Table des Matières

Publications & Communications

(2015 - 2017)

Publications liées à la thèse

1. A Tebani , I Schmitz-Afonso, L Abily-Donval, B Heron, M Piraud, J Ausseil, F Zermiche, A Brassier, P De Lonlay, FM Vaz, BJ Gonzalez, S Marret, C Afonso, S Bekri. **Urinary metabolic phenotyping of mucopolysaccharidoses combining untargeted and targeted strategies with data modelling.** ***Submitted***

2. A Tebani, C Afonso, S Bekri. **Advances in metabolome information retrieval: turning chemistry into biology. Part I: Analytical chemistry of the metabolome.** *J Inherit Metab Dis*. 2017. ***Submitted***

3. A Tebani, C Afonso, S Bekri. **Advances in metabolome information retrieval: turning chemistry into biology. Part II: Biological information recovery.** *J Inherit Metab Dis*. 2017. ***Submitted***

4. Tebani A, Afonso C, Marret S, Bekri S. **Omics-Based Strategies in Precision Medicine: Toward a Paradigm Shift in Inborn Errors of Metabolism Investigations**. Int J Mol Sci. 2016 Sep 14;17(9).

5. Tebani A, Abily-Donval L, Afonso C, Marret S, Bekri S. **Clinical Metabolomics: The New Metabolic Window for Inborn Errors of Metabolism Investigations in the Post-Genomic Era**. *Int. J. Mol. Sci.* **2016**, *17*, 1167.

6. A Tebani, I Schmitz-Afonso, DN Rutledge, BJ Gonzalez, S Bekri, C Afonso. **Optimization of a Liquid Chromatography Ion Mobility-Mass Spectrometry method for untargeted metabolomics using experimental design and multivariate data analysis.** *913, **2016** Mar; 55-62*

Autres publications

1. P Gaildrat, S Lebbah, **A Tebani**, B Sudrié-Arnaud, I Tostivint, G Bollee, H Tubeuf, T Charles, A Bertholet-Thomas, A Goldenberg, F Barbey, A Martins, P Saugier-Veber, T Frébourg, B Knebelmann, S Bekri. **Clinical and molecular characterization of Cystinuria in a French cohort: relevance of assessing large-scale rearrangements and splicing variants. Mol Genet Genomic Med. 2017** May 16.

2. A Tebani, L Zanoutene-Cheriet, Z Adjtoutah, L Abily-Donval, C Brasse-Lagnel, S Marret, A Chalabi-Benabdellah, S Bekri. **Clinical and molecular characterization of patients with mucopolysacharidosis type I in an Algerian series**. Int. J. Mol. Sci. **2016**, 17, 743

3. F Marguet, H Barakizou, **A Tebani**, L Abily-Donval, S Torre, F Bayoudh, S Jebnoun, M Brasseur-Daudruy, S Marret, A Laquerriere, S Bekri. **Pyridoxine-dependent epilepsy: report on three families with neuropathology**. Metab Brain Dis. **2016** Jul 20.

4. J Lanzini, D Dargère, A Regazzetti, **A Tebani**, O Laprévote, N Auzeil. **Changing in lipid profile induced by the mutation of Foxn1 gene: A lipidomic analysis of Nude mice skin**. Biochimie. **2015** Nov;118:234-43

Communications orales

1. **A Tebani**. Metabolic phenotyping of mucopolysaccharidoses using untargeted liquid chromatography ion mobility mass spectrometry-based strategy. 13th Annual WORLDSymposium. San Diego. USA. February 13th - 17th, **2017.**

2. A Tebani. Stratégie d'optimisation par design expérimental des méthodes métabolomiques globales basées sur la spectrométrie de masse et la mobilité ionique. Ecole Thématique CNRS –17 -20 octobre 2016 – Cabourg. France.

3. **A Tebani.** Stratégie d'optimisation par design expérimental et analyses multivariées des méthodes métabolomiques globales basées sur la chromatographie liquide et la mobilité ionique couplées à la spectrométrie de masse. Waters MS Day. Paris. France. 4 octobre **2016.**

4. **A Tebani**. Mass spectrometry based metabolomics: a promising tool for the diagnosis of Inborn Errors of Metabolism. Journées Ecole Doctorale ED NBISE. Caen – France. 16 - 17 mars **2016.**

5. **A Tebani.** Metabolomics: the Clinical Chemistry of the 21st century. Sepcial seminar. Institute of Infectious Disease & Molecular Medicine. University of Cape Town. South Africa. February 16th, **2016.**

6. **A Tebani**. La spectrométrie de masse dans le diagnostic des Maladies Héréditaires du Métabolisme. 1ères Journées Scientifiques des Maladies Héréditaires du Métabolisme : Du diagnostic au traitement. Faculté de Médecine. Marrakech. Maroc. 14 au 15 janvier **2016.**

Communications affichées

1. **A Tebani**, L Abily-Donval, I Schmitz-Afonso, DN Rutledge, BJ Gonzalez, S Marret, C Afonso, S Bekri. Implementation of an untargeted liquid chromatography ion mobility-mass spectrometry-based metabolomics method for inherited metabolic diseases investigation.13th Annual WORLDSymposium. San Diego. USA. February 13 – 17, 2017.
2. **A Tebani**, L Abily-Donval, I Schmitz-Afonso, B Heron, M Piraud, J Ausseil, F Zermiche, A Brassier, P De Lonlay, T Levade, S Marret, C Afonso, S Bekri. Metabolic phenotyping of mucopolysaccharidoses using untargeted liquid chromatography ion mobility mass spectrometry-based strategy. 13th Annual WORLDSymposium. San Diego. USA. February 13 – 17, 2017.
3. **A Tebani**, L Abily-Donval, I Schmitz-Afonso, S Marret, C Afonso, S Bekri. Mass spectrometry based metabolomics as a promising tool for the diagnosis of Inborn Errors of Metabolism. Journée Normande de Recherche Biomédicale. Rouen. 16 septembre **2016.**
4. **A Tebani**, I Schmitz-Afonso, D N Rutledge, B J Gonzalez, S Bekri, C Afonso. A new optimization approach for liquid chromatography ion mobility-mass spectrometry untargeted metabolomics method using experimental design. Rome. Italie. SSIEM 6 - 9 Septembre **2016**
5. **A Tebani**, L Abily-Donval, I Schmitz-Afonso, S Marret, C Afonso, S Bekri. Mass spectrometry based metabolomics: a promising tool for the diagnosis of inborn errors of metabolism. Rome. Italie. SSIEM 6 - 9 Septembre **2016.**
6. **A Tebani**, L Abily-Donval, I Schmitz-Afonso, S Marret, C Afonso, S Bekri. L'approche métabolomique dans le diagnostic des Maladies Héréditaires du Métabolisme. Congrès de la Société Française de Pédiatrie. Lille. 20 au 20 mai **2016**.
7. **A Tebani**, I Zanoutene-Cheriet, Z Adjtoutah, I Abili-Donval, S Marret, A Chalabi Benabdellah, S Bekri. Caractérisation clinique et moléculaire de la mucopolysaccharidose de type I dans une cohorte algérienne. Congrès de la Société Française de Pédiatrie. Lille. 20 au 20 mai **2016**.
8. **A Tebani,** L Abily-Donval, I Schmitz-Afonso, S Marret, C Afonso, S Bekri. Mass spectrometry based metabolomics: a promising tool for the diagnosis of Inborn Errors of Metabolism. Journées Ecole Doctorale ED NBISE. Caen **2016.**
9. **A Tebani**, I Schmitz-Afonso, BJ Gonzalez, S Bekri, C Afonso**.** Optimization of a UHPLC-MS method for untargeted metabolomics using an experimental design approach and multivariate data analysis.9ème Journées Scientifiques du Réseau Francophone de Fluxomique et Métabolomique. Lille, France. **2015.**
10. **A Tebani**, I Schmitz-Afonso, BJ Gonzalez, S Bekri, C Afonso**.** Chemometric approaches for optimization of a high-resolution mass spectrometry based method for untargeted metabolomics. 4ème Journée de l'IRIB. Rouen, France, **2015.**

Liste des figures

Liste des principales abréviations

ADN	Acide désoxyribonucléique
ARN	Acide ribonucléique
ARNm	ARN messager
CCS	Collision Cross Section
DoE	Design of Experiments
DTIMS	Drift Tube Ion Mobility Spectrometry
Dt	Drift time
EIC	Extracted Ion Chromatogram
EIM	Erreur Innées du Métabolisme
ESI	ElectroSpray Ionisation
FAIMS	Field Asymmetric Waveforme Ion Mobility Spectrometry
FDR	Flase Discovery Rate
FT-ICR	Fourrier Transform-Ion Cyclotronic Resonance
FT-MS	Fourrier Transform-Mass Spectrometry
GC	Gas Chromatography
HDMS	High Definition Mass Spectrometry
HGMD	Human Gene Mutation Database
HMDB	Human Metabolome DataBase
IMS	Ion Mobility Spectrometry
LC	Liquid Chromatography
min	Minute
MLS	Maladies de Surcharge Lysosomale
MPS	Mucopolysaccharidoses
MS	Mass spectrometry
ms	Millisecond
MS/MS	Tandem Mass spectrometry
m/z	Rapport masse sur charge
NBS	NewBorn Screening
NMR	Nuclear Magnetic Resonnance
OMIM	Online Mendelian Inheritance in Man
OPLSDA	Orthogonal Partial Least Square Discriminant Analysis
PCA	Principal Component Analysis
PC	Principal Component
Q	Quadripole
QC	Quality Control
RE	Réticulum Endoplasmique
s	Second
TCSH	Transplantation de Cellules Souches Hématopoiétiques
TES	Thérapie Enzymatique Substitutive
TOF	Time of Flight
tR	Temps de rétention
TWIMS	Travelling Wave Ion Mobility Spectrometry
UHPLC	Ultra-High Pression Liquid Chromatography
VIP	Variable Influence on Projection

Partie I : Introduction Générale

Préambule

La médecine de précision (PM) est un nouveau paradigme qui est en train de révolutionner la pratique médicale actuelle et remodèle complètement la médecine de demain. La PM aspire à placer le patient au centre du parcours de soins en y intégrant les données médicales et biologiques individuelles tout en tenant compte de la grande diversité interindividuelle. Il est désormais bien établi que les interactions complexes entre les gènes et l'environnement façonnent les processus physiologiques et pathologiques aussi bien à l'échelle individuelle que populationnelle [1]. La prédiction des états physiologiques et pathologiques chez les patients nécessite une compréhension dynamique et systémique de ces interactions. La médecine systémique, fil conducteur de la médecine de précision, est un nouveau concept basé sur des approches globales pour le diagnostic et le suivi des maladies. L'idée de base de ces approches est qu'un système complexe est plus intelligible si on le considère dans sa globalité en tenant compte des échelles spatiales et temporelles [2]. Les erreurs innées du métabolisme (EIM) sont des troubles génétiques résultant de défauts dans une voie biochimique donnée en raison de la déficience ou de l'anomalie d'une enzyme, son cofacteur ou un transporteur. Ce déficit conduit à une accumulation du substrat ou de la diminution voire l'absence du produit en aval. Ces EIM sont un modèle approprié pour les études de médecine systémique parce que la base biologique sous-jacente de ces maladies a été, au moins en partie, largement investiguée. La première description de ces troubles a été faite par Sir Archibald Garrod, qui a initié le paradigme « un gène - une protéine - une maladie ». Cependant, il existe un manque manifeste de corrélation génotype-phénotype dans la plupart des EIM. En outre, la même variation génétique peut aboutir à différents phénotypes dans la même famille [3]. Ces observations remettent en question le paradigme de Garrod et suggèrent l'influence de facteurs génétiques ou environnementaux. Les EIM ne sont plus considérées comme des maladies monogéniques, ce qui ajoute une autre couche de complexité à leur caractérisation et à leur diagnostic. L'avènement des approches «omiques» offre une opportunité exceptionnelle pour fournir de nouveaux outils efficaces pour le dépistage, le diagnostic, le traitement et la surveillance de ces maladies. Les technologies omiques permettent des observations globales des molécules constitutives d'un système biologique. Elles visent essentiellement à extraire, d'une manière non ciblée, et sans hypothèse préalable, les informations biologiques exprimées par les gènes (génomique), l'ARNm (transcriptomique), les protéines (protéomique) et les métabolites (métabolomique). Ces stratégies contrastent clairement avec les approches conventionnelles intrinsèquement réductionnistes et principalement axées sur une hypothèse prédéfinie. Pour bien comprendre les processus pathologiques, une approche d'investigation globale doit être appliquée à de multiples niveaux de l'information biologique. Depuis les origines de la médecine, le corps humain a toujours été considéré comme une collection de composants distincts et indépendants, et ainsi, les médecins traitaient les maladies en essayant d'identifier une anomalie liée à une seule composante déficiente. Cette approche manque d'information contextuelle qui est vitale pour la compréhension mécanistique de la physiopathologie et par conséquent, pour la conception de stratégies thérapeutiques. En effet, la caractérisation complète d'un système biologique devrait inclure sa structure, son organisation et sa description fonctionnelle. La structure comprend les composants fondamentaux (gènes, protéines et métabolites). Le schéma d'organisation indique comment ces composants sont liés les uns aux autres et comment ils sont organisés topologiquement (par exemple, séquence linéaire ou ramifiée des réactions) et morphologiquement (lié ou pas à la membrane, compartimentation fonctionnelle). Et enfin, la fonction décrit comment l'ensemble du système se comporte dans l'espace et dans le temps en ce qui concerne les flux métaboliques et la réponse aux stimuli [4-6]. Deux avancées ont permis l'émergence de la biologie des systèmes : la génération de données biologiques à haut débit et leur modélisation. D'une part, la montée en puissance des technologies omiques à haut débit a permis de récupérer une information biologique globale. D'autre part, le développement des capacités informatiques a permis la modélisation des systèmes complexes et leur visualisation. Les classifications génétiques des maladies sont maintenant bien établies, étant donné les outils génomiques modernes qui peuvent fournir de riches informations sur les cohortes de patients. Cependant, d'autres approches hautement complémentaires basées sur l'information protéomique et métabolomique peuvent aider les investigateurs à contextualiser biochimiquement ou physiologiquement l'information génétique sous-jacente, aidant ainsi à mieux appréhender le phénotype et à permettre la stratification des patients. Les approches métabolomiques sont particulièrement pertinentes pour les EIM compte tenu de leur physiopathologie de base qui est étroitement liée au métabolisme. Ces maladies présentant des symptômes cliniques non spécifiques et des tests de laboratoire appropriés sont cruciaux pour établir leur diagnostic. Cependant, les procédures classiques de diagnostic biologique sont basées sur une série d'essais biochimiques séquentiels et segmentés sur différentes plateformes analytiques séparées. Cette approche est lente, fastidieuse et complexe, alors qu'une prise en charge optimale des patients nécessite une amélioration de la rapidité des rendus des examens biochimiques pour permettre un diagnostic précoce et une meilleure surveillance des EIM. Pour répondre à ce besoin de dépistage plus rapide, le profil métabolique peut être un candidat prometteur.

Les objectifs de ce travail de thèse sont définis comme suit :

1. Mise en place une méthodologie métabolomique non ciblée basée sur une stratégie analytique multidimensionnelle comportant la spectrométrie de masse à haute résolution couplée à la chromatographie liquide ultra-haute performance et la mobilité ionique.
2. Mise en place de la méthodologie de prétraitement, d'analyse et d'exploitation des données
3. Application à l'exploration des maladies héréditaires du métabolisme (Mucopolysaccharidoses)

Le manuscrit de cette thèse est structuré en trois parties, Introduction (**Partie I**), travail expérimental (**Partie II**) et conclusion et perpestives (**Partie III**).

L'introduction est subdivisée en cinq chapitres. Le chapitre I est une introduction générale aux maladies héréditaires du métabolisme. Le chapitre II reprend les éléments fondamentaux et clinico-biologiques des maladies de surcharge lysosomales. Le chapitre III présente la description des différentes mucopolysccharidoses, modèle d'étude EIM de cette thèse. Le chapitre IV est une revue systématique des différentes approches omiques et leur apport à la médecine systémique ainsi que leurs applications et les défis inhérents à leur implémentation dans le contexte clinique. Ce chapitre est basé sur une revue : **Article 1** (*Tebani A, Afonso C, Marret S, Bekri S. Omics-Based Strategies in Precision Medicine: Toward a Paradigm Shift in Inborn Errors of Metabolism Investigations. Int J Mol Sci. 2016 Sep 14;17(9)*). Le chapitre V présente une description des différentes stratégies métabolomiques, leurs avantages ainsi que leurs limites. Ce chapitre est basé sur deux revues : **Article 2** (*Tebani A, Afonso C, Bekri S. Advances in metabolome information retrieval : from chemistry to biology. Part I: Analytical Chemistry of the Metabolome. J Inherit Metab Dis. 2017.* ***Submitted****.*) et **Article 3** (*Tebani A, Afonso C, Bekri S. Advances in metabolome information retrieval: from chemistry to biology. Part II: Biological Information retrieval. J Inherit Metab Dis. 2017.* ***Submitted****.*). Le chapitre VI est une perspective sur le potentiel de la métabolomique dans le contexte clinique en général et dans l'exploration des EIM en particulier. Ce chapitre est basé sur une revue : **Article 4** (*Tebani A, Abily-Donval L, Afonso C, Marret S, Bekri S. Clinical Metabolomics: The New Metabolic Window for Inborn Errors of Metabolism Investigations in the Post-Genomic Era. Int. J. Mol. Sci. 2016, 17, 1167*).

La **partie II** du manuscrit correspond à la partie expérimentale et comporte deux chapitres. Dans le premier chapitre nous décrivons une nouvelle méthodologie d'optimisation des méthodes métabolomiques non ciblées par une approche multivariée en utilisant des plans d'expériences et la chimiométrie. Ce chapitre est basé sur l'**Article 5** (*A Tebani, I Schmitz-Afonso, DN Rutledge, BJ Gonzalez, S Bekri, C Afonso. Optimization of a Liquid Chromatography Ion Mobility-Mass Spectrometry method for untargeted metabolomics using experimental design and multivariate data analysis. 913, 2016 Mar ; 55-62).* Le chapitre II décrit l'application de la méthodologie dans l'exploration des mucopolysaccharidoses. Ce chapitre est basé sur l'**Article 6** (*A Tebani, I Schmitz-Afonso, L Abily-Donval, B Heron, M Piraud, J Ausseil, F Zermiche, A Brassier, P De Lonlay, FM Vaz, BJ Gonzalez, S Marret, C Afonso, S Bekri. Urinary metabolic phenotyping of mucopolysaccharidoses combining untargeted and targeted strategies with data modeling.* ***Submitted****).*

La partie III résume les principaux résultats présentés dans cette thèse et les diverses perspectives de recherches complémentaires notamment l'application des méthodes d'analyse de données proposées dans l'exploration des maladies héréditaires du métabolisme.

1. Barabasi, A.-L.; Gulbahce, N.; Loscalzo, J. Network medicine: A network-based approach to human disease. *Nat Rev Genet* **2011**, *12*, 56-68.
2. Hood, L.; Balling, R.; Auffray, C. Revolutionizing medicine in the 21st century through systems approaches. *Biotechnology Journal* **2012**, *7*, 992-1001.
3. Lanpher, B.; Brunetti-Pierri, N.; Lee, B. Inborn errors of metabolism: The flux from mendelian to complex diseases. *Nat Rev Genet* **2006**, *7*, 449-460.
4. Aon, M.A. Complex systems biology of networks: The riddle and the challenge. In *Systems biology of metabolic and signaling networks*, Springer: 2014; pp 19-35.
5. Kitano, H. Computational systems biology. *Nature* **2002**, *420*, 206-210.
6. Aon, M.A.; Cortassa, S. Systems biology of the fluxome. *Processes* **2015**, *3*, 607-618.

CHAPITRE I : LES MALADIES HEREDITAIRES DU METABOLISME

1. Définition et classification

Le métabolisme est un réseau complexe, interconnecté et finement régulé. Il est composé de réactions biochimiques qui transforment des substrats, d'origine endogène ou exogène, en produits indispensables à l'intégrité fonctionnelle de la cellule, du tissu et de l'organisme. De ce fait, une dérégulation de cette homéostasie sous-tend les mécanismes physiopathologiques de différentes maladies [1]. Une altération d'une voie métabolique peut être liée à des facteurs nutritionnels, environnementaux ou génétiques. Les troubles métaboliques qui ont une base génétique sont appelés erreurs innées du métabolisme (EIM) ou maladies héréditaires du métabolisme (MHM). Les EIM représentent un groupe d'environ 500 maladies rares avec une incidence globale d'environs 1/2500. Elles sont principalement dues à un défaut génétique d'enzymes ou de cofacteurs participant à une voie métabolique ou encore au transport intra- ou intercellulaire de métabolites. Les EIM ont été décrits pour la première fois par Sir Archibald Garrod en 1908, au cours des Croonian lectures on the inborn errors of metabolism. Il y décrit l'alkaptonuira, la cystinurie, l'albinisme et la pentosurie [2,3]. Depuis, le domaine des EIM n'a cessé d'évoluer à la fois en termes de diagnostic de nouveaux phénotypes, mais aussi bien dans leurs stratégies thérapeutiques. Les EIM peuvent se manifester à n'importe quel âge [4]. Les symptômes qui en résultent apparaissent pendant la période anténatale et néonatale. Certaines formes sont tardives et peuvent apparaître dans l'enfance et même à l'âge adulte [5]. Dans les maladies tardives, les symptômes sont souvent chroniques et progressifs. L'un des principaux problèmes dans la compréhension des troubles métaboliques héréditaires est leur grande variabilité métabolique et hétérogénéité clinique. Les EIM sont individuellement rares, mais elles forment, collectivement, un groupe important de maladies en particulier dans le contexte pédiatrique. De nombreux types d'EIM peuvent donner lieu à des symptômes similaires et compliquer ainsi la démarche diagnostique. Les phénotypes cliniques d'une même maladie peuvent également varier. Les troubles métaboliques sont classés en fonction de leur présentation principale aiguë, intermédiaire ou chronique [4]. Cette approche s'avère également appropriée pour orienter les stratégies thérapeutiques. Trois groupes sont, ainsi, définis :

Troubles du type « intoxication » : une substance toxique est présente en excès comme dans les troubles du cycle de l'urée, aminoacidopathies ou aciduries organiques. Ces maladies présentent souvent un intervalle libre sans symptômes qui est suivi de symptômes de type intoxication. Ces symptômes peuvent être aigus ou chroniques, mais souvent intermittents et provoqués par un agent intercurrent ou des changements alimentaires. Le traitement, s'il est disponible, repose souvent sur une thérapie nutritionnelle qui permet d'utiliser des voies alternes et d'éviter la voie métabolique altérée.

Troubles du métabolisme énergétique : Les présentations cliniques peuvent être aiguës ou chroniques ou à rechutes. Toutes les voies du métabilsme ébnargétiques peuvent être concernées ; ces déficits peuvent être mitochondriaux (troubles de la chaîne respiratoire, cycle de Krebs et oxydation du pyruvate, défauts d'oxydation des acides gras et corps cétoniques), ou bien affectant les processus énergétiques cytoplasmiques (par exemple, la gluconéogenèse et la glycolyse). Dans ces troubles, les possibilités de traitement dépendent de la nature du défaut de production d'énergie. La plupart des défauts mitochondriaux sont, à ce jour, impossibles à traiter, alors que les défauts de la gluconéogenèse et de la glycolyse sont souvent moins sévères et traitables.

Troubles impliquant les organelles intracellulaires et la surcharge : ces troubles concernent la synthèse ou la dégradation des molécules complexes tels que les maladies lysosomales de surcharge (MLS), les maladies de peroxysomes et les défauts de synthèse de cholestérol. Ces troubles ont une évolution chronique et progressive et ne sont pas liés à l'alimentation ou d'autres facteurs intercurrents. Auparavant, ces troubles étaient longtemps dépourvus de traitements, mais actuellement, des stratégies thérapeutiques sont disponibles pour certaines MLS telles que la maladie de Fabry, Gaucher et Pompe et mucopolysaccharidoses I, II, IVA et VI [6].

2. Stratégies thérapeutiques

Deux stratégies thérapeutiques principales ont été développées pour traiter les EIM. La première est basée sur la réduction de la concentration du substrat accumulé soit en favorisant son élimination, soit par la réduction de sa production. La seconde est une stratégie de complémentation de la protéine déficiente comprenant l'allo-transplantation de cellules souches hématopoïétiques (TCSH), la thérapie enzymatique substitutive (TES), la thérapie pharmacologique de chaperon, la thérapie génique et la thérapie cellulaire. Certaines de ces approches peuvent être combinées pour améliorer leur efficacité.

La **Figure 1** présente le schéma général des bases biochimiques des EIM et leurs stratégies thérapeutiques.

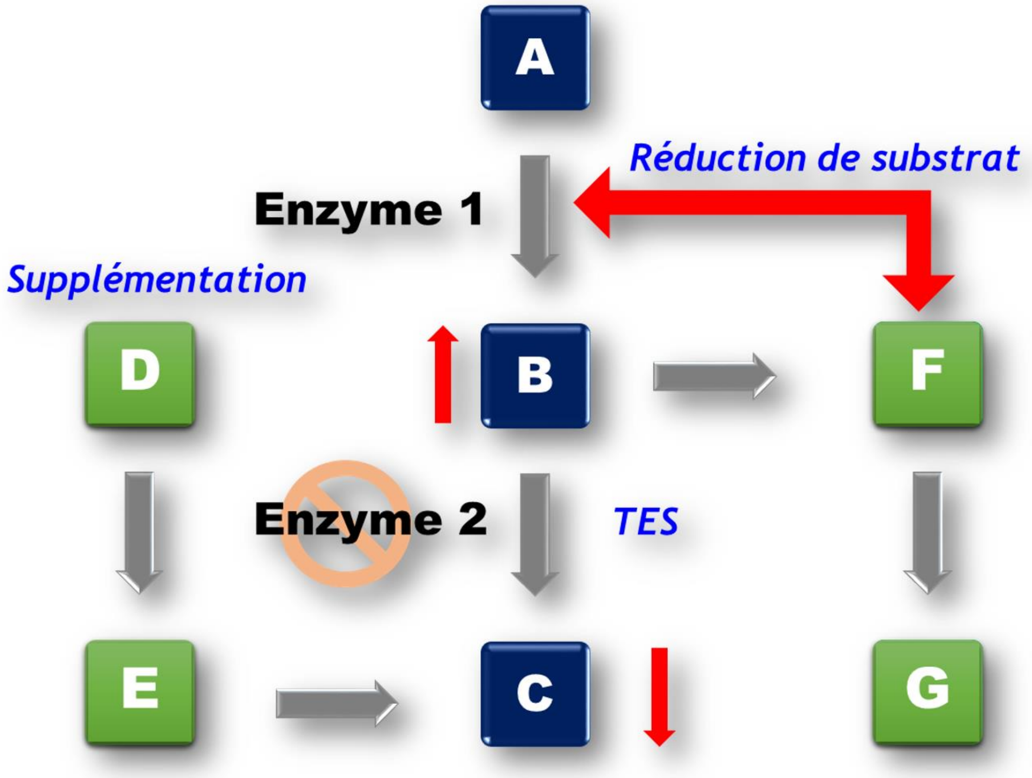


Figure 1. Bases biochimiques générales des Maladies Héréditaires du Métabolisme. En considérant une voie métabolique simple : substrat A est transformé en produit B via l'enzyme 1 qui est ensuite transformé en produit C via l'enzyme 2. Un déficit en enzyme 2 conduit à une augmentation du substrat B et carence en produit B. Trois stratégies thérapeutiques majeures en découlent : réduire le substrat B par des voies métaboliques alternatives ou en bloquant l'activité de l'enzyme qui le produit. Supplémentation directe ou indirecte du produit C. Remplacement de l'enzyme déficiente par voie pharmacologique ou thérapie enzymatique substitutive (TES).

2.1. Stratégie substrat-dépendante

2.1.1. Limitation des apports du substrat accumulé

Une restriction de l'apport exogène du substrat accumulé permet de réduire sa toxicité (diminution des apports en phénylalanine dans la phénylcétonurie).

2.1.2. Epuration du substrat accumulé

Cette stratégie peut être illustrée par l'utilisation d'épurateurs dans les déficits du cycle de l'urée pour épurer l'ammoniaque accumulé. Diverses molécules sont utilisées telles que le benzoate de sodium ou le phénylbutyrate de sodium. Récemment, des cyclodextrines ont été utilisées pour solubiliser des composés lipidiques qui s'accumulent dans les endosomes tardifs dans les cellules Niemann-Pick C. En effet, la maladie de Niemann-Pick C se caractérise par une exportation lysosomale défectueuse du cholestérol et d'autres lipides, entraînant une détérioration neurologique progressive. Les cyclodextrines sont des oligosaccharides cycliques hydrophiles qui contiennent une poche hydrophobe dans laquelle le cholestérol est séquestré [7].

2.1.3. Traitement par réduction de la production du substrat

Cette stratégie vise à diminuer la production du substrat accumulé et à empêcher, ainsi, son stockage. La réduction du substrat peut reposer sur l'inhibition des enzymes en amont de l’accumulation. En effet, l’inhibition de la biosynthèse des substrats permet de rétablir l'équilibre entre la synthèse et la dégradation par les enzymes lysosomales [8]. Le premier inhibiteur qui a été commercialisé est le Miglustat pour la maladie de Gaucher type I (type non-neurologique) [9]. Miglustat est une petite molécule qui inhibe la première enzyme (Glucosylcéramide synthase) de la voie de biosynthèse des gangliosides. Cette approche peut être utile pour d’autres maladies de surcharge lysosomale partageant la même voie de biosynthèse. Ainsi, le même inhibiteur de substrat pourrait être utilisé pour un groupe de maladie tel que le Miglustat dans la maladie de Tay-Sachs [10] et Niemann-Pick type C [11] Sandhoff [12]. Récemment, une nouvelle molécule, Eliglustat, a été décrite pour la maladie de Gaucher type 1 [13].

2.2. Stratégie de complémentation

2.2.1. Transplantation de cellules souches hématopoïétiques (TCSH)

La TCSH consiste à administrer des cellules souches hématopoïétiques (moelle osseuse ou sang de cordon ombilical) par voie intraveineuse à un patient préalablement soumis à une myéloablation. Les cellules injectées deviennent une source de production de l’enzyme manquante non seulement au niveau de la moelle osseuse, mais également au niveau du système nerveux central par le biais des cellules microgliales [14,15]. La transplantation de cellules souches hématopoïétiques a été utilisée pour traiter des sphingolipidoses telles que la maladie de Krabbe et les mucopolysaccharidoses tel que la MPS I. Les cellules souches hématopoïétiques peuvent permettre la complémentation de fonctions lysosomales déficientes et certaines cellules générées migrent vers le SNC [16]. Par ailleurs, la TCSH peut être associée à une enzymothérapie substitutive débutée dès le diagnostic, avant la greffe et poursuivie jusqu’à récupération d’une activité enzymatique satisfaisante.

2.2.2. Thérapie enzymatique substitutive

Elle consiste à remplacer les enzymes manquantes ou défectueuses par des enzymes exogènes administrées par voie intraveineuse. Ces enzymes sont internalisées par les cellules et adressées aux lysosomes pour exercer leur activité catabolique [6,17]. Le rationnel de la thérapie enzymatique substitutive (TES) repose sur l'administration systémique de protéines lysosomales synthétiques. Dans la cellule, l'enzyme nouvellement synthétisée est adressée à l'endosome tardif par liaison de ses résidus mannose-6-phosphate (M6P) au récepteur du M6P dans l'appareil de Golgi. Le récepteur M6P est recyclé soit vers l'appareil de Golgi, soit vers la membrane plasmique. Le principe de la TES repose sur l'internalisation de l'enzyme exogène à l'aide du récepteur M6P situé sur la membrane plasmique. Il est à noter que le résidu mannose-6-phosphate n'est pas présent dans l'α-glucosidase endogène (maladie de Gaucher). L'α-glucosidase synthétisée à des fins ERT est modifiée et présente avec un résidu Man-6-P qui est utilisé pour cibler cette enzyme vers l'endosome précoce [18]. L'efficacité de la TES est variable en fonction des organes ciblés et est limitée par l'accès de l'enzyme à certains tissus atteints (SNC, os). Divers événements indésirables ont été signalés, y compris les problèmes d'immunogénicité. Il est à noter la TES est administré par voie intraveineuse puisque ces protéines sont susceptibles d'être dégradées dans le tractus gastro-intestinal [19].

2.2.3. Supplémentation en cofacteurs

Parfois, la supplémentation en cofacteur d'une enzyme peut permettre d'utiliser de façon optimale l'enzyme résiduelle et obtenir, ainsi, une activité enzymatique suffisante (par exemple la forme B12-sensible de l'acidémie méthylmalonique).

2.2.4. Chaperonnes pharmacologiques

Certains variants de gènes peuvent conduire à la synthèse de protéines mal repliées dans le réticulum endoplasmique. Ces molécules instables sont dégradées par le protéasome (ERAD - Entedplasmic Reticulum Associated Degradation). Les chaperonnes pharmacologiques agissent comme des inhibiteurs de l'enzyme cible. Ces analogues du substrat enzymatique ont une forte affinité pour le site actif. Les chaperonnes pharmacologiques se lient et stabilisent l'enzyme mal repliée et permettent, ainsi, son internalisation au lysosome. La molécule chaperonne est libérée de l'enzyme dans le lysosome en raison du pH acide et de la différence de rapport de concentration entre le substrat non dégradé et la chaperonne pharmacologique [20].

2.2.5. Reprise traductionnelle en aval d'un codon stop prématuré

La présence d'un codon stop prématuré abouti putativement à une protéine tronquée qui sera dégradée. Certaines molécules permettent de stabiliser le ribosome et d'incorporer un acide aminé en position du codon stop et ainsi la poursuite de la traduction jusqu'au codon stop naturel. La Gentamicine [21] et le PTC124 (Ataluren) [22] ont été décrits pour la mucoviscidose.

2.2.6. Inhibiteurs de protéases

L'homéostasie des protéines se réfère au contrôle de la concentration de protéines, de leur conformation et de leurs interactions souvent par des modifications transcriptionnelles et/ou traductionnelles. Par conséquent, pour maintenir cette homéostasie, différents mécanismes impliquant des réponses cellulaires compartimentées, tel que comme la réponse aux heat-shock protéines dans le cytoplasme et la réponse des protéines mal repliées dans le réticulum endoplasmique. Ces mécanismes sont nécessaires pour limiter les défauts de repliement et l'agrégation des protéines. Les maladies dues aux anomalies conformationnelles des protéines sont généralement caractérisées par une induction inefficace de ces réponses. L'utilisation de régulateurs de cette homéostasie semble être une stratégie thérapeutique prometteuse [23]. Deux types de régulateurs sont décrits, inhibiteurs du protéasome et régulateurs du calcium [9]. Le fait que les régulateurs de la protéostasie agissent sur l'homéostasie des protéines plutôt que directement sur une protéine déterminée rend leur spectre d'action potentiel plus large. En outre, comme décrit pour BCM-95 et hydroxypropyl-β-cyclodextrine [24], ces régulateurs sont susceptibles d'agir de manière synergique [25,26].

2.2.7. Inducteurs de l'autophagie

L'autophagie est la principale voie catabolique pour les agrégats protéiques et les organites endommagés, et son activation ou son inhibition a été liée à diverses maladies telle que le cancer [27], les maladies neurodégénératives [28] et le vieillissement [29]. De ce fait, plusieurs stratégies pharmacologiques ont été développées pour améliorer l'activité de cette voie dans le contexte pathologique [30]. Plusieurs études ont démontré que l'autophagie est altérée dans la plupart des maladies lysosomales de surcharge [31,32]. Différents composés ont été décrits comme des inducteurs d'autophagie, tels que la génistéine et les hydroxypropyl-β-cyclodextrines, dont le dernier est efficace dans un modèle de lipofushcinose infantile tardive [33] et ainsi que le Niemann-Pick type C [34]. Il est à noter que la plupart des inducteurs autophagiques agissent en inhibant la voie mTOR. L'induction autophagique par l'inhibition de la voie mTOR a été liée à plusieurs effets négatifs limitant la dose et a incité à découvrir de nouveaux modulateurs de l'autophagie indépendants de mTOR [35].

CHAPITRE II : LES MALADIES DE SURCHARGE LYSOSOMALE

Le lysosome, décrit par De Duve et al. en 1955, est un organite cellulaire cytoplasmique présente dans toutes les cellules eucaryotes nucléées [36]. Son fonctionnement normal est important pour la dégradation des macromolécules et l'homéostasie cellulaire. Cet organite a également un rôle dans les processus de phagocytose et de présentation des antigènes, qui sont nécessaires pour la régulation de l'inflammation et le contrôle de l'auto-immunité. Le système lysosome-endosome est intimement impliqué dans la régulation de l'autophagie, de l'apoptose et de la mort cellulaire par transduction du signal et des facteurs d'exocytose impliqués dans l'inflammation, l'oncogenèse et la maladie neurodégénérative ainsi que dans le recyclage des récepteurs pour la régulation de la neurotransmission et dans la pigmentation de la peau [37,38]. Lorsque l'un des composants du lysosome (enzymes, protéines membranaires, cofacteurs, etc.) est déficient, cette défaillance peut provoquer l'apparition de processus pathologiques. Cliniquement, ces maladies ont commencé à être reconnues au cours du XIXe siècle par une série de médecins dont les noms de famille nous sont encore familiers, puisqu'ils ont été utilisés plus tard pour nommer les troubles que chacun a décrits pour la première fois tel que Bernard Sachs [39], Philippe Gaucher [40] ou encore Johannes Fabry [41]. Les maladies lysosomale de surcharge (MLS) représentent un groupe d'environ 50 maladies héréditaires dues à des déficits de protéines lysosomales. Cela conduit à une accumulation progressive de composés dans le lysosome. Ce stockage de molécules non dégradées provoque diverses défaillances d'organes et une mort cellulaire prématurée [42]. Dans ce chapitre sont présentées la physiologie lysosomale et les MLS.

1. Le système lysosomal : physiologie et pathologie

Le lysosome est un organite délimité par une membrane lipidique monocouche et présente un pH acide interne de 4,5 à 5,0. Ce pH est maintenu par une pompe à protons dépendant du triphosphate d'adénosine (ATP). Morphologiquement, c'est un organite très hétérogène. La taille, la forme et le nombre de lysosomes par cellule sont variables selon les types cellulaires. Il peut être sphérique, ovoïde, ou parfois de forme tubulaire, et peut varier entre 0,1 et 2 µm de taille. Les lysosomes sont le centre de dégradation de la cellule et sont principalement responsables de la dégradation des protéines, des polysaccharides, glycosaminoglycanes et des lipides complexes en leurs molécules élémentaires respectivement : les acides aminés, les monosaccharides et les acides gras libres. Les lysosomes contiennent plus de 60 types différents d'hydrolases indispensables pour la fonction catabolique lysosomale [43]. La batterie d’enzymes hydrolytiques localisées dans lumière lysosomale est optimalement active à un pH acide et a la capacité de dégrader la plupart des macromolécules comprenant les protéines, les glucides, les lipides, l'ARN et l'ADN.

Les produits de dégradation lysosomaux sont transportés à l'extérieur du lysosome via des transporteurs spécifiques localisés dans la membrane lysosomale [44] ou via le trafic membranaire vésiculaire pour la réutilisation dans les voies biosynthétiques [45]. Le lysosome constitue une importante plaque tournante métabolique et joue un rôle régulateur majeur dans les cellules eucaryotes. Le lysosome fait partie du système endosome-lysosome qui intervient dans l'import et l'export de molécules de et vers la cellule, du déplacement de la cellule et de la dégradation des molécules exogènes et endogènes (autophagie) [43]. D'autres systèmes fonctionnent en parallèle du système lysosomal comme les mécanismes protéolytiques tels que le système ubiquitine-protéasome (UPS), qui contribue au turnover efficace des protéines. Le coordinateur central de ce réseau intracellulaire est finalement le lysosome lui-même, organite acide lié à la membrane qui fonctionne pour dégrader et retraiter une large gamme de molécules cellulaires. Bien que traditionnellement représenté comme un compartiment terminal, son rôle dans le recyclage des précurseurs moléculaires confirme l'importance du cycle lysosomal. La dégradation catabolique médiée par le lysosome est un processus adaptatif régulé par le statut nutritionnel et la signalisation cellulaire [37]. Les lysosomes reçoivent les composants extracellulaires ou de surface cellulaire par endocytose et reçoivent des composants intracellulaires par autophagie [43]. Considéré dans son ensemble, le système lysosomal fonctionne donc au centre même de l'homéostasie métabolique cellulaire. Avec la découverte d'un réseau de régulation globale de gènes appelé CLEAR (Coordinated Lysosomal Expression and Regulation) et son facteur de transcription EB (TFEB), de nombreux composants du système lysosomal ont été décrits être liés au niveau transcriptionnel [46]. En effet, ces études établissent encore le système lysosomal comme un réseau hautement efficace et coordonné. Ainsi, une fonction lysosomale adéquate est essentielle, car le déficit de ce système entraîne, inexorablement, des conséquences délétères pour les cellules, les organes et l'organisme, avec près de 60 types différents de maladies lysosomales documentés à ce jour [47]. L'accumulation lysosomale active une variété de cascades pathogéniques qui aboutissent à des phénotypes complexes caractérisées par une expression clinique multi-systémique [48-51]. Actuellement, et même si certains modèles ont été suggérés, nous manquons encore d'une image claire des événements moléculaires et cellulaires pertinents reliant les variants pathogènes aux symptômes de la maladie qui sont déterminés par des mécanismes fonctionnant non seulement au niveau cellulaire, mais aussi dans les tissus et organes. L'un des plus grands défis est de comprendre comment le l'accumulation de substrat affecte la fonction des cellules, des tissus et des organes, provoquant *in fine* la pathogenèse de la maladie. La **Figure 2** présente le modèle proposé pour les mécanismes pathogéniques mis en jeu dans MLS. La **Figure 3** présente l'exemple de la pathogénie des MLS et les mécanismes secondaires conduisant à la mort cellulaire dans les neurones.

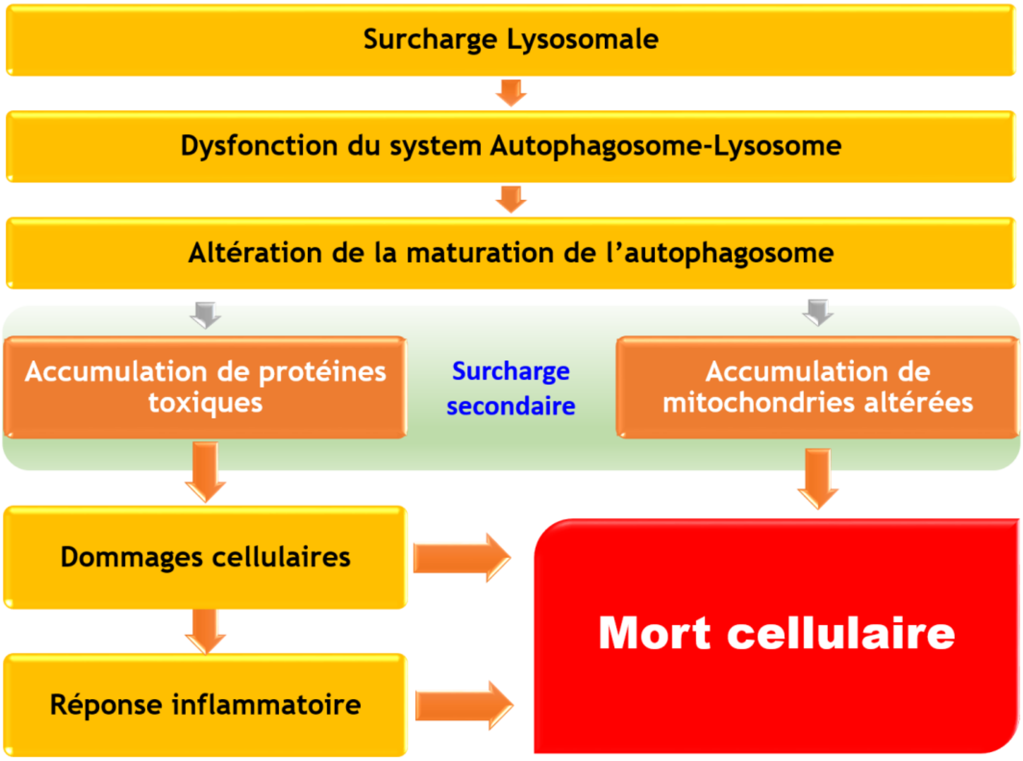

Surcharge Lysosomale
Dysfonction du system Autophagosome-Lysosome
Altération de la maturation de l'autophagosome
Accumulation de protéines toxiques
Surcharge secondaire
Accumulation de mitochondries altérées
Dommages cellulaires
Mort cellulaire
Réponse inflammatoire

Figure 2. Modèle proposé pour la pathogenèse des MLS. Le stockage des lysosomes conduit à une capacité réduite des lysosomes à fusionner avec les autophagosomes. Il en résulte un blocage (au moins partiel) de la maturation autophagique et de la dégradation défectueuse. En conséquence, les substrats autophagiques tels que les agrégats de protéines polyubiquitinés et les mitochondries non fonctionnelles s'accumulent et favorisent la mort cellulaire. La réponse inflammatoire aux lésions cellulaires contribue encore à la mort cellulaire.

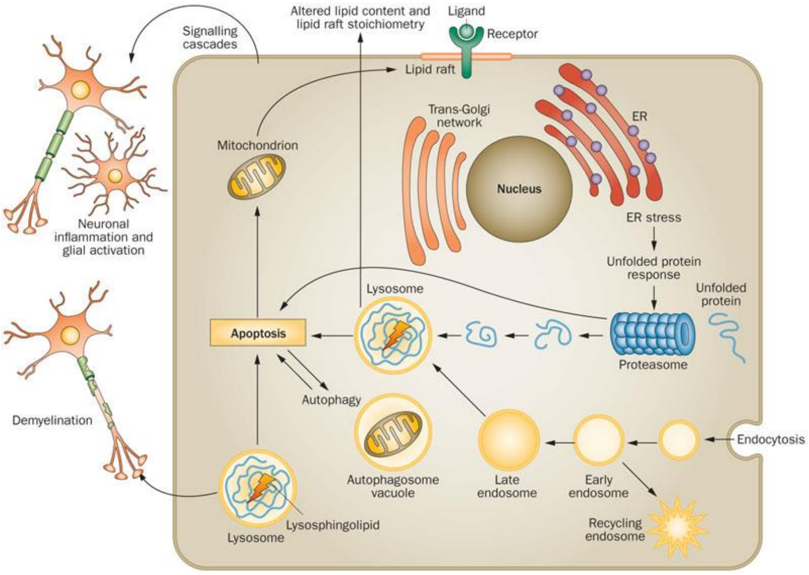


Figure 3. Les MLS et mécanismes secondaires conduisant à la mort cellulaire dans les neurones. Le stockage lysosomal dans les neurones est accompagné par l'amplification des processus cellulaires avec des effets négatifs sur les neurones, la glie, les cellules de Schwann et le cerveau. Ces processus incluent l'activation de voies de transduction de signal médiées par un récepteur ligand, l'altération du trafic de lipides et la teneur en stéroïde, une augmentation du stress ER et une réponse protéique dépliée médiée par le calcium, le stockage de lysosphingolipides et l'activation gliale avec neuroinflammation qui conduit à une démyélinisation. Ces processus aboutissent collectivement à une apoptose accélérée et à une dérégulation de l'autophagie. Abréviation : ER, réticulum endoplasmique. D'après [38].

2. Maladies du lysosomales de surcharge : définition et classification

2.1. Définition

Le concept de MLS a été développé en 1963, suite à la découverte que la maladie de Pompe qui est causée par un déficit en α-glucosidase, une enzyme lysosomale qui décompose l'amidon en glucose [49]. De nombreux déficits de protéines membranaires et lysosomales ont été décrits. Désormais, les MLS représentent un groupe d'environ 50 maladies héréditaires dues à des déficits de protéines lysosomales. Cela conduit à une accumulation progressive de composés dans le lysosome. La plupart sont transmises selon un mode autosomique récessif, bien que certaines soient liées au chromosome X (Hunter, Fabry, Danon) [38].

2.2. Classification

Les MLS sont le fréquemment classées selon le principal composé de surcharge donc non dégradé. Cliniquement, cette classification est très utile et bien acceptée. Ainsi, les troubles dans lesquels prévaut l'accumulation de fragments de glycosaminoglycanes sont classés comme des mucopolysaccharidoses, ceux qui sont dominés par la surcharge des lipides sont nommés des lipidoses. L'accumulation de sphingolipides prédomine dans les sphingolipidoses et oligosaccharidoses sont dues à l'accumulation d'oligosaccharides [38]. Il faut toutefois souligner que, dans la plupart des maladies lysosomales, plus d'un composé s'accumule et que dans certains troubles, pour diverses raisons, le substrat non dégradé peut être assez hétérogène. Ainsi, un certain nombre de glycosidases lysosomales ne sont pas spécifiques d'un substrat unique, mais plutôt d'un résidu de sucre et de la stéréochimie de sa liaison. Ce résidu et cette liaison peuvent se produire dans les glycosaminoglycanes ainsi que dans les lipides, de sorte qu'une déficience de l'enzyme entraîne une accumulation des deux. Le Tableau 1 présente les différentes MLS, la protéine déficiente et le substrat accumulé.

2.3. Aspects cliniques et diagnostic clinique

En général, les MLS sont des troubles chroniques et progressifs, qui peuvent présenter un large spectre de gravité et de symptômes. Les MLS sont multisystémiques, affectant divers organes, y compris le système squelettique, les muscles, le foie, la rate, le cœur, les poumons et, surtout, le système nerveux central (SNC) qui est fortement affecté dans les formes les plus sévères. En fait, les MLS sont la cause la plus fréquente des maladies neurodégénératives pédiatriques [49]. Les traitements des MLS offrent de meilleurs résultats chez les patients qui en bénéficient au début de la maladie avant le développement de séquelles irréversibles. En conséquence, les cliniciens sont amenés à faire un diagnostic rapide pour permettre à leurs patients de bénéficier des effets d'une thérapeutique précoce. En l'absence du dépistage néonatal pour le diagnostic précoce des MLS, le clinicien joue un rôle majeur dans l'orientation diagnostique. Chez certains patients, la présentation peut être anténatale avec une anasarque, alors que dans d'autres cas, avec le même déficit enzymatique avec les mêmes variant, l'apparition des signes cliniques peut être à un âge plus tardif. Pour de nombreux patients, l'apparition des symptômes peut être dans les premiers mois ou années de vie. Les premiers signes peuvent être un ralentissement du développement ou d'autres anomalies neurologiques. Chez d'autres patients, une organomégalie ou des traits grossiers peuvent être présents. La reconnaissance de ces signes cliniques facilitera le choix des tests de diagnostic les plus appropriés (**Figure 3**). Il est à noter que le diagnostic à l'âge adulte des MLS est de plus en plus observé. Dans ce groupe, la présentation peut être atypique et les manifestations neuropsychiatriques en l'absence de caractéristiques dysmorphiques sont beaucoup plus fréquentes chez ces patients [52-54].

2.4. Diagnostic biologique

La caractéristique biochimique commune des MLS est l'accumulation de substrats partiellement ou non dégradés dans le lysosome. Plusieurs mécanismes peuvent entraver le catabolisme normal des molécules dans le lysosome, ou la sortie des molécules dégradées de l'organite : i) **les défauts de dégradation** qui peuvent concerner plusieurs substrats tels que des glycosaminoglycanes, des lipides ou des protéines; ii) **le défaut de transport** à travers la membrane lysosomale; ou iii) **le défaut de trafic endosome-lysosome** [51]. En fait, toute interruption de la fonction lysosomale peut conduire à l'accumulation de substrat(s) non dégradé(s) dans les endosomes et les lysosomes, compromettant finalement la fonction cellulaire [43,50]. Le diagnostic biochimique et génétique des MLS doit être effectué dans des laboratoires spécialisés. Différents échantillons biologiques peuvent être utilisés pour l'analyse, tels que le sang, l'urine, le liquide amniotique, les fibroblastes cutanés et les biopsies tissulaires [55]. La mesure de la concentration du substrat accumulé est souvent la première approche lorsqu'une MLS est suspectée ou dans des programmes de dépistage systématique. Des modifications secondaires spécifiques peuvent également se produire dans les cellules associées à un déficit d'une enzyme lysosomale. De tels changements peuvent entraîner la modification d'autres protéines ou composants cellulaires qui peuvent être utiles comme marqueurs de maladies spécifiques ou pour suivre la progression de la maladie. L'identification de la base moléculaire d'un trouble peut permettre l'utilisation de tests plus spécifiques, tels que l'évaluation des enzymes lysosomales et des analyses moléculaires. Récemment, l'introduction de prélèvements de sang ou d'urine séchée sur du papier-filtre facilite le prélèvement, le transport et le stockage. Par ailleurs, l'analyse de l'ADN sur sang séché est possible est son utilisation est en cours d'extension [56]. Les tests effectués pour le diagnostic peuvent être divisés en trois approches et sont décrits ci-dessous.

Table 1. Maladies lysosomales de surcharges.

Maladie	Neur	Enzyme ou protéine déficiente
Sphingolipidoses		
Fabry disease	Y	α-Galactosidase A
Farber lipogranulomatosis	N	Ceramidase
Gaucher disease type I	N	β-Glucosidase
Gaucher disease types II and III	Y	Saposin-C activator
Niemann–Pick disease types A and B	Y	Sphingomyelinase
GM1-gangliosidosis: infantile, juvenile and adult variants	Y	β-Galactosidase
GM2-gangliosidosis (Sandhoff): infantile and juvenile	Y	β-Hexosaminidase A and B
GM2-gangliosidosis (Tay–Sachs): infantile, juvenile and adult variants	Y	β-Hexosaminidase A
GM2-gangliosidosis (GM2-activator deficiency)	Y	GM2-activator protein
GM3-gangliosidosis	Y	GM3 synthase
Metachromatic leukodystrophy (late infantile, juvenile and adult)	Y	Arylsulphatase A
Sphingolipid-activator deficiency	Y	Sphingolipid activator
Mucopolysaccharidoses		
MPS I (Scheie, Hurler–Scheie and Hurler disease)	Y	α-Iduronidase
MPS II (Hunter)	Y	Iduronidase-2-sulphatase
MPS IIIA (Sanfilippo A)	Y	Heparan N-sulphatase (sulphamidase)
MPS IIIB (Sanfilippo B)	Y	N-acetyl-α-glucosaminidase
MPS IIIC (Sanfilippo C)	Y	Acetyl-CoA; α-glucosamide N-acetyltransferase
MPS IIID (Sanfilippo D)	Y	N-acetylglucosamine-6-sulphatase
MPS IVA (Morquio syndrome A)	Y	N-acetylgalactosamine-6-sulphate sulphatase
MPS IVB (Morquio syndrome B)	N	β-Galactosidase
MPS VI (Maroteaux–Lamy)	Y	N-acetylgalactosamine-4-sulphatase (arylsulphatase B)
MPS VII (Sly disease)	Y	β-Glucuronidase
MPS IX	Y	Hyaluronidase
Glycogen storage disease		
Pompe (glycogen storage disease type II)	Y	α-Glucosidase
Oligosaccharidoses		
α-Mannosidosis	Y	α-Mannosidase
β-Mannosidosis	Y	β-Mannosidase
Fucosidosis	Y	α-Fucosidase
Aspartylglucosaminuria	Y	Aspartylglucosaminidase
Schindler disease	Y	α-N-acetylgalactosaminidase
Sialidosis	Y	α-Neuraminidase
Galactosialidosis	Y	Lysosomal protective protein
Mucolipidosis II (I-cell disease); mucolipidosis III	Y	Urine diphosphate-N-acetylglucosamine; lysosomal enzyme N-acetylglucosaminyl-1-phosphotransferase
Integral membrane protein disorders		
Cystinosis	N	Cystinosin
Danon disease	Y	Lysosome-associated membrane protein 2
Action myoclonus–renal failure syndrome	N	Lysosome membrane protein 2
Salla disease	Y	Sialin
Niemann–Pick disease type C1	Y	NPC-1 , NPC-2
Mucolipidosis IV	Y	Mucolipin
Additional disease types		
Multiple sulphatase deficiency	Y	Sulphatase-modifying factor 2
Niemann–Pick disease type C2	Y	NPC-2
Wolman disease (infantile); cholesteryl ester storage disease	N	Lysosomal acid lipase
Galactosialidosis	Y	Cathepsin A

Abbreviations:
MPS, mucopolysaccharidosis; NPC-1, Niemann–Pick disease type C1 protein; NPC-2, Niemann–Pick disease type C2 protein.
Neur: Neurological involvlement. N : No Y : Yes
D'après [38]

2.4.1. Dosage des substrats

a. Oligosaccharides

Le dosage des oligosaccharides urinaires est réalisé usuellement par chromatographie en couche mince haute performance (HPTLC), en utilisant la technique décrite pour la première fois par Humbel et collègues en 1975 [57]. Cette méthode, cependant, présente de beaucoup limites. La quantification absolue des métabolites n'est pas possible et l'identification des métabolites est approximative, car uniquement basée sur l'évaluation de la migration des métabolites. En outre, plusieurs médicaments ou situations physiologiques peuvent produire un profil anormal rendant l'interprétation des résultats difficile. Les exemples de troubles qui présentent un profil anormal sont l'aspartylglucosaminurie, la gangliosidose GM1, la gangliosidose GM2, la sialidose, la galactosialidose, la fucosidose, l'α-mannosidose et la maladie de stockage de glycogène de type II. Actuellement les techniques de chromatographie couplées à la spectrométrie de masse en tandem sont de plus adoptées compte tenu de leur sensibilité et spécificité (identification positive des métabolites) rendant les techniques chromatographiques sur couche mince obsolètes [58-61].

b. Glycosaminoglycanes urinaires

Le diagnostic des Mucopolysaccharidoses (MPS) peut être orienté par l'analyse des glycosaminoglycanes urinaires (GAG). La technique usuelle consiste à purifier les GAG par précipitation répétée avec du chlorure de cétylpyridinium. Puis, les GAG sont isolés par centrifugation et quantifiés par un dosage colorimétrique, souvent en utilisant de l'harmine [62,63]. Cette méthode évalue la quantité d'acide hexuronique contenue dans les GAG extraits et peut donner des résultats faussement négatifs lorsque le GAG accumulé est du sulfate de kératine, qui contient du galactose au lieu de l'acide hexuronique. Le deuxième piège lors de la quantification des GAG est lié à la diminution progressive des concentrations de GAG urinaires qui se produit avec l'âge chez les patients MPS [64]. Pour permettre une identification des GAG accumulés, une électrophorèse uni- ou bidimensionnelle est utilisée pour séparer les principales classes de GAG : sulfate de chondroïtine (CS), sulfate de kératane (KS), sulfate de dermatane (DS) et sulfate d'héparane (HS) [65]. Dans la situation physiologique, seul le sulfate de chondroïtine est détectable, sauf dans l'urine des nouveau-nés, où une fine bande correspondant au sulfate d'héparane peut être présente. L'analyse des GAG urinaires peut être affectée par des facteurs nutritionnels et environnementaux. Compte tenu de ces limites analytiques, de nouvelles méthodes basées sur la chromatographie liquide couplée à la spectrométrie de masse en tandem (LC-MS/MS) sont actuellement préférées pour le dosage de GAG [66-76].

c. Dosage de substrats spécifiques

Des métabolites, autres que ceux accumulés suite au déficit primaire, peuvent être informatifs tels que le lysoGb3 dans la maladie de Fabry. Le développement de la spectrométrie de masse en tandem pour l'identification et la quantification des substrats lysosomaux et des métabolites a apporté un progrès significatif dans le diagnostic des LSD [60,67,76]. Pour exemple, une étude chez 47 patients a permis d'analyser les glycosphingolipides et les oligosaccharides par spectrométrie de masse et a montré des profils différents correspondant à 12 maladies [77].

2.4.2. Etude fonctionnelle

a. Mesure de l'activité enzymatique

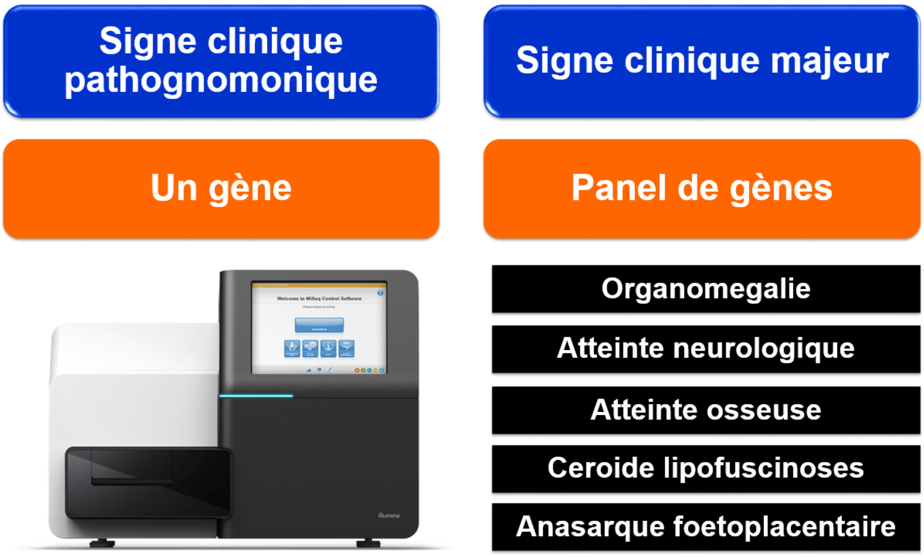
Actuellement, plusieurs activités enzymatiques peuvent être mesurées. Ce dosage se faisait habituellement sur culot leucytaire ou fibroblastes en utilisant des techniques fluorimétriques ou radio-immunologiques. Récemment, des méthodes à l'aide de taches de sang séchées sur papier buvard ont été développées soit en fluorimétrie [78] soit en spectrométrie de masse [79]. Les techniques fluorimétrique, le procédé est automatisé en utilisant les systèmes microfluidiques avec des microvolumes [80]. Ce procédé a été appliqué au dépistage multiplex du nouveau-né pour les maladies de Fabry, Gaucher, MPS I et II et Pompe [81]. L'autre alternative est d'utiliser des méthodes basées sur la détection par spectrométrie de masse en tandem. De nouveaux substrats enzymatiques et des standards internes ont été synthétisés avec des masses mutuellement exclusives pour le dosage des enzymes lysosomales par chromatographie liquide couplée à la spectrométrie de masse en tandem. Cette technique est très sensible et permet le dosage simultané de plusieurs enzymes à partir d'une tache de sang séchées [79,82]. Diverses institutions ont adopté cette technique pour le diagnostic des MLS à l'instar du Laboratoire de Biochimie Métabolique du CHU de Rouen et ce depuis 2012 [83-88].

b. Dosage des protéines

La protéomique est une autre stratégie alternative pour le dépistage préliminaire des MLS. La mesure des changements dans la concentration des enzymes lysosomales et des protéines secondaires au déficit primaire a longtemps été utilisée pour dépister des maladies spécifiques ou comme aide au diagnostic comme la chitotriosidase et la maladie de Gaucher. L'immuno-détection des changements de concentration d'un panel de protéines lysosomales, incluant les saposines et les protéines lysosomales associées à la membrane (LAMPs), a été la base d'une méthode de dépistage des maladies lysosomales chez les nouveau-nés [89]. La détermination quantitative d'un panel de peptides tryptiques marqueurs de protéines lysosomales par LC-MS/MS serait potentiellement une méthode sensible et rapide pour le criblage d'une maladie individuelle ou un sous-groupe ou toutes les MLS [90].

2.4.3. Analyse moléculaire

La plupart, voir la totalité, des gènes liés à des pathologies lysosomales ont été clonés et des tests génétiques moléculaires sont utilisés en routine pour confirmer un diagnostic biochimique [91]. L'analyse moléculaire est également utilisée pour clarifier l'effet d'une pseudodéficience d'une enzyme lysosomale ou pour établir un diagnostic pour les troubles pour lesquels il n'existe pas un test biochimique simple. Ainsi, l'identification des variants dans le cas index permet d'effectuer des tests précis, rapides et fiables auprès d'autres membres de la famille, y compris le diagnostic prénatal. Par ailleurs, les hétérozygotes ne peuvent être détectés de manière fiable que par des techniques moléculaires, ce qui est particulièrement important pour les troubles liés au X. Avec l'avènement du séquençage haut débit (Next Generation Sequencing NGS), le défi réside dans l'interprétation des variants détectés pour permettre de ne signaler que les variants pathogènes. La signification pathologique d'un nouveau variant doit être étudiée avant son application au diagnostic ou au conseil génétique. Ceci est particulièrement important dans le contexte du dépistage néonatal, qui révèle de nombreux variants qui conduisent à l'apparition tardive de signes cliniques ou de variants bénins. Les tests génétiques moléculaires n'ont pas été largement utilisés comme test diagnostique de première intention pour les MLS, mais cela peut changer avec l'avènement de méthodes de haut débit qui sont rapides, fiables et abordables telles que le séquençage de l'ADN à haut débit (Next Generation Sequencing) [92,93]. Bien qu'avec une couverture incomplète et une efficacité variable, cette technologie permet le séquençage de l'ensemble du génome (tous les gènes) ou de l'exome (tous les exomes), ou des panels de gènes sélectionnés. Le séquençage complet de l'exome (Whole Genome Sequencing WES) implique la capture sélective de l'exome, son amplification et séquençage, puis l'identification des variants potentiellement pathogènes à l'aide d'algorithmes bioinformatiques [94]. Bien que le WES n'identifie pas tous les variants pour diverses raisons techniques [95], il a été appliqué avec succès à la résolution de plusieurs problèmes de diagnostic dans les MLS. La première application de WES aux MLS était la détection d'une délétion homozygote à 6 paires de bases (6pb) dans le gène *GNPTG* dans une famille consanguine présentant une dysplasie spondylo-épiphysaire et une rétinite pigmentaire, une combinaison des symptômes habituellement associés à MLIIIγ [96]. L'utilisation du WES pour établir des diagnostics dans des cas atypiques élargira le spectre clinique des MLS. En plus d'élargir le spectre clinique d'une maladie, le WES facilitera l'identification de nouveaux gènes associés à un phénotype clinique [92]. Les variants pathogènes des gènes autres que ceux impliqués dans la fonction lysosomale seront identifiés par WES et/ou le séquençage complet du génome (WGS). Ceux-ci peuvent être sans rapport avec le phénotype clinique à l'étude et peuvent prédire la sensibilité à une autre affection, ou, ils peuvent être des modificateurs du phénotype exploré. Un exemple de ce dernier est la découverte de la base génétique possible du phénotype de la maladie de Gaucher et l'hémopathie maligne [97]. Il est à noter que la survenue de découvertes accidentelles dans les explorations NGS soulève aussi des questions éthiques, cliniques et pratiques considérables [93]. Le séquençage de panels de gènes sélectionnés peut éviter certains de ces problèmes, aux prix d'une couverture partielle et au risque de perdre des informations importantes. Actuellement, les centres de diagnostic développent des panels de gènes lysosomaux et apparentés pour le dépistage moléculaire. Les aspects techniques, cliniques et éthiques sont revus dans l'article I intitulé « *Omics-Based Strategies in Precision Medicine: Toward a Paradigm Shift in Inborn Errors of Metabolism Investigations* » (Tebani *et* al. Int J Mol Sci 2016. 17(9)). Par ailleurs, une proposition d'une nouvelle stratégie de l'exploration des maladies héréditaires du métabolisme dont font partie les MLS y est présentée [93]. Le laboratoire de Biochimie Métabolique du CHU de Rouen a mis en place une méthode NGS comprenant 52 gènes lysosomaux. Les figures 4A et 4B présentent la stratégie NGS adoptée par le laboratoire de Biochimie métabolique du CHU de Rouen pour le diagnostic des MLS.



Signe clinique pathognomonique

Signe clinique majeur

Un gène

Panel de gènes

Organomegalie

Atteinte neurologique

Atteinte osseuse

Ceroide lipofuscinoses

Anasarque foetoplacentaire

Figure 4A. Stratégie d'implémentation du panel NGS pour le diagnostic des MLS.

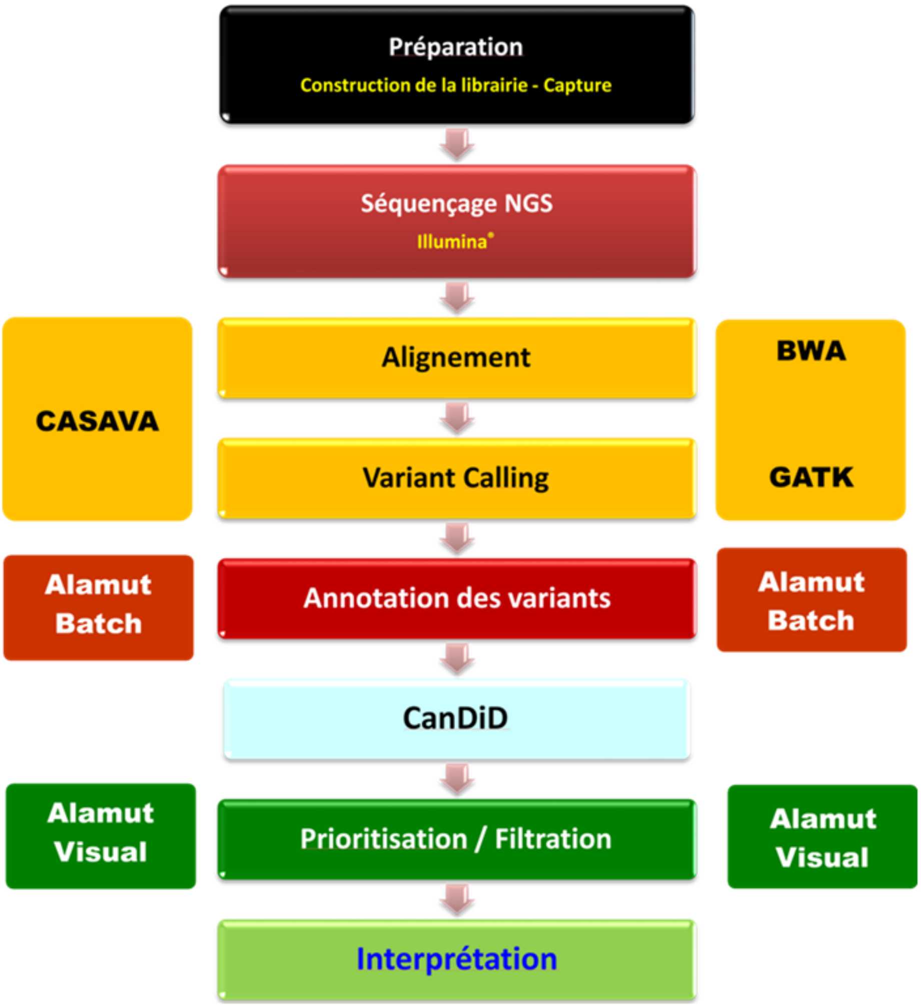


Figure 4B. Pipeline bio-informatique d'analyse NGS implémenté au niveau de la plateforme Génomique du CHU de Rouen.

2.4.4. Diagnostic prénatal

L'analyse moléculaire est devenue la méthode préférée pour le diagnostic prénatal. Cette analyse peut se faire soit un prélèvement de trophoblaste ou sur liquide amniotique. Le diagnostic génétique préimplantatoire (DPI) a été réalisé chez des couples à risque de plusieurs MLS, dont les maladies de Tay-Sachs, Gaucher, Morquio A, Sandhoff, MPS I et Niemann-Pick A/B. Les ovocytes ou les blastomères uniques d'un embryon à six cellules sont analysés pour des variants spécifiques connus [98]. Le DPI est une option attrayante pour certaines familles, car des tests peuvent être effectués sur des ovocytes avant la fécondation in vitro. La détection de variants connues dans l'ADN fœtal dans le sang maternel peut constituer une méthode non invasive de diagnostic prénatal à l'avenir [99].

CHAPITRE III : LES MUCOPOLYSACCHARIDOSES

1. Introduction

Les mucopolysaccharidoses (MPS) sont un groupe MLS causées par un déficit en enzymes catalysant la dégradation progressive des glycosaminoglycanes (GAG) incluant des GAG sulfatés, dermatane sulfate (DS), héparane sulfate (HS), kératane sulfate (KS), chondroïtine sulfate (CS) et des GAG non sulfatés, le hyaluronane [100,101]. Dans les MPS, les GAG non dégradées sont stockés dans les lysosomes et la matrice extracellulaire (MEC) dans divers tissus, puis sécrétés dans la circulation sanguine, et excrétés dans l'urine. Les GAG accumulés conduisent à un dysfonctionnement cellulaire et aboutissent à une structure anormale de la MEC, causant des dommages progressifs multi-tissus et multi-organes [101]. Les GAG sont des composants importants de la matrice extracellulaire (MEC) et se trouvent dans divers tissus [100,102]. Ce sont des polymères liés par covalence à des protéines pour produire des protéoglycanes (PG) ou restent comme des polysaccharides libres. Les PG sont associées à diverses fonctions physiologiques telles que l'hydratation et la régulation de la formation de fibrilles de collagène, et le site anionique principal est responsable de la sélectivité de charge dans les glomérules filtration. Les GAG ont aussi un rôle mécanique pour absorber les forces de compression exercées sur les tissus [103]. Les profils de sulfatation des GAG définissent leur rôle physiologique. Les GAG sont constituées de modules disaccharidiques répétés contenant des fractions amino-osidiques acétylées (N-acétyl-galactosamine ou N-acétyl-glucosamine) et principalement de l'acide uronique (acide D-glucoronique ou acide L-iduronique). Le KS comporte des disaccharides répétés contenant du galactose (4N-acétyl-glucosamine-β1, 3-galactose-β1). Les GAG sulfatés (CS, DS et HS) sont liés à leurs copules protéiques respectives via des résidus de serine. Le KS peut être synthétisé sur des oligosaccharides N-liés et sur des oligosaccharides liés à l'oxygène de serine/thréonine, qui sont distincts des régions de liaison des autres GAG. Le HS contient également des résidus de N-sulfate [100,102]. Les protéines n'ont pas seulement un rôle structural pour les GAG, elles peuvent avoir diverses activités biologiques telles que le repliement des glycoprotéines et leur dégradation et la reconnaissance immunologique [103]. A l'exception de l'acide hyaluronique, les GAG sont les produits de dégradation des protéoglycanes qui existent dans la MEC et sont clivés par protéolyse, donnant naissance à des GAG qui rejoignent le lysosome pour la dégradation intracellulaire. Selon la molécule à dégrader, il existe quatre voies différentes de dégradation lysosomale des GAG : DS, HS, KS et CS. La dégradation progressive des GAG nécessite 11 enzymes différentes : cinq glycosidases, cinq sulfatases et une transferase. Les carences de chacune de ces enzymes ont été rapportées et se traduisent par onze MPS différentes (Cf. Table 2). La **Figure 5** présente les structures des différents GAG leurs enzymes de dégradation avec les MPS correspondantes.

Onze déficits enzymatiques ont été décrits, résultant en sept formes distinctes de MPS [42]. L'incidence globale des MPS est supérieure à 1 : 25.000 à 1 : 30.000 naissances vivantes. La transmission génétique de ces maladies est autosomique récessive à l'exception de la MPS II qui est liée à l'X. En période néonatale, la plupart des patients atteints de MPS sont asymptomatiques. Le phénotype des MPS est hétérogène et son évolution est lente. Les manifestations cliniques commune aux MPS sont musculo-squelettiques: déformations de la colonne vertébrale, de la cage thoracique, des hanches, des genoux, du crâne et/ou des mains. Une petite taille, anomalies et douleurs articulaires [47]. Plusieurs stratégies thérapeutiques sont utilisées en clinique ou en cours de développement pour certains types de MPS. L'espérance de vie moyenne chez les patients MPS non traités est d'une à deux décennies.

Transplantation de cellules souches hématopoïétiques (TCSH)

Ce traitement est indiqué pour la MPS I de forme sévère (maladie de Hurler). Pour les autres types de MPS I, MPS II, MPS IV voire MPS VI, les indications peuvent se discuter au cas par cas au sein du CETMPS (Comité d'Evaluation Thérapeutique des Mucopolysaccharidoses). Par ailleurs, la TCSH peut être associée à une enzymothérapie substitutive débutée dès le diagnostic, avant la greffe et poursuivie jusqu'à récupération d'une activité enzymatique satisfaisante

Thérapie enzymatique substitutive (TES)

Quatre MPS bénéficient déjà de cette stratégie thérapeutique : MPS I, MPS II, MPS IV et MPS VI.

Thérapie de réduction de substrat (TRS)

Le flavonoïde Génistéine a été proposé comme traitement pour la MPS III [104].

Table 2. Liste des mucopolysaccharidoses.

Pathologie		OMIM	Gene	Protéine déficiente	ENZYME COMMISSIO N	Phénotype biochimique
MPS I	Syndrome de Hurler Syndrome de Scheie	607014 607015 607016	*IDUA*	α-L-Iduronidase	EC 3.2.1.76	Dermatane sulfate Héparane sulfate
MPS II	Syndrome de Hunter	309900	*IDS*	Iduronate-2-sulfatase	EC 3.1.6.13	Dermatane sulfate Héparane sulfate
MPS III A	Syndrome de Sanfilippo A	252900	*SGSH*	Héparane-N-sulfatase	EC 3.10.1.1	Héparane sulfate
MPS III B	Syndrome de Sanfilippo B	252920	*NAGLU*	α-N-acétylglucosaminidase	EC 3.2.1.50	Héparane sulfate
MPS III C	Syndrome de Sanfilippo C	252930	*HGSNAT*	AcétylCoA :α-glucosaminide N-acétyltransférase	EC 2.3.1.3	Héparane sulfate
MPS III D	Syndrome de Sanfilippo D	252940	*GNS*	N-acétylglucosamine-6-sulfatase	EC 3.1.6.14	Héparane sulfate
MPS IV A	Syndrome de Morquio A	253000	*GALNS*	N-acétylgalactosamine-6-sulfatase	EC 3.1.6.14	Héparane sulfate
MPS IV B	Syndrome de Morquio B	253010	*GLB1*	β-galactosidase	EC 3.2.1.23	Kératane sulfate Chondroïtine 6-sulfate
MPS VI	Syndrome de Maroteaux-Lamy	253200	*ARSB*	N-acétylgalactosamine-4-sulfatase	EC 3.1.6.12	Dermatane sulfate
MPS VII	Syndrome de Sly	253220	*GUSB*	β-glucuronidase	EC 3.2.1.31	Dermatane sulfate Chondroïtine 4,6-sulfates Héparane sulfate
MPS IX		601492	*HYAL1*	Hyaluronidase	EC 3.2.1.35	Hyalurane

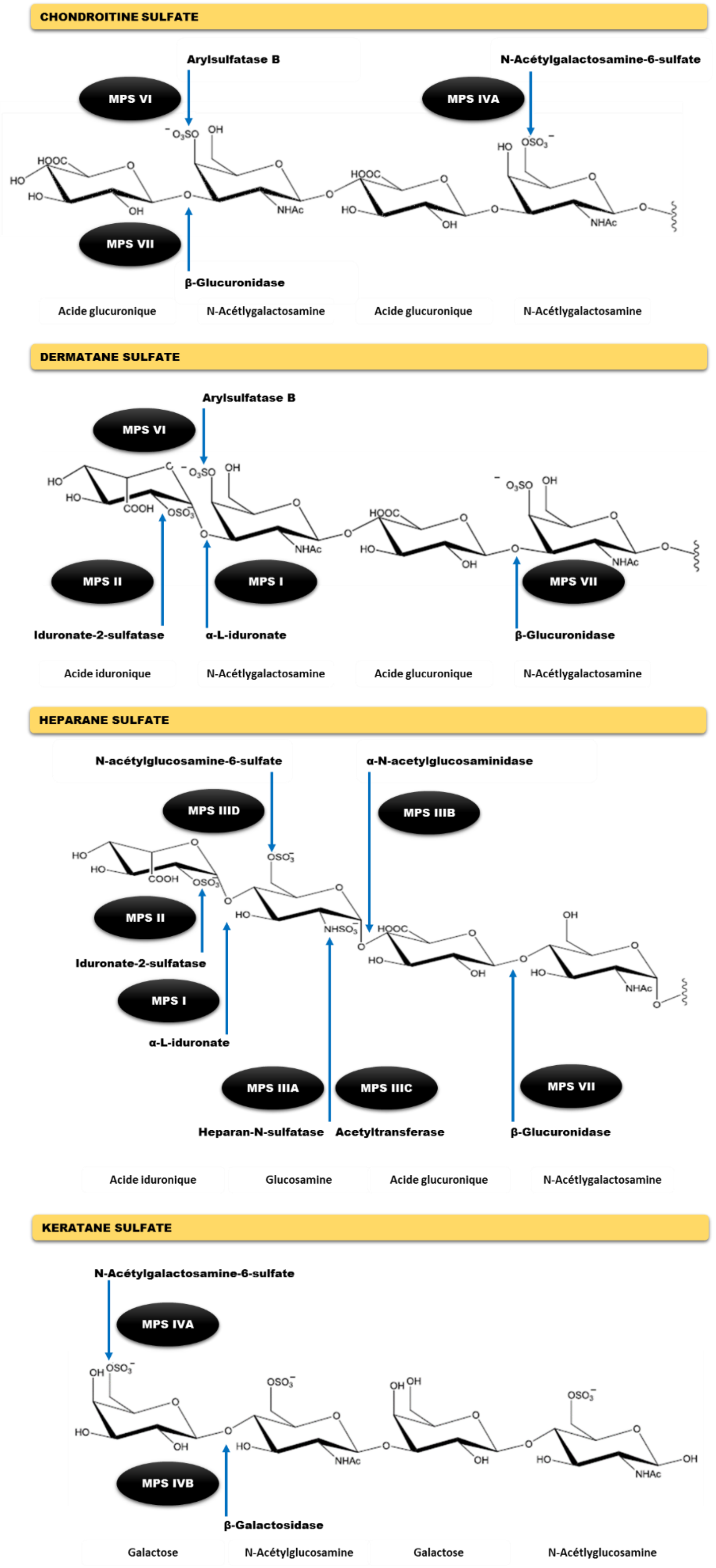


Figure 5. Structure et sites de clivage enzymatique des GAG. Les différentes MPS sont annotées.

2. MPS I

2.1. Définition

La MPS I a historiquement été divisée en trois phénotypes en fonction du type, la gravité et la progression des symptômes. Le syndrome de Hurler (OMIM : 607014) est le terme pour la forme la plus sévère et prend son nom de Gertrude Hurler qui a décrit un garçon et une fille atteints de MPS I en 1919 [105]. En 1962, le Dr Scheie, un ophtalmologiste, a décrit des patients avec opacification de la cornée qui ont été légèrement touchés et ont été diagnostiqués comme atteints de syndrome de Scheie (OMIM : 607016) [106]. Le syndrome de Scheie a été initialement considéré comme une forme différente du Hurler. Le déficit enzymatique a été découvert en 1971 et a été clairement établi que les syndromes de Scheie et de Hurler avaient le même déficit enzymatique [107]. Plus tard, un certain nombre de patients ont été décrits dont la maladie était de sévérité intermédiaire, et qui ne s'intègrent pas clairement dans Hurler ou Scheie, ceux-ci ont été classés historiquement comme des patients Hurler-Scheie (OMIM : 607015). Basée sur la compréhension actuelle de l'enzyme et de son gène, la MPS I comprend un large spectre de sévérité et que les individus peuvent être classés dans un continuum phénotypique allant de la forme la plus grave à une forme atténuée [5]. Les classifications Hurler, Hurler-Scheie et syndrome de Scheie sont connues pour être des simplifications qui ne reflètent pas de façon adéquate l’hétérogénéité clinique et la progression de la maladie.

2.2. Épidémiologie

L'incidence estimée est d'environ 1 cas pour 100 000 naissances.

2.3. Clinique

Le déficit en alpha-L-iduronidase (IDUA) peut entraîner un large éventail d'atteintes phénotypiques conduisant à trois grandes entités cliniques reconnues : les syndromes Hurler (MPS IH), Scheie (MPS IS; 607016) et Hurler-Scheie (MPS IH / S; 607015). Les syndromes de Hurler et de Scheie représentent des phénotypes extrêmes sévère et atténué du spectre clinique de MPS I, respectivement. Les patients IH ne présentent pas habituellement de signes à la naissance, le phénotype apparait progressivement dès les 1ers mois de vie. L’un âge de diagnostic est entre 4 et 18 mois devant des signes morphologiques, viscéraux et orthopédiques (hernies inguinales/ombilicales, cyphose, enraidissement articulaire, macrocéphalie, infections ORL récidivantes ou encombrement respiratoire chronique). Le phénotype s’élargit progressivement pour arriver à une atteinte multiviscérale et neurodégénérative. Pour le type IS : les signes apparaissent après 5 ans avec un diagnostic fait le plus souvent tardivement dans la 2ème décennie devant un enraidissement articulaire invalidant, des opacités cornéennes, une surdité, parfois une atteinte valvulaire cardiaque et une compression médullaire. Pour le type I-HS, qui est de gravité intermédiaire, les patients sont possiblement atteints de déficience intellectuelle modérée avec des manifestations physiques, osseuses et viscérales intermédiaires entre le types IH et IS.

2.4. Enzyme déficiente et substrat accumulé

- L'enzyme déficitaire est l'alpha-L-iduronidase (IDUA).
- Substrat accumulé : DS et HS.

2.5. Génétique

La MPS I se transmet selon un mode autosomique récessif et affecte les deux sexes. Le gène *IDUA* est localisé sur le chromosome 4p16.3 avec une taille de 19kb et comporte 14 exons. Un intron de 566 pb sépare les deux premiers exons. Un grand intron d'environ 13 kb suit. Les 12 derniers exons sont regroupés dans 4,5 kb. Le gène *IDUA* code pour une protéine précurseur de 653 acides aminés, qui est glycosylée et ensuite transformée en sa forme mature. Jusqu'à présent, 201 variants ont été rapportés dans la base HGMD publique [108]. La plupart des variants sont privés et quatre sont spécifiques à certaines populations (p.W402X, p.Q70X, p.P533R, p.G51D).

2.6. Traitement

- TES : Laronidase® (Sanofi-Genzyme)
- TCSH (Forme Hurler)

3. MPS II

3.1. Définition

La première description d'une MPS a été faite par Charles Hunter, un médecin canadien qui, en 1917, a décrit une maladie rare chez 2 frères [109]. La cause biochimique du syndrome de Hunter (OMIM : 309900) est une déficience de l'activité de l'enzyme lysosomale, l'iduronate-2-sulfatase (IDS) qui catalyse l'élimination du groupe sulfate en position 2 de l'acide L-iduronique dans les DS et HS. Le syndrome de Hunter est l'un des MPS les plus fréquentes [110].

3.2. Épidémiologie

L'incidence estimée est d'environ 1 :80.000 garçons.

3.3. Clinique

Deux types cliniques sont décrits MPS IIA et MPS IIB. Une atteinte neurodégénérative est observée chez le MPS IIA avec des troubles comportementaux importants dès la 1ère année de vie. Les atteintes viscérales et orthopédiques sont moins marquées que le MPS I. Le diagnostic est habituellement plus tardif entre 18 mois et 4 ans. L'aspect neurocognitif comporte une dégradation neurologique entre 6 et 10 ans. Une atteinte dermatologique caractéristique peut être observée est possible avec des lésions nodulaires de couleur ivoire situées au niveau de la région sacrée, des fesses et des membres supérieurs. Pour le MPS IIB, ne présente pas d'atteinte neurodégénérative. Il est à noter l'absence d'opacités cornéennes dans la MPS II.

3.4. Enzyme déficiente et substrat accumulé

- L'enzyme déficitaire est Iduronate 2-sulfatase (IDS).
- Substrat accumulé : DS et HS.

3.5. Génétique

La MPS II se transmet selon un mode lié à l'X. Le gène *IDS* est localisé sur le chromosome Xq27.3-q28 avec une taille de 24 Kb et contient 9 exons. Un pseudogène avec des régions homologues est localisé à approximativement 25 Kb télomériques du gène fonctionnel. Environ 480 variants ont été décrits. La plupart (70%) étant des variants privées. Il est à noter, 10 à 20% de patients MPS II présentent des altérations génétiques de grande taille, y compris réarrangements et délétions totales du gène *IDS*.

3.6. Traitement

- TES : Idursulfase® (Shire)
- TCSH

4. MPS III

4.1. Définition

Ce syndrome a été mentionné pour la première fois en 1961 dans un rapport des manifestations cliniques d'une fille atteinte d'hépatosplénomégalie, d'une évaluation squelettique normale et d'une excrétion urinaire de grandes quantités d'HS [111]. Deux ans plus tard, le pédiatre Silvestre Sanfilippo décrit huit enfants atteints d'un retard mental et d'une excrétion élevée d'un seul type de GAG, HS [112], désormais connu sous le nom de syndrome de Sanfilippo (MPS III). La MPS III est caractérisée par une détérioration mentale progressive et des troubles de comportement. Quatre sous-types différents, MPS III type A (OMIM : 252900), type B (OMIM : 252920), type C (OMIM : 252930) et type D (OMIM : 252940) sont décrits sur la base du déficit enzymatique. Les patients atteints de MPS III ont d'abord été identifiés sur la base d'études métaboliques au début des années 1960. Plus tard, quatre déficits enzymatiques liés au catabolisme de l'HS ont été identifiés [113].

4.2. Epidemiology

L'incidence globale est estimée à 1:147 000 naissances

4.3. Clinique

La MPS III est cliniquement caractérisée par une dégénérescence sévère du système nerveux central (SNC) avec une atteinte somatique tardive. Les symptômes commencent généralement entre 2 et 6 ans, avec un retard de langage, une hyperactivité, un comportement agressif, des retards de développement et un hirsutisme. Les troubles du sommeil sont une caractéristique très commune, faisant partie du phénotype comportemental [114]. L'hépatosplénomégalie est inconstante et tardive par rapport aux signes de régression neurodéveloppementale. À la fin de la première décennie de vie, il y a une lente perte de compétences, le développement de troubles de la démarche et l'apparition de signes pyramidaux, qui conduisent à un état végétatif et la mort, en général au début de la 3ème décennie [5].

4.4. Enzyme déficiente et substrat accumulé

Syndrome	OMIM	Gène	Enzyme déficiente	Substrat accumulé
MPS IIIA	252900	*SGSH*	Heparan-N-sulfatase	HS
MPS IIIB	252920	*NAGLU*	N-acetyl-α-glucosaminidase	HS
MPS IIIC	252930	*HGSNAT*	Acétly-CoA:α-glucosaminidaseN-acetyltransferase	HS
MPS IIID	252940	*GNS*	N-acetylglucosamine 6-sulfatase	HS

4.5. Génétique

Les MPS III sont transmises selon un mode autosomique récessif.

- Le gène *SGSH* code pour la sulfamidase ou l'héparane N-sulfatase, dont le déficit conduit au **MPS IIIA**. Il est localisé sur le chromosome 17q25.3 et s'étend sur 11 kb et contient huit exons. Au total, 137 variants ont été décrits dans la base HGMD publique, la plupart sont des variants faux-sens (77,3%) [115]. Le gène NAGLU code pour l'α-N-acétylglucosaminidase, responsable de la **MPS IIIB**. Il est localisé sur le chromosome 17q21.2 et s'étend sur 8,5 kb avec six exons. Un total de 152 variants dans la base HGMD publique. La MPS IIIB est plus fréquente en Europe du Sud [116].
- Le gène *HGSNAT* spécifique de l'acétyl CoA: α-glucosaminidase N-acétyltransférase, responsable de la MPS IIIC, a été localisé dans une région péricentromérique dans le chromosome 8p11.21 [117]. Soixante-quatre variants ont été décrits dans la base HGMD publique.
- Le déficit de N-acétylglucosamine 6-sulfatase qui conduit à MPS IIID est causé par des variations dans le gène *GNS* situé sur le chromosome 12q14.3. [118]. Vingt-trois variants ont été décrits dans la base HGMD publique.

4.6. Traitement

- Pas de traitement spécifique

5. MPS IV

5.1. Définition

Les symptômes cliniques de la maladie de Morquio ou MPS IV ont d'abord été décrits par le Dr Luis Morquio en 1929 [119], mais ce n'est que dans les années 1960 que l'on a découvert des concentrations élevées de GAG dans l'urine des patients et qu'on a utilisé le terme MPS IV [120]. En 1976, Singh et al. ont trouvé que l'enzyme N-acétylgalactosamine-6 sulfatase était déficitaire dans le Morquio A et l'année suivante le Morquio B a été identifié comme une entité distincte [120]. Ainsi, la MPS IV inclut deux sous-types : type A (Morquio A) et type B (Morquio B). Le MPS IVA (OMIM : 253000) est due au déficit en N-acétylgalactosamine-sulfate sulfatase (*GALNS)* Morquio A se caractérise par une accumulation de CS et KS dans de nombreux tissus et organes. Cette accumulation entraîne des atteintes cliniques multi-systémiques, notamment des anomalies musculo-squelettiques, une stature courte, un dysfonctionnement pulmonaire et cardiaque, une perte auditive et une opacité cornéenne [121]. La MPS IVB B (OMIM : 253010) est due au déficit en béta-galactosidase (*GLB1*). Les patients Morquio B ont un phénotype plus atténué par rapport aux patients Morquio A. Il est à noter qu'il existe une forme allélique du déficit en beta-galactosidase, la gangliosidose GM1. Cette pathologie est distincte du Morquio B et se présente sous forme d'une encéphalopathie avec des phénotypes précoces infantiles, juvéniles ou des phénotypes adultes.

5.2. Épidémiologie

L'incidence globale est estimée à 1 : 300 000 naissances en France. Elle est de 1 : 76, 320 en Irlande du Nord et 1 : 641 178 en Australie [122].

5.3. Clinique

Pour la MPS IVA, des formes sévères dont les premiers signes peuvent être visibles en anténatal (anasarque foetoplacentaire) ou avant l'âge de 1 an. Des formes intermédiaires à progression lente qui peuvent apparaître au-à l'âge adulte. L'âge moyen au diagnostic est de 5 ans. L'atteinte orthopédique est prédominante avec une dysostose, un retard statural sévère, micromélie, une cyphose, une déformation thoracique, un genu valgum marqué et une hyperlaxité articulaire. Pour la MPS IVB présente une déformation rachidienne et retard statural variable. Il n'y a pas d'atteinte neurodégénérative ni d'opacité cornéenne. Néanmoins, les apprentissages et l'autonomie sont perturbés à cause des problèmes sensoriels et la limitation motrice [123].

5.4. Enzyme déficiente et substrat accumulé

Syndrome	OMIM	Gène	Enzyme déficiente	Substrat accumulé
MPS IVA	253000	*GALNS*	N-acétylgalactosamine-6 sulfatase	CS et KS
MPS IVB	253010	*GLB1*	β-galactosidase	KS

5.5. Génétique

- Le gène *GALNS* code pour la N-acétylgalactosamine-6 sulfatase. Il est localisé sur le chromosome 16q24.3 et s'étend sur 50 kb et contient 14 exons. Cent quatre-vingt-treize variants ont été décrits dans la base HGMD publique.
- Le gène *GLB1* code pour la β-galactosidase. Il est localisé sur le chromosome 16q24.3 et s'étend sur 62.5 kb et contient 16 exons. Cent soixante-dix-huit variants ont été décrits dans la base HGMD publique.

5.6. Traitement

- TES : Elosulfase alfa (Biomarin) pour MPS IVA

6. MPS VI

6.1. Définition

La MPS VI ou syndrome de Maroteaux-Lamy (OMIM : 253200) est décrite pour la première fois en 1963 par les médecins français Pierre Maroteaux et Maurice Lamy [124]. La maladie est due à un déficit de l'enzyme arylsulfatase B ou N-acétylgalactosamine-4-sulfatase (*ARSB*) et conduit à l'accumulation multisystémique de DS. Les manifestations cliniques sont liées à l'accumulation progressive de DS et d'oligosaccharides sulfatés dérivés à la fois de DS et de CS dans les lysosomes, les cellules et les tissus. La maladie est cliniquement hétérogène. Un phénotype sévère est souvent observé. Parfois, la maladie est plus atténuée et progresse lentement [4,125].

6.2. Épidémiologie

L'incidence varie de 1 : 43.261 naissances en Allemagne à 1 dans 1 : 505 160 naissances en Suède [126]. Elle est estimée à 1 : 600 000 naissances en France.

6.3. Clinique

Le diagnostic se fait en général avant l'âge de 2 ans pour les formes sévères, devant une cassure de la courbe de croissance, des déformations squelettiques, des opacités cornéennes, des traits du visage épais. Une obstruction des voies aériennes supérieures est aussi observée. Par ailleurs, les atteintes sensorielles et motrices peuvent retentir sur les apprentissages et l'autonomie. Des formes lentement progressives avec atteintes cardiopulmonaire et ostéo-articulaire ont été observées [126].

6.4. Enzyme déficiente et substrat accumulé

- Enzyme déficiente : Arylsulfatase B
- Substrat accumulé : DS

6.5. Génétique

Le gène *ARSB* est localisé sur le chromosome 5q11–q13 et s'étend sur 44 kb avec huit exons. Cent cinquante variants pathogènes ont été identifiés.

6.6. Traitement

- Galsulfase (Biomarin)

7. MPS VII

7.1. Définition

La MPS VII ou syndrome de Sly (OMIM : 253220) se caractérise par un déficit de l’activité de l'enzyme β-glucuronidase (*GUSB* : β-D-glucuronoside glucuronosohydrolase) [127]. En l'absence de GUS, le CS, DS et HS ne sont que partiellement dégradés et s'accumulent dans les lysosomes de nombreux tissus, conduisant finalement à un dysfonctionnement cellulaire et organique. La MPS VII est rare. La MPS VII provoque un retard mental, une hépatosplénomégalie et une dysplasie squelettique. Les patients atteints de MPS VII présentaient un large éventail de phénotype clinique, allant de la forme la plus sévère avec anasarque foetoplacentaire aux phénotypes plus légers avec apparition plus tardive et intelligence normale.

7.2. Épidémiologie

L’incidence globale est estimée à 1 : 250 000 naissances.

7.3. Clinique

L’âge de révélation est le plus souvent anténatal par une anasarque foeto-placentaire et le pronostic est sombre. La MPS VII provoque un retard mental, une hépatosplénomégalie et une atteinte musculo-squelettique.

7.4. Enzyme déficiente et substrat accumulé

- Enzyme déficiente : β-glucuronidase
- Substrat accumulé : CS, DS et HS. Cependant, le phénotype est hétérogène, pour certains cas seule la CS est augmentée.

7.5. Génétique

Le gène *GUSB* est localisé sur le chromosome 7q21.11 et s'étend sur 21 kb avec 12 exons. Cinquante-quatre variants pathogènes ont été rapportés dans la base HGMD publique.

7.6. Traitement

- TES : Recombinant Human Beta-glucuronidase (rhGUS) Ultrageneix

8. MPS IX

8.1. Définition

La MPS IX (OMIM : 601492) est la forme la plus rare de MPS. Elle est causée par le déficit de l'enzyme hyaluronidase 1 (HYAL-1) qui dégrade l'hyaluronane (acide hyaluronique : HA). L'hyaluronane, est un polymère d'acide D-glucuronique et de N-acétylglucosamine qui est synthétisé dans la membrane plasmique des cellules et maturé dans le cytoplasme. Diverses fonctions biologiques lui sont associées, y compris la modulation de la prolifération cellulaire, la migration et la différenciation, ainsi que la régulation de l'eau extracellulaire et de l'homéostasie des protéines. Il est également un composant structurel intégral du cartilage et d'autres tissus et agit comme un lubrifiant dans les articulations. A ce jour, seulement quatre patients ont été signalés. La première patiente a été rapportée en 1996. Elle avait des masses de tissus mous péri-articulaires et des érosions acétabulaires sans caractéristiques classiques de MPS comme l'atteinte neurologique ou viscérale [128]. Trois autres patients étaient des enfants de parents consanguins du Moyen-Orient et tous présentaient une arthrite idiopathique juvénile (JIA) [129].

8.2. Épidémiologie

La MPS IX est exceptionnelle. Compte tenu de son extrême rareté, Il n'y a pas d'information dans la littérature concernant cette enquête de prévalence. Quatre patients sont décrits.

8.3. Clinique

Tous les patients signalés avec MPS de type IX ont présenté des problèmes articulaires et squelettiques. Par conséquent, la MPS IX peut être facilement diagnostiquée comme JIA. Elle est caractérisée par de multiples masses périarticulaires dans les tissus mous et des kystes synoviaux.

8.4. Enzyme déficiente et substrat accumulé

- Enzyme déficiente : Hyaluronidase 1
- Substrat accumulé : HA

8.5. Génétique

Le gène *HYAL1* est localisé sur le chromosome 3p21 et s'étend sur 3.5 kb avec 3 exons. Trois variants pathogènes ont été rapportés dans la base HGMD publique.

8.6. Traitement

- Pas de traitement spécifique

REFERENCES

1. Lammert, E.; Zeeb, M. *Metabolism of human diseases: Organ physiology and pathophysiology*. Springer Vienna: 2014.
2. Garrod, A. The incidence of alkaptonuria : A study in chemical individuality. *The Lancet* **1902**, *160*, 1616-1620.
3. Scriver, C.R. Garrod's croonian lectures (1908) and the charter 'inborn errors of metabolism': Albinism, alkaptonuria, cystinuria, and pentosuria at age 100 in 2008. *J Inherit Metab Dis* **2008**, *31*, 580-598.
4. Tada, K.; Buist, N.R.M.; Fernandes, J.; Saudubray, J.M.; van den Berghe, G. *Inborn metabolic diseases: Diagnosis and treatment*. Springer Berlin Heidelberg: Germany, 2013.
5. Hoffmann, G.F.; Zschocke, J.; Nyhan, W.L. *Inherited metabolic diseases: A clinical approach*. Springer Berlin Heidelberg: Germany, 2016.
6. Parenti, G.; Andria, G.; Ballabio, A. Lysosomal storage diseases: From pathophysiology to therapy. *Annual review of medicine* **2015**, *66*, 471-486.
7. Vance, J.E.; Karten, B. Niemann-pick c disease and mobilization of lysosomal cholesterol by cyclodextrin. *Journal of lipid research* **2014**, *55*, 1609-1621.
8. Coutinho, M.F.; Santos, J.I.; Alves, S. Less is more: Substrate reduction therapy for lysosomal storage disorders. *International Journal of Molecular Sciences* **2016**, *17*, 1065.
9. Matalonga, L.; Gort, L.; Ribes, A. Small molecules as therapeutic agents for inborn errors of metabolism. *J Inherit Metab Dis* **2017**, *40*, 177-193.
10. Shapiro, B.E.; Pastores, G.M.; Gianutsos, J.; Luzy, C.; Kolodny, E.H. Miglustat in late-onset tay-sachs disease: A 12-month, randomized, controlled clinical study with 24 months of extended treatment. *Genetics in Medicine* **2009**, *11*, 425-433.
11. Wraith, J.E.; Vecchio, D.; Jacklin, E.; Abel, L.; Chadha-Boreham, H.; Luzy, C.; Giorgino, R.; Patterson, M.C. Miglustat in adult and juvenile patients with niemann–pick disease type c: Long-term data from a clinical trial. *Molecular genetics and metabolism* **2010**, *99*, 351-357.
12. Villamizar-Schiller, I.T.; Pabón, L.A.; Hufnagel, S.B.; Serrano, N.C.; Karl, G.; Jefferies, J.L.; Hopkin, R.J.; Prada, C.E. Neurological and cardiac responses after treatment with miglustat and a ketogenic diet in a patient with sandhoff disease. *European journal of medical genetics* **2015**, *58*, 180-183.
13. Smid, B.E.; Ferraz, M.J.; Verhoek, M.; Mirzaian, M.; Wisse, P.; Overkleeft, H.S.; Hollak, C.E.; Aerts, J.M. Biochemical response to substrate reduction therapy versus enzyme replacement therapy in gaucher disease type 1 patients. *Orphanet Journal of Rare Diseases* **2016**, *11*, 28.
14. Aldenhoven, M.; Wynn, R.F.; Orchard, P.J.; O'Meara, A.; Veys, P.; Fischer, A.; Valayannopoulos, V.; Neven, B.; Rovelli, A.; Prasad, V.K. Long-term outcome of hurler syndrome patients after hematopoietic cell transplantation: An international multicenter study. *Blood* **2015**, *125*, 2164-2172.
15. Tomatsu, S.; Azario, I.; Sawamoto, K.; Pievani, A.S.; Biondi, A.; Serafini, M. Neonatal cellular and gene therapies for mucopolysaccharidoses: The earlier the better? *J Inherit Metab Dis* **2016**, *39*, 189-202.
16. Kim, S.U. Lysosomal storage diseases: Stem cell-based cell- and gene-therapy. *Cell Transplant* **2014**.
17. Muenzer, J. Early initiation of enzyme replacement therapy for the mucopolysaccharidoses. *Molecular genetics and metabolism* **2014**, *111*, 63-72.
18. Beutler, E. Enzyme replacement in gaucher disease. *PLoS medicine* **2004**, *1*, e21.
19. Baldo, B.A. Enzymes approved for human therapy: Indications, mechanisms and adverse effects. *BioDrugs : clinical immunotherapeutics, biopharmaceuticals and gene therapy* **2015**, *29*, 31-55.
20. Fan, J.Q.; Ishii, S.; Asano, N.; Suzuki, Y. Accelerated transport and maturation of lysosomal alpha-galactosidase a in fabry lymphoblasts by an enzyme inhibitor. *Nature medicine* **1999**, *5*, 112-115.
21. Clancy, J.; BEBÖK, Z.; Ruiz, F.; King, C.; Jones, J.; Walker, L.; Greer, H.; Hong, J.; Wing, L.; Macaluso, M. Evidence that systemic gentamicin suppresses premature stop mutations in patients with cystic fibrosis. *American journal of respiratory and critical care medicine* **2001**, *163*, 1683-1692.
22. Kerem, E.; Hirawat, S.; Armoni, S.; Yaakov, Y.; Shoseyov, D.; Cohen, M.; Nissim-Rafinia, M.; Blau, H.; Rivlin, J.; Aviram, M. Effectiveness of ptc124 treatment of cystic fibrosis caused by nonsense mutations: A prospective phase ii trial. *The Lancet* **2008**, *372*, 719-727.
23. Muntau, A.C.; Leandro, J.; Staudigl, M.; Mayer, F.; Gersting, S.W. Innovative strategies to treat protein misfolding in inborn errors of metabolism: Pharmacological chaperones and proteostasis regulators. *J Inherit Metab Dis* **2014**, *37*, 505-523.
24. Tatti, M.; Motta, M.; Scarpa, S.; Di Bartolomeo, S.; Cianfanelli, V.; Tartaglia, M.; Salvioli, R. Bcm-95 and (2-hydroxypropyl)-β-cyclodextrin reverse autophagy dysfunction and deplete stored lipids in sap c-deficient fibroblasts. *Human molecular genetics* **2015**, ddv153.
25. Williams, I.M.; Wallom, K.-L.; Smith, D.A.; Al Eisa, N.; Smith, C.; Platt, F.M. Improved neuroprotection using miglustat, curcumin and ibuprofen as a triple combination therapy in niemann–pick disease type c1 mice. *Neurobiology of disease* **2014**, *67*, 9-17.
26. Mu, T.-W.; Ong, D.S.T.; Wang, Y.-J.; Balch, W.E.; Yates, J.R.; Segatori, L.; Kelly, J.W. Chemical and biological approaches synergize to ameliorate protein-folding diseases. *Cell* **2008**, *134*, 769-781.
27. Santana-Codina, N.; Mancias, J.D.; Kimmelman, A.C. The role of autophagy in cancer. *Annual Review of Cancer Biology* **2017**, *1*.
28. Menzies, F.M.; Fleming, A.; Caricasole, A.; Bento, C.F.; Andrews, S.P.; Ashkenazi, A.; Füllgrabe, J.; Jackson, A.; Sanchez, M.J.; Karabiyik, C. Autophagy and neurodegeneration: Pathogenic mechanisms and therapeutic opportunities. *Neuron* **2017**, *93*, 1015-1034.
29. García-Prat, L.; Martínez-Vicente, M.; Perdiguero, E.; Ortet, L.; Rodríguez-Ubreva, J.; Rebollo, E.; Ruiz-Bonilla, V.; Gutarra, S.; Ballestar, E.; Serrano, A.L. Autophagy maintains stemness by preventing senescence. *Nature* **2016**, *529*, 37-42.
30. Morel, E.; Mehrpour, M.; Botti, J.; Dupont, N.; Hamaï, A.; Nascimbeni, A.C.; Codogno, P. Autophagy: A druggable process. *Annual review of pharmacology and toxicology* **2017**, *57*, 375-398.
31. Settembre, C.; Fraldi, A.; Rubinsztein, D.C.; Ballabio, A. Lysosomal storage diseases as disorders of autophagy. *Autophagy* **2008**, *4*, 113-114.
32. Chévrier, M.; Brakch, N.; Céline, L.; Genty, D.; Ramdani, Y.; Moll, S.; Djavaheri-Mergny, M.; Brasse-Lagnel, C.; Annie Laquerrière, A.L.; Barbey, F. Autophagosome maturation is impaired in fabry disease. *Autophagy* **2010**, *6*, 589-599.
33. Song, W.; Wang, F.; Lotfi, P.; Sardiello, M.; Segatori, L. 2-hydroxypropyl-β-cyclodextrin promotes transcription factor eb-mediated activation of autophagy implications for therapy. *Journal of Biological Chemistry* **2014**, *289*, 10211-10222.
34. García-Robles, A.A.; Company-Albir, M.J.; Megías-Vericat, J.E.; Fernández-Megía, M.J.; Pérez-Miralles, F.C.; López-Briz, E.; Alcalá-Vicente, C.; Galeano, I.; Casanova, B.; Poveda, J.L. Use of 2 hydroxypropyl-beta-cyclodextrin therapy in two adult niemann pick type c patients. *Journal of the neurological sciences* **2016**, *366*, 65.
35. Kuo, S.-Y.; Castoreno, A.B.; Aldrich, L.N.; Lassen, K.G.; Goel, G.; Dančík, V.; Kuballa, P.; Latorre, I.; Conway, K.L.; Sarkar, S. Small-molecule enhancers of autophagy modulate cellular disease phenotypes suggested by human genetics. *Proceedings of the National Academy of Sciences* **2015**, *112*, E4281-E4287.
36. De Duve, C. The lysosome turns fifty. *Nature cell biology* **2005**, *7*, 847-849.
37. Settembre, C.; Fraldi, A.; Medina, D.L.; Ballabio, A. Signals from the lysosome: A control centre for cellular clearance and energy metabolism. *Nature reviews Molecular cell biology* **2013**, *14*, 283-296.
38. Boustany, R.-M.N. Lysosomal storage diseases[mdash]the horizon expands. *Nat Rev Neurol* **2013**, *9*, 583-598.
39. Sachs, B. On arrested cerebral development, with special reference to its cortical pathology. 1. *The Journal of Nervous and Mental Disease* **1887**, *14*, 541-553.
40. Chen, M.; Wang, J. Gaucher disease: Review of the literature. *Archives of pathology & laboratory medicine* **2008**, *132*, 851-853.

