# Utilisation du microbiome intestinal dans la prédiction de l'état de santé de l'hôte

**Mémoire**

**Thomas Deschênes**

**Maîtrise en nutrition - avec mémoire**
Maître ès sciences (M. Sc.)

Québec, Canada

# Utilisation du microbiome intestinal dans la prédiction de l'état de santé de l'hôte

**Mémoire**

**Thomas Deschênes**

Sous la direction de :

Frédéric Raymond, directeur de recherche
Vincenzo Di Marzo, codirecteur recherche

# Résumé

Durant les dernières décennies, la recherche sur le microbiome intestinal a positionné ce dernier comme important régulateur de nombreux processus physiologiques chez l'humain. Propulsée par les technologies de séquençage à haut débit, la recherche sur l'écologie microbienne a connu un important changement de paradigme. Les méthodes d'isolation et de mise en culture de bactéries d'intérêt sont maintenant, de manière générale, remplacées par le séquençage génétique de communautés microbiennes complètes directement dans leur environnement. Ce type d'analyse, la métagénomique, a révélé l'immense catalogue de gènes bactériens présents dans l'environnement intestinal et a levé le voile sur la majorité silencieuse du microbiote : les microorganismes non-cultivables. Ce vaste catalogue de gènes microbiens représente une véritable mine d'information dans un contexte où la recherche tente de trouver des mécanismes moléculaires expliquant la relation entre le microbiome et la santé des individus. Dans ce contexte, l'apprentissage automatique, qui permet l'analyse de données complexes, peut être utilisé pour pointer vers des effecteurs microbiens d'intérêt. L'objectif du projet est d'utiliser les données métagénomiques d'individus malades et en santé dans une tâche de classification. Plus précisément, notre but est de comparer le pouvoir prédictif de différentes représentations du microbiome, toutes dérivées des données de séquençage en métagénomique non-ciblée. Notre étude a démontré que dans un contexte de classification de phénotype de l'hôte, les méthodes de représentation qui utilisent toute l'information génique séquencée permettent de meilleures performances de prédiction que celles qui utilisent exclusivement l'information contenue dans les banques de données de référence, comme les profils taxonomique et fonctionnel. Nos résultats suggèrent que l'utilisation exclusive de l'information *a priori* dans un contexte d'apprentissage automatique limite, d'une certaine façon, la possibilité de trouver de nouveaux effecteurs microbiens inconnus des banques de données.

# Abstract

During the last decades, research positioned the gut microbiome as a major regulator of numerous physiological processes in humans. Propelled by next-generation sequencing technologies, the research on microbial ecology has undergone a significant paradigm shift; generally, bacteria isolation and cultivation are now being replaced by genetic sequencing of whole bacterial communities directly from their environment. This type of analysis, referred as metagenomics, revealed the large catalog of microbial genes comprised in the gut environment and lifted the veil on the microbiota's silent majority: non-cultivable microorganisms. This vast catalog of genes represents a real mine of information in a context where research aims at finding molecular mechanisms to explain the relation between microbiome and host health. In this context, machine learning, which allows the analysis of complex data, can be used to point toward promising microbial features. The objective of this project is to use metagenomics data from healthy and diseased individuals in a classification task. More precisely, our goal is to compare the predictive power of different microbiome representations, all derived from untargeted metagenomics data. Our study has shown that in a context of host phenotype classification, representation methods that use all the available sequenced information allow better prediction performances than those that are based on reference databases, like the taxonomic and functional profiles. Our results suggest that the exclusive use of *a priori* information, in a machine learning context, limits, in a way, the possibility of finding new microbial effectors unknown from reference databases.

# Table des matières

*Résumé* ..... *ii*

*Abstract* ..... *iii*

*Table des matières* ..... *iv*

*Liste des figures* ..... *vi*

*Liste des abréviations, sigles, acronymes* ..... *vii*

*Avant-propos* ..... *viii*

*Introduction* ..... *1*

**1. Le microbiome et la santé humaine** ..... **1**
1.1. Rôle physiologique du microbiome ..... 1
1.2. Maladie et microbiome ..... 2
1.2.1. Cancer colorectal ..... 3
1.2.2. Cirrhose hépatique ..... 4
1.2.3. Diabète et obésité ..... 5
1.3. Mécanismes moléculaires entre le microbiome et le métabolisme ..... 6
1.3.1. Endotoxémie métabolique ..... 6
1.3.2. Endocannabinoïdome ..... 7
1.3.2.1 Définition de l'endocannabinoïdome ..... 7
1.3.2.2. Interaction entre l'endocannabinoïdome et le microbiome intestinal ..... 9

**2. Séquençage du microbiome** ..... **9**
2.1. Séquençage de nouvelle génération ..... 10
2.2. Métagénomique ..... 11
2.2.1. Séquençage du gène de la sous-unité ribosomale 16S (ARNr 16S) ..... 11
2.2.2. Le séquençage métagénomique de type « shotgun » ..... 13
2.2.3. Données dérivées du séquençage métagénomique « shotgun » ..... 14
2.2.3.1. Alignement des séquences brutes ..... 14
2.2.3.2. Les différents niveaux d'inconnu du microbiome ..... 15
2.2.3.3. Assemblage ..... 16

**3. Apprentissage automatique** ..... **18**
3.1. Apprentissage supervisé ..... 19
3.1.2. Généralisation, surapprentissage et sous-apprentissage ..... 20
3.1.3. Validation croisée ..... 21
3.2. Algorithmes communs ..... 22
3.2.1. Modèles linéaires ..... 22
3.2.1.1. Séparateurs à vastes marges ..... 22
3.2.1.2. Régression logistique ..... 22
3.2.2. Set covering machine ..... 23
3.2.3. Arbres de décisions ..... 23
3.2.4. Forêt aléatoire ..... 25
3.3. Apprentissage automatique en bio-informatique ..... 25

*Mise en contexte, hypothèse et objectifs* ..... *27*

*Approche méthodologique* ..... *28*

*Chapitre 1 : Reference-free microbiome representation enhances host phenotype classification* ..... *29*

**Résumé** ..... **30**

**Abstract** ..... **32**

**Introduction** ..... **33**

**Results** ..... **34**

The COGs outperform the taxonomic and functional profiles ..... 34

Subsets of COGs with specific functions ..... 36

Potential implication of microbial gut-brain axis effectors ..... 37

The impact of algorithms and data types on prediction performance ..... 38

**Discussion and conclusion** ..... **39**

**Methods** ..... **41**

Production of the different data types ..... 41

Machine learning protocol ..... 41

**Figures and tables** ..... **43**

**References** ..... **44**

*Conclusion* ..... *48*

*Bibliographie* ..... *51*

# Liste des figures

## Introduction

**Figure 1**. Format des données d'entrée et représentation des concepts d'exemple, de variable et d'étiquette.....19
**Figure 2**. Schéma des étapes d'une validation croisée à 5 plis .....21

## Chapitre 1

**Figure 1**. Heatmaps of the results for every combination of algorithm and data type .....43

# Liste des abréviations, sigles, acronymes

2-AG : 2-arachidonoyl-glycérol
2-OG : 2-oléoyl-glycérol
2-PG : 2-palmitoyl-glycérol
ACP : analyse en composante principale
AEA : N-arachidonoyl-éthanolamine
ARNr 16S: gène de la sous-unité ribosomale 16S
ASV : *amplicon sequence variant*
$CB_1$ : récepteur cannabinoïde 1
$CB_2$ : récepteur cannabinoïde 2
CD14 : *Cluster of Differentiation 14*
COG : *Clusters of Orthologous Groups of proteins*
CRC : cancer colorectal
DGL$\alpha$ : *sn-1-specific diacylglycerol lipase-alpha*
DGL$\beta$ : *sn-1-specific diacylglycerol lipase-beta*
DT : arbre de décision
FAAH : fatty acid amide hydrolase 1
FadA : *Fusobacterium adhesin A*
GPCR : *G protein-coupled receptors*
GPR119 : *G protein-coupled receptor 119*
GPR55 : *G protein-coupled receptor 55*
HMM : *Hidden Markov model*
LC : cirrhose hépatique
LEA : N-linoléoyl-éthanolamine
LPS : lipopolysaccharides
LR : regression logistique
MAGL : *monoacylglycerol lipase*
NAE : N-acyl-éthanolamines
NAPE-PLD : N-acyl-phosphatidylethanolamine-hydrolyzing phospholipase D
NGS : séquençage de nouvelle génération
NOD : *nucleotide oligomerization domain receptors*
OEA : N-oleoyl-éthanolamine
OTU : *operational taxonomic unit*
PCR : *polymerase chain reaction*
PEA : N-palmitoyl-éthanolamine
PGN : peptidoglycane
PPAR$\alpha$ : *peroxisome proliferator-activated receptor alpha*
RF : forêt aléatoire
SCM : *set covering machine*
SEA : *N*-stéaroyl-éthanolamine
SVM *:* séparateurs à vastes marges
T2D : diabète de type II
$\Delta^9$-THC : delta-9-tétrahydrocannabinol
TLR4 : *Toll Like Receptor 4*
TRPV1 : *transient receptor potential vanilloid 1*

# Avant-propos

Ce mémoire présente les travaux que j'ai réalisé durant mes deux années de maîtrise. Mon projet portait sur l'analyse génomique du microbiome intestinal en lien avec l'état de santé des individus. Plus particulièrement, nous nous sommes intéressés à l'utilisation de l'apprentissage automatique pour l'analyse de données dérivées de la métagénomique non-ciblée. Les résultats de ces analyses seront publiés prochainement dans un article intitulé « Reference-free microbiome representation enhances host phenotype classsification », dont je suis le premier auteur. Pour cet article, les co-auteurs sont Fred Wilfried Elom Tohoundjona, Pier-Luc Plante, Vincenzo Di Marzo et Frédéric Raymond. Fred Wilfried Elom Tohoundjona a participé à la préparation et au traitement des données. Pier-Luc Plante a participé à la réalisation du protocole expérimental et à l'analyse et l'interprétation des résultats. Vincenzo Di Marzo a supervisé l'analyse et l'interprétation des résultats ainsi que la réalisation du protocole expérimental. Frédéric Raymond a participé à la réalisation du protocole expérimental, a supervisé le déroulement de chaque étape et a participé à la rédaction de l'article. Pour ma part, j'ai participé aux étapes de préparation et de traitement des données, à la réalisation du protocole expérimental, à l'analyse et à l'interprétation des résultats ainsi qu'à la rédaction de l'article. La réalisation de cet article n'aurait pas été possible sans la contribution des co-auteurs, que je tiens à remercier. L'article est actuellement en révision par les co-auteurs et la version inclue dans ce mémoire sera donc différente de la version finale.

La réalisation de mon projet de maîtrise a été possible grâce à la Chaire d'excellence en recherche du Canada sur l'axe microbiome-endocannabinoïdome dans la santé métabolique (CERC-MEND), dirigée par son titulaire, Vincenzo Di Marzo. Je suis reconnaissant d'avoir fait partie de ce groupe de recherche d'envergure. Faire partie d'une équipe de gens curieux, déterminés et rigoureux a été une expérience motivante.

J'ai eu la chance de réaliser mon premier stage au baccalauréat en sciences biomédicales dans le laboratoire de Frédéric Raymond. Ce stage m'a permis de confirmer mon intérêt pour la recherche en bio-informatique et pour l'informatique en général; j'ai donc continué mon parcours à la maîtrise dans ce laboratoire. Je tiens d'ailleurs à remercier sincèrement Frédéric Raymond de m'avoir fait confiance et de m'avoir permis d'intégrer la branche bio-informatique de son équipe de recherche malgré mon expérience de départ limitée, voire inexistante dans ce domaine. Je suis aussi reconnaissant de la liberté et de l'autonomie qu'il m'a accordées tout au long de mon parcours, ce qui m'a permis de

m'améliorer et de cheminer comme étudiant-chercheur. Son écoute et ses conseils ont aussi pu m'aiguiller dans la bonne direction lors des moments de doute.

Merci à Pier-Luc pour le partage de son expertise en bio-informatique et en apprentissage automatique, pour sa positivité contagieuse à l'égard du projet et pour ses conseils lors des satanés « bugs » ou autres problèmes informatiques.

Je tiens également à remercier certains membres de l'équipe. Premièrement, merci à Fredy Alexander de m'avoir permis d'aiguiser mes compétences de « débogage » dans R. Mais merci surtout pour sa bonne compagnie et son humour. Je tiens également à remercier Fred Wilfried pour son aide avec la production des données et des scripts, ce qui m'a sauvé du temps précieux. Merci également à Edi pour nos conversations captivantes dans le local à l'INAF et pour son aide lors de mes débuts avec Python.

Merci à mes amis, qui sont toujours là pour me faire décrocher, mais aussi pour m'encourager.

Merci à ma copine Anne-Marie pour son support constant tout au long de ma maîtrise, particulièrement dans les moments de doute, et de s'être intéressée à mon projet un tantinet compliqué.

Finalement, merci à ma famille et, surtout, à mes parents, qui m'encouragent depuis toujours et sans qui je ne serais pas là où j'en suis aujourd'hui.

# Introduction

Ce chapitre est divisé en sections qui traitent, respectivement, du concept de microbiome intestinal, du séquençage de ce dernier et de l'apprentissage automatique. Le microbiome intestinal est présenté en mettant l'emphase sur son lien avec la santé humaine et diverses maladies. Les différentes maladies explicitées dans ce chapitre sont des maladies à haute prévalence dont le lien avec le microbiome a été démontré dans la littérature. Ces maladies sont aussi celles qui sont étudiées dans l'article présenté au chapitre suivant. De plus, en raison du rôle important de l'endocannabinoïdome – le système endocannabinoïde élargi – dans la santé métabolique, les interactions moléculaires entre ce système et le microbiome seront aussi traitées dans cette section. La seconde section présente les différents types de données dérivées du séquençage du microbiome et leurs caractéristiques. Finalement, la troisième section présente l'apprentissage supervisé, un type d'analyse utile pour modéliser le système complexe qu'est le microbiome et qui tient compte du nombre élevé de variables produites par le séquençage.

## 1. Le microbiome et la santé humaine

Le microbiome est un concept dont la définition a été souvent remaniée. Les définitions actuelles s'entendent généralement pour décrire le microbiome comme un véritable écosystème. Aujourd'hui, la définition la plus citée est celle de Lederberg et coll. et décrit le microbiome comme une « communauté écologique de microorganismes commensaux, symbiotiques et pathogéniques compris dans une partie du corps ou dans un autre environnement » (1,2). Dans un contexte bio-informatique, le microbiome est plutôt défini comme étant l'ensemble des gènes présents dans le microbiote. Il est aussi généralement accepté que le terme microbiote renvoie aux microorganismes contenus dans un microbiome.

### 1.1. Rôle physiologique du microbiome

Chez l'humain, la majeure partie du microbiote présent dans l'organisme se retrouve dans la partie distale du tractus gastro-intestinal (3). Il est composé de bactéries, d'archées, de virus et de microorganismes eucaryotes (4). La majorité du microbiote intestinal est composée de microorganismes commensaux ou mutualistes (3). Ce réseau dense de microorganismes est le résultat de la coévolution entre les communautés microbiennes et leur hôte. Parmi les fonctions acquises au cours de cette coévolution se trouvent : la maturation et la constante éducation du système

immunitaire, la régulation de certaines fonctions endocrines au sein de l'intestin, la biogénèse d'énergie, de vitamines, de neurotransmetteurs, de molécules bioactives ayant des rôles encore indéterminés, le métabolisme des sels biliaires, l'interaction avec certains médicaments et l'élimination de toxines (3). Certaines de ces fonctions ont été amplement décrites dans la littérature. Par exemple, il est maintenant connu que le microbiote peut dégrader et métaboliser les polysaccharides alimentaires en acides gras à courte chaîne, comme le butyrate, qui est une source d'énergie importante pour les cellules du colon, entre autres (5). Cela dit, le degré d'importance du microbiome dans certaines de ces fonctions reste à être démontré et la recherche portant sur la relation microbiome-hôte demeure un domaine en pleine effervescence. L'éventail de fonctions mentionnées précédemment est très large et les domaines d'études qui étudient maintenant le microbiome sont tout aussi variés.

### 1.2. Maladie et microbiome

Au début de l'ère de la recherche sur le lien entre le microbiome et diverses maladies, le rôle causal du microbiome a souvent été remis en question. Une des hypothèses était que les changements au sein du microbiome étaient un phénomène subséquent au développement des maladies. Or, une étude mécanistique de Bakhed et coll, en 2004, a mis en lumière un rôle causal du microbiome en montrant que, chez la souris, le microbiome intestinal régulait la capacité de l'hôte à extraire l'énergie contenue dans la diète en plus d'être impliqué dans le stockage de l'énergie (6). Depuis, avec l'évolution des techniques d'analyses, la plus grande capacité de stockage de données et le développement concomitant d'outils d'analyse bio-informatiques, beaucoup d'études ont contribué à la caractérisation du lien hôte-microbiome. Par exemple, avec l'avènement du séquençage de nouvelle génération (NGS), des études d'association sur de grandes cohortes ont montré que le profil microbien des personnes atteintes de maladies était différent de celui des individus en santé (7–10).

