Reconstitution de pan-génomes microbiens par séquençage métagénomique aléatoire : Application à l'étude du microbiote intestinal humain

Florian Plaza Onate

université PARIS-SACLAY

NNT : 2018SACLV068

Thèse de doctorat

UNIVERSITÉ DE VERSAILLES ST-QUENTIN-EN-YVELINES

université PARIS-SACLAY

Reconstitution de pan-génomes microbiens par séquençage métagénomique aléatoire

Application à l'étude du microbiote intestinal humain

Thèse de doctorat de l'Université Paris-Saclay préparée à l'Université de Versailles Saint-Quentin en Yvelines

ECOLE DOCTORALE N°577 - STRUCTURE ET DYNAMIQUE DES SYSTEMES VIVANTS (SDSV)
SPECIALITE DE DOCTORAT : SCIENCES DE LA VIE ET DE LA SANTE

Thèse présentée et soutenue à Gif-Sur-Yvette, le 10 Décembre 2018 par

Florian Plaza Oñate

Sous la supervision de Stanislav Dusko Ehrlich et Frédéric Magoulès

Composition du Jury :

Claudine Médigue Directeur de Recherche, CEA - CNRS	Présidente
Didier Debroas Professeur, Université Clermont Auvergne – CNRS	Rapporteur
Pierre Peterlongo Chargé de Recherche, INRIA Rennes Bretagne Atlantique	Rapporteur
Rayan Chikhi Chargé de Recherche, Université de Lille 1 - CNRS	Examinateur
Stanislav Dusko Ehrlich Directeur de recherche, INRA	Directeur de thèse
Frédéric Magoulès Professeur, Ecole CentraleSupelec	Co-Directeur de thèse

Table des matières

Remerciements

Cette thèse est le fruit d'une collaboration entre l'unité MetaGenoPolis (INRA) et de la société Enterome dans le cadre du dispositif CIFRE 2014/0057. Elle a été financée par l'Association Nationale de la Recherche et de la Technologie (ANRT), Enterome et MetaGenoPolis grâce au programme d'Investissements d'Avenir.

Je tiens dans un premier temps à remercier mon directeur de thèse, Stanislav Dusko Ehrlich pour son accueil au sein de l'unité MetaGenoPolis, pour ses conseils avisés et pour avoir stimulé ma curiosité scientifique. Je remercie également mon co-directeur de thèse Frédéric Magoulès pour m'avoir motivé et poussé à toujours aller de l'avant malgré les difficultés.

Je tiens également à remercier toute la société Enterome et son directeur Pierre Bélichard pour m'avoir accueilli durant cette thèse et m'avoir pleinement considéré comme un des leurs malgré ma présence quelques jours par semaine. Je remercie toute l'équipe « Biomarkers Discovery » et plus particulièrement Alessandra Cervino, Jonathan Plassais et William Farrin qui m'ont donné l'opportunité de valoriser mon travail dans le cadre de l'étude Baria. Je remercie aussi Rachel Morra pour être régulièrement venue aux nouvelles, y compris dans les moments de creux.

Je remercie tout particulièrement mon encadrant Matthieu Pichaud pour m'avoir suivi tout le long de cette thèse malgré l'Océan Atlantique qui nous séparait. Merci pour ton soutien, ton implication dans le projet et ta dose de bonne humeur quasi-quotidienne. Cette thèse ne serait pas ce qu'elle est sans toi.

Je remercie chaleureusement Florence Haimet qui a cru en moi et m'a permis de terminer cette thèse dans les meilleures conditions possibles.

Je remercie aussi Emmanuelle Le Chatelier pour le temps et l'énergie dépensés à améliorer et à valoriser mon travail. Tu as été une des premières à voir tout le potentiel de MSPminer et avoir fait sa promotion dans l'équipe.

Je remercie la présidente du jury, Claudine Médigue ; Pierre Peterlongo et Didier Debroas, rapporteurs de cette thèse ; ainsi que Rayan Chikhi, examinateur et membre de mon comité de suivi.

Merci à Hanna Julienne et Florence Thirion pour leur travail sur l'analyse fonctionnelle des MSPs. Ces résultats m'ont permis de valoriser grandement mes travaux de thèse.

Je remercie vivement Nicolas Pons, Anne-Sophie Alvarez, Mathieu Almeida et Victoria Mesliers pour la relecture attentive de ce manuscrit.

Je n'oublie pas toute l'équipe InfoBioStats de MetaGenoPolis (plantes comprises) pour tous ces bons moments passés ensemble que je n'oublierais jamais.

Un grand merci à ma famille, mes parents, mon frère et tout particulièrement ma compagne Mélanie pour sa patience et son soutien inconditionnel tout le long de cette thèse.

Enfin, je dédie cette thèse à mon fils Isaac qui a pointé le bout de son nez quelques jours avant que j'achève la rédaction de ce manuscrit. J'espère que tu seras fier de ton père

1. Introduction

1.1 Le microbiote intestinal humain

1.1.1 Structure, diversité et variabilité

Le tractus gastro-intestinal humain abrite une communauté microbienne complexe appelée microbiote. Chez un individu en bonne santé, la masse totale du microbiote intestinal est d'environ 200g. Il représente un nombre de cellules de l'ordre de 10^{13} équivalent à celui du nombre de cellules humaines [1]. Le microbiote intestinal est composé d'une grande variété de microorganismes. Les bactéries y sont les plus abondantes [2] bien que des archées [3], protozoaires, champignons [4,5] et virus [6] soient aussi présents mais moins étudiés jusqu'ici. En 2018, plus d'un millier d'espèces procaryotes colonisant le tractus gastro-intestinal humain ont été recensées [7].

1.1.1.1 *Variabilité temporelle*

Microbiote des enfants

A la naissance, l'intestin humain est quasiment stérile. Les enfants naissant par voie basse sont rapidement colonisés par le microbiote vaginal et intestinal de la mère tandis que ceux nés par césarienne le sont par des microorganismes associés à la peau de la mère ou au milieu hospitalier [8]. Par la suite, la colonisation est influencée par des facteurs comme la prématurité, le régime alimentaire de l'enfant (lait en poudre ou lait maternel), l'hygiène, la prise d'antibiotiques et, en cas d'allaitement au sein, par la diète et l'état de santé de la mère [8]. Le microbiote intestinal des nouveaux nés est immature. Il est caractérisé par une faible diversité microbienne et une abondance élevée des phyla *Proteobacteria* et *Actinobacteria* généralement sous-dominants chez les adultes sains. Suite à l'introduction de nourriture solide et au sevrage, le microbiote gagne en diversité et l'abondance des phyla *Firmicutes* et *Bacteroidetes* croit progressivement jusqu'à ce qu'ils deviennent majoritaires. On considère que le microbiote intestinal des enfants est comparable à celui d'un adulte à l'âge de 3 ou 4 ans.

Microbiote des adultes

Le microbiote intestinal des adultes occidentaux est dominé par les phyla *Firmicutes*, *Bacteroidetes* et *Actinobacteria* avec une abondance relative cumulée de 90% [2]. On dénombre plusieurs dizaines d'autres phyla minoritaires dont les plus abondants sont *Verrucomicrobia*, *Euryarchaeota* et *Proteobacteria.* Le microbiote intestinal des adultes en bonne santé est stable dans le temps et résilient même si cet équilibre peut être perturbé par des changements alimentaires [9], les voyages [10], la prise de médicaments [11] et l'utilisation d'antibiotiques [12].

Microbiote et vieillissement

Comparé aux jeunes adultes, le microbiote intestinal des personnes âgées est caractérisé par une faible diversité microbienne et par une baisse de l'abondance du phylum *Firmicutes* couplée à une hausse des *Bacteroidetes* [13]. De plus, ce changement de composition est marqué par une augmentation des Enterobactéries et à un risque accru d'infection par des pathogènes comme *Clostridioides difficile*. On note aussi une baisse significative de l'espèce *Faecalibacterium prausnitzii* productrice de molécules anti-inflammatoires. Finalement, ces altérations créent un environnement pro-inflammatoire qui affaiblit le système immunitaire favorisant l'apparition de maladies intestinales.

1.1.1.2 *Variabilité spatiale*

Variabilité transversale

Le tractus digestif est constitué d'une succession d'organes ayant des caractéristiques physicochimiques qui leur sont spécifiques (pH, taux de dioxygène, motilité). La composition et la

densité cellulaire du microbiote varient suivant la région du tractus digestif considérée [1] : on parle de variabilité transversale.

La concentration microbienne est faible dans l'estomac (10^3 cellules/ml) principalement à cause de l'acidité du milieu (pH 2) résultant de la présence de sucs gastriques. Seules quelques espèces acidophiles comme *Helicobacter pylor*i peuvent s'y développer. La concentration microbienne croit ensuite progressivement dans l'intestin grêle et passe de 10^4 cellules/ml dans le duodénum à 10^8 cellules/ml dans l'iléon mais reste faible car le mucus qui le tapisse est riche en composés antimicrobiens et les nutriments y transitent rapidement. Finalement, elle atteint un maximum dans le colon (10^{11} cellules/ml) car une grande partie de l'eau y est absorbée et les résidus de nourriture y restent longtemps (jusqu'à 30 heures) ce qui permet aux microorganismes de se développer. De plus, la couche externe du mucus facilite l'établissement d'une communauté microbienne résidente qui adhère et dégrade la mucine. Ainsi, le colon est le principal contributeur à la population totale du microbiote intestinal (> 99%).

Variabilité horizontale

A cette variabilité transversale s'ajoute une variabilité horizontale. Les microorganismes colonisent essentiellement la lumière du colon et la surface du mucus (10^{11} cellules/ml). La concentration microbienne décroît dans le mucus d'autant plus que l'on se rapproche de l'épithélium. Ainsi elle passe de 10^6 cellules/ml dans la couche de mucus externe à 10^5 dans la couche interne. Une communauté microbienne spécifique colonise le mucus car, en comparaison avec la lumière intestinale, le milieu est plus visqueux, beaucoup plus pauvre en dioxygène et riche en mucine [14].

1.1.1.3 *Variabilité inter-individus*

Variabilité des espèces

On note une grande variabilité inter-individus des espèces constituant le microbiote intestinal ainsi que de leurs abondances respectives. En effet, seulement quelques dizaines d'espèces sont détectées chez plus de 90% des individus. L'abondance relative de ces dernières peut varier d'un facteur 1000 d'un individu à l'autre [15]. Malgré une forte variabilité inter-individus, trois grands types de compositions microbiennes appelés entérotypes dominés respectivement par les genres *Bacteroides*, *Prevotella* et *Ruminococcus* ont été mis en évidence [16].

La composition du microbiote intestinal varie fortement suivant l'origine géographique et l'ethnie des individus. Bien qu'une association entre le microbiote intestinal et patrimoine génétique ait été montrée en analysant des jumeaux monozygotes et dizygotes, l'environnement est le principal facteur expliquant ces différences [17]. En effet, le régime alimentaire et plus généralement le mode de vie [18] façonnent le microbiote intestinal. Chez les chasseurs cueilleurs, une alimentation riche en fibres (tubercules, baies), en protéines (poisson, gibier), en acides gras de qualité (noix) et pauvre en glucides explique la très forte diversité de leur microbiote. L'abondance élevée des taxons *Prevotella* ou *Treponema* facilite la digestion et l'extraction de nutriments provenant de végétaux fibreux. On détecte généralement la présence de microorganismes pathogènes et d'autres parasites car ces populations n'ont accès ni à l'eau potable ni à la médecine moderne. Dans les communautés paysannes rurales, la diversité du microbiote décroît ainsi que l'abondance du genre *Prevotella*, car l'alimentation est moins variée. Cette baisse de diversité est encore plus marquée dans les sociétés urbaines du fait d'une alimentation pauvre en fibres et riche en glucides facilement digérables combinée à une hygiène renforcée et à l'usage de médicaments et d'antibiotiques [18].

Variabilité des souches

L'analyse de cohortes de grande taille a révélé que la composition du microbiote intestinal varie significativement d'un individu à l'autre car ces derniers sont non seulement porteurs d'espèces

différentes mais aussi de souches différentes de la même espèce. Ces dernières se distinguent les unes des autres par les gènes qui les composent [19], par un polymorphisme nucléotidique [20] et par la variabilité du nombre de nombre de copies de certains gènes [21].

Ces différences expliquent que de nombreux traits phénotypiques microbiens ne sont pas spécifiques d'une espèce mais de seulement certaines souches. Par exemple, seules deux des trois sous-espèces identifiées de *Eubacterium rectale* possèdent un opéron codant pour un flagelle ayant une activité pro inflammatoire [22]. De même, une étude a révélé que la présence de certaines souches de *Prevotella copri* est associée au développement de la polyarthrite rhumatoïde, maladie inflammatoire qui touche les articulations, tandis que d'autres souches sont abondantes chez des individus en bonne santé consommant une grande quantité de fibres [23].

1.1.1.4 *Le pan-génome microbien*

Le séquençage de plusieurs isolats d'*Escherichia coli,* qui pour des raisons historiques est l'espèce microbienne la plus étudiée, a mis pour la première fois en évidence une grande variabilité du contenu en gènes au sein de souches d'une même espèce. En effet, parmi les 5 000 gènes composant en moyenne un génome d'*E. coli*, seuls 2 000 (40%) sont présents chez toutes les souches tandis que le répertoire total de l'espèce est composé de plusieurs dizaines de milliers de gènes [24].

De cette observation est née la notion de pan-génome défini comme le répertoire de gènes d'une espèce. Dans un pan-génome, on distingue les gènes core présents chez toutes les souches de l'espèce et les gènes accessoires (parfois appelés gènes optionnels ou dispensables) présents chez seulement certaines souches [25].

Par la suite, le développement des techniques de séquençage a permis la caractérisation du pan-génome d'autres espèces. Ces études ont révélé l'existence d'espèces au pan-génome dit « ouverts » et d'espèces au pan-génome « fermé ». La taille d'un pan-génome ouvert augmente sans cesse avec l'ajout des gènes provenant de souches nouvellement séquencées tandis que la taille des pan-génomes fermés atteint un plateau après le séquençage de seulement deux ou trois souches [26].

Les espèces au pan-génome ouvert comme *E. coli* ont un contenu en gènes très variable. Les gènes accessoires qu'elles portent leur fournissent des avantages sélectifs leur permettant de coloniser diverses niches écologiques, de faire face à différentes sources de stress (antibiotiques, attaque phagique, compétition avec d'autres espèces) et d'adapter leur métabolisme pour tirer parti de plusieurs sources d'énergie [27]. Elles disposent de mécanismes de transfert horizontal permettant l'acquisition de matériel génétique provenant d'autres microorganismes et accroissent ainsi constamment la taille de leur répertoire de gènes. Les espèces au pan-génome fermé comme *Campylobacter jejuni* ont au contraire un contenu en gènes très stable. Elles sont confinées à une seule niche écologique par l'absence de mécanismes efficaces permettant d'acquérir du matériel génétique dans leur environnement.

Enfin, les études de génomique des populations comparant plusieurs souches de la même espèce [28,29] montrent que la prévalence des gènes d'un pan-génome ouvert suit une distribution bimodale en U le plus souvent asymétrique à gauche (voir **Figure 38**, page 72). Ainsi, Koonin et Wolf [30] proposent de classifier les gènes d'un pan-génome dans non plus deux (core et accessoire) mais quatre catégories):

1. Les gènes (hard)-core sont présents dans tous les génomes de toutes les souches
2. Les gènes soft-core sont quasiment présents dans toutes les souches (≥ 95%). Il s'agit vraisemblablement de gènes core absents dans quelques génomes incomplets.
3. Les gènes shell sont des gènes accessoires de prévalence intermédiaire.

4. Les gènes cloud sont des gènes accessoires rares présents dans seulement quelques souches voire une seule.

Le fait que les gènes cloud soient les plus nombreux suggère que les espèces acquièrent en permanence de nouveaux gènes par transfert horizontal et dans une moindre mesure par duplication [31] mais que seule une petite proportion d'entre eux apportent un avantage sélectif suffisant pour être fixés dans la population.

1.1.2 Le microbiote intestinal et la santé humaine

Le microbiote intestinal joue un rôle clef dans la santé humaine et à ce titre, il est considéré comme un organe à part entière [32].

1.1.2.1 *Rôle du microbiote intestinal*

Les microorganismes colonisant le côlon utilisent comme source d'énergie les protéines et glucides complexes qui n'ont pas été absorbés par l'hôte au niveau de l'intestin grêle [33]. Cette fermentation produit des gaz comme le dioxyde de carbone, le sulfure d'hydrogène, le méthane et le dihydrogène ainsi que des acides gras à chaîne courte (AGCC) tels que le propionate, l'acétate et le butyrate [34].

Les AGCC jouent un rôle clé dans la santé humaine car ils constituent une source d'énergie importante pour notre organisme. Le propionate, décrit comme inhibiteur de la liponéogénèse et la synthèse de cholestérol, se révèle être un métabolite clé dans la prévention de l'obésité et du diabète [35]. Le butyrate est un autre métabolite très étudié présentant un effet bénéfique sur l'hôte. Il est décrit comme un métabolite particulièrement important dans la prévention du cancer colorectal, des maladies inflammatoires intestinales, de par ses propriétés anti-inflammatoires et sa participation au maintien de l'intégrité de la barrière intestinale [36]. Principal nutriment des cellules de la muqueuse colique (colonocytes), il favorise ainsi leur bonne différenciation et prolifération cellulaire. Le microbiote intestinal permet donc à l'hôte d'accroitre la quantité d'énergie provenant de l'alimentation et contribue à la bonne santé de la muqueuse.

En outre, le microbiote intestinal synthétise certaines vitamines du groupe K et du groupe B et notamment la vitamine B12 indispensable au bon fonctionnement de l'organisme de l'hôte. Ainsi, il contribue à plus de 25% des apports journaliers pour quatre des huit vitamines B [37].

Enfin, le microbiote intestinal contribue à la protection de l'hôte contre les bactéries pathogènes en stimulant le développement du système immunitaire, en assurant un effet de barrière empêchant la colonisation de la muqueuse et en renforçant l'imperméabilité de la barrière épithéliale [38].

1.1.2.2 *Maladies et microbiote intestinal*

La dysbiose, en opposition avec l'eubiose, est un terme désignant un microbiote altéré dont la composition est déséquilibrée. La dysbiose du microbiote intestinal se caractérise par une diminution de la richesse et la diversité microbienne, une baisse de l'abondance du phylum *Firmicutes* au profit des *Bacteroidetes*, et une augmentation de l'abondance d'espèces potentiellement pathogènes. Du point de vue clinique, un microbiote dysbiotique entraine une hyper perméabilité de la barrière intestinale, un état inflammatoire et un stress oxydatif [39].

De nombreuses maladies ont été associées à une dysbiose intestinale ou plus généralement à un changement compositionnel du microbiote par rapport à des individus sains (**Tableau 1**). Par exemple, le microbiote intestinal des individus cirrhotiques est enrichi en bactéries orales appartenant aux genres *Veillonella*, *Streptococcus* et *Haemophilus* [40]. Chez les individus atteints d'une maladie inflammatoire chronique de l'intestin, on note un appauvrissement des genres bactériens ayant un

effet protecteur pour la muqueuse intestinale dont *Faecalibacterium* et *Roseburia* ainsi qu'une hausse des espèces pro-inflammatoires comme *Ruminococcus gnavus* et diverses Enterobactéries.

Type de maladie	Maladie
Maladies métaboliques	Obésité [41]
	Diabète de type 2 [42]
	Stéatohépatite non-alcoolique [43] et cirrhose [40]
	Athérosclérose [44]
Maladies immunitaires	Maladie inflammatoires chroniques de l'intestin [45] (rectocolite hémorragique et maladie de Crohn)
	Spondylarthrite ankylosante [46]
	Diabète de type 1 [47]
	Maladie cœliaque [48]
	Allergies et asthme [49]
Maladies neurodégénératives	Maladie de Parkinson [50]
	Maladie d'Alzheimer [51]
Troubles de l'humeur et du comportement	Anxiété [52]
	Dépression [52]
	Autisme [53]
Autres maladies	Cancer colorectal [54]
	Syndrome de l'intestin irritable [55]

***Tableau 1** : Liste de maladies associées à un changement compositionnel du microbiote intestinal*

1.1.2.3 *Vers une médecine stratifiée et personnalisée*

Modulation du microbiote

Toutes ces découvertes ouvrent la voie au développement de nouvelles stratégies thérapeutiques visant à moduler la composition microbiote pour traiter des maladies ou en atténuer les symptômes.

Les probiotiques sont définis par la FAO et l'OMS comme des « micro-organismes vivants qui, lorsqu'ils sont ingérés en quantité suffisante, exercent des effets positifs sur la santé, au-delà des effets nutritionnels traditionnels » [56]. Aujourd'hui, de nombreux probiotiques sont disponibles à la vente et certains ont même le statut de médicaments. Par exemple, Enterogermina® constitué d'un cocktail de souches de l'espèce *Bacillus clausii* ou l'Ultralevure® *Saccharomyces boulardii* sont indiqués dans le traitement de la diarrhée. Le marché des probiotiques est en pleine croissance et par conséquent de nombreuses entreprises investissent dans ce domaine. Par exemple, la société Nextbiotix développe un probiotique constitué d'une souche de l'espèce *Faecalibacterium prausnitzii* allongeant les périodes de rémission chez les patients atteints de maladies inflammatoires chroniques de l'intestin. Seres Therapeutics et Vedenta développent des cocktails de souches (SER-10 et VE303) ramenant le microbiote à un état sain suite à une infection par *Clostridioides difficile*.

L'identification des molécules actives produites par des souches qui engendrent une amélioration de l'état de santé ouvre la voie au développement de nouveaux médicaments. Plutôt que d'ingérer des probiotiques, on propose d'administrer directement leurs molécules actives produites *in vitro*. Ainsi, Enterome développe EB110, un médicament composé d'un métabolite sécrété par une bactérie qui ralentit ou bloque la progression de la maladie de Crohn grâce à un effet anti-inflammatoire.

Il est aussi envisagé de moduler la composition du microbiote intestinal en développant des composés antimicrobiens spécifiques de certaines espèces. EB8018 développé par Enterome est une molécule destinée aux patients atteints la maladie de Crohn qui inhibe l'adhésine FimH produite par des souches d'*Escherichia coli* à effet pro-inflammatoire (AIEC) et les empêche ainsi de coloniser la muqueuse iléale.

De même, la société C3J Therapeutics développe de phages issus de la biologie de synthèse détruisant spécifiquement certains pathogènes.

Une autre technique prometteuse pour moduler le microbiote intestinal est la transplantation fécale. Elle consiste à administrer de la matière fécale prélevée chez un sujet sain et à l'administrer à un patient sujet à une dysbiose intestinale. A ce jour, l'Agence nationale de sécurité du médicament (ANSM) n'autorise cette pratique que pour traiter des infections à *Clostridioides difficile*. La société MaaT Pharma développe quant à elle une technique d'autotransplantation où on administre à un patient sa propre matière fécale prélevée avant un traitement ayant un effet délétère sur le microbiote comme la chimiothérapie ou l'antibiothérapie.

Diagnostic, diagnostic compagnon et outils de monitoring

Après avoir mis en évidence une association entre la composition du microbiote et une maladie, il est possible de développer un test pour la diagnostiquer et suivre son évolution de façon non invasive par prélèvement de selles. Ainsi, l'INRA a développé en collaboration avec des équipes chinoises un test breveté fiable à plus de 90% pour diagnostiquer la cirrhose du foie en quantifiant seulement 7 espèces bactériennes. La société Enterome développe quant à elle un outil de monitoring (IBD 110) permettant le suivi de l'évolution de la maladie de Crohn en déterminant le niveau d'altération de la muqueuse à partir de la composition du microbiote et de la quantification de certains métabolites.

Enfin, le diagnostic compagnon est un test qui détermine si un patient répondra ou non à la d'administration d'un médicament. Il est particulièrement utile lorsque le traitement est cher ou a de nombreux effets indésirables. Dans de nombreux cas, le développement d'un diagnostic compagnon basé sur la composition du microbiote intestinal permettrait la stratification des patients en groupes de répondeurs et de non répondeurs. Par exemple, il a récemment été montré que la composition du microbiote d'un patient influence l'efficacité d'un traitement anti-cancéreux par immunothérapie [57]. De même, il a été montré que certains patients sont porteurs d'une sous-espèce d'*Eggerthella lenta* qui produit une molécule rendant inefficace la digoxine [58], un médicament utilisé pour traiter des maladies cardiaques.

1.2 Caractérisation du microbiote intestinal

La compréhension du lien entre le microbiote et la santé humaine nécessite d'identifier, quantifier et caractériser les différents microorganismes colonisant l'intestin et de comprendre la façon dont ils interagissent avec l'hôte. Nous présenterons ici les outils les plus utilisés à ce jour pour étudier le microbiote.

1.2.1 Prélèvement d'échantillons

La caractérisation du microbiote intestinal nécessite le prélèvement d'échantillons par biopsie ou par collecte de selles.

Les biopsies permettent l'étude du microbiote *in situ*. Elles ont par exemple mis en évidence la variabilité spatiale du microbiote ou la présence de pathogènes difficiles à détecter dans les fèces [59]. Bien qu'elle permette d'obtenir des informations précises sur le microbiote dans un contexte spatial donné, cette méthode requiert une intervention anxiogène, inconfortable et potentiellement risquée pour le patient. Elle est donc principalement employée pour des raisons médicales.

Un échantillon de selles peut être récolté lors de la défécation sans intervention ni présence médicale. Ainsi, l'auto-prélèvement ne présente pas les inconvénients de la biopsie pour le patient. De plus, il est peu coûteux car il requiert des moyens matériels limités. A ce jour, il s'agit du mode de prélèvement le plus utilisé. Cependant, la selle agrège en quelque sorte un historique des différentes étapes de la digestion mais ne permet pas une véritable étude *in situ* comme la biopsie.

1.2.2 Caractérisation par culture microbienne

La culture microbienne est une technique de laboratoire permettant la multiplication contrôlée de microorganismes sur un milieu déterminé. Pour caractériser un écosystème, on cherche en général à obtenir des cultures pures c'est-à-dire des populations de cellules issues d'une seule cellule. Pour cela, on réalise des repiquages successifs sur des milieux sélectifs ou enrichis. Une fois isolés, les microorganismes sont caractérisés par microscopie, divers tests biochimiques, par caractérisation de gènes essentiels et dans l'idéal, par le séquençage de leur génome.

1.2.2.1 *Limites historiques de la culture microbienne*

En 2010, on estimait que seules 20% des espèces du microbiote intestinal humain avaient été isolées, cultivées et séquencées [60]. On considérait alors que les 80% restantes étaient pour leur majorité « non cultivables ».

La faible proportion d'espèces cultivées était due à plusieurs facteurs [61]. Tout d'abord, le substrat et les conditions de cultures étaient inadaptés par manque de connaissance des espèces ciblées. En effet, le microbiote intestinal est composé d'une majorité d'espèces anaérobies strictes très sensibles à l'oxygène. D'autres espèces dites microaérophiles ont besoin d'oxygène pour croître mais à un taux inférieur à celui de l'atmosphère. D'autres espèces en dormance ou sous forme de spores dans les selles étaient souvent manquées. Le temps d'incubation était trop court pour qu'elles puissent entrer en phase de croissance et que la colonie atteigne une taille suffisante pour être détectée. Finalement, certains microorganismes ne peuvent croître efficacement en étant isolés. Dans leur état naturel, certains se regroupent sous forme d'agrégats ou de biofilms. D'autres s'inscrivent dans des chaines alimentaires complexes impliquant des échanges de métabolites entre microorganismes.

1.2.2.2 *Progrès récents de la culture microbienne*

Récemment, les méthodes de culture microbiennes ont réalisé des progrès considérables. En quelques années, le nombre d'espèces recensées dans le microbiote intestinal humain a quasiment doublé [7]. Ces progrès sont majoritairement dus à l'utilisation de chambres anaérobies dites chambres de Freiter ainsi qu'à l'essai systématique de plusieurs conditions de cultures en variant les substrats (agar, gomme gellane, liquide ruminal, sang etc.), la température, l'acidité ainsi que le taux de dioxygène. Par ailleurs, diverses techniques ont été développées pour éliminer la population microbienne dominante et se focaliser sur les espèces peu abondantes. Par exemple, l'usage d'antibiotiques permet la sélection des espèces résistantes au détriment des espèces sensibles. De plus, l'ajout de bactériophages permet de lyser spécifiquement les cellules de certains microorganismes. La filtration par des membranes micrométriques (de 5 à 0.2 µm) ne conserve quant à elle que les espèces avec des cellules de petite taille. D'autres part, les espèces à croissance lente sont isolées grâce à des systèmes optiques détectant des colonies microscopiques. En outre, un système de criblage haut-débit basé sur de la spectroscopie de masse (technologie MALDI-TOF [62]) discrimine les espèces connues de celles inconnues en comparant le spectre obtenu à une base de données de référence. Ce processus de sélection permet de concentrer l'effort de séquençage sur les espèces non répertoriées. Enfin, la cytométrie en flux (FACS) facilite l'isolation d'espèces d'intérêt après marquage de leurs cellules avec des marqueurs fluorescents (technique FISH).

1.2.2.3 *Avantages et inconvénients de la culture microbienne*

La culture microbienne est nécessaire pour découvrir et valider expérimentalement les attributs fonctionnels des microorganismes et ainsi comprendre leur interaction avec l'hôte. Elle permet d'identifier des microorganismes faiblement abondants qui ne peuvent être détectés ou assemblés par séquençage métagénomique. Malgré les progrès notables réalisés ces dernières années, la culture microbienne reste un processus long et laborieux.

1.2.3 Caractérisation par séquençage d'amplicons

Le séquençage d'amplicons consiste à séquencer uniquement un groupe de gènes orthologues que l'on suppose universels dans un taxon donné et suffisamment variables pour être considérés comme marqueurs phylogénétiques. Cette méthode ciblée permet de recenser les taxons présents dans un échantillon, d'estimer leur abondance relative et d'établir leurs liens de parenté (phylogénie).

1.2.3.1 *Méthodologie*

Choix des gènes à séquencer

Pour étudier les eucaryotes intestinaux (champignons et protozoaires), on séquence le gène codant pour l'ARN ribosomique 18S ou les espaceurs internes transcrits (Internal Transcribed Spacer ou ITS). Lorsque l'on s'intéresse aux procaryotes (archées et bactéries), le choix se porte sur le gène codant pour l'ARN ribosomique 16S. La majorité des études sur le microbiote intestinal humain se concentrent sur les espèces procaryotes car elles sont de loin les plus abondantes. Par conséquent, on ne décrira par la suite que le séquençage de l'ARNr 16S.

La longueur du gène codant pour l'ARNr 16S est d'environ 1500 paires de bases. Il est constitué de 9 régions variables numérotées de V1 à V9 dont la longueur varie de 30 à 200 paires de bases. Ces régions variables sont séparées par 9 régions conservées composées de séquences complémentaires qui après transcription s'apparient pour former une tige ; cette contrainte d'appariement expliquant le faible taux de mutation. Les régions variables forment quant à elles des boucles où les contraintes évolutives sont beaucoup plus faibles [63].

Amplification et séquençage

Après extraction de l'ADN, le gène codant pour l'ARN 16S est amplifié par PCR. Pour ce faire, on utilise une paire d'amorces universelles (de 15 à 25bp) visant des régions conservées du gène et encadrant plusieurs régions hypervariables. On adjoint à ces amorces des séquences nécessaires au séquençage (adaptateurs). Ensuite, on procède à 20 à 30 cycles d'amplification après quoi l'ADN qui ne provient pas du gène codant pour l'ARNr 16S a une abondance relative très faible.

Le gène de l'ARN 16S n'est en général pas séquencé en intégralité. En effet, la technologie Illumina MiSeq à ce jour la plus abordable et la plus couramment utilisée produit des lectures dont la taille n'excède pas 300 paires de bases. Néanmoins, on peut générer des lectures pairées qui se chevaucheront en sélectionnant une taille d'insert adéquate. Par la suite, les lectures pairées sont fusionnées avec logiciel dédié pour obtenir des lectures d'environ 500 paires de bases.

1.2.3.2 *Traitement bioinformatique*

Le traitement bioinformatique consiste à regrouper les séquences ayant un fort niveau d'identité en clusters appelés Operational Taxonomic Units (OTUs). Ces OTUs sont ensuite annotés en recherchant des séquences similaires dans les bases de données de gènes de l'ARN 16S dont les plus connues sont RDP [64], SILVA [65] et Greengenes [66]. On établit éventuellement les liens de parenté entre les différents OTUs en procédant à un alignement multiple à partir duquel on infère un arbre phylogénétique. Finalement, on aligne les lectures séquencées sur la banque d'OTUs générée. Ainsi, on estime l'abondance de chaque OTU dans chaque échantillon.

1.2.3.3 *Biais et limites du séquençage du gène de l'ARNr 16s*

Le séquençage du gène de l'ARNr 16S est sujet à plusieurs biais techniques ou biologiques.

Biais de PCR

Lors de l'étape d'amplification par PCR, des séquences chimériques composées de fragments de plusieurs gènes de l'ARN 16S sont produites [67]. Ceci entraine une surestimation de la richesse microbienne de l'échantillon étudié et des abondances d'OTUs incorrectes. De plus, les multiples cycles

de PCR favorisent l'amplification des séquences peu abondantes qui ont statistiquement moins de chances de se réapparier [68]. Ainsi, l'abondance relative des OTUs peu abondants sera surestimée.

Haut degré de conservation du gène

L'annotation taxonomique de séquences provenant du gène de l'ARN 16s est problématique lorsque celui-ci n'est que partiellement séquencé. Bien souvent, la séquence est similaire à celles de plusieurs espèces disponibles dans les bases de données [69] car les régions ciblées ne sont pas suffisamment discriminantes bien qu'elles soient qualifiées d'hypervariables. Par conséquent, une annotation au niveau espèce est souvent impossible et la plupart des études se contentent d'une résolution au niveau genre. Le séquençage de l'intégralité du gène grâce à des technologies comme PacBio permettent d'atteindre une meilleure résolution [70] en contrepartie d'un coût de séquençage plus élevé.

Variabilité du nombre de copies du gène

L'étude de 1 690 génomes bactériens a montré que le nombre de copies du gène de l'ARN 16S variait d'une espèce à l'autre [69]. La plupart des génomes avaient entre 1 et 7 copies, le maximum observé étant de 15. Remarquablement, seuls 15% des génomes en possédaient seulement une. Dans certains cas, une variabilité importante du nombre de copies a été observée au sein de souches d'une même espèce. Au final, l'abondance relative des espèces avec un nombre de copies élevé sera surestimée. Une normalisation par le nombre de copies permet de limiter ce biais [71]. Toutefois, cette information n'est pas toujours disponible car certaines espèces n'ont aucun génome séquencé.

Prédiction du potentiel fonctionnel

Après séquençage du gène codant pour l'ARNr 16S, l'outil piCRUST [72] infère le contenu fonctionnel d'un échantillon métagénomique à partir de la liste des OTUs détectées. En particulier, piCRUST prédit les attributs fonctionnels des microorganismes sans génome de référence disponible à partir des génomes de microorganismes apparentées. Toutefois, cette approche est limitée par le fait que deux espèces apparentées peuvent avoir un profil fonctionnel distinct ainsi que par l'existence de fonctions spécifiques à certaines souches [73].

1.2.4 Caractérisation par séquençage métagénomique shotgun

Le séquençage métagénomique shotgun, parfois appelé séquençage non ciblé, aléatoire, global ou complet, consiste à tirer aléatoirement des fragments d'ADN dans une communauté microbienne puis à les lire avec un séquenceur pour déterminer les nucléotides qui les composent. Contrairement au séquençage d'amplicons qui se focalise sur un seul gène marqueur, le séquençage shotgun cible l'ensemble du matériel génétique présent dans un échantillon sans *a priori*.

1.2.4.1 *Extraction de l'ADN*

La première étape consiste à extraire l'ADN contenu dans les cellules des microorganismes. Pour cela, les membranes cellulaires sont détruites (lyse) par des méthodes physiques (agitation de billes, choc thermique), enzymatiques (lysozyme) ou chimiques (utilisation de sels et dénaturants). Ensuite, différentes impuretés sont éliminées et l'ADN est séparé des autres composés cellulaires par centrifugation. Pour terminer, on évalue la quantité d'ADN extrait et on détermine la taille des fragments à des fins d'assurance qualité. Ces étapes sont cruciales pour le succès de l'étude. En effet, le choix des techniques employées influe sur la capacité à extraire des ADNs représentatifs de la composition microbienne de l'échantillon [74].

1.2.4.2 *Préparation la bibliothèque de séquençage*

L'ADN extrait est cassé en fragments d'une centaine de paires de bases par utilisation d'ultrasons (sonication). L'information sur le lien physique entre les fragments avant sonication est perdue. Ainsi, certains auteurs comparent un échantillon métagénomique à « une centaine de puzzles dont les pièces

ont été mélangées dans une seule boite » [75]; les puzzles étant les génomes des microorganismes et les pièces les fragments d'ADN. Pour permettre la fixation des fragments sur le séquenceur, on ajoute à leur extrémité des courtes séquences appelées adaptateurs. En cas de séquençage simultané de plusieurs échantillons (multiplexing), on adjoint d'autres courtes séquences nommées codes-barres (barcodes) pour identifier l'échantillon dont provient le fragment *a posteriori*. Cet ensemble de fragments liés à des courtes séquences est appelé bibliothèque, banque ou librairie de séquençage.

1.2.4.3 *Séquençage*

Finalement, des séquenceurs d'ADN dits à haut débit (HTS : High Throughput Sequencing) ou nouvelle génération (NGS : Next Generation Sequencing) déterminent la composition nucléotidique de quelques dizaines de millions de ces fragments tirés aléatoirement [76].

Pour ce faire, les fragments d'ADN à lire sont déposés dans des puits ou sur des lames. Ils sont ensuite dupliqués par PCR lors d'une étape appelée amplification monoclonale. L'amplification monoclonale permet de produire un signal suffisamment fort pour être détectable par la machine lors de la deuxième étape appelée séquençage par synthèse. Le séquençage par synthèse consiste à générer le brin complémentaire du fragment d'ADN dont on souhaite connaitre la séquence en déterminant la nature et l'ordre des nucléotides intégrés. A ce jour, les technologies Illumina et Ion Torrent dans une moindre mesure dominent le marché du séquençage shotgun. La technologie Illumina utilise quatre types de nucléotides modifiés qui émettent un signal lumineux spécifique lorsqu'ils sont incorporés au brin complémentaire. La technologie IonTorrent détecte quant à elle l'émission de protons lors de l'incorporation d'un nucléotide dans le brin d'ADN complémentaire. Les séquences obtenues à la fin du processus de séquençage sont appelées lectures ou plus communément reads. Les reads ainsi que des scores sur la qualité de chaque nucléotide les composant sont stockés dans un fichier texte au format FASTQ.

1.3 Analyse de données de séquençage métagénomique shotgun

1.3.1 Comparaison d'échantillons sans référence

Les méthodes sans référence permettent de comparer des échantillons métagénomiques sans avoir recours à des données externes comme des génomes ou des catalogues de gènes.

1.3.1.1 *Approches par comparaison des lectures*

Compareads [77] calcule la distance entre deux échantillons métagénomiques en estimant le nombre de lectures qu'ils partagent. Le programme repose sur une heuristique qui considère qu'une lecture est présente dans un échantillon si les mots de longueurs k (k-mers) la composant y sont trouvés. Cette stratégie sans alignement permet de traiter des millions de lectures 30 fois plus rapidement que BLASTn avec une sensibilité similaire. Compareads s'appuie une structure de données probabiliste basée sur les filtres de Bloom pour indexer des centaines de millions de k-mers avec seulement 4 Go de RAM et un faible taux d'erreur. COMMET [78] est une amélioration de Compareads qui indexe les lectures sur une structure de données conçue pour accélérer les comparaisons et limiter la taille des résultats stockés sur disque.

1.3.1.2 *Approches par comparaison des profils en k-mers*

Technique exhaustive

Plutôt que de compter le nombre de lectures partagées, Simka [79] estime la distance entre n échantillons métagénomiques à partir de leur profils d'abondance en k-mers. L'outil repose une stratégie de comptage des k-mers consommant peu de mémoire vive et ayant un fort potentiel de parallélisation [80]. La matrice de comptage des k-mers n'est pas calculée directement ce qui permet de limiter considérablement la quantité d'espace disque nécessaire. Au final, Simka est bien plus rapide et consomme moins de mémoire vive que les outils basés sur une comparaison de lectures.

Technique de réduction de dimensionnalité

L'outil Mash [81] calcule une signature (sketch) pour chaque échantillon métagénomique en utilisant la technique MinHash. La signature d'un échantillon est composée d'un sous-ensemble de quelques milliers de k-mers présents dans ses lectures. Les échantillons sont finalement comparés en calculant une distance de Jaccard entre leurs signatures. Mash surpasse de loin toutes les autres méthodes sans référence aussi bien sur le plan la vitesse de traitement que sur l'utilisation des ressources de calcul. Toutefois, cela se fait peut-être au détriment de la précision car les signatures en k-mers ne sont pas exhaustives et ne prennent pas en compte leur abondances relatives.

1.3.1.3 *Avantages et limites*

Les méthodes sans référence à l'exception des techniques de réduction de dimensionnalité présentent l'avantage de considérer toutes les lectures d'un échantillon. En comparaison, une grande partie du signal biologique peut être manqué en utilisant une méthode s'appuyant sur une référence qui n'est pas suffisamment exhaustive. Ainsi, les méthodes sans référence sont particulièrement utiles pour étudier des écosystèmes composés d'une majorité d'espèces inconnues. Elles implémentent des algorithmes à complexité linéaire qui en quelques heures comparent deux à deux des centaines d'échantillons métagénomiques et groupent ceux présentant des caractéristiques similaires. Dans une démarche d'assurance qualité, elles permettent de détecter rapidement des échantillons inversés, de mauvaise qualité ou contaminés [82]. Cependant, elles ne fournissent pas d'informations sur la composition taxonomique ni le potentiel fonctionnel des communautés microbiennes.

1.3.2 Profilage taxonomique

Le profilage taxonomique identifie les microorganismes présents dans un échantillon métagénomique (approche qualitative) et estime leurs abondances relatives (approche quantitative).

1.3.2.1 *Profilage taxonomique par classification exhaustive des lectures*

Cette première catégorie d'outils tente s'assigner chaque lecture d'un échantillon métagénomique à un taxon donné ; idéalement à une espèce.

Approches probabilistes

Phymm [83] s'appuie sur des modèles de Markov interpolés (IMMs) pour attribuer un label taxonomique aux lectures.

Dans des modèles de Markov d'ordre k, chaque nucléotide d'une séquence composant une séquence d'ADN a une loi de probabilité dépendante des k derniers nucléotides. Les IMMs sont une extension des modèles d'ordre fixe où la longueur des k-mers peut varier et où seules les positions les plus informatives sont prises en compte pour calculer la loi de probabilité conditionnelle.

Durant la phase d'entraînement, Phymm reconstruit pour chaque taxon (espèce, genre) une IMM représentative des motifs nucléotidiques qui lui sont spécifiques. Durant la phase de classification, l'outil analyse les nucléotides composant une lecture puis utilise les IMMs pour calculer la probabilité que cette lecture provienne d'un taxon donné.

Au final, Phymm fournit des résultats de médiocre qualité lorsqu'il traite des lectures d'une centaine de bases. De plus, il peut difficilement traiter de gros jeux de données car il classifie à peine une centaine de reads par minute sur des machines de calcul modernes.

Approches par alignement

Centrifuge [84] effectue le profilage taxonomique d'un échantillon métagénomique en alignant les lectures contre la base de données de génomes RefSeq avec bowtie2. L'outil consomme très peu de mémoire vive car il génère un index compressé (4,2 Go) regroupant les séquences RefSeq ayant plus

de 99% d'identité nucléotidique. Après qu'un label taxonomique ait été assigné à chaque lecture, Centrifuge estime les abondances relatives des espèces ou des taxons de rangs supérieurs avec un algorithme espérance-maximisation (ou EM pour Expectation/Maximization) qui limite les fausses détections.