41. Fabry, J. Ein beitrag zur kenntniss der purpura haemorrhagica nodularis (purpura papulosa haemorrhagica hebrae). *Archiv für Dermatologie und Syphilis* **1898**, *43*, 187-200.
42. Ballabio, A.; Gieselmann, V. Lysosomal disorders: From storage to cellular damage. *Biochimica et biophysica acta* **2009**, *1793*, 684-696.
43. Xu, H.; Ren, D. Lysosomal physiology. *Annual review of physiology* **2015**, *77*, 57-80.
44. Saftig, P.; Klumperman, J. Lysosome biogenesis and lysosomal membrane proteins: Trafficking meets function. *Nature reviews Molecular cell biology* **2009**, *10*, 623-635.
45. Ruivo, R.; Anne, C.; Sagné, C.; Gasnier, B. Molecular and cellular basis of lysosomal transmembrane protein dysfunction. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* **2009**, *1793*, 636-649.
46. Sardiello, M.; Palmieri, M.; di Ronza, A.; Medina, D.L.; Valenza, M.; Gennarino, V.A.; Di Malta, C.; Donaudy, F.; Embrione, V.; Polishchuk, R.S. A gene network regulating lysosomal biogenesis and function. *Science (New York, N.Y.)* **2009**, *325*, 473-477.
47. Mehta, A.B.; Winchester, B. *Lysosomal storage disorders: A practical guide*. John Wiley & Sons: 2012.
48. Eisenstein, M. Myriad maladies. *Nature* **2016**, *537*, S146-S147.
49. Coutinho, M.F.; Alves, S. From rare to common and back again: 60 years of lysosomal dysfunction. *Molecular genetics and metabolism* **2016**, *117*, 53-65.
50. Vitner, E.B.; Platt, F.M.; Futerman, A.H. Common and uncommon pathogenic cascades in lysosomal storage diseases. *Journal of Biological Chemistry* **2010**, *285*, 20423-20427.
51. Ballabio, A.; Gieselmann, V. Lysosomal disorders: From storage to cellular damage. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* **2009**, *1793*, 684-696.
52. Manger, B.; Mengel, E.; Schaefer, R.M. Rheumatologic aspects of lysosomal storage diseases. *Clinical rheumatology* **2007**, *26*, 335-341.
53. Cimaz, R.; Coppa, G.V.; Koné-Paut, I.; Link, B.; Pastores, G.M.; Elorduy, M.R.; Spencer, C.; Thorne, C.; Wulffraat, N.; Manger, B. Joint contractures in the absence of inflammation may indicate mucopolysaccharidosis. *Pediatric Rheumatology* **2009**, *7*, 18.
54. Sévin, M.; Lesca, G.; Baumann, N.; Millat, G.; Lyon-Caen, O.; Vanier, M.T.; Sedel, F. The adult form of niemann–pick disease type c. *Brain* **2007**, *130*, 120-133.
55. Bekri, S. Diagnostic biologique des maladies lysosomales. *Annales de Biologie Clinique* **2006**, *64*, 592-600.
56. Shen, J.; Zhou, Y.; Lu, T.; Peng, J.; Lin, Z.; Huang, L.; Pang, Y.; Yu, L.; Huang, Y. An integrated chip for immunofluorescence and its application to analyze lysosomal storage disorders. *Lab on a chip* **2012**, *12*, 317-324.
57. Humbel, R.; Collart, M. Oligosaccharides in urine of patients with glycoprotein storage diseases. I. Rapid detection by thin-layer chromatography. *Clinica chimica acta; international journal of clinical chemistry* **1975**, *60*, 143-145.
58. Bonesso, L.; Piraud, M.; Caruba, C.; Van Obberghen, E.; Mengual, R.; Hinault, C. Fast urinary screening of oligosaccharidoses by maldi-tof/tof mass spectrometry. *Orphanet J Rare Dis* **2014**, *9*, 19.
59. Ramsay, S.L.; Meikle, P.J.; Hopwood, J.J.; Clements, P.R. Profiling oligosaccharidurias by electrospray tandem mass spectrometry: Quantifying reducing oligosaccharides. *Analytical biochemistry* **2005**, *345*, 30-46.
60. Ramsay, S.L.; Maire, I.; Bindloss, C.; Fuller, M.; Whitfield, P.D.; Piraud, M.; Hopwood, J.J.; Meikle, P.J. Determination of oligosaccharides and glycolipids in amniotic fluid by electrospray ionisation tandem mass spectrometry: In utero indicators of lysosomal storage diseases. *Molecular genetics and metabolism* **2004**, *83*, 231-238.
61. Xia, B.; Asif, G.; Arthur, L.; Pervaiz, M.A.; Li, X.; Liu, R.; Cummings, R.D.; He, M. Oligosaccharide analysis in urine by maldi-tof mass spectrometry for the diagnosis of lysosomal storage diseases. *Clin Chem* **2013**, *59*, 1357-1368.
62. Gray, G.; Claridge, P.; Jenkinson, L.; Green, A. Quantitation of urinary glycosaminoglycans using dimethylene blue as a screening technique for the diagnosis of mucopolysaccharidoses: An evaluation. *Annals of clinical biochemistry* **2007**, *44*, 360-363.
63. de Jong, J.G.; Wevers, R.A.; Liebrand-van Sambeek, R. Measuring urinary glycosaminoglycans in the presence of protein: An improved screening procedure for mucopolysaccharidoses based on dimethylmethylene blue. *Clin Chem* **1992**, *38*, 803-807.
64. Afyoncu, E.; Yilmaz, G.; Yilmaz, F.M.; Yucel, D. Performance of different screening methods for the determination of urinary glycosaminoclycans. *Clinical chemistry and laboratory medicine* **2013**, *51*, 347-350.
65. Piraud, M.; Boyer, S.; Mathieu, M.; Maire, I. Diagnosis of mucopolysaccharidoses in a clinically selected population by urinary glycosaminoglycan analysis: A study of 2,000 urine samples. *Clinica chimica acta; international journal of clinical chemistry* **1993**, *221*, 171-181.
66. Zamfir, A.D. Applications of capillary electrophoresis electrospray ionization mass spectrometry in glycosaminoglycan analysis. *Electrophoresis* **2016**, *37*, 973-986.
67. Mashima, R.; Sakai, E.; Tanaka, M.; Kosuga, M.; Okuyama, T. The levels of urinary glycosaminoglycans of patients with attenuated and severe type of mucopolysaccharidosis ii determined by liquid chromatography-tandem mass spectrometry. *Molecular Genetics and Metabolism Reports* **2016**, *7*, 87-91.
68. Kubaski, F.; Osago, H.; Mason, R.W.; Yamaguchi, S.; Kobayashi, H.; Tsuchiya, M.; Orii, T.; Tomatsu, S. Glycosaminoglycans detection methods: Applications of mass spectrometry. *Molecular genetics and metabolism* **2016**.
69. Auray-Blais, C.; Lavoie, P.; Maranda, B.; Boutin, M. Evaluation of urinary keratan sulfate disaccharides in mps iva patients using uplc-ms/ms. *Bioanalysis* **2016**, *8*, 179-191.
70. Zhang, H.; Wood, T.; Young, S.P.; Millington, D.S. A straightforward, quantitative ultra-performance liquid chromatography-tandem mass spectrometric method for heparan sulfate, dermatan sulfate and chondroitin sulfate in urine: An improved clinical screening test for the mucopolysaccharidoses. *Molecular genetics and metabolism* **2015**, *114*, 123-128.
71. Langereis, E.J.; Wagemans, T.; Kulik, W.; Lefeber, D.J.; van Lenthe, H.; Oussoren, E.; van der Ploeg, A.T.; Ruijter, G.J.; Wevers, R.A.; Wijburg, F.A., *et al.* A multiplex assay for the diagnosis of mucopolysaccharidoses and mucolipidoses. *PloS one* **2015**, *10*, e0138622.
72. Tomatsu, S.; Shimada, T.; Mason, R.W.; Montano, A.M.; Kelly, J.; LaMarr, W.A.; Kubaski, F.; Giugliani, R.; Guha, A.; Yasuda, E., *et al.* Establishment of glycosaminoglycan assays for mucopolysaccharidoses. *Metabolites* **2014**, *4*, 655-679.
73. Tomatsu, S.; Shimada, T.; Mason, R.W.; Kelly, J.; LaMarr, W.A.; Yasuda, E.; Shibata, Y.; Futatsumori, H.; Montano, A.M.; Yamaguchi, S., *et al.* Assay for glycosaminoglycans by tandem mass spectrometry and its applications. *J Anal Bioanal Tech* **2014**, *2014*, 006.
74. Tomatsu, S.; Kubaski, F.; Sawamoto, K.; Mason, R.W.; Yasuda, E.; Shimada, T.; Montaño, A.M.; Yamaguchi, S.; Suzuki, Y.; Orii, T. Newborn screening and diagnosis of mucopolysaccharidoses: Application of tandem mass spectrometry. *Nihon Masu Sukuriningu Gakkai shi = Journal of Japanese Society for Mass-Screening* **2014**, *24*, 19-37.
75. Osago, H.; Shibata, T.; Hara, N.; Kuwata, S.; Kono, M.; Uchio, Y.; Tsuchiya, M. Quantitative analysis of glycosaminoglycans, chondroitin/dermatan sulfate, hyaluronic acid, heparan sulfate, and keratan sulfate by liquid chromatography–electrospray ionization–tandem mass spectrometry. *Analytical biochemistry* **2014**, *467*, 62-74.
76. Zhang, H.; Young, S.P.; Auray-Blais, C.; Orchard, P.J.; Tolar, J.; Millington, D.S. Analysis of glycosaminoglycans in cerebrospinal fluid from patients with mucopolysaccharidoses by isotope-dilution ultra-performance liquid chromatography-tandem mass spectrometry. *Clin Chem* **2011**, *57*, 1005-1012.
77. Meikle, P.J.; Ranieri, E.; Simonsen, H.; Rozaklis, T.; Ramsay, S.L.; Whitfield, P.D.; Fuller, M.; Christensen, E.; Skovby, F.; Hopwood, J.J. Newborn screening for lysosomal storage disorders: Clinical evaluation of a two-tier strategy. *Pediatrics* **2004**, *114*, 909-916.
78. Civallero, G.; Michelin, K.; de Mari, J.; Viapiana, M.; Burin, M.; Coelho, J.C.; Giugliani, R. Twelve different enzyme assays on dried-blood filter paper samples for detection of patients with selected inherited lysosomal storage diseases. *Clinica Chimica Acta* **2006**, *372*, 98-102.
79. Spáčil, Z.k.; Elliott, S.; Reeber, S.L.; Gelb, M.H.; Scott, C.R.; Tureček, F.e. Comparative triplex tandem mass spectrometry assays of lysosomal enzyme activities in dried blood spots using fast liquid chromatography: Application to newborn screening of pompe, fabry, and hurler diseases. *Analytical chemistry* **2011**, *83*, 4822-4828.

80. Sista, R.S.; Eckhardt, A.E.; Wang, T.; Graham, C.; Rouse, J.L.; Norton, S.M.; Srinivasan, V.; Pollack, M.G.; Tolun, A.A.; Bali, D., *et al.* Digital microfluidic platform for multiplexing enzyme assays: Implications for lysosomal storage disease screening in newborns. *Clinical Chemistry* **2011**, *57*, 1444-1451.
81. Sista, R.S.; Wang, T.; Wu, N.; Graham, C.; Eckhardt, A.; Winger, T.; Srinivasan, V.; Bali, D.; Millington, D.S.; Pamula, V.K. Multiplex newborn screening for pompe, fabry, hunter, gaucher, and hurler diseases using a digital microfluidic platform. *Clinica Chimica Acta* **2013**, *424*, 12-18.
82. Spacil, Z.; Tatipaka, H.; Barcenas, M.; Scott, C.R.; Turecek, F.; Gelb, M.H. High-throughput assay of 9 lysosomal enzymes for newborn screening. *Clinical chemistry* **2013**, *59*, 502-511.
83. Elliott, S.; Buroker, N.; Cournoyer, J.J.; Potier, A.M.; Trometer, J.D.; Elbin, C.; Schermer, M.J.; Kantola, J.; Boyce, A.; Turecek, F., *et al.* Pilot study of newborn screening for six lysosomal storage diseases using tandem mass spectrometry. *Molecular genetics and metabolism* **2016**, *118*, 304-309.
84. Cho, S.E.; Kwak, J.R.; Lee, H.; Seo, D.H.; Song, J. Triplex tandem mass spectrometry assays for the screening of 3 lysosomal storage disorders in a korean population. *Clinica chimica acta; international journal of clinical chemistry* **2016**, *454*, 20-27.
85. Liao, H.C.; Chiang, C.C.; Niu, D.M.; Wang, C.H.; Kao, S.M.; Tsai, F.J.; Huang, Y.H.; Liu, H.C.; Huang, C.K.; Gao, H.J., *et al.* Detecting multiple lysosomal storage diseases by tandem mass spectrometry--a national newborn screening program in taiwan. *Clinica chimica acta; international journal of clinical chemistry* **2014**, *431*, 80-86.
86. Brand, G.D.; Matos, H.C.; Cruz, G.C.; Fontes Ndo, C.; Buzzi, M.; Brum, J.M. Diagnosing lysosomal storage diseases in a brazilian non-newborn population by tandem mass spectrometry. *Clinics (Sao Paulo, Brazil)* **2013**, *68*, 1469-1473.
87. Mechtler, T.P.; Stary, S.; Metz, T.F.; De Jesus, V.R.; Greber-Platzer, S.; Pollak, A.; Herkner, K.R.; Streubel, B.; Kasper, D.C. Neonatal screening for lysosomal storage disorders: Feasibility and incidence from a nationwide study in austria. *Lancet* **2012**, *379*, 335-341.
88. Gelb, M.H.; Turecek, F.; Scott, C.R.; Chamoles, N.A. Direct multiplex assay of enzymes in dried blood spots by tandem mass spectrometry for the newborn screening of lysosomal storage disorders. *J Inherit Metab Dis* **2006**, *29*, 397-404.
89. Meikle, P.J.; Grasby, D.J.; Dean, C.J.; Lang, D.L.; Bockmann, M.; Whittle, A.M.; Fietz, M.J.; Simonsen, H.; Fuller, M.; Brooks, D.A., *et al.* Newborn screening for lysosomal storage disorders. *Molecular genetics and metabolism* **2006**, *88*, 307-314.
90. Manwaring, V.; Heywood, W.E.; Clayton, R.; Lachmann, R.H.; Keutzer, J.; Hindmarsh, P.; Winchester, B.; Heales, S.; Mills, K. The identification of new biomarkers for identifying and monitoring kidney disease and their translation into a rapid mass spectrometry-based test: Evidence of presymptomatic kidney disease in pediatric fabry and type-i diabetic patients. *Journal of proteome research* **2013**, *12*, 2013-2021.
91. Filocamo, M.; Morrone, A. Lysosomal storage disorders: Molecular basis and laboratory testing. *Human genomics* **2011**, *5*, 156.
92. Komlosi, K.; Sólyom, A.; Beck, M. The role of next-generation sequencing in the diagnosis of lysosomal storage disorders. *Journal of Inborn Errors of Metabolism and Screening* **2016**, *4*, 2326409816669376.
93. Tebani, A.; Afonso, C.; Marret, S.; Bekri, S. Omics-based strategies in precision medicine: Toward a paradigm shift in inborn errors of metabolism investigations. *Int J Mol Sci* **2016**, *17*.
94. Goodwin, S.; McPherson, J.D.; McCombie, W.R. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **2016**, *17*, 333-351.
95. Ashley, E.A. Towards precision medicine. *Nature Reviews Genetics* **2016**, *17*, 507-522.
96. Schrader, K.A.; Heravi-Moussavi, A.; Waters, P.J.; Senz, J.; Whelan, J.; Ha, G.; Eydoux, P.; Nielsen, T.; Gallagher, B.; Oloumi, A. Using next-generation sequencing for the diagnosis of rare disorders: A family with retinitis pigmentosa and skeletal abnormalities. *The Journal of pathology* **2011**, *225*, 12-18.
97. Lo, S.M.; Choi, M.; Liu, J.; Jain, D.; Boot, R.G.; Kallemeijn, W.W.; Aerts, J.M.; Pashankar, F.; Kupfer, G.M.; Mane, S. Phenotype diversity in type 1 gaucher disease: Discovering the genetic basis of gaucher disease/hematologic malignancy phenotype by individual genome analysis. *Blood* **2012**, *119*, 4731-4740.
98. Altarescu, G.; Beeri, R.; Eiges, R.; Epsztejn-Litman, S.; Eldar-Geva, T.; Elstein, D.; Zimran, A.; Margalioth, E.J.; Levy-Lahad, E.; Renbaum, P. Prevention of lysosomal storage diseases and derivation of mutant stem cell lines by preimplantation genetic diagnosis. *Molecular biology international* **2012**, *2012*.
99. Mackie, F.; Hemming, K.; Allen, S.; Morris, R.; Kilby, M. The accuracy of cell-free fetal DNA-based non-invasive prenatal testing in singleton pregnancies: A systematic review and bivariate meta-analysis. *BJOG: An International Journal of Obstetrics & Gynaecology* **2017**, *124*, 32-46.
100. Schaefer, L.; Schaefer, R.M. Proteoglycans: From structural compounds to signaling molecules. *Cell and Tissue Research* **2009**, *339*, 237.
101. Muenzer, J. Overview of the mucopolysaccharidoses. *Rheumatology (Oxford, England)* **2011**, *50 Suppl 5*, v4-12.
102. Iozzo, R.V.; Schaefer, L. Proteoglycan form and function: A comprehensive nomenclature of proteoglycans. *Matrix Biology* **2015**, *42*, 11-55.
103. Varki, A. Biological roles of glycans. *Glycobiology* **2017**, *27*, 3-49.
104. Piotrowska, E.; Jakóbkiewicz-Banecka, J.; Tylki-Szymanska, A.; Liberek, A.; Maryniak, A.; Malinowska, M.; Czartoryska, B.; Puk, E.; Kloska, A.; Liberek, T. Genistin-rich soy isoflavone extract in substrate reduction therapy for sanfilippo syndrome: An open-label, pilot study in 10 pediatric patients. *Current Therapeutic Research* **2008**, *69*, 166-179.
105. Berliner, M.L. Lipin keratitis of hurler's syndrome (gargoylism or dysostosis multiplex): Clinical and pathologic report. *Archives of Ophthalmology* **1939**, *22*, 97-105.
106. Gifford, S.R.; Scheie, H.G.; Hambrick Jr, G.W.; Barness, L.A. A newly recognized forme fruste of hurler's disease (gargoylism)*. *American Journal of Ophthalmology* **1962**, *53*, 753-769.
107. Bach, G.; Friedman, R.; Weissmann, B.; Neufeld, E.F. The defect in the hurler and scheie syndromes: Deficiency of α-l-iduronidase. *Proceedings of the National Academy of Sciences* **1972**, *69*, 2048-2051.
108. Tebani, A.; Zanoutene-Cheriet, L.; Adjtoutah, Z.; Abily-Donval, L.; Brasse-Lagnel, C.; Laquerriere, A.; Marret, S.; Chalabi Benabdellah, A.; Bekri, S. Clinical and molecular characterization of patients with mucopolysaccharidosis type i in an algerian series. *Int J Mol Sci* **2016**, *17*.
109. Bach, G.; Eisenberg, F.; Cantz, M.; Neufeld, E.F. The defect in the hunter syndrome: Deficiency of sulfoiduronate sulfatase. *Proceedings of the National Academy of Sciences* **1973**, *70*, 2134-2138.
110. Martin, R.; Beck, M.; Eng, C.; Giugliani, R.; Harmatz, P.; Muñoz, V.; Muenzer, J. Recognition and diagnosis of mucopolysaccharidosis ii (hunter syndrome). *Pediatrics* **2008**, *121*, e377-e386.
111. Harris, R. Mucopolysaccharide disorder: A possible new genotype of hurler's syndrome. *Am J Dis Child* **1961**, *102*, 741.
112. Sanfilippo, S.J.; Podosin, R.; Langer, L.; Good, R.A. Mental retardation associated with acid mucopolysacchariduria (heparitin sulfate type). *The Journal of Pediatrics* **1963**, *63*, 837-838.
113. Valstar, M.J.; Ruijter, G.J.G.; van Diggelen, O.P.; Poorthuis, B.J.; Wijburg, F.A. Sanfilippo syndrome: A mini-review. *J Inherit Metab Dis* **2008**, *31*, 240-252.
114. Cross, E.M.; Hare, D.J. Behavioural phenotypes of the mucopolysaccharide disorders: A systematic literature review of cognitive, motor, social, linguistic and behavioural presentation in the mps disorders. *J Inherit Metab Dis* **2013**, *36*, 189-200.
115. Andrade, F.; Aldamiz-Echevarria, L.; Llarena, M.; Couce, M.L. Sanfilippo syndrome: Overall review. *Pediatrics international : official journal of the Japan Pediatric Society* **2015**, *57*, 331-338.
116. Poorthuis, B.J.; Wevers, R.A.; Kleijer, W.J.; Groener, J.E.; de Jong, J.G.; van Weely, S.; Niezen-Koning, K.E.; van Diggelen, O.P. The frequency of lysosomal storage diseases in the netherlands. *Human genetics* **1999**, *105*, 151-156.
117. Fan, X.; Zhang, H.; Zhang, S.; Bagshaw, R.D.; Tropak, M.B.; Callahan, J.W.; Mahuran, D.J. Identification of the gene encoding the enzyme deficient in mucopolysaccharidosis iiic (sanfilippo disease type c). *American journal of human genetics* **2006**, *79*, 738-744.
118. Robertson, D.A.; Freeman, C.; Morris, C.P.; Hopwood, J.J. A cdna clone for human glucosamine-6-sulphatase reveals differences between arylsulphatases and non-arylsulphatases. *The Biochemical journal* **1992**, *288 (Pt 2)*, 539-544.

119. Brailsford, J.F. Chondro-osteo-dystrophy: Roentgenographic & clinical features of a child with dislocation of vertebrae. *The American Journal of Surgery* **1929**, *7*, 404-410.
120. Hendriksz, C.; Harmatz, P.; Beck, M.; Jones, S.; Wood, T.; Lachman, R.; Gravance, C.; Orii, T.; Tomatsu, S. Review of clinical presentation and diagnosis of mucopolysaccharidosis iva. *Molecular genetics and metabolism* **2013**, *110*, 54-64.
121. Montano, A.M.; Tomatsu, S.; Gottesman, G.S.; Smith, M.; Orii, T. International morquio a registry: Clinical manifestation and natural course of morquio a disease. *J Inherit Metab Dis* **2007**, *30*, 165-174.
122. Leadley, R.M.; Lang, S.; Misso, K.; Bekkering, T.; Ross, J.; Akiyama, T.; Fietz, M.; Giugliani, R.; Hendriksz, C.J.; Hock, N.L., *et al.* A systematic review of the prevalence of morquio a syndrome: Challenges for study reporting in rare diseases. *Orphanet J Rare Dis* **2014**, *9*, 173.
123. Hendriksz, C.J.; Harmatz, P.; Beck, M.; Jones, S.; Wood, T.; Lachman, R.; Gravance, C.G.; Orii, T.; Tomatsu, S. Review of clinical presentation and diagnosis of mucopolysaccharidosis iva. *Molecular genetics and metabolism* **2013**, *110*, 54-64.
124. Maroteaux, P.; Leveque, B.; Marie, J.; Lamy, M. [a new dysostosis with urinary elimination of chondroitin sulfate b]. *Presse Med* **1963**, *71*, 1849-1852.
125. Giugliani, R.; Herber, S.; Lapagesse, L.; de Pinto, C.; Baldo, G. Therapy for mucopolysaccharidosis vi: (maroteaux-lamy syndrome) present status and prospects. *Pediatric endocrinology reviews : PER* **2014**, *12 Suppl 1*, 152-158.
126. Valayannopoulos, V.; Nicely, H.; Harmatz, P.; Turbeville, S. Mucopolysaccharidosis vi. *Orphanet Journal of Rare Diseases* **2010**, *5*, 5.
127. Sly, W.S.; Quinton, B.A.; McAlister, W.H.; Rimoin, D.L. Beta glucuronidase deficiency: Report of clinical, radiologic, and biochemical features of a new mucopolysaccharidosis. *J Pediatr* **1973**, *82*, 249-257.
128. Natowicz, M.R.; Short, M.P.; Wang, Y.; Dickersin, G.R.; Gebhardt, M.C.; Rosenthal, D.I.; Sims, K.B.; Rosenberg, A.E. Clinical and biochemical manifestations of hyaluronidase deficiency. *The New England journal of medicine* **1996**, *335*, 1029-1033.
129. Imundo, L.; Leduc, C.A.; Guha, S.; Brown, M.; Perino, G.; Gushulak, L.; Triggs-Raine, B.; Chung, W.K. A complete deficiency of hyaluronoglucosaminidase 1 (hyal1) presenting as familial juvenile idiopathic arthritis. *J Inherit Metab Dis* **2011**, *34*, 1013-1022.

CHAPITRE IV : PLACE DES STRATEGIES OMIQUES DANS LA MEDECINE DE PRECISION

Les techniques omiques offrent des points de vue complémentaires pour l'exploration des processus physiopathologique des maladies. Les cinq principales dans les sciences médicales sont la génomique, la transcriptomique, la protéomique, la métabolomique et l'épigénomique. Ces techniques visent la détection globale des gènes, des ARNm, des protéines, des métabolites et profile de méthylation, respectivement. Le génome offre beaucoup d'informations sur la susceptibilité d'un individu aux maladies génétiques et la prédiction de la réponse au traitement pour un patient donné. Il peut également aider à élucider les mécanismes moléculaires de la maladie, identifier les cibles thérapeutiques potentielles, concevoir des médicaments. Cependant, génome seul est partiellement informatif. Souvent, l'information génomique seule n'est pas adéquate pour prédire l'apparition de la maladie. Cela motive la nécessité d'augmenter l'information offerte par la génomique avec les autres technologies omiques pour obtenir des informations au niveau fonctionnel des cellules et des tissus tel que le transcriptome, le protéome, le métabolome ou l'épigénome. L'utilisation de ces différents niveaux de l'information biologique permet de mieux cerner le phénotype clinique du patient.

Ce chapitre est une revue systématique présentant les différentes approches omiques et leur apport à la médecine systémique ainsi que leurs applications et les défis inhérents à leur implémentation dans le contexte clinique.

Ce chapitre est présenté sous forme de revue publiée dans le journal *International Journal of Molecular Sciences.*

Article I : *Tebani A, Afonso C, Marret S, Bekri S. Omics-Based Strategies in Precision Medicine: Toward a Paradigm Shift in Inborn Errors of Metabolism Investigations. Int J Mol Sci. 2016 Sep 14;17(9).*

International Journal of
Molecular Sciences

Review

Omics-Based Strategies in Precision Medicine: toward a Paradigm Shift in Inborn Errors of Metabolism Investigations

Abdellah Tebani [1,2,3], Carlos Afonso [3], Stéphane Marret [2,4] and Soumeya Bekri [1,2,*]

[1] Department of Metabolic Biochemistry, Rouen University Hospital, 76031 Rouen, France; abdellah.tebani@chu-rouen.fr
[2] Normandie University, UNIROUEN, INSERM, CHU Rouen, Laboratoire NeoVasc ERI28, 76000 Rouen, France; stephane.marret@chu-rouen.fr
[3] Normandie University, UNIROUEN, INSA Rouen, CNRS, COBRA, 76000 Rouen, France; carlos.afonso@univ-rouen.fr
[4] Department of Neonatal Pediatrics, Intensive Care and Neuropediatrics, Rouen University Hospital, 76031 Rouen, France
* Correspondence: soumeya.bekri@chu-rouen.fr; Tel.: +33-2-32-88-81-24; Fax: +33-2-32-88-83-41

Academic Editor: William Chi-shing Cho
Received: 27 July 2016; Accepted: 7 September 2016; Published: 14 September 2016

Abstract: The rise of technologies that simultaneously measure thousands of data points represents the heart of systems biology. These technologies have had a huge impact on the discovery of next-generation diagnostics, biomarkers, and drugs in the precision medicine era. Systems biology aims to achieve systemic exploration of complex interactions in biological systems. Driven by high-throughput omics technologies and the computational surge, it enables multi-scale and insightful overviews of cells, organisms, and populations. Precision medicine capitalizes on these conceptual and technological advancements and stands on two main pillars: data generation and data modeling. High-throughput omics technologies allow the retrieval of comprehensive and holistic biological information, whereas computational capabilities enable high-dimensional data modeling and, therefore, accessible and user-friendly visualization. Furthermore, bioinformatics has enabled comprehensive multi-omics and clinical data integration for insightful interpretation. Despite their promise, the translation of these technologies into clinically actionable tools has been slow. In this review, we present state-of-the-art multi-omics data analysis strategies in a clinical context. The challenges of omics-based biomarker translation are discussed. Perspectives regarding the use of multi-omics approaches for inborn errors of metabolism (IEM) are presented by introducing a new paradigm shift in addressing IEM investigations in the post-genomic era.

Keywords: omics; next-generation sequencing; mass spectrometry; machine learning; chemometrics; data integration; bioinformatics; biomarkers; inborn errors of metabolism; precision medicine

1. Introduction

Precision medicine (PM) is a disruptive concept that takes into account both individual variability and population characteristics to provide personalized care; this approach widens biological knowledge and explores the great diversity of individuals [1]. PM comprises the customization of healthcare for an individual on the basis of measurements obtained at the individual level. However, it also uses the data and learning retrieved from the rest of the population. Hence, PM relies on both biological individuality and population knowledge to provide tailored healthcare. One of the goals of PM is to use the ever-growing understanding of biology to provide patients with accurate and personalized interventions. All PM strategies include the use of decision-making processes based on biomarker-driven approaches. Genes, gene expression products (i.e., transcripts and proteins), and metabolites are the main biomarker families. Given this molecular diversity of biomarkers, the increase in high-throughput omics technologies offers an amazing opportunity to capture the whole picture of biological systems in a hypothesis-free and unbiased mode. These global strategies are, conceptually, clearly disruptive compared to the current ones, which are mainly hypothesis-driven and, thus, intrinsically reductionist. Holistic investigative methods need to be applied to multiple levels of biological information to deeply understand disease processes.

The prediction of normal and pathological states in patients is based on a dynamic understanding of gene–environment interactions on individual and population scales [2]. The new concept of systems medicine relies on global and integrative approaches for patient care. A biological system can be fully understood only if the space and time scales are considered. Figure 1 gives an overview of the multi-scale perspective of systems medicine.

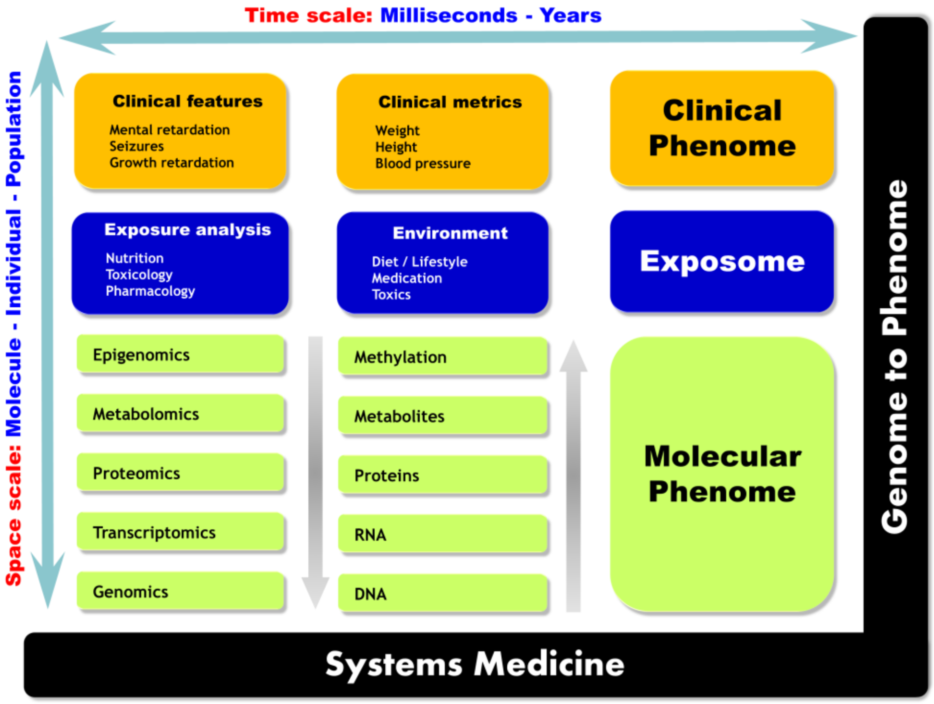


Figure 1. Multi-scale biology overview of systems medicine. Three main drivers define phenotype: (i) the molecular phenome, which is defined by the underlying molecular supports of biological information. The different omics strategies enable to interrogate these supports for information retrieval; (ii) environmental effects spanning from exposures to toxic substances or drugs to diet define the exposome; and (iii) the different clinical metrics used to define the clinical phenome. These different biological and clinical metrics should be approached in a multi-dimensional fashion and should take into account the inherent spatial and temporal scales of both measurement technologies and disease dynamics from the molecular to the population level.

For centuries, biological sciences independently addressed the different parts of life systems and physicians viewed and addressed diseases. Global information retrieval allows contextual pathophysiology understanding of the disease for better diagnosis and treatment [2,3]. Structure, organization, and function descriptions should be considered for a complete understanding of a given biological system. The structure involves basic biomolecules (genes, gene expression products, proteins, and metabolites). The topological connections between these molecules define the organization. The function reflects how the system evolves with regard to metabolic fluxes and environmental stimuli [4,5].

Inborn errors of metabolism (IEM) are an appealing model for systems medicine because the disrupted pathways underlying these diseases have been described at least to some extent. IEM clinical presentations are often non-specific; therefore, appropriate laboratory tests are pivotal for making a diagnosis [6]. However, the widespread routine laboratory diagnosis strategies are mainly represented by sequential investigation assays. This approach is slow and lacks an integrated overview of the generated data. For faster and effective IEM screening and diagnosis, a paradigm shift in investigation strategies is urgently needed. A part of the answer may be found in the new field of systems medicine that capitalizes on omics surge, bioinformatics, and computational advancements to translate the huge amount of data generated by high-throughput omics technologies into effective clinically actionable tools to aid medical decision-making.

In this review, omics technologies that allow holistic biological information retrieval are described. Furthermore, the huge potential of multi-omics data integration strategies within the clinical context is described, as is its role as a key driver for the clinical actionability of omics-based biomarkers. Challenges facing their clinical implementation are then discussed. There is a focus on the relevance of the use of these strategies in IEM.

2. Omics Revolution in Translational and Clinical Contexts

Since the discovery of the DNA structure [7], great advances have been made in understanding genome complexity; these advances have led to sequencing the whole human genome using international endeavors such as the Human Genome Project [8]. Genomics approaches have been widely adopted in biomedical research and have successfully identified the genes and genetic loci involved in the development of human diseases [9–11]. These findings revealed the complexity of biological systems and provided insights for new approaches to disease diagnosis, treatment, and prevention [12–15]. Additionally, other high-throughput omics technologies have been developed to measure other biomolecules, such as epigenomics for epigenetic markers, proteomics for proteins and peptides, and metabolomics for low-molecular-weight metabolites. High-throughput analytical methods allow us to study a large number of omics markers simultaneously. In many ways, omics association studies are similar because they search for omics biomarkers connected with phenotype by unbiased ome-wide screening. Given the uneven maturity of the different omics technologies, genomics seems to be the closest next-generation sequencing (NGS)-based technology introduced into the clinic compared to transcriptomics and epigenomics, which are still promising. Regarding mass spectrometry (MS)-based omics, metabolomics seems to be closer than proteomics to being introduced into clinical practice because metabolite analyses using MS are already routinely adopted in clinical laboratories for drug monitoring and IEM screening. In this review, we mainly focus on mature omics technologies that are actively involved in clinical practice to achieve the promise of PM. However, an overview is also given regarding all omics methods.

2.1. Omics Technologies

2.1.1. High-Throughput Sequencing (HTS) Technologies

Next-generation sequencing (NGS) techniques using a massive parallel sequencing strategy have profoundly changed the clinical genomic landscape. HTS techniques can be classified according to their applications for investigating genomes, epigenomes, or transcriptome. NGS-based strategies that could be used in medical diagnostics vary according to the size of the interrogated genome. These strategies include capturing the few protein-coding regions of a selected panel of genes (tens to hundreds), sequencing of the entire genetic code of a person, which is called whole-genome sequencing (WGS), and sequencing parts of the genome that contain exonic regions, which is called whole-exome sequencing (WES). WGS and WES are used to discover variants associated with a cell function or a disease [16–18]. However, NGS-based transcriptome analysis (RNA-seq) [19] entails quantitative gene expression profiling, whereas epigenomic methods focus on chromatin structure [20].

Genomics

The genome is the complete set of DNA of an organism. This genetic material is mainly found in the nucleus of the human cell (nuclear DNA). Mitochondria contain their own genome (mitochondrial DNA). Fredrick Sanger described the chain-termination strategy to replicate a nucleotide sequence of a DNA fragment (500–1000 bases) [21]. Sanger used chemically modified nucleotide bases and radioactive labeling, along with DNA polymerase, primers, chain-terminating nucleotides, and electrophoresis. Since then, sequencing chemistries evolved using fluorophore-labeled dideoxynucleotides and thermostable DNA polymerases allowed cycled sequencing. Electrophoresis automation and laser detection enhanced the sensitivity of this method [22–24]. The replicated DNA fragments produce signals (electropherogram peaks) related to the nucleotide sequence. Thereafter, these reads undergo an alignment step with a reference genome to identify variants and define their genomic origin. Of note, the Sanger method is still considered the standard method for DNA sequencing accuracy of approximately 1 in 10,000 bases. The first human genome sequence was achieved in 2001 [25,26]. The genome sequences of several model organisms were determined soon thereafter. These endeavors were accomplished with Sanger DNA sequencing, which involves high costs and low throughput. These drawbacks limited the potential of DNA sequencing for healthcare translation. Several HTS technologies were developed soon after the release of the human genome sequence [27] and high-throughput analysis became widely available for genomics. NGS-based platforms provide the ability to replicate, in parallel, many overlapping short DNA fragments (50–500) derived from already prepared libraries. There are different innovative approaches to the special separation of fragments on arrays or beads [8,28,29]. Simultaneous DNA replication of each fragment during the reaction cycles produces billions of short elongations of the DNA sequence. These short stretches are called reads. Hence, each base is synthesized several times. The lowest number of times that each base being monitored is incorporated into an overlapping fragment is called depth of coverage. At the end of the cycle, all the short reads are assembled according to a reference sequence that allows for reconstruction of the original sequence, ranging from a small exon to an entire genome. Innovative high-throughput NGS-based methods have the ability to conserve the genome information and the redundancy of the sequenced genome through their depth of coverage. Different commercial HTS platforms exist. These platforms differ mainly in their sequencing strategies (ligation versus synthesis), amplification by polymerase chain reaction (PCR) of the DNA fragments (flow cell bridge PCR versus bead emulsion PCR), and finally in their adopted targeted approach (PCR amplification versus hybrid capture) [8].

HTS are powerful technologies for personal genome and transcriptome sequencing [8]. Variants can only be interpreted with a good clinical history, family history, and physical examination. These preliminary steps allow physicians to assess whether there are similar or related phenotypes in other family members; if so, then the inheritance pattern can be evaluated. However, clinical validity is the most challenging aspect of NGS. According to the size of the interrogated genomic information, three strategies could be used for diagnosis purposes. Targeted gene sequencing panels are useful tools for analyzing specific genes in a given clinical condition and are widely used in current clinical practice [14,15]. WES focuses on the more functional and informative part of the genome and is being adopted for genetic studies of IEM for gene identification and clinical diagnosis [10,11,30,31]. This approach might shortly replace targeted approaches. WGS provides a unique window to investigate genetic or somatic variations, thus leading to new avenues for exploration of normal and disease phenotypes. However, the inherent data management and interpretation issues hamper its clinical implementation [32]. From a clinical perspective, comparing different genomic diagnostics approaches is of great interest but requires standard and adopted metrics [33]. Sanger sequencing is the gold standard and allows confident calling of genotypes. Because non-inferiority is a prerequisite for clinical adoption of any new medical innovation, Goldfeder et al. recently proposed an interesting metric to quantify the clinical grade reporting standard of sequencing technologies [34].

Epigenomics

Chemical modifications of DNA, histones, non-histone chromatin proteins, and nuclear RNA define the epigenome. These changes affect gene expression without altering the base sequence. Epigenetics usually refers to the structural adaptation of chromosomal regions. These epigenetic marks may be transient or inherited through cell division [35]. They are due to environmental exposures at various developmental stages throughout the life span [36]. The four main actors of epigenetic machinery include DNA methylation, histone modification, microRNA (miRNA) expression and processing, and chromatin condensation [37,38]. Epigenomic modifications depend on spatial and time-related factors. Therefore, they can be tissue-specific in response to environmental or disease-related modifiers. These modifications could regulate gene expression and, thus, affect cell homeostasis. Comprehensive mapping of epigenetic makeup in many cell types and tissues has been reported [39]. Different strategies have been developed to assess the epigenome [20]. Epigenomics methods generally focus on chromatin structure and include histone modification ChIP-seq (chromatin immunoprecipitation sequencing), thus allowing the identification of DNA-associated protein-binding sites [40]. DNase-seq combines DNase I digestion of chromatin with HTS to identify regulatory regions of the genome [41]. DNA methylation [42] and ATAC-seq (assay for transposase-accessible chromatin sequencing) allow the mapping of chromatin accessibility genome-wide [43]. For more technical details, the reader may refer to a recent review [44]. Recently, an epigenome-wide study suggested that interindividual variations in high-density lipoprotein (HDL) particle metabolism rely on epigenome modifications [45].

Transcriptomics

The gene expression pattern in a cell/tissue can broadly reflect its functional state. The transcriptome is the complete set of RNA transcripts, including ribosomal RNA (rRNA), messenger RNA (mRNA) that represents only 1.5 to 2 percent of the transcriptome, transfer RNA (tRNA), miRNA, and other non-coding RNA (ncRNA). Quantitative analyses of the transcriptome can be performed with either microarrays (Chips) or RNA sequencing (RNAseq). Microarrays are based on specific hybridization of RNA transcripts to DNA probes, and HTS-based expression profiling by RNA-seq allows comprehensive qualitative and quantitative mapping of all transcripts [19]. The massively parallel capabilities of HTS have dramatically widened the transcriptional landscape with small quantities of total RNA [46]. Transcriptome-based studies have been applied to some IEM such as McArdle disease [47], Hunter disease [48], lysinuric protein intolerance [49], Lesch–Nyhan disease [50], and Niemann–Pick C disease [51].

2.1.2. Mass Spectrometry-Based Omics

MS analyzers are instruments that weigh molecules and separate them according to their mass-to-charge ratios. There are several MS analyzers with different analytical technologies and, thus, various performance levels regarding resolution, accuracy, throughput and chemical coverage. MS analysis could be semi-quantitative in an untargeted fashion using high-resolution MS instruments or quantitative through targeted analysis using tandem MS [52–54]. MS instruments could be combined with separation methods such as liquid or gas chromatography, capillary electrophoresis, or ion mobility. These combinations aim to enhance the dynamic range, sensitivity, specificity, and chemical coverage [55,56]. Given the chemical diversity of proteins and metabolites and the high sensitivity of this technology, MS has proven its superiority in metabolomics and proteomics.

Proteomics

The proteome consists of all the proteins expressed by a biological system [57]. Posttranslational modifications rely on a highly specialized enzymatic arsenal specific to each cellular type, which leads to the generation of different proteomes from the same genome. These modifications add layers in proteome complexity and, thus, broaden their functionalities [58]. Hence, proteins exhibit different conformation, localization, and interactions depending on space and time factors. The development of proteomics assays is triggered by these complexity challenges. The proteome can mainly be analyzed using MS or protein microarrays [53,59]. However, MS and protein separation allow rapid and accurate detection of hundreds of human proteins and peptides from a small amount of body fluid or tissue [59–61]. Recent studies showed promising results using proteome analysis to explore cystinuria [62], mucopolysaccharidoses [63], and liver mitochondrial functions [64]. Despite increasing analytical performances, proteomics has not been used in routine clinical laboratory practice [65].

Metabolomics

The idea behind metabolomics or metabolic profiling has been empirically used in the past; for example, urine organoleptic characteristics (taste, odor, or color) aided in the diagnosis of medical conditions [66]. The metabolome is defined as the set of metabolites present in a given biological system, fluid, cell, or tissue at a given time [67]. Metabolomics is an omics approach based on biochemical characterizations of the metabolites and their fluctuations related to internal (genetic) and external factors (environment) [68]. The metabolomics approach has been applied in many disease studies [69,70]. MS and nuclear magnetic resonance (NMR) are the main analytical techniques used in metabolomics [71]. However, MS is already adopted in clinical laboratories. New advances in analytical technologies such as ion mobility spectrometry (IMS) combined with high-resolution MS have allowed better coverage of the metabolome [56]. Because IEM are related to metabolism disruption, metabolomics is indicated in assessing these diseases. The future of IEM diagnoses relies on simultaneous quantitative metabolic profiling of many metabolites in biological fluids. Targeted MS-based metabolomics is already widely used and implemented in IEM newborn screening national programs worldwide [72]. Untargeted approaches have also been tested and have shown promising results [73,74]. An integrated strategy for IEM assessment using both targeted and untargeted approaches has been recently proposed by Miller et al. This strategy provides useful and actionable diagnostic information for IEM. The authors have successfully diagnosed 21 IEM disorders using plasma metabolite measurements through metabolomics [75]. Aygen et al. performed a multi-center clinical study in 14 clinical centers in Turkey using NMR-based platforms. The urine samples of 989 neonates were analyzed. A set of specific metabolites that varies in patients compared with healthy individuals was characterized and predictive models were developed. Furthermore, a reference NMR database has been built [74]. For a deeper overview of the potential of metabolomics in IEM investigations, refer to a recent comprehensive review reporting underlying metabolic profiling technologies with limits and advantages and their applications in IEM [76].

2.1.3. Phenomics

Phenome is a term that describes the measurable physical and chemical outcomes of the interactions between genes and the environment that are experienced by individuals and influence their phenotypes [77]. Hence, phenotypes could be retrieved through precise, quantitative analysis [78]. Phenomics, which is a branch of science that explores the basis of how our genes respond to environmental changes, is an emerging and powerful approach to revealing important human attributes at the molecular level. It aims to explain how we adapt and why we are affected by diseases [79]. In other words, phenomics approaches capture our personalized experience with our environment [80–84].

Two main pillars build phenomics: deep phenotyping (DP) and phenomics analysis (PA). DP refers to a strategic and comprehensive approach to data acquisition that includes clinical assessment, laboratory analyses, pathology, and imaging. PA involves the evaluation of patterns and relationships between individuals with related phenotypes and/or between genotype–phenotype associations. PA relies on both clinical data and high-dimensional data integration [85], analysis, and visualization [81,86,87].

In a recent work, Kochinke et al. provided a curated database of 746 currently known genes involved in intellectual disability (ID). The genes were classified according to ID-associated clinical features. This work allowed systematic insights into the clinical and molecular landscapes of ID disorders [88]. Kim et al. introduced the integrative phenotyping framework (iPF) for disease subtype identification. This solution allows accessible visualization of multi-omics data following effective dimension reduction. The strategy has been successfully applied to chronic obstructive lung disease (COPD) [89]. The Monarch initiative is an impressive global endeavor that provides computational tools for genotype–phenotype analysis, genomic diagnostics, and PM across broad areas of disease. Thus, the Monarch initiative illustrates the importance of phenomics [90].

For more details, the reader may refer to a recent review reporting state-of-the art phenome-wide association studies [79].

2.2. *Multi-Omics Strategies, or When the Whole Is More than the Sum of Its Parts*

Although each omics technology is able to measure one family of biomolecules accurately and comprehensively, they are all limited by the functional roles of each type of molecule in a biological system. With the significant advancement of high-throughput technologies and diagnostic techniques described here, the molecular basis of many disorders has been unveiled and their integrative consideration could help solve this issue. However, translation of a patient-specific molecular mechanism into personalized clinical applications remains a challenging task that requires integration of multi-dimensional molecular and clinical data into patient-centric models. For example, family history, clinical history, and physical examination are mandatory for the interpretation of variants and laboratory results. However, in NGS, reporting the result is a very tricky task. NGS test accuracy is at its best when the considered variant in a given gene has been previously associated with the patient’s condition and when a conclusive functional test has revealed the gene’s function abnormalities. Furthermore, few functional studies are available regarding the biological effect of individual variants. This largely impedes effective and comprehensive interpretation of NGS data. In this regard, PM combining multilayer molecular information and specific clinical phenotypes for a given patient may be an answer to this limit [1]. Applying the PM concept to omics and clinical data is a challenging and exciting task. This integrative view of disease modeling is an emerging knowledge-based paradigm in translational and clinical research that capitalizes on the ever-growing power of computational methods to collect, store, integrate, model, and interpret curated disease information across multi-scale biology from molecules to phenotypes [85,91]. With the tremendous amount of available biological and clinical data, the development of appropriate data mining tools is mandatory to extract the hidden information, thereby allowing its translation into actionable clinical tools [91,92]. As technologies keep evolving and datasets grow in volume, velocity, and variety in the big data era, a strong informatics infrastructure will be essential to embrace the PM promise of improved healthcare derived from personal data. Different computational solutions using machine learning and dimension reduction methods have been developed for omics integration [93]. Recent studies have shown the potential of multi-omics studies to provide insightful biological inferences [64,94–99] and to help determine definitive diagnoses in the IEM field [10,11].

2.3. *Issues and Limitations of Omics Analysis*

2.3.1. Technical Limitations

Experimental and Analytical Noise

Reproducibility and repeatability are prerequisites for obtaining consistent results [100]. These important validation steps are hampered by the so-called batch effects. In addition, this drawback can be an important confounder in association studies and potentially causes spurious associations unrelated to the outcomes of interest. Multiple technical platforms from different manufacturers are usually available for the same type of omics profiling. For example, multiple versions of microarray and sequencing platforms are available for genomics, transcriptomics, and epigenomics association studies. They usually have different coverage of the sequenced regions [44]. MS platforms for proteomics and metabolomics have different sensitivities and chemical space coverage [61,76]. This is due to the differences in MS analyzer technology in terms of ionization method, resolution power, measurement accuracy, multi-dimensional separation, scan speed, dynamic range, and analysis throughput [55,56,76]. Such technical heterogeneity often makes meta-analysis and data fusion of different omics studies very challenging. Batch effects issues can be handled by using harmonized Standard Operating Procedures (SOPs) [101–104]. Furthermore, using standard quality control (QC) processes and metrics to normalize intra-laboratory and inter-laboratory omics measurement variations [105,106] and applying consistent statistical correction methods [107–109] and appropriate computational tools [110] can address some technical variation issues.

Analytical Accuracy and Clinical Relevance

Historically, genomics has tightly evolved along with reference sequence GRCh38 [111]. The genome contains approximately 20,000 protein-coding genes, and these vary enormously, spanning from eight base pairs (a transfer RNA) to millions of base pairs. For a given gene, the exon number spans from one to hundreds. Furthermore, a gene's GC richness is a great challenge, especially for capture chemistry-based targeted sequencing. This great genome complexity presents challenges for NGS sequencing strategies regarding accuracy, which is the mandatory prerequisite for clinical implementation. For example, given the intrinsic short-reads sequencing strategies of HTS, simple repeats that are shorter than the read could be determined with NGS. However, if the read length is shorter than the repeat stretch, then the size of the repeated region is difficult to define [112]. Another clinically relevant challenge to using short reads is the lack of phase information, which is the parental chromosomal origin. The characterization of compound heterozygosity (two identified variants in the same gene) is challenging and, thus, illustrates this limitation. A variant-calling algorithm solution has been developed to handle such issues [113]. To solve some of these challenges, long-read sequencing strategies such as using either longer molecule barcoding fragments combined with short-read sequencing and in silico assembly [114] or longer molecule direct sequencing may be of interest [115]. Such sequencing strategies may provide a more accurate view of the genome. Chaisson et al. provided seminal evidence for the utility of long-read sequencing to generate high-quality reference genomes. The authors closed euchromatic gaps in the GRCh37 human reference genome using long-read sequencing [115]. Another major drawback of existing NGS strategies is the need for time-consuming library preparation and DNA enrichment. More automation of this step would enhance the workload and turnover and dramatically change the adoption of NGS into the clinical environment, which requires a high standard of accuracy and rapid reporting of results. The use of nanopores is a promising technology that could overcome this limitation by directly sequencing DNA fragments by passing through nanopores using either nanophotonic chambers [116] or a protein nanopore [117].

Regarding MS-based omics, there are still great challenges regarding their widespread use in the clinical environment. For metabolomics, the great drawback is still metabolite identification, particularly for untargeted approaches [118]. Accurate curated spectral repositories are essential to their clinical adoption and compliance with regulatory issues. Furthermore, harmonization of data reporting and data visualization in clinically accessible formats limits their clinical implementations [76]. Proteomic analysis is technically challenging and has major drawbacks due to splice variants and post-translational modifications. The post-translational modifications interfere with DNA and RNA measurements of protein level predictions [119,120]. Better proteomic measurements require unbiased identification and quantification of proteins by direct measurements using methods analyzing their unique structure, mass, and charge with high specificity [106,121,122]. Furthermore, subtle changes in the detection of low-abundance proteins, which are often important in early-stage disease screening, might be affected by MS sensitivity limits. To overcome this limitation, different approaches have been proposed. One approach is immunocapture enrichment of low-abundance proteins prior to their MS detection [123] at the expense of additional steps in the analytical process, which may affect the throughput. The MS-based omics community is aware of these limitations and actively strives to overcome them [54,104,124].

Omics Informatics Pipelines in the Clinical Environment

A bioinformatics pipeline is a sequential series of computationally complex data analysis processes spanning from raw data retrieval to final results output. The series includes processing, data analysis, and interrogation of reference databases. Two main pipelines are discussed here, NGS and MS-based pipelines. Figure 2 represents a schematic overview of both pipelines.

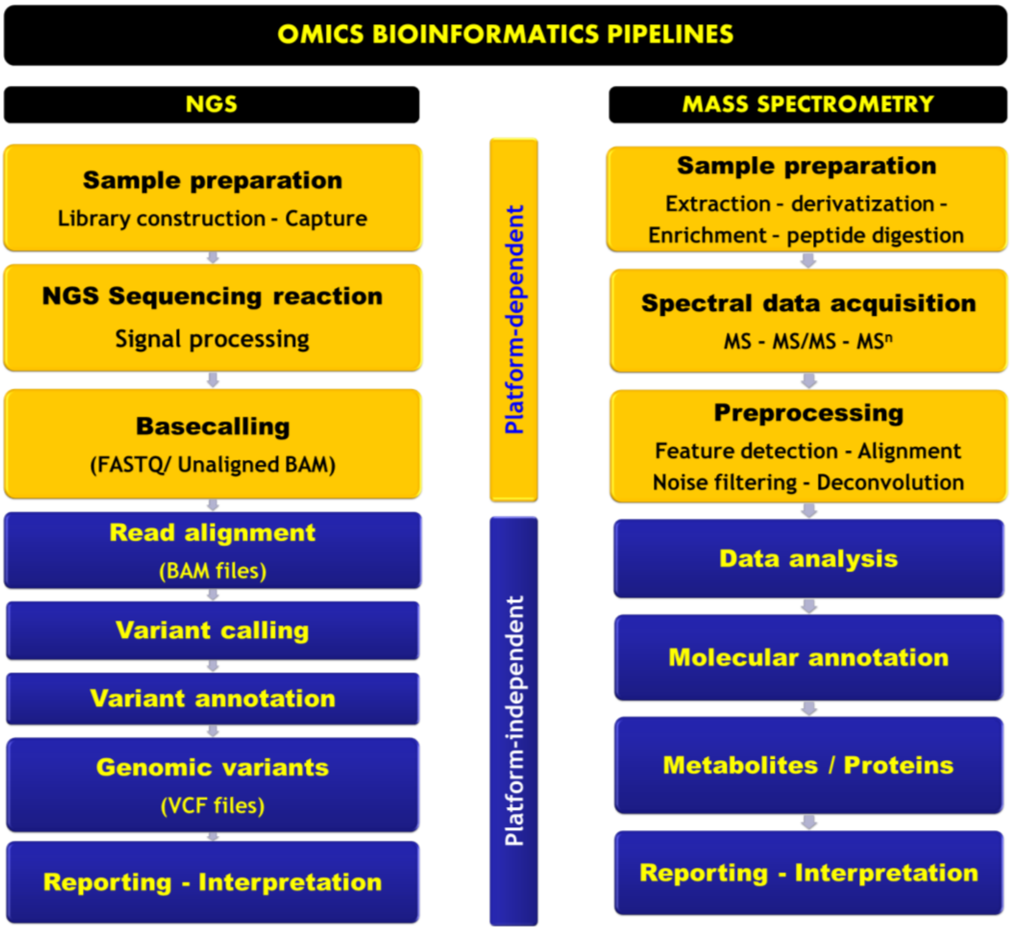


Figure 2. Schematic illustration of bioinformatics pipelines in next-generation sequencing (NGS) and mass spectrometry (MS)-based omics. **Left**: The NGS pipeline comprises library construction and capture, sequencing reaction, and signal processing. Then, a base-calling step is performed to define the unaligned nucleotide sequence. The data are stored in FASTAQ file format containing quality scores. Subsequently, read alignment to a reference sequence is performed, followed by variant calling and annotation. The final output is a list of variants in VCF format for visualization and interpretation; **Right**: MS pipeline starts with sample preparation, depending on the MS instruments and the combined separation method. Data acquisition is performed according to the chosen mode (full scan or tandem MS). Subsequently, a pre-processing step is needed for feature extraction and data cleaning. The result is a list of features that will undergo data analysis, molecular annotation, and identification before biological interpretation. Signal processing is platform-dependent in NGS; however, open source solutions are available for pre-processing MS data.

NGS Informatics Pipeline

Millions of reads are generated by most common NGS platforms using short reads that overlap either the whole genome (WGS) or a specific region (targeted sequencing). An NGS pipeline includes platform-specific software to generate the sequence derived from the primary instrument signal; this step is called base calling. Subsequently, alignment is performed against a reference human genome sequence of the overlapping reads. Several alignment algorithms have been developed with different performance results regarding sequence variation detection [125,126]. The aligned reads are used as input files for single-nucleotide variant (SNV) detection, copy number variation (CNV), indels, and large rearrangements using open source or commercial tools. This step is called variant calling. Several variant callers are available, such as Atlas [127], MuTect [128], VarScan2 [129], and Genome Analysis Toolkit (GATK) [130]. It should be noted that these variant callers exhibit different performances depending on the platform and variant types used [131]. Thus, using different variant callers for a wider coverage of variants capture is recommended. Further annotation of the detected variant is performed using clinical data, Human Genome Variation Society annotations, genome–phenotype correlation, pathway analysis, and predicted effect (on transcription and translation). The quality and consistency of the interrogated online databases are crucial for this step. To avoid misdiagnoses, interpreting variants should be approached as a dynamic big data problem because these online databases are constantly evolving as disease knowledge evolves [132]. NGS is rapidly making its way into the clinic, and its smooth integration upstream and downstream of a sequencing analysis is becoming an important issue. Informatics challenges facing the implementation of NGS in clinical environments range from data acquisition to data reporting, including data validation, data analytics, data storage, and interoperability with already existing laboratory systems and clinical informatics infrastructures. Sample tracking and workflow management logistics are the core of any clinical grade laboratory; this should also be true for NGS. However, uncommon but important downstream offline steps should be consistently tracked such as nucleic acids extraction, library preparation, sequencing runs along with upstream steps such as bioinformatics analysis, quality assurance documentation, data interpretation, and results reporting. All these steps add complexity and error sources to the workflow. To overcome interoperability problems that face in-house custom solutions, such as fragmenting of the workflow, the ideal informatics solution should be fully integrated with the laboratory information system (LIS) so it is able to track samples from order receipt to results reporting. Of note, the generated NGS data range from 10 GB for WES to 150 GB for WGS. Hence, data storage solutions need to be addressed before implementation. Data analysis challenges not only include the computationally heavy burden of the NGS bioinformatics pipeline but also involve handling the huge amount of background data related to wet laboratory steps, sample meta-data, sample processing and tracking, reports, and QC data. With all the high-dimensional data management issues, NGS clinical implementation should be approached with big data analytics solutions [133].

Mass Spectrometry-Based Omics Informatics Pipeline

MS-based processing methods involve four main steps: (i) data acquisition; (ii) data pre-processing; (iii) data analysis using chemometrics; and (iv) identification, network, and pathway analysis [76]. Data files are acquired with proprietary software depending on each platform. Various proprietary data formats have been developed by MS manufacturers to handle MS data, but this raised sharing and processing limits between platforms. To address this problem, open formats have been developed such as netCDF, mzDATA, mzXML, and mzML [134]. Data pre-processing include peak detection, peak alignment, which is a drift time correction step in separative methods (gas chromatography-MS, liquid chromatography-MS, capillary electrophoresis-MS, and ion mobility-MS). During alignment for untargeted analysis, it is crucial to match peaks corresponding to the same analytes in different samples. Subsequently, baseline correction and spectral deconvolution for visualization are performed. Depending on the algorithm used, the order of these steps might be different [135]. The output of these steps is a matrix containing feature concentration or intensity across the different samples. Different output formats such as txt, csv, or an Excel spreadsheet could be used. Subsequently, before data analysis and modeling, different filters, transformations, and normalization methods could be applied to the generated matrix to handle the noise and clean the data. Then, various pattern recognition and machine learning techniques are applied to extract the important features (metabolites or proteins) for the next identification step and pathway and network analysis [76]. MS-based bioinformatics pipeline challenges are the same ones described for big data scaling and interoperability issues with Laboratory Information System (LIS) in a clinical environment. However, some limitations are specific to these platforms in particular, such as sample extraction and/or derivatization, which are offline processes that should be consistently tracked. The metabolite or protein identification steps still lack smooth and streamlined informatics solutions for direct database interrogation. From an informatics perspective, NGS seems to be much more advanced to be included in clinical practice. Therefore, many endeavors are needed to enhance MS informatics infrastructures to a clinical grade [136], and some initiatives have already begun [104,137,138].

2.3.2. Biological Variation

Biological variation is another source of discrepancies in omics studies. Except for genetic profiles being identical across tissues and cell types, all other omics profiles depend on sample type. Tissue and cell-type specificity lead to two important issues in multi-omics approaches: tissue and cell type selection and heterogeneity of tissues. The most accessible specimen in human samples is peripheral blood. Blood-based specimens such as plasma, serum, and leukocytes are commonly used in omics studies. Although the use of blood as a surrogate tissue is sometimes relevant, the biological relevance of blood omics profiles may not be apparent for many human diseases. Using blood-based specimens is a convenient start for searching novel disease-related biomarkers; however, using blood as a surrogate tissue requires cautious validation and interpretation to unravel disease mechanisms [139–141]. Furthermore, diet, circadian rhythm, and drugs may interfere. Another issue is cell heterogeneity; a tissue sample always involves several cell types, with each having a unique omics profile. Depending on the location of a tissue sample or the individual physiological condition, the proportions of the different cell types can change substantially. Statistical methods have been developed to adjust for potential confounding effects due to cell-type heterogeneity [142–144]. However, measuring the omics profile of each purified cell type is an ideal solution that could directly infer the molecular mechanism of a disease [145].

3. Omics and Biomarkers: From Bench to Bedside

3.1. Definitions

A biomarker has been defined as a trait that can be objectively measured and evaluated; therefore, it can be used as an indicator of biological processes (normal versus disease) or of pharmacologic response upon a therapeutic intervention [146]. The FDA defined a biomarker as a measurable endpoint that may be used as an indicator of a disease or physiological state of an organism. According to these definitions, several indicators may be included, such as imaging-based or laboratory-measured biomarkers [147]. The Institute of Medicine Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials defines omics as the study of related sets of biological molecules in a comprehensive fashion. Omics-based tests are defined as “an assay composed of, or derived from, multiple molecular measurements and interpreted by a fully specified computational model to produce a clinically actionable result” [148].

3.2. Biomarker Development

To be used for diagnostics or drug development, an ideal biomarker needs to be highly specific and sensitive [147]. Biomarkers can be classified as pharmacodynamic by indicating the outcome of the interaction between a drug and a target [149] or as prognostic/predictive by stratifying the patient population to responders and non-responders [150]. Another classification that includes three types has been suggested by the Biomarkers and Surrogate End Point Working Group [146,151]: type 0 biomarkers indicate the natural history of disease and correlate with clinical indices, type I biomarkers track the effects of intervention associated with the drug mechanism of action, and type II biomarkers are surrogate end points that predict clinical benefit.

Biomarker development and translational strategies have four main issues that need to be addressed: analytical validity, clinical validity, clinical utility, and regulatory and ethical compliance. Analytical validity includes evidence of assay accuracy, reliability, and reproducibility. Clinical validity denotes evidence regarding the statistical association of biomarkers with the clinical outcome. Clinical utility assesses the benefit of the biomarker in terms of public health. Regulatory and ethical issues address guidelines and requirements compliance of the previous development steps with regulatory bodies and societal challenges, respectively [152]. Figure 3 represents the different pillars of a biomarker discovery pipeline.

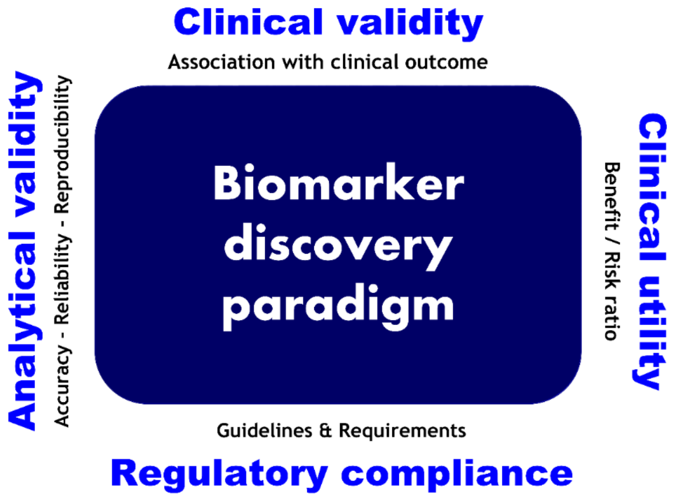

Clinical validity
Association with clinical outcome
Analytical validity
Accuracy - Reliability - Reproducibility
Biomarker discovery paradigm
Clinical utility
Benefit / Risk ratio
Guidelines & Requirements
Regulatory compliance

Figure 3. Biomarker development pipeline milestones.

3.3. Criteria for Omics-Based Biomarkers in Clinical Context

Three main aspects entail omics-based test development: analytical development, computational modeling of the predictor, and its clinical utility validation. Given the multi-dimensional and rich information generated by omics data, mathematical modeling is the key to building classifiers for effective medical decision-making. Because omics data are high-dimensional, machine learning and chemometric methods are needed to obtain insights from the data [91]. These methods may be divided into two main classes: unsupervised and supervised methods [76]. Unsupervised methods are exploratory and track patterns in the data; they include principal component analysis [153], independent component analysis [154], k-means clustering [155], hierarchical cluster analysis [156], and self-organizing maps [157]. Supervised methods are mainly predictive and explanatory. They model the dataset so that the class label of separate validation set samples can be predicted based on a series of mathematical models derived from the original data, namely the training set. Different supervised methods such as PLS discriminant analysis (PLS-DA) [158] and orthogonal PLS-DA (OPLS-DA) [159], as well as support vector machines [160], could be applied. For more details, the reader may refer to a recent review [76]. Figure 4 presents a schematic view of the two main computational modeling strategies using machine learning techniques for omics-based biomarker implementation.

The high-dimensionality characteristic of omics data requires new approaches for omics-based biomarkers development. McShane et al. described the main issues to take into account during omics-based biomarker development, including samples, analytical development of assays, computational model development, clinical utility assessment, and ethical and regulatory issues. The authors suggested criteria that should be assessed for effective biomarker validation [161]. All these steps raise specific challenges regarding validation practices and determine the use of these omics-based tests [100]. A stepwise approach of using machine learning methods for clinical phenotypes prediction and omics-based predictor development spanning from data collection to large-scale clinical validation are presented in Figure 5.

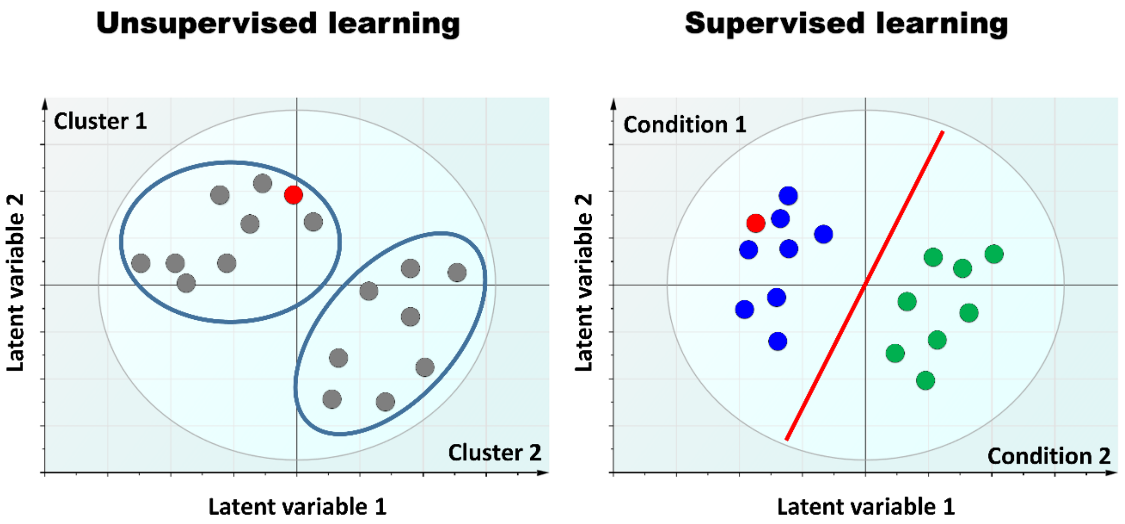
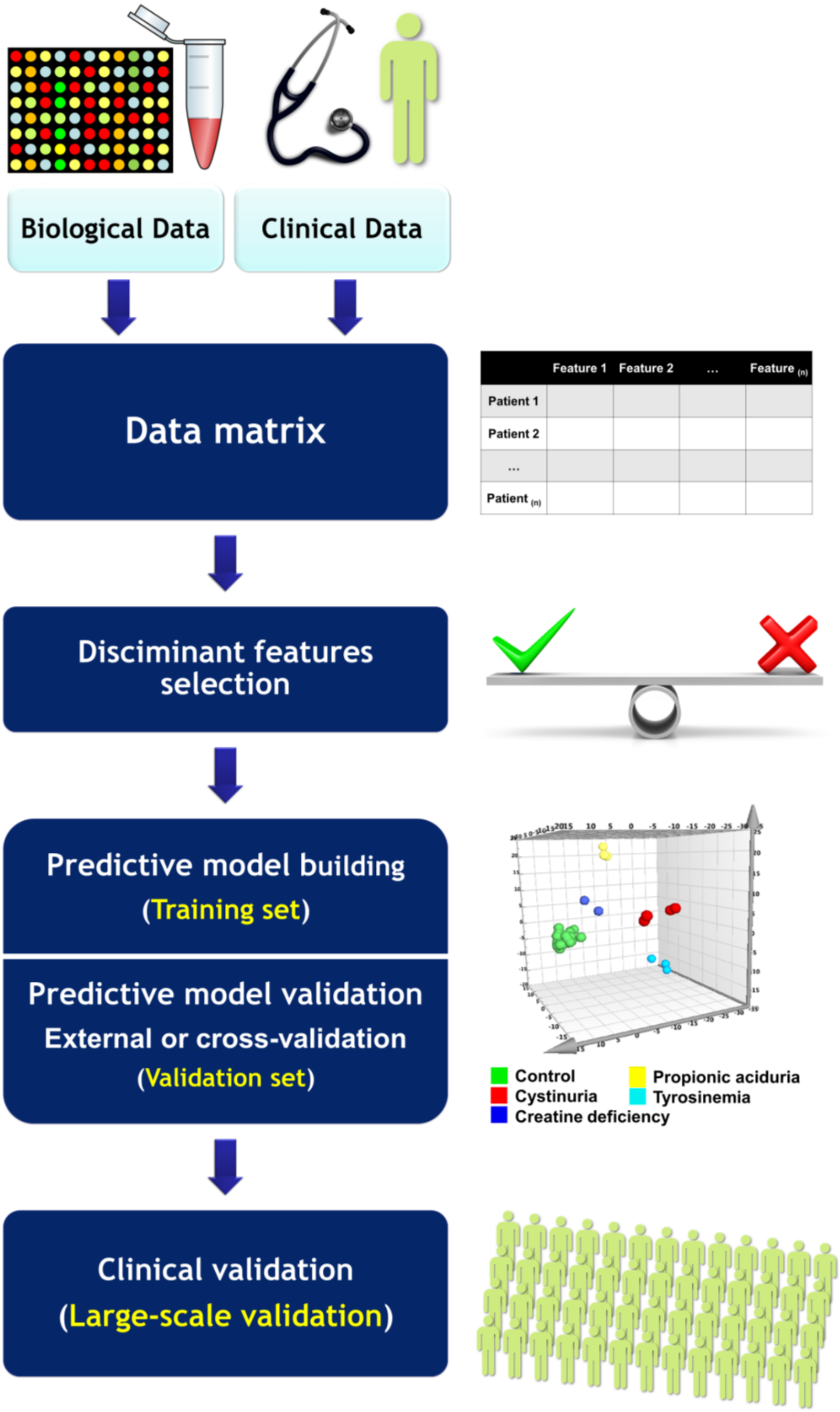


Figure 4. Illustration of the two main machine learning techniques on which omics-based biomarker strategies rely. **Left**: All samples are unlabeled in unsupervised learning. A model separates samples into different clusters based on their biological similarity. A new sample (red circle) is classified according to its similarity to a particular cluster; **Right**: In supervised learning, a training dataset of samples with known class labels is used to build a model (blue circle for condition 1 and green circle for condition 2). The model maximizes the difference between samples from condition 1 and condition 2. Based on this learning, a label for a new sample (red circle) is determined.

3.4. Omics Integration and the Curse of Dimensionality

Biomedical data are becoming quantitatively (number of samples) and qualitatively (data heterogeneity) complex. The number of samples is driven by the ever-growing high throughput of data acquisition technologies and their digitization, whereas heterogeneity entails biological features (biomolecules, diseases) and related metadata (sampling metadata and clinical data). Furthermore, data could be acquired through different platforms, thus adding bias, complexity, and noise. For these issues, machine learning methods are suitable for data modeling and integration [162]. Data integrative methods can holistically analyze multiple data types to provide systems-level biological insights [91]. Dimensionality reduction techniques have been widely used to handle the biomedical big data deluge, but on a large scale they are computationally intensive. To handle these issues, topological data analysis (TDA) methods may help. TDA methods have emerged recently, but the concept goes back to Leonhard Euler and his work with algebraic topology in the 16th century. TDA methods acquire insight from data by analyzing their shapes (patterns) with geometric dimensional conversions [163–165]. These methods have shown good performance in finding hidden patterns when other standard methods fail [95,163,166]. Parsimony phylogenetic analysis is another promising method to handle the omics data deluge [167]. Disease subtype classification for patient stratification is both data-dependent and method-dependent. Thus, it is urgent to have a representative and consistent reference dataset that can be used for the comparison and evaluation of methods.



	Feature 1	Feature 2	...	Feature $_{(n)}$
Patient 1				
Patient 2				
...				
Patient $_{(n)}$				

Figure 5. A stepwise approach to using machine learning methods for the prediction of clinical phenotypes. A training dataset is first collected. Then, a subset of features associated with the phenotype of interest is selected. Based on these features, a multi-variate model is built by the training data. A validation set acquired using the same omics profiling methods is collected and treated as new input to the established multi-variate model. The predictions provided by the model are used to assess the classification performance of the test input by comparing the model output and the actual clinical phenotypes of the patients in the validation set.

4. Perspectives and Challenges in Translational and Clinical Contexts

4.1. Data Integrity, Standardization, and Sharing

Data quality, integrity, and security are the keys to retrieving and maintaining the flow of data and are essential for achieving the promise of "precise" medicine. Data sharing can allow a study to proceed despite the low number of participants, which is often the case in IEM studies. However, the key drivers of data sharing are data and meta-data standards. These are essential for successful data integration and exchange. The lack of such standards or their inconsistent use, especially in omics, are the main drawbacks [102]. Furthermore, in addition to global harmonization, new adapted regulatory approaches for these new omics strategies are urgently needed [168,169].

Large amounts of acquired data raise complex challenges for healthcare stakeholders, including patients. These challenges include the following: (i) sample collection, handling, storage, and transport; (ii) data analyses using multi-omics integration techniques; and (iii) collecting electronic medical record data. The integration of medical record data with biological data and their analysis are other issues. Finally, data sharing within the scientific community raises controversial legal, ethical, and privacy concerns as well [170,171].

4.2. Turning Data into Knowledge

Although molecular biomarkers have helped to unveil the underlying pathophysiological mechanisms of disease, only a few of the currently known biomarkers are clinically actionable [172]. When introducing a biomarker to the clinic, it is important to consider its functional characterization through pathways and network analysis, along with its implementation feasibility in terms of public health. Despite the progress in patient phenotyping and stratification, new methods are needed to address the PM era challenges, including analyses of large data [173], integration of multi-type data [174], and simulation of disease behaviors across multi-scale modeling in space and time [91,175–177].

4.3. Clinical Research Enterprise and Embracing Multi-Disciplinary Sciences

The new omics revolution will play a central role in the post-genomics era of healthcare. To achieve this promise, it is necessary to combine expertise from multiple disciplines, including clinicians, medical laboratory professionals, data scientists, computational biologists, biostatisticians, and lawyers. This observation increases the necessity for new PM teams with new skill sets to develop overlapping expertise for more effective medical interactions across all healthcare partners. Hence, the skill sets of medical professionals need to be diverse; clinical, biological, and computational knowledge to achieve the promises of PM. Training the new generation of the medical workforce to manage and interpret omics data is one solution, and inception of such thinking has already started [178–180]. Clinical bioinformatics provides a bridge between omics sciences and clinical practices [181]. We are facing an urgent need to transform all aspects of the healthcare system.

4.4. Informatics and New Pathways to Clinical Actionability

Informatics research and innovation are key drivers of the science underlying PM [181]. Actionable biomarkers that aid in clinical decision-making will be envisioned by new frameworks to navigate multi-level evidence regarding whether and how a detected molecular abnormality might be a clinically relevant biomarker [11]. Thanks to databases, accurate annotations with contextual and actionable clinical information will enable the emergence of decision support systems to provide intuitive and patient-specific actionable reports [87,182,183]. Urgent areas to be addressed by clinical bioinformatics research may include biomarker discovery, computational phenotyping, and frameworks for evaluating clinical actionability and utility [181,184,185]. Furthermore, standardization and harmonization-related barriers might trap interoperability and integration by making data aggregation a challenging task [87].

5. Paradigm Shift in IEM Investigations

Because IEM are linked to a genetic defect, their current characterization addresses the mutated gene and its products. However, genotype–phenotype correlation is lacking in several IEM, which leads to consideration of the influence of genetic or environmental modifying factors and the impact of an altered pathway on metabolic flux as a whole. These diseases are related to the disruption of specific interactions in a highly organized metabolic network [91,186]. Thus, the impact of a given disruption is not easily predictable [6,187]. Therefore, a functional overview integrating both space and time dimensions is needed to assess the actors of the altered pathway and the potential interactions of each actor [4]. Systemic approaches may address IEM complexity and allow their diagnosis [10,91]. The effectiveness of such approaches has been recently illustrated by van Karnebeek et al. These authors observed a disruption of the *N*-acetylneuraminic acid pathway in patients with severe developmental delay and skeletal dysplasia by using both genomics and metabolomics approaches. As a result, variations in the *NANS* gene encoding the synthase for *N*-acetylneuraminic acid have been identified [10].

Omics-generated data and clinical data integration allow a paradigm shift in IEM handling. An innovative global approach that involves extracting the useful and actionable information may change screening and diagnosis practices. Therefore, a disruptive move from sequential and hypothesis-driven approaches to a global and hypothesis-generating approach is mandatory to embrace the PM era. The core idea of the paradigm shift in the IEM diagnosis workflow is presented in Figure 6.

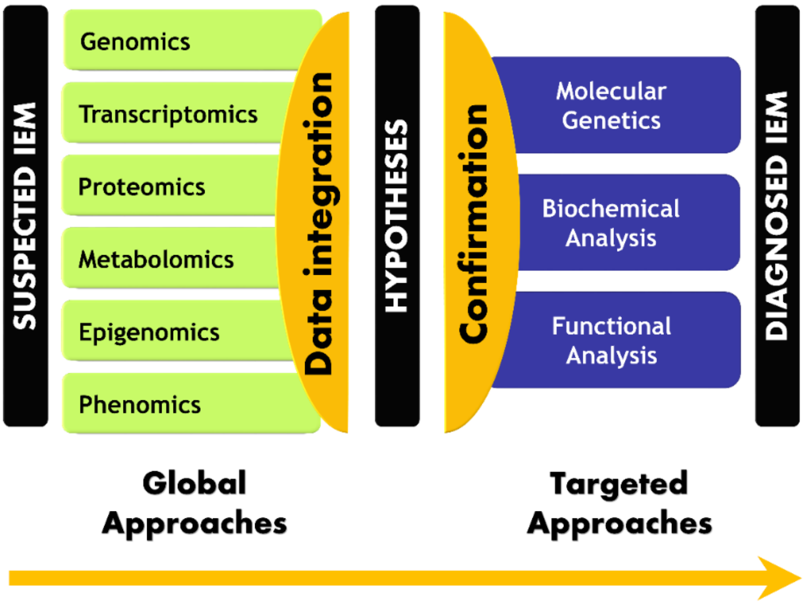


Figure 6. Paradigm shift in Inborn Errors of Metabolism (IEM) diagnosis workflow. Laboratory workflow using high-throughput analytical technologies, integrative bioinformatics, and computational frameworks recovers molecular information for more effective medical decision-making.

6. Conclusions

Current medical practice is being undermined and PM is profoundly reshaping the future of medicine through recent technological advances. Omics technologies are enabling the simultaneous measurement of a huge number of biochemical entities, including genes, genes expressions, proteins, and metabolites. After decades of reductionism, holistic approaches have begun to address inborn errors of metabolism in a systemic fashion [9,64,91]. Despite some existing drawbacks, genomics and metabolomics seem to be taking the lead in the race to get into clinical practice. However, challenges such as data quality/integrity, reproducibility, and study sample sizes have to be addressed. The small number of multi-omics datasets in the field of IEM and the lack of standardized and harmonized protocols affect the wide dissemination of these approaches. To overcome these drawbacks, attention should be given to validation strategies at all stages. Moreover, the development of new analytical and machine learning methods will facilitate analysis of multi-tissue and multi-organ data, thus enabling a real investigation of systemic effects [95,141,163]. Extended and effective resources for biobanking are also essential to ensure consistency. Addressing these challenges will improve healthcare management of IEM by moving from a reactive, targeted, and reductionist approach to a more proactive, global, and integrative one.

Upgrading laboratory informatics infrastructures and a new medical workforce trained in biomedical big data management are necessary for the successful integration of omics-based strategies. However, the potential of these strategies in the investigation of IEM has yet to be unveiled to all IEM stakeholders worldwide. Laboratory workflows with high-quality data acquisition, mining, and visualization are fundamental for fully embracing the four Ps (predictive, personalized, preventive, and participatory) of PM [188] and effectively translating the underlying biological knowledge into clinically actionable tools.

Acknowledgments: This work was supported by Normandy University, the Institut National de la Santé et de la Recherche Médicale (INSERM), the Conseil Régional de Normandie, Labex SynOrg (ANR-11-LABX-0029), and the European Regional Development Fund (ERDF 31708).

Author Contributions: Abdellah Tebani performed the literature review and wrote the manuscript, including tables and figures. Carlos Afonso critically revised and edited the manuscript. Stéphane Marret critically revised and edited the manuscript. Soumeya Bekri conceived the topic under review and critically revised and edited the manuscript. All authors approved the final manuscript.

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

Abbreviations

ATAC-seq	Assay for transposase-accessible chromatin next-generation sequencing
BAM	Binary alignment map
ChIP-seq	Chromatin immunoprecipitation next-generation sequencing
CT	Computerized tomography
DNA	Deoxyribonucleic acid
DNase-seq	DNase I digestion of chromatin combined with next-generation sequencing
FDA	Food and Drug Administration
HTS	High-throughput sequencing
ICA	Independent component analysis
IEM	Inborn errors of metabolism
iPF	Integrative phenotyping framework
miRNA	microRNA
ML	Machine learning
MRI	Magnetic resonance imaging
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
ncRNA	Non-coding RNA
NGS	Next-generation sequencing
OPLSDA	Orthogonal partial least squares discriminant analysis
PCA	Principal component analysis
PLSDA	Partial least squares discriminant analysis
PM	Precision medicine
QC	Quality control
RNA	Ribonucleic acid
rRNA	Ribosome RNA
SAM	Sequence alignment map
SNP	Single-nucleotide polymorphisms
SOM	Self-organizing maps
SOP	Standard operating procedure
SVM	Support vector machines

TDA	Topological data analysis
tRNA	Transfer RNA
VCF	Variant call format
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

References

1. Collins, F.S.; Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **2015**, *372*, 793–795.
2. Ahn, A.C.; Tewari, M.; Poon, C.S.; Phillips, R.S. The limits of reductionism in medicine: Could systems biology offer an alternative? *PLoS Med.* **2006**, *3*, e208.
3. Van Regenmortel, M.H. Reductionism and complexity in molecular biology: Scientists now have the tools to unravel biological and overcome the limitations of reductionism. *EMBO Rep.* **2004**, *5*, 1016–1020.
4. Aon, M.A. Complex systems biology of networks: The riddle and the challenge. In *Systems Biology of Metabolic and Signaling Networks*; Springer Berlin Heidelberg: Heidelberg, Germany, 2014; pp. 19–35.
5. Kitano, H. Systems biology: A brief overview. *Science* **2002**, *295*, 1662–1664.
6. Lanpher, B.; Brunetti-Pierri, N.; Lee, B. Inborn errors of metabolism: The flux from mendelian to complex diseases. *Nat. Rev. Genet.* **2006**, *7*, 449–460.
7. Watson, J.D.; Crick, F.H. The structure of DNA. *Cold Spring Harb. Symp. Quant. Biol.* **1953**, *18*, 123–131.
8. Goodwin, S.; McPherson, J.D.; McCombie, W.R. Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **2016**, *17*, 333–351.
9. Yang, Y.; Muzny, D.M.; Reid, J.G.; Bainbridge, M.N.; Willis, A.; Ward, P.A.; Braxton, A.; Beuten, J.; Xia, F.; Niu, Z.; et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* **2013**, *369*, 1502–1511.
10. Van Karnebeek, C.D.; Bonafe, L.; Wen, X.Y.; Tarailo-Graovac, M.; Balzano, S.; Royer-Bertrand, B.; Ashikov, A.; Garavelli, L.; Mammi, I.; Turolla, L.; et al. Nans-mediated synthesis of sialic acid is required for brain and skeletal development. *Nat. Genet.* **2016**, *48*, 777–784.
11. Tarailo-Graovac, M.; Shyr, C.; Ross, C.J.; Horvath, G.A.; Salvarinova, R.; Ye, X.C.; Zhang, L.H.; Bhavsar, A.P.; Lee, J.J.; Drogemoller, B.I.; et al. Exome sequencing and the management of neurometabolic disorders. *N. Engl. J. Med.* **2016**, *374*, 2246–2255.
12. Worthey, E.A.; Mayer, A.N.; Syverson, G.D.; Helbling, D.; Bonacci, B.B.; Decker, B.; Serpe, J.M.; Dasu, T.; Tschannen, M.R.; Veith, R.L.; et al. Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet. Med.* **2011**, *13*, 255–262.
13. Benson, M. Clinical implications of omics and systems medicine: Focus on predictive and individualized treatment. *J. Intern. Med.* **2016**, *279*, 229–240.
14. Yohe, S.; Hauge, A.; Bunjer, K.; Kemmer, T.; Bower, M.; Schomaker, M.; Onsongo, G.; Wilson, J.; Erdmann, J.; Zhou, Y.; et al. Clinical validation of targeted next-generation sequencing for inherited disorders. *Arch. Pathol. Lab. Med.* **2015**, *139*, 204–210.
15. Yubero, D.; Brandi, N.; Ormazabal, A.; Garcia-Cazorla, A.; Perez-Duenas, B.; Campistol, J.; Ribes, A.; Palau, F.; Artuch, R.; Armstrong, J.; et al. Targeted next generation sequencing in patients with inborn errors of metabolism. *PLoS ONE* **2016**, *11*, e0156359.
16. Cirulli, E.T.; Goldstein, D.B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **2010**, *11*, 415–425.
17. Stranneheim, H.; Wedell, A. Exome and genome sequencing: A revolution for the discovery and diagnosis of monogenic disorders. *J. Intern. Med.* **2016**, *279*, 3–15.
18. Meienberg, J.; Zerjavic, K.; Keller, I.; Okoniewski, M.; Patrignani, A.; Ludin, K.; Xu, Z.; Steinmann, B.; Carrel, T.; Rothlisberger, B.; et al. New insights into the performance of human whole-exome capture platforms. *Nucleic Acids Res.* **2015**, *43*, e76.
19. Mortazavi, A.; Williams, B.A.; McCue, K.; Schaeffer, L.; Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **2008**, *5*, 621–628.
20. Mensaert, K.; Denil, S.; Trooskens, G.; van Criekinge, W.; Thas, O.; de Meyer, T. Next-generation technologies and data analytical approaches for epigenomics. *Environ. Mol. Mutagen.* **2014**, *55*, 155–170.
21. Sanger, F.; Nicklen, S.; Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 5463–5467.

22. Marsh, M.; Tu, O.; Dolnik, V.; Roach, D.; Solomon, N.; Bechtol, K.; Smietana, P.; Wang, L.; Li, X.; Cartwright, P.; et al. High-throughput DNA sequencing on a capillary array electrophoresis system. *J. Capill. Electrophor.* **1997**, *4*, 83–89.
23. McBride, L.J.; Koepf, S.M.; Gibbs, R.A.; Salser, W.; Mayrand, P.E.; Hunkapiller, M.W.; Kronick, M.N. Automated DNA sequencing methods involving polymerase chain reaction. *Clin. Chem.* **1989**, *35*, 2196–2201.
24. Prober, J.M.; Trainor, G.L.; Dam, R.J.; Hobbs, F.W.; Robertson, C.W.; Zagursky, R.J.; Cocuzza, A.J.; Jensen, M.A.; Baumeister, K. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **1987**, *238*, 336–341.
25. Venter, J.C.; Adams, M.D.; Myers, E.W.; Li, P.W.; Mural, R.J.; Sutton, G.G.; Smith, H.O.; Yandell, M.; Evans, C.A.; Holt, R.A.; et al. The sequence of the human genome. *Science* **2001**, *291*, 1304–1351.
26. Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921.
27. Reuter, J.A.; Spacek, D.V.; Snyder, M.P. High-throughput sequencing technologies. *Mol. Cell* **2015**, *58*, 586–597.
28. Head, S.R.; Komori, H.K.; LaMere, S.A.; Whisenant, T.; van Nieuwerburgh, F.; Salomon, D.R.; Ordoukhanian, P. Library construction for next-generation sequencing: Overviews and challenges. *Biotechniques* **2014**, *56*, 61.
29. Mardis, E.R. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem.* **2013**, *6*, 287–303.
30. Lim, E.C.; Brett, M.; Lai, A.H.; Lee, S.P.; Tan, E.S.; Jamuar, S.S.; Ng, I.S.; Tan, E.C. Next-generation sequencing using a pre-designed gene panel for the molecular diagnosis of congenital disorders in pediatric patients. *Hum. Genom.* **2015**, *9*, 33.
31. Taylor, R.W.; Pyle, A.; Griffin, H.; Blakely, E.L.; Duff, J.; He, L.; Smertenko, T.; Alston, C.L.; Neeve, V.C.; Best, A.; et al. Use of whole-exome sequencing to determine the genetic basis of multiple mitochondrial respiratory chain complex deficiencies. *JAMA* **2014**, *312*, 68–77.
32. Howard, H.C.; Knoppers, B.M.; Cornel, M.C.; Wright Clayton, E.; Senecal, K.; Borry, P. Whole-genome sequencing in newborn screening? A statement on the continued importance of targeted approaches in newborn screening programmes. *Eur. J. Hum. Genet.* **2015**, *23*, 1593–1600.
33. Ashley, E.A. Towards precision medicine. *Nat. Rev. Genet.* **2016**, *17*, 507–522.
34. Goldfeder, R.L.; Ashley, E.A. A precision metric for clinical genome sequencing. *bioRxiv* **2016**, 051490.
35. Bird, A. Perceptions of epigenetics. *Nature* **2007**, *447*, 396–398.
36. Huang, B.; Jiang, C.; Zhang, R. Epigenetics: The language of the cell? *Epigenomics* **2014**, *6*, 73–88.
37. Sadakierska-Chudy, A.; Filip, M. A comprehensive view of the epigenetic landscape. Part II: Histone post-translational modification, nucleosome level, and chromatin regulation by ncRNAs. *Neurotox. Res.* **2015**, *27*, 172–197.
38. Sadakierska-Chudy, A.; Kostrzewa, R.M.; Filip, M. A comprehensive view of the epigenetic landscape part I: DNA methylation, passive and active DNA demethylation pathways and histone variants. *Neurotox. Res.* **2015**, *27*, 84–97.
39. Kundaje, A.; Meuleman, W.; Ernst, J.; Bilenky, M.; Yen, A.; Heravi-Moussavi, A.; Kheradpour, P.; Zhang, Z.; Wang, J.; Ziller, M.J.; et al. Integrative analysis of 111 reference human epigenomes. *Nature* **2015**, *518*, 317–330.
40. Barski, A.; Cuddapah, S.; Cui, K.; Roh, T.Y.; Schones, D.E.; Wang, Z.; Wei, G.; Chepelev, I.; Zhao, K. High-resolution profiling of histone methylations in the human genome. *Cell* **2007**, *129*, 823–837.
41. Yaragatti, M.; Basilico, C.; Dailey, L. Identification of active transcriptional regulatory modules by the functional assay of DNA from nucleosome-free regions. *Genome Res.* **2008**, *18*, 930–938.
42. Lister, R.; O'Malley, R.C.; Tonti-Filippini, J.; Gregory, B.D.; Berry, C.C.; Millar, A.H.; Ecker, J.R. Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell* **2008**, *133*, 523–536.
43. Buenrostro, J.D.; Wu, B.; Chang, H.Y.; Greenleaf, W.J. Atac-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **2015**, *109*, doi:10.1002/0471142727.mb2129s109.
44. Meyer, C.A.; Liu, X.S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.* **2014**, *15*, 709–721.
45. Guay, S.P.; Voisin, G.; Brisson, D.; Munger, J.; Lamarche, B.; Gaudet, D.; Bouchard, L. Epigenome-wide analysis in familial hypercholesterolemia identified new loci associated with high-density lipoprotein cholesterol concentration. *Epigenomics* **2012**, *4*, 623–639.

46. Wang, Z.; Gerstein, M.; Snyder, M. RNA-seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57–63.
47. Nogales-Gadea, G.; Consuegra-Garcia, I.; Rubio, J.C.; Arenas, J.; Cuadros, M.; Camara, Y.; Torres-Torronteras, J.; Fiuza-Luces, C.; Lucia, A.; Martin, M.A.; et al. A transcriptomic approach to search for novel phenotypic regulators in mcardle disease. *PLoS ONE* **2012**, *7*, e31718.
48. Mazzoccoli, G.; Tomanin, R.; Mazza, T.; D'Avanzo, F.; Salvalaio, M.; Rigon, L.; Zanetti, A.; Pazienza, V.; Francavilla, M.; Giuliani, F.; et al. Circadian transcriptome analysis in human fibroblasts from hunter syndrome and impact of iduronate-2-sulfatase treatment. *BMC Med. Genom.* **2013**, *6*, 37.
49. Tringham, M.; Kurko, J.; Tanner, L.; Tuikkala, J.; Nevalainen, O.S.; Niinikoski, H.; Nanto-Salonen, K.; Hietala, M.; Simell, O.; Mykkanen, J. Exploring the transcriptomic variation caused by the finnish founder mutation of lysinuric protein intolerance (LPI). *Mol. Genet. Metab.* **2012**, *105*, 408–415.
50. Dauphinot, L.; Mockel, L.; Cahu, J.; Jinnah, H.A.; Ledroit, M.; Potier, M.C.; Ceballos-Picot, I. Transcriptomic approach to Lesch–Nyhan disease. *Nucleosides Nucleotides Nucleic Acids* **2014**, *33*, 208–217.
51. Cluzeau, C.V.; Watkins-Chow, D.E.; Fu, R.; Borate, B.; Yanjanin, N.; Dail, M.K.; Davidson, C.D.; Walkley, S.U.; Ory, D.S.; Wassif, C.A.; et al. Microarray expression analysis and identification of serum biomarkers for niemann-pick disease, type c1. *Hum. Mol. Genet.* **2012**, *21*, 3632–3646.
52. Cajka, T.; Fiehn, O. Toward merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics. *Anal. Chem.* **2015**, *88*, 524–545.
53. Scherl, A. Clinical protein mass spectrometry. *Methods* **2015**, *81*, 3–14.
54. Kusebauch, U.; Campbell, D.S.; Deutsch, E.W.; Chu, C.S.; Spicer, D.A.; Brusniak, M.-Y.; Slagel, J.; Sun, Z.; Stevens, J.; Grimes, B.; et al. Human srmatlas: A resource of targeted assays to quantify the complete human proteome. *Cell* **2016**, *166*, 766–778.
55. May, J.C.; McLean, J.A. Advanced multidimensional separations in mass spectrometry: Navigating the big data deluge. *Annu. Rev. Anal. Chem.* **2016**, *9*, 387–409.
56. Tebani, A.; Schmitz-Afonso, I.; Rutledge, D.N.; Gonzalez, B.J.; Bekri, S.; Afonso, C. Optimization of a liquid chromatography ion mobility-mass spectrometry method for untargeted metabolomics using experimental design and multivariate data analysis. *Anal. Chim. Acta* **2016**, *913*, 55–62.
57. James, P. Protein identification in the post-genome era: The rapid rise of proteomics. *Quart. Rev. Biophys.* **1997**, *30*, 279–331.
58. Khoury, G.A.; Baliban, R.C.; Floudas, C.A. Proteome-wide post-translational modification statistics: Frequency analysis and curation of the swiss-prot database. *Sci. Rep.* **2011**, *1*, 90.
59. Betzen, C.; Alhamdani, M.S.S.; Lueong, S.; Schröder, C.; Stang, A.; Hoheisel, J.D. Clinical proteomics: Promises, challenges and limitations of affinity arrays. *Proteom. Clin. Appl.* **2015**, *9*, 342–347.
60. Sabbagh, B.; Mindt, S.; Neumaier, M.; Findeisen, P. Clinical applications of ms-based protein quantification. *Proteom. Clin. Appl.* **2016**, *10*, 323–345.
61. Lassman, M.E.; McAvoy, T.; Chappell, D.L.; Lee, A.Y.; Zhao, X.X.; Laterza, O.F. The clinical utility of mass spectrometry based protein assays. *Clin. Chim. Acta* **2016**, *459*, 155–161.
62. Kovacevic, L.; Lu, H.; Goldfarb, D.S.; Lakshmanan, Y.; Caruso, J.A. Urine proteomic analysis in cystinuric children with renal stones. *J. Pediatr. Urol.* **2015**, *11*, 217.e1–217.e6.
63. Heywood, W.E.; Camuzeaux, S.; Doykov, I.; Patel, N.; Preece, R.L.; Footitt, E.; Cleary, M.; Clayton, P.; Grunewald, S.; Abulhoul, L.; et al. Proteomic discovery and development of a multiplexed targeted mrm-lc-ms/ms assay for urine biomarkers of extracellular matrix disruption in mucopolysaccharidoses I, II, and VI. *Anal. Chem.* **2015**, *87*, 12238–12244.
64. Williams, E.G.; Wu, Y.; Jha, P.; Dubuis, S.; Blattmann, P.; Argmann, C.A.; Houten, S.M.; Amariuta, T.; Wolski, W.; Zamboni, N.; et al. Systems proteomics of liver mitochondria function. *Science* **2016**, *352*, aad0189.
65. Martens, L. Bringing proteomics into the clinic: The need for the field to finally take itself seriously. *Proteom. Clin. Appl.* **2013**, *7*, 388–391.
66. Holmes, E.; Wilson, I.D.; Nicholson, J.K. Metabolic phenotyping in health and disease. *Cell* **2008**, *134*, 714–717.
67. Oliver, S.G.; Winson, M.K.; Kell, D.B.; Baganz, F. Systematic functional analysis of the yeast genome. *Trends Biotechnol.* **1998**, *16*, 373–378.

68. 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.
69. Nicholson, J.K.; Holmes, E.; Kinross, J.M.; Darzi, A.W.; Takats, Z.; Lindon, J.C. Metabolic phenotyping in clinical and surgical environments. *Nature* **2012**, *491*, 384–392.
70. Suhre, K.; Raffler, J.; Kastenmüller, G. Biochemical insights from population studies with genetics and metabolomics. *Arch. Biochem. Biophys.* **2016**, *589*, 168–176.
71. Alonso, A.; Marsal, S.; Julia, A. Analytical methods in untargeted metabolomics: State of the art in 2015. *Front. Bioeng. Biotechnol.* **2015**, *3*, 23.
72. Therrell, B.L.; Padilla, C.D.; Loeber, J.G.; Kneisser, I.; Saadallah, A.; Borrajo, G.J.; Adams, J. Current status of newborn screening worldwide: 2015. *Semin. Perinatol.* **2015**, *39*, 171–187.
73. Denes, J.; Szabo, E.; Robinette, S.L.; Szatmari, I.; Szonyi, L.; Kreuder, J.G.; Rauterberg, E.W.; Takats, Z. Metabonomics of newborn screening dried blood spot samples: A novel approach in the screening and diagnostics of inborn errors of metabolism. *Anal. Chem.* **2012**, *84*, 10113–10120.
74. Aygen, S.; Durr, U.; Hegele, P.; Kunig, J.; Spraul, M.; Schafer, H.; Krings, D.; Cannet, C.; Fang, F.; Schutz, B.; et al. NMR-based screening for inborn errors of metabolism: Initial results from a study on turkish neonates. *JIMD Rep.* **2014**, *16*, 101–111.
75. Miller, M.; Kennedy, A.; Eckhart, A.; Burrage, L.; Wulff, J.; Miller, L.D.; Milburn, M.; Ryals, J.; Beaudet, A.; Sun, Q.; et al. Untargeted metabolomic analysis for the clinical screening of inborn errors of metabolism. *J. Inherit. Metab. Dis.* **2015**, *38*, 1029–1039.
76. Tebani, A.; Abily-Donval, L.; Afonso, C.; Marret, S.; Bekri, S. Clinical metabolomics: The new metabolic window for inborn errors of metabolism investigations in the post-genomic era. *Int. J. Mol. Sci.* **2016**, *17*, 1167.
77. Houle, D.; Govindaraju, D.R.; Omholt, S. Phenomics: The next challenge. *Nat. Rev. Genet.* **2010**, *11*, 855–866.
78. Plomin, R.; Haworth, C.M.; Davis, O.S. Common disorders are quantitative traits. *Nat. Rev. Genet.* **2009**, *10*, 872–878.
79. Bush, W.S.; Oetjens, M.T.; Crawford, D.C. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nat. Rev. Genet.* **2016**, *17*, 129–145.
80. Bilder, R.M.; Sabb, F.W.; Cannon, T.D.; London, E.D.; Jentsch, J.D.; Parker, D.S.; Poldrack, R.A.; Evans, C.; Freimer, N.B. Phenomics: The systematic study of phenotypes on a genome-wide scale. *Neuroscience* **2009**, *164*, 30–42.
81. Freimer, N.; Sabatti, C. The human phenome project. *Nat. Genet.* **2003**, *34*, 15–21.
82. Gerlai, R. Phenomics: Fiction or the future? *Trends Neurosci.* **2002**, *25*, 506–509.
83. Oetting, W.S.; Robinson, P.N.; Greenblatt, M.S.; Cotton, R.G.; Beck, T.; Carey, J.C.; Doelken, S.C.; Girdea, M.; Groza, T.; Hamilton, C.M.; et al. Getting ready for the human phenome project: The 2012 forum of the human variome project. *Hum. Mutat.* **2013**, *34*, 661–666.
84. Groza, T.; Kohler, S.; Moldenhauer, D.; Vasilevsky, N.; Baynam, G.; Zemojtel, T.; Schriml, L.M.; Kibbe, W.A.; Schofield, P.N.; Beck, T.; et al. The human phenotype ontology: Semantic unification of common and rare disease. *Am. J. Hum. Genet.* **2015**, *97*, 111–124.
85. Ritchie, M.D.; Holzinger, E.R.; Li, R.; Pendergrass, S.A.; Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* **2015**, *16*, 85–97.
86. Tracy, R.P. "Deep phenotyping": Characterizing populations in the era of genomics and systems biology. *Curr. Opin. Lipidol.* **2008**, *19*, 151–157.
87. Shameer, K.; Badgeley, M.A.; Miotto, R.; Glicksberg, B.S.; Morgan, J.W.; Dudley, J.T. Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. *Brief. Bioinform.* **2016**, doi:10.1093/bib/bbv118.
88. Kochinke, K.; Zweier, C.; Nijhof, B.; Fenckova, M.; Cizek, P.; Honti, F.; Keerthikumar, S.; Oortveld Merel, A.W.; Kleefstra, T.; Kramer, J.M.; et al. Systematic phenomics analysis deconvolutes genes mutated in intellectual disability into biologically coherent modules. *Am. J. Hum. Genet.* **2016**, *98*, 149–164.
89. Kim, S.; Herazo-Maya, J.D.; Kang, D.D.; Juan-Guardela, B.M.; Tedrow, J.; Martinez, F.J.; Sciurba, F.C.; Tseng, G.C.; Kaminski, N. Integrative phenotyping framework (iPF): Integrative clustering of multiple omics data identifies novel lung disease subphenotypes. *BMC Genom.* **2015**, *16*, 1–11.

90. Mungall, C.J.; Washington, N.L.; Nguyen-Xuan, J.; Condit, C.; Smedley, D.; Kohler, S.; Groza, T.; Shefchek, K.; Hochheiser, H.; Robinson, P.N.; et al. Use of model organism and disease databases to support matchmaking for human disease gene discovery. *Hum. Mutat.* **2015**, *36*, 979–984.
91. Argmann, C.A.; Houten, S.M.; Zhu, J.; Schadt, E.E. A next generation multiscale view of inborn errors of metabolism. *Cell Metab.* **2016**, *23*, 13–26.
92. Gligorijevic, V.; Przulj, N. Methods for biological data integration: Perspectives and challenges. *J. R. Soc. Interface* **2015**, *12*, 112.
93. Wanichthanarak, K.; Fahrmann, J.F.; Grapov, D. Genomic, proteomic, and metabolomic data integration strategies. *Biomark. Insights* **2015**, *10*, 1–6.
94. Wahl, S.; Vogt, S.; Stuckler, F.; Krumsiek, J.; Bartel, J.; Kacprowski, T.; Schramm, K.; Carstensen, M.; Rathmann, W.; Roden, M.; et al. Multi-omic signature of body weight change: Results from a population-based cohort study. *BMC Med.* **2015**, *13*, 48.
95. Liu, W.; Bai, X.; Liu, Y.; Wang, W.; Han, J.; Wang, Q.; Xu, Y.; Zhang, C.; Zhang, S.; Li, X.; et al. Topologically inferring pathway activity toward precise cancer classification via integrating genomic and metabolomic data: Prostate cancer as a case. *Sci. Rep.* **2015**, *5*, 13192.
96. Chen, R.; Mias, G.I.; Li-Pook-Than, J.; Jiang, L.; Lam, Hugo, Y.K.; Chen, R.; Miriami, E.; Karczewski, K.J.; Hariharan, M.; Dewey, F.E.; et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **2012**, *148*, 1293–1307.
97. Bartel, J.; Krumsiek, J.; Schramm, K.; Adamski, J.; Gieger, C.; Herder, C.; Carstensen, M.; Peters, A.; Rathmann, W.; Roden, M.; et al. The human blood metabolome-transcriptome interface. *PLoS Genet.* **2015**, *11*, e1005274.
98. Shin, S.Y.; Fauman, E.B.; Petersen, A.K.; Krumsiek, J.; Santos, R.; Huang, J.; Arnold, M.; Erte, I.; Forgetta, V.; Yang, T.P.; et al. An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **2014**, *46*, 543–550.
99. Petersen, A.K.; Zeilinger, S.; Kastenmuller, G.; Romisch-Margl, W.; Brugger, M.; Peters, A.; Meisinger, C.; Strauch, K.; Hengstenberg, C.; Pagel, P.; et al. Epigenetics meets metabolomics: An epigenome-wide association study with blood serum metabolic traits. *Hum. Mol. Genet.* **2014**, *23*, 534–545.
100. Ioannidis, J.P.; Khoury, M.J. Improving validation practices in "omics" research. *Science* **2011**, *334*, 1230–1232.
101. Kolker, E.; Ozdemir, V.; Martens, L.; Hancock, W.; Anderson, G.; Anderson, N.; Aynacioglu, S.; Baranova, A.; Campagna, S.R.; Chen, R.; et al. Toward more transparent and reproducible omics studies through a common metadata checklist and data publications. *Omics* **2014**, *18*, 10–14.
102. Tenenbaum, J.D.; Sansone, S.A.; Haendel, M. A sea of standards for omics data: Sink or swim? *JAMIA* **2014**, *21*, 200–203.
103. Chitayat, S.; Rudan, J.F. Chapter 10—Phenome centers and global harmonization. In *Metabolic Phenotyping in Personalized and Public Healthcare*; Academic Press: Boston, MA, USA, 2016; pp. 291–315.
104. Rocca-Serra, P.; Salek, R.M.; Arita, M.; Correa, E.; Dayalan, S.; Gonzalez-Beltran, A.; Ebbels, T.; Goodacre, R.; Hastings, J.; Haug, K.; et al. Data standards can boost metabolomics research, and if there is a will, there is a way. *Metabolomics* **2016**, *12*, 14.
105. Dunn, W.B.; Wilson, I.D.; Nicholls, A.W.; Broadhurst, D. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis* **2012**, *4*, 2249–2264.
106. Walzer, M.; Pernas, L.E.; Nasso, S.; Bittremieux, W.; Nahnsen, S.; Kelchtermans, P.; Pichler, P.; van den Toorn, H.W.; Staes, A.; Vandenbussche, J.; et al. Qcml: An exchange format for quality control metrics from mass spectrometry experiments. *Mol. Cell. Proteom.* **2014**, *13*, 1905–1913.
107. Issaq, H.J.; Waybright, T.J.; Veenstra, T.D. Cancer biomarker discovery: Opportunities and pitfalls in analytical methods. *Electrophoresis* **2011**, *32*, 967–975.
108. Jonsson, P.; Wuolikainen, A.; Thysell, E.; Chorell, E.; Stattin, P.; Wikstrom, P.; Antti, H. Constrained randomization and multivariate effect projections improve information extraction and biomarker pattern discovery in metabolomics studies involving dependent samples. *Metabolomics* **2015**, *11*, 1667–1678.
109. Scherer, A. *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*; John Wiley & Sons: Chichester, UK, 2009; Volume 868.

