Apport de la modélisation pour une meilleure stratification des populations à risque

Stéphanie Monnerie

INRA - UNIVERSITE CLERMONT AUVERGNE
ECOLE DOCTORALE SCIENCES FONDAMENTALES

THESE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE CLERMONT AUVERGNE

Spécialités : Bio-informatique, bio-analyse, métabolomique

Présentée et soutenue publiquement le 13 novembre 2019 par

Stéphanie MONNERIE

Apport de la modélisation pour une meilleure stratification des populations à risque

Membres du Jury :

Fabien JOURDAN (Rapporteur)	Directeur de Recherche, INRA Toulouse
Serge RUDAZ (Rapporteur)	Professeur, Université de Genève
Christine DES ROSIERS (Examinateur)	Professeur, Université de Montréal
Anne-Marie DELORT (Examinateur)	Directeur de Recherche, Université Clermont Auvergne
Estelle PUJOS-GUILLOT (Directrice de thèse)	Ingénieur de Recherche HDR, INRA Clermont-Ferrand
Pierrette GAUDREAU (Co encadrante, invitée)	Professeur, Université de Montréal
Blandine COMTE (Invitée)	Directeur de Recherche, INRA Clermont-Ferrand

REMERCIEMENTS

Dans une carrière scientifique, réaliser une thèse est une étape qui peut se révéler éprouvante comme très gratifiante. Toutefois, il serait difficile de l'achever sans être parfaitement entouré…

Je tenais à remercier avant tout ma directrice de thèse **Estelle PUJOS-GUILLOT** pour son soutien permanent au cours de ces 3 années. Grâce à ton encadrement je ne me suis jamais sentie démunie au cours de cette thèse. J'estime avoir énormément appris tant sur le plan scientifique que sur le plan personnel en développant un peu plus chaque jour mes capacités. Ta disponibilité au quotidien fut, à mon sens, l'un des éléments clé de la réussite de cette thèse. Je te remercie de m'avoir fait confiance et de m'avoir confié ce beau projet. Outre l'aspect professionnel, je garderai également un excellent souvenir de ces 3 années et des moments d'échanges que nous avons pu avoir à l'occasion de nos différents déplacements ou de petits moments de pauses autour de bons repas. Il va de soi que j'adresse le même type de remerciements à **Blandine COMTE** qui fut une véritable co-encadrante au quotidien, faisant preuve du même dévouement qu'Estelle.

Je tenais également à remercier **Pierrette GAUDREAU** pour son implication outre Atlantique dans le bon déroulement de ma thèse et du projet de recherche. Je tenais à te remercier de ton chaleureux accueil lors de ma venue au Québec, ainsi que de ton aide tout au long de ces 3 années.

Je souhaitais remercier les membres de mon comité de suivi de thèse : **Jean-Philippe ANTIGNAC**, **Julien BOCCARD** et **Anne-Marie DELORT**, pour leurs précieux commentaires et suggestions, ainsi que pour les échanges scientifiques que nous avons pu avoir. Ainsi que les membres de mon jury de thèse qui ont accepté d'évaluer mon travail : **Anne-Marie DELORT**, **Christine DES ROSIERS**, **Fabien JOURDAN** et **Serge RUDAZ**.

J'adresse ensuite mes sincères remerciements aux équipes avec lesquelles j'ai eu le plaisir de travailler. Avant tout les membres de la plateforme d'exploration du métabolisme de Theix : **Marion BRANDOLINI-BUNLON**, **Delphine CENTENO**, **Christophe DUPERIER**, **Stéphanie DURAND**, **Franck GIACOMONI**, **Charlotte JOLY**, **Bernard LYAN**, **Carole MIGNE**, **Nils PAULHE**, **Mélanie PETERA** et **Corinne POUYET**. Vous avez tous contribué à ce projet par différents biais et cette thèse n'aurait pas été la même sans votre bonne humeur et votre soutien. Je n'oublie pas **Céline DALLE** qui fut également très présente aux côtés des membres de la plateforme et a tenu un rôle tout aussi important. J'adresse également un remerciement tout particulier à **Mathieu RAMBEAU** qui fut mon collègue de bureau durant presque 2 ans : ta gentillesse et ta bienveillance m'ont permis de me sentir facilement à ma place.

Je remercie chaleureusement les différents collaborateurs avec qui j'ai échangé tout au long de ma thèse : les membres de la cohorte NuAge **José MORAIS**, **Hélène PAYETTE**, **Nancy PRESSE** ; **Daniela ZIEGLER** de l'Université de Montréal; et les membres de l'infrastructure metaboHUB.

Je tenais également à adresser une pensée chaleureuse à tous les participants aux côtés desquels j'ai vécu l'aventure *Ma Thèse en 180 Secondes*, cette parenthèse dans ma thèse fut une expérience incroyable qui n'aurait jamais pu être la même sans vous.

Je ne peux clore ces remerciements sans citer tous mes proches…

Avant tout mes amis qui m'ont offert de précieux moments de partage et de joie. Je pense inévitablement à ceux qui ont dépassé le stade des simples amis et qui sont à mes yeux des membres de la famille : **Alyson et Valentin GUEUGNEAU**. Vous m'avez soutenu au cours de cette thèse avant tout parce que vous savez vous aussi ce que cela représente, mais surtout car vous êtes une véritable bouffée d'oxygène. Je pense également à ceux qui ont partagé avec nous les dernières vacances avant la fin de cette thèse : **Mégane LAPALUS** et **Guillaume PRETRE**, et quel voyage… ! Je remercie aussi nos amis les Auvergnoux : **Julia BECHAUX**, **Arnaud CASTRES**, **Jeanne DANON**, **Vincent MADER**, **Clémentine PASQUET** et **Maxime SARKIS** pour les francs moments de rigolade et le fameux bivouac.

J'adresse des remerciements des plus sincères à tous les membres de ma famille qui m'auront encouragé et soutenu au cours de cette thèse, mais également de tout mon parcours. Mes parents, mes grands-parents, mon petit frère, ma tante et mes cousins. On n'oublie jamais là d'où l'on vient et ce que l'on y a appris. Votre amour m'a fait grandir, prendre confiance en moi et m'épanouir… C'est probablement le plus beau cadeau que vous puissiez tous me faire.

Enfin je terminerais par la personne exceptionnelle qui m'a accompagné et supporté tout au long de ces 7 dernières années et qui deviendra prochainement mon mari : **Antoine DEFFROMONT**. Ton soutien dans les bons comme dans les mauvais moments aura été indispensable. Sans toi, je n'aurais pas pu tout mener de front (avouons que construire une maison durant sa thèse n'est pas forcément une mince affaire). Je te remercie d'avoir su me remotiver quand je doutais, de m'avoir fait sourire quand j'en avais besoin et de m'avoir encouragé tout au long de ces 3 ans. Je ne pourrai jamais te remercier assez d'être à mes côtés chaque jour.

RESUME

Titre : Apport de la modélisation pour une meilleure stratification des populations à risque.

Mots clés : Bio-informatique, Traitement de données, Métabolomique, Syndrome métabolique.

Résumé :

Le projet de thèse s'inscrit dans une démarche d'épidémiologie des systèmes qui consiste à identifier les contributeurs de pathologies complexes, à de multiples niveaux, ainsi que leurs interactions et ce en utilisant une approche système. Cette dernière combine généralement des données de type omique à des données épidémiologiques observationnelles pour répondre un objectif fixé. Le but à long terme est de pouvoir utiliser la métabolomique pour reclassifier ces pathologies. Pour cela, il est nécessaire de modéliser les phénotypes des populations à risque, par des approches statistiques/mathématiques.

Dans ce contexte, ce projet s'intéresse à la caractérisation du syndrome métabolique (SMet) chez la personne âgée par des méthodes de phénotypages multidimensionnelles (métabolomique, lipidomique, phénotypique, nutritionnelle…). Différents sous-objectifs bio-informatiques ont été définis pour parvenir à cette fin. Ils concernent d'une part la prise en charge et la gestion de larges volumes de données complexes (métabolomique/lipidomique et épidémiologique), et d'autre part, des aspects d'extraction de connaissance à partir de ces données multidimensionnelles, dans un processus multi étapes et itératif.

Sur le plan bio-informatique, ce projet a tout d'abord permis la création d'une base de données de biomarqueurs du SMet issus de la littérature. Il a également permis la mise en place d'un cahier des charges pour la gestion des données des futurs projets de métabolomique.

Par la suite, un workflow complet et reproductible, d'extraction de connaissance a été développé, visant à extraire une signature (ensemble de biomarqueurs) à partir de données métabolomiques/lipidomiques non ciblées multi-plateformes. Il a inclus le développement d'un outil de filtration des corrélations analytiques rencontrées en métabolomique lors de la génération des données ; la mise en place d'une stratégie de sélection de variables utilisant différents modèles statistiques afin d'arriver à proposer une signature du SMet chez la personne âgée ; ou encore la mise en œuvre d'outils bio-informatiques permettant d'aller plus loin dans l'interprétation biologique des données à travers la visualisation dans des réseaux métaboliques.

Enfin, une démarche semblable a été mise en place pour modéliser le spectre phénotypique du SMet. En particulier, elle a consisté à étudier les potentiels sous-phénotypes du SMet pour tenter de proposer une reclassification moléculaire des individus basée sur les données métabolomiques.

ABSTRACT

Title: Contribution of modelling for a better risk population stratification**.**

Key words: bioinformatics, data analysis, metabolomics, metabolic syndrome

Abstract:

This PhD project was part of a system epidemiologic approach which aim to identify contributors of complex pathologies, at multiple levels, and their interactions by using system approach. This usually combine data from omics analysis and observational epidemiologic data to rich a fixed goal. The long term objective is to be able to use metabolomics to re-classify those pathologies. To this end, it is necessary to modelling phenotypes of at risk population using statistical and mathematical approaches.

In this context, this project aim to characterise Metabolic Syndrome (MetS) in elderly people using multidimensional phenotyping (metabolomics, lipidomics, phenotyping, nutrion…). Different bio-informatics goals have been defined. They involve care and management of big volume of complex data, and knowledge extraction from multidimensional data in a multi-step and iterative process.

From bio-informatics field, this project first allow the creation of a MetS biomarkers from literature database. It also allow the establishment of guidelines for management of data within future metabolomics projects.

Thereafter, an entire and reproducible workflow for knowledge extraction have been developed, aiming to provide a signature (set of biomarkers) from multi-platforms non-targeted metabolomics/lipidomics data. It include the development of a new tool to manage analytical correlation generated on metabolomics data during acquisition and filter it ; deployment of variable selection strategies using different statistical models to provide a MetS signature for elderly people ; or execution of bio-informatics tools to go further in biological interpretation of data across visualization on metabolomics networks.

Finally, a similar approach was design for sub-phenotyping of MetS modelling. It specially consist to study potential sub-phenotypes of Mets to propose a molecular re-classifying of individuals based on metabolomics data.

SOMMAIRE

ABREVIATIONS

ACorF : Analytic Correlation Filtration

ACP : Analyse en Composante Principale

AFM : Analyse Factorielle Multiple

APCI : Atmospheric Pressure Chemical Ionization (ionisation chimique à pression atmosphérique)

BP : blood pressure (pression artérielle)

CAH : Classification Ascendante Hiérarchique

ChEBI : Chemical Entities of Biological Interest

CI : Chemical Ionization (ionisation chimique)

CV : coefficients de variation

Da : Dalton

DEXA : Dual X-ray Absorptiometry (absorptiométrie à rayon X en double énergie)

EI : Electronic Impact (impact électronique)

ESI : Electrospray Ionisation (électrospray ou électronébulisation)

FA : fatty acids (acides gras)

FAIR : Findable Accessible Interoperable Reusable

GC : Gas Chromatography (chromatographie en phase gazeuse)

HCPC : Classification Hiérarchique sur Composantes Principales

HDL-C : High Density Lipoprotein Cholesterol

HeTOP : Health Terminology/Ontology Portal

HILIC : chromatographie d'interaction hydrophile

IMC : Indice de Masse Corporelle

InChi : IUPAC International Chemical Identifier

IRM : Imagerie par Résonance Magnétique

IUPAC CPEP : International Union of Pure and Applied Chemistry, Committee on Printed and Electronical Publications

KEGG : Kyoto Encyclopedia of Genes and Genomes

kNN : k-nearest neighbors (k plus proches voisins)

LC : Liquid Chromatography (chromatographie en phase liquide)

m/z : masse/charge

MeSH : Medical Subject Headings

NCEP ATP III : National Cholesterol Education Program Adult Treatment Panel III

netCDF : Network Common Data Form

O-PLS : Orthogonal Partial Least Squares

PC : phosphatidylcholines

PE : phosphatidylethanolamines

PFEM : Plateforme d'Exploration du Métabolisme

PLS : Partial Least Squares

PLS-DA : Partial Least Squares – Discriminant Analysis

ppm : parties par million

Q : quadripôle

QC : Quality Controls (contrôles qualité)

RF : Random Forests

RMN : Résonance Magnétique Nucléaire

RMN-1D : Résonance Magnétique Nucléaire 1-Dimension

RMN-2D : Résonance Magnétique Nucléaire 2-Dimension

SMILES : Simplified Molecular Input Line Entry Specification

SVM : Support Vector Machine

TG : triglycérides

TI : trappe d'ions

ToF : Time of Flight (temps de vol)

uma : unité de masse atomique

VIP : Variable Importance in Projection

W4M : Workflow4Metabolomics

WC : Waist Circumference (tour de taille)

INTRODUCTION

I. Contexte scientifique

La mondialisation des marchés alimentaires, l'urbanisation, la croissance économique ou encore la sédentarité sont autant de facteurs qui ont été associés à l'augmentation de la prévalence des maladies métaboliques chroniques telles que le diabète de type 2 au cours du vieillissement.

Du fait du coût de traitement de telles pathologies sur les systèmes de santé et de leur incidence sur l'espérance de vie, leur prise en charge est devenue un enjeu de santé publique à l'échelle mondiale. Dans un contexte de recherche multidisciplinaire innovante ayant un impact sur le Bien Vieillir, une meilleure compréhension de ces états de santé et l'identification du risque de l'évolution vers ces maladies est indispensable pour envisager une gestion plus adaptée et plus précoce, ainsi qu'une meilleure prévention.

1. Le Syndrome Métabolique

1.1. Définition des critères

Le syndrome métabolique (SMet) est défini comme un état physiopathologique impliquant différents critères de dysfonctions, tous découlant de problèmes cardio-métaboliques, conduisant au développement de maladies cardiovasculaires et/ou au diabète de type 2 et aux complications qui lui sont associées [1]. La définition de ces critères est soumise à discussion depuis de nombreuses années. Ainsi, le **Tableau 1** présente l'évolution des seuils au cours du temps.

	WHO (1998)	EGIR (1999)	NCEP ATP III (2001)	AACE (2003)	IDF (2005)	AHA (2005)
Critères	**1 critère obligatoire + 2 parmi les 6 autres**	**1 critère obligatoire + un ou plusieurs autres**	**Au moins 3 des 5 critères**	**-**	**1 critère obligatoire + au moins 2 parmi les 4 autres**	**Au moins 3 des 5 critères**
Obésité	Ratio taille/hanche > 90cm (H), > 85cm (F) ou IMC > 30 kg/m²	Tour de taille ≥ 94 cm (H), ≥ 80 cm (F)	Tour de taille > 102 cm (H), > 88cm (F)	IMC ≥ 25 kg/m²	Obésité abdominale requise (dépendante de l'origine ethnique)	Tour de taille > 102 cm (H), > 88cm (F)
Insulino-résistance	Insulino-résistance ou diabète de type 2 obligatoire	Insuline plasmatique > 75ème percentile		Mauvaise tolérance au glucose ou hyperglycémie à jeun		
Glucose	Mauvaise tolérance au glucose ou hyperglycémie à jeun ou diabète de type 2	Mauvaise tolérance au glucose ou hyperglycémie à jeun (exclus diabète de type 2)	Glucose à jeun ≥ 6,1mM (incluant les diabétiques)	Mauvaise tolérance au glucose ou hyperglycémie à jeun (exclus diabète de type 2)	Glucose à jeun ≥ 5,5 mM ou diabète de type 2	Glucose à jeun ≥ 5,5mM ou traitement hypoglycémiant
Hypertriglycéridémie	Triglycérides ≥ 1,7 mM	Triglycérides ≥ 1,7 mM	Triglycérides ≥ 1,7 mM	Triglycérides ≥ 1,7 mM	Triglycérides ≥ 1,7 mM ou traitement hypolipémiant	Triglycérides ≥ 1,7 mM ou traitement hypolipémiant
Hypo-HDL cholestérolémie	Cholestérol HDL < 1 mM	Cholestérol HDL < 1 mM	Cholestérol HDL < 1 mM (H), < 1,2 mM (F)	Cholestérol HDL < 1 mM (H), < 1,2 mM (F)	Cholestérol HDL < 1 mM (H), < 1,2 mM (F) ou traitement hypolipémiant	Cholestérol HDL < 1 mM (H), < 1,3 mM (F) ou traitement hypolipémiant
Hypertension	Pression systolique > 140 mmHg ou pression diastolique > 90 mmHg	Pression systolique > 140 mmHg ou pression diastolique > 90 mmHg ou traitement antihypertenseur	Pression systolique >130 mmHg ou pression diastolique >85 mmHg	Pression systolique > 130 mmHg ou pression diastolique > 85 mmHg	Pression systolique > 130 mmHg ou pression diastolique > 85 mmHg ou traitement antihypertenseur	Pression systolique > 130 mmHg ou pression diastolique > 85 mmHg ou traitement antihypertenseur
Autre	Micro albuminurie : excrétion d'albumine urinaire ≥20 µg/min ou ratio albumine/créatinine ≥30 µg/g	Historique familial (diabète de type 2, hyper uricémie…)		Autres caractéristiques de l'insulino-résistance		

Tableau 1 : les différents critères du Syndrome Métabolique selon les six définitions faites de ce dernier. Adapté de Grundy *et al*, 2005 [2]
M : hommes, F : femmes ; IMC : Indice de Masse Corporelle.

La première définition date de 1998. L'Organisation Mondiale de la Santé (OMS/WHO) avait alors définit le SMet comme un état physiopathologique caractérisé tout d'abord par une résistance à l'insuline, soit une déficience dans la réponse des organes périphériques à l'insuline et donc par conséquent, une augmentation de la glycémie. Cette insulino-résistance devait être accompagnée d'au moins deux des cinq critères suivants : un Indice de Masse Corporelle (IMC) ou un tour de taille élevé, une hypertriglycéridémie, une diminution du niveau d'HDL-Cholestérol (HDL-C) circulant, de l'hypertension et une concentration urinaire en albumine élevée [3].

En 1999, l'European Group for Study of Insulin Resistance (EGIR) a proposé des modifications de la définition faite par l'OMS. Il a notamment été proposé d'éliminer les sujets atteints de diabète de type 2 et de n'inclure que ceux présentant une résistance à l'insuline (considérant que cette dernière était précurseur de l'état diabétique). Le terme Syndrome d'Insulino-Résistance (SIR) est alors utilisé [4].

Par la suite, en 2001, la définition du National Cholesterol Education Program Adult Treatment Panel III (NCEP ATP III) est publiée, supprimant la notion d'hyper-albuminémie urinaire et remplaçant la notion de résistance à l'insuline indispensable, par celle de l'hyperglycémie à jeun. De plus, les seuils pour la définition de l'hypertension sont diminués de 10 points pour la tension artérielle systolique et de 5 pour la diastolique. Il est alors établi que la présence d'au moins trois critères sur les cinq énumérés précédemment est nécessaire à la validation de cet état métabolique [5].

En 2003, l'American Association of Clinical Endocrinologists (AACE) a apporté des modifications à la définition faite en 1998 dans le but de replacer l'insulino-résistance au cœur de la définition. Le terme SIR est à nouveau utilisé afin de désigner cet état pré-diabétique. Il n'est, pour cette définition, pas spécifié de nombre de critère minimal, cette donnée étant laissée à l'appréciation des cliniciens. Il y est par contre, recommandé de tenir compte des antécédents familiaux du patient [6].

En 2005, l'International Diabetes Federation (IDF) tente de mettre en place un consensus entre toutes ces définitions. Elle considère la mesure du tour de taille comme un critère indispensable et suffisamment corrélée à l'insulino-résistance pour ne pas nécessiter de mesures supplémentaires de cette dernière. Toutefois, cette corrélation varie en fonction des origines ethniques des populations; des critères de tour de taille ont donc été déterminés. Ainsi, pour les Européens, il faut une valeur ≥ 94 cm pour les hommes, et ≥ 80 cm pour les femmes, tandis que pour les populations Asiatiques, le seuil masculin passe à ≥ 90 cm, exception faite de la population Japonaise qui elle, voit la tendance inversée avec une limite à ≥ 85 cm et ≥ 90 cm pour respectivement, les hommes et les femmes. Pour valider le diagnostic du SMet, la présence de 2 critères supplémentaires parmi les 4 autres est nécessaire [7].

Enfin, la définition de l'American Heart Association/National Heart, Lung and Blood Institute (AHA/NHLBI) conforte celle de la NCEP ATP III dont elle conserve la majorité des critères en apportant simplement une modification du seuil de glycémie abaissé suite aux nouvelles recommandations de l'American Diabetes Association (ADA) en 2003 [2].

Des comparaisons entre ces différentes définitions ont été faites à plusieurs reprises, permettant ainsi de dégager une forme de consensus qui reste très proche de la définition faite par le NCEP ATP III, l'une des plus communément admise jusqu'en 2005 [8-15]. Pour être considéré comme atteint de SMet selon cette définition consensus [16], un sujet devra donc présenter au moins 3 des 5 critères suivants :

- Un tour de taille ≥ 102 cm chez les hommes et ≥ 88 cm chez les femmes.
- Une pression artérielle systolique ≥ 130 mm Hg ou une pression artérielle diastolique ≥ 85 mm Hg ou un traitement antihypertenseur.
- Un niveau de glucose sanguin à jeun ≥ 100 mg/dL (5,5 mM) ou un traitement hypoglycémiant.
- Un niveau sanguin de triglycérides à jeun ≥ 150 mg/dL (1,7 mM) ou un traitement hypolipémiant.
- Un niveau d'HDL-C à jeun < 40 mg/dL (1,03 mM) chez les hommes et < 50 mg/dL (1,3 mM) chez les femmes ou un traitement hypolipémiant.

Il est possible de rencontrer des modifications de seuil pour certains critères, dépendamment des populations étudiées (*e.g.* âge, genre, origines ethniques), notamment pour le tour de taille.

1.2. Prévalence du Syndrome Métabolique

En 2004, Cameron *et al* ont montré la forte prévalence du SMet à travers le monde. En utilisant la définition NCEP ATP-III, ils ont comparé différentes études menées dans le monde entier entre 1987 et 2003 [17]. Leurs résultats (**Tableau 2**) montrent qu'il touche des populations de tous âges (de 20 à 80 ans) et des deux sexes : de 8% à 44% pour les hommes, et de 7% à 57% pour les femmes. Il existe de fortes différences en fonction des origines ethniques comme le montre les différentes études réalisées par Meig *et al [18]* aux Etats-Unis portant sur des populations incluant des sujets de même âge, vivant dans la même zone géographique mais d'origines différentes (hispanique, afro-américaine).

Pays	Années	Groupe âge	Référence bibliographique	Prévalence (%)	
				Homme	Femme
Mauritanie	1987	>24	Cameron *et al* [19]	16.6	14.7
Finlande	1988-1989	42-60	Laaksonen *et al* [20]	13.7	-
États-Unis	1988-1994	>19	Ford *et al* [21]	24.2	23.5
États-Unis (Amérindiens)	1988-1994	45-49	Resnick *et al* [22]	43.6	56.7
États-Unis	1991-1995	30-79	Meigs *et al* [18]	26.9	21.4
Mexique	1992-1993	20-69	Aguilar-Salinas *et al* [23]	26.6	
États-Unis (blancs non hispaniques)	1992-1996	30-79	Meigs *et al* [18]	24.7	21.3
États-Unis (Américano-Mexicains)	1992-1996	30-79	Meigs *et al* [18]	29	32.8
États-Unis (Américains Philippins)	1992-1999	50-69	Araneta *et al* [24]	-	34.3
France	1996	30-64	Balkau *et al*[25]	10	7
Australie	1999-2000	>24	Données non publiées	19.5	17.2
Iran	1999-2001	>20	Azizi *et al* [26]	24	42
Turquie	2000	>31	Onat *et al* [27]	27	38.6
Sultanat d'Oman	2001	>20	Al-Lawati *et al* [28]	19.5	23
Inde	2002	20-75	Deepa *et al* [29]	36.4	46.5
Inde	2003	>20	Gupta *et al* [30]	7.9	17.5
Irlande	2003	50-69	Villegas *et al* [31]	21.8	21.5

Tableau 2 : Prévalence du SMet à travers le monde en 2004, selon la définition de la NCEP ATP-III. D'après Cameron *et al* [17]

Une étude de 2009 a montré que la prévalence du SMet aux États-Unis chez les sujets de plus de 20 ans était de 34%, et qu'elle augmentait avec l'âge (41% des hommes et 37 % des femmes de 40 à 59 ans, et 52% des hommes et 54% des femmes de plus de 60 ans) [32]. Ce même phénomène a également été observé en Chine avec une prévalence de 24,5% chez les sujets de plus de 15 ans, de 13,9% chez les 15-39ans, de 26,4% chez les 40-59 ans et de 32,4% chez les sujets de plus de 60 ans [33].

La revue systématique menée par Van Vliet-Ostaptchouk *et al* en 2014, étudiant 10 cohortes différentes (soit 163 517 individus) en Europe, a montré que la prévalence, tous sexes confondus, chez les sujets obèses, varie entre 42,7% chez les Italiens et 78,2% chez les Finlandais, avec une fréquence plus élevée chez les femmes : 24% en Italie et 64,8% en Finlande. Alors que la prévalence du SMet est plus élevée chez les hommes que chez les femmes dans les populations vivant aux Etats-Unis, il semble que le phénomène soit inversé chez les individus européens, tout du moins lorsque ces derniers sont obèses. Si la prévalence du SMet est plus élevée chez les sujets obèses, il ne s'agit pas pour autant de la seule population touchée par cet état de santé. En effet, plusieurs études menées chez des sujets non obèses ont montré que la prévalence, tous sexes confondus, y était également importante : elle

atteint 17,3% chez les Polonais et les Norvégiens [34], 15,1% chez les sujets Indiens-Asiatiques [35] et 29,6% dans la population Brésilienne [36].

Chez les sujets obèses Européens de l'étude de Van Vliet-Ostaptchouk *et al*, l'hypertension artérielle est le critère du SMet le plus représenté après l'obésité, avec 60 à 85% des individus validant ce dernier. A l'inverse, l'hyperglycémie est le critère le moins rencontré [37]. Dans la population Brésilienne, c'est l'hypo-HDL-cholestérolémie (59,3%) et l'hypertension (52,5%) qui sont les composantes les plus présentes et non l'obésité [36]. Au sein de la population Chinoise, la composante la plus présente chez les femmes est l'obésité (46,1%), tandis qu'il s'agit de l'hypertension (52,8%) chez les hommes [33]. Toutes ces études illustrent le fait que dépendamment de la population étudiée la répartition des critères du SMet est différente.

Par ailleurs, une revue systématique réalisée en 2013 par Friend *et al*, a montré une prévalence mondiale inquiétante du SMet chez les enfants : elle représente 3,3% de la population infantile. Parmi les enfants en surpoids, 11,9% sont identifiés avec le SMet et le pourcentage grimpe à 29,2% chez les sujets obèses. Il a également été remarqué, comme chez les adultes, que les garçons sont plus touchés que les filles (5,1% contre 3,0%). Les données suggèrent des différences de prévalence entre les différents groupes ethniques sans pour autant les rendre quantifiables [38]. Tout comme observé dans la plupart des populations adultes, il a été montré, dans une cohorte d'enfants Pakistanais, que la composante du SMet la plus observée (54%) était l'hypertension, la seconde composante étant l'hypo-HDL-cholestérolémie (36,5%) [39].

1.3. Une caractérisation encore variable

Les contours de la définition du SMet ainsi que les seuils de validation des critères restent à ce jour encore variables. Les différentes études menées récemment ne font pas toutes appel à la même définition et l'uniformisation de cette dernière n'est pas encore une réalité. D'autre part, sa définition fait intervenir plusieurs critères laissant entrevoir la complexité de cet état de santé ainsi que l'existence de différents sous-phénotypes au sein des populations touchées. En effet, au vu de toutes les combinaisons de critères possibles, on peut aisément imaginer que chacune d'entre elle diffère des autres et donne lieu à un état métabolique particulier. Cela en fait donc un état physiopathologique encore mal caractérisé et compris.

Sa prévalence croissante, le coût sur les systèmes de santé et la qualité de vie des populations font de sa prise en charge et sa compréhension l'une des clés importantes de l'amélioration de la santé

mondiale. Si le SMet a été initialement associé en priorité aux populations à risque telles que les adultes en surpoids ou obèses, il semble maintenant que les populations cibles s'élargissement (*e.g.* sujets non obèses, enfants, personnes âgées).

Finalement, la difficulté à caractériser le SMet et le manque de connaissances concernant sa pathogénèse soulèvent plusieurs questions. Existe-il de meilleurs critères pour une utilisation clinique courante que ceux proposés dans les actuelles définitions ? La prise en charge du SMet dans son ensemble diffère-t-elle de celle utilisée pour chacune des composantes prise individuellement ? Tous les sujets atteints de SMet doivent-ils être considérés de la même façon ou faut-il prendre en compte la combinaison de critères comme un critère en lui-même pour différencier les sous-phénotypes ? Autant de questions auxquelles il est impossible de répondre aujourd'hui sans chercher à comprendre les phénomènes métaboliques liés à cet état de santé.

Dans ce contexte, il est devenu nécessaire de disposer d'outils de phénotypage pour permettre une meilleure caractérisation des individus concernés. Les progrès réalisés dans les domaines analytique (métabolomique et lipidomique), statistique et bio-informatique permettent maintenant d'envisager des approches systèmes pour enrichir les connaissances autours du SMet.

II. La métabolomique

1. Définitions

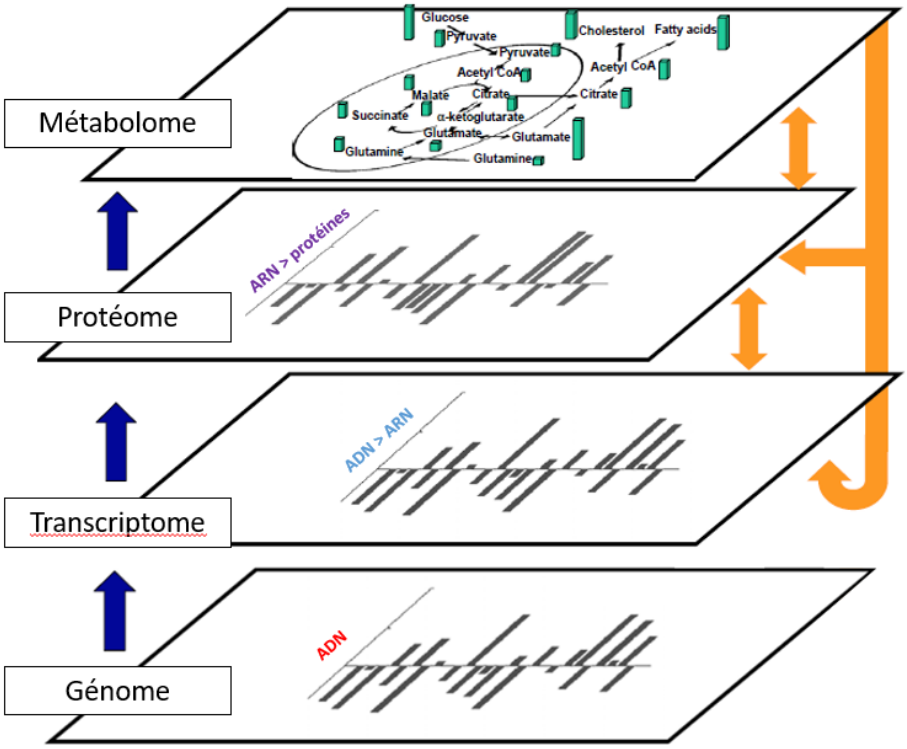


Figure 1 : Schéma des interactions entre les différents niveaux de régulation cellulaire. D'après Kelleher *et al*, 2003 [40]

Alors que le génome représente l'ensemble du matériel génétique pouvant être transcrit ou non en ARN puis traduit en protéines capables d'assurer une multitude de fonctions dans les cellules vivantes, le métabolome, lui, est le résultat final de l'expression de ces gènes dans un environnement donné. De par son rôle d'intermédiaire dans de nombreuses réactions chimiques et biologiques impliquant de nombreuses protéines, il est le lien final avec le phénotype d'un individu ou d'un organisme vivant. La **Figure 1** représente les différentes interactions entre les différents niveaux d'expression des gènes. Elle illustre également le fait qu'il existe de nombreux stades de rétrocontrôle entre les différents niveaux, notamment par l'intermédiaire des métabolites qui sont capables d'interagir par exemple avec des éléments régulateurs de l'expression des gènes.

Le métabolome est défini comme étant l'ensemble des métabolites présent dans un système biologique (dans un biofluide, un tissu ou une cellule d'un organisme vivant). Les métabolites sont de

petites molécules de faible poids moléculaire, intermédiaires de réactions du métabolisme permettant, par un ensemble de réactions chimiques, le maintien en vie de tout système vivant. Les métabolites peuvent être séparés en deux catégories : les métabolites endogènes qui sont synthétisés et utilisés au sein du système biologique étudié, et les métabolites exogènes qui proviennent de l'extérieur de l'organisme (*e.g.* alimentation). Concernant ces derniers, l'une des catégories majeures est celle des xénobiotiques qui regroupe les contaminants alimentaires, les composés synthétiques, les polluants environnementaux et les médicaments. L'exposition à ces composés est extrêmement variable d'un individu à un autre et peut être observée à travers les profils métaboliques.

Le métabolome comprend une large gamme de molécules (sucres, acides organiques, acides aminés, terpènes, alcaloïdes, toxines, etc…). Elles sont définies comme ayant un poids moléculaire < 1500 Da, et possèdent des propriétés physico-chimiques variées et des concentrations très différentes. La base de données « The Human Metabolome » (HMDB, version 4.0) [41, 42] l'illustre parfaitement. Elle contient aujourd'hui environ 114 000 métabolites différents rencontrés au sein de l'espèce humaine, chiffre que l'on pense encore sous-estimé. Pour les plantes, on estime qu'au moins 200 000 métabolites différents existent et qu'entre 7 000 et 15 000 seraient présent dans chaque espèce. Mieux identifier et comprendre ces petites molécules au sein des organismes vivant reste un enjeu majeur aujourd'hui.

Dans ce but, la métabolomique a été définie comme la mesure quantitative des réponses métaboliques multivariées de cellules, tissus, ou d'un organisme à un stimulus physio(patho)logique ou environnemental, ou à une modification génétique [43]. La lipidomique quant à elle est une approche similaire dédiée à l'analyse des lipides [44]. Elle a permis d'ouvrir de nouvelles perspectives de phénotypage métabolique et de faire évoluer les recherches vers une approche intégrative. Elle permet la comparaison de profils d'individus sains et malades pour mettre en évidence des différences d'états métaboliques ; la quantification de composés ; l'identification de molécules d'intérêt biologique potentiel et l'obtention d'informations structurales et moléculaires.

Ce type d'approches est de plus en plus utilisé dans le domaine de la santé pour plusieurs raisons : 1) 95% des diagnostics médicaux sont réalisés par dosages de petites molécules ; 2) la grande majorité des médicaments utilisés sont de petites molécules ; 3) les métabolites servent de co-facteurs à plusieurs milliers de protéines ; 4) un grand nombre de maladies génétiques sont dues à un dérèglement dans le métabolisme de petites molécules ; 5) la métabolomique donne une image du métabolisme à un instant donné qui permet de mieux comprendre les mécanismes de fonctionnement d'un organisme. La métabolomique est particulièrement utilisée pour caractériser de nouvelles pathologies par le biais de l'identification de nouveaux biomarqueurs. Ces derniers sont définis comme

des caractéristiques biologiques mesurables liées à un état de santé permettant son dépistage, son diagnostic et l'évaluation d'un traitement médical [45].

2. Approches

La métabolomique représente un challenge analytique important. En effet, il s'agit de développer des méthodes permettant la détection d'un grand nombre de métabolites, d'une grande diversité chimique (pKa, polarité, masse) et présents à des concentrations très variables dans des matrices complexes. Pour se faire, il existe 2 types d'approches : ciblées et non ciblées. Ces dernières sont représentées sur la **Figure 2**.

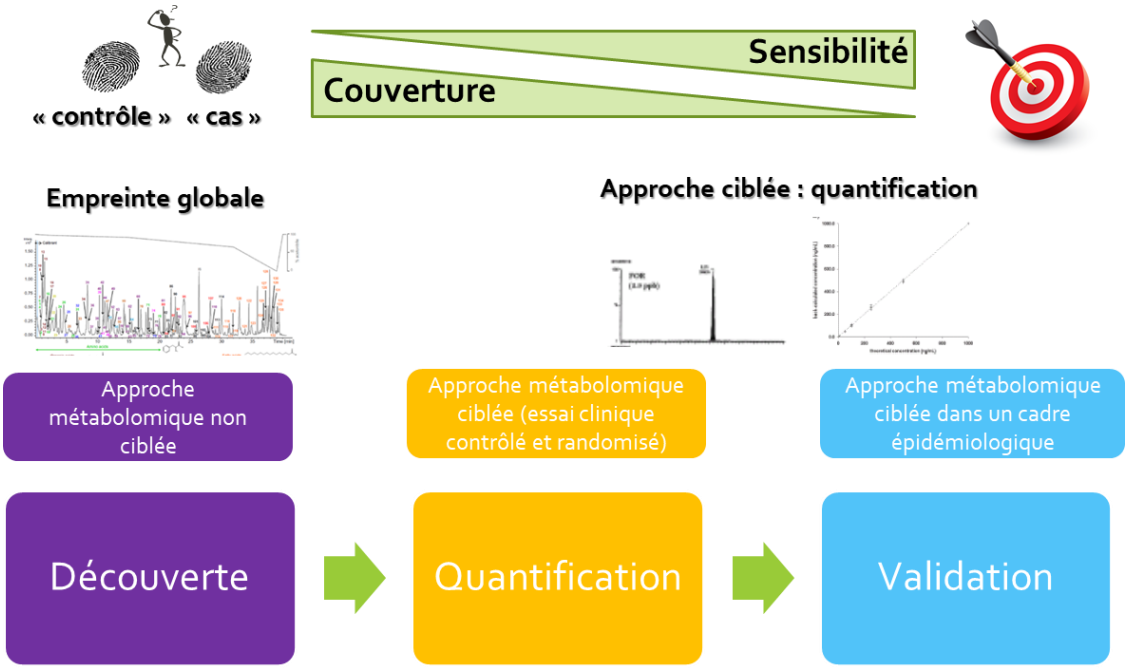


Figure 2 : Différences entre les approches de métabolomique ciblée et non ciblée. D'après Ismail *et al*, Electrophoresis 2013

Les approches ciblées permettent la détection et la quantification d'un petit nombre de métabolites préalablement identifiés et connus, grâce à une sensibilité importante mais offrent une couverture moindre. Elles peuvent parfois servir à la détection de nouveaux biomarqueurs, mais sont essentiellement utilisées pour la quantification de métabolites d'intérêt préalablement connus et la validation de biomarqueurs préalablement identifiés [46, 47].

Les approches non ciblées consistent, elles, à mesurer simultanément un très grand nombre de métabolites de manière plus exploratoire. Elles permettent ainsi d'avoir une couverture très importante et de détecter un grand nombre de métabolites présents dans un échantillon en ne fournissant que des informations semi-quantitatives de sensibilité moindre. Lorsque l'on utilise ces dernières, il n'est donc pas nécessaire de disposer d'hypothèses précises sur des groupes de métabolites donnés pouvant présenter un intérêt. Cela permet donc d'avoir une approche davantage dirigée par les données et non par les hypothèses basées sur les connaissances scientifiques préalables. Elles sont utilisées pour l'identification de nouveaux biomarqueurs grâce à la couverture importante qu'elles offrent. Toutefois, l'une des limites de ces méthodes réside dans le fait qu'il est généralement très difficile d'identifier la grande quantité de métabolites détectés [46, 47].

3. Acquisition des données

La métabolomique s'appuie essentiellement sur l'utilisation de 2 techniques d'analyse : la spectrométrie de masse (MS) et la spectroscopie par résonance magnétique nucléaire (RMN). Ces 2 méthodes d'analyse permettent l'une comme l'autre, d'identifier la structure chimique des métabolites et de mesurer des concentrations relatives ou absolues. [48].

3.1. Spectrométrie de masse

La spectrométrie de masse se base sur la transformation des molécules de leur état naturel en ions chargés à l'état gazeux. Elle permet l'acquisition de données sous la forme d'un rapport masse/charge (m/z) et d'une concentration (approche ciblée) ou d'une intensité relative (approche non ciblée) pour chaque composé mesuré. La masse est exprimée en unité de masse atomique (uma) ou en dalton (Da). La charge elle est équivalente au nombre de charges portées par l'ion (n fois la charge de l'électron).

Un spectromètre de masse est constitué en 4 blocs majeurs :

- Un dispositif d'introduction de l'échantillon
- Une source qui est chargée de la transformation des composés à analyser en ions
- Un ou des analyseurs qui séparent les espèces chargées en fonction du rapport m/z.
- Un détecteur qui quantifie les ions.

Le dispositif d'introduction et de séparation :

Pour réduire la complexité de l'échantillon biologique avant l'analyse, des méthodes de séparation sont généralement utilisées en amont du spectromètre de masse. Elles permettent de répartir dans le temps l'injection des molécules dans le spectromètre pour permettre une meilleure détection des molécules. La chromatographie en phase liquide (LC) et la chromatographie en phase gazeuse (GC) sont les 2 méthodes séparatives les plus communément utilisées en métabolomique. Elles sont appliquées selon les propriétés des métabolites étudiés. L'échantillon contenant un mélange complexe de métabolites est entraîné par un courant de phase mobile (gaz ou liquide) au contact d'une phase stationnaire (silice, polymère, silice greffée etc) présent dans une colonne. La colonne chromatographique à travers laquelle l'échantillon va passer va retenir les métabolites dépendamment des interactions qui s'établissent, selon l'affinité des métabolites avec ces deux phases. Ils ne vont donc pas tous migrer à travers la colonne à la même vitesse et détéctés en fonction de leurs propriétés chimiques. Le temps que met chaque métabolite à parcourir la colonne est appelé temps de rétention.

La source :

Il existe plusieurs types de sources : en métabolomique, principalement des sources sous vide et des sources à pression atmosphérique. Les sources d'ionisation à pression atmosphérique permettent l'ionisation des molécules sur une grande gamme de poids moléculaire (de 10^1 à 10^5 Da) et de nature polaire à très polaire tandis que les sources fonctionnant sous vide permettent d'ioniser des composés de plus faible poids moléculaire (10^1-10^3 Da) et de nature plutôt apolaire. Pour pallier à cette limitation, il est possible de réaliser une réaction de dérivation pour masquer les fonctions polaires des molécules, rendre les composés thermostables et volatiles et permettre leur ionisation sous vide. Cette opération peut également permettre de gagner en sensibilité pour obtenir des formules brutes plus spécifique et favoriser la fragmentation.

Les sources à ionisation sous vide sont principalement associées avec une séparation par chromatographie gazeuse. Parmi elles, on distingue celle à ionisation par impact électronique (EI) de celles par ionisation chimique (CI). Dans le cas des sources EI, la production des ions se fait par collision des molécules avec les électrons émis par un filament électrique. Ce type d'ionisation a tendance à générer beaucoup de fragments et les résultats obtenus sont extrêmement reproductibles d'un appareil à l'autre. Pour les sources CI, l'ionisation se fait par transfert d'électrons *via* un gaz réactant. Le choix du gaz a une influence sur la nature des molécules qui seront le plus facilement ionisées. Elle peut se faire en mode positif (formation de [M+H]+), ou en mode négatif (formation de [M-H]-). Comparativement à l'ionisation par EI elle a l'avantage d'être plus douce, de produire moins de fragments, et d'améliorer le rapport signal sur bruit.

Les sources fonctionnant à pression atmosphérique sont, elles, couplées avec la chromatographie en phase liquide. Le fait de s'affranchir des conditions de vide tout en permettant l'utilisation de phases mobiles liquides pouvant contenir de l'eau a permis l'essor de la MS dans le domaine de la biologie. Il existe plusieurs types de sources dont : électrospray ou électronébulisation (ESI) ou à ionisation chimique à pression atmosphérique (APCI). Les sources ESI reposent sur la formation d'ions en solution dans une chambre à pression atmosphérique sous l'effet d'un champ électrostatique. Elles n'engendrent pas de fragmentation et donnent accès aux molécules polaires de très haut poids moléculaire par l'intermédiaire d'ions multichargés $[MHn]^{n+}$. La composition de la phase mobile au moment de l'entrée dans la source influence la capacité à ioniser. Les sources APCI, elles, évaporent le solvant et vaporisent les analytes à pression atmosphérique sous l'effet de la chaleur. L'ionisation se produit alors en phase gazeuse selon un mécanisme proche de celui de l'ionisation chimique sous vide. Elles permettent l'ionisation des composés neutres ou moyennement polaires, volatiles et de poids moléculaire faible. La composition de la phase mobile ainsi que son débit d'arrivée dans la source influencent l'ionisation.

Les analyseurs :

Les ions produits par la source sont extraits et parcourent le spectromètre sous vide sans se décharger, ni entrer en collision avec d'autres molécules. Pour permettre la libre circulation des ions, il est important de disposer de conditions de vide poussées. Les ions sont alors séparés soit dans l'espace, soit dans le temps en fonction de leur rapport m/z.

La séparation des ions selon des critères spatiaux se fait généralement en métabolomique *via* un analyseur quadripôle (Q) ou un analyseur temps de vol (ToF). L'analyseur quadripôle est formé de 4 barres parallèles associées électriquement 2 à 2. Grâce à des variations de tension et fréquence, les

ions vont entrer en résonnance successivement pour atteindre le détecteur les uns après les autres (une fréquence correspond à un m/z). Il est possible de l'utiliser en balayant toutes les fréquences pour conserver tous les ions (mode SCAN) ou en choisissant une ou quelques fréquences d'intérêt pour isoler certains ions (mode SIM). Enfin, le principe de l'analyseur ToF est de mesurer le temps que met l'ion à se déplacer dans un tube de vol sous vide suite à son accélération par une tension connue : les ions les plus légers vont plus vite que les lourds. Il permet donc d'analyser simultanément des paquets d'ions.

La séparation des ions selon des critères temporels peut, elle, se faire *via* un analyseur à trappe d'ions (TI). L'analyseur de type trappe d'ions piège les ions formés dans une trappe, puis émet des fréquences pour relâcher les ions dépendamment de leur m/z. Comme le quadripôle, il peut être utilisé en mode SCAN ou en mode SIM

La résolution de la mesure faite par l'analyseur correspond à la qualité de distinction qu'est capable de faire l'analyseur entre 2 ions de m/z proche. Si les 2 ions sont confondus en un seul pic, alors la résolution est faible, en revanche, plus ils seront séparés en 2 pics distincts, plus la résolution est importante (**Figure 3**). La sensibilité, elle, correspond à la capacité de l'appareil à détecter des métabolites malgré une présence en faible concentration. La gamme dynamique et la masse exacte sont deux éléments essentiels à la mesure.

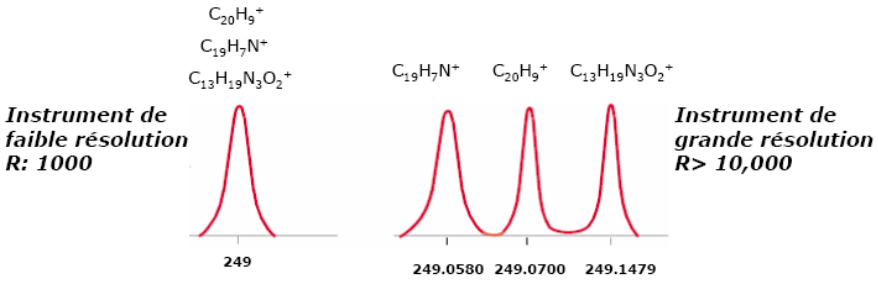


Figure 3 : Illustration de la résolution de mesure

Afin d'obtenir des informations plus complètes sur les métabolites d'un échantillon, il est possible de réaliser une analyse en tandem MS/MS. Cela consiste à coupler 2 analyseurs successivement en les séparant par une cellule de collision qui va fragmenter les ions issus du premier analyseur. La cellule de collision est le plus souvent constituée d'un gaz inerte et d'une source d'excitation afin de produire une fragmentation modérée. Le plus souvent, un premier analyseur est utilisé pour filtrer un ion qui est ensuite fragmenté et analysé dans le second analyseur (**Figure 4**). On

peut ainsi obtenir des informations structurales concernant le métabolite d'origine. L'une des configuration MS/MS les plus fréquemment utilisée en métabolomique est le LTQ-Orbitrap. Il est constitué de 2 analyseurs : un analyseur trappe linéaire et un orbitrap. Un tel couplage permet de bénéficier d'une haute résolution et d'une forte précision en masse pour identifier, sans ambiguïté, la présence de certains éléments (soufre, oxygène, azote). La précision de la mesure de masse est de l'ordre du ppm permettant d'améliorer considérablement la qualité des recherches dans les bases de données ou la caractérisation structurale des molécules : détermination des formules brutes et fragmentations MS/MS permettant de cumuler des informations complémentaires dans le cas d'interprétations structurales difficiles ou de caractérisation de métabolites inconnus.

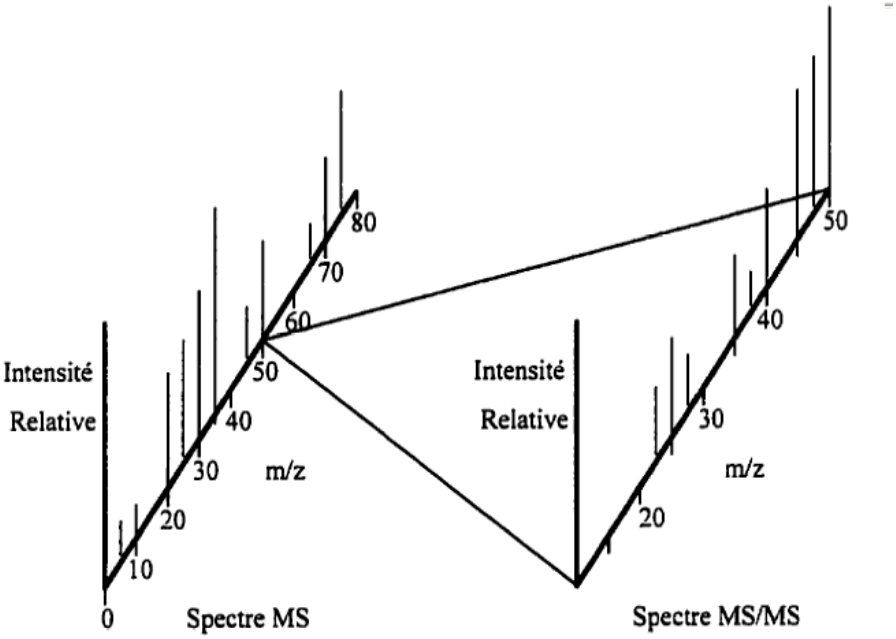


Figure 4 : Spectre MS et spectre MS/MS obtenu à partir d'un échantillon.

Le détecteur :

Il existe différents types de détecteurs, tous basés sur des principes physiques différents, mais leur rôle reste le même, compter les ions. Le détecteur va ainsi transformer le signal ionique en signal électrique. Ainsi, les données LC/MS ou GC/MS utilisées en métabolomique sont tridimensionnelles : chaque variable (ions) est caractérisée par son temps de rétention, sa masse et son intensité (**Figure 5**). Le chromatogramme représente l'intensité en fonction du temps tandis que le spectre correspond à un « zoom » sur un pic du chromatogramme pour le représenté sous forme d'intensité en fonction du m/z.

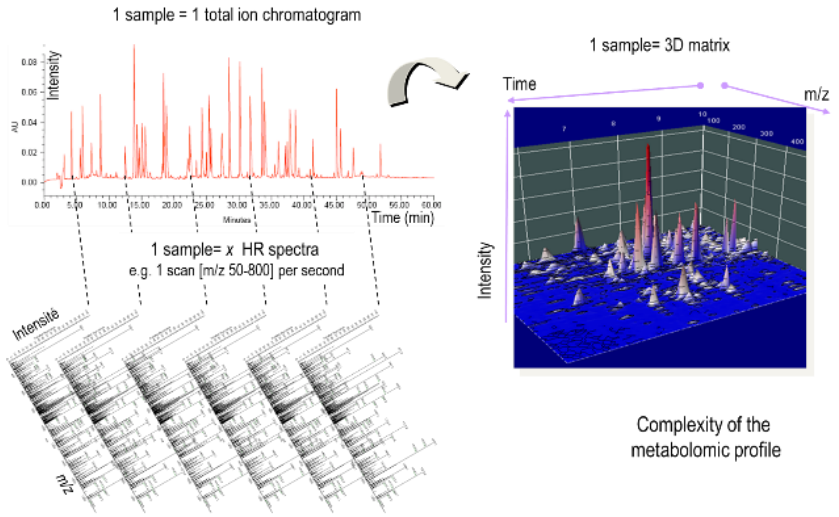


Figure 5 : Représentation d'un ion selon ses 3 dimensions (temps, m/z et intensité)

3.2. Résonance magnétique nucléaire

La spectroscopie RMN se base sur les propriétés physiques (spin) du noyau atomique observé (souvent l'hydrogène car le plus abondant) à absorber puis restituer de l'énergie lorsqu'il est soumis à une onde radio fréquence en présence du champ magnétique créé par le spectromètre [49]. Ce retour à l'état initial se traduit par une fréquence émise proportionnelle au champ magnétique perçu qui est différent selon l'environnement électronique du noyau observé (atomes voisins). La fréquence de résonance de chaque noyau se traduit sur le spectre RMN par un déplacement chimique (δ) normalisé par rapport à une référence interne présente dans l'échantillon RMN, il est alors indépendant du champ magnétique généré par le spectromètre et est exprimé en parties par million (δ ppm). Il est alors possible de comparer des spectres RMN obtenus sur des spectromètres de champs différents réalisés dans les mêmes conditions (solvant, pH, etc..). La **Figure 6** représente un spectre obtenu par RMN du proton pour un composé unique. Chaque pic (raie) observé est défini par un déplacement chimique, une aire (intégrale) qui est proportionnelle au nombre d'atomes impliqués et à une multiplicité (couplage scalaire) dépendante du nombre de noyaux voisins impliqués dans l'interaction [50].

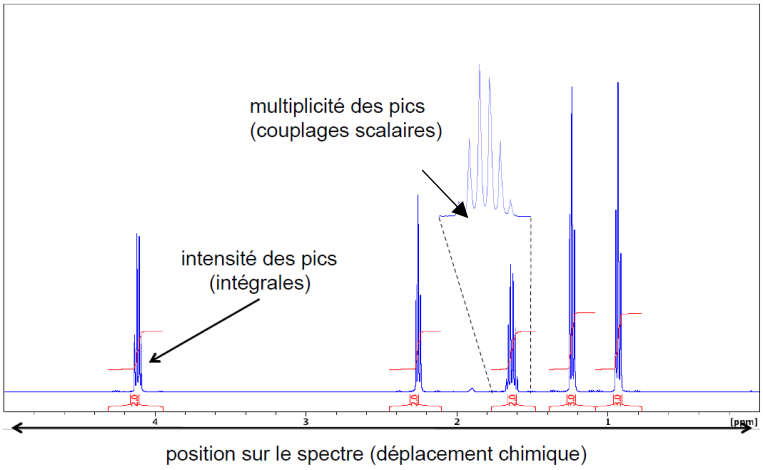


Figure 6 : Exemple de spectre en RMN du proton

La RMN 1 dimension (RMN-1D) est la plus répandue. Elle représente un axe unique de fréquence où les pics des molécules sont placés en fonction de leur fréquence de résonance (exprimée en ppm) utilisée pour l'acquisition des profils

La RMN 2 dimensions (RMN-2D), elle, utilise 2 axes de fréquence et est principalement utilisée pour les composés qui ne peuvent pas être identifiés *via* la RMN-1D. La seconde dimension peut correspondre à une mesure du même type de noyau (*e.g.* proton/proton sur la **Figure 7**) ou à une mesure d'un second type de noyau (*e.g.* proton/carbone). Elle permet de séparer les signaux des métabolites se chevauchant et donc d'apporter des informations complémentaires sur les propriétés chimiques des molécules [51]. A l'image de la MS/MS, elle permet l'identification des métabolites.

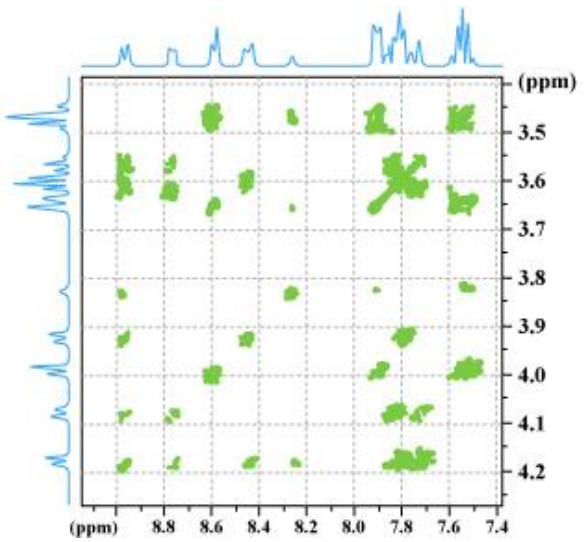


Figure 7 : Exemple de spectre en RMN du proton 2D

La RMN se base sur les propriétés magnétiques de certains noyaux atomiques pour mesurer la présence de métabolites sans appliquer nécessairement un traitement physique ou chimique à l'échantillon qui pourrait l'altérer (comme l'ionisation lors des analyses par spectrométrie de masse). Ceci lui confère un côté robuste et permet une bonne répétabilité et reproductibilité des mesures. Cela fait de la RMN une technique très précise capable de détecter de faibles variations biologiques dans le cadre d'une approche en métabolomique non ciblée. Les spectromètres RMN à haut champs sont utilisés afin d'augmenter la sensibilité et la résolution permettant d'étudier des échantillons solides (biopsies), liquides ou encore vivants (IRM = Imagerie par Résonance Magnétique). Le fait que la RMN soit une technique analytique non destructive permet de ne pas altérer l'échantillon et autorise ainsi des études directes par exemple sur des biopsies de tissus qui pourront ensuite être utilisées pour d'autres analyses [48]. De même il est possible d'effectuer sur un même échantillon différentes séquences analytiques afin d'augmenter les informations sur les composés étudiés. En plus de permettre l'acquisition d'informations quantitatives dans un mélange biologique complexe, la RMN permet également d'obtenir des informations sur la structure chimique des molécules de par les déplacements chimiques, ainsi que les interactions entre noyaux observées lors des différentes séquences d'analyse possibles (RMN-1D, RMN-2D). Cette technique reste malgré tout limitée en sensibilité.

3.3. Qualité des données

Lors de l'analyse d'échantillons par MS, il est important d'assurer la qualité des données générées. Pour se faire, 2 éléments sont primordiaux : le passage dans un ordre randomisé des échantillons, l'introduction de contrôles qualité entre les échantillons.

Lors des analyses, les appareils utilisés peuvent être amenés à s'encrasser légèrement en retenant une partie des molécules d'un échantillon après son passage. Ce phénomène s'il n'est pas maîtrisé, peut avoir un impact fort sur l'interprétation des données et la découverte de biomarqueurs, notamment quand le nombre d'échantillons analysés devient important [46, 52-54]. Il est d'autant plus essentiel de prendre en compte cet effet lorsque l'étude implique la comparaison de 2 groupes entre eux (cas/témoins, Temps de mesure X/Temps de mesure Y…). En effet, si tous les échantillons des individus cas sont par exemple analysés en premier et les échantillons témoins ensuite, l'état de l'encrassement du spectromètre est forcément moindre pour l'analyse des cas que pour celle des témoins. Par conséquent, nous ne pourrons pas comparer les 2 groupes d'échantillons sans introduire de biais relatif à l'analyse. Pour pallier à cela, il convient de passer les échantillons de manière

aléatoire/randomisée. Pour se faire, un simple tirage au sort de l'ordre de passage est possible, néanmoins il est préférable de s'assurer que le hasard ne regroupe pas les échantillons selon des critères fortement impactant comme la période de prélèvement des échantillons, ou d'autres potentiels facteurs confondants… Une méthode basée sur celle de construction de carrés latins peut alors être utilisée.

Tout comme l'étape de randomisation des échantillons, il est indispensable d'introduire des contrôles qualité (QC) tout au long de l'analyse. Les QC correspondent à l'injection d'échantillons de référence (pool des échantillons ou standards connus). Ils permettent d'assurer un bon suivi de la performance du système d'analyse (contrôles des temps de rétention, des intensités et de la calibration de la masse), d'avoir un aperçu de la variabilité de la méthode si plusieurs échantillons QC sont préparés indépendamment pour chaque batch (lot d'échantillon analysé successivement et sans interruption sur la machine) et de calculer la fonction de dérivation basée sur les échantillons QC pour supprimer les tendances systématiques et la dérive du signal.

III. Aspects bio-informatiques dans le domaine de la métabolomique et de la lipidomique

1. Production des données

1.1 Les données brutes

Les analyses métabolomiques et lipidomiques réalisées donnent lieu à la création de fichiers de données brutes (fichiers raw). Une analyse génère 1 fichier raw par échantillon. Ils permettent le stockage de l'information directement telle qu'elle est mesurée, soit 1 point de mesure par unité de temps définie (chromatogramme) correspondant lui-même à la mesure de plusieurs masses (spectre de masse). Ceci représente un volume de données important : d'une centaine de mégaoctets à 2,5 gigas par échantillon soit des données brutes pouvant représenter plusieurs centaines de gigas par projet. Le volume de ces fichiers est dépendant du temps d'acquisition défini par le protocole d'analyse, de la résolution de l'appareil et du mode d'acquisition (continuum conservant tous les points de mesure, ou centroïde transformant les pics Gaussiens en bâton pour simplifier l'information).

Les fichiers sont générés directement par les machines d'acquisition et sont donc dans des formats propriétaires. Toutefois, des formats libres existent et ont été développés pour faciliter l'interopérabilité des données ; ces derniers seront présentés par la suite.

1.2. Prétraitement des données

Afin de rendre les données brutes interprétables, il est nécessaire de réaliser un prétraitement ayant pour but de transformer l'ensemble des spectres en matrice, en éliminant les sources de variations analytiques. Cette étape permet d'extraire l'information pertinente et de résumer les fichiers multiples en une seule table contenant les intensités mesurées pour chaque variable dans chaque échantillon (**Figure 8**) [55].

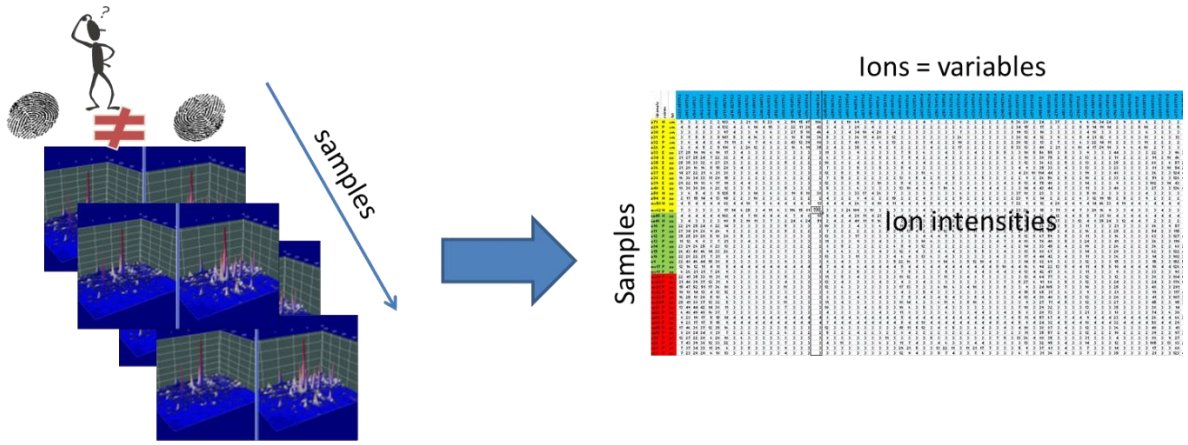


Figure 8 : Illustration du prétraitement permettant de passer de multiples fichiers à une table unique.

Cette étape complexe peut inclure un alignement des éléments spectraux, la détection des pics et leur intégration, la soustraction du bruit et des blancs, suppression des isotopes, la déconvolution des pics ou encore la normalisation. L'objectif est de prendre en compte les dérives instrumentales inévitables pour assurer la comparabilité des intensités entre elles. Le prétraitement peut se faire avec les logiciels commerciaux vendus par les constructeurs ou par le biais de solutions open-sources plus flexibles après conversion des données en format ouvert (CDF, mzML ou mzXML par exemple) [56].

a. Extraction des données brutes

A. Extraction des données de spectrométrie de masse

Extraction des pics :

Cette étape consiste à intégrer les pics chromatographiques de façon automatisée pour obtenir un tableau de données. Généralement, l'utilisateur est amené à paramétrer :

- La valeur de largeur minimale, à mi-hauteur, du pic chromatographique (exprimé en unité de temps)
- La valeur seuil d'intensité qui permet l'extraction de l'ion comparativement au bruit mesuré
- La largeur de la fenêtre de m/z pour l'extraction des pics chromatographiques

Il existe de nombreux outils pour réaliser cette étape comme MZmine [57], MetAlign [58] ou XCMS [59]. Cette étape isole les ions un a un sans chercher à reconstituer d'information globale sur le composé dont ils sont issus et elle génère donc de nombreuses redondances [60].

Alignement des ions entre échantillons :

L'extraction initiale se faisant échantillon par échantillon, pour obtenir une matrice finale, il est important de déterminer quels ions sont communs entre les échantillons. Pour ce faire, les stratégies se basent le plus souvent sur des recoupements de masses et de temps de rétention proches, qui peuvent être complétés par des paramètres tels que les proportions des échantillons dans lequel un ion doit être retrouvé ou la largeur du pic chromatographique. A l'issue de cette étape, la matrice peut contenir des valeurs manquantes pour les échantillons dans lesquels un ion n'a pas été détecté. Il existe cependant des méthodologies permettant d'aller récupérer des ions manquants en ciblant les parties du chromatogramme concerné [60].

L'alignement des temps de rétention :

Malgré la qualité des méthodes chromatographiques déployées, il est possible de rencontrer des décalages de temps de rétention entre les différents échantillons (dues à des variations de

température, de pH de pression…) [56]. L'alignement va permettre de corriger ce décalage. Pour se faire, les outils alignent les spectres des échantillons sur un spectre de référence. Le choix de ce spectre servant de référence à l'alignement aura le plus souvent un impact sur ce dernier [60].

L'alignement peut se faire à partir des données brutes en rendant l'axe du temps de rétention commun ou après détection des différents éléments en cherchant à trouver un consensus entre les différents spectres pour positionner l'échelle de temps.

Des outils comme SpecConnect, MetaboliteDetector, MetAlign [58], MZmine [57], TagFinder, XCMS [59], MeltDB ou encore GAVIN permettent de réaliser cet alignement.

Suppression et imputation des valeurs manquantes :

Le terme « valeurs manquantes » désigne ici l'absence de valeurs spécifiques au sein d'un échantillon donné, du fait de la non détection du métabolite. Elles sont dues à l'analyse en elle-même, ou à l'absence du métabolite dans l'échantillon biologique [61, 62]. Pour faire la différence entre celles liées à l'analyse et celles biologiquement expliquées, il est important de prendre en compte les informations relatives aux individus (prise en compte du statut cas/témoins par exemple) pour établir un seuil raisonnable reflétant sa cause. Il est ensuite possible de supprimer les variables pour lesquels le taux de valeurs manquantes est trop important et non lié à la variabilité biologique [60].

Pour limiter le nombre de valeurs manquantes, il est possible de réaliser de l'imputation de données. En métabolomique, les méthodes les plus utilisées sont l'imputation par zéro, la méthode des k plus proches voisins (kNN) et les forêts d'arbres décisionnelles [63].

B. Extraction des données de RMN

La transformation de Fourier :

Cette première étape d'extraction des données de RMN consiste à transformer les données basées sur le temps, en données basées sur la fréquence. La transformation mathématique de Fourier permet ce passage du premier type de données au second en se basant sur le signal sinusoïdal des données.

Correction de la ligne de base :

La ligne de base doit idéalement être une ligne horizontale d'intensité zéro. Cependant, comme elle représente le bruit moyen du spectre, elle est souvent divergente de zéro. Il convient donc de corriger cette divergence en identifiant des zones du spectre où les signaux sont clairement distinguables du bruit pour ensuite projeter ce seuil de bruit au reste du spectre.

Cette étape est le plus souvent réalisée à l'aide du logiciel fourni par le constructeur pour limiter les biais créés par le paramétrage de l'utilisateur.

Correction de phasage :

Afin d'affecter la bonne forme de pics aux données, il est nécessaire de réaliser une étape de phasage. Elle corrige les erreurs instrumentales qui surviennent durant l'acquisition des données brutes spectrales.

Binning ou bucketing :

Les données spectrales sont affectées par de nombreuses variations externes qui peuvent invalider les résultats des analyses statistiques. L'une de ces variations est la position des pics entre les échantillons, même après une étape de calibration de fréquence. En général, cela est dû à des différences de pH entre les échantillons, mais ce phénomène peut également être lié à la composition en ions métalliques, aux échanges chimiques et interactions entre les protéines et les métabolites, à une prise de médication, à l'alimentation ou à des conditions de santé particulières.

Cette étape de binning/bucketing va créer des bin ou buckets qui sont des variables obtenues soit par la moyenne, le maximum ou l'intégrale de l'intensité de tous les points contenus dans la fenêtre correspondant au bin (en générale 0,04ppm de large). Elle va permettre de réduire la taille de la matrice de données en ne conservant que les signaux les plus importants.

Réduction de données :

Durant l'analyse des spectres RMN en analyse non ciblées, la réduction des données est réalisée en supprimant les régions non informatives du spectre comme le signal de l'eau (4.60 – 5.00 ppm) ou les zones contenant exclusivement du bruit (de -0.50 à 0.70 ppm et > 9.00 ppm). Cette étape de réduction est souvent réalisée par binning ou bucketing.

Intégration :

L'intégration des signaux du spectre vise à quantifier de manière absolue ou relative la concentration des métabolites. L'intensité intégré des signaux en ^{1}H-RMN est directement corrélée à la concentration et au nombre d'atomes d'hydrogène. L'intégration en RMN se fait toujours de manière relative, puis un standard interne est nécessaire pour déterminer la concentration absolue.

b. Suppression des variables non pertinentes (blancs, bruit)

La pratique courante pour le nettoyage des données est d'utiliser les blancs pour substituer le bruit créé par les agents de dérivation, les solvants, les contaminants, la colonne… L'identification et l'élimination du bruit permet par la suite, de focaliser les analyses sur les variables réellement pertinentes.

La filtration peut également se faire en se basant sur des critères analytiques en considérant la variance des variables à travers les injections répétées de QC. De la même manière, il est également possible de réaliser une dilution en série des QC, pour ensuite exclure les variables qui ne montrent pas de tendance linéaire [60].

c. Correction de l'effet batch (en MS)

De la variabilité non désirée peut être introduite dans les données par le management des échantillons ou des variabilités instrumentales entre les batchs d'analyses. L'effet batch peut par exemple, résulter d'une dérive dans la mesure de l'abondance des ions du fait de l'encrassement de la source. Il est donc primordial de prendre en compte cet effet batch et de le corriger. Un état des lieux peut ainsi être réalisé *via* une ACP, dans la mesure où l'effet « dérive analytique » peut être visible dès la première composante. La stabilité instrumentale étant nettement plus forte en RMN, il est rarement nécessaire de devoir apporter une telle correction, contrairement aux données MS.

Dans le cas de données de MS ciblées, il est fortement recommandé d'utiliser un standard pour corriger cet effet. En revanche, pour les données non ciblées, il est approprié d'utiliser les QC pour modéliser et normaliser la variation d'intensité d'une injection à l'autre. Ainsi, l'évolution des QC permet de modéliser la tendance à l'aide de méthodes de régression telles que la régression locale de

type LOESS. Il est également possible de le faire en se basant sur la tendance observée sur les intensités de la population plutôt que des QC. Toutefois cette méthode n'est pas privilégiée car la tendance peut être affectée par des effets autres que la dérive analytique, en particulier en cas d'effet confondant entre l'ordre d'injection et un facteur biologique [60].

d. Transformation des données : normalisation et mise à l'échelle

Les données de métabolomique présentes dans les matrices obtenues sont des variables quantitatives continues. La plupart d'entre elles suivent une distribution normale ou log-normale, mais ce n'est pas systématiquement le cas. Les statistiques classiques se basent essentiellement sur la distribution normale des données. Assumer cette dernière permet donc d'accéder à un éventail très large d'outils statistiques. Pour valider cette condition, il est donc possible de transformer les données afin qu'elles observent une distribution normale ou tout du moins une distribution symétrisée.

Cette normalisation peut être nécessaire notamment car les échantillons analysés, selon leur nature biologique, peuvent varier entre eux dépendamment de facteurs tels que l'heure de prélèvement, la santé du sujet, l'alimentation, le genre, l'âge… Les échantillons sanguins (sérum ou plasma) ne nécessitent généralement pas de normalisation du fait du volume sanguin relativement constant et contrôlé. Toutefois, la quantité d'échantillon injecté doit être constante pour limiter les variations d'un échantillon à l'autre, la préparation des échantillons à l'aide d'un robot peut diminuer cette variabilité. En revanche, pour des bio-fluides tels que l'urine, les concentrations en métabolites peuvent grandement différer d'un échantillon à l'autre avec des facteurs pouvant aller de 1 à 20. Ces différences résultent souvent du contrôle de l'homéostasie de l'organisme et sont liés à des volumes urinaires très variables d'un individu à l'autre. La normalisation devient donc indispensable.

Elle peut se faire de façon pré-analytique : au moment de la préparation des échantillons en ajustant les concentrations par des dilutions. Le plus souvent celles-ci se basent sur la concentration en métabolites spécifiques comme la créatinine, la mesure de l'osmolalité ou encore le volume urinaire. Toutefois, elle est plus couramment réalisée après analyse des échantillons. Elle se base le plus souvent sur des facteurs de mise à l'échelle spécifique de chaque échantillon comme la moyenne ou la médiane de l'intensité, en partant du postulat que les signaux au sein d'un échantillon sont échelonnés de manière similaire [60].