Approches par pseudo-alignement

Kraken [85] met en œuvre une stratégie de "pseudo-alignement" reposant sur un index qui associe à chaque k-mer le label taxonomique correspondant au plus petit ancêtre commun des génomes dans lesquels il est présent. Un algorithme classifie ensuite une lecture à partir de la liste des taxons assignés aux k-mers la composant.

Kraken repose sur une stratégie d'indexation efficace qui regroupe les k-mers ayant une forte probabilité d'être consécutifs dans une séquence. Ainsi, les résultats des requêtes sont directement chargés depuis la mémoire cache du processeur plutôt que sur la mémoire vive. Grâce à cette optimisation, Kraken est bien plus rapide que les outils basés sur un alignement de séquences tout en atteignant une sensibilité et une précision similaire au niveau espèce. Néanmoins, la recherche des k-mers exacts composant une séquence nécessite de charger en mémoire vive un index de très grande taille dépassant parfois 100 Go. Par conséquent, l'outil est inutilisable sur des ordinateurs de bureau.

Dans le même ordre d'idées, CLARK [86] indexe uniquement les k-mers spécifiques d'une taxon donné (espèce/genre) pour simplifier l'algorithme de classification, limiter le nombre de fausses détections et réduire la consommation de mémoire vive.

Limites

Les méthodes de classification des lectures génèrent des profils taxonomiques potentiellement biaisés par la variabilité de la taille des génomes des espèces présentes. Comme le nombre de lectures alignées sur un génome est proportionnel à sa longueur, l'abondance relative des espèces sera faussée si ce paramètre n'est pas pris en compte. De plus, la présence de plasmides en de multiples copies peut amener à surestimer l'abondance d'une espèce.

Finalement, ces méthodes sont impactées par la représentativité de la base de données utilisée. Si la souche présente dans un échantillon est distante de celles utilisées pour construire la base, certaines lectures ne seront pas classifiées ce qui peut amener à sous-estimer l'abondance de l'espèce.

1.3.2.2 *Profilage taxonomique par recensement de gènes marqueurs*

Cette deuxième catégorie d'outils réalise le profilage taxonomique d'un échantillon en utilisant uniquement les lectures s'alignant sur un catalogue de gènes marqueurs prédéfini.

Utilisation de gènes universels

La comparaison à grande échelle de génomes séquencés a mis en évidence des familles de gènes orthologues communs à tous les êtres vivants [87]. Ils codent des protéines (protéine ribosomale, sous-unité de l'ADN polymérase…) ou des ARN non codants (précurseurs d'ARN de transfert…) essentiels au vivant. Parmi ces gènes, certains peuvent être utilisés comme des marqueurs phylogénétiques robustes car ils sont en général présents en une seule copie dans un génome et sont peu sujets aux transferts horizontaux. De plus, ils sont composés de régions suffisamment variables pour effectuer un profilage taxonomique jusqu'au niveau espèce.

MetaPhyler [88] aligne avec blastn les lectures d'un échantillon métagénomique sur 30 familles de gènes orthologues présents dans tous les domaines du vivants (bactéries, archées, eucaryotes). L'abondance d'un taxon est estimée à partir de la couverture verticale de ses gènes plutôt que par le nombre de lectures qui lui sont assignées. Cette normalisation permet de supprimer le biais lié à la

longueur des gènes utilisés. MetaPhyler s'appuie sur une référence composée de seulement 40 000 gènes marqueurs ce que lui permet d'être bien plus rapide et de consommer moins de mémoire que les outils alignant les lectures contre des génomes de référence.

Utilisation de gènes clade-spécifiques

L'outil MetaPhlan [89,90] s'appuie quant à lui sur une collection de marqueurs dits clade-spécifiques permettant d'identifier et de quantifier sans ambigüité des espèces microbiennes ou des rangs taxonomique de plus au niveau allant du genre jusqu'au règne. Ces marqueurs sont des gènes fortement conservés dans un clade donné et pour lesquels on ne retrouve pas de séquence similaire dans des génomes d'espèces extérieures au clade. A ce jour, le catalogue de MetaPhlan couvre plus de 7 500 espèces avec 200 marqueurs disponibles par espèce en moyenne auxquels s'ajoute plus de 115 000 marqueurs spécifiques à des rangs taxonomiques plus élevés. Ce grand nombre de marqueurs permet à MetaPhlan d'être plus sensible que les méthodes s'appuyant sur quelques dizaines de gènes universels tout en étant plus spécifique.

La base de gènes marqueurs de MetaPhlan est particulièrement compacte car elle ne couvre que 4% des gènes des 7500 génomes. Une fois indexée, elle ne pèse que 1,2Go ce qui permet d'utiliser l'outil sur une simple machine de bureau. La procédure d'alignement des lectures est particulièrement rapide car seule une petite fraction provient de marqueurs. De plus, contrairement à d'autres outils basés sur blastn, MetaPhlan s'appuie sur bowtie2 [91] qui est optimisé pour l'alignement de courtes lectures.

1.3.3 Profilage fonctionnel

Le profilage fonctionnel détermine le potentiel fonctionnel (ou capacités métaboliques) d'une communauté microbienne à partir de données de séquençage métagénomique shotgun. Il s'agit d'une étape essentielle à la compréhension le rôle du microbiote et de son interaction avec l'hôte. Intuitivement, le profilage taxonomique répond à la question « qui est là ? » tandis que le profilage fonctionnel répond à la question « que peuvent-ils faire » ? Les microorganismes d'une communauté peuvent partager plusieurs fonctions biologiques mais avoir des rôles qui leur sont propres. Par exemple, les espèces *Faecalibacterium prausnitzii*, *Eubacterium rectale* et *Ruminococcus bromii* produisent des acides gras à chaine courte par fermentation de sucres complexes tandis que l'archée *Methanobrevibacter smithii* est une des rares espèces à produire du méthane.

L'outil HUMAnN [55] détecte et quantifie les modules fonctionnels présents dans un échantillon en utilisant directement les lectures produites lors d'un séquençage shotgun. Ces dernières sont traduites à la volée en fragments protéiques en testant les 6 cadres de lecture possibles puis alignées sur la base protéique KEGG [92] en utilisant l'outil mBLASTX [93]. Finalement, HUMAnN calcule l'abondance d'une famille protéique à partir du nombre de lectures alignées sur ses gènes représentatifs tout en pondérant les résultats par la longueur des gènes et la qualité des alignements produits (e-value).

En utilisant une méthode similaire, GOmixer [94] propose de quantifier uniquement les modules fonctionnels spécifiques du microbiote intestinal humain dont l'existence a été validée par la littérature scientifique. GOmixer est accessible via une interface web mettant à disposition des outils de visualisation.

RAMMCAP [95] réduit considérablement le temps de calcul nécessaire pour comparer les lectures avec un catalogue de familles de protéines connues. Pour cela, les lectures ayant une forte homologie nucléotidique sont au préalable regroupées avec CD-HIT. Finalement, RAMMCAP est 100x plus rapide qu'une stratégie consistant à aligner individuellement chaque séquence.

ShotMAP [96] ajuste quant à lui automatiquement les paramètres du pipeline d'annotation en fonction de la nature des données à traiter. Par exemple, le bitscore au-delà duquel un alignement est considéré

varie en fonction de la longueur des lectures. ShotMAP propose de prédire directement les ORFs dans les lectures séquencées plutôt que de tester naïvement les 6 cadres de lectures possibles. A partir de données simulées, les auteurs montrent que cette stratégie permet d'estimer plus précisément l'abondance relative des différentes familles de protéines, dès lors que la longueur des lectures excède 70 bp.

Enfin, FishTaco [97] est un outil établissant un lien entre les profils taxonomiques et les profils fonctionnels calculés dans un ensemble d'échantillons métagénomiques. Après avoir mis en évidence des modules fonctionnels différentiellement abondants dans deux populations, l'outil identifie les taxons expliquant cette différence en utilisant un algorithme issu de la théorie des jeux.

1.3.4 Caractérisation au niveau souche

Les outils présentés jusqu'alors détectent et quantifient avec précision les espèces présentes dans un échantillon tout en déterminant leur potentiel métabolique. Toutefois, étant donné la grande variabilité génétique et phénotypique observée au sein des souches d'une même espèce (voir 1.1.1.3), une résolution allant au-delà du niveau espèce est requise en épidémiologie, en recherche clinique ou en génétique des populations.

Ainsi, des outils récemment développés ont révélé tout le potentiel du séquençage métagénomique shotgun en caractérisant les souches présentes dans un échantillon sans besoin d'assembler leur génome au préalable.

1.3.4.1 *Approches par construction d'haplotypes*

La première catégorie d'outil tire parti de la grande qualité des lectures produites par les séquenceurs pour recenser des polymorphismes nucléotidiques (SNV) à partir desquels est inférée une signature de la souche présente dans l'échantillon étudié.

Ainsi, StrainPhlan [98] et MIDAS [60] alignent les lectures sur des gènes marqueurs spécifiques d'une espèce (voir 1.3.2.2) ; ces marqueurs ayant en pratique une variabilité génétique équivalente à celle de l'ensemble des gènes core. Cette stratégie réduit le temps de traitement informatique en comparaison avec des outils comme metaSNV [99] alignant les lectures sur l'intégralité d'un génome.

En appliquant ce type d'outil à chaque échantillon d'une cohorte, un arbre phylogénétique des souches de chaque espèce est construit pour identifier celles associées à des traits phénotypiques comme l'origine géographique ou l'état de santé. Par ailleurs, un traçage longitudinal permet de vérifier la persistance temporelle des souches tandis qu'un traçage vertical peut révéler le transfert d'une souche d'un individu à un autre, comme par exemple entre la mère et l'enfant [100].

Les approches présentées jusqu'alors supposent qu'une seule souche d'une espèce est présente dans un échantillon. Or, l'analyse de la fréquence des variants nucléotidiques dans des échantillons métagénomiques a révélé la présence simultanée de plusieurs souches de la même espèce dans certains échantillons. Dans le microbiote intestinal, on observe souvent la présence d'une souche dominante et d'une ou plusieurs souches sous-dominantes. Le rapport d'abondance entre la souche dominante et la souche sous dominante est de l'ordre de $7/1$ [20].

StrainEst [101] détecte et quantifie l'ensemble des souches d'une espèce coexistant dans un échantillon métagénomique. Pour cela, l'outil identifie dans le core génome de l'espèce les positions où se situent des variants nucléotidiques. Pour chacune de ces positions, les fréquences des 4 allèles possibles sont calculées puis ces multiples vecteurs de fréquences sont agrégés dans une matrice. Enfin, un algorithme de régression pénalisée (Lasso) caractérise chaque souche présente comme une combinaison de variants nucléotidiques (haplotype).

Ces outils présentent l'avantage de caractériser les souches d'une espèce même lorsqu'un seul génome séquencé est disponible. Néanmoins, une couverture des souches d'intérêt supérieure à 3x est requise pour distinguer les variants nucléotidiques des erreurs de séquençage [98]. Ainsi, une telle analyse n'est possible que pour les espèces dominantes du microbiote intestinal. Enfin, une discrimination basée sur le polymorphisme nucléotidique ne permet pas de connaitre le potentiel fonctionnel spécifique à chaque souche. Or cette information est cruciale comme lors de la caractérisation des facteurs de virulence de souches pathogènes [102].

1.3.4.2 *Approches par étude de la variabilité du contenu en gènes*

La deuxième catégorie d'outil s'appuie sur le caractère non ciblé du séquençage métagénomique shotgun pour caractériser les souches à partir de leur contenu en gènes accessoires.

Pour ce faire, PanPhlan [20] ou MIDAS [60] alignent les lectures d'un échantillon contre le pan-génome d'une espèce construit en extrayant les gènes de tous les génomes disponibles pour cette espèce. Par la suite, la souche présente dans l'échantillon est caractérisée en identifiant les gènes du pan-génome présents ou absents. En cas de mélange de plusieurs souches de la même espèce dans un échantillon, seule la souche dominante est considérée. En effet, les gènes spécifiques des souches sous-dominantes sont écartés car leur couverture est inférieure à celle attendue. Lorsqu'une annotation fonctionnelle est disponible, le profil de contenu en gènes permet de caractériser avec précision le potentiel métabolique et pathogène de la souche considérée. La comparaison des souches dans plusieurs échantillons permet l'identification de sous-espèces potentiellement associées avec le phénotype de l'hôte (état de santé, origine géographique). Ces souches caractérisées par métagénomique peuvent par la suite être comparées aux génomes de référence pour identifier des sous-espèces inconnues à ce jour.

La performance de ces outils est dépendante de l'exhaustivité des pan-génomes construits c'est-à-dire du nombre de génomes de référence disponibles par espèce. De plus, une couverture de séquençage supérieure ou égale à 2x est nécessaire pour détecter la quasi-totalité des gènes constituant un microorganisme [20]. En deçà, de nombreux gènes seront manqués (faux négatifs) ce qui biaisera les résultats de l'analyse.

1.3.5 Métagénomique quantitative par utilisation d'un catalogue de gènes

1.3.5.1 *Limites des approches basées sur les génomes de référence*

La majorité des outils d'analyse de données de métagénomique shotgun présentés s'appuient sur les génomes séquencés et sont donc limités par leur disponibilité. En effet, la diversité microbienne s'étend bien au-delà du contenu des génomes de référence faisant des échantillons métagénomiques un gisement inexploité d'informations.

Malgré les progrès récents des méthodes de culture microbienne, un grand nombre d'espèces colonisant l'intestin humain demeurent probablement inconnues à ce jour et sont couramment désignées par le terme de « matière noire microbienne ». En 2016, on estimait que 50% des espèces microbiennes colonisant l'intestin humain d'individus occidentaux n'avaient pas été isolées, cultivées et séquencées. Cette proportion atteignait 85% chez les individus ayant un mode de vie non occidentalisé [60]. De plus, les espèces potentiellement pathogènes présentant un intérêt majeur pour la santé publique comme *Escherichia coli*, *Klebsiella pneumoniae* ou *Enterococcus faecium* sont représentées par des centaines voire des milliers de génomes dans les bases de données publiques (GenBank, PATRIC) tandis que seulement quelques souches sont disponibles pour la majorité des espèces commensales (**Figure 1** et **Tableau 2**). Par conséquent, des variants génétiques et de nombreux gènes accessoires associés à des traits phénotypiques pourraient manquer dans les répertoires de gènes construits à partir de génomes de référence.

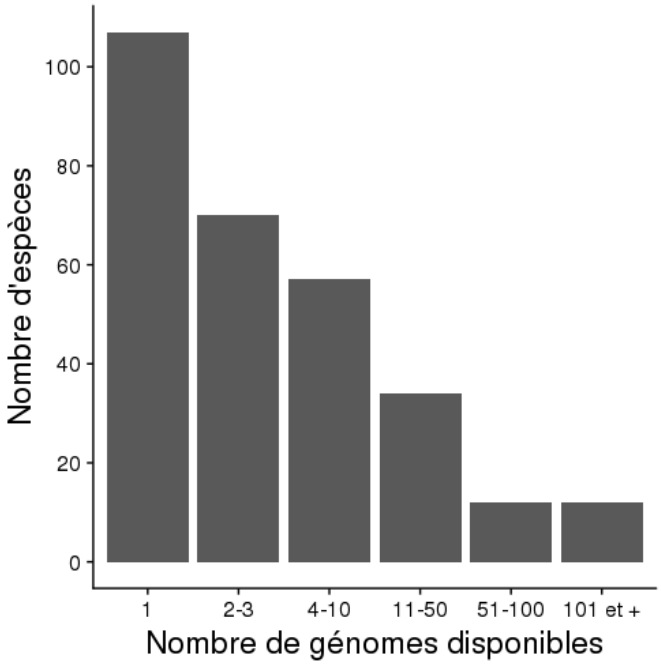


***Figure 1** : Histogramme de génomes déposés sur GenBank (Mai 2018) pour 289 espèces du microbiote intestinal humain.*

Espèce	**Nombre de génomes déposés sur GenBank**
Escherichia coli	10877
Klebsiella pneumoniae	3948
Enterococcus faecium	1009
Methanobrevibacter smithii	26
Bacteroides vulgatus	14
Ruminococcus champanellensis	2
Paraprevotella clara	1
Adlercreutzia equolifaciens	1

***Tableau 2** : Nombre de génomes déposés sur GenBank en Mai 2018 pour 8 espèces du microbiote intestinal humain. On constate une grande disparité entre les espèces d'intérêt clinique potentiellement pathogènes et les espèces commensales.*

1.3.5.2 *Création du catalogue de gènes*

L'assemblage métagénomique est une technique puissante pour dépasser les limites des méthodes reposant sur des génomes de référence. Il consiste à combiner les lectures chevauchantes en des séquences plus longues pour reconstituer les génomes des microorganismes présents dans un échantillon.

L'assemblage dans un contexte métagénomique est une tâche complexe et des outils dédiés ont été développés pour l'effectuer en un temps raisonnable tout en limitant la consommation de mémoire. Les plus populaires sont à ce jour MEGAHIT [103] et metaSPAdes [104]. MEGAHIT est le programme le plus rapide et ayant la plus faible emprunte mémoire tandis que metaSPAdes produit les assemblages les plus contigus [105].

L'assemblage est effectué en décomposant dans un premier temps les lectures en une succession de sous-mots de longueur k appelés k-mers. Dans un deuxième temps, un graphe de De Bruijn est créé :

ses sommets sont des k-mers et ses arrête sont des (k-1) mers représentatifs du chevauchement de deux k-mers dans une lecture. Le graphe De Bruijn est un graphe bidirigé où les arrêtes sont pondérées par le nombre de fois qu'un chevauchement est observé. Dans un troisième temps, les régions du graphe ayant une couverture trop faible sont éliminées car elles correspondent vraisemblablement à des erreurs de séquençage. Enfin, les séquences finales sont créées en recherchant des chemins super-eulériens dans le graphe, c'est-à-dire un ensemble de parcours passant par chaque arrête un nombre de fois au plus égal à sa pondération.

L'assemblage ne reconstitue en général que des génomes incomplets éclatés en plusieurs fragments nommés contigs [106] et ceci même pour les espèces abondantes ayant une couverture de séquençage importante. Cette fragmentation s'explique en partie par la présence de régions contenant des motifs répétés. Ces régions sont particulièrement difficiles à assembler quand leur longueur totale excède celle des lectures. En métagénomique, la coexistence de microorganismes fortement apparentés accroit cette fragmentation ou mène à la génération de séquences chimériques [107].

Dans le but d'obtenir une référence aussi exhaustive que possible, l'assemblage métagénomique est effectué sur plusieurs échantillons. Il est aussi envisageable de combiner les lectures de ces échantillons pour effectuer un seul co-assemblage ce qui augmente les chances de reconstituer des contigs issus de microorganismes sous-dominants [108]. Toutefois, cette stratégie doit être utilisée avec précaution car elle augmente le risque de générer des séquences chimériques. Dans un second temps, des gènes sont identifiés dans les contigs de chaque échantillon en utilisant des prédicteurs de gènes comme Prodigal [109] ou MetaGeneMark [110]. Pour détecter des séquences codantes, ces outils analysent la composition nucléotidique ainsi que l'usage des codons dans les contigs et recherchent des courtes séquences correspondant au site de fixation du ribosome (motif Shine Dalgarno ou Ribosome Binding Site). Par la suite, les gènes prédits dans les différents échantillons sont mis en commun et on ajoute éventuellement à cet ensemble des gènes issus de génomes de référence. Enfin, lors d'une étape appelée suppression de la redondance, les gènes ayant une forte similitude (généralement 95% d'identité sur 90% de leur longueur) sont regroupés pour être finalement représentés par un seul gène dans un catalogue dit non-redondant (**Figure 2**). Cette redondance s'explique par la duplication de certains gènes dans les génomes mais surtout par la présence du même gène dans plusieurs échantillons. La variabilité d'un gène dans les différents échantillons est due au polymorphisme nucléotidique au sein des différentes souches d'une même espèce et plus rarement à des erreurs commises lors du séquençage ou de l'assemblage. La suppression de la redondance est généralement réalisée avec l'outil CD-HIT [111] qui, plutôt que de comparer toutes les séquences deux à deux, implémente un algorithme glouton qui trie les gènes par longueur décroissante puis choisit comme nouveau représentant le gène le plus long qui n'a pas encore été clusterisé.

***Figure 2** : Chaine de traitements pour créer un catalogue de gènes non redondant*

Au final, les catalogues de gènes non-redondants du microbiote intestinal contiennent d'autant plus de gènes que le nombre d'échantillons utilisés pour les construire est important et que leurs traits phénotypiques sont variés (**Tableau 3**).

Catalogue	Nombre d'échantillons	Origine géographique	Etat de santé	Nombre de gènes
Qin *et al.* 2010 [108]	124	Europe	Sains, obèses, IBD	3,3 millions
Nielsen *et al.* 2014 [112]	396	Europe	Sains, obèses, IBD	3,9 millions
Li *et al.* 2014 [113]	1 267	Europe + USA + Chine	Sains, obèses, IBD, diabète type 2	9,9 millions
Wen et al. 2017 [46]	1 478	Europe + USA + Chine	Sains, obèses, IBD, t2d, spondylarthrite ankylosante	10,4 millions

***Tableau 3** : Taille des catalogues de gènes du microbiote intestinal humain en fonction du nombre d'échantillons et de la diversité de leurs traits phénotypiques (origine géographique et état de santé)*

1.3.5.3 *Pipeline de métagénomique quantitative*

Une fois les gènes du microbiote intestinal recensés, on les quantifie dans un ensemble d'échantillons métagénomiques n'ayant pas nécessairement servi à la construction du catalogue : on parle alors de métagénomique quantitative. La métagénomique quantitative est généralement utilisée dans le cadre d'études d'association métagénomique (MGWAS : metagenome-wide association studies) qui cherchent un lien entre la composition du microbiote intestinal et le phénotype des individus donneurs [114]. Le phénotype d'intérêt est par exemple l'état de santé (sain contre malade), la réponse à un traitement médical ou l'origine géographique.

Plusieurs chaines de traitements ou « pipelines » de métagénomique quantitative ont été proposés pour automatiser le traitement bioinformatique des données de séquençage et répartir la charge de calcul sur plusieurs serveurs ou sur une infrastructure de type cloud (**Figure 3**). Parmi elles, on trouve par exemple MOCAT [115], NG-meta-profiler [116] ou Meteor [117] développé à l'INRA.

La première étape d'un pipeline de métagénomique quantitative appelée trimming consiste à raccourcir où à éliminer les lectures composées de nucléotides dont score le qualité PHRED est inférieur à un certain seuil. L'outil Trimmomatic [118] est à ce jour le plus utilisé pour traiter des données Illumina.

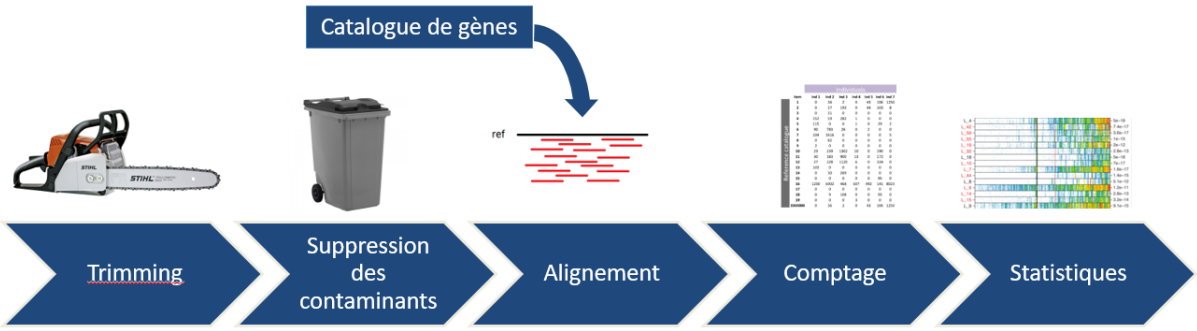


***Figure 3** : Chaine de traitement de métagénomique quantitative*

Dans un second temps, les lectures considérées comme des contaminants sont éliminées. Il peut s'agir d'adaptateurs ou de code-barres utilisés pour créer la bibliothèque de séquençage ou de lectures provenant du génome de l'hôte (ici l'humain) et non du microbiote. La proportion de lectures d'origine

humaine est faible chez les individus sains mais peut être beaucoup plus importante chez ceux sujets à une inflammation de l'intestin.

La troisième étape consiste à aligner les courtes lectures sur un catalogue de gènes de référence. Ce traitement plus couramment appelé mapping permet d'identifier le gène dont provient vraisemblablement une lecture. Il est réalisé par des outils spécialement conçus pour l'alignement de courtes séquences dont les plus populaires sont Bowtie 2 [91] et BWA [119]. Tous deux reposent sur la transformée de Burrows-Wheeler pour identifier rapidement des graines d'alignement. Ces outils sont en général utilisés en mode bout-à-bout (end-to-end) de telle sorte que seuls les alignements couvrant l'intégralité d'une lecture sont conservés. En fin de traitement, les résultats sont stockés dans un fichier au format SAM (ou BAM pour son équivalent binaire) qui est aujourd'hui un standard *de facto* en bioinformatique.

La quatrième étape appelée comptage détermine l'abondance des gènes du catalogue dans un échantillon donné. Dans un premier temps, le fichier SAM de l'échantillon est filtré pour ne conserver que alignements dont le pourcentage d'identité excède le seuil choisi lors de la suppression de la redondance, soit en général 95% d'identité. Ensuite, pour chaque lecture, seul l'alignement ayant la distance d'édition la plus faible par rapport au gène cible est conservé. Pour finir, on calcule le nombre de lectures alignées sur chacun des gènes du catalogue. Parfois, certaines lectures ont plusieurs meilleurs alignements car elles correspondent à une séquence conservée retrouvée dans plusieurs gènes. Trois stratégies existent pour gérer ces ambiguïtés. Le comptage unique ne prend en compte que les lectures ayant un seul meilleur alignement. Le comptage partagé incrémente quant à lui de $1/n$ les comptages des n gènes touchés par une lecture. Enfin, le comptage partagé « intelligent » s'inspire des algorithmes espérance-maximisation. Il consiste à distribuer chaque lecture alignée sur plusieurs gènes proportionnellement à leurs comptages uniques. Les avantages et inconvénients de chacune de ces stratégies seront évoqués dans la discussion de cette thèse (voir 8.2.4).

Enfin, les résultats sont concaténés en une table (ou matrice) de comptage donnant le nombre de lectures alignées sur chacun des gènes dans les échantillons de la cohorte étudiée. Cette matrice fait ensuite l'objet de différentes analyses statistiques dont la plus courante est la recherche de gènes différentiellement abondants dans deux catégories d'échantillons comme par exemple entre les sains et les malades. Cette recherche s'effectue grâce au test non paramétrique de Wilcoxon-Mann–Whitney ou avec des méthodes plus complexes conçues à l'origine pour l'analyse de données de transcriptomique (RNA-Seq) [120].

1.4 Structuration d'un catalogue de gènes

1.4.1 Pourquoi structurer un catalogue ?

Lors de la recherche de biomarqueurs différentiellement abondants dans deux populations, chacun des millions de gènes du catalogue de référence est un candidat. L'approche exhaustive consistant à tester l'association de millions de variables avec un phénotype d'intérêt ne permet pas d'obtenir d'information biologique réellement exploitable car une majorité des gènes découverts ne seront pas assignés à un microorganisme répertorié. Même si ces derniers ont une annotation taxonomique au niveau espèce, il pourrait s'agir de gènes communs à toutes les souches de l'espèce (core) ou de gènes souche-spécifiques (accessoires).

Cette approche souffre aussi d'un manque de puissance statistique. En effet, il est recommandé d'ajuster les p-valeurs avec des procédures telles que Benjamini–Hochberg [121] une fois les tests effectués pour limiter le nombre de fausses découvertes. Or, étant donné le très grand nombre de variables testées, peu de tests resteront significatifs après ajustement. De plus, on s'attend à ce que la majorité des gènes provenant de la même entité biologique aient des profils d'abondance similaires.

Ainsi, considérer chacun de ces gènes individuellement équivaut à tester plusieurs fois la même variable [122]. Enfin, aligner les lectures sur quelques milliers de gènes d'un microorganisme plutôt que sur son génome revient à « diluer » le signal. Ainsi, les gènes provenant des espèces sous-dominantes sont faiblement détectés voire manqués à cause du caractère aléatoire du séquençage. Par conséquent, l'estimation de leur abondance est peu précise.

Partant du principe que des gènes physiquement liés devraient avoir une abondance directement proportionnelle entre les échantillons, le regroupement de gènes co-abondants a été proposé pour organiser les catalogues en clusters de gènes provenant de la même entité biologique. Cette méthode puissante s'appuyant uniquement sur des données quantitatives reconstitue le répertoire de gènes d'espèces inconnues et permet ainsi de caractériser leur potentiel fonctionnel. De plus, les répertoires d'espèces connues sont potentiellement enrichis en gènes absents des génomes disponibles publiquement. Ce processus de réduction de la dimensionnalité augmente aussi la puissance statistique car il limite le nombre de tests à effectuer, et ce, sans perte d'information car les gènes regroupés dans un cluster ont un profil d'abondance similaire. Ainsi, les gènes les plus appropriés pour détecter et quantifier les espèces sont clairement mis en évidence. Mieux, la combinaison des signaux de plusieurs gènes apporte un gain en sensibilité permettant de détecter les espèces sous-dominantes et d'estimer plus précisément leur abondance.

1.4.2 Méthodes de regroupement des gènes co-abondants

1.4.2.1 Clustering partiel

Le clustering exhaustif consistant à comparer deux à deux les profils d'abondance de tous les gènes est une tâche dont le temps de calcul croit comme le carré du nombre de gènes à traiter (complexité quadratique). Ainsi, traiter des millions de gènes nécessite une infrastructure de calcul conséquente.

Pour réduire le nombre de comparaisons à effectuer, certains auteurs ont effectué le clustering uniquement sur le sous-ensemble de gènes statistiquement significatifs [41,42] ce qui n'améliore pas la puissance de l'analyse.

D'autres proposent de grouper uniquement les gènes assignés à 10 familles de marqueurs phylogénétiques (mOTUs , metagenomic Operational Taxonomic Units) des clusters nommés mOTU-LGs (mOTUs Linkage Groups) [123]. Ces marqueurs fournissent une résolution taxonomique plus élevée que l'ARNr 16S car ils permettent de distinguer clairement des espèces microbiennes (voir 1.3.2.2). Cette approche est particulièrement rapide car elle réduit considérablement le nombre de gènes à clustériser. Par exemple, seuls 0,18% (18 173) des 9,9 millions de gènes du catalogue IGC sont assignés à l'un des 10 mOTUs. Cependant, les génomes des espèces sous dominantes sont souvent fragmentés et incomplets après assemblage métagénomique et *in fine*, seule une petite proportion de leurs marqueurs sont présents dans les catalogues. Par conséquent, ces espèces peu abondantes peuvent être manquées en se focalisant uniquement les mOTUs. De plus, même si les mOTU-LGs permettent de réaliser le profilage taxonomique d'échantillons métagénomiques, le profilage fonctionnel n'est pas directement réalisable car les espèces sont représentées par seulement 10 gènes.

1.4.2.2 *Clustering exhaustif*

L'outil MetaProf [124] développé par l'INRA et la société AS+ dans le cadre du projet openGPU [125] calcule une matrice de corrélation entre les profils d'abondance de tous les gènes d'un catalogue. Cette tâche est effectuée sur des accélérateurs matériels GPUs car leur architecture de type SIMT (single instruction multiple threads) est particulièrement adaptée au calcul de corrélations tout en offrant un meilleur rapport $\frac{\text{puissance de calcul}}{\text{consommation énergétique}}$ [126]. Par la suite, l'outil postMetaprof importe les corrélations calculées puis effectue le clustering à proprement parler grâce à un algorithme de type

single linkage. Cependant, la génération d'une matrice de corrélations est un calcul très lourd qui n'est pas nécessaire au clustering. Pire, ce calcul génère un gros volume d'informations redondantes qui complexifie le traitement en aval.

De plus, plusieurs heuristiques ont été développées pour clustériser l'ensemble des gènes d'un catalogue sans effectuer une comparaison exhaustive de tous les vecteurs d'abondance des gènes. Le programme Metagenomic Linkage Group (MLG) [44] s'appuie sur le coefficient de corrélation de Spearman pour évaluer la similarité entre les vecteurs d'abondance des gènes puis effectue le clustering proprement dit avec l'algorithme « CHAMELEON » [127]. L'outil Metagenomic Clusters (MGC) [64] groupe quant à lui uniquement les gènes vus dans au moins 10 échantillons. Il repose sur l'algorithme de clustering de graphes MCL (Markov cluster algorithm) [128]. Dans ce graphe, les nœuds sont des gènes et les arrêtes correspondent aux distances entre des paires de gènes calculées avec le coefficient de corrélation de Spearman. Enfin, une méthode basée sur une variante de l'algorithme de clustering Canopy [129] regroupe les gènes en calculant elle aussi des corrélations de Pearson.

A ce jour, seule la méthode de clustering Canopy est disponible publiquement. Appliqué en 2014 à un catalogue de 3,9 millions de gènes du microbiote intestinal quantifiés dans 396 échantillons [112], cet outil a permis la découverte de 741 espèces métagénomiques (MGS) dont seulement 115 (15,5%) provenaient d'une espèce répertoriée à l'époque. Les MGS ont facilité l'assemblage de génomes sans utiliser de référence et ont permis la découverte d'associations entre la composition du microbiote intestinal humain et des maladies chroniques comme le diabète de type 2 ou la maladie de Crohn.

1.4.3 Limites des méthodes existantes

Malgré leurs apports significatifs, les outils présentés ci-dessus implémentent des algorithmes de clustering génériques qui ne prennent pas en compte la spécificité des objets traités. Bien qu'utilisés pour regrouper des gènes à partir de leur profil de comptage, ils pourraient s'appliquer à d'autres types de données.

Comme nous le verrons en détail dans cette thèse (2.2), les outils existants ne prennent pas en considération les caractéristiques des comptages de gènes telles que la surdispersion, la richesse en zéros, l'asymétrie et la présence de faux positifs. Pour détecter des gènes co-abondants, ces méthodes s'appuient sur les coefficients de corrélation de Pearson ou de Spearman qui évaluent respectivement une relation linéaire avec une ordonnée à l'origine potentiellement non nulle ou n'importe quelle relation monotone. Or, nous soutenons que ces types de relations ne sont pas suffisamment spécifiques car une relation de proportionnalité directe devrait être observée entre des gènes co-abondants.

De plus, ils ne font pas d'hypothèses biologiques sur la nature des clusters générés qui devraient pour la plupart correspondre à des groupes de gènes provenant de la même espèce microbienne. Pour rappel, le répertoire de gènes d'une espèce nommé pan-génome est composé de l'ensemble des gènes retrouvés dans les souches de cette espèce. En particulier, on distingue les gènes core présents dans toutes les souches des gènes accessoires présents dans seulement certaines. Bien que les méthodes existantes regroupent dans un même cluster les gènes core et gènes accessoires très prévalents d'une même espèce, ils manquent des gènes accessoires de prévalence plus faible ou les assignent à des petits clusters distincts [130]. La dépendance entre les petits clusters core et les clusters accessoires peut être évaluée dans un traitement en aval grâce à un test de cooccurrence comme le test exact de Fisher basé sur les profils de présence/absence des clusters dans les échantillons [112]. Toutefois, cette stratégie ne vérifie pas si les gènes des petits clusters satellites sont coabondants avec les gènes du cluster core. De plus, elle est peu performante pour traiter des clusters de gènes accessoires rares ou associés à des clusters core d'espèces très prévalentes.

Pour finir, ces outils ne prennent pas en compte les caractéristiques de l'écosystème étudié comme la coexistence de plusieurs souches de la même espèce dans un échantillon ou la présence d'espèces apparentées dont les gènes sont difficilement distinguables en utilisant les seuils d'identité nucléotidiques traditionnels (95%).

1.4.4 Objectifs de la thèse

L'objectif de cette thèse est de proposer un nouveau cadre d'analyse pour caractériser plus finement le microbiote intestinal. Ainsi, le but est de recenser et quantifier précisément les différents microorganismes dont ceux inconnus à ce jour mais surtout de pousser l'analyse jusqu'au niveau de la souche en considérant les gènes accessoires.

Pour cela, nous proposons une nouvelle méthode de clustering qui reconstitue le répertoire de gènes d'espèces microbiennes en regroupant des gènes co-abondants au sein d'un ensemble d'échantillons métagénomiques. Contrairement aux outils existants, il capture et distingue non seulement les gènes core des espèces mais aussi ses gènes accessoires. Cet outil s'appuie sur une nouvelle mesure de la proportionnalité prenant en compte les caractéristiques spécifiques des comptages de gènes.

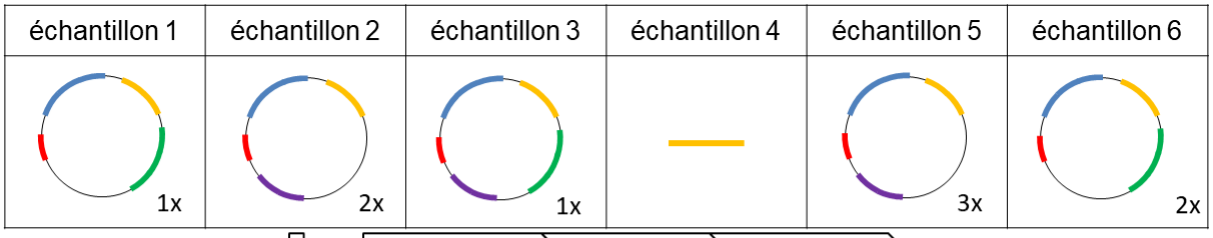
2. Préambule

2.1 Raisonnement et hypothèses de travail

Comme exposé en introduction, le pan-génome d'une espèce microbienne est un répertoire composé de gènes *core* présents chez toutes les souches de cette espèce et de gènes *accessoires* présents chez seulement certaines. Dans un contexte de séquençage métagénomique global, nous définissons comme *partagés* des gènes de l'espèce présents dans des échantillons où les gènes core de cette dernière ne sont pas détectés.

Une souche est une instance du pan-génome de l'espèce : elle est composée de l'ensemble de ses gènes core (partagés) et d'un sous-ensemble de ses gènes accessoires (partagés). Ainsi, les gènes core sont adaptés pour détecter et quantifier l'espèce, tandis que les gènes accessoires permettent de comparer ses différentes souches. Les gènes étiquetés comme partagés doivent être utilisés avec précaution. Il peut s'agir soit de gènes détectés à tort dans des échantillons (faux positifs) ou de gènes sujets à des transferts horizontaux inter-espèces.

Nous avons supposé que les gènes core devraient être systématiquement détectés dans les échantillons où l'espèce est présente à condition que la profondeur de séquençage soit suffisante (co-occurrence). De plus, l'abondance des gènes core devrait être proportionnelle dans ces échantillons (co-abondance). Remarquablement, un gène core et un gène accessoire devraient avoir une abondance proportionnelle uniquement dans le sous-ensemble d'échantillons porteurs d'une souche avec ce gène accessoire (**Figure 4**).



		éch. 1	éch. 2	éch. 3	éch. 4	éch. 5	éch. 6	
gène core 1	—	1	2	1	0	3	2	Gènes co-abondants
gène core 2	—	3	6	3	0	9	6	
gène core partagé	—	2	4	2	5*	6	4	**Partiellement** co-abondants avec les gènes core
gène accessoire 1	—	3	**0**	3	0	**0**	6	
gène accessoire 2	—	**0**	4	2	0	6	**0**	

***Figure 4** : Modèle simplifié illustrant le raisonnement sur lequel s'appuie la méthode*

On considère 6 échantillons dont tous à l'exception du 4ème sont porteurs d'une souche d'une même espèce microbienne représentée par un chromosome circulaire. Chaque souche possède différents gènes matérialisés par des arcs de cercles colorés.

Le pan-génome de l'espèce est composé de 2 gènes core (rouge et bleu) présents dans toutes les souches, de deux gènes accessoires (vert et violet) présents dans seulement certaines ainsi que d'un

gène core partagé (jaune) présent dans l'échantillon 4 où l'espèce est absente. Les gènes ont pour longueur 1 (gène rouge), 2 (gène jaune et gène violet) ou 3 (gène bleu et gène vert) suivant une unité de longueur arbitraire.

On effectue un séquençage métagénomique global de chaque échantillon en générant des lectures de longueur 1 (longueur du gène le plus court). La profondeur de séquençage indiquée en bas à droite du chromosome de chaque souche varie de 1 à 3. Finalement, une table d'abondance comptabilisant le nombre de lectures alignées sur chaque gène dans chaque échantillon est générée.

Une relation de proportionnalité directe est observée entre les vecteurs d'abondance de deux gènes core ; le coefficient de proportionnalité étant égal au ratio de leurs longueurs respectives. Comme le gène core bleu est 3 fois plus long que le gène core rouge, ses comptages sont 3 fois plus élevés. En revanche, une telle relation entre un gène core et un gène accessoire n'est observée que dans le sous-ensemble d'échantillons où le gène accessoire est présent. Pour chaque gène accessoire, les échantillons à écarter pour observer la relation de proportionnalité sont mis en évidence par les zéros en gras. Finalement, le gène core partagé à des comptages proportionnels à ceux des 2 gènes core même s'il est détecté de façon inattendue dans l'échantillon 4.

2.2 Nature des comptages

2.2.1 Caractère aléatoire du séquençage et hétéroscédasticité

Après un séquençage global d'un échantillon métagénomique, l'abondance d'un gène est mesurée par le nombre de lectures alignées sur celui-ci. Le modèle présenté dans la **Figure 4** est simplificateur car le séquençage n'est pas un processus déterministe où le nombre de lectures produites par un gène est exactement proportionnel à sa longueur et à l'abondance du microorganisme dont il provient.

En réalité, le séquençage est la réalisation d'un processus aléatoire (ou stochastique) où des fragments d'ADN lus par le séquenceur sont tirés aléatoirement dans l'échantillon métagénomique. En effet, si l'on séquençait plusieurs fois le même échantillon en générant l lectures (réplicat technique), le nombre de lectures alignées sur un gène d'abondance relative a suivrait une loi de Poisson de moyenne $a \cdot l$. Or, la variance d'une loi de Poisson est égale à sa moyenne [131]. Ainsi, la variance des comptages n'est pas constante : on dit qu'il y a hétéroscédasticité. Plus les comptages sont forts, plus l'erreur absolue commise est élevée.

Le tirage aléatoire des fragments d'ADN n'est pas la seule source de variabilité. A cette dernière s'ajoutent par exemple le stockage et le traitement des échantillons, la préparation de la bibliothèque de séquençage, les biais de séquençage et les traitements bio-informatiques. L'étude des comptages de gènes dans des échantillons métagénomiques réels montre que la variance croit plus vite que la moyenne : on parle de surdispersion [132]. Pour prendre en compte cette variabilité supplémentaire, on suggère de modéliser les comptages par une loi quasi-poisson (variance proportionnelle à l'espérance) ou une loi binomiale négative (variance proportionnelle au carré de l'espérance) [133].

2.2.2 Valeurs nulles surreprésentées

La plupart des espèces du microbiote intestinal humain ne sont détectées que dans un faible pourcentage d'individus. De plus, la majorité de leurs gènes accessoires sont rares car ils ne sont présents que dans seulement quelques souches [30]. Par conséquent, les gènes du microbiote intestinal sont pour la grande majorité détectés dans peu d'échantillons. A titre d'exemple, la table quantifiant les 9,9 millions de gènes du catalogue IGC [113] dans 1267 échantillons est constituée à 92% de zéros. En effet, 79% des gènes du catalogue IGC sont détectés dans moins de 10% des échantillons (**Figure 5**). En tirant aléatoirement un gène, on s'attend à ce que son vecteur d'abondance dans un ensemble d'échantillons contienne essentiellement des zéros. On parle dans la littérature de données de comptages « gonflées en zéros » (zero-inflated) [134].