110. Vivian, J.; Rao, A.; Nothaft, F.A.; Ketchum, C.; Armstrong, J.; Novak, A.; Pfeil, J.; Narkizian, J.; Deran, A.D.; Musselman-Brown, A.; et al. Rapid and efficient analysis of 20,000 RNA-Seq samples with toil. *bioRxiv* **2016**, doi:10.1101/062497.
111. Church, D.M.; Schneider, V.A.; Steinberg, K.M.; Schatz, M.C.; Quinlan, A.R.; Chin, C.-S.; Kitts, P.A.; Aken, B.; Marth, G.T.; Hoffman, M.M.; et al. Extending reference assembly models. *Genome Biol.* **2015**, *16*, 1–5.
112. Goldfeder, R.L.; Priest, J.R.; Zook, J.M.; Grove, M.E.; Waggott, D.; Wheeler, M.T.; Salit, M.; Ashley, E.A. Medical implications of technical accuracy in genome sequencing. *Genome Med.* **2016**, *8*, 24.
113. Tewhey, R.; Bansal, V.; Torkamani, A.; Topol, E.J.; Schork, N.J. The importance of phase information for human genomics. *Nat. Rev. Genet.* **2011**, *12*, 215–223.
114. Zheng, G.X.; Lau, B.T.; Schnall-Levin, M.; Jarosz, M.; Bell, J.M.; Hindson, C.M.; Kyriazopoulou-Panagiotopoulou, S.; Masquelier, D.A.; Merrill, L.; Terry, J.M.; et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **2016**, *34*, 303–311.
115. Chaisson, M.J.; Huddleston, J.; Dennis, M.Y.; Sudmant, P.H.; Malig, M.; Hormozdiari, F.; Antonacci, F.; Surti, U.; Sandstrom, R.; Boitano, M. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **2015**, *517*, 608–611.
116. Foquet, M.; Samiee, K.T.; Kong, X.; Chauduri, B.P.; Lundquist, P.M.; Turner, S.W.; Freudenthal, J.; Roitman, D.B. Improved fabrication of zero-mode waveguides for single-molecule detection. *J. Appl. Phys.* **2008**, *103*, 034301.
117. Clarke, J.; Wu, H.-C.; Jayasinghe, L.; Patel, A.; Reid, S.; Bayley, H. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **2009**, *4*, 265–270.
118. Vinaixa, M.; Schymanski, E.L.; Neumann, S.; Navarro, M.; Salek, R.M.; Yanes, O. Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *TrAC Trends Anal. Chem.* **2016**, *78*, 23–35.
119. Wu, L.; Candille, S.I.; Choi, Y.; Xie, D.; Jiang, L.; Li-Pook-Than, J.; Tang, H.; Snyder, M. Variation and genetic control of protein abundance in humans. *Nature* **2013**, *499*, 79–82.
120. Vogel, C.; Marcotte, E.M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **2012**, *13*, 227–232.
121. Bittremieux, W.; Valkenborg, D.; Martens, L.; Laukens, K. Computational quality control tools for mass spectrometry proteomics. *Proteomics* **2016**, doi:10.1002/pmic.201600159.
122. Deutsch, E.W.; Overall, C.M.; van Eyk, J.E.; Baker, M.S.; Paik, Y.-K.; Weintraub, S.T.; Lane, L.; Martens, L.; Vandenbrouck, Y.; Kusebauch, U.; et al. Human proteome project mass spectrometry data interpretation guidelines 2.1. *J. Proteome Res.* **2016**, doi:10.1021/acs.jproteome.6b00392.
123. Whiteaker, J.R.; Zhao, L.; Anderson, L.; Paulovich, A.G. An automated and multiplexed method for high throughput peptide immunoaffinity enrichment and multiple reaction monitoring mass spectrometry-based quantification of protein biomarkers. *Mol. Cell. Proteom.* **2010**, *9*, 184–196.
124. Fehniger, T.E.; Boja, E.S.; Rodriguez, H.; Baker, M.S.; Marko-Varga, G. Four areas of engagement requiring strengthening in modern proteomics today. *J. Proteome Res.* **2014**, *13*, 5310–5318.
125. Shang, J.; Zhu, F.; Vongsangnak, W.; Tang, Y.; Zhang, W.; Shen, B. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. *BioMed Res. Int.* **2014**, *2014*, 309650.
126. Pabinger, S.; Dander, A.; Fischer, M.; Snajder, R.; Sperk, M.; Efremova, M.; Krabichler, B.; Speicher, M.R.; Zschocke, J.; Trajanoski, Z. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.* **2014**, *15*, 256–278.
127. Evani, U.S.; Challis, D.; Yu, J.; Jackson, A.R.; Paithankar, S.; Bainbridge, M.N.; Jakkamsetti, A.; Pham, P.; Coarfa, C.; Milosavljevic, A. Atlas2 cloud: A framework for personal genome analysis in the cloud. *BMC Genom.* **2012**, *13*, S19.
128. Cibulskis, K.; Lawrence, M.S.; Carter, S.L.; Sivachenko, A.; Jaffe, D.; Sougnez, C.; Gabriel, S.; Meyerson, M.; Lander, E.S.; Getz, G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **2013**, *31*, 213–219.
129. Koboldt, D.C.; Zhang, Q.; Larson, D.E.; Shen, D.; McLellan, M.D.; Lin, L.; Miller, C.A.; Mardis, E.R.; Ding, L.; Wilson, R.K. Varscan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **2012**, *22*, 568–576.
130. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M. The genome analysis toolkit: A mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2010**, *20*, 1297–1303.

131. Spencer, D.H.; Tyagi, M.; Vallania, F.; Bredemeyer, A.J.; Pfeifer, J.D.; Mitra, R.D.; Duncavage, E.J. Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. *J. Mol. Diagn.* **2014**, *16*, 75–88.
132. Manrai, A.K.; Funke, B.H.; Rehm, H.L.; Olesen, M.S.; Maron, B.A.; Szolovits, P.; Margulies, D.M.; Loscalzo, J.; Kohane, I.S. Genetic misdiagnoses and the potential for health disparities. *N. Engl. J. Med.* **2016**, *375*, 655–665.
133. Gullapalli, R.R.; Desai, K.V.; Santana-Santos, L.; Kant, J.A.; Becich, M.J. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. *J. Pathol. Inform.* **2012**, *3*, 40.
134. Deutsch, E.W. File formats commonly used in mass spectrometry proteomics. *Mol. Cell. Proteom.* **2012**, *11*, 1612–1621.
135. Misra, B.B.; van der Hooft, J.J. Updates in metabolomics tools and resources: 2014–2015. *Electrophoresis* **2016**, *37*, 86–110.
136. Annesley, T.; Diamandis, E.; Bachmann, L.; Hanash, S.; Hart, B.; Javahery, R.; Singh, R.; Smith, R. A spectrum of views on clinical mass spectrometry. *Clin. Chem.* **2016**, *62*, 30–36.
137. Lathrop, J.T.; Jeffery, D.A.; Shea, Y.R.; Scholl, P.F.; Chan, M.M. US food and drug administration perspectives on clinical mass spectrometry. *Clin. Chem.* **2016**, *62*, 41–47.
138. Levin, N.; Salek, R.M.; Steinbeck, C. Chapter 11—From databases to big data. In *Metabolic Phenotyping in Personalized and Public Healthcare*; Academic Press: Boston, MA, USA, 2016; pp. 317–331.
139. GTEx Consortium. The genotype-tissue expression (GTEX) pilot analysis: Multitissue gene regulation in humans. *Science* **2015**, *348*, 648–660.
140. Torell, F.; Bennett, K.; Cereghini, S.; Rannar, S.; Lundstedt-Enkel, K.; Moritz, T.; Haumaitre, C.; Trygg, J.; Lundstedt, T. Multi-organ contribution to the metabolic plasma profile using hierarchical modelling. *PLoS ONE* **2015**, *10*, e0129260.
141. Do, K.T.; Kastenmüller, G.; Mook-Kanamori, D.O.; Yousri, N.A.; Theis, F.J.; Suhre, K.; Krumsiek, J. Network-based approach for analyzing intra- and interfluid metabolite associations in human blood, urine, and saliva. *J. Proteome Res.* **2015**, *14*, 1183–1194.
142. McGregor, K.; Bernatsky, S.; Colmegna, I.; Hudson, M.; Pastinen, T.; Labbe, A.; Greenwood, C.M.T. An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. *Genome Biol.* **2016**, *17*, 1–17.
143. Buettner, F.; Natarajan, K.N.; Casale, F.P.; Proserpio, V.; Scialdone, A.; Theis, F.J.; Teichmann, S.A.; Marioni, J.C.; Stegle, O. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-Sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **2015**, *33*, 155–160.
144. Houseman, E.A.; Molitor, J.; Marsit, C.J. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* **2014**, *30*, 1431–1439.
145. Bock, C.; Farlik, M.; Sheffield, N.C. Multi-omics of single cells: Strategies and applications. *Trends Biotechnol.* **2016**, *34*, 605–608.
146. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.* **2001**, *69*, 89–95.
147. Halim, A.-B. *Biomarkers in Drug Development: A Useful Tool but Discrepant Results May Have a Major Impact*; INTECH Open Access Publisher: Rijeka, Croatia, 2011.
148. Micheel, C.M.; Nass, S.J.; Omenn, G.S. *Evolution of Translational Omics: Lessons Learned and the Path Forward*; National Academies Press: Washington, DC, USA, 2012.
149. Feuerstein, G.; Dormer, C.; Ruffolo, R.; Stiles, G.; Walsh, F.; Rutkowski, J. Translational medicine perspectives of biomarkers in drug discovery and development. Part I. Target selection and validation-biomarkers take center stage. *Int. Drug Discov.* **2007**, *2*, 36–43.
150. Brünner, N. What is the difference between "predictive and prognostic biomarkers"? Can you give some examples. *Connection* **2009**, *13*, 18.
151. Frank, R.; Hargreaves, R. Clinical biomarkers in drug discovery and development. *Nat. Rev. Drug Discov.* **2003**, *2*, 566–580.
152. Horvath, A.R.; Lord, S.J.; StJohn, A.; Sandberg, S.; Cobbaert, C.M.; Lorenz, S.; Monaghan, P.J.; Verhagen-Kamerbeek, W.D.J.; Ebert, C.; Bossuyt, P.M.M. From biomarkers to medical tests: The changing landscape of test evaluation. *Clin. Chim. Acta* **2014**, *427*, 49–57.

153. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 417.
154. Rutledge, D.N.; Bouveresse, D.J.-R. Independent components analysis with the jade algorithm. *TrAC Trends Anal. Chem.* **2013**, *50*, 22–32.
155. Hartigan, J.A.; Wong, M.A. Algorithm as 136: A k-means clustering algorithm. *J. R. Stat. Soc. Ser. C* **1979**, *28*, 100–108.
156. Johnson, S.C. Hierarchical clustering schemes. *Psychometrika* **1967**, *32*, 241–254.
157. Kohonen, T. The self-organizing map. *Proc. IEEE* **1990**, *78*, 1464–1480.
158. Wold, S.; Sjöström, M.; Eriksson, L. Pls-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.
159. Trygg, J.; Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom.* **2002**, *16*, 119–128.
160. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.
161. McShane, L.M.; Cavenagh, M.M.; Lively, T.G.; Eberhard, D.A.; Bigbee, W.L.; Williams, P.M.; Mesirov, J.P.; Polley, M.-Y.C.; Kim, K.Y.; Tricoli, J.V.; et al. Criteria for the use of omics-based predictors in clinical trials. *Nature* **2013**, *502*, 317–320.
162. Satagopam, V.; Gu, W.; Eifes, S.; Gawron, P.; Ostaszewski, M.; Gebel, S.; Barbosa-Silva, A.; Balling, R.; Schneider, R. Integration and visualization of translational medicine data for better understanding of human diseases. *Big Data* **2016**, *4*, 97–108.
163. Offroy, M.; Duponchel, L. Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry. *Anal. Chim. Acta* **2016**, *910*, 1–11.
164. Lum, P.; Singh, G.; Lehman, A.; Ishkanov, T.; Vejdemo-Johansson, M.; Alagappan, M.; Carlsson, J.; Carlsson, G. Extracting insights from the shape of complex data using topology. *Sci. Rep.* **2013**, *3*, 1236.
165. Carlsson, G. Topology and data. *Bull. Am. Math. Soc.* **2009**, *46*, 255–308.
166. Nielson, J.L.; Paquette, J.; Liu, A.W.; Guandique, C.F.; Tovar, C.A.; Inoue, T.; Irvine, K.-A.; Gensel, J.C.; Kloke, J.; Petrossian, T.C.; et al. Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nat. Commun.* **2015**, *6*, 8581.
167. Salazar, J.; Amri, H.; Noursi, D.; Abu-Asab, M. Computational tools for parsimony phylogenetic analysis of omics data. *Omics J. Integr. Biol.* **2015**, *19*, 471–477.
168. Altman, R.B.; Khuri, N.; Salit, M.; Giacomini, K.M. Unmet needs: Research helps regulators do their jobs. *Sci. Transl. Med.* **2015**, *7*, 315ps22.
169. Zerhouni, E.; Hamburg, M. The need for global regulatory harmonization: A public health imperative. *Sci. Transl. Med.* **2016**, *8*, 338ed6.
170. Jiang, X.; Zhao, Y.; Wang, X.; Malin, B.; Wang, S.; Ohno-Machado, L.; Tang, H. A community assessment of privacy preserving techniques for human genomes. *BMC Med. Inform. Decis. Mak.* **2014**, *14*, 1–10.
171. Shoenbill, K.; Fost, N.; Tachinardi, U.; Mendonca, E.A. Genetic data and electronic health records: A discussion of ethical, logistical and technological considerations. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 171–180.
172. Poste, G. Bring on the biomarkers. *Nature* **2011**, *469*, 156–157.
173. Gligorijevic, V.; Malod-Dognin, N.; Przulj, N. Integrative methods for analyzing big data in precision medicine. *Proteomics* **2016**, *16*, 741–758.
174. Li, L.; Cheng, W.Y.; Glicksberg, B.S.; Gottesman, O.; Tamler, R.; Chen, R.; Bottinger, E.P.; Dudley, J.T. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Transl. Med.* **2015**, *7*, 311ra174.
175. Asai, Y.; Abe, T.; Li, L.; Oka, H.; Nomura, T.; Kitano, H. Databases for multilevel biophysiology research available at physiome.jp. *Front. Physiol.* **2015**, *6*, 251.
176. Garny, A.; Cooper, J.; Hunter, P.J. Toward a VPH/physiome toolkit. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2010**, *2*, 134–147.
177. Clancy, C.E.; An, G.; Cannon, W.R.; Liu, Y.; May, E.E.; Ortoleva, P.; Popel, A.S.; Sluka, J.P.; Su, J.; Vicini, P.; et al. Multiscale modeling in the clinic: Drug design and development. *Ann. Biomed. Eng.* **2016**, *44*, 2591–2610.
178. Henricks, W.H.; Karcher, D.S.; Harrison, J.H.; Sinard, J.H.; Riben, M.W.; Boyer, P.J.; Plath, S.; Thompson, A.; Pantanowitz, L. Pathology informatics essentials for residents: A flexible informatics curriculum linked to accreditation council for graduate medical education milestones. *J. Pathol. Inform.* **2016**, *7*, 27.

179. Louis, D.N.; Feldman, M.; Carter, A.B.; Dighe, A.S.; Pfeifer, J.D.; Bry, L.; Almeida, J.S.; Saltz, J.; Braun, J.; Tomaszewski, J.E. Computational pathology: A path ahead. *Arch. Pathol. Lab. Med.* **2015**, *140*, 41–50.
180. Louis, D.N.; Gerber, G.K.; Baron, J.M.; Bry, L.; Dighe, A.S.; Getz, G.; Higgins, J.M.; Kuo, F.C.; Lane, W.J.; Michaelson, J.S. Computational pathology: An emerging definition. *Arch. Pathol. Lab. Med.* **2014**, *138*, 1133–1138.
181. Sirintrapun, S.J.; Zehir, A.; Syed, A.; Gao, J.; Schultz, N.; Cheng, D.T. Translational bioinformatics and clinical research (biomedical) informatics. *Clin. Lab. Med.* **2016**, *36*, 153–181.
182. Miotto, R.; Li, L.; Kidd, B.A.; Dudley, J.T. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* **2016**, *6*, 26094.
183. Soualmia, L.F.; Lecroq, T. Bioinformatics methods and tools to advance clinical care. Findings from the yearbook 2015 section on bioinformatics and translational informatics. *Yearb. Med. Inform.* **2015**, *10*, 170–173.
184. Tenenbaum, J.D.; Avillach, P.; Benham-Hutchins, M.; Breitenstein, M.K.; Crowgey, E.L.; Hoffman, M.A.; Jiang, X.; Madhavan, S.; Mattison, J.E.; Nagarajan, R.; et al. An informatics research agenda to support precision medicine: Seven key areas. *JAMIA* **2016**, *23*, 791–795.
185. Altman, R.B.; Prabhu, S.; Sidow, A.; Zook, J.M.; Goldfeder, R.; Litwack, D.; Ashley, E.; Asimenos, G.; Bustamante, C.D.; Donigan, K.; et al. A research roadmap for next-generation sequencing informatics. *Sci. Transl. Med.* **2016**, *8*, 335ps310.
186. Sahoo, S.; Franzson, L.; Jonsson, J.J.; Thiele, I. A compendium of inborn errors of metabolism mapped onto the human metabolic network. *Mol. BioSyst.* **2012**, *8*, 2545–2558.
187. Cho, D.-Y.; Kim, Y.-A.; Przytycka, T.M. Chapter 5: Network biology approach to complex diseases. *PLoS Comput. Biol.* **2012**, *8*, e1002820.
188. Hood, L.; Balling, R.; Auffray, C. Revolutionizing medicine in the 21st century through systems approaches. *Biotechnol. J.* **2012**, *7*, 992–1001.

CHAPITRE V : LA METABOLOMIQUE : DE LA CHIMIE A LA BIOLOGIE

1. Introduction

L'idée originelle de la métabolomique remonte à la Grèce antique, où les médecins utilisaient les caractéristiques organoleptiques de l'urine à des fins diagnostiques. Le gout de l'urine était utilisé pour détecter le glucose élevé dans le diabète. De telles caractéristiques organoleptiques, chimiques par nature, sont naturellement d'origine métabolique. Le mot métabolome a été utilisé pour la première fois par Olivier et al. en 1998 et est défini comme l'ensemble des métabolites synthétisés par un organisme donné [1]. Ainsi, le métabolome désigne l'ensemble de tous les métabolites présents dans un système, un fluide, une cellule ou un tissu donné. Les métabolites peuvent être définis comme de petites molécules organiques impliquées dans des réactions enzymatiques et la métabolomique est l'une des technologies "omique" basées sur la caractérisation biochimique et moléculaire du métabolome et ses changements liés à des facteurs génétiques, environnementaux et nutritionnels. La métabolomique permet de caractériser les interactions de ces facteurs et d'évaluer de façon systémique les mécanismes biochimiques impliqués dans ces changements. En effet, les métabolites remplissent le critère clé dans la mesure où leur concentration est modulée en réponse à des changements physiologiques et peuvent générer des informations vitales sur les voies biochimiques qui sont modifiées dans les conditions étudiées [2]. La métabolomique a trouvé des applications dans divers domaines biologiques [3]. A l'avenir, il est possible que le diagnostic des EIM fasse appel prioritairement à une analyse métabolomique. L'article II “The New Metabolic Window for Inborn Errors of Metabolism Investigations in the Post-Genomic Era (Tebani et al. *Int. J. Mol. Sci.* **2016**, *17*(7), 1167) » est une revue de l'état de l'art de la métabolomique et ses applications cliniques en particulier dans l'exploration des EIM [2]. Dans le chapitre qui suit sont présentés les aspects conceptuels et techniques d'une analyse métabolomique.

2. Stratégies analytiques et instrumentation

2.1. Préparation des échantillons

Le choix de la méthode de préparation des échantillons est extrêmement important en métabolomique. Cette étape affecte directement la qualité de l'analyse métabolomique sur le plan quantitatif, qualitatif et la qualité de l'interprétation biologique des données obtenues. Une méthode de préparation des échantillons idéale devrait être la moins sélective possible pour assurer une couverture maximale de métabolites et la plus simple et la plus rapide pour éviter toute perte de métabolite et/ou de dégradation au cours la procédure de préparation, adaptable au haut débit, et évidement reproductible.

2.2. Homogénéisation

Préalablement à l'extraction, les matrices biologiques natives sont souvent homogénéisées afin de réduire la variabilité biologique. Bien que diverses méthodes d'homogénéisation mécaniques ont été développées, la cryopulverisation [4] et le broyage avec billes sont particulièrement utiles pour la préparation des échantillons solides (e. g. tissus et culot cellulaire). En fait, la cryopulverization implique la congélation des échantillons à l'aide d'azote liquide suivie d'une pulvérisation mécanique réalisée en utilisant un mortier et un pilon ou la compression en utilisant des pistons inertes. Cette méthode maintient la stabilité de l'échantillon grâce à la température basse et produit des échantillons finement homogénéisés sous la forme de poudre. Une homogénéisation secondaire des échantillons dans une solution contenant des billes inertes permet d'augmenter la surface de contact entre le solvant et l'échantillon, et ainsi l'amélioration de l'efficacité d'extraction.

2.3. Extraction

L'extraction des métabolites à partir de tissus ou d'autres matrices biologiques est une étape clé dans l'analyse métabolomique en général et sa qualité influence directement celle des résultats. Idéalement, l'extraction des métabolites vise à libérer efficacement les métabolites de l'échantillon, éliminer les interférents qui rendent l'analyse difficile (par exemple, des sels et des protéines), rendre l'extrait compatible avec la technique analytique, et, lorsque cela est nécessaire, concentrer les métabolites traces avant l'analyse.

La sélectivité de l'extraction dépendra de l'objectif de l'étude et le protocole d'extraction dépend principalement de la l'échantillon biologique [5]. L'objectif d'extraction diffère entre l'approche globale et ciblée. Dans les analyses globales, une extraction est réalisée pour isoler le maximum de métabolites possible. Pour l'approche ciblée, l'extraction des métabolites spécifiques est utile pour éliminer les composants indésirables de la matrice [6,7]. Pour les métabolites polaires, la précipitation des protéines par le méthanol est la plus répandue [8]. L'extraction de métabolites lipophiles fait intervenir souvent un système biphasique de solvants organiques non miscibles à l'eau. Les solvants d'extraction sont généralement choisis pour optimiser les performances d'extraction et de minimiser l'altération des métabolites d'intérêt à extraire. Les différents protocoles d'extraction par solvant pour l'analyse des lipides comportent généralement un solvant organique primaire, un solvant modificateur et une composante aqueuse (**Tableau 6**). Ces systèmes biphasiques aboutissent à une répartition des métabolites entre les deux phases liquides non miscibles, de sorte que les composés polaires et lipophiles soient répartis entre les phases organique et aqueuse. Il est aussi possible de faire appel à une extraction solide-liquide et enrichir l'extrait de manière sélective avec les analytes cibles [9]. Par ailleurs, certaines matrices telles que les urines peuvent être analysées sans extraction préalable avec une simple centrifugation et dilution le cas échéant [10].

Tableau 6. Principaux protocoles d'extraction (Liste non exhaustive).

Méthode	Solvant organique	Solvant aqueux	Modificateur	Composition (v /v/v)	Ref.
Polson et *al*	Méthanol	Eau	/	1 : 3	[11]
Folch et *al*	Chloroforme	Eau	Méthanol	2 : 1 : 0.6	[12]
Bligh Dyer et *al*	Chloroforme	Eau	Méthanol	1 : 1 : 0.9	[13]
Cequier-Sanchez et *al*	Dichloromethane	Eau	Méthanol	2 : 1 : 0.6	[14]
Smedes et *al*	Cyclohexane	Eau	Isopropanol	10 : 8 : 11	[15-17]
Mataysh et *al*	Methyl-tert-butyl ether	Eau	Méthanol	10 : 3 : 2.5	[18]

2.4. Conservation

Il est essentiel que les échantillons biologiques soient prélevés avec le minimum de stress ou de traumatisme, pour éviter la lipolyse *in vivo*. Tous les tissus, quelle que soit leur origine, devraient idéalement être extraits immédiatement après le prélèvement de l'organisme vivant, de sorte qu'il y ait peu de modifications des composants lipidiques en particulier. En cas de préparation différée, le tissu doit être congelé aussi rapidement que possible, dans de la glace carbonique ou l'azote liquide et stocké dans des récipients scellés. Les lipides sont très labiles et ainsi très sensibles aux conditions de conservation. Par ailleurs, tous les acides gras insaturés peuvent subir une oxydation de leurs doubles liaisons carbone-carbone par le dioxygène. Enfin, les plasmalogènes peuvent être clivés en milieu acide [19]. Un stockage à -80°C à l'abri de la lumière est conseillé.

2.5. Instrumentation

Durant la dernière décennie, les applications de la métabolomique a suscité un intérêt croissant. Domaine interdisciplinaire par excellence, la métabolomique combinent la biologie, la chimie analytique et des outils d'analyse des données avancés pour extraire l'information chimique et biologique pertinente [2]. Les avancées des technologies analytiques et de traitement de données en ont permis un développement remarquable. De nombreuses plates-formes d'analyse ont été proposées, dont la spectrométrie de masse (MS) et la Résonnace Magnétique Nucléaire (RMN) occupent une place de choix.

Ces deux techniques étant complémentaires, leur application parallèle est souvent souhaitable voir indispensable pour une large couverture métabolique. Aujourd'hui, la MS et la RMN représentent plus de 80% de toutes les études de métabolomique publiées [20] (**Figure 6**).

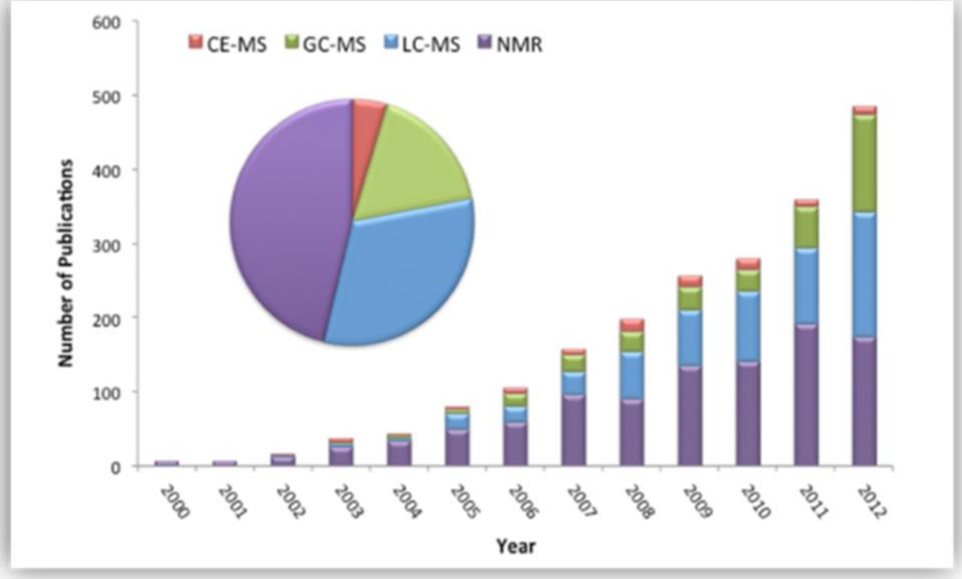


Figure 6. Etude bibliographique des articles de recherche publiés sur la métabolomique. Mot clés : "metabolom * ou metabonom *. RMN (46%), LC/MS (32%), GC/MS (17%), et CE/MS (5%) [20].

2.5.1. La Résonance Magnétique Nucléaire

La Résonance Magnétique Nucléaire (RMN) est la première méthode à être utilisée pour l'analyse globale du métabolome [21]. Elle est basée sur la détection des transitions de spin entre les états énergétiques des noyaux atomiques ayant un moment magnétique non nul (^{1}H, ^{13}C, ^{15}N ou ^{31}P) et placés dans un champ magnétique intense et homogène. L'adoption précoce de la RMN en métabolomique [21-24] découle de sa grande robustesse, son aspect quantitatif, sa reproductibilité et son caractère non destructif et non invasif permettant une analyse rapide des échantillons biologiques avec une préparation d'échantillon minimale. La richesse des spectres en informations structurales en est un atout majeur. Cependant, sa faible sensibilité, comparée à la spectrométrie de masse, en constitue une limite et en justifie la complémentarité analytique souvent mentionnée avec la spectrométrie de masse.

2.5.2. La spectrométrie de masse

La spectrométrie de masse est une méthode d'analyse largement adoptée en bioanalyse. En mesurant le rapport *m/z* des entités chimiques analysées, elle permet la détection simultanée d'analytes multiples avec une sensibilité élevée (micromole voire femtomole). Elle s'est imposée comme une plate-forme puissante et incontournable en métabolomique [6,25,26].

De manière générale, une analyse par spectrométrie de masse se déroule selon les étapes suivantes :

1. **Injection :** l'introduction de l'échantillon dans la source d'ionisation peut être soit directe sans séparation préalable soit après une étape séparative grâce à un couplage avec une méthode séparative chromatographique ou électrophorétique.

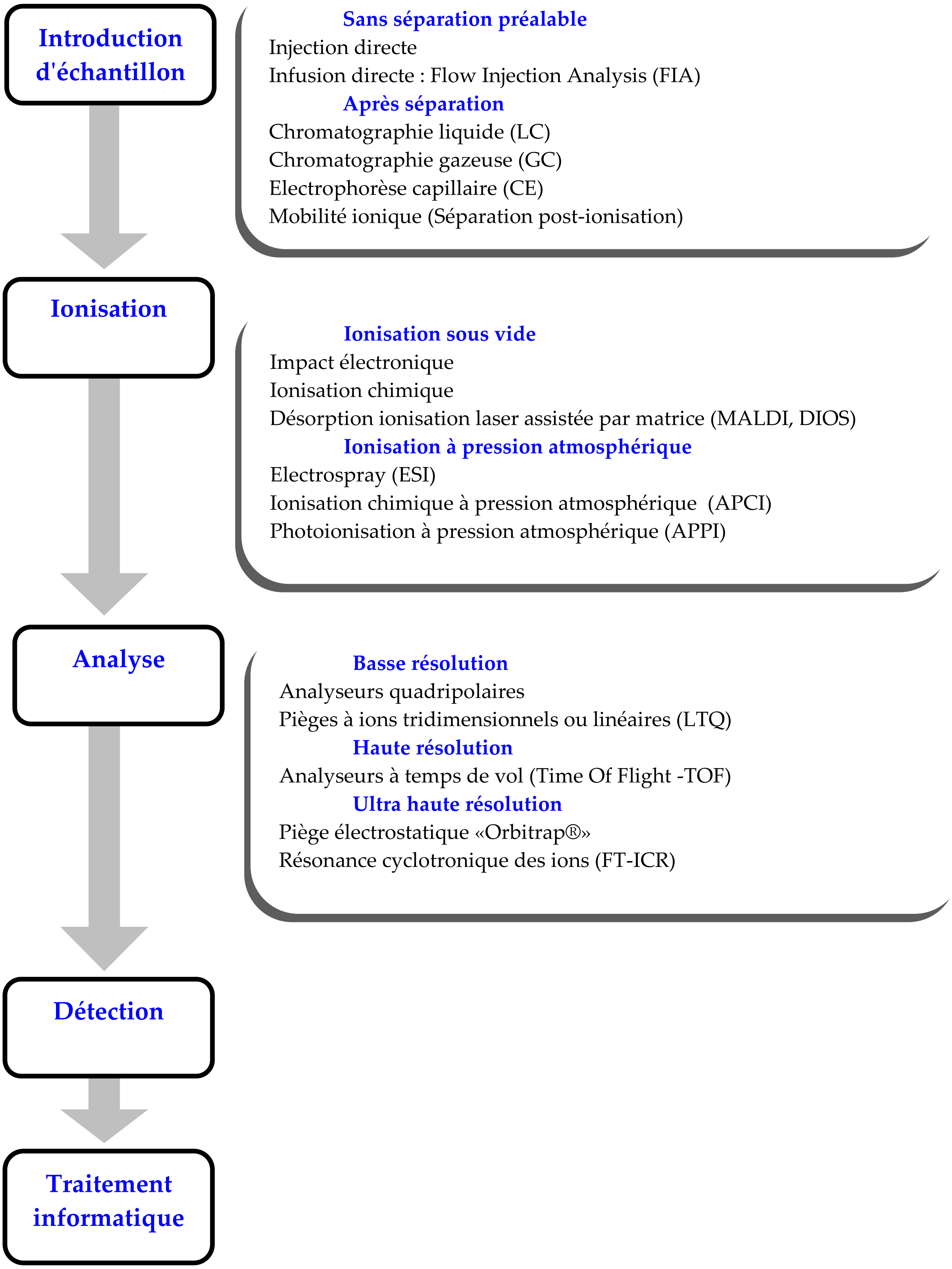
2. **Ionisation** : au niveau de la source, les analytes sont vaporisés et ionisés par diverses méthodes sous vide ou à pression atmosphérique. Certaines méthodes d’ionisation sont très énergiques et induisent des fragmentations, tandis que d'autres sont plus douces et ne produisent que des espèces moléculaires intactes. Les propriétés physico-chimiques de la molécule à analyser sont très importantes à ce stade, car c'est l'étape d'ionisation qui détermine le type d'échantillons qui sera analysé par spectrométrie de masse. Plusieurs modes d'ionisation peuvent être envisagés : electrospray (ESI), ionisation chimique à pression atmosphérique (APCI), photoionisation à pression atmospheric (APPI), désorption/ionisation laser assisté par matrice (MALDI) [25,27-29]. L'ionisation électronique (EI) et l'ionisation par *electrospray* sont couramment utilisées en chromatographie gazeuse (GC-MS) et chromatographie liquide (LC-MS) respectivement.
3. **Analyse** : aussitôt formés, les ions sont extraits de la source puis focalisés et accélérés par des lentilles électroniques, pour accroître leur énergie cinétique. Par la suite, ils sont filtrés suivant leur rapport masse/charge (*m/z*) par l’analyseur de masse. Divers types d'analyseurs de performances différentes sont disponibles. Le Tableau 7 présente une analyse comparative des différents analyseurs.

Tableau 7. Tableau comparatif des différents analyseurs de masse.

	Quadrupole	Ion Trap	Time-of-Flight	Time-of-Flight Reflectron	Magnetic Sector	Q-FTMS	Q-TOF
Précision (ppm)	100	100	200	3-10	<5	0.1-5	3-10
Résolution	4000	4000	8000	15000	30000	100000	10000
Gamme *m/z*	4000	4000	>300,000	10000	10000	10000	10000
Vitesse de balayage	secondes	~ 100 ms	µsecondes	millisecondes	secondes	secondes	secondes
Tandem MS	MS^2 (QQQ)	MS^n	MS	MS^2	MS^2	MS^2	MS^2

4. **Détection** : une fois que les ions sont séparés par l'analyseur de masse, ils atteignent le détecteur d'ions qui génère un signal sous forme de courant électrique à partir des ions incidents. Le détecteur le plus couramment utilisé est le multiplicateur d'électrons associé à une dynode de conversion, qui transfère l'énergie cinétique des ions incidents sur une surface (dynode) qui génère à son tour des électrons secondaires qui sont multipliés par le multiplicateur d’électron avec un gain de 10^6-10^7. Cependant, d'autres approches sont utilisées pour détecter les ions en fonction du type de spectromètre de masse. . Ainsi les instruments à transformés de Fourier, enregistres le signal alternatif induit par les ions lors de leur mouvement qui est numérisé et transformé en spectre de fréquence par la transformée de Fourier. Les instruments à temps de vol utilisent des galettes de microcannaux qui sont des arrangements miniaturisés de multiplicateurs d’électrons.

La **Figure 7** présente les différentes étapes d'une analyse en spectrométrie de masse.



Introduction d'échantillon

Sans séparation préalable

Injection directe

Infusion directe : Flow Injection Analysis (FIA)

Après séparation

Chromatographie liquide (LC)

Chromatographie gazeuse (GC)

Electrophorèse capillaire (CE)

Mobilité ionique (Séparation post-ionisation)

Ionisation

Ionisation sous vide

Impact électronique

Ionisation chimique

Désorption ionisation laser assistée par matrice (MALDI, DIOS)

Ionisation à pression atmosphérique

Electrospray (ESI)

Ionisation chimique à pression atmosphérique (APCI)

Photoionisation à pression atmosphérique (APPI)

Analyse

Basse résolution

Analyseurs quadripolaires

Pièges à ions tridimensionnels ou linéaires (LTQ)

Haute résolution

Analyseurs à temps de vol (Time Of Flight -TOF)

Ultra haute résolution

Piège électrostatique «Orbitrap®»

Résonance cyclotronique des ions (FT-ICR)

Détection

Traitement informatique

Figure 7. Schéma général d'une analyse par spectrométrie de masse. Description des quatre étapes principales : introduction de l'échantillon, ionisation, analyse, détection et traitement des données.

– **Méthodes de spectrométrie de masse par introduction directe**

L'échantillon est introduit directement dans le spectromètre de masse sous forme gazeuse, liquide (infusion directe) ou solide (dépôt sur plaque MALDI-TOF ou sur une canne d'introduction directe) [30]. L'utilisation de l'introduction directe permet des analyses rapides, à haut débit, aussi bien des échantillons natifs que des extraits. Bien que cette technique permette de s'affranchir des biais et de la lourdeur de la préparation d'échantillon, sa limite réside dans l'effet matrice qui est souvent observé [31-33].

– **Couplage de la spectrométrie de masse aux méthodes séparatives**

En raison de la complexité des échantillons biologiques et des effets matrice potentiels, le couplage de la spectrométrie de masse à une méthode séparative permet de s'affranchir, même partiellement, de ces limites. De meilleurs résultats sont ainsi obtenus en spectrométrie de masse lorsqu'elle est est couplée à une étape de séparation appropriée, telle que la chromatographie gazeuse (GC), la chromatographie liquide (LC), l' électrophorèse capillaire (CE) [20,25,34,35] ou la spectrométrie de mobilité ionique [36] qui est une méthode séparative post-ionisation. Le choix de la technique de séparation est dicté par les propriétés physicochimiques des métabolites à analyser et l'application envisagée. La GC-MS est plus appropriée pour les analytes volatiles et semi-volatiles. La LC-MS est adaptée pour les composés en solution liquide, mais ne permet de s'affranchir de l'effet matrice. Pour les analytes chargés, la CE est indiquée. Si l'on considère que la spectrométrie de masse est en soit une méthode séparative, un couplage LC-MS constitue une séparation bidimensionnelle dont la capacité de séparation globale correspond au produit de la capacité de séparation de chacune des méthodes. Outre ce gain en séparation, le couplage de la MS avec une méthode séparative permet de séparer les espèces isomériques qui ne peuvent par définition pas être différentiées par mesure de masse.

2.6. Stratégies analytiques multidimensionnelles en métabolomique

Conceptuellement, l'approche analytique pour appréhender la complexité d'un système consiste à caractériser ces composants et leurs interactions en transformant cette complexité en données informatives et intelligibles. En chimie analytique, il s'agit d'élucider la structure chimique d'un analyte à partir d'un échantillon initialement inconnu ou, inversement, de caractériser un échantillon complexe par les descripteurs chimiques de ses constituants moléculaires. Pour la spectrométrie de masse, l'approche conventionnelle de l'identification moléculaire consiste à utiliser la mesure de masse précise pour réduire les millions de formules moléculaires possibles à quelques centaines voir moins pour une caractérisation initiale [37]. D'autres éléments informationnels orthogonaux sont, ensuite, utilisés de manière complémentaire pour attribuer une identité structurale à l'analyte inconnu et améliorer, ainsi, la sélectivité analytique. Typiquement, une étape initiale de fractionnement et de pré-ionisation, de l'échantillon tel que la chromatographie ou l'électrophorèse capillaire est utilisée pour réduire la complexité, mais également pour éviter les effets de suppression spectrale qui sont inhérents aux sources d'ionisations utilisées en spectrométrie de masse [38]. Cette dernière limite est importante, car les séparations post-ionisation telles que la mobilité ionique ne peuvent pas compenser la nécessité de séparations en phase condensée (électrophorèse ou chromatographie), qui permettent d'atténuer les limitations de la source d'ions elle-même en termes de suppression spectrale due à l'effet matrice. Théoriquement, une fois la mesure de masse précise est obtenue, le nombre de structures possibles peut être réduit de manière significative. A une résolution de 1 ppm avec des règles de filtrage fondées sur l'appariement des profils isotopiques, des règles de chimie et les structures probables [37,39], une formule brute unique peut être attribuée à l'analyte avec une certitude relativement élevée. Marshall et al a démontré qu'une précision de l'ordre de 0,1 mDa (0,2 ppm à 500 Da) est nécessaire pour assigner sans ambiguïté une formule moléculaire basée sur la mesure de masse seule [40]. Ce niveau de précision de masse élevée a été démontré pour le FT-ICR [41] et l'Orbitrap [42]. Aujourd'hui, le développement de spectromètres de masse à transfomé de Fourier permettant de dépasser en routine des résolutions de 106 permettent d'obtenir la structure isotopique fine en séparant les différents isotopologues. Dans ce cas, une attribution non ambiguë des formules brutes peut être obtenue [43] Bien que ces exemples démontrent qu'il est possible d'attribuer une formule moléculaire unique basée sur la mesure de masse précise seule, il est à noter que la formule moléculaire n'est pas un descripteur spécifique de l'analyte et qu'une formule brute peut représenter des milliers d'isomères. Ainsi, pour transformer une formule brute en une information structurale unique, des méthodes qui apportent d'autres informations orthogonales sont indispensables. L'intégration de dimensions de séparation supplémentaires avec la MS est donc nécessaire pour appréhender les difficultés inhérentes à la complexité des échantillons biologiques. De nombreuses combinaisons de séparations en phase condensée et en phase gazeuse avec la MS ont été décrites [44]. Par exemple, le couplage de la chromatographie liquide à la MS nécessite une source d'ions continue à pression ambiante qui permet d'ioniser le flux liquide telle que l'electrospray (ESI) ou l'ionisation chimique à pression atmosphérique (APCI) [45]. D'autre part, l'imagerie par MS est classiquement couplée à un analyseur de masse au moyen d'un faisceau ionique pulsé ou d'une source laser pour fournir une ionisation à grande vitesse avec une résolution spatiale sur la surface de l'échantillon [46]. La spectrométrie de mobilité ionique couplée à la MS est une autre dimension séparative post-ionisation prometteuse dans l'identification structurale [47]. Malgré quelques limitations, de nombreuses combinaisons de séparations à la MS ont été décrites, chacune d'entre elles fournissant un niveau d'information unique [48]. La **Figure 8** présente les performances analytiques en termes de capacité de pics et de production de pics pour les différentes approches multidimensionnelles.

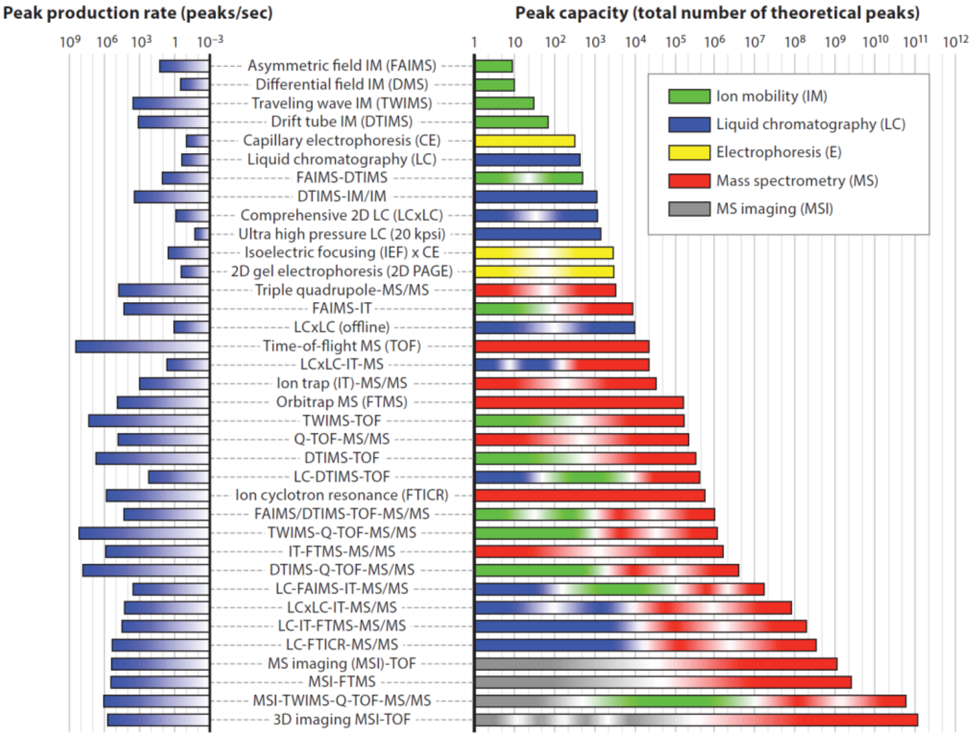


Figure 8. Capacités (à droite) et vitesse de production de pics (à gauche) pour la spectrométrie de masse multidimensionnelle hybride et les techniques connexes. D'après [49].

De façon générale, le couplage multidimensionnel de différentes techniques séparatives nécessite que la résolution obtenue à partir de chaque séparation antérieure soit largement conservée à mesure que les analytes passent aux dimensions suivantes. Ceci est particulièrement difficile lorsque tous les analytes parcourent le même chemin pendant l'analyse, comme c'est le cas pour les techniques tempo-dispersives. Ainsi, la solution consiste à augmenter progressivement la fréquence d'échantillonnage de chaque dimension de dispersion temporelle suivante de sorte que des mesures multiples soient obtenues à l'intérieur d'un intervalle temporel fixe. De cette façon, le temps d'arrivée dans chaque dimension antérieure peut être réassemblé sur la base du signal intégré de dimensions ultérieures. Cette stratégie est couramment utilisée lors du couplage de séparations de phases condensées telles que GC ou LC à la MS. Grâce à cette intercalement temporal, la dimension IMS (échelle de l'ordre de la microseconde) peut être parfaitement nichée entre la dimension chromatographique dont l'échelle temporale est de l'ordre de la minute et celle de la masse avec une échelle de l'ordre de la milliseconde [50]. La **Figure 9** illustre la puissance analytique du couplage des différentes dimensions de séparation qui sont décalées d'un ou plusieurs ordres de grandeur dans le temps. Ce chapitre présente les bases conceptuelles et analytiques des méthodes utilisées dans ce travail de thèse à savoir : la spectrométrie de masse à temps de vol couplé à la chromatographie liquide ultra-haute performance et la mobilité ionique.

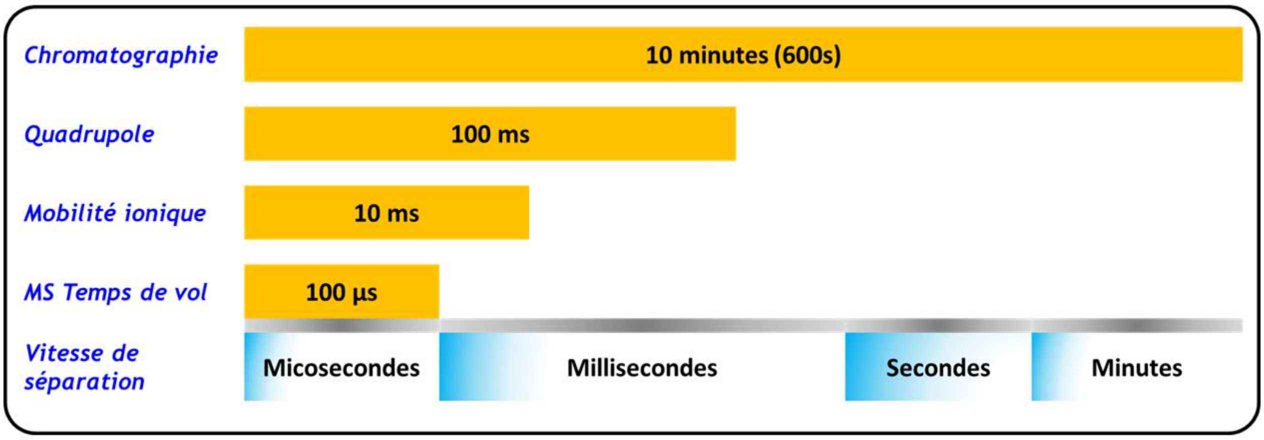


Figure 9. Echelles temporelles analytiques basées sur la vitesse de séparation obtenues en nichant les dimensions de séparation analytique suivantes : chromatographie, quadripôle, mobilité ionique, spectrométrie de masse à temps de vol.

2.7. Méthodologie analytique

2.7.1. Ionisation par *electrospray* (ESI)

Le concept d'ionisation par *electrospray* (ESI) a été introduit par Dole et al. en 1968 [51]. Cependant, le véritable potentiel de l'ESI n'a été mis en évidence dans les années 80, lorsque Fenn a développé l'ESI comme une véritable interface pour la spectrométrie de masse [52]. Fenn a démontré que des ions chargés de façon multiple sont obtenus à partir de grandes biomolécules, telles que des protéines, directement à partir d'une solution permettant, ainsi, leur analyse par spectrométrie de masse avec des analyseurs de gamme *m/z* relativement faible. L'ESI a également permis le couplage de diverses méthodes séparatives tel que la chromatographie liquide (LC) avec la MS. L'apport du développement de *l'electrospray* a été significatif pour la spectrométrie de masse moderne, en biologie en particulier, car il permet la génération d'ions monochargés et/ou à charges multiples à partir de molécules semi-volatiles d'intérêt biologique tels que les protéines et les métabolites [53] d'où le développement fulgurant des différentes sciences omiques basées sur la MS telle que la protéomique et la métabolomique [54]. Le principe de l'ESI est représenté dans la **Figure 10**. En 2002 John Fenn a reçu un quart du prix Nobel de chimie en raison de l'impact significatif de l'electrospray dans le domaine de l'analyse des biomolécules.

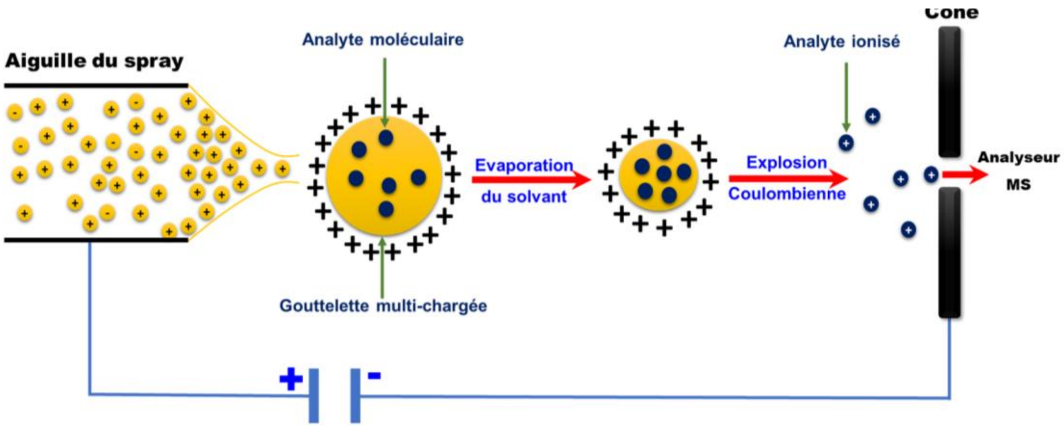


Figure 10. Schéma général du principe de l'ionisation par électrospray. Le soluté à analyser est introduit dans la source à travers un capillaire auquel une tension élevée est appliquée. Un cône de Taylor avec un excès de charge, positive ou négative en fonction du mode, se forme sur sa surface en raison du gradient du champ électrique entre le capillaire et la contre-électrode. Des gouttelettes chargées sont formées à partir de la pointe du cône de Taylor et s'évaporent avant l'entrée dans le spectromètre de masse pour produire des molécules d'analytes libres chargées pouvant être séparées et analysées par leur rapport masse/charge (*m*/*z*).

2.7.2. La spectrométrie de masse haute résolution à temps de vol (TOF)

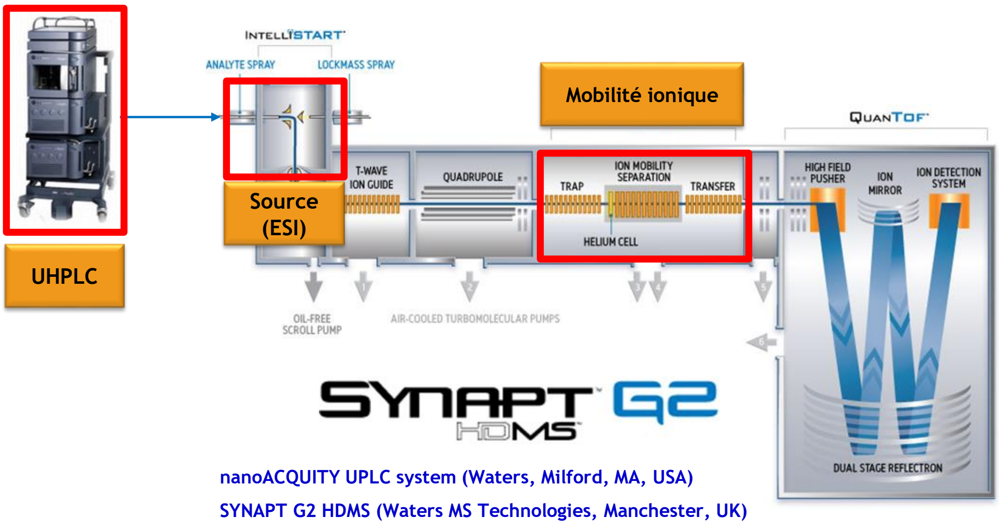
Au cours de la dernière décennie, l'utilisation de la spectrométrie de masse à haute résolution pour la détection moléculaire dans les applications environnementales, chimiques et biologiques à connu un grand essor [53]. Des améliorations remarquables ont été apportées à la performance analytique et à la fonctionnalité des spectromètres de masse actuels, ce qui a sensiblement augmenté la résolution instrumentale, la sensibilité et la vitesse de balayage. Par ailleurs, des avancées dans les logiciels pour l'analyse des données générées ont permis des analyses ciblées et non ciblées à partir d'une seule acquisition [55-58]. Contrairement aux instruments de champ magnétique à haute résolution, où la résolution est souvent limité pour quelques analytes cibles en raison des contraintes de la rapidité de balayage du champ magnétique, les instruments haute résolution à temps de vol (Time of Flight TOF) fonctionnent avec une vitesse de scan plus rapide. Les cadences d'acquisition des analyseurs TOF-MS actuels offrent simultanément une haute résolution, des vitesses de scan élevées et une haute sensibilité, permettant d'obtenir des méthodes d'acquisition riche en information spectrale aussi bien pour les analyses ciblées que non ciblées. Par ailleurs, les données TOF-MS peuvent subir diverses étapes de prétraitement post-acquisition. Il existe de nombreux domaines de recherche dans lesquels les méthodologies d'analyse de données sont utilisées pour comparer entre deux et plusieurs milliers d'analyses en haute résolution à balayage complet afin de déterminer la corrélation non ciblée entre les composés détectés dans les divers échantillons. Ces méthodes de prétraitement seront détaillées plus loin.

Le principe de fonctionnement d'un analyseur de masse TOF utilisé dans ce travail peut être décrit comme suit : en quittant la source d'ionisation par le cône d'échantillonnage (sample cone), les ions entrent dans une zone sous vide primaire puis traversent le cône d'extraction avant d'être pris en charge par un guide d'ion à onde progressive (Travelling-wave ion guide ou TWIG) qui focalisera le faisceau d'ions avant son entrée dans le premier quadripôle. Celui-ci, contrairement aux quadripôles présents dans les TQ, n'effectue pas de balayages en vue de l'acquisition d'un spectre de masse. Il sert principalement de guide d'ion en mode MS et de filtre en mode MS/MS. En sortie du quadripôle, les ions entrent dans la Triwave®, constituée par l'enchaînement d'une cellule TWIG dite trappe, d'une cellule de mobilité ionique et d'une cellule TWIG dite de transfert. Les ions atteignent ensuite l'analyseur à temps de vol orthogonal. Sous l'effet du haut champ électrique perpendiculaire au faisceau d'ions, ces derniers sont injectés dans le TOF avec une distribution d'énergie cinétique réduite. Les paquets d'ions se dirigent vers le réflectron à deux étages qui les renvoient vers le détecteur. La **Figure11** présente un schéma général du SYNAPT G2 HDMS.

L'analyseur en temps de vol peut opérer suivant trois modes nommés ci-après :

- **Sensitivity** : la trajectoire décrite par l'ion est un V. La sensibilité est améliorée, en revanche, ce mode propice à la quantification ne permet pas de dépasser une résolution de 10000.
- **Resolution**: la trajectoire est toujours en V. La résolution obtenue est de 20000. C'est le mode le plus couramment utilisé et c'est celui-ci qui a été utilisé dans cette étude.
- **High Resolution** : l'angle d'injection des ions dans le TOF est modifié de telle sorte que les ions puissent décrire une trajectoire en W avec l'aide d'un deuxième miroir électrostatique situé entre le pusher et le détecteur. Ce mode est le plus résolutif (30000), mais au prix d'une diminution de la transmission et donc de la sensibilité qui est divisée par 10 par rapport au mode « Resolution ».

Le Q-TOF est donc un instrument hybride polyvalent offrant la possibilité de quantifier des composés avec un potentiel dans l'analyse structurale, une résolution en masse de 30 000 et une exactitude inférieure à 3 ppm en routine. Cette performance est possible grâce à l'analyse alternative de l'échantillon et d'un calibrant de masse (Lockspray™), utilisé pour compenser toute dérive de la calibration du TOF au cours de l'analyse.



nanoACQUITY UPLC system (Waters, Milford, MA, USA)
SYNAPT G2 HDMS (Waters MS Technologies, Manchester, UK)

Figure 11. Schéma général du nanoAcquity SYNAPT G2 HDMS (Waters).
LC : chromatographie liquide. TOF : Time of flight.

2.7.3. La chromatographie liquide ultra haute performance (UHPLC)

Alors que les démarches analytiques appliquées en biologie ont été longtemps dominées par les analyses ciblées, l'acquisition des données spectrales globale est de plus en plus nécessaire. Ce changement méthodologique vers l'analyse globale a évolué parallèlement avec les améliorations instrumentales relatives à la vitesse de balayage et de la gamme dynamique ainsi que la sensibilité. Cependant, la complexité des matrices est à l'origine de problèmes analytiques en raison de la diversité quantitative (concentration) et qualitative (métabolites, protéines, xénobiotiques, toxiques) des analytes qu'elles contiennent. Cette complexité soulève un problème majeur quant à l'étendue de la couverture souhaitée et la stratégie analytique choisie. Actuellement, aucune méthode analytique unique ne peut couvrir la totalité de l'espace chimique. De plus, la nature complexe des échantillons biologiques exige une pré-séparation avant l'analyse par spectrométrie de masse pour éviter les effets de charge et de suppression spectrale au niveau de la source. Ainsi, le couplage de dimensions analytiques à performances orthogonales permettrait d'étendre cette couverture. A cet effet, les méthodes chromatographiques ont été largement adoptées par leur apport d'une dimension séparative supplémentaire [59]. Les techniques chromatographiques ont constitué la majeure partie des fondements des sciences séparatives et sont donc au cœur de la chimie analytique. Ces dernières années, la chromatographie liquide haute performance (HPLC) est devenue la méthode de choix pour les séparations et l'analyse des échantillons biologiques [60]. La nécessité d'une séparation de mélanges complexes avec des propriétés physico-chimiques variables, ainsi que les avancées technologiques, ont conduit à l'apparition de variantes de techniques de chromatographie en phase liquide telles que le fluide supercritique ultra-haute pression, LC à interaction hydrophile (HILIC) et chromatographie échangeuse d'ions [61,62]. La polyvalence offerte par les techniques de chromatographie liquide pour l'analyse de composés chimiquement divers est possible grâce choix de phases stationnaires. La sélectivité pour les analytes est obtenue en chromatographie liquide par la sélection des colonnes en conjonction avec la modification de la composition des phases mobiles. La chromatographie liquide est généralement associée à la MS pour l'analyse à haut débit d'échantillons complexes en raison de son pouvoir résolutif élevé, de sa robustesse et de sa facilité de couplage avec l'ESI. La chromatographie liquide en phase inverse est la forme la plus couramment utilisée de séparations LC et est basée sur le principe des interactions hydrophobes entre les analytes et une phase stationnaire apolaire. Une phase aqueuse modérément polaire mobile telle que de l'eau combinée à un solvant organique (méthanol, acétonitrile, etc.) est utilisée pour séparer les analytes en fonction de leur hydrophobicité [59]. Il en résulte une élution rapide de molécules polaires alors que des molécules moins polaires sont retenues. Une élution en gradient peut être appliquée à ce système dans lequel le pourcentage de solvant organique dans l'éluant est graduellement augmenté. L'affinité des analytes hydrophobes vis-à-vis de la phase stationnaire est ainsi réduite, ce qui conduit à des temps d'élution plus rapides pour ces analytes. L'optimisation du pourcentage du solvant organique et du profil de gradient est critique pour assurer une séparation maximale des analytes de tout mélange complexe avant l'analyse spectrale de masse. Ces dernières années ont vu l'évolution de la HPLC traditionnelle à travers des progrès techniques donnant lieu à la chromatographie liquide ultra-haute performance (UHPLC) [60]. L'UHPLC utilise des particules de taille plus petite (typiquement sous 2 µm) ou des particules superficielles poreuses pour une efficacité accrue de la séparation chromatographique. La contre-pression générée par les colonnes UHPLC est inversement proportionnelle à la taille des particules du milieu utilisé pour remplir la colonne. Au cours de ces dernières années, le développement de phases stationnaires a permis d'améliorer les performances chromatographiques et de réduire les temps d'analyse. Ceci est réalisé grâce à des pompes et des instruments de haute performance qui peuvent supporter des contre-pressions élevées obtenues avec par des colonnes UHPLC. Dans les colonnes HPLC, à mesure que la taille des particules diminue à moins de 2,5 µm, il y a une augmentation significative de l'efficacité chromatographique qui ne diminue pas même à des débits de phase mobiles accrus. En raison de l'utilisation d'une plus petite taille de particules, la vitesse et la capacité maximale peuvent être augmentées. L'UHPLC a tiré profit des progrès de la chimie des particules combinés à la instrumentation qui est capable de fonctionner à des pressions plus élevées par rapport à celle utilisée dans les systèmes HPLC. La chromatographie liquide ultra-performante emploie des particules de 2 µm de matériaux tels que l'éthylène hybride ponté (BEH) C18 comme matériau d'emballage (Waters Corporation, Manchester, UK). Ceci a été appelé UHPLC ou UPLC ™ (Waters Corporation, Manchester, UK). Cette technologie est utilisée dans ce travail de thèse.

2.7.4. La mobilité ionique

La spectrométrie par mobilité ionique (IMS) est une technique de séparation en phase gazeuse qui sépare les ions en fonction de leur mobilité relative en présence d'un champ électrique externe et d'un gaz tampon [63]. Le champ électrique (uniforme ou non uniforme) facilite l'accélération des ions, alors que les collisions avec le gaz tampon les décelèrent, donnant lieu à une vitesse de dérive constante [64]. La mobilité d'un ion est fonction de la section efficace de collision (CCS) (Ω) (c'est-à-dire de sa taille et de sa forme), de sa charge ionique et de sa masse réduite [47,65]. La mobilité est déterminée en mesurant la vitesse de dérive de l'ion sous l'influence d'un gradient de champ électrique et en présence d'un gaz tampon dans un tube de dérive. Par conséquent, la mobilité ionique d'un ion (analyte) particulier est caractéristique de l'analyte et du gaz tampon donné. L'IMS peut être utilisée comme technique de séparation sélective pour un composant spécifique à partir d'un mélange. Une mesure de mobilité ionique consiste donc à déterminer le temps nécessaire, appelé temps de dérive (*tD*) qui est souvent de l'ordre de la milliseconde, de chaque ion, pour traverser la cellule de mobilité ionique. L'état de charge des ions prédomine sur la section efficace de collision se traduisant ainsi par des temps de dérive des molécules de bas état de charge plus élevés que les molécules de plus haut état de charge. Pour un même état de charge, les ions seront séparés en fonction de leur section efficace de collision. En effet, lorsque la conformation de l'ion est dépliée, c'est-à-dire sa section efficace de collision est élevée, les collisions avec le gaz tampon seront nombreuses et donc le temps de dérive sera élevé. Par conséquent, un ion de conformation compacte arrivera plus rapidement au détecteur qu'un ion de conformation plus dépliée. Conceptuellement, les techniques d'IMS modernes sont subdivisées en trois classes principales: i) tempo-dispersive, ii) spacio-dispersive et iii) à confinement (piégeage) et libération sélective [47]. Les méthodes de mobilité ionique tempo-dispersives génèrent un spectre de temps d'arrivée, avec tous les ions dérivant le long d'un parcours similaire. Elles incluent DTIMS (drift time ion mobility spectrometer) et TWIMS (Traveling-wave IMS) [66]. Les méthodes de mobilité ionique spatio-dispersives séparent les ions le long de trajectoires de dérive différentes en fonction des différences dans leur mobilité, mais ne confèrent aucune dispersion significative dans le temps. Une caractéristique clé des techniques spatio-dispersives est qu'un voltage est balayé afin d'obtenir un spectre de mobilité ionique à large bande. Celles-ci incluent des techniques de mobilité ionique à haut-bas champ modulées (FAIMS et DMS) [67] analyseurs de mobilité différentielle à champ uniforme (DMA) [68], et la mobilité ionique à modulation transversale [69]. Les méthodes de confinement et de libération des ions piègent les ions au sein d'une région sous pression et éjectent sélectivement ces ions en fonction des différences de mobilité telles que le TIMS [70]. Ces méthodes de mobilité à base de pièges ioniques sont d'apparition récente, car les capacités nécessaires pour contrôler la position des ions sous des conditions de pression élevées n'ont été que récemment maîtrisées. La **Figure 12** représente les différentes stratégies analytiques de la mobilité ionique.

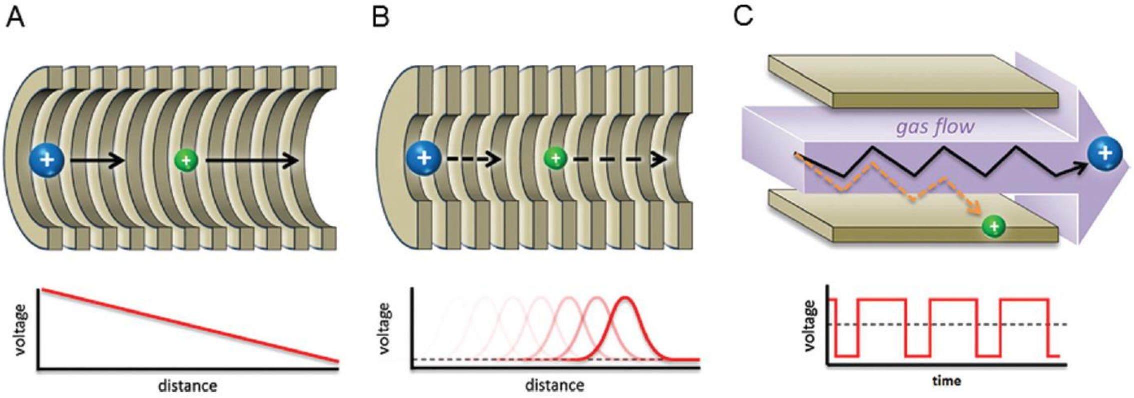


Figure 12. Schéma des différents principes de spectrométrie à mobilité ionique. A) Dans la DTIMS les ions sont introduits dans une chambre remplie de gaz et séparés en fonction de leur dérive différentielle, un potentiel de qui diminue de façon linéaire continue. B) Dans le TWIMS les ions sont sont séparés dans une chambre remplie de gaz en utilisant une le déplacement d'une vague de potentiel. C) Dans la FAIMS les ions sont séparés en fonction de leur migration différentielle orthogonale à un flux de gaz de balayage. Adapté de [66]

Le Synapt G2 HDMS (Waters®) a été utilisé pour ce présent travail. La partie se situant entre le quadripôle et l'analyseur TOF et qui comporte la cellule de mobilité ionique est appelée le *Tri-Wave*. Cette partie du spectromètre de masse est composée de quatre guides d'ions SRIG (stacked ring ion guide), à savoir la cellule d'accumulation (*trap cell*), la cellule d'hélium, la cellule TWIMS et la cellule de transfert (cf. Figure 5). Le Synapt G2 HDMS peut être utilisé selon deux modes, à savoir le mode MS (Q-TOF) et le mode IMS-MS. Les ions vont toujours traverser le *Tri-Wave* quel que soit le mode utilisé. Les principales différences concernent les pressions des gaz dans les différentes cellules du *Tri-Wave*. En mode MS, les pressions sont très faibles dans l'ensemble des cellules constituant le *Tri-Wave*. La bonne transmission des ions à travers cette partie du spectromètre est assurée par des *T-Waves* de très faibles amplitudes propagées à de faibles vitesses. Dans ce cas, la cellule de mobilité ionique sert uniquement à transférer les ions. En mode IMS, le flux d'ions est focalisé à la sortie du quadripôle dans la cellule d'accumulation remplie d'argon à une pression de 10-3 mbar où est appliquée une *T-Wave*. Elle est constituée de 33 paires d'électrodes pour une longueur de 10 cm. Un courant continu est appliqué de façon pulsée sur la dernière électrode permettant ainsi d'accumuler les ions et de les libérer par paquet dans la cellule d'hélium à une pression de 2,5 mbar. Cette cellule d'hélium permet de maintenir une bonne transmission des ions, mais également de refroidir les ions en diminuant leur énergie au centre de masse afin de limiter leur fragmentation dans la cellule de mobilité ionique. La cellule TWIMS du Synapt G2 HDMS est constituée de 79 électrodes pour une longueur de 25,4 cm et la résolution en mobilité ionique est de 40. Une fois séparés en fonction de leur mobilité en phase gazeuse, les ions pénètrent dans la cellule de transfert qui est constituée de 33 paires électrodes pour une longueur de 10 cm et remplie d'argon à une pression de 10-3 mbar. Une *T-Wave* est également appliquée dans cette cellule afin de maintenir la séparation des ions avant leur injection dans l'analyseur TOF. La spectrométrie de mobilité ionique permet la séparation très rapide des ions en phase gazeuse, mais permet également de mesurer les sections efficaces de collision (CCS) des ions qui sont relatives à leur conformation en phase gazeuse. Cette mesure représente l'aire effective qui peut entrer en collision avec une molécule de gaz tampon. La valeur de CCS est ainsi caractéristique d'un ion et est exprimée en Å^2. La valeur de CCS expérimentale peut être comparée à des bases de données [71,72] afin de confirmer une attribution ou encore être comparée à des valeurs de CCS théoriques déterminées à partir de structures théoriques obtenues par modélisation moléculaire. La détermination des CCS doit passer par une étape d'étalonnage en mesurant les temps de dérive de composés références dont la CCS a été déterminée au préalable par principalement dans l'hélium, mais également dans l'azote [73].

3. Traitement des données

Le besoin croissant d'une extraction efficace de l'information chimique et biologique a donné naissance à diverses plates-formes logicielles pour couvrir les étapes de l'analyse métabolomique. Différents logiciels de traitement de données sont actuellement disponibles [74-76]. Certains sont commerciaux, fournis par les fabricants d'instruments, d'autres sont des logiciels à accès libre. Dans ce dernier cas, il est parfois possible d'accéder aux algorithmes pour en modifier les paramètres et/ou les optimiser ce qui leur offre plus de flexibilité. Ces logiciels diffèrent aussi par les approches mises en jeu. Alors que la soustraction du bruit de fond repose souvent sur des algorithmes de filtrage classiquement utilisés en traitement du signal, de grandes différences son observées au niveau des étapes d'extraction et d'alignement des signaux. Le Tableau 8 présente une liste non exhaustive de quelques logiciels utilisés dans les traitements de données MS.

Tableau 8. Exemples de logiciels de traitement des données métabolomiques.*

Logiciels	Fournisseur	Traitements des données	Normalisation	Référence
Gratuits				
XCMS	http://metlin.scripps.edu/xcms/	Filtration, détection et alignement des pics Annotation , visualisation	OUI	[77-79]
MetaboAnalyst	http://www.metaboanalyst.ca	Filtration, détection et alignement des pics Visualisation	OUI	[80,81]
Mzmine	http://mzmine.github.io/	Détection par deconvolution et alignement des pics	OUI	[82]
metaP-server	http://metabolomics.helmholtz-muenchen.de/metap2/	Filtration, détection et alignement des pics Visualisation	NON	[83]
MeltDB	http://www.cebitec.uni-bielefeld.de/groups/brf/software/meltdb_info/	Détection par deconvolution et alignement des pics	NON	[84]
MetAlign	http://www.wageningenur.nl/en/show/MetAlign-1.htm	Correction de la ligne de base et du bruit de fond, Détection et alignement des pics	OUI	[85]
Commerciaux				
Bluefuse	BlueGnome	Filtration, détection et alignement des pics	NON	
Progenesis QI	Waters®	Correction de la ligne de base et du bruit de fond, Détection et co-alignement des pics	OUI	
UNIFI	Waters®	Correction de la ligne de base et du bruit de fond, Détection et alignement des pics	OUI	
MarkerView	Applied Biosystems®	Détection par alignement des pics	OUI	
Sieve®	ThermoFisher Scientific	Détection des pics directement à partir des données brutes et Alignement des pics	NON	
MassHunter	Agilent Technologies	Détection et alignement des pics	OUI	
Metabolyzer	Metabolon	Traitement automatique Alignement des pics	OUI	
Phenomenome Profiler	Phenomenome Discoveries	Détection et alignement des pics	OUI	

* Liste non exhaustive

3.1. Conversion des fichiers

L'acquisition des données brutes est le point de départ pour le prétraitement des données en métabolomique en particulier si des logiciels à libre accès sont utilisés. Les données LC-IM-MS sont un ensemble de vecteurs de points enregistrés au cours d'intervalles de temps successifs. Chaque point se compose d'un rapport *m/z*, d'un temps de rétention, d'une CCS et d'une intensité. Les formats de fichiers sont souvent fournisseur et instruments dépendants. Cependant, le format du fichier à utiliser dépend du logiciel de prétraitement. Pour pallier ce problème, divers convertisseurs sont disponibles pour convertir les fichiers en format ouvert tel que mzXML, NetCDF ou mzML [86].

3.2. Correction de la ligne de base

Les algorithmes de correction de la ligne de base estiment la fréquence de base, puis soustraient la valeur estimée à partir du signal. Un filtre Savitzky-Golay de faible degré peut être utilisé pour supprimer la ligne de base à partir d'un signal LC-MS [87].

3.3. Filtrage

Le filtre est utilisé pour éliminer le bruit de fond acquis avec les données. Selon les paramètres, ces filtres permettent d'améliorer le rapport signal/bruit. L'exigence majeure pour le filtre consiste à supprimer le bruit tout en conservant l'information pertinente initiale. Différentes méthodes de filtrage sont décrites telles que le filtrage médian ou le filtrage par moyenne mobile. L'application aux données observées une régression locale avec une fonction polynomiale d'ordre supérieur (i. e. Savitzky-Golay) s'avère particulièrement efficace dans la préservation de la forme des pics [87,88]. Plusieurs autres méthodes pour l'élimination du bruit et la détection des pics basés sur les maximas locaux ou la transformée en ondelettes [89-91].

3.4. Détection des pics

La détection des pics est une transformation qui convertit les données brutes continues en forme centroïde donc en données discrètes sous forme de pic de sorte que chaque ion soit représenté par un pic. Cette transformation offre deux avantages : une suppression d'une partie du bruit contenu dans les données brutes et une réduction la dimension des données sans perte notable d'information. La détection des pics est en général effectuée en deux étapes en commençant par le calcul des centroïdes des pics dans la gamme *m*/*z*, puis la recherche dans toute la gamme des temps de rétention des pics chromatographiques et/ou du spectre de mobilité ionique. Pour le pic centroïde dans la gamme *m*/*z*, de nombreux fabricants d'instruments MS fournissent des logiciels spécifiques qui permettent à l'utilisateur d'acquérir directement les données en centroïde. Actuellement, les principaux efforts des algorithmes de détection de pic se concentrent sur la centroïdisation (centroïding) en fonction du temps de rétention. Compte tenu de la variation discrète des *m*/*z* par rapport aux autres dimensions séparatives (temps de rétention, temps de dérive), la détection des pics est généralement réalisée sur des chromatogrammes d'ions extraits (Extracted Ion Chromatograms EIC) [77,92] ou Ion Mobility Spectra pour la mobilité ionique, qui est, en fait, un signal 2-D de l'intensité en fonction du temps de rétention sur un petit intervalle *m*/*z*. Dans la majorité des algorithmes, les EIC sont acquis par binning sur la gamme *m*/*z* avec un petit intervalle (exemple 10-30 ppm). Cependant, une des limites du binning est qu'un ion peut être subdivisé en deux pics voisins. Après l'extraction des EIC, ils sont analysés pour déterminer la présence de pics ainsi que leurs limites en utilisant un filtre adapté [77]. D'autres algorithmes sont décrits pour améliorer la détection des pics grâce à une meilleure modélisation des pics chromatographiques [92,93].

3.5. Alignement

Le groupement des pics et l'alignement des temps de rétention (tR) et/ou temps de dérive (tD) permettent la comparaison des données LC-IM-MS dans les différents échantillons analysés. En effet, ces dimensions séparatives peuvent présenter des dérives dans les différents échantillons à cause des fluctuations instrumentales au cours de l'analyse. Ainsi, après la détection de pics, un alignement sur tous les profils est nécessaire pour générer un ensemble combiné de caractéristiques interprétables. L'alignement s'effectue par *m*/*z* ainsi que les autres dimensions de séparation [94,95]. La plupart des méthodes existantes incluent une étape d'estimation de l'écart de temps non linéaire et fournissent des temps corrigés [94,96]. L'alignement par paires est ensuite complété par référence au profil avec un nombre maximal de caractéristiques détectées et tous les autres profils sont alignés par rapport à la référence par paires en utilisant divers algorithmes [95-99]. Pour les études métabolomiques, l'alignement du temps de rétention est utilisé pour corriger la dérive des temps de rétention et veiller à ce que le même ion soit comparable entre les différents échantillons. L'une des méthodes proposées pour la correction du temps de rétention est d'ajouter des composés de référence dans les échantillons et les utiliser comme point de repère pour aligner les pics [6]. Toutefois, ces composés de référence doivent être soigneusement choisis pour avoir une couverture suffisante de l'intervalle de temps de rétention et d'éviter, ainsi, un chevauchement lors de l'analyse des métabolites. Cependant, la présence excessive d'étalons internes peut également provoquer une suppression spectrale. À cause de ces limites, les approches d'alignement n'utilisant pas de composés de référence sont préférables. Une de ces approches utilise les résultats de détection de pics et essaie de trouver et de faire correspondre les pics similaires [77]. Par exemple, XCMS utilise d'abord une estimation pour regrouper les pics avec les mêmes valeurs *m*/*z* et temps de rétention et à travers l'ensemble de données. Après regroupement des meilleurs pics, des groupes de pics qui n'ont pas été attribués dans très peu d'échantillons sont utilisés comme repères pour l'alignement. Une régression est effectuée entre les écarts entre les temps de rétention de ces pics repères à partir de leurs valeurs médianes au sein des groupes de pointe et le temps de rétention. Les régions sur le chromatogramme sans pics de référence peuvent être interpolées et alignées. Les pics alignés sont regroupés une seconde fois, pour adapter les pics avec les temps derétention corrigés. Cette procédure est habituellement effectuée de manière itérative deux ou trois fois pour s'assurer que la dérive du temps de rétention est suffisamment corrigée. L'autre catégorie d'approches utilise les données LC-MS brutes pour l'alignement de temps de rétention tels que l'ion EIC ou le TIC. Une évaluation critique de plusieurs méthodes d'alignement conclut que XCMS donne la meilleure performance pour l'alignement des données LC-MS [100]. Cependant, il est à souligner que la performance d'une méthode dépend fortement du choix des paramètres appropriés [101,102]. L'optimisation des paramètres de prétraitement est une étape cruciale dans l'extraction des données [101,103-105]. IPO [106] et xMSanalyzer [105] sont des exemples de packages développés sous l'environnement R qui permettent l'optimisation des paramètres XCMS et d'autres logiciels de prétraitement. La limite majeure de ses solutions et le temps de calcul nécessaire pour exécuter les extractions multiples, intégrer des données et évaluer leur qualité.

3.6. Regroupement des ions et leurs adduits

Cette étape groupe les ions qui sont susceptibles de provenir d'un même composé. Lors de l'utilisation des méthodes séparatives couplées à la MS, un métabolite est souvent représenté par des pics multiples et distincts portant des valeurs *m/z* distinctes, mais avec un même temps de rétention, et ceci en raison de la présence des isotopes, adduits et ions fragments. Quand la vitesse de balayage est correctement ajustée et le nombre de points acquis est suffisant pour définir les pics chromatographiques, les ions du même composé forment des profils d'élution similaires qui peuvent être représentés par leurs EIC. L'annotation d'ions peut être réalisée par regroupement des profils d'élution semblables. Une méthode d'annotation d'ions a été développée dans laquelle les ions sont regroupés sur la base de la corrélation de Pearson de leur EIC [107]. Si la corrélation entre deux ions est supérieure à un seuil prédéfini et la différence *m/z* entre les deux ions peut s'expliquer par une information connue (adduits, isotopes ou ion fragment), les deux ions sont considérés comme provenant du même métabolite. Toutefois, dans la méthode décrite ci-dessus, le choix du seuil de corrélation de Pearson est largement empirique, sans interprétation statistique. En outre, lorsque les profils d'élution de deux ions ont un grand chevauchement, la corrélation de Pearson est généralement élevée et pas assez sensible pour capter les différences subtiles dans EIC. Une approche statistique rigoureuse a été proposée pour tester si deux ions mesurés par TOF-MS sont originaires de la même entité chimique [108]. Dans cette approche, le signal observé est modélisé comme une distribution de Poisson. Si deux ions sont dérivés du même composé, la distribution de l'intensité observée suit une loi binomiale. Le test Pearson χ^2 a été utilisé pour évaluer la qualité de la corrélation de l'observation de la distribution binomiale à laquelle une *p*-value est associée. Il a été montré que cette approche permet de réduire de 6% le taux de faux positifs dans l'annotation d'ion par rapport aux 50% obtenus par la méthode de corrélation de Pearson en conservant le même niveau de sensibilité [108].

4. Identification des métabolites en spectrométrie de masse

L'identification des métabolites est une étape fondamentale en métabolomique qui permet de traduire les données analytiques acquises en information biologique interprétable dans le contexte étudié. L'introduction des spectromètres de masse à haute résolution et l'obtention de mesures de masses précises, donnant accès à la formule brute des pics détectés, a considérablement accéléré cette étape [109]. L'utilisation conjointe de pièges ioniques pour la réalisation d'expériences de fragmentations séquentielles permet d'obtenir des informations structurales complémentaires, indispensables à l'identification des métabolites d'intérêts [110]. Cependant, la spectrométrie de masse associée à des sources d'ionisation à pression atmosphérique présentant une forte variabilité dans les profils de fragmentation générés sur différents appareils [111] limitant, ainsi, la construction de banques de données spectrales universelles à l'instar de celles obtenues par ionisation électronique ou par RMN [112]. Cette difficulté est lié en particulier au manque de standardisation au niveau des modes de dissociation entre les types d'instruments (piège à ions, Q-TOF…), du type de gaz de collision, de la pression de gaz, mais également par des différences entre les constructeurs de spectromètres de masse. En spectrométrie de masse, une ou plusieurs formules élémentaires peuvent être générée(s) si des instruments de haute résolution sont utilisés, ce qui fournit un premier élément pour effectuer une interrogation des bases de données existantes. L'acquisition de spectres de fragmentation, le plus souvent par spectrométrie de masse en tandem, permet à ce stade de discriminer les réponses obtenues précédemment sur la base d'ions produits ou de pertes de neutres, caractéristiques de groupements chimiques [109]. Identifier les variables d'intérêt isolées par l'analyse statistique multivariée est l'ultime étape critique de l'analyse métabolomique. Compte tenu de l'importance de l'étape d'identification, des éléments de standardisation ont été proposés pour harmoniser les données d'identification des métabolites. Quatre niveaux d'identification ont été définis dans le cadre de la « Metabolomics Standards Initiative » en fonction des informations disponibles sur le métabolite à identifier [113,114].

1 **Métabolite identifié** Un minimum de deux paramètres physico-chimiques indépendants identiques à ceux du standard dans les mêmes conditions analytiques. En LC/MS cela peut correspondre à la masse précise, au temps de rétention ou encore le spectre de MS^n.

2 **Métabolite putativement annoté** En cas d'indisponibilité du standard, l'identification peut être basée sur les propriétés physico-chimiques (i.e. tR, CCS) et/ou les similarités spectrales (ex : MS/MS) avec les informations des bases de données publiques ou privées.

3 **Métabolite putativement caractérisé** L'identification peut être basée sur les propriétés physicochimiques d'une classe de composés et/ou les similarités spectrales (ex : MS/MS)

4 **Métabolite inconnu** Bien que non identifiés, ces métabolites peuvent être différenciés à partir de données spectrales qui peuvent permettre une quantification relative.

L'annotation d'un métabolite est définie comme une identification putative et ne doit pas être confondue avec l'identification chimique. L'identification chimique est requise pour les pics d'intérêt, mais elle peut être difficile et soumise à des critères de certitude différents [115]. L'identification des métabolites est l'étape limitante majeure dans la métabolomique non ciblée. En effet, dans le cas de l'ionisation par *electrospray* (ESI), chaque composé chimique spécifique donne lieu à une ou plusieurs espèces d'ions, qui sont inclus dans le même spectre de masse. Ces espèces ioniques comprennent ; isotope, fragments et adduits. La présence de tous ces ions représentant un seul composé pose les problèmes de redondance de signaux pendant l'analyse de données. Ces dernières années, plusieurs méthodes telles que AStream [116], CAMERA [117], ProbMetab [118] et MetAssign [119] ont été développées pour l'annotation de métabolites. La plupart de ces méthodes utilisent rapport *m/z*, le temps de rétention, les modèles d'adduits, les isotopes, et les corrélations et les similitudes séparatives entre les métabolites pour l'annotation des métabolites. Des méthodes utilisant d'autres sources d'information, telles que les associations de voies métaboliques, peuvent améliorer l'identification [120,121]. Le développement d'algorithmes qui utilisent l'apprentissage automatique pour prédire le temps de rétention, les CCS, les probabilités d'adduits et d'isotopes, l'intensité relative et diverses propriétés physiques des métabolites, des modes d'ionisation et des colonnes préalablement validées pourraient améliorer les performances des méthodes de prédiction d'identité [122,123].

- **Caractérisation multi-vectorielle des métabolites**

En LC-IM-MS, l'identification des composés reste une tâche difficile et souvent considérée comme le défi majeur dans l'analyse et l'interprétation des données métabolomiques. Les trois principales stratégies pour l'identification de métabolites sont :

1/ L'identification basée sur la masse précise en utilisant des spectromètres de masse à haute résolution. En combinant la mesure de masse précise, l'allure du massif isotopique, l'accès à la composition élémentaire du composé est possible.

2/ L'application de la spectrométrie de masse en tandem où l'appareil effectue l'enregistrement d'un spectre de masse, et puis en sélectionnant un ou plusieurs ions pour des analyses en mode MS^2, MS^n ou sans sélection préalable des ions comme dans le cas du mode MS^E de la société Waters. Cette approche permet l'accès à l'information structurale d'un composé en exploitant les profils de fragmentation.

3/ La comparaison des temps de rétention, des mesures de CCS et des spectres de masse des métabolites avec ceux obtenus à partir des standards commerciaux.

Les méthodes de caractérisation d'ions décrites ci-dessus constituent une base pour désigner les ions par leur positionnement dans un espace multivectoriel. Une telle approche nécessite l'assemblage de données pour des ions non identifiés dans des bases de données spectrales [124]. Des mesures robustes seront nécessaires pour définir la masse précise, le temps de rétention, les spectres MS^2, les sections efficaces de collision (CCS) et leur chiralité. Chaque paramètre fournissant un vecteur pour définir de manière unique un ion dans un espace multivectoriel.

4.1. Masse précise

Les analyseurs de masse sont désormais disponibles pour fournir des mesures de *m/z* à 1 ppm. De telles informations peuvent être particulièrement utiles en tant que caractéristique robuste pour décrire des ions non identifiés. La résolution de masse est importante pour s'assurer que le rapport *m/z* reflète un ion unique et l'étalonnage de masse est essentiel pour assurer la précision du *m/z* indiqué.

4.2. Temps de rétention

En LC, les produits chimiques sont séparés en fonction de leur répartition entre la phase stationnaire et la phase mobile. Les produits chimiques co-élués possèdent généralement des propriétés similaires, telles que la lipophilie, l'hydrophobie, la force ionique et la constante de dissociation des acides. Par conséquent, les temps de rétention des produits chimiques possédant des structures et des propriétés physicochimiques connues pourraient servir de référence pour déduire des propriétés physico-chimiques qualitatives pour prédire un métabolite inconnu [125]. De plus, si un *m/z* non identifié est détecté avec deux techniques de séparation chromatographique orthogonales telle qu'en phases inversées (C18) et HILIC, ou échange d'anions et C18, les associations de métabolites et l'indexation du temps de rétention peuvent être validées entre plusieurs plateformes.

4.3. Caractérisation des ions par les spectres MS^2

De nombreux outils existent pour caractériser les modèles de fragmentation MS^2 et prédire l'identification basée sur des caractéristiques spectrales. L'interprétation est complétée par une combinaison de différentes stratégies et d'approches informatiques. Lorsqu'on utilise l'information spectrale pour caractériser les pics inconnus détectés, il est utile de classer la méthodologie en deux groupes : top-down (méthode *in silico*) ou de *botom-up* (élucidation structurale). Les approches top-down utilisent des modèles théoriques, souvent étalonnés pour recueillir des données expérimentales de MS^2, pour prédire les modèles de fragmentation basés sur les énergies de dissociation des liaisons, le réarrangement et les groupes fonctionnels moléculaires. Bien qu'il n'existe actuellement aucun algorithme disponible fournissant un spectre MS^2 de haute précision pour toutes les différentes méthodes de dissociation, il existe plusieurs méthodes heuristiques qui fournissent suffisamment de spectres qui peuvent être utilisés pour améliorer la confiance dans l'annotation et la classification des caractéristiques des ions *m/z* fragmentés. De nombreuses approches combinent la fragmentation *in silico* avec des données MS^2 collectées expérimentalement pour l'annotation des fragments et le classement de la probabilité d'une identification correcte [122,126,127]. Les réseaux moléculaires des données MS^2 [128,129] suscitent beaucoup d’intérêt et sont des outils importants pour le développement de la métabolomique. Les réseaux moléculaires constituent une technique qui a été initialement développée comme stratégie de déréplication pour l'identification des produits naturels [129], et utilise un réseau de similarité déterminé à partir de la relation du spectre de fragmentation pour identifier des produits chimiques structurellement similaires. Le réseau résultant peut être utilisé pour identifier des produits chimiques partageant des composants structurels et des biotransformations similaires [130].

4.4. Section efficace de collision

La CCS fournit un important descripteur moléculaire complémentaire pour faciliter l'identification moléculaire. De même, pour une caractérisation non ambiguë des ions non identifiés, le CCS fournit une caractéristique utile qui est indépendante des spectres MS^2 et du temps de rétention et partiellement indépendante de la mesure de la masse *m/z*. En effet la masse d’une molécule reste corrélée à sa taille. De surcroît, la CCS est une propriété moléculaire unique et prédictible. Comme la séparation est basée sur des principes différents que la séparation chromatographique, l'IMS-MS peut fournir une caractérisation supplémentaire et orthogonale qui n'est pas accessible par les autres dimensions analytiques. Plusieurs algorithmes de calcul, tels que la méthode de la trajectoire, la méthode de diffusion de la sphère dure exacte et la méthode d'approximation de la projection, ont été développés pour déterminer les valeurs de CCS à partir de modèle moléculaires théoriques [65]. Des études récentes ont montré que la combinaison des valeurs expérimentales de CCS obtenues avec les techniques de modélisation moléculaire et les valeurs théoriques de CCS obtenues en utilisant MOBCAL ou Sigma peuvent aider à l'identification structurale de métabolites de médicaments, de lipides, de petites molécules et d'isomères structurels inconnus [131-136]. Par ailleurs, il est à noter un besoin urgent de développer des processus automatiques de détermination de la CCS en combinaison avec des méthodes de calcul pour faciliter grandement l'identification des ions détectés.

5. Bases de données

Afin de faciliter l'étape d'identification des métabolites, de nombreuses bases de données ont été développées comprenant notamment les propriétés physico-chimiques et biologiques des composés chimiques. Le **Tableau 9** présente une liste non exhaustive. Enfin, le site OMICtools fournit une description très riche des logiciels pouvant être utilisés pour l'analyse des données métabolomiques, ainsi que d'autres ressources omiques [137].

Tableau 9. Bases de données utilisées en métabolomique. [138-140]. (Liste non exhaustive)

Base de données	URL	Commentaires
Données métaboliques et biochimiques		
HumanCyc (Encylopedia of Human Metabolic Pathways)	http://humancyc.org/	Métabolisme humain
KEGG (Kyoto Encyclopedia of Genes and Genomes)	http://www.genome.jp/kegg/	Voies métaboliques Couvre de nombreux organismes
The Medical Biochemistry Page	http://themedicalbiochemistrypage.org/	
MetaCyc (Encyclopedia of Metabolic Pathways)	http://metacyc.org/	Semblable à KEGG Pas de voies métaboliques pathologiques ou de médicaments
Reactome (A Curated Knowledgebase of Pathways)	http://www.reactome.org/	
Roche Applied Sciences Biochemical Pathways Chart	http://www.expasy.org/cgi-bin/search-biochem-index	Métabolisme humain
Small Molecule Pathway Database (SMPDB)	http://www.smpdb.ca/	Métabolisme humain Pas de voies métaboliques pathologiques ou de médicaments Outils d'analyse et de visualisation
Chemicals Entities of Biological Interest (ChEBI)	http://www.ebi.ac.uk/chebi/	Couvre les métabolites et les médicaments d'intérêt biologique Mettre l'accent sur l'ontologie et la nomenclature (pas la biologie)
Données métabolomiques, chimiques et spectrales		
ChemSpider	http://www.chemspider.com/	Meta-database contenant des données chimiques
Golm Metabolome Database	http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html	Données MS ou GC-MS uniquement Pas de données biologiques Spécifique aux plantes
Human Metabolome Database	http://www.hmdb.ca	Métabolome humain uniquement
KNApSAcK	http://kanaya.naist.jp/KNApSAcK/	Données phytochimiques
LipidMaps	http://www.lipidmaps.org/	Lipidomique Nomenclature standard
METLIN Metabolite Database	http://metlin.scripps.edu/	Métabolites Nom, structure et ID
PubChem	http://pubchem.ncbi.nlm.nih.gov/	Base de données contenant 27 millions de produits chimiques uniques avec des liens vers les résumés PubMed
NIST National institute for standard and technology (USA)	www.nist.gov/srd/nist1a.htm	
Données toxicologiques		
ACToR (Aggregated Computation Toxicology Resource)	http://actor.epa.gov/actor/faces/ACToRHome.jsp	2.500.000 produits chimiques (structure, valeurs physico-chimique, les données d'essai toxicologiques in vitro et in vivo)
CTD (Comparative Toxicogenomic Database)	http://ctd.mdibl.org/	Données sur les produits chimiques et leurs interactions avec les gènes
SuperToxic	http://bioinformatics.charite.de/supertoxic/	60 000 composés toxiques Données chimiques et toxicologiques
T3DB (Toxin, Toxin-Target Database)	http://www.t3db.org/	3100 toxines communes 1400 protéines cibles Données structurales, physiologiques, mécanistiques, biochimiques et médicales

6. Analyses des données

6.1. Evaluation de la qualité des données et stratégies de corrections

Des outils Web et les packages R tels que MetaboAnalyst [141], xMSanalyzer [105] et MSPrep [142] et QCScreen [143] fournissent des utilitaires pour l'évaluation de la qualité de l'extraction des données et sa correction le cas échéant. La qualité des caractéristiques individuelles des échantillons peut être évaluée en fonction du coefficient de variation (CV) des pics dans les réplicats techniques. Le pourcentage de valeurs manquantes, le rapport signal sur bruit, la précision de la mesure de masse, les effets de dérive sont des paramètres usuellement utilisés pour évaluer la reproductibilité analytique [141,144]. Des échantillons de contrôle qualité (QC) sont utilisés pour évaluer la qualité des données acquises, souvent via une analyse en composantes principales pour identifier les valeurs aberrantes. Diverses méthodes ont été développées pour traiter des problèmes d'effet batch (effet de dérive inter-séries) souvent observés dans les analyses métabolomique à grande échelle [145]. L'erreur de mesure de masse précise est une autre source d'erreur et peut se produire en raison de variations de température et d'un étalonnage incorrect de l'instrument. La précision de la masse joue un rôle critique lors de l'alignement et de l'annotation des pics. Au cours du processus d'annotation des pics, le rapport *m/z* mesuré est comparé au *m/z* théorique et seuls les métabolites qui sont dans le niveau de tolérance de masse défini par l'utilisateur sont sélectionnés. Le nombre de faux positifs peut augmenter de façon sensible à mesure que l'exactitude de la masse se détériore [146-148]. Des étalons internes ou les pics annotés basés sur les métabolites de référence peuvent être utilisés pour le suivi de la précision de masse et l'estimation de l'erreur de mesure de masse. En résumé, différentes approches sont disponibles pour améliorer l'extraction d'informations sur les ions mesurés par spectrométrie de masse. Ces approches fournissent une évaluation de la qualité et une correction des données pour l'utilisation de la métabolomique non ciblée. Cependant, en termes de détection chimique, les limites doivent être prises en considération. Pour développer la détection des entités chimique à faible abondance, des efforts supplémentaires doivent être axés sur l'identification et l'amélioration de la sensibilité des instruments.

6.2. Normalisation

Les données omiques partagent plusieurs caractéristiques intrinsèques tel que leur distribution asymétrique [149] leur grande dispersion [150], la proportion substantielle de bruit instrumental, analytique et biologique [151,152] et la variabilité des cohortes [153,154]. L'objectif de la normalisation des données est de supprimer les biais expérimentaux concernant l'abondance des ions détectés entre les différents échantillons, tout en conservant les variations biologiques. La plupart des méthodes sont inspirées des précédentes stratégies omiques (génomique et transcriptomique) qui souffrent des mêmes biais expérimentaux [54]. En effet, la diversité chimique des métabolites et les variations interindividuelles entraînant des modifications des rendements d'extraction et d'ionisation, rendant difficile la séparation des variations d'intérêt biologique des biais analytiques (instrumentations, opérateurs, réactifs). Les stratégies de normalisation des empreintes métaboliques peuvent être divisées en deux catégories, les approches statistiques et les approches chimiques.

- *Approches statistiques* : basées sur des modèles statistiques utilisés pour définir des facteurs de correction propres à chaque échantillon à partir du jeu de données complet [155], tels que la normalisation par l'écart-type [156], par l'intensité moyenne globale [157], normalisation par quantile [158], Probabilistic Quotient Normalization [159], Linear Baseline Scaling [160], Non-Linear Baseline Normalization [161] normalisation de contraste [162], cubic splines [163], Cyclic Loess [164], QC-robust spline batch correction (QC-RLSC) [165] signal total utile (MS Total Useful Signal MSTUS) [166] et support vector regression [167]. Des analyses comparatives des différentes méthodes de normalisation a été récemment présentées [155].
- *Approches chimiques* : reposant sur un ou plusieurs composés de référence [168-170], standards internes, composés endogènes ou xénobiotiques, utilisés pour normaliser l'ensemble du chromatogramme (composé unique) ou certaines régions du chromatogramme, en normalisant chaque zone grâce à un standard qui y est élué.

D'autres stratégies basées sur des caractéristiques de la matrice étudiée telle que, la masse sèche des échantillons, le volume (e. g. diurèse), osmolalité, le taux de protéines ou créatininurie peuvent être aussi utilisées [171].

6.3. Transformation

La transformation des données métabolomiques est parfois nécessaire pour modifier la distribution des données pour les préparer aux analyses statistiques ultérieures. Les transformations conduisant à une distribution normale des données ou à une réduction de la variation dynamique des données sont souvent utilisées. Ces transformations des intensités des pics sont généralement empiriques. Diverses méthodes sont décrites [172]; Z-score, transformation logarithmique, et la transformation racine carrée (square-root transformation) sont des choix récurrents. Le **Tableau 7** résume les différentes méthodes de transformations utilisées. La **Figure 13** présente les effets de la transformation sur les données.

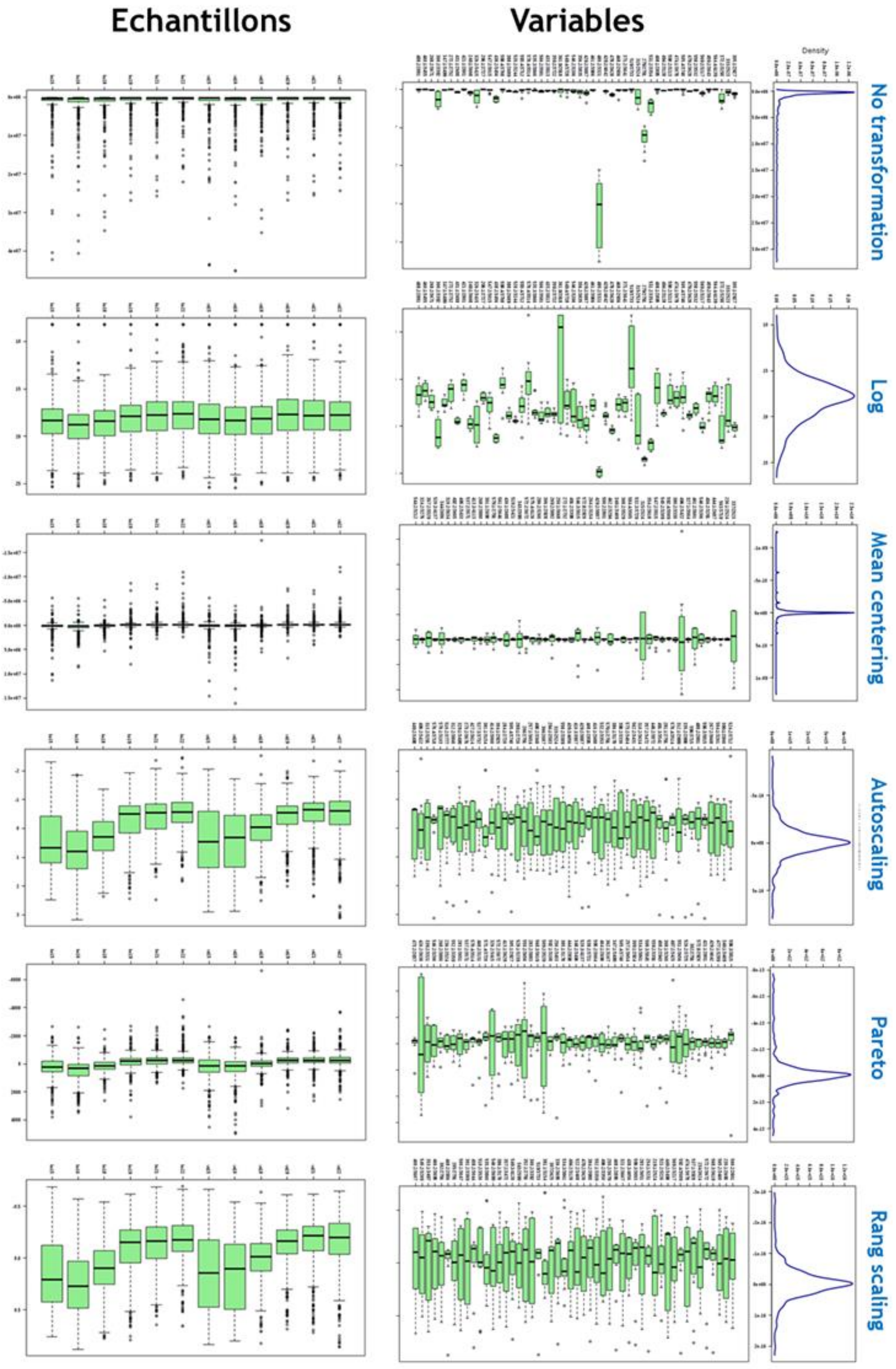


Figure 13. Effets de différentes transformations sur les données.

Tableau 7. Méthodes de transformation de données. D'après [172].

Méthode	Formule	Unité	But	Avantages	Limites
Centering	$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$		Centrer les données	Surprime le décalage	En cas d'hétérosédasticité, cette correction n'est pas suffisante
Autoscaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$	/	Comparer les données	Toutes les données ont la même unité	Augmentation de l'erreur
Range scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\left(x_{i_{\max}} - x_{i_{\min}}\right)}$	/	Comparer par rapport à l'intervalle de réponse biologique	L'importance de tous les métabolites est équivalente. La normalisation est liée à la biologie	Augmentation de l'erreur Sensible aux points aberrants
Pareto scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$	*O*	Réduit l'importance des valeurs élevée en conservant la structure des données	Plus conservateur que l'autoscaling	Sensible aux grands écarts (Fold change)
Vast scaling	$\tilde{x}_{ij} = \frac{\left(x_{ij} - \bar{x}_i\right)}{s_i} \cdot \frac{\bar{x}_i}{s_i}$ $\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$	/	Focalise sur les métabolites avec de faibles variations	Offre plus de robustesse	Pas adéquate pour les larges variations intragroupes
Level scaling		/	Focalise sur les réponses proportionnelles	Convient à l' identification	Augmentation de l'erreur
Log transformation	$\tilde{x}_{ij} =^{10} \log\left(x_{ij}\right)$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	**Log *O***	Corrige l'hétéroscedasticité Rend additif les modèles multiplicatifs	Réduit l'hétéroscedasticité. Les effets multiplicatifs deviennent additifs	Difficultés pour les valeurs ayant un grand écart-type et les zéros
Power transformation	$\tilde{x}_{ij} = \sqrt{\left(x_{ij}\right)}$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	**√*O***	Corrige l'hétéroscedasticité	Réduit l'hétéroscedasticité Prend en charge les valeurs faibles	Le choix du degré de la racine est arbitraire

Unité : correspond àl'unité desdonnéesaprèstransformation.**O**représentel'unité d'origine, et« / »représente les donnéessans dimension.

$\tilde{x}$ et $\hat{x}$ représentent les données après transformation.

Médiane = $s_i = \sqrt{\frac{\sum_{j=1}^{J}\left(x_{ij} - \bar{x}_i\right)^2}{J-1}}$ Déviation Standard = $\bar{x}_i = \frac{1}{J}\sum_{j=1}^{J} x_{ij}$

Il est à noter que différentes méthodes peuvent être utilisées simultanément. Des analyses comparatives pour évaluer les performances des différentes méthodes de normalisation et de transformation ont été récemment décrites dans la littérature [155,173].

6.4. Analyse statistique

6.4.1. Analyse statistique univariée

Le but de cette étape est de réduire la dimension des données acquises en filtrant les variables initialement détectées et aboutir un nombre réduit portant l'information potentiellement pertinente pour les analyses statistiques ultérieures. Après avoir choisi le test statistique adéquat au plan expérimental et aux jeux de données acquises en fonction de la distribution des données, paramétrique ou non paramétrique, les variables sont classées et ne sont retenues que les variables montrant des variations statistiquement significatives entre les groupes étudiés par comparaison un seuil préalablement défini [174]. Il est à noter que les résultats doivent être analysés sous forme de médianes au lieu des moyennes si le test non paramétrique est choisi. Dans le cas des comparaisons multiples, il est impératif d'indiquer le nombre de faux positifs après une correction FDR – False Discovery Rate – telle que la correction Bonferroni ou correction Benjamini-Hochberg [175,176]. Tracer les histogrammes de distribution de fréquence des *p*-values pour avoir une vue d'ensemble et apprécier si les données contiennent des différences significatives et définir un seuil FDR. Un seuil de 5% est souvent défini pour le FDR. Divers paramètres permettent cette sélection tels que la significativité des tests univariés, le rapport de variation (Fold Change), le coefficient de variation des contrôles qualité et la corrélation entre différentes dilutions du contrôle qualité.

6.4.2. Analyse statistique multivariée

Les analyses métabolomiques génèrent des quantités de données importantes aussi bien sur le plan qualitatif que quantitatif. L'analyse de ces données nécessite des outils mathématiques et statistiques répondant à cet impératif de multidimensionnalité. Les analyses statistiques multivariées sont les mieux adaptées pour organiser, extraire et interpréter l'information biologique complexe associée. La visualisation des signatures métaboliques est d'une importance primordiale pour analyser la complexité biologique. La flexibilité des méthodes PLS en général et de O-PLS permet d'améliorer l'analyse des données complexes, ce qui facilite l'extraction des informations relatives aux processus biologiques. Ces approches sont utilisées pour résoudre les problèmes présents dans les ensembles de données biologiques complexes et multifactorielles.

6.4.2.1. Analyse multivariée descriptive : Approche non supervisée

6.4.2.1.1. Analyse en Composante Principale (ACP)

L'objectif principal de cette méthode est la compression des données et leur exploration globale. Le principe commun à toutes les méthodes multivariées est de considérer les échantillons comme étant des points dans un espace défini par des variables et que les coordonnées d'une observation sont données par les valeurs de chacune de ces variables. Le principe de l'Analyse en Composante Principale (PCA) est de trouver les directions de plus grande dispersion des observations dans cet espace. L'idée étant que les directions de la plus grande dispersion sont les directions contenant le plus de variance et donc d'information [177-179]. Si les données ne contiennent que du bruit, les individus seront dispersés de façon homogène et uniforme dans toutes les directions. Une direction qui s'écarte d'une telle répartition sphérique contiendrait de l'information potentiellement pertinente. Mathématiquement, la PCA calcule des combinaisons linéaires des variables de départ donnant de nouveaux axes qui contiennent la plus grande partie de la variabilité de la matrice de données de départ. La PCA est une méthode non supervisée, car aucune hypothèse préalable n'est faite concernant les relations éventuelles entre les individus et entre les variables. Pour éviter d'avoir la même «information» dans plusieurs Composantes Principales, celles-ci doivent toutes être orthogonales les unes par rapport aux autres. La décomposition matricielle de la PCA permet d'obtenir des matrices des

coordonnées factorielles (ou «**scores**») et des **contributions factorielles** (ou «**loadings**»), à partir de la matrice de données originale, pour avoir les coordonnées factorielles des individus et les contributions factorielles des variables sur les Composantes Principales. La quantité de variance (information) contenue dans chaque PC est proportionnelle à sa valeur propre («eigenvalue»). Les composantes principales sont simplement des entités mathématiques qui peuvent représenter, après un choix d'un ensemble de variables représentatif, la matrice de départ. D'un point de vue géométrique, la PCA peut être plus facilement comprise comme une méthode de rotation des données pour que l'observateur soit le mieux placé pour comprendre les relations entre les individus. Les coordonnées factorielles permettent de projeter les individus sur des plans construits à partir des PC, où l'on peut éventuellement détecter des répartitions structurées des objets, la formation de groupes ou la présence d'individus aberrants (**Figure 14**).

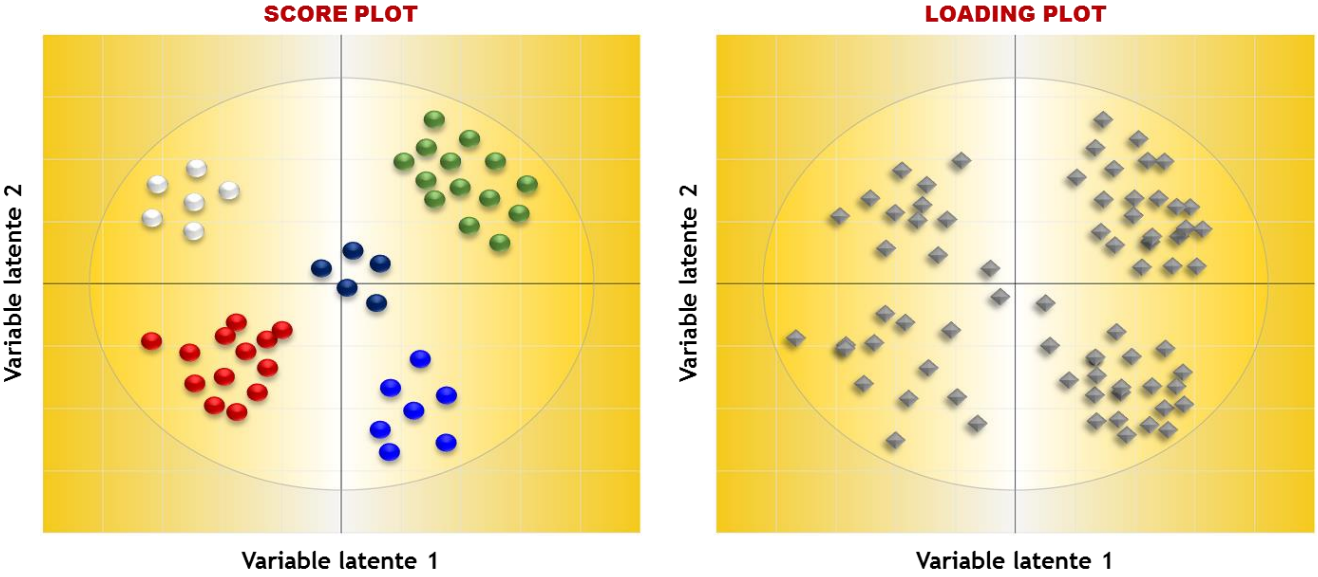


Figure 14. Les diagrammes de dispersion des observations (scores plot à gauche) et des loadings (à droite) doivent être interprétés simultanément afin d'analyser les relations entre les tendances de regroupement des observations et quelles variables - métabolites – en sont responsables. La position des variables peut être superposée à celle des observations pour l'interprétation des relations entre les variables et observations.

6.4.2.1.2. Identification des Outliers

Les échantillons aberrants « Outliers » sont facilement identifiés en analyse multivariée. Une observation avec une déviation significative sera évidente sur le score plot avec une ellipse qui indique l'intervalle de confiance (e. g. 95%) établie en utilisant une généralisation multivariée du test de Student, appelé Hotelling T^2. Une autre façon des détecter les valeurs aberrantes modérées est d'utiliser la distance au modèle défini dans l'espace X comme DModX, implémentée dans SIMCA Umerics, ce qui équivaut à la déviation standard de l'observation. Quand DModX est plus grande que la distance critique choisie (Dcrit > 95%), l'observation est considérée comme une valeur aberrante modérée [179] (**Figure 15**).

