HAL
open science

Metaproteomics analysis to study functionalities of the gut microbiota in large cohorts

Ariane Bassignani

To cite this version:

Ariane Bassignani. Metaproteomics analysis to study functionalities of the gut microbiota in large cohorts. Genomics [q-bio.GN]. Sorbonne Université, 2019. English. NNT : 2019SORUS043 . tel-01255116v2

HAL Id: tel-01255116

https://theses.hal.science/tel-01255116v2

Submitted on 18 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

SORBONNE UNIVERSITÉS

DOCTORAL THESIS

Metaproteomics analysis to study functionalities of the gut microbiota in large cohorts

Author:
Ariane BASSIGNANI

Directed by:
Dr. Catherine JUSTE
Supervised by:
Dr. Sandra PLANCADE
Dr. Magali BERLAND

Thesis jury:
President Pr. Alessandra CARBONE
Referee Dr. Jean ARMENGAUD
Referee Dr. Delphine PFLIEGER
Examiner Pr. Laurence SABATIER
Examiner Pr. Alain DENISE

A thesis submitted in fulfilment of the requirements for the degree of PhD in Bioinformatics

in the

Unités MetaGenoPolis, Micalis and MaIAGE
Institut National de la Recherche Agronomique

Publicly defended the 30th of September 2019

"Qui acquiert science s'acquiert du travail et du tourment."

Pierre Charron, *Le traité de la sagesse*

"Pour frayer un sentier nouveau, il faut être capable de s'égarer."

Jean Rostand, *Inquiétudes d'un biologiste*

"Je ne crois pas qu'il ait besoin d'être immortel. Je crois que tout ce qui lui faut, c'est écrire la bonne histoire. Parce que certaines histoires sont immortelles."

Stephen King, *Le Chant de Susannah*

SORBONNE UNIVERSITÉS

Abstract

École Doctorale Physiologie, Physiopathologie et Thérapeutique
Institut National de la Recherche Agronomique

PhD

Metaproteomics analysis to study functionalities of the gut microbiota in large cohorts

by Ariane BASSIGNANI

Metaproteomics focuses on identifying and quantifying proteins in complex biological samples such as the human gut microbiota. Currently, only few studies on human gut microbiota exceed tens of subjects. The analysis of several hundred of samples is nevertheless of interest given the growing recognition of the gut microbiota as a health partner. However, the methods and protocols used so far in proteomics and metaproteomics are not suitable for large-scale studies, whether in terms of time/memory consumption or calibration of parameters. We have therefore developed algorithms, evaluated and compared several identification approaches for peptides and proteins and proposed systematic evaluation criteria, with a particular interest in the replicability of identifications, in order to develop a pre-treatment pipeline suitable for wide-ranging studies. The systematic comparison of these approaches of identification as well as the study of the replicability bring a methodological base so far missing in the field of the metaproteomics of the human gut microbiota. Quantification of peptides and proteins by eXtracted Ion Chromatogram has never been performed on this type of data, we have also compared normalization methods and developed a methodology for imputing missing data to refine the abundance estimations obtained by the more classical method known as "Spectral Counting". This thesis work has highlighted microbial biomarkers of potential interest for predicting the response to a slimming diet, or to characterize various phenotypes of inflammatory bowel disease. We have also been able to analyse the metaproteome of more than 200 patients in the framework of the ProteoCardis ANR, which is ancillary to the European project MetaCardis, and which focuses on the potential link between gut microbiota and cardiovascular diseases. The search for proteins of interest among these data should allow us to discover protective or aggravating candidate biomarkers of cardiovascular diseases.

SORBONNE UNIVERSITÉS

Résumé

École Doctorale Physiologie, Physiopathologie et Thérapeutique
Institut National de la Recherche Agronomique

PhD

Analyse métaprotéomique pour l'étude des fonctionnalités du microbiote intestinal dans de grandes cohortes

par Ariane BASSIGNANI

La métaprotéomique s'attache à identifier et quantifier les protéines d'échantillons biologiques complexes comme le microbiote intestinal humain. Actuellement, peu d'études sur le microbiote intestinal humain dépassent quelques dizaines de sujets. L'analyse de plusieurs centaines d'échantillons revêt pourtant un intérêt évident compte tenu de la reconnaissance croissante du microbiote intestinal en tant que partenaire santé. Cependant, les méthodes et protocoles utilisés jusqu'à ce jour en protéomique et métaprotéomique ne sont pas adaptés à des études de grande ampleur, que ce soit en terme de temps/mémoire ou de calibrage des paramètres. Nous avons donc développé des algorithmes, évalué et comparé plusieurs approches d'identification des peptides et protéines et proposé des critères d'évaluation systématiques, avec un intérêt particulier porté sur la réplicabilité des identifications, afin de développer un pipeline de prétraitement adapté à des études d'envergure. La comparaison systématique de ces approches d'identification ainsi que l'étude de la réplicabilité apportent un socle méthodologique jusqu'ici manquant dans le domaine de la métaprotéomique du microbiote intestinal humain. La quantification des peptides et protéines par eXtracted Ion Chromatogram n'ayant jamais été réalisée sur ce type de données, nous avons également comparé des méthodes de normalisation et développé une méthodologie d'imputation des données manquantes permettant d'affiner les estimations d'abondances obtenues par la méthode plus classique dite « de Spectral Counting ». Ce travail de thèse a permis de mettre en évidence des biomarqueurs microbiens potentiellement d'intérêt pour prédire la réponse à un régime amaigrissant, ou encore pour caractériser différents phénotypes de maladies inflammatoires chroniques de l'intestin. Nous avons également pu analyser le métaprotéome de plus de 200 patients dans le cadre de l'ANR ProteoCardis adossée au projet européen MetaCardis, et s'intéressant au lien possible entre microbiote intestinal et maladies cardiovasculaires. La recherche de protéines d'intérêt parmi ces données devrait permettre de découvrir des candidats biomarqueurs protecteurs ou aggravants de maladies cardiovasculaires.

Remerciements

Tout d'abord, je souhaite remercier les membres de mon jury, qui ont accepté de venir écouter le fruit de mes trois dernières années de travail. Je suis particulièrement reconnaissante envers Jean Armengaud et Delphine Pflieger, qui ont la lourde tâche de relire ce mémoire. J'espère que la lecture sera agréable, et que ma présentation vous intéressera.

Merci à Catherine, qui a dirigé ma thèse avec beaucoup d'implication. Je serai toujours impressionnée par ta volonté de vouloir tout comprendre et tout essayer, et par la passion qui t'anime dans ton travail.

Merci à Sandra pour ta patience, surtout quand il s'agit de me faire comprendre des concepts mathématiques. Merci à Magali pour ton côté calmant, et pour voir toujours le bon côté des choses.

ch Merci à Olivier et Melisande, mes co-encadrants "non-officiels". Je ne comprendrais probablement encore rien à la protéomique si je n'avais pas pu apprendre grâce à vous. Cela vaut aussi pour Michel, Thierry, Filippo, et Marlène. En dehors de l'apport scientifique, je m'engage à passer plus souvent vous voir pour prendre un café et philosopher sur le sens de la vie ou le week-end de Marlène, mais pas aller à la cantine, car (i) c'est dégueulasse et (ii) Thierry ne parle que de spectrométrie de masse en mangeant. Merci aux autres membres du Moulon pour leur accueil, les discussions à la cafet', les food-truck et le Père-Noël secret : Alice, Emmanuelle, Gaëlle, Julie (petit ange parti trop tôt dans le sud), Yannick, Anthony, Harry, Dorian et Pierre. Je n'oublie pas le coin fumeur, avec Xavier, Arnaud et Adrien.

Vient le tour de MetaGenoPolis, ou MGP, ou "la famille". Parce que oui, c'est comme ça qu'on se sent là-bas. Je n'aurais pu rêver meilleurs collègues pour passer cette thèse, car dans le meilleur comme dans le pire, aller au labo a toujours été un plaisir, et vous voir tous les jours parfois un soulagement. Merci à Nicolas Maziers pour les discussions philosophiques au retour de la cantine, à Kevin Kweiszer pour m'avoir sauvé du pétrin plus d'une fois (rm -r * tmtc), Florence Thirion pour ses réflexions acérées. Merci à Florence Haimet, qui s'occupe de nous comme d'une maman et dont les compétences en botanique sont impressionnantes.

Merci à Sébastien pour avoir toujours le sourire. Merci à Léonard pour nous faire partager ses idées complètement cinglées (et bon courage pour ta thèse). Merci à Parfait pour ses anecdotes toujours délirantes; sérieux, on pourrait faire un film avec tes aventures. Merci à Océane, dont la capacité à bitcher donne toujours des pauses clopes hilarantes. Merci à Matthieu, qui a une capacité incroyable à me faire taper des fous-rires. Merci à Samar pour ta gentillesse et ton soutien moral.

Merci à Susie pour être aussi drôle, j'espère te revoir à La Réunion, ou quand tu reviendras sur Paris. Merci à Victoria pour tes conseils et ton franc-parler. Merci

à Laurie pour être une personne sincère, honnête et drôle (même si je ne sais pas comment tu fais pour dormir si peu). Merci à Anne-Sophie pour ta gentillesse, et pour ne pas nous avoir oublié après être partie du labo, ça me fait toujours autant plaisir de te voir pendant les pique-nique du soir. Merci à Hugo pour être aussi passionné par ton travail (mais arrête d'en parler quand tu as trop bu).

Merci à Alexandre pour les pauses-clopes et les soirées. Merci à Fatou pour être aussi épuisée par le fait de vivre, ça me fait toujours autant rire. Merci à Florian pour être une personne avec qui on peut avoir des discussions extrêmement sérieuse, et extrêmement délirante. Je compte sur toi pour la création future du poulailler de MGP. Merci à Franck pour tes enseignements en bash, pour le badminton avec tout ce qui nous passe sous la main, pour evotar et pour tous nos petits délires. Merci à Hanna pour être un sage chinois un peu dingo. Merci à Nicolas Pons pour ton aide dans le travail, les BBQ et le week-end MGP, mais aussi parce que tu es une bonne personne. Mais arrête de travailler autant.

Merci à Magali (encore), pour ta patience dans le travail, et surtout de toujours avoir cru en moi, même quand je n'y croyais pas moi-même. Pour ton soutien moral, pour ta gentillesse, pour avoir toujours été là quand ça n'allait pas au cours de ces trois ans et demi. Et parce que tu peux passer 3 semaines en vacances avec moi sans s'engueuler et en se marrant. Merci à Marie, sans qui j'aurais probablement fait une dépression depuis bien longtemps. Je ne vais pas m'étendre là-dessus, je crois que tout le monde sait à quel point nous sommes proches, et j'espère que ça le sera toujours. Je crois que ne plus travailler à côté de toi va être une putain d'épreuve. Mais je compte sur toi pour me skyper de l'Amérique du Sud pour me raconter tes aventures. Te voir était ma principale source de motivation depuis 3 ans pour me lever le matin et aller au travail.

Bref, à tous ceux de MGP: Merci d'être qui vous êtes.

Merci ensuite, à tous mes amis hors-MGP (oui, il y en a). Kevin Da Silva déjà (parce que t'es un peu MGP aussi), parce qu'on est tellement semblables malgré les apparences. Merci Cécile pour ton soutien moral, et pour être fidèle à toi-même. Merci à Hugues et Maxime, pour les petits restos tranquillou bilou. Merci à Maurice pour avoir été présent quand j'en ai eu besoin, et pour ta capacité à te plaindre de tout. TMTC, c'est la famille. Merci à Quentin, qui est prêt à me ramener un futon au labo à 22h30 pour que je passe pas la nuit allongée sur ma serviette de sport. Merci à Thomas pour avoir été là quand j'avais besoin de parler, et d'être là depuis plus de dix ans maintenant. Et pour être là, toujours. Merci à mes copains de fac, Alexandre, Steepan, Arnault, Mathilde, Tiphaine et Guillaume, pour les sortie bar rares mais qualitatifs. Désolée d'être arrivée en retard au mariage d'Alexandre, mais au moins comme ça Steepan a pu voir les nichons de Tiphaine quand elle se changeait dans la voiture, et ça, ça n'a pas de prix. Enfin, merci à Caro, Caro et Wendy. Toutes les trois, si vous n'aviez pas été là, je ne crois pas que j'y serais arrivée. Quand le bateau

tanguait pendant la tempête, vous m'avez donné la force de m'accrocher au mât.

Merci à ma famille de m'avoir soutenu ces dernières années, même si vous ne savez toujours pas prononcer "bioinformatique" et que vous n'avez aucune idée de ce que je fais dans la vie.

Merci à Misty. Oui, ça peut sembler stupide de remercier un chien qui est intelligent mais pas au point de lire cette thèse, mais cette petite chose a été la seule chose stable ces trois dernières années. J'aurais probablement perdu ma santé mentale, si je n'avais pas dû faire d'efforts pour rester à flôt et m'occuper d'elle.

Enfin, merci à Alexandre. Avec toi, je peux être qui je suis, et tu acceptes toutes mes facettes. Et aussi ma chienne, même si ce n'était pas gagné au début. Je ne sais pas ce que me réserve mon futur, mais quand j'angoisse, je me dis que tant que tu es à mes côtés, ce sera forcément heureux.

Contents

Abstract iii

Résumé v

Remerciements vii

Synopsis xxiii

1 The gut microbiota: a challenging complexity 1

1.1 The gut microbiota in humans . . . 1
1.2 Metagenomics . . . 3
1.2.1 Different approaches . . . 3
1.2.1.1 Targeted metagenomics . . . 3
1.2.1.2 Shotgun metagenomics . . . 5
1.2.2 Sequencing technologies . . . 6
1.2.2.1 Illumina technology . . . 6
1.2.2.2 Ion Torrent^{TM} technology . . . 9
1.2.3 Assembly . . . 11
1.3 Metaproteomics . . . 14
1.3.1 Metaproteomics experimental workflow . . . 14
1.3.2 LC-MS/MS analyses . . . 15
1.3.3 Interpretation of LC-MS/MS data . . . 17
1.3.3.1 Peptide-spectrum matching . . . 17
1.3.3.2 Protein identification and grouping . . . 19
1.3.3.3 Spectral Counting quantification . . . 20
1.3.3.4 eXtracted Ion Chromatogram quantification . . . 23
1.3.4 The rise of metaproteomics in the last decade . . . 24
1.3.5 The challenges of metaproteomics . . . 25
1.3.5.1 Sample preparation . . . 25
1.3.5.2 Bioinformatics analyses . . . 26
1.4 Cardiovascular artery diseases . . . 28
1.5 Evidence for a relationship between gut microbiota and CAD . . . 30
1.6 The ProteoCardis project . . . 31
1.7 Thesis objectives . . . 32

2 Mass spectra interpretation without individual metagenomes **35**
2.1 The ObOmics study . . . 35
2.2 Scientific questions . . . 36
2.3 Methods . . . 36
2.3.1 Samples preparation and injection . . . 36
2.3.2 Interrogated databases . . . 37
2.3.3 Interrogation strategies . . . 38
2.3.3.1 Classical identification . . . 38
2.3.3.2 Iterative identification in two steps . . . 38
2.3.3.3 Iterative identification in three steps . . . 39
2.3.3.4 Iterative identification used in the experiments . . . 40
2.3.4 Construction of the datasets . . . 41
2.3.5 Peptide and subgroup quantification . . . 42
2.3.6 Evaluation criteria . . . 42
2.4 Results . . . 44
2.4.1 Gain of identification with MetaHIT 9.9 . . . 44
2.4.2 Identifications specific to each database . . . 45
2.4.3 Reproducibility of the identifications with MetaHIT 3.3 and MetaHIT 9.9 . . . 47
2.4.4 Gain of identification with the iterative strategy . . . 50
2.4.5 Identifications specific to each interrogation strategy . . . 51
2.4.6 Reproducibility of the identifications with the classical and the iterative strategy . . . 52
2.5 Conclusion . . . 54

3 Two examples of clinical data interpretation with MetaHIT databases **57**
3.1 Metaproteomic features related to weight loss . . . 57
3.1.1 Scientific context . . . 57
3.1.2 Methods . . . 57
3.1.3 Results . . . 58
3.1.4 Conclusion . . . 59
3.2 Metaproteomic features related to intestinal bowel diseases . . . 60
3.2.1 Scientific context . . . 60
3.2.2 Methods . . . 61
3.2.3 Results . . . 61
3.2.3.1 Metaproteomic profiling of stool samples . . . 61
3.2.3.2 Search for IBD signatures in stool samples . . . 65
3.2.3.3 Search for signatures between IBD phenotypes . . . 67
3.2.4 Conclusion . . . 69

4 Mass spectra interpretation in the context of the ProteoCardis cohort **71**
4.1 Methods . . . 71
4.1.1 Metagenomics sequencing . . . 71

4.1.2 Metaproteomic analyses . . . 72
4.2 Development of MetaRaptor . . . 73
4.3 Assembly of the individual metagenomes of the ProteoCardis cohort . 77
4.4 Performance of the individual catalogues . . . 78
4.5 Metaproteome landscape in cardiovascular diseases . . . 82

5 XIC quantification 93
5.1 Challenges of XIC quantification in metaproteomics . . . 93
5.2 Chromatographic alignment . . . 94
5.3 Correction of XIC . . . 97
5.4 Imputation of missing data . . . 102
5.4.1 Imputation of missing values in classic proteomics . . . 102
5.4.2 Imputation of missing values in metaproteomics . . . 103
5.4.3 Imputation implemented in the ProteoCardis study . . . 103

6 Can the ProteoCardis data be improved by normalization methods? 109
6.1 Ascertainment of the batch effect . . . 109
6.2 Normalization methods . . . 111
6.2.1 Methods for SC normalization . . . 111
6.2.2 Methods for XIC normalization . . . 112
6.3 Evaluation of the normalizations . . . 113
6.3.1 Normalization of SC . . . 113
6.3.2 Normalization of XIC . . . 117
6.4 Conclusion on the correction of technical variability . . . 119

7 Exploration of statistical approaches for biomarker discovery 121
7.1 Methods . . . 121
7.1.1 Multiple testing approach . . . 121
7.1.1.1 Resampled FDR . . . 121
7.1.1.2 Modelling of SC . . . 122
7.1.1.3 Modelling of XIC . . . 124
7.1.2 Random forests approach . . . 124
7.1.2.1 Principle . . . 124
7.1.2.2 Typical preprocessings . . . 126
7.1.2.3 Implemented parameters and validation scheme . . . 126
7.2 Results . . . 127
7.2.1 Preliminary: evaluation of the batch effect . . . 127
7.2.2 Results with multiple testing . . . 128
7.2.3 Results with random forests . . . 131
7.2.4 Relationship between the two approaches . . . 133
7.3 Perspectives on statistical analysis . . . 134

8 Conclusion and perspectives 135

A Polymerase Chain Reaction **139**

B Physiopathology of atherosclerosis **141**

C Methods for the ObOmics study **145**
C.1 Stool Sample Collection and Processing 145
C.2 Protein Digestion and Peptide Desalting 145
C.3 LC-MS/MS Analysis 146
C.4 LC-MS/MS interpretation 146
C.5 Taxonomic and functional annotation 147

D Methods for the MICI-Pep study **149**
D.1 Volunteers and Sample collection 149
D.2 Preparation of microbiota 149
D.3 Metaproteomic analyses 150
D.4 LC-MS/MS analyses 150
D.5 Search for contrasts 151
D.6 Taxonomic and functional annotation 151

E Parameter validation of assembly on a mock community **153**

F Scientific contributions **157**
F.1 Poster communications 157
F.2 Oral communications 158

Bibliography **159**

List of Figures

1.1 Inventory of scientific papers recorded each year in Pubmed for gut metagenomics and gut metaproteomics. 3
1.2 Variability of the regions of the 16S gene of *Pseudomonas*. 5
1.3 Bridge amplification and Illumina sequencing. 8
1.4 Library and sequencing matrix preparation for Ion Torrent technology. 10
1.5 Incorporation of a dNTP to DNA with DNA polymerase at the 3′ end. 10
1.6 Sequencing with Ion Torrent technology. 11
1.7 Genome reconstruction from reads with a De Bruijn graph. 13
1.8 Gasification of samples in liquid phase by electrospray at the input of the mass spectrometer. 15
1.9 Scheme of the Orbitrap Fusion^{TM} Lumos^{TM} Tribrid^{TM} mass spectrometer. 17
1.10 Example of grouping of 6 proteins. 20
1.11 Example of quantification of four peptiz by SC and XIC. 23
1.12 Description of the ProteoCardis datasets. 32

2.1 Preparation and injection of the samples. 37
2.2 Classical strategy of interrogation. 38
2.3 Two-steps iterative strategy of interrogation. 39
2.4 Three-steps iterative strategy of interrogation. 40
2.5 Iterative strategy of interrogation implemented in our work. 41
2.6 Total identifications with two databases. 44
2.7 Individual identifications with two databases and two strategies. . . . 45
2.8 Number of peptides identified by MetaHIT 3.3 and MetaHIT 9.9. . . . 46
2.9 Number of peptides identified in each sample specifically by one of the two MetaHIT databases. 46
2.10 Fraction of common proteins identified in each pair of replicates with MetaHIT 3.3 and MetaHIT 9.9 47
2.11 Fraction of common peptides identified in each pair of replicates with MetaHIT 3.3 and MetaHIT 9.9 48
2.12 Log_2 of specific spectral counts per protein, as a function of the number of replicates in which they were identified. 49
2.13 Probability to get a zero abundance for a protein in a replicate, as a function of its abundance observed in another replicate. 49

2.14 Proteins diversity of the ObOmics samples with MetaHIT 3.3 and MetaHIT 9.9. 50
2.15 Peptide and protein diversity as a function of the number of biological samples with the classical and iterative interrogation strategies. 51
2.16 Number of peptides and proteins identified by the classical workflow and the iterative workflow. 51
2.17 Number of peptides and proteins identified by one of the strategies only. 52
2.18 Fraction of common proteins identified in each pair of replicates, with the classical and the iterative strategies. 53
2.19 Fraction of common peptides identified in each pair of replicates, with the classical and the iterative strategies. 53
2.20 Proteins diversity. 54

3.1 Heatmap of proteins positively or negatively correlated with weight loss at a timepoint of the study. 58
3.2 Taxonomic distribution of the 27 proteins whose correlation with weight loss was accentuated over time. 59
3.3 Pearson's correlation between metaproteomic profiles of fresh and frozen samples. 62
3.4 Clustering of samples based on their subgroup counting. 63
3.5 Taxonomic distribution of the microbiome of the MICI-Pep patients. . 65
3.6 Proteins overabundant in IBD patients compared to control. 66
3.7 Heatmap of 101 subgroups overabundant in CDIC samples compared to all other IBD samples. 67
3.8 Proteins overabundant in all CDC compared to all UC samples. 68
3.9 Tree of decision for diagnostic of IBD based on our metaproteomic analysis of microbiota envelope-enriched fractions. 69

4.1 Schema of the Metaraptor's steps. 76
4.2 Distribution of the number of genes for the ProteoCardis stool samples. 77
4.3 Identifications with MetaHIT 9.9 and individual catalogues. 78
4.4 Venn diagram of the number of peptides identified by MetaHIT 9.9 and the individual catalogues. 79
4.5 Length of the proteins identified with the MetaHIT 9.9 catalogue and the individual catalogues. 80
4.6 Number of peptides per subgroup for subgroups identified with MetaHIT 9.9 and the individual catalogues. 81
4.7 Metabolic pathways identified in each fraction of the ProteoCardis samples. 84
4.8 Venn diagrams of the number of peptides and proteins identified in fractions of the ProteoCardis samples. 84
4.9 Clustering of the cytosolic fractions of the non-bariatric samples (n=188) 85

4.10 Clustering of the envelope fractions of the non-bariatric samples (n=188) 86
4.11 Taxonomic distribution of subgroups in the cytosolic fractions of the non-bariatric samples. 87
4.12 Taxonomic distribution of subgroups in the envelope-enriched fractions of the non-bariatric samples. 88
4.13 Protein diversity, defined as the mean number of subgroups identified, observed with an increasing number of samples 90
4.14 Taxonomic-functional diversity, defined as the mean number of the combination KEGG-species terms, identified with an increasing number of samples 91

5.1 Ratio of abundances between two chromatogram alignments. 96
5.2 Quantification of peptiz of a test sample aligned with two references. . 96
5.3 Intensity profiles of the samples along the chromatographic retention time. 98
5.4 Standard deviation of the retention times for the non-bariatric patients, cytosolic fraction. 99
5.5 Distance ratio between replicated and non-replicated samples after filtering XIC data. 100
5.6 Percentage of missing values after the XIC filtering. 101
5.7 Number of peptides and proteins after the XIC filtering. 101
5.8 Relationship between the mean of quantification of peptides over 12 to 14 standard samples and its standard deviation. 104
5.9 Linear modelisation on XIC values. 104
5.10 Linear or nonparametric modelling of the relationship between mean and standard deviation of the quantification of peptides with low XIC mean values. 105
5.11 Interpolation of sd by splines considering the number of knots for the splines modelling for non-bariatric and bariatric, cytosolic and envelope fractions of ProteoCardis samples. 106

6.1 Sum of SC of the proteins, before any normalization (raw data). 110
6.2 Global intensity profiles of XIC before normalization 110
6.3 Total Ion Current of the ProteoCardis samples, patient non-bariatric, cytosolic fraction. 112
6.4 Correlation between TIC and sum of SC of each sample. 112
6.5 Principal Component Analysis of raw Spectral Counts of proteins. . . 114
6.6 Principal Component Analysis of Spectral Counts of proteins after the normalization of batches by the sum of SC of the standards, the MS2 of the standards and linear regression 115
6.7 Principal Component Analysis of Spectral Counts of proteins after the normalization of total SC abundance by TIC and TMM. 116
6.8 Sum of SC of the proteins, after TIC normalization. 116

6.9 Ratio of the distances between replicated and non-replicated samples, before normalization and after each normalization evaluated. 117
6.10 Global intensity profiles of XIC after normalization by percentage, median and median-RT. 118

7.1 Difference of AIC and BIC between the negative binomial model and the zero-inflated negative binomial model. 123
7.2 Empirical SC of three groups containing 89%, 41% and 12% of values greater than 0, and their estimation by zero-inflated negative binomial model. 123
7.3 Distribution of XIC abundances for the observed values or the imputed values. 124
7.4 Resampled p-values of the 740 protein groups in a model without or with batch effect when the effect of patient group was tested. 128
7.5 P-values of SC and XIC for the subgroups, cytosolic fraction, and for the comparison between aCAD and controls. 131
7.6 Out-Of-Bag accuracy for the classification of aCAD and controls for each threshold of occurrence of important variables. 132
7.7 Abundance of the 16 groups that better classify the aCAD and control groups. 133
7.8 Results on spectral counting of groups with random forest and multiple testing 133

A.1 The different steps of a PCR. 139

B.1 Healthy artery layers 141
B.2 Stages of the development of atherosclerosis. 144

E.1 Coverage of the reference databases of the mock community by contigs and genes assembled with MetaRaptor. 154
E.2 Relative abundance of each genus present in the even HMP mock community. 155

List of Tables

1.1 How to count MS2 spectra at different levels. 22

3.1 Proteins diversity with functional and taxonomic annotation in the MICI-Pep samples 64

4.1 Results of iterative interrogation of MetaHIT 9.9 in the ProteoCardis cohort. 83
4.2 Taxonomic distribution of proteins in the cytosolic fractions of the ProteoCardis samples 89
4.3 Taxonomic distribution of proteins in the envelope-enriched fractions of the ProteoCardis samples 89

5.1 Summary of XIC data. 97
5.2 Retention time thresholds retained for the 4 ProteoCardis datasets. . . 98
5.3 Values computed for the imputation of XIC missing values. 106

7.1 Number of variable detected for each dataset and each test, for three levels of FDR, based on SC. 129
7.2 Number of variable detected for each dataset and each test, for three levels of FDR, based on XIC. 130
7.3 Accuracy of the forests for different preprocessings for the classification of aCAD and controls. 132

B.1 The six stages of atherosclerosis and age of occurrence 143

List of Abbreviations

aCAD	**a**cute **CA**rdiovascular **D**isease
BS	**B**ariatric **S**urgery
CAD	**CA**ardiovascular **D**isease
cCAD	**c**hronic **CA**rdiovascular **D**isease
cCADic	**c**hronic **CA**rdiovascular **D**isease with **c**ardiac **i**nsufficiency
cCADic ncor	heart failure unrelated to CAD
CD	**C**rohn's **D**isease
CDC	**C**olic **C**rohn's **D**isease
CDIC	**I**leo-**C**olic **C**rohn's **D**isease
FDR	**F**alse **D**iscovery **R**ate
HPLC	**H**igh-**P**erformance **L**iquid **C**hromatography
IBD	**I**nflammatory **B**owel **D**isease
LC-MS/MS	**L**iquid **C**hromatography coupled to tandem **M**ass **S**pectrometry
m/z	**M**ass/**C**harge ratio
NA	**N**ot **A**vailable (missing values)
NB	**N**egative **B**inomial
PCR	**P**olymerase **C**hain **R**eaction
PSM	**P**eptide-**S**pectrum **M**atch
SC	**S**pectral **C**ounting
SOP	**S**tandard **O**perational **P**rocedure
SRM	**S**elected **R**eaction **M**onitoring
UC	**U**lcerative **C**olitis
XIC	e**X**tracted **I**on **C**hromatogram
ZINB	**Z**ero-**I**nflated **N**egative **B**inomial

Synopsis

Metaproteomics is the analysis of all the proteins present in an ecosystem. Initiated a bit more than ten years ago, metaproteomics goes beyond the genetic potential of metagenomics, gives an overview of the genomic expression of the microorganisms that make up an ecosystem, and thus detects and quantifies their real activity. Even with so much interest, progress in metaproteomics has been slowed by technological bottlenecks, and therefore still remains an emerging field. With new developments in mass spectrometry, and increasing coverage of metagenomes, essential for mass spectra interpretation, it now becomes possible to analyse metaproteome of ecosystems as complex as as the human intestinal microbiota. Like any emerging field, implementation of metaproteomic analysis protocols, whether in terms of sample preparation, analysis by mass spectrometry, or bioinformatic analyses for data processing, are developing and diversifying rapidly.

The aim of the research work developed in this thesis, entitled "Metaproteomics analysis to study functionalities of the gut microbiota in large cohorts", is to make a contribution to this still challenging field, with a focus on bioinformatic and analysis of such complex data. The manuscript is organised in eight chapters.

Chapter 1 introduces the fundamental principles of the main tools to study the human intestinal microbiota, as well as the accompanying challenges. It also presents the context of the thesis and the rational of the project on which I mainly worked.

Chapter 2 examines the methodologies used to identify peptides and proteins on a set of 48 human intestinal metaproteomes. I compared several identification methodologies and evaluated their performance from the point of view of the number of peptide and protein identifications, but also of their reproducibility, an aspect that is rarely addressed in the context of metaproteomics. These preliminary results made it possible to define an optimal identification methodology that will be used in our following analyses.

Chapter 3 presents two applications of metaproteomic analysis of the human intestinal microbiota. These studies focus on weight loss during a low-calorie diet, and on inflammatory diseases of the digestive tract. The results obtained in the previous chapter allowed us, with simple analyses, to identify proteins of potential interest for the prediction of weight loss and the diagnosis of inflammatory diseases of the digestive tract.

Chapter 4 focuses on the identification of peptides and proteins in the large-scale cohort on which I mainly worked during my thesis. I evaluated interests and limitations of using individual metagenomes for mass spectra interpretation. The results of identification on a large-scale cohort allowed us to have a first overview of the patients' metaproteomes.

Chapter 5 evaluates the feasibility of quantifying peptides by eXtracted Ion Chromatogram (XIC) over a large-scale cohort. In this chapter, I took special care to XIC

data cleaning, and developed a methodology for imputing missing data adapted to such large datasets.

Chapter 6 proposes and evaluates corrections of the count and XIC data to reduce technical variability observed on these data, and thus to increase robustness of the statistical analyses used in the next chapter.

Chapter 7 presents the mathematical concepts of the approaches used to discover biomarkers (peptides and/or proteins) of interest in large-scale cohorts. First statistical results are presented and ways for further analyses are considered.

Chapter 8 finally draws the general conclusions and perspectives of my thesis work.

Chapter 1

The gut microbiota: a challenging complexity

1.1 The gut microbiota in humans

The gut microbiota is the microbial ecosystem of the digestive tract. This ecosystem consists of about a hundred thousand billions of commensal bacteria, viruses, archaea and fungi, and is estimated to weight between 200g and up to 2kg [1]. More than a thousand different species constitute the human intestinal microbiota, and each individual's microbiota is made up of about 500 of these species. There is a great diversity of species between individuals, diversity being defined by the number and abundance of the different species present in the microbiota of each individual. Several metagenomic studies converge to show that the most abundant phyla in stool samples are Bacteroidetes and Firmicutes, representing about 60% of the total bacteria, and Actinobacteria and Proteobacteria representing about 10% of the total bacteria [2, 3]. The remaining part of the bacteria phyla are more variable between individuals.

At the genus level, human microbiota can be clustered in three enterotypes, regardless the geographic origin of the subjects. These enterotypes, mainly driven by the genera *Bacteroides*, *Prevotella* and *Ruminococcus*, define distinct microbial communities of the human gut microbiota [3].

The gut microbiota can also be clustered based on its richness into "High Gene Count" (HGC) and "Low Gene Count" (LGC), presenting specificities in term of taxonomic and functional composition. At the taxonomic levels, HGC communities are enriched in Verrucomicrobia and Actinobacteria at the phylum level, in *Faecalibacterium* and *Bifidobacterium* at the genus level, and in *Faecalibacterium prausnitzii* at the specie level. Conversely, LGC communities are enriched in *Proteobacteria* and *Bacteroidetes* at the phylum level, in *Bacteroides* and *Parabacteroides* at the genus level, and *Ruminococcus gnavus* at the specie level. Clinical data suggest that individuals with LGC gut microbiota are more exposed to metabolic disturbances known to bring multiple cardiovascular disease risk factors (increased insulin resistance, free fatty acids and percentage of fat mass, and decreased High Density Lipoprotein - HDL -) [4].

The taxonomic diversity corresponds to a considerable functional diversity, functions carried by the genes. The core functionalities of the gut microbiota, like carbohydrate metabolism and aromatic amino acid metabolism, remain stable between individuals [2], but the enormous functional diversity and specificity of human gut microbiota is revealed by the sequencing of their metagenomes. Indeed, even with more than a thousand of people, new individual-specific genes are still discovered; the total number of genes increases with the number of sequenced metagenomes, and never reaches a plateau. Only few genes are shared by several people, most of them being individual-specific [5]. This diversity will influence relations between the members of the ecosystem, but also the microbiota-host-food relations, and ultimately the health of the host. The study of the gut microbiota is therefore of interest in clinical conditions, but due to its enormous complexity, this study is challenging.

However, over the past twenty years, technological progress in the fields of sequencing and bioinformatics now enable to finely study the gut microbiota. Metagenomics played a pioneering role in this domain, first for the molecular characterization of taxonomies through 16S metagenomics, then for the deep characterization of the genetic potential of gut microbial communities through shotgun metagenomics with the introduction of next-generation sequencing (developed in Section 1.2.2). However, it was not until the 2010's, when mass spectrometry has really taken off, that metaproteomics, which is defined as the large-scale profiling of the protein complement of the metagenome, entered the field through label-free shotgun approach (developed in Section 1.3). Even though the obvious interest of intestinal metaproteomics is well recognized, and has been tightly argued in a number of interesting reviews [6–11], this scientific field seems to be having some trouble emerging. This is illustrated by Figure 1.1, which compares inventories of intestinal metagenomics and metaproteomics studies since 1997 to present. The lag between metaproteomics and metagenomics development can be explained by the relatively recent evolution in Liquid Chromatography coupled to tandem Mass Spectrometry (LC-MS/MS) technology, the lack of sufficiently representative search databases, but also by the multiplicity and the complexity of steps in a shotgun metaproteomics experiment. All these steps are further explained in Section 1.3.

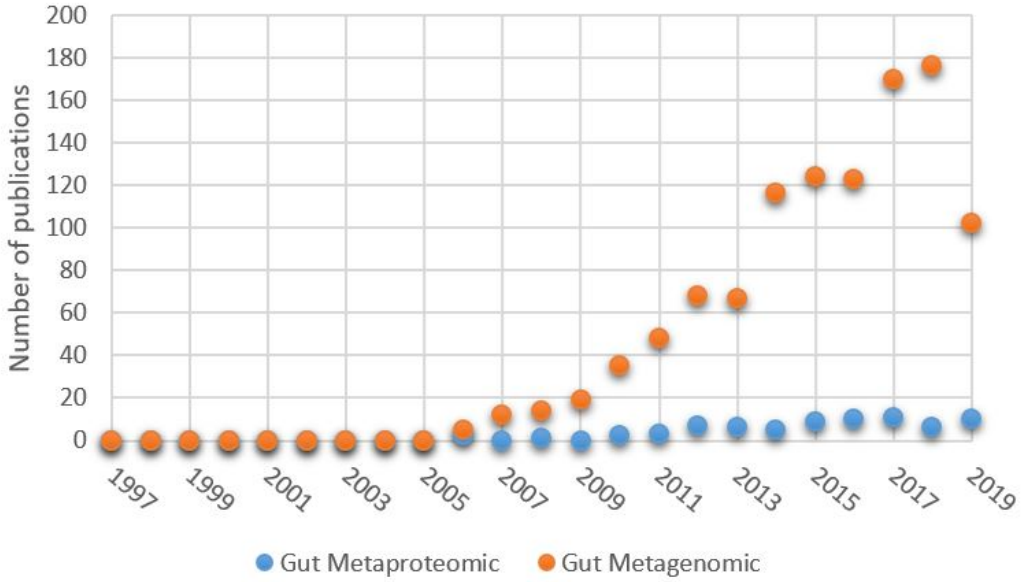


FIGURE 1.1 – Inventory of scientific papers recorded each year in Pubmed containing the terms: ["metagenome" OR "metagenomic"] AND ["gut" or "intestine" OR "intestinal"] IN ["title" or "abstract"]. The same researched was performed with ["metaproteome" OR "metaproteomic"].

Recent developments of tools meant to manage high-throughput data enables the in-depth study of the gut microbiota communities, and in particular their functional activity, with different -omics approaches. We present in the following chapter the technical principles of -omics tools - metagenomics and metaproteomics - that we used for deciphering metaproteomes of the cohorts studied in this thesis work.

1.2 Metagenomics

1.2.1 Different approaches

Metagenomics is the most widely used method for analysis of a microbial community - or microbiome -, as high-throughput sequencing technologies have seen an unprecedented development in recent years. Next-generation sequencing technologies can now achieve high-throughput and accurately study ecosystems as complex as the intestinal microbiota. Unlike genomics, which focuses on studying DNA from single organisms, the purpose of metagenomics is to quantify the DNA of a multitude of species in a particular ecosystem at once. Two sequencing approaches are used today: targeted metagenomics and shotgun metagenomics.

1.2.1.1 Targeted metagenomics

Targeted metagenomics consists in amplifying and sequencing a single gene present in several species of the environment under study. The candidate gene, while shared

by several species of interest, must be variable enough to be able to discriminate the species that carry it. The ribosome is a complex composed of proteins and RNA that translates messenger RNAs into proteins. This structure, extremely conserved during evolution, is composed of two subunits and is present in eukaryotes and prokaryotes:

- In eukaryotes, the large subunit is composed of 5S, 28S and 5.8S ribosomal RNAs and 49 ribosomal proteins. The small subunit is composed of 18S ribosomal RNA and 33 ribosomal proteins.
- In prokaryotes, the large subunit is composed of 5S and 23S ribosomal RNAs and 34 ribosomal proteins. The small subunit is composed of 16S ribosomal RNA and 21 ribosomal proteins.

Preservation during the evolution of the gene coding for 16S ribosomal RNA (rRNA) makes it a target of choice for primers used in molecular biology and targeted metagenomics. Indeed, being shared by bacteria and archaea, it allows targeting these two domains while excluding eukaryotes and thus contamination by human DNA present in the stool samples. However, targeted metagenomics excludes the study of fungi (from the eukaryotic domain) and viruses belonging to the gut microbiota ecosystem. Although highly conserved, 16S rRNA gene contains hypervariable regions (regions V1 to V9, figure 1.2), which discriminates species based on their sequence. The first targeted metagenomic studies showed that most of the existing bacterial species in various environments (freshwater, seawater, soil, ...) had not been identified before because they were not cultivable by standard methods in laboratories [12, 13]. This observation is even more valid for the gut microbiota. Indeed, their anaerobic nature, their interactions with other species of the community and the lack of knowledge about their optimal culture conditions make them hard to cultivate and amplify by conventional Polymerase Chain Reaction (PCR) methods (the procedure of PCR is further explained in Appendix A), although new methods emerge [14]. The 16S rRNA gene is therefore particularly useful for the study of this ecosystem.

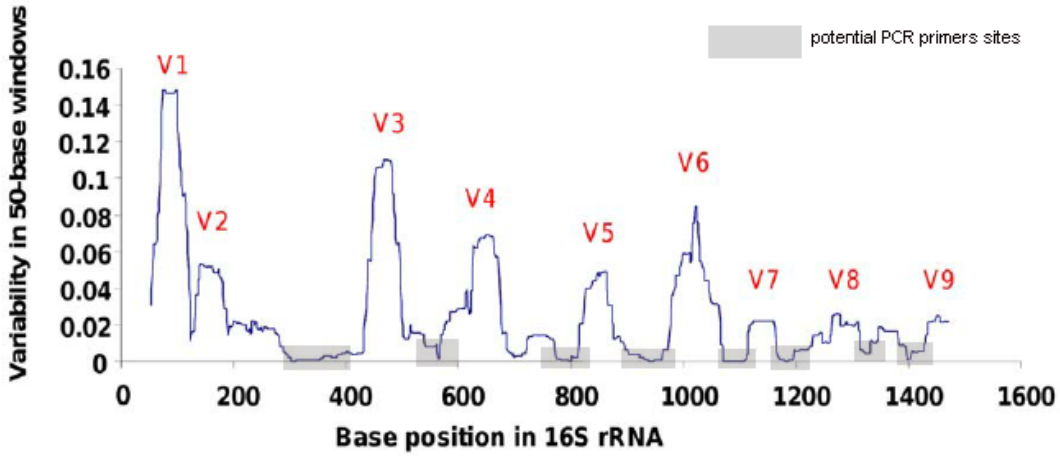


FIGURE 1.2 – Variability of the regions of the 16S gene of *Pseudomonas*. V1 to V9 are hypervariable regions, used in taxonomy annotation. The highly conserved regions can be used as primers sites in PCR. Adapted from Bodilis et al. [15] in accordance to the licence CC BY 4.0.

However, within some genera or families, 16S rRNA gene sequencing does not allow for species determination because the genes' sequences have a very high similarity. This is the case, for example, within the families Enterobacteriaceae and Peptostreptococcaceae, whose constituent species have 16S rRNA gene sequences with a similarity greater than 97%. It has been shown that for 47% of bacterial genera, there are specie level assignment difficulties using 16S rRNA gene sequencing alone [16]. To identify microorganisms at a finer taxonomical rank, shotgun sequencing methods have been developed.

1.2.1.2 Shotgun metagenomics

Shotgun metagenomics consists in sequencing all the DNA from all organisms in the sample. Since it is not limited to a particular organism, this sequencing method makes it possible to capture bacterial and archaeal genomes (in the same way as targeted metagenomics) but also host, fungi and viruses. It thus allows to have a finer overview of the microorganisms composing the ecosystem.

Moreover, since no part of the genome is preferentially sequenced, this method enables to identify new genomes that had never been observed. Indeed, in the targeted metagenomics method, species identification is based on 16S rRNA genes already identified in databases since it uses primers based on the already known 16S rRNA gene sequences. Although the most abundant organisms are the most represented in shotgun metagenomics results, the random nature of shotgun sequencing ensures that low-abundant organisms of the gut microbiota are still represented.

Lastly, shotgun metagenomics gives direct information on the potential functions encoded by the metagenome. This makes it possible to identify metabolic pathways potentially used by the gut microorganisms.

However shotgun metagenomics requires a much larger sequencing depth (number of reads) in order to capture the genomes of low-abundant microorganisms. Reads are fragments of sequenced DNA; their generation is fully developed in Section 1.2.2.2. While targeted metagenomics requires 50 000 - 100 000 reads to identify bacterial in a sample, shotgun metagenomics requires several millions of reads. It results in much higher sequencing costs, as well as much heavier downstream bioinformatic processing. In addition, genomes' coverage variability may be a barrier to taxonomic assignment, which can be done at the species level only if the genome coverage is sufficient. Nevertheless, shotgun metagenomics is a method offering the greatest potential for identification of bacterial species and their functional potentials [16].

In the context of this thesis work, our aim was to study the metaproteome functionalities of the gut microbiota, and reference metagenomic databases are a prerequisite for the study of the whole protein composition of a microbial community (Section 1.3). We therefore used shotgun metagenomics for the sequencing of the metagenomes of the patients in the ProteoCardis cohort, a project developed in Section 1.6. Several sequencing tools can be used to sequence the metagenome, and they rely on different technologies and measurement methods. We will present here the two main sequencing approaches: Illumina and Ion Torrent.

1.2.2 Sequencing technologies

1.2.2.1 Illumina technology

The Illumina sequencing technology has the particularity of amplifying extracted DNA through a bridge technique, and sequencing it through the detection of photons during the polymerisation.

First, the double-stranded DNA is fragmented into pieces of about 200 kilobases (kb) by transposomes, which also allow for the attachment of primers to the ends. Then short amplification cycles allow for the attachment of the adapters with two distinct oligonucleotides sequences. Amplification is performed by PCR.

Both types of adapters are attached by covalent bind to a flowcell. Then, DNA strands (denatured to become single-stranded) randomly hybridise to the adapters by complementarity. The reverse strand is synthesized through a polymerase, including the adapter located at the other end of the DNA. The denatured complementarity DNA is bound to the flowcell and bridges are created between the adapters attached to the flowcell. These bridges are then amplified, which creates clusters with high density amplified sequences. This type of amplification is called "bridge amplification".

After amplification, the reverse strands are cleaved from their adapter, leaving on the flowcell only the forward strands that are sequenced. The four types of nucleotides are added to the flowcell, marked with different colours. Incorporation

of the complementary nucleotide into the sequenced strand is determined by the colour emitted by the cluster after excitation by a laser. Incorporation of the four nucleotides at each cycle and detection of the emitted colour allows for strand sequence computation. Nucleotide incorporated at each position is detected in parallel on all the clusters of the flowcell, which allows an extremely fast sequencing (Figure 1.3).

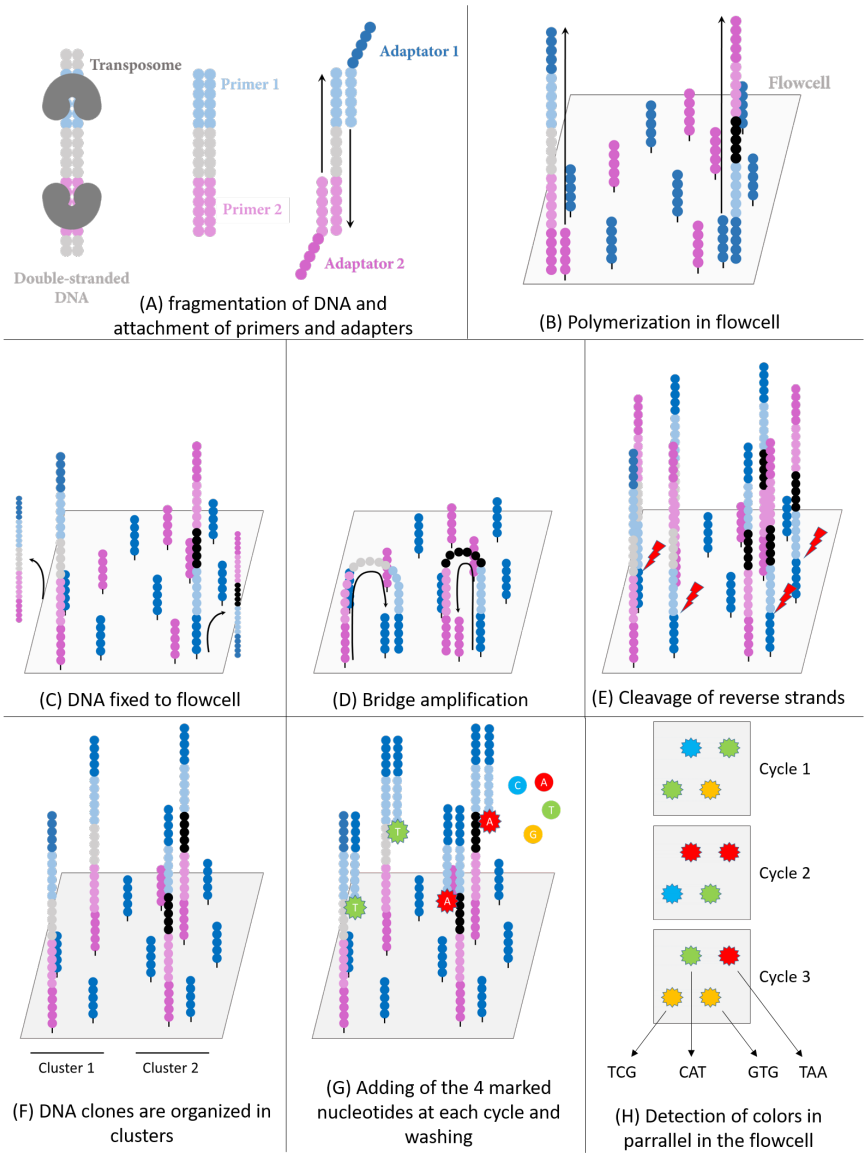


FIGURE 1.3 – Bridge amplification and Illumina sequencing. (A) DNA is fragmented into 200kb fragments and primers are attached by transposomes. The amplification adds two types of adaptors at each end. (B) The DNA fragments are floated onto a flowcell and elongated by DNA polymerase. (C) The unattached strands are washed by denaturation. (D) The strands form bridges at the surface of the flowcell and are amplified by cycles of polymerisation/denaturation. (E) The antisense strands are cut and washed, leaving only sense strands. (F) The sense strands form clusters of the same sequence of DNA, here a cluster of grey DNA and a cluster of black DNA. (G) At each cycle, the four types of nucleotides are introduced into the flowcell and are incorporated by polymerisation. After washing, a laser excites the last nucleotide incorporated, which emits a distinctive colour. (H) At each cycle, the clusters are sequenced in parallel thanks to the colour they emit. The succession of the colours at each cycle determines the sequence of the DNA of each cluster.

1.2.2.2 Ion Torrent^{TM} technology

DNA fragmentation can be done by physical methods (acoustic shearing or sonication) or enzymatic methods (non-specific endonuclease cocktails). DNA fragments are then selected according to their size. The desired fragment sizes is determined by NGS instrumentation's limitations and by the specific sequencing application. Selected fragments are ligated with primers P1 and A at their end and amplified by four or five PCR cycles. The obtained DNA fragments with adapters constitute the library (Figure 1.4 A-B).

In order to obtain enough DNA to reach a threshold of detection of the necessary and sufficient signal to perform the sequencing, the library undergoes an emulsion PCR: this is the clonal amplification step. For this PCR, primer A coupled to biotin, and primer P1 coupled to the adapter B are used. These primers are introduced into microreactors in the presence of a sphere and a single DNA fragment. PCR in these microreactors amplifies the introduced DNA fragment, thus managing a monoclonal population of this fragment. Adapter B binds fragments to the sphere. At the end of the PCR, the sphere will be covered with clones of the fragment initially introduced. Under ideal conditions, each microreactor would initially contain one DNA fragment and one sphere. In order to eliminate spheres on which there has been no clonal amplification, for example because there was no fragment initially introduced into the microreactor, streptavidin is used. Magnetic beads coated with streptavidin will bind to biotin that has been coupled with primer A, and then, thanks to a magnet, only spheres bound to biotin, and thus to DNA fragments, will be recovered. The set of spheres covered with clones of DNA fragments constitute the sequencing matrix, which will be used in the Ion Torrent sequencer (Figure 1.4 C-D).

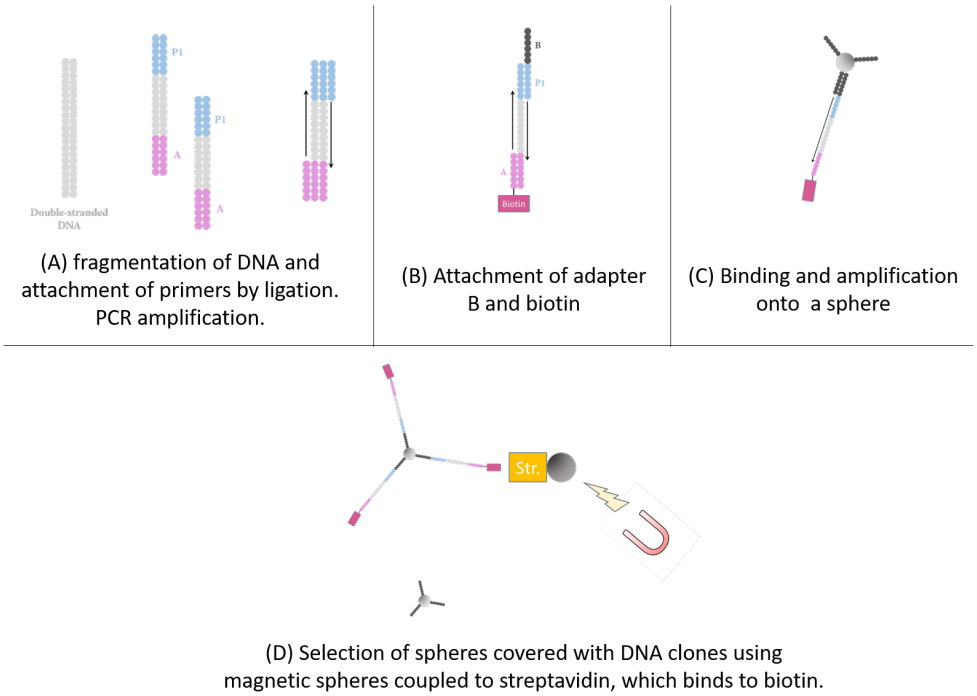


FIGURE 1.4 – Library and sequencing matrix preparation for Ion Torrent technology. Str.: streptavidin.

The sequencing matrix is introduced into microwells on a semiconductor chip, so that a single sphere linked to numerous copies of a single DNA fragment is introduced into a single well, in the presence of DNA polymerase. Then the well is flooded alternately with a solution containing one of the four deoxyribonucleotides (dNTPs) at a pH of 7.8. When the correct dNTP is incorporated by the DNA polymerase to synthesize the complementary strand to the fragment of interest, the formation of the new phosphodiester bond releases an H+ ion as shown on Figure 1.5.

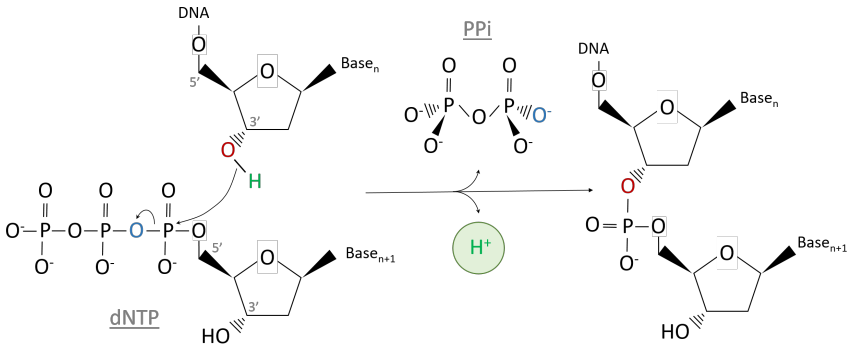


FIGURE 1.5 – Incorporation of a dNTP to DNA with DNA polymerase at the 3′ end. The phosphodiester bond creation releases a pyrophosphate (PPi) and a H+ ion which will be used to determine the sequence of the fragment.

This ions release modifies the pH of the solution, detected by the sequencer

thanks to a hypersensitive pH meter placed under each well, and therefore deduces the sequence of the fragment. Quality of the base incorporation signal gives a quality score to each incorporated dNTP, corresponding to a probability of base miscall (sequencing error). This probability of sequencing error is called a Phred score, and is encoded with ASCII symbols in the output of the sequencing. The sequencing with Ion Torrent technology is illustrated on Figure 1.6.