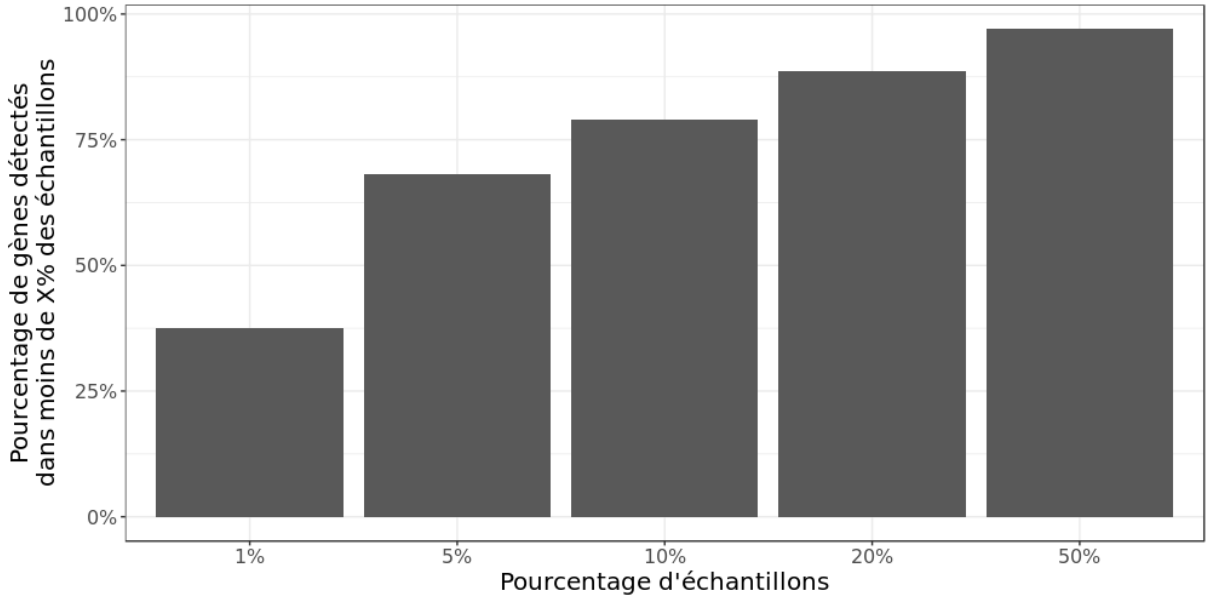


Figure 5 *: Fonction de répartition de la prévalence des gènes du catalogue IGC dans 1267 échantillons.*

2.2.3 Asymétrie de la distribution

Le nombre de lectures alignées sur un gène varie significativement d'un échantillon à l'autre. Si l'on omet les zéros, les comptages s'échelonnent parfois sur quatre ordres de grandeur, soit de un à quasiment 10 000 (**Figure 6**). Le plus souvent, la distribution des comptages d'un gène est asymétrique biaisée vers la gauche car les comptages faibles sont beaucoup plus nombreux que les comptages forts.

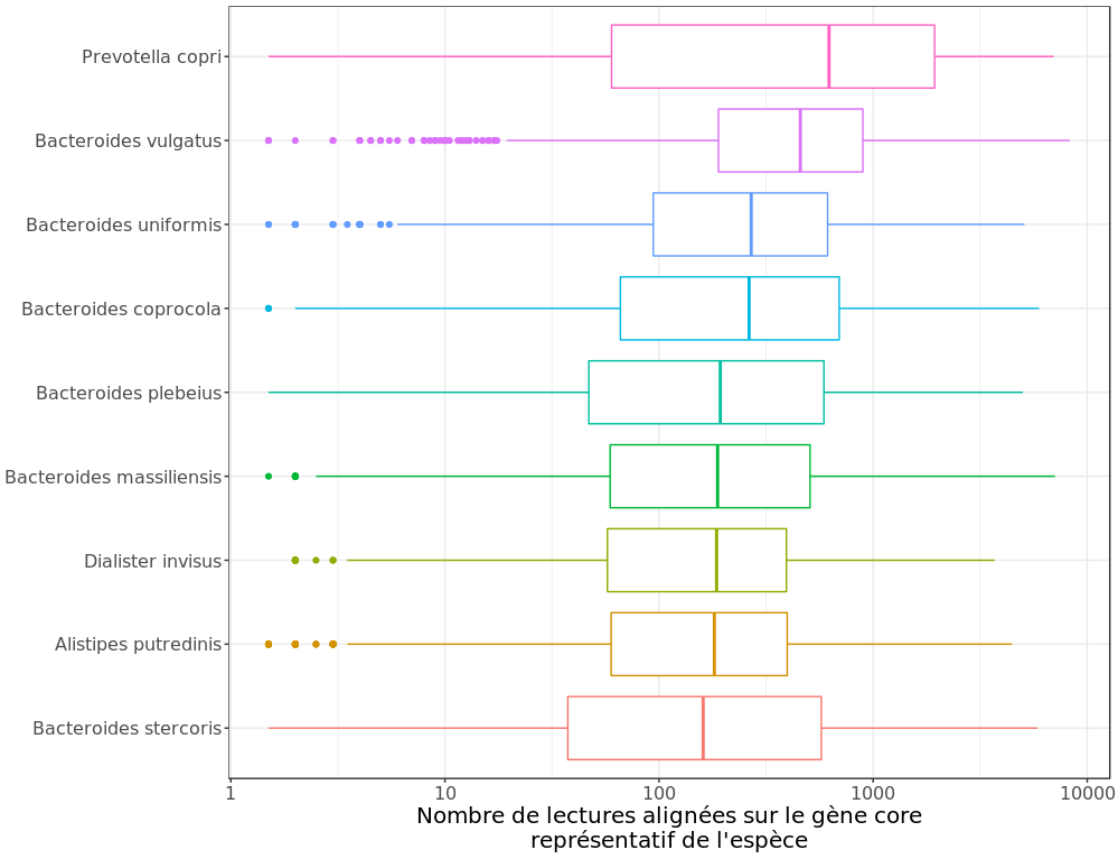


Figure 6 *: Nombre de lectures alignées sur les gènes core de 9 espèces du microbiote intestinal humain (échelle logarithmique) dans les 1267 échantillons du catalogue IGC. On remarque que les comptages*

s'étalent sur quatre ordres de grandeur. Pour chaque gène, les échantillons avec un comptage nul n'ont pas été considérés

Cette variation a d'abord une origine écologique car l'abondance relative d'un microorganisme et donc de ses gènes varie jusqu'à 3 ordres de grandeur [108] d'un individu à l'autre. De plus, le nombre de lectures alignées sur des gènes est proportionnel à la profondeur séquençage. Or, ce paramètre peut varier d'un facteur 10 d'un échantillon à l'autre suivant le but visé (**Figure 7**). En effet, une profondeur importante est requise pour effectuer un assemblage *de novo* tandis qu'un profilage taxonomique ou fonctionnel peut être réalisé avec un nombre de lectures plus faible si l'on dispose d'un catalogue de gènes suffisamment représentatif de l'écosystème étudié. Finalement, le nombre de lectures générées n'est pas parfaitement maitrisable car il est impacté par des facteurs comme la quantité et de la qualité de l'ADN extrait, la préparation de la bibliothèque de séquençage ainsi que la température ambiante de la pièce où est entreposé le séquenceur.

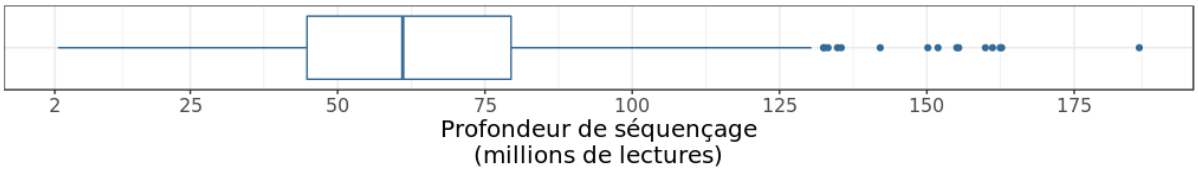


***Figure 7** : Nombre de lectures produites lors du séquençage des 1267 échantillons du catalogue IGC*

3. Détection de gènes co-abondants

3.1 Coefficients de corrélations traditionnels

3.1.1 Corrélation de Pearson

La corrélation de Pearson est le coefficient le plus couramment utilisé pour détecter des gènes co-abondants. Ce coefficient évalue une relation linéaire du type $y = a \cdot x + b$ entre deux variables x et y (Formule 1). Pour rappel, le coefficient de corrélation de Pearson s'échelonne entre -1 et 1. Plus le coefficient est proche de 1 en valeur absolue, plus la relation linéaire est forte. Un coefficient positif (respectivement négatif) indique une relation linéaire croissante (respectivement décroissante) donc une constante a positive (respectivement négative). Dans le contexte de la recherche de gènes co-abondants, seuls les coefficients positifs nous intéressent.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Formule 1 : Coefficient de corrélation de Pearson de deux jeux de données $\{x_1, \dots, x_n\}$ *et* $\{y_1, \dots, y_n\}$ *où :*

- n *correspond aux nombres d'observations*
- x_i et y_i correspondent respectivement à la i-ème observation des variables x et y
- $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ et $\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i$

Le coefficient de corrélation de Pearson n'est pas idéal pour comparer des vecteurs de comptages bruts de gènes. Premièrement, une étude sur des données simulées montre qu'il sous-estime la « véritable » corrélation lorsque la majorité des comptages sont nuls [135]. Deuxièmement, il est peu robuste lorsque les données traitées ne suivent pas une distribution normale et *a fortiori* asymétrique. Lors du calcul du coefficient, les comptages forts auront un poids plus important que les comptages faibles ce qui augmente le risque de détecter de fausses associations [136]. De plus, il est sensible à l'hétéroscédasticité [137]. En effet, son calcul nécessite d'estimer des écarts types (c.f. dénominateur dans la Formule 1) à partir d'un estimateur qui suppose que la variance des comptages est constante (homoscédasticité). Enfin, le coefficient de corrélation de Pearson sera sous-estimé en présence de valeurs aberrantes et ce d'autant plus qu'elles seront nombreuses et éloignées des valeurs attendues [138].

3.1.2 Corrélation de Spearman

D'autres auteurs utilisent le coefficient de corrélation de Spearman pour détecter des gènes co-abondants [15,41]. Ce coefficient consiste à calculer une corrélation de Pearson sur les rangs des comptages plutôt que sur les comptages eux-mêmes. Contrairement au coefficient de corrélation de Pearson, il n'est pas impacté par l'asymétrie de la distribution et beaucoup moins par l'hétéroscédasticité. Néanmoins, il sous-estime lui aussi la « véritable » corrélation lorsque les valeurs nulles sont surreprésentées [135]. Il est aussi impacté par la présence de valeurs aberrantes même si l'écart entre la valeur observée et la valeur attendue a potentiellement moins d'importance en travaillant sur des rangs. Enfin, il permet de détecter n'importe quelle relation monotone alors que l'on ne s'intéresse ici qu'aux relations de proportionnalité directe.

3.2 Impact de la transformation des comptages

Certains auteurs proposent de transformer les données de comptage avant de calculer des corrélations pour limiter l'asymétrie et/ou l'hétéroscédasticité.

3.2.1 Transformation logarithmique

L'application d'une fonction log (généralement de base 10) sur des données de comptage est couramment utilisée pour limiter l'asymétrie de la distribution [130]. Cette transformation « étale » les comptages et permet, si on omet les valeurs nulles, de s'approcher d'une distribution normale (**Figure 8.B** et **Tableau 4**). La relation de proportionnalité entre deux gènes notée $g_2 = \alpha \cdot g_1$ devient en $\log_{10}(g_2) = \log_{10}(\alpha \cdot g_1) \leftrightarrow \log_{10}(g_2) = \log_{10}(\alpha) + \log_{10}(g_1)$. Cependant, la fonction log ne stabilise pas la variance et introduit un biais important pour les comptages faibles [133]. Après la transformation, il y a toujours hétéroscédasticité car la variance est d'autant plus grande que les comptages sont faibles (**Figure 9.B**). De plus, on doit introduire un pseudo-comptage car la fonction log n'est pas définie en 0. Or, si l'on ajoute 1 à tous les comptages, la relation de proportionnalité n'est plus vérifiée car $\log_{10}(g_2 + 1) \neq \log_{10}(\alpha) + \log_{10}(g_1 + 1)$. La différence est négligeable lorsque les comptages sont élevés mais est importante lorsqu'ils sont proches de 1.

3.2.2 Transformation racine carrée

La fonction racine carrée est une transformation simple où la relation de proportionnalité entre deux gènes $g_2 = \alpha \cdot g_1$ devient $\sqrt{g_2} = \sqrt{\alpha \cdot g_1} \leftrightarrow \sqrt{g_2} = \sqrt{\alpha}\sqrt{g_1}$. Comme expliqué en 2.2.1, les données de comptage suivent en première approximation une distribution de Poisson. Or, on démontre que la fonction racine carrée transforme une loi de Poisson de moyenne λ et de variance λ^2 en une loi normale de moyenne $\sqrt{\lambda}$ et de variance $1/4$ [139]. Après transformation, la variance est une constante qui ne dépend plus de λ : il y a donc homoscédasticité. Cependant, les comptages de gènes sont surdispersés car la variance croit plus vite que le moyenne. Par conséquent, il y a toujours hétéroscédasticité après application de la racine carrée. Comme illustré par la **Figure 9.B**, la variance est stable lorsque les comptages sont faibles puis croit lorsque ceux-ci augmentent. Néanmoins, l'hétéroscédasticité est beaucoup moins marquée qu'avec des comptages bruts ou log-transformés.

Deuxièmement, la transformation racine carrée est adaptée aux distributions asymétriques à droite car elle « compresse » la distribution vers la gauche (**Figure 8.C** et **Tableau 4**). Même si la fonction log semble plus efficace pour limiter une asymétrie à gauche, la performance de la fonction racine carrée reste acceptable.

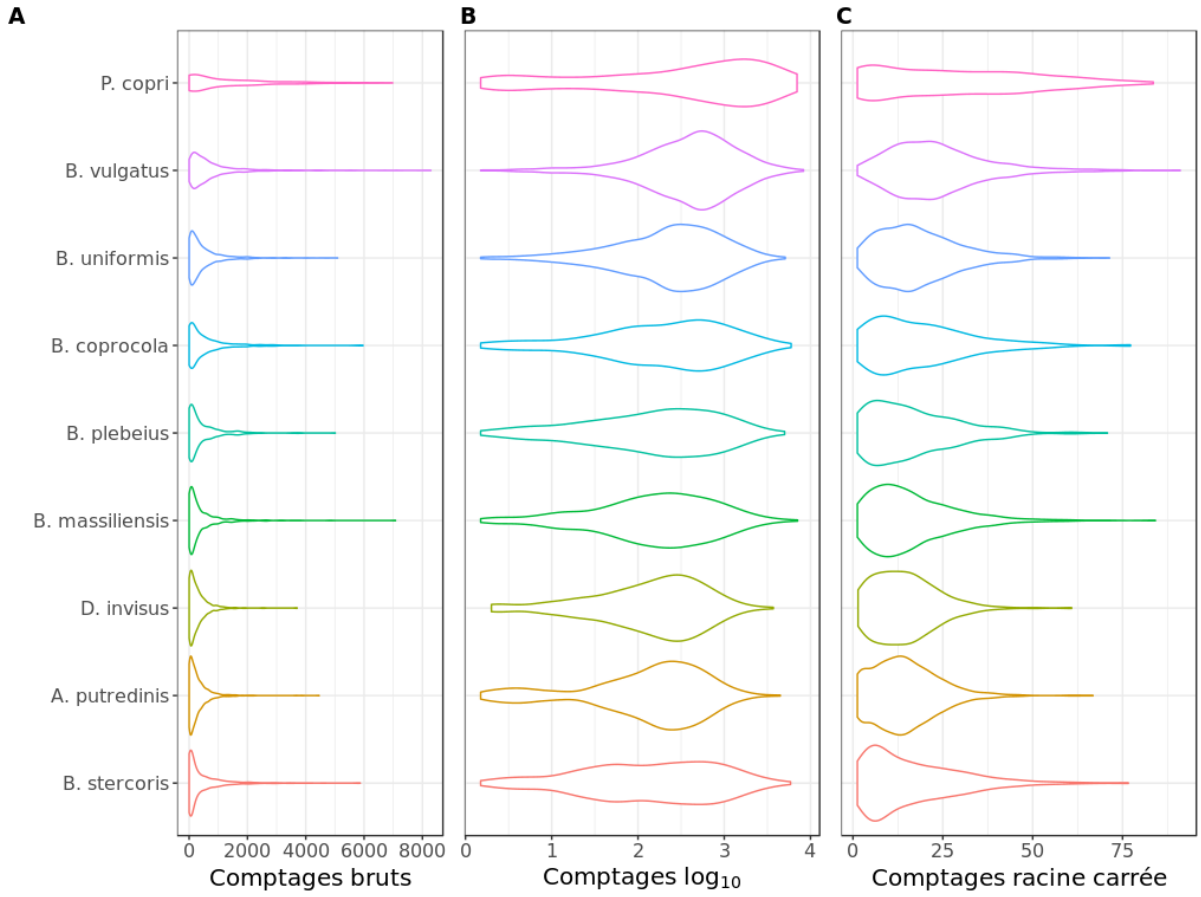


Figure 8 *: Distribution du nombre de lectures alignées sur les gènes core de 9 espèces du microbiote intestinal humain. Les 1267 échantillons du catalogue IGC ont été utilisés. Pour chaque gène, les échantillons avec un comptage nul ont été filtrés.*

A. Comptages bruts. La distribution des comptages est fortement biaisée à gauche. On a une majorité de comptages faibles et peu de comptages forts.

B. Comptages log transformés. La transformation « étale » les comptages. On s'approche d'une distribution normale.

C. Comptages ayant subi un transformation racine carrée. La transformation étale les comptages même si la distribution reste biaisée à gauche.

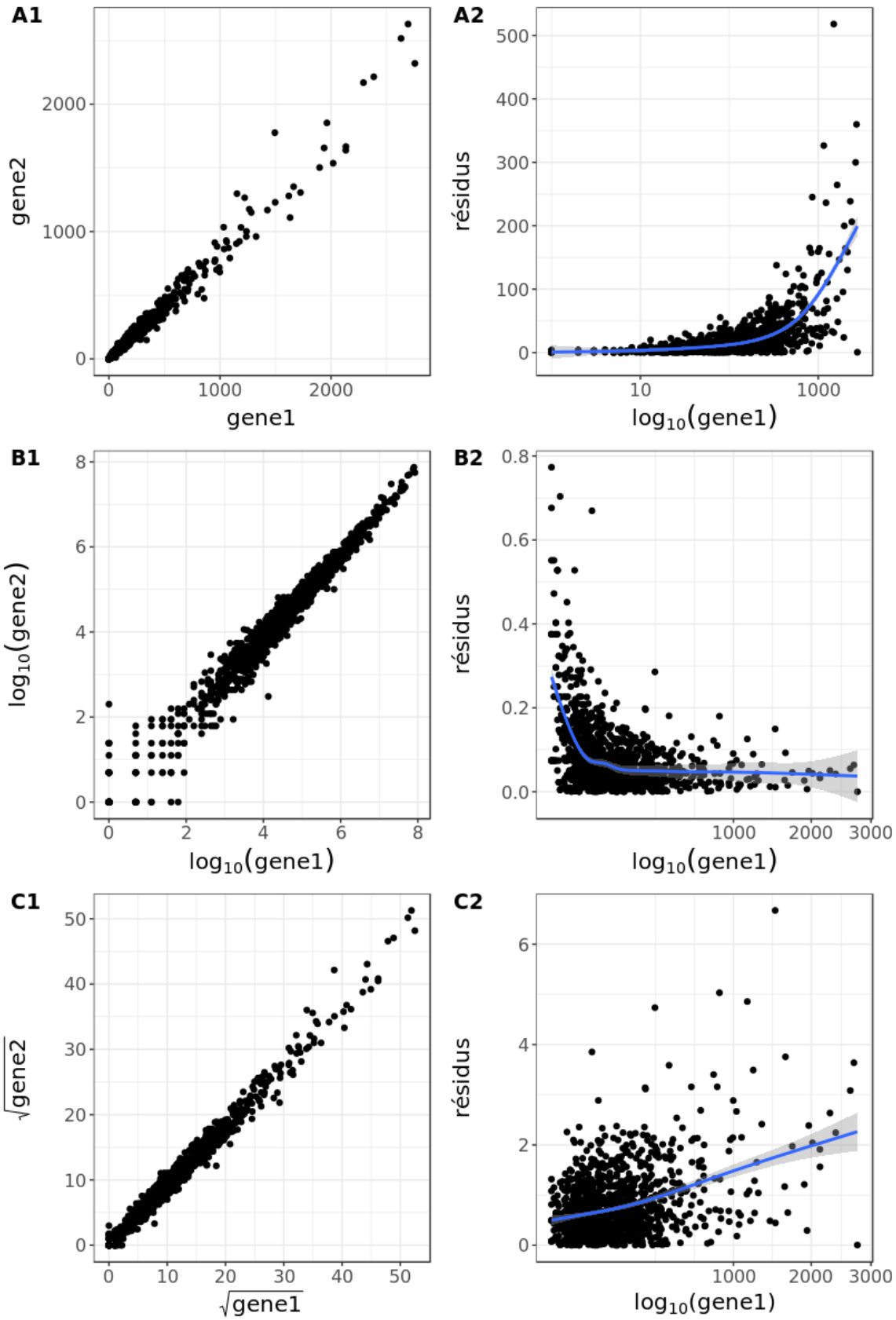


Figure 9 *: Comparaison des comptages de deux gènes core de l'espèce Parabacteroides distasonis. (Axe des abscisses : MH0020_GL0053675 ; axes des ordonnées : V1.UC58-4_GL0199806) dans les 1267 échantillons du catalogue IGC.*

A. Comptages bruts. A gauche (A1) : nuage de points des comptages des deux gènes. A droite (A2) : résidus en valeur absolue de la régression $g_2 = \alpha \cdot g_1$. Plus les comptages sont forts, plus l'erreur est importante.

B. Comptages log transformés. A gauche (B1) : nuage de points des comptages des deux gènes. A droite (B2) : résidus en valeur absolue de la régression $log_{10}(g_2) = log_{10}(\alpha) + log_{10}(g_1)$. Plus les comptages sont faibles, plus l'erreur est importante.

C. Comptages ayant subi un transformation racine carrée. A gauche (C1) : nuage de points des comptages des deux gènes. A droite (C2) : résidus en valeur absolue de la régression $\sqrt{g_2} = \sqrt{\alpha} \cdot \sqrt{g_1}$. Plus les comptages sont forts, plus l'erreur est importante même si elle croit moins vite qu'avec les comptages bruts.

	Asymétrie des comptages bruts	Asymétrie des comptages $\log_{10}$	Asymétrie des comptages $\sqrt{}$
Prevotella copri	1.48	-0.80	0.50
Bacteroides vulgatus	3.23	-0.88	1.20
Bacteroides uniformis	2.80	-0.77	1.01
Bacteroides coprocola	3.15	-0.68	1.16
Bacteroides plebeius	3.05	-0.55	1.10
Bacteroides massiliensis	4.38	-0.57	1.47
Dialister invisus	3.88	-0.83	1.06
Alistipes putredinis	4.11	-0.96	1.02
Bacteroides stercoris	3.27	-0.37	1.27

***Tableau 4** : Asymétrie de la distribution des comptages des gènes core de 9 espèces du microbiote intestinal humain*

L'asymétrie est estimée avec le coefficient d'asymétrie de Pearson. Une valeur nulle indique une distribution symétrique, une valeur positive une asymétrique à gauche et une valeur négative une asymétrie à droite. Plus la valeur absolue est importante, plus l'asymétrie est forte [140]*. La transformation log_{10} limite plus fortement l'asymétrie que la transformation racine carrée.*

3.2.3 Raréfaction

Pour atténuer la variabilité technique des comptages, certains auteurs ont suggéré de procéder à leur raréfaction (ou downsizing) [112]. Cette méthode consiste à ramener l'ensemble des échantillons à une même profondeur de séquençage en tirant des lectures dans chacun d'eux aléatoirement sans remise. Le seuil de raréfaction doit être inférieur ou égal au nombre de lectures de l'échantillon ayant la profondeur de séquençage la plus faible. Cependant, la raréfaction diminue de façon limitée l'asymétrie de la distribution (**Figure 10.A**) car la variabilité des comptages est essentiellement d'origine biologique. De plus, en omettant une partie des données disponibles dans les échantillons où la profondeur excède le seuil fixé, la raréfaction ajoute artificiellement de l'incertitude. Le regroupement par co-abondance sera moins performant [132], en particulier pour les gènes provenant d'espèces sous dominantes car leur signal sera plus faible voire nul (**Figure 10.B**).

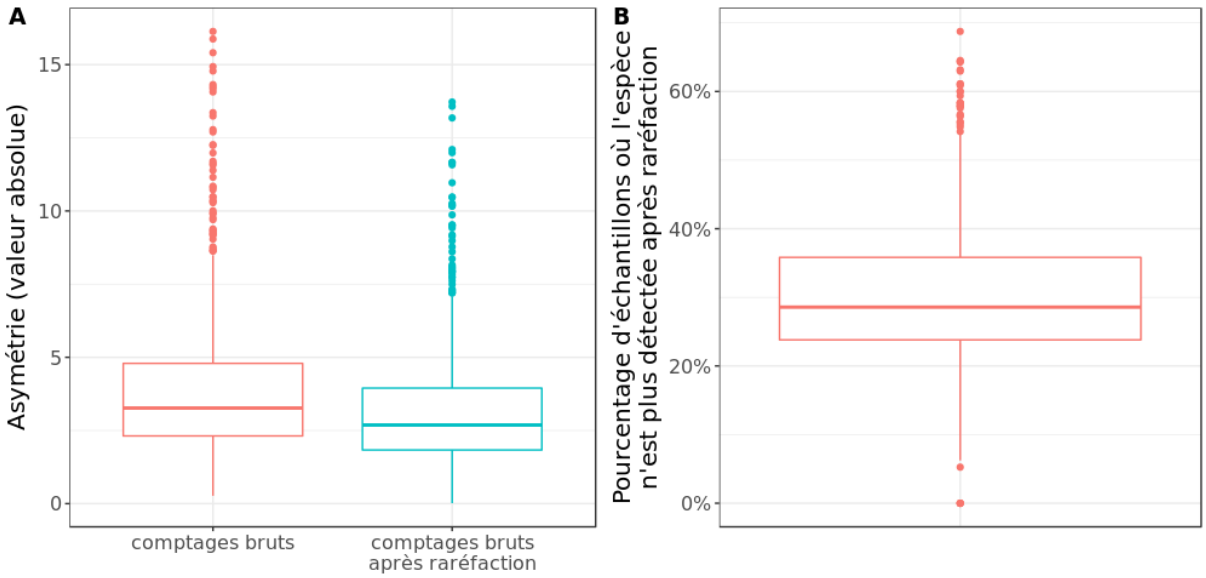


Figure 10 *: Effet de la raréfaction sur la distribution des comptages des gènes core de 1135 espèces du microbiote intestinal humain.*

Les 1267 échantillons du catalogue IGC [113] *ont été utilisés. Le seuil de raréfaction a été fixé à 15 millions de lectures pour tous les échantillons. Seules les espèces détectées dans au moins 10 échantillons avant raréfaction ont été considérées.*

A. Boxplots donnant pour chacune des 1135 espèces l'asymétrie de la distribution des comptages avant et après raréfaction. La raréfaction diminue légèrement l'asymétrie (médiane avant raréfaction = 3.26, médiane après raréfaction = 2.69)

B. Boxplot donnant pour chacune des 1135 espèces le pourcentage d'échantillons où elle était détectée avant raréfaction mais plus après (médiane = 28.6%)

3.3 Méthode proposée

Bien que la transformation racine carrée et le filtrage des zéros permettraient d'utiliser le coefficient de corrélation de Pearson pour détecter des gènes co-abondants, nous avons choisi de ne pas l'utiliser. En effet, ce coefficient permet de détecter n'importe quelle relation linéaire du type $y = a \cdot x + b$ entre deux variables x et y (**Figure 11**). Or, une relation de proportionnalité directe du type $y = a \cdot x$ avec une pente positive ($a > 0$) et une ordonnée à l'origine nulle ($b = 0$) est attendue entre des gènes co-abondants comme l'ont suggéré de précédentes études [141].

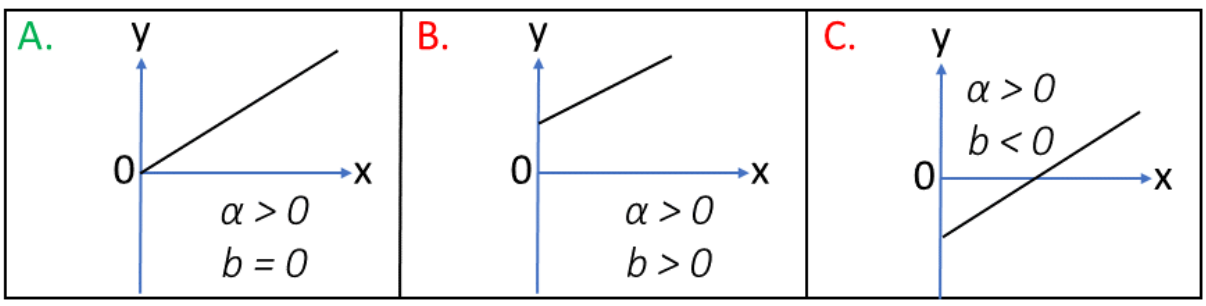


Figure 11 *: 3 types de relations linéaires ($y = \alpha \cdot x + b$) correspondant à un coefficient de corrélation de Pearson strictement positif. Seul le cas illustré par la sous-figure A nous intéresse.*

Ci-dessous, nous décrivons deux nouvelles mesures développées lors de cette thèse qui évaluent le lien de proportionnalité entre les comptages de deux gènes.

3.3.1 Notations

Soit $E = \{e_1, e_2, \ldots, e_s\}$ un ensemble non vide de s échantillons métagénomiques.

Soient $g_1 = \left(g_{1,e_1}, g_{1,e_2}, \ldots, g_{1,e_s}\right)$ et $g_2 = \left(g_{2,e_1}, g_{2,e_2}, \ldots, g_{2,e_s}\right)$ les vecteurs contenant le nombre de lectures alignées sur les deux gènes à comparer dans l'ensemble d'échantillons métagénomiques E.

Dans un premier temps, la méthode proposée estime le coefficient de proportionnalité supposé (α) entre g_1 et g_2. Ensuite, la proportionnalité entre g_1 et g_2 est évaluée en fonction du coefficient de proportionnalité α précédemment calculé (mesure non robuste de la proportionnalité, p_{nr}). Sinon, la proportionnalité est évaluée après élimination des échantillons avec des comptages aberrants (mesure robuste de la proportionnalité, p_r).

3.3.2 Estimation du coefficient de proportionnalité

Admettons qu'il existe un lien de proportionnalité directe entre g_1 et g_2 noté $g_2 = \alpha \cdot g_1$ où α est une constante strictement positive correspondant au coefficient de proportionnalité. Si l'on suppose que le nombre de lectures alignées sur un gène est proportionnel à sa longueur, alors on s'attend à ce que le coefficient α soit égal au rapport des longueurs de g_2 et g_1. Toutefois, ce rapport n'est pas toujours un bon estimateur par exemple lorsqu'un un gène est dupliqué ou lorsque sa couverture de séquençage n'est pas uniforme (**Figure 12**).

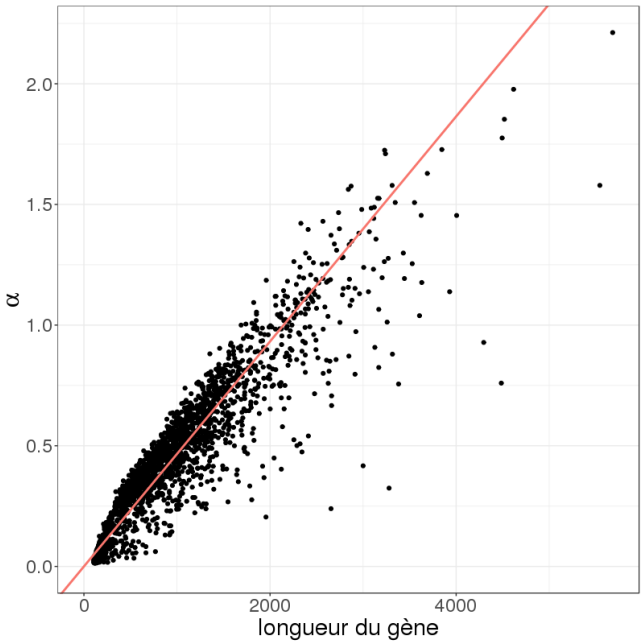


Figure 12 *: Comparaison de la longueur des 1921 gènes core de l'espèce Parabacteroides distasonis (axe des abscisses) et de leur coefficient de proportionnalité α respectif (axe des ordonnées).*

La longueur des gènes est exprimée en paires de bases. Le coefficient de proportionnalité α est estimé en comparant le profil d'abondance de chaque gène avec le profil d'abondance médian des 30 meilleurs gènes core de l'espèce. Les gènes ont été quantifiés dans les 1267 échantillons du catalogue IGC. La ligne rouge correspond à la tendance calculée avec une régression linéaire robuste en forçant une ordonnée à l'origine nulle.

On constate que le coefficient α est proportionnel à la longueur des gènes (corrélation de Pearson = 0.91, p-valeur=0). Néanmoins, les points en dessous de la ligne rouge indiquent que certains gènes ont une couverture plus faible qu'attendue.

Par conséquent, on propose d'estimer α en calculant la médiane des rapports des comptages des gènes :

$$\alpha = \text{médiane}\left(\frac{g_{2,e}}{g_{1,e}}\right) e \in E | \left(g_{1,e} \geq t \wedge g_{2,e} \geq t\right)$$

La médiane permet d'obtenir un estimateur robuste tolérant théoriquement jusqu'à 50% de valeurs aberrantes. Dans cette formule, seuls les échantillons où g_1 et g_2 ont des comptages au-dessus d'un seuil t sont pris en compte (**Figure 14**). L'estimation du coefficient de proportionnalité à partir de ce sous-ensemble d'échantillons présente les avantages suivants :

1. On exclut les échantillons où les deux gènes ont un comptage nul car ils n'apportent aucune information quantitative.
2. On exclut les échantillons où les deux gènes ont des comptages faibles car ils ne permettent pas une estimation précise du coefficient de proportionnalité.
3. On exclut les échantillons où un des deux gènes a un comptage nul car ceci permet de détecter un lien de proportionnalité concernant uniquement un sous-ensemble d'échantillons.

Les gènes ne sont pas comparés si moins de 3 échantillons sont disponibles pour estimer α.

3.3.3 Classifications des comptages nuls

Dans un échantillon, un comptage nul pour un gène est soit un zéro structurel soit un zéro d'échantillonnage [142]. Dans le premier cas, le gène est réellement absent de l'échantillon. Dans le second cas, le gène est présent dans l'échantillon mais il n'est pas observé car la profondeur de séquençage est trop faible pour le détecter. Lors de la reconstitution d'un pan-génome, il est crucial de distinguer ces deux types de zéros pour classifier avec précision un gène en tant que core, accessoire ou partagé.

Un gène avec un comptage nul dans un échantillon est classifié comme zéro structurel si l'autre gène a un comptage supérieur au seuil de quantification soit formellement :

$$e \in E \mid \left(g_{1,e} \geq t \wedge g_{2,e} = 0\right) \vee \left(g_{1,e} = 0 \wedge g_{2,e} \geq t\right)$$

Dans les autres cas, le zéro est classifié comme indéterminé. Il est impossible de déterminer si le gène est réellement absent de l'échantillon ou si la profondeur de séquençage est trop faible pour le détecter.

Le seuil de quantification t est fixé à 6 par défaut. On s'appuie sur le raisonnement suivant pour le justifier. Supposons que le séquençage d'un échantillon métagénomique génère n lectures et que cet échantillon possède un gène d'abondance relative a. Si la seule source de variabilité est liée au tirage aléatoire des fragments d'ADN, alors la variable aléatoire X correspondant au nombre de lectures à alignées sur ce gène suit une loi de Poisson de moyenne $n \cdot a$ soit $X \sim Poisson(\lambda = n \cdot a)$

Si l'on pose $n \cdot a = 6$ alors $P(X = 0) = 0.0025$ (**Tableau 5**). Autrement dit, on a 0.25% de chance de classifier à tort comme zéro structurel en choisissant un seuil t égal à 6. En pratique, cette probabilité est légèrement supérieure car les comptages sont surdispersés (voir 2.2.1)

n.a	1	2	3	4	5	6
P(X=0| n.a)	36.8%	13.5%	4.98%	1.83%	0.67%	0.25%

***Tableau 5** : Probabilité d'obtenir un zéro d'échantillonnage (2ème ligne) en fonction du nombre théorique n.a de lectures alignées sur le gène (1ère ligne). La variable X suit une loi de Poisson de paramètre n.a*

Si l'on considère un gène de 1000 paires de bases et des lectures de 100 paires de bases, le seuil de quantification $t = 6$ équivaut à une couverture minimale de $\frac{longueur\ des\ lectures \times t}{longueur\ du\ gène} = \frac{100*6}{1000} = 0{,}6$.

3.3.4 Adaptation des seuils de quantification en fonction du coefficient de proportionnalité

Lorsque que l'on compare les comptages de deux gènes co-abondants qui n'ont pas la même longueur, le coefficient de proportionnalité α est différent de 1 car l'un deux gènes produit des comptages plus élevés que l'autre. Ceci peut mener à classifier à tort un zéro comme structurel.

A titre d'exemple, considérons deux gènes g_1 et g_2 ayant pour longueur respectives 200 et 2000 paires de bases. Si l'on suppose que le coefficient de proportionnalité α est égal au rapport des longueurs des deux gènes alors $g_2 = \frac{2000}{200} \cdot g_1 = 10 \cdot g_1$. Dans un échantillon, on pourrait observer un comptage supérieur au seuil de quantification pour g_2 mais nul pour g_1 alors que ce dernier était bien présent (**Figure 13**). En utilisant le seuil $t = 6$, le comptage nul de g_1 serait classifié à tort comme un zéro structurel.

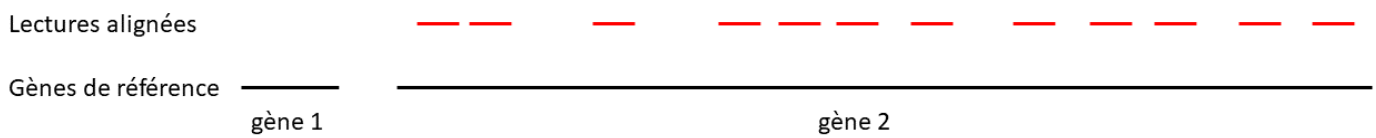


***Figure 13** : Lectures alignées (traits rouges) sur deux gènes (traits noirs) d'une espèce séquencée*

Le gène 2 est 10 fois plus long que le gène 1. 12 lectures sont alignées sur le gène 2 mais aucune sur le gène 1 alors que ce dernier est présent dans l'échantillon. Ici, la profondeur de séquençage n'est pas suffisante pour le détecter : il s'agit d'un zéro d'échantillonnage.

Pour résoudre ce problème, on utilise des seuils de quantification différents pour g_1 et g_2 nommés respectivement t_1 et t_2 dont la valeur varie en coefficient de proportionnalité (**Figure 14**) :

$$t_1 = \max\left(t, \frac{t}{\alpha}\right) \text{ et } t_2 = \max(t, \alpha \cdot t)$$

Finalement, un gène avec un comptage nul est classifié comme un zéro structurel si l'autre gène a un comptage supérieur à son seuil de quantification soit :

$$e \in E \mid \left(g_{1,e} \geq t_1 \ \wedge \ g_{2,e} = 0\right) \vee \left(g_{1,e} = 0 \ \wedge \ g_{2,e} \geq t_2\right)$$

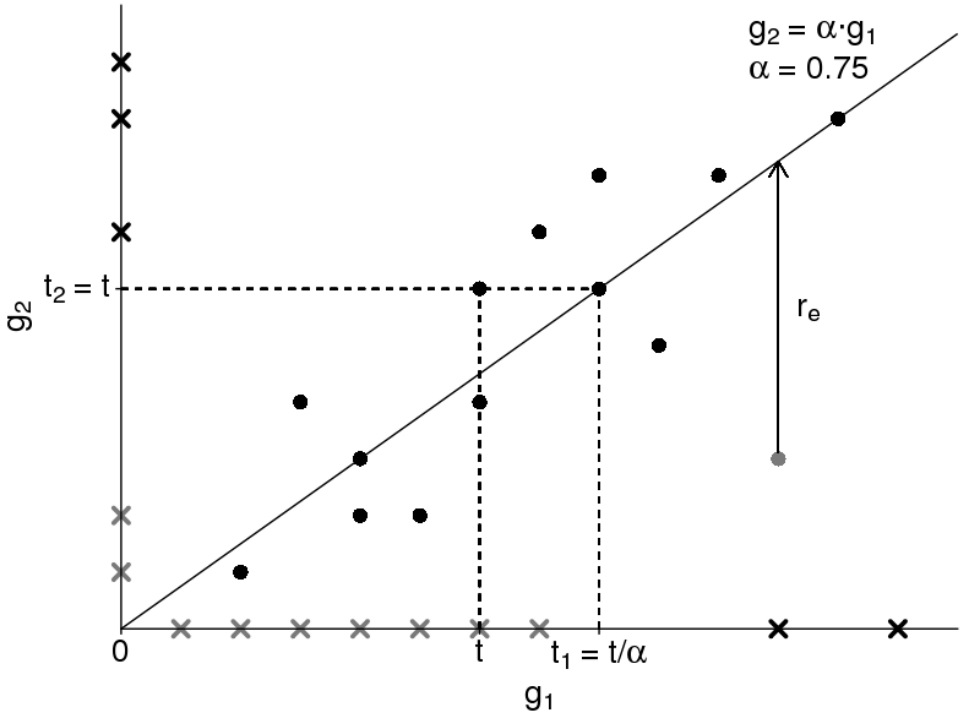


***Figure 14** : Illustration de la méthode pour comparer les profils d'abondance d'une paire de gènes*

Les comptages du gène g_2 sont comparés à ceux du gène g_1 dans un ensemble d'échantillon métagénomiques (données simulées). Le coefficient de proportionnalité α entre g_1 et g_2 est estimé à 0.75. La ligne pleine de pente α correspond aux comptages attendus. Les lignes en pointillés représentent les seuils de quantification t_1 et t_2 avant et après ajustement en fonction du coefficient de proportionnalité. Les croix noires et grises correspondent respectivement aux zéros structurels et aux zéros indéterminés. Les points noirs et gris correspondent respectivement aux échantillons « inliers » et aux échantillons « outliers » (valeurs aberrantes). La distance entre l'unique outlier et le comptage attendu est égal au résidu r_e.

3.3.5 Mesure de proportionnalité

L'accord de g_1 et g_2 avec une relation de proportionnalité directe de coefficient α est évalué en utilisant le coefficient de concordance de Lin [143]. Au préalable, une transformation racine carrée est appliquée aux comptages pour stabiliser leur variance et limiter l'asymétrie de leur distribution [131].

Le coefficient de concordance de Lin a été à l'origine conçu pour évaluer la reproductibilité soit une relation du type $y = x$. La formule de ce coefficient est :

$$\rho = \frac{2 \cdot cov(x,y)}{\sigma_x^2 + \sigma_y^2 + (\overline{x} - \overline{y})^2}$$

où $\bar{x}$ et $\bar{y}$ sont les moyennes, σ_x^2 and σ_y^2 les variances et $cov(x,y)$ la covariance de x et y.