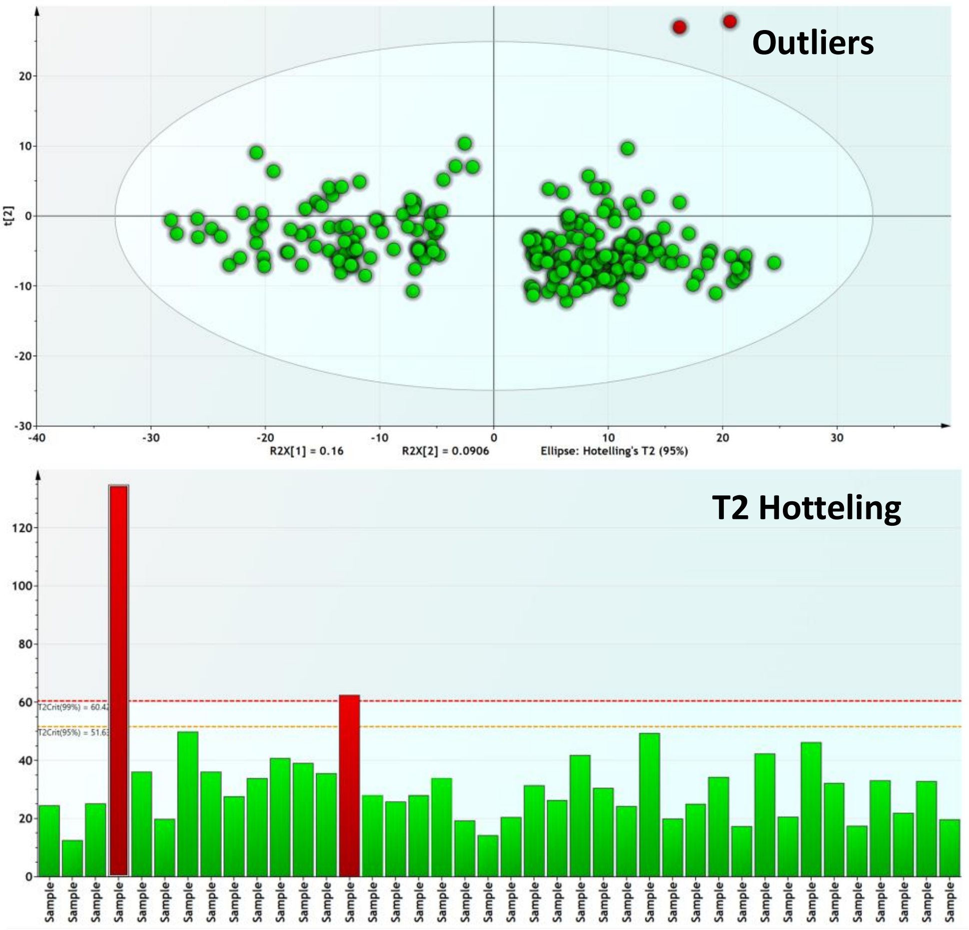


Figure 15. Score plot en deux dimensions de l'analyse en composante principale montrant la dispersion des observations. En rouge sont montrées les observations aberrantes (outliers).

6.4.2.2. Analyse multivariée explicative : Approche supervisée

6.4.2.2.1. Partial Least Squares Regression (PLS)

L'objectif de cette méthode est de construire des modèles de régression multivariée. La régression PLS, ou régression au sens des moindres carrées partielles ou régression par projection su s structures latentes (PLS) cherche à trouver les relations entre deux matrices à travers un modèle linéaire multivarié. Elle permet l'analyse des données avec des variables colinéaires, bruitées ou incomplètes dans les deux matrices [180-182]. Les variables prédictives (**X**) sont souvent des spectres, mais peuvent aussi être des mesures physico-chimiques, la plupart du temps avec une forte colinéarité. Les réponses (matrice **Y**) peuvent être de natures variées, dont des concentrations des substances à prédire, mais aussi des propriétés physico-chimiques, des activités biologiques. En métabolomique, les variables (**X**) correspondent aux couples Dimension1_Dimension2_Dimension(n) [exemple de dimensions : tR, *m*/*z*, CCS] et les réponses **Y** correspondent aux intensités.

6.4.2.2.2. PLS-DA (PLS-Discriminant Analysis)

En biologie, il est courant que la variable réponse « **y** » soit catégorielle, définissant une appartenance à une classe ou groupe, par exemple Témoins/Traités, Patients/Controles. L'analyse discriminante PLS-DA (PLS-DA) n'est en fait qu'une régression PLS classique où la variable réponse « **y** » est catégorielle. L'objectif de la PLS-DA est d'accentuer la séparation entre les groupes d'observations et de déterminer les variables qui portent l'information permettant la séparation des classes en examinant les loadings des variables correspondant sur les composantes latentes ayant permis la séparation [182,183]. La **Figure 16** présente un exemple de score plot de PLSDA.

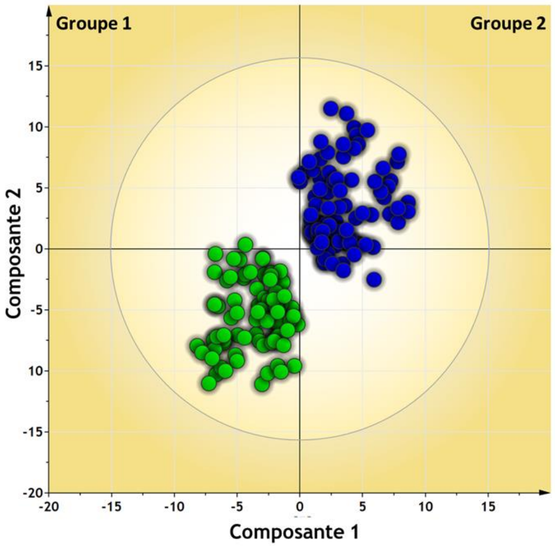


Figure 16. Exemple d'une représentation PLS-DA montrant la séparation des groupes 1 et 2.

6.4.2.2.3. OPLS-DA (Orthogonal Partial Least Square Discriminant Analysis)

L'Orthogonal Partial Least Square Discriminant Analysis (OPLS-DA) est une modification de l'algorithme initial de la PLS avec l'objectif d'éliminer la variation de X qui n'est pas corrélée avec y [184-187]. Les données analytiques contiennent souvent des variations systématiques qui ne sont pas liées à la réponse Y. Cette variation systématique peut être expérimentale ou biologique. Il est ainsi important de séparer la variation X non corrélée avec la réponse y, car elle affecte les performances prédictives des modèles statistiques générés. L'OPLS fournit une méthode pour supprimer la variation de X indépendante de Y, donc orthogonale et améliore l'interprétation des modèles PLS en réduisant la complexité du modèle (**Figure 17**).

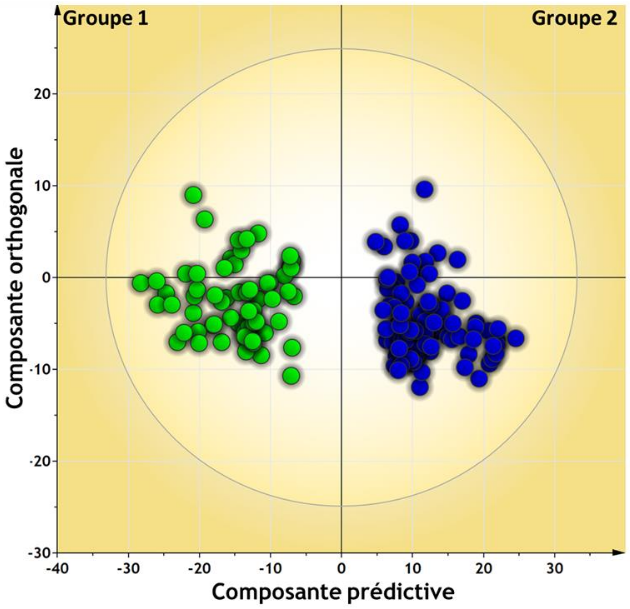


Figure 17. Interprétation de la variabilité intra et inter-groupes en OPLS-DA est facilitée par la séparation de la composante prédictive des composantes orthogonales.

6.4.3. Validation des modèles

La validation des modèles générés est une étape importante et indispensable pour l'analyse de données. De la validation du modèle dépendra sa fiabilité prédictive ultérieure [182].

6.4.3.1. Validation croisée

Cette étape permet de définir le nombre optimal de variables latentes donc la dimensionnalité optimale du modèle. On peut construire des modèles de plus en plus proches des données simplement en augmentant le nombre de variables latentes utilisées. Ceci entraîne une diminution des résidus avec l'augmentation du nombre de variables latentes. Dans le cas de l'utilisation de la PLS comme modèle de prédiction, l'objectif est la diminution des écarts par rapport à la prédiction. Le défi est de s'approcher le plus des valeurs de la variable réponse, en utilisant seulement l'information pertinente dans le jeu de données servant à construire un modèle de prédiction minimisant les écarts pour les nouveaux échantillons. La variabilité intrinsèque (bruit) des échantillons sans rapport avec la variance informative pour la description des phénomènes étudiés ne doit pas être incluse dans le modèle, sous peine de surajustement («**overfitting**») du modèle. Cet effet diminue la performance des prédictions. Le sous-ajustement («**underfitting**») correspond à un modèle avec un nombre de variables latentes inférieur à ce qu'il faut pour inclure toute la variabilité nécessaire pour minimiser les erreurs de prédiction [178]. De ce fait, la PLS est fortement dépendante du choix judicieux du nombre de variables latentes à inclure dans le modèle, ce qui implique un équilibre entre l'extraction de la variabilité des matrices (ajustement) et un bon pouvoir prédictif. Plusieurs méthodes existent pour choisir le nombre de variables d'un modèle de régression et pour évaluer l'incertitude des modèles, dont la validation croisée [188,189], le «jackknifing »[190], le «bootstrapping»[191] et le test des permutations[192]. La validation croisée «cross-validation» estime l'exactitude des prédictions pour des modèles avec un nombre de variables latentes croissant. Dans son implémentation la plus simple, la matrice est partitionnée plusieurs fois en deux parties, une partie (**échantillons d'étalonnage** ou **Training Set**) est utilisée pour construire des modèles d'étalonnage avec différents nombres de variables latentes, et l'autre (**échantillons de validation** ou **Validation Set**) pour déterminer les erreurs de prédiction. Après avoir appliqué cette procédure à tous les échantillons, on calcule une valeur moyenne pour la somme des carrées des erreurs résiduelles de prédiction, ou PRESS («Predicted Residual Error Sum of Squares») pour les modèles avec différents nombres de variables latentes. La dimensionnalité optimale pour le modèle de régression PLS est celle qui minimise cette valeur de PRESS.

6.4.3.2. Test de permutation

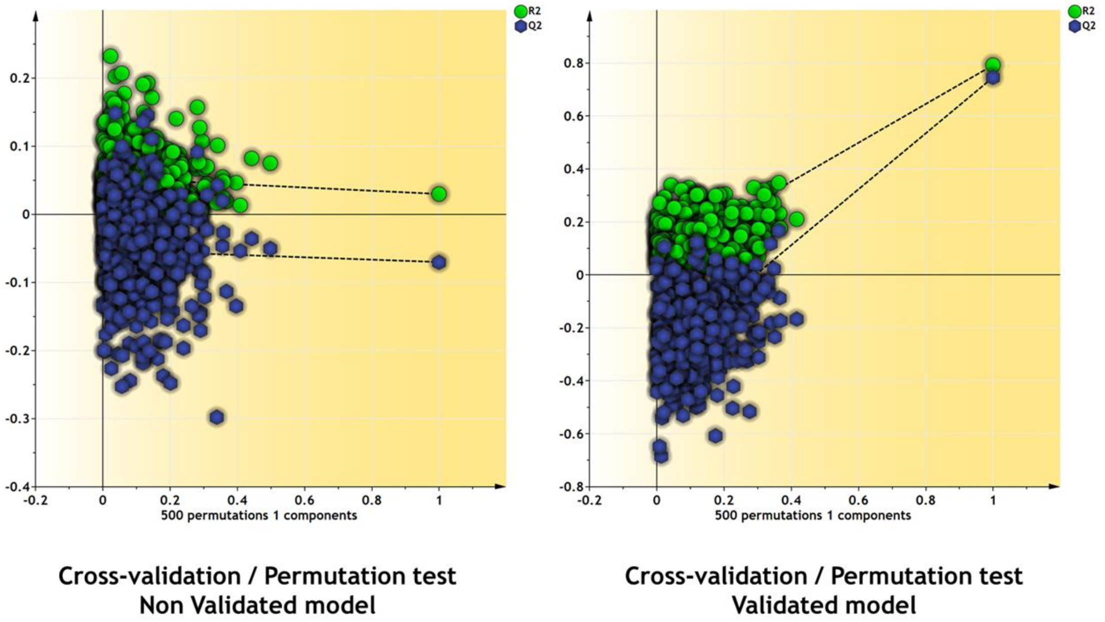
Pour avoir une estimation de la confiance sur un modèle de régression, avec un certain nombre de variables latentes dans le cas de la PLS, un test de permutations peut être utilisé. Donc, une fois la dimensionnalité optimale déterminée par validation croisée, ce test peut être appliqué pour estimer le niveau de signification du modèle. Les étapes de ce test sont comme suit :

1. Changer aléatoirement toutes les positions des valeurs de **y** et établir le modèle entre la matrice **X** et ce vecteur **y** perturbé pour un jeu d'échantillons d'étalonnage et calculer les valeurs prédites avec le jeu de validation.
2. Retenir les erreurs de prédiction (PRESS) pour le meilleur modèle
3. Répéter la procédure de permutation plusieurs fois (jusqu'à 999 fois sur SIMCA 14) et retenir les erreurs PRESS de tous les modèles pour chacun des couples de groupes d'étalonnage/prédiction
4. Comparer la valeur de PRESS du «vrai» modèle de validation croisée avec la distribution des erreurs des modèles «perturbés».

NB/ Dans le cas d'une discrimination, on peut utiliser la distance entre les barycentres des groupes, avec ou sans permutation des appartenances.

En résumé, la validation d'un modèle est appréciée par le degré de variation décrite par ce modèle, et par la précision de sa prédiction.

Dans le logiciel **SIMCA 14.0 Umertrics™**, utilisé dans ce travail, aussi bien pour la PCA, PLS-DA et la OPLS-DA, la variation expliquée par une composante dans le modèle est donnée par **R^2** qui a une valeur comprise entre **0** et **1**. Le **R^2** est défini par :**R^2 = 1 - RSS / SXXtot.corr.**, où RSS est la somme des carrés des résidus des données de la composante, et SXXtot.corr. est la variation totale de la matrice X centrée réduite. Plus R^2 est proche de 1, plus la variation de la matrice de données est expliquée par le modèle [182]. La validation des modèles PLS génère deux valeurs de **R^2** définis comme **R^2X** et **R^2Y**, décrivant la variation du modèle dans les matrices X et Y. En augmentant le nombre de composantes dans un modèle, le **R^2** peut être augmenté, mais cela risque d'aboutir à un surajustement « overfitting » du modèle, ce qui peut être évité par une validation croisée. Le terme **Q^2 = 1-PRESS/SXXtot.corr.**, où **PRESS** qui est la somme des moindres carrés entre les valeurs prédites et réelles (Predicted Residual Sum Of Squares) est utilisé comme un indicateur de la performance du modèle prédictif. Pour toutes les composantes successives construites pour un modèle, la validation croisée est réalisée jusqu'à ce que l'ajout de nouvelles composantes n'améliore plus le modèle. Un modèle prédictif est considéré significatif à partir d'un **Q^2 supérieur à 0,4 [182].**



Cross-validation / Permutation test
Non Validated model

Cross-validation / Permutation test
Validated model

Figure. 18. Exemples des résultats de test de permutation réalisé sur le logiciel SIMCA 14. Les axes Y représentent les valeurs R2Y et Q2Y de chaque modèle. Les axes X représentent le coefficient de corrélation entre la valeur « réelle » Y et la valeur « permutée » Y.

Droite : le modèle PLS est significatif et validé. Les valeurs de R2Y et Q2Y du modèle réel sont toujours supérieures aux valeurs après permutation.

Gauche : le modèle PLS n'est pas significatif donc pas validé. R2Y et Q2Y du modèle réel ne sont pas toujours supérieurs aux valeurs après permutation.

6.4.3.3. CV-ANOVA

CV-ANOVA est une analyse de variance des résidus de prédiction issus de la validation croisée. C'est un outil de diagnostic pour évaluer la fiabilité de la PLS et OPLS. Les avantages d'utiliser le CV-résidus sont qu'aucun calcul supplémentaire n'est nécessaire et que cette procédure garantit raisonnablement des données indépendantes et des estimations de la variance. Le test ANOVA est effectué sur la taille de la somme des résidus. Cela signifie que le but est de tester si le modèle (PLS/OPLS) présente une variabilité significative des résidus prédictifs et non pas une simple variation autour la moyenne. En résumé, le CV-ANOVA est un test de significativité de l'hypothèse nulle de résidus égaux des deux modèles comparés [193].

6.4.3.4. Validation externe

Alors que les méthodes de validation, décrites ci-dessus, peuvent nous éclairer sur la qualité des modèles, la prédiction d'une donnée externe peut apporter un niveau supérieur de validation quant à la qualité du modèle. Selon l'objectif de l'expérience, la validité d'un modèle peut être testée en permettant de prédire des échantillons qui ont été acquis ou traités indépendamment des données ayant permis de construire le modèle comme différents instruments et par différents opérateurs. Ces échantillons ne doivent pas avoir été utilisés pour la construction du modèle. La sensibilité et la spécificité peuvent être évaluées en cas d'analyse discriminante (classification). Par ailleurs, les mesures de l'erreur de prédiction peuvent être calculées en cas de modèle basé sur la régression (i. e. Erreur quadratique moyenne de la prédiction, RMSEP).

6.4.4. Sélection des variables discriminantes

Après validation du modèle, la sélection des variables discriminantes se fait selon diverses stratégies. Les méthodes classiques reposent sur l'application de tests statistiques (i. e. tests de student, Mann-Whitney, ANOVA). Cependant, d'autres métriques multivariées peuvent être envisagées couplées ou pas aux approches univariées. En métabolomique, le but du processus de modélisation et de trouver la combinaison la plus simple de métabolites qui peut prédire le phénomène biologique observé. Comme le processus de découverte des biomarqueurs implique deux paramètres principaux : l'utilité des biomarqueurs et le nombre de métabolites utilisés dans le modèle prédictif. Les principaux défis sont la sélection des prédicteurs (biomarqueurs) et l'évaluation de la puissance prédictive du modèle construit. La sélection des variables (métabolites) vise à extraire des métabolites importants des signaux détectés qui expliquent et prédisent le mieux la question biologique à l'étude. Statistiquement, c'est une approche d'optimisation qui récupère la meilleure combinaison de variables à partir des données. Différentes techniques de sélection de variables ont été décrites. Certaines de ces stratégies suggérées sont fondées sur des propriétés statistiques univariées ou multivariées de variables utilisées comme filtres (loadings, importance de variable sur les scores de projection (VIP) ou coefficients de régression), d'autres sont basées sur des algorithmes d'optimisation [194,195]. Les méthodes de filtrage classent les sous-ensembles de variables dans l'ordre d'importance avant de former les modèles puis de répéter le processus de modélisation en utilisant les métabolites les plus prédictifs. Chaque modèle de sous-ensemble est alors évalué jusqu'à atteindre la performance requise [196]. Dans le logiciel SIMCA, la représentation S-plot permet une sélection intuitive des variables potentiellement discriminantes basée sur les VIP ainsi que la covariance et la corrélation des variables (**Figure 19**). L'axe horizontal la covariance de la variable initiale, plus elle est éloignée du centre plus son intensité subit une forte variation entre les deux groupes. L'axe vertical représente la corrélation entre l'intensité de la variable et le groupe de l'échantillon, ainsi, plus on s'éloigne du centre, plus l'intensité de la variable est homogène au sein du groupe. Les variables discriminantes et potentiellement informatives se retrouvent sur les deux coins ; **supérieur droit** et **inférieur gauche** du S-plot qui représentent, en valeur absolue, le maximum de corrélation et de covariance des variables [187]. Une fois les variables choisies, elles sont classées en fonction de leur importance et de leur contribution dans la construction le modèle, donc de leur pouvoir discriminant. Plusieurs paramètres permettent de classer les variables. Dans ce travail, la classification par VIP (Variable importance). Le VIP est un coefficient qui résume la relation entre les variables Y et X. Ces coefficients sont analogues aux coefficients obtenus à partir de régression multiple. Le VIP Plot décrit quelles variables X qui caractérisent le mieux le bloc de variables X et corrélées avec Y. Ainsi, la valeur VIP résume l'ensemble des contributions de chaque variable X, sur toutes les autres composantes et pondéré en fonction de la variation Y représentée par chaque composante. Le score VIP est un outil de contrôle d'une importance capitale sur le choix des ions importants dans un ensemble complexe de données de métabolomique. Plus le score VIP est élevé (supérieur à 1 [197]), plus la significativité de la différence d'intensités de l'ion entre les groupes étudiés est élevée.

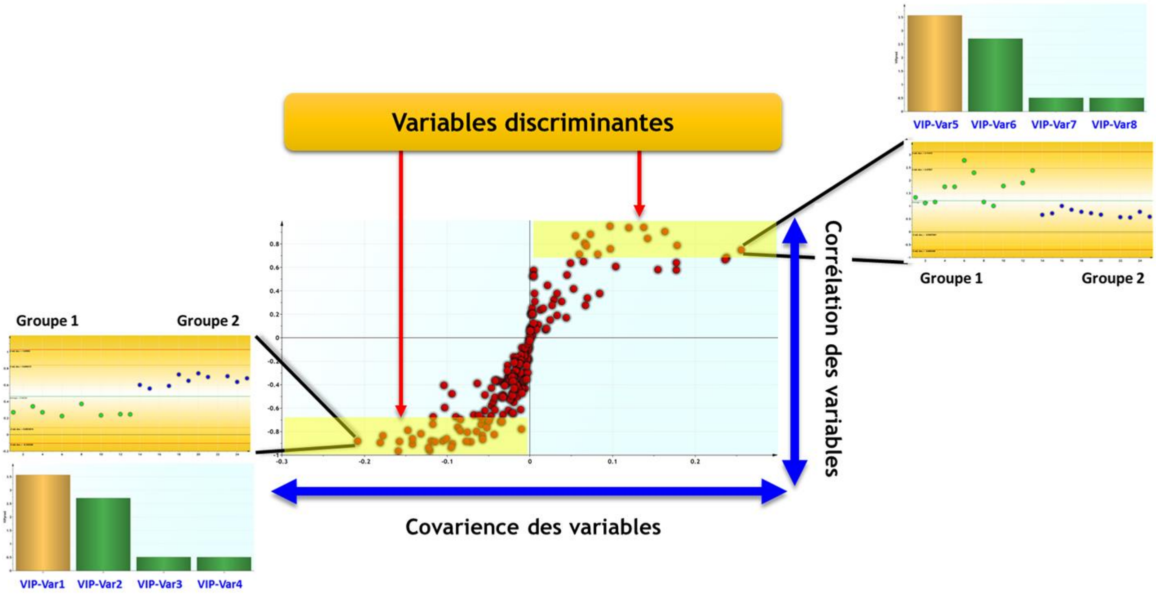


Figure 19. Représentation en S-Plot d'une OPLS-DA permettant la sélection des variables potentiellement discriminantes entre les groupes étudiés. Deux exemples de variables avec leurs intensités dans les échantillons et les VIP correspondants sont présentés. **Gauche**) Variable diminuée dans le Groupe 1. **Droite**) Variable augmentée dans le Groupe 1.

L'article II (A Tebani, C Afonso, S Bekri. Advances in metabolome information retrieval: from chemistry to biology. Part I: Analytical Chemistry of the Metabolome. JIMD. 2017. *Submitted*) et l'article III (A Tebani, C Afonso, S Bekri. Advances in metabolome information retrieval: from chemistry to biology. Part I: Biological Information Recovery. JIMD. 2017. *Submitted*) présente l'état de l'art des stratégies analytiques pour l'extraction de l'information métabolique d'un système biologique et les outils chimiométriques et bioinformatiques pour son interprétation et, enfin, sa transformation en support décisionnel qu'il soit translationel ou clinique.

RÉFÉRENCES

1. Oliver, S.G.; Winson, M.K.; Kell, D.B.; Baganz, F. Systematic functional analysis of the yeast genome. *Trends in biotechnology* **1998**, *16*, 373-378.
2. Tebani, A.; Abily-Donval, L.; Afonso, C.; Marret, S.; Bekri, S. Clinical metabolomics: The new metabolic window for inborn errors of metabolism investigations in the post-genomic era. *Int J Mol Sci* **2016**, *17*.
3. Bekri, S. The role of metabolomics in precision medicine. *Expert Review of Precision Medicine and Drug Development* **2016**.
4. Weckwerth, W.; Wenzel, K.; Fiehn, O. Process for the integrated extraction, identification and quantification of metabolites, proteins and rna to reveal their co-regulation in biochemical networks. *Proteomics* **2004**, *4*, 78-83.
5. Milne, S.B.; Mathews, T.P.; Myers, D.S.; Ivanova, P.T.; Brown, H.A. Sum of the parts: Mass spectrometry-based metabolomics. *Biochemistry* **2013**.
6. Dettmer, K.; Aronov, P.A.; Hammock, B.D. Mass spectrometry-based metabolomics. *Mass spectrometry reviews* **2007**, *26*, 51-78.
7. Vuckovic, D. Current trends and challenges in sample preparation for global metabolomics using liquid chromatography-mass spectrometry. *Anal Bioanal Chem* **2012**, *403*, 1523-1548.
8. Sitnikov, D.G.; Monnin, C.S.; Vuckovic, D. Systematic assessment of seven solvent and solid-phase extraction methods for metabolomics analysis of human plasma by lc-ms. *Scientific Reports* **2016**, *6*, 38885.
9. Shearer, G.C.; Harris, W.S.; Pedersen, T.L.; Newman, J.W. Detection of omega-3 oxylipins in human plasma and response to treatment with omega-3 acid ethyl esters. *Journal of lipid research* **2010**, *51*, 2074-2081.
10. Lewis, M.R.; Pearce, J.T.M.; Spagou, K.; Green, M.; Dona, A.C.; Yuen, A.H.Y.; David, M.; Berry, D.J.; Chappell, K.; Horneffer-van der Sluis, V., *et al.* Development and application of uplc-tof ms for precision large scale urinary metabolic phenotyping. *Analytical chemistry* **2016**.
11. Polson, C.; Sarkar, P.; Incledon, B.; Raguvaran, V.; Grant, R. Optimization of protein precipitation based upon effectiveness of protein removal and ionization effect in liquid chromatography-tandem mass spectrometry. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences* **2003**, *785*, 263-275.
12. Folch, J.; Lees, M.; Sloane Stanley, G.H. A simple method for the isolation and purification of total lipides from animal tissues. *The Journal of biological chemistry* **1957**, *226*, 497-509.
13. Bligh, E.G.; Dyer, W.J. A rapid method of total lipid extraction and purification. *Can J Biochem Phys* **1959**, *37*, 911-917.
14. Cequier-Sanchez, E.; Rodriguez, C.; Ravelo, A.G.; Zarate, R. Dichloromethane as a solvent for lipid extraction and assessment of lipid classes and fatty acids from samples of different natures. *J Agr Food Chem* **2008**, *56*, 4297-4303.
15. Smedes, F. Determination of total lipid using non-chlorinated solvents. *Analyst* **1999**, *124*, 1711-1718.
16. Smedes, F.; Thomasen, T.K. Evaluation of the bligh & dyer lipid determination method. *Mar Pollut Bull* **1996**, *32*, 681-688.
17. Smedes, F.; Askland, T.K. Revisiting the development of the bligh and dyer total lipid determination method. *Mar Pollut Bull* **1999**, *38*, 193-201.
18. Matyash, V.; Liebisch, G.; Kurzchalia, T.V.; Shevchenko, A.; Schwudke, D. Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. *Journal of lipid research* **2008**, *49*, 1137-1146.
19. Wolf, C.; Quinn, P.J. Lipidomics: Practical aspects and applications. *Progress in lipid research* **2008**, *47*, 15-36.
20. 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.
21. Nicholson, J.K.; Wilson, I.D. High-resolution proton magnetic-resonance spectroscopy of biological-fluids. *Prog Nucl Mag Res Sp* **1989**, *21*, 449-501.
22. Holmes, E.; Nicholson, J.K.; Nicholls, A.W.; Lindon, J.C.; Connor, S.C.; Polley, S.; Connelly, J. The identification of novel biomarkers of renal toxicity using automatic data reduction techniques and pca of proton nmr spectra of urine. *Chemometr Intell Lab* **1998**, *44*, 245-255.
23. Lenz, E.M.; Bright, J.; Wilson, I.D.; Morgan, S.R.; Nash, A.F.P. A h-1 nmr-based metabonomic study of urine and plasma samples obtained from healthy human subjects. *J Pharmaceut Biomed* **2003**, *33*, 1103-1115.
24. Nicholson, J.K.; Connelly, J.; Lindon, J.C.; Holmes, E. Metabonomics: A platform for studying drug toxicity and gene function. *Nature Reviews Drug Discovery* **2002**, *1*, 153-161.
25. Pulfer, M.; Murphy, R.C. Electrospray mass spectrometry of phospholipids. *Mass spectrometry reviews* **2003**, *22*, 332-364.
26. Theodoridis, G.; Gika, H.G.; Wilson, I.D. Mass spectrometry-based holistic analytical approaches for metabolite profiling in systems biology studies. *Mass spectrometry reviews* **2011**, *30*, 884-906.
27. Köfeler, H.C.; Fauland, A.; Rechberger, G.N.; Trötzmüller, M. Mass spectrometry based lipidomics: An overview of technological platforms. *Metabolites* **2012**, *2*, 19-38.
28. Ivanova, P.T.; Milne, S.B.; Myers, D.S.; Brown, H.A. Lipidomics: A mass spectrometry based systems level analysis of cellular lipids. *Curr Opin Chem Biol* **2009**, *13*, 526-531.
29. Han, X.L.; Yang, K.; Gross, R.W. Multi-dimensional mass spectrometry-based shotgun lipidomics and novel strategies for lipidomic analyses. *Mass spectrometry reviews* **2012**, *31*, 134-178.

30. Fuchs, B.; Suss, R.; Schiller, J. An update of maldi-tof mass spectrometry in lipid research. *Progress in lipid research* **2011**, *50*, 132.
31. Schwudke, D.; Liebisch, G.; Herzog, R.; Schmitz, G.; Shevchenko, A. Shotgun lipidomics by tandem mass spectrometry under data-dependent acquisition control. *Method Enzymol* **2007**, *433*, 175-+.
32. Han, X.L.; Yang, J.Y.; Cheng, H.; Yang, K.; Abendschein, D.R.; Gross, R.W. Shotgun lipidomics identifies cardiolipin depletion in diabetic myocardium linking altered substrate utilization with mitochondrial dysfunction. *Biochemistry* **2005**, *44*, 16684-16694.
33. Habchi, B.; Alves, S.; Paris, A.; Rutledge, D.N.; Rathahao-Paris, E. How to really perform high throughput metabolomic analyses efficiently? *TrAC Trends in Analytical Chemistry* **2016**, *85*, 128-139.
34. Loizides-Mangold, U. On the future of mass spectrometry based lipidomics. *The FEBS journal* **2013**.
35. Murphy, R.C.; Gaskell, S.J. New applications of mass spectrometry in lipid analysis. *Journal of Biological Chemistry* **2011**, *286*, 25427-25433.
36. Goodwin, C.R.; Sherrod, S.D.; Marasco, C.C.; Bachmann, B.O.; Schramm-Sapyta, N.; Wikswo, J.P.; McLean, J.A. Phenotypic mapping of metabolic profiles using self-organizing maps of high-dimensional mass spectrometry data. *Analytical chemistry* **2014**, *86*, 6563-6571.
37. Kind, T.; Fiehn, O. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical reviews* **2010**, *2*, 23-60.
38. Annesley, T.M. Ion suppression in mass spectrometry. *Clinical chemistry* **2003**, *49*, 1041-1044.
39. Kind, T.; Fiehn, O. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics* **2007**, *8*, 105.
40. Kim, S.; Rodgers, R.P.; Marshall, A.G. Truly "exact" mass: Elemental composition can be determined uniquely from molecular mass measurement at~ 0.1 mda accuracy for molecules up to~ 500da. *International Journal of Mass Spectrometry* **2006**, *251*, 260-265.
41. Savory, J.J.; Kaiser, N.K.; McKenna, A.M.; Xian, F.; Blakney, G.T.; Rodgers, R.P.; Hendrickson, C.L.; Marshall, A.G. Parts-per-billion fourier transform ion cyclotron resonance mass measurement accuracy with a "walking" calibration equation. *Analytical chemistry* **2011**, *83*, 1732-1736.
42. Scheltema, R.A.; Kamleh, A.; Wildridge, D.; Ebikeme, C.; Watson, D.G.; Barrett, M.P.; Jansen, R.C.; Breitling, R. Increasing the mass accuracy of high-resolution lc-ms data using background ions–a case study on the ltq-orbitrap. *Proteomics* **2008**, *8*, 4647-4656.
43. Miladinović, S.M.; Kozhinov, A.N.; Gorshkov, M.V.; Tsybin, Y.O. On the utility of isotopic fine structure mass spectrometry in protein identification. *Analytical chemistry* **2012**, *84*, 4042-4051.
44. 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. *Journal of Chromatography A* **2013**, *1292*, 51-65.
45. Yergey, A.L.; Edmonds, C.G.; Lewis, I.A.; Vestal, M.L. *Liquid chromatography/mass spectrometry: Techniques and applications*. Springer Science & Business Media: 2013.
46. Spengler, B. Mass spectrometry imaging of biomolecular information. *Analytical chemistry* **2015**, *87*, 64-82.
47. May, J.C.; McLean, J.A. Ion mobility-mass spectrometry: Time-dispersive instrumentation. *Analytical chemistry* **2015**, *87*, 1422-1436.
48. Schrimpe-Rutledge, A.C.; Codreanu, S.G.; Sherrod, S.D.; McLean, J.A. Untargeted metabolomics strategies—challenges and emerging directions. *Journal of The American Society for Mass Spectrometry* **2016**, *27*, 1897-1905.
49. May, J.C.; McLean, J.A. Advanced multidimensional separations in mass spectrometry: Navigating the big data deluge. *Annual review of analytical chemistry* **2016**, *9*, 387-409.
50. Tebani, A.; Schmitz-Afonso, I.; Rutledge, D.N.; Gonzalez, B.J.; Bekri, S.; Afonso, C. Optimization of a liquid chromatography ion mobility-mass spectrometry method for untargeted metabolomics using experimental design and multivariate data analysis. *Analytica Chimica Acta* **2016**, *913*, 55-62.
51. Dole, M.; Hines, R.; Mack, L.; Mobley, R.; Ferguson, L.; Alice, M. Gas phase macroions. *Macromolecules* **1968**, *1*, 96-97.
52. Whitehouse, C.M.; Dreyer, R.; Yamashita, M.; Fenn, J. Electrospray ionization for mass-spectrometry of large biomolecules. *Science (New York, N.Y.)* **1989**, *246*, 64-71.
53. Lee, M.S. *Mass spectrometry handbook*. Wiley: 2012.
54. Tebani, A.; Afonso, C.; Marret, S.; Bekri, S. Omics-based strategies in precision medicine: Toward a paradigm shift in inborn errors of metabolism investigations. *Int J Mol Sci* **2016**, *17*.
55. Andra, S.S.; Austin, C.; Patel, D.; Dolios, G.; Awawda, M.; Arora, M. Trends in the application of high-resolution mass spectrometry for human biomonitoring: An analytical primer to studying the environmental chemical space of the human exposome. *Environment international* **2017**.
56. Maurer, H.H.; Meyer, M.R. High-resolution mass spectrometry in toxicology: Current status and future perspectives. *Archives of toxicology* **2016**, *90*, 2161-2172.
57. Lesur, A.; Domon, B. Advances in high-resolution accurate mass spectrometry application to targeted proteomics. *Proteomics* **2015**, *15*, 880-890.
58. Ibanez, C.; Simo, C.; Garcia-Canas, V.; Acunha, T.; Cifuentes, A. The role of direct high-resolution mass spectrometry in foodomics. *Anal Bioanal Chem* **2015**, *407*, 6275-6287.

59. Haggarty, J.; Burgess, K.E.V. Recent advances in liquid and gas chromatography methodology for extending coverage of the metabolome. *Current Opinion in Biotechnology* **2017**, *43*, 77-85.
60. Kaufmann, A. Combining uhplc and high-resolution ms: A viable approach for the analysis of complex samples? *TrAC Trends in Analytical Chemistry* **2014**, *63*, 113-128.
61. 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. *Journal of chromatography. A* **2013**, *1292*.
62. Tang, D.Q.; Zou, L.; Yin, X.X.; Ong, C.N. Hilic-ms for metabolomics: An attractive and complementary approach to rplc-ms. *Mass spectrometry reviews* **2014**.
63. Hill, H.H., Jr.; Siems, W.F.; St Louis, R.H.; McMinn, D.G. Ion mobility spectrometry. *Analytical chemistry* **1990**, *62*, 1201A-1209A.
64. Clemmer, D.E.; Jarrold, M.F. Ion mobility measurements and their applications to clusters and biomolecules. *Journal of Mass Spectrometry* **1997**, *32*, 577-592.
65. May, J.C.; Morris, C.B.; McLean, J.A. Ion mobility collision cross section compendium. *Analytical chemistry* **2016**.
66. Kliman, M.; May, J.C.; McLean, J.A. Lipid analysis and lipidomics by structurally selective ion mobility-mass spectrometry. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **2011**, *1811*, 935-945.
67. Kolakowski, B.M.; Mester, Z. Review of applications of high-field asymmetric waveform ion mobility spectrometry (faims) and differential mobility spectrometry (dms). *The Analyst* **2007**, *132*, 842-864.
68. de la Mora, J.F.; Ude, S.; Thomson, B.A. The potential of differential mobility analysis coupled to ms for the study of very large singly and multiply charged proteins and protein complexes in the gas phase. *Biotechnol J* **2006**, *1*, 988-997.
69. Vidal-de-Miguel, G.; Macía, M.; Cuevas, J. Transversal modulation ion mobility spectrometry (tm-ims), a new mobility filter overcoming turbulence related limitations. *Analytical chemistry* **2012**, *84*, 7831-7837.
70. Michelmann, K.; Silveira, J.A.; Ridgeway, M.E.; Park, M.A. Fundamentals of trapped ion mobility spectrometry. *Journal of The American Society for Mass Spectrometry* **2015**, *26*, 14-24.
71. Clemmer, D.E. http://www.indiana.edu/~clemmer/Research/CrossSectionDatabase/cs_database.php.
72. Bush, M.F. http://depts.washington.edu/bushlab/ccsdatabase
73. Smith, D.P.; Knapman, T.W.; Campuzano, I.; Malham, R.W.; Berryman, J.T.; Radford, S.E.; Ashcroft, A.E. Deciphering drift time measurements from travelling wave ion mobility spectrometry-mass spectrometry studies. *European journal of mass spectrometry (Chichester, England)* **2009**, *15*, 113-130.
74. Katajamaa, M.; Oresic, M. Data processing for mass spectrometry-based metabolomics. *Journal of chromatography. A* **2007**, *1158*, 318-328.
75. Sugimoto, M.; Kawakami, M.; Robert, M.; Soga, T.; Tomita, M. Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. *Current bioinformatics* **2012**, *7*, 96-108.
76. Misra, B.B.; der Hooft, J.J. Updates in metabolomics tools and resources: 2014–2015. *Electrophoresis* **2016**, *37*, 86-110.
77. 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. *Analytical chemistry* **2006**, *78*, 779-787.
78. Tautenhahn, R.; Patti, G.J.; Kalisiak, E.; Miyamoto, T.; Schmidt, M.; Lo, F.Y.; McBee, J.; Baliga, N.S.; Siuzdak, G. Metaxcms: Second-order analysis of untargeted metabolomics data. *Analytical chemistry* **2011**, *83*, 696-700.
79. Benton, H.P.; Wong, D.M.; Trauger, S.A.; Siuzdak, G. Xcms2: Processing tandem mass spectrometry data for metabolite identification and structural characterization. *Analytical chemistry* **2008**, *80*, 6382-6389.
80. Xia, J.; Mandal, R.; Sinelnikov, I.V.; Broadhurst, D.; Wishart, D.S. Metaboanalyst 2.0--a comprehensive server for metabolomic data analysis. *Nucleic acids research* **2012**, *40*, W127-133.
81. Xia, J.; Broadhurst, D.I.; Wilson, M.; Wishart, D.S. Translational biomarker discovery in clinical metabolomics: An introductory tutorial. *Metabolomics* **2013**, *9*, 280-299.
82. 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.
83. Kastenmuller, G.; Romisch-Margl, W.; Wagele, B.; Altmaier, E.; Suhre, K. Metap-server: A web-based metabolomics data analysis tool. *Journal of biomedicine & biotechnology* **2011**, *2011*.
84. Neuweger, H.; Albaum, S.P.; Dondrup, M.; Persicke, M.; Watt, T.; Niehaus, K.; Stoye, J.; Goesmann, A. Meltdb: A software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics* **2008**, *24*, 2726-2732.
85. Lommen, A. Metalign: Interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal Chem* **2009**, *81*, 3079-3086.
86. Castillo, S.; Gopalacharyulu, P.; Yetukuri, L.; Orešič, M. Algorithms and tools for the preprocessing of lc–ms metabolomics data. *Chemometr Intell Lab* **2011**, *108*, 23-32.
87. Daszykowski, M.; Walczak, B. Use and abuse of chemometrics in chromatography. *Trac-Trend Anal Chem* **2006**, *25*, 1081-1096.
88. Savitzky, A.; Golay, M.J.E. Smoothing + differentiation of data by simplified least squares procedures. *Anal Chem* **1964**, *36*, 1627-&.

89. Szymańska, E.; Brodrick, E.; Williams, M.; Davies, A.N.; van Manen, H.-J.; Buydens, L.M.C. Data size reduction strategy for the classification of breath and air samples using multicapillary column-ion mobility spectrometry. *Analytical chemistry* **2015**, *87*, 869-875.
90. Bader, S.; Urfer, W.; Baumbach, J.I. Preprocessing of ion mobility spectra by lognormal detailing and wavelet transform. *International Journal for Ion Mobility Spectrometry* **2008**, *11*, 43-49.
91. Urbas, A.A.; Harrington, P.B. Two-dimensional wavelet compression of ion mobility spectra. *Analytica Chimica Acta* **2001**, *446*, 391-410.
92. Tautenhahn, R.; Bottcher, C.; Neumann, S. Highly sensitive feature detection for high resolution lc/ms. *BMC bioinformatics* **2008**, *9*, 504.
93. Yu, T.W.; Peng, H.S. Quantification and deconvolution of asymmetric lc-ms peaks using the bi-gaussian mixture model and statistical model selection. *BMC bioinformatics* **2010**, *11*.
94. Lange, E.; Tautenhahn, R.; Neumann, S.; Gropl, C. Critical assessment of alignment procedures for lc-ms proteomics and metabolomics measurements. *BMC Bioinf.* **2008**, *9*, 375.
95. Szymanska, E.; Davies, A.; Buydens, L. Chemometrics for ion mobility spectrometry data: Recent advances and future prospects. *The Analyst* **2016**.
96. Yu, T.; Park, Y.; Johnson, J.M.; Jones, D.P. Aplcms--adaptive processing of high-resolution lc/ms data. *Bioinformatics (Oxford, England)* **2009**, *25*, 1930.
97. Tautenhahn, R.; Bottcher, C.; Neumann, S. Highly sensitive feature detection for high resolution lc/ms. *BMC Bioinf.* **2008**, *9*, 504.
98. Mahieu, N.G.; Spalding, J.L.; Patti, G.J. Warpgroup: Increased precision of metabolomic data processing by consensus integration bound analysis. *Bioinformatics (Oxford, England)* **2016**, *32*, 268.
99. Khayamian, T.; Sajjadi, S.M.; Mirmahdieh, S.; Mardihallaj, A.; Hashemian, Z. Simultaneous analysis of bifenthrin and tetramethrin using corona discharge ion mobility spectrometry and tucker 3 model. *Chemometrics and Intelligent Laboratory Systems* **2012**, *118*, 88-96.
100. Lange, E.; Tautenhahn, R.; Neumann, S.; Gropl, C. Critical assessment of alignment procedures for lc-ms proteomics and metabolomics measurements. *BMC bioinformatics* **2008**, *9*, 375.
101. Eliasson, M.; Rannar, S.; Madsen, R.; Donten, M.A.; Marsden-Edwards, E.; Moritz, T.; Shockcor, J.P.; Johansson, E.; Trygg, J. Strategy for optimizing lc-ms data processing in metabolomics: A design of experiments approach. *Analytical chemistry* **2012**, *84*, 6869-6876.
102. Brodsky, L.; Moussaieff, A.; Shahaf, N.; Aharoni, A.; Rogachev, I. Evaluation of peak picking quality in lc-ms metabolomics data. *Anal Chem* **2010**, *82*, 9177-9187.
103. Libiseller, G.; Dvorzak, M.; Kleb, U.; Gander, E.; Eisenberg, T.; Madeo, F.; Neumann, S.; Trausinger, G.; Sinner, F.; Pieber, T., *et al.* Ipo: A tool for automated optimization of xcms parameters. *BMC bioinformatics* **2015**, *16*, 118.
104. Tebani, A.; Schmitz-Afonso, I.; Rutledge, D.N.; Gonzalez, B.J.; Bekri, S.; Afonso, C. Optimization of a liquid chromatography ion mobility-mass spectrometry method for untargeted metabolomics using experimental design and multivariate data analysis. *Anal Chim Acta* **2016**, *913*, 55-62.
105. Uppal, K.; Soltow, Q.A.; Strobel, F.H.; Pittard, W.S.; Gernert, K.M.; Yu, T.; Jones, D.P. Xmsanalyzer: Automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. *BMC Bioinf.* **2013**, *14*, 15.
106. Zschocke, J. Disorders of the biosynthesis and breakdown of complex molecules. In *Inherited metabolic diseases: A clinical approach*, Hoffmann, G.F.; Zschocke, J.; Nyhan, W.L., Eds. Springer Berlin Heidelberg: Berlin, Heidelberg, 2017; pp 9-12.
107. R. Tautenhahn, C.B.t.a.S.N. Annotation of lc/esi-ms mass signals. In *Bioinformatics research and development*, Wagner, S.H.a.R., Ed. Springer: Berlin/Heidelberg, 2007 Vol. 4414, pp 371–380.
108. Ipsen, A.; Want, E.J.; Lindon, J.C.; Ebbels, T.M. A statistically rigorous test for the identification of parent-fragment pairs in lc-ms datasets. *Anal Chem* **2010**, *82*, 1766-1778.
109. Roux, A.; Lison, D.; Junot, C.; Heilier, J.F. Applications of liquid chromatography coupled to mass spectrometry-based metabolomics in clinical chemistry and toxicology: A review. *Clinical biochemistry* **2011**, *44*, 119-135.
110. Lafaye, A.; Junot, C.; Ramounet-Le Gall, B.; Fritsch, P.; Tabet, J.C.; Ezan, E. Metabolite profiling in rat urine by liquid chromatography/electrospray ion trap mass spectrometry. Application to the study of heavy metal toxicity. *Rapid communications in mass spectrometry : RCM* **2003**, *17*, 2541-2549.
111. Bogusz, M.J.; Maier, R.D.; Kruger, K.D.; Webb, K.S.; Romeril, J.; Miller, M.L. Poor reproducibility of in-source collisional atmospheric pressure ionization mass spectra of toxicologically relevant drugs. *Journal of Chromatography A* **1999**, *844*, 409-418.
112. Cui, Q.; Lewis, I.A.; Hegeman, A.D.; Anderson, M.E.; Li, J.; Schulte, C.F.; Westler, W.M.; Eghbalnia, H.R.; Sussman, M.R.; Markley, J.L. Metabolite identification via the madison metabolomics consortium database. *Nature biotechnology* **2008**, *26*, 162-164.
113. Sumner, L.W.; Urbanczyk-Wochniak, E.; Broeckling, C.D. Metabolomics data analysis, visualization, and integration. *Methods Mol Biol* **2007**, *406*, 409-436.

114. Sumner, L.W.; Amberg, A.; Barrett, D.; Beale, M.H.; Beger, R.; Daykin, C.A.; Fan, T.W.M.; Fiehn, O.; Goodacre, R.; Griffin, J.L., *et al.* Proposed minimum reporting standards for chemical analysis. *Metabolomics : Official journal of the Metabolomic Society* **2007**, *3*, 211-221.
115. Schymanski, E.L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H.P.; Hollender, J. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environ. Sci. Technol.* **2014**, *48*, 2097.
116. 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 (Oxford, England)* **2011**, *27*, 1339.
117. 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.
118. Silva, R.R.; Jourdan, F.; Salvanha, D.M.; Letisse, F.; Jamin, E.L.; Guidetti-Gonzalez, S.; Labate, C.A.; Vencio, R.Z. Probmetab: An r package for bayesian probabilistic annotation of lc-ms-based metabolomics. *Bioinformatics (Oxford, England)* **2014**, *30*, 1336.
119. Daly, R.; Rogers, S.; Wandy, J.; Jankevics, A.; Burgess, K.E.; Breitling, R. Metassign: Probabilistic annotation of metabolites from lc-ms data using a bayesian clustering approach. *Bioinformatics (Oxford, England)* **2014**, *30*, 2764.
120. Li, S.; Park, Y.; Duraisingham, S.; Strobel, F.H.; Khan, N.; Soltow, Q.A.; Jones, D.P.; Pulendran, B. Predicting network activity from high throughput metabolomics. *PLoS Comput. Biol.* **2013**, *9*, e1003123.
121. Pirhaji, L.; Milani, P.; Leidl, M.; Curran, T.; Avila-Pacheco, J.; Clish, C.B.; White, F.M.; Saghatelian, A.; Fraenkel, E. Revealing disease-associated pathways by network integration of untargeted metabolomics. *Nat Meth* **2016**, *13*, 770-776.
122. Kind, T.; Fiehn, O. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal. Rev.* **2010**, *2*, 23.
123. Zhou, Z.; Shen, X.; Tu, J.; Zhu, Z.-J. Large-scale prediction of collision cross-section values for metabolites in ion mobility-mass spectrometry. *Analytical chemistry* **2016**.
124. Mallard, W.G.; Andriamaharavo, N.R.; Mirokhin, Y.A.; Halket, J.M.; Stein, S.E. Creation of libraries of recurring mass spectra from large data sets assisted by a dual-column workflow. *Anal. Chem.* **2014**, *86*, 10231.
125. Boswell, P.G.; Schellenberg, J.R.; Carr, P.W.; Cohen, J.D.; Hegeman, A.D. Easy and accurate high-performance liquid chromatography retention prediction with different gradients, flow rates, and instruments by back-calculation of gradient and flow rate profiles. *J. Chromatogr. A* **2011**, *1218*, 6742.
126. Vaniya, A.; Fiehn, O. Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics. *TrAC, Trends Anal. Chem.* **2015**, *69*, 52.
127. Ma, Y.; Kind, T.; Yang, D.; Leon, C.; Fiehn, O. Ms2analyzer: A software for small molecule substructure annotations from accurate tandem mass spectra. *Anal. Chem.* **2014**, *86*, 10724.
128. Yang, J.Y.; Sanchez, L.M.; Rath, C.M.; Liu, X.; Boudreau, P.D.; Bruns, N.; Glukhov, E.; Wodtke, A.; de Felicio, R.; Fenner, A., *et al.* Molecular networking as a dereplication strategy. *J. Nat. Prod.* **2013**, *76*, 1686.
129. Allard, P.-M.; Péresse, T.; Bisson, J.; Gindro, K.; Marcourt, L.; Pham, V.C.; Roussi, F.; Litaudon, M.; Wolfender, J.-L. Integration of molecular networking and in-silico ms/ms fragmentation for natural products dereplication. *Analytical chemistry* **2016**, *88*, 3317-3323.
130. Quinn, R.A.; Phelan, V.V.; Whiteson, K.L.; Garg, N.; Bailey, B.A.; Lim, Y.W.; Conrad, D.J.; Dorrestein, P.C.; Rohwer, F.L. Microbial, host and xenobiotic diversity in the cystic fibrosis sputum metabolome. *ISME J.* **2016**, *10*, 1483.
131. Lapthorn, C.; Pullen, F.; Chowdhry, B.Z. Ion mobility spectrometry-mass spectrometry (ims-ms) of small molecules: Separating and assigning structures to ions. *Mass Spectrom. Rev.* **2013**, *32*, 43.
132. Campuzano, I.; Bush, M.F.; Robinson, C.V.; Beaumont, C.; Richardson, K.; Kim, H.; Kim, H.I. Structural characterization of drug-like compounds by ion mobility mass spectrometry: Comparison of theoretical and experimentally derived nitrogen collision cross sections. *Anal. Chem.* **2012**, *84*, 1026.
133. Reading, E.; Munoz-Muriedas, J.; Roberts, A.D.; Dear, G.J.; Robinson, C.V.; Beaumont, C. Elucidation of drug metabolite structural isomers using molecular modeling coupled with ion mobility mass spectrometry. *Anal. Chem.* **2016**, *88*, 2273.
134. Paglia, G.; Williams, J.P.; Menikarachchi, L.; Thompson, J.W.; Tyldesley-Worster, R.; Halldórsson, S.; Rolfsson, O.; Moseley, A.; Grant, D.; Langridge, J., *et al.* Ion mobility derived collision cross sections to support metabolomics applications. *Analytical chemistry* **2014**, *86*, 3985-3993.
135. Paglia, G.; Kliman, M.; Claude, E.; Geromanos, S.; Astarita, G. Applications of ion-mobility mass spectrometry for lipid analysis. *Anal. Bioanal. Chem.* **2015**, *407*, 4995.
136. Paglia, G.; Angel, P.; Williams, J.P.; Richardson, K.; Olivos, H.J.; Thompson, J.W.; Menikarachchi, L.; Lai, S.; Walsh, C.; Moseley, A., *et al.* Ion mobility-derived collision cross section as an additional measure for lipid fingerprinting and identification. *Analytical chemistry* **2015**, *87*, 1137-1144.
137. Henry, V.J.; Bandrowski, A.E.; Pepin, A.-S.; Gonzalez, B.J.; Desfeux, A. Omictools: An informative directory for multi-omic data analysis. *Database* **2014**, *2014*, bau069.
138. Wishart, D.S. Chapter 3: Small molecules and disease. *Plos Comput Biol* **2012**, *8*.
139. Werner, E.; Heilier, J.F.; Ducruix, C.; Ezan, E.; Junot, C.; Tabet, J.C. Mass spectrometry for the identification of the discriminating signals from metabolomics: Current status and future trends. *J Chromatogr B* **2008**, *871*, 143-163.
140. Go, E.P. Database resources in metabolomics: An overview. *J Neuroimmune Pharm* **2010**, *5*, 18-30.

141. Xia, J.; Sinelnikov, I.V.; Han, B.; Wishart, D.S. Metaboanalyst 3.0-making metabolomics more meaningful. *Nucleic acids research* **2015**.
142. Hughes, G.; Cruickshank-Quinn, C.; Reisdorph, R.; Lutz, S.; Petrache, I.; Reisdorph, N.; Bowler, R.; Kechris, K. Msprep--summarization, normalization and diagnostics for processing of mass spectrometry-based metabolomic data. *Bioinformatics (Oxford, England)* **2014**, *30*, 133.
143. Simader, A.M.; Kluger, B.; Neumann, N.K.; Bueschl, C.; Lemmens, M.; Lirk, G.; Krska, R.; Schuhmacher, R. Qcscreen: A software tool for data quality control in lc-hrms based metabolomics. *BMC bioinformatics* **2015**, *16*, 341.
144. Naz, S.; Vallejo, M.; García, A.; Barbas, C. Method validation strategies involved in non-targeted metabolomics. *Journal of Chromatography A* **2014**, *1353*, 99-105.
145. Mertens, B.J. Transformation, normalization, and batch effect in the analysis of mass spectrometry data for omics studies. In *Statistical analysis of proteomics, metabolomics, and lipidomics data using mass spectrometry*, Springer: 2017; pp 1-21.
146. Johnson, C.H.; Ivanisevic, J.; Benton, H.P.; Siuzdak, G. Bioinformatics: The next frontier of metabolomics. *Anal. Chem.* **2015**, *87*, 147.
147. Cajka, T.; Fiehn, O. Toward merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics. *Analytical chemistry* **2015**.
148. Kind, T.; Fiehn, O. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinf.* **2006**, *7*, 234.
149. De Livera, A.M.; Dias, D.A.; De Souza, D.; Rupasinghe, T.; Pyke, J.; Tull, D.; Roessner, U.; McConville, M.; Speed, T.P. Normalizing and integrating metabolomics data. *Analytical chemistry* **2012**, *84*, 10768-10776.
150. Xia, J.; Wishart, D.S. Web-based inference of biological patterns, functions and pathways from metabolomic data using metaboanalyst. *Nature protocols* **2011**, *6*, 743-760.
151. Mak, T.D.; Laiakis, E.C.; Goudarzi, M.; Fornace, A.J. Selective paired ion contrast analysis: A novel algorithm for analyzing postprocessed lc-ms metabolomics data possessing high experimental noise. *Analytical chemistry* **2015**, *87*, 3177-3186.
152. Grun, D.; Kester, L.; van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat Meth* **2014**, *11*, 637-640.
153. Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M., *et al.* Ncbi geo: Archive for functional genomics data sets—update. *Nucleic acids research* **2012**, *41*, D991-D995.
154. Haug, K.; Salek, R.M.; Conesa, P.; Hastings, J.; de Matos, P.; Rijnbeek, M.; Mahendraker, T.; Williams, M.; Neumann, S.; Rocca-Serra, P., *et al.* Metabolights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic acids research* **2012**, *41*, D781-D786.
155. Li, B.; Tang, J.; Yang, Q.; Cui, X.; Li, S.; Chen, S.; Cao, Q.; Xue, W.; Chen, N.; Zhu, F. Performance evaluation and online realization of data-driven normalization methods used in lc/ms based untargeted metabolomics analysis. *Sci Rep* **2016**, *6*, 38881.
156. Scholz, M.; Gatzek, S.; Sterling, A.; Fiehn, O.; Selbig, J. Metabolite fingerprinting: Detecting biological features by independent component analysis. *Bioinformatics* **2004**, *20*, 2447-2454.
157. Wang, W.X.; Zhou, H.H.; Lin, H.; Roy, S.; Shaler, T.A.; Hill, L.R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C.H. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal Chem* **2003**, *75*, 4818-4826.
158. Lee, J.; Park, J.; Lim, M.S.; Seong, S.J.; Seo, J.J.; Park, S.M.; Lee, H.W.; Yoon, Y.R. Quantile normalization approach for liquid chromatography-mass spectrometry-based metabolomic data from healthy human volunteers. *Analytical sciences : the international journal of the Japan Society for Analytical Chemistry* **2012**, *28*, 801-805.
159. Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1h nmr metabonomics. *Analytical chemistry* **2006**, *78*, 4281-4290.
160. Bolstad, B.M.; Irizarry, R.A.; Astrand, M.; Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)* **2003**, *19*, 185-193.
161. Li, C.; Wong, W.H. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America* **2001**, *98*, 31-36.
162. Astrand, M. Contrast normalization of oligonucleotide arrays. *J Comput Biol* **2003**, *10*, 95-102.
163. Workman, C.; Jensen, L.J.; Jarmer, H.; Berka, R.; Gautier, L.; Nielser, H.B.; Saxild, H.-H.; Nielsen, C.; Brunak, S.; Knudsen, S. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology* **2002**, *3*, research0048.0041-research0048.0016.
164. Dudoit, S.; Yang, Y.H.; Callow, M.J.; Speed, T.P. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica sinica* **2002**, 111-139.
165. Kirwan, J.; Broadhurst, D.; Davidson, R.; Viant, M. Characterising and correcting batch variation in an automated direct infusion mass spectrometry (dims) metabolomics workflow. *Analytical and bioanalytical chemistry* **2013**, *405*, 5147-5157.

166. Warrack, B.M.; Hnatyshyn, S.; Ott, K.-H.; Reily, M.D.; Sanders, M.; Zhang, H.; Drexler, D.M. Normalization strategies for metabonomic analysis of urine samples. *Journal of Chromatography B* **2009**, *877*, 547-552.
167. Shen, X.; Gong, X.; Cai, Y.; Guo, Y.; Tu, J.; Li, H.; Zhang, T.; Wang, J.; Xue, F.; Zhu, Z.-J. Normalization and integration of large-scale metabolomics data using support vector regression. *Metabolomics* **2016**, *12*, 89.
168. Hermansson, M.; Uphoff, A.; Kakela, R.; Somerharju, P. Automated quantitative analysis of complex lipidomes by liquid chromatography/mass spectrometry. *Anal Chem* **2005**, *77*, 2166-2175.
169. Bijlsma, S.; Bobeldijk, I.; Verheij, E.R.; Ramaker, R.; Kochhar, S.; Macdonald, I.A.; van Ommen, B.; Smilde, A.K. Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. *Anal Chem* **2006**, *78*, 567-574.
170. Sysi-Aho, M.; Katajamaa, M.; Yetukuri, L.; Oresic, M. Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC bioinformatics* **2007**, *8*, 93.
171. Wu, Y.; Li, L. Sample normalization methods in quantitative metabolomics. *Journal of Chromatography A* **2016**, *1430*, 80-95.
172. van den Berg, R.A.; Hoefsloot, H.C.; Westerhuis, J.A.; Smilde, A.K.; van der Werf, M.J. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC genomics* **2006**, *7*, 142.
173. Di Guida, R.; Engel, J.; Allwood, J.W.; Weber, R.J.M.; Jones, M.R.; Sommer, U.; Viant, M.R.; Dunn, W.B. Non-targeted uhplc-ms metabolomic data processing methods: A comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics* **2016**, *12*, 93.
174. Vinaixa, M.; Samino, S.; Saez, I.; Duran, J.; Guinovart, J.J.; Yanes, O. A guideline to univariate statistical analysis for lc/ms-based untargeted metabolomics-derived data. *Metabolites* **2012**, *2*, 775-795.
175. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* **1995**, 289-300.
176. Benjamini, Y.; Cohen, R. Weighted false discovery rate controlling procedures for clinical trials. *Biostatistics* **2016**, kxw030.
177. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemometr Intell Lab* **1987**, *2*, 37-52.
178. PINTO, R.J.D.S.C. Développement de nouvelles méthodes chimiométriques d'analyse application à la caractérisation spectroscopique de la qualité des aliments. l'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech), Paris. France, 2009.
179. Eriksson, L.; Byrne, T.; Johansson, E.; Trygg, J.; Vikström, C. *Multi- and megavariate data analysis basic principles and applications*. MKS Umetrics: Sweden, 2013.
180. Wold, S.; Sjostrom, M.; Eriksson, L. Pls-regression: A basic tool of chemometrics. *Chemometr Intell Lab* **2001**, *58*, 109-130.
181. Höskuldsson, A. Pls regression methods. *Journal of Chemometrics* **1988**, *2*, 211-228.
182. Eriksson I, J.E., Kettaneh-Wold N, Wold S. *Multi- and megavariate data analysis. Principles and applications*. Umetrics Acedemy: Umea, Sweden, 2001; Vol. 1.
183. Barker, M.; Rayens, W. Partial least squares for discrimination. *Journal of Chemometrics* **2003**, *17*, 166-173.
184. Trygg, J.; Wold, S. Orthogonal projections to latent structures (o-pls). *Journal of Chemometrics* **2002**, *16*, 119-128.
185. Tapp, H.S.; Kemsley, E.K. Notes on the practical utility of opls. *Trac-Trend Anal Chem* **2009**, *28*, 1322-1327.
186. Bylesjo, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J.K.; Holmes, E.; Trygg, J. Opls discriminant analysis: Combining the strengths of pls-da and simca classification. *Journal of Chemometrics* **2006**, *20*, 341-351.
187. Wiklund, S.; Johansson, E.; Sjostrom, L.; Mellerowicz, E.J.; Edlund, U.; Shockcor, J.P.; Gottfries, J.; Moritz, T.; Trygg, J. Visualization of gc/tof-ms-based metabolomics data for identification of biochemically interesting compounds using opls class models. *Anal Chem* **2008**, *80*, 115-122.
188. Vandeginste, B.G.M.M., D.L. Buydens, L.M.C.; De Jong, S. Lewi, P.J. Smeyers-Verbeke, J. . *Handbook of chemometrics and qualimetrics – part b.* Elsevier: Amsterdam, 1998.
189. Martens, H.N., T. *Multivariate calibration*. John Wiley & Sons: New York, 1989
190. Martens, H.; Martens, M. Modified jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (plsr). *Food Qual Prefer* **2000**, *11*, 5-16.
191. Wehrens, R.; Putter, H.; Buydens, L.M.C. The bootstrap: A tutorial. *Chemometr Intell Lab* **2000**, *54*, 35-52.
192. Dijksterhuis, G.B.; Heiser, W.J. The role of permutation tests in exploratory multivariate data analysis. *Food Qual Prefer* **1995**, *6*, 263-270.
193. Eriksson, L.; Trygg, J.; Wold, S. Cv-anova for significance testing of pls and opls (r) models. *Journal of Chemometrics* **2008**, *22*, 594-600.
194. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics (Oxford, England)* **2007**, *23*, 2507-2517.
195. Yi, L.; Dong, N.; Yun, Y.; Deng, B.; Ren, D.; Liu, S.; Liang, Y. Chemometric methods in data processing of mass spectrometry-based metabolomics: A review. *Analytica Chimica Acta* **2016**, *914*, 17-34.
196. Yun, Y.-H.; Deng, B.-C.; Cao, D.-S.; Wang, W.-T.; Liang, Y.-Z. Variable importance analysis based on rank aggregation with applications in metabolomics for biomarker discovery. *Analytica Chimica Acta* **2016**, *911*, 27-34.
197. Chan, E.C.; Pasikanti, K.K.; Nicholson, J.K. Global urinary metabolic profiling procedures using gas chromatography-mass spectrometry. *Nature protocols* **2011**, *6*, 1483-1499.

7. Etat de l'art des stratégies métabolomiques : de la chimie à la biologie (Articles II et III)

Les **articles II et III** présentent l'état de l'art des stratégies analytiques pour l'extraction de l'information métabolique d'un système biologique et les outils chimiométriques et bioinformatiques pour son interprétation et sa traduction en supports décisionnels.

Article II: A Tebani, C Afonso, S Bekri. Advances in metabolome information retrieval: turning chemistry into biology. Part I: Analytical chemistry of the metabolome. *J Inherit Metab Dis*. 2017. ***Submitted***

Article III: A Tebani, C Afonso, S Bekri. Advances in metabolome information retrieval: turning chemistry into biology. Part II: Biological information recovery. *J Inherit Metab Dis*. 2017. ***Submitted***

Advances in metabolome information retrieval: turning chemistry into biology

Part I: Analytical Chemistry of the Metabolome

Abdellah Tebani [1,2,3], Carlos Afonso [3], Soumeya Bekri [1,2,*]

[1] Department of Metabolic Biochemistry, Rouen University Hospital, Rouen 76000, France;
[2] Normandie Univ, UNIROUEN, CHU Rouen, IRIB, INSERM U1245, Rouen 76000, France
[3] Normandie Univ, UNIROUEN, INSA Rouen, CNRS, COBRA, Rouen 76000, France

Abstract

Metabolites are small molecules produced by enzymatic reactions in a given organism. Metabolomics or metabolic phenotyping is a well-established omics aimed at comprehensively assessing metabolites in biological systems. These comprehensive analyses use analytical platforms, mainly nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry, along with associated separation methods to gather qualitative and quantitative data. Metabolomics holistically evaluates biological systems in an unbiased, data-driven approach that may ultimately support generation of hypotheses. The approach inherently allows the molecular characterization of a biological sample with regard to both internal (genetics) and environmental (exosome, microbiome) influences. Metabolomics workflows are based on whether the investigator knows a priori what kind of metabolites to assess. Thus, a targeted metabolomics approach is defined as a quantitative analysis (absolute concentrations are determined) or a semiquantitative analysis (relative intensities are determined) of a set of metabolites that are possibly linked to common chemical classes or a selected metabolic pathway. An untargeted metabolomics approach is a semiquantitative analysis of the largest possible number of metabolites contained in a biological sample. This is the part I of a review intending to give an overview of the state of the art of major metabolic phenotyping technologies. Furthermore, their inherent analytical advantages and limits regarding experimental design, sample handling, standardization and workflow challenges are discussed.

Keywords: omics; metabolomics; metabolome; mass spectrometry; nuclear magnetic resonance; chemometrics.

1. Introduction: a historical perspective

Systems biology is a new scientific paradigm aimed at unveiling the systemic function of biology and bridging the gap between biological information and its context. Systems biology can be defined as a global and systemic analysis of complex system interconnections and their functional interrelationships [1-4]. Two seminal inputs have facilitated the emergence of systems biology: data generation and data modeling. High-throughput omics technologies allowed the recovery of a holistic and comprehensive biological information, but the development of computational capabilities have allowed sophisticated systems modeling and convenient visualization tools [5-7]. Omics strategies aim at a comprehensive assessment of entire classes of biomolecules (genes, proteins, metabolites, etc.) of a biological tissue, cell, fluid, or organism. Conceptually, metabolomics has its roots in the practices of ancient Greek doctors who used the organoleptic characteristics of urine for diagnosis; for example, urine sweetness reveals the high glucose levels in diabetes. Such organoleptic chemical features are, of course, linked to metabolism. Olivier et al. coined the metabolome in 1998 and defined it as the set of metabolites synthesized by an organism [8]. Metabolome refers to all metabolites present in a given biological system, fluid, cell, or tissue [9]. Other terms have been used, including metabolic fingerprinting, metabolic footprinting, metabotyping and metabolic phenotyping, with the latter being increasingly accepted. Metabolites can be defined as organic small molecules produced by enzymatic reactions. Thus, metabolomics is one of the "omic" technologies. It is based on biochemical and molecular characterizations of the metabolome and the changes in metabolites related to genetic, environmental, drug, or dietary variables in addition to other factors [10-13]. Metabolomics has found different applications in many disease studies and in complex diseases, with promising perspectives in screening, diagnosis, prognosis, patient stratification, and treatment follow-up [14,15]. Metabolomics is the study of the complete biochemical profile, and the main analytical platforms are nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) paired with separation methods such a high performance liquid chromatography (HPLC). Metabolomics holistically investigates biological systems using an unbiased, data-driven approach that may ultimately lead to generation of hypotheses. In this review, major metabolic phenotyping technologies and their characteristics will be presented along with that challenges associated with data analysis. A discussion on current trends and the requirements for biomarker discovery will be also presented. Finally, we address the current state of the art with respect to standardization and workflow challenges and the gaps in preclinical and clinical environments that hinder translation of metabolic signatures into clinically useful tools.

2. Analytical strategies and chemical information extraction

A few highly reliable metabolites could be sufficient to a certain extent for diagnostic or monitoring purposes. However, the use of more metabolites for a broader overview is more appropriate for assessing, for example, a biochemical pathway. Thus, metabolomics is obviously an interesting tool to support answering biological questions, especially in biomarker discovery. By definition, a metabolic signature contains a set of disrupted metabolites rather than just a single metabolite, which is plausible because of the relevance of affected metabolic pathways and the network theory underpinning biological systems [14,16]. Thus, analytical technologies need to be reliable and robust for high-throughput routine analyses [17]. Furthermore, metabolites have qualitatively and quantitatively heterogenic characteristics. To our best knowledge, no single methodology can separate, detect, and quantify the whole metabolome. Thus, multiple analytical techniques and sample preparation strategies are necessary to recover most of the metabolome [10]. Metabolomics workflows are based on whether the investigator knows a priori what kind of metabolites to assess. A targeted metabolomics approach is defined as a quantitative analysis (absolute concentrations are determined) or a semiquantitative analysis (relative intensities are determined) of a set of metabolites that might be linked to common chemical classes or a selected metabolic pathway. An untargeted metabolomics approach is primarily based on the qualitative or semiquantitative analysis of the largest possible number of metabolites from diverse chemical and biological classes contained in a biological sample. The metabolomics workflow (Fig. 1) comprises the comparative sequential steps of both targeted and untargeted metabolomics analyses. Typical metabolomics experiments aim to analyze as many metabolites as possible in a biological specimen. Several established analytical platforms can enable the semiquantitative assessment (relative intensities) of metabolites. However, the field of metabolomics is increasingly embracing the absolute quantitation of metabolites. The acquired data are extensive and need to be processed and mined to extract insightful biological interpretations. Hence, multivariate data analyses are routinely used to extract information from large metabolomics data sets [18]. The data can be used to build hypotheses or to explain observations. The identified metabolites associated with an observation provide a holistic overview about the interrogated biological system. The metabolomics workflow generally includes biological problem formulation and experimental design, sample preparation, data acquisition, data preprocessing, data pretreatment, data analysis, network and pathway analysis, and finally biological interpretation (cf. Fig. 1).

2.1. Experimental design

The design of a metabolomics experiment requires consideration of various aspects including sample type, number of samples, replication, data analysis strategy, cost, and time, along with the allocation of human and technical resources. The decisions will lead to answers to different questions. Therefore, a well-defined strategy regarding the tools that will be used for data analysis and interpretation is fundamental, and it should set objective questions and recover appropriate answers. Experimental design and data analysis are tightly related, and the first step in any metabolomics workflow is the clear formulation of the biological problem to be addressed. Depending of the biological problem, the investigator must define the metabolomics approach (targeted vs. untargeted), biological samples (biofluids, tissues, cells, and/or intact organisms), sample size, pooling, experimental conditions (i.e., observational studies, exploratory studies, time series), sampling conditions (frequency of sample collection, quenching to stop enzymatic activity, storage), analytical platforms, and sample preparation protocols. It should be noted that most metabolomics studies are intrinsically comparative; therefore, a group of control samples (samples that did not undergo the investigated condition) and test samples (samples exposed to the investigated condition) are rigorously defined in the experimental design with clear inclusion and exclusion criteria [19]. Indeed, subjects are heterogeneous with respect to demographic and lifestyle factors, and this factor is especially important in defining healthy controls. Sample randomization along with sample analysis order and instrument conditions are means to reduce the correlation of confounders. Moreover, advanced statistical experimental design strategies can be used to handle these issues [20-22].

2.2. Biological Samples

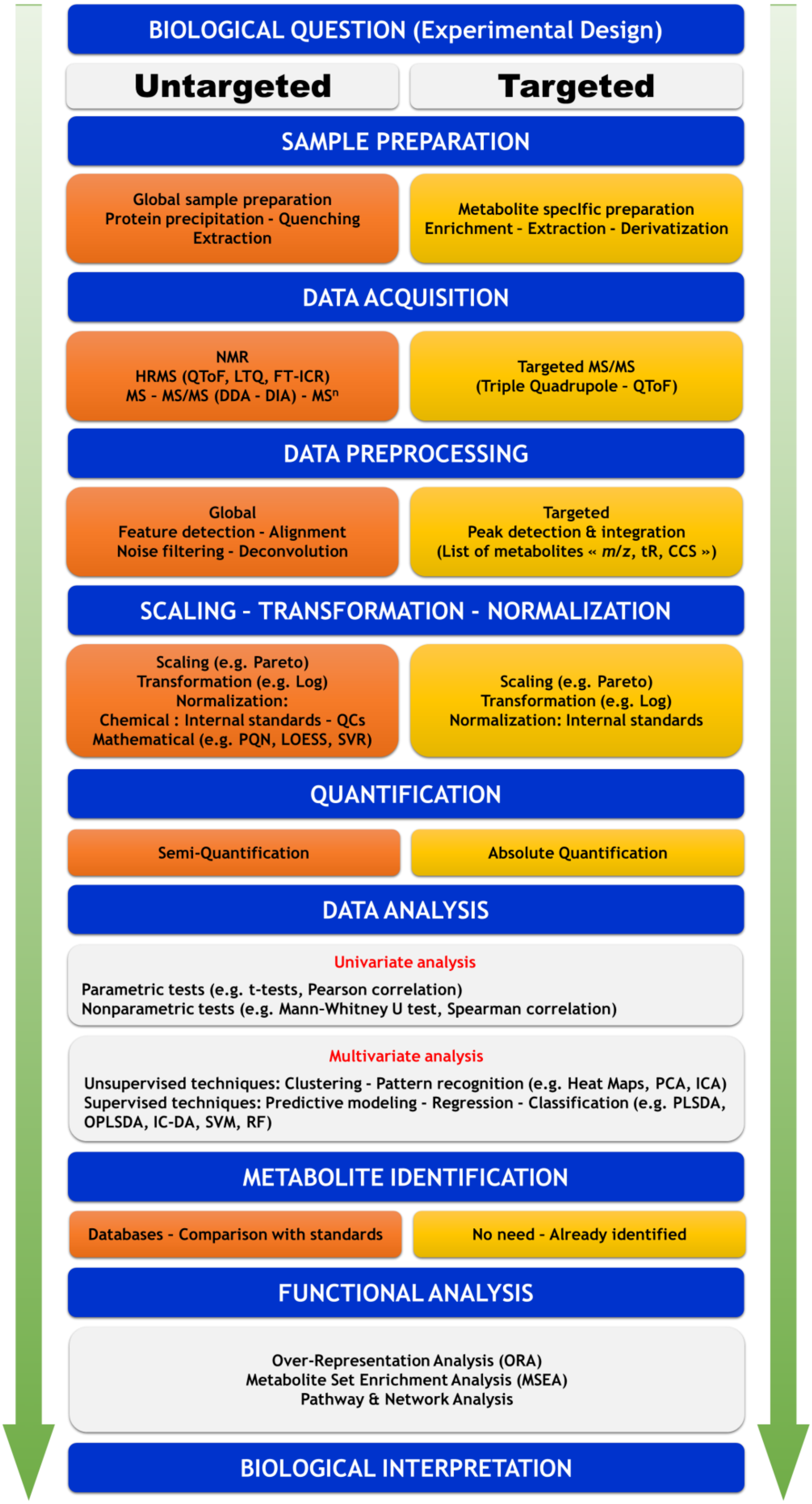
Sample collection (time and type), storage, and handling have important impacts on the retrieved metabolic profile [23-25]. Thus, these factors have to be standardized to avoid spurious biomarker discovery interpretation [26,27]. Therefore, careful consideration of sampling conditions and handling is needed to provide a reliable metabolic snapshot of the sample at the time it is collected. The primary objective should be to ensure qualitatively and quantitatively consistent and representative samples prior to their collection. Samples can be separated into two general classes: (i) metabolically active samples (intracellular metabolome) and (ii) metabolically inactive samples (extracellular metabolome) [28]. The sample type is chosen based on the biological question being investigated. However, in some studies, the preferred sample type cannot be collected, and a surrogate sample type, such as urine or blood (serum or plasma), has to be used. Compared with intact or extracted tissues, urine and serum or plasma are the most commonly studied biofluids in clinical practice because they are easily obtained and prepared. However, other specialized fluids can be used, including cerebrospinal fluid [29,30], saliva [31-33], sweat [34], and even breath [35,36]. Dried blood (and other biofluid) spots have also been investigated [24,37-39] and offer an interesting alternative to conventional liquid samples for generating metabolite profiles. Given their very practical advantages, including low volume, low cost, and handling convenience, dried blood spots are drawing interest as a sampling option for metabolic profiling [37,40-42]. Of note, most metabolomics studies, particularly in clinical metabolomics, include data from a single biofluid, most often blood or urine. However, the biochemical signature in a biofluid denotes complex interactions with different organs, which adds to the interpretative complexity of metabolomics data. This complexity can only be understood by investigating pathophysiological states from a metabolic network perspective, taking into account the local metabolome and its contribution to the systemic metabolome. Different data-driven approaches have been described for handling these issues by using metabolomics data modeling [43,44].

2.3. Sample preparation

For untargeted metabolomics, minimal sample preparation is generally recommended to avoid metabolite loss. Tissues are often homogenized using manual techniques such as a mortar and pestle or ball grinding with silica particles or stainless steel. The homogenization process is often performed with an extraction solvent, which leads to cell lysis and extraction of the metabolites. Monophasic (water/methanol, water/acetonitrile) or biphasic (water and methanol often along with a nonpolar solvent such as chloroform, dichloromethane, or methyl tert-butyl ether) solvent can be used in extraction systems depending on the planned analysis [45]. The choice of solvent systems depends on whether polar or nonpolar molecules are to be investigated. For cell metabolomics, a quenching step is required before extraction to minimize metabolite modifications. Extraction from cells must be performed as quickly as possible to avoid enzymatic reactions and to improve reproducibility [46]. The preparation and analysis may be costly in terms of time, which may be quite limiting in the clinical environment. The ability to collect data without sample preparation combined with real-time data analysis would allow rapid clinical decision-making, which would place metabolomics at a higher clinically actionable level (i.e., surgery, pathology). For example, the intelligent knife (iKnife) represents a significant advance in *in vivo* sampling and real-time metabolomics technology. This process vaporizes tissue and the resultant smoke is transferred into a mass spectrometer to provide real-time clinical decision-making in the operating room [47].

2.4. Analytical platforms

The analysis of the metabolome raises different challenges compared with other omics analyses, which are based on profiling large molecules built with a simple and limited set of subunits, such as nucleotides for genomics and transcriptomics and amino acids for proteomics. For identification and functional analysis of DNA, RNAs, and proteins, subunit order is what matters; it represents the observed biological complexity. Hence, analytical strategies based on sequencing essentially rely on the incremental detection of the subunits [48]. However, a sequencing concept cannot be applied to metabolites in complex biofluids because the analytical challenge does not lie in cracking any order code; there is no order. The metabolome requires a more complex analytical strategy that allows individual and selective differentiation of metabolites across a wide qualitative and quantitative chemical space. The physicochemical heterogeneity of metabolites adds another layer of complexity to metabolomics studies. In an early scientific paper in the field, Pauling and colleagues described a method using gas chromatographic separation with flame ionization detection to analyze the breath [49]. Impressive analytical developments have occurred since then. The metabolic profiling technologies that are mainly used now include are NMR spectroscopy and MS, sometimes in combination with a gas phase or liquid phase separation method [18]. These technologies retrieve global, unbiased, and comprehensive chemical information from complex mixtures. For information translation, the resultant high-dimensional spectral data are typically analyzed using chemometric techniques to identify informative metabolic combinations that can be used for either global biomarker discovery or sample classification.



BIOLOGICAL QUESTION (Experimental Design)

Untargeted

Targeted

SAMPLE PREPARATION

Global sample preparation
Protein precipitation - Quenching
Extraction

Metabolite specIfic preparation
Enrichment - Extraction - Derivatization

DATA ACQUISITION

NMR
HRMS (QToF, LTQ, FT-ICR)
MS - MS/MS (DDA - DIA) - MS^n

Targeted MS/MS
(Triple Quadrupole - QToF)

DATA PREPROCESSING

Global
Feature detection - Alignment
Noise filtering - Deconvolution

Targeted
Peak detection & integration
(List of metabolites « *m/z*, tR, CCS »)

SCALING - TRANSFORMATION - NORMALIZATION

Scaling (e.g. Pareto)
Transformation (e.g. Log)
Normalization:
Chemical : Internal standards - QCs
Mathematical (e.g. PQN, LOESS, SVR)

Scaling (e.g. Pareto)
Transformation (e.g. Log)
Normalization: Internal standards

QUANTIFICATION

Semi-Quantification

Absolute Quantification

DATA ANALYSIS

Univariate analysis

Parametric tests (e.g. t-tests, Pearson correlation)
Nonparametric tests (e.g. Mann-Whitney U test, Spearman correlation)

Multivariate analysis

Unsupervised techniques: Clustering - Pattern recognition (e.g. Heat Maps, PCA, ICA)
Supervised techniques: Predictive modeling - Regression - Classification (e.g. PLSDA, OPLSDA, IC-DA, SVM, RF)

METABOLITE IDENTIFICATION

Databases - Comparison with standards

No need - Already identified

FUNCTIONAL ANALYSIS

Over-Representation Analysis (ORA)
Metabolite Set Enrichment Analysis (MSEA)
Pathway & Network Analysis

BIOLOGICAL INTERPRETATION

Figure 1. General metabolomics workflow.

2.4.1.Nuclear magnetic resonance spectroscopy

NMR spectroscopy is rapid and nondestructive, and it has the advantage of being highly reproducible and robust. It is based on the absorption and re-emission of energy by the atom nuclei due to variations in an external magnetic field. Different types of metabolomics data can be generated depending on the targeted atom nuclei. However, in the analysis of biological samples, hydrogen-1 is the most commonly used type of nuclei (^{1}H-NMR) because of its naturally high abundance in these samples. Other nuclei, such as carbon-13 (^{13}C-NMR) and phosphorus-31 (^{31}P NMR), can also be used to provide additional information on specific metabolite types. ^{31}P NMR is useful for studies of cellular energy states *in vivo* and *ex vivo*, but a limitation is the overlapping of ^{31}P signals from phosphorylated compounds. NMR spectroscopy is a powerful technology that offers atom-centered information that is crucial for elucidating molecular structures [26]. The resulting spectral data allow quantification and identification of the metabolites. Peak areas are used for quantification, whereas the spectral patterns permit metabolite identification. The spectral data generated by NMR techniques can be divided into two NMR strategies regarding the frequency axis used. Frequency axes are referenced by the chemical shift expressed in parts per million (ppm). The chemical shift is calculated as the difference between the metabolite resonance frequency and that of a reference substance [50]. One-dimensional NMR (1D-NMR) spectra are based on a single frequency axis, where the peaks of each molecule occur within the resonant frequencies of that axis. This method is the most used in high-throughput metabolomics. Two-dimensional NMR (2D-NMR), which is based on two frequency axes, can be used to complement 1D-NMR. Signals are either binned and then analyzed or fitted to patterns of signals corresponding to the metabolites expected to be present in the mixture. ^{13}C NMR signals are better resolved, but they exhibit low sensitivity due to a low natural abundance of ^{13}C [51]. In 2D-NMR, the second dimension allows separation of overlapping spectral peaks and therefore provides additional and orthogonal chemical information on the investigated metabolites within the analyzed matrix [52]. 2D-NMR methods include ^{1}H-^{1}H COSY (correlated spectroscopy), ^{1}H–^{1}H TOCSY (total correlation spectroscopy), and ^{1}H–^{13}C HSQC (heteronuclear single-quantum correlation) [26]. Of note, nuclei with low natural abundance, including ^{2}H (deuteron), ^{13}C, and ^{15}N, may serve as excellent metabolic tracers [53]. Despite its relatively low sensitivity, NMR spectroscopy offers many advantages because it allows rigorous quantification of highly abundant metabolites present in biological fluids, cell extracts, and tissues with minimal or no sample preparation [54]. NMR spectroscopy is useful for molecules that are difficult to ionize or require derivatization for MS analysis. NMR spectroscopy also allows the identification of isomeric molecules, and it is the gold standard for determining structures of unknown compounds. Using stable isotope labels, NMR spectroscopy can be used for dynamic assessment of compartmentalization of metabolic pathways, such as metabolite transformations and drug metabolism. Finally, intact tissue NMR imaging and spectroscopy are very appealing for in vivo metabolic investigations [55]. The main drawback of NMR methods is its low sensitivity and resolution compared with MS-based methods [56].

2.4.2.Mass spectrometry

Mass spectrometry is an analytical technique that retrieves chemical data from the gas-phase ions produced from a sample. The ions generate different peak patterns that define the fingerprint of the original molecule in the form of a mass-to-charge ratio (*m*/*z*) and a relative intensity of the measured chemical features (e.g., metabolites). The sample is introduced into the mass spectrometer via the sample inlet, an ion source generates gas-phase ions, a mass analyzer separates the ions according to their *m*/*z*, and a detector generates an electric current from the incident ions that is proportional to their abundances [57]. A sample can be directly injected into a mass spectrometer such as in direct infusion mass spectrometry (DIMS) [58]. The major drawback is the ion suppression effect, which leads to metabolite information loss and prohibits separation of isomers. Mass analyzers can be used alone or in combination with the same type of mass analyzer or with different mass analyzers (hybrid instruments). Such combinations are the foundation for the analytical mode of tandem mass spectrometry (MS/MS). In MS/MS, the ions that arrive at the first mass analyzer (precursor ions) are selected, then fragmented in a collision cell. The fragmented ions are separated according to their *m*/*z* in a second mass analyzer and then detected. Different operation modes are possible, including data dependent analysis (DDA) and data independent analysis (DIA). In DDA, a fixed number of precursor ions whose *m*/*z* values were recorded in a survey scan are selected using predetermined rules and are subjected to a second stage of mass selection in an MS/MS analysis [59]. Modes include single reaction monitoring (SRM) or multiple reaction monitoring (MRM), which is the application of SRM with parallel detection of all transitions in a single analysis. In DIA, all precursor ions within a defined *m*/*z* window undergo fragmentation. The analysis is repeated as the mass spectrometer progresses through the full selected *m*/*z* range [60]. This process yields accurate metabolite quantification without being limited to profiling predefined metabolites of interest [61]. For some mass analyzers, such as quadrupole ion traps, several steps of MS/MS can be performed. For example, the fragmented ions can be further fragmented and detected. The experiment is called multiple-stage mass spectrometry (MS^n, n refers to the number of MS steps). MS/MS and MS^n improve structural identification, combining information from both molecular and fragmented ions generated from precursor ions. The main performance characteristics of a mass analyzers are (1) mass accuracy, or mass resolving power, which is related to the ability of an MS analyzer to generate distinct signals for two ions with a small *m*/*z* difference; (2) mass range, which is the range of *m*/*z* over which a mass spectrometer can detect ions to record a mass spectrum; (3) sensitivity; (4) scan speed; and (5) duty cycle time, which is the fraction of ions that effectively reach the detector in the mass spectrometer. The mass analyzer choice is mainly based on the type of metabolomics approach to be carried out, targeted or untargeted. Single quadrupole (Q), triple quadrupole (QqQ), quadrupole ion trap (QIT), and Orbitrap (OT) are suitable for targeted metabolomics because of their sensitivity and duty cycle characteristics. In comparison, dynamic range, mass accuracy, and resolution power are the main characteristics of a mass analyzer to be used in untargeted metabolomics studies. Time of flight (TOF), quadrupole time of flight (QTOF), Fourier transform ion cyclotron resonance (FTICR), and OT are the most used mass analyzers for this purpose. The principle underlying TOF and QToF involves the time required for ions to travel a flight tube. Ions are accelerated in an electric field, reaching a linear velocity that depends on their *m*/*z* ratio. The velocity can reach 10,000 per second scan speed, with a mass error of 5 ppm. A QTOF mass analyzer is a hybrid instrument that can generate high-resolution MS/MS spectra [62]. FTICR is an ultra-high-resolution (10^5–10^7 depending on the detection time and magnetic field) mass analyzer that uses cyclotron frequency in a fixed magnetic field to measure *m*/*z* ions at the cost of relatively slow acquisition rates (typically 1 Hz). In the same way, the OT is also a FTMS instrument, which is based on harmonic ion oscillations in an electrostatic field. Ions are trapped around a central electrode, and ion oscillation frequencies are used to measure the *m*/*z* values. The OT provides high mass resolution (>100,000 FWHM), high mass accuracy (2–5 ppm), and an acceptable dynamic range. However, the scan speed is inversely related to mass resolution. Recently, the high-field Orbitrap has provided a resolution above 1,000,000 at *m*/*z* 300–400 with 3 s detection time, using an absorption mode [63]. A wide range of instrumental and technical variants are currently available for MS spectrometry. These variants are mainly characterized by different ionization and mass selection methods [64]. Because of the matrix effect limit and potential isomers, MS is generally preceded by a separation step in metabolomics. This step reduces the complexity of a biological sample and allows sequential MS analysis of the different molecules. Different separation methods coupled to MS have been described, such as liquid chromatography (LC-MS) [65,66], gas chromatography (GC-MS) [67], and capillary electrophoresis (CE-MS) [68]. Thus, metabolites with different chemical properties will spend different amounts of time (retention time, t_R) in the separation dimension. This different separation methods enhance the sensitivity and the dynamic range of MS and provide complementary and orthogonal molecular information.

– **Liquid chromatography**

LC-MS is widely used in metabolomics because of its analytical versatility, covering separation performance of different classes of molecules, from very polar to very lipophilic compounds. This high versatility is achieved through the variety of chromatographic columns along with stationary phases [69]. The LC separation basics depend on physico-chemical properties, such as hydrophobicity, molecular size, and polarity. The separation of compounds occurs in a chromatographic column composed of a stationary phase with polar or lipophilic properties. When polar stationary phase columns are used, the method is referred to as normal-phase liquid chromatography (NPLC); when nonpolar stationary phase columns are used, the method is called reversed-phase liquid chromatography (RPLC). The choice of LC columns depends on the polarity of the metabolites and the analytical scope. To analyze nonpolar and/or weakly polar metabolites, nonpolar C18 and C8 columns are mostly used for untargeted metabolomics [62]. However, for hydrophilic, ionic, and polar compounds, hydrophilic interaction liquid chromatography (HILIC) is recommended. HILIC is similar to NPLC, but it differs because of the mobile phase, which is composed of a polar and/or aprotic organic solvent miscible in water that is easier to use with electrospray-mass spectrometry [70]. Multiple-column strategies could be used for more extensive metabolome coverage [71]. Recently, RPLC and HILIC columns with a smaller internal diameter (e.g., 1 mm) and shorter length are drawing interest in metabolomics. Thus, instruments that can operate at very high pressure—ultra-performance liquid chromatography (UHPLC)—coupled to mass spectrometry have been introduced to improve metabolite coverage and detection. UHPLC methods allow increased resolution, better sensitivity, and lower ion suppression. As a result, better metabolome coverage is obtained in comparison with conventional HPLC. Moreover, lower solvent consumption is observed because of the low flow rate (150–250 μL/min) [72].

– **Gas chromatography**

GC-MS is often used for analysis of volatile compounds and molecules with low vapor pressure, such as lipids, long-chain alcohols, amides, alkaloids, sugar alcohols, and organic acids However, using derivative techniques widens the coverage of GC-MS. GC-MS has been accepted as a robust metabolomics platform because of its selective separation, reproducibility, and robustness. The greatest advantage of GC-MS is that its ionization mode is highly reproducible and standardized (based on electron ionization at 70 eV) across GC-MS systems worldwide and across different vendors [73], which has allowed comprehensive GC-MS mass spectral libraries such as NIST and FiehnLab to be established [74]. As a result, GC-MS has been a set and reliable platform for MS-based metabolomics. The main limitation of GC-MS is the necessary derivatization step for some metabolite classes. In metabolomics, derivatization usually uses oximation and a silylation/chloroformate reagent. This step is time consuming, hampers the throughput, and can introduce error by adding analytical variability [75]. Moreover, GC-MS metabolome coverage is limited by the stationary phase stability as well as the thermal stability of metabolites and their derivatives [76].

– **Capillary electrophoresis**

Capillary electrophoresis (CE) offers an orthogonal separation mechanism. CE-specific characteristics, such as high efficiency and resolution, high throughput, and, importantly, the ability to assess the most polar compounds without derivatization, have made CE an attractive method for metabolomics [77]. CE-MS was the last pre-ionization separation technique to be paired with MS in metabolomics. Capillary zone electrophoresis (CZE) is the simplest and most commonly used CE mode because of its principle of separation and its broad application to the analysis of diverse samples, spanning small to large biomolecules. In CZE, analytes are separated according to their intrinsic differential electrophoretic mobility in a capillary filled with separation buffer under the influence of an electric field. The mobilities depend on the ion *m*/*z* and the viscosity of the medium [77]. The main drawback of CZE is that neutral molecules are not separated. To overcome this disadvantage, other CE modes have been developed, such as micellar electrokinetic chromatography, capillary isotachophoresis, capillary isoelectric focusing based on pH gradient, capillary electrochromatography, capillary gel electrophoresis, and affinity capillary electrophoresis. Because of its simplicity, CZE is the preferred CE mode in metabolomics. Despite the recent technical advances of CE-MS, its use in metabolomics is still limited compared with NMR spectroscopy and chromatography-based methods. For more details about CE-MS applications in metabolomics, the reader may refer to a recent review [78].

- **Ion mobility and multidimensional strategies**

Another gas phase separation, ion mobility spectrometry (IMS), [79], is drawing interest in metabolomics [80-86]. In general, the multidimensional coupling of different separation techniques requires that the resolution obtained from each anterior separation must be largely preserved as the analytes pass to the following dimensions. This preservation is particularly difficult when all analytes travel along the same path during the analysis, as is the case for tempo-dispersive techniques. Thus, the solution is to incrementally increase the sampling frequency of each subsequent time dimension so that multiple measurements are obtained within a fixed time interval. In this way, the arrival time in each anterior dimension can be reassembled based on the integrated signal of subsequent dimensions. This strategy is commonly used when coupling condensed phase separations such as GC, LC, or CE to MS. IMS is an appealing post-ionization separation method that is based on molecular size, shape, and charge. It is typically performed on a millisecond timescale, which can be perfectly nested between chromatography (seconds) and high-resolution MS detection (microseconds) timescales. Hence, coupling IMS with high-resolution mass spectrometry and chromatography (LC-IMS-MS) provides additional analytic selectivity without significantly compromising the speed of MS-based measurements. As a result, the MS dimension affords accurate mass information, while the IMS dimension provides molecular, structural, and conformational information through the determination of the ion collision cross-section (CCS), which is a valuable and predictable chemical descriptor. Indeed, ion mobility spectrometry adds a separation dimension to the hybrid MS instruments, allowing a higher analytical coverage of complex biological mixtures [82,87-90]. One important feature of IMS is its ability to separate isomers [91]; the predictability of the CCS and peak width for one isomer mainly depend on ion diffusion [92-94]. Furthermore, exploring a multivectorial space containing retention time, accurate mass, and CCS obtained by the combination of multiple separation methods with MS allows valuable measurement integration, which enhances molecular identification and consequently biomarker discovery [95,96].

- **Towards real-time MS-based metabolomics**

Recent introduction of ambient ionization sources has significantly increased the high throughput of global metabolic profiling analysis. These techniques permit direct sampling of complex matrices under ambient conditions, and they include atmospheric solids analysis probe [97], desorption electrospray ionization [98-100], and rapid evaporative ionization MS methods [47,101]. These techniques can provide real-time, interpretable MS data on biofluids and tissues, *in vivo* and *ex vivo*, and they are reshaping high-throughput real-time metabolome analysis in different areas [102,103]. For example, in many surgeries, visually distinguishing between healthy and diseased tissues is often difficult. It requires time-consuming biopsies and immuno-staining procedures to be performed by experienced trained histopathologists during surgery. By eliminating this need for external tissue histotyping, techniques such as the iKnife could open the way to true real-time precision surgery. For more details about the use of ambient MS in clinical diagnosis, refer to a recent and detailed review by Ifa and Eberlin [104]. Table 1 presents a comparison between different analytical strategies used in metabolomics.

Table 1. Comparison of main analytical technologies in metabolomics.

Platform	Technique	Identification dimensions	Principle	Advantages	Limits
Nuclear magnetic resonance	1 Dimension 2 Dimensions	Chemical shift Chemical shift × chemical shift	Uses interaction of spin active nuclei (^{1}H, ^{13}C, ^{31}P) with electromagnetic fields, yielding structural, chemical, and molecular environment information	Nondestructive Highly reproducible Exact quantification possible Minimal sample preparation Molecular dynamic and compartmental information using diffusional methods Relatively high throughput Availability of databases for identification	High instrumentation cost Overlap of metabolites Low sensitivity
Mass spectrometry	Direct injection (DI-MS)	*m/z*	Uses a nanospray source directly coupled to MS detector. It does not require chromatographic separation.	Very high throughput High sensitivity No cross-sample contamination No column carryover Low-cost analysis Automated analysis Low sample volume requirement Allows MS imaging	Samples not recoverable (destructive) No retention time information, which gives limited specificity Inability to separate isomers Subjected to significant ion suppression phenomenon High ionization discrimination (ESI)
	Liquid chromatography (LC-MS)	Time × *m/z*	Uses chromatographic columns that enables liquid phase chromatographic separation of molecules followed by MS detection (suitable for polar to hydrophobic compounds)	Minimal sample preparation (protein precipitation or dilution of biological sample) High-throughput capability UHPLC can be coupled to any type of MS Flexibility in column chemistry widening the range of detectable compounds High sensitivity	Samples not recoverable (destructive) Very polar molecules need specific chromatographic conditions Retention times are highly dependent on exact chromatographic conditions Batch analysis Lack of large metabolite databases High ionization discrimination (ESI)
	Gas chromatography (GC-MS)	Time × *m/z*	Uses chromatographic columns that enables gas phase chromatographic separation of molecules followed by MS detection (suited for apolar and volatile compounds)	Structure information obtained through in-source fragmentation Availability of universal databases for identification High sensitivity Reproducible	Samples not recoverable (destructive) Requires more extensive sample preparation Only volatile compounds are detected Polar compounds need derivatization Low ionization discrimination
	Capillary electrophoresis (CE-MS)	Time × *m/z*	Uses electrokinetic separation of polar molecules paired with a mass spectrometry detector	Excellent for polar analysis in aqueous samples Measures inorganic and organic anions Low running costs	Samples not recoverable (destructive) Relatively low throughput profiling
	Ion mobility spectrometry (IMS-MS)	Time × *m/z* (CCS × *m/z*)	Uses a uniform or periodic electric field and a buffer gas to separate ions based on charge, size, and shape paired with mass spectrometry	Very robust and reproducible (ability to determine collision cross-section, which is a robust chemical descriptor) High peak capacity High selectivity Separation of isomeric and isobaric compounds Very high throughput	Samples not recoverable (destructive) CCS and mass are highly correlated parameters, which limits the orthogonality of the method

3. Conclusion

Substantial advances have occurred in analytical chemistry for metabolomics strategies for better chemical data extraction from biological samples. These advances have had a substantial impact on metabolomics workflows by simplifying analytical protocols and introducing more robust systems. However, to go a step further to translate metabolomics into an actionable exploratory and ultimately a diagnostic tool, issues that need to be addressed include streamlining and automating sample preparation, improving analytical throughput by using faster separation (or no separation, if using DIMS), and introducing orthogonal analytical dimensions such as IMS-MS in metabolomics. NMR and chromatography-based platforms are still the well-established technologies for metabolomics studies. LC-MS and GC-MS are the most adopted analytical platforms in clinical metabolomics. Still, for a more comprehensive metabolome coverage, implementation of multiplatform approaches is necessary. To reach next-generation metabolomics, further advances are urgently needed in analytical strategies for reliable identification and absolute quantification. Finally, standardization regarding sample handling and analytical procedures is a big issue for larger clinical studies and wide adoption of metabolomics, particularly, in clinical environments.

Acknowledgments: This work was supported by the Normandy University, the Institut National de la Santé et de la Recherche Médicale (INSERM), the Conseil Régional de Normandie, Labex SynOrg (ANR-11-LABX-0029), and the European Regional Development Fund (ERDF 31708).

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

REFERENCES

1. Weston, A.D.; Hood, L. Systems biology, proteomics, and the future of health care: Toward predictive, preventative, and personalized medicine. *Journal of proteome research* **2004**, *3*, 179-196.
2. Ehrenberg, M.; Elf, J.; Aurell, E.; Sandberg, R.; Tegner, J. Systems biology is taking off. *Genome research* **2003**, *13*, 2377-2380.
3. Kitano, H. Looking beyond the details: A rise in system-oriented approaches in genetics and molecular biology. *Current genetics* **2002**, *41*, 1-10.
4. Kitano, H. Systems biology: A brief overview. *Science (New York, N.Y.)* **2002**, *295*, 1662-1664.
5. Tenenbaum, J.D.; Avillach, P.; Benham-Hutchins, M.; Breitenstein, M.K.; Crowgey, E.L.; Hoffman, M.A.; Jiang, X.; Madhavan, S.; Mattison, J.E.; Nagarajan, R., *et al.* An informatics research agenda to support precision medicine: Seven key areas. *J Am Med Inform Assoc* **2016**.
6. McMurry, J.; Kohler, S.; Balhoff, J.; Borromeo, C.; Brush, M.; Carbon, S.; Conlin, T.; Dunn, N.; Engelstad, M.; Foster, E., *et al.* Navigating the phenotype frontier: The monarch initiative. *bioRxiv* **2016**.
7. Ritchie, M.D.; Holzinger, E.R.; Li, R.; Pendergrass, S.A.; Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet* **2015**, *16*, 85-97.
8. Oliver, S.G.; Winson, M.K.; Kell, D.B.; Baganz, F. Systematic functional analysis of the yeast genome. *Trends in biotechnology* **1998**, *16*, 373-378.
9. 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; the fate of foreign compounds in biological systems* **1999**, *29*, 1181-1189.
10. Dunn, W.B.; Broadhurst, D.I.; Atherton, H.J.; Goodacre, R.; Griffin, J.L. Systems level studies of mammalian metabolomes: The roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chemical Society reviews* **2011**, *40*, 387-426.

11. Fiehn, O. Metabolomics--the link between genotypes and phenotypes. *Plant molecular biology* **2002**, *48*, 155-171.
12. Holmes, E.; Wilson, I.D.; Nicholson, J.K. Metabolic phenotyping in health and disease. *Cell* **2008**, *134*, 714-717.
13. Benton, H.P.; Want, E.; Keun, H.C.; Amberg, A.; Plumb, R.S.; Goldfain-Blanc, F.; Walther, B.; Reily, M.D.; Lindon, J.C.; Holmes, E., *et al.* Intra- and interlaboratory reproducibility of ultra performance liquid chromatography-time-of-flight mass spectrometry for urinary metabolic profiling. *Analytical chemistry* **2012**, *84*, 2424-2432.
14. Bekri, S. The role of metabolomics in precision medicine. *Expert Review of Precision Medicine and Drug Development* **2016**.
15. Tebani, A.; Abily-Donval, L.; Afonso, C.; Marret, S.; Bekri, S. Clinical metabolomics: The new metabolic window for inborn errors of metabolism investigations in the post-genomic era. *Int J Mol Sci* **2016**, *17*.
16. Ravasz, E.; Somera, A.L.; Mongru, D.A.; Oltvai, Z.N.; Barabasi, A.L. Hierarchical organization of modularity in metabolic networks. *Science (New York, N.Y.)* **2002**, *297*, 1551.
17. Zampieri, M.; Sekar, K.; Zamboni, N.; Sauer, U. Frontiers of high-throughput metabolomics. *Current Opinion in Chemical Biology* **2017**, *36*, 15-23.
18. Alonso, A.; Marsal, S.; Julia, A. Analytical methods in untargeted metabolomics: State of the art in 2015. *Frontiers in bioengineering and biotechnology* **2015**, *3*, 23.
19. Broadhurst, D.; Kell, D. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* **2006**, *2*, 171-196.
20. Jonsson, P.; Wuolikainen, A.; Thysell, E.; Chorell, E.; Stattin, P.; Wikström, P.; Antti, H. Constrained randomization and multivariate effect projections improve information extraction and biomarker pattern discovery in metabolomics studies involving dependent samples. *Metabolomics* **2015**, *11*, 1667-1678.
21. Boccard, J.; Rudaz, S. Exploring omics data from designed experiments using analysis of variance multiblock orthogonal partial least squares. *Analytica chimica acta* **2016**, *920*, 18-28.
22. Dunn, W.B.; Wilson, I.D.; Nicholls, A.W.; Broadhurst, D. The importance of experimental design and qc samples in large-scale and ms-driven untargeted metabolomic studies of humans. *Bioanalysis* **2012**, *4*, 2249-2264.
23. Yin, P.; Lehmann, R.; Xu, G. Effects of pre-analytical processes on blood samples used in metabolomics studies. *Analytical and Bioanalytical Chemistry* **2015**, *407*, 4879-4892.
24. Prentice, P.; Turner, C.; Wong, M.C.Y.; Dalton, R.N. Stability of metabolites in dried blood spots stored at different temperatures over a 2-year period. *Bioanalysis* **2013**, *5*, 1507-1514.
25. Burton, G.J.; Sebire, N.J.; Myatt, L.; Tannetta, D.; Wang, Y.L.; Sadovsky, Y.; Staff, A.C.; Redman, C.W. Optimising sample collection for placental research. *Placenta* **2014**, *35*, 9-22.
26. Emwas, A.-H.; Salek, R.; Griffin, J.; Merzaban, J. Nmr-based metabolomics in human disease diagnosis: Applications, limitations, and recommendations. *Metabolomics* **2013**, *9*, 1048-1072.
27. Dunn, W.B.; Broadhurst, D.; Begley, P.; Zelena, E.; Francis-McIntyre, S.; Anderson, N.; Brown, M.; Knowles, J.D.; Halsall, A.; Haselden, J.N., *et al.* Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature protocols* **2011**, *6*, 1060-1083.
28. Chetwynd, A.J.; Dunn, W.B.; Rodriguez-Blanco, G. Collection and preparation of clinical samples for metabolomics. In *Metabolomics: From fundamentals to clinical applications*, Sussulini, A., Ed. Springer International Publishing: Cham, 2017; pp 19-44.
29. Graham, S.F.; Chevallier, O.P.; Roberts, D.; Hölscher, C.; Elliott, C.T.; Green, B.D. Investigation of the human brain metabolome to identify potential markers for early diagnosis and therapeutic targets of alzheimer’s disease. *Analytical chemistry* **2013**, *85*, 1803-1811.
30. Wuolikainen, A.; Hedenstrom, M.; Moritz, T.; Marklund, S.L.; Antti, H.; Andersen, P.M. Optimization of procedures for collecting and storing of csf for studying the metabolome in als. *Amyotrophic lateral sclerosis : official publication of the World Federation of Neurology Research Group on Motor Neuron Diseases* **2009**, *10*, 229-236.
31. Dame, Z.; Aziat, F.; Mandal, R.; Krishnamurthy, R.; Bouatra, S.; Borzouie, S.; Guo, A.; Sajed, T.; Deng, L.; Lin, H., *et al.* The human saliva metabolome. *Metabolomics* **2015**, 1-20.
32. Kawasaki, G.; Ichikawa, Y.; Yoshitomi, I.; Umeda, M. Metabolomics of salivary biomarkers in yusho patients. *Fukuoka igaku zasshi = Hukuoka acta medica* **2015**, *106*, 144-148.
33. Mikkonen, J.J.; Singh, S.P.; Herrala, M.; Lappalainen, R.; Myllymaa, S.; Kullaa, A.M. Salivary metabolomics in the diagnosis of oral cancer and periodontal diseases. *Journal of periodontal research* **2015**, *16*.
34. Mena-Bravo, A.; Luque de Castro, M.D. Sweat: A sample with limited present applications and promising future in metabolomics. *Journal of Pharmaceutical and Biomedical Analysis* **2014**, *90*, 139-147.
35. Bach, J.-P.; Gold, M.; Mengel, D.; Hattesohl, A.; Lubbe, D.; Schmid, S.; Tackenberg, B.; Rieke, J.; Maddula, S.; Baumbach, J.I., *et al.* Measuring compounds in exhaled air to detect alzheimer's disease and parkinson?S disease. *PloS one* **2015**, *10*, e0132227.
36. Pijls, K.E.; Smolinska, A.; Jonkers, D.M.A.E.; Dallinga, J.W.; Masclee, A.A.M.; Koek, G.H.; van Schooten, F.-J. A profile of volatile organic compounds in exhaled air as a potential non-invasive biomarker for liver cirrhosis. *Scientific Reports* **2016**, *6*, 19903.
37. Koulman, A.; Prentice, P.; Wong, M.C.; Matthews, L.; Bond, N.J.; Eiden, M.; Griffin, J.L.; Dunger, D.B. The development and validation of a fast and robust dried blood spot based lipid profiling method to study infant metabolism. *Metabolomics* **2014**, *10*, 1018-1025.

