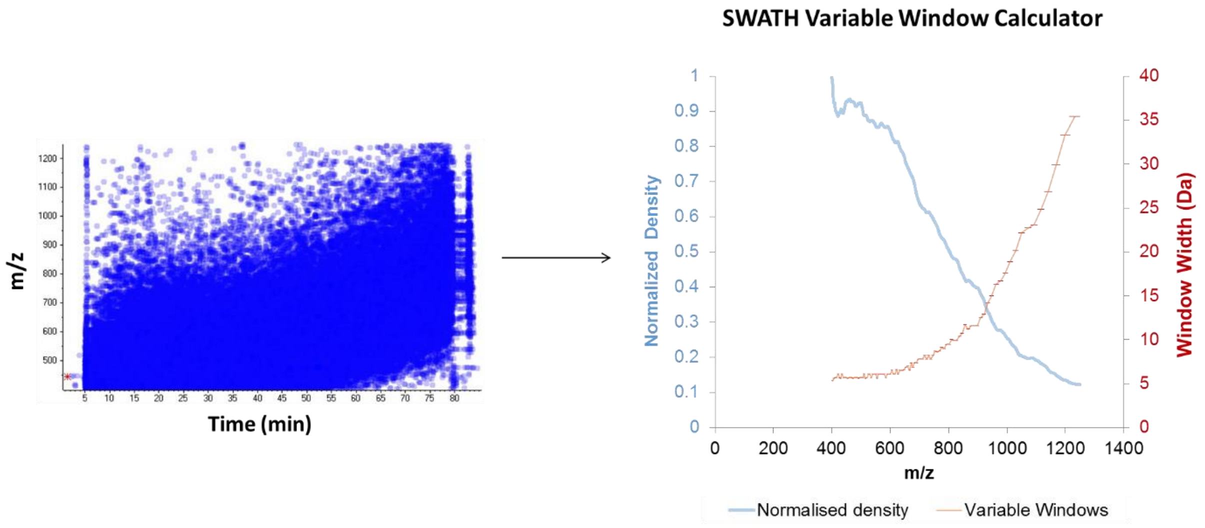


***Figure 55 : Generation of a variable window SWATH method from DDA data.***
*First, a DDA analysis of a representative sample is performed to probe the peptides density over the m/z range. At left, the peptides are represented according to their m/z and retention time. Then, the SWATH Variable Window Calculator uses the peptides density information to generate an optimal isolation window scheme, which can be directly imported into the acquisition method.*

The use of variable isolation windows was evaluated using a K562 human cell line protein extract digest, which is part of the recently introduced quality control kit for Triple TOF systems. After having spiked retention time standards, the digest was analysed on a microLC-Triple TOF 6600 coupling using either a 100 x 8.5 Da windows or a 100 variable windows (from 5 to 49 Da, optimised by SCIEX) setup. Data were extracted with the optimised parameters using the spectral library provided by SCIEX, containing 56 777 proteotypic peptides corresponding to 6 840 proteins, and the number of identifications using both acquisition methods was compared (**Figure 56**).

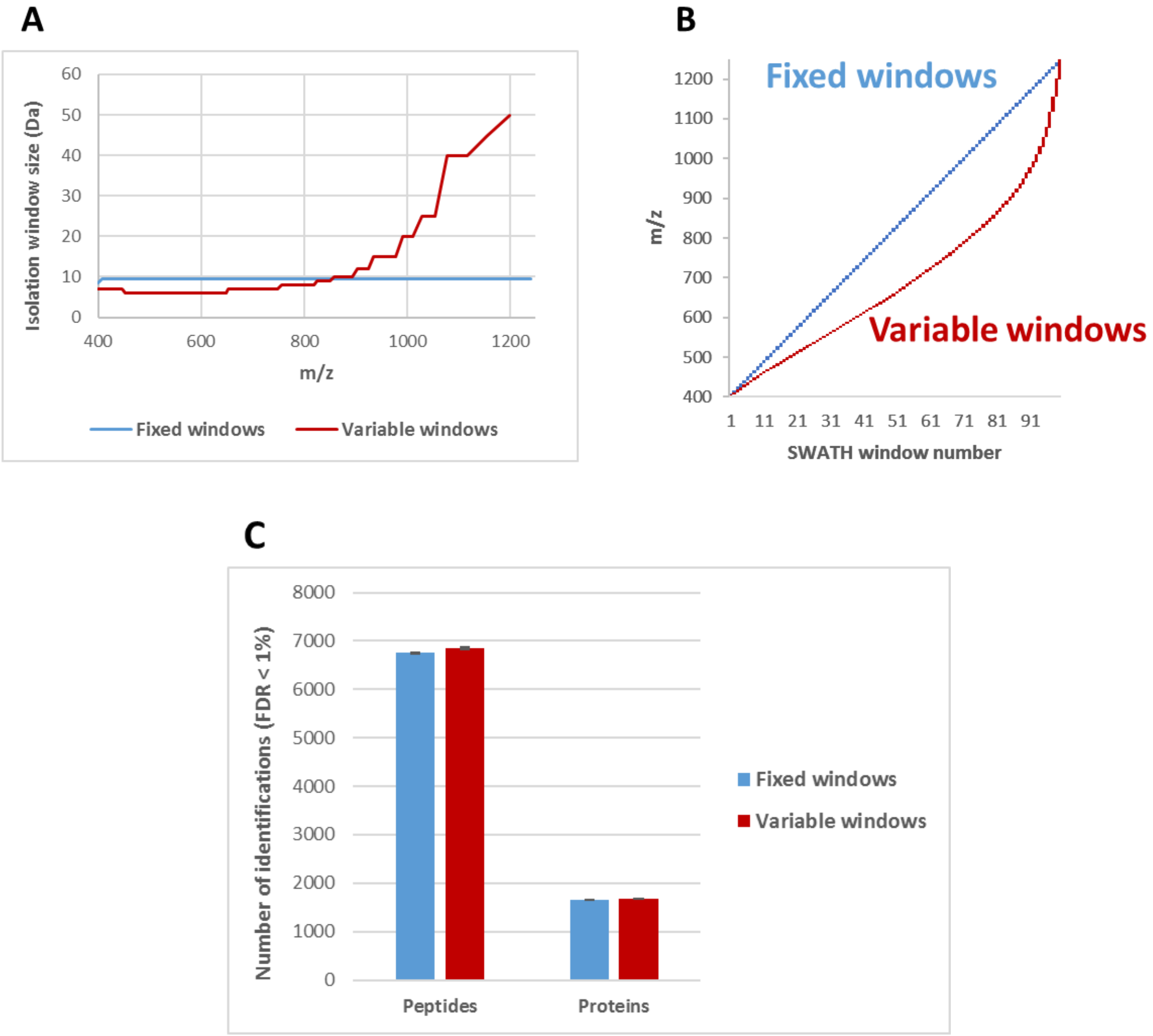


***Figure 56 : Evaluation of the use of variable SWATH windows.***
*A. The optimised window scheme is presented, with the window size plotted against the m/z range. B. The m/z coverage of each acquisition method is presented. C. The number of identified peptides and proteins using both acquisition methods are displayed.*

The fixed and variable windows method showed equivalent proteome coverage, allowing the identification of ≈ 6 800 peptides and ≈ 1 660 proteins (+ 1% more peptides and proteins using the variable windows method). This poor improvement can be due to the high number of isolation windows of both methods. Indeed, using the fixed windows method, the windows are 8.5 Da wide, while using the variable windows method they ranged from 5 to 49 Da, and it seems that the gain in specificity was not sufficient to be translated into a gain in sensitivity. In conclusion, the use of variable windows should have a more appreciable effect for acquisition methods with a lower number of windows, e.g. when a higher accumulation time is preferred.

It is of note that usually, when an acquisition method is optimised for a given sample, its versatility is reduced: the optimised acquisition method is better suited for a given sample, but not necessarily for a different sample. However, we noticed for different proteome samples that usually the lower part of the m/z range (400-850 m/z) is more crowded compared to higher m/z region (850-1250 m/z), which was also observed by others[208], and therefore optimised variable windows methods should not be fundamentally different between different proteomes. The optimised windows scheme becomes different for less complex samples, like partially purified proteins.

### C. *Conclusion*

Using DIA, MS/MS data are collected for all peptides within a defined m/z range, whatever the method settings. However, the data quality is defined by the fine tuning of the DIA window's setup, which will condition the specificity and the sensitivity of the assay and finally the proteome coverage. We showed that using a high number of windows increased the specificity of the assay by reducing interferences, and even if no significant results were observed, an increased specificity can still be reached by using variable windows.

## **III.** Sample preparation

The sample preparation step is often underestimated, but it will greatly condition the sensitivity, accuracy and reproducibility of the assay. For an optimal reproducibility, the sample preparation must be as simple and as fast as possible, because each step can introduce variability. Ideally, the sample must stay unfractionated to prevent any bias in the quantification. Since samples are often available in limited amount, we evaluated the impact of the protein load on a stacking gel and of the peptide load on the column prior to DIA analysis.

### A. *Workflow*

A HepaRG human cell line protein extract was used to optimise the sample preparation step. We loaded from 10 to 100 µg of HepaRG proteins on stacking gels. After in-gel digestion using trypsin, retention time standards were added to the extracted peptides, and 100 ng to 5 µg of samples were analysed on the nanoLC-Triple TOF 5600 coupling in DIA-SWATH mode using the 34 x 25 Da windows method. Previously optimised data analysis parameters were used to extract DIA data using the homemade spectral library.

### B. *Load on stacking gel*

We realised 10 stacking gels with protein load ranging from 10 to 100 µg, but a constant amount of 1 µg of sample was loaded on column. The digests were analysed using a 34 x 25 Da windows method and data were extracted as previously optimised (**Figure 57**).

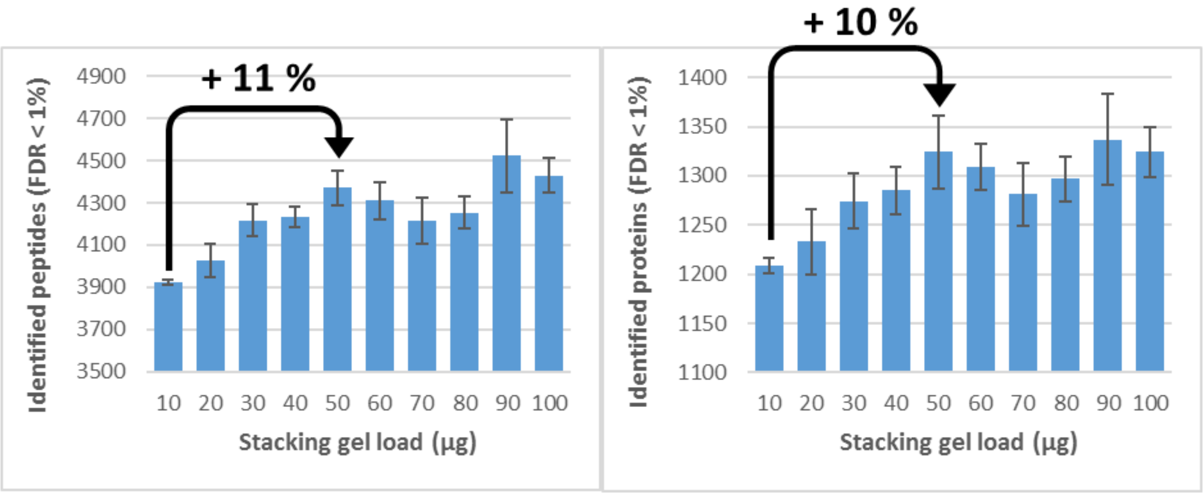


*Figure 57 : Effect of the stacking gel load on proteome coverage.*

The number of identified peptides and proteins slightly increased with increasing load in the stacking gel up to 50 µg, from 3 925 peptides and 1 209 proteins to 4 371 peptides (+ 11%) and 1 324 proteins (+ 10%), for 10 µg and 50 µg respectively. Between 50 and 100 µg loaded on stacking gels, no significant difference in the proteome coverage could be observed. Therefore, at least **50 µg** of proteins should be loaded on the stacking gel to reach optimal sensitivity.

The lower performances reached with lower protein amounts could be explained by the coating of peptides after digestion onto the plastic tubes or 96-well plates. For low quantities, the proportion of coated peptides is high and can induce significant peptide loss. This coating could be reduced by the use of low-binding material.

### C. *Injected amount*

The amount of sample that should be loaded is conditioned by the internal diameter of the LC column. In the laboratory, we mainly use nanoLC systems which are best suited for very low quantity of available sample amount, which is the case for the majority of our projects. For nanoLC systems, we use to load 1 µg of sample onto the column. Here we assessed a range of sample amounts loaded onto the column, from 100 ng to 5 µg (**Figure 58**).

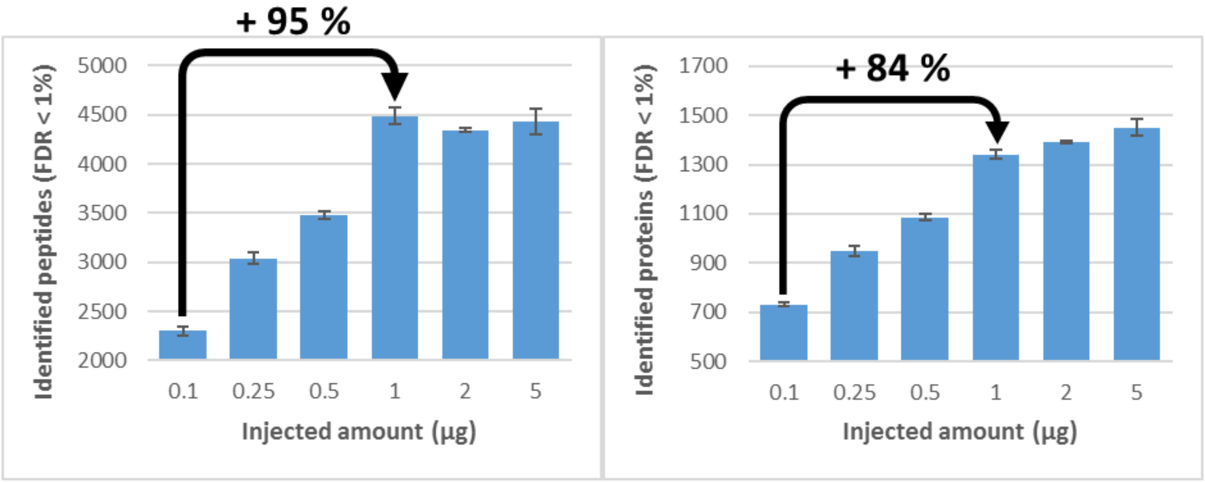


*Figure 58 : Effect of the column load on proteome coverage.*

The more sample was loaded onto the column, the more peptides and proteins were identified up to **1 µg**, from 2 302 peptides and 730 proteins to 4 490 (+95%) peptides and 1 342 (+84%) proteins, for 100 ng and 1 µg, respectively. From 1 µg, a plateau was reached, and loading more than 1 µg did not increase the sensitivity of the assay, but it may damage the column and dirty the mass spectrometer. However, it is of note that loading 1 µg onto the column seems optimal only for nanoLC systems and for this type of sample complexity (cell line protein extract). Indeed, lower sample amounts should be loaded for less complex samples because the total protein amount is divided into less proteins, and in the same way higher amounts may be loaded for more complex samples like metaproteomes. Indeed, what is important is the sample amount that is analysed over time, not the total amount that is injected, and it is true for both the identifications and the instrument stability point of views.

### D. *Conclusion*

The sample preparation is often overlooked, but in fact it is one the most important step which will condition many aspects of subsequent analysis. Indeed, we showed that at least 50 µg of proteins should be loaded on stacking gels to provide an optimal sensitivity. Moreover, if a cell line protein extract is analysed, 1 µg of the resulting peptides should be loaded onto the LC column on a nanoLC coupling to provide optimal sensitivity and material sustainability. Additionally, the sample preparation step will condition the reproducibility of the assay.

## IV. Comparison between DDA and DIA

Compared to shotgun approaches performed using DDA, DIA approaches promise to be more sensitive, specific and reproducible due to the systematic acquisition of complete MS/MS data for all peptides present in the sample. Despite these promising characteristics, the use of DIA remains marginal, mainly due to the very complex and challenging DIA data analysis step. Moreover, it is rapidly evolving and many workflows arise, while shotgun DDA approaches are well-established.

But today, what are the performances of DIA compared to DDA? If DIA is still being developed, can we already obtain better results when compared to DDA? We tried to answer these questions by comparing a classic shotgun DDA analysis to a DIA-SWATH analysis on a yeast sample. We analysed a yeast sample in technical triplicates on a nanoLC-Triple TOF 5600 coupling, either in DDA or in DIA mode. For DDA data, we performed a classic Mascot search, and validated the identifications with Proline at 1% FDR. For DIA data, we performed targeted data extraction with the previously optimised parameters using our homemade spectral library in Skyline, and validated the identifications with mProphet at 1% FDR (**Figure 59**).

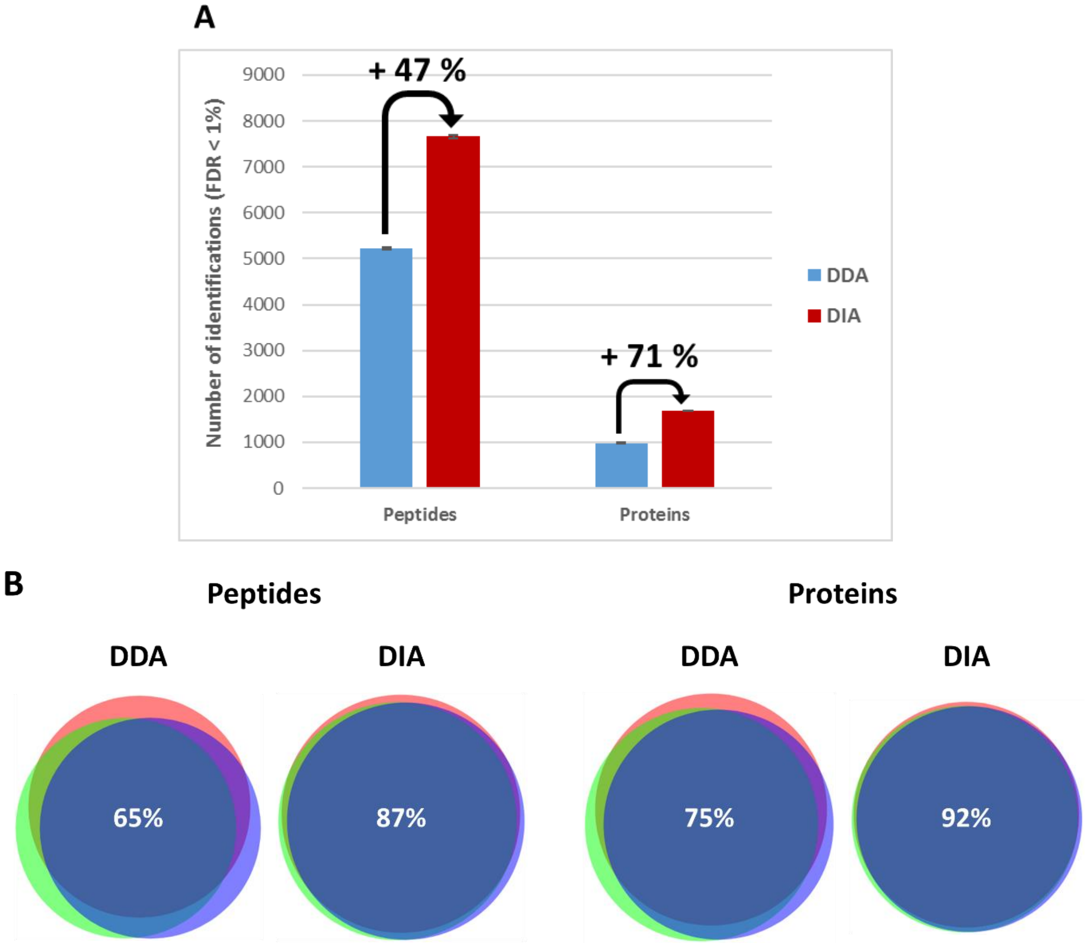


***Figure 59 : Comparison of the identification performances of DDA and DIA modes.***
*A. The number of identified peptides and proteins using DDA and DIA modes are presented. B. The identification reproducibility was evaluated using Venn diagrams for peptides and proteins identified in each technical replicate. The indicated value in each diagram is the percentage of common identifications between the three replicates.*

We showed that for this sample, the identification performances of DIA mode are clearly better than those of DDA mode. Indeed, DIA mode allowed the identification of 42% more peptides and 31% more proteins when compared to DDA mode. Moreover, the identification reproducibility is also better in DIA mode, with a coverage between technical triplicates of 87% for the peptides and 92% for the proteins, compared to 65% for the peptides and 75% for the proteins using DDA mode, showing the undersampling effect is real when using DDA. However, cross identifications are still possible for DDA analysis, using retention time alignment and precursor ion chromatograms to cross-identify a peptide that has been identified in one replicate but not in the other[55].

Then, we wanted to compare the quantification performances of XIC MS1 using DDA data and DIA. DDA data were analysed in Skyline using usual parameters for XIC MS1 quantification: retention time tolerance of +/- 3 min around the retention time that allowed peptide identification, extraction window equal to the resolution (as we observed this led to better performances in I.A.3), and extraction of three isotopes per peptide (P, P+1 and P+2). DIA data were analysed in Skyline as previously optimised. The areas under the curves of the transitions were summed to quantify peptides, and the number of peptides and proteins that were reproducibly quantified with a coefficient of variation below 20% between technical triplicates were compared (**Figure 60**).

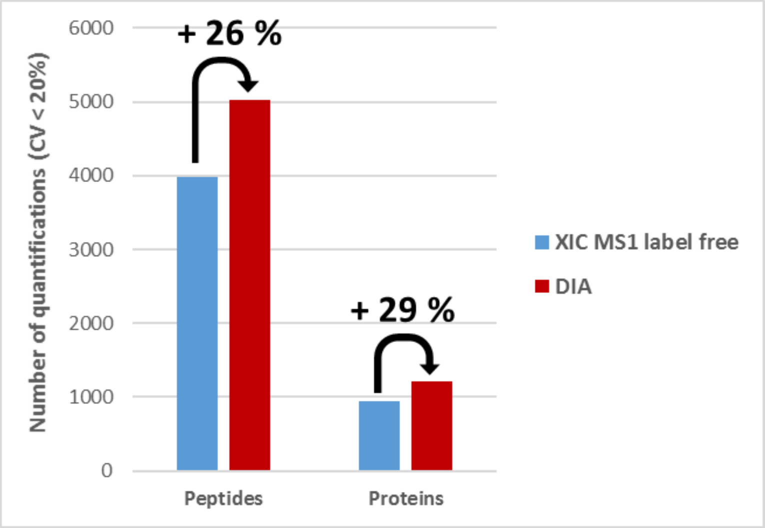


***Figure 60 : Comparison of the quantification performances of DDA and DIA modes.***
*The number of peptides and proteins quantified with a coefficient of variation between technical triplicates below 20% (applied on the peak area values) were compared between XIC MS1 label free (from DDA data) and DIA quantifications.*

These results show that DIA mode allowed the quantification of 26% more peptides and 29% more proteins when compared to XIC MS1 label free quantification. However, the comparison between DDA and DIA modes should be deepened, and for instance comparisons are ongoing on the yeast sample spiked with known quantities of UPS1 proteins presented in I.A.1, but with additional spiked UPS1 amounts, in order to estimate the accuracy and the dynamic range of DDA versus DIA quantifications.

In conclusion, we showed that a DIA-SWATH approach coupled to a peptide-centric data analysis outperformed the classical shotgun workflow based on DDA mode in terms of number and reproducibility of identifications, but also in its ability to reproducibly quantify peptides and proteins. These results are consistent with previous observations[162].

## V. Conclusion

The objective of this part of my PhD was to learn and understand all steps of the whole DIA workflow, in order to highlight best practice and define an optimised DIA-SWATH workflow.

First, the data analysis was the biggest challenge, because it is today the major bottleneck for DIA approaches. Using a well-defined sample, we optimised a peptide-centric data analysis workflow consisting in the use of Skyline for targeted data extraction, with optimal retention time tolerance (empirically determined), extraction window width (equal to the MS/MS peak resolution), number of transitions to use per peptide (6) and false discovery rate threshold (1%). We showed that today the use of publicly available spectral libraries, even if they contain high quality data, is compromised by the too large number of peptides present, which render the use of a peptide-centric target decoy approach inefficient.

Then, we optimised the data acquisition using different windows setup. We showed that using a high number of windows reduced the number of interfered signals by increasing the specificity of the assay, which ultimately led to a significant increase in sensitivity. We could not demonstrate the benefit of using variable windows over fixed windows, but this was probably due to the light differences between both acquisition methods, because of their large number of isolation windows (100) and comparable width. However, the use of variable windows, by linearising the peptide density among the windows, should reduce the interferences in crowded m/z regions (typically 400-850 m/z) and improve the specificity, and in turn the sensitivity of the assay.

The sample preparation step was then assayed with range of sample loadings on stacking gels and on column prior to DIA analysis. We showed that, if possible, at least 50 µg proteins should be loaded on a stacking gel, otherwise the sensitivity is slightly reduced, maybe due to subsequent peptides adsorption on the plastics. Moreover, we showed that for nanoLC couplings, which are the most used in the laboratory, 1 µg of sample should be injected to provide optimal sensitivity while not dirtying the mass spectrometer.

Finally, we compared the optimised DIA workflow to a classical shotgun approach using DDA, and we showed that DIA allows much betters proteome coverage and reproducibility than DDA, for both protein identification and quantification.

In conclusion, this chapter aims to provide a comprehensive guide to DIA workflow understanding and development. We also showed that DIA clearly surpasses DDA in terms of proteome coverage and reproducibility. Therefore, DIA keeps its promises and the further improvements in instrumental performances for scan speed, sensitivity and resolution, as well as the improvements for DIA data analysis will surely make DIA the reference approach for bottom-up proteomics in the coming years.

# Chapter IV Development of cutting edge mass spectrometry approaches to monitor host cell protein impurities during bioprocess development

Today, host cell protein (HCP) impurities are usually quantified by ELISA, which is the gold standard method for HCP monitoring during bioprocess development, manufacturing and final product purity assessment[43-45]. However, ELISA suffers from major drawbacks, including (i) its limited HCP coverage because it is based on the use of an anti-HCP antibodies solution which cannot cover the entire HCP population, and (ii) it does not provide information on the identity of the detected HCP. An increasing number of papers show that ELISA does not provide a comprehensive HCP monitoring[43-45, 56-58]. Therefore, there is an urgent need for alternative methods development, among which mass spectrometry (MS) approaches are the most promising, because they can provide individual HCP monitoring without the biases inherent to ELISA[44-45].

In this context, we aimed to develop cutting edge MS-based quantification approaches to monitor HCP impurities in monoclonal antibody (mAb) samples. For this project, I realised the whole workflow, from the sample preparation at University College London to the quantification of HCP by MS at University of Strasbourg.

The sample preparation was performed in the Department of Biochemical Engineering at University College London, where I followed extensive trainings on cell culture maintenance and monitoring, mAb purification and HCP quantification. For this project, UCB Pharma provided a mAb-producing CHO cell line to be used as a model. I cultivated these cells at small scale in shake flasks, collected the harvest material and performed the most common first step of the mAb purification, namely protein A affinity purification, which removes the vast majority of HCP[59]. Thereby, I generated a range of sample obtained from different steps of the purification process, i.e. clarified cell culture fluid (CCCF) and post protein A (PPA) fractions, using different production conditions, i.e. different cell culture duration, shear stresses and protein A purification protocols. In total, 4 CCCF and 8 PPA samples were collected, representing more than 600 aliquots.

Back to the Laboratory of BioOrganique Mass Spectrometry (LSMBO) in Strasbourg, I developed a range of MS-based analytical methods to quantify HCP impurities in the produced samples. After having built a comprehensive HCP spectral library, we developed an original data independent acquisition (DIA) approach, combining global HCP profiling and absolute quantification of key HCP impurities within a single analysis. The global HCP profiling was performed using Top 3 estimations, assuming that the signal of the three best responding peptides per mole of protein is constant within a coefficient of variation of less than 10%[60]. Absolute quantification of key HCP was performed using isotope dilution (ID). This method, named Top 3-ID-DIA, was benchmarked against the gold standard methods ELISA and SRM coupled to isotope dilution (ID-SRM). Overall, HCP were quantified over 5 orders of magnitude and down to sub-ppm level. The Top 3-ID-DIA showed equivalent sensitivity, accuracy and precision when compared to ID-SRM.

In conclusion, the developed Top 3-ID-DIA method could provide strong support to bioprocess development and product purity assessment.

This work has been submitted to the Analytical Chemistry journal of the American Chemical Society, and the submitted article is attached, comprising the manuscript and the supplementary information, except the Supplementary Table 1 which is too large to be included in this manuscript.

# Dual Data Independent Acquisition approach combining global HCP profiling and absolute quantification of key impurities during bioprocess development

**Authors list**

Gauthier Husson[1], Aurélie Delangle[2], John O'Hara[3], Annick Gervais[2], Alain Van Dorsselaer[1], Dan Bracewell[4], Christine Carapito[1*]

**Affiliations**

[1]Laboratoire de Spectrométrie de Masse BioOrganique, Université de Strasbourg, CNRS, IPHC, UMR 7178, F-67000 Strasbourg, France

[2]Department of Analytical Sciences Biologicals, UCB Pharma s.a., Chemin du Foriest, B-1420 Braine L'alleud, Belgium

[3]Department of Analytical Sciences Biologicals, UCB Pharma s.a., 216 Bath Road, Slough SL1 4EN, UK

[4]Dept. Biochemical Engineering, University College London, Gower Street, London, WC1E 6BT, UK

*Corresponding author: Christine Carapito, Laboratoire de Spectrométrie de Masse BioOrganique, Université de Strasbourg, CNRS, IPHC, UMR 7178, 25 Rue Becquerel, F-67087 Strasbourg, France

**Keywords**

Bioprocess development - Host cell proteins - Mass spectrometry characterization - Data Independent Acquisition – Absolute quantification

## Abstract

Host cell proteins (HCP) are a major class of impurities derived from recombinant protein production process. While HCP are usually monitored by ELISA, mass spectrometry (MS) based approaches are emerging as promising orthogonal methods. Here, we developed an original method relying on data independent acquisition (DIA) coupled to global HCP amounts estimation (Top 3) and absolute quantification with isotope dilution (ID). The method named Top 3-ID-DIA was benchmarked against ELISA and a gold standard selected reaction monitoring assay (ID-SRM). Both MS-methods were applied on various samples generated at different steps and conditions of the purification process, including different culture durations, harvest procedures and purification protocols. Overall, HCP were quantified over 5 orders of magnitude and down to sub-ppm level. The Top 3-ID-DIA strategy proved to be equivalent to the gold standard ID-SRM in terms of sensitivity (1-10 ppm), accuracy and precision. Moreover, 81% of the Top 3 estimations were accurate within a factor of 2 when compared to ID-SRM. Thus, our approach aggregates global HCP profiling for comprehensive process understanding with absolute quantification of key HCP within a single analysis, and provides an efficient support for bioprocess development and product purity assessment.

Host cell proteins (HCP) constitute a major class of impurities that must be monitored and efficiently removed during recombinant protein purification process[1]. Remaining HCP in the final drug product can reduce the drug efficacy[2-4] or induce adverse patient reactions[5-6]. Therefore, HCP amounts in the final drug product must be provided to the regulatory authorities[7]. As a rule of thumb, HCP must be quantified below 100 ppm in the final product by enzyme-linked immunosorbent assay (ELISA)[8]. However, there are more and more evidences that ELISA does not provide comprehensive HCP quantification since it only detects HCP that induced immune response in animals during ELISA development and provides only total HCP amounts without any information on the identity of the detected HCP[1, 9-13]. Finally, developing a specific ELISA is costly and time consuming[10].

As an alternative, mass spectrometry (MS) approaches recently revealed to be most promising to characterise HCP contents as they allow unbiased quantification and individual HCP monitoring. Recent advances in the MS field, notably the use of MS2 signals for quantification by targeted methods (selected reaction monitoring SRM[14] or parallel reaction monitoring PRM[15]) or data independent acquisition (DIA) methods, allowed a 2- to 8-fold gain in sensitivity[16], and a significant gain in specificity and dynamic range when compared to the use of MS1 signals. These features are particularly crucial in the HCP field in which very low abundant proteins have to be quantified besides a highly abundant predominant protein. The targeted approach using SRM conducted on triple-quadrupole type instruments coupled to isotope dilution has, for long, been the gold standard MS-based quantification technique offering highest sensitivity, accuracy and robustness. However, targeted approaches are still limited in multiplexing to a few tens of proteins[14]. Besides, DIA modes based on the collection of MS2 information for all detectable species have been recently introduced on high resolution/accurate mass (HRAM) instruments in order to extract valuable quantitative information from whole complex proteome maps[16]. For instance, two-dimensional liquid chromatography coupled to DIA-type MS$^E$ (2D-LC MS$^E$) technology has been used in a few studies to quantify HCP in monoclonal antibody (mAb) solutions. A Top 3 quantification strategy, which assumes that the signal of the three best responding peptides per mole of protein is constant within a coefficient of variation of less than 10%[17], was used in these studies to estimate absolute amounts of HCP down to a lower limit of quantification (LLOQ) of about 10 ppm[18-23]. However, at least 10 hours were necessary for this type of analysis, which is not easily compatible with real time process support. Alternatively, a 1D-LC sequential windowed acquisition of all theoretical fragment ion mass spectra (SWATH) DIA approach was recently shown by Walker and co-workers[24] to achieve equivalent sensitivity in only one hour.

In this work, we developed MS-based quantification approaches to characterise and profile HCP contents in a variety of mAb samples obtained from different steps and conditions of the purification process. We propose an original dual DIA-based HCP quantification approach allowing both global HCP profiling and absolute quantification of a subset of key HCP, thereby leveraging the advantages of global and targeted approaches within a single analysis. Our method was benchmarked against ELISA and a gold standard isotope dilution SRM assay (ID-SRM).

## Experimental Section

***Cell culture***. An IgG4 A33 mAb producing CHO-DG44 cell line (provided by UCB Pharma, Brussels, Belgium) was cultivated in batch mode using a protein free and chemically defined CD CHO medium (Thermo Fisher Scientific, Waltham, MA, USA) supplemented with 6 mM Glutamine (Thermo Fisher Scientific) and 5 nM Methotrexate (Merck, Darmstadt, Germany). Cells were grown in 1 L Erlenmeyer flasks (300 mL working volume) and incubated at 36.5°C with 5% CO2 on an orbital shaker (123 rpm). The cell concentration and viability were monitored every day using a Vi-Cell (Beckman Coulter, Brea, CA, USA). Viable cells were distinguished from dead cells using the trypan blue dye exclusion method.

***Samples production.*** The ultra scale-down (USD) shearing, harvest procedure and clarified cell culture fluid (CCCF) fractions collection protocols are described in Supporting Information.

***Protein A chromatography.*** mAbs were purified using 1 mL HiTrap MabSelect SuRe columns (GE Healthcare Life Sciences, Pittsburgh, PA) on an AKTA Pure system (GE Healthcare Life Sciences). Two purification types were performed at 1 mL/min using either (i) a standard protocol[25] or (ii) a modified protocol[26]. (i) Standard protocol: equilibration step (5 column volumes (CV) of PBS, pH 7.4) followed by loading of an appropriate volume of CCCF for 20 mg mAb. The column was washed with loading buffer, and the mAb was eluted (0.1 M citrate pH 3.6). (ii) Modified protocol: equilibration step (5 CV 25 mM Tris, 100 mM NaCl pH 7.4) followed by loading of an appropriate volume of CCCF for 20 mg mAb. The column was washed with loading buffer. An intermediate wash (5 CV 25 mM Tris, 10% isopropanol, 1 M urea, pH 9) and a pre-elution wash (3 CV 50 mM citrate, pH 4.4) were performed before mAb elution (100 mM acetate, pH 3.6). After elution, the post protein A (PPA) fractions were directly neutralised to pH 6 using 2 M Tris HCl pH 8.8. A dedicated new column was used for each purification.

***Protein quantification.*** mAb and global protein quantifications were performed as described in Supporting Information.

***HCP-ELISA***. The HCP were quantified using the CHO HCP ELISA kit, 3G (Cygnus Technologies, Southport, NC, USA) in technical triplicates according to the manufacturer’s protocol.