Pour évaluer avec la concordance de Lin une relation de proportionnalité directe du type $y = k \cdot x$ où k est une constante, on doit ajuster l'échelle des variables x et y. Pour ce faire, on multiplie la variable x par k pour que x et y aient des comptages du même ordre de grandeur.

Dans la formule ci-dessus, on substitue x par $k \cdot x$. Comme $cov(k \cdot x, y) = k \cdot cov(x,y)$, $var(k \cdot x) = k^2 \cdot var(x)$ et $\overline{k \cdot x} = k \cdot \bar{x}$, on obtient:

$$\rho = \frac{2k \cdot cov(x,y)}{k^2 \cdot \sigma_x^2 + \sigma_y^2 + (k \cdot \overline{x} - \overline{y})^2}$$

Comme attendu, cette formule est symétrique. On aurait pu laisser la variable x inchangée et substituer la variable y par $\frac{1}{k} \cdot y$. En effet :

$$\rho = \frac{2k \cdot cov(x,y)}{k^2 \cdot \sigma_x^2 + \sigma_y^2 + (k \cdot \overline{x} - \overline{y})^2} =$$

$$\frac{\frac{1}{k^2} \cdot \left(2k \cdot cov(x,y)\right)}{\frac{1}{k^2} \cdot \left(k^2 \cdot \sigma_x^2 + \sigma_y^2 + k^2 \cdot \left(\overline{x} - \frac{1}{k} \cdot \overline{y}\right)^2\right)} =$$

$$\frac{\frac{2}{k} \cdot cov(x,y)}{\sigma_x^2 + \frac{1}{k^2} \cdot \sigma_y^2 + \left(\overline{x} - \frac{1}{k} \cdot \overline{y}\right)^2}$$

En substituant la variable x par g_1, la variable y par g_2 et la constante k par le coefficient de proportionnalité α précédemment estimé, on obtient la formule de la *mesure non-robuste de la proportionnalité* :

$$\rho_{nr} = \frac{2\alpha \cdot cov(g_1, g_2)}{\alpha \cdot \sigma_{g_1}^2 + \sigma_{g_2}^2 + (\alpha \cdot \overline{g_1} - \overline{g_2})^2}$$

Seuls les échantillons où les deux gènes ont des comptages non nuls sont utilisés pour calculer ρ_{nr}.

3.3.6 Détection des valeurs aberrantes et mesure robuste de la proportionnalité

La présence d'échantillons avec des comptages incohérents nommés ci-dessous valeurs aberrantes (outliers en anglais) peut faire décroitre significativement la mesure de la proportionnalité (**Figure 15**). Si les valeurs aberrantes ne sont pas traitées de façon appropriée, certaines associations pertinentes peuvent être manquées [144]. Ces valeurs aberrantes résultent du mélange de souches de la même espèce dans un échantillon ou d'erreurs commises lors de l'alignement des lectures sur le catalogue de gènes.

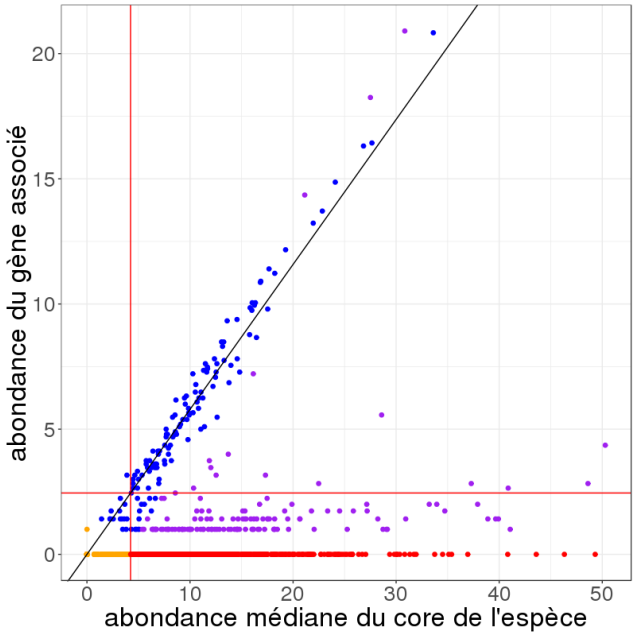


Figure 15 *: Comparaison du vecteur d'abondance médian du core génome de l'espèce Parabacteroides distasonis (axe des abscisses) et du vecteur d'abondance d'un de ses gènes accessoires (axe des ordonnées) dans les 1267 échantillons du catalogue IGC.*

Les points violets correspondent à des échantillons avec des comptages faibles très inférieurs à ceux attendus. En excluant les valeurs aberrantes, la mesure de proportionnalité passe de 0.14 à 0.97.

Pour faire face à ce problème, nous avons développé une version robuste de la mesure de la proportionnalité (ρ_r). Dans un premier temps, le coefficient de proportionnalité α est estimé en utilisant la procédure décrite ci-dessus (3.3.2). Ensuite, les résidus (r_e) définis comme la différence entre les comptages observés et ceux attendus en cas de proportionnalité sont calculés dans les échantillons où les deux gènes ont des comptages supérieurs à leur seuil de quantification respectif (**Figure 15**):

$$R = \{r_e = g_{2,e} - \alpha \cdot g_{1,e}\}\, e \in E \mid (g_{2,e} \geq t_2 \wedge g_{1,e} \geq t_1)$$

Par la suite, les valeurs aberrantes sont détectées avec la méthode de Tukey [145]. Notons respectivement Q_1 et Q_3 le premier et le troisième quartile des résidus calculés (R) et IQR l'écart interquartile défini par $IQR = Q_3 - Q_1$. Parmi l'ensemble d'échantillons E' où les deux gènes ont un comptage non nul, ceux avec des résidus supérieurs à $Q_3 + 1.5 \cdot \mathrm{IQR}$ ou inférieurs à $Q_1 - 1.5 \cdot \mathrm{IQR}$ sont étiquetés en tant que valeurs aberrantes (O). Formellement :

$Q_1 = 1^{er} quartile(R)$ et $Q_3 = 3^{\text{ème}} quartile(R)$

$IQR = Q_3 - Q_1$

$lwr\_thr = Q_1 - 1.5 \cdot IQR$ et $upr\_thr = Q_3 + 1.5 \cdot IQR$

$E' = \{e \in E \mid (g_{2,e} > 0 \wedge g_{1,e} > 0)\}$

$O = \{e \in E' \mid (r_e < lwr\_thr \vee r_e > upr\_thr)\}$ et $I = S' \backslash O$

Finalement, la mesure robuste de la proportionnalité p_r est calculée à partir des comptages des échantillons ne correspondant pas à des valeurs aberrantes (ensemble I défini dans le cadre ci-dessus) en utilisant la même formule que p_{nr} (3.3.5). Pour éviter la détection de fausses associations, cette

statistique n'est pas calculée si la proportion de valeurs aberrantes excède 30% soit formellement $|O| > (|S'| - 5) \cdot 0.3$.

3.3.7 Critère de co-occurrence pour la détection de fausses associations

Lors du calcul de la mesure de proportionnalité, nous avons fait le choix d'écarter les échantillons où au moins un des deux gènes comparés possède un comptage nul. Ainsi, on accroit la sensibilité car des relations de proportionnalité concernant uniquement un sous-ensemble d'échantillons peuvent être détectées. Malheureusement, ceci se fait au détriment de la spécificité car de fausses associations peuvent apparaitre. Ceci se produit quand il existe un très grand nombre de zéros structurels et peu d'échantillons au-dessus des seuils de quantification (**Figure 16**).

Jusqu'ici, l'unique critère pour décréter une paire de gènes comme associée était basé sur la co-abondance détectée avec la mesure de proportionnalité. Pour écarter les faux-positifs, on ajoute un critère basé sur la co-occurrence comparant les profils de présence/absence des gènes dans les différents échantillons. Ainsi, on peut évaluer l'interdépendance entre deux gènes. Intuitivement, on mesure à quel point il est inattendu d'observer simultanément une paire de gènes dans un ensemble d'échantillons.

Le critère de cooccurrence s'appuie sur le test exact de Fisher unilatéral à droite (alternative = "greater" sous R) [146]. Pour ce faire, on remplit la table de contingence suivante :

Gène 1 \ Gène 2	Présent	Absent
Présent	Nombre d'échantillons où les deux gènes sont détectés	Nombre d'échantillons où seul le gène 1 est détecté
Absent	Nombre d'échantillons où seul le gène 2 est détecté	Nombre d'échantillons où les deux gènes ne sont pas détectés

En pratique, on utilise les seuils de quantification pour considérer uniquement les échantillons où les gènes sont détectés avec certitude et écarter les zéros d'échantillonnage. En reprenant les notations utilisées en 3.3.1, la table de contingence utilisée pour le test de Fisher devient :

Gène 1 \ Gène 2	Présent	Absent
Présent	$\left\lvert\left\{ \begin{matrix} e \in E \text{ tels que} \\ g_{1,e} \geq t_1 \text{et } g_{2,e} \geq t_2 \end{matrix} \right\}\right\rvert$	$\left\lvert\left\{ \begin{matrix} e \in E \text{ tels que} \\ g_{1,e} \geq t_1 \text{et } g_{2,e} = 0 \end{matrix} \right\}\right\rvert$
Absent	$\left\lvert\left\{ \begin{matrix} e \in E \text{ tels que} \\ c_{1,e} = 0 \text{ et } c_{2,e} \geq t_2 \end{matrix} \right\}\right\rvert$	$\left\lvert\left\{ \begin{matrix} e \in E \text{ tels que} \\ g_{1,e} = 0 \text{ et } g_{2,e} = 0 \end{matrix} \right\}\right\rvert$

Plus la p-valeur obtenue est proche de 0, plus la probabilité que l'association observée soit due au hasard est faible. Nous avons fixé empiriquement le seuil de significativité d'une association à 10^{-10}.

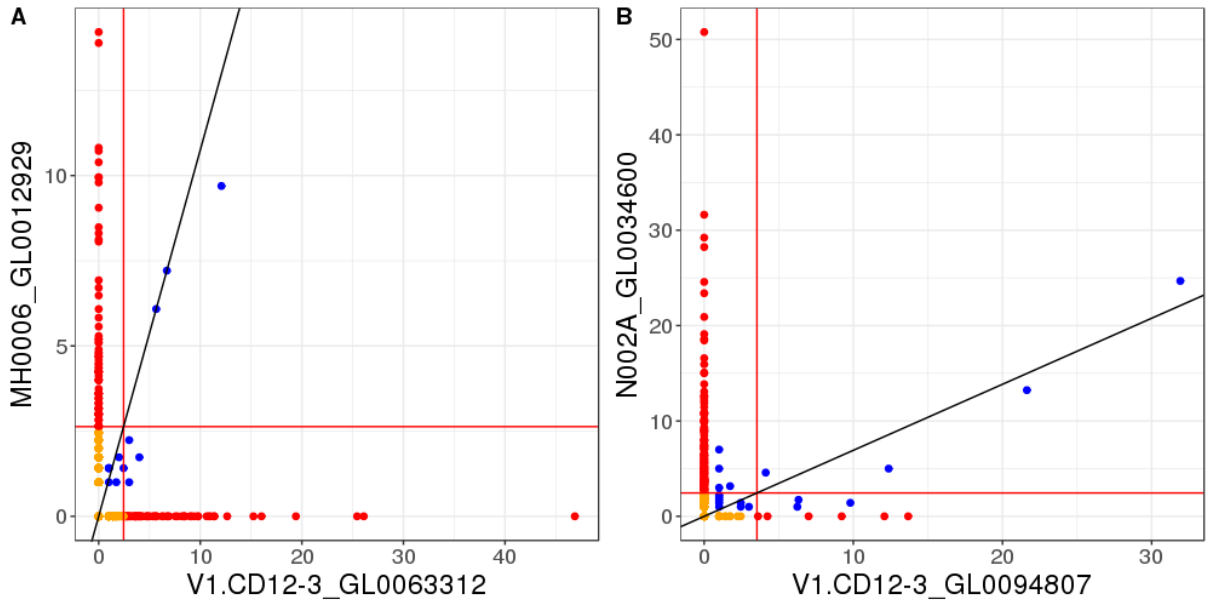


Figure 16 *: Exemples d'associations correspondant vraisemblablement à des faux positifs*

A. Comparaison des vecteurs d'abondance du gène V1.CD12-3_GL0063312 assigné à Blautia hansenii *(axe des abscisses) et du gène MH0006_GL0012929 (axe des ordonnées) assigné* à Blautia sp. Marseille-P3087 *quantifiés dans les 1267 échantillons du catalogue IGC. Les deux gènes sont détectés simultanément dans seulement 3 échantillons mais il existe 148 zéros structurels (84 pour l'axe des abscisses, 64 pour l'axe des ordonnées). La mesure non-robuste de la proportionnalité est de 0.90 mais la p-valeur du test exact de Fisher vaut 0.15.*

B. Comparaison du vecteur d'abondance du gène V1.CD12-3_GL0094807 assigné à Clostridium bolteae *(axe des abscisses) et du gène assigné à* Escherichia coli *(axe des ordonnées) quantifiés dans les 1267 échantillons du catalogue IGC. Les deux gènes sont détectés simultanément dans seulement 4 échantillons mais il existe 156 zéros structurels (6 pour l'axe des abscisses, 150 pour l'axe des ordonnées). La mesure non-robuste de la proportionnalité est de 0.88 mais la p-valeur du test exact de Fisher vaut 0.06.*

3.4 Evaluation des mesures de proportionnalité sur un jeu de données simulées

3.4.1 Création du jeu de données simulées

Pour évaluer les mesures de proportionnalité, nous avons généré une table d'abondance simulant les comptages des gènes d'une espèce virtuelle. Le pan-génome de cette espèce est composé de 1 000 gènes core détectés dans toutes les souches et de 6 000 gènes accessoires présents chez seulement certaines. Les longueurs des gènes ont été tirées aléatoirement (min = 100, max = 5 000 paires de bases) ainsi que la prévalence des gènes accessoires (min = 2,5%, max = 99,5%).

200 échantillons porteurs d'une souche différente d'une même espèce ont été générés. La couverture de séquençage d'une souche a été tirée à partir d'une loi uniforme (min=0.6, max=20). La longueur des lectures a été fixée à 100 paires de bases. Dans un échantillon donné, le nombre théorique de lectures alignées sur un gène est non nul s'il est présent dans la souche et est proportionnel à la longueur du gène et à la couverture de séquençage. Finalement, les comptages de gènes observés ont été tirés à partir de distributions de Poisson de moyennes égales aux comptages théoriques.

3.4.2 Comparaison aux coefficients de corrélation de Pearson et Spearman

Cette table a été utilisée pour comparer les performances du coefficient de corrélation de Pearson, du coefficient de corrélation de Spearman et de la mesure non-robuste de proportionnalité (p_{nr}) pour

détecter une relation entre le vecteur d'abondance du core génome de l'espèce et les vecteurs d'abondance de chacun de ses gènes, accessoires compris.

Cette simulation montre que les coefficients de corrélation Pearson et de Spearman décroissent d'autant plus que la prévalence du gène testé est faible tandis que p_{nr} reste élevée (supérieure à 0,8) et relativement stable (**Figure 17**). En effet, p_{nr} est calculée uniquement sur le sous-ensemble d'échantillons où le gène testé et le core génome de l'espèce sont simultanément détectés alors que les coefficients de corrélation prennent en compte tous les échantillons. Or, les échantillons où l'espèce est présente mais où le gène accessoire testé est absent génèrent des zéros structurels qui font décroitre le coefficient de corrélation d'autant plus que la prévalence du gène est faible. Par conséquent, des associations pertinentes entre le core génome d'une espèce et de nombreux gènes accessoires seront manquées en utilisant les coefficients de corrélations usuels. Dans l'exemple ci-dessous, les gènes accessoires dont la prévalence est inférieure à 75% ne seront pas associés au core génome de l'espèce si l'on fixe un seuil minimal d'inclusion à 0,8.

Néanmoins, les performances des coefficients de corrélation auraient été comparables à celle de p_{nr} si les zéros avaient été filtrés. Cependant, p_{nr} ne détecte que des relations de proportionnalité directe tandis que les coefficients de corrélation identifient soit des relations linéaires (corrélation de Pearson) soit n'importe quelle relation monotone (corrélation de Spearman). Ainsi, les relations identifiées grâce à p_{nr} le seront aussi avec les coefficients de corrélations usuels mais la réciproque n'est pas vraie. En effet, de fausses relations peuvent être détectées avec les coefficients de corrélation lorsque la prévalence des gènes comparés est faible. A l'inverse, la mesure de proportionnalité reconstitue le répertoire de gènes d'espèces rares et détecte des gènes accessoires peu prévalents tout en garantissant une spécificité élevée.

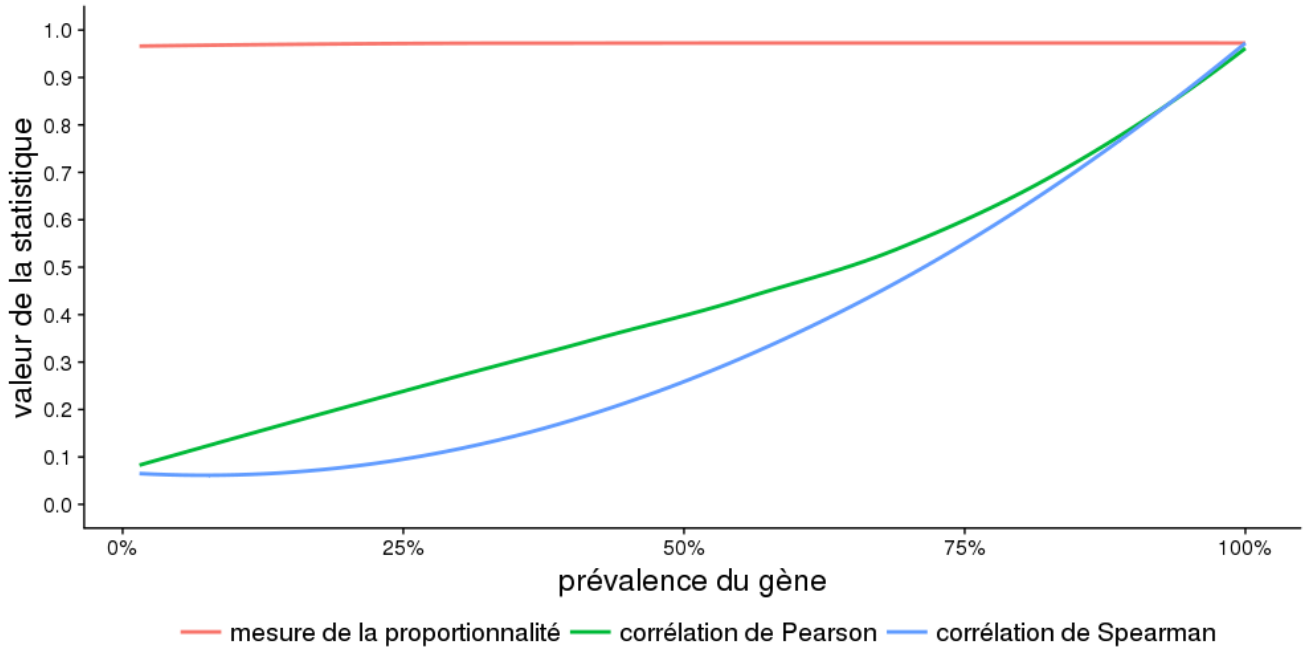


***Figure 17** : Comparaison des performances de la mesure non-robuste de proportionnalité (rouge), du coefficient de corrélation de Pearson (vert) et du coefficient de corrélation de Spearman (bleu) pour détecter un lien entre le vecteur d'abondance médian de l'espèce simulée et les vecteurs d'abondance de chacun de ses gènes.*

L'axe des abscisses correspond à la prévalence du gène, c'est-à-dire le pourcentage d'échantillons dans lequel il est détecté. Les gènes core ont une prévalence de 100% tandis que les autres sont des gènes accessoires d'autant plus rares que leur prévalence est faible.

L'axe des ordonnées correspond à l'intensité du lien détecté entre le vecteur d'abondance du gène et le vecteur d'abondance du core génome de l'espèce. Plus la valeur est proche de 1, plus l'intensité du lien est forte.

3.4.3 Impact de la longueur des gènes et de la couverture de séquençage

Dans un second temps, nous avons fait varier les paramètres de la simulation pour évaluer l'impact de la longueur des gènes et de la couverture de séquençage sur p_{nr}. Cette simulation montre que la sensibilité de p_{nr} est plus élevée pour les gènes longs et ceux ayant une couverture de séquençage variant fortement d'un échantillon à l'autre (**Figure 18**).

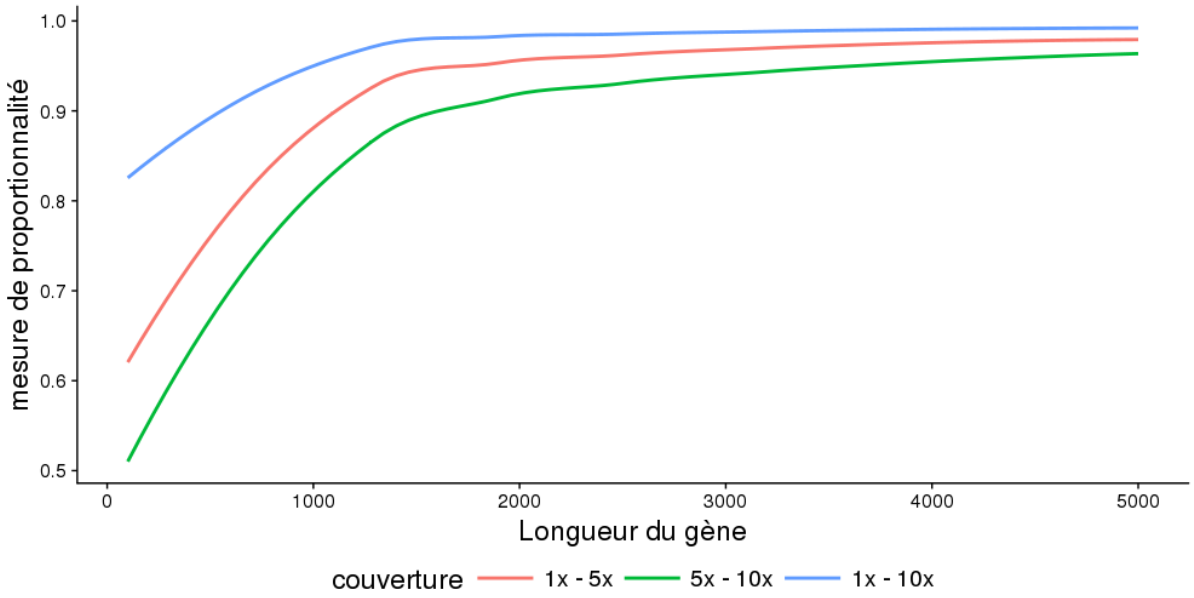


***Figure 18** : Impact de la longueur des gènes et de la couverture de séquençage sur p_{nr}*

Ce graphique représente l'intensité du lien détecté entre le vecteur d'abondance d'un gène core de l'espèce et le vecteur d'abondance médian de son core génome (axe des ordonnées) en fonction de la longueur du gène core testé (axe des abscisses). La couverture de séquençage des gènes core varie soit de 1x à 5x (rouge), soit de 5x à 10x (vert) ou soit de 1x à 10x (bleu).

p_{nr} est plus élevée pour les gènes longs car leurs comptages sont plus « étalés » et moins dispersés (**Figure 19** et **Tableau 6**). Remarquablement, une normalisation par la longueur des gènes n'atténue pas cet effet.

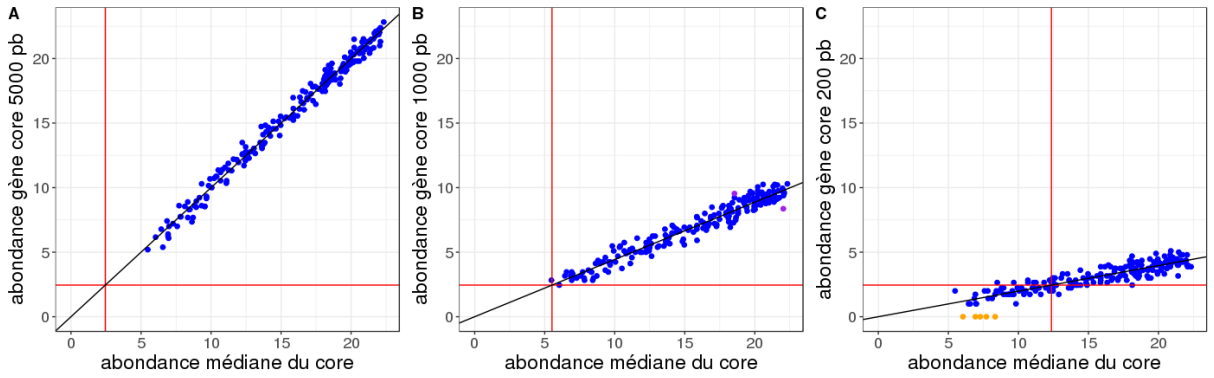


Figure 19 *: Comparaison du vecteur d'abondance médian du core génome de l'espèce virtuelle (axe des abscisses) avec les vecteurs d'abondance de 3 gènes core (axe des ordonnées) en échelle racine carrée.*

Les gènes core représentés en A, B et C ont pour longueurs respectives 5000, 1000 et 200 paires de bases. Plus les gènes sont courts, plus l'étendue de leurs comptages est faible et plus p_{nr} est faible. On remarque que le gène le plus court (C) a des comptages nuls dans certains échantillons où il est présent (points jaunes) : il s'agit de zéros d'échantillonnage.

Comparaison	Longueur du gène	Comptage minimum	Comptage maximum	Ecart type	Mesure de la proportionnalité (p_{nr})
A	1000 pb	5,1	22,8	4,7	0,99
B	5000 pb	2,4	10,3	2,1	0,97
C	200 pb	0	5,1	1,1	0,89

Tableau 6 *: On indique pour chacun des gènes décrits ci-dessus le comptage le plus faible, le comptage le plus fort, l'écart type des comptages et la mesure de proportionnalité. Plus le gène est long, plus l'écart type des comptages est grand et plus p_{nr} est élevée.*

De même, la mesure de proportionnalité est plus élevée pour les gènes dont la couverture est hautement variable car leurs comptages sont plus « étalés » (**Figure 20** et **Tableau 7**).

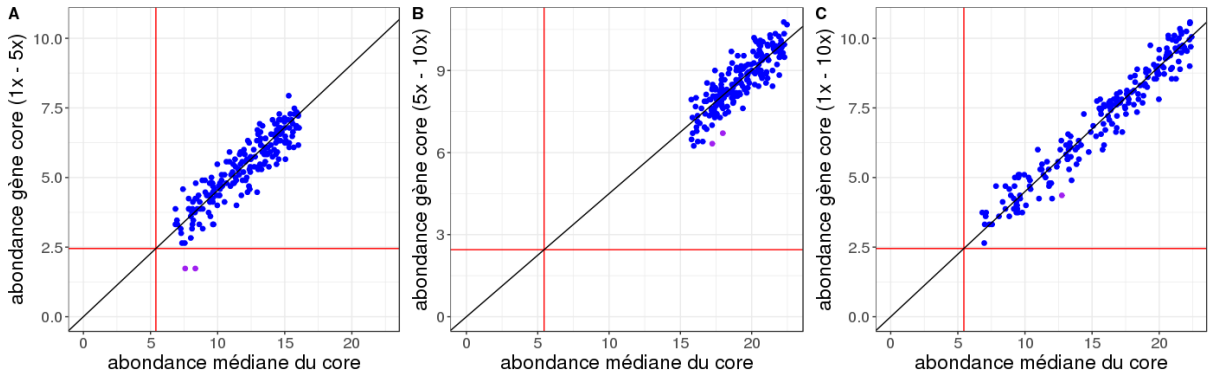


Figure 20 *: Comparaison du vecteur d'abondance médian du core génome de l'espèce (axe des abscisses) avec le vecteur d'abondance d'un gène core (axe des ordonnées) en échelle racine carrée.*

On fait varier la couverture de séquençage de l'espèce dans les différents échantillons (A = 1x – 5x ; B = 5x -10x ; C = 1x – 10x). Plus la couverture est variable, plus p_{nr} est forte.

Comparaison	Couverture	Comptage minimum	Comptage maximum	Ecart type	Mesure de la proportionnalité (p_{nr})
A	**1x – 5x**	2,6	7,8	1,3	0,92
B	**5x – 10x**	6,2	10,9	0,99	0,85
C	**1x – 10x**	3	10,8	1,9	0,96

***Tableau 7** : On indique pour chacun des gènes décrits ci-dessus le comptage le plus faible, le comptage le plus fort, l'écart type des comptages et la mesure de proportionnalité. Plus la couverture est variable, plus l'écart type des comptages est grand et plus p_{nr} est élevée.*

3.4.4 Comparaison de la version non-robuste et de la version robuste de la mesure de proportionnalité

Finalement, nous avons comparé les capacités de la mesure robuste (p_r) et de la mesure non robuste (p_{nr}) à identifier une relation de proportionnalité directe malgré des valeurs aberrantes. Pour ce faire, nous avons ajouté un pourcentage croissant de valeurs aberrantes (5%, 10% puis 20%) dans les vecteurs d'abondance de chaque gène core de l'espèce. Les valeurs aberrantes ont été générées en multipliant les comptages des gènes par ¼, ⅓, 2 ,3 ou 4. Pour un pourcentage de valeurs aberrantes donné, nous avons comparé le vecteur d'abondance bruité de chaque gène core au vecteur d'abondance médian non bruité du core de l'espèce en utilisant soit p_r soit p_{nr}.

Cette simulation montre que p_{nr} décroit d'autant plus que le pourcentage de valeurs aberrantes est important alors que p_r reste élevée (**Figure 21**). Ainsi, p_r permet de détecter un lien de proportionnalité malgré la présence de valeurs aberrantes.

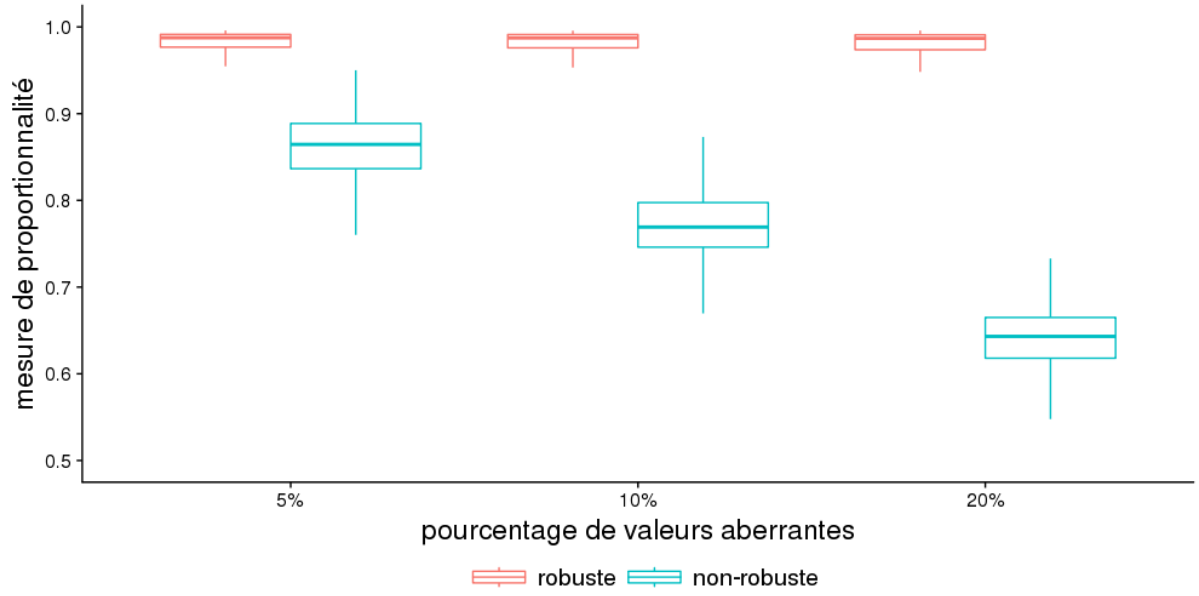


***Figure 21** : Capacités de la mesure non-robuste (boxplots bleus) et de la mesure robuste (boxplots rouges) pour identifier une relation de proportionnalité directe entre les vecteurs d'abondance bruité des gènes core de l'espèce simulée et le vecteur d'abondance médian non bruité du core.*

On augmente progressivement la proportion de valeurs aberrantes à 5%, 10% puis 20% (axe des abscisses). L'axe des ordonnées correspond à l'intensité du lien détecté entre le vecteur d'abondance bruité du gène et le vecteur d'abondance non bruité du core génome de l'espèce.

4. Reconstitution *in silico* de pan-génomes microbiens

4.1 Méthodes

Dans cette partie, nous décrivons la méthode qui reconstitue des pan-génomes microbiens en regroupant des gènes aux comptages proportionnels. Cette méthode s'appuie sur les mesures de proportionnalité (p_{nr}et p_r) décrites précédemment.

4.1.1 Regroupement des gènes co-abondants et co-occurrents

La première étape consiste à rassembler les gènes ayant des signaux très similaires. Parmi ces groupes de gènes, ceux de grande taille ont une forte probabilité de correspondre aux gènes core d'une espèce microbienne. Plus précisément, on rassemble au sein d'un même groupe les gènes ayant des comptages proportionnels (co-abondance) et détectés dans les même échantillons (co-occurrence). Ces groupes de gènes nommés *seeds* par la suite sont créées grâce à l'algorithme glouton suivant :

1. Toutes les paires de gènes sont comparées deux-à-deux en utilisant la mesure non-robuste de la proportionnalité (p_{nr}). Ces comparaisons sont sauvegardées dans une liste (**Figure 22.1**).
2. Seules les paires de gènes dont p_{nr} est supérieur ou égal à 0.8 et pour lesquelles aucun échantillon n'est classifié comme zéro structurel sont conservées dans la liste (**Figure 22.2**).
3. Les paires de gènes sélectionnées sont triées par p_{nr} décroissante (**Figure 22.3**).
4. La paire de gènes ayant le p_{nr} le plus élevé est sélectionnée comme centroïde (gènes en gras dans la **Figure 22.4**). Les gènes associés à un des gènes du centroïde sont groupés dans une nouvelle seed
5. Les paires de gènes contenant au moins un gène présent dans la nouvelle seed sont supprimées de la liste. On réitère la procédure d'agglomération (étape 4) jusqu'à ce que la liste soit vide (**Figure 22.5**).
6. Finalement, les seeds dont la taille est inférieure à 30 gènes sont écartées.

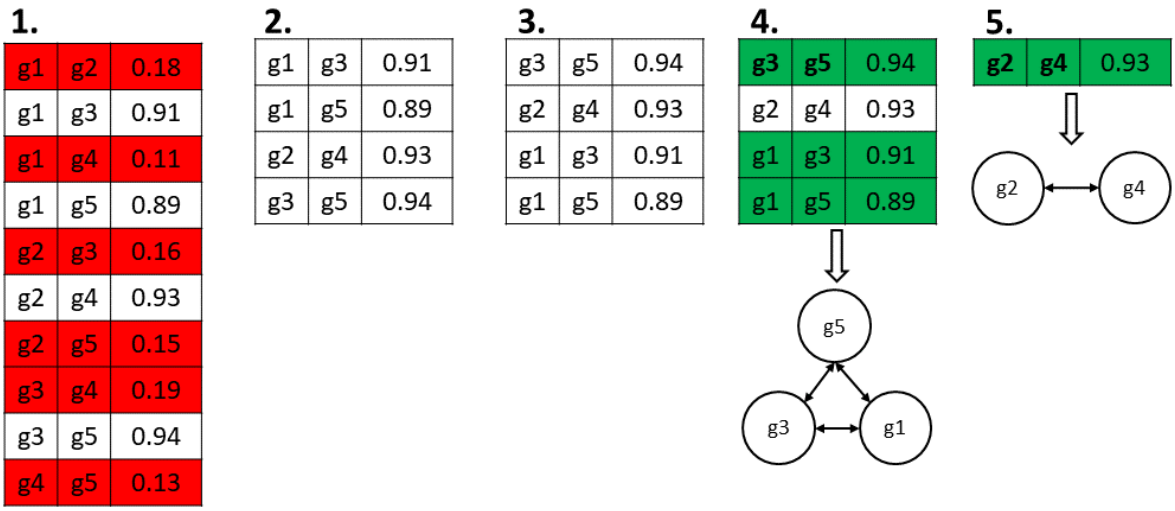


***Figure 22** : Illustration des différentes étapes de la création des seeds décrites ci-dessus.*

Ici, on considère 5 gènes (g1, g2, g3, g4 et g5) regroupés en deux seeds distinctes. La première seed contient les gènes g1, g3 et g5, la deuxième les gènes g2 et g4. Pour simplifier, on considère que tous les gènes d'une même seed sont liés. Les paires de gènes en gras correspondent aux centroïdes.

4.1.2 Calcul du représentant d'une seed

Comme une seed regroupe des gènes avec un signal très similaire, on peut résumer son contenu grâce à un seul gène que l'on nommera *représentant*. Par la suite, ce représentant sera utilisé pour comparer des seeds deux à deux.

Le calcul du représentant de la seed s'effectue grâce à l'algorithme suivant :

1. Le représentant de la seed est défini comme le vecteur médian des comptages de l'ensemble de ses gènes.
2. L'ensemble des gènes de la seed est comparé à son représentant en utilisant la mesure non-robuste de la proportionnalité (p_{nr}).
3. Le représentant final de la seed correspond au vecteur médian des comptages des 30 gènes avec la p_{nr} la plus élevée. En pratique, ces gènes sont ceux avec le signal le plus fort et le moins dispersé (voir 3.4.3)

Le nombre de gènes sélectionnés pour calculer le représentant de la seed (30 par défaut) a été fixé empiriquement. Néanmoins, nous avons vérifié que l'utilisation de valeurs différentes (20 - 40 - 50) modifie marginalement les résultats obtenus en aval.

Lorsqu'une seed possède moins de 30 gènes mais plus de 10, elle est représentée par le vecteur médian des comptages de l'ensemble de ses gènes. Si elle possède moins de 10 gènes, son représentant est un des deux gènes qui a servi de centroïde lors de sa création car il s'agit de la paire de gènes ayant la mesure de la proportionnalité la plus élevée.

4.1.3 Pré-regroupement des gènes

La complexité de l'algorithme décrit ci-dessus est quadratique car toutes les paires de gènes sont comparées. En effet, si g correspond au nombre de gènes alors $\frac{g\cdot(g-1)}{2}$ comparaisons sont effectuées.

Il n'est pas envisageable d'effectuer toutes les comparaisons deux à deux lorsque le nombre de gènes à traiter atteint plusieurs millions. Le temps de calcul serait bien trop important à moins de disposer d'une infrastructure de calcul conséquente à laquelle peu de laboratoires ont accès. De plus, une majorité des comparaisons effectuées seraient inutiles car peu de gènes sont associés deux à deux.

Une stratégie s'appuyant sur le patron de programmation Map/Reduce (**Figure 23**) permet de réduire sensiblement le nombre de comparaison à effectuer. L'étape Map répartit les gènes dans plusieurs paniers. Dans chaque panier, on crée des seeds en utilisant l'algorithme en 4.1.3. Néanmoins, cette étape peut placer dans des paniers différents des gènes associés qui auraient dû être regroupés dans la même seed. Pour remédier à ce problème, l'étape Reduce récupère d'abord les seeds créées dans l'ensemble des paniers puis fusionne celles éclatées par l'étape Map.

Notons g le nombre de gènes traités et n le nombre de paniers créés. Si chaque panier possède le même nombre de gènes, on effectue $\frac{\left(\frac{g}{n}\right)\cdot\left(\frac{g}{n}-1\right)}{2}$ comparaisons dans chacun d'entre eux soit $n\cdot\frac{\left(\frac{g}{n}\right)\cdot\left(\frac{g}{n}-1\right)}{2}=\frac{1}{n}\cdot\frac{g\cdot(g-n)}{2}$ comparaisons au total. Si le nombre de panier est très inférieur au nombre de gènes alors $\frac{1}{n}\cdot\frac{g\cdot(g-n)}{2}\approx\frac{1}{n}\cdot\frac{g\cdot(g-1)}{2}$. On en déduit que l'étape Map divise le nombre de comparaisons proportionnellement au nombre de paniers. Néanmoins, le nombre de paniers ne doit pas être trop important car on risque d'éclater les seeds en de multiples seeds de petite taille et ainsi de rendre l'étape de fusion très couteuse.

Répartir les gènes aléatoirement est la solution la plus simple pour obtenir des paniers de taille équivalente. Cependant, nous avons choisi de regrouper dans un même panier les gènes dont le comptage le plus élevé provient du même échantillon. On obtient ainsi un nombre de paniers égal au nombre d'échantillons. Cette répartition est plus astucieuse car elle augmente la probabilité de placer dans un même panier des gènes qui seront regroupés dans une seed. On limite ainsi la fragmentation des seeds ce qui réduit ainsi le nombre de fusions à réaliser en aval.

Toutefois, cette stratégie répartit les gènes de façon déséquilibrée entre les différents paniers. En effet, les paniers associés à un échantillon avec une profondeur de séquençage élevée agglomèreront plus de gènes car les comptages y sont globalement plus élevés. Pour supprimer ce biais, on divise les vecteurs de comptages de chaque gène par le nombre de total de lectures alignées dans les échantillons. Grâce à cette normalisation, on s'attend à ce que la taille des paniers soit indépendante de la profondeur de séquençage. De plus, les gènes devraient être répartis plus équitablement entre les différents paniers. Notons que les comptages normalisés sont utilisés uniquement pour répartir les gènes dans les différents paniers et pas dans les étapes ultérieures.