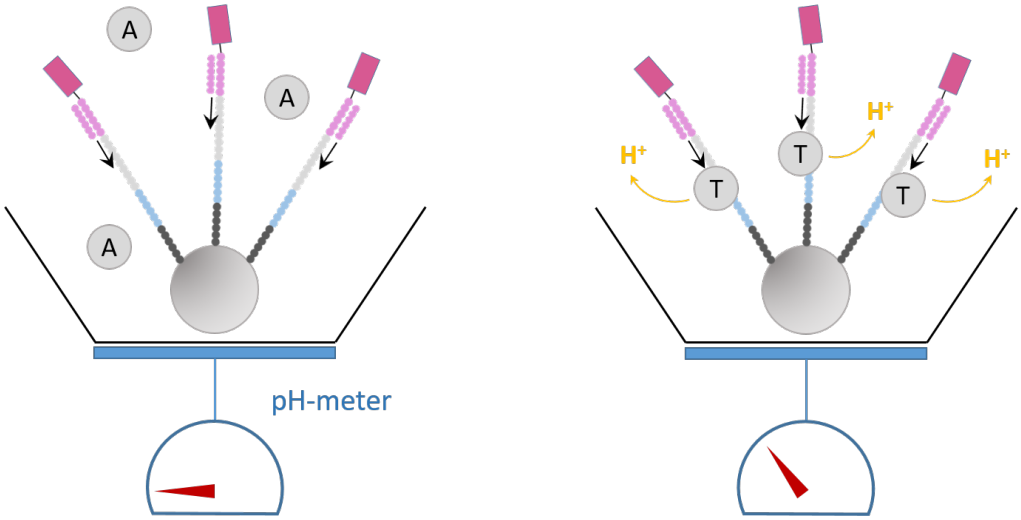


FIGURE 1.6 – Sequencing with Ion Torrent technology. A single sphere coated with clonal DNAs is introduced into each well. Each dNTP is introduced sequentially. If the dNTP is incorporated by polymerisation, it releases H^+ ions which are detected by pH-meter under each well.

If several dNTPs are incorporated, *i.e.* if the fragment contains several identical bases next to each other, pH change is greater, which is also detected by the pH meter that deduces the number of incorporated dNTPs. Thus, the DNA strands introduced into each well are sequenced simultaneously in several millions of wells. The resulting sequencing output is therefore millions of sequenced fragments, called reads. Although this is a fast and efficient method of sequencing, one limitation is the retranscription of homopolymers (a large number of identical dNTPs next to each other). Indeed, since the release of H+ ions is proportional to the number of integrated bases, it is difficult to accurately measure the number of bases incorporated when they are numerous. The sequencing errors of this technology thus lies mainly in the counting of dNTPs in homopolymers.

1.2.3 Assembly

The assembly of the metagenome is a step that leads, from the sequencing reads, to a catalogue of genes that is crucial for metaproteomics data analysis (Section 1.3.3.1). The microbiota being a complex ecosystem whose all components are poorly known, its metagenome is assembled "*de novo*", which means that it will be built without

any prior knowledge of its bacterial species. The *de novo* assembly differs from the reference-guided assembly, in which the reads are aligned on reference genomes of the studied ecosystem and the contigs reconstruction is performed by inference thanks to the reference genome. A contig is a set of reads whose sequences are overlapping, thus defining a long consensus DNA sequence. The reference-guided assembly requires reference genomes representing the ecosystem, and alignment to the reference genome is highly time-consuming. This method of assembly is therefore particularly suited for genomics, where few genomes are studied and the number of reads is limited. In the case of metagenomics, where the number of reads is counted in tens of millions, the time devoted to the alignment is a first obstacle to the use of this method. In addition, the number of reference genomes is limited in the context of gut microbiota. *De novo* assembly is therefore preferred, which makes it possible to capture the genome of still unknown microbial species.

Before the actual assembly, several steps are necessary to obtain high-quality reads to be assembled. Then, assembled reads need to be cured to obtain a high-quality gene catalogue.

Filtering contaminants and low-quality bases

This step is meant to obtain high-quality reads for the downstream assembly. The nucleotides of low-quality are removed, as well as reads with a low mean Phred score and too short reads. Shotgun methods sequence any DNA present in the sample, regardless of its origin (Section 1.2.1.2). However, since we are interested in the specific assembly of bacterial metagenomes, it is also necessary to eliminate reads that are not from bacterial origin. The parameter choices for these steps have been calibrated in MetaGenoPolis to obtain optimal quality reads to assemble.

***De novo* assembly based on De Bruijn graphs**

The reads are assembled into contigs. The current *de novo* metagenome assemblers rely on the principle of De Bruijn oriented graphs. These graphs represent the overlaps of length n-1 between the words of length n of a given alphabet. The nodes of these graphs are the words of length n, and the edges represent the overlap of an n-1 size of two words. In the context of the assembly, reads resulting from the sequencing are decomposed into k-mers, which are subsequences of length k contained in the reads. The De Bruijn graph is reconstructed with (k-1)-mers as nodes, and the observed k-mers as the edges. For example, a CCGTC sequence read can be decomposed into 3 k-mers of length 3: CCG, CGT, and GTC. The De Bruijn graph representing the k-mer CCG will consist of two nodes of size (k-1), CC and CG, which will be connected by the edge representing the k-mer CCG. After generating the De Bruijn graph with all the sequenced reads, the reconstruction of the contigs is an Eulerian path search in the graph, which is the path to go through all the edges of the graph. This reconstruction is illustrated on Figure 1.7.

The choice of the k-mers size is a trade-off. Indeed, in the case where it is too

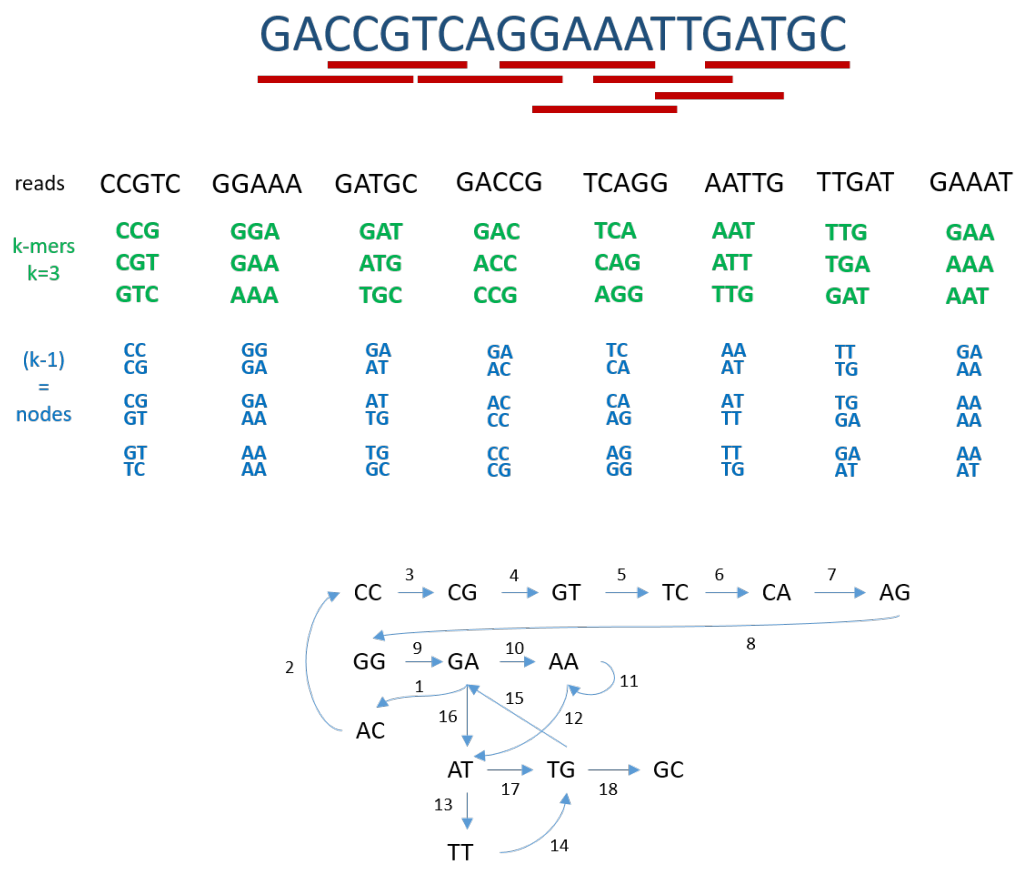


FIGURE 1.7 – Genome reconstruction from reads with a De Bruijn graph. The genome is sequenced with 8 reads (in red), whose sequences are in black. The reads are fragmented in k-mers of length 3, in green. The (k-1)-mers, in blue, are the nodes of the De Bruijn graph. They are linked by edges which are the k-mers observed. The Eulerian path reconstructs the genome, here symbolized by the numbers at the edges.

small, the graph can become too complex and the Eulerian path too difficult to reconstruct, especially in the case of repeated sequences. With a k too large, the genome is split, making it impossible to rebuild the genome.

Prediction of genes and building of the catalogue

Genes are predicted on the reconstructed contigs. In the context of the gut microbiota, gene predictors dedicated to bacterial genomes is preferred. Following genes prediction on the contigs, their redundancy is eliminated, classically by clustering genes with a high identity percentage. To generate the catalogue, the clustering of the predicted genes is first performed separately on each metagenomic sample to eliminate intra-sample redundancy. Then, the merged catalogue is computed by clustering the genes of all the samples. Clustering of the genes' catalogues, while adapted to study the gut metagenome, is a critical step when the catalogue is aimed to be used as a reference in metaproteomics; this challenge is further explained is the

Section 1.3.3.1.

Various software solutions are available for performing each of these steps, however a standard pipeline that integrates tools has been so far missing. In collaboration with colleagues, I developed MetaRaptor, a pipeline for metagenome assembly and gene prediction designed to handle many metagenomic samples, presented in Section 4.2.

The intestinal microbiota is mainly studied *via* metagenomics, which brings knowledge of the genetic potential of the microbiome. The rise of functional metagenomics, which studies the functional capacity of ecosystems, provided evidence of distorted functional potential of the gut microbiota in pathologic conditions [17]. However, metagenomics does not provide direct information on gene expression. The identification and quantification of functions expressed in a particular physiopathological context requires an in-depth and without *a priori* deciphering of the metaproteome, which reveals metabolic and cellular functions actually expressed by the microbiome and their possible association with the hosts' health.

1.3 Metaproteomics

Compared to metagenomics, which provides taxonomic structural information and genomic potential profiling of the microbiome, metaproteomics gives us access to the active part of the microbiome, whether in term of protein species inventories reflecting diversity, or relative abundances of the proteins that are operating in the system at a given time. Taxonomic and functional annotations of all the proteins detected enable us to know "who is doing what in the system". Shotgun label-free metaproteomics is precisely dedicated to the deciphering of metaproteomes without *a priori*, allowing us to approach metabolic and cellular functions, as well as microbial actors, actually operating in the system. Shotgun metaproteomics is performed by LC-MS/MS and is based on the quantification of ions characterized by their mass-to-charge ratio (m/z) and their retention time in the chromatographic column. In our context, the ions analysed are peptides (6 to 40 amino acids) obtained by tryptic digestion of the gut microbiota proteins extracted from stool samples.

1.3.1 Metaproteomics experimental workflow

A label-free shotgun metaproteomics experimental workflow typically consists of (i) sample collection, (ii) optional extraction of the microbiome, (iii) protein extraction with optional fractionation, (iv) protein digestion (most usually tryptic digestion), (v) peptide cleaning, (vi) peptide separation and analysis using LC-MS/MS, (vii) matching of the experimental mass spectra to the mass spectra library, (viii) filtering of the identification results, (ix) grouping of proteins, and (x) quantification of all peptides, proteins and protein groups identified. Finally, the peptide/protein information is used for downstream statistical analyses and taxonomic and functional

annotations. Two recent reviews focuses on some, but not all, of these steps [7, 9] with the first guidelines published in April 2019 [9]. I was involved in the bioinformatics processing of the analysis, *i.e* steps (vii-x), and present hereafter the principles of metaproteomic analysis as well as the downstream bioinformatic processings.

1.3.2 LC-MS/MS analyses

To date, mass spectrometry remains the analytical platform of choice for metaproteomics. Bulk of peptides from the microbial proteome are introduced and separated in an HPLC column (High-Performance Liquid Chromatography) in liquid phase. As they are eluted, they are transferred to the gas phase by electrospray. Briefly, at the end of the needle, the surface tension as well as the application of an electric field between it and the opposite electrode located in front of the needle allows the formation of a Taylor cone, which creates droplets of liquid phase with a diameter of about one micrometre. The electric field also charges the peptides when they are still in the liquid phase, thus ions are already formed during the liquid phase. The charged peptides are called "peptiz", and the different charges of one peptide can lead to several different peptiz. The electric field, coupled with a high temperature (200°C) also allows the solvent to evaporate, the droplets of samples cracking until reaching the gaseous phase (desolvation) [18]. This process is illustrated on Figure 1.8.

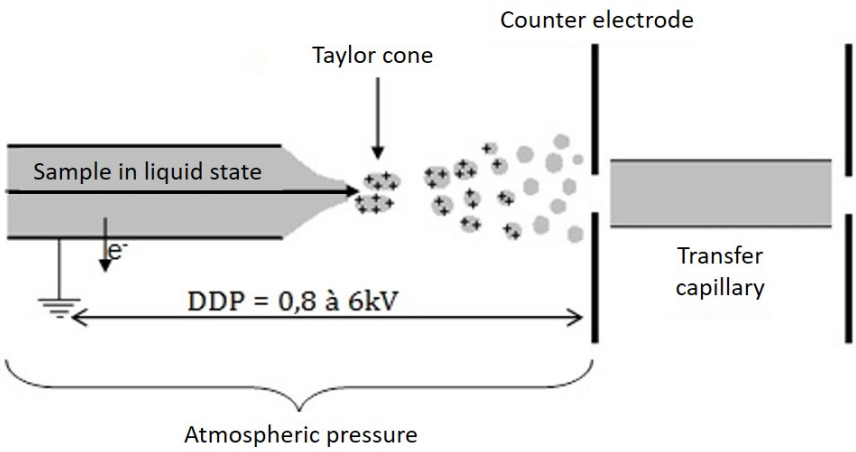


FIGURE 1.8 – Gasification of samples in liquid phase by electrospray at the input of the mass spectrometer. The liquid phase forms a Taylor cone at the end of the needle, and the electric potential coupled with the high temperature allows the sample to be desolvated, switching to a gas phase [18].

Isolated ions are guided in a beam guide (①, Figure 1.9), which has a bent structure. An electric field allows correctly ionized peptides to be guided in the bend, the non-ionized peptides colliding at the bend of the beam guide. The correctly ionized peptides thus arrive at the quadrupole (②, Figure 1.9). This structure composed of

four electrically connected rods applies a radio frequency that vibrates the ions passing through it. Ions with an unstable trajectory or a m/z ratio that is not compatible with this frequency will collide with the rods, which selects the ions entering a certain range of m/z ratio. This method allows in particular the selection of the ions corresponding to peptides and the limitation of contamination by other molecules.

The mass spectrometers used for the datasets analysed in this thesis were an Q-Exactive for the ObOmics study and an Orbitrap Fusion^{TM} Lumos^{TM} Tribrid^{TM} for the MICI-Pep and ProteoCardis studies. Both are based on the analysis of the ions by Orbitrap. The selected ions are stored in the C-trap (③, Figure 1.9), which is a curved ion trap. The ions are then pulsed towards the Orbitrap (④, Figure 1.9), which analyses the mass spectra. The Orbitrap is an ion trap composed of two electrodes, with a hollow one inside which is placed coaxially the second spindle-shaped electrode. Ions are introduced between the two electrodes. The imposition of an electrostatic field allows ions to fly between the two electrodes, with a circular movement around and an oscillatory movement along the central electrode. The current induced by the oscillation converts the signal into frequency then into m/z ratio with a Fourier transformation. This first analysis step, called MS1, allows the m/z and the intensity of the so-called "parent" peptides to be measured.

The most intense peptides are selected by their m/z thanks to the quadrupole, and stored in a linear ion trap (⑤, Figure 1.9) located behind the C-trap. This trap isolates (thanks to the same system as the quadrupole) but also fragments the peptides of interest by exciting them electrically to give them an internal energy sufficiently large so that the ions are fragmented. Peptide fragments are thus routed to the Orbitrap in order to proceed to the second analysis and recording step of the m/z, called MS2. These very specific MS2 spectra are more easily identifiable than the MS1 spectra. The double analysis with intermediate fragmentation is called tandem mass-spectrometry. The scheme of the Orbitrap Fusion^{TM} Lumos^{TM} Tribrid^{TM} mass spectrometer used in the analysis of the MICI-Pep and ProteoCardis samples is illustrated on Figure 1.9.

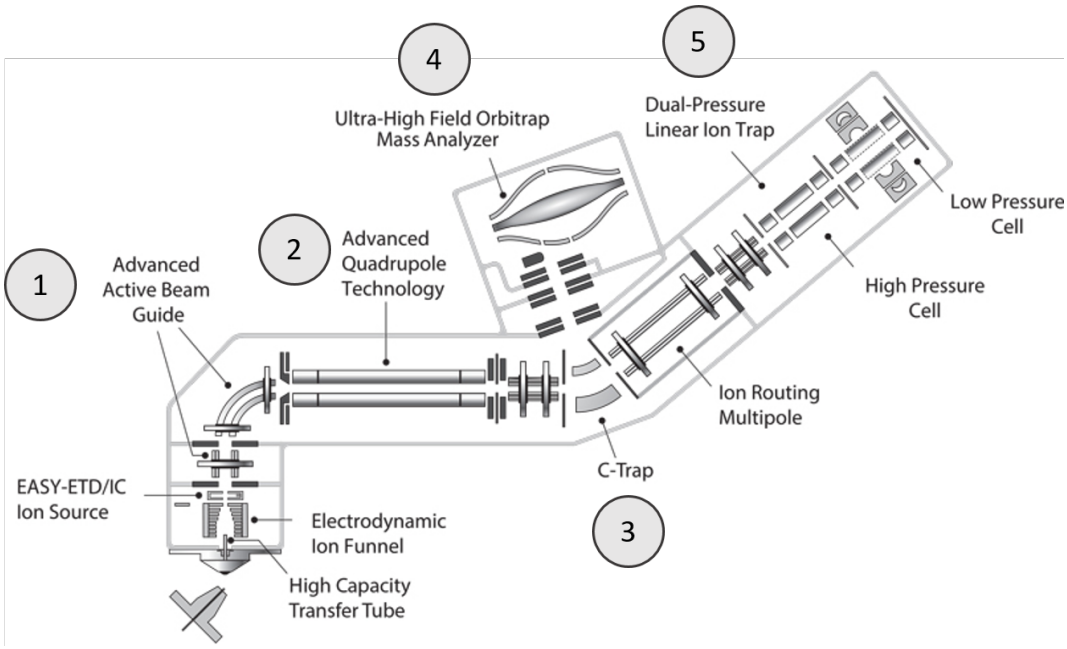


FIGURE 1.9 – Scheme of the Orbitrap Fusion^{TM} Lumos^{TM} Tribrid^{TM} mass spectrometer. The sample is electrosprayed to the transfer tube, and filtered in ① the beam guide to exclude non-ionic molecules. ② The quadrupole selects the ions based on a m/z range. ③ The C-trap traps the ions and pulse them towards ④ the Orbitrap, which performs the first mass analysis (MS1). The ions of interest are selected and trapped in ⑤ the linear ion trap, which fragments them. The fragments are routed to the Orbitrap for the second analysis (MS2) [19].

LC-MS/MS thus allows for the acquisition of a very large number of MS1 and MS2 spectra (for example, about 15 000 MS1 and 100 000 MS2 spectra were observed in our data with 3 hours runs of analysis with Orbitrap Fusion^{TM} Lumos^{TM} Tribrid^{TM}), characterized by their m/z ratio, their intensity, and their retention time in the HPLC. These spectra must be interpreted in order to understand to what peptides they correspond.

1.3.3 Interpretation of LC-MS/MS data

1.3.3.1 Peptide-spectrum matching

Mass spectrometry data interpretation is based on the comparison of the acquired experimental spectra with theoretical spectra from a protein database, namely the target database. The protein database to interrogate is a gene catalogue generated by metagenome assemblies (Section 1.2.3), translated into proteins. Adequacy and completeness of the database used for the assignment of mass spectra are crucial for the matching of theoretical and experimental spectra. Ideally, complete individual metagenomes should be sequenced and translated into proteins, since they would

represent all the genetic information of the individual gut microbiota. I devoted the Chapter 4 of my thesis to this crucial aspect in the context of large-scale experiments.

Proteins composing the database are digested with trypsin *in silico*, in order to obtain all the theoretically identifiable peptides, whose spectra are generated for comparison with experimental spectra. The classical method of database interrogation consists in comparing each experimental spectra to the theoretical spectra database and selecting the Peptide-Spectrum Matches (PSM) having an e-value score lower than a given value. This score calculation may vary from one identification software to another.

Some of the most used identification software solutions are Mascot [20], OMSSA [21] and X!Tandem [22]. Taking into account speed (which is challenging in our high-scale project) and sensitivity, X!Tandem was proven to outperform the other software solutions [23]. Moreover, in the case of Mascot, the calculation of the e-value score is highly correlated to the search space, *i.e.* to the size of the database used, as well as the identification parameters chosen, contrary to X!Tandem [24].

The X!Tandem algorithm eliminates technical artefacts and noise, *in silico* digests the proteins of the reference database, and searches for post-translational modifications. A hyperscore between theoretical and experimental peptides is calculated, based on the dot-product between experimental spectrum and predicted spectrum peaks [25]. The hyperscores distribution of the set of spectra permits to compute the probability that a hyperscore occurs by mistake, and thus to transform the peptides hyperscores into e-values [26].

The proportion of false-positive identifications is usually controlled using a decoy database, *i.e.* a database of sequences known to be incorrect, built by inverting the protein sequences of the original database. Spectra are queried against a concatenation of the database of interest (target database) and the decoy database; this type of query is called "target-decoy interrogation" [27]. This interrogation allows us to evaluate global false hits or error rate, and thus to calculate for a threshold of e-value a false discovery rate (FDR) following the equation of Käll [28]:

$$FDR = \frac{number\,of\,decoy\,PSMs}{number\,of\,target\,PSMs} \tag{1.1}$$

This FDR can be used as a filter, where the e-value threshold is calculated not to exceed a fixed FDR threshold, or as a quality control of the results for a fixed e-value threshold. Using FDR as a filter with X!Tandem is nevertheless more complicated as the search engine requires an e-value as the threshold. Interrogation of databases can be processed in one step or successive stages. I devoted the Chapter 2 of my thesis to the optimization of large databases interrogation for high-scale metaproteomic experiments.

1.3.3.2 Protein identification and grouping

Identifications provide for each experimental spectrum the set of PSMs having passed the e-value threshold, resulting in a very large dimension dataset. The complexity of these data is reduced *via* a grouping step which generates an exhaustive and parsimonious list of the peptides and proteins present in the studied samples. I used the grouping algorithm included in X!TandemPipeline [29], a software developed by the PAPPSO platform. First, a minimum of two distinct peptides identified across all samples in the dataset is set to validate a protein, in order to exclude proteins with weak proof of presence. Second, the presence of a protein is attested if it contains at least one specific peptide, which is a peptide that is not seen in any other protein. Lastly, proteins identified are assembled into subgroups and groups:

- If a protein has no specific peptide, it is eliminated (no proof of presence)
- Proteins identified with the same set of peptides are assembled into subgroups because one cannot distinguish which of these proteins is/are present. One of the proteins is arbitrarily chosen as a representative of the subgroup, but the output of X!TandemPipeline gives access to the entire list. Thus the number of subgroups in an experiment is the minimal number of proteins present in the samples.
- Groups are formed by gathering two-by-two the subgroups that share peptide(s). Thus groups are defined by a set of peptides that are not shared with other groups. Proteins in a same group have been observed to have a similar function and to belong to closely related members of the ecosystem [29, 30].

In this thesis, I computed from each mass spectra datasets the lists of peptides and of inferred subgroups and groups. An example of grouping is shown on Figure 1.10.

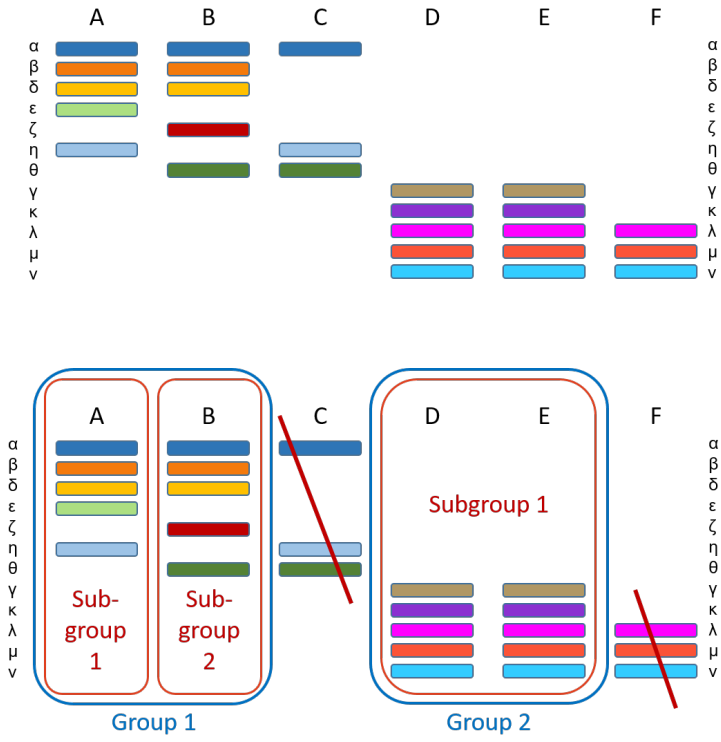


FIGURE 1.10 – Example of grouping of 6 proteins. The identified proteins are represented by letters (A-F) (top). These proteins were identified by inference of peptides identified by mass spectrometry (coloured bands). After grouping (down), the proteins F and C are eliminated because they do not have specific peptides (no proof of presence). The D and E proteins are identified with the same set of peptides, it is impossible to distinguish which of the two is present, so they form a subgroup for which D is arbitrarily selected as the representative. Proteins A and B each have two specific peptides (light green and light blue for A and dark red and dark green for B), and thus form two distinct subgroups. A and B have three peptides in common (dark blue, orange and yellow) and therefore belong to the same group. D and E do not share peptides with other proteins, so the subgroup is alone in its group. By inference, we can say that this sample is composed of at least 3 proteins: A, B and D/E.

1.3.3.3 Spectral Counting quantification

Identification in MS2 also provides a quantification of peptides by Spectral Counting (SC), which is the number of MS2 spectra observed for each peptide. SC quantification takes integer values, which infer semi-quantitative abundances. Peptide SC is a direct result of the mass spectrometry analysis. However, peptide counts matrix is usually sparse in metaproteomics context, making differential analysis difficult. In addition, the large size of the matrices can be a brake on some approaches requiring significant computation time. As subgroups and groups are defined by sets of peptides, we can infer their abundances from abundances of their related peptides. As groups do not share any peptides, the computing of their abundances is

straightforward. But this is not the case for subgroups, therefore we considered two different ways of inferring the subgroup abundances, arbitrarily named "protein SC" and "subgroup SC":

- **protein SC**: the sum of SC of peptides (specific and shared) which belong to the subgroup. This counting overestimates the subgroup abundances, since the abundance of shared peptides is duplicated as many times as the peptides are shared. It uses all the information of the identification. It is used in this thesis when we are counting the number of subgroups identified in each sample of a dataset.

- **subgroup SC**: the sum of SC of **specific** peptides which belong to the subgroup. This counting underestimates the subgroup abundances, since the shared peptides (which bring bias in abundance [31]) are not taken into account. Although the subgroup SC does not reflect the abundance of all proteins present in the sample, it provides a minimal but robust image of the proteome/metaproteome landscape. This counting is used in this thesis to represent taxonomical distribution of the subgroups in the samples, as well as for the research of differential abundances of the subgroups of the MICI-Pep, ObOmics and ProteoCardis studies. Of note, the total number of subgroups identified in a whole dataset in subgroup SC and protein SC are equal ; only the abundances differ. Both of them can therefore be used to count the total number of subgroups identified in a dataset.

- **group SC**: the sum of all peptide SC which belong to the group. Since we repeatedly observed that a given group was in the vast majority of cases, associated to a unique molecular function [29, 30], we can use the counting of groups to detect, identify and quantify the different functions expressed in the samples.

An example of the different SC quantification is presented in Tables 1.1 A-D, the peptides and proteins being those illustrated on Figure 1.10.

We chose to eliminate subgroups that do not have an SC equal of greater than 2 in at least one sample of the dataset considered in protein SC. Indeed, proteins with low abundances have a weak reproducibility in term of detection. This point will be established in Section 2.4.6.

TABLE 1.1 – (A) The number of MS2 peaks for each peptide in each sample is retrieved. The peptides are identified in the subgroups indicated between brackets. (B) The sum of all the peptides in a subgroup infers the protein SC. Here protein A is inferred from the counting of α,β,δ,ϵ and η. Protein B is inferred from the counting of α,β,δ,ζ and θ. Protein D is inferred from the counting of γ,κ,λ,μ and ν. (C) The sum of the specific peptides of a subgroup infers the subgroup SC. Here protein A is inferred from the counting of ϵ and η. Protein B is inferred from the counting of ζ and θ. Protein D is inferred from the counting of γ,κ,λ,μ and ν. (D) The sum of all the peptides in a group infers the group SC. Here SC of group 1 is inferred from the counting of α,β,δ,ϵ,ζ,η and θ. SC of group 2 is inferred from the counting of γ,κ,λ,μ and ν.

(A) Peptide SC.

peptides	sample 1	sample 2	sample 3
$\alpha_{(A,B)}$	1	2	0
$\beta_{(A,B)}$	2	0	0
$\delta_{(A,B)}$	0	0	2
$\epsilon_{(A)}$	0	1	0
$\zeta_{(B)}$	1	0	1
$\eta_{(A)}$	0	0	2
$\theta_{(B)}$	0	0	2
$\gamma_{(D)}$	1	0	1
$\kappa_{(D)}$	1	0	2
$\lambda_{(D)}$	1	0	0
$\mu_{(D)}$	2	0	0
$\nu_{(D)}$	0	2	1

(B) Protein SC.

subgroups	sample 1	sample 2	sample 3
A	3	3	4
B	4	2	5
D	5	2	4

(C) Subgroup SC.

subgroups	sample 1	sample 2	sample 3
A	0	1	2
B	1	0	3
D	5	2	4

(D) Group SC.

group	sample 1	sample 2	sample 3
1	4	3	7
2	5	2	4

1.3.3.4 eXtracted Ion Chromatogram quantification

eXtracted Ion Chromatogram (XIC) is an alternative quantification of peptiz (a peptide with a given charge) abundances, based on the area under their MS1 curve (the intensity of the chromatographic signal) over time. While SC method only quantifies the peptiz that have been fragmented in MS2 and then identified, XIC method also quantifies peptiz that have not been fragmented in MS2. Quantification by XIC of those peptides relies on their identification in another sample. By concordance of their m/z and their retention time (after alignment of the retention times between the samples), non-fragmented peptiz can be identified and quantified by XIC. Moreover, XIC quantification have better accuracy and repeatability than SC quantification when high-resolution analysis is performed [32] (which is the case in the experiments performed for this thesis). However the use of XIC quantification is challenging in metaproteomics, a topic developed in Section 5.1.

This measurement allows us to obtain continuous values (Figure 1.11). XIC missing value may correspond to (i) a peptiz which is really missing in the sample, (ii) a peptiz which has an abundance under the detection capacity of the mass spectrometer, (iii) a peptiz for which the area cannot be computed because its MS1 peak is not detected at different timepoints, and therefore not quantifiable. Since it is difficult to identify the reason which led to a missing value, they are filled with a missing value (NA).

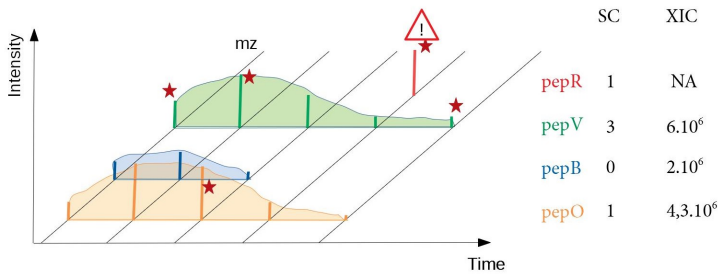


FIGURE 1.11 – Example of quantification of four peptiz by SC and XIC. At each time (x axis), the MS1 peak (coloured bands) with the greater abundance (y axis) is fragmented into MS2. The peak fragmented is symbolized by a red star. pepB is not fragmented, therefore it has a null quantification by SC but its presence is detectable by XIC and quantifiable if it has been fragmented and identified in another sample. The relative abundance of pepV and pepO is more accurate by XIC than by SC. PepR, which is fragmented in MS2 and therefore identified and quantified by SC, has no quantifiable MS1 area because the peptide is not detected at another timepoint.

XIC abundances at the peptide level is defined as the sum of the abundances of their peptiz, and XIC abundances of subgroups and groups are computed similarly to SC, by summing XIC abundances of their specific and total peptides, respectively.

1.3.4 The rise of metaproteomics in the last decade

The ability of metaproteomics to decipher ecosystems from a meaningful postgenomic point of view was first demonstrated in the early on activated sludge [33] and in a natural microbial biofilm [34]. A couple of years later in 2007, reproducible two-dimensional gels coupled with extraction of proteins and attempted identification using MALDI-TOF-MS, demonstrated the applicability of proteomics to complex intestinal ecosystems [35], although insufficient microbiome sequence information was a real bottleneck for protein identification at that time.

In the succeeding decade, say over the 2007-2018 period, there was an increasing number of small-scale proof-of-principle and methodological studies that demonstrated the feasibility and the relevance of intestinal metaproteomics approaches in areas as diverse as human physiopathology [36–44] including preterm infants [45–49], rodent modelling [40, 50–54], and feeding, health and well-being of livestock animals [55–59]. This was made possible with ongoing expansion in metagenomic databases and bioinformatics tools. The common goal was a better knowledge of the key players that drive cellular and metabolic activities of the intestinal microbiome, how they are linked with the heath of the host, and might ultimately serve as candidate biomarkers or targets to design therapeutic intervention systems for various fields of application.

Interestingly, two studies published in 2017 [39] and 2019 [60], compared metagenomics and metaproteomics profilings of the same stool samples, and found considerable divergences between genetic potential and functional activity of the human gut microbiome. Clearly, metaproteomes displayed a higher plasticity compared to the lower inter-individual variability of metagenome profiles in 15 healthy volunteers [39]. This was confirmed at the individual level, in a series of eight stool samples longitudinally collected from a unique Crohn's patient repeatedly observed over a 4.5-year period [60]. These pioneering multi-omics studies provide preliminary demonstration that DNA-to-protein correlations seem to be low in complex microbial systems, with consistent but also highly divergent trends between the two data types. This well emphasizes that expression and not only potential of intestinal functions must be integrated in a near future, and that metaproteomics profiling might be a powerful mean to better reflect disease-related changes.

Finally, only one label-free shotgun metaproteomics study published in 2018 [61] included for the first time a significant number of subjects. It focused on Intestinal Bowel Diseases (IBD), human pathologies known to be accompanied by serious dysbiosis of the intestinal bacterial populations based on targeted metagenomics [62] and shotgun metagenomics [63]. Forty-seven treatment-naive paediatric patients diagnosed with IBD (25 Crohn's disease and 22 ulcerative colitis), and 24 healthy controls were enrolled. It was found that microbial proteins related to oxidative stress responses were overrepresented at the mucosal-intestinal interface of IBD patients. At the same time, the bulk of human proteins related to antimicrobial activities was

increased in extravesicular vesicles derived from intestinal immune cells of those patients, and correlated with alteration of microbial functions.

1.3.5 The challenges of metaproteomics

Each of the multiple steps of the metaproteomics workflow can have profound impact on metaproteomics results. On the one hand, sample preparation (in which I did not contribute) can be processed with multiple methods. On the other hand, bioinformatic analyses face computational and accuracy non-trivial challenges.

1.3.5.1 Sample preparation

Stool sample collection: samples are self-collected by the participants and can be temporarily stored before processing. Ideally, freshly collected specimens are prepared as soon as possible, typically within 2 hours after emission. If this is not possible, a common practice is storage of crude faecal specimens in a -80°C standard freezer, since the faecal matrix is highly cryoprotective.

Pretreatment: given the high complexity of stools that contain bacterial, dietary and host proteins, most metaproteomics studies first extract the microbiota from the faecal matrix, to focus on microbial proteomes. Pretreatment of samples can be performed by differential centrifugation or gradients made of Nycodenz®, Histodenz^{TM} or Iodixanol (commercial name OptiPrep^{TM}). Differential centrifugation was shown to lead to a higher number of peptide and protein identifications, and thus a higher taxonomic and functional diversity, when compared to raw samples [42]. However, great differences in the relative abundance of several gut microbial taxa were also observed after differential centrifugation, notably a significant increase in the Firmicutes/Bacteroidetes ratio, as well as a decrease in some important microbial functional categories, including cell surface enzymes, membrane-associated proteins, extracellular proteins and flagellins. Importantly, all samples analysed in my thesis were pre-treated by a sophisticated Nycodenz or Iodixanol gradient (detailed in the Appendices), which preserved anaerobiosis essential to the most oxygen-sensitive species all along microbiota extractions. This extraction method was shown to preserve microbial diversity of the total stool samples [30]. It also discards various unknown or unexpected chemicals, which might affect the efficiency of enzymatic digestion in the downstream.

Protein extraction: protein extraction from microbiome samples is challenging due to the resistant structure of Gram-positive cell walls, which can be even more resistant in natural ecosystems than in pure cultures. One can proceed with mechanical or chemical lysis or a combination of both. A previous study has shown that combining sodium dodecyl sulfate (SDS)-based lysis buffer and ultrasonic probe, was the most efficient approach to recover proteins from both Gram-positive and -negative bacteria [64]. SDS is a strong anionic detergent that efficiently assists in the

disruption of biological membranes, but prevents downstream fractionation into cytosolic and envelope-enriched subcellular compartments. For this reason, we chose not to use it and proceeded with an ultrasonic cell disruption only. This allowed us to fractionate every lysate analysed in this thesis into its cytosolic and envelope-enriched fractions, which were separately treated and analysed in the downstream analyses, in order to increase depth of metaproteome analysis. Such a "divide and rule" strategy considerably improves the coverage of the metaproteome, and advantageously replaces the more common protein bulk fractionation into 1D-PAGE bands, which necessitates as many LC-MS/MS analyses as gel bands, and cannot be applied in large-scale studies such as those presented in this thesis.

Protein digestion and peptide cleaning: it can be performed either in-gel or in-solution. In-gel digestion has long been routine in proteomics and presents the advantage of getting rid of any chemicals that might affect the efficiency of enzymatic digestion, as only proteins will move through the gel. A second advantage is high accessibility of all protein species to proteolysis, as they are spread into the gel network. But in-gel digestion has a low throughput. Therefore, in-solution digestion after cleaning by acetone or acetonitrile precipitation is preferred in high-scale studies like those presented in this thesis work. This allowed us to digest batches of 18 to 22 samples according to well-defined, completely randomized designs. Peptide cleaning is the last step before LC-MS/MS injection, and consists in eliminating all reagents for keeping only a clean, desalted peptide bulk. This is performed by solid phase extraction, traditionally on C18-modified silica-based sorbents, and more recently on diverse polymeric sorbents.

1.3.5.2 Bioinformatics analyses

Data complexity: in metaproteomics, unlike proteomics, several hundred microorganisms are studied. The great complexity of these data therefore requires powerful algorithms to process the data, in computation time and memory capacity. Thus, algorithms that can perform well in proteomics (or in small-sized metaproteomics) cannot always be used in a large-scale metaproteomics context. In particular, protein databases used in metaproteomics have a much broader dimension than those used in proteomics, notably in the context of the human gut microbiota, where large inter-sample variability requires very large databases. In addition, exhaustive identification of peptides based on mass spectral library matching would require that the protein sequences in the database are exactly the ones observed experimentally, which is unrealistic taking into account the huge different microbial strains and mutations within a single microbiome. Besides, increasing the database size reduces the sensitivity of identification algorithms [65]. Therefore, the choice of a reference database is not trivial since it must be a trade-off between identifying a maximum of peptides present in the sample with an exhaustive database, and ensuring a better sensitivity with a database not too wide. This is particularly challenging in the

context of the gut microbiota, where the component species are extremely variable between individuals. Recently, an iterative method of identification has been developed to interrogate the huge databases used in metaproteomics, with the aim of increasing the identification sensitivity. This well illustrates that specific bioinformatic developments, in which I participated in this thesis, are necessary in metaproteomics.

Redundancy of the protein sequences: several proteins can be identified with the same set of peptides (and thus form a subgroup), so we chose to work mainly with the representative protein of each subgroup since it is impossible to distinguish which protein(s) is/are actually present in the sample. Since it is difficult to assign the counting of the peptides shared between different subgroups [31], we chose, when differential abundances between subgroups are searched, to quantify subgroups only with their specific peptides (subgroup SC) [66–68]. On the contrary, we chose to take into account all peptides (protein SC) when searching for the subgroups presence in the samples [41, 69].

Statistical analysis of differentially abundant subgroups: it can be challenging due to the highly sparsity and low abundance of subgroup SC values. XIC quantification, which has better accuracy and repeatability in proteomics and generates less sparse matrices [32], could be useful in the detection of differential abundances of subgroups. However, XIC quantification requires the alignment of all chromatographic peaks of all samples in the experiment in order to quantify peptides that have not been identified in MS2 (such as peptide B, Figure 1.11). In the context of metaproteomics, where the number of chromatographic peaks is very high, alignment is the most challenging since deviation of the retention time of a few seconds can lead to misalignment with another peptide ion having the same mass/charge ratio. XIC quantification of the human gut microbiota and the particular challenges linked to XIC alignment are studied in Chapter 5 of the thesis.

Metaproteomics is interested in learning about metabolically active microbial members, which communicate not only to each other, but also with their host with the best and worst repercussions on health. Previous studies show us that it is possible to use metaproteomics as a tool to study different physiopathologies of the human gut microbiota. It has been possible, through a metaproteomics approach, to identify signature proteins of IBD that were not suspected in metagenomics [61]. The possible link between intestinal microbiota and obesity and between intestinal microbiota and type 1 diabetes (two important risk factors for cardiovascular diseases) was reported in a small-scale study [38] and case report [44]. We were therefore interested on deciphering, with a metaproteomic approach, the possible link between gut microbiota and cardiovascular diseases, which are complex diseases with many risk factors.

1.4 Cardiovascular artery diseases

Cardiovascular diseases (CAD) are now the leading cause of death in the world, with a higher prevalence in western countries. In the most recent reports, 31.5% of annual deaths are caused by cardiovascular diseases, corresponding to more than 17 million deaths [70]. In addition to its mortality burden, CAD is a leading cause of morbidity and loss of quality of life. This makes CAD a major public health scourge with dramatic economic costs. The major cause of cardiovascular diseases is atherosclerosis, a vessel disease characterized by the development of a plaque reducing its lumen. The physiopathology of atherosclerosis is developed in Appendix B.

Considerable progress has been made in defining clinical risk factors for the development of cardiovascular complications. Subjects with metabolic syndrome, obese or diabetic subjects represent subpopulations at high risk. The classical risk factors are presented below.

Metabolic syndrome

This syndrome is defined by a set of phenotypic and physiological characteristics, proposed by the International Diabetes Federation (IDF) [71] in 2005. To have a metabolic syndrome, a patient must present:

- Obesity, defined by waist circumference (depending on gender and ethnicity) or if the Body Mass Index (BMI) is greater than $30 kg/m^3$

as well as 2 or more of these factors:

- High triglyceride level: $\geq 150 mg/dL$
- Low High-Density Lipoprotein (HDL) cholesterol level: $\leq 40 mg/dL$ for men, $\leq 50 mg/dL$ for women
- Hypertension: $\geq 130 mmHg$ for systolic blood pressure and ≥ 85 mmHg for diastolic blood pressure
- High blood glucose level: ≥ 100 mg/dL while fasting, or previously diagnosed type 2 diabetes

Among these characteristics, hypertension is an atherosclerosis risk factor known for a long time, as suggested by the fact that atheromatous plaques are mainly developed in areas subject to disturbances of blood flow. Hypertension may weaken the endothelium by increasing the hemodynamic pressure, and induce the expression of proteins that promote the infiltration of monocytes into the endothelium and the triggering of inflammatory processes via angiotensin II. Angiotensin II stimulates the production of free radicals and the expression of pro-inflammatory cytokines.

Diabetes

Diabetes, inducing blood hyperglycaemia, triggers an inflammatory response to the endothelium through the production of advanced glycation products. These

products modify Low-Density Lipoprotein (LDL), giving them pro-atherogenic properties similar to oxidized LDL (LDL-ox) [72].

Pro-inflammatory proteins

High-Sensitivity C-Reactive Protein (hs-CRP) is expressed in an inflammatory context and promotes phagocytosis by macrophages. The elevated concentration of this protein in blood is associated with the number of thin cap atheroma. Moreover, the concentration of the protein is significantly higher in patients who suddenly died from severe coronary artery disease. Other inflammatory molecules are known to be linked to atherosclerosis, like matrix metalloproteinase 9, interleukin-6 and interleukin-18. On the contrary, the high concentration of interleukin-10, an anti-inflammatory protein, has a more favourable prognosis in patients with high hs-CRP levels and acute coronary syndromes. The balance of pro-inflammatory and anti-inflammatory proteins is therefore crucial for the development of atherosclerosis [73].

Environmental factors

Multiple environmental factors, linked to the lifestyle, can influence the risk for cardiovascular disease. Physical activity has various effects on multiple other cardiovascular risk factors, like increasing HDL production, and improve myocardial efficiency, glycogen storage capacity and insulin sensitivity. Conversely, physical activity can decrease cholesterol, body fat, platelet aggregation and blood pressure. Physical activity has therefore anti-atherogenic effects, and physical inactivity increases CAD risk [74].

The diet has a great influence on the risk of CAD. The high consumption of fat leads to a higher risk of atherosclerosis by increasing cholesterol in the blood and developing obesity, another risk factor [75, 76]. The studies about the effect of consumption of antioxidants are however contradictory. Reviews support nevertheless the protective effect of the consumption of antioxidants for atherosclerosis [77, 78].

A study on humans has shown that cigarette smoking accelerates the development of atherosclerosis, and temporarily reduces vessel vasodilation abilities in young smokers. For chronic smokers, the dilation induced by blood flow is reduced, and the frequency and duration of ischaemia are increased. The effect of cigarette smoking is explained by the production of reactive oxygen species which induce an endothelium dysfunction, and production of superoxide anions (O^{2-}) which sequester free nitric oxide (NO), an important anti-atherogenic agent [79].

As we have seen, the development of cardiovascular disease is multi-factorial. Despite this, nearly a quarter of cardiovascular risk remains unexplained [80]. Therefore, there remains a research space for innovative studies aimed at improving our understanding of the mechanisms and risk factors associated with this disease, with an obvious public health issue. Among the ways to be explored, the influence of the gut microbiota is an interesting research field. Indeed, it is now clear that a microbial

imbalance, or dysbiosis, is associated with some intestinal or metabolic disorders, such as obesity [81] or diabetes [82].

1.5 Evidence for a relationship between gut microbiota and CAD

It is well recognized that the intestinal microbiota is involved in several inflammatory diseases, in particular in IBD [83]. More recent researches brought convincing elements of the impact of the gut microbiome structure (targeted metagenomics) and functional potential (shotgun metagenomics) on host genome expression, notably in the context of obesity and other diseases with inflammatory components [4, 30, 84–87]. Studies on the intestinal metabolism of neutral and acidic steroid molecules are additional arguments in favour of a key role of commensal bacteria in the onset and development of cardiometabolic diseases [88, 89].

Evidence towards a causative role of the gut microbiota for fat storage is sustained by a set of elegant experiments in germ-free, conventional and gnotobiotic mice. Germ-free mice are resistant to diet-induced obesity [90] and germ-free born mice colonized with microbiota from obesity-prone animals eat more food, gain more weight, and become more obese than their counterparts colonized with microbiota from obesity-resistant animals [91]. Less elegant but not less convincing is a recent case report of a woman successfully treated with FMT (faecal microbiota transplantation) in the State of Rhode Island, who developed new-onset obesity after receiving stool from a healthy but overweight donor [92].

A plethora of potential mechanisms are invoked to explain weight gain linked to the gut microbiota, including food intake behaviour, increase in the intestinal glucose absorption, energy extraction from non-digestible food components and concomitant higher glycaemia and insulinemia, two key metabolic factors that regulate lipogenesis [93, 94].

Gut microbiota has also been directly linked to CAD, some types of microbiota being flagged as pro-atherogenic because they generate metabolites such as TMAO (trimethylamine N-oxide) which promote atherosclerosis through up-regulation of multiple macrophage scavenger receptors [95] or have a high potential of peptidoglycan biosynthesis which could prime the innate immune system and enhance neutrophil function [96]. Other types of microbiota could be anti-atherogenic, notably through the production of antioxidants [96].

Finally, the idea of manipulating the intestinal microbiota of especially at-risk individuals to help them remain healthy, and highlighting connections between microbiomic metabolic/cellular pathways and CAD risk to assist in the development of new strategies for maintaining or restoring health, would be of high interest. The ProteoCardis study proposes an unprecedented gut metaproteome-wide association

study of CAD on top of a metagenomic-wide one in the context of the MetaCardis study.

1.6 The ProteoCardis project

The ProteoCardis project supported by the Agence Nationale de la Recherche (ANR), is an association study between the intestinal metaproteome and CAD. ProteoCardis has selected more than 200 patients from the 2000 included in the European project MetaCardis, whose data are put into perspective with patient records, metabolic features, complete cardio-vascular exams and outcomes that are acquired in the FP7 MetaCardis framework (2013-2018). The gut metagenomes of the MetaCardis subjects are being characterized by high throughput sequencing of the total faecal DNA. The accompanying challenge proposed by ProteoCardis is a holistic metaproteomics approach to get closer to the real functionality of the gut microbiome by exploring the expression of metabolic and cellular pathways. We interpreted the metaproteomic data collected on this cohort of unprecedented size in this scientific field, and with the most powerful instruments at this time.

Without any *a priori* assumption of the metabolic and/or cellular pathways that can accompany the disease, we search for metaproteomic variables associated with CAD. The change of protein signals before and six months after bariatric surgery, an intervention known to reduce the cardiovascular risk, is also examined (Figure 1.12). The clinical status of the patients are the following:

- 49 patients with a first recent (< 2 weeks) event of coronary artery disease and normal cardiac function defined according to cardiac ultrasound with a left ventricular ejection fraction $\geq$ 45% (aCAD)
- 50 patients with chronic CAD with normal cardiac function defined as above (cCAD)
- 16 patients with chronic CAD and congestive heart failure defined according to cardiac ultrasound with a left ventricular ejection fraction $\leq$ 45% (cCADic)
- 23 patients with heart failure unrelated to coronary artery disease (cCADicncor)
- 50 healthy controls without metabolic syndrome, type 2 diabetes, CAD, chronic inflammatory or infectious disease (CTRL)
- 30 severely obese subjects (body mass index $\geq$ 40) that had bariatric surgery (BS), with an observation before BS (BS1) and six months after BS (BS2)

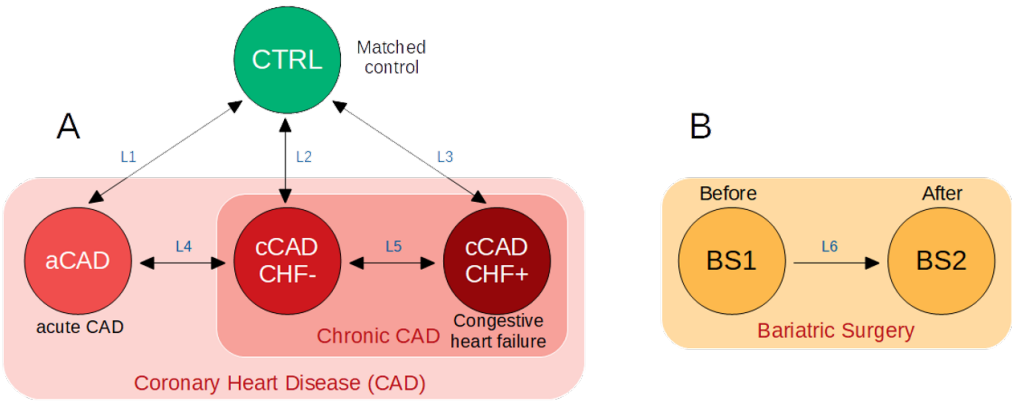


FIGURE 1.12 – Description of the ProteoCardis datasets. 188 individuals constitutes the dataset A, which comprises 138 patients suffering from CAD and 50 controls. 30 obese patients constitutes the dataset B, for whom a sample is collected before and 6 months after a bariatric surgery. The variables L are searched to discriminate the different patient groups.

The search of metaproteomic markers provides a set of relevant variables from those obtained in both contexts of aggravation (CAD *vs.* CTRL) and improvement (BS2 vs. BS1) of the clinical status. We quantified the thousands of peptides and proteins identified, and implemented, adapted, and developed robust statistical tests adapted to this type of data in order to extract reliable signatures of the studied pathology.

The next step of the ProteoCardis study is a multiplexed Selected Reaction Monitoring (SRM) assay, targeting and precisely quantifying the peptides/proteins of interest. The candidates will be first validated to test their predictive value on 20 CAD patients and 20 matched controls. The presence of the potential markers will then be examined in additional individuals at high risk of CAD (50 subjects with metabolic syndrome, 50 obese and 50 type 2 diabetic subjects), and association of these markers with any risk factors as well as complications or adverse cardiovascular outcomes of the subjects over a four year period will be investigated.

1.7 Thesis objectives

Incorporated within the framework of the ProteoCardis project, the main goal of this thesis was to optimize the process of peptides and proteins identifications and quantification for large cohorts. I was interested in all the technical and methodological challenges raised by large numbers of metaproteomic samples.

The first step was to compare several identification methodologies. In particular, I considered individual metagenomes to generate databases adapted to the metaproteomic context. Construction of such databases is not trivial and required a particular analysis on the parameter tuning. The performance of these databases had also to be evaluated.

The second step concerned the quantification stage: widely used in proteomics, the XIC quantification of peptides/proteins exceeds the precision offered by the spectral counts (SC), but its possibilities and limitations remained unexplored in the context of large-scale metaproteomics studies. The results of this work, which is therefore mainly methodological, is applicable to any metaproteomics project and goes well beyond the scope of the ProteoCardis study.

The third step was to prepare the count/abundance data for the subsequent statistical analyses. I explored various normalization strategies, on both XIC and SC quantification. Finally, the search for biomarkers has been initiated by partners of the ProteoCardis project. I did not directly performed these analyses, but I participated in the discussions and interpretations.

At the end, this work enabled on the one hand to produce abundances matrices as accurate as possible for statistical analysis in the ProteoCardis project; on the other hand to define the optimal preprocessing strategies for metaproteomics data in the context of large cohorts.

Chapter 2

Mass spectra interpretation without individual metagenomes

The identification of peptides in metaproteomics is based on the completeness of the database against which the mass spectra are interrogated. In the case where the metagenomes of the samples are not available, we use general databases, generated from samples of several hundred individuals. In this chapter, we evaluate the identification performance of two databases as well as two interrogation strategies on the ObOmics dataset. These results allowed us to interrogate, with the most efficient method, the ObOmics and MICI-Pep samples and to carry out simple analyses to highlight proteins of interest for these two studies.

2.1 The ObOmics study

The metaproteomics is still a recent field, profiling of metaproteomes thus requires an evaluation and adaptation of bioinformatics tools originally developed for proteomics. Identification of peptides and proteins composing the gut metaproteome is indeed challenging due to its complexity. We propose to compare several workflows for interpreting the ObOmics metaproteomics MS/MS dataset related to 48 individual samples of the MICRO-Obes ANR project.

These samples were analysed by LC-MS/MS on a Q-Exactive spectrometer over the years 2014-2015, and were therefore available from the start of my thesis, at a time where the ProteoCardis datasets were not yet produced. I used the ObOmics dataset as a basis for my first bioinformatic work.

The expertise and lessons learned will then be applied to the ProteoCardis cohort. Indeed, the ObOmics cohort is a quarter the size of the ProteoCardis cohort, and each individual MS/MS datafile is third the size of individual ProteoCardis datafile because acquisition was performed on a less sophisticated spectrometer. Therefore, the data are smaller and more suited for the development part of this thesis work. Interestingly, the ObOmics study includes a sample which was analysed fourteen times over the 2014-2015 period. We thus studied the repeatability of metaproteomic analyses, a study that had never been performed before on data of such a complexity.