***Sample preparation.*** Samples were separated using SDS-PAGE for spectral library generation (pooled CCCF and pooled PPA fractions), or stacked in a single band for HCP quantification. Retention time standards (iRT, Biognosys, Zurich, Switzerland) and four accurately quantified standard proteins (on-column 100 fmol ADH (yeast alcohol dehydrogenase P00330), 20 fmol PYGM (phosphorylase b P00489), 5 fmol BSA (bovin serum albumin P02769) and 2 fmol ENL (yeast enolase P00924) from the MassPREP Digestion Standard Kit, Waters, Milford, MA, USA) were spiked in each samples. For absolute quantification experiments, a concentration-balanced mixture of 20 accurately quantified stable isotope labelled peptides (AQUA peptides, Thermo Fisher Scientific) were spiked.

***Mass spectrometry analysis.*** Data dependent acquisition (DDA) and data independent acquisition-sequential windowed acquisition of all theoretical fragment ion mass spectra (DIA-SWATH) analyses were performed on an Eksigent NanoLC 400 system operated in microLC-mode and coupled to a TripleTOF 6600 quadrupole-time of flight mass spectrometer (both from SCIEX, Framingham, MA, USA). Selected reaction monitoring (SRM) analyses were performed on a Dionex UltiMate 3000 operated in microLC-mode and coupled to a TSQ Vantage triple quadrupole mass spectrometer (both from Thermo Fisher Scientific). On both couplings, 8 µg of peptides were separated on a ZORBAX 300SB-C18 column (150 mm x 300 µm with 3.5 µm diameter particles, Agilent Technologies). All chromatographic gradient and MS settings are given in Supporting Information.

***DIA-SWATH targeted data extraction.*** A spectral library was generated as described in Supporting Information. DIA data were processed using Skyline[27] (version 3.5.9.10061). Validated proteotypic peptides from the spectral library were extracted with following parameters (based on previous work[28] and in-house optimisations on standard samples, data not shown): the 6 most intense 1+ b- and y-type product ions were extracted, from ion 3 to last ion – 1, while the precursors with less than 3 transitions were excluded. Resolving power was set to 50 000, and a retention time tolerance of 5 min (+/- 2.5 min) was used. Retention times were predicted with iRT standards (Biognosys). Peaks were reintegrated using the target decoy approach (reverse sequences) of the mProphet peak-scoring model, and a Q-value was assigned to each peak. Peak integrations were manually checked and curated for HCP of interest.

Fragment areas, detection Q value and library dot product were exported for each peptide in .csv files.

***Top 3 estimation.*** Only peptides with Q-value below 0.01 (corresponding to a false discovery rate of 1%) and dot-product above 0.6 were kept. The fragment areas were summed for each peptide and the 3 best responding peptides were summed for each protein. Only proteins quantified in at least two replicates in at least one sample were kept, independently for CCCF and PPA fractions. The universal signal response factor[17] (signal / mol of protein) was calculated using PYGM and was used to estimate mol quantities of all proteins. Using molecular weights and mAb quantifications, individual HCP amounts in ppm were estimated. Only quantifications with a coefficient of variation (CV) below 20% between technical triplicates were used to build a heat map (see Supplementary Table S1) and calculate total HCP amounts in each sample.

***Selection of 10 HCP and their proteotypic peptides.*** Ten HCP were chosen based on their potential immunogenicity, proteolytic activity, purification behaviour, or estimated abundance using preliminary data acquired on the samples. The selected HCP and their proteotypic peptides are described in Supporting Information and Supplementary Table S2.

***Absolute quantification.*** Six transitions were analysed for each precursor ion for both SRM and DIA approaches. If comprised in the linear range of the assay as determined by calibration curves (detailed procedures for calibration curves and LLOQ determination are provided in Supporting Information and Supplementary Figure S1 and S2). The ratios between endogen and stable isotope labelled AQUA peptides were used to calculate the mol amounts of endogenous peptides, which were averaged to calculate the mol amounts of corresponding proteins. Using molecular weights and injected mAb quantities, individual HCP amounts in ppm were calculated.

## Results

### *Overview of MS approaches developed to monitor HCP contents during bioprocess development*

MS-based HCP quantification strategies were evaluated according to the global workflow presented in Figure 1. First, the effect of the cell culture duration and cell viability at harvest on the HCP content were investigated, as it was shown to induce significant changes in harvest HCP composition[12, 29-32]. Cell culture fluid was thus collected at days 7 and 10, corresponding to cell viability of 71% and 8%, respectively (Supplementary Figure S3). Then, different shear stress conditions during harvest were compared using the ultra scale down (USD) shear device[33-34] developed at University College London. Finally, two protein A purification protocols were compared: a standard protocol[25], and a modified protocol[26] including a high pH wash with a combination of 1 M urea and 10% isopropanol in order to disrupt mAb – HCP interactions while preserving mAb – protein A bindings. Overall, 4 clarified cell culture fluid (CCCF) and 8 post protein A (PPA) fractions were collected (Figure 2).

MS-based quantification approaches were developed and benchmarked for HCP monitoring on all samples: an original Top 3-ID-DIA approach combining DIA and Top 3 quantification of all detected HCP (using a single reference protein spiked in known amounts) with isotope dilution for absolute quantification of a subset of 10 selected HCP, and a gold standard isotope dilution SRM assay to absolutely quantify the same 10 selected HCP.

For global profiling, a spectral library was generated as described in Experimental Section and used to extract signals for all detectable HCP. For Top 3 amount estimations, the PYGM protein was used as the reference protein, and ADH, BSA and ENO were used as quantification controls. Stable isotope labelled peptides were also spiked into all samples to allow accurate absolute quantification of the 10 selected HCP with both Top 3-ID-DIA and ID-SRM methods.

### *Global HCP contents estimation over bioprocess steps*

First, we generated a comprehensive HCP spectral library from the mAb-producing CHO cell line containing 25 338 proteotypic peptides corresponding to 3 220 proteins. While it can be easier to use a null CHO cell line to build an HCP spectral library to avoid interferences from the overwhelming mAb peptides, the use of an SDS-PAGE separation prior to LC-MS/MS analyses overcame this issue, and allowed us to build an HCP spectral library that specifically corresponds to the producing cell line. Then, to reduce potential interferences and achieve highest specificity in DIA analysis, two distinct acquisition methods using 75 variable windows were optimised for CCCF and PPA fractions (Supplementary Table S3). Transition groups specific to all peptides contained in the spectral library were extracted using their predicted retention times (thanks to retention time standards) as described in the Experimental Section. Noteworthy is that the specificity of PPA samples required an additional dot product threshold to be applied to remove highly interfered peptides. Indeed, in these samples a very low number of targeted peptides are effectively present which makes the differentiation between targets and decoys challenging and thus the false discovery rate strategy suboptimal[35]. The universal response factor[17] (signal / mol of protein) allowed the quantification controls ADH, BSA and ENO spiked at 100 fmol, 5 fmol and 2 fmol, to be estimated at 140 ± 12 fmol, 7 ± 2 fmol and 0.5 ± 0.2 fmol, respectively. In the end, only quantifications achieved with a coefficient of variation (CV) of less than 20% between triplicates were summed to calculate total HCP amounts for each sample.

On average, 1 454 HCP were quantified in the CCCF fractions representing 288 513 to 389 657 ppm and 119 HCP in the PPA fractions representing 2 646 to 5 386 ppm (Figure 2). These global HCP amounts are in accordance with previous studies focused on HCP quantification by MS[19, 23]. We could estimate individual HCP amounts ranging from 0.5 to 16 192 ppm in the CCCF fractions, and from 0.1 to 731 ppm in the PPA fractions, thus covering a dynamic range of 5 orders of magnitude (Supplementary Table S1). In parallel, HCP were quantified in PPA fractions using ELISA from 276 to 959 ppm, which is significantly lower when compared to Top 3-ID-DIA estimations (on average 8 times lower).

### *Absolute quantification of 10 selected HCP*

***ID-SRM quantification.*** Working with the same LC gradient as for the Top 3-ID-DIA analyses, a time-scheduled ID-SRM method was developed, first, using 20 crude stable isotope labelled synthetic peptides spiked in CCCF and PPA matrices. Six specific transitions were chosen for each peptide, and collision energy values were optimised for each transition. Once the method optimised, accurately quantified stable isotope labelled AQUA peptides were spiked in known amounts into the samples for absolute quantification. Calibration curves were realised for each peptide to determine the linear quantification range and LLOQ of the assay (Supplementary Figure S1). FDA-approved criteria[36] were applied for calibration curves interpretation as detailed in Supporting Information. LLOQ values could be determined for 13 out of the 20 peptides. Absolute quantification could be obtained for 8 out of the 10 targeted HCP, ranging from 1.7 to 23 681 ppm thus covering a dynamic range of 4.1 orders of magnitude (Supplementary Table S4, Supplementary Figure S4). As expected, Pyruvate kinase was found very abundant in CCCF fractions (from 13 674 to 23 681 ppm), while cytoplasmic Isoleucyl-tRNA synthetase was very low abundant in PPA fractions (below 18 ppm). Difficult to remove HCPs were detected in the PPA fractions from 1.7 to 106 ppm with the exception of Pyruvate kinase found at 157 and 536 ppm in PPA 7 and 8 fractions, respectively. HEAT repeat-containing protein 3 and Eukaryotic translation initiation factor 3 subunit L were not quantified because no valid calibration curve could be built for their corresponding peptides.

***Top 3-ID-DIA quantification.*** The same 10 HCP were absolutely quantified using isotope dilution within the Top 3-ID-DIA experiment. Identical criteria were applied as for the ID-

SRM approach to build calibration curves and determine LLOQ in DIA mode. Calibration curves and LLOQ values could be determined for 17 out of the 20 peptides (Supplementary Figure S2). For a fair comparison with the ID-SRM approach, the same peptides were used to quantify the HCP, except for HEAT repeat-containing protein 3 and Eukaryotic translation initiation factor 3 subunit L which were not quantified by ID-SRM (Supplementary Table S4, Supplementary Figure S4). The 10 HCP were accurately quantified from 0.7 to 26 017 ppm thus covering a dynamic range of 4.6 orders of magnitude. Again, as expected, Pyruvate kinase was found highly abundant in CCCF fractions (from 15 494 to 26 017 ppm), while cytoplasmic Isoleucyl-tRNA synthetase was very low abundant in PPA fractions (below 20 ppm). Difficult to remove HCPs were consistently detected in PPA fractions from 0.7 to 120 ppm, excepted Pyruvate kinase which was quantified at 172 and 456 ppm in PPA 7 and 8 fractions, respectively.

## Discussion

### *Benchmarking of MS methods for HCP quantification*

First, the sensitivity of the Top 3-ID-DIA method was evaluated by comparing LLOQ values (obtained using calibration curves of AQUA peptides) to LLOQ values achieved with the gold standard ID-SRM assay (Supplementary Figure S5a). Most LLOQ determined for both modes were below 10 ppm, with minima at 0.3 ppm for ID-SRM and 0.1 ppm for ID-DIA. These sensitivities are consistent, even better, when compared to previous works that published LLOQ values of 10 ppm using 1D LC-SWATH[24] and 2D-LC MS[E23]. Then, the accuracy of both Top 3 estimations and ID-DIA absolute quantifications simultaneously achievable with the Top 3-ID-DIA method, was assessed using pairwise comparisons to ID-SRM absolute quantifications (Supplementary Figure S5b). Absolute quantifications achieved by ID-SRM and ID-DIA were all consistent within a factor of 2. Top 3 estimations presented wider errors attributable to both acquisition mode and quantification strategy changes. Nevertheless, it is of note that 81% of the Top 3 estimations were consistent with ID-SRM quantification values within a factor of 2. These results are also in line with previous evaluations of the Top 3 estimation strategy for HCP quantification[22, 24]. Finally, the precision of quantification was probed using coefficients of variation (CV) between technical triplicates (Supplementary Figure S5c). All three approaches, Top 3 estimation, ID-DIA and ID-SRM displayed equivalent and good precision with a vast majority of CV values between technical triplicates below 5%.

Overall, our results demonstrate that Top 3-DIA estimations are not so far from accurate quantifications and constitute a good compromise with limited method setup requirements and limited cost for stable isotope labelled standards, while providing a wide view and understanding of the HCP content. However, when accurate quantification of specific HCP of interest needs to be provided (which is certainly the case for problematic HCP such as for instance recognised immunogenic ones), the use of stable isotope dilution combined to a targeted MS assay like ID-SRM is still recommended. In the present work, we proved that combining both strategies within a single Top 3-ID-DIA approach is possible without compromising performances. By spiking a reference protein and an optimised mixture of stable isotope labelled AQUA peptides corresponding to key HCP into the samples, we reached equivalent performances compared to the gold standard ID-SRM approach for a subset of HCP (ID-DIA) in addition to provide estimations of all detected HCP amounts within a single analysis. The combined Top 3-ID-DIA strategy thus constitutes a solution of choice that could be generalised in the HCP characterisation field.

### *Benchmarking of MS quantification against ELISA quantification*

In PPA fractions, total HCP contents were quantified by the Top 3-ID-DIA approach and by a generic ELISA. Overall, the MS-HCP quantification raised on average an 8 fold higher total HCP content, which is in line with previous reports[37-39]. This can be explained by the biases intrinsic to ELISA quantification (i) the anti-HCP antibodies only detect a subset of HCP (those who elicit an immune response during ELISA development) while MS allows unbiased quantification of all detectable HCP, (ii) intracellular enzymes including proteases released at harvest (increased probability with increased shear stress[33]) may degrade HCP and thus prevent their recognition by ELISA, while degraded HCP can still be detected by MS[2-4, 40], and (iii) the generic ELISA standard HCP sample that is used to generate the standard curve does not contain the same HCP population as the tested samples thus biasing the quantification. Interestingly, the ratio obtained between MS and ELISA quantifications increases when the samples diverge from a "standard" sample. Indeed, the MS over ELISA quantification ratio is about 4 for "standard" samples (PPA 1, 2 and 3) generated using standard protocols, while the ratio increases for "non-standard" samples, up to 14.4 for PPA 8 fraction. In fact, 46 HCP were detected uniquely in the PPA 8 fraction representing 904 ppm over a total of 5 386 ppm for 154 HCP. This observation again argues that ELISA targets only a subset of all possible HCP. Ultimately, this argument could also be raised for our Top 3-ID-DIA approach as it is limited to HCP present in the spectral library used to extract the data. However, building an HCP spectral library is less demanding of time and resources than developing a new ELISA assay, and improving an HCP spectral library by adding newly identified peptides is possible anytime (provided retention time standards are spiked in newly analysed samples). Thus, an HCP spectral library can be considered as an evolving resource that can be easily shared and implemented to ultimately reach the largest proteome coverage of the concerned cell line. On the other hand, non-library-based algorithms are currently being developed by the computational proteomics community to interpret DIA data and, even if the results are still not reaching the quality levels of library-based approaches, the output of these solutions has recently significantly improved in terms of proteome coverage and false discovery proportions control[28, 41]. Besides, HCP characterisation using MS techniques would greatly benefit from a better curated CHO protein database[42], as the one that is currently available on public resources contains 99% unreviewed and high redundancy sequences (mostly UniProtKB-TrEMBL entries). Actually, database redundancy is in the end the most limiting factor as only unique peptides are considered for quantification and therefore numerous peptides are unnecessarily discarded based on non-unicity criteria.

### ***MS allows better understanding of process-related behaviours***

Beyond global HCP contents estimation and unlike ELISA, the Top 3-ID-DIA approach also allowed precise identification of about 1 450 HCP in CCCF fractions and 120 HCP in PPA fractions. Precisely identifying and individually quantifying HCP is of crucial importance if one aims to understand ongoing mechanisms and eventually improve bioprocess.

For instance, global HCP contents were estimated gradually higher when cell culture fluids were exposed to low or high shear stress or when cells were cultivated for an extended duration. Both observations are in line with previous studies[11-12, 29-30]. However, this tendency was not observed in PPA fractions and it can be tentatively explained by looking at specific HCP behaviours (Supplementary Figure S6a). As an example, a gradual and strong enrichment of ribosomal proteins was observed with increasing shear stress: from 1 804 ppm without shear stress (CCCF 1) to 3 619 and 12 409 ppm with low (CCCF 2) and high (CCCF 3) shear stresses, respectively. Indeed, the shear stress is known to induce cell breakage and therefore intracellular content release among which ribosomal proteins are highly abundant[33]. However, these differences were not observed among PPA fractions where ribosomal proteins were quantified around 50 ppm regardless their originating CCCF fraction, demonstrating that the protein A purification step efficiently removed these abundant intracellular proteins. Such behaviour can be extended to the majority of HCP, as the protein A purification step removes the vast majority of impurities and remaining HCP in the PPA fraction are known to be mainly "hitchhiker"

HCP bound to the mAb, thus affecting the differences observed upstream[19, 43-46].

Moreover, we showed that an extended cell culture duration led to overrepresentation of heat shock proteins family in the CCCF fractions (Supplementary Figure S6b): Heat shock protein (tr|A0A061ID29|A0A061ID29_CRIGR) was quantified at 24 ppm at 7 days versus 124 ppm at 10 days, Endoplasmin (tr|G3HQM6|G3HQM6_CRIGR) at 8 798 ppm at 7 days versus 16 192 ppm at 10 days, and 78 kDa glucose-regulated protein (tr|G3I8R9|G3I8R9_CRIGR) at 6 059 ppm at 7 days versus 12 009 ppm at 10 days. While Heat shock protein is totally removed (not detected) and 78 kDa glucose-regulated protein is partially removed from all PPA fractions (average 54 ppm), Endoplasmin remains more abundant in PPA fractions obtained at 10 days (on average 110 ppm in all PPA fractions from 7 days but 172 and 405 ppm in PPA 7 and 8 fractions from 10 days). On the contrary, several proteins are underrepresented in CCCF fractions after an extended culture duration, like Annexins (tr|A0A061IML2|A0A061IML2_CRIGR, tr|G3I5L3|G3I5L3_CRIGR, tr|G3IG05|G3IG05_CRIGR) which were quantified at 245, 797 and 324 ppm at 7 days versus 0, 51 and 18 ppm at 10 days. Annexins were efficiently removed by the protein A purification and were not detected in any PPA fraction. Conversely, several HCP keep constant over time in CCCF fractions but are significantly more abundant in PPA fractions obtained after 10 days of culture like Pyruvate kinase (tr|A0A098KXC0|A0A098KXC0_CRIGR) which was quantified on average at 60 ppm in PPA fractions from 7 days versus 183 and 445 ppm in PPA 7 and 8 fractions from 10 days, 6-phosphogluconate dehydrogenase, decarboxylating (tr|G3IHY5|G3IHY5_CRIGR) which was not detected in PPA fractions except in PPA 7 and 8 fractions from 10 days at 32 and 67 ppm, respectively, Heat shock cognate 71 kDa protein (sp|P19378|HSP7C_CRIGR) which was quantified on average at 21 ppm in PPA fractions from 7 days versus 94 and 250 ppm in PPA 7 and 8 fractions from 10 days, respectively. These behaviours are not easy to understand but one could hypothesise that the presence of specific HCP after 10 days in the CCCF fraction could help other HCP to co-purify with the mAb. Thereby, an HCP that is known to be easily removed could become challenging in the presence of certain cofactors.

Finally, several HCP were more efficiently removed by the modified protein A purification protocol compared to the standard purification protocol (Supplementary Figure S6c) such as Heterogeneous nuclear ribonucleoprotein U-like protein 1 (tr|G3IA10|G3IA10_CRIGR) which was quantified on average at 298 ppm in PPA fractions obtained using the standard purification protocol, but at 59 and 14 ppm in PPA 4 and 8 fractions obtained using the modified purification protocol; Putative phospholipase B-like 2 (tr|G3I6T1|G3I6T1_CRIGR) which was quantified on average at 62 ppm in PPA fractions obtained using the standard purification protocol, but at only 12 ppm using the modified purification protocol. Phospholipase B-like 2, or PLBL-2, was also absolutely quantified by ID-SRM and ID-DIA which confirmed the Top 3-DIA estimations as shown in Figure 3. This is particularly interesting since PLBL-2 is known for its immunogenicity[5-6] and currently constitutes a major purification challenge[47].

More generally, it becomes obvious that after the protein A purification step, downstream purification process has to face with co-purifying HCP that are in majority specifically bound the mAb[19, 43-46]. In this context also, the specific identification of "difficult to remove" HCP by MS and the consecutive development of robust targeted quantification methods constitute tools of choice to help in designing an appropriate HCP clearance strategy. Altogether, these examples show that MS-based quantification approaches, and especially the proposed combined Top 3-ID-DIA approach, provide additional valuable information to ELISA: HCP contents can be precisely monitored in a more comprehensive manner with high throughput, and can lead process development to release cleaner and safer products.

## Acknowledgements

This project was supported by the «Association Nationale de la Recherche et de la Technologie » and UCB Pharma via the CIFRE fellowship of GH. This work was also supported by the “Agence Nationale de la Recherche” (ANR) and the French Proteomic Infrastructure (ProFI; ANR-10-INBS-08-03).

## References

1. Bracewell, D. G.; Francis, R.; Smales, C. M., The future of host cell protein (HCP) identification during process development and manufacturing linked to a risk-based management for their control. *Biotechnol Bioeng* **2015,** *112* (9), 1727-37.
2. Gao, S. X.; Zhang, Y.; Stansberry-Perkins, K.; Buko, A.; Bai, S.; Nguyen, V.; Brader, M. L., Fragmentation of a highly purified monoclonal antibody attributed to residual CHO cell protease activity. *Biotechnol Bioeng* **2011,** *108* (4), 977-82.
3. Bee, J. S.; Tie, L.; Afdahl, C. D.; Jusino, K. C.; Johnson, D.; Dimitrova, M. N., Trace levels of the CHO host cell protease cathepsin D caused particle formation in a monoclonal antibody product. *Biotechnol Prog* **2015**.
4. Robert, F.; Bierau, H.; Rossi, M.; Agugiaro, D.; Soranzo, T.; Broly, H.; Mitchell-Logean, C., Degradation of an Fc-fusion recombinant protein by host cell proteases: Identification of a CHO cathepsin D protease. *Biotechnol Bioeng* **2009,** *104* (6), 1132-41.
5. Hanania, N. A.; Noonan, M.; Corren, J.; Korenblat, P.; Zheng, Y.; Fischer, S. K.; Cheu, M.; Putnam, W. S.; Murray, E.; Scheerens, H.; Holweg, C. T.; Maciuca, R.; Gray, S.; Doyle, R.; McClintock, D.; Olsson, J.; Matthews, J. G.; Yen, K., Lebrikizumab in moderate-to-severe asthma: pooled data from two randomised placebo-controlled studies. *Thorax* **2015,** *70* (8), 748-56.
6. Fischer, S. K.; Cheu, M.; Peng, K.; Lowe, J.; Araujo, J.; Murray, E.; McClintock, D.; Matthews, J.; Siguenza, P.; Song, A., Specific Immune Response to Phospholipase B-Like 2 Protein, a Host Cell Impurity in Lebrikizumab Clinical Material. *The AAPS journal* **2017,** *19* (1), 254-263.
7. ICH *Guidance for Industry Q6B Specifications: Test Procedures and Acceptance Criteria for Biotechnological/Biological products*; 1999.
8. Chon, J. H.; Zarbis-Papastoitsis, G., Advances in the production and downstream processing of antibodies. *N Biotechnol* **2011,** *28* (5), 458-63.
9. Tscheliessnig, A. L.; Konrath, J.; Bates, R.; Jungbauer, A., Host cell protein analysis in therapeutic protein bioprocessing - methods and applications. *Biotechnol J* **2013,** *8* (6), 655-70.
10. Wang, X.; Hunter, A. K.; Mozier, N. M., Host cell proteins in biologics development: Identification, quantitation and risk assessment. *Biotechnol Bioeng* **2009,** *103* (3), 446-58.
11. Hogwood, C. E.; Bracewell, D. G.; Smales, C. M., Measurement and control of host cell proteins (HCPs) in CHO cell bioprocesses. *Curr Opin Biotechnol* **2014,** *30C*, 153-160.
12. Jin, M.; Szapiel, N.; Zhang, J.; Hickey, J.; Ghose, S., Profiling of host cell proteins by two-dimensional difference gel electrophoresis (2D-DIGE): Implications for downstream process development. *Biotechnol Bioeng* **2010,** *105* (2), 306-16.
13. Zhu-Shimoni, J.; Yu, C.; Nishihara, J.; Wong, R. M.; Gunawan, F.; Lin, M.; Krawitz, D.; Liu, P.; Sandoval, W.; Vanderlaan, M., Host cell protein testing by ELISAs and the use of orthogonal methods. *Biotechnol Bioeng* **2014,** *111* (12), 2367-79.
14. Picotti, P.; Aebersold, R., Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat Methods* **2012,** *9* (6), 555-66.
15. Gallien, S.; Duriez, E.; Crone, C.; Kellmann, M.; Moehring, T.; Domon, B., Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. *Mol Cell Proteomics* **2012,** *11* (12), 1709-23.
16. Gillet, L. C.; 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.

17. Silva, J. C.; Gorenstein, M. V.; Li, G. Z.; Vissers, J. P.; Geromanos, S. J., Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics* **2006,** *5* (1), 144-56.

18. Doneanu, C. E.; Anderson, M.; Williams, B. J.; Lauber, M. A.; Chakraborty, A.; Chen, W., Enhanced Detection of Low-Abundance Host Cell Protein Impurities in High-Purity Monoclonal Antibodies Down to 1 ppm Using Ion Mobility Mass Spectrometry Coupled with Multidimensional Liquid Chromatography. *Analytical chemistry* **2015,** *87* (20), 10283-91.

19. Zhang, Q.; Goetze, A. M.; Cui, H.; Wylie, J.; Tillotson, B.; Hewig, A.; Hall, M. P.; Flynn, G. C., Characterization of the co-elution of host cell proteins with monoclonal antibodies during protein A purification. *Biotechnol Prog* **2016,** *32* (3), 708-17.

20. Farrell, A.; Mittermayr, S.; Morrissey, B.; McLoughlin, N.; Navas Iglesias, N.; Marison, I. W.; Bones, J., Quantitative Host Cell Protein Analysis using Two Dimensional Data Independent LC-MS^E. *Analytical chemistry* **2015**.

21. Doneanu, C. E.; Xenopoulos, A.; Fadgen, K.; Murphy, J.; Skilton, S. J.; Prentice, H.; Stapels, M.; Chen, W., Analysis of host-cell proteins in biotherapeutic proteins by comprehensive online two-dimensional liquid chromatography/mass spectrometry. *mAbs* **2012,** *4* (1), 24-44.

22. Schenauer, M. R.; Flynn, G. C.; Goetze, A. M., Identification and quantification of host cell protein impurities in biotherapeutics using mass spectrometry. *Anal Biochem* **2012,** *428* (2), 150-7.

23. Zhang, Q.; Goetze, A. M.; Cui, H.; Wylie, J.; Trimble, S.; Hewig, A.; Flynn, G. C., Comprehensive tracking of host cell proteins during monoclonal antibody purifications using mass spectrometry. *mAbs* **2014,** *6* (3), 659-70.

24. Walker, D. E.; Yang, F.; Carver, J.; Joe, K.; Michels, D. A.; Yu, X. C., A modular and adaptive mass spectrometry-based platform for support of bioprocess development toward optimal host cell protein clearance. *mAbs* **2017,** *9* (4), 654-663.

25. Liu, H. F.; Ma, J.; Winter, C.; Bayer, R., Recovery and purification process development for monoclonal antibody production. *MAbs* **2010,** *2* (5), 480-99.

26. Shukla, A. A.; Hinckley, P., Host cell protein clearance during protein A chromatography: development of an improved column wash step. *Biotechnol Prog* **2008,** *24* (5), 1115-21.

27. MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J., Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010,** *26* (7), 966-8.

28. Navarro, P.; Kuharev, J.; Gillet, L. C.; Bernhardt, O. M.; MacLean, B.; Rost, H. L.; Tate, S. A.; Tsou, C. C.; Reiter, L.; Distler, U.; Rosenberger, G.; Perez-Riverol, Y.; Nesvizhskii, A. I.; Aebersold, R.; Tenzer, S., A multicenter study benchmarks software tools for label-free proteome quantification. *Nat Biotechnol* **2016**.

29. Tait, A. S.; Hogwood, C. E.; Smales, C. M.; Bracewell, D. G., Host cell protein dynamics in the supernatant of a mAb producing CHO cell line. *Biotechnol Bioeng* **2012,** *109* (4), 971-82.

30. Grzeskowiak, J. K.; Tscheliessnig, A.; Toh, P. C.; Chusainow, J.; Lee, Y. Y.; Wong, N.; Jungbauer, A., 2-D DIGE to expedite downstream process development for human monoclonal antibody purification. *Protein Expr Purif* **2009,** *66* (1), 58-65.

31. Hogwood, C. E.; Bracewell, D. G.; Smales, C. M., Measurement and control of host cell proteins (HCPs) in CHO cell bioprocesses. *Curr Opin Biotechnol* **2014,** *30*, 153-60.

32. Hogwood, C. E.; Ahmad, S. S.; Tarrant, R. D.; Bracewell, D. G.; Smales, C. M., An ultra scale-down approach identifies host cell protein differences across a panel of mAb producing CHO cell line variants. *Biotechnol J* **2016,** *11* (3), 415-24.

33. Lau, E. C.; Kong, S.; McNulty, S.; Entwisle, C.; McIlgorm, A.; Dalton, K. A.; Hoare, M., An ultra scale-down characterization of low shear stress primary recovery stages to enhance selectivity of fusion protein recovery from its molecular variants. *Biotechnol Bioeng* **2013,** *110* (7), 1973-83.

34. Boychyn, M.; Yim, S. S.; Bulmer, M.; More, J.; Bracewell, D. G.; Hoare, M., Performance prediction of industrial

centrifuges using scale-down models. *Bioprocess Biosyst Eng* **2004,** *26* (6), 385-91.
35. Elias, J. E.; Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **2007,** *4* (3), 207-14.
36. FDA, Guidance for industry: bioanalytical method validation, Draft. *US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research: Rockville, MD* **2013**.
37. Kreimer, S.; Gao, Y.; Ray, S.; Jin, M.; Tan, Z.; Mussa, N. A.; Tao, L.; Li, Z.; Ivanov, A. R.; Karger, B. L., Host Cell Protein Profiling by Targeted and Untargeted Analysis of Data Independent Acquisition Mass Spectrometry Data with Parallel Reaction Monitoring Verification. *Analytical chemistry* **2017**.
38. Farrell, A.; Mittermayr, S.; Morrissey, B.; Mc Loughlin, N.; Navas Iglesias, N.; Marison, I. W.; Bones, J., Quantitative Host Cell Protein Analysis Using Two Dimensional Data Independent LC-MS. *Anal Chem* **2015**.
39. Henry, S. M.; Sutlief, E.; Salas-Solano, O.; Valliere-Douglass, J., ELISA reagent coverage evaluation by affinity purification tandem mass spectrometry. *MAbs* **2017**, 1-11.
40. Carter-Franklin, J. N.; Victa, C.; McDonald, P.; Fahrner, R., Fragments of protein A eluted during protein A affinity chromatography. *J Chromatogr A* **2007,** *1163* (1-2), 105-11.
41. Tsou, C. C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A. C.; Nesvizhskii, A. I., DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* **2015,** *12* (3), 258-64, 7 p following 264.
42. Meleady, P.; Hoffrogge, R.; Henry, M.; Rupp, O.; Bort, J. H.; Clarke, C.; Brinkrolf, K.; Kelly, S.; Muller, B.; Doolan, P.; Hackl, M.; Beckmann, T. F.; Noll, T.; Grillari, J.; Barron, N.; Puhler, A.; Clynes, M.; Borth, N., Utilization and evaluation of CHO-specific sequence databases for mass spectrometry based proteomics. *Biotechnol Bioeng* **2012,** *109* (6), 1386-94.
43. Levy, N. E.; Valente, K. N.; Choe, L. H.; Lee, K. H.; Lenhoff, A. M., Identification and characterization of host cell protein product-associated impurities in monoclonal antibody bioprocessing. *Biotechnol Bioeng* **2014,** *111* (5), 904-12.
44. Tarrant, R. D.; Velez-Suberbie, M. L.; Tait, A. S.; Smales, C. M.; Bracewell, D. G., Host cell protein adsorption characteristics during protein A chromatography. *Biotechnol Prog* **2012,** *28* (4), 1037-44.
45. Sisodiya, V. N.; Lequieu, J.; Rodriguez, M.; McDonald, P.; Lazzareschi, K. P., Studying host cell protein interactions with monoclonal antibodies using high throughput protein A chromatography. *Biotechnology Journal* **2012,** *7* (10), 1233-1241.
46. Nogal, B.; Chhiba, K.; Emery, J. C., Select host cell proteins coelute with monoclonal antibodies in protein A chromatography. *Biotechnol Prog* **2012,** *28* (2), 454-8.
47. Tran, B.; Grosskopf, V.; Wang, X.; Yang, J.; Walker, D., Jr.; Yu, C.; McDonald, P., Investigating interactions between phospholipase B-Like 2 and antibodies during Protein A chromatography. *J Chromatogr A* **2016,** *1438*, 31-8.

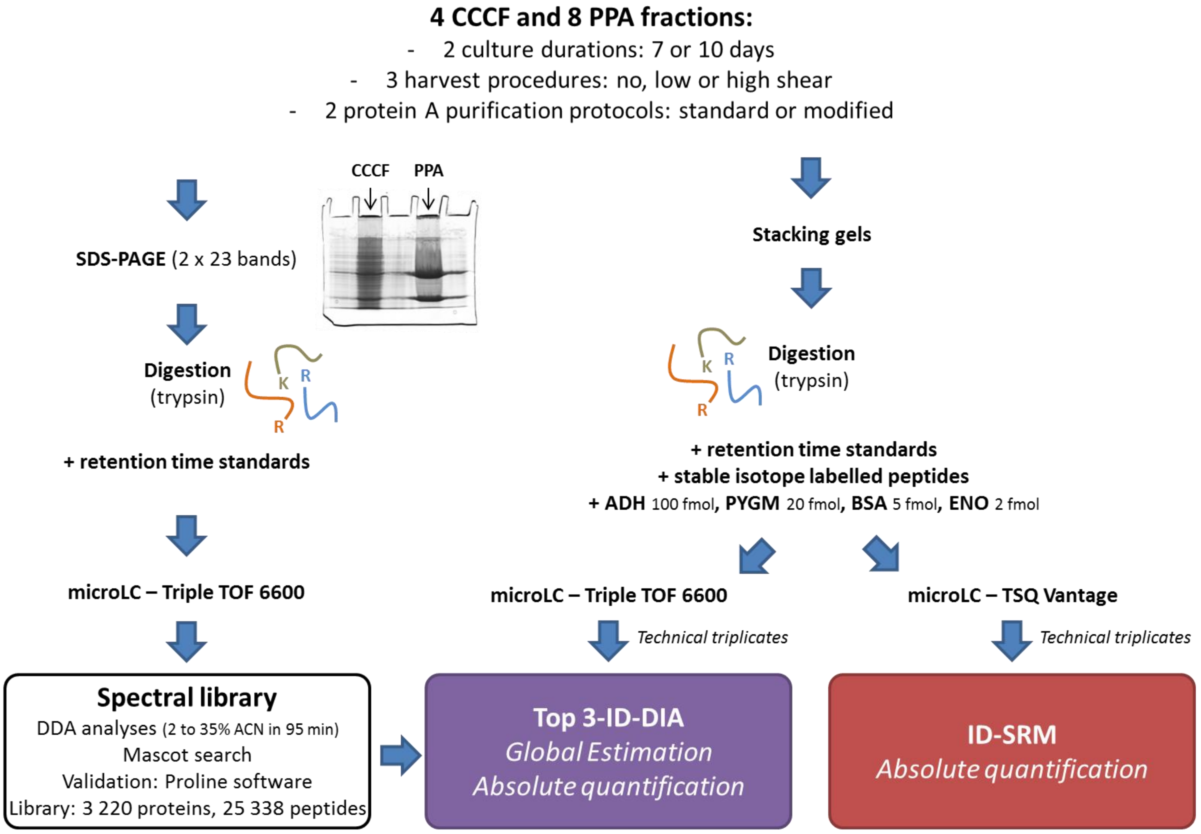
## Figure legends

***Figure 1:*** Overview of MS-based quantification strategies developed for HCP monitoring.

***Figure 2:*** Estimations of global HCP contents using Top 3-ID-DIA (and ELISA quantification for PPA fractions). Quantification was performed using the 3 best responding peptides per protein relative to a standard spiked protein (20 fmol PYGM), deriving ppm values and summing all proteins amounts to obtain a total HCP content in ppm. ELISA quantification was obtained using a generic ELISA kit according to the manufacturer's protocol.