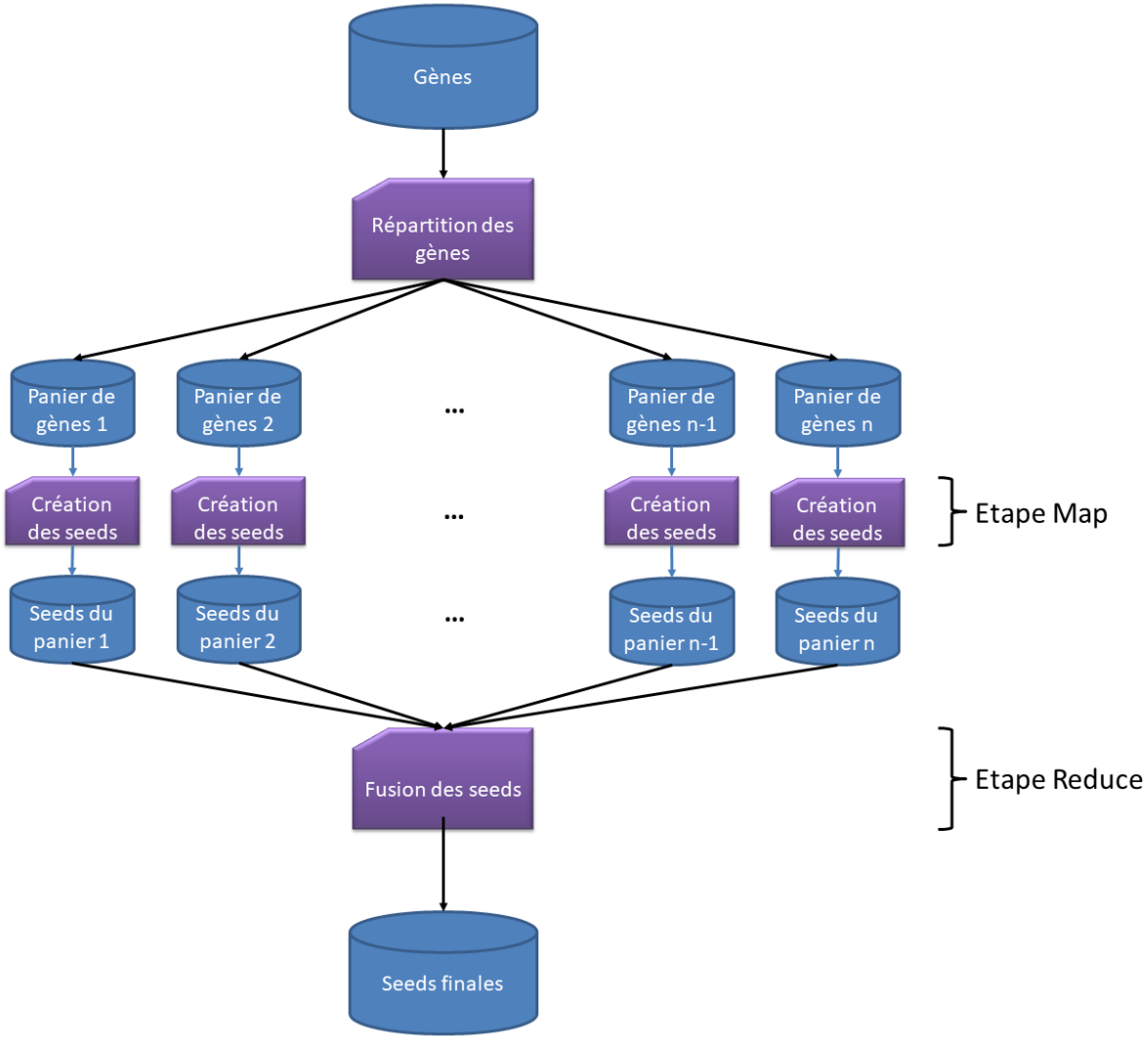


***Figure 23** : Schéma de la stratégie Map/Reduce mise en œuvre. Les cylindres bleus correspondent aux données, les rectangles violets aux traitements.*

4.1.4 Fusion des seeds

Comme expliqué précédemment, l'étape de pré-regroupement peut placer dans différents paniers des gènes associés qui devraient être regroupés dans la même seed. Ceci se produit par exemple lorsque les échantillons avec les comptages les plus forts ont des valeurs proches. L'étape Reduce résout ce problème en fusionnant les seeds fragmentées.

La fusion des seeds s'effectue grâce à l'algorithme suivant :

1. Les seeds générées dans chaque panier sont regroupées, triées par nombre de gènes décroissant puis sauvegardées dans une liste de seeds à traiter.
2. La seed qui contient le plus de gènes est sélectionnée comme centroïde puis est supprimée de la liste de seeds à traiter.
3. Le représentant de la seed centroïde est comparé aux représentants respectifs des seeds à traiter avec la mesure non robuste de proportionnalité (ρ_{nr}).
4. Les seeds à traiter avec $\rho_{nr} \geq 0.8$ et pour lesquelles aucun échantillon n'est classifié comme zéro structurel sont fusionnées avec la seed centroïde dans une nouvelle seed.
5. On calcule le représentant de la nouvelle seed (voir 4.1.2) puis on sauvegarde cette dernière dans une liste des seeds finales.
6. On supprime les seeds fusionnées de la liste des seeds à traiter. S'il reste des seeds à traiter, on retourne à l'étape 2.
7. Les seeds finales dont la taille est inférieure à 150 gènes sont écartées.

4.1.5 Identification des seeds cores

Parmi les seeds finales, celles regroupant des gènes core d'espèces microbiennes seront nommées seeds core. Les autres seeds finales peuvent être classifiées dans 2 catégories.

La première correspond à des seeds associées à des seed core. Il s'agit soit de gènes accessoires, soit de gènes partagés par plusieurs espèces ou éventuellement de prophages. La deuxième correspond à des seeds indépendantes dont le signal ne peut être associé à celui de seeds core. Il peut s'agir de virus ou d'éléments génétiques mobiles.

En ne gardant que les seeds dont la taille excède 150 gènes, on suppose que l'on a éliminé celles appartenant à la deuxième catégorie. Ensuite, on fait l'hypothèse que parmi un ensemble de seeds finales, la plus grande correspond à une seed core. Il s'agit du critère utilisé dans l'algorithme suivant pour identifier les seeds core :

1. Les seeds finales sont triées par nombre de gènes décroissant puis sauvegardées dans une liste de seeds core potentielles.
2. La seed la plus grande est sauvegardée dans la liste des seeds core.
3. Le représentant de la nouvelle seed core est comparé aux représentants respectifs des seeds core potentielles en utilisant la mesure non robuste de proportionnalité ρ_{nr}
4. Les seeds à traiter avec $\rho_{nr} \geq 0.8$ et une p-valeur au test exact de Fisher inférieure à 10^{-10} sont considérées comme associées à la nouvelle seed core.
5. Les seeds associées sont supprimées de la liste seeds core potentielles.
6. S'il reste des seeds à traiter, on retourne à l'étape 2.

4.1.6 Récupération des gènes associés

Récupérer uniquement les gènes provenant de seeds associés à une seed core n'est pas suffisant. En effet, de nombreux gènes associés proviennent de seeds de moins de 150 gènes qui ont été éliminées.

D'autres gènes présents dans un sous-ensemble d'échantillons qui leur est spécifique et forment par conséquent des seeds singletons.

Pour récupérer le plus possible de gènes associés, les représentants de chaque seed core sont comparés à tous les gènes du catalogue. Ceux dont la mesure robuste de la proportionnalité (ρ_r) est supérieure ou égale à 0.8 et dont la p-valeur au test exact de Fisher inférieure à 10^{-10} sont considérés comme associés à la seed core.

4.1.7 Classification des gènes associés

Soient $g_1 = (g_{1,e_1}, g_{1,e_2}, \dots, g_{1,e_s})$ et $g_2 = (g_{2,e_1}, g_{2,e_2}, \dots, g_{2,e_s})$ les vecteurs contenant respectivement le nombre de lectures alignées sur le représentant d'une seed core et sur un de ses gènes associés.

Le gène associé est assigné à une des 4 catégories suivant la présence de zéros structurels :

1. Gène core : le gène associé est présent dans tous les échantillons où la seed core est détectée et uniquement dans ceux-là (**Figure 24.A**) :

$$\forall e \in E \mid (g_{1,e} \geq t_1 \rightarrow g_{2,e} \neq 0) \wedge (g_{2,e} \geq t_2 \rightarrow g_{1,e} \neq 0)$$

2. Gène accessoire : le gène associé est présent uniquement dans un sous-ensemble d'échantillons où la seed core est détectée (**Figure 24.B**) :

$$(\exists e \in E \mid g_{1,e} \geq t_1 \wedge g_{2,e} = 0) \wedge (\forall e \in E \mid g_{2,e} \geq t_2 \rightarrow g_{1,e} \neq 0)$$

3. Gène core partagé : le gène associé est détecté dans tous les échantillons où la seed core est présente et dans d'autres échantillons où la core seed est absente (**Figure 24.C**) :

$$(\forall e \in E \mid g_{1,e} \geq t_1 \rightarrow g_{2,e} \neq 0) \wedge (\exists e \in E \mid g_{2,e} \geq t_2 \wedge g_{1,e} = 0)$$

4. Gène accessoire partagé : le gène associé est détecté dans un sous-ensemble d'échantillons où la seed core est présente et dans d'autres échantillons où la seed core est absente (**Figure 24.D**).

$$(\exists e \in E \mid g_{1,e} \geq t_1 \wedge g_{2,e} = 0) \wedge (\exists e \in E \mid g_{2,e} \geq t_2 \wedge g_{1,e} = 0)$$

Le terme *gène partagé* peut porter à confusion car un tel gène n'est pas forcément associé à plusieurs MSPs. Il s'agit en réalité d'un gène détecté dans des échantillons où le core de la MSP ne l'est pas. Ainsi, le terme *gène non spécifique* aurait pu convenir.

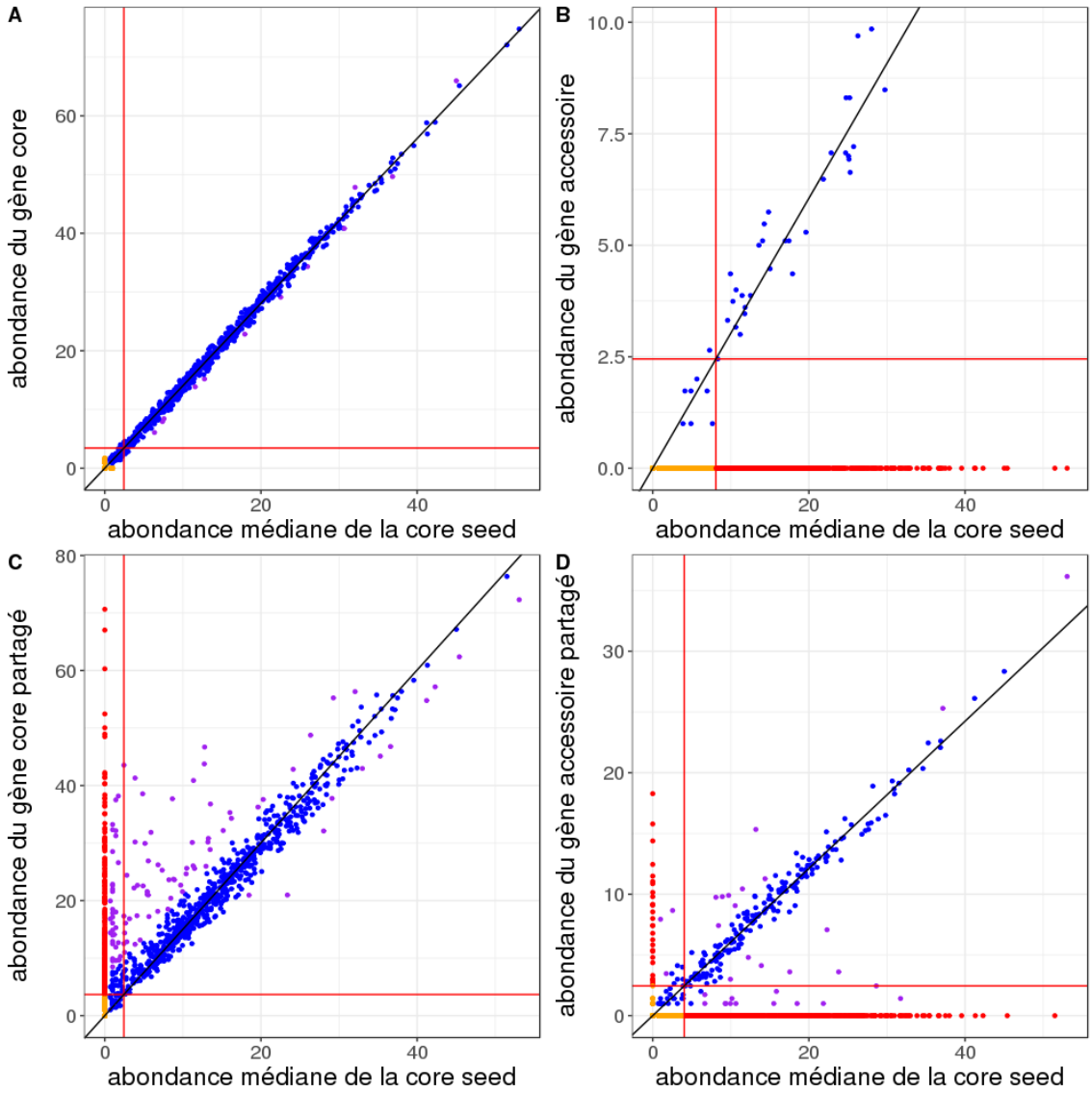


Figure 24 *: Illustration des 4 catégories de gènes dans une MSP.*

Le vecteur d'abondance médian des 30 meilleurs gènes représentatifs d'une seed core (axe des abscisses) est comparé aux vecteurs d'abondance de 4 gènes qui lui sont associés (axe des ordonnées). Les gènes sont quantifiés dans les 1 267 échantillons du catalogue IGC et les abondances sont représentés suivant une échelle racine carrée. La ligne noire a pour pente le coefficient de proportionnalité α. Les lignes rouges verticales et horizontales correspondent aux seuils de quantification. Elles ont pour équations respectives $y = t_1$ *et* $x = t_2$*. Les points bleus sont sélectionnés pour calculer la mesure robuste de proportionnalité alors que les points violets sont exclus car classifiés en tant valeurs aberrantes. Les points jaunes et rouges correspondent respectivement aux zéros structurels et aux zéros indéterminés. Seuls les zéros structurels sont utilisés pour affecter les gènes à une catégorie donnée.*

A. *Le gène est classifié comme core. Il est présent dans tous les échantillons où la seed core est détectée et uniquement dans ceux-là.* ***B.*** *Le gène est classifié comme accessoire. Il est présent dans un sous-ensemble d'échantillons (7.2%) où la seed core est détectée.* ***C.*** *Le gène est classifié comme core partagé. Il est présent dans tous les échantillons où la seed core est détectée ainsi que dans 286 échantillons où la seed core est absente.* ***D.*** *Le gène est classifié comme accessoire partagé. Il est présent dans un sous-ensemble d'échantillons (33.5%) où la seed core est détectée ainsi que dans 28 échantillons où la seed core est absente.*

4.1.8 Création des MSPs

Les gènes core, accessoires, core partagés et accessoires partagés associés à une seed core sont assemblés dans un objet appelé MSP pour Metagenomic Species Pan-genome (**Figure 25**).

Les gènes core sont comparés au représentant de la seed core et triés par mesure de la proportionnalité décroissante. Dans chaque catégorie à l'exception des gènes core, une procédure similaire à celle utilisée pour la création des seeds est exécutée (voir 4.1.1). On identifie ainsi des modules de gènes co-occurrents. Ces gènes interdépendants peuvent par exemple correspondre à des opérons ou à des gènes chevauchants. Les gènes non clustérisés sont sauvegardés comme des modules singletons.

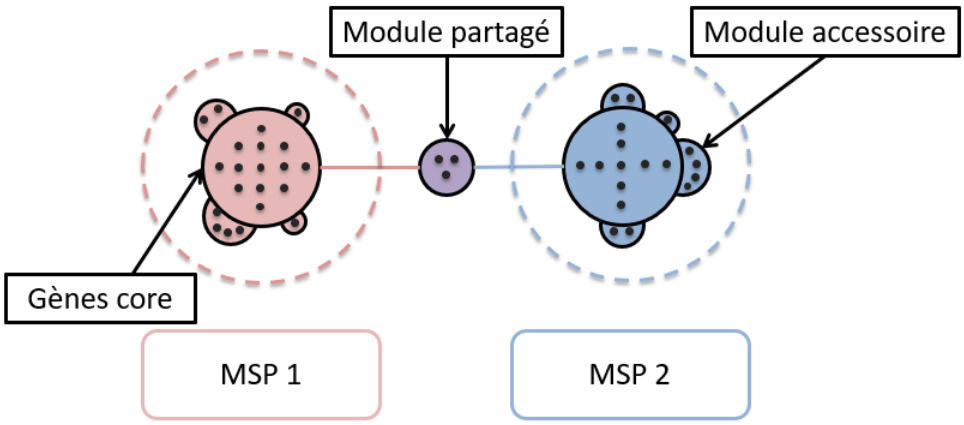


***Figure 25** : Structure des MSPs (Metagenomic Species Pan-genomes)*

On représente ici deux MSPs : la MSP 1 est couleur corail et la MSP 2 est bleue. Chaque point correspond un gène. Les gènes regroupés dans des cercles à trait plein représentent des modules. Les 2 MSPs partagent le module violet.

4.1.9 Implémentation

La méthode décrite ci-dessus a été implémentée dans un programme écrit en C++ nommé MSPminer (pour Metagenomic Species Pan-genomes miner). Le langage C++ permet de générer un exécutable natif exploitant au mieux les ressources matérielles disponibles sur une machine. Lorsque cela était possible, certaines parties du programme ont été parallélisées en utilisant l'interface de programmation OpenMP [147]. Ainsi, on lance plusieurs fils d'exécutions (threads) pour exploiter pleinement le potentiel des processeurs multicœurs.

Les comptages des gènes sont stockés sous la forme de flottants simple précision. Cependant, les calculs sont effectués en double précision car MSPminer manipule des nombres dont les ordres de grandeur sont très variables. On limite ainsi les erreurs d'arrondi.

Une attention toute particulière a été portée à assurer la reproductibilité des résultats. A partir d'un même jeu de données, MSPminer produit exactement les mêmes MSPs d'une exécution à l'autre et ce quel que soit le nombre de threads lancés ou la machine sur laquelle il est exécuté. Pour ce faire, on utilise uniquement des algorithmes déterministes et stables. Des barrières de synchronisation assurent que les threads écrivent les résultats toujours dans le même ordre.

4.2 Evaluation du pré-regroupement des gènes

L'utilisation de MSPminer sur un jeu de données réel confirme l'impact positif du pré-regroupement des gènes (4.1.3) sur les performances. Celui-ci diminue non seulement le nombre de comparaisons à

effectuer (**Figure 26.A**) mais augmente la probabilité que des gènes co-abondants et co-occurrents soient placés dans le même panier par comparaison avec une répartition aléatoire (**Figure 26.B**).

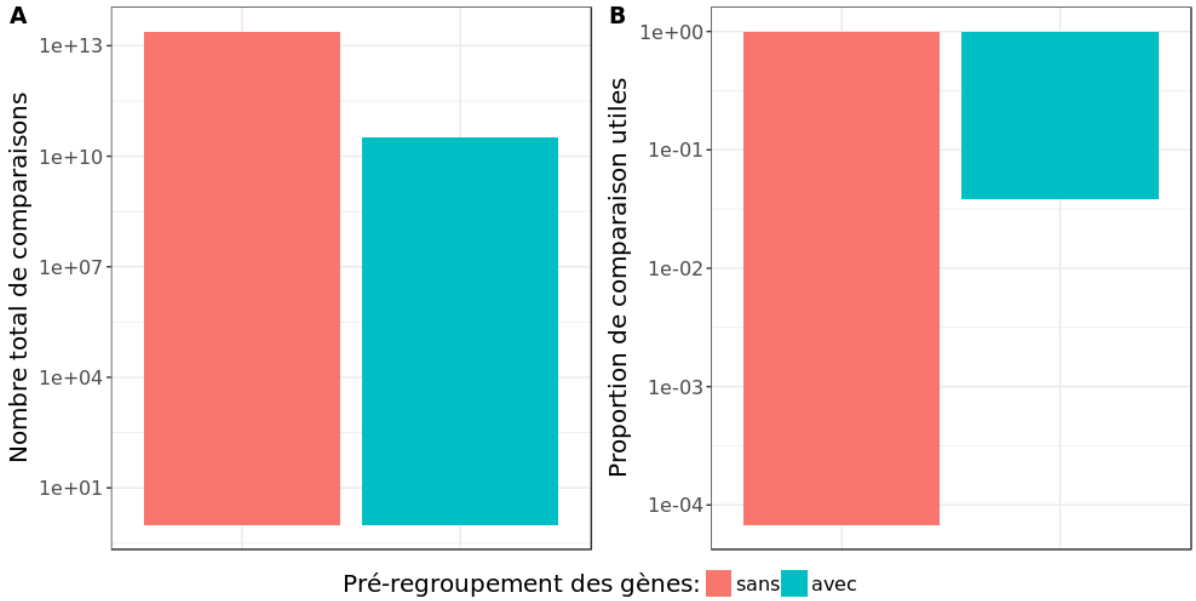


***Figure 26** : Performance du pré-regroupement de gènes sur un jeu de données réel. On considère la matrice de comptage du catalogue IGC composée de 6 971 229 gènes quantifiés sur 1267 échantillons.*

A. Nombre total de paires de gènes comparées avec ou sans pré-regroupement. Le pré-regroupement des gènes divise le nombre de comparaison par 763.

B. Proportion de comparaisons utiles pour la création de seeds avec ou sans pré-regroupement des gènes. Le pré-regroupement augmente la proportion de comparaisons utiles par 562. Pour rappel, une comparaison utile met en jeu une paire de gènes co-abondants et co-occurrents. La proportion sans pré-regroupement a été estimée par une simulation de Monte Carlo où un milliard de paires de gènes ont été tirées aléatoirement.

Comme attendu, la taille des paniers est indépendante du nombre de lecture mappées dans les échantillons après normalisation (**Figure 27**).

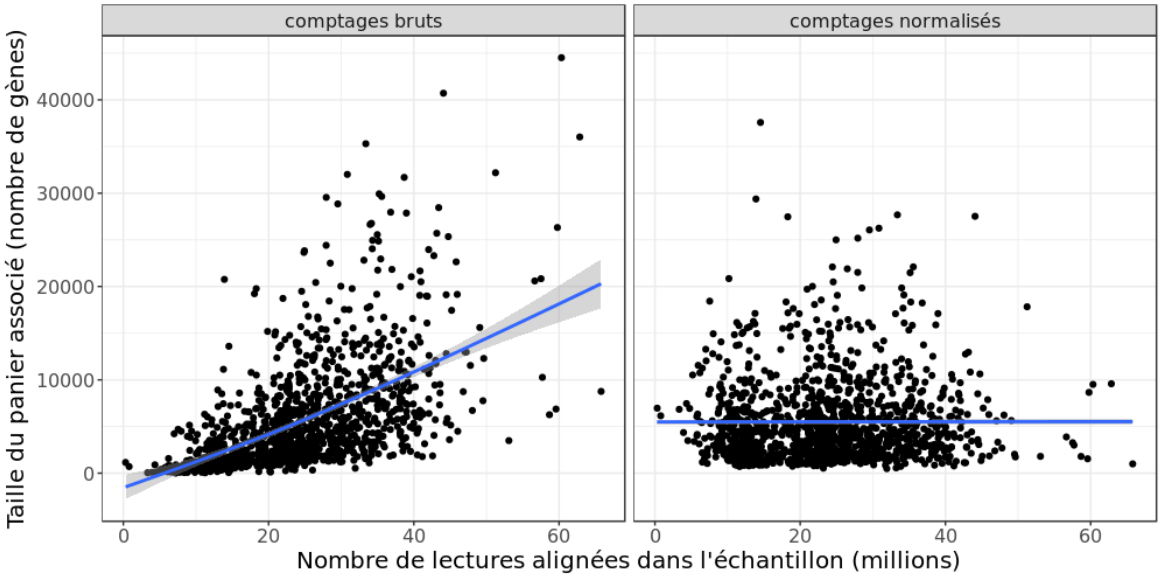


Figure 27 *: Taille du panier associé à un échantillon en fonction du nombre de lectures dans ce dernier.*

Les 1267 échantillons du catalogue IGC ont été considérés. Les gènes détectés dans moins de 3 échantillons ont été filtrés.

A. Paniers générés à partir de comptages bruts. La courbe de tendance indique que les paniers contiennent d'autant plus de gènes que le nombre de lectures alignées dans l'échantillon est important. Ceci est confirmé par le coefficient de corrélation de Spearman (rhô = 0.68, p-valeur <2.10^{16})

B. Paniers générés à partir de comptages normalisés. La courbe de tendance indique qu'il n'y a plus de lien entre le nombre de gènes dans un panier et le nombre de lectures alignées dans l'échantillon. Ceci est confirmé par le coefficient de corrélation de Spearman (rhô = 0.04, p-valeur = 0.145)

On constate que la taille des différents paniers est plus équilibrée après normalisation par la profondeur de séquençage. Ainsi, la charge de calcul est mieux distribuée entre les différents fils d'exécution (**Figure 28**).

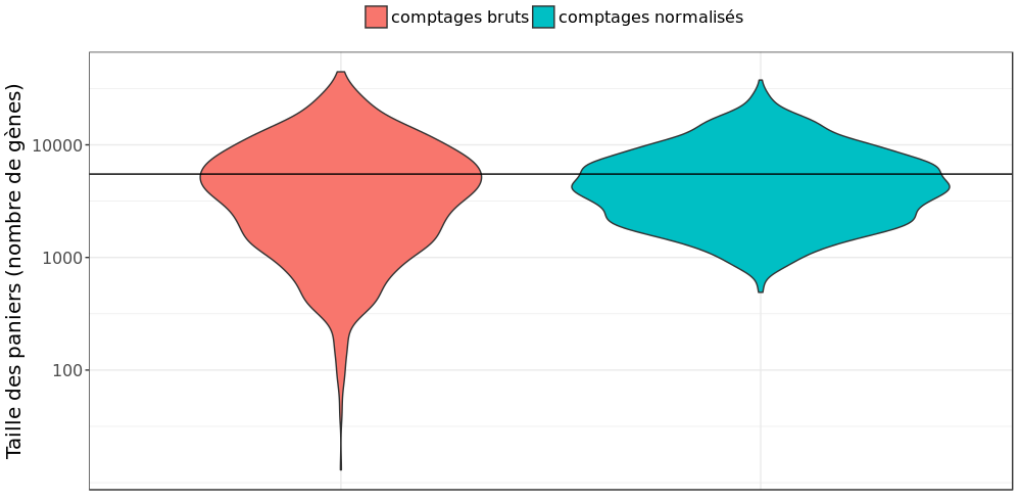


Figure 28 *Taille des paniers obtenus en utilisant des comptages bruts ou des comptages normalisés par le nombre de lectures alignées dans chaque échantillon.*

Les 1267 échantillons du catalogue IGC [113] *ont été considérés. Les gènes détectés dans moins de 3 échantillons ont été filtrés.*

La droite noire correspond à la taille de panier optimale (6 971 229 gènes / 1 267 paniers =5 502 gènes).

4.3 Evaluation de la méthode de clustering

Pour évaluer la méthode de clustering implémentée dans MSPminer, nous avons réutilisé le pangénome de l'espèce simulée décrite en 3.4.1

4.3.1 Impact de la prévalence de l'espèce

Dans un premier temps, nous avons généré plusieurs tables d'abondance des gènes de l'espèce en faisant progressivement décroitre le nombre d'échantillons où celle-ci est détectée (200, 100 puis 50).

Cette simulation montre que la prévalence de l'espèce a peu, voire pas d'impact sur le clustering de ses gènes core et accessoires fortement prévalents. Ainsi, on s'attend à ce qu'une MSP soit générée pour toute espèce détectée dans au moins 3 échantillons. Cependant, les gènes accessoires moins prévalents sont groupés dans la MSP de l'espèce seulement lorsque celle-ci est présente dans un nombre suffisamment important d'échantillons (**Figure 29**).

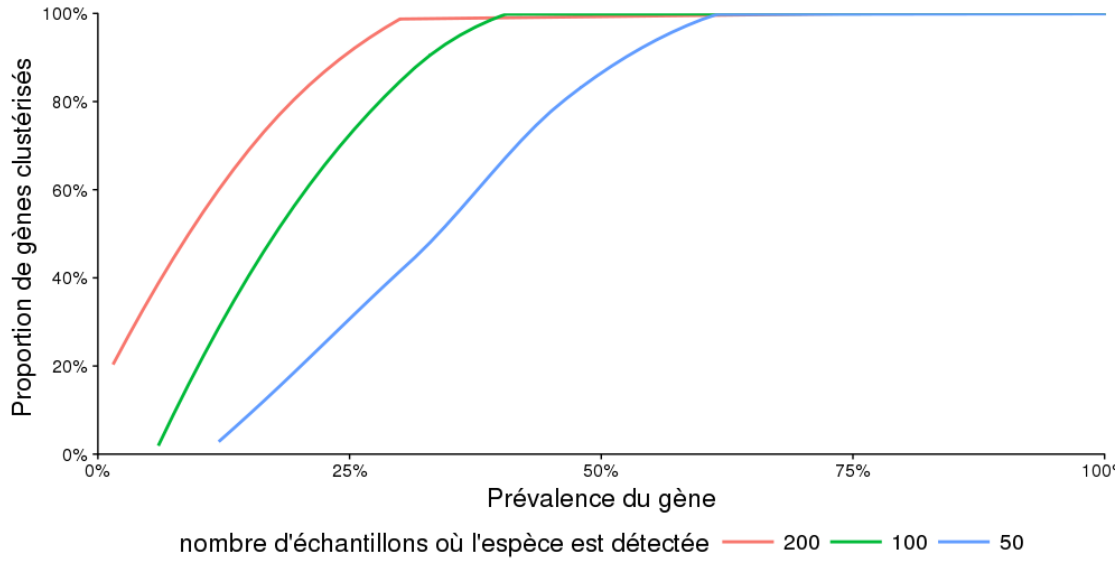


***Figure 29** : Impact du nombre d'échantillons où une espèce est détectée sur la complétion de sa MSP*

Ce graphique représente la proportion de gènes clustérisés dans la MSP associée à l'espèce (axe des ordonnées) en fonction de leur prévalence (axe des abscisses). Les gènes core ont une prévalence de 100%. Les autres gènes sont des gènes accessoires d'autant plus rares que leur prévalence est faible. Dans cette simulation, on fait décroitre progressivement le nombre d'échantillons où le core génome de l'espèce est détecté (couleur du trait).

4.3.2 Impact du mélange de souches

Dans un second temps, nous avons simulé des échantillons porteurs de deux souches de la même espèce. Nous avons supposé que la souche dominante est 5 à 10 fois plus abondante que la sous-dominante comme observé dans de nombreux échantillons fécaux [98]. Lorsqu'il est occasionnel (< 50% des échantillons), le mélange de souches a peu d'impact sur la performance du clustering. Cependant, si le mélange de souches est fréquent, de nombreux gènes accessoires de prévalence faible ou intermédiaire sont manqués (**Figure 30**).

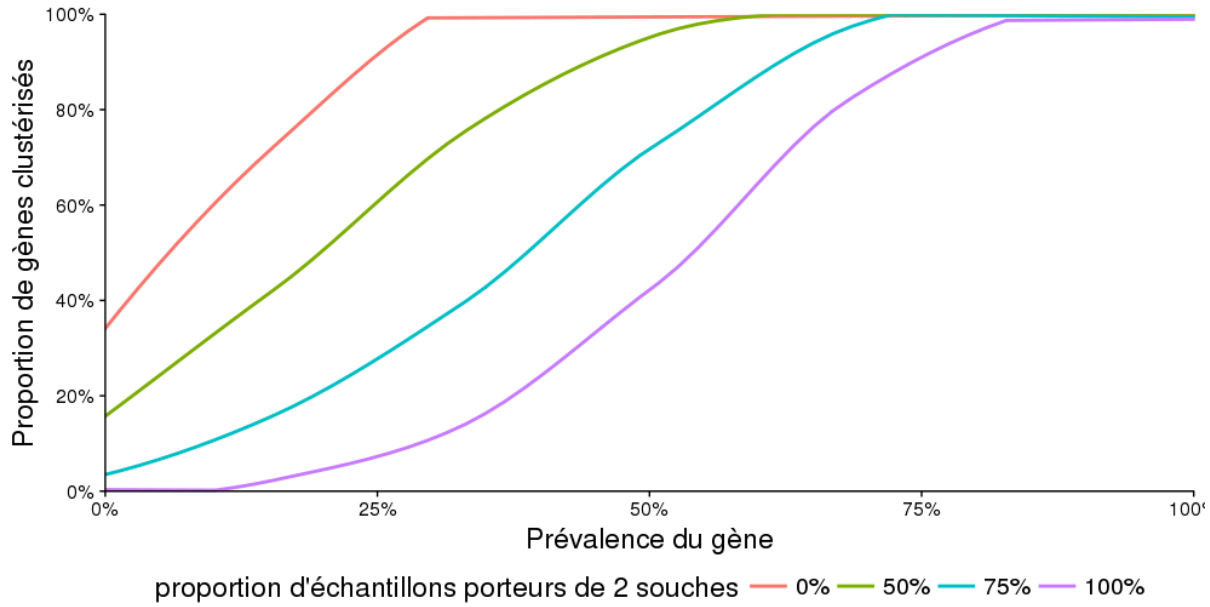


Figure 30 *: Impact du mélange de souches sur le clustering*

Ce graphique représente la proportion de gènes clustérisés dans la MSP associée à l'espèce (axe des ordonnées) en fonction de leur prévalence (axe des abscisses). Les gènes core ont une prévalence de 100%. Les autres gènes sont des gènes accessoires d'autant plus rares que leur prévalence est faible.

Dans cette simulation, on augmente progressivement la proportion d'échantillons porteurs de deux souches de l'espèce (couleur du trait).

En effet, lorsque qu'un gène accessoire n'est présent que dans la souche dominante d'un échantillon, son abondance est légèrement plus faible que s'il avait aussi été présent dans la souche sous-dominante. Dans la majorité des cas, un tel échantillon ne sera pas classifié comme valeur aberrante car la différence d'abondance est trop faible pour perturber la détection de la proportionnalité par ρ_r. Cependant, si un gène accessoire n'est présent que dans la souche sous-dominante, son abondance sera nettement plus faible que s'il avait aussi été présent dans la souche dominante. Par conséquent, cet échantillon sera classifié comme valeur aberrante.

Ainsi, plus la proportion d'échantillons porteurs de deux souches de l'espèce est importante, plus la proportion de valeurs aberrantes sera grande. Si la proportion de valeurs aberrantes est inférieure à 30%, ρ_r détectera et filtrera les échantillons concernés (points violets **Figure 31.B** et **Figure 31.C**). Si cette proportion dépasse 30%, ρ_r ne sera pourra pas détecter la relation de proportionnalité directe (**Figure 31.D**).

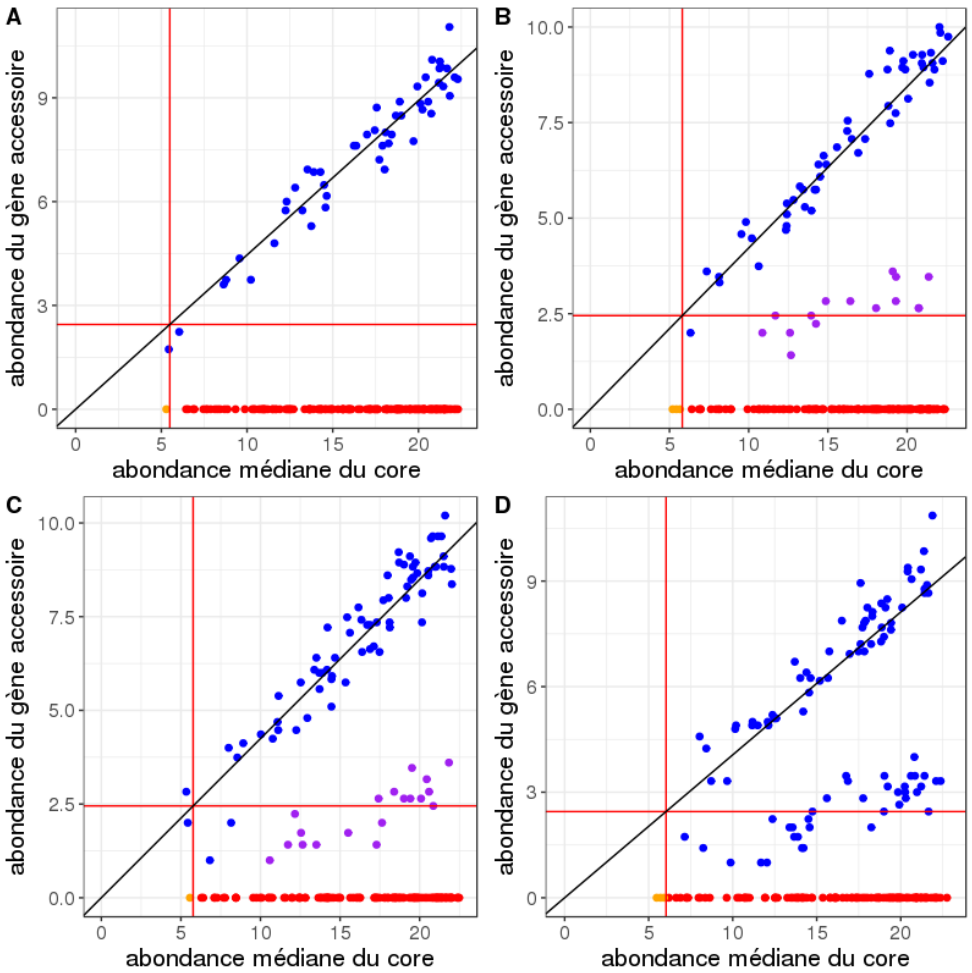


Figure 31 *: Comparaison du vecteur d'abondance médian du core génome de l'espèce (axe des abscisses) avec le vecteur d'abondance d'un gène accessoire présent dans 25% des échantillons (axe des ordonnées).*

On augmente progressivement la proportion d'échantillons porteurs de deux souches (A = 0% ; B = 50%; C = 75%, D = 100%). Le tableau ci-dessous montre que plus la proportion d'échantillons porteurs de deux souches est importante, plus la proportion de valeurs aberrantes augmente (B et C) jusqu'à ce que la proportionnalité ne soit plus détectée (D).

Comparaison	Proportion d'échantillons porteurs de deux souches	Proportion de valeurs aberrantes	Mesure non-robuste de proportionnalité	Mesure robuste de proportionnalité
A	0%	0%	0.96	0.96
B	50%	23%	0.56	0.97
C	75%	27%	0.49	0.95
D	100%	NA	0.30	0.30

4.4 Evaluation des performances de MSPminer

4.4.1 Temps de calcul

Grâce aux choix techniques effectués (langage de programmation, parallélisme) et aux différentes optimisations mises en œuvre, MSPminer traite des tables de comptages de très grande dimension composées de plusieurs millions de gènes et de centaines (voire milliers) d'échantillons en quelques heures sur un seul serveur de calcul (**Tableau 8**).

MSPminer tire avantage des architectures multicœur. L'efficacité du parallélisme est bonne (> 0.85) mais n'atteint pas 1 car certaines parties ne sont pas parallélisées et la charge de calcul n'est pas répartie de façon parfaitement équitable. Au niveau des barrières de synchronisation, les threads auxquels on a affecté la charge de calcul la plus faible attendent ceux qui ont la charge la plus importante. De plus, sur la machine de test, le nombre de cœurs logiques excède le nombre de cœurs physiques ce qui peut mener à sous-estimer l'efficacité réelle du programme.

Nombre de threads (fils d'exécution)	**1**	**2**	**4**	**6**	**8**	**10**	**12**
Temps réel de calcul (minutes)	2129	1078	571	396	304	244	206
Accélération	1	1.97	3.73	5.38	7.00	8.73	10.33
Efficacité du parallélisme	1	0.99	0.93	0.90	0.88	0.87	0.86

***Tableau 8** : Performances du parallélisme implémenté dans MSPminer*

Les tests ont été effectués sur une machine dotée d'un processeur Intel® Xeon® E5-2630 (6 cœurs physiques) et 128Go de mémoire vive. La donnée d'entrée est la matrice de comptage du catalogue IGC composée de 9.9M gènes quantifiés sur 1267 échantillons.

On mesure le temps d'exécution réel (wall-time) avec la commande UNIX time. L'accélération correspond au temps de calcul avec n threads divisé par celui avec un seul thread. L'efficacité est égale à l'accélération divisée par le nombre de threads. Dans l'idéal, le temps de calcul est inversement proportionnel au nombre de threads. Dans ce cas, l'accélération est alors égale au nombre de threads et l'efficacité vaut 1.

4.4.2 Consommation de mémoire vive

MSPminer est relativement gourmand en RAM car l'intégralité de la matrice de comptage est chargée en mémoire. Ainsi, la consommation de mémoire vive est proportionnelle au nombre d'échantillons et au nombre de gènes (**Figure 32**).

Actuellement, il n'est pas envisageable de traiter une table d'abondance de gènes de grande taille sur un ordinateur de bureau car la quantité de mémoire vive insuffisante. Par exemple, une machine de

calcul disposant d'au moins 60 Go de RAM est requise pour traiter la matrice du catalogue IGC qui quantifie 9 879 896 millions de gènes dans 1 267 échantillons.

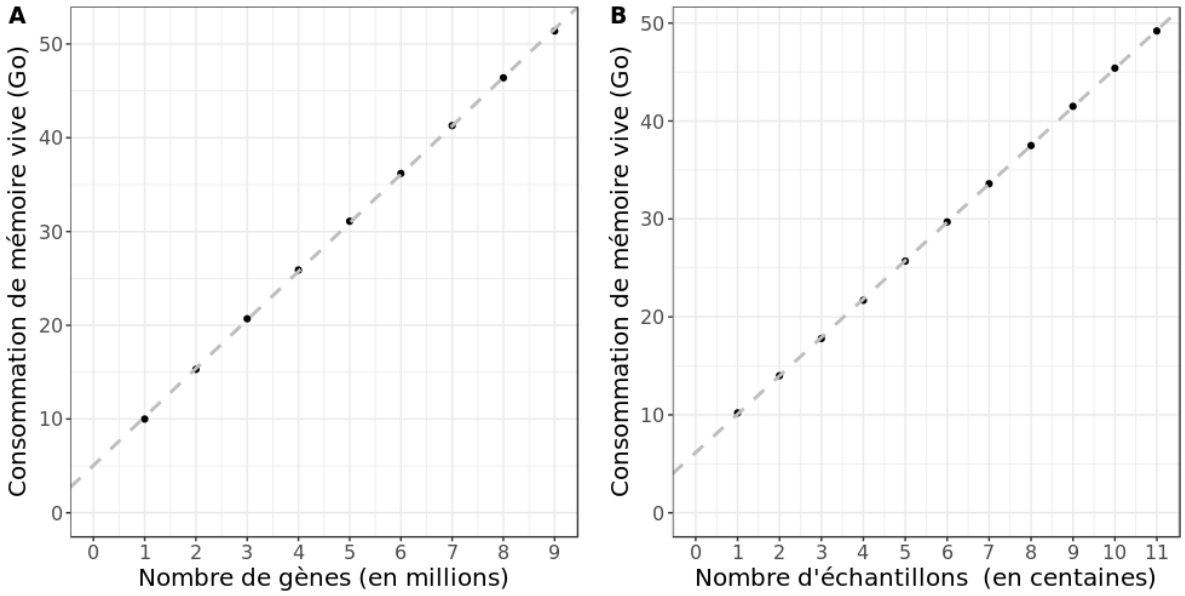


***Figure 32** : Consommation de mémoire vive de MSPminer*

A. Consommation de mémoire vive (axe des ordonnées) en fonction du nombre de gènes (axe des abscisses). Le nombre d'échantillons est fixé à 1267.

B. Consommation de mémoire vive (axe des ordonnées) en fonction du nombre d'échantillons (axe des abscisses). Le nombre de gènes est fixé à 9 879 896.

Les droites en pointillés gris correspondent au meilleur ajustement linéaire (A: $y = 5.17 \cdot x + 5.05$; B: $y = 3.92 \cdot x + 6.14$)

5. Compendium du microbiote intestinal humain

MSPminer a été appliqué à la table d'abondance publiée avec le catalogue de gènes du microbiote intestinal humain IGC [148]. Il s'agit de la plus grande table disponible publiquement à ce jour : 9 879 896 gènes y sont quantifiés dans 1267 échantillons de fèces prélevés chez 1018 individus différents ; certains individus ayant été prélevés à des temps différents. Ces individus ont une grande diversité phénotypique de par leur origine géographique (européens, chinois, états-uniens) et leur état de santé (sains, diabétiques, obèses, atteints de maladies inflammatoires chroniques de l'intestin etc.)

La majorité des gènes du catalogue IGC ont été prédits sur des contigs provenant d'assemblages *de novo* d'échantillons métagénomiques de différentes études (HMP, MetaHIT etc.). Les autres gènes proviennent de 511 génomes d'espèces procaryotes du microbiote intestinal. Ainsi, on obtient des gènes de meilleure qualité et moins fragmentés. De plus, on couvre des espèces détectables par séquençage métagénomique global mais trop peu abondantes pour être correctement assemblées. Le catalogue IGC est non redondant. Les gènes ayant plus de 95% d'identité nucléotidique sur 80% de leur longueur sont regroupés et sont représentés dans le catalogue par celui le plus long.