D'autre part, les techniques statistiques multivariés sont particulièrement sensibles à l'échelle des variations entre les différentes variables. Le centrage et la réduction de données peuvent alors

être réalisé pour faciliter la comparaison (notamment lorsque les données analysées sont issues de plusieurs jeux de données métabolomique acquis sur plusieurs technologies). Ils permettent d'assigner à une variable une moyenne nulle et une écart-type de 1, les variables ainsi transformer sont donc toute indépendantes de l'unité ou de l'échelle de mesure et deviennent donc comparables de manière relative. Il est également possible de réaliser une transformation logarithmique des données pour rendre symétrique la distribution des données d'intensité. Elle est recommandée lorsque les données présentent des valeurs de variance hétérogènes.

1.3. Des données métabolomique complexes et redondantes

Les données générées par les analyses métabolomiques et lipidomiques sont complexes. Le prétraitement permet de gérer une partie de cette complexité en prenant en charge la réduction du bruit et l'imputation des données manquantes. Malgré tout, ces approches génèrent énormément d'information avec un nombre de variables métaboliques extrêmement important en comparaison du nombre d'échantillons. De plus, les fichiers contiennent un fort niveau d'informations redondantes. En effet, un même métabolite donnera plusieurs signaux après analyse, et plusieurs métabolites peuvent être très fortement corrélés, par exemple si ces derniers appartiennent à une même voie métabolique qui est modulée. Extraire de la connaissance de ces données nécessite donc des outils spécifiques permettant par exemple la gestion de cette redondance [56].

Il est également important de prendre en compte le fait que la métabolomique permet l'étude de tous les métabolites présents dans un échantillon. Or, ces métabolites peuvent être d'origines diverses et il est nécessaire, lors de l'analyse et l'interprétation de ces données, de conserver à l'esprit cette information. En effet, elle peut représenter un véritable biais d'interprétation rendant difficile le fait de distinguer les causes de variations métaboliques observées (pathologie, alimentation, médication…). La gestion de cette problématique passe essentiellement par l'utilisation de tests statistiques et d'étude de corrélation, l'utilisation de bases de données informatives sur l'origine des métabolites ou par la visualisation dans des réseaux métaboliques compartimentés offrant un aperçu de la provenance des molécules. Elle peut également se faire par le couplage des données métabolomiques avec des données omiques d'une autre nature comme des données de transcriptomiques ou de protéomiques.

2. Management et prise en charge des données

2.1. Partager les données avec la communauté scientifique mondiale

La mise à disposition des données de métabolomique et lipidomique est aujourd'hui une vraie problématique collective. Les enjeux sont pluriels : la réutilisation des données générées pour répondre à de nouvelles problématiques biologiques, l'ouverture à la validation de données dans un contexte où les scandales de données falsifiées sont de plus en plus courants, la possibilité d'évaluer la reproductibilité des analyses… Sur les milliers de publications faites dans le domaine de la métabolomique, seules quelques données sont disponibles auprès de la communauté scientifique [64]. Pour pousser l'évolution des pratiques, des institutions comme l'European Bioinformatics Institute (EMBL-EBI) [65] et le National Center for Biotechnology Information (NCBI) tentent de proposer des solutions de stockage des données à long terme (au-delà des limites courantes de 3 à 5 années) [66]. Un exemple dans le domaine de la métabolomique est le répertoire de référence MetaboLights [67] créé par l'EMBL-EBI en 2012 pour regrouper les données issues de tout type d'organisme, et dont l'activité croit constamment avec de plus en plus de données déposées. D'autres bases de données ont également vues le jour dans ce même objectif, comme Metabolomics Workbench [68], Metabolomic Repository Bordeaux (MeRy-B) [69] ou Metabolonote [70]. Une instance créée par l'EMBL-EBI, Metabolome Xchange, offre la possibilité de consulter et mettre en réseau ces bases de données [66].

Dans ce contexte, des questions se posent concernant la nature des données à partager. En effet le format, le vocabulaire utilisé, la précision de la description des protocoles… peuvent être extrêmement différents d'une étude à l'autre.

La notion de données FAIR (Findable, Accessible, Interoperable, Reusable) est récemment apparue pour tenter de proposer un certain nombre de recommandations pour uniformiser les données. Toutefois, la standardisation est un processus qui demande le consentement d'une communauté entière et est parfois difficile à obtenir. La création de données FAIR n'est pas, à l'heure actuelle, standardisée dans tous les domaines de recherche et de nombreux consortia travaillent encore à l'écriture de recommandations dans ce but. Néanmoins Wilkinson *et al* ont publié un article décrivant ce principe et fournissant des recommandations illustrées d'exemples [71].

- **Findable** : Pour être trouvables, les données doivent posséder un identifiant unique et persistant, être décrites par le biais de métadonnées qui doivent être claires et

explicitent incluant l'identifiant des données et enfin être enregistrées ou indexées de manière à être présentes dans un lieu consultable.

- **Accessible** : Les données doivent être accessibles *via* leur identifiant et consultables de manière libre ou contrôlée par des procédure d'identification ou d'autorisation. Si les données ne sont pas accessibles librement, les métadonnées, elles, doivent demeurer consultables.
- **Interoperable** : Les données doivent être interopérables, c'est-à-dire compatibles avec différents types et systèmes informatiques. Elles doivent être écrites dans un format accessible, partagé par une majorité internationale, et dans un langage largement applicable. Elles doivent respecter un vocabulaire suivant le principe FAIR et inclure des références qualifiées aux autres données ou métadonnées.
- **Reusable** : Enfin, les données doivent être réutilisables grâce à des descriptions riches et complètes, vérifiées et bien écrites. Elles doivent être publiées avec une licence d'utilisation claire, une provenance détaillée et un format collant au mieux aux standards établis par la communauté du domaine concerné.

2.2. Uniformisation des formats de données brutes

Les analyses par techniques de métabolomique et lipidomique produisent des données complexes et massives. Bien que les étapes d'extraction et de prétraitement des données soient globalement identiques quel que soit la technique d'analyse utilisée, il existe en réalité de véritables disparités entre les formats de données brutes générées, chaque constructeur utilisant son propre format. Les uniformiser au travers des standards acceptés par la communauté est primordial pour encourager le partage de données à l'échelle de la communauté scientifique internationale [66].

Dans les années 1990, le comité de la IUPAC CPEP (International Union of Pure and Applied Chemistry, Committee on Printed and Electronical Publications) a proposé un format JCAMP de standardisation des données MS et RMN [72]. Il proposait une uniformisation de la liste de pics produite ainsi que des métadonnées associées pour les rendre plus facilement lisibles et interprétables. Il permet une bonne lecture humaine mais n'est pas informatiquement adéquat (plus difficilement lisible par un ordinateur). A cette même période, un format de données appelé netCDF (The Network Common Data Form) a vu le jour [73]. Il offre la possibilité de représenter les données par des vecteurs et des tableaux et non plus par une simple liste d'ions.

Il y a une dizaine d'années, 2 types de format XML ont été développés séparément : mzXML [74] et mzData [75]. Pour tenter d'arriver à un consensus, en 2009, les avantages des 2 formats ont été conciliés en une proposition unique : mzML [76]. Pour que celui-ci soit unanimement utilisé, il doit être supporté par les différents logiciels des fournisseurs d'équipements (ou qu'il soit possible de convertir gratuitement les fichiers constructeurs en mzML) de plus, il doit impérativement exister des librairies open source permettant de parser ces fichiers. Son existence permet de rendre les utilisateurs indépendants des constructeurs pour le partage des données. Malgré tout, de nouveaux formats continuent de voir le jour, souvent basés sur mzML, pour proposer de nouvelles fonctionnalités comme mz5 [77] qui utilise un conteneur permettant des actions plus rapides sur les données.

2.3. Standardisation des données expérimentales et des métadonnées

Posséder un format de fichier standardisé est un premier pas vers le partage, mais la question des métadonnées accompagnant ces données brutes est tout aussi importante. En effet, il est primordial d'avoir accès aux données expérimentales, aux différentes étapes de prétraitements ainsi qu'aux données biologiques ayant servies à la construction de l'étude (état de santé des sujets, mesures cliniques annexes…) [64]. Il n'est pas évident de standardiser ce type d'informations de manière générique tout en conservant la cohérence, l'exactitude et la reproductibilité des données [64]. Au cours du temps, plusieurs recommandations ont été publiées comme les recommandations SMRS (Standard Metabolic Reporting Structure) [78] ou CIMR (Core Information for Metabolomics Reporting) publiées par la Metabolomics Standards Initiative (MSI) [79].

Le format ISA-Tab [80] est un format de plus en plus utilisé qui fut créé en suivant une partie des recommandations citées précédemment. Il comprend un ensemble de tables délimitées à la manière d'une feuille de calcul qui décrivent une question de recherche posée dans une ou plusieurs études comprenant un ensemble d'échantillons et une ou plusieurs analyses. Des outils comme ISAcreator [81] permettent de faciliter la création de ce type de fichier. Toutefois, il est important que des bonnes pratiques soient mise en place dès le début d'un projet pour collecter les informations nécessaires à la construction de ces fichiers sur des supports informatiques communs et non sur des cahiers de laboratoire ou des documents informatiques répartis entre les différents intervenants du projet. Il est donc idéal que ces fichiers ISA soient construits au fur et à mesure des analyses réalisées.

L'une des améliorations à apporter aux formats actuellement existants réside dans le fait d'être capable de faire le lien entre des formats de faible granulométrie (contenant des informations générales) et des formats de forte granulométrie (où le niveau de détail est nettement supérieur).

2.4. Construction de workflow d'analyse

L'uniformatisation des formats de données permet d'assurer l'interopérabilité de ces dernières, mais leur caractère réutilisable dépend de la description des protocoles de production qui est fourni. Il est essentiel d'accompagner les données des informations sur le protocole expérimental suivi, mais également sur le protocole de traitement de données qui lui fait suite et aboutit aux conclusions biologiques d'une étude.

Pour répondre à cela, il est nécessaire de construire des workflows d'analyse *via* des outils publics et partagés. Dans ce but, des instances Galaxy comme Workflow4Metabolomics [82, 83] peuvent être utilisées. Elles regroupent de nombreux outils open source permettant l'analyse des données et offrent la possibilité de partager publiquement le workflow et les données associées.

3. Traitement/analyse des données en métabolomique et lipidomique

L'interprétation des données métabolomiques représente un challenge qui nécessite des approches statistiques robustes et fiables pour en extraire l'information pertinente [56, 84, 85]. L'évolution des concentrations des métabolites peut être très importante ou plus subtile mais seules les statistiques permettent de déterminer si elles sont significativement différentes. Le type d'analyse statistique utilisé dépend énormément de la question de recherche posée [56].

Un exemple d'application la plus courante de la métabolomique dans le domaine des sciences médicales est la découverte de biomarqueurs. Un biomarqueur est une molécule qui doit être mesurable par les cliniciens dans des fluides biologiques facilement accessibles et dont le coût de prélèvement est faible. Sa qualité se mesure également à sa capacité à distinguer les changements subtils dans le phénotype d'un individu (par exemple la transition d'un patient pré-diabétique vers le diabète). Enfin, sa validation dépend de la possibilité de faire le lien avec la sévérité de la pathologie ou ses conséquences comme la mortalité, et ainsi de fournir aux cliniciens la capacité de mieux prendre

en charge le patient. Un biomarqueur seul ne peut quasiment jamais répondre à tous ces critères, notamment à cause de la complexité des variations qui peuvent avoir des origines autres que la pathologie. Pour cette raison, il est donc maintenant reconnu qu'une combinaison de ces derniers plutôt que des molécules individuelles, est plus pertinent et efficace pour décrire un statut ou une maladie [86, 87]. Pour les identifier à partir des matrices d'intensités ou de concentrations, des approches statistiques de classification sont déployées (souvent supervisées). Elles permettent de construire des modèles en s'appuyant sur les différences métaboliques observées entre deux groupes de sujets (*e.g.* étude cas/témoins).

Les méthodes statistiques disponibles sont variées et permettent d'extraire des informations très différentes : aperçu global de la structure des données, identification de facteurs confondants, mise en évidence de corrélations, discrimination cas/témoins… Dans le contexte de l'étude des données métabolomiques, on sépare généralement les méthodes en 2 grandes catégories : les méthodes univariées et les méthodes multivariées. Cette dénomination s'éloigne du sens traditionnellement utilisé en statistique : on appellera méthodes univariées celles qui considèrent les variables métabolomiques (ions, buckets) une à une de façon indépendante, pour les confronter à une ou plusieurs autres variables (par exemple le statut cas/témoins). En revanche, lorsque l'on parle d'approches multivariées, on sous-entend les méthodes qui vont prendre en considération plusieurs variables métabolomiques face à une ou plusieurs variables d'intérêt. Le découpage des méthodes qui sera fait par la suite répond à ce prérequis, mais elles auraient pu être classées différemment d'un point de vue purement statistique (*e.g.* l'ANOVA est une approche qui peut être univariée ou multivariée).

3.1. Méthodes d'analyse univariées

Les méthodes univariées analysent les variables métabolomiques (les ions) indépendamment les unes des autres. Elles présentent l'avantage d'être relativement faciles à réaliser et à interpréter. Toutefois, leur principal désavantage réside dans le fait qu'elles ne prennent pas en considération de potentielles interactions entre les variables. En effet, il peut exister des corrélations entre les ions issus d'un même métabolite, mais également entre les métabolites d'une même voie biologique. De plus, il peut exister de potentiels facteurs confondants tels que le genre, l'alimentation, l'IMC… qui ne sont pas pris en compte [88-90]. Il existe un grand nombre de méthodes univariées, mais la sélection de la méthode doit être dépendante des propriétés statistiques de la distribution des données [84, 91].

Pour évaluer la différence entre 2 (ou plus) groupes, les tests paramétriques comme le test de Student ou l'analyse de variance (ANOVA) sont couramment utilisés après vérification de la normalité des données (par le test de normalité de Kolmogorov-Smirnov ou le test d'homogénéité des variances de Barlett). Si celle-ci ne peut être réalisée, ce sont alors des tests non-paramétriques qui sont appliqués (par le test U de Mann-Whitney ou l'analyse unidirectionnelle de variance de Kruskal-Wallis).

Lorsque l'on réalise plusieurs tests, même si le risque individuel est maitrisé, il est toujours possible d'avoir des variables considérées comme d'intérêt par erreur. Ce risque augmente lorsque l'on répète le nombre de test. Pour pallier à cela, il existe de nombreuses méthodes de correction : la correction de Bonferroni qui va minimiser la probabilité d'avoir un faux positif dans l'ensemble des données, la correction de Benjamini et Hochberg (BH) [92] qui va elle minimiser la proportion de faux positifs sur le nombre total de positifs.

3.2. Méthodes d'analyse des corrélations

Comme évoqué précédemment, les analyses univariées ne permettent pas de prendre en compte les liens existants entre les ions. Pour se faire, et en complément d'une analyse univariée, il est intéressant de calculer les corrélations entre variables quantitatives. Elles permettent de représenter l'intensité des liens existants entre 2 variables X et Y.

Plusieurs types de coefficient de corrélation peuvent être calculés : le coefficient de Pearson met en évidence les ions qui évoluent de la même manière (métabolites impliqués dans la même voie, fragments, adduits…) et de façon linéaire. Le coefficient de Spearman est utilisé pour représenter des corrélations non linéaires. A la différence du précédent, il étudie non pas les valeurs mais les rangs des variables.

Ces corrélations peuvent être utilisées pour former des groupes de variables aux comportements similaires et faciliter les étapes d'interprétation biologiques par exemple.

3.3. Méthodes d'analyses multivariées

Par opposition avec les méthodes univariées, les méthodes multivariées prennent en compte chacune des variables mesurées (ions) de manière simultanée. Par conséquent, elles permettent d'identifier les motifs des relations existantes entre elles. Il existe 2 approches : non supervisées et supervisées.

a. Méthodes non supervisées

Les méthodes statistiques non supervisées permettent le traitement des données sans informations *a priori* comme par exemple le statut cas/témoins de sujets. En absence de ces informations, le jeu de données est considéré comme un ensemble d'objets analogues. Les méthodes non supervisées vont alors tenter de retrouver des motifs de répartition ou de classification des échantillons et de mettre en avant les variables qui sont responsables de ces relations [56].

<u>L'analyse en composante principale (ACP) :</u>

L'ACP est la méthode la plus utilisée en métabolomique [93, 94]. Elle permet de représenter dans l'espace la corrélation entre les variables par projection sur des composantes, réduisant ainsi un système à n variables corrélées de même importance à un système à n composantes principales indépendantes et d'importance décroissante [95]. Elle représente une première phase exploratoire permettant de représenter les données et de focaliser l'attention sur un petit nombre de variables non corrélées qui expliquent une large partie de la variance totale du jeu de données [56].

<u>La Classification Ascendante Hiérarchique (CAH) :</u>

Il est également très courant d'utiliser la Classification Ascendante Hiérarchique. Elle permet de constituer des groupes d'individus « similaires » (de même classe) sur la base de leur description par un ensemble de variables quantitatives. Cette méthode est particulièrement adaptée lorsque le nombre de clusters souhaité n'est pas connu *a priori*. Le clustering/regroupement se fait le plus souvent par agglomération (les individus sont considérés individuellement puis regroupés au fur et à mesure) mais il peut également se faire par division (on part d'un groupe unique subdivisé ensuite

étape par étape). La CAH repose sur des calculs de distance entre les clusters formés. Les résultats sont présentés sous forme de dendrogramme offrant une visualisation de cette distance [56, 96].

b. Méthodes supervisées

Pour ce qui est des approches supervisées, les identifiants des échantillons sont utilisés pour mettre en évidence les variables ou combinaisons de variables qui sont le plus associées au phénotype d'intérêt. Ce sont des approches de statistiques décisionnelles qui permettent par exemple de prédire une variable d'intérêt ou l'appartenance à des classes [97].

Régression logistique :

La régression logistique est une méthode statistique qui permet de mesurer l'association entre la survenue d'un événement à expliquer Y (statut cas/témoins par exemple) et de potentielles variables explicatives Xi). Elle établit pour cela un modèle qui permet d'effectuer des prévisions de la variable Y à l'aide de la connaissance des covariables Xi incluses dans le modèle. Elle mesure la probabilité d'appartenance d'un individu à un groupe ou un autre. Elle est beaucoup utilisée dans les domaines de l'épidémiologie, mais peut tout à fait être utilisée en métabolomique pour construire des modèles sélectionnant un petit nombre de variables prédictives d'un critère.

Afin d'éviter d'être trop optimiste sur la valeur que l'on accorde au modèle on choisit de considérer un taux d'erreur cross validé plutôt que le taux d'erreur obtenu sur les données totales. La cross-validation consiste à réaliser le modèle sur une partie de l'échantillon qui sera donc l'échantillon test, puis de réaliser une validation sur le reste des individus permettant ainsi d'obtenir un taux d'erreurs dit cross-validé.

La régression des moindres carrés partiels (PLS) :

La régression des moindres carrés partiels ou partial least squares regression (PLS) [98] est l'une des méthodes supervisées les plus utilisées en métabolomique notamment pour des analyses de régression (avec une variable d'intérêt quantitative) ou comme classifier binaire (Partial Least Squares – Discriminant Analysis (PLS-DA) avec une variable d'intérêt binaire). Contrairement à l'ACP, les composantes de la PLS ne maximisent pas la variabilité/variance du jeu de données mais la covariance

entre la variable d'intérêt et les données métaboliques. Le loading plot des composants de la PLS correspondent à une mesure de la contribution des variables métabolomiques à la discrimination des différents groupes d'échantillons. L'une des faiblesses de la PLS réside dans le fait que certaines variables métabolomiques non corrélées avec la variable d'intérêt peuvent malgré tout influencer le résultat. Pour pallier à cela, la PLS orthogonale (O-PLS) [99] a été développée. Elle factorise la variance des données en 2 composantes : la première est corrélée avec la variable d'intérêt et la seconde contient les variables non corrélées. A l'heure actuelle différentes comparaison entre PLS et O-PLS ont été faite, mais il est difficile de trancher en faveur de l'un ou l'autre des modèles [100]. Toutefois, l'O-PLS est de plus en plus utilisé dans le domaine de la métabolomique, au détriment de la PLS [98].

Il est nécessaire de valider le modèle de classification pour mesurer sa performance lorsqu'il est appliqué à de nouvelles données. Cette étape est particulièrement requise lorsque le nombre d'échantillons de l'étude est faible pour écarter un potentiel sur-paramétrage (création d'un modèle qui en voulant correspondre au mieux aux données, entraine la création d'erreurs non présentes dans un modèle plus simple). Pour se faire, on réalise un test de permutation. Il détermine la probabilité d'observer des performances égales ou meilleures par pure chance. Le principe est de permuter de multiple fois les classes de groupes d'échantillons (variable d'intérêt) et de calculer les statistiques pour chaque jeu permuté. On évalue ensuite si le modèle réel se révèle meilleur que les modèles permutés.

<u>Les méthodes dites de wrappers :</u>

Les méthodes de Wrapper ("Support Vector Machine" (SVM) et "Random Forests" (RF)) [101] sont utilisées afin d'ordonner les variables et sélectionner les meilleures d'entre elles pour identifier les variables les plus discriminantes et prédictives. Ces deux classifieurs sont parmi les plus utilisés dans la littérature [102-104]. Ils s'appuient sur des mesures de distance. Ils sont définis comme suit :

- La technique SVM est une généralisation des classifieurs linéaires. C'est un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression.
- La technique RF est un apprentissage supervisé qui combine une technique d'agrégation et une technique particulière d'induction d'arbres de décision.

Ces classifieurs issues de SVM ou RF sont plus difficiles à interpréter que ceux de la PLS et de l'O-PLS.

Evaluation de la performance des modèles :

L'évaluation de la performance mesure comment la prédiction faite par le modèle concorde avec la réalité. Il existe de nombreuses mesures pour l'évaluer : la précision (pourcentage de sujets correctement classifiés), la sensibilité (pourcentage de vrais positifs qui sont correctement classifiés) et la spécificité (pourcentage de vrais négatifs qui sont correctement classifiés). L'utilisation de ces mesures pour évaluer la qualité du modèle est fortement influencée par les seuils limites fixés par l'utilisateur.

La courbe ROC est un moyen très répandu pour limiter ce biais. Il s'agit d'une procédure non-paramétrique consistant à comparer la spécificité à la sensibilité en tenant compte d'une limite de décision spécifique. Ces courbes ROC sont souvent résumées par une valeur d'aire sous la courbe (AUC en anglais) sous la courbe ROC. Elle représente la probabilité que le modèle classe un exemple positif aléatoire au-dessus d'un exemple négatif aléatoire. Un modèle dont 100% des prédictions sont fausses aura une AUC de 0 alors qu'un modèle avec 100% de prédictions correctes aura une AUC de 1. On fixe généralement un seuil minimal de 0.7 pour avoir une performance correcte dans le cadre d'un futur biomarqueur clinique [97]. Il existe de nombreux outils permettant de réaliser ce type de courbes ROC et de les interpréter : par exemple les packages R ROCR [105] et pROC [106], l'application web ROCCET [97].

3.4. Enrichissement des données de métabolomique

Il s'agit de la dernière étape des workflows d'analyse de métabolomique. Elle confère aux données une signification clinique et permet d'en exploiter le potentiel maximal.

a. Bases de données

Les bases de données sont des ressources essentielles pour le processus d'enrichissement (identification, interprétation biologique…). Une base de données est définie comme une « collection organisée d'informations ». Il s'agit d'un regroupement de données et d'un système informatique chargé de les stocker. En comparaison, une banque de données se réfère aux données seules et une

librairie est un moyen de représenter cette information. Toutefois, en bio-informatique, ces termes sont souvent utilisés de manière interchangeable [107]. La qualité et le nombre de données stockées est critique pour la performance des algorithmes d'identification et d'annotation. Le nombre de bases de données disponibles est de plus en plus important et le nombre de métabolites qu'elles contiennent est en constante augmentation [108, 109]. En métabolomique on distingue 5 types de bases de données [60] :

- Les banques de composés chimiques
- Les librairies spectrales
- Les bases de données de reconstruction métabolique basées sur le génome
- Les bases de données de connaissances
- Les répertoires de références

Le choix de la base de données à utiliser lors de l'identification des métabolites n'est pas simple, bien qu'elle impacte directement la qualité des résultats et donc de l'interprétation biologique [60].

Les banques de composés chimiques contiennent des informations sur un nombre très vaste de composés chimiques d'origines très diverses (être vivants, chimie industrielle, environnement naturel…). PubChem [110] est la plus grande collection d'information sur des composés chimiques qui soit accessible gratuitement. Elle compte plus de 96 millions de composés d'origines très variées. Sur le même principe, ChemSpider (www.chemspider.com) répertorie 74 millions de composés chimiques et permet des interrogations sur la base des poids moléculaires, masses moyennes ou mono-isotopiques. Il existe également des bases de données « orientées chimie » mais qui focalisent leur contenu sur des thématiques ou organismes précis. ChEBI (Chemical Entities of Biological Interest) [111] est l'une des bases de référence en métabolomique puisqu'elle regroupe les composés d'intérêt biologique. Elle compte près de 41 000 éléments qui sont manuellement annotés pour fournir des données de haute qualité. Elle inclut une ontologie propre qui répertorie la structure chimique et les fonctions biologiques. KNApSAcK (http://kanaya.naist.jp/KNApSAck/) est une autre base de données de composés chimiques qui se focalise sur les métabolites des plantes. Elle regroupe près de 20 000 plantes et 50 000 métabolites. Sa principale caractéristique réside dans le fait qu'elle tente de mettre en relation les espèces métaboliques. Elle contient des informations sur le poids moléculaire, la formule chimique ou encore les spectres MS.

Les bases de données spectrales contiennent des spectres de métabolites de références accompagnés d'informations relatives à l'annotation de ces derniers ainsi que des métadonnées concernant par exemple, l'appareil d'acquisition du spectre de référence. Certaines bases de données

peuvent également contenir des informations annexes sur le métabolite comme : ses interactions avec d'autres métabolites, les espèces biologiques où il est retrouvé, de potentielles implications cliniques chez l'homme, des identifiants pour permettre le passage vers d'autres bases de données… Leur nombre est important, mais, du fait de leur complémentarité, il est recommandé de les utiliser de manière combinée. Bien qu'elles soient généralement de taille importante avec plusieurs dizaines de milliers de données spectrales, elles restent encore relativement incomplètes et sont en constante incrémentation. De plus, les variabilités très fortes de mesure dépendamment des méthodes et appareils d'analyse (particulièrement en LC-MS) ont tendance à favoriser la création de bases de données internes propres aux protocoles utilisés au sein d'un laboratoire [60]. MassBank [112] est un premier exemple de ce type de bases. Elle contient des spectres de masse de haute résolution avec plus de 41 000 spectres MS et 47 000 composés. Elle se base sur la collaboration avec un large réseau de contributeurs, pour fournir les données spectrales et leur annotation en assurant une qualité maximale. Pour se faire, un travail important d'harmonisation du vocabulaire et du format a été fait. Différentes bases ont été développées dans le domaine de la chimie, telle que Pubchem [110], ChemSpider (www.chemspider.com) ou encore la librairie spectrale de NIST (https://chemdata.nist.gov), et possèdent des données pour l'identification des spectres GC/MS. Plus récemment, des banques de métabolites ont été développées. METLIN [113] contient 64 000 structures chimiques différentes d'origines très variées et plus de 59 000 spectres MS de haute résolution. mzCloud (https://www.mzcloud.org/) regroupe, elle, des spectres MS/MS obtenus par Obitrap et organisés en arbres. Il existe encore de nombreuses autres bases spectrales comme Golm Metabolome Database (GDM) [114], The Fiehn library [115].

Les bases de données dites de connaissances regroupent des informations diverses pouvant être issues de plusieurs bases de données différentes. HMDB [41] fait référence dans le domaine de la métabolomique humaine. Elle contient plus de 114 000 métabolites pour lesquels elle fournit des « metabocard » contenant les informations de propriétés chimiques et biochimiques, des données cliniques, différents liens vers d'autres bases d'informations des données de MS et RMN téléchargeables et acquise à l'aide de différents instruments en haute et basse résolution. Lorsqu'aucun spectre expérimental n'est fourni, elle propose un spectre prédit *in silico*. FooDB (www.FooDB.ca) regroupe elle des informations sur les métabolites liés à l'alimentation. Elle associe notamment les métabolites à leurs propriétés alimentaires : goût, couleur, saveur, texture, arôme… Dans le même esprit, PhytoHUB (www.phytoHUB.eu) se focalise sur les composés phytochimiques alimentaires en fournissant des informations sur les principales sources alimentaires où ils sont retrouvés. Elle fournit des informations issues de la littérature, des spectres MS et est manuellement curée à l'échelle internationale.

Les répertoires de référence regroupent eux les données relatives aux différentes études métabolomique publiées. Un exemple de ce type de répertoire est MetaboLights (MTBLS) [67].

Le lien entre toutes ces bases de données peut généralement se faire à l'aide d'identifiants. Certains sont des identifiants propres à une base de données (identifiants ChEBI, HMDB, KEGG…) repris par d'autres bases pour faire le lien, et d'autres sont des identifiants plus universels que l'on retrouve dans la plupart de ces bases comme l'InChi ou le SMILES qui sont relatifs à la structure chimique du composé. Le passage d'une identifiant à l'autre peut parfois se faire à l'aide d'outils de conversion comme BridgeDB (https://www.bridgedb.org/), Chemical Translation Service (CTS) [116].

b. Identification des métabolites

L'identification des métabolites détectés est l'un des challenges majeurs en métabolomique. Elle représente une étape absolument indispensable pour fournir une signification biologique aux variables. Cette identification nécessite souvent plusieurs outils analytiques complémentaires. Après analyse des échantillons biologiques par LC-QToF, de nombreux signaux sont révélés par les analyses statistiques. Une fois que les ions moléculaires sont isolés, la masse exacte obtenue est alors recherchée dans les bases de données. Si cette masse correspond à un métabolite connu alors une identification putative est possible mais si ce n'est pas le cas, une étape de caractérisation, permettant de générer une formule brute chimique à partir de la masse mesurée est envisagée. Alors, quelques échantillons peuvent être ré-analysés par des techniques haute résolution telle que le LTQ-Orbitrap. Dans le cas des données de type LC/MS, les spectres obtenus ne sont pas parfaitement comparables. Dans ce cas, il est alors plus adapté de comparer les informations de masse mesurée et potentiellement de temps de rétention (avec une tolérance plus ou moins importante) à celles présentes dans les bases. Toutefois, par cette méthode, il est fréquent de générer de nombreux faux positifs du fait de l'incertitude de la mesure de la masse et surtout du nombre parfois important d'ions et métabolites partageant la même masse.

En GC/MS, les données étant hautement reproductibles, l'identification se fait par comparaison des spectres à des spectres de référence. Elle se fait après une étape de déconvolution des pics qui, à partir des pics chromatographiques mesurés, extrait des spectres purs en séparant les uns des autres les pics de masses issus de métabolites différents. Les spectres purs peuvent ainsi être comparés à des bases de données de composés référents. Les outils permettant de réaliser cette

déconvolution sont nombreux : AMDIS (http://www.amdis.net), ADAP qui a été inclus dans MZmine [57], eRah [117] ou encore MetaboliteDetector [118].

En RMN, comme en GC/MS, l'identification se fait après une phase de déconvolution permettant d'isoler les pics relatifs à un métabolite unique. Elle consiste également à comparer le spectre pur à une base de données de composés de référence. MetaboHunter [119] est un outil en ligne qui permet ce type d'identification. Toutefois, on voit émergé des approches basées sur un concept de validation de cluster qui permettent de limiter le nombre de faux positifs. Elles prennent également en compte des informations liées à l'intensité des pics et aux corrélations d'intensité entre les échantillons pour renforcer la correspondance des données [120, 121].

4 niveaux d'identification ont été établi par la communauté scientifique en métabolomique, dépendamment du degré de confiance pouvant être accordé à l'annotation [79] :

- Niveau 1 : les composés formellement identifiés par analyse du standard pur de référence dans des conditions analytiques identiques (concordance d'au moins 2 données indépendantes et relatives au composé (ex : RT, m/z, profil isotopique, fragments source, MS/MS))
- Niveau 2 : les composés annotés putativement (sans validation par analyse de standard chimique) mais dont les propriétés physico-chimiques (formule chimique, fragmentation…) et/ou les similarités de spectres avec des librairies prouvent l'identité.
- Niveau 3 : les composés caractérisés par leur classe (basés sur des caractéristiques physico-chimiques de classes de composés, par similarité spectrales à l'échelle des classes de composés).
- Niveau 4 : les composés inconnus.

c. Interprétation des données annotées

Dans le but de comprendre et interpréter au maximum les données préalablement traitées, considérer les voies et les réseaux métaboliques constitue l'étape finale. Ces approches exploitent les propriétés relationnelles existantes entre les données métabolomiques. L'analyse des voies utilise en priorité les connaissances biologiques pour analyser les schémas métaboliques d'un point de vue intégratif. Les réseaux, eux, permettent d'utiliser les informations relatives aux forts degrés de

corrélation présents entre les données pour construire des représentations qui permettent ensuite de caractériser les relations complexes pouvant exister au sein des jeux de données.

L'analyse des voies métaboliques :

Les voies métaboliques sont des ensembles de réactions enzymatiques entre métabolites au sein d'un même processus biologique. Ces composés chimiques sont donc en relation les uns avec les autres et peuvent être connectés entre eux directement ou indirectement par diverses réactions. Les métabolites peuvent réagir pour former de plus grandes molécules (anabolisme) ou dégradés en plus petites molécules (catabolisme).

Lorsque l'on cherche à identifier ces voies au sein des jeux de données métabolomiques, on parle d'analyse d'enrichissement. Il existe de nombreuses bases de données qui regroupent les informations relatives à ces voies pour faciliter cette étape d'enrichissement : Kyoto Encyclopedia of Genes and Genomes (KEGG) [122], Small Molecule Patway Database (SMPDB) [123], EHMN, WikiPathways [124] ou encore MetaCyc [125].

Il existe également de nombreux outils pour visualiser les données enrichies comme Paintomics [126], Vanted [127], Cytoscape [128] ou encore Metaboanalyst [129]. Les métabolites sont ainsi mappés sur des voies métaboliques prédéfinies.

L'analyse de réseaux métaboliques :

Les réseaux métaboliques sont construits à partir des réactions chimiques impliquant les métabolites et les enzymes associées. Ils correspondent à une projection visuelle de ces dernières dans les voies métaboliques. La visualisation de ces informations a pour objectif de fournir une représentation plus ou moins fidèle de la réalité biologique, permettant aux experts (bio-informaticiens et biologistes) de construire puis de valider de nouvelles hypothèses.

Au sein des réseaux, les nœuds représentent les métabolites qui sont liés entre eux par des réactions enzymatiques impliquant de potentiels cofacteurs. La construction de ces derniers prend appui sur le génotypage des organismes vivants. En effet, le séquençage génétique offre la possibilité de reconstituer les réseaux de réactions biochimiques dans de nombreux organismes. Initialement, ce type de représentation a été fait pour des organismes eucaryotes simples [130], mais la démarche a été étendue en 2013, avec notamment la création d'un réseau global chez l'homme [131].

Certains réseaux peuvent être construits sur la base des corrélations puis enrichis par les voies métaboliques. Ils sont alors construits par rapport aux schémas de la relation observée dans les données expérimentales. Dans le réseau qui en résulte, chaque métabolite représente un nœud, mais les liens entre les arrêtes représentent le niveau de corrélation mathématique entre les paires de métabolites. En utilisant des corrélations classiques, on obtient des réseaux encombrés dans lesquels les associations directes et indirectes ne sont pas différenciées [132]. Il est possible de pallier à cela en utilisant des corrélations partielles [133, 134]. La corrélation entre 2 métabolites sera alors également impactée par la corrélation de ces métabolites avec d'autre métabolites, différenciant ainsi les corrélations directes des indirectes. Il existe différents moyens de donner par la suite, des scores à ces liens entre métabolites pour ajouter des éléments supplémentaires d'information (probabilité de voir apparaitre une telle relation à partir des connaissances biologiques…).

Plusieurs outils permettent la construction de ce type de réseaux : MetExplore [135], KEGG Pathway [122], Metscape [136] de Cytoscape ou encore BioCyc [137].

IV. Objectif du travail de recherche

Mon projet de thèse s'inscrit dans une démarche d'épidémiologie des systèmes [138], qui consiste à identifier les contributeurs de pathologies complexes, à de multiples niveaux, ainsi que leurs interactions et ce en utilisant une approche système. Cette dernière combine généralement des données de type omique à des données épidémiologiques observationnelles pour répondre à un objectif fixé. Le but à long terme est de pouvoir utiliser la métabolomique pour reclassifier ces pathologies. Pour cela, il est nécessaire de modéliser les phénotypes des populations à risque, par des approches statistiques/mathématiques.

Dans ce contexte, mon projet s'intéresse à la caractérisation du syndrome métabolique chez la personne âgée par des méthodes de phénotypages multidimensionnelles (métabolomique, lipidomique, phénotypique, nutritionnelle…). Compte tenu de l'état de l'art présenté précédemment, différents sous-objectifs bio-informatiques ont été définis pour parvenir à cette fin. Ils concernent d'une part la prise en charge et la gestion de larges volumes de données complexes (métabolomique/lipidomique et épidémiologique), et d'autre part, des aspects d'extraction de connaissance à partir de ces données multidimensionnelles, dans un processus multi étapes et itératif (**Figure 9**).

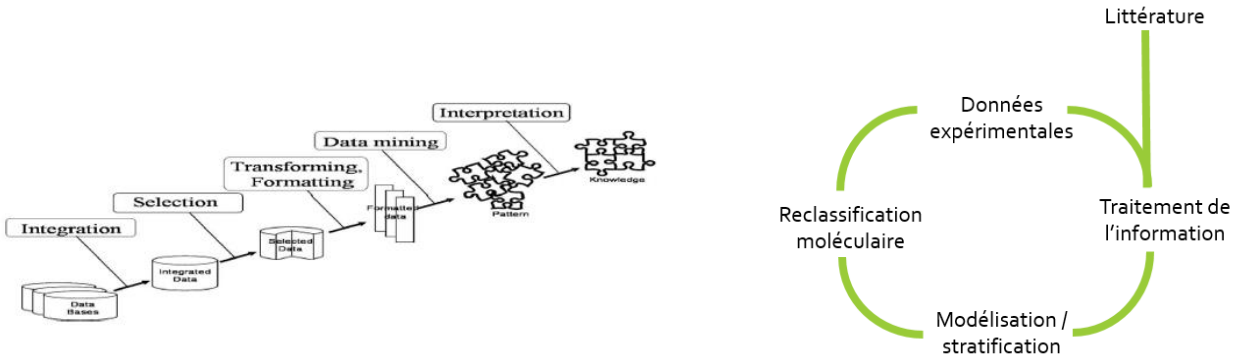


Figure 9 : Schématisation d'une démarche d'extraction de connaissance des bases de données (à droite) et de recherche itérative en biologie des systèmes (à gauche).

Le cadre conceptuel de ce projet est représenté ci-dessous (**Figure 10**). Il sera basé sur :

- **Des données issues de la littérature : l'objectif sera de développer une base de données des biomarqueurs du SMet précédemment identifiés par des approches métabolomiques et lipidomiques et décrits dans la littérature.** Construire une telle base permet de mettre à jour l'état des connaissances à partir des travaux publiés, mais surtout d'élaborer un cahier des charges pour la gestion des métadonnées épidémiologiques, ainsi que de consolider les résultats obtenus dans notre propre étude.
- **Des données expérimentales issues de l'analyse combiné de 6 méthodes d'analyse métabolomique.** Six jeux de données issus d'analyses métabolomiques et lipidomiques non ciblées, seront générés dans le cadre d'une étude cas/témoins sur le syndrome métabolique. Les participants à l'étude seront sélectionnés à partir de la cohorte québécoise NuAge portant sur la nutrition comme déterminant d'un vieillissement réussi.

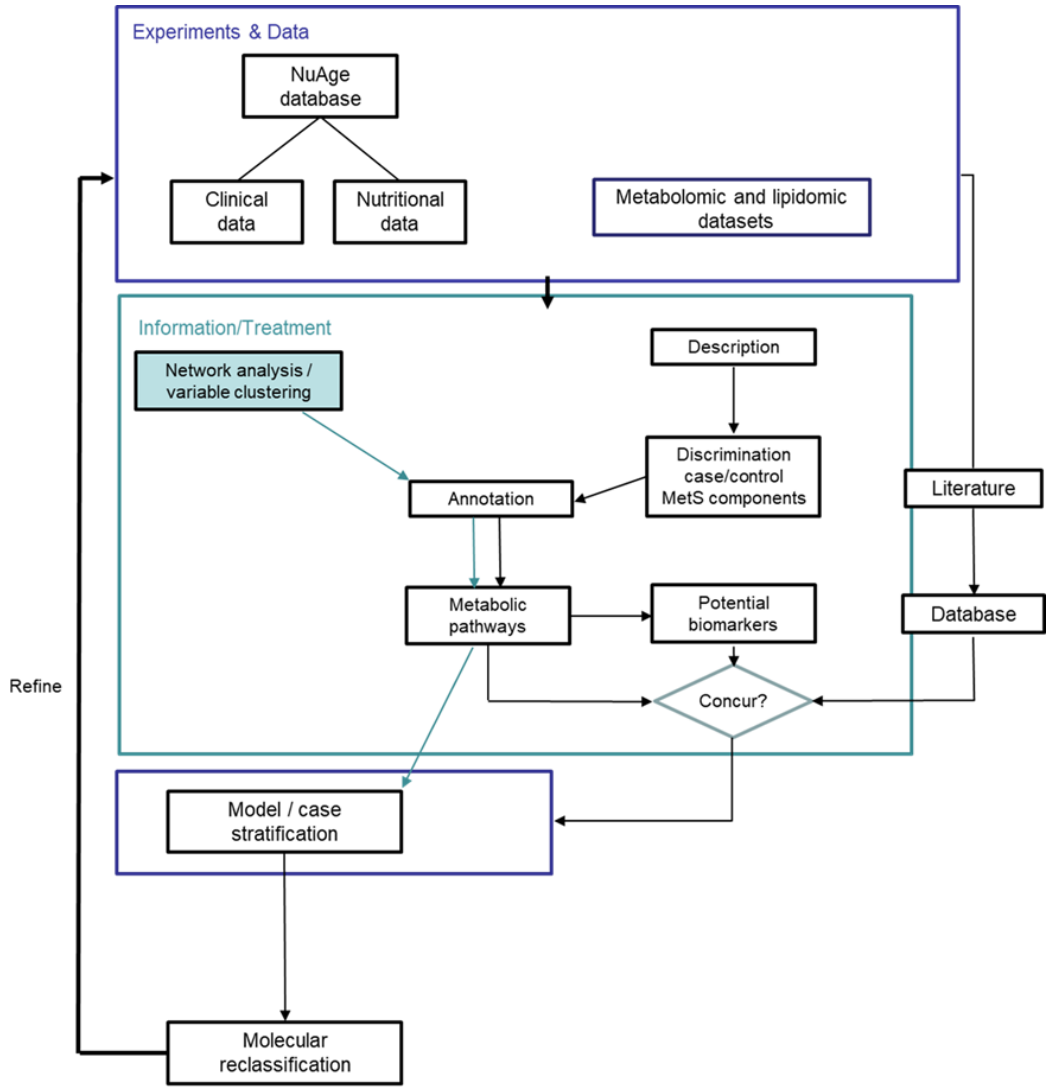


Figure 10 : Schéma conceptuel du projet de recherche

La première étape a pour but de développer des outils de management des données épidémiologiques pour permettre leur curation et leur gestion. Le projet s'appuiera sur la banque de données de la cohorte NuAge regroupant plus de 2 000 variables disponibles à plusieurs temps de mesure, et dont la qualité est validée. Il sera donc indispensable de développer des scripts permettant leur mise au format et la tenue d'un dictionnaire de variables.

La seconde étape consiste à mettre en place un workflow d'extraction de connaissance. Pour permettre l'interprétation des données métabolomiques issues des 6 jeux de données générés, il sera indispensable de réaliser un workflow de traitement uniformisé pour répondre à plusieurs objectifs biologiques. Les objectifs découlant de cette volonté sont pluriels :

- **Développer des outils pour faciliter la filtration des jeux de données métabolomiques.** Les données de métabolomique et lipidomiques génèrent un nombre important de variables qui contiennent de la redondance. Certains outils gèrent une partie de cette redondance mais peu sont réellement focaliser sur sa filtration. Développer un outil assurant cette tâche est donc nécessaire à la mise en place d'une stratégie d'analyse de ces données.
- **Créer un workflow d'analyse statistique permettant la détection d'une signature du SMet.** Les outils permettant l'analyse des données métabolomiques sont nombreux mais il est important de concevoir un workflow pertinent permettant de passer de plusieurs jeux de données métabolomiques contenant plusieurs milliers de variables à une signature de quelques métabolites permettant de prédire au mieux ce phénotype.
- **Enrichir les données métabolomiques en utilisant les réseaux biologiques.** Utiliser les réseaux biologiques reconstruits est indispensable pour permettre l'interprétation des données mais cela génère des problématiques bio-informatiques liées notamment à l'identification et au manque d'interopérabilité des outils et bases de données disponibles.
- **Appliquer la démarche mise en place pour modéliser le spectre phénotypique du SMet et en reclassifier les sous-populations.** Cette étape fait partie des grands enjeux de la médecine de précision qui vise à redéfinir les pathologies complexes de façon plus fine. Elle contribuera au développement d'une prise en charge et d'une prévention des populations à risque, de façon plus personnalisée.

REFERENCES DE L'INTRODUCTION

1. Mottillo S, Filion KB, Genest J, Joseph L, Pilote L, Poirier P, et al. The metabolic syndrome and cardiovascular risk a systematic review and meta-analysis. J Am Coll Cardiol. 2010;56(14):1113-32.
2. Grundy SM, Cleeman JI, Daniels SR, Donato KA, Eckel RH, Franklin BA, et al. Diagnosis and management of the metabolic syndrome: an American Heart Association/National Heart, Lung, and Blood Institute Scientific Statement. Circulation. 2005;112(17):2735-52.
3. Alberti KG, Zimmet PZ. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. Diabet Med. 1998;15(7):539-53.
4. Balkau B, Charles MA. Comment on the provisional report from the WHO consultation. European Group for the Study of Insulin Resistance (EGIR). Diabet Med. 1999;16(5):442-3.
5. National Cholesterol Education Program Expert Panel on Detection E, Treatment of High Blood Cholesterol in A. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. Circulation. 2002;106(25):3143-421.
6. Einhorn D, Reaven GM, Cobin RH, Ford E, Ganda OP, Handelsman Y, et al. American College of Endocrinology position statement on the insulin resistance syndrome. Endocr Pract. 2003;9(3):237-52.
7. Alberti KG, Zimmet P, Shaw J, Group IDFETFC. The metabolic syndrome--a new worldwide definition. Lancet. 2005;366(9491):1059-62.
8. Ford ES, Giles WH. A comparison of the prevalence of the metabolic syndrome using two proposed definitions. Diabetes Care. 2003;26(3):575-81.
9. Liao Y, Kwon S, Shaughnessy S, Wallace P, Hutto A, Jenkins AJ, et al. Critical evaluation of adult treatment panel III criteria in identifying insulin resistance with dyslipidemia. Diabetes Care. 2004;27(4):978-83.
10. Marchesini G, Forlani G, Cerrelli F, Manini R, Natale S, Baraldi L, et al. WHO and ATPIII proposals for the definition of the metabolic syndrome in patients with Type 2 diabetes. Diabet Med. 2004;21(4):383-7.
11. Rodriguez A, Muller DC, Engelhardt M, Andres R. Contribution of impaired glucose tolerance in subjects with the metabolic syndrome: Baltimore Longitudinal Study of Aging. Metabolism. 2005;54(4):542-7.
12. Wyszynski DF, Waterworth DM, Barter PJ, Cohen J, Kesaniemi YA, Mahley RW, et al. Relation between atherogenic dyslipidemia and the Adult Treatment Program-III definition of metabolic syndrome (Genetic Epidemiology of Metabolic Syndrome Project). Am J Cardiol. 2005;95(2):194-8.
13. Tan CE, Ma S, Wai D, Chew SK, Tai ES. Can we apply the National Cholesterol Education Program Adult Treatment Panel definition of the metabolic syndrome to Asians? Diabetes Care. 2004;27(5):1182-6.
14. Enkhmaa B, Shiwaku K, Anuurad E, Nogi A, Kitajima K, Yamasaki M, et al. Prevalence of the metabolic syndrome using the Third Report of the National Cholesterol Educational Program Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (ATP III) and the modified ATP III definitions for Japanese and Mongolians. Clin Chim Acta. 2005;352(1-2):105-13.
15. Hunt KJ, Resendez RG, Williams K, Haffner SM, Stern MP, San Antonio Heart S. National Cholesterol Education Program versus World Health Organization metabolic syndrome in relation to all-cause and cardiovascular mortality in the San Antonio Heart Study. Circulation. 2004;110(10):1251-7.
16. Alberti KG, Eckel RH, Grundy SM, Zimmet PZ, Cleeman JI, Donato KA, et al. Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. Circulation. 2009;120(16):1640-5.

17. Cameron AJ, Shaw JE, Zimmet PZ. The metabolic syndrome: prevalence in worldwide populations. Endocrinol Metab Clin North Am. 2004;33(2):351-75, table of contents.
18. Meigs JB, Wilson PW, Nathan DM, D'Agostino RB, Sr., Williams K, Haffner SM. Prevalence and characteristics of the metabolic syndrome in the San Antonio Heart and Framingham Offspring Studies. Diabetes. 2003;52(8):2160-7.
19. Cameron AJ SJ, Zimmet PZ, Chitson P, Alberti KGGM, Tuomilehto J. Comparison of WHO and NCEP metabolic syndrome definitions over 5 years in Mauritius. Diabetologia. 2003;46:A3068.
20. Laaksonen DE, Lakka HM, Niskanen LK, Kaplan GA, Salonen JT, Lakka TA. Metabolic syndrome and development of diabetes mellitus: application and validation of recently suggested definitions of the metabolic syndrome in a prospective cohort study. Am J Epidemiol. 2002;156(11):1070-7.
21. Ford ES, Giles WH, Dietz WH. Prevalence of the metabolic syndrome among US adults: findings from the third National Health and Nutrition Examination Survey. Jama. 2002;287(3):356-9.
22. Resnick HE, Strong Heart Study I. Metabolic syndrome in American Indians. Diabetes Care. 2002;25(7):1246-7.
23. Aguilar-Salinas CA, Rojas R, Gomez-Perez FJ, Valles V, Rios-Torres JM, Franco A, et al. Analysis of the agreement between the World Health Organization criteria and the National Cholesterol Education Program-III definition of the metabolic syndrome: results from a population-based survey. Diabetes Care. 2003;26(5):1635.
24. Araneta MR, Wingard DL, Barrett-Connor E. Type 2 diabetes and metabolic syndrome in Filipina-American women : a high-risk nonobese population. Diabetes Care. 2002;25(3):494-9.
25. Balkau B, Vernay M, Mhamdi L, Novak M, Arondel D, Vol S, et al. The incidence and persistence of the NCEP (National Cholesterol Education Program) metabolic syndrome. The French D.E.S.I.R. study. Diabetes Metab. 2003;29(5):526-32.
26. Azizi F, Salehi P, Etemadi A, Zahedi-Asl S. Prevalence of metabolic syndrome in an urban population: Tehran Lipid and Glucose Study. Diabetes Res Clin Pract. 2003;61(1):29-37.
27. Onat A, Ceyhan K, Basar O, Erer B, Toprak S, Sansoy V. Metabolic syndrome: major impact on coronary risk in a population with low cholesterol levels--a prospective and cross-sectional evaluation. Atherosclerosis. 2002;165(2):285-92.
28. Al-Lawati JA, Mohammed AJ, Al-Hinai HQ, Jousilahti P. Prevalence of the metabolic syndrome among Omani adults. Diabetes Care. 2003;26(6):1781-5.
29. Deepa R, Shanthirani CS, Premalatha G, Sastry NG, Mohan V. Prevalence of insulin resistance syndrome in a selected south Indian population--the Chennai urban population study-7 [CUPS-7]. Indian J Med Res. 2002;115:118-27.
30. Gupta A, Gupta R, Sarna M, Rastogi S, Gupta VP, Kothari K. Prevalence of diabetes, impaired fasting glucose and insulin resistance syndrome in an urban Indian population. Diabetes Res Clin Pract. 2003;61(1):69-76.
31. Villegas R, Creagh D, Hinchion R, O'Halloran D, Perry IJ. Prevalence and lifestyle determinants of the metabolic syndrome. Ir Med J. 2004;97(10):300-3.
32. Ervin RB. Prevalence of metabolic syndrome among adults 20 years of age and over, by sex, age, race and ethnicity, and body mass index: United States, 2003-2006. Natl Health Stat Report. 2009(13):1-7.
33. Li R, Li W, Lun Z, Zhang H, Sun Z, Kanu JS, et al. Prevalence of metabolic syndrome in Mainland China: a meta-analysis of published studies. BMC Public Health. 2016;16:296.
34. Suliga E, Koziel D, Gluszek S. Prevalence of metabolic syndrome in normal weight individuals. Ann Agric Environ Med. 2016;23(4):631-5.
35. Geetha L, Deepa M, Anjana RM, Mohan V. Prevalence and clinical profile of metabolic obesity and phenotypic obesity in Asian Indians. J Diabetes Sci Technol. 2011;5(2):439-46.
36. de Carvalho Vidigal F, Bressan J, Babio N, Salas-Salvado J. Prevalence of metabolic syndrome in Brazilian adults: a systematic review. BMC Public Health. 2013;13:1198.
37. van Vliet-Ostaptchouk JV, Nuotio ML, Slagter SN, Doiron D, Fischer K, Foco L, et al. The prevalence of metabolic syndrome and metabolically healthy obesity in Europe: a collaborative analysis of ten large cohort studies. BMC Endocr Disord. 2014;14:9.

38. Friend A, Craig L, Turner S. The prevalence of metabolic syndrome in children: a systematic review of the literature. Metab Syndr Relat Disord. 2013;11(2):71-80.
39. Iqbal AZ, Basharat S, Basharat A, Basharat S. Prevalence of the metabolic syndrome and its component abnormalities among school age Pakistani children. J Ayub Med Coll Abbottabad. 2014;26(2):194-9.
40. Kelleher JK. Probing metabolic pathways with isotopic tracers: insights from mammalian metabolic physiology. Metab Eng. 2004;6(1):1-5.
41. Nicholson JK, Lindon JC, Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica. 1999;29(11):1181-9.
42. Cajka T, Fiehn O. Comprehensive analysis of lipids in biological systems by liquid chromatography-mass spectrometry. Trends Analyt Chem. 2014;61:192-206.
43. Strimbu K, Tavel JA. What are biomarkers? Curr Opin HIV AIDS. 2010;5(6):463-6.
44. Dunn WB, Broadhurst DI, Atherton HJ, Goodacre R, Griffin JL. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. Chem Soc Rev. 2011;40(1):387-426.
45. Griffin JL. Metabolic profiles to define the genome: can we hear the phenotypes? Philos Trans R Soc Lond B Biol Sci. 2004;359(1446):857-71.
46. Nicholson JK, Lindon JC. Systems biology: Metabonomics. Nature. 2008;455(7216):1054-6.
47. Bothwell JH, Griffin JL. An introduction to biological nuclear magnetic resonance spectroscopy. Biol Rev Camb Philos Soc. 2011;86(2):493-510.
48. Blümich B, Callaghan PT. Principle of nuclear magnetic resonance microscopy1993.
49. Ward JL, Baker JM, Beale MH. Recent applications of NMR spectroscopy in plant metabolomics. FEBS J. 2007;274(5):1126-31.
50. Brunius C, Shi L, Landberg R. Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction. Metabolomics. 2016;12(11):173.
51. Dunn WB, Wilson ID, Nicholls AW, Broadhurst D. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. Bioanalysis. 2012;4(18):2249-64.
52. Sangster T, Major H, Plumb R, Wilson AJ, Wilson ID. A pragmatic and readily implemented quality control strategy for HPLC-MS and GC-MS-based metabonomic analysis. Analyst. 2006;131(10):1075-8.
53. Hilario M, Kalousis A, Pellegrini C, Muller M. Processing and classification of protein mass spectra. Mass Spectrom Rev. 2006;25(3):409-49.
54. Boccard J, Veuthey JL, Rudaz S. Knowledge discovery in metabolomics: an overview of MS data handling. J Sep Sci. 2010;33(3):290-304.
55. Pluskal T, Castillo S, Villar-Briones A, Oresic M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics. 2010;11:395.
56. Lommen A, Kools HJ. MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware. Metabolomics. 2012;8(4):719-26.
57. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem. 2006;78(3):779-87.
58. Ulaszewska MM, Weinert CH, Trimigno A, Portmann R, Andres Lacueva C, Badertscher R, et al. Nutrimetabolomics: An Integrative Action for Metabolomic Analyses in Human Nutritional Studies. Mol Nutr Food Res. 2019;63(1):e1800384.
59. Gromski PS, Xu Y, Kotze HL, Correa E, Ellis DI, Armitage EG, et al. Influence of missing values substitutes on multivariate analysis of metabolomics data. Metabolites. 2014;4(2):433-52.
60. Wei R, Wang J, Su M, Jia E, Chen S, Chen T, et al. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. Sci Rep. 2018;8(1):663.

61. Kumar N, Hoque MA, Shahjaman M, Islam SM, Mollah MN. Metabolomic Biomarker Identification in Presence of Outliers and Missing Values. Biomed Res Int. 2017;2017:2437608.
62. Rocca-Serra P, Salek RM, Arita M, Correa E, Dayalan S, Gonzalez-Beltran A, et al. Data standards can boost metabolomics research, and if there is a will, there is a way. Metabolomics. 2016;12:14.
63. Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R. The European Bioinformatics Institute in 2016: Data growth and integration. Nucleic Acids Res. 2016;44(D1):D20-6.
64. Haug K, Salek RM, Steinbeck C. Global open data management in metabolomics. Curr Opin Chem Biol. 2017;36:58-63.
65. Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, et al. MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Res. 2013;41(Database issue):D781-6.
66. Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, et al. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Res. 2016;44(D1):D463-70.
67. Ferry-Dumazet H, Gil L, Deborde C, Moing A, Bernillon S, Rolin D, et al. MeRy-B: a web knowledgebase for the storage, visualization, analysis and annotation of plant NMR metabolomic profiles. BMC Plant Biol. 2011;11:104.
68. Ara T, Enomoto M, Arita M, Ikeda C, Kera K, Yamada M, et al. Metabolonote: a wiki-based database for managing hierarchical metadata of metabolome analyses. Front Bioeng Biotechnol. 2015;3:38.
69. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
70. Lampen P, Hillig H, Davies A, Linscheid M. JCAMP-DX for mass-spectrometry. Applied Spectroscopy. 1994.
71. Rew R, Davis G. NetCDF: an interface for scientific data access. IEEE Computer Graphics and Applications. 1990;10(4):76-82.
72. Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, et al. A common open representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol. 2004;22(11):1459-66.
73. Orchard S, Taylor C, Hermjakob H, Zhu W, Julian R, Apweiler R. Current status of proteomic standards development. Expert Rev Proteomics. 2004;1(2):179-83.
74. Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, et al. mzML--a community standard for mass spectrometry data. Mol Cell Proteomics. 2011;10(1):R110 000133.
75. Wilhelm M, Kirchner M, Steen JA, Steen H. mz5: space- and time-efficient storage of mass spectrometry data sets. Mol Cell Proteomics. 2012;11(1):O111 011379.
76. Lindon JC, Nicholson JK, Holmes E, Keun HC, Craig A, Pearce JT, et al. Summary recommendations for standardization and reporting of metabolic analyses. Nat Biotechnol. 2005;23(7):833-8.
77. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics. 2007;3(3):211-21.
78. Members MSIB, Sansone SA, Fan T, Goodacre R, Griffin JL, Hardy NW, et al. The metabolomics standards initiative. Nat Biotechnol. 2007;25(8):846-8.
79. Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, et al. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics. 2010;26(18):2354-6.
80. Giacomoni F, Le Corguille G, Monsoor M, Landi M, Pericard P, Petera M, et al. Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. Bioinformatics. 2015;31(9):1493-5.
81. Guitton Y, Tremblay-Franco M, Le Corguille G, Martin JF, Petera M, Roger-Mele P, et al. Create, run, share, publish, and reference your LC-MS, FIA-MS, GC-MS, and NMR data analysis workflows with

the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics. Int J Biochem Cell Biol. 2017;93:89-101.
82. Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics. 2006;2.
83. Holmes E, Antti H. Chemometric contributions to the evolution of metabonomics: mathematical solutions to characterising and interpreting complex biological NMR spectra. Analyst. 2002;127(12):1549-57.
84. Serkova NJ, Standiford TJ, Stringer KA. The emerging field of quantitative blood metabolomics for biomarker discovery in critical illnesses. Am J Respir Crit Care Med. 2011;184(6):647-55.
85. Morrow DA, de Lemos JA. Benchmarks for the assessment of novel cardiovascular biomarkers. Circulation. 2007;115(8):949-52.
86. Rasmussen L, Savorani F, Larsen T, Dragsted L, Astrup A, Engelsen S. Standardization of factors that influence humain urine metabolomics. Metabolomics. 2011;7.
87. Winnike JH, Busby MG, Watkins PB, O'Connell TM. Effects of a prolonged standardized diet on normalizing the human metabolome. Am J Clin Nutr. 2009;90(6):1496-501.
88. Townsend MK, Clish CB, Kraft P, Wu C, Souza AL, Deik AA, et al. Reproducibility of metabolomic profiles among men and women in 2 large cohort studies. Clin Chem. 2013;59(11):1657-67.
89. Vinaixa M, Samino S, Saez I, Duran J, Guinovart JJ, Yanes O. A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data. Metabolites. 2012;2(4):775-95.
90. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. 1995;57:289-300.
91. Bro R, Smilde AK. Principal component analysis. Analytical methods. 2014(9).
92. World S, Esbensen K, Geladi P. Principal component analysis. Chemometrics and intelligent laboratory systems. 1987;2(1-3):37-52.
93. Hotelling H. Analysis of a complex of statistical variables into principal components. Journal of educational psychology. 1933;24(6):417-41.
94. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998;95(25):14863-8.
95. Xia J, Broadhurst DI, Wilson M, Wishart DS. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. Metabolomics. 2013;9(2):280-99.
96. Fonville JM, Richards SE, RBarton RH, Boulange CL, Ebbels TMD, Nicholson JK, et al. The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping. Journal of chemometrics. 2010;24(11-12):636-349.
97. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). Journal of chemometrics. 2002;16(3):119-28.
98. Tapp H, Kemsley EK. Notes on the practical utility of OPLS. Trends in Analytical Chemistry. 2009;28(11):1322-7.
99. Gromski PS, Muhamadali H, Ellis DI, Xu Y, Correa E, Turner ML, et al. A tutorial review: Metabolomics and partial least squares-discriminant analysis--a marriage of convenience or a shotgun wedding. Anal Chim Acta. 2015;879:10-23.
100. Kim Y, Koo I, Jung BH, Chung BC, Lee D. Multivariate classification of urine metabolome profiles for breast cancer diagnosis. BMC Bioinformatics. 2010;11 Suppl 2:S4.
101. Luts J, Ojeda F, Van de Plas R, De Moor B, Van Huffel S, Suykens JA. A tutorial on support vector machine-based methods for classification problems in chemometrics. Anal Chim Acta. 2010;665(2):129-45.
102. Mahadevan S, Shah SL, Marrie TJ, Slupsky CM. Analysis of metabolomic data using support vector machines. Anal Chem. 2008;80(19):7562-70.
103. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21(20):3940-1.
104. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.

105. Lacroix Z, Critchlow T. Bioinformatics: Managing Scientific Data: Morgan Kaufmann Publishers Inc; 2003.
106. Ellinger JJ, Chylla RA, Ulrich EL, Markley JL. Databases and Software for NMR-Based Metabolomics. Curr Metabolomics. 2013;1(1).
107. Fukushima A, Kusano M. Recent progress in the development of metabolome databases for plant systems biology. Front Plant Sci. 2013;4:73.
108. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019;47(D1):D1102-D9.
109. Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res. 2013;41(Database issue):D456-63.
110. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom. 2010;45(7):703-14.
111. Tautenhahn R, Cho K, Uritboonthai W, Zhu Z, Patti GJ, Siuzdak G. An accelerated workflow for untargeted metabolomics using the METLIN database. Nat Biotechnol. 2012;30(9):826-8.
112. Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmuller E, et al. GMD@CSB.DB: the Golm Metabolome Database. Bioinformatics. 2005;21(8):1635-8.
113. Kind T, Wohlgemuth G, Lee DY, Lu Y, Palazoglu M, Shahbaz S, et al. FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. Anal Chem. 2009;81(24):10038-48.
114. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, et al. HMDB 3.0--The Human Metabolome Database in 2013. Nucleic Acids Res. 2013;41(Database issue):D801-7.
115. Wohlgemuth G, Haldiya PK, Willighagen E, Kind T, Fiehn O. The Chemical Translation Service--a web-based tool to improve standardization of metabolomic reports. Bioinformatics. 2010;26(20):2647-8.
116. Domingo-Almenara X, Brezmes J, Vinaixa M, Samino S, Ramirez N, Ramon-Krauel M, et al. eRah: A Computational Tool Integrating Spectral Deconvolution and Alignment with Quantification and Identification of Metabolites in GC/MS-Based Metabolomics. Anal Chem. 2016;88(19):9821-9.
117. Hiller K, Hangebrauk J, Jäger C, Schreiber K, Schomburg D. MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis. Anal Chem. 2009;81(9):3429-39.
118. Tulpan D, Leger S, Belliveau L, Culf A, Cuperlovic-Culf M. MetaboHunter: an automatic approach for identification of metabolites from 1H-NMR spectra of complex mixtures. BMC Bioinformatics. 2011;12:400.
119. Jacob D, Deborde C, Moing A. An efficient spectra processing method for metabolite identification from 1H-NMR metabolomics data. Anal Bioanal Chem. 2013;405(15):5049-61.
120. Mercier P, Lewis MJ, Chang D, Baker D, Wishart DS. Towards automatic metabolomic profiling of high-resolution one-dimensional proton NMR spectra. J Biomol NMR. 2011;49(3-4):307-23.
121. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40(Database issue):D109-14.
122. Jewison T, Su Y, Disfany FM, Liang Y, Knox C, Maciejewski A, et al. SMPDB 2.0: big improvements to the Small Molecule Pathway Database. Nucleic Acids Res. 2014;42(Database issue):D478-84.
123. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, et al. WikiPathways: building research communities on biological pathways. Nucleic Acids Res. 2012;40(Database issue):D1301-7.
124. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 2008;36(Database issue):D623-31.
125. Garcia-Alcalde F, Garcia-Lopez F, Dopazo J, Conesa A. Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. Bioinformatics. 2011;27(1):137-9.

126. Rohn H, Junker A, Hartmann A, Grafahrend-Belau E, Treutler H, Klapperstuck M, et al. VANTED v2: a framework for systems biology applications. BMC Syst Biol. 2012;6:139.
127. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011;27(3):431-2.
128. Xia J, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS. MetaboAnalyst 2.0--a comprehensive server for metabolomic data analysis. Nucleic Acids Res. 2012;40(Web Server issue):W127-33.
129. Kim J, Kim I, Han SK, Bowie JU, Kim S. Network rewiring is an important mechanism of gene essentiality change. Sci Rep. 2012;2:900.
130. Thiele I, Swainston N, Fleming RM, Hoppe A, Sahoo S, Aurich MK, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol. 2013;31(5):419-25.
131. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
132. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Syst Biol. 2011;5:21.
133. Valcarcel B, Wurtz P, Seich al Basatena NK, Tukiainen T, Kangas AJ, Soininen P, et al. A differential network approach to exploring differences between biological states: an application to prediabetes. PLoS One. 2011;6(9):e24702.
134. Cottret L, Frainay C, Chazalviel M, Cabanettes F, Gloaguen Y, Camenen E, et al. MetExplore: collaborative edition and exploration of metabolic networks. Nucleic Acids Res. 2018;46(W1):W495-W502.
135. Gao J, Tarcea VG, Karnovsky A, Mirel BR, Weymouth TE, Beecher CW, et al. Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. Bioinformatics. 2010;26(7):971-3.
136. Latendresse M, Karp PD. Web-based metabolic network visualization with a zooming user interface. BMC Bioinformatics. 2011;12:176.
137. Dammann O, Gray P, Gressens P, Wolkenhauer O, Leviton A. Systems Epidemiology: What's in a Name? Online J Public Health Inform. 2014;6(3):e198.

CHAPITRE 1 : BASE DE DONNEES DES BIOMARQUEURS EXISTANT DU SMET

I. Conduite d'une revue systématique

L'**objectif** de ce premier travail a été **de développer une base de données des biomarqueurs du SMet précédemment identifiés** par des approches métabolomiques et lipidomiques et décrits dans la littérature.

Pour se faire, nous avons réalisé une revue systématique de la littérature à l'aide d'une méthodologie de référence [139], afin **d'inventorier l'ensemble des métabolites définis comme biomarqueurs du SMet ou de ses composantes chez les sujets adultes**. Cette base de données permettra d'effectuer un état des connaissances déjà publiées dans le cadre de notre problématique, d'établir un cahier des charges pour la future gestion des données épidémiologiques, et enfin de consolider les observations faites au sein des données de métabolomique et lipidomique de notre étude en apportant entre autre une aide dans les tâches d'annotation des métabolites détectés.

Une revue systématique a pour but de résumer de façon exhaustive les connaissances publiées à un temps donné avec précision et fiabilité. Elle rassemble toutes les informations de la littérature qui répondent à des critères préalablement définis et à une question de recherche spécifique. Pour ce faire, des méthodes précises sont utilisées pour minimiser les biais et garantir des informations fiables et pertinentes. Dans certains cas, des méta-analyses, analyses statistiques basées sur les données collectées, permettent de combiner les résultats d'études indépendantes afin de tirer de nouvelles conclusions.

Pour faciliter la conduite de ce type de revue, un protocole de recommandations appelé PRISMA a été établi en 2009 [139, 140]. Ce dernier permet de garantir une approche fiable et méthodique inventoriant de manière exhaustive et pertinente la littérature existante sur un sujet donné. Il décrit 4 phases permettant de conduire au mieux la recherche d'articles et propose une liste de vérifications de 27 éléments pour faciliter la rédaction de la revue.

- La première phase de la recherche consiste à identifier les articles répondant à la question posée *via* notamment la construction d'une requête précise basée sur des mots clés définis. C'est durant cette étape que les bases de données bibliographiques et la littérature grise sont interrogées. La requête doit être construite par l'utilisateur de manière à cibler au maximum la thématique donnée. Si elle est trop large, elle manquera de pertinence et engendrera une analyse longue et fastidieuse des

résultats. Si elle est trop précise, elle pourrait ne pas satisfaire au but premier de la revue systématique : rassembler l'intégralité des connaissances publiées sur la question posée.

Pour aborder au mieux la construction de cette dernière, il est recommandé de commencer par **définir des concepts**. Ces derniers sont issus de la question à laquelle la revue cherche à répondre : quels sont les métabolites déjà cités dans la littérature comme associé au SMet ou à ses composantes ? Dans notre cas, nous avons orienté cette question de façon à optimiser la construction de notre banque de données de biomarqueurs du SMet. Un premier niveau représente l'objet central de notre banque de données : les biomarqueurs. Le second correspond à la méthode de mesure utilisée pour détecter ces biomarqueurs. Notre objectif étant d'enrichir l'annotation des données métabolomiques, nous nous sommes focalisés sur les méthodes de détection faisant appel à la MS et à la RMN. Enfin le dernier concept représente notre cas d'étude, l'état de santé auquel nous nous intéressons et que les biomarqueurs caractérisent : le syndrome métabolique ainsi que toutes les composantes qui lui sont associées.

Il faut ensuite déterminer **les mots clés liés à ces concepts**, ce qui peut représenter une difficulté plus ou moins grande selon le domaine. Pour faciliter leur recherche, des bases de données de synonymes ou de classification comme MeSH (Medical Subject Headings), EMTREE ou HeTOP (Health Terminology/Ontology Portal), qui sont ici, propres au domaine de la biologie médicale, peuvent être utilisées. Pour le concept « biomarqueurs », nous avons associé les termes « biomarkers », « metabolites » ou encore « lipids » avec « risk stratification » puisque, associé à la notion de métabolomique/lipidomique, il permet d'inclure des études visant à construire des modèles prédictifs sur la base de biomarqueurs. Le second concept regroupe des mots clés définissant les méthodes d'analyse utilisées en métabolomique/lipidomique. Comme précisé dans le chapitre d'introduction, elles peuvent correspondre à des associations entre des techniques de séparation et des analyseurs de type variable (en MS ou en MS/MS) ou à l'utilisation de la RMN. Les notions de signature ou d'empreinte métabolique sont incluses dans ce concept car elles résultent des analyses métabolomiques. Le dernier concept portant sur la pathologie étudiée contient comme mots clés le SMet et ses synonymes (comme syndrome X), mais également des termes associés au diabète de type 2 ainsi qu'au pré-diabète du fait qu'il représente la trajectoire qui suit le développement du SMet et qu'il est inclut dans certaines de ces définitions. Pour lier les mots clés ainsi que les concepts les uns aux autres, il convient d'utiliser des opérateurs booléens comme « ET » et « OU ».

Finalement, le choix des bases de données dans lesquelles sera effectuée la requête est primordial. Plus la couverture est large, meilleure sera la qualité de la revue. La requête établie précédemment peut nécessiter d'être adaptée en fonction de la base de données, notamment concernant l'écriture

et la reconnaissance de certains termes qui peuvent n'être pris en charge que dans certaines bases (par exemple les termes MeSH au sein de PubMed). Pour notre revue, nous avons interrogé 5 bases de données dont le domaine de couverture était la biologie ou la santé : MEDLINE, EMBASE, EMB Review, CINHAL Complete et PubMed.

- La seconde phase consiste à parcourir les publications (généralement uniquement le titre et le résumé), après suppression des doublons, pour éliminer celles qui ne répondent pas à la question posée. Cette étape, ainsi que les suivantes, nécessitent une lecture croisée par 2 personnes ou plus, afin de garantir la pertinence de cette première filtration. Elle nécessite de déterminer des critères d'exclusion.

Dans le cadre de notre revue, la requête étant large, nous avons obtenu une liste de plus de 8 000 publications.