Les technologies de séquençage ont levé le voile sur la grande diversité microbienne et sur le large potentiel fonctionnel présent au sein du microbiome. À l'aide de cette véritable mine d'informations, les études tendent maintenant à trouver les mécanismes moléculaires expliquant le rôle physiologique du microbiome dans l'organisme, mais aussi le rôle qu'il joue dans le développement de certaines maladies. Les prochaines sections portent sur l'état des connaissances de la relation hôte-microbiome dans le contexte de maladies et sur des mécanismes de communication entre le métabolisme et le microbiome.

### 1.2.1. Cancer colorectal

Le cancer colorectal (CRC) est le troisième type de cancer le plus commun à l'échelle mondiale et est responsable du plus grand nombre de décès reliés au cancer (11). L'étiologie de la maladie est toutefois encore inconnue et 90% des cas de CRC sont sporadiques. Comme plusieurs cancers, certains facteurs génétiques et environnementaux peuvent être des facteurs de risque : sédentarité, diète occidentale (haute en aliments transformés, riche en graisses animales et en sucres, faible en fibres, etc.), tabagisme, consommation d'alcool et obésité. Ces facteurs environnementaux sont aussi connus pour avoir un impact sur la composition du microbiome intestinal.

De études ont rapporté que les individus atteints de CRC présentaient une diversité microbienne réduite dans le microbiome dérivés d'échantillons fécaux ainsi qu'au niveau des muqueuses intestinales, comparativement aux individus en santé (12,13). L'accumulation d'études mesurant la composition taxonomique des individus atteints de CRC suggère que la perte de certaines bactéries productrices de butyrate au profit de pathogènes opportunistes pro-inflammatoires puisse exacerber la dysbiose et ultimement contribuer au développement de tumeurs (14,15). De plus, il a été démontré que la dysbiose associée aux stades plus précoces de la maladie était différente de celle associée aux stades plus avancés, ce qui suggère que le microbiome puisse avoir un rôle dans la progression du CRC (16).

Il est aussi connu que le microbiome intestinal produit des molécules bioactives jouant un possible rôle dans le développement du CRC. Par exemple, les polyamines, produites par la fermentation de protéines, sont des molécules qui stimulent la croissance normale des cellules, mais qui peuvent avoir des effets cancérigènes lorsque leur métabolisme est dérégulé (17). D'ailleurs, une méta-analyse des jeux de données de métagénomique non-ciblée a rapporté que le microbiome des individus atteints de CRC était associé à des altérations dans le métabolisme des polyamines, ce qui suggère un mécanisme microbien dans le développement du CRC (18). Une étude de 2013 a aussi démontré un mécanisme par lequel *Fusobacterium nucleatum* pouvait adhérer sur les cellules hôtes et induire une réponse pro-inflammatoire et stimuler la croissance d'adénomes via sa protéine de surface FadA (19).

Aussi, un possible rôle du virome - le contenu génique provenant des phages intestinaux - dans le CRC a été mis en lumière dans une étude (20). En effet, en utilisant les technologies de séquençage à haut débit, il a été démontré que la diversité virale était augmentée chez les individus atteints de

CRC. De plus, le profil viral des microbiomes a pu être utilisé pour prédire le statut de CRC en ségrégant les individus atteints selon les stages précoces ou avancés de la maladie.

### 1.2.2. Cirrhose hépatique

La cirrhose hépatique est un stade avancé de fibrose – ou cicatrisation - du foie qui peut être causée par plusieurs autres maladies ou conditions. Parmi ces causes, les plus fréquentes sont la stéatohépatite non alcoolique qui se veut être une comorbidité fréquente associée à l'obésité et au diabète et qui est aussi la maladie du foie la plus fréquente dans le monde, l'infection au virus de l'hépatite et la consommation excessive d'alcool (21). Aux stades les plus avancés de la maladie, ses complications ont un mauvais pronostic et la transplantation est souvent requise (22). Le foie est en étroite interaction avec le tractus gastro-intestinal via le système porte hépatique et la sécrétion biliaire. La progression de la cirrhose du foie est associée à la translocation de bactéries et de leurs composantes à travers la barrière épithéliale de l'intestin (21). Par contre, les changements au sein du microbiome qui sont associés au développement de la maladie sont encore peu connus. De manière générale, la dysbiose associée à la cirrhose hépatique, comme dans plusieurs autres contextes pathologiques, est associée à une augmentation de bactéries pathogènes et à une proportion réduite de bactéries considérées bénéfiques (22). Notamment, l'augmentation de la famille *Prevotellaceae* au niveau du microbiome intestinal est associée à la cirrhose alcoolique, lorsque comparée avec la cirrhose associée au virus de l'hépatite B ou des individus en santé (23). Certaines études ont aussi montré que les altérations dans la composition du microbiome avaient un impact important surtout dans les complications de la maladie, lors des stades les plus avancés (24). La dysbiose a aussi été associé au maintien de la détérioration du foie durant les stades moins avancés de la maladie, par exemple dans la stéatohépatite non alcoolique (25). Une étude récente a utilisé l'apprentissage automatique pour discriminer entre des individus sains et des individus atteints de cirrhose associée à une stéatohépatite non alcoolique en se basant le profil microbien de leur microbiote. Ils ont pu discerner une signature microbienne composée de 19 espèces bactériennes (26). De plus, leur modèle de prédiction s'est avéré être relativement efficace dans la détection de cirrhoses d'origine étiologiquement différente dans des cohortes indépendantes. Ce fait suggère que certains changements au sein du microbiome sont communs aux différents types de cirrhose et cela est prometteur pour comprendre la relation entre le microbiome et le développement de cette maladie.

### 1.2.3. Diabète et obésité

L'obésité et ses nombreuses comorbidités, comme le diabète de type 2 (T2D), exercent une pression énorme sur les systèmes de santé à l'échelle mondiale (27). Bien que l'accumulation du tissu adipeux viscéral soit globalement reconnue comme un facteur de risque du développement de nombreuses maladies, les mécanismes inhérents à leur développement ne sont pas encore complètement élucidés (28). Il est toutefois accepté que l'inflammation chronique serait un déclencheur des dérèglements métaboliques associés à l'obésité et au T2D (27). Cette inflammation est principalement le résultat d'un surplus nutritionnel et métabolique impliquant des sentiers métaboliques similaires à ceux retrouvés dans l'inflammation dite « classique ». Une hypothèse intéressante propose que cette similarité proviendrait d'un héritage évolutif; les organes impliqués dans le métabolisme et l'immunité auraient évolué à partir de structures ancestrales communes avant d'acquérir des fonctions distinctes (27,29). De plus, la composante immunitaire serait importante dans le développement de l'inflammation chronique, chez les animaux supérieurs, en raison de la présence des cellules immunitaires dans la fraction stroma-vasculaire du tissu adipeux et du foie, notamment. Aussi, des études chez l'animal et chez l'homme ont montré que plusieurs cytokines pro-inflammatoires et d'autres médiateurs du système immunitaire seraient grandement impliqués dans les dérèglements métaboliques (27,30). Avec la transition rapide – à l'échelle évolutive – du mode de vie des humains vers la sédentarité et l'abondance nutritionnelle, il est possible de penser que la composition nouvelle du microbiome, qui résulte de ces nouvelles habitudes de vie, vient déséquilibrer la relation mutualiste du microbiome avec son hôte, acquise par évolution. Ce déséquilibre pourrait résulter en un gain de fonctions, au possible détriment de fonctions bénéfiques pour l'hôte.

Grâce à l'essor des technologies de séquençage, des études d'association ont mis en lumière que la composition du microbiote intestinal des individus obèses et/ou diabétiques était différente de celle retrouvée chez des individus en santé (7–10). Parmi celles-ci, une étude de 2013 a révélé que la richesse en gènes du microbiome intestinal était généralement plus faible chez des individus obèses lorsque comparés avec des individus minces. Aussi, ce nombre réduit de gènes corrélait avec un risque élevé d'obésité, de résistance à l'insuline et d'inflammation chronique (9). En 2012, la première étude d'association métagénomique a été réalisée pour étudier la corrélation entre le microbiome intestinal et le T2D. Cette étude a permis de démontrer une différence significative dans la composition des profils taxonomique et fonctionnel dans le groupe des individus diabétiques lorsque comparés aux individus en santé. Les différences taxonomiques incluaient *Akkermansia municiphila*, *Clostridium*

*hathewayi* et *Escherichia coli.* Les profils fonctionnels incluaient des différences au niveau du transport des sucres et des acides aminés à chaînes ramifiés et du métabolisme du méthane, entre autres (10).

## 1.3. Mécanismes moléculaires entre le microbiome et le métabolisme

Depuis l'étude phare de Bakhed et coll, en 2004, le lien entre le microbiome et les maladies métaboliques a connu un important gain en popularité et ce sujet est encore au cœur de nombreuses recherches aujourd'hui. Depuis, certains mécanismes moléculaires ont été mis de l'avant pour mieux comprendre la relation hôte-microbiome. Les prochaines sections portent sur l'endotoxémie métabolique et l'endocannabinoïdome.

### 1.3.1. Endotoxémie métabolique

Une hypothèse relativement récente suggère que le microbiote intestinal soit une source importante d'antigènes microbiens et que sa composition puisse affecter le système immunitaire par l'initiation de mécanismes complexes via le système immunitaire inné. Cette hypothèse provient d'une étude de Brugman et coll., où ils ont observé que des souris génétiquement prédisposées à développer le diabète de type 1 avaient été partiellement protégées contre le développement de la maladie après avoir subi un traitement antibiotique (31). Ce traitement aurait diminué la charge d'antigènes au sein du microbiote. En réponse à cette hypothèse, Cani et coll. ont cherché à mettre en lumière un facteur provenant du microbiote qui pourrait enclencher et maintenir une réponse inflammatoire chronique chez des souris, lorsque nourries avec une diète riche en gras (32). Ils ont étudié l'effet de cette diète sur les niveaux plasmatiques de lipopolysaccharides (LPS) bactériens. Les LPS sont produits par la grande majorité des bactéries Gram négatives du microbiote et peuvent migrer vers les capillaires de l'intestin via un mécanisme TLR4-dépendant (33). De plus, les LPS peuvent atteindre certains tissus internes via des mécanismes de transports des lipides alimentaires et induire une cascade pro-inflammatoire (34). En résumé, l'étude de Cani et coll. a démontré que les niveaux de LPS plasmatiques étaient augmentés par la diète riche en gras, que les LPS plasmatiques, à une concentration induite par la diète, enclenchent les désordres métaboliques associés à l'obésité et au T2D et que CD14, le récepteur principal des LPS, agit comme important régulateur de ces désordres métaboliques. Les auteurs ont défini le concept d'endotoxémie métabolique comme étant l'élévation, induite par la diète, des niveaux de LPS plasmatiques. Le lien entre l'inflammation chronique et le LPS a par la suite été corroboré par d'autres études (35,36) et, plus particulièrement, chez les humains (37). Une deuxième étude du même groupe a par la suite démontré que l'endotoxémie métabolique

était régulée par le microbiote intestinal et que ce dernier influençait aussi la perméabilité intestinale, ce qui peut potentiellement expliquer l'élévation des composantes bactériennes dans le milieu interne de l'hôte (36). Depuis la découverte de l'endotoxémie métabolique, d'autres études se sont aussi penchées sur des composantes microbiennes qui pouvaient jouer un rôle dans l'inflammation chronique. C'est le cas du peptidoglycane (PGN), une composante de la paroi bactérienne. Des études ont montré que le système de détection du système immunitaire inné PGN-NOD contribue à la protection contre l'inflammation systémique et que ce système régule les interactions microbiome-hôte en contrôlant le niveau de translocation de composantes bactériennes (38–41). En plus de l'activation du système immunitaire inné, d'autres mécanismes ont été suggérés ces dernières années dans la relation hôte-microbiome. Parmi ces mécanismes se trouvent ceux impliquant l'endocannabinoïdome.

### 1.3.2. Endocannabinoïdome

L'endocannabinoïdome est un large système moléculaire impliqué dans de nombreux processus physiologiques. Au niveau du système digestif, il est impliqué dans la régulation de la vidange de l'estomac, dans la motilité gastro-intestinale et dans l'inflammation. Aussi, les endocannabinoïdes sont produits en quantité importante dans les organes impliqués dans la régulation de l'homéostasie de l'énergie comme le cerveau, le foie, le tissu adipeux, les muscles et le pancréas (42). Au niveau du tissu adipeux, l'endocannabinoïdome occupe des rôles importants dans la régulation de l'adipogénèse et de l'inflammation. Ces exemples de fonctions font de l'endocannabinoïdome un sujet de recherche prometteur pour la découverte de mécanismes expliquant les désordres métaboliques.

#### 1.3.2.1 Définition de l'endocannabinoïdome

Avant la découverte des cannabinoïdes endogènes – les endocannabinoïdes -, la recherche sur les cannabinoïdes se concentrait surtout sur la composante psychoactive de la plante *Cannabis Sativa* : le delta-9-tétrahydrocannabinol ($\Delta^9$-THC). Les deux principaux récepteurs des cannabinoïdes, les récepteurs cannabinoïdes de types 1 et 2 ($CB_1$ et $CB_2$), sont des récepteurs transmembranaires couplés aux protéines G (GPCR). L'activation des récepteur $CB_1$, présents principalement dans le cerveau, est responsable des effets psychotropes, euphoriques et de stimulation de l'appétit tandis que l'activation des récepteurs $CB_2$, surtout présents dans les cellules immunitaires, est plutôt responsable de la modulation du système immunitaire (43). Les principaux ligands endogènes de ces deux types de récepteurs ont été découverts dans les années 1990. Ces deux molécules sont dérivées de l'acide arachidonique, un acide gras polyinsaturé oméga-6. Il s'agit du *N*-arachidonoyl-éthanolamine (AEA) et du 2-arachidonoyl-glycérol (2-AG). Ces deux endocannabinoïdes ont une haute affinité pour chacun

des récepteurs $CB_1$ et $CB_2$ et peuvent les activer de manière efficace. Le système endocannabinoïde est défini comme étant l'ensemble des récepteurs $CB_1$ et $CB_2$, leurs principaux ligands – AEA et 2-AG -, ainsi que toutes les molécules impliquées dans les sentiers cataboliques et anaboliques de ces molécules. La production de l'AEA est assurée par l'enzyme NAPE-PLD (*N-acyl-phosphatidylethanolamine-hydrolyzing phospholipase D*) et la production du 2-AG est assurée par les enzymes DGL$\alpha$ (*sn-1-specific diacylglycerol lipase-alpha*) et DGL$\beta$. Les endocannabinoïdes sont produits « sur demande » à partir de précurseurs phospholipidiques provenant de la membrane cellulaire. Ils sont étroitement régulés et sont rapidement hydrolysés par la FAAH (*fatty acid amide hydrolase*) et la MAGL (*monoacylglycerol lipase*), respectivement, en des composés qui sont inactifs aux récepteurs cannabinoïdes (44).

En plus de leur affinité pour les récepteurs $CB_1$ et $CB_2$, les endocannabinoïdes peuvent aussi interagir avec les récepteurs PPAR$\alpha$, PPARγ et GPR55 (45). De plus, l'AEA peut interagir avec le récepteur TRPV1. En plus des deux principaux endocannabinoïdes, il existe plusieurs autres molécules structurellement semblables à ces derniers et qui peuvent interférer avec la réponse du système endocannabinoïde, sans toutefois activer directement les récepteurs $CB_1$ et $CB_2$. Parmi ces molécules analogues, on retrouve notamment d'autres *N*-acétyl-éthanolamines (NAE), comme le palmitoyl-éthanolamine (PEA) le *N*-oleoyl-éthanolamine (OEA), *N*-stéaroyl-éthanolamine (SEA) et le *N*-linoléyl-éthanolamine (LEA). De plus, on retrouve d'autres acylglycérols comme le palmitoylglycérol (2-PG) et le oléoylglycérol (2-OG) (42). Ces lipides bioactifs peuvent emprunter certaines portions des sentiers métaboliques des endocannabinoïdes « classiques » et pourraient augmenter l'activité de ces derniers en inhibant leur inactivation. Par exemple, le PEA et l'OEA peuvent activer les récepteurs PPAR$\alpha$ et TRPV1, tandis que le OEA, le LEA et le 2-OG peuvent activer le GPR119 (46). En raison de ce large éventail d'effecteurs du système endocannabinoïde, il est maintenant plus approprié de désigner l'ensemble des molécules impliquées dans la réponse endocannabinoïde par le terme « endocannabinoïdome ». Comme le microbiome, l'endocannabinoïdome est impliqué dans une multitude de processus physiologiques dans l'organisme et les travaux des dernières années ont démontré l'existence d'une communication entre ces deux systèmes, formant l'axe microbiome-endocannabinoïdome.