***Figure 3:*** Quantification results obtained for Phospholipase B-like 2 protein in PPA fractions. PPA 1, 2, 3, 5, 6 and 7 fractions were obtained with the standard purification protocol, while PPA 4 and 8 fractions were obtained with the modified purification protocol.

**Figure 1**


4 CCCF and 8 PPA fractions:
- 2 culture durations: 7 or 10 days
- 3 harvest procedures: no, low or high shear
- 2 protein A purification protocols: standard or modified
CCCF
PPA
SDS-PAGE (2 x 23 bands)
Digestion (trypsin)
K
R
R
+ retention time standards
microLC – Triple TOF 6600
Spectral library
DDA analyses (2 to 35% ACN in 95 min)
Mascot search
Validation: Proline software
Library: 3 220 proteins, 25 338 peptides
Stacking gels
Digestion (trypsin)
+ retention time standards
+ stable isotope labelled peptides
+ ADH 100 fmol, PYGM 20 fmol, BSA 5 fmol, ENO 2 fmol
microLC – Triple TOF 6600
Technical triplicates
microLC – TSQ Vantage
Technical triplicates
Top 3-ID-DIA
Global Estimation
Absolute quantification
ID-SRM
Absolute quantification

## Figure 2

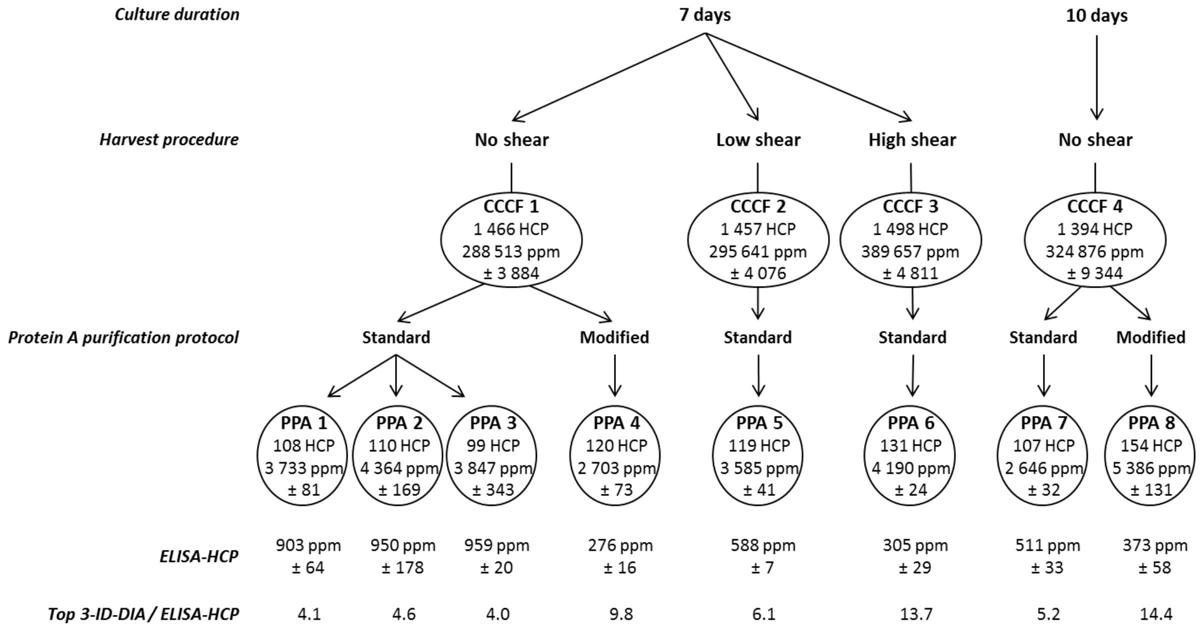
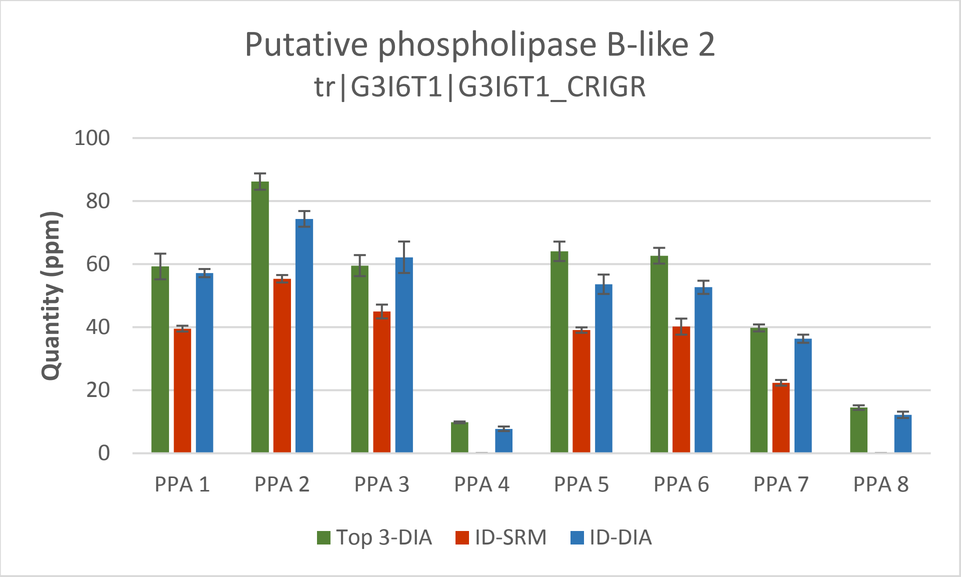

Culture duration
7 days
10 days
Harvest procedure
No shear
Low shear
High shear
No shear
CCCF 1
1 466 HCP
288 513 ppm
± 3 884
CCCF 2
1 457 HCP
295 641 ppm
± 4 076
CCCF 3
1 498 HCP
389 657 ppm
± 4 811
CCCF 4
1 394 HCP
324 876 ppm
± 9 344
Protein A purification protocol
Standard
Modified
Standard
Standard
Standard
Modified
PPA 1
108 HCP
3 733 ppm
± 81
PPA 2
110 HCP
4 364 ppm
± 169
PPA 3
99 HCP
3 847 ppm
± 343
PPA 4
120 HCP
2 703 ppm
± 73
PPA 5
119 HCP
3 585 ppm
± 41
PPA 6
131 HCP
4 190 ppm
± 24
PPA 7
107 HCP
2 646 ppm
± 32
PPA 8
154 HCP
5 386 ppm
± 131
ELISA-HCP
903 ppm
± 64
950 ppm
± 178
959 ppm
± 20
276 ppm
± 16
588 ppm
± 7
305 ppm
± 29
511 ppm
± 33
373 ppm
± 58
Top 3-ID-DIA / ELISA-HCP
4.1
4.6
4.0
9.8
6.1
13.7
5.2
14.4

**Figure 3**


Putative phospholipase B-like 2
tr|G3I6T1|G3I6T1_CRIGR
Quantity (ppm)
100
80
60
40
20
0
PPA 1
PPA 2
PPA 3
PPA 4
PPA 5
PPA 6
PPA 7
PPA 8
Top 3-DIA
ID-SRM
ID-DIA

## Supporting Information

### Dual Data Independent Acquisition approach combining global HCP profiling and absolute quantification of key impurities during bioprocess development

Gauthier Husson[1], Aurélie Delangle[2], John O'Hara[3], Annick Gervais[2], Alain Van Dorsselaer[1], Dan Bracewell[4], Christine Carapito[1*]

[1]Laboratoire de Spectrométrie de Masse BioOrganique, Université de Strasbourg, CNRS, IPHC, UMR 7178, F-67000 Strasbourg, France
[2]Department of Analytical Sciences Biologicals, UCB Pharma s.a., Chemin du Foriest, B-1420 Braine L'alleud, Belgium
[3]Department of Analytical Sciences Biologicals, UCB Pharma s.a., 216 Bath Road, Slough SL1 4EN, UK
[4]Dept. Biochemical Engineering, University College London, Gower Street, London, WC1E 6BT, UK

## Table of Contents

***Detailed experimental procedures***

***Supplementary Table S1:*** Heat map representing all detected HCP quantifications by the Top 3-ID-DIA approach. Only HCP quantified with a coefficient of variation (CV) below 20% are shown.

***Supplementary Table S2:*** List of 10 selected HCP for targeted absolute quantification and rationale of their selection. *: abundance selection criterion based on results obtained from preliminary DIA analyses (data not shown). **: only one specific peptide, SAPATGGVK, could be selected for Histone H3 family (protein sets tr|G3H2T7|G3H2T7_CRIGR and tr|G3HDT2|G3HDT2_CRIGR). Peptide AGLQFPVGR is shared between histones H2A, H3 and H4 (protein sets tr|A0A061IIJ6|A0A061IIJ6_CRIGR, tr|G3H2T7|G3H2T7_CRIGR, tr|G3HDS3|G3HDS3_CRIGR, tr|G3HDT6|G3HDT6_CRIGR, tr|G3HPV6|G3HPV6_CRIGR, tr|G3I968|G3I968_CRIGR).

***Supplementary Table S3:*** DIA-SWATH windows setup optimised for CCCF (a) and PPA (b) fractions.

***Supplementary Table S4:*** Quantification of 10 selected HCP by Top 3 estimation (green), absolute ID-SRM (red) and absolute ID-DIA (blue). Presented LLOQ values were averaged across samples. Peptide quantifications falling in the linear range of the assay were averaged to obtain protein quantification that are provided as ppm values. Top 3-DIA estimations were obtained using 3 peptides per protein, while ID-SRM and ID-DIA absolute quantifications were obtained using 2 peptides per protein except values with * for which only one peptide was used. "<LLOQ" means that measured intensities were below the LLOQ and "ND" means that no signal could be detected. A graphic view of these results is presented in Supplementary Figure S4.

***Supplementary Figure S1:*** Peptides' calibration curves and LLOQ obtained by ID-SRM.

***Supplementary Figure S2:*** Peptides' calibration curves and LLOQ obtained by Top 3-ID-DIA.

***Supplementary Figure S3:*** Cell culture monitoring. The cell count (a), cell viability (b) and mAb concentration (c) were monitored during cell culture.

***Supplementary Figure S4:*** Graphic view of Supplementary Table S4. HCP were quantified by Top 3 estimations (green), absolute ID-SRM (red) and absolute ID-DIA (blue).

***Supplementary Figure S5:*** Benchmarking of sensitivity, accuracy and precision of the Top 3-ID-DIA method. (a) Sensitivity of Top 3-ID-DIA and ID-SRM were compared using LLOQ values obtained with calibration curves of AQUA peptides spiked in CCCF and PPA matrix. (b) Accuracy was assessed by comparing quantification values obtained by Top 3 estimation and ID-DIA absolute quantification to ID-SRM absolute quantification. (c) Precision was probed using calculated coefficients of variation of technical triplicates for Top 3 estimations, ID-DIA and ID-SRM absolute quantification.

***Supplementary Figure S6:*** Interesting process-related behaviour of several HCP linked to (a) the shear stress during harvest procedure, (b) the cell culture duration or (c) the protein A purification protocol.

## Detailed experimental procedures

***Cell culture***. An IgG4 A33 mAb producing CHO-DG44 cell line (provided by UCB Pharma, Brussels, Belgium) was cultivated in batch mode using a protein free and chemically defined CD CHO medium (Thermo Fisher Scientific, Waltham, MA, USA) supplemented with 6 mM Glutamine (Thermo Fisher Scientific) and 5 nM Methotrexate (Merck, Darmstadt, Germany). Cells were grown in 1 L Erlenmeyer flasks (300 mL working volume) and incubated at 36.5°C with 5% CO2 on an orbital shaker (123 rpm). The cell concentration and viability were monitored every day using a Vi-Cell (Beckman Coulter, Brea, CA, USA). Viable cells were distinguished from dead cells using the trypan blue dye exclusion method.

***USD Shear device***. The ultra scale-down (USD) shear device[1] developed at University College London generated a shear environment used to mimic the surface adsorption and shear forces encountered at industrial scale in the feed zone of a disc stack centrifuge. The shear device was filled with cell culture fluid and operated at 6 000 rpm (for a low shear stress that may be experienced in a hydrohermetic feed zone of a disc stack centrifuge) or 12 000 rpm (for a high shear stress experienced in a non-hermetic feed zone) for 20 sec, equivalent to, respectively, a maximum shear strain rate[2] of 1.59 x 104 $s^{-1}$ or 3.17 x 104 $s^{-1}$, or a maximum energy dissipation rate[3] of 0.019 x 106 W.kg$^{-1}$ or 0.37 x 106 W.kg$^{-1}$.

***Harvest procedure***. The cell culture fluid was centrifuged at 3 300 g during 10 min to pellet the cells. The pH and conductivity of the supernatant were adjusted to meet the protein A chromatography equilibration buffer characteristics. The supernatant was then filtered through a 0.22 µm Stericup filter unit (Merck), and the clarified cell culture fluid (CCCF) fraction was collected.

***Protein A chromatography.*** mAbs were purified using 1 mL HiTrap MabSelect SuRe columns (GE Healthcare Life Sciences, Pittsburgh, PA) on an AKTA Pure system (GE Healthcare Life Sciences). Two purification types were performed at 1 mL/min using either (i) a standard protocol[4] or (ii) a modified protocol[5]. (i) Standard protocol: equilibration step (5 column volumes (CV) of PBS, pH 7.4) followed by loading of an appropriate volume of CCCF for 20 mg mAb. The column was washed with loading buffer, and the mAb was eluted (0.1 M citrate pH 3.6). (ii) Modified protocol: equilibration step (5 CV 25 mM Tris, 100 mM NaCl pH 7.4) followed by loading of an appropriate volume of CCCF for 20 mg mAb. The column was washed with loading buffer. An intermediate wash (5 CV 25 mM Tris, 10% isopropanol, 1 M urea, pH 9) and a pre-elution wash (3 CV 50 mM citrate, pH 4.4) were performed before mAb elution (100 mM acetate, pH 3.6). After elution, the post protein A (PPA) fractions were directly neutralised to pH 6 using 2 M Tris HCl pH 8.8. A dedicated new column was used for each purification.

***Samples.*** Four CCCF and eight PPA fractions were generated (Figure 2). CCCF 1, 2 and 3 fractions were obtained after 7 days of culture. For CCCF 1, the cells did not undergo shear stress, while low and high shear stresses were applied to CCCF 2 and 3 fractions, respectively. CCCF 4 fraction was collected after 10 days of culture, without shearing. CCCF 1 fraction was purified using the standard protein A purification protocol in triplicate resulting in PPA 1, 2 and 3 fractions, and using the modified protocol[5] giving PPA 4 fraction. PPA 5 and 6 were obtained using the standard protein A purification protocol of CCCF 2 and 3 fractions, respectively. CCCF 4 fraction was purified using the standard protocol giving PPA 7 fraction, and using the modified protocol giving PPA 8 fraction. Aliquots were collected for CCCF and PPA fractions, and mixed with 4 volumes of cold acetone. After 1h30 incubation at -20 °C, the samples were centrifuged at 14 000 g for 10 min. The supernatant was discarded and the protein pellets were stored at -80°C.

***mAb quantification.*** The mAb titre was determined using a High Pressure Liquid Chromatography (HPLC) Agilent 1100 Series HPLC System (Agilent Technologies, Santa Clara, CA, USA) and a 1 mL HiTrap Protein G HP column (GE Healthcare Life Sciences). 100 µL of sample were loaded onto the column at 1 mL/min. The column was washed with 20 mM sodium phosphate, pH 7, and the mAb was eluted with 20 mM glycine, pH 2.8. Peaks were integrated and the mAb concentration was determined using a standard curve of purified mAb (data not shown).

***HCP-ELISA***. The HCP were quantified using the CHO HCP ELISA kit, 3G (Cygnus Technologies, Southport, NC, USA) in technical triplicates according to the manufacturer's protocol.

***Global protein quantification.*** The protein pellets were resuspended in gel loading buffer (10 mM Tris, 1 mM EDTA, 5% β-Mercaptoethanol, 5% SDS, 10% glycerol, pH 6.8), and the total protein concentration was determined using the RC DC Protein Assay kit (Bio-Rad Laboratories, Hercules, CA, USA) following manufacturer's instructions.

***SDS-PAGE.*** Proteins were loaded on an SDS-PAGE (sodium dodecyl sulfate – polyacrylamide gel electrophoresis) gel and separated in 23 bands for spectral library generation (pooled CCCF and pooled PPA fractions), or stacked in a single band for HCP quantification. Bands were excised and cut, and proteins were in-gel reduced (10 mM dithiothreitol in 25 mM NH4HCO3), alkylated (55 mM iodoacetamide in 25 mM NH4HCO3) and digested overnight using modified porcine trypsin (Promega) at 37 °C. Peptides were extracted from the gel using acetonitrile, vacuum dried and resolubilised with an adequate volume of 98% water, 2% acetonitrile, 0.1% formic acid. Retention time standards (iRT, Biognosys, Zurich, Switzerland), four accurately quantified standard proteins (on-column 100 fmol ADH (yeast alcohol dehydrogenase P00330), 20 fmol PYGM (phosphorylase b P00489), 5 fmol BSA (bovin serum albumin P02769) and 2 fmol ENL (yeast enolase P00924) from the MassPREP Digestion Standard Kit, Waters, Milford, MA, USA) and a concentration-balanced mixture of 20 accurately quantified stable isotope labelled peptides (AQUA peptides, Thermo Fisher Scientific) were spiked into each sample.

***Mass spectrometry analysis.*** Data dependent acquisition (DDA) and data independent acquisition-sequential windowed acquisition of all theoretical fragment ion mass spectra (DIA-SWATH) analyses were performed on an Eksigent NanoLC 400 system operated in microLC-mode and coupled to a TripleTOF 6600 quadrupole-time of flight mass spectrometer (both from SCIEX, Framingham, MA, USA). Selected reaction monitoring (SRM) analyses were performed on a Dionex UltiMate 3000 operated in microLC-mode and coupled to a TSQ Vantage triple quadrupole mass spectrometer (both from Thermo Fisher Scientific). On both couplings, 8 µg of peptides were separated on a ZORBAX 300SB-C18 column (150 mm x 300 µm with 3.5 µm diameter particles, Agilent Technologies). The solvents consisted of 0.1% formic acid in water (solvent A) and 0.1% formic acid in acetonitrile (solvent B). The DDA, DIA and SRM analyses were performed using the same LC gradient: peptides were loaded on column and eluted at 5 µL/min with a linear gradient from 2 to 35% B in 95 min, 35 to 80% B in 1 min, isocratic at 80% B for 5 min, down to 2% B in 1 min and isocratic at 2% B for 13 min.

For DDA analyses, the MS1 spectra were collected from 350 to 1250 m/z for 250 ms. The most intense precursor ions with charge states 2-4 were selected for fragmentation, and MS2 spectra were collected in high sensitivity mode from 200 to 1600 m/z using dynamic accumulation, with an accumulation time for high intensity peaks of 25 ms and a total cycle time of 2.8 sec. After fragmentation, the precursor ions were excluded for 18 sec.

For DIA analyses, we used two distinct 75 variable windows SWATH methods, optimised either for CCCF or PPA samples. The methods were generated using the SWATH Variable Window Calculator (SCIEX) applied to DDA analyses of pooled samples, to allocate the same precursor ion density to all windows covering the 350-1250 m/z range, with an overlap of 1 m/z. The optimised windows setup is described in Supplementary Table S3. MS2 spectra were acquired in high sensitivity mode from 200 to 1600 m/z for 40 ms. An additional MS1 scan per cycle was recorded for 150 ms, resulting in a total cycle time of 3.2 sec. The collision energy was calculated using the equation of doubly charged precursor ions (collision energy = 0.049 x precursor m/z - 1) with m/z in the middle of the isolation window, and a collision energy spread of 5 volts was applied around the calculated value.

For SRM analyses, we developed a scheduled SRM method with 4 minutes time windows and 3 seconds cycle time using crude stable isotope labelled peptides (PEPotec peptides, Thermo Fisher Scientific). Collision energies were optimised for each transition by ramping collision energy around the calculated value (5 steps of 2 V each).

For both DIA and SRM analyses, accurately quantified stable isotope labelled peptides (AQUA peptides, Thermo Fisher Scientific) were used for accurate absolute quantification. Signals and retention times of AQUA peptides were used as internal controls to ensure system stability. With an average chromatographic peak duration of 30 sec, the cycle time was set to 3 sec to obtain about 10 data points per chromatographic peak and ensure precise quantification. Samples were analysed in technical triplicates.

***Spectral library generation.*** Profile-mode .wiff files from DDA analyses of the 23 gel bands of both CCCF and PPA pools were processed using Protein Pilot 5.0 software (SCIEX) and the recalibrated peak lists were exported as .mgf files. Peptides and proteins were searched with Mascot 2.5.1 search engine (Matrix Science, London, UK) against a custom protein database containing all sequences of the Chinese Hamster (taxonomy ID = 10 029) extracted from UniProtKB/TrEMBL, the retention time standards (iRT peptides concatenated as a unique protein sequence), the four standard proteins (ADH, PYGM, BSA, ENO), the heavy and light chains of the A33 mAb, common contaminants and all corresponding reverse sequences concatenated using the MSDA software tool (https://msda.unistra.fr[6]). The following parameters were used: trypsin digestion with 1 missed cleavage allowed, MS tolerance of 15 ppm, MS/MS tolerance of 0.05 Da, cysteine carbamidomethylation as fixed modification, and methionine oxidation as variable modification. The search results were validated using the in-house developed ProlineStudio 1.4 software (http://proline.profiproteomics.fr/[7]) to keep only identifications with a Mascot Ion Score above 25, a pretty rank (as defined by Mascot) equal to 1, and a false discovery rate below 1% at the peptide (on e-value) and protein (on Mascot modified MudPIT score) levels. In total, 25 338 unique peptides were identified, corresponding to 3 220 protein sets.

***DIA targeted data extraction.*** DIA data were processed using Skyline[8] (version 3.5.9.10061). The spectral library was generated as described and validated proteotypic peptides were extracted with the following parameters (based on previous work[9] and in-house optimisations on standard samples, data not shown): the 6 most intense 1+ b- and y-type product ions were extracted, from ion 3 to last ion – 1, while the precursors with less than 3 transitions were excluded. Resolving power was set to 50 000, and a retention time tolerance of 5 min (+/- 2.5 min) was used. Retention times were predicted with iRT standards (Biognosys). Peaks were reintegrated using the target decoy approach (reverse sequences) of the mProphet peak-scoring model, and a Q-value was assigned to each peak. Peak integrations were manually checked and curated for HCP of interest. Total fragment area, detection Q value and library dot-product were exported for each peptide in .csv files.

***Top 3 estimation.*** Only peptides with Q-value below 0.01 (corresponding to a false discovery rate of 1%) and dot-product above 0.6 were kept. The total fragment areas were summed among charge states for each peptide, and the 3 best responding peptides were summed for each protein. Only proteins quantified in at least two replicates in at least one sample were kept, independently for CCCF and PPA fractions. The universal signal response factor[10] (signal / mol of protein) was calculated using PYGM and was used to estimate mol quantities of all proteins. Using molecular weights and injected mAb quantities, individual HCP amounts in ppm were estimated. Quantifications with a coefficient of variation (CV) below 20% among technical triplicates were used to build a heat map (Supplementary Table S1) and calculate total HCP amounts in each sample.

***Selection of 10 HCP and their proteotypic peptides.*** Ten HCP were chosen based on their potential immunogenicity, proteolytic activity, purification behaviour, or estimated abundance using preliminary data acquired on the samples (data not shown). Elongation factor 1-alpha 1, Pyruvate kinase, Histone H3, Clusterin, Eukaryotic translation initiation factor 3 subunit L, Putative phospholipase B-like 2, Serine protease HTRA1, Cathepsin L1 were reported as difficult to remove in previous studies[11-17]. Serine protease HTRA1 and Cathepsin L1 were selected as proteases can affect the mAb product integrity[18-19]. In preliminary analyses, Pyruvate kinase was detected as very abundant in CCCF fractions, while cytoplasmic Isoleucyl-tRNA synthetase and HEAT repeat-containing protein 3 were very low abundant in PPA fractions. Finally, Phospholipase B-like 2 is known to induce a specific immune response in patients[20-21] in addition to being of big concern for purification process[17]. Two peptides were chosen per selected protein among peptides previously identified in DDA analyses, according to their proteotypicity, absence of missed trypsin cleavage, absence of amino acid residues prone to posttranslational modification, sequence length, and MS2 spectra quality. It is to note that for Histone H3, due to high sequence homologies among the histones family, only one specific peptide could be selected (SAPATGGVK) while the second (AGLQFPVGR) is shared among Histone H2A, H3 and H4. The chosen peptides were synthesised with stable isotope labelled C-terminal amino acids and used for accurate absolute quantification (Supplementary Table S2).

***Accurate quantification.*** Six transitions were analysed for each precursor ion for both SRM and DIA approaches. If comprised in the linear range of the assay as determined by calibration curves, the ratios between endogen and stable isotope labelled AQUA peptides were used to calculate the mol amounts of endogenous peptides, which were averaged to calculate the mol amounts of corresponding proteins. Using molecular weights and injected mAb quantities, individual HCP amounts in ppm were calculated.

***Calibration curves and LLOQ determination.*** For DIA and SRM acquisitions, calibration curves were realised for each stable isotope labelled standard peptide to determine the linearity range and the lower limit of quantification (LLOQ). Different amounts of stable isotope labelled peptides were spiked into a representative matrix (CCCF or PPA pool). The matrix effect was found to be very limited and thus only one merged calibration curve was built for each peptide using results for CCCF and PPA fractions. To be included in the calibration curve, data points must fulfil following criteria: show a CV precision below 20% among technical triplicates; the coefficient of determination $R^2$ must be higher than 0.99 between the total fragment area and the injected amount; the coefficient of determination $R^2$ must be higher than 0.99 between the back calculated and the real injected amounts; calibration points must show an accuracy between 80 and 120 % by back calculating expected injected amounts using regression equations after logarithmic transformation. Finally, calibration curves must comprise at least 3 data points. Thereby, we determined a quantification linearity range for each peptide from 1.3 to 4.7 orders of magnitude for SRM approach, and from 2 to 4.7 orders of magnitude for DIA approach. The LLOQ corresponds the lowest point of the calibration curve. Proteins' LLOQ are the lowest LLOQ of their corresponding peptides.

## Supplementary Table S2

| Protein name<br>(Protein ID) | Selection criteria | Peptides sequences |
|---|---|---|
| Elongation factor 1-alpha 1<br>(sp\|P62629\|EF1A1_CRIGR) | Difficult to remove[11, 13] | LPLQDVYK<br>QLIVGVNK |
| Pyruvate kinase<br>(tr\|A0A098KXC0\|A0A098KXC0_CRIGR) | Difficult to remove[11, 13, 15], very high abundance in CCCF fractions* | LDIDSAPITAR<br>NTGIICTIGPASR |
| Histone H3**<br>(tr\|G3H2T7\|G3H2T7_CRIGR) | Difficult to remove[16] | AGLQFPVGR<br>SAPATGGVK |
| HEAT repeat-containing protein 3<br>(tr\|G3H5M8\|G3H5M8_CRIGR) | Very low abundance in PPA fractions* | LGPLLLDSSLAVR<br>SQAEIINAILK |
| Clusterin<br>(tr\|G3HNJ3\|G3HNJ3_CRIGR) | Difficult to remove[11-12, 15] | EIQNAVQGVK<br>LTQQYNELLHSLQTK |
| Isoleucyl-tRNA synthetase, cytoplasmic<br>(tr\|G3HP24\|G3HP24_CRIGR) | Very low abundance in PPA fractions* | ESIDHLTIPSR<br>QLSSEELEQFQK |
| Eukaryotic translation initiation factor 3 subunit L<br>(tr\|G3I505\|G3I505_CRIGR) | Difficult to remove[11] | LAGFLDLTEQEFR<br>LHSLLGDYYQAIK |
| Putative phospholipase B-like 2<br>(tr\|G3I6T1\|G3I6T1_CRIGR) | Immunogenic[20-21], difficult to remove[14-15, 17] | LALDGATWADIFK<br>SVLLDAASGQLR |
| Serine protease HTRA1<br>(tr\|G3IBF4\|G3IBF4_CRIGR) | Protease, difficult to remove[11, 13, 15] | LPVLLLGR<br>VTAGISFAIPSDK |
| Cathepsin L1<br>(tr\|G3INC5\|G3INC5_CRIGR) | Protease, difficult to remove[13] | GLDSEESYPYEAK<br>QLVNGYK |

## Supplementary Table S3

(a)

| SWATH Exp Index: | Start Mass (Da) | Stop Mass (Da) |
|---|---|---|
| SWATH Exp 1: | 349.50 | 360.90 |
| SWATH Exp 2: | 359.90 | 371.60 |
| SWATH Exp 3: | 370.60 | 381.50 |
| SWATH Exp 4: | 380.50 | 390.10 |
| SWATH Exp 5: | 389.10 | 398.20 |
| SWATH Exp 6: | 397.20 | 405.80 |
| SWATH Exp 7: | 404.80 | 413.50 |
| SWATH Exp 8: | 412.50 | 420.20 |
| SWATH Exp 9: | 419.20 | 426.50 |
| SWATH Exp 10: | 425.50 | 432.80 |
| SWATH Exp 11: | 431.80 | 439.10 |
| SWATH Exp 12: | 438.10 | 445.00 |
| SWATH Exp 13: | 444.00 | 450.80 |
| SWATH Exp 14: | 449.80 | 457.10 |
| SWATH Exp 15: | 456.10 | 463.40 |
| SWATH Exp 16: | 462.40 | 469.30 |
| SWATH Exp 17: | 468.30 | 475.60 |
| SWATH Exp 18: | 474.60 | 482.30 |
| SWATH Exp 19: | 481.30 | 489.10 |
| SWATH Exp 20: | 488.10 | 495.40 |
| SWATH Exp 21: | 494.40 | 501.70 |
| SWATH Exp 22: | 500.70 | 508.00 |
| SWATH Exp 23: | 507.00 | 514.30 |
| SWATH Exp 24: | 513.30 | 520.60 |
| SWATH Exp 25: | 519.60 | 526.90 |
| SWATH Exp 26: | 525.90 | 533.20 |
| SWATH Exp 27: | 532.20 | 539.00 |
| SWATH Exp 28: | 538.00 | 545.30 |
| SWATH Exp 29: | 544.30 | 551.70 |
| SWATH Exp 30: | 550.70 | 558.00 |
| SWATH Exp 31: | 557.00 | 564.30 |
| SWATH Exp 32: | 563.30 | 571.00 |
| SWATH Exp 33: | 570.00 | 577.80 |
| SWATH Exp 34: | 576.80 | 584.50 |
| SWATH Exp 35: | 583.50 | 591.30 |
| SWATH Exp 36: | 590.30 | 598.00 |
| SWATH Exp 37: | 597.00 | 604.30 |
| SWATH Exp 38: | 603.30 | 611.10 |
| SWATH Exp 39: | 610.10 | 617.80 |
| SWATH Exp 40: | 616.80 | 624.60 |
| SWATH Exp 41: | 623.60 | 631.30 |
| SWATH Exp 42: | 630.30 | 638.50 |
| SWATH Exp 43: | 637.50 | 645.70 |
| SWATH Exp 44: | 644.70 | 653.40 |
| SWATH Exp 45: | 652.40 | 661.00 |
| SWATH Exp 46: | 660.00 | 669.60 |
| SWATH Exp 47: | 668.60 | 678.10 |
| SWATH Exp 48: | 677.10 | 686.70 |
| SWATH Exp 49: | 685.70 | 696.10 |
| SWATH Exp 50: | 695.10 | 705.60 |
| SWATH Exp 51: | 704.60 | 715.50 |
| SWATH Exp 52: | 714.50 | 725.40 |
| SWATH Exp 53: | 724.40 | 735.70 |
| SWATH Exp 54: | 734.70 | 746.50 |
| SWATH Exp 55: | 745.50 | 757.80 |

| | | |
|---|---|---|
| SWATH Exp 56: | 756.80 | 769.90 |
| SWATH Exp 57: | 768.90 | 782.50 |
| SWATH Exp 58: | 781.50 | 796.50 |
| SWATH Exp 59: | 795.50 | 810.90 |
| SWATH Exp 60: | 809.90 | 826.60 |
| SWATH Exp 61: | 825.60 | 842.40 |
| SWATH Exp 62: | 841.40 | 859.00 |
| SWATH Exp 63: | 858.00 | 876.10 |
| SWATH Exp 64: | 875.10 | 893.20 |
| SWATH Exp 65: | 892.20 | 910.80 |
| SWATH Exp 66: | 909.80 | 927.00 |
| SWATH Exp 67: | 926.00 | 943.60 |
| SWATH Exp 68: | 942.60 | 959.80 |
| SWATH Exp 69: | 958.80 | 975.60 |
| SWATH Exp 70: | 974.60 | 994.90 |
| SWATH Exp 71: | 993.90 | 1019.20 |
| SWATH Exp 72: | 1018.20 | 1055.70 |
| SWATH Exp 73: | 1054.70 | 1102.00 |
| SWATH Exp 74: | 1101.00 | 1170.40 |
| SWATH Exp 75: | 1169.40 | 1249.60 |

(b)