6 971 229 (70.6%) gènes avec des comptages supérieurs ou égaux à 6 dans au moins 3 échantillons ont été conservés par MSPminer. Parmi les gènes conservés, 3 288 928 (47,2%) ont été organisés en 1 661 MSPs avec au moins 150 gènes.

5.1 Taxonomie

5.1.1 Annotation taxonomique des MSPs

L'annotation taxonomique des MSPs a été effectuée en alignant leurs gènes core et accessoires contre les bases nt et WGS du NCBI (version de Septembre 2017) avec BLASTn [149] (version 2.7.1). Pour accélérer le traitement, nous avons restreint les comparaisons aux taxons *Bacteria*, *Archaea*, *Fungi*, *Virus* et *Blastocystis*. Aussi, nous avons utilisé l'algorithme megablast (word_size = 16) qui est significativement plus rapide malgré une légère baisse de sensibilité. Finalement, les 20 meilleurs résultats ont été conservés pour chaque gène.

Une annotation niveau espèce a été donnée aux MSPs dont au moins 50% des gènes correspondaient à la souche type de l'espèce avec une identité moyenne supérieure à 95% et une couverture supérieure à 90%. Les MSPs restantes ont été assignées aux rangs de plus haut niveau (genre, famille, ordre, classe, phylum et domaine) si au moins 50% de leurs gènes avaient la même annotation.

5.1.2 Diversité taxonomique des MSPs

Lorsque l'on considère le rang taxonomique le plus bas assigné aux MSPs, 642 (38,7 %) sont annotées au niveau espèce, 315 (19,0 %) au niveau genre, 525 (31,6 %) à un rang taxonomique de plus haut niveau (famille, ordre, classe, phylum et domaine) et 179 (10,8%) n'ont pas d'annotation. Ceci montre que la majorité des espèces du microbiote intestinal humain n'ont pas de génome déposé dans les bases de données publiques.

Les MSPs annotées sont pour leur très grande majorité des bactéries (99%) représentées par les phyla *Firmicutes* (1 016 MSPs), *Bacteroidetes* (263 MSPs), *Proteobacteria* (94 MSPs) et *Actinobacteria* (46 MSPs). Parmi les 1% des MSPs restantes, une est associée *Homo sapiens*, 4 sont eucaryotes unicellulaires du genre *Blastocystis* et 8 sont des archées.

Parmi les 642 MSPs annotées au niveau espèce, seules 304 correspondent à des espèces bien définies validées par le *Code International de la Nomenclature Bactérienne* (ICNB). 56 MSPs sont associées à des génomes de souches isolées et cultivées en laboratoire avec un nom d'espèce imprécis (sp., c.f.)

ou pas encore validé par l'ICNB (Candidatus). Les 282 MSPs restantes sont associées à des génomes reconstruits à partir d'échantillons métagénomiques (*CAG* ou *MGS*) ou résultant d'un séquençage de cellule unique (*single-cell*).

Finalement, la majorité des MSPs assignées à des espèces bien définies correspondent à des souches types selon l'ICNB et/ou à des génomes de référence dans la base RefSeq.

5.2 Phylogénie

5.2.1 Construction de l'arbre phylogénétique

Pour étudier les liens de parenté entre les espèces du microbiote intestinal, nous avons construit un arbre phylogénétique à partir des 1 661 MSPs et de 360 génomes représentatifs des espèces identifiées lors de l'annotation taxonomique.

Pour cela, nous avons extrait 40 gènes marqueurs phylogénétiques universels dans les génomes et les MSPs avec l'outil fetchMG [123]. Les MSPs possédant moins de 5 marqueurs ont été écartées car elles ne peuvent être placées avec suffisamment de précision dans l'arbre. Les gènes associés au même marqueur ont été concaténés puis alignés avec MUSCLE [150]. Chaque alignement a été rogné avec trimAl [151] pour extraire les régions les plus pertinentes pour la construction de l'arbre. Les 40 alignements ont ensuite été fusionnés en insérant un gap lorsqu'un marqueur manquait dans une MSP ou un génome. Finalement, l'arbre phylogénétique a été calculé avec FastTreeMP [152] puis visualisé avec l'application web iTOL [153].

5.2.2 Diversité phylogénétique des MSPs

Au total, 1502 (90,4%) des MSPs sont placées dans l'arbre phylogénétique. Les 59 MSPs restantes n'ont pas été considérées car elles possédaient moins de 5 gènes marqueurs non dupliqués.

Globalement, la phylogénie obtenue est cohérente avec l'annotation taxonomique des MSPs. En effet, les blocs uniformément colorés indiquent que les MSPs et génomes assignés au même phylum sont regroupés dans une même branche de l'arbre (**Figure 33**)

Néanmoins, l'arbre met en évidence quelques erreurs dans l'annotation taxonomique. Par exemple, la msp_0314 a été assignée à l'espèce *Faecalibacterium prausnitzii* (phylum *Firmicutes*) car ses gènes core avaient un très fort pourcentage d'identité (98,7%) avec la souche *F. prausnitzii* 2789STDY5834930. Or, la msp_0314 matérialisée dans l'arbre par une bande orange en haut à gauche (Firmicutes) est insérée dans le bloc bleu (*Bacteroides*). Ceci montre que la msp_0314 est mal annotée et qu'elle ne correspond pas à *F. prausnitzii* mais à une espèce du phylum *Bacteroides*. En effet, le génome *F. prausnitzii* 2789STDY5834930 est vraisemblablement mal annoté et cette erreur s'est propagée à la msp_0314 lors du processus d'annotation taxonomique.

De plus, la phylogénie apporte des informations complémentaires à la taxonomie (**Figure 34**) car 145 MSPs sans annotation matérialisées par des bandes grises sont placées dans l'arbre. Par exemple, celles insérées dans le bloc coloré orange appartiennent vraisemblablement au phylum *Firmicutes*. A ce jour, ces améliorations et corrections de l'annotation taxonomique sont faites manuellement sur des critères visuels mais on pourrait envisager de les automatiser pour traiter des niveaux taxonomiques plus bas que le phylum.

Enfin, nous avons ajouté un cercle coloré externe indiquant si les MSPs correspondent à une espèce isolée, cultivée et séquencée. Les régions majoritairement vertes correspondent à des groupes d'espèces déjà caractérisés. Les nombreuses régions à bandes rouges correspondent à taxons regroupant des espèces qui n'ont pas encore été isolées et cultivées. Ceci illustre l'intérêt de l'approche métagénomique et des MSPs pour caractériser le microbiote intestinal.

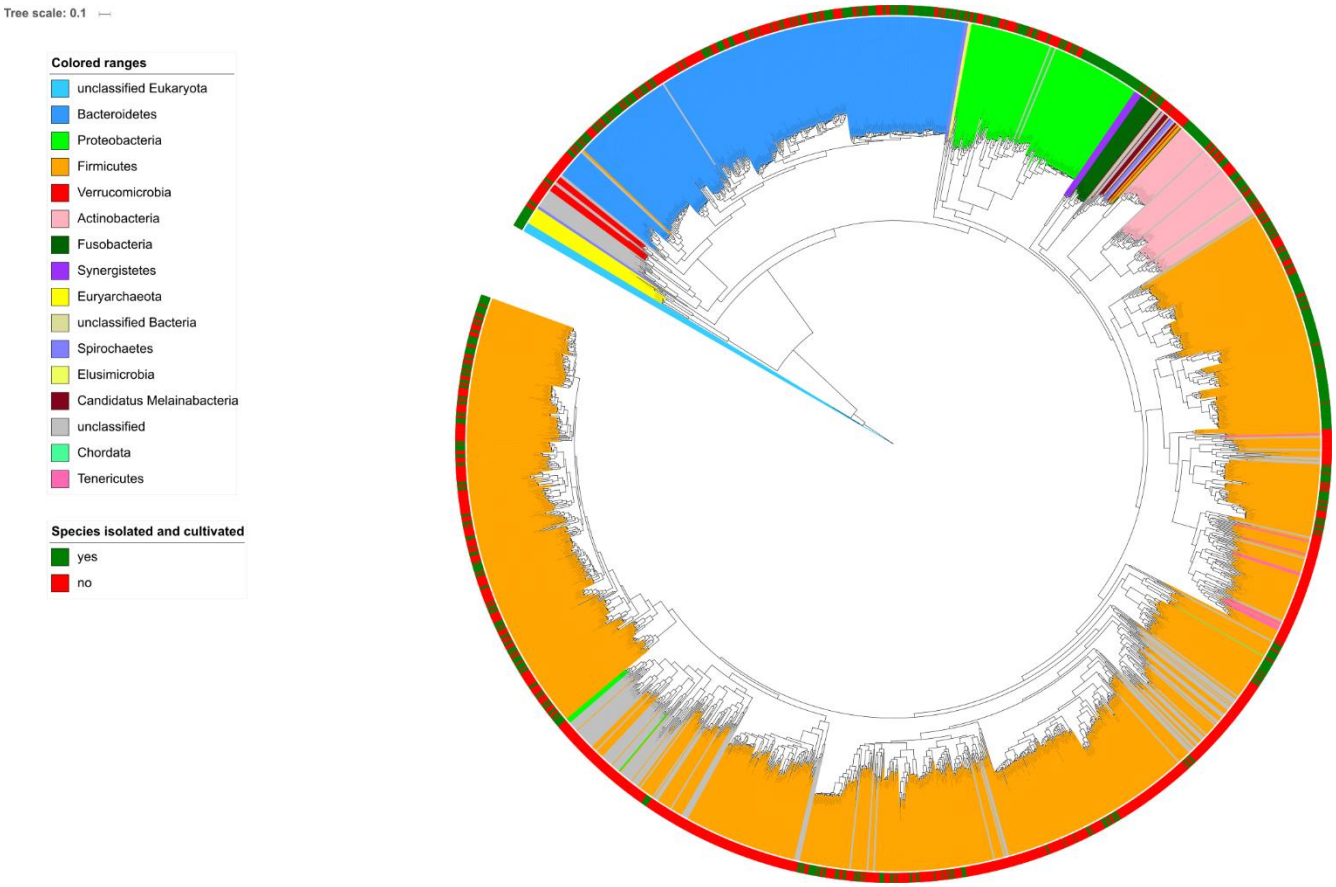


Figure 33 *: Phylogénie des 1 661 MSPs et de 360 génomes. Les couleurs internes correspondent à l'annotation au niveau phylum des MSPs et des génomes. Les couleurs externes indiquent si la MSP est représentative d'une espèce isolée, cultivée et séquencée. L'image est disponible en haute résolution à l'adresse URL suivante :* https://itol.embl.de/tree/9092247973557615244112 34

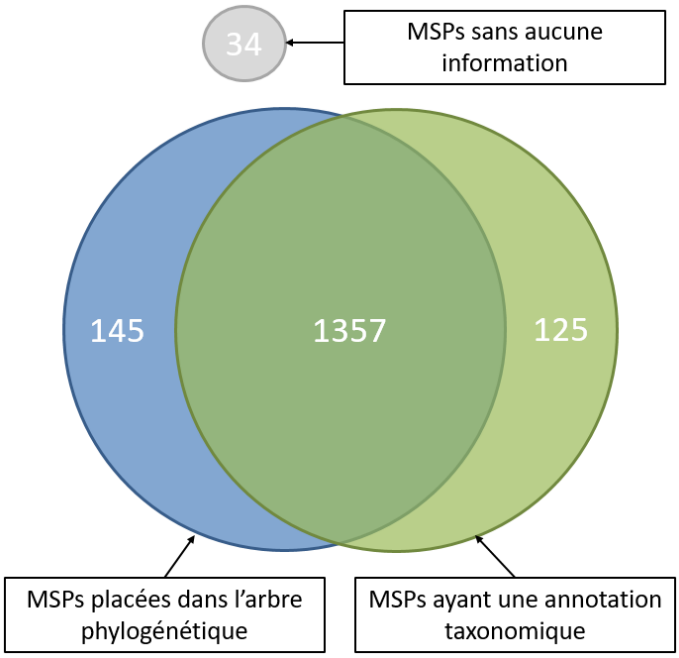


Figure 34 *: Comparaison des informations apportées par l'annotation taxonomique et l'arbre phylogénétique.*

Le cercle bleu correspond à l'ensemble des MSPs placées dans l'arbre phylogénétique, le cercle vert aux MSPs ayant une annotation taxonomique (niveau règne ou inférieur) et le cercle gris aux MSPs ni placées dans l'arbre ni annotées taxonomiquement.

5.3 Taille des MSPs

La majorité des MSPs obtenues sont de petite taille puisque le nombre médian de gènes est de 1 822. Néanmoins, 51 MSPs possèdent plus de 5000 gènes (**Figure 35.A** et **Tableau 9**).

Comme attendu, on observe une forte corrélation entre le nombre total de gènes dans une MSP et son nombre de gènes accessoires (**Figure 35.B**). Cependant, 4 MSPs correspondant aux procaryotes du genre *Blastocystis* ont un grand nombre de gènes (> 10 000) mais une faible proportion classifiée comme accessoire. Ceci suggère que les génomes des Eucaryotes du microbiote intestinal ont un nombre de gènes plus important et un contenu en gène plus stable que celui des Procaryotes. Parmi les MSPs de grande taille, on retrouve aussi plusieurs espèces procaryotes dont le contenu en gène a été signalé comme très variable dans des études de génomique des populations comme par exemple *Escherichia coli* (msp_0005 = 10 768 gènes), *Clostridium bolteae* (msp_0008 = 9 055 gènes) ou *Akkermansia muciniphila* (msp_0023 = 6 641 gènes) [154–156].

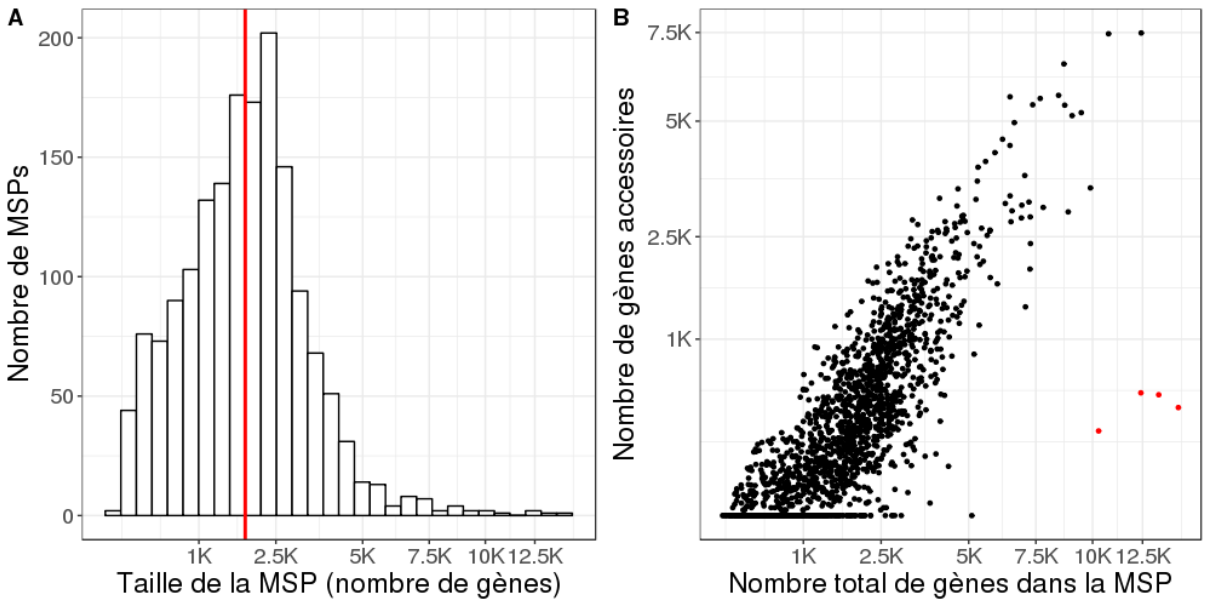


***Figure 35** : Taille et contenu en gènes des MSPs*

A. Histogramme représentatif du nombre de gènes dans les MSPs. La ligne verticale rouge représente le nombre de gènes médian dans une MSP (1 822 gènes)

B. Comparaison du nombre total de gènes dans une MSP (axe des abscisses) et son nombre de gènes accessoires (axe des ordonnées). Une forte corrélation entre le nombre total de gènes dans une MSP et son nombre de gènes accessoires est observée (coefficient de corrélation de Pearson = 0.78, p-valeur = 0). Les 4 MSPs avec un grand nombre de gènes mais peu d'accessoires sont mises en évidence en rouge. En les écartant, le coefficient de corrélation de Pearson atteint 0.84.

msp	Annotation taxonomique	Nombre de gènes	Pourcentage de gènes accessoires
msp_0001	*Blastocystis sp. subtype 3*	14467	2,7%
msp_0002	*Blastocystis sp. subtype 1*	13372	3,5%
msp_0003	*Bacteroides cellulosilyticus*	12431	85,8%
msp_0004	*Blastocystis sp. subtype 2*	12414	3,9%
msp_0005	*Escherichia coli*	10768	70,4%
msp_0006	*Blastocystis sp. subtype 4*	10287	2,4%
msp_0007	*Bacteroides intestinalis*	9473	87,8%
msp_0008	*Clostridium bolteae*	9055	58,9%
msp_0009	*Bacteroides fragilis*	8880	65,7%
msp_0010	*Bacteroides thetaiotaomicron*	8726	74,1%

***Tableau 9** : Liste des 10 MSPs de plus grande taille. Les MSPs associées au genre* Blastocystis *ont un grand nombre de gènes mais peu sont classifiés comme accessoires.*

5.4 Potentiel fonctionnel des MSPs

Après avoir annoté le catalogue IGC avec la base de données KEGG, nous avons recherché dans les MSPs des fonctions surreprésentées parmi les gènes core ou les gènes accessoires. Pour ce faire, nous avons utilisé un test binomial suivi d'une correction des p-valeurs avec la méthode de Benjamini-Hochberg. Les fonctions ayant une p-valeur ajustée inférieure à 10^{-2} ont été considérées comme significativement surreprésentées.

Comme attendu, des fonctions essentielles à la vie comme la traduction, la transcription, la réplication et la réparation de l'ADN sont significativement plus présentes parmi les gènes core de 446, 93 et 67 MSPs. La détection du quorum synchronisant l'expression de gènes au sein d'une population, les systèmes de sécrétion bactérienne ainsi que le métabolisme des glucides sont surreprésentés parmi les gènes accessoires d'une cinquantaine de MSPs.

Par la suite, nous avons recherché des gènes de résistance aux antibiotiques dans les MSPs en s'appuyant sur la base de données mustARD [157]. En utilisant blastn et des seuils de 95% d'identité sur 90% de la longueur, nous avons retrouvé la quasi-totalité (99,7%) des gènes mustARD dans le catalogue IGC. Remarquablement, 81% (4 907/6 075) d'entre eux sont groupés dans une MSP dont (94% 4 615/4 907) est classifiée comme core. Par conséquent, ces gènes sont vraisemblablement localisés sur des chromosomes et non sur des éléments génétiques mobiles comme les plasmides. Parmi les MSPs possédant des gènes de résistance, on retrouve des pathogènes connus dont *Clostridium scindens*, *Fusobacterium mortiferum*, *Enterococcus faecium* mais surtout des espèces commensales comme *Akkermansia muciniphila*, *Bacteroides vulgatus, Faecalibacterium prausnitzii* ou *Prevotella copri.* Ces résultats soutiennent l'hypothèse des auteurs selon laquelle les gènes de résistance font intrinsèquement partie de la flore commensale et qu'il est peu probable qu'ils soient transférés à des espèces pathogènes.

5.5 Prévalence et abondance des MSPs

La majorité des MSPs sont rares car elles sont détectées dans peu d'échantillons ce qui est cohérent avec les observations faites pour les gènes (page 31). 596 MSPs (35,9%) sont détectées dans moins de 1% des échantillons et 1110 (66,2%) dans moins de 5%. Seules 82 MSPs (4,9%) ont une prévalence supérieure à 50% montrant que le core microbien de l'intestin humain est limité à quelques dizaines d'espèces (**Tableau 10**). Parmi ces 82 MSPs, 28 ne sont pas annotées au niveau espèce. Ceci montre que des espèces prévalentes n'ont pas encore été séquencées à ce jour malgré les progrès des techniques de culture microbienne. Fait intéressant, aucune MSP n'est détectée dans tous les

échantillons probablement à cause de la taille importante de la cohorte et de la grande diversité de phénotypes.

msp	Annotation taxonomique	Prévalence dans les 1267 échantillons du catalogue IGC	Abondance moyenne dans les échantillons où le core de la MSP est détecté (%)
msp_0071	*Bacteroides vulgatus*	97,5% (1 235)	7,3%
msp_0044	*Bacteroides uniformis*	94,0% (1 191)	4,1%
msp_0079	*Blautia wexlerae*	93,8% (1 189)	0,9%
msp_0011	*Parabacteroides distasonis*	89,3% (1 131)	1,3%
msp_0204	*Ruminococcus sp. Marseille-P328*	86,9% (1 101)	0,2%
msp_0113	*Anaerostipes hadrus*	86,1% (1 091)	0,6%
msp_0489	*Dorea formicigenerans*	83,1% (1 053)	0,2%
msp_0151	*Fusicatenibacter saccharivorans*	82,6% (1 047)	0,5%
msp_0452	*Firmicutes* non classifiée	80,4% (1 019)	0,2%
msp_0377	*Odoribacter splanchnicus*	80,3% (1 018)	0,4%

***Tableau 10** : Liste des 10 MSPs les plus prévalentes dans les 1267 échantillons du catalogue IGC.*

Nous n'avons pas trouvé de lien évident entre la prévalence d'une MSP et son abondance (**Figure 36**). Remarquablement, les deux MSPs correspondant à *Bacteroides vulgatus* et *Bacteroides uniformis* sont à la fois très prévalentes (détectées respectivement dans 97,5% et 94,0% des échantillons) et abondantes (abondances relatives moyennes de 7,3% et 4,1%). La MSP correspondant à *Prevotella copri* est quant à elle détectée que dans un tiers de la cohorte mais est très abondante lorsqu'elle est présente (abondance relative moyenne de 12,7%).

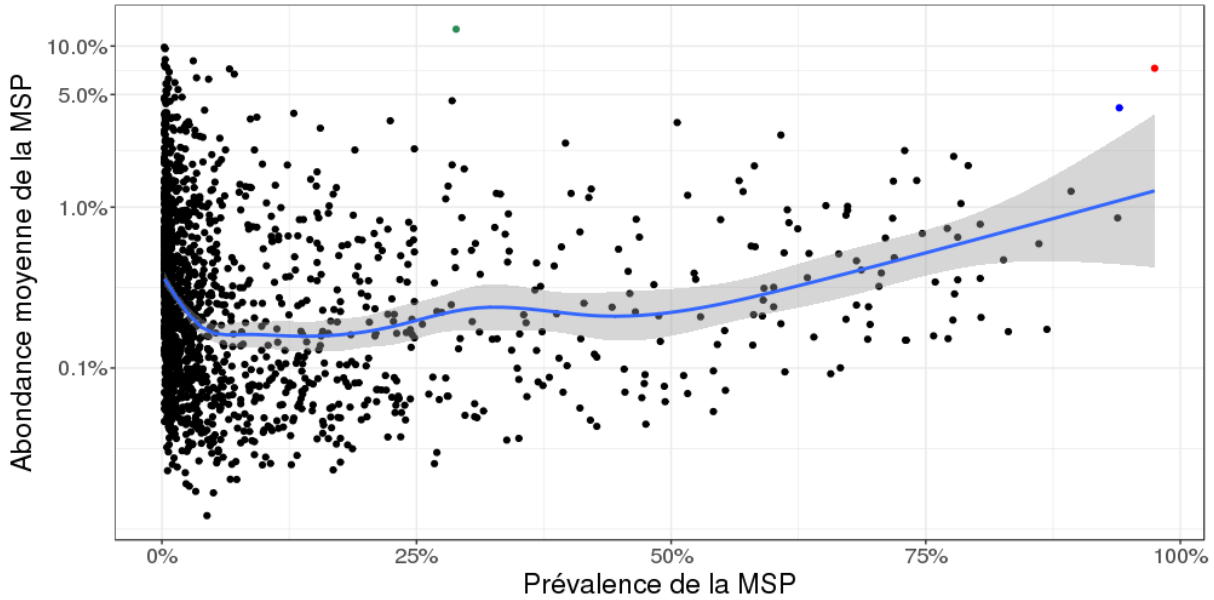


***Figure 36** : Abondance moyenne des MSPs dans les échantillons où elles sont détectées en fonction de leur prévalence.*

L'axe des abscisses représente le pourcentage d'échantillons du catalogue IGC où le core de la MSP est détecté. L'axe des ordonnées correspond à l'abondance moyenne de la MSP dans les échantillons où elle est présente (échelle $\log_{10}$). Les points rouge, bleu et vert correspondent respectivement aux MSPs associées à Bacteroides vulgatus*,* Bacteroides uniformis *et* Prevotella copri*. La courbe bleue correspond à la courbe de tendance calculée avec un modèle additif généralisé (méthode GAM)*

On note aussi la présence de MSPs très rares mais abondantes dans les échantillons où elles sont présentes (**Tableau 11**). Ces MSPs sont pour la plupart associées au genre *Prevotella*. Ainsi, il est important de prendre en considération les espèces rares lors des analyses car elles peuvent jouer un rôle significatif dans certains échantillons.

msp	Annotation taxonomique	Prévalence dans les 1267 échantillons du catalogue IGC	Abondance moyenne dans les échantillons où le core de la MSP est détecté (%)
msp_1524	unclassified *Prevotella*	0,32% (4)	9,63%
msp_0989	*Prevotella sp. CAG:1320*	0,32% (4)	8,26%
msp_1460	unclassified *Prevotella*	0,47% (6)	7,36%
msp_0545	*Prevotella sp. CAG:1124*	0,63 (8)	6,28%
msp_0744	*Prevotella sp. CAG:1185*	0,47 (6)	5,06%

***Tableau 11** : Exemple de MSPs très rares mais abondantes dans les échantillons où elles sont détectées*

Globalement, les MSPs annotées au niveau espèce sont détectées dans un plus grand nombre d'échantillons que les MSPs ayant une annotation moins précise (**Figure 37.A**). Ainsi, les espèces du microbiote intestinal humain séquencées à ce jour sont globalement les plus prévalentes. A contrario, nous n'avons pas trouvé de différence significative entre l'abondance des MSPs annotées au niveau espèce et celles qui ne le sont pas (**Figure 37.B**).

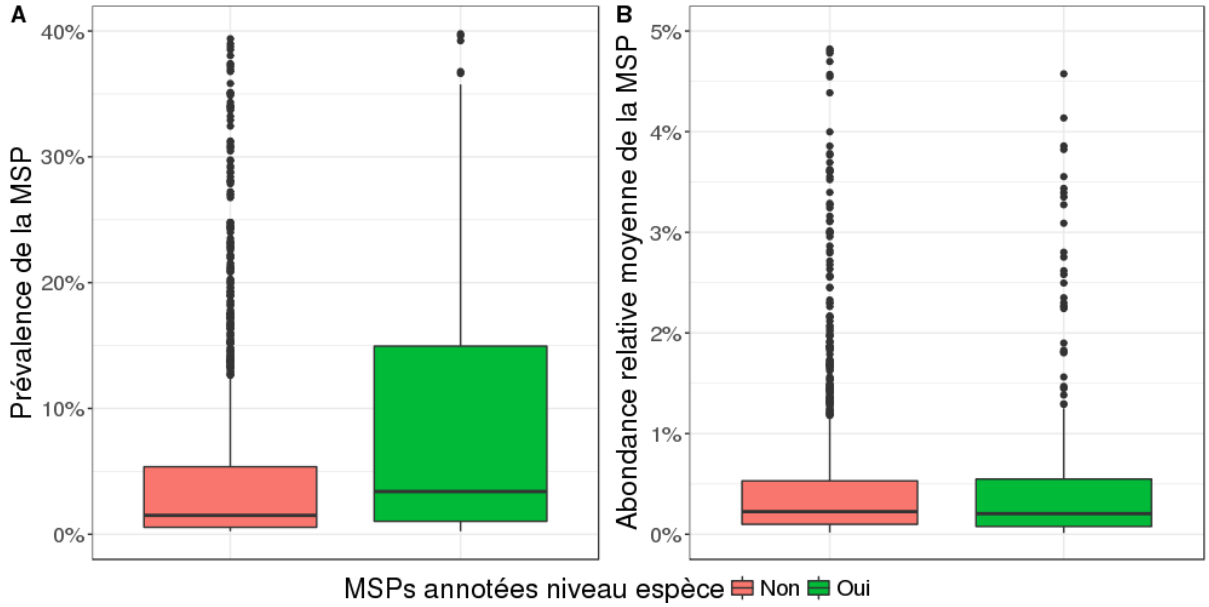


***Figure 37** : Comparaison de la prévalence (à gauche) et de l'abondance (à droite) des MSPs annotées au niveau espèce (boxplots verts) et de celles qui ne le sont pas (boxplots rouges).*

A. Les MSPs annotées au niveau espèce sont significativement plus prévalentes que celles qui ne le sont pas (prévalence médiane de 5,4% contre 1,7%, p-valeur=1,4.10-21 au test U de Mann-Whitney). Les MSPs avec une prévalence supérieure à 50% n'ont pas été représentées pour faciliter la lisibilité du graphique.

B. Il n'existe pas de différence significative entre l'abondance relative des MSPs annotées niveau espèce et celles qui ne le sont pas (abondance médiane de 0,21% contre 0,23%, p-valeur=0,21 au test U de Mann-Whitney). Seuls les échantillons où les gènes core des MSPs sont détectés ont été considérés pour le calcul de l'abondance. Les MSPs avec une abondance moyenne supérieure à 5% n'ont pas été représentées pour faciliter la lisibilité du graphique.

Nous avons aussi constaté que la prévalence des gènes dans une MSP suit souvent une distribution bimodale en U comme exposé en 1.1.1.4. Les gènes ont soit une prévalence élevée (gènes core + soft core) ou faible (gènes cloud) mais plus rarement intermédiaire (gènes shell) (**Figure 38**).

Finalement, on observe une forte corrélation entre la prévalence d'une MSP et son nombre de gènes accessoires (coefficient de corrélation de Spearman = 0.86, p-valeur = 0). Comme appréhendé dans la simulation, plus une MSP est détectée dans un nombre important d'échantillons, plus MSPminer récupère un nombre important de gènes accessoires et plus particulièrement les gènes rares (cloud) qui sont les plus nombreux.

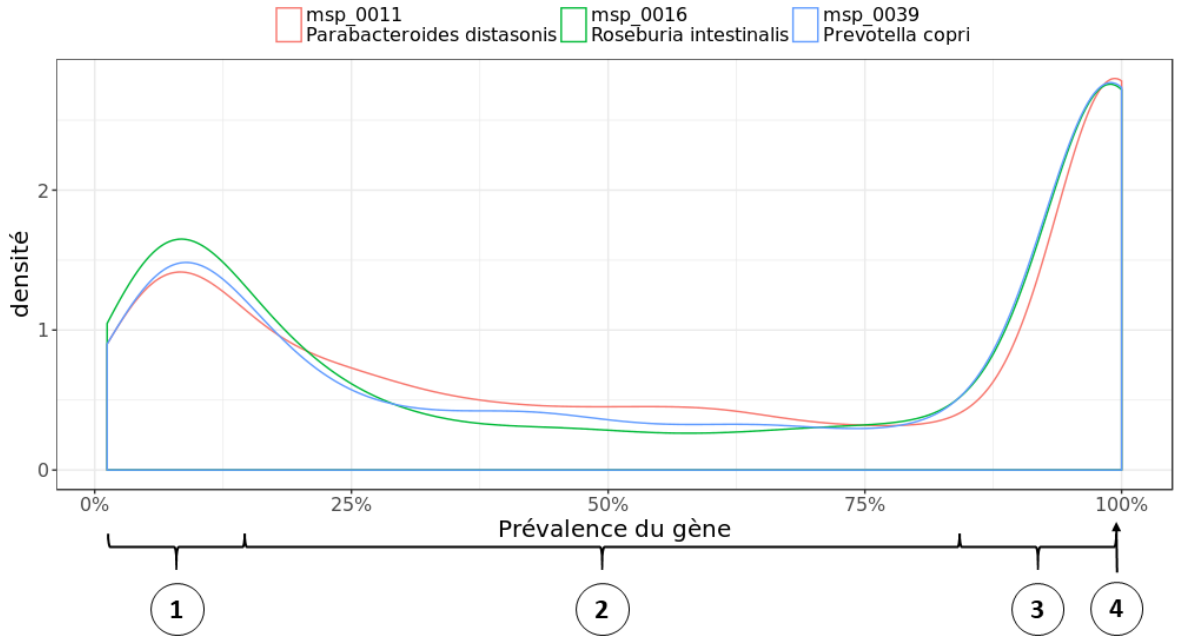


***Figure 38** : Prévalence des gènes de 3 MSPs dans les échantillons du catalogue IGC où leurs gènes core respectifs sont détectés.*

Ces MSPs ont été sélectionnées car elles sont détectées dans un grand nombre d'échantillons ce qui augmente la probabilité d'y trouver des gènes accessoires rares. Pour chaque MSP, on représente sur l'axe des abscisses la prévalence du gène et en ordonnées la proportion de gènes ayant cette prévalence.

On met ainsi en évidence 4 catégories de gènes décrits par Koonin et Wolf [30] *:*

1. Cloud : gènes accessoires rares détectés dans une faible proportion d'échantillons où l'espèce est présente. Ici, la proportion de gènes cloud est sous-estimée car les plus rares ne vérifient pas le critère de sélection basé sur la co-occurrence.

2. Shell : gènes accessoires de prévalence intermédiaire

3. Soft core : gènes accessoires très prévalents présents dans quasiment tous les échantillons où l'espèce est présente.

4. (Hard) core : gènes présents dans tous les échantillons où l'espèce est présente.

5.6 Exploration du contenu d'une MSP

Pour explorer le contenu d'une MSP et établir un lien avec les souches séquencées, nous avons comparé le génome complet de la souche *Parabacteroides distasonis* ATCC 8503 [158] à la msp_0011 représentative de cette espèce. Notre choix s'est porté sur l'espèce *Parabacteroides distasonis* pour deux raisons. Premièrement, c'est une espèce très prévalente dans le microbiote intestinal. En effet, elle est présente dans 90% des individus de la cohorte du catalogue IGC. Deuxièmement, la MSP qui

lui correspond est constituée d'un grand nombre de gènes (8 690) dont 78% sont d'accessoires. Ceci suggère que le contenu en gènes de cette espèce varie fortement d'une souche à l'autre.

Parmi les 3 850 gènes prédits dans le génome, 3 781 (98%) ont un homologue proche (95% d'identité sur 90% de la longueur) dans le catalogue IGC et 3 442 (89%) sont regroupés dans la MSP. Comme attendu, quasiment tous les gènes core de la MSP sont présents dans le génome (1 867 / 1 921, 97%) ainsi qu'une fraction importante de gènes accessoires dont la prévalence excède 80% (522 / 599, 87%). Seule une petite fraction de gènes accessoires moins prévalents est retrouvée dans le génome (1 371/ 5 090, 27%).

256 gènes classifiés comme accessoires dans la MSP sont associés à l'origine géographique des échantillons (voir 7.1.2). Parmi ces 256 gènes, 224 (87,5%) sont plus fréquents chez les individus américains que chez les individus d'autres origines. Ceci est attendu car la souche ATCC 8503 a été isolée dans les selles d'un individu américain.

Les gènes accessoires détectés dans les mêmes échantillons et groupés dans un module sont dans la majorité des cas physiquement proches sur le génome. En effet, la distance médiane entre leurs gènes successifs d'un module est inférieure à 150 paires de bases dans 82% des cas. Cette proximité physique est en partie due à la présence de gènes chevauchants. En effet, 28% des modules sont composés uniquement de tels gènes et 46% en possèdent au moins une paire. Il est probable que les gènes proches non chevauchants d'un module assurent conjointement une fonction biologique et soient organisés en opérons. Remarquablement, certains gènes singletons sont entourés de gènes regroupés dans le même module. Ceci montre que le critère strict de co-occurrence utilisé par MSPminer peut éclater dans plusieurs modules des gènes avec des motifs de présence/absence légèrement différents alors qu'ils auraient pu être regroupés (**Figure 39**).

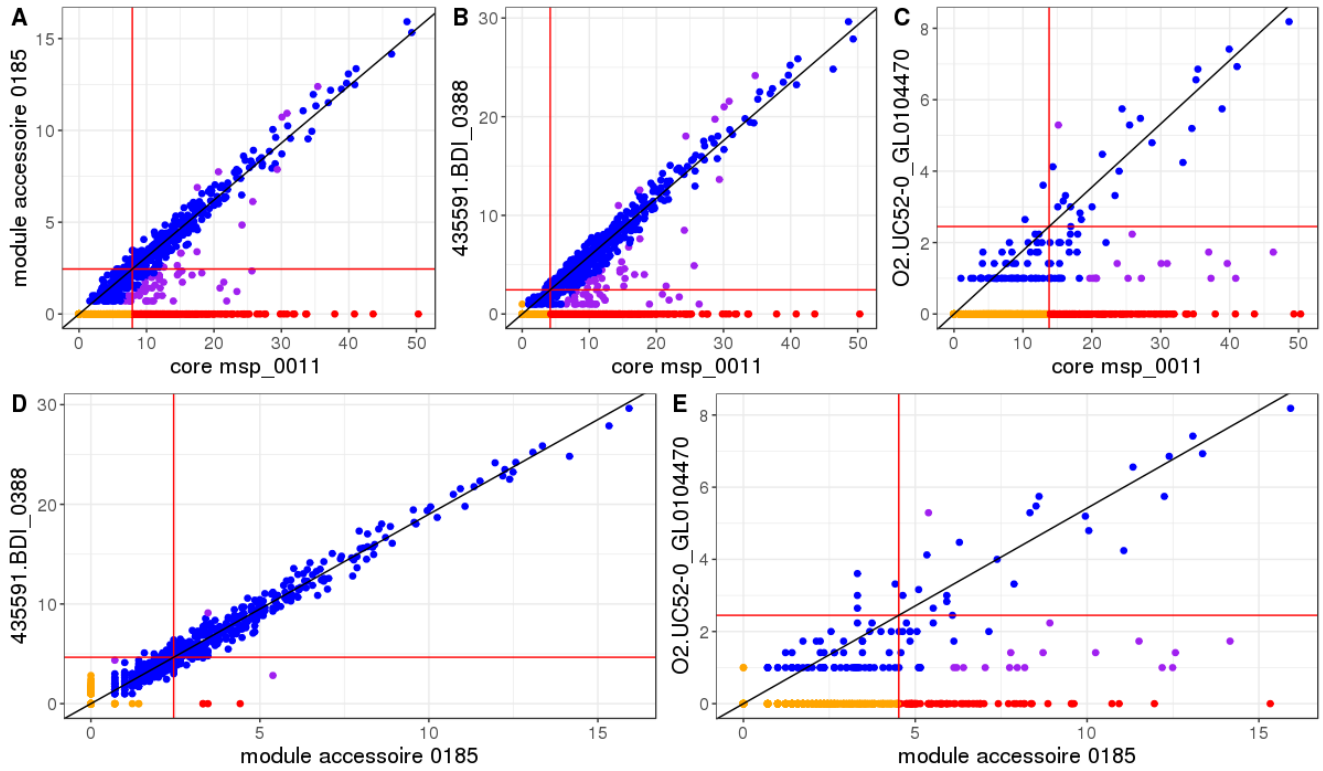


Figure 39 *: Gènes retrouvés dans une région génomique de* P. distasonis *ATCC 8503 délimitée par le module accessoire 0185*

Le module accessoire 0185 est composé de 4 gènes détectés dans le génome de Parabacteroides distasonis ATCC 8503 entre les positions 441 398 et 442 762. Les gènes 435591.BDI_0388 et O2.UC52-0_GL0104470 sont détectés dans cette région (positions 441903 - 442697 et 442 472 - 442 697 respectivement) mais ne sont pas groupés dans le module.

A. B. et C. : Comparaison du signal du core génome de la msp_0011 correspondant à Parabacteroides distasonis (axe des abscisses) avec celui du module accessoire 0185 (A), du gène 435591.BDI_0388 (B) et O2.UC52-0_GL0104470 (C) (axes des ordonnées) dans les 1267 échantillons du catalogue IGC. L'association entre le core de la MSP, le module accessoire et les deux gènes singletons est claire.

D. : Comparaison du signal du module accessoire 0185 et du gène singleton 435591.BDI_0388 dans les 1267 échantillons du catalogue IGC. Ici, le gène n'est pas clustérisé dans le module parce qu'il manque dans 1,3% des échantillons où le module est présent. Ici, on peut considérer que le critère de regroupement est trop strict.

E. : Comparaison du signal du signal du module accessoire 0185 et du gène singleton O2.UC52-0_GL0104470 dans les 1267 échantillons du catalogue IGC. Ici, le gène n'est pas clustérisé dans le module parce qu'il n'est pas détecté dans 70% des échantillons où le module est présent. Ici, exclure le gène du module est pertinent.

Certaines régions génomiques contiennent des gènes qui ne sont pas groupés dans la MSP. Bien que certains gènes exclus puissent être de faux négatifs, un grand nombre d'entre eux sont exclus à raison car ils sont observés dans trop peu d'échantillons où leurs comptages ne vérifient pas le critère le regroupement basé sur l'hypothèse de proportionnalité (**Figure 40**). Fait intéressant, certaines de ces régions sont annotées comme des éléments génétiques mobiles tels que des transposons ou des prophages.

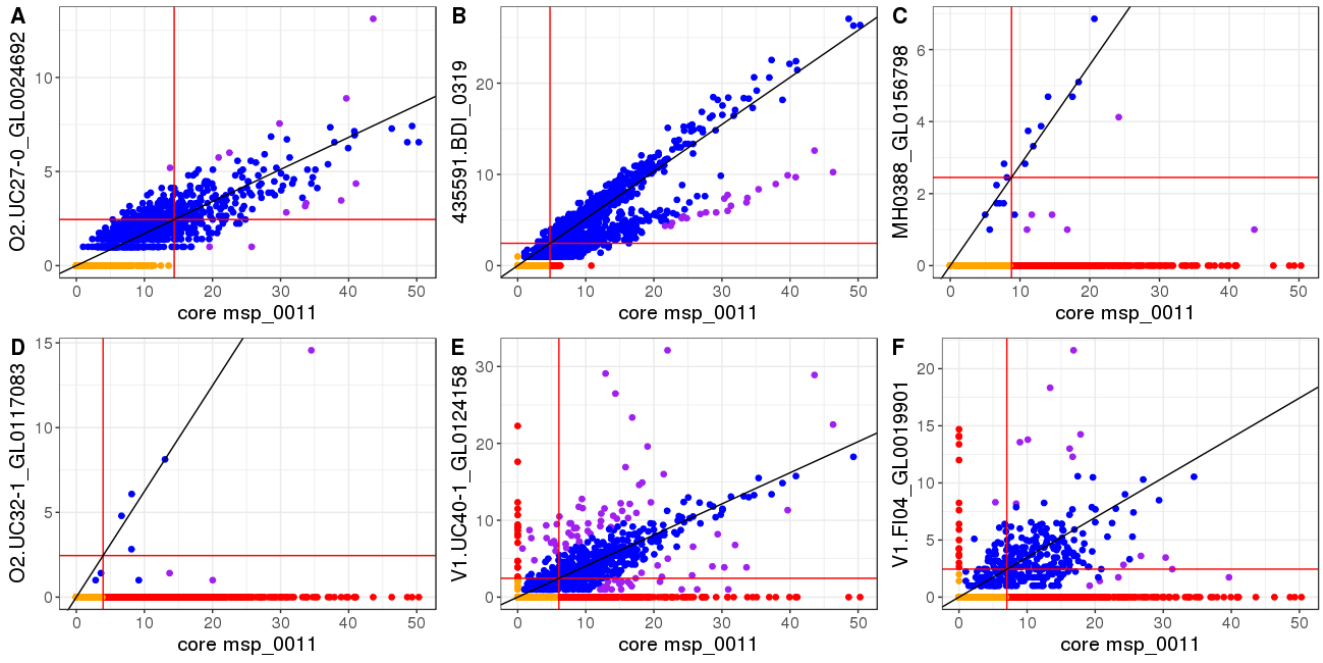


***Figure 40** : Gènes de Parabacteroides distasonis ATCC 8503 non regroupés dans la msp_0011*

Dans chaque sous-figure, on compare le signal du core génome de la msp_0011 (axe des abscisses) avec celui du gène (axe des ordonnées) dans les 1267 échantillons du catalogue IGC.