Nous avons réalisé cette étape à l'aide du gestionnaire de bibliographie EndNote® qui offre la possibilité d'éliminer les doublons issus des différentes bases de données, mais également d'effectuer différentes recherches de mots clé à travers les titres et les résumés des articles. C'est notamment grâce à cette dernière possibilité que la sélection a pu être raffinée. L'utilisation de mots clés a permis le regroupement de certains articles qui ont ensuite été éliminés sur la base des critères de recherche. Par exemple, les titres contenant le mot « traitement » ont été recherchés pour identifier les articles faisant référence à des études sur l'effet d'un traitement, ils ont ensuite été éliminés des résultats après validation de la cohérence par les relecteurs.

Les critères d'exclusion fixés sont les suivants : langue de publication, revues et livres, articles ne portant pas sur modèle humain, articles ne portant pas sur des études de cohortes, études d'intervention, études sur les enfants, les adolescents ou les femmes enceintes, études sur des sujets non caucasiens, les articles dont le sujet central n'est ni le SMet ni l'une de ces composantes et enfin les articles se rapportant à des études de génomiques/transcriptomiques et/ou protéomiques.

- La troisième phase correspond à la détermination de l'éligibilité des articles, c'est-à-dire déterminer si leur contenu détaillé répond à des critères précis pour permettre leur inclusion dans l'étude suite à leur lecture. Cette étape nécessite donc la lecture complète des documents pour en extraire les informations nécessaires à la prise de décision. Elle est réalisée par 3 personnes pour augmenter la fiabilité de la décision. Des critères d'inclusion sont également déterminés tel que la durée de suivi des sujets ou les caractéristiques de populations disponibles (*e.g.* âge, critères su SMet).

Les études, pour être incluses doivent être réalisées sur un minimum de 20 sujets par groupe et les données de description au regard du SMet et de ses composantes doivent être disponibles. En effet,

la notion de biomarqueur est associée à un état de santé donné, dans une population donnée, il est donc primordial que ces informations soient correctement décrites pour pouvoir utiliser et compiler les résultats.

- La dernière phase correspond à l'extraction des données : elle permet de collecter les informations issues des publications qui seront incluses dans la revue. Le rôle des auteurs est de regrouper et organiser les compétences issues de chaque publication pour fournir à la communauté un aperçu détaillé des connaissances.

Nous avons donc extrait de chaque publication les conclusions faites par les auteurs ainsi que les métabolites cités comme liés au SMet ou à ses composantes. Nous avons distingué les métabolites discriminant du statut SMet, des métabolites pour lesquels une simple corrélation statistique a été faite avec le SMet ou l'une de ses composantes.

Nous avons également extrait les informations descriptives de la population : le nombre de sujets inclus dans l'étude, l'objectif initial de cette étude (cas/contrôle sur le SMet, trajectoire du diabète…), les valeurs cliniques mesurées pour chacun des 5 critères du SMet (tension artérielle, dysglycémie, tour de taille, dyslipidémie (HDL-cholestérolémie et triglycéridémie)) afin de pouvoir étudier la proportion de chaque critère au sein des différentes études, et les éléments pouvant avoir une incidence sur le SMet, comme le sexe (du fait des différences métaboliques fortes existantes entre les hommes et les femmes) ou l'âge. Nous avons également extrait les informations relatives au type d'échantillons (origine, prélèvement à jeun ou non), à la méthode d'analysé utilisées (LC/MS, GC/MS, RMN…) ainsi qu'aux méthodes statistiques employées.

Nous avons ensuite organisé toutes ces informations sous forme de tableaux pour faciliter la lecture des résultats et la constitution de la banque de données « métabolite centrée ». Nous avons également réalisé des représentations en diagramme de Venn des métabolites significativement corrélés aux composantes du SMet (**Figure 11**), ainsi que des métabolites associés au diabète de type 2 prévalent (déjà existant) ou incident (nouveaux cas) et/ou à la composante glycémique du SMet (**Figure 12**). Ces diagrammes ont permis de visualiser dans quel type de population les biomarqueurs du SMet (ou du diabète de type 2) ont été observés.

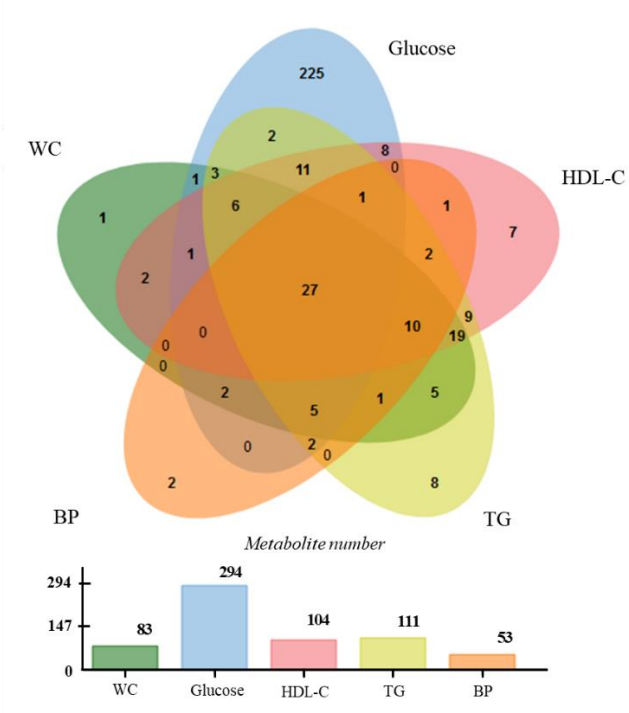


Figure 11 : Diagramme de Venn représentant le nombre de métabolites significativement corrélées avec une ou plusieurs composantes du SMet
WC = tour de taille; BP = pression sanguine; TG = triglycérides; HDL-C = HDL-cholestérol.

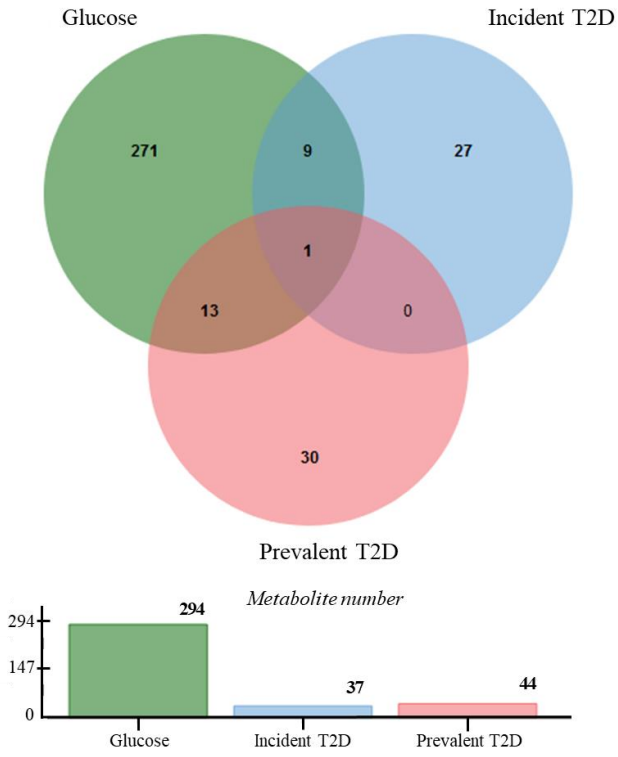


Figure 12 : Diagramme de Venn représentant le nombre de métabolites significativement modulés en cas de diabète de type 2 prévalent ou incident ainsi que le nombre de métabolites associés à la composante glycémique du SMet.

Publication n°1

Monnerie S, Comte B, Ziegler D, Morais JA, Pujos-Guillot E, Gaudreau P.

Metabolomic and lipidomic signatures of metabolic syndrome and its physiological components in aging: a systematic review.

Article en cours de révision dans le journal « Scientific reports »

METABOLOMIC AND LIPIDOMIC SIGNATURES OF METABOLIC SYNDROME AND ITS PHYSIOLOGICAL COMPONENTS IN ADULTS: A SYSTEMATIC REVIEW

Stéphanie Monnerie[1], Blandine Comte[1], Daniela Ziegler[2], José A. Morais[3], Estelle Pujos-Guillot[1,4*], Pierrette Gaudreau[5,6].

[1]Université Clermont Auvergne, INRA, UNH, Mapping, F-63000 Clermont Ferrand, France; [2]Centre Hospitalier de l'Université de Montréal (CHUM), Montréal, Canada; [3]Département de Gériatrie, Université McGill, Montréal, Canada; [4]Université Clermont Auvergne, INRA, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, F-63000 Clermont-Ferrand, France; [5]Centre de Recherche du CHUM; [6]Département de médecine, Université de Montréal, Montréal, Canada.

*Corresponding author: Estelle Pujos-Guillot

ABSTRACT

The aim of this work was to conduct a systematic review of human studies on metabolite/lipid biomarkers of metabolic syndrome (MetS) and its components, and provide recommendations for future studies. The search was performed in MEDLINE, EMBASE, EMB Review, CINHAL Complete, PubMed, and on grey literature, for population studies identifying MetS biomarkers from metabolomics/lipidomics. Extracted data included population, design, number of subjects, sex/gender, clinical characteristics and main outcome. Data were collected regarding biological samples, analytical methods, and statistics. Metabolites were compiled by biochemical families including listings of their significant modulations. Finally, results from the different studies were compared. The search yielded 31 eligible studies (2005-2019). A first category of articles identified prevalent and incident MetS biomarkers using mainly targeted metabolomics. Even though the population characteristics were quite homogeneous, results were difficult to compare in terms of modulated metabolites because of the lack of methodological standardization. A second category, focusing on MetS components, allowed comparing more than 300 metabolites, mainly associated with the glycemic component. Finally, this review included also publications studying type 2 diabetes as a whole set of metabolic risks,, raising the interest of reporting metabolomics/lipidomics signatures to reflect the metabolic phenotypic spectrum in systems approaches.

INTRODUCTION

Metabolic syndrome (MetS) is a complex health condition responsible for the concurrence of several metabolic abnormalities and cardiovascular disturbances. Despite a lack of unified definition among health organizations (*e.g.* National Cholesterol Education Program (NCEP), International Diabetes Federation (IDF), World Health Organization (WHO)), MetS comprises glucose metabolism dysregulation due to insulin resistance, central obesity, dyslipidemia, including increased blood triglycerides (TG) and decreased high-density lipoprotein cholesterol (HDL-C), and hypertension [16, 141-143]. This combination of risk factors favor adverse outcomes such as type 2 diabetes (T2D) and cardiovascular disease (CVD) and increased mortality rate by approximately 1.5-fold [1]. It is generally accepted that the prevalence of MetS is on the rise in accordance with increasing body mass index (BMI) and aging of the population [32]. Because several clinical definitions co-exist, the true prevalence of MetS is difficult to establish. In spite of this, U.S. surveys indicate that one-third of adults [144-146], including young adults [147] have MetS. Moreover, by the age of 60, the prevalence reaches 42% compared to 7% for young adults [148]. Europe has not been spared from such epidemic, with also a sharp increase of MetS among older adults [37]. Therefore, it is now accepted that MetS represents a global public health concern with a worldwide prevalence ranging from 10 to 84%, depending on the ethnicity, age and sex/gender [149, 150].

MetS is recognized as a progressive pathophysiological state, being part of the trajectory leading to pre-diabetes, T2D and CVD [151]. In fact, MetS is not only a precursor but also a predictor of T2D development [152-155]. Risks of adverse health outcomes increase substantially with accumulation of MetS clinical components and deleterious environmental factors (*e.g.* inactivity, Western-type diet). In this context, it is important to better characterize intermediate phenotypes associated with metabolic abnormalities. Biomarkers are considered useful to disentangle the exposure-disease relationships in chronic metabolic disorders and provide sensitive tools for a better identification and stratification of high-risk individuals [156]. Timely identification of MetS physiological disturbances should allow pinpointing individuals at highest risk to develop T2D, CVD, and multi-organ damage. Moreover, studies of their trajectories should provide insights into key periods for lifestyle intervention, risk factor management, and robustness of pharmacological treatment.

Over the last few years, omics technologies allowed obtaining an integrated view of biological systems, bridging the genotype-to-phenotype gap using a systems biology approach to better define the phenotype. In chronic metabolic diseases, the phenotype is complex and dynamic, because of the occurrence of multiple interactions among genetic and environmental factors [157]. In this setting, metabolomics, introduced by Nicholson *et al*. 1999 [43], aiming at measuring all small molecules/metabolites present in a biological system and accessible to analysis, represents a powerful phenotyping tool. Indeed, it provides metabolic profiles that represent an integrated view of metabolism because it allows a sensitive detection of molecular changes over time, resulting from the interaction between intrinsic and extrinsic factors [158]. Metabolites, used as single targets or in combination within

a comprehensive signature, are thus promising biomarkers to reveal early metabolic dysfunctions, when conventional clinical markers have a limited ability for risk assessment and stratification. Metabolomics has therefore been widely applied for metabolic disease diagnosis and candidate biomarker discovery as well as pathophysiological exploration of underlying mechanisms, and prognosis and prediction [159, 160].

Because the human metabolome is complex (*e.g.* large concentration ranges, high number of metabolites, chemical diversity), different analytical strategies and methods have been developed. The approach can be untargeted, as a data driven approach dedicated to biomarker discovery, or targeted when it is focused on the detection and quantification of specific classes of compounds, or subsets of known metabolic pathways [161]. For example, lipidomics has been described as a subsection of metabolomics dedicated to lipid analysis, even if there is a continuum of polarity between lipophilic and hydrophilic metabolites [162]. To cover this wide diversity of metabolites present in a given biological sample, diverse analytical platforms are used. Mass Spectrometry (MS) coupled with gas or liquid chromatography (GC- or LC-, respectively) and Nuclear Magnetic Resonance (NMR) Spectroscopy are the two main analytical techniques used. NMR is non-destructive, rapid, and highly robust, which is convenient for a rapid screening of biological sample [163] but suffers from limited sensitivity (less than 100 metabolites in most biological samples by current methods). Advances in MS and its hyphenated techniques, particularly the increase of their respective resolving and separation powers, significantly impacted metabolomics research allowing for higher sensitivity and broader metabolome coverage [164]. Nonetheless, these MS-based techniques still lack standardization and throughput. In addition, technical factors (time of sampling, sample type, stability) have to be considered for metabolome investigations and the results of different studies need to be compared. Interestingly, certified commercial targeted LC-MS based assays or platforms became available during the last years (*e.g.* Biocrates, Metabolon).

Considering the diversity of experimental design and analytical methods to characterize the multifaceted physiopathology of MetS, it is necessary to rigorously analyse the scientific literature to answer the general question “Do metabolomic/lipidomic profiles of MetS and/or its clinical components allow distinguishing from healthy individuals and do they expand the current knowledge about MetS phenotypes?” The aim of this work was therefore to conduct a systematic review of human studies on metabolite/lipid markers of MetS and its individual clinical components and provide recommendations for improving the experimental design and result reports of MetS biomarkers.

RESULTS

Search results

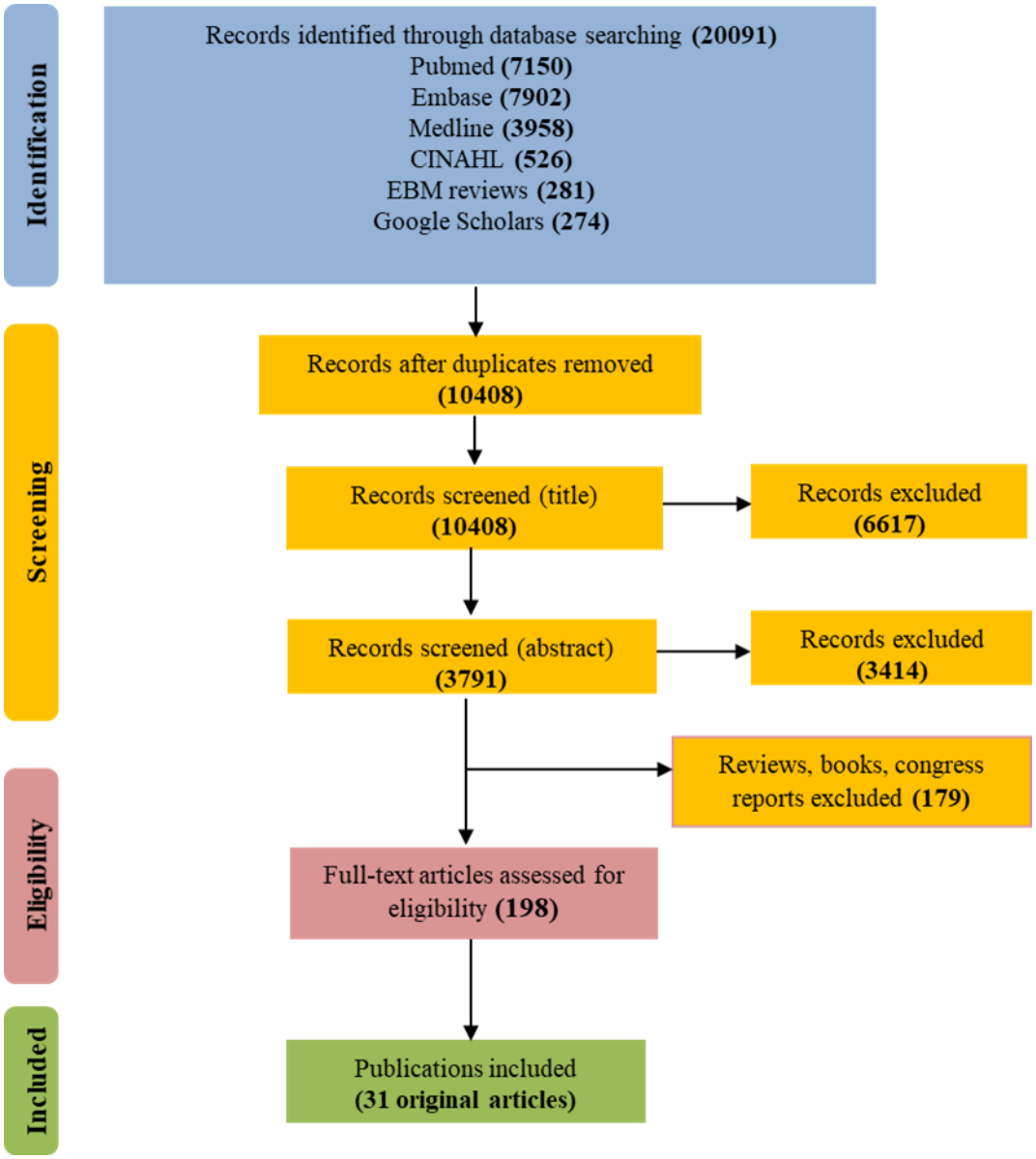


Figure 1: Flow diagram of reviewed citations modified from PRISMA flow diagram 2009 [139].

The primary search identified 20,091 records from five databases (Fig. 1). After removing duplicates, 10,408 original publications were screened for titles and abstracts. Following title screening, 6617 of them were discarded and an additional 3414 were excluded after reading the abstracts, in accordance with the identified inclusion/exclusion criteria. Among the 377 remaining articles, 97 were excluded because they were reviews and 82 more because they were books, congress reports and proceedings. Finally, the full content of 198 original articles was read and analysed for eligibility by three independent authors, and 31 of them were retained for the present review.

These articles were published between 2005 and 2019, 30 out of 31 were published over the last 7 years, and 19 since 2016. Three categories of articles were identified, depending on the main outcome and study design, and the same article could be classified in more than one category depending on the approaches. Twelve of them were case/control studies on MetS with the objective of identifying prevalent or incident MetS biomarkers (Table 1). Sixteen were focussing on MetS components and studied the correlations/associations between identified metabolites and MetS criteria (Table 2). Finally, four articles identified prevalent T2D biomarkers (Table 3) and four others were prospective studies of associations between metabolites and incident T2D (Table 4).

MetS biomarkers: results from case/control studies

Sixteen articles were included in the first section of the systematic review, on prevalent MetS biomarkers identified in case/control studies (Table 1). They provide population characteristics and statistical parameters with cofactors. Most of the studies were performed in populations aged between 40 and 60 years. Generally, MetS cases exhibited three criteria: a high WC combined with two of the following, high glucose, high TG or hypertension. They were compared with healthy controls. These sixteen articles described 409 different modulated metabolites in blood or urine, each one discriminating MetS patients and controls from a single studied population for the discovery. Ninety of them are amino acids and derivatives, 90 others, di- and tri-glycerides, and around 70 glycerophospholipids. No replication/validation was performed and these biomarkers were mostly identified using targeted MS metabolomics. The metabolites are presented in Supplemental Table 1a with associated references and classified by metabolite families and direction of variation (*i.e.* positive or negative). A total of twenty-four different metabolites families were found to be involved. The main classes are amino acids and derivatives, carbohydrates and derivatives, glycolysis related metabolites, glycerophospholipids, glycerolipids, sphingolipids, fatty acids, cholesterol and oxysterols, steroids, and peptides.

Two other publications described biomarkers of incident MetS in prospective studies including only men. Nineteen metabolites were identified as belonging to the following chemical families: amino acids and derivatives, carbohydrates and derivatives, carnitines, fatty acids and derivatives, glycerophospholipids, peptides and steroids (Supplemental Table 1b). It is noteworthy that seven among these metabolites were already described as markers of prevalent MetS, namely alanine, glutamic acid, phenylalanine, tyrosine, oleic acid, total and free testosterone.

Metabolites associated with MetS clinical components

Sixteen articles were included in the second section of the systematic review and are presented in Table 2. In these publications, the main outcome was not only MetS, but also associated components (*e.g.* obesity, cardio-metabolic risk). Each study correlated metabolites and MetS criteria using different statistical approaches (Spearman/Pearson correlations or linear regression). In terms of clinical

characteristics, data were generally provided regarding the whole studied populations and therefore are quite heterogeneous within the age range of 36 to 69 years and BMI of 25 to 33 kg/m^2.

Over three hundred metabolites (361) were described as being significantly correlated with one or several MetS criteria, independently (Supplemental Table 2), including 22 metabolite families. Twenty seven of them are correlated with all MetS components (Fig. 2): alanine, choline, glutamate, glutamine, glutamine/glutamate ratio, glycine, isoleucine, L-carnitine, leucine, methionine, phenylalanine, proline, tyrosine, valine, glycerol, 9 TGs , testosterone, alpha-hydroxybutyric acid, and Cer(20:3). Of interest, nineteen of them have already been reported to be prevalent MetS biomarkers in case/control studies (alanine, L-carnitine, choline, glutamate, glutamine, isoleucine, leucine, phenylalanine, proline, tyrosine, valine, and 8 TGs).

Around 10% of the metabolites were common to three of the MetS criteria (all combinations of them). More specifically, about 60% of the identified metabolites showed levels correlated with HDL-C, TG, and glycemia criteria. In addition, this review highlights that some metabolite levels were found to be specifically correlated to each of the MetS criteria (Supplemental Table 3). Seventeen of them were previously described as prevalent MetS biomarkers: 3-hydroxybutyrate, nitric oxides, 5 phospholipids, and 10 TGs.

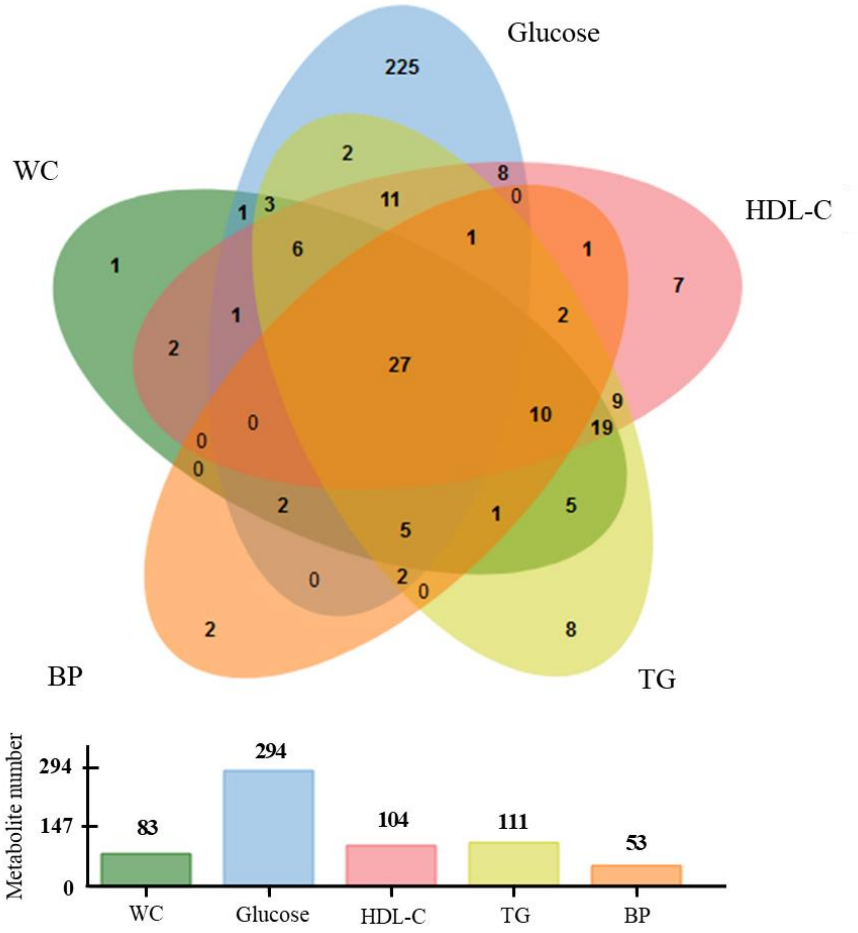


Figure 2: Venn diagram showing the number of metabolites significantly correlated with MetS components, together with respective histogram representing the number of significant metabolites for each clinical MetS components. WC = waist circumference; BP = blood pressure; TG = triglycerides; HDL-C = high-density lipoprotein cholesterol.

The glycemic component: towards T2D

Considering that MetS can lead to T2D and was included in some criteria definition (IDF), we also analyzed articles highlighting an association between prevalent and incident T2D and metabolite dysregulations. A large body of literature was found regarding the investigation of T2D using metabolomics. However, we only selected publications including available clinical data about MetS criteria. Four original articles were selected with case/control design aiming at identifying prevalent T2D markers (Table 3). Four other prospective studies have assessed metabolites associated with incident T2D (Table 4). All these studies have included hypertensive older adults (48 to 70 years) with some cases having a BMI around 30 compared to controls (BMI around 27). Fifty-two metabolites were

positively modulated with prevalent T2D from 10 different metabolite families (Supplemental Table 4), identified using targeted MS approaches, predominantly, performed on plasma or serum. The incident markers of T2D were more frequently investigated using untargeted MS approaches and were validated within a replication study in different cohorts, revealing 39 modulated blood metabolites (Supplemental Table 5) from 11 chemical families. Of particular interest, three studies used multivariate statistical analyses to define a metabolic signature of T2D-related early metabolic disturbances. Among the individual markers, only isoleucine was already reported as a marker of prevalent T2D.

The prevalent and incident T2D markers were then compared to those previously described as being associated with the glucose component (Fig. 3). Thirteen metabolites (mostly amino acids, total hexoses and lipid derivatives) are shared by the prevalent T2D and the glucose component whereas 9 metabolites (mostly amino acids) are shared by the incident T2D and the glucose component of MetS. Of particular interest, the amino acid isoleucine is the only shared metabolite by all these glycemic states.

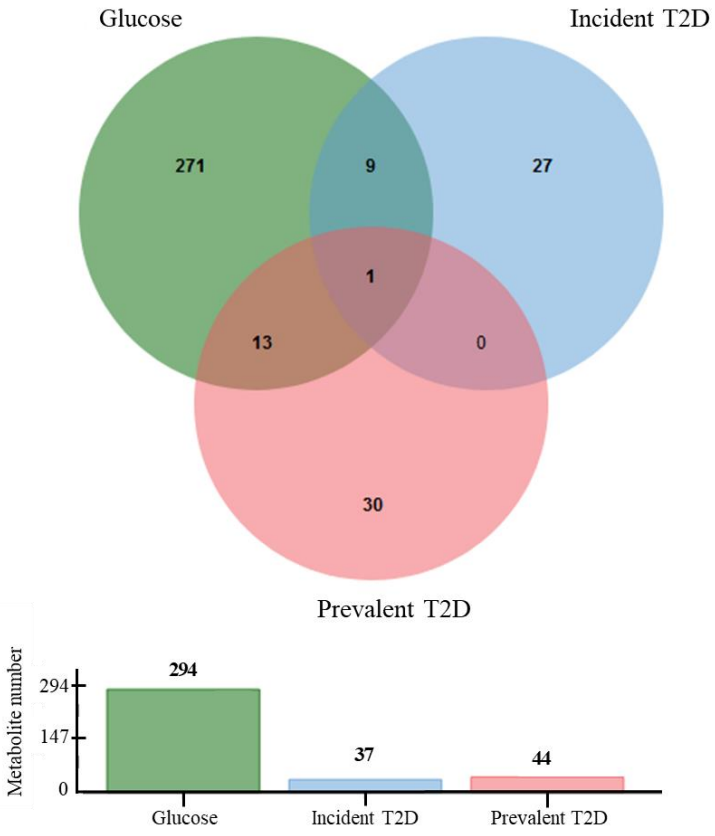


Figure 3: Venn diagram showing the numbers of metabolites significantly modulated with prevalent and incident T2D and the number of metabolites associated with glycemia, together with respective histogram representing the number of significant metabolites for each outcome.

DISCUSSION

MetS biomarkers: results from case/control studies

In the present systematic review, a first category of publications identified prevalent MetS biomarkers in adults using mainly targeted metabolomics approaches. Even if the population characteristics were clearly presented and quite homogeneous, results were difficult to compare in terms of modulated metabolites because of the limited metabolome detected by each single targeted analytical method. However, if the same samples were subjected to different complementary analyses or techniques, some additional metabolites would have been detected. This point is highlighted in two included recent publications that performed semi-targeted approaches that allowed identifying hundreds of modulated metabolites [165, 166]. This comparison of throughput and coverage in targeted and non-targeted metabolomics have extensively been discussed in the literature, showing the interest of using multi-platform approaches [167-169] to obtain a broader scope of the metabolome related to specific phenotypes. However, due to the high costs of analyses, limited biofluid sample volumes and complexity of resulting data treatments, this strategy is still not a current practice.

Metabolites associated with MetS clinical components

The second category of articles focusing on MetS individual components allowed us comparing metabolites associated with clinical data defining MetS. Amino acids, glycerolipids and glycerophospholipids are the major metabolite classes reported as being correlated. Among lipid species, results were particularly difficult to report and to compare, due to the diversity in notations of lipid structures. In fact, even if several consortia proposed guidelines [170, 171], there is still different levels of annotations (from lipid class to stereoisomers) and different ontologies among the databases in use.

In these publications, the diversity of outcome, related to cardiometabolic risk was found to be important. Moreover, the lack of description regarding either other MetS criteria or characteristics of controls, together with the absence of additional phenotypic data (*e.g.* physical activity, nutrition) in some publications, prevented us from including them in this review. For example, plasma metabolite concentrations are known to be highly influenced by physical activity and/or microbiota [172-174] and plasma phospholipids were proposed to be indicative of both food habits and metabolic changes [175]. It has been recognized that publication of all the metadata (data about the samples) along with the metabolomic data is a good practice to assess the quality of the models and the drawn conclusions. Despite the existing data repositories in the field (MetaboLights [176], Metabolomics Workbench [68])

and available guidelines provided by the metabolomics standards initiative (MSI) [79, 177], such good practice is still quite rare.

The glycemic component: towards T2D

Among all the MetS criteria, elevated fasting blood glucose was by far the most studied phenotype using metabolomics/lipidomics, because of its direct link with T2D. Studies on dysglycemia have been among the main drivers in this research field using global metabolomic approaches for biomarker discovery and validation. This review allows getting an overview of the publications considering this specific component among a whole set of metabolic risks, which is of great interest, in the context of systems approaches. Metabolomics has been shown as a powerful tool for the identification of relevant pattern of hundreds of detected metabolites that could be used to predict future development of T2D. However, metabolic profiles acquired with semi- or non-targeted approaches are complex and required dedicated variable selection to build powerful predictive models of specific prediabetic phenotypes [178]. As the analysis of data is one of the most challenging steps in the metabolomics approach due to high data dimensionality and limited number of samples, recommendations as well as appropriate statistical workflows have been proposed. They often include a combination of univariate and multivariate analyses and highlighted the importance of feature/variable selection and external validation to minimize the risk of overfitting [179, 180]. In most publications included in the present review, statistical approaches were not described in detail and limited to univariate analyses, which are the most commonly used due to their easiness of interpretation. However, in the context of metabolomics/lipidomics, multivariate methods are of great relevance as they make use of all variables simultaneously and deal with the relationship between variables, reflecting orchestrated biological processes [181].

Limitations and recommendations for further studies

An important limitation concerning this review is the intrinsic issue of selecting a targeted metabolomic/lipidomic approach or interpreting the resulting data in connection with the study design and the phenotypes of interest. Such a strategy can lead to difficulties in interpretation due to missing acquired data on relevant pathways from this context. Moreover, in the selected articles, even if confounding factors have been often considered in study designs, data description and analysis of these potentially interacting factors were frequently lacking. Such biases have often been identified and statistical approaches have been developed to avoid false discoveries in metabolomics [180]. Beyond this aspect, multiple ontologies used to describe metabolites/lipids [182] and the semi quantitative property of most of the analytical methods, are still major bottlenecks of the field.

Despite these limitations, it is now recognized that metabolomics is a powerful tool allowing metabolic stratification of patients and prognosis [183]. Indeed a metabolic signature would lead to a molecular definition of MetS [184], as exemplified by Wiklung et *al.* [185] and Pujos-Guillot et *al.*[186]. Clinically speaking, the interest of subtyping MetS has been shown since the prevalence and risk for further

cardiovascular disease and T2D is associated with different combinations of its components [151]. More recently, Sperling et *al.* [187] highlighted the need of identifying subtypes of MetS on the basis of pathophysiology, as well as studying the evolution of its stages for a more efficient prevention and therapy. In this context, metabolomic and lipidomic signatures are suitable systems approaches not only to identify biomarkers of sub-phenotypes but also for hypothesis generation of the underlying pathogenic mechanisms.

CONCLUSION

The present review indicates that relatively few articles have been published so far on MetS biomarkers identification using metabolomics and lipidomics in adults. Unfortunately, due to many limitations previously highlighted, it is difficult to compare conclusions from the available data. Moreover, individual MetS clinical components were not specifically investigated, despite the fact that metabolomics/lipidomics are recognized as being powerful phenotyping tools in chronic metabolic diseases. Since studies on T2D have been among the main drivers in this research field using these global approaches for biomarker discovery and validation, it can be concluded that metabolomics and lipidomics signatures could be the strategy of choice for a deeper investigation and characterization of MetS and its sub-phenotypes. Considering future research, a number of key recommendations can be made. First, untargeted methods must be performed using multiplatform approaches for a wide detection of metabolite diversity enabling new biomarker discovery. Second, the complexity of metabolomic/lipidomic data has to be investigated using dedicated univariate and multivariate statistics and data reporting has to follow the FAIR principle [71], concerning both population characteristics and marker metadata. This issue is crucial to ensure the reliability, validity and inter-comparability of experimental results. Such effort should allow transferring knowledge from basic research to clinical practices.

MATERIALS AND METHODS

Methodology for review of published literature

The systematic review of the literature was performed according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for conducting systematic reviews [139].

A specific request was made through several bibliographic electronic databases in August 2019. All databases were chosen in line with the application field studied in the review, namely health research and biology, and five were retained: MEDLINE (from 1946 onwards), EMBASE (from 1974 onwards), EMB Review (from 1991 onwards), CINHAL Complete (from 1937 onwards) and PubMed. To ensure that information collected was complete, the request was also performed on grey literature ((CADTH, Clinical Trials, National Guideline Clearing House, National Institute for Health and Care Excellence (NICE), MedNar, Google Scholar and Open Grey). The request combined words and expressions for three conceptual groups: “Metabolomics/lipidomics”, “Metabolic Syndrome” and “metabolites/biomarkers” (Supplemental Material 1). For each database, words and expressions from controlled vocabulary (MeSH, EMTREE and others) and free-text searching were used. Snowballing techniques and Handsearching was also used to identify other references. Duplicate publications were deleted.

Study selection and data extraction

Initially, titles and abstracts were screened by two authors using the following inclusion and exclusion criteria: 1) articles had to be published in English; 2) publications had to contain original data, therefore reviews, book chapters, and editorials were excluded; 3) studies on non-human models (*e.g.* animals, plants, cells) were excluded; human studies were restricted to case/control, observational, and prospective designs; intervention studies were excluded. Finally, population was restricted to adult/aging Caucasian subjects; thus articles on children, adolescents or pregnant women were excluded; 4) the primary outcome had to be the MetS and/or its components, including T2D, and 5) articles referring to genetic/transcriptomic markers or proteomics were also excluded. These two authors resolved disagreements. To determine publication relevance, three authors independently screened all titles and abstracts to assess their eligibility against the following more restrictive criteria: Eligible publications in the review had to include a minimum of 20 subjects per group and available clinical data regarding the MetS criteria: fasting glucose, TG, HDL-C concentrations, waist circumference, systolic and diastolic blood pressures. Concerning the number of subjects considered as minimum per study, it is generally admitted that 30 subjects is a limit to be able to perform common methods in statistics, in relation to a normal distribution. Moreover, because of the diversity/complexity of the MetS metabolic phenotypes, influenced by numerous factors (gender, age, diet…), taking a population of 40 subjects (i.e. 20 subjects per group for a case/control study) was considered as a minimum requirement. Disagreements in abstracts inclusion were resolved after consensual decision involving a fourth author.

Pertinent data from papers were then extracted, including, author names, publication year, study population and design, number of subjects, gender/sex, baseline clinical characteristics and main outcome. The experimental measures were collected regarding the nature of the biological samples, the analytical approach and techniques, and information regarding statistical methods and covariates when relevant. The results were analysed and compiled by biochemical family including significantly modulated metabolites ($p\text{-value}<0.05$), metabolite listings with levels of change according to the outcome and/or MetS clinical criteria. Finally, results from different studies were compared using Venn diagrams [188] to obtain a more synthetic view.

Ethics statement: This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1 Alberti, K. G. *et al.* Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. *Circulation* **120**, 1640-1645, doi:10.1161/CIRCULATIONAHA.109.192644 (2009).

2 Alberti, K. G., Zimmet, P. & Shaw, J. Metabolic syndrome--a new world-wide definition. A Consensus Statement from the International Diabetes Federation. *Diabet Med* **23**, 469-480, doi:10.1111/j.1464-5491.2006.01858.x (2006).

3 Day, C. Metabolic syndrome, or What you will: definitions and epidemiology. *Diab Vasc Dis Res* **4**, 32-38, doi:10.3132/dvdr.2007.003 (2007).

4 Lam, D. W. & LeRoith, D. in *Endotext* (eds L. J. De Groot *et al.*) (2000).

5 Mottillo, S. *et al.* The metabolic syndrome and cardiovascular risk a systematic review and meta-analysis. *J Am Coll Cardiol* **56**, 1113-1132, doi:10.1016/j.jacc.2010.05.034 (2010).

6 Ervin, R. B. Prevalence of metabolic syndrome among adults 20 years of age and over, by sex, age, race and ethnicity, and body mass index: United States, 2003-2006. *Natl Health Stat Report*, 1-7 (2009).

7 Ford, E. S., Mannino, D. M., National, H. & Nutrition Examination Survey Epidemiologic Follow-up, S. Prospective association between lung function and the incidence of diabetes: findings from the National Health and Nutrition Examination Survey Epidemiologic Follow-up Study. *Diabetes Care* **27**, 2966-2970 (2004).

8 Beltran-Sanchez, H., Harhay, M. O., Harhay, M. M. & McElligott, S. Prevalence and trends of metabolic syndrome in the adult U.S. population, 1999-2010. *J Am Coll Cardiol* **62**, 697-703, doi:10.1016/j.jacc.2013.05.064 (2013).

9 Saklayen, M. G. The Global Epidemic of the Metabolic Syndrome. *Curr Hypertens Rep* **20**, 12, doi:10.1007/s11906-018-0812-z (2018).

10 Nolan, P. B., Carrick-Ranson, G., Stinear, J. W., Reading, S. A. & Dalleck, L. C. Prevalence of metabolic syndrome and metabolic syndrome components in young adults: A pooled analysis. *Prev Med Rep* **7**, 211-215, doi:10.1016/j.pmedr.2017.07.004 (2017).

11 Ford, E. S., Giles, W. H. & Dietz, W. H. Prevalence of the metabolic syndrome among US adults: findings from the third National Health and Nutrition Examination Survey. *JAMA* **287**, 356-359 (2002).

12 van Vliet-Ostaptchouk, J. V. *et al.* The prevalence of metabolic syndrome and metabolically healthy obesity in Europe: a collaborative analysis of ten large cohort studies. *BMC Endocr Disord* **14**, 9, doi:10.1186/1472-6823-14-9 (2014).

13 Kaur, J. A comprehensive review on metabolic syndrome. *Cardiol Res Pract* **2014**, 943162, doi:10.1155/2014/943162 (2014).

14 Ranasinghe, P., Mathangasinghe, Y., Jayawardena, R., Hills, A. P. & Misra, A. Prevalence and trends of metabolic syndrome among adults in the asia-pacific region: a systematic review. *BMC public health* **17**, 101, doi:10.1186/s12889-017-4041-1 (2017).

15 Wilson, P. W., D'Agostino, R. B., Parise, H., Sullivan, L. & Meigs, J. B. Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus. *Circulation* **112**, 3066-3072, doi:10.1161/CIRCULATIONAHA.105.539528 (2005).

16 Stern, M. P., Williams, K., Gonzalez-Villalpando, C., Hunt, K. J. & Haffner, S. M. Does the metabolic syndrome improve identification of individuals at risk of type 2 diabetes and/or cardiovascular disease? *Diabetes Care* **27**, 2676-2681 (2004).

17 Li, C. & Ford, E. S. Definition of the Metabolic Syndrome: What's New and What Predicts Risk? *Metab Syndr Relat Disord* **4**, 237-251, doi:10.1089/met.2006.4.237 (2006).

18 Grundy, S. M. Pre-diabetes, metabolic syndrome, and cardiovascular risk. *J Am Coll Cardiol* **59**, 635-643, doi:10.1016/j.jacc.2011.08.080 (2012).

19 Poon, V. T., Kuk, J. L. & Ardern, C. I. Trajectories of metabolic syndrome development in young adults. *PLoS One* **9**, e111647, doi:10.1371/journal.pone.0111647 (2014).
20 Steinbrecher, A. & Pischon, T. The potential use of biomarkers in the prevention of Type 2 diabetes. *Expert Review of Endocrinology and Metabolism* **8**, 217-219, doi:http://dx.doi.org/10.1586/eem.13.11 (2013).
21 Ramautar, R., Berger, R., Greef, J. v. d. & Hankemeier, T. Human metabolomics: Strategies to understand biology. (2013).
22 Nicholson, J. K., Lindon, J. C. & Holmes, E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* **29**, 1181-1189, doi:10.1080/004982599238047 (1999).
23 Liggi, S. & Griffin, J. L. Metabolomics applied to diabetes-lessons from human population studies. *Int J Biochem Cell Biol* **93**, 136-147, doi:https://dx.doi.org/10.1016/j.biocel.2017.10.011 (2017).
24 Zhang, A. H., Qiu, S., Xu, H. Y., Sun, H. & Wang, X. J. Metabolomics in diabetes. *Clin Chim Acta* **429**, 106-110, doi:10.1016/j.cca.2013.11.037 (2014).
25 Park, S., Sadanala, K. C. & Kim, E. K. A Metabolomic Approach to Understanding the Metabolic Link between Obesity and Diabetes. *Mol Cells* **38**, 587-596, doi:10.14348/molcells.2015.0126 (2015).
26 Bain, J. R. *et al.* Metabolomics applied to diabetes research: moving from information to knowledge. *Diabetes* **58**, 2429-2443 (2009).
27 Cajka, T. & Fiehn, O. Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics. *Anal Chem* **88**, 524-545, doi:10.1021/acs.analchem.5b04491 (2016).
28 Duarte, I. F., Diaz, S. O. & Gil, A. M. NMR metabolomics of human blood and urine in disease research. *J Pharm Biomed Anal* **93**, 17-26, doi:10.1016/j.jpba.2013.09.025 (2014).
29 Forcisi, S. *et al.* Liquid chromatography-mass spectrometry in metabolomics research: mass analyzers in ultra high pressure liquid chromatography coupling. *J Chromatogr A* **1292**, 51-65, doi:10.1016/j.chroma.2013.04.017 (2013).
30 Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. & Group, P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *J Clin Epidemiol* **62**, 1006-1012, doi:10.1016/j.jclinepi.2009.06.005 (2009).
31 Capel, F. *et al.* Metabolomics reveals plausible interactive effects between dairy product consumption and metabolic syndrome in humans. *Clin Nutr*, doi:10.1016/j.clnu.2019.06.013 (2019).
32 Surowiec, I. *et al.* Metabolomic and lipidomic assessment of the metabolic syndrome in Dutch middle-aged individuals reveals novel biological signatures separating health and disease. *Metabolomics* **15**, 23, doi:10.1007/s11306-019-1484-7 (2019).
33 Becker, S., Kortz, L., Helmschrodt, C., Thiery, J. & Ceglarek, U. LC-MS-based metabolomics in the clinical laboratory. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences.*, doi:http://dx.doi.org/10.1016/j.jchromb.2011.10.018.
34 Alonso, A., Marsal, S. & Julia, A. Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in bioengineering and biotechnology* **3**, 23, doi:10.3389/fbioe.2015.00023 (2015).
35 Sas, K. M., Karnovsky, A., Michailidis, G. & Pennathur, S. Metabolomics and diabetes: analytical and computational approaches. *Diabetes* **64**, 718-732 (2015).
36 Liebisch, G. *et al.* Shorthand notation for lipid structures derived from mass spectrometry. *J Lipid Res* **54**, 1523-1530, doi:10.1194/jlr.M033506 (2013).
37 Sud, M. *et al.* LMSD: LIPID MAPS structure database. *Nucleic Acids Res* **35**, D527-532, doi:10.1093/nar/gkl838 (2007).

38 Gonzalez-Franquesa, A., Burkart, A. M., Isganaitis, E. & Patti, M. E. What Have Metabolomics Approaches Taught Us About Type 2 Diabetes? *Current diabetes reports* **16**, 74, doi:10.1007/s11892-016-0763-1 (2016).
39 Palau-Rodriguez, M. *et al.* Metabolomic insights into the intricate gut microbial-host interaction in the development of obesity and type 2 diabetes. *Frontiers in microbiology* **6**, 1151, doi:10.3389/fmicb.2015.01151 (2015).
40 Shapiro, H., Suez, J. & Elinav, E. Personalized microbiome-based approaches to metabolic syndrome management and prevention. *Journal of diabetes* **9**, 226-236, doi:10.1111/1753-0407.12501 (2017).
41 Forouhi, N. G. *et al.* Differences in the prospective association between individual plasma phospholipid saturated fatty acids and incident type 2 diabetes: the EPIC-InterAct case-cohort study. *The lancet* **Diabetes & endocrinology. 2**, 810-818 (2014).
42 Kale, N. S. *et al.* MetaboLights: An Open-Access Database Repository for Metabolomics Data. *Curr Protoc Bioinformatics* **53**, 14 13 11-18, doi:10.1002/0471250953.bi1413s53 (2016).
43 Sud, M. *et al.* Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res* **44**, D463-470, doi:10.1093/nar/gkv1042 (2016).
44 Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **3**, 211-221, doi:10.1007/s11306-007-0082-2 (2007).
45 Hardy, N. W. & Taylor, C. A roadmap for the establishment of standard data exchange structures for metabolomics. *Metabolomics* **3**, 243-248 (2007).
46 Lehmann, R. Diabetes subphenotypes and metabolomics: The key to discovering laboratory markers for personalized medicine? (2013).
47 Grissa, D. *et al.* Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data. *Front Mol Biosci* **3**, 30, doi:10.3389/fmolb.2016.00030 (2016).
48 Broadhurst, D. I. & Kell, D. B. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* **2**, 171-196 (2006).
49 Saccenti, E., Hoefsloot, H. C. J., Smilde, A. K., Westerhuis, J. A. & Hendriks, M. M. W. Reflections on univariate and multivariate analysis of metabolomics data. *metabolomics* **10**, 361-374 (2014).
50 Vinaixa, M. *et al.* Mass spectral databases for LC/MS and GC/MS-based metabolomics: state of the field and future prospects. *Trends in Analytical Chemistry*, doi:http://dx.doi.org/doi:10.1016/j.trac.2015.09.005 (2015).
51 Lindon, J. C. & Nicholson, J. K. The emergent role of metabolic phenotyping in dynamic patient stratification. *Expert Opin Drug Metab Toxicol* **10**, 915-919, doi:10.1517/17425255.2014.922954 (2014).
52 Dumas, M. E., Kinross, J. & Nicholson, J. K. Metabolic phenotyping and systems biology approaches to understanding metabolic syndrome and fatty liver disease. *Gastroenterology* **146**, 46-62, doi:10.1053/j.gastro.2013.11.001 (2014).
53 Wiklund, P. K. *et al.* Serum metabolic profiles in overweight and obese women with and without metabolic syndrome. *Diabetol Metab Syndr* **6**, 40, doi:10.1186/1758-5996-6-40 (2014).
54 Pujos-Guillot, E. *et al.* Systems Metabolomics for Prediction of Metabolic Syndrome. *J Proteome Res* **16**, 2262-2272, doi:10.1021/acs.jproteome.7b00116 (2017).
55 Sperling, L. S. *et al.* The CardioMetabolic Health Alliance: Working Toward a New Care Model for the Metabolic Syndrome. *J Am Coll Cardiol* **66**, 1050-1067, doi:10.1016/j.jacc.2015.06.1328 (2015).
56 Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018, doi:10.1038/sdata.2016.18 (2016).
57 Bardou, P., Mariette, J., Escudie, F., Djemiel, C. & Klopp, C. jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics* **15**, 293, doi:10.1186/1471-2105-15-293 (2014).

58 Caimi, G. *et al.* Evaluation of nitric oxide metabolites in a group of subjects with metabolic syndrome. *Diabetes Metab Syndr* **6**, 132-135, doi:10.1016/j.dsx.2012.09.012 (2012).
59 James-Todd, T. M., Huang, T., Seely, E. W. & Saxena, A. R. The association between phthalates and metabolic syndrome: the National Health and Nutrition Examination Survey 2001-2010. *Environ Health* **15**, 52, doi:10.1186/s12940-016-0136-x (2016).
60 Kulkarni, H. *et al.* Variability in associations of phosphatidylcholine molecular species with metabolic syndrome in Mexican-American families. *Lipids* **48**, 497-503, doi:10.1007/s11745-013-3781-7 (2013).
61 Ntzouvani, A. *et al.* Amino acid profile and metabolic syndrome in a male Mediterranean population: A cross-sectional study. *Nutr Metab Cardiovasc Dis*, doi:10.1016/j.numecd.2017.07.006 (2017).
62 Olszanecka, A., Kawecka-Jaszcz, K. & Czarnecka, D. Association of free testosterone and sex hormone binding globulin with metabolic syndrome and subclinical atherosclerosis but not blood pressure in hypertensive perimenopausal women. *Arch Med Sci* **12**, 521-528, doi:10.5114/aoms.2016.59925 (2016).
63 Ramakrishanan, N., Denna, T., Devaraj, S., Adams-Huet, B. & Jialal, I. Exploratory lipidomics in patients with nascent Metabolic Syndrome. *Journal of Diabetes and its Complications* **32**, 791-794, doi:http://dx.doi.org/10.1016/j.jdiacomp.2018.05.014 (2018).
64 Shim, K., Gulhar, R. & Jialal, I. Exploratory metabolomics of nascent metabolic syndrome. *Journal of Diabetes and its Complications* **33**, 212-216, doi:http://dx.doi.org/10.1016/j.jdiacomp.2018.12.002 (2019).
65 Tremblay-Franco, M. *et al.* Effect of obesity and metabolic syndrome on plasma oxysterols and fatty acids in human. *Steroids* **99**, 287-292, doi:10.1016/j.steroids.2015.03.019 (2015).
66 Antonio, L. *et al.* Associations between sex steroids and the development of metabolic syndrome: A longitudinal study in European men. *Journal of Clinical Endocrinology and Metabolism* **100**, 1396-1404, doi:http://dx.doi.org/10.1210/jc.2014-4184 (2015).
67 Barrea, L. *et al.* Trimethylamine-N-oxide (TMAO) as novel potential biomarker of early predictors of metabolic syndrome. *Nutrients* **10**, doi:http://dx.doi.org/10.3390/nu10121971 (2018).
68 Blouin, K. *et al.* Contribution of age and declining androgen levels to features of the metabolic syndrome in men. *Metabolism* **54**, 1034-1040, doi:10.1016/j.metabol.2005.03.006 (2005).
69 Cheng, S. *et al.* Metabolite profiling identifies pathways associated with metabolic risk in humans. *Circulation* **125**, 2222-2231, doi:10.1161/circulationaha.111.067827 (2012).
70 Favennec, M. *et al.* The kynurenine pathway is activated in human obesity and shifted toward kynurenine monooxygenase activation. *Obesity (Silver Spring)* **23**, 2066-2074, doi:10.1002/oby.21199 (2015).
71 Gao, X., Tian, Y., Randell, E., Zhou, H. & Sun, G. Unfavorable associations between serum trimethylamine N-oxide and L-carnitine levels with components of metabolic syndrome in the Newfoundland population. *Frontiers in Endocrinology* **10**, doi:http://dx.doi.org/10.3389/fendo.2019.00168 (2019).
72 Ho, J. E. *et al.* Metabolomic Profiles of Body Mass Index in the Framingham Heart Study Reveal Distinct Cardiometabolic Phenotypes. *PLoS One* **11**, e0148361, doi:10.1371/journal.pone.0148361 (2016).
73 Huynh, K. *et al.* High-Throughput Plasma Lipidomics: Detailed Mapping of the Associations with Cardiometabolic Risk Factors. *Cell Chemical Biology* **26**, 71-84.e74, doi:http://dx.doi.org/10.1016/j.chembiol.2018.10.008 (2019).
74 Liu, J. *et al.* A Mendelian Randomization Study of Metabolite Profiles, Fasting Glucose, and Type 2 Diabetes. *Diabetes* **66**, 2915-2926, doi:10.2337/db17-0199 (2017).
75 Marchand, G. B. *et al.* Increased body fat mass explains the positive association between circulating estradiol and insulin resistance in postmenopausal women. *American journal of physiology* **Endocrinology and metabolism.** **314**, E448-E456, doi:http://dx.doi.org/10.1152/ajpendo.00293.2017 (2018).

76 Neeland, I. J. *et al.* Relation of plasma ceramides to visceral adiposity, insulin resistance and the development of type 2 diabetes mellitus: the Dallas Heart Study. *Diabetologia* **61**, 2570-2579, doi:http://dx.doi.org/10.1007/s00125-018-4720-1 (2018).

77 Ottosson, F., Smith, E., Melander, O. & Fernandez, C. Altered asparagine and glutamate homeostasis precede coronary artery disease and type 2 diabetes. *Journal of Clinical Endocrinology and Metabolism* **103**, 3060-3069, doi:http://dx.doi.org/10.1210/jc.2018-00546 (2018).

78 Wang-Sattler, R. *et al.* Novel biomarkers for pre-diabetes identified by metabolomics. *Mol Syst Biol* **8**, 615, doi:10.1038/msb.2012.43 (2012).

79 Lind, P. M., Zethelius, B. & Lind, L. Circulating levels of phthalate metabolites are associated with prevalent diabetes in the elderly. *Diabetes Care* **35**, 1519-1524, doi:10.2337/dc11-2396 (2012).

80 Liu, J. *et al.* Metabolomics based markers predict type 2 diabetes in a 14-year follow-up study. (2017).

81 Meikle, P. J. *et al.* Plasma Lipid Profiling Shows Similar Associations with Prediabetes and Type 2 Diabetes. *PLoS ONE* **8 (9) (no pagination)**, doi:http://dx.doi.org/10.1371/journal.pone.0074341 (2013).

82 Peddinti, G. *et al.* Early metabolic markers identify potential targets for the prevention of type 2 diabetes. *Diabetologia*, doi:10.1007/s00125-017-4325-0 (2017).

83 Suvitaival, T. *et al.* Lipidome as a predictive tool in progression to type 2 diabetes in Finnish men. *Metabolism*, doi:10.1016/j.metabol.2017.08.014 (2017).

84 Yengo, L. *et al.* Impact of statistical models on the prediction of type 2 diabetes using non-targeted metabolomics profiling. (2016).

Acknowledgements: S. Monnerie is recipient of a doctoral fellowship from the INRA DID'IT metaprogramme. The *Centre Hospitalier de l'Université de Montréal* is acknowledged for the salary support of D. Ziegler.

Author Contributions: BC, EPG and PG designed the study; DZ performed the search; SM, BC, EPG and PG analyzed the data; JAM gave advices on clinical aspects of selected articles; SM, BC, EPG and PG wrote the manuscript and all authors reviewed the content.

Additional information: *Supplementary information* accompanies this paper. *Competing interests:* All authors have declared no competing interests.

Table 1: Characteristics of case/control studies on MetS.

Reference (Study, population location)	Study design	Outcome (MetS definition)	N	Age range	Gender	Population sample characteristics									Methods			Results
						N	Type	Age	BMI	WC (cm)	Sys BP Dia BP (mmHg)	Glucose (mM)	TG (mM)	HDL-C (mM)	Bio-logical fluid	Data production	Statistical method (covariates in fully adjusted model)	Family with significantly modulated metabolites
Caimi,2012 [189] (Italy)	Case/ Control	MetS (IDF) + T2D (IDF)	160	-	M+W	106	MetS	54±9	32±5	107 ±11	132±16 81±10	6.3±2.5	2.5±1.7	1.0±0.3	Blood	Nitric oxide measurement (Micro-method)	Student's t test	Nitric oxides
						54	non-MetS	No population description										
Capel_2018 [165] (Mona Lisa survey, France)	Case/ Control	MetS (Alberti 2009)	298	35-74	M+W	61	MetS	54±8	30±5	102±10	141±20 88±12	5.7±0.6	2.0±0.8	1.3±0.3	Plasma	Semi targeted LC/MS-MS (Metabolon® platform)	ANOVA (dairy consumption), p< 0.05, q-value <0.1	Amines, amino acids and derivatives, bile acids, carbohydrates and derivatives, carnitines, cholesterol and oxysterols, cofactors and vitamins, eicosanoids, fatty acids, glycerolipids, glycerophospholipids, glycolysis related metabolites, imidazoles, lipids, organic acids, organonitrogen compoinds, peptides, purines and derivatives, pyrimidines and derivatives, sphingolipids, steroids, ureas
						237	non-MetS	48±8	24±3	85±10	122±16 77±10	5.1±0.4	1.0±0.4	1.6±0.3				
James-Todd_2016 [190] (NHANES, USA)	Case/ Control	MetS (NCEP ATP III)	1338	20-80	M	464	MetS	52±22	33±7	114±22	129±22 74±22	6.7±4.3	2.8±4.3	1.1±0.4	Urine	Targeted LC/MS-MS	Logistic regression	Phtalates
						924	non-MetS	43±30	27±6	96±30	119±30 70±30	5.6±1.2	1.4±0.9	1.1±0.6				
	Case/ Control	MetS (NCEP ATP III)	1331	20-81	W	501	MetS	53±22	33±9	107±22	126±22 71±22	6.4±2.2	2.1±2.2	1.3±0.5				
						830	non-MetS	43±29	27±6	89±29	115±29 69±12	5.1±5.8	1.1±0.9	1.6±0.3				
Kulkarni_2013 [191] (SAFHS, USA)	Case/ Control	MetS (IDF)	1358	22-56	M+W	1358	total pop	39±17	29±7	95±17	120±19 71±10	5.6±2.5	1.7±1.2	1.3±0.3	Plasma	Targeted LC/MS-MS	Polygenic regression (Age, age², sex/gender, age x sex/gender interaction, and age² x sex/gender interactions), correction for multiple testing using Li and Ji method.	Glycerophospholipids
Ntzouvani_2017 [192] (Greece)	Case/ Control	MetS (IDF)	100	over 30	M	56	MetS	58* (47;64)	29* (27;32)	105* (100;112)	134* (126;138) 85* (79;90)	5.5* (5.0; 6.1)	1.9* (1.4;2.5)	1.0* (0.9;1.2)	Plasma	Targeted LC/MS-MS + GC/MS-MS	Mann-Whitney U test	Amino acids and derivatives

						44	non-MetS	54* (47;57)	25* (24;27)	91* (87;93)	124* (116;131) 80* (71;86)	5.1* (4.8; 5.4)	1.1* (0.8;1.4)	1.3* (1.1;1.5)				
Olszanecka_2016 [193] (Poland)	Case/ Control	MetS (IDF)	152	40-60	W	63	MetS	51±3	29±3	90±7	163±20 93±12	5.3±0.6	2.3±1.2	1.3±0.3	Serum	Microparticle Enzyme Immunoassay (MEIA Kits, Abbott)	Student's t test, p<0.05	Steroids
						89	non-MetS	51±2	26±3	84±8	151±13 89±11	4.9±0.4	1.2±0.8	1.7±0.3				
Ramakrishanan_2018 [194] (USA)	Case/ Control	MetS (NCEP ATP III)	50	24-72	M+W	30	MetS	53±9	35±6	109±14	132±11 80±9	5.4±0.7	1.7	1.0±0.3	Urine	Targeted LC/MS-MS	Wilcoxon Rank Sum test, p<0.05	Glycerophospholipids
						20	non-MetS	48±13	30±6	92±14	117±12 14±9	4.8±0.4	0.7	1.3±0.3				
Shim_2019 [195] (USA)	Case/ Control	MetS (NCEP ATP III)	50	24-72	M+W	30	MetS	53±9	35±6	109±14	132±11 80±9	5.4±0.7	1.7	1.0±0.3	Urine	Targeted GC/MS	Wilcoxon Rank Sum test, p<0.05	Amino acids and derivatives
						20	non-MetS	48±13	30±6	92±14	117±12 14±9	4.8±0.4	0.7	1.3±0.3				
Surowiec_2018 [166] (Leiden Longevity Study, Netherlands)	Case/ Control	MetS (NCEP ATP III)	115	-	M+W	50	MetS	64±6	NA	106±10	147±18 85±9	6.9±3	2.3±1.3	1.1±0.3	Plasma	Semi targeted LC/MS + GC/MS	Linear regression, p-value < 7.5E-4 after correction for multiple testing	Amines, amino acids and derivatives, carbihydrates and derivatives, carnitines, glycerolipids, glycerophospholipids, glycolysis related metabolites, organic acids, sphingolipids, ureas
						65	non-MetS	62±7	NA	96±12	130±18 77±9	5.4±1.3	1.2±0.5	1.6±0.4				
Tremblay-Franco_2015 [196] (Finland)	Case/ Control	MetS (NCEP ATP III) + obesity	345	around 40	M+W	75	MetS	46±10	35±6	NA	135±14 87±9	NA	1.6±0.8	1.2±0.3	Plasma	Targeted GC/MS	Kruskal-Wallis test; False discovery rate adjusted p-value	Cholesterol and oxysterols
						210	non-MetS	42±11	25±2	NA	120±12 78±8	NA	1.0±0.4	1.5±0.4				
Wiklund_2014 [185] (EWI-study, Finland)	Case/ Control	MetS (Alberti 2009)	78	around 40	W	36	MetS	44±6	31±3	99±6	136±11 84±7	5.5±0.7	2.0±0.9	1.4±0.3	Serum	Non-targeted NMR	ANCOVA (Age, fat mass and WC); false discovery adjusted p<5E-4	Amino acids and derivatives, Carbohydrates and derivatives, Glycolysis related metabolites, Glycerophospholipids, Fatty acids, Cholesterol and oxysterols
						42	non-MetS	40±8	29±3	96±9	122±7 78±6	5.1±0.3	1.0±0.3	1.6±0.3				
Antonio_2015 [197] (EMAS, Europe)	Prospec-tive (4 years follow-up)	MetS (NCEP ATP III) prediction	1651	40-79	M	289	MetS	59±10	28±3	101±8	147±21 88±13	5.5±1.0	1.5±0.8	1.4±0.4	Serum	Targeted LC/MS + GC/MS	Logistic regression (unadjusted or adjusted for age, study center, alcohol intake, current smoking status, physical activity, general health)	Steroids
						1362	non-MetS	59±11	26±3	93±9	142±20 85±11	5.3±0.8	1.2±0.6	1.5±0.4				
Pujos-Guillot_2017 [186] (GAZEL, France)	Prospec-tive (5 years follow-up)	MetS (NCEP ATP III) prediction	112	52-64	M	56	MetS	59±3	27±1	95±4	137±14 80±8	6.6±1.3	1.2±0.5	1.5±0.3	Serum	Non-targeted LC/MS	Two-way ANOVA, p<0.05 (BMI)	Amino acids and derivatives, Carbohydrates and derivatives, Carnitines, Fatty acids and derivatives, Glycerophospholipids, Peptides
						56	non-MetS	59±3	27±1	92±5	129±12 78±8	5.5±0.5	1.0±0.4	1.5±0.4				

BMI = body mass index; WC = waist circumference; BP = blood pressure (sys = systolic; dia = diastolic); TG = triglycerides; HDL-C = high-density lipoprotein cholesterol. Mean values ± SD; *Median value (25^{th}; 75^{th} percentiles)

Table 2: Characteristics of studies investigating correlations between metabolites and MetS criteria.

Reference (Study, population location)	Study design	Outcome (definition)	N	Age range	Gender	Population sample characteristics								Methods			Results
						Mean type (available or calculated)	Age	BMI	WC (cm)	Sys BP Dia BP (mmHg)	Glucose (mM)	TG (mM)	HDL-C (mM)	Biological fluid	Data production	Statistical method (covariates in fully adjusted model)	Family with significantly modulated metabolites
Barrea_2018 [198] (Italy)	-	MetS (NCEP ATP III)	137	20-63	M+W	Calculated	36	33	109	126 80	5.5	1.6	1.1	Serum	Targeted LC/MS	Pearson correlation p<0.05 (Gender, BMI, smocking, physical activity and total energy intake)	Aminoxides
Blouin_2005 [199] (Quebec family study (QFS), Quebec (CAN))	-	MetS (NCEP ATP III)	130	20-71	M	Available	43±15	27±5	93±14	117±16 73±10	5.5±1.1	1.5±0.8	1.1±0.3	Plasma	Targeted LC/MS + GC/MS	Spearman correlation (Age or visceral adipose tissue area)	Steroids
Caimi_2012 [189] (Italy)	Case/ Control	MetS ± T2D (IDF)	160	-	M+W	All MetS	54±9	32±5	107 ±11	132±16 81±10	6.3±2.5	2.5±1.7	1.0±0.3	Blood	Nitric oxide measurement (Micromethod)	Linear regression p<0.05	Nitric oxides
Cheng_2012 [200] (Framingham Heart Study (FHS), USA) (Malmö Diet and Cancer Study (MDC), Sweden)	Case/ Control	Cardio-metabolic risk	1015	47-65	M+W	Available	56±9	28±5	96±14	129±18 76±10	5.4±0.6	1.8±1.2	1.2±0.4	Plasma	Semi-targeted LC/MS	Linear regression (Age and sex/gender)	Amino acids and derivatives, Carnitines, Glycerophospholipids, Imidazoles, Indoles and derivatives, Peptides, Purines and derivatives
	Case/ Control	Cardio-metabolic risk	746	53-65	M+W	Available	59±6	27±4	88±13	147±19 90±9	5.1±0.5	1.3±NA	1.3±0.3				
Favennec_2015 [201] (D.E.S.I.R. cohort, France)	Case/ Control	T2D	1048	37-60	M+W	Calculated	48	25	85	NA	5.5	NA	NA	Serum	Targeted LC/MS + LC/MS-MS + GC/MS-MS	Logistic regression (Age, sex/gender BMI)	Amino acids and derivatives, Pyridines and derivatives
(Biological Atlas of Severe Obesity (ABOS), France)	Case/ Control	Obesity	109	26-56	W	Calculated	46	25	121	NA	6.6	NA	NA			Spearman correlation (Age, sex/gender BMI)	
Gao_2019 [202] (CODING, Canada)	-	MetS	536	-	M	Available	42±13	28±5	99±13	133±15 84±10	5.3±0.7	1.5±1	1.2±0.3	Serum	Targeted LC/MS-MS	Partial correlation (Age, total calorie intake, physical activity level, medicine status, alcohol intake and smoking status)	Carnitines
			545	-	W	Available	45±11	27±5	91±15	123±16 80±11	5.1±0.7	1.2±0.7	1.5±0.4			Partial correlation (Age, total calorie intake, physical activity level, medical status, alcohol intake, smoking status and menopausal status)	
Ho_2016 [203] (Framingham Heart Study (FHS), USA)	-	BMI	2383	45-65	M+W	Available	55±10	28±5	NA	126±19 75±10	5.3* (4.9;5.7)	1.4* (1.0;2.0)	1.2* (1.0;1.5)	Plasma	Targeted LC/MS	Linear regression (Age, sex/gender, batch, BMI (except for BMI and WC), and log-TG concentrations (except for TG analyses)) Bonferroni corrected threshold p< 0.00023	Amino acids and derivatives, Carbohydrates and derivatives, Carnitines, Ceramides, Cholines, Fatty acids, Glycerolipids, Glycerophospholipids, Glycolysis related metabolites, Peptides, Purines and derivatives

Huynh_2019 [204] (AusDiab, Australia)	-	Cardio-metabolic risk	389	-	M+W	Available	55±12	27±4	NA	131±18 71±11	5.3±0.4	1.5±0.9	1.46±0.4	Plasma	Targeted LC/MS	Linear regression (Age, gender, total cholesterol, HDL-C and TG)	Carnitines, Cholesterols and oxysterols, Glycerolipids, Glycerophospholipids, Sphingolipids
Liu_2017 [205] (ERF, Netherlands)	Case/ Control	T2D	2776	-	M+W	Calculated	49	27	NA	140 80	4.7	1.2	1.3	Plasma	Targeted LC/MS, NMR	Partial correlation (Age, sex and lipid-lowering medication)	Cholesterols and oxysterolsAmino acids and derivatives, Carbohydrates and derivatives, Carnitines, Glycerolipids, Glycerophospholipids, Glycolysis related metabolites, Organic acids
Marchand_2018 [206] (Quebec (CAN))	-	Insulin resistance	101	48-68	W	Available	57±4	28±5	89±12	130±15 82±7	5.6±0.8	1.3±0.7	1.4±0.4	Plasma	Targeted LC/MS	Pearson correlation	Steroids
Neeland_2018 [207] (DHS, USA)	-	T2D	3072	18-65	M+W	Available	43±10	28	NA	119 NA	5	5.2	2.7	Plasma	Targeted LC/MS	Spearman correlation (Age, sex and ethnicity)	Sphingolipids
Ntzouvani_2017 [192] (Greece)	Case/ Control	MetS (IDF)	100	over 30	M	Calculated	56	27	NA	130 83	5.3	1.5	1.1	Plasma	Targeted LC/MS-MS + GC/MS-MS	Correlation of principal component analysis factors (age)	Amino acids and derivatives
Ottosson_2018 [208] (Malmö Preventive Project, Sweden)	-	T2D	1084	-	M+W	Calculated	69	27	NA	147 NA	5.5	1.3	1.3	Plasma	Targeted LC/MS	Spearman correlation (False-discovery rate)	Amines, Amino acids and derivatives, Carnitines, Cholines
Ramakrishanan_2018 [194] (USA)	Case/ Control	MetS (NCEP ATP III)	50	24-72	M+W	Calculated	51	33	102	126 78	5.2	1.3	1.2	Urine	Targeted LC/MS-MS	Spearman correlation	Glycerophospholipids
Shim_2019 [195] (USA)	Case/ Control	MetS (NCEP ATP III)	50	24-72	M+W	Calculated	51	33	102	126 78	5.2	1.3	1.2	Urine	Targeted GC/MS	Spearman correlation	Amino acids and derivatives
Wang-Satler_2012 [209] (KORA, Germany)	Case/ Control	T2D	1297	58-72	M+W	Calculated	64	28	NA	135 NA	5.6	1.5	1.5	Serum	Targeted LC/MS (AbsoluteIDQ® p180 kit: Biocrates)	Pearson correlation	Amino acids and derivatives, Carbohydrates and derivatives, Glycerophospholipids, Shingolipids

BMI = body mass index; WC = waist circumference; BP = blood pressure (sys = systolic; dia = diastolic); TG = triglycerides; HDL-C = high-density lipoprotein cholesterol; NA = not available; 'Calculated mean type' refers to clinical variable means that were calculated, when missing, from the available data in the publication. Mean values ± SD; *Median value (25^{th}; 75^{th} percentiles)

Table 3: Characteristics of case/control studies on T2D.

Reference (Study, population location)	Study design	Outcome	N	Age range	Gender	Population sample characteristics									Methods			Results
						N	Type	Age	BMI	WC (cm)	Sys BP Dia BP (mmHg)	Glucose (mM)	TG (mM)	HDL-C (mM)	Biological fluid	Data production	Statistical method (covariates in fully adjusted model)	Family with significantly modulated metabolites
Lind_2012 [210] (PIVUS, Sweden)	Case/ Control	T2D	1016	70	M+W	119	T2D	70	29±5	98±11	155±24 80±12	8.4±3.1	1.5±0.8	1.4±0.4	Serum	Targeted LC/MS	Logistic regression (Sex/gender, serum cholesterol and TG, BMI, smoking and exercise habits, educational levels)	Phtalates
						897	non-T2D	70	27±4	90±11	149±22 79±10	4.9±0.5	1.3±0.6	1.5±0.4				
Liu_2017 [211] (ERF, Netherland)	Case/ Control	T2D	2776	48-60	M+W	212	T2D	60±12	30±6	99±14	154±21 83±10	7.4±2.2	1.6* (1.1;1.9)	1.1±0.3	Plasma	Targeted LC/MS-MS + NMR	Logistic regression (Age, sex/gender and lipid-lowering medication)	Amino acids and derivatives, Carbohydrates and derivatives, Cholesterol and oxysterols, Glycerolipids, Glycerophospholipids
						2564	non-T2D	48±14	27±5	87±13	139±20 80±10	4.5±0.7	1.2* (0.8;1.6)	1.3±0.4				Glycolysis related metabolites, Organic acids, Peptides
Meikle_2013 [212] (AusDiab, Australia)	Case/ Control	T2D	287	52-73	M+W	117	T2D	62* (52;73)	28* (26;31)	97* (89;104)	143*(131;154) NA	6.9* (5.7; 7.4)	1.9* (1.3; 2.9)	1.2* (1.0;1.5)	Plasma	Targeted LC/MS	Logistic regression (Age, sex/gender, WC and SBP) BH corrected p-value < 0.05	Ceramides, Cholesterol and oxysterols, Glycerolipids, Glycerophospholipids
						170	non-T2D	60* (49;72)	26* (24;28)	90* (83;98)	133*(121;146) NA	5.3* (5.1;5.6)	1.2* (0.9;1.6)	1.4* (1.2;1.7)				
Wang-Satler_2012 [209] (KORA, Germany)	Case/ Control	T2D	957	58-72	M+W	91	T2D	66±5	30±4	NA	147±22 NA	7.4±1.8	1.9±1.2	1.3±0.4	Serum	Targeted LC/MS (AbsoluteIDQ® p180 kit: Biocrates)	Logistic regression (Age, sex/gender, BMI, physical activity, alcohol intake, smoking, SBP and HDL-C + fasting glucose)	Amino acids and derivatives, Carbohydrates and derivatives, Glycerophospholipids
						866	non-T2D	64±6	28±4	NA	132±19 NA	5.3±0.4	1.4±0.8	1.6±0.4				

BMI = body mass index; WC = waist circumference; BP = blood pressure (sys = systolic; dia = diastolic); TG = triglycerides; HDL-C = high-density lipoprotein cholesterol. Mean values ± SD; *Median value (25^{th}; 75^{th} percentiles)

Table 4: Characteristics of prospective studies on T2D.