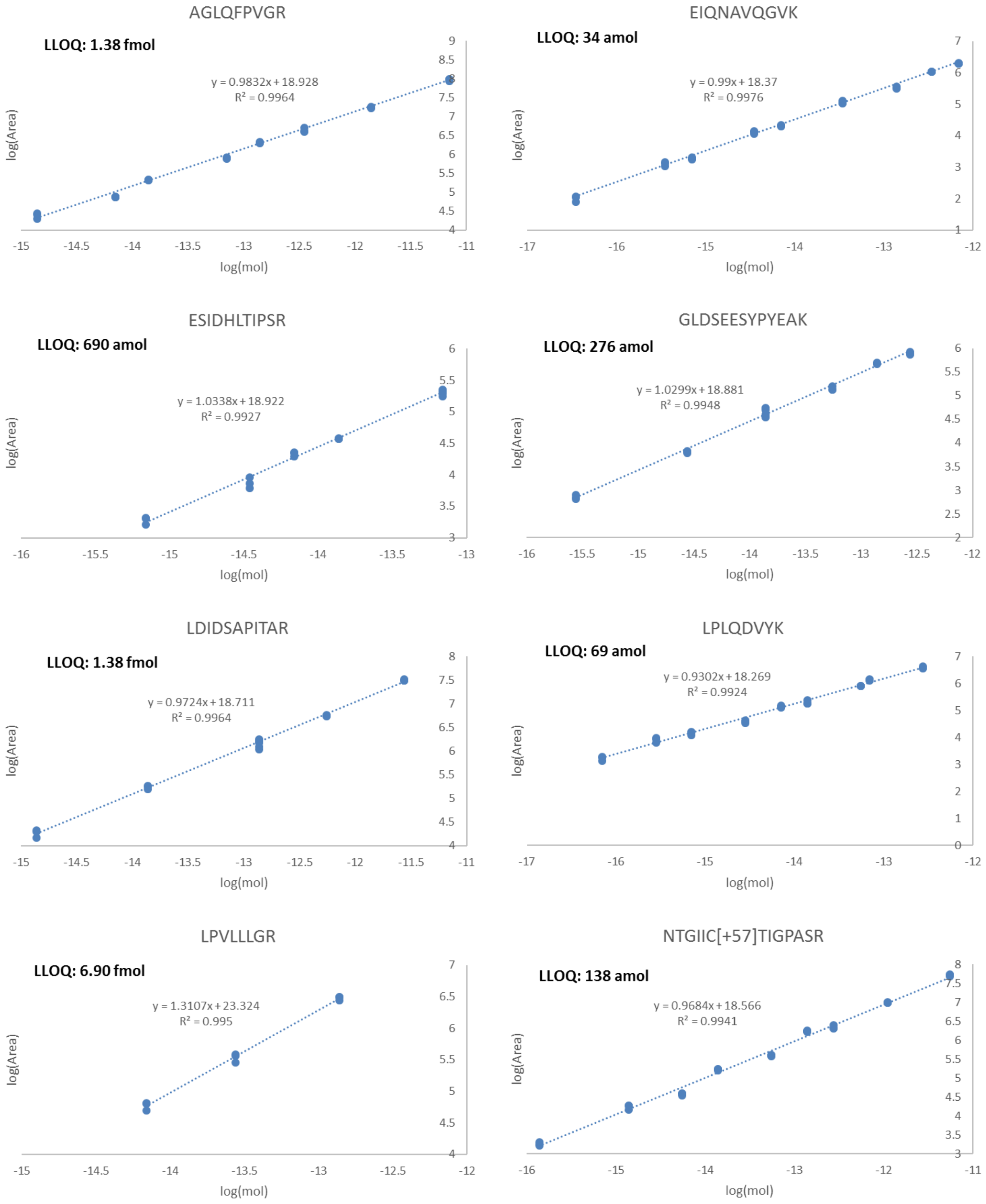
| SWATH Exp Index: | Start Mass (Da) | Stop Mass (Da) |
|---|---|---|
| SWATH Exp 1: | 349.50 | 365.40 |
| SWATH Exp 2: | 364.40 | 377.50 |
| SWATH Exp 3: | 376.50 | 387.40 |
| SWATH Exp 4: | 386.40 | 394.10 |
| SWATH Exp 5: | 393.10 | 400.40 |
| SWATH Exp 6: | 399.40 | 405.80 |
| SWATH Exp 7: | 404.80 | 410.80 |
| SWATH Exp 8: | 409.80 | 416.20 |
| SWATH Exp 9: | 415.20 | 421.10 |
| SWATH Exp 10: | 420.10 | 426.10 |
| SWATH Exp 11: | 425.10 | 430.60 |
| SWATH Exp 12: | 429.60 | 435.50 |
| SWATH Exp 13: | 434.50 | 440.00 |
| SWATH Exp 14: | 439.00 | 444.10 |
| SWATH Exp 15: | 443.10 | 449.00 |
| SWATH Exp 16: | 448.00 | 454.40 |
| SWATH Exp 17: | 453.40 | 461.20 |
| SWATH Exp 18: | 460.20 | 469.70 |
| SWATH Exp 19: | 468.70 | 479.60 |
| SWATH Exp 20: | 478.60 | 491.30 |
| SWATH Exp 21: | 490.30 | 499.90 |
| SWATH Exp 22: | 498.90 | 508.00 |
| SWATH Exp 23: | 507.00 | 515.60 |
| SWATH Exp 24: | 514.60 | 523.30 |
| SWATH Exp 25: | 522.30 | 530.00 |
| SWATH Exp 26: | 529.00 | 536.30 |
| SWATH Exp 27: | 535.30 | 542.60 |
| SWATH Exp 28: | 541.60 | 549.00 |
| SWATH Exp 29: | 548.00 | 554.80 |
| SWATH Exp 30: | 553.80 | 561.10 |
| SWATH Exp 31: | 560.10 | 568.80 |
| SWATH Exp 32: | 567.80 | 576.40 |
| SWATH Exp 33: | 575.40 | 584.50 |

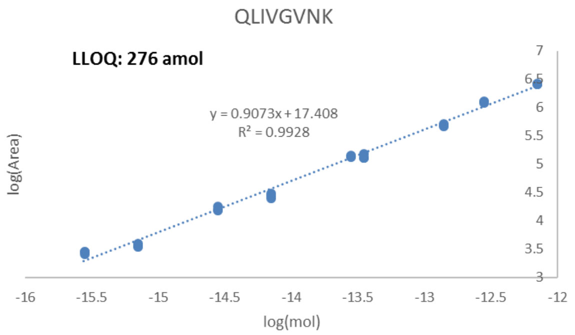
| | | |
|---|---|---|
| SWATH Exp 34: | 583.50 | 592.60 |
| SWATH Exp 35: | 591.60 | 600.30 |
| SWATH Exp 36: | 599.30 | 607.50 |
| SWATH Exp 37: | 606.50 | 614.70 |
| SWATH Exp 38: | 613.70 | 621.00 |
| SWATH Exp 39: | 620.00 | 627.70 |
| SWATH Exp 40: | 626.70 | 634.50 |
| SWATH Exp 41: | 633.50 | 641.70 |
| SWATH Exp 42: | 640.70 | 648.40 |
| SWATH Exp 43: | 647.40 | 655.20 |
| SWATH Exp 44: | 654.20 | 662.40 |
| SWATH Exp 45: | 661.40 | 670.50 |
| SWATH Exp 46: | 669.50 | 679.50 |
| SWATH Exp 47: | 678.50 | 689.80 |
| SWATH Exp 48: | 688.80 | 701.50 |
| SWATH Exp 49: | 700.50 | 713.20 |
| SWATH Exp 50: | 712.20 | 726.30 |
| SWATH Exp 51: | 725.30 | 740.70 |
| SWATH Exp 52: | 739.70 | 756.90 |
| SWATH Exp 53: | 755.90 | 778.50 |
| SWATH Exp 54: | 777.50 | 808.60 |
| SWATH Exp 55: | 807.60 | 833.80 |
| SWATH Exp 56: | 832.80 | 850.50 |
| SWATH Exp 57: | 849.50 | 866.20 |
| SWATH Exp 58: | 865.20 | 881.10 |
| SWATH Exp 59: | 880.10 | 896.80 |
| SWATH Exp 60: | 895.80 | 913.00 |
| SWATH Exp 61: | 912.00 | 922.90 |
| SWATH Exp 62: | 921.90 | 931.90 |
| SWATH Exp 63: | 930.90 | 940.90 |
| SWATH Exp 64: | 939.90 | 949.50 |
| SWATH Exp 65: | 948.50 | 957.10 |
| SWATH Exp 66: | 956.10 | 964.30 |
| SWATH Exp 67: | 963.30 | 971.50 |
| SWATH Exp 68: | 970.50 | 979.20 |
| SWATH Exp 69: | 978.20 | 988.60 |
| SWATH Exp 70: | 987.60 | 999.40 |
| SWATH Exp 71: | 998.40 | 1012.00 |
| SWATH Exp 72: | 1011.00 | 1030.00 |
| SWATH Exp 73: | 1029.00 | 1061.10 |
| SWATH Exp 74: | 1060.10 | 1134.40 |
| SWATH Exp 75: | 1133.40 | 1249.60 |

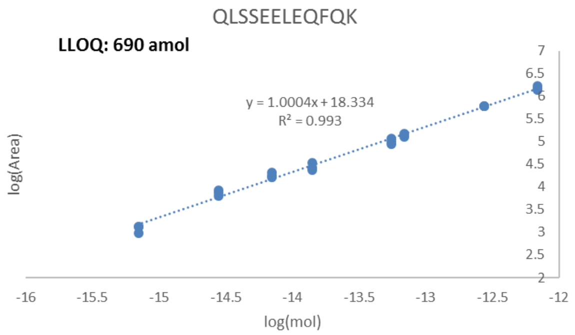
## Supplementary Table S4

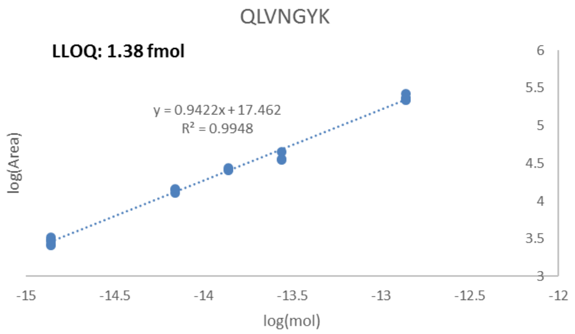
| | LLOQ (ppm) in CCCF / PPA | CCCF 1 | CCCF 2 | CCCF 3 | CCCF 4 | PPA 1 | PPA 2 | PPA 3 | PPA 4 | PPA 5 | PPA 6 | PPA 7 | PPA 8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Elongation factor 1-alpha 1<br>sp\|P62629\|EF1A1_CRIGR | - | 3 186 ± 60 | 3 996 ± 75 | 8 733 ± 237 | 3 140 ± 106 | 65 ± 2 | 113 ± 2 | 79 ± 2 | 63 ± 3 | 62 ± 2 | 45 ± 2 | 49 ± 2 | 106 ± 3 |
| | 1.5 / 0.5 | 2 851 ± 59 | 2 880 ± 109 | 3 995 ± 156 | 1 900 ± 100 | 54 ± 0 | 106 ± 4 | 72 ± 4 | 21 ± 1 | 35 ± 2 | 24 ± 2 | 17 ± 1 | 21 ± 1 |
| | 0.6 / 0.2 | 2 613 ± 115 | 2 699 ± 65 | 3 778 ± 205 | 1 732 ± 29 | 68 ± 2 | 118 ± 5 | 87 ± 3 | 34 ± 2 | 50 ± 4 | 33 ± 2 | 29 ± 4 | 37 ± 5 |
| Pyruvate kinase<br>tr\|A0A098KXC0\|A0A098KXC0_CRIGR | - | 9 116 ± 365 | 10 709 ± 318 | 15 640 ± 735 | 10 846 ± 139 | 62 ± 4 | 77 ± 3 | 63 ± 3 | 52 ± 1 | 60 ± 3 | 47 ± 6 | 183 ± 10 | 445 ± 10 |
| | 4.8 / 1.6 | 13 674 ± 207 | 15 948 ± 765 | 23 681 ± 588 | 16 839 ± 823 | 37 ± 2 | 56 ± 2 | 44 ± 1 | 42* ± 4 | 28 ± 2 | 24 ± 1 | 157 ± 7 | 536 ± 34 |
| | 4.8 / 1.6 | 15 494 ± 376 | 17 308 ± 401 | 26 017 ± 531 | 19 627 ± 290 | 54 ± 2 | 70 ± 3 | 57 ± 3 | 36 ± 1 | 42 ± 3 | 35 ± 1 | 172 ± 9 | 456 ± 26 |
| Histone H3<br>tr\|G3H2T7\|G3H2T7_CRIGR | - | 2 500 ± 33 | 4 457 ± 123 | 6 829 ± 117 | 6 868 ± 392 | 2 ± 0 | 2 ± 0 | 2 ± 0 | 1 ± 0 | 1 ± 0 | 1 ± 0 | 1 ± 0 | 2 ± 0 |
| | 17.3 / 6.0 | 2 204 ± 39 | 4 414 ± 241 | 6 795 ± 348 | 6 722 ± 193 | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ |
| | 1.7 / 0.6 | 2 271 ± 32 | 4 761 ± 61 | 7 768 ± 134 | 7 694 ± 316 | 5* ± 0 | 6* ± 0 | 6* ± 1 | 3* ± 1 | 2* ± 1 | 3* ± 0 | 2* ± 0 | 2* ± 0 |
| HEAT repeat-containing protein 3<br>tr\|G3H5M8\|G3H5M8_CRIGR | - | 80 ± 3 | 96 ± 1 | 120 ± 1 | 77 ± 4 | ND | ND | ND | ND | ND | ND | ND | ND |
| | - | No valid calibration curve for both peptides | | | | | | | | | | | |
| | 36 / 12 | 100* ± 6 | 91* ± 9 | 75* ± 6 | 68* ± 6 | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ |
| Clusterin<br>tr\|G3HNJ3\|G3HNJ3_CRIGR | - | 2 341 ± 88 | 1 693 ± 61 | 1 518 ± 52 | 1 269 ± 56 | 43 ± 1 | 58 ± 2 | 42 ± 0 | 38 ± 2 | 38 ± 3 | 19 ± 2 | 21 ± 1 | 42 ± 1 |
| | 0.8 / 0.3 | 4 054* ± 161 | 1 551* ± 95 | 1 183* ± 40 | 1 226* ± 56 | 61* ± 1 | 86* ± 3 | 54* ± 3 | 13* ± 0 | 20* ± 1 | 12* ± 1 | 7* ± 0 | 10* ± 1 |
| | 7.6 / 2.6 | 4 253* ± 129 | 1 747* ± 22 | 1 393* ± 12 | 1 424* ± 40 | 90* ± 3 | 120* ± 3 | 76* ± 2 | 23* ± 3 | 37* ± 2 | 22* ± 1 | 15* ± 3 | 18* ± 2 |
| Isoleucyl-tRNA synthetase, cytoplasmic<br>tr\|G3HP24\|G3HP24_CRIGR | - | 350 ± 17 | 483 ± 12 | 777 ± 4 | 464 ± 18 | ND | ND | ND | ND | ND | ND | ND | 35 ± 1 |
| | 42 / 15 | 178 ± 5 | 235 ± 26 | 370 ± 19 | 222 ± 4 | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | 18* ± 4 |
| | 42 / 15 | 231 ± 13 | 294 ± 15 | 492 ± 9 | 294 ± 13 | <LLOQ | 15* ± 1 | <LLOQ | 14* ± 5 | <LLOQ | <LLOQ | <LLOQ | 20* ± 3 |
| Eukaryotic translation initiation factor 3 subunit L<br>tr\|G3I505\|G3I505_CRIGR | - | 670 ± 16 | 759 ± 34 | 921 ± 34 | 633 ± 114 | 48 ± 1 | 53 ± 1 | 35 ± 2 | 41 ± 5 | 49 ± 3 | 46 ± 1 | 49 ± 2 | 60 ± 2 |
| | - | No valid calibration curve for both peptides | | | | | | | | | | | |
| | 89 / 31 | 386* ± 34 | 373* ± 29 | 257* ± 22 | 283* ± 19 | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ |
| Putative phospholipase B-like 2<br>tr\|G3I6T1\|G3I6T1_CRIGR | - | 1 018 ± 11 | 763 ± 17 | 618 ± 19 | 667 ± 13 | 59 ± 4 | 86 ± 3 | 60 ± 3 | 10 ± 0 | 64 ± 3 | 63 ± 3 | 40 ± 1 | 14 ± 1 |
| | 38 / 13 | 1 078* ± 98 | 730* ± 35 | 591* ± 27 | 639* ± 26 | 40* ± 1 | 55* ± 1 | 45* ± 2 | <LLOQ | 39* ± 1 | 40* ± 3 | 22* ± 1 | <LLOQ |
| | 3.8 / 1.3 | 889* ± 36 | 689* ± 30 | 583* ± 63 | 584* ± 14 | 57* ± 1 | 74* ± 2 | 62* ± 5 | 8* ± 1 | 54* ± 3 | 53* ± 2 | 36* ± 1 | 12* ± 1 |
| Serine protease HTRA1<br>tr\|G3IBF4\|G3IBF4_CRIGR | - | 240 ± 8 | 156 ± 10 | 140 ± 2 | 52 ± 1 | 16 ± 0 | 20 ± 1 | 16 ± 1 | 4 ± 0 | 10 ± 1 | 9 ± 1 | 2 ± 0 | 2 ± 0 |
| | 84 / 29 | 194* ± 14 | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ |
| | 1.7 / 0.6 | 160* ± 4 | 92* ± 2 | 43* ± 1 | 26* ± 3 | 18* ± 1 | 19* ± 1 | 13* ± 1 | 1* ± 0 | 6* ± 0 | 4* ± 0 | 1* ± 0 | <LLOQ |
| Cathepsin L1<br>tr\|G3INC5\|G3INC5_CRIGR | - | 534 ± 4 | 421 ± 4 | 408 ± 9 | 339 ± 9 | ND | ND | ND | ND | ND | ND | ND | ND |
| | 4.4 / 1.5 | 686 ± 19 | 368 ± 14 | 463* ± 17 | 347 ± 10 | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | <LLOQ | 2* ± 0 | 2* ± 0 |
| | 0.2 / 0.1 | 600 ± 37 | 366 ± 20 | 288 ± 8 | 322 ± 12 | 2 ± 0 | 2 ± 0 | 2 ± 1 | 2 ± 0 | 1 ± 1 | 2 ± 0 | 2 ± 1 | 2 ± 1 |

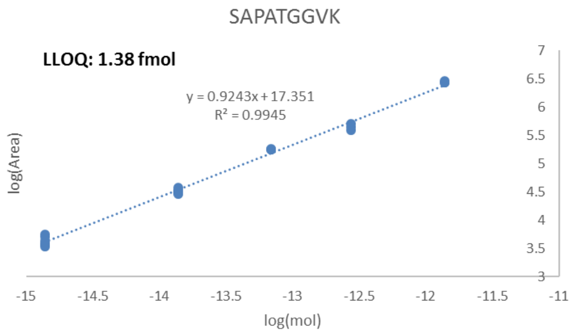
## Supplementary Figure S1

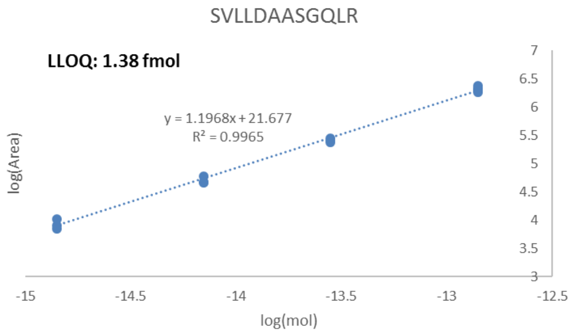



QLIVGVNK
LLOQ: 276 amol
y = 0.9073x + 17.408
R² = 0.9928
log(Area)
log(mol)
-16
-15.5
-15
-14.5
-14
-13.5
-13
-12.5
-12
3
3.5
4
4.5
5
5.5
6
6.5
7

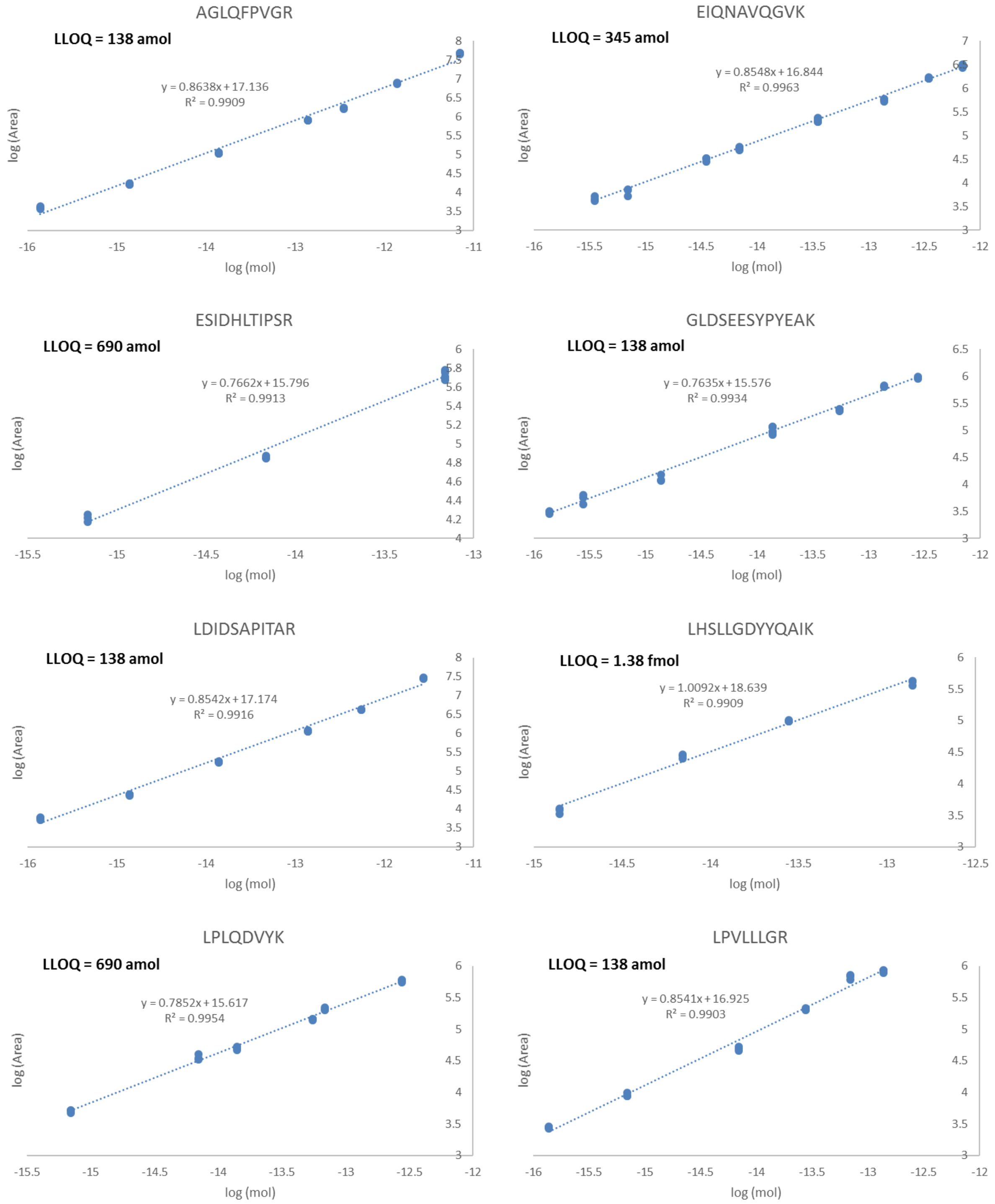

QLSSEELEQFQK
LLOQ: 690 amol
y = 1.0004x + 18.334
R² = 0.993
log(Area)
log(mol)
-16
-15.5
-15
-14.5
-14
-13.5
-13
-12.5
-12
2
2.5
3
3.5
4
4.5
5
5.5
6
6.5
7

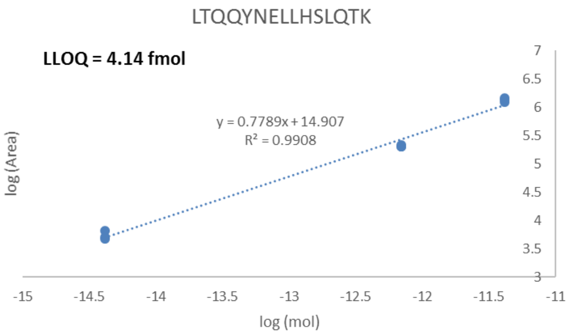

QLVNGYK
LLOQ: 1.38 fmol
y = 0.9422x + 17.462
R² = 0.9948
log(Area)
log(mol)
-15
-14.5
-14
-13.5
-13
-12.5
-12
3
3.5
4
4.5
5
5.5
6

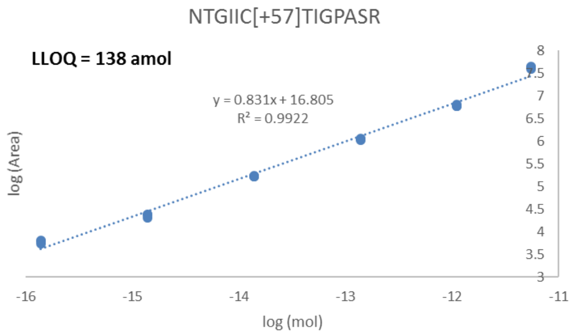

SAPATGGVK
LLOQ: 1.38 fmol
y = 0.9243x + 17.351
R² = 0.9945
log(Area)
log(mol)
-15
-14.5
-14
-13.5
-13
-12.5
-12
-11.5
-11
3
3.5
4
4.5
5
5.5
6
6.5
7

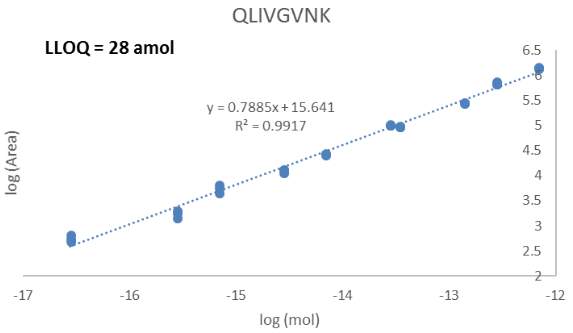

SVLLDAASGQLR
LLOQ: 1.38 fmol
y = 1.1968x + 21.677
R² = 0.9965
log(Area)
log(mol)
-15
-14.5
-14
-13.5
-13
-12.5
3
3.5
4
4.5
5
5.5
6
6.5
7

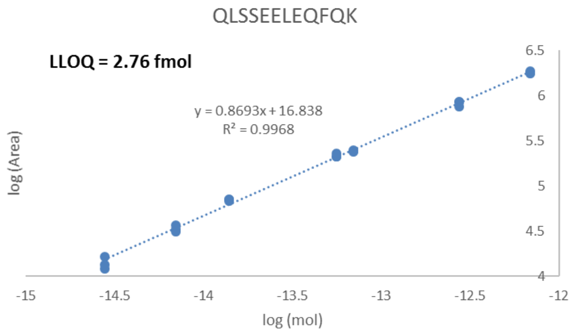
## Supplementary Figure S2

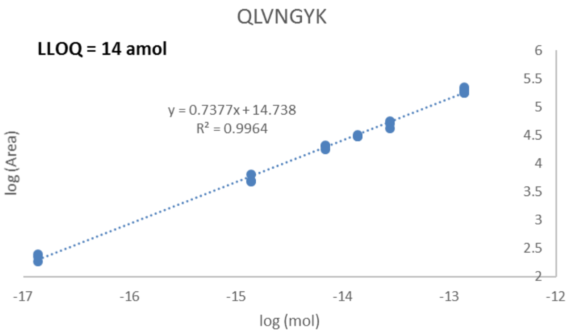


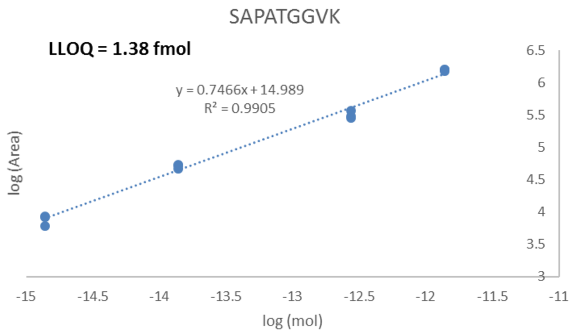

LTQQYNELLHSLQTK
LLOQ = 4.14 fmol
y = 0.7789x + 14.907
R² = 0.9908
log (Area)
log (mol)
7
6.5
6
5.5
5
4.5
4
3.5
3
-15
-14.5
-14
-13.5
-13
-12.5
-12
-11.5
-11

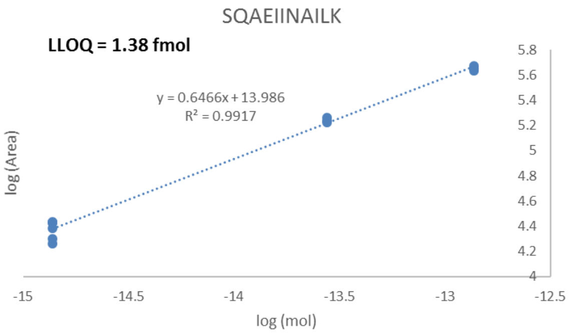

NTGIIC[+57]TIGPASR
LLOQ = 138 amol
y = 0.831x + 16.805
R² = 0.9922
log (Area)
log (mol)
8
7.5
7
6.5
6
5.5
5
4.5
4
3.5
3
-16
-15
-14
-13
-12
-11

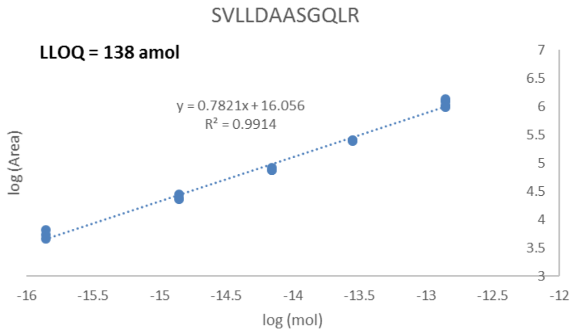

QLIVGVNK
LLOQ = 28 amol
y = 0.7885x + 15.641
R² = 0.9917
log (Area)
log (mol)
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
-17
-16
-15
-14
-13
-12

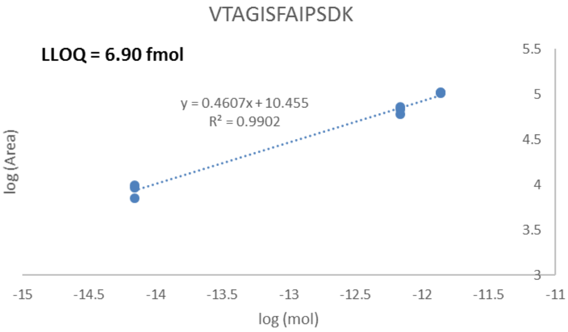

QLSSEELEQFQK
LLOQ = 2.76 fmol
y = 0.8693x + 16.838
R² = 0.9968
log (Area)
log (mol)
6.5
6
5.5
5
4.5
4
-15
-14.5
-14
-13.5
-13
-12.5
-12


QLVNGYK
LLOQ = 14 amol
y = 0.7377x + 14.738
R² = 0.9964
log (Area)
log (mol)
6
5.5
5
4.5
4
3.5
3
2.5
2
-17
-16
-15
-14
-13
-12


SAPATGGVK
LLOQ = 1.38 fmol
y = 0.7466x + 14.989
R² = 0.9905
log (Area)
log (mol)
6.5
6
5.5
5
4.5
4
3.5
3
-15
-14.5
-14
-13.5
-13
-12.5
-12
-11.5
-11

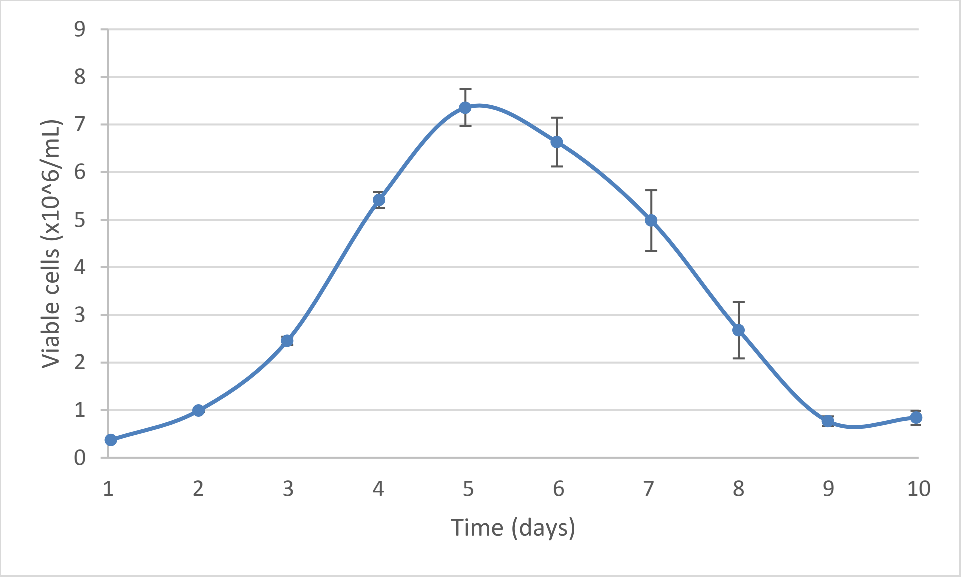

SQAEIINAILK
LLOQ = 1.38 fmol
y = 0.6466x + 13.986
R² = 0.9917
log (Area)
log (mol)
5.8
5.6
5.4
5.2
5
4.8
4.6
4.4
4.2
4
-15
-14.5
-14
-13.5
-13
-12.5


SVLLDAASGQLR
LLOQ = 138 amol
y = 0.7821x + 16.056
R² = 0.9914
log (Area)
log (mol)
7
6.5
6
5.5
5
4.5
4
3.5
3
-16
-15.5
-15
-14.5
-14
-13.5
-13
-12.5
-12

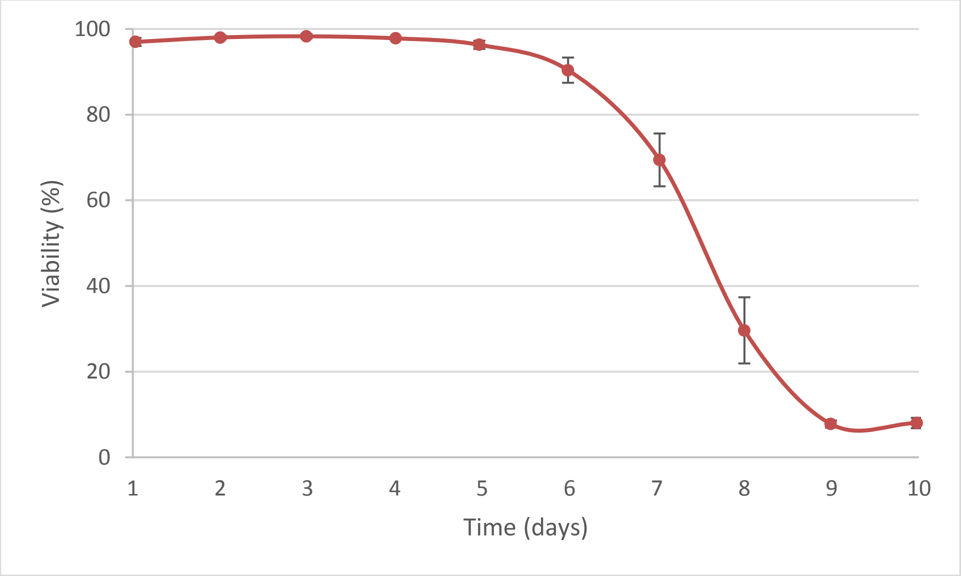

VTAGISFAIPSDK
LLOQ = 6.90 fmol
y = 0.4607x + 10.455
R² = 0.9902
log (Area)
log (mol)
5.5
5
4.5
4
3.5
3
-15
-14.5
-14
-13.5
-13
-12.5
-12
-11.5
-11

## Supplementary Figure S3

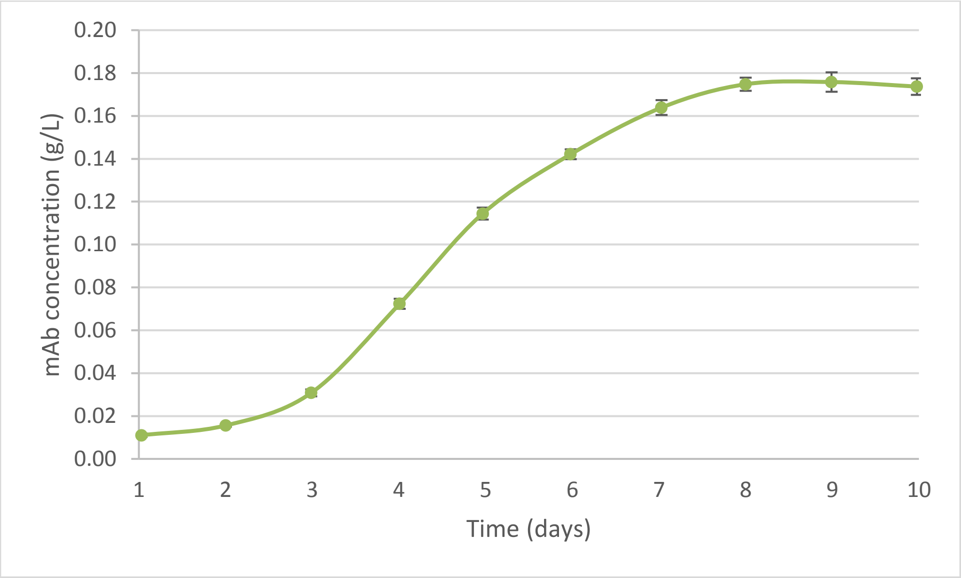
(a)



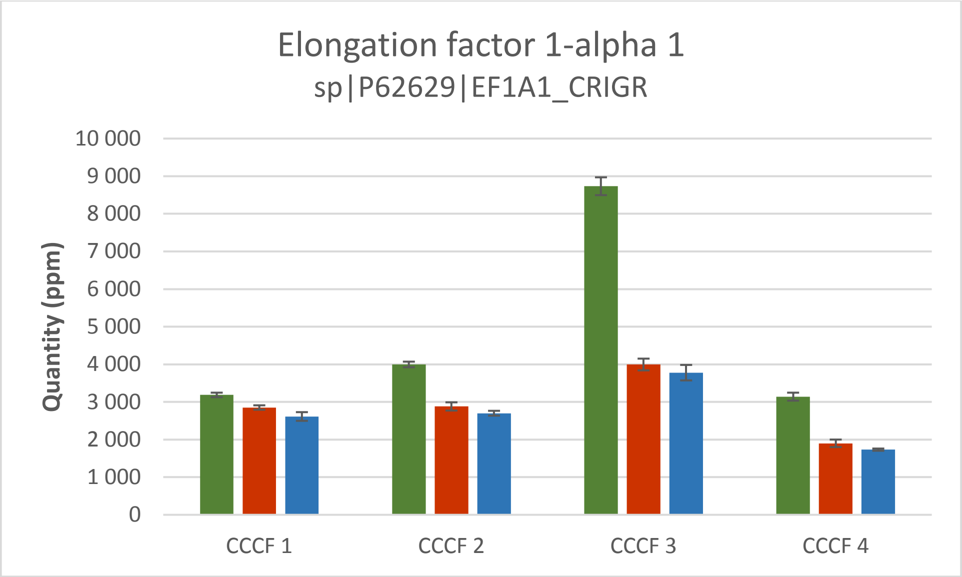
(b)