#### 1.3.2.2. Interaction entre l'endocannabinoïdome et le microbiome intestinal

Une des premières preuves scientifiques de l'existence du lien entre un membre précis du microbiome intestinal et l'endocannabinoïdome a été publiée en 2007. L'administration d'un probiotique - une souche de *Lactobacillus acidophilus* - avait modulé l'expression des récepteurs des cannabinoïdes et des récepteurs aux opioïdes dans les cellules intestinales, diminuant la douleur abdominale chez des rats (42,47). De plus, comme mentionné précédemment, l'endotoxémie métabolique est maintenant reconnue comme un facteur contributoire aux dérèglements métaboliques associés à l'obésité et au T2D. La translocation des LPS à partir de la lumière intestinale vers le milieu interne de l'organisme est régie, entre autres, par la fonction de barrière intestinale. Une étude de 2010 a démontré que l'endocannabinoïdome jouait un rôle dans les fonctions de barrière intestinale et d'adipogénèse en utilisant des molécules agonistes et antagonistes des récepteurs $CB_1$ dans des souris minces et obèses. De plus, cette même étude a suggéré que les LPS interféreraient avec des processus d'adipogénèse régis par l'endocannabinoïdome. En résumé, les auteurs ont affirmé qu'étant donné que l'obésité est communément associée à une réponse endocannabinoïde élevée, à un taux élevé de LPS plasmatiques, à une composition altérée du microbiote et à un métabolisme altéré du tissu adipeux, un important axe de communication existe fort probablement entre les systèmes de l'endocannabinoïdome, du microbiome intestinal et du métabolisme énergétique (48). En 2017, Cohen et coll. ont montré que les bactéries du microbiome encodaient des enzymes capables de former des molécules analogues aux endocannabinoïdes, des N-acylamides, et que ces dernières avaient la capacité de se lier aux récepteurs GPCR (*G protein coupled receptors*) de l'hôte. Les GPCR forment la plus grande famille de récepteurs membranaires chez les eucaryotes. Ceux avec lesquels les molécules dérivées du microbiome interagissent, notamment le récepteur GPR119 faisant partie de l'endocannabinoïdome, sont impliqués dans une multitude de conditions pathologiques incluant les maladies métaboliques (49). Cette étude a montré le potentiel des analyses bio-informatiques dans la recherche de molécules bioactives dérivées du microbiome et a tracé une voie prometteuse pour la caractérisation du lien hôte-microbiome (50).

## 2. Séquençage du microbiome

Durant les dernières décennies, la recherche portant sur l'écologie microbienne a connu un important changement de paradigme, tant au niveau de notre compréhension des communautés microbiennes que de la façon dont ces dernières sont étudiées. Les avancées en biologie moléculaire ont permis de

lever le voile sur l'immense catalogue de gènes contenus dans les échantillons dérivés de l'environnement, démontrant du même coup que la grande majorité des microorganismes contenus dans les différentes niches écologiques n'ont pas encore été cultivés. Ces observations ont mené au développement de la métagénomique, un champ de recherche dédié à l'analyse génomique de tous les microorganismes contenus dans les échantillons, incluant ceux non-cultivables. Les prochaines sections portent sur des travaux majeurs dans ce domaine, sur différentes technologies maintenant utilisées en métagénomique et sur différentes méthodes d'analyse bio-informatiques maintenant essentielles pour traiter les données volumineuses produites par ces nouvelles technologies.

## 2.1. Séquençage de nouvelle génération

Les premières technologies de séquençage de l'ADN datent des années 1970 (51,52). La méthode de Sanger consiste à déterminer l'ordre d'incorporation des nucléotides sur un fragment d'ADN. Elle est donc basée sur la réaction de polymérisation de l'ADN par l'ADN polymérase. Sommairement, l'incorporation séquentielle de nucléotides radiomarqués couplée à une analyse par électrophorèse sur gel de polyacrylamide, permet de déduire la séquence du brin d'ADN amplifié (53). La technique développée par Sanger et coll. est celle qui a été la plus utilisée et a pavé la voie pour le développement du séquençage à haut débit - aussi appelé séquençage de nouvelle génération (NGS).

Essentiellement, les techniques de NGS mesurent aussi l'incorporation séquentielle de nucléotides. Elles diffèrent généralement de par leur méthode de parallélisation du processus et de détection du signal produit par l'incorporation des différents nucléotides. Actuellement, avec les technologies de troisième génération, ces technologies sont les plus utilisées par la communauté scientifique. En raison de la compétition entre les compagnies offrant ces différentes technologies, le prix des séquenceurs a diminué avec le temps et cela a permis de démocratiser le séquençage à haut débit. Les technologies de NGS sont caractérisées par un parallélisme efficace, un meilleur rendement, une facilité d'utilisation, un prix par nucléotide séquencé plus faible, mais ont aussi l'inconvénient de produire des séquences plus courtes (54). L'avantage principal des technologies de troisième génération est la production de séquences plus longues, ce qui peut aider à produire des assemblages de meilleure qualité ou simplement diminuer le besoin en ressources de calcul informatique pour l'identification des séquences. Les désavantages des technologies de troisième génération sont le prix encore élevé de la technologie et un taux d'erreurs de séquençage plus élevé. Pour ces raisons, ces technologies sont moins populaires que celles de NGS. On peut toutefois noter qu'il existe de nouvelles technologies qui

produisent de longues séquences et qui sont plus abordables, comme *Oxford Nanopore MinION* et *PacBio* (54).

Il existe, en général, deux méthodes de séquençage pour obtenir le profil microbien des échantillons : le séquençage du gène de la sous-unité ribosomale 16S (ARNr 16S) et le séquençage métagénomique de type « shotgun », aussi appelé métagénomique non-ciblée. Ces deux méthodes ont bénéficié des technologies de séquençage à haut débit et sont maintenant grandement utilisées dans la recherche portant sur le microbiome.

## 2.2. Métagénomique

La métagénomique a révolutionné la recherche en écologie microbienne en permettant l'analyse génomique directe, sans mise en culture, de populations microbiennes présentes dans un échantillon. En comparaison, les méthodes classiques d'analyse en génomique microbienne dépendent de l'isolation, de la mise en culture et de l'amplification de l'ADN de souches individuelles (55). Bien que ces techniques soient encore utiles pour la caractérisation génomique précise et exacte de microorganismes, elles comportent plusieurs limites pour l'analyse de communautés microbiennes d'échantillons dérivés de l'environnement. Premièrement, une grande partie des communautés microbiennes en provenance de l'environnement, comme celles du microbiote intestinal humain, sont composées de microorganismes non-cultivables. De plus, ces techniques ne permettent pas d'estimer la composition des populations microbiennes dans leur niche écologique étant donné que les méthodes de mise en culture introduisent d'importants biais de sélection (56). Dans une certaine mesure, la métagénomique peut être utilisée pour pallier ces problèmes. Le concept de métagénomique a été introduit par Handelsman et coll. en 1998 (57); il a été décrit comme étant l'étude des microorganismes présents dans un échantillon dérivé d'une niche écologique, incluant ceux non-cultivables, basée sur leur matériel génétique. Cela dit, avant le séquençage de tous les gènes microbiens des échantillons dérivés de l'environnement – incluant ceux codant pour des protéines -, la grande diversité microbienne a d'abord été démontré avec le séquençage du gène marqueur de l'ARNr 16S.

### 2.2.1. Séquençage du gène de la sous-unité ribosomale 16S (ARNr 16S)

Le séquençage du gène de l'ARNr 16S est la méthode qui a été la plus utilisée jusqu'à maintenant pour estimer la composition taxonomique d'échantillons en raison de son efficacité et de son coût

relativement faible. Puisque ce type de séquençage informe exclusivement sur le profil taxonomique des échantillons, il est plus approprié de considérer ce type de séquençage comme une sous-catégorie de la métagénomique : la métataxonomique. Woese et coll., en 1980, ont été les premiers à utiliser le gène de l'ARNr 16S pour évaluer les relations phylogénétiques des bactéries (58). Cela a alors grandement révolutionné les méthodes plus conventionnelles, qui étaient basées sur des informations comme la morphologie et la croissance, entre autres. Le gène de l'ARNr 16S possède des caractéristiques qui permettent de s'en servir comme gène marqueur pour étudier la phylogénie et estimer la composition taxonomique de populations bactériennes au sein d'échantillons biologiques. Premièrement, en raison de son rôle primordial dans la synthèse de protéines, il est présent chez tous les procaryotes et est conservé au niveau de l'espèce, en plus d'être suffisamment différent entre les différents taxons. Le gène de l'ARNr 16S est composé de 9 séquences nucléotidiques hypervariables (V1 à V9) intercalées entre des régions conservées. Ces régions hypervariables sont généralement suffisantes pour distinguer les différents taxons et estimer la composition taxonomique des populations (59). Les régions ciblées dépendent généralement de l'environnement duquel proviennent les échantillons et donc des taxons retrouvés dans cet environnement. Par exemple, lors de choix des amorces, il est conseillé d'inclure la région V4 pour le séquençage d'échantillons dérivés du microbiote intestinal (60). C'est en utilisant les portions conservées du gène que Pace et coll. ont mis au point un protocole de séquençage du gène en utilisant la transcriptase inverse et des séquences nucléotidiques « amorces », complémentaires aux séquences conservées (61). L'avantage principal de ce protocole est la facilité d'obtention de plusieurs séquences différentes grâce à l'universalité, chez les procaryotes, des séquences conservées. Ensuite, grâce aux techniques d'amplification par PCR (*polymerase chain reaction*), cette méthode a été grandement démocratisée et le champ de recherche sur le gène de l'ARNr 16S a connu un important gain de popularité, augmentant ainsi le nombre de bactéries documentées dans les banques de données et du même coup la qualité et la quantité des assignations taxonomiques possibles dans le futur.

Il existe plusieurs stratégies pour construire le profil taxonomique d'échantillons à partir des séquences du gène ARNr 16S. La construction du profil taxonomique se fait généralement en 3 étapes : le contrôle-qualité des séquences (1), le regroupement des séquences (2) et l'identification taxonomique des groupes de séquences par l'alignement avec des banques de données de référence (3). Il existe plusieurs pipelines bio-informatiques pour réaliser ces différentes étapes. Par exemple, mothur (62), QIIME (63), USEARCH (64) et DADA2 (65) font partie des pipelines qui sont couramment utilisés dans

la littérature. Ces outils offrent généralement des fonctions similaires assurées par divers logiciels tiers, mais peuvent différer à certaines étapes. Par exemple, les créateurs de mothur, QIIME et USEARCH proposent l'utilisation de banques de données de références différentes pour l'assignation taxonomique, soit Silva (66), GreenGenes (67) et RDP (68), respectivement. De plus, ces 3 derniers pipelines reposent sur le regroupement des séquences en OTUs (*operational taxonomic units*), tandis que DADA2 propose d'utiliser les séquences individuelles, aussi appelées ASVs (*amplicon sequence variants*). Bien que le regroupement des séquences en OTUs permette, dans une certaine mesure, d'amortir l'effet des erreurs de séquençage et de limiter l'interprétation de ces erreurs comme étant des variations biologiques, les auteurs de DADA2 estiment que le potentiel du séquençage est sous-utilisé avec cette technique. Ils ont donc développé un protocole, utilisant un algorithme d'apprentissage automatique, qui corrige les erreurs de séquençage en se basant sur le taux d'erreur spécifique à chaque lot séquencé et qui permet donc l'analyse précise des variations biologiques du gène, parfois même au niveau des souches bactériennes (65).

### 2.2.2. Le séquençage métagénomique de type « shotgun »

Alternativement, au lieu de séquencer un seul gène marqueur, le séquençage en métagénomique non-ciblée, aussi appelé « shotgun », séquence tout l'ADN contenu dans un échantillon, de manière aléatoire. Ce séquençage produit donc un type de données d'une plus haute complexité que le séquençage du gène ARNr 16s et permet une analyse plus complète et approfondie du microbiome. En d'autres mots, la métataxonomique permet de répondre à la question « Qui est là? », tandis que la métagénomique non-ciblée permet de répondre aux questions « Qui est là? » et « Que peuvent-ils faire ? ». Donc, comme pour le séquençage du gène ARNr 16S, il est possible de constituer le profil taxonomique des échantillons mais ce, de manière plus précise. En effet, il est possible qu'entre certaines espèces bactériennes, les différences dans leur séquence du gène ARNr 16S ne soient pas assez significatives pour les différencier phylogénétiquement. Par contre, d'autres éléments dans leur génome peuvent être amplement suffisants pour y parvenir (69). Bien qu'il soit possible de construire le profil taxonomique des échantillons, une des valeurs ajoutées de ce type de données, lorsque comparée au séquençage du gène ARNr 16S, est la possibilité de construire le profil fonctionnel des échantillons. Comme pour le séquençage du gène ARNr 16s, il existe différentes méthodes d'analyse bio-informatique pour étudier les séquences produites par la méthode « shotgun ».

### 2.2.3. Données dérivées du séquençage métagénomique « shotgun »

Afin de structurer cette prochaine section, les méthodes d'analyse des données de séquençage en métagénomique non-ciblée sont séparées en deux catégories générales :

i) l'alignement des séquences avec des banques de données de référence;

ii) la reconstitution des génomes (assemblages *de novo*).

#### 2.2.3.1. Alignement des séquences brutes

La première méthode, comme son nom l'indique, dépend de l'alignement des séquences obtenues par séquençage avec les séquences contenues dans une banque de données de référence, dont les séquences sont systématiquement traitées « à la main » et annotées. Un des avantages de l'utilisation de l'alignement des séquences est l'exactitude de l'annotation obtenue.

Les méthodes basées sur l'homologie de séquence utilisent généralement l'alignement par paires, dans lequel les séquences sont comparées avec des génomes déjà séquencés pour trouver des sections où les séquences correspondent. Aussi, la taxonomie peut être déduite à partir du génome correspondant. L'outil qui a longtemps été le plus populaire pour effectuer cette méthode est BLAST (*Basic Local Alignment Search Tool*), qui utilise une approche basée sur des *k*-mers pour réaliser l'alignement contre une banque de données de génomes. Les *k*-mers sont des sous-divisions des séquences nucléotidiques, de longueur *k*. Les meilleurs scores d'alignement sont ensuite utilisés pour assigner la meilleure annotation possible. Les méthodes basées sur l'homologie, comme BLAST, sont très précises lorsque la séquence à annoter fait partie d'un génome présent dans la banque de données. Or, leur principal défaut est évidemment l'impossibilité d'annotation lorsque la séquence est absente des banques de données utilisées. De plus, la comparaison de *k*-mers contre des milliers de génomes est un processus qui est lent étant donné que le séquençage d'échantillon produit, de façon générale, plusieurs millions de séquences. Pour cette raison, de nouvelles technologies d'annotation basée sur l'homologie ont été développées. Essentiellement, elles sont plus rapides en raison de l'utilisation de k-mers courts et ne produisent pas d'alignement par paires. Cependant, malgré l'augmentation importante de vitesse d'exécution, ces nouvelles méthodes, comme elles ne produisent pas d'alignement, ne permettent l'identification précise de gènes ou de variations dans une région précise des génomes. Parmi ces outils ultra-rapides, Kraken fait partie de ceux qui sont les plus utilisés. Il utilise une banque de données qui associe chaque *k*-mer à un certain clade, soit le plus petit ancêtre

commun à tous les organismes contentant ce *k*-mer. Le profil taxonomique peut donc être créé en utilisant les clades qui résultent de l'analyse (70).

Il existe aussi des méthodes de classification qui produisent directement l'abondance taxonomique des échantillons en utilisant les séquences brutes. Elles sont basées sur l'homologie et utilisent l'alignement par paires, mais sur des banques de données de tailles réduites. Ces banques de données sont composées de gènes-marqueurs spécifiques à certains clades, de manière à ce que les séquences qui s'alignent avec ces gènes puissent être associées à un clade en toute confiance. Un des outils les plus populaires utilisant cette méthode est MetaPhlAn, dont la banque de données est maintenant rendue à la troisième version (71). Afin d'estimer l'abondance de certaines voies métaboliques dans les échantillons, à partir des séquences brutes, HUMAnN est aussi très utilisé dans la littérature. Il s'agit d'un pipeline composé de plusieurs étapes servant à vérifier quelle information génétique est présente dans l'échantillon séquencé, tant au niveau des gènes que des voies métaboliques, et à déterminer leur abondance relative (72). Malgré la popularité et l'efficacité relative de ces différents outils, l'alignement comporte aussi des limites. Dans les cas des séquences codantes, une similarité de séquence ne garantit pas que la fonction sera la même entre deux gènes. Par exemple, les protéines contenues dans une grande familles peuvent partager une certaine similarité de séquence, mais ces familles peuvent être constituées de plusieurs sous-divisions, qui à leur tour peuvent être régulées de manières différentes dans les cellules et avoir, finalement, des fonctions différentes (73).