2.2 Scientific questions

Most of the metaproteomics studies involve small cohorts and/or low-complexity samples, such as *in vitro*-produced communities containing only a few dozen species (Mock communities). More precisely, metaproteomic studies on human faeces concerned 1 to 29 patients [7] until recently, where the metaproteomes of 56 patients suffering from acute leukaemia were analysed [97]. The analysis of several hundred of samples is therefore particularly challenging because protocols and methods that have been implemented to explore metaproteomes are not suited for very large-scale studies. However, in the final few months of my thesis, first recommendations have just began to be proposed [9].

Apart from the preparation of samples and the acquisition of mass data in a sufficient depth, a main challenge is the maximization of mass spectra interpretation to produce a description of the system that is sufficiently detailed and representative. As mentioned in Section 1.3.3.1, interpretation of mass spectra into peptides then proteins goes through interrogation of databases, which in the context of metaproteomics are of large-scale and consequently requires extensive computation time. In addition to increasing the computation time, matching large mass spectra datasets to a large metaproteomic database may also dramatically decrease the identification rates at a given FDR threshold [65].

In the study hereafter, we aimed to compare two strategies of interrogation and two public human gut metagenomic databases for mass spectra interpretation in the context of a large scale metaproteomics study of the human gut microbiome. The comparisons are based on a series of well-defined qualitative and quantitative criteria detailed in Section 2.3.6.

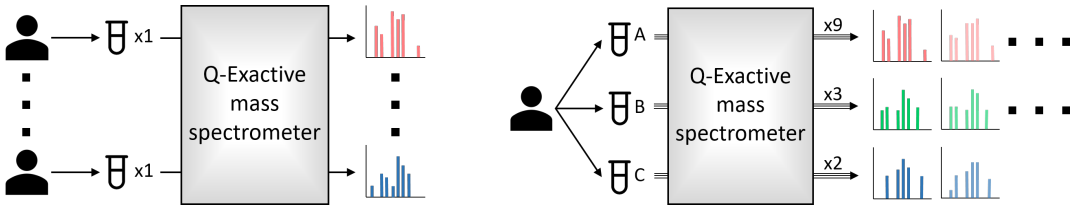
2.3 Methods

2.3.1 Samples preparation and injection

Stool samples were self-collected by the 48 overweight/obese subjects of the MICRO-Obes cohort. Faecal samples were transferred to a biobank at -80°C within two hours of collection. Then about 1g stool aliquots were cut frozen and the microbiota was separated from the faecal matrix by flotation in a preformed Nycodenz continuous gradient as detailed in Appendix C. The extracted microbiota were lysed on ice with a probe sonicator in an anti-protease cocktail containing buffer. Then the suspensions were centrifuged at 5000 x *g* for 30 min at 4°C to remove unbroken cells and large cellular debris. The supernatants were finally ultracentrifuged at 220 000 x *g* for 30 min at 4°C to pellet cell envelopes. These cell envelope-enriched fractions were acetone delipidated, trypsin digested, and the desalted peptide bulks were analysed by LC-MS/MS, all steps detailed in Appendix C.

Sample preparation and LC-MS/MS analysis were carried out only once for 47 of the stool samples and repeated multiple times for one sample (sample S32) for a

study of reproducibility (Figure 2.1). More precisely, sample S32 was prepared in triplicates, from microbiota extraction up to resolubilization of the peptide mixture in LC buffer. These preparations, called A, B and C, were injected nine, three and two times, respectively. The operator for the preparation and injection of the S32 replicates according to well-defined standard operating procedures (SOPs) changed over time, so that we measured an overall reproducibility, even though reproducibility for a same preparation is expected to be better than between different preparations. In total, we thus performed 61 LC-MS/MS runs of which 47 corresponded to non-replicated stool samples and 14 technical replicates corresponding to the S32 replicates.



(A) non-replicated samples (B) replicated sample

FIGURE 2.1 – Preparation and injection of the samples. (A) 47 individual biological samples were each prepared and injected only once. (B) One biological sample from one individual was prepared in triplicates A, B and C, which were injected nine, three and two times, respectively.

2.3.2 Interrogated databases

The LC-MS/MS data were searched against two translated human gut microbiota gene databases, MetaHIT 3.3 [98] and MetaHIT 9.9 [5]. The former, published in 2010, contains 3.3 millions of gut microbiota genes from 124 individuals from Denmark and Spain (available at `www.bork.embl.de/~arumugam/Qin_et_al_2009/`). The latter, published in 2014, contains 9.9 millions of genes from the metagenomic sequencing of 1 267 individual samples from Europe, United States and China, including the samples used in the construction of MetaHIT 3.3 (available at `meta.genomics.cn/meta/dataTools`) plus a selection of sequenced gut bacterial genomes.

The size of the database used for mass spectra identification, as explained earlier, may influence the results of the identification; moreover, the computation time can be extremely long if the database is too large. The objective is to define the interest of using MetaHIT 9.9 compared to MetaHIT 3.3. Indeed, if MetaHIT 3.3 is complete enough to identify a large number of peptides and proteins in our samples, we hypothesised that using MetaHIT 9.9 could only unnecessarily lengthen our computational time. On the other hand, if MetaHIT 3.3 is not exhaustive enough to cover the complexity of the samples, it will justify the use of MetaHIT 9.9 even if the computation time is more important.

The data were concurrently searched against the *Homo sapiens* protein catalogue from Uniprot (April 2018) including canonical and isoforms proteins from Swissprot and TrEMBL, and the contaminant database, which includes 58 sequences of common contaminants of spectrometry experiments, such as keratins, BSA, and trypsin.

2.3.3 Interrogation strategies

2.3.3.1 Classical identification

To interpret mass spectra, we used either a classical or an iterative interrogation strategy. The classical strategy consisted of a one-step target-decoy interrogation of the database (Figure 2.2).

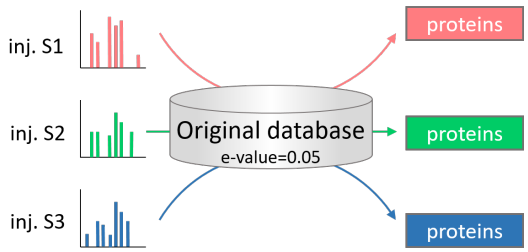


FIGURE 2.2 – Classical strategy of interrogation. In the classical strategy, the MS/MS data are interpreted in one step against the target-decoy database with an *e*-value cutoff of 0.05.

2.3.3.2 Iterative identification in two steps

To address the issue of completeness and size of the database, a strategy of interrogation known as iterative database search has been proposed by Jagtap et al. in 2012 [99]. This method showed a higher number of peptides and proteins identified, compared to the classical method conventionally used in proteomics.

The principle of iterative identification is to first interrogate large metaproteomic databases with relaxed FDR thresholds, which generate reduced individual databases including all possible proteins at the individual level, and finally to re-interrogate these sub-databases with stringent thresholds. Several variants of this strategy have been successfully applied to interpret human salivary metaproteomes [99, 100] and gut metaproteomes of mice and humans [41]. When coupled with the interrogation of the human proteome, the iterative database search also identifies many human proteins that may be highly relevant in clinical contexts [99, 100].

The first iterative interrogation approach on a complex ecosystem provides a two-step workflow on six human saliva samples. The first step aims to refine a human microbiome saliva used in the second more stringent iteration (Figure 2.3). This first paper highlights the possibility to manage databases of very large scale without affecting the sensitivity of metaproteomic analyses [65]. The authors also

showed that this method can identify a large number of human proteins, which are particularly interesting in the context of the human microbiota.

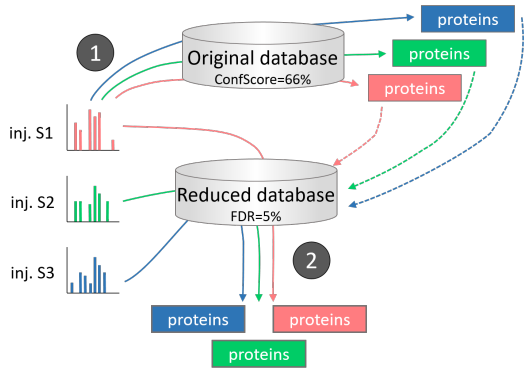


FIGURE 2.3 – Two-steps iterative strategy of interrogation. In the two-steps iterative strategy, ① the LC-MS/MS spectra are interrogated against a large database with relaxed thresholds. ② The proteins identified in the first step are merged to generate a reduced database for the second target-decoy interrogation with stringent thresholds. Inj.: injection. The threshold values given as an example are from Jagtap et al. [99].

2.3.3.3 Iterative identification in three steps

Another implementation of the iterative method was published in 2016 by Zhang et al. [41], this time using three interrogation steps and looking at the gut microbiota of eight mice and eight humans (Figure 2.4). The first step refines a human gut microbiota database and creates sub-databases specific to each sample. The second step interrogates the data against these multiple individual sub-databases with stringent parameters. Finally, the last step merges together the proteins identified in each sample to create an unified sub-database used in the stringent final interrogation. The generation of this last database is used to have the same identification performance rate across all samples, since the varying size of individual metaproteomes could affect the identification results.

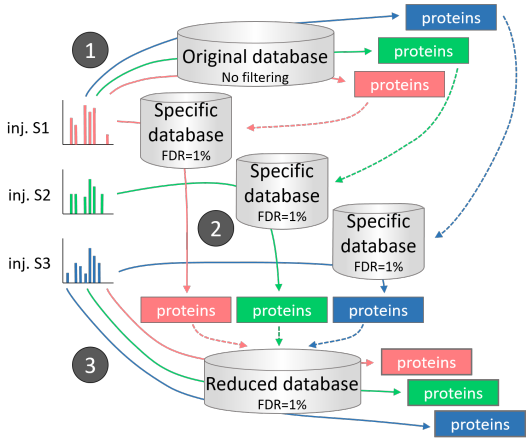


FIGURE 2.4 – Three-steps iterative strategy of interrogation. In the three-steps iterative strategy, ① the LC-MS/MS spectra are interrogated against a large database with relaxed parameters. ② LC-MS/MS spectra were searched again against their specific database and their decoy with an stringent FDR cutoff. The proteins identified in this second step were put together to create a reduced database specific to the sample group. ③ The LC-MS/MS spectra were interrogated against this target-decoy reduced database with a stringent FDR cutoff to identify the proteins. Inj.: injection. The parameters illustrated are from Zhang et al. [41].

Since then, the iterative database search has been used in a few metaproteomic studies [57, 60, 61] and starts to be included in automatic identification software solutions [101, 102]. But in spite of its increasing popularity, this strategy of interrogation has been evaluated only on small datasets (<10 samples) [60, 100], and never within the context of large-scale metaproteomics experiments.

2.3.3.4 Iterative identification used in the experiments

To interpret mass spectra, we used either a classical or our own iterative interrogation strategy. The classical strategy consisted of a one-step target-decoy interrogation of the database translated from either MetaHIT 3.3 or MetaHIT 9.9 together with the *Homo sapiens* database and the contaminant database. The *e*-value thresholds for peptides and proteins were set to 0.05 (Figure 2.2).

The iterative strategy was adapted from Jagtap et al. [99, 100] and Zhang et al. [41]. We began as in the classical interrogation strategy, but the peptide and protein e-value thresholds were set to 10 and the decoy was not used, as in the abovementioned studies. We then used the identified bacterial proteins to build sub-databases specific to each sample. These subdatabases were used in the second step, where individual MS/MS data were searched against their own subdatabase concatenated with the *Homo sapiens* database and the contaminant database, and their decoy. At this step, peptide and protein e-value thresholds were set to 0.05. In the third and final step, all bacterial proteins identified in the second step were combined into a reduced database together with the *Homo sapiens* and the contaminant databases.

Target-decoy identification was performed by using this reduced database with peptide and protein *e*-value thresholds set to 0.05. Of note, by using X!Tandem, we had to filter the identifications based on an *e*-value threshold. We computed FDR at the end of the workflow as a quality control of the identifications only, contrary to the abovementioned methods which use the FDR as a filter.

The data were searched using the X!Tandem software [22] version 2015.04.01.1. against three databases: (i) each of the two human gut microbiota protein databases computed from the two MetaHIT catalogues previously mentioned, (ii) the *Homo sapiens* database and (iii) the contaminants database. For all identifications, five types of modifications were searched: carbamidomethylation of cysteines (fixed modification), oxidation of methionines, excision of the N-term methionine with or without acetylation, excision of the 1-50th N-term amino acids, and cyclization of N-term (potential modifications). The mass tolerance was set to 10 ppm for the parent peptide and 0.02 Da for the fragments. One miscleavage was allowed. The *e*-value threshold for peptides and proteins were set to 0.05 for the classical identification and the second and third step of the iterative method (Figure 2.2 and Figure 2.5), and set to 10 for the first step of the iterative method (Figure 2.5).

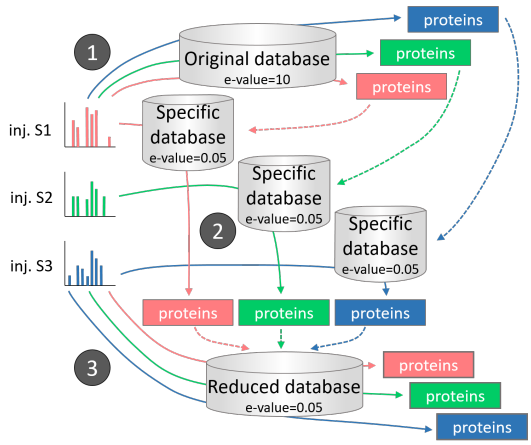


FIGURE 2.5 – Iterative strategy of interrogation implemented in our work. The steps were the same as those implemented by Zhang et al., Figure 2.4, except that we chose a threshold by e-value and the decoy search was not used in the second step.

2.3.4 Construction of the datasets

Importantly, the grouping of peptides (Section 1.3.3.2) depends on the whole set of samples. Thus, to mimic the framework of a real experiment where the sample is not replicated, we searched the MS/MS data of each S32 replicate together with the MS/MS data of the 47 non-replicated samples. The 14 independent peptide and subgroup identifications obtained for the 14 replicates of S32 constituted the replicate dataset that served to measure reproducibility in the actual context of large-scale experiments. One replicate of sample S32 was randomly selected and its MS/MS

data were searched together with the MS/MS data of the 47 non-replicated samples. The resulting peptide and protein identifications constituted the complete dataset. Finally, to ensure that the method comparison was not affected by the cohort size, five non-replicated samples were randomly selected and their MS/MS data were searched together, constituting the reduced dataset.

2.3.5 Peptide and subgroup quantification

Peptides were quantified by Spectral Counting (SC), namely the number of MS2 spectra per peptide per sample. Protein SC was subsequently computed by summing the SC of all their peptides (Section 1.3.3.3). This counting is used when looking for the presence of subgroups in the samples, which is mainly what is studied hereinafter. The subgroup SC was computed by summing the SC of the specific peptides, *i.e.* by excluding shared peptides which bear information difficult to deconvolve [31]. This counting is used when we are interested in the abundances of the subgroups.

2.3.6 Evaluation criteria

A database and a workflow define a method. To compare the identification results obtained after interrogating two different databases or using two interrogation strategies, we considered various criteria.

Diversity of the samples

First, we considered that the more identified peptides and subgroups, the better, since it estimates the diversity of the samples. We performed this comparison for each sample (equation 2.1) as well as for overall datasets (equation 2.2). Interpolation and extrapolation of the total number of peptides and subgroups identified with an increasing number of samples was performed with iNEXT [103].

For a technical sample i and method $meth$ (a method being here defined by the catalogue and workflow used), N_i^{meth} counts the number of peptides or subgroups identified exclusively by the method $meth$, and N_i^{share} the number of peptides or subgroups identified by both methods in sample i. To compare the identifications between two methods, the peptides are defined by their sequence and modification, and the subgroups are defined by the sequence of their representative protein. The subgroups are hereinafter referred to as "proteins" for simplification.

N^{meth} counts the number of peptides or proteins identified by method $meth$ in at least one sample of the dataset, but never identified by the other method in any sample of the dataset, and N^{share} the number of peptides or proteins identified by both methods in at least one sample (not necessarily the same sample for both methods).

The gain in peptides or proteins of method 2 with respect to method 1 in sample i is defined as:

$$\frac{N_i^{meth2} - N_i^{meth1}}{N_i^{meth1} + N_i^{share}} \times 100 \tag{2.1}$$

The overall gain in peptides or proteins of method 2 with respect to method 1 is defined as:

$$\frac{N^{meth2} - N^{meth1}}{N^{meth1} + N^{share}} \times 100 \tag{2.2}$$

Specificity of identification

The number of peptides and proteins specifically identified with one database or one interrogation strategy was also considered. The differences were tested with paired t-test when normality assumption was not rejected (p>0.05, Shapiro test) and paired Wilcoxon test otherwise.

Quality of identification

The peptide-level false discovery rate (FDR) was used to quantify the quality of the identifications. The FDR is defined in the Section 1.3.3.1.

Reproducibility

Lastly, the reproducibility was evaluated based on the proportion of common peptides and proteins identified in each pair of S32 replicates: the higher this proportion is, the more reproducible are the identifications. The proportion of proteins identified in the pair of replicates i, j for method *meth* is defined as:

$$P_{i,j}^{meth} = \frac{M_{share}^{meth}}{M_i^{meth} + M_j^{meth} + M_{share}^{meth}} \times 100 \tag{2.3}$$

where M_i^{meth} counts the number of proteins or peptides identified in sample i and not in sample j with method *meth*; M_j^{meth} the number of proteins or peptides identified in sample j and not in sample i; M_{share}^{meth} the number of proteins or peptides identified in both i and j.

In addition, we displayed the diversity captured by an increasing number of replicates, with interpolation and extrapolation performed with iNEXT [103]. The reproducibility of protein identifications was further evaluated by the probability to get a zero abundance for a protein in a replicate, as a function of its SC abundance observed in another replicate. For each positive integer a, we estimate the probability P_a of not identifying a protein in another replicate, given that the protein has SC a in the original technical sample:

$$P_a = P[Y_j > 0 | X_j = a] \tag{2.4}$$

with X_j and Y_j the abundance of the protein j in the original technical sample and its replicate, respectively. Noting that

$$P_a = \frac{P[Y_j > 0, X_j = a]}{P[X_j = a]}$$

P_a is estimated by replacing the numerator and denominator by their empirical counterparts, computed over all pairs of replicates.

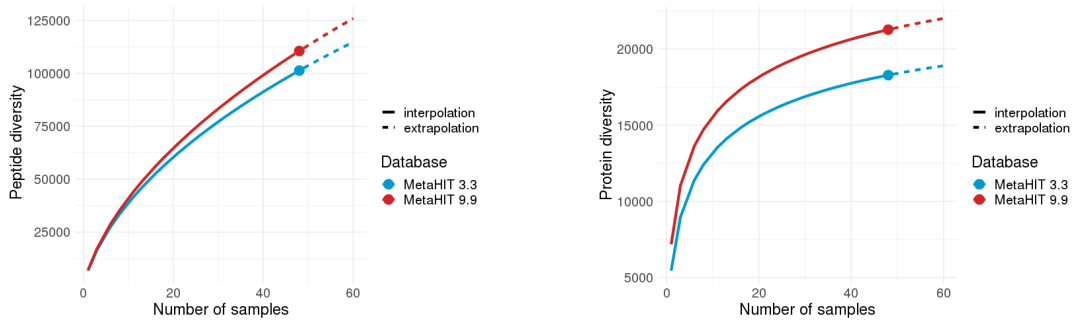
All computations were performed with RStudio version 1.1.383 and R version 3.3.3 [104].

2.4 Results

2.4.1 Gain of identification with MetaHIT 9.9

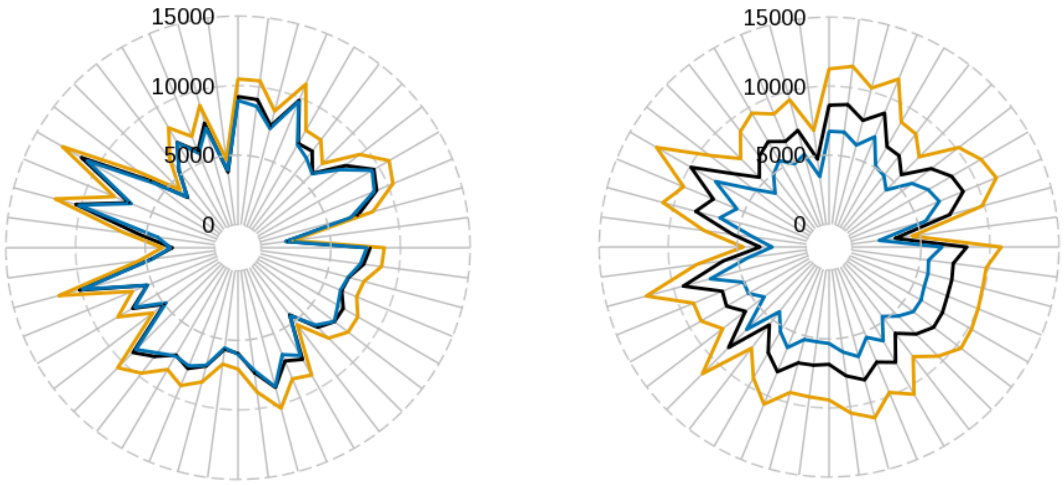
We compared the results obtained from the interrogation of the MetaHIT 3.3 and MetaHIT 9.9 databases using the classical interrogation strategy on the complete dataset which contained the identification results of the 48 stool samples.

Compared to MetaHIT 3.3, MetaHIT 9.9 gene database allowed us to identify more peptides and proteins (defined as the representative of the subgroups and counted with protein SC) overall (+9.12% and +16.32% for peptides and proteins, respectively, interpolation endpoint of Figure 2.6), as well as in most of the samples taken individually. However in this last case, the gain was much more substantial for proteins (+32.74% $\pm$ 5.56, p=1.7e^{-09}, paired Wilcoxon test) than for peptides (+2.97% $\pm$ 4.73, p=1.6e^{-07}, paired Wilcoxon test) (Figure 2.7). The between-sample variations observed in the gain brought by MetaHIT 9.9 were likely due to a difference in proteome representativeness. Indeed this database gathers more information than MetaHIT 3.3 because it was built with more individuals that come from diverse geographic origin and with whole-genome bacterial species. Of note, the difference between the two databases was lower for the samples exhibiting few identified peptides and proteins. As these samples also showed low numbers of MS2, we assume that this result is due to a low richness of the microbiome.



(A) Number of peptides (B) Number of proteins

FIGURE 2.6 – Total identifications with two databases. Number of peptides (A) and proteins (B) identified as a function of the number of biological samples from the complete dataset with either MetaHIT 3.3 or MetaHIT 9.9. Each radius corresponds to a sample. Interpolation (number of samples $\leq$48): mean over sampling of samples. Extrapolation (number of samples >48): estimations for a higher number of samples.



(A) Number of peptides per sample

(B) Number of proteins per sample

FIGURE 2.7 – Individual identifications with two databases and two strategies. Number of peptides (A) and proteins (B) identified per sample of the complete dataset with MetaHIT 3.3/classical interrogation strategy (blue), MetaHIT 9.9/ classical interrogation strategy (black) and MetaHIT 9.9/iterative strategy (yellow).

With the same *e*-value filters applied, the FDR returned by X!TandemPipeline was lower with MetaHIT 9.9 (0.046%) than with MetaHIT 3.3 (0.129%) while the search space was three times larger in the former than in the latter. If we had increased the e-value threshold in the MetaHIT 9.9 search to reach the MetaHIT 3.3 FDR, then the number of identifications with MetaHIT 9.9 would have been even higher and more drastically increased compared to the MetaHIT 3.3 identification results. Therefore, MetaHIT 9.9 gave access to a greater peptide and protein diversity at both the individual and the cohort level, while being more accurate (lower FDR).

2.4.2 Identifications specific to each database

Overall, 71.69% of the peptides identified with either database were identified with both of them. In addition, 10.41% and 17.90% were identified exclusively with MetaHIT 3.3 and MetaHIT 9.9, respectively (Figure 2.8). Therefore, although MetaHIT 3.3 is three times smaller than MetaHIT 9.9, and resulted in an overall lower rate of peptide identification, it has its own peptide matches.

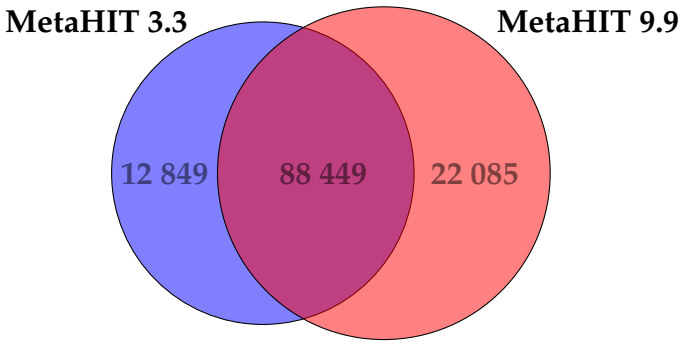


FIGURE 2.8 – Number of peptides identified by MetaHIT 3.3 (blue) and MetaHIT 9.9 (red).

Most of these peptides were not identified with MetaHIT 9.9 because they were not tryptic in MetaHIT 9.9 or because their *e*-value was higher than 0.05. However, almost a quarter (3 315 out of 12 849 identified with MetaHIT 3.3 only) were really missing in MetaHIT 9.9. Therefore, some proteins identified with MetaHIT 3.3 may have been missed with MetaHIT 9.9. Since proteins are different in the two databases, the comparison performed on peptides could not be transposed to proteins. We also verified that the number of peptides identified in each sample with only one of the two databases was higher with MetaHIT 9.9 in most samples (Figure 2.9).

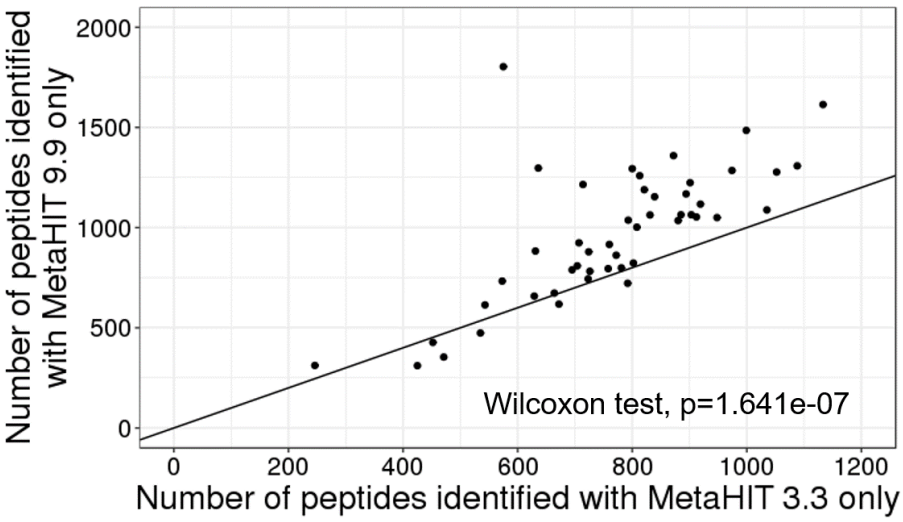


FIGURE 2.9 – Number of peptides identified in each sample specifically by one of the two MetaHIT databases. A great majority of points fell above the line y=x, indicating more numerous identifications with MetaHIT 9.9.

2.4.3 Reproducibility of the identifications with MetaHIT 3.3 and MetaHIT 9.9

The proportion of protein subgroups shared by pairs of replicates was significantly greater with MetaHIT 9.9 (p<2.2e^{-16}, paired t-test, Figure 2.10). For peptides, even if this difference was strongly significant with a paired t-test (p<2.2e^{-16}), the order of magnitude of reproducibility was similar with the two databases (Figure 2.11). Lastly, peptide detection was globally far less reproducible than protein detection, with a median proportion of shared peptides equal to 34.6% (Figure 2.11) *versus* 67.9% for proteins (Figure 2.10).

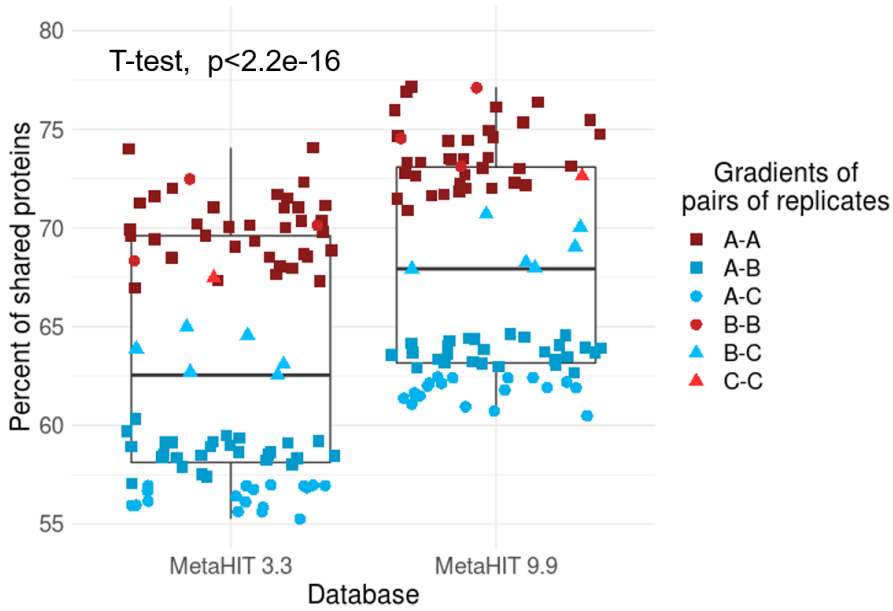


FIGURE 2.10 – Fraction of common proteins identified in each pair of replicates with MetaHIT 3.3 and MetaHIT 9.9.

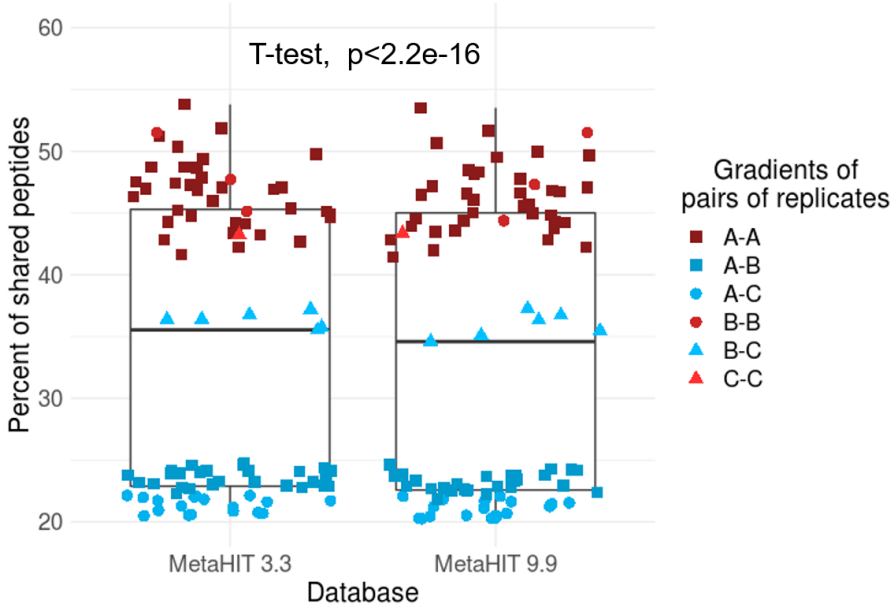


FIGURE 2.11 – Fraction of common peptides identified in each pair of replicates with MetaHIT 3.3 and MetaHIT 9.9.

Our level of reproducibility of protein and peptide identification in replicated samples fell within the range of values published earlier [105], based on a commercial mixture of 48 human proteins or a protein extract of *S. cerevisiae* repeatedly analysed with the same type of Thermo Q-Exactive instrument as ours. Thus, the reproducibility achieved in our study, where technical replicates included (i) repeated extractions of the microbial populations from raw faecal material by different staff members, (ii) subsequent LC-MS/MS analyses spread over time (including pre-column and column changes and involving different platform engineers), and finally (iii) individual grouping with 47 other biological samples (all steps based on SOPs), can be considered as highly satisfactory. Of note, the distribution of the fraction of common proteins identified in pairs of replicates displayed a bimodal distribution (Figure 2.10), with the highest values corresponding to the same sample preparations repeatedly injected, and the lowest to different sample preparations.

Furthermore, we investigated the reproducibility of protein identifications as a function of abundance (computed with subgroup SC, *i.e.* excluding the SC of shared peptides) when either sample preparation and injection or only injection were repeated. MetaHIT 3.3 and MetaHIT 9.9 displayed similar trends. As expected, the proteins identified in a larger number of replicates displayed higher abundances (Supplementary Figure 2.12). Moreover, the probability to get a zero abundance for a protein in a replicate as a function of its abundance was around 65% for a SC equal to 1 and dropped below 10% for a SC of 6 when the replication included repetition of the preparation. When the replicate originated from the same preparation, the probability was around 50% for a SC of 1 and as low as 3% for a SC of 6 (Figure 2.13). This confirms that the preparation did introduce a loss in reproducibility.

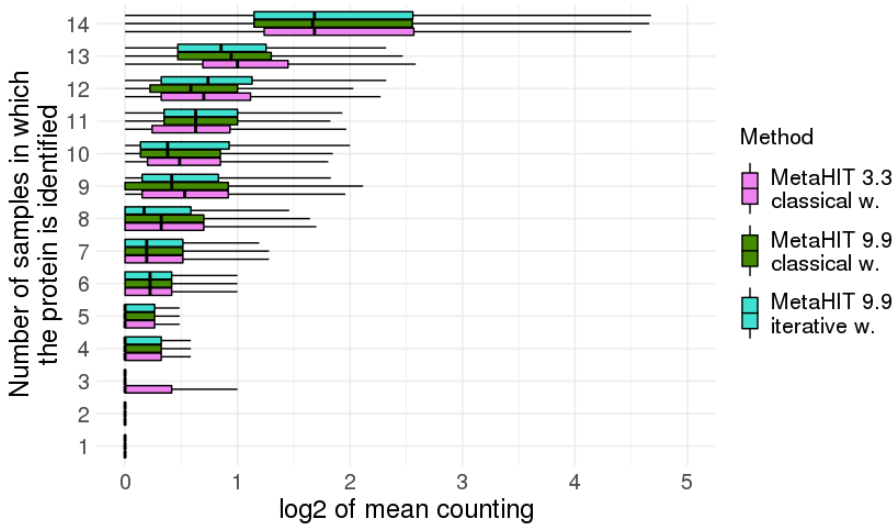


FIGURE 2.12 – Log_2 of specific spectral counts per protein, as a function of the number of replicates in which they were identified.

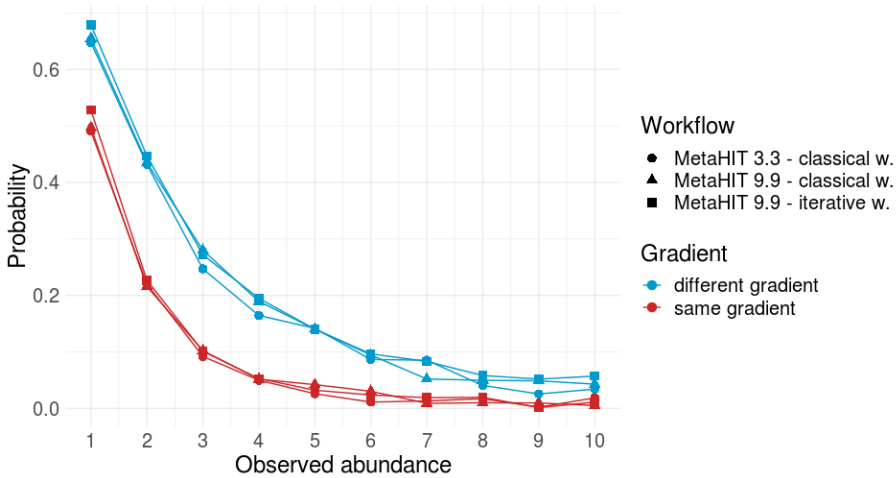


FIGURE 2.13 – Probability to get a zero abundance for a protein in a replicate, as a function of its abundance observed in another replicate. Abundance of a protein is calculated as the sum of its specific spectra.

Figure 2.14 displays the average number of proteins identified where the number of replicates increases, with an extrapolation to 20 replicates; the results for 14 non-replicated samples randomly selected from the complete dataset (excluding the replicated sample S32) are superimposed for comparison. The diversity curves did not reach a plateau, indicating the difficulty to capture the diversity of such complex samples, even with a large number of replicates. However, as expected, the diversity increased much slower when cumulating replicates rather than biological samples (Figure 2.14).

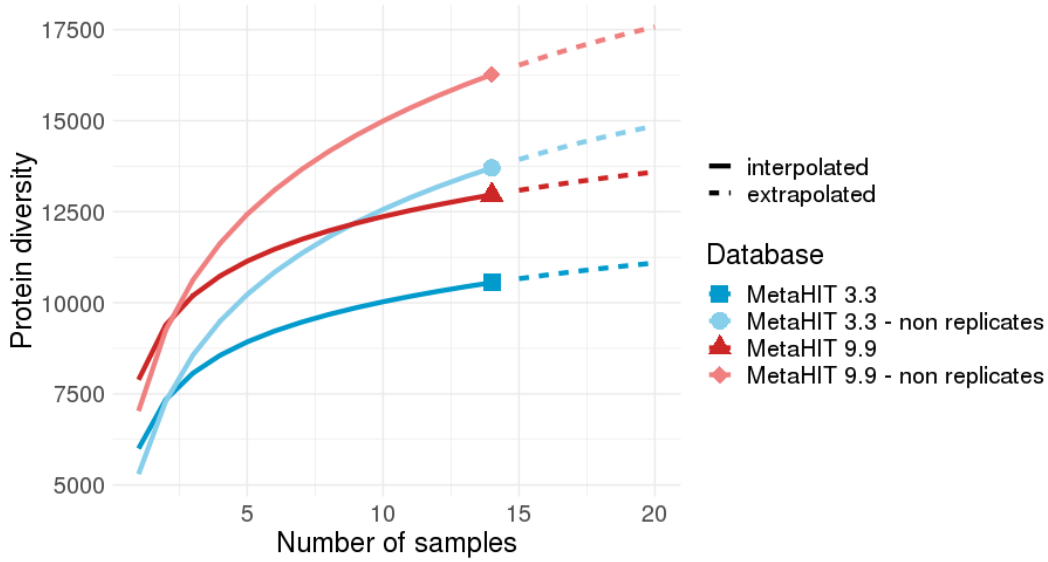


FIGURE 2.14 – Proteins diversity, defined as the average number of proteins discovered as a function of an increasing number of replicates with MetaHIT 3.3 and MetaHIT 9.9. Curves obtained for an increasing number of biological samples are also plotted for comparison.

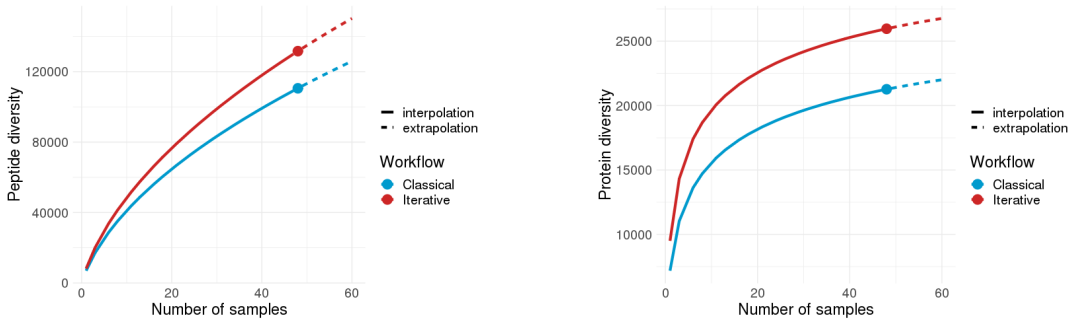
Thus MetaHIT 9.9 not only allowed us to identify more peptides and proteins, but also yielded protein identifications that were more reproducible compared to MetaHIT 3.3. Although the computation time and memory-consumption are more important when using the MetaHIT 9.9 database (three times longer in our experiment), it is the most effective database to use when searching proteins of interest in the gut microbiota. In view of these results, we used MetaHIT 9.9 as the reference database for comparison of identification strategies in the following section.

2.4.4 Gain of identification with the iterative strategy

We compared the results obtained from the interrogation of the MetaHIT 9.9 database using the classical and the iterative interrogation strategy on the complete dataset.

The iterative strategy allowed us to identify significantly more peptides (+17.56% $\pm$ 2.77 ; p<2.2e^{-16}, paired t-test) and proteins (+32.69% $\pm$ 3.07 ; p=1.7e−09, paired Wilcoxon test) in each sample (Figure 2.7).

The FDR was lower with the iterative strategy: 0.0270% *versus* 0.0458% with the classical interrogation strategy. The overall number of peptides and proteins identified within the cohort also increased when using the iterative strategy, as shown on Figure 2.15 (+19.08% and +22.09% for peptides and proteins, respectively).



(A) Number of peptides (B) Number of proteins

FIGURE 2.15 – Peptide (A) and protein (B) diversity as a function of the number of biological samples with the classical and iterative interrogation strategies. Interpolation (number of samples $\leq$48): mean of peptides and proteins identified depending on the number of samples considered. Extrapolation (number of samples >48): estimations for a higher number of samples.

2.4.5 Identifications specific to each interrogation strategy

Overall, only few peptides and proteins were identified only by the classical strategy (0.26% and 5.03% respectively of the union of total peptides/proteins identified with either the classical or iterative strategy), while a substantial number of peptides and proteins were identified by the iterative strategy only (16.24% and 22.21% respectively) (Figures 2.16).

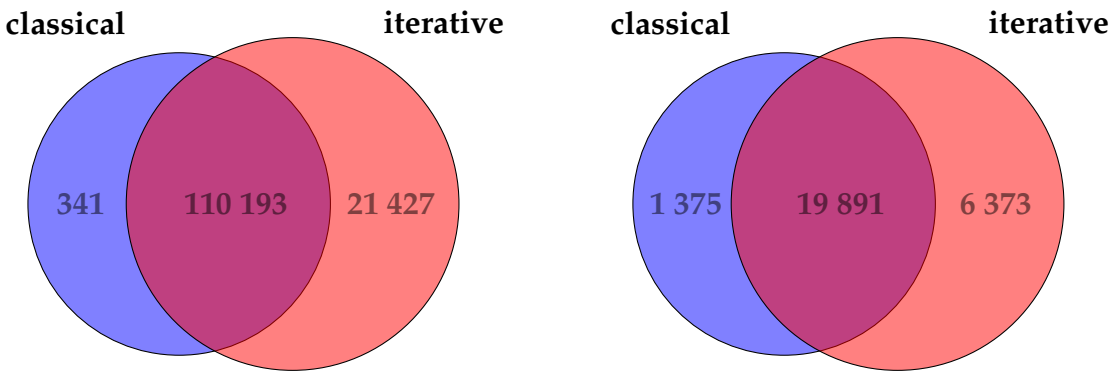
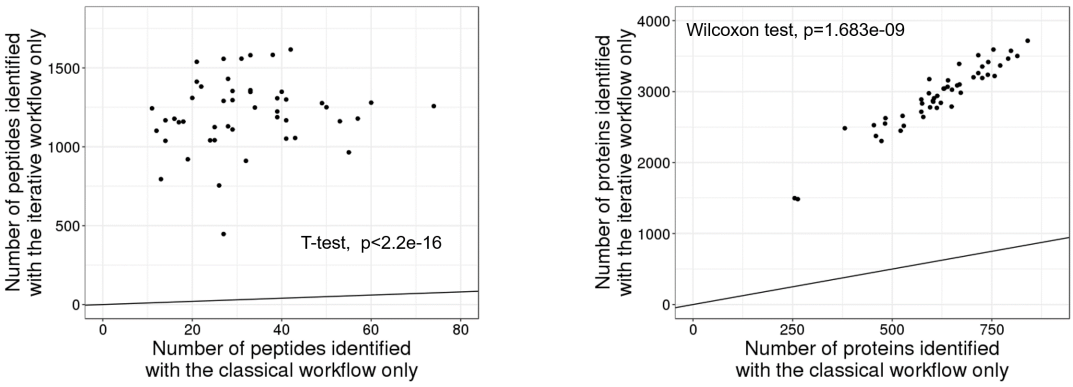


FIGURE 2.16 – Number of peptides (left) and proteins (right) identified by the classical workflow (blue) and the iterative workflow (red).

Moreover, the number of peptides and proteins per sample exclusively identified by one of the workflows was significantly higher with the iterative workflow (Figure 2.17, p<2.2e^{-16}, paired t-test for peptides, p=1.7e^{-09}, paired Wilcoxon test for proteins).



(A) Number of peptides

(B) Number of proteins

FIGURE 2.17 – Number of peptides (A) and proteins (B) identified by one of the strategies only. For all the samples, the number of peptides and proteins identified with the iterative strategy only was higher than with the classical strategy. Line: y=x.

Since the last step of the iterative strategy consisted in pooling the proteins identified in all samples, we checked if the number of samples in the experiment could have influenced the gain from the iterative strategy. Therefore, we performed the same comparisons on the reduced dataset of five randomly selected biological samples. The gain per sample with the iterative strategy was in the same order of magnitude (+22.64% and +30.13% for peptides and proteins, respectively) as the one obtained with the complete dataset. The overall gain was also similar (+22.84% and +25.52% for peptides and proteins, respectively). The gain brought by the iterative strategy was thus visible even within a small cohort, and had the same order of magnitude.

2.4.6 Reproducibility of the identifications with the classical and the iterative strategy

The proportion of common proteins identified in pairs of replicates did not significantly differ between the two strategies (p=0.53, paired t-test, Figure 2.18). Thus, while being more sensitive, the iterative strategy displayed a similar reproducibility at the protein level. Although the reproducibility of peptide identifications was lower with the iterative strategy, the order of magnitude of reproducibility was similar with the two strategies ($p<2.2e^{-16}$, paired Wilcoxon test, Figure 2.19).

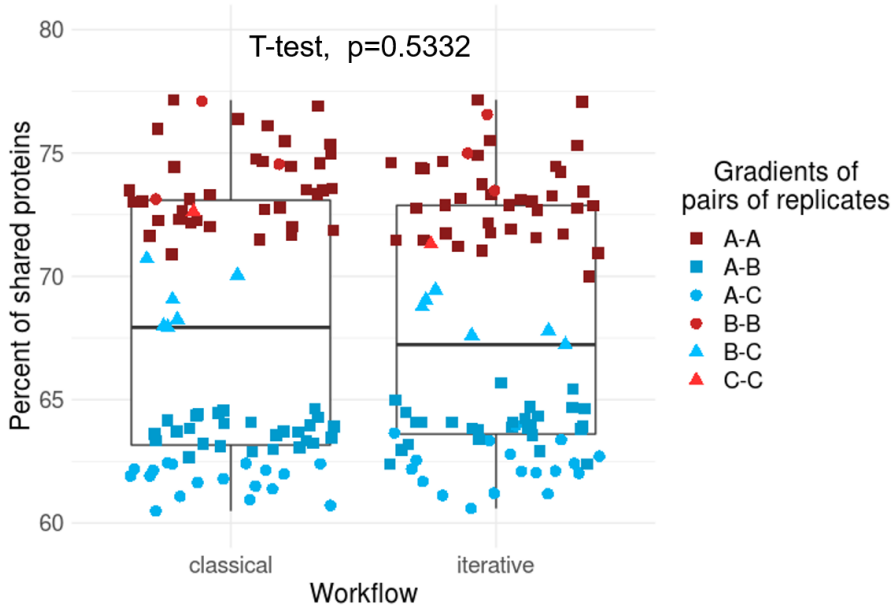


FIGURE 2.18 – Fraction of common proteins identified in each pair of replicates, with the classical and the iterative strategies.

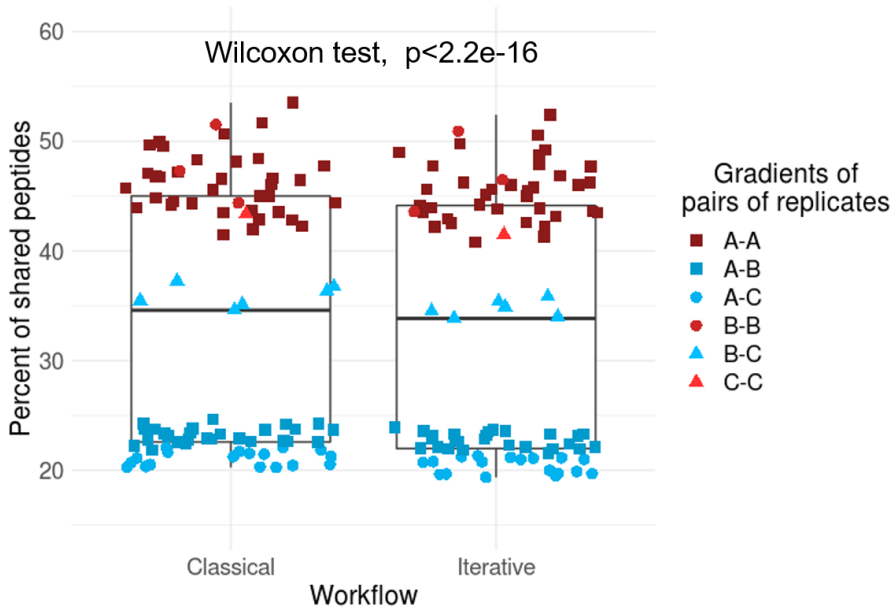


FIGURE 2.19 – Fraction of common peptides identified in each pair of replicates, with the classical and the iterative strategies.

As previously observed with the classical strategy, reproducibility of protein identifications across the 14 replicates increased with protein abundance (Figure 2.12). Also, the probability to get a null abundance for a protein in a replicate, as a function of its abundance observed in another replicate was similar to that previously reported for the classical strategy (Figure 2.13).

Figure 2.20 displays the average number of proteins identified with an increasing number of replicates, with an extrapolation to 20 replicates; the results for 14 non-replicated samples randomly selected among the complete dataset (excluding the replicate sample S32) were superimposed for comparison. Neither of the curves related to the replicated samples reached a plateau, and with as many as 14 replicates, we cannot predict which of the two strategies would reach a plateau first.

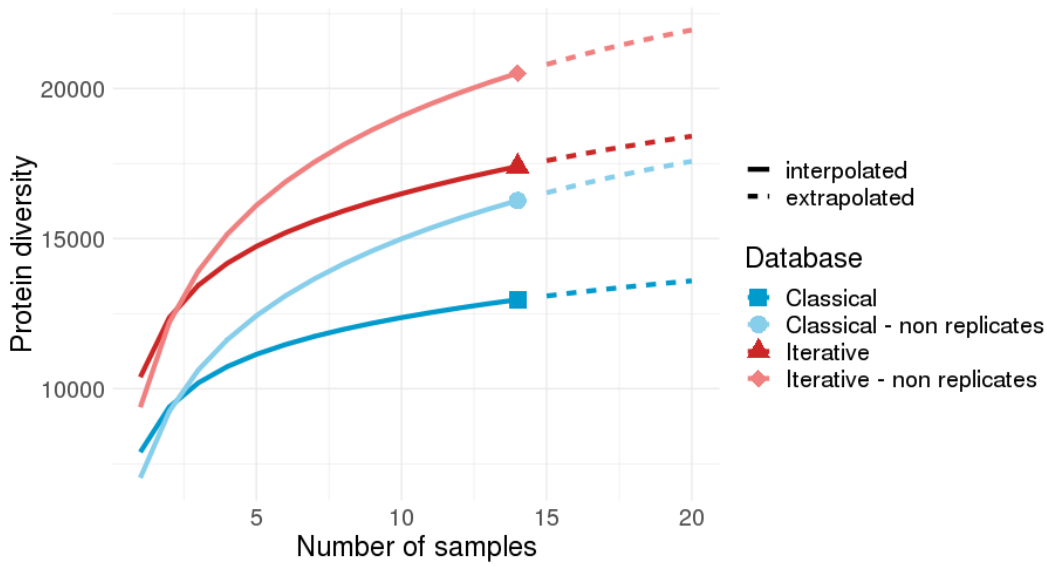


FIGURE 2.20 – Proteins diversity, defined as the mean number of proteins identified, observed with an increasing number of replicates. Curves for an increasing number of biological samples are also plotted for comparison.

Overall, the iterative strategy provided a higher number of peptide and protein identifications than the classical strategy with a higher accuracy and a similar reproducibility in large or small-sized metaproteomic datasets.

2.5 Conclusion

Based on the analysis of peptide and subgroup counting, both at the individual and the cohort level, as well as on a large number of technical replicates, we recommend to perform identification by interrogation of the MetaHIT 9.9 database using the iterative strategy. This strategy clearly gives access to the greatest peptide and protein diversity with a good reproducibility. The same conclusion holds for experiment on large- and small-sized cohorts, as demonstrated by our analysis of a sub-cohort. While a greater number of identifications could have been expected to lead to a greater proportion of false discovery, we showed that the method which brought the higher peptide and protein diversity also resulted in the lowest FDR. **This work was submitted as an original article to Journal of Proteome Research on 13/05/2019 and is currently under review.**

To achieve better results, more stringent parameters for gene clustering in MetaHIT 3.3 and MetaHIT 9.9 could be investigated. Currently set to 95% identity threshold with aligned sequence length covering over 90% of the shorter gene [98], higher values would lead to conserve a greater diversity of the genes, and therefore of the translated proteins, which is a crucial point to assign spectra in metaproteomics. In this study, the spectral assignment was mostly in the range of 30-40%, a truly honourable score when compared to pure bacterial strains (50-60%). Sometimes, metagenome sequence information is available and can be used for mass spectral interpretation [39, 41]. This was not the case in the present study, but another lead of interest would be the comparison of the iterative interrogation of MetaHIT 9.9 and the interrogation of matched individual metagenomes within the context of large-scale studies, as previously done for a few number of samples [41]. This will be assessed in Chapter 4.

Chapter 3

Two examples of clinical data interpretation with MetaHIT databases

3.1 Metaproteomic features related to weight loss

3.1.1 Scientific context

The gut microbiota is one of the multiple components which might contribute to responses after dietary intervention. Notably, gut microbiota is involved in energy harvest, lipid and sterol metabolism, and inflammation of tissues, including adipocytes [106]. In the MICRO-Obes ANR project, it was therefore hypothesized that profiling of the structure and functional potential of the host microbiota at baseline (just before starting the diet) could be useful for predicting subsequent weight loss. And in fact, using 16S rRNA gene-targeting metagenomics, a higher number of the bacterial members from the *Lactobacillus*, *Leuconostoc* and *Pediococcus* group was reported in subjects who lost less weight [87], while shotgun metagenomics revealed that low gene counts was associated with aggravation of diverse biological parameters predisposing to obesity [86].

Here, we used the ObOmics SC computed in the precedent chapter by iterative interrogation of MetaHIT 9.9, to look at the relationship between the gut metaproteomic profiles of 48 subjects just before they start a slimming diet, and their weight trajectory over the next twelve weeks. We searched, by a simple association study, if some metaproteomic features could have predicted the weight trajectory of the subjects.

3.1.2 Methods

The 48 overweight volunteers of the ObOmics study accepted to go on a well-controlled energy-restricted (approximately 1 100 kcal per day), high-protein diet for six weeks. Then for an additional six weeks, the patients were subjected to a period of stabilization with a diet of about 1 400 kcal per day. Their weight was measured at the beginning of the experiment, 7 days after the beginning of the experiment, at

the end of the first period (t = 6 weeks) and at the end of the second period (t = 12 weeks). Their stool samples were collected at the beginning of the experiment. The methods for stool sample collection and processing are fully developed in Appendix C.

3.1.3 Results

With iterative interrogation of MetaHIT 9.9, we identified 131 620 peptides and 26 264 subgroups, whose abundance were approaches by summing the SC of their specific peptides. Spearman's correlation with weight loss over the observation period ranged from -0.53 to 0.57. We first selected subgroups whose correlation with weight loss was greater than 0.4 or less than -0.4 for at least one observation time. A list of 140 proteins were selected (Figure 3.1).