Reference (Study, population location)	Study design	Follow-up time (years)	Outcome	N	Age range	Gender	Population sample characteristics									Methods			Results
							N	Type	Age	BMI	WC (cm)	Sys BP Dia BP (mmHg)	Glucose (mM)	TG (mM)	HDL-C (mM)	Biological fluid	Data production	Statistical method (covariates in fully adjusted model)	Family with significantly modulated metabolites
Peddinti_2017 [213] (Botnia, Finland + DESIR, France)	Case/ Control	10	T2D prediction	543	48-52	M+W	146	T2D	52±1	29±0.4	96±1	139±2 84±1	5.9±0.05	1.7±0.08	1.3±0.03	Plasma	Semi-targeted LC/MS + GC/MS	Conditional logistic regression FDR q<0.05 (Age, sex/gender, BMI, fasting glucose level and family history of T2D) p- values < 0.05 multivariate logistic regression	Amino acids and derivatives, Bilirubins, Carbohydrates and derivatives, Fatty acids and derivatives, Quinones and hydroquinones
							397	non-T2D	48±1	26±0.2	88±1	130±1 79±1	5.6±0.03	1.3±0.04	1.4±0.01				
Suvitaival_2017 [214] (METSIM (discovery set), Denmark)	Case/ Control	5	T2D prediction	323	53-65	M	107	T2D	59±6	29±4	102±10	143±16 90±9	6.0±0.5	1.9±1.2	1.3±0.4	Plasma	Non-targeted LC/MS	Logistic regression Model (Age and BMI)	Glycerolipids, Glycerophospholipids
							216	non-T2D	60±5	26±2	95±7	133±15 85±9	5.2±0.2	1.1±0.5	1.5±0.4				
Wang-Satler_2012 [209] (KORA, Germany)	Case/ Control	10	T2D prediction	876	58-72	M+W	91	T2D	66±5	30±4	NA	138±19 NA	5.9±0.6	1.7±0.8	1.3±0.3	Serum	Targeted LC/MS (AbsoluteIDQ® p180 kit: Biocrates)	Logistic regression (Age, sex/gender, BMI, physical activity, alcohol intake, smoking, SBP, HDL cholesterol Hb1Ac, fasting glucose and fasting insulin)	Glycerophospholipids
							785	non-T2D	63±5	28±4	NA	132±19 NA	5.4±0.5	1.4±0.8	1.6±0.4				
Yengo_2016 [215] (DESIR, Europe)	Case/ Control	9	T2D prediction (ADA)	1067	37-60	M+W	231	T2D	51±9	28±4	94±11	139±17 84±9	5.9±0.6	1.7±1.2	1.5±0.4	Plasma	Non-targeted LC/MS + GC/MS	Logistic and Cox regressions	Amino acids and derivatives, Carbohydrates and derivatives, Carnitines, Fatty acids and derivatives, Glycerolipids, Glycerophospholipids, Peptides, Purines and derivatives, Steroids
							836	non-T2D	47±10	25±4	83±11	131±16 80±10	5.3±0.7	1.1±0.7	1.6±0.4				

BMI = body mass index; WC = waist circumference; BP = blood pressure (sys = systolic; dia = diastolic); TG = triglycerides; HDL-C = high-density lipoprotein cholesterol.

.

II. L'importance d'une approche globale et multi techniques analytiques

La conduite de cette revue systématique a révélé plusieurs éléments qu'il est intéressant de discuter. Le premier concerne le type d'études menées. Bon nombre des publications analysées présentent les résultats de méthodes d'analyses ciblées. Celles-ci ayant des couvertures analytiques souvent très différentes, il est difficile de comparer les résultats obtenus. Il semble pertinent d'envisager de combiner plusieurs méthodes d'analyse (ciblées ou non) pour maximiser la couverture obtenue et ainsi faciliter la détection de nouveaux biomarqueurs peut être inconnus. Ce type d'approche reste rare notamment du fait de son coût et des quantités importantes d'échantillons nécessaires. Malgré tout, elle reste l'un des meilleurs moyens de caractériser de manière pertinente l'état métabolique des individus et donc d'établir une signature précise d'un état de dérèglement métabolique comme le SMet.

D'autre part, les publications présentant les biomarqueurs « individuels » constituent la majorité des articles retenus. Toutefois, la complexité d'un état de santé comme le SMet est souvent mieux traduite par l'utilisation combinée de plusieurs de ces biomarqueurs et ce sous forme de signature. Il semble donc primordial que de futures études abordent ces aspects, idéalement en complément de l'utilisation de méthodes d'analyse les plus larges et complémentaires possibles, comme évoqué précédemment.

Enfin, cette revue systématique a permis la construction d'une banque de données de métabolites décris comme associés au SMet ou à l'un ou plusieurs de ses composants. Elle contient à ce jour plus de 700 métabolites. Cette revue fut également l'opportunité de dresser un cahier des charges pour faciliter la gestion et l'inclusion des métadonnées au sein des études. Cette dernière étape a mis en avant plusieurs problématiques récurrentes :

- En premier lieu, un manque flagrant de **description des populations** étudiées dans les publications. Certains articles présentant des études cas/contrôles du SMet ne mettent pas à disposition les caractéristiques cliniques des sujets pour les 5 critères du syndrome. Outre les données relatives au SMet, celles de description générale sont parfois manquantes : âge, sexe, ethnicité, potentiels facteurs confondantes, contexte d'habitudes alimentaires, informations détaillées sur les statistiques, etc… Or, comme expliqué auparavant, la notion de biomarqueur ne peut pas être dissociée de la population dans laquelle il est identifié et validé. Ces données sont donc primordiales pour permettre la réutilisation des résultats. L'une des informations les mieux décrites est souvent la méthode de détection avec des matériels et méthodes souvent

détaillés. Il est dommage que la rigueur utilisée pour les données méthodologiques ne soit pas transposée aux autres types de métadonnées de l'étude.

- Nous avons également constaté un manque **d'uniformisation des nomenclatures données aux métabolites**. En effet, certains articles font référence aux mêmes molécules mais utilisent des termes différents pour les nommer (par exemple le 1,5 anhydrohexitol est synonyme du déoxyglucose). Certaines correspondances entre 2 noms peuvent se faire facilement car ce sont des synonymes courants (exemple du glutamate et de l'acide glutamique), mais il arrive que certaines soit plus difficiles à relier, notamment lorsque les nomenclatures viennent de domaines d'étude différents (par exemple le glucose peut également être appelée glucopyranose, et selon son origine, on peut également parler de dextrose ou de cérélose). Il devient alors indispensable de posséder une certaine expertise biologique/chimique pour permettre la détection des redondances entre les différents noms et donc la comparaison des différents résultats obtenus.
- Finalement, nous avons constaté **qu'aucun article ne fournissait d'identifiant de bases de données** (HMDB, ChEBI, InChi…) pour les métabolites identifiés. Ce dernier point aurait permis de résoudre partiellement la problématique précédente tout en facilitant la construction d'une banque de données robuste.

Pour construire la banque de données centrée sur les biomarqueurs, toutes les données descriptives collectées ont été utilisées. Elles permettent, pour un métabolite donné, d'accéder aux différentes études l'ayant décrit comme biomarqueur du SMet, et pour chacune d'entre elles, de connaitre les détails de la population et des analyses réalisées. La mise à disposition de ces connaissances facilite l'interprétation de nouveaux résultats en offrant la possibilité de les confronter à des études précédemment réalisées.

Pour enrichir ces informations, 2 experts chimiste et biologiste ont proposé une uniformisation des nomenclatures rencontrées afin de regrouper ensemble les composés identiques mais nommés différemment, et ce en leur assignant également une appartenance à une famille de molécules (acides aminés, phospholipides, dérivés de la créatinine…). A partir de cette nomenclature nous avons tenté d'interroger des bases de données internes et externes et d'utiliser des outils de conversion pour retrouver un maximum d'identifiants de base de données pour ses composés.

Cette étape se révèle complexe notamment car certaines familles de molécules comme les lipides sont mal différenciées dans les bases de données, il est donc difficile d'obtenir un identifiant précis. Par exemple, la famille des céramides phospho-éthanolamine est à ce jour très mal décrite [216] et absente des bases de données. De plus, les nomenclatures lipidiques du type CerPE(38:2),

CerPE(34:1) ou encore CerPE(36:2), telles qu'utilisées dans la publication de Fall *et al [217]*, indiquent le nombre de carbones présents dans la molécule ainsi que le nombre de double liaison. Ces dernières ne permettant de connaitre ni la répartition des carbones sur une ou plusieurs chaînes, ni la position des doubles liaisons, il peut donc exister plusieurs molécules sous un même nom. De ce fait, la banque de données n'a pas pu être enrichie en totalité (**Figure 13**).

Référence	Metabolite family	Metabolite name	ChEBI	ChemSpider	KEGG	InChI Code	mz
Liu_2017 ; Yengo_201	Carbohydrates and derivatives	1,5-anhydroglucitol	CHEBI:16070	18497858	C07326	InChI=1S/C6H12O5	164.068473
Yengo_2016	Glycerophospholipids	1-linoleoyl-GPC	CHEBI:73869	No result	No result	InChI=1S/C26H50N	519.33249
Yengo_2016	Glycerolipids	1-palmitoylglycerol	CHEBI:69081	No result	No result	InChI=1S/C19H38O	330.27701
Tremblay-Franco_201	cholesterol and oxysterols	25-hydroxycholesterol	No result	58604	C15519	InChI=1S/C27H46O	402.349781
Ho_2016	Amino acids and derivatives	2-aminoadipic acid	CHEBI:37024	456	No result	InChI=1S/C6H11NO	161.068808
Liu_2017	Organic acids	2-hydroxybutyrate	CHEBI:50613	389701	C05984	InChI=1S/C4H8O3/	104.047344
Fall_2016	Fatty acids and derivatives	2-Ketohexanoic acid	CHEBI:17308	140384	C00902	InChI=1S/C6H10O3	130.062994
Fall_2016 ; Fall_2016	Carnitines	2-Methylbutyroylcarnitine	CHEBI:73026	4932320	No result	InChI=1S/C12H23N	245.162708
Liu_2017	Glycolysis related metabolites	2-oxoglutaric acid	CHEBI:30915	50	C00026	InChI=1S/C5H6O5/	146.021523
Yengo_2016	Fatty acids and derivatives	3-hydroxyisobutyrate	CHEBI:18064	85	C01188	InChI=1S/C4H8O3/	104.047344
Tremblay-Franco_201	cholesterol and oxysterols	4-alpha-hydroxycholesterol	No result	19990805	No result	InChI=1S/C27H46O	402.349781
Tremblay-Franco_201	cholesterol and oxysterols	4-beta-hydroxycholesterol	CHEBI:85778	2497521	No result	InChI=1S/C27H46O	402.349781

Figure 13 : Extrait de la banque de données représentant le nom des métabolites associé ou non à des identifiants de base externes.

Posséder des identifiants de bases de données externes au sein de la banque de données permet d'avoir accès aux informations de masse, de conformations, etc… Il devient alors possible d'établir une requête pour parvenir à l'annotation des composés sur la base de cette banque personnalisée notamment grâce à des outils comme Bank Inhouse disponible dans W4M. Il permet de passer en entrée la banque de données personnalisée contenant les noms des métabolites et leur m/z, et de comparer l'information à une liste de masses à annoter. Le fichier de sortie correspond à un fichier annoté dans lequel, pour chaque métabolite annoté, les différentes correspondances obtenues sont listées et séparées par un « | ». Chaque correspondance se présente sous la forme : delta de masse observé # masse correspondante dans la banque de données # nom du métabolite # identifiant ChEBI # identifiant ChemSpider # identifiant KEGG # mz # [M+H]+ # [M-H]- # référence du ou des articles où le métabolite est référencé comme lié au SMet. Un exemple de fichier de résultat est présenté sur la **Figure 14**.

name	mz	DELTA_mass(0.005Da)#MASS_Result#Metabolite name#ChEBI#ChemSpider#KEGG#mz#[M+H]+#[M-H]-#Reference
M100T1212	100.0016	No_result_found_in_bank_inhouse
M102T381	102.056	0.000167415321#102.0561285#Aminoisobutyric acid#CHEBI:27971#No result#C03665#103.0633285#104.0705285#102.0561285#Cheng_2012|0.0001
M103T568	103.0036	No_result_found_in_bank_inhouse
M112T190	111.979	No_result_found_in_bank_inhouse
M112T195	112.0516	No_result_found_in_bank_inhouse
M113T99	113.0608	No_result_found_in_bank_inhouse
M114T376	114.0559	0.000181514496#114.0561285#Proline#CHEBI:17203#128566#C00148#115.0633285#116.0705285#114.0561285#Cheng_2012 ; Ho_2016
M115T567	115.0036	No_result_found_in_bank_inhouse
M115T94	115.0764	No_result_found_in_bank_inhouse
M116T268	116.0352	No_result_found_in_bank_inhouse
M116T506	116.0465	No_result_found_in_bank_inhouse
M116T310	116.0716	0.000153646635#116.0717786#Betaine#CHEBI:17750#242#C00719#117.0789786#118.0861786#116.0717786#Cheng_2012|0.000153646635#116.0717

Figure 14 : Exemple de fichier de résultats issus de Bank Inhouse

En conclusion, la construction d'une banque de données similaire à celle que nous avons mise en place devrait être le point de départ de toute étude métabolomique, et ce quel que soit la problématique posée. Elle nécessite de passer du temps à la conception d'une requête correspondant à la structure que l'on souhaite lui donner, ainsi qu'à la lecture des différents articles retenus, mais elle permet également de créer une base de connaissance qui facilitera par la suite, l'annotation des composés ainsi que l'interprétation des résultats obtenus. Pour que cette banque de données gère au mieux les métadonnées, il est nécessaire de les collecter avec un maximum de précision en ne focalisant pas la lecture uniquement sur les informations d'intérêt directes comme la description du statut cas/témoin. Les recommandations énoncées précédemment permettent d'élaborer un cahier des charges qui ne s'applique pas uniquement au contexte de notre étude. De plus, elles peuvent également servir à guider la rédaction d'une future publication qui devra contenir un maximum de métadonnées et posséder des identifiants clairs et pertinents pour rendre le travail de recherche le plus réutilisable possible par une large communauté scientifique.

REFERENCES DU CHAPITRE 1

1. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol. 2009;62(10):1006-12.
2. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol. 2009;62(10):e1-34.
3. Panevska A, Skocaj M, Krizaj I, Macek P, Sepcic K. Ceramide phosphoethanolamine, an enigmatic cellular membrane sphingolipid. Biochim Biophys Acta Biomembr. 2019;1861(7):1284-92.
4. Fall T, Salihovic S, Brandmaier S, Nowak C, Ganna A, Gustafsson S, et al. Non-targeted metabolomics combined with genetic analyses identifies bile acid synthesis and phospholipid metabolism as being associated with incident type 2 diabetes. Diabetologia. 2016;59(10):2114-24.

CHAPITRE 2 : PRESENTATION DES DONNEES ET MANAGEMENT

I. Les objectifs du projet

Comme énoncé lors de la présentation des objectifs de mon projet de thèse, ce dernier s'inscrit dans un projet plus vaste faisant intervenir différentes disciplines et ayant pour objectif d'identifier des marqueurs multidimensionnels stables du SMet chez la personne âgée en utilisant des approches métabolomiques et lipidomiques afin de permettre une meilleure stratification des populations à risque.

Afin de répondre à cette problématique, une étude cas/témoins portant sur le SMet a été construite. Comme évoqué dans la revue systématique présentée précédemment, les méthodes analytiques disponibles à ce jour ne permettent pas, lorsqu'utilisées individuellement, de couvrir l'intégralité des métabolites, bien souvent ces dernières sont donc complémentaires. Pour cette raison, cette étude a été conçue pour proposer une méthodologie couplant plusieurs approches de métabolomique et lipidomique non ciblées afin de permettre une couverture maximale des métabolites et de leur modulation dans le sérum des individus étudiés.

L'étude utilise des participants issus de la cohorte NuAge, ainsi que sur l'infrastructure d'excellence MetaboHUB et ses plateformes d'analyses en métabolomique et lipidomique. Des échantillons de sérum, prélevé chez chacun des individus à 3 ans d'intervalle (T1 et T4), ont été analysé par 6 méthodes d'analyses métabolomiques non ciblées LC-HRMS (C18 en mode positif et négatif, chromatographie d'interaction hydrophile (HILIC) et lipidomiques), GC-QToF, et RMN. Les données issues de ces analyses ont été ensuite prétraitées suivant un protocole standardisé pour chacune des méthodes. Ces données ont été statistiquement analysées afin d'extraire les métabolites significativement modulés entre cas et témoins mais qui demeurent stables durant la période étudiée. Dans un second temps, une signature caractérisant le syndrome a pu être établie à partir des métabolites stables uniquement. Enfin, une reclassification moléculaire des participants a été établie dans le but d'étudier d'éventuels sous-phénotypes du SMet.

II. Les données de la cohorte NuAge

1. Présentation de la cohorte

L'étude longitudinale québécoise sur la nutrition comme déterminant d'un vieillissement réussi (NuAge) menée sur 5 ans, a été conçue pour étudier le rôle de la nutrition dans le vieillissement en santé. La cohorte NuAge se compose de 1 793 hommes et femmes âgés de 68 à 82 ans lors de leur recrutement au sein de la cohorte en 2003, en bonne santé générale et vivant de façon autonome. Chacun d'eux a été suivi sur une période de 3 ans après leur inclusion dans l'étude.

Au cours de ce suivi, des échantillons de sang, de salive et d'urine ont été prélevés annuellement et stockés à -80°C. Diverses mesures ont été faites sur ces échantillons pour fournir des données biologiques (*e.g.* formule sanguine complète, glucose, insuline, cortisol, bilan lipidique) afin de caractériser *a minima* l'état métabolique des sujets. Un pool d'échantillons sérique a également été préparé au début de l'étude pour évaluer dans le temps la qualité de la conservation et la reproductibilité des différentes mesures. En parallèle, plus de 1 000 mesures complémentaires (quantitatives et qualitatives) ont été faites permettant une évaluation nutritionnelle (*e.g.* fréquence de consommation d'aliments, rappels alimentaires, prise de suppléments), anthropométrique (*e.g.* mesures et composition corporelles), fonctionnelle (*e.g.* force, capacités, activité physique, performances physiques…), médicale (*e.g.* santé physique, mental et cognitive) et sociale (*e.g.* réseau, soutient ou encore participation à la vie sociale) [218].

La cohorte NuAge a été choisie pour mener ce projet car elle représente une base de données unique sur les sujets caucasiens âgés. Elle présente l'avantage de disposer de prélèvements sanguins réalisés en début et en fin de suivi et dont la conservation est contrôlée d'années en années garantissant la fiabilité des analyses futures. De plus, le nombre de métadonnées concernant les sujets (nutritionnelles, sociales, cliniques, etc…) est extrêmement important, permettant ainsi une sélection de sujets adéquats pour répondre à la question de recherche, mais ouvrant également la voie à de nombreuses analyses complémentaires jumelant par exemple observations métaboliques et données cliniques et nutritionnelles.

2. Sélection des sujets

Pour mener notre étude cas/contrôle au sein de la cohorte NuAge, nous avons dû développer une stratégie de sélection de sujets. Elle se base sur la présence et le nombre de critères du SMet, et leur stabilité au cours du temps. Elle a été menée en amont de mon travail de thèse par différents membres de MetaboHUB et de la cohorte NuAge.

Pour permettre la sélection des sujets, plusieurs variables disponibles dans la base de données de la cohorte ont été utilisées, fournissant un certain nombre de descripteurs : identifiant du sujet, variables démographiques ou encore variables cliniques (*e.g.* pression artérielle, glycémie à jeun), renseignées à plusieurs temps. Ces données ont été complétées par des dosages lipidiques pour augmenter les possibilités d'inclusion de sujets dans l'étude. Leur sélection s'est faite en plusieurs étapes présentées au sein du workflow ci-dessous (**Figure 15**).

Figure 15 : Présentation du workflow de sélection des sujets parmi les 1 793 individus initialement inclus dans la cohorte NuAge.

Les différences métaboliques entre hommes et femmes sont connues comme par exemple au niveau du métabolisme des lipides [219]. De plus, d'après le ministère de la santé et des services sociaux du gouvernement du Québec, les hommes présentent une prévalence du diabète supérieure aux femmes (tout comme c'est le cas dans de nombreux autres pays) [220-222]. Pour ces deux raisons, il a été décidé de ne considérer que des individus masculins dans notre étude. Leur nombre s'élève à 853 au sein de la cohorte NuAge.

Dans un premier temps, les variables **quantitatives** du SMet disponibles (glycémie, tour de taille et pression artérielle) ont été utilisées pour réaliser un premier filtre. Les sujets présentant des valeurs manquantes pour l'une ou plus de ces 3 variables à T1 et T4 ont été exclus.

Pour chaque critère du SMet, la positivité ou la négativité au critère a été établie en se basant sur : les 3 variables quantitatives disponibles, les valeurs qualitatives binaire pour les critères de l'hypertriglycéridémie et de l'hypo-HDL-cholestérolémie ainsi que les données de type descriptives relatives à la prise de médicaments. Les critères utilisés pour définir le statut sont les suivants [16, 223] : au moins 3 critères sur les 5 pour être considéré comme ayant le SMet :

- Un tour de taille ≥ 102 cm chez les hommes et ≥ 88 cm chez les femmes.
- Une pression artérielle systolique ≥ 130 mm Hg ou une pression artérielle diastolique ≥ 85 mm Hg ou un traitement antihypertenseur.
- Un niveau de glucose sanguin à jeun ≥ 100 mg/dL (5,5 mM) ou un traitement hypoglycémiant.
- Un niveau sanguin de triglycérides à jeun ≥ 150 mg/dL (1,7 mM) ou un traitement hypolipémiant.
- Un niveau d'HDL-C à jeun < 40 mg/dL (1,03 mM) chez les hommes et < 50 mg/dL (1,3 mM) chez les femmes ou un traitement hypolipémiant.

La stabilité à travers le temps du statut défini défini précédemment a été étudié. Les individus pour lesquels le statut a évolué entre T1 et T4 ont été écartés et 208 sujets ont été considérés éligibles.

Des mesures lipidiques (HDL-cholestérol et triglycérides) ont été réalisées pour compléter les mesures quantitatives déjà disponibles pour les 3 premiers critères. Les individus pour lesquels ses mesures sont manquantes à T1 et T4 pour des raisons expérimentales ont été écartés laissant un groupe de 182 sujets disposant des mesures nécessaires à la poursuite de la sélection.

Le statut définitif (SMet ou non) des sujets a ensuite été établi sur la base de ces nouvelles mesures lipidiques quantitatives. Certains sujets possèdent des valeurs manquantes à l'un des 2 temps

T2 ou T3 pour l'un des 5 critères sur SMet, et leur statut à ces points de mesure peut donc être non défini. Si c'est le cas, ils ont été écartés. Seul 161 sujets ont donc un statut établi au 4 temps de mesure.

Les individus atteints du SMet sont définis comme ayant moins de 3 critères parmi les 5, on obtient donc une présélection de 88 témoins et 73 cas incidents. Sur la base des nouveaux critères quantitatifs, les sujets au statut instable entre T1 et T4 ont une nouvelle fois été écartés. Au final un panel de 61 cas a été retenus.

Concernant les 88 individus témoins, il est important d'exclure les sujets extrêmes qui pourraient générer de faux négatifs. En accord avec les cliniciens, les individus prenant plus de 7 médicaments ont été exclus. De plus, les « outliers valeurs » ont été étudiés : les sujets présentant une valeur anormalement élevée ou faible pour une variable donnée aux 4 temps de mesure ont été éliminés. Sont considéré « outliers valeurs » les sujets situés à l'extérieur de l'intervalle définie par la médiane plus ou moins 1.5 fois l'écart interquartile. Ce sont finalement 62 individus témoins qui ont été conservés permettant d'atteindre un équilibre de nombre de sujets entre les groupes cas et témoins.

Une publication visant à valoriser ce travail de sélection a été soumise au journal « Metabolomics » : Lenuzza N, Pétéra M, **Monnerie S**, Morais JA, Payette H, Gaudreau P, Thévenot E, Comte B, Pujos-Guillot E. The importance of an optimized selection process for biomarker discovery in epidemiological studies. Cette dernière est disponible en **Annexe 1**.

3. Caractéristiques des sujets

Les caractéristiques cliniques des 123 individus sélectionnés sont présentées dans le **Tableau 3**. Pour évaluer la significativité des variables descriptives par rapport au statut cas/témoins ainsi qu'au facteur temps, une ANOVA à mesure répétée a été réalisée en prenant également en compte 59 variables complémentaires considérés comme descriptives de la population : 22 variables biochimiques (hémoglobine, albumine…), 8 variables issues d'examens cliniques (poids, IMC…), 25 variables nutritionnelles (essentiellement relatives aux apports en macronutriments) et 4 scores en lien avec l'activité physique.

On observe que les variables cliniques associées au SMet sont plutôt stables dans le temps avec toutefois de légères décroissances significatives pour la pression artérielle systolique, la glycémie

à jeun et les TG à jeun. En revanche les 6 variables cliniques relatives aux 5 critères sont bien significatives du statut (p-value corrigées de 10^{-5} à 10^{-10}).

	Contrôles	Cas	Témoins		Cas		p-value temps corrigée (BH)	p-value statut corrigée (BH)
			T1	T4	T1	T4		
n	124	122	62	62	61	61	-	-
Age (années)			73.5 ± 4.1 (62)	-	74.1 ± 3.6 (61)	-	1.0	0.34
Poids (kg)			71.0 ± 8.0 (62)	69.9 ± 7.8 (62)	87.7 ± 12.5 (61)	87.4 ± 13.3 (61)	**0.04**	**6.2e-14**
IMC (kg/m^2)			25.1 ± 2.3 (62)	24.8 ± 2.4 (62)	30.5 ± 3.7 (61)	30.6 ± 3.7 (61)	0.37	**1.5e-16**
Tour de taille (cm)	93.1 ± 6.9 (124)	110.3 ± 9.2 (122)	93.3 ± 6.9 (62)	92.8 ± 6.9 (62)	109.9 ± 8.9 (61)	110.8 ± 9.5 (61)	0.67	**6.2e-21**
Glycémie à jeun (mM)	4.97 ± 0.52 (124)	6.6 ± 1.3 (122)	5.08 ± 0.44 (62)	4.86 ± 0.58 (62)	6.66 ± 1.45 (61)	6.54 ± 1.21 (61)	**0.04**	**2.2e-15**
TG à jeun (mM)	1.20 ± 0.44 (103)	2.10 ± 0.95 (112)	1.23 ± 0.47 (50)	1.18 ± 0.40 (53)	2.23 ± 1.01 (51)	1.94 ± 0.86 (51)	**0.04**	**1.6e-8**
HDL-C à jeun (mM)	1.47 ± 0.4 (103)	1.15 ± 0.28 (112)	1.43 ± 0.45 (50)	1.50 ± 0.34 (53)	1.13 ± 0.29 (56)	1.16 ± 0.26 (56)	0.74	**1.1e-5**
Pression artérielle systolique (mmHg)	123.6 ± 17.6 (124)	136.0 ± 17.8 (122)	126.2 ± 16.6 (62)	120.9 ± 18.4 (62)	138.4 ± 15.8 (61)	133.7 ± 19.3 (61)	**0.02**	**4.4e-5**
Pression artérielle diastolique (mmHg)	72.8 ± 9.1 (124)	74.2 ± 9.1 (122)	71.8 ± 9.9 (62)	73.9 ± 8.1 (62)	74.7 ± 8.9 (61)	73.62 ± 9.4 (61)	0.69	0.47

Tableau 3 : Caractéristiques des 123 sujets retenus pour l'étude. P-value de l'ANOVA à mesure répétée (après correction BH pour les 59 paramètres) < 0.05.
Moyenne ± Ecart-type (effectif)
Cas et témoins : moyenne (T1-T4)

Il est établi que le profil métabolique est modifié avec l'âge, il a donc été vérifié qu'il n'existait aucune différence significative de l'âge entre les cas et les témoins pour écarter un potentiel biais. Pour cela, un découpage en classes d'âges de l'effectif a été réalisé selon la distribution expérimentale. Trois classes ont été définies : les individus de 65 ans ± 2 ans, 70 ans ± 2 ans et 75 ans ± 2 ans. La balance des tailles des classes ainsi formées a été vérifiée dans chaque groupe par un test exact de Fisher (**Tableau 4**).

Proportions	67 - 72 ans	73 - 77 ans	78 - 84 ans	
Témoins	25	22	15	=62
Cas	22	24	15	=61
p-value (Fisher's exact test		*0,8729*		

Tableau 4 : Répartition des effectifs cas et témoins en fonction des classes d'âges à l'entrée dans l'étude.

Les effectifs des sujets en fonction de la positivité des critères du SMet sont présentés dans le **Tableau 5**. Il est important de noter que la quasi-totalité des sujets cas sont hypertendus. Le critère d'hypo-HDL-cholestérolémie est lui le moins représenté parmi les 5 critères du SMet chez les cas. Il est également important de noter que les sujets témoins ne sont pas tous exempts de critères du SMet : 46% d'entre eux font de l'hypertension. Ceci est lié à la méthode de sélection des sujets pour laquelle nous avons souhaité avoir des cas et des témoins partageant certains critères et non pas des témoins obligatoirement exempts de tous critères. Ainsi, nous avons favorisé la sélection d'un échantillon le plus représentatif possible de la population.

	Tour de taille (cm) à T1	**Tension artérielle systolique (mmHg) à T1**	**Taux de glucose (mM) à T1**	**Taux de TG (mM) à T1**	**Taux de HDL-cholestérole (mM) à T1**
Témoins	5	29	7	7	4
	8,06%	46,77%	11,29%	11,29%	6,45%
Cas	52	59	46	47	24
	85,25%	96,72%	75,41%	77,05%	39,34%

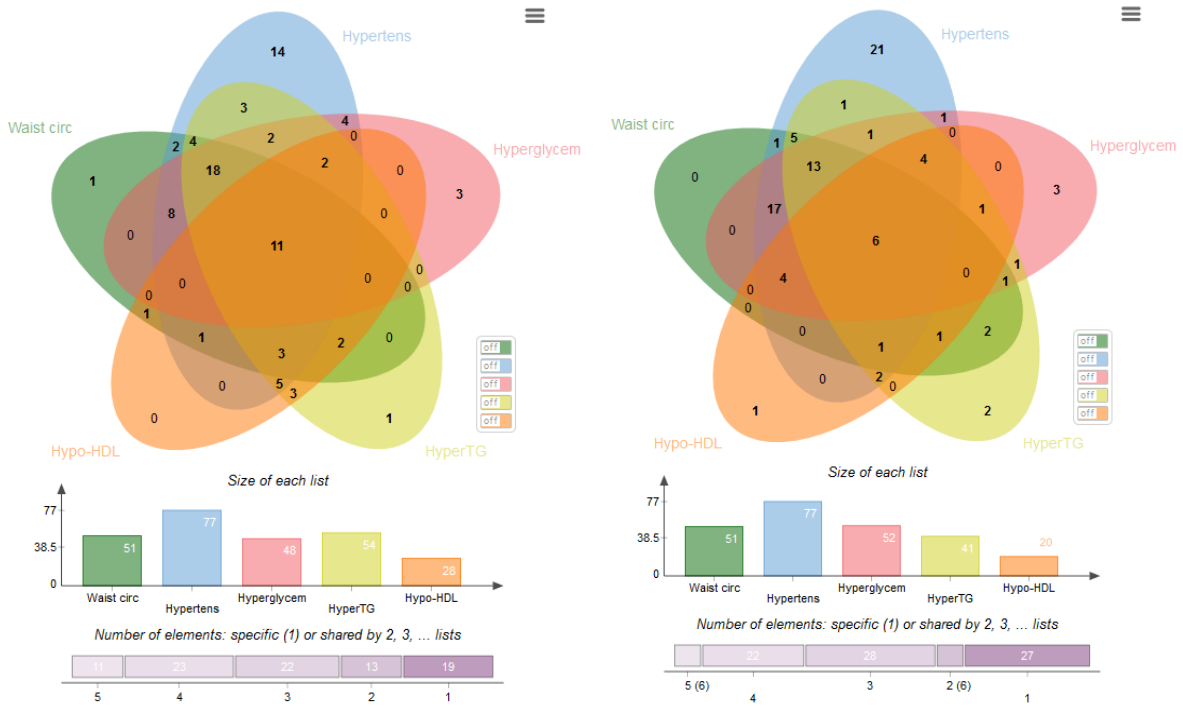
Tableau 5 : Répartition des effectifs en fonction des critères (positifs) du SMet, chez les cas et les témoins.

La sélection des sujets réalisée a donc permis de conserver la diversité des phénotypes qui constituent le SMet avec des cas présentant des nombres de critères variables (>3), positifs pour différentes combinaisons de critères, mais également des contrôles non exempts de critères, par exemple étant positifs pour 2 et pouvant être assez proches d'un basculement vers la pathologie. On peut ainsi parler d'un continuum d'individus au sein desquels la différence entre cas et témoin n'est pas une limite nette. Cela nous permettra de nous intéresser à un potentiel sous-découpage de la population en sous phénotypes. La répartition des individus dépendamment des critères qu'ils possèdent est présenté dans le **Figure 16**. La construction de ces diagrammes de Venn met en évidence la présence de 5 sous-groupes majeurs d'individus au sein de notre échantillon représentant différents sous-phénotypes cliniques du SMet (présent à T1 comme à T4) :

- Les témoins exempts de critères du SMet.
- Les témoins avec hypertension.
- Les cas présentant les 5 critères du SMet.
- Les cas présentant les 4 critères suivants : Hyperglycémie, Hypertension, Tour de taille élevé, hyper TG.

- Et enfin les cas présentant 3 critères : Hyperglycémie, Hypertension, Tour de taille élevé.

La mise en évidence de ces 5 sous-groupes cliniques majeurs soulève la question de l'existence de différences métaboliques entre ces différents groupes d'individus. Cela suggère ainsi qu'il serait pertinent de s'intéresser aux caractéristiques métaboliques de chacun des sous-groupe et de tenter de proposer une reclassification moléculaire du SMet.



T1 T4

Figure 16 : Diagramme de Venn de la répartition des critères du SMet (considérés comme mesure binaire : positif ou négatif) dans les groupes de sujets cas et témoins.

4. Variables disponibles

La cohorte NuAge possède une base de données comportant plus de 5 900 variables répartis aux 4 temps de mesures du suivi. Elles sont réparties en 8 catégories : données sociologiques, nutrition, compositions corporelle, activités physiques, état de santé physique, état de santé cognitive et mentale, états fonctionnels, paramètres biologiques et données autres (fardeau de la maladie,

fragilité, décès et autres commentaires). Une sélection d'environ 2 000 variables a été faite pour permettre de répondre aux différentes problématiques du projet. Pour un total de 123 sujets analysés, cela représente environ 250 000 valeurs. Par conséquent, un état des lieux de ces variables est indispensable avant toute poursuite du projet.

Dans un premier temps, nous avons cherché à savoir s'il existait des variables pour lesquelles des mesures étaient absentes à l'un de nos deux temps d'intérêt (T1 et T4). La **Tableau 6A** présente un aperçu général du nombre de variables disponibles par type, en fonction des différents temps de suivi disponibles. On remarque que les données de type nutritionnelles et de composition corporelle (absorptiométrie à rayon X en double énergie, DEXA) sont présentes dans des proportions très différentes entre T1 et T4 et qu'il risque donc d'être difficile, voire impossible, d'analyser leur évolution dans le temps. Concernant les variables nutritionnelles ceci s'explique par le fait que les questionnaires de fréquence n'ont été réalisé que l'année de l'entrée dans l'étude. En ce qui concerne les données de composition corporelle mesurées *via* la DEXA, les mesures n'ont été faite qu'à T1 et à T3. La question de considérer ces mesures faites à T3 comme un reflet d'un état à T4 s'est posé, cependant, du fait de la possibilité de changements majeurs avec le vieillissement, il a été décidé de ne pas prendre les valeurs de T3 pour substituer les données manquantes.

Le **Tableau 6B** illustre le fait que 392 variables sur les 1 027 présentes à T1 ou T4 sont présentes aux deux temps pour permettre l'analyse de l'évolution des sujets. Dans la catégorie des variables nutritionnelles, 271 variables sur les 821 disponibles permettront une étude de la stabilité du comportement et des habitudes alimentaires. Concernant les données de composition corporelles, comme indiqué plus tôt, aucunes données n'étant disponible à T4, le suivi des individus ne sera donc pas possible.

Catégories	T1/S1	T2	T3	T4	TOTAL
Autres (variables ajoutées)	31	18	18	21	88
Socio-démographie	8	2	3	3	16
Nutrition	812/189	94	93	280	1468
Composition anthropométrique	26	18	18	18	80
Composition corporelle (DEXA)	42	0	42	0	84
Activité physique actuelle (PASE)	1	1	1	1	4
Etat de santé physique	64	62	62	62	250
Etat de santé mentale	4	4	4	4	16
Paramètres biologiques	22	1	1	20	44
TOTAL	1009 / 199	200	242	409	**2060**

A : Nombre de variable par catégorie et par temps de suivit

Catégorie	Uniq T1	T1 et T4	Uniq T4	TOTAL
Autres (variables ajoutées)	10	21	0	31
Socio-démographie	5	3	0	8
Nutrition	541	271	9	821
Composition anthropométrique	10	16	2	28
Composition corporelle (DEXA)	42	0	0	42
Activité physique actuelle (PASE)	0	1	0	1
Etat de santé physique	8	56	6	70
Etat de santé mentale	0	4	0	4
Paramètres biologiques	2	20	0	22
TOTAL	618	392	17	1027

B : Nombre de variable par catégorie en fonction de leur présence à T1 seul, T4 seul ou au 2 temps confondus.

Tableau 6 : Etude des catégories de variables et de leur nombre dépendamment du temps de mesure.

Sur les 253 380 valeurs renseignées pour les 2 060 variables choisis, tous temps de mesure confondus, 19 445 valeurs sont identifiées comme « manquantes » (égales à « NA ») et 26 787 sont « absentes » (égales à « - ») (c'est-à-dire qu'il est normale de n'avoir aucune valeur, souvent à cause de questionnaires où les questions sont imbriquées, la réponse à une première question conditionnant la réponse ou non à une seconde). Sur les 2 060 variables disponibles, 40 d'entre elles ne possèdent aucune valeur. A l'exception de 6 variables, toutes possèdent essentiellement des valeurs absentes attendues. Les 6 variables ne possédant que des « NA » sont en fait des équivalences en nombre de portions de groupe alimentaires non renseignés pour 2 types de portions sur 10 normalement disponibles.

La **Figure 17** représente le pourcentage de variables retenus en fonction d'un seuil de valeurs manquantes de type « NA » maximum évoluant de 0 à 100%. On constate que 34% des variables disponibles (701 sur les 2 060) sont exemptes de valeurs manquantes. Toutefois, 84% des variables présentent moins de 10% de valeurs manquantes. Il est également important de remarquer que presque 7% des variables possèdent entre 40 et 100% de valeurs manquantes, les rendant difficilement exploitables.

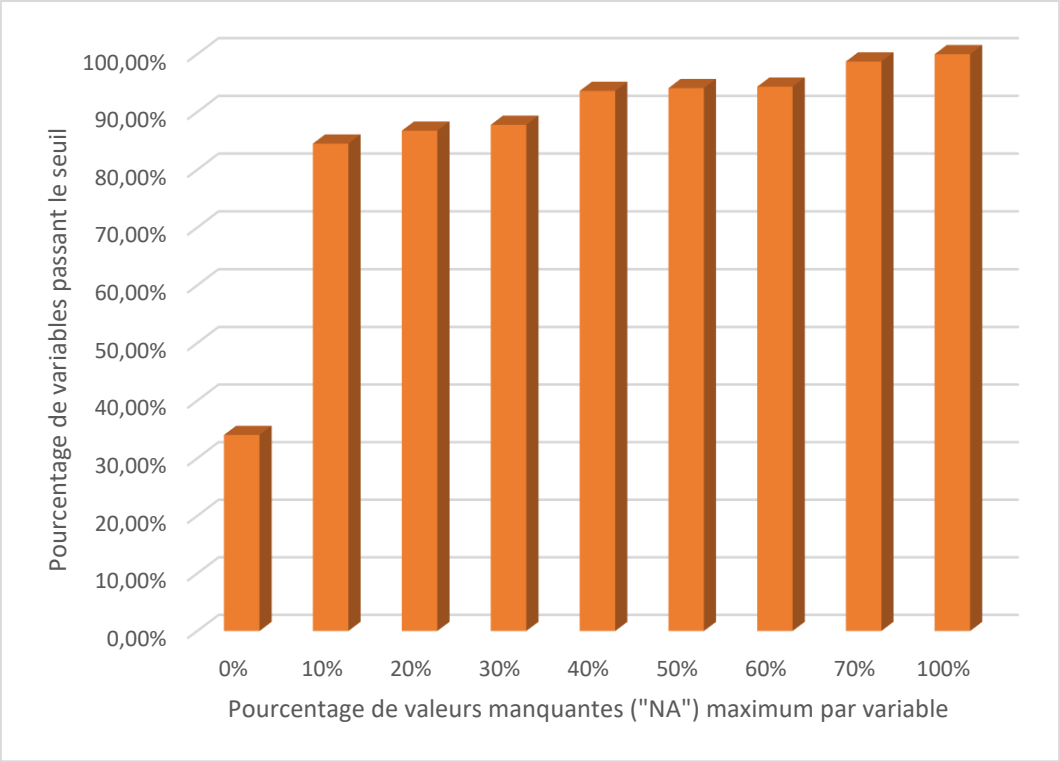


Figure 17 : Pourcentage de variables en fonction d'un seuil de valeurs manquantes (« NA ») maximum par variable évoluant de 0 à 100%.

Pour aller plus loin dans l'identification des valeurs manquantes, nous nous somme intéressée aux variables de manière catégorielle (nutritionnelles, paramètres biologiques, santé physique…) afin de mieux appréhender les catégories touchées par le manque d'information et plus particulièrement les valeurs manquantes « NA ». La **Figure 18** représente le pourcentage de valeurs manquantes en fonction de la catégorie de variables ; la **Figure 19** représente le pourcentage de variables possédant moins de 10% de valeurs manquantes en fonction de la catégorie de variables. Ce seuil correspond à un taux idéal de valeurs manquantes pour pouvoir traiter les données ; il reste acceptable d'avoir un pourcentage plus élevé dépendamment de la problématique étudiée, mais un seuil de 10% permet un bon aperçu des variables facilement analysables en comparaison des autres. On constate sur la **Figure 18** que les variables cliniques, du C-HEI à partir de rappel de 24h et de DEXA sont les plus touchées par les valeurs manquantes « NA ». La fixation d'un seuil de valeurs manquantes inférieur à 10% révèle sur la **Figure 19** la difficulté d'exploitation des données de DEXA et d'habitudes alimentaires pour lesquelles le pourcentage de variables possédant moins de 10% de valeurs manquantes est faible.

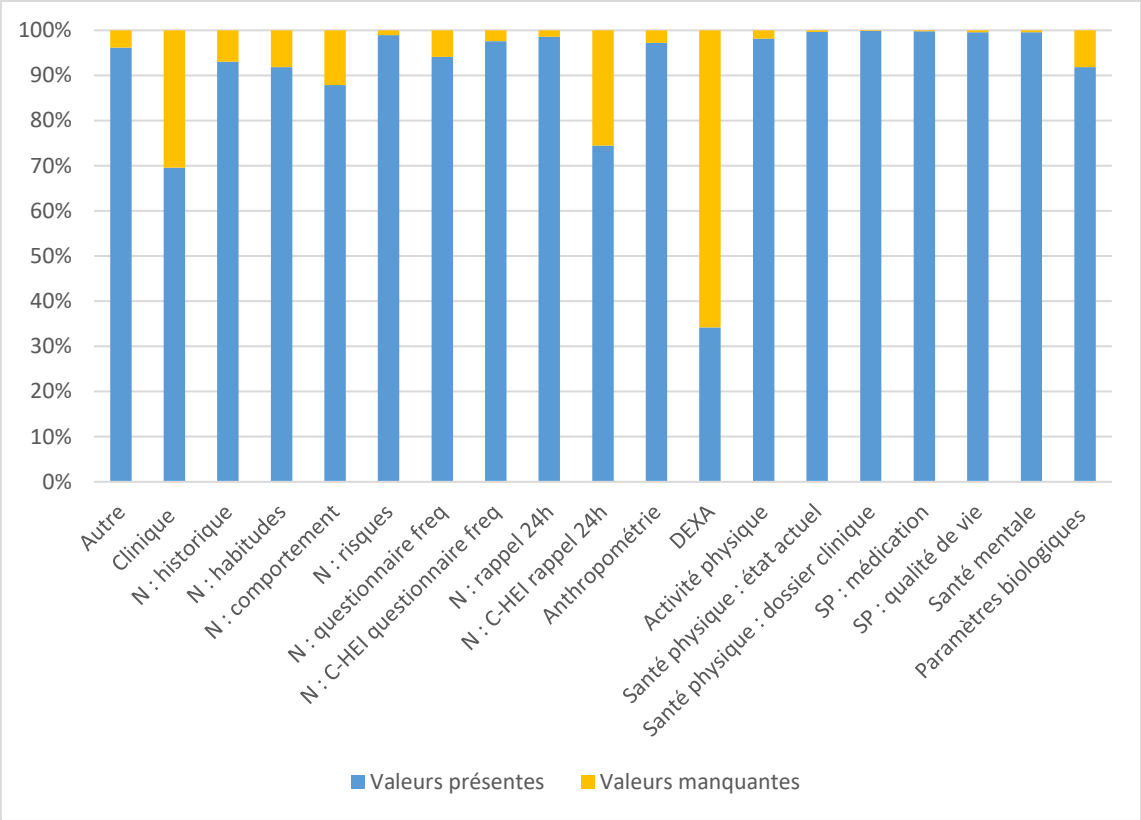


Figure 18 : Pourcentage de valeurs manquantes en fonction de la catégorie de variables

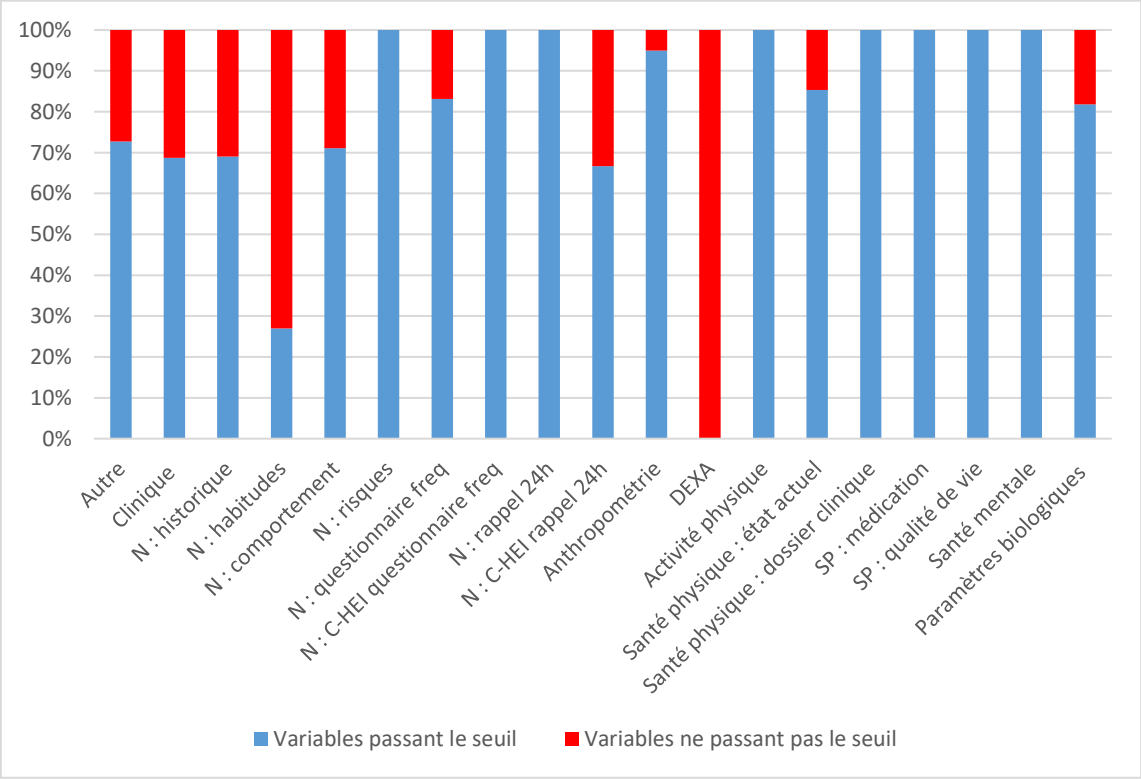


Figure 19 : Pourcentage de variables possédant moins de 10% de valeurs manquantes en fonction de la catégorie de variables

Au cours de ma thèse les principales variables de la cohorte NuAge qui seront traité (outre les variables relatives aux critères) sont l'ensemble des 59 variables décrites précédemment lors de la caractérisation des sujets. Le reste des variables sera traité dans une seconde phase du projet qui interviendra après la fin de mon travail de thèse.

III. Les données métabolomiques et lipidomiques

Le projet dans lequel s'inscrit mon travail de thèse implique comme second acteur majeur l'Infrastructure MetaboHUB, chargée de réaliser les analyses métabolomiques et lipidomiques sur les échantillons de sérum de la cohorte NuAge.

1. L'infrastructure Française d'excellence MetaboHUB

Les enjeux scientifiques actuels, que ce soit dans le domaine de la santé, de la nutrition, de l'agronomie, de la microbiologie, des biotechnologies, de l'écotoxicologie ou encore de l'environnement amènent une demande croissante d'identification et de quantification des métabolites à grande échelle. L'objectif est souvent de produire des représentations des fonctionnements dynamiques et des réseaux biologiques dans lesquels les métabolites sont impliqués dans le but d'expliquer et d'optimiser ces processus biologiques. Pour les équipes de recherche, il est donc devenu nécessaire d'avoir accès à des structures regroupant les expertises requises pour l'acquisition, le traitement et l'intégration des données dans une approche de biologie des systèmes et capables de prendre en charge des projets de grande ampleur. C'est donc dans ce contexte que MetaboHUB a vu le jour en 2013 dans le cadre d'un programme « Investissement d'Avenir ».

L'Infrastructure Française d'excellence MetaboHUB a donc pour objectif de fournir des outils technologiques de pointe et des services en métabolomique et fluxomique aux équipe de recherche académiques et à des partenaires dans les domaines de la santé, de la nutrition, de l'agriculture, de l'environnement et des biotechnologies. Pour parvenir à cet objectif, elle rassemble des outils analytiques complémentaires (équipements, techniques analytiques, outils de traitement de données) au sein de 4 plateformes localisées à Paris-Saclay, Bordeaux, Toulouse et Clermont-Ferrand. Elle a pour ambition de développer des méthodes de pointe en métabolomique et en fluxomique afin de relever les défis technologiques rencontrés dans les domaines scientifiques cités précédemment.

2. Des techniques d'analyses multiples

Les métabolites mesurables au sein des échantillons représentent une large gamme de molécules de propriétés physico-chimiques variées, et présentes à des concentrations différentes. De ce fait, il est primordial de disposer de techniques d'analyses complémentaires permettant de garantir une bonne couverture analytique des différentes espèces chimiques présentes. La **Figure 20** illustre les techniques d'analyse en fonction de la nature des métabolites étudiés et de l'échantillon (nature, concentration du métabolite au sein de ce dernier).

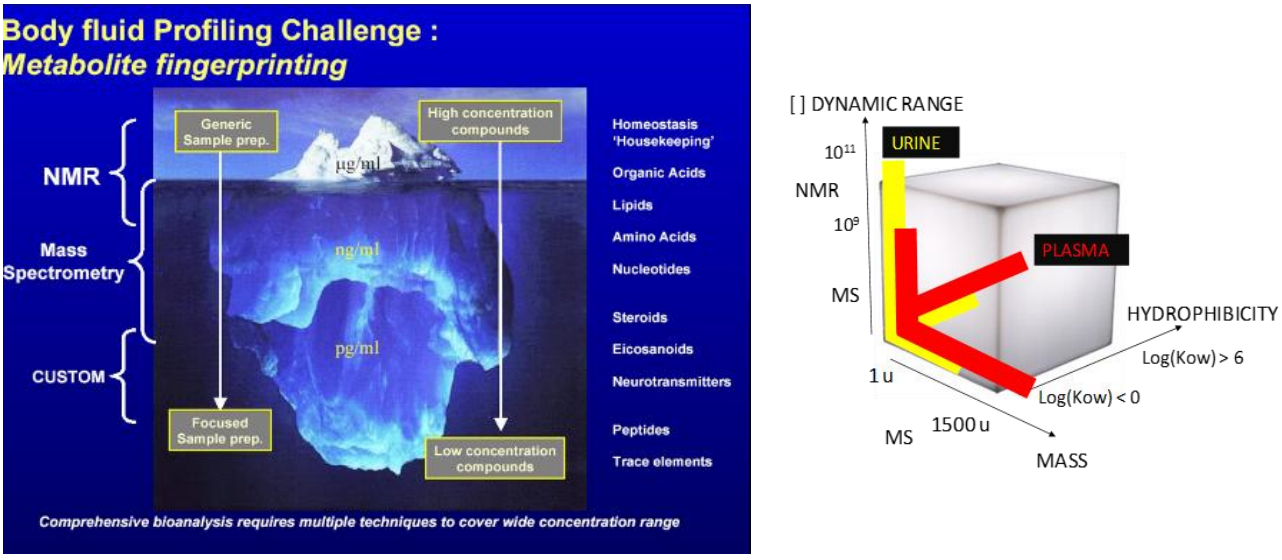


Figure 20 : Répartition des techniques d'analyse métabolomique en fonction de la nature chimique des composés étudiés.

L'existence de 4 sites au sein de l'infrastructure MetaboHUB nous a permis d'avoir accès à plusieurs types de technologies d'analyse. Les échantillons ont donc été analysés selon 6 méthodes différentes et complémentaires, et ce grâce à 3 des 4 sites de l'instance nationale : Clermont-Ferrand, Saclay et Toulouse. Ces 6 méthodes d'analyse font appel à des couplages LC/MS, GC/MS et à la RMN. On peut les diviser en 2 approches : métabolomique et lipidomique. En effet, la lipidomique est une sous-section de la métabolomique. Comme cela a pu être discuté dans la revue de Cajka et Fiehn [44], il existe un continuum de polarités entre les métabolites lipophiles et hydrophiles. Pour couvrir cette large diversité de molécules, des méthodes spécifiques en termes de préparation des échantillons et d'analyse ont été développées donnant lieu à l'apparition de deux domaines que sont la métabolomique et la lipidomique. Actuellement les 2 communautés scientifiques sont relativement

séparées, notamment au niveau des valorisations par publication, mais elles restent complémentaires en termes de résultats et il est tout à fait pertinent de coupler ces 2 approches dans un projet tel que le nôtre.

La RMN étant une technologie non destructive, rapide et robuste, elle offre la possibilité d'analyser les échantillons avec un prétraitement minimal pour détecter les métabolites de grande taille [163]. En complément, des méthodes d'analyse par spectrométrie de masse ont été utilisées. Elles permettent une analyse plus sensible des échantillons pour la détection d'espèces chimiques minoritaires en utilisant des méthodes chromatographiques liquide et gazeuses complémentaires [164]. Un premier couplage de type MS haute résolution et chromatographie en phase inverse (C18-LC/HRMS) avec acquisition en mode d'ionisation positif, puis négatif a rendu possible la détection d'une large gamme de métabolites polarisés, donnant lieu à l'acquisition de deux jeux de données distincts dépendant du mode d'ionisation. Un second couplage MS et chromatographie d'interaction hydrophile (HILIC/MS) a mesuré les métabolites de polarité intermédiaire à élevée. Un couplage MS et chromatographie en phase gazeuse (GC/MS) a permis la détection des métabolites volatiles de faible poids moléculaire. Enfin, un couplage MS haute résolution (Orbitrap) et chromatographie liquide à ultra-haute performance, en mode d'ionisation positif et négatif, a offert la possibilité de détecter des lipides appartenant à 5 grandes familles (acides gras libres, sphingolipides, glycérophospholipides, glycérolipides, et stérols et dérivés).

Les informations techniques relatives aux 6 méthodes utilisées ainsi que les noms des jeux de données sont détaillés dans le **Tableau 7**.

Nom du jeu de données	Technique d'analyse	Plateforme	Appareil de mesure
C18-Pos	Chromatographie liquide en phase inverse/MS	Theix	Quadripôle temps de vol (QToF Bruker), équipé d'une source « electrospray »
C18-Neg	Chromatographie liquide en phase inverse/MS	Theix	Quadripôle temps de vol (QToF Bruker), équipé d'une source « electrospray »
GC	Chromatographie en phase gazeuse/MS	Theix	QToF, équipé d'une source à impact électronique
HILIC	Chromatographie d'interaction hydrophile/MS	Saclay	Orbitrap Fusion, équipé d'une source « electrospray »
Lipido	Lipidomique globale par chromatographie liquide à ultra-haute performance/MS haute résolution	Saclay	Orbitrap Fusion, équipé d'une source « electrospray »
RMN	Résonnance magnétique nucléaire	Toulouse	Bruker Avance III HD spectrometer, opérant à 600.13 MHz (^{1}H) et équipé avec une détection inverse 5 mm CQPCI ^{1}H-^{31}P-^{13}C-^{15}N cryoprobe

Tableau 7 : Détails des méthodes d'analyse utilisées par les 3 plateformes MetaboHUB et identification des jeux de données.

3. Protocole d'analyse commun (randomisation des échantillons et contrôles qualité)

Pour permettre une analyse commune des résultats issus des différentes plateformes et méthodes d'analyse, il est important d'uniformiser au maximum les étapes de préparation et d'analyse des échantillons. La préparation des échantillons est restée dépendante de chacune des méthodes du fait des contraintes analytiques souvent très différentes. Toutefois, l'ordre de passage des échantillons sur les machines, les contrôles qualité et le prétraitement des données en sortie ont été uniformisé au maximum.

3.1. Randomisation des échantillons

La randomisation permet de s'assurer que des critères fortement impactant sur le statut cas/témoins ne soient pas confondus avec la variabilité analytique. L'étape de randomisation est limitée par le nombre de facteurs utilisables, le nombre de sujets et le nombre de groupes. Dans notre cas, elle a été faite *via* une stratégie basée sur les carrés latins de Williams. Le facteur principal considéré est celui de l'étude (le SMet) par le biais de la somme du nombre de critères aux différents temps de mesure (subdivisé en 4 groupes : 0-3 ; 3-7 ; 11-14 ; 15-20). Ce choix est dû à la variabilité importante dans le nombre de critères au sein des 2 groupes cas et témoins. Il est intéressant de considérer cette hétérogénéité au regard de la capacité de la métabolomique à révéler les sous-phénotypes potentiels. D'autre part, la randomisation prend en compte un facteur pouvant avoir un potentiel impact sur les données métabolomiques : la région de prélèvement (Montréal, Sherbrooke ou Laval). La date de prélèvement a également été considérée (4 groupes correspondant aux 4 saisons), mais non prise en compte directement dans le processus de randomisation : les échantillons ayant été collecté tout au long des saisons, cela peut avoir un effet sur les données métabolomiques notamment dû aux variations dans l'alimentation.

L'ordre établi suite à cette randomisation a servi à la fois lors de la préparation des échantillons et des séquences analytiques des analyses métabolomiques et lipidomiques.

3.2. Uniformisation des contrôles de qualité

Tout comme l'étape de randomisation des échantillons, les contrôles qualité (QC) effectués lors des analyses ont été uniformisés. Ils ont été réalisés à l'aide d'une série de dilutions (8 ; 4 ; 2 ; 0) du pool des échantillons. Les QC sont injectés dans l'ordre de la plus forte à la plus faible dilution, chaque QC dilué étant injecté 3 fois de suite sur la machine avant de passer au suivant. Les signaux détectés ayant un coefficient de corrélation avec la dilution > 0.7 seront conservés, la qualité des mesures est considérée comme étant bonne.

3.3. Extraction et prétraitement des données

L'extraction des données de spectrométrie de masse a été faite en utilisant l'outil XCMS disponible dans l'instance Galaxy Workflow4Metabolomics. L'un des paramètres primordiaux lors de cette extraction concerne l'étape d'alignement des ions entre échantillons. Il s'agit du paramètre appelé « minfrac » : il définit la proportion minimale d'échantillons dans laquelle doit être retrouvé un pic pour être conservé. Ce dernier a donc été définit de manière uniforme pour toutes les méthodes et fixé à une valeur de 0.2 signifiant qu'un pic doit être retrouvé dans au minimum 20% des échantillons.

La dernière étape avant l'analyse des données, le prétraitement, a également été uniformisée. Elle a notamment pour but de retravailler les données pour écarter les informations incohérentes ou non utiles, qui peuvent être dues à l'analyse, et ne doivent pas être interprétées par la suite comme des variations biologiques. Pour commencer, les signaux présents dans les blancs de façon équivalente ou plus intense que dans les échantillons biologiques sont écartés (les ions dont l'intensité moyenne n'est pas au moins 3 fois supérieure à celle trouvée dans les blancs ne sont pas conservés, paramètre « mean fold-change »). Ensuite, l'effet potentiel des batchs d'analyse a été corrigé en uniformisant les intensités obtenues dans chacun d'entre eux. L'étape suivante consiste en l'étude des coefficients de variation (CV). D'une part les CV des pools doivent être inférieurs à 0,3, d'autre part, le CV des pools ne doit pas être supérieur à celui des échantillons. Le filtre sur la base des corrélations aux dilutions de pools est également appliqué. Finalement, il a parfois été nécessaire d'appliquer un filtre d'intensité moyenne minimum pour écarter certaines variables difficilement mesurables, le seuil alors choisi est dépendant de l'appareil d'analyse.

Les données de RMN ont elle été importées dans le logiciel constructeur Amix (version 3.9.15, Bruker, Rheinstetten, Germany) pour réaliser l'intégration. Un bucketing de taille variable a été réalisé en se basant sur des motifs graphiques et chaque bucket a ensuite été intégré.

4. Variables métabolomiques

Une fois les données de métabolomique et lipidomique générées et prétraitées, comme pour les données issues de la cohorte NuAge, une synthèse des données était indispensable. Elle a porté aussi bien sur les corrélations entre jeux de données, que sur les individus et échantillons présents dans les données, ou le comportement des variables mesurées.

4.1. Individus manquants

Nous avons constaté que les analyses avaient conduit à la suppression de certains échantillons dans certains jeux de données car ces derniers étaient manquant ou de mauvaise qualité (problème de préparation, ou d'analyse). La suppression de ces échantillons aura un impact sur les analyses qui seront faites par la suite. En effet, la concaténation des différents jeux de données issues des différentes méthodes d'analyse, requiert des données pour chacun des échantillons présents dans chaque jeu de données. Les échantillons disparus/éliminés dans certains jeux devront donc être écartés des autres jeux, conduisant *in fine* à une diminution du nombre total d'échantillons. De plus, lors d'analyse statistique prenant en compte le facteur temporel des données (exemple ANOVA à mesure répétée), il est indispensable que la donnée soit présente aux 2 temps de mesure. Par conséquent, lorsqu'un échantillon est écarté par le prétraitement, l'échantillon apparié par le temps à ce dernier ne devra pas être pris en compte lors de ces analyses.

Les échantillons écartés dépendamment de la méthode d'analyse sont présentés dans le **Tableau 8**. Au total, 11 échantillons ont été écartés dans au moins l'une des méthodes d'analyse, ce qui représente 4.5 % des échantillons analysés.

Méthode d'analyse	Echantillons écartés
C18-Pos	GP094_T1
C18-Neg	
GC	GP106_T4
HILIC	GP043_T1 ; GP046_T4 ; GP048_T1 ; GP063_T1 ; GP094_T1 ; GP119_T1
Lipido	GP023_T4 ; GP094_T1 ; GP096_T4 ; GP099_T4
RMN	

Tableau 8 : Présentation des échantillons écartés dépendamment de la méthode d'analyse

4.2. Corrélations entre les jeux de données

Dans un premier temps, nous avons réalisé un état des lieux des corrélations entre les 6 jeux de données. Le couplage des 6 méthodes doit normalement permettre de couvrir un large éventail de métabolites, avec potentiellement des recoupements. Etant donné qu'à cette étape, il n'est pas envisageable d'annoter l'ensemble des variables, nous avons, pour étudier cela, réalisé une analyse des corrélations entre les variables issues de tous les jeux de données après une suppression des redondances analytiques (détails de la méthode de suppression de ces redondances données dans le chapitre suivant). Les jeux de données étudiés représentent 2 915 variables : 1 091 issus du C18-Pos, 249 du C18-Neg, 612 de l'Hilic, 571 de la Lipidomique, 345 de la GC et 47 de RMN. Pour se faire, nous avons utilisé un outil de calcul de corrélation de Spearman (Between-table correlation, disponible dans W4M) calculant la significativité de chacune des valeurs de corrélation 2 à 2, et appliquant ensuite une correction de type BH. Seul 7% des corrélations sont supérieures à 0.7. Ces variables sont représentées dans le réseau de corrélation de la **Figure 21**. 30% de ces corrélations sont > 0.9, suggérant des corrélations de type analytiques plutôt que biologiques (métabolites impliqués dans des voies communes). Ce réseau illustre parfaitement la complémentarité entre les différentes méthodes de par le faible nombre de variables corrélées entre des méthodes d'analyse différentes.

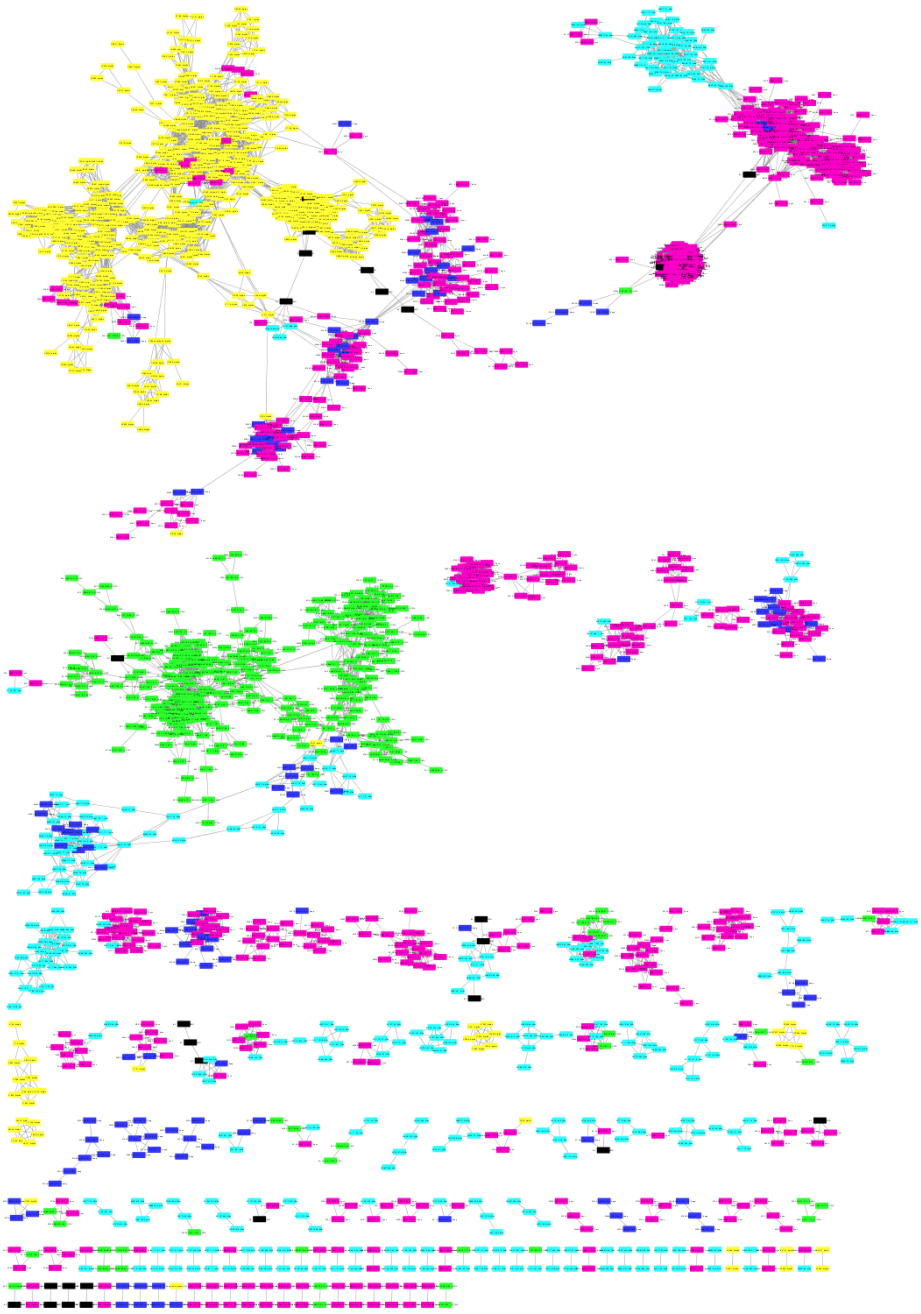


Figure 21 : Réseaux des corrélations > 0.7 entre les 6 méthodes d'analyse. Bleu foncé = C18-neg ; Rose = C18-pos ; Vert = GC ; Bleu clair = Hilic ; Jaune = Lipido et Noir = RMN.

Afin de valider le fait que les méthodes d'analyse sont bien complémentaires les unes des autres nous avons réalisé une Analyse Factorielle Multiple (AFM). Elle a pour but de comparer entre eux des blocs de données, ici nos 6 méthodes. Le coefficient RV permet de traduire la similarité entre ces blocs, il est ici au maximum de 0,21 (**Tableau 9**).

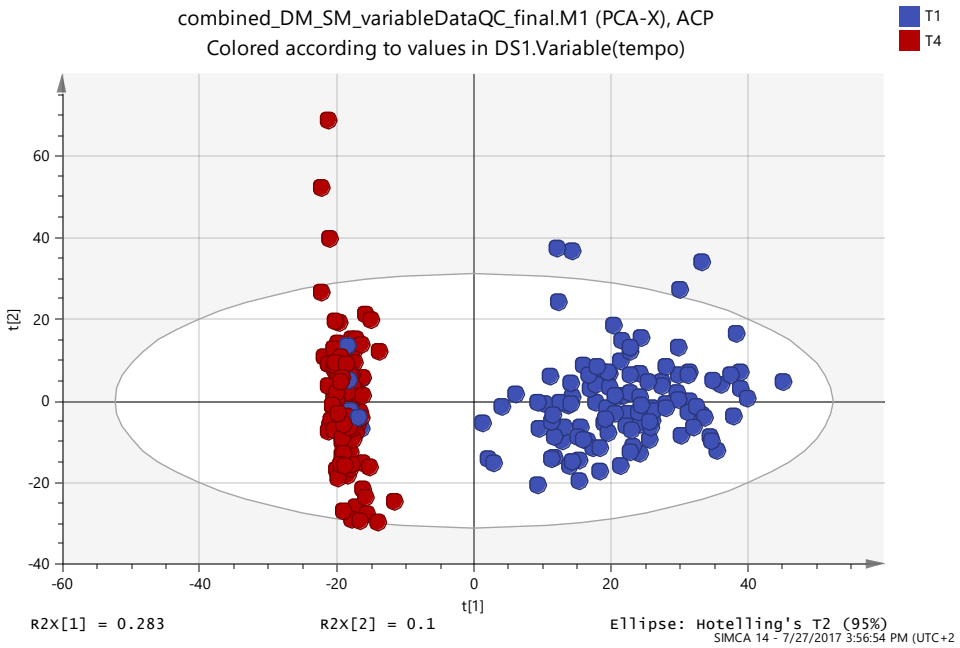
blocks	RV.HILICneg	RV.LIPIDO	RV.RMN	RV.C18neg	RV.C18pos	RV.GCMS
HILICneg	1	0.17	0.11	0.11	0.18	0.07
LIPIDO	0.17	1	0.18	0.11	0.14	0.04
RMN	0.11	0.18	1	0.04	0.09	0.02
C18neg	0.11	0.11	0.04	1	0.21	0.02
C18pos	0.18	0.14	0.09	0.21	1	0.03
GCMS	0.07	0.04	0.02	0.02	0.03	1

Tableau 9 : Similarité entre les blocs de méthodes évaluée par l'intermédiaire du coefficient RV lors de l'AFM.