#### 2.2.3.2. Les différents niveaux d'inconnu du microbiome

La limite la plus évidente des méthodes d'alignement est que les banques de données de référence ne sont pas exhaustives. Ainsi, ce ne sont pas toutes les séquences qui peuvent être alignées avec celles connues de la littérature. Dans le contexte actuel où de plus en plus d'organismes diversifiés sont séquencés, les efforts d'identification de nouveaux génomes sont devenus encore plus importants pour faire croître la taille des banques de données de référence et ainsi faire avancer la recherche.

Actuellement, une grande portion du microbiome humain demeure encore inconnue. Malgré l'utilisation grandissante des technologies de séquençage et la diversité des échantillons, certains obstacles freinent l'expansion de la portion connue du microbiome. Premièrement, certains microbes sont peu abondants et passent donc sous le seuil de détection des technologies de séquençage.

Deuxièmement, lorsque des microbes sont phylogénétiquement éloignés de ceux connus, une portion importante de leur génome ne peut pas être alignée avec les génomes connus. Troisièmement, étant donné que les espèces microbiennes sont souvent représentées par un nombre restreint de génomes dans les banques de données, le pangénome des espèces n'est pas toujours bien représenté. Pour cette raison, des portions de génome accessoire des nouvelles souches demeurent inconnues puisqu'elles ne peuvent pas être alignées. Finalement, l'association des gènes à une fonction biologique est un obstacle supplémentaire. Segata et coll., en 2019, ont rapporté que 39,6% de l'*Integrated Gene Catalogue* (IGC) (74) – composé de 9.9 millions de gènes microbiens dérivés du microbiome intestinal – ne s'alignait pas avec les banques de données fonctionnelles et que 15 à 20% des 60.4% restants avaient déjà été observés mais l'étiquette de « fonction inconnue » leur avait été attribué (75).

Pour pallier certains de ces obstacles, l'assemblage *de novo* de métagénomes peut être utilisé. Un des avantages dans l'utilisation des assemblages *de novo* est la possibilité d'isoler et de caractériser des génomes de manière non-biaisée, indépendante des banques de données, pour créer de nouvelles connaissances et élargir les banques de données de référence, entre autres.

#### 2.2.3.3. Assemblage

L'assemblage génomique est un processus informatique par lequel les courts fragments bruts produits par le séquençage sont réassemblés en plus longues séquences, appelées contigs. Habituellement, l'assemblage est réalisé en deux étapes : les fragments sont joints via leur section chevauchante, ce qui produit un fragment plus long (le contig) (1) puis en utilisant l'information du séquençage en paires, où les fragments d'ADN sont séquencés dans les deux sens (5' -> 3' et 3' -> 5'), les contigs peuvent être joints, parfois avec un écart en eux, en échafauds (2). Les algorithmes utilisés pour réaliser ces tâches sont sophistiqués et leur fonctionnement ne sera pas abordé dans ce document.

Une fois l'assemblage *de novo* réalisé, il est possible d'étudier les gènes présents sur chacun des contigs. Il existe en général deux méthodes pour identifier les gènes présents sur un contig donné. Encore une fois, l'alignement peut être utilisé pour trouver des gènes qui ont déjà été étudiés ou des séquences qui sont phylogénétiquement proches de celles-ci. Par contre, cette méthode ne permet pas la découverte de nouveaux gènes lorsque ceux-ci sont éloignés de ceux présents dans les banques de données, ce qui est problématique dans un contexte où l'étude des fonctions du

microbiome intestinal, par exemple, nécessite inévitablement la découverte de nouveaux gènes. C'est pourquoi il existe des méthodes d'identification des gènes qui ne reposent pas sur l'alignement avec des banques de données de référence. Globalement, ces outils de détection utilisent différentes caractéristiques – composition G :C, HMMs et autres statistiques – des séquences codantes pour les distinguer des séquences non-codantes. Par exemple, Prodigal utilise l'apprentissage automatique pour apprendre les caractéristiques du génome à l'étude pour ensuite prédire, sur le génome, la position de gènes et des sites d'initiation de la traduction pour obtenir les séquences peptidiques potentielles (76).

À l'échelle des communautés microbiennes, une fois les gènes microbiens trouvés, il est intéressant de continuer les analyses sans utiliser les banques de données de références pour maximiser le potentiel de découvrir de nouvelles connaissances dans les données. Dans ce contexte, pour comparer le contenu génique entre différents génomes, il existe plusieurs techniques de génomique comparative. Une étape primordiale dans pratiquement tous les protocoles de génomique comparative est le regroupement des gènes en groupes d'orthologues (COGs). Essentiellement, il s'agit du regroupement par homologie – ou identité de séquence – de tous les gènes présents dans les échantillons à l'étude pour ensuite évaluer leur distribution. Le choix du seuil d'identité de séquence dépend habituellement du design expérimental de l'étude et de la question de recherche. Par exemple, un seuil d'identité de >99% peut être utile pour savoir si des gènes précis sont présents en plusieurs copies dans un génome (paralogues), ou présents dans des génomes de taxons distincts (orthologues). Dans le contexte des assemblages *de novo* de métagénomes, les COGs peuvent être utilisés pour obtenir l'abondance relative des familles de gènes à l'échelle populationnelle. Ce type de données permet donc une analyse fonctionnelle qui ne requiert pas directement l'annotation via des banques de données de références.

Lorsque les gènes microbiens prédits *in silico* n'alignent pas avec ceux des banques de données de référence, il est possible de prédire leur fonction en utilisant des outils spécialisés dans la recherche de domaines protéiques. Généralement, ces outils utilisent des informations statistiques – comme les profils HMM, par exemple - qui sont caractéristiques aux séquences qui contiennent les domaines protéiques en question. Un score est ensuite attribué aux prédictions faites sur les nouveaux gènes. Cette technique est utile pour vérifier la distribution de certaines fonctions au sein d'échantillon et permet d'aller au-delà des limitations imposées par les banques de données de gènes. Une banque

de données de domaines protéiques souvent utilisée est Pfam (77), qui contient les profils HMM de 19,179 familles de protéines en date de rédaction de ce document.

Bien que plusieurs mécanismes moléculaires aient été démontré dans la relation hôte-microbiome, une multitude de mécanismes restent à être découverts par la recherche. Les efforts grandissants de la communauté scientifique dans le séquençage d'échantillons dérivés d'individus, combinés avec la mise à la disposition du public de ces données, contribuent à la mise en valeur de ces données. À l'ère de l'intelligence artificielle, les données volumineuses produites par les assemblages *de novo*, ainsi que tous les types de données dérivés de ces derniers via les outils mentionnés ci-haut, forment la matière essentielle pour les méthodes d'analyse par apprentissage automatique.

## 3. Apprentissage automatique

L'apprentissage automatique est un champ des sciences informatiques qui utilise des algorithmes spécialisés dans l'extraction de connaissances à partir de données. Il est à la jonction des statistiques, de l'intelligence artificielle et de l'informatique. À l'ère de la connectivité et de la génération massive de données, l'apprentissage automatique est devenu omniprésent au quotidien. Un exemple populaire d'application est celui des recommandations personnalisées proposées par la plupart des moteurs de recherche d'aujourd'hui. Ces logiciels sont arrivés à ce niveau de sophistication en bonne partie en raison de l'apprentissage automatique, pierre angulaire du traitement des données d'utilisation. Mis à part le secteur commercial, l'apprentissage automatique est aussi utilisé dans les champs de recherche où il y a génération substantielle de données. Par exemple, en biologie, un des intérêts de l'utilisation de l'apprentissage automatique est la découverte de connaissances scientifiques sur des phénomènes qui sont éminemment complexes à modéliser. Cela dit, la présence d'une grande quantité de données n'est pas suffisante à elle seule pour appliquer l'apprentissage automatique sur un problème donné. L'application de l'apprentissage automatique repose avant tout sur l'hypothèse qu'il existe une relation entre les données à l'étude et le résultat à prédire (78). Les prochaines sections portent sur les types de problèmes où l'apprentissage automatique peut être utilisé, sur les caractéristiques que doivent posséder les données ainsi que sur les étapes de conception d'une tâche d'apprentissage automatique.

## 3.1. Apprentissage supervisé

L'apprentissage supervisé est utilisé lorsque le but de la modélisation est de prédire un résultat à partir d'échantillons donnés. Il peut être utilisé à condition d'avoir un jeu de données composé de paires d'exemples et d'étiquettes. Par exemple, dans un jeu de données portant sur le microbiome, les exemples correspondent aux échantillons et les variables correspondent aux différentes bactéries (voir Figure 1).

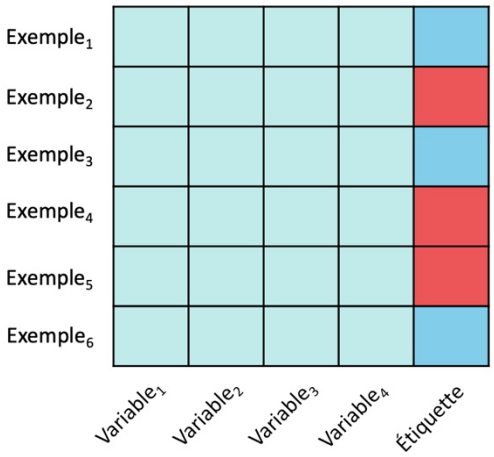


**Figure 1**. Format des données d'entrée et représentation des concepts d'exemple, de variable et d'étiquette

Dans un protocole d'apprentissage automatique, ces paires d'exemples et d'étiquettes – appelés données d'entraînement - serviront, comme leur nom l'indique, à entraîner le modèle. L'expertise humaine est très importante à l'étape de la préparation des données d'entraînement puisque la qualité des décisions prises par le modèle dépend directement de l'exactitude des étiquettes associées aux exemples. Il existe deux sous-types d'apprentissage supervisé : la classification et la régression. Les problèmes de classification consistent à prédire à quelle classe appartiennent les échantillons; la valeur à prédire est donc discrète.

**Ex :** Prédire si les individus sont diabétiques ou non en utilisant le profil bactérien de leur microbiote intestinal.

La classification s'applique souvent dans un contexte binaire, mais peut aussi être utilisée dans un contexte de classification avec de multiples classes. Les modèles de régression sont plutôt construits dans le but de prédire un nombre réel, soit une valeur continue.

**Ex :** Prédire la concentration d'un marqueur sanguin chez des individus en utilisant le profil bactérien de leur microbiote intestinal.

### 3.1.2. Généralisation, surapprentissage et sous-apprentissage

En apprentissage supervisé, on utilise les données d'entraînement pour produire un modèle qui a la capacité de faire des prédictions sur de nouvelles données jamais vues par le modèle. Lorsqu'un modèle performe bien sur de nouvelles données, il est dit que le modèle offre de bonnes performances de généralisation. Le but en apprentissage automatique est donc de maximiser la généralisation des modèles. Comme mentionné précédemment, la seule façon d'évaluer la généralisation d'un modèle est de mesurer ses performances de prédiction sur un ensemble de données de validation. Ces données n'ont jamais été vues par le modèle, mais leur étiquette est connue de l'utilisateur, ce qui permet à ce dernier de vérifier la performance du modèle en comparant les valeurs de prédiction et les valeurs véritables. La validation est une étape importante, puisqu'un modèle qui performe très bien sur les données d'entraînement ne garantit pas d'aussi bonnes performances sur de nouvelles données. En général, deux problèmes peuvent expliquer une mauvaise généralisation. Premièrement, il est possible que le modèle apprenne une fonction qui est spécifique aux données d'entraînement et à leurs particularités; il s'agit du phénomène de surapprentissage. À l'inverse, lorsqu'un modèle apprend une fonction trop simple qui ne capte pas les différentes variabilités dans le jeu de données, il est dit que le modèle sous-apprend. Le modèle qui offrira de bonnes performances de généralisation sera donc celui qui prend en compte la complexité des données d'entraînement sans pour autant apprendre la spécificité des exemples contenus dans les données d'entraînement. En général, ce juste équilibre peut être obtenu en utilisant de bonnes pratiques dans la conception d'un protocole d'apprentissage automatique. Par exemple, bien que les algorithmes d'apprentissage automatique soient conçus pour être utilisés indépendamment des jeux de données, certains de leurs paramètres, appelés hyperparamètres, nécessitent un réglage par l'utilisateur pour optimiser l'apprentissage des modèles. Généralement, la combinaison optimale d'hyperparamètres dépend du jeu de données à l'étude. Il existe des stratégies pour trouver les bonnes combinaisons d'hyperparamètres, comme par exemple la validation croisée.

### 3.1.3. Validation croisée

La validation croisée fait partie des bonnes pratiques à inclure dans un protocole d'apprentissage automatique. Cette méthode est utile pour simuler la généralisation en divisant les données d'entraînement en différents plis (ou sous-groupes). Par exemple, en divisant les données d'entraînement en *n* plis de taille égale (en général, 5 ou 10 plis sont utilisés en validation croisée), le modèle peut apprendre sur *n*-1 plis et mesurer les performances de prédiction sur le pli restant. Ainsi, dans cet exemple de *n* plis, on peut répéter le processus *n* fois en alternant les plis utilisés pour l'entraînement et celui utilisé pour la validation (voir Figure 2). Ce processus produit donc un ensemble de *n* mesures de performance de prédiction. Une manière d'estimer la généralisation du modèle est d'utiliser la moyenne des résultats sur les *n* différents plis. Par exemple, on peut s'attendre qu'un modèle ayant une précision moyenne de 90% en validation croisée aie une performance similaire en utilisant de nouvelles données jamais vues.

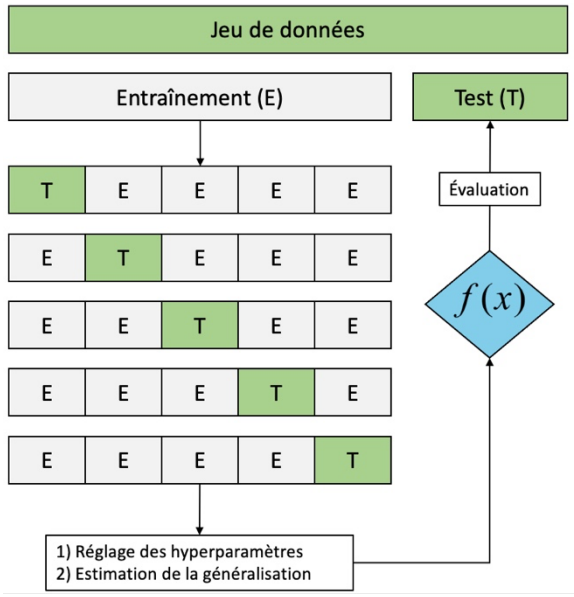


**Figure 2**. Schéma des étapes d'une validation croisée à 5 plis

Il existe plusieurs avantages à utiliser la validation croisée. Premièrement, étant donné que les plis sont formés aléatoirement, qu'ils sont différents les uns des autres et que la performance du modèle est moyennée sur le nombre de plis, cela empêche de sous-estimer ou de surestimer les performances du modèle. Par exemple, si un pli est composé seulement d'exemples qui sont faciles à prédire et que

les exemples qui se retrouvent dans le pli de validation sont difficiles à prédire, les performances du modèle vont paraître médiocres. En diversifiant la composition des différents plis, on s'assure que la moyenne des résultats soit, de manière générale, représentative du jeu de données à l'étude.

La prochaine section présente certains algorithmes utilisés fréquemment dans la littérature et qui sont aussi utilisés dans l'article présenté au chapitre 2.

## 3.2. Algorithmes communs

### 3.2.1. Modèles linéaires

Les modèles linéaires reposent sur l'hypothèse que la variable à prédire peut être prédite, ou estimée, par une combinaison linéaire des variables d'entrée. Pour les problèmes de régression, la prédiction sera égale à la somme des variables pondérées, tandis que pour les problèmes de classification, une classe donnée sera prédite si la somme des variables pondérées est inférieure ou supérieure à un seuil. De manière générale, c'est la méthode de calcul des coefficients de l'équation linéaire qui diffère entre les différents modèles linéaires. Cette section porte sur deux algorithmes classiques pour produire des modèles linéaires.

#### 3.2.1.1. Séparateurs à vastes marges

Les séparateurs à vastes marges (SVM) (79) visent à apprendre une frontière de décision entre les classes qui maximise la distance – ou marge - entre les exemples qui sont les plus proches de la frontière de décision. Ces exemples sont ceux qui sont utilisés par le modèle et sont aussi appelés vecteurs supports. Les vecteurs supports sont donc les exemples qui sont les plus difficiles à discriminer entre les classes. La frontière de décision, dans un espace à *n* dimension(s), est représentée par un hyperplan de *n-1* dimension(s). Le positionnement optimal de l'hyperplan est donc celui qui est le plus loin des vecteurs supports. Lorsque les classes ne sont pas linéairement séparables dans l'espace original de données, des fonctions noyau peuvent être utilisées afin de projeter les données dans un espace de plus hautes dimensions et ainsi trouver une nouvelle discrimination linéaire. Or, l'utilisation de fonctions noyau produit des modèles de plus haute complexité et rend l'étape de l'interprétation du modèle plus difficile.