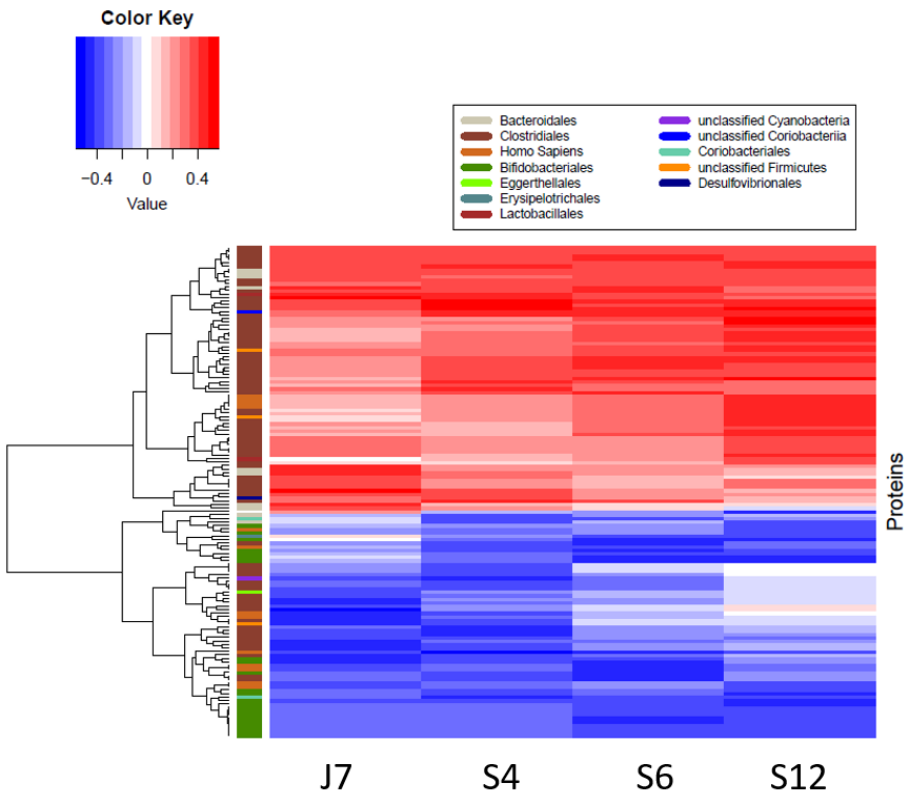


FIGURE 3.1 – Heatmap of proteins positively or negatively correlated with weight loss at a timepoint of the study.

The taxonomic annotation (Appendix C.5) of these proteins showed that most of the proteins positively correlated with weight loss belong to the order Clostridiales. Interestingly, among the proteins correlated with weight loss, all those belonging to the order Bifidobacteriales were found to be negatively correlated.

We then selected the subgroups whose correlation with weight loss was accentuated during the course of the dietary intervention. Twenty seven proteins were

eligible, of which 4 were negatively correlated and 23 positively correlated with weight loss. Among those 23 proteins, 4 were from human origin. The remaining 19 were essentially members of the phylum Firmicutes and, to a lesser extent, of the genus *Prevotella*. Three of the four proteins with increasing negative correlation with weight loss over time, were members of the genus *Bifidobacterium*, the latter being unclassified (Figure 3.2).

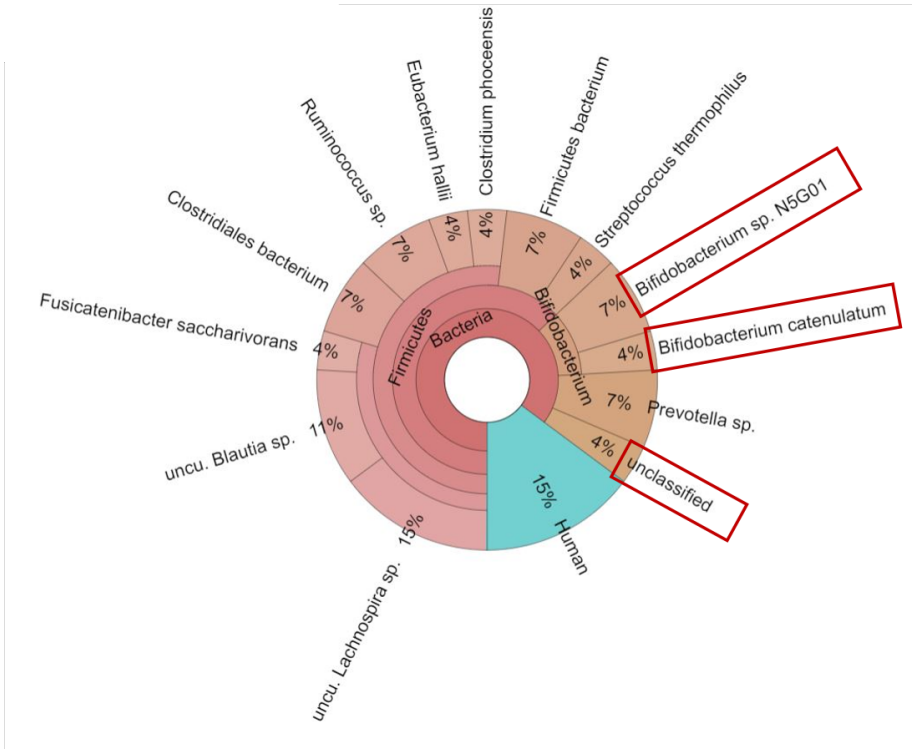


FIGURE 3.2 – Taxonomic distribution of the 27 proteins whose correlation (positive or negative) with weight loss was accentuated over time. The proteins negatively correlated are framed. The Krona chart was computed with KronaTools [107].

The functional annotation (Appendix C.5) of the 27 proteins revealed proteins involved in amino acid metabolism and carbohydrate metabolism. The three proteins negatively correlated with weight loss and belonging to the order of Bifidobacteriales were annotated as formate C-acetyltransferase, raffinose/stachyose/melibiose transport system substrate-binding proteins, and large subunit ribosomal protein L13. Interestingly, the four positively correlated proteins from human origin are monoamine oxidases in the Uniprot database. The inhibition of these enzymes which catabolize monoamines in the brain and the gut is known to induce weight gain [108].

3.1.4 Conclusion

We provided evidence that, even with a basic correlation analysis, we can extract from a highly complex metaproteomic dataset, a coherent whole from a taxonomic

and functional point of view, to answer key public health questions. With more sophisticated statistical analysis like the one presented in the last chapter of this manuscript, we should be able to target the most relevant metaproteomic predictors of weight loss and proneness to weight regain (perspectives of the work, Section 8). Provided that these candidates were confirmed by independent techniques, this could help in the development of strategies for customized nutrition intervention.

3.2 Metaproteomic features related to intestinal bowel diseases

3.2.1 Scientific context

Intestinal Bowel Diseases (IBD) are a set of chronic inflammatory disorders of the intestinal tract. They include Crohn's disease (CD) and ulcerative colitis (UC), two autoimmune diseases that develop in relapses. Both have similar symptoms, such as bloody diarrhoea, significant weight loss and abdominal pain. Although these two diseases have very similar symptoms, they affect the digestive tract in different ways.

Ulcerative colitis is characterized by superficial and continuous ulcerations of the colonic mucosa. In the period of remission, the mucosa may look normal. In flare-up, the clinical phenotype is heterogeneous, which makes this disease difficult to diagnose [109].

Crohn's disease is also characterized by ulcers, but also fistulas of the mucosa. Unlike UC, the inflammation is discontinuous and all parts of the intestine can be affected, although it is more common in the ileum. Inflammation is transmural, which means that it affects different layers of the intestinal wall, while UC is limited to the mucosa [110].

The correct diagnosis of these diseases permits to adapt the pharmacological treatment specific to each IBD and its severity [111]. Importantly, 70 to 75% of CD patients require bowel resection when symptoms become life-threatening (intestinal perforation, refractory bleeding), when only 25 to 30% of UC patients undergo this surgery [109, 111]. Improving an early diagnosis allows more effective medical care of the patients.

Although the causes of IBD are poorly understood, it has been observed that the patients' intestinal microbiota showed a reduction in the diversity of *Firmicutes* and *Bacteroidetes*. Efficiency of antibiotics as well as studies in mice show that the host-microbiome interaction could be a triggering factor for IBD [112]. Today, diagnosis of IBD requires blood tests, endoscopic and imaging procedures (Computerized tomography scan to search for perforated colon, magnetic resonance imaging to evaluate fistula). The objective of the MICI-Pep study was to assist in the diagnosis of IBD phenotypes thanks to specific intestinal metaproteomic traits of patients, with a simple stool collection, instead of invasive investigations. To our knowledge, IBD

markers has never been searched at the metaproteomic level, except in one recent study focused on the mucosal-luminal interface of paediatric IBD patients [61].

3.2.2 Methods

The main problematic of the IBD diseases being the differentiation between the IBD phenotypes (CDC for Colic Crohn's disease, CDIC for Ileo-Colic Crohn's disease and UC for Ulcerative Colitis), we were interested in highlighting proteins whose abundance differed between healthy controls and IBD patients. Sample collection and processing of this study is presented in Appendix D.

The stool samples (n=40; 8 controls, 7 UC, 3 CDC and 2 CDIC, each fresh and frozen) were collected and analysed as developed in Appendix D. In this study, focus was on the envelope-enriched metaproteome as the first line of interaction with the host mucosae. Considering the circumstances where the patients are far from the diagnosis centre, we were interested in the discovery of markers that could be used for diagnosis from either fresh or frozen samples. The envelope-enriched metaproteome was thus extracted twice for each sample, either freshly collected or after freezing at -80°C for two months. All analyses were performed on a single batch of 40 samples on an Orbitrap Fusion^{TM} Lumos^{TM} Tribrid^{TM}. We implemented the classical interrogation of MetaHIT 3.3 and compared the results with those obtained with the iterative interrogation of MetaHIT 9.9.

3.2.3 Results

3.2.3.1 Metaproteomic profiling of stool samples

Using the classical interrogation of MetaHIT 3.3 concatenated with the human proteome database, we identified a total of 190 893 peptides and 31 348 subgroups across all samples. These numbers were increased up to 231 500 (+21.3%) and 43 521 (+38.8%) by iterative interrogation of MetaHIT 9.9 plus the human proteome database, thus confirming on another dataset analysed with a higher resolution spectrometer the valuable potential of this strategy for metaproteome deciphering. Of note, the FDR was the same for the two database interrogations (0.0023% with MetaHIT 9.9, 0.0026% with MetaHIT 3.3). Therefore, we will use MetaHIT 9.9 to mine the MICI-Pep data in the following.

We first compared the metaproteomic landscape of fresh and frozen samples, based on abundances of all the 43 521 subgroups in the 40 samples. The correlation matrix on Figure 3.3 shows a strong correlation ($r > 0.9$) in all samples but two, between metaproteome profiles obtained from either fresh or frozen aliquots. The low correlation ($r = 0.56$) observed for sample S09 comes from contamination of the fresh bloody sample, but not the settled frozen one, by erythrocyte proteins that we could identify as highly abundant in the dataset. Unsupervised clustering of samples confirmed that pairs of fresh and frozen samples were closely related (Figure

3.4). Therefore, we considered that the same candidate biomarkers could apply to fresh as well as frozen samples, which were grouped together in the following.

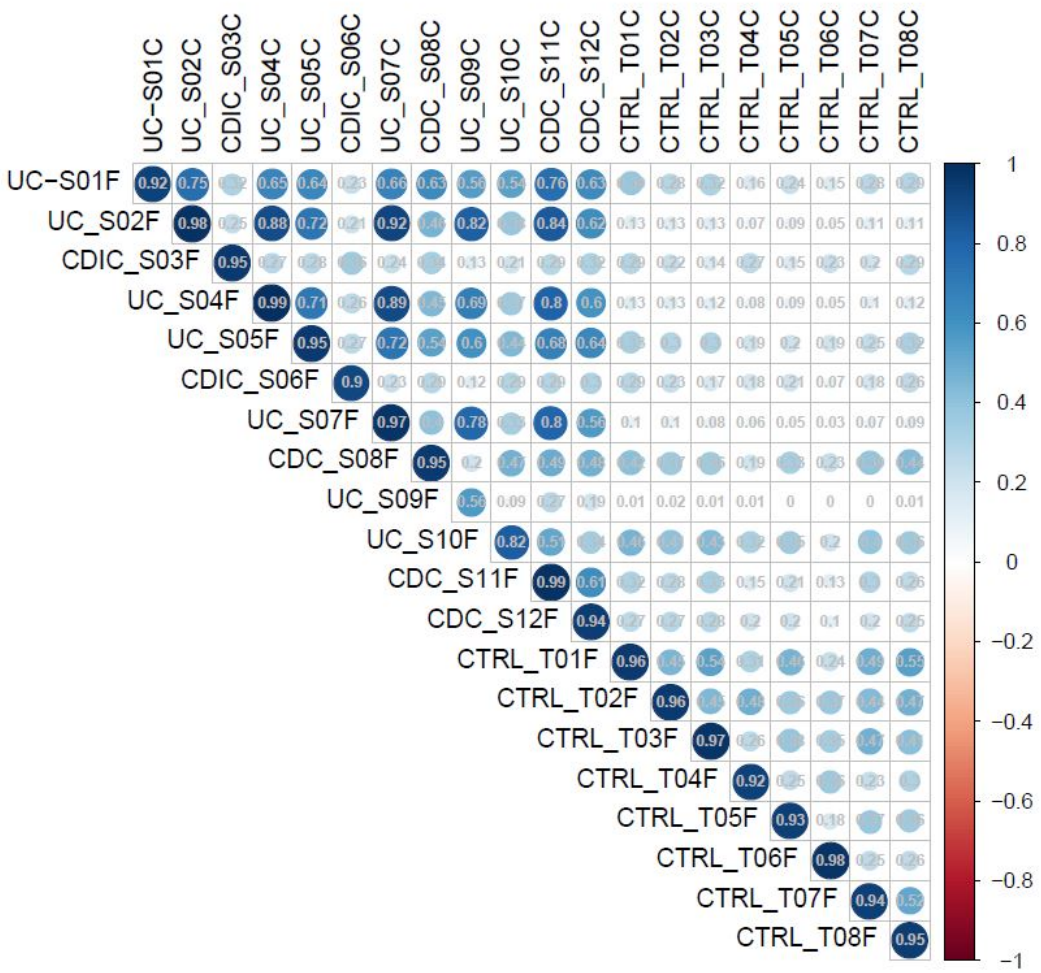


FIGURE 3.3 – Pearson's correlation between metaproteomic profiles (abundance of each of the 43 521 subgroups) of fresh and frozen samples. CTRL: controls; CDC: Colic Crohn's disease; CDIC: Ileo-Colic Crohn's disease; UC: Ulcerative Colitis.

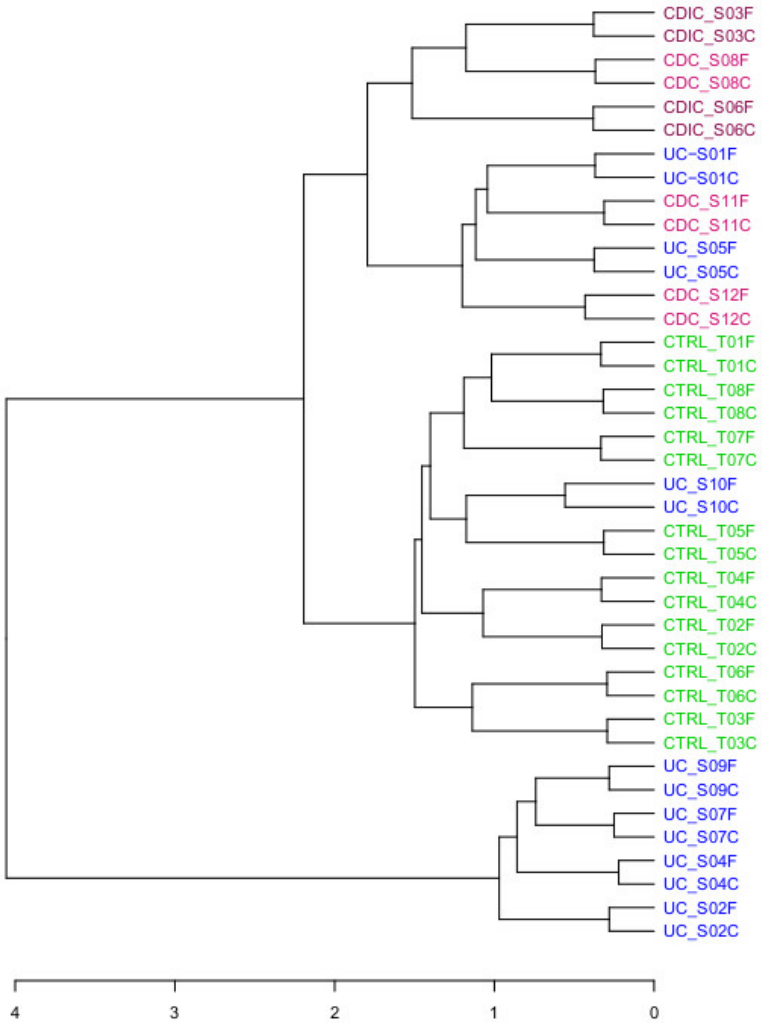


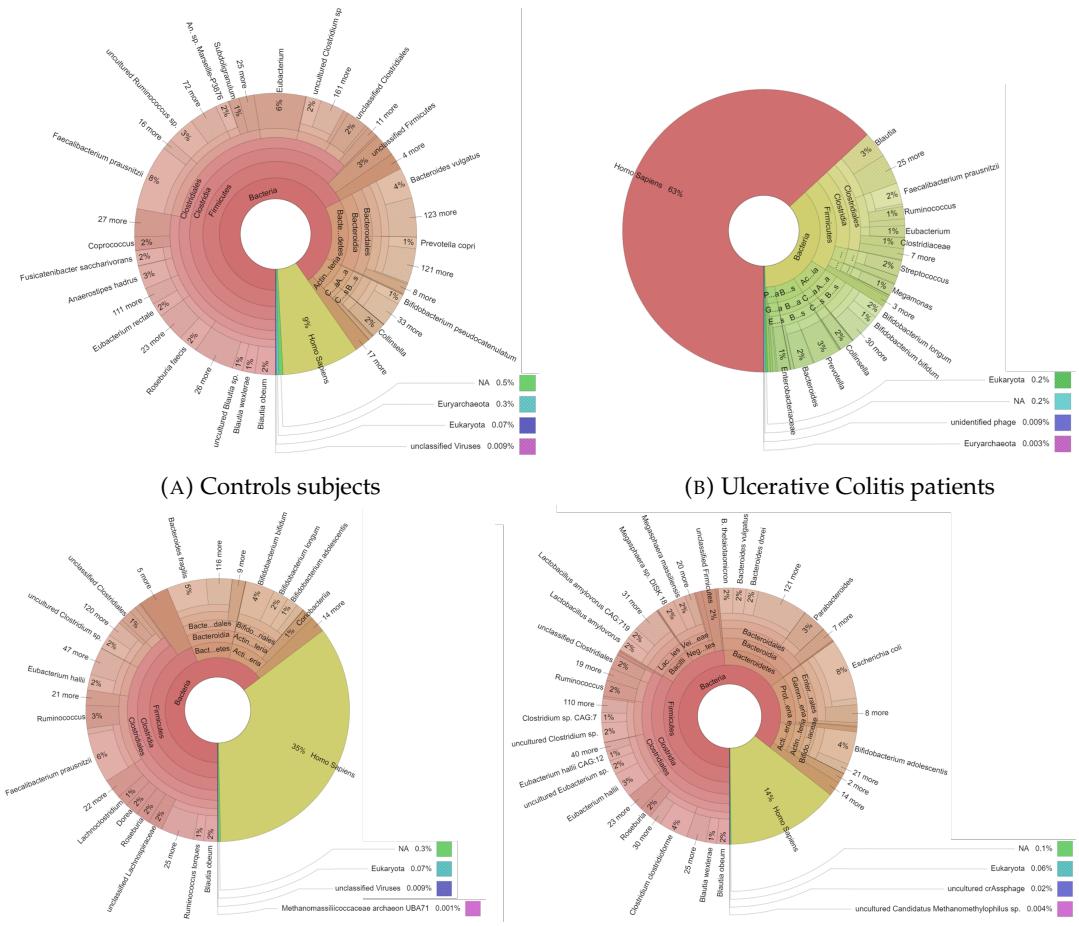
FIGURE 3.4 – Clustering of samples based on their subgroup counting. CTRL: controls; CDC: Colic Crohn's disease; CDIC: Ileo-Colic Crohn's disease; UC: Ulcerative Colitis.

Table 3.1 below summarizes taxonomic and functional diversity of proteins per sample within each group of subjects, revealing a loss of diversity in the three patient groups compared to the controls. Of note, this loss was highly heterogeneous in the UC samples (ranging from 4 655 to 22 022 subgroups per sample). The proportion of human proteins as well as the total number of bacterial species and KEGG annotation are reported in the Table 3.1. The present study is the first demonstration of a loss in diversity of IBD microbiomes at the metaproteomic level.

TABLE 3.1 – Proteins diversity with functional and taxonomic annotation in the MICI-Pep samples.

Group	Number of samples	Proteins per sample	% human	Bacterial species	KEGG orthology (bacterial)
CTRL	16	22 592.88 ± 449.23	1.94	2 770	1 714
UC	14	13 657.50 ± 6 436.61	8.12	2 696	1 523
CDC	6	19 210.00 ± 2 224.04	5.87	2 548	1 211
CDIC	4	17 758.50 ± 2 289.51	2.69	2 516	1 328

A detailed taxonomic distribution of the microbiome subgroups, also including contribution of bacteria-coating subgroups from human origin, is given by the Krona charts on Figure 3.5.



(A) Controls subjects (B) Ulcerative Colitis patients

(C) Colic Crohn's Disease patients (D) Ileo-Colic Crohn's Disease patients

FIGURE 3.5 – Taxonomic distribution of subgroups in the envelope-enriched fractions of microbiota of the MICI-Pep volunteers, based on abundances of subgroups (computed as the sum of SC of their specific peptides).

Interestingly, while human protein species were in a minority (Table 3.1), their contribution in term of abundance, estimated by the sum of their specific spectral counts, was much higher, from 9% in the controls up to 63% in the UC patients (Figure 3.5). Pancreatic and intestinal enzymes were in the top list of the most abundant proteins in all groups. A subset of seven human proteins, including calprotectins and other serine proteases were in the top list of patients only. Although we cannot disclose their detailed list, all of them are linked with inflammation, regulation of autophagy and innate immunity, or have an antimicrobial activity.

3.2.3.2 Search for IBD signatures in stool samples

Given the low number of samples, we applied a highly stringent selection of candidate biomarkers and retained only subgroups that were strictly over or underrepresented in all samples (either fresh or frozen) from a group of subjects compared to all samples from another group. To protect confidential information, only a restraint

naming of the candidate markers are provided in the following. We first searched subgroups which may segregate controls and all IBD patients. We identified seven subgroups from human origin for which SC were systematically in a greater abundance in patients than in controls (Figure 3.6).

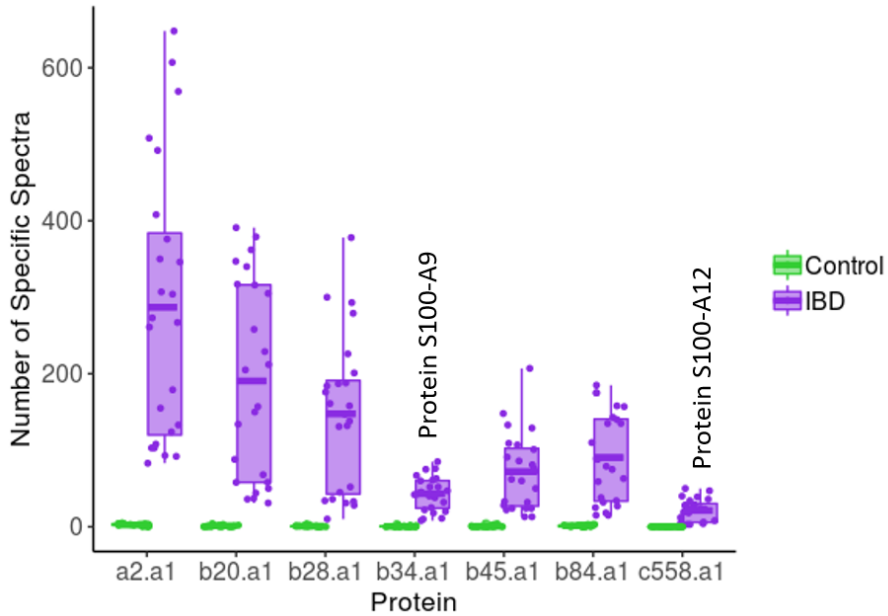


FIGURE 3.6 – Proteins overabundant in IBD patients compared to controls.

Two of them (b43.a1 and c558.a1) are the well-known neutrophil-derived proteins S100-A9 and S100-A12 (also called calprotectins), that have a strong antibacterial and antifungal properties. Calprotectins are activated when inflammation occurs (for whatever reason). Thus, when abdominal symptoms exist, dosage of faecal calprotectin can be used to identify an inflammatory bowel condition and determine the next course of action in diagnosis and treatment. Importantly, faecal calprotectin is a useful and cost-effective marker to help differentiate between IBD and IBS (Irritable Bowel Syndrome), but does not differentiate between different IBD phenotypes, as also proved for the first time by our metaproteomics approach.

Interestingly, we found five additional immune cell-derived proteins that were even more increased than calprotectins in all patients compared to all controls. They are all related to host defence against bacterial infections, and some of them were reported to support the differentiation of chronic IBD from IBS and correlate with the severity of IBD inflammation.

Therefore, our findings well reflect the release of a number of proteins by activated and degranulated immune cells, which is consistent with exacerbation of the host defence system in IBD. This clearly demonstrates that gut metaproteomics can be a powerful tool for relevant marker discovery, not only from bacterial but also from human origin.

Looking for proteins overexpressed in control samples revealed one protein from bacterial origin, which was absent in all patients but detected at low levels in all controls.

3.2.3.3 Search for signatures between IBD phenotypes

A main purpose of this research was to find proteins that could differentiate between CD and UC, and ideally between CDIC, CDC and UC. As CDIC is the most serious IBD phenotype, we started with the selection of the proteins whose abundance was systematically increased or decreased in all CDIC samples compared to all CDC or UC samples. We could identify 101 proteins, which were specifically overabundant in CDIC samples (Figure 3.7). Among these, 97 were from bacterial origin, shared between the three phyla Proteobacteria (n = 68, most of them from *Escherichia coli*), Firmicutes (n = 20, most of them from *Clostridium clostridioforme*) and Bacteroidetes (n = 9). The four remaining were host enzymes involved in the lipid metabolic process. Of note, 87 of these 101 abovementioned proteins also emerged when CDIC samples were compared to all other samples, including the controls. They are delineated above the horizontal line on Figure 3.7. Interestingly, we found only one host protein which was specifically less abundant in CDIC samples compared to CDC and UC samples. This protein is known to regulate intestinal epithelial cell survival in response to pro-inflammatory stimuli.

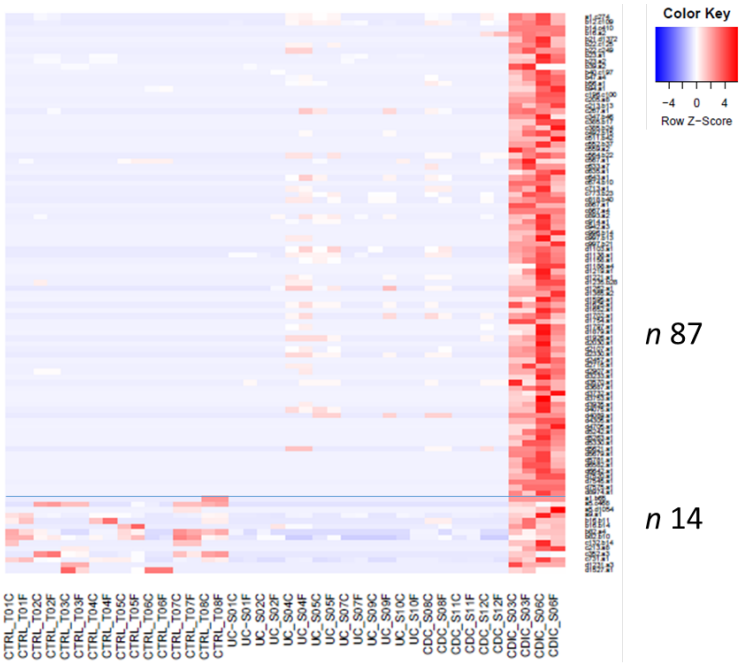


FIGURE 3.7 – Heatmap of 101 subgroups overabundant in CDIC samples compared to all other IBD samples. Over these subgroups, 87 of them are also overabundant compared to controls. The data are log-transformed and zero-centered.

CDIC samples therefore differ from CDC and UC samples, or even from controls, by an invasion of bacterial proteins mainly from *Escherichia coli*, and to a lesser extent from *Clostridium clostridioforme*. Invasion of a number of opportunistic pathogens such as *E. coli* and *C. clostridioforme* was already reported in CD patients, but it was based on metagenomic shotgun sequencing and the comparison did not distinguish between ileocolic (CDIC) and exclusive colonic (CDC) localization of Crohn's disease [113, 114].

We are now at the point where we can suspect an inflammatory bowel condition based on a group of seven abundant immune cell-derived proteins and where we can reasonably suspect a Crohn's disease with ileocolic localization based on an invasion of protein entities from *E. coli* and *C. clostridioforme*. We still have to distinguish between CDC and UC. We identified six proteins that were more abundant in all CDC samples compared to all UC samples. Four of them were from *Faecalibacterium* species (referred to as a1.c148, b78.a1, c278.b28 and d3555.a1 on Figure 3.8), and the two others were pancreatic proenzymes (referred as to c113.a1 and c113.a2 on the same figure). We found no protein specifically less abundant in all CDC samples compared to all UC samples.

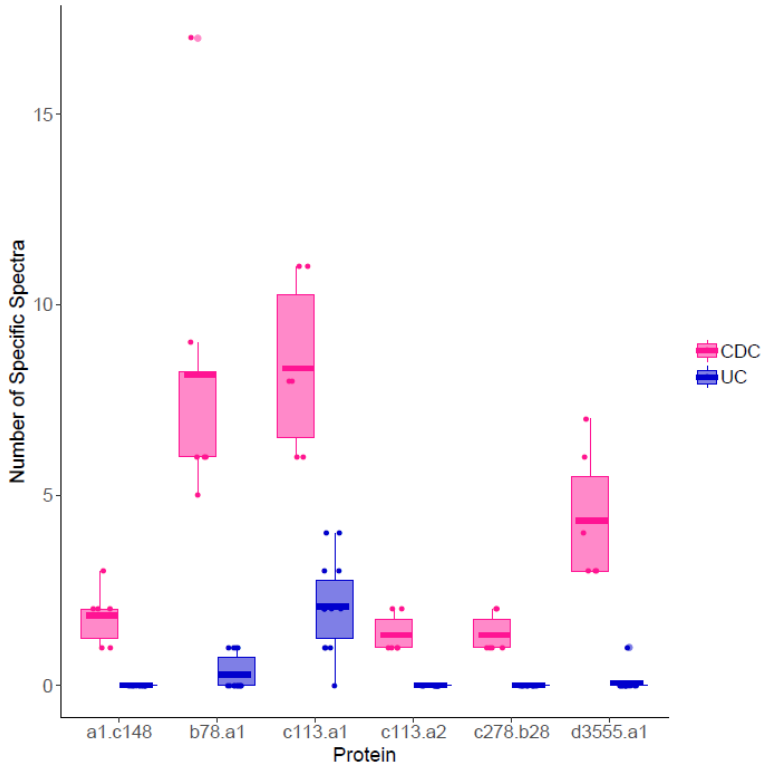


FIGURE 3.8 – Proteins overabundant in all CDC compared to all UC samples.

3.2.4 Conclusion

Based on metaproteomic analysis of the envelope-enriched fractions of microbiota extracted from some IBD patients and controls, together with a highly stringent selection of subgroups that were specifically either over- or underrepresented in the different IBD phenotypes, we can propose leads for earlier diagnosis of these different inflammatory bowel flares. A gut metaproteomics-based decision tree is reported on Figure 3.9 below.

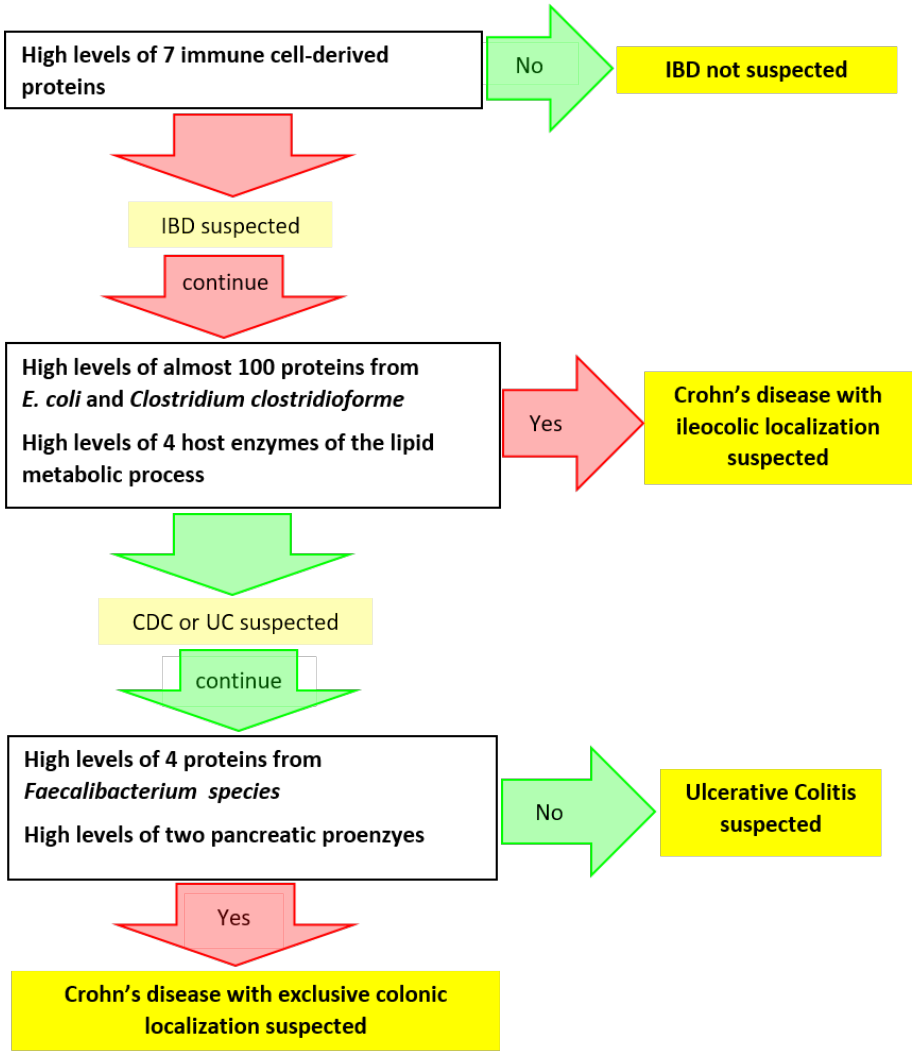


FIGURE 3.9 – Tree of decision for diagnostic of IBD based on our metaproteomic analysis of microbiota envelope-enriched fractions.

Chapter 4

Mass spectra interpretation in the context of the ProteoCardis cohort

The choice of a suitable database is not trivial because we have to make a trade-off between the size of the database and its completeness (Section 1.3.5). Without an individual metagenome, as in the case of the ObOmics and MICI-Pep studies, the general databases of human gut microbiomes, MetaHIT 3.3 and MetaHIT 9.9, perform well for the identification of metaproteomics mass spectral data. However, when the self-metagenome of the sample is sequenced, we hypothesized that it should be the database of choice, since it combines a reduced size and sequences specific to the sample. In this chapter, we were interested in the assembly of individual metagenomes, that we used to interpret the corresponding metaproteomes. We measure performances of this strategy compared to the use of the generalist database MetaHIT 9.9. The comparison lies on the sequencing reads of individual metagenomes acquired from 188 patients enrolled in the FP7 MetaCardis European program and selected for the ProteoCardis ANR program. The ProteoCardis project included two cohorts, whose metagenomes were sequenced by shotgun metagenomics and metaproteomes were analysed by LC-MS/MS. The first cohort included 138 individuals with acute or chronic heart disease as well as 50 healthy individuals. The second cohort was composed of 30 severe obese patients who were observed before and after bariatric surgery. Therefore a total of 248 stool samples were collected for the analysis of their individual metagenomes and metaproteomes.

4.1 Methods

4.1.1 Metagenomics sequencing

Total DNA was extracted from stool samples of all individuals on the SAMBO platform of MetaGenoPolis. The DNA were fragmented by endonuclease and the fragments size threshold was here set to a minimum of 150 base pairs (bp). The sequencing was performed on an Ion Proton sequencer (Ion Torrent^{TM} technology), giving a minimum of 20 millions reads per metagenome.

MetaRaptor was developed for the filtering, assembly and gene prediction of the 248 individual metagenomes. Although construction of metagenomes is done routinely at MetaGenoPolis, the context of our study where metagenomes are intended to metaproteomics data interpretation, adjustments of the parameters used were required. Indeed, the identification of peptides has technical limitations due to the calculation time and the memory used for this step, both increasing with the size of the database used. Furthermore, the database has to be simultaneously reduced in size to ensure a good sensitivity, and sufficiently exhaustive to identify accurately the peptides. This last requirement is particularly challenging in metaproteomic context due to the extreme complexity of the samples and the possible mutations affecting the bacteria. As a result, each step involved in the construction of individual metagenomes is critical in this work.

4.1.2 Metaproteomic analyses

Sample preparation is identical to the preparation of the MICI-Pep samples, detailed in Appendix D.2 and D.3. Briefly, the microbiota were extracted through a gradient, and the cytosolic fractions and envelope were separately analysed. This separation permits to enrich the envelope fractions, whose proteins are in direct interaction with the host, but which are usually underrepresented in whole-cell preparations. Extraction was randomized using a block design on patient groups, by dividing each group of patients equally in the batches of preparation.

LC-MS/MS analyses were performed as for the MICI-Pep study (Appendix D.4). The injections were also randomized thanks to an efficient block design, hence a potential batch effect would not lead to bias in the differential analyses between groups of patients. A cleaning of the mass spectrometer with a blank was performed between each injection, and a deep cleaning was performed between each batch to ensure a high sensitivity all along the injections (a batch including 18 to 22 injections of samples). Eight of the one hundred and eighty-eight subjects from the first cohort were injected seven times for a study of reproducibility; two from the aCAD group, two from the cCAD group, two from the cCADic group and two controls. The samples from the second cohort (bariatric surgery) were not replicated, but the samples from the same patient (before and after bariatric surgery) were injected in the same batch to avoid a potential batch effect. To evaluate the potential batch effect, a sample from a subject which is not included in the study was prepared and injected at each end of a batch. This replicated sample, injected fourteen times, is called a standard.

All the study design and preparation followed in advance the guidelines proposed by Zhang and Figeys in July 2019 [9], in particular independent randomization of sample preparations and injections and a standard sample run throughout the experiment. Bioinformatics development and presentation also anticipated the guidelines and recommendation of this paper: a clear reporting of the databases used, the strategies of database search, the definition of specific peptides, shared

peptides, subgroups and groups, and the addition of human protein database for identification.

4.2 Development of MetaRaptor

MetaRaptor is a software solution which perform in-parallel assembly of many metagenomic samples from reads to general catalogue. I is dedicated to bacterial analysis and can be used in both metagenomic context or single-cell experiment. It is accessible to naive users, but also modifiable by experienced users for implementing future NGS methods and analyses. Here we present key software solutions that make up the MetaRaptor steps.

Quality filtering

The first steps of MetaRaptor is to remove primers at reads' ends, and nucleotides and reads of poor quality. The software AlienTrimmer [115] detects and eliminates primer sequences located at the ends of the reads. The number of primer sequences used being very limited, AlienTrimmer eliminates these sequences in a few minutes, based on the detection of k-mers. This software also contains features to eliminate poor quality reads, calculated using the Phred score associated with each nucleotide. It starts with the elimination of reads with a too high proportion of nucleotides with a low quality score. Then, it removes low quality nucleotides that may occur in 5' or 3' ends. Finally, the reads too short after these filtering are eliminated. In the context of ProteoCardis, I eliminated reads with more than 40% nucleotides with a Phred score of less than 20. This cutoff score was also used to eliminate poor-quality nucleotides at the ends of the reads. Finally, reads having a length less than 100 nucleotides were removed. All of these parameters were previously benchmarked in the team, and were recognized as the parameters allowing the better quality of assembly.

Contaminant filtering

The removing of host's genome contamination constitutes the second step of MetaRaptor. For that goal, MetaRaptor uses the software Bowtie2 [116]. This software aligns reads to long reference genomes. This alignment step, which is usually very long, is extremely fast and memory-efficient with Bowtie2, thanks to the use of indexes. The contaminant genomes must therefore be indexed before launching MetaRaptor. Bowtie2 also lets the creation of indexes with a small memory footprint, based on a Burrows-Wheeler transformation. It is thus a software particularly adapted to the search for contaminated reads in a metagenomic context. The user can eliminate reads belonging to one or more organisms (pig, mouse, chicken, ...) by tuning the parameters as long as they have the complete contaminant genome(s). In the context of ProteoCardis, since we are working on stool samples, there is a high contamination by human DNA on sequencing reads. I therefore remove reads that aligned on human DNA.

Optimization of reads data

A software which aims to reduce the redundancy of reads has been included in MetaRaptor: Khmer [117]. Indeed, due to the considerable amount of reads sequenced by NGS technologies, the downstream assembly is highly time-consuming. Khmer proposes to eliminate redundant reads while retaining sufficient information for the assembly, based on the recognition of k-mers. Although Khmer was an option included in MetaRaptor, I did not use it in the context of ProteoCardis since the parameters was not yet benchmarked.

***De novo* assembly**

For the assembly step, we included in MetaRaptor the Spades software [118], an algorithm that perform assemblies with different k-mer sizes and merge the results to obtain an optimal assembly. This algorithm showed high performances compared to other assembly softwares [119, 120]. For ProteoCardis, I used the parameters of k-mers used conventionally.

Genes prediction

Several predictors can be used to predict bacterial genomes, such as Prodigal [121]. This software uses a dynamic programming with a log-likelihood function to compute scores for prokaryotic gene recognition and initiation of translation sites identification, with the goal to reduce the false positive rate. It is particularly accurate for gene prediction on assembled contigs, and is fast and easy to use. Although it has poor performances on contigs with less than 200bp [122], it shows high accuracy on high-GC-content sequences, which is not the case for other gene prediction software solutions [121]. I therefore chose this gene predictor for this thesis work.

Genes filtering

After assembly and gene prediction, several filters select the higher quality genes (in-house scripts). MetaRaptor can remove short genes (whose threshold is determined by the user) and incomplete genes. For ProteoCardis, I first eliminated genes whose length was less than 60 nucleotides. The proteins corresponding to these genes will then have a size greater than or equal to 20 amino acids. This reduces the size of the database by eliminating proteins that are unlikely to be identified by mass spectrometry. Indeed, the probability to identify short proteins is low since few number of peptide ions (6-40 amino-acids residues) can be detected in those proteins. Then, I eliminated the proteins whose start and stop have not been identified. Although Prodigal uses a probabilistic algorithm for the detection of bacterial genes, and therefore performs well even if start and/or stop are missing, I chose to eliminate the genes whose neither the start nor the stop were detected by Prodigal, because the detection of these genes is still less reliable compared to sequences with a start and/or a stop.

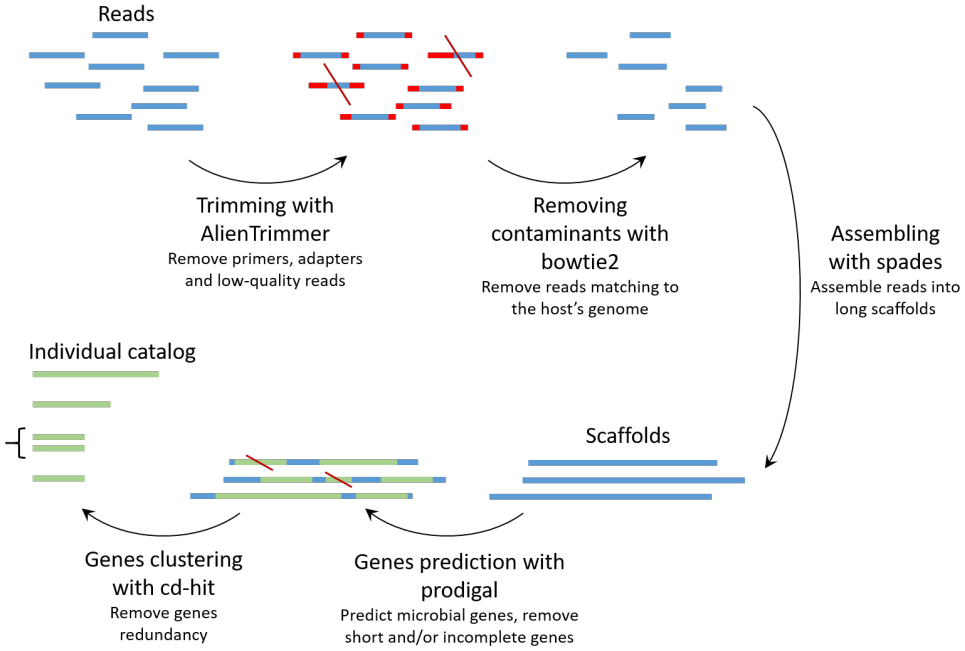
Genes clustering

After assembly and the different filtering steps, clustering of the genes in each sample reduces the size of the database by eliminating gene redundancy. Classically, the clustering parameters used are 90% coverage of the smallest sequence and 95% percent identity between the two sequences. The algorithm cd-hit [123] is widely used to eliminate the redundancy of metagenomic catalogues [5, 61, 124]. Based on k-mers indexation of the sequences to compare, cd-hit can quickly cluster a large number of sequences. The sequences are first sorted according to their length, and the longest sequence becomes representative of the first cluster. Then a new sequence to be clustered is compared to the representative of this cluster. If this new sequence has a percentage of identity greater than a given threshold, it is part of the cluster, the longest sequence remaining representative of the cluster. If it is not the case, it forms a new cluster of which it is the representative. This algorithm thus accelerates the clustering by avoiding the comparison of each sequence with each other, since the comparison is made only with the representative of each cluster. By default, the cluster representative is the longest sequence of the cluster. The clustering result is the set of cluster representatives.

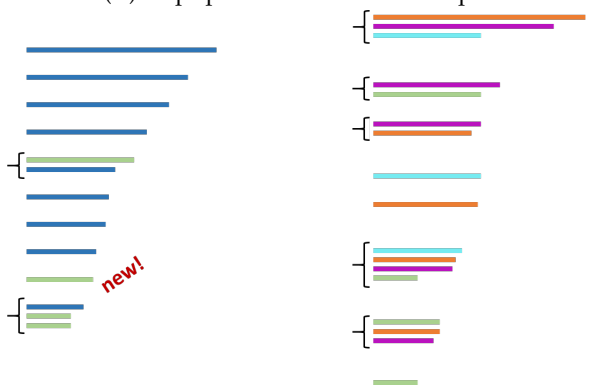
In the case of proteomics, identification is sequence dependent. That is, if the peptide sequence in the database differs by only one amino acid from the sequence whose spectrum has been recorded by LC-MS/MS, the spectrum cannot be interpreted. Therefore, elimination of the redundancy is not desirable in the particular case of proteomics. Hence, I chose to cluster the genes with 100% sequence identity and 100% coverage of the smallest sequence. As a result, only genes whose sequence was exactly included in another one were clustered (and thus eliminated from the list of genes in the sample), while all others that could contain Single Nucleotide Polymorphisms (SNPs) were preserved.

All these steps are performed in-parallel for each sample of a metagenomic study. The tuning parameters for each software solution included in MetaRaptor are modifiable by the user. After all the samples have been processed into individual catalogues, the last step of MetaRaptor is to compute (i) an enrichment of a preexisting catalogue or (ii) a new catalogue representative of the metagenomic samples studied. This computation is a clustering (i) of individual catalogues with the preexisting catalogue or (ii) of the individual catalogues together. For ProteoCardis, I chose to compute a brand new catalogue, with the clustering of all individual samples with the most stringent parameters cited above.

The performance of MetaRaptor with the parameters used for ProteoCardis was validated on a mock community, whose results are presented in Appendix E. All the steps detailed above are summarized on Figure 4.1.



(A) Steps performed for each sample



(B) Merging of results to enrich preexisting catalogue or to create general catalogue

FIGURE 4.1 – Schema of the Metaraptor's steps. (A) Each read file of the samples is treated simultaneously and construct an individual catalogue from the reads. (B) The individual catalogues can enrich a pre-existing catalogue (left, not used in this thesis) or create a catalogue compiling the results of the samples of interest (right, here four samples represented with colours).

4.3 Assembly of the individual metagenomes of the ProteoCardis cohort

In consideration of the results obtained on the mock community (Appendix E), I assembled the metagenomes of the 248 microbiomes of ProteoCardis with the same parameters as those validated hereinbefore and presented in the methods. I thus obtained a catalogue of genes for each of the 248 ProteoCardis samples, containing on average 238 121.9 $\pm$ 70 819.71 genes (Figure 4.2). N50 and L50 are statistics widely used in metagenomics to evaluate the quality of assembly. The N50 is the size of the contig which, along with the larger contigs, contains half of the total number of nucleotides assembled. The L50 is the smallest number of contigs whose length sum contains half of the total number of nucleotides assembled. The mean N50 of the scaffolds was 3 454.34 $\pm$ 2 145.04 and the mean L50 was 9 206.91 $\pm$ 5 011.93.

Once each sample has been assembled, the last step in constructing a gene catalogue is the gene clustering of all samples to produce a non-redundant catalogue representative of all the samples. Here again, in our context, we need to conserve all the genetic variability. The goal is simply to eliminate the genes whose sequence is exactly included in another longer sequence. However, despite the speed of the algorithm cd-hit used for gene clusterization, the huge number of genes (more than 53 millions of genes) exceeded the possible computation time. So I eliminated the genes which had exactly the same sequence with a home-made python script. This reduced the number of genes to 40 millions, but the resulting catalogue was too wide to be used for mass spectra interpretation with realistic computation times, so I only used individual catalogues in the work presented hereafter.

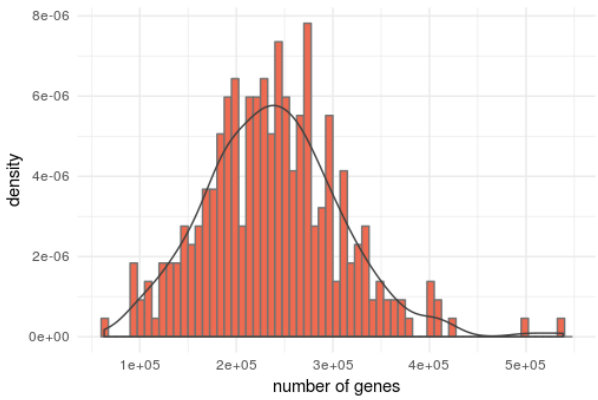


FIGURE 4.2 – Distribution of the number of genes for the ProteoCardis stool samples.

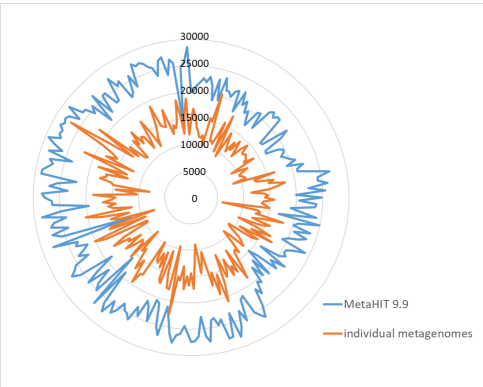
4.4 Performance of the individual catalogues

We assessed the performance of mass spectra interpretation based on self-metagenome interrogation compared to MetaHIT 9.9 interrogation, on a subset of the ProteoCardis cohort. A total of 236 MS/MS datasets corresponding to the cytosolic fraction of the 188 non-bariatric cohort and the 7 replicates from 8 of these samples were used to compare coverage of the metaproteome by each of the method.

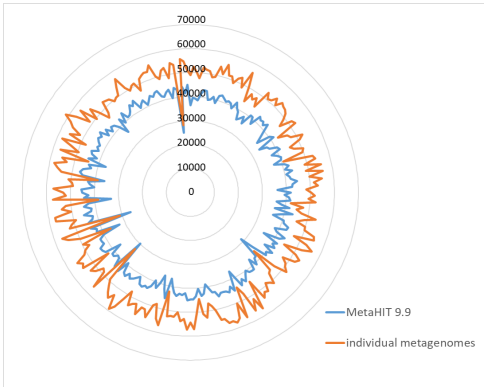
For each of LC-MS/MS datafiles, I interrogated the own individual metagenomic database translated into proteins, together with the *Homo Sapiens* protein database and the common contaminants protein database. The interrogation was in one step with the use of the decoy database and a stringent peptide e-value of 0.01 in order to reduce the number of peptides and proteins discovered in the context of this assessment. I then performed a grouping of the 236 identification files taken together, and eliminated peptides and proteins shared with the contaminant database.

I applied the same parameters for querying the 236 MS/MS datasets against MetaHIT 9.9, *Homo Sapiens* and contaminant database, and performed the grouping of all identification files taken together.

The criteria used to compare the performances of the two approaches were the number of peptides and subgroups identified. The subgroups being represented by their representative protein, they are referred as "proteins" hereinafter, and defined by the sequence of their representative protein. The peptides were defined by their sequence and modification. We took into account the total number of peptides and proteins identified across all samples, as well as the number of peptides and proteins identified per sample. The results of the number of peptides and proteins identified for each sample are presented on Figure 4.3.



(A) Number of peptides



(B) Number of proteins

FIGURE 4.3 – Number of peptides (A) and proteins (B) identified per sample with MetaHIT 9.9 (blue) or the individual metagenomes (orange), in 236 cytosolic metaproteomes of the ProteoCardis cohort. Each radius corresponds to a sample, for which the number of peptides (A) or proteins (B) identified are indicated in orange or blue depending on the database considered.

The MetaHIT 9.9 catalogue always identified more peptides than the individual catalogue within each samples (average increase of 70.3%), with a lower FDR (0.016% with MetaHIT 9.9, compared to 0.101% with the individual databases). Considering all the samples together, MetaHIT 9.9 identified 551 503 peptides against 463 115 with the individual databases, *i.e.* a gain of 19% only, due to the redundancy of peptide identifications between samples.

In most metaproteomics and proteomics studies using algorithms other than X!tandem, the identification results are filtered based on a fixed FDR [40, 60, 125]. In the particular case of the X!Tandem algorithm, filtering is on a defined e-value for peptides, while we can calculate the FDR simply to assess the overall quality of identifications. Here, if we could have imposed the same FDR for searches in MetaHIT 9.9 and individual databases (which is not trivial), the gain with MetaHIT 9.9 would have been even much greater.

However, a significant number of peptides were identified only with the individual catalogues (14.3% of all the peptides identified with the two databases, Figure 4.4). I suggest four hypotheses to explain this result:

- These peptides are indeed absent from MetaHIT 9.9 due to the non-redundancy of the catalogue
- These peptides had a higher e-value with MetaHIT 9.9 that exceeded the threshold
- The mass spectra were matched to a different peptide with MetaHIT 9.9 due to a lower e-value
- The peptides are non-tryptic in MetaHIT 9.9 (their N-terminal peptide is neither arginine nor lysine) and therefore their sequence is discarded from the search, because only tryptic peptides are considered for the *in silico* digestion of the reference database

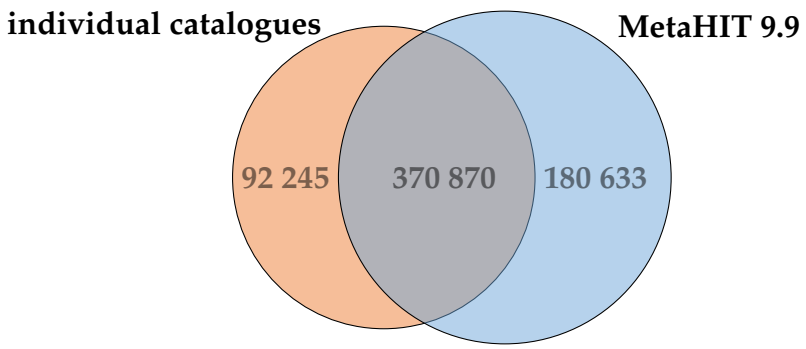


FIGURE 4.4 – Venn diagram of the number of peptides identified by MetaHIT 9.9 and the individual catalogues, all samples taken together.

Due to the high complexity and specificities of individual gut microbiota, a high proportion of the peptides identified with individual databases only are probably absent from the MetaHIT 9.9 database. A simple motif research showed that, of these 92 945 peptides, 68 587 were indeed absent from MetaHIT 9.9. The individual metagenomes thus contain peptide and protein valuable information that are not found in the non-redundant generalist database.