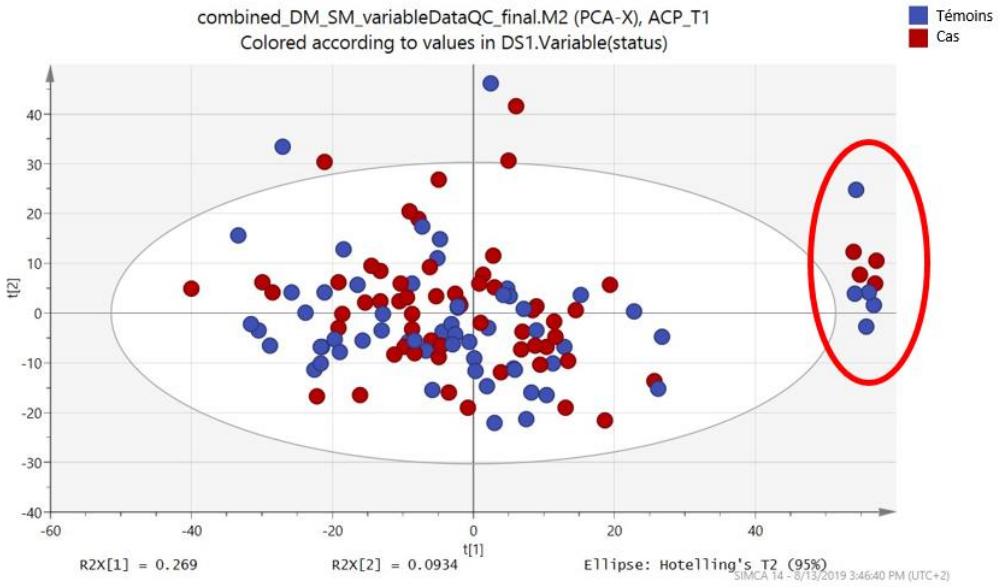
4.3. Analyse des variables

Par la suite, afin d'obtenir un aperçu du comportement des données de métabolomique et lipidomique, des analyses multivariées ont été réalisées. L'objectif est ici de mettre en évidence des facteurs pouvant être par exemple très lié au statut cas/témoins ou qui pourraient engendrer de potentiels biais dans nos analyses s'ils ne sont pas pris en compte. Cet état des lieux a été fait à l'identique pour toutes les méthodes d'analyse à l'aide du logiciel SimCA®.

Tout d'abord, des ACP ont été réalisées sur les données à T1 et T4 réunis, puis à T1 et T4 séparément, et ceux pour chacun des jeux de données. Des colorations ont été appliquées aux ACP afin de détecter de potentiels facteurs d'influence : statut, temps, critères du SMet, lieu et date de prélèvement, IMC, âge… Ces analyses n'ont révélé aucun effet majeur, excepté pour les données de C18-Pos. Un effet de séparation très marqué a pu être observé sur la composante 1 et associé au paramètre du temps (**Figure 22A**). Il semblerait que cet effet soit dû à des différences fortes au niveau de certaines variables lipidiques. Toutefois, ce type d'effet n'ayant pas été retrouvé dans les autres jeux de données (notamment le jeu Lipidomique), on ne peut pas l'associer à une dégradation des lipides au cours du temps dans les échantillons. Il pourrait s'agir d'un biais d'analyse ou d'un effet spécifique lié par exemple à des facteurs alimentaires et traduit uniquement sur au niveau des lipides mesurés en C18-Pos. De plus, quelques échantillons prélevés à T1 semblent être atypiques, se plaçant parmi les échantillons T4. Le caractère atypique de ces individus est d'autant plus visible lorsque l'on regarde l'ACP réalisée à T1 (**Figure 22B**). Malheureusement nous n'avons pas trouvé d'explication particulière à ce comportement à partir des données épidémiologiques disponibles.



A : ACP T1+T4 (coloration dépendante du temps)



B : ACP à T1 (coloration dépendante du statut cas/témoins)

Figure 22 : ACP sur les données de C18-Pos à T1 et T4 confondus (A), puis à T1 (B).

Dans un second temps, des PLS et des O-PLS ont été réalisées. Ces dernières ont été effectuées sur : toutes les variables liées au statut des individus (les 5 critères du SMet), l'âge, la condition physique (IMC, score PASE…), les conditions de prélèvement (lieux, saison) et de traitement des

échantillons (batch). Aucun facteur influençant n'a été détecté, la plupart des modèles n'étant pas valides (qualité du modèle faible et test de permutation non concluant).

IV. Le management des données, la mise en place d'un système d'information : OmicM

Ma thèse et le cas pratique que représente le projet dans lequel elle s'intègre ont permis de soulever des problématiques de stockage et gestion des données générées. En effet, la complexité et la quantité de données rapidement accumulées au début de ma thèse, m'ont poussées à me rapprocher de mes collègues informaticiens et bio-informaticiens afin de discuter de la possible mise en place d'un système de gestion. En effet, dans un projet visant à croiser de nombreuses données de sources et formats différents comme c'est le cas ici, il est raisonnable de vouloir mettre en place un tel système. Ce dernier doit impérativement être sécurisé, sauvegardé, accessibles aux utilisateurs, sachant gérer des données hétérogènes de par leur provenance et leur format et surtout requêtable/interrogeable afin de pouvoir potentiellement en extraire des sous-jeux de données.

Un tel système permet d'éviter la gestion de données au travers de fichiers à plat comme des fichiers Excel®. En effet, cette pratique entraine souvent des erreurs, notamment lors d'une gestion de projet à long terme. Les différentes versions d'un fichier vont s'accumuler au fur et à mesure du rajout d'informations. De plus, il n'est pas rare que 2 utilisateurs travaillent en parallèle sur 2 copies du fichier entrainant souvent la perte d'informations lors de la mise en commun. Posséder un système unique, centralisant les données et accessible à tous les utilisateurs impliqués dans le projet permet donc de limiter ce type de biais.

Dans cet objectif, un entrepôt de données (OmicM) a été développé par l'équipe d'informaticiens et bio-informaticiens de la plateforme d'exploration du métabolisme de Clermont-Ferrand/Theix (PFEM). Dans la mise en place de ce système, j'ai participé à l'élaboration du cahier des charges, ainsi qu'au formatage des données d'entrée lorsque nécessaire.

OmicM doit permettre la gestion des données métabolomiques produites par la plateforme, comme des données scientifiques (informations sur les sujets d'étude, mesures autres que métabolomiques, résultats d'analyses statistiques…). L'objectif est de structurer les processus de génération de données au sein de la plateforme, d'avoir une gestion plus fine des volumes d'informations produits, de pouvoir calculer des indicateurs et des comptes rendus sur les données générées et de faciliter le suivi des workflows d'analyse, et enfin de permettre le stockage à long terme

des projets épidémiologiques impliquant de la métabolomique, afin de pouvoir réutiliser et enrichir la connaissance produite.

1. Des données d'origines très variées

Les données devant être présentes dans le système d'information sont par définition très variées, introduisant une première complexité dans la gestion de l'information :

- Les données de métabolomique et lipidomique issues des analyses par spectrométrie de masse et RMN. Il s'agit à la fois des données brutes (fichiers raw) obtenues des analyses sur les machines, et des données aux différentes étapes de prétraitement. Ces dernières sont des données orientées sur les échantillons, avec pour un ion mesuré des valeurs d'intensité propres à chaque échantillon (2 échantillons par sujet). De plus, chaque ion détecté par le prétraitement possède également des données qui lui sont propres comme sa masse, son temps de rétention, etc…
- Les données « descriptives » issues de la base de données NuAge. Il s'agit des variables nutritionnelles, cliniques, de capacité fonctionnelles, etc… fournies pour chacun des sujets. Ces dernières doivent être subdivisées en sous bases pour faciliter l'accès à certaines catégories de données par exemple, les données nutritionnelles ou d'activité physique.
- Les résultats (p-value et indicateurs de filtration par exemple) des différentes analyses statistiques menées dans le cadre du projet pour aboutir à la signature finale du SMet ou à l'étude de ces sous-composantes.
- Les données issues des bases de données extérieures. Il s'agit des données d'annotation, des potentiels liens vers des bases de données externes, des informations de la base de données des biomarqueurs de la littérature créée suite à l'écriture de la revue systématique…

Certaines données sont disponibles à l'échelle de l'individu : un sujet possède un âge, un lieu de naissance… Ce même individu peut posséder des informations temporalisées, par exemple sa glycémie à T1 et à T4, on a donc les valeurs d'une même variable à plusieurs temps. D'autre part, chaque individu possède 2 échantillons sanguins pour lesquels des mesures ont été faites. Ces valeurs mesurées sur les échantillons sanguins sont également temporalisées (présentes à T1 ou T4). La mise

en relation de toutes ces informations nécessite donc la construction d'une architecture informatique pertinente prenant en compte ces différents aspects.

Devant la quantité très importante de variables et leur hétérogénéité, le système d'information doit pouvoir fournir à ses utilisateurs une description des variables disponibles. A cet effet, un dictionnaire de variables a été réalisé. Ce dernier contient le nom de toutes les variables ainsi que leur définition en français et en anglais, l'unité de mesure si existante et le type de variables (métadonnée, qualitative, quantitative). Pour que le système d'information soit le plus pertinent possible, ce dictionnaire doit pouvoir être implémenté en permanence dès l'ajout de nouvelles variables dans le système d'informations, que ce soit de manière manuelle par l'utilisateur *via* l'interface ou par l'ajout d'un fichier de nouvelles variables venant s'ajouter aux précédentes.

2. Mise en place de différents indicateurs

Un tel système de gestion permet la mise en place de différents indicateurs graphiques ou numérique : taille des données générées, nombre de variables sauvées dans la base, pourcentage et nature des valeurs manquantes, nombre d'échantillons, de sujets, traçabilité des actions utilisateur, historique des requêtes… Ces indicateurs peuvent être développés au cours de la vie du système pour coller au mieux aux besoins utilisateurs.

3. Formatage des fichiers d'entrée

Afin d'intégrer les différentes données au sein d'OmicM, il est nécessaire de passer par l'ajout de fichiers au format préétabli et respectant certaines règles de lecture. En effet, les données qui ont par exemple été fournis par la cohorte NuAge possèdent leur propre formatage.

3.1. Anonymisation des sujets

La première des choses à faire concernant la prise en charge des données issues de la base de données NuAge a été de créer une table de conversion des identifiants afin d'anonymiser totalement

les échantillons. En effet les données transmises par les gestionnaires de la cohorte comportaient un identifiant unique par individu. Afin de sécuriser les données et de rendre impossible la correspondance entre les données générées dans le cadre du projet et les données déjà disponibles dans la base de données Québécoise pour des personnes extérieures au projet, ces identifiant ont été remplacés et une table de correspondance a été créée. Celle-ci est conservée dans un lieu de stockage différent de celui des données et ce de manière sécurisée.

3.2. Reformatage et script de curation

Au sein des données NuAge figurent des variables exprimées en unités nord-américaines (taille en pouce, poids en livres…). Il a donc fallu convertir toutes ces données en unités internationales. De plus, il a également fallu prendre en charge la protection des variables texte par des guillemets, les valeurs manquantes ou encore l'encodage des fichiers pour conserver certains caractères spéciaux.

Pour permettre la mise en forme de ces données (curation, uniformisation…), j'ai mis au point un script simple codé en langage Perl. Ce dernier sépare les valeurs par des « ; » ou des tabulations selon le choix de l'utilisateur. Il protège également les variables textuelles par des guillemets. Afin de ne pas supprimer des caractères comme les accents présents dans les phrases de réponse aux questionnaires rédigées en français, il encode le fichier au format utf8. Il vérifie la cohérence du dictionnaire de variables (vérifie que toutes les variables présentes dans le fichier de donnée soient bien présentes dans le dictionnaire), étape qui devrait par la suite être directement géré par OmicM lors de l'import.

Enfin, il a fallu différencier les valeurs manquantes des valeurs absentes. En effet, il arrive, pour certaines variables, que ces dernières dépendent d'une variable précédemment renseignée. Par exemple si une variable A est « vous souvenez-vous de votre poids à 20 ans » réponse : NON, et que la variable B est « quel était votre poids à 20 ans ? » il est normal de ne pas avoir de valeur à la variable B, c'est donc une valeur absente et non une valeur manquante. Les valeurs manquantes sont matérialisées par des « NA » tandis que les valeurs absentes sont matérialisées par des « - ». Pour se faire, les variables concernées ont été identifiées. Leur petit nombre a permis un travail par filtration dans le logiciel Excel®.

4. Un système d'information requêtable facilement

Finalement, le système d'information doit permettre aux utilisateurs de requêter les bases de données pour extraire des sous-jeux de données à volonté. La mise en place d'un système de requête efficace va permettre de rendre les données réutilisables dans différents buts et par différentes personnes. La construction de la requête doit pouvoir se faire le plus naturellement possible depuis une interface graphique web, et à l'aide d'opérateurs booléens (ET, OU, SAUF…) permettant l'imbrication de plusieurs requêtes entre elles. Plus la construction des requêtes sera intuitive, plus facile sera l'utilisation du système d'information. La F**igure 23** présente un exemple de construction de requête complexe qui traduit l'information suivante : « Je veux sélectionner les individus âgés de 60 à 65 ans et avec une glycémie à jeun strictement supérieure à 1g/L et avec soit un indice de masse corporel supérieur ou égale à 25, soit un poids d'exactement 70 kg ».

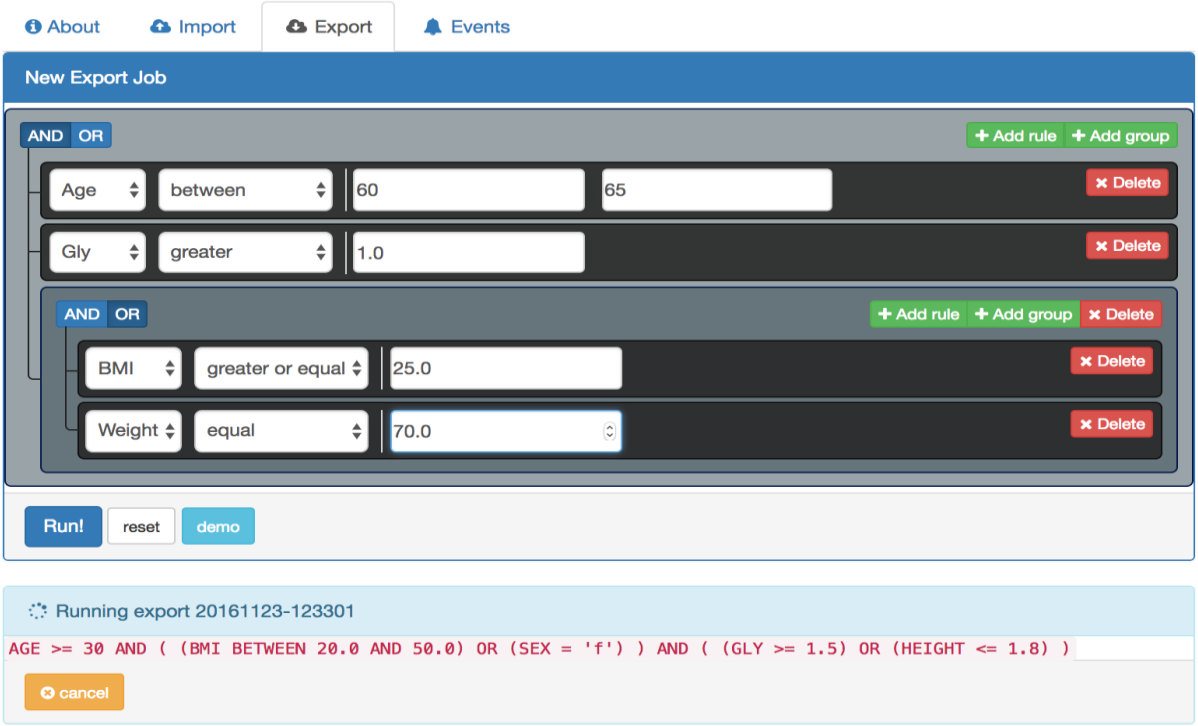


Figure 23 : Exemple d'une requête de sous jeu de données dans l'interface web d'OmicM.

Le système doit également permettre de générer automatiquement des fichiers orientés individus ou échantillons. En effet, le type de fichiers nécessaires sera différent selon l'échelle considéré (**Figure 24**). Pour réaliser un état des lieux des variables cliniques par exemple, l'utilisateur

aura besoin de générer un sous fichier contenant ces variables pour chaque individu. Il contiendra donc par exemple les variables IMC@T1 et IMC@T4 pour chaque individu (c'est-à-dire l'IMC des individus à chaque temps de mesure). En revanche, pour analyser des données relatives aux échantillons comme par exemple les données métabolomiques tout en prenant en compte le facteur IMC du patient, il faudra que le fichier généré possède des informations non plus pour chaque individu, mais pour chaque échantillon (l'un à T1, l'autre à T4). Par conséquent les variables IMC@T1 et IMC@T4 doivent fusionner pour devenir la variable IMC@T représentant le poids de l'individu dépendamment de l'échantillon considéré et non plus du sujet.

ID_pfem	status	IMC@T1	IMC@T4	C1@T1
GP001	témoin	25.9	25.3	Non/No
GP002	témoin	23.4	22.7	Oui/Yes
GP003	cas	33.9	35.3	Non/No
GP004	témoin	23.9	23.4	Non/No
GP005	cas	28.0	28.4	Oui/Yes
GP006	cas	26.4	26.9	Oui/Yes
GP007	témoin	22.1	21.7	Oui/Yes
GP008	témoin	20.8	20.8	Non/No
GP009	témoin	27.8	25.9	Oui/Yes
GP010	témoin	24.8	23.8	Oui/Yes
GP011	témoin	30.6	28.6	Non/No
GP012	cas	31.5	30.0	Oui/Yes
GP013	cas	29.3	30.7	Non/No
GP014	cas	32.9	32.4	Non/No

Fichier orienté sujets

ID_LIMS	ID_pfem	ID_pfem_temps	temps	status	IMC@T	C1@T
Ec52230	GP001	GP001_T1	T1	témoin	25.9	Non/No
Ec52406	GP001	GP001_T4	T4	témoin	25.3	Oui/Yes
Ec52319	GP002	GP002_T1	T1	témoin	23.4	Oui/Yes
Ec52392	GP002	GP002_T4	T4	témoin	22.7	Oui/Yes
Ec52308	GP003	GP003_T1	T1	cas	33.9	Non/No
Ec52238	GP003	GP003_T4	T4	cas	35.3	Non/No
Ec52262	GP004	GP004_T1	T1	témoin	23.9	Oui/Yes
Ec52367	GP004	GP004_T4	T4	témoin	23.4	Non/No
Ec52439	GP005	GP005_T1	T1	cas	28.0	Non/No
Ec52300	GP005	GP005_T4	T4	cas	28.4	Oui/Yes
Ec52221	GP006	GP006_T1	T1	cas	26.4	Non/No
Ec52358	GP006	GP006_T4	T4	cas	26.9	Oui/Yes
Ec52445	GP007	GP007_T1	T1	témoin	22.1	Non/No
Ec52343	GP007	GP007_T4	T4	témoin	21.7	Non/No

Fichier orienté échantillons

Figure 24 : Translation entre un fichier orienté sujet et un fichier orienté échantillon.
ID_pfem : identifiant du sujet ; ID_LIMS : identifiant de l'échantillon

Par ailleurs, et pour faciliter la sélection des variables notamment lorsqu'elles sont nombreuses, il doit pouvoir permettre de sélectionner des groupes de variables prédéfinis (par exemple les sous-bases correspondantes aux différents types de données de la cohorte NuAge ou toutes les données métabolomiques issues de l'analyse en C18-Pos…).

Finalement, l'interface de requête doit permettre d'accéder au dictionnaire sans perdre la requête en cours de construction afin de vérifier la nature des variables ou encore leur nom tout au long de la construction. De plus, le téléchargement des sous-jeux de données doit s'accompagner du téléchargement d'un dictionnaire personnalisé contenant uniquement les variables extraites de la base pour assurer la compréhension des variables quel que soit l'utilisateur qui extrait les données du système.

5. Etat de développement d'OmicM

La mise en place et la construction d'un tel système d'information nécessite du temps et la mobilisation de plusieurs personnes. Les enjeux informatiques sont divers que ce soit le lieu de stockage des données, leur sécurisation, le choix des langages de programmation utilisés pour construire les différentes bases de données, etc… Pour ces raisons, OmicM n'a pas pu être totalement finalisé au court de ma thèse et reste actuellement en développement. Il n'a donc pas directement permis de manager toutes les phases du projet. Les données issues de la cohorte NuAge ont pu être intégrées au système et les données brutes issues des analyses de la plateforme d'exploration du métabolisme de Clermont-Ferrand/Theix ont également été stockées à l'état de fichiers bruts, les données prétraitées ne pouvant pas encore être prise en charge. La gestion à « l'échelle échantillon » des données n'ayant pas été mise en place, les résultats des analyses relatives à la comparaison de ces derniers ainsi que les résultats des différentes analyses statistiques sur les données métabolomiques n'ont pas pu être stockés sur le système.

Les fonctionnalités intégrant les informations issues de bases de données extérieures (HMDB, ChEBI…) ou interne (base de données des biomarqueurs de la littérature concernant le SMet, PeakForest…) n'ont à ce jour pas encore été mise en place.

Pour assurer au mieux le management des données, en attendant la mise en fonction de OmicM, j'ai dû développer de petits scripts Perl. Le premier, présenté précédemment, permet de formater au mieux les données, le second lui permet de générer des sous-jeux de données à partir des fichiers reformatés. Cette génération de sous-jeux de données fut nécessaire pour permettre la randomisation des échantillons, les différentes analyses statistiques ou encore l'interprétation des données. L'objectif du script est de simplifier au maximum cette tâche et de l'automatiser pour limiter le risque d'erreurs lié à une intervention humaine. Le script prend en entrée un jeu de données à filtrer, ainsi qu'une liste de variables ou d'individus d'intérêt relative au futur sous-jeu de données. L'utilisateur paramètre la filtration souhaitée : il indique si sa liste d'entrée contient des variables ou des sujets, puis s'il souhaite conserver ou à l'inverse supprimer les variables/sujets présents dans la liste. Le script renvoie alors le nouveau sous-jeu de données correspondant.

Tous ces outils mis en place pour faciliter la gestion des données permettent de passer à l'étape suivante qui consiste à en extraire des connaissances pour répondre à la ou les problématique(s) de recherche qui peuvent être posées.

REFERENCES DU CHAPITRE 2

1. Gaudreau P, Morais JA, Shatenstein B, Gray-Donald K, Khalil A, Dionne I, et al. Nutrition as a determinant of successful aging: description of the Quebec longitudinal study Nuage and results from cross-sectional pilot studies. Rejuvenation Res. 2007;10(3):377-86.
2. Blaak E. Gender differences in fat metabolism. Curr Opin Clin Nutr Metab Care. 2001;4(6):499-502.
3. Statistique Canada [a]. Proportion des personnes de 12 ans et plus ayant reçu un diagnostic d'hypertension, selon le groupe d'âge et selon le sexe, Québec, 2013-2014.
4. Statistique Canada [b]. Tableau 13-10-0096-20 Indice de masse corporelle, embonpoint ou obèse, autodéclaré corrigé, adulte, selon le groupe d'âge (18 ans et plus).
5. Statistique Canada [c]. Proportion des personnes de 12 ans et plus ayant reçu un diagnostic d'hypertension, selon le groupe d'âge et selon le sexe, Québec, 2013-2014.
6. Alberti KG, Eckel RH, Grundy SM, Zimmet PZ, Cleeman JI, Donato KA, et al. Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. Circulation. 2009;120(16):1640-5.
7. Tan S, Avalos G, Dineen B, Burke A, Gavin J, Brennan M, et al. Traveller health: prevalence of diabetes, pre diabetes and the metabolic syndrome. Ir Med J. 2009;102(6):176-8.
8. Cajka T, Fiehn O. Comprehensive analysis of lipids in biological systems by liquid chromatography-mass spectrometry. Trends Analyt Chem. 2014;61:192-206.
9. Duarte IF, Diaz SO, Gil AM. NMR metabolomics of human blood and urine in disease research. J Pharm Biomed Anal. 2014;93:17-26.
10. Forcisi S, Moritz F, Kanawati B, Tziotis D, Lehmann R, Schmitt-Kopplin P. Liquid chromatography-mass spectrometry in metabolomics research: mass analyzers in ultra high pressure liquid chromatography coupling. J Chromatogr A. 2013;1292:51-65.

CHAPITRE 3 : TRAITEMENT DES DONNEES

Notre objectif premier est d'établir une signature métabolique du SMet stable dans le temps, à partir de l'analyse des 6 jeux de données métabolomiques. Concrètement, il s'agit de déterminer un ensemble de métabolites/biomarqueurs qui permettront la construction d'un modèle statistique capable de différencier les individus SMet des témoins et dont la stabilité dans le temps est vérifiée. Différentes recommandations ont été émises dans la littérature quant à la définition et à la qualification d'un biomarqueur [87]. En particulier, pour qu'une signature (ou ensemble de biomarqueurs) soit utilisable d'un point de vue clinique, elle doit (i) contenir des métabolites facilement mesurables pour les cliniciens (méthodes d'analyse fiables et répétables) ; (ii) apporter de nouvelles informations concernant la pathophysiologie; (iii) permettre une bonne prise en charge des patients [87]. Dans le cas présent, étant dans une phase de découverte de cette signature, l'objectif est d'identifier les métabolites modulés par le SMet stable dans le temps, puis de constituer une signature possédant un nombre restreint de métabolites pour qu'elle puisse être, après validation, utilisée en clinique. Le travail bio-informatique consiste par conséquent à construire un workflow répétable et reproductible pour arriver à cet objectif.

Dans le cadre d'un projet utilisant plusieurs techniques d'analyse dont les résultats devront par la suite être intégrés, il est important d'adopter une méthodologie de traitement des résultats uniforme et commune. Au même titre que l'ordre de passage des échantillons a été déterminé de façon à être le même sur tous les équipements et que les étapes de prétraitement ont été uniformisées, le traitement des données jusqu'à l'obtention d'une signature du SMet l'a également été.

La construction de workflow peut passer par l'utilisation de différents outils. Dans le domaine des sciences de la vie, leur mise à disposition au sein d'instances informatiques accessibles à tout type d'utilisateur a notamment été possible grâce à la création de Galaxy [224]. Ce gestionnaire de workflow permet le regroupement d'outils dans le but de les chaîner, permettant ensuite d'appliquer cet enchaînement à différents jeux de données en offrant la possibilité de le partager avec la communauté. Différentes instances Galaxy ont été créées pour répondre aux besoins parfois très spécifiques de certains domaines (génomique, transcriptomique, protéomique). L'instance W4M [82] a été créée pour répondre aux besoins analytiques spécifiques à la métabolomique. Elle regroupe des outils communément utilisés par la communauté comme XCMS ou CAMERA pour l'extraction des données, ainsi que des outils développés spécifiquement pour les analyses statistiques ou encore l'annotation des métabolites. Nous avons donc choisi de construire au maximum notre workflow d'analyse au sein

de cette instance afin de pouvoir facilement reproduire les différentes étapes du traitement à tous les jeux de données générés et créer une méthodologie applicable à de futurs projets similaires.

I. Métabolites et lipides modulés par le SMet

1. Données d'entrée

Comme présenté dans le chapitre précédent, les paramètres de prétraitement ont été choisis de manière à être communs aux 6 méthodes d'analyse. Cette étape a donné lieu à la création de 6 jeux de données sous la forme de 3 fichiers pour chacun : (i) une dataMatrix qui correspond à la table des intensités pour chaque ion ou bucket mesuré, et ce pour chaque échantillon ; (ii) une variableMetadata qui contient , pour chaque ion/bucket, les variables le caractérisant (m/z, temps de rétention, variables issues du prétraitement…) ; et (iii) une sampleMetadata qui contient les informations relatives à chaque échantillon (statut du sujet, temps de mesure…). L'ensemble de ces fichiers contient un nombre important de variables : 13 902 tous jeux confondus (1 656 en C18-Pos, 606 en C18-Neg, 1 124 en HILIC, 6 697 en Lipidomique, 3 745 en GC et 74 en RMN).

2. Filtration des redondances analytiques

Comme évoqué dans l'introduction, lors des analyses d'échantillons par spectrométrie de masse, les données générées présentent un fort taux de redondance d'information. Au cours de la phase d'ionisation, un métabolite va donner plusieurs ions dupliquant ainsi l'information initiale. Selon les métabolites le nombre de fragments varie ne donnant pas le même taux de redondance. Cette dernière, que l'on pourrait qualifier d'analytique, peut influencer les résultats des analyses statistiques, en accordant plus de poids à certains métabolites hautement fragmentés en comparaison d'autres peu ou pas fragmentés. Elle peut notamment avoir un impact fort sur les corrections de test multiple (BH par exemple) ou certains indicateurs générés par des méthodes de fouille de données comme SVM et Random Forest.

Pour gérer cette redondance, il est nécessaire **d'identifier de la manière la plus fiable possible les ions issus d'un même métabolite**. Cette étape est possible par le biais des annotations des

métabolites, quel que soit le niveau d'annotation. Toutefois, compte tenu du nombre très important de variables générées dans un projet comme le nôtre, faisant appel à plusieurs méthodes d'analyse non ciblées, cette approche n'est pas envisageable.

Il existe des outils qui peuvent être utilisés pour détecter les ions provenant d'un même métabolite. Toutefois, ces derniers prennent souvent en entrée des données brutes et non des données tabulées comme celles que nous possédons après le prétraitement. Ils prennent en compte des paramètres telles que les corrélations spectrales ou le temps de rétention, qu'il n'est pas toujours possible de paramétrer, et laissent souvent de côté une information primordiale : la différence de masse entre les ions, pouvant entrainer des erreurs de regroupement. En effet, les ions issus d'un même métabolite apparaissent au même temps de rétention, possèdent des masses différentes et sont hautement corrélées (corrélation > 0,9). La différence de masse entre les ions correspond à des fragments, des adduits ou des isotopes du métabolite d'origine. De plus, il est fréquent que plusieurs métabolites soient ionisés à un même temps de rétention (co-élution). Il est alors facile de confondre 2 ions issus d'un même métabolite et 2 ions issus de 2 molécules différentes. Seule l'information de la différence de masse et de la corrélation dans le jeu de données permet cette distinction. Pour toutes ces raisons, nous avons décidé de développer un nouvel outil, prenant en entrée des données tabulées, s'insérant dans un workflow et permettant d'avoir accès aux différents paramètres de regroupement. Cet outil, nommé Analytic Correlation Filtration (ACorF), a été codé en langage Perl. Il permet de grouper les ions analytiquement corrélés en se basant sur leur taux de corrélation (calculé selon la méthode choisie par l'utilisateur en amont de l'outil), le temps de rétention et une liste de différences de masses exactes préétablie, identifiant les fragments, adduits et isotopes connus. Cette liste a été construite au sein de MetaboHUB et reflète l'expertise acquise par les analystes au fil des années. Une fois les ions regroupés, l'outil propose un représentant pour chaque groupe afin de ne conserver que ce dernier, supprimant ainsi ces redondances. Le choix de ce représentant peut se faire de différentes façons selon l'option choisie par l'utilisateur.

Pour valider l'efficacité de l'outil, nous avons comparé ses fonctionnalités ainsi que les résultats obtenus sur un jeu de données publié, à ceux obtenus avec un outil effectuant également du regroupement d'ion : CAMERA (dans sa version disponible sous W4M).

Dans le but de de construire des workflows d'analyse sous Galaxy en incluant cette étape de filtration de corrélations analytiques, ACorF a été « Galaxifié » (construction de l'interface graphique adaptée à Galaxy) pour être rendu accessible dans l'instance dédiée à la métabolomique W4M. Cette étape a notamment impliqué la création d'un fichier « xml » définissant les caractéristiques de l'interface graphique et imbriquant les fonctionnalités de lancement du script.

Publication n°2

Monnerie S, Pétéra M, Gaudreau P, Comte B, Pujos-Guillot E.

Analytic correlation filtration: A new tool to reduce analytical complexity of metabolomic datasets

Article soumis au journal « Metabolites »

metabolites

MDPI

Article

Analytic Correlation Filtration: A New Tool to Reduce Analytical Complexity of Metabolomic Datasets

Stephanie Monnerie [1,*], Melanie Petera [2], Bernard Lyan [2], Pierrette Gaudreau [3,4], Blandine Comte [1], Estelle Pujos-Guillot [1,2,*]

[1] Université Clermont Auvergne, INRA, UNH, Mapping, F-63000 Clermont Ferrand, France; blandine.comte@inra.fr

[2] Université Clermont Auvergne, INRA, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, F-63000 Clermont-Ferrand, France; melanie.petera@inra.fr (M.P.); bernard.lyan@inra.fr (B.L.)

[3] Centre de Recherche du Centre Hospitalier de l'Université de Montréal, Montréal, QC H2X 3E4, Canada; pierrette.gaudreau@umontreal.ca

[4] Département de médecine, Université de Montréal, Montréal, QC H3T 1J4, Canada

* Correspondence: stephanie.monnerie@inra.fr (S.M.); estelle.pujos-guillot@inra.fr. (E.P.-G.)

Received: 3 October 2019; Accepted: 22 October 2019; Published: date

Abstract: Metabolomics generates massive and complex data. Redundant different analytical species and the high degree of correlation in datasets is a constraint for the use of data mining/statistical methods and interpretation. In this context, we developed a new tool to detect analytical correlation into datasets without confounding them with biological correlations. Based on several parameters, such as a similarity measure, retention time, and mass information from known isotopes, adducts, or fragments, the algorithm principle is used to group features coming from the same analyte, and to propose one single representative per group. To illustrate the functionalities and added-value of this tool, it was applied to published datasets and compared to one of the most commonly used free packages proposing a grouping method for metabolomics data: 'CAMERA'. This tool was developed to be included in Galaxy and will be available in Workflow4Metabolomics (http://workflow4metabolomics.org). Source code is freely available for download under CeCILL 2.1 license at https://services.pfem.clermont.inra.fr/gitlab/grandpa /tool-acf and implement in Perl.

Keywords: metabolomics; data filtration; high-resolution mass spectrometry

1. Introduction

Metabolomics is described as an approach allowing the description of small molecules/metabolites present in a biological system by identifying and possibly quantifying them [43]. By the global study of low molecular weight compounds, it allows determining metabolic phenotypes that may vary between individuals depending on several factors: genetics, environment exposure, and life habits. As it gives an integrated vision of the health status, metabolomics has been shown in recent years as a powerful tool to better understand the biological mechanisms involved in the pathophysiological processes and to identify disease biomarkers [225]. However, this approach generates massive and complex data that need adequate analyses to extract the biologically meaningful information.

In metabolomics, mass spectrometry (MS) is one of the most common analytical platforms. Upstream to the ionization and detection steps, metabolites from complex biological samples are separated most frequently using liquid chromatography (LC) or gas chromatography (GC) [164, 226]. During the analysis, metabolites are ionised in the mass spectrometer source to produce several ions/analytical features (parent ion, isotopes, adducts, and fragments) that are part of the original molecule. Raw data obtained from metabolic profiles are processed to yield a data matrix containing retention times (time between sample injection and appearance of the maximum ion signal), masses,

and peak intensities. This step results in thousands of features present in the final dataset with a high degree of correlation [[227]]. In addition to this analytical redundancy, biological correlation issued from modulated metabolites coming from the same pathways does also exist. This high degree of correlation in datasets is a constraint for the use of various data mining and statistical methods. For example, analytical redundancy highly affects multiple testing correction. Indeed, having non-independent variables (coming from the same metabolite) leads to an over-correction of data that can hide potentially relevant information. Moreover, for biological interpretation, experts are mainly focusing on metabolites rather than on the different analytical features. For all these reasons, considering metabolite as a unique entity instead of the individual ions, as a variable, is more relevant.

In order to handle this data complexity and, in particular, identify subgroups of related features, tools have been developed, mostly for annotation purposes [168]. In fact, the use of correlations between ions contributes to determining chemical identities and isolating species part of the same metabolite from those of other coeluting compounds. Generally, the available packages consist of a two-step process: first, a grouping of all features derived from the same analyte, and then, an annotation of the ion species. The first category of tools is based on a grouping approach consisting of using chromatographic peak-shape similarity of coeluting features in raw data. Among these tools, CAMERA is a Bioconductor R package that is widely used in the field of mass spectrometry metabolomics [228], designed to post-process XCMS [59] feature lists and to collect all features related to a compound. In an iterative process, it first selects the most intense feature not yet assigned to a compound spectrum and determines an associated retention time (RT) window. All features within this range are then included in a new compound spectrum. Next, the algorithm excludes unfitting features using the chromatographic peak shape similarity of the extracted ion chromatogram (EIC) of each feature, as well as a Pearson correlation between intensities inside the chromatographic peak boundaries for all pairs of features within the compound spectrum. Finally, in the last step, it allows annotation of adducts, common neutral losses using a combination of lists of observable ions. Alternatively, some computational tools involve intensity correlation analysis across multiple samples as the basis of feature grouping. AStream [229]], a tool designed for LC/MS annotation, is based on a grouping method using intensity correlations, retention time, and adduct, isotope, and fragment identification. This tool, available as an R package, uses the intensity correlations across samples for all the features present in the data instead of analysing the individual chromatogram correlations. Therefore, it does not require the raw LC/MS data but only features intensities. More recently, cliqueMS [230] proposed an integrated approach building a feature similarity network from coeluting profiles. xMSannotator [231] also incorporates multi-criterion scoring (both analytical and biological correlations, matches in databases, etc.) for improving the annotation of high-resolution metabolomics data.

Even though a grouping step does exist in these packages, most of the time all the generated information (correlation coefficient, retention time, mass difference between features) are not used together to form the groups. Moreover, users do not have the possibility of accessing the grouping information nor obtaining representative features among groups, which is essential for data reduction. Finally, most of those tools are not giving the possibility to build a workflow with other processing tools, as they require either a specific input format or specific approach for measuring similarities.

In the context of contemporary e-Science, it has been recognised that data have to be Findable, Accessible, Interoperable, and Reusable in the long-term [71]. Workflow4Metabolomics (W4M, http://workflow4metabolomics.org; [82, 83] has been developed within this objective, for comprehensive metabolomics data pre-processing, statistical analysis, and interpretation. It is a fully open-source virtual research environment (VRE; [232]) built upon the Galaxy environment [224] for bioinformatics developers and metabolomics users and allows user-friendly functionalities for workflow management.

In this context, we proposed a new stand-alone tool dedicated to data reduction based on the removal of analytical redundancies of MS-based metabolomics datasets. As a key element within the metabolomics data analysis workflow, as well as to ensure reproducible computational analyses, we

made it available via Galaxy and provided it for W4M, with generic input files and different output files for visualisation and further data analysis steps within workflows.

2. Materials and Methods

Our aim was to detect analytical correlations into MS-based metabolomics datasets (tabular files) without confounding them with biological ones that may exist within samples. To achieve our goal, we developed a Perl tool supported by metabolomics experts to translate and understand the chemical complexity of datasets as well as possible.

The algorithm principle is to group features coming from the same metabolite and to suggest one single representative per group. In optimal settings, the grouping criteria include a similarity measure, retention time, and mass information from a reference list containing isotopes, adducts, and fragments. Thresholds for all these criteria can be fixed, and the representative feature can be determined following four methods according to the user's needs and the analytical technology used, either LC- or GC-MS. As the output, the module returns the input file with new columns in relation to resulting groups (representative feature choice, grouping information, and annotation of features), as well as a .sif file allowing correlation network visualisation of the dataset of interest. The present tool "Analytic correlation filtration" (ACorF) is available via the web interface Galaxy as a single module and can be chained with other W4M modules.

As CAMERA is also available in W4M; the present tool was compared to this package by using a published dataset, demonstrating its utility and various possibilities of use.

2.1. Algorithm Description

Major steps of the algorithm are presented in Figure 1.

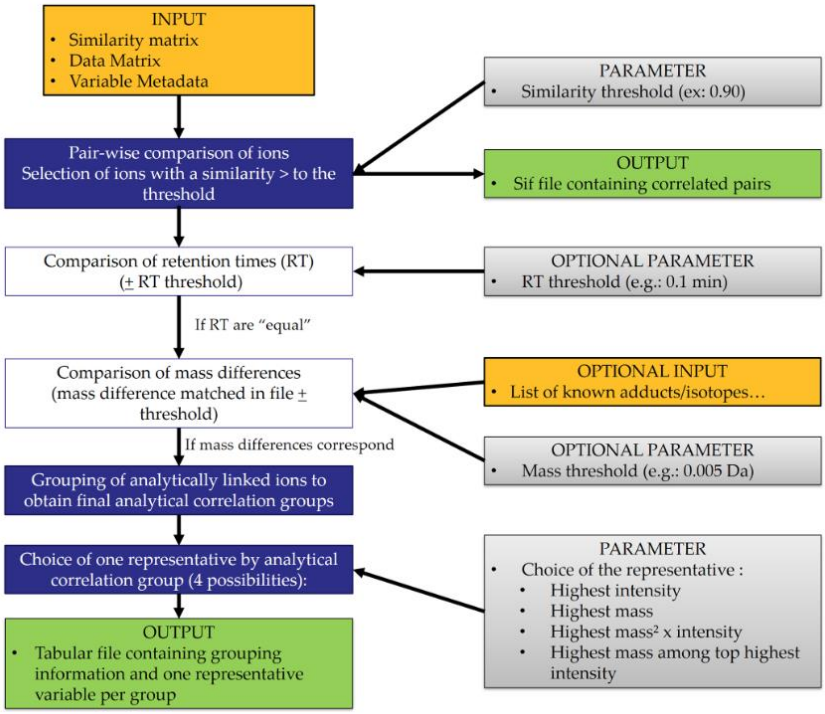


Figure 1. Flowchart representing the major steps of the algorithm (input and output files, parameters, and calculation steps). The key mandatory steps are filled in blue, while optional ones are in white.

2.1.1. Input Files

The ACorF tool takes 3 files related to collected data as input, in tabular format (see Supplemental Figure 1). The first file, referred to as data matrix, consists in a table containing intensities of each variable (each ion detected on the mass spectrum) per sample; the second file, referred to as variable

metadata, consists in descriptive additional metadata of variables (e.g., m/z, retention time). The tool also takes, as input, a third file, the similarity matrix: a table representing pair-wise similarity within the dataset, in CSV or tabular format. This table generation is not included in the tool to allow more flexibility: there is a large variety of similarity measures (Pearson/Spearman correlation, Clustering, Partial correlation, et al.), whose relevance can vary depending on the filtering goal. The similarity matrix can be obtained either using W4M (*e.g.*, Metabolites Correlation Analysis, Between Table Correlation, et al.) or any external tool.

The last file, containing a list of known adducts, fragments, and isotopes, and their associated masses, is needed when choosing the mass comparison option.

2.1.2. Processing

The first step of the algorithm is performing a pair-wise comparison of the different variables. The similarity matrix is read, and only pairs having a similarity coefficient higher than the chosen threshold are selected.

The next two steps are optional but highly recommended to increase analytical relevance. In a pair-wise process once again, the retention times of variables within the selected pairs are compared. If the ions have an identical RT (more or less a delta fixed by the user), their mass difference can be taken into account. Indeed, the user can specify the use of a list of known isotope, adduct, and fragment mass differences. In case the user does not provide a personal uploaded list, a default one is available within ACorF. The mass difference between two variables is compared to this list with a tolerance defined by the user, to confirm the chemical link between them. If a match is found, the two ions are considered as coming from the same metabolite and will be put in the same group. Those steps are repeated for each selected pair to obtain analytical correlation groups.

The last step consists of choosing a representative variable for each group. The user can choose among four options to allow the best choice of the quantifier depending on its technology and method (ensuring good signal to noise ratio and specificity).

1) Retaining the ion with the highest intensity
2) Retaining the ion with the highest mass
3) Retaining the ion with the highest 'mass^2 × average intensity'
4) Retaining the highest mass among the top highest average intensities of the group. For this last option, the user determines the number of ions considered in the top list (top 5, top 3, top 10, etc.).

2.1.3. Output Files

The correlated pairs are used to create the first output, a *.sif file containing pair-wise correlation rate. This file allows correlation network visualisation using tools such as Cytoscape [233].

Then, ACorF returns a second output file consisting in the variable metadata file (in tabular format) with additional result columns. This new file includes (i) an 'ACorF_groups' column that contains the group name; (ii) a 'isotopes_adducts_fragments' column that proposes annotations based on the list of known isotope, adducts, and fragments (relatively for each ions contained in the same group); (iii) a column entitled 'ACorF_filter' that indicates if the variables have to be conserved or deleted for a filtration step; (iv) a 'representative' column that contains the name of the variables selected as representatives of their correlation groups (the name of this column will indicate the chosen representative option); and finally (v) an 'annotation_relative_to_representative' column that proposes annotation of the ion comparatively to the representative ion selected for the concerned group. If no analytical correlation is found for a given ion, it is assigned to an individual group, and remaining cells are filled with '-'.

2.2. Examples of Use

To illustrate the ACorF functionalities and the results obtainable on typical experimental data, datasets publicly available on W4M were used as examples.

The first dataset, named 'Sacurine' (W4M00002_Sacurine-comprehensive), was obtained from LC-HRMS analyses (negative ionisation mode) of human urine samples in Guitton et al. [83] (DOI:10.15454/1.481114233733302E12). The present test was performed on a subset of the initial 7456 ion dataset after noise elimination. This subset contains a total of 184 samples and 3120 ions after various steps of pre-processing using XCMS and noise filtration. The ACorF tool was applied using parameters as close as possible to the ones of CAMERA to allow a better comparison: (i) a Pearson correlation as similarity measurement with a threshold of 0.75 (as it is used as default setting in CAMERA); (ii) a RT threshold of 0.1 min; (iii) the default list of known adducts and isotopes, with (iv) a mass threshold of 0.002 Da, and (v) the representative ion selected as the one with the highest intensity.

The second dataset, named 'Algae' (W4M00004_GCMS-Algae), was obtained from GC-MS analyses of algae samples (DOI:10.15454/1.4811272313071519E12) in Guitton et al. [83]. This dataset contains a total of 12 samples and 2908 ions after various steps of pre-processing using XCMS. The ACorF tool was applied using GC-MS recommendations: (i) a Pearson correlation as a similarity measurement with a threshold of 0.90; (ii) a RT threshold of 0.1 min; (iii) a list of known adducts and isotopes, with (iv) a mass threshold of 0.2 Da due to low resolution of the quadrupole mass spectrometer, and (v) the representative ion selected as the one with the mass among those with highest intensity.

3. Results and Discussion

3.1. Functionalities

We chose to compare the present tool to one of the most commonly used free packages proposing a grouping method and available as a W4M module: 'CAMERA.annotate'.

Before comparing the results, the functionalities of both tools were listed in Table 1. The first important difference concerns the input format. The CAMERA R package requires the use of XCMS to pre-process raw data. Therefore, the CAMERA.annotate module takes an .RData as the input file resulting from the xcmsfill.peaks function. The advantage of the ACorF tool is to propose more universal inputs with three files that can easily be generated from any type of metabolomics data, whatever the software used to obtain the data. Another difference concerns the algorithm itself and in particular, the grouping step. To group ions, CAMERA defines an RT window for each peak, based on the highest intensity ion chromatogram, using 2 different parameters ([sigma] and [perfwhm]) related to the chromatographic peak characteristics. Even if the user can optimise those parameters, CAMERA does not give the possibility to have an overview of their impact on each defined peak. The present tool is based on a different algorithm that proposes to the user to determine an RT window. In addition, in CAMERA, a list of mass differences for known adduct/isotopes is not used for the pc-group formation but only for annotation purposes. In ACorF, the user can choose a default list or his own list to group ions together with similarity and RT criteria to validate the fact that the redundancy observed most likely has an analytical origin.

Finally, the major attractive feature of the present tool compared to CAMERA is its selection of a representative ion for each formed group, based on 4 alternative methods. This representative ion proposition allows dataset filtration by removing analytic correlations.

Table 1. Comparison of the CAMERA.annotate and ACorF tool functionalities.

-	**CAMERA.annotate (W4M version)**	**"Analytic Correlation Filtration" (ACorF) Tool**
Interface	Galaxy (W4M)	Galaxy (W4M)
Language	R	Perl
Version	Galaxy version 2.1.3	-
Input files	.Rdata output from XCMS Galaxy pre-processing	DataMatrix, variableMetadata and similarity matrix
Parameters	-	-
Mandatory	-	Correlation rate Representative selection method
Optional	Correlation rate RT window determination variables	Mass difference list Retention time tolerance delta
Correlation information	Calculation of correlation is included in the tool	A correlation table has to be obtained before using the tool
Correlation type	Pearson correlation	Any type of correlation are possible
Possibility to set a Correlation threshold	Yes—only for the second step of grouping	Yes
Retention time (RT) window	Calculated for each peak	Defined by a threshold
Parameter settings	Two different parameters ([sigma] and [perfwhm]) are available	The user can set the RT tolerance delta value
Comparison to a mass defect list	Conditioned by obtained group but not used for grouping	When used, directly impact the group determination
Isotope identification	Yes—performed in a previous step	Yes—if the isotope mass difference is included in the list
Existing default list	Yes	Yes
Possibility to upload a personal list	Yes	Yes
Possibility of setting a mass difference tolerance value	No	Yes
Possibility of selecting a representative ion for each group	No	Yes
Output files	variableMetadata with additional columns	variableMetadata with additional columns and a .sif file for network visualisation
Optional output files	EIC for main pc-group visualisation pdf file	-

3.2. Example of Use: The Sacurine Dataset

To illustrate the ACorF functionalities and the results obtainable on typical experimental data, the ACorF tool was applied to the Sacurine dataset, using parameters as close as possible to the ones of CAMERA to allow a better comparison.

Within the 3120 ions, ACorF allowed the creation of 2697 groups, meaning that 14% of ions are proposed to be filtered from the dataset because of analytical redundancies. Using the generated *.sif file, a quick network visualisation was performed with Cytoscape (Figure 2). It represents the existing correlations >0.75 between features for groups containing more than 10 ions. An overview of all the correlations >0.75 existing in the analysed subset is available in Supplemental Figure 2.

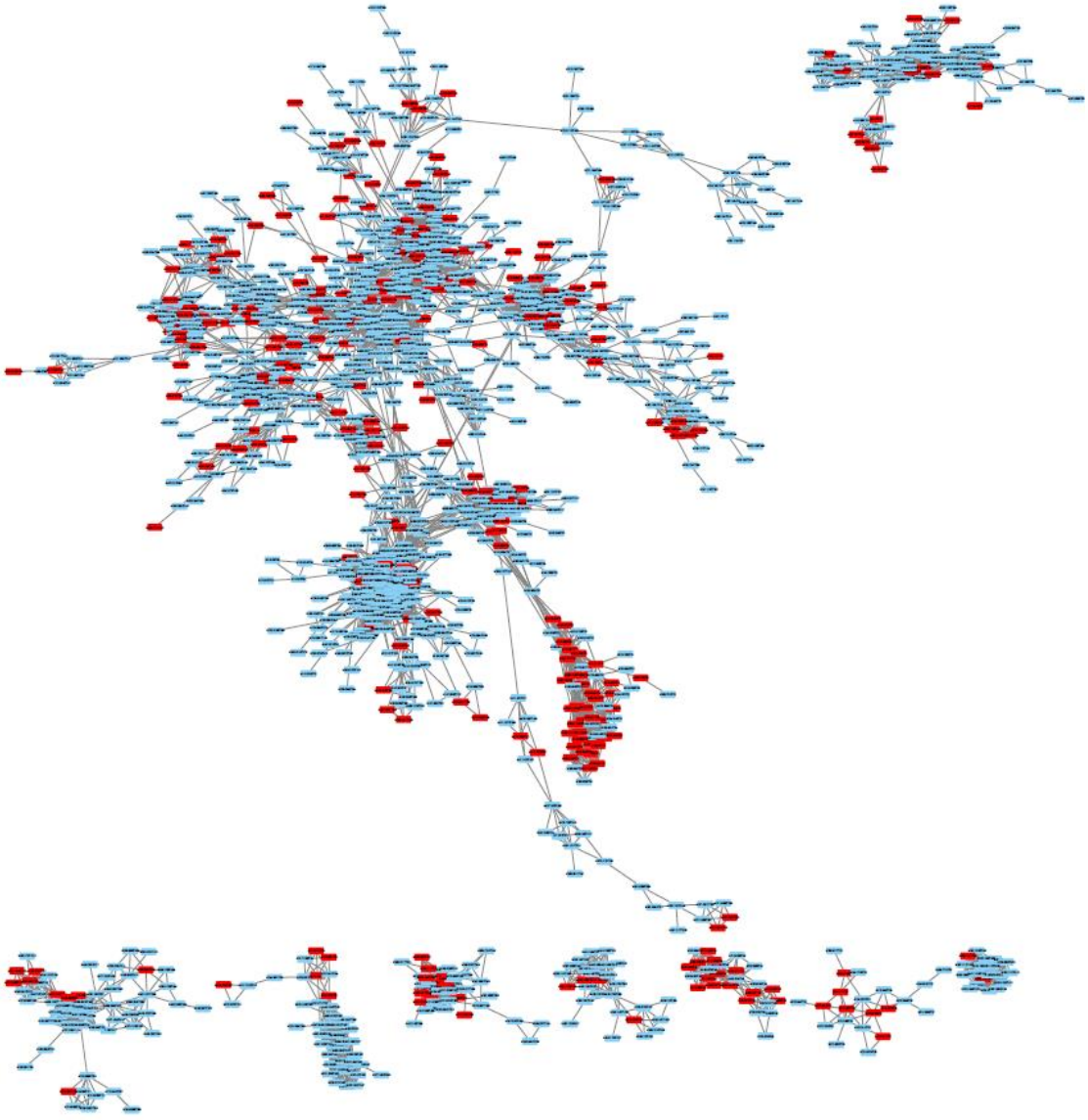


Figure 2. Correlation network of ions that have correlation coefficients above 0.75 in the Sacurine dataset. This network was obtained with Cytoscape using the *.sif output file of the ACorF tool. It represents groups containing more than 10 ions. Red features are identified as being redundant and tagged as deleted by the ACorF tool.

3.3. Result Comparisons

To illustrate the performance of the developed tool, we compared its results obtained using the Sacurine dataset to those obtained using CAMERA.annotate in the published workflow.

While CAMERA proposed 2238 pc-groups, ACorF proposed 2697 groups. With CAMERA, 89% of the pc-groups contained only one ion (1995) versus 77% for ACorF (2397). The numbers of ions in the other groups for both tools are presented in Figure 3. These histograms showed that the groups formed using the ACorF tool are smaller than the ones obtained from CAMERA. While CAMERA generated more than 20 groups of more than 10 ions, the proposed tool subdivided them into smaller ones corresponding to individual annotated metabolites.

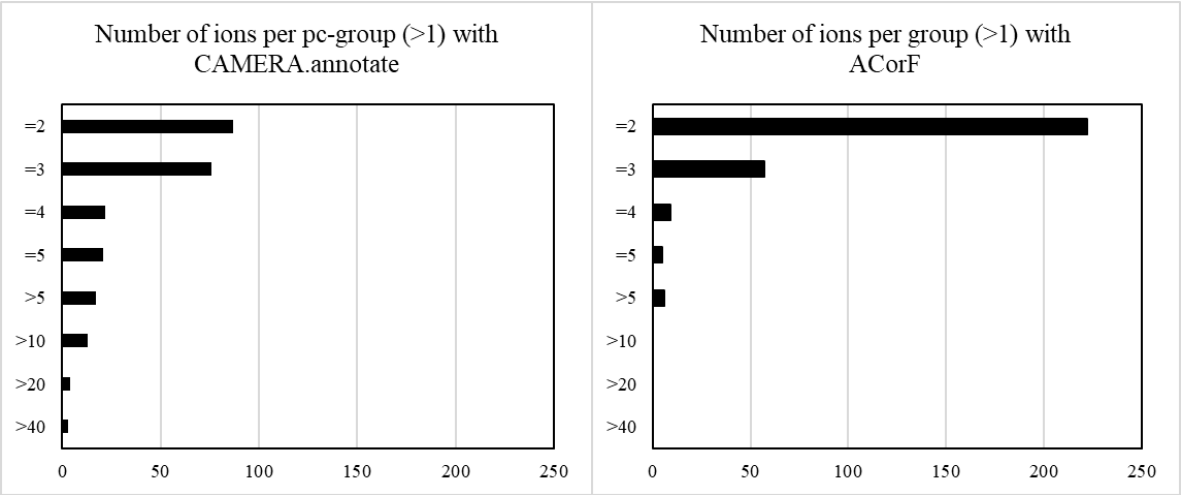


Figure 3. Bar diagram presenting the number of groups (x-axis) by the group size (y-axis: number of ions per group).

To explore these results, we focused on the largest group formed by CAMERA, the pc-group#70, containing 50 ions, all eluting at the beginning of the chromatogram (small values of RT). With the ACorF tool, this pc-group is divided into 44 groups using all the parameters. Among those groups, 32 are single-ion ones that could not be linked to other ions as being part of the same metabolite. The other 12 resulting groups of 2 or more ions are presented in Figure 4: six of those groups are fully included into the initial pc-group#70, and the 6 others include additional ions from other pc-groups, raising the interest of the present approach.

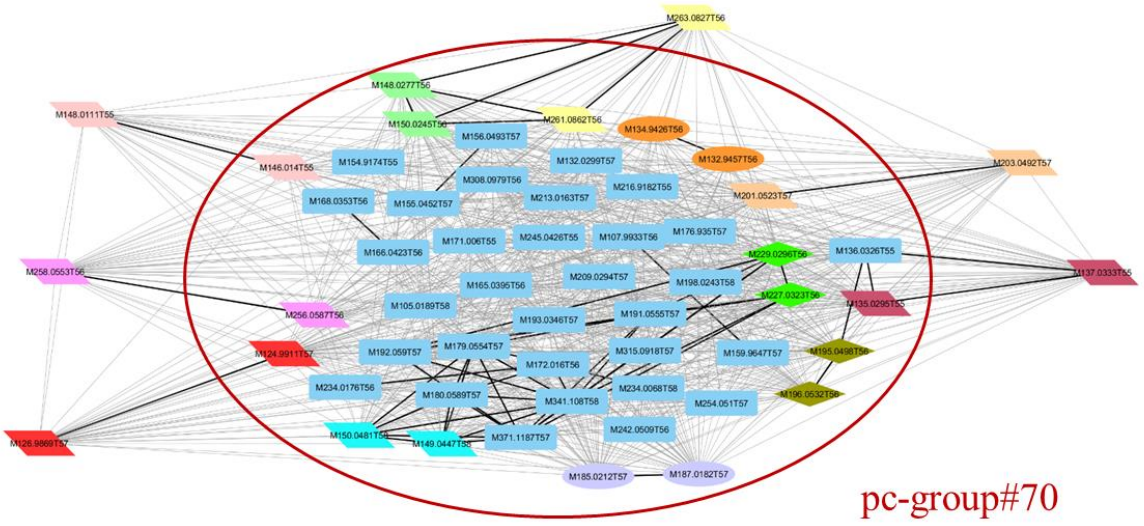


Figure 4. Correlation network of the largest pc-group formed by CAMERA (pc-group#70) shown in the red circle. Ions in blue are those put in individual groups by the ACorF tool. Other coloured ions are those put into groups of ≥2 ions by ACorF. Squared nodes are those that are not grouped by ACorF ; ellipse nodes are grouped by considering correlation coefficient only, diamond nodes are grouped considering correlation coefficient + retention time; and parallelogram nodes are grouped considering correlation coefficient + retention time + mass differences.

To go deeper into interpretation, the impact of the 3 major steps of the ACorF algorithm was evaluated regarding this group subdivision. When only using the correlation as grouping criteria, 24 groups are formed, whereas 39 are created when adding the RT criteria and 5 more are obtained using the mass difference parameter. In this early-eluted compound area, many metabolites are detected within a very narrow RT, increasing the interest of taking this RT factor into account in the grouping step. This point was also highlighted on identified compounds from plasma samples [186]] (see Supplemental Figure S3).

To illustrate the added value of the mass difference parameter, we focused on the pc-group #93 obtained from CAMERA that contains 15 of more retained ions, eluting at around 5 min.

ACorF divided the pc-group#93 into 8 groups using all parameters. In particular, the sub-group#41 contains 10 annotated ions: 8 within the pc-group#93 and 2 more that are included in the pc-group#2729 (M292.0134T309 and M292.0214T309). An example of the annotation of mass defect between features using ACorF (mass threshold 0.002 Da) is presented for sub-group#41 in Figure 5 and compared to an expert annotated raw spectrum, illustrating the validity of the mass defect grouping. Moreover, on this particular example, ACorF allowed for the annotation of sulfate and phosphate moieties that are not identified by CAMERA.

variableMetadata	ACorF_groups	isotopes_adducts_fragments_[#id|annotation(delta_annotation)]	ACorF_filter	intensity_repres	annotation_relative_to_representative
M150.0555T309	group41	#M194.0449T309|-(CO2)(0.000410)|	0	M194.0449T309	[M -(CO2)]|
M194.0449T309	group41	#M150.0555T309|+(CO2)(0.000410)|#M195.0485T309|-(13C)(0.000212)|#M240.0513T309|-(HCOOH)(0.000842)|#M292.0214T309|-(H3PO4)(0.000399)|#M292.0134T309|-(H2SO4)(0.001136)|	1	M194.0449T309	M
M195.0485T309	group41	#M194.0449T309|+(13C)(0.000212)|#M241.0536T309|-(HCOOH)(0.000333)|	0	M194.0449T309	[M +(13C)]|
M240.0513T309	group41	#M194.0449T309|+(HCOOH)(0.000842)|#M241.0536T309|-(13C)(0.000963)|	0	M194.0449T309	[M +(HCOOH)]|
M241.0536T309	group41	#M195.0485T309|+(HCOOH)(0.000333)|#M391.1047T309|-(C5H10O5)(0.001790)|#M240.0513T309|+(13C)(0.000963)|	0	M194.0449T309	-
M292.0134T309	group41	#M194.0449T309|+(H2SO4)(0.001136)|	0	M194.0449T309	[M +(H2SO4)]|
M292.0214T309	group41	#M194.0449T309|+(H3PO4)(0.000399)|	0	M194.0449T309	[M +(H3PO4)]|
M389.0963T309	group41	#M391.1047T309|-(13C2)(0.001640)|#M390.1012T309|-(13C)(0.001499)|	0	M194.0449T309	-
M390.1012T309	group41	#M391.1047T309|-(13C)(0.000141)|#M389.0963T309|+(13C)(0.001499)|	0	M194.0449T309	-
M391.1047T309	group41	#M389.0963T309|+(13C2)(0.001640)|#M241.0536T309|+(C5H10O5)(0.001790)|#M390.1012T309|+(13C)(0.000141)|	0	M194.0449T309	-

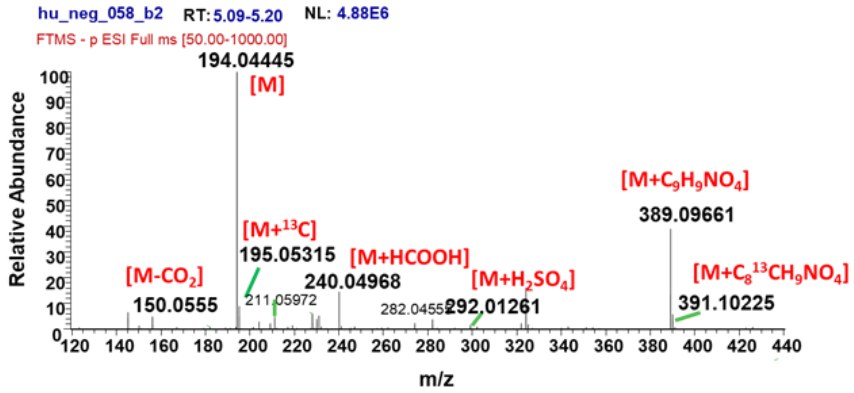


Figure 5. Example of annotation of mass difference between features using ACorF (mass threshold 0.002 Da) and the comparison of an expert annotated raw spectrum.

On a global point of view concerning the Sacurine dataset annotation, ACorF provided in-source annotation from the ions of the dataset for 100% of the grouped variables, as it is one parameter of the grouping process, whereas CAMERA proposed annotation using hypothetic observable ions for 70% of its grouped features.

3.4. Use and Configuration

ACorF can be used to process either LC- or GC-MS data; recommendations for different parameter settings for a successful filtration are the following:

For UPLC/HRMS data, default parameters can be the following: (i) if a Pearson correlation is used, the default threshold can be set at 0.90; (ii) a delta RT of 0.1 min or adjusted depending on chromatographic systems; (iii) the use of the list of known adduct/isotope mass differences with a mass delta of 0.005 Da or adjusted depending on MS resolution; and (iv) the choice of the ion with the highest intensity as the representative ion.

For GC/HRMS datasets, we recommend to first filter the dataset to remove unspecific fragments of derivative agents. Then, the same parameters as above can be used, but we recommend choosing the ion with the highest mass among the top highest intensity as representative. As an example, ACorF was also applied to process a GC-MS dataset publicly available on W4M named 'Algae' (W4M00004_GCMS-Algae). For GC-MS use, ACorF also allowed for the subdivision of some pc-groups into smaller ones (see supplemental Figure S4). However, the main added value compared to CAMERA is the ability to perform the grouping using annotation (versus 40% annotations in grouped features with CAMERA) and choosing an adequate quantifier ion, more specific than the highest intensity ion.

Different output files are produced for further data analysis steps within workflows. To perform dataset filtration following the ACorF utilisation, we encourage users to work with tools available in the W4M instance. The "Generic Filter" can be used to exclude all lines with a "0" value in the ACorF_filter column of the variableMetadata and to provide the filtered DataMatrix for further statistical analysis in W4M.

Finally, the users can make use of group information for metabolite annotation, especially for LC-MS, by generating sub-files corresponding to the different groups or performing queries in databases using the DataMatrix filtered with representative ions.

4. Conclusions

We introduced a new tool, ACorF, which allows identifying and filtering the analytical redundancy within metabolomics LC- and GC-MS datasets. The developed algorithm uses three independent key parameters (any similarity measurement, retention time, and mass difference between features) to group ions part of the same metabolites and the intensity information to propose a representative feature of the group. Finally, as a key element within metabolomics data analysis workflow, this tool will be available via the web-based galaxy platform W4M with generic input tabular files and propose different output files for visualisation for further data analysis within workflows.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: Total overview of correlation > 0.75 within the Sacurine dataset, Figure S2: Presentation of the 3 input files for the analytic correlation filtration tool. Figure S3. Example of annotation of mass difference between features for early coeluting compounds using ACorF (mass threshold 0.002 Da) in comparison to CAMERA. Figure S4: Bar diagram presenting the number of groups (x-axis) by the group size (y-axis: number of ions per group) obtained from the GC-MS dataset (W4M00004_GCMS-Algae).

Author Contributions: study design: M.P., E.P.-G.; software: S.M.; data interpretation: M.P., B.L.; work supervision: E.P.-G, P.G. and B.C.; original draft preparation: S.M., M.P., E.P.-G. and B.C.; all authors contributed to the manuscript, read, and approved the final version of the article.

Funding: This work was supported by (i) the INRA DID'IT metaprogramme: S. Monnerie is the recipient of a doctoral fellowship, (ii) the Agence Nationale de la Recherche (MetaboHUB national infrastructure for metabolomics and fluxomics, ANR-11-INBS-0010 grant).

Acknowledgments: We thank Yann Guitton for the reviewing of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nicholson, J.K.; Lindon, J.C.; Holmes, E. 'Metabonomics': Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* **1999**, *29*, 1181–1189.
2. Ramautar, R.; Berger, R.; van der Greef, J.; Hankemeier, T. Human metabolomics: Strategies to understand biology. *Curr. Opin. Chem. Biol*. **2013**, *17*, 841–846.
3. Forcisi, S.; Moritz, F.; Kanawati, B.; Tziotis, D.; Lehmann, R.; Schmitt-Kopplin, P. Liquid chromatography-mass spectrometry in metabolomics research: Mass analyzers in ultra-high pressure liquid chromatography coupling. *J. Chromatogr. A* **2013**, *1292*, 51–65.
4. Kuehnbaum, N.L.; Britz-McKibbin, P. New advances in separation science for metabolomics: Resolving chemical diversity in a post-genomic era. *Chem. Rev*. **2013**, *113*, 2437–2468.
5. Fuhrer, T.; Zamboni, N. High-throughput discovery metabolomics. *Curr. Opin. Biotechnol.* **2015**, *31*, 73–78.
6. Alonso, A.; Marsal, S.; Julia, A. Analytical methods in untargeted metabolomics: State of the art in 2015. *Front. Bioeng. Biotechnol*. **2015**, *3*, 23.
7. Kuhl, C.; Tautenhahn, R.; Bottcher, C.; Larson, T.R.; Neumann, S. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem*. **2012**, *84*, 283–289.
8. Smith, C.A.; Want, E.J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem*. **2006**, *78*, 779–787.
9. Alonso, A.; Julia, A.; Beltran, A.; Vinaixa, M.; Diaz, M.; Ibanez, L.; Correig, X.; Marsal, S. AStream: An R package for annotating LC/MS metabolomic data. *Bioinformatics* **2011**, *27*, 1339–1340.
10. Senan, O.; Aguilar-Mogas, A.; Navarro, M.; Capellades, J.; Noon, L.; Burks, D.; Yanes, O.; Guimerà,, R.; Sales-Pardo, M. CliqueMS: A computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network. *Bioinformatics* **2019**, *35*, 4089–4097.
11. Uppal, K.; Walker, D.I.; Jones, D.P. xMSannotator: An R package for network-based annotation of high-resolution metabolomics data. *Anal. Chem.* **2017**, *89*, 1063–1067.
12. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; Bourne, P.E.; Bouwman, J.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018.
13. Guitton, Y.; Tremblay-Franco, M.; le Corguillé, G.; Martin, J.F.; Pétéra, M.; Roger-Mele, P.; Delabrière, A.; Goulitquer, S.; Monsoor, M.; Canlet, C; et al. Create, run, share, publish, and reference your LC-MS, FIA-MS, GC-MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics. *Int. J. Biochem. Cell Biol*. **2017**, *93*, 89–101.
14. Giacomoni, F.; le Corguillé, G.; Monsoor, M.; Landi, M.; Pericard, P.; Pétéra, M.; Duperier, C.; Tremblay-Franco, M.; Martin, J.-E.; Goulitquer, S.; et al. Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics. *Bioinformatics* **2015**, *31*, 1493–1495.
15. Carusi, A.; Reimer, T. *Virtual Research Environment Collaborative Landscape Study*; UK's Joint Information Systems Committee (JISC): *Bristol*, UK, 2010.
16. Goecks, J.; Nekrutenko, A.; Taylor, J.; Galaxy, T. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **2010**, *11*, R86.
17. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504.
18. Pujos-Guillot, E.; Brandolini, M.; Pétéra, M.; Grissa, D.; Joly, C.; Lyan, B.; Herquelot, É.; Czernichow, S.; Zins, M.; Comte, B.; et al. Systems metabolomics for prediction of metabolic syndrome. *J. Proteome Res.* **2017**, *16*, 2262–2272.

Si ACorF a initialement été développé pour distinguer les corrélations analytiques, il peut être utilisé par la suite pour regrouper les redondances autres (souvent d'origine biologique). En effet, une fois les redondances analytiques supprimées, il est possible de l'utiliser à nouveau en ne prenant plus en compte les paramètres de masse et de temps de rétention. Se baser uniquement sur les corrélations restantes revient à regrouper entre elles les variables biologiquement liées (voies métaboliques communes notamment).

Au cours du développement de l'outil ACorF nous avons réalisé des tests sur des jeux de données d'origine analytique différente afin de produire des « recommandations utilisateur ». Nous avons appliqué ces dernières à une partie de nos jeux de données : C18-pos, C18-neg, Hilic et GC. Pour se faire, nous avons généré une matrice de corrélation à partir d'un outil Galaxy disponible sur une instance locale : « Between Table Correlation » en prenant en compte les corrélations de Pearson entre les variables et en appliquant une correction BH aux p-value relatives à la significativité de cette corrélation. ACorF a ensuite été utilisé avec les paramètres suivants : (i) un taux de corrélation > 0,9, (ii) un seuil de temps de rétention de +/- 0,1 minutes, (iii) des différences de masses égales à celles de la liste à +/- 0,005 Da, (iv) la liste de différences de masse produite par la PFEM et (v) un choix du représentant, dépendant du jeu de données : le plus intense pour les jeux de données C18_pos, C18-neg et Hilic, l'ion de plus forte masse parmi les 3 plus intenses pour la GC.

La redondance analytique a également été filtrée pour le jeu de données Lipidomique et la RMN mais par le biais des annotations. En effet, il s'agit des seuls jeux pour lesquels l'annotation a été faite dès la fin du prétraitement. Les ions annotés comme étant le même métabolite ont donc été éliminés pour ne conserver qu'un représentant : le plus intense en Lipidomique et le plus pur en RMN. Avec cette dernière, le petit nombre de variables (74 buckets) et la facilité d'annotation par comparaison des spectres purs aux bases de données justifie cette approche. Pour la lipidomique, l'annotation putative est faite très tôt du fait de l'existence d'une base de données interne très fournie au sein de MetaboHUB ; elle permet ce type de filtration mais cela est difficilement envisageable pour les autres jeux de données.

Cette étape de filtration analytique a permis de réduire le nombre de variables à 2 915. Les résultats détaillés par méthodes sont présentés dans la **Figure 25**. Sur les jeux de données filtrés par ACorF, 100% des variables présentant des corrélations de Pearson > 0,9 ont été regroupées avec au moins une autre variable pour des raisons analytiques.

	C18-Pos	C18-Neg	Hilic	Lipido	GC	RMN	TOTAL
Prétraitement	1656	606	1124	6697	3745	74	13 902
Nombre de corrélations > 0,9 au sein de la matrice de corrélation	1177	56	70	1523	210	5	3 041
Filtration des redondances analytiques	1091 (ACorF)	249 (ACorF)	612 (ACorF)	571 (filtration annotation)	345 (ACorF)	47 (filtration annotation)	2 915

Figure 25 : Récapitulatif de l'étape de filtration des corrélations analytiques à l'aide de l'outil ACorF.

3. ANOVA à mesure répétée

Nous avons ensuite cherché à identifier les ions stables entre les 2 temps de mesure et significatifs ou non du statut cas/témoins. Pour ce faire, nous avons réalisé une ANOVA à mesure répétée sur chacun des jeux de données. Le choix de cette méthode statistique résulte de la volonté de prendre en compte la dimension temporelle de nos données avec des mesures répétées dans le temps pour chacun des sujets. En effet, l'analyse de variance sur mesures répétées est basée sur le même principe que l'ANOVA classique, mais la variable d'intérêt est mesurée plusieurs fois et chacune de ces mesures est appelée répétition. Tout comme en ANOVA classique, il est possible d'inclure dans ce type de modèle des facteurs d'interaction. L'un des facteurs pouvant être généralement associé au SMet dans un modèle ANOVA est l'IMC, qui est connu comme facteur potentiellement confondant du fait de sa corrélation forte avec le statut. Dans notre cas des analyses préalables (ANOVA à 2 facteurs et prise en compte de l'IMC comme facteur confondant) ont démontrées qu'il n'était pas nécessaire de prendre en compte ce dernier.

L'ANOVA à mesure répétée a permis d'obtenir les résultats présentés dans le **Tableau 10**. Trois types d'informations peuvent être extraits de cette analyse :

- Les métabolites non significatifs du temps c'est-à-dire stables entre T1 et T4.
- Les métabolites significatifs du statut cas/témoins.
- Les métabolites significatifs du statut cas/témoins ET non significatifs du temps, c'est-à-dire stables dans le temps.