#### 3.2.1.2. Régression logistique

Malgré son nom, l'algorithme de régression logistique est utilisé à des fins de classification. Par défaut, cet algorithme ne peut être utilisé que pour des tâches de classification binaires. La fonction au cœur de l'algorithme de régression logistique est la fonction logistique. À l'origine, la régression logistique est utilisée pour déterminer la relation entre une ou plusieurs variables indépendantes et une variable dépendante. Sommairement, les coefficients de l'équation linéaire trouvés par l'algorithme de régression logistique est basé sur une mesure de probabilité. Cet algorithme est l'un des plus populaires pour réaliser des tâches de classification binaire portant sur le microbiome. Il a précédemment été utilisé pour établir des signatures microbiennes dans diverses maladies en plus d'offrir des bonnes performances de prédiction (80,81).

### 3.2.2. *Set covering machine*

Le *set covering machine* (SCM) (82) est un algorithme basé sur des règles. Cet algorithme apprend des combinaisons logiques de règles simples, soit des conjonctions et des disjonctions. Chaque règle est sous forme booléenne, qui retourne soit *vrai* ou *faux.* Pour un modèle qui apprend une conjonction de règles, la classe positive est prédite si toutes les règles retournent *vrai*, tandis que pour un modèle apprenant une disjonction de règles, la classe positive est prédite si au moins une règle retourne *vrai*. Cet algorithme a précédemment été utilisé en génomique bactérienne dans la prédiction de phénotypes bactériens et un de ses avantages est l'obtention de modèles parcimonieux et facilement interprétables; les règles obtenues sont simples et peu nombreuses (83,84).

### 3.2.3. Arbres de décisions

Les arbres de décisions peuvent être utilisés pour des tâches de classification ou de régression. Comme le SCM, ils sont aussi basés sur des règles. Ils sont composés d'une hiérarchie de règles sous forme booléenne, qui mènent à une décision. Afin de trouver le bon modèle, l'algorithme cherche la série de règles qui mène au résultat le plus rapidement possible. Pour ce faire, l'algorithme de l'arbre de décision recherche, parmi toutes les possibilités de règles, celle qui est la plus informative sur le jeu de données. Les règles considérées informatives sont celles qui séparent le mieux les classes à prédire. Ainsi, en répondant à la règle trouvée, les exemples sont séparés selon la réponse. Par exemple, une règle qui séparerait tous les exemples de la classe « 0 » de ceux de la classe « 1 » serait considérée parfaite et l'arbre de décision serait constitué d'une seule règle. Or, les jeux de données sont souvent plus complexes et nécessitent davantage de règles pour distinguer les classes. Donc,

après la séparation du jeu de données en deux sous-groupes d'exemples, le même processus est répété dans chaque sous-groupe, de manière récursive, jusqu'à ce que les exemples puissent être assignés à leur classe respective. C'est ce processus récursif qui produit un arbre de décision. Les règles qui séparent parfaitement les exemples dans leur classe respective sont appelées feuilles pures.

La complétion du processus jusqu'à l'obtention de feuilles pures produit un arbre de décision généralement d'une haute complexité et qui performe parfaitement sur les données d'entraînement (100% de précision). Afin de limiter ce surapprentissage, il est possible de contrôler certains paramètres lors de la construction de l'arbre. Par exemple, il est possible d'arrêter le processus récursif précocement. Cela peut être réalisé en contrôlant certains critères comme la profondeur de l'arbre, le nombre maximum de nœuds, ou imposer un nombre minimal d'exemples dans un nœud avant de continuer la création de nouveaux nœuds. Il est aussi possible d'élaguer l'arbre de décision après sa complétion en éliminant des nœuds qui contiennent peu d'information.

Une des forces des arbres de décision est son interprétation intuitive. En effet, une fois l'arbre de décision affiché, il est possible d'observer quelles variables sont les plus importantes selon leur position dans l'arbre et comment les nœuds divisent le jeu de données. Or, dans le cas d'un arbre de décision de grande taille et complexe, il devient difficile de tirer des conclusions par son observation. Une manière de pallier ce problème est d'utiliser la propriété d'importance des variables. Cette propriété résume en quelque sorte le fonctionnement des arbres de décision. Elle peut être calculée de différentes manières, mais la méthode par défaut est l'impureté de Gini, qui attribue un score entre 0 et 1 pour chaque variable et où la somme de tous les scores est égale à 1. La valeur 0 correspond à une variable qui n'est pas utilisée par l'arbre de décision tandis que la valeur 1 correspond à une variable qui prédit parfaitement la classe. Il est à noter qu'une valeur d'importance faible ne signifie pas nécessairement que la variable n'apporte pas d'information. Il se peut qu'une autre variable encode la même information et que cette autre variable soit préférentiellement sélectionnée par l'algorithme. Malgré la possibilité d'élagage, le défaut principal des arbres de décision est qu'ils tendent à surapprendre. Afin de pallier ce problème et obtenir de meilleures performances de généralisation, les méthodes par ensemble peuvent être utilisées.

Les méthodes par ensemble utilisent plusieurs algorithmes d'apprentissage automatique et surpassent généralement les algorithmes utilisés seuls. Parmi les méthodes par ensemble qui se sont révélées

efficaces se trouvent celles qui utilisent de multiples arbres de décision. Cette section porte sur une méthode par ensemble efficace et vastement utilisée : la forêt aléatoire.

### 3.2.4. Forêt aléatoire

La forêt aléatoire, comme son nom l'indique, est composée d'arbres de décision. Cette méthode a surtout été développée pour diminuer le surapprentissage des arbres de décisions employés seuls. L'idée générale des forêts aléatoires est que chaque arbre de décision qui compose la forêt est différent et que chacun des arbres peut être utilisé pour la prédiction, mais ceux-ci, comme mentionné dans la section précédente, sont prédisposés à surapprendre. Or, en utilisant plusieurs arbres de décisions qui performent généralement bien et qui surapprennent de manière différente, il est possible d'obtenir de bonnes performances de généralisation en utilisant la moyenne de prédiction de tous les arbres qui composent la forêt aléatoire. Afin de produire une multitude d'arbres de décisions différents, l'algorithme de forêt aléatoire utilise principalement deux stratégies. Les arbres sont, premièrement, construits à partir de sous-ensembles différents d'exemples. De cette façon, certains exemples ne sont pas présents dans le sous-ensemble et certains peuvent être présents plus d'une fois. Ce processus est répété une fois par arbre de décision qui compose la forêt. La deuxième méthode utilisée pour générer des arbres de décisions différents est la sélection aléatoire de variables lors de la création des nœuds. Ainsi, à chaque nœud, la recherche du meilleur test est effectuée sur un sous-ensemble de variables.

## 3.3. Apprentissage automatique en bio-informatique

Dans un contexte bio-informatique, les données ont généralement des caractéristiques qui peuvent soulever certains problèmes lors de l'application d'un protocole d'apprentissage automatique. Premièrement, il est dit que les données bio-informatiques sont aux prises avec la « malédiction de la dimensionnalité ». Ce problème peut être défini comme étant un très grand nombre de variables pour un nombre relativement faible d'exemples. Par exemple, pour un seul échantillon dérivé du microbiote intestinal d'un individu, des milliers, voire des millions d'éléments génétiques peuvent être considérés comme des variables. Le problème principal que cause cette caractéristique est l'augmentation du risque de surapprentissage puisque beaucoup de bruit est introduit dans les données et les algorithmes risquent d'apprendre des informations trop spécifiques aux exemples. Pour atténuer ce problème, il est possible d'utiliser des méthodes spécialisées dans la réduction de la dimensionnalité, comme l'analyse en composante principale (ACP), par exemple. Il est important de noter qu'il existe une

multitude de méthodes de sélection de variables et que l'expertise du domaine est souvent requise pour choisir les bonnes variables.

Aussi, en bio-informatique, étant donné que le nombre d'échantillon est souvent relativement faible, il est possible de répéter la validation croisée plusieurs fois afin d'atténuer le biais de sélection de données - par exemple un ensemble de validation qui permet de bonnes performances de prédiction, mais qui n'est pas représentatif du jeu de données complet. Pour ce faire, à chaque répétition, le jeu de données est divisé de manière aléatoire en un ensemble d'entraînement et un en ensemble de validation, de manière qu'à chaque répétition, la composition de chacun de ces ensembles ne soit pas la même. Ce processus est appelé validation croisée de Monte-Carlo.

# Mise en contexte, hypothèse et objectifs

Mon projet de maîtrise s'inscrit dans un contexte où de plus en plus de données sont produites dans le but de caractériser la relation du microbiome avec son hôte et où, concomitamment, les avancées en intelligence artificielle permettent une analyse approfondie de ces données. Par exemple, il est maintenant possible de prédire l'état de santé des individus en utilisant les données provenant de leur microbiome. Plusieurs facteurs peuvent influencer la qualité des prédictions faites par les algorithmes d'apprentissage automatique, comme le type de données utilisé en entrée. Actuellement, les méthodes de représentation les plus utilisées dans la littérature – l'abondance taxonomique et l'abondance des voies métaboliques – reposent sur l'annotation des séquences avec des banques de données de référence. Or, cette annotation, bien qu'elle permette d'obtenir de l'information *a priori* sur les échantillons, n'utilise qu'une fraction de l'information séquencée. De fait, les données métagénomiques de type « shotgun » contiennent une majorité de séquences dites *de novo*, qui sont inconnues des banques de données.

L'hypothèse du projet est que l'utilisation de toute l'information obtenue par séquençage de type « shotgun » permet d'obtenir de meilleures performances dans un contexte de prédiction de l'état de santé de l'hôte.

Le premier objectif spécifique du projet est de proposer une méthode de représentation de données qui tient compte de toute l'information obtenue par le séquençage de type « shotgun » sans utiliser les banques de données de référence.

Le deuxième objectif spécifique est de tester notre méthode de représentation dans un contexte de prédiction de la santé de l'hôte et de vérifier si elle permet d'offrir de meilleures performances de prédiction comparativement aux types de données plus conventionnels de l'abondance taxonomique et de l'abondance des sentiers métaboliques.

# Approche méthodologique

Pour répondre aux objectifs du projet, la méthode du regroupement des séquences codantes en groupes d'orthologues (COG) a été sélectionnée. Afin de tester ce type de représentation de données, quatre jeux de données publics ont été utilisés. Ces jeux de données proviennent d'études sur le lien hôte-microbiome et sont composés de données métagénomiques de type « shotgun ». Les études sont de type cas-contrôle et un protocole de classification binaire a donc pu être utilisé. Les jeux de données inclus dans le projet portent sur l'obésité (n=265), le T2D (n=199), la LC (n=237) ainsi que sur le CRC (n=149). Pour des fins de comparaison avec la méthode mise de l'avant avec ce projet, les types de données plus conventionnels – l'abondance taxonomique et l'abondance des sentiers métaboliques - ont aussi été produits. Dans le même ordre d'idée, les quatre jeux de données ont été précédemment utilisés dans la littérature, aussi dans un contexte de prédiction de l'état de santé et il a donc été possible de comparer les résultats du présent projet avec ceux d'autres groupes pour valider nos résultats.

Ainsi, sur chacun des jeux de données, un protocole d'apprentissage automatique a été appliqué. Pour chacun des types de données (l'abondance taxonomique, l'abondance des sentiers métaboliques et le regroupement en groupes d'orthologues), 8 algorithmes différents ont été appliqués : trois séparateurs à vastes marges (SVM), deux régressions logistiques (LR), un arbre de décision (DT), une forêt aléatoire (RF) et un Set Covering Machine (SCM).

Finalement, la filtration de certains COG a permis de mettre en évidence le possible rôle de certaines catégories fonctionnelles de gènes microbiens dans le développement des divers phénotypes à l'étude.

# Chapitre 1 : Reference-free microbiome representation enhances host phenotype classification

## Résumé

Avec les progrès concomitants de la recherche sur le microbiome et en apprentissage automatique, la prédiction du phénotype de l'hôte est maintenant une avenue intéressante pour la découverte potentielle de biomarqueurs ou pour la prédiction de l'état de santé de l'hôte. Les données de métagénomique non-ciblée dérivées du microbiome intestinal humain sont composées d'un ensemble de caractéristiques microbiennes de grande dimension. La modélisation de la relation hôte-microbiome à partir de ce type de données complexe peut être difficile avec les méthodes statistiques plus conventionnelles. Une des limites de ce type de données est la présence de gènes *de novo*, pour lesquels l'annotation avec des banques de données de référence n'est pas possible. Par contre, l'information encodée dans cette « matière noire du microbiome » pourrait potentiellement pointer vers des caractéristiques microbiennes qui jouent un rôle dans le phénotype de l'hôte mais qui sont encore inconnues. Dans cette étude, nous avons comparé les performances de prédiction de différentes approches d'apprentissage automatique selon différents types de données dérivées de la métagénomique non-ciblée. Les différents types de données incluent celles qui sont basées sur l'annotation avec des banques de données de référence, comme les profils taxonomique et fonctionnel, et une méthode de représentation plus granulaire, indépendante des banques de données de référence : les groupes de gènes orthologues (COGs). Pour les 4 jeux de données de type cas-contrôle inclus dans notre étude (diabète de type II (T2D), obésité, cirrhose hépatique (LC) et cancer colorectal (CRC)), les COGs, qu'ils soient utilisés seuls ou en combinaison avec les types de données annotées avec des banques de données de référence, ont permis de meilleures performances de prédiction que les autres types de données. De plus, l'utilisation de sous-ensembles de COGs provenant de certaines catégories fonctionnelles de gènes peut mettre en évidence l'importance de certaines fonctions microbiennes dans le phénotype de l'hôte.

# Reference-free microbiome representation enhances host phenotype classification

Thomas Deschênes[1,3,6], Fred Wilfried Elom Tohoundjona[1,3,6], Pier-Luc Plante[1,6], Vincenzo Di Marzo[1,2,3,4,5,6], Frédéric Raymond[1,3,6]

1. Centre Nutrition, Santé et Société (NUTRISS) - Institut sur la Nutrition et les Aliments Fonctionnels (INAF), Université Laval, Québec, Canada;
2. Centre de recherche de l'Institut universitaire de cardiologie et de pneumologie de Québec (IUCPQ), Québec, Canada;
3. École de nutrition, Faculté des sciences de l'agriculture et de l'alimentation (FSAA), Université Laval, Québec, Canada;
4. Département de médecine, Faculté de Médecine, Université Laval, Québec, Canada;
5. Joint International Unit on Chemical and Biomolecular Research on the Microbiome and its Impact on Metabolic Health and Nutrition (UMI-MicroMeNu),
6. Canada Research Excellence Chair in the Microbiome-Endocannabinoidome mediators Axis in Metabolic Health (CERC-MEND)

**Corresponding author**:

Frédéric Raymond, PhD
Centre Nutrition, Santé et Société (NUTRISS) - Institut sur la Nutrition et les Aliments Fonctionnels (INAF), Université Laval
2440, boulevard Hochelaga, Québec, Québec (Canada), G1V 0A6
Email: frederic.raymond@fsaa.ulaval.ca Telephone : 418.656.2131 p.402529

**Short title**: Microbiome representations influence machine learning predictions

## Abstract

With the concomitant advances in both the microbiome research and the machine learning fields, the prediction of host phenotype has become of great interest for the potential discovery of biomarkers to be used in the prediction of the host's health status. Shotgun metagenomics data derived from the human microbiome is composed of a high-dimensional set of microbial features. The use of such complex data for the modeling of host-microbiome interactions remains a challenge. A drawback in the use of shotgun metagenomics data is the vast content of *de novo* genes for which the annotation using catalogs of known genes is not possible. On the other hand, the information encoded in this "microbial dark matter" might shed light on previously uncharacterised microbial features related to the host phenotype. In this study, we compared the prediction performances of machine learning approaches according to different types of data derived from shotgun metagenomics. The different data types range from the more frequently used, annotation-based methods, like the taxonomic and functional profiles, to the more granular, annotation-free Clusters of Orthologous Groups of proteins (COGs) method. For the four case-control datasets used in the study (type 2 diabetes (T2D), obesity, liver cirrhosis (LC) and colorectal cancer (CRC)), the COGs, whether used alone or in combination with reference-based data types, allowed better classification performances than the taxonomic and functional profiles. In addition, we showed that using subsets of COGs from specific functional categories of genes can highlight the importance of these functions on the host phenotype.

## Introduction

Associations between microbiome-derived features and human diseases have been extensively documented in recent years, especially in chronic diseases (1-4). These efforts have led to an increasing number of publicly available datasets of microbiome in health and disease. Numerous studies have applied machine learning on a wide range of these microbiome datasets to harvest hidden knowledge and better understand the implications of the microbiome in health and disease (5-10). In the context of the inference of host phenotype for disease prediction, such modeling can be used to point toward important microbial features (i.e. biomarkers) or serve as potential diagnostic tools.