38. Wilson, I. Global metabolic profiling (metabonomics/metabolomics) using dried blood spots: Advantages and pitfalls. *Bioanalysis* **2011**, *3*, 2255-2257.
39. Michopoulos, F.; Theodoridis, G.; Smith, C.J.; Wilson, I.D. Metabolite profiles from dried blood spots for metabonomic studies using uplc combined with orthogonal acceleration tof-ms: Effects of different papers and sample storage stability. *Bioanalysis* **2011**, *3*, 2757-2767.
40. Denes, J.; Szabo, E.; Robinette, S.L.; Szatmari, I.; Szonyi, L.; Kreuder, J.G.; Rauterberg, E.W.; Takats, Z. Metabonomics of newborn screening dried blood spot samples: A novel approach in the screening and diagnostics of inborn errors of metabolism. *Analytical chemistry* **2012**, *84*, 10113-10120.
41. Wagner, M.; Tonoli, D.; Varesio, E.; Hopfgartner, G. The use of mass spectrometry to analyze dried blood spots. *Mass spectrometry reviews* **2014**, 1–78.
42. Oliveira, R.V.; Henion, J.; Wickremsinhe, E.R. Automated high-capacity on-line extraction and bioanalysis of dried blood spot samples using liquid chromatography/high-resolution accurate mass spectrometry. *Rapid Communications in Mass Spectrometry* **2014**, *28*, 2415-2426.
43. Do, K.T.; Kastenmüller, G.; Mook-Kanamori, D.O.; Yousri, N.A.; Theis, F.J.; Suhre, K.; Krumsiek, J. Network-based approach for analyzing intra- and interfluid metabolite associations in human blood, urine, and saliva. *Journal of Proteome Research* **2015**, *14*, 1183-1194.
44. Torell, F.; Bennett, K.; Cereghini, S.; Rannar, S.; Lundstedt-Enkel, K.; Moritz, T.; Haumaitre, C.; Trygg, J.; Lundstedt, T. Multi-organ contribution to the metabolic plasma profile using hierarchical modelling. *PloS one* **2015**, *10*, e0129260.
45. Sitnikov, D.G.; Monnin, C.S.; Vuckovic, D. Systematic assessment of seven solvent and solid-phase extraction methods for metabolomics analysis of human plasma by lc-ms. *Scientific Reports* **2016**, *6*, 38885.
46. Ser, Z.; Liu, X.; Tang, N.N.; Locasale, J.W. Extraction parameters for metabolomics from cultured cells. *Analytical biochemistry* **2015**, *475*, 22-28.
47. Balog, J.; Sasi-Szabo, L.; Kinross, J.; Lewis, M.R.; Muirhead, L.J.; Veselkov, K.; Mirnezami, R.; Dezso, B.; Damjanovich, L.; Darzi, A., *et al.* Intraoperative tissue identification using rapid evaporative ionization mass spectrometry. *Sci. Transl. Med.* **2013**, *5*, 11.
48. Athersuch, T. Metabolome analyses in exposome studies: Profiling methods for a vast chemical space. *Archives of Biochemistry and Biophysics* **2016**, *589*, 177-186.
49. Pauling, L.; Robinson, A.B.; Teranishi, R.; Cary, P. Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. *Proceedings of the National Academy of Sciences of the United States of America* **1971**, *68*, 2374-2376.
50. Nagana Gowda, G.A.; Raftery, D. Recent advances in nmr-based metabolomics. *Analytical chemistry* **2017**, *89*, 490-510.
51. Markley, J.L.; Brüschweiler, R.; Edison, A.S.; Eghbalnia, H.R.; Powers, R.; Raftery, D.; Wishart, D.S. The future of nmr-based metabolomics. *Current Opinion in Biotechnology* **2017**, *43*, 34-40.
52. Larive, C.K.; Barding, G.A., Jr.; Dinges, M.M. Nmr spectroscopy for metabolomics and metabolic profiling. *Analytical chemistry* **2015**, *87*, 133-146.
53. Fan, T.W.; Lane, A.N.; Higashi, R.M. Stable isotope resolved metabolomics studies in ex vivo tissue slices. *Bio-protocol* **2016**, *6*.
54. Fan, T.W.M.; Lane, A.N. Applications of nmr spectroscopy to systems biochemistry. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2016**, *92–93*, 18-53.
55. Verma, A.; Kumar, I.; Verma, N.; Aggarwal, P.; Ojha, R. Magnetic resonance spectroscopy — revisiting the biochemical and molecular milieu of brain tumors. *BBA Clinical* **2016**, *5*, 170-178.
56. Emwas, A.H. The strengths and weaknesses of nmr spectroscopy and mass spectrometry with particular focus on metabolomics research. *Methods in molecular biology (Clifton, N.J.)* **2015**, *1277*, 161-193.
57. Murray, K.K.; Boyd, R.K.; Eberlin, M.N.; Langley, G.J.; Li, L.; Naito, Y. Definitions of terms relating to mass spectrometry (iupac recommendations 2013). *Pure and Applied Chemistry* **2013**, *85*, 1515-1609.
58. González-Domínguez, R.; Sayago, A.; Fernández-Recamales, Á. Direct infusion mass spectrometry for metabolomic phenotyping of diseases. *Bioanalysis* **2017**, *9*, 131-148.
59. Mann, M.; Hendrickson, R.C.; Pandey, A. Analysis of proteins and proteomes by mass spectrometry. *Annual review of biochemistry* **2001**, *70*, 437-473.
60. Plumb, R.S.; Johnson, K.A.; Rainville, P.; Smith, B.W.; Wilson, I.D.; Castro-Perez, J.M.; Nicholson, J.K. Uplc/mse; a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Communications in Mass Spectrometry* **2006**, *20*, 1989-1994.
61. Zhou, J.; Li, Y.; Chen, X.; Zhong, L.; Yin, Y. Development of data-independent acquisition workflows for metabolomic analysis on a quadrupole-orbitrap platform. *Talanta* **2017**, *164*, 128-136.
62. 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. *Journal of Chromatography A* **2013**, *1292*, 51-65.
63. Denisov, E.; Damoc, E.; Lange, O.; Makarov, A. Orbitrap mass spectrometry with resolving powers above 1,000,000. *International Journal of Mass Spectrometry* **2012**, *325*, 80-85.

64. Glish, G.L.; Vachet, R.W. The basics of mass spectrometry in the twenty-first century. *Nat Rev Drug Discov* **2003**, *2*, 140-150.
65. Want, E.J.; Masson, P.; Michopoulos, F.; Wilson, I.D.; Theodoridis, G.; Plumb, R.S.; Shockcor, J.; Loftus, N.; Holmes, E.; Nicholson, J.K. Global metabolic profiling of animal and human tissues via uplc-ms. *Nature protocols* **2013**, *8*, 17-32.
66. Want, E.J.; Wilson, I.D.; Gika, H.; Theodoridis, G.; Plumb, R.S.; Shockcor, J.; Holmes, E.; Nicholson, J.K. Global metabolic profiling procedures for urine using uplc-ms. *Nature protocols* **2010**, *5*, 1005-1018.
67. Chan, E.C.Y.; Pasikanti, K.K.; Nicholson, J.K. Global urinary metabolic profiling procedures using gas chromatography-mass spectrometry. *Nature protocols* **2011**, *6*, 1483-1499.
68. Ramautar, R.; Somsen, G.W.; de Jong, G.J. Ce-ms for metabolomics: Developments and applications in the period 2012-2014. *Electrophoresis* **2015**, *36*, 212-224.
69. 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.
70. Tang, D.Q.; Zou, L.; Yin, X.X.; Ong, C.N. Hilic-ms for metabolomics: An attractive and complementary approach to rplc-ms. *Mass spectrometry reviews* **2014**.
71. Haggarty, J.; Burgess, K.E.V. Recent advances in liquid and gas chromatography methodology for extending coverage of the metabolome. *Current Opinion in Biotechnology* **2017**, *43*, 77-85.
72. Kaufmann, A. Combining uhplc and high-resolution ms: A viable approach for the analysis of complex samples? *TrAC Trends in Analytical Chemistry* **2014**, *63*, 113-128.
73. Kopka, J.; Schauer, N.; Krueger, S.; Birkemeyer, C.; Usadel, B.; Bergmüller, E.; Dörmann, P.; Weckwerth, W.; Gibon, Y.; Stitt, M. Gmd@ csb. Db: The golm metabolome database. *Bioinformatics (Oxford, England)* **2005**, *21*, 1635-1638.
74. Vinaixa, M.; Schymanski, E.L.; Neumann, S.; Navarro, M.; Salek, R.M.; Yanes, O. Mass spectral databases for lc/ms- and gc/ms-based metabolomics: State of the field and future prospects. *TrAC Trends in Analytical Chemistry* **2016**, *78*, 23-35.
75. Moros, G.; Chatziioannou, A.C.; Gika, H.G.; Raikos, N.; Theodoridis, G. Investigation of the derivatization conditions for gc–ms metabolomics of biological samples. *Bioanalysis* **2017**, *9*, 53-65.
76. Kaal, E.; Janssen, H.-G. Extending the molecular application range of gas chromatography. *Journal of Chromatography A* **2008**, *1184*, 43-60.
77. García, A.; Godzien, J.; López-Gonzálvez, Á.; Barbas, C. Capillary electrophoresis mass spectrometry as a tool for untargeted metabolomics. **2016**.
78. Rodrigues, K.T.; Cieslarová, Z.; Tavares, M.F.M.; Simionato, A.V.C. Strategies involving mass spectrometry combined with capillary electrophoresis in metabolomics. In *Metabolomics: From fundamentals to clinical applications*, Springer: 2017; pp 99-141.
79. Hill, H.H., Jr.; Siems, W.F.; St Louis, R.H.; McMinn, D.G. Ion mobility spectrometry. *Analytical chemistry* **1990**, *62*, 1201A-1209A.
80. Paglia, G.; Angel, P.; Williams, J.P.; Richardson, K.; Olivos, H.J.; Thompson, J.W.; Menikarachchi, L.; Lai, S.; Walsh, C.; Moseley, A., *et al.* Ion mobility-derived collision cross section as an additional measure for lipid fingerprinting and identification. *Analytical chemistry* **2015**, *87*, 1137-1144.
81. Maldini, M.; Natella, F.; Baima, S.; Morelli, G.; Scaccini, C.; Langridge, J.; Astarita, G. Untargeted metabolomics reveals predominant alterations in lipid metabolism following light exposure in broccoli sprouts. *Int J Mol Sci* **2015**, *16*, 13678-13691.
82. Paglia, G.; Williams, J.P.; Menikarachchi, L.; Thompson, J.W.; Tyldesley-Worster, R.; Halldórsson, S.; Rolfsson, O.; Moseley, A.; Grant, D.; Langridge, J., *et al.* Ion mobility derived collision cross sections to support metabolomics applications. *Analytical chemistry* **2014**, *86*, 3985-3993.
83. Wickramasekara, S.I.; Zandkarimi, F.; Morre, J.; Kirkwood, J.; Legette, L.; Jiang, Y.; Gombart, A.F.; Stevens, J.F.; Maier, C.S. Electrospray quadrupole travelling wave ion mobility time-of-flight mass spectrometry for the detection of plasma metabolome changes caused by xanthohumol in obese zucker (fa/fa) rats. *Metabolites* **2013**, *3*, 701-717.
84. Dwivedi, P.; Schultz, A.J.; Hill, H.H. Metabolic profiling of human blood by high resolution ion mobility mass spectrometry (im-ms). *Int J Mass Spectrom* **2010**, *298*, 78-90.
85. Hauschild, A.C.; Frisch, T.; Baumbach, J.I.; Baumbach, J. Carotta: Revealing hidden confounder markers in metabolic breath profiles. *Metabolites* **2015**, *5*, 344-363.
86. Smolinska, A.; Hauschild, A.C.; Fijten, R.R.; Dallinga, J.W.; Baumbach, J.; van Schooten, F.J. Current breathomics--a review on data pre-processing techniques and machine learning in metabolomics breath analysis. *J Breath Res* **2014**, *8*, 027105.
87. Fenn, L.; Kliman, M.; Mahsut, A.; Zhao, S.; McLean, J. Characterizing ion mobility-mass spectrometry conformation space for the analysis of complex biological samples. *Analytical and Bioanalytical Chemistry* **2009**, *394*, 235-244.
88. Fenn, L.; McLean, J. Biomolecular structural separations by ion mobility–mass spectrometry. *Analytical and Bioanalytical Chemistry* **2008**, *391*, 905-909.
89. Kliman, M.; May, J.C.; McLean, J.A. Lipid analysis and lipidomics by structurally selective ion mobility-mass spectrometry. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **2011**, *1811*, 935-945.

90. Tebani, A.; Schmitz-Afonso, I.; Rutledge, D.N.; Gonzalez, B.J.; Bekri, S.; Afonso, C. Optimization of a liquid chromatography ion mobility-mass spectrometry method for untargeted metabolomics using experimental design and multivariate data analysis. *Analytica Chimica Acta* **2016**, *913*, 55-62.
91. Domalain, V.; Tognetti, V.; Hubert-Roux, M.; Lange, C.M.; Joubert, L.; Baudoux, J.; Rouden, J.; Afonso, C. Role of cationization and multimers formation for diastereomers differentiation by ion mobility-mass spectrometry. *J Am Soc Mass Spectrom* **2013**, *24*, 1437-1445.
92. Harper, B.; Neumann, E.K.; Stow, S.M.; May, J.C.; McLean, J.A.; Solouki, T. Determination of ion mobility collision cross sections for unresolved isomeric mixtures using tandem mass spectrometry and chemometric deconvolution. *Anal Chim Acta* **2016**, *939*, 64-72.
93. Zhou, Z.; Shen, X.; Tu, J.; Zhu, Z.-J. Large-scale prediction of collision cross-section values for metabolites in ion mobility-mass spectrometry. *Analytical chemistry* **2016**.
94. Jeanne Dit Fouque, K.; Afonso, C.; Zirah, S.; Hegemann, J.D.; Zimmermann, M.; Marahiel, M.A.; Rebuffat, S.; Lavanant, H. Ion mobility-mass spectrometry of lasso peptides: Signature of a rotaxane topology. *Analytical chemistry* **2015**, *87*, 1166-1172.
95. May, J.C.; Goodwin, C.R.; McLean, J.A. Ion mobility-mass spectrometry strategies for untargeted systems, synthetic, and chemical biology. *Curr Opin Biotechnol* **2015**, *31*, 117-121.
96. Sherrod, S.D.; McLean, J.A. Systems-wide high-dimensional data acquisition and informatics using structural mass spectrometry strategies. *Clin Chem* **2015**, *62*, 77-83.
97. Twohig, M.; Shockcor, J.P.; Wilson, I.D.; Nicholson, J.K.; Plumb, R.S. Use of an atmospheric solids analysis probe (asap) for high throughput screening of biological fluids: Preliminary applications on urine and bile. *Journal of proteome research* **2010**, *9*, 3590-3597.
98. Eberlin, L.S.; Norton, I.; Orringer, D.; Dunn, I.F.; Liu, X.; Ide, J.L.; Jarmusch, A.K.; Ligon, K.L.; Jolesz, F.A.; Golby, A.J., *et al.* Ambient mass spectrometry for the intraoperative molecular diagnosis of human brain tumors. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, 1611-1616.
99. Ferreira, C.R.; Jarmusch, A.K.; Pirro, V.; Alfaro, C.M.; Gonzalez-Serrano, A.F.; Niemann, H.; Wheeler, M.B.; Rabel, R.A.; Hallett, J.E.; Houser, R., *et al.* Ambient ionisation mass spectrometry for lipid profiling and structural analysis of mammalian oocytes, preimplantation embryos and stem cells. *Reproduction, fertility, and development* **2015**, *27*, 621-637.
100. Kerian, K.S.; Jarmusch, A.K.; Pirro, V.; Koch, M.O.; Masterson, T.A.; Cheng, L.; Cooks, R.G. Differentiation of prostate cancer from normal tissue in radical prostatectomy specimens by desorption electrospray ionization and touch spray ionization mass spectrometry. *Analyst* **2015**, *140*, 1090-1098.
101. Balog, J.; Kumar, S.; Alexander, J.; Golf, O.; Huang, J.; Wiggins, T.; Abbassi-Ghadi, N.; Enyedi, A.; Kacska, S.; Kinross, J., *et al.* In vivo endoscopic tissue identification by rapid evaporative ionization mass spectrometry (reims). *Angew Chem Int Ed Engl* **2015**, *54*, 11059–11062
102. Dunham, S.J.B.; Ellis, J.F.; Li, B.; Sweedler, J.V. Mass spectrometry imaging of complex microbial communities. *Accounts of Chemical Research* **2017**, *50*, 96-104.
103. Arentz, G.; Mittal, P.; Zhang, C.; Ho, Y.Y.; Briggs, M.; Winderbaum, L.; Hoffmann, M.K.; Hoffmann, P. Chapter two - applications of mass spectrometry imaging to cancer. In *Advances in cancer research*, Richard, R.D.; Liam, A.M., Eds. Academic Press: 2017; Vol. Volume 134, pp 27-66.
104. Ifa, D.R.; Eberlin, L.S. Ambient ionization mass spectrometry for cancer diagnosis and surgical margin evaluation. *Clinical Chemistry* **2016**, *62*, 111-123.

Advances in metabolome information retrieval: turning chemistry into biology

Part II: Biological Information Recovery

Abdellah Tebani [1,2,3], Carlos Afonso [3], Soumeya Bekri [1,2,*]

[1] Department of Metabolic Biochemistry, Rouen University Hospital, Rouen 76000, France;
[2] Normandie Univ, UNIROUEN, CHU Rouen, IRIB, INSERM U1245, Rouen 76000, France
[3] Normandie Univ, UNIROUEN, INSA Rouen, CNRS, COBRA, Rouen 76000, France

Abstract

This work reports the second part of a review intending to give the state of the art of major metabolic phenotyping strategies. It particularly deals with inherent advantages and limits regarding data analysis issues and biological information retrieval tools along with translational challenges. This Part starts with introducing the main data preprocessing strategies of the different metabolomics data. Then, it describes the main data analysis techniques including univariate and multivariate aspects. It also addresses the challenges related to metabolite annotation and characterization. Finally, functional analysis including pathway and network strategies are discussed. The last section of this review is devoted to practical considerations and current challenges and pathways to bring metabolomics into clinical environments.

.

Keywords: Omics; metabolomics; metabolome; mass spectrometry; nuclear magnetic resonance; chemometrics.

1. Introduction

Addressing biology as an informational science is a key driver to translate biological data into actionable knowledge. This requires innovative tools that allow to extract information from high dimensional data. Bioinformatics is the field that is born to tackle this challenge [1]. Bioinformatics applies informatics techniques such as applied mathematics, computer science and statistics to retrieve the organization the biological information. In short, bioinformatics is a management information system for a biological system [2]. The metabolomic data requires adapted statistical tools to retrieve as much as possible chemical information to translate it into biological knowledge. The major challenge is to reduce the dimensionality by selecting informative signals from the noise. To achieve this goal, chemometric tools are widely used. Chemometrics is the science of extracting useful information from chemical systems using data-driven means [3]. It is inherently interdisciplinary, borrowing methods from data-analytic disciplines such as statistics, signal processing, applied mathematics, and computer science. Descriptive and predictive problems could be addressed using chemical data. This second part of the review intends to give the state of the art of metabolomics data handling strategies along with their inherent advantages and limits regarding data analysis issues. Furthermore, biological information retrieval tools and their translational challenges into actionable results are described. Finally, practical considerations and current challenges to bring metabolomics into clinical environment are discussed. The general metabolomics workflow is presented in Fig. 1.

2. Biological information recovery

The analytical performance improvements associated with metabolomics platforms have led to the generation of complex and high-dimensional data sets. Handling the huge amount of generated data in a smoothly high-throughput fashion is a very important issue for transforming the data into clinically actionable knowledge.

2.1. Preprocessing

Targeted metabolomics aims to process data sets retrieved from a subset of the metabolome. It contains predefined, chemically characterized, and biochemically annotated metabolites. The main advantages of targeted metabolomics are that no analytical artifacts are carried throughout the downstream analysis; only a set of selected metabolites are analyzed. However, if numerous metabolites are involved, data analysis is quite time-consuming. Different automated processes have been developed [4-6] along with commercial solutions from instrument vendors. In contrast, the untargeted approach attempts a comprehensive analysis of all measurable metabolites in a given sample, including unknowns. It requires a holistic analysis of high-dimensional raw data sets, which in turn requires reducing the data into more computationally manageable formats without significantly compromising the contained chemical information. Because of noise, sample variation, or analytical/instrument factors, NMR and MS spectra often show differences in width, position, and peak shape. The goal of preprocessing is to correct these differences for better quantification of metabolites and enhanced intersample comparability. Several preprocessing considerations and methods can be applied to both NMR and MS data [7]. MS data preprocessing includes some or all of the following steps: noise filtering, baseline correction, peak detection, peak alignment, and spectral deconvolution. The order of the steps may differ between algorithms. Noise filtering is often applied to MS data to improve peak detection, and many different noise filters exist, including Gaussian, Savitzky–Golay, and wavelet-based filters [8]. The aim of the peak detection and deconvolution step is to identify and quantify the signals that correspond to the analytes (metabolites) in a given sample. Peak detection algorithms follow two strategies: derivative techniques or matched filter response [8,9]. A deconvolution step is used to separate overlapping peaks in order to improve peak detection [10]. Furthermore, a de-isotoping step is used to cluster the isotopic peaks corresponding to the same chemical feature to clean the data matrix. Alignment of the detected features across different samples aims to remove intersample shifts, and several alignment algorithms have been developed [9,11]. The data dimensionality has to be reduced to make them applicable to instruments paired with MS. Different strategies enable data compression such as binning and the "search of regions of interest (ROI)" methods that are the most adequate hyphenated MS data sets.

NMR data preprocessing typically includes baseline correction, alignment, and binning. Baseline correction aims to correct systematic baseline distortion. Some spectral regions, such as that of water, are often removed. Peak shifts due to differences in instrumental factors such as salt concentrations, temperature, and pH changes can be corrected by alignment procedures [12]. Binning or bucketing is a dimension reduction method that splits the spectra into segments or bins and assigns a representative value to each bin. However, binning can hamper spectral resolution. A comparison of some peak-picking algorithms used in untargeted MS-based metabolomics have been reported [13]. The typical output of the preprocessing step is a data matrix that contains the detected features and the corresponding intensity (abundance) in each sample.

2.2. Normalization

As with other omics, metabolomics data have several intrinsic characteristics, such as their asymmetric distribution [14] and a substantial proportion of instrumental, analytical, and biological noise [15,16]. Thus, the goal of data normalization is to eliminate experimental biases related to the abundance of detected features between samples without compromising biological variations. Most of the methods are inspired by previous omic strategies (genomics and transcriptomics) that suffer from similar experimental biases [17]. Indeed, the chemical diversity of metabolites and interindividual variations lead to changes in extraction and MS ionization yields, making it difficult to distinguish changes of biological interest from analytical biases (instrumentation, operators, and reagents). Strategies for normalization of metabolomics data can be divided into statistical approaches and chemical approaches. Statistical approaches are based on statistical models that define correction factors specific to each sample from the complete data set [18], such as normalization by standard deviation [19], mean global intensity [20], quantile normalization [21], probabilistic quotient normalization [22], linear baseline scaling [23], nonlinear baseline normalization [24], contrast normalization [25], cubic splines [26], cyclic loess [27], QC-robust spline batch correction [28], MS total useful signal [29], and support vector regression [30]. Chemical approaches are based on one or more reference compounds [31-33], internal standards, or endogenous or exogenous compounds that are used to normalize the entire chromatogram (single compound) or certain regions of the chromatogram by normalizing each zone according to a standard that is eluted in that region. Other strategies based on the characteristics of the studied matrix, such as dry mass of the samples, volume (e.g., 24-hour urine), and osmolality. Protein or creatinine levels can also be used [34]. A comprehensive comparison of state-of-the-art normalization techniques was recently reported [18].

2.3. Transformation, centering, and scaling

Statistical methods assume that the data under analysis have a specific type of probability distribution. Thus, the inferences made from the data depend on the chosen distribution. If the data under examination do not exhibit that distribution, then the inferences could be false or misleading. Most parametric methods in metabolomics assume that the data have a Gaussian distribution, which means that the built model's errors (residuals) are normally distributed with a homogeneous variance (homoscedastic noise). However, in metabolomics, MS and NMR data are hampered by heteroscedastic noise from different sources. Furthermore, the feature distributions can be skewed. So, transformations aim to correct for heteroscedasticity and skewness before statistical analysis. To build statistically meaningful and interpretable models in metabolomics, data are often transformed before modelling. Different mathematical transformations can be used, such as log transformation and power transformation [35]. Multivariate analytical methods are based on latent variable projections that extract information from the data by projecting observations onto the direction of the maximum variance. Hence, NMR and MS data analysis by these methods mainly focuses on the average spectrum. This approach may mask underlying biological variation because more abundant metabolites will exhibit high values in the data matrix and subsequently show large differences among samples compared to less abundant metabolites. Data scaling methods divide each data point for a given feature by a scaling factor that is a measure of data dispersion for that feature. Therefore, scaling the data aims to remove the offset from the data and focus on the biological variation regarding similarities and dissimilarities of samples. There are several scaling methods such as auto-scaling (unit variance scaling), in which the mean and the standard deviation of the feature are calculated. The feature is first mean-centered, and each observation on the mean-centered feature is then divided by the standard deviation. The aim of auto-scaling is to give equal weights to all features, but this method is very sensitive to large deviations from the sample mean. Thus, pareto scaling is most popular alternative in metabolomics. It is similar to auto-scaling, but each observation in the mean-centered feature is divided by the square root of the standard deviation. Pareto scaling is a compromise between mean-centering and auto-scaling. Other methods have also been described [35].

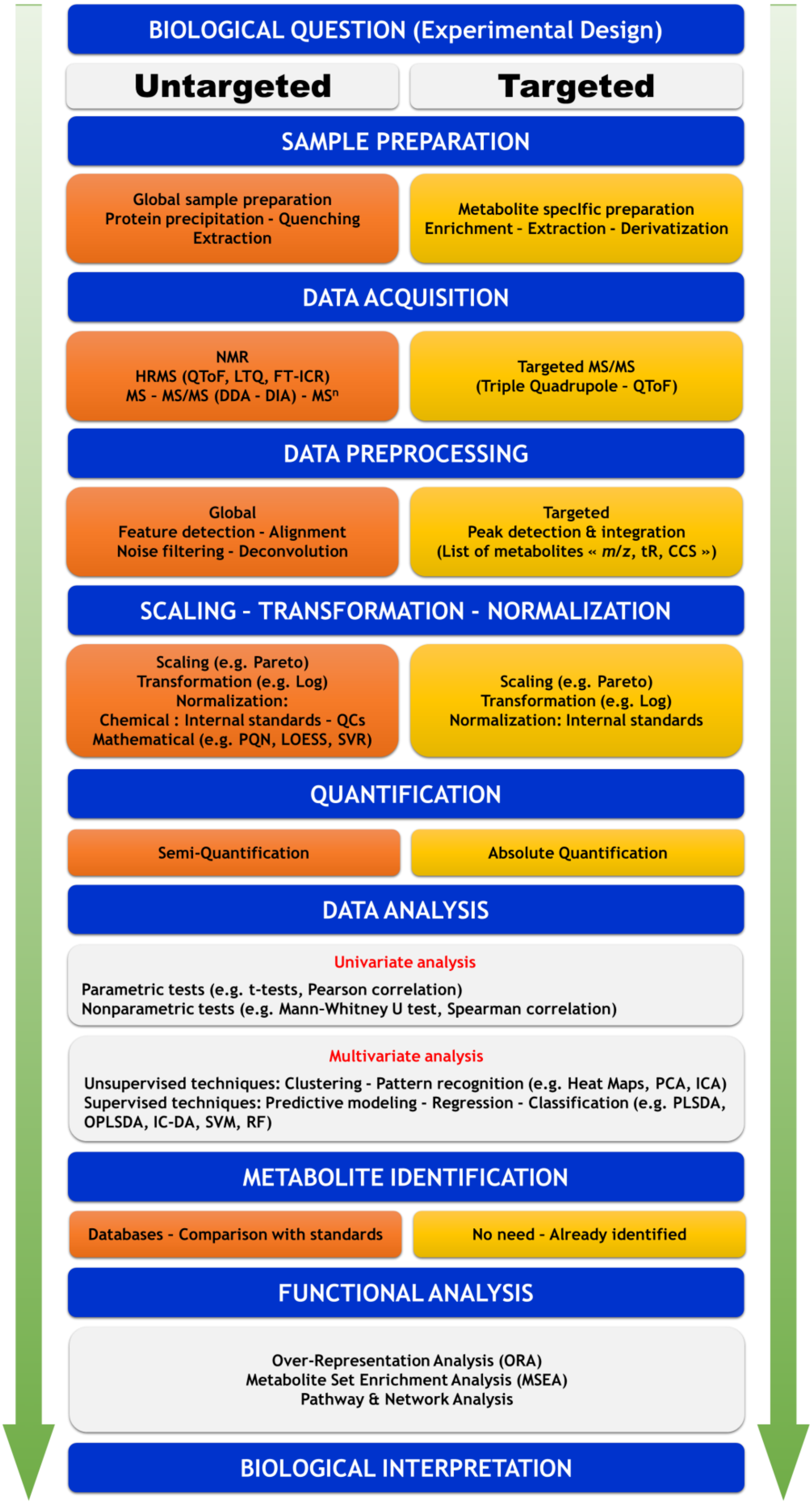

BIOLOGICAL QUESTION (Experimental Design)
Untargeted
Targeted
SAMPLE PREPARATION
Global sample preparation
Protein precipitation - Quenching
Extraction
Metabolite specific preparation
Enrichment - Extraction - Derivatization
DATA ACQUISITION
NMR
HRMS (QToF, LTQ, FT-ICR)
MS - MS/MS (DDA - DIA) - MSⁿ
Targeted MS/MS
(Triple Quadrupole - QToF)
DATA PREPROCESSING
Global
Feature detection - Alignment
Noise filtering - Deconvolution
Targeted
Peak detection & integration
(List of metabolites « m/z, tR, CCS »)
SCALING - TRANSFORMATION - NORMALIZATION
Scaling (e.g. Pareto)
Transformation (e.g. Log)
Normalization:
Chemical : Internal standards - QCs
Mathematical (e.g. PQN, LOESS, SVR)
Scaling (e.g. Pareto)
Transformation (e.g. Log)
Normalization: Internal standards
QUANTIFICATION
Semi-Quantification
Absolute Quantification
DATA ANALYSIS
Univariate analysis
Parametric tests (e.g. t-tests, Pearson correlation)
Nonparametric tests (e.g. Mann-Whitney U test, Spearman correlation)
Multivariate analysis
Unsupervised techniques: Clustering - Pattern recognition (e.g. Heat Maps, PCA, ICA)
Supervised techniques: Predictive modeling - Regression - Classification (e.g. PLSDA, OPLSDA, IC-DA, SVM, RF)
METABOLITE IDENTIFICATION
Databases - Comparison with standards
No need - Already identified
FUNCTIONAL ANALYSIS
Over-Representation Analysis (ORA)
Metabolite Set Enrichment Analysis (MSEA)
Pathway & Network Analysis
BIOLOGICAL INTERPRETATION

Figure 1. General metabolomics workflow.

3. Data analysis

3.1. Univariate data analysis

Univariate statistical methods can be used in metabolomics. Their main limitation is that they consider only one variable at a time, which may not be convenient for high-dimensional data. Parametric tests such as Student's *t*-test and ANOVA are commonly applied to assess the differences between two or more groups, respectively, provided that the normality assumption is verified [36]. Otherwise, if normality is not assumed, nonparametric test such as Mann–Whitney *U* test or Kruskal–Wallis one-way ANOVA can be used. Another important issue is that applying multiple univariate tests in parallel with a high-dimensional data set raises the multiple testing problem. Since a large number of features are simultaneously analyzed in metabolomics, the probability of accidentally finding a statistically significant difference (i.e., true positive) is high. Different correction methods can be used to handle this multiple testing issue. In the Bonferroni correction, the significance level for a hypothesis is divided by the number of hypotheses simultaneously being tested [36]. Hence, the Bonferroni correction is considered a conservative correction method. Less conservative methods are available and are based on lowering the false-discovery rate (FDR). FDR-based methods minimize the expected proportion of false positives among the total number of positives [37]. Less restrictive approaches such as FDR methods seem to be more useful. It should be noted that potential confounding factors such as sex, age, or diet may lead to spurious results if not properly addressed. Furthermore, the main disadvantage of univariate methods is their lack of feature correlations and insights about interactions. Hence, advanced multivariate approaches are more suitable for in-depth inferences.

3.2. Multivariate data analysis

To translate biological data into knowledge, biology should be addressed as an informational science, using tools that allow tracking of information on a large scale. To do so, an entire field, bioinformatices, was created [1]. Bioinformatics allows tracking biology from an informational perspective. It permits data collection, analysis, parsing, and contextual interpretation, and it supports decision-making on those bases. Thus, bioinformatics can be defined as conceptualizing biology in terms of molecular components and by applying "informatics techniques" borrowed from disciplines such as applied mathematics, computer science, and statistics to understand and organize information on a large scale [2]. Similar to any other biological data, metabolomics data fall into this definition. However, metabolomics data require statistical tools to be adapted to permit retrieval of as much chemical information as possible from the data in order to translate it into actionable knowledge. The major challenge is to reduce the dimensionality by selecting informative metabolic signals from the highly noisy raw data. Chemometric tools are widely used to achieve this goal. Chemometrics is defined as the science of extracting useful information from chemical systems by data-driven means [3]. It includes (i) the analysis of chemical data for pertinent information retrieval and (ii) the design of chemical experiments that will yield data that provide optimal information about initially formulated questions. It is intrinsically interdisciplinary, using multivariate statistics, applied mathematics, and computing. Chemometrics may be applied to solve both descriptive and predictive problems, using biochemical data. Chemometrics methods are mainly divided into unsupervised and supervised methods. In unsupervised methods, no assumptions are made about the samples and the aim is mainly exploratory. In supervised methods, samples are assigned to classes or each sample is associated with a specific outcome value, and the aim is mainly explanatory and predictive. In multivariate methods, representive samples are presented as points in the space of the initial variables. The samples can then be projected into a lower dimensionality space based on components or latent variables, such as a line, a plane, or a hyperplane, which can be seen as the shadow of the initial data set viewed from its best perspective. The sample coordinates of the newly defined latent variables are the scores, while the directions of variance to which they are projected are the loadings. The loadings vector for each latent variable contains the weights of each of the initial variables (metabolites) for that latent variable. Unsupervised methods attempt to reveal patterns or clustering trends in the data that underpin relationships between the samples. These methods also highlight the variables that are responsible for these relationships, using visualization means. In metabolomics data, metabolic similarity shapes the observed clustering. Principal component analysis [38] is a widely used pattern recognition method; it is a projection-based method that reduces the dimensionality of the data by creating components. Principal component analysis allows a two- or three-dimensional visualization of the data. Because it contains no assumptions on the data, it is used as an initial visualization and exploratory tool to detect trends, groups, and outliers. It allows simpler global visualization by representing the variance in a small number of uncorrelated latent variables. Independent component analysis (ICA) is another unsupervised method that is a blind source separation method that separates multivariate signals into additive subcomponents. ICA attempts to recover the original signals by estimating a linear transformation, using a criterion that measures statistical independence among the sources [39]. Its interpretation is similar to PCA, but instead of orthogonal components, it calculates non-Gaussian and mutually independent components. ICA does not order the components according to variance, and the number of components influences the structure of the components themselves. Thus, determining the right number of components is very important. Different algorithms have been described for determining ICA components [40,41]. Compared to PCA, ICA as a linear method could provide potential benefits for untargeted metabolomics; more meaningful components would be extracted by optimizing the independence condition instead of the variance maximization used in PCA. Independence conditions detected by ICA involve both orthogonality (linear independence) and independence, while classical PCA only ensures orthogonality between components. Therefore, ICA could potentially extract hidden information from the data set. ICA has been successfully used in metabolomics [42-44]. Other unsupervised methods, such as clustering, aim to identify naturally occurring clusters in the data set by using similarity measures defined by distance and linkage metrics [45]. The method starts by viewing each single object as a cluster. Then it iteratively finds the minimal distance between two objects and cluster them, then three objects, and so on, updating the clusters after each iteration until all objects are part of the same cluster. A dendrogram or a heat map can then be created to visualize the similarities between samples. Commonly used clustering methods include correlation matrix, k-means clustering [46], hierarchical cluster analysis [47], and self-organizing maps [48,49]. Supervised methods handle data sets that have response variables. When the variables are discrete (e.g., control group versus diseased group), the task is called classification. When the variables are continuous (e.g., metabolite concentration) the task is called regression. The main purposes of supervised techniques are (i) to determine the association between the response variable and the predictors (metabolites) and (ii) to make accurate predictions based on the predictors. In supervised methods, the multivariate data sets can be modeled so that the class label of separate samples (validation set) can be predicted based on mathematical models derived from the original data (training set). In metabolomics biomarker discovery, within the modelling process, it is important to find the simplest combination of metabolites that can produce a suitably effective predictive outcome. Thus, the biomarker discovery process should involve two parameters, the biomarker utility and the number of metabolites used in the predictive model. The main challenges are therefore predictor selection and the evaluation of the fitness and predictive power of the built model. The predictor selection is also called the feature or variable selection step. Feature selection aims to identify important metabolites from among the detected ones that best explain and predict the biological question under consideration. Statistically, it is an optimization approach for recovering the best variable combination from the data. Different feature selection techniques have been described. Some of these suggested strategies are based on univariate or multivariate statistical proprieties of variables used as filters (loading weights, variable importance on projection scores, or regression coefficients), while others are based on optimization algorithms [8,50]. Filter methods rank subsets of variables in order of importance before training the models and then repeat the modelling process using the top metabolites. Each subset model is then evaluated, producing the requisite performance [51]. The wrapper method ranks subsets of variables by running every trained model on the test data set and selecting the model (subset of metabolites) with the best performance. A new method that has been recently described uses resampling, ranking of variable importance, and significance assessment by permutation of the feature values [52]. Recently, another elegant method has been reported that essentially combines estimation of Mahalanobis distances with principal component analysis and variable selection using a penalty metric instead of dimension reduction [53]. This method was successfully applied for diagnostic purposes. We next need goodness-of-fit metrics to assess the model predictive power. Commonly used statistics may include root mean square error (RMSE) for regression problems and sensitivity, specificity, and the area under the receiver-operating characteristic (ROC) curve for classification models. However, the gold standard is to use independent test data sets to assess the predictive power and over-fitting of the model. Sometimes, data collection may be expensive or hampered by limited samples such as in rare diseases. In this case, various resampling methods are used to efficiently use the available data set. These methods include cross-validation, bootstrapping, and jackknifing [54]. Regarding the supervised methods, various techniques can be used in metabolomics. Some of the most used techniques include linear discriminant analysis (LDA) [55,56] and partial least squares (PLS) methods such as PLS-discriminant analysis (PLS-DA) [57] and orthogonal-PLS-DA (OPLS-DA) [58,59], as well as support vector machines [60,61] and random forest [62,63]. Recently, Habchi et al. proposed an innovative supervised method based on ICA called IC-DA. This method has been successfully applied to analyze DIMS metabolomics data [64]. Furthermore, new methods based on topology data analysis are drawing interest and seem promising for data analysis because of their intrinsic flexibility and exploratory and predictive abilities [65,66]. Projection-based linear methods (PLSDA, OPLSDA, and LDA) are popular because of their interpretation convenience. Nonlinear methods, such as neural networks, support vector machines, and random forests, are less common in metabolomics when interpretation is needed. They are mostly used for prediction of new samples for classification/regression purposes. Recently, a new method, called statistical health monitoring (SHM), has been adapted from industrial statistical process control (SPC). In this SHM approach, an individual metabolic profile is compared to a healthy one in a multivariate fashion. Abnormal metabolite patterns are thus detected, and more intelligible interpretation is enabled [67]; however, interpretation is dependent on data analysis. Hence, the aim of metabolomics studies and the data analysis strategy are highly interdependent. Moreover, multivariate and univariate data analysis pipelines are not mutually exclusive, and they are often to used together to enhance the quality of the information recovery. For further details on data analysis techniques and tools used in metabolomics, refer to recent reviews on this issue [68-70].

4. Metabolite annotation and characterization

The identification of the discriminant metabolites is an important step in metabolomics that allows the translation of the potential biomarker into actionable biological information. The introduction of high-resolution mass spectrometers and accurate mass measurements that facilitate access to the chemical formula of the detected peaks has considerably accelerated this step. The combined use of quadrupole ion traps for sequential fragmentation experiments provides additional structural information needed to identify metabolites of interest. However, MS at atmospheric pressure exhibits high variability in the fragmentation profiles generated on different devices, thus limiting the construction of universal spectral data banks such as those obtained by electron ionization or NMR [71]. Indeed, in MS, one or more chemical formulas can be generated if high-resolution instruments are used, which provides a first element for carrying out an interrogation of the existing databases. The acquisition of fragmentation spectra at this stage makes it possible to discriminate the responses obtained previously on the basis of ions produced or neutral losses, characteristic of chemical groups. Given the importance of the identification step, standardization elements have been proposed to harmonize metabolite identification data. Thus, identification standards have been defined within the framework of the Metabolomics Standards Initiative according to the available information on the metabolite to be characterized [72]. Computational tools such as CAMERA [73], ProbMetab [74], AStream [75], and MetAssign [76] have been developed for metabolite annotation. These methods mainly use *m*/*z*, retention time, adduct patterns, isotope patterns, and correlation methods for metabolite annotation. However, in MS the detected *m*/*z* ion and MS database matching is insufficient for unambiguous charcterization. Alhough annotation and retention time prediction are still used to improve identification confidence, complementary orthogonal information is required for reliable assignment of chemical identity, such as retention time matching and molecular dissociation patterns compared to authentic standards [72]. For reliable characterization, a solution may be in a multidimensionnal framework based on orthogonal information integration, which may include accurate mass *m/z*, chromatographic retention time, MS/MS spectra patterns, CCS, chiral form, and peak intensity. Furthermore, hybrid strategies, including pathway network and analysis methods, could enhance metabolite characterization through different metrics integration, including data-driven network topology, chemical features correlation, omics data, and biological databases. Such a multidimensional approach may permit the chemical characterization by merging both extended chemical information and biological context.

5. Functional analysis: translating information into knowledge

One of the fundamental difficulties pathophysiological studies is that diseases might be caused by various genetic and environmental factors and their combinations. In addition, if a disease is caused by a combinatorial effect of many factors, the individual effects of each component might be low and thus hard to unveil. So, considering systems approaches to get deeper and informative biological insights is appealing. Any biological network can be pictured as a collection of linked nodes. The nodes may be genes, proteins, metabolites, diseases, or even individuals. The links or edges represent the interactions between the nodes: metabolic reactions, protein–protein interactions, gene–protein interactions, or interactions between individuals. The distribution of nodes ranges from random to highly clustered. However, biological networks are not random. They are collections of nodes and links that evolve as clusters; therefore, biological networks are referred to as scale-free, which means that they contain few highly-connected nodes called hubs. The core idea of the biological network theory is the modularity structure. Three distinct modules can be defined: topological, functional, and disease modules [77]. A topological module represents a local subset of nodes and links in the network; in this module, nodes have a higher tendency to link to nodes within the same local neighborhood. A functional module is a collection of nodes with similar or correlated function in the same network zone. Finally, a disease module represents a group of network components that together contribute to a cellular function whose disruption results in a disease phenotype. Of note, these three modules are correlated and overlap. Computational biology is gaining increasingly more space in modern biology to embrace this new network perspective. It can be divided into two main fields: knowledge discovery (or data-mining) and simulation-based analysis. The former generates hypotheses by extracting hidden patterns from high-dimensional experimental data. However, the latter tests hypotheses with *in silico* experiments, yielding predictions to be confirmed by *in vitro* and *in vivo* studies [78]. Thus, pathway and network analysis strategies rely on the information generated by metabolomics studies for biological inference [79,80]. Both approaches exploit the interrelationships contained in the metabolomic data. Network modeling and pathway-mapping tools help to decipher the roles of metabolite interactions in a biological disturbance [80]. Conceptual framework of pathway analysis is illustrated in Fig. 2 using experimental data and biological databases (Table 1).

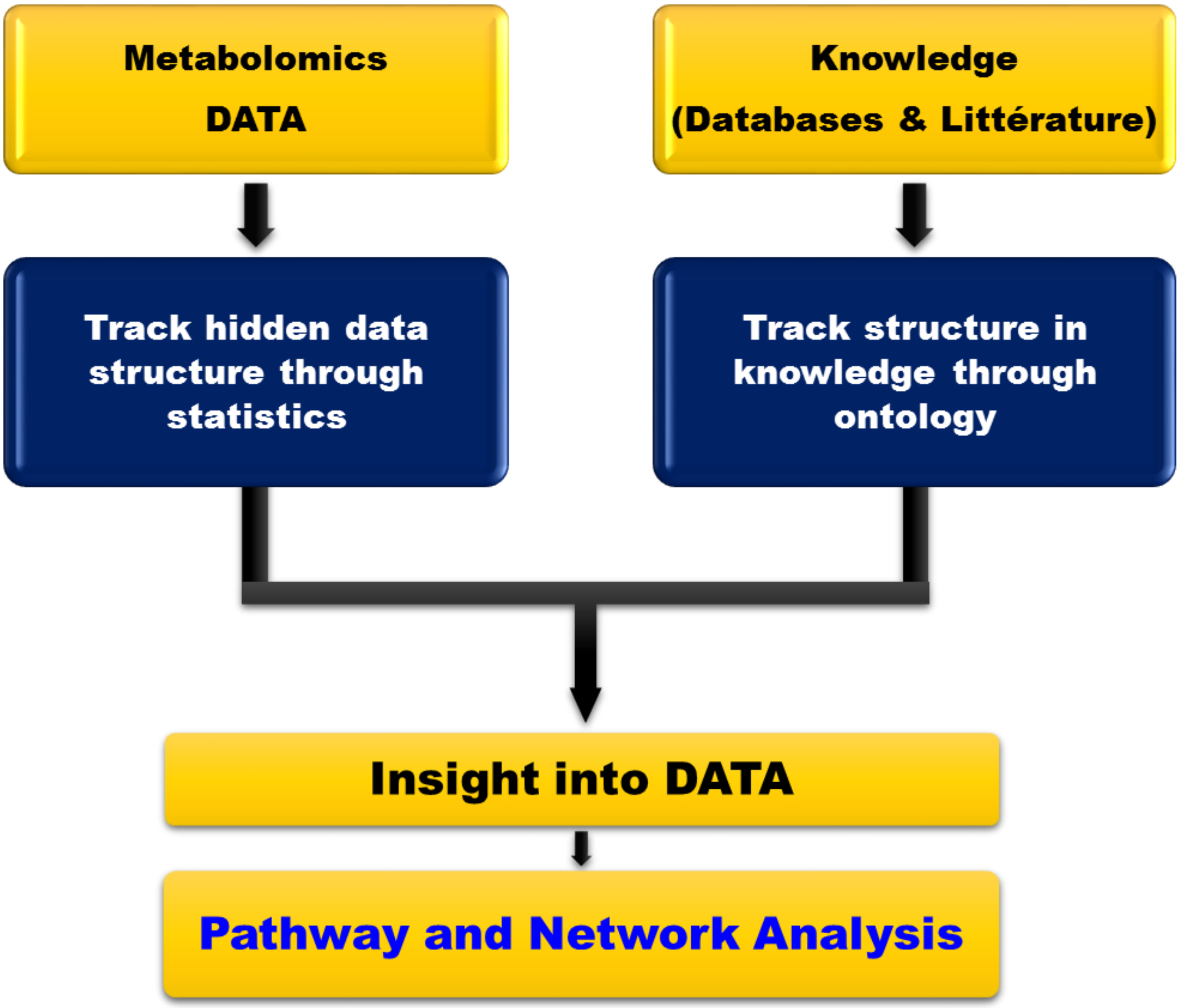

Metabolomics DATA
Knowledge (Databases & Littérature)
Track hidden data structure through statistics
Track structure in knowledge through ontology
Insight into DATA
Pathway and Network Analysis

Figure 2. An illustration of pathway analysis strategies. Metabolome pathway analysis is designed to uncover significant pathway–phenotype relationships within a large data set. In one hand, it unveils hidden data structure in experimental data through differential expression using statistical metrics. In the other hand, it uses prior knowledge retrieved through biological databases and literature. Pathway analysis combines these two pillars to interpret the experimental findings.

Indeed, pathway analysis or metabolite set enrichment analysis (MSEA) are methodologically based on the gene set enrichment analysis approach, previously developed for pathway analysis of gene-expression data [81,82]. There are three distinct methods for performing MSEA [81,83]: overrepresentation analysis (ORA), quantitative enrichment analysis (QEA), and single-sample profiling (SSP). In ORA, the relevant pathways can be detected if the proportion of differentially expressed metabolites within a given pathway is significant. A hypergeometric test or a Fisher’s exact test is used to evaluate the statistical significance of whether the metabolite belongs to the pathway. The final result from an ORA method consists of a list of the most relevant pathways, ranked by *p*-value and/or a *p*-value corrected by multiple hypothesis testing. Due to the selected cutoff method for statistical significance, potentially important components could be omitted in the analysis. QEA uses absolute concentrations of a set of quantified metabolites from multiple samples. Different tools can be used to enrich pathways, such as the Wilcoxon-based test [84], global test [85], or global Ancova [86]. Enriched pathways include pathways in which a set of metabolites are significantly changed or pathways in which a large number of metabolites are significantly changed [83,87]. SSP is used at the sample level, and it requires a list of metabolite concentrations in biofluids (i.e., urine, blood, and cerebrospinal fluid), tissue, or cell type and a database with the normal concentration ranges. Thus, SSP identifies from the data the set of metabolites presenting significantly different levels compared to the normal ranges [83,87]. An important advantage of computational metabolomics lies in the use of correlations among feature signals to map chemical identity. Since metabolites are interconnected by a series of biochemical reactions to build network of metabolites, they can be interrogated using network-based analytical tools. In metabolomics, network analysis uses the high degree of correlation in metabolomics data to build metabolic networks based on the complex relationships of the measured metabolites. Based on the observed relationship patterns in the experimental data, correlation-based methods allow building metabolic networks in which each metabolite represents a node. However, unlike the pathway analysis, the links between nodes denote the level of mathematical correlation between each metabolite pair and are called edge. High correlation coefficients are frequent in metabolomics data, which is due to the presence of systemic associations [88]. Classical correlation coefficients lead to overcrowded networks, and direct and indirect associations are not distinguished. To overcome this problem, partial correlation can be used [88-90]. In this method, the correlation between two metabolites is conditioned against the correlation with the remaining metabolites. Thus, the link between two metabolites is scored according to the differences in ratios between the corresponding metabolites in the two sample groups. Therefore, the related network topology is based on the metabolic differences between the two studied phenotypes. These data-driven strategies have been successfully applied for the reconstruction of metabolic networks from metabolomics data [88,91,92]. Biological inference often needs prior identification of metabolites. Since this step is challenging, a novel approach, named Mummichog, has been proposed by Li et al. to reboot the conventional metabolomic workflow [93]. This method predicts biological activity directly from MS-based untargeted metabolomics data without a priori identification of metabolites. The idea behind this strategy is combining network analysis and metabolite prediction under the same computational framework, which significantly reduces the metabolomics workflow time. Based on spectral peaks, the computational prediction of metabolites yields several hits; thus, a "null" distribution can be estimated by how these predicted metabolites, retrieved from a metabolomics experiment, map to all known metabolite reactions through interrogating databases. Despite most annotations being false, the biological meaning underpinning the data drives enrichment of the metabolites. The metabolite enrichment pattern of real metabolites compared to the null distribution is then statistically assessed. This method has been elegantly illustrated in an exploration of innate immune cell activation, which revealed that glutathione metabolism is modified by viral infection driven by constitutive nitric oxide synthases [94]. Recently, Mummichog has been used for metabolic pathway analysis in a population by untargeted metabolomics. Hoffman et al. identified metabolic pathways linked to age, sex, and genotype, including glycerophospholipid, neurotransmitters, metabolism carnitine shuttle, and amino acid metabolism [95]. Tyrosine metabolism was found to be associated with nonalcoholic fatty liver [96]. Recently, Pirhaji et al. described a new network-based approach using a prize-winning Steiner forest algorithm for integrative analysis of untargeted metabolomics (PIUMet). This method infers molecular pathways via integrative analysis of metabolites without prior identification. Furthermore, PIUMet enabled elucidating putative identities of altered metabolites and inferring experimentally undetected, disease-associated metabolites and dysregulated proteins [97]. Compared to Mummichog, PIUMet also allows system-level inference by integrating other omics data.

Contextualization of metabolomics information is also important in pathophysiological investigations. From a metabolic network stand point, flux is defined as the rate (i.e., quantity per unit time) at which metabolites are converted or transported between different compartments [98]. Thus, metabolic fluxes, or the fluxome, represent a unique and functional readout of the phenotype. Indeed, the fluxome presents the metabolome in its contextual and multilevel functional aspects through tracking gene–environment interactions [98,99]. Thus, from a network view of metabolism, one or more metabolic fluxes could be altered in a given metabolic disorder depending on the complexity of the disease [100]. To interrogate these fluxes, fluxome network modeling can be achieved using constraints of mass and charge conservation along with stoichiometric and thermodynamic limitations [101-104]. Based on the stoichiometry of the reactants and products of biochemical reactions, flux balance analysis can estimate metabolic fluxes without knowledge about the kinetics of the participating enzymes [98,99]. Recently, Cortassa et al. suggested a new approach, distinct from flux balance analysis or metabolic flux analysis, that takes into account kinetic mechanisms and regulatory interactions [105]. A wide variety of software tools are available for analyzing metabolomic data at the pathway and network levels. Table 1 presents different functional analysis tools for both pathway analysis and visualization.

Table 1. Biological databases and functional analysis tools.

Tools	Websites	References
	Biological databases	
KEGG (Kyoto Encyclopedia of Genes and Genomes)	http://www.genome.jp/kegg	[106]
HumanCyc (Encylopedia of Human Metabolic Pathways)	http://humancyc.org	[107]
MetaCyc (Encyclopedia of Metabolic Pathways)	http://metacyc.org	[108]
Reactome (A Curated Knowledgebase of Pathways)	http://www.reactome.org	[109]
SMPDB (Small Molecule Pathway Database)	http://www.smpdb.ca	[110]
Virtual Metabolic Human Database	https://vmh.uni.lu	[79]
Wikipathways	http://www.wikipathways.org	[111]
	Pathway and networks analysis and visualization	
BioCyc—Omics Viewer	http://biocyc.org	[112]
iPath	http://pathways.embl.de	[113]
MetScape	http://metscape.ncibi.org	[114]
Paintomics	http://www.paintomics.org	[115]
Pathos	http://motif.gla.ac.uk/Pathos	[116]
Pathvisio	http://www.pathvisio.org	[117]
VANTED	http://vanted.ipk-gatersleben.de	[118]
IMPaLA	http://impala.molgen.mpg.de	[119]
MBROLE 2.0	http://csbg.cnb.csic.es/mbrole2	[120]
MPEA	http://ekhidna.biocenter.helsinki.fi/poxo/mpea	[121]
Mummichog	http://clinicalmetabolomics.org/init/default/software	[94]
PIUMet	http://fraenkel-nsf.csbi.mit.edu/PIUMet/	[97]
3Omics	http://3omics.cmdm.tw/	[122]
InCroMAP	http://www.ra.cs.uni-tuebingen.de/software/InCroMAP/	[123]
	Multifunctional tools	
MetaboAnlayst	http://www.metaboanalyst.com	[87]
XCMS online	https://xcmsonline.scripps.edu	[124]
MASSyPup	http://www.bioprocess.org/massypup	[125]
Workflow4Metabolomics	http://workflow4metabolomics.org	[126]
Metabox	https://github.com/kwanjeeraw/metabox	[127]

6. Challenges in Metabolomics Clinical Translation

6.1. Metabolite Identification

Metabolite identification remains a central issue in metabolomics prior to embracing complete basic and clinical translation. Despite spectral information becoming available in the literature or in spectral databases, metabolite identification is still a challenging task [128]. No software is currently available to automate the identification step. Often integration of multiplatform data (NMR and MS) is needed for the reliable characterization of metabolites. Furthermore, metabolite identification is mandatory for absolute quantitation especially in MS-based methods requiring the use of stable isotope-labelled internal standards. Some data-driven alternatives have been developed to elucidate metabolite structure associations such as correlation-based network and modularity analysis. The association structure can be used to identify MS ions derived from the same metabolite [129] or to identify biotransformations [130]. However, these knowledge-based approaches may be hampered by their limits for addressing the entire chemical space and limited coverage of metabolome databases. Another limitation lies in the cost for targeted analyses, which cannot reasonably be expected to support measurement of tens of thousands of chemicals in large populations. Thus, more efforts are needed to overcome this issue. However, in IEM, a few hundred key metabolites may be defined for large-scale screening.

6.2. Standardization

Standardized and validated protocols are a prerequisite for metabolic phenotyping technologies. Harmonization of the sample preparation, processing, analysis, and reporting, using validated and standardized protocols, is mandatory [131,132]. Despite substantial efforts to standardize metabolomics methods, there are still no universally adopted protocols, particularly for MS-based strategies. This situation is due to the diverse and ever-changing analytical platform. Standardized protocols are particularly helpful for untargeted metabolomics. In targeted methods, since each analyte is known and quantified, technology versatility is less important. The community and journals may take a lead in standardization by aligning it to community-published standards, such as the Metabolomics Standards Initiative [72], and data repisotories to encourage open metabolomic data, such as MetaboLights database at the EBI. All these endeavors aim to develop infrastructures and frameworks standardize terminology, data structure, and analytical workflows [133]. Finally, addressing these standardization issues is essential for regulatory compliance, which is a prerequisite for any clinical implementation.

6.3. Informatics and Automation

Automation at different stages, at instrument and pre- and post-analytic levels, is an important issue for broader use of metabolomics technologies. Automation enhances throughput, reproducibility, and reliability. Direct infusion MS-based methods are taking the lead from a translational perspective, such as the iKnife, which would allow real-time cancer diagnosis [56], and breathomics strategies for lung and respiratory diseases based on breath signatures [134]. Furthermore, metabolomics generates a huge amount of data that require comprehensive analysis and integration with other omics and metadata to infer the topology and dynamics of the underlying biological networks. Advanced statistical and computational tools along with effective data visualization are required to smoothly handle the diversity and quantity of the data and metabolite mapping [135,136]. In this regard, combining genomic and metabolic information may enhance biological inference and even clinical diagnostics [137,138]. Despite these promising steps, further advances in computational tools are needed for more efficient storage and integration [139]. Figure 3 shows how laboratory workflow using high-throughput analytical technologies, integrative bioinformatics and computational frameworks will reshape IMD investigations. This integrative approach will allow intelligible molecular and clinical information recovery for a more effective medical decision-making in inherited metabolic diseases.

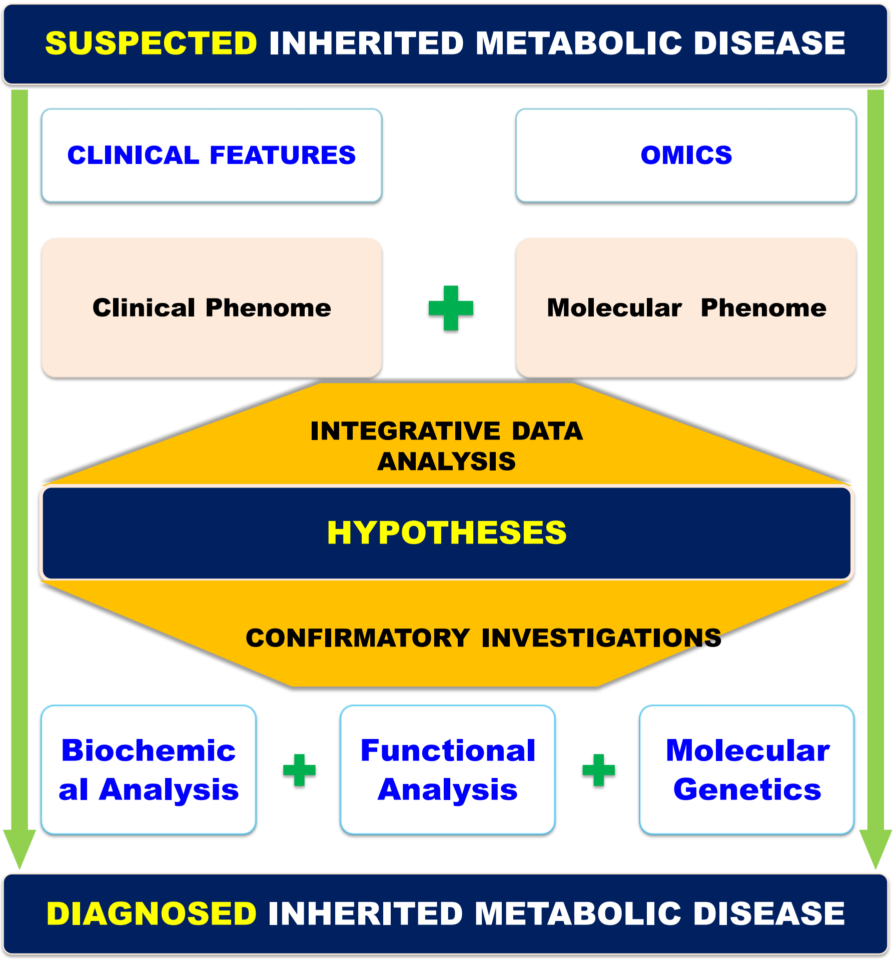

SUSPECTED INHERITED METABOLIC DISEASE
CLINICAL FEATURES
OMICS
Clinical Phenome
Molecular Phenome
INTEGRATIVE DATA ANALYSIS
HYPOTHESES
CONFIRMATORY INVESTIGATIONS
Biochemical Analysis
Functional Analysis
Molecular Genetics
DIAGNOSED INHERITED METABOLIC DISEASE

Figure 3. Paradigm shift in inherited metabolic diseases investigation. High-throughput analytical technologies, integrative bioinformatics and medical computational frameworks will allow intelligible molecular and clinical information recovery and effective medical decision-making.

7. Conclusion

Translating metabolomic data into actionable knowledge is the ultimate goal. Particular attention should be paid to computational tools for multidimensional data processing. There is an urgent need for more databases with validated and curated MRM transitions for targeted metabolites. Furthermore, for untargeted metabolomics, larger libraries and curated MS/MS spectra for metabolite identification are needed. Hybrid strategies including pathway and network analysis methods could enhance metabolite characterization through integration of different metrics, including data-driven network topology, chemical features correlation, omics data, and biological databases. Such multidimensional approaches may improve the chemical characterization by combining both extended chemical information and biological context. With all the high-dimensional data management issues, like other omics, metabolomics clinical implementation should be tackled using big data handling strategies for efficient storage, integration, visualization and sharing of metabolomics data. To achieve the promise of the Precision Medicine era, it is crucial to combine expertise from multiple disciplines, including clinicians, medical laboratory professionals, data scientists, computational biologists and biostatisticians. This raises the urgent need to think new teams with new skill sets and overlapping expertise for more effective medical interactions cross all healthcare partners for the management of IMD. Training next generation medical workforce to manage and interpret omics data is a way to go.

Acknowledgments: This work was supported by the Normandy University, the Institut National de la Santé et de la Recherche Médicale (INSERM), the Conseil Régional de Normandie, Labex SynOrg (ANR-11-LABX-0029), and the European Regional Development Fund (ERDF 31708).

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

REFERENCES

1. Hogeweg, P. The roots of bioinformatics in theoretical biology. *PLoS Comput Biol* **2011**, *7*, e1002021.
2. Luscombe, N.M.; Greenbaum, D.; Gerstein, M. What is bioinformatics? A proposed definition and overview of the field. *Methods of information in medicine* **2001**, *40*, 346-358.
3. Brereton, R.G. A short history of chemometrics: A personal view. *Journal of Chemometrics* **2014**, *28*, 749-760.
4. Cai, Y.; Weng, K.; Guo, Y.; Peng, J.; Zhu, Z.-J. An integrated targeted metabolomic platform for high-throughput metabolite profiling and automated data processing. *Metabolomics* **2015**, *11*, 1575-1586.
5. Tsugawa, H.; Ohta, E.; Izumi, Y.; Ogiwara, A.; Yukihira, D.; Bamba, T.; Fukusaki, E.; Arita, M. Mrm-diff: Data processing strategy for differential analysis in large scale mrm-based lipidomics studies. *Front Genet* **2014**, *5*, 471.
6. Tsugawa, H.; Arita, M.; Kanazawa, M.; Ogiwara, A.; Bamba, T.; Fukusaki, E. Mrmprobs: A data assessment and metabolite identification tool for large-scale multiple reaction monitoring based widely targeted metabolomics. *Analytical chemistry* **2013**, *85*, 5191-5199.
7. Vettukattil, R. Preprocessing of raw metabonomic data. In *Metabonomics: Methods and protocols*, Bjerrum, J.T., Ed. Springer New York: New York, NY, 2015; pp 123-136.
8. Yi, L.; Dong, N.; Yun, Y.; Deng, B.; Ren, D.; Liu, S.; Liang, Y. Chemometric methods in data processing of mass spectrometry-based metabolomics: A review. *Analytica Chimica Acta* **2016**, *914*, 17-34.
9. Szymanska, E.; Davies, A.; Buydens, L. Chemometrics for ion mobility spectrometry data: Recent advances and future prospects. *The Analyst* **2016**.
10. Johnsen, L.G.; Skou, P.B.; Khakimov, B.; Bro, R. Gas chromatography mass spectrometry data processing made easy. *Journal of Chromatography A*.
11. Smith, R.; Ventura, D.; Prince, J.T. Lc-ms alignment in theory and practice: A comprehensive algorithmic review. *Briefings in bioinformatics* **2013**, *16*, 104-117.

12. Smolinska, A.; Blanchet, L.; Buydens, L.M.; Wijmenga, S.S. Nmr and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. *Anal Chim Acta* **2012**, *750*, 82-97.
13. Rafiei, A.; Sleno, L. Comparison of peak-picking workflows for untargeted liquid chromatography/high-resolution mass spectrometry metabolomics data analysis. *Rapid Communications in Mass Spectrometry* **2015**, *29*, 119-127.
14. De Livera, A.M.; Dias, D.A.; De Souza, D.; Rupasinghe, T.; Pyke, J.; Tull, D.; Roessner, U.; McConville, M.; Speed, T.P. Normalizing and integrating metabolomics data. *Analytical chemistry* **2012**, *84*, 10768-10776.
15. Mak, T.D.; Laiakis, E.C.; Goudarzi, M.; Fornace, A.J. Selective paired ion contrast analysis: A novel algorithm for analyzing postprocessed lc-ms metabolomics data possessing high experimental noise. *Analytical chemistry* **2015**, *87*, 3177-3186.
16. Grun, D.; Kester, L.; van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat Meth* **2014**, *11*, 637-640.
17. Tebani, A.; Afonso, C.; Marret, S.; Bekri, S. Omics-based strategies in precision medicine: Toward a paradigm shift in inborn errors of metabolism investigations. *Int J Mol Sci* **2016**, *17*.
18. Li, B.; Tang, J.; Yang, Q.; Cui, X.; Li, S.; Chen, S.; Cao, Q.; Xue, W.; Chen, N.; Zhu, F. Performance evaluation and online realization of data-driven normalization methods used in lc/ms based untargeted metabolomics analysis. *Sci Rep* **2016**, *6*, 38881.
19. Scholz, M.; Gatzek, S.; Sterling, A.; Fiehn, O.; Selbig, J. Metabolite fingerprinting: Detecting biological features by independent component analysis. *Bioinformatics* **2004**, *20*, 2447-2454.
20. Wang, W.X.; Zhou, H.H.; Lin, H.; Roy, S.; Shaler, T.A.; Hill, L.R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C.H. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal Chem* **2003**, *75*, 4818-4826.
21. Lee, J.; Park, J.; Lim, M.S.; Seong, S.J.; Seo, J.J.; Park, S.M.; Lee, H.W.; Yoon, Y.R. Quantile normalization approach for liquid chromatography-mass spectrometry-based metabolomic data from healthy human volunteers. *Analytical sciences : the international journal of the Japan Society for Analytical Chemistry* **2012**, *28*, 801-805.
22. Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1h nmr metabonomics. *Analytical chemistry* **2006**, *78*, 4281-4290.
23. Bolstad, B.M.; Irizarry, R.A.; Astrand, M.; Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)* **2003**, *19*, 185-193.
24. Li, C.; Wong, W.H. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America* **2001**, *98*, 31-36.
25. Astrand, M. Contrast normalization of oligonucleotide arrays. *J Comput Biol* **2003**, *10*, 95-102.
26. Workman, C.; Jensen, L.J.; Jarmer, H.; Berka, R.; Gautier, L.; Nielser, H.B.; Saxild, H.-H.; Nielsen, C.; Brunak, S.; Knudsen, S. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology* **2002**, *3*, research0048.0041-research0048.0016.
27. Dudoit, S.; Yang, Y.H.; Callow, M.J.; Speed, T.P. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica sinica* **2002**, 111-139.
28. Kirwan, J.; Broadhurst, D.; Davidson, R.; Viant, M. Characterising and correcting batch variation in an automated direct infusion mass spectrometry (dims) metabolomics workflow. *Analytical and bioanalytical chemistry* **2013**, *405*, 5147-5157.
29. Warrack, B.M.; Hnatyshyn, S.; Ott, K.-H.; Reily, M.D.; Sanders, M.; Zhang, H.; Drexler, D.M. Normalization strategies for metabonomic analysis of urine samples. *Journal of Chromatography B* **2009**, *877*, 547-552.
30. Shen, X.; Gong, X.; Cai, Y.; Guo, Y.; Tu, J.; Li, H.; Zhang, T.; Wang, J.; Xue, F.; Zhu, Z.-J. Normalization and integration of large-scale metabolomics data using support vector regression. *Metabolomics* **2016**, *12*, 89.
31. Hermansson, M.; Uphoff, A.; Kakela, R.; Somerharju, P. Automated quantitative analysis of complex lipidomes by liquid chromatography/mass spectrometry. *Anal Chem* **2005**, *77*, 2166-2175.
32. Bijlsma, S.; Bobeldijk, I.; Verheij, E.R.; Ramaker, R.; Kochhar, S.; Macdonald, I.A.; van Ommen, B.; Smilde, A.K. Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. *Anal Chem* **2006**, *78*, 567-574.
33. Sysi-Aho, M.; Katajamaa, M.; Yetukuri, L.; Oresic, M. Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC bioinformatics* **2007**, *8*, 93.
34. Wu, Y.; Li, L. Sample normalization methods in quantitative metabolomics. *Journal of Chromatography A* **2016**, *1430*, 80-95.
35. van den Berg, R.A.; Hoefsloot, H.C.; Westerhuis, J.A.; Smilde, A.K.; van der Werf, M.J. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC genomics* **2006**, *7*, 142.
36. Broadhurst, D.I.; Kell, D.B. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* **2006**, *2*, 171-196.
37. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **1995**, *57*, 289-300.
38. Hotelling, H. *Analysis of a complex of statistical variables into principal components*. Warwick & York: 1933.
39. Bouveresse, D.J.-R.; Rutledge, D. Independent components analysis: Theory and applications. *Resolving Spectral Mixtures: With Applications from Ultrafast Time-Resolved Spectroscopy to Super-Resolution Imaging* **2016**, *30*, 7225.
40. Wang, G.; Ding, Q.; Hou, Z. Independent component analysis and its applications in signal processing for analytical chemistry. *TrAC Trends in Analytical Chemistry* **2008**, *27*, 368-376.
41. Al-Saegh, A. Independent component analysis for separation of speech mixtures: A comparison among thirty algorithms. *Iraqi Journal for Electrical & Electronic Engineering* **2015**, *11*.
42. Liu, Y.; Smirnov, K.; Lucio, M.; Gougeon, R.D.; Alexandre, H.; Schmitt-Kopplin, P. Metica: Independent component analysis for high-resolution mass-spectrometry based non-targeted metabolomics. *BMC bioinformatics* **2016**, *17*, 1-14.
43. Li, X.; Hansen, J.; Zhao, X.; Lu, X.; Weigert, C.; Häring, H.-U.; Pedersen, B.K.; Plomgaard, P.; Lehmann, R.; Xu, G. Independent component analysis in non-hypothesis driven metabolomics: Improvement of pattern discovery and simplification of biological data interpretation demonstrated with plasma samples of exercising humans. *Journal of Chromatography B* **2012**, *910*, 156-162.
44. Monakhova, Y.B.; Godelmann, R.; Kuballa, T.; Mushtakova, S.P.; Rutledge, D.N. Independent components analysis to increase efficiency of discriminant analysis methods (fda and lda): Application to nmr fingerprinting of wine. *Talanta* **2015**, *141*, 60-65.
45. Wiwie, C.; Baumbach, J.; Rottger, R. Comparing the performance of biomedical clustering methods. *Nature methods* **2015**, *12*, 1033-1038.
46. Hartigan, J.A.; Wong, M.A. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **1979**, *28*, 100-108.

47. Johnson, S.C. Hierarchical clustering schemes. *Psychometrika* **1967**, *32*, 241-254.
48. Kohonen, T. The self-organizing map. *Proceedings of the IEEE* **1990**, *78*, 1464-1480.
49. Goodwin, C.R.; Sherrod, S.D.; Marasco, C.C.; Bachmann, B.O.; Schramm-Sapyta, N.; Wikswo, J.P.; McLean, J.A. Phenotypic mapping of metabolic profiles using self-organizing maps of high-dimensional mass spectrometry data. *Analytical chemistry* **2014**, *86*, 6563-6571.
50. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics (Oxford, England)* **2007**, *23*, 2507-2517.
51. Yun, Y.-H.; Deng, B.-C.; Cao, D.-S.; Wang, W.-T.; Liang, Y.-Z. Variable importance analysis based on rank aggregation with applications in metabolomics for biomarker discovery. *Analytica Chimica Acta* **2016**, *911*, 27-34.
52. Rinaudo, P.; Boudah, S.; Junot, C.; Thévenot, E.A. Biosigner: A new method for the discovery of significant molecular signatures from omics data. *Frontiers in Molecular Biosciences* **2016**, *3*.
53. Engel, J.; Blanchet, L.; Engelke, U.; Wevers, R.; Buydens, L. Sparse statistical health monitoring: A novel variable selection approach to diagnosis and follow-up of individual patients. *Chemometrics and Intelligent Laboratory Systems* **2017**.
54. Westad, F.; Marini, F. Validation of chemometric models – a tutorial. *Analytica Chimica Acta* **2015**, *893*, 14-24.
55. Ouyang, M.; Zhang, Z.; Chen, C.; Liu, X.; Liang, Y. Application of sparse linear discriminant analysis for metabolomics data. *Analytical Methods* **2014**, *6*, 9037-9044.
56. Balog, J.; Sasi-Szabo, L.; Kinross, J.; Lewis, M.R.; Muirhead, L.J.; Veselkov, K.; Mirnezami, R.; Dezso, B.; Damjanovich, L.; Darzi, A., *et al.* Intraoperative tissue identification using rapid evaporative ionization mass spectrometry. *Sci. Transl. Med.* **2013**, *5*, 11.
57. Wold, S.; Sjöström, M.; Eriksson, L. Pls-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **2001**, *58*, 109-130.
58. Trygg, J.; Wold, S. Orthogonal projections to latent structures (o-pls). *Journal of Chemometrics* **2002**, *16*, 119-128.
59. Manwaring, V.; Boutin, M.; Auray-Blais, C. A metabolomic study to identify new globotriaosylceramide-related biomarkers in the plasma of fabry disease patients. *Analytical chemistry* **2013**, *85*, 9039-9048.
60. Cortes, C.; Vapnik, V. Support-vector networks. *Machine Learning* **1995**, *20*, 273-297.
61. Lin, X.; Wang, Q.; Yin, P.; Tang, L.; Tan, Y.; Li, H.; Yan, K.; Xu, G. A method for handling metabonomics data from liquid chromatography/mass spectrometry: Combinational use of support vector machine recursive feature elimination, genetic algorithm and random forest for feature selection. *Metabolomics* **2011**, *7*, 549-558.
62. Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5-32.
63. Huang, J.-H.; Fu, L.; Li, B.; Xie, H.-L.; Zhang, X.; Chen, Y.; Qin, Y.; Wang, Y.; Zhang, S.; Huang, H., *et al.* Distinguishing the serum metabolite profiles differences in breast cancer by gas chromatography mass spectrometry and random forest method. *RSC Advances* **2015**, *5*, 58952-58958.
64. Habchi, B.; Alves, S.; Jouan-Rimbaud Bouveresse, D.; Moslah, B.; Paris, A.; Lécluse, Y.; Gauduchon, P.; Lebailly, P.; Rutledge, D.N.; Rathahao-Paris, E. An innovative chemometric method for processing direct introduction high resolution mass spectrometry metabolomic data: Independent component–discriminant analysis (ic–da). *Metabolomics* **2017**, *13*, 45.
65. Offroy, M.; Duponchel, L. Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry. *Analytica Chimica Acta* **2016**, *910*, 1-11.
66. Liu, W.; Bai, X.; Liu, Y.; Wang, W.; Han, J.; Wang, Q.; Xu, Y.; Zhang, C.; Zhang, S.; Li, X., *et al.* Topologically inferring pathway activity toward precise cancer classification via integrating genomic and metabolomic data: Prostate cancer as a case. *Scientific Reports* **2015**, *5*, 13192.
67. Engel, J.; Blanchet, L.; Engelke, U.F.; Wevers, R.A.; Buydens, L.M. Towards the disease biomarker in an individual patient using statistical health monitoring. *PloS one* **2014**, *9*, e92452.
68. Ren, S.; Hinzman, A.; Kang, E.; Szczesniak, R.; Lu, L. Computational and statistical analysis of metabolomics data. *Metabolomics* **2015**, *11*, 1492-1513.
69. Gromski, P.S.; Muhamadali, H.; Ellis, D.I.; Xu, Y.; Correa, E.; Turner, M.L.; Goodacre, R. A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta* **2015**, *879*, 10-23.
70. Misra, B.B.; van der Hooft, J.J. Updates in metabolomics tools and resources: 2014-2015. *Electrophoresis* **2016**, *37*, 86-110.
71. Cui, Q.; Lewis, I.A.; Hegeman, A.D.; Anderson, M.E.; Li, J.; Schulte, C.F.; Westler, W.M.; Eghbalnia, H.R.; Sussman, M.R.; Markley, J.L. Metabolite identification via the madison metabolomics consortium database. *Nature biotechnology* **2008**, *26*, 162-164.
72. Sumner, L.W.; Amberg, A.; Barrett, D.; Beale, M.H.; Beger, R.; Daykin, C.A.; Fan, T.W.M.; Fiehn, O.; Goodacre, R.; Griffin, J.L., *et al.* Proposed minimum reporting standards for chemical analysis. *Metabolomics : Official journal of the Metabolomic Society* **2007**, *3*, 211-221.
73. 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.
74. Silva, R.R.; Jourdan, F.; Salvanha, D.M.; Letisse, F.; Jamin, E.L.; Guidetti-Gonzalez, S.; Labate, C.A.; Vencio, R.Z. Probmetab: An r package for bayesian probabilistic annotation of lc-ms-based metabolomics. *Bioinformatics (Oxford, England)* **2014**, *30*, 1336.
75. 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 (Oxford, England)* **2011**, *27*, 1339.
76. Daly, R.; Rogers, S.; Wandy, J.; Jankevics, A.; Burgess, K.E.; Breitling, R. Metassign: Probabilistic annotation of metabolites from lc-ms data using a bayesian clustering approach. *Bioinformatics (Oxford, England)* **2014**, *30*, 2764.
77. Barabasi, A.-L.; Gulbahce, N.; Loscalzo, J. Network medicine: A network-based approach to human disease. *Nat Rev Genet* **2011**, *12*, 56-68.
78. Kitano, H. Computational systems biology. *Nature* **2002**, *420*, 206-210.
79. Thiele, I.; Swainston, N.; Fleming, R.M.; Hoppe, A.; Sahoo, S.; Aurich, M.K.; Haraldsdottir, H.; Mo, M.L.; Rolfsson, O.; Stobbe, M.D., *et al.* A community-driven global reconstruction of human metabolism. *Nat Biotechnol* **2013**, *31*, 419-425.
80. Cazzaniga, P.; Damiani, C.; Besozzi, D.; Colombo, R.; Nobile, M.S.; Gaglio, D.; Pescini, D.; Molinari, S.; Mauri, G.; Alberghina, L., *et al.* Computational strategies for a system-level understanding of metabolism. *Metabolites* **2014**, *4*, 1034-1087.
81. Garcia-Campos, M.A.; Espinal-Enriquez, J.; Hernandez-Lemus, E. Pathway analysis: State of the art. *Front Physiol* **2015**, *6*, 383.

82. Khatri, P.; Sirota, M.; Butte, A.J. Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput Biol* **2012**, *8*, e1002375.
83. Xia, J.; Wishart, D.S. Msea: A web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic acids research* **2010**, *38*, W71-W77.
84. Adjaye, J.; Huntriss, J.; Herwig, R.; BenKahla, A.; Brink, T.C.; Wierling, C.; Hultschig, C.; Groth, D.; Yaspo, M.L.; Picton, H.M., *et al.* Primary differentiation in the human blastocyst: Comparative molecular portraits of inner cell mass and trophectoderm cells. *Stem cells (Dayton, Ohio)* **2005**, *23*, 1514-1525.
85. Goeman, J.J.; van de Geer, S.A.; de Kort, F.; van Houwelingen, H.C. A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics (Oxford, England)* **2004**, *20*, 93-99.
86. Hummel, M.; Meister, R.; Mansmann, U. Globalancova: Exploration and assessment of gene group effects. *Bioinformatics (Oxford, England)* **2008**, *24*, 78-85.
87. Xia, J.; Sinelnikov, I.V.; Han, B.; Wishart, D.S. Metaboanalyst 3.0--making metabolomics more meaningful. *Nucleic acids research* **2015**, *43*, W251-257.
88. Krumsiek, J.; Suhre, K.; Illig, T.; Adamski, J.; Theis, F.J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC systems biology* **2011**, *5*, 21.
89. Do, K.T.; Kastenmüller, G.; Mook-Kanamori, D.O.; Yousri, N.A.; Theis, F.J.; Suhre, K.; Krumsiek, J. Network-based approach for analyzing intra- and interfluid metabolite associations in human blood, urine, and saliva. *Journal of Proteome Research* **2015**, *14*, 1183-1194.
90. Valcarcel, B.; Wurtz, P.; Seich al Basatena, N.K.; Tukiainen, T.; Kangas, A.J.; Soininen, P.; Jarvelin, M.R.; Ala-Korpela, M.; Ebbels, T.M.; de Iorio, M. A differential network approach to exploring differences between biological states: An application to prediabetes. *PLoS One* **2011**, *6*, e24702.
91. Bartel, J.; Krumsiek, J.; Schramm, K.; Adamski, J.; Gieger, C.; Herder, C.; Carstensen, M.; Peters, A.; Rathmann, W.; Roden, M., *et al.* The human blood metabolome-transcriptome interface. *PLoS genetics* **2015**, *11*, e1005274.
92. Shin, S.Y.; Fauman, E.B.; Petersen, A.K.; Krumsiek, J.; Santos, R.; Huang, J.; Arnold, M.; Erte, I.; Forgetta, V.; Yang, T.P., *et al.* An atlas of genetic influences on human blood metabolites. *Nature genetics* **2014**, *46*, 543-550.
93. Li, S.; Park, Y.; Duraisingham, S.; Strobel, F.H.; Khan, N.; Soltow, Q.A.; Jones, D.P.; Pulendran, B. Predicting network activity from high throughput metabolomics. *PLoS Comput. Biol.* **2013**, *9*, e1003123.
94. Li, S.; Park, Y.; Duraisingham, S.; Strobel, F.H.; Khan, N.; Soltow, Q.A.; Jones, D.P.; Pulendran, B. Predicting network activity from high throughput metabolomics. *PLoS Comput Biol* **2013**, *9*, e1003123.
95. Hoffman, J.M.; Tran, V.; Wachtman, L.M.; Green, C.L.; Jones, D.P.; Promislow, D.E. A longitudinal analysis of the effects of age on the blood plasma metabolome in the common marmoset, callithrix jacchus. *Experimental gerontology* **2016**, *76*, 17-24.
96. Jin, R.; Banton, S.; Tran, V.T.; Konomi, J.V.; Li, S.; Jones, D.P.; Vos, M.B. Amino acid metabolism is altered in adolescents with nonalcoholic fatty liver disease-an untargeted, high resolution metabolomics study. *J Pediatr* **2016**, *172*, 14-19.e15.
97. Pirhaji, L.; Milani, P.; Leidl, M.; Curran, T.; Avila-Pacheco, J.; Clish, C.B.; White, F.M.; Saghatelian, A.; Fraenkel, E. Revealing disease-associated pathways by network integration of untargeted metabolomics. *Nat Meth* **2016**, *13*, 770-776.
98. Aon, M.A.; Cortassa, S. Systems biology of the fluxome. *Processes* **2015**, *3*, 607-618.
99. Cascante, M.; Marin, S. Metabolomics and fluxomics approaches. *Essays in biochemistry* **2008**, *45*, 67-82.
100. Lanpher, B.; Brunetti-Pierri, N.; Lee, B. Inborn errors of metabolism: The flux from mendelian to complex diseases. *Nat Rev Genet* **2006**, *7*, 449-460.
101. Cortassa, S.; Aon, M.A. Computational modeling of mitochondrial function. *Methods in molecular biology (Clifton, N.J.)* **2012**, *810*, 311-326.
102. Kell, D.B.; Goodacre, R. Metabolomics and systems pharmacology: Why and how to model the human metabolic network for drug discovery. *Drug discovery today* **2014**, *19*, 171-182.
103. Winter, G.; Kromer, J.O. Fluxomics - connecting 'omics analysis and phenotypes. *Environmental microbiology* **2013**, *15*, 1901-1916.
104. Aurich, M.K.; Thiele, I. Computational modeling of human metabolism and its application to systems biomedicine. *Methods in molecular biology (Clifton, N.J.)* **2016**, *1386*, 253-281.
105. Cortassa, S.; Caceres, V.; Bell, L.N.; O'Rourke, B.; Paolocci, N.; Aon, M.A. From metabolomics to fluxomics: A computational procedure to translate metabolite profiles into metabolic fluxes. *Biophysical journal* **2015**, *108*, 163-172.
106. Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. Kegg as a reference resource for gene and protein annotation. *Nucleic acids research* **2016**, *44*, D457-462.
107. Romero, P.; Wagg, J.; Green, M.L.; Kaiser, D.; Krummenacker, M.; Karp, P.D. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology* **2005**, *6*, R2-R2.
108. Caspi, R.; Foerster, H.; Fulcher, C.A.; Kaipa, P.; Krummenacker, M.; Latendresse, M.; Paley, S.; Rhee, S.Y.; Shearer, A.G.; Tissier, C., *et al.* The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research* **2008**, *36*, D623-631.
109. Vastrik, I.; D'Eustachio, P.; Schmidt, E.; Gopinath, G.; Croft, D.; de Bono, B.; Gillespie, M.; Jassal, B.; Lewis, S.; Matthews, L., *et al.* Reactome: A knowledge base of biologic pathways and processes. *Genome Biol* **2007**, *8*, R39.
110. Jewison, T.; Su, Y.; Disfany, F.M.; Liang, Y.; Knox, C.; Maciejewski, A.; Poelzer, J.; Huynh, J.; Zhou, Y.; Arndt, D., *et al.* Smpdb 2.0: Big improvements to the small molecule pathway database. *Nucleic acids research* **2014**, *42*, D478-484.
111. Kelder, T.; van Iersel, M.P.; Hanspers, K.; Kutmon, M.; Conklin, B.R.; Evelo, C.T.; Pico, A.R. Wikipathways: Building research communities on biological pathways. *Nucleic acids research* **2012**, *40*, D1301-1307.
112. Caspi, R.; Billington, R.; Ferrer, L.; Foerster, H.; Fulcher, C.A.; Keseler, I.M.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Mueller, L.A., *et al.* The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research* **2016**, *44*, D471-480.
113. Yamada, T.; Letunic, I.; Okuda, S.; Kanehisa, M.; Bork, P. Ipath2.0: Interactive pathway explorer. *Nucleic acids research* **2011**, *39*, W412-415.
114. Karnovsky, A.; Weymouth, T.; Hull, T.; Tarcea, V.G.; Scardoni, G.; Laudanna, C.; Sartor, M.A.; Stringer, K.A.; Jagadish, H.V.; Burant, C., *et al.* Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics (Oxford, England)* **2012**, *28*, 373-380.
115. 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 (Oxford, England)* **2011**, *27*, 137-139.

116. Leader, D.P.; Burgess, K.; Creek, D.; Barrett, M.P. Pathos: A web facility that uses metabolic maps to display experimental changes in metabolites identified by mass spectrometry. *Rapid communications in mass spectrometry : RCM* **2011**, *25*, 3422-3426.
117. Kutmon, M.; van Iersel, M.P.; Bohler, A.; Kelder, T.; Nunes, N.; Pico, A.R.; Evelo, C.T. Pathvisio 3: An extendable pathway analysis toolbox. *PLoS Comput Biol* **2015**, *11*, e1004085.
118. Rohn, H.; Junker, A.; Hartmann, A.; Grafahrend-Belau, E.; Treutler, H.; Klapperstück, M.; Czauderna, T.; Klukas, C.; Schreiber, F. Vanted v2: A framework for systems biology applications. *BMC systems biology* **2012**, *6*, 1-13.
119. Kamburov, A.; Cavill, R.; Ebbels, T.M.D.; Herwig, R.; Keun, H.C. Integrated pathway-level analysis of transcriptomics and metabolomics data with impala. *Bioinformatics (Oxford, England)* **2011**, *27*, 2917-2918.
120. Lopez-Ibanez, J.; Pazos, F.; Chagoyen, M. Mbrole 2.0-functional enrichment of chemical compounds. *Nucleic acids research* **2016**, *44*, W201-204.
121. Kankainen, M.; Gopalacharyulu, P.; Holm, L.; Oresic, M. Mpea--metabolite pathway enrichment analysis. *Bioinformatics (Oxford, England)* **2011**, *27*, 1878-1879.
122. Kuo, T.-C.; Tian, T.-F.; Tseng, Y.J. 3omics: A web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Systems Biology* **2013**, *7*, 64.
123. Wrzodek, C.; Eichner, J.; Büchel, F.; Zell, A. Incromap: Integrated analysis of cross-platform microarray and pathway data. *Bioinformatics (Oxford, England)* **2013**, *29*, 506-508.
124. Tautenhahn, R.; Patti, G.J.; Rinehart, D.; Siuzdak, G. Xcms online: A web-based platform to process untargeted metabolomic data. *Analytical chemistry* **2012**, *84*, 5035-5039.
125. Winkler, R. An evolving computational platform for biological mass spectrometry: Workflows, statistics and data mining with massypup64. *PeerJ* **2015**, *3*, e1401.
126. Giacomoni, F.; Le Corguille, G.; Monsoor, M.; Landi, M.; Pericard, P.; Petera, M.; Duperier, C.; Tremblay-Franco, M.; Martin, J.F.; Jacob, D., *et al.* Workflow4metabolomics: A collaborative research infrastructure for computational metabolomics. *Bioinformatics (Oxford, England)* **2015**, *31*, 1493-1495.
127. Wanichthanarak, K.; Fan, S.; Grapov, D.; Barupal, D.K.; Fiehn, O. Metabox: A toolbox for metabolomic data analysis, interpretation and integrative exploration. *PloS one* **2017**, *12*, e0171046.
128. Goodacre, R.; Broadhurst, D.; Smilde, A.K.; Kristal, B.S.; Baker, J.D.; Beger, R.; Bessant, C.; Connor, S.; Capuani, G.; Craig, A. Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics* **2007**, *3*, 231-241.
129. Broeckling, C.D.; Afsar, F.A.; Neumann, S.; Ben-Hur, A.; Prenni, J.E. Ramclust: A novel feature clustering method enables spectral-matching-based annotation for metabolomics data. *Analytical chemistry* **2014**, *86*, 6812-6817.
130. Kind, T.; Fiehn, O. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical reviews* **2010**, *2*, 23-60.
131. Chitayat, S.; Rudan, J.F. Chapter 10 - phenome centers and global harmonization. In *Metabolic phenotyping in personalized and public healthcare*, Academic Press: Boston, 2016; pp 291-315.
132. Kohler, I.; Verhoeven, A.; Derks, R.J.; Giera, M. Analytical pitfalls and challenges in clinical metabolomics. *Bioanalysis* **2016**, *8*, 1509-1532.
133. Levin, N.; Salek, R.M.; Steinbeck, C. Chapter 11 - from databases to big data. In *Metabolic phenotyping in personalized and public healthcare*, Academic Press: Boston, 2016; pp 317-331.
134. Hauschild, A.C.; Frisch, T.; Baumbach, J.I.; Baumbach, J. Carotta: Revealing hidden confounder markers in metabolic breath profiles. *Metabolites* **2015**, *5*, 344-363.
135. Alyass, A.; Turcotte, M.; Meyre, D. From big data analysis to personalized medicine for all: Challenges and opportunities. *BMC Medical Genomics* **2015**, *8*, 1-12.
136. Ritchie, M.D.; Holzinger, E.R.; Li, R.; Pendergrass, S.A.; Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet* **2015**, *16*, 85-97.
137. Tarailo-Graovac, M.; Shyr, C.; Ross, C.J.; Horvath, G.A.; Salvarinova, R.; Ye, X.C.; Zhang, L.H.; Bhavsar, A.P.; Lee, J.J.; Drogemoller, B.I., *et al.* Exome sequencing and the management of neurometabolic disorders. *The New England journal of medicine* **2016**, *374*, 2246-2255.
138. van Karnebeek, C.D.; Bonafé, L.; Wen, X.-Y.; Tarailo-Graovac, M.; Balzano, S.; Royer-Bertrand, B.; Ashikov, A.; Garavelli, L.; Mammi, I.; Turolla, L. Nans-mediated synthesis of sialic acid is required for brain and skeletal development. *Nature genetics* **2016**.
139. Perez-Riverol, Y.; Bai, M.; da Veiga Leprevost, F.; Squizzato, S.; Park, Y.M.; Haug, K.; Carroll, A.J.; Spalding, D.; Paschall, J.; Wang, M., *et al.* Discovering and linking public omics data sets using the omics discovery index. *Nat Biotech* **2017**, *35*, 406-409.

CHAPITRE VI : POTENTIEL DE LA METABOLOMIQUE CLINIQUE (ARTICLE IV)

Les progrès significatifs constatés dans les sciences séparatives, la spectrométrie de masse et la spectroscopie par résonance magnétique nucléaire ont conforté les bases analytiques la caractérisation et la mesure des métabolites dans les échantillons biologiques. Il est largement admis que la surveillance globale des processus métaboliques est nécessaire pour une compréhension fondamentale de la genèse et de la progression des maladies, en particulier pour les EIM. Les stratégies basées sur la métabolomique dans les études cliniques modernes ont permis une meilleure compréhension des conditions pathologiques et des mécanismes pathogéniques. Ceci a permis l'émergence d'outils innovants pour le diagnostic et de pronostic des maladies. Le chapitre VI est présenté sous forme d'article décrivant l'état de l'art de la métabolomique dans le contexte clinique en général, et dans l'exploration des EIM en particulier. Par ailleurs, les avancées ainsi que les défis inhérents aux exigences cliniques sont présentés.

Article IV : Tebani A, Abily-Donval L, Afonso C, Marret S, Bekri S. Clinical Metabolomics: The New Metabolic Window for Inborn Errors of Metabolism Investigations in the Post-Genomic Era. Int J Mol Sci. 2016 Jul 20;17(7)

International Journal of
Molecular Sciences

MDPI

Review

Clinical Metabolomics: The New Metabolic Window for Inborn Errors of Metabolism Investigations in the Post-Genomic Era

Abdellah Tebani [1,2,3], Lenaig Abily-Donval [2,4], Carlos Afonso [3], Stéphane Marret [2,4] and Soumeya Bekri [1,2,*]

[1] Department of Metabolic Biochemistry, Rouen University Hospital, Rouen 76031, France; abdellah.tebani@chu-rouen.fr

[2] Normandie Univ, UNIROUEN, INSERM, CHU Rouen, IRIB, Laboratoire NeoVasc ERI28, Rouen 76000, France; lenaig.abily-donval@chu-rouen.fr (L.A.-D.); stephane.marret@chu-rouen.fr (S.M.)

[3] Normandie Univ, UNIROUEN, INSA Rouen, CNRS, COBRA, Rouen 76000, France; carlos.afonso@univ-rouen.fr

[4] Department of Neonatal Pediatrics and Intensive Care, Rouen University Hospital, Rouen 76031, France

[*] Correspondence: soumeya.bekri@chu-rouen.fr; Tel.: +33-2-3288-8124; Fax: +33-2-3288-8341

Academic Editor: Ting-Li (Morgan) Han
Received: 11 June 2016; Accepted: 15 July 2016; Published: 20 July 2016

Abstract: Inborn errors of metabolism (IEM) represent a group of about 500 rare genetic diseases with an overall estimated incidence of 1/2500. The diversity of metabolic pathways involved explains the difficulties in establishing their diagnosis. However, early diagnosis is usually mandatory for successful treatment. Given the considerable clinical overlap between some inborn errors, biochemical and molecular tests are crucial in making a diagnosis. Conventional biological diagnosis procedures are based on a time-consuming series of sequential and segmented biochemical tests. The rise of "omic" technologies offers holistic views of the basic molecules that build a biological system at different levels. Metabolomics is the most recent "omic" technology based on biochemical characterization of metabolites and their changes related to genetic and environmental factors. This review addresses the principles underlying metabolomics technologies that allow them to comprehensively assess an individual biochemical profile and their reported applications for IEM investigations in the precision medicine era.

Keywords: metabolomics; inborn errors of metabolism; screening; diagnosis; systems medicine; precision medicine

1. Introduction

The new field of precision medicine is revolutionizing current medical practice and reshaping future medicine. Precision medicine aspires to put the patient as the central driver of healthcare by broadening biological knowledge and acknowledging the greate diversity of individuals [1]. It is well established that complex gene–environment interactions shape normal physiological and disease processes at both the individual and population scale. Predicting normal and pathological states in patients requires dynamic and systematic understanding of these interactions. Systems medicine is a new concept based on holistic approaches for disease diagnosis and monitoring. The basic idea of these approaches is that a complex system is more comprehensively understood if considered as a whole at both the spatial and temporal scales.

Inborn errors of metabolism (IEM) are an appropriate model for systems medicine studies because the biological basis underlying these diseases has been, at least partly, revealed. IEM represent a group of about 500 rare genetic diseases with an overall estimated incidence of 1/2500. Even though these disorders are individually rare, they are collectively more common and cause a significant childhood morbidity and mortality. IEM are genetic disorders resulting from defects in a given biochemical pathway due to the deficiency or abnormality of an enzyme, its cofactor, or a transporter, leading to an accumulation of a substrate or lack of the product. Hence, the diversity of metabolic pathways involved explains the difficulties in establishing a diagnosis.

Autosomal recessive transmission is most frequent, but autosomal dominant and X-linked disorders have also been described. IEM may involve mutations in mitochondrial DNA. The pathogenesis of IEM can be explained by mechanisms such as deficiency of an essential product or enzyme, systemic toxic effects of circulating metabolites, and activation or inhibition of alternative metabolism [2]. Based on these pathophysiological traits, several IEM therapies have been developed, including dietary restriction, toxic product clearance, or biotherapies (enzyme replacement and gene therapy) [3]. Initiating these treatments at birth or at early stages is usually mandatory for optimal patient management. The first description of these disorders was made by Sir Archibald Garrod [4], who initiated the "one gene–one disease" paradigm. However, there is a lack of genotype–phenotype correlation in IEM. Furthermore, for the same genetic variation, different phenotypes have been observed in the same family [2]. These observations challenge Garrod's paradigm and suggest the influence of either genetic or environmental modifying factors. Thus, IEM are more than monogenic diseases, which adds another layer of complexity to disease characterization and diagnosis.

The rise of "omic" approaches, enabled by the tremendous technological shift in both multiscale biological information capture and data management, offers an amazing opportunity to provide new effective tools for screening, diagnosis, treatment, and monitoring of these diseases. Omic technologies offer global views on the basic molecules that build a biological system at the cell, tissue, or organism level. Primarily, they aim to recover, in an untargeted, unbiased, and hypothesis-free fashion, the biological information carried by genes (genomics), mRNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics). These holistic strategies clearly contrast with conventional studies, which are mainly hypothesis-driven and reductionist. To truly understand disease processes, a global investigative approach needs to be applied at multiple biological informational levels.

Since the early days of medicine, the human body is viewed as a collection of separate and independent components, and thus, physicians typically treated disease by trying to identify the single abnormality related to a single component. This approach lacks contextual information which is vital for mechanistic understanding of pathophysiology and, thus, for designing treatment strategies [5,6]. Indeed, the complete characterization of a biological system should include a structural, an organizational pattern and a functional description [7]. The structure comprises the fundamental actor components (genes, proteins and metabolites). The organization pattern denotes how these actors are linked to each other and how they are organized topologically (e.g., linear or branched sequence of reactions) and morphologically (membrane-bound or functional compartmentalization). The function describes how the whole system behaves in space and time with regard to metabolic fluxes and response to stimuli [8–10].

Systems biology is a new scientific field that tries to achieve this systemic understanding of biology and to fill in the gap between information and context from a biological standpoint. Systems biology can be defined as a holistic and systemic analysis of complex system inter-connections and their functional interrelationships [11–14]. Two vital pillars supported the emergence of systems biology: data generation and data modeling. On the one hand, the surge of high-throughput omics technologies allowed the retrieval of a global and comprehensive biological information. On the other hand, the amazing development of computational capabilities allowed complicated systems modeling and convenient and intuitive visualization. Furthermore, these informatics advancements are crucial for comprehensive integration and insightful interpretation of the complex biological information [15–17].

The patient-centric approach is essential to achieve the promise of personal and stratified medicine. Indeed, unlike conventional medical biology practice based primarily on sequential studies of genes, proteins, and metabolites, the great challenge of modern biology is to apprehend a disease as a complex, integrated, and dynamic network. The dynamic view refers to the quantitative and qualitative assessment of changes and interactions between the different layers of the biological information [7,18–21]. The genetic classifications of disease are now well established, given the modern genomic tools that can provide rich information about large patient cohorts. However, other highly complementary approaches based on proteomic and metabolic information can help researchers to biochemically or physiologically contextualize the underlying genetic information, thus helping to get closer to the phenotype and allowing patient stratification [22]. Thanks to disruptive technological jumps, a revolutionary vision was pioneered by Lee Hood, who coined the term P4 medicine [19], which is aimed to be predictive, preventive, personalized, and participatory. This new shift defines a new healthcare strategy in which each person serves as his or her own control over time [23].

The omics surge presents an amazing opportunity to provide new innovative tools for rapid diagnosis of IEM. Furthermore, metabolomics approaches are relevant for IEM because their basic pathophysiology is tightly related to metabolism. These diseases present with nonspecific clinical symptoms and appropriate laboratory tests are crucial in making a diagnosis. However, conventional biological diagnosis procedures are based on a series of sequential and segmented biochemical tests on various separated analytical platforms. This approach is slow, time-consuming, and complex, whereas optimal patient management requires improved speed of biochemical tests to allow early diagnosis and better monitoring of IEM. To address this need of faster screening and diagnosis strategies, metabolic profiling is a promising candidate.

In this review, we describe basic principles underlying metabolic phenotyping and metabolomic approaches that can be used to comprehensively assess an individual biochemical profile and their reported applications in IEM. Data for this review were identified by searches of PubMed and references from relevant articles using the search terms “metabolomics”, “metabonomics”, “metabolic profiling”, “inborn errors of metabolism”, and “inherited metabolic diseases”.

2. Metabolomics

2.1. Metabolites and Metabolome

The idea behind metabolomics goes back to ancient Greece, where doctors used the organoleptic characteristics of urine to link them to different medical conditions. Urine sweetness has been used to detect high glucose in diabetes [24]. Such organoleptic features are, of course, metabolic in origin. The word metabolome was coined by Olivier et al. in 1998 and defined as the set of metabolites synthesized by an organism [25]. Metabolome refers to the comprehensive complement of all metabolites present in a given biological system, fluid, cell, or tissue [26]. Metabolites can be defined as organic small molecules involved in enzymatic reactions. Thus, metabolomics is one of the “omic” technologies based on biochemical and molecular characterizations of the metabolome and the changes in metabolites related to genetic, environmental, drug or dietary, and other factors.

Metabolomics allows researchers to characterize these interactions and to evaluate the biochemical mechanisms involved in such changes in a systematic fashion. Indeed, metabolites fulfill the key criterion in that they change rapidly in response to physiological changes and may generate vital information about biochemical pathways that are modified in patients and in treated patients. Hence, metabolic profiling is highly informative since metabolites act as substrates or products in biochemical metabolic pathways [22,27–29].

Metabolomics has found applications in many disease studies and in complex interacting systems [22]. The possibility of predicting drug effects from baseline metabolic profiles has been demonstrated and gave rise to pharmacometabonomics as a potential effector for patient stratification and personalized medicine [30–35]. It is possible that the future of IEM diagnosis may be found in the developing area of metabolomics by doing simultaneous quantitative metabolic profiling of many metabolites in biological fluids.

2.2. Analytical Strategies and Chemical Information Retrieval

2.2.1. Biological Samples

For biological information recovery, metabolomics generally uses biofluids, cells, or tissue extracts as primary sources of metabolic fingerprint data. Compared with intact or extracted tissues, urine and serum or plasma are the most commonly studied biofluids in clinical practice, because they are easily obtained and prepared [36–39]. However, other specialized fluids could be used, including cerebrospinal fluid [40,41] or saliva [42–44] and even breath [45,46]. Dried blood (and other biofluids) spots samples (DBS) have also been investigated [47–50] and were shown to be an interesting alternative to conventional liquid samples for generating metabolite profiles. Given their practical advantages such as low volume, low cost, and handling convenience, DBS is gaining interest as a sampling support for metabolic profiling in IEM [47,51–53]. Of note, most metabolomics studies, particularly in clinical metabolomics, include data from a single biofluid, most often blood or urine. However, biochemical signature in a biofluid denotes complex interrelationships from different organs, which add another complexity layer for metabolomics data interpretation. This could be only understood by investigating pathophysiological states from a metabolic interactions perspective taking into account local metabolome specificities and their contribution to systemic metabolome. Different data-driven approaches have been described to handle these issues using multiple biofluids sampling and metabolomics data modeling [54,55].

2.2.2. Analytical Technologies

The human metabolome is a complex, highly responsive, and dynamic system. Thus, it raises different analytical challenges compared to other omics analysis approaches that are based on profiling large molecules built with a simple and limited set of subunits, such as nucleotides for genomics and transcriptomics and amino acids for proteomics. Thus, for identification and functional analysis of DNA, RNAs and proteins, the order combination of the subunits is what matters. It is the order of subunits that embodies the observed complexity that carries the biological information. Sequencing technologies rely basically on an incremental detection of these subunits. Researchers must figure out the order of the subunits to decode the carried biological information [56]. However, the same sequencing approach cannot be used to analyze metabolites in complex biofluids, because the analytical challenge is not simply to crack the order code, as there is no obvious order.

To retrieve the metabolic information, the metabolome requires a more complex analysis of chemical mixtures that allows components to be individually and selectively differentiated, identified and measured across a wide qualitative and quantitative chemical space.

The diversity of the physicochemical properties of the various metabolites groups adds another layer of complexity to metabolomics studies. This supplemental challenge has been the key driver for the development of various analytical protocols and platforms. Indeed, scientists tackled this analytical challenge even before the term metabolomics was coined. The first scientific article about metabolomics was published by Pauling and colleagues, in which they described a method using gas chromatographic separation with flame ionization detection to analyze the breath [57]. The authors referred to orthomolecular medicine linking the detected biochemical signature to phenotypes.

Since then, huge development has been made. The mainly used metabolic profiling technologies are nuclear magnetic resonance (NMR) spectroscopy [58–60] and mass spectrometry (MS), either combined or not to a gas phase or liquid phase separation method [27,51]. These technologies are suitable for metabolomics studies because they deliver global, unbiased, and comprehensive chemical information from complex mixtures. For information recovery, the multivariate spectroscopic data produced are typically analyzed using chemometric techniques to identify informative metabolic combinations that can be used for either sample classification or global biomarker discovery [51,61,62]. NMR spectroscopy is rapid and nondestructive and has the advantage of being highly reproducible. It is a powerful spectroscopic technology that offers atom-centered information that is crucial for molecular structure elucidation [63]. High-resolution NMR using stronger magnetic fields or two-dimensional NMR allows higher information recovery. The major drawback of NMR is its lack of sensitivity. However, MS offers complementary molecular information and is, by far, more sensitive than NMR. Hence, it allows higher metabolome coverage. The use of separation methods coupled to MS, such as liquid chromatography [38,39], gas chromatography [64], or capillary electrophoresis [65], allows a molecular separation step before MS detection. This enhances sensitivity and the dynamic range and provides complementary molecular information using the separation dimension.

Recently, approaches using another gas phase separation, ion mobility spectrometry (IMS) [66], has been gaining interest in metabolomics [67–73]. Indeed, IMS is a well-established post-ionization separation method based on size, shape, and charge performed on a millisecond timescale, which represents an intermediate timescale between chromatography (seconds) and high-resolution MS detection (microseconds). Coupled with high-resolution mass spectrometry and chromatography (LC-IM-MS), IMS provides additional analyte selectivity without significantly compromising the speed of MS-based measurements. The MS dimension affords accurate mass information, while the IMS dimension provides molecular, structural, and conformational information through the determination of the ion collision cross section. Indeed, ion mobility spectrometry adds a separation dimension to the hybrid MS instruments allowing, thus, a more comprehensive analysis of complex biological mixtures [69,74–77]. Furthermore, accessing retention time, accurate mass, and collision cross section obtained by the combination of LC-IM-MS allows measurement integration, which enhances molecular identification and consequently biomarker discovery [78,79].

Fourier transform mass spectrometry is another group of ultra-high-resolution methods that offer the highest resolving power, resolution, and mass-to-charge ratio (m/z) measurement accuracy and, hence, better metabolome coverage [80]. However, given their high cost, these methods are limited to only a few research groups.

Recently, to increase the high-throughput of global metabolic profiling analysis, ambient ionization sources were introduced. They are capable of direct sampling for complex matrices under ambient conditions. For example, atmospheric solids analysis probe [81], desorption electrospray ionization (DESI) [82–84], and rapid evaporative ionization MS methods [85,86] have been demonstrated to provide real-time, interpretable MS data on biofluids and tissues, in vivo and ex vivo, and will certainly reshape the future for high-throughput real-time metabolome analysis. In many surgeries, it is often difficult to distinct visually between the healthy and diseased tissues, and this requires time-consuming biopsies and immuno-staining procedures to be performed by histopathologists during surgery. By eliminating this need for external tissue histotyping, the iKnife could open the way to true real-time precision surgery. For more details about the use of ambient MS in clinical diagnosis, refer to a recent and detailed review by Ifa et al. [87].

Table 1 presents a comparison between different analytical strategies used in metabolomics with potential interest for IEM. Given the already existing chemical biomarker infrastructure and growing adoption of MS in clinical laboratories, its relatively low cost compared to NMR instruments, and the analytical performance of current mass spectrometers in terms of sensitivity and resolution in particular, MS-based metabolomics is a very promising tool in clinical biochemistry in the near future [88].

2.3. Metabolomics Workflows: Targeted vs. Untargeted

Metabolomics analysis is typically described as two complementary analytical approaches: targeted and untargeted. The first one aims to define the metabolic profile of the groups to study; subsequently, multivariate statistical analysis is undertaken to define the discriminating metabolites (potential biomarkers) between groups. Second, predictive mathematical models based on multivariate statistical analysis can be built. These models predict a subsequent classification of unknown biological samples (e.g., healthy versus diseased, treated versus untreated). The targeted approach focuses on identifying and quantifying selected metabolites according to their involvement in a metabolic pathway or their specific chemical or biochemical proprieties.

Table 1. Comparison of main metabolomics analytical technologies with particular potential in inborn errors of metabolism.