Jeux de données	Nombre de variables avant modèle mixte	Nombre de variables stables dans le temps	Nombre de variables significatives du statut	Nombre de variables significatives du statut et stables dans le temps
C18-Pos	1 091	646	156	105
C18-Neg	249	221	55	54
Hilic	612	498	93	69
Lipido	571	446	255	197
GC	345	338	35	35
RMN	47	27	28	16

Tableau 10 : Résultat des ANOVA à mesure répétée sur chacun des jeux de données (p-value corrigée (BH) < 0,05)

Les ANOVA à mesure répétée ont permis de faire ressortir 476 variables significatives du statut cas/témoins qui sont stables dans le temps (soit 16% des données) : 105 en C18-Pos, 54 en C18-Neg, 69 en HILIC, 197 en Lipidomique, 35 en GC et 16 en RMN. Tous ces métabolites sont modulés dans le cadre du SMet, il convient donc de les considérer pour mieux comprendre la pathologie et les mécanismes qu'elle implique.

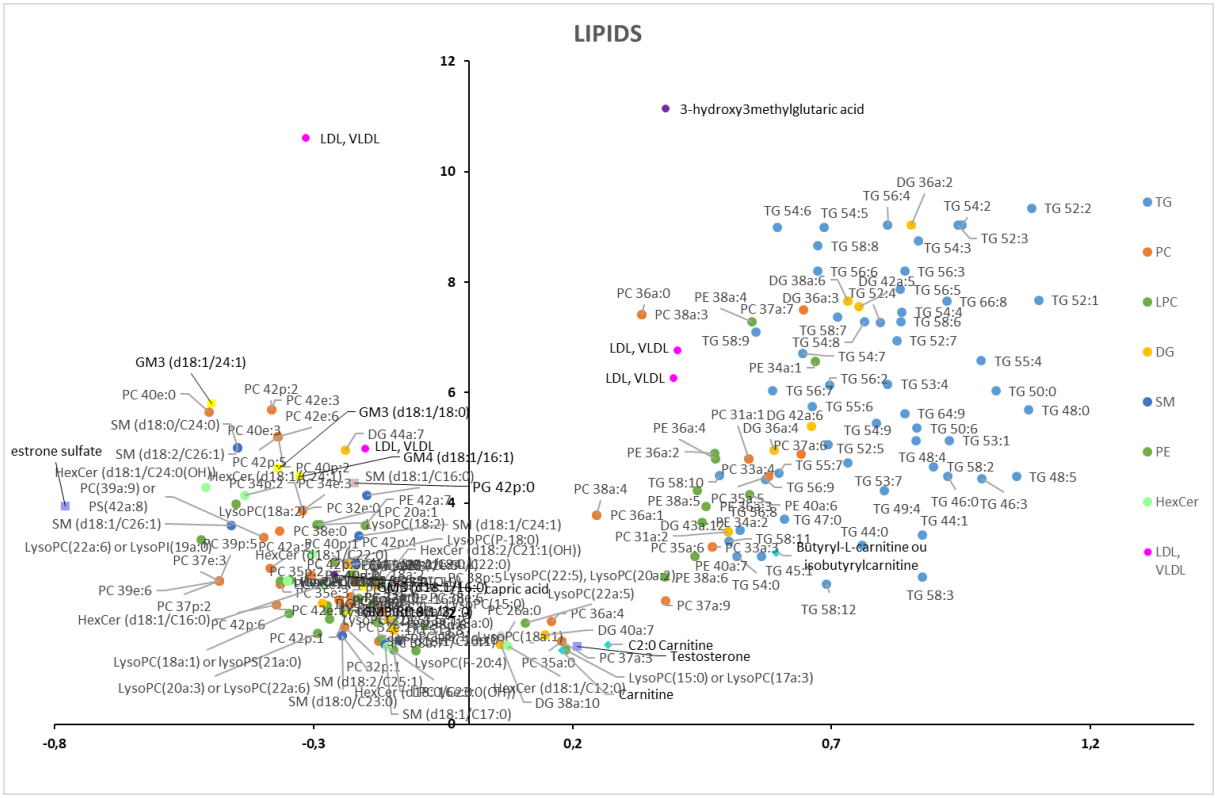
II. Interprétation des résultats

1. Annotation des ions et buckets

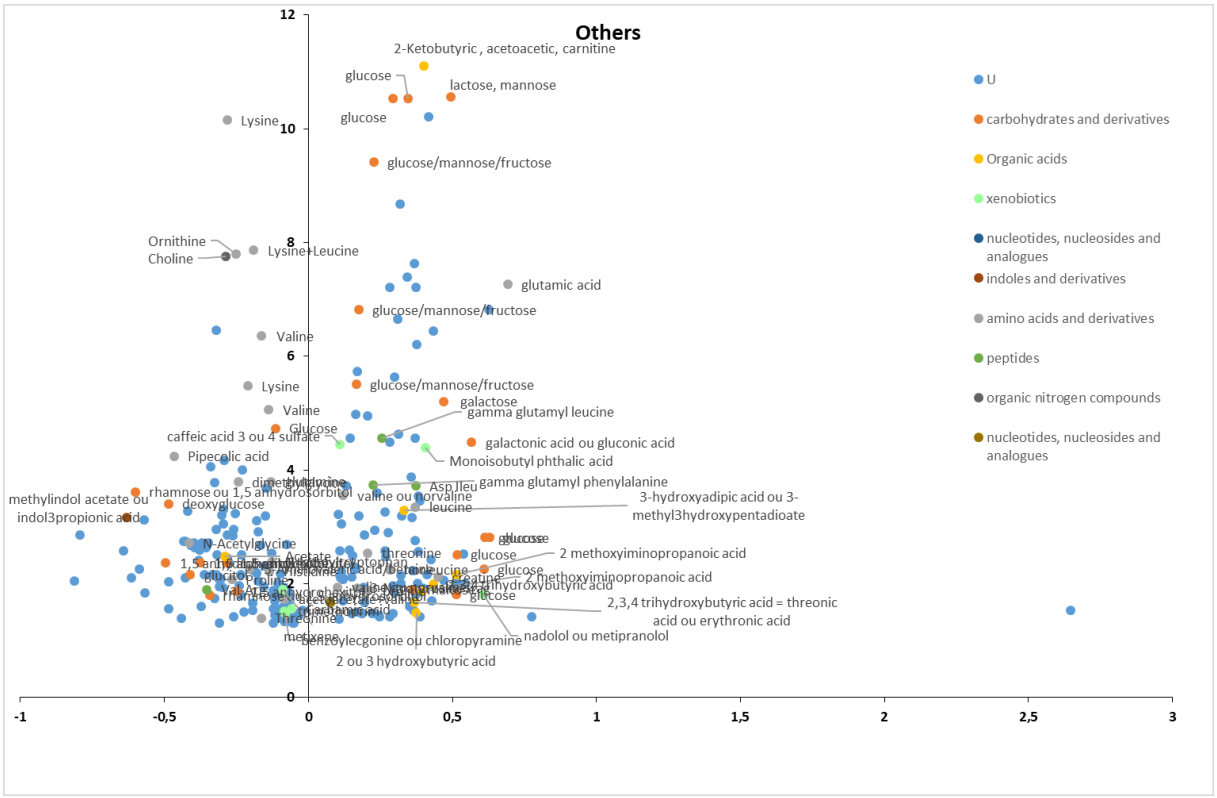
Pour interpréter ces résultats il convient de donner une signification biologique aux variables grâce à leur annotation. Celle-ci a été réalisée à l'aide de banques internes (PFEM et MetaboHUB), de bases de données externes telles que HMDB, Lipidmaps, Metlin, etc… mais également de celle que nous avons créée à partir des biomarqueurs de la littérature identifiés lors de la revue systématique.

Cette étape a permis l'annotation de 150 métabolites parmi les 476 significatifs du statut cas/témoins (soit 31%), dont 72 lipides, avec des degrés de certitude variables (niveaux 1, 2, 3) [79]. Il est à ce stade possible que des métabolites soient présents plusieurs fois du fait de la redondance inter-méthodes. Les identifications sont en cours de validation par des experts chimistes et biologistes. Pour chacune des 150 annotations, des identifiants de base (HMDB, KEGG ou ChEBI) ont été associés.

La **Figure 26** met en évidence l'ampleur des changements observés pour les 476 métabolites (228 lipides **Figure 26A** et 248 autres **Figure 26B**) significatifs du statut cas/témoins et ce en fonction de leur valeur de p-value. Environ 50% de ces molécules sont positivement associées au SMet, avec une significativité forte (p-value corrigée < 10^{-5}) : les glucides, les acides aminés et certains lipides (principalement les TG) déjà identifiés. Parmi les molécules négativement modulées, nous avons pu identifier des métabolites provenant de la nourriture (exemple avec le 1,5-anhydrohexitol), du métabolisme microbien (acide indole propionique, acide pipécolique) et différents dérivés des acides aminés ainsi que des peptides (p-values allant de 10^{-3} à 10^{-5}). Lorsque les lipides sont regroupés en classes, seul les TGs et les gangliosides sont significatifs (respectivement 7,4 x 10^{-10} et 9,10 x 10^{-6}). D'autres classes de lipides, comme les phosphatidylcholines (PC), les phosphatidylethanolamines (PE) ou les acides gras (FA) sont très hétérogènes avec des composés individuels très significativement modulés (l'acide 3-hydroxy-3-methylglutarique, PC(38a :3), PE(38a :4)).



A : Volcanoplot représentant les lipides stables dans le temps et significativement modulés par le SMet



B : Volcanoplot des métabolites non lipidiques stables dans le temps et significativement modulés par le SMet

Figure 26 : Volcano plot représentant les valeurs de p-value des 476 métabolites significatifs du statut cas/témoins et stable au cours du temps, en fonction de leurs valeurs de fold change.

2. Intégration des données dans les réseaux métaboliques

Comme évoqué dans l'introduction, la biologie des systèmes permet une meilleure compréhension du fonctionnement des êtres vivants. Elle étudie les interactions entre les différents composants d'un système biologique, d'un être vivant. Pour se faire, elle tente d'intégrer différentes informations à différentes échelles : cellules, tissus, gènes, protéines, métabolites… Pour parvenir à une meilleure appréhension des interdépendances existantes, l'un des moyens les plus performants pour compiler ces informations est la construction de réseaux biologiques complexes de plusieurs niveaux hiérarchiques. Il reflète une lecture fonctionnelle de l'état de l'organisme. De ce fait, la génération de représentation au sein de réseaux métaboliques enrichis est aujourd'hui l'un des enjeux majeurs en métabolomique afin de permettre une interprétation la plus complète possible.

Nous avons envisagé d'interpréter les variations observées en utilisant l'outil KEGG-pathway qui se base sur les identifiants KEGG obtenus lors de l'annotation. Toutefois, il offre des représentations en 2 dimensions figées des réseaux. La principale conséquence réside dans le fait qu'il est parfois difficile de se rendre compte de la distance graphique entre 2 métabolites représentés par KEGG-pathway. En effet, 2 métabolites reliés par une réaction mais impliqués dans 2 voies différentes peuvent être très éloignés physiquement dans le réseau. C'est par exemple, le cas dans la voie de dégradation du glycogène en glucose-6-phosphate où des métabolites directement reliés par une réaction ont des positions projetées dans la visualisation du réseau éloignées (**Figure 27**).

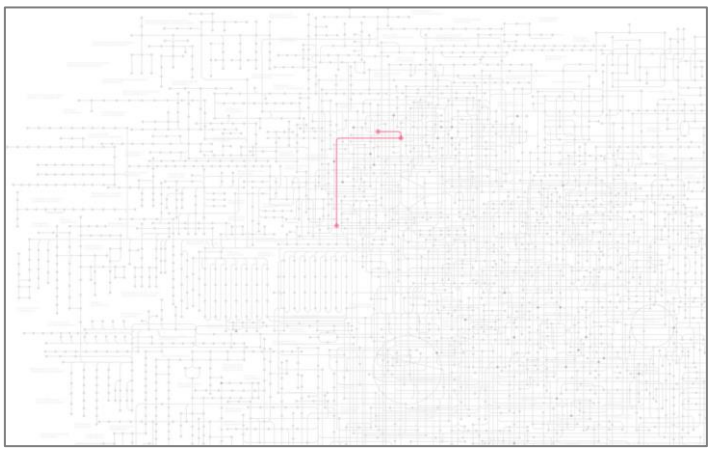


Figure 27 : Réseau métabolique humain issu de KEGG pathway avec coloration de la voie de dégradation du glycogène en glucose-6-phosphate.

Alternativement, MetExplore [234] est un serveur web en ligne permettant de lier les métabolites identifiés en métabolomique non ciblée dans des réseaux métaboliques reconstruits à l'échelle du génome. La plupart des outils de visualisation replace les métabolites au sein de voies métaboliques séparées et déconnectées, négligeant ainsi la plupart des connexions entre les voies. MetExplore utilise des réseaux reconstruits de différentes espèces biologiques (KEGG [235], BioCYC [236], Recon [131]...) pour insérer les métabolites identifiés d'intérêt, puis applique des méthodes fondées sur la théorie des graphes, permettant de les lier avec des distances les plus faibles possibles entre eux. Il propose ensuite des outils de visualisation 2D dynamique pour augmenter la capacité d'analyse et d'interprétation des données. Nous avons choisi de l'utiliser pour générer des réseaux et aider à l'interprétation des informations issues de la liste des métabolites modulés par le SMet.

2.1. La problématique des identifiants de métabolites

La métabolomique génère de grandes quantités de connaissance relatives aux entités chimiques que sont les métabolites. Néanmoins, celle-ci peut parfois manquer de robustesse dans la manière dont elle est représentée. En effet, comme mentionné dans les chapitres précédents, il est fréquent que les noms donnés aux molécules varient d'une publication à une autre, et que les identifiants ne permettent pas de lien entre bases de données. La notion d'ontologie, vocabulaire contrôlé et hiérarchisé décrivant des concepts dans un domaine précis, existante par exemple dans le domaine de la génomique et de la transcriptomique, demeure complexe en métabolomique. En effet, c'est un domaine qui se situe à la frontière entre plusieurs spécialités (chimie, biochimie, biologie…) dont les vocabulaires et ontologies sont parfois déjà établis et ne sont pas toujours similaires d'un domaine à l'autre. Ce manque d'uniformité dans les appellations engendre une complexité supplémentaire dans le mapping des données métabolomiques au sein des réseaux avec notamment de nombreuses redondances d'appellation parfois difficiles à détecter. En 2015, Schlegel *et al* ont mis en évidence la nécessité de définir l'utilisation de termes communs [237].

En attendant la mise en place d'une nomenclature unique, la problématique des identifiants en métabolomique est devenue un facteur limitant pour le mapping des métabolites. En effet, lors de l'annotation des composés, les nomenclatures utilisées par les utilisateurs peuvent être diverses : noms communs liés au domaine de la biologie/santé, noms chimiques de l'Union Internationale de Chimie Pure et Appliquée (UICPA ou IUPAC en anglais), etc… Des identifiants se basant sur la structure chimique des composés ont notamment été mis en place. Ils ont été créés par la communauté des chimistes afin de traduire de façon plus précise des informations sur la constitution et la conformation

des molécules (formule chimique, nombre et type de liaisons carbones…) et ce dans un format lisible informatiquement. Les plus répandus dans le domaine de la métabolomique sont les formats SMILES (Simplified Molecular Input Line Entry Specification), InChI (IUPAC International Chemical Identifier) et sa forme raccourcie l'InChIKey.

En parallèle, de nombreuses bases de données (présentées dans d'introduction) ont vues le jour pour faciliter les étapes d'annotation des métabolites. Elles ont pour but de rassembler des informations spectrales, chimiques ou encore biologiques pour faciliter l'interprétation des données. Chacune d'elles couvre en général un type, un domaine, ou un organisme précis : HMDB couvre les métabolites présents dans l'espèce humaine ; KEGG Compound (Kyoto Encyclopedia of Genes and Genomes - Compound) regroupe les petites molécules impliquées dans les réseaux métaboliques d'une grande variété d'organismes ; ChEBI regroupe les composés chimiques naturels et synthétiques de petites masses.

Les 150 métabolites précédemment identifiés sont tous associés à un identifiant HMDB. Afin d'obtenir les identifiants ChEBI et/ou KEGG, nous avons utilisé un script Perl permettant de récupérer ces informations directement par lecture du fichier « xml » de chaque identifiants HMDB disponible en ligne. Nous avons ensuite fait le choix de représenter nos métabolites dans un réseaux issu de Recon 2.03 [131] enrichi et non compartimenté disponible dans MetExplore [234] : le réseau numéro 3223. En effet, il s'agit de l'un des réseaux les plus complets disponible dans MetExplore pour l'espèce humaine, tout en étant le seul réseau non compartimenté et donc adapté à la représentation de métabolites mesurés dans des biofluides comme le sérum. Pour effectuer ce travail, il est nécessaire d'avoir recours à des identifiants spécifiques des réseaux tel que Recon. Des outils ont été développés (Metabolite Identifier Matcher de MetExplore, Chemical Translation Service, Enrichment Analysis dans MetaAnalyst…) pour permettre les liens et conversions entre les bases de données d'annotation et les identifiants de réseaux.

name_HMDB	HMDB	ChEBI	KEGG	Recon via ChEBI		Recon via KEGG
				ID	Distance ChEBI	
L-Leucine	HMDB0000687	15603	C00123	M_leu_L	0	M_leu_L
L-Methionine	HMDB0000696	16643	C00073	M_met_L	0	M_met_L
Butyrylcarnitine	HMDB0002013	21949	C02862	M_c4crn	1	M_HC02150 ; M_c4crn
cis-Aconitic acid	HMDB0000072	32805	C00417	No match	-	M_HC00342
L-Lysine	HMDB0000182	18019	C00047	M_lys_L	-0,1	M_lys_L
2-Ketobutyric acid	HMDB0000005	30831	C00109	M_2obut	-0,1	M_2obut

Tableau 11 : Résultat de la conversion des identifiants HMDB en identifiants Recon par l'intermédiaire des identifiants ChEBI ou KEGG.

Tous ces outils se basent sur les correspondances établies entre différents identifiants : InChiKey, ChEBI, KEGG, SMILES, PubChem, HMDB, LipidMap ou encore SwissLipids pour effectuer les conversions. En complément à cela, Metabolite Identifier Matcher utilise les ontologies proposées par la base de données ChEBI pour tenter de représenter par une famille le métabolite recherché lorsque ce dernier n'est pas disponible directement dans Recon. Le **Tableau 11** illustre quelques exemples de métabolites pour lesquels la conversion d'un ID HMDB vers un Recon a pu être faite en utilisant soit les identifiants ChEBI, soit les KEGG. Il est fréquent que le matching se fasse parfaitement avec une correspondance entre les résultats ChEBI et KEGG. Le fait d'utiliser 2 types d'identifiants augmente les chances de trouver une correspondance Recon. Par exemple, pour l'acide cis-aconitique aucun résultat n'est trouvé à partir du ChEBI alors que le KEGG permet d'obtenir un identifiant Recon. De plus, il est possible que les résultats diffèrent ou soient multiples comme pour la butyrylcarnitine pour laquelle on trouve 2 matchs dans Recon en utilisant l'identifiant KEGG (le M_c4crn correspond à la butyrylcarnitine et le M_HC02150 à la butanoylcarnitine). L'utilisation de l'identifiant ChEBI ne donnant qu'une seule correspondance, elle aide à trancher en faveur de l'identifiant M_c4crn. Finalement, il est possible qu'il soit nécessaire d'utiliser les ontologies ChEBI pour remonter à des identifiants distants lorsqu'aucune correspondance n'est trouvée. L'exemple présenté sur la **Figure 28** est celui de la L-lysine dont l'identifiant ChEBI est le 18019. Lorsque l'on tente de le faire correspondre avec Recon, aucun métabolite n'est proposé. Toutefois, en activant l'option « Class identifier for Chebi » de l'outil, il propose en correspondance non exacte l'identifiant 32551 qui correspond au L-lysinium (+1), permettant alors de trouver l'identifiant Recon « M_lys_L ». La L-lysinium (1+) est en fait un conjugué de la L-lysine qui se trouve être représenté dans Recon alors que cette dernière ne l'est pas. L'outil propose donc cette substitution en indiquant une correspondance non exacte avec une distance dans l'arbre d'ontologie de -0,1 (le signe représentant le sens d'évolution dans l'arbre d'ontologie, - indique que l'on descend, + que l'on remonte ; les chiffres avant la virgule correspondent aux niveaux dans l'arbre, les chiffres après à des conformations chimiques différentes).

Nous avons donc utilisé l'outil Metabolite Identifier Matcher de MetExplore pour convertir les identifiants de bases de données en identifiants Recon.

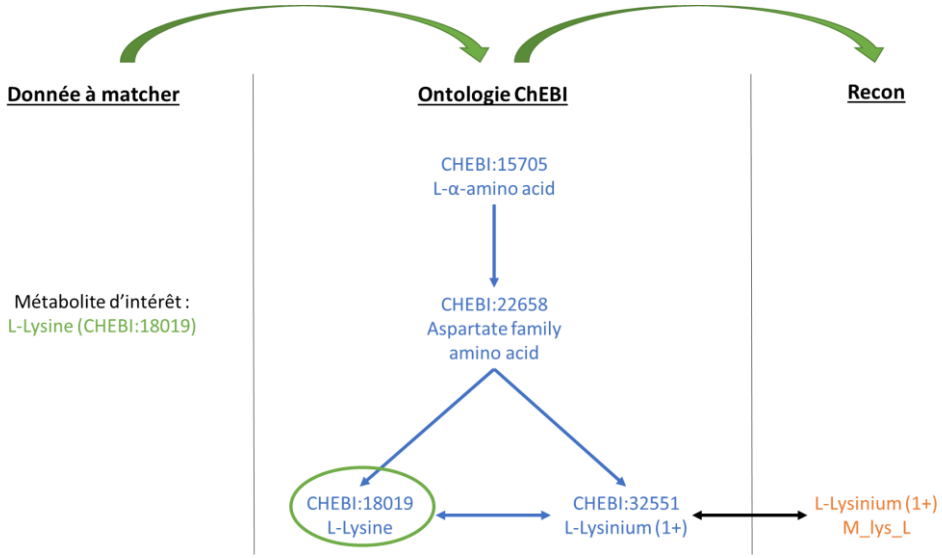


Figure 28 : Illustration de l'utilisation des ontologies ChEBI pour trouver des correspondances Recon.

Le récapitulatif des étapes d'annotation et de transition entre identifiants est présenté dans la **Figure 29**.

	Métabolites hors lipides	Lipides
Nombre de variables stables dans le temps et significatives du statut	265	211
Nombre de variables identifiées	74	76
Nombre de variables avec identifiant HMDB	74	76
Nombre de variables avec identifiant ChEBI et/ou KEGG	72	75
Nombre de variables avec identifiant Recon	48	58
Nombre d'identifiants Recon uniques	27	12

Figure 29 : Récapitulatif du nombre de variables/identifiants obtenu aux différentes étapes de d'annotation et conversion d'identifiants entre la détection des variables stables et significatives du statut et la représentation dans les réseaux.

Parmi les 150 métabolites ayant pu être identifiés, 76 sont des lipides. Ces derniers résultent dans l'obtention de seulement 12 identifiants Recon, ce qui est très faible (16%). Cette différence s'explique par le fait que certains métabolites, et plus particulièrement les lipides, ne sont pas représentés dans Recon de manière unique, mais sous forme de famille. Par exemple les triglycérides comme TAG (44:1) et TAG (44:0) sont regroupés sous un même identifiant Recon « M_tag_hs ».

Au final, sur les 476 métabolites initiaux, seul 18% peuvent être représentés dans un réseau et ce avec une simplification de l'information importante puisque la moitié d'entre eux sont inclus dans une famille. Il est important de tenir compte de ces regroupements en famille lors de l'interprétation des données puisque des métabolites d'une même famille peuvent avoir des variations différentes. Par exemple, une étude des corrélations au sein de la famille des phosphatidylcholines a, par exemple, mis en évidence 2 groupes de variables distincts et non corrélés (**Figure 30**). D'autre part, certains métabolites mesurés sont exogènes comme le triméthoprime ou le methixène qui sont des médicaments ; il est donc cohérent de ne pas les retrouver dans les identifiants Recon.

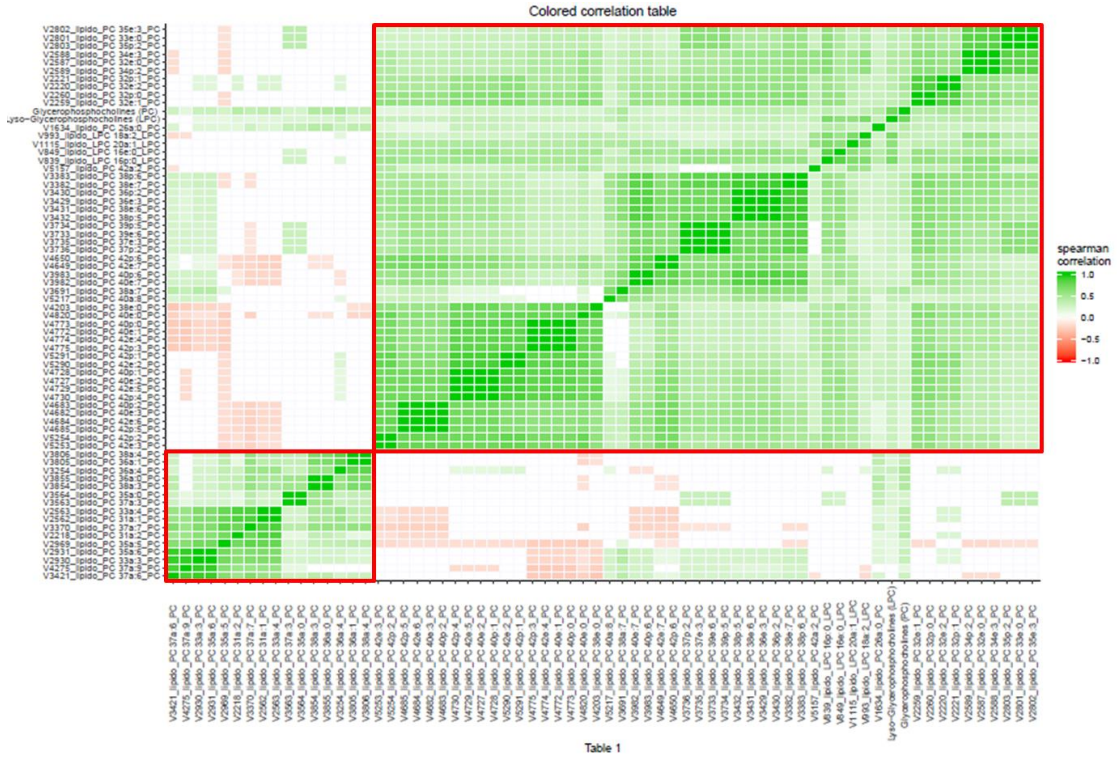


Figure 30 : Représentation graphique de la matrice de corrélations entre les métabolites de la famille des phosphatidylcholines.
Vert : corrélation positive ; Rouge : corrélation négative (significative p-value corrigé (BH) <0,05)

Toutefois, même grâce aux ontologies, il n'est pas toujours possible de faire de lien entre les identifiants. Il devient alors impossible de construire un réseau métabolique complet **basé sur l'entièreté des données (métabolites) annotées.** Ceci constitue l'une des principales limites actuelles dans l'intégration de la métabolomique dans des approches systèmes.

2.2. Génération d'un réseau *via* l'outil KEGG Pathway Database

Suite au travail de traduction des identifiants, 45 métabolites ont été annotées avec un identifiant KEGG, et ont été insérés dans le réseau humain KEGG grâce à l'outil KEGG Pathway. Trente-six d'entre eux ont pu être mappés dans les voies métaboliques : de façon globale, 15 dans les voies des transporteurs ABC, 13 dans les voies de biosynthèse des acides aminés, 12 dans les voies de biosynthèse des aminoacyl-tRNA (**Figure 30**).

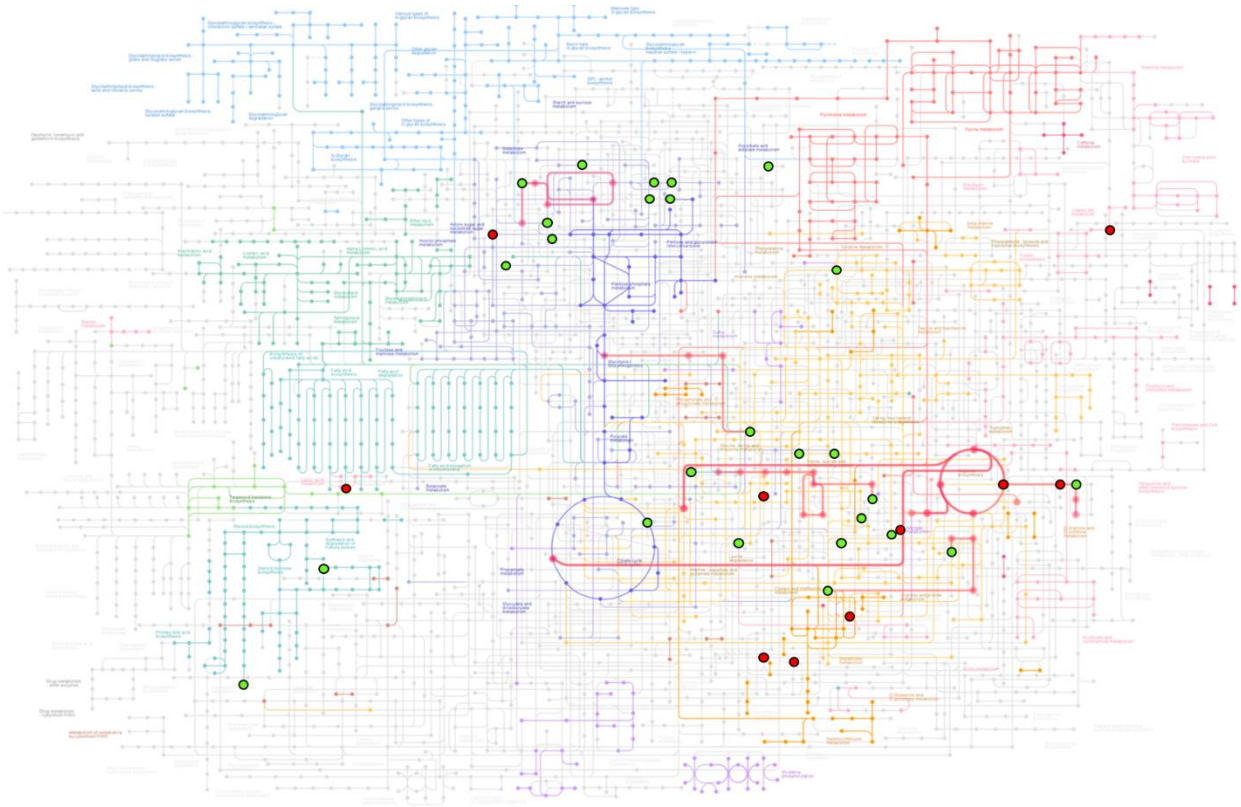


Figure 30 : Visualisation des 36 métabolites mappés dans les voies métaboliques du réseau KEGG homo sapiens.
Vert : positivement modulé ; Rouge : négativement modulé.

Afin de mieux interpréter nos processus d'intérêt impliquant plusieurs voies métaboliques et d'effectuer un travail d'enrichissement pour mettre en évidence les voies sur et sous-représentées, l'outil MetExplore a été utilisé.

2.3. Génération d'un réseau *via* l'outil MetExplore

Comme évoqué précédemment, nous avons fait le choix de construire nos réseaux dans l'outil MetExplore, à partir du réseau numéro 3223. Nous avons ensuite utilisé 3 stratégies différentes pour tenter de connecter avec le plus de pertinence possible nos 38 identifiants Recon :

- La première méthode a consisté à réaliser un enrichissement de voies à partir d'un réseau filtré sur les réactions impliquant au moins l'un des 38 métabolites. Cette méthode recherche le moyen le plus optimisé de relier entre eux les métabolites d'intérêt, en supprimant les réactions qui complexifient l'information. Malgré tout, elle engendre souvent la construction de réseaux encore très complexes et difficiles d'interprétation.
- La seconde méthode a filtré la liste des réactions à l'aide de l'outil « Subnetwork Extraction » développé par l'équipe de MetExplore. Ce dernier recherche lechemin le plus léger entre les 398 métabolites en se basant sur une méthode qui prend en compte le suivit des atomes. Elle permet de recréer ensuite un réseau à partir des réactions permettant ces liens.
- Enfin, la dernière méthode utilise la matrice de distance fournie par l'outil « Subnetwork Extraction » lors de l'analyse précédente, construit une heatmap à partir de ces valeurs pour ensuite détecter la présence de sous-groupes de métabolites proches en distance. Nous avons cherché à construire des sous-réseaux en appliquant une filtration des réactions non pas sur la base des 38 métabolites, mais sur les sous-groupes observés. Ces sous-réseaux ont pour but de faciliter la tâche d'interprétation des réseaux lorsque les 2 premiers réseaux sont trop gros ou trop complexes.

Le réseau obtenu par la méthode 1, après suppression des « sides compounds » (composés chimiques intervenant dans un grand nombre de réactions en tant que cofacteurs) est présentée dans la **Figure 31**.

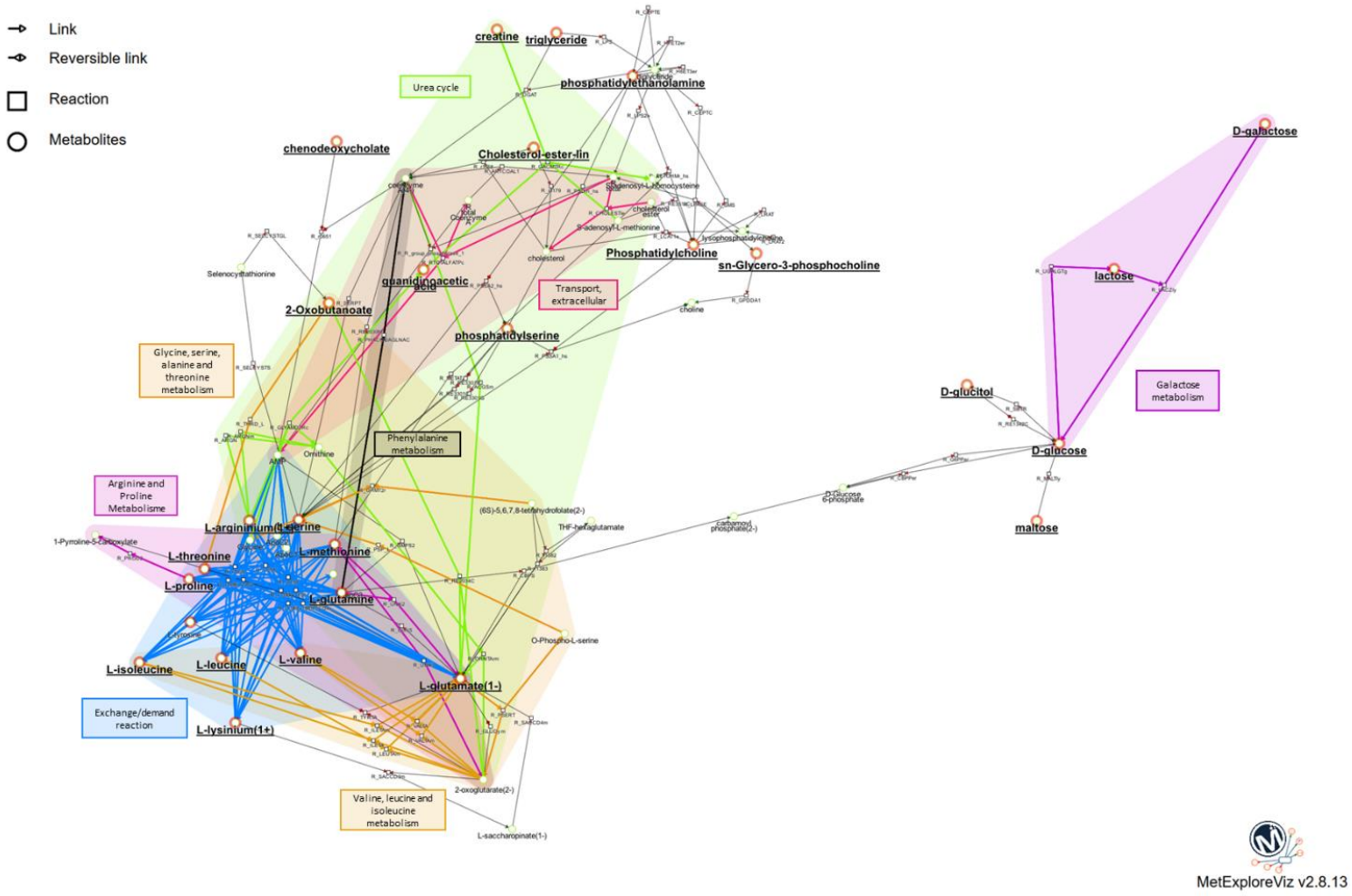


Figure 31 : Réseau métabolique obtenu avec MetExplore.
Coloration des voies métaboliques significativement différentes entre cas et témoins.

Le **Tableau 12** présente les principales voies impliquées ainsi que les résultats de l'enrichissement. L'utilisation de MetExplore nous a permis de mettre en évidence l'importance centrale du métabolisme des acides aminés au carrefour des changements associés au SMet. Le métabolisme des lipides est également identifié comme étant modulé, bien que seules les familles lipidiques aient pu être mappées. D'autre part, le métabolisme des sucres (galactose) est également modulé avec un lien relativement direct au métabolisme énergétique et des acides aminés *via* le glutamate et la glutamine.

Nom de la voie métabolique	Nombre de réactions	Couverture du mapping sur la voie	Nombre de métabolites mappés	p-value corrigée (Bonferroni)	p-value corrigée (BH)
Exchange/demand reaction	49	24.59	15	3.7721E-14	3.77E-14
Urea cycle	66	8.86	7	0.006609093	0.003304547
Glycerophospholipid metabolism	74	7.79	6	0.043871034	0.014623678
Arginine and Proline Metabolism	41	8.62	5	0.080635061	0.020158765
Glycine, serine, alanine and threonine metabolism	37	7.81	5	0.125843112	0.025168622
Galactose metabolism	12	13.04	3	0.252387821	0.042064637
Valine, leucine, and isoleucine metabolism	42	7.55	4	0.412602802	0.052287991
Phenylalanine metabolism	10	10.71	3	0.443875005	0.052287991
Transport, extracellular	25	7.27	4	0.470591917	0.052287991

Tableau 12 : Voies métaboliques significativement différentes entre cas et moins.

III. Obtention d'une signature du SMet

Cette partie du travail a pour but d'identifier une signature du SMet, stable dans le temps, constituée d'un nombre restreint de métabolites pour qu'elle puisse être à terme, après validation, utilisable en clinique. Elle a nécessité la construction d'un workflow dédié utilisant W4M.

La démarche globale d'obtention d'une signature à partir de plusieurs jeux de données est présentée sur la **Figure 32.**

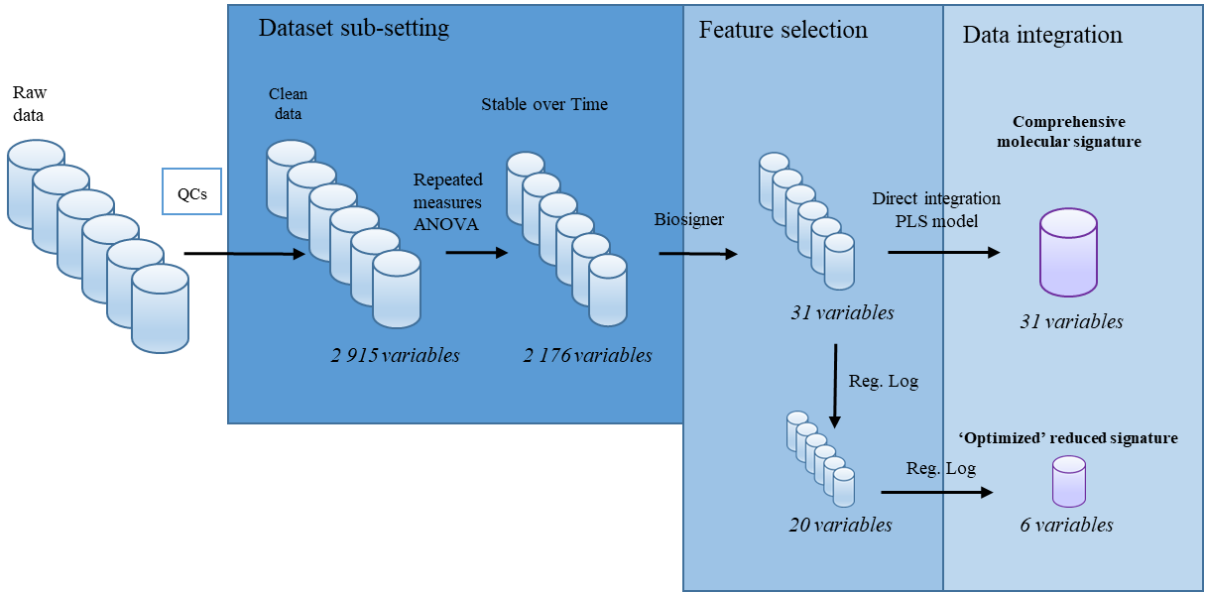


Figure 32 : Présentation de la démarche l'obtention d'une signature métabolique à partir de plusieurs jeux de données

Les données sont d'abord considérées individuellement (jeu de données par jeu de données) dans l'objectif d'effectuer l'étape de sélection de variables. Elles sont avant tout nettoyées, notamment de leur redondance analytique. Les métabolites stables dans le temps sont ensuite sélectionnés au regard de la question biologique, indépendamment de leur lien avec le statut cas/témoin. L'étape suivante est appelée **étape de sélection de variables** et précède l'intégration des jeux de données entre eux. Un petit nombre de métabolites est sélectionné sur la base de leur caractère prédictif par des méthodes multivariées à l'aide de l'outil BioSigner. En effet, ce type de méthodes est particulièrement adapté dans le cas des processus biologiques complexes et orchestrés. Il est alors possible de réaliser l'intégration des différents jeux de données à 2 niveaux. La première option consiste à le faire directement *via* la construction d'un modèle PLS-DA, méthode fréquemment utilisé dans le domaine de la métabolomique [101]. Cette approche conduit à l'obtention d'une signature relativement importante en nombre de métabolites avec des modèles souvent très performants. En revanche, elle implique de mesurer plusieurs dizaines de métabolites pour établir un diagnostic qui peut devenir difficile et coûteux. La seconde option consiste à sélectionner d'avantage les variables avant de réaliser l'intégration. Elle s'appuie sur l'utilisation de régressions logistiques, méthode souvent utilisée en épidémiologie. Elle s'effectue en 2 étapes : d'abord réaliser les régressions logistiques sur les différents jeux de données séparément pour réduire le nombre de variables à inclure dans le modèle unique final. En effet, il est recommandé d'inclure moins de 20 variables lors de la réalisation d'une régression logistique. Il est ensuite possible d'intégrer les 6 jeux de données dans un modèle unique à l'aide d'une régression logistique finale permettant d'obtenir une signature réduite en termes de nombre de métabolites. Le choix des métabolites à inclure fait souvent appel à une expertise biologique pour driver la pertinence de la sélection.

1. Sélection de variables

La sélection de variables représente l'une étape les plus clé dans l'exploration des données métabolomiques [238, 239]. Elle consiste à détecter la combinaison de métabolites la plus simple possible, mais aussi et surtout la plus pertinente pour décrire un état de santé. Elle peut être réalisée par une combinaison de différentes approches (machine learning, méthodes multidimensionnelles, visualisation de données, régressions, clustering, classification…) [179]. Malgré le fait que chacune présente des avantages et des inconvénients, les méthodes multivariées sont à privilégier dans le contexte de ce type de maladies de systèmes où des processus biologiques multiples sont en jeu. Elles

permettent de maximiser l'information qui sera extraite des données en considérant les variables dans leur ensemble et non de façon isolée.

L'étape de sélection des variables intervient à partir des jeux de données contenant les variables stables dans le temps, mises en évidence par le biais de l'ANOVA à mesure répétée. Les données d'entrée comprennent donc 2 176 variables, ce qui représente 75% des variables disponibles après prétraitement, réparties en six jeux de données : 646 en C18-Pos, 221 en C18-Neg, 498 en HILIC, 446 en Lipidomique, 338 en GC et 27 en RMN.

1.1. Filtration des corrélations fortes restantes

Au sein de ces jeux de données, la redondance analytique a été supprimée. Cependant, il existe encore une colinéarité intra-jeu importante, due à la redondance biologique, qu'il est important de limiter. En effet, au même titre que la redondance analytique, la redondance biologique peut avoir un impact sur les méthodes statistiques employées. De plus, la signature finale obtenue après les analyses ne doit pas contenir d'information répétitive (plusieurs métabolites intervenant en cascade dans la même voie par exemple) mais bien un ensemble de métabolites les plus indépendants possible les uns des autres et permettant la description d'un état de SMet. Pour ceci, nous avons à nouveau utilisé l'outil ACorF avec un paramétrage moins restrictif. Nous avons tenu compte uniquement d'un taux de corrélation > 0,8 pour regrouper entre eux, les ions potentiellement liés par une voie métabolique ou une quelconque interaction biochimique. Nous avons ensuite conservé comme représentant, l'ion le plus intense pour tous les jeux de données, exception faite de la RMN dans lequel le bucket le plus pur sert de représentant. Ce critère de sélection de l'ion représentatif permet de choisir l'ion le plus correctement mesurable au sein du groupe, critère indispensable pour établir une signature pertinente [87].

Nous avons ainsi réduit le jeu de données C18-Pos à 338 ions, le C18-Neg à 182 ions, l'Hilic à 431, la Lipidomique à 95 ions, la GC à 194 ions et la RMN à 23 buckets, soit une filtration de l'ordre de 42% des variables pour un total de 1 263 variables restantes.

1.2. Méthode de sélection

L'étape suivante consiste à sélectionner un sous-jeu de variables qui soit le plus restreint possible. Pour ce faire, nous avons utilisé l'outil Biosigner [240] dans sa version disponible sous W4M. Il permet de réduire au maximum la taille du jeu de données, en conservant les variables qui contribuent significativement à la bonne performance du modèle. Pour ce faire, il se base sur :

1. Des ré-échantillonnage successifs (bootstrap) : par défaut, l'outil réalise 50 ré-échantillonnages, c'est-à-dire qu'il sélectionne aléatoirement un jeu de données d'entrainement à partir des données fournies et répète l'opération autant de fois que paramétré.
2. Un classement des variables en fonction de leur importance dans le modèle : la classification est dépendante de la méthode avec par exemple, le « Variable Importance in Projection » (VIP) pour la PLS-DA [241], l'importance des variables pour RF [242] et le poids au carré pour SVM [243].
3. Une évaluation de la significativité par des tests de permutation : il s'agit de l'étape qui permet de restreindre le sous-ensemble de données. Elle évalue la significativité des variables une à une et élimine celles qui présentent un rang trop faible.
4. La construction d'un modèle final : les 3 premières étapes sont répétées jusqu'à ce que les variables candidates retenues soient toutes significatives ou qu'il ne reste aucune variable à tester. Le modèle final est ensuite obtenu à partir de toutes les observations faites sur chacune des variables retenues.

Biosigner se base sur 3 méthodes de classification binaires : PLS-DA, Random Forest et SVM qui permettent d'avoir des performances spécifiques dépendamment de la structure des données. Il offre une méthode rapide, automatique et fortement restrictive pour générer des signatures métaboliques. Les variables retenues par les modèles finaux sont classées en catégories dépendamment de leur importance dans le modèle de prédiction. Les variables S représentent le modèle final, ce sont celles qui sont retenues à chacune des itérations du bootstrap, puis l'importance des variables est décroissante des catégories A jusqu'à E.

Biosigner a été utilisé sur des jeux sans redondance, ce qui se révèle primordial pour optimiser au mieux l'analyse. En effet, les classifieurs tels que RF ou SVM sont très sensibles à cette dernière. Le fait de posséder, au sein d'un jeu de données, des informations très corrélées va perturber le processus de sélection et de classement des variables dans les modèles. Si plusieurs variables sont fortement corrélées et que toutes ont un fort pouvoir de prédiction, elles se retrouvent en compétition lors du

processus de sélection. Par exemple, dans la méthode RF, à chaque construction d'arbre, le bootstraping des échantillons influence le choix des variables. Lorsqu'il n'y a qu'une seule variable discriminante qui n'est pas en compétition avec d'autres corrélées, elle sera sélectionnée dans plusieurs itérations du bootstrap. *In fine*, cette dernière aura un score sélectif plus élevé que les variables corrélées entre elles qui ne sont pas sélectionnées de la même façon à chaque itération. Cependant son pouvoir discriminant n'est pas forcément plus élevé. Il est donc important de limiter la redondance et les corrélations au sein des jeux de données avant d'utiliser un outil comme Biosigner.

Nous avons appliqué l'algorithme de Biosigner à nos 6 jeux de données indépendamment, en répétant l'analyse 5 fois. Nous avons fixé le paramètre « seed » à 5 valeurs différentes pour permettre la répétabilité de l'analyse. Il s'agit d'un paramètre fixé normalement aléatoirement et permettant le bootstraping, le fait de lui assigner une valeur permet de pouvoir répéter l'analyse autant de fois que souhaité en obtenant toujours les mêmes résultats. Nous avons conservé toutes les variables S sélectionnées dans au moins l'un des 3 classifieurs et commun aux 5 itérations. Cette approche est considérée comme plus stringente qu'une PLS-DA unique, notamment grâce à sa capacité à combiner un nombre important de permutations de données au sein de 3 méthodes multivariées différentes.

1.3. Résultats de la sélection

La sélection a permis de construire des modèles robustes incluant entre 1 et 11 variables selon les jeux de données pour un nombre total de 31, tous jeux de données confondus, soit : 11 en C18-Pos, 5 en C18-Neg, 1 en HILIC, 4 en Lipidomique, 8 en GC et 2 en RMN.

Les méthodes SVM, RF et PLS utilisées par Biosigner se basent sur des structures de modèle différentes et donc ne sélectionnent pas nécessairement les mêmes variables. Nous avons donc choisi de conserver toutes les variables sélectionnées par une ou plusieurs méthodes afin de ne pas écarter un potentiel de prédiction qui serait propre à une méthode spécifique. La **Figure 33** illustre les variables sélectionnées par chacune des 3 méthodes au sein du jeu de données C18-pos lors de la première itération (seed à 19). On constate effectivement que 11 ions sont sélectionnés par SVM et ne le sont par les 2 autres méthodes.

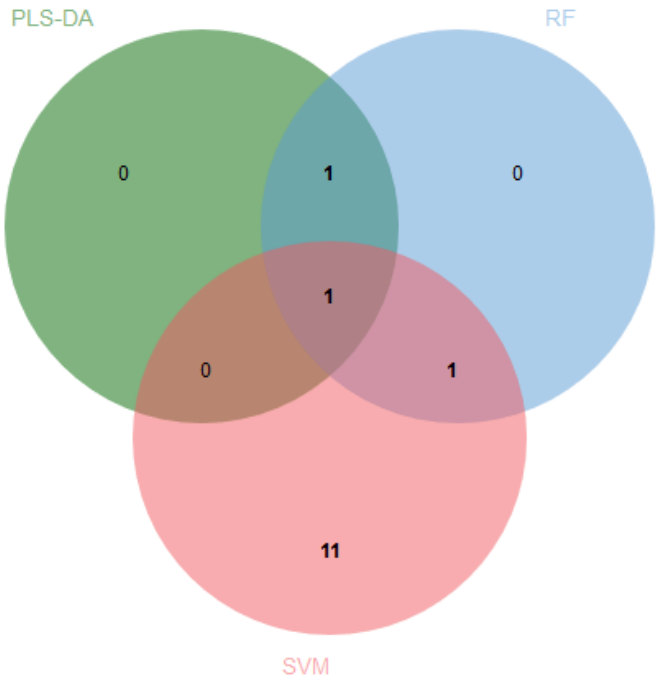


Figure 33 : Diagramme de Venn des variables sélectionnées par chacune des 3 méthodes. PLS-DA = Partial Least Squares – Discriminant Analysis ; RF = Random Forests ; SVM = Support Vector Machine.

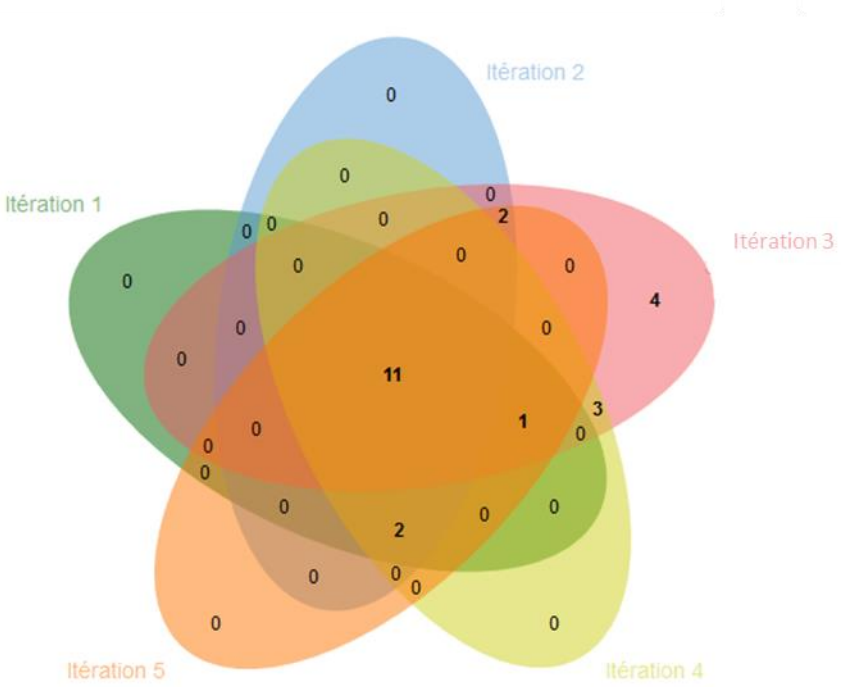


Figure 34 : Diagramme de Venn des variables sélectionnées en catégories S pour chacune des itérations de Biosigner.

Dans la mesure où la méthode de sélection par bootstrap se base sur une part d'aléatoire, les résultats observés entre les 5 itérations sont plus ou moins répétables en fonction des jeux de données. Ainsi, dans le cas du jeu de données C18-pos (**Figure 34**), on constate que la majorité des variables (11 ions) sont communément sélectionnées, quelque unes sont spécifiques à 1 à 4 itérations différentes. Notre choix de conserver uniquement les variables communes à toutes les itérations permet de favoriser une sélection de variables plus robuste.

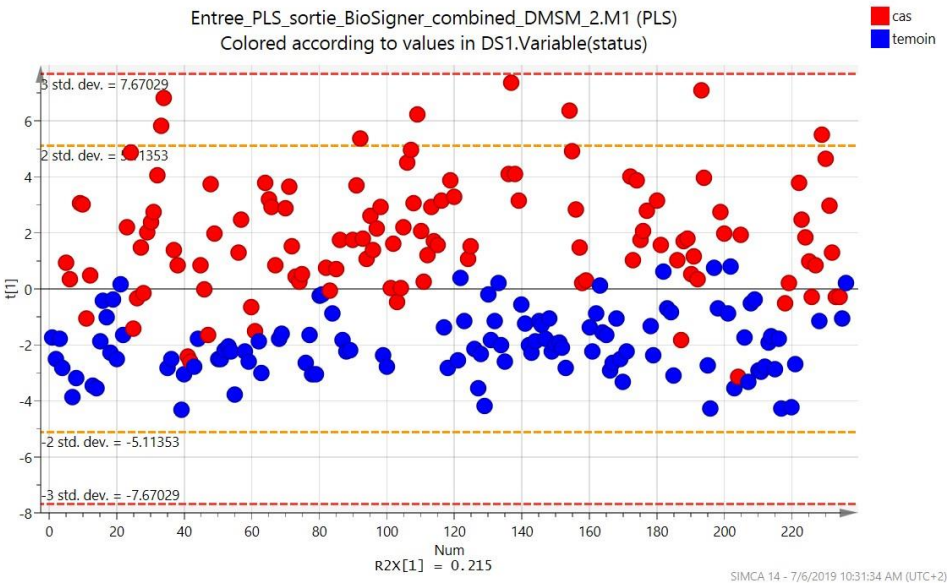
2. Intégration des jeux de données

2.1. Intégration directe

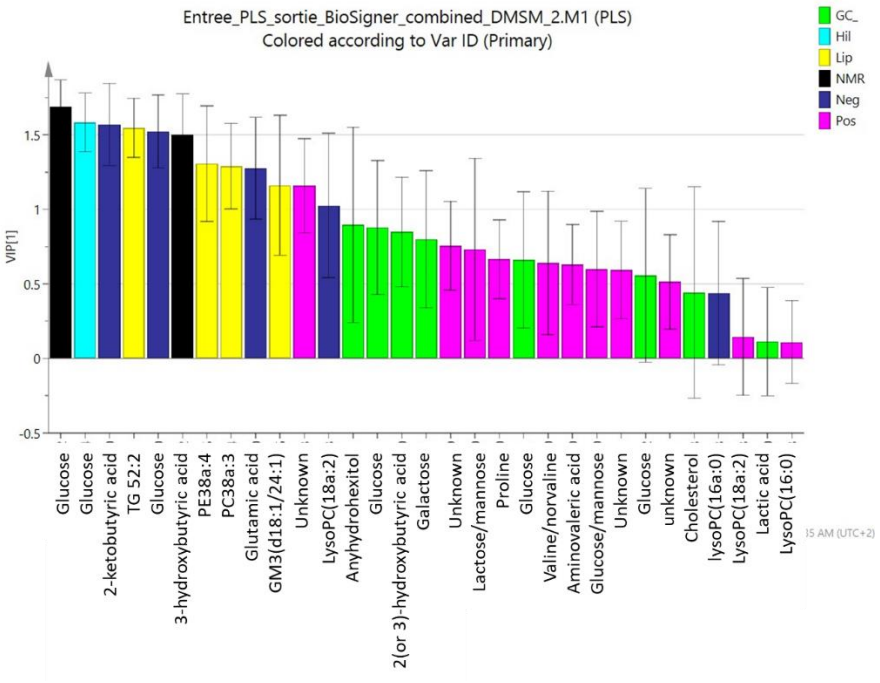
Comme indiqué dans la **Figure 32**, il est possible de réaliser l'intégration des jeux de données immédiatement après l'utilisation de Biosigner. Le moyen le plus couramment utilisé en métabolomique pour réaliser cette dernière et évaluer la capacité de prédiction des variables consiste à réaliser une PLS-DA [101]. Cette dernière a été faite grâce à l'outil « Multivariate » présent dans W4M, toujours dans le souci de produire un workflow entièrement réutilisable et transposable à d'autres projets.

La signature moléculaire qui en découle contient donc les 31 variables. Le modèle validé permet une bonne discrimination du statut cas/témoins (R^2Y=0,58 ; Q^2Cum=0,56 ; taux d'erreur de 11% ; AUC = 0,95 ; CI:[0,92-0,98]) comme illustré dans la **Figure 35**.

Cette intégration directe présente le désavantage d'aboutir à un modèle au sein duquel peut subsister de la redondance d'information entre les différents jeux de données comme illustré par la **Figure 35B**. A titre d'exemple, le glucose apparait 5 fois dans la signature en étant mesuré par 4 méthodes d'analyse différentes.



A : Score plot



B : VIP

	Témoins prédits	Cas prédits
Témoins observés	109	8
Cas observés	18	101

C : Table de confusion

Figure 35 : Résultats de la PLS-DA sur les variables intégrées après Biosigner.

Toutefois, si l'objectif d'un tel modèle est de permettre la détection de la pathologie en clinique, sa taille reste trop importante et il peut être intéressant de proposer plutôt une signature ne contenant que quelques molécules.

2.2. Intégration après sélection

Dans le but de réduire au maximum cette signature nous avons donc poursuivi le workflow pour ajouter des étapes de sélection et intégrer les jeux de données non pas en sortie de Biosigner, mais plus tardivement.

a. Régressions logistiques sur jeu de données individuel

Nous avons tout d'abord fait le choix de réaliser des régressions logistiques sur chacun des jeux de données séparément. Cette méthode de sélection de variables est couramment utilisée dans le domaine de l'épidémiologie, contrairement à la métabolomique où la PLS-DA est majoritairement choisie comme évoqué précédemment. Cependant, l'une des limites de cette méthode réside dans le fait qu'elle n'est adaptée qu'à un nombre restreint de variables explicatives. Ainsi, pour une taille de population d'une centaine d'individus, considérer 31 variables dans un même modèle n'est pas recommandé, justifiant le choix d'analyser séparément les 6 jeux de données.

La régression logistique a pour but de réduire le nombre de variables d'entrée (ici entre 1 et 11 selon le jeu de données) à un plus petit nombre tout en garantissant la construction d'un modèle de prédiction du statut cas/témoins qui soit de qualité.

Elles ont été réalisées à l'aide d'un script en langage R développé par la PFEM visant à réaliser des régressions logistiques après vérification de la multi-colinéarité entre les variables. Les différents modèles générés ont permis de réduire à 25 le nombre d'ions à considérés : de 11 on est passé à 7 variables en C18-Pos, 1 variable a été écartée en C18-Neg ainsi qu'en GC. Le nombre de variables reste inchangé pour les autres jeux de données (Hilic, Lipidomique et RMN).

b. Filtration des corrélations entres jeux de données

L'étape suivante consiste à intégrer ensemble les 6 jeux de données pour pouvoir réaliser par la suite une régression logistique. Toutefois, au-delà de la complémentarité analytique entre les 6 jeux de données, il existe une certaine redondance au regard de quelques métabolites qui sont mesurés par plusieurs méthodes. Pour cette raison, et parce qu'à l'instar des méthodes de wrappers, la régression logistique est sensible aux corrélations, nous avons cherché à détecter cette redondance d'information, non plus au sein d'un jeu de données, mais entre les 6 méthodes d'analyse. Pour ce faire, nous avons une nouvelle fois utilisé l'outil ACorF suite à la construction d'une matrice de corrélation toujours avec l'outil « Beetwen Table Correlation », en prenant des corrélations de Spearman afin de prendre en charge l'hétérogénéité des matrices des différentes méthodes, et une correction BH des p-value. Le paramétrage d'ACorF prend en compte des corrélations > 0,8, sans prise en charge des informations de RT et de différences de masse et choisit comme représentant des groupes la variable la plus intense. En effet, le but de notre analyse étant d'arriver *in fine* à une signature facilement mesurable, nous avons préféré privilégier les variables de forte intensité.

Cette filtration a réduit l'ensemble de 25 à 20 variables (7 en C18-Pos, 4 en C18-Neg, 1 en Hilic, 4 en Lipidomique, 7 en GC et 2 en RMN), permettant ainsi d'avoir un nombre optimal de variables pour réaliser une régression logistique intégrant les données des 6 méthodes.

Malgré une réduction déjà efficace du nombre de variables, il peut persister entre elles de la multi-colinéarité pouvant parfois impacter négativement des modèles de régression multiple. Pour cette raison, le script effectuant la régression logistique prend en charge une étape de détection et filtration de cette dernière si nécessaire.

c. Régression logistique commune et signature finale

Comme évoqué dans la section précédente, la dernière étape pour obtenir une signature de taille réduite est la réalisation d'une régression logistique après intégration des 6 méthodes. Cette dernière a une nouvelle fois, été faite à l'aide du script R de la plateforme. Elle permet d'obtenir une sélection de 6 variables formant un modèle de discrimination de bonne qualité : taux d'erreur de 0,155 ; AUC 0,93 ; CI:[0,9;0,96] (**Figure 36**). Les métabolites du modèle final ont été annotés, la

signature réduite contient donc : l'acide kétobutyrique, le glutamate, la proline, l'acide hydroxybutyrique, le 1,5-anhydrohexytol et la valine.

En comparaison avec la PLS réalisée sur les 31 métabolites, la qualité de prédiction du modèle est légèrement plus faible : taux d'erreur de 15,5% contre 11% avec la PLS-DA et AUC de 0,93 contre 0,95. Toutefois, l'augmentation du taux d'erreur est liée à une augmentation du taux de faux négatif (témoins prédit cas à tort). Dans le cas d'une méthode de screening/diagnostic, il est préférable de voir apparaitre ce type d'erreur plus que de faux positifs (cas prédits témoins à tort), une méthode de confirmation de diagnostic pouvant être mise en place. De plus, l'utilisation de la signature de 6 métabolites reste plus pertinente du fait de sa taille réduite.

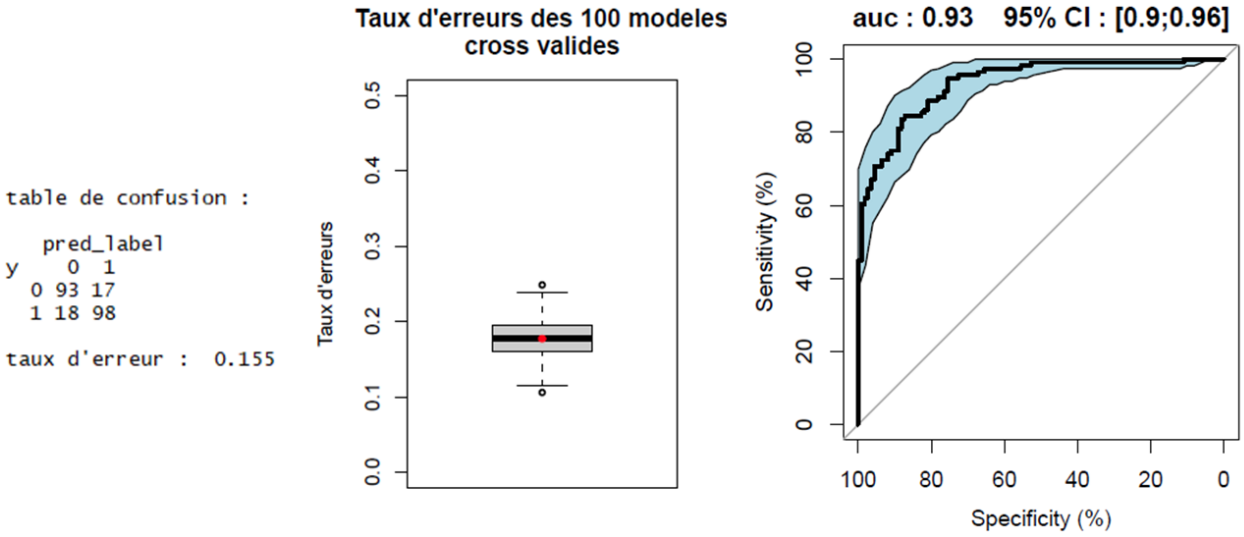


Figure 36 : Résultat de la régression logistique permettant l'obtention de la signature finale. 0 = témoin ; 1 = cas

IV. Conclusions associées à la construction d'un workflow de traitement et sélection de données

Dans un premier temps, la démarche d'analyse et le workflow mis en place allant jusqu'à la construction de réseaux métaboliques plus ou moins complexes et à l'enrichissement des voies associées, vont permettre une meilleure interprétation et compréhension du SMet, en donnant une vision intégrée des processus biologiques concertés.

Dans un second temps, le workflow a permis la sélection d'une signature de taille réduite caractérisant le SMet. A partir de 13 902 variables après prétraitement, on arrive à en isoler 2 915 sans redondance. Parmi ces variables, 476 sont significatives du statut cas/témoins et stables dans le temps,

2 176 sont, elles, uniquement stables (sans prise en compte du statut). A partir de ces dernières, on arrive à isoler 31 variables pour construire un modèle PLS, puis 6 pour obtenir une signature réduite du SMet. Les résultats obtenus à chaque étape du workflow permettant l'obtention de la signature réduite sont présentés dans le **Figure 37**.

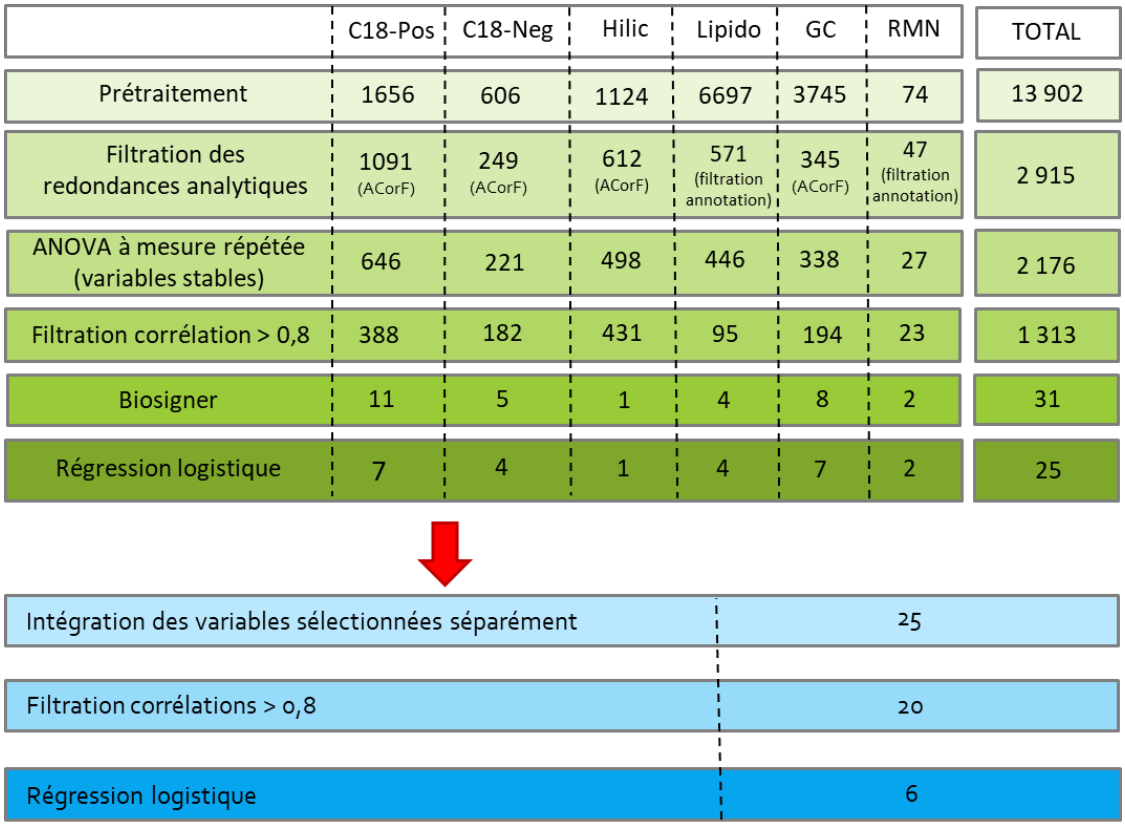


Figure 37 : Récapitulatif du nombre de variables sélectionnées à chacune des étapes du workflow pour l'obtention de la signature réduite.

Le workflow d'analyse général est présenté sur la **Figure 38.** Il permet le passage d'une analyse métabolique multiplateformes, à une signature de taille réduite pour discriminer 2 états, dans notre cas, SMet ou non SMet, mais également l'interprétation des variations observées d'un point de vue métabolique entre les 2 groupes. Ce workflow fait intervenir plusieurs outils disponibles sur différentes plateformes d'analyse. Ceci engendre plusieurs difficultés d'un point de vue bio-informatique, que ce soit dans l'enchaînement même des outils, ou l'utilisation de certains d'entre eux :

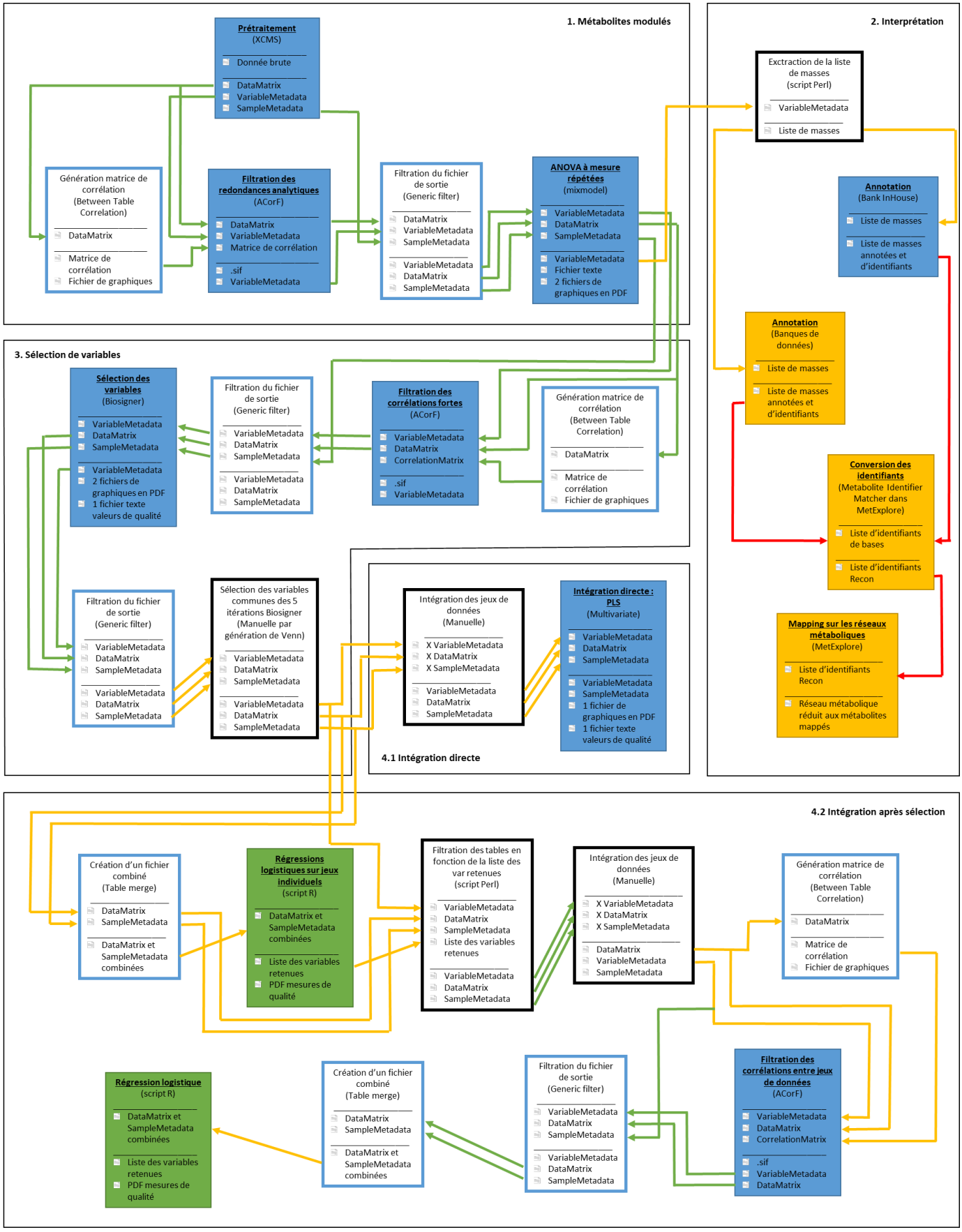


Figure 38 : Présentation générale du workflow d'analyse.
Les cases en bleu correspondent aux étapes réalisées à partir de W4M ; en jaune, celles à l'aide d'une application web ; en vert, celles à partir de scripts R et en orange, celle à partir de logiciels indépendants. Les flèches vertes indiquent qu'aucune difficulté particulière n'est rencontrée ; les oranges, qu'il existe des problématiques de format de données et/ou de mise en place de filtration ; les rouges, qu'une étape est critique et engendre le plus souvent une perte importante d'information.