A. Le gène O2.UC27-0_GL0024692 correspond à un faux négatif dont les comptages sont surdispersés. La mesure robuste de proportionnalité est en dessous du seuil par défaut utilisé par MSPminer.

B. Le gène 435591.BDI_0319 est un faux négatif dont le nombre de copies varie d'un échantillon à l'autre. La mesure robuste de proportionnalité est en dessous du seuil par défaut utilisé par MSPminer.

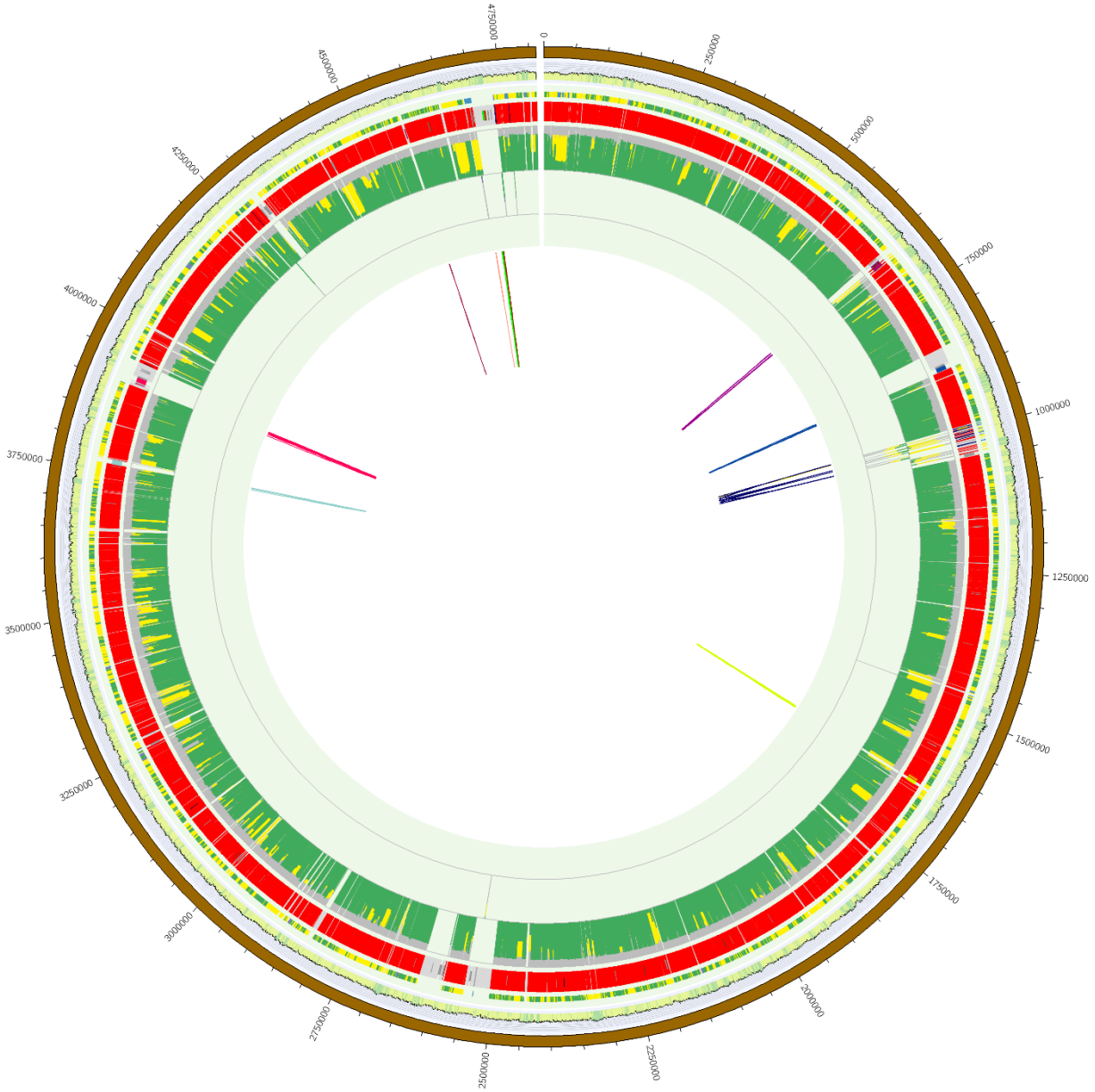
C. Le gène MH0388_GL0156798 correspond probablement à un faux négatif. Le nombre d'échantillons où le gène est détecté est trop faible pour détecter un lien avec le signal du core de la MSP.

D. Le gène O2.UC32-1_GL0117083 pourrait être un faux négatif. Néanmoins, le nombre d'échantillons dans lequel il est détecté est beaucoup trop faible pour établir un lien avec le signal du core de la MSP.

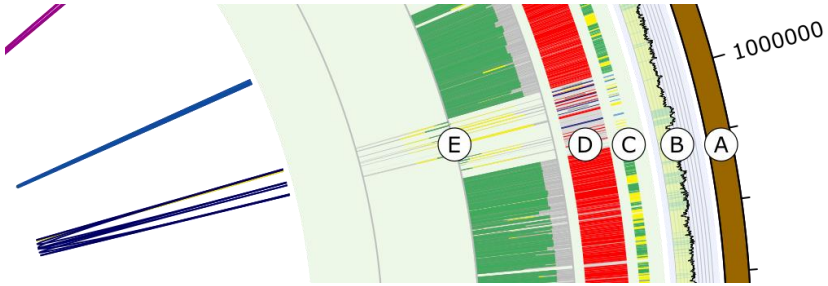
E. Le gène V1.UC40-1_GL0124158 correspond à un vrai négatif car il n'est pas clairement associé avec le core de la MSP. Il fait partie d'un élément génétique mobile et est annoté comme une protéine de transposon conjugatif.

F. Le gène V1.FI04_GL0019901 correspond à un vrai négatif car il n'est pas clairement associé avec le core de la MSP. Fait intéressant, il est classifié comme gène accessoire partagé dans la msp_0025 correspondant à Parabacteroides merdae.

Finalement, nous avons développé un script s'appuyant sur l'utilitaire Circos [159] pour visualiser la projection d'un génome sur les MSPs. En appliquant ce script à *Parabacteroides distasonis* ATCC 8503, on obtient la **Figure 41.** Celle-ci met évidence les régions contenant des gènes qui ne sont pas groupés dans la msp_0011 ainsi que des gènes consécutifs qui ont la même prévalence (modules)



: Projection du génome de P. distasonis ATCC 8503 sur les MSPs



Description des cercles de l'extérieur vers l'intérieur :

A. Position sur le chromosome
B. Taux de GC (format : histogramme)
C. Type de gène dans la MSP:
   vert=core, jaune=accessoire,
   bleu=core partagé, violet=accessoire partagé
D. MSP :
   Bande rouge=gène groupé dans la msp_0011
   Autre couleur=gène groupé dans une autre MSP (bande large) ou une seed (bande étroite)
E. Attribution des 1267 échantillons (format : histogramme):
   - Positionnement :
     Vers l'extérieur si le gène est dans la msp_0011, vers l'intérieur sinon.
   - Couleurs pour la présence/absence des échantillons :
     gris=gène et MSP non détectés
     vert=gène et MSP détectés
     jaune=MSP détectée mais pas le gène
     violet=gène détecté mais pas la MSP

6. Evaluation et validation des MSPs

6.1 Précision

Dans un premier temps, nous avons évalué la spécificité des MSPs en calculant l'homogénéité taxonomique des gènes les composant. On s'assure qu'une MSP contient uniquement les gènes d'une espèce microbienne et qu'elle n'est pas contaminée par des gènes provenant d'une autre entité.

Par la suite, seules les 304 MSPs assignées à des espèces bien définies ont été considérées. On calcule pour chacune d'entre-elles la proportion de gènes assignés à l'espèce dominante. Si on exclut les gènes non annotés, l'homogénéité taxonomique est très élevée pour toutes les catégories de gènes (moyenne > 0,98) à l'exception des gènes accessoires partagés (**Tableau 12**).

	tous les gènes	**core uniquement**	**accessoires uniquement**	**core partagés uniquement**	**acc. partagés uniquement**
gènes non annotés inclus	0,83 - 0,93 -0,98	0,98 - 1,0 - 1,0	0,67 - 0,81 - 0,94	0,94 - 0,98 - 1,0	0,50 - 0,70 - 0,88
gènes non annotés exclus	0,98 - 1,0 - 1,0	1,0 - 1,0 - 1,0	0,99 - 1,0 - 1,0	0,98 - 1,0 - 1,0	0,80 - 0,92 - 0,99

***Tableau 12** : Homogénéité taxonomique des 304 MSPs annotées niveau espèce. Les valeurs sont : 1er quartile, médiane et 3ème quartile.*

On calcule pour chaque MSP la proportion de gènes assignés à l'espèce dominante. La ligne d'en-tête indique la catégorie de gènes considérée. La colonne d'en-tête indique si l'on considère ou non les gènes sans annotation

On constate que la proportion de gènes non annotés est plus importante parmi les gènes accessoires. Ceci suggère que MSPminer découvre des gènes facultatifs qui ne sont pas présents dans les souches séquencées à ce jour. Par exemple, les MSPs représentatives des espèces *Bacteroides plebeius* (msp_0015), *Ruminococcus bicirculans* (msp_0012) et *Eubacterium eligens* (msp_0027) possèdent de nombreux gènes accessoires non annotés (respectivement 2 888, 2 821 et 2 399) ce qui est cohérent avec le faible nombre de génomes disponibles pour ces espèces (respectivement 2, 2 et 4).

Pour confirmer que ces gènes non annotés font effectivement partie du pan-génome des espèces, nous avons téléchargé les contigs provenant des assemblages *de novo* des 1 267 échantillons du catalogue IGC. En moyenne, 80% des nouveaux gènes sont détectés dans des contigs assignés aux espèces d'intérêt. Plus précisément, ces contigs contiennent le nouveau gène et au moins deux gènes connus assignés à l'espèce d'intérêt. Les nouveaux gènes qui n'ont pas pu être validés sont détectés dans des contigs de trop petite taille ou qui ne contiennent que des gènes non annotés.

Inversement, 99% des gènes provenant de la MSP représentative de *Escherichia coli* (msp_0005) sont annotés car des milliers de génomes sont disponibles pour cette espèce.

6.2 Sensibilité

Par la suite, nous avons aligné 3 143 génomes représentatifs de 322 espèces du microbiote intestinal humain avec les gènes du catalogue IGC. Pour chaque génome, la sensibilité de MSPminer a été définie comme rapport entre le nombre total de ses gènes groupés dans la MSP la plus représentative et le nombre de ses gènes détectés dans la catalogue. Globalement, la sensibilité pondérée par le nombre de génomes disponibles par espèce est élevée (médiane = 77%, c.f. **Figure 42**).

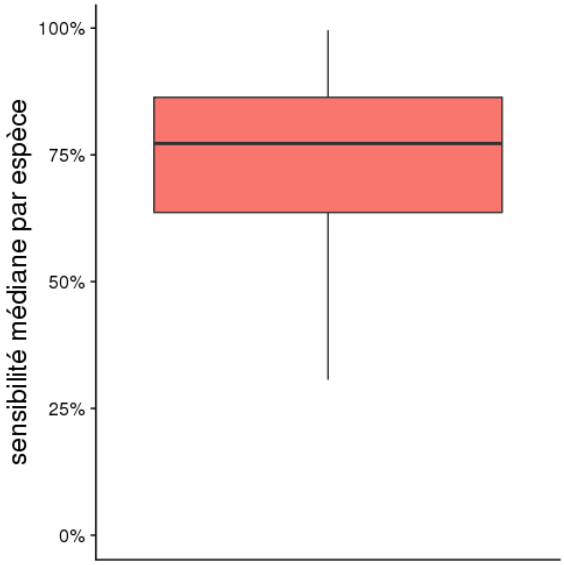


Figure 42 *: Sensibilité de MSPminer pour 322 espèces du microbiote intestinal humain.*

Ce graphique n'est pas biaisé par le nombre de génomes disponibles par espèce. Pour une espèce donnée, on calcule la médiane des sensibilités de chacun de ses génomes. Par exemple, une seule valeur (83,4%) représente les 1 127 génomes d'Escherichia coli.

Nous avons aussi constaté que les gènes regroupés dans les MSPs sont significativement plus longs que ceux qui ne le sont pas (**Figure 43**). En effet, les gènes courts sont plus difficilement regroupés dans les MSPs car leurs comptages sont plus faibles et plus dispersés (voir 8.2.4)

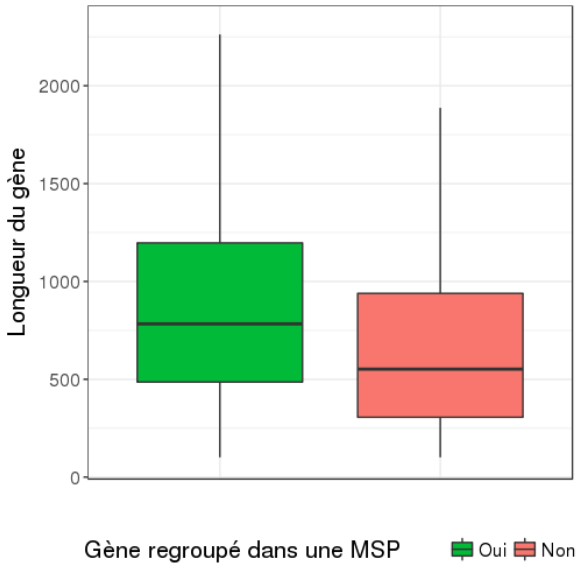


Figure 43 *: Comparaison de la longueur des gènes regroupés dans les MSPs et de ceux qui ne le sont pas.*

Les gènes regroupés dans les MSPs sont significativement plus longs que ceux qui ne le sont pas (longueur médiane de 783 paires de bases contre 552 paires de bases, p-valeur=0 au test U de Mann-Whitney). Les gènes dont la longueur excède 2 300 pb ne sont pas représentés pour faciliter la lisibilité du graphique.

De plus, nous avons remarqué que la proportion de gènes complets (c'est-à-dire ceux qui commencent par un codon START et qui finissent par un CODON STOP) est plus importante parmi les gènes regroupés dans les MSPs par rapport à ceux qui ne le sont pas (65% de gènes complets contre 50%, p-valeur = 0 au test exact du khi-deux). Ceci s'explique par le fait que les gènes complets sont globalement plus longs que les gènes incomplets (médiane = 711 pb vs 495 bp, p-valeur = 0 au test de Wilcoxon-Mann-Whitney)

Par la suite, nous nous sommes focalisés sur l'espèce *Escherichia coli* représentée ici par 1 127 génomes de souches isolées à partir de fèces humains ou de biopsies de l'intestin. Ces génomes sont bien couverts par la msp_0005 (q1 = 80% - médiane = 83,5% - q3 = 86,7%). 95% des gènes détectés dans 99% des génomes ou plus sont étiquetés comme gènes core dans la msp_0005. Ce résultat illustre la robustesse de la classification effectuée par MSPminer. Cependant, 32 078 gènes du catalogue IGC détectés dans les génomes de *E. coli* ne sont pas regroupés dans la msp_0005. 85% de ces gènes sont présents dans moins de 5% des échantillons métagénomiques où *E. coli* est détectée indiquant que MSPminer manque les gènes accessoires les plus rares (gènes cloud) qui sont très nombreux.

6.3 Validation des MSPs par recensement de gènes marqueurs

Ci-après, on recense les gènes marqueurs dans chaque MSP pour s'assurer qu'elles correspondent à des espèces microbiennes. Ceci est particulièrement utile pour valider les MSPs qui ne sont pas annotées au niveau espèce. Le pourcentage de gènes marqueurs détectés dans une MSP permet quant à lui d'estimer la complétion du core génome. Enfin, on s'assure que les marqueurs regroupés dans des MSPs sont des gènes core pour valider la classification effectuée par MSPminer.

Les 40 marqueurs phylogénétiques universels décrits par Sunagawa *et al.* ont été extraits du catalogue IGC avec l'outil fetchMG [123]. 84% des gènes marqueurs détectés dans au moins 3 échantillons sont assignés à des MSPs indiquant que ces dernières capturent une grande proportion du signal biologique au niveau espèce. 648 (39%) MSPs ont au moins 38 marqueurs détectés et 915 (55%) en ont au moins 30 (**Figure 44.A**). 93% des marqueurs recensés dans les MSPs sont classifiés comme core. Parmi les marqueurs qui ne sont pas des gènes core, 70% sont des accessoires très prévalents présents dans plus de 90% des échantillons où la MSP est détectée.

De même, la même analyse a été effectuée en extrayant avec l'outil fetch-cscg [160] les 139 (respectivement 162) marqueurs du règne des archées (respectivement bactéries) décrits par Rinke et al. [161] . 647 (39%) MSPs ont au moins 90% des marqueurs et 838 (50%) en ont au moins 80% (**Figure 44.B**). 93,7% des marqueurs recensés dans les MSPs sont classifiés comme gènes core, les 6,3% restants étant aux trois quarts des accessoires dont la prévalence excède 90%.

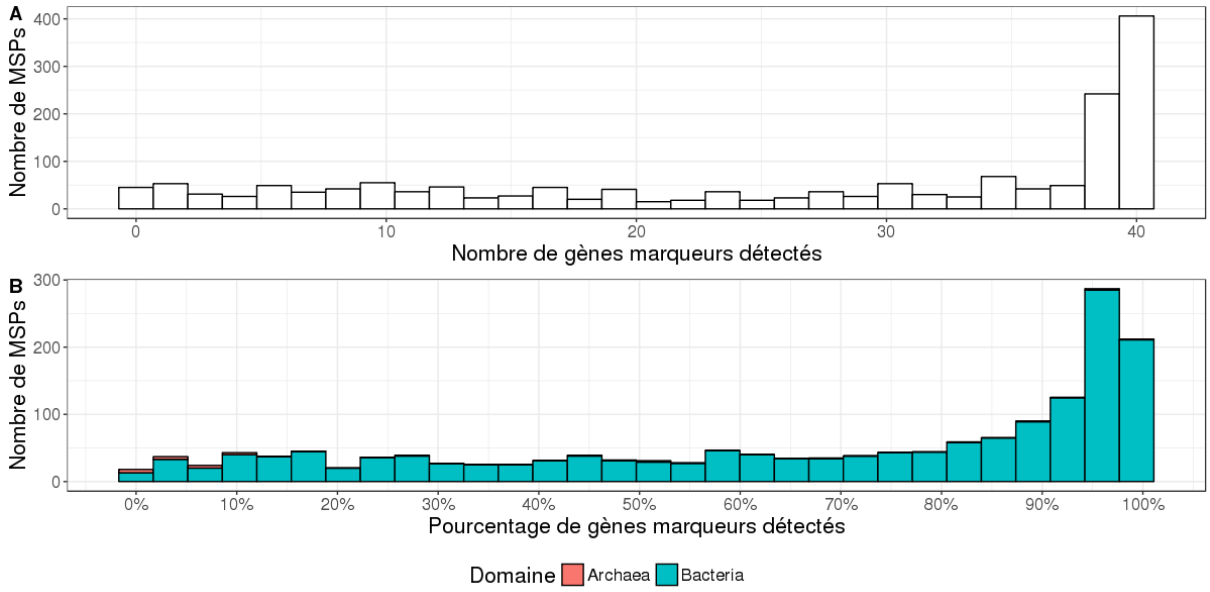


Figure 44 *: Recensement des gènes marqueurs dans les MSPs.*

A. Histogramme du nombre de gènes marqueurs universels (Sunagawa et al.) détectés dans les MSPs.

B. Histogramme du pourcentage de gènes marqueurs domaine-spécifique (Rinke et al.) détectés dans les MSPs. Les barres bleues (respectivement rouges) correspondent aux MSPs annotées comme bactéries (respectivement archées). Les MSPs sont pour leur très grande majorité des bactéries ce qui explique la finesse des barres rouges.

Les résultats obtenus par l'intermédiaire des deux méthodes sont similaires. A partir du pourcentage de gènes marqueurs détectés dans les MSPs, on déduit que 50% d'entre elles ont un core génome couvert quasiment intégralement. Comme la grande majorité des marqueurs sont des gènes core ou des gènes accessoires très prévalents, on conclut que classification empirique des gènes core réalisée par MSPminer est fiable.

6.4 Validation des MSPs par analyse de la composition nucléotidique

On appelle k-mers l'ensemble des sous-séquences de longueur k composant une séquence d'ADN. De nombreuses études ont montré une similitude des fréquences en k-mers dans des fragments d'ADN de quelques kilobases provenant de la même espèce. Cette signature compositionnelle reflète le biais d'usage des codons c'est-à-dire de l'utilisation préférentielle par une espèce microbienne de certains triplets de nucléotides pour coder un acide aminé. Dans un contexte métagénomique, ces signatures sont utilisées pour regrouper les contigs provenant de la même entité biologique [162,163].

Pour montrer que les MSPs qui n'ont pas d'annotation au niveau espèce sont vraisemblablement constituées de gènes provenant de la même espèce microbienne, nous avons mis en évidence l'homogénéité de la composition en k-mers des gènes groupés dans une même MSP par comparaison avec des groupes de gènes tirés aléatoirement. Ici, nous avons choisi de recenser les 4-mers en représentation canonique où un 4-mer et son complémentaire inverse représentent le même objet. Les 4-mers (ou tétranucléotides) ont l'avantage d'être bien plus discriminants que les 2-mers (ou dinucléotides) [164]. De plus, leur faible nombre (136) permet un traitement informatique rapide et peu gourmand en mémoire.

Dans un premier temps, nous avons généré 1 661 clusters composés de 1 800 gènes tirés aléatoirement dans le catalogue IGC. Par la suite, nous avons calculé les profils en 4-mers de chacun

des gènes d'un cluster (MSP ou cluster aléatoire) puis nous avons comparé ces profils deux à deux en utilisant le coefficient de corrélation de Pearson. Enfin, nous avons calculé la corrélation moyenne de chaque cluster puis nous avons comparé les moyennes obtenues pour les MSPs assignées à une espèce connue, les MSPs assignées à un rang taxonomique supérieur et les clusters aléatoires.

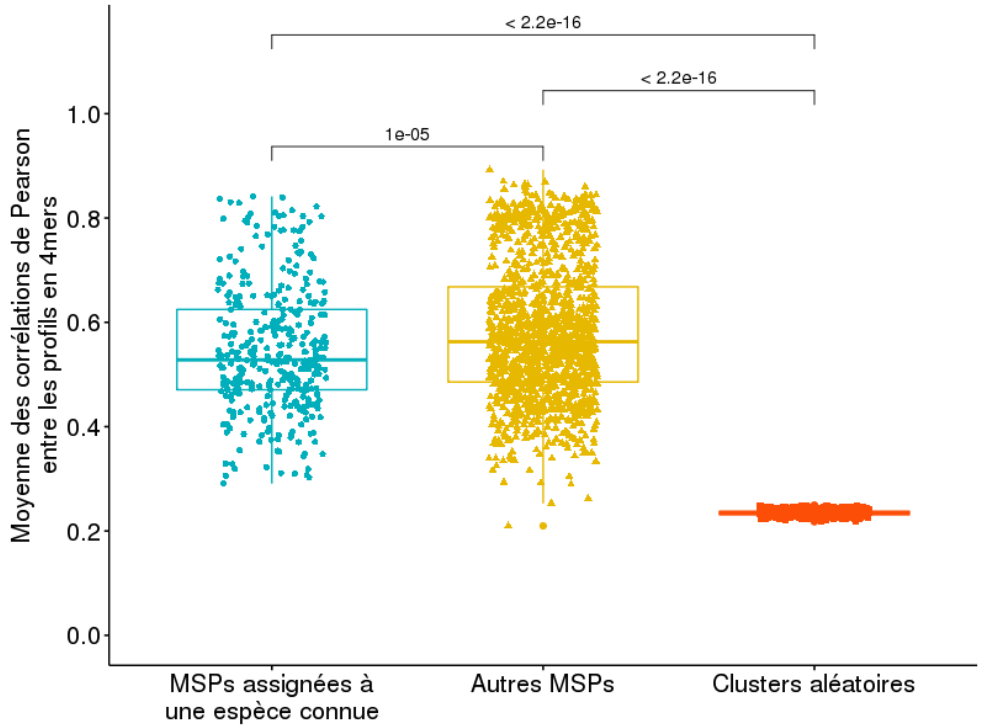


Figure 45 *: Comparaison des moyennes des corrélations de Pearson entre les profils en 4-mers des gènes groupés dans les MSPs assignées à une espèce connue (bleu), dans les MSPs ayant une annotation taxonomique moins précise (jaune) et dans les clusters aléatoires (rouge).*

Les gènes groupés dans une MSP ont des profils en tétranucléotides significativement plus corrélés que les gènes des clusters aléatoires (moyenne des corrélations moyennes par cluster = 0.57 contre 0,23 ; p-valeur < 2.2e-16 au test de Wilcoxon-Mann-Whitney). Néanmoins, les corrélations des profils en 4-mers dans les MSPs sont bien plus faibles que celles observées pour des contigs dont la longueur excède 2 500 paires de bases (r > 0,9) car les gènes sont généralement trop courts pour obtenir des profils en tétranucléotides robustes représentatifs de l'espèce.

Remarquablement, les corrélations des profils en 4-mers dans les MSPs assignées à une espèce connue sont du même ordre de grandeur que celles des MSPs ayant une annotation taxonomique moins précise (**Figure 45**). Ceci suggère que les MSPs n'ayant pas d'annotation au niveau espèce regroupent les gènes d'espèces microbiennes inconnues à ce jour. On constate même que les corrélations dans les MSPs correspondant à une espèce connue sont légèrement plus faibles que celles des autres MSPs (0,54 contre 0,57) probablement parce les MSPs de la première catégorie sont composés de gènes globalement plus courts que ceux de la deuxième (moyenne des longueurs médianes des gènes dans les MSPs = 838 contre 927, p-valeur < 2.2e-16 au test de Wilcoxon-Mann-Whitney).

6.5 Validation des MSPs par analyse du lien physique entre les gènes

Partant du principe que les MSPs sont constituées de gènes provenant de la même espèce, on s'attend à ce que ces derniers soient liés physiquement. Autrement dit, les gènes d'une même MSP devraient être observés simultanément sur des contigs obtenus par assemblage *de novo*.

Pour le vérifier, nous avons téléchargé les assemblages des 1 267 échantillons ayant servi à la construction du catalogue IGC. Dans un second temps, nous avons recherché avec BLASTn [149] les gènes constituant les MSPs dans les 61 749 489 contigs téléchargés. Seuls les hits couvrant le gène cible à plus 90% avec au moins 95% d'identité nucléotidique ont été conservés. Nous avons ensuite calculé la connectivité de chaque gène groupé dans une MSP, c'est à dire le nombre d'autre gènes de la MSP avec lesquels le gène a été observé simultanément au moins une fois sur un contig. Enfin, nous avons calculé la connectivité médiane des gènes de chaque MSP.

Nous avons effectué le même traitement sur les clusters aléatoires décrits en 6.4 puis nous avons comparé la connectivité des gènes de ces clusters avec celle calculée pour les MSPs.

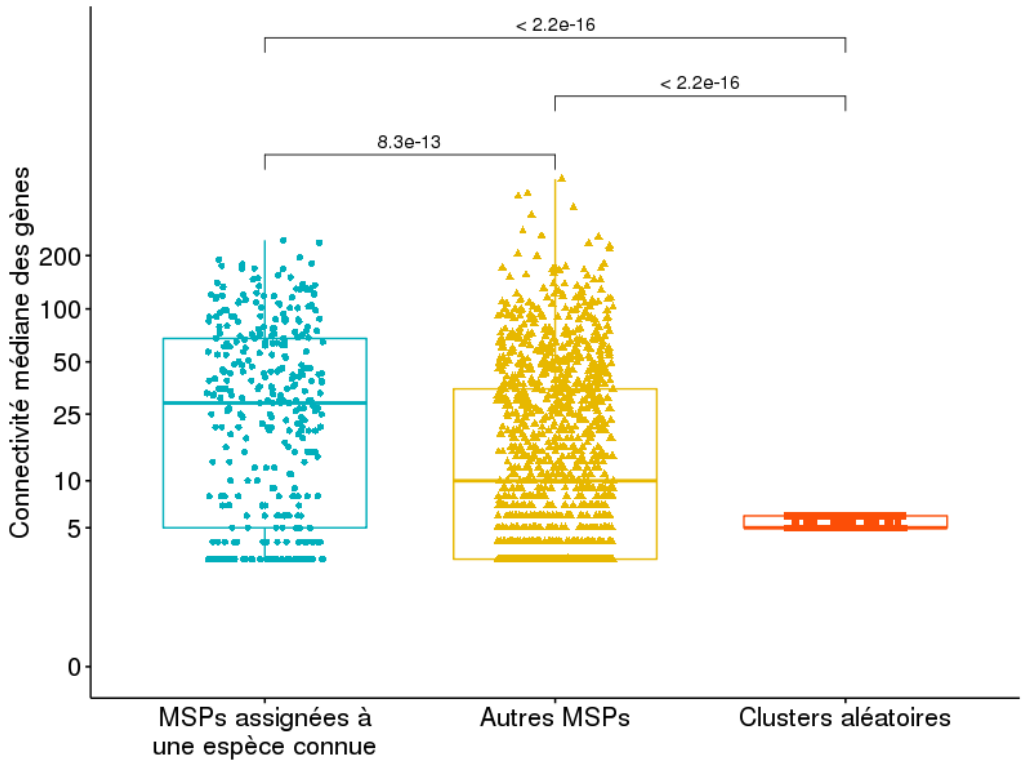


***Figure 46** : Comparaison de la connectivité médiane des gènes groupés dans les MSPs assignées à une espèce connue (bleu), dans les MSPs ayant une annotation taxonomique moins précise (jaune) et dans les clusters aléatoires (rouge). Pour faciliter la lisibilité du graphique, la connectivité médiane des gènes (axe des ordonnées) est représentée en échelle log10.*

La connectivité des gènes dans les MSPs est significativement plus élevée que celle observée dans les clusters aléatoires (moyennes des connectivités médianes 28,7 contre 3,5 ; p-valeur = 0 au test de Wilcoxon-Mann-Whitney). Ainsi, les gènes groupés dans une MSP proviennent vraisemblablement de la même espèce.

Cependant, 652 MSPs (39,4%) ont une connectivité médiane équivalente à celle observée dans les clusters aléatoires (≤ 4) (**Figure 46**). Parmi ces MSPs faiblement connectées, 258 sont assignées à une espèce connue ce qui laisse penser qu'une majorité d'entre elles ne sont pas chimériques. De plus, la connectivité des gènes dans les MSPs est positivement corrélée à leur prévalence (rho de Spearman = 0,45 ; p-valeur < 2.2e-16) et à leur abondance (rho de Spearman = 0,61 ; p-valeur < 2.2e-16). Par conséquent, la faible connectivité de certaines MSPs s'explique par leur faible prévalence et/ou une faible abondance. De même, la différence de connectivité entre les MSPs annotées au niveau espèce et les autres MSPs s'explique par une différence de prévalence (voir 5.5). En effet, les contigs

provenant d'une espèce peu prévalente sont reconstitués dans peu d'échantillons et les assemblages d'une espèce peu abondante sont plus fragmentés ce qui limite la probabilité d'observer des connections entre des gènes.

6.6 Qualité du profilage d'échantillons métagénomiques avec les MSPs

6.6.1 Proportion du signal capturé par les MSPs

En moyenne, 72% des lectures alignées sur le catalogue IGC touchent des gènes regroupés dans les MSPs (**Figure 47.B**). Parmi les 1 267 échantillons considérés, on en recense seulement 2 où la proportion de lectures capturées par les MSPs est inférieure à 50%. Par conséquent, on pourra s'appuyer sur les MSPs pour analyser des échantillons métagénomiques du microbiote intestinal. La majorité du signal biologique sera capturé et ainsi les résultats obtenus seront valides.

On constate que le pourcentage de gènes détectés et groupés dans les MSPs est plus faible que la proportion de lectures capturées par les MSPs (moyenne = 61%, **Figure 47.A**) car cette statistique donne plus de poids aux gènes capturant beaucoup de signal.

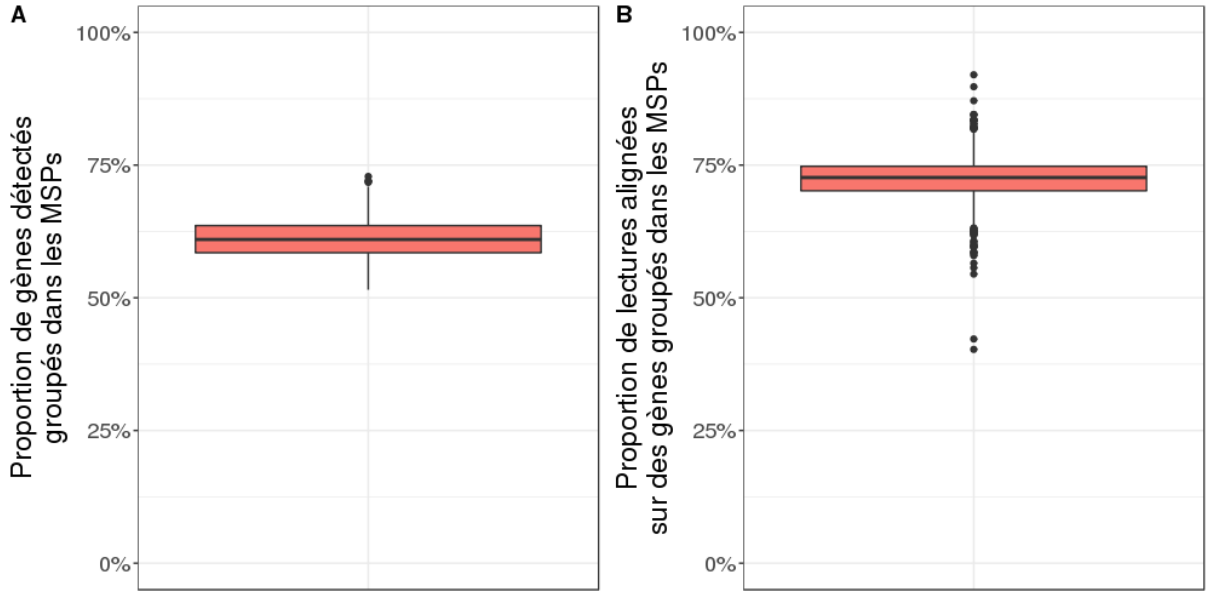


***Figure 47** : Proportion du signal capturé par les MSPs dans les 1267 échantillons du catalogue IGC*

A. Proportion de gènes détectés dans un échantillon (comptage non nul) et groupés dans une MSP.

B. Proportion de lectures alignées sur le catalogue IGC touchant des gènes groupés dans une MSP.

6.6.2 Proportion du signal provenant d'espèces inconnues

Contrairement aux outils de profilage métagénomique s'appuyant sur des génomes de référence comme par exemple MetaPhlAn 2, les MSPs permettent de détecter et quantifier les espèces qui n'ont pas été séquencés jusqu'alors. En moyenne, seules 50% des MSPs détectées dans un échantillon de fèces humain ont déjà été isolées, cultivées et séquencées (**Figure 48.A**). Cependant, 65% (moyenne) des cellules présentes dans des échantillons de fécès sont des microorganismes connus. En effet, les espèces les plus abondantes sont pour l'essentiel déjà connues (**Figure 48.B**).

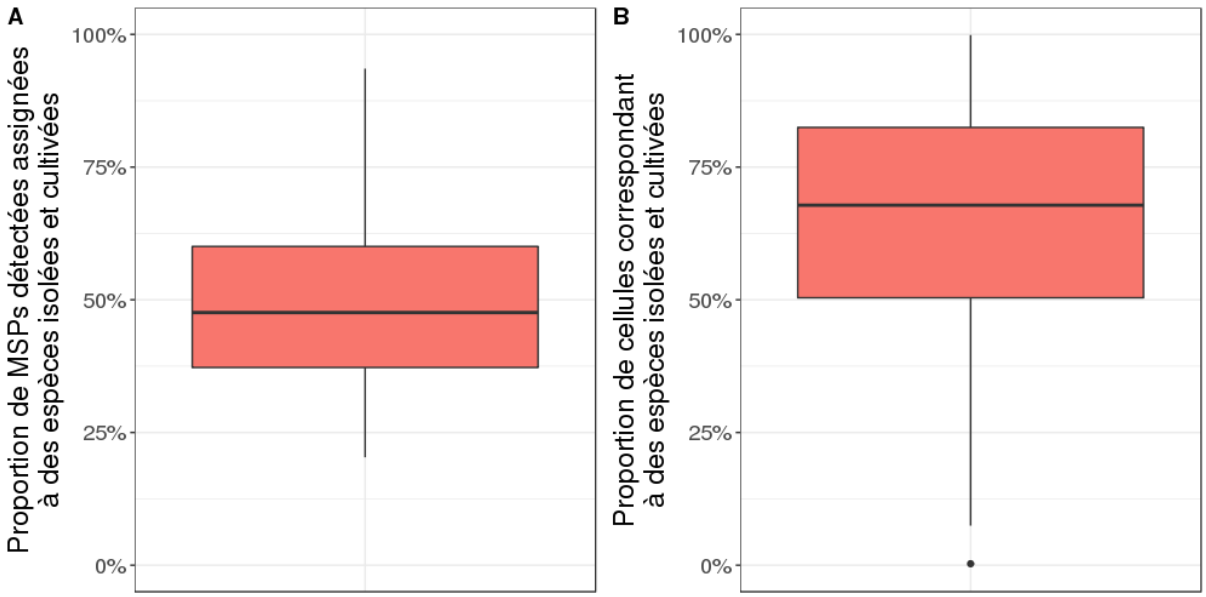


Figure 48 *: Proportion du signal capturé par les MSPs assignées à des espèces connues.*

A. Proportion de MSPs détectées correspondant à des espèces isolées, cultivées et séquencées dans les 1 267 échantillons du catalogue IGC

B. Proportion de cellules détectées correspondant à des espèces isolées, cultivées et séquencées dans les 1 267 échantillons du catalogue IGC.

6.7 Comparaison avec l'algorithme de clustering Canopy

L'algorithme de clustering Canopy [112] a été comparé à MSPminer en appliquant chaque outil à la table d'abondance de gènes du catalogue IGC [113]. Des comptages normalisés ont été utilisés avec Canopy car d'expérience l'outil fournit de meilleurs résultats avec ce type de données.

6.7.1 MSPs vs CAGs : comparaison des objets et de leur contenu en gènes

Au total, MSPminer regroupe 17,8% de gènes de plus que Canopy (3 288 928 contre 2 704 552 gènes) bien que MSPminer ait un critère de sélection des gènes plus strict (6 971 229 contre 7 304 439 gènes considérés lors du clustering).

Ce gain de gènes clustérisés à deux origines. Premièrement, 178 MSPs englobant 154 617 gènes n'ont pas d'équivalent parmi les clusters créés par Canopy que nous appellerons par la suite CAGs (pour Co-Abundance gene Groups). Il faut noter que les clusters ne sont pas manqués par Canopy du fait de leur rareté car leurs gènes respectifs n'ont pas été filtrés. Remarquablement, plus de la moitié de ces MSPs sont composées d'au moins 700 gènes et 60% possèdent plus de la moitié des 40 gènes marqueurs universels [123] ce qui suggère qu'elles correspondent bien à des espèces microbiennes. Remarquablement, pour 75% de ces MSPs, les 3 échantillons avec les comptages les plus forts représentent plus de 90% de la somme des comptages sur les 1267 échantillons (**Figure 49**). Par défaut, Canopy ne considère pas les gènes dont la distribution des comptages est fortement asymétrique à gauche (paramètre *filter_max_dominant_obs*) pour limiter le nombre de faux positifs. Contrairement à Canopy qui s'appuie sur le coefficient de corrélation de Pearson (2 degrés de liberté), MSPminer utilise la mesure de proportionnalité (1 degré de liberté) couplée à une transformation des données qui limite l'asymétrie de la distribution (voir 3.2.2). Ainsi, MSPminer reconstitue des espèces manquées par Canopy tout en gardant une spécificité élevée.

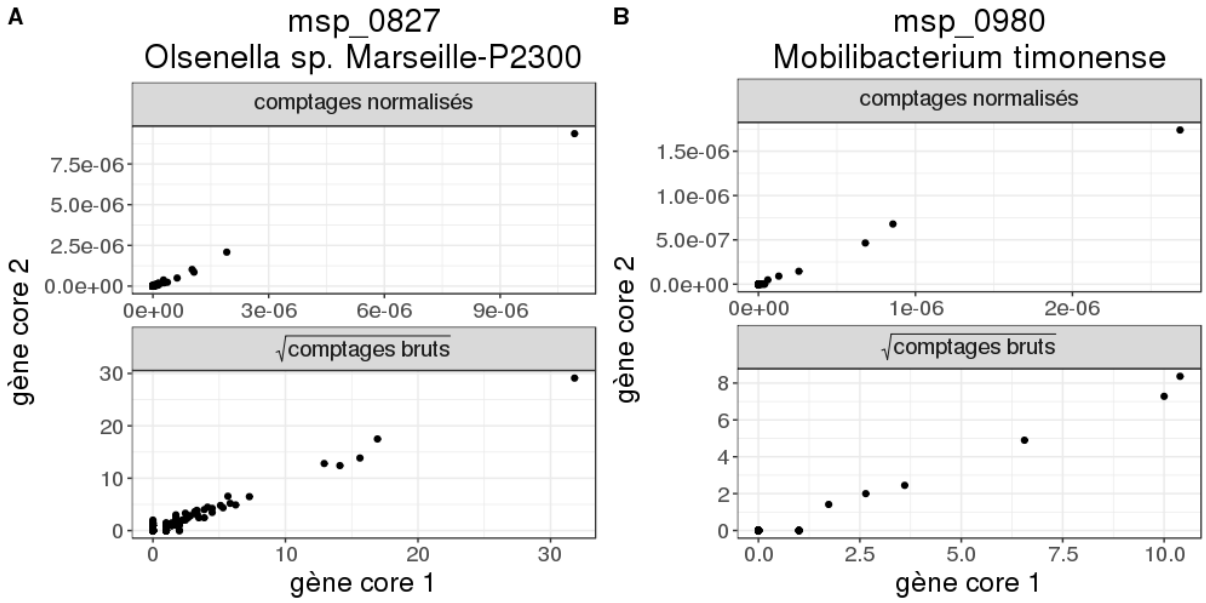


Figure 49 *: Profil d'abondance de 2 MSPs qui n'ont pas d'équivalent parmi les CAGs.*

A gauche on représente respectivement la msp_0827 assignée à Olsenella sp. Marseille-P2300 *; et à droite la msp_0980 assignée à* Mobilibacterium timonense*. Pour chaque MSP, on compare le vecteur d'abondance de deux gènes core. En haut on utilise des comptages normalisés ; et en bas des comptages bruts ayant subi une transformation racine carrée.*

Pour ces deux MSPs, la distribution des comptages des gènes est fortement asymétrique. Quelques échantillons ont des comptages très forts, les autres ont des comptages faibles. Comme exposé en 3.2.2, la transformation racine carrée compresse la distribution et limite l'asymétrie

Deuxièmement, Canopy groupe dans les CAGs de grande taille les gènes core et les gènes accessoires très prévalents des espèces microbiennes. La majorité des gènes accessoires de prévalence plus faible ne sont pas clustérisés. Néanmoins, certains gènes accessoires sont regroupés dans des CAGs distincts de petite taille (**Figure 50** et **Figure 51**). Par conséquent, Canopy produit plus d'objets de plus de 150 gènes que MSPminer (2 010 CAGs vs 1 661 MSPs) car de nombreuses espèces sont éclatées en de multiples clusters. A l'opposé, MSPminer génère une seule MSP par espèce dans la plupart des cas ce qui facilite et augmente la puissance des analyses statistiques réalisées en aval.