Platform	Technique	Identification Dimensions	Principle	Advantages	Limits
Nuclear Magnetic Resonance based methods (NMR)	1 Dimension 2 Dimensions	Chemical shift Chemical shift x Chemical shift	Uses interaction of spin active nuclei (^{1}H, ^{13}C, ^{31}P) with electromagnetic fields which gives structural, chemical and molecular environment information	Nondestructive Highly reproducible Exact quantification possible Minimal sample preparation Molecular dynamic and compartmental information using diffusional methods Relative high throughput Availability of databases for identification	High instrumentation cost Overlap of metabolites Low sensitivity
Mass spectrometry based methods	Direct Injection (DI-MS)	m/z	Uses a nanospray source directly coupled to MS detector. It does not require chromatographic separation	Very high throughput High sensitivity No cross sample contamination No column carryover Low cost analysis Automated analysis Low sample volume requirement	Samples not recoverable (destructive) No retention time information which gives limited specificity Inability to separate isomers and isobaric species Subjected to significant ion suppression phenomenon High ionization discrimination (ESI)
	Liquid chromatography (LC-MS)	Time x m/z	Uses chromatographic columns that enables liquid phase chromatographic separation of molecules followed by MS detection (Suitable for polar to hydrophobic compounds)	Minimal sample preparation (protein precipitation or dilution of biological sample) High throughput capability UPLC can be coupled to any type of MS Flexibility in column chemistry widening the range of detectable compounds High sensitivity	Samples not recoverable (destructive) Very polar molecules need specific chromatographic conditions Retention times are highly dependent of exact chromatographic conditions Batch analysis Lack of large metabolite databases High ionization discrimination (ESI)
	Gas Chromatography (GC-MS)	Time x *m/z*	Uses chromatographic columns that enables gas phase chromatographic separation of molecules followed by MS detection (Suited for apolar and volatiles compounds)	Structure information obtained through in-source fragmentation Availability of universal databases for identification High sensitive Reproducible	Samples not recoverable (destructive) Requires higher sample preparation Only volatile compounds are detected Polar compounds need derivatization Low ionization discrimination
	Capillary Electrophoresis (CE-MS)	Time x m/z	Uses electrokinetic separation of polar molecules hyphenated with a mass spectrometry detector	Excellent for polar analysis in aqueous samples Measures inorganic and organic anions Low running costs	Samples not recoverable (destructive) Relatively low throughput profiling
	Ion Mobility (IM-MS)	Time x m/z (CCS x m/z)	Uses a uniform or periodic electric field and a buffer gas, to separate ions based on size and shape which is hyphenated with mass spectrometry	Very robust and reproducible (ability to determine Collision Cross Section which is a robust chemical descriptor) High peak capacity High selectivity Separation of isomeric and isobaric compounds Very high throughput	Samples not recoverable (destructive) CCS and mass are highly correlated parameters which limits the orthogonality of the method

In general, a metabolomic analysis involves mainly four steps. Step 1 is a preparatory step on both analytical and conceptual aspects. It is initiated by the biological question to consider and the definition of the study aim. It also defines the most informative biological matrix and the experimental design to implement. In addition, this step defines the appropriate sample preparation according to the considered analytical study. Step 2 includes analytical and instrumental strategy choices. During this step, data are collected and processed and then statistical analysis is performed. Step 3 involves the putative annotation, identification, and confirmation of the potential biomarkers generated by the data analysis. Chemical, biochemical, and spectral databases are queried. Step 4 aims to build a predictive mathematical model based on the identified biomarkers. This model is then validated analytically and clinically. This final step involves the integration of experimental data and their interpretation in the studied biological or clinical context [89]. Figure 1 illustrates the general workflow of translational metabolomics.

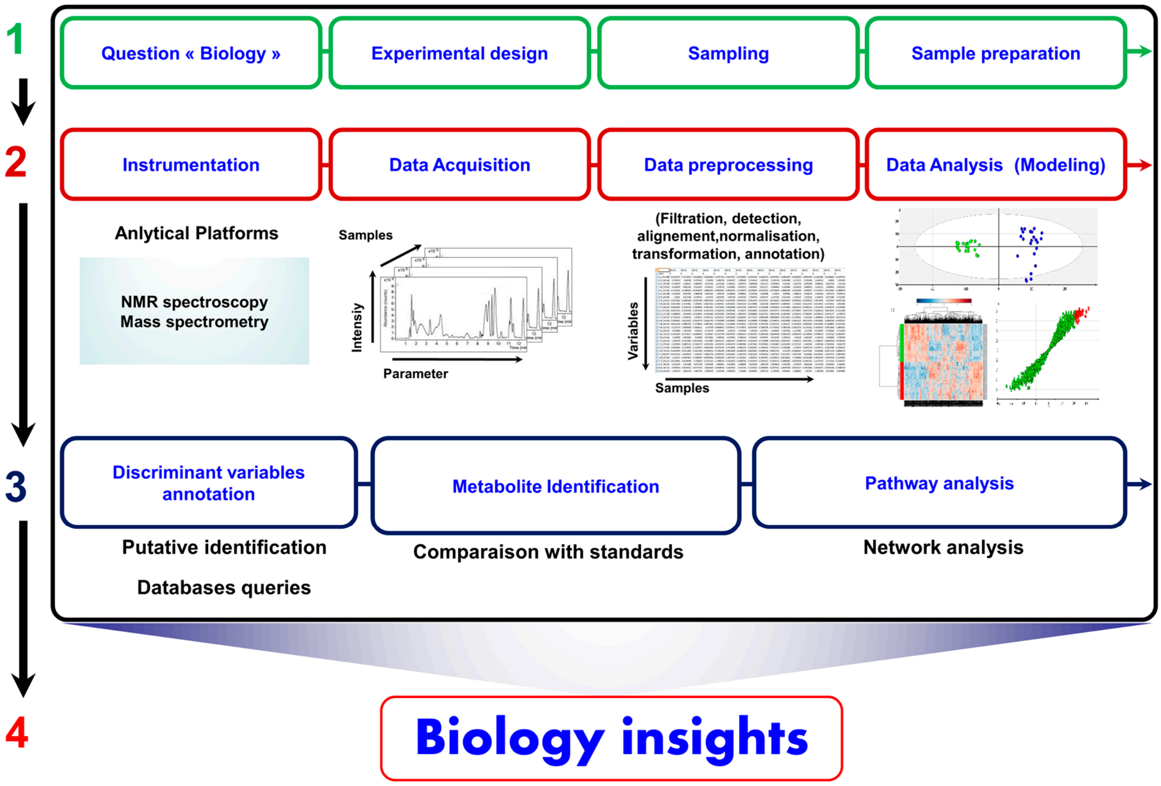


Figure 1. Translational metabolomics workflow.

2.4. Data Analysis, Information Recovery, and the Curse of Dimensionality

Few highly reliable metabolites could be, at some extent, sufficient for diagnostic or monitoring purposes. However, a broader overview using more metabolites is more appropriate to assess, for example, a biochemical pathway. Thus, the choice of the most appropriate data modeling strategy is an important issue and is dependent on the underlying question to be addressed. In mechanistic studies, the structural data descriptions and the underlying extracted information yielded by the built model are more important than its predictive ability to classify new samples. However, in diagnosis applications, the predictive performances of the model are vital regarding samples classification. Hence, the clear and precise definition of the study aims has to be intelligible and purpose driven.

The analytical performance improvements associated with metabolomics platforms led to the generation of complex and high-dimensional datasets. Handling, in a smoothly high-throughput fashion, the huge amount of generated data is a very important issue for transforming the data into clinically actionable knowledge.

2.4.1. Univariate Data Analysis

Metabolomics data analysis can be approached from a univariate perspective using traditional statistical methods that consider only one variable at a time. Univariate methods are common statistical analysis tools and their main advantage is the convenient use and interpretation. To assess the differences between two or more groups, parametric tests such as Student's *t*-test and ANOVA are commonly applied, respectively. However, normality assumptions should be verified for consistent conclusions [90]. Otherwise, non-parametric test such as Mann–Whitney *U* test or Kruskal–Wallis one-way analysis of variance could be used if normality is not assumed. Another important issue is that applying multiple univariate tests in parallel to a high dimensional dataset raises the multiple testing problem. In metabolomics studies, a large number of features are simultaneously analyzed. Thus, the probability to find a statistically significant difference accidentally (i.e., true positive) is high. In order to handle this multiple testing issue, different correction methods could be used. Each method tries to balance between avoiding false metabolite associations (i.e., false positives) and discarding true associations (i.e., false negatives). In the Bonferroni correction, the significance level for a hypothesis is divided by the number of hypotheses tested simultaneously [90]. Hence, the Bonferroni correction is considered a stringent correction method. Other less conservative methods are available and are mostly based on the minimization of the false positives or false-discovery rate (FDR). FDR-based methods minimize the expected proportion of false positives on the total number of positives [91]. Gene expression microarray data analysis has matured most of these methods, where thousands of genes are simultaneously tested. Similarly, in untargeted metabolomics studies large sets of metabolites are measured in parallel. The use of less restrictive approaches such as FDR methods seems to be more useful.

Furthermore, it should be noted that potential confounding factors like gender, age or diet may affect the output results if not properly addressed. Furthermore, the main limit of these approaches is their lack of handling the correlations and interactions between the different metabolic features. Hence, advanced multivariate approaches are more suitable.

2.4.2. Multivariate Data Analysis

Translating biological data into knowledge requires addressing biology as an informational science using tools that allow to track the information at large scales. To do so, an entire field was born "Bioinformatics" [92]. Bioinformatics can be defined as mean of conceptualizing biology in terms of molecules and by applying "informatics techniques" borrowed to disciplines such as applied mathematics, computer science and statistics to understand and organize the information related to these molecules, on a large scale. In short, bioinformatics is a management information system for a biological system [93].

The high-dimensionality of metabolic data requires adapted statistical tools to retrieve as much as possible chemical information from the data to translate it into biological knowledge. The major challenge is to reduce the dimensionality by selecting relevant signals from the noisy raw data. To achieve this goal, chemometric tools are widely used. Chemometrics is the science of extracting useful information from chemical systems using data-driven means [94]. It is inherently interdisciplinary, borrowing methods from data-analytic disciplines such as multivariate statistics, applied mathematics, and computer science. Thus, chemometrics is applied to solve both descriptive and predictive problems using biochemical data.

The data analysis methods are mainly divided into two types: unsupervised and supervised methods. The former are mainly exploratory, whereas the latter are explanatory and predictive. Unsupervised methods are used to analyze the behavior of the observations in the data set without taking into account any related outcome. Because there is no class labeling or response, the data set is considered as a collection of analogous objects. Unsupervised learning methods track patterns or clustering trends in the data to understand any spontaneous relationships between the samples. It can also highlight the variables that are responsible for these relationships. Based on effective visualization means, unsupervised learning helps to reveal categories of samples or variables that naturally cluster together based on their underlying similarities. In metabolomics data, it is the metabolic similarity that shapes the clustering. Principal component analysis [95] is a widely used pattern recognition method; it is a projection-based method that reduces the dimensionality of the data by creating components or latent variables. Principal component analysis allows a two- or three-dimensional visualization of the data. However, clustering methods aim to identify clusters in the dataset using similarity measures. A dendrogram or a heat map can be then formed to visualize the samples similarities. The commonly used clustering methods are k-means clustering [96], hierarchical cluster analysis [97], and self-organizing maps [98]. Correlation matrix could also be used to get an overview of the data. Because the main goal in metabolomics, especially in clinical context, is to differentiate between groups (healthy versus diseased, treated versus control), a sample can be classified according to its spectral patterns. The metabolic features responsible for the classification can then be identified. The metabolic features intensities in the dataset matrix can be considered as a multidimensional space of metabolites coordinates. Thus, each spectrum is a point in a multidimensional metabolic hyperspace.

In supervised methods, the multivariate datasets can be modeled so that the class label of separate samples known as a validation set can be predicted based on a series of mathematical models derived from the original data, namely the training set. Various supervised methods could be used in metabolomics, including partial least squares (PLS) methods such as PLS-Discriminant Analysis (PLS-DA) [99] and Orthogonal-PLS-DA (OPLS-DA) [100], as well as support vector machines [101]. Methods based on topology data analysis are gaining great interests and seem promising for data analysis because of their intrinsic flexibility and exploratory and predictive abilities [102]. It must be noted that the retrieved information from the raw data and the generated outputs are highly dependent on the chosen data analysis strategy. Hence, the aim of metabolomics research and the data analysis step are mutually dependent.

Of note, multivariate and univariate data analysis pipelines are not mutually exclusive and it is often recommended to use both to maximize the quality of the information extraction from metabolomics data.

For further details on data analysis techniques and tools in metabolomics, refer to recent reviews on this issue [103–105].

2.5. Pathway and Network Analysis: From Information to Knowledge

The integration of experimental data and computational tools is mandatory to understand complex biological systems. This gave birth to computational biology which could be divided into two distinct branches: knowledge discovery or data-mining, and simulation-based analysis. The former extracts the hidden patterns from huge amount of experimental data, generating hypotheses. However, the latter tests hypotheses with in silico experiments, providing predictions to be confirmed by in vitro and in vivo studies [9].

One of the biggest challenges of any metabolomics study is linking the identified metabolites to biology, which is a crucial step to move from biomarkers towards more mechanistic insights. To achieve this purpose, pathway and network analysis approaches aim to capitalize on the information generated by metabolomics studies to get insightful inference [106,107]. Both approaches exploit the interrelationships properties contained in the metabolomic data. Network modeling and pathway-mapping tools help to decipher metabolites interactions roles in a biological disturbance [107].

Metabolic pathways are sets of metabolites that are connected to the same biological process, and that are linked by one or multiple enzymatic reactions directly or indirectly. Biological databases are therefore seminal enablers providing rich information of different of metabolic pathways (Table 2). Indeed, pathway analysis (PA) uses prior biological knowledge to analyze metabolic patterns from an integrative point of view. Pathway-based methods are currently known as metabolite set enrichment analysis (MSEA), and are methodologically based on the gene set enrichment analysis (GSEA) approach, previously developed for pathway analysis of gene-expression data [108,109].

Table 2. Biological databases.

Databases	Websites	Ref.
KEGG (Kyoto Encyclopedia of Genes and Genomes)	http://www.genome.jp/kegg	[110]
HumanCyc (Encylopedia of Human Metabolic Pathways)	http://humancyc.org	[111]
MetaCyc (Encyclopedia of Metabolic Pathways)	http://metacyc.org	[112]
Reactome (A Curated Knowledgebase of Pathways)	http://www.reactome.org	[113]
SMPDB (Small Molecule Pathway Database)	http://www.smpdb.ca	[114]
Virtual Metabolic Human Database	https://vmh.uni.lu	[106]
Wikipathways	http://www.wikipathways.org	[115]

There are mainly three distinct methods to perform MSEA [108,116]:

Overrepresentation analysis (ORA): The basic hypothesis in this method is that relevant pathways can be detected if the proportion of differential expressed metabolites, within a given pathway, exceeds the proportion of metabolites that could be randomly expected. A hypergeometric test or a Fisher's Exact test is used to evaluate the statistical significance of whether the metabolite belongs to the pathway. The final result from an ORA method consists in a list of the most relevant pathways, ranked by p-value and/or a multiple-hypothesis-test-corrected p-value. The ORA main advantage over non-knowledge-driven (i.e., purely data-driven) analysis is that it gives metabolomic data a biological context. This allows formulating a hypothesis that could subsequently be test experimentally. Hence, ORA turns data analysis into a knowledge generation cycle, proper of the Systems Biology approach. However, PA exhibits some limits. Due to the selected cut-off method for statistical significance potentially important components could be omitted in the analysis. Furthermore, PA assume that pathways are independent from each other, which is contrary to the admitted interaction and overlapping between pathways [108]. Other methods have been developed to overcome these limits.

Quantitative enrichment analysis (QEA): In this approach, the input data are a set of quantified metabolite from multiple samples. Thus, absolute concentrations are used. Enriched pathways can be identified using different approaches like the Wilcoxon-based test [117], globaltest [118] or globalAncova [119]. Enriched pathways include pathways where a set of metabolites that are significantly changed or pathways where a large number of metabolites that significantly changed [116,120].

Single-sample profiling (SSP): Unlike the previous methods that are designed for studies involving multiple samples, this method is used at the sample level. In this case, SSP requires a list of metabolite concentrations in biofluids (i.e., urine, blood and CSF), tissue, or cell type and a database with the normal concentration ranges of the chosen metabolites in the analyzed sample. Thus, SSP identifies, from the data, the set of metabolites presenting significantly different levels compared to the normal ranges [116,120].

For better interpretability of pathway analysis outputs, MSEA results could be combined with pathway topological analysis (PTA). PTA measures assess the impact of the disturbed metabolites within the pathway. First, single impacts are evaluated using the degree and betweenness network centrality measures of each metabolite. This represents the number of shortest paths passing through a certain node to estimate its centrality (importance). Subsequently, the overall impact (i.e., pathway impact) is calculated as the sum of the single impact measures of the disturbed metabolites normalized by the sum of the measures of the impact of all the metabolites within the considered pathway [121]. Indeed, changes in the most important nodes within a network generate a more significant impact on the system than changes in bordering or solitary nodes.

From a topological standpoint, a metabolic network can be considered as an interconnected ensemble of nodes presented by metabolites, and edges representing reactions catalyzed by enzymes. Thus, unlike PA, network analysis uses the high degree of correlation in metabolomics data to build metabolic networks that characterize the complex relationships the measured metabolites.

Biological data exhibit a high level of correlation that exists between the different biological components (i.e., DNA, mRNAs, proteins and metabolites). Indeed, a given metabolite may be connected to different metabolic pathways and, thus, show correlation patterns. In other cases, the observed correlations may be due to other causes such as global changes (i.e., diurnal variation in time series studies) or specific changes due to the intrinsic variability of metabolomic data [54,122]. These patterns can provide valuable information about the underlying metabolic network associated to a specific biological process [54,123].

Based on the observed relationship patterns present in the experimental data, correlation-based methods allow building metabolic networks in which each metabolite represents a node. However, unlike the pathway analysis, the links between nodes denotes the level of mathematical correlation between each metabolites pair and called edge. High correlation coefficients are frequent in metabolomics data which is due to the presence of systemic associations [123]. Hence, using classical correlation coefficients leads to overcrowded networks. In addition, direct and indirect associations are not distinguished. To overcome this problem partial correlation could be used [54,123,124]. In partial correlation approach, the correlation between two metabolites is conditioned against the correlation with the remaining metabolites. Consequently, partial correlation allows discriminating between direct and indirect metabolite correlations. In this method, the link between two metabolites is scored according to the ratios differences between the corresponding metabolites in the two sample groups. Therefore, the related network topology is based on the metabolic differences between the two studied phenotypes. These data-driven strategies have been successfully applied for reconstruction of metabolic networks from metabolomics data [123,125,126].

Metabolite identification is a challenging and time consuming task. Thus, a novel approach, named Mummichog, has been proposed by Li et al. for network analysis. This method predicts biological activity directly from mass spectrometry based untargeted metabolomics data without a priori identification of metabolites. The idea behind this strategy is combining network analysis and metabolite prediction under the same computational framework reducing significantly the metabolomics workflow time. This method has been elegantly illustrated by exploring the activation of innate immune cells. It yielded that glutathione metabolism is modified by viral infection driven by constitutive nitric oxide synthases [127].

A wide variety of software tools are available to analyze metabolomic data at the pathway and network level. Table 3 presents different functional analysis tools for both pathway analysis and visualization.

Table 3. Functional analysis and biological interpretation tools.

Tools	Websites	Ref.
	Pathway and Networks Analysis and Visualization	
BioCyc—Omics Viewer	http://biocyc.org	[128]
iPath	http://pathways.embl.de	[129]
Metscape	http://metscape.ncibi.org	[130]
Paintomics	http://www.paintomics.org	[131]
Pathos	http://motif.gla.ac.uk/Pathos	[132]
Pathvisio	http://www.pathvisio.org	[133]
VANTED	http://vanted.ipk-gatersleben.de	[134]
IMPaLA	http://impala.molgen.mpg.de	[135]
MBROLE 2.0	http://csbg.cnb.csic.es/mbrole2	[136]
MPEA	http://ekhidna.biocenter.helsinki.fi/poxo/mpea	[137]
Mummichog	http://clinicalmetabolomics.org/init/default/software	[127]
	Multifunctional Tools	
MetaboAnlayst	http://www.metaboanalyst.com	[120]
XCMS online	https://xcmsonline.scripps.edu	[138]
MASSyPup	http://www.bioprocess.org/massypup	[139]
Workflow4Metabolomics	http://workflow4metabolomics.org	[140]
MetaboLyzer	https://sites.google.com/a/georgetown.edu/fornace-lab-informatics/home/metabolyzer	[141]

Contextual interpretation is crucial to fully embrace the potential of metabolomics. Indeed, metabolites carry out precious contextual biological information. In a metabolic network, flux is defined as the rate (i.e., quantity per unit time) at which metabolites are converted or transported between different compartments [10]. Thus, metabolic fluxes, or fluxome, represent a unique and functional readout of the phenotype. The fluxome captures the metabolome in its ultimate functional interactions with the environment and the genome [10,142]. As such, the fluxome integrates information on different cellular processes, and hence it is a unique spatiotemporal phenotypic signature of cells. Thus, one or more metabolic fluxes could be altered in a metabolic disorder depending on the complexity of the disease [2]. Different strategies are used to translate metabolomics data into fluxomic insights by modeling of metabolic networks. The network modeling can be achieved using constraints of mass and charge conservation along with stoichiometric and thermodynamic ones [34,143–145]. Based on the stoichiometry of the reactants and products of biochemical reactions, flux balance analysis (FBA) can estimate metabolic fluxes without knowledge about the kinetics of the participating enzymes [10,142]. Recently, Cortassa et al. suggested a new approach, distinct from FBA or metabolic flux analysis, which takes into account kinetic mechanisms and regulatory interactions [146].

3. Potential Integration of Metabolomics in Laboratory Medicine Frameworks

Metabolites embody physiological end-points and regulatory processes directly connected to the fluxome. Hence, the metabolome is very time sensitive and is constantly changing. Therefore, changes in metabolite concentrations are usually more suitable to describe the biochemical state of a biological system. Because metabolomics is the ultimate expression of the genes' influences and proteins' use of metabolites, it offers a rich and tremendous view on the phenotype. Indeed, metabolites carry out precious contextual biological information that could be used to assess pathophysiological states.

Metabolic profiling as a diagnostic tool opens an informative metabolic window into disease, which makes metabolomics an appealing ally in disease diagnosis.

What makes metabolomics a key driver in the post-genomic era is its tight relationship with the phenotype, whether the phenotype is driven by a monogenic or a multifactorial complex condition. Linking metabolic profile modulation with particular genetic variation [126] and/or environmental factors such as the microbiome [147], diet [148], toxics [56], or therapies [34] offers an exciting opportunity to rationalize diagnostics and translate a more comprehensive information into clinical actionable knowledge. The above-cited factors that influence the metabolome, and then phenotype(s), remind us that assessing metabolites, as chemical supporters of life, is the core for knowledge building that will shape clinical decisions.

Early biochemists such as Cori, Warburg, Meyerhof, and Krebs made seminal contributions to map most fundamental aspects of metabolic pathways and physiology. Therefore, urine chemical properties guided early physicians in founding the concept of IEM [4]. Sir Garrod's idea suggested that a biochemical fingerprint within biofluids was a product of human variation and, hence, could be a surrogate for distinct diseases. Garrod argued that the IEM that he was able to observe "were merely extreme examples of variations of chemical behavior which are probably everywhere present in minor degrees" [149]. In other words, he believed that there were phenotypes that could be associated with specific biochemicals. However, given the limited technical sensitivity back then, he was not able to affirm this idea. Recently, his hypothesis was elegantly confirmed with metabolomics approaches and metabolic modeling [126].

With the expected improvements in the metabolic profiling scope and data quality, metabolomics is destined to play a major and disruptive role in the near future as an efficient screening and diagnostic tool [150]. There are mainly two ways that metabolomics could be implemented in clinical context and laboratory medicine: chemometrics or a quantitative approach. For the former, direct statistical analysis is applied to spectral patterns and signal intensity data, and identification of metabolites may be performed in the last step if needed. This method captures metabolic snapshots and builds pattern-recognition-based models using machine learning techniques to sort samples (subjects) according to their metabolic patterns. This approach is eloquently embodied by the intelligent scalpel (iKnife) introduced by Takatz et al. [85], which instantaneously classifies, in vivo and ex vivo, cancerous and noncancerous tissues. This compelling technology aims to help surgeons during cancer surgery [85,86]. In contrast, the quantitative approach targets a set of metabolites and then analyzes the quantitative data directly. This approach affords an absolute quantitation of a set of chosen metabolites (e.g., amino acids, carnitines and acylcarnitines, or organic acids). A multivariate predictive model can be built based on the absolute concentration of these metabolites to predict clinical status or intervention outcomes. Compared to quantitative metabolomics, the key advantage of chemometric profiling is its capability of automated and unbiased assessment of metabolomics data. However, it requires a large number of spectra and sample uniformity, which are less of a concern in quantitative metabolomics. Nevertheless, the multivariate data analysis strategies underlying the two strategies are quite similar. Figure 2 illustrates the two clinical workflows.

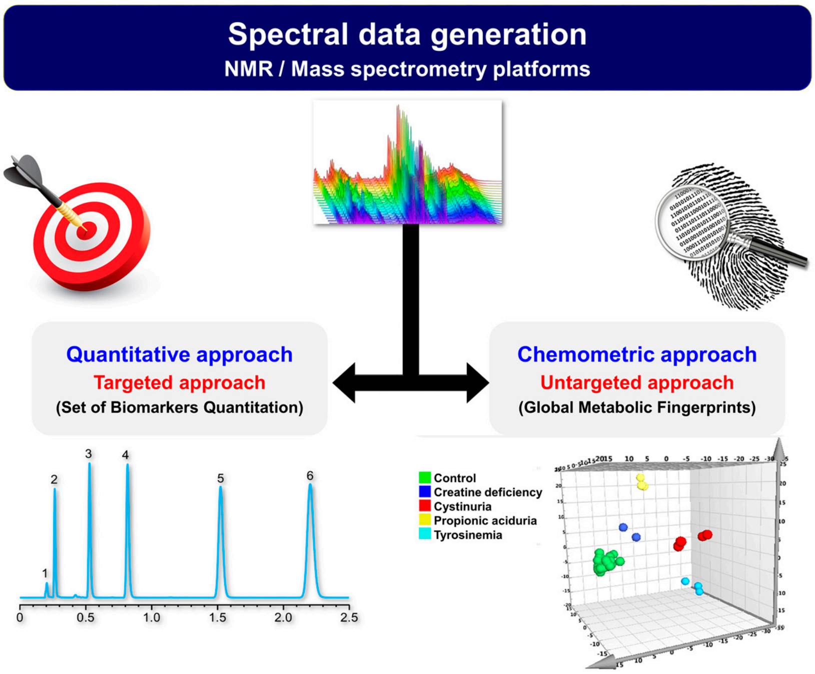


Figure 2. Clinical metabolomics implementation strategies.

4. Applications of Metabolomics in Inborn Errors of Metabolism (IEM) Investigations

IEM being tightly connected with metabolism, the inherent pathophysiological changes are the main determinant of the metabolome and functional understanding of the disease. Hence, due to its intrinsic multidisciplinary nature, integrating biochemistry, analytical chemistry, advanced statistics and bioinformatics, metabolomics analysis represents a promising tool to achieve improved understanding and better diagnosis of IEM in the post-genomic and precision medicine era. For years, MS has been used in the assessment of inherited metabolic diseases. Several IEM are currently diagnosed using targeted MS-based metabolomics methods such as aminoacidopathies, organic acidurias, and fatty acid oxidation disorders [151–155]. Furthermore, MS is now widely implemented in IEM newborn screening national programs worldwide [156]. However, the combination of the already existing tools with data analysis strategies is compelling for better biological information recovery.

To assess different IEM including aminoacidopathies, organic aciduria, and mitochondrial disorders, Janeckova et al. used targeted analysis combined with multivariate data analysis. Their work showed how combining chemometrics modeling and absolute quantification are eloquently complementary in the assessment of IEMs [157]. Drecksen et al. used a similar approach to assess isovaleric aciduria (IVA) based on 86 urine samples: 10 untreated and 10 treated IVA cases, 12 heterozygotes, 22 children controls, and 32 adult controls. The work succeeded in producing a comprehensive profile of metabolites of practical significance in IVA [158]. Osterman et al. described a matrix-assisted laser desorption/ionization MS-based method for acylcarnitine and organic acid analysis on DBS. The method enabled the identification and quantification of metabolites involved in different organic aciduria and beta oxidation deficiencies [159]. Using targeted metabolomics addressing complex lipids, Fan and colleagues showed that some sphingolipids species were elevated in Niemann–Pick Type C1 subjects. These lipid biomarkers may be used for monitoring the efficacy of specific therapy [160].

Given the increasing potential of metabolomics in IEM, different groups published work regarding the usefulness of untargeted-metabolomics-based approaches in IEM in disease characterization, diagnosis, and biomarker discovery. For characterization of disease biosignatures, respiratory chain deficiencies have been investigated by several research groups to track specific metabolic signatures using metabolomics [161–163]. Wikoff et al. used MS-based untargeted metabolomics in plasma to characterize methylmalonic acidemia and propionic aciduria. Propionylcarnitine, a known biomarker, was retrieved using untargeted strategy which illustrates the potential of metabolic profiling in biomarker detection. Five additional plasma acylcarnitine metabolites presented significant differences between patients and control individuals. In addition, γ-butyrobetaine was highly increased in a subset of patients. This demonstrates that metabolomics can widen the range of metabolites associated with IEM [164]. Auray-Blais et al. used MS-based untargeted metabolomics for biomarker discovery in Fabry disease which led to the discovery of seven globotriaosylceramide (Gb3) analogues as biomarkers that are now suggested biomarkers for the screening and the follow-up of Fabry disease [61,62,165].

Sholmi et al. presented an elegant computational approach for assessing metabolic profiles of red blood cells enzyme deficiencies. The developed predictive method yielded biomarkers for red blood cells alterations and revealed a strong correlation with disrupted metabolic concentrations. Over 200 metabolites were identified as potential biomarkers due to 176 enzyme deficiencies. Furthermore, already known disease indicators were retrieved by the developed prediction method. Importantly, potential novel biomarkers were also predicted. This approach proved to dramatically increase biomarker discovery performance [166].

Because the metabolome is highly influenced by nutritional factors, diet monitoring also has been investigated using metabolomics. Phenylketonuria is an interesting example of diet monitoring in IEM. Using metabolomics, Mutze et al. showed that a long-term dietary fatty acid restriction influences mitochondrial beta-oxidation intermediates. No functional influence on unsaturated fatty acid metabolism and platelet aggregation in patients with phenylketonuria was detected [167].

Regarding the use of metabolomics platforms as a diagnosis tool, several teams proposed metabolomics workflows. Using NMR and DESI-MS methods, Pan et al. clearly discriminated six patients with IEMs from six controls based on their respective urine metabolic profiles, identifying argininosuccinic aciduria, classic homocystinuria, classic methylmalonic acidemia, maple syrup urine disease, phenylketonuria, and type II tyrosinemia [168]. Later, Denes et al. proposed a method based on high-resolution MS with high throughput using DBS and direct flow injection analysis. Their method has been tested on 500 controls and 66 abnormal samples and showed a clear discrimination of the various assessed metabolic diseases [51]. Ilya et al. also proposed another method based on high-resolution MS coupled to liquid chromatography for the assessment of IEM. Their method resolved highly polar as well as hydrophobic analytes under reverse-phase conditions, enabling analysis of a wide range of chemicals in an untargeted fashion. Their work provides a tailored high-resolution MS platform for IEM and covers various metabolites usually quantified by a combination of different separate instrumentation [169]. Miller et al. described a comprehensive global strategy to assess IEM using liquid chromatography and gas chromatography MS-based metabolomics platforms combining both targeted and untargeted analysis. In total, 120 plasma samples from patients with a confirmed IEM and those of 70 controls were assessed. This strategy allowed, elegantly, comprehensive pathway analysis that provides useful diagnostic information of IEM [170].

Regarding NMR-based platforms, Aygen and colleagues conducted a multi-center clinical study in 14 clinical centers in Turkey. Urine samples from 989 neonates were collected and investigated using NMR spectroscopy in two different laboratories to assess reproducibility. The objectives of their study were twofold: (1) to explore the metabolite variations to set pathological thresholds of specific metabolites in comparison with healthy neonates to develop predictive models; and (2) to build a NMR database from a healthy population of neonates for IEM metabolite identification [171].

5. Clinical and Translational Metabolomics Challenges

5.1. Metabolite Identification

Metabolite identification is the main bottleneck of metabolomics for large adoption in both translational and clinical context. Despite spectral information becomes available in the literature or in spectral databases, metabolites identification is still a challenging task [172]. To the best of our knowledge, there is no software currently available to fully and smoothly facilitate the identification process. Especially, the integration of NMR and MS data, which is essential for the reliable identification of metabolites. Furthermore, metabolite identification is mandatory for absolute quantitation especially in MS based methods requiring the use of labeled isotope. Thus, more efforts are needed to enhance this drawback of metabolomics.

5.2. Standardization and Harmonization

Standardization is a vital aspect for a wide spread of any new technology. Thus, for clinical metabolomics, harmonization of the sample preparation, processing, analysis and reporting using validated and standardized protocols is mandatory [173,174]. This is important since biological samples change over time. The lack of harmonization in protocols for sample handling, MS and NMR data generation and data reporting could lead to poor reproducibility and, thus, to data misinterpretation, particularly in population metabolic profiling. This is a fundamental obstacle for clinical translation. A definition of normal or reference samples is also important to build reference databases. This will rely on signatures derived from the complete characterization of the considered disease, IEM in the scope of this review. Finally, addressing these standardization issues is essential for regulatory compliance, which is a prerequisite for clinical implementation and adoption.

5.3. Automation, Data Visualization and Clinical Actionability

Automation at different stages, instrument-, pre- and post-analytic levels are a very important issue for large clinical adoption of any diagnostic innovation. Metabolomics workflow automation is a key enabler regarding high-throughput, reproducibility and reliability which are pillars of modern laboratory medicine practice. To address this limit, current efforts are promising like the iKnife, which would allow real-time cancer diagnosis [85] and breathomics strategies for lung and respiratory diseases based on breath signatures [72]. Data fusion and integration of omics and other biological and clinical data is another great challenge to fully unveil the potential of metabolomics [17,175]. With this regard, combining genomic and metabolic profiling information to enhance clinical diagnostics and to enable patient stratification and monitoring of interventional pathways patient journeys is a promising field [22,176]. The clinical actionability would involve advanced mathematical modeling of genomic and metabolic data sets in relation to patient clinical data using machine learning and expert systems. Intuitive visualization tools of the data in clinical accessible formats are needed to support effective clinical decision making. Figure 3 presents the main challenges in clinical metabolomics.

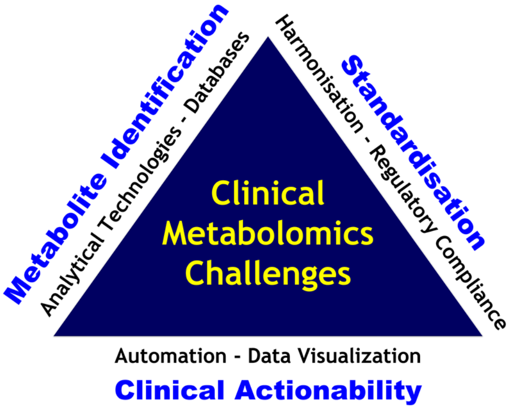

Metabolite Identification
Analytical Technologies - Databases
Harmonisation - Regulatory Compliance
Standardisation
Clinical Metabolomics Challenges
Automation - Data Visualization
Clinical Actionability

Figure 3. Metabolomics challenges for effective clinical implementation.

6. Conclusions

It is common to perform early diagnosis of IEM by assessing specific metabolic biomarkers related to a genetic defect. However, the original paradigm of "one gene–one enzyme–one disease" is no longer viewed as a reality for IEM. The impact of an altered protein on metabolic flux is not easily predictable. Indeed, the metabolic pathways are not linear and metabolites are tightly linked with several interactions within a highly organized network [21,177]. Depending on the complexity of the disease, one primary metabolite flux or an entire network of metabolite fluxes might be affected [2,20]. Therefore, a complete contextual, multilayer, network-based functional overview is needed to effectively assess all the actors of a given pathway in a holistic fashion [8]. Systemic approaches are needed to understand IEM complexity and to effectively diagnose and treat them [21]. To achieve such a goal, metabolomics is a key driver in the systems medicine based strategy. The great potential of metabolomics integration with other omics data will allow systems biology and clinical data to be linked. This paves the way for a paradigm shift in medical practice from cohort evidence-based medicine to algorithm-based precision medicine. This will in turn enhance clinicians' abilities to be more pre-emptive and thus, more efficient in handling IEM.

Metabolomics is still in its infancy with regard to the investigation of IEM, and its great potential has yet to be explored worldwide at both the basic and clinical sides. Improving workflows for high-quality data acquisition, processing, and visualization is an important issue for effectively translating the biological information into actionable knowledge under clinically accessible formats for effective healthcare management. However, this innovative global approach also requires a paradigm shift in our practice at different levels. A complete change is needed in our screening and diagnosis strategies. Thus, a disruptive move from a hypothesis-driven approach to a more data-driven and hypothesis-generating approach is crucial to address the challenges of the post-genomic era. The core idea of the paradigm shift in IEM laboratory investigation is presented in Figure 4.

Furthermore, totally new investigative thinking is needed to transform all aspects of the laboratory medicine enterprise, including education, research, and healthcare. Upgrading medical practitioners' skill sets on both the clinical and laboratory sides is needed to smoothly achieve the full potential of systems medicine. These skills integrate biology, computing and data analytics to develop common communication channels for more effective medical interactions. This ongoing high digitization of the individual biological and clinical information offers a tremendous and exciting opportunity to fully embrace the promising era of precision medicine.

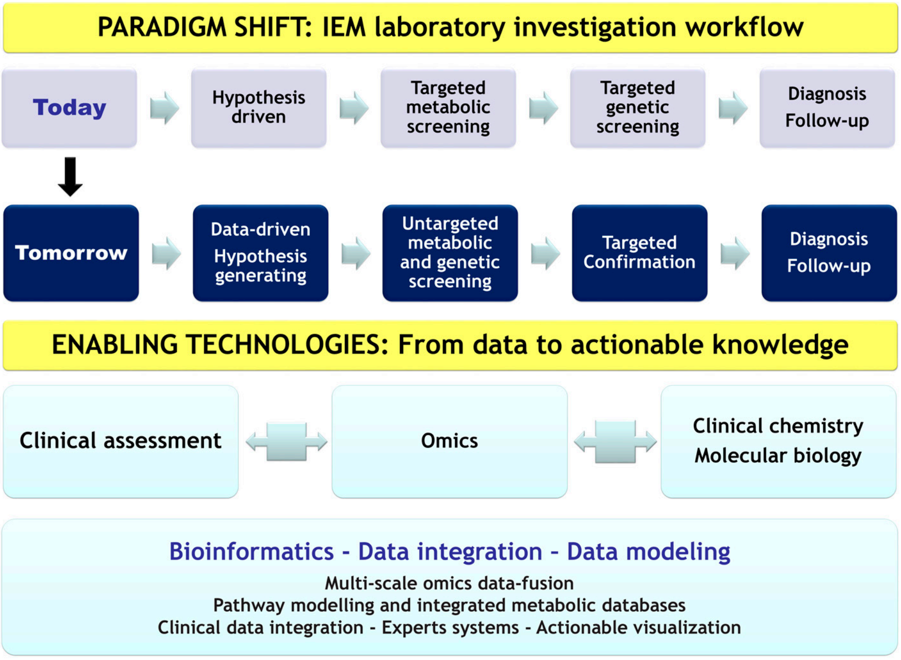

PARADIGM SHIFT: IEM laboratory investigation workflow
Today
Hypothesis driven
Targeted metabolic screening
Targeted genetic screening
Diagnosis Follow-up
Tomorrow
Data-driven Hypothesis generating
Untargeted metabolic and genetic screening
Targeted Confirmation
Diagnosis Follow-up
ENABLING TECHNOLOGIES: From data to actionable knowledge
Clinical assessment
Omics
Clinical chemistry Molecular biology
Bioinformatics - Data integration - Data modeling
Multi-scale omics data-fusion
Pathway modelling and integrated metabolic databases
Clinical data integration - Experts systems - Actionable visualization

Figure 4. Paradigm shift in Inborn Errors of Metabolism diagnosis workflow. The change in molecular information recovery in laboratory investigation workflow is driven by advancing analytical technologies and bioinformatics systems for a more effective medical practice using an integrative computational framework. IEM: Inborn Errors of Metabolism.

Acknowledgments: This work was supported by the Normandy University, the Institut National de la Santé et de la Recherche Médicale (INSERM), the Conseil Régional de Normandie, Labex SynOrg (ANR-11-LABX-0029) and the European Regional Development Fund (ERDF 31708).

Author Contributions: Abdellah Tebani performed literature review and wrote the manuscript including tables and figures. Lenaig Abily-Donval performed literature search. Carlos Afonso critically revised and edited the manuscript. Stéphane Marret critically revised and edited the manuscript. Soumeya Bekri conceived the topic of the review and critically revised and edited the manuscript. All authors approved the final manuscript.

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

References

1. Collins, F.S.; Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **2015**, *372*, 793–795. [CrossRef] [PubMed]
2. Lanpher, B.; Brunetti-Pierri, N.; Lee, B. Inborn errors of metabolism: The flux from mendelian to complex diseases. *Nat. Rev. Genet.* **2006**, *7*, 449–460. [CrossRef] [PubMed]
3. Vernon, H.J. Inborn errors of metabolism: Advances in diagnosis and therapy. *JAMA Pediatr.* **2015**, *169*, 778–782. [CrossRef] [PubMed]
4. Garrod, A. The incidence of alkaptonuria: A study in chemical individuality. *Lancet* **1902**, *160*, 1616–1620. [CrossRef]
5. Ahn, A.C.; Tewari, M.; Poon, C.S.; Phillips, R.S. The limits of reductionism in medicine: Could systems biology offer an alternative? *PLoS Med.* **2006**, *3*, e208. [CrossRef] [PubMed]
6. Regenmortel, M.H.V.V. Reductionism and complexity in molecular biology. *EMBO Rep.* **2004**, *5*, 1016–1020. [CrossRef] [PubMed]
7. Aon, M.A.; Lloyd, D.; Saks, V. From physiology, genomes, systems, and self-organization to systems biology: The historical roots of a twenty-first century approach to complexity. In *Systems Biology of Metabolic and Signaling Networks: Energy, Mass and Information Transfer*; Aon, A.M., Saks, V., Schlattner, U., Eds.; Springer: Berlin & Heidelberg, Germany, 2014; pp. 3–17.

8. Aon, M.A. Complex systems biology of networks: The riddle and the challenge. In *Systems Biology of Metabolic and Signaling Networks*; Springer: Berlin & Heidelberg, Germany, 2014; pp. 19–35.
9. Kitano, H. Computational systems biology. *Nature* **2002**, *420*, 206–210. [CrossRef] [PubMed]
10. Aon, M.A.; Cortassa, S. Systems biology of the fluxome. *Processes* **2015**, *3*, 607–618. [CrossRef]
11. Weston, A.D.; Hood, L. Systems biology, proteomics, and the future of health care: Toward predictive, preventative, and personalized medicine. *J. Proteome Res.* **2004**, *3*, 179–196. [CrossRef] [PubMed]
12. Ehrenberg, M.; Elf, J.; Aurell, E.; Sandberg, R.; Tegner, J. Systems biology is taking off. *Genome Res.* **2003**, *13*, 2377–2380. [CrossRef] [PubMed]
13. Kitano, H. Looking beyond the details: A rise in system-oriented approaches in genetics and molecular biology. *Curr. Genet.* **2002**, *41*, 1–10. [CrossRef] [PubMed]
14. Kitano, H. Systems biology: A brief overview. *Science* **2002**, *295*, 1662–1664. [CrossRef] [PubMed]
15. Tenenbaum, J.D.; Avillach, P.; Benham-Hutchins, M.; Breitenstein, M.K.; Crowgey, E.L.; Hoffman, M.A.; Jiang, X.; Madhavan, S.; Mattison, J.E.; Nagarajan, R.; et al. An informatics research agenda to support precision medicine: Seven key areas. *JAMIA* **2016**, *23*, 791–795. [CrossRef] [PubMed]
16. McMurry, J.; Kohler, S.; Balhoff, J.; Borromeo, C.; Brush, M.; Carbon, S.; Conlin, T.; Dunn, N.; Engelstad, M.; Foster, E.; et al. Navigating the phenotype frontier: The monarch initiative. *bioRxiv* **2016**. [CrossRef]
17. Ritchie, M.D.; Holzinger, E.R.; Li, R.; Pendergrass, S.A.; Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* **2015**, *16*, 85–97. [CrossRef] [PubMed]
18. Sperisen, P.; Cominetti, O.; Martin, F.-P.J. Longitudinal omics modeling and integration in clinical metabonomics research: Challenges in childhood metabolic health research. *Front. Mol. Biosci.* **2015**. [CrossRef] [PubMed]
19. Hood, L.; Balling, R.; Auffray, C. Revolutionizing medicine in the 21st century through systems approaches. *Biotechnol. J.* **2012**, *7*, 992–1001. [CrossRef] [PubMed]
20. Cho, D.-Y.; Kim, Y.-A.; Przytycka, T.M. Chapter 5: Network biology approach to complex diseases. *PLoS Comput. Biol.* **2012**, *8*, e1002820. [CrossRef] [PubMed]
21. Argmann, C.A.; Houten, S.M.; Zhu, J.; Schadt, E.E. A next generation multiscale view of inborn errors of metabolism. *Cell Metab.* **2016**, *23*, 13–26. [CrossRef] [PubMed]
22. Nicholson, J.K.; Holmes, E.; Kinross, J.M.; Darzi, A.W.; Takats, Z.; Lindon, J.C. Metabolic phenotyping in clinical and surgical environments. *Nature* **2012**, *491*, 384–392. [CrossRef] [PubMed]
23. Chen, R.; Mias, G.I.; Li-Pook-Than, J.; Jiang, L.; Lam, H.Y.K.; Chen, R.; Miriami, E.; Karczewski, K.J.; Hariharan, M.; Dewey, F.E.; et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **2012**, *148*, 1293–1307. [CrossRef] [PubMed]
24. Nicholson, J.K.; Lindon, J.C. Systems biology: Metabonomics. *Nature* **2008**, *455*, 1054–1056. [CrossRef] [PubMed]
25. Oliver, S.G.; Winson, M.K.; Kell, D.B.; Baganz, F. Systematic functional analysis of the yeast genome. *Trends Biotechnol.* **1998**, *16*, 373–378. [CrossRef]
26. 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. [CrossRef] [PubMed]
27. Dunn, W.B.; Broadhurst, D.I.; Atherton, H.J.; Goodacre, R.; Griffin, J.L. Systems level studies of mammalian metabolomes: The roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem. Soc. Rev.* **2011**, *40*, 387–426. [CrossRef] [PubMed]
28. Fiehn, O. Metabolomics—The link between genotypes and phenotypes. *Plant Mol. Biol.* **2002**, *48*, 155–171. [CrossRef] [PubMed]
29. Holmes, E.; Wilson, I.D.; Nicholson, J.K. Metabolic phenotyping in health and disease. *Cell* **2008**, *134*, 714–717. [CrossRef] [PubMed]
30. Clayton, T.A.; Lindon, J.C.; Cloarec, O.; Antti, H.; Charuel, C.; Hanton, G.; Provost, J.P.; le Net, J.L.; Baker, D.; Walley, R.J.; et al. Pharmaco-metabonomic phenotyping and personalized drug treatment. *Nature* **2006**, *440*, 1073–1077. [CrossRef] [PubMed]
31. Kaddurah-Daouk, R.; Kristal, B.S.; Weinshilboum, R.M. Metabolomics: A global biochemical approach to drug response and disease. *Annu. Rev. Pharmacol. Toxicol.* **2008**, *48*, 653–683. [CrossRef] [PubMed]
32. James, L.P. Metabolomics: Integration of a new "omics" with clinical pharmacology. *Clin. Pharmacol. Ther.* **2013**, *94*, 547–551. [CrossRef] [PubMed]

33. Kaddurah-Daouk, R.; Weinshilboum, R.M. Pharmacometabolomics: Implications for clinical pharmacology and systems pharmacology. *Clin. Pharmacol. Ther.* **2014**, *95*, 154–167. [CrossRef] [PubMed]
34. Kell, D.B.; Goodacre, R. Metabolomics and systems pharmacology: Why and how to model the human metabolic network for drug discovery. *Drug Discov. Today* **2014**, *19*, 171–182. [CrossRef] [PubMed]
35. Everett, J.R. Pharmacometabonomics in humans: A new tool for personalized medicine. *Pharmacogenomics* **2015**, *16*, 737–754. [CrossRef] [PubMed]
36. Dunn, W.B.; Broadhurst, D.; Begley, P.; Zelena, E.; Francis-McIntyre, S.; Anderson, N.; Brown, M.; Knowles, J.D.; Halsall, A.; Haselden, J.N.; 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*, 1060–1083. [CrossRef] [PubMed]
37. Nunes de Paiva, M.J.; Menezes, H.C.; de Lourdes Cardeal, Z. Sampling and analysis of metabolomes in biological fluids. *Analyst* **2014**, *139*, 3683–3694. [CrossRef] [PubMed]
38. Want, E.J.; Masson, P.; Michopoulos, F.; Wilson, I.D.; Theodoridis, G.; Plumb, R.S.; Shockcor, J.; Loftus, N.; Holmes, E.; Nicholson, J.K. Global metabolic profiling of animal and human tissues via UPLC-MS. *Nat. Protoc.* **2013**, *8*, 17–32. [CrossRef] [PubMed]
39. Want, E.J.; Wilson, I.D.; Gika, H.; Theodoridis, G.; Plumb, R.S.; Shockcor, J.; Holmes, E.; Nicholson, J.K. Global metabolic profiling procedures for urine using UPLC-MS. *Nat. Protoc.* **2010**, *5*, 1005–1018. [CrossRef] [PubMed]
40. Graham, S.F.; Chevallier, O.P.; Roberts, D.; Hölscher, C.; Elliott, C.T.; Green, B.D. Investigation of the human brain metabolome to identify potential markers for early diagnosis and therapeutic targets of Alzheimer's disease. *Anal. Chem.* **2013**, *85*, 1803–1811. [CrossRef] [PubMed]
41. Wuolikainen, A.; Hedenstrom, M.; Moritz, T.; Marklund, S.L.; Antti, H.; Andersen, P.M. Optimization of procedures for collecting and storing of CSF for studying the metabolome in ALS. *Amyotroph. Lateral Scler.* **2009**, *10*, 229–236. [CrossRef] [PubMed]
42. Dame, Z.; Aziat, F.; Mandal, R.; Krishnamurthy, R.; Bouatra, S.; Borzouie, S.; Guo, A.; Sajed, T.; Deng, L.; Lin, H.; et al. The human saliva metabolome. *Metabolomics* **2015**, *11*, 1864–1883. [CrossRef]
43. Kawasaki, G.; Ichikawa, Y.; Yoshitomi, I.; Umeda, M. Metabolomics of salivary biomarkers in yusho patients. *Hukuoka Acta Med.* **2015**, *106*, 144–148. [PubMed]
44. Mikkonen, J.J.; Singh, S.P.; Herrala, M.; Lappalainen, R.; Myllymaa, S.; Kullaa, A.M. Salivary metabolomics in the diagnosis of oral cancer and periodontal diseases. *J. Periodontal Res.* **2015**. [CrossRef] [PubMed]
45. Bach, J.-P.; Gold, M.; Mengel, D.; Hattesohl, A.; Lubbe, D.; Schmid, S.; Tackenberg, B.; Rieke, J.; Maddula, S.; Baumbach, J.I.; et al. Measuring compounds in exhaled air to detect Alzheimer's disease and parkinson? S disease. *PLoS ONE* **2015**, *10*, e0132227. [CrossRef] [PubMed]
46. Pijls, K.E.; Smolinska, A.; Jonkers, D.M.A.E.; Dallinga, J.W.; Masclee, A.A.M.; Koek, G.H.; van Schooten, F.-J. A profile of volatile organic compounds in exhaled air as a potential non-invasive biomarker for liver cirrhosis. *Sci. Rep.* **2016**. [CrossRef] [PubMed]
47. Koulman, A.; Prentice, P.; Wong, M.C.; Matthews, L.; Bond, N.J.; Eiden, M.; Griffin, J.L.; Dunger, D.B. The development and validation of a fast and robust dried blood spot based lipid profiling method to study infant metabolism. *Metabolomics* **2014**, *10*, 1018–1025. [CrossRef] [PubMed]
48. Wilson, I. Global metabolic profiling (metabonomics/metabolomics) using dried blood spots: Advantages and pitfalls. *Bioanalysis* **2011**, *3*, 2255–2257. [CrossRef] [PubMed]
49. Michopoulos, F.; Theodoridis, G.; Smith, C.J.; Wilson, I.D. Metabolite profiles from dried blood spots for metabonomic studies using UPLC combined with orthogonal acceleration TOF-MS: Effects of different papers and sample storage stability. *Bioanalysis* **2011**, *3*, 2757–2767. [CrossRef] [PubMed]
50. Prentice, P.; Turner, C.; Wong, M.C.Y.; Dalton, R.N. Stability of metabolites in dried blood spots stored at different temperatures over a 2-year period. *Bioanalysis* **2013**, *5*, 1507–1514. [CrossRef] [PubMed]
51. Denes, J.; Szabo, E.; Robinette, S.L.; Szatmari, I.; Szonyi, L.; Kreuder, J.G.; Rauterberg, E.W.; Takats, Z. Metabonomics of newborn screening dried blood spot samples: A novel approach in the screening and diagnostics of inborn errors of metabolism. *Anal. Chem.* **2012**, *84*, 10113–10120. [CrossRef] [PubMed]
52. Wagner, M.; Tonoli, D.; Varesio, E.; Hopfgartner, G. The use of mass spectrometry to analyze dried blood spots. *Mass Spectrom. Rev.* **2014**, *35*, 361–438. [CrossRef] [PubMed]

53. Oliveira, R.V.; Henion, J.; Wickremsinhe, E.R. Automated high-capacity on-line extraction and bioanalysis of dried blood spot samples using liquid chromatography/high-resolution accurate mass spectrometry. *Rapid Commun. Mass Spectrom.* **2014**, *28*, 2415–2426. [CrossRef] [PubMed]
54. Do, K.T.; Kastenmüller, G.; Mook-Kanamori, D.O.; Yousri, N.A.; Theis, F.J.; Suhre, K.; Krumsiek, J. Network-based approach for analyzing intra- and interfluid metabolite associations in human blood, urine, and saliva. *J. Proteome Res.* **2015**, *14*, 1183–1194. [CrossRef] [PubMed]
55. Torell, F.; Bennett, K.; Cereghini, S.; Rannar, S.; Lundstedt-Enkel, K.; Moritz, T.; Haumaitre, C.; Trygg, J.; Lundstedt, T. Multi-organ contribution to the metabolic plasma profile using hierarchical modelling. *PLoS ONE* **2015**, *10*, e0129260. [CrossRef] [PubMed]
56. Athersuch, T. Metabolome analyses in exposome studies: Profiling methods for a vast chemical space. *Arch. Biochem. Biophys.* **2016**, *589*, 177–186. [CrossRef] [PubMed]
57. Pauling, L.; Robinson, A.B.; Teranishi, R.; Cary, P. Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. *Proc. Natl. Acad. Sci. USA* **1971**, *68*, 2374–2376. [CrossRef] [PubMed]
58. Jimenez, B.; Montoliu, C.; MacIntyre, D.A.; Serra, M.A.; Wassel, A.; Jover, M.; Romero-Gomez, M.; Rodrigo, J.M.; Pineda-Lucena, A.; Felipo, V. Serum metabolic signature of minimal hepatic encephalopathy by (1) h-nuclear magnetic resonance. *J. Proteome Res.* **2010**, *9*, 5180–5187. [CrossRef] [PubMed]
59. Wijeyesekera, A.; Selman, C.; Barton, R.H.; Holmes, E.; Nicholson, J.K.; Withers, D.J. Metabotyping of long-lived mice using 1 h NMR spectroscopy. *J. Proteome Res.* **2012**, *11*, 2224–2235. [CrossRef] [PubMed]
60. Larive, C.K.; Barding, G.A., Jr.; Dinges, M.M. NMR spectroscopy for metabolomics and metabolic profiling. *Anal. Chem.* **2015**, *87*, 133–146. [CrossRef] [PubMed]
61. Auray-Blais, C.; Boutin, M. Novel GB(3) isoforms detected in urine of fabry disease patients: A metabolomic study. *Curr. Med. Chem.* **2012**, *19*, 3241–3252. [CrossRef] [PubMed]
62. Manwaring, V.; Boutin, M.; Auray-Blais, C. A metabolomic study to identify new globotriaosylceramide-related biomarkers in the plasma of fabry disease patients. *Anal. Chem.* **2013**, *85*, 9039–9048. [CrossRef] [PubMed]
63. Emwas, A.-H.; Salek, R.; Griffin, J.; Merzaban, J. Nmr-based metabolomics in human disease diagnosis: Applications, limitations, and recommendations. *Metabolomics* **2013**, *9*, 1048–1072. [CrossRef]
64. Chan, E.C.Y.; Pasikanti, K.K.; Nicholson, J.K. Global urinary metabolic profiling procedures using gas chromatography-mass spectrometry. *Nat. Protoc.* **2011**, *6*, 1483–1499. [CrossRef] [PubMed]
65. Ramautar, R.; Somsen, G.W.; de Jong, G.J. CE-MS for metabolomics: Developments and applications in the period 2012–2014. *Electrophoresis* **2015**, *36*, 212–224. [CrossRef] [PubMed]
66. Hill, H.H., Jr.; Siems, W.F.; st Louis, R.H.; McMinn, D.G. Ion mobility spectrometry. *Anal. Chem.* **1990**, *62*, 1201A–1209A. [CrossRef] [PubMed]
67. Paglia, G.; Angel, P.; Williams, J.P.; Richardson, K.; Olivos, H.J.; Thompson, J.W.; Menikarachchi, L.; Lai, S.; Walsh, C.; Moseley, A.; et al. Ion mobility-derived collision cross section as an additional measure for lipid fingerprinting and identification. *Anal. Chem.* **2015**, *87*, 1137–1144. [CrossRef] [PubMed]
68. Maldini, M.; Natella, F.; Baima, S.; Morelli, G.; Scaccini, C.; Langridge, J.; Astarita, G. Untargeted metabolomics reveals predominant alterations in lipid metabolism following light exposure in broccoli sprouts. *Int. J. Mol. Sci.* **2015**, *16*, 13678–13691. [CrossRef] [PubMed]
69. Paglia, G.; Williams, J.P.; Menikarachchi, L.; Thompson, J.W.; Tyldesley-Worster, R.; Halldórsson, S.; Rolfsson, O.; Moseley, A.; Grant, D.; Langridge, J.; et al. Ion mobility derived collision cross sections to support metabolomics applications. *Anal. Chem.* **2014**, *86*, 3985–3993. [CrossRef] [PubMed]
70. Wickramasekara, S.I.; Zandkarimi, F.; Morre, J.; Kirkwood, J.; Legette, L.; Jiang, Y.; Gombart, A.F.; Stevens, J.F.; Maier, C.S. Electrospray quadrupole travelling wave ion mobility time-of-flight mass spectrometry for the detection of plasma metabolome changes caused by xanthohumol in obese zucker (fa/fa) rats. *Metabolites* **2013**, *3*, 701–717. [CrossRef] [PubMed]
71. Dwivedi, P.; Schultz, A.J.; Hill, H.H. Metabolic profiling of human blood by high resolution ion mobility mass spectrometry (IM-MS). *Int. J. Mass Spectrom.* **2010**, *298*, 78–90. [CrossRef] [PubMed]
72. Hauschild, A.C.; Frisch, T.; Baumbach, J.I.; Baumbach, J. Carotta: Revealing hidden confounder markers in metabolic breath profiles. *Metabolites* **2015**, *5*, 344–363. [CrossRef] [PubMed]
73. Smolinska, A.; Hauschild, A.C.; Fijten, R.R.; Dallinga, J.W.; Baumbach, J.; van Schooten, F.J. Current breathomics—A review on data pre-processing techniques and machine learning in metabolomics breath analysis. *J. Breath Res.* **2014**. [CrossRef] [PubMed]

74. Fenn, L.; Kliman, M.; Mahsut, A.; Zhao, S.; McLean, J. Characterizing ion mobility-mass spectrometry conformation space for the analysis of complex biological samples. *Anal. Bioanal. Chem.* **2009**, *394*, 235–244. [CrossRef] [PubMed]
75. Fenn, L.; McLean, J. Biomolecular structural separations by ion mobility–mass spectrometry. *Anal. Bioanal. Chem.* **2008**, *391*, 905–909. [CrossRef] [PubMed]
76. Kliman, M.; May, J.C.; McLean, J.A. Lipid analysis and lipidomics by structurally selective ion mobility-mass spectrometry. *Biochim. Biophys. Acta* **2011**, *1811*, 935–945. [CrossRef] [PubMed]
77. Tebani, A.; Schmitz-Afonso, I.; Rutledge, D.N.; Gonzalez, B.J.; Bekri, S.; Afonso, C. Optimization of a liquid chromatography ion mobility-mass spectrometry method for untargeted metabolomics using experimental design and multivariate data analysis. *Anal. Chim. Acta* **2016**, *913*, 55–62. [CrossRef] [PubMed]
78. May, J.C.; Goodwin, C.R.; McLean, J.A. Ion mobility-mass spectrometry strategies for untargeted systems, synthetic, and chemical biology. *Curr. Opin. Biotechnol.* **2015**, *31*, 117–121. [CrossRef] [PubMed]
79. Sherrod, S.D.; McLean, J.A. Systems-wide high-dimensional data acquisition and informatics using structural mass spectrometry strategies. *Clin. Chem.* **2015**, *62*, 77–83. [CrossRef] [PubMed]
80. Junot, C.; Madalinski, G.; Tabet, J.C.; Ezan, E. Fourier transform mass spectrometry for metabolome analysis. *Analyst* **2010**, *135*, 2203–2219. [CrossRef] [PubMed]
81. Twohig, M.; Shockcor, J.P.; Wilson, I.D.; Nicholson, J.K.; Plumb, R.S. Use of an atmospheric solids analysis probe (ASAP) for high throughput screening of biological fluids: Preliminary applications on urine and bile. *J. Proteome Res.* **2010**, *9*, 3590–3597. [CrossRef] [PubMed]
82. Eberlin, L.S.; Norton, I.; Orringer, D.; Dunn, I.F.; Liu, X.; Ide, J.L.; Jarmusch, A.K.; Ligon, K.L.; Jolesz, F.A.; Golby, A.J.; et al. Ambient mass spectrometry for the intraoperative molecular diagnosis of human brain tumors. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 1611–1616. [CrossRef] [PubMed]
83. Ferreira, C.R.; Jarmusch, A.K.; Pirro, V.; Alfaro, C.M.; Gonzalez-Serrano, A.F.; Niemann, H.; Wheeler, M.B.; Rabel, R.A.; Hallett, J.E.; Houser, R.; et al. Ambient ionisation mass spectrometry for lipid profiling and structural analysis of mammalian oocytes, preimplantation embryos and stem cells. *Reprod. Fertil. Dev.* **2015**, *27*, 621–637. [CrossRef] [PubMed]
84. Kerian, K.S.; Jarmusch, A.K.; Pirro, V.; Koch, M.O.; Masterson, T.A.; Cheng, L.; Cooks, R.G. Differentiation of prostate cancer from normal tissue in radical prostatectomy specimens by desorption electrospray ionization and touch spray ionization mass spectrometry. *Analyst* **2015**, *140*, 1090–1098. [CrossRef] [PubMed]
85. Balog, J.; Sasi-Szabo, L.; Kinross, J.; Lewis, M.R.; Muirhead, L.J.; Veselkov, K.; Mirnezami, R.; Dezso, B.; Damjanovich, L.; Darzi, A.; et al. Intraoperative tissue identification using rapid evaporative ionization mass spectrometry. *Sci. Transl. Med.* **2013**. [CrossRef] [PubMed]
86. Balog, J.; Kumar, S.; Alexander, J.; Golf, O.; Huang, J.; Wiggins, T.; Abbassi-Ghadi, N.; Enyedi, A.; Kacska, S.; Kinross, J.; et al. In vivo endoscopic tissue identification by rapid evaporative ionization mass spectrometry (REIMS). *Angew. Chem. Int. Ed. Engl.* **2015**, *54*, 11059–11062. [CrossRef] [PubMed]
87. Ifa, D.R.; Eberlin, L.S. Ambient ionization mass spectrometry for cancer diagnosis and surgical margin evaluation. *Clin. Chem.* **2016**, *62*, 111–123. [CrossRef] [PubMed]
88. Annesley, T.; Diamandis, E.; Bachmann, L.; Hanash, S.; Hart, B.; Javahery, R.; Singh, R.; Smith, R. A spectrum of views on clinical mass spectrometry. *Clin. Chem.* **2016**, *62*, 30–36. [CrossRef] [PubMed]
89. Wishart, D.S.; Jewison, T.; Guo, A.C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; et al. Hmdb 3.0—The human metabolome database in 2013. *Nucleic Acids Res.* **2013**, *41*, D801–D807. [CrossRef] [PubMed]
90. Broadhurst, D.I.; Kell, D.B. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* **2006**, *2*, 171–196. [CrossRef]
91. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **1995**, *57*, 289–300.
92. Hogeweg, P. The roots of bioinformatics in theoretical biology. *PLoS Comput. Biol.* **2011**, *7*, e1002021. [CrossRef] [PubMed]
93. Luscombe, N.M.; Greenbaum, D.; Gerstein, M. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf. Med.* **2001**, *40*, 346–358. [PubMed]
94. Brereton, R.G. A short history of chemometrics: A personal view. *J. Chemom.* **2014**, *28*, 749–760. [CrossRef]
95. Hotelling, H. *Analysis of a Complex of Statistical Variables into Principal Components*; Warwick & York: Oradell, NJ, USA, 1933.

96. Hartigan, J.A.; Wong, M.A. Algorithm as 136: A k-means clustering algorithm. *J. R. Stat. Soc.* **1979**, *28*, 100–108. [CrossRef]
97. Johnson, S.C. Hierarchical clustering schemes. *Psychometrika* **1967**, *32*, 241–254. [CrossRef] [PubMed]
98. Kohonen, T. The self-organizing map. *Proc. IEEE* **1990**, *78*, 1464–1480. [CrossRef]
99. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130. [CrossRef]
100. Trygg, J.; Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom.* **2002**, *16*, 119–128. [CrossRef]
101. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
102. Offroy, M.; Duponchel, L. Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry. *Anal. Chim. Acta* **2016**, *910*, 1–11. [CrossRef] [PubMed]
103. Ren, S.; Hinzman, A.; Kang, E.; Szczesniak, R.; Lu, L. Computational and statistical analysis of metabolomics data. *Metabolomics* **2015**, *11*, 1492–1513. [CrossRef]
104. Gromski, P.S.; Muhamadali, H.; Ellis, D.I.; Xu, Y.; Correa, E.; Turner, M.L.; Goodacre, R. 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. [CrossRef] [PubMed]
105. Misra, B.B.; van der Hooft, J.J. Updates in metabolomics tools and resources: 2014–2015. *Electrophoresis* **2016**, *37*, 86–110. [CrossRef] [PubMed]
106. Thiele, I.; Swainston, N.; Fleming, R.M.; Hoppe, A.; Sahoo, S.; Aurich, M.K.; Haraldsdottir, H.; Mo, M.L.; Rolfsson, O.; Stobbe, M.D.; et al. A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* **2013**, *31*, 419–425. [CrossRef] [PubMed]
107. Cazzaniga, P.; Damiani, C.; Besozzi, D.; Colombo, R.; Nobile, M.S.; Gaglio, D.; Pescini, D.; Molinari, S.; Mauri, G.; Alberghina, L.; et al. Computational strategies for a system-level understanding of metabolism. *Metabolites* **2014**, *4*, 1034–1087. [CrossRef] [PubMed]
108. Garcia-Campos, M.A.; Espinal-Enriquez, J.; Hernandez-Lemus, E. Pathway analysis: State of the art. *Front. Physiol.* **2015**. [CrossRef] [PubMed]
109. Khatri, P.; Sirota, M.; Butte, A.J. Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput. Biol.* **2012**, *8*, e1002375. [CrossRef] [PubMed]
110. Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. Kegg as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **2016**, *44*, D457–D462. [CrossRef] [PubMed]
111. Romero, P.; Wagg, J.; Green, M.L.; Kaiser, D.; Krummenacker, M.; Karp, P.D. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.* **2005**. [CrossRef] [PubMed]
112. Caspi, R.; Foerster, H.; Fulcher, C.A.; Kaipa, P.; Krummenacker, M.; Latendresse, M.; Paley, S.; Rhee, S.Y.; Shearer, A.G.; Tissier, C.; et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res.* **2008**, *36*, D623–D631. [CrossRef] [PubMed]
113. Vastrik, I.; D'Eustachio, P.; Schmidt, E.; Gopinath, G.; Croft, D.; de Bono, B.; Gillespie, M.; Jassal, B.; Lewis, S.; Matthews, L.; et al. Reactome: A knowledge base of biologic pathways and processes. *Genome Biol.* **2007**. [CrossRef] [PubMed]
114. Jewison, T.; Su, Y.; Disfany, F.M.; Liang, Y.; Knox, C.; Maciejewski, A.; Poelzer, J.; Huynh, J.; Zhou, Y.; Arndt, D.; et al. Smpdb 2.0: Big improvements to the small molecule pathway database. *Nucleic Acids Res.* **2014**, *42*, D478–D484. [CrossRef] [PubMed]
115. Kelder, T.; van Iersel, M.P.; Hanspers, K.; Kutmon, M.; Conklin, B.R.; Evelo, C.T.; Pico, A.R. Wikipathways: Building research communities on biological pathways. *Nucleic Acids Res.* **2012**, *40*, D1301–D1307. [CrossRef] [PubMed]
116. Xia, J.; Wishart, D.S. MSEA: A web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.* **2010**, *38*, W71–W77. [CrossRef] [PubMed]
117. Adjaye, J.; Huntriss, J.; Herwig, R.; BenKahla, A.; Brink, T.C.; Wierling, C.; Hultschig, C.; Groth, D.; Yaspo, M.L.; Picton, H.M.; et al. Primary differentiation in the human blastocyst: Comparative molecular portraits of inner cell mass and trophectoderm cells. *Stem Cells* **2005**, *23*, 1514–1525. [CrossRef] [PubMed]
118. Goeman, J.J.; van de Geer, S.A.; de Kort, F.; van Houwelingen, H.C. A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics* **2004**, *20*, 93–99. [CrossRef] [PubMed]
119. Hummel, M.; Meister, R.; Mansmann, U. Globalancova: Exploration and assessment of gene group effects. *Bioinformatics* **2008**, *24*, 78–85. [CrossRef] [PubMed]

120. Xia, J.; Sinelnikov, I.V.; Han, B.; Wishart, D.S. Metaboanalyst 3.0—Making metabolomics more meaningful. *Nucleic Acids Res.* **2015**, *43*, W251–W257. [CrossRef] [PubMed]
121. Xia, J.; Wishart, D.S. METPA: A web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics* **2010**, *26*, 2342–2344. [CrossRef] [PubMed]
122. Steuer, R. Review: On the analysis and interpretation of correlations in metabolomic data. *Brief. Bioinform.* **2006**, *7*, 151–158. [CrossRef] [PubMed]
123. Krumsiek, J.; Suhre, K.; Illig, T.; Adamski, J.; Theis, F.J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol.* **2011**. [CrossRef] [PubMed]
124. Valcarcel, B.; Wurtz, P.; Seich al Basatena, N.K.; Tukiainen, T.; Kangas, A.J.; Soininen, P.; Jarvelin, M.R.; Ala-Korpela, M.; Ebbels, T.M.; de Iorio, M. A differential network approach to exploring differences between biological states: An application to prediabetes. *PLoS ONE* **2011**, *6*, e24702. [CrossRef] [PubMed]
125. Bartel, J.; Krumsiek, J.; Schramm, K.; Adamski, J.; Gieger, C.; Herder, C.; Carstensen, M.; Peters, A.; Rathmann, W.; Roden, M.; et al. The human blood metabolome-transcriptome interface. *PLoS Genet.* **2015**, *11*, e1005274. [CrossRef] [PubMed]
126. Shin, S.Y.; Fauman, E.B.; Petersen, A.K.; Krumsiek, J.; Santos, R.; Huang, J.; Arnold, M.; Erte, I.; Forgetta, V.; Yang, T.P.; et al. An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **2014**, *46*, 543–550. [CrossRef] [PubMed]
127. Li, S.; Park, Y.; Duraisingham, S.; Strobel, F.H.; Khan, N.; Soltow, Q.A.; Jones, D.P.; Pulendran, B. Predicting network activity from high throughput metabolomics. *PLoS Comput. Biol.* **2013**, *9*, e1003123. [CrossRef] [PubMed]
128. Caspi, R.; Billington, R.; Ferrer, L.; Foerster, H.; Fulcher, C.A.; Keseler, I.M.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Mueller, L.A.; et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res.* **2016**, *44*, D471–D480. [CrossRef] [PubMed]
129. Yamada, T.; Letunic, I.; Okuda, S.; Kanehisa, M.; Bork, P. Ipath2.0: Interactive pathway explorer. *Nucleic Acids Res.* **2011**, *39*, W412–W415. [CrossRef] [PubMed]
130. Karnovsky, A.; Weymouth, T.; Hull, T.; Tarcea, V.G.; Scardoni, G.; Laudanna, C.; Sartor, M.A.; Stringer, K.A.; Jagadish, H.V.; Burant, C.; et al. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics* **2012**, *28*, 373–380. [CrossRef] [PubMed]
131. 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*, 137–139. [CrossRef] [PubMed]
132. Leader, D.P.; Burgess, K.; Creek, D.; Barrett, M.P. Pathos: A web facility that uses metabolic maps to display experimental changes in metabolites identified by mass spectrometry. *Rapid Commun. Mass Spectrom.* **2011**, *25*, 3422–3426. [CrossRef] [PubMed]
133. Kutmon, M.; van Iersel, M.P.; Bohler, A.; Kelder, T.; Nunes, N.; Pico, A.R.; Evelo, C.T. Pathvisio 3: An extendable pathway analysis toolbox. *PLoS Comput. Biol.* **2015**, *11*, e1004085. [CrossRef] [PubMed]
134. Rohn, H.; Junker, A.; Hartmann, A.; Grafahrend-Belau, E.; Treutler, H.; Klapperstück, M.; Czauderna, T.; Klukas, C.; Schreiber, F. Vanted v2: A framework for systems biology applications. *BMC Syst. Biol.* **2012**, *6*, 1–13. [CrossRef] [PubMed]
135. Kamburov, A.; Cavill, R.; Ebbels, T.M.D.; Herwig, R.; Keun, H.C. Integrated pathway-level analysis of transcriptomics and metabolomics data with impala. *Bioinformatics* **2011**, *27*, 2917–2918. [CrossRef] [PubMed]
136. Lopez-Ibanez, J.; Pazos, F.; Chagoyen, M. Mbrole 2.0-functional enrichment of chemical compounds. *Nucleic Acids Res.* **2016**, *44*, W201–W204. [CrossRef] [PubMed]
137. Kankainen, M.; Gopalacharyulu, P.; Holm, L.; Oresic, M. Mpea—Metabolite pathway enrichment analysis. *Bioinformatics* **2011**, *27*, 1878–1879. [CrossRef] [PubMed]
138. Tautenhahn, R.; Patti, G.J.; Rinehart, D.; Siuzdak, G. Xcms online: A web-based platform to process untargeted metabolomic data. *Anal. Chem.* **2012**, *84*, 5035–5039. [CrossRef] [PubMed]
139. Winkler, R. An evolving computational platform for biological mass spectrometry: Workflows, statistics and data mining with massypup64. *PeerJ* **2015**, *3*, e1401. [CrossRef] [PubMed]
140. Giacomoni, F.; le Corguille, G.; Monsoor, M.; Landi, M.; Pericard, P.; Petera, M.; Duperier, C.; Tremblay-Franco, M.; Martin, J.F.; Jacob, D.; et al. Workflow4metabolomics: A collaborative research infrastructure for computational metabolomics. *Bioinformatics* **2015**, *31*, 1493–1495. [CrossRef] [PubMed]

141. Mak, T.D.; Laiakis, E.C.; Goudarzi, M.; Fornace, A.J., Jr. Metabolyzer: A novel statistical workflow for analyzing postprocessed LC-MS metabolomics data. *Anal. Chem.* **2014**, *86*, 506–513. [CrossRef] [PubMed]
142. Cascante, M.; Marin, S. Metabolomics and fluxomics approaches. *Essays Biochem.* **2008**, *45*, 67–82. [CrossRef] [PubMed]
143. Cortassa, S.; Aon, M.A. Computational modeling of mitochondrial function. *Methods Mol. Biol.* **2012**, *810*, 311–326. [PubMed]
144. Winter, G.; Kromer, J.O. Fluxomics—Connecting "omics" analysis and phenotypes. *Environ. Microbiol.* **2013**, *15*, 1901–1916. [CrossRef] [PubMed]
145. Aurich, M.K.; Thiele, I. Computational modeling of human metabolism and its application to systems biomedicine. *Methods Mol. Biol.* **2016**, *1386*, 253–281. [PubMed]
146. Cortassa, S.; Caceres, V.; Bell, L.N.; O'Rourke, B.; Paolocci, N.; Aon, M.A. From metabolomics to fluxomics: A computational procedure to translate metabolite profiles into metabolic fluxes. *Biophys. J.* **2015**, *108*, 163–172. [CrossRef] [PubMed]
147. Cho, I.; Blaser, M.J. The human microbiome: At the interface of health and disease. *Nat. Rev. Genet.* **2012**, *13*, 260–270. [CrossRef] [PubMed]
148. Holmes, E.; Loo, R.L.; Stamler, J.; Bictash, M.; Yap, I.K.; Chan, Q.; Ebbels, T.; de Iorio, M.; Brown, I.J.; Veselkov, K.A.; et al. Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature* **2008**, *453*, 396–400. [CrossRef] [PubMed]
149. Garrod, A.E. *The Inborn Factors in Disease*; Clarendon Press: Oxford, UK, 1931.
150. Beebe, K.; Kennedy, A.D. Sharpening precision medicine by a thorough interrogation of metabolic individuality. *Comput. Struct. Biotechnol. J.* **2016**, *14*, 97–105. [CrossRef] [PubMed]
151. Auray-Blais, C.; Maranda, B.; Lavoie, P. High-throughput tandem mass spectrometry multiplex analysis for newborn urinary screening of creatine synthesis and transport disorders, triple H syndrome and otc deficiency. *Clin. Chim. Acta* **2014**, *436*, 249–255. [CrossRef] [PubMed]
152. Pitt, J.J. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *Clin. Biochem. Rev.* **2009**, *30*, 19–34. [PubMed]
153. Pitt, J.J. Newborn screening. *Clin. Biochem. Rev.* **2010**, *31*, 57–68. [PubMed]
154. Pitt, J.J.; Eggington, M.; Kahler, S.G. Comprehensive screening of urine samples for inborn errors of metabolism by electrospray tandem mass spectrometry. *Clin. Chem.* **2002**, *48*, 1970–1980. [PubMed]
155. Spacil, Z.; Tatipaka, H.; Barcenas, M.; Scott, C.R.; Turecek, F.; Gelb, M.H. High-throughput assay of 9 lysosomal enzymes for newborn screening. *Clin. Chem.* **2013**, *59*, 502–511. [CrossRef] [PubMed]
156. Therrell, B.L.; Padilla, C.D.; Loeber, J.G.; Kneisser, I.; Saadallah, A.; Borrajo, G.J.; Adams, J. Current status of newborn screening worldwide: 2015. *Semin. Perinatol.* **2015**, *39*, 171–187. [CrossRef] [PubMed]
157. Janeckova, H.; Hron, K.; Wojtowicz, P.; Hlidkova, E.; Baresova, A.; Friedecky, D.; Zidkova, L.; Hornik, P.; Behulova, D.; Prochazkova, D.; et al. Targeted metabolomic analysis of plasma samples for the diagnosis of inherited metabolic disorders. *J. Chromatogr. A* **2012**, *1226*, 11–17. [CrossRef] [PubMed]
158. Dercksen, M.; Koekemoer, G.; Duran, M.; Wanders, R.J.A.; Mienie, L.J.; Reinecke, C.J. Organic acid profile of isovaleric acidemia: A comprehensive metabolomics approach. *Metabolomics* **2013**, *9*, 765–777. [CrossRef]
159. Ostermann, K.M.; Dieplinger, R.; Lutsch, N.M.; Strupat, K.; Metz, T.F.; Mechtler, T.P.; Kasper, D.C. Matrix-assisted laser desorption/ionization for simultaneous quantitation of (acyl-)carnitines and organic acids in dried blood spots. *Rapid Commun. Mass Spectrom.* **2013**, *27*, 1497–1504. [CrossRef] [PubMed]
160. Fan, M.; Sidhu, R.; Fujiwara, H.; Tortelli, B.; Zhang, J.; Davidson, C.; Walkley, S.U.; Bagel, J.H.; Vite, C.; Yanjanin, N.M.; et al. Identification of niemann-pick c1 disease biomarkers through sphingolipid profiling. *J. Lipid Res.* **2013**, *54*, 2800–2814. [CrossRef] [PubMed]
161. Reinecke, C.J.; Koekemoer, G.; Westhuizen, F.H.; Louw, R.; Lindeque, J.Z.; Mienie, L.J.; Smuts, I. Metabolomics of urinary organic acids in respiratory chain deficiencies in children. *Metabolomics* **2011**, *8*, 264–283. [CrossRef]
162. Smuts, I.; Westhuizen, F.H.; Louw, R.; Mienie, L.J.; Engelke, U.F.H.; Wevers, R.A.; Mason, S.; Koekemoer, G.; Reinecke, C.J. Disclosure of a putative biosignature for respiratory chain disorders through a metabolomics approach. *Metabolomics* **2012**, *9*, 379–391. [CrossRef]
163. Venter, L.; Lindeque, Z.; Jansen van Rensburg, P.; van der Westhuizen, F.; Smuts, I.; Louw, R. Untargeted urine metabolomics reveals a biosignature for muscle respiratory chain deficiencies. *Metabolomics* **2014**, *11*, 111–121. [CrossRef]

164. Wikoff, W.R.; Gangoiti, J.A.; Barshop, B.A.; Siuzdak, G. Metabolomics identifies perturbations in human disorders of propionate metabolism. *Clin. Chem.* **2007**, *53*, 2169–2176. [CrossRef] [PubMed]

165. Auray-Blais, C.; Boutin, M.; Gagnon, R.; Dupont, F.O.; Lavoie, P.; Clarke, J.T. Urinary globotriaosylsphingosine-related biomarkers for fabry disease targeted by metabolomics. *Anal. Chem.* **2012**, *84*, 2745–2753. [CrossRef] [PubMed]

166. Shlomi, T.; Cabili, M.N.; Ruppin, E. Predicting metabolic biomarkers of human inborn errors of metabolism. *Mol. Syst. Biol.* **2009**. [CrossRef] [PubMed]

167. Mutze, U.; Beblo, S.; Kortz, L.; Matthies, C.; Koletzko, B.; Bruegel, M.; Rohde, C.; Thiery, J.; Kiess, W.; Ceglarek, U. Metabolomics of dietary fatty acid restriction in patients with phenylketonuria. *PLoS ONE* **2012**, *7*, e43021. [CrossRef] [PubMed]

168. Pan, Z.; Gu, H.; Talaty, N.; Chen, H.; Shanaiah, N.; Hainline, B.E.; Cooks, R.G.; Raftery, D. Principal component analysis of urine metabolites detected by nmr and desi-ms in patients with inborn errors of metabolism. *Anal. Bioanal. Chem.* **2007**, *387*, 539–549. [CrossRef] [PubMed]

169. Gertsman, I.; Gangoiti, J.A.; Barshop, B.A. Validation of a dual LC-HRMS platform for clinical metabolic diagnosis in serum, bridging quantitative analysis and untargeted metabolomics. *Metabolomics* **2014**, *10*, 312–323. [CrossRef] [PubMed]

170. Miller, M.; Kennedy, A.; Eckhart, A.; Burrage, L.; Wulff, J.; Miller, L.D.; Milburn, M.; Ryals, J.; Beaudet, A.; Sun, Q.; et al. Untargeted metabolomic analysis for the clinical screening of inborn errors of metabolism. *J. Inherit. Metab. Dis.* **2015**, 1–11. [CrossRef] [PubMed]

171. Aygen, S.; Dürr, U.; Hegele, P.; Kunig, J.; Spraul, M.; Schäfer, H.; Krings, D.; Cannet, C.; Fang, F.; Schütz, B.; et al. NMR-based screening for inborn errors of metabolism: Initial results from a study on turkish neonates. *JIMD Rep.* **2014**, *16*, 101–111. [PubMed]

172. Goodacre, R.; Broadhurst, D.; Smilde, A.K.; Kristal, B.S.; Baker, J.D.; Beger, R.; Bessant, C.; Connor, S.; Capuani, G.; Craig, A. Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics* **2007**, *3*, 231–241. [CrossRef]

173. Chitayat, S.; Rudan, J.F. Chapter 10—Phenome centers and global harmonization. In *Metabolic Phenotyping in Personalized and Public Healthcare*; Academic Press: Boston, MA, USA, 2016; pp. 291–315.

174. Kohler, I.; Verhoeven, A.; Derks, R.J.; Giera, M. Analytical pitfalls and challenges in clinical metabolomics. *Bioanalysis* **2016**, *8*, 1509–1532. [CrossRef] [PubMed]

175. Alyass, A.; Turcotte, M.; Meyre, D. From big data analysis to personalized medicine for all: Challenges and opportunities. *BMC Med. Genom.* **2015**, *8*, 1–12. [CrossRef] [PubMed]

176. Tarailo-Graovac, M.; Shyr, C.; Ross, C.J.; Horvath, G.A.; Salvarinova, R.; Ye, X.C.; Zhang, L.H.; Bhavsar, A.P.; Lee, J.J.; Drogemoller, B.I.; et al. Exome sequencing and the management of neurometabolic disorders. *N. Engl. J. Med.* **2016**, *374*, 2246–2255. [CrossRef] [PubMed]

177. Sahoo, S.; Franzson, L.; Jonsson, J.J.; Thiele, I. A compendium of inborn errors of metabolism mapped onto the human metabolic network. *Mol. Biosyst.* **2012**, *8*, 2545–2558. [CrossRef] [PubMed]

Partie II : Travail expérimental

CHAPITRE I : OPTIMISATION ANALYTIQUE EN METABOLOMIQUE NON CIBLEE

1. Introduction

Le terme «optimisation» se réfère à l'amélioration de la performance d’un processus analytique, c'est-à-dire la détermination des conditions expérimentales qui permettent d’obtenir la meilleure réponse. En chimie analytique, l'optimisation est une étape critique pour trouver la valeur que chaque facteur doit avoir pour produire la meilleure réponse possible. L’optimisation doit assurer la bonne performance dans les méthodes analytiques développées en laboratoire, modifiées à partir de méthodes officielles standardisées ou obtenues à partir de la littérature scientifique. Deux stratégies d'optimisation peuvent être distinguées : les approches univariées et les approches multivariées. Dans la première, un seul facteur est modifié à la fois tandis que les autres facteurs restent constants. Cette procédure classiquement appliquée ne tient pas compte des interactions entre les facteurs. De plus, le nombre d'expériences est important lorsque le nombre de facteurs augmente. Dans cette approche le domaine expérimental exploré est généralement plus restreint comparé à celui examiné avec l'approche multivariée. D'autre part, dans la stratégie multivariée, plusieurs facteurs sont étudiés simultanément dans un nombre prédéfini d'expériences, souvent réduit, en variant ensemble les niveaux de tous les facteurs impliqués dans le processus [1]. Dans ce contexte, la conception multivariée des expériences ou plan d’expériences (Design of Experiments DOE) est une démarche séduisante parce qu'elle nécessite moins de temps, d'efforts et de ressources que les procédures univariées qui sont étonnamment encore largement utilisées dans le développement de méthodes de routine. Le DOE facilite la collecte de grandes quantités de données tout en minimisant le nombre d'expériences [2]. Le DOE et la méthodologie de la surface de réponse (RSM) se sont révélés remarquablement utiles pour le développement, l'amélioration et l'optimisation des processus [1]. Quand un grand nombre de réponses doivent être optimisées, la fonction de désirabilité est l'outil le plus souvent appliqué [3]. En chimie analytique, le rôle majeur de la DOE concerne les optimisations de méthodes dont le but principal est de définir les conditions expérimentales qui permettent d'obtenir les meilleures performances analytiques possibles. Deux étapes peuvent être envisagées lors de l'optimisation de la méthode : i) une étape de screening où de nombreux facteurs sont étudiés pour identifier ceux qui ont des effets significatifs sur les variables critiques et ii) l'optimisation, où les facteurs sont examinés afin de déterminer les meilleures conditions analytiques en rapport avec la performance analytique souhaitée. En outre, le DOE est également utilisé en chimie analytique pour évaluer la robustesse dans la validation de la méthode pour examiner les effets que de petites modifications des conditions de la méthode analytique ont sur les réponses et pour construire des matériaux d'étalonnage et de validation à utiliser aux fins d'étalonnage [4]. La procédure multivariée présente plusieurs avantages par rapport à la stratégie univariée. L'approche multivariée permet une compréhension plus globale du système étudié dans tout le domaine expérimental considéré. A partir des résultats obtenus, un modèle mathématique peut être construit pour relier la réponse aux conditions expérimentales choisies. Par ailleurs, la réponse pour tout point du domaine expérimental peut être prédite après une estimation des coefficients du modèle. De plus, le nombre d'expériences est inférieur au nombre d'expériences requises dans l'approche univarié, ce qui réduit le coût, l'effort et le temps. Il est aussi possible d'étudier les interactions entre les facteurs et les relations non linéaires avec les réponses. En général, il est possible de trouver l'optimum absolu dans le domaine étudié, alors que l'approche univariée peut trouver un optimum local qui dépend des conditions initiales de l'analyse.

2. Etapes du design expérimental

- **Définition du problème**

  Il est nécessaire d'avoir une idée claire de la question en cours et des objectifs d'optimisation. La conception expérimentale est un outil qui permet de trouver des solutions à des problèmes analytiques bien définis. L'objectif de l'étude doit être clairement identifié et précisé. Par ailleurs, le temps et le coût de l'expérimentation doivent être évalués.

- **Sélection des variables de réponse**

  Une variable qui peut fournir les informations nécessaires dans l'évaluation de la performance analytique de la méthode doit être sélectionnée pour être soumise à la procédure d'optimisation. Cette variable est appelée réponse et selon l'objectif, il peut être nécessaire d'observer plusieurs réponses. Plusieurs indicateurs analytiques peuvent être utilisés comme variables ou réponses par exemple le rendement d'extraction, le facteur de préconcentration, la surface du pic, la résolution chromatographique, l'écart-type relatif, le temps de migration, de rétention ou le nombre de pics détectés en métabolomique.

- **Sélection des facteurs et de leurs niveaux**

  Tous les facteurs qui peuvent affecter le processus doivent être soigneusement sélectionnés et examinés en fonction de l'expérience de l'analyste et/ou des données de la littérature. Le domaine expérimental doit être défini pour chaque facteur et un mode de contrôle et de mesure doit être établi. Les facteurs peuvent être divisés en quantitatifs, qualitatifs ou liés au mélange pour le volume de solvants. Puisque le nombre de facteurs à considérer peut être important, il est nécessaire d'effectuer des expériences de screening afin de déterminer les variables expérimentales et les interactions qui ont une influence significative et pertinente sur une ou plusieurs réponses. Dans la phase de screening, les facteurs sont habituellement examinés à deux niveaux (-1, +1). L'intervalle entre les niveaux est l'intervalle le plus large dans lequel le facteur peut être modifié pour le système étudié et est choisi sur la base des informations de la littérature ou des connaissances antérieures de l'analyste.

- **Sélection d'un plan expérimental**

  Il convient de s'intéresser aux questions à prendre en considération pour la sélection du meilleur design expérimental pour chaque étape. Ces considérations incluent l'objectif défini : type de problèmes et d'informations connus, le nombre de facteurs et des interactions à étudier, la validité statistique et l'efficacité de chaque design, les limites de fonctionnement, de coût et de temps. Enfin, la facilité de compréhension et de mise en œuvre de la complexité de chaque design.

- **Exécution des expériences et détermination des réponses**

  Dans cette étape, il est recommandé de prendre en compte la considération suivante, les observations et les erreurs doivent être des variables aléatoires indépendantes. Lorsque le nombre d'expériences dépasse la quantité qui peut être réalisée en un temps imparti, les expériences doivent être effectuées en blocs pour prendre en compte l'effet de batch. Il est, évidemment, nécessaire d'appliquer des calculs statistiques pour estimer les effets des facteurs sur une réponse. Une estimation est toujours sujette à erreur et pour décider si l'effet est ou non significatif, l'écart-type de l'effet doit être connu. Cet écart-type est également appelé erreur-type sur un effet et il peut être calculé à partir de l'écart-type d'une mesure expérimentale (erreur expérimentale, erreur aléatoire et incertitude expérimentale). L'erreur aléatoire est due aux causes courantes du processus et représente la variabilité observée de la réponse qui ne peut pas être expliquée à travers les facteurs étudiés. Cette erreur peut inclure l'effet des facteurs non étudiés et des erreurs de l'opérateur commises lors de l'exécution des expériences. Si la variabilité due aux deux derniers cas est importante, il n'est pas possible de distinguer quel est l'effet réel que le facteur étudié exerce sur la réponse. Pour cette raison, il est important que les erreurs de l'opérateur restent faibles ou négligeables pour éviter les variations de tout facteur ayant une influence significative sur la réponse. Si, au contraire, ces erreurs ne sont pas connues auparavant, il existe différentes façons de les estimer. Une première approche consiste à reproduire les points centraux et/ou autres points du design et à estimer la variance. Dans ce cas, il est nécessaire que les répétitions soient mesurées dans des conditions de précision intermédiaire. Les mesures effectuées dans des conditions de répétabilité conduisent à sous-estimer l'erreur expérimentale.

3. Application : Stratégie d'optimisation de la méthodologie analytique (Article V)

Les études de métabolomique incluent des étapes séquentielles et intégrées allant de la question biologique à l'interprétation des données. Elles comprennent le stockage des échantillons, le prétraitement des échantillons, l'acquisition et le traitement des données, la modélisation statistique multivariée, la validation et l'interprétation. Le résultat final dépend fortement de la qualité de chaque étape. La métabolomique étant essentiellement un outil générant des hypothèses à partir des données, des protocoles optimisés et normalisés pour la plupart de ces étapes sont essentiels pour obtenir des informations fiables, reproductibles et interprétables. Les protocoles d'optimisation des méthodes métabolomiques non ciblées précédemment décrits sont basés principalement sur l'évaluation des métabolites endogènes ou des signaux de standards . Cependant, ces approches d'optimisation sont, dans une certaine mesure, restrictives et dépendent des standards internes utilisés ou des métabolites endogènes choisis. Pour conserver la propriété intrinsèque et vitale de la métabolomique non ciblée - qui est de couvrir autant que possible les métabolites présents ajoutés dans l'échantillon étudié - une approche d'optimisation globale est nécessaire pour prendre en compte tous les métabolites détectables. Nous proposons ici une stratégie pour l'optimisation des méthodes non ciblées UHPLC-IM-MS en utilisant une approche de design expérimental pour résoudre ce problème. Comme décrit ci-dessus, le design expérimental est une procédure formalisée dans laquelle des modifications spécifiques et contrôlées sont apportées à un système donné de variables d'entrée afin de créer des modèles mathématiques prédictifs qui permettent l'optimisation des variables de réponse surveillées du système en fonction des modifications appliquées. Le principal avantage de l'utilisation du design expérimental est sa capacité à générer des paramètres expérimentaux optimisés avec un minimum d'expériences. Par conséquent, cette approche est une solution efficace et économique pour la modélisation expérimentale et l'optimisation des méthodes métabolomiques. Le but de l’étude présentée dans l’article IV est d'optimiser les conditions analytiques d'un système LC-IM-MS en utilisant une approche design expérimental. Une nouvelle stratégie simple pour optimiser une méthode de métabolomique non ciblée basée sur l'UHPLC-IM-MS est proposée. Un plan factoriel fractionné à deux niveaux a été utilisé pour étudier et optimiser sept facteurs clés pour les paramètres sources et un plan factoriel complet à deux niveaux a été utilisé pour deux paramètres clés de la mobilité ionique. Pour les paramètres UHPLC-MS, les résultats ont montré que la tension du cône de l'échantillon, la température de désolvatation, le débit de gaz de désolvatation et la tension du cône d'extraction ont les effets positifs les plus significatifs sur le nombre total de pics. De plus, la vitesse des vagues était un facteur important dans l'analyse de l'IMS. Cette étude met en évidence l'avantage de l'utilisation des outils chimiométriques dans le développement et l’optimisation des méthodes métabolomiques non ciblée par spectrométrie de masse. Une compréhension fine des paramètres instrumentaux qui influencent la réponse des analytes en métabolomique est cruciale pour obtenir une réponse optimale et des données robustes et reproductibles.

Ce chapitre est présenté sous forme d’article publié dans le journal Analytica Chimica Acta.

Article V : Tebani A, Schmitz-Afonso I, Rutledge DN, Gonzalez BJ, Bekri S, Afonso C. Optimization of a liquid chromatography ion mobility-mass spectrometry method for untargeted metabolomics using experimental design and multivariate data analysis. Anal Chim Acta. 2016 Mar 24;913:55-62

RÉFÉRENCES

1. Leardi, R. Experimental design in chemistry: A tutorial. *Analytica Chimica Acta* **2009**, *652*, 161-172.
2. Lundstedt, T.; Seifert, E.; Abramo, L.; Thelin, B.; Nyström, Å.; Pettersen, J.; Bergman, R. Experimental design and optimization. *Chemometrics and intelligent laboratory systems* **1998**, *42*, 3-40.
3. Bekele, E.A.; Annaratone, C.E.P.; Hertog, M.L.A.T.M.; Nicolai, B.M.; Geeraerd, A.H. Multi-response optimization of the extraction and derivatization protocol of selected polar metabolites from apple fruit tissue for gc–ms analysis. *Analytica Chimica Acta* **2014**, *824*, 42-56.
4. G. Brereton, R. Multilevel multifactor designs for multivariatecalibration. *The Analyst* **1997**, *122*, 1521-1529.

Optimization of a Liquid Chromatography Ion Mobility-Mass Spectrometry method for untargeted metabolomics using experimental design and multivariate data analysis

Abdellah TEBANI,[1,2,3] * Isabelle SCHMITZ-AFONSO, [1] Douglas N. RUTLEDGE, [4] Bruno J. GONZALEZ, [2] Soumeya BEKRI, [2,3] Carlos AFONSO, [1] *

[1] Normandie Univ, COBRA, UMR 6014 and FR 3038; Université de Rouen; INSA Rouen; CNRS, IRCOF, 1 Rue Tesnière, 76821 Mont-Saint-Aignan Cedex, France

[2] Region-Inserm Team NeoVasc ERI28, Laboratory of Microvascular Endothelium and Neonatal Brain Lesions, Institute of Research for Innovation in Biomedicine, Normandy University, Rouen, France.

[3] Department of Metabolic Biochemistry, Rouen University Hospital, Rouen, France.

[4] UMR Genial, AgroParisTech, INRA, Université Paris-Saclay, 91300, Massy, France

Abstract

High-resolution mass spectrometry coupled with pattern recognition techniques is an established tool to perform comprehensive metabolite profiling of biological datasets. This paves the way for new, powerful and innovative diagnostic approaches in the post-genomic era and molecular medicine. However, interpreting untargeted metabolomic data requires robust, reproducible and reliable analytical methods to translate results into biologically relevant and actionable knowledge. The analyses of biological samples were developed based on ultra-high performance liquid chromatography (UHPLC) coupled to ion mobility - mass spectrometry (IM-MS). A strategy for optimizing the analytical conditions for untargeted UHPLC-IM-MS methods is proposed using an experimental design approach. Optimization experiments were conducted through a screening process designed to identify the factors that have significant effects on the selected responses (total number of peaks and number of reliable peaks). For this purpose, full and fractional factorial designs were used while partial least squares regression was used for experimental design modeling and optimization of parameter values. The total number of peaks yielded the best predictive model and is used for optimization of parameters setting.

Keywords: Experimental design; Mass spectrometry; Ion mobility; Metabolomics, Chemometrics

1. Introduction

The concept of "metabolome" refers to the comprehensive analysis of all metabolites present in a given biological system, fluid, cell or tissue [1,2]. Metabolites can be defined as small organic molecules involved in or resulting from enzymatic reactions. So, metabolomics is one of the "omics" approaches based on biochemical and molecular characterizations of the metabolome and the changes in metabolites related to genetics, environment, drugs or diet and other factors [3]. Two different analytical approaches may be used in metabolomics studies: targeted and untargeted. The targeted approach relies on the measurements of a specific subset of metabolites, focusing typically on pathways of interest. However, the untargeted approach has the advantage of simultaneously measuring as many metabolites as possible in a biological sample. To achieve this goal, different analytical strategies have been developed. Most of them are based on nuclear magnetic resonance (NMR) or mass spectrometry (MS) [4]. However, due to the superior sensitivity of mass spectrometry [5], the predominant analytical methods are nowadays based on hyphenated approaches combining chromatographic separation and MS. In metabolomics, the separation step prior to MS analysis reduces the high biological sample complexity and gives access to the differentiation of isomeric species that cannot be easily done by MS alone. This also decreases ion suppression effects and enhances sensitivity. Liquid and gas chromatography are the most commonly used separation techniques [6]. Recently, approaches using gas phase separation, namely ion mobility spectrometry (IMS) [7] are gaining in interest [8-14]. Indeed, integrated with high resolution mass spectrometry (HRMS) and liquid chromatography (LC-IM-MS), IMS provides additional analyte selectivity without significantly compromising the speed of MS-based measurements. The MS dimension affords accurate mass information while the IMS dimension provides molecular, structural and conformational information. Indeed, combining ion mobility spectrometry with hybrid mass spectrometry instruments offers an additional separation dimension for more comprehensive analysis of complex mixtures [10,15-17]. In addition, the drift time determined from the IMS analysis can be converted to the ion collision cross section (CCS) which is an intrinsic property of the analyzed ion and is therefore a very robust parameter that can be used together with the *m/z* determination for compound identification. Furthermore, having access to retention time, mass and molecular density obtained by the combination of LC-IM-MS allows integration of measurements that enhances molecular identification [18]. UHPLC-IM-MS heat map showing the multidimensionality of the data acquisition is presented in Fig. 1.

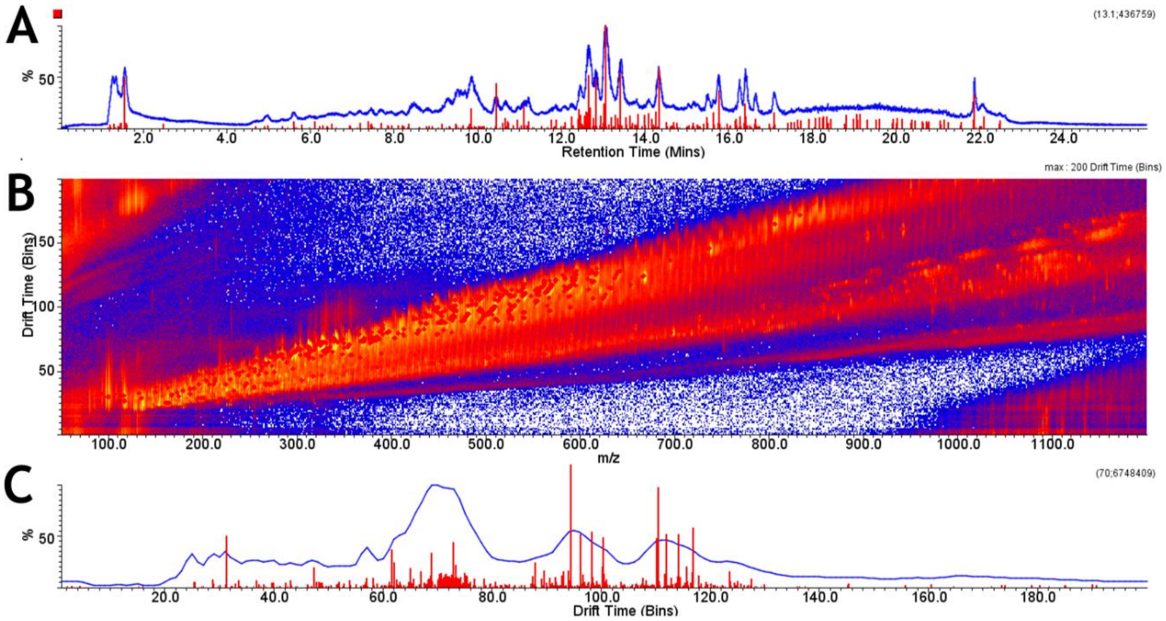


Fig. 1. UHPLC-IM-MS heat map showing the multidimensionality of the data acquisition. A) Chromatographic dimensions (intensity and retention time). B and C) Ion mobility and mass spectrometry dimensions (*m*/*z* and drift time).

So far, HRMS coupled with pattern recognition techniques is an established tool to obtain comprehensive metabolite profiling from biological datasets [19-21]. However, interpreting untargeted metabolomics data requires robust, reproducible and reliable analytical methods to translate results into biologically relevant and actionable knowledge [22]. This paves the way for new, powerful and innovative diagnostic approaches in the post-genomic era [23-25]. Metabolomics studies include sequential and integrated steps spanning from the biological question to data interpretation. It includes sample storage, sample pretreatment, data acquisition and processing, multivariate statistical modeling, validation, and interpretation. The final result is highly dependent on the quality of each step [3]. Metabolomics being primarily a data-driven and hypothesis generating tool, optimized and standardized protocols for most of these steps are essential to retrieve reliable, reproducible and interpretable information from generated data [26]. Previously described optimization protocols of MS based untargeted metabolomic methods are mainly based on the assessment of endogenous metabolites or signals of added standards [27-30]. However, these optimization approaches are, to some extent, restrictive and dependent on the standards used or endogenous metabolites chosen. To keep the intrinsic and vital property of untargeted metabolomics - which is to cover as much as possible metabolites present in the studied sample - a global optimization approach is needed to take into account all the detectable metabolites. Here we propose a strategy for optimizing untargeted UHPLC-IM-MS methods using an experimental design approach to address this issue. Design of experiments (DoE) is a formalized procedure in which specific and controlled modifications are made to a given system of input variables in order to create predictive mathematical models that allow the optimization of the monitored response variables of the system as a function of the applied modifications. DoE may thus be used to explain the system's changes. The main advantage of using DoE is its ability to generate optimized experimental parameters with a minimum of experiments [31-33]. Hence, DoE is an effective and economical solution to experimental modeling and optimization. A general workflow of an experimental design approach is shown in Fig. 2. The aim of the present study was to optimize the analytical conditions of an LC-IM-MS system using a DoE approach.

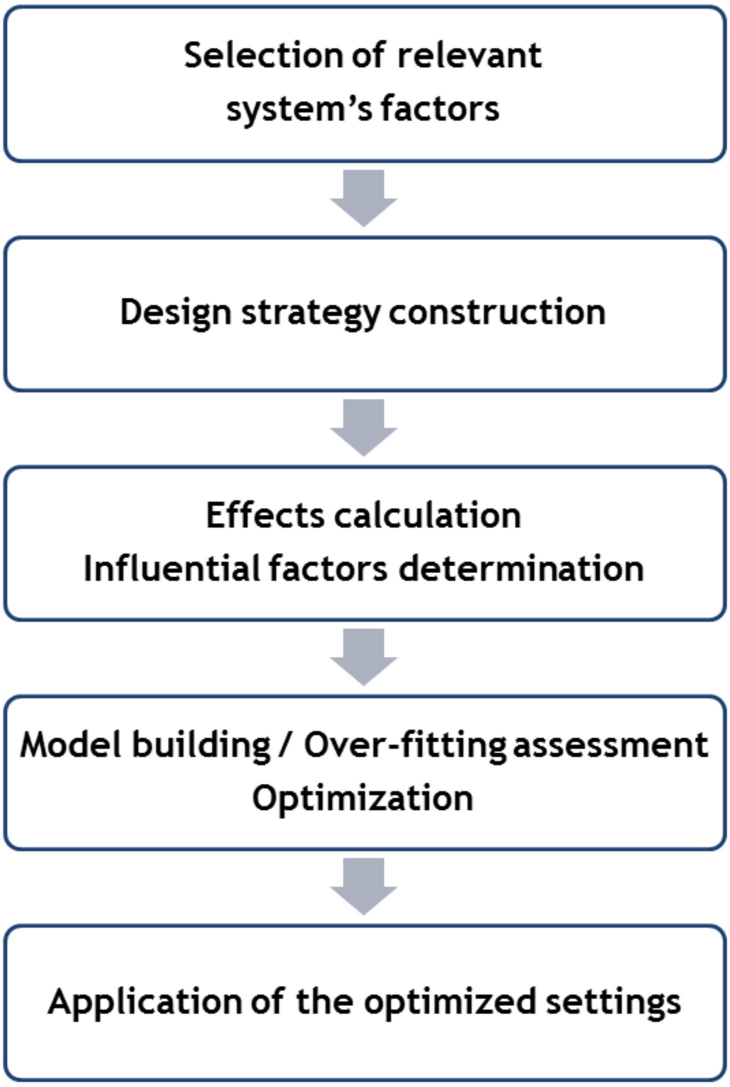

Selection of relevant system's factors
Design strategy construction
Effects calculation
Influential factors determination
Model building / Over-fitting assessment
Optimization
Application of the optimized settings

Fig. 2. Overview of an experimental design workflow.

2. Experimental

2.1. Reagents and chemicals

Methanol and acetonitrile were purchased from VWR Chemicals (France), ultrapure water (18 MX) from Millipore (Molsheim, France) and formic acid from Fluka (Saint Quentin Fallavier, France). The chemicals used were of analytical grade. Leucine Enkephalin (Sigma–Aldrich) at a concentration of 2 ng/L (in acetonitrile/water, 50/50) was used as reference for mass measurements.

2.2. Sample preparation

Four different human sera from volunteers in our laboratory staff were used to create a pooled sample. A small volume of each of the individual samples was mixed into a pooled sample. Two dilutions (1/2 and 1/4) were made from this pool and 100 µL of the pooled sample and dilutions were treated with 300 µL of methanol. The resulting samples were then mixed using a vortex mixer for 20 s, left on ice at 4 °C for 60 min to allow protein precipitation, then centrifuged for 15 min at 15,000 × g. Supernatant of each sample was dried. Dried extracts were suspended with 100 µL of Acetonitrile/Water 50/50.

2.3. Instrumentation

2.3.1. Chromatographic conditions

The chromatography was performed on a Waters NanoAcquity UPLC module (Saint Quentin en Yvelines, France) upgraded to work with 1 mm columns. Separation was carried out at 40 °C using a 1.0 x 100 mm, 1.8 µm Acquity UPLC HSS T3 column (Waters), with a particle size of 1.8 µm, equipped with a 0.2 µm prefilter. Serum was eluted from the LC column using the following linear gradient (curve number 6): 0–2 min: 100% A; 2–15 min: 0–100% B; 15–20 min: 100% B; 20–21 min 0-100% A, 21–26 min 100% A for re-equilibration. Solvent A was water and solvent B was acetonitrile, both solvents contained 0.1% formic acid. Injection volume was set to 3 µL. The chromatography flow rate was optimized.

2.3.2. Mass spectrometry

The U-HPLC system was coupled to a hybrid quadrupole orthogonal time-of-flight (TOF) mass spectrometer (SYNAPT G2 HDMS, Waters MS Technologies, Manchester, UK). The negative mode was investigated. The mass spectrometer was operated in the negative electrospray ionization mode (ESI-). The sample cone voltage, extraction cone voltage, source temperature, desolvation temperature, desolvation gas flow and cone gas flow were optimized. Leucine enkephalin was used as the lock mass $[M-H]^-$ at *m/z* 554.2615. Sodium formate solution was used for external instrument calibration.

2.3.3. Ion mobility

Synapt G2 HDMS (Waters MS Technologies, Manchester, UK) was used in our study for Ion Mobility optimization. It is equipped with a travelling wave "Triwave™" geometry in which the ion mobility cell (IMS T-wave) is placed between two traveling wave ion guides (trap T-wave and transfer T-wave). After ionization in the source and transfer through the quadrupole, the ions arrive at the first travelling-wave ion guide that acts as an ion trap, namely "trap TWIG". In this region, the ions are accumulated before being released in packets and accelerated using the trap-bias voltage to the second cell "IMS-TWIG" for mobility separation. In the IMS-TWIG a travelling wave is continuously applied at a given wave height and velocity to enhance separation through the mobility cell, which is filled with a gas. In this study, the IMS drift gas flow (nitrogen) and the wave velocity settings were assessed and optimized. The helium cell gas flow, wave height, Trap Bias and IMS wave delay were set at 180 mL/min, 40 V, 45 V and 450 µs respectively. The optimized settings of sample cone voltage, extraction cone voltage, source temperature, desolvation temperature, desolvation gas flow, cone gas flow and chromatography flow rate were used for this step. The TOF analyzer was operated in the *V* resolution mode with an average mass resolution of m/Δm 20,000 (full width at half-maximum definition). Data acquisition of an ion mobility experiment consisted of 200 bins. Data acquisition and processing were performed using MassLynx v4.1 and Driftscope v2.2.

2.4. Raw data pretreatment

Raw data files were converted into NetCDF format (Network Common Data Format) using the Waters DataBridge software. All LC–MS data were processed using XCMS [34,35] to yield a data matrix containing retention times, accurate masses and peak intensities. Peak detection and peak matching across samples were performed using the centWave algorithm [36]. Retention time (tR) correction and chromatographic alignment were performed using the OBI-warp method [37] before applying peak-picking on the integrated XICs (Extracted Ion Chromatograms). The preprocessing step resulted in an X-matrix where tR and *m*/*z* values were concatenated into ''tR_m/z'' features (in columns) present in each sample (in rows) with corresponding peak areas. The XCMS parameters were optimized using the Isotopologue Parameter Optimization (IPO) package implemented in the R environment [38]. The used parameters are Peakwidth = c(5, 25), ppm = 20, noise = 0, snthresh = 10, mzdiff = 0.0045, prefilter = c(3, 100), bw=0.25, mzwid=0.0266, minfrac=0.5, minsamp=1, max=50. IM-MS data were processed using DriftScope V2.2.