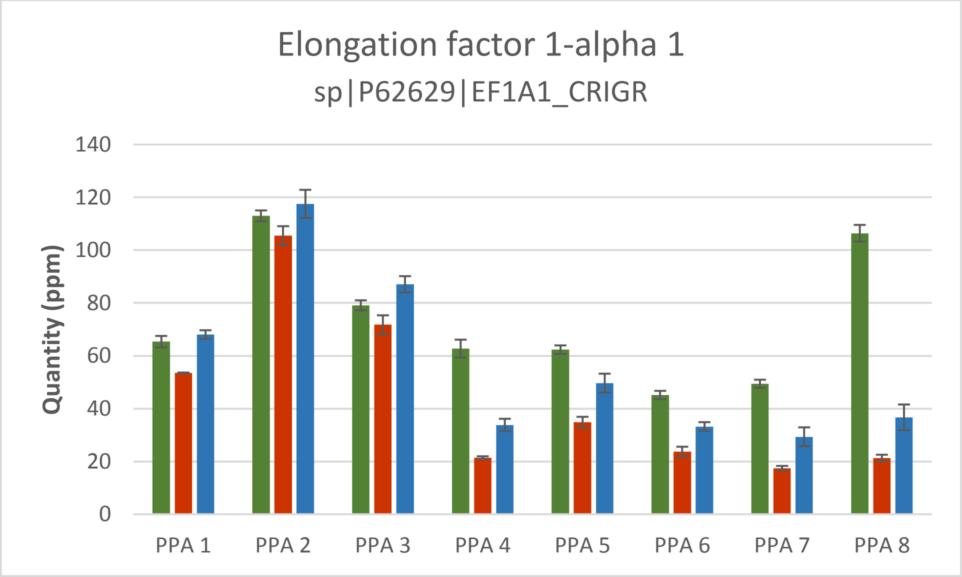


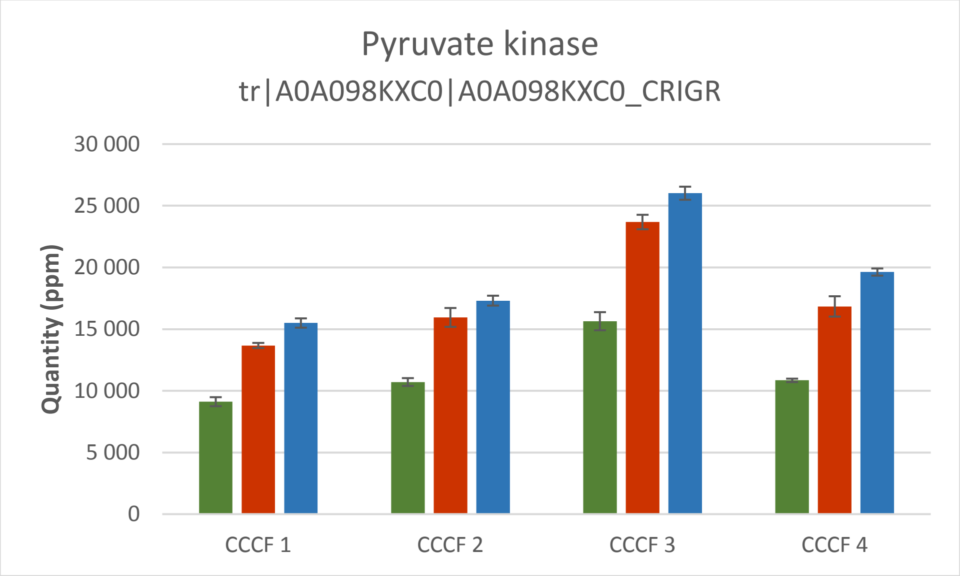
(c)

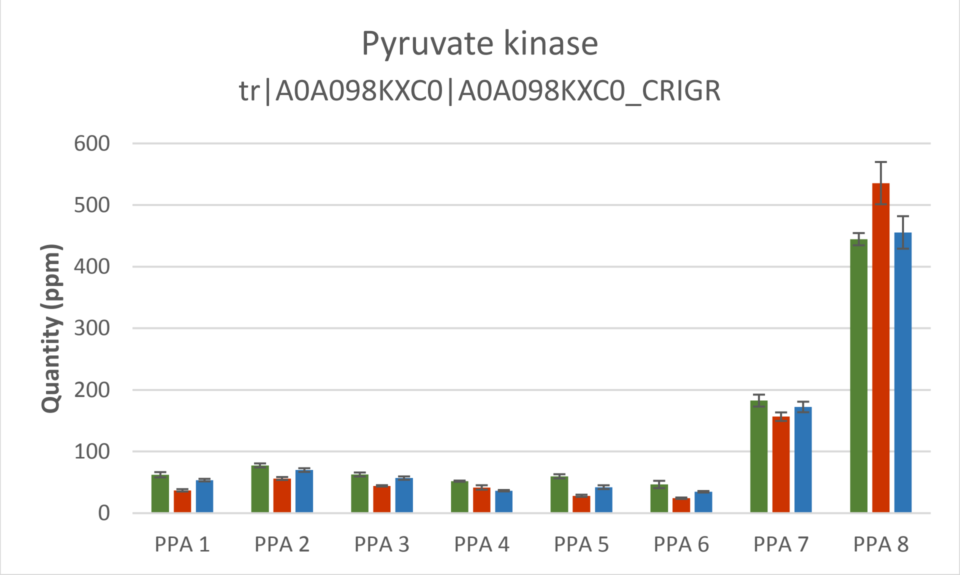

mAb concentration (g/L)
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
1
2
3
4
5
6
7
8
9
10
Time (days)

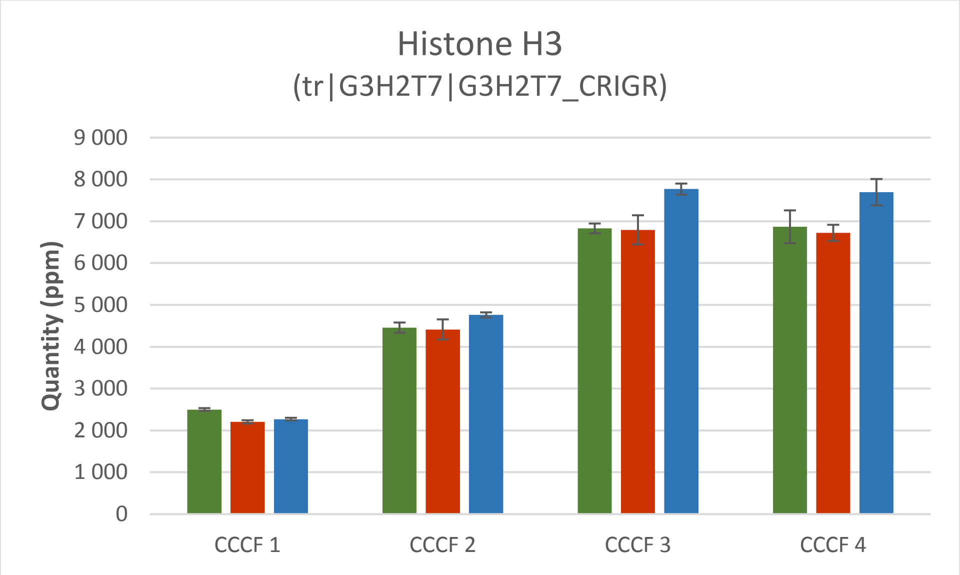
## Supplementary Figure S4

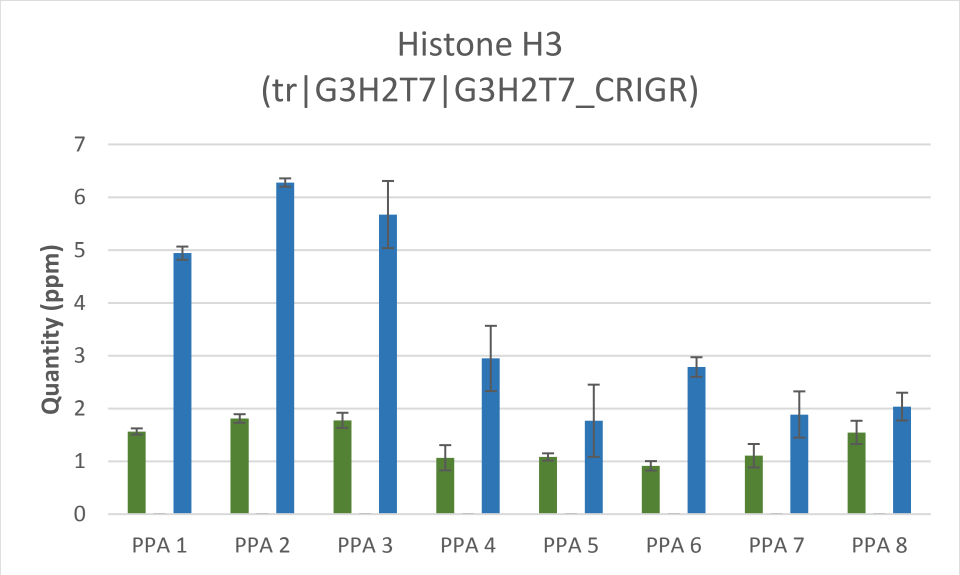


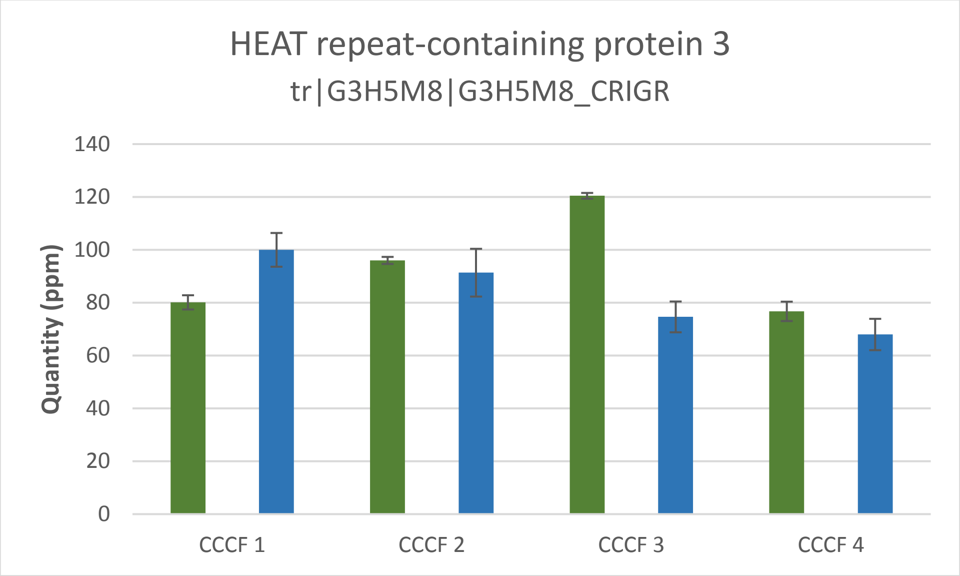


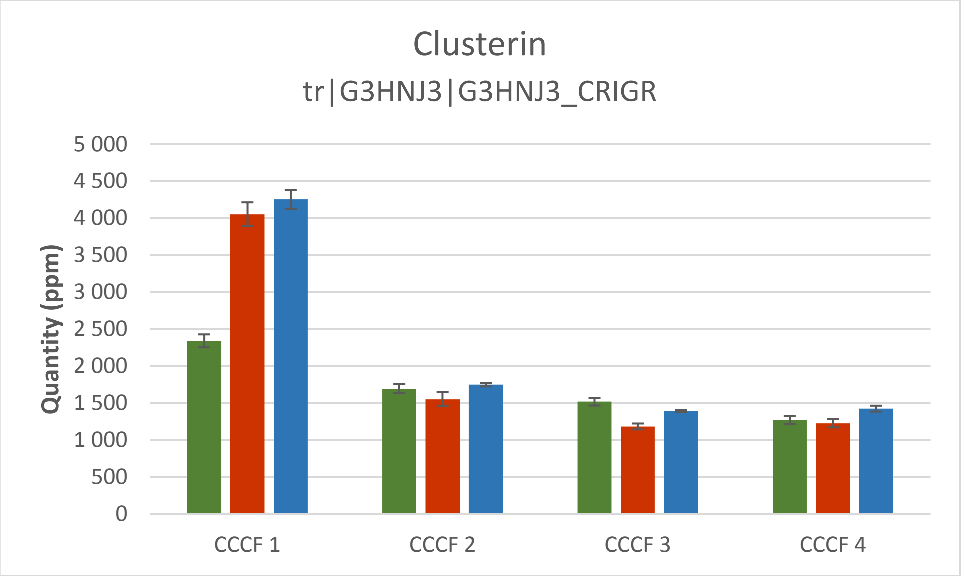

Pyruvate kinase
tr|A0A098KXC0|A0A098KXC0_CRIGR
Quantity (ppm)
30 000
25 000
20 000
15 000
10 000
5 000
0
CCCF 1
CCCF 2
CCCF 3
CCCF 4

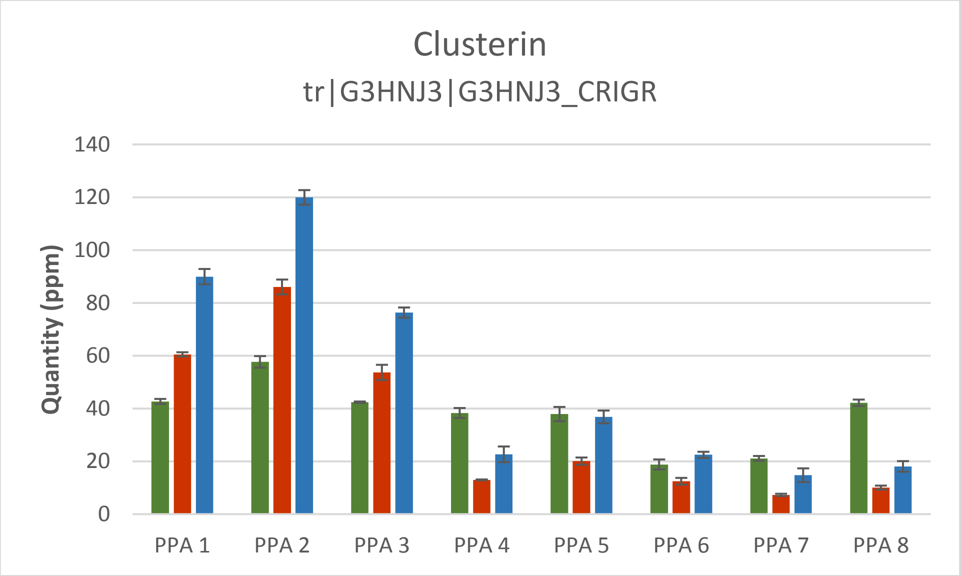

Pyruvate kinase
tr|A0A098KXC0|A0A098KXC0_CRIGR
Quantity (ppm)
600
500
400
300
200
100
0
PPA 1
PPA 2
PPA 3
PPA 4
PPA 5
PPA 6
PPA 7
PPA 8

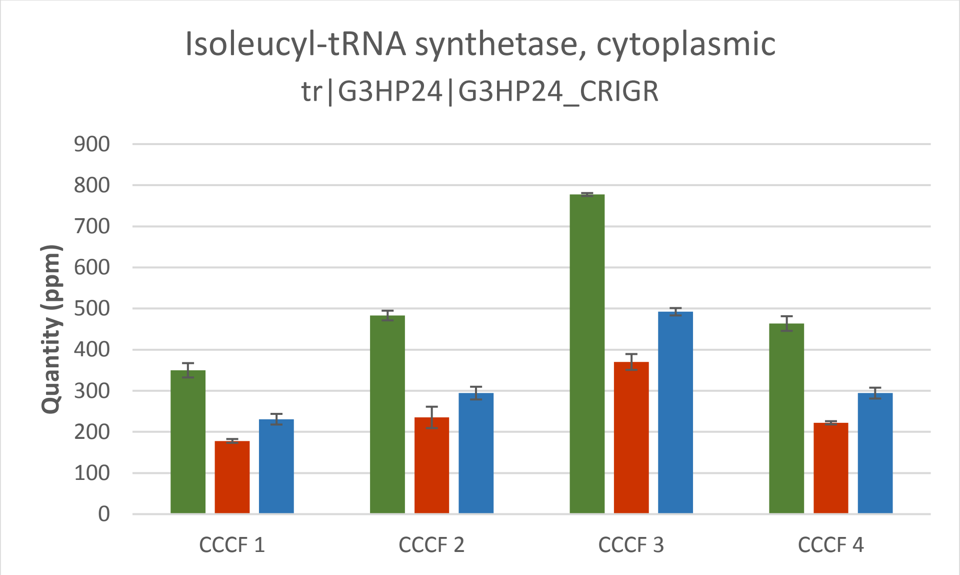

Histone H3
(tr|G3H2T7|G3H2T7_CRIGR)
Quantity (ppm)
9 000
8 000
7 000
6 000
5 000
4 000
3 000
2 000
1 000
0
CCCF 1
CCCF 2
CCCF 3
CCCF 4

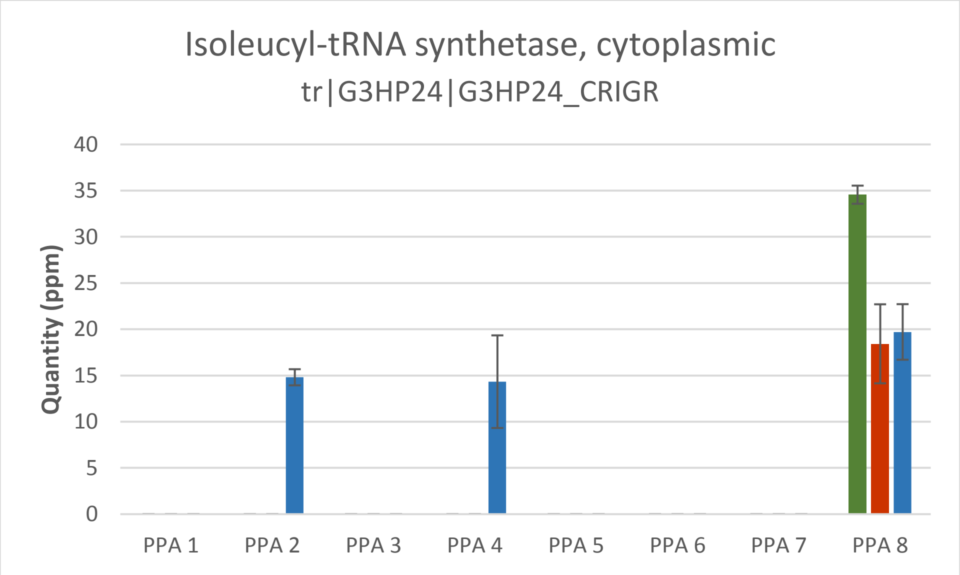

Histone H3
(tr|G3H2T7|G3H2T7_CRIGR)
Quantity (ppm)
7
6
5
4
3
2
1
0
PPA 1
PPA 2
PPA 3
PPA 4
PPA 5
PPA 6
PPA 7
PPA 8

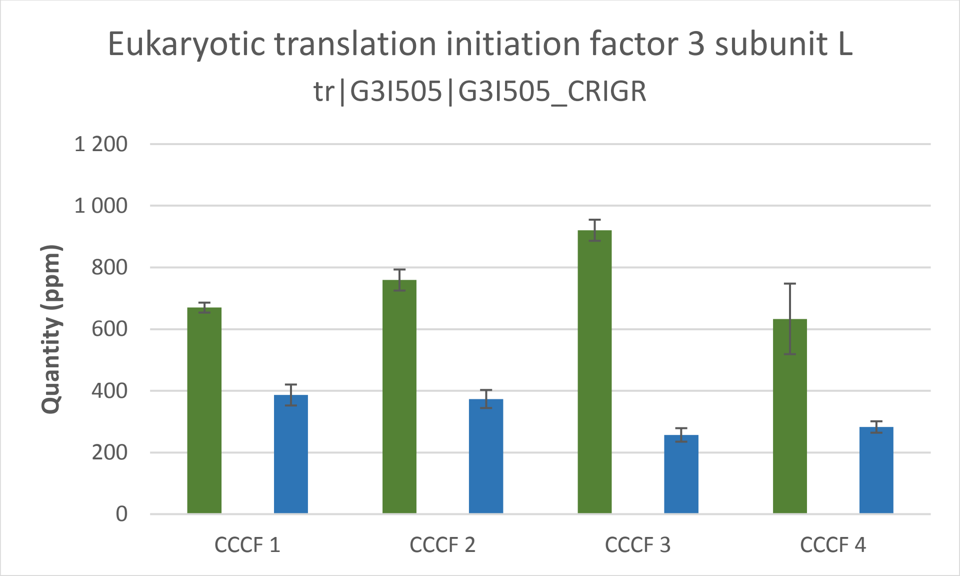

HEAT repeat-containing protein 3
tr|G3H5M8|G3H5M8_CRIGR
Quantity (ppm)
140
120
100
80
60
40
20
0
CCCF 1
CCCF 2
CCCF 3
CCCF 4

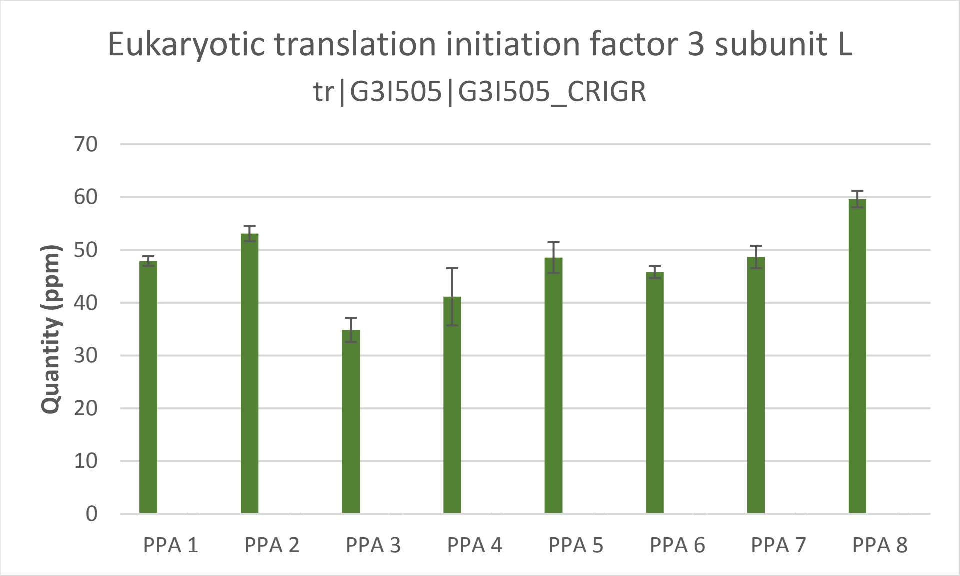

Clusterin
tr|G3HNJ3|G3HNJ3_CRIGR
Quantity (ppm)
5 000
4 500
4 000
3 500
3 000
2 500
2 000
1 500
1 000
500
0
CCCF 1
CCCF 2
CCCF 3
CCCF 4

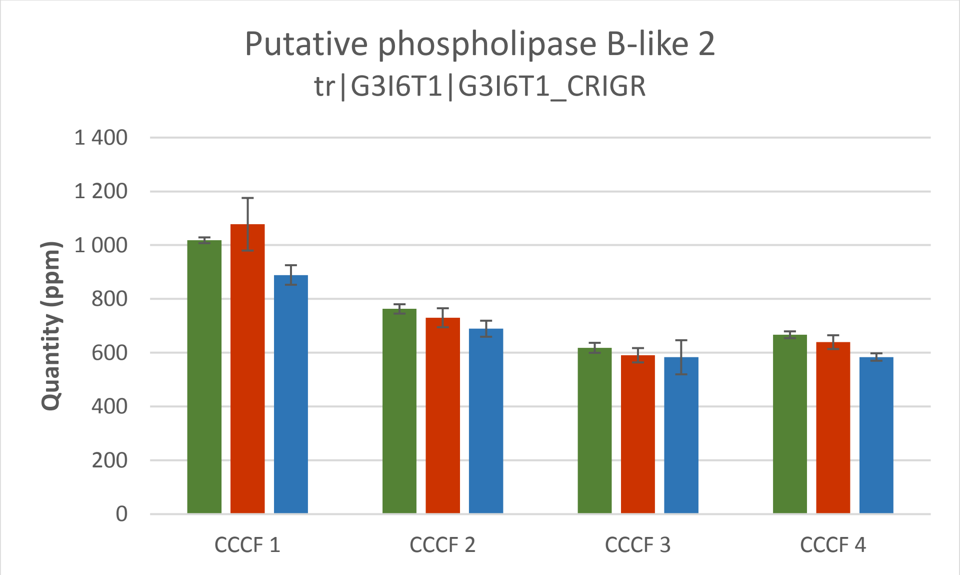

Clusterin
tr|G3HNJ3|G3HNJ3_CRIGR
Quantity (ppm)
140
120
100
80
60
40
20
0
PPA 1
PPA 2
PPA 3
PPA 4
PPA 5
PPA 6
PPA 7
PPA 8

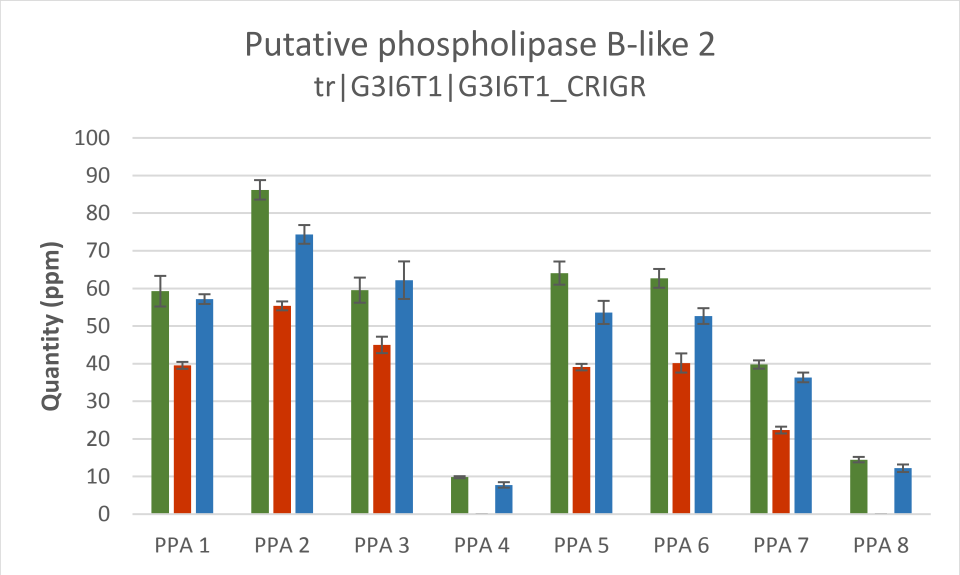

Isoleucyl-tRNA synthetase, cytoplasmic
tr|G3HP24|G3HP24_CRIGR
Quantity (ppm)
900
800
700
600
500
400
300
200
100
0
CCCF 1
CCCF 2
CCCF 3
CCCF 4

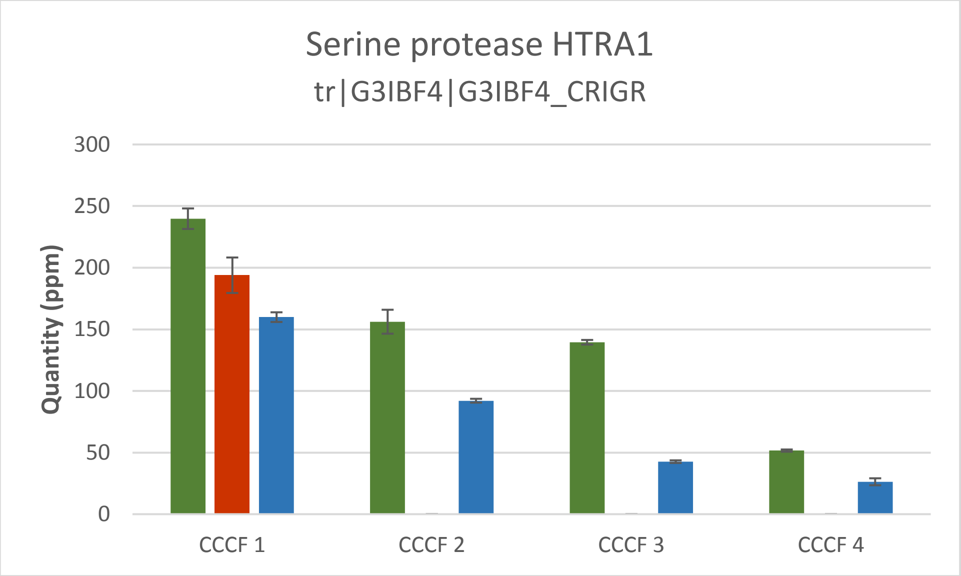

Isoleucyl-tRNA synthetase, cytoplasmic
tr|G3HP24|G3HP24_CRIGR
Quantity (ppm)
40
35
30
25
20
15
10
5
0
PPA 1
PPA 2
PPA 3
PPA 4
PPA 5
PPA 6
PPA 7
PPA 8

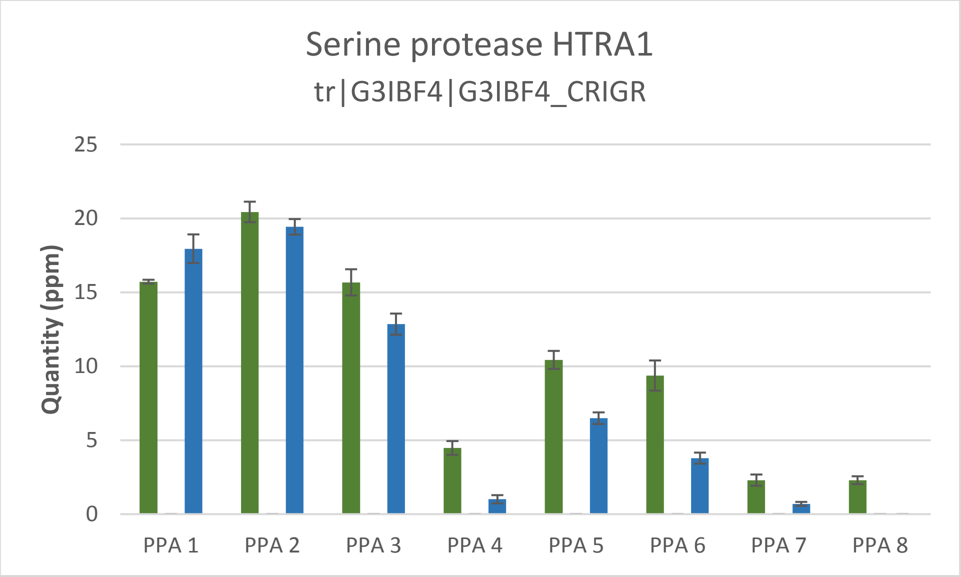

Eukaryotic translation initiation factor 3 subunit L
tr|G3I505|G3I505_CRIGR
Quantity (ppm)
1 200
1 000
800
600
400
200
0
CCCF 1
CCCF 2
CCCF 3
CCCF 4

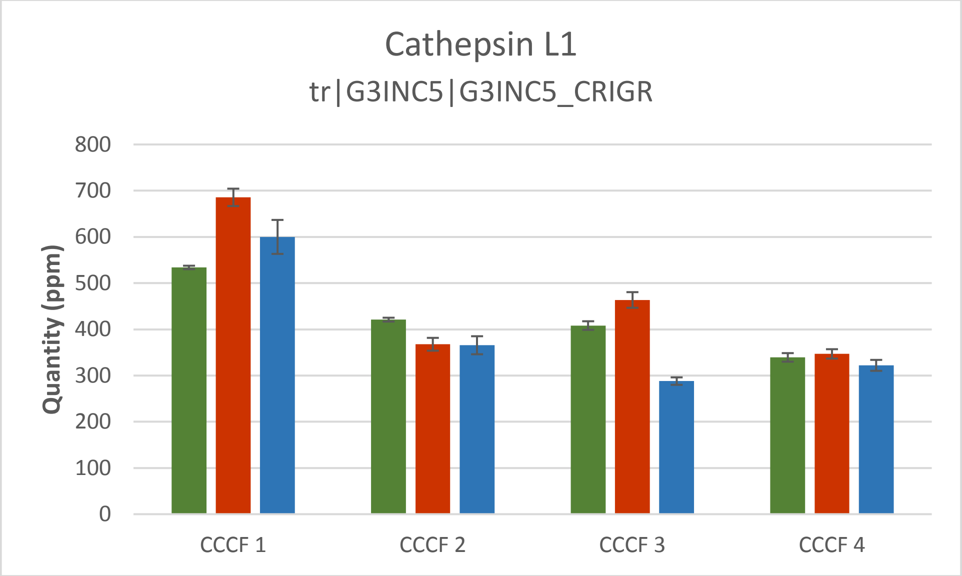

Eukaryotic translation initiation factor 3 subunit L
tr|G3I505|G3I505_CRIGR
Quantity (ppm)
70
60
50
40
30
20
10
0
PPA 1
PPA 2
PPA 3
PPA 4
PPA 5
PPA 6
PPA 7
PPA 8

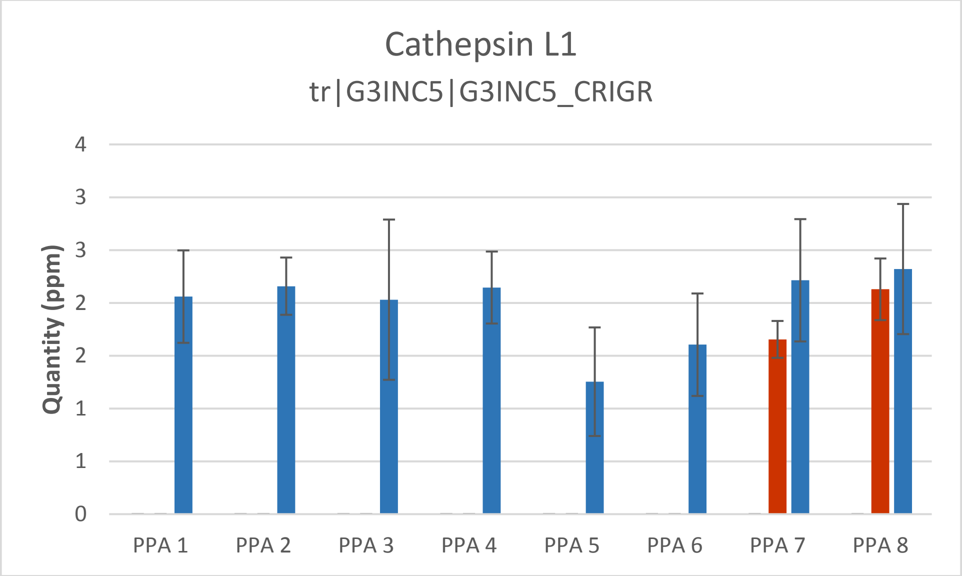

Putative phospholipase B-like 2
tr|G3I6T1|G3I6T1_CRIGR
Quantity (ppm)
1 400
1 200
1 000
800
600
400
200
0
CCCF 1
CCCF 2
CCCF 3
CCCF 4


Putative phospholipase B-like 2
tr|G3I6T1|G3I6T1_CRIGR
Quantity (ppm)
100
90
80
70
60
50
40
30
20
10
0
PPA 1
PPA 2
PPA 3
PPA 4
PPA 5
PPA 6
PPA 7
PPA 8


Serine protease HTRA1
tr|G3IBF4|G3IBF4_CRIGR
Quantity (ppm)
300
250
200
150
100
50
0
CCCF 1
CCCF 2
CCCF 3
CCCF 4