Les principales difficultés bio-informatiques rencontrées sont les suivantes :

- **Outils de gestion de format et de contenu.** Malgré tout, il reste certaines limitations avant de pouvoir réaliser le workflow de sélection (hors étape d'interprétation biologique) entièrement dans Galaxy… Comme évoqué dans le Chapitre 2, lors de l'intégration des jeux de données il est nécessaire de fusionner les 6 jeux tout en s'assurant de disposer du même nombre d'échantillons. Il peut donc être nécessaire de supprimer certains échantillons dans certains jeux de données. La gestion de cette étape pourrait se faire au sein de W4M, mais il faudrait pour cela qu'un nouvel outil de fusion de jeux de données soit mis à disposition en permettant l'ajout de plusieurs trios DataMatrix/SampleMetadate/VariableMetadata issus des différentes analyses pour produire un nouveau trio unique. Son absence nous oblige donc à nous affranchir de W4M pour réaliser cette étape indépendamment.
- **Passage d'un outil à l'autre lorsque les 2 ne sont pas disponibles au même endroit (W4M)**. Le passage de W4M (où est effectuée la plupart des étapes de prétraitement et de traitement) à de nouveaux outils comme MetExplore où le script de régression logistique entraîne la nécessité de transformer les formats des données pour les rendre compatibles. Par exemple, transposer des colonnes, fusionner 2 fichiers en 1, supprimer des échantillons ou des sujets…
- **Ajouter ou convertir de l'information.** Outre la transformation de format à proprement parler, qu'il est possible de gérer par le biais de petits outils/scripts de transformation, la problématique du rajout ou de la conversion d'information peut se poser. En effet, comme évoqué lors de son utilisation, des outils comme MetExplore nécessitent des annotations les plus précises possibles, qu'il est parfois difficile d'obtenir, mais aussi et surtout de convertir les identifiants obtenus en identifiants spécifiques des réseaux comme Recon. Malgré les outils existants, cette conversion est souvent difficile et engendre une grande perte d'information.
- **Des outils à intégrer au sein de l'instance Galaxy W4M pour permettre la construction d'un workflow de sélection complet au sein de l'instance.** Certains des outils utilisés dans notre workflow, comme le script de régression logistique, ne sont pas intégrés à l'instance W4M. Par conséquent, il est nécessaire de sortir de l'instance puis d'y revenir ce qui représente une forme de faiblesse. Galaxyfier ce(s) script(s) permettrait donc d'offrir la possibilité d'avoir un outil complet en ligne sur la même instance, et le rendrait beaucoup plus facilement réutilisable.

V. Publication associée

Une publication est en cours de préparation et sera soumise prochainement au journal « Molecular Systems Biology ».

Comte B, **Monnerie S**, Brandolini M, Canlet C, Castelli C, Colsch B, Fenaille F, Joly C, Lenuzza N, Lyan B, Martin JF, Migné C, Morais JA, Pétéra M, Thévenot E, Junot C, Gaudreau P, Pujos-Guillot E. Integrative multiplatform metabolomics for metabolic syndrome exploration.

REFERENCES DU CHAPITRE 3

1. Morrow DA, de Lemos JA. Benchmarks for the assessment of novel cardiovascular biomarkers. Circulation. 2007;115(8):949-52.
2. Goecks J, Nekrutenko A, Taylor J, Galaxy T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11(8):R86.
3. Giacomoni F, Le Corguille G, Monsoor M, Landi M, Pericard P, Petera M, et al. Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. Bioinformatics. 2015;31(9):1493-5.
4. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics. 2007;3(3):211-21.
5. Cottret L, Wildridge D, Vinson F, Barrett MP, Charles H, Sagot MF, et al. MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. Nucleic Acids Res. 2010;38(Web Server issue):W132-7.
6. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27-30.
7. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 2016;44(D1):D471-80.
8. Thiele I, Swainston N, Fleming RM, Hoppe A, Sahoo S, Aurich MK, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol. 2013;31(5):419-25.
9. Schlegel DR, Ruttenberg A, P.L. E. Ontologies in Metabolomics. Metabolomics. 2015;5:e137.
10. Gromski PS, Muhamadali H, Ellis DI, Xu Y, Correa E, Turner ML, et al. A tutorial review: Metabolomics and partial least squares-discriminant analysis--a marriage of convenience or a shotgun wedding. Anal Chim Acta. 2015;879:10-23.
11. Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3. 2003.
12. Liu H, Motoda H. Feature selection for knowledge discovery and data mining: The Springer International Series in Engineering and Computer Science; 1998.
13. Grissa D, Petera M, Brandolini M, Napoli A, Comte B, Pujos-Guillot E. Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data. Front Mol Biosci. 2016;3:30.
14. Rinaudo P, Boudah S, Junot C, Thevenot EA. biosigner: A New Method for the Discovery of Significant Molecular Signatures from Omics Data. Front Mol Biosci. 2016;3:26.
15. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometrics and intelligent laboratory systems. 2001;58:109-30.
16. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.
17. Guyon I, Weston J, Barnhill SMD, Vapnik V. Gene Selection for Cancer Classification usingSupport Vector Machines. Machine Learning. 2002;46:389-422.

CHAPITRE 4 : ETUDE DES SOUS-PHENOTYPES

Comme évoqué dans le Chapitre 2 lors de la caractérisation des sujets, nous avons mis en évidence l'existence de 5 sous-groupes cliniques définis par des combinaisons de critères du SMet, chez les sujets étudiés. Ceci soulève donc la question de l'existence de potentiels sous-phénotypes du SMet qui pourraient se traduire par des profils métaboliques discriminants. L'objectif de ce travail est d'évaluer la capacité de la métabolomique à caractériser et modéliser le spectre phénotypique du SMet. Cet objectif fait partie des grands enjeux de la médicine de précision pour une prise en charge à terme, plus personnalisée des populations.

D'un point de vue bio-analytique, ce travail consiste à mettre en place une stratégie de traitement des données. Nous avons choisi de nous focaliser sur T1 pour appréhender la complexité des sous-phénotypes dans un contexte simplifié tout d'abord (sans prise en compte de l'aspect longitudinal).

I. Sous-phénotypes basés sur les données cliniques

1. Sous-phénotypes cliniques observés lors de l'analyse des caractéristiques des sujets

La première approche que nous avons mise en place est celle présenté dans le Chapitre 2 qui consiste à étudier les combinaisons de critères cliniques possibles, lorsque l'on considère ces derniers de façon classique, comme des valeurs binaires (présence/absence). Elle a été menée sur un sous-ensemble de 106 individus pour lesquels aucune valeur manquante n'est existante concernant la présence/absence des 5 critères à T1. Ceci a permis la mise en évidence de 5 sous-groupes majeurs regroupant un total de 69 individus, pour rappel :

- **Groupe A** : les cas présentant les 5 critères du SMet (11 individus).
- **Groupe B** : les cas présentant les 4 critères suivants : Hyperglycémie, Hypertension, Tour de taille élevé, hyper TG (18 individus).
- **Groupe C** : les témoins avec hypertension (14 individus).
- **Groupe D** : les cas présentant 3 critères : Hyperglycémie, Hypertension, Tour de taille élevé (8 individus).

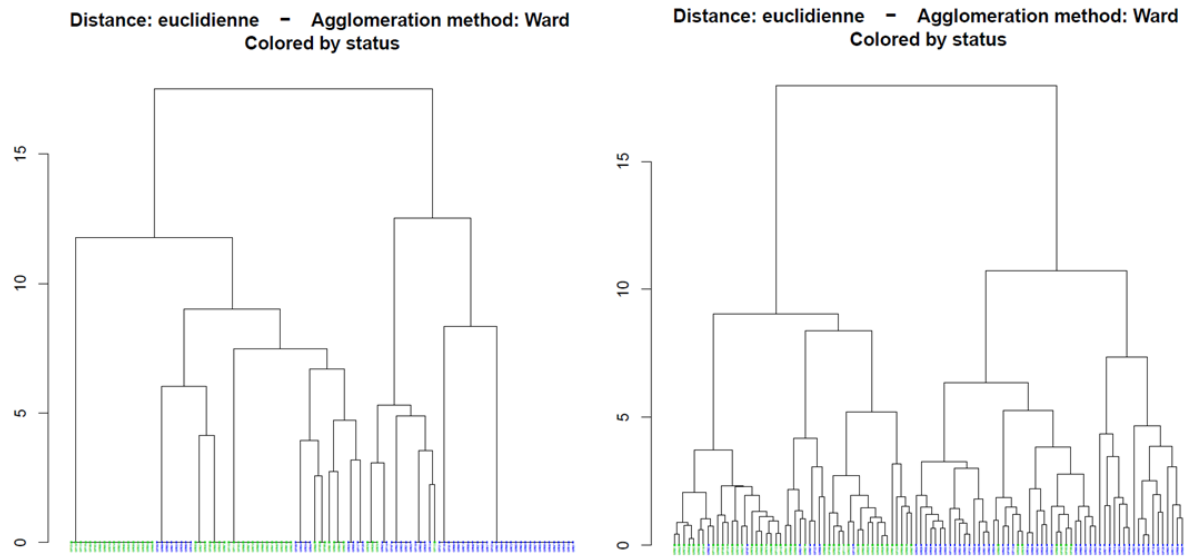
- **Groupe E** : les témoins exempts de critères du SMet (18 individus).

Les individus restants représentent des sous-groupes de taille trop petites pour être analyser.

2. Comparaison entre approches sur critères binaires et sur critères quantitatifs

Nous avons ensuite cherché à étudier les groupes d'individus formés par Classification Ascendante Hiérarchique (CAH) en se basant soit sur ces mêmes valeurs binaires, soit sur les valeurs quantitatives mesurées pour chaque critère. Les CAH ont été réalisées à l'aide d'un script R développé au sein de la plateforme d'exploration du métabolisme de Theix, paramétré pour effectuer la CAH après centrage réduction des données, avec utilisation des distances euclidiennes et *via* la méthode d'agrégation dite Ward. Cette étape nous a permis d'étudier l'effet des seuils établis par la définition du SMet sur la répartition des individus en clusters.

La **Figure 39** compare les résultats obtenus par les 2 approches. Ils mettent clairement en évidence des clusters très nets dans le cas des valeurs binaires, où les cas et les témoins sont bien séparés. En revanche, lorsque l'on supprime la notion de seuil, on constate que les clusters formés sont un peu plus hétérogènes et que les distances entre individus augmentent. La prise en compte des valeurs binaires semble cacher une certaine complexité qui apparait lorsque l'on considère les valeurs quantitatives formant ainsi des groupes d'individus plus complexes.



A : Dendrogramme généré par la CAH sur les données de critères binaires N=106

B : Dendrogramme généré par la CAH sur les données de critères quantitatifs N=106

Figure 39 : Dendrogrammes issus des CAH réalisées sur les critères du SMet binairisés ou sous forme de valeurs quantitatives à T1.
Coloration en fonction du statut (bleu = témoins; vert = cas).

Dans un second temps, nous avons souhaité comparer les résultats obtenus sur les critères quantitatifs, aux groupes formés par simple observation de répartition des critères binaires. Pour se faire, nous avons fait une CAH sur le même nombre d'individus que ceux inclus dans les 5 groupes A-E à savoir 69 sujets. La **Figure 40** illustre la sélection à 5 clusters effectuée à partir du dendogramme obtenu.

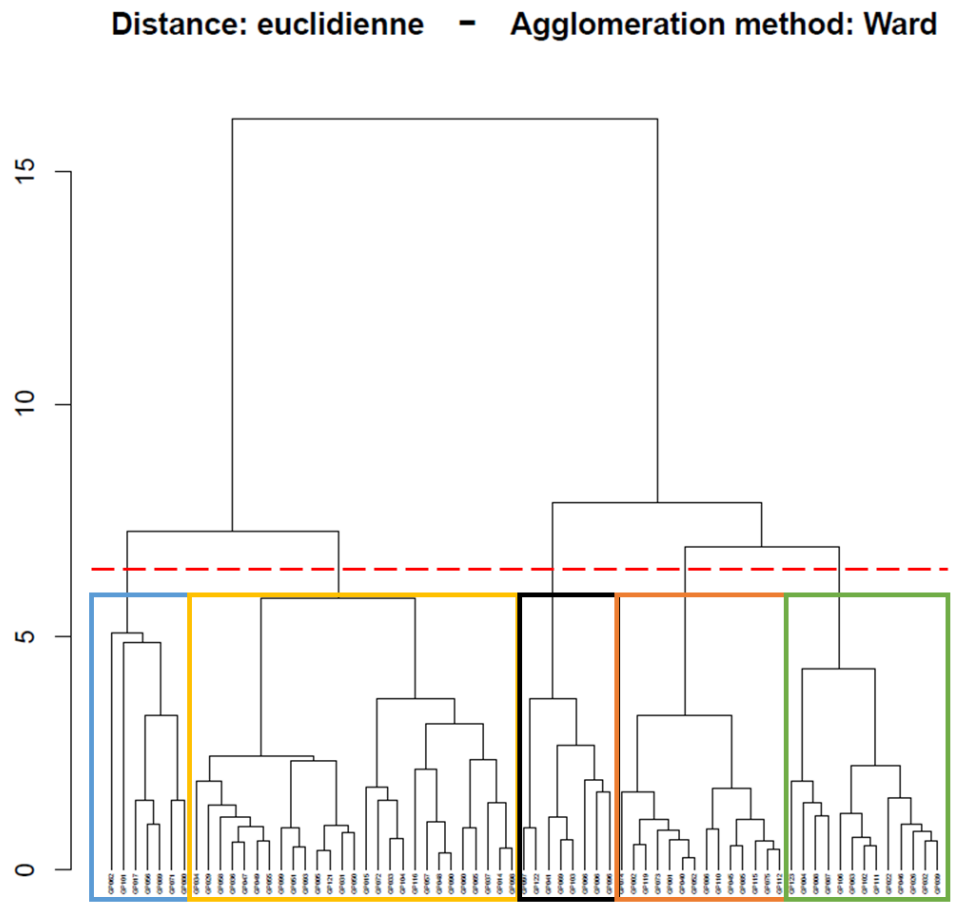


Figure 40 : Dendrogramme généré par la CAH sur les données de critères quantitatifs (nombre de sujets = 69).
Bleu = cluster 1 ; Jaune = cluster 2 ; Noir = cluster 3 ; Orange = cluster 4 ; Vert = cluster 5.

La comparaison entre les groupes de critères binaires et les 5 clusters obtenus à partir des critères quantitatifs est présentée sur la **Figure 41**. Elle confirme le fait que considérer les critères en mesures quantitatives n'aboutit pas au même type de classification que lorsque l'on considère un seuil de positivité ou négativité pour ces derniers. En effet, si certaines similitudes sont observables, la majorité des clusters de la CAH ne correspondent pas aux groupes de la CAH sur données binaires qui sont souvent constitués d'individus issus de plusieurs clusters.

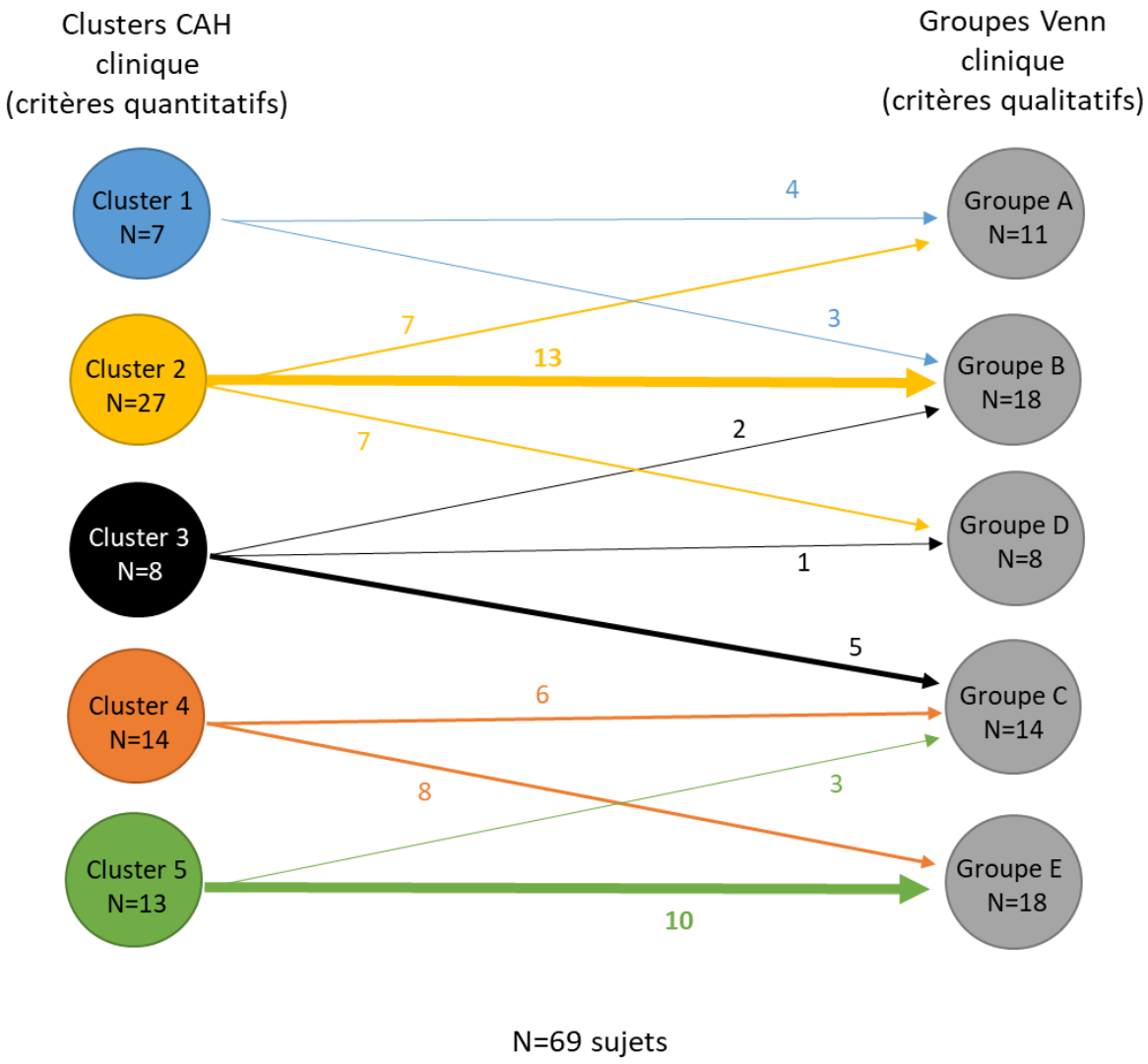
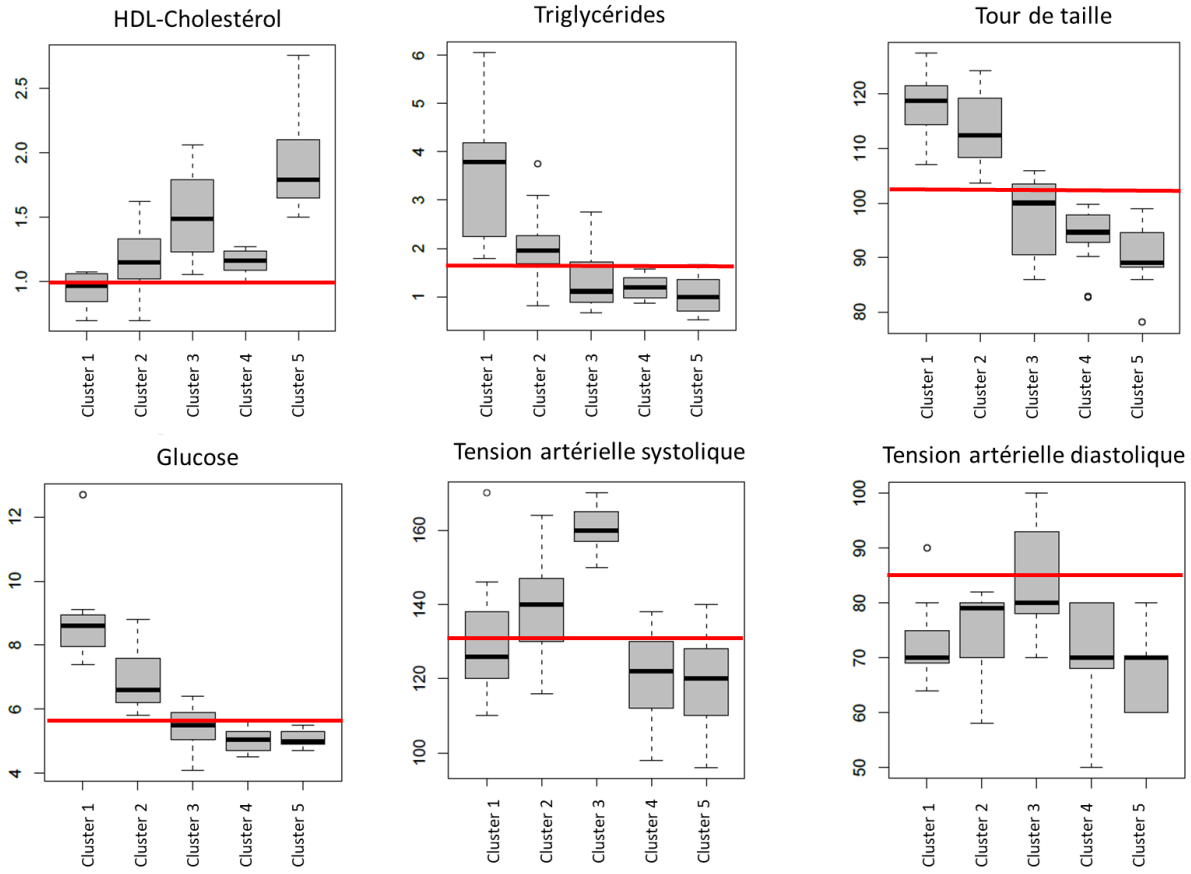


Figure 41 : Comparaison entre les groupes de critères binaires et les clusters sur critères quantitatifs.

Les caractéristiques cliniques des individus présents dans ces 5 clusters sont présentées dans la **Figure 42**. La description des individus de ces différents clusters confirme la différence existante entre l'approche sur critères binaires et l'approche sur critères quantitatifs.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
HDL-Cholestérol	+	-	--	-	--
Triglycérides	++	+	-	-	-
Tour de taille	++	++	-	-	--
Glucose	++	+	-	-	-
Tension artérielle systolique	-	+	++	-	-
Tension artérielle diastolique	--	-	o	--	--

A : Tableau de positivité aux critères du SMet.
+ = légèrement au-dessus du seuil de positivité ; ++ = très clairement au-dessus du seuil de positivité ; - = légèrement au-dessous du seuil de négativité ; -- = très clairement au-dessous du seuil de négativité ; o = à cheval avec le seuil.



B : Box plot ; Ligne rouge = seuil du critère fixé par la définition du SMet.

Figure 42 : Caractéristiques des sujets présents dans les différents clusters en fonction des 5 critères du SMet pris en compte.

3. ACP puis Classification Hiérarchique sur Composantes Principales (HCPC)

Dans un troisième temps, nous avons cherché à observer la classification des individus à partir des critères quantitatifs mais en se basant simplement sur les principales variabilités observées. Ainsi, on peut s'affranchir de l'effet de seuil des variables qualitatives tout en limitant l'impact des faibles variabilités individuelles difficiles à prendre en charge sur seulement une centaine de sujets.

Pour ce faire, nous avons réalisé une ACP sur les 5 critères quantitatifs dans le but d'isoler les composantes principales (ici au nombre de 2). Nous avons ensuite réalisé une classification hiérarchique sur composantes principales (HCPC). Cette analyse a été réalisée à l'aide du package R FactoMineR. Ce dernier a suggéré un nombre optimal de clusters à 3. Le dendrogramme issu de l'HCPC est présenté en **Figure 43**.

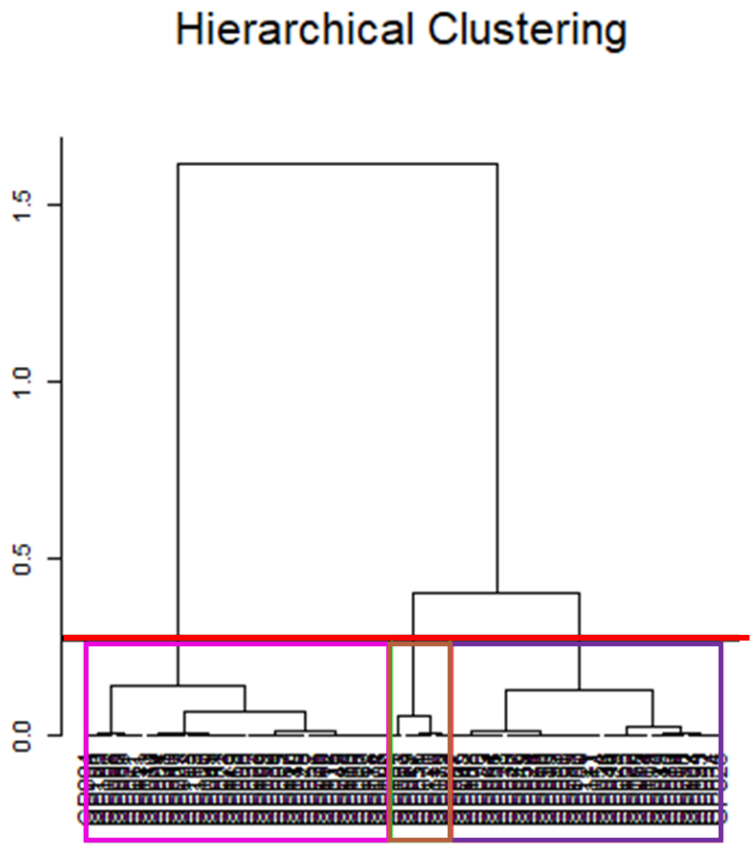


Figure 43 : Dendrogramme issu de l'HCPC sur les critères quantitatifs (nombre de sujets = 106). Violet = cluster A ; Rose = cluster B ; Marron = cluster C.

Les données utilisées pour générer ces nouveaux clusters étant également les critères quantitatifs, nous avons cherché à comparer les résultats obtenus avec ceux de la CAH simple. Pour ce faire nous avons cette-fois ci découpé notre dendrogramme en 3 clusters et non plus 5 pour les rendre plus facilement comparable à l'HCPC (**Figure 44**).

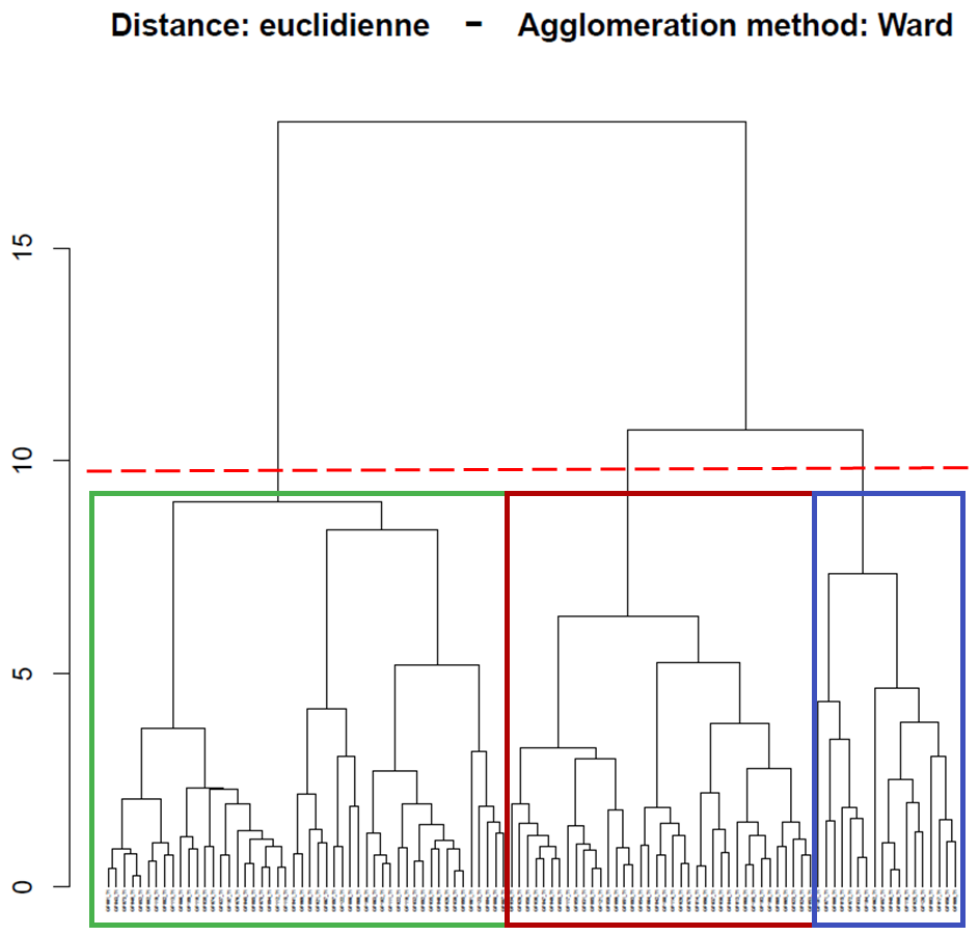


Figure 44 : Dendrogramme généré par la CAH sur les données de critères quantitatifs (nombre de sujets = 106).
Vert = cluster 1 ; Rouge = cluster 2 ; Bleu = cluster 3.

L'illustration de cette comparaison est présentée sur la **Figure 45**. Les clusters formés par HCPC sont sensiblement les mêmes que ce formés par CAH directe. Ceci soulève donc l'idée de considérer 3 sous-phénotypes majeurs du syndrome pour effectuer notre modélisation à partir des données métabolomiques.

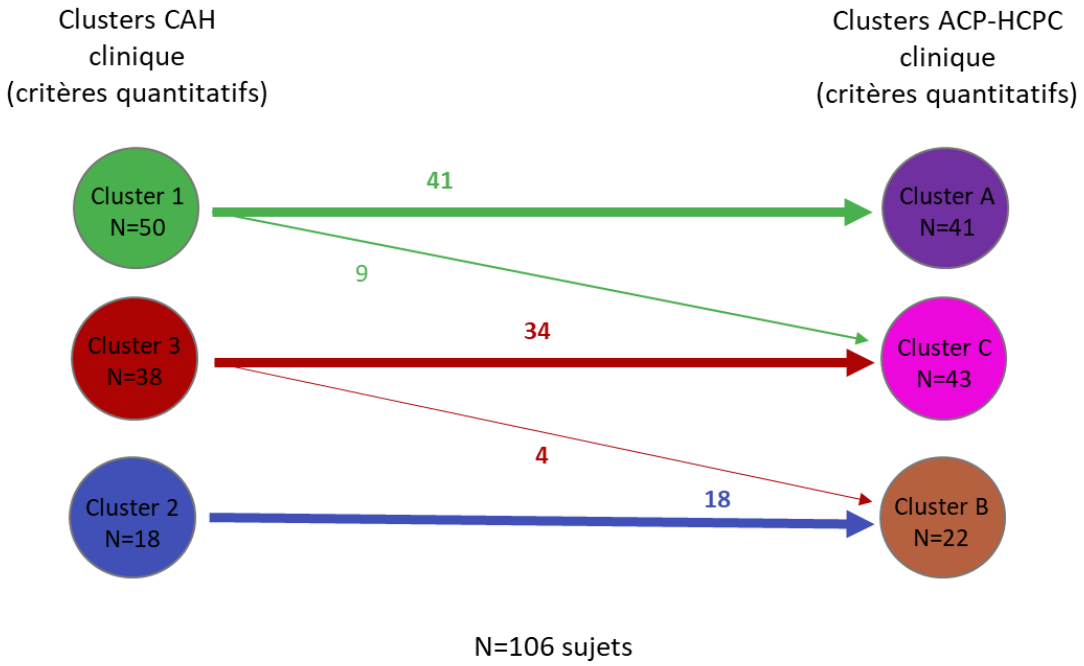
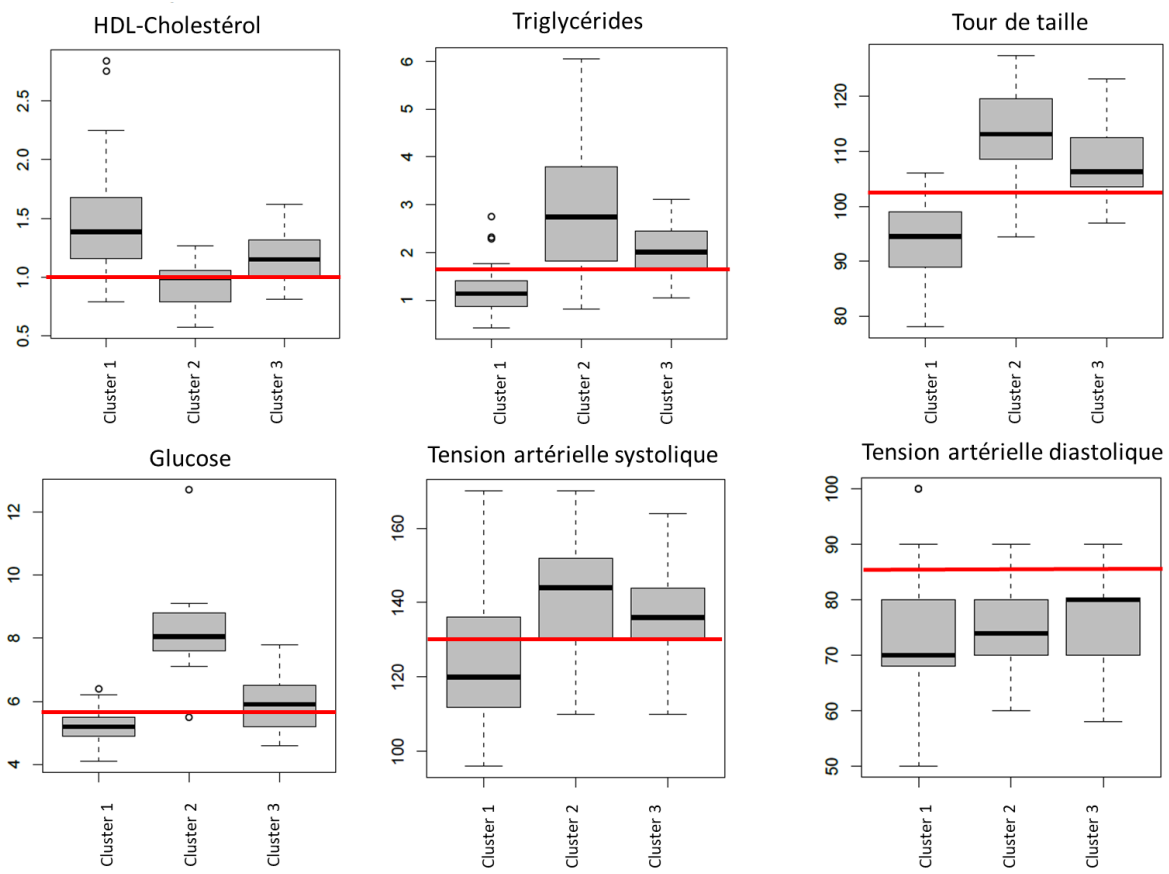


Figure 45 : Comparaison entre les clusters formés par HCPC et les clusters formés par CAH directe sur les critères quantitatifs.

Les caractéristiques cliniques des individus présents dans les 3 clusters de la CAH sont présentées sous forme de box-plot dans la **Figure 46**.

	Cluster 1	Cluster 2	Cluster 3
HDL-Cholestérol	-	o	-
Triglycérides	-	++	+
Tour de taille	-	++	+
Glucose	-	++	+
Tension artérielle systolique	-	+	+
Tension artérielle diastolique	-	-	-

A : Tableau de positivité aux critères du SMet.
+ = légèrement au-dessus du seuil de positivité ; ++ = très clairement au-dessus du seuil de positivité ; - = légèrement au-dessous du seuil de négativité ; -- = très clairement au-dessous du seuil de négativité ; o = à cheval avec le seuil.



B : Box plot ; Ligne rouge = seuil du critère fixé par la définition du SMet.

Figure 46 : Caractéristique des sujets présents dans les différents clusters

II. Prédiction des sous-phénotypes à partir des données métabolomiques

Suite à l'étude des sous-phénotypes basés sur les critères cliniques du SMet, nous nous sommes demandé s'il était possible d'observer une classification identique à partir des données métabolomiques.

1. Réduction du nombre de variables d'intérêt

L'utilisation des données métabolomiques pour étudier de potentiel sous-phénotypes implique avant tout de sélectionner un nombre de variables restreintes et pertinentes/portant

l'information. Nous avons choisi de considérer les variables issues des 6 jeux de données, après l'étape de suppression des corrélations analytiques évoquée dans le Chapitre 3. Le nombre de variable s'élève à 2 915. Nous avons décidé de baser la sélection de variables sur leurs corrélations aux 5 critères du SMet.

Des matrices de corrélation entre les critères quantitatifs du SMet, et les variables métabolomiques de chaque jeu de données ont été réalisées à l'aide de l'outil « Beetwen Table Correlation » (corrélation de Spearman et avec une correction de test multiple de type BH). Nous avons ensuite sélectionné les variables pour lesquelles une corrélation > 0,5 ou < -0,5 avec au moins l'un des critères était observée.

La considération des cas et des témoins de façon distincte ou non entraine une sélection de variables différentes. En effet, les processus biologiques sous-jacents peuvent différer dans les deux phénotypes, modifiant les liens entre variables. Dans le but de n'écarter aucune variable d'intérêt, nous avons choisi de considérer les 230 variables uniques sélectionnées par l'union des deux approches. Afin d'avoir un aperçu des liens existant entre ces variables, une nouvelle matrice de corrélation (corrélations de Spearman avec correction BH), est présentée **Figure 47**. On remarque que près de 80% des variables ont au moins une corrélation > 0,8 ou < - 0,8 ; les corrélations > 0,8 ou < -0,8 représentant seulement 6% de la matrice.

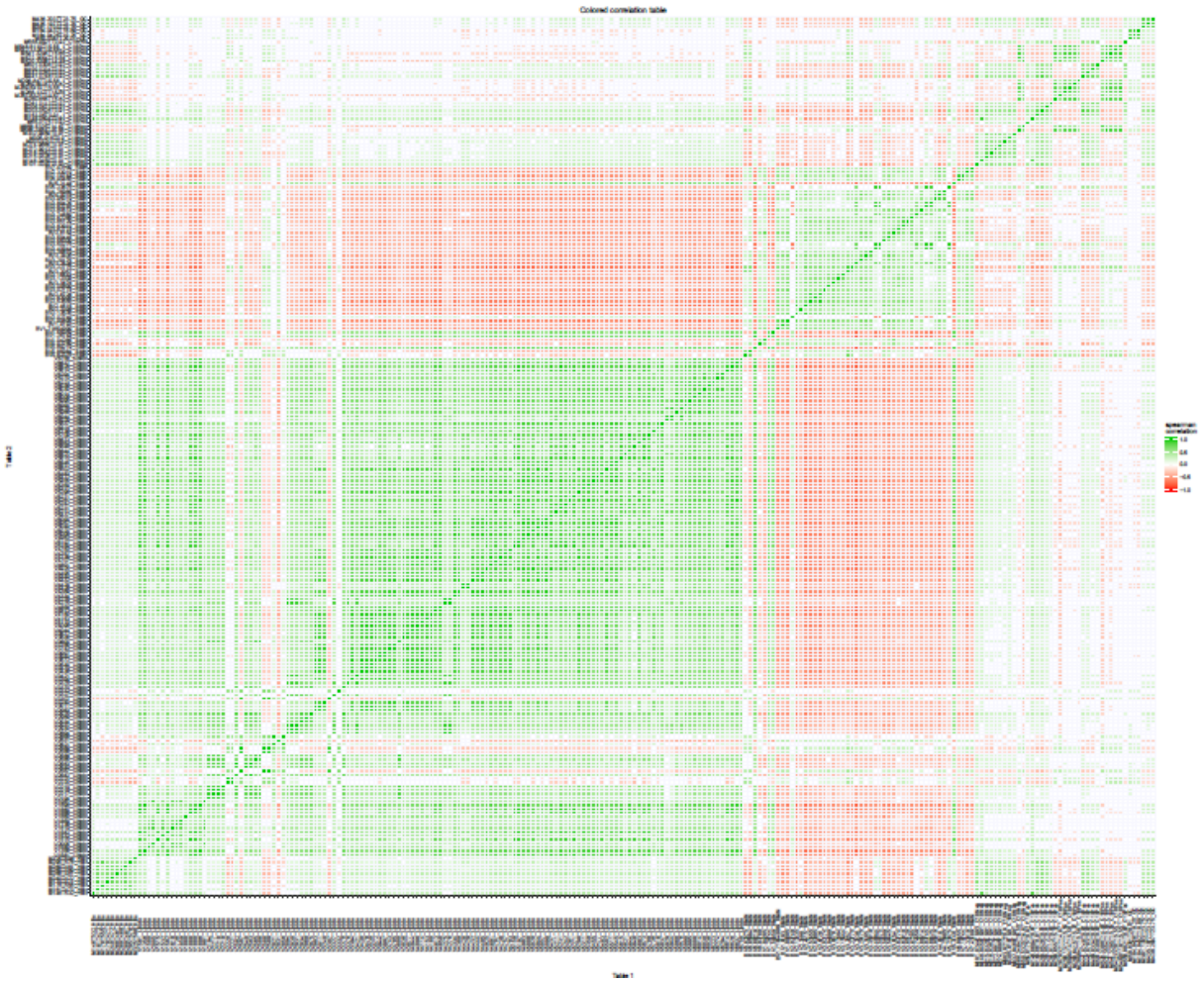


Figure 47 : Représentation graphique de la matrice de corrélation des 230 variables métabolomiques entre elles.

2. Construction de modèles de prédiction des clusters issus de la CAH

Avant de proposer une reclassification du SMet entièrement à partir de la métabolomique, nous avons étudié la possibilité de réaliser un modèle prédictif des 3 clusters issus de la dernière CAH sur les données quantitatives, à partir des données métabolomiques sélectionnées. Pour cela, nous nous sommes basés sur la construction d'un modèle PLS. Cette dernière analyse, constituée de 2 composantes, montre des indicateurs de qualité plutôt médiocres (R^2 = 0,538, Q^2 = 0,374 et test de permutations peu concluant, lié au manque de séparation du cluster 3). Toutefois, le score plot sur les 2 composantes laisse entrevoir une discrimination possible entre les 3 clusters 2 à 2 sur la base des données métabolomiques (**Figure 48**).

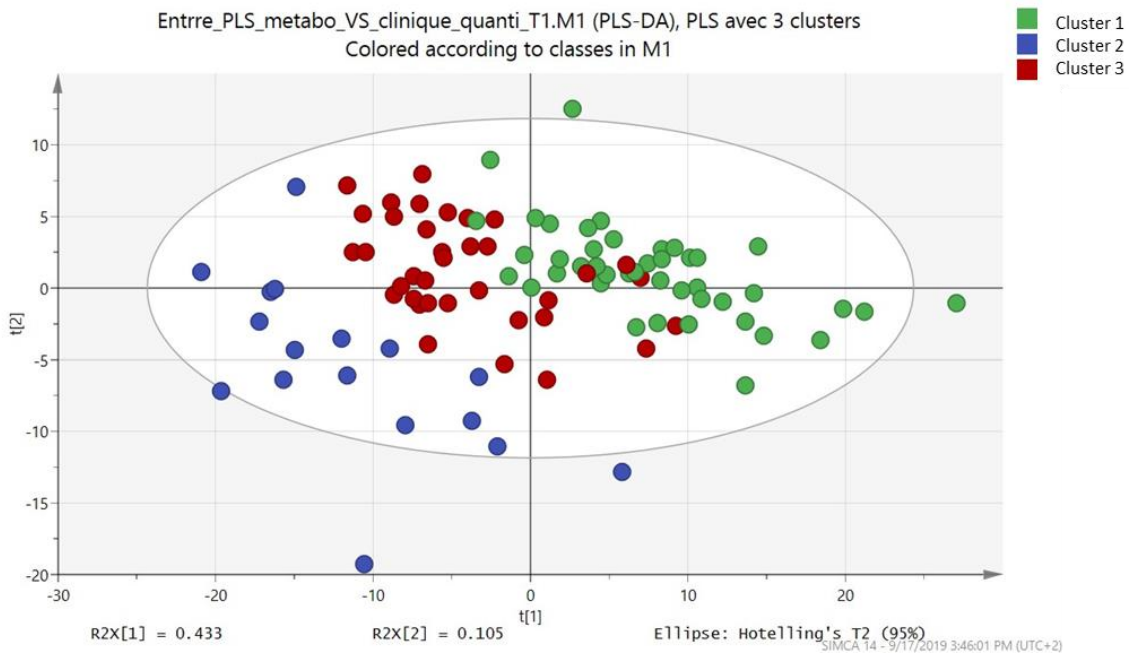
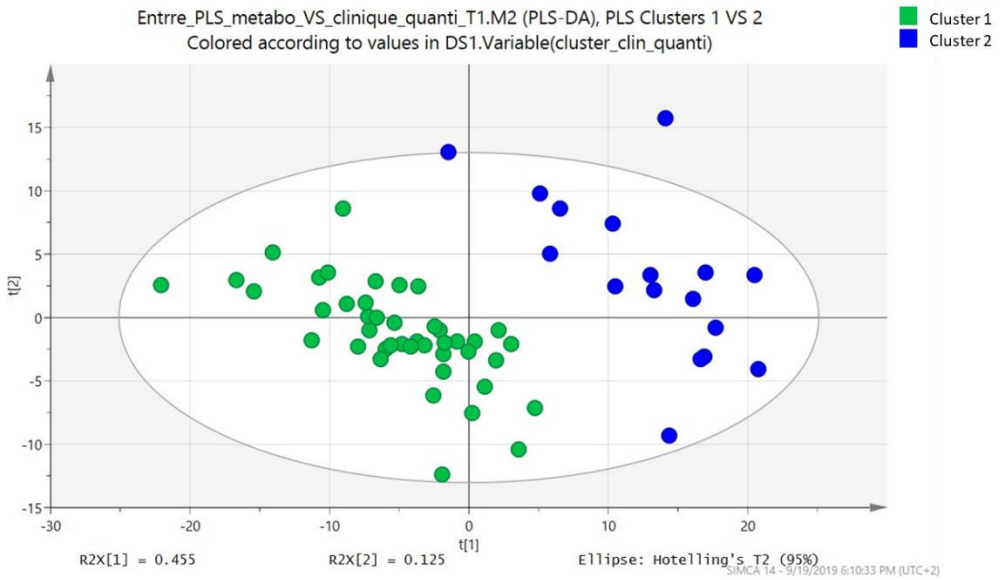
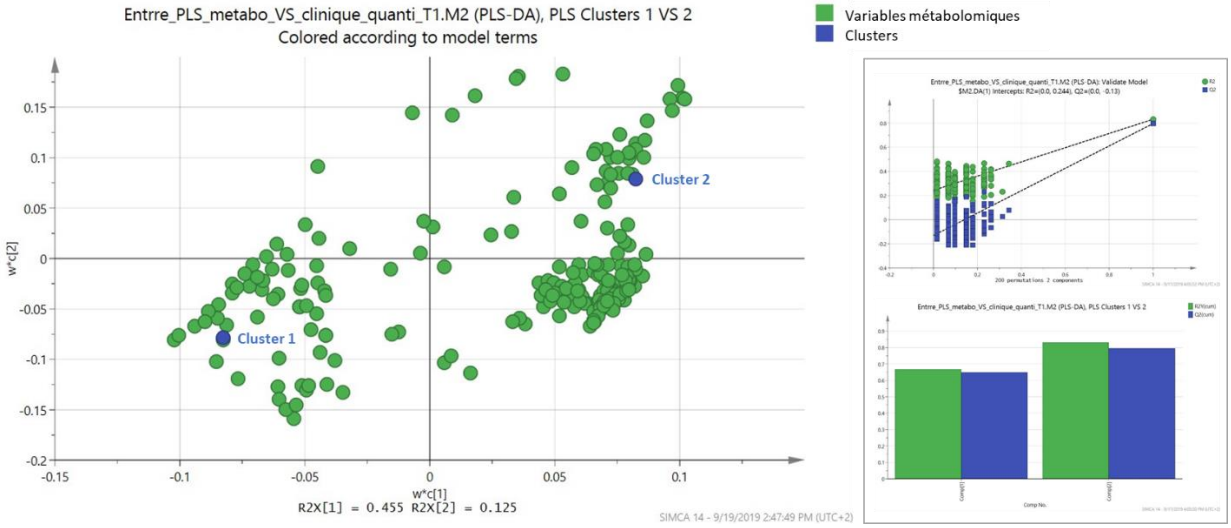


Figure 48 : Score plot de la PLS réalisée afin de prédire les 3 clusters de la CAH sur les critères quantitatifs.

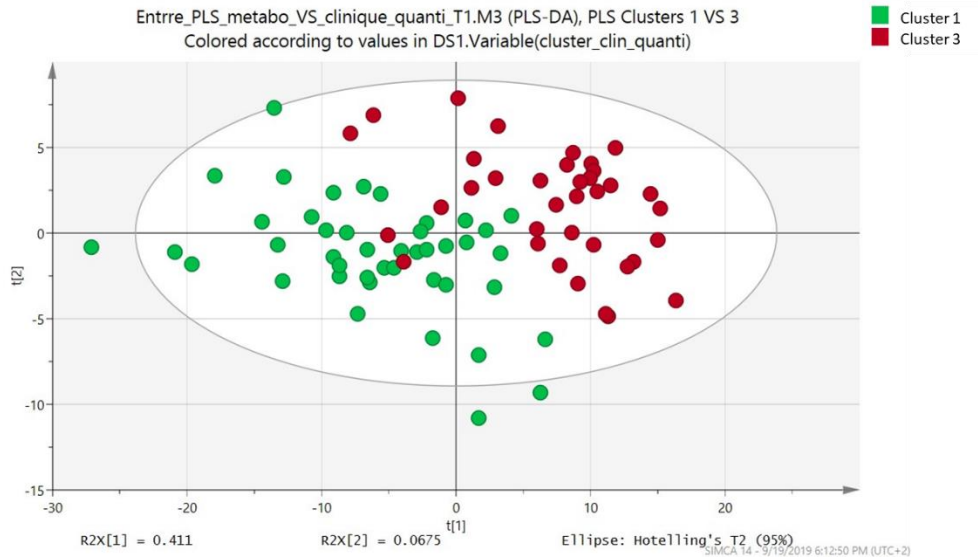
Au vu de ces résultats encourageants, nous avons réalisé des PLS de prédiction des clusters 2 à 2. Les résultats des 3 PLS sont présentés dans la **Figure 49**. Ces dernières montrent de très bons résultats de prédiction, notamment la PLS sur les clusters 1 et 2. Les variables métabolomiques ayant un score VIP > 1,25 (projection de l'importance de la variable) pour chacune des PLS ont été sélectionnées et sont présentées en **Annexe 2**. L'annotation de ces variables permettra de proposer une interprétation des différences entre les clusters qui complétera les informations cliniques connus sur les phénotypes des individus les composants.

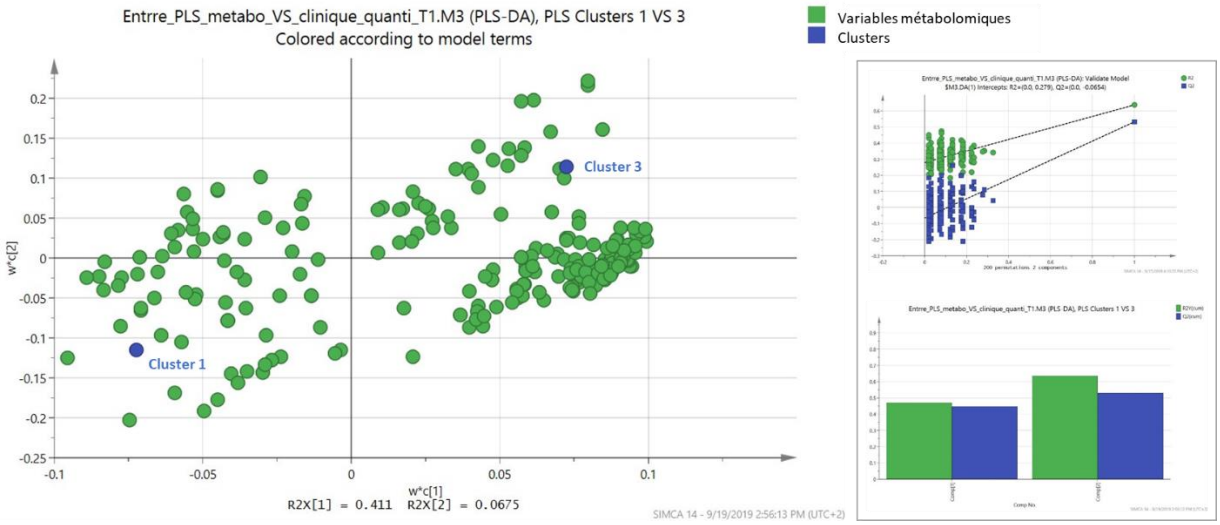




	Cluster 1 prédits	Cluster 2 prédits
Cluster 1 observés	43	0
Cluster 2 observés	1	16

49A : PLS cluster 1 et cluster 2.
R^2 = 0,58 ; Q^2 = 0,796 ; taux d'erreurs = 1,7 %

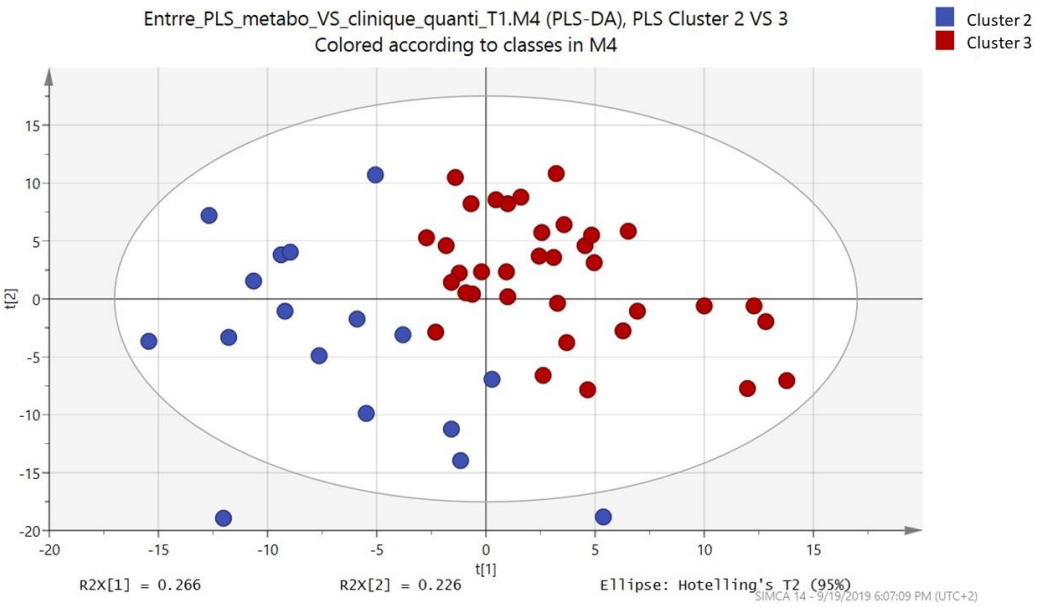


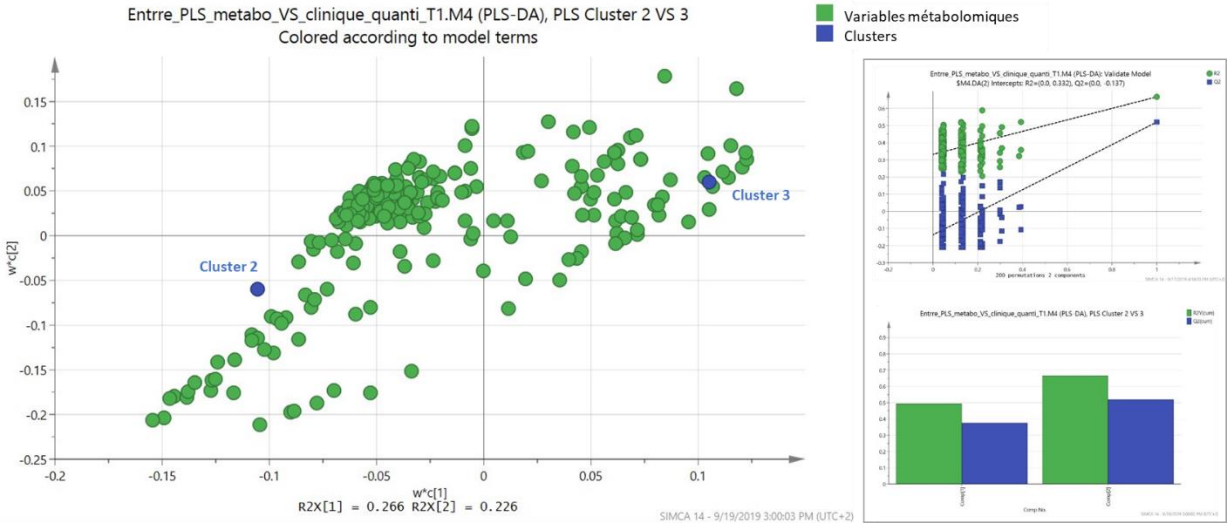


	Cluster 1 prédits	Cluster 3 prédits
Cluster 1 observés	39	4
Cluster 3 observés	2	34

49B : PLS cluster 1 et cluster 3

R^2 = 0,478 ; Q^2 = 0,531 ; taux d'erreurs = 7,6 %





	Cluster 2 prédits	Cluster 3 prédits
Cluster 2 observés	16	1
Cluster 3 observés	1	35

49C : PLS cluster 2 et cluster 3
R^2 = 0,491 ; Q^2 = 0,52 ; taux d'erreurs = 3,8 %

Figure 49 : Résultats des 3 PLS discriminantes des clusters 2 à 2.
Score scatter plots, loading plots, test de permutations, résumés des composantes et tables de confusion.

3. Méthode de clustering basée sur les réseaux

De nombreuses méthodes peuvent être envisagées pour arriver à répondre à notre objectif de reclassification moléculaire du SMet. En complément du travail présenté précédemment, nous avons fait le choix de nous focaliser sur une méthode qui nous a semblé intéressante et particulièrement originale : le clustering basé sur la construction de réseaux. Cette méthode de clustering a notamment été développé dans le domaine de la génomique par l'équipe du professeur Jan Baumbach de l'Université Technique de Munich, en vue de la reclassification de pathologies complexes. Je me suis donc rendu 5 semaines au sein de son laboratoire pour échanger sur ces méthodes et voir quelles pistes d'application seraient possibles.

KeyPathwayMiner (KPM) [244] est à l'origine un outil d'enrichissement pour des données multi-omics permettant des représentations sous forme de réseaux. Il permet d'enrichir des données de génomique ou de protéomique en construisant des réseaux. Différents réseaux sont générés pour chaque individu et peuvent ainsi être comparés. KPM propose alors un réseau d'union qui représente les principales voies rencontrées chez une majorité de sujets. L'étude des sous-réseaux proposée avant la création du réseau d'union représente donc un moyen d'étudier les groupes de sujets présentant des réseaux très proches et donc une forme de clustering. La **Figure 50** illustre le principe avec des réseaux formés légèrement différents permettant de mettre en évidence 2 clusters d'individus différents en bleu et en vert.

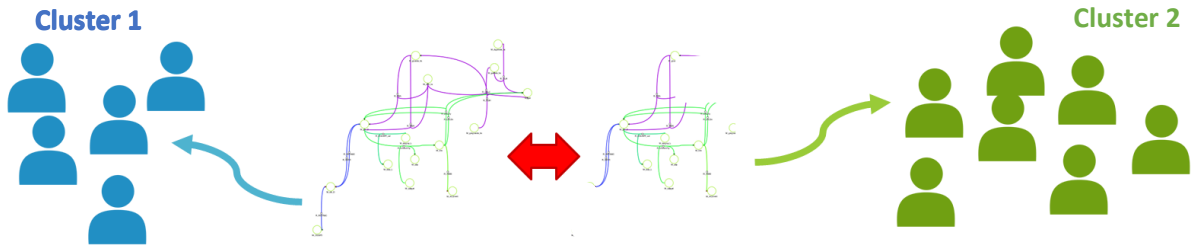


Figure 50 : Principe simplifié de la clusterisation par le biais des réseaux.

Malheureusement, l'outil a été développé pour des réseaux de gènes et de protéines et l'utilisation de données de métabolomiques n'est à ce jour pas possible. Nous avons toutefois cherché à tester malgré tout cette méthode à partir de nos données. Pour se faire, nous avons avant annoté un maximum des 230 variables sélectionnées, toujours en utilisant les méthodes d'interrogation de bases de données évoquées dans les chapitres précédents. Nous avons une nouvelle fois fait face à la

problématique des identifiants rencontrée précédemment lors de la construction des réseaux métaboliques. Les lipides constituent à nouveau la catégorie de métabolites la plus complexe à étudier. Nous avons ensuite envisagé de relier les noms de métabolites à des réactions (par le biais de réseaux comme Recon) afin d'être capable de les rattacher à des protéines pour permettre leur représentation au sein de KPM. Toutefois, il est délicat de réduire un métabolite à une seule et unique réaction, la véracité biologique de cette approximation étant faible. Il est donc très difficile de réaliser ce « changement d'échelle » du métabolite vers la protéine. Pour ces raisons, nous avons pris la décision de mettre en pause l'exploration de cette approche qui nécessite du temps et la mise en place d'une collaboration à plus long terme avec le laboratoire du professeur Baumbach. L'objectif d'une telle collaboration serait de transposer la démarche de clustering par réseaux métaboliques à partir de nouveaux outils ou en transformant des outils déjà existant.

4. Conclusions et perspectives

Les différentes approches de clustering réalisées ont démontré : l'importante différence entre le fait de considérer les critères avec des valeurs binaires dépendantes d'un seuil et les critères sous forme quantitative. Le fait de considérer de nouveaux sous-phénotypes du SMet semble offrir une alternative plus proche de la réalité métabolique.

L'exploration des sous-phénotypes n'en est qu'à ces débuts et nous avons souhaité montrer à travers ce chapitre les méthodes applicables et le potentiel que représente la métabolomique pour la reclassification des phénotypes.

Concernant l'exploration du clustering basé sur les réseaux, nous espérons qu'une future collaboration permettra l'exploration plus poussée d'une éventuelle application au domaine de la métabolomique.

REFERENCES DU CHAPITRE 4

1. Alcaraz N, Kücük H, Weil J, Wipat A, Baumbach J. KeyPathwayMiner: Detecting Case-Specific Biological Pathways Using Expression Data. Internet Mathematics. 2011;7(4):299-313.

CONCLUSIONS ET PERSPECTIVES

I. Conclusions

Mon projet de thèse visait à identifier les déterminants du syndrome métabolique, à de multiples niveaux, ainsi que leurs interactions et ce en utilisant une approche systémique [138]. Ce type d'approches combinent à la fois des données de type omique, et des données de type épidémiologiques observationnelles. Elles ont pour objectif de construire des modèles de prédiction du phénotype étudié à partir de multiples jeux de données par le biais de méthodes statistiques/mathématiques. La sélection de variables est un élément clé pour parvenir à construire ces modèles et proposer une stratification des individus étudiés. Différentes hypothèses peuvent alors émerger suite à cette stratification réalisée et être en lien plus ou moins étroit avec les données épidémiologiques, nutritionnelles... En particulier des sous-phénotypes peuvent être identifiés, apportant un nouveau regard sur la diversité phénotypique de la pathologie/de son installation et permettent la génération de nouvelles hypothèses, qui pourront être étudiées par la suite (**Figure 51**).

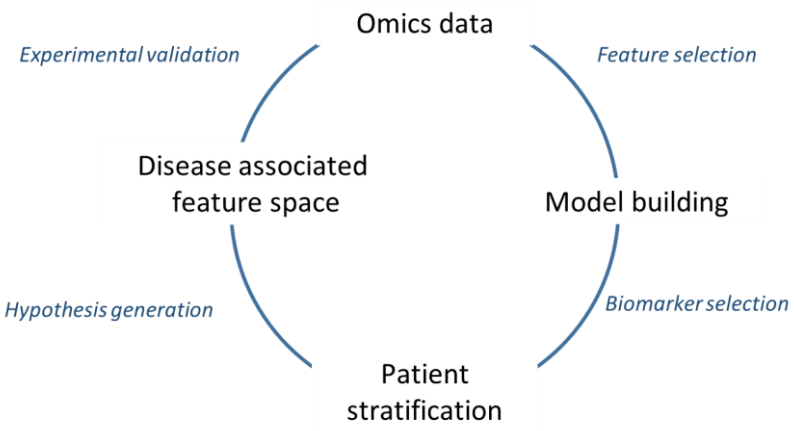


Figure 51 : Illustration d'une démarche d'analyse systémique

Dans le cadre de mon projet, les données utilisées sont des données de métabolomique non ciblées, issues d'une approche multi-plateformes, pouvant être associées à des données épidémiologiques relatives à la nutrition, la santé mentale, l'activité physique… de sujets âgés. L'objectif à long terme de ce type de projet est d'utiliser la métabolomique pour reclassifier les individus atteints de pathologies métaboliques chroniques comme le SMet. Ces 3 années de recherche se sont découpées en plusieurs objectifs qui ont pu être atteints avec plus ou moins de difficultés.

Dans un premier temps la réalisation de la revue systématique et la création de la banque de données des biomarqueurs du SMet a constitué le point de départ de mon travail de recherche. Plus

de 700 métabolites ont pu être référencés et associés aux informations descriptives des études constituant une base solide de connaissance pour la suite du projet. Toutefois, cette démarche a mis en évidence l'une des principales difficultés rencontrée dans le traitement et la gestion des données en métabolomique : le manque d'uniformité dans la nomenclature des métabolites. Avoir été confrontée à cette limitation dès le début de ma thèse m'a permis de garder en tête ce point et de tenter de le contourner au maximum en collectant notamment des informations telles que les identifiants de bases de données afin de recouper certains synonymes sous un seul identifiant.

En parallèle de la conduite de la revue systématique, la question de la gestion des données descriptives (métadonnées) des individus ainsi que des données métabolomiques s'est rapidement posée. La plateforme réalisant des développements bio-informatiques pour une meilleure gestion des métadonnées d'études complexes, les données de ce projet ont été utilisées pour élaborer le cahier des charges d'un futur système d'information (OmicM). Ce dernier n'a pas pu être déployé dans sa totalité au cours de ma thèse, notamment à cause de la complexité d'architecture engendrée par la diversité des données et de leurs relations. Toutefois, avoir participé à l'élaboration du cahier des charges m'a permis de mettre en place différentes routines de gestion et curation de fichiers, mais également de mieux appréhender la taille et la complexité des données générées par ce type de projets.

L'analyse des données métabolomiques, dans l'objectif de construction d'un workflow complet et reproductif, visant à extraire une signature du SMet chez la personne âgée a constitué l'un des volets centraux de ma thèse. Elle a nécessité de nombreuses étapes aux limitations variées.

- Tout d'abord, concernant la filtration des données, j'ai proposé le développement d'un outil bio-informatique (ACorF) visant à limiter les redondances analytiques. Ce développement m'a directement confronté aux problématiques analytiques présentes au sein des données métabolomiques (fragmentation des métabolites engendrant une redondance, corrélations fortes entre les métabolites issus de mêmes voies biologiques...) et à leur complexité due aux technologies multiples employées. Il a fallu adapter l'algorithme pour maximiser la détection de ces redondances. Pour rendre l'outil accessible à la communauté et utilisable au sein de workflow d'analyses, il a été nécessaire de le porter sous l'instance Galaxy W4M, ce qui m'a permis de réaliser l'ensemble des étapes de valorisation.
- Le second point a concerné le développement d'une stratégie de sélection de variables utilisant différents modèles statistiques afin d'arriver à proposer une signature du SMet. Nous avons ainsi été capable d'obtenir une première signature constituée de 31

métabolites, ainsi qu'une signature plus réduite de 6 métabolites et dont les applications en clinique seront plus facilement envisageables du fait du faible nombre de variables à mesurer. Les analyses ayant permis la constitution de ces signatures ont été réalisées sous la forme d'un workflow à l'aide de l'instance Galaxy W4M.

Toutefois, sa mise en place a soulevé plusieurs problématiques bio-informatiques, notamment : l'absence de moyen de conversion de certains formats de données et surtout de fusion entre des fichiers (intégration des 6 jeux de données entre eux) ; l'absence de certains outils au sein de W4M obligeant à tronquer le workflow en sortant les fichiers de Galaxy pour certaines étapes.

Au-delà de la mise en évidence d'une signature du SMet, l'un des objectifs était également de tenter d'apporter une meilleure compréhension de cet état de santé. Pour se faire, la représentation des métabolites modulés au sein de réseaux métaboliques est une étape indispensable. C'est à cette étape que nous avons rencontré le plus de difficultés. En effet, la conversion des identifiants entre les différentes bases de données et les réseaux est une étape encore très complexe entrainant la perte d'une quantité considérable d'information et ce malgré l'existence de certains outils de conversion. Cette difficulté est à la fois liée aux nomenclatures/ontologies non uniformisées comme évoqué lors de la conduction de la revue systématique, mais également à l'absence de lien entre les différentes bases de données. De plus, de nombreuses familles de métabolites (notamment les lipides) sont encore mal référencées et considérées encore comme un ensemble de métabolites variant dans le même sens et donc représentés par un identifiant unique au sein des réseaux, allant ainsi à l'encontre de la réalité biologique. Ces limitations représentent un énorme défi pour le domaine de la métabolomique et sont complexes à lever. Elles ne dépendent pas d'un utilisateur ou d'un outil donné mais bien de la communauté entière. Il est extrêmement difficile de proposer un système d'uniformisation du fait de la caractéristique très multidisciplinaire du domaine (chimie, biologie, bio-informatique) donnant lieu à autant d'habitudes et de nomenclatures que de communautés scientifiques. Malgré tout, certaines bases de données sont en train de devenir des références dans le domaine (*e.g.* ChEBI) et s'avèrent être une base solide pour tenter de constituer un socle d'information suffisamment détaillé pour limiter au maximum la perte d'information.

Le dernier objectif de ce travail de thèse était d'appliquer la démarche mise en place pour modéliser le spectre phénotypique du SMet. Plus précisément, il s'agissait d'étudier les potentiels sous-phénotypes du SMet pour tenter de proposer une reclassification moléculaire des individus basée sur les données métabolomiques. Nous avons initié cette démarche et pu mettre en évidence l'aspect réducteur des critères binaires définissant la pathologie. En effet, considérer les critères comme définis par des seuils simplifie une réalité biologique plus complexe chez des individus dont l'état de santé

n'est pas binaire. Nous avons réussi à proposer une reclassification des individus en 3 groupes basés sur l'utilisation des critères dans leur forme quantitative. Ces 3 clusters d'individus peuvent être prédits à l'aide de quelques dizaines de variables métabolomiques. Afin de valider l'existence d'une réalité biologique, l'exploration d'une méthode de clustering basée sur la construction en amont de réseaux métaboliques a été envisagée et restera une perspective d'intérêt.

II. Perspectives

Les perspectives ouvertes par ce travail de thèse sont nombreuses. Dans un premier temps, il serait intéressant d'approfondir l'étude des sous-phénotypes. L'approche de clustering à partir des réseaux métaboliques pourrait être développée en transposant la logique appliquée aux gènes et aux protéines aux données métabolomiques. Il serait également pertinent d'appliquer aux données métabolomiques d'autres stratégies de sélection de variables et d'investiguer d'autres méthodes de clustering.

Dans un second temps, il serait extrêmement intéressant de s'intéresser aux données épidémiologiques, nutritionnelles, d'activité physique, etc… disponibles sur les sujets de l'étude. Intégrer ces informations (notamment les informations nutritionnelles et activité physique), permettraient d'établir des profils de sujets multi-échelles et d'étudier les liens possibles entre métabolites et habitudes de vie. Cela apporterait une nouvelle dimension intégrée de connaissance pour une meilleure compréhension des déterminants de la pathologie et l'identification de cibles potentielles d'action.

Enfin, la validation des 2 signatures obtenues au sein d'un nouvel échantillon de personnes âgées serait une perspective particulièrement pertinente pour une applicabilité dans le domaine clinique.

D'un point de vue bio-informatique, des outils complémentaires de traitement automatisé pourraient être mis en place au sein d'instances comme W4M pour faciliter l'utilisation de workflows d'analyses des données métabolomiques multi-plateformes dans le cadre de projets d'épidémiologie des systèmes.

REFERENCES CONCLUSIONS ET PERSPECTIVES

1. Dammann O, Gray P, Gressens P, Wolkenhauer O, Leviton A. Systems Epidemiology: What's in a Name? Online J Public Health Inform. 2014;6(3):e198.

VALORISATION DU TRAVAIL DE THESE

Ce travail de thèse a fait l'objet de plusieurs valorisations scientifiques citées ci-dessous :

Publications scientifiques :

1. **Publié**

   - **Monnerie S**, Pétéra M, Gaudreau P, Comte B, Pujos-Guillot E. Analytic correlation filtration: a new tool to reduce analytical complexity of metabolomic datasets. *Metabolites.*

2. **En révision**

   - **Monnerie S**, Comte B, Ziegler D, Morais JA, Pujos-Guillot E, Gaudreau P. Metabolomic and lipidomic signatures of metabolic syndrome and its physiological components in aging: a systematic review. Submitted to *Scientific reports*.