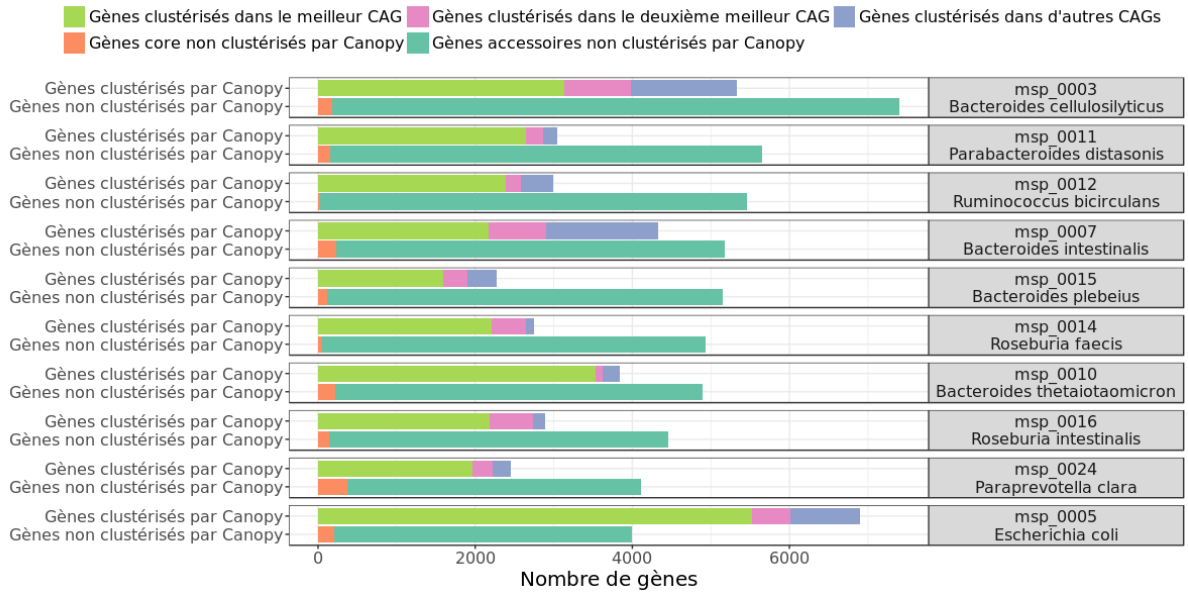


Figure 50 *: Projection des gènes regroupés dans les MSPs sur les CAGs générés par Canopy.*

Pour chaque MSP, on recense les CAGs contenant un sous-ensemble ses gènes. Ces CAGs sont ensuite triés par nombre décroissant de gènes provenant de la MSP. Le CAG regroupant le plus de gènes de la MSP est étiqueté comme meilleur CAG. Chaque diagramme en barres encadré illustre la répartition des gènes d'une MSP dans les CAGs. On distingue systématiquement les gènes clustérisés par Canopy de ceux qui ne le sont pas.

Les gènes regroupés dans une MSP et clusterisés par Canopy proviennent pour la majorité d'un CAG de grande taille (barre verte). Les autres gènes sont regroupés dans des CAGs distincts de plus petite taille. Les gènes regroupés dans une MSP mais manqués par Canopy sont essentiellement des gènes accessoires (barre turquoise).

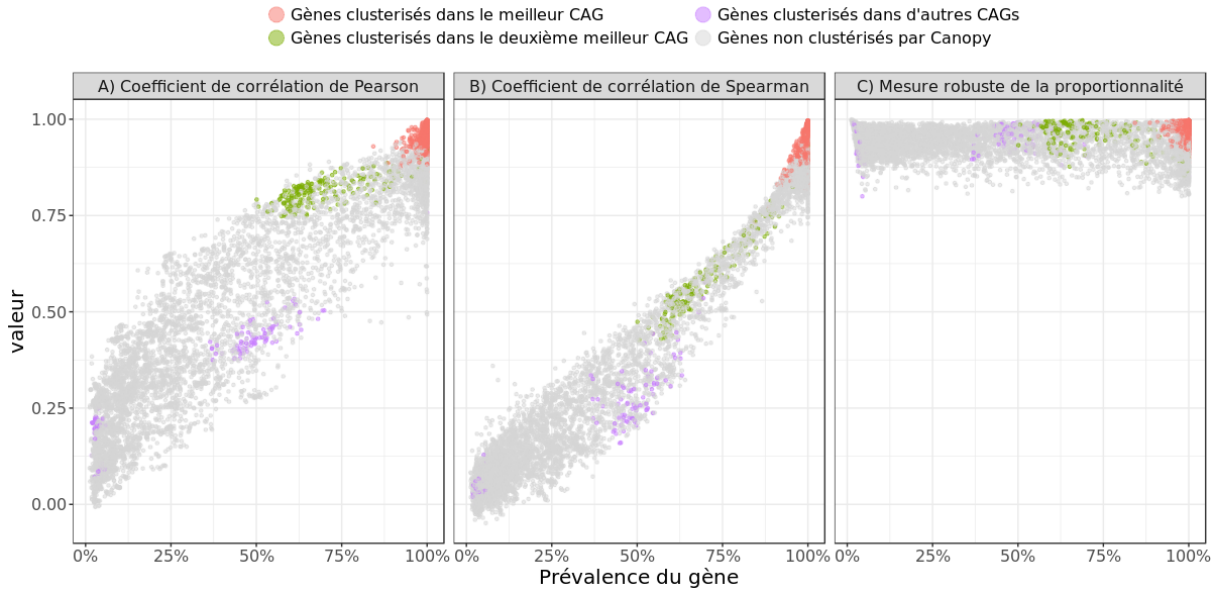


Figure 51 *: Comparaison des vecteurs d'abondance de chaque gène de la msp_0011 (Parabacteroides distasonis) et du vecteur d'abondance médian de ses gènes core en utilisant le coefficient de corrélation de Pearson (A), le coefficient de corrélation de Spearman (B) ou la mesure robuste de proportionnalité*

ρ_r pour détecter un lien entre les vecteurs d'abondance de chaque gène de l'espèce simulée et le vecteur d'abondance médian de ses gènes core.

L'axe des abscisses correspond à la prévalence du gène, c'est-à-dire le pourcentage d'échantillons dans lequel il est détecté. L'axe des ordonnées correspond à l'intensité du lien détecté entre le vecteur d'abondance du gène et le vecteur d'abondance du core génome de l'espèce.

Le CAG contenant le plus de gènes de la msp_0011 regroupe des gènes core et des gènes accessoires très prévalents (>90%) de P. distasonis *(points rouges). Les gènes accessoires moins prévalents ne sont pas clustérisés (points gris) ou groupés dans des CAGs distincts de petite taille (points verts et violets).*

6.7.2 MSPs vs CAGs : comparaison de la spécificité et de la précision

Nous avons comparé la spécificité (homogénéité taxonomique) des CAGs et des MSPs en reprenant la méthodologie définie en 6.1. Pour éviter de pénaliser Canopy, seuls les CAGs de plus de 150 gènes ont été considérés. Globalement, les deux outils ont une précision très élevée (moyenne > 98%). Ceci indique que Canopy et MSPminer génèrent des objets cohérents qui regroupent des gènes provenant de la même espèce.

Ensuite, nous avons comparé la sensibilité (couverture des génomes) des CAGs et des MSPs en s'appuyant sur la méthodologie et les 3 143 génomes décrits en 6.2. Dans un premier temps, nous avons considéré pour chaque génome uniquement le CAG et la MSP le couvrant le mieux. Après avoir pondéré les résultats en fonction du nombre de génomes disponibles par espèce, on remarque que MSPminer est significativement plus sensible que Canopy (médiane : 77% contre 62%, p-valeur test des rangs signés de Wilcoxon = $1{,}7.10^{-16}$, **Figure 52.A**)

Dans la partie 6.7.1, nous avons remarqué que Canopy éclatait le répertoire de gènes d'une espèce en de multiples clusters : les gènes core sont regroupés dans un CAG de grande taille et les gènes accessoires dans de multiples CAGs distincts de petite taille. Ainsi, la différence de sensibilité observée entre les MSPminer et Canopy pourrait être due au fait que l'on considère uniquement les CAGs dominants. En effet, le gain de sensibilité de MSPminer est plus faible si l'on considère tous les CAGs contenant au moins un gène du génome d'intérêt même s'il est toujours significatif (médiane : 77% contre 72%, p-valeur test des rangs signés de Wilcoxon = $9{,}4.10^{-8}$, **Figure 52.B**).

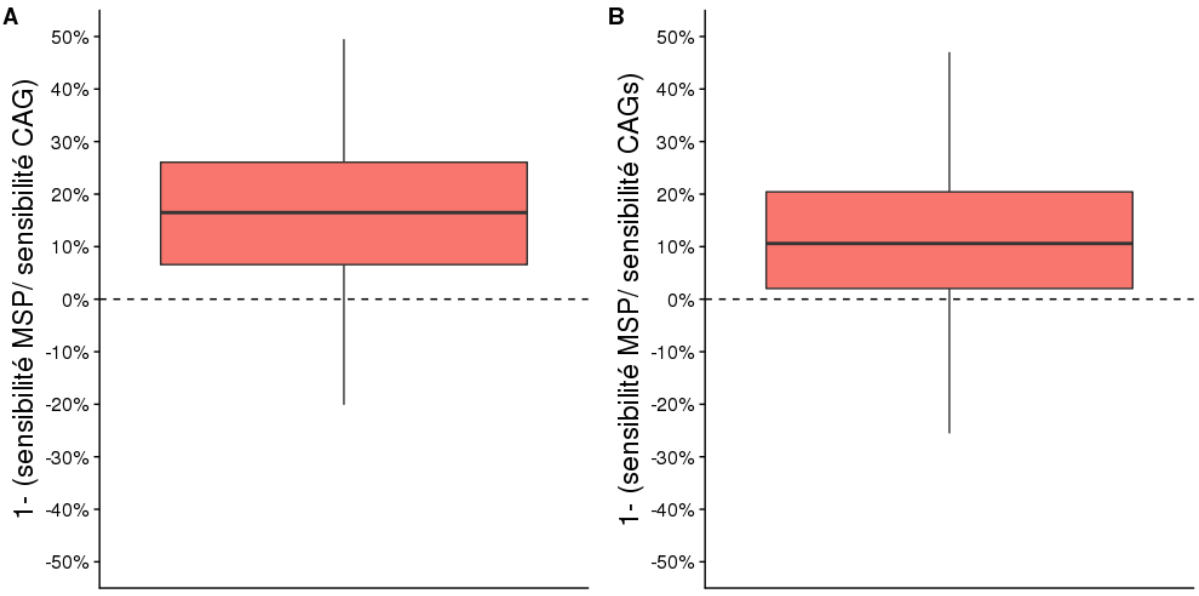


***Figure 52** : Rapport entre la sensibilité des MSPs et de la sensibilité des CAGs pour 322 espèces du microbiote intestinal.*

La sensibilité de chaque espèce est égale à la médiane des sensibilités des génomes la représentant. Une valeur supérieure (respectivement inférieure) à 0% (ligne en pointillés horizontale) indique qu'une espèce est mieux couverte par les MSPs (respectivement les CAGs). Les valeurs supérieures à 50% ont été considérées pour générer les graphiques mais n'ont pas été représentées pour faciliter sa lisibilité.

A. Pour chaque génome, le calcul de la sensibilité est basé uniquement sur la MSP et le CAG le couvrant le mieux.

B. Pour chaque génome, le calcul de la sensibilité est basé sur la MSP le couvrant le mieux et tous les CAGs qui contiennent au moins un de ses gènes.

6.7.3 MSPs vs CAGs : comparaison du potentiel fonctionnel

Par la suite nous avons comparé les fonctions des gènes de 3 143 génomes (c.f. 6.2) avec celles des gènes regroupés leurs CAG/MSP correspondante. Pour ce faire, nous avons aligné les séquences protéiques de chaque gène du catalogue IGC contre la base KEGG (version 82 datée d'Avril 2017) avec blastp. Les résultats dont le bitscore était inférieur à 60 ou dont la e-value excédait 0.01 ont été éliminés. Chaque gène a finalement été assigné au groupe fonctionnel (KEGG Ortholog abrégé KO) associé à l'alignement ayant généré le bitscore le plus élevé. La même procédure d'annotation fonctionnelle a été appliquée aux 3 143 génomes.

Le profil fonctionnel des génomes, CAGs et MSPs a été calculé en évaluant la complétion des 132 modules fonctionnels définis dans GOmixer [94] qu'on appellera par la suite modules GMM. Ensuite, le profil fonctionnel des génomes de référence a été comparé à celui de son CAG/MSP correspondant en utilisant la distance de Manhattan. Lorsqu'une espèce était représentée par plusieurs génomes, une distance égale à la médiane des distances de ses génomes a été calculée.

Comparativement aux CAGs, le profil fonctionnel des MSPs est significativement plus proche de celui des génomes de référence (distance médiane = 3,55 contre 6,65 ; p-valeur au test des rangs signés de Wilcoxon = $3{,}3.10^{-23}$; cf. **Figure 53**)

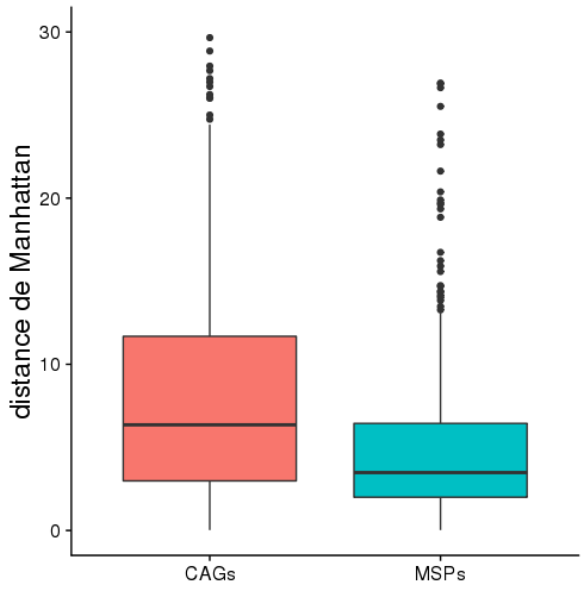


***Figure 53** : Distance de Manhattan entre le profil fonctionnel des génomes de référence et le profil fonctionnel de leur CAG/MSP correspondant.*

Le profil fonctionnel est le vecteur de complétion des 132 modules GMM.

Lorsqu'une espèce est représentée par plusieurs génomes, on ne représente qu'une seule valeur égale à la médiane des distances de ces génomes.

Plus la distance est faible, plus le CAG/MSP a un profil fonctionnel proche du génome de référence.

Certains modules fonctionnels détectés dans les génomes manquent dans les CAGs alors qu'ils sont quasiment tous présents dans les MSPs (nombre de modules manquants médian = 4 contre 1 ; p-valeur au test des rangs signés de Wilcoxon = $8{,}87.10^{-22}$, cf. **Figure 54.A**). Ceci est cohérent avec les résultats obtenus en 6.7.2 montrant que les MSPs couvrent mieux les génomes que les CAGs .

Cependant, certains modules fonctionnels sont détectés dans les MSPs alors qu'ils ne sont pas présents dans les génomes de référence. Remarquablement, ce phénomène se produit moins souvent pour les CAGs (nombre médian de modules inattendus = 0 contre 1 ; p-valeur au test des rangs signés de Wilcoxon = $1{,}41.10^{-5}$). A première vue, on pourrait penser que la spécificité des MSPs est plus faible mais on peut rejeter cette hypothèse avec les résultats obtenus en 6.7.2. Cette différence est probablement due à la présence de gènes accessoires dans les MSPs qui ne sont pas présents dans les génomes testés.

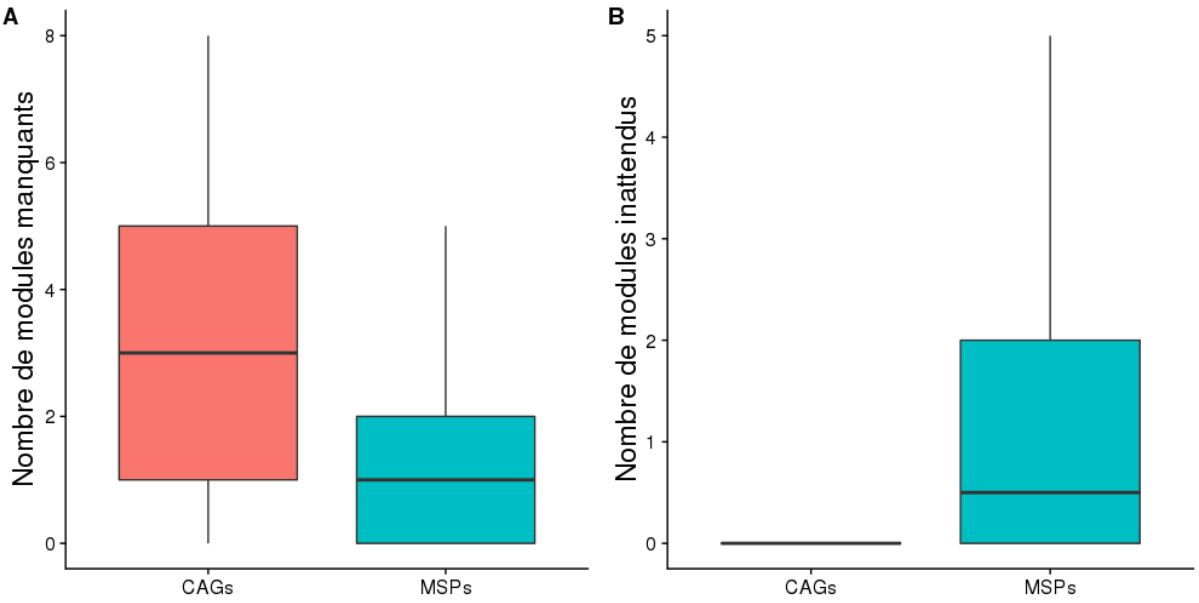


***Figure 54** : Comparaison des profils fonctionnels des génomes de référence et de leur CAG/MSP correspondant.*

A. Recensement des modules GMM détectés dans les génomes mais pas dans leur CAG/MSP représentatifs.

B. Recensement des modules GMM non détectés dans les génomes mais présents dans leur CAG/MSP représentatifs.

Lorsqu'une espèce est représentée par plusieurs génomes, on ne représente qu'une seule valeur médiane.

A titre d'exemple, nous avons comparé le profil fonctionnel du génome de *Methanobrevibacter smithii* TS147C avec ceux de sa MSP (msp_0871) et sa CAG (CAG01222) correspondantes (**Figure 55**). On constate que de nombreux modules d'intérêt du génome comme ceux de la méthanogénèse ou du métabolisme de l'hydrogène sont présents dans la MSP mais pas le CAG.

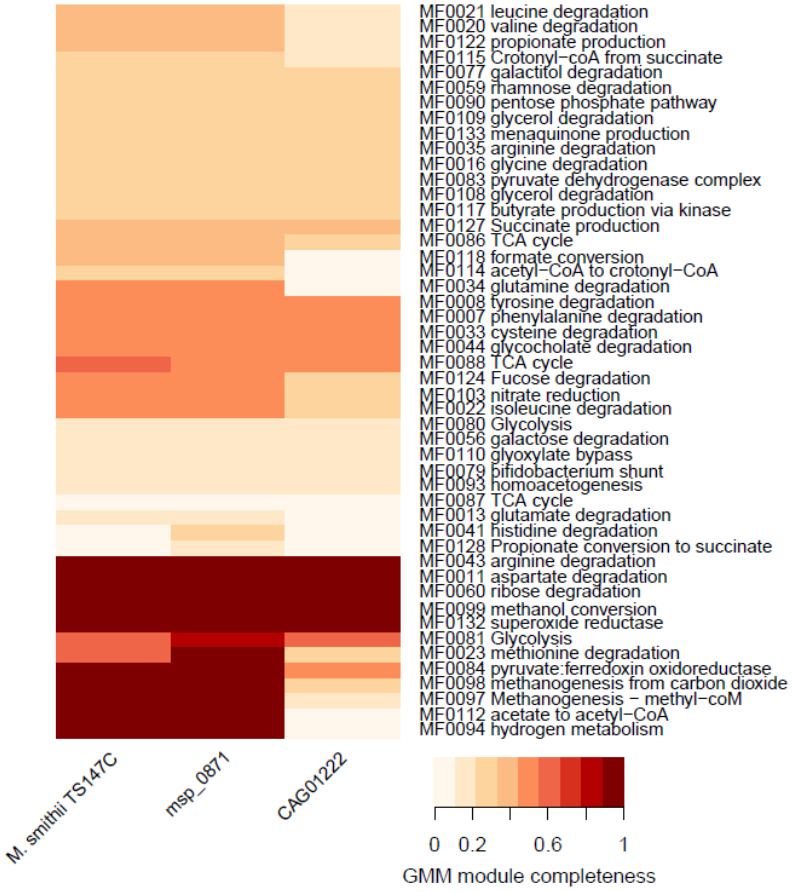


***Figure 55** : Comparaison du profil fonctionnel du génome* M. smithii *TS147C avec ceux de la msp_0871 et la CAG01222.*

Chaque ligne correspond à un module fonctionnel GMM. Un gradient de couleur indique sa complétion dans l'objet considéré.

6.7.4 Comparaison des performances informatiques

Finalement, les performances informatiques de MSPminer ont été comparées à celles de l'algorithme de clustering Canopy [112]. Pour ce faire, les deux outils ont été appliqués à la matrice du catalogue IGC.

MSPminer consomme 3 fois moins de mémoire que Canopy (**Tableau 13**). Cette différence a plusieurs causes. Premièrement, Canopy stocke les comptages de gènes sous la forme de flottants double précision (type double 64 bits) alors que MSPminer utile des flottants simple précision (type float 32 bits) ce qui double la consommation mémoire. Deuxièmement, Canopy précalcule une matrice de corrélation et stocke ces résultats sous la forme de flottants double précision. Les développeurs de Canopy ont choisi de diminuer le temps de calcul au détriment de la consommation mémoire. Finalement, Canopy mappe la totalité du fichier texte en mémoire (fonction mmap) avant de réaliser

les conversions ASCII vers flottant. Ainsi, plusieurs dizaines de Go de mémoire vive supplémentaires sont consommés inutilement car on pourrait lire la matrice ligne par ligne.

Le clustering réalisé par MSPminer nécessite 2h de calcul alors que celui réalisé par Canopy nécessite presque deux jours (**Tableau 13**). Cette différence est essentiellement due à la stratégie de parallèlisation de type Map/Reduce implémentée dans MSPminer (voir 4.1.3) qui permet d'identifier rapidement des clusters de taille importante (> 100 gènes).

Méthode	**Temps de calcul**	**Pic de consommation de mémoire**
MSPminer	2h 05min	74Go
Canopy	41h 52min (× 20)	231Go (× 3,1)

***Tableau 13** : Comparaison des performances de MSPminer et de Canopy pour traiter la matrice de comptage du catalogue IGC (9.9 millions de gènes x 1 267 échantillons).*

Le calcul a été réalisé sur un serveur avec 2 CPUs Intel E5-2690 (2x12 cores) fonctionnant sur le système d'exploitation Centos7. Les performances ont été relevées avec l'utilitaire GNU time.

7. Applications

7.1 Découvertes de biomarqueurs associés à l'origine géographique

7.1.1 MSPs associées à l'origine géographique

Pour démontrer que MSPminer est utile pour la découverte de biomarqueurs, nous avons recherché des MSPs différentiellement abondantes en fonction de l'origine géographique des échantillons.

7.1.1.1 *Méthode*

L'abondance de chaque MSP a été estimée à partir de l'abondance relative médiane de ses 30 meilleurs gènes core (voir 4.1.2). Par la suite, les abondances relatives d'une MSP dans chaque population ont été comparées avec un test de Wilcoxon-Mann-Whitney bilatéral. Les p-valeurs obtenues ont été ajustées avec la procédure de Benjamini-Hochberg [121]. De plus, le ratio $\log_2$-transformé (FC_{log2}) entre les abondances relatives moyennes de la MSP dans les deux populations comparées a été calculé. Les MSPs avec une p-valeur ajustée inférieure à 10^{-2} et un ratio $\log_2$-transformé supérieur à 1 en valeur absolue ont été étiquetées comme différentiellement abondantes.

7.1.1.2 *Résultats*

En s'appuyant sur la cohorte du catalogue IGC, nous avons trouvé 343 MSPs discriminantes entre les échantillons Occidentaux (Européens + Américains) et les échantillons Chinois (**Figure 56**). Toutes celles appartenant au phylum *Proteobacteria* (*Klebsiella pneumoniae*, *Klebsiella quasipneumoniae*, *Escherichia coli* et *Bilophila wadsworthia*) sont plus abondantes chez les Chinois ce qui est cohérent avec des résultats précédemment publiés [113]. Les MSPs annotées comme *Akkermansia muciniphila*, *Dorea longicatena et Methanobrevibacter smithii* font partie des espèces les plus abondantes chez les Occidentaux. Observation intéressante, trois MSPs assignées à l'espèce *Faecalibacterium prausnitzii* sont discriminantes mais deux sont plus abondantes chez les Occidentaux et une chez les Chinois.

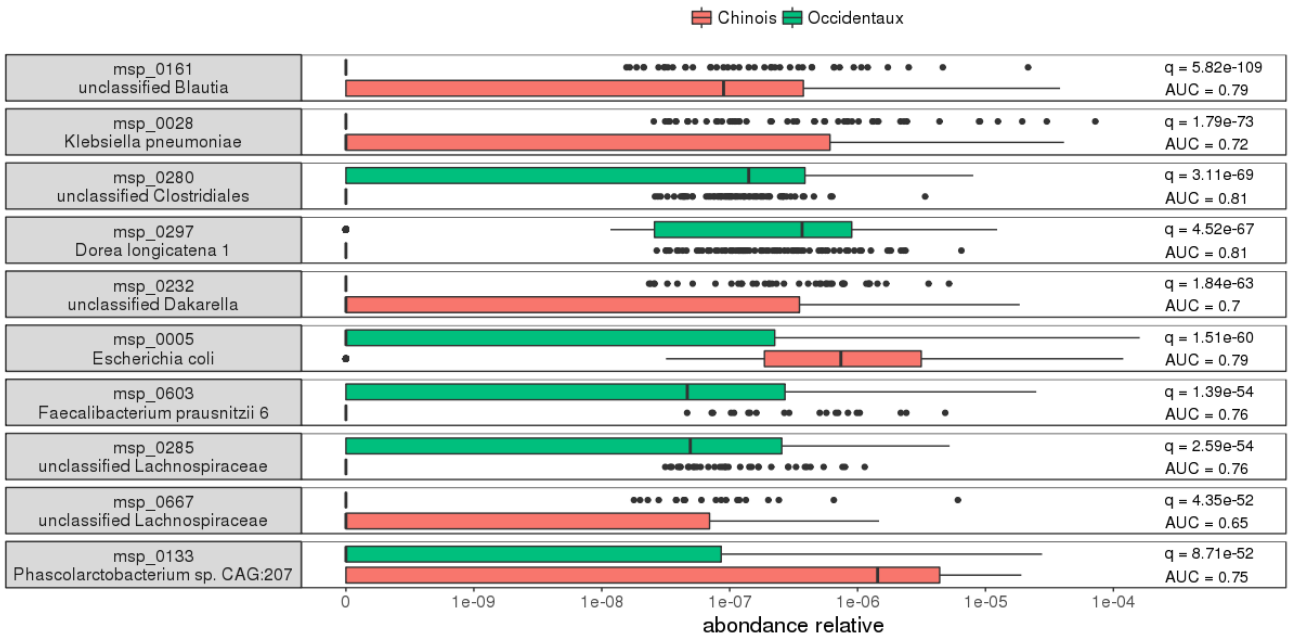


***Figure 56** : Comparaison de l'abondance relative des 10 MSPs les plus discriminantes entre les échantillons Chinois et Occidentaux (Américains + Européens) du catalogue IGC.*

L'abondance relative d'une MSP dans un échantillon correspond à l'abondance médiane de ses 30 meilleurs gènes core. A droite, la q-valeur est la p-valeur du test de Wilcoxon-Mann-Whitney ajustée avec la méthode FDR et l'AUC à l'aire sous la courbe ROC.

De plus, 134 MSPs discriminantes entre les échantillons Européens et Américains ont été découvertes dont 119 (89%) sont plus abondantes chez les Européens (**Figure 57**). Ce résultat est cohérent avec de précédentes études montrer une diversité du microbiote intestinal plus élevée chez les Européens que chez les Américains [123]. 3 MSPs associées du genre *Bacteroides (Bacteroides vulgatus*, *Bacteroides thetaiotaomicron* et *Bacteroides stercoris*) font partie des rares espèces plus abondantes chez les Américains.

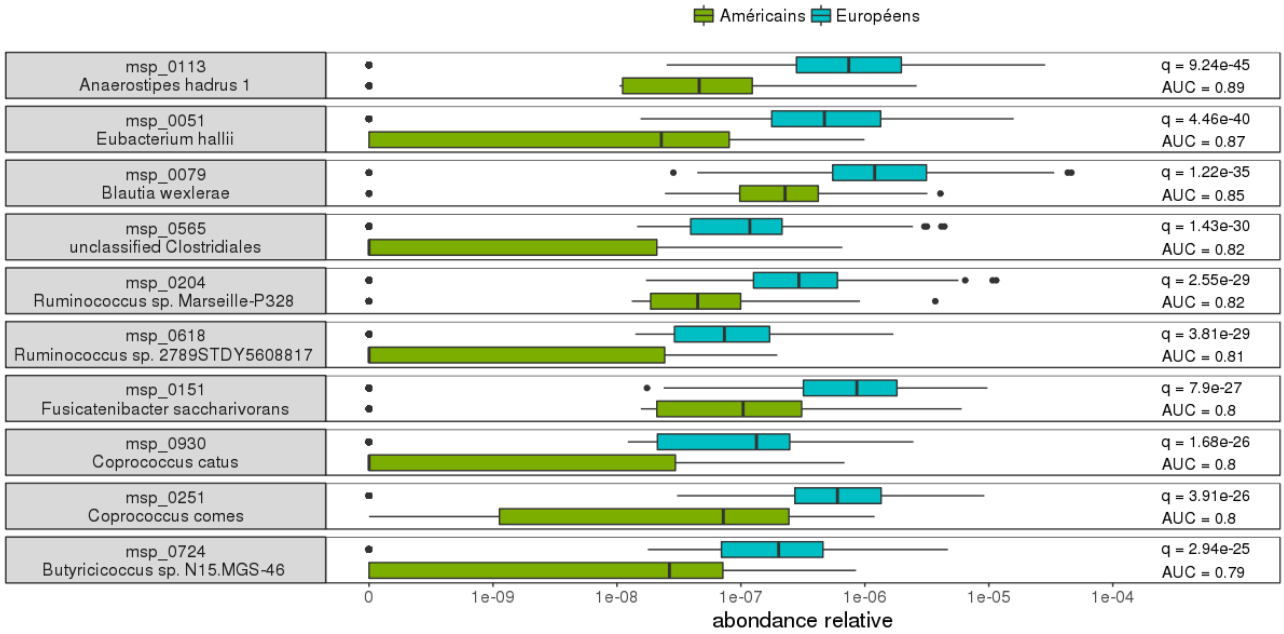


***Figure 57** : Comparaison de l'abondance relative des 10 MSPs les plus discriminantes entre les échantillons Américains et Européens du catalogue IGC.*

L'abondance relative d'une MSP dans un échantillon correspond à l'abondance médiane de ses 30 meilleurs gènes core. A droite, la q-valeur est la p-valeur du test de Wilcoxon-Mann-Whitney ajustée avec la méthode FDR et l'AUC à l'aire sous la courbe ROC.

7.1.2 Gènes accessoires associés à l'origine géographique

Dans un second temps, nous avons effectué une analyse au niveau souche grâce aux MSPs. Pour ce faire, nous avons cherché des gènes accessoires clustérisés dans les MSPs plus fréquents dans des échantillons d'une origine donnée.

7.1.2.1 *Méthode*

• Soit $P = \{p_1, p_2, .., p_n\}$, l'ensemble des n traits phénotypiques d'un caractère. Ici, le caractère considéré est l'origine géographique et les traits associés sont $P = \{\text{Chinois, Européen, Américain}\}$

• Soit $E = \{e_1, e_2, .., e_s\}$, un ensemble de s échantillons métagénomiques et $\{E_1, E_2, .., E_n\}$, une partition de E composée des sous-ensembles d'échantillons disjoints associés à chaque trait de caractère. Dans notre cas, E_1, E_2 et E_3 correspondent respectivement aux échantillons Chinois, Européens et Américains.

• Soit M la MSP contenant les gènes accessoires à tester.

• Notons $g_1 = \left(g_{1,e_1}, g_{1,e_2}, \dots, g_{1,e_s}\right)$ le vecteur du nombre de lectures alignées sur le représentant du core génome de la MSP (voir 4.1.2) dans les échantillons métagénomiques E.

• Notons $g_2 = \left(g_{2,e_1}, g_{2,e_2}, \dots, g_{2,e_s}\right)$ le vecteur du nombre de lectures alignées sur un gène accessoire de la MSP les échantillons métagénomiques E.

On cherche si la présence du gène accessoire est dépendante des traits de caractère associés aux échantillons où la MSP est détectée. Pour ce faire, on crée une table de contingence de dimensions $2 \times n$ suivante :

Caractère / Gène accessoire	Trait p_1	…	Trait p_n
Présent	Nombre d'échantillons possédant le trait p_1 où la MSP et gène accessoire sont simultanément présent	…	Nombre d'échantillons possédant le trait p_n où la MSP et gène accessoire sont simultanément présent
Absent	Nombre d'échantillons possédant le trait p_1 où la MSP est présente mais pas le gène accessoire	…	Nombre d'échantillons possédant le trait p_n où la MSP est présente mais pas le gène accessoire

Soient t_1 et t_2 les seuils de quantification associés respectivement à g_1 et à g_2 (3.3.4). Ces seuils sont utilisés pour considérer uniquement les échantillons où le gène accessoire et le core MSP sont détectés avec certitude et écarter les zéros d'échantillonnage (**Figure 14**, page 42). Formellement, la table de contingence devient :

Caractère / Gène accessoire	Trait p_1	…	Trait p_n
Présent	$\left\lvert \begin{Bmatrix} e \in E_1 \; tels \; que \\ g_{1,e} \geq t_1 \wedge g_{2,e} \geq t_2 \end{Bmatrix} \right\rvert$	…	$\left\lvert \begin{Bmatrix} e \in E_n \; tels \; que \\ g_{1,e} \geq t_1 \wedge g_{2,e} \geq t_2 \end{Bmatrix} \right\rvert$
Absent	$\left\lvert \begin{Bmatrix} e \in E_1 \; tels \; que \\ g_{1,e} \geq t_1 \wedge g_{2,e} = 0 \end{Bmatrix} \right\rvert$	…	$\left\lvert \begin{Bmatrix} e \in E_n \; tels \; que \\ g_{1,e} \geq t_1 \wedge g_{2,e} = 0 \end{Bmatrix} \right\rvert$

Par la suite, un test du chi-deux prenant en entrée la table de contingence est effectué. Finalement, on considère que la présence du gène accessoire est dépendante du caractère considéré lorsque la p-valeur obtenue est inférieure à 10^{-10}. Cette procédure est utilisée pour tester chaque gène accessoire de chaque MSP.

7.1.2.2 *Résultats*

Nous avons découvert 54 MSPs possédant au moins 200 gènes accessoires associés à l'origine géographique des échantillons (**Tableau 14**). Parmi ces MSP, 24 ne sont pas annotées au niveau espèce montrant qu'une analyse au niveau souche est possible y compris pour les espèces non séquencées à ce jour.

msp	**Annotation taxonomique**	**Nombre de gènes accessoires associés à l'origine géographique des échantillons**
msp_0027	*Eubacterium eligens*	1313
msp_0187	*Coprobacillus* non classifiée	1215
msp_0014	*Roseburia faecis*	1153
msp_0149	*Ruminococcus torques*	1152
msp_0070	*Eubacterium rectale*	1148
msp_0058	non annotée	788
msp_0078	non annotée	712
msp_0011	*Parabacteroides distasonis*	689
msp_0143	*Ruminococcus sp. SR1/5*	619
msp_0086	*Firmicutes* non classifiée	594

***Tableau 14** : Liste des 10 MSPs possédant le plus de gènes accessoires associés à l'origine géographique des échantillons.*

A titre d'exemple, la **Figure 58** illustre les gènes accessoires de la msp_0027 dont la présence est dépendante de l'origine géographique des échantillons. On remarque que les souches des individus Chinois ont un contenu en gènes différent de celles des individus Occidentaux. Ceci révèle l'existence de deux sous-espèces d'*Eubacterium eligens* associées à l'origine géographique.

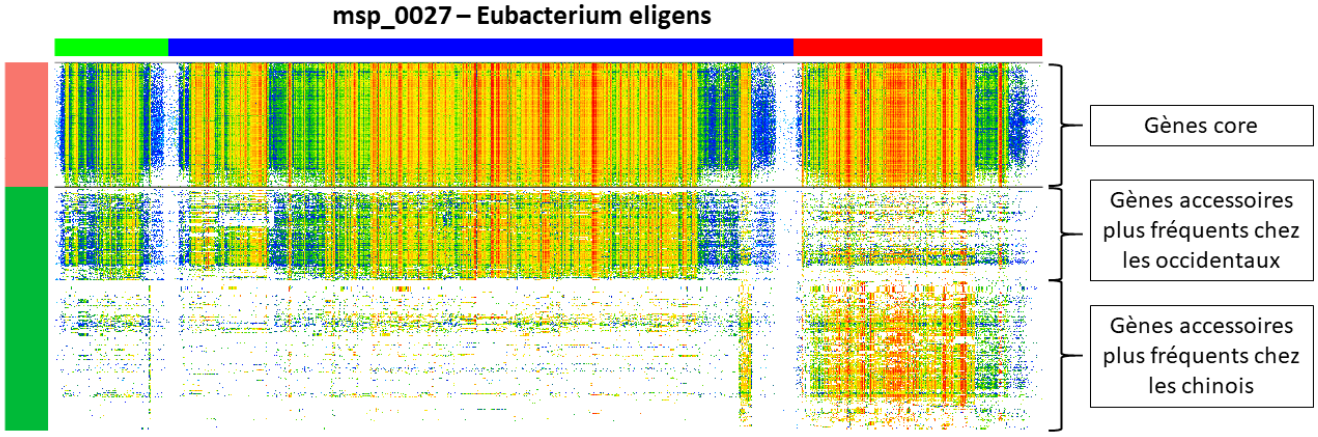


***Figure 58** : Heatmap représentative de l'abondance relative des gène core et des gènes accessoires associés à l'origine géographique de la msp_0027 (*Eubacterium eligens*)*

Chaque colonne représente un échantillon. Les échantillons bleus, rouges et verts correspondent respectivement aux Européens, aux Chinois et aux Américains. Chaque ligne représente un gène. Les gènes rouges, kakis, cyans et violets correspondent respectivement aux gènes core, accessoires, core partagés et accessoires partagés. Chaque case de la heatmap indique l'abondance relative d'un gène dans un échantillon donné suivant un gradient de couleur (blanc=absence, bleu=abondance faible, rouge=abondance élevée)

7.2 Caractérisation du microbiote intestinal après chirurgie bariatrique

7.2.1 Préambule

L'obésité est une maladie définie par l'OMS comme « une accumulation anormale ou excessive de graisse corporelle qui peut nuire à la santé » [165]. Elle est associée à de nombreuses autres maladies comme le diabète de type 2, l'hyperlipidémie, l'hypertension et l'apnée du sommeil. Au cours des dernières décennies, la prévalence de l'obésité a augmenté spectaculairement dans le monde entier. Ainsi, l'obésité est aujourd'hui considérée comme une épidémie mondiale.

La chirurgie bariatrique (chirurgie du traitement de l'obésité) est à ce jour la stratégie la plus efficace pour les patients atteints d'obésité morbide (IMC ≥ 40 kg/m²) avec une perte de poids significative et un taux de réussite supérieur à 66%. Les deux principales interventions chirurgicales sont la gastrectomie par laparoscopie (LSG) ou la dérivation gastrique de Roux-en-Y (LRYGB) (**Figure 59**).

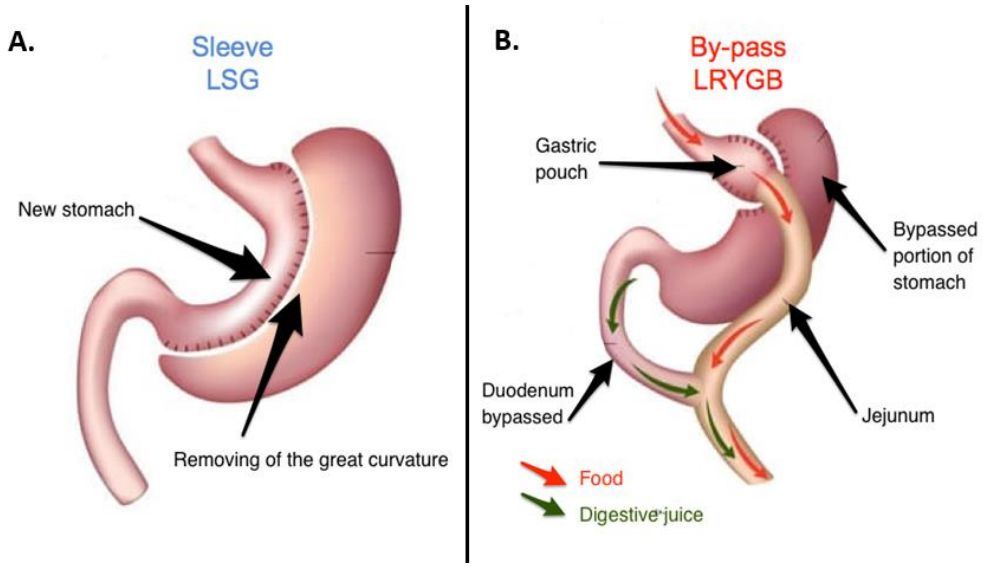


***Figure 59** : Représentation schématique des deux principaux types de chirurgie bariatrique*

A. Principe de la gastrectomie par laparoscopie ou en anglais Laparoscopic Sleeve Gastrectomy (LSG). Cette opération consiste à retirer une grande partie de l'estomac le long de la grande courbure. La LSG est une procédure irréversible qui n'implique pas de réarrangement de l'intestin.

B. Principe de la dérivation gastrique de Roux-en-Y ou en anglais Laparoscopic Roux-En-Y Gastric Bypass (LRYGB). Cette opération consiste à diviser l'estomac en une petite poche supérieure et une grande poche inférieure. Ensuite, la partie centrale de l'intestin grêle (jejunum) est directement reliée à la petite poche supérieure. Après dérivation, la nourriture ne circule ni dans la grande poche inférieure ni dans le segment initial de l'intestin grêle (duodénum).