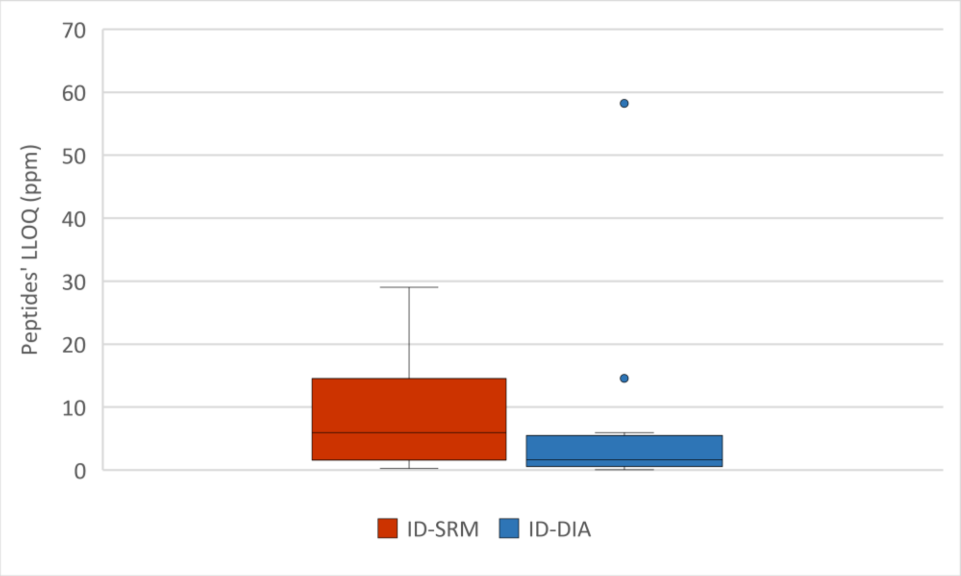

Serine protease HTRA1
tr|G3IBF4|G3IBF4_CRIGR
Quantity (ppm)
25
20
15
10
5
0
PPA 1
PPA 2
PPA 3
PPA 4
PPA 5
PPA 6
PPA 7
PPA 8


Cathepsin L1
tr|G3INC5|G3INC5_CRIGR
Quantity (ppm)
800
700
600
500
400
300
200
100
0
CCCF 1
CCCF 2
CCCF 3
CCCF 4

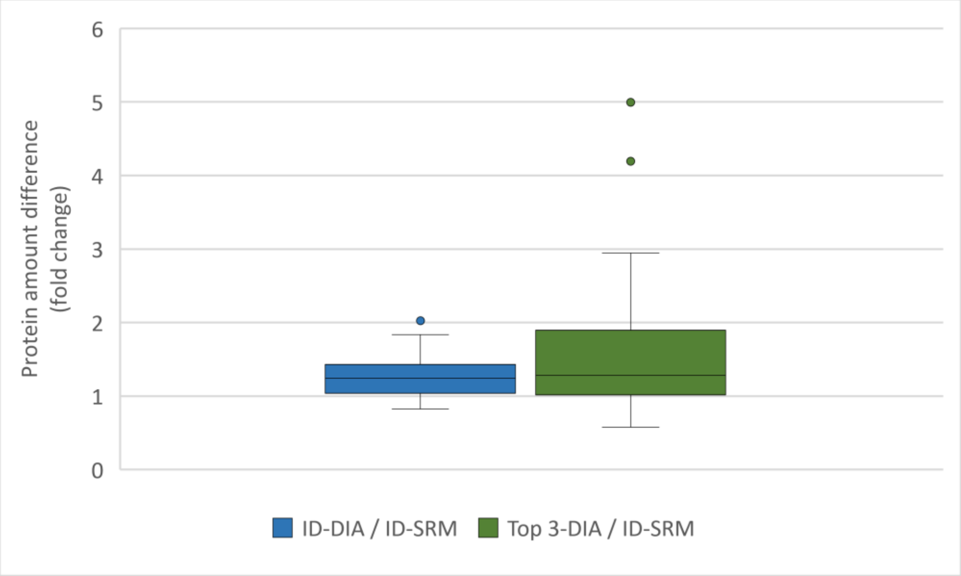

Cathepsin L1
tr|G3INC5|G3INC5_CRIGR
Quantity (ppm)
4
3
3
2
2
1
1
0
PPA 1
PPA 2
PPA 3
PPA 4
PPA 5
PPA 6
PPA 7
PPA 8

## Supplementary Figure S5

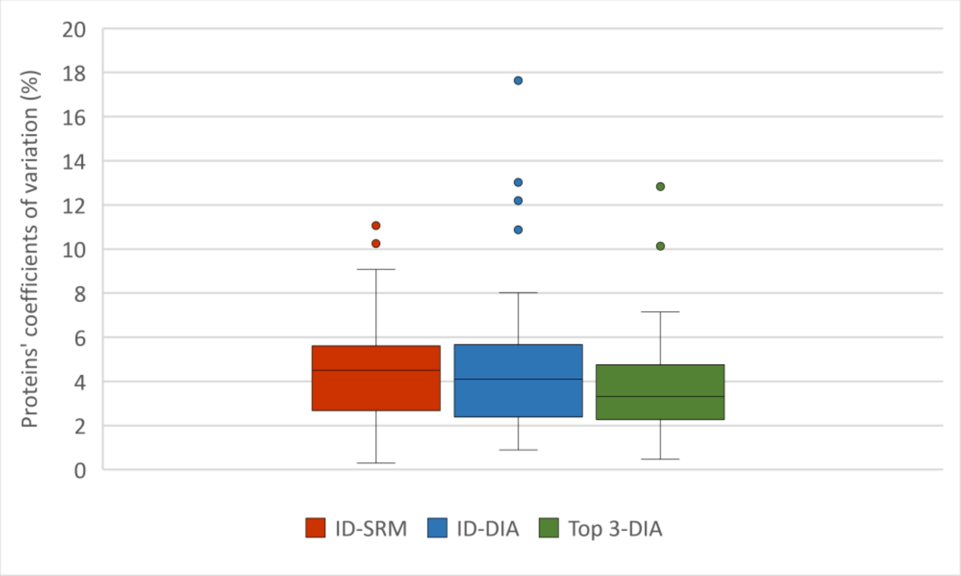
(a)



(b)

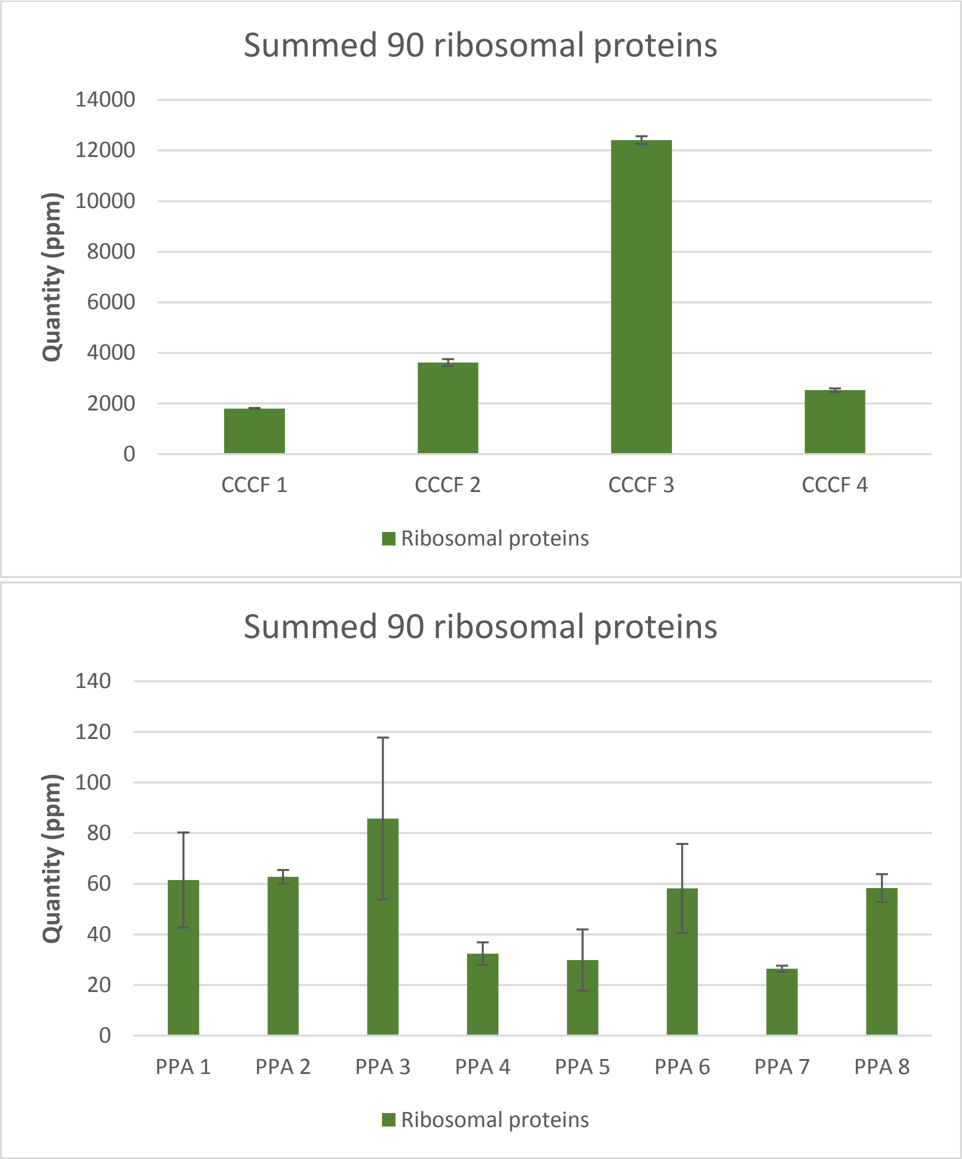


(c)

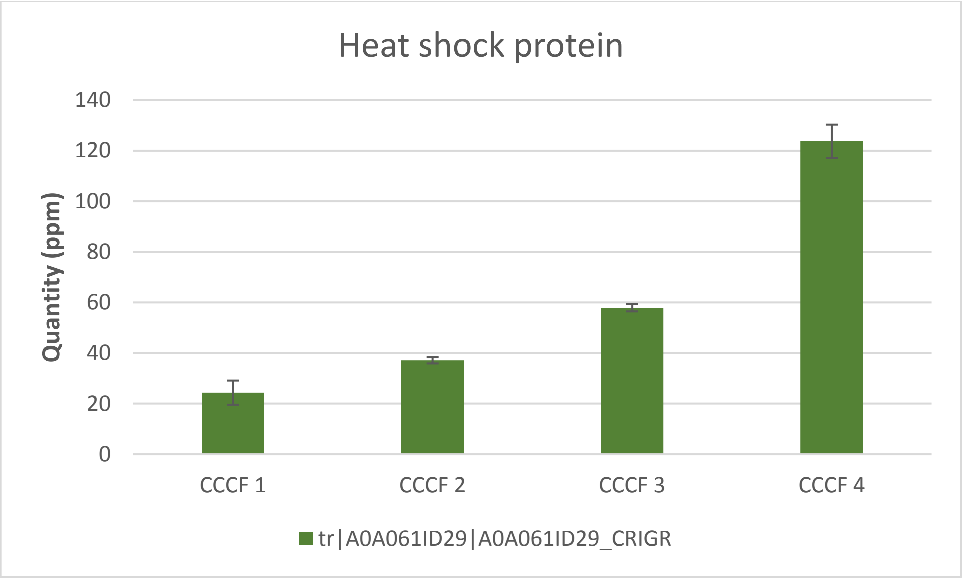

Proteins' coefficients of variation (%)
20
18
16
14
12
10
8
6
4
2
0
ID-SRM
ID-DIA
Top 3-DIA

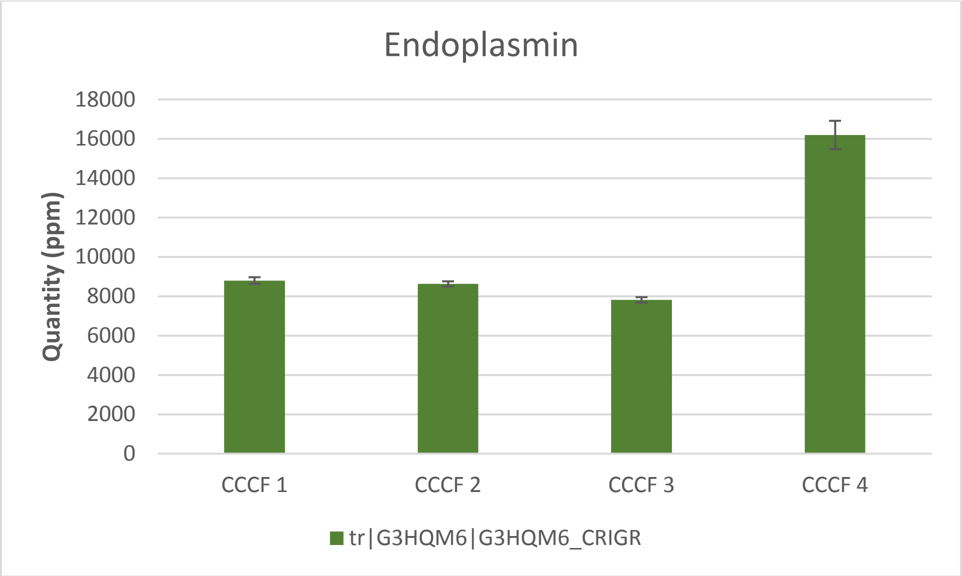
**Supplementary Figure S6**

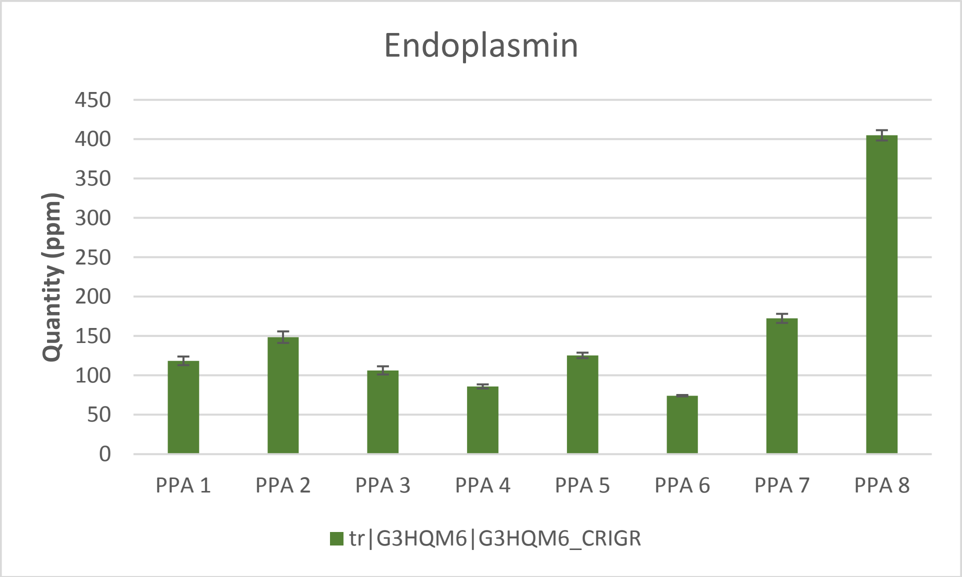
(a)

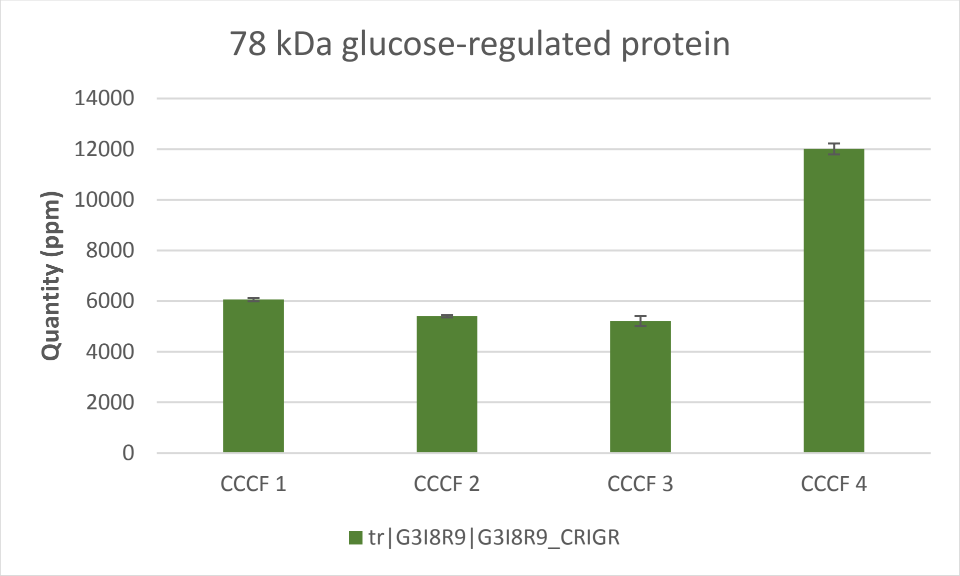


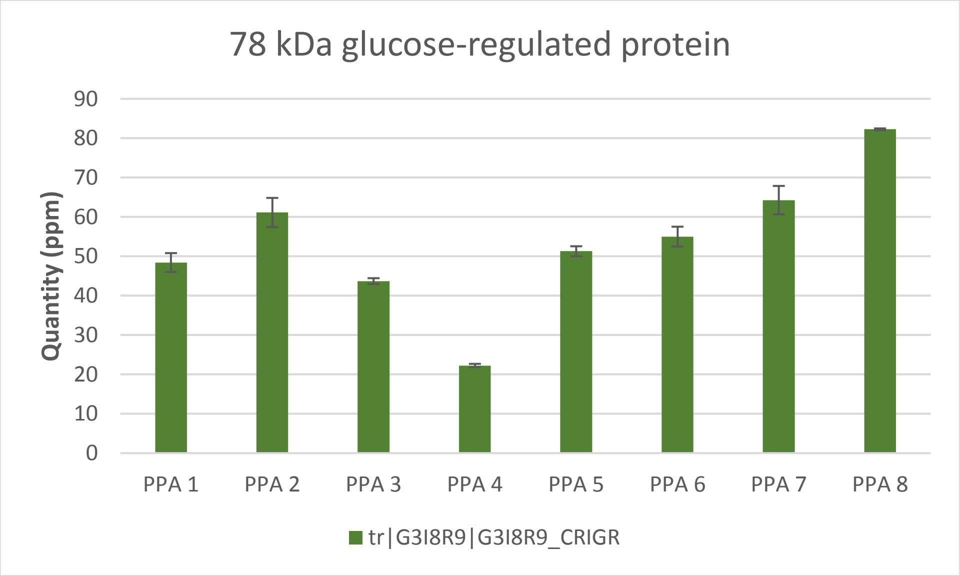
(b)

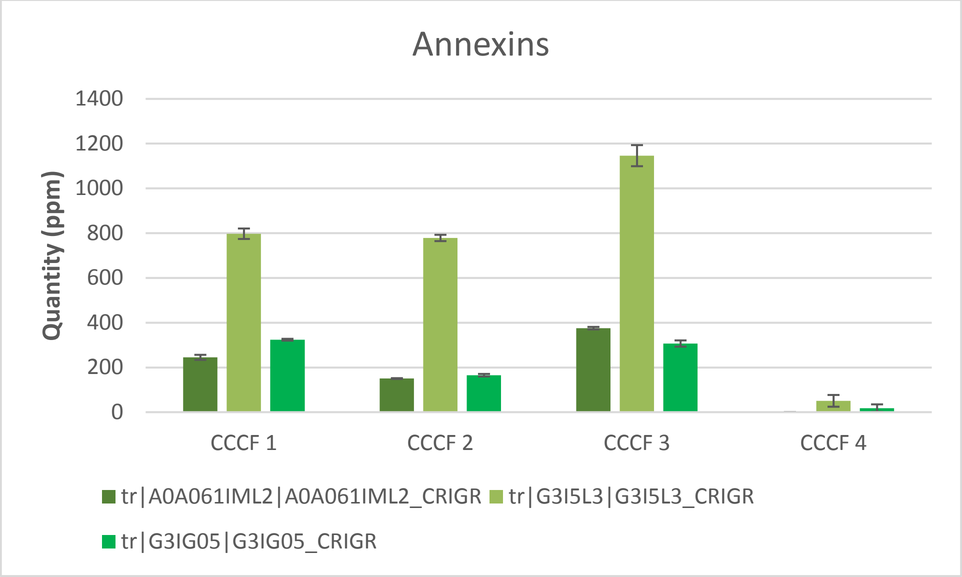

Heat shock protein
Quantity (ppm)
0
20
40
60
80
100
120
140
CCCF 1
CCCF 2
CCCF 3
CCCF 4
🟩 tr|A0A061ID29|A0A061ID29_CRIGR

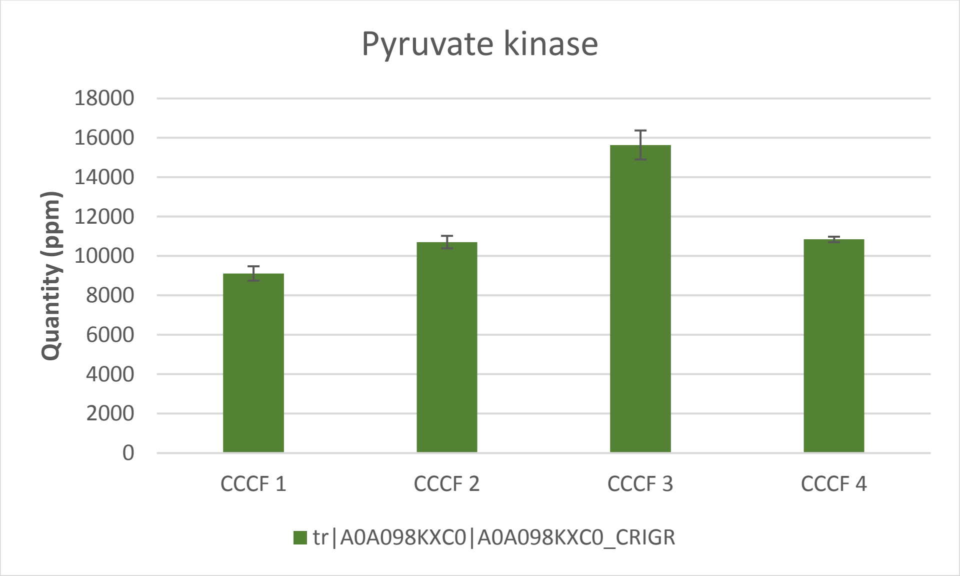

Endoplasmin
Quantity (ppm)
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
CCCF 1
CCCF 2
CCCF 3
CCCF 4
🟩 tr|G3HQM6|G3HQM6_CRIGR

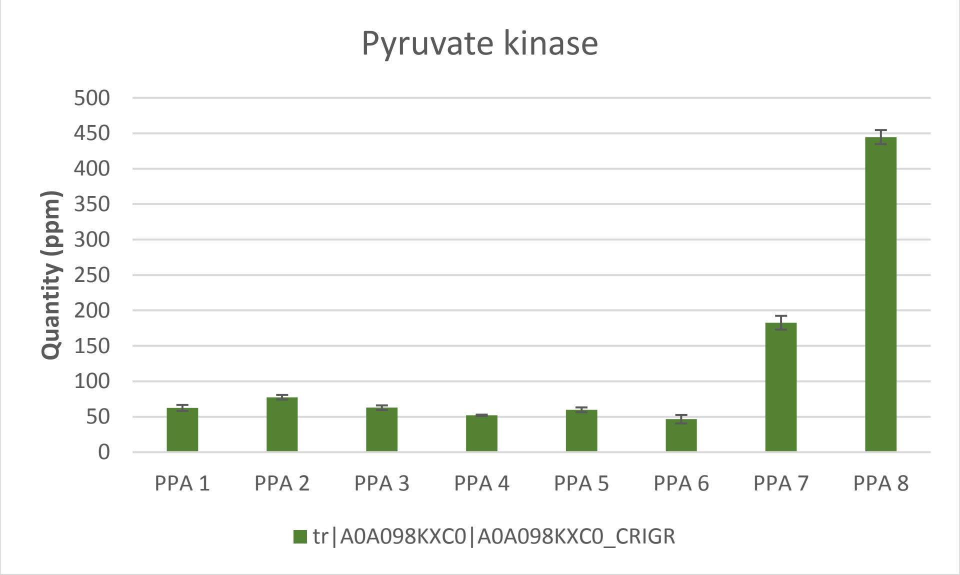

Endoplasmin
Quantity (ppm)
0
50
100
150
200
250
300
350
400
450
PPA 1
PPA 2
PPA 3
PPA 4
PPA 5
PPA 6
PPA 7
PPA 8
🟩 tr|G3HQM6|G3HQM6_CRIGR

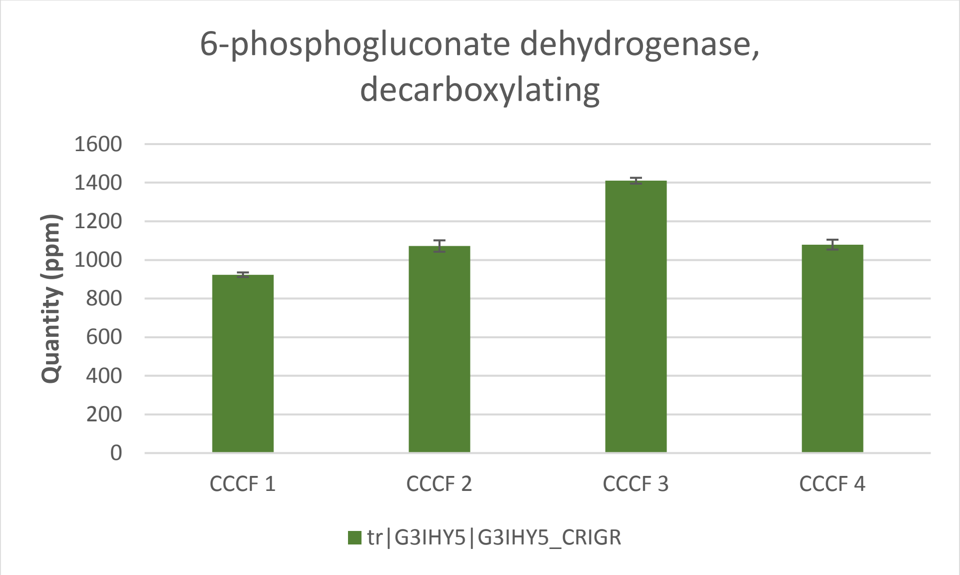

78 kDa glucose-regulated protein
Quantity (ppm)
14000
12000
10000
8000
6000
4000
2000
0
CCCF 1
CCCF 2
CCCF 3
CCCF 4
🟩 tr|G3I8R9|G3I8R9_CRIGR

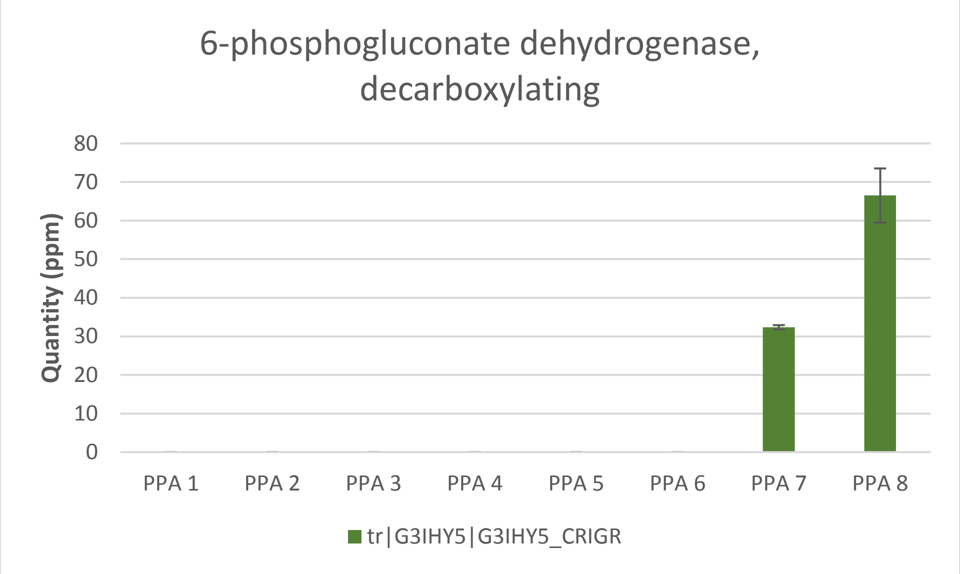

78 kDa glucose-regulated protein
Quantity (ppm)
90
80
70
60
50
40
30
20
10
0
PPA 1
PPA 2
PPA 3
PPA 4
PPA 5
PPA 6
PPA 7
PPA 8
🟩 tr|G3I8R9|G3I8R9_CRIGR

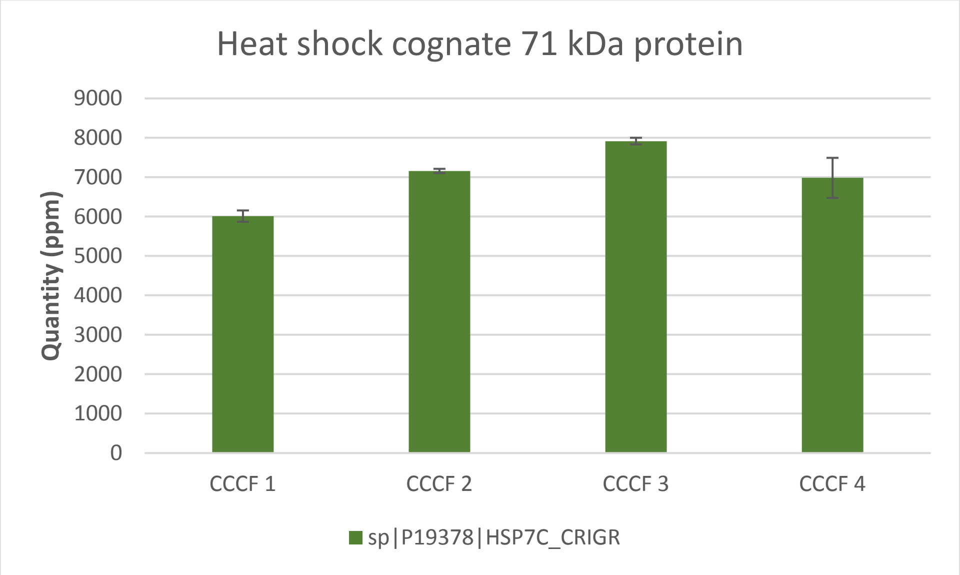

Annexins
Quantity (ppm)
1400
1200
1000
800
600
400
200
0
CCCF 1
CCCF 2
CCCF 3
CCCF 4
tr|A0A061IML2|A0A061IML2_CRIGR
tr|G3I5L3|G3I5L3_CRIGR
tr|G3IG05|G3IG05_CRIGR

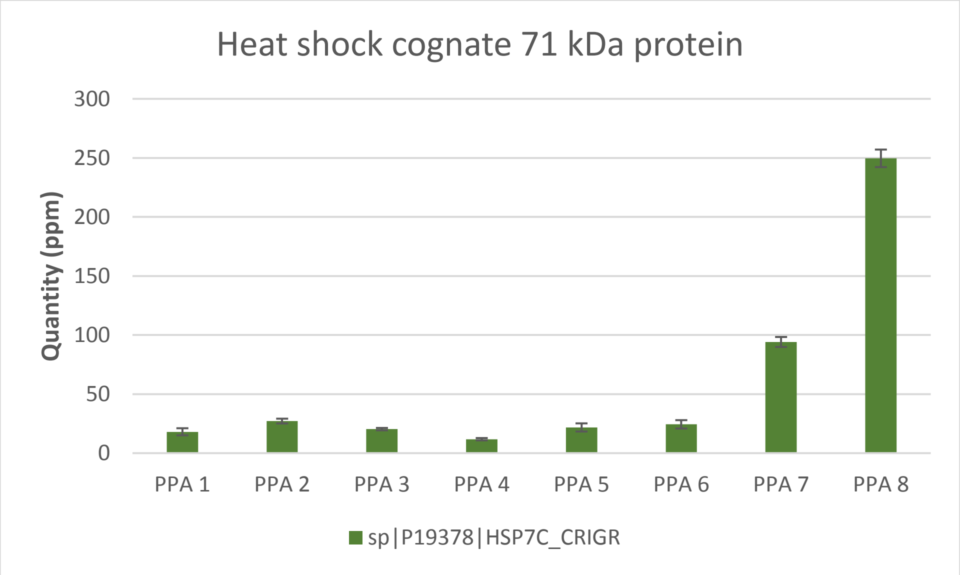

Pyruvate kinase
Quantity (ppm)
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
CCCF 1
CCCF 2
CCCF 3
CCCF 4
tr|A0A098KXC0|A0A098KXC0_CRIGR


Pyruvate kinase
Quantity (ppm)
500
450
400
350
300
250
200
150
100
50
0
PPA 1
PPA 2
PPA 3
PPA 4
PPA 5
PPA 6
PPA 7
PPA 8
tr|A0A098KXC0|A0A098KXC0_CRIGR

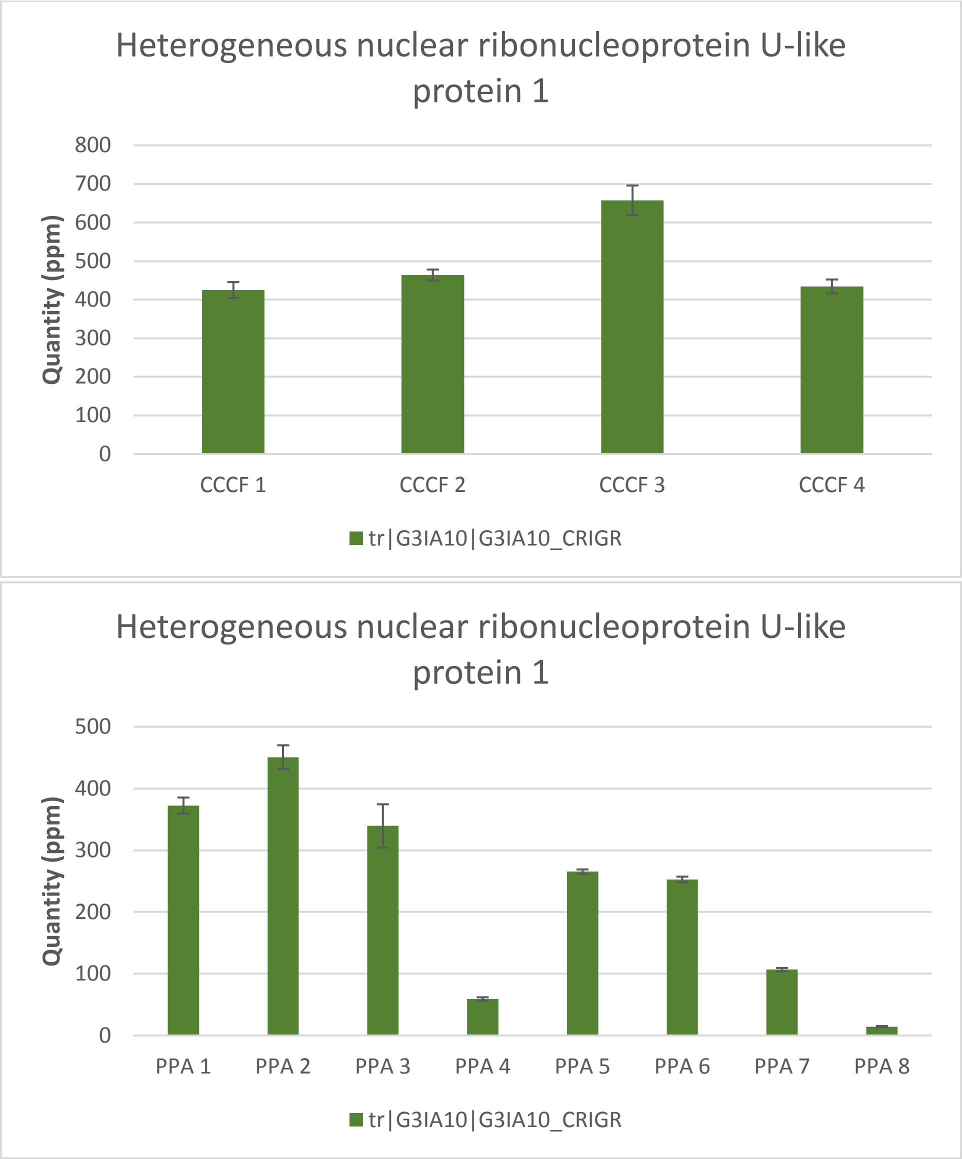

6-phosphogluconate dehydrogenase, decarboxylating
Quantity (ppm)
1600
1400
1200
1000
800
600
400
200
0
CCCF 1
CCCF 2
CCCF 3
CCCF 4
tr|G3IHY5|G3IHY5_CRIGR