High-throughput sequencing technologies allow to measure DNA, including that from non-cultivable microbes, directly from the environment (11). Despite the wide range of microorganisms sequenced in the last decades, multiple obstacles can explain why a considerable part of the microbiome remains unknown (12). Some members of the microbiome remain undetected due to their low abundance; the detection threshold is one of the first obstacle when trying to address the unknown part of the microbiome, known as the "microbial dark matter" (13). Another challenge in this regard is the sequencing of taxa that are phylogenetically far from known genomes and thus difficult to map to these genomes, although sequencing efforts tend to decrease this gap. Plus, because species are often represented by a few numbers of genomes in public databases, a considerable part (i.e., the accessory genome) of the newly sequenced strains remains unmappable to reference genomes, thus losing potentially valuable genetic information (14). Maybe the greatest challenge in building the functional profiles of microbial communities is the difficulty to associate a gene to a specific function. For example, as pointed out in (12), 39.6% of the comprehensive Integrated Gene Catalogue (IGC) of the human gut microbiome – which contains 9.9 million genes (15) - were unmapped to functional databases and 15-20% of the remaining 60.4% have been observed before but had unknown function. Also, the clustering of genes into broad functional categories (or pathways) based primarily on homology might introduce bias in the data and using solely *a priori* information could limit the creation of new knowledge (16). The use of *de novo* assembly of metagenomes can help overcome these limitations; the information encoded in assemblies being not limited to the content of databases. For example, microbial gene prediction, combined with protein function prediction, can point toward new genes implicated in certain phenotypes. On the other hand, this approach yields a high-dimensional set of genes that can be challenging to analyze. However, with the right analysis, this data type can be used effectively to bring

out important microbial features. One such analysis technique is the use of supervised classification to infer relationships between gut microbiome components and the health status of individuals.

In this work, we compared performances of classification models on different data types given in entry - all derived from shotgun metagenomics. Shotgun metagenomics studies often use referenced-based data types like the taxonomic and functional profiles (5). Here, we compared these representation methods with a granular, annotation-free and *de novo* assembly-derived data type that are the clusters of orthologous groups (COGs). A similar method has been previously used in the context of phenotype prediction in infants – that is age, sex, country of origin, delivery type, breastfeeding status, and antibiotic usage. For this cohort, the authors observed that *de novo*, gene-level information yielded better classification performances compared to reference-based taxonomies (6). Here, we systematically applied our protocol in four different cohorts from different case-control studies to evaluate the applicability of reference-free representations independently of context in addition to studying the predictive performances of certain families of genes. This work takes place in a context where an increasing number of microbiome studies use untargeted metagenomics. Our goal is to demonstrate the predictive power of this high-resolution data type and by the same token show that reference-free microbiome representations hold the potential to create new knowledge.

## Results

### The COGs outperform the taxonomic and functional profiles

We selected four human microbiome datasets of shotgun metagenomics from case-control studies with publicly available raw data. The four datasets included in our study were from T2D (n=199) (2), obesity (n=265) (1), LC (n=237) (3) and CRC (n=149) (4). For every dataset, we evaluated the effect of data representation in a prediction task that consisted of a binary classification of healthy and diseased individuals. The different data types used in our study were the taxonomic profile (MetaPhlAn3) (17), the potential metabolic functional profile (HUMAnN3) (17), the COGs and the concatenation of all three, thereafter named early fusion. For every data type in entry, we assessed the predictive performance of 8 classifiers: two support vector machines (SVM) with linear kernels (L1- and L2-regularized), a SVM with a radial basis function kernel (SVM-rbf), two logistic regressions (LR) (L1- and L2-regularized), a decision tree (DT), a random forest (RF) and a set covering machine (SCM).

For every dataset, we applied the machine learning protocol – in which the training phase is based on a 5-fold cross-validation – on every data type, making sure that the same samples were used in the same data split across the different data types. Because the datasets included in the study had imbalanced classes, we evaluated the models with the balanced accuracy. We also included the area under the receiver-operating characteristic curve (AUROC).

In T2D, for the taxonomic and functional profiles, the RF achieved the highest balanced accuracy (on 10 repetitions) of 0.728 $\pm$ 0.044 (mean $\pm$ SD) and 0.713 $\pm$ 0.062, respectively. The results are summarized in Figure 1. In comparison, with the COGs and the early fusion, the RF achieved a score of 0.769 $\pm$ 0.057 and 0.7781 $\pm$ 0.079. For these last two data types, the highest scores were obtained by linear models; the L2-regularized LR achieved a score of 0.789 $\pm$ 0.101 for the COGs and the L2-regularized SVM reached 0.806 $\pm$ 0.076 for the early fusion.

In obesity, the best performing models for the taxonomic and functional profiles were respectively the L2-regularized SVM (0.647 $\pm$ 0.095) and the L1-regularized SVM (0.635 $\pm$ 0.108). For both the COGs and the early fusion, the SVM-rbf achieved the best prediction performance with respective scores of 0.801 $\pm$ 0.088 and 0.786 $\pm$ 0.069.

In liver cirrhosis, for the taxonomic and functional profiles, the highest balanced accuracy was obtained with the RF (0.877 $\pm$ 0.058 for the taxonomic abundance and 0.868 $\pm$ 0.060 for the metabolic pathways abundance). For the COGs, the L1-regularized SVM reached 0.938 $\pm$ 0.033 and the SVM-rbf achieved a score of 0.928 $\pm$ 0.038 for the early fusion.

Lastly, on the CRC dataset, the RF and the L2-regularized LR were the best performing models for the taxonomic and functional profiles at 0.819 $\pm$ 0.065 and 0.814 $\pm$ 0.069, respectively. The COGs, when using the SVM-rbf, achieved a score of 0.844 $\pm$ 0.092. For the early fusion, the L2-regularized LR achieved a balanced accuracy of 0.859 $\pm$ 0.086.

Although our protocol was not optimized for the AUROC, we can use this metric to compare our results with those from the well cited work from (5), as they also applied their protocol on the four datasets that we used. In LC, their best performing SVM had an AUROC of 0.963 $\pm$ 0.027 for marker genes presence (from MetaPhlan2). In comparison, our best AUROC was obtained by the L1-regularized LR at 0.984

$\pm$ 0.009 with the early fusion data type. In CRC, their group reached an AUROC of 0.881 $\pm$ 0.067 for the taxonomic abundance (also from MetaPhlan2) with a RF. With the early fusion and a SVM-rbf, we reached an AUROC of 0.927 $\pm$ 0.053. In obesity, our L2-regularized SVM achieved a score of 0.863 $\pm$ 0.119 for the early fusion and their SVM, on the marker genes presence, had a AUROC of 0.659 $\pm$ 0.073. In T2D, for the presence of marker genes, their SVM had a score of 0.757 $\pm$ 0.056 and our best model, for the early fusion, was the L2-regularized SVM with a score of 0.893 $\pm$ 0.073.

Our learning protocol and theirs had some differences. For example, in T2D, we used only one phase of the cohort (n=199) and they used two phases (n=345). We also used a more recent version of MetaPhlAn and used 5-fold cross-validation (CV) instead of 10-fold CV. In addition to using different algorithms, the hyperparameters grids were also different during CV. To control for these differences, we can look at the AUROC score for the taxonomic abundance data type as we both used it. For the group of (5), the RF yielded the best AUROC for every dataset; 0.946 ± 0.035 in LC, 0.881 ± 0.067 in CRC, 0.656 ± 0.072 in obesity and 0.745 ± 0.056 in T2D. In comparison, with the RF, our group achieved comparable AUROCs of 0.937 ± 0.034 in LC, 0.863 ± 0.060 in CRC, 0.664 ± 0.083 in obesity and 0.797 ± 0.063 in T2D. Consequently, the performance differences are mostly driven by the data type given in entry and not by the design of the protocols.

## Subsets of COGs with specific functions

Despite the high dimensionality of the COGs, the features selected with the initial random forest allowed generally great prediction performances across the different datasets. On the other hand, because of the high content of *de novo* sequences in the COGs, the selected microbial features can be from unknown sources, and this can lead to models that are difficult to interpret or to generalize on new datasets. To correct this problem, we tested our protocol on specific groups of COGs, thus by turning to annotation but keeping the same data structure. Each group consisted of COGs that contained at least one sequence from databases composed of known genes with confirmed function. The databases used in our protocol were from specific functional categories of genes: enzymes implicated in the use of carbohydrates (from the Carbohydrate-Active Enzymes (CAZy) database (18)), genes of antibiotic resistance (from the Mobile Elements and Resistance Genes Enhanced for Metagenomics (MERGEM) database (19)), insertion sequences (from MERGEM), biosynthetic gene clusters (from the Minimum Information about a Biosynthetic Gene cluster (MIBiG) database (20)), enzymes from the comprehensive enzymes database BRENDA (21) and genes containing protein domains potentially

implicated in the production or the degradation of neurotransmitters and endocannabinoids analogs (see Methods). The advantages of filtering the COGs table are twofold; firstly, it reduces the size of the table and decreases the needs in computing resources and secondly, it allows to assess the implication of more specific microbial functions in the development of the host's phenotype.

In T2D, when using only the COGs corresponding to carbohydrate-utilizing enzymes, the L2-regularized LR reached a balanced accuracy of $0.789 \pm 0.072$ and $0.797 \pm 0.093$ when using COGs matching with the BRENDA database. These scores are comparable with the early fusion and with all the COGs.

In obesity, the use of COGs corresponding to the BRENDA database led to a balanced accuracy of $0.756 \pm 0.088$ with the L2-regularized LR. The same algorithm reached a score of $0.702 \pm 0.105$ when the COGs were filtered by the CAZy database. Although these scores are lower than the ones obtained with all the COGs and the early fusion, these subsets of genes achieved better scores than the taxonomic and functional profiles.

In LC, the filtration with BRENDA led to a balanced accuracy of $0.929 \pm 0.036$ with a L2-regularized LR. With the COGs filtered by the CAZy database, the SVM-rbf achieved a score of $0.867 \pm 0.047$. The COGs filtered with the insertion sequences led to a comparable score of $0.849 \pm 0.055$ with the L1-regularized LR, a score comparable with the taxonomic abundance.

In CRC, the filtration by BRENDA used with a L2-regularized SVM reached a score of $0.850 \pm 0.085$. When using the CAZy database, the SVM-rbf reached a score of $0.823 \pm 0.077$.

In the four datasets, the biosynthetic gene clusters, the insertion sequences and the antibiotic resistance genes did not perform as well as the COGs – whether filtered or not with the BRENDA and the CAZy databases - and the early fusion. On the other hand, some of these categories of genes did achieve scores comparable to the taxonomic and functional profiles.

## Potential implication of microbial gut-brain axis effectors

The human gut microbiota and the host central nervous system interact by various means of communication that include the neural, endocrine and immune systems (22). Studies of animal models subjected to gut microbiota-altering treatments suggested a role of gut microbiota in the modulation of

anxiety, mood cognition and pain (23). In humans, high-throughput sequencing technology allowed to reveal associations between various pathological states and gut microbiota (24, 25). The synthesis potential of neuroactive mediators by gut prokaryotes has been shown previously (22). Also, the endocannabinoids (2-arachidonoylglycerol (2-AG) and N-arachidonoylethanolamide (AEA)) and their congeners – the comprehensive set of these molecules, the enzymes implicated in their anabolic and catabolic pathways and their receptors now being referred as the endocannabinoidome (26) – are important regulators of both the intestinal function and the gut-brain axis (27, 28). It has also been shown that commensal bacteria can produce endocannabinoids analogs and therefore have the potential to interact with gut function (29). Knowing this, we hypothesized that the abundance of microbial genes containing Pfam (30) domains from enzymes implicated in the production and degradation of endocannabinoids analogs and other neurotransmitters could encode new information about the crosstalk between dysbiosis, the expanded endocannabinoid system and disease. This could also highlight the potential roles of microbial gut-brain axis effectors in contexts other than psychiatric and neurological disorders.

We used only the subset of COGs corresponding to the aforementioned genes to realise the same prediction task on every dataset. In T2D, the use of this subset of COGs led to a balanced accuracy of 0.774 ± 0.060 with a L2-regularized LR. In obesity and CRC, the SVM-rbf reached a score of 0.761 ± 0.076 and 0.824 ± 0.089, respectively. In LC, the L1-regularized SVM reached 0.917 ± 0.059. Compared to the non-filtered COGs and the early fusion, this subset of COGs yielded close prediction performances despite the reduced number of features. This indicates that microbial gut-brain axis effectors might play a significant role in the development of disease.

## The impact of algorithms and data types on prediction performance

As noticeable in Figure 1, we observed that the data types that include COGs – all COGs, early fusion and some of the filtered COGs tables - improved the prediction performance of most of the models, in each dataset, compared to taxonomic or functional representations. Also, for these data types, when comparing linear models, we see that the best performing models tend to be more complex. In T2D, obesity and CRC, the L2-regularized SVM and LR had the highest scores compared to other models. Compared to L1 regularization, which forces the weights of uninformative features to be zero and therefore produced models that are sparser, L2 regularization penalizes weights toward zero without making them equal to zero. Therefore, with L2 regularization, the resulting models include generally

more features. Unfortunately, the generally worst performing models are the DT and the SCM. These rule-based models are known for their high interpretability. The SCM, for example, is known to produce sparse models made of a small number or rules and its interpretation can point toward interesting features. The fact that more complex models perform generally better might indicate that the characteristic microbe-microbe interactions in the different classes are difficult to model and therefore more features are needed. For the taxonomic and functional profiles, RF was the best performing model in T2D, LC and CRC.

## Discussion and conclusion

In this study, we realised a binary classification task that consisted in the prediction of diseased and healthy individuals based on shotgun metagenomics data derived from feces. The main objective was to assess the predictive power of shotgun metagenomics data when used at a level of high granularity, that is at the gene family level. We showed that clustering all the genes present in metagenome *de novo* assemblies, by sequence identity, allowed generally an important gain in prediction performances compared to taxonomic and functional profiles and that this observation was independent of the dataset used.

This gain in prediction performance might be due to several characteristics of this type of data. One of the benefits of using the COGs without annotation from common databases is that the biological roles of the genes are not limited to *a priori* information. Therefore, the COGs data type is likely to encode the information whether a gene is biologically implicated in unsuspected but important functions. For example, a gene might be aggregated and fall under a specific KEGG pathway when using reference-based functional annotation, but that same gene could have another role in an unknown microbial function. Plus, because *de novo* genes with unknown function are not eliminated from the data, the potentially important biological information encoded in these genes might increase the quality of the predictions. Also, the good predictive performances of gene-level data type in host phenotype prediction might indicate that microbial functions are in a closer relationship with the phenotype than the microbes themselves.

Another advantage of using the COGs is the possibility of testing specific COGs across different datasets. For example, to evaluate a model on new samples, their sequences can be clustered according to the COGs of interest. This facilitates the application of models on distinct and unrelated

datasets. Also, isolated bacterial genomes can be screened for specific COGs in the context of mechanistic, in-vitro studies.

However, a risk when using this high-dimensional data type is the learning of subject-specific features instead of class-defining features. To control for this problem, we used an embedded feature selection technique that consisted in training a random forest, ranking the features according to the resulting model, and learning on subsets of these ranked features, a method that has been used previously (5, 10). It allowed to decrease the noise in the data and in the end, limit overfitting.

A potential limit of this study is the absence of a cross-cohort validation of the models. In the context of disease prediction as a potential helper in diagnostics, this step is an effective way to verify the generalization performances of the models (5). However, publicly available datasets from the same diseases are difficult to access or the study from which they are derived have a specific experimental design that makes the transfer of the models impossible (i.e. shotgun metagenomics vs 16S rRNA sequencing). Arguably, though, in the context of biomarker discovery or creation of new knowledge, the cross-cohort validation is not a necessity. For example, the use of gene-level data in a specific disease holds the potential to point toward interesting microbes related to the disease and their subsequent isolation and cultivation could confirm the observation made *in silico*.

Another potential drawback of this analysis is the presence of numerous correlated features. For example, even at the taxonomy level, measuring pairwise correlation can be computationally challenging. As the number of possible two-features interactions for a dataset with $n$ features is $(n*(n-1))/2$, this often results in millions of calculations since microbiomes are composed of thousands of microorganisms. Plus, because microbes live in community, it is likely that there are three-feature, four-feature interactions and more (31). These problems are exacerbated when using microbiome data at a gene-level granularity; the genes being carried by microbes, the aforementioned microbe-microbe interactions are also present at gene level. The topic of dimensionality reduction for gene-level metagenomics data might be a good continuation to this study.