Although MetaHIT 9.9 catalogue identified more peptides than individual catalogues, it identified fewer proteins. I first thought that this result could be explained by a better coverage of MetaHIT 9.9 proteins by more peptides identified with this catalogue. However, I discarded this hypothesis by verifying that the coverage of proteins was similar for both catalogues. A second hypothesis was that a number of proteins of the individual catalogues are fragments of longer proteins in MetaHIT 9.9. In this case, a set of peptides would identify only one protein in MetaHIT 9.9 but several of its fragments in the individual catalogues, thus artificially increasing the number of proteins identified. I computed the length of the proteins identified with the two catalogues (taking into account all the proteins contained in the subgroups and not only the representative), and a Wilcoxon test showed that they were indeed significantly different in size (p-value <2.2e-16), MetaHIT 9.9 including a lot more long-sized proteins (Figure 4.5).

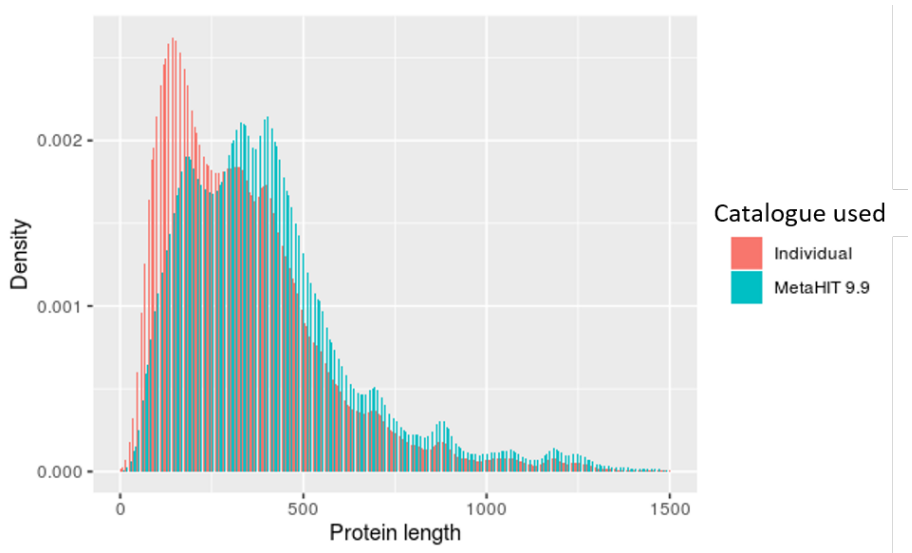


FIGURE 4.5 – Length of the proteins identified with the MetaHIT 9.9 catalogue and the individual catalogues.

To support this hypothesis, I blasted all the bacterial proteins of the catalogue of one individual on the MetaHIT 9.9 database. I chose the individual database of the patient for whom the number of proteins identified with each databases differed the most. I set the e-value threshold at 10^{-2}. Out of the 62 689 bacterial proteins identified in this sample (including all the proteins in the subgroups) with the individual catalogue, only 1 554 blasted against 9 189 proteins identified with MetaHIT

9.9 (also including all the proteins in the subgroups). Out of these 9 189 proteins against which the individual proteins blasted, 76.4% of the proteins were blasted by only one protein from the individual catalogue. The hypothesis that the individual catalogue proteins are fragments of the MetaHIT 9.9 base proteins is therefore not sufficient to explain the results. A deep examination of the peptides and proteins and comparison of the results obtained for one sample would help us to better explain this phenomena.

The subgroups identified with MetaHIT 9.9 are nevertheless identified with more peptides than with the individual catalogues, as illustrated on Figure 4.6 and confirmed by a Wilcoxon test ($p<2.2e^{-16}$). We computed this result by taking into account only the representative protein of each subgroup.

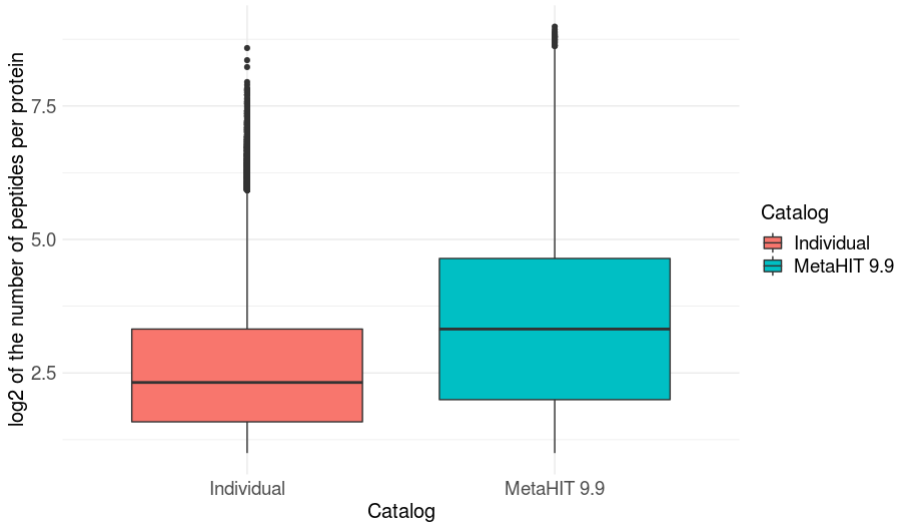


FIGURE 4.6 – Number of peptides per subgroup for subgroups identified with MetaHIT 9.9 and the individual catalogues. Log scale.

We then identified peptides and proteins within each sample by combining the MetaHIT 9.9 database and its individual database. We computed this identification on 40 samples and compared the results with those obtained using either MetaHIT 9.9 or the individual database alone. This combination brought only 3.5% of supplementary peptides compared to the identification with MetaHIT 9.9 only. Moreover, the grouping of the 40 samples took 10 minutes with the individual databases only, 50 minutes with MetHIT 9.9 only, and 280 minutes (more than 4 hours) with the combination, which makes it impossible to implement with more than the 200 samples of the ProteoCardis cohort. This explosion of computation time is probably due to the higher complexity of data. To reduce complexity and computation time, we could try to first combine MetaHIT 9.9 and the individual databases and remove the redundant sequences, and then use this concatenated database in an iterative search approach. Such a strategy was shown to enrich MetaHIT 9.9 with 6.9% more genes

in a recent study [61] including 28 individual mucosal-luminal interface gut metagenomes. However, the gain in peptide and protein identifications compared to the use of MetaHIT 9.9 only was not documented in this work.

The results obtained with the individual catalogues thus showed us that, although these catalogues were generated to represent as accurately as possible the metagenome of the subjects, they were less effective than the generalist database in terms of identification of peptides. This may be due to an insufficient sequencing depth of the samples, which did not capture the entire diversity of individual metagenomes. We can also consider improving the assembly by cleaning up potential sequencing errors in reads, an issue that has not been addressed in my thesis work. Finally, the use of the generalist database seemed the most reliable for characterizing metaproteomes of the ProteoCardis samples

4.5 Metaproteome landscape in cardiovascular diseases

The deciphering of intestinal metaproteomes and the accompanying challenging search for metaproteomic signatures of this disease were the ultimate goal of my thesis work. Thanks to the results obtained above, we interpreted all MS/MS data of the ProteoCardis cohort and the standards (a sample outside the study, injected at each end of the LC-MS/MS batches) by iterative interrogation of MetaHIT 9.9 database concatenated with the *Homo sapiens* database and the contaminants database. In the particular case of bariatric datasets, in addition to these three databases, the human oral database [126], containing more than 3 millions of sequences, was added, since bypass of the stomach and part of the intestine makes it possible that oral bacteria are present in stools. In all cases, peptide e-value threshold was set at 0.05 (except for the first step of interrogation, where the threshold was set to 10), and peptides and proteins with less than 2 counts across all samples assembled within a same dataset were discarded.

Four datasets were generated, including all replicated samples and standards :

- non-bariatric cytosolic metaproteomes (188 individuals and replicates/standards, corresponding to 250 LC-MS/MS files)
- non-bariatric envelope metaproteomes (188 individuals and replicates/standards, corresponding to 250 LC-MS/MS files)
- bariatric cytosolic metaproteomes (30 individuals at two timepoints and standards, corresponding to 74 LC-MS/MS files)
- bariatric envelope metaproteomes (30 individuals at two timepoints and standards, corresponding to 74 LC-MS/MS files)

The number of peptides and proteins identified in the samples as shown in Table 4.1, where the protein SC was considered.

TABLE 4.1 – Results of iterative interrogation of MetaHIT 9.9 in the ProteoCardis cohort.

	non-baria cyto	**non-baria env**	**baria cyto**	**baria env**
number of peptides	294 397	352 090	254 547	260 748
number of subgroups	57 044	64 107	48 424	48 400
mean number of peptides per sample	8 408.68 ± 1 256.61	8 029.70 ± 1 597.60	15 908.35 ± 1 830.80	13 110.99 ± 2 699.83
mean number of subgroups per sample	12 206.95 ± 1 441.77	10 753.90 ± 1 638.23	18 325.14 ± 1 880.47	14 703.85 ± 2 204.21
FDR	0.0123%	0.0172%	0.0106%	0.0193%

The identification of the peptides and subgroups in all the ProteoCardis samples opens a window on the metaproteome landscape in cardiovascular context. In the following section, we removed the identifications of the 14 standards and we randomly selected one of the seven replicates for each of the eight replicated samples. Peptides and subgroups identified in these samples only were consequently removed from the datasets.

Overall, we observed 54 913 and 61 629 subgroups throughout the 188 non-bariatric cytosolic and the 188 non-bariatric envelope-enriched individual metaproteomes, respectively. They were inferred respectively from 281 011 and 332 491 peptides. The functional annotation by KEGG showed that 16% and 22% of the cytosolic and envelope (respectively) fraction subgroups were not annotated. Subgroups annotation showed that some functions were identified only in one of the fractions (cytosolic or envelope, Figure 4.7). Functions identified in the envelope fractions only would probably have been missed if the processing of the samples would not include a fractionation and enrichment of this subcellular compartments. Figure 4.8 confirmed that peptides and proteins identified in each fraction are mostly different; fractionation therefore allows for a higher number of peptides and proteins identified in the downstream bioinformatic analyses.

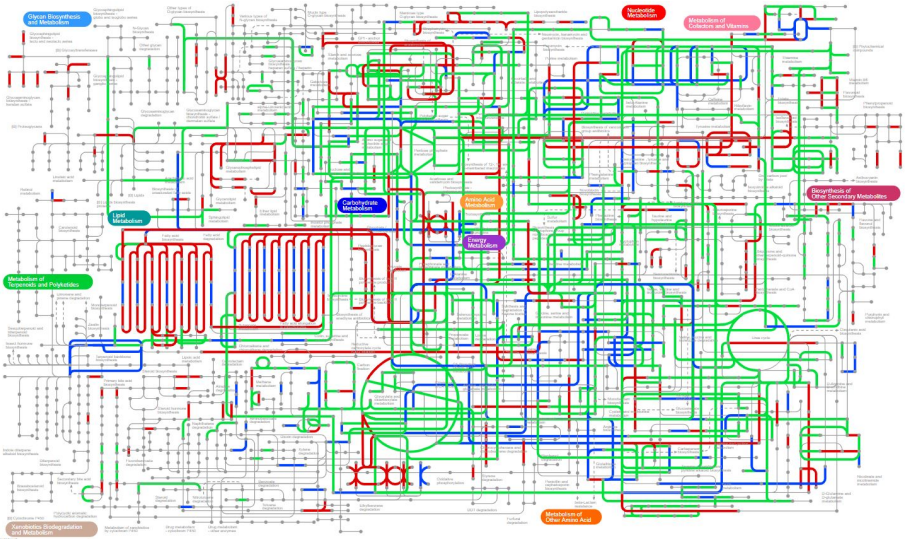


FIGURE 4.7 – Metabolic pathways identified in each fraction of the ProteoCardis samples. Blue: cytosolic fraction only. Red: envelope-enriched fraction only. Green: both fractions. This diagram was computed on KEGG annotations with iPath [127].

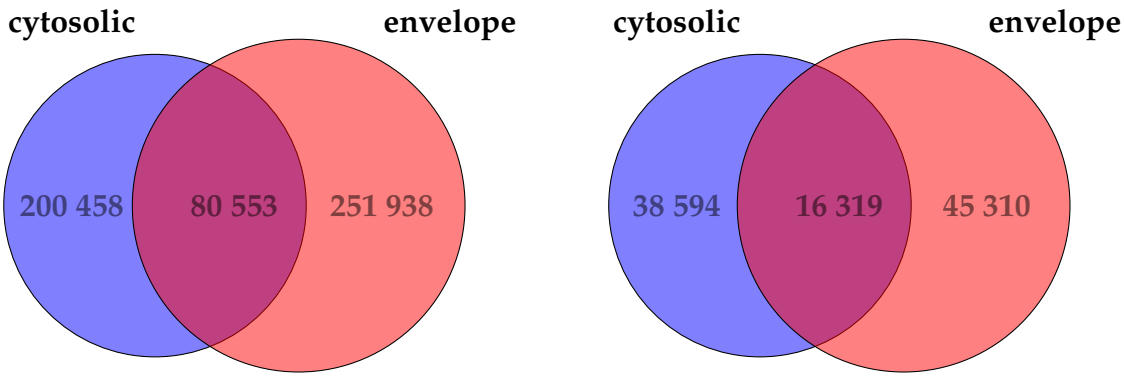


FIGURE 4.8 – Venn diagrams of the number of peptides (left) and proteins (right) identified in fractions of the ProteoCardis samples.

We then drew the cluster trees of all cytosolic metaproteomes on the one hand, and all envelope-enriched metaproteomes on the other hand, as it is a convenient means of quickly viewing similarities between samples, based on the abundance of all subgroups (here approached by the number of their specific spectra). The distances for the tree was computed as $dist = 1 - cor$ where cor is the Spearman's correlation of subgroups abundances between the samples. The clustering was performed with Ward's minimum variance method which minimizes the within-cluster variance. As illustrated by Figures 4.9 and 4.10, samples did not clusterized by colourized patient group. Clearly, patients from a same group could be either close or distant from each other's, in the same manner as patients from different groups.

The image is therefore very different from that we obtained in IBD diseases where healthy and IBD subjects were fairly well separated.

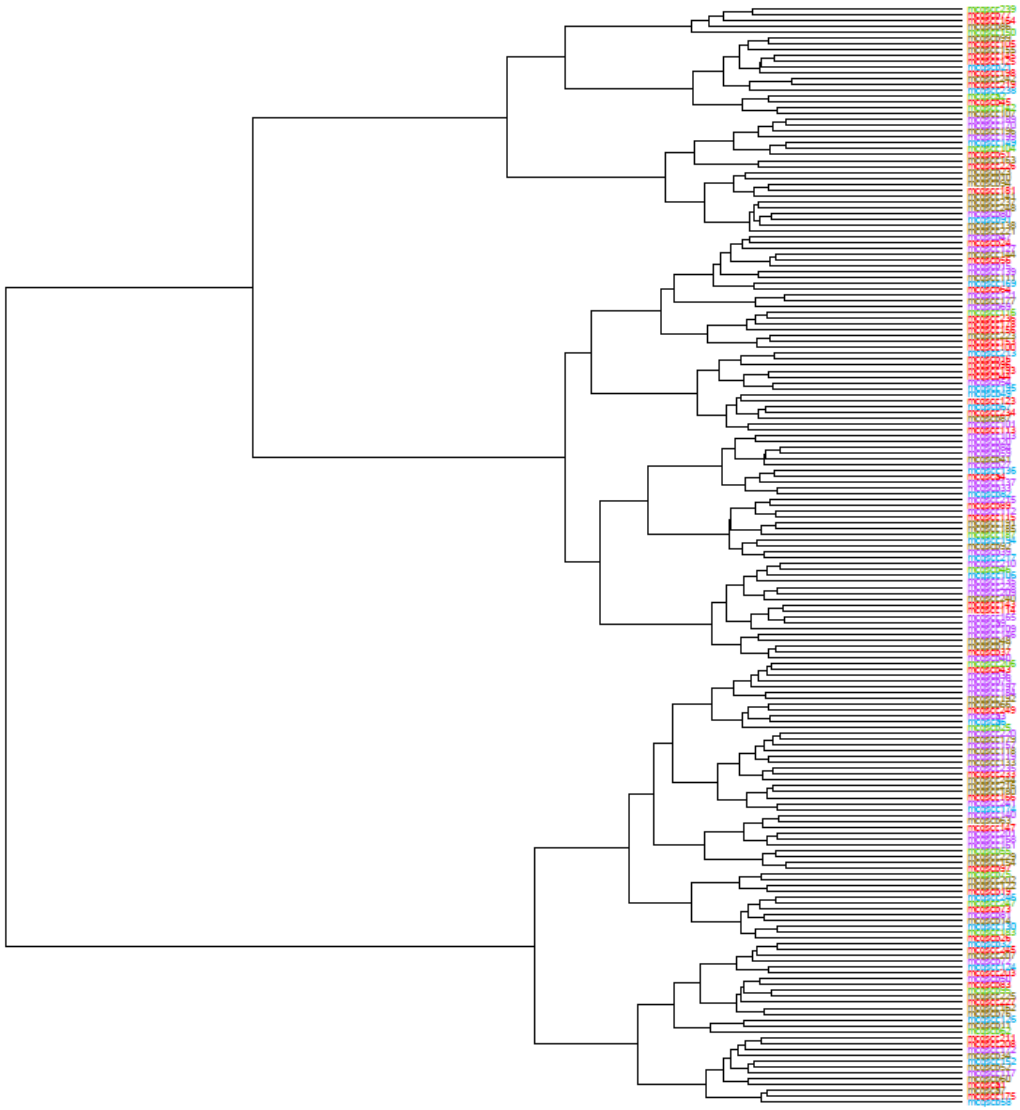


FIGURE 4.9 – Clustering of the cytosolic fractions of the non-bariatric samples (n=188). Colouration by patient group.

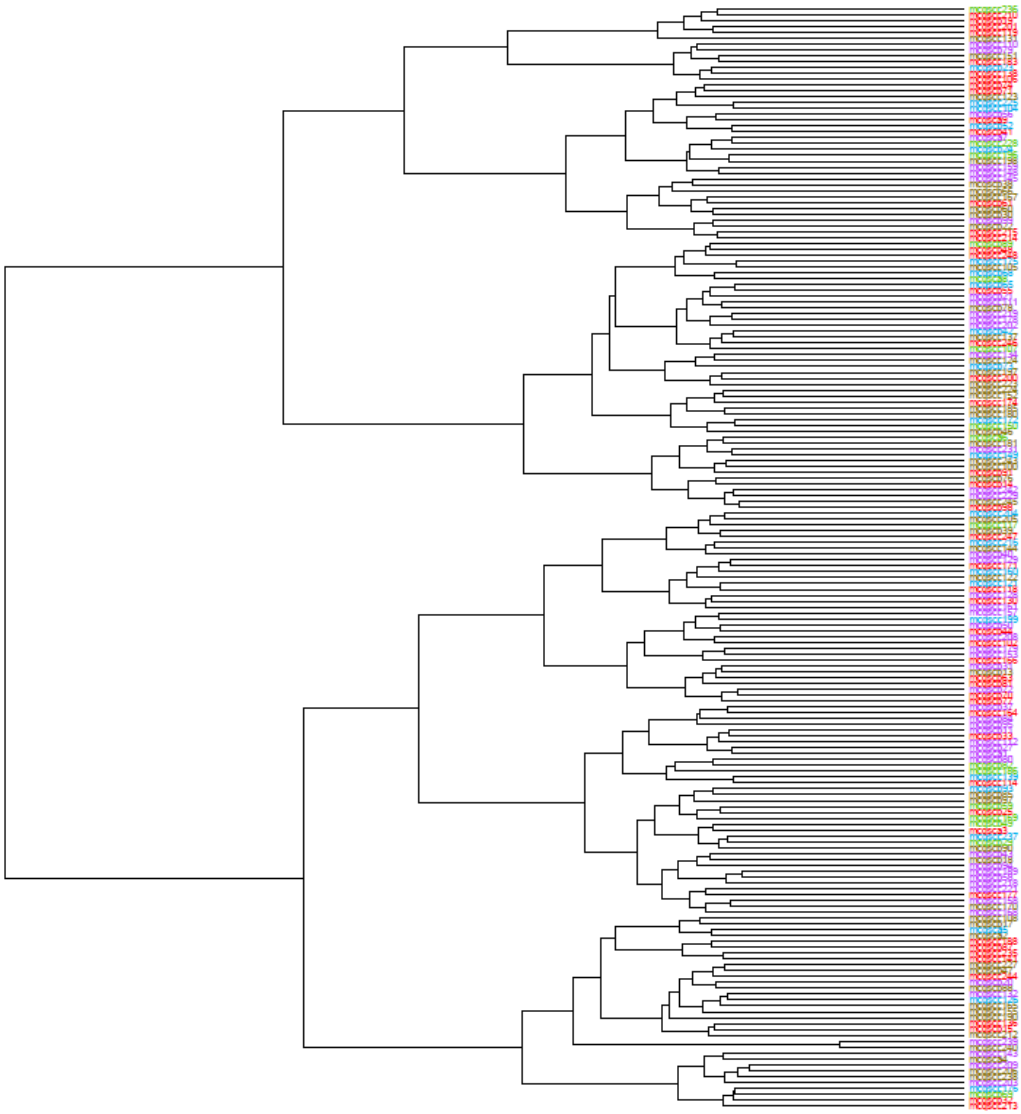
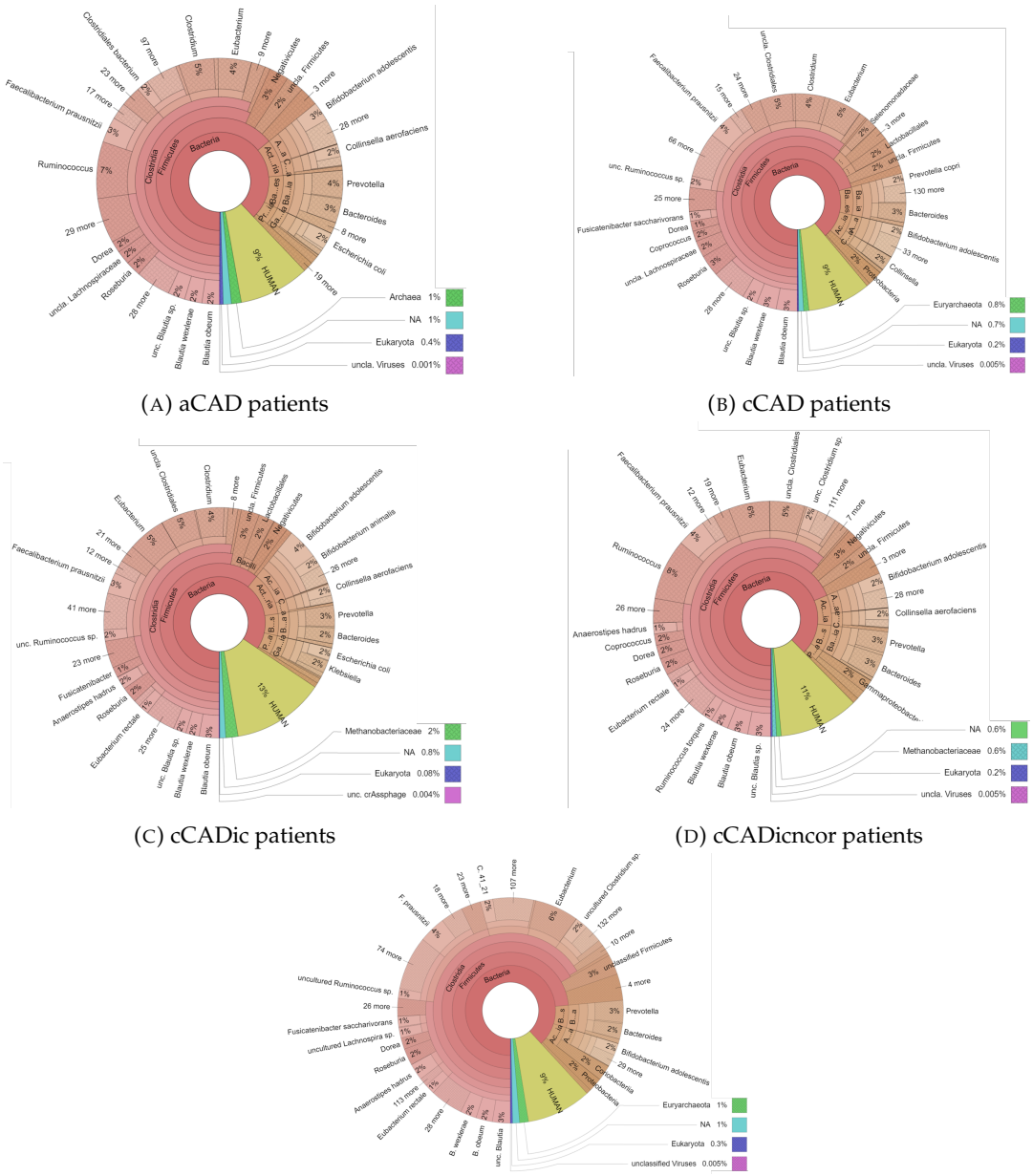


FIGURE 4.10 – Clustering of the envelope fractions of the non-bariatric samples (n=188). Colouration by patient group.

We were then interested in the distribution of subgroups among the different taxa. For that purpose, we mapped all representative proteins of each dataset (cytosolic and envelope-enriched fractions) on the non-redundant NCBI database as described in the Appendix D.6. A total of 53 472 (*i.e.* 98.09% of all microbial proteins) and 58 860 proteins (*i.e.* 97.21% of all microbial proteins) could be annotated down to the species level for the 188 cytosolic and the 188 envelope-enriched fractions, respectively. The taxonomic cytosolic- and envelope-related Krona charts for each patient group are illustrated on Figures 4.11 and 4.12 and main data at the phylum level are summarized in Tables 4.2 and 4.3.



(A) aCAD patients

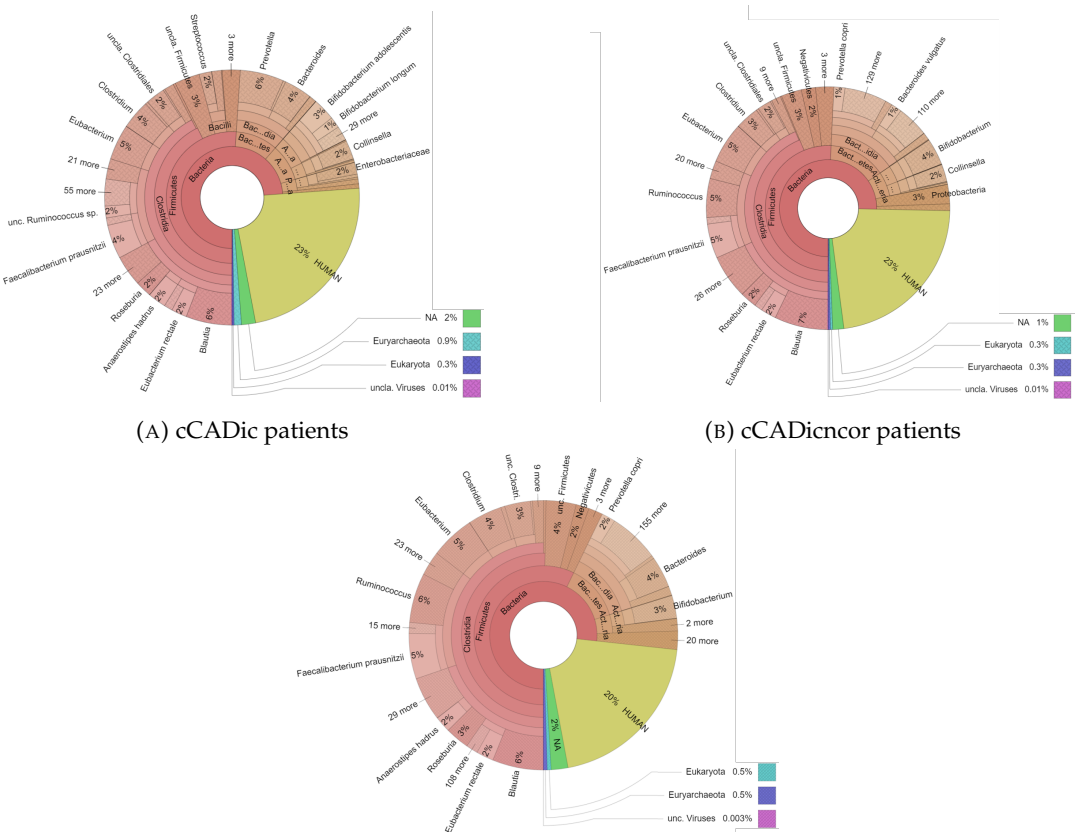
(B) cCAD patients

(C) cCADic patients

(D) cCADicncor patients

(E) Controls

FIGURE 4.11 – Taxonomic distribution of subgroups in the cytosolic fractions of the non-bariatric samples. Contributions are based on abundance of subgroups computed as the sum of SC of their specific peptides.



(A) cCADic patients

(B) cCADicncor patients

(C) Controls

FIGURE 4.12 – Taxonomic distribution of subgroups in the envelope-enriched fractions of the non-bariatric samples. Contributions are based on abundance of subgroups computed as the sum of SC of their specific peptides. The results for the envelope-enriched fractions of the patient groups aCAD and cCAD has not been computed due to memory constraints.

TABLE 4.2 – Taxonomic distribution of proteins in the cytosolic fractions of the ProteoCardis samples. Contributions are based on abundance of subgroups computed as the sum of SC of their specific peptides. aCAD: acute CAD, cCAD: chronic CAD, cCADic: chronic CAD with congestive heart failure, cCADicncor: heart failure unrelated to CAD.

% of all proteins	aCAD	cCAD	cCADic	cCADicncor	Controls
NA	1	0.7	0.8	0.6	1
Viruses	0.001	0.005	0.004	0.005	0.005
Archaea	1	0.8	2	0.6	1
Human	9	9	13	11	9
Other Eukaryota	0.4	0.2	0.08	0.2	0.3
Bacteria	88	89	84	88	88
Firmicutes	64	71	61	69	74
Bacteroidetes	8	8	6	7	6
Actinobacteria	9	8	12	8	6
Proteobacteria	5	2	5	3	2

TABLE 4.3 – Taxonomic distribution of proteins in the envelope-enriched fractions of the ProteoCardis samples. Contributions are based on abundance of subgroups computed as the sum of SC of their specific peptides. The results for the envelope-enriched fractions of the patient groups aCAD and cCAD has not been computed due to memory constraints. cCADic: chronic CAD with congestive heart failure, cCADicncor: heart failure unrelated to CAD.

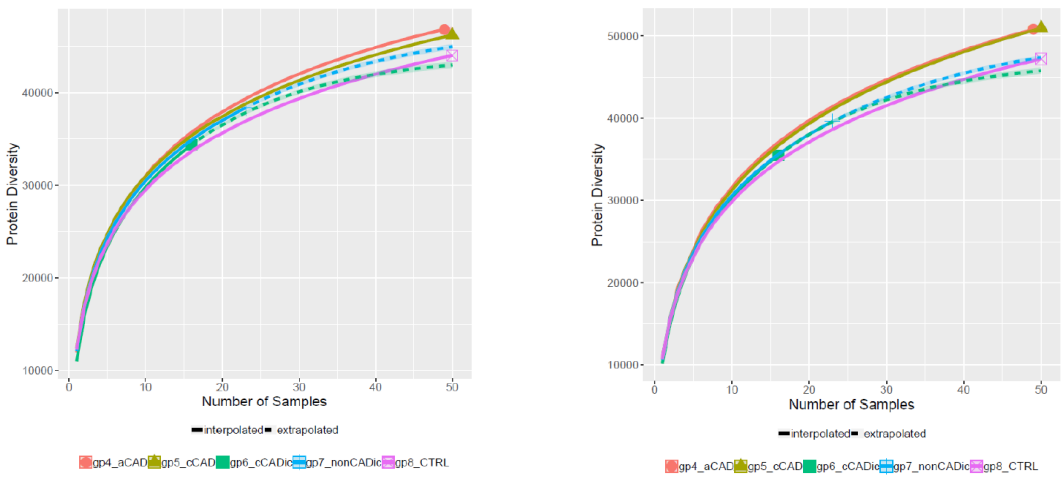
% of all proteins	cCADic	cCADicncor	Controls
NA	2	1	2
Viruses	0.01	0.01	0.003
Archaea	0.9	0.3	0.5
Human	23	23	20
Other Eukaryota	0.3	0.3	0.5
Bacteria	74	75	77
Firmicutes	51	51	57
Bacteroidetes	10	15	13
Actinobacteria	9	6	4
Proteobacteria	3	3	1

When considering the taxonomic annotation of subgroups without taking into account their abundance, no obvious difference appeared in the taxonomic distribution of proteins between the different patient groups (data not shown). Within

cytosolic fractions, proteins from bacterial origin accounted for 96-97% of all proteins, and human proteins for 1-2%. When considering the abundances of those proteins, estimated by the sum of their specific spectral counts, contribution of human proteins was about tenfold increased (9-13%, Figure 4.11) while bacterial proteins decreased by ten points. These results well coincides with an original article just published [128]. Once again, the overall profiles appeared similar in all patient groups. Since a brief overview of the taxonomic distribution showed no difference between the patient groups, the potential dysbiosis associated to gut microbiota require more in-depth research to reveal potential disease signatures.

Compared to the cytosolic fractions, the envelope-enriched fractions contained a higher diversity of human proteins (2-4% of all proteins), which accounted for up to 23% of the total protein content of the envelope-enriched fractions. Interestingly, relative abundance of proteins from the Firmicutes phylum was lower and that of the Bacteroidetes phylum was higher in the envelope-enriched fractions compared to the cytosolic fractions, suggesting an especially high metabolic activity of the former group.

We also looked at the protein diversity in the five non-bariatric patient groups. As illustrated by Figure 4.13, protein diversity was the highest in the two groups aCAD and cCAD, in both cytosolic and envelope-enriched fractions. It was lower in the other three groups, with a slight trend towards an earlier plateau for cCADic and CADicncor than for Controls.

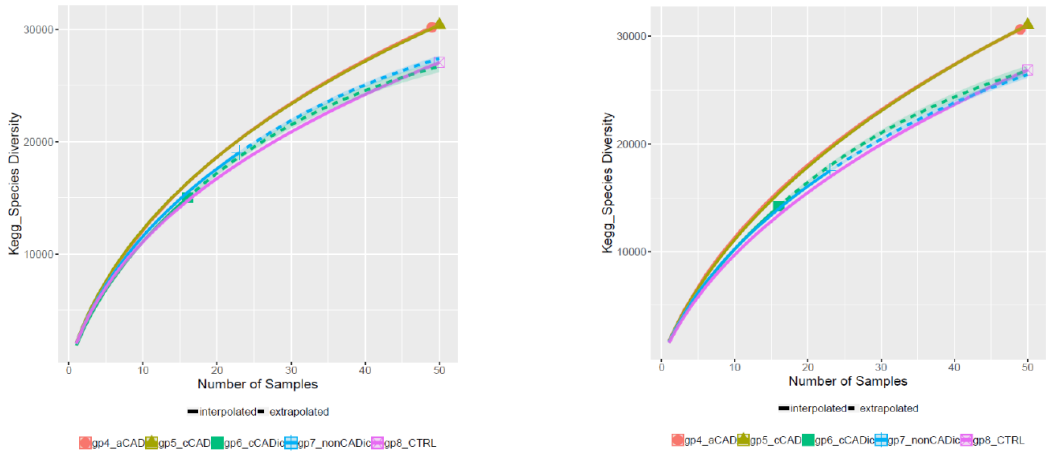


(A) cytosolic fraction (B) envelope fraction

FIGURE 4.13 – Proteins diversity, defined as the mean number of subgroups identified with an increasing number of samples. (A) cytosolic fractions. (B) envelope fractions.

We finally produced the presence-absence matrices for a KEGG/species term combination in order to view the diversity in a more physiologically meaningful way. This is illustrated on Figure 4.14, which reached the same conclusion, *i.e.* a higher taxonomic-functional diversity in groups aCAD and cCAD. This also proves

that the grouping algorithm of X!Tandem Pipeline based solely on shared and specific peptides to construct subgroups and groups of proteins, perfectly reflect the physiological activity of the microbial members, *i.e.* well inform us on "who does what in the system". That said, we could have expected the diversity to be higher in the healthy control group since low gene counts of the gut microbiome has been associated several times with pathological conditions such as obesity and IBD [4, 129].



(A) cytosolic fraction (B) envelope fraction

FIGURE 4.14 – Taxonomic-functional diversity, defined as the mean number of the combination KEGG-species terms, identified with an increasing number of samples. (A) cytosolic fraction, (B) envelope fraction.

To conclude, we showed that fractionation of sample into cytosolic and envelope-enrichment fractions brought complementary information at the functional level, which would have been probably missed without this method for sample preparation. Furthermore, although human proteins was in low diversity, they were in high abundances, as observed in another study [61]. However, inspection of the gut metaproteome landscape in cardiovascular diseases showed no evidence of a profound distortion of the structure and functions of the microbiome. This enables us to foresee a tricky discovery of robust markers that will require a careful selection of the most appropriate statistical methods. Therefore, to maximize our chance of getting a successful outcome of statistical analyses aimed at the discovery of metaproteomic candidate markers of the CAD risk, we carefully examined raw data. This is the topic of the last three chapters with emphasis on XIC quantification, correction of technical effects and statistical analyses.

Chapter 5

XIC quantification

Quantification by eXtracted Ion Chromatograms promises remarkable potential in large-scale studies. Indeed, this MS1-based quantification method makes possible to infer the abundances of peptides across many samples, which is particularly interesting for quantification of peptides of low-abundances, which can be poorly quantified by Spectral Counting. This capacity of inference significantly reduces the amount of missing data compared with SC, especially when the mass spectrometer is of high resolution as it is the case in the ProteoCardis study [130]. However, the high chemical noise is a difficulty to properly quantify the peptides by XIC. In the following chapter, I will present the challenges of XIC quantification as well as the methods that we used to handle them.

5.1 Challenges of XIC quantification in metaproteomics

XIC quantification requires the extraction of the precursor ions chromatograms from MS1 scans to compute peptide intensities from the area under the curve (Section 1.3). The data acquired on fragmented ions in MS2 scans are used to assign MS2 spectra to peptides and proteins but are not involved in the quantification.

The alignment of all the chromatograms in the experiment allows the quantification of peptides that are not identified in MS2, based on their identification in another sample. From one sample to another, the m/z of a given peptide remains the same, but its retention time can deviate. Indeed, reproducibility of LC separation is hardly achievable due to random variations; this variability can profoundly affect the quality of quantification [130]. Quantification by XIC therefore requires alignment of retention times to identify peaks that have not been fragmented. Peak matching based on m/z and retention time of different samples allows us to quantify the corresponding peptide in all the samples, if it has been fragmented and identified in at least one of the samples. However, the alignment of retention times can be difficult in complex samples because the huge number of peaks can lead to a misalignment.

A study of Cheng et al. quantified by XIC and SC the peptides of 32 LC-MS/MS runs of intestinal microbiota of mice under high-fat or low-fat diet [101]. This study revealed differences between the two dietary conditions, which were more evident

with XIC quantification. The study also showed that, for low-abundance proteins, the trend was indistinguishable by SC but observable with XIC quantification. To our knowledge, quantification by XIC has never been implemented on samples of human gut microbiota.

Due to the very high complexity of our samples as well as their high number, we performed a series of analyses to estimate the quality of the alignment of the retention times across all samples.

5.2 Chromatographic alignment

A robust chromatographic alignment is critical to accurately associate corresponding peaks across all runs of the experiment without any mismatching. When a peptide is identified in MS2 in two samples, it will serve as an "anchor" to calculate the deviation of the retention time between all peptides of these two samples. Since the deviation of the retention time is not uniform throughout the LC-MS/MS runs, multiple anchors are necessary for a good alignment.

Quantification by XIC is traditionally used in the context of proteomics. Indeed, in the case of a single organism, the vast majority of proteins will be common to all samples. The alignment of the retention times is therefore easier because there are many anchor points between different samples.

Conversely, in the context of metaproteomics, the huge diversity and variability of micro-organisms and microbial functions in each individual microbiome mean that both quality and quantity of the microbial proteins highly differ between samples. This results in a much smaller number of common multi-sample peptides that could serve as anchors.

Alignment of retention times requires the selection of a reference sample, from which all other samples will be compared to align retention times. Traditionally, in proteomics, the reference sample is the one for which the most peptides have been identified. In our context, it is important that the reference shares as many peptides as possible with all other samples. We selected one reference for each of the four ProteoCardis datasets. These references shared a minimum of 880, 827, 4994 and 1817 peptides with any other samples for the non-bariatric cytosolic, non-bariatric envelope, bariatric cytosolic and bariatric envelope dataset, respectively. This variability in the minimum number of common peptides is in line with our previous observations which showed that non-bariatric samples have a higher diversity than bariatric samples (Table 4.1). Therefore, the higher the diversity the smaller the number of anchor peptides.

We wondered if a limited number of anchor peptides could affect the reliability of chromatographic alignment of highly complex samples and, consequently, the quantification by XIC. To answer this question, I used two different reference samples (referred to as "reference 1" and "reference 2" hereinafter) to align and quantify a

same sample taken from the non-bariatric cytosolic dataset (referred to as "test sample" hereinafter). This test sample and reference 1 was chosen because they share only 880 peptides (the limit of anchor that we want to evaluate), and the reference 2 was chosen because it is the sample which shares the maximum of peptides with the test sample (1 324 peptides). We computed the XIC with MassChroQ, a tool developed by PAPPSO to align the chromatograms, extract and quantify the XIC [131]. The parameters used for XIC extraction and quantification were those usually used at the PAPPSO platform for proteomics. I verified that these parameters were correct thanks to a visual validation with the MassChroQ graphical interface. I quantified the natural isotopes representing at least 80% of the total theoretical intensity, which allowed us to take into account several isotopes and not only the most represented one.

As developed in Section 1.3.3.4, peptiz (a peptide with a given charge) is the direct measurement of XIC. We therefore considered their quantification. As developed in Section 1.3.2, a peptide can be ionized in several peptiz; each peptide feature is therefore quantified by the sum of its peptiz intensities.

The test sample had 7 825 identified peptides, reference 1 had 8 762 identified peptides and reference 2 had 10 375 identified peptides.

When the chromatogram of the test sample was aligned with the chromatogram of reference 1, with which it shared the less peptides, 16 943 peptiz were quantified in the two samples, including 11 417 peptiz in the test sample. With reference 2, 19 565 peptiz were quantified in the two samples, including 12 373 peptiz in the test sample. Of the peptiz quantified in the test sample, 9 003 were quantified with both references. Among them, 8 709 (*i.e.* 96.7%) have a strictly identical quantification regardless of the reference used. Only 118 peptiz had a higher quantification with reference 1, and 176 a higher quantification with reference 2 (a four times greater median). These differences were likely due to misalignments with the references, which accounted for less than 5% of the total peptiz quantifications.

However, when we looked at the 9 003 peptiz quantified in the test sample with each reference, we noted that they corresponded to 8 046 peptides, of which 7 761 had been identified in MS2 in the test sample, and 285 had not been identified. These 285 peptides corresponded to 290 peptiz, of which only 11 peptiz had the same quantification when the test sample was aligned with reference 1 and reference 2. The abundance ratio is illustrated on Figure 5.1 and shows that the abundance can highly vary depending on the reference used for the chromatographic alignment. This clearly demonstrates that XIC quantification of peptides not identified in MS2 may be highly hazardous in metaproteomics. These results are illustrated on Figure 5.2.

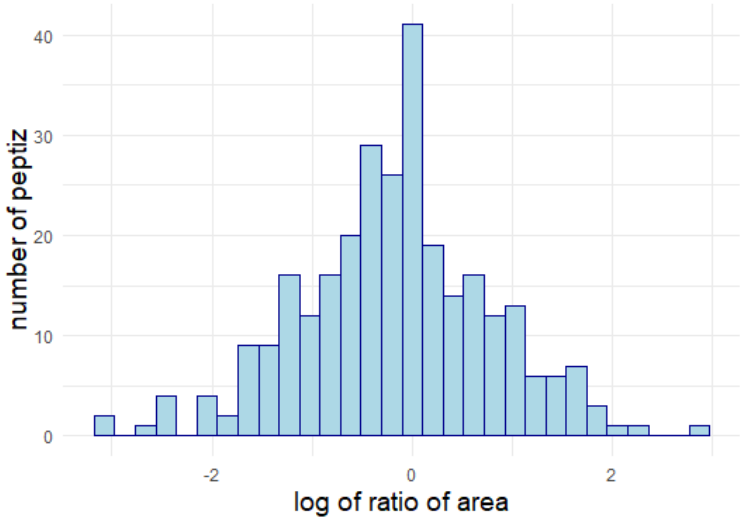


FIGURE 5.1 – $\log_{10}$ of the ratio of peptiz abundances between the test sample aligned with reference 1 or reference 2. If the log ratio is greater than 0, the abundance of the peptiz is higher when the test sample is aligned with reference 1, and *vice versa*.

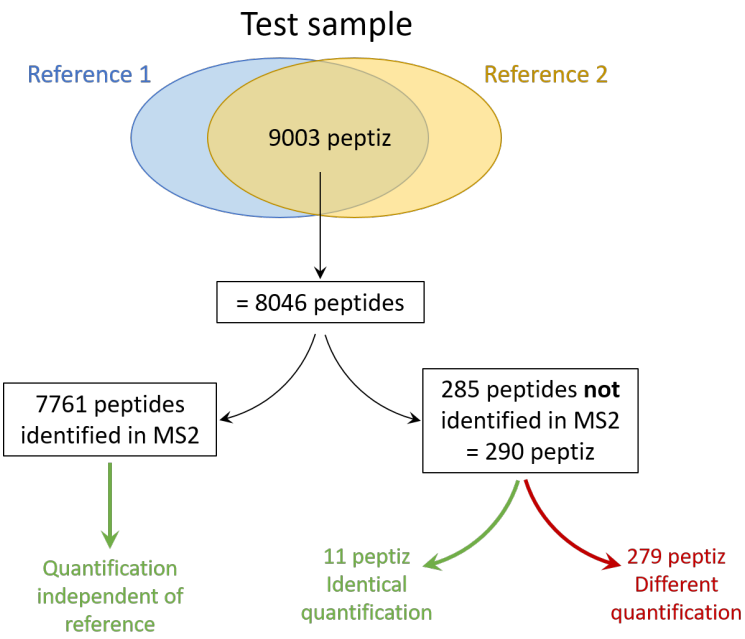


FIGURE 5.2 – Quantification of peptiz of a test sample aligned with two references. After XIC alignment of the test sample with each of the reference, 9 003 peptiz were quantified in the test sample whatever the reference considered. Most of these peptiz corresponds to peptides identified in MS2, for which the quantification is independent on the reference used. Over the peptiz for which the peptide was not identified in MS2, and therefore the quantification by XIC would be possible through alignment, large majority had quantification dependent of the reference used, thus probably misaligned with at least one of the two reference.

Thus, while in classical proteomics, the XIC method successfully quantifies peptides across all biological samples even if they were identified only once in MS2 throughout the entire dataset, this could not be the case in metaproteomics. The XIC quantification obtained for all the datasets and the subsequent statistical analyses should therefore be treated with caution.

We computed the XIC on each of the four datasets of the ProteoCardis study with the same tool and parameters as those used above. We quantified the peptiz of all the samples, including the standards (14 injections of the same sample at the end of each batch) and the replicates (8 samples replicated 7 times) in order to normalize the data, a problematic addressed in the following section. Table 5.1 summarizes the results obtained for each dataset. This table shows that the percentage of missing values is higher than 50% for the four datasets; as a matter of comparison, a proteomics experiment usually have 10-20% of missing values [130, 132]. This high proportion of missing values may be problematic from the statistical point of view to discover peptides/proteins differentially abundant between patient groups.

TABLE 5.1 – Summary of XIC data.

	non-baria cyto	**non-baria env**	**baria cyto**	**baria env**
Number of XIC	30 089 856	33 268 760	9 090 818	8 163 671
Number of peptiz	348 793	412 093	307 115	308 325
Number of peptides	294 355	352 058	254 471	260 685
Number of proteins	72 182	84 094	105 407	122 779
Number of subgroups	57 044	64 107	48 424	48 400
% missing value	60.38	62.72	56.89	59.18

In order to carry out statistical studies, we wanted to pre-process the XIC data to remove potential technical variabilities and misalignments. We also performed an imputation of the missing XIC data. Indeed, replacing the missing data (NA) with zero is impossible in this context because this NA can correspond to a positive value below the detection threshold of the mass spectrometer. Thus, since the values of XIC are of the range of 10^6, inference of the NA by zeros would be highly underestimated. In addition, for some statistical tests, it is better not to have a large number of identical values in the dataset.

5.3 Correction of XIC

Although we have taken the greatest care to ensure reproducible preparation and LC-MS/MS analyses of the samples (Section 4.1.2), to achieve reproducible peptide and protein quantification throughout the entire experiment, perfectly reproducible LC separation is not achievable especially for large cohorts, due to a series of technical variability factors such as random variations of separation conditions, fluctuation of environmental temperature, column aging, and systematic RT shifts over

time [130]. This can never be completely eliminated, but can be well controlled by a completely randomized experiment plan both at the sample preparation and LC-MS/MS analyses level, and inclusion of standards and replicated samples, as it has been done in the ProteoCardis study.

A batch effect linked to the technical variability of mass spectrometric analyses is frequently observed in XIC quantification. This can result in quantification variability between several injections. In order to clean the XIC data, a number of filters must be applied to the data before normalization.

First, the XIC are commonly filtered according to their retention time (RT). Indeed, the spectra at the beginning and the end of the chromatogram have an unstable intensity. So we defined the RT thresholds below and above which XIC intensities became highly variable, based on the profile of intensity values along the RT. For example, the intensity profile of the non-bariatric cytosolic samples presented on Figure 5.3 shows high variation of intensity before 1 100s and after 10 300s. We applied this filter, and the Table 5.2 presents the retained RT for each dataset, as well as the number of peptides and proteins eliminated as a result of this filter.

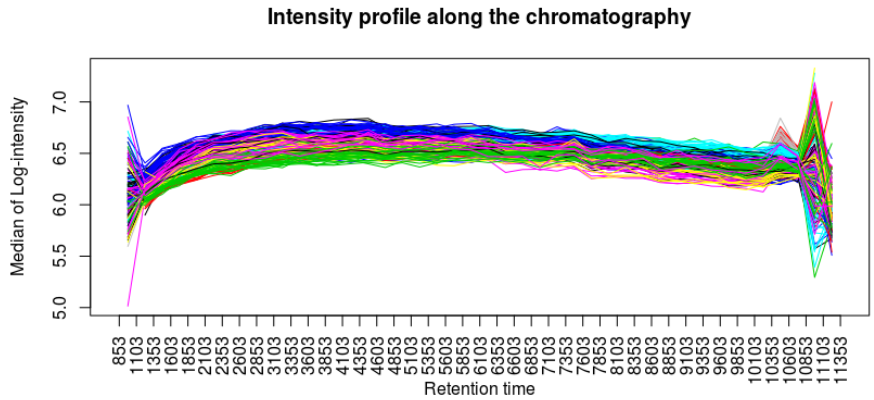


FIGURE 5.3 – Intensity profiles of the samples along the chromatographic retention time. Example of the non-bariatric cytosolic dataset.

TABLE 5.2 – Retention time thresholds retained for the 4 ProteoCardis datasets.

	non-baria cyto	**non-baria env**	**baria cyto**	**baria env**
RT min (s)	1 200	1 050	1 180	1 050
RT max (s)	10 300	8 800	10 000	8 080
Number of peptides removed	12 929	30 465	17 550	32 534
Number of proteins removed	45	232	101	1 060

I then applied a second filter on the retention time to avoid any mismatching, since precise peptide matching among samples is a prerequisite for reliable XIC quantification. I evaluated several filters, where the standard deviation (sd) of the retention times of peptides should not exceed 20, 50, 100, 150 or 200 seconds (the smaller the RT deviation threshold, the more stringent the filter). The profile of standard deviations of retention times is shown on Figure 5.4 and presented no obvious disconnection that could help us identify the limit between good matches and mismatches. I also experienced a third filter where we kept only the intensities of the peptides which have been identified in MS2. Indeed, only the quantifications of the peptides identified in SC are completely reliable (Section 5.2). For this third filter that we called "SC filter", I replaced with missing values the quantifications of any peptide that has not been identified in MS2.

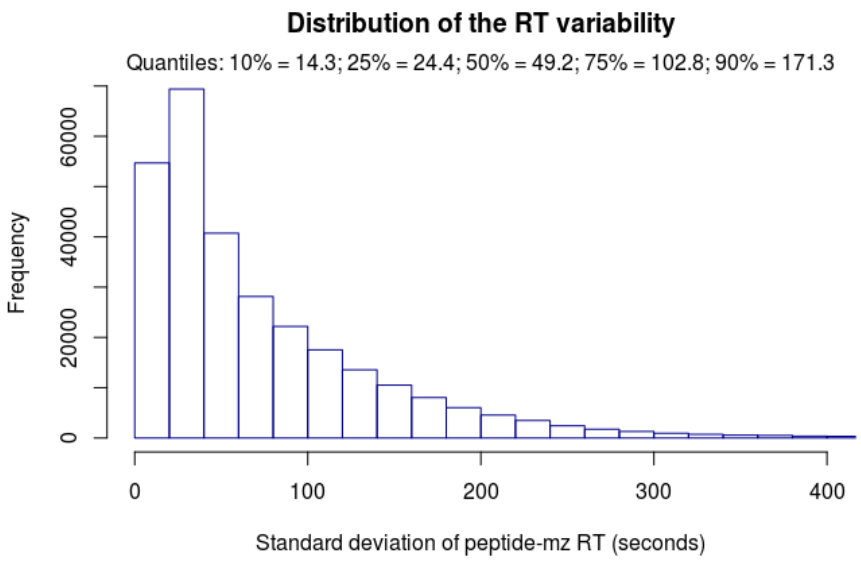


FIGURE 5.4 – Standard deviation of the retention times for the non-bariatric patients, cytosolic fractions.

We used replicated and non-replicated samples to compare results of filtering on RT deviation thresholds and on MS2 identification ("SC" filter). We computed the mean distance between replicated and non-replicated samples with four metrics of distance (Jensen-Shannon, Bray-Curtis, Jaccard and Spearman's correlation). The corresponding ratio (distance between replicates/distance between biological samples) has to be minimized to have a trade-off between replicability of the samples and separation of biological (non-replicated) samples. This ratio is our criteria to define the best filter. Figure 5.5 shows that this ratio was minimized by the "SC" filter with the four metrics, and was rather stable with either no filtering or RT filtering between 200 and 50s. We also considered the percentage of missing values throughout the matrix, which can be a drag on statistical analyses. Figure 5.6 clearly shows that

for the "SC" filter and a RT filter of 20s, the percentage of missing values dramatically increased. These two filters were therefore not suitable in our case. Figure 5.7 shows that the number of peptides and proteins preserved after the different filters clearly decreased for a RT filter < 150s. Of note, the number of peptides and proteins for the "SC" filter stayed high because we considered here the total number of peptides and proteins across all samples. With the "SC" filter, we obtained a matrix with the same number of peptides and proteins than without a filter, but the matrix was filled with more than 90% of missing values. Considering these data, we chose to filter the XICs with an RT threshold of 150 seconds, since this method was the best compromise for preserving a high number of peptides and proteins and avoiding a high number of missing values throughout the matrix.

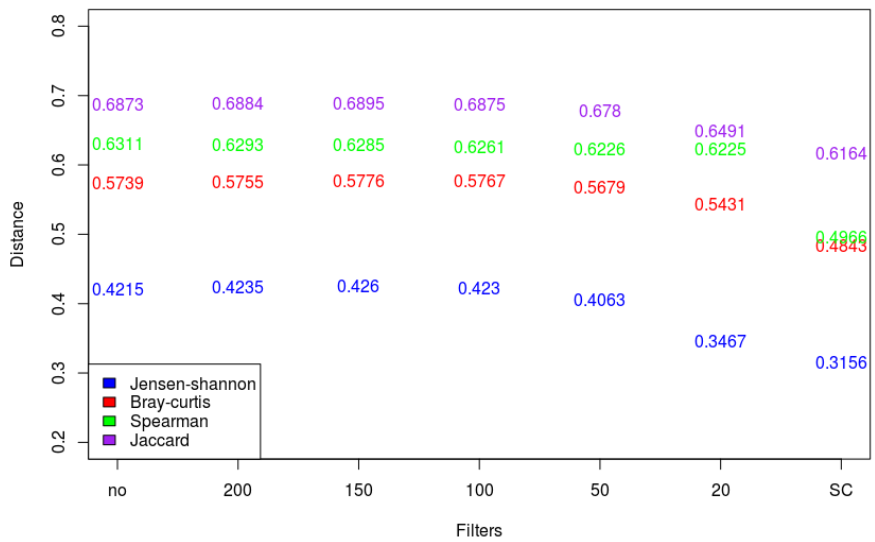


FIGURE 5.5 – Distance ratio between replicated and non-replicated samples after filtering XIC data. X-axis: the XIC were not filtered ("no"), or filtered depending on the standard deviation of the retention time, or filtered depending on the identification in MS2 ("SC").