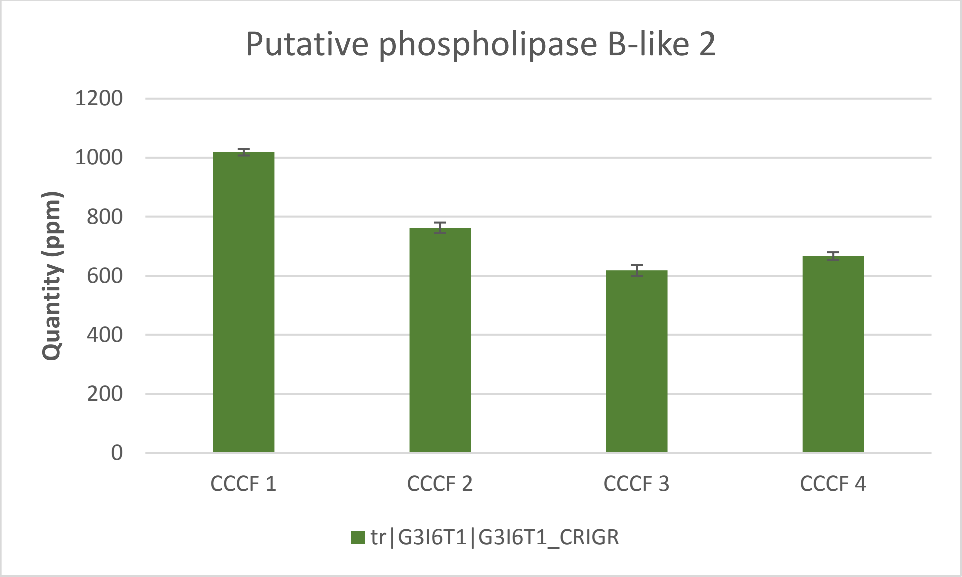

6-phosphogluconate dehydrogenase, decarboxylating
Quantity (ppm)
80
70
60
50
40
30
20
10
0
PPA 1
PPA 2
PPA 3
PPA 4
PPA 5
PPA 6
PPA 7
PPA 8
tr|G3IHY5|G3IHY5_CRIGR

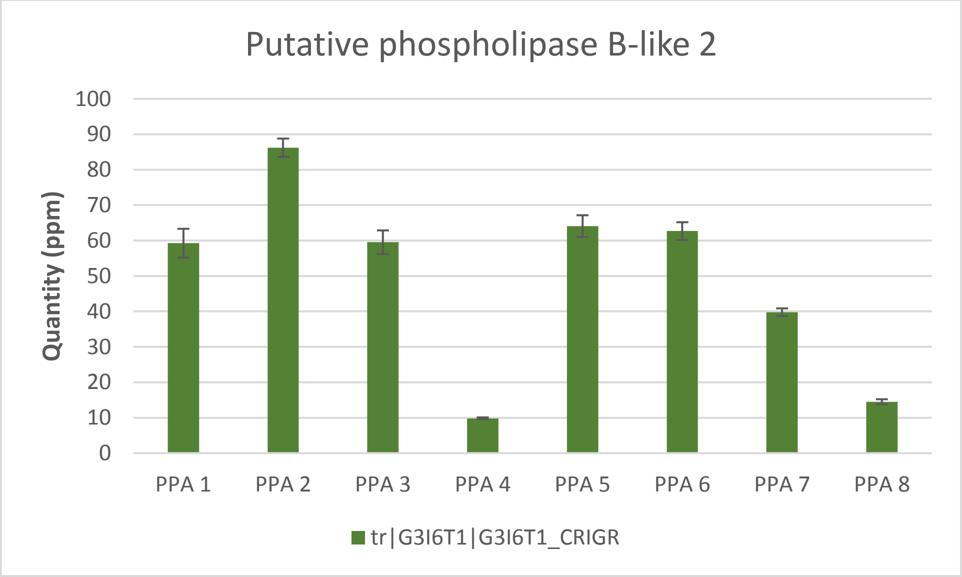

Heat shock cognate 71 kDa protein
Quantity (ppm)
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
CCCF 1
CCCF 2
CCCF 3
CCCF 4
sp|P19378|HSP7C_CRIGR


Heat shock cognate 71 kDa protein
Quantity (ppm)
300
250
200
150
100
50
0
PPA 1
PPA 2
PPA 3
PPA 4
PPA 5
PPA 6
PPA 7
PPA 8
sp|P19378|HSP7C_CRIGR

(c)


Heterogeneous nuclear ribonucleoprotein U-like protein 1
Quantity (ppm)
800
700
600
500
400
300
200
100
0
CCCF 1
CCCF 2
CCCF 3
CCCF 4
🟩 tr|G3IA10|G3IA10_CRIGR
Heterogeneous nuclear ribonucleoprotein U-like protein 1
Quantity (ppm)
500
400
300
200
100
0
PPA 1
PPA 2
PPA 3
PPA 4
PPA 5
PPA 6
PPA 7
PPA 8
🟩 tr|G3IA10|G3IA10_CRIGR


Putative phospholipase B-like 2
Quantity (ppm)
1200
1000
800
600
400
200
0
CCCF 1
CCCF 2
CCCF 3
CCCF 4
tr|G3I6T1|G3I6T1_CRIGR


Putative phospholipase B-like 2
Quantity (ppm)
100
90
80
70
60
50
40
30
20
10
0
PPA 1
PPA 2
PPA 3
PPA 4
PPA 5
PPA 6
PPA 7
PPA 8
tr|G3I6T1|G3I6T1_CRIGR

## References

1. Boychyn, M.; Yim, S. S.; Bulmer, M.; More, J.; Bracewell, D. G.; Hoare, M., Performance prediction of industrial centrifuges using scale-down models. *Bioprocess Biosyst Eng* **2004,** *26* (6), 385-91.
2. Tavakoli-Keshe, R.; Phillips, J. J.; Turner, R.; Bracewell, D. G., Understanding the relationship between biotherapeutic protein stability and solid-liquid interfacial shear in constant region mutants of IgG1 and IgG4. *J Pharm Sci* **2014,** *103* (2), 437-44.
3. Lau, E. C.; Kong, S.; McNulty, S.; Entwisle, C.; McIlgorm, A.; Dalton, K. A.; Hoare, M., An ultra scale-down characterization of low shear stress primary recovery stages to enhance selectivity of fusion protein recovery from its molecular variants. *Biotechnol Bioeng* **2013,** *110* (7), 1973-83.
4. Liu, H. F.; Ma, J.; Winter, C.; Bayer, R., Recovery and purification process development for monoclonal antibody production. *MAbs* **2010,** *2* (5), 480-99.
5. Shukla, A. A.; Hinckley, P., Host cell protein clearance during protein A chromatography: development of an improved column wash step. *Biotechnol Prog* **2008,** *24* (5), 1115-21.
6. 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), 1014-9.
7. Carapito, C.; Lane, L.; Benama, M.; Opsomer, A.; Mouton-Barbosa, E.; Garrigues, L.; Gonzalez de Peredo, A.; Burel, A.; Bruley, C.; Gateau, A.; Bouyssie, D.; Jaquinod, M.; Cianferani, S.; Burlet-Schiltz, O.; Van Dorsselaer, A.; Garin, J.; Vandenbrouck, Y., Computational and Mass-Spectrometry-Based Workflow for the Discovery and Validation of Missing Human Proteins: Application to Chromosomes 2 and 14. *J Proteome Res* **2015,** *14* (9), 3621-34.
8. MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J., Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010,** *26* (7), 966-8.
9. Navarro, P.; Kuharev, J.; Gillet, L. C.; Bernhardt, O. M.; MacLean, B.; Rost, H. L.; Tate, S. A.; Tsou, C. C.; Reiter, L.; Distler, U.; Rosenberger, G.; Perez-Riverol, Y.; Nesvizhskii, A. I.; Aebersold, R.; Tenzer, S., A multicenter study benchmarks software tools for label-free proteome quantification. *Nat Biotechnol* **2016**.
10. Silva, J. C.; Gorenstein, M. V.; Li, G. Z.; Vissers, J. P.; Geromanos, S. J., Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics* **2006,** *5* (1), 144-56.
11. Doneanu, C. E.; Xenopoulos, A.; Fadgen, K.; Murphy, J.; Skilton, S. J.; Prentice, H.; Stapels, M.; Chen, W., Analysis of host-cell proteins in biotherapeutic proteins by comprehensive online two-dimensional liquid chromatography/mass spectrometry. *mAbs* **2012,** *4* (1), 24-44.
12. Levy, N. E.; Valente, K. N.; Choe, L. H.; Lee, K. H.; Lenhoff, A. M., Identification and characterization of host cell protein product-associated impurities in monoclonal antibody bioprocessing. *Biotechnol Bioeng* **2014,** *111* (5), 904-12.
13. Pezzini, J.; Joucla, G.; Gantier, R.; Toueille, M.; Lomenech, A. M.; Le Senechal, C.; Garbay, B.; Santarelli, X.; Cabanne, C., Antibody capture by mixed-mode chromatography: a comprehensive study from determination of optimal purification conditions to identification of contaminating host cell proteins. *J Chromatogr A* **2011,** *1218* (45), 8197-208.
14. Joucla, G.; Le Senechal, C.; Begorre, M.; Garbay, B.; Santarelli, X.; Cabanne, C., Cation exchange versus multimodal cation exchange resins for antibody capture from CHO supernatants: identification of contaminating host cell proteins by mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci* **2013,** *942-943*, 126-33.
15. Aboulaich, N.; Chung, W. K.; Thompson, J. H.; Larkin, C.; Robbins, D.; Zhu, M., A novel approach to monitor clearance of host cell proteins associated with monoclonal antibodies. *Biotechnol Prog* **2014,** *30* (5), 1114-24.
16. Gagnon, P.; Nian, R.; Lee, J.; Tan, L.; Latiff, S. M.; Lim, C. L.; Chuah, C.; Bi, X.; Yang, Y.; Zhang, W.; Gan, H. T., Nonspecific interactions of chromatin with immunoglobulin G and protein A, and their impact on purification performance. *J Chromatogr A* **2014,** *1340*, 68-78.

17. Tran, B.; Grosskopf, V.; Wang, X.; Yang, J.; Walker, D., Jr.; Yu, C.; McDonald, P., Investigating interactions between phospholipase B-Like 2 and antibodies during Protein A chromatography. *J Chromatogr A* **2016,** *1438*, 31-8.

18. Gao, S. X.; Zhang, Y.; Stansberry-Perkins, K.; Buko, A.; Bai, S.; Nguyen, V.; Brader, M. L., Fragmentation of a highly purified monoclonal antibody attributed to residual CHO cell protease activity. *Biotechnol Bioeng* **2011,** *108* (4), 977-82.

19. Bee, J. S.; Tie, L.; Afdahl, C. D.; Jusino, K. C.; Johnson, D.; Dimitrova, M. N., Trace levels of the CHO host cell protease cathepsin D caused particle formation in a monoclonal antibody product. *Biotechnol Prog* **2015**.

20. Fischer, S. K.; Cheu, M.; Peng, K.; Lowe, J.; Araujo, J.; Murray, E.; McClintock, D.; Matthews, J.; Siguenza, P.; Song, A., Specific Immune Response to Phospholipase B-Like 2 Protein, a Host Cell Impurity in Lebrikizumab Clinical Material. *The AAPS journal* **2017,** *19* (1), 254-263.

21. Hanania, N. A.; Noonan, M.; Corren, J.; Korenblat, P.; Zheng, Y.; Fischer, S. K.; Cheu, M.; Putnam, W. S.; Murray, E.; Scheerens, H.; Holweg, C. T.; Maciuca, R.; Gray, S.; Doyle, R.; McClintock, D.; Olsson, J.; Matthews, J. G.; Yen, K., Lebrikizumab in moderate-to-severe asthma: pooled data from two randomised placebo-controlled studies. *Thorax* **2015,** *70* (8), 748-56.

# General conclusion

The objective of my PhD work was to improve proteome characterisation by quantitative mass spectrometry, and to develop mass spectrometry-based approaches to monitor host cell protein (HCP) impurities in monoclonal antibody (mAb) samples.

The first part of this manuscript is a bibliographic introduction, presenting (i) the state of the art of bottom-up proteomics, with a thorough description of the whole workflow, including the sample preparation step, the data acquisition by mass spectrometry, and data analysis. The different strategies are exposed, including data dependent acquisition (DDA) for shotgun proteomics, selected reaction monitoring (SRM), parallel reaction monitoring (PRM) or multiple reaction monitoring in high resolution (MRM HR) for targeted proteomics, and data independent acquisition (DIA) which promises to combine the advantages of both shotgun and targeted approaches. The bibliographic introduction also presents (ii) the field of mAbs, with a brief description of the mAb manufacturing process, and the state of the art of HCP monitoring.

In this context, the objectives of my PhD work presented in the second part of the manuscript were (i) to improve proteome characterisation by shotgun proteomics in DDA mode, (ii) to benchmark several targeted proteomics possibilities that are available in the laboratory, (iii) optimise the whole DIA workflow, including sample preparation, data acquisition and data analysis, and (iv) develop cutting edge mass spectrometry approaches to monitor the HCP impurities present in mAb samples.

Shotgun proteomics was improved by optimising a DDA method for our newly acquired microLC-Triple TOF 6600 coupling. The LC peptides separation, the source parameters and the mass spectrometry part were optimised with an emphasis on the mass spectrometry parameters. This allowed to provide the laboratory an optimised DDA method for both peptides and proteins identification and quantification using the XIC MS1 strategy. Several key parameters were highlighted to optimise a DDA method, among which the LC gradient was the most important, but also the desolvation gas and the use of dynamic accumulation.

Targeted proteomics aims to monitor ≈ 50-100 proteins of interest in large cohorts of samples with high sensitivity, specificity and reproducibility. The gold standard approach for targeted proteomics is SRM performed on a triple quadrupole instrument, but recently several approaches like PRM performed on quadrupole-orbitrap or MRM HR performed on quadrupole-time-of-flight instruments emerged, providing a higher specificity and an easier method development. Moreover, microLC is usually preferred to nanoLC because of its better robustness, though the necessary sample amount is higher and the sensitivity is reduced. In order to help in the decision making when a targeted approach is envisaged, I compared four targeted MS configurations, including nanoLC systems versus microLC systems, and the gold standard SRM method versus PRM and MRM HR. Globally, the four evaluated configurations performed equivalently in terms of sensitivity, accuracy and precision. The decision making should therefore be based (i) on the sample amount available, preferring the use of microLC systems over nanoLC systems because of their better robustness, (ii) the instrument dedication, because last generation HR/AM are capable of doing shotgun or DIA analyses, while low resolution triple quadrupole are usually dedicated to SRM, and (iii) if discovery and validation steps are performed on the same instrument, for instance if shotgun proteomics is performed on a HR/AM instrument, targeted proteomics can be performed on the same instrument, rendering the method development easier.

DIA approaches promise to combine the advantages of both shotgun and targeted proteomics approaches, allowing the quantification of all detected proteins with high sensitivity, specificity and reproducibility. However, DIA data analysis is today the major bottleneck of this approach. Therefore, we optimised the whole DIA workflow using a range of samples: we highlighted guidelines for sample preparation, data acquisition and data analysis. In particular, we extensively optimised the peptide-centric DIA data extraction parameters to improve the sensitivity and specificity of the workflow. We showed that a homemade library is still today the most efficient way to treat DIA data. However, developments are still needed in order to provide a better peak picking and interference management. Then, we showed that DIA offers much betters proteome coverage and reproducibility when compared to a classic shotgun approach using DDA, and equivalent sensitivity, specificity and reproducibility were reached compared to targeted approaches. Indeed, because of its data acquisition mode, DIA is unlimited in multiplexing, and absolutely all detectable data are collected, while DDA suffers from a strong undersampling effect, and targeted approaches are limited to ≈ 50-100 targeted proteins. Moreover, we showed that the systematic analysis of all peptides, combined with the use of MS/MS signals and HR/AM instruments allow DIA to reach equivalent sensitivity, specificity and reproducibility when compared to targeted approaches. In conclusion, though further improvements in DIA data analysis are still needed, notably for data analysis (spectrum centric approach), DIA keeps its promises, and may become the method of choice for mass spectrometry-based proteomics in the coming years.

Thereby, I produced a range of mAb samples from different process steps and conditions, and developed an innovative dual DIA approach, called Top 3-ID-DIA, allowing both global HCP profiling using Top 3 estimations, and absolute quantification of key HCP using isotope dilution (ID). We quantified HCP within a dynamic range of 5 orders of magnitude, and down to sub ppm level. This method was benchmarked against reference methods for HCP quantification (ELISA) and accurate quantification by mass spectrometry (ID-SRM). Overall, the Top 3-ID-DIA approach reached equivalent sensitivity, accuracy and precision when compared to ID-SRM. Moreover, it allowed an unbiased and more comprehensive HCP characterisation when compared to ELISA.

This method could be readily transferred to industry, and could be applicable in only 2 months, which is to be compared with ELISA development which takes usually more than one year (**Figure 61**).

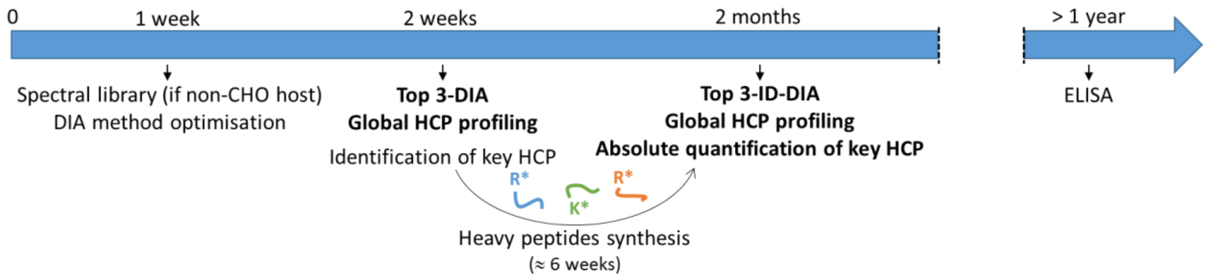


***Figure 61 : Timescale of the Top 3-ID-DIA method transfer to industry.***
*If CHO host cells are used, the spectral library generated during this PhD could be used. If different host cells are used (e.g. yeast or E. Coli), a new spectral library can be generated and the DIA method could be optimised within one week. A global HCP profiling could be obtained after an additional week, allowing the identification of key HCP. Stable isotope labelled peptides could be ordered, which typically takes ≈6 weeks, and after only 2 months the Top 3-ID-DIA method could be ready to use. On the other hand, ELISA development usually takes more than one year.*

Even if this methodology is of lower throughput (12 samples per day *vs* 21 samples per day for ELISA) and more expensive (≈ 600 k€ for a LC-MS coupling *vs* ≈ 200 k€ to develop an ELISA), it should be kept in mind that ELISA needs to be re-developed when there are no more ELISA reagents (e.g. anti-HCP antibodies), which also leads to bridging issues due to the inconsistency of ELISA results between kits (different HCP mixtures injected into animals leading to different anti-HCP antibodies and different results). Moreover, it should be balanced with the amount of information provided by MS when compared to ELISA, with more than 3 000 HCP quantified in this work, while ≈ 1 000 HCP can be quantified by ELISA. MS also provides identification and individual quantification of HCP, allowing a comprehensive risk assessment, while ELISA only provides total HCP amounts without information on the number nor identity of the detected HCP.

In conclusion, the Top 3-ID-DIA method could allow the release of cleaner and safer biotherapeutics, and could also provide a real time support to bioprocess development which is not provided by ELISA. In the short term, the Top 3-ID-DIA method could provide complementary results to ELISA, and in the long term totally replace it.

# References

1. Milo, R., What is the total number of protein molecules per cell volume? A call to rethink some published values. *BioEssays* **2013,** *35* (12), 1050-1055.
2. Zubarev, R. A., The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics* **2013,** *13* (5), 723-6.
3. Steen, H.; Mann, M., The abc's (and xyz's) of peptide sequencing. *Nat Rev Mol Cell Biol* **2004,** *5* (9), 699-711.
4. Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M.-C.; Yates, J. R., Protein Analysis by Shotgun/Bottom-up Proteomics. *Chemical reviews* **2013,** *113* (4), 2343-2394.
5. Beranova-Giorgianni, S., Proteome analysis by two-dimensional gel electrophoresis and mass spectrometry: strengths and limitations. *TrAC Trends in Analytical Chemistry* **2003,** *22* (5), 273-281.
6. Zhang, Z.; Wu, S.; Stenoien, D. L.; Pasa-Tolic, L., High-throughput proteomics. *Annu Rev Anal Chem (Palo Alto Calif)* **2014,** *7*, 427-54.
7. Marko-Varga, G.; Fehniger, T. E., Proteomics and disease--the challenges for technology and discovery. *J Proteome Res* **2004,** *3* (2), 167-78.
8. Aebersold, R.; Mann, M., Mass-spectrometric exploration of proteome structure and function. *Nature* **2016,** *537* (7620), 347-55.
9. Richards, A. L.; Hebert, A. S.; Ulbrich, A.; Bailey, D. J.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J., One-hour proteome analysis in yeast. *Nat Protoc* **2015,** *10* (5), 701-14.
10. Blueggel, M.; Chamrad, D.; Meyer, H. E., Bioinformatics in proteomics. *Curr Pharm Biotechnol* **2004,** *5* (1), 79-88.
11. Nesvizhskii, A. I.; Aebersold, R., Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **2005,** *4* (10), 1419-40.
12. Li, H.; Han, J.; Pan, J.; Liu, T.; Parker, C. E.; Borchers, C. H., Current trends in quantitative proteomics - an update. *Journal of mass spectrometry : JMS* **2017,** *52* (5), 319-341.
13. Bakalarski, C. E.; Kirkpatrick, D. S., A Biologist's Field Guide to Multiplexed Quantitative Proteomics. *Mol Cell Proteomics* **2016**.
14. Schilling, B.; Rardin, M. J.; MacLean, B. X.; Zawadzka, A. M.; Frewen, B. E.; Cusack, M. P.; Sorensen, D. J.; Bereman, M. S.; Jing, E.; Wu, C. C.; Verdin, E.; Kahn, C. R.; Maccoss, M. J.; Gibson, B. W., 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), 202-14.
15. Tabb, D. L.; Vega-Montoto, L.; Rudnick, P. A.; Variyath, A. M.; Ham, A.-J. L.; Bunk, D. M.; Kilpatrick, L. E.; Billheimer, D. D.; Blackman, R. K.; Cardasis, H. L.; Carr, S. A.; Clauser, K. R.; Jaffe, J. D.; Kowalski, K. A.; Neubert, T. A.; Regnier, F. E.; Schilling, B.; Tegeler, T. J.; Wang, M.; Wang, P.; Whiteaker, J. R.; Zimmerman, L. J.; Fisher, S. J.; Gibson, B. W.; Kinsinger, C. R.; Mesri, M.; Rodriguez, H.; Stein, S. E.; Tempst, P.; Paulovich, A. G.; Liebler, D. C.; Spiegelman, C., Repeatability and Reproducibility in Proteomic Identifications by Liquid Chromatography−Tandem Mass Spectrometry. *Journal of Proteome Research* **2010,** *9* (2), 761-776.
16. Mermelekas, G.; Vlahou, A.; Zoidakis, J., SRM/MRM targeted proteomics as a tool for biomarker validation and absolute quantification in human urine. *Expert Rev Mol Diagn* **2015,** *15* (11), 1441-54.
17. Dupin, M.; Fortin, T.; Larue-Triolet, A.; Surault, I.; Beaulieu, C.; Gouel-Cheron, A.; Allaouchiche, B.; Asehnoune, K.; Roquilly, A.; Venet, F.; Monneret, G.; Lacoux, X.; Roitsch, C. A.; Pachot, A.; Charrier, J. P.; Pons, S., Impact of Serum and Plasma Matrices on the Titration of Human Inflammatory Biomarkers Using Analytically Validated SRM Assays. *J Proteome Res* **2016,** *15* (8), 2366-78.

18. Reis-de-Oliveira, G.; Garcia, S.; Guest, P. C.; Cassoli, J. S.; Martins-de-Souza, D., A Selected Reaction Monitoring Mass Spectrometry Protocol for Validation of Proteomic Biomarker Candidates in Studies of Psychiatric Disorders. *Advances in experimental medicine and biology* **2017,** *974*, 213-218.

19. Prochazkova, I.; Lenco, J.; Fucikova, A.; Dresler, J.; Capkova, L.; Hrstka, R.; Nenutil, R.; Bouchal, P., Targeted proteomics driven verification of biomarker candidates associated with breast cancer aggressiveness. *Biochim Biophys Acta* **2017,** *1865* (5), 488-498.

20. Percy, A. J.; Yang, J.; Chambers, A. G.; Mohammed, Y.; Miliotis, T.; Borchers, C. H., Protocol for Standardizing High-to-Moderate Abundance Protein Biomarker Assessments Through an MRM-with-Standard-Peptides Quantitative Approach. *Advances in experimental medicine and biology* **2016,** *919*, 515-530.

21. Rauniyar, N.; Peng, G.; Lam, T. T.; Zhao, H.; Mor, G.; Williams, K. R., Data-Independent Acquisition and Parallel Reaction Monitoring Mass Spectrometry Identification of Serum Biomarkers for Ovarian Cancer. *Biomarker Insights* **2017,** *12*, 1177271917710948.

22. Kim, H. J.; Lin, D.; Lee, H. J.; Li, M.; Liebler, D. C., Quantitative Profiling of Protein Tyrosine Kinases in Human Cancer Cell Lines by Multiplexed Parallel Reaction Monitoring Assays. *Mol Cell Proteomics* **2016,** *15* (2), 682-91.

23. Thomas, S. N.; Harlan, R.; Chen, J.; Aiyetan, P.; Liu, Y.; Sokoll, L. J.; Aebersold, R.; Chan, D. W.; Zhang, H., Multiplexed Targeted Mass Spectrometry-Based Assays for the Quantification of N-Linked Glycosite-Containing Peptides in Serum. *Anal Chem* **2015,** *87* (21), 10830-8.

24. Gallien, S.; Duriez, E.; Crone, C.; Kellmann, M.; Moehring, T.; Domon, B., Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. *Mol Cell Proteomics* **2012,** *11* (12), 1709-23.

25. Schilling, B.; MacLean, B.; Held, J. M.; Sahu, A. K.; Rardin, M. J.; Sorensen, D. J.; Peters, T.; Wolfe, A. J.; Hunter, C. L.; MacCoss, M. J.; Gibson, B. W., 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), 10222-9.

26. Tong, L.; Zhou, X. Y.; Jylha, A.; Aapola, U.; Liu, D. N.; Koh, S. K.; Tian, D.; Quah, J.; Uusitalo, H.; Beuerman, R. W.; Zhou, L., Quantitation of 47 human tear proteins using high resolution multiple reaction monitoring (HR-MRM) based-mass spectrometry. *J Proteomics* **2015,** *115*, 36-48.

27. Gillet, L. C.; 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.

28. Tsou, C. C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A. C.; Nesvizhskii, A. I., DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* **2015,** *12* (3), 258-64, 7 p following 264.

29. Navarro, P.; Kuharev, J.; Gillet, L. C.; Bernhardt, O. M.; MacLean, B.; Rost, H. L.; Tate, S. A.; Tsou, C. C.; Reiter, L.; Distler, U.; Rosenberger, G.; Perez-Riverol, Y.; Nesvizhskii, A. I.; Aebersold, R.; Tenzer, S., A multicenter study benchmarks software tools for label-free proteome quantification. *Nat Biotechnol* **2016**.

30. Ecker, D. M.; Jones, S. D.; Levine, H. L., The therapeutic monoclonal antibody market. *MAbs* **2015,** *7* (1), 9-14.

31. Reichert, J. M., Antibodies to watch in 2017. *MAbs* **2017,** *9* (2), 167-181.

32. Doig, A.; Ecker, D.; Ransohoff, T., Monoclonal antibody targets and indications. *American Pharmaceutical Review 15* **2015,** *177490*.

33. Suzuki, M.; Kato, C.; Kato, A., Therapeutic antibodies: their mechanisms of action and the pathological findings they induce in toxicity studies. *Journal of Toxicologic Pathology* **2015,** *28* (3), 133-139.

34. Publishing, L. M. *2016 Sales of Recombinant Therapeutic Antibodies & Proteins*; Mar 2017 2017.

35. Jayapal, K. P.; Wlaschin, K. F.; Hu, W. S.; Yap, M. G. S., Recombinant protein therapeutics from CHO Cells - 20 years and counting. *Chemical Engineering Progress* **2007,** *103* (10).

36. Kim, J. Y.; Kim, Y.-G.; Lee, G. M., CHO cells in biotechnology for production of recombinant proteins: current state and further potential. *Applied Microbiology and Biotechnology* **2012,** *93* (3), 917-930.

37. ICH *Guidance for Industry Q6B Specifications: Test Procedures and Acceptance Criteria for Biotechnological/Biological products*; 1999.

38. Gao, S. X.; Zhang, Y.; Stansberry-Perkins, K.; Buko, A.; Bai, S.; Nguyen, V.; Brader, M. L., Fragmentation of a highly purified monoclonal antibody attributed to residual CHO cell protease activity. *Biotechnol Bioeng* **2011,** *108* (4), 977-82.

39. Bee, J. S.; Tie, L.; Afdahl, C. D.; Jusino, K. C.; Johnson, D.; Dimitrova, M. N., Trace levels of the CHO host cell protease cathepsin D caused particle formation in a monoclonal antibody product. *Biotechnol Prog* **2015**.

40. Robert, F.; Bierau, H.; Rossi, M.; Agugiaro, D.; Soranzo, T.; Broly, H.; Mitchell-Logean, C., Degradation of an Fc-fusion recombinant protein by host cell proteases: Identification of a CHO cathepsin D protease. *Biotechnol Bioeng* **2009,** *104* (6), 1132-41.

41. Hanania, N. A.; Noonan, M.; Corren, J.; Korenblat, P.; Zheng, Y.; Fischer, S. K.; Cheu, M.; Putnam, W. S.; Murray, E.; Scheerens, H.; Holweg, C. T.; Maciuca, R.; Gray, S.; Doyle, R.; McClintock, D.; Olsson, J.; Matthews, J. G.; Yen, K., Lebrikizumab in moderate-to-severe asthma: pooled data from two randomised placebo-controlled studies. *Thorax* **2015,** *70* (8), 748-56.

42. Fischer, S. K.; Cheu, M.; Peng, K.; Lowe, J.; Araujo, J.; Murray, E.; McClintock, D.; Matthews, J.; Siguenza, P.; Song, A., Specific Immune Response to Phospholipase B-Like 2 Protein, a Host Cell Impurity in Lebrikizumab Clinical Material. *The AAPS journal* **2017,** *19* (1), 254-263.

43. Zhu-Shimoni, J.; Yu, C.; Nishihara, J.; Wong, R. M.; Gunawan, F.; Lin, M.; Krawitz, D.; Liu, P.; Sandoval, W.; Vanderlaan, M., Host cell protein testing by ELISAs and the use of orthogonal methods. *Biotechnol Bioeng* **2014,** *111* (12), 2367-79.

44. Bracewell, D. G.; Francis, R.; Smales, C. M., The future of host cell protein (HCP) identification during process development and manufacturing linked to a risk-based management for their control. *Biotechnol Bioeng* **2015,** *112* (9), 1727-37.

45. Tscheliessnig, A. L.; Konrath, J.; Bates, R.; Jungbauer, A., Host cell protein analysis in therapeutic protein bioprocessing - methods and applications. *Biotechnol J* **2013,** *8* (6), 655-70.

46. Meleady, P.; Hoffrogge, R.; Henry, M.; Rupp, O.; Bort, J. H.; Clarke, C.; Brinkrolf, K.; Kelly, S.; Muller, B.; Doolan, P.; Hackl, M.; Beckmann, T. F.; Noll, T.; Grillari, J.; Barron, N.; Puhler, A.; Clynes, M.; Borth, N., Utilization and evaluation of CHO-specific sequence databases for mass spectrometry based proteomics. *Biotechnol Bioeng* **2012,** *109* (6), 1386-94.

47. Vidova, V.; Spacil, Z., A review on mass spectrometry-based quantitative proteomics: Targeted and data independent acquisition. *Anal Chim Acta* **2017,** *964*, 7-23.

48. Picotti, P.; Aebersold, R., Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat Methods* **2012,** *9* (6), 555-66.

49. Nakamura, K.; Hirayama-Kurogi, M.; Ito, S.; Kuno, T.; Yoneyama, T.; Obuchi, W.; Terasaki, T.; Ohtsuki, S., Large-scale multiplex absolute protein quantification of drug-metabolizing enzymes and transporters in human intestine, liver, and kidney microsomes by SWATH-MS: Comparison with MRM/SRM and HR-MRM/PRM. *Proteomics* **2016,** *16* (15-16), 2106-17.

50. Percy, A. J.; Tamura-Wells, J.; Albar, J. P.; Aloria, K.; Amirkhani, A.; Araujo, G. D. T.; Arizmendi, J. M.; Blanco, F. J.; Canals, F.; Cho, J.-Y.; Colomé-Calls, N.; Corrales, F. J.; Domont, G.; Espadas, G.; Fernandez-Puente, P.; Gil, C.; Haynes, P. A.; Hernáez, M. L.; Kim, J. Y.; Kopylov, A.; Marcilla, M.; McKay, M. J.; Mirzaei, M.; Molloy, M. P.; Ohlund, L. B.; Paik, Y.-K.; Paradela, A.; Raftery, M.; Sabidó, E.; Sleno, L.; Wilffert, D.; Wolters, J. C.; Yoo, J. S.; Zgoda, V.; Parker, C. E.; Borchers, C. H., Inter-laboratory evaluation of instrument platforms and experimental workflows for quantitative accuracy and reproducibility assessment. *EuPA Open Proteomics* **2015,** *8*, 6-15.

51. Percy, A. J.; Chambers, A. G.; Yang, J.; Domanski, D.; Borchers, C. H., Comparison of standard- and nano-flow liquid chromatography platforms for MRM-based quantitation of putative plasma biomarker proteins. *Anal Bioanal Chem* **2012,** *404* (4), 1089-101.

52. MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J., Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010,** *26* (7), 966-8.

53. Rosenberger, G.; Koh, C. C.; Guo, T.; Rost, H. L.; Kouvonen, P.; Collins, B. C.; Heusel, M.; Liu, Y.; Caron, E.; Vichalkovski, A.; Faini, M.; Schubert, O. T.; Faridi, P.; Ebhardt, H. A.; Matondo, M.; Lam, H.; Bader, S. L.; Campbell, D. S.; Deutsch, E. W.; Moritz, R. L.; Tate, S.; Aebersold, R., A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci Data* **2014,** *1*, 140031.

54. Reiter, L.; Rinner, O.; Picotti, P.; Huttenhain, R.; Beck, M.; Brusniak, M. Y.; Hengartner, M. O.; Aebersold, R., mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat Methods* **2011,** *8* (5), 430-5.

55. 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), 1367-72.

56. Wang, X.; Hunter, A. K.; Mozier, N. M., Host cell proteins in biologics development: Identification, quantitation and risk assessment. *Biotechnol Bioeng* **2009,** *103* (3), 446-58.

57. Hogwood, C. E.; Bracewell, D. G.; Smales, C. M., Measurement and control of host cell proteins (HCPs) in CHO cell bioprocesses. *Curr Opin Biotechnol* **2014,** *30C*, 153-160.

58. Jin, M.; Szapiel, N.; Zhang, J.; Hickey, J.; Ghose, S., Profiling of host cell proteins by two-dimensional difference gel electrophoresis (2D-DIGE): Implications for downstream process development. *Biotechnol Bioeng* **2010,** *105* (2), 306-16.

59. Liu, H. F.; Ma, J.; Winter, C.; Bayer, R., Recovery and purification process development for monoclonal antibody production. *MAbs* **2010,** *2* (5), 480-99.