# Methods

## Production of the different data types

The taxonomic and functional profiles produced respectively with MetaPhlAn3 and HUMAnN3 (17) were downloaded via the R package curatedMetagenomicData (version 1.20.0) (32). The COGs matrices were produced following three steps: the assembly of reads into contigs using MEGAHIT (version 1.2.9) (33), the prediction of protein-coding genes and their subsequent translation into amino acids using Prodigal with the default parameters (version 2.6.3) (34) and the clustering of the resulting amino acids sequences at a 70% identity threshold using CD-HIT (version 4.7) (35) with parameters -c 0.70 -aS 0.50 -d 0 -M 0 -T 0 -g 1 -G 0. To produce the filtered COGs tables, we first clustered together both the sequences from the databases and the sequences of the samples. The sequences used for this analysis were from the CAZy, MERGEM, BRENDA and MIBiG databases. Next, we kept every COG that contained at least one sequence from the databases and one sequence from the samples. To select the COGs containing genes potentially implicated in the production or the degradation of neurotransmitters and endocannabinoids analogs, we constructed a local database of Hidden Markov Model (HMM) profiles of protein domains (from Pfam) found in enzymes included in the pathways of synthesis and degradation of neurotransmitters, endocannabinoids and endocannabinoids analogs. We then searched for these HMM profiles in the sequences of each COG's representative sequence with the command hmmsearch (-E 0.001) from HMMER (36) and kept every COG whose representative sequence matched at least one of the HMM profiles.

## Machine learning protocol

The first step of our pipeline is the application of a simple RF - *scikit-learn*'s *RandomForestClassifier* with 200 decision trees (*n_estimators*) and all other hyperparameters set at default – on all remaining features after the preprocessing steps. We then sort the features according to the "*features_importances_"* method, which is a score derived from the Gini impurity, where the higher the score is, the more important is the feature. We then apply the downstream steps of our protocol with top *k*-th subsets of these ranked features, by different *k* increments in the set {1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, …, 490, 500}. For every subset of features, we applied our main protocol for every tested algorithm, which is divided in four steps. The scores reported in Figure 1 are those obtained with the best combination of hyperparameters, including *k*, the number of selected features. The first step of the protocol is to divide the dataset into independent training and testing sets, using 80% and 20%

of the data, respectively. The second step is to apply a 5-fold CV on the training set to find the best combination of hyperparameters. The third step is to retrain the best performing model on all the training set. Finally, the fourth step is to use the testing set to assess the generalizability of the model on unseen data by computing the following metrics: accuracy, balanced accuracy, F1-score, precision, recall and AUROC. This protocol was repeated 10 times, using different random partitions, to obtain more precise estimates of the generalization performance.

SVMs are algorithms that aim at finding the hyperplane that maximizes the margin between the samples from each class (37). When the classes are not linearly separable, kernel functions can be used to map data to a higher dimensional space to find the hyperplane. For hyperparameters tuning, the regularization parameter C was chosen in the set $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.25, 0.5, 0.8, 0.9, 1, 10\}$ for both the SVM with linear and rbf kernels. The same values were used for the C parameter in L1 and L2-regularized LR. For the DT, the maximum depth of the tree was in {1, 3, 5, 10, 25}. The minimum number of samples required to split a node was in {2, 5, 10} and the minimum number of samples required to be at a leaf node was in {1, 2, 4}. For RF, the number of trees, the minimum number of samples required to split a node, the maximum depth of the trees and the minimum number of samples required to be at a leaf node were in {100, 200, 500}, {2, 5, 10}, {1, 3, 5, 10, 25} and {1, 2, 4}, respectively. The SCM is a rule-based algorithm that learns conjunctions (logical-AND) and disjunctions (logical-OR) which are logical combinations of rules (38, 39). For SCM, the trade-off parameter for the utility function was in {0.5, 1, 2}, the maximum number of rules was in {1, 2, 3, 4, 5} and the type of model varied between conjunction and disjunction.

## Acknowledgment

T.D., P.L.P., V.M. and F.R conceived and design the work. T.D. and F.W.E.T. carried out the experimental work. T.D., P.L.P., V.M. and F.R. contributed to data analysis and interpretation. T.D. and F.R. drafted the manuscript.

## Figures and tables

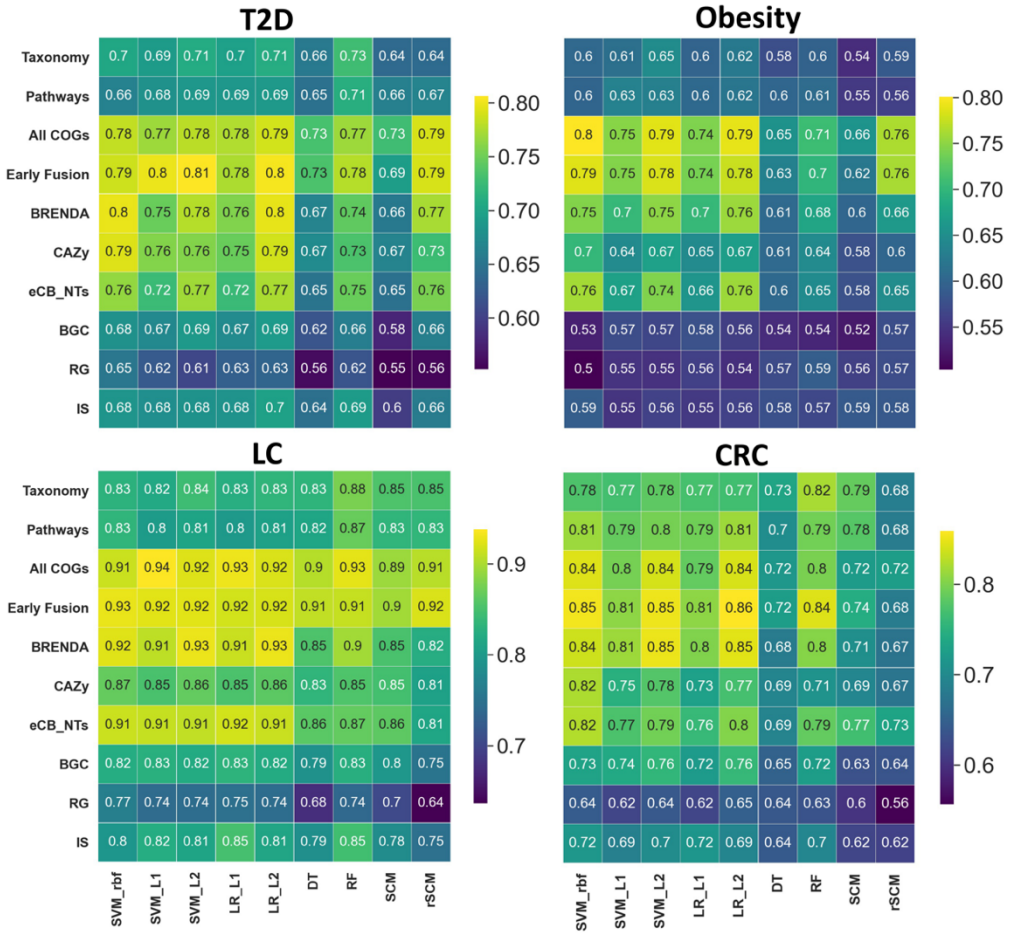


**Figure 1**. Heatmaps of the results for every combination of algorithm and data type

Algorithms are on the X axis and data types are on the Y axis. Performance evaluated with mean balanced accuracy on 10 different data splits. T2D: Type 2 diabetes; LC: Liver cirrhosis; CRC: Colorectal cancer. Taxonomy : MetaPhlAn3 taxonomic profile; Pathways : HUMAnN3 functional profile; All COGs: All orthologs; Early Fusion: concatenation of Taxonomy, Pathways and All COGs; BRENDA: intersection of BRENDA database and All COGs; CAZy: intersection of CAZy database and All COGs; eCB_NTs: COGs containing protein domains of interest as described in Methods; BGC: intersection of MiBIG database and All COGs; RG: intersection of antibiotic resistance genes from MERGEM database and All COGs; intersection of mobile elements from MERGEM database and All COGs

## References


1. MetaHIT consortium, Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, et al. Richness of human gut microbiome correlates with metabolic markers. Nature. 2013 Aug 29;500(7464):541–6.

2. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012 Oct;490(7418):55–60.

3. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature. 2014 Sep 4;513(7516):59–64.

4. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol. 2014 Nov;10(11):766.

5. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. Eisen JA, editor. PLOS Comput Biol. 2016 Jul 11;12(7):e1004977.

6. Le Goallec A, Tierney BT, Luber JM, Cofer EM, Kostic AD, Patel CJ. A systematic machine learning and data type comparison yields metagenomic predictors of infant age, sex, breastfeeding, antibiotic usage, country of origin, and delivery type. Segata N, editor. PLOS Comput Biol. 2020 May 11;16(5):e1007895.

7. Beck D, Foster JA. Machine Learning Techniques Accurately Classify Microbial Communities by Bacterial Vaginosis Characteristics. White BA, editor. PLoS ONE. 2014 Feb 3;9(2):e87830.

8. Wu H, Cai L, Li D, Wang X, Zhao S, Zou F, et al. Metagenomics Biomarkers Selected for Prediction of Three Different Diseases in Chinese Population. BioMed Res Int. 2018;2018:1–7.

9. Oh TG, Kim SM, Caussy C, Fu T, Guo J, Bassirian S, et al. A Universal Gut-Microbiome-Derived Signature Predicts Cirrhosis. Cell Metab. 2020 Nov;32(5):878-888.e6.

10. Baxter NT, Ruffin MT, Rogers MAM, Schloss PD. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Med. 2016 Dec;8(1):37.

11. Handelsman J. Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiol Mol Biol Rev. 2004 Dec;68(4):669–85.

12. Thomas AM, Segata N. Multiple levels of the unknown in microbiome research. BMC Biol. 2019 Dec;17(1):48.

13. Marcy Y, Ouverney C, Bik EM, Losekann T, Ivanova N, Martin HG, et al. Dissecting biological “dark matter” with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc Natl Acad Sci. 2007 Jul 17;104(29):11889–94.

14. McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T, et al. Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations. Hughes D, editor. PLOS Genet. 2016 Sep 12;12(9):e1006280.

15. MetaHIT Consortium, Li J, Jia H, Cai X, Zhong H, Feng Q, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol. 2014 Aug;32(8):834–41.

16. Heintz-Buschart A, Wilmes P. Human Gut Microbiome: Function Matters. Trends Microbiol. 2018 Jul;26(7):563–74.

17. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife. 2021 May 4;10:e65088.

18. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res. 2009 Jan;37(Database issue):D233-238.

19. Déraspe M. Développement d’une base de données sur la résistance aux antibiotiques et son utilisation en génomique [Master’s thesis]. Université Laval; 2015.

20. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res. 2019 Oct 15;gkz882.

21. Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, et al. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res. 2021 Jan 8;49(D1):D498–508.

22. Valles-Colomer M, Falony G, Darzi Y, Tigchelaar EF, Wang J, Tito RY, et al. The neuroactive potential of the human gut microbiota in quality of life and depression. Nat Microbiol. 2019 Apr;4(4):623–32.

23. Cryan JF, Dinan TG. Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. Nat Rev Neurosci. 2012 Oct;13(10):701–12.

24. Song Y, Liu C, Finegold SM. Real-Time PCR Quantitation of Clostridia in Feces of Autistic Children. Appl Environ Microbiol. 2004 Nov;70(11):6459–65.

25. Arneth BM. Gut–brain axis biochemical signalling from the gastrointestinal tract to the central nervous system: gut dysbiosis and altered brain function. Postgrad Med J. 2018 Aug;94(1114):446–52.

26. Veilleux A, Di Marzo V, Silvestri C. The Expanded Endocannabinoid System/Endocannabinoidome as a Potential Target for Treating Diabetes Mellitus. Curr Diab Rep. 2019 Nov;19(11):117.

27. Iannotti FA, Di Marzo V. The gut microbiome, endocannabinoids and metabolic disorders. J Endocrinol. 2021 Feb;248(2):R83–97.

28. Sharkey KA, Wiley JW. The Role of the Endocannabinoid System in the Brain–Gut Axis. Gastroenterology. 2016 Aug;151(2):252–66.

29. Cohen LJ, Esterhazy D, Kim S-H, Lemetre C, Aguilar RR, Gordon EA, et al. Commensal bacteria make GPCR ligands that mimic human signalling molecules. Nature. 2017 Sep 7;549(7670):48–53.

30. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D412–9.

31. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME J. 2016 Jul;10(7):1669–81.

32. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, et al. Accessible, curated metagenomic data through ExperimentHub. Nat Methods. 2017 Nov;14(11):1023–4.

33. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015 May 15;31(10):1674–6.

34. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010 Dec;11(1):119.

35. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012 Dec;28(23):3150–2.

36. Eddy SR. Accelerated Profile HMM Searches. Pearson WR, editor. PLoS Comput Biol. 2011 Oct 20;7(10):e1002195.

37. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995 Sep;20(3):273–97.

38. Marchand M, Shawe-Taylor J. The Set Covering Machine. Journal of Machine Learning Research. 2002 Dec;723–46.

39. Drouin A, Letarte G, Raymond F, Marchand M, Corbeil J, Laviolette F. Interpretable genotype-to-phenotype classifiers with performance guarantees. Sci Rep. 2019 Dec;9(1):4071.

# Conclusion

Les travaux réalisés dans la cadre de mon projet de maîtrise ont permis de démontrer l'impact de la représentation des données dérivées du microbiote dans un contexte d'apprentissage automatique. Plus concrètement, nous avons pu démontrer le potentiel des données de séquençage de type « shotgun » lors d'une tâche de classification de métagénomes dérivés d'individus en santé ou malades. Notre hypothèse était que ce type de données, d'une haute granularité et contenant de l'information encore inconnue des banques de données, permettait aux algorithmes de déceler plus de relations significatives entre les données qu'avec les méthodes de représentation plus conventionnelles. En conséquence, nous avons supposé que les performances de classification puissent être augmentées, lorsque comparées avec les méthodes de représentation plus classiques. Notre hypothèse a pu être confirmée en répondant aux deux objectifs spécifiques du projet : la production d'un type de données qui prend en compte le plus possible la haute dimensionnalité des données dérivées du séquençage en métagénomique non-ciblée (1) et l'utilisation de ce type de représentation dans des tâches de classification d'individus en santé ou malades et ce, dans des contextes différents (2).

Le premier objectif spécifique du projet a été atteint en utilisant le regroupement, par homologie de séquence, des gènes des échantillons prédits *in-silico*. Ainsi, en ne filtrant pas les gènes obtenus avec des banques de données de références comme c'est le cas dans plusieurs pipelines retrouvés dans la littérature, les groupes de gènes orthologues (COGs) contiennent à la fois l'information connue des banques de données de référence, mais aussi une quantité importante de gènes *de novo*, dont le rôle est encore inconnu. Malgré l'avantage que représente l'utilisation de gènes *de novo*, un autre avantage d'utilisation des COGs est la possibilité d'utiliser des banques de données de référence pour filtrer et annoter les COGs. Ainsi, après avoir démontré le bon pouvoir prédictif des COGs pris dans leur ensemble, nous avons pu mettre en évidence le pouvoir prédictif de certaines catégories fonctionnelles de gènes. Cette étape a permis de réaliser que malgré l'utilisation exclusive de gènes orthologues à ceux contenus dans les banques de données, la prédiction était tout de même supérieure, pour certaines catégories fonctionnelles de gènes, comme les gènes d'utilisation des glucides et les possibles effecteurs de l'axe cerveau-intestin, à celle obtenue avec les profils taxonomiques et fonctionnels dérivés des séquences brutes. Cela suggère donc que l'utilisation de gènes individuels,

ou de familles de gènes, encode plus d'information utile pour la prédiction que lorsque les gènes sont agglomérés en un seul groupe pour former un sentier (ou module) métabolique.

Le deuxième objectif de mon projet, qui était d'utiliser la méthode de représentation proposée au premier objectif dans une tâche de classification, a été réalisé dans quatre contextes différents, soit quatre maladies aux étiologies différentes, afin de vérifier l'applicabilité de la méthode proposée indépendamment du type de maladie à l'étude. Ainsi, nous avons utilisé des jeux de données publics qui portent sur le diabète de type 2, l'obésité, la cirrhose hépatique et le cancer colorectal. Afin de comparer notre méthode avec ce qui est réalisé dans la littérature, nous avons comparé les performances de nos modèles avec les profils taxonomiques et fonctionnels produits respectivement par MetaPhlAn3 et HUMAnN3. En appliquant le même protocole d'apprentissage automatique pour tous les jeux de données et pour toutes les représentations utilisées, nous avons remarqué un gain en performance de classification en utilisant les COGs.

Une des forces de notre étude est qu'elle fait partie, à notre connaissance, des premiers travaux qui adressent le problème de l'utilisation exclusive de l'information *a priori* en métagénomique non ciblée, dans un contexte de classification en apprentissage supervisé. Nos travaux s'inscrivent donc dans un contexte de mise en valeur des données, dont le but est évidemment l'avancement des connaissances scientifiques sur l'impact de l'écologie microbienne sur la santé humaine. À l'ère de la grande capacité de calcul informatique des laboratoires et des universités, le grand volume d'information contenue dans les données métagénomiques représente une véritable source de savoir à porter de main. Or, il est difficile d'utiliser ce type de données à son plein potentiel avec les analyses statistiques traditionnelles, dû à leur grande dimensionnalité et à leur complexité sous-jacente. Ainsi, en combinant les expertises de microbiologie et d'apprentissage automatique, les modèles produits peuvent servir à pointer vers des gènes ou des bactéries qui peuvent par la suite être sujets à une isolation, une mise en culture et à des expériences pour confirmer les observations faites à l'ordinateur.