3. **En cours**

   - Comte B, **Monnerie S**, Brandolini M, Canlet C, Castelli C, Colsch B, Fenaille F, Joly C, Lenuzza N, Lyan B, Martin JF, Migné C, Morais JA, Pétéra M, Thévenot E, Junot C, Gaudreau P, Pujos-Guillot E. Integrative multiplatform metabolomics for metabolic syndrome exploration. In preparation; to be submitted to *Molecular Systems Biology.*

   - Lenuzza N, Pétéra M, **Monnerie S**, Morais JA, Payette H, Gaudreau P, Thévenot E, Comte B, Pujos-Guillot E. The importance of an optimized selection process for biomarker discovery in epidemiological studies. Submitted to *Metabolomics.*

Présentation en congrès scientifiques (souligné : orateur) **:**

1. **Oraux**

   - **Monnerie S**, Ziegler D, Morais JA, Payette H, Comte B, Pujos-Guillot E, Gaudreau P. Biomarqueurs du syndrome métabolique et signatures métaboliques comme outil de caractérisation d'un vieillissement en santé : une revue systématique. *11ème Congrès International Francophone de Gérontologie et Gériatrie (CIFGG)*, June 13-15 2018, Montreux (SWITZERLAND).

   - Pujos-Guillot E, **Monnerie S**, Thévenot E, Junot C, Morais JA, Payette H, Gaudreau P, Comte B. Apport d'un phénotypage multidimensionnel dans la stratification du

syndrome métabolique. *11ème Congrès International Francophone de Gérontologie et Gériatrie (CIFGG)*, June 13-152018, Montreux (SWITZERLAND).

2. Flash poster

- **Monnerie S**, Pétéra M, Comte B, Pujos-Guillot E. A new tool to reduce analytical complexity of metabolomic dataset. 12èmes journées scientifiques du Réseau Francophone de Métabolomique et Fluxomique (RFMF), May 21-23 2019, Clermont-Ferrand (FRANCE).

3. Posters

3.1. Outil de filtration des corrélations ACorF

- **Monnerie S**, Pétéra M, Comte B, Pujos-Guillot E. A new tool to reduce analytical complexity of metabolomic dataset. *MetaboMeeting 2018*, December 17-19 2018, Nottingham (UNITED KINGDOM).

- **Monnerie S**, Pétéra M, Comte B, Pujos-Guillot E. A new tool to reduce analytical complexity of metabolomic dataset. *12èmes journées scientifiques du RFMF*, May 21-23 2019, Clermont-Ferrand (FRANCE).

- **Monnerie S**, Pétéra M, Comte B, Pujos-Guillot E. A new tool to reduce analytical complexity of metabolomic dataset. *15th Annual Conference of the Metabolomics Society – Metabolomics 2019*, June 23-27 2019, The Hague (NETHERLANDS).

3.2. Sélection des sujets dans le cadre du projet

- Lenuzza N, Thévenot E, Pétéra M., **Monnerie S**, Morais JA, Payette H, Gaudreau P, Comte B, Pujos-Guillot E. Optimized selection process to identify a metabolic syndrome metabolomic/lipidomic signature in older adults of the NuAge cohort. *13th International Conference of the European Union Geriatric Medicine Society*, September 20-22 2017, Nice (FRANCE).

- Lenuzza N, Thévenot E, Pétéra M, **Monnerie S**, Morais JA, Payette H, Gaudreau P, Comte B, Pujos-Guillot E. Optimized selection process to identify a metabolic syndrome metabolomic/lipidomic signature in older adults of the NuAge cohort. *46th Annual Scientific and Educational Meeting, Evidence for Action in an Aging World*, *Canadian Association of Gerontology 2017*, October 19-21 2017, Winnipeg (MB, CANADA).

3.3. Présentation globale du projet

- Pujos-Guillot E, **Monnerie S**, Thévenot E, Junot C, Morais JA, Payette H, Gaudreau P, Comte B. Input of multidimensional phenotyping in the metabolic syndrome stratification. *13th International Conference of the European Union Geriatric Medicine Society*, September 20-22 2017, Nice (FRANCE).

- Pujos-Guillot E, **Monnerie S**, Thévenot E, Junot C, Morais JA, Payette H, Gaudreau P, Comte B. Input of multidimensional phenotyping in the metabolic syndrome stratification. *46th Annual Scientific and Educational Meeting, Evidence for Action in an Aging World, Canadian Association of Gerontology 2017*, October 19-21 2017, Winnipeg (MB, CANADA).

- Pujos-Guillot E, **Monnerie S**, Thévenot E, Junot C, Morais JA, Payette H, Gaudreau P, Comte B. Input of multidimensional phenotyping in the metabolic syndrome stratification. *16th annual scientific sessions of the Society for Heart and Vascular Metabolism*, September 30 – October 3 2018, Isle of Palms – Charleston (SC, UNITED STATES).

- Comte B, **Monnerie S**, Canlet C, Castelli F, Colsch B, Fenaille F, Joly C, Junot C, Lenuzza N, Lyan B, Martin JF, Migné C, Pétéra M, Thevenot E, Tremblay-Franco M, Morais JA, Gaudreau P, Pujos-Guillot E. Metabolic characterization of metabolic syndrome: evidence from cross-sectional and longitudinal data. *19th International Conference on Systems Biology*, October 28 – November 1 2018, Lyon (FRANCE).

- Pujos-Guillot E, **Monnerie S**, Ziegler D, Pétéra M, Morais JA, Gaudreau P, Comte B. Far from the FAIR principle? Results from a systematic review on metabolomic / lipidomic signatures of metabolic syndrome. *MetaboMeeting 2018*, December 17-19 2018, Nottingham (UNITED KINGDOM).

3.4. Intégration des données cliniques et métabolomiques

- Brandolini-Bunlon M, Pétéra M, **Monnerie S**, Joly C, Morais J, Payette H, Gaudreau P, Comte B, Boccard J, Pujos-Guillot E. Evaluation de méthodes statistiques pour l'intégration de données métabolomiques, cliniques et alimentaires. *Congrès de Spectrométrie de Masse, Métabolomique et Analyse Protéomique 2017*, October 2-5 2017, Paris (FRANCE).

Autres valorisations

1. Prix et récompenses

- "Poster presentation Award" : **Monnerie S**, Pétéra M, Comte B, Pujos-Guillot E. A new tool to reduce analytical complexity of metabolomic dataset. *MetaboMeeting 2018*, December 17-19 2018, Nottingham (UNITED KINGDOM).

- Premier prix du Jury lors de la finale régionale de Ma Thèse en 180 secondes à Clermont-Ferrand le 7 mars 2019. Participation à la demi-finale nationale à Paris du 4 au 6 avril 2019.

- Bourse d'échange dans le cadre du COST15120 (Open MultiMed) pour un séjour de 5 semaines à Munich au sein du laboratoire de bio-informatique expérimentale de l'Université Technique de Munich du 11mars au 13 avril 2019.

- Bourse de voyage du Réseau Francophone de Métabolomique et Fluxomique pour assister à la conférence de la Métabolomique Society à La Haye du 23 au 27 juin 2019.

2. Vulgarisation scientifique associée à Ma Thèse en 180 Secondes

- Participation à l'émission de radio « Les décodeurs » sur France Bleu Pays d'Auvergne, animée par Jean-Luc Guillet, 7 mai 2019.

- Présentation ma thèse en 180 secondes lors des journées scientifiques de l'Unité de Nutrition Humaine, 4 Juin 2019, *INRA centre de Theix*.

- Présentation ma thèse en 180 secondes lors Rencontres Régionales de l'Enseignement Supérieur, de la Recherche et de l'Innovation, 21 juin 2019, *Centre de la Région Auvergne-Rhône-Alpes – Lyon.*

3. Autres

- Membre du parcours doctoral de l'École internationale de recherche d'Agreenium (EIR-A) et obtention du label Agreenium suite à la réalisation de 2 séjours à l'étranger de 5 semaines chacun :
    - Préparation de la revue systématique : équipe du Professeur Pierrette GAUDREAU, mars et avril 2017.
    - Etude des sous-phénotypes du SMet par approche réseau : équipe du Professeur Jan Baumbach, mars et avril 2019.

ANNEXE 1

Publication n°3

Lenuzza N, Pétéra M, **Monnerie S**, Morais JA, Payette H, Gaudreau P, Thévenot E, Comte B, Pujos-Guillot E.

The importance of an optimized selection process for biomarker discovery in epidemiological studies

Article soumis au journal « Metabolomics »

THE IMPORTANCE OF AN OPTIMIZED EXPERIMENTAL DESIGN FOR BIOMARKER DISCOVERY IN COHORT STUDIES USING METABOLOMICS

Natacha Lenuzza[1], Mélanie Pétéra[2], Stéphanie Monnerie[3], José Morais[4], Pierrette Gaudreau[5], Etienne Thévenot[1], Blandine Comte[3], Estelle Pujos-Guillot[2,3*]

[1]CEA, LIST, Laboratory for Data Sciences and Decision, MetaboHUB, Gif-sur-Yvette, France, [2]Université lermont Auvergne, INRA, UNH, CRNH Auvergne, F-63000 Clermont Ferrand, France; Université Clermont Auvergne, INRA, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, CRNH Auvergne, F-63000 Clermont-Ferrand, France; [3]Université Clermont Auvergne, INRA, UNH, CRNH Auvergne, F-63000 Clermont Ferrand, France, [4]Division of Geriatric Medicine, McGill University, Montreal, Canada, [5]Centre de Recherche du Centre hospitalier de l'Université de Montréal, Montreal, Canada; Département de médecine, Université de Montréal, Montreal, Canada.

*corresponding author: estelle.pujos-guillot@inra.fr

Abstract (100 mots)

Introduction: Metabolomic data reflects variations from a very large number of sources. Thus, designing metabolomics studies require special attention.

Objectives: Within a context of biomarker discovery using cohorts, the objective was to highlight important points to consider for an optimal experimental design.

Methods: Using a 2-timepoint case-control study on metabolic syndrome, subject selection and experimental design were investigated and discussed.

Results: An optimized design should consider: limiting missing values, defining status using quantitative parameters, studying outliers and potential confounding factors.

Conclusion: Applying this strategy and sharing associated results will contribute to a metabolomics open science.

Keywords

Please provide 4 to 6 keywords which can be used for indexing purposes:

case-control study, subject selection, metabolomics, systems epidemiology

1 - Introduction

The recent development of omics approaches allows in-depth studies of systems biology, aiming at a better understanding of the complexity of regulations in interaction with environmental factors. In particular, metabolomics, defined as the global study of small molecules/metabolites, provides an integrated view of metabolism. This powerful phenotyping tool has been therefore widely applied in epidemiology for metabolic disease diagnosis and candidate biomarker discovery, pathophysiological exploration of underlying mechanisms and for prediction and prognosis [159, 160].

However, performing large-scale metabolomics requires special attention when designing studies, in particular because of the high dimensionality of data [245]. First, to ensure high quality studies, investigators should at least define the general, biological pathways, metabolomic endpoints, as well as appropriate design [246]. Particular attention must therefore be paid to selection of subjects, collection and storage of samples, standard operating procedures for data production and data processing [247] to be able to extract useful information [248]. Case/control studies are commonly used for biomarker discovery when using metabolomics, as the relationship between metabolites and disease is usually strong when samples are all collected at the time of diagnosis. However, this type of design is also likely to be affected by bias [180, 249] and careful selection of subjects can become even more critical.

In the context of non-communicable diseases (NCDs), longitudinal study design (i.e., comparing each individual with himself/herself across time) are of major interest as it increases the ability to identify early biomarkers. The evolution of health in such systems diseases is represented by a continuum of transitions, involving multifaceted processes at multiple levels. Therefore, there is a major interest for biomarkers that can better characterize metabolic phenotypes and identify at-risk sub-populations. Within this objective of biomarker identification, the use of such a design in combination with metabolomics data generation requires specific guidelines to be able to perform adequate data treatment and obtain reliable models.

In this manuscript, a description of some general recommendations for subject selection within a 2-timepoint case-control study is presented, with the requirement for designing metabolomics analyses in a human cohort. In the framework of NCDs, we illustrate this particular issue with a study involving metabolic syndrome (MetS), known as predictor of type 2 diabetes (T2D) in different populations [250].

2 - Materials and methods

Study population: The present study was designed within the 5-year longitudinal observation study of NuAge (Quebec Longitudinal Study on Nutrition and Successful Aging) constituted of 1,793 healthy men and women, selected from three age groups (68-72, 73-77, 78-82) at recruitment [218]. French or English speaking community-dwelling participants were committed to give fasting blood annually and to answer questionnaires related to food and health biannually. The NuAge database comprises large qualitative and quantitative data related to nutrition/dietary intakes, physical activity, and numerous markers of physical and cognitive status, functional autonomy and social functioning. To ensure the quality and integrity of the collected samples, all sera of the NuAge subjects have been frozen rapidly in liquid nitrogen and kept at -80°C until use. Thawing was allowed only once at the time of sample preparation for a given analyte. At inclusion, a pool of fasting sera was created from the NuAge research team. This external standard has been used in multiple assays performed since then and reproducible results have been obtained.

Subject selection: A case control study on MetS was designed within the NuAge cohort, with serum samples collected at two time points (recruitment 2003-2005 (T1) and 3 years later (T4)), with the objective to identify a metabolic signature of MetS, stable over time, and to describe potential sub-phenotypes using a multiplatform lipidomic/metabolomics approach. In this context, an optimized participant selection strategy was developed.

Outcome definition: MetS was defined using the following criteria and thresholds defined for men (*Tan et al, 2009*; *Alberti et al, 2013*): elevated waist circumference (≥ 102 cm); high blood pressure (systolic > 130 mmHg and/or diastolic > 85 mmHg) or antihypertensive drug treatment with history of hypertension elevated fasting glucose (≥ 5.7 mM) or drug treatment of elevated glucose; high circulating triglyceride levels (≥ 1.7 mM) or drug treatment (fibrates, nicotinic acid) or high dose of ω3 fatty acids (3-4g/day); and reduced-HDL-cholesterol (< 1.0

mM) or drug treatment (fibrates, nicotinic acid) or high dose of ω3 fatty acids (3-4g/day). Assessment of variable stability over time (T1 to T4) was performed using a Kruskal-Wallis test.

Inclusion criteria: Cases were defined as having three or more of the MetS criteria. Two alternative strategies are possible for selecting control subjects, either using a random sample selection of at-risk subjects at the time of case incidence or by selecting completely disease-free subjects at a fixed time interval.

Exclusion criteria: One critical point concerns missing values of variables defining the outcome. Due to the 3-out-of-5 condition for MetS criteria, a missing value does not imply mandatory exclusion. Nonetheless, specific rules had to be defined to guaranty non-ambiguous status of subjects. Another critical point lies in subjects with very particular profiles regarding the criteria variables, both in terms of values and variation. In addition to the complexity of MetS definition, the challenge of status variation across time and the representativeness of sampling must be considered. In this study, a common threshold of 1.5 times the interquartile range of total individual values was used to define outlier values.

Subjects' characteristics: Since metabolomics is integrating a lot of environmental factors (nutrition, physical activity), the potential number of interacting factors is important and can lead to false conclusion. Distribution of a variable of interest between case and control classes can correlate with other uncontrolled variables. In particular, age, gender, diet, lifestyle, drug use, have been extensively described as confounding factors in metabolomics [180, 251]. In order to identify these potential biases, a list of potential cofounding variables was studied using multivariate statistical analyses. Principal component analysis (PCA) was performed to assess variation and show any trends or outlying data using the SIMCA v14 software (Umetrics, Umeå, Sweden).

Sample size: Estimation of statistical power and sample size is a key element of experimental design. However, in untargeted metabolomics, there is currently no standard approach, mostly because of the unknown nature of the expected effect. The variability of a large subset of metabolites and the consequences for epidemiologic studies were investigated previously [252]. The authors concluded that metabolomics requires large but feasible sample sizes to detect the moderate effect sizes typical for epidemiologic studies. As untargeted MS metabolomics is a data-driven approach, which consists in detecting thousands of metabolites in biological samples with various analytical and inter-individual variability, it is not relevant to calculate a unique statistical power to determine the number of subjects to be analyzed. However, different approaches and methods were proposed and authors generally recommend designing pilot studies to investigate this issue [253, 254]. In the present context, previous studies demonstrated that metabolomics allows risk assessment of pathologies using relatively limited sample size of about 100 subjects.

Randomization of biological samples

Following sample selection and in perspective of multiplatform analyses, sample preparation and analytical sequence have to be carefully built. In metabolomics, analytical sequences are usually randomized using a Williams Latin Square strategy defined according to the main factors of the study, as well as potential confounding factors linked to sampling conditions. Moreover, quality control samples have to be designed and prepared to control potential bias due to sample preparation or analytical drifts. Since in untargeted metabolomics hundreds/thousands of metabolites are detected, the use of internal standards is almost impossible and pooled quality control samples are recognized to be the most adequate solution [255].

3 Results and discussion

Subject selection:

The optimized subject selection was based on presence and number of MetS criteria, and their stability over three years (Figure 1). The selection was performed among the 853 males. In fact, it has been recognized that in Quebec, men have more risk factors of MetS than women do [256-258].

First, the three quantitative available MetS variables (namely glycemia, waist circumference, and blood pressure) were considered. Subjects with missing values both at T1 and T4 regarding these criteria were excluded. Then, subjects with unstable or uncertain MetS status (using the 5 criteria) over time were excluded.

At this step, 208 individuals were found eligible. Regarding dyslipidemia, to complement available qualitive data from questionnaires, quantitative measurements (triglyceride and HDL-cholesterol levels) were performed for these pre-selected subjects. Those with missing values due to experimental issues, at T1 and T4 were then excluded.

Finally, using all the quantitative MetS criteria, subjects with undefined MetS status at any time (due to remaining missing values) were also excluded. Regarding the study objective, only stable subjects over time were then included, considering their number of MetS criteria. Controls were defined as having less than three MetS criteria over time. It resulted in identifying 61 incident cases and 88 controls. Concerning control individuals, it was important to exclude extreme subjects that could generate false negative results. Therefore, in agreement with clinicians, controls with seven or more drug treatments were excluded. Moreover, value outliers were analyzed. Because no time effect was observed for the quantitative variables defining MetS, individuals with mean extreme values for MetS biological variables over time, outside the range defined by the mean (T1 to T4) ± 1.5 interquartile range (IQR) were excluded.

The present selection strategy was compared to a more common approach, which would have consisted in selecting from the 853 men of the cohort, all the individuals with an identified and stable MetS status at T1 and T4. This approach ended up with 108 cases and 135 controls. Half of this total of 243 subjects would have included individuals with at least one of the following limitations: missing values for some criteria (even with a defined MetS status), unstable MetS status at T2 and T3, outlier state concerning standard clinical measures. To obtain a design with around 120 subjects, a random selection would have been performed within the 2 groups, leading to more heterogeneous population with potential outliers. This would drastically increase the complexity of data interpretation.

Subject characteristics:

Regarding MetS criteria, cases showed mainly high waist circumference, hypertension and hyperglycemia. Moreover, it is important to note that all the controls are not free of MetS criteria, as described above. In fact, almost 50% of them are positive for hypertension and at risk for the other criteria. Because it is known that metabolomic profiles are modified by age, it was checked that there was no significant age difference between cases and controls to avoid a potential bias. To do so, three experimental classes were defined according to the age distribution (67-72 years old (n=25 *vs* 22), 73-77 years old (n=22 *vs* 24), 78-84 years old (n=15 *vs* 15)), and the size balance between age class in both groups was checked using Fisher's Exact Test.

Fifty-eight quantitative variables in total were considered to describe the selected population using PCA analysis: 22 biochemical parameters, 8 variables from clinical examination, 25 nutritional data (essentially related to macronutrient intake and selected nutrients described as related with MetS), and finally 4 scores in link with physical activity. Scores and loadings from the PCA of these parameters (after UV-scaling) from the 123 selected subjects are presented in Figure 2A and 2B, respectively. The results showed a discrimination between cases and controls mostly explained by the second principal component (11.7% of the total variance), with still a partial overlap. We could expect a limited putative confounding effect of each clinical criterion. The PCA also confirmed that no outlier subject had been selected regarding these parameters. The loading plot showed the absence of a major bias associated with a particular variable. If it has been the case, it would have led to consider this additional variable and associated thresholds for a new subject selection, in an iterative process. Moreover, in complement to univariate analyses, PCA allows studying and visualizing variable interactions, which give a first overview of key parameters to further explore using metabolomics data.

Randomization of biological samples

To avoid confounding effects between analytical and biological origins, a randomization of biological samples was performed. This process is always limited in the number of usable factors, by the subject and the group numbers. Therefore, it is important to investigate the subject metadata in advance. In the present case, samples were randomized with a Williams-Latin–Square-based strategy defined first according to the main factor of the study (MetS), considering the sum of the number of MetS criteria between the two time points (divided in 4 groups: 0-3; 3-7; 11-14; 15-20). This choice was made because of the important variability in the number of criteria in both groups. This heterogeneity is interesting to consider regarding the capacity of metabolomics to reveal sub-phenotypes. Second, the randomization was performed using a factor we expected to have a potential impact on metabolomics data, namely the region of sampling (Montreal, Sherbrooke, and Laval). Sampling dates were also considered but not used in the randomization process. In fact, samples were collected all over the seasons, that can

be reflected in metabolic profiles. When looking at a metabolic disease such as MetS in which nutrition plays a key role, this parameter is important to be included. Therefore, samples dates were categorized into 4 seasons to balance this factor in the sample sequence, and still avoid a possible bias. All this design was used both for sample preparation and for analytical sequences of metabolomic/lipidomic analyses.
These different factors were highlighted on the PCA score plots.
Figure 2C represents the projections on the PCA of the different factors taken into account in the randomization/sample design process. First, regarding the sum of positive MetS criteria along the 5-year study period, there was an effect corresponding to the discrimination between cases and controls. Second, no effect was observed when looking at the time of considered sampling (T1 vs T4). This is consistent with the fact that selection was performed on subjects who were stables for the MetS diagnosis between the two time points of sampling. Finally, Figure 2C showed that the effect of the sampling location and the season of blood sampling do not seem to have a major impact on the 58 variables selected. These analyses allowed obtaining a first view of the variability of the selected subjects. However, these factors will be further studied based on metabolomics data.

In this study, a large panel of variables expected to have an effect on metabolism or subject variability were controlled from the beginning. This approach was based on combined expertise concerning MetS as well as metabolomics on human samples. It allowed limiting excessive heterogeneity in the subject samples that could have left to difficulties to interpret the resulting metabolomic data. In case these key elements were not considered prior to subject selection, these variables could still have been studied in the perspective of sample randomisation. However, an excessive variability for some of these variables observed at this step could have make the randomisation more difficult. Indeed, the number of factors to consider for randomisation is still limited by the number of samples with the risk to end up with analytical bias confounded with some subtle biological effects. In this study, given the design and the number of subjects, it would have been difficult to consider even just one additional variable in the randomisation.

4 Conclusion and recommendations

In conclusion, when looking at a biological question regarding chronic diseases in cohorts using metabolomics, for better understanding pathophysiological mechanisms and/or with the objective to identify candidate biomarkers, it is important to optimize the study design. In particular, in order to ensure meaningful experiments and results, several points must be addressed by a collaborative and multidisciplinary consortium of epidemiologists, chemists, and biostatisticians. Beyond a well-defined, characterized population with relevant metadata from which subjects can be selected, a clear outcome will allow the definition of the inclusion, exclusion and confounding factors. Moreover, following this, an adequate selection process of the different groups excluding the presence of outliers is required to limit heterogeneity in the data. Secondly, concerning the samples to be analyzed, it is essential that sufficient information is provided to be able to check the sample quality requirements, and to study potential confounding factors to design randomization for sample preparation and analytical sequences. Finally, in order to allow comparisons of results between studies and apply good practices for an open science, all the subject characteristics and sample metadata have to be reported following the standards defined in the field [66, 79].

Figure 1: Flow chart for study subject selection

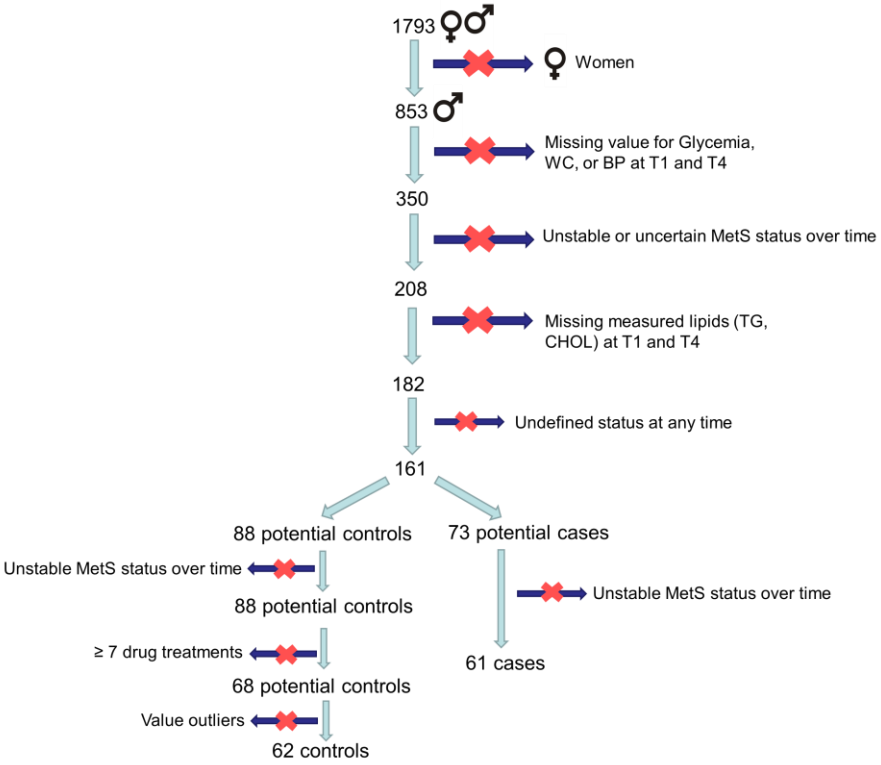
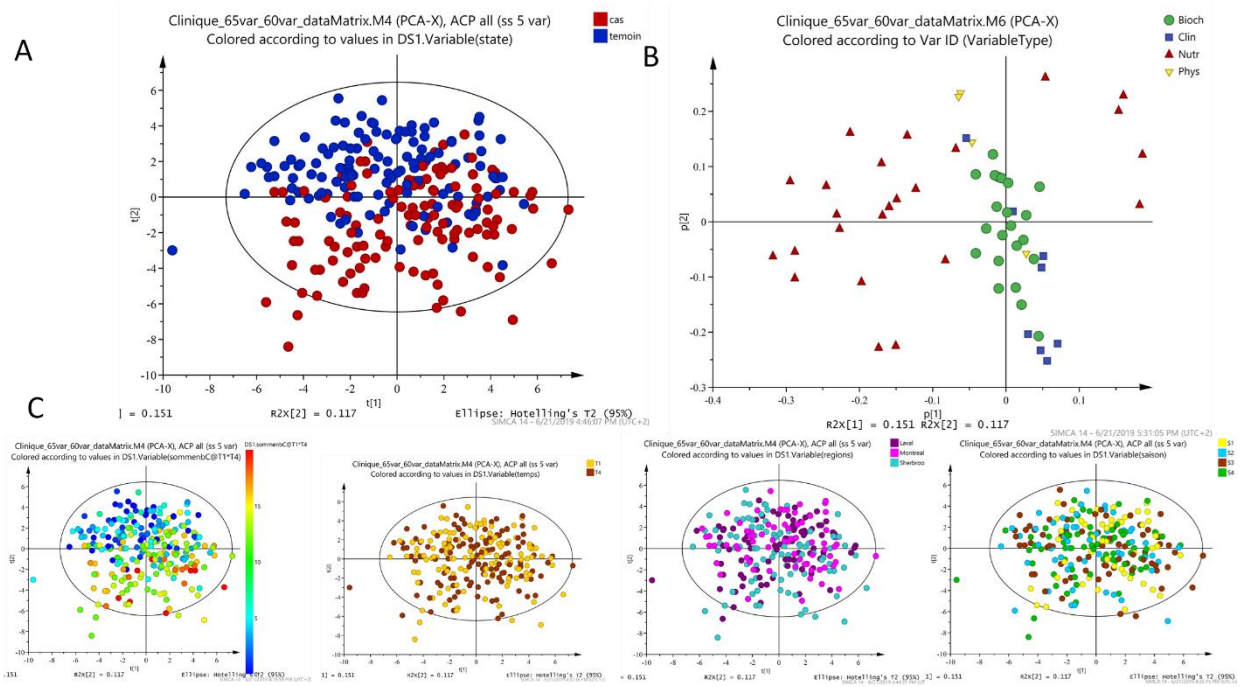


Figure 2: Principal component analysis derived from the 22 biochemical parameters, 9 variables from clinical examination, and 4 scores in link with physical activity, and 25 nutritional data considered from the 123 studied subjects. A: score plot (PC1: 15.1% of variance; PC2: 11.7% of variance); B: loading plot (PC1-PC2); C: projections of main confounding factors of the study design.



References

1. Mottillo, S., et al., *The metabolic syndrome and cardiovascular risk a systematic review and meta-analysis.* J Am Coll Cardiol, 2010. **56**(14): p. 1113-32.
2. Grundy, S.M., et al., *Diagnosis and management of the metabolic syndrome: an American Heart Association/National Heart, Lung, and Blood Institute Scientific Statement.* Circulation, 2005. **112**(17): p. 2735-52.
3. Alberti, K.G. and P.Z. Zimmet, *Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation.* Diabet Med, 1998. **15**(7): p. 539-53.
4. Balkau, B. and M.A. Charles, *Comment on the provisional report from the WHO consultation. European Group for the Study of Insulin Resistance (EGIR).* Diabet Med, 1999. **16**(5): p. 442-3.
5. National Cholesterol Education Program Expert Panel on Detection, E. and A. Treatment of High Blood Cholesterol in, *Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report.* Circulation, 2002. **106**(25): p. 3143-421.
6. Einhorn, D., et al., *American College of Endocrinology position statement on the insulin resistance syndrome.* Endocr Pract, 2003. **9**(3): p. 237-52.
7. Alberti, K.G., et al., *The metabolic syndrome--a new worldwide definition.* Lancet, 2005. **366**(9491): p. 1059-62.
8. Ford, E.S. and W.H. Giles, *A comparison of the prevalence of the metabolic syndrome using two proposed definitions.* Diabetes Care, 2003. **26**(3): p. 575-81.
9. Liao, Y., et al., *Critical evaluation of adult treatment panel III criteria in identifying insulin resistance with dyslipidemia.* Diabetes Care, 2004. **27**(4): p. 978-83.
10. Marchesini, G., et al., *WHO and ATPIII proposals for the definition of the metabolic syndrome in patients with Type 2 diabetes.* Diabet Med, 2004. **21**(4): p. 383-7.
11. Rodriguez, A., et al., *Contribution of impaired glucose tolerance in subjects with the metabolic syndrome: Baltimore Longitudinal Study of Aging.* Metabolism, 2005. **54**(4): p. 542-7.
12. Wyszynski, D.F., et al., *Relation between atherogenic dyslipidemia and the Adult Treatment Program-III definition of metabolic syndrome (Genetic Epidemiology of Metabolic Syndrome Project).* Am J Cardiol, 2005. **95**(2): p. 194-8.
13. Tan, C.E., et al., *Can we apply the National Cholesterol Education Program Adult Treatment Panel definition of the metabolic syndrome to Asians?* Diabetes Care, 2004. **27**(5): p. 1182-6.
14. Enkhmaa, B., et al., *Prevalence of the metabolic syndrome using the Third Report of the National Cholesterol Educational Program Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (ATP III) and the modified ATP III definitions for Japanese and Mongolians.* Clin Chim Acta, 2005. **352**(1-2): p. 105-13.
15. Hunt, K.J., et al., *National Cholesterol Education Program versus World Health Organization metabolic syndrome in relation to all-cause and cardiovascular mortality in the San Antonio Heart Study.* Circulation, 2004. **110**(10): p. 1251-7.
16. Alberti, K.G., et al., *Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity.* Circulation, 2009. **120**(16): p. 1640-5.
17. Cameron, A.J., J.E. Shaw, and P.Z. Zimmet, *The metabolic syndrome: prevalence in worldwide populations.* Endocrinol Metab Clin North Am, 2004. **33**(2): p. 351-75, table of contents.
18. Meigs, J.B., et al., *Prevalence and characteristics of the metabolic syndrome in the San Antonio Heart and Framingham Offspring Studies.* Diabetes, 2003. **52**(8): p. 2160-7.
19. Cameron AJ, S.J., Zimmet PZ, Chitson P, Alberti KGGM, Tuomilehto J, *Comparison of WHO and NCEP metabolic syndrome definitions over 5 years in Mauritius.* Diabetologia, 2003. **46**: p. A3068.

20. Laaksonen, D.E., et al., *Metabolic syndrome and development of diabetes mellitus: application and validation of recently suggested definitions of the metabolic syndrome in a prospective cohort study.* Am J Epidemiol, 2002. **156**(11): p. 1070-7.
21. Ford, E.S., W.H. Giles, and W.H. Dietz, *Prevalence of the metabolic syndrome among US adults: findings from the third National Health and Nutrition Examination Survey.* Jama, 2002. **287**(3): p. 356-359.
22. Resnick, H.E. and I. Strong Heart Study, *Metabolic syndrome in American Indians.* Diabetes Care, 2002. **25**(7): p. 1246-7.
23. Aguilar-Salinas, C.A., et al., *Analysis of the agreement between the World Health Organization criteria and the National Cholesterol Education Program-III definition of the metabolic syndrome: results from a population-based survey.* Diabetes Care, 2003. **26**(5): p. 1635.
24. Araneta, M.R., D.L. Wingard, and E. Barrett-Connor, *Type 2 diabetes and metabolic syndrome in Filipina-American women : a high-risk nonobese population.* Diabetes Care, 2002. **25**(3): p. 494-9.
25. Balkau, B., et al., *The incidence and persistence of the NCEP (National Cholesterol Education Program) metabolic syndrome. The French D.E.S.I.R. study.* Diabetes Metab, 2003. **29**(5): p. 526-32.
26. Azizi, F., et al., *Prevalence of metabolic syndrome in an urban population: Tehran Lipid and Glucose Study.* Diabetes Res Clin Pract, 2003. **61**(1): p. 29-37.
27. Onat, A., et al., *Metabolic syndrome: major impact on coronary risk in a population with low cholesterol levels--a prospective and cross-sectional evaluation.* Atherosclerosis, 2002. **165**(2): p. 285-92.
28. Al-Lawati, J.A., et al., *Prevalence of the metabolic syndrome among Omani adults.* Diabetes Care, 2003. **26**(6): p. 1781-5.
29. Deepa, R., et al., *Prevalence of insulin resistance syndrome in a selected south Indian population--the Chennai urban population study-7 [CUPS-7].* Indian J Med Res, 2002. **115**: p. 118-27.
30. Gupta, A., et al., *Prevalence of diabetes, impaired fasting glucose and insulin resistance syndrome in an urban Indian population.* Diabetes Res Clin Pract, 2003. **61**(1): p. 69-76.
31. Villegas, R., et al., *Prevalence and lifestyle determinants of the metabolic syndrome.* Ir Med J, 2004. **97**(10): p. 300-3.
32. Ervin, R.B., *Prevalence of metabolic syndrome among adults 20 years of age and over, by sex, age, race and ethnicity, and body mass index: United States, 2003-2006.* Natl Health Stat Report, 2009(13): p. 1-7.
33. Li, R., et al., *Prevalence of metabolic syndrome in Mainland China: a meta-analysis of published studies.* BMC Public Health, 2016. **16**: p. 296.
34. Suliga, E., D. Koziel, and S. Gluszek, *Prevalence of metabolic syndrome in normal weight individuals.* Ann Agric Environ Med, 2016. **23**(4): p. 631-635.
35. Geetha, L., et al., *Prevalence and clinical profile of metabolic obesity and phenotypic obesity in Asian Indians.* J Diabetes Sci Technol, 2011. **5**(2): p. 439-46.
36. de Carvalho Vidigal, F., et al., *Prevalence of metabolic syndrome in Brazilian adults: a systematic review.* BMC Public Health, 2013. **13**: p. 1198.
37. van Vliet-Ostaptchouk, J.V., et al., *The prevalence of metabolic syndrome and metabolically healthy obesity in Europe: a collaborative analysis of ten large cohort studies.* BMC Endocr Disord, 2014. **14**: p. 9.
38. Friend, A., L. Craig, and S. Turner, *The prevalence of metabolic syndrome in children: a systematic review of the literature.* Metab Syndr Relat Disord, 2013. **11**(2): p. 71-80.
39. Iqbal, A.Z., et al., *Prevalence of the metabolic syndrome and its component abnormalities among school age Pakistani children.* J Ayub Med Coll Abbottabad, 2014. **26**(2): p. 194-9.
40. Kelleher, J.K., *Probing metabolic pathways with isotopic tracers: insights from mammalian metabolic physiology.* Metab Eng, 2004. **6**(1): p. 1-5.

41. Wishart, D.S., et al., *HMDB 3.0--The Human Metabolome Database in 2013.* Nucleic Acids Res, 2013. **41**(Database issue): p. D801-7.
42. Wishart, D.S., et al., *HMDB: a knowledgebase for the human metabolome.* Nucleic Acids Res, 2009. **37**(Database issue): p. D603-10.
43. Nicholson, J.K., J.C. Lindon, and E. Holmes, *'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data.* Xenobiotica, 1999. **29**(11): p. 1181-9.
44. Cajka, T. and O. Fiehn, *Comprehensive analysis of lipids in biological systems by liquid chromatography-mass spectrometry.* Trends Analyt Chem, 2014. **61**: p. 192-206.
45. Strimbu, K. and J.A. Tavel, *What are biomarkers?* Curr Opin HIV AIDS, 2010. **5**(6): p. 463-6.
46. Dunn, W.B., et al., *Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy.* Chem Soc Rev, 2011. **40**(1): p. 387-426.
47. Griffin, J.L., *Metabolic profiles to define the genome: can we hear the phenotypes?* Philos Trans R Soc Lond B Biol Sci, 2004. **359**(1446): p. 857-71.
48. Nicholson, J.K. and J.C. Lindon, *Systems biology: Metabonomics.* Nature, 2008. **455**(7216): p. 1054-6.
49. Bothwell, J.H. and J.L. Griffin, *An introduction to biological nuclear magnetic resonance spectroscopy.* Biol Rev Camb Philos Soc, 2011. **86**(2): p. 493-510.
50. Blümich, B. and P.T. Callaghan, *Principle of nuclear magnetic resonance microscopy*. Oxford University Press. Vol. 492. 1993.
51. Ward, J.L., J.M. Baker, and M.H. Beale, *Recent applications of NMR spectroscopy in plant metabolomics.* FEBS J, 2007. **274**(5): p. 1126-31.
52. Brunius, C., L. Shi, and R. Landberg, *Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction.* Metabolomics, 2016. **12**(11): p. 173.
53. Dunn, W.B., et al., *The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans.* Bioanalysis, 2012. **4**(18): p. 2249-64.
54. Sangster, T., et al., *A pragmatic and readily implemented quality control strategy for HPLC-MS and GC-MS-based metabonomic analysis.* Analyst, 2006. **131**(10): p. 1075-8.
55. Hilario, M., et al., *Processing and classification of protein mass spectra.* Mass Spectrom Rev, 2006. **25**(3): p. 409-49.
56. Boccard, J., J.L. Veuthey, and S. Rudaz, *Knowledge discovery in metabolomics: an overview of MS data handling.* J Sep Sci, 2010. **33**(3): p. 290-304.
57. Pluskal, T., et al., *MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data.* BMC Bioinformatics, 2010. **11**: p. 395.
58. Lommen, A. and H.J. Kools, *MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware.* Metabolomics, 2012. **8**(4): p. 719-726.
59. Smith, C.A., et al., *XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification.* Anal Chem, 2006. **78**(3): p. 779-87.
60. Ulaszewska, M.M., et al., *Nutrimetabolomics: An Integrative Action for Metabolomic Analyses in Human Nutritional Studies.* Mol Nutr Food Res, 2019. **63**(1): p. e1800384.
61. Gromski, P.S., et al., *Influence of missing values substitutes on multivariate analysis of metabolomics data.* Metabolites, 2014. **4**(2): p. 433-52.
62. Wei, R., et al., *Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data.* Sci Rep, 2018. **8**(1): p. 663.
63. Kumar, N., et al., *Metabolomic Biomarker Identification in Presence of Outliers and Missing Values.* Biomed Res Int, 2017. **2017**: p. 2437608.
64. Rocca-Serra, P., et al., *Data standards can boost metabolomics research, and if there is a will, there is a way.* Metabolomics, 2016. **12**: p. 14.
65. Cook, C.E., et al., *The European Bioinformatics Institute in 2016: Data growth and integration.* Nucleic Acids Res, 2016. **44**(D1): p. D20-6.

66. Haug, K., R.M. Salek, and C. Steinbeck, *Global open data management in metabolomics.* Curr Opin Chem Biol, 2017. **36**: p. 58-63.
67. Haug, K., et al., *MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data.* Nucleic Acids Res, 2013. **41**(Database issue): p. D781-6.
68. Sud, M., et al., *Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools.* Nucleic Acids Res, 2016. **44**(D1): p. D463-70.
69. Ferry-Dumazet, H., et al., *MeRy-B: a web knowledgebase for the storage, visualization, analysis and annotation of plant NMR metabolomic profiles.* BMC Plant Biol, 2011. **11**: p. 104.
70. Ara, T., et al., *Metabolonote: a wiki-based database for managing hierarchical metadata of metabolome analyses.* Front Bioeng Biotechnol, 2015. **3**: p. 38.
71. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Sci Data, 2016. **3**: p. 160018.
72. Lampen, P., et al., *JCAMP-DX for mass-spectrometry.* Applied Spectroscopy, 1994.
73. Rew, R. and G. Davis, *NetCDF: an interface for scientific data access.* IEEE Computer Graphics and Applications, 1990. **10**(4): p. 76-82.
74. Pedrioli, P.G., et al., *A common open representation of mass spectrometry data and its application to proteomics research.* Nat Biotechnol, 2004. **22**(11): p. 1459-66.
75. Orchard, S., et al., *Current status of proteomic standards development.* Expert Rev Proteomics, 2004. **1**(2): p. 179-83.
76. Martens, L., et al., *mzML--a community standard for mass spectrometry data.* Mol Cell Proteomics, 2011. **10**(1): p. R110 000133.
77. Wilhelm, M., et al., *mz5: space- and time-efficient storage of mass spectrometry data sets.* Mol Cell Proteomics, 2012. **11**(1): p. O111 011379.
78. Lindon, J.C., et al., *Summary recommendations for standardization and reporting of metabolic analyses.* Nat Biotechnol, 2005. **23**(7): p. 833-8.
79. Sumner, L.W., et al., *Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI).* Metabolomics, 2007. **3**(3): p. 211-221.
80. Members, M.S.I.B., et al., *The metabolomics standards initiative.* Nat Biotechnol, 2007. **25**(8): p. 846-8.
81. Rocca-Serra, P., et al., *ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level.* Bioinformatics, 2010. **26**(18): p. 2354-6.
82. Giacomoni, F., et al., *Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics.* Bioinformatics, 2015. **31**(9): p. 1493-5.
83. Guitton, Y., et al., *Create, run, share, publish, and reference your LC-MS, FIA-MS, GC-MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics.* Int J Biochem Cell Biol, 2017. **93**: p. 89-101.
84. Broadhurst, D.I. and D.B. Kell, *Statistical strategies for avoiding false discoveries in metabolomics and related experiments.* Metabolomics, 2006. **2**.
85. Holmes, E. and H. Antti, *Chemometric contributions to the evolution of metabonomics: mathematical solutions to characterising and interpreting complex biological NMR spectra.* Analyst, 2002. **127**(12): p. 1549-57.
86. Serkova, N.J., T.J. Standiford, and K.A. Stringer, *The emerging field of quantitative blood metabolomics for biomarker discovery in critical illnesses.* Am J Respir Crit Care Med, 2011. **184**(6): p. 647-55.
87. Morrow, D.A. and J.A. de Lemos, *Benchmarks for the assessment of novel cardiovascular biomarkers.* Circulation, 2007. **115**(8): p. 949-52.
88. Rasmussen, L., et al., *Standardization of factors that influence humain urine metabolomics.* Metabolomics, 2011. **7**.

89. Winnike, J.H., et al., *Effects of a prolonged standardized diet on normalizing the human metabolome.* Am J Clin Nutr, 2009. **90**(6): p. 1496-501.
90. Townsend, M.K., et al., *Reproducibility of metabolomic profiles among men and women in 2 large cohort studies.* Clin Chem, 2013. **59**(11): p. 1657-67.
91. Vinaixa, M., et al., *A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data.* Metabolites, 2012. **2**(4): p. 775-95.
92. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the royal statistical society, 1995. **57**: p. 289-300.
93. Bro, R. and A.K. Smilde, *Principal component analysis.* Analytical methods, 2014(9).
94. World, S., K. Esbensen, and P. Geladi, *Principal component analysis.* Chemometrics and intelligent laboratory systems, 1987. **2**(1-3): p. 37-52.
95. Hotelling, H., *Analysis of a complex of statistical variables into principal components.* Journal of educational psychology, 1933. **24**(6): p. 417-441.
96. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.
97. Xia, J., et al., *Translational biomarker discovery in clinical metabolomics: an introductory tutorial.* Metabolomics, 2013. **9**(2): p. 280-299.
98. Fonville, J.M., et al., *The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping.* Journal of chemometrics, 2010. **24**(11-12): p. 636-349.
99. Trygg, J. and S. Wold, *Orthogonal projections to latent structures (O-PLS).* Journal of chemometrics, 2002. **16**(3): p. 119-128.
100. Tapp, H. and E.K. Kemsley, *Notes on the practical utility of OPLS.* Trends in Analytical Chemistry, 2009. **28**(11): p. 1322-1327.
101. Gromski, P.S., et al., *A tutorial review: Metabolomics and partial least squares-discriminant analysis--a marriage of convenience or a shotgun wedding.* Anal Chim Acta, 2015. **879**: p. 10-23.
102. Kim, Y., et al., *Multivariate classification of urine metabolome profiles for breast cancer diagnosis.* BMC Bioinformatics, 2010. **11 Suppl 2**: p. S4.
103. Luts, J., et al., *A tutorial on support vector machine-based methods for classification problems in chemometrics.* Anal Chim Acta, 2010. **665**(2): p. 129-45.
104. Mahadevan, S., et al., *Analysis of metabolomic data using support vector machines.* Anal Chem, 2008. **80**(19): p. 7562-70.
105. Sing, T., et al., *ROCR: visualizing classifier performance in R.* Bioinformatics, 2005. **21**(20): p. 3940-1.
106. Robin, X., et al., *pROC: an open-source package for R and S+ to analyze and compare ROC curves.* BMC Bioinformatics, 2011. **12**: p. 77.
107. Lacroix, Z. and T. Critchlow, *Bioinformatics: Managing Scientific Data*. 2003: Morgan Kaufmann Publishers Inc.
108. Ellinger, J.J., et al., *Databases and Software for NMR-Based Metabolomics.* Curr Metabolomics, 2013. **1**(1).
109. Fukushima, A. and M. Kusano, *Recent progress in the development of metabolome databases for plant systems biology.* Front Plant Sci, 2013. **4**: p. 73.
110. Kim, S., et al., *PubChem 2019 update: improved access to chemical data.* Nucleic Acids Res, 2019. **47**(D1): p. D1102-D1109.
111. Hastings, J., et al., *The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013.* Nucleic Acids Res, 2013. **41**(Database issue): p. D456-63.
112. Horai, H., et al., *MassBank: a public repository for sharing mass spectral data for life sciences.* J Mass Spectrom, 2010. **45**(7): p. 703-14.
113. Tautenhahn, R., et al., *An accelerated workflow for untargeted metabolomics using the METLIN database.* Nat Biotechnol, 2012. **30**(9): p. 826-8.

114. Kopka, J., et al., *GMD@CSB.DB: the Golm Metabolome Database.* Bioinformatics, 2005. **21**(8): p. 1635-8.
115. Kind, T., et al., *FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry.* Anal Chem, 2009. **81**(24): p. 10038-48.
116. Wohlgemuth, G., et al., *The Chemical Translation Service--a web-based tool to improve standardization of metabolomic reports.* Bioinformatics, 2010. **26**(20): p. 2647-8.
117. Domingo-Almenara, X., et al., *eRah: A Computational Tool Integrating Spectral Deconvolution and Alignment with Quantification and Identification of Metabolites in GC/MS-Based Metabolomics.* Anal Chem, 2016. **88**(19): p. 9821-9829.
118. Hiller, K., et al., *MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis.* Anal Chem, 2009. **81**(9): p. 3429-39.
119. Tulpan, D., et al., *MetaboHunter: an automatic approach for identification of metabolites from 1H-NMR spectra of complex mixtures.* BMC Bioinformatics, 2011. **12**: p. 400.
120. Jacob, D., C. Deborde, and A. Moing, *An efficient spectra processing method for metabolite identification from 1H-NMR metabolomics data.* Anal Bioanal Chem, 2013. **405**(15): p. 5049-61.
121. Mercier, P., et al., *Towards automatic metabolomic profiling of high-resolution one-dimensional proton NMR spectra.* J Biomol NMR, 2011. **49**(3-4): p. 307-23.
122. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular data sets.* Nucleic Acids Res, 2012. **40**(Database issue): p. D109-14.
123. Jewison, T., et al., *SMPDB 2.0: big improvements to the Small Molecule Pathway Database.* Nucleic Acids Res, 2014. **42**(Database issue): p. D478-84.
124. Kelder, T., et al., *WikiPathways: building research communities on biological pathways.* Nucleic Acids Res, 2012. **40**(Database issue): p. D1301-7.
125. Caspi, R., et al., *The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.* Nucleic Acids Res, 2008. **36**(Database issue): p. D623-31.
126. Garcia-Alcalde, F., et al., *Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data.* Bioinformatics, 2011. **27**(1): p. 137-9.
127. Rohn, H., et al., *VANTED v2: a framework for systems biology applications.* BMC Syst Biol, 2012. **6**: p. 139.
128. Smoot, M.E., et al., *Cytoscape 2.8: new features for data integration and network visualization.* Bioinformatics, 2011. **27**(3): p. 431-2.
129. Xia, J., et al., *MetaboAnalyst 2.0--a comprehensive server for metabolomic data analysis.* Nucleic Acids Res, 2012. **40**(Web Server issue): p. W127-33.
130. Kim, J., et al., *Network rewiring is an important mechanism of gene essentiality change.* Sci Rep, 2012. **2**: p. 900.
131. Thiele, I., et al., *A community-driven global reconstruction of human metabolism.* Nat Biotechnol, 2013. **31**(5): p. 419-25.
132. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.
133. Krumsiek, J., et al., *Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data.* BMC Syst Biol, 2011. **5**: p. 21.
134. Valcarcel, B., et al., *A differential network approach to exploring differences between biological states: an application to prediabetes.* PLoS One, 2011. **6**(9): p. e24702.
135. Cottret, L., et al., *MetExplore: collaborative edition and exploration of metabolic networks.* Nucleic Acids Res, 2018. **46**(W1): p. W495-W502.
136. Gao, J., et al., *Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks.* Bioinformatics, 2010. **26**(7): p. 971-3.
137. Latendresse, M. and P.D. Karp, *Web-based metabolic network visualization with a zooming user interface.* BMC Bioinformatics, 2011. **12**: p. 176.

138. Dammann, O., et al., *Systems Epidemiology: What's in a Name?* Online J Public Health Inform, 2014. **6**(3): p. e198.
139. Moher, D., et al., *Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.* J Clin Epidemiol, 2009. **62**(10): p. 1006-12.
140. Liberati, A., et al., *The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration.* J Clin Epidemiol, 2009. **62**(10): p. e1-34.
141. Alberti, K.G., P. Zimmet, and J. Shaw, *Metabolic syndrome--a new world-wide definition. A Consensus Statement from the International Diabetes Federation.* Diabet Med, 2006. **23**(5): p. 469-80.
142. Day, C., *Metabolic syndrome, or What you will: definitions and epidemiology.* Diab Vasc Dis Res, 2007. **4**(1): p. 32-8.
143. Lam, D.W. and D. LeRoith, *Metabolic Syndrome*, in *Endotext*, L.J. De Groot, et al., Editors. 2000: South Dartmouth (MA).
144. Ford, E.S., et al., *Prospective association between lung function and the incidence of diabetes: findings from the National Health and Nutrition Examination Survey Epidemiologic Follow-up Study.* Diabetes Care, 2004. **27**(12): p. 2966-70.
145. Beltran-Sanchez, H., et al., *Prevalence and trends of metabolic syndrome in the adult U.S. population, 1999-2010.* J Am Coll Cardiol, 2013. **62**(8): p. 697-703.
146. Saklayen, M.G., *The Global Epidemic of the Metabolic Syndrome.* Curr Hypertens Rep, 2018. **20**(2): p. 12.
147. Nolan, P.B., et al., *Prevalence of metabolic syndrome and metabolic syndrome components in young adults: A pooled analysis.* Prev Med Rep, 2017. **7**: p. 211-215.
148. Ford, E.S., W.H. Giles, and W.H. Dietz, *Prevalence of the metabolic syndrome among US adults: findings from the third National Health and Nutrition Examination Survey.* JAMA, 2002. **287**(3): p. 356-9.
149. Kaur, J., *A comprehensive review on metabolic syndrome.* Cardiol Res Pract, 2014. **2014**: p. 943162.
150. Ranasinghe, P., et al., *Prevalence and trends of metabolic syndrome among adults in the asia-pacific region: a systematic review.* BMC Public Health, 2017. **17**(1): p. 101.
151. Wilson, P.W., et al., *Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus.* Circulation, 2005. **112**(20): p. 3066-72.
152. Stern, M.P., et al., *Does the metabolic syndrome improve identification of individuals at risk of type 2 diabetes and/or cardiovascular disease?* Diabetes Care, 2004. **27**(11): p. 2676-81.
153. Li, C. and E.S. Ford, *Definition of the Metabolic Syndrome: What's New and What Predicts Risk?* Metab Syndr Relat Disord, 2006. **4**(4): p. 237-51.
154. Grundy, S.M., *Pre-diabetes, metabolic syndrome, and cardiovascular risk.* J Am Coll Cardiol, 2012. **59**(7): p. 635-43.
155. Poon, V.T., J.L. Kuk, and C.I. Ardern, *Trajectories of metabolic syndrome development in young adults.* PLoS One, 2014. **9**(11): p. e111647.
156. Steinbrecher, A. and T. Pischon, *The potential use of biomarkers in the prevention of Type 2 diabetes.* Expert Review of Endocrinology and Metabolism, 2013. **8**(3): p. 217-219.
157. Ramautar, R., et al., *Human metabolomics: Strategies to understand biology.* 2013.
158. Liggi, S. and J.L. Griffin, *Metabolomics applied to diabetes-lessons from human population studies.* International Journal of Biochemistry & Cell Biology, 2017. **93**: p. 136-147.
159. Zhang, A.H., et al., *Metabolomics in diabetes.* Clin Chim Acta, 2014. **429**: p. 106-10.
160. Park, S., K.C. Sadanala, and E.K. Kim, *A Metabolomic Approach to Understanding the Metabolic Link between Obesity and Diabetes.* Mol Cells, 2015. **38**(7): p. 587-96.
161. Bain, J.R., et al., *Metabolomics applied to diabetes research: moving from information to knowledge.* Diabetes, 2009. **58**(11): p. 2429-2443.
162. Cajka, T. and O. Fiehn, *Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics.* Anal Chem, 2016. **88**(1): p. 524-45.

163. Duarte, I.F., S.O. Diaz, and A.M. Gil, *NMR metabolomics of human blood and urine in disease research.* J Pharm Biomed Anal, 2014. **93**: p. 17-26.
164. Forcisi, S., et al., *Liquid chromatography-mass spectrometry in metabolomics research: mass analyzers in ultra high pressure liquid chromatography coupling.* J Chromatogr A, 2013. **1292**: p. 51-65.
165. Capel, F., et al., *Metabolomics reveals plausible interactive effects between dairy product consumption and metabolic syndrome in humans.* Clin Nutr, 2019.
166. Surowiec, I., et al., *Metabolomic and lipidomic assessment of the metabolic syndrome in Dutch middle-aged individuals reveals novel biological signatures separating health and disease.* Metabolomics, 2019. **15**(2): p. 23.
167. Becker, S., et al., *LC-MS-based metabolomics in the clinical laboratory.* Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences.
168. Alonso, A., S. Marsal, and A. Julia, *Analytical methods in untargeted metabolomics: state of the art in 2015.* Front Bioeng Biotechnol, 2015. **3**: p. 23.
169. Sas, K.M., et al., *Metabolomics and diabetes: analytical and computational approaches.* Diabetes, 2015. **64**(3): p. 718-732.
170. Liebisch, G., et al., *Shorthand notation for lipid structures derived from mass spectrometry.* J Lipid Res, 2013. **54**(6): p. 1523-30.
171. Sud, M., et al., *LMSD: LIPID MAPS structure database.* Nucleic Acids Res, 2007. **35**(Database issue): p. D527-32.
172. Gonzalez-Franquesa, A., et al., *What Have Metabolomics Approaches Taught Us About Type 2 Diabetes?* Curr Diab Rep, 2016. **16**(8): p. 74.
173. Palau-Rodriguez, M., et al., *Metabolomic insights into the intricate gut microbial-host interaction in the development of obesity and type 2 diabetes.* Front Microbiol, 2015. **6**: p. 1151.
174. Shapiro, H., J. Suez, and E. Elinav, *Personalized microbiome-based approaches to metabolic syndrome management and prevention.* J Diabetes, 2017. **9**(3): p. 226-236.
175. Forouhi, N.G., et al., *Differences in the prospective association between individual plasma phospholipid saturated fatty acids and incident type 2 diabetes: the EPIC-InterAct case-cohort study.* The lancet, 2014. **Diabetes & endocrinology. 2**(10): p. 810-818.
176. Kale, N.S., et al., *MetaboLights: An Open-Access Database Repository for Metabolomics Data.* Curr Protoc Bioinformatics, 2016. **53**: p. 14 13 1-18.
177. Hardy, N.W. and C. Taylor, *A roadmap for the establishment of standard data exchange structures for metabolomics.* Metabolomics, 2007. **3**: p. 243-248.
178. Lehmann, R., *Diabetes subphenotypes and metabolomics: The key to discovering laboratory markers for personalized medicine?* 2013.
179. Grissa, D., et al., *Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data.* Front Mol Biosci, 2016. **3**: p. 30.
180. Broadhurst, D.I. and D.B. Kell, *Statistical strategies for avoiding false discoveries in metabolomics and related experiments.* Metabolomics, 2006. **2**(4): p. 171-196.
181. Saccenti, E., et al., *Reflections on univariate and multivariate analysis of metabolomics data.* metabolomics, 2014. **10**: p. 361-374.
182. Vinaixa, M., et al., *Mass spectral databases for LC/MS and GC/MS-based metabolomics: state of the field and future prospects.* Trends in Analytical Chemistry, 2015.
183. Lindon, J.C. and J.K. Nicholson, *The emergent role of metabolic phenotyping in dynamic patient stratification.* Expert Opin Drug Metab Toxicol, 2014. **10**(7): p. 915-9.
184. Dumas, M.E., J. Kinross, and J.K. Nicholson, *Metabolic phenotyping and systems biology approaches to understanding metabolic syndrome and fatty liver disease.* Gastroenterology, 2014. **146**(1): p. 46-62.
185. Wiklund, P.K., et al., *Serum metabolic profiles in overweight and obese women with and without metabolic syndrome.* Diabetol Metab Syndr, 2014. **6**(1): p. 40.

186. Pujos-Guillot, E., et al., *Systems Metabolomics for Prediction of Metabolic Syndrome.* J Proteome Res, 2017. **16**(6): p. 2262-2272.
187. Sperling, L.S., et al., *The CardioMetabolic Health Alliance: Working Toward a New Care Model for the Metabolic Syndrome.* J Am Coll Cardiol, 2015. **66**(9): p. 1050-67.
188. Bardou, P., et al., *jvenn: an interactive Venn diagram viewer.* BMC Bioinformatics, 2014. **15**: p. 293.
189. Caimi, G., et al., *Evaluation of nitric oxide metabolites in a group of subjects with metabolic syndrome.* Diabetes Metab Syndr, 2012. **6**(3): p. 132-5.
190. James-Todd, T.M., et al., *The association between phthalates and metabolic syndrome: the National Health and Nutrition Examination Survey 2001-2010.* Environ Health, 2016. **15**: p. 52.
191. Kulkarni, H., et al., *Variability in associations of phosphatidylcholine molecular species with metabolic syndrome in Mexican-American families.* Lipids, 2013. **48**(5): p. 497-503.
192. Ntzouvani, A., et al., *Amino acid profile and metabolic syndrome in a male Mediterranean population: A cross-sectional study.* Nutr Metab Cardiovasc Dis, 2017.
193. Olszanecka, A., K. Kawecka-Jaszcz, and D. Czarnecka, *Association of free testosterone and sex hormone binding globulin with metabolic syndrome and subclinical atherosclerosis but not blood pressure in hypertensive perimenopausal women.* Arch Med Sci, 2016. **12**(3): p. 521-8.
194. Ramakrishanan, N., et al., *Exploratory lipidomics in patients with nascent Metabolic Syndrome.* Journal of Diabetes and its Complications, 2018. **32**(8): p. 791-794.
195. Shim, K., R. Gulhar, and I. Jialal, *Exploratory metabolomics of nascent metabolic syndrome.* Journal of Diabetes and its Complications, 2019. **33**(3): p. 212-216.
196. Tremblay-Franco, M., et al., *Effect of obesity and metabolic syndrome on plasma oxysterols and fatty acids in human.* Steroids, 2015. **99**(Pt B): p. 287-92.
197. Antonio, L., et al., *Associations between sex steroids and the development of metabolic syndrome: A longitudinal study in European men.* Journal of Clinical Endocrinology and Metabolism, 2015. **100**(4): p. 1396-1404.
198. Barrea, L., et al., *Trimethylamine-N-oxide (TMAO) as novel potential biomarker of early predictors of metabolic syndrome.* Nutrients, 2018. **10**(12).
199. Blouin, K., et al., *Contribution of age and declining androgen levels to features of the metabolic syndrome in men.* Metabolism, 2005. **54**(8): p. 1034-40.
200. Cheng, S., et al., *Metabolite profiling identifies pathways associated with metabolic risk in humans.* Circulation, 2012. **125**(18): p. 2222-31.
201. Favennec, M., et al., *The kynurenine pathway is activated in human obesity and shifted toward kynurenine monooxygenase activation.* Obesity (Silver Spring), 2015. **23**(10): p. 2066-74.
202. Gao, X., et al., *Unfavorable associations between serum trimethylamine N-oxide and L-carnitine levels with components of metabolic syndrome in the Newfoundland population.* Frontiers in Endocrinology, 2019. **10**(MAR).
203. Ho, J.E., et al., *Metabolomic Profiles of Body Mass Index in the Framingham Heart Study Reveal Distinct Cardiometabolic Phenotypes.* PLoS One, 2016. **11**(2): p. e0148361.
204. Huynh, K., et al., *High-Throughput Plasma Lipidomics: Detailed Mapping of the Associations with Cardiometabolic Risk Factors.* Cell Chemical Biology, 2019. **26**(1): p. 71-84.e4.
205. Liu, J., et al., *A Mendelian Randomization Study of Metabolite Profiles, Fasting Glucose, and Type 2 Diabetes.* Diabetes, 2017. **66**(11): p. 2915-2926.
206. Marchand, G.B., et al., *Increased body fat mass explains the positive association between circulating estradiol and insulin resistance in postmenopausal women.* American journal of physiology, 2018. **Endocrinology and metabolism. 314**(5): p. E448-E456.
207. Neeland, I.J., et al., *Relation of plasma ceramides to visceral adiposity, insulin resistance and the development of type 2 diabetes mellitus: the Dallas Heart Study.* Diabetologia, 2018. **61**(12): p. 2570-2579.
208. Ottosson, F., et al., *Altered asparagine and glutamate homeostasis precede coronary artery disease and type 2 diabetes.* Journal of Clinical Endocrinology and Metabolism, 2018. **103**(8): p. 3060-3069.

209. Wang-Sattler, R., et al., *Novel biomarkers for pre-diabetes identified by metabolomics.* Mol Syst Biol, 2012. **8**: p. 615.
210. Lind, P.M., B. Zethelius, and L. Lind, *Circulating levels of phthalate metabolites are associated with prevalent diabetes in the elderly.* Diabetes Care, 2012. **35**(7): p. 1519-24.
211. Liu, J., et al., *Metabolomics based markers predict type 2 diabetes in a 14-year follow-up study.* 2017.
212. Meikle, P.J., et al., *Plasma Lipid Profiling Shows Similar Associations with Prediabetes and Type 2 Diabetes.* PLoS ONE, 2013. **8 (9) (no pagination)**(e74341).
213. Peddinti, G., et al., *Early metabolic markers identify potential targets for the prevention of type 2 diabetes.* Diabetologia, 2017.
214. Suvitaival, T., et al., *Lipidome as a predictive tool in progression to type 2 diabetes in Finnish men.* Metabolism, 2017.
215. Yengo, L., et al., *Impact of statistical models on the prediction of type 2 diabetes using non-targeted metabolomics profiling.* 2016.
216. Panevska, A., et al., *Ceramide phosphoethanolamine, an enigmatic cellular membrane sphingolipid.* Biochim Biophys Acta Biomembr, 2019. **1861**(7): p. 1284-1292.
217. Fall, T., et al., *Non-targeted metabolomics combined with genetic analyses identifies bile acid synthesis and phospholipid metabolism as being associated with incident type 2 diabetes.* Diabetologia, 2016. **59**(10): p. 2114-24.
218. Gaudreau, P., et al., *Nutrition as a determinant of successful aging: description of the Quebec longitudinal study Nuage and results from cross-sectional pilot studies.* Rejuvenation Res, 2007. **10**(3): p. 377-86.
219. Blaak, E., *Gender differences in fat metabolism.* Curr Opin Clin Nutr Metab Care, 2001. **4**(6): p. 499-502.
220. Statistique Canada, [a.], *Proportion des personnes de 12 ans et plus ayant reçu un diagnostic d'hypertension, selon le groupe d'âge et selon le sexe, Québec, 2013-2014.*
221. Statistique Canada, [b.], *Tableau 13-10-0096-20 Indice de masse corporelle, embonpoint ou obèse, autodéclaré corrigé, adulte, selon le groupe d'âge (18 ans et plus).*
222. Statistique Canada, [c.], *Proportion des personnes de 12 ans et plus ayant reçu un diagnostic d'hypertension, selon le groupe d'âge et selon le sexe, Québec, 2013-2014.*
223. Tan, S., et al., *Traveller health: prevalence of diabetes, pre diabetes and the metabolic syndrome.* Ir Med J, 2009. **102**(6): p. 176-8.
224. Goecks, J., et al., *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.* Genome Biol, 2010. **11**(8): p. R86.
225. Ramautar, R., et al., *Human metabolomics: strategies to understand biology.* Curr Opin Chem Biol, 2013. **17**(5): p. 841-6.
226. Kuehnbaum, N.L. and P. Britz-McKibbin, *New advances in separation science for metabolomics: resolving chemical diversity in a post-genomic era.* Chem Rev, 2013. **113**(4): p. 2437-68.
227. Fuhrer, T. and N. Zamboni, *High-throughput discovery metabolomics.* Curr Opin Biotechnol, 2015. **31**: p. 73-8.
228. Kuhl, C., et al., *CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets.* Anal Chem, 2012. **84**(1): p. 283-9.
229. Alonso, A., et al., *AStream: an R package for annotating LC/MS metabolomic data.* Bioinformatics, 2011. **27**(9): p. 1339-40.
230. Senan, O., et al., *CliqueMS: A computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network.* Bioinformatics, 2019.
231. Uppal, K., D.I. Walker, and D.P. Jones, *xMSannotator: An R Package for Network-Based Annotation of High-Resolution Metabolomics Data.* Anal Chem, 2017. **89**(2): p. 1063-1067.
232. Carusi, A. and T. Reimer, *Virtual Research Environment Collaborative Landscape Study: A JISC funded project.* 2010.

233. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
234. Cottret, L., et al., *MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W132-7.
235. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.
236. Caspi, R., et al., *The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.* Nucleic Acids Res, 2016. **44**(D1): p. D471-80.
237. Schlegel, D.R., A. Ruttenberg, and E. P.L., *Ontologies in Metabolomics.* Metabolomics, 2015. **5**: p. e137.
238. Guyon, I. and A. Elisseeff, *An Introduction to Variable and Feature Selection.* Journal of Machine Learning Research 3, 2003.
239. Liu, H. and H. Motoda, *Feature selection for knowledge discovery and data mining*. 1998: The Springer International Series in Engineering and Computer Science.
240. Rinaudo, P., et al., *biosigner: A New Method for the Discovery of Significant Molecular Signatures from Omics Data.* Front Mol Biosci, 2016. **3**: p. 26.
241. Wold, S., M. Sjöström, and L. Eriksson, *PLS-regression: a basic tool of chemometrics.* Chemometrics and intelligent laboratory systems, 2001. **58**: p. 109-130.
242. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.
243. Guyon, I., et al., *Gene Selection for Cancer Classification usingSupport Vector Machines.* Machine Learning, 2002. **46**: p. 389-422.
244. Alcaraz, N., et al., *KeyPathwayMiner: Detecting Case-Specific Biological Pathways Using Expression Data.* Internet Mathematics, 2011. **7**(4): p. 299-313.
245. Brown, M., et al., *A metabolome pipeline: from concept to data to knowledge.* Metabolomics, 2005. **1**(1): p. 39-51.
246. Haznadar, M., et al., *Navigating the road ahead: addressing challenges for use of metabolomics in epidemiology studies.* Metabolomics, 2014. **10**(2): p. 176-178.
247. Dunn, W.B., et al., *Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry.* Nat Protoc, 2011. **6**(7): p. 1060-83.
248. Alyass, A., M. Turcotte, and D. Meyre, *From big data analysis to personalized medicine for all: challenges and opportunities.* BMC Med Genomics, 2015. **8**: p. 33.
249. Lubin, J.H. and M.H. Gail, *Biased selection of controls for case-control analyses of cohort studies.* Biometrics, 1984. **40**(1): p. 63-75.
250. Ford, E.S., C. Li, and N. Sattar, *Metabolic syndrome and incident diabetes: current state of the evidence.* Diabetes Care, 2008. **31**(9): p. 1898-904.
251. Fages, A., et al., *Investigating sources of variability in metabolomic data in the EPIC study: the Principal Component Partial R-square (PC-PR2) method.* Metabolomics, 2014. **10**(6): p. 1074-1083.
252. Sampson, J.N., et al., *Metabolomics in epidemiology: sources of variability in metabolite measurements and implications.* Cancer Epidemiol Biomarkers Prev, 2013. **22**(4): p. 631-40.
253. Blaise, B.J., et al., *Power Analysis and Sample Size Determination in Metabolic Phenotyping.* Anal Chem, 2016. **88**(10): p. 5179-88.
254. Trutschel, D., et al., *Experiment design beyond gut feeling: statistical tests and power to detect differential metabolites in mass spectrometry data.* Metabolomics, 2015. **11**(4): p. 851-860.
255. Broadhurst, D., et al., *Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies.* Metabolomics, 2018. **14**(6): p. 72.
256. Statistique Canada[a], *Proportion des personnes de 12 ans et plus ayant reçu un diagnostic d'hypertension, selon le groupe d'âge et selon le sexe, Québec, 2013-2014.*
257. Statistique Canada[b], *Tableau 13-10-0096-20 Indice de masse corporelle, embonpoint ou obèse, autodéclaré corrigé, adulte, selon le groupe d'âge (18 ans et plus).*

258. Statistique Canada[c], *Proportion des personnes de 12 ans et plus ayant reçu un diagnostic d'hypertension, selon le groupe d'âge et selon le sexe, Québec, 2013-2014.*

ANNEXE 2

Les variables métabolomiques ayant un score VIP > 1,25 (projection de l'importance de la variable) pour chacune des PLS réalisé sur les clusters 2 à 2 sont présentées ci-dessous.

Identifiant de la variable	Score VIP
M179T471_Hilic	1.66749
M101.0244T0.93_C18Neg	1.64768
M161T473_Hilic	1.64336
M240T472_Hilic	1.58304
M269T469_Hilic	1.55388
BV2.13535_RMN	1.43445
M269T461_Hilic	1.40974
BV2.4515_RMN	1.40787
BV5.23675_RMN	1.3511
M295T576_Hilic	1.33391
BV3.59445_RMN	1.31025
M178T555_Hilic	1.28206
BV3.3405_RMN	1.27813
M441.0748T0.92_C18Pos	1.2761
M203.0526T0.91_C18Pos	1.26311

A : PLS des clusters 1 et 2

Identifiant de la variable	Score VIP
M179T471_Hilic	1.88703
M161T473_Hilic	1.85053
M287.6321T14.67_1_C18Pos	1.7377
M240T472_Hilic	1.63569
M101.0244T0.93_C18Neg	1.60967
M146T532_Hilic	1.55995
M289.6415T15.32_C18Pos	1.53781
BV3.22775_RMN	1.484
BV0.85964_RMN	1.43052
M280.655T15.32_C18Pos	1.41835
M269T469_Hilic	1.41332
V6223_Lipido	1.29593
V5231_Lipido	1.29514
V5261_Lipido	1.2879
V5990_Lipido	1.27189

B : PLS des clusters 1 et 3

Identifiant de la variable	Score VIP
M203.0526T0.91_C18Pos	2.19575
M101.0244T0.93_C18Neg	2.13577
M269T461_Hilic	2.03789
M161T473_Hilic	2.01204
M179T471_Hilic	1.95331
M441.0748T0.92_C18Pos	1.93121
M442.0783T0.92_C18Pos	1.86854
M240T472_Hilic	1.81817
M268T461_Hilic	1.77519
BV5.23675_RMN	1.77357
M499.0334T0.92_C18Pos	1.75512
M501.0308T0.92_C18Pos	1.72294
BV0.85964_RMN	1.69649
M269T469_Hilic	1.69642
BV3.50275_RMN	1.61002
M163.06T0.91_1_C18Pos	1.60494
BV3.59445_RMN	1.60261
BV4.0018_RMN	1.60099
BV3.83745_RMN	1.5911
BV3.56309_RMN	1.57473
BV3.29175_RMN	1.52076
BV3.2057_RMN	1.4897
BV3.40944_RMN	1.4852
BV2.02749_RMN	1.47252
M305.918T0.9_C18Neg	1.46503
BV7.33019_RMN	1.45989
M216.0362T0.91_C18Neg	1.44962
M272.9913T0.91_C18Neg	1.42966
M556.9921T0.92_C18Pos	1.4262
BV2.13535_RMN	1.39683
BV3.22775_RMN	1.39592
M325T598_Hilic	1.3954
BV4.25985_RMN	1.38079
BV3.89425_RMN	1.35785
BV3.3405_RMN	1.3443
M295T576_Hilic	1.317
BV3.74559_RMN	1.30064
BV2.4515_RMN	1.29234
M215.0328T0.91_C18Neg	1.28572
M55.0175T0.92_C18Pos	1.27002

B : PLS des clusters 1 et 3