60. Silva, J. C.; Gorenstein, M. V.; Li, G. Z.; Vissers, J. P.; Geromanos, S. J., Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics* **2006,** *5* (1), 144-56.

61. Smith, L. M.; Kelleher, N. L., Proteoform: a single term describing protein complexity. *Nat Meth* **2013,** *10* (3), 186-187.

62. Wang, E. T.; Sandberg, R.; Luo, S.; Khrebtukova, I.; Zhang, L.; Mayr, C.; Kingsmore, S. F.; Schroth, G. P.; Burge, C. B., Alternative isoform regulation in human tissue transcriptomes. *Nature* **2008,** *456* (7221), 470-6.

63. Gstaiger, M.; Aebersold, R., Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat Rev Genet* **2009,** *10* (9), 617-27.

64. Han, X.; Jin, M.; Breuker, K.; McLafferty, F. W., Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons. *Science* **2006,** *314* (5796), 109-12.

65. Cai, W.; Tucholski, T.; Chen, B.; Alpert, A. J.; McIlwain, S.; Kohmoto, T.; Jin, S.; Ge, Y., Top-Down Proteomics of Large Proteins up to 223 kDa Enabled by Serial Size Exclusion Chromatography Strategy. *Anal Chem* **2017,** *89* (10), 5467-5475.

66. Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M.; Wu, C.; Sweet, S. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L., Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **2011,** *480* (7376), 254-8.

67. Catherman, A. D.; Durbin, K. R.; Ahlf, D. R.; Early, B. P.; Fellers, R. T.; Tran, J. C.; Thomas, P. M.; Kelleher, N. L., Large-scale top-down proteomics of the human proteome: membrane proteins, mitochondria, and senescence. *Mol Cell Proteomics* **2013,** *12* (12), 3465-73.

68. Durbin, K. R.; Fornelli, L.; Fellers, R. T.; Doubleday, P. F.; Narita, M.; Kelleher, N. L., Quantitation and Identification of Thousands of Human Proteoforms below 30 kDa. *J Proteome Res* **2016,** *15* (3), 976-82.

69. Wu, S.-L.; Hühmer, A. F. R.; Hao, Z.; Karger, B. L., On-Line LC−MS Approach Combining Collision-Induced Dissociation (CID), Electron-Transfer Dissociation (ETD), and CID of an Isolated Charge-Reduced Species for the Trace-Level Characterization of Proteins with Post-Translational Modifications. *Journal of Proteome Research* **2007,** *6* (11), 4230-4244.

70. Gregorich, Z. R.; Ge, Y., Top-down proteomics in health and disease: challenges and opportunities. *Proteomics* **2014,** *14* (10), 1195-210.

71. Compton, P. D.; Zamdborg, L.; Thomas, P. M.; Kelleher, N. L., On the scalability and requirements of whole protein mass spectrometry. *Anal Chem* **2011,** *83* (17), 6868-74.
72. Wu, C.; Tran, J. C.; Zamdborg, L.; Durbin, K. R.; Li, M.; Ahlf, D. R.; Early, B. P.; Thomas, P. M.; Sweedler, J. V.; Kelleher, N. L., A Protease for Middle Down Proteomics. *Nature methods* **2012,** *9* (8), 822-824.
73. Switzar, L.; Giera, M.; Niessen, W. M., Protein digestion: an overview of the available techniques and recent developments. *J Proteome Res* **2013,** *12* (3), 1067-77.
74. Picotti, P.; Bodenmiller, B.; Mueller, L. N.; Domon, B.; Aebersold, R., Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* **2009,** *138* (4), 795-806.
75. Kota, U.; Stolowitz, M. L., Improving Proteome Coverage by Reducing Sample Complexity via Chromatography. *Advances in experimental medicine and biology* **2016,** *919*, 83-143.
76. Bradford, M. M., A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* **1976,** *72*, 248-54.
77. Noble, J. E.; Bailey, M. J., Quantitation of protein. *Methods Enzymol* **2009,** *463*, 73-95.
78. Lowry, O. H.; Rosebrough, N. J.; Farr, A. L.; Randall, R. J., Protein measurement with the Folin phenol reagent. *The Journal of biological chemistry* **1951,** *193* (1), 265-75.
79. Peterson, G. L., Review of the Folin phenol protein quantitation method of Lowry, Rosebrough, Farr and Randall. *Anal Biochem* **1979,** *100* (2), 201-20.
80. Mesmin, C.; van Oostrum, J.; Domon, B., Complexity reduction of clinical samples for routine mass spectrometric analysis. *Proteomics Clin Appl* **2016,** *10* (4), 315-22.
81. Laemmli, U. K., Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **1970,** *227* (5259), 680-5.
82. Lu, X.; Zhu, H., Tube-gel digestion: a novel proteomic approach for high throughput analysis of membrane proteins. *Mol Cell Proteomics* **2005,** *4* (12), 1948-58.
83. Muller, L.; Fornecker, L.; Dorsselaer, A.; Cianferani, S.; Carapito, C., Benchmarking sample preparation/digestion protocols reveals tube-gel being a fast and repeatable method for quantitative proteomics. *Proteomics* **2016**.
84. Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M., Universal sample preparation method for proteome analysis. *Nat Meth* **2009,** *6* (5), 359-362.
85. Wisniewski, J. R., Quantitative Evaluation of Filter Aided Sample Preparation (FASP) and Multienzyme Digestion FASP Protocols. *Anal Chem* **2016,** *88* (10), 5438-43.
86. Coleman, O.; Henry, M.; Clynes, M.; Meleady, P., Filter-Aided Sample Preparation (FASP) for Improved Proteome Analysis of Recombinant Chinese Hamster Ovary Cells. *Methods Mol Biol* **2017,** *1603*, 187-194.
87. Kachuk, C.; Stephen, K.; Doucette, A., Comparison of sodium dodecyl sulfate depletion techniques for proteome analysis by mass spectrometry. *Journal of Chromatography A* **2015,** *1418*, 158-166.
88. Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M., Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal Chem* **1996,** *68*.
89. Rosenfeld, J.; Capdevielle, J.; Guillemot, J. C.; Ferrara, P., In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis. *Anal Biochem* **1992,** *203* (1), 173-9.
90. Shen, Y.; Zhao, R.; Berger, S. J.; Anderson, G. A.; Rodriguez, N.; Smith, R. D., High-efficiency nanoscale liquid chromatography coupled on-line with mass spectrometry using nanoelectrospray ionization for proteomics. *Anal Chem* **2002,** *74* (16), 4235-49.
91. Karas, M.; Hillenkamp, F., Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical chemistry* **1988,** *60* (20), 2299-301.
92. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M., Electrospray ionization for mass spectrometry of large biomolecules. *Science (New York, N.Y.)* **1989,** *246* (4926), 64-71.
93. Meher, A. K.; Chen, Y. C., Electrospray Modifications for Advancing Mass Spectrometric Analysis. *Mass spectrometry (Tokyo, Japan)* **2017,** *6* (Spec Iss), S0057.

94. Sleno, L.; Volmer, D. A., Ion activation methods for tandem mass spectrometry. *J Mass Spectrom* **2004,** *39* (10), 1091-112.

95. Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F., Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A* **2004,** *101* (26), 9528-33.

96. Zubarev, R. A.; Horn, D. M.; Fridriksson, E. K.; Kelleher, N. L.; Kruger, N. A.; Lewis, M. A.; Carpenter, B. K.; McLafferty, F. W., Electron capture dissociation for structural characterization of multiply charged protein cations. *Anal Chem* **2000,** *72* (3), 563-73.

97. Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A., Mobile and localized protons: a framework for understanding peptide dissociation. *J Mass Spectrom* **2000,** *35* (12), 1399-406.

98. Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M., Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods* **2007,** *4* (9), 709-12.

99. Biemann, K., Appendix 5. Nomenclature for peptide fragment ions (positive ions). *Methods Enzymol* **1990,** *193*, 886-7.

100. Frese, C. K.; Zhou, H.; Taus, T.; Altelaar, A. F.; Mechtler, K.; Heck, A. J.; Mohammed, S., Unambiguous phosphosite localization using electron-transfer/higher-energy collision dissociation (EThcD). *J Proteome Res* **2013,** *12* (3), 1520-5.

101. 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. *Journal of Proteome Research* **2011,** *10* (4), 1785-1793.

102. Eng, J. K.; McCormack, A. L.; Yates, J. R., 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), 976-89.

103. Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999,** *20* (18), 3551-67.

104. Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H., Open mass spectrometry search algorithm. *J Proteome Res* **2004,** *3* (5), 958-64.

105. Craig, R.; Cortens, J. P.; Beavis, R. C., Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* **2004,** *3* (6), 1234-42.

106. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M., Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **2011,** *10* (4), 1794-805.

107. Shteynberg, D.; Nesvizhskii, A. I.; Moritz, R. L.; Deutsch, E. W., Combining results of multiple search engines in proteomics. *Mol Cell Proteomics* **2013,** *12* (9), 2383-93.

108. Balgley, B. M.; Laudeman, T.; Yang, L.; Song, T.; Lee, C. S., Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. *Mol Cell Proteomics* **2007,** *6* (9), 1599-608.

109. Nesvizhskii, A. I., Proteogenomics: concepts, applications and computational strategies. *Nat Meth* **2014,** *11* (11), 1114-1125.

110. O'Leary, N. A.; Wright, M. W.; Brister, J. R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; Astashyn, A.; Badretdin, A.; Bao, Y.; Blinkova, O.; Brover, V.; Chetvernin, V.; Choi, J.; Cox, E.; Ermolaeva, O.; Farrell, C. M.; Goldfarb, T.; Gupta, T.; Haft, D.; Hatcher, E.; Hlavina, W.; Joardar, V. S.; Kodali, V. K.; Li, W.; Maglott, D.; Masterson, P.; McGarvey, K. M.; Murphy, M. R.; O'Neill, K.; Pujar, S.; Rangwala, S. H.; Rausch, D.; Riddick, L. D.; Schoch, C.; Shkeda, A.; Storz, S. S.; Sun, H.; Thibaud-Nissen, F.; Tolstoy, I.; Tully, R. E.; Vatsan, A. R.; Wallin, C.; Webb, D.; Wu, W.; Landrum, M. J.; Kimchi, A.; Tatusova, T.; DiCuccio, M.; Kitts, P.; Murphy, T. D.; Pruitt, K. D., Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **2016,** *44* (D1), D733-45.

111. McEntyre, J., Linking up with Entrez. *Trends Genet* **1998,** *14* (1), 39-40.

112. Wu, C. H.; Yeh, L. S.; Huang, H.; Arminski, L.; Castro-Alvear, J.; Chen, Y.; Hu, Z.; Kourtesis, P.; Ledley, R. S.; Suzek, B. E.; Vinayaka, C. R.; Zhang, J.; Barker, W. C., The Protein Information Resource. *Nucleic Acids Res* **2003,** *31* (1), 345-7.

113. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res* **2000,** *28* (1), 235-42.
114. Squizzato, S.; Park, Y. M.; Buso, N.; Gur, T.; Cowley, A.; Li, W.; Uludag, M.; Pundir, S.; Cham, J. A.; McWilliam, H.; Lopez, R., The EBI Search engine: providing search and retrieval functionality for biological data from EMBL-EBI. *Nucleic Acids Res* **2015,** *43* (W1), W585-8.
115. Mashima, J.; Kodama, Y.; Kosuge, T.; Fujisawa, T.; Katayama, T.; Nagasaki, H.; Okuda, Y.; Kaminuma, E.; Ogasawara, O.; Okubo, K.; Nakamura, Y.; Takagi, T., DNA data bank of Japan (DDBJ) progress report. *Nucleic Acids Res* **2016,** *44* (D1), D51-7.
116. Benson, D. A.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W., GenBank. *Nucleic Acids Res* **2015,** *43* (Database issue), D30-5.
117. UniProt: a hub for protein information. *Nucleic Acids Res* **2015,** *43* (Database issue), D204-12.
118. Elias, J. E.; Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **2007,** *4* (3), 207-14.
119. Elias, J. E.; Gygi, S. P., Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol Biol* **2010,** *604*, 55-71.
120. Navarro, P.; Vazquez, J., A refined method to calculate false discovery rates for peptide identification using decoy databases. *J Proteome Res* **2009,** *8* (4), 1792-6.
121. Omenn, G. S.; Lane, L.; Lundberg, E. K.; Beavis, R. C.; Nesvizhskii, A. I.; Deutsch, E. W., 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), 3452-60.
122. Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; 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), 376-86.
123. Miyagi, M.; Rao, K. C., Proteolytic 18O-labeling strategies for quantitative proteomics. *Mass spectrometry reviews* **2007,** *26* (1), 121-36.
124. Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R., Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **1999,** *17* (10), 994-9.
125. Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J., Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **2004,** *3* (12), 1154-69.
126. Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, A. K.; Hamon, C., Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical chemistry* **2003,** *75* (8), 1895-904.
127. Nahnsen, S.; Bielow, C.; Reinert, K.; Kohlbacher, O., Tools for label-free peptide quantification. *Mol Cell Proteomics* **2013,** *12* (3), 549-56.
128. Liu, H.; Sadygov, R. G.; Yates, J. R., 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **2004,** *76* (14), 4193-201.
129. Ramus, C.; Hovasse, A.; Marcellin, M.; Hesse, A. M.; Mouton-Barbosa, E.; Bouyssie, D.; Vaca, S.; Carapito, C.; Chaoui, K.; Bruley, C.; Garin, J.; Cianferani, S.; Ferro, M.; Van Dorssaeler, A.; Burlet-Schiltz, O.; Schaeffer, C.; Coute, Y.; Gonzalez de Peredo, A., Benchmarking quantitative label-free LC-MS data processing workflows using a complex spiked proteomic standard dataset. *J Proteomics* **2016,** *132*, 51-62.
130. Bouyssie, D.; Gonzalez de Peredo, A.; Mouton, E.; Albigot, R.; Roussel, L.; Ortega, N.; Cayrol, C.; Burlet-Schiltz, O.; Girard, J. P.; 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), 1621-37.
131. Mallick, P.; Schirle, M.; Chen, S. S.; Flory, M. R.; Lee, H.; Martin, D.; Ranish, J.; Raught, B.; Schmitt, R.; Werner, T.; Kuster, B.; Aebersold, R., Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol* **2007,** *25* (1), 125-31.

132. Maclean, B.; Tomazela, D. M.; Abbatiello, S. E.; Zhang, S.; Whiteaker, J. R.; Paulovich, A. G.; Carr, S. A.; Maccoss, M. J., Effect of collision energy optimization on the measurement of peptides by selected reaction monitoring (SRM) mass spectrometry. *Anal Chem* **2010,** *82* (24), 10116-24.

133. Holstein Sherwood, C. A.; Gafken, P. R.; Martin, D. B., Collision energy optimization of b- and y-ions for multiple reaction monitoring mass spectrometry. *J Proteome Res* **2011,** *10* (1), 231-40.

134. Gerber, S. A.; Rush, J.; Stemman, O.; Kirschner, M. W.; Gygi, S. P., Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A* **2003,** *100* (12), 6940-5.

135. Pratt, J. M.; Simpson, D. M.; Doherty, M. K.; Rivers, J.; Gaskell, S. J.; Beynon, R. J., Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat Protoc* **2006,** *1* (2), 1029-43.

136. 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), 2139-49.

137. Armbruster, D. A.; Pry, T., Limit of Blank, Limit of Detection and Limit of Quantitation. *The Clinical Biochemist Reviews* **2008,** *29* (Suppl 1), S49-S52.

138. Kuzyk, M. A.; Smith, D.; Yang, J.; Cross, T. J.; Jackson, A. M.; Hardie, D. B.; Anderson, N. L.; Borchers, C. H., Multiple reaction monitoring-based, multiplexed, absolute quantitation of 45 proteins in human plasma. *Mol Cell Proteomics* **2009,** *8* (8), 1860-77.

139. Gallien, S.; Bourmaud, A.; Kim, S. Y.; Domon, B., Technical considerations for large-scale parallel reaction monitoring analysis. *J Proteomics* **2014,** *100*, 147-59.

140. Peterson, A. C.; Russell, J. D.; Bailey, D. J.; Westphall, M. S.; Coon, J. J., Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics. *Molecular & Cellular Proteomics : MCP* **2012,** *11* (11), 1475-1488.

141. 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), 964-80.

142. 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), 307-21.

143. Meyer, J. G.; Schilling, B., Clinical applications of quantitative proteomics using targeted and untargeted data-independent acquisition techniques. *Expert Review of Proteomics* **2017,** *14* (5), 419-429.

144. Gillet, L. C.; Leitner, A.; Aebersold, R., Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annu Rev Anal Chem (Palo Alto Calif)* **2016**.

145. Rardin, M. J.; Schilling, B.; Cheng, L.-Y.; MacLean, B. X.; Sorensen, D. J.; Sahu, A. K.; MacCoss, M. J.; Vitek, O.; Gibson, B. W., MS1 Peptide Ion Intensity Chromatograms in MS2 (SWATH) Data Independent Acquisitions. Improving Post Acquisition Analysis of Proteomic Experiments. *Molecular & Cellular Proteomics : MCP* **2015,** *14* (9), 2405-2419.

146. Purvine, S.; Eppel, J. T.; Yi, E. C.; Goodlett, D. R., Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics* **2003,** *3* (6), 847-50.

147. Silva, J. C.; Denny, R.; Dorschel, C. A.; Gorenstein, M.; Kass, I. J.; Li, G. Z.; McKenna, T.; Nold, M. J.; Richardson, K.; Young, P.; Geromanos, S., Quantitative proteomic analysis by accurate mass retention time pairs. *Anal Chem* **2005,** *77* (7), 2187-200.

148. Geiger, T.; Cox, J.; Mann, M., Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol Cell Proteomics* **2010,** *9* (10), 2252-61.

149. Venable, J. D.; Dong, M. Q.; Wohlschlegel, J.; Dillin, A.; Yates, J. R., Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods* **2004,** *1* (1), 39-45.

150. Panchaud, A.; Scherl, A.; Shaffer, S. A.; von Haller, P. D.; Kulasekara, H. D.; Miller, S. I.; Goodlett, D. R., Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Anal Chem* **2009,** *81* (15), 6481-8.

151. Panchaud, A.; Jung, S.; Shaffer, S. A.; Aitchison, J. D.; Goodlett, D. R., Faster, quantitative, and accurate precursor acquisition independent from ion count. *Anal Chem* **2011,** *83* (6), 2250-7.

152. Weisbrod, C. R.; Eng, J. K.; Hoopmann, M. R.; Baker, T.; Bruce, J. E., Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification. *J Proteome Res* **2012,** *11* (3), 1621-32.

153. Egertson, J. D.; Kuehn, A.; Merrihew, G. E.; Bateman, N. W.; MacLean, B. X.; Ting, Y. S.; Canterbury, J. D.; Marsh, D. M.; Kellmann, M.; Zabrouskov, V.; Wu, C. C.; MacCoss, M. J., Multiplexed MS/MS for Improved Data Independent Acquisition. *Nature methods* **2013,** *10* (8), 744-746.

154. Shliaha, P. V.; Bond, N. J.; Gatto, L.; Lilley, K. S., Effects of traveling wave ion mobility separation on data independent acquisition in proteomics studies. *J Proteome Res* **2013,** *12* (6), 2323-39.

155. Distler, U.; Kuharev, J.; Navarro, P.; Levin, Y.; Schild, H.; Tenzer, S., Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nat Methods* **2014,** *11* (2), 167-70.

156. Martin, L. B.; Sherwood, R. W.; Nicklay, J. J.; Yang, Y.; Muratore-Schroeder, T. L.; Anderson, E. T.; Thannhauser, T. W.; Rose, J. K.; Zhang, S., Application of wide selected-ion monitoring data-independent acquisition to identify tomato fruit proteins regulated by the CUTIN DEFICIENT2 transcription factor. *Proteomics* **2016,** *16* (15-16), 2081-94.

157. Prakash, A.; Peterman, S.; Ahmad, S.; Sarracino, D.; Frewen, B.; Vogelsang, M.; Byram, G.; Krastins, B.; Vadali, G.; Lopez, M., Hybrid data acquisition and processing strategies with increased throughput and selectivity: pSMART analysis for global qualitative and quantitative analysis. *J Proteome Res* **2014,** *13* (12), 5415-30.

158. Moseley, M. A.; Hughes, C. J.; Juvvadi, P. R.; Soderblom, E. J.; Lennon, S.; Perkins, S. R.; Thompson, J. W.; Steinbach, W. J.; Geromanos, S. J.; Wildgoose, J.; Langridge, J. I.; Richardson, K.; Vissers, J. P. C., Scanning Quadrupole Data Independent Acquisition - Part A. Qualitative and Quantitative Characterization. *J Proteome Res* **2017**.

159. Ting, Y. S.; Egertson, J. D.; Payne, S. H.; Kim, S.; MacLean, B.; Kall, L.; Aebersold, R.; Smith, R. D.; Noble, W. S.; MacCoss, M. J., Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Mol Cell Proteomics* **2015,** *14* (9), 2301-7.

160. Egertson, J. D.; MacLean, B.; Johnson, R.; Xuan, Y.; MacCoss, M. J., Multiplexed Peptide Analysis using Data Independent Acquisition and Skyline. *Nature protocols* **2015,** *10* (6), 887-903.

161. Rost, H. L.; Rosenberger, G.; Navarro, P.; Gillet, L.; Miladinovic, S. M.; Schubert, O. T.; Wolski, W.; Collins, B. C.; Malmstrom, J.; Malmstrom, L.; Aebersold, R., OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol* **2014,** *32* (3), 219-23.

162. Bruderer, R.; Bernhardt, O. M.; Gandhi, T.; Miladinovic, S. M.; Cheng, L. Y.; Messner, S.; Ehrenberger, T.; Zanotelli, V.; Butscheid, Y.; Escher, C.; Vitek, O.; Rinner, O.; Reiter, L., Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol Cell Proteomics* **2015,** *14* (5), 1400-10.

163. Schubert, O. T.; Gillet, L. C.; Collins, B. C.; Navarro, P.; Rosenberger, G.; Wolski, W. E.; Lam, H.; Amodei, D.; Mallick, P.; MacLean, B.; Aebersold, R., Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nat Protoc* **2015,** *10* (3), 426-41.

164. Bruderer, R.; Bernhardt, O. M.; Gandhi, T.; Reiter, L., High-precision iRT prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation. *Proteomics* **2016,** *16* (15-16), 2246-56.

165. Caron, E.; Espona, L.; Kowalewski, D. J.; Schuster, H.; Ternette, N.; Alpizar, A.; Schittenhelm, R. B.; Ramarathinam, S. H.; Lindestam Arlehamn, C. S.; Chiek Koh, C.; Gillet, L. C.; Rabsteyn, A.; Navarro, P.; Kim, S.; Lam, H.; Sturm, T.; Marcilla, M.; Sette, A.; Campbell, D. S.; Deutsch, E. W.; Moritz, R. L.; Purcell, A. W.; Rammensee, H. G.; Stevanovic, S.; Aebersold, R., An open-source computational and data resource to analyze digital maps of immunopeptidomes. *Elife* **2015,** *4*.

166. Picotti, P.; Clement-Ziza, M.; Lam, H.; Campbell, D. S.; Schmidt, A.; Deutsch, E. W.; Rost, H.; Sun, Z.; Rinner, O.; Reiter, L.; Shen, Q.; Michaelson, J. J.; Frei, A.; Alberti, S.; Kusebauch, U.; Wollscheid,

B.; Moritz, R. L.; Beyer, A.; Aebersold, R., A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* **2013,** *494* (7436), 266-70.

167. Tsou, C. C.; Tsai, C. F.; Teo, G. C.; Chen, Y. J.; Nesvizhskii, A. I., Untargeted, spectral library-free analysis of data-independent acquisition proteomics data generated using Orbitrap mass spectrometers. *Proteomics* **2016,** *16* (15-16), 2257-71.

168. Kohler, G.; Milstein, C., Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* **1975,** *256* (5517), 495-7.

169. Elgundi, Z.; Reslan, M.; Cruz, E.; Sifniotis, V.; Kayser, V., The state-of-play and future of antibody therapeutics. *Advanced drug delivery reviews* **2016**.

170. Liu, J. K. H., The history of monoclonal antibody development – Progress, remaining challenges and future innovations. *Annals of Medicine and Surgery* **2014,** *3* (4), 113-116.

171. Frenzel, A.; Hust, M.; Schirrmann, T., Expression of Recombinant Antibodies. *Frontiers in Immunology* **2013,** *4*, 217.

172. Dumont, J.; Euwart, D.; Mei, B.; Estes, S.; Kshirsagar, R., Human cell lines for biopharmaceutical manufacturing: history, status, and future perspectives. *Critical reviews in biotechnology* **2016,** *36* (6), 1110-1122.

173. Sommerfeld, S.; Strube, J., Challenges in biotechnology production—generic processes and process optimization for monoclonal antibodies. *Chemical Engineering and Processing: Process Intensification* **2005,** *44* (10), 1123-1137.

174. Kunert, R.; Reinhart, D., Advances in recombinant antibody manufacturing. *Appl Microbiol Biotechnol* **2016,** *100* (8), 3451-61.

175. Shukla, A. A.; Suda, E., HARVEST AND RECOVERY OF MONOCLONAL ANTIBODIES: CELL REMOVAL AND CLARIFICATION. In *Process Scale Purification of Antibodies, Second Edition*, John Wiley & Sons, Inc.: 2017; pp 55-79.

176. Chon, J. H.; Zarbis-Papastoitsis, G., Advances in the production and downstream processing of antibodies. *N Biotechnol* **2011,** *28* (5), 458-63.

177. Follman, D. K.; Fahrner, R. L., Factorial screening of antibody purification processes using three chromatography steps without protein A. *J Chromatogr A* **2004,** *1024* (1-2), 79-85.

178. Jensen, K., A NORMALLY OCCURRING STAPHYLOCOCCUS ANTIBODY IN HUMAN SERUM. *Acta Pathologica Microbiologica Scandinavica* **1958,** *44* (4), 421-428.

179. Hober, S.; Nord, K.; Linhult, M., Protein A chromatography for antibody purification. *J Chromatogr B Analyt Technol Biomed Life Sci* **2007,** *848* (1), 40-7.

180. Gronemeyer, P.; Ditz, R.; Strube, J., Trends in upstream and downstream process development for antibody manufacturing. *Bioengineering* **2014,** *1* (4), 188-212.

181. Henry, S. M.; Sutlief, E.; Salas-Solano, O.; Valliere-Douglass, J., ELISA reagent coverage evaluation by affinity purification tandem mass spectrometry. *MAbs* **2017**, 1-11.

182. Sblattero, D.; Berti, I.; Trevisiol, C.; Marzari, R.; Tommasini, A.; Bradbury, A.; Fasano, A.; Ventura, A.; Not, T., Human recombinant tissue transglutaminase ELISA: an innovative diagnostic assay for celiac disease. *The American journal of gastroenterology* **2000,** *95* (5), 1253-7.

183. Griffin, J. F.; Spittle, E.; Rodgers, C. R.; Liggett, S.; Cooper, M.; Bakker, D.; Bannantine, J. P., Immunoglobulin G1 enzyme-linked immunosorbent assay for diagnosis of Johne's Disease in red deer (Cervus elaphus). *Clinical and diagnostic laboratory immunology* **2005,** *12* (12), 1401-9.

184. Porcelli, B.; Ferretti, F.; Vindigni, C.; Terzuoli, L., Assessment of a Test for the Screening and Diagnosis of Celiac Disease. *Journal of clinical laboratory analysis* **2016,** *30* (1), 65-70.

185. Winkler, I. G.; Lochelt, M.; Levesque, J. P.; Bodem, J.; Flugel, R. M.; Flower, R. L., A rapid streptavidin-capture ELISA specific for the detection of antibodies to feline foamy virus. *J Immunol Methods* **1997,** *207* (1), 69-77.

186. Zhu, M.; Gong, X.; Hu, Y.; Ou, W.; Wan, Y., Streptavidin-biotin-based directional double Nanobody sandwich ELISA for clinical rapid and sensitive detection of influenza H5N1. *Journal of translational medicine* **2014,** *12*, 352.

187. Peng, J.; Song, S.; Xu, L.; Ma, W.; Liu, L.; Kuang, H.; Xu, C., Development of a Monoclonal Antibody-Based Sandwich ELISA for Peanut Allergen Ara h 1 in Food. *International Journal of Environmental Research and Public Health* **2013,** *10* (7), 2897-2905.

188. Wang, S.-Y.; Li, Z.; Wang, X.-J.; Lv, S.; Yang, Y.; Zeng, L.-Q.; Luo, F.-H.; Yan, J.-H.; Liang, D.-F., Development of Monoclonal Antibody-Based Sandwich ELISA for Detection of Dextran. *Monoclonal Antibodies in Immunodiagnosis and Immunotherapy* **2014,** *33* (5), 334-339.

189. MacPhee, D. J., Methodological considerations for improving Western blot analysis. *Journal of pharmacological and toxicological methods* **2010,** *61* (2), 171-7.

190. Mahmood, T.; Yang, P.-C., Western Blot: Technique, Theory, and Trouble Shooting. *North American Journal of Medical Sciences* **2012,** *4* (9), 429-434.

191. Magdeldin, S.; Enany, S.; Yoshida, Y.; Xu, B.; Zhang, Y.; Zureena, Z.; Lokamani, I.; Yaoita, E.; Yamamoto, T., Basics and recent advances of two dimensional- polyacrylamide gel electrophoresis. *Clinical Proteomics* **2014,** *11* (1), 16.

192. Simpson, R. J., Rapid coomassie blue staining of protein gels. *Cold Spring Harbor protocols* **2010,** *2010* (4), pdb.prot5413.

193. Chevallet, M.; Luche, S.; Rabilloud, T., Silver staining of proteins in polyacrylamide gels. *Nature Protocols* **2006,** *1* (4), 1852-1858.

194. Grzeskowiak, J. K.; Tscheliessnig, A.; Toh, P. C.; Chusainow, J.; Lee, Y. Y.; Wong, N.; Jungbauer, A., 2-D DIGE to expedite downstream process development for human monoclonal antibody purification. *Protein Expr Purif* **2009,** *66* (1), 58-65.

195. Hayduk, E. J.; Choe, L. H.; Lee, K. H., A two-dimensional electrophoresis map of Chinese hamster ovary cell proteins based on fluorescence staining. *Electrophoresis* **2004,** *25* (15), 2545-56.

196. Bailey-Kellogg, C.; Gutierrez, A. H.; Moise, L.; Terry, F.; Martin, W. D.; De Groot, A. S., CHOPPI: a web tool for the analysis of immunogenicity risk from host cell proteins in CHO-based protein production. *Biotechnol Bioeng* **2014,** *111* (11), 2170-82.

197. Doneanu, C. E.; Anderson, M.; Williams, B. J.; Lauber, M. A.; Chakraborty, A.; Chen, W., Enhanced Detection of Low-Abundance Host Cell Protein Impurities in High-Purity Monoclonal Antibodies Down to 1 ppm Using Ion Mobility Mass Spectrometry Coupled with Multidimensional Liquid Chromatography. *Anal Chem* **2015,** *87* (20), 10283-91.

198. Zhang, Q.; Goetze, A. M.; Cui, H.; Wylie, J.; Tillotson, B.; Hewig, A.; Hall, M. P.; Flynn, G. C., Characterization of the co-elution of host cell proteins with monoclonal antibodies during protein A purification. *Biotechnol Prog* **2016,** *32* (3), 708-17.

199. Farrell, A.; Mittermayr, S.; Morrissey, B.; McLoughlin, N.; Navas Iglesias, N.; Marison, I. W.; Bones, J., Quantitative Host Cell Protein Analysis using Two Dimensional Data Independent LC-MS^E. *Analytical chemistry* **2015**.

200. Doneanu, C. E.; Xenopoulos, A.; Fadgen, K.; Murphy, J.; Skilton, S. J.; Prentice, H.; Stapels, M.; Chen, W., Analysis of host-cell proteins in biotherapeutic proteins by comprehensive online two-dimensional liquid chromatography/mass spectrometry. *mAbs* **2012,** *4* (1), 24-44.

201. Schenauer, M. R.; Flynn, G. C.; Goetze, A. M., Identification and quantification of host cell protein impurities in biotherapeutics using mass spectrometry. *Anal Biochem* **2012,** *428* (2), 150-7.

202. Zhang, Q.; Goetze, A. M.; Cui, H.; Wylie, J.; Trimble, S.; Hewig, A.; Flynn, G. C., Comprehensive tracking of host cell proteins during monoclonal antibody purifications using mass spectrometry. *mAbs* **2014,** *6* (3), 659-70.

203. Walker, D. E.; Yang, F.; Carver, J.; Joe, K.; Michels, D. A.; Yu, X. C., A modular and adaptive mass spectrometry-based platform for support of bioprocess development toward optimal host cell protein clearance. *MAbs* **2017,** *9* (4), 654-663.

204. Escher, C.; Reiter, L.; MacLean, B.; Ossola, R.; Herzog, F.; Chilton, J.; MacCoss, M. J.; Rinner, O., Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* **2012,** *12* (8), 1111-21.

205. Teleman, J.; Rost, H. L.; Rosenberger, G.; Schmitt, U.; Malmstrom, L.; Malmstrom, J.; Levander, F., DIANA--algorithmic improvements for analysis of data-independent acquisition MS data. *Bioinformatics* **2015,** *31* (4), 555-62.

206. Bilbao, A.; Zhang, Y.; Varesio, E.; Luban, J.; Strambio-De-Castillia, C.; Lisacek, F.; Hopfgartner, G., Ranking Fragment Ions Based on Outlier Detection for Improved Label-Free Quantification in Data-Independent Acquisition LC-MS/MS. *J Proteome Res* **2015,** *14* (11), 4581-93.
207. Keller, A.; Bader, S. L.; Shteynberg, D.; Hood, L.; Moritz, R. L., Automated Validation of Results and Removal of Fragment Ion Interferences in Targeted Analysis of Data-independent Acquisition Mass Spectrometry (MS) using SWATHProphet. *Mol Cell Proteomics* **2015,** *14* (5), 1411-8.
208. Scherl, A.; Shaffer, S. A.; Taylor, G. K.; Kulasekara, H. D.; Miller, S. I.; Goodlett, D. R., Genome-specific gas-phase fractionation strategy for improved shotgun proteomic profiling of proteotypic peptides. *Anal Chem* **2008,** *80* (4), 1182-91.

# Gauthier HUSSON

**Development of host cell protein impurities quantification methods by mass spectrometry to control the quality of biopharmaceuticals**

## Résumé

Les récents progrès instrumentaux en spectrométrie de masse, notamment en terme de rapidité de balayage et de résolution, ont permis l'émergence de l'approche « data independent acquisition » (DIA). Cette approche promet de combiner les points forts des approches « shotgun » et ciblées, mais aujourd'hui l'analyse des données DIA reste compliquée.

L'objectif de cette thèse a été de développer des méthodes innovantes de spectrométrie de masse, et en particulier d'améliorer l'analyse des données DIA. De plus, nous avons développé une approche originale Top 3-ID-DIA, permettant à la fois un profilage complet des protéines de la cellule hôte (HCP) ainsi qu'une quantification absolue d'HCP clés dans les échantillons d'anticorps monoclonaux (mAb), au sein d'une même analyse.

Cette méthode est prête à être implémentée en industrie, et pourrait fournir un support en temps réel aux développements du procédé de production de mAb, ainsi que pour évaluer la pureté des biomédicaments.

Mots clé : Spectrométrie de masse, Analyse protéomique quantitative, Data independent acquisition, Anticorps monoclonaux, Protéines de la cellule hôte

## Résumé en anglais

Recent instrumental developments in mass spectrometry, notably in terms of scan speed and resolution, allowed the emergence of “data independent acquisition” (DIA) approach. This approach promises to combine the strengths of both shotgun and targeted proteomics, but today DIA data analysis remains challenging.

The objective of my PhD was to develop innovative mass spectrometry approaches, and in particular to improve DIA data analysis. Moreover, we developed an original Top 3-ID-DIA approach, allowing both a global profiling of host cell proteins (HCP) and an absolute quantification of key HCP in monoclonal antibodies samples, within a single analysis.

This method is ready to be transferred to industry, and could provide a real time support for mAb manufacturing process development, as well as for product purity assessment.

Keywords : Mass spectrometry, Quantitative proteomics, Data independent acquisition, Monoclonal antibodies, Host cell proteins