Une des limites de l'étude est l'absence de validation des modèles dans des cohortes différentes portant sur les mêmes maladies. Cette forme de validation peut être utilisée pour s'assurer que les variables considérées par les modèles ne sont pas spécifiques aux individus présents dans la cohorte utilisée pour entraîner les modèles. De surcroît, le type de données très granulaires des COGs est sujet à produire des modèles biaisés par des variables qui sont simplement spécifiques aux individus et qui ne sont pas caractéristiques de la classe à laquelle ces derniers appartiennent. Cela dit, ce

dernier problème a été adressé en utilisant une étape de sélection de variables, ce qui diminue les chances de surapprentissage. Aussi, bien que la validation des modèles dans des cohortes différentes soit nécessaire dans le contexte de développement d'un outil de prédiction de maladies, cela est différent dans un contexte de découverte de nouvelles connaissances. Par exemple, une fois que les modèles ont des bonnes performances de prédiction sur un jeu de données, il est possible d'interpréter le modèle afin de vérifier quels éléments du microbiome sont importants dans la prise de décision du modèle. Ensuite, avec des analyses informatiques et de laboratoire, il est possible d'étudier en profondeur certaines observations faites par les modèles et possiblement créer de nouvelles connaissances en relation avec la maladie en question.

En conclusion, ce projet de maîtrise a pu montrer le potentiel de l'information encodée dans les données produites par métagénomique non-ciblée dans un contexte d'apprentissage supervisé. Toutefois, la méthode de représentation présentée dans ce mémoire produit un type de données d'une haute dimensionnalité, ce qui, généralement, peut limiter les performances de généralisation des modèles produits. Étant donné la grande taille des matrices utilisées, la réduction de dimensionnalité représente un défi considérable qui peut demander beaucoup de ressources de calcul. En continuation de ce projet, des méthodes de réduction de dimensionnalité pourraient être investiguées davantage. Il serait aussi intéressant, dans le futur, de mettre l'emphase sur l'interprétation des modèles. En effet, la découverte possible d'effecteurs microbiens méconnus pourrait augmenter la portée des données métagénomiques de type « shotgun ».

# Bibliographie


1. Lederberg J, Mccray A. `Ome Sweet `Omics--A genealogical treasury of words. The Scientist. 2001;15(7):8-8.

2. Berg G, Rybakova D, Fischer D, Cernava T, Vergès M-CC, Charles T, et al. Microbiome definition re-visited: old concepts and new challenges. Microbiome. 2020 Dec;8(1):103.

3. Lynch SV, Pedersen O. The Human Intestinal Microbiome in Health and Disease. Phimister EG, editor. N Engl J Med. 2016 Dec 15;375(24):2369–79.

4. Fan Y, Pedersen O. Gut microbiota in human metabolic health and disease. Nat Rev Microbiol. 2021 Jan;19(1):55–71.

5. Silva YP, Bernardi A, Frozza RL. The Role of Short-Chain Fatty Acids From Gut Microbiota in Gut-Brain Communication. Front Endocrinol. 2020 Jan 31;11:25.

6. Backhed F, Ding H, Wang T, Hooper LV, Koh GY, Nagy A, et al. The gut microbiota as an environmental factor that regulates fat storage. Proc Natl Acad Sci. 2004 Nov 2;101(44):15718–23.

7. Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. Obesity alters gut microbial ecology. Proc Natl Acad Sci. 2005 Aug 2;102(31):11070–5.

8. Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Human gut microbes associated with obesity. Nature. 2006 Dec;444(7122):1022–3.

9. MetaHIT consortium, Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, et al. Richness of human gut microbiome correlates with metabolic markers. Nature. 2013 Aug 29;500(7464):541–6.

10. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012 Oct;490(7418):55–60.

11. Sánchez-Alcoholado L, Ramos-Molina B, Otero A, Laborda-Illanes A, Ordóñez R, Medina JA, et al. The Role of the Gut Microbiome in Colorectal Cancer Development and Therapy Response. Cancers. 2020 May 29;12(6):1406.

12. Chen W, Liu F, Ling Z, Tong X, Xiang C. Human Intestinal Lumen and Mucosa-Associated Microbiota in Patients with Colorectal Cancer. Moschetta A, editor. PLoS ONE. 2012 Jun 28;7(6):e39743.

13. Saffarian A, Mulet C, Regnault B, Amiot A, Tran-Van-Nhieu J, Ravel J, et al. Crypt- and Mucosa-Associated Core Microbiotas in Humans and Their Alteration in Colon Cancer Patients. Parkhill J, editor. mBio [Internet]. 2019 Aug 27 [cited 2021 Jul 26];10(4). Available from: https://journals.asm.org/doi/10.1128/mBio.01315-19

14. Shen XJ, Rawls JF, Randall TA, Burcall L, Mpande C, Jenkins N, et al. Molecular characterization of mucosal adherent bacteria and associations with colorectal adenomas. Gut Microbes. 2010 May;1(3):138–47.

15. Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. Nat Commun. 2015 May;6(1):6528.

16. Viljoen KS, Dakshinamurthy A, Goldberg P, Blackburn JM. Quantitative Profiling of Colorectal Cancer-Associated Bacteria Reveals Associations between Fusobacterium spp., Enterotoxigenic Bacteroides fragilis (ETBF) and Clinicopathological Features of Colorectal Cancer. McDowell A, editor. PLOS ONE. 2015 Mar 9;10(3):e0119462.

17. Louis P, Hold GL, Flint HJ. The gut microbiota, bacterial metabolites and colorectal cancer. Nat Rev Microbiol. 2014 Oct;12(10):661–72.

18. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med. 2019 Apr;25(4):667–78.

19. Rubinstein MR, Wang X, Liu W, Hao Y, Cai G, Han YW. Fusobacterium nucleatum Promotes Colorectal Carcinogenesis by Modulating E-Cadherin/β-Catenin Signaling via its FadA Adhesin. Cell Host Microbe. 2013 Aug;14(2):195–206.

20. Nakatsu G, Zhou H, Wu WKK, Wong SH, Coker OO, Dai Z, et al. Alterations in Enteric Virome Are Associated With Colorectal Cancer and Survival Outcomes. Gastroenterology. 2018 Aug;155(2):529-541.e5.

21. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature. 2014 Sep 4;513(7516):59–64.

22. Schnabl B, Brenner DA. Interactions Between the Intestinal Microbiome and Liver Diseases. Gastroenterology. 2014 May;146(6):1513–24.

23. Chen Y, Yang F, Lu H, Wang B, Chen Y, Lei D, et al. Characterization of fecal microbial communities in patients with liver cirrhosis. Hepatology. 2011 Aug;54(2):562–72.

24. Garcia-Tsao G, Wiest R. Gut microflora in the pathogenesis of the complications of cirrhosis. Best Pract Res Clin Gastroenterol. 2004 Apr;18(2):353–72.

25. De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, et al. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc Natl Acad Sci. 2010 Aug 17;107(33):14691–6.

26. Oh TG, Kim SM, Caussy C, Fu T, Guo J, Bassirian S, et al. A Universal Gut-Microbiome-Derived Signature Predicts Cirrhosis. Cell Metab. 2020 Nov;32(5):878-888.e6.

27. Hotamisligil GS. Inflammation and metabolic disorders. Nature. 2006 Dec;444(7121):860–7.

28. Anhê FF, Jensen BAH, Varin TV, Servant F, Van Blerk S, Richard D, et al. Type 2 diabetes influences bacterial tissue compartmentalisation in human obesity. Nat Metab. 2020 Mar;2(3):233–42.

29. Søndergaard L. Homology between the mammalian liver and the Drosophila fat body. Trends Genet. 1993 Jun;9(6):193–193.

30. Xu H, Barnes GT, Yang Q, Tan G, Yang D, Chou CJ, et al. Chronic inflammation in fat plays a crucial role in the development of obesity-related insulin resistance. J Clin Invest. 2003 Dec 15;112(12):1821–30.

31. Brugman S, Klatter FA, Visser JTJ, Wildeboer-Veloo ACM, Harmsen HJM, Rozing J, et al. Antibiotic treatment partially protects against type 1 diabetes in the Bio-Breeding diabetes-prone rat. Is the gut flora involved in the development of type 1 diabetes? Diabetologia. 2006 Sep;49(9):2105–8.

32. Cani PD, Amar J, Iglesias MA, Poggi M, Knauf C, Bastelica D, et al. Metabolic Endotoxemia Initiates Obesity and Insulin Resistance. Diabetes. 2007 Jul 1;56(7):1761–72.

33. Neal MD, Leaphart C, Levy R, Prince J, Billiar TR, Watkins S, et al. Enterocyte TLR4 Mediates Phagocytosis and Translocation of Bacteria Across the Intestinal Barrier. J Immunol. 2006 Mar 1;176(5):3070–9.

34. Vreugdenhil ACE, Rousseau CH, Hartung T, Greve JWM, van ‘t Veer C, Buurman WA. Lipopolysaccharide (LPS)-Binding Protein Mediates LPS Detoxification by Chylomicrons. J Immunol. 2003 Feb 1;170(3):1399–405.

35. Poggi M, Bastelica D, Gual P, Iglesias MA, Gremeaux T, Knauf C, et al. C3H/HeJ mice carrying a toll-like receptor 4 mutation are protected against the development of insulin resistance in white adipose tissue in response to a high-fat diet. Diabetologia. 2007 May 3;50(6):1267–76.

36. Cani PD, Bibiloni R, Knauf C, Waget A, Neyrinck AM, Delzenne NM, et al. Changes in Gut Microbiota Control Metabolic Endotoxemia-Induced Inflammation in High-Fat Diet-Induced Obesity and Diabetes in Mice. Diabetes. 2008 Jun 1;57(6):1470–81.

37. Dandona P, Ghanim H, Bandyopadhyay A, Korzeniewski K, Ling Sia C, Dhindsa S, et al. Insulin Suppresses Endotoxin-Induced Oxidative, Nitrosative, and Inflammatory Stress in Humans. Diabetes Care. 2010 Nov 1;33(11):2416–23.

38. Anhê FF, Jensen BAH, Perazza LR, Tchernof A, Schertzer JD, Marette A. Bacterial Postbiotics as Promising Tools to Mitigate Cardiometabolic Diseases. J Lipid Atheroscler. 2021;10(2):123.

39. Clarke TB, Davis KM, Lysenko ES, Zhou AY, Yu Y, Weiser JN. Recognition of peptidoglycan from the microbiota by Nod1 enhances systemic innate immunity. Nat Med. 2010 Feb;16(2):228–31.

40. Schertzer JD, Tamrakar AK, Magalhaes JG, Pereira S, Bilan PJ, Fullerton MD, et al. NOD1 Activators Link Innate Immunity to Insulin Resistance. Diabetes. 2011 Sep 1;60(9):2206–15.

41. Denou E, Lolmède K, Garidou L, Pomie C, Chabo C, Lau TC, et al. Defective NOD 2 peptidoglycan sensing promotes diet-induced inflammation, dysbiosis, and insulin resistance. EMBO Mol Med. 2015 Mar;7(3):259–74.

42. Cani PD, Plovier H, Van Hul M, Geurts L, Delzenne NM, Druart C, et al. Endocannabinoids — at the crossroads between the gut microbiota and host metabolism. Nat Rev Endocrinol. 2016 Mar;12(3):133–43.

43. Veilleux A, Di Marzo V, Silvestri C. The Expanded Endocannabinoid System/Endocannabinoidome as a Potential Target for Treating Diabetes Mellitus. Curr Diab Rep. 2019 Nov;19(11):117.

44. Di Marzo V, Matias I. Endocannabinoid control of food intake and energy balance. Nat Neurosci. 2005 May;8(5):585–9.

45. Ryberg E, Larsson N, Sjögren S, Hjorth S, Hermansson N-O, Leonova J, et al. The orphan receptor GPR55 is a novel cannabinoid receptor: GPR55, a novel cannabinoid receptor. Br J Pharmacol. 2007 Dec;152(7):1092–101.

46. De Petrocellis L, Di Marzo V. Non-CB1, Non-CB2 Receptors for Endocannabinoids, Plant Cannabinoids, and Synthetic Cannabimimetics: Focus on G-protein-coupled Receptors and Transient Receptor Potential Channels. J Neuroimmune Pharmacol. 2010 Mar;5(1):103–21.

47. Rousseaux C, Thuru X, Gelot A, Barnich N, Neut C, Dubuquoy L, et al. Lactobacillus acidophilus modulates intestinal pain and induces opioid and cannabinoid receptors. Nat Med. 2007 Jan;13(1):35–7.

48. Muccioli GG, Naslain D, Bäckhed F, Reigstad CS, Lambert DM, Delzenne NM, et al. The endocannabinoid system links gut microbiota to adipogenesis. Mol Syst Biol. 2010 Jan;6(1):392.

49. Pacher P, Kunos G. Modulating the endocannabinoid system in human health and disease - successes and failures. FEBS J. 2013 May;280(9):1918–43.

50. Cohen LJ, Esterhazy D, Kim S-H, Lemetre C, Aguilar RR, Gordon EA, et al. Commensal bacteria make GPCR ligands that mimic human signalling molecules. Nature. 2017 Sep 7;549(7670):48–53.

51. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci. 1977 Dec 1;74(12):5463–7.

52. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci. 1977 Feb 1;74(2):560–4.

53. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016 Jan;107(1):1–8.

54. Liao X, Li M, Zou Y, Wu F-X, Yi-Pan, Wang J. Current challenges and solutions of de novo assembly. Quant Biol. 2019 Jun;7(2):90–109.

55. Wooley JC, Godzik A, Friedberg I. A Primer on Metagenomics. Bourne PE, editor. PLoS Comput Biol. 2010 Feb 26;6(2):e1000667.

56. Handelsman J. Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiol Mol Biol Rev. 2004 Dec;68(4):669–85.

57. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol. 1998 Oct;5(10):R245–9.

58. Fox G, Stackebrandt E, Hespell R, Gibson J, Maniloff J, Dyer T, et al. The phylogeny of prokaryotes. Science. 1980 Jul 25;209(4455):457–63.

59. Clarridge JE. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. Clin Microbiol Rev. 2004 Oct;17(4):840–62.

60. MetaHIT Consortium, Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010 Mar;464(7285):59–65.

61. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci. 1985 Oct 1;82(20):6955–9.

62. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl Environ Microbiol. 2009 Dec;75(23):7537–41.

63. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010 May;7(5):335–6.

64. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010 Oct 1;26(19):2460–1.

65. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016 Jul;13(7):581–3.

66. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2012 Nov 27;41(D1):D590–6.

67. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol. 2006 Jul;72(7):5069–72.

68. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol. 2007 Aug 15;73(16):5261–7.

69. Escobar-Zepeda A, Godoy-Lozano EE, Raggi L, Segovia L, Merino E, Gutiérrez-Rios RM, et al. Analysis of sequencing strategies and tools for taxonomic annotation: Defining standards for progressive metagenomics. Sci Rep. 2018 Dec;8(1):12034.

70. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46.

71. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife. 2021 May 4;10:e65088.

72. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, et al. Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. Eisen JA, editor. PLoS Comput Biol. 2012 Jun 13;8(6):e1002358.

73. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. Valencia A, editor. PLoS Comput Biol. 2009 Dec 11;5(12):e1000605.

74. MetaHIT Consortium, Li J, Jia H, Cai X, Zhong H, Feng Q, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol. 2014 Aug;32(8):834–41.

75. Thomas AM, Segata N. Multiple levels of the unknown in microbiome research. BMC Biol. 2019 Dec;17(1):48.

76. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010 Dec;11(1):119.

77. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D412–9.

78. Müller AC, Guido S. Introduction to Machine Learning with Python. O'Reilly. 2017. 380 p.

79. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995 Sep;20(3):273–97.

80. Beck D, Foster JA. Machine Learning Techniques Accurately Classify Microbial Communities by Bacterial Vaginosis Characteristics. White BA, editor. PLoS ONE. 2014 Feb 3;9(2):e87830.

81. Wu H, Cai L, Li D, Wang X, Zhao S, Zou F, et al. Metagenomics Biomarkers Selected for Prediction of Three Different Diseases in Chinese Population. BioMed Res Int. 2018;2018:1–7.

82. Marchand M, Shawe-Taylor J. The Set Covering Machine. Journal of Machine Learning Research. 2002 Dec;723–46.

83. Drouin A, Giguère S, Déraspe M, Marchand M, Tyers M, Loo VG, et al. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. BMC Genomics. 2016 Dec;17(1):754.

84. Drouin A, Letarte G, Raymond F, Marchand M, Corbeil J, Laviolette F. Interpretable genotype-to-phenotype classifiers with performance guarantees. Sci Rep. 2019 Dec;9(1):4071.