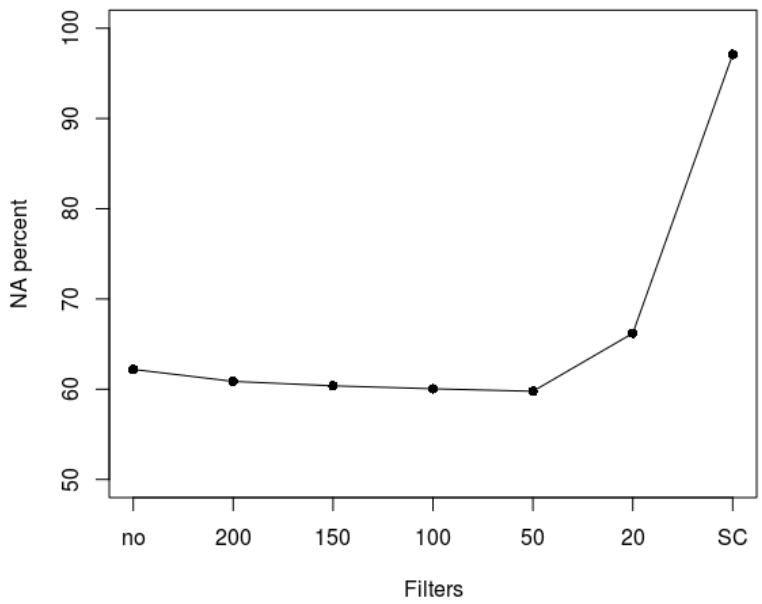


FIGURE 5.6 – Percentage of missing values after the XIC filtering. X-axis: the XIC were not filtered ("no"), or filtered depending on the standard deviation of the retention time, or filtered depending on the identification in MS2 ("SC").

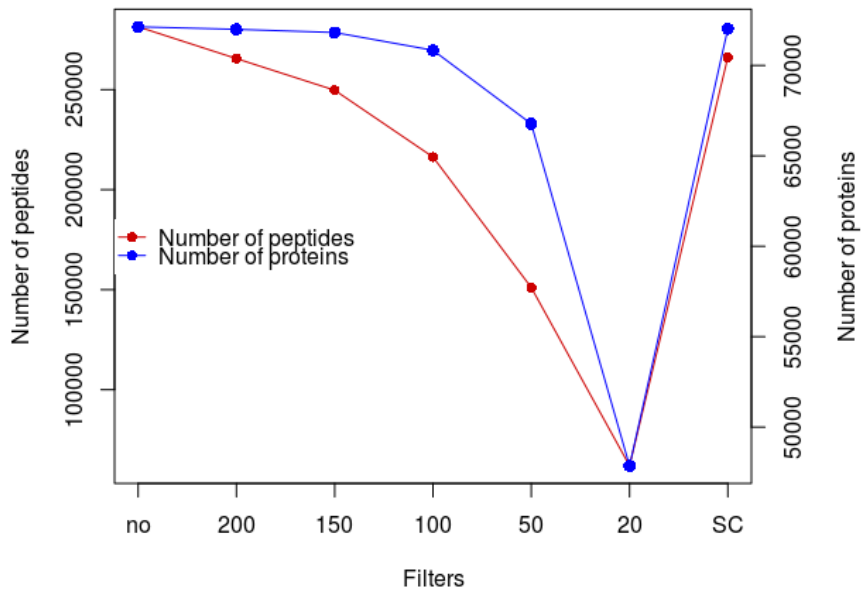


FIGURE 5.7 – Number of peptides and proteins after the XIC filtering. X-axis: the XIC were not filtered ("no"), or filtered depending on the standard deviation of the retention time, or filtered depending on the identification in MS2 ("SC").

Missing information prevents the full, complete, and accurate extraction of quantitative protein and functional information. Thus, one of the major challenges of global proteomic studies is to deal with these missing data appropriately, since they may be real missing values, but also low intensities beyond the detection capacity (which is not a fixed value) of the mass spectrometer. I proposed an imputation of missing values, detailed in the following section.

5.4 Imputation of missing data

The metaproteomics data have a high rate of missing values, as seen in results of XIC quantification on the ProteoCardis samples (Section 5.2). Briefly, three types of missing values were defined by Rubin in 1976 [133] :

- **Missing Completely At Random** (MCAR) when the peptide was randomly missed, due to stochastic fluctuations and independently of its nature or abundance.
- **Missing At Random** (MAR) when the propensity for a value to be missing is not related to the missing value itself, but is related to some of the observed value (conditional dependencies). In proteomics, it is assumed that MAR values are also MCAR [132].
- **Missing Not At Random** (MNAR) when the value is missing because the abundance of the peptide is close to the limit of detection of the mass spectrometer, or even really absent.

Since the reason for a value to be missing is indistinguishable between MCAR, MAR and MNAR, the main approach to replace the missing values in proteomics is imputation of non-zeros values. Indeed, filling missing values with zeros generates biases as does not take into account the correlation structure in the data [134, 135].

5.4.1 Imputation of missing values in classic proteomics

Several methods for imputing missing values exist in proteomics. First, the single-value approaches propose to replace the missing values by a fixed value, determined from the measured values. Typically, the smallest value observed in the experiment estimates the detection limit of the mass spectrometer, and a fraction of this value is used to define the replaced missing values. However, this approach is not appropriate when statistical tests are sensitive to ex-aequo.

Local Similarity Approaches use the abundance of similar peptides (peptides with correlated intensity profiles in the same dataset) to impute missing values. Briefly, it determines the most similar peptides of a given peptide, and uses the abundances of these peptides to estimate the missing values. However, this method is not suitable in the case of metaproteomics because of the large proportion of missing values in the datasets. Indeed, our XIC results show a proportion of missing

values around 60%. Moreover, these approaches make the assumption that proteins are regulated dependently and that highly correlated abundances are normally observed with co-regulated proteins [136]. This assumption is not necessarily correct in the case of metaproteomics, where proteins can be regulated differently between gut microbiomes, which are highly specific to individuals. Finally, this type of imputation can impute high abundances, which is not desirable in the case where the missing values are values too small to be detected by the spectrometer.

Finally, Global-Structure Approaches use a decomposition of the data matrix and then iteratively reconstruct the missing values. These methods have been shown to be less effective than Local Similarity Approaches [136], and are extremely time-consuming, even in the proteomics context.

5.4.2 Imputation of missing values in metaproteomics

The methods of imputation in metaproteomics are based on the methods developed in proteomics. For example, Zhang et al. uses the K-nearest neighbours method (KNN) to search for the most similar peptides (Local Similarity Approach), after filtering the subgroups that are present in 50% of the samples [61]. The KNN method requires less than 30% missing data to be effective [137]. For this reason and others explained herebefore, this method is not suitable in our case. Other studies uses a Global-Structure Approach that is hardly applicable to large-scale datasets [40, 138].

In some metaproteomics study, a variant of the single value method have been proposed, in which the missing values are imputed from a normal distribution centered on the inferred detection threshold [125, 139]. Although this method is interesting because it allows us to obtain (i) non-identical values and (ii) low values representing many missing data, the distribution parameters are not discussed. I decided to use this approach, and I determined the parameters using our data.

5.4.3 Imputation implemented in the ProteoCardis study

For our NA imputation, we considered peptide abundances, which are defined by the sum of the corresponding peptiz intensities. The values were determined according to a normal distribution, whose mean and standard deviation parameters must be fixed. NA values were considered as values under the detection limit of the mass spectrometer (MNAR). The average value of the normal law was fixed to the quantile 10^{-5} of the set of values observed in the quantification.

In XIC data, we observed that the standard deviation (sd) tended to increase with the mean of the peptides abundance across samples (Figure 5.8). The sd linked to technical variability can be estimated thanks to the peptides with a mean abundance close to the quantile 10^{-5} in replicated samples, since the variations in these samples have no biological cause. However for these peptides with low abundances, the high number of missing values makes their sd poorly reliable. We therefore inferred the

sd by extrapolation of the sd observed on peptides with few missing values, those quantified in at least 12 to 14 samples of the replicated standards.

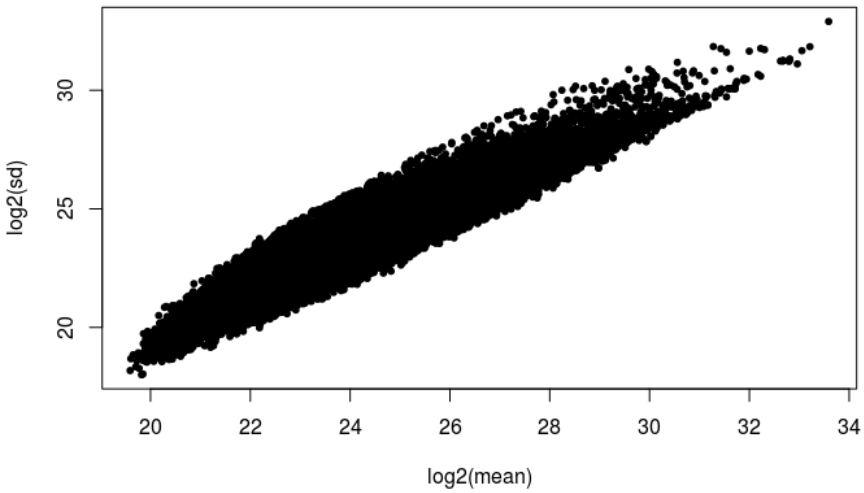
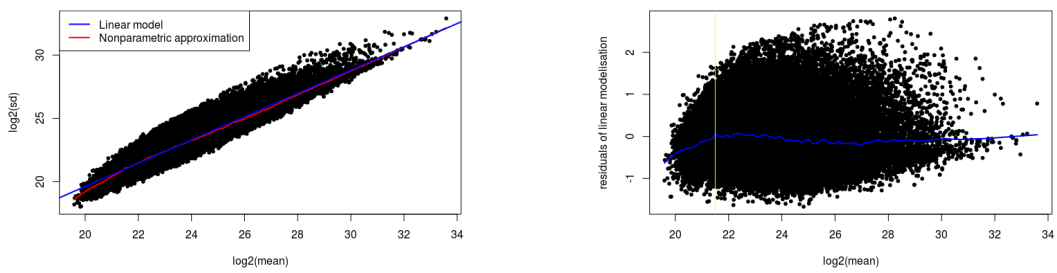


FIGURE 5.8 – Mean quantification of peptides quantified in 12 to 14 standard samples and their standard deviation. Log scale.

We observed that the curve on Figure 5.8 was approximately linear, except for the low mean values, which are the values for which we want to infer the sd. The implementation of a linear model and non-parametric approximation (Figure 5.9A) and the residuals of the linear model (Figure 5.9B) confirmed that the linear model was not adapted for low mean values and overestimated the sd. The implementation of the linear model only on the low mean values (log_2(mean)<21.5) presented the same overestimation of the sd (Figure 5.10).



(A) Linear model and non-parametric approximation

(B) Residuals of the linear model. yellow: quantile 10% of the mean.

FIGURE 5.9 – (A) Linear or non-parametric modelling of the relationship between mean and standard deviation of XIC peptides values. (B) Residuals of the linear model.

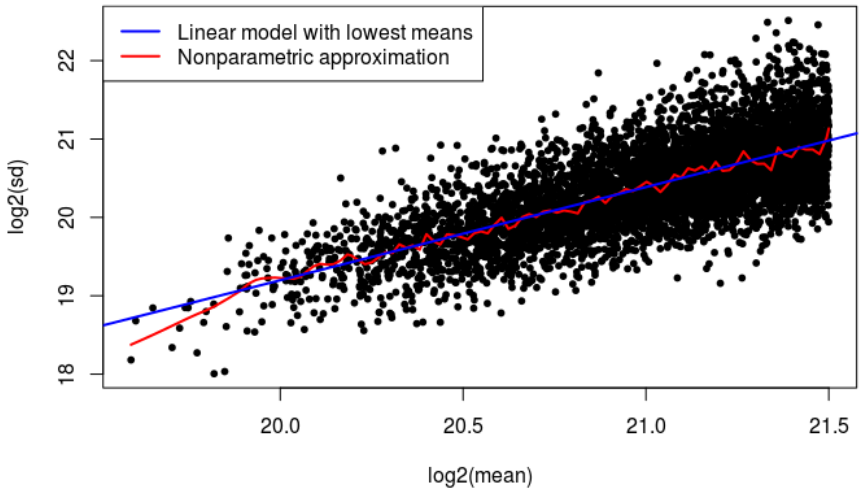
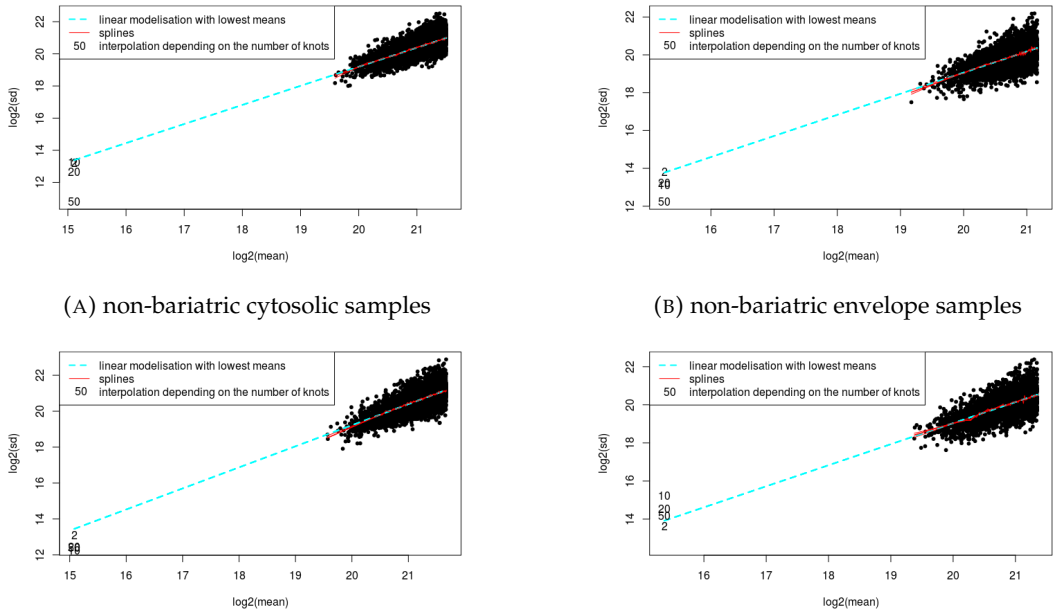


FIGURE 5.10 – Linear or nonparametric modelling of the relationship between mean and standard deviation of the quantification of peptides with low XIC mean values.

We therefore fitted a linear spline model on the 10% lower quantile of the mean values. Linear spline is a piecewise function consisting of a polynomial of degree 1 (straight) on each interval between knots. We considered 2, 10, 20 and 50 knots spaced equally to the quantiles of the distribution. This spline modelling interpolates the value of sd expected for a given mu, Figure 5.11.



(A) non-bariatric cytosolic samples

(B) non-bariatric envelope samples

(C) bariatric cytosolic samples

(D) bariatric envelope samples

FIGURE 5.11 – Interpolation of sd by splines considering the number of knots for the splines modelling for non-bariatric (top) and bariatric (bottom), cytosolic (left) and envelope (right) fractions of ProteoCardis samples.

For each dataset, we retained the number of knots for which the model had the lowest probability to select a negative value. The values of mu, sd and the probability of sampling a negative value for each dataset are summarized in Table 5.3.

TABLE 5.3 – Values computed for the imputation of XIC missing values. non-baria: non-bariatric. baria: bariatric. cyt: cytosolic fraction. env: envelope fraction.

	non-baria cyto	**non-baria env**	**baria cyto**	**baria env**
best number of knots	50	50	10	2
μ	35 380.15	39 364.83	34 648.21	42 249.47
sd	1 790.41	4 827.85	4 936.40	12 642.81
probability to have a negative value	3.23×10^{-87}	1.77×10^{-16}	1.12×10^{-12}	4.16×10^{-4}

So I replaced the missing values in each dataset with a random sampling following the normal law with the parameters in Table 5.3, and the negative values were replaced by zeros. Of note, negative values were only sampled in the bariatric envelope dataset, representing only 0.02% of all values in this dataset.

In this chapter, we therefore determined a model that best fitted to our data, in order to determine the optimum parameters to impute missing values. This method imputes values consistent with the range of values of the datasets. However, imputation with sampling assume that most of the missing values are MNAR, and thus impute low-abundance values. In the case of MAR and MCAR, the range of missing values is not representative of the true range of the missing values. The use of XIC with imputed values may therefore be difficult, especially when more than a half of the values have been imputed.

Chapter 6

Can the ProteoCardis data be improved by normalization methods?

When the samples are injected in the mass spectrometer, the needle becomes clogged due to the non-volatile molecules in the injected samples. This clogging decreases the sensitivity of the spectrometer and increases the background noise, despite a deep cleaning every 18-22 injections (Section 4.1.2). The last sample injected in each batch corresponds to the standard sample. Since clogging and cleaning can influence the abundances of identified peptides, in this chapter I evaluated and corrected the abundances of SC and XIC through different normalization methods.

6.1 Ascertainment of the batch effect

The total abundance of SC and XIC can vary between injection batches. The abundance of the peptides can be modified between each batch ("batch effect"), as well as within batches where the sensitivity of the spectrometer is better on the first injections than on the last ones. In addition, apart from the batch effect, an injection order effect can be observed, where the total abundances can deviate as injections occur despite cleaning. These variations are solely due to technical effects, which we want to minimize. I chose to consider the batches as categorical variables without taking into account the injection orders, since normalizations of the batch effects are already existing methods.

I conducted the analyses on the non-bariatric ProteoCardis dataset, cytosolic fraction, and on the protein SC (which sums all the peptides counts). We used the SC of the proteins, which reduces the datasets size compared to the SC of the peptides. A representation of the sum of the SC ranged by injection order showed that it seems to increase in the course of the experiment, suggesting that a batch effect exists (Figure 6.1).

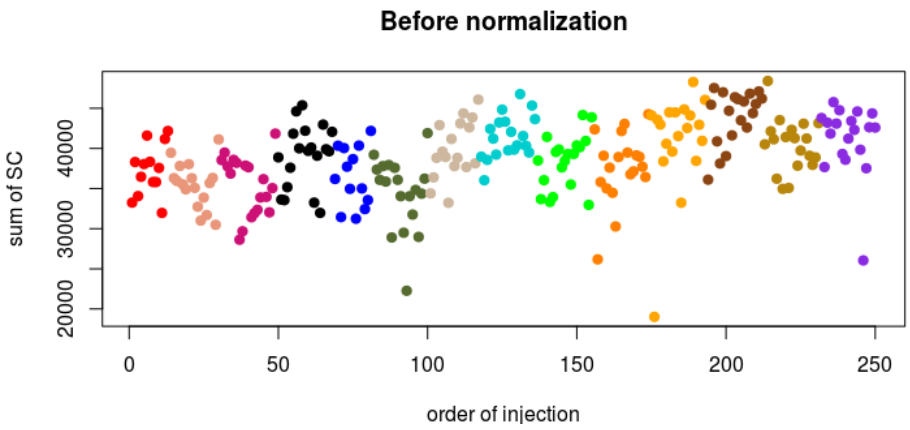


FIGURE 6.1 – Sum of SC of the proteins, before any normalization (raw data) for the dataset of non-bariatric patients, cytosolic fractions. Colouration by batch.

We often observed a batch effect on the intensity of XIC (Section 5.3). Figure 6.2 shows that the batch effect on XIC was low on the cytosolic fraction of non-bariatric patients, although we observed a slight drop in XIC intensities at the 4th and 7th batches.

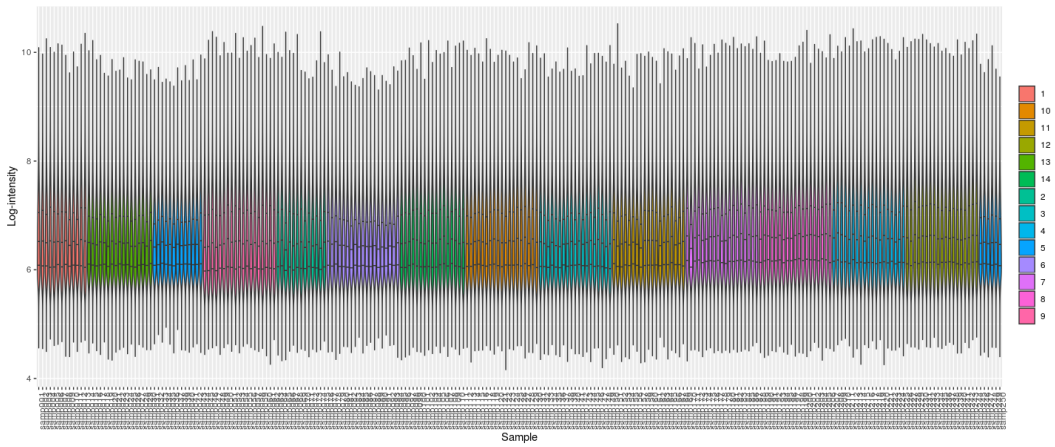


FIGURE 6.2 – Global intensity profiles of XIC before normalization

Although the batch effect appears to be weak at the SC or XIC level, we normalized the data using several methods to obtain the cleanest data possible for the statistical analyses that follow.

6.2 Normalization methods

6.2.1 Methods for SC normalization

For the SC normalization, we have first considered normalization by batch.

Sum of Spectral Counts of Standards: I computed a normalization factor based on the sum of the spectral counts of the standards present in each batch. We rely on the hypothesis that since the variations observed between the standard samples are solely due to technical variation, the sum of its spectral counts must be identical between the batches. This hypothesis makes it possible to calculate a normalization factor specific to each batch. This factor is then applied to all SC of the samples in each batch.

MS2 of Standards: I computed a normalization factor based on the number of MS2 spectra of the standards. The hypothesis is that the sum of SC previously performed is dependent on identification; however, the number of MS2 is only dependent on LC-MS/MS analysis. The number of MS2 of the standards must therefore theoretically be identical between batches. The normalization factor is applied to all SC of the samples in each batch.

Linear Regression: A commonly used method to suppress the batch effect in RNA-seq experiments is linear regression. It is based on the hypothesis that the abundance of the variables have a Gaussian distribution, whose average is dependent on the batch. Since RNA-seq analyses are also based on count data, I wanted to evaluate this method on our metaproteomics data.

In the following, I also considered a normalization on abundances without taking into account the batches.

Total Ion Current: The Total Ion Current (TIC) is the total amount of intensity in precursor spectra, and therefore estimates the total signal available in the mass spectrometer. It includes contributions from peptides, contaminants, and noise. The hypothesis underlying this normalization is that all samples should have the same TIC because the amount of protein introduced into the mass spectrometer is identical between samples. However, the TIC can decrease with time due to the clogging of the spray needle. There is indeed a variation of the TIC according to the injections (Figure 6.3). A decrease in TIC for a given sample indicates that the amount of proteins introduced is lower for this sample. The number of MS1 and therefore MS2 could thus be reduced and induce technical variability, which we seek to reduce. There was indeed a correlation between the TIC and the sum of SC (Figure 6.4), with a correlation coefficient of 0.68 (Spearman's correlation). I therefore computed a specific normalization factor for each sample, proportional to TIC, which was applied to each peptiz SC of the sample.

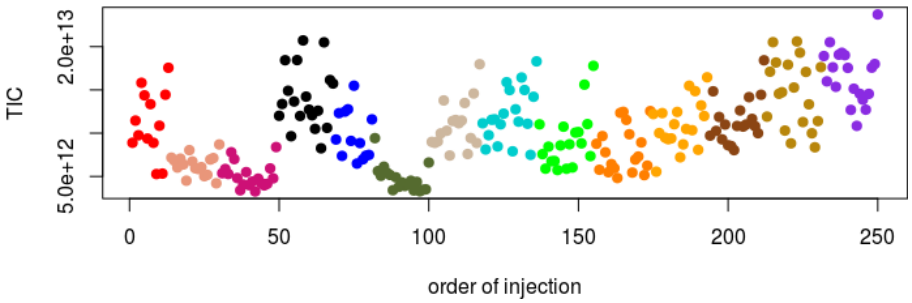


FIGURE 6.3 – Total Ion Current of the ProteoCardis samples, patient non-bariatric, cytosolic fraction. Colouration by batch.

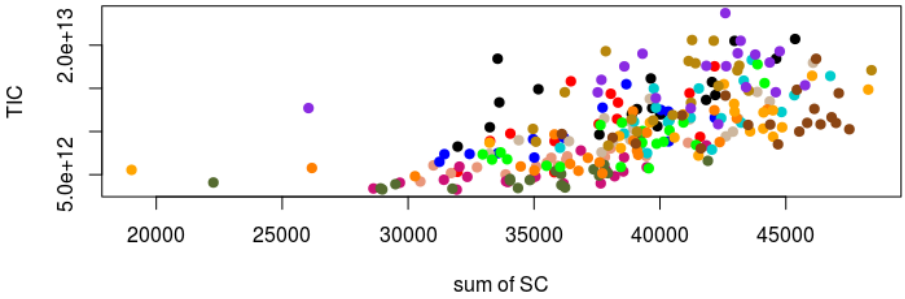


FIGURE 6.4 – Correlation between TIC and sum of SC of each sample. Colouration by batch.

Trimmed Mean of M-values: Trimmed Mean of M-values (TMM), was proposed by Robinson and Oshlack in 2010 [140]. This method, initially developed for the RNA-Seq, proposes to take into account the composition of the samples to compute a normalization factor. Indeed, two samples sequenced to the same depth and expressing the same number of transcripts for a given gene will not have the same RNA-Seq count if one of the two samples transcribes a greater diversity of genes. This problem is also found in the case of tandem mass spectrometry, because a limited number of spectra can be analysed on the time of acquisition, and the selection of the ions to be fragmented has a random component. Since our samples can have extremely heterogeneous diversities, we therefore wanted to evaluate the TMM normalization on our metaproteomic samples.

6.2.2 Methods for XIC normalization

XIC normalization methods are all based on the hypothesis that most peptides have the same intensity across all samples, the intensities are thus corrected by a factor.

Normalization by percentage: The percentage method consists of dividing the intensities of the peptiz into each sample by the sum of the intensities in the sample. Then these intensities are multiplied by the average intensity calculated on all the samples. This method preserves the relative abundance of peptiz within each sample.

Normalization by median: The intensities of the peptiz in each sample are divided by their intensity in a reference sample (if the peptiz is also quantified in the reference), giving a ratio for each common peptiz between the reference and the sample considered. Then, in each sample, the median of these ratios is calculated to define a median ratio for all the peptiz of a sample relative to the reference sample. Each original intensity of peptiz is then divided by this median ratio, specific to each sample. We chose as a reference for this normalization the sample which had the greatest minimum number of common peptides with any other sample (peptides identified in MS2 in both samples).

Normalization by median-RT: Throughout the acquisition by mass spectrometry, the intensity deviations are not always uniform. Median-RT normalization is a method derived from median normalization; in the median-RT normalization, the median ratio is not calculated uniformly. The ratios are ordered according to the retention time of their peptiz, then the values are smoothed thanks to a cubic smoothing spline, which is a piecewise curve made up of polynomials of the third degree. The intensities of each peptiz are then divided by the smoothed ratio values corresponding to their retention time. As for median normalization, we chose as a reference the sample that had the greatest minimum number of common peptide with any other sample.

6.3 Evaluation of the normalizations

6.3.1 Normalization of SC

We evaluated the variability between samples using a Principal Component Analysis (PCA). The objective of the PCA is to reduce the number of explanatory variables (here, the peptides/proteins) while preserving the distances between individuals. This makes it possible to synthesize the information and to explore the links between individuals. It relies in particular on the transformation of the correlated variables into new uncorrelated variables, called principal components. The number of principal components is less than or equal than the number of original variables. These principal components make it possible to determine the main axes on which the individuals can be represented. The interpretation of these graphs makes it possible to understand the structure of the analysed data. The use of replicates allowed us to evaluate *via* the PCA the quality of a batch correction. The more the replicates were grouped on the PCA, the more we considered that the technical variability was reduced. I also calculated the ratio between the average distances of the technical replicates and the biological samples, as in Section 5.3. The objective was to minimize this ratio, so that the technical replicates are as close as possible while preserving the heterogeneity of biological samples. I therefore used a visual criterion (PCA) and a numerical one (ratio) to judge the normalization.

PCA on unnormalized data is shown on Figure 6.5, where we observe that the replicates are very close to each other. The technical variability is therefore hardly visible on the SC data of this dataset.

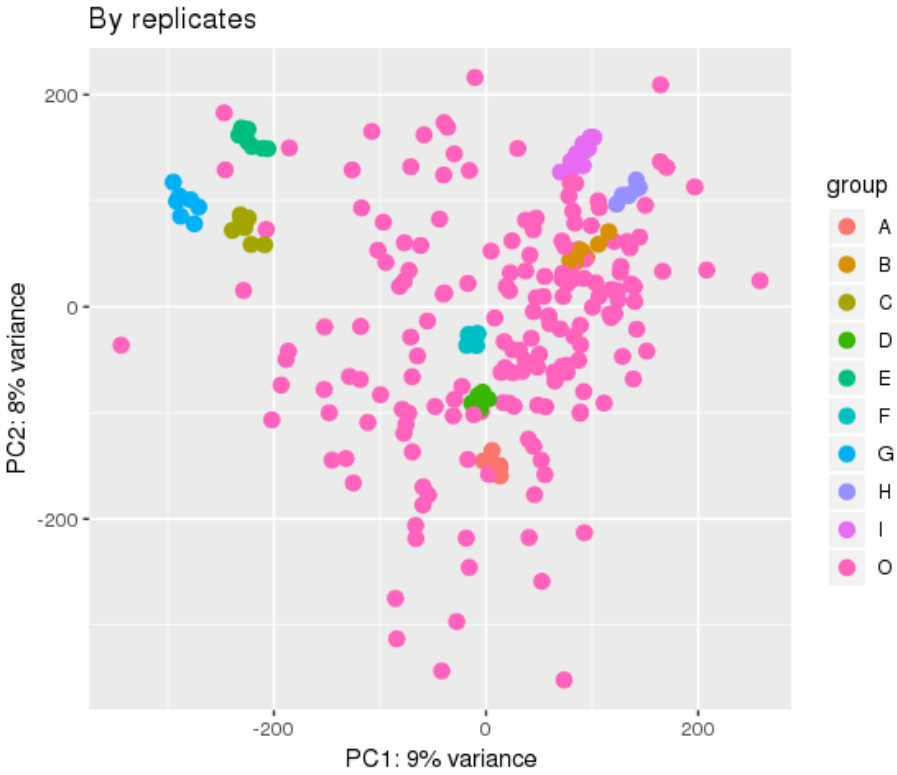
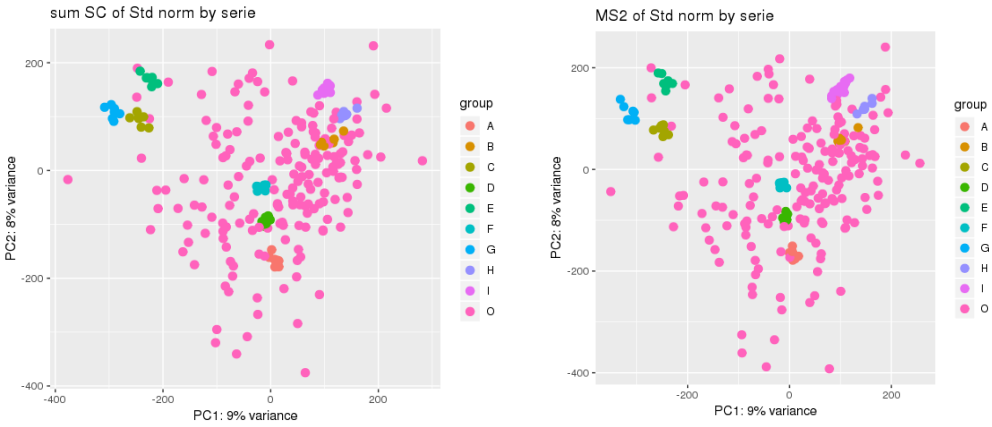


FIGURE 6.5 – Principal Component Analysis of raw Spectral Counts of proteins. The samples are coloured by replicates. A-H: replicates, each replicated seven times. I: Standard, replicated fourteen times. O: samples not replicated.

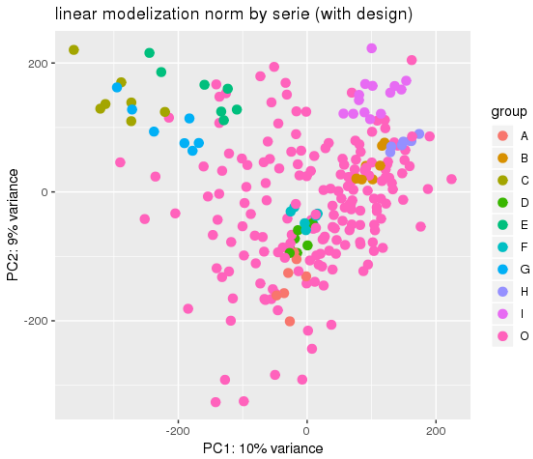
The PCA resulting from the different normalizations by batch (sum of SC of standards, MS2 of standards and linear regression) are illustrated on Figure 6.6. For normalization with the sum of the SC or MS2 of the standards, the PCA showed replicates still as close as before the normalization, with some replicates that got slightly closer or farther away according to the normalization applied compared to unnormalized data (B, C, G, I). Conversely, normalization by linear regression completely burst the replicates and is therefore not adapted to our datasets. This effect is due to the distribution of metaproteomic data, compared to those observed in RNA-seq. Indeed, the SC of the proteins are mostly equal to zero. Therefore the distribution of the data cannot be Gaussian in metaproteomics and this method is thus not transposable to metaproteomics.

I then assessed normalizations on total abundance (TIC and TMM). The PCAs resulting from these normalizations are illustrated on Figure 6.7.



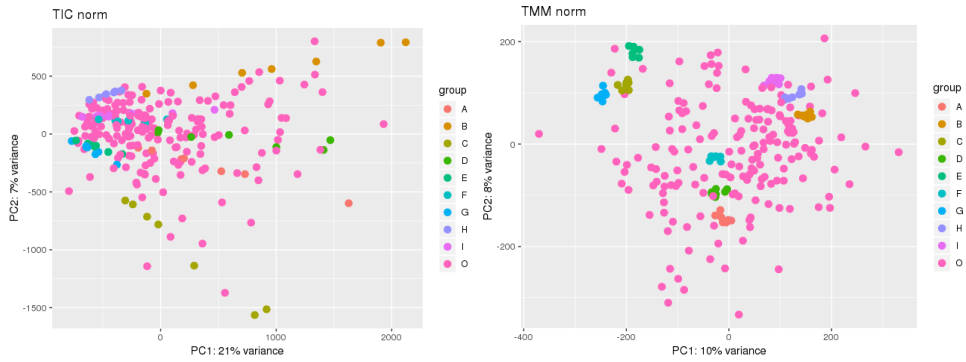
(A) Sum of SC of standards

(B) MS2 of standards



(C) Linear regression

FIGURE 6.6 – Principal Component Analysis of Spectral Counts of proteins after the normalization of batches by (A) the sum of SC of the standards, (B) the MS2 of the standards and (C) linear regression. The samples are coloured by replicates. A-H: replicates, each replicated seven times. I: Standard, replicated fourteen times. O: samples not replicated.



(A) TIC (B) TMM

FIGURE 6.7 – Principal Component Analysis of Spectral Counts of proteins after the normalization of total SC abundance by (A) TIC and (B) TMM. The samples are coloured by replicates. A-H: replicates, each replicated seven times. I: Standard, replicated fourteen times. O: samples not replicated.

PCA after TIC normalization showed that it completely burst the replicates. The representation of SC abundance before and after TIC normalization showed that this normalization introduced a batch effect (Figure 6.8) which was weak in the raw data (Figure 6.1). In contrast, the replicates are well clustered on the PCA after TMM normalization. However, one of the hypotheses of the TMM method is that most proteins have the same abundance across all samples. In the case of metaproteomics, this hypothesis is not verified.

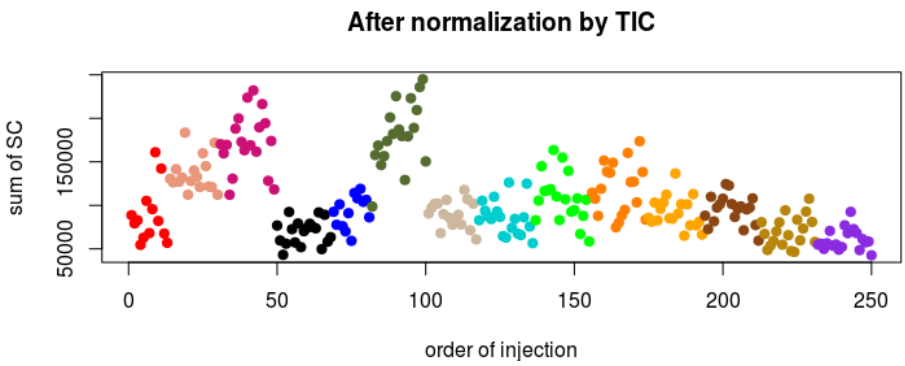


FIGURE 6.8 – Sum of SC of the proteins, after TIC normalization. Colouration by batch.

Four distance metrics (Jensen-Shannon, Bray-Curtis, Spearman correlation, and Jaccard) were used to calculate the average distance between biological samples and replicates. The corresponding ratio (distance between replicates/distance between biological samples), which we seek to minimize, was calculated before and after each normalization. Figure 6.9 shows that regardless of the distance metric used, unnormalized data and TMM normalization gave the best results.

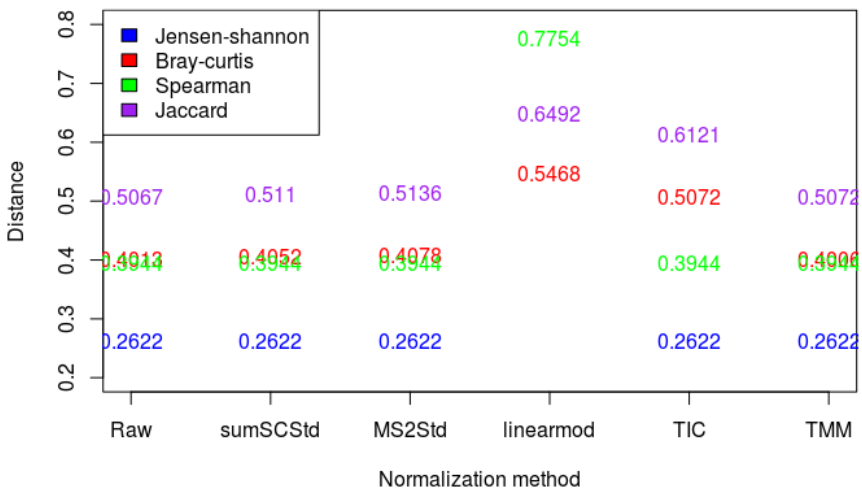
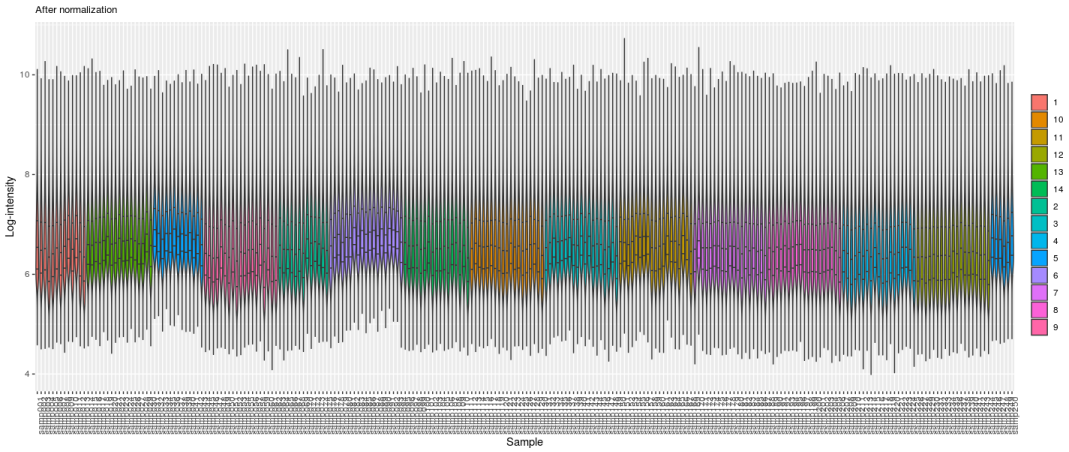


FIGURE 6.9 – Ratio of the distances between replicates and non-replicates samples, before normalization and after each normalization evaluated.

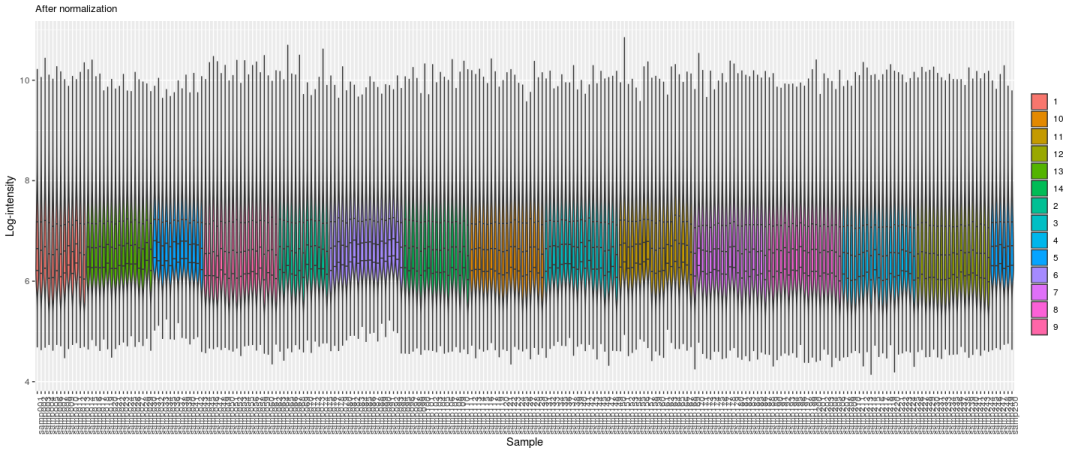
In order to not introduce bias and in view of the observed results on the unnormalized data, where the batch effect seemed weak, I decided not to normalize the SC data.

6.3.2 Normalization of XIC

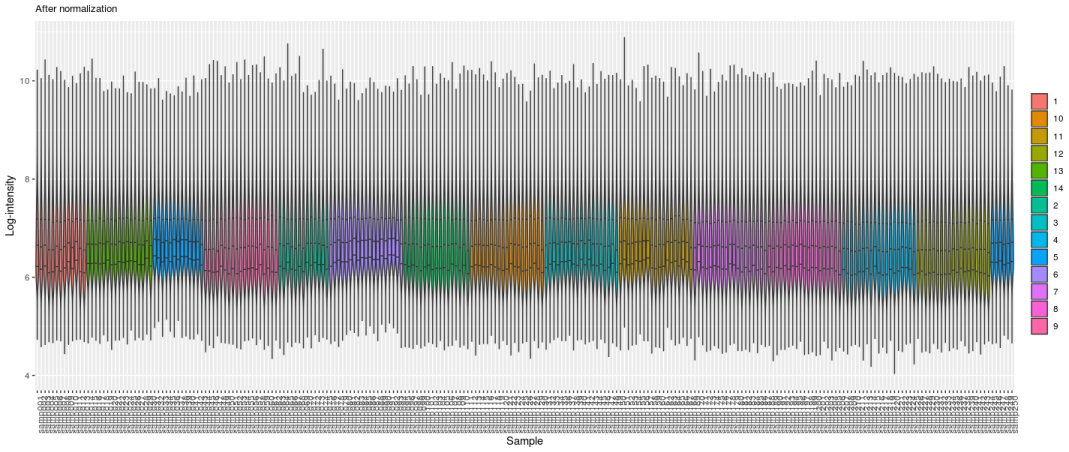
I examined the variability of the total intensities visually, thanks to the distribution of the peptiz intensities in each sample which has to be homogeneous. The distribution of peptiz intensity after computation of the three normalizations (percent, median, and median-RT) is shown on Figure 6.10. Normalization by percentage did not decrease the batch effect, and even tended to accentuate it. Median and median-RT normalization were less detrimental than percentage normalization but still tended to accentuate the small batch effect observed on raw intensities.



(A) Percentage



(B) Median



(C) Median-RT

FIGURE 6.10 – Global intensity profiles of XIC after normalization by (A) percentage, (B) median and (C) median-RT. Report to Figure 6.2 for unnormalized data.

As explained in Section 6.2.2, XIC normalization methods are all based on the hypothesis that most peptides have the same intensity across all samples. This statement is true in proteomics, but not necessarily in metaproteomics. In addition, the

low proportion of common peptides between several samples is a brake on these normalization methods. Since our samples seem to have a fairly weak batch effect (Figure 6.2) and the experimental design includes randomization of patient groups, we chose not to normalize the XIC data.

6.4 Conclusion on the correction of technical variability

We observed technical variability on SC and XIC abundances on the cytosolic fraction of the ProteoCardis non-bariatric dataset. Although observable, this variability is weak. No correction to eliminate this technical variability was effective, either for SC or XIC. However, the patient groups are randomized at all levels in the experimental design (from protein extraction to injection into the mass spectrometer ; see Section 4.1.2), *i.e.* all the manipulations of the samples have been planned so that samples from each patient group were evenly distributed. The correction of the technical variability is therefore not essential, and the eventual batch effect does not generate any bias in the statistical analyses on the groups of patients. Ideally, the correction of batch effect could increase statistical power, but only if we had a relevant correction method. Our choice is therefore not to normalize the SC and XIC data, and to take into account the batch effect in the statistical analyses. This is possible in simple statistical analyses, but can be difficult or impossible with some machine-learning models.

Chapter 7

Exploration of statistical approaches for biomarker discovery

We implemented two statistical approaches on the ProteoCardis datasets in order to discover biomarkers of CAD. The first, multiple testing, focuses on finding significantly different variables across patient groups. The second, random forest, is a classification method that retrieves a restricted list of important variables to predict the patient group. These two approaches apprehend the data in a complementary way: on the one hand the multiple testing interrogates each variable separately, on the other hand the random forests brings a more global vision of the variables and their structure. The results of these two methods seems therefore relevant to mine the data from different points of view.

7.1 Methods

7.1.1 Multiple testing approach

7.1.1.1 Resampled FDR

In the multiple testing approach, we implemented a test by variable (subgroup, peptide...) to test the association between this variable and the patient group. In the case where the number k of variables is high, if we set a test level α, if no variable is associated with the patient group, a proportion α of variables will be erroneously detected. Thus, if 10 000 variables are tested at level 5%, an average of 500 variables will be erroneously detected. To take this phenomenon into account, the procedure of Benjamini-Hochberg [141] controls the proportion of false discovery, or FDR (False Discovery Rate).

This procedure works as follows:

1. the p-values of the variables are ordered in ascending order: $p_{(1)} \leq p_{(2)} ... \leq p_{(k)}$ where (l) is the index of the variable of rank l. For example, (1) is the index of the variable that has the smallest p-value.

2. we define the "corrected" p-values

$$p_{(l)}^{adj} = p_{(l)} \frac{k}{l} \tag{7.1}$$

3. we select variables such as $p_j^{BH} \leq \alpha$

Under certain hypotheses, we are guaranteed to have on average a false positive proportion less than or equal to α. Nevertheless, the hypotheses necessary for this result are complex and difficult to verify in practice (independence of the variables or absence of a particular structure of correlation). This procedure is widely used in practice without questioning these hypotheses, which can lead to bias. One way to overcome these hypotheses is resampled FDR procedure [142].

First, groups of patients are permuted, which forms "false" groups that should not be associated with any variable; the matrix of variables remains unchanged, which preserves the correlation structure between the variables. The vector of the p-values is calculated from these permuted data; this group permutation is carried out $N = 10$ times. For each permutation $r \in [1...N]$ and each variable $j \in [1, ...k]$, the p-value of j after the permutation r is written $p_j^{perm,r}$. The original p-values are rescaled from the p-values of the permuted data (Equation 7.2).

$$p_j^{rescale} = \frac{1}{N}\frac{1}{k}\sum_{i=1}^{k}\sum_{r=1}^{N} \mathbb{1}_{\{p_i^{perm,r} \leq p_j\}} \tag{7.2}$$

Finally, we apply the classic procedure of Benjamini-Hochberg to the rescaled p-values $(p_1^{rescale}, \ldots, p_k^{rescale})$

7.1.1.2 Modelling of SC

We considered a Zero-Inflated Negative Binomial (ZINB) model, classic to model count data with a high proportion of zeros. We examined the relevance of this model, compared to the simpler model of Negative Binomial (NB). We performed our tests on SC groups for non-bariatric patients, cytosolic fraction. The data were filtered to eliminate groups present in less than 16 samples, which is the number of patients in the smallest patient group. This left 3 609 protein groups to analyse.

The ZINB model has a distribution defined on the Equation 7.3, where only μ depends on the patient group.

$$\mathbb{P}[X = j|\pi, \mu, \alpha] = \begin{cases} \pi + (1-\pi) f_{NB}(0|\mu, \alpha) & \text{if} \quad j = 0 \\ (1-\pi) f_{NB}(j|\mu, \alpha) & \text{else} \end{cases} \tag{7.3}$$

These models were chosen due to the strong presence of zeros in our SC data. To compare the models, we used the Akaike Information Criterion (AIC), which uses maximum likelihood by penalizing models with too many variables, as well as the Bayesian Information Criterion (BIC) whose penalty also depends on the size of the samples. The lower the AIC and BIC criteria, the more relevant the model.

The comparison of the models showed that the AIC is favourable to the ZINB model for the non sparse protein groups (containing few zeros), but does not decide for the sparse protein groups. Conversely, the BIC criterion, which is more conservative, does not decide for the non-sparse groups and favours NB for the sparse groups (Figure 7.1). Although the gain in terms of fit of the ZINB model is not clearly attested by the chosen criteria, this model fit correctly the empirical values as illustrated in the Figure 7.2 with 3 groups selected randomly with different sparsity. In addition, the ZINB model has only one additional parameter, so the statistical power loss is moderate. So we chose to implement a ZINB model, and to use a NB model only if the ZINB model did not converge.

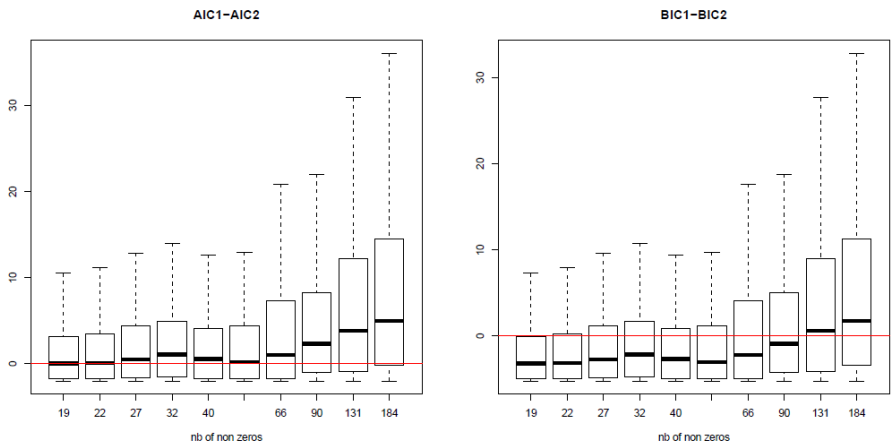


FIGURE 7.1 – Difference of AIC (left) and BIC (right) between the negative binomial model (model 1) and the zero-inflated negative binomial model (model2). The groups are clustered depending on their sparsity. Values under 0 means that the NB model was judged more relevant, and *vice-versa*.

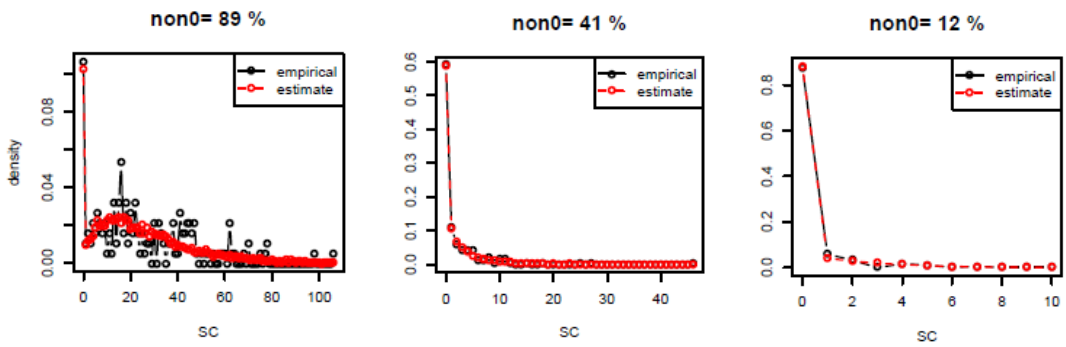


FIGURE 7.2 – Empirical SC of three groups containing 89% (left), 41% (center) and 12% (right) of values greater than 0, and their estimation by zero-inflated negative binomial model.

7.1.1.3 Modelling of XIC

Figure 7.3 shows the bimodal distribution of XIC values of specific peptides after imputation, for the non-bariatric dataset, cytosolic fraction.

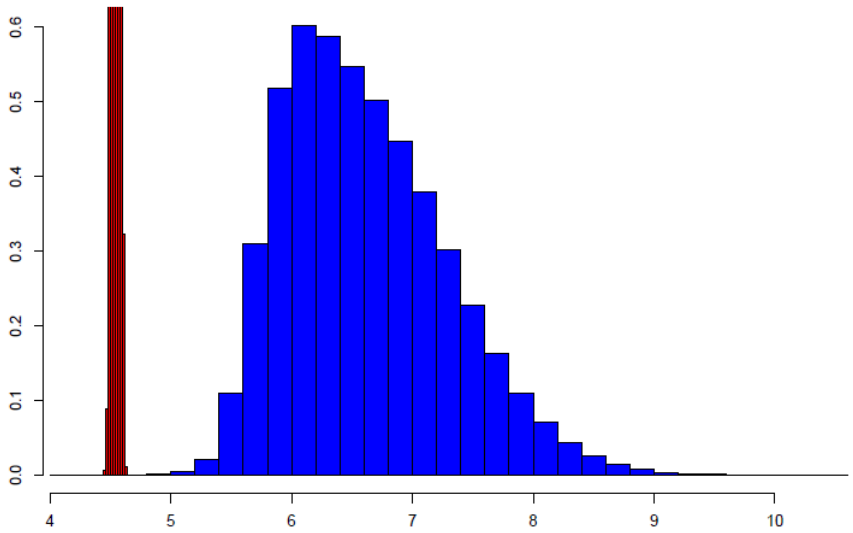


FIGURE 7.3 – Distribution of XIC abundances for the observed values (blue) or the imputed values (red). Log-scale.

Using a parametric model seemed difficult in this context. We therefore considered the non-imputed data and used non-parametric tests by combining a Fisher test and a Kruskall-Wallis test. For each variable, a Fisher test was performed for the presence/absence of the variables (where the variable is 1 if there is no missing value), and a Kruskall-Wallis test calculated after removal of the missing data. The minimum of these p-values, designated S, was considered the p-value of our test. The distribution of S under the null hypothesis is obtained by the repeatedly permutation of the patient groups, and by aggregating the variables with the same number of missing values.

7.1.2 Random forests approach

7.1.2.1 Principle

Random forests are a classification method that builds a large number of decision trees [143]. Decision trees allows us to divide a population into homogeneous groups (here, the different disease groups) according to a set of variables (here, the counts/XIC of the peptides/subgroups). They thus permit, according to different discriminant variables, to predict the response variable. To classify a new sample from a set of variables, the input is submitted to each of the trees in the forest. Each tree gives a classification (the trees "votes" for that patient group -or class-). The forest chooses the classification with the most votes.

The construction of the random forest works as followed. For each tree:

- The construction is performed using a bootstrap sample from the original data. About one-third of the cases are left out of the bootstrap sample and not used in the construction. This procedure enables the calculation of the out-of-bag (OOB) error that provides an unbiased estimate of the test set error.
- At each node, a fixed number of variables is selected at random out of the input variables and the best split on these is used to split the node. The number of variables sampled is held constant during the forest growing, this number is tuned based on the OOB error rate calculated with increasing number of variables sampled.
- There is no pruning: the growth is performed to the largest extent possible.