2.5. Experimental design

2.5.1. Optimization protocol

A pooled sample is used in our proposed DoE optimization approach. Using a pooled sample ensures that all metabolites are present during the optimization procedure. Thus, by diluting this pooled sample, an associated concentration vector can be calculated that can be used to separate reliable peaks from unreliable ones, since for informative features the relationship between peak area and sample concentration are expected to be close to linear [31]. A small volume of each of the individual serum samples was mixed into a pooled sample. Three dilution levels were analyzed with UHPLC-MS, using the different parameter settings as defined by the DoE 1. This generated raw data in a three-dimensional array (sample, retention time, *m*/*z*). Then, the raw data were processed with XCMS using optimized parameter settings. Each UHPLC-MS setting generated a two-dimensional table (sample, peaks) with integrated peak areas. For each UHPLC-MS setting and each peak, the coefficient of determination, R^2, between peak area and concentration vector was calculated. For each UHPLC-MS setting, the total number of detected peaks, the number of reliable peaks, *i.e.* those with high correlations ($R^2 \geq 0.9$), were counted, generating two response vectors for each setting. Example plots of included and excluded features are presented in Fig. SI-1. The quality of each processed data set and corresponding parameter settings were then assessed by evaluating two different responses: total number of peaks and number of peaks with $R^2 \geq 0.9$.

2.5.2. Factorial designs

The optimization protocol was divided into two steps. The first tackled the optimization of UHPLC-MS parameters. The second step addressed the ion mobility separation parameters. For UHPLC-MS parameters, the applied chemometric approach is based on a fractional factorial design (FFD) with resolution IV including the selected parameters (DoE 1) as presented in Table 1. The factor levels were chosen based on experimental knowledge of the instrument. Detailed information of the FFD used are shown in Table SI-1. Data descriptive statistics are shown in Table SI-7A. Fractional factorial designs are designs of experiments consisting of a fraction of the runs of a full-factorial design. They allow to perform fewer experimental runs to determine the most important variables [39]. In this paper, the FFD was applied to determine the influence of the operational factors on the chosen responses. Designs with factors that are set at two levels implicitly assume that the effect of the factors on the dependent variable of interest is linear within the range of the two levels. So, including runs where all factors are set at their center point has two main purposes. On the one hand, it allows to check the system stability and its inherent variability. On the other hand, it allows to check the curvature, to know whether the relation between the response and the factors is in fact linear. Such runs are called center point runs since they are, in a sense, in the center of the design. Four central points were

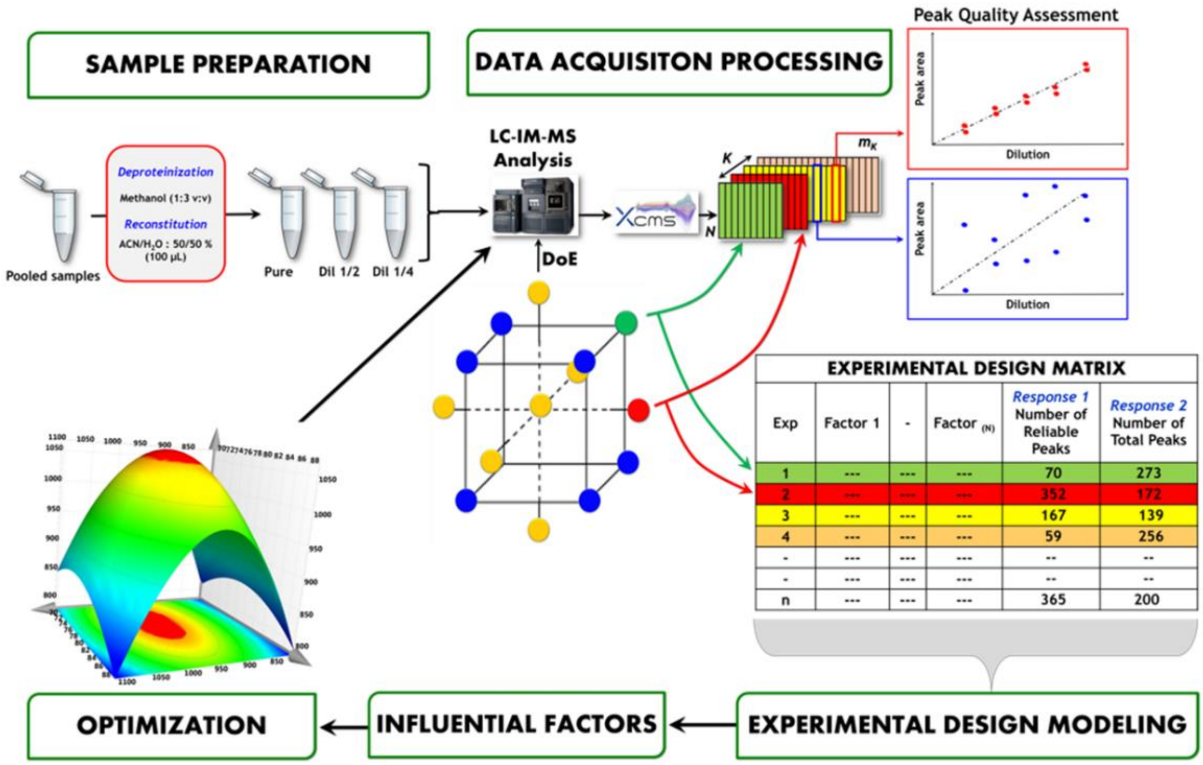
included in our design. Therefore, a total of 20 experiments were performed according to the FFD, with sixteen experimental runs plus four central points.

Table 1. Coded Fractional Factorial design matrix for LC-MS optimization (DoE 1).

Run N°	Run Order	Sample cone voltage (V)	Source Temperature (°C)	Desolvation Temperature (°C)	Desolvation Gas (L/h)	Cone Gas Flow (L h^{-1})	LC Flow rate (µL min^{-1})	Extraction Voltage (V)
1	20	-1	-1	-1	-1	-1	-1	-1
2	9	1	-1	-1	-1	-1	1	1
3	17	-1	1	-1	-1	1	-1	1
4	10	1	1	-1	-1	1	1	-1
5	6	-1	-1	1	-1	1	1	1
6	16	1	-1	1	-1	1	-1	-1
7	13	-1	1	1	-1	-1	1	-1
8	4	1	1	1	-1	-1	-1	1
9	19	-1	-1	-1	1	1	1	-1
10	15	1	-1	-1	1	1	-1	1
11	1	-1	1	-1	1	-1	1	1
12	2	1	1	-1	1	-1	-1	-1
13	14	-1	-1	1	1	-1	-1	1
14	7	1	-1	1	1	-1	1	-1
15	11	-1	1	1	1	1	-1	-1
16	12	1	1	1	1	1	1	1
17	18	0	0	0	0	0	0	0
18	3	0	0	0	0	0	0	0
19	5	0	0	0	0	0	0	0
20	8	0	0	0	0	0	0	0

2.5.3. Investigated factors and multivariate data modeling

The experimental data were modeled by Partial Least Squares regression (PLS). PLS was used to determine the optimal parameter settings with regards to each chosen response using the optimizer implemented in MODDE v10 (MKS Umetrics). The output from the optimizer is the parameter setting combination yielding the best response value. The DoE strategy is shown in Fig. 3.



EXPERIMENTAL DESIGN MATRIX

Exp	Factor 1	-	Factor (N)	Response 1 Number of Reliable Peaks	Response 2 Number of Total Peaks
1	---	---	---	70	273
2	---	---	---	352	172
3	---	---	---	167	139
4	---	---	---	59	256
-	---	---	---	--	--
-	---	---	---	--	--
n	---	---	---	365	200

Fig. 3. Experimental design strategy. The pooled sample is processed then diluted to different concentrations. The samples in the dilution series are analyzed according to the corresponding DoE and then processed in XCMS. From the resulting data sets, one for each design experiment, the peaks are binned according to the correlation between peak area and the dilution factor of the samples in the dilution series. The correlation values for each data set are then used to calculate the number of either total or reliable peaks. The parameters of the DoE are then optimized with respect to maximizing the response.

The selection of the investigated factors and their range is based on prior knowledge of the liquid chromatography-mass spectrometry system and on experimental constraints. Thus, the effects of seven factors were investigated: chromatographic flow rate is related to chromatographic separation efficiency; source temperature, desolvation temperature, desolvation gas flow and cone gas flow are related to ion desolvation efficiency; sample cone voltage and extraction cone voltage are related to ion transmission efficiency. In particular, the upper value for sample cone voltage was chosen to minimize the in-source fragmentations. The chromatographic flow rate was limited to 80 µL/min due to the system high pressure limit. The levels chosen for the factors, in coded and real values, are shown in Table 2. PLS coefficients are used to assess the factor effects. To make the coefficients comparable between responses, the centered, scaled and normalized coefficient values are used. Negative values indicate that increasing the value of this term decreases the response, and positive values indicate an increase in response upon increasing the corresponding term. The greater the absolute value of a term is, the greater is its influence on analyte response. Several methods can be used to assess the significance of the factor effect. One method is to use the sum of squares to perform an analysis of variance (ANOVA) on the PLS regression coefficients. For a statistically significant effect of a factor, the model p-value of ANOVA should not exceed 0.05 [39].

Table 2. Factors and factor levels investigated during the optimization study.

Factors			Coded values		
	Abbreviation	Unit	-1	0	+1
DoE 1					
Sample Cone Voltage	Samp	V	10	19	35
Extraction Cone Voltage	Extr	V	2	3	5
Source Temperature	SourT	°C	100	123	150
Desolvation Temperature	DesT	°C	300	424	600
Desolvation Gas flow	DesG	L/h	400	632	1000
Cone Gas flow	Cone	L/h	10	22	50
LC Flow rate	Flow	µL/min	50	63	80
DoE 2					
Drift gas flow	GFlow	mL/min	70	80	90
Wave velocity	Wave	m/s	700	850	1000

2.6. Software and data analysis

All data analyzes and modeling were done using MODDE 10.0 (Umetrics, Umeå, Sweden). MODDE 10.0 was also used to generate the different experimental design matrices. The responses (Y matrix) values were log transformed and scaled to unit variance. PLS1 was used to relate the experimental design matrix X to the responses (Y matrix). A single cross-validation with seven cross-validation groups was used throughout to determine the number of components. R^2X is the cumulative modeled variation in X, R^2Y is the cumulative modeled variation in Y, and Q^2Y is the cumulative predicted variation in Y, based on the cross-validation. The range of these parameters is 0-1, where 1 indicates a perfect fit [40].

3. Results and discussion

3.1. LC-MS parameters settings optimization

In our study, sample cone voltage, desolvation temperature, desolvation gas flow and extraction cone voltage showed the most significant positive effects on the total number of peaks. Flow rate, source temperature and cone gas had no significant effect on this response. Regarding number of reliable peaks, only desolvation gas flow showed a significant negative effect, meaning that other parameters have little influence of the number of reliable peaks and do not bring any statistically significant difference on this number. The effects on the different responses are shown in Fig. 4. Contour plots are shown in Fig. SI-2 for total number peaks as a function of the different factors assessed (Supplementary data). The values of the coded regression coefficients and their p-values are given in Table SI-2 (Supplementary data). The fitted model and the experimental data are within 95% confidence level. The fitted models and ANOVA results of each model are presented in Table SI-3 (Supplementary data). In Fig. 3, the prediction plot between the predicted and observed values of total number of peaks and number of reliable peaks shows, for 20 data points, a cumulative modeled variation (R^2X) value of 88% and 66% and a cumulative predicted variation (Q^2Y) value of 42% and 38% respectively, indicating good agreement. According to the model performance assessment, total number of peaks yielded the best predictive model with three principal components. Hence it was used to select the influential parameters and optimize the parameter settings. These results are in agreement with the experimental knowledge on ionization mechanisms. Thus, desolvation temperature and desolvation gas flow are directly related to the efficiency of the ion desolvation. Optimum voltages, such as sample cone voltage and extraction cone voltage, vary depending on the ion transmission efficiency and ion gas phase stability.

Our method allows to optimize these voltages in a non-targeted way. Optimized parameter settings are presented in Table 3.

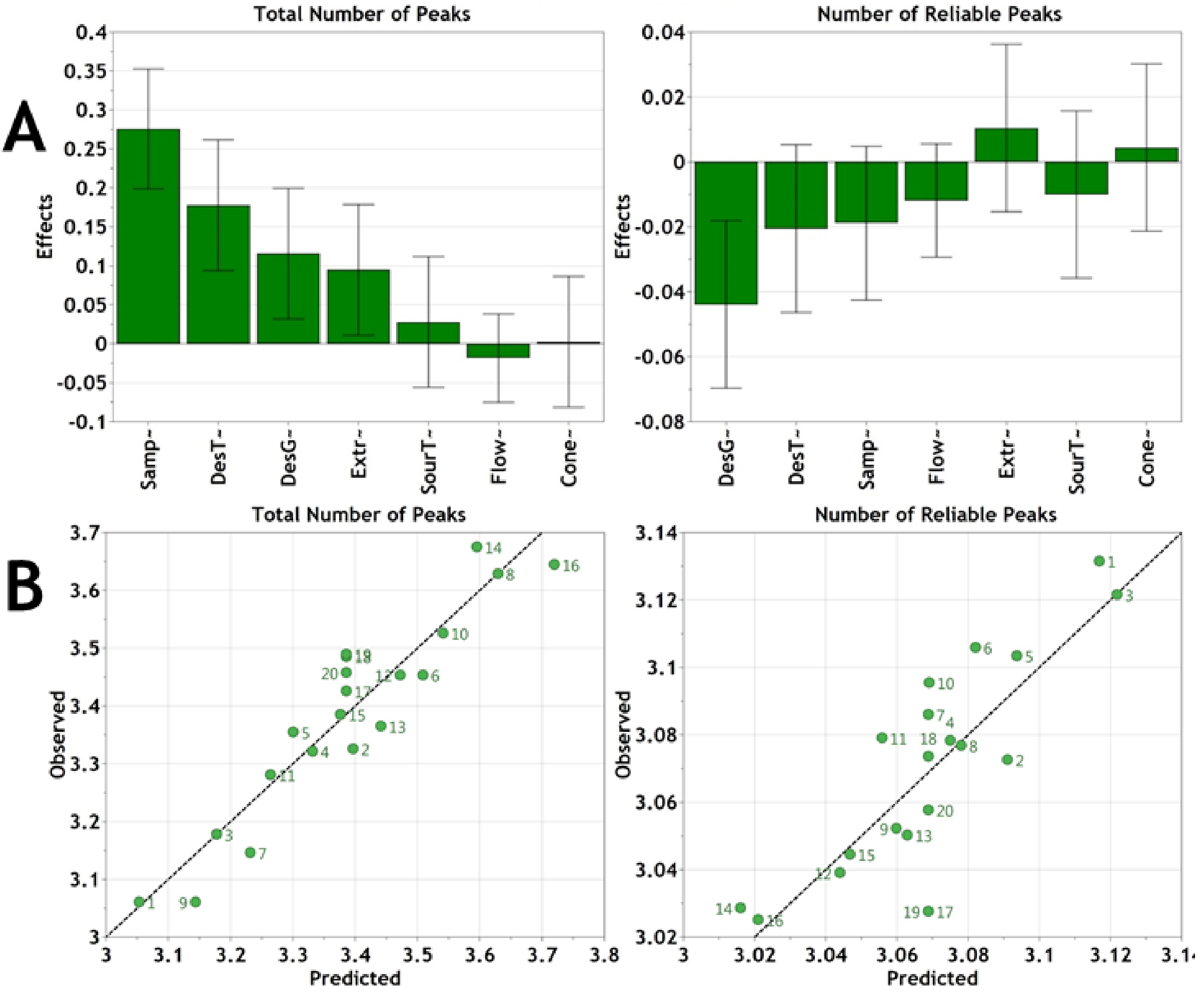


Fig. 4. A) Representation of the effects of the different factors on the total number of peaks and the number of reliable peaks. B) Prediction plot between the predicted and observed log-transformed values of total number of peaks and number of reliable peaks shows a correlation coefficient (R^2) value of 88% and 66% respectively for 20 data points, indicating good agreement.

3.2. Ion mobility parameters settings optimization

Regarding ion mobility settings optimization, a second experimental design was performed based on UHPLC-IM-MS analyzes. A full-factorial design including selected parameters was created (DoE 2) as presented in Table 4. Data descriptive statistics are shown in Table SI-7B. Five center points were used. When conditions are not optimized, the ions will not traverse through the ion mobility cell effectively, and their journey may take longer than the time required for the next ion packet to be released into the ion mobility cell. As a result, a new ion packet will be released from the Trap region before the previous packet has been released to the pusher region. This will lead to a 'wrap-around' effect, in which the peaks observed in the first part of the ion mobility spectrum is identical to those in the tailing edge. Thus, two parameters have been assessed; IMS drift gas flow (nitrogen) and the wave velocity. These parameters have been selected as they are the most important in regard to ion mobility separation efficiency [41]. The number of detected peaks, using the DriftScope software, was chosen as a response factor. The number of detected peaks in the ion mobility was chosen to avoid the ’wrap-around’ effect. Fig. SI-6 shows an example of the selection strategy. The levels chosen for these factors, in coded and real values, are shown in Table 2. A detailed matrix design is presented in Table SI-4 (Supplementary data). Based on the acquired data, a model with two components was built and showed a cumulative modeled variation (R^2X) and a cumulative predicted variation (Q^2Y) value of 91% and 57% respectively for 13 data points, indicating a good agreement as presented in Fig. SI-4 and Fig. SI-5. The values of the coded regression coefficients and their *p*-values are given in Table SI-5 (Supplementary data). The fitted model and the experimental data are within 95% confidence limits. The fitted models and ANOVA results for each model are presented in Table SI-6 (Supplementary data). Based on this model, the effect Pareto plot, presented in Fig. 5, shows a significant effect of wave velocity on the total number of detected peaks. However, drift gas flow showed a lower effect in the studied experimental conditions. As for UHPLC-MS parameters, our method allows a non-targeted optimization of the wave velocity for a high number of variables. It is to be noted that a higher wave velocity offers a better resolution for small molecules at the expense of higher *m*/*z* ratios. Contour plots of the investigated parameters are presented in Fig. SI-7. The yielded model was used for optimization and the optimized settings for IMS parameters are presented in Table 3.

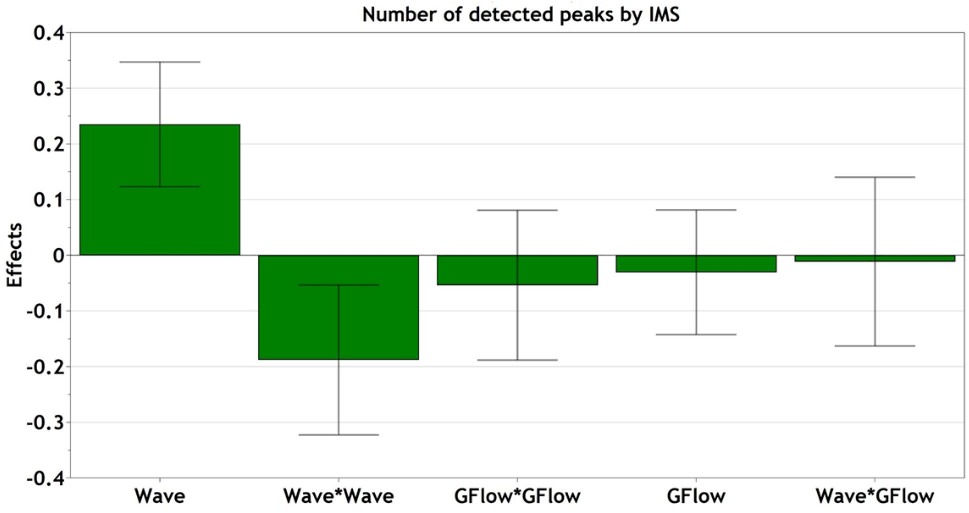


Fig. 5. Representation of the effects of the wave velocity and drift gas flow on the number of peaks detected by ion mobility spectrometry.

Table 3. Optimized parameter settings in negative mode.

Factor	Abbreviation	Unit	Value
Sample Cone Voltage	Samp	V	35
Extraction Cone Voltage	Extr	V	5
Source Temperature	SourT	°C	120
Desolvation Temperature	DesT	°C	330
Desolvation Gas flow	DesG	L/h	400
Cone Gas flow	Cone	L/h	50
LC Flow rate	Flow	µL/min	80
Wave velocity	Wave	m/s	954
Drift gas flow	DGFlow	mL/min	80

Table 4. Coded Full Factorial design matrix for IMS optimization (DoE 2).

Run N°	Run Order	Wave Velocity (m/s)	Drift gas flow mL/min
1	2	-1	-1
2	13	1	-1
3	1	-1	1
4	8	1	1
5	5	-1	0
6	3	1	0
7	12	0	-1
8	10	0	1
9	7	0	0
10	4	0	0
11	11	0	0
12	9	0	0
13	6	0	0

4. Conclusion

A new and simple strategy to optimize an untargeted UHPLC-IM-MS based metabolomics method based on an experimental design approach is proposed. A two-level fractional factorial design was used to study and optimize seven key factors for source parameters and a two-level full-factorial design for two ion mobility key settings. For UHPLC-MS parameters, the design showed that sample cone voltage, desolvation temperature, desolvation gas flow and extraction cone voltage have the most significant positive effects on the total number of peaks. However, wave velocity was an important factor in IMS analysis. This study highlights the benefit of using chemometrics for efficient experimental designs in untargeted mass spectrometry based metabolomics method development. Simple factorial designs are demonstrated to be useful to elucidate the influence of instrumental parameters in untargeted metabolomics studies. It is to be noted that fractional factorial designs are very efficient in reducing the number of experiments when the number of factors is large. It was noted in this study that it was effective to increase the overall ion mobility metabolome coverage at the expense of high CCS features. A good understanding of the way in which instrumental parameters influence analyte response in metabolomics is crucial in order to obtain optimal response, robust and reproducible data

Acknowledgements

The authors gratefully acknowledge the Region Haute-Normandie, the Labex SynOrg (ANR-11-LABX-0029) and the European Regional Development Fund (ERDF 31708) for financial support.

RÉFÉRENCES

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; the fate of foreign compounds in biological systems* **1999**, *29*, 1181-1189.
2. Fiehn, O. Metabolomics--the link between genotypes and phenotypes. *Plant molecular biology* **2002**, *48*, 155-171.
3. Beisken, S.; Eiden, M.; Salek, R.M. Getting the right answers: Understanding metabolomics challenges. *Expert review of molecular diagnostics* **2014**, 1-13.
4. Alonso, A.; Marsal, S.; Julia, A. Analytical methods in untargeted metabolomics: State of the art in 2015. *Frontiers in bioengineering and biotechnology* **2015**, *3*, 23.
5. Emwas, A.-H.; Salek, R.; Griffin, J.; Merzaban, J. Nmr-based metabolomics in human disease diagnosis: Applications, limitations, and recommendations. *Metabolomics* **2013**, *9*, 1048-1072.
6. Cajka, T.; Fiehn, O. Toward merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics. *Analytical chemistry* **2015**.
7. Hill, H.H., Jr.; Siems, W.F.; St Louis, R.H.; McMinn, D.G. Ion mobility spectrometry. *Analytical chemistry* **1990**, *62*, 1201A-1209A.
8. Paglia, G.; Angel, P.; Williams, J.P.; Richardson, K.; Olivos, H.J.; Thompson, J.W.; Menikarachchi, L.; Lai, S.; Walsh, C.; Moseley, A., *et al.* Ion mobility-derived collision cross section as an additional measure for lipid fingerprinting and identification. *Analytical chemistry* **2015**, *87*, 1137-1144.
9. Maldini, M.; Natella, F.; Baima, S.; Morelli, G.; Scaccini, C.; Langridge, J.; Astarita, G. Untargeted metabolomics reveals predominant alterations in lipid metabolism following light exposure in broccoli sprouts. *Int J Mol Sci* **2015**, *16*, 13678-13691.
10. Paglia, G.; Williams, J.P.; Menikarachchi, L.; Thompson, J.W.; Tyldesley-Worster, R.; Halldórsson, S.; Rolfsson, O.; Moseley, A.; Grant, D.; Langridge, J., *et al.* Ion mobility derived collision cross sections to support metabolomics applications. *Analytical chemistry* **2014**, *86*, 3985-3993.
11. Wickramasekara, S.I.; Zandkarimi, F.; Morre, J.; Kirkwood, J.; Legette, L.; Jiang, Y.; Gombart, A.F.; Stevens, J.F.; Maier, C.S. Electrospray quadrupole travelling wave ion mobility time-of-flight mass spectrometry for the detection of plasma metabolome changes caused by xanthohumol in obese zucker (fa/fa) rats. *Metabolites* **2013**, *3*, 701-717.
12. Dwivedi, P.; Schultz, A.J.; Hill, H.H. Metabolic profiling of human blood by high resolution ion mobility mass spectrometry (im-ms). *Int J Mass Spectrom* **2010**, *298*, 78-90.
13. Hauschild, A.C.; Frisch, T.; Baumbach, J.I.; Baumbach, J. Carotta: Revealing hidden confounder markers in metabolic breath profiles. *Metabolites* **2015**, *5*, 344-363.
14. Smolinska, A.; Hauschild, A.C.; Fijten, R.R.; Dallinga, J.W.; Baumbach, J.; van Schooten, F.J. Current breathomics--a review on data pre-processing techniques and machine learning in metabolomics breath analysis. *J Breath Res* **2014**, *8*, 027105.
15. Fenn, L.; Kliman, M.; Mahsut, A.; Zhao, S.; McLean, J. Characterizing ion mobility-mass spectrometry conformation space for the analysis of complex biological samples. *Analytical and Bioanalytical Chemistry* **2009**, *394*, 235-244.
16. Fenn, L.; McLean, J. Biomolecular structural separations by ion mobility–mass spectrometry. *Analytical and Bioanalytical Chemistry* **2008**, *391*, 905-909.
17. Kliman, M.; May, J.C.; McLean, J.A. Lipid analysis and lipidomics by structurally selective ion mobility-mass spectrometry. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **2011**, *1811*, 935-945.
18. May, J.C.; Goodwin, C.R.; McLean, J.A. Ion mobility-mass spectrometry strategies for untargeted systems, synthetic, and chemical biology. *Curr Opin Biotechnol* **2015**, *31*, 117-121.
19. Jonsson, P.; Bruce, S.J.; Moritz, T.; Trygg, J.; Sjostrom, M.; Plumb, R.; Granger, J.; Maibaum, E.; Nicholson, J.K.; Holmes, E., *et al.* Extraction, interpretation and validation of information for comparing samples in metabolic lc/ms data sets. *The Analyst* **2005**, *130*, 701-707.
20. Junot, C.; Fenaille, F.; Colsch, B.; Bécher, F. High resolution mass spectrometry based techniques at the crossroads of metabolic pathways. *Mass spectrometry reviews* **2014**, *33*, 471-500.
21. Boccard, J.; Rudaz, S. Harnessing the complexity of metabolomic data with chemometrics. *Journal of Chemometrics* **2014**, *28*, 1-9.
22. Dunn, W.B.; Wilson, I.D.; Nicholls, A.W.; Broadhurst, D. The importance of experimental design and qc samples in large-scale and ms-driven untargeted metabolomic studies of humans. *Bioanalysis* **2012**, *4*, 2249-2264.
23. Yin, P.; Xu, G. Current state-of-the-art of nontargeted metabolomics based on liquid chromatography-mass spectrometry with special emphasis in clinical applications. *Journal of chromatography. A* **2014**, *1374*, 1-13.

24. Balog, J.; Sasi-Szabo, L.; Kinross, J.; Lewis, M.R.; Muirhead, L.J.; Veselkov, K.; Mirnezami, R.; Dezso, B.; Damjanovich, L.; Darzi, A., *et al.* Intraoperative tissue identification using rapid evaporative ionization mass spectrometry. *Sci. Transl. Med.* **2013**, *5*, 194ra193.
25. Nicholson, J.K.; Holmes, E.; Kinross, J.M.; Darzi, A.W.; Takats, Z.; Lindon, J.C. Metabolic phenotyping in clinical and surgical environments. *Nature* **2012**, *491*, 384-392.
26. Reinke, S.; Broadhurst, D. Moving metabolomics from a data-driven science to an integrative systems science. *Genome Medicine* **2012**, *4*, 85.
27. Danielsson, A.H.; Moritz, T.; Mulder, H.; Spégel, P. Development of a gas chromatography/mass spectrometry based metabolomics protocol by means of statistical experimental design. *Metabolomics* **2012**, *8*, 50-63.
28. Pereira, H.; Martin, J.-F.; Joly, C.; Sébédio, J.-L.; Pujos-Guillot, E. Development and validation of a uplc/ms method for a nutritional metabolomic study of human plasma. *Metabolomics* **2010**, *6*, 207-218.
29. Riter, L.S.; Vitek, O.; Gooding, K.M.; Hodge, B.D.; Julian, R.K. Statistical design of experiments as a tool in mass spectrometry. *Journal of Mass Spectrometry* **2005**, *40*, 565-579.
30. Zhou, Y.; Song, J.-Z.; Choi, F.F.-K.; Wu, H.-F.; Qiao, C.-F.; Ding, L.-S.; Gesang, S.-L.; Xu, H.-X. An experimental design approach using response surface techniques to obtain optimal liquid chromatography and mass spectrometry conditions to determine the alkaloids in meconopsi species. *Journal of Chromatography A* **2009**, *1216*, 7013-7023.
31. Eliasson, M.; Rannar, S.; Madsen, R.; Donten, M.A.; Marsden-Edwards, E.; Moritz, T.; Shockcor, J.P.; Johansson, E.; Trygg, J. Strategy for optimizing lc-ms data processing in metabolomics: A design of experiments approach. *Analytical chemistry* **2012**, *84*, 6869-6876.
32. Gerretzen, J.; Szymańska, E.; Jansen, J.J.; Bart, J.; van Manen, H.-J.; van den Heuvel, E.R.; Buydens, L.M. Simple and effective way for data preprocessing selection based on design of experiments. *Analytical chemistry* **2015**, *87*, 12096-12103.
33. Eriksson, L. *Design of experiments: Principles and applications*. Umetrics: 2008.
34. 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. *Analytical chemistry* **2006**, *78*, 779-787.
35. Benton, H.P.; Want, E.J.; Ebbels, T.M. Correction of mass calibration gaps in liquid chromatography-mass spectrometry metabolomics data. *Bioinformatics (Oxford, England)* **2010**, *26*, 2488-2489.
36. Tautenhahn, R.; Bottcher, C.; Neumann, S. Highly sensitive feature detection for high resolution lc/ms. *BMC bioinformatics* **2008**, *9*, 504.
37. Prince, J.T.; Marcotte, E.M. Chromatographic alignment of esi-lc-ms proteomics data sets by ordered bijective interpolated warping. *Analytical chemistry* **2006**, *78*, 6140-6152.
38. Libiseller, G.; Dvorzak, M.; Kleb, U.; Gander, E.; Eisenberg, T.; Madeo, F.; Neumann, S.; Trausinger, G.; Sinner, F.; Pieber, T., *et al.* Ipo: A tool for automated optimization of xcms parameters. *BMC bioinformatics* **2015**, *16*, 1-10.
39. Montgomery, D.C. *Design and analysis of experiments*. John Wiley & Sons: 2008.
40. Eriksson, L.; Trygg, J.; Wold, S. A chemometrics toolbox based on projections and latent variables. *Journal of Chemometrics* **2014**, *28*, 332-346.
41. Michaelevski, I.; Kirshenbaum, N.; Sharon, M. T-wave ion mobility-mass spectrometry: Basic experimental procedures for protein complex analysis. *J Vis Exp* **2010**, 1985.

Supplementary material for:

Optimization of a Liquid Chromatography Ion Mobility-Mass Spectrometry method for untargeted metabolomics using experimental design and multivariate data analysis

Abdellah TEBANI,[1,2,3]* Isabelle SCHMITZ-AFONSO,[1] Douglas N. RUTLEDGE,[4] Bruno J. GONZALEZ,[2] Soumeya BEKRI,[2,3] Carlos AFONSO,[1]*

[1] Normandie Univ, COBRA, UMR 6014 and FR 3038; Université de Rouen; INSA Rouen; CNRS, IRCOF, 1 Rue Tesnière, 76821 Mont-Saint-Aignan Cedex, France

[2] Region-Inserm Team NeoVasc ERI28, Laboratory of Microvascular Endothelium and Neonatal Brain Lesions, Institute of Research for Innovation in Biomedicine, Normandy University, Rouen, France.

[3] Department of Metabolic Biochemistry, Rouen University Hospital, Rouen, France.

[4] UMR Genial, AgroParisTech, INRA, Université Paris-Saclay, 91300, Massy, France

Contents

Tables

Figures

Table SI-1A. Fractional Factorial design matrix with responses values (DoE 1).

Run N°	Run Order	Sample cone voltage (V)	Source Temperature (°C)	Desolvation Temperature (°C)	Desolvation Gas (L/h)	Cone Gas Flow (L/h)	LC Flow rate (µL/min)	Extraction Voltage (V)	Number of Reliable Peaks	Total Number of Peaks
1	20	10	100	300	400	10	50	2	1152	1354
2	9	35	100	300	400	10	80	5	1182	2115
3	17	10	150	300	400	50	50	5	1323	1507
4	10	35	150	300	400	50	80	2	1198	2096
5	6	10	100	600	400	50	80	5	1269	2261
6	16	35	100	600	400	50	50	2	1276	2839
7	13	10	150	600	400	10	80	2	1219	1401
8	4	35	150	600	400	10	50	5	1194	4252
9	19	10	100	300	1000	50	80	2	1128	1151
10	15	35	100	300	1000	50	50	5	1246	3356
11	1	10	150	300	1000	10	80	5	1200	1912
12	2	35	150	300	1000	10	50	2	1094	2841
13	14	10	100	600	1000	10	50	5	1123	2322
14	7	35	100	600	1000	10	80	2	1068	4733
15	11	10	150	600	1000	50	50	2	1108	2431
16	12	35	150	600	1000	50	80	5	1060	4421
17	18	19	122	424	632	22	63	3	1066	2668
18	3	19	122	424	632	22	63	3	1185	3056
19	5	19	122	424	632	22	63	3	1066	3091
20	8	19	122	424	632	22	63	3	1142	2872

Table SI-1B. Coded Fractional Factorial design matrix with responses values (DoE 1).

Run N°	Run Order	Sample cone voltage (V)	Source Temperature (°C)	Desolvation Temperature (°C)	Desolvation Gas (L/h)	Cone Gas Flow (L/h)	LC Flow rate (µL/min)	Extraction Voltage (V)	Number of Reliable Peaks	Total Number of Peaks
1	20	-1	-1	-1	-1	-1	-1	-1	1152	1354
2	9	1	-1	-1	-1	-1	1	1	1182	2115
3	17	-1	1	-1	-1	1	-1	1	1323	1507
4	10	1	1	-1	-1	1	1	-1	1198	2096
5	6	-1	-1	1	-1	1	1	1	1269	2261
6	16	1	-1	1	-1	1	-1	-1	1276	2839
7	13	-1	1	1	-1	-1	1	-1	1219	1401
8	4	1	1	1	-1	-1	-1	1	1194	4252
9	19	-1	-1	-1	1	1	1	-1	1128	1151
10	15	1	-1	-1	1	1	-1	1	1246	3356
11	1	-1	1	-1	1	-1	1	1	1200	1912
12	2	1	1	-1	1	-1	-1	-1	1094	2841
13	14	-1	-1	1	1	-1	-1	1	1123	2322
14	7	1	-1	1	1	-1	1	-1	1068	4733
15	11	-1	1	1	1	1	-1	-1	1108	2431
16	12	1	1	1	1	1	1	1	1060	4421
17	18	0	0	0	0	0	0	0	1066	2668
18	3	0	0	0	0	0	0	0	1185	3056
19	5	0	0	0	0	0	0	0	1066	3091
20	8	0	0	0	0	0	0	0	1142	2872

Table SI-2. PLS model regression coefficients generated by DoE 1.

Model 1

Total Number of Peaks

	Coefficient	**SD**	***p***
Sample cone voltage	0.12661	0.0176572	0.00001
Source Temperature	0.01268	0.0176574	0.48613
Desolvation Temperature	0.08155	0.0176572	0.00059
Desolvation Gas	0.05314	0.0176572	0.01087
Cone Gas Flow	0.00103	0.0176572	0.95425
Flow	-0.01249	0.0176572	0.49383
Extraction Voltage	0.04347	0.0176572	0.02992

N = 20	Q2 = 0.418	
DF = 12	R2 = 0.881	RSD = 0.0769
Comp. = 3	R2 adj. = 0.812	
		Confidence = 0.95

Model 2

Number of Reliable Peaks

	Coefficient	**SD**	***p***
Sample cone voltage	-0.00863	0.00543	0.13813
Source Temperature	-0.00460	0.00543	0.41328
Desolvation Temperature	-0.00942	0.00543	0.10825
Desolvation Gas	-0.02016	0.00543	0.00297
Cone Gas Flow	0.002047	0.00543	0.71290
Flow	-0.00803	0.00543	0.16511
Extraction Voltage	0.004806	0.00543	0.39379

N = 20	Q2 = 0.389	
DF = 12	R2 = 0.658	RSD = 0.0236
Comp. = 3	R2 adj. = 0.460	
		Confidence = 0.95

DF: degrees of freedom, SD: standard deviation, MS: Mean Sum of squares (Variance), F: Fisher test, comp: number of components, R2: cumulative modeled variation, Q2: cumulative predicted variation, RSD: Residual Standard Deviation

Table SI-3. ANOVA test results of the generated PLS models for LC-MS design of experiments generated by DoE 1. Model 1 with total number of peaks as a response. Model 2 with number of reliable peaks as a response.

Model 1 : Total Number of Peaks

	DF	SS	MS (variance)	F	*p*
Total	20	229.915	11.4957		
	1	229.317	229.317		
Total corrected					
Regression	19	0.59761	0.03145		
Residual	7	0.52652	0.07521	12.6976	0.000
	12	0.07108	0.00592		
	N = 20	Q2 =	0.418		
	DF = 12	R2 =	0.881	RSD =	0.0769
	Comp. = 3	R2 adj. =	0.812		

Model 2 : Number of Reliable Peaks

	DF	SS	MS (variance)	F	*p*
Total	20	188.384	9.41922		
	1	188.365	188.365		
Total corrected					
Regression	19	0.01971	0.00103		
Residual	7	0.01298	0.00185	3.30614	0.033
	12	0.00673	0.00056		
	N = 20	Q2 =	0.389		
	DF = 12	R2 =	0.658	RSD =	0.0236
	Comp. = 3	R2 adj. =	0.460		

DF: degrees of freedom, SS: Sum of squares, MS: Mean Sum of squares (Variance), F: Fisher test, Comp: number of components, R2: cumulative modeled variation, Q2: cumulative predicted variation

Table SI-4A. Full Factorial design matrix with response values (DoE 2).

Run N°	Run Order	Wave Velocity (m/s)	Drift gas flow (mL/min)	Number of detected peaks
1	2	700	70	131
2	13	1000	70	1002
3	1	700	90	498
4	8	1000	90	910
5	5	700	80	589
6	3	1000	80	900
7	12	850	70	793
8	10	850	90	925
9	7	850	80	1009
10	4	850	80	893
11	11	850	80	1029
12	9	850	80	899
13	6	850	80	941

Table SI-4B. Coded Full Factorial design matrix with response values (DoE 2).

Run N°	Run Order	Wave Velocity (m/s)	Drift gas flow (mL/min)	Number of detected peaks
1	2	-1	-1	131
2	13	1	-1	1002
3	1	-1	1	498
4	8	1	1	910
5	5	-1	0	589
6	3	1	0	900
7	12	0	-1	793
8	10	0	1	925
9	7	0	0	1009
10	4	0	0	893
11	11	0	0	1029
12	9	0	0	899
13	6	0	0	941

Table SI-5. PLS model regression coefficients generated by DoE 2 (IMS design of experiments).

Total Number of Peaks	Coefficient	SD	*p*	
Wave Velocity	0.078610	0.01409	0.00140	
Drift gas flow	-0.01026	0.01409	0.49373	
Wave*Wave	-0.04202	0.01228	0.01412	
GFlow*GFlow	-0.01202	0.01228	0.36550	
Wave*GFlow	-0.00255	0.01384	0.85970	
N = 12	Q2 =	0.569		
DF = 6	R2 =	0.909	RSD =	0.0418
Comp. = 2	R2 adj. =	0.832		
			Confidence = 0.95	

DF: degrees of freedom, SS: Sum of squares, MS: Mean Sum of squares (Variance), F: Fisher test, *p*: *p*-value, Comp: number of components, R2: cumulative modeled variation, Q2: cumulative predicted variation

Table SI-6. ANOVA test results of the PLS model generated by DoE 2 (IMS design of experiments).

Total Number of Peaks	DF	SS	MS (variance)	F	*p*
Total	12	103.78	8.64829		
	1	103.665	103.665		
Total corrected					
Regression	11	0.11486	0.01044		
Residual	5	0.10436	0.02087	11.9249	**0.005**
	6	0.01050	0.001750		
	N = 12	Q2 =	0.569		
	DF = 6	R2 =	0.909	RSD =	0.0418
	Comp. = 2	R2 adj. =	0.832		

DF: degrees of freedom, SS: Sum of squares, MS: Mean Sum of squares (Variance), F: Fisher test, Comp: number of components, R2: cumulative modeled variation, Q2: cumulative predicted variation

Table SI-7A. Data descriptive statistics of DoE 1.

	Number of Reliable Peaks	Total Number of Peaks
N	20	20
Min	1.78533	3.06108
Max	1.89209	3.67514
Mean	1.82944	3.38613
Q(25%)	1.80276	3.30144
Q(75%)	1.85122	3.48763
Median	1.83251	3.40599
Std. dev.	0.0323668	0.177351
Min/Max	0.943573	0.832915
Std. dev./Mean	0.0176922	0.0523758

Table SI-7B. Data descriptive statistics of DoE 2.

	Total Number of Peaks
N	12
Min	2.69723
Max	3.02449
Mean	2.93917
Q(25%)	2.93273
Q(75%)	3.0002
Median	2.9715
Std. dev.	0.102189
Min/Max	0.891798
Std. dev./Mean	0.0347679

Figure SI-1. Examples of correlation plots of some included and excluded features with their respective correlation coefficient R². A) Included features. B) Excluded features.

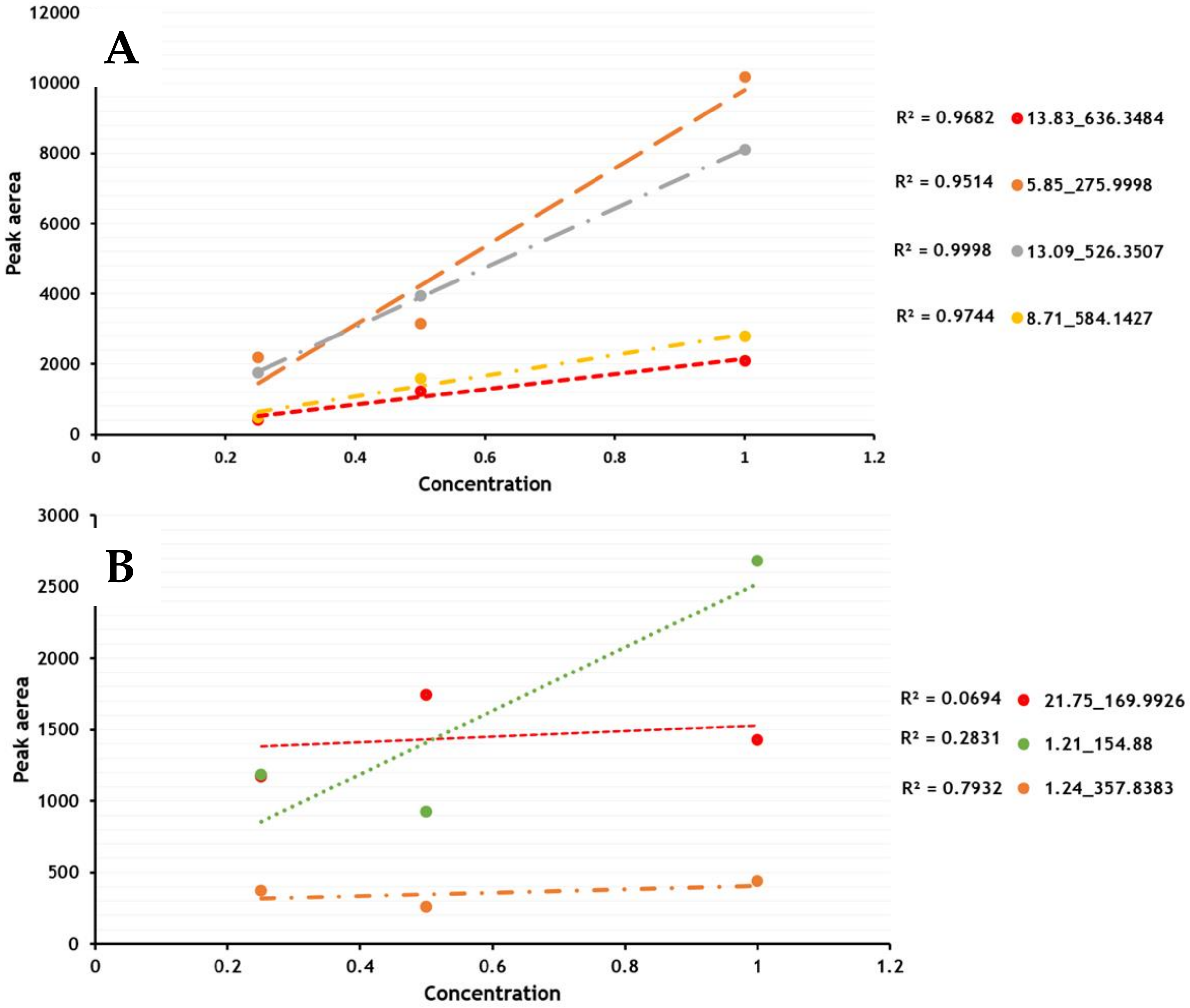
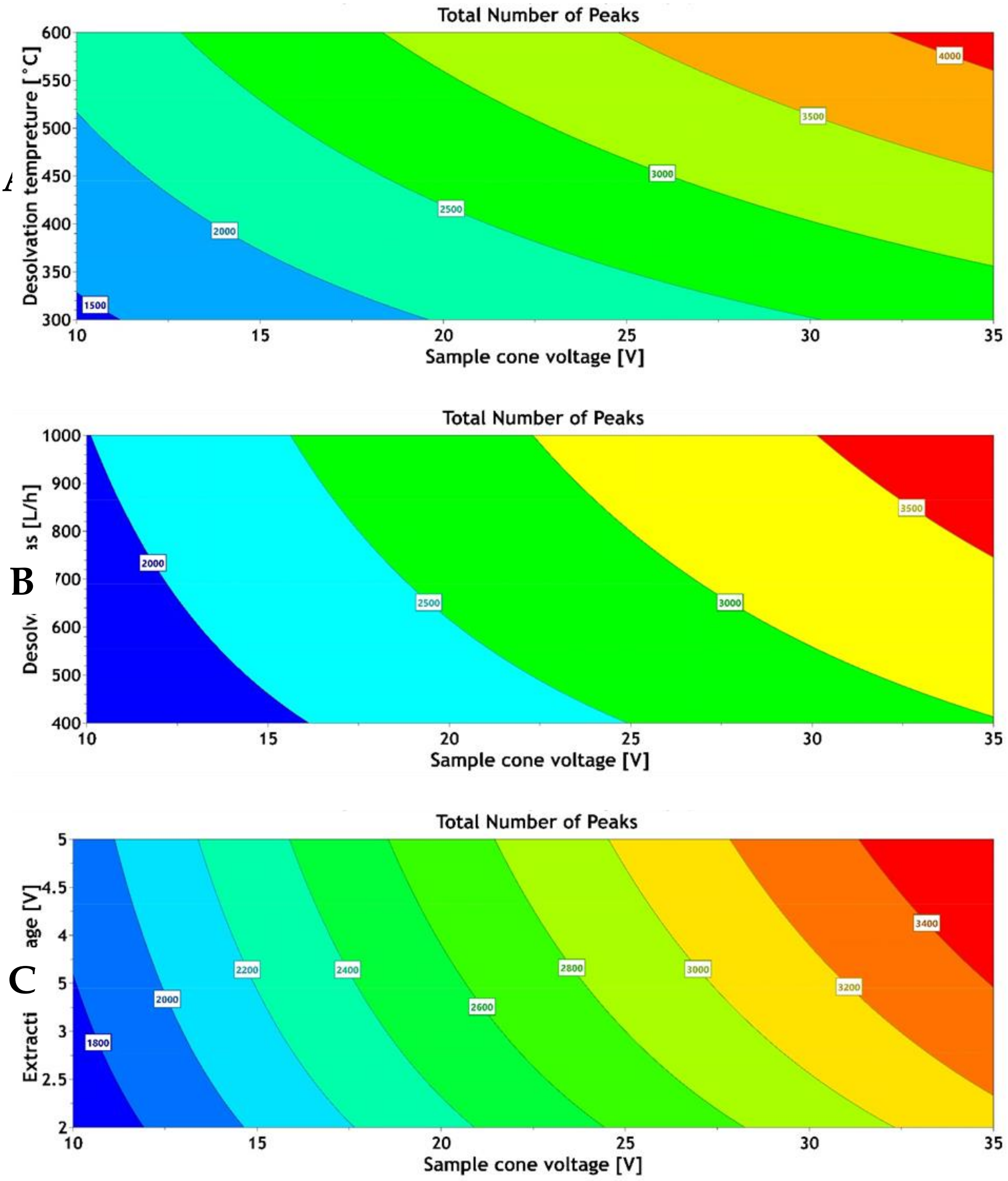


Figure SI-2. Contour plot of the different factors with total number of peaks as response. A) Sample cone voltage vs Desolvation temperature. B) Sample cone voltage vs Desolvation gas. C) Sample cone voltage vs Extraction voltage. D) Extraction voltage *vs* Desolvation gas. E) Desolvation temperature vs Desolvation gas. F) Extraction voltage vs Desolvation temperature.



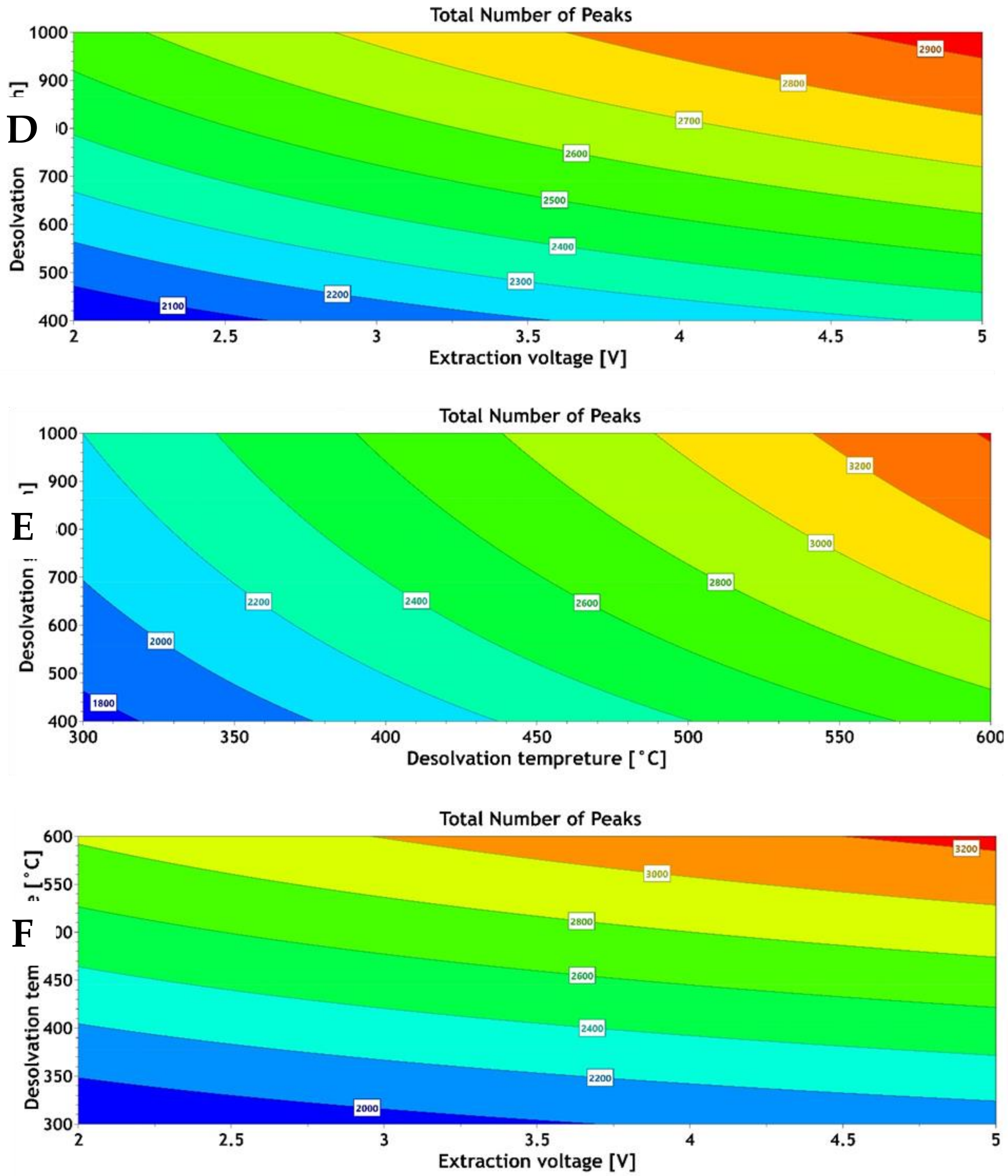

D
Total Number of Peaks
Desolvation
h]
1000
900
0
700
600
500
400
2100
2200
2300
2400
2500
2600
2700
2800
2900
2
2.5
3
3.5
4
4.5
5
Extraction voltage [V]
E
Total Number of Peaks
Desolvation
]
1000
900
00
700
600
500
400
1800
2000
2200
2400
2600
2800
3000
3200
300
350
400
450
500
550
600
Desolvation tempreture [°C]
F
Total Number of Peaks
Desolvation tem
[°C]
600
550
00
450
400
350
300
2000
2200
2400
2600
2800
3000
3200
2
2.5
3
3.5
4
4.5
5
Extraction voltage [V]

Figure SI-3. PLS model summary for the investigated response models in DoE 1, total number of peaks and number of reliable peaks. Cumulative values of regression coefficients and predictive coefficients are presented in green and blue bars respectively according to the number of included components in the model.

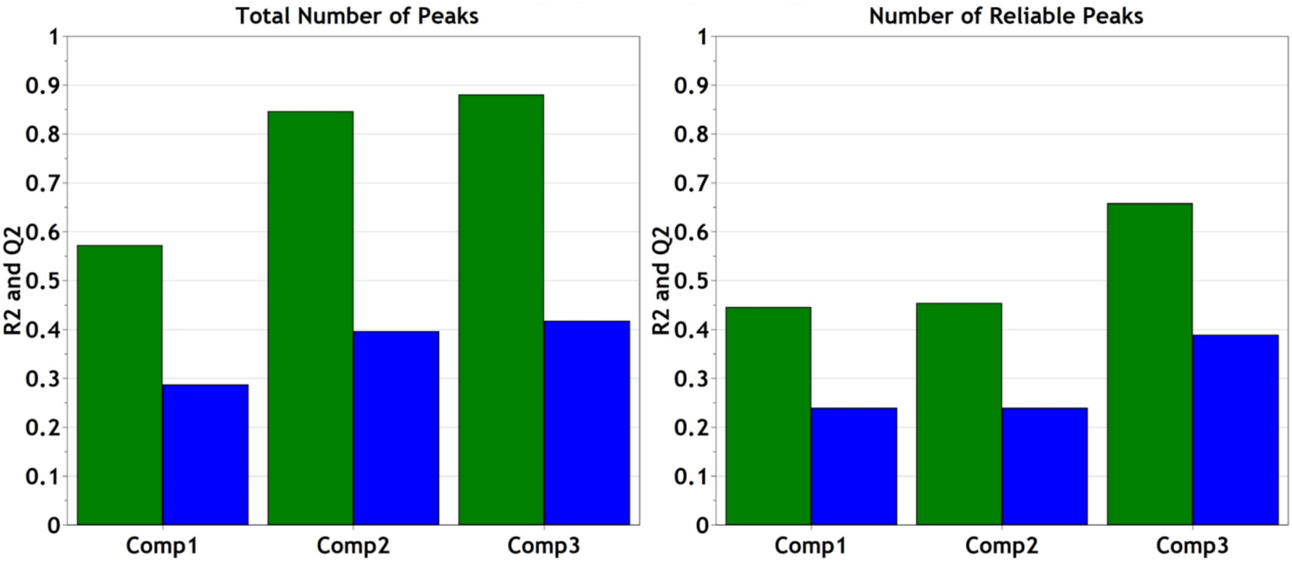


Figure SI-4. Plot the predicted versus the observed values of a number of detected peaks. Cumulative correlation (R^2) of 91% for 13 data points, indicating a good agreement.

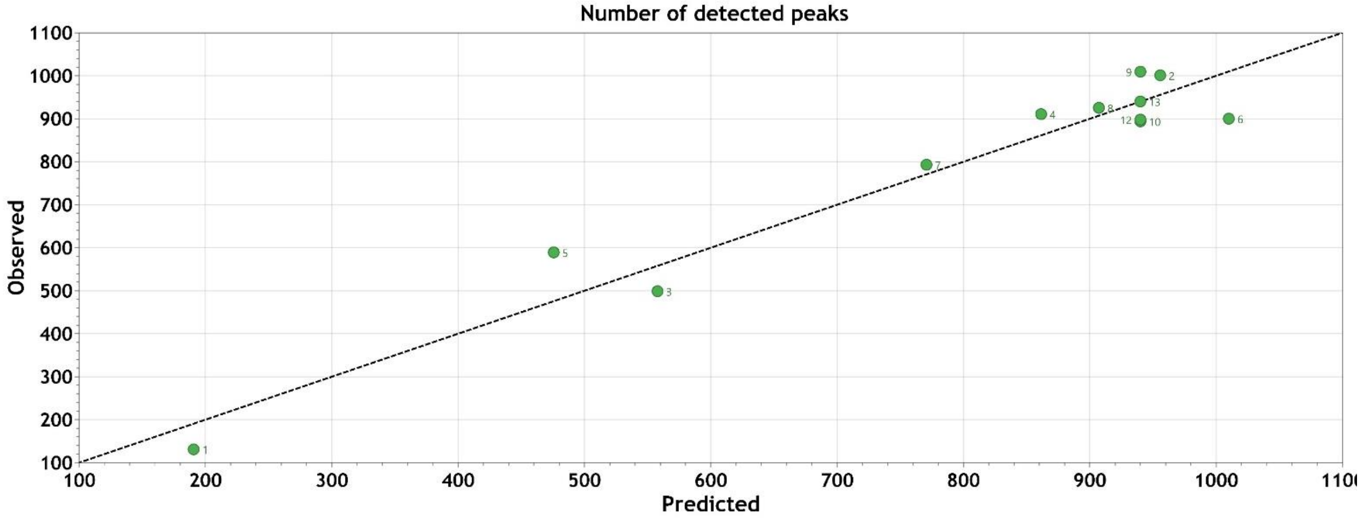


Figure SI-5. PLS model summary for the investigated response model total number of peaks (DoE 2). Cumulative values of regression coefficients and predictive coefficients are presented in green and blue bars respectively according to the number of included components in the model.

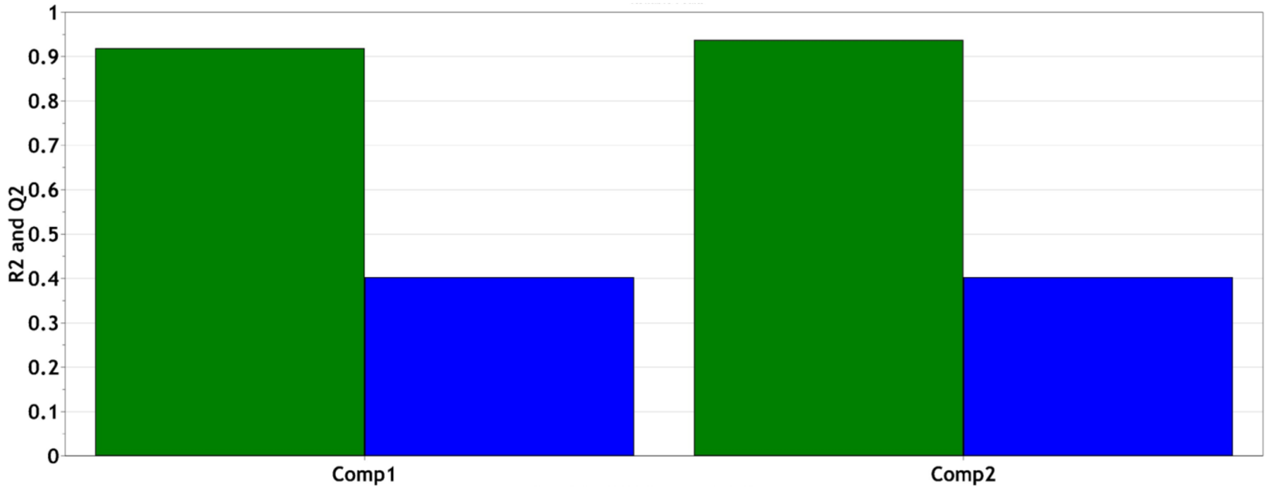


Figure SI-6. Peak selection strategy for the ion mobility experimental design optimization. Peaks are chosen to avoid the potential ion mobility cell 'wrap-around' effect.

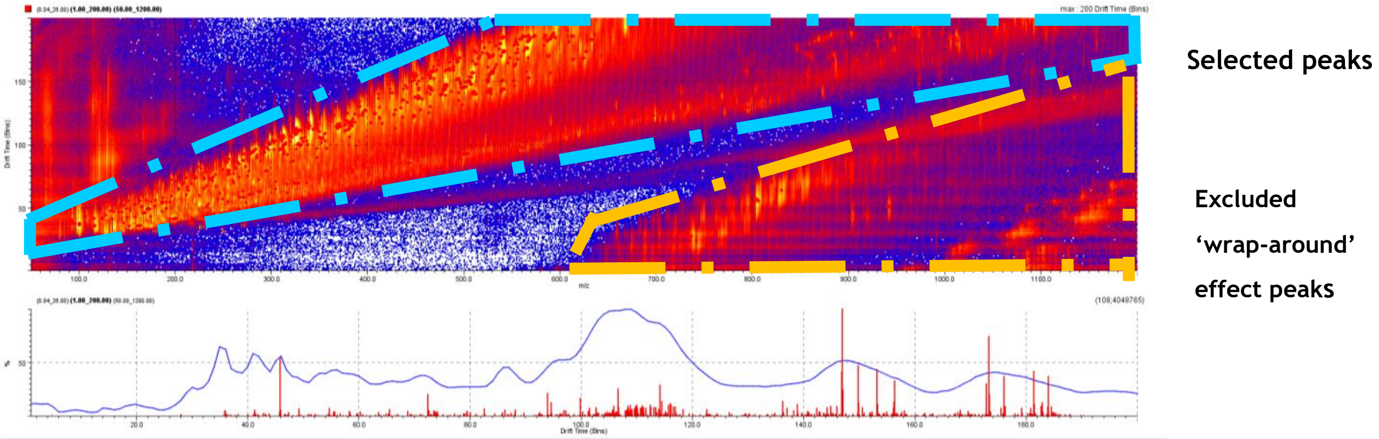
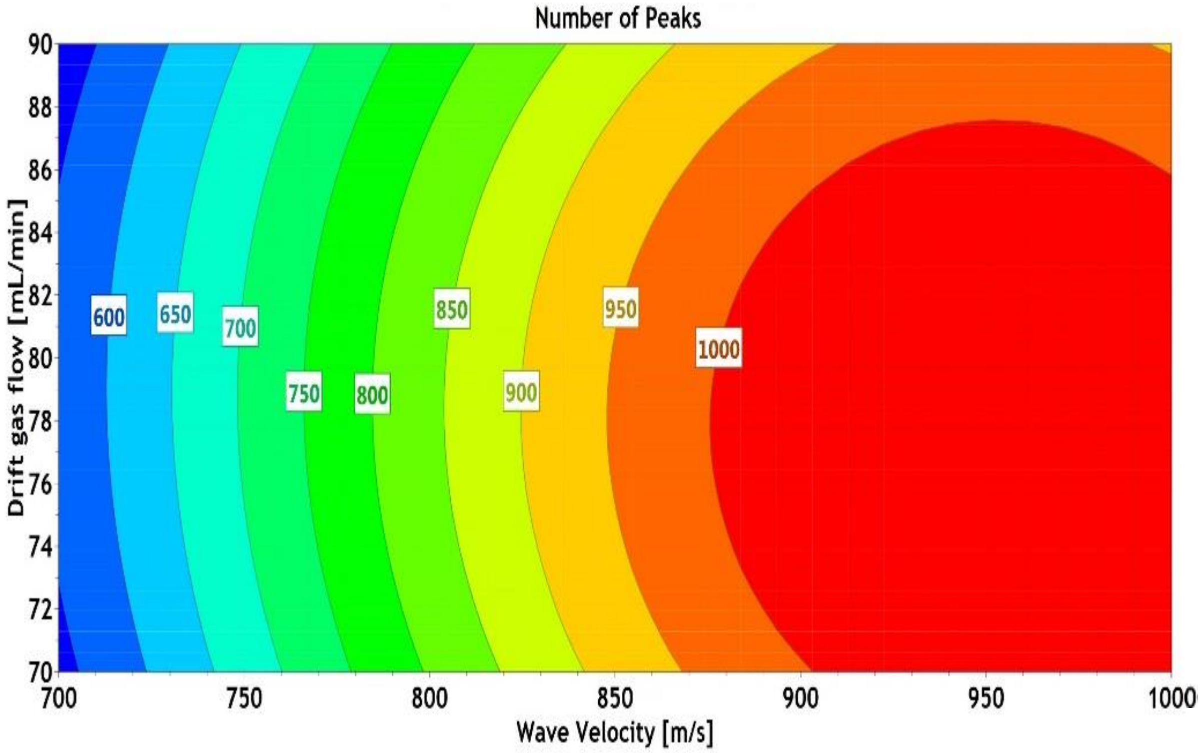


Figure SI-7. Contour plot of the wave velocity, drift gas flow with the number of detected peaks as response.



CHAPITRE II : ANALYSE METABOLOMIQUE DES MUCOPOLYSACCHARIDOSES

1. Projet METALYS

Comme décrit dans les chapitres précédents, la métabolomique dans son approche ciblée a depuis longtemps, utilisée dans l'évaluation des EIM en utilisant le plus souvent de la spectrométrie de masse.. En effet, plusieurs IEM sont actuellement diagnostiquées par cette approche tel que les aminoacidopathies, aciduries organiques, les troubles de l'oxydation des acides gras [1-5]. Cependant, peu de recherches métabolomiques ont été publiées dans le domaine des maladies lysosomales de surcharge (MLS). Les MLS représentent un groupe d'environ 50 maladies héréditaires dues à des déficits de protéines lysosomales. Cela conduit à une accumulation progressive de composés dans le lysosome. Ce stockage de métabolites induit diverses défaillances d'organes et une mort prématurée [6]. Dans le projet METALYS, nous proposons de développer une analyse métabolomique, à savoir une comparaison des profils métaboliques globale pour, d'une part, la recherche de biomarqueurs et d'autre part la compréhension de la physiopathologie. Comme preuve de concept, nous nous sommes d'abord concentrés sur les Mucopolysaccharidoses (MPS) qui appartiennent au groupe des MLS.

Le projet METALYS vise la validation clinique de cette approche globale en utilisant des échantillons de patients et de contrôle. Concernant les maladies extrêmement rares ou à faible prévalence, ce qui est le cas de notre étude, il est conseillé de suréchantillonner population témoin afin d'obtenir un certain degré de diversité dans l'échantillon test. Dans notre étude 66 échantillons contrôles ont été collectés. Les échantillons de patients ont été obtenus à partir de collections des hôpitaux (Rouen, Necker-Enfants Malades, Trousseau, Lille, Amiens et Lyon) incluant les différents types MPS (Figure 20).

2. Objectifs

L'objectif principal du projet METALYS est d'étudier l'association de profils métaboliques MPS en utilisant des techniques de modélisation statistiques multivariées. Cette analyse différentielle permettra de construire des modèles prédictifs basés les empreintes métaboliques urinaires spécifiques qui pourraient être utilisées à des fins de diagnostic et de suivi. Ces modèles devraient, donc, permettre la discrimination entre les échantillons MPS et les échantillons non-MPS (Contrôles *vs* Patients). Pour les MPS I, cette approche peut être aussi utilisée pour l'évaluation des effets métaboliques du traitement en discriminant les échantillons MPS I des échantillons MPS I traités. La méthodologie analytique a été développée et validée comme décrit dans le chapitre précèdent [7].

Par ailleurs, l'analyse métabolomique non ciblée permet d'identifier des métabolites discriminants et l'analyse des réseaux métaboliques permet de visualiser les voies métaboliques modulées. Cet aspect du projet porte sur la découverte de biomarqueurs. En effet, le profiling métabolique exploratoire offre un puissant moyen pour extraire des informations systémiques qui reflètent les influences génétiques et environnementales. Il est très informatif de déchiffrer les perturbations et les voies métaboliques liées au MPS. L'exploitation des données déjà acquises par la réalisation des analyses différentielles binaires de chaque sous-type MPS présente une voie prometteuse pour dévoiler la pathophysiologie de ces maladies. En effet, cette étude différentielle permet d'étudier les profils métaboliques spécifiques qui pourraient être utilisés pour le diagnostic et le suivi des objectifs des différents MPS. Ces profils peuvent fournir des indications sur la physiopathologie de la maladie par l'exploitation des caractéristiques métaboliques discriminantes qui peuvent constituer des biomarqueurs spécifiques potentiels. En outre, la stratégie analytique mise en œuvre utilisant la spectrométrie de masse à haute résolution, la chromatographie et la mobilité ionique offrent une combinaison puissante pour l'identification des métabolites et, par conséquent, un outil prometteur pour la découverte de biomarqueurs. Par ailleurs, les résultats du projet METALYS peuvent être testés sur la cohorte Radico-MPS dont l'objectif est de caractériser l'épidémiologie et l'histoire naturelle des MPS en construisant une collection rétrospective et prospective, avec de nombreuses données phénotypiques d'une cohorte française MPS des patients atteints de MPS. Les objectifs exploratoires du projet METALYS sont mettre en place de nouveaux biomarqueurs pour les essais cliniques à venir, en établissant une liste de biomarqueurs en termes de prise en charge thérapeutique et de trouver une meilleure prise en charge des patients. En effet, le projet METALYS est directement aligné avec les objectifs et perspectives Radico-MPS.

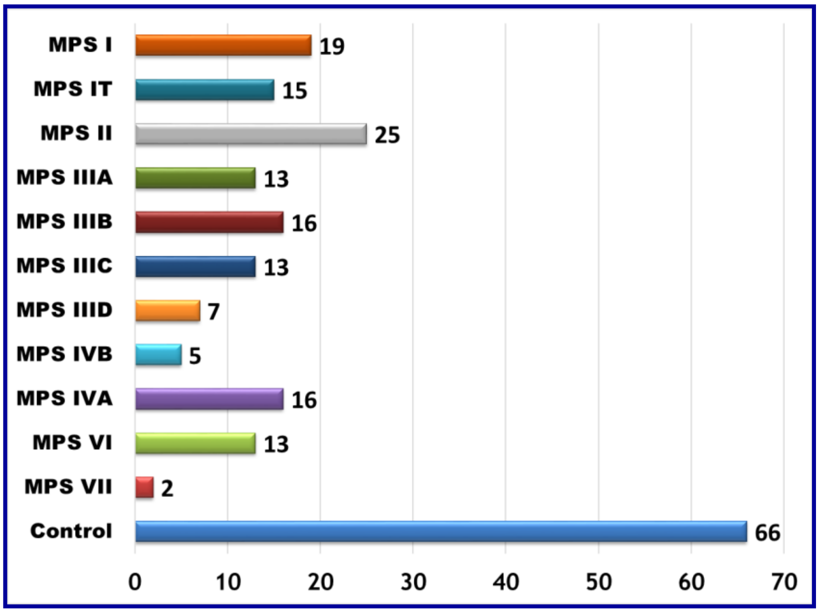


Figure 20. Répartition des échantillons de la cohorte du projet METALYS.

3. Partenaires

Ce travail a bénéficié de l'expertise du Service de Biochimie Métabolique- CHU de Rouen, du Laboratoire COBRA - UMR 6014 -IRCOF, Mont-Saint-Aignan, et de l'équipe "Genetics and Physiopathology of Neurodevelopmental Disorders" INSERM U1245, ainsi que les autres centres de références des EIM partenaires de ce projet à savoir Amiens, Lyon, Lille, Paris (Necker-Enfants Malades et Trousseau).

4. Analyse métabolomique urinaire de la mucopolysaccharidose de type I (Article VI)

Dans ce manuscrit, nous présentons les résultats du projet METALYS concernant l'analyse de la MPS I comparée à un groupe contrôle ainsi qu'un groupe de patients MPS I traités. Cette étude a permis de décrire des voies métaboliques impliquées dans le remodelage métabolique de la MPS I à savoir les métabolismes de l'arginine et du glutathion.

Ce chapitre, détaillant ces résultats, est présenté sous forme d'un article.

A Tebani , I Schmitz-Afonso, L Abily-Donval, B Heron, M Piraud, J Ausseil, F Zermiche, A Brassier, P De Lonlay, FM Vaz, BJ Gonzalez, S Marret, C Afonso, S Bekri. Urinary metabolic phenotyping of mucopolysaccharidoses combining untargeted and targeted strategies with data modelling. ***Submitted***

RÉFÉRENCES

1. Auray-Blais, C.; Maranda, B.; Lavoie, P. High-throughput tandem mass spectrometry multiplex analysis for newborn urinary screening of creatine synthesis and transport disorders, triple h syndrome and otc deficiency. *Clinica chimica acta; international journal of clinical chemistry* **2014**, *436*, 249-255.
2. Pitt, J.J. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *The Clinical biochemist. Reviews / Australian Association of Clinical Biochemists* **2009**, *30*, 19-34.
3. Pitt, J.J. Newborn screening. *The Clinical biochemist. Reviews / Australian Association of Clinical Biochemists* **2010**, *31*, 57-68.
4. Pitt, J.J.; Eggington, M.; Kahler, S.G. Comprehensive screening of urine samples for inborn errors of metabolism by electrospray tandem mass spectrometry. *Clinical chemistry* **2002**, *48*, 1970-1980.
5. Spacil, Z.; Tatipaka, H.; Barcenas, M.; Scott, C.R.; Turecek, F.; Gelb, M.H. High-throughput assay of 9 lysosomal enzymes for newborn screening. *Clin Chem* **2013**, *59*, 502-511.
6. Ballabio, A.; Gieselmann, V. Lysosomal disorders: From storage to cellular damage. *Biochimica et biophysica acta* **2009**, *1793*, 684-696.
7. Tebani, A.; Schmitz-Afonso, I.; Rutledge, D.N.; Gonzalez, B.J.; Bekri, S.; Afonso, C. Optimization of a liquid chromatography ion mobility-mass spectrometry method for untargeted metabolomics using experimental design and multivariate data analysis. *Analytica Chimica Acta* **2016**.

Urinary metabolic phenotyping of mucopolysaccharidosis type I combining untargeted and targeted strategies with data modelling

Abdellah TEBANI[1,2,3], Isabelle SCHMITZ-AFONSO[3], Lenaig ABILY-DONVAL[2,4], Bénédicte HERON[5], Monique PIRAUD[6], Jérôme AUSSEIL[7], Anais BRASSIER[8], Pascale De LONLAY[8], Farid ZERIMECH[9], Frédéric M VAZ[10], Bruno J. GONZALEZ[2], Stéphane MARRET[2,4], Carlos AFONSO[3], Soumeya BEKRI[1, 2,*]

[1] Department of Metabolic Biochemistry, Rouen University Hospital, Rouen, 76000, France

[2] Normandie Univ, UNIROUEN, CHU Rouen, INSERM U1245, 76000 Rouen, France

[3] Normandie Univ, UNIROUEN, INSA Rouen, CNRS, COBRA, 76000 Rouen, France

[4] Department of Neonatal Pediatrics and Intensive Care, Rouen University Hospital, Rouen, 76031, France

[5] Departement of Pediatric Neurology, Reference Center of Lysosomal Diseases, Trousseau Hospital, APHP, and GRC ConCer-LD, Sorbonne Universities, UPMC University 06, Paris, France

[6] Service de Biochimie et Biologie Moléculaire Grand Est, Unité des Maladies Héréditaires du Métabolisme et Dépistage Néonatal, Centre de Biologie et de Pathologie Est CHU de Lyon, Lyon, France.

[7] INSERM U1088, Laboratoire de Biochimie Métabolique, Centre de Biologie Humaine, CHU Sud, 80054, Amiens Cedex, France.

[8] Reference Center of Inherited Metabolic Diseases, Imagine Institute, Hospital Necker Enfants Malades, APHP, University Paris Descartes, Paris, France.

[9] Laboratoire de Biochimie et Biologie Moléculaire, Université de Lille et Pôle de Biologie Pathologie Génétique du CHRU de Lille, 59000, Lille, France.

[10] Laboratory Genetic Metabolic Diseases, Department of Clinical Chemistry and Pediatrics, Academic Medical Center, Amsterdam, The Netherlands

List of abbreviations: IEM, Inborn errors of metabolism; LSD, lysosomal storage diseases; MPS, Mucopolysaccharidoses; GAGs, Glycosaminoglycans; MPSI, mucopolysaccharidosis type I; MPSIT, treated mucopolysaccharidosis type I; UPLC-IM-MS, Ultraperformance liquid chromatography-ion mobility mass spectrometry; CCS, Collision Cross Section; ERT, enzyme replacement therapy; QC, Quality control; HS, Heparan sulfate; DS, Dermatan sulfate; KS, Keratan sulfate; ROC, Receiver operating characteristic; FDR, False discovery rate; PCA, Principal Component Analysis; OPLS-DA, Orthogonal Partial Least-Squares-Discriminant Analysis; VIP, Variable influence in projection; AUC, Area under curve; mTORC1, mammalian target of rapamycin complex 1

Abstract

BACKGROUND: Application of metabolic phenotyping could expand the pathophysiological knowledge of mucopolysaccharidoses (MPS) and reveal the comprehensive metabolic impairments in MPS. However, few studies applied this approach to MPS.

METHODS: We applied targeted and untargeted metabolic profiling in urine samples obtained from a French cohort comprising 19 MPSI and 15 MPSI treated patients along with 66 controls. For that purpose, we used ultra-high performance liquid chromatography combined with ion mobility and high resolution mass spectrometry following a protocol designed for large-scale metabolomics studies regarding robustness and reproducibility. Furthermore, 24 amino acids have been quantified using liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS). Keratan sulfate, Heparan sulfate and Dermatan sulfate concentrations have also been measured using an LC-MS/MS method. Univariate and multivariate data analyses have been used to select discriminant metabolites. The mummichog algorithm has been used for pathway analysis.

RESULTS: The studied groups yielded distinct biochemical phenotypes using multivariate data analysis. Univariate statistics also revealed metabolites that differentiated the groups. Specifically, metabolites related to the amino acid metabolism. Pathway analysis revealed that several major amino acid pathways were dysregulated in MPS. Comparison of targeted and untargeted metabolomics data with *in silico* results yielded arginine, proline and glutathione metabolisms being the most affected.

CONCLUSION: This study constitutes one of the first metabolic phenotyping studies of MPSI. The findings might help to generate new hypotheses about MPS pathophysiology and to develop further targeted studies of a smaller number of potentially key metabolites.

Keywords: Metabolomics; mucopolysaccharidosis type I; lysosomal storage diseases; mass spectrometry; ion mobility; inborn errors of metabolism

1. Introduction

Inborn errors of metabolism (IEM) represent a group of about 500 rare diseases with an overall estimated incidence of 1/2500. The diversity of involved metabolisms explains the difficulties in establishing their diagnosis. Optimal management of these patients requires then improved speed of biochemical investigation to allow early diagnosis and better monitoring. The rise of "omic" approaches offered a growing hope to provide new effective tools for screening, diagnosis, treatment and monitoring of these diseases. Indeed, unlike the conventional medical biology practice based primarily on the sequential study of genes, proteins and metabolites, the great challenge of modern biology is to understand disease as a complex, integrated and dynamic network [1]. The concept of "metabolome" refers to the comprehensive complement of all metabolites present in a given biological system, fluid, cell or tissue [2]. So, metabolomics is one of the "omic" technologies based on biochemical characterizations of the metabolome and its changes related to genetic and environmental factors. Metabolomics allows to characterize these interactions and to evaluate the biochemical mechanisms involved in such changes in a systematic fashion [3,4]. Given the strong link between IEM and metabolism, metabolomics is a very appealing tool to explore these diseases [5]. For years, mass spectrometry has been used in the assessment of inherited metabolic diseases [6-8]. However, few metabolomic research has been published in lysosomal storage diseases (LSD) field. LSDs represent a group of about 50 inherited disorders due to lysosomal proteins deficiencies which leads to a progressive accumulation of compounds within the lysosome. This metabolite storage causes various organ failures and premature death [9]. Mucopolysaccharidoses (MPS) belong to the LSD group. They are caused by impaired catabolism of glycosaminoglycans (GAGs), leading to their accumulation in lysosomes and extracellular matrix [10]**.** Then, they are secreted into the bloodstream to be eliminated in urine. Accumulated GAGs causes progressively multiple tissues and organ damages [11]. There are 11 known enzyme deficiencies, resulting in seven distinct forms of MPS [9]. Overall incidence is more than 1 in 30,000 live births [12]. Most MPS patients are asymptomatic after birth, however, prenatal symptoms may be observed in MPSI, MPSIVA, MPSVI and more frequently in MPSVII. MPS symptoms and severity vary with patients and MPS subtypes. Several MPS treatments are in clinical use or being investigated under clinical trials for patients [13]. MPSI is a rare autosomal recessive disorder caused by α-L-Iduronidase (IDUA, EC 3.2.1.76) deficiency. IDUA degrades complex polysaccharides by removing a single α-L-iduronyl residue from heparan sulfate and dermatan sulfate. The symptoms range from the severe Hurler form [MPSIH - OMIM #67014] to the more attenuated Hurler–Scheie (MPSIH/S - OMIM #607015) and Scheie (MPSIS - OMIM #67016) phenotypes. The classification is mainly based on the age at first symptoms and the presence or not of mental retardation [14]. The average survival age is of 28 years which imply patient's shift from pediatrics to adults [15]. Two specific treatments are available: hematopoietic stem cell transplantations (from bone marrow or blood cord donors) since the 1980s, and enzyme replacement therapy (ERT) (Laronidase, ALDURAZYME) since the 2000s. The aim of this study is to apply both targeted and untargeted metabolic profiling on MPSI patients compared to controls and to treated MPSI patients (MPSIT) to assess metabolic changes in this condition.

2. Materials and Methods

2.1. Urine samples

Random urine samples were collected from MPS patients in whom the diagnosis had been confirmed by demonstrating marked enzyme deficiency in leucocytes and/or by molecular analysis. Pseudodeficiencies have been ruled out. Urine samples were collected within seven reference centers for inherited metabolic diseases in France. Nineteen untreated MPSI patients were evaluated: 18 males (age range from 1 to 43.6 years, mean age: 22 years) and 1 female (age 5.5 years). Control urine samples from 66 healthy subjects, 27 males and 39 females (age range from 5.5 to 70 years, mean age: 40.8 years). Fifteen samples from MPSI treated patients (MPSIT) with enzyme replacement therapy, 11 males and 4 females (ages range from 1.3 to 39.3 years, mean age: 11.5 years) were analyzed for comparison. All samples were frozen at -80°C until analysis. This project was approved by the Research Ethics Board of Rouen University Hospital (CERNI E2016-21).

2.2. Metabolic phenotyping

2.2.1. Sample preparation

For untargeted metabolomics, urine samples were processed by transferring 200 µL of urines to 1.5 mL tubes and centrifuged at 4°C for 10 min at 13,000 g then 100 µL ultrapure water were added to 100 µL of supernatant and mixed. For amino acids and GAGs analysis, detailed protocols are presented in supporting information.

2.2.2. Untargeted analysis

Ultraperformance liquid chromatography-ion mobility mass spectrometry (UPLC-IM-MS) and data-independent MS acquisition with simultaneous analysis of molecular fragmentation (MS^E) were performed on a Waters Synapt G2 HDMS (Saint-Quentin-en-Yvelines, France) mass spectrometer equipped with a Waters nano-Acquity UPLC system and autosampler (Saint-Quentin-en-Yvelines, France), as previously described [16]. Injection volume was 2 µL and the autosampler was thermostated at 5°C. Metabolites separation was carried out at 45 °C using a 1.0 x 100 mm, Acquity UPLC HSS T3 column (Waters), with a particle size of 1.8 µm, equipped with a 0.2 µm prefilter. Urine was eluted from the LC column using the following linear gradient (curve number 6): 0–1 min: 99% A; 1–3 min, 99–85% A; 3–6 min, 85-50% A; 6–9 min, 50-0% A; 9–12 min, 100% B, 12–16 min, 99% A for re-equilibration. Solvent A was water and solvent B was acetonitrile, both solvents contained 0.1% formic acid. For ion mobility, the helium cell gas flow, wave height, Trap Bias and IMS wave delay, wave velocity and nitrogen pressure were set at 180 mL/min, 40 V, 45 V, 450 µs, 857 m/s and 80 ml/min respectively. The TOF analyzer was operated in the *V* resolution mode with an average mass resolution of m/Δm 20,000 (full-width at half-maximum definition). Data acquisition of an ion mobility experiment consisted of 200 bins. Collision Cross Section (CCS) values, obtained in nitrogen, were experimentally determined using singly charged Poly-DL-alanine oligomers as the TWIM calibrant species for ESI+. CCS values were derived according to previously reported procedures [17]. The ion mobility resolution was ~40 Ω/ΔΩ (fwhm). The CCS values reported were determined at the apex of the ion-mobility peak. Detailed instrument settings are presented in Table S-1 (Supporting Information). The mass spectrometer was operated in positive electrospray ionization mode. A mass range of *m*/*z* 50−1200 was used. The sample cone voltage, extraction cone voltage, source temperature, desolvation temperature, desolvation gas flow and cone gas flow were optimized and were as follows respectively: 25V, 5V, 120°C, 500°C, 400 L/h, 50 L/h Leucine enkephalin was used as the lock mass $[M+H]^+$ at *m*/*z* 556.2771. Sodium formate solution was used for external instrument calibration.

2.2.3. Raw data preprocessing

All LC-IM–MS raw data files, data processing, peak detection and peak matching across samples using retention time (t_R) correction and chromatographic alignment along with drift time and CCS calculation were performed using Progenesis QI (Waters MS Technologies, Manchester, UK) to yield a data matrix containing retention times, accurate masses (*m*/*z*), CCS and peak intensities. The preprocessing step resulted in an X-matrix where t_R, CCS and *m*/*z* values were concatenated into ʻʻt_R_m/z_CCSʼʼ features (in columns) present in each sample (in rows) with corresponding peak areas.

2.2.4. Quality Control

Ten µL of each urine sample are mixed together to generate a pooled quality control sample (QCs). QCs and solvent blank samples (mobile phase) were injected sequentially in-between the urine samples. In addition, a dilution series of QC samples (6%, 12.5%, 25%, 50% and 100% original concentration) are used to assess the quality of the extracted features. An analysis sequence is presented in Fig. S1. In this study, we used a filter strategy in which the features intensity must be correlated to the matrix concentration in a series of diluted QC samples. Furthermore, RSD values derived from repeated measurements of a pooled QC sample were used. The threshold was set to RSD<25%. More details are presented as supporting information.

2.2.5. Targeted analysis

2.2.5.1. Amino acids quantification

The analysis of free amino acid profiles in urine was based on a liquid chromatography coupled to tandem mass spectrometry method and the aTRAQ reagent. The aTRAQ kit allows to quantify 24 proteinogenic and non-proteinogenic free amino acids, in a range of biological fluids. The analyses were performed using the liquid chromatography instrument Prominence Shimadzu LC system consisting of a DGU-20A3 degasser, a LC-20AB pump, a SIL-20ACHT autosampler, and a CTO-20AC oven (Shimadzu, Prominence, Kyoto, Japan) coupled to the 4000 QTRAP mass spectrometer (Sciex, Framingham, MA, USA) with an electrospray ion source. The detailed description of the applied LC-ESI-MS/MS methodology for amino acid determination is presented in supplementary material. The amino acids concentrations were normalized by dividing them by creatininuria of the same urine sample.

2.2.5.2. Glycosaminoglycan quantification (HS, DS and KS)

Total urinary GAGs were measured with the dimethyl methylene blue-binding assay [18]. GAG-derived disaccharides in urine were analyzed using a multiplex assay as previously described by Eveline et al [19]. Briefly, Heparan sulfate, dermatan sulfate and keratan sulfate were enzymatically digested into disaccharides. The disaccharides were separated on a Thermo Hypercarb HPLC column (100×2.1 mm, 5 µm) using an Acquity UPLC system and quantified on a Waters Quattro Premier XE (tandem) mass spectrometer (Waters Corporation, Milford, MA, USA) with an ESI source using MRM acquisition mode. The detailed description of the protocol is presented in supplementary material. The GAG concentrations were normalized on creatinine levels.

2.3. Statistical Analyses

A one-way ANOVA test was applied to each selected variable in order to confirm their actual difference between the three groups. A t-test is used when binary comparison is applied. Furthermore, the Benjamini and Hochberg false discovery rate (FDR) method was used for calculating the false-positive rate associated with multiple comparisons, and provides corrected q-values with a 0.05 significance level (FDR 5%). A Receiver operating characteristic curve (ROC) has been used to assess the diagnostic performance of the chosen classifiers.

2.4. Data analysis and modelling

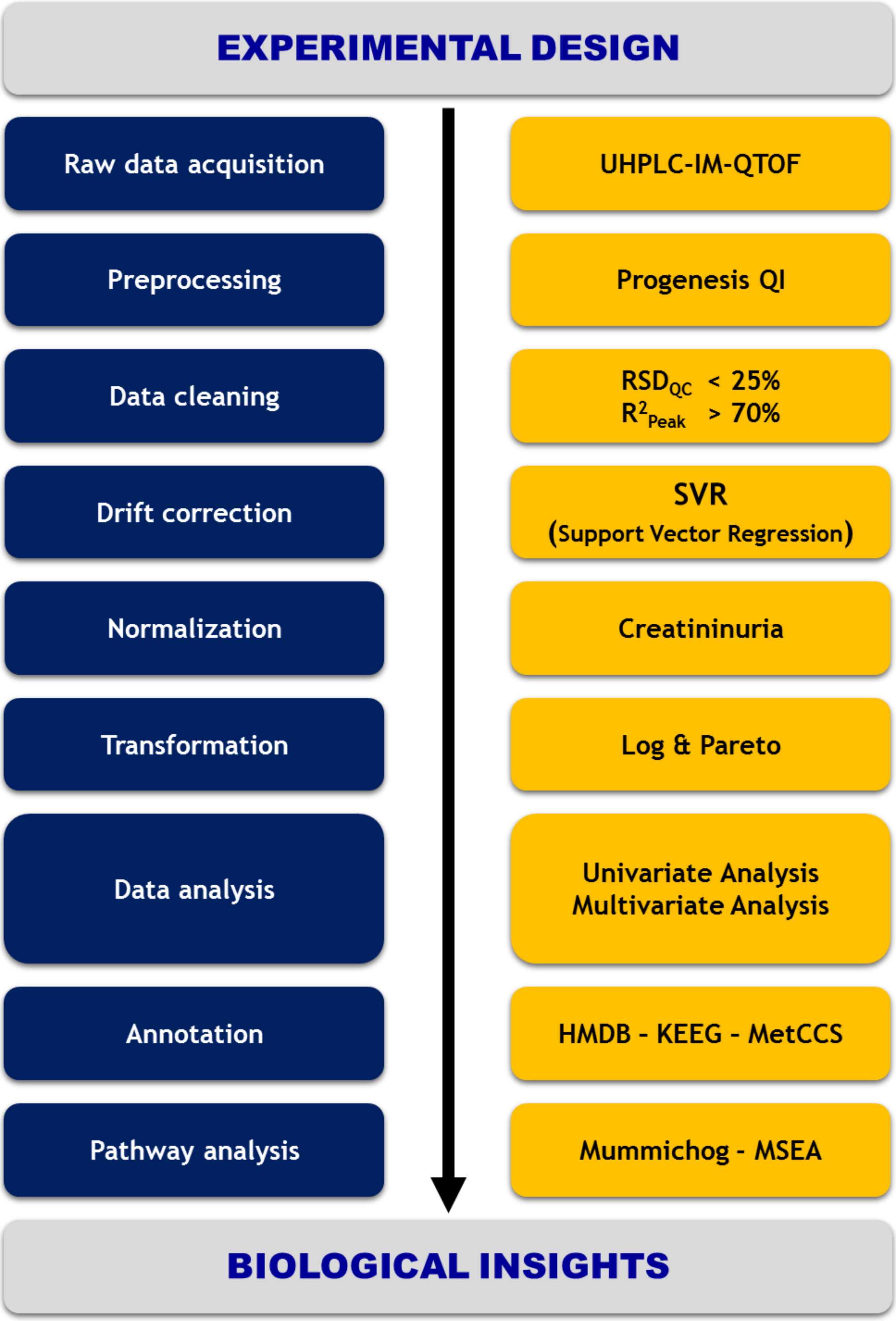
Support vector regression normalization method was applied using the MetNormalizer R package [20] before any data analysis, to remove the unwanted intra- and inter-batch measurement analytical variations. The effect of this normalization step on the raw data is shown in Fig.s S-2, S-3 and S-4 (Supporting information). Then, the normalized data matrix has been log-transformed and pareto-scaled. All data analyzes and modeling were done using SIMCA 14.0 (MKS DAS, Umeå, Sweden) and R software. First, hierarchical cluster analysis and Principal Component Analysis (PCA) were used as exploratory unsupervised methods [21]. Orthogonal Partial Least-Squares-Discriminant Analysis (OPLS-DA) was used as a supervised method for predictive modelling purposes. Details regarding data modeling, characteristics and validation results from all OPLS-DA models are provided in supplementary material.

2.5. Feature selection and annotation

To select the most discriminant variables for the separation of groups, S-Plot was used. The S-plot combines the covariances and correlations between the **X** matrix and O-PLS scores for a given model component. The covariance values give the magnitude of contribution of a variable while the correlation values reflect the effect and reliability of the variable for the model component scores. Variables with both very high correlation and covariance are important for the explanation power of the model. Selection of discriminant variables was achieved using the VIP (Variable Influence in Projection) scores procedure for each validated OPLS-DA model [22]. Putative annotation of detected features was performed using accurate mass comparison using freely available metabolite databases HMDB, LipidBlast, KEGG, and Metlin. Furthermore, CCS values were also compared to the MetCCS database [23].

2.6. Pathway and network analysis

In order to provide a broader understanding of metabolic changes in MPSI, we explored the biochemical pathways with a network analysis approach using the Mummichog software. This Python package highlights pathways that are significantly impacted in the studied groups. Significantly impacted biochemical pathways are those exhibiting an adjusted p-value <0.05. For this comparison, we focused on features that significantly changed (q-values = 0.05 and FDR = 5%). Mummichog annotates metabolites based on accurate mass *m/z* and tests significant pathway enrichment within a reference metabolic network using a Fisher's exact test [24]. To protect against incorrect pathway selection, redundant pathways or those enriched by less than two metabolites were excluded. MetaboAnalyst [25] has been used for Metabolite Set Enrichment Analysis on the amino acid concentration matrix. **The Fig. 1** presents an overview of the adopted metabolomics workflow.



EXPERIMENTAL DESIGN

Raw data acquisition	UHPLC-IM-QTOF
Preprocessing	Progenesis QI
Data cleaning	RSD_{QC} < 25% R^2_{Peak} > 70%
Drift correction	SVR (Support Vector Regression)
Normalization	Creatininuria
Transformation	Log & Pareto
Data analysis	Univariate Analysis Multivariate Analysis
Annotation	HMDB - KEEG - MetCCS
Pathway analysis	Mummichog - MSEA

BIOLOGICAL INSIGHTS

Figure 1. Illustration of the untargeted metabolomics wororflow spanning from experimental design to pathway analysis and biological interpretation. HMDB: Human Metabolome Database. KEEG: Kyoto Encyclopedia of Genes and Genomes. MetCCS: Metabolite CCS database. MSEA: Metabolite Set Enrichment Analysis. RSD: Relative Standard Deviation.

3. Results

Untargeted analysis

The untargeted analysis of urine samples of control individuals, MPSI and MPSIT patients yielded 854 features. The analysis by independent ANOVA test resulted in 511 metabolites above the $p < 0.05$ cut-off (FDR 5%). A hierarchical clustering analysis was first applied to group samples with similar profiles of variable intensity. The heatmap in **Fig. 2A** represents the top 100 features ranked by analysis of variance (ANOVA). The results show that all samples belonging to the same group were correctly clustered together. The dendrogram structure highlights two main clusters of variable intensities represented by its two longest branches (maximum dissimilarity according to the Euclidean distance). According to the color gradient, the intensity differences between groups are substantial. This first analysis allowed us to easily detect natural clusters in the data, although it did not facilitate extraction of discriminant variables among the dataset. To further explore natural separation between metabolic profiles and reduce data dimensionality, the dataset underwent a principal component analysis (PCA). The number of significant components was estimated using internal cross-validation with seven exclusion groups giving a three-component PCA model accounting for 21% of the total variance. The resulting scores plot was used to identify trends, groups and potential outliers within the data. **Fig. 2B** shows PCA scores plot. There is a clear separation between MPSI and MPSIT samples. However, there is an overlap with the control samples. Thus, to address our classification purposes, supervised methods are more suitable since they allow to accurately model the relationship between controls, MPS I and MPSIT samples. OPLS-DA classification was first applied to the dataset. Samples were labelled according to the corresponding groups, MPSI, MPSIT and Control. A model was considered predictive if the Q2 (cross-validation measure of the predictive power) regression line intercept resulting from the permutation test was negative. This means that the random labeled models exhibit lower predictive performance than the true one. The final model had an R2 = 0.96 and Q2= 0.54. The OPLS-DA scores plot (**Fig. 2C**) revealed that each class was well separated, suggesting that the OPLS-DA model successfully discriminated samples according to their underlying metabolic profile. This model was internally validated both by CV-ANOVA (p-value = 4 x 10^{-20}) and by the permutation test (999 permutations gave a negative Q2 intercept). Model validation details are shown in supplementary information (**Fig. S5**). To go further in data modelling, binary OPLS-DA classification models have been built. The first OPLS-DA model was built using a dataset including Control and MPSI samples. The model had one predictive and two orthogonal components, and its validation parameters were as follows: R2 = 0.94, Q2 = 0.63 and CV-ANOVA p-value = 1.75 x 10^{-15} (**Fig. S6**). The corresponding score plot is shown in **Fig. 2E.** It exhibited a clear separation between the two classes on the predictive component. A second OPLS-DA model was built using a dataset including MPSI and MPSIT samples following the same procedure. The OPLS-DA model had 1 + 3 component model with R2 = 0.97, Q2 = 0.63 and CV-ANOVA p-value = 5 x 10^{-4} (**Fig. 2D**). Selection of discriminant variables was achieved using the VIP scores procedure for each validated OPLS-DA model. Based on 1 as a cutoff value, 216 features out of the 854 were selected for the MPSI *vs.* Control model and 169 for the MPSI *vs.* MPSIT model. We then refined the two lists of variables by retaining only the most discriminant variables along with their putative annotation. The list included Carnitine, Arginine, Tetrahydrocorticosteron, Prolyl-Lysine, S-(5'-Adenosyl)-L-Methionine, oleic acid and Phenylalanylalanine. These discriminant variables are presented in **Table 1** for both models along with their respective statistical metrics and annotation accuracy. Boxplot of the main discriminant features are presented in **Fig. S8**. The discriminant performance of these features is also assessed using area under the receiver operating characteristic (ROC) curves. Carnitine has the highest AUC (AUC = 0.93). The overall ROC results are shown are shown in **Fig. S9**. Furthermore, to explore the underlying pathways dysregulated in MPS I we used the Mummichog software to look for significant pathways related to variation in the significantly disturbed features. Different metabolism pathways were affected such as glycerophospholipid metabolism, vitamins and amino acids are shown in **Table 2**. Interestingly, a series of amino acid metabolic pathways were markedly dysregulated.

Table 2. Significantly dysregulated pathways in MPS I.

MPS I vs. Control

Pathway	Overlap size	*p*-value (FDR = 5%)
Aspartate and asparagine metabolism	13	0.0075
Lysine metabolism	9	0.0128
Glutathione Metabolism	3	0.0145
Vitamin H (biotin) metabolism	3	0.0145
De novo fatty acid biosynthesis	6	0.0280
Tyrosine metabolism	19	0.0357
Ascorbate (Vitamin C) and Aldarate Metabolism	4	0.0359
Omega-3 fatty acid metabolism	3	0.0449
Vitamin B5 - CoA biosynthesis from pantothenate	3	0.0449

MPS I vs. MPS I treated

Pathway	Overlap size	*p*-value (FDR = 5%)
Lysine metabolism	13	0.0014
Glycerophospholipid metabolism	11	0.0028
Methionine and cysteine metabolism	8	0.0123
Tyrosine metabolism	26	0.0216
Biopterin metabolism	7	0.0222
Urea cycle/amino group metabolism	14	0.0260
Ascorbate (Vitamin C) and Aldarate Metabolism	5	0.0264
Arginine and Proline Metabolism	8	0.0341
Glutathione Metabolism	3	0.0388
Vitamin H (biotin) metabolism	3	0.0388

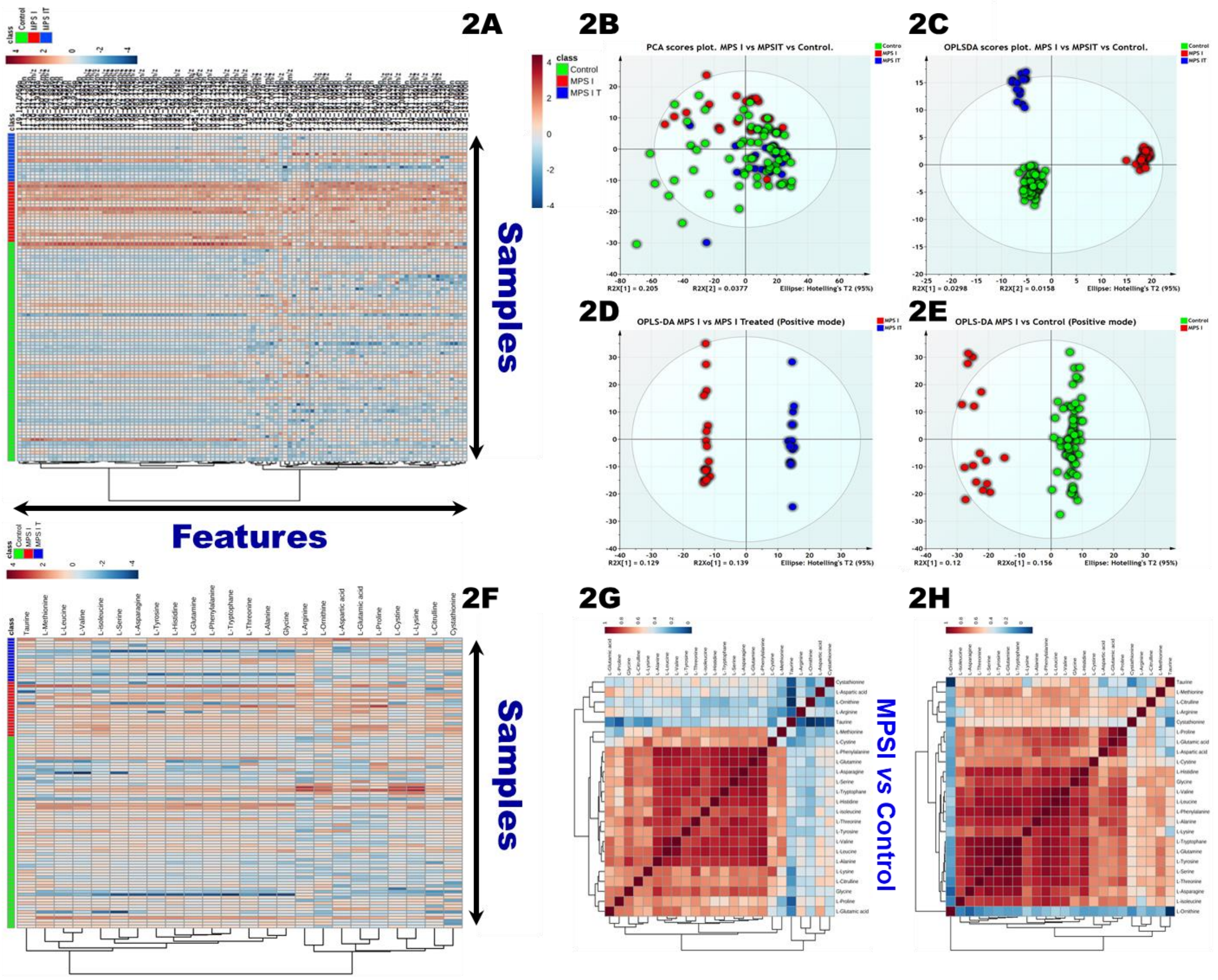


Fig. 2. 2A) Hierarchical cluster analysis and heat map visualization of top 100 variables (y-axis) ranked by ANOVA. The urine sample class within are represented along the x-axis. The color code was used to represent log-scaled intensities of features between -4 (blue) and +4 (brown), showing the features relative abundance according to the groups. **2B)** PCA scores plot of the normalized dataset. The three groups are represented by different colors. MPSI and MPSIT samples are well separated on PC1 according to their class membership. However, control samples show an overlap. **2C)** OPLSDA scores plot (R2 = 0.96, Q2= 0.54) shows a clear separation between the different groups. PC 1 separates the MPSI samples from the controls. However, PC2 separates treated MPSI from control samples. **2D)** Clear separation between treated MPSI and MPSIT samples is observed (R2 = 0.97, Q2= 0.63). **2E)** Clear separation between MPSI samples from the controls is observed (R2 = 0.94, Q2= 0.63) . Detailed model characteristics and validation are given in supporting information. **2F)** Heat map representing the clustering of 24 amino acids across the 3 groups of samples (MPSI, MPSIT and Controls). Columns represent individual samples and rows refer to amino acid. Shades of red or blue represent elevation or decrease, respectively, of an amino acid. **2G** and **2H)** Spearman rank-order correlation matrix 24 amino acids based on their concentrations profiles across all samples. Shades of red or blue represent low-to-high correlation coefficient between markers. **G)** MPSI vs Control. **H)** MPSI vs MPSIT.

Table 1. Some discriminant features extracted by OPLS-DA models allowing the discrimination of control subjects, MPSI and MPSIT.

MPSI *vs.* Control

HMDB	Putative annotation	Formula	*M*	*m/z*	Adduct	Δ *m/z* (ppm)	t_R (min)	tD (ms)	CCS (Å²)	FDR	%RSD	VIP
HMDB00062	Carnitine	$C_7H_{15}NO_3$	161.1053	203.1518	M+ACN+H	0.48	1.41	2.43	140.4	3.22E-09	10.0	1.90
HMDB00207	Oleic acid	$C_{18}H_{34}O_2$	282.2546	283.2616	M+H	-4.36	10.75	4.00	183.6	9.20E-07	24.53	1.85
HMDB29022	Prolyl-Lysine	$C_{11}H_{21}N_3O_3$	243.1595	282.1226	M+K	4.88	1.49	3.08	158.1	7.43E-08	5.97	1.73
HMDB00268	Tetrahydrocorticosteron	$C_{21}H_{34}O_4$	350.2465	351.2538	M+H	2.33	10.75	4.48	194.2	6.40E-05	17.26	1.51
HMDB00517	Arginine	$C_6H_{14}N_4O_2$	174.1112	175.1212	M+H	-2.70	1.23	2.11	130.7	3.21E-02	22.62	2.38

MPSI *vs.* MPSIT

HMDB	Putative annotation	Formula	*M*	*m/z*	Adduct	Δ *m/z* (ppm)	t_R (min)	tD (ms)	CCS (Å²)	FDR	%RSD	VIP
HMDB00517	Arginine	$C_6H_{14}N_4O_2$	174.1112	175.1212	M+H	-2.70	1.23	2.11	130.7	6.12E-03	22.62	2.38
HMDB00207	Oleic acid	$C_{18}H_{34}O_2$	282.2546	283.2616	M+H	-4.36	10.75	4.00	183.6	6.96E-04	24.53	1.93
HMDB00062	Carnitine	$C_7H_{15}NO_3$	161.1053	203.1518	M+ ACN+H	0.48	1.41	2.43	140.4	4.62E-04	9.97	1.93
HMDB28988	Phenylalanylalanine	$C_{12}H_{16}N_2O_3$	236.1152	237.1225	M+H	-3.62	7.67	2.70	147.8	2.07E-03	10.51	1.91
HMDB29022	Prolyl-Lysine	$C_{11}H_{21}N_3O_3$	243.1595	282.1226	M+K	4.88	1.49	3.08	158.1	1.70E-03	5.97	1.77
HMDB00268	Tetrahydrocorticosteron	$C_{21}H_{34}O_4$	350.2465	351.2538	M+H	2.33	10.75	4.48	194.2	4.00E-03	17.26	1.61

M: monoisotopic mass, ppm: parts per million t_R: Retention time, tD: Drift time, CCS: Cross collision section, VIP: variable influence on projection

Targeted analysis

The first targeted analysis addressed urinary glycosaminoglycans concentrations. As expected, total GAGs, dermatan sulfate (DS) and Heparan sulfate (HS) were significantly elevated in MPSI patients (**Table 3** and **Fig. S11**). Of note, Keratan sulfate (KS) is slightly elevated in MPSI patients compared to control samples. Given the results of the above untargeted approach, we also performed a targeted amino acid profiling on all the samples. Free amino acid profiles in urine samples of MPSI, MPSIT patients and the control group were obtained. Twenty-four amino acids were quantified in all samples and their concentrations were subjected to subsequent statistical and pathway analysis. **Table S3** presents absolute urine concentrations of amino acids. Boxplot of normalized amino acid concentrations are presented in **Fig. S10**. The statistical analysis of amino acids is listed in **Table 3**. Regarding Control *vs.* MPSI comparison, thirteen amino acids have shown significant difference between the two groups; Arginine, Aaspartic acid, Glutamic acid, Proline, Valine, Tryptophane, Lysine, Alanine, Leucine, Histidine, Threonine, Glutamine and Glycine. Besides, six amino acids showed statistically different concentrations between MPSI and MPSIT samples; Glutamic acid, Aspartic acid, Valine, Alanine and Isoleucine. To determine the amino acids profile differences between controls and MPSI and MPSIT patients, the 24 amino acids were first analyzed by an ANOVA test. The analysis yielded eight amino acids above the $p < 0.05$ cut-off (FDR 5%). A hierarchical clustering analysis was first applied to group samples with similar profiles. The heatmap in **Fig. 2F** represents the 24 amino acids ranked by analysis of variance (ANOVA). Even though, there is no an obvious visual pattern, the results show that all samples belonging to the same group were correctly clustered together. The dendrogram structure highlights two main clusters of variables represented by its two longest branches (maximum dissimilarity according to the Euclidean distance). Furthermore, a correlation analysis of the overall concentrations matrix has been performed. **Fig. 2G** and **2H** presents the heatmap of the correlation analysis. Both Figures show a clear cluster of variables that have high correlation. **Fig. 2G** (MPSI *vs* Control) showed a main cluster including Alanine, Leucine, Valine, Tyrosine, Threonine, Isoleucine, Histidine, Tryptophane, Serine, Asparagine, Glutamine, Phenylalanine. Regarding MPSI vs MPSIT, **Fig. 2H** shows two main clusters: The first one includes Isoleucine, Asparagine, Threonine, Serine, Tyrosine, Glutamine and Tryptophane; the second includes Alanine, Phenylalanine, Leucine, Valine, Glycine and Histidine. To assess the diagnostic performance of the different amino acids, univariate ROC curve analyses for MPSI *vs.* Control groups indicated four amino acids with high AUC above 0.80 and are: Arginine (AUC = 0.90), Glutamic acid (AUC = 0.86), Aspartic acid (AUC = 0.83) and Proline (AUC = 0.81). The same procedure has been performed for MPSI *vs.* MPSIT groups and indicated two amino acids with high AUC above 0.80 and were: Aspartic acid (AUC = 0.85) and Isoleucine (AUC = 0.82). The overall univariate and ROC analysis results are presented in **Table 3**. A comparison of different combinations of the main significant amino acids using a PLSDA model with three components each is presented in **Fig. S12**. Using these quantitative data, we performed pathway analysis that yielded the main impaired metabolisms. For MPSI *vs.* Control analysis, Arginine and Proline, Malate-Aspartate Schuttle, Cysteine, Urea cycle and alanine metabolism were the most affected pathways.