This method is particularly useful for multifactorial spaces, as several decorrelated trees can be generated. Indeed, if a small number of predictors tend to dominate the others, it makes them appear each time close to the root of the tree, creating correlated trees. Taking random subsets of variables allows us to build decorrelated trees which reduces the variance and the forest error rate. The robustness of the random forest is estimated by the OOB error, calculated as the mean error of the classification when using the observations not included in the model construction. The calculation of the accuracy is used to estimate the prediction error of OOB samples. It is calculated using the Equation 7.4. This metric is only relevant when the possible outcomes are of equal numbers. An unbalanced dataset makes this performance metric unreliable because the random forest classifier tends to be biased towards the majority class (*i.e.* the group of patients with the higher number of samples).

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative} \tag{7.4}$$

The importance of variables can be evaluated to see their impact on the construction of the trees with the calculation of the Mean Decrease Accuracy (MDA), using the OOB samples. After the construction of a tree, the accuracy of the OOB samples are computed. Then the values of each variable are randomly permuted between the OOB samples, and the accuracy is recalculated. This procedure is performed for all trees and averaged to give the MDA, which allows to estimate the importance of each variable in the random forest. Indeed, if the variable has little importance in the decision process, the permutation of its values between samples will have little influence on the result of the decision by the tree. Conversely, the permutation of the values of a variable essential to classification will lead to a sharp decrease in accuracy.

7.1.2.2 Typical preprocessings

Before entering the classification step, some typical preprocessing may be applied on the input data:

- Remove zero- and near-zero variance predictors: in some situations, the data generating mechanism can create predictors that only have a single unique value (i.e. a "zero-variance predictor"). Similarly, predictors might have only a handful of unique values that occur with very low frequencies. These predictors may become zero-variance predictors when the data are split into bootstrap sub-samples and a few samples may have an undue influence on the model.

- Remove correlated predictors and/or linear dependencies: when variables are highly correlated, the selection of the predictor to create a node will be random, and the importance of each of these variables will be diminished. This dilutes the importance of each of the correlated predictors and may make the variable importance measure less helpful.

7.1.2.3 Implemented parameters and validation scheme

Preprocessing: In the case of ProteoCardis, the removal of near-zero variance predictors, correlated predictors (cutoff = 0.75) and linear dependencies has been performed with the preprocessing functions of the caret package [144], with default parameters. These tree preprocessings have been performed separately in order to be compared.

Parameters: The random forest has been performed with the randomForest package [143]. The number of trees in the forest has been fixed to 500 (default parameter). The number of variables randomly selected at each split ("mtry" parameter) has been tuned on the OOB samples with the caret package based on a grid of length 10.

Set up of a robust validation scheme: since the random forest algorithm relies on bootstrapping and random variable selection, we noticed that the variability of the results for several repetitions was high. Although the general accuracy of the model was stable, the ranking of the important variables was different from one repetition to the other. Moreover, in a feature selection perspective, it was challenging to determine the optimal number of important variables to consider as all the variables are ranked. The main objectives of the set up scheme was to (i) determine the optimal number of important variables and (ii) to converge toward a list of common variables.

In order to ensure the external validation of the selected variables, we split the dataset as following: - 10% of the samples were kept apart for the "gold validation", - 90% of the samples were used for the construction of the models.

To determine the optimal number of important variables, 1% of artificial variables were added to the input variables. The artificial variables were randomly selected among the input variables, then their counts have been shuffled in order to break any link with the response variable.

Classification models and occurrent important variables: 50 repetitions of random forest classification models have been computed. Each time, the important variables were those with a MDA higher than the first artificial variable encountered. Then, the occurrence of each variable in these 50 repetitions has been calculated.

7.2 Results

For all statistical studies, we eliminated standard samples and peptides and proteins identified only in these samples. We also kept only one copy of the replicate samples, randomly selected. In the same way, we eliminated the identified peptides and proteins only present in the removed samples. We first wanted to evaluate the batch effect on the SC results, because the models with random batch effect (taking into account the batch effect in the tests) is computationally heavy.

7.2.1 Preliminary: evaluation of the batch effect

We considered a ZINB model with fixed effect of the patient group and random effect of the batch. Calculations were implemented on the SC protein groups, from which the groups missing in more than half of the samples were removed (this filtering was applied only in this preliminary section). We thus studied the batch effect on 740 protein groups. We performed the resampled FDR procedure detailed in the Section 7.1.1.1. The resampled FDR procedure detected a batch effect for 16 protein groups among the 740 with a 10% FDR, and 308 proteins groups with a 50% FDR. The batch effect is therefore existing but not very significant.

The test of the effect of the patient groups by including or not the batch effect showed that the resampled p-values were very similar (Figure 7.4).

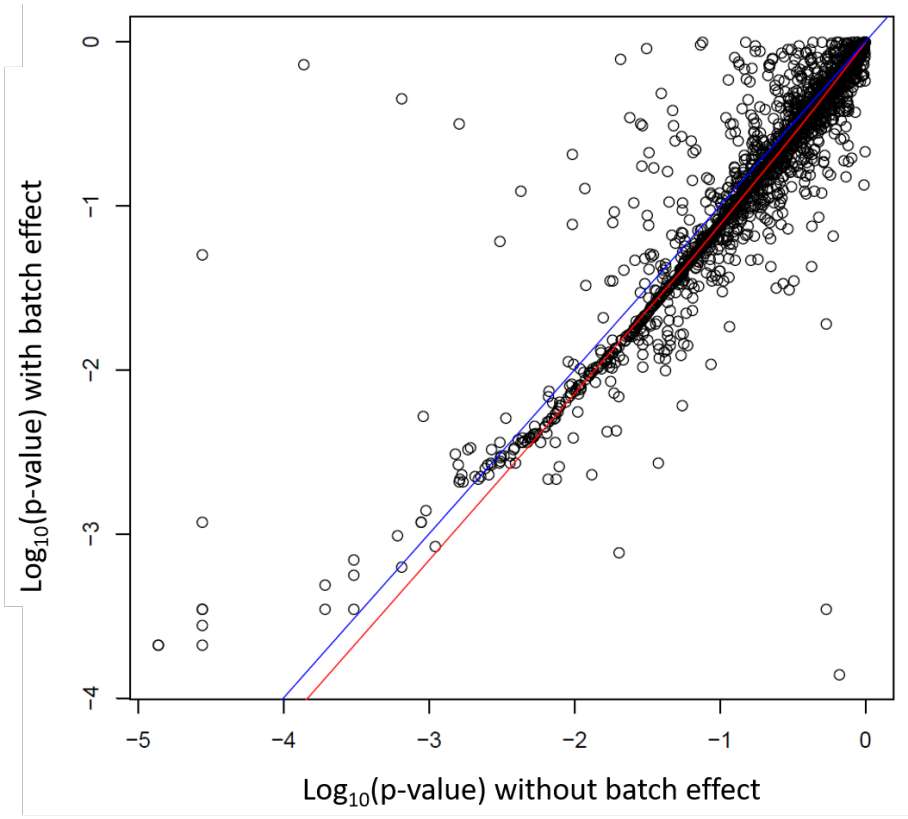


FIGURE 7.4 – Resampled p-values of the 740 proteins groups in a model without or with batch effect when the effect of patient group was tested. Red: Lowess interpolation. Blue: y=x. Log-scale.

The batch effect is therefore very weak and taking it into account is computationally heavy, especially in the context of the resampled FDR which include numerous iterations. So we decided to not consider the batch effect in future analyses.

7.2.2 Results with multiple testing

The effect of the patient groups was tested for each fraction of the SC and XIC datasets (cytosolic and envelope), and on specific peptides, subgroups (which sums only the specific peptides). Thus we did not take into account the shared peptides that can be problematic in the analyses, as developed in the section 1.3.3.3. We have implemented the following tests:

- Global group effect
- All pathologies *versus* controls
- Each pathology *versus* controls

For each test, the variables present in less than X samples are removed, where X is defined as 80% of the size of the smallest patient group. Tables 7.1 and 7.2 give

the number of variables detected with SC and XIC respectively, for each dataset and each test, for three FDR levels, for non-bariatric patients.

TABLE 7.1 – Number of variable detected for each dataset and each test, for three levels of FDR, based on SC.

Test	Dataset	FDR 0.05	FDR 0.1	FDR 0.5
Global group effect	Specific peptides cyt	3	4	476
	Specific peptides env	3	7	242
	Subgroups cyt	1	5	201
	Subgroups env	0	1	227
All pathologies *vs* controls	Specific peptides cyt	15	41	658
	Specific peptides env	27	53	674
	Subgroups cyt	6	18	287
	Subgroups env	0	54	322
aCAD *vs* controls	Specific peptides cyt	2	14	216
	Specific peptides env	6	7	272
	Subgroups cyt	6	18	114
	Subgroups env	14	34	135
cCAD *vs* controls	Specific peptides cyt	3	4	47
	Specific peptides env	1	1	175
	Subgroups cyt	0	0	75
	Subgroups env	3	10	118
cCADic *vs* controls	Specific peptides cyt	0	0	472
	Specific peptides env	0	0	125
	Subgroups cyt	0	0	125
	Subgroups env	0	2	80
cCADic ncor *vs* controls	Specific peptides cyt	21	43	372
	Specific peptides env	12	13	150
	Subgroups cyt	7	7	97
	Subgroups env	0	11	116

TABLE 7.2 – Number of variable detected for each dataset and each test, for three levels of FDR, based on XIC.

Test	Dataset	FDR 0.05	FDR 0.1	FDR 0.5
Global group effect	Specific peptides cyt	0	3	52
	Specific peptides env	0	0	64
	Subgroups cyt	2	2	4
	Subgroups env	0	1	75
All pathologies *vs* controls	Specific peptides cyt	0	0	2 833
	Specific peptides env	0	1	801
	Subgroups cyt	6	8	677
	Subgroups env	18	33	1 375
aCAD *vs* controls	Specific peptides cyt	0	0	1 044
	Specific peptides env	2	6	2 285
	Subgroups cyt	11	12	52
	Subgroups env	13	21	1 372
cCAD *vs* controls	Specific peptides cyt	0	1	99
	Specific peptides env	2	2	1 019
	Subgroups cyt	3	3	3
	Subgroups env	0	0	290
cCADic *vs* controls	Specific peptides cyt	0	0	1 121
	Specific peptides env	0	0	6
	Subgroups cyt	0	0	293
	Subgroups env	0	0	0
cCADic ncor *vs* controls	Specific peptides cyt	2	2	167
	Specific peptides env	2	2	76
	Subgroups cyt	0	1	58
	Subgroups env	0	5	42

Whatever the type of quantification considered (SC or XIC), few variables were detected for a low FDR value of 0.05, which did not make it possible to identify variables that have a high probability of being linked to patient group. However, a large number of variables were detected at a FDR of 50%. Figure 7.5 shows the p-values of SC and XIC for comparison of the aCAD group *versus* controls at the subgroup level. The filtering at 80% of the smallest population considered (here 39 because the aCAD group contains 49 patients) preserved 620 subgroups in SC and 36 411 subgroups in XIC (because the SC data contain much more zeros than XIC data contain missing values), including 600 common subgroups shown in Figure 7.5. We observed that these p-values were only slightly correlated (correlation = 0.35), the results on the data SC and XIC are therefore inconsistent. The study of the concordances between the results of SC and XIC would be an interesting work to

carry out to understand if the two types of results can be complementary.

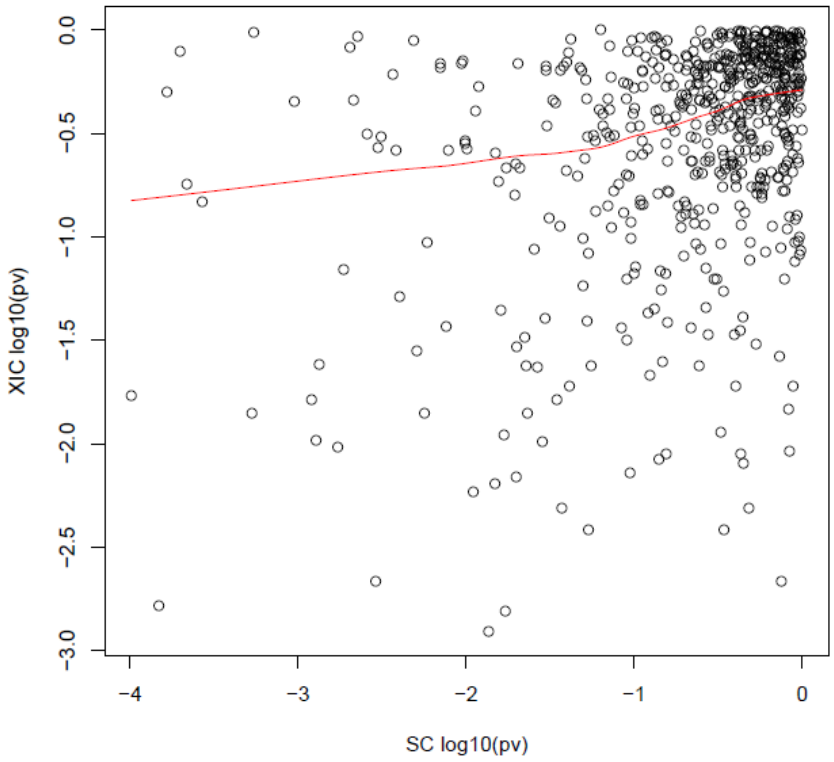


FIGURE 7.5 – P-values of SC and XIC for the subgroups, cytosolic fraction, and for the comparison between aCAD and controls. Red line: LOWESS interpolation. Log-scale.

7.2.3 Results with random forests

Different strategies exist to handle imbalanced classes, not yet implemented in this work. In our first analyses, we worked on the group SC of patient groups with balanced size, *i.e.* the group aCAD (49 patients) *versus* controls (50 individuals).

In order to converge toward the minimal list of variables important for the classification, we considered each threshold of occurrence of the important variables computed previously in the 50 repetitions. We constructed a random forest model with the important variables that have been observed a more or equal times than this threshold. For example, all variables that were judged important 12 times or more out of the 50 forests were used to construct a forest, whose OOB accuracy is reported at x=12 on Figure 7.6. This calculation makes it possible to determine which threshold maximizes the OOB accuracy, and therefore the minimal list of important variables to take into account to correctly separate two groups. It was observed here that the accuracy was maximized for x=37, so the variables observed more than 37 times in the 50 repetitions are used to build the final model. The model with the maximal OOB accuracy was kept and its performance on the gold validation dataset has been evaluated; the results are presented in Table 7.3.

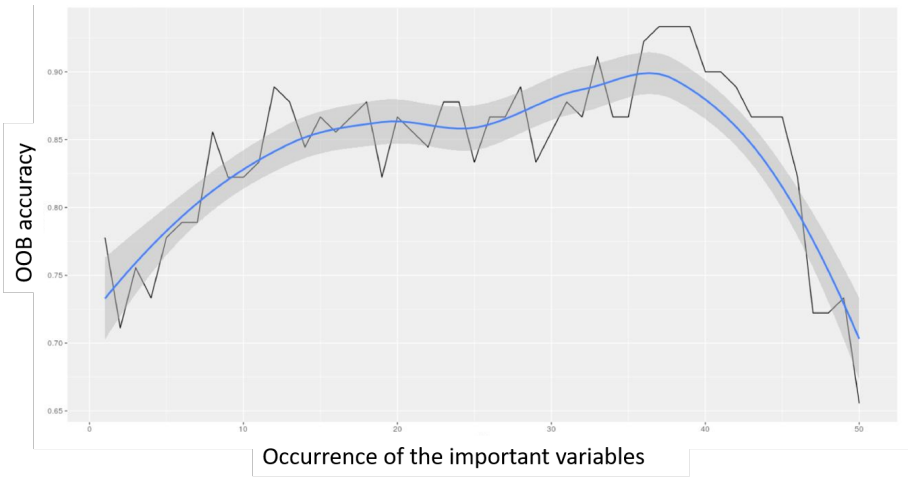


FIGURE 7.6 – Out-Of-Bag accuracy for the classification of aCAD and controls for each threshold of occurrence of important variables. In this example, the forests were computed with no preprocessing.

TABLE 7.3 – Accuracy of the forests for different preprocessings for the classification of aCAD and controls. The computation was made on group SC data. GV: Gold-Validation. nzv: removing of near-zero variance predictors. col. pred.: removing of colinear predictors. lin. dep.: removing of linear dependencies.

Preprocessing	**OOB accuracy**	**GV accuracy**	**Number of variables used in the forest**
no	0.93	0.78	16
nzv	0.93	0.78	16
col. pred.	0.92	0.56	29
lin. dep.	0.81	0.89	35
all above	0.84	0.89	72

In the context of the classification of aCAD and controls with group SC, the accuracy computed without preprocessing showed a more limited number of variables to take into account, for a OOB and Gold-Validation accuracies of more than 75%. The distribution of the 16 variables used to construct this forest is shown on Figure 7.7. Taken independently, most of them did not differ in abundance (tested with a Wilcoxon test) between the aCAD and control groups. Multiple testing approach would not have selected them as significant on the ProteoCardis study, but taken together they classify with a good accuracy the aCAD patients and the controls.

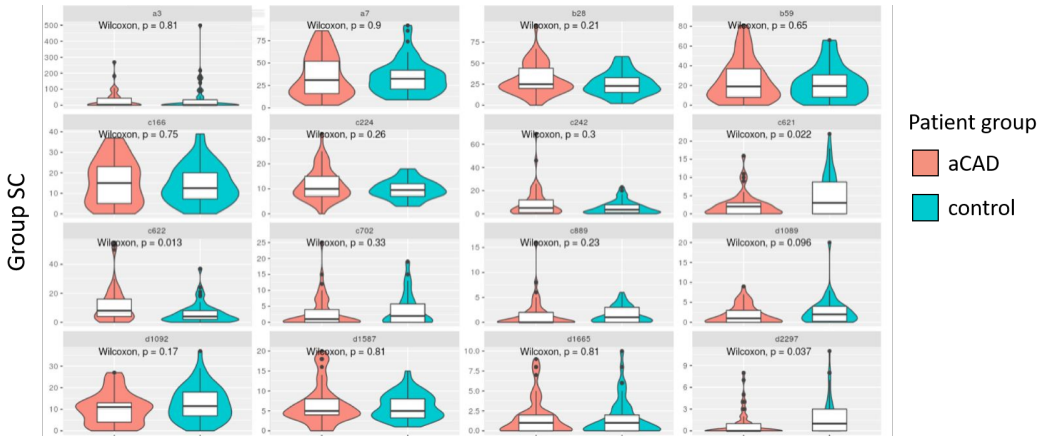


FIGURE 7.7 – Abundance of the 16 groups that better classify the aCAD and control groups.

7.2.4 Relationship between the two approaches

We were interested on linking the results obtained in multiple testing and on RF. The multiple testing approach was implemented for the group SC of patient groups aCAD *versus* control to compare with the results obtained with RF. The results showed no clear correlation between the importance of protein groups in RF and their p-value on multiple testing (Figure 7.8).

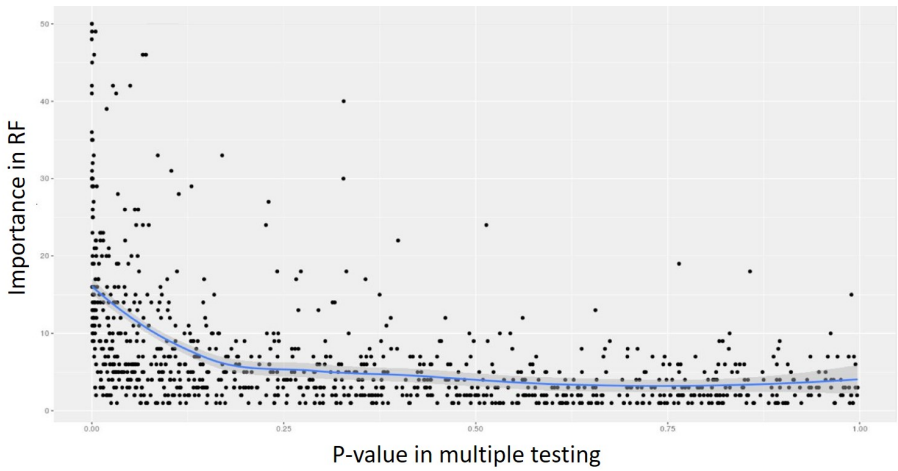


FIGURE 7.8 – Results on spectral counting of groups of the aCAD patients and control, with random forest and multiple testing. RF: random forest.

7.3 Perspectives on statistical analysis

As we have seen, the variables detected significant by the multiple testing approach are not systematically detected as important variables in the classification by random forests. However, the multiple testing approach can select the variables with the best chance of having a differential abundance between the groups of patients tested. One of the perspectives of this work is therefore to use the multiple testing approach to select, with a high FDR threshold, a limited number of variables. These variables could then be used in random forests to determine which are the most important for classification.

Chapter 8

Conclusion and perspectives

The work performed in this thesis has made possible to explore the different stages of identification of mass spectra and their quantification, with the aim of optimizing them to explore the metaproteome of hundreds of patients. My work therefore focused on bioinformatics processing of the data meant to reveal protein biomarkers for cardiovascular diseases, the downstream statistical analyses still needing development.

Since individual metagenomes are not always available, I first focused on the identification performance with two publicly available generalist databases. Although the most recent and complete database (MetaHIT 9.9) can be seen naturally as the most powerful, this performance had to be evaluated in metaproteomics, where the size of the database can be an obstacle to an accurate identification. In addition, the older and smaller database (MetaHIT 3.3) could have been sufficient to identify most peptides and proteins present in the samples. This study has shown that although using the MetaHIT 9.9 database increases the calculation time, it identifies a large number of additional proteins while keeping a higher accuracy and not reducing replicability. It is therefore a database of choice for identifying human gut microbiota proteins. Interestingly, although the number of additional proteins identified with MetaHIT 9.9 is high, the number of additional peptides identified is limited. In addition, several thousand peptides have been specifically identified with MetaHIT 3.3. This finding opens interesting avenues of work for the construction of generalist databases in the context of metaproteomics analyses of the human gut microbiota, where elimination of redundancies from metagenomic catalogues may have a tremendous influence in metaproteomics.

When the individual metagenomes are available, they represent the more natural database for the identification of peptides and proteins in the samples. I participated in the development of MetaRaptor, which I used to generate individual metagenomes by assembling the sequencing reads of the ProteoCardis cohort. The particular set of parameters, designed to generate high-quality catalogues for metaproteomics downstream processings, was tuned on a mock community. However, this community had been sequenced with a different technology than the one used for ProteoCardis. A smaller mock community, sequenced by Ion Torrent^{TM} technology was generated at MetaGenoPolis during my thesis. Assessing the assembler's

performances on this type of data would complement the assessment performed in this thesis. Our team also plans to sequence the mock with the most recent sequencing technology producing longer reads (MinIon, Oxford Nanopore Technologies). The validation of MetaRaptor for assembling metagenomes sequenced with various sequencing technologies is meant to prepare a public release of the software, which could be used by the community and advantageously serve metaproteomics for individual metagenome assembling.

The individual catalogues assembled with MetaRaptor were used as reference databases for the identification of peptides and proteins in the ProteoCardis samples and the results were compared with those obtained with MetaHIT 9.9. Surprisingly, our results indicate that the generalist database identified more peptides but less proteins than the individual databases. Despite our efforts to explain the reason why individual catalogues identified so many proteins with less peptides, we could not solve it, and this remains an open question, which has to be further explored. I also tested identification with a combination of the generalist database with the individual personal database. However, the explosion of computation time of the grouping step observed on 40 samples made it impossible to implement for the whole ProteoCardis cohort, and the interest of such approach seemed limited considering the few supplementary peptides (compared with MetaHIT 9.9 only) identified in 40 samples. Another option to explore is the merge of all the individual catalogues into a wide one, which therefore represents all the samples of a cohort. This merged database identified more peptides and proteins than identification with individual databases in the study of three human stool samples [40]. Nevertheless, we chose in our approach to keep all the individual variability present in the samples, which made it impossible to use the merged database due to its size (more than 40 millions of genes). A possibility would therefore to remove the redundancy from the merged database and to evaluate its interest compared to the individual databases and MetaHIT 9.9. In alternative to the strict individual metagenomes, we could consider a catalogue built with metagenomics and metatranscriptomics data [44], which would reflect specificity of individual metagenomes but with a better coverage.

I also compared different interrogation strategies, using a simple query ("classical strategy") or an iterative interrogation in three steps ("iterative strategy"). Iterative interrogations had already been shown to be more efficient for interpreting the metaproteomic mass spectra, but the magnitude of this efficiency had never been assessed on several tens of samples. The iterative strategy in three steps showed higher peptide and protein identifications than the classical interrogation, while having a higher accuracy and did not reduce the replicability. The higher performance of iterative interrogation and the MetaHIT 9.9 database was also validated on the MICI-Pep samples.

I studied the replicability of the identifications on 14 technical replicates; such a study had never been conducted on so many replicates, and showed that the depth

of metaproteomic analyses is still not efficient enough to capture all the sample diversity, even with as many as 14 repeated mass spectrometry analyses. This study showed that the replicability of protein identification was higher with MetaHIT 9.9 than with MetaHIT 3.3, but it was not the case for peptides. Replicability was not affected with an iterative interrogation, which had never been questioned. Finally, we proposed an estimator of the probability of reproducibility as a function of the SC abundance; this quantity could be used to filter out observations with weak signal.

MS1 ion current measurement is the method of choice for quantification in label-free shotgun proteomics, but still in its infancy in metaproteomics. I showed that chromatogram alignment was challenging in metaproteomics because the number of spectra and the sample complexity are very wide. The matrices of XIC quantification produced in this context may therefore lack reliability. In order to improve chromatogram alignment, we could perform iterative alignments between samples with the most common identified peptides, instead of using a unique reference for all the remaining samples.

I evaluated the performance of several SC and XIC normalization methods, but our results not clearly indicate a superiority of any of the methods tested. In fact, thanks to extreme precautions taken for preparation and injection of samples, monitoring of mass spectrometer performances, and routine cleaning of the spray needle and column, the technical variabilities were weak, particularly illustrated by the very low variability of XIC abundances. Therefore, I demonstrated that normalization was not necessary or could even be deleterious in our particular case. This emphasizes the importance of a thorough examination of raw and normalized data before any decision on pretreatment. In addition, the balanced experimental design guarantees unbiased statistical analyses, even if weak batch effects cannot be corrected.

I performed early descriptive analyses on two studies of the human gut microbiota, in the context of weight loss during a low-calorie diet and inflammatory bowel diseases. These studies address problematics related to prediction (for MICI-Pep) or longitudinal studies (for ObOmics). Extensive statistical analyses of these studies may provide biomarkers of interest.

In the ProteoCardis study, the problematic is to extract variables of interest in the context of CAD patients. The statistical analyses conducted on unnormalized data confirmed that the weak batch effect did not induce bias in our results. Both univariate (multiple testing) and multivariate (random forests) statistical approaches have highlighted a signal to differentiate groups of patients. However, the variables of interest obtained with the two approaches as well as those obtained with the two quantification methods (SC and XIC) did not overlap. The functional and taxonomical annotation of the proteins could help us to decipher the possible links between them. One of the opportunity considered is to use multiple testing procedure with a high threshold of FDR to select a set of candidate variables, that will be used in a

classification model. Combining (i) statistical approaches, (ii) quantification methods and (iii) different levels of analysis (peptides, subgroups, groups) is currently the main challenge to overcome. Among the potential proteins of interest that will be discovered, the functional annotation will help us selecting peptides/proteins related to CAD, with the aim of selecting about 50 peptides of interest. Their predictive value will be assessed on an independent patient group, beyond this thesis work.

To conclude, this thesis enabled to in-depth explore the methodologies of peptides and proteins identification and quantification in the context of large-scale metaproteomic studies of the intestinal microbiota. With the growing development of metaproteomics and the significant interest given to the intestinal microbiota as a health partner, it lays the necessary methodological bases for other studies similar to ProteoCardis. It also opens interesting avenues of research to further optimize the identification of protein and peptide biomarkers, and for the statistical analysis of such large-scale and complex studies.

Appendix A

Polymerase Chain Reaction

PCR is a method of genic amplification *in vitro*. It's goal is to exponentially duplicate DNA segments. This method is based on thermal cycles to drive polymerisation by DNA polymerase and separation of double-stranded DNA by denaturation.

The PCR consists of twenty to forty repeated cycles, each cycle consisting in three steps:

1. **Denaturation:** The solution is heated to 95°C for 30 seconds to denature the double-stranded DNA.

2. **Annealing:** The solution is cooled to 50-65°C for 1 minute to anneal the primers to the DNA, targeting the DNA region of interest. Usually, sens and anti-sens primers are used to target the complementary region of interest. The DNA polymerase also binds to the primers.

3. **Elongation:** The solution is heated to 72°C for 1-2 minutes, which allows the polymerase to synthesize the complementary DNA with free dNTPs in the solution. The time of this step depends on the length of the DNA to amplify. At the end of this step, an original double stranded DNA has therefore been duplicated in two double-stranded DNA.

The PCR procedure is illustrated in Figure A.1

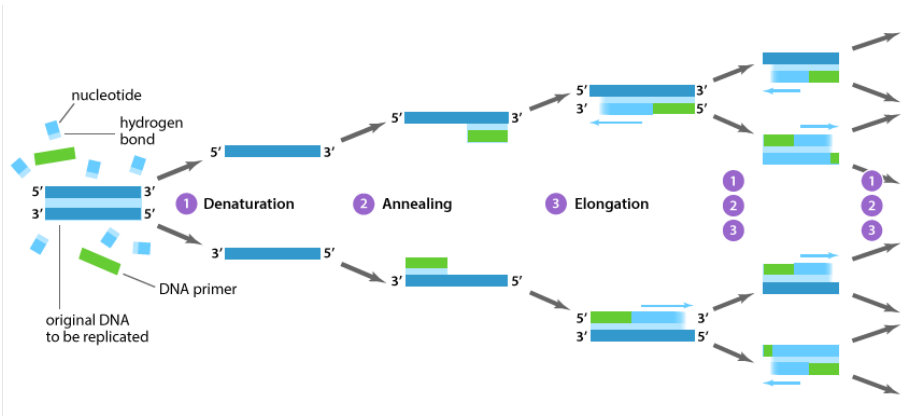


FIGURE A.1 – The different steps of a PCR, from [145].

Appendix B

Physiopathology of atherosclerosis

The healthy vessels are composed of three layers (Figure B.1) , from inner to outer layer:

- **The intima**, composed by a single layer of epithelium cells and connective tissues supporting the internal elastic lamina.
- **The media**, consisting of elastic fibres, collagen, smooth contractile muscles and external elastic lamina. The smooth muscles controls the local blood pressure by contraction, modifying the diameter of the lumen. The external elastic lamina is absent in veins.
- **The adventitia**, composed of connective tissues, elastic fibres, nerves and nutrient capillaries (*vasa vasorum*).

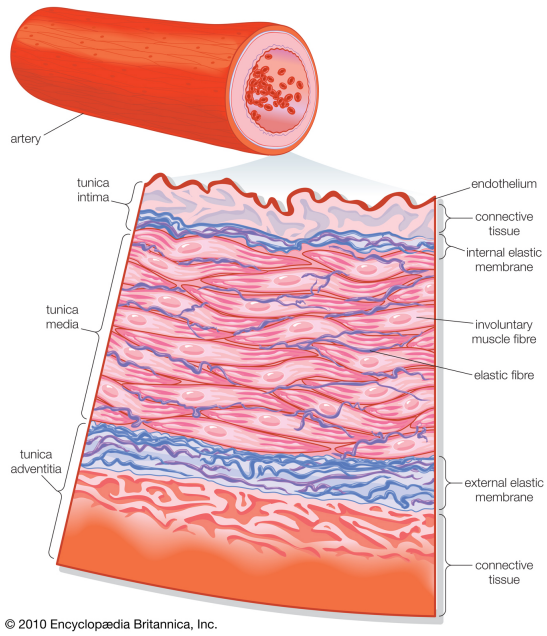


FIGURE B.1 – Healthy artery layers, from Britannica [146]

Atherosclerosis is an inflammatory disease which develops in six evolutionary stages, the early stages developing in childhood, the latter usually after 40 years.

Stage I of atherosclerosis is characterized by the presence of some foam cells in the intima. It begins especially in vessels where the blood flow can be non-uniform (curvatures or branches), which induces a shear stress on the inner wall of the arteries, the intima. This stress induces inflammatory lesions, which causes nitric oxide (NO) production and activation of integrins and adhesion factors to reduce the inflammation. In this context, the endothelium cells also secrete extracellular matrix proteins and metalloproteinases, in order to repair the tissue. However, the permeability of endothelial lesion and the retaining of low density atherogenic lipoproteins (LDL) by extracellular matrix induce infiltration of LDL into the intima. The integrins and adhesion factors permits monocytes to also enter in the intima, where they differentiate into macrophages due to the inflammatory process. The LDL is oxidized in LDL-ox by free radicals secreted by macrophages, which internalize the LDL-ox and become foam cells.[147, 148]. The non-foamy macrophages produce inflammatory cytokines and metalloproteinases.

Stage II of atherosclerosis is characterized by intracellular lipid accumulation, forming fatty streaks. The process of the stage I self-perpetuates, as LDL-ox in the intima leads to an inflammatory response which retains and recruit more LDL and monocytes. The foam cells have a pro-atherogenic activity by secreting pro-atherogenic molecules which maintain the inflammation and shifts the phenotype of the smooth muscle cells (SMC) of media from a contractile to a pro-inflammatory phenotype. In healthy context, SMC mainly produces proteins involved in the contractile function, while in the context of atherosclerosis, SMC migrates from the media to the intima, becomes proliferative, and express extracellular matrix and cytokines, maintaining the pro-inflammatory state. They also internalize lipids, becoming foam cells and contributing to the formation of the fatty steaks [72, 149].

Stage III of atherosclerosis is characterized by the aggregation of droplets of lipids in the extracellular matrix. The foam cells apoptosis, driven by LDL-ox, causes the aggregation of the cell debris in the intercellular matrix, which lower the amount of pro-atherogenic cells, but leads to the accumulation of pro-inflammatory metabolites like lipids. These lipids aggregate as droplets. Moreover, apoptotic macrophages that are not rapidly ingested by nearby phagocytes may become necrotic (post-apoptotic macrophage necrosis), which is a source of proinflammatory stimuli and thus can elicit an inflammatory response and cause damage to nearby cells [150].

Stage IV of atherosclerosis occurs when the lipids droplets form a lipid core which contains crystals of cholesterols, few giant foam cells and cellular debris from apoptotic and necrotic macrophages. The lipid core is also called the atheroma.

Stage V of atherosclerosis is characterized by a fibrotic layer around the lipid core. The fibrotic layer (or cap) is formed by the production of collagen and proteoglycans by SMCs. The atheroma increases the thickness of the intima, which narrows the diameter of the vessel lumen. The atheroma with a fibrotic layer is called a plaque, which can be stable if the fibrotic layer is thick enough, because it ensures the stability of the plaque [72].

Stage	Description	Age of occurence (years)
I	Isolated macrophage foam cells	0-15
II	Fatty steaks ; intracellular lipid accumulation	0-15
III	Droplets of lipids in the extracellular matrix	15-35
IV	Lipid core	15-35
V	Lipid core and fibrotic layer	40+
VI	Breaking of atheroma, thrombus	40+

TABLE B.1 – The six stages of atherosclerosis and age of occurrence [150, 151]

.

Stage VI of atherosclerosis occurs when the atheroma breaks, causing the lipid core be in contact with the blood. The break can be caused by the degradation of a thin fibrotic layer by metalloproteinases. The lipid core is rich in thrombogenic elements such as cellular debris of apoptosis and tissue factors, which initiate coagulation by activate platelets, generating a thrombus. This thrombus can occludes partially (non-occlusive wall thrombosis) or completely (occlusive thrombosis) the vessel, which triggers acute vascular events [72, 148].

The stages of atherosclerosis are summarized in the table B.1, and a summary of the plaque formation is illustrated in the figure B.2.

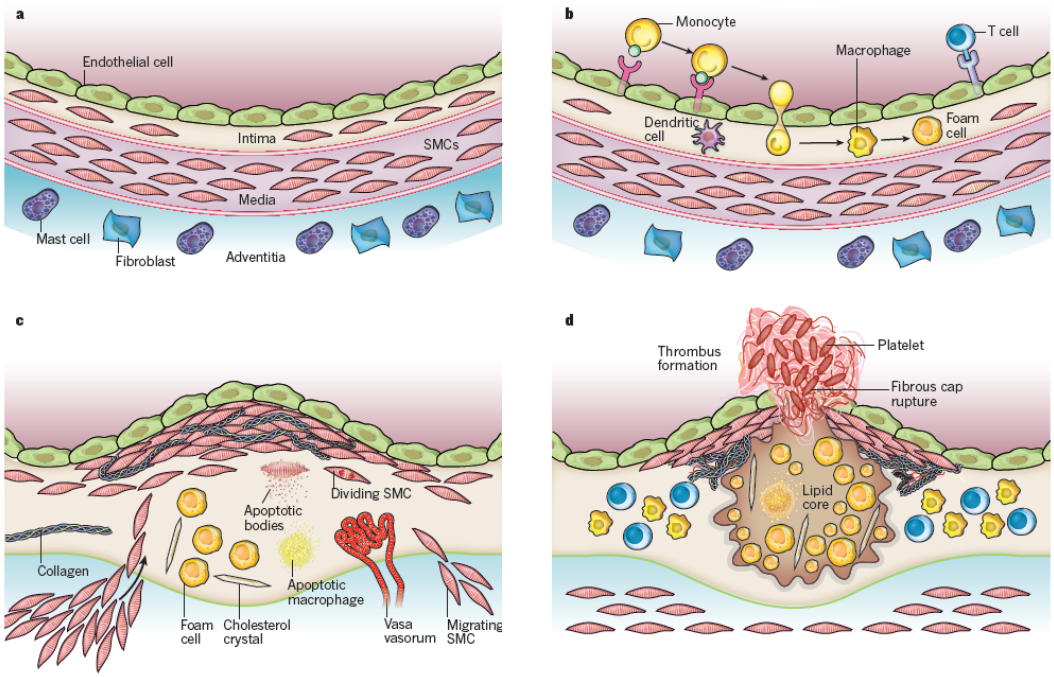


FIGURE B.2 – Stages of the development of atherosclerosis. **(a)** The healthy artery layers. **(b)** In early stages, due to the presence of LDL-ox, monocytes migrates to the intima layer, maturing into macrophages. They forms foam cells by the uptake of LDL-ox and secrete pro-inflammatory molecules. **(c)** Due to the pro-inflammatory context, SMC migrates from the media to the intima, secretes extracellular matrix macromolecules and becomes proliferative. The extracellular matrix macromolecules forms the fibrous cap of the atheroma. The death by apoptosis of SMC and foam cells free lipids in the extracellular matrix, which can form a lipid core. **(d)** The disruption of the atherosclerotic plaque put in contact the lipidic core and the blood components, which triggers the formation of a thrombus by coagulation. The thrombus, extending to the lumen of the vessel, can disrupt the blood flow. From Libby, Ridker, and Hansson [152].

Appendix C

Methods for the ObOmics study

C.1 Stool Sample Collection and Processing

All samples were self-collected as previously detailed [30]. About 1g stool aliquots were cut frozen and the microbiota was separated from the faecal matrix by flotation in a preformed Nycodenz continuous gradient according to a variant of the method previously detailed by Juste et al. Here, we just reduced the size of the gradient. Briefly, stool specimens were supplemented up to 2.82 g with 1X PBS-0.03% w/v Na-deoxycholate, then with 8 ml of Nycodenz 60%, and 6.5 ml of this suspension were loaded below a preformed gradient which has been prepared with 5 ml of a 23% w/v Nycodenz solution in an Ultra-Clear centrifuge tube (14.5 X 95mm, Beckman Instruments, CA, USA). During low-speed ultracentrifugation in a swinging SW 40 Ti rotor (Beckman, 14,567 × g, 45 min, 4°C), bacterial cells migrated up to their buoyant density (d 1.110-1.190) while the unwanted faecal matrix sedimented. After washing in cold Tris saline (20 mM Tris, 138 mM NaCl, 2.7 mM KCl, 0.03% w/v Na-deoxycholate, pH 7.4), the extracted microbiota were frozen in liquid nitrogen then kept at –80°C in 2 mL screw cap Sarstedt tubes. For bacterial lysis, 1.5 mL of cold saline Tris-EDTA buffer (50 mM Tris-HCl, pH 7.8 containing 150 mM NaCl and 1 mM EDTA, and extemporally supplemented with PMSF at a final concentration of 2mM and protease inhibitor cocktail (cOmplete™, EDTA-free Protease Inhibitor Cocktail, ROCHE) at a final concentration of 1.3X), was directly added to each frozen bacterial pellet. The pellets were dispersed by vigorous vortexing and sonicated on ice using a 3 mm diameter probe in short intervals of 10 sec ON / 10 sec OFF, with 20% amplitude, and for two 5 min periods separated by a 15 min break on ice with periodic vigorous vortexing. Finally, the suspension were centrifuged at 5000 × g for 30 min at 4°C to remove unbroken cells and large cellular debris. The supernatant was ultracentrifuged in a swinging rotor (SW 55 Ti, Beckman) at 220,000 × g for 30 min at 4°C to separate cell envelopes (pellet) and cytosolic fractions (supernatant).

C.2 Protein Digestion and Peptide Desalting

Cell envelope-enriched pellets were resuspended in 100 µL MilliQ water with a Microman pipette fitted with a 50-µL capillary piston, and then sonicated and vortexed

until homogeneous suspensions were obtained. Proteins were then precipitated on ice with 6 volumes of pure ice-cold (-20°C) acetone. The suspension was chilled down to -20°C with vigorous vortexing at 10 min intervals over the first 30 min, and then left at -20°C overnight. The next morning, suspensions were transferred for 10 min at -80°C before being centrifuged at 16,000 × g, 4°C for 10 min. Supernatants were decanted and the protein pellets were washed once with ice-cold (-20°C) acetone 80% in MilliQ water. Protein concentration was determined at this stage using the 2D-Quant kit from GE Healthcare, and aliquots equivalent to 50 µg protein were frozen in liquid nitrogen then kept at -80°C. Reduction, alkylation, and liquid digestion in the presence of Trypsin Gold in a trypsin-to-protein ratio of 1:50 (w:w), and of ProteaseMax as the surfactant (final concentration 0.05%), was essentially as recommended in the Promega technical bulletin. Finally, peptide mixtures were desalted on 360 mg Sep-Pak® Plus Short tC18 cartridges with 35% acetonitrile in the final elution step.

C.3 LC-MS/MS Analysis

HPLC was performed on an Eksigent NanoLC-Ultra system (Eksigent, Les Ulis, France). Trypsic digestion products (7 µg) were loaded, concentrated and desalted on a precolumn cartridge (BIOSPHERE C18, 5 µm; column: 100 µm i.d., 2 cm; NanoSeparations, Nieuwkoop, The Netherlands) with 0.1% HCOOH at 7.5 µl.min-1 for 3 min. The precolumn cartridge was connected to the separating column (Acclaim PepMap100, 3 µm, 100 Å, 75 µm i.d. × 50 cm, Thermo fisher) and the peptides were eluted with a non-linear gradient from 5 to 35% ACN in 0.1% HCOOH for 180 min at 300 nL.min-1.
On line analysis of peptides was performed with a Q-exactive mass spectrometer (Thermo Fisher Scientific, USA), using a nanoelectrospray ion source (non-coated capillary probe, 10 µ i.d.; New Objective, Woburn, MA, USA). Peptide ions were analyzed using Xcalibur 2.1 with the following data-dependent acquisition steps: (1) full MS scan (mass-to-charge ratio (m/z) 300 to 1,400, resolution 70,000) and (2) MS/MS (normalized collision energy = 30%, resolution 17 500). Step 2 was repeated for the 12 major ions detected in step 1. Dynamic exclusion was set to 60 s.

C.4 LC-MS/MS interpretation

For all identifications, four types of modifications were searched: carbamidomethylation of cysteines (fixed modification), oxidation of methionines, excision of the N-term methionine with or without acetylation, and cyclization of N-term (potential modifications). The mass tolerance was set to 10 ppm for the parent peptide and 0.02 Da for the fragments. One miscleavage was allowed.

C.5 Taxonomic and functional annotation

The proteins were annotated taxonomically with Diamond. Their nucleic sequence were blasted against the non-redundant database of NCBI with a e-value threshold of 10^{-4}. The taxonomic assignation of the hit with the better bitscore was designated as the taxonomic assignation of the subject protein.

The functional annotation was performed with KEGG, which is a database which includes genomic, gene products and biological pathways.

Appendix D

Methods for the MICI-Pep study

D.1 Volunteers and Sample collection

We conducted a cross-sectional study including twelve patients with active Intestinal Bowel Disease (IBD; ten women and two men, aged 24 through 51 years) and eight healthy controls (CTRL) matched for age, sex and weight. Patients were followed and hospitalized in the Hepato-Gastro-Enterology Department of the Saint-Antoine hospital (Paris). We made a rigorous selection of different phenotypes for this pilot study : seven patients were diagnosed for an active ulcerative colitis (UC), and five patients for an active Crohn's disease (CD), either with ileo-colic (CDIC, n=2) or exclusive colonic (CDC, n=3) localization. Exclusion criterion was the use of antibiotics within the preceding 2 months, but all patients were treated with either salicylic derivatives, or immunosuppressants, or anti-TNF or monoclonal antibodies, or a combination of these therapies. The control group comprised healthy volunteers with neither symptoms nor a family history of gastrointestinal disease, and with no use of medication. All participants gave oral and written consent to the protocol that was approved by the ethics committee of the hospital. In addition to the plasma samples collected in the course of this study, twenty-six additional plasma samples were provided by the biobank of the Saint-Antoine hospital.

D.2 Preparation of microbiota

Every participant was asked to provide a single fresh stool sample collected in a Stomacher 400 plastic bag (Seward Medical), which was left open in a one-litre hermetic plastic box containing a catalyst (Anaerocult, Merck, Darmstadt, Germany) to generate anaerobic conditions. This faecal material was maintained in a coolbox and transferred within 2 hours into an anaerobic chamber (90% N_2, 5% H_2 and 5% CO_2) for processing. The microbiota were extracted immediately from the fresh donations and the extraction was repeated from the same stool specimens that had been frozen for two months at -80°C, in order to select markers that are valid in the case where the samples should be routed to a distant diagnosis centre. The extraction procedure was that previously detailed in Appendix C, except that Nycodenz® was

replaced by OptiPrep™ in the gradients. Here again, focus was on the envelope-enriched fractions of the microbiota according to the same fractionation method as previously described (Appendix C), as this subcellular fraction does represent the first line of interaction with the host.

D.3 Metaproteomic analyses

Purification and digestion of proteins were according to SOPs previously detailed (Appendix C) except that the trypsin enhancer surfactant ProteoaseMAX™ was replaced by the non-ionic surfactant ALS-400 (Progenta™). Forty microbiota LC-MS/MS analyses (twenty from freshly extracted microbiota and as many from post-freezing extractions, 4 µg proteins injected) were carried out in a completely randomized design, with five additional well-distributed bulk samples, and a blank between each injection.

D.4 LC-MS/MS analyses

The analyses of peptides was obtained using UltiMate™ 3000 RSLCnano System (Thermo Fisher Scientific) coupled either to Orbitrap Fusion™ Lumos™ Tribrid™ mass spectrometer (Thermo Fischer Scientific). Trypsic digestion products (5 μ g) were loaded, concentrated and desalted on a precolumn cartridge (stationary phase: C18 PepMap 100, 5 μm; column: 300 μm x 5 mm) and desalted with a loading buffer 2% ACN and 0.08% TFA. After 4 min, the precolumn cartridge was connected to the separating RSLC PepMap C18 column (stationary phase: RSLC PepMap 100, 3 μm; column: 75 μm x 500 mm). Elution buffers were A: 2% ACN in 0.1% formic acid (HCOOH) and B: 80% ACN in 0.1% HCOOH. The peptide separation was achieved with a gradient from 0 to 35% B for 160 min at 300 nL/min, then 50% B for 170 min at 300 nL/min . One run took 195 min, including the regeneration and the equilibration steps at 98% B. Peptide ions were analysed using Xcalibur 4.1.5 with the following data-dependent acquisition steps: (1) full MS scan (mass-to-charge ratio (m/z) 400 to 1 600, resolution 120 000) and (2) MS/MS (HCDOT, collision energy = 30%, resolution 15 000). Step 2 was repeated in top speed mode with a cycle time equal to 3 seconds. Dynamic exclusion was set to 60 s. Mass data interpretation was carried out as detailed in Section 2.3.3, i.e., either by one-step interrogation of the concatenated databases MetaHIT 3.3, *Homo sapiens* Swiss-Prot-TrEMBL (release April 2018) and contaminants, or three-step interrogation of the concatenated databases MetaHIT 9.9, Homo sapiens Swiss-Prot-TrEMBL (release April 2018) and contaminant, with the same peptide e-value as that previously used (0.05). Importantly, in the case of plasma mass data interpretation, albumin was removed from the contaminant database. The grouping of proteins was done as previously described in Chapter 1.3.3.2. For all identifications, four types of modifications were searched: carbamidomethylation of cysteines (fixed modification), oxidation of methionines,

excision of the N-term methionine with or without acetylation, and cyclization of N-term (potential modifications). The mass tolerance was set to 5 ppm for the parent peptide and 10 ppm for the fragments. One miscleavage was allowed.

D.5 Search for contrasts

Abundance of proteins was approached by the sum of their specific spectral counts. In a preliminary analysis, we found that the protein profiles of faecal microbiota prepared from the same sample either fresh or frozen, were closely related as illustrated by the correlation matrix in Figure 3.4. Therefore, all searches for contrasts were done on the pooled freshly and post-freezing prepared microbiota, giving unique lists of markers that can be useful for further developments of routine clinical tests for all types of samples. Given the low number of individual faecal samples, we applied a highly stringent selection, only retaining those proteins that were either strictly overrepresented or underrepresented in one group compared to another group. An iterative strategy was applied, starting with the search for markers that distinguished between all IBD samples and all CTRL, then refining search for contrasts between the three IBD phenotypes. In the particular case of plasmas, we applied a somewhat less stringent selection based upon quantile calculation, retaining proteins whose abundance was strictly higher or lower in 95 to 80% of the samples from one group compared to the higher or lower value of another group.

D.6 Taxonomic and functional annotation

The proteins were annotated taxonomically with Diamond. Their nucleic sequence were blasted against the non-redundant database of NCBI with a e-value threshold of 10^{-4}. The taxonomic assignation of the hit with the better bitscore was designated as the taxonomic assignation of the subject protein.

The functional annotation was performed with KEGG, which is a database which includes genomic, gene products and biological pathways.

Appendix E

Parameter validation of assembly on a mock community

A mock community is a mixture of bacterial DNA created *in vitro* whose composition is known, to simulate a bacterial sample while controlling the composition. I used MetaRaptor with the parameters discussed herebefore for assembling the sequencing reads of a mock community taken as a reference and provided by the assembly software MOCAT [153]. This community is composed of 20 bacterial species, one archaea (*Methanobrevibacter smithii*), and one eukaryotic specie (*Candida albicans*) classically found in human microbial communities [154], and whose reference genomes are known. The DNA of this bacterial mixture was sequenced for the MOCAT publication and publicly available, so I used the reads from this sequencing to test the performance of our assembly software, MetaRaptor.

The assembly of this mock sample predicted 45 432 scaffolds from 56 to 61 283 nucleotides, for a total of 30 259 499 nucleotides assembled. N50 and L50 are statistics widely used in metagenomics to evaluate assemblies. The N50 is the size of the contig which, along with the larger contigs, contains half of the total number of nucleotides assembled. The L50 is the smallest number of contigs whose length sum contains half of the total number of nucleotides assembled. The N50 of our assembly is 1 057 and the L50 is 5 468. These scaffolds are composed of 45 434 contigs, so contig scaffold reconstruction has been inefficient because there are as many scaffolds as contigs. 41 284 bacterial genes were predicted on these scaffolds, including 10 835 complete genes, that is to say having a start and a stop.

I first evaluated the quality of the scaffolds produced by mapping the raw reads on the scaffolds. 84.4% of the reads were mapped to scaffolds and the scaffolds coverage, *i.e.* the percentage of nucleotides' scaffolds on which reads mapped, was 99.9%; the scaffolds thus represent the original reads. Unmapped reads could be reads that contained sequencing errors.

I then assessed the quality of the predicted genes by performing a BLAST between these genes and the reference genomes of the microbial species present in the mock community. Among the 10 835 complete genes predicted, 98% were blasted on at least one reference genome with a e-value threshold of 10^{-2}. In comparison, the results obtained with the MOCAT genome assembler on the same data set were 1

042 complete predicted genes, of which 89.3% blasted against the reference genomes [153]. We thus obtain ten times more genes, these genes being of good quality since they were indeed present on the reference genomes.

Finally, I evaluated the depth of the genome reconstruction by blasting the contigs and genes produced by the assembly on the reference genomes and I calculated the coverage of the genomes. The results are shown on Figure E.1.

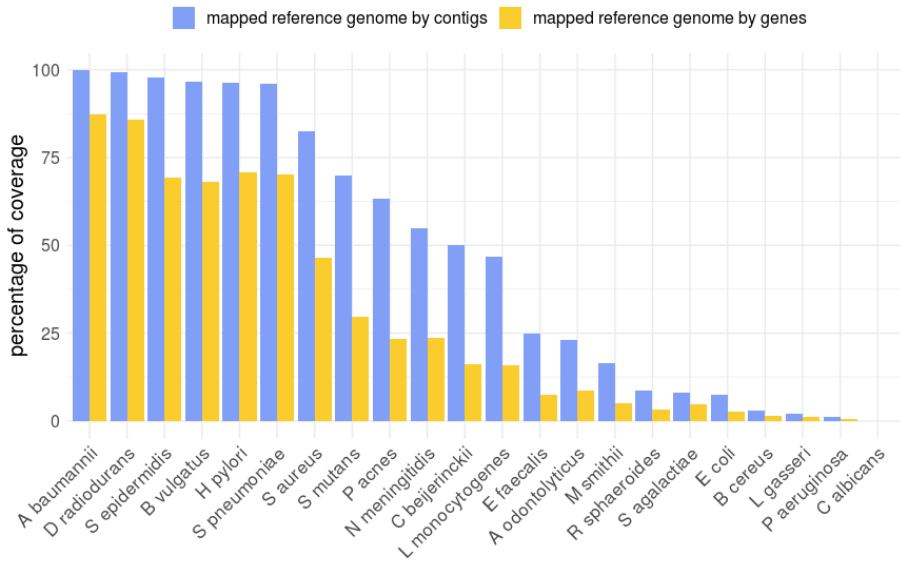


FIGURE E.1 – Coverage of the reference databases of the mock community by contigs and genes assembled with MetaRaptor.

We observed that some bacterial species were well represented by contigs. In contrast, other species (*E. faecalis*, *A. odontolyticus*, *R. sphaeroides*, *S. agalactiae*, *E. coli*, *B. cereus*, *L. gasseri*, and *P. aeruginosa*) were almost absent (less than 10% of the genome) after assembly. The archaea had a low genome coverage by genes (5.1%), which was still much lower for the eukaryota (0.004%). This is explained by the choice of the gene predictor, specific to bacteria.

In the species covered at more than 90% by contigs, the genes cover 70 to 88% of the reference genome. These results are consistent with the percentage of coding regions of prokaryotic DNA [155].

Thus, some bacterial species were less represented than others by the genes and contigs resulting from sequencing and assembly. We do not know if this result is due to sequencing or assembly bias. They are nevertheless consistent with the experiments conducted by Kultima et al. in the publication of MOCAT [153], which shows the differences between the relative abundances of the species present in the mock (Figure E.2, log scale).

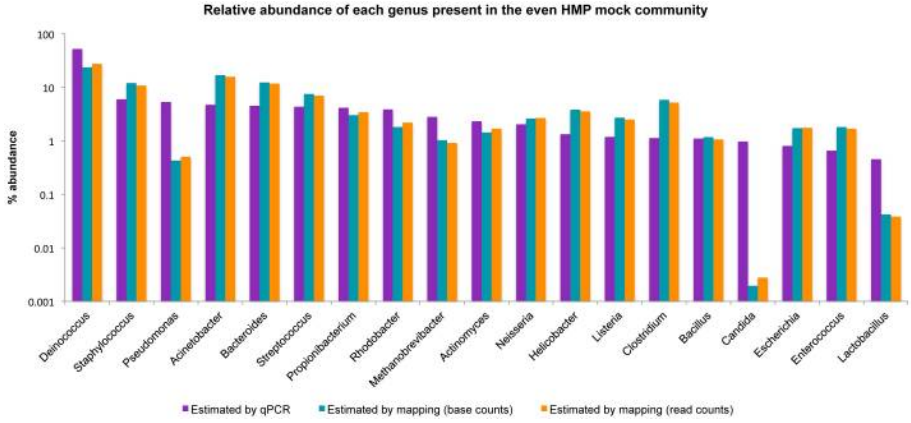


FIGURE E.2 – Relative abundance of each genus present in the even HMP mock community, from Kultima et al. [153].