Regarding, MPSI *vs.* MPSIT analysis, Alanine, Malate-Aspartate Schuttle, branched amino acids metabolisms were the most affected. The overall results are shown in **Fig. 3A** and 3**B** for all the studied groups.

Table 3. T-test statistics, fold change and area under the curve of the receiver operating curves (ROC) for 24 amino acids and the GAGs. (*p < 0.05).* Significant features are highlighted in bold. (FDR = 5%)

	Control vs MPS I			MPS IT vs MPS I		
	AUC	q-value (FDR)	Fold Change	AUC	q-value (FDR)	Fold Change
Amino acids						
L-Arginine	0.904	**3.14E-06**	-3.75	0.51	4.78E-01	-0.41
L-Aspartic acid	0.83	**2.25E-04**	-2.19	0.86	**1.52E-02**	1.3
L-Glutamic acid	0.858	**4.78E-04**	-1.85	0.75	**3.58E-02**	0.28
L-Proline	0.816	**4.79E-04**	-1.89	0.73	7.34E-02	0.43
L-Valine	0.786	**1.50E-02**	-1.87	0.78	**3.33E-02**	0.5
L-Tryptophane	0.752	**2.41E-02**	-2.72	0.66	2.08E-01	0.16
L-Lysine	0.751	**2.41E-02**	-2.56	0.75	7.34E-02	0.47
L-Alanine	0.784	**2.41E-02**	-2.13	0.76	**3.58E-02**	0.83
L-Leucine	0.734	**2.51E-02**	-1.45	0.76	**4.72E-02**	0.35
L-Histidine	0.733	**4.30E-02**	-2.53	0.61	2.82E-01	-0.05
L-Threonine	0.687	**4.77E-02**	-1.13	0.67	2.01E-01	0.12
L-Glutamine	0.746	**4.77E-02**	-2.08	0.68	2.01E-01	0.15
Glycine	0.754	**4.77E-02**	-2.05	0.58	2.01E-01	0.04
Cystathionine	0.709	7.17E-02	-1.85	0.65	3.65E-01	0.15
L-Serine	0.718	7.19E-02	-2.06	0.68	2.82E-01	0.4
L-isoleucine	0.685	7.19E-02	-1.1	0.83	**1.85E-02**	1.81
L-Phenylalanine	0.704	7.80E-02	-1.3	0.75	1.51E-01	0.7
L-Ornithine	0.652	1.23E-01	-1.54	0.57	7.94E-01	-0.19
L-Citrulline	0.657	1.72E-01	-1	0.66	3.96E-01	0.19
L-Tyrosine	0.636	2.15E-01	-0.96	0.69	2.08E-01	0.55
L-Cystine	0.53	2.50E-01	-0.98	0.58	5.65E-01	-0.31
L-Asparagine	0.638	2.61E-01	-0.73	0.7	2.31E-01	0.74
Taurine	0.549	3.06E-01	-1.22	0.59	4.00E-01	-1.95
L-Methionine	0.524	3.65E-01	-0.48	0.65	1.93E-01	0.09
GAGs						
Total GAGs	0.92	**4.76E-04**	-3.74	0.79	6.39E-01	1.56
Dermatan sulfate	0.92	**1.37E-06**	-5.14	0.78	3.65E-01	2.48
Keratan sulfate	0.91	**2.75E-02**	-1.84	0.77	5.65E-01	1.24
Heparan sulfate	0.89	**1.14E-03**	-3.13	0.75	6.02E-01	1.46

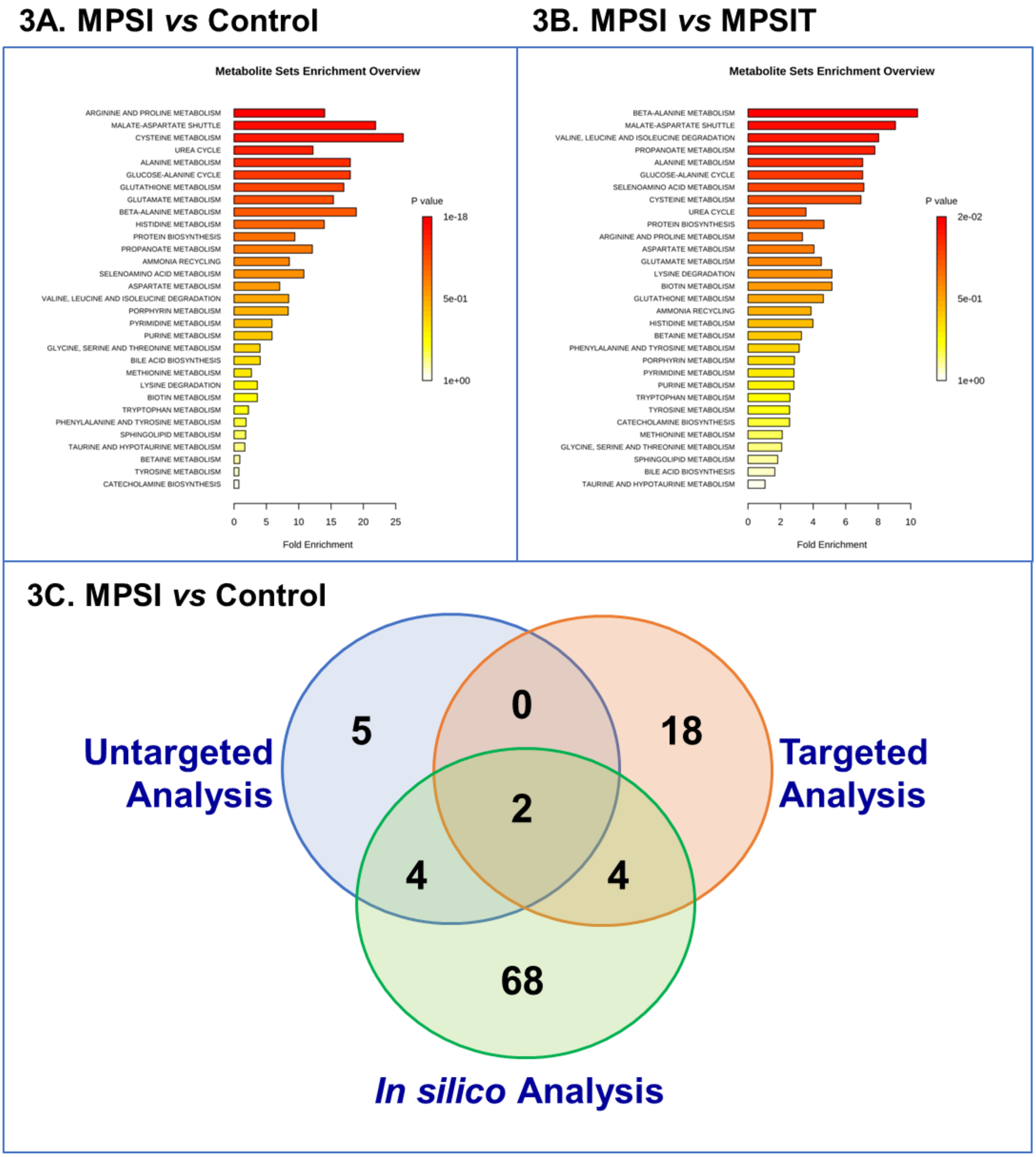


Fig. 3. Metabolite Set Enrichment Analysis using amino acid concentrations. **3A)** MPSI vs. Control. **3B)** MPSI vs. MPSIT. **3C)** Venn diagram of the significant pathways retrieved from untargeted, targeted approaches and in silico systems biology approach from Salazar DA et al [25]. The diagram shows two common metabolisms: Arginine-Proline metabolism and Cysteine-Glutathione metabolism. Detailed pathway information is given in supporting information (**Table S4**).

4. Discussion

In this study, the potential of metabolomics to identify biomarkers related to MPSI in urine was investigated. The data demonstrates that lysosomal accumulation of GAGs triggers deep metabolic turnover in MPSI patients. Urinary global metabolomics profiling may provide better understanding MPSI disease mechanisms and may pave the way for potential biomarkers. The seen metabolic alterations were mainly relevant to amino acid pathways contributing significantly to the clear discrimination of the different studied groups, MPSI, MPSIT and Control samples using their metabolic differences. Indeed, based on the untargeted urinary metabolic profiles retrieved from the different studied groups, we were able to build a predictive model that clearly separates the different studied groups, MPSI, MPSIT and Control samples using their metabolic differences. This study showed metabolic impairments mainly in amino acid metabolism and related metabolisms such vitamin and Glutathione metabolisms. In the light of these results, we performed a targeted analysis focusing on amino acids profiles which confirmed the amino acids profiles alterations. This pathway analysis yielded different dysregulated metabolic pathways. Furthermore, we performed a comparative analysis between the pathway analysis results from both untargeted and targeted results along with the recently *in silico* systems analysis data reported by Salazar *et al*. [26]. Salazar *et al.* performed a system biology approach using a genome-scale human metabolic reconstruction to understand the effect of metabolism alterations in MPS. The *in silico* MPSI model was generated by silencing *IDUA* gene then this model was analyzed through a flux balance and variability analysis. Thus, to depict the interrelationships between our untargeted and targeted results along with these *in silico* metabolic impairment data, we used a Venn diagram approach (**Fig. 3C**). Thus, two main metabolisms were identified: Arginine-Proline metabolism and Cysteine-Glutathione metabolism. Detailed data are presented in **Table S4.** Arginine-proline metabolism is depicted in **Fig. S13** and cysteine-glutathione metabolism is presented in **Fig. S14**. The later metabolism is tightly linked to oxidative stress, recent studies have shown oxidative damage involvement in the pathophysiology of several genetic diseases, including LSD [27]. The GAGs biosynthesis requires recycled substrates, however, the lack of recycled substrates in MPSs may lead to an increase in cellular energy needs [28]. This energy requirement may trigger the active mitochondrion turnover and results in an excess production of reactive oxidative species. Oxidative stress has been recognized as a mechanism of cell damage in MPSs and has been reported in MPS patients undergoing ERT [29-33]. Moreover, oxidative stress has been observed in MPSI mouse model [34,35]. Pereira *et al.* assessed oxidative stress in MPSI patients, compared with control subjects [29]. The authors detected a decrease in superoxide dismutase activity in erythrocytes from MPSI patients after ERT, while Catalase activity increased after ERT compared to baseline levels. These findings could suggest that potential antioxidants might be included as adjuvants for current MPS therapies. Regarding arginine, its classification performance is interestingly comparable with that of the quantified GAGs (**Table 3**). Arginine is an amino acid which is involved in several key metabolisms, urea cycle, nitric oxid, polyamins, glutamate, proline and homoarginine. Furthermore, changes in arginine levels act as nutritional sensors and regulate cellular metabolism through its interaction with mammalian target of rapamycin complex 1 (mTORC1) [36]. Recently, Chantranupong et al demonstrated that arginine sensing by mTORC1 is specific and depends on CASTOR1. The later interacts with GATOR2 to inhibit mTORC1 in low arginine condition. In the presence of arginine, CASTOR1 is bound to arginine, and thus free up GATOR2 and activate mTORC1 [37]. mTORC1 is already defined as a key regulator of protein synthesis, cell growth and autophagy [38]. Autophagy has been described as impaired in several LSDs [39] which may be attributed to lysosome dysfunction but may also be linked to arginine metabolism impairment. Woloszynek *et al*. observed profound metabolic alterations in energy expenditure in MPSI mice, similar to those observed in the current study with an increase in most amino acids concentrations including dipeptides, amino acid derivatives, and urea [40]. The authors attributed these changes to an increase of protein catabolism and an autophagy disruption as a consequence of lysosome dysfunction. Interestingly, autophagic vacuoles number are increased in several LSDs, and reduced in MPSI mice fed a high-fat diet [40].

5. Conclusion

Metabolic phenotyping might help generate new hypotheses about MPS pathophysiology and develop further targeted studies of a smaller number of potentially important metabolites. These findings enabled us to unveil profound metabolic impairments beyond the primary deficiency in MPSI. The understanding of disease pathophysiological bases may open new therapeutic strategies such as antioxidants adjuvants and diet intervention as complementary treatments for MPS and maybe for other LSDs.

Acknowledgements

The authors are grateful to Carine Pilon, Thomas Plichet and Tony Pereira (Rouen University Hospital, France) for their technical assistance. The authors gratefully acknowledge the Region Haute-Normandie, the Labex SynOrg (ANR-11-LABX-0029), the European Regional Development Fund (ERDF 31708), Shire, Sanofi-Genzyme and Vaincre les Maladies Lysosomales for financial support.

RÉFÉRENCES

1. Tebani, A.; Afonso, C.; Marret, S.; Bekri, S. Omics-based strategies in precision medicine: Toward a paradigm shift in inborn errors of metabolism investigations. *Int J Mol Sci* **2016**, *17*.
2. 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; the fate of foreign compounds in biological systems* **1999**, *29*, 1181-1189.
3. Bekri, S. The role of metabolomics in precision medicine. *Expert Review of Precision Medicine and Drug Development* **2016**.
4. Benton, H.P.; Want, E.; Keun, H.C.; Amberg, A.; Plumb, R.S.; Goldfain-Blanc, F.; Walther, B.; Reily, M.D.; Lindon, J.C.; Holmes, E., *et al.* Intra- and interlaboratory reproducibility of ultra performance liquid chromatography-time-of-flight mass spectrometry for urinary metabolic profiling. *Analytical chemistry* **2012**, *84*, 2424-2432.
5. Tebani, A.; Abily-Donval, L.; Afonso, C.; Marret, S.; Bekri, S. Clinical metabolomics: The new metabolic window for inborn errors of metabolism investigations in the post-genomic era. *Int J Mol Sci* **2016**, *17*.
6. Pitt, J.J. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *The Clinical biochemist. Reviews / Australian Association of Clinical Biochemists* **2009**, *30*, 19-34.
7. Pitt, J.J. Newborn screening. *The Clinical biochemist. Reviews / Australian Association of Clinical Biochemists* **2010**, *31*, 57-68.
8. Spacil, Z.; Tatipaka, H.; Barcenas, M.; Scott, C.R.; Turecek, F.; Gelb, M.H. High-throughput assay of 9 lysosomal enzymes for newborn screening. *Clin Chem* **2013**, *59*, 502-511.
9. Ballabio, A.; Gieselmann, V. Lysosomal disorders: From storage to cellular damage. *Biochimica et biophysica acta* **2009**, *1793*, 684-696.
10. Wraith, J.E. The mucopolysaccharidoses: A clinical review and guide to management. *Archives of disease in childhood* **1995**, *72*, 263-267.
11. Neufeld EF, M.J. The mucopolysaccharidoses. In *The metabolic and molecular basis of inherited disease*, Scrive C, B.A., Sly W, Vaele D, Ed. Mc Graw-Hill: New York, NY, 2001; pp 3421–3452.
12. Baehner, F.; Schmiedeskamp, C.; Krummenauer, F.; Miebach, E.; Bajbouj, M.; Whybra, C.; Kohlschutter, A.; Kampmann, C.; Beck, M. Cumulative incidence rates of the mucopolysaccharidoses in germany. *J Inherit Metab Dis* **2005**, *28*, 1011-1017.
13. Parenti, G.; Andria, G.; Ballabio, A. Lysosomal storage diseases: From pathophysiology to therapy. *Annu Rev Med* **2015**, *66*, 471-486.
14. Tebani, A.; Zanoutene-Cheriet, L.; Adjtoutah, Z.; Abily-Donval, L.; Brasse-Lagnel, C.; Laquerriere, A.; Marret, S.; Chalabi Benabdellah, A.; Bekri, S. Clinical and molecular characterization of patients with mucopolysaccharidosis type i in an algerian series. *Int J Mol Sci* **2016**, *17*.
15. OA, B. Clinical characteristics of mps i patients in the mps i registry. 2007 san diego. *The American Society of Human Genetics* **2007**.
16. Tebani, A.; Schmitz-Afonso, I.; Rutledge, D.N.; Gonzalez, B.J.; Bekri, S.; Afonso, C. Optimization of a liquid chromatography ion mobility-mass spectrometry method for untargeted metabolomics using experimental design and multivariate data analysis. *Analytica Chimica Acta* **2016**, *913*, 55-62.
17. Paglia, G.; Angel, P.; Williams, J.P.; Richardson, K.; Olivos, H.J.; Thompson, J.W.; Menikarachchi, L.; Lai, S.; Walsh, C.; Moseley, A., *et al.* Ion mobility-derived collision cross section as an additional measure for lipid fingerprinting and identification. *Analytical chemistry* **2015**, *87*, 1137-1144.
18. de Jong, J.G.; Wevers, R.A.; Liebrand-van Sambeek, R. Measuring urinary glycosaminoglycans in the presence of protein: An improved screening procedure for mucopolysaccharidoses based on dimethylmethylene blue. *Clin Chem* **1992**, *38*, 803-807.
19. Langereis, E.J.; Wagemans, T.; Kulik, W.; Lefeber, D.J.; van Lenthe, H.; Oussoren, E.; van der Ploeg, A.T.; Ruijter, G.J.; Wevers, R.A.; Wijburg, F.A., *et al.* A multiplex assay for the diagnosis of mucopolysaccharidoses and mucolipidoses. *PloS one* **2015**, *10*, e0138622.
20. Shen, X.; Gong, X.; Cai, Y.; Guo, Y.; Tu, J.; Li, H.; Zhang, T.; Wang, J.; Xue, F.; Zhu, Z.-J. Normalization and integration of large-scale metabolomics data using support vector regression. *Metabolomics* **2016**, *12*, 89.
21. Eriksson, L.; Trygg, J.; Wold, S. A chemometrics toolbox based on projections and latent variables. *Journal of Chemometrics* **2014**, *28*, 332-346.
22. Galindo-Prieto, B.; Eriksson, L.; Trygg, J. Variable influence on projection (vip) for orthogonal projections to latent structures (opls). *Journal of Chemometrics* **2014**.
23. Zhou, Z.; Xiong, X.; Zhu, Z.J. Metccs predictor: A web server for predicting collision cross-section values of metabolites in ion mobility-mass spectrometry based metabolomics. *Bioinformatics (Oxford, England)* **2017**.
24. Li, S.; Park, Y.; Duraisingham, S.; Strobel, F.H.; Khan, N.; Soltow, Q.A.; Jones, D.P.; Pulendran, B. Predicting network activity from high throughput metabolomics. *PLoS Comput. Biol.* **2013**, *9*, e1003123.
25. Xia, J.; Sinelnikov, I.V.; Han, B.; Wishart, D.S. Metaboanalyst 3.0-making metabolomics more meaningful. *Nucleic acids research* **2015**.

26. Salazar, D.A.; Rodriguez-Lopez, A.; Herreno, A.; Barbosa, H.; Herrera, J.; Ardila, A.; Barreto, G.E.; Gonzalez, J.; Almeciga-Diaz, C.J. Systems biology study of mucopolysaccharidosis using a human metabolic reconstruction network. *Molecular genetics and metabolism* **2016**, *117*, 129-139.
27. Donida, B.; Jacques, C.E.D.; Mescka, C.P.; Rodrigues, D.G.B.; Marchetti, D.P.; Ribas, G.; Giugliani, R.; Vargas, C.R. Oxidative damage and redox in lysosomal storage disorders: Biochemical markers. *Clinica Chimica Acta* **2017**, *466*, 46-53.
28. Woloszynek, J.C.; Kovacs, A.; Ohlemiller, K.K.; Roberts, M.; Sands, M.S. Metabolic adaptations to interrupted glycosaminoglycan recycling. *The Journal of biological chemistry* **2009**, *284*, 29684-29691.
29. Pereira, V.G.; Martins, A.M.; Micheletti, C.; D'Almeida, V. Mutational and oxidative stress analysis in patients with mucopolysaccharidosis type i undergoing enzyme replacement therapy. *Clinica chimica acta; international journal of clinical chemistry* **2008**, *387*, 75-79.
30. Jacques, C.E.; Donida, B.; Mescka, C.P.; Rodrigues, D.G.; Marchetti, D.P.; Bitencourt, F.H.; Burin, M.G.; de Souza, C.F.; Giugliani, R.; Vargas, C.R. Oxidative and nitrative stress and pro-inflammatory cytokines in mucopolysaccharidosis type ii patients: Effect of long-term enzyme replacement therapy and relation with glycosaminoglycan accumulation. *Biochimica et biophysica acta* **2016**, *1862*, 1608-1616.
31. Negretto, G.W.; Deon, M.; Biancini, G.B.; Burin, M.G.; Giugliani, R.; Vargas, C.R. Glycosaminoglycans can be associated with oxidative damage in mucopolysaccharidosis ii patients submitted to enzyme replacement therapy. *Cell biology and toxicology* **2014**, *30*, 189-193.
32. Filippon, L.; Wayhs, C.A.; Atik, D.M.; Manfredini, V.; Herber, S.; Carvalho, C.G.; Schwartz, I.V.; Giugliani, R.; Vargas, C.R. DNA damage in leukocytes from pretreatment mucopolysaccharidosis type ii patients; protective effect of enzyme replacement therapy. *Mutation research* **2011**, *721*, 206-210.
33. Donida, B.; Marchetti, D.P.; Biancini, G.B.; Deon, M.; Manini, P.R.; da Rosa, H.T.; Moura, D.J.; Saffi, J.; Bender, F.; Burin, M.G., *et al.* Oxidative stress and inflammation in mucopolysaccharidosis type iva patients treated with enzyme replacement therapy. *Biochimica et biophysica acta* **2015**, *1852*, 1012-1019.
34. Simonaro, C.M.; D'Angelo, M.; He, X.; Eliyahu, E.; Shtraizent, N.; Haskins, M.E.; Schuchman, E.H. Mechanism of glycosaminoglycan-mediated bone and joint disease: Implications for the mucopolysaccharidoses and other connective tissue diseases. *The American journal of pathology* **2008**, *172*, 112-122.
35. Reolon, G.K.; Reinke, A.; de Oliveira, M.R.; Braga, L.M.; Camassola, M.; Andrades, M.É.; Moreira, J.C.F.; Nardi, N.B.; Roesler, R.; Dal-Pizzol, F. Alterations in oxidative markers in the cerebellum and peripheral organs in mps i mice. *Cellular and molecular neurobiology* **2009**, *29*, 443-448.
36. Morris, S.M., Jr. Arginine metabolism revisited. *The Journal of nutrition* **2016**, *146*, 2579s-2586s.
37. Chantranupong, L.; Scaria, S.M.; Saxton, R.A.; Gygi, M.P.; Shen, K.; Wyant, G.A.; Wang, T.; Harper, J.W.; Gygi, S.P.; Sabatini, D.M. The castor proteins are arginine sensors for the mtorc1 pathway. *Cell* **2016**, *165*, 153-164.
38. Goberdhan, D.C.; Wilson, C.; Harris, A.L. Amino acid sensing by mtorc1: Intracellular transporters mark the spot. *Cell metabolism* **2016**, *23*, 580-589.
39. Chévrier, M.; Brakch, N.; Céline, L.; Genty, D.; Ramdani, Y.; Moll, S.; Djavaheri-Mergny, M.; Brasse-Lagnel, C.; Annie Laquerrière, A.L.; Barbey, F. Autophagosome maturation is impaired in fabry disease. *Autophagy* **2010**, *6*, 589-599.
40. Woloszynek, J.C.; Kovacs, A.; Ohlemiller, K.K.; Roberts, M.; Sands, M.S. Metabolic adaptations to interrupted glycosaminoglycan recycling. *The Journal of biological chemistry* **2009**, *284*, 29684-29691.

Supplementary material for:

Urinary metabolic phenotyping of mucopolysaccharidosis type I combining untargeted and targeted strategies with data modelling

Abdellah TEBANI[1,2,3], Isabelle SCHMITZ-AFONSO[3], Lenaig ABILY-DONVAL[2,4], Bénédicte HERON[5], Monique PIRAUD[6], Jérôme AUSSEIL[7], Anais BRASSIER[8], Pascale De LONLAY[8], Farid ZERIMECH[9], Frédéric M VAZ[10], Bruno J. GONZALEZ[2], Stéphane MARRET[2,4], Carlos AFONSO[3], Soumeya BEKRI[1, 2,*]

[1] Department of Metabolic Biochemistry, Rouen University Hospital, Rouen, 76000, France

[2] Normandie Univ, UNIROUEN, CHU Rouen, INSERM U1245, 76000 Rouen, France

[3] Normandie Univ, UNIROUEN, INSA Rouen, CNRS, COBRA, 76000 Rouen, France

[4] Department of Neonatal Pediatrics and Intensive Care, Rouen University Hospital, Rouen, 76031, France

[5] Departement of Pediatric Neurology, Reference Center of Lysosomal Diseases, Trousseau Hospital, APHP, and GRC ConCer-LD, Sorbonne Universities, UPMC University 06, Paris, France

[6] Service de Biochimie et Biologie Moléculaire Grand Est, Unité des Maladies Héréditaires du Métabolisme et Dépistage Néonatal, Centre de Biologie et de Pathologie Est CHU de Lyon, Lyon, France.

[7] INSERM U1088, Laboratoire de Biochimie Métabolique, Centre de Biologie Humaine, CHU Sud, 80054, Amiens Cedex, France.

[8] Reference Center of Inherited Metabolic Diseases, Imagine Institute, Hospital Necker Enfants Malades, APHP, University Paris Descartes, Paris, France.

[9] Laboratoire de Biochimie et Biologie Moléculaire, Université de Lille et Pôle de Biologie Pathologie Génétique du CHRU de Lille, 59000, Lille, France.

[10] Laboratory Genetic Metabolic Diseases, Department of Clinical Chemistry and Pediatrics, Academic Medical Center, Amsterdam, The Netherlands

Contents

Technical details

Tables

Figures

1. Reagents and chemicals

Acetonitrile was purchased from VWR Chemicals (France), ultrapure water (18 MX) from Millipore (Molsheim, France) and formic acid from Fluka (Saint Quentin Fallavier, France). The chemicals used were of analytical grade. Leucine Enkephalin (Sigma–Aldrich) at a concentration of 2 ng/µL (in acetonitrile/water, 50/50) was used as reference for mass measurements. Poly-DL-alanine was prepared in 50:50 (v/v) water/acetonitrile at 10 mg/L and used for ion mobility cell calibration. The aTRAQ Kit for Amino acid Analysis of Physiological Fluids (Ref. 4442674) was purchased from Sciex (Life Science Holdings, France). HPLC gradient grade methanol was purchased from VWR Chemicals (Fontenay-sous-Bois, France).

2. GAG-derived disaccharides analysis (Adapted from [1])

Heparan sulfate, dermatan sulfate and keratan sulfate were enzymatically digested into disaccharides in a mixture containing 100 mM NH_4Ac (pH7.0), 10 mM Ca(Ac)2, 2 mM DTT, 5 mIU each of heparinase I, II, III, 50 mIU chondroitinase B, 10µl of KerII and 50 µL urine diluted to 2 mM creatinine in a final volume of 150 µL. After 2 h of incubation at 30°C, 15 µL of 150 mM EDTA (pH7.0) was added along with 125 ng of the internal standard, 4UA-2S-GlcNCOEt-6S (HD009, Iduron, Manchester, UK), and the reaction was stopped by boiling for 5 min to denature the proteins. The reaction mixture was centrifuged at 20,000×g for 5 min at room temperature. The supernatant was subsequently applied to an Amicon Ultra 30 K centrifugal filter (Millipore) and centrifuged at 14,000×g for 15 min at 25°C. The filtrate was stored at -20°C until analysis. The disaccharides were quantified on an Acquity UPLC system (UPLC-MS/MS) coupled to a Waters Quattro Premier XE (tandem) mass spectrometer (Waters Corporation, Milford, MA, USA) with a ESI source operated in negative ionization mode. Source parameters were as follows: capillary voltage 3.5 kV, source temperature 130°C, desolvation temperature 350°C, cone gas flow 50 L/hr, desolvation gas flow 900 L/hr. Collision gas pressure was 2.5*10−3 mbar. The disaccharides were separated on a Thermo Hypercarb HPLC column (100×2.1 mm, 5 µm). The mobile phase consisted of 10 mM NH_4HCO_3 (pH10), and the disaccharides were eluted with an acetonitrile gradient of 0% to 20% for 2.5 min, held at 20% for the next 2.5 min, with 2 min of equilibration at 0% before the next injection; the flow rate was 0.2 mL/min, and the total run time was 7.1 min. All disaccharides were detected and quantified in the MRM acquisition mode, using the transition *m/z* 378.1>175.1, cone voltage of 25 V and collision energy of 14 V for D0A0, *m/z* 416.1>138.0, cone voltage of 40 V and collision energy of 22 V for D0S0, *m/z* 458.1>97.0, cone voltage of 40 V and collision energy of 34 V for D0A6 and D2A0, *m/z* 496.0>416.0, cone voltage of 25 V and collision energy of 16 V for D2S0 and D0S6, *m/z* 458.0>299.9, cone voltage of 35 V and collision energy of 22 V for D0a4, *m/z* 538.0>458.0, cone voltage of 20 V and collision energy of 15 V for D0a10, *m/z* 462.1>361, cone voltage of 55 V and collision energy of 25 V for g0A6 and g6A6 (which could be separated by their retention time of 4.7 and 5.2 min, respectively) and *m/z* 472.0>97.0, cone voltage of 45 V and collision energy of 25 V for the 4UA-2S-GlcNCOEt-6S internal standard.

The identity of the peaks for the KS disaccharides was confirmed by analysis of samples spiked with 13C-labelled g0A6 (Glycosyn, New Zealand) and g6A6 and for HS and DS disaccharides with samples spiked with unlabeled disaccharide solutions. The concentration of the disaccharides was calculated using a calibration curve of each disaccharide with 4UA-2S-GlcNCOEt-6S (HD009, Iduron) as an internal standard. The following disaccharides were analyzed and used to calculate GAG concentrations: D0A0, D0S0, D0A6, D2A0, D0S6 and D2S0 (HS), D0a4 and D0a10 (DS), g0A6 and G6A6 (KS).

3. LC-ESI-MS/MS methodology for determination of amino acids

The aTRAQ Kit for Amino acid Analysis of Physiological Fluids was purchased from Sciex (Framingham, MA, USA). It consisted of amine-modifying labeling aTRAQ reagent Δ8, aTRAQ internal standard set of amino acids labeled with the aTRAQ reagent Δ0, 10 % sulfosalicylic acid, borate buffer of pH 8.5, 1.2 % hydroxylamine and mobile phase modifiers – formic acid and heptafluorobutyric acid. HPLC gradient grade methanol was purchased from J.T. Baker (Center Valley, PA, USA). Deionized water obtained from Millipore Simplicity UV water purification system (Waters Corporation, Milford, MA, USA) was used. Amino acids standards were purchased from SIGMA-ALDRICH (acidic and neutral amino acids Ref: A6407-5ML. Basic amino acids Ref: A6282-5ML).

The following protocol was used for preparation of urine samples. The urine were first diluted using 50/50 v/v using deionized water. An aliquot of 40 µl of the sample was added to 10 µl of 10% sulfosalicylic acid in order to precipitate proteins. After mixing and centrifugation (10 000 x g for 2 min) the supernatant was mixed with 40 µl of borate buffer. Next, an aliquot of 10 µl the obtained solution was labeled with aTRAQ reagent solution (aTRAQ reagent Δ8), mixed and centrifuged. After 30 min of incubation at room temperature the labeling reaction was stopped by addition of 5 µl 1.2% hydroxylamine solution and the sample was incubated at room temperature for 15 min. In the next step, 32 µl of the internal standard solution was added to the sample. After mixing and centrifugation the sample was evaporated in a vacuum concentrator for 15 min in order to reduce volume to about 20 µl. Then the residue was diluted with 20 µl of water. Each determined amino acid had its corresponding internal standard (the same amino acid labeled with the aTRAQ reagent Δ0). Two non-proteinogenic amino acids (norleucine and norvaline) were used to evaluate the labeling efficiency and recovery. A calibration curve was constructed using five concentrations derived from the peak area ratio of each amino acid and the internal standard. A calibration curve has been used as follows:

Amino acids (µM)	**5**	**25**	**100**	**250**	**500**	**1200**
Glutamine (µM)	**10**	**50**	**200**	**500**	**1000**	**2500**
Cystine (µM)	**5**	**25**	**100**	**250**	**500**	

Concentrations were calculated from these area ratios using the calibration curve established by simple regression. The determination of free amino acid levels was conducted using the liquid chromatography instrument Prominence Shimadzu UFLC system consisting of a DGU-20A3 degasser, a LC-20AB pump, a SIL-20ACHT autosampler, and a CTO-20AC oven (Shimadzu, Prominence, Kyoto, Japan) coupled to the 4000 QTRAP mass spectrometer (Sciex, Framingham, MA, USA) with an electrospray ion source. The chromatographic separation was achieved on Sciex C18 column (5 µm, 4.6 mm x 150 mm) maintained at 50 °C with a flow rate of 800 µL/min. A mobile phase gradient of eluent A (0.1% formic acid and 0.01% heptafluorobutyric acid in water) and eluent B (0.1% formic acid and 0.01% heptafluorobutyric acid in methanol) was applied. A gradient profile was the following: from 2% to 40% of B from 0 till 6 min, maintained at 40% of B for 4 min, then increased to 90% of B till 11 min and held at 90% of B for 1 min. After 12 min the gradient decreased to 2% of B. From 13 to 18 min, the mobile phase composition was unaltered. The injection volume was set at 2 µl. The ion source settings were: curtain gas, 20 psig; ion spray voltage, 4500 V; source temperature, 600 °C; ion source gas 1, 60 psig and ion source gas 2, 50 psig. The mass spectrometer operated in positive ionization mode with the following parameters: entrance potential, 10 V; declustering potential, 30 V and collision cell exit potential, 5 V. Collision energy of 30 eV was applied. The list of measured MRM transitions are presented in **Table S1**. Scheduled multiple reaction monitoring mode was used with nitrogen as a collision gas. A system suitability test was conducted before each batch of the samples (analysis of a standard mixture) to warm up the LC-MS/MS system and check the inter-day performance of the system. Data acquisition and processing were performed using the Analyst 1.5 software (Sciex, Framingham, MA, USA).

Table S1. MRM transitions for each amino acid and its corresponding internal standard.

	Analyte		Internal standard	
Amino acid	**Q1 (*m/z*)**	**Q2 (*m/z*)**	**Q1 (*m/z*)**	**Q2 (*m/z*)**
Alanine	238.2	121.1	230.2	113.1
Arginine	323.2	121.1	315.2	113.1
Asparagine	281.2	121.1	273.2	113.1
Aspartic acid	282.1	121.1	274.1	113.1
Citruline	324.2	121.1	316.2	113.1
Cysteine	537.2	121.1	521.2	113.1
Glutamate	296.2	121.1	288.2	113.1
Glutamine	295.2	121.1	287.2	113.1
Glycine	224.1	121.1	216.1	113.1
Histidine	304.2	121.1	296.2	113.1
Isoleucine	280.2	121.1	272.2	113.1
Leucine	280.2	121.1	272.2	113.1
Lysine	443.3	121.1	427.3	113.1
Methionine	298.2	121.1	290.2	113.1
Ornitine	429.3	121.1	413.3	113.1
Phenylalanine	314.2	121.1	306.2	113.1
Proline	264.2	121.1	256.2	113.1
Serine	254.2	121.1	246.2	113.1
Taurine	274.1	121.1	266.1	113.1
Threonine	268.2	121.1	260.2	113.1
Tryptophane	353.2	121.1	345.2	113.1
Tyrosine	330.2	121.1	322.2	113.1
Valine	266.2	121.1	258.2	113.1

4. UHPLC-IM-MS analysis

4.1.1.1.1. Sample handling

The sample handling component was a Waters 2777C sample manager (Waters Corp., Milford, MA, USA) equipped with a 25 µl Hamilton syringe, a 2 µl loop used for full loop injections of prepared sample, and a 2-drawer sample chamber thermo-stated at 4°C with a constant flow of dry nitrogen gas to prevent the buildup of condensation.

4.1.1.1.2. Chromatographic conditions

The chromatography was performed on a Waters NanoAcquity UPLC module (Saint Quentin en Yvelines, France) upgraded to work with 1 mm columns and composed with a binary solvent manager and column heater/cooler module. Separation was carried out at 45 °C using a 1.0 x 100 mm, Acquity UPLC HSS T3 column (Waters), with a particle size of 1.8 µm, equipped with a 0.2 µm prefilter. Urine was eluted from the LC column using the following linear gradient (curve number 6): 0–1 min: 99% A; 1–3 min, 99–85% A; 3–6 min, 85-50% A; 6–9 min, 50-0% A; 9–12 min, 100% B, 12–16 min, 99% A for re-equilibration. Solvent A was water and solvent B was acetonitrile, both solvents contained 0.1% formic acid. The duration of column equilibration was adjusted to provide sufficient retention and chromatographic precision of early eluting species in subsequent analyses at the minimal expense of time. Sample analysis order has been randomized to avoid potential for confounding critical variables with analytical run order effects. Peak splitting and column overload was avoided by using small injection volumes (2 µL) for LC-IM-MS analysis.

4.1.1.1.3. Mass spectrometry

The U-HPLC system was coupled to a hybrid quadrupole orthogonal time-of-flight (TOF) mass spectrometer (SYNAPT G2 HDMS, Waters MS Technologies, Manchester, UK). The mass spectrometer was operated in positive electrospray ionization mode. A mass range of *m*/*z* 50−1200 was used in both modes. The sample cone voltage, extraction cone voltage, source temperature, desolvation temperature, desolvation gas flow and cone gas flow were optimized and were as follows respectively: 25V, 5V, 120°C, 500°C, 400 L/h, 50 L/h. Leucine enkephalin was used as the lock mass $[M+H]^+$ at *m*/*z* 556.2771. Sodium formate solution was used for external instrument calibration.

4.1.1.1.4. Ion mobility

Synapt G2 HDMS (Waters MS Technologies, Manchester, UK) was used in our study for Ion Mobility. It is equipped with a traveling wave "Triwave™" geometry in which the ion mobility cell (IMS T-wave) is placed between two traveling wave ion guides (trap T-wave and transfer T-wave). After ionization in the source and transfer through the quadrupole, the ions arrive at the first traveling-wave ion guide that acts as an ion trap, namely "trap TWIG". In this region, the ions are accumulated before being released in packets and accelerated using the trap-bias voltage to the second cell "IMS-TWIG" for mobility separation. In the IMS-TWIG a traveling wave is continuously applied at a given wave height and velocity to enhance separation through the mobility cell, which is filled with a gas. In this study, the IMS drift gas flow (nitrogen) and the wave velocity settings were assessed and optimized. The helium cell gas flow, wave height, Trap Bias and IMS wave delay were set at 180 mL/min, 40 V, 45 V and 450 µs respectively. The TOF analyzer was operated in the *V* resolution mode with an average mass resolution of m/Δm 20,000 (full-width at half-maximum definition). Data acquisition of an ion mobility experiment consisted of 200 bins. CCS values, obtained in nitrogen, were experimentally determined using singly charged Poly-DL-alanine oligomers as the TWIM calibrant species for ESI+. CCS values were derived according to previously reported procedures [2]. The ion mobility resolution was ~40 Ω/ΔΩ (fwhm). The N_2 CCS values reported were determined at the apex of the ion-mobility peak. Detailed instrument settings are presented in Table S-2.

4.1.1.1.5. Raw data preprocessing

All LC-IM–MS raw data files data processing, peak detection and peak matching across samples using retention time (tR) correction and chromatographic alignment along with drift time and CCS calculation were performed using Progenesis QI (Waters MS Technologies, Manchester, UK) to yield a data matrix containing retention times, accurate masses, CCS and peak intensities. The preprocessing step resulted in an X-matrix where tR, CCS and *m/z* values were concatenated into ''tR_m/z_CCS'' features (in columns) present in each sample (in rows) with corresponding peak areas.

4.1.1.1.6. Quality Control

Aliquoted 10 µL of each urine sample are mixed together to generate a pooled quality control sample (QCs). QCs and solvent blank samples (mobile phase) were injected sequentially in-between the urine samples. In addition, a dilution series of QC samples (6%, 12.5%, 25%, 50% and

100% original concentration) are used to assess the quality of the extracted features. An analysis sequence is presented in Figure. S-1. Indeed, feature extraction algorithms including automatic peak detection, grouping, and integration often yield a data matrix containing analytical system noise such as mobile phase chemical contaminants signals. Depending on the software and the used parameters for feature extraction, such noise can represent a significant portion of the total number of detected features. Therefore, this may mislead further data analysis such as transformation, normalization and data modeling. Simple noise filtering strategies such as the minimum fraction filter may remove infrequently observed signals within the considered distinct sample classes. In this study, we used a filter strategy in which the features intensity must be correlated to the matrix concentration in a series of diluted QC samples in order to be included in further analysis. Beyond its role as a system noise marker, the dilution series filter is very useful to assess the informative quality of the extracted features. It ensures that the observed signal of a given feature and its relative concentration in the sample are positively correlated. This approach is used to identify feature groups that are not correlated to the gradient of concentration generated by the dilutions series and therefore should not be considered as reliable features. Thus, feature groups with correlation coefficient of less than 0.7 were removed from the dataset. Furthermore, datasets are refined by removal of feature groups that do not meet threshold of peak area measurement precision prior to data analysis. This approach uses RSD values derived from repeated measurements of a pooled QC sample. The threshold was set to RSD<25%. Thus enhancing the biological interpretation of metabolomics data. The system stability assessment has been done using the Principal Component Analysis and assessing the clustering of the QC samples. Figures are shown in supporting information (Figures S-2) presenting the PCA score plot derived from the metabolomics analysis of all the urine samples and QCs replicates. In the PCA score plot, each point corresponds to a different sample or QC. The tightness of the clustering reveals the similarity of the samples (QCs), and thus the system's stability. The QCs are tightly clustered indicating a good instrumental stability over the metabolomics analysis.

Table S2. Instrumental settings for UHPLC-IM-MS analysis.

Chromatography		Column Temperature	45 (°C)
		Flow rate	80 (µL/min)
		Injection volume	2 µL
ESI-MS		Capillary voltage	2 kV
		Sampling cone voltage	25 V
		Extraction cone voltage	5
		Source temperature	120 (°C)
		Desolvation temperature	500 (°C)
		Desolvation gas flow	400 (liters/h)
		Cone gas flow	50
		Optic mode	Resolution
		MS scan rate	0.2 scan/s
		Lock mass solution	Leu–enk (2 µg/ml)
		Lock mass flow rate	6 µL/min
Triwave DC	Trap DC	Entrance	3
		Bias	45
		Trap DC	0
		Exit	3
	IMS DC	Entrance	25
		Helium cell DC	35
		Helium exit	−5
		Bias	3
		Exit	0
	Transfer DC	Entrance	4
		Exit	15
Gas controls	IMS gas	Nitrogen	80 (ml/min)
	Helium cell	Nitrogen	180 (ml/min)
Triwave	Trap	Wave velocity	300 (m/s)
		Wave height	0.5 (V)
	IMS	Wave velocity	857 (m/s)
		Wave height	40 (V)
	Transfer	Wave velocity	300 (m/s)
		Wave height	5 (V)

5. Data analysis and modeling

Support vector regression normalization method was applied using the MetNormalizer R package [3] before any data analysis, to remove the unwanted intra- and inter-batch measurements analytical variations. The effect of this normalization step on the raw data is shown in Figures S-2, S-3 and S-4 (Supporting information). Then normalized data matrix has been log-transformed and pareto-scaled. All data analyzes and modeling were done using SIMCA 14.0 (MKS DAS, Umeå, Sweden). First, hierarchical cluster analysis has been applied to the data set to get an overview of the clustering trends of samples with similar profiles of variable intensity. Furthermore, multivariate data analysis and modeling is performed using Principal Component Analysis (PCA) as an unsupervised method. PCA was first applied to get an overview of the data and identify potential severe outliers which are defined as observations whose scores mapped outside the Hotelling's T2 ellipse (confidence interval = 0.95) in a cross-validated seven-component model. The DmodX was used to detect moderate outliers [4]. Orthogonal Partial Least-Squares-Discriminant Analysis (OPLS-DA) is used as a supervised method. To select the most relevant features, the training group has been repeatedly split into a training set and a test set. A permutation test (999 iterations) was performed to prevent the OPLS-DA over fitting of the model by comparing diagnostic statistic metrics of the generated model with those of randomly generated models. R2X is the cumulative modeled variation in X (X = features), R2Y is the cumulative modeled variation in Y (Y = sample groups), and Q2Y is the cumulative predicted variation in Y, based on the cross-validation. The range of these parameters is between 0 and 1, where 1 indicates a perfect fit. Furthermore, cross-validated analysis of variance (CV-ANOVA) was systematically performed based on the cross-validated model [5]. The Figure S-2 shows a PCA score plot of the raw data compared to normalized, log-transformed and pareto-scaled data which highlights the importance of these data pretreatment steps before data modeling. The X matrix was 100 x 854 variables. The Y matrix was 100 analyses × 3 groups. Characteristics and validation results from all OPLS-DA models are provided in supplemental material (Figures S-5, S-6 and S-7).

6. Feature selection and annotation

To select the most discriminant variables for the separation of groups, S-Plot was used. The S-plot combines the covariances and correlations between the **X** matrix and OPLS scores for a given model component. The covariance values give the magnitude of contribution of a variable while the correlation values reflect the effect and reliability of the variable for the model component scores. Variables with both very high correlation and covariance are important for the explanation power of the model. Furthermore, selection of discriminant variables was achieved using the VIP score procedures for each validated OPLS-DA model [6]. Putative annotation of detected features was performed using both accurate mass comparison using freely available metabolite databases HMDB, LipidBlast, KEGG, and Metlin. Furthermore, CCS values were also compared to the MetCCS database [7].

7. Pathway and network analysis

In order to provide a broader understanding of metabolic changes in MPS I, we also explored the biochemical pathways using a network analysis approach using Mummichog (v.1.0.5) which allows pathway enrichment analyses. The idea behind this metabolic network prediction strategy assumes that metabolite concentration alterations are more likely to occur within a metabolic connected network rather than in a random fashion. This Mummichog python package highlights pathways that are significantly impacted in the studied groups. Significantly impacted biochemical pathways are those exhibiting an adjusted p-value <0.05. For this comparison, we focused on features that significantly changed (511 features with q-values = 0.05 and FDR = 5%). Mummichog annotates metabolites based on accurate mass *m*/*z* (5 ppm mass error was used) and tests significant pathway enrichment within a reference metabolic network using a Fisher's exact test [8]. The matched candidates were then mapped to reference human metabolic networks from the KEGG, MetaCyc, Recon and Edinburgh Human Metabolic Network. The null distribution in pathway analysis was obtained from 1000 set of randomly permutated *m*/*z* lists draw from all features detected in the whole metabolomic dataset and modeled by Gamma distribution. To protect against incorrect pathway selection, redundant pathways or those enriched by less than two metabolites were excluded. MetaboAnalyst [9] has been used for Metabolite Set Enrichment Analysis using the amino acid concentration matrix.

RÉFRÉRENCES

1. Langereis, E.J.; Wagemans, T.; Kulik, W.; Lefeber, D.J.; van Lenthe, H.; Oussoren, E.; van der Ploeg, A.T.; Ruijter, G.J.; Wevers, R.A.; Wijburg, F.A., *et al.* A multiplex assay for the diagnosis of mucopolysaccharidoses and mucolipidoses. *PloS one* **2015**, *10*, e0138622.
2. Paglia, G.; Angel, P.; Williams, J.P.; Richardson, K.; Olivos, H.J.; Thompson, J.W.; Menikarachchi, L.; Lai, S.; Walsh, C.; Moseley, A., *et al.* Ion mobility-derived collision cross section as an additional measure for lipid fingerprinting and identification. *Analytical chemistry* **2015**, *87*, 1137-1144.
3. Shen, X.; Gong, X.; Cai, Y.; Guo, Y.; Tu, J.; Li, H.; Zhang, T.; Wang, J.; Xue, F.; Zhu, Z.-J. Normalization and integration of large-scale metabolomics data using support vector regression. *Metabolomics* **2016**, *12*, 89.
4. Eriksson, L.; Trygg, J.; Wold, S. A chemometrics toolbox based on projections and latent variables. *Journal of Chemometrics* **2014**, *28*, 332-346.
5. Eriksson, L.; Trygg, J.; Wold, S. Cv-anova for significance testing of pls and opls® models. *Journal of Chemometrics* **2008**, *22*, 594-600.
6. Galindo-Prieto, B.; Eriksson, L.; Trygg, J. Variable influence on projection (vip) for orthogonal projections to latent structures (opls). *Journal of Chemometrics* **2014**.
7. Zhou, Z.; Xiong, X.; Zhu, Z.J. Metccs predictor: A web server for predicting collision cross-section values of metabolites in ion mobility-mass spectrometry based metabolomics. *Bioinformatics (Oxford, England)* **2017**.
8. Li, S.; Park, Y.; Duraisingham, S.; Strobel, F.H.; Khan, N.; Soltow, Q.A.; Jones, D.P.; Pulendran, B. Predicting network activity from high throughput metabolomics. *PLoS Comput. Biol.* **2013**, *9*, e1003123.
9. Xia, J.; Sinelnikov, I.V.; Han, B.; Wishart, D.S. Metaboanalyst 3.0-making metabolomics more meaningful. *Nucleic acids research* **2015**.
10. Salazar, D.A.; Rodriguez-Lopez, A.; Herreno, A.; Barbosa, H.; Herrera, J.; Ardila, A.; Barreto, G.E.; Gonzalez, J.; Almeciga-Diaz, C.J. Systems biology study of mucopolysaccharidosis using a human metabolic reconstruction network. *Molecular genetics and metabolism* **2016**, *117*, 129-139

Table S3. The normalized concentrations of free amino acids in urine samples of the three studied groups.

(µM amino acid / mM creatinine)

	Control group (n=66)			MPS I group (n = 19)			MPS IT group (n = 15)		
	Mean	Median	Range	Mean	Median	Range	Mean	Median	Range
L-Alanine	22.4	16.9	0.1 - 83.3	63.0	43.7	5.9 - 244.5	26.3	14.0	0.8 - 122.5
L-Arginine	2.2	1.0	0 - 27.9	12.3	9.0	1.6 - 50.7	9.4	7.3	0 - 28.3
L-Citrulline	1.9	1.0	0 - 17.1	4.2	3.9	0 - 12.9	2.0	1.4	0 - 10
L-Cystine	4.8	3.6	0 - 22.7	9.2	4.0	0.1 - 65.6	4.2	1.7	0.3 - 15.8
L-Glutamine	38.4	30.5	0.2 - 138.8	103.6	93.3	2.7 - 528.1	48.2	32.5	2.4 - 184.2
L-Glutamic acid	3.5	2.4	0 - 19.5	10.0	6.5	1.5 - 25.8	5.0	2.6	1 - 20.8
Glycine	112.3	78.0	0.5 - 448.5	265.1	164.8	37 - 1322.7	157.3	136.3	5.5 - 436.7
L-Histidine	52.2	43.5	0.3 - 147.7	133.5	68.6	7.6 - 810.9	72.7	67.8	2.7 - 267.5
L-isoleucine	1.8	1.2	0 - 11.8	4.3	2.5	0.1 - 12.5	0.9	0.3	0 - 5
L-Leucine	3.3	2.4	0 - 19.2	7.1	5.2	1.3 - 35.9	3.3	2.0	0.4 - 17.5
L-Lysine	13.6	7.5	0.2 - 70	24.8	23.5	4.2 - 66.8	15.9	5.0	1.6 - 82.5
L-Methionine	0.8	0.5	0 - 6.1	1.1	0.4	0.1 - 8.6	0.3	0.4	0 - 1
L-Ornithine	5.1	1.2	0 - 91	6.1	1.7	0.1 - 27.1	7.6	2.9	0 - 42.5
L-Phenylalanine	5.1	4.1	0 - 23.7	13.0	8.0	0.1 - 82.8	4.4	2.8	0 - 22.5
L-Proline	2.0	1.1	0 - 14.1	7.0	3.6	0.4 - 44.6	3.0	1.6	0.5 - 10.6
L-Serine	32.2	25.9	0 - 119.8	84.4	68.4	0.1 - 379.7	41.5	25.1	0 - 174.2
L-Tyrosine	8.3	6.2	0 - 39.4	19.9	11.8	0.1 - 117.2	6.3	4.1	0 - 31.7
L-Valine	4.2	3.2	0 - 22.6	10.8	6.8	1.7 - 42.2	4.4	2.6	0.8 - 20
L-Aspartic acid	1.2	0.4	0 - 8.8	2.8	2.7	0.4 - 6.8	1.9	0.5	0 - 20
Taurine	33.3	31.0	0 - 106.8	148.4	27.9	0.1 - 1704.7	111.4	40.0	1.6 - 821.7
L-Threonine	13.2	8.4	0.8 - 71	27.4	17.4	3 - 136.2	15.4	6.2	1 - 70.8
L-Asparagine	11.9	6.7	0 - 89.6	22.3	13.2	0.1 - 67.7	9.1	5.9	0 - 48.7
Cystathionine	2.9	2.5	0 - 9.8	11.4	7.3	0.1 - 48.4	3.2	2.8	0 - 9.2
L-Tryptophane	4.4	4.2	0 - 9.6	13.2	8.7	0.6 - 94.5	5.2	3.6	0.2 - 23.3

Table S4. Data for venn diagram of the significantly pathways retrieved from untargeted, targeted approaches and *in silico* systems biology approach from Salazar DA et al [10].

Data	N°	Pathways
In silico **Targeted Untargeted**	**2**	Glutathione Metabolism Arginine and Proline Metabolism
In silico **Targeted**	**4**	Glycerophospholipid Metabolism Tyrosine Metabolism Methionine and Cysteine Metabolism Lysine Metabolism
In silico **Untargeted**	**4**	Histidine Metabolism Taurine and Hypotaurine Metabolism Propanoate Metabolism Butanoate Metabolism
Untargeted	**5**	Ascorbate (Vitamin C) And Aldarate Metabolism Vitamin H (Biotin) Metabolism Biopterin Metabolism Urea Cycle/Amino Group Metabolism
Targeted	**18**	Lysine Biosynthesis Glycine; Serine and Threonine; Metabolism Aminoacyl-trna Biosynthesis Valine; Leucine and Isoleucine Degradation; Nicotinate and Nicotinamide Metabolism; Nitrogen Metabolism; Cyanoamino acid Metabolism; Selenoamino acid Metabolism; Beta-alanine Metabolism; Porphyrin Metabolism; Valine, Leucine and Isoleucine Biosynthesis; Cysteine and Methionine Metabolism; D-arginine and D-ornithine Metabolism; Pyrimidine Metabolism; Purine Metabolism; Pantothenate and CoA Biosynthesis; D-glutamine and D-glutamate Metabolism; Alanine, Aspartate and Glutamate Metabolism.
In silico	**68**	Squalene and Cholesterol Synthesis; Sphingolipid Metabolism Fatty Acid Oxidation; Purine Synthesis; Phosphatidylinositol Phosphate Metabolism; Glycosphingolipid Metabolism; Vitamin B6 Metabolism; Vitamin A Metabolism; Tryptophan Metabolism; Glycolysis/Gluconeogenesis Transport; Lysosomal N-glycan Degradation; Phenylalanine Metabolism; Thiamine Metabolism; CoA Synthesis Transport; Endoplasmic Reticular Vitamin C Metabolism; Cholesterol Metabolism; NAD Metabolism; Pentose Phosphate Pathway; Pyruvate Metabolism; Keratan Sulfate Degradation; Pyrimidine Catabolism; Fructose and Mannose Metabolism; Tetrahydrobiopterin Metabolism; Aminosugar Metabolism; Pyrimidine Synthesis; Galactose Metabolism; Folate Metabolism Transport; Extracellular Oxidative Phosphorylation; Glyoxylate and Dicarboxylate Metabolism; Glutamate Metabolism; ROS Detoxification, Heme Synthesis; Citric Acid Cycle; Inositol Phosphate Metabolism; Keratan Sulfate Synthesis; Purine Catabolism; Bile Acid Synthesis; Biotin Metabolism N-glycan Synthesis Valine, Leucine, and Isoleucine Metabolism Vitamin B2 Metabolism; Urea Cycle; Triacylglycerol Synthesis; Nucleotide Salvage Pathway; Fatty Acid Synthesis; Steroid Metabolism; Androgen and Estrogen Synthesis and Metabolism; Alanine and Aspartate Metabolism; Eicosanoid Metabolism; Cysteine Metabolism; C5-branched Dibasic Acid Metabolism; CoA Catabolism; Starch and Sucrose Metabolism; Glycine, Serine, Alanine and Threonine Metabolism; D-alanine Metabolism.

Figure S1. Illustration of the analysis sequence. Ten µL of each urine sample are mixed together to generate a pooled quality control sample (QCs). QCs and solvent blank samples (mobile phase) were injected sequentially in-between the urine samples. In addition, a dilution series of QC samples (6%, 12.5%, 25%, 50% and 100% of the original concentration) are used to assess the quality of the extracted features. A conditioning step is used to condition the column using ten QC injections. Sample injection order has been orthogonalized.

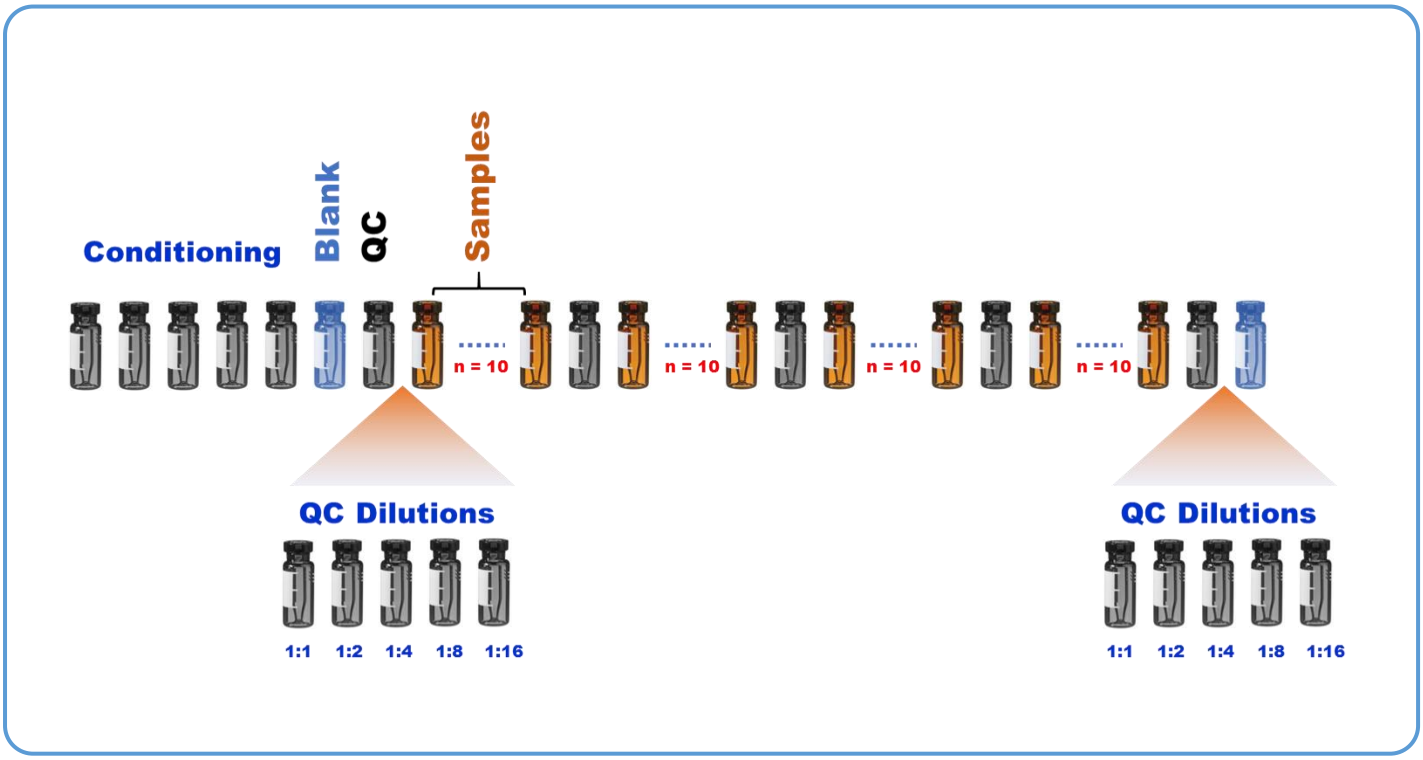


Figure S2. PCA score plot showing the tight clustering of quality control (QC) samples before and after normalization. Grey circle: samples. Green circle: QC samples. Left: raw data. Right: normalized data. The QC samples are tightly clustered after normalization.

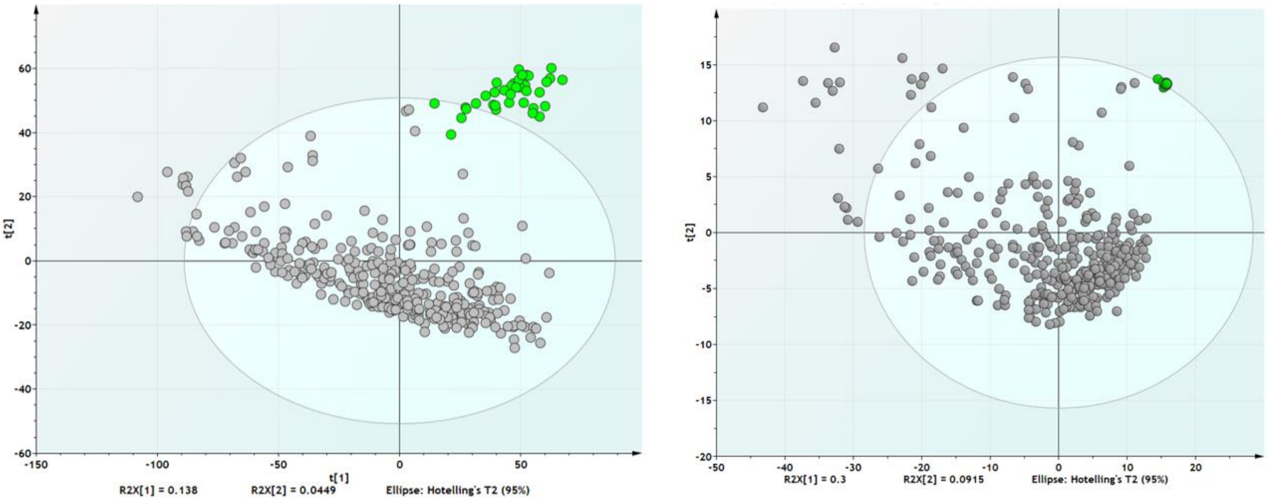


Figure S3 PCA **s**cores showing the effect of the support vector normalization step on the QC samples scores and the drift correction. Blue: first component. Red: second component. Green: third component

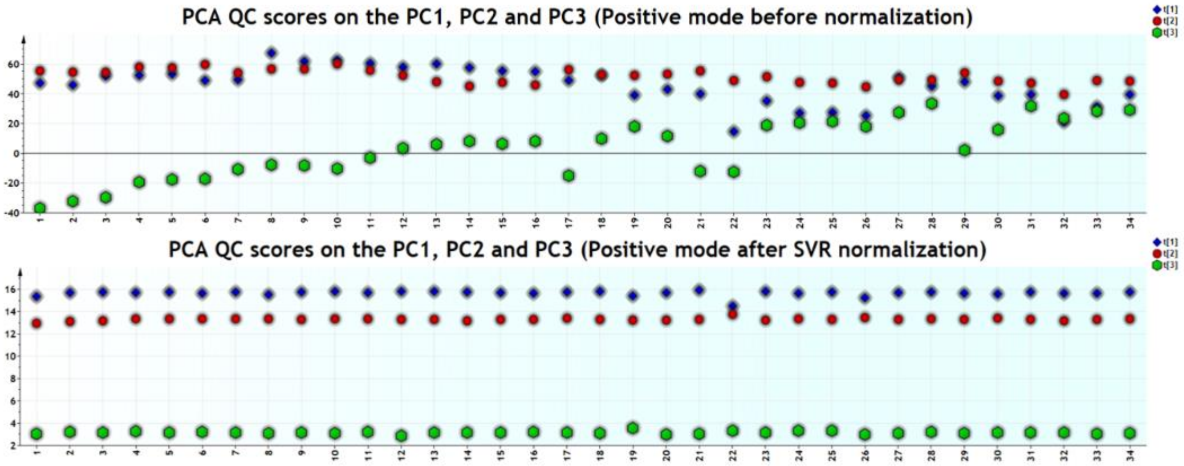


Figure S4. The RSD distribution bar plots before and after support vector normalization step.

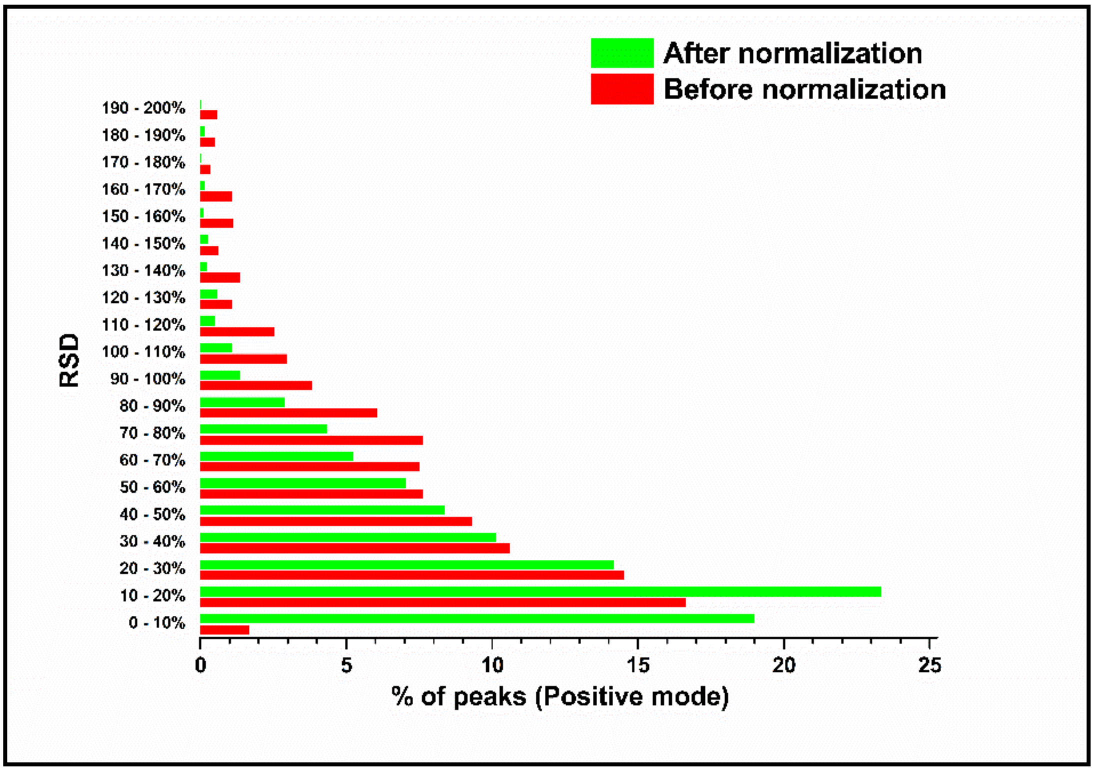
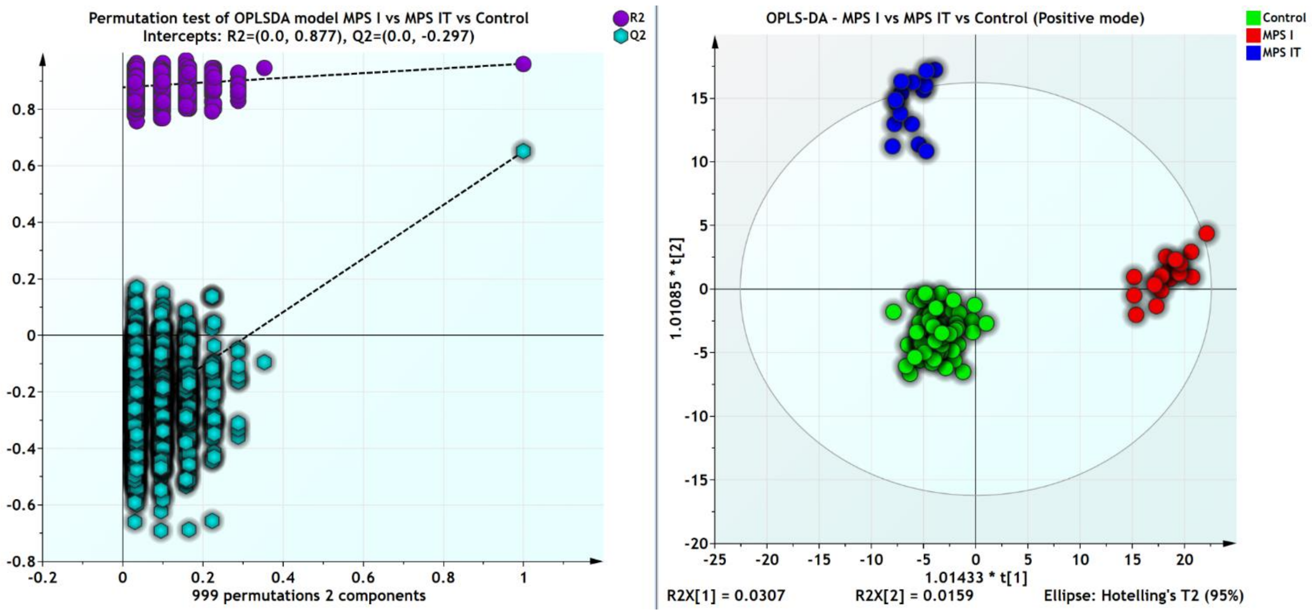


Figure S5. OPLSDA model validation including the three groups: MPS I, MPSIT and Controls. Below the figure, model parameters and CV-ANOVA results are presented.



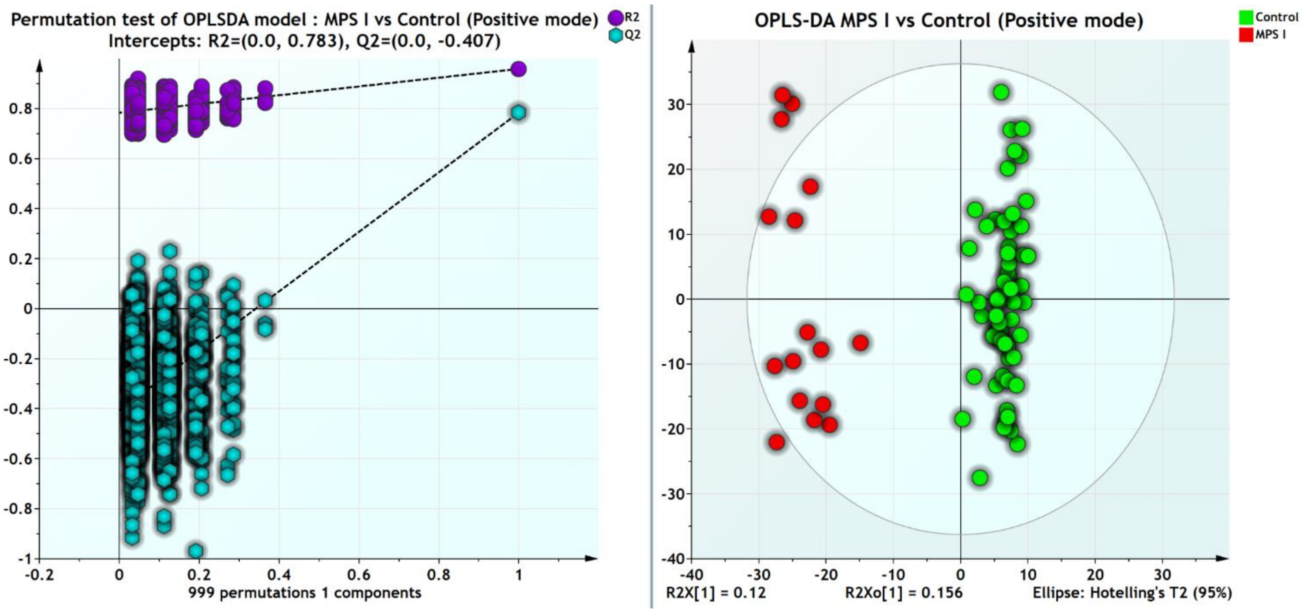
MPS I *vs*. MPSIT *vs*. Control - OPLSDA model parameters.

Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)
Model		0.328			0.961			0.542		1
Predictive		0.0456			0.961			0.542		1
P1	0.0298	0.0298	3.04	0.51	0.51	0.296	0.01	0.296	0.52	0.52
P2	0.0158	0.0456	1.61	0.451	0.961	0.247	0.01	0.542	0.48	1
Orthogonal in X(OPLS)		0.283			0					
O1	0.198	0.198	20.2	0	0					
O2	0.0258	0.224	2.64	0	0					
O3	0.0209	0.245	2.13	0	0					
O4	0.0221	0.267	2.25	0	0					
O5	0.0158	0.283	1.61	0	0					

CV-ANOVA test results.

	SS	DF	MS	F	*p*	SD
Total corr.	202	202	1			1
Regression	110.774	24	4.615	9.005	3.99e-020	2.148
Residual	91.226	178	0.512			0.715

Figure S6. OPLSDA model validation for MPS I *vs*. Control.
Below the figure, model parameters and CV-ANOVA results are presented.



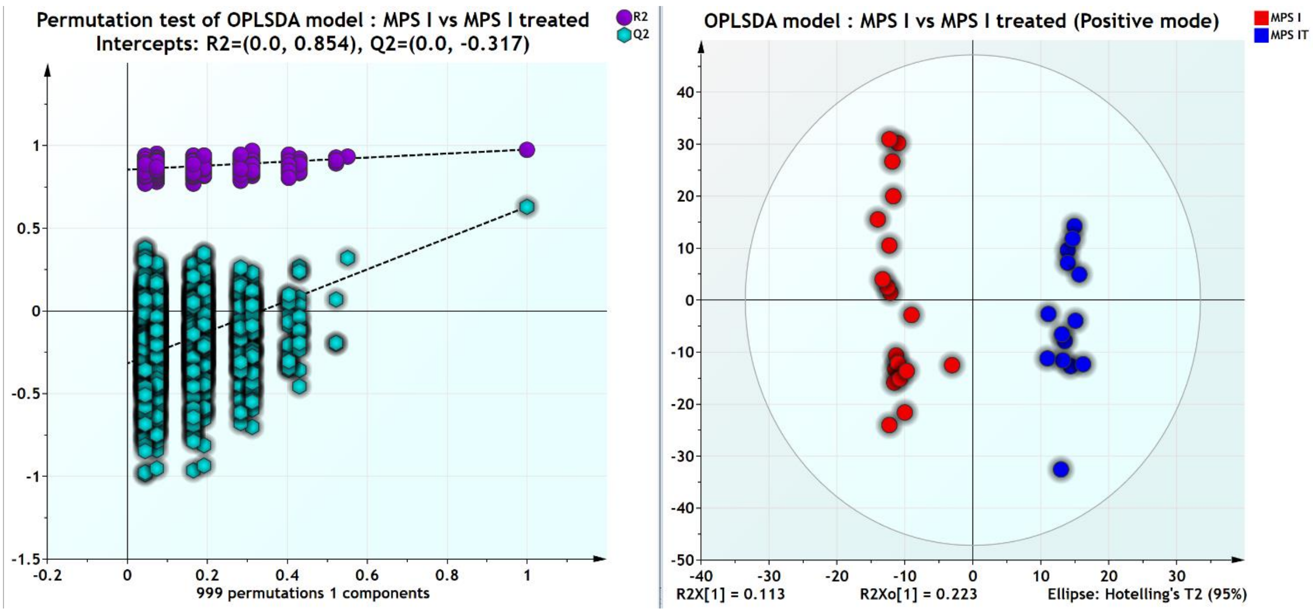
MPS I *vs*. Control - OPLSDA model parameters.

Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)
Model		0.271			0.945			0.631		1
Predictive		0.0329			0.945			0.631		1
P1	0.0329	0.0329	2.86	0.945	0.945	0.631	0.01	0.631	1	1
Orthogonal in X(OPLS)		0.239			0					
O1	0.21	0.21	18.3	0	0					
O2	0.0282	0.239	2.45	0	0					

MPS I *vs*. Control - CV-ANOVA test results.

	SS	DF	MS	F	*p*	SD
Total corr.	**86**	**86**	**1**			**1**
Regression	**54.235**	**6**	**9.039**	**22.766**	**1.75e-015**	**3.006**
Residual	**31.764**	**80**	**0.397**			**0.630**

Figure S7. OPLSDA model validation for MPS I *vs*. MPSIT.
Below the figure, model parameters and CV-ANOVA results are presented.



MPS I *vs*. MPSIT - OPLSDA model parameters.

Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)
Model		0.488			0.975			0.63		1
Predictive		0.113			0.975			0.63		1
P1	0.113	0.113	3.83	0.975	0.975	0.63	0.01	0.63	1	1
Orthogonal in X(OPLS)		0.376			0					
O1	0.223	0.223	7.59	0	0					
O2	0.115	0.338	3.9	0	0					
O3	0.0375	0.376	1.28	0	0					

MPS I *vs*. MPSIT - CV-ANOVA test results.

	SS	DF	MS	F	*p*	SD
Total corr.	33	33	1			1
Regression	20.785	8	2.598	5.3179	0.000580	1.611
Residual	12.214	25	0.488			0.698

Figure S8. Boxplots of selected discriminant features in the three groups: MPS I, MPS IT and Control samples.

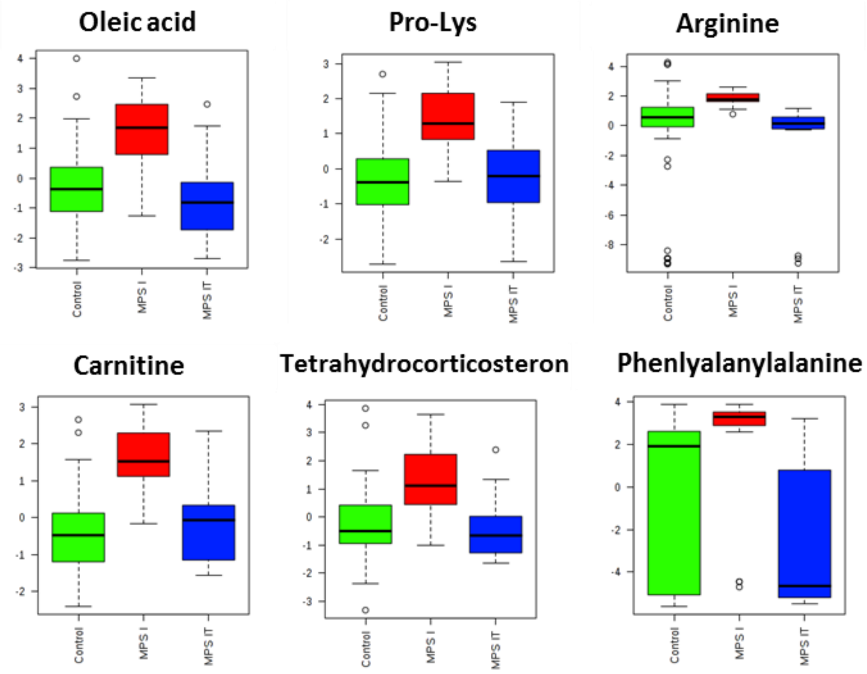
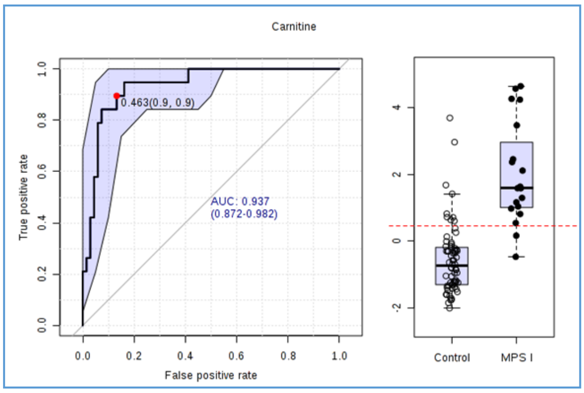
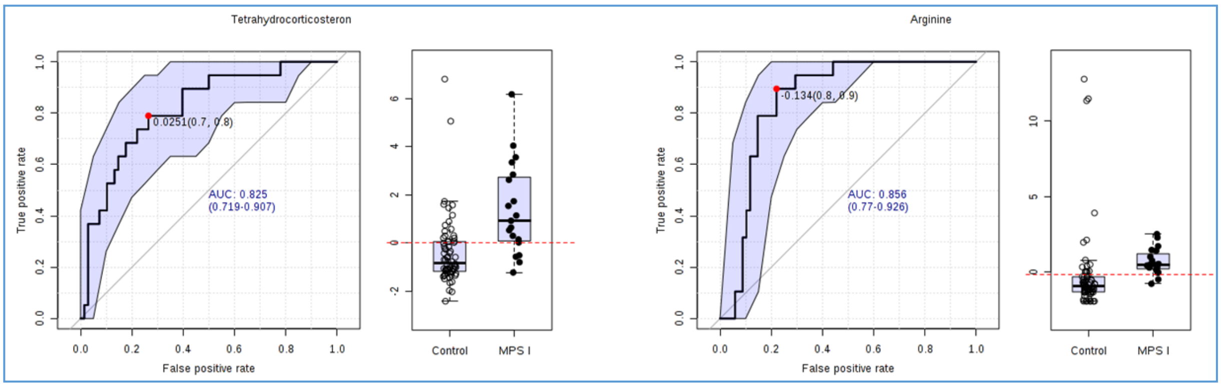


Figure S9. Area under the receiver operating characteristic (ROC) curves, comparing diagnostic performance of different discriminant features retrieved from untargeted metabolomics to differentiate MPS I from Control samples. Up) Arginine, Carnitine and Tetrahydrocorticosteron. Down) Comparison of different combinations of features using PLSDA model with three components each. Combining features does not show improvement in AUC.

AUC: Area under the curve. False positive rate = 100 – Specificity. True positive rate = Sensitivity.





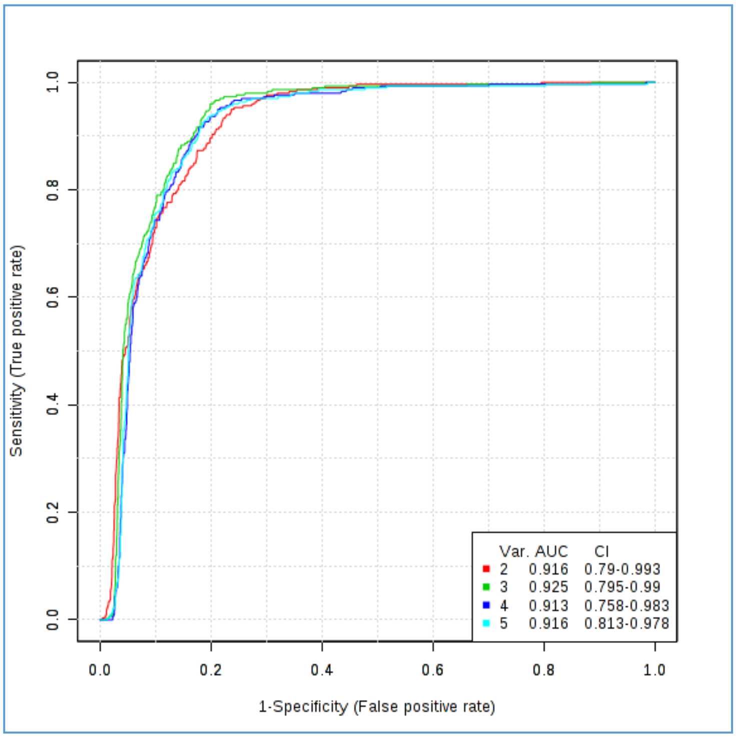


Figure S10. Boxplots of amino acid concentrations across the three studied groups: MPS I, MPS IT and Control samples.

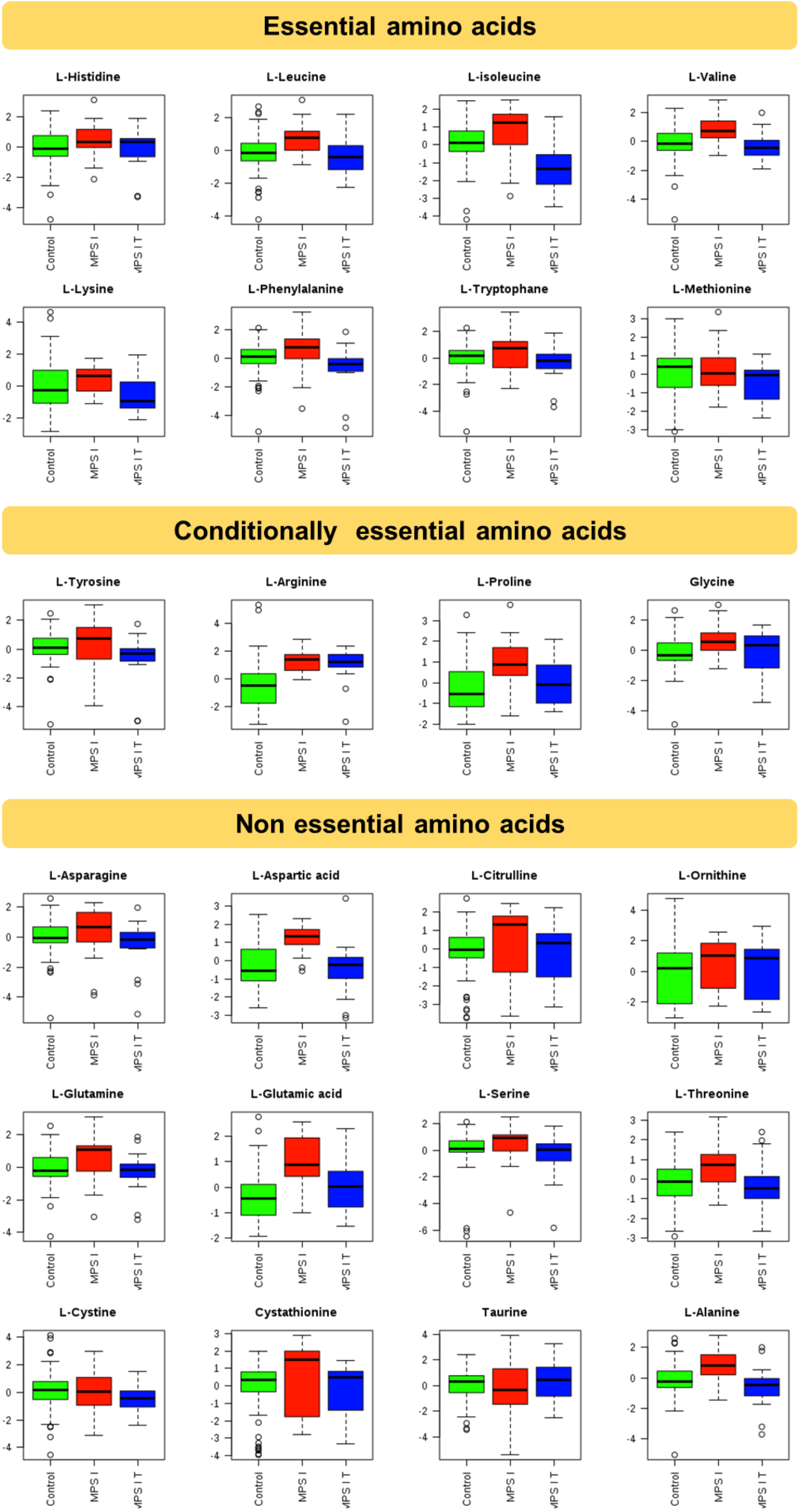


Figure S11. Boxplots of Heparan sulfate, Keratan sulfate, Dermatan sulfate and Total Glycoaminoglycanes (Total GAGs) in the three groups: MPS I, MPS IT and Control samples.

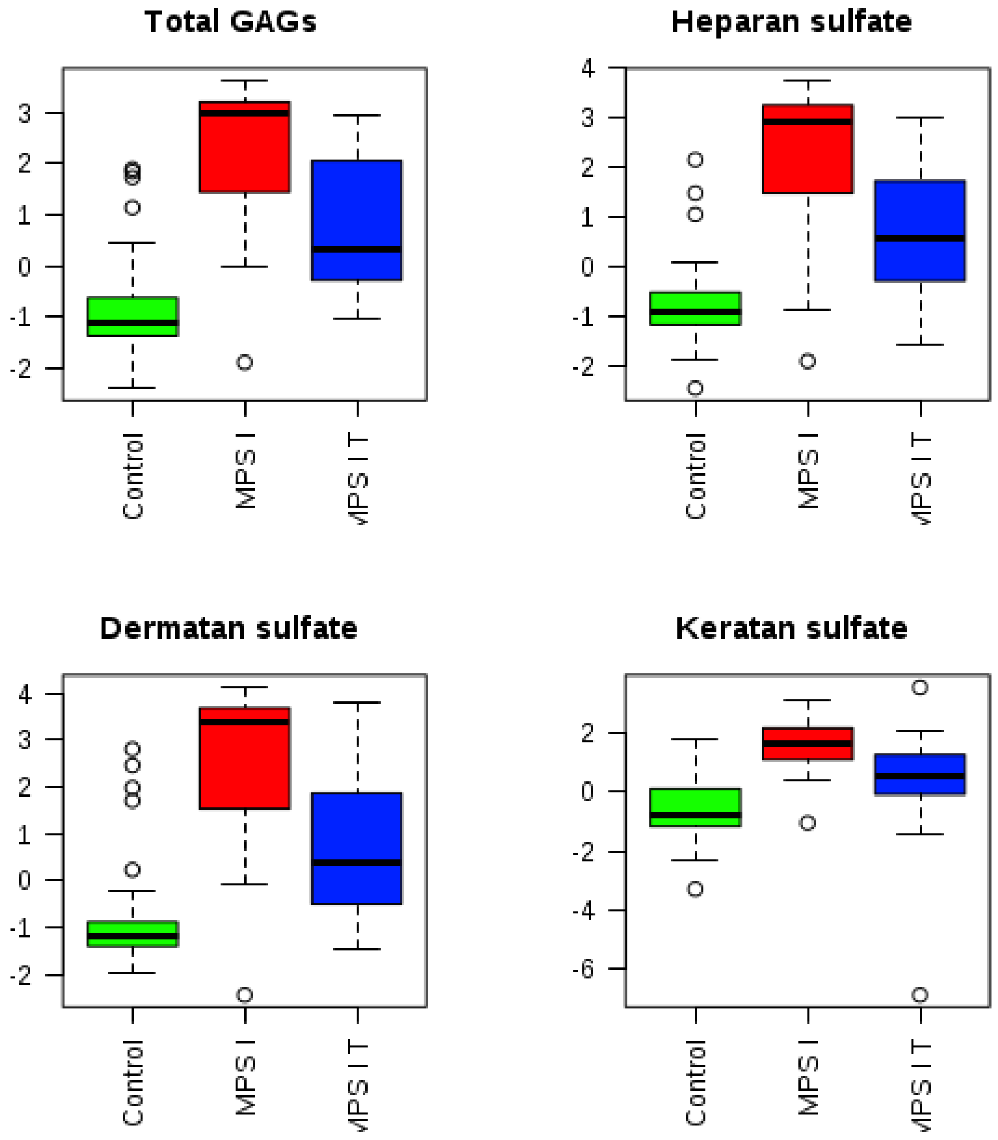


Figure S12. Area under the receiver operating characteristic (ROC) curves, comparing diagnostic performance of the most significant quantified amino acids to differentiate MPS I from Control samples. A) Arginine, Proline, Aspartic acid, and Glutamic acid. B) A comparison of different combinations of the four amino acids using a PLSDA model with three components each is presented. Combining amino acids does not show significant improvement in AUC.
AUC, Area under the curve. False positive rate = 100 – Specificity. True positive rate = Sensitivity.

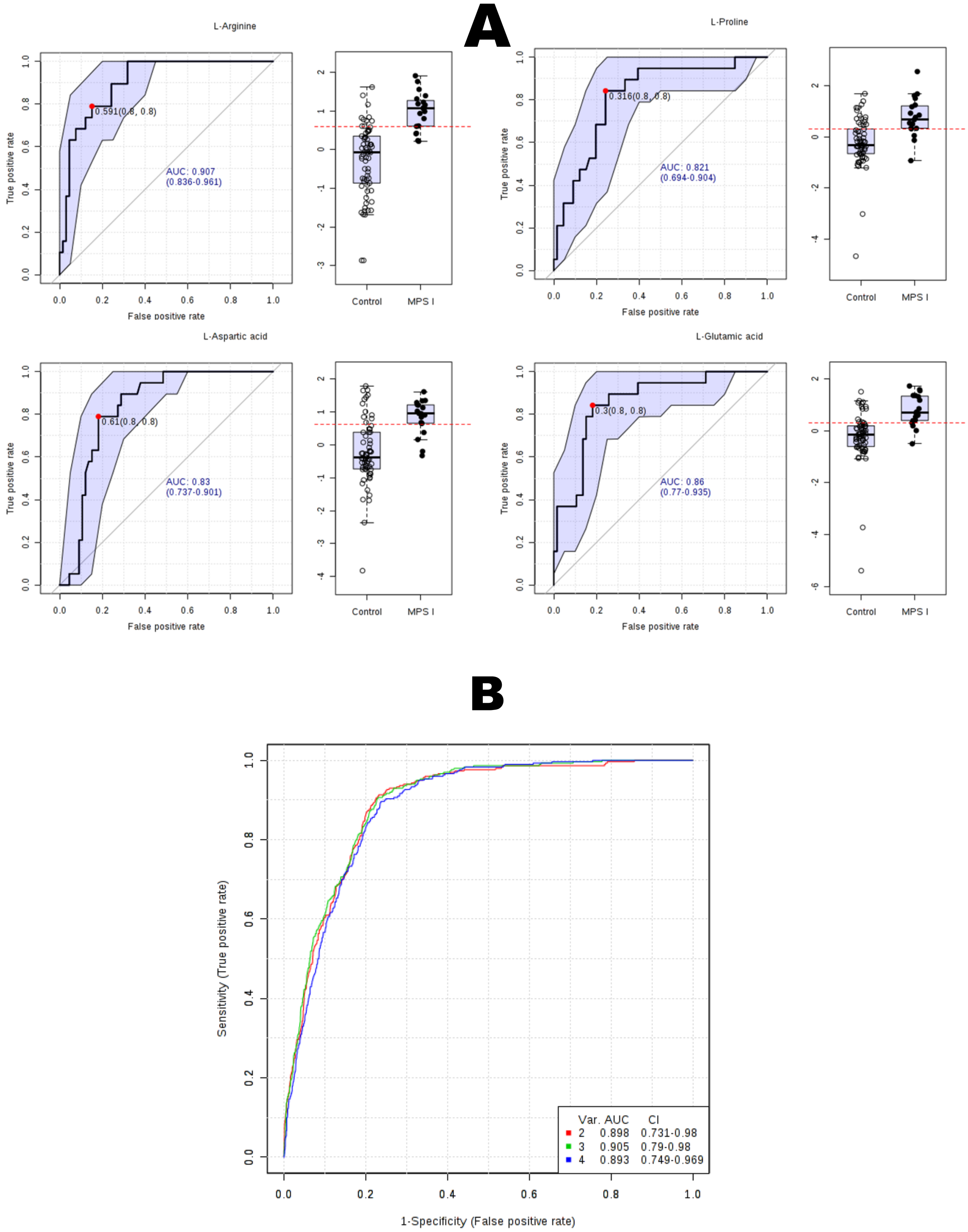


Figure S13. Pathways of arginine metabolism. Enzymes that catalyze the indicated reactions are as follows: 1: argininosuccinic lyase, 2: NO synthases. 3: arginine:glycine amidinotransferase. 4: arginase. 5, arginine decarboxylase. 6: agmatinase (agmatine ureohydrolase). 7: guanidinoacetate N-methyltransferase. 9: ornithine aminotransferase. 10: pyrroline- 5-carboxylate reductase. 11: pyrroline- 5-carboxylate dehydrogenase. 12: glutamate dehydrogenase. 13: alanine aminotransferase, aspartate aminotransferase, or branched-chain amino acid aminotransferase. 14: glutamine synthetase. 15: glutaminase. 16: ornithine decarboxylase. 17: spermidine synthase. 18: spermine synthase. 19: diamine oxidase. 20, aldehyde dehydrogenase. 21: glutamate decarboxylase. Complete oxidation of arginine-derived a-ketoglutarate occurs via the citric acid cycle. Step 8 is a spontaneous, nonenzymatic reaction. DCAM: decarboxylated S-adenosylmethionine. Glu: L-glutamate. MTA: methylthioadenosine. SAHC: S-adenosylhomocysteine. SAM: S-adenosylmethionine. aKG: a-ketoglutarate. The metabolites that were measured in the present study are shown in red boxes. Elevated metabolites in MPSI urine are highlighted in orange.

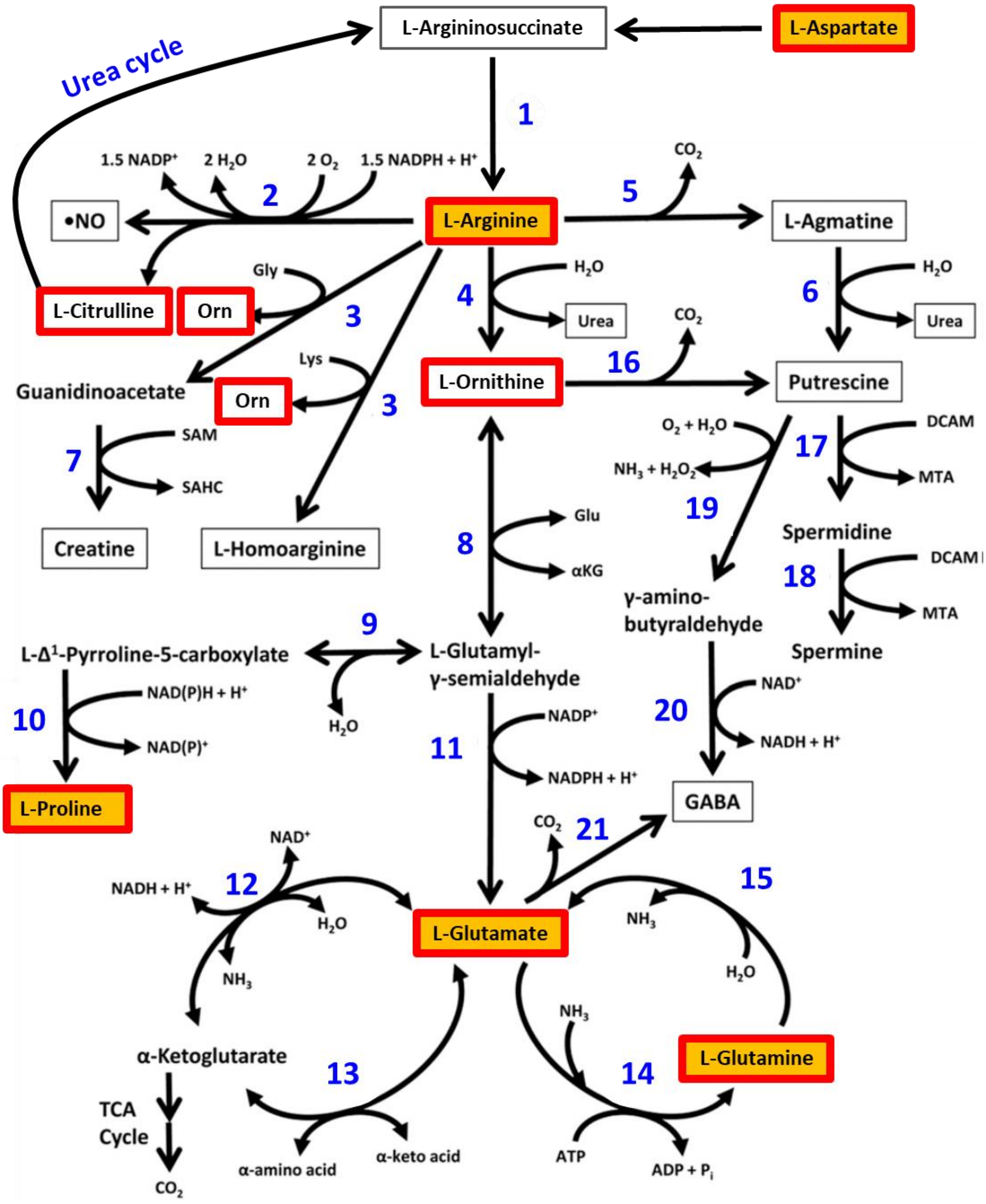
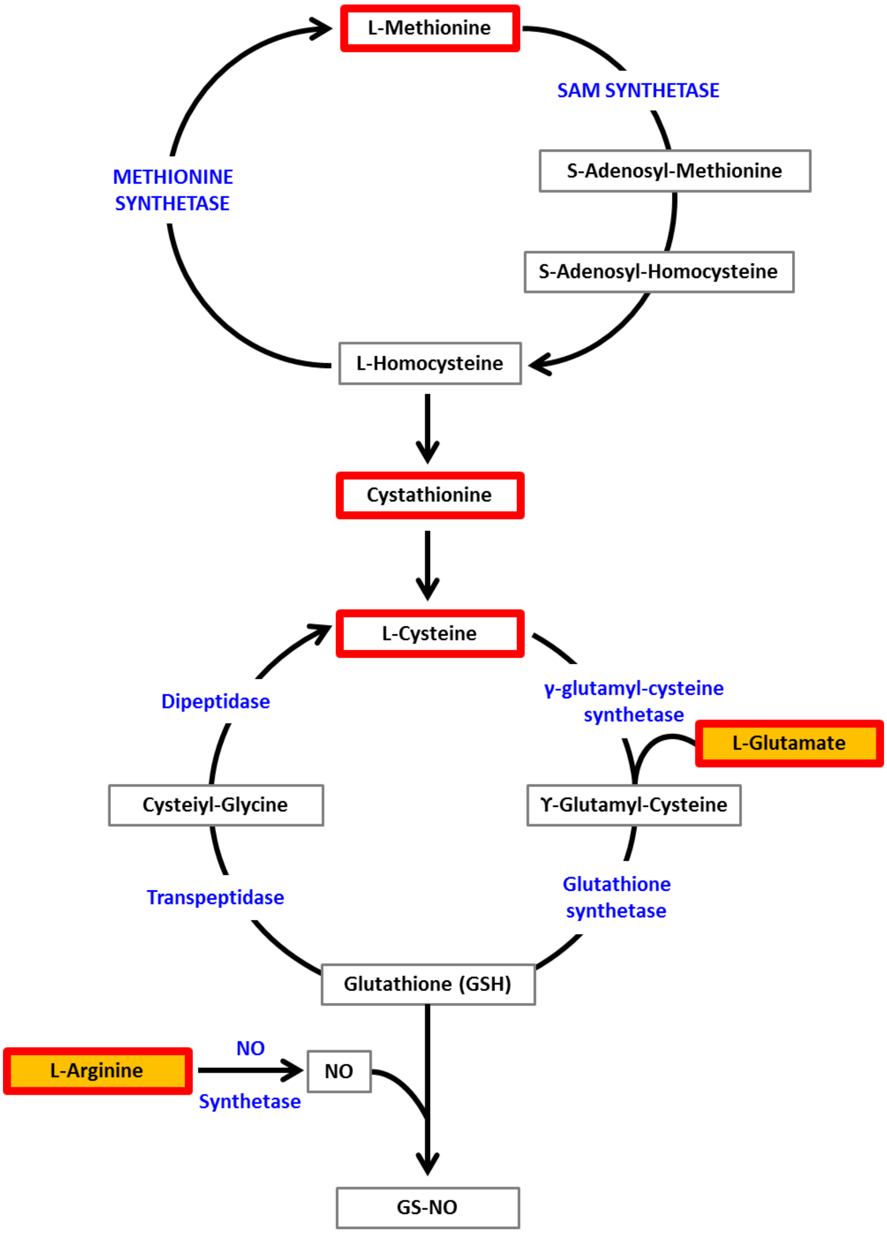


Figure S14. Schematic representation of aminothiol synthesis and glutathione metabolism. GSH is synthesized from cysteine, glutamate and glycine by the sequential action of GCL (glutamate–cysteine ligase) and GS (glutathione synthase). Intercellular transport of GSH involves its breakdown by the action of γ-glutamyl transpeptidase and dipeptidase, forming Cys-Gly and cysteine respectively. Alternatively, GSH may be oxidized to its disulphide, GSSG, and either GSH or GSSG may interact with protein cysteinyl thiols. Analogously, cysteine may be oxidized to cystine. Cysteine and Cys-Gly may also interact with proteins. Methionine is the precursor for cysteine synthesis, in which homocysteine and cystathionine are intermediates. Homocysteine is the substrate for the regeneration of methionine by methionine synthase. Methionine and homocysteine may also be oxidized to methionine sulphoxide and homocystine respectively as a result of ROS attacks. SAM (S-adenosylmethionine) and SAH (S-adenosylhomocysteine), which are intermediates in the conversion of methionine into homocysteine, also serve as methyl group donors. Arginine also interacts with GSH metabolism through nitric oxide (NO) metabolism. The metabolites that were measured in the present study are shown in red boxes. Elevated metabolites are highlighted in orange.



Conclusion et perspectives

CONCLUSIONS ET PERSPECTIVES

Le métabolisme humain est un réseau intégré, dans lequel les principales voies métaboliques les glucides, les lipides, les protéines et les macromolécules qui les constituent se combinent pour permettre d'utiliser les divers ingrédients de l'alimentation pour fournir à l'organisme l'énergie et les éléments essentiels à la vie de ses cellules. En outre, les vitamines et les minéraux fournissent les cofacteurs requis, et les hormones médient la régulation de ces voies métaboliques. Chaque organe dispose des voies spécifiques en fonction des substrats et de l'influence hormonale pour assurer son homéostasie. La compréhension des modulations métaboliques qui sou-tendent les IEM permettra à terme de proposer des stratégies thérapeutiques qui englobe l'ensemble des voies dérégulées et ne se focalisent pas sur la seule étape concernée par le déficit enzymatique ou de transport. Le travail présenté dans cette thèse représente un potentiel prometteur pour les diverses applications des méthodes de biologie des systèmes tel que la métabolomique dans le domaine des IEM.

La revue systématique de la littérature a montré un manque et un besoin d'exploration métabolomique des maladies lysosomales en générale, et des MPS en particulier. Ainsi, le travail présenté dans cette thèse vise à comprendre les mécanismes métaboliques et leur perturbation dans les mucopolysaccharidoses en utilisant des approches métabolomiques globales. Ceci a permis de générer des hypothèses métaboliques quant à la mécanistique métabolique sous-jacent les MPS. En effet, les résultats ont montré un remodellage métabolique profond dans la MPS I. Deux métabolismes majeurs ont été identifiés à savoir le métabolisme de l'Arginine et le Glutathion acteur majeur du stress oxydant. Les perspectives directes de ce travail consistent à mieux comprendre les modifications de ces métabolismes dans des modèles cellulaires et murins de MPS I.

Par la suite, les données générées au cours de cette étude pour les autres MPS (MPS II, MPS IIIA, MPS IIIB, MPS IIIC, MPS IVA et MPS VI) seront analysées pour valider des modèles prédictifs, identifier les molécules discriminantes et caractériser les voies métaboliques modulées dans chacune de ces pathologies.

Le monde des EIM est en perpétuel développement grâce aux innovations technologiques qui nourrissent la créativité thérapeutique, mais aussi conceptuelle dans la compréhension de ces maladies. Ces avancées imposent une mise à jour de nos stratégies réductionnistes et de les intégrer avec les approches systémiques qui présentent un potentiel évident aussi bien pour la recherche fondamentale que pour la recherche clinique. Ce changement de paradigme aboutirait à terme à une meilleure caractérisation de ces maladies, plus facile et fiable, objectif ultime pour entrer pleinement dans l'ère de la Médecine de Précision.

“I may not have gone where I intended to go, but I think I have ended up where I needed to be.”

— Douglas Adams

Analyse métabolomique multidimensionnelle : application aux Erreurs Innées du Métabolisme

Résumé : La médecine de précision (MP) est un nouveau paradigme qui révolutionne la pratique médicale actuelle et remodèle complètement la médecine de demain. La MP aspire à placer le patient au centre du parcours de soins en y intégrant les données médicales et biologiques individuelles tout en tenant compte de la grande diversité interindividuelle. La prédiction des états pathologiques chez les patients nécessite une compréhension dynamique et systémique. Les erreurs innées du métabolisme (EIM) sont des troubles génétiques résultant de défauts dans une voie biochimique donnée en raison de la déficience d'une enzyme, de son cofacteur ou d'un transporteur. Les EIM ne sont plus considérées comme des maladies monogéniques, mais tendent à être plus complexes et multifactorielles. Le profil métabolomique permet le dépistage d'une pathologie, la recherche de biomarqueurs et l'exploration des voies métaboliques mises en jeu. Dans ce travail de thèse, nous avons utilisé l'approche métabolomique qui est particulièrement pertinente pour les EIM compte tenu de leur physiopathologie de base qui est étroitement liée au métabolisme. Ce travail a permis la mise en place d'une méthodologie métabolomique non ciblée basée sur une stratégie analytique multidimensionnelle comportant la spectrométrie de masse à haute résolution couplée à la chromatographie liquide ultra-haute performance et la mobilité ionique. La mise en place de la méthodologie de prétraitement, d'analyse et d'exploitation des données générées avec des outils de design expérimental et d'analyses multivariées ont été aussi établies. Enfin, cette approche a été appliquée pour l'exploration des EIM avec les mucopolysaccharidoses comme preuve de concept. Les résultats obtenus suggèrent un remodelage majeur du métabolisme des acides aminés en particulier l'arginine ainsi que du stress oxydant via le métabolisme du glutathion dans la mucopolysaccharidose de type I. En résumé, la métabolomique pourrait être un outil complémentaire pertinent en appui à l'approche génomique dans l'exploration des EIM.

Mots clés : Erreurs innées du métabolisme, mucopolysaccharidoses, omiques, métabolomique, spectrométrie de masse, chromatographie liquide, mobilité ionique, chimiométrie

Multidimensional metabolomics analysis: application to Inborn Errors of Metabolism

Abstract: The new field of precision medicine is revolutionizing current medical practice and reshaping future medicine. Precision medicine intends to put the patient as the central driver of healthcare by broadening biological knowledge and acknowledging the great diversity of individuals. The prediction of physiological and pathological states in patients requires a dynamic and systemic understanding of these interactions. Inborn errors of metabolism (IEM) are genetic disorders resulting from defects in a given biochemical pathway due to the deficiency of an enzyme, its cofactor or a transporter. IEM are no longer considered to be monogenic diseases, which adds another layer of complexity to their characterization and diagnosis. To meet this need for faster screening, the metabolic profile can be a promising candidate given its ability in disease screening, biomarker discovery and metabolic pathway investigation. In this thesis, we used a metabolomic approach which is particularly relevant for IEM given their basic pathophysiology that is tightly related to metabolism. This thesis allowed the implementation of an untargeted metabolomic methodology based on a multidimensional analytical strategy including high-resolution mass spectrometry coupled with ultra-high-performance liquid chromatography and ion mobility. This work also set a methodology for preprocessing, analysis and interpretation of the generated data using experimental design and multivariate data analysis. Finally, the strategy is applied to the exploration of IEM with mucopolysaccharidoses as a proof of concept. The results suggest a major remodeling of the amino acid metabolisms, particularly, arginine as well as oxidative stress via glutathione metabolism in mucopolysaccharidosis type I. In summary, metabolomic is a relevant complementary tool to support the genomic approach in the functional investigations and diagnosis of IEM.

Keywords**:** Inborn errors of metabolism, mucopolysaccharidoses, omics, metabolomics, mass spectrometry, liquid chromatography, ion mobility, chemometrics