When the reads were mapped to the reference genomes, the least abundant species correspond to the least covered species in our experiment (*Candida albicans*, *Lactobacillus gasseri*, *Pseudomonas aeruginosa*, *Bacilus cereus*, *Methanobrevibacter smithii*, *Actinomyces odontolyticus*, *Escherichia coli*, *Enterococcus faecalis*, *Rhodobacter sphaeroides*). Since these reads are those that were used to test the MetaRaptor assembly software, the bias observed on Figure E.1 is probably due to a sequencing bias.

Appendix F

Scientific contributions

F.1 Poster communications

"Metaraptor: An integrated pipeline to build a gene catalog directly from NGS reads" - 18-*ième Journées de l'ED394* - Paris (France) - April 2017

"Peptide identification in the gut microbiota: contribution of patient-specific gene catalogs" - *International Metaproteomics Symposium (IMS)* - Porto Conte (Italy) - June 2017

"A pipeline dedicated to Metaproteomics for large cohorts" (second author) - *Spectrométrie de Masse, Métabolomique et Analyse Protéomique (SMMAP)* - Marne-la-Vallée (France) - October 2017

"ProteoCardis: a transdisciplinary study to investigate the link between cardiometabolic diseases and gut microbiota proteome" - *Séminaire interne MetaGenoPolis* - Jouy-en-Josas (France) - May 2018

"Peptide identification in the gut microbiota: Contribution of patient-specific gene catalogs" - 19-*ième Journées de l'ED394* - Paris (France) - May 2018

"Interpretation of mass spectrometry-based metaproteomics: How much can we trust the MetaHIT 9.9 catalog?" *International Human Microbiome Consortium (IHMC)* - Cork (Ireland) - June 2018

"Interpretation of mass spectrometry-based metaproteomics: How much can we trust the MetaHIT 9.9 catalog?" *Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM)* - Marseille (France) - July 2018

"Metaproteomic of the human gut microbiota in physiological and pathological context" - 20-*ième Journées de l'ED394* - Paris (France) - May 2019

"Metaproteomics of the human intestinal microbiota in physiological and pathological conditions" (not the talker) - *American Society for Mass Spectrometry (ASMS) Conference* - Atlanta (USA) - June 2019

"ProteoCardis: an intestinal metaproteome-wide association study of coronary artery disease" - *Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM)* - Nantes (France) - July 2019

F.2 Oral communications

"ProteoCardis: a transdisciplinary study to investigate the link between cardiometabolic diseases and gut microbiota proteome" - *Journée des bioinformaticiens de Jouy* - Jouy-en-Josas (France) - November 2017

"Défis du big data pour l'identification des protéines en métaprotéomique" - *Colloque interne Génétique Quantitative et Évolution* - Gif-sur-Yvette (France) - February 2018

"Défis du big data pour l'identifiiation des protéines en métaprotéomique" - *Réunions StatInfOmics* - Jouy-en-Josas (France) - November 2018

"ProteoCardis: an intestinal metaproteome-wide association study of coronary artery disease" - *International Metaproteomics Symposium (IMS)* - Leipzig (Germany) - December 2018

Bibliography

[1] Valeria D'Argenio and Francesco Salvatore. "The role of the gut microbiome in the healthy adult status". en. In: *Clinica Chimica Acta* 451 (Dec. 2015), pp. 97–102. ISSN: 00098981. DOI: 10.1016/j.cca.2015.01.003. URL: http://linkinghub.elsevier.com/retrieve/pii/S0009898115000170 (visited on 08/04/2017).

[2] The Human Microbiome Project Consortium. "Structure, function and diversity of the healthy human microbiome". en. In: *Nature* 486.7402 (June 2012), pp. 207–214. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11234. URL: http://www.nature.com/articles/nature11234 (visited on 03/12/2019).

[3] Manimozhiyan Arumugam et al. "Enterotypes of the human gut microbiome". In: *Nature* 473.7346 (May 2011), pp. 174–180. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature09944. URL: http://www.nature.com/doifinder/10.1038/nature09944 (visited on 08/04/2017).

[4] Le Chatelier Emmanuelle et al. "Richness of human gut microbiome correlates with metabolic markers". en. In: *Nature* 500.7464 (Aug. 2013), pp. 541–546. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature12506. URL: http://www.nature.com/articles/nature12506 (visited on 03/13/2019).

[5] Junhua Li et al. "An integrated catalog of reference genes in the human gut microbiome". en. In: *Nature Biotechnology* 32.8 (Aug. 2014), pp. 834–841. ISSN: 1087-0156. DOI: 10.1038/nbt.2942. URL: http://www.nature.com/nbt/journal/v32/n8/full/nbt.2942.html.

[6] Sven-Bastiaan Haange and Nico Jehmlich. "Proteomic interrogation of the gut microbiota: potential clinical impact". en. In: *Expert Review of Proteomics* 13.6 (June 2016), pp. 535–537. ISSN: 1478-9450, 1744-8387. DOI: 10.1080/14789450.2016.1190652. URL: https://www.tandfonline.com/doi/full/10.1080/14789450.2016.1190652 (visited on 07/24/2019).

[7] Pey Yee Lee et al. "Metaproteomic analysis of human gut microbiota: where are we heading?" en. In: *Journal of Biomedical Science* 24.1 (Dec. 2017), p. 36. ISSN: 1423-0127. DOI: 10.1186/s12929-017-0342-z. URL: http://jbiomedsci.biomedcentral.com/articles/10.1186/s12929-017-0342-z (visited on 05/15/2019).

[8] Weili Xiong et al. "Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota". en. In: *PROTEOMICS* 15.20 (Oct. 2015), pp. 3424–3438. ISSN: 16159853. DOI: 10.1002/pmic.201400571. URL: http://doi.wiley.com/10.1002/pmic.201400571 (visited on 07/24/2019).

[9] Xu Zhang and Daniel Figeys. "Perspective and Guidelines for Metaproteomics in Microbiome Studies". en. In: *Journal of Proteome Research* 18.6 (June 2019), pp. 2370–2380. ISSN: 1535-3893, 1535-3907. DOI: 10.1021/acs.jproteome.9b00054. URL: http://pubs.acs.org/doi/10.1021/acs.jproteome.9b00054 (visited on 07/08/2019).

[10] Simon Deusch et al. "News in livestock research — use of Omics -technologies to study the microbiota in the gastrointestinal tract of farm animals". en. In: *Computational and Structural Biotechnology Journal* 13 (2015), pp. 55–63. ISSN: 20010370. DOI: 10.1016/j.csbj.2014.12.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S2001037014000555 (visited on 07/24/2019).

[11] Gerren P. Hobby et al. "Chronic kidney disease and the gut microbiome". en. In: *American Journal of Physiology-Renal Physiology* 316.6 (June 2019), F1211–F1217. ISSN: 1931-857X, 1522-1466. DOI: 10.1152/ajprenal.00298.2018. URL: https://www.physiology.org/doi/10.1152/ajprenal.00298.2018 (visited on 07/24/2019).

[12] R. I. Amann, W. Ludwig, and K. H. Schleifer. "Phylogenetic identification and in situ detection of individual microbial cells without cultivation". eng. In: *Microbiological Reviews* 59.1 (Mar. 1995), pp. 143–169. ISSN: 0146-0749.

[13] Philip Hugenholtz, Brett M Goebel, and Norman R Pace. "Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity". en. In: *J. BACTERIOL.* 180 (1998), p. 10.

[14] Emma Allen-Vercoe. "Bringing the gut microbiota into focus through microbial culture: recent progress and future perspective". en. In: *Current Opinion in Microbiology* 16.5 (Oct. 2013), pp. 625–629. ISSN: 13695274. DOI: 10.1016/j.mib.2013.09.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S1369527413001598 (visited on 03/15/2019).

[15] Josselin Bodilis et al. "Variable Copy Number, Intra-Genomic Heterogeneities and Lateral Transfers of the 16S rRNA Gene in Pseudomonas". en. In: *PLoS ONE* 7.4 (Apr. 2012). Ed. by Katy C. Kao, e35647. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0035647. URL: https://dx.plos.org/10.1371/journal.pone.0035647 (visited on 03/19/2019).

[16] Juan Jovel et al. "Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics". In: *Frontiers in Microbiology* 7 (Apr. 2016). ISSN: 1664-302X. DOI: 10.3389/fmicb.2016.00459. URL: http://journal.frontiersin.org/Article/10.3389/fmicb.2016.00459/abstract (visited on 03/18/2019).

[17] Ludmila Chistoserdova. "Functional Metagenomics: Recent Advances and Future Challenges". en. In: *Biotechnology and Genetic Engineering Reviews* 26.1 (Jan. 2009), pp. 335–352. ISSN: 0264-8725, 2046-5556. DOI: 10.5661/bger-26-335. URL: http://www.tandfonline.com/doi/abs/10.5661/bger-26-335 (visited on 09/15/2017).

[18] *Principe de la source ionisation électrospray*. Mar. 2016. URL: http://masse-spec.fr/electrospray (visited on 08/07/2017).

[19] *Thermo Fisher :: Orbitrap :: Orbitrap Fusion Lumos*. URL: http://planetorbitrap.com/orbitrap-fusion-lumos#tab:schematic (visited on 04/05/2019).

[20] David N. Perkins et al. "Probability-based protein identification by searching sequence databases using mass spectrometry data". en. In: *Electrophoresis* 20.18 (Dec. 1999), pp. 3551–3567. ISSN: 0173-0835, 1522-2683. DOI: 10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2. URL: http://doi.wiley.com/10.1002/%28SICI%291522-2683%2819991201%2920%3A18%3C3551%3A%3AAID-ELPS3551%3E3.0.CO%3B2-2 (visited on 04/05/2019).

[21] Lewis Y. Geer et al. "Open Mass Spectrometry Search Algorithm". en. In: *Journal of Proteome Research* 3.5 (Oct. 2004), pp. 958–964. ISSN: 1535-3893, 1535-3907. DOI: 10.1021/pr0499491. URL: http://pubs.acs.org/doi/abs/10.1021/pr0499491 (visited on 04/05/2019).

[22] R Craig and RC Beavis. "TANDEM: matching proteins with tandem mass spectra." In: *Bioinformatics* 20 (June 2004), pp. 1466–1467. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bth092. URL: http://bioinformatics.oxfordjournals.org/content/20/9/1466.

[23] Clemens Blank et al. "Disseminating Metaproteomic Informatics Capabilities and Knowledge Using the Galaxy-P Framework". en. In: *Proteomes* 6.1 (Jan. 2018), p. 7. ISSN: 2227-7382. DOI: 10.3390/proteomes6010007. URL: http://www.mdpi.com/2227-7382/6/1/7 (visited on 04/05/2019).

[24] Markus Brosch et al. "Comparison of Mascot and X!Tandem Performance for Low and High Accuracy Mass Spectrometry and the Development of an Adjusted Mascot Threshold". en. In: *Molecular & Cellular Proteomics* 7.5 (May 2008), pp. 962–970. ISSN: 1535-9476, 1535-9484. DOI: 10.1074/mcp.M700293-MCP200. URL: http://www.mcponline.org/lookup/doi/10.1074/mcp.M700293-MCP200 (visited on 04/01/2019).

[25] D. Zosso et al. "Tandem mass spectrometry protein identification on a PC grid". eng. In: *Studies in Health Technology and Informatics* 126 (2007), pp. 3–12. ISSN: 0926-9630.

[26] Malik N. Akhtar et al. "Evaluation of Database Search Programs for Accurate Detection of Neuropeptides in Tandem Mass Spectrometry Experiments". en.

In: *Journal of Proteome Research* 11.12 (Dec. 2012), pp. 6044–6055. ISSN: 1535-3893, 1535-3907. DOI: 10.1021/pr3007123. URL: http://pubs.acs.org/doi/10.1021/pr3007123 (visited on 04/05/2019).

[27] Suruchi Aggarwal and Amit Kumar Yadav. "False Discovery Rate Estimation in Proteomics". In: *Statistical Analysis in Proteomics*. Ed. by Klaus Jung. Vol. 1362. New York, NY: Springer New York, 2016, pp. 119–128. ISBN: 978-1-4939-3105-7 978-1-4939-3106-4. DOI: 10.1007/978-1-4939-3106-4_7. URL: http://link.springer.com/10.1007/978-1-4939-3106-4_7 (visited on 04/08/2019).

[28] Lukas Käll et al. "Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases". en. In: *Journal of Proteome Research* 7.1 (Jan. 2008), pp. 29–34. ISSN: 1535-3893, 1535-3907. DOI: 10.1021/pr700600n. URL: http://pubs.acs.org/doi/abs/10.1021/pr700600n (visited on 04/08/2019).

[29] Olivier Langella et al. "X!TandemPipeline: A Tool to Manage Sequence Redundancy for Protein Inference and Phosphosite Identification". en. In: *Journal of Proteome Research* 16.2 (Feb. 2017), pp. 494–503. ISSN: 1535-3893, 1535-3907. DOI: 10.1021/acs.jproteome.6b00632. URL: http://pubs.acs.org/doi/abs/10.1021/acs.jproteome.6b00632 (visited on 11/30/2017).

[30] Catherine Juste et al. "Bacterial protein signals are associated with Crohn's disease". en. In: *Gut* 63.10 (Oct. 2014), pp. 1566–1577. ISSN: 0017-5749, 1468-3288. DOI: 10.1136/gutjnl-2012-303786. URL: http://gut.bmj.com/lookup/doi/10.1136/gutjnl-2012-303786 (visited on 12/19/2017).

[31] Mélisande Blein-Nicolas et al. "Including shared peptides for estimating protein abundances: A significant improvement for quantitative proteomics". en. In: *PROTEOMICS* 12.18 (Sept. 2012), pp. 2797–2801. ISSN: 16159853. DOI: 10.1002/pmic.201100660. URL: http://doi.wiley.com/10.1002/pmic.201100660 (visited on 04/26/2019).

[32] Mélisande Blein-Nicolas and Michel Zivy. "Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics". en. In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1864.8 (Aug. 2016), pp. 883–895. ISSN: 15709639. DOI: 10.1016/j.bbapap.2016.02.019. URL: https://linkinghub.elsevier.com/retrieve/pii/S1570963916300322 (visited on 07/12/2019).

[33] Paul Wilmes and Philip L. Bond. "The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms". en. In: *Environmental Microbiology* 6.9 (Sept. 2004), pp. 911–920. ISSN: 1462-2912, 1462-2920. DOI: 10.1111/j.1462-2920.2004.00687.x. URL: http://doi.wiley.com/10.1111/j.1462-2920.2004.00687.x (visited on 07/24/2019).

[34] RJ Ram. "Community proteomics of a natural microbial biofilm." In: *Science* 308 (June 2005), pp. 1915–20. ISSN: 00368075, 10959203. DOI: 10.1126/science.1109070.

[35] E. S. Klaassens, W. M. de Vos, and E. E. Vaughan. "Metaproteomics Approach To Study the Functionality of the Microbiota in the Human Infant Gastrointestinal Tract". en. In: *Applied and Environmental Microbiology* 73.4 (Feb. 2007), pp. 1388–1392. ISSN: 0099-2240. DOI: 10.1128/AEM.01921-06. URL: http://aem.asm.org/cgi/doi/10.1128/AEM.01921-06 (visited on 07/24/2019).

[36] Carolin A. Kolmeder et al. "Comparative Metaproteomics and Diversity Analysis of Human Intestinal Microbiota Testifies for Its Temporal Stability and Expression of Core Functions". en. In: *PLoS ONE* 7.1 (Jan. 2012). Ed. by Adam Driks, e29913. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0029913. URL: https://dx.plos.org/10.1371/journal.pone.0029913 (visited on 07/24/2019).

[37] Koos Rooijers et al. "An iterative workflow for mining the human intestinal metaproteome". en. In: *BMC Genomics* 12.1 (Dec. 2011), p. 6. ISSN: 1471-2164. DOI: 10.1186/1471-2164-12-6. URL: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-12-6 (visited on 07/24/2019).

[38] Ester Hernández et al. "Functional consequences of microbial shifts in the human gastrointestinal tract linked to antibiotic treatment and obesity". en. In: *Gut Microbes* 4.4 (July 2013), pp. 306–315. ISSN: 1949-0976, 1949-0984. DOI: 10.4161/gmic.25321. URL: http://www.tandfonline.com/doi/abs/10.4161/gmic.25321 (visited on 07/24/2019).

[39] Alessandro Tanca et al. "Potential and active functions in the gut microbiota of a healthy human cohort". en. In: *Microbiome* 5.1 (Dec. 2017), p. 79. ISSN: 2049-2618. DOI: 10.1186/s40168-017-0293-3. URL: http://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-017-0293-3 (visited on 04/10/2019).

[40] Alessandro Tanca et al. "The impact of sequence database choice on metaproteomic results in gut microbiota studies". en. In: *Microbiome* 4.1 (Dec. 2016), p. 51. ISSN: 2049-2618. DOI: 10.1186/s40168-016-0196-8. URL: http://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-016-0196-8 (visited on 06/03/2019).

[41] Xu Zhang et al. "MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota". en. In: *Microbiome* 4.1 (Dec. 2016). ISSN: 2049-2618. DOI: 10.1186/s40168-016-0176-z. URL: http://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-016-0176-z (visited on 11/25/2016).

[42] Alessandro Tanca et al. "Enrichment or depletion? The impact of stool pretreatment on metaproteomic characterization of the human gut microbiota". en. In: *PROTEOMICS* 15.20 (Oct. 2015), pp. 3474–3485. ISSN: 16159853. DOI: 10.1002/pmic.201400573. URL: http://doi.wiley.com/10.1002/pmic.201400573 (visited on 07/24/2019).

[43] Christopher S. Reigstad and Purna C. Kashyap. "Beyond phylotyping: understanding the impact of gut microbiota on host biology". en. In: *Neurogastroenterology & Motility* 25.5 (May 2013), pp. 358–372. ISSN: 13501925. DOI: 10.1111/nmo.12134. URL: http://doi.wiley.com/10.1111/nmo.12134 (visited on 07/24/2019).

[44] Anna Heintz-Buschart et al. "Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes". en. In: *Nature Microbiology* 2.1 (Jan. 2017), p. 16180. ISSN: 2058-5276. DOI: 10.1038/nmicrobiol.2016.180. URL: http://www.nature.com/articles/nmicrobiol2016180 (visited on 07/24/2019).

[45] Romy D. Zwittink et al. "Metaproteomics reveals functional differences in intestinal microbiota development of preterm infants". en. In: *Molecular & Cellular Proteomics* 16.9 (Sept. 2017), pp. 1610–1620. ISSN: 1535-9476, 1535-9484. DOI: 10.1074/mcp.RA117.000102. URL: http://www.mcponline.org/lookup/doi/10.1074/mcp.RA117.000102 (visited on 07/24/2019).

[46] Brandon Brooks et al. "Strain-resolved microbial community proteomics reveals simultaneous aerobic and anaerobic function during gastrointestinal tract colonization of a preterm infant". In: *Frontiers in Microbiology* 6 (July 2015). ISSN: 1664-302X. DOI: 10.3389/fmicb.2015.00654. URL: http://journal.frontiersin.org/article/10.3389/fmicb.2015.00654 (visited on 07/24/2019).

[47] Maria Guirro et al. "Multi-omics approach to elucidate the gut microbiota activity: Metaproteomics and metagenomics connection". en. In: *ELECTROPHORESIS* 39.13 (July 2018), pp. 1692–1701. ISSN: 01730835. DOI: 10.1002/elps.201700476. URL: http://doi.wiley.com/10.1002/elps.201700476 (visited on 07/24/2019).

[48] Weili Xiong et al. "Genome-resolved metaproteomic characterization of preterm infant gut microbiota development reveals species-specific metabolic shifts and variabilities during early life". en. In: *Microbiome* 5.1 (Dec. 2017), p. 72. ISSN: 2049-2618. DOI: 10.1186/s40168-017-0290-6. URL: http://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-017-0290-6 (visited on 07/24/2019).

[49] Jacque C. Young et al. "Metaproteomics reveals functional shifts in microbial and human proteins during a preterm infant gut colonization case". en. In: *PROTEOMICS* 15.20 (Oct. 2015), pp. 3463–3473. ISSN: 16159853. DOI: 10.

1002/pmic.201400563. URL: http://doi.wiley.com/10.1002/pmic.201400563 (visited on 07/24/2019).

[50] Sven-Bastiaan Haange et al. "Metaproteome Analysis and Molecular Genetics of Rat Intestinal Microbiota Reveals Section and Localization Resolved Species Distribution and Enzymatic Functionalities". en. In: *Journal of Proteome Research* 11.11 (Nov. 2012), pp. 5406–5417. ISSN: 1535-3893, 1535-3907. DOI: 10.1021/pr3006364. URL: http://pubs.acs.org/doi/10.1021/pr3006364 (visited on 07/24/2019).

[51] Stefano Levi Mortera et al. "Metaproteomic investigation to assess gut microbiota shaping in newborn mice: A combined taxonomic, functional and quantitative approach". en. In: *Journal of Proteomics* 203 (July 2019), p. 103378. ISSN: 18743919. DOI: 10.1016/j.jprot.2019.103378. URL: https://linkinghub.elsevier.com/retrieve/pii/S1874391919301502 (visited on 07/24/2019).

[52] Xinxin Ke et al. "Synbiotic-driven improvement of metabolic disturbances is associated with changes in the gut microbiome in diet-induced obese mice". en. In: *Molecular Metabolism* 22 (Apr. 2019), pp. 96–109. ISSN: 22128778. DOI: 10.1016/j.molmet.2019.01.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S2212877818312699 (visited on 07/24/2019).

[53] Monika Schaubeck et al. "Dysbiotic gut microbiota causes transmissible Crohn's disease-like ileitis independent of failure in antimicrobial defence". en. In: *Gut* 65.2 (Feb. 2016), pp. 225–237. ISSN: 0017-5749, 1468-3288. DOI: 10.1136/gutjnl-2015-309333. URL: http://gut.bmj.com/lookup/doi/10.1136/gutjnl-2015-309333 (visited on 07/24/2019).

[54] Alessandro Tanca et al. "Caloric restriction promotes functional changes involving short-chain fatty acid biosynthesis in the rat gut microbiota". en. In: *Scientific Reports* 8.1 (Dec. 2018), p. 14778. ISSN: 2045-2322. DOI: 10.1038/s41598-018-33100-y. URL: http://www.nature.com/articles/s41598-018-33100-y (visited on 07/24/2019).

[55] Johanna Tröscher-Mußotter et al. "Analysis of the Bacterial and Host Proteins along and across the Porcine Gastrointestinal Tract". en. In: *Proteomes* 7.1 (Jan. 2019), p. 4. ISSN: 2227-7382. DOI: 10.3390/proteomes7010004. URL: http://www.mdpi.com/2227-7382/7/1/4 (visited on 07/24/2019).

[56] Daniel Borda-Molina, Jana Seifert, and Amélia Camarinha-Silva. "Current Perspectives of the Chicken Gastrointestinal Tract and Its Microbiome". en. In: *Computational and Structural Biotechnology Journal* 16 (2018), pp. 131–139. ISSN: 20010370. DOI: 10.1016/j.csbj.2018.03.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S2001037017301162 (visited on 07/24/2019).

[57] Antonio Palomba et al. "Multi-Omic Biogeography of the Gastrointestinal Microbiota of a Pre-Weaned Lamb". en. In: *Proteomes* 5.4 (Dec. 2017), p. 36.

ISSN: 2227-7382. DOI: 10.3390/proteomes5040036. URL: http://www.mdpi.com/2227-7382/5/4/36 (visited on 03/28/2019).

[58] Bruno Tilocca et al. "Variations of Phosphorous Accessibility Causing Changes in Microbiome Functions in the Gastrointestinal Tract of Chickens". en. In: *PLOS ONE* 11.10 (Oct. 2016). Ed. by Gunnar Loh, e0164735. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0164735. URL: http://dx.plos.org/10.1371/journal.pone.0164735 (visited on 07/24/2019).

[59] Yue Tang et al. "Metaproteomics Analysis Reveals the Adaptation Process for the Chicken Gut Microbiota". en. In: *Applied and Environmental Microbiology* 80.2 (Jan. 2014), pp. 478–485. ISSN: 0099-2240, 1098-5336. DOI: 10.1128/AEM.02472-13. URL: http://aem.asm.org/lookup/doi/10.1128/AEM.02472-13 (visited on 07/24/2019).

[60] Robert H. Mills et al. "Evaluating Metagenomic Prediction of the Metaproteome in a 4.5-Year Study of a Patient with Crohn's Disease". en. In: *mSystems* 4.1 (Feb. 2019). Ed. by Marcus J. Claesson. ISSN: 2379-5077. DOI: 10.1128/mSystems.00337-18. URL: http://msystems.asm.org/lookup/doi/10.1128/mSystems.00337-18 (visited on 03/28/2019).

[61] Xu Zhang et al. "Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease". en. In: *Nature Communications* 9.1 (Dec. 2018), p. 2873. ISSN: 2041-1723. DOI: 10.1038/s41467-018-05357-4. URL: http://www.nature.com/articles/s41467-018-05357-4 (visited on 06/03/2019).

[62] Aleksandar D. Kostic, Ramnik J. Xavier, and Dirk Gevers. "The Microbiome in Inflammatory Bowel Disease: Current Status and the Future Ahead". en. In: *Gastroenterology* 146.6 (May 2014), pp. 1489–1499. ISSN: 00165085. DOI: 10.1053/j.gastro.2014.02.009. URL: https://linkinghub.elsevier.com/retrieve/pii/S0016508514002200 (visited on 07/24/2019).

[63] Arnau Vich Vila et al. "Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome". en. In: *Science Translational Medicine* 10.472 (Dec. 2018), eaap8914. ISSN: 1946-6234, 1946-6242. DOI: 10.1126/scitranslmed.aap8914. URL: http://stm.sciencemag.org/lookup/doi/10.1126/scitranslmed.aap8914 (visited on 07/24/2019).

[64] Xu Zhang et al. "Assessing the impact of protein extraction methods for human gut metaproteomics". en. In: *Journal of Proteomics* 180 (May 2018), pp. 120–127. ISSN: 18743919. DOI: 10.1016/j.jprot.2017.07.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S187439191730235X (visited on 07/24/2019).

[65] Thilo Muth et al. "Navigating through metaproteomics data: A logbook of database searching". en. In: *PROTEOMICS* 15.20 (Oct. 2015), pp. 3439–3453.

ISSN: 16159853. DOI: 10.1002/pmic.201400560. URL: http://doi.wiley.com/10.1002/pmic.201400560 (visited on 04/12/2018).

[66] Shao-En Ong and Matthias Mann. "Mass spectrometry–based proteomics turns quantitative". en. In: *Nature Chemical Biology* 1.5 (Oct. 2005), pp. 252–262. ISSN: 1552-4450, 1552-4469. DOI: 10.1038/nchembio736. URL: http://www.nature.com/articles/nchembio736 (visited on 06/12/2019).

[67] Cheng Chang et al. *LFAQ: towards unbiased label-free absolute protein quantification by predicting peptide quantitative factors*. en. preprint. Bioinformatics, May 2018. DOI: 10.1101/328864. URL: http://biorxiv.org/lookup/doi/10.1101/328864 (visited on 06/12/2019).

[68] Bing He et al. "Label-free absolute protein quantification with data-independent acquisition". en. In: *Journal of Proteomics* 200 (May 2019), pp. 51–59. ISSN: 18743919. DOI: 10.1016/j.jprot.2019.03.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S1874391919300867 (visited on 06/12/2019).

[69] Nathan C Verberkmoes et al. "Shotgun metaproteomics of the human distal gut microbiota". en. In: *The ISME Journal* 3.2 (Feb. 2009), pp. 179–189. ISSN: 1751-7362, 1751-7370. DOI: 10.1038/ismej.2008.108. URL: http://www.nature.com/articles/ismej2008108 (visited on 07/12/2019).

[70] Emelia J. Benjamin et al. "Heart Disease and Stroke Statistics—2017 Update: A Report From the American Heart Association". en. In: *Circulation* 135.10 (Mar. 2017). ISSN: 0009-7322, 1524-4539. DOI: 10.1161/CIR.0000000000000485. URL: https://www.ahajournals.org/doi/10.1161/CIR.0000000000000485 (visited on 01/21/2019).

[71] K George MM Alberti, Paul Zimmet, and Jonathan Shaw. "The metabolic syndrome—a new worldwide definition". en. In: *The Lancet* 366.9491 (Sept. 2005), pp. 1059–1062. ISSN: 01406736. DOI: 10.1016/S0140-6736(05)67402-8. URL: https://linkinghub.elsevier.com/retrieve/pii/S0140673605674028 (visited on 03/01/2019).

[72] Martine Glorian and Isabelle Limon. "L'athérosclérose, une maladie inflammatoire". fr. In: *Revue Francophone des Laboratoires* 2007.389 (Feb. 2007), pp. 43–48. ISSN: 1773035X. DOI: 10.1016/S1773-035X(07)80061-X. URL: https://linkinghub.elsevier.com/retrieve/pii/S1773035X0780061X (visited on 02/11/2019).

[73] J.-C. Fruchart. "New Risk Factors for Atherosclerosis and Patient Risk Assessment". en. In: *Circulation* 109.23_suppl_1 (June 2004), pp. III–15–III–19. ISSN: 0009-7322, 1524-4539. DOI: 10.1161/01.CIR.0000131513.33892.5b. URL: http://circ.ahajournals.org/cgi/doi/10.1161/01.CIR.0000131513.33892.5b (visited on 02/28/2019).

[74] Andrew P. Goldberg. "Aerobic and resistive exercise modify risk factors for coronary heart disease:" en. In: *Medicine & Science in Sports & Exercise* 21.6 (Dec. 1989), p. 669. ISSN: 0195-9131. DOI: 10.1249/00005768-198912000-00008. URL: https://insights.ovid.com/crossref?an=00005768-198912000-00008 (visited on 03/01/2019).

[75] Nils P. Larsen. "Diet and Atherosclerosis: A Field Study". en. In: *A.M.A. Archives of Internal Medicine* 100.3 (Sept. 1957), p. 436. ISSN: 0888-2479. DOI: 10.1001/archinte.1957.00260090092012. URL: http://archinte.jamanetwork.com/article.aspx?doi=10.1001/archinte.1957.00260090092012 (visited on 03/04/2019).

[76] A. Keys. "Human Atherosclerosis and the Diet". en. In: *Circulation* 5.1 (Jan. 1952), pp. 115–118. ISSN: 0009-7322, 1524-4539. DOI: 10.1161/01.CIR.5.1.115. URL: http://circ.ahajournals.org/cgi/doi/10.1161/01.CIR.5.1.115 (visited on 03/04/2019).

[77] Marco Matteo Ciccone et al. "Dietary Intake of Carotenoids and Their Antioxidant and Anti-Inflammatory Effects in Cardiovascular Care". en. In: *Mediators of Inflammation* 2013 (2013), pp. 1–11. ISSN: 0962-9351, 1466-1861. DOI: 10.1155/2013/782137. URL: http://www.hindawi.com/journals/mi/2013/782137/ (visited on 03/04/2019).

[78] Volha Summerhill et al. "Vasculoprotective Role of Olive Oil Compounds via Modulation of Oxidative Stress in Atherosclerosis". In: *Frontiers in Cardiovascular Medicine* 5 (Dec. 2018). ISSN: 2297-055X. DOI: 10.3389/fcvm.2018.00188. URL: https://www.frontiersin.org/article/10.3389/fcvm.2018.00188/full (visited on 03/04/2019).

[79] Carlos Mercado and Edgar A. Jaimes. "Cigarette smoking as a risk factor for atherosclerosis and renal disease: Novel pathogenic insights". en. In: *Current Hypertension Reports* 9.1 (Mar. 2007), pp. 66–72. ISSN: 1522-6417, 1534-3111. DOI: 10.1007/s11906-007-0012-8. URL: http://link.springer.com/10.1007/s11906-007-0012-8 (visited on 03/04/2019).

[80] T.H.S. Dent. "Predicting the risk of coronary heart disease". en. In: *Atherosclerosis* 213.2 (Dec. 2010), pp. 345–351. ISSN: 00219150. DOI: 10.1016/j.atherosclerosis.2010.06.019. URL: https://linkinghub.elsevier.com/retrieve/pii/S0021915010004375 (visited on 07/22/2019).

[81] Herbert Tilg and Arthur Kaser. "Gut microbiome, obesity, and metabolic dysfunction". en. In: *Journal of Clinical Investigation* 121.6 (June 2011), pp. 2126–2132. ISSN: 0021-9738. DOI: 10.1172/JCI58109. URL: http://www.jci.org/articles/view/58109 (visited on 11/25/2016).

[82] Nadja Larsen et al. "Gut Microbiota in Human Adults with Type 2 Diabetes Differs from Non-Diabetic Adults". en. In: *PLoS ONE* 5.2 (Feb. 2010). Ed. by Stefan Bereswill, e9085. ISSN: 1932-6203. DOI: 10.1371/journal.pone.

0009085. URL: http://dx.plos.org/10.1371/journal.pone.0009085 (visited on 08/04/2017).

[83] R Balfour Sartor and Sarkis K Mazmanian. "Intestinal Microbes in Inflammatory Bowel Diseases". In: *The American Journal of Gastroenterology Supplements* 1.1 (July 2012), pp. 15–21. ISSN: 1948-9498, 1948-9501. DOI: 10.1038/ajgsup.2012.4. URL: http://www.nature.com/doifinder/10.1038/ajgsup.2012.4 (visited on 11/25/2016).

[84] Ruth E. Ley et al. "Human gut microbes associated with obesity". en. In: *Nature* 444.7122 (Dec. 2006), pp. 1022–1023. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/4441022a. URL: http://www.nature.com/articles/4441022a (visited on 07/22/2019).

[85] Junjie Qin et al. "A metagenome-wide association study of gut microbiota in type 2 diabetes". In: *Nature* 490.7418 (Sept. 2012), pp. 55–60. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11450. URL: http://www.nature.com/doifinder/10.1038/nature11450 (visited on 02/04/2019).

[86] Aurélie Cotillard et al. "Dietary intervention impact on gut microbial gene richness". In: *Nature* 500.7464 (Aug. 2013), pp. 585–588. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature12480. URL: http://www.nature.com/doifinder/10.1038/nature12480 (visited on 12/19/2017).

[87] Ling Chun Kong et al. "Insulin resistance and inflammation predict kinetic body weight changes in response to dietary weight loss and maintenance in overweight and obese subjects by using a Bayesian network approach". en. In: *The American Journal of Clinical Nutrition* 98.6 (Dec. 2013), pp. 1385–1394. ISSN: 0002-9165, 1938-3207. DOI: 10.3945/ajcn.113.058099. URL: https://academic.oup.com/ajcn/article/98/6/1385/4577252 (visited on 07/06/2019).

[88] Philippe Gerard et al. "Gnotobiotic rats harboring human intestinal microbiota as a model for studying cholesterol-to-coprostanol conversion". en. In: *FEMS Microbiology Ecology* 47.3 (Mar. 2004), pp. 337–343. ISSN: 01686496, 15746941. DOI: 10.1016/S0168-6496(03)00285-X. URL: https://academic.oup.com/femsec/article-lookup/doi/10.1016/S0168-6496(03)00285-X (visited on 07/22/2019).

[89] P. Gerard et al. "Bacteroides sp. Strain D8, the First Cholesterol-Reducing Bacterium Isolated from Human Feces". en. In: *Applied and Environmental Microbiology* 73.18 (Sept. 2007), pp. 5742–5749. ISSN: 0099-2240. DOI: 10.1128/AEM.02806-06. URL: http://aem.asm.org/cgi/doi/10.1128/AEM.02806-06 (visited on 03/11/2019).

[90] Fredrik Bäckhed et al. "Mechanisms underlying the resistance to diet-induced obesity in germ-free mice". en. In: *Proceedings of the National Academy of Sciences* 104.3 (Jan. 2007), pp. 979–984. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/

pnas.0605374104. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.0605374104 (visited on 03/11/2019).

[91] Federation of American Societies for Experimental Biology (FASEB). "Gut organisms could be clue in controlling obesity risk." In: *ScienceDaily* (Apr. 2012). URL: www.sciencedaily.com/releases/2012/04/120423162223.htm.

[92] N. Alang and C. R. Kelly. "Weight Gain After Fecal Microbiota Transplantation". en. In: *Open Forum Infectious Diseases* 2.1 (Feb. 2015), ofv004–ofv004. ISSN: 2328-8957. DOI: 10.1093/ofid/ofv004. URL: https://academic.oup.com/ofid/article-lookup/doi/10.1093/ofid/ofv004 (visited on 07/22/2019).

[93] F. Backhed et al. "The gut microbiota as an environmental factor that regulates fat storage". en. In: *Proceedings of the National Academy of Sciences* 101.44 (Nov. 2004), pp. 15718–15723. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0407076101. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.0407076101 (visited on 07/22/2019).

[94] Peter J. Turnbaugh et al. "An obesity-associated gut microbiome with increased capacity for energy harvest". en. In: *Nature* 444.7122 (Dec. 2006), pp. 1027–1031. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature05414. URL: http://www.nature.com/articles/nature05414 (visited on 03/11/2019).

[95] Zeneng Wang et al. "Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease". In: *Nature* 472.7341 (Apr. 2011), pp. 57–63. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature09922. URL: http://www.nature.com/doifinder/10.1038/nature09922 (visited on 08/04/2017).

[96] Fredrik H. Karlsson et al. "Symptomatic atherosclerosis is associated with an altered gut metagenome". en. In: *Nature Communications* 3.1 (Jan. 2012). ISSN: 2041-1723. DOI: 10.1038/ncomms2266. URL: http://www.nature.com/articles/ncomms2266 (visited on 03/11/2019).

[97] Julia Rechenberger et al. "Challenges in Clinical Metaproteomics Highlighted by the Analysis of Acute Leukemia Patients with Gut Colonization by Multidrug-Resistant Enterobacteriaceae". en. In: *Proteomes* 7.1 (Jan. 2019), p. 2. ISSN: 2227-7382. DOI: 10.3390/proteomes7010002. URL: https://www.mdpi.com/2227-7382/7/1/2 (visited on 07/20/2019).

[98] Junjie Qin et al. "A human gut microbial gene catalogue established by metagenomic sequencing". In: *Nature* 464.7285 (Mar. 2010), pp. 59–65. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature08821. URL: http://www.nature.com/doifinder/10.1038/nature08821 (visited on 09/28/2017).

[99] Pratik Jagtap et al. "Deep metaproteomic analysis of human salivary supernatant". en. In: *PROTEOMICS* 12.7 (Apr. 2012), pp. 992–1001. ISSN: 16159853. DOI: 10.1002/pmic.201100503. URL: http://doi.wiley.com/10.1002/pmic.201100503 (visited on 03/06/2018).

[100] Pratik Jagtap et al. "A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies". en. In: *PROTEOMICS* 13.8 (Apr. 2013), pp. 1352–1357. ISSN: 16159853. DOI: 10.1002/pmic.201200352. URL: http://doi.wiley.com/10.1002/pmic.201200352 (visited on 03/06/2018).

[101] Kai Cheng et al. "MetaLab: an automated pipeline for metaproteomic data analysis". In: *Microbiome* 5 (Dec. 2017). ISSN: 2049-2618. DOI: 10.1186/s40168-017-0375-2. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5712144/ (visited on 03/06/2018).

[102] Doruk Beyter et al. "ProteoStorm: An Ultrafast Metaproteomics Database Search Framework". en. In: *Cell Systems* 7.4 (Oct. 2018), 463–467.e6. ISSN: 24054712. DOI: 10.1016/j.cels.2018.08.009. URL: https://linkinghub.elsevier.com/retrieve/pii/S2405471218303569 (visited on 03/28/2019).

[103] T. C. Hsieh, K. H. Ma, and Anne Chao. "iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers)". en. In: *Methods in Ecology and Evolution* 7.12 (Dec. 2016). Ed. by Greg McInerny, pp. 1451–1456. ISSN: 2041210X. DOI: 10.1111/2041-210X.12613. URL: http://doi.wiley.com/10.1111/2041-210X.12613 (visited on 02/25/2019).

[104] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2013. URL: http://www.R-project.org/.

[105] David L. Tabb et al. "Repeatability and Reproducibility in Proteomic Identifications by Liquid Chromatography-Tandem Mass Spectrometry". en. In: *Journal of Proteome Research* 9.2 (Feb. 2010), pp. 761–776. ISSN: 1535-3893, 1535-3907. DOI: 10.1021/pr9006365. URL: http://pubs.acs.org/doi/abs/10.1021/pr9006365 (visited on 01/31/2019).

[106] Nathalie M. Delzenne and Patrice D. Cani. "Interaction Between Obesity and the Gut Microbiota: Relevance in Nutrition". en. In: *Annual Review of Nutrition* 31.1 (Aug. 2011), pp. 15–31. ISSN: 0199-9885, 1545-4312. DOI: 10.1146/annurev-nutr-072610-145146. URL: http://www.annualreviews.org/doi/10.1146/annurev-nutr-072610-145146 (visited on 07/06/2019).

[107] Brian D Ondov, Nicholas H Bergman, and Adam M Phillippy. "Interactive metagenomic visualization in a Web browser". en. In: *BMC Bioinformatics* 12.1 (Dec. 2011), p. 385. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-385. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-385 (visited on 07/22/2019).

[108] M. Fava. "Weight gain and antidepressants". eng. In: *The Journal of Clinical Psychiatry* 61 Suppl 11 (2000), pp. 37–41. ISSN: 0160-6689.

[109] Lieke M Spekhorst. "Performance of the Montreal classification for inflammatory bowel diseases". en. In: *World Journal of Gastroenterology* 20.41 (2014), p. 15374. ISSN: 1007-9327. DOI: 10.3748/wjg.v20.i41.15374. URL: http://www.wjgnet.com/1007-9327/full/v20/i41/15374.htm (visited on 06/17/2019).

[110] Bruno Rafael Ramos de Mattos et al. "Inflammatory Bowel Disease: An Overview of Immune Mechanisms and Biological Treatments". en. In: *Mediators of Inflammation* 2015 (2015), pp. 1–11. ISSN: 0962-9351, 1466-1861. DOI: 10.1155/2015/493012. URL: http://www.hindawi.com/journals/mi/2015/493012/ (visited on 06/17/2019).

[111] Daniel C. Baumgart. "The Diagnosis and Treatment of Crohn's Disease and Ulcerative Colitis". In: *Deutsches Aerzteblatt Online* (Feb. 2009). ISSN: 1866-0452. DOI: 10.3238/arztebl.2009.0123. URL: https://www.aerzteblatt.de/10.3238/arztebl.2009.0123 (visited on 06/17/2019).

[112] Clara Abraham and Judy H. Cho. "Inflammatory Bowel Disease". en. In: *New England Journal of Medicine* 361.21 (Nov. 2009), pp. 2066–2078. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMra0804647. URL: http://www.nejm.org/doi/abs/10.1056/NEJMra0804647 (visited on 06/17/2019).

[113] Qing He et al. "Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients". en. In: *GigaScience* 6.7 (July 2017). ISSN: 2047-217X. DOI: 10.1093/gigascience/gix050. URL: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/gix050/3888814 (visited on 06/20/2019).

[114] Eric A. Franzosa et al. "Gut microbiome structure and metabolic activity in inflammatory bowel disease". en. In: *Nature Microbiology* 4.2 (Feb. 2019), pp. 293–305. ISSN: 2058-5276. DOI: 10.1038/s41564-018-0306-4. URL: http://www.nature.com/articles/s41564-018-0306-4 (visited on 06/20/2019).

[115] Alexis Criscuolo and Sylvain Brisse. "AlienTrimmer: A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads". en. In: *Genomics* 102.5-6 (Nov. 2013), pp. 500–506. ISSN: 08887543. DOI: 10.1016/j.ygeno.2013.07.011. URL: https://linkinghub.elsevier.com/retrieve/pii/S0888754313001481 (visited on 03/26/2019).

[116] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2". en. In: *Nature Methods* 9.4 (Apr. 2012), pp. 357–359. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.1923. URL: http://www.nature.com/articles/nmeth.1923 (visited on 03/26/2019).

[117] Michael R. Crusoe et al. "The khmer software package: enabling efficient nucleotide sequence analysis". en. In: *F1000Research* 4 (Sept. 2015), p. 900. ISSN: 2046-1402. DOI: 10.12688/f1000research.6924.1. URL: https://f1000research.com/articles/4-900/v1 (visited on 07/24/2019).

[118] Anton Bankevich et al. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing". en. In: *Journal of Computational Biology* 19.5 (May 2012), pp. 455–477. ISSN: 1066-5277, 1557-8666. DOI: 10.1089/cmb.2012.0021. URL: http://online.liebertpub.com/doi/abs/10.1089/cmb.2012.0021 (visited on 01/19/2017).

[119] Sergey Nurk et al. "metaSPAdes: a new versatile metagenomic assembler". en. In: *Genome Research* 27.5 (May 2017), pp. 824–834. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.213959.116. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.213959.116 (visited on 07/23/2019).

[120] Esmaeil Forouzan et al. "Practical evaluation of 11 de novo assemblers in metagenome assembly". en. In: *Journal of Microbiological Methods* 151 (Aug. 2018), pp. 99–105. ISSN: 01677012. DOI: 10.1016/j.mimet.2018.06.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167701218301210 (visited on 07/23/2019).

[121] Doug Hyatt et al. "Prodigal: prokaryotic gene recognition and translation initiation site identification". en. In: *BMC Bioinformatics* 11.1 (Dec. 2010), p. 119. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-119. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-119 (visited on 06/07/2019).

[122] William L Trimble et al. "Short-read reading-frame predictors are not created equal: sequence error causes loss of signal". en. In: *BMC Bioinformatics* 13.1 (Dec. 2012), p. 183. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-183. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-183 (visited on 06/07/2019).

[123] W. Li and A. Godzik. "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences". en. In: *Bioinformatics* 22.13 (July 2006), pp. 1658–1659. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btl158. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl158 (visited on 06/11/2019).

[124] Chengping Wen et al. "Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis". en. In: *Genome Biology* 18.1 (Dec. 2017), p. 142. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1271-6. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1271-6 (visited on 06/13/2019).

[125] Boris L. Zybailov et al. "Metaproteomics reveals potential mechanisms by which dietary resistant starch supplementation attenuates chronic kidney disease progression in rats". en. In: *PLOS ONE* 14.1 (Jan. 2019). Ed. by Daotai Nie, e0199274. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0199274. URL: http://dx.plos.org/10.1371/journal.pone.0199274 (visited on 06/03/2019).

[126] T. Chen et al. "The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information". en. In: *Database* 2010.0 (July 2010), baq013–baq013. ISSN: 1758-0463. DOI: 10.1093/database/baq013. URL: https://academic.oup.com/database/article-lookup/doi/10.1093/database/baq013 (visited on 07/08/2019).

[127] T. Yamada et al. "iPath2.0: interactive pathway explorer". In: *Nucleic Acids Res* 39 (2011). DOI: 10.1093/nar/gkr313. URL: https://doi.org/10.1093/nar/gkr313.

[128] J. Alfredo Blakeley-Ruiz et al. "Metaproteomics reveals persistent and phylum-redundant metabolic functional stability in adult human gut microbiomes of Crohn's remission patients despite temporal variations in microbial taxa, genomes, and proteomes". en. In: *Microbiome* 7.1 (Dec. 2019), p. 18. ISSN: 2049-2618. DOI: 10.1186/s40168-019-0631-8. URL: https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0631-8 (visited on 07/22/2019).

[129] Chaysavanh Manichanh et al. "The gut microbiota in IBD". en. In: *Nature Reviews Gastroenterology & Hepatology* 9.10 (Oct. 2012), pp. 599–608. ISSN: 1759-5045, 1759-5053. DOI: 10.1038/nrgastro.2012.152. URL: http://www.nature.com/articles/nrgastro.2012.152 (visited on 07/15/2019).

[130] Xue Wang et al. "MS1 ion current-based quantitative proteomics: A promising solution for reliable analysis of large biological cohorts". en. In: *Mass Spectrometry Reviews* (Mar. 2019), mas.21595. ISSN: 0277-7037, 1098-2787. DOI: 10.1002/mas.21595. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mas.21595 (visited on 07/15/2019).

[131] Benoît Valot et al. "MassChroQ: A versatile tool for mass spectrometry quantification". en. In: *PROTEOMICS* 11.17 (Sept. 2011), pp. 3572–3577. ISSN: 1615-9861. DOI: 10.1002/pmic.201100120. URL: http://onlinelibrary.wiley.com/doi/10.1002/pmic.201100120/abstract (visited on 02/07/2017).

[132] Cosmin Lazar et al. "Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies". en. In: *Journal of Proteome Research* 15.4 (Apr. 2016), pp. 1116–1125. ISSN: 1535-3893, 1535-3907. DOI: 10.1021/acs.jproteome.5b00981. URL: http://pubs.acs.org/doi/10.1021/acs.jproteome.5b00981 (visited on 05/29/2019).

[133] Donald B. Rubin. "Inference and missing data". en. In: *Biometrika* 63.3 (1976), pp. 581–592. ISSN: 0006-3444, 1464-3510. DOI: 10.1093/biomet/63.3.581. URL: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/63.3.581 (visited on 07/16/2019).

[134] T. Aittokallio. "Dealing with missing values in large-scale studies: microarray data imputation and beyond". en. In: *Briefings in Bioinformatics* 11.2 (Mar. 2010), pp. 253–264. ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbp059. URL: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbp059 (visited on 07/16/2019).

[135] Amit Patel. "Development and Evaluation of Statistical Approaches in Proteomic Biomarker Discovery". English. PhD thesis. 2012. URL: https://pdfs.semanticscholar.org/9dac/3a5c9f27647d14cf8473807eb567670db655.pdf (visited on 07/16/2019).

[136] Bobbie-Jo M. Webb-Robertson et al. "Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics". en. In: *Journal of Proteome Research* 14.5 (May 2015), pp. 1993–2001. ISSN: 1535-3893, 1535-3907. DOI: 10.1021/pr501138h. URL: http://pubs.acs.org/doi/10.1021/pr501138h (visited on 05/29/2019).

[137] P. Jonsson and C. Wohlin. "An evaluation of k-nearest neighbour imputation using likert data". In: *10th International Symposium on Software Metrics, 2004. Proceedings*. Chicago, IL, USA: IEEE, 2004, pp. 108–118. ISBN: 978-0-7695-2129-9. DOI: 10.1109/METRIC.2004.1357895. URL: http://ieeexplore.ieee.org/document/1357895/ (visited on 06/03/2019).

[138] Tauno Metsalu and Jaak Vilo. "ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap". en. In: *Nucleic Acids Research* 43.W1 (July 2015), W566–W570. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkv468. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv468 (visited on 06/03/2019).

[139] Yanbao Yu et al. "Comprehensive Metaproteomic Analyses of Urine in the Presence and Absence of Neutrophil-Associated Inflammation in the Urinary Tract". en. In: *Theranostics* 7.2 (2017), pp. 238–252. ISSN: 1838-7640. DOI: 10.7150/thno.16086. URL: http://www.thno.org/v07p0238.htm (visited on 06/03/2019).

[140] Mark D Robinson and Alicia Oshlack. "A scaling normalization method for differential expression analysis of RNA-seq data". en. In: *Genome Biology* 11.3 (2010), R25. ISSN: 1465-6906. DOI: 10.1186/gb-2010-11-3-r25. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-3-r25 (visited on 06/18/2019).

[141] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". en. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (Jan. 1995), pp. 289–300. ISSN: 00359246. DOI: 10.1111/j.2517-6161.1995.tb02031.x. URL: http://doi.wiley.com/10.1111/j.2517-6161.1995.tb02031.x (visited on 07/10/2019).

[142] A. Reiner, D. Yekutieli, and Y. Benjamini. "Identifying differentially expressed genes using false discovery rate controlling procedures". en. In: *Bioinformatics* 19.3 (Feb. 2003), pp. 368–375. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btf877. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btf877 (visited on 07/26/2019).

[143] Leo Breiman. "Classification and regression based on a forest of trees using random inputs." In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 08856125. DOI: 10.1023/A:1010933404324. URL: http://link.springer.com/10.1023/A:1010933404324 (visited on 07/20/2019).

[144] Max Kuhn. "Building Predictive Models in R Using the caret Package". en. In: *Journal of Statistical Software* 28.5 (2008). ISSN: 1548-7660. DOI: 10.18637/jss.v028.i05. URL: http://www.jstatsoft.org/v28/i05/ (visited on 07/20/2019).

[145] *#157 PCR and gel electrophoresis | Biology Notes for A level*. July 2016. URL: http://biology4alevel.blogspot.com/2016/07/157-pcr-and-gel-electrophoresis.html (visited on 06/28/2019).

[146] Encyclopaedia Britannica. *Transverse section of an artery.* 2010. URL: https://www.britannica.com/science/artery/images-videos/media/1/36874/121565.

[147] Adam Zmysłowski and Arkadiusz Szterk. "Current knowledge on the mechanism of atherosclerosis and pro-atherosclerotic properties of oxysterols". en. In: *Lipids in Health and Disease* 16.1 (Dec. 2017). ISSN: 1476-511X. DOI: 10.1186/s12944-017-0579-2. URL: http://lipidworld.biomedcentral.com/articles/10.1186/s12944-017-0579-2 (visited on 02/11/2019).

[148] P. Duriez. "Mécanismes de formation de la plaque d'athérome". fr. In: *La Revue de Médecine Interne* 25 (June 2004), S3–S6. ISSN: 02488663. DOI: 10.1016/j.revmed.2004.04.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S0248866304001250 (visited on 02/11/2019).

[149] Amanda C. Doran, Nahum Meller, and Coleen A. McNamara. "Role of Smooth Muscle Cells in the Initiation and Early Progression of Atherosclerosis". en. In: *Arteriosclerosis, Thrombosis, and Vascular Biology* 28.5 (May 2008), pp. 812–819. ISSN: 1079-5642, 1524-4636. DOI: 10.1161/ATVBAHA.107.159327. URL: https://www.ahajournals.org/doi/10.1161/ATVBAHA.107.159327 (visited on 02/11/2019).

[150] Tracie Seimon and Ira Tabas. "Mechanisms and consequences of macrophage apoptosis in atherosclerosis". en. In: *Journal of Lipid Research* 50.Supplement (Apr. 2009), S382–S387. ISSN: 0022-2275, 1539-7262. DOI: 10.1194/jlr.R800032-JLR200. URL: http://www.jlr.org/lookup/doi/10.1194/jlr.R800032-JLR200 (visited on 02/11/2019).

[151] Angelos Tsipis et al. "Novel Oral Anticoagulants in Peripheral Arterial and Coronary Artery Disease". en. In: *Cardiovascular & Hematological Agents in Medicinal Chemistry* 12.1 (Dec. 2014), pp. 21–25. ISSN: 18715257. DOI: 10.2174/1871525712011412010926 28. URL: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1871-5257&volume=12&issue=1&spage=21 (visited on 02/12/2019).

[152] Peter Libby, Paul M Ridker, and Göran K. Hansson. "Progress and challenges in translating the biology of atherosclerosis". en. In: *Nature* 473.7347 (May 2011), pp. 317–325. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature10146. URL: http://www.nature.com/articles/nature10146 (visited on 03/14/2019).

[153] Jens Roat Kultima et al. "MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit". en. In: *PLoS ONE* 7.10 (Oct. 2012). Ed. by Jack Anthony Gilbert, e47656. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0047656. URL: http://dx.plos.org/10.1371/journal.pone.0047656 (visited on 09/13/2017).

[154] Sarah Highlander. "Mock Community Analysis". en. In: *Encyclopedia of Metagenomics*. Ed. by Karen E. Nelson. New York, NY: Springer New York, 2014, pp. 1–7. ISBN: 978-1-4614-6418-1. DOI: 10.1007/978-1-4614-6418-1_54-1. URL: http://link.springer.com/10.1007/978-1-4614-6418-1_54-1 (visited on 09/12/2017).

[155] Yubo Hou and Senjie Lin. "Distinct Gene Number-Genome Size Relationships for Eukaryotes and Non-Eukaryotes: Gene Content Estimation for Dinoflagellate Genomes". en. In: *PLoS ONE* 4.9 (Sept. 2009). Ed. by Rosemary Jeanne Redfield, e6978. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0006978. URL: https://dx.plos.org/10.1371/journal.pone.0006978 (visited on 05/21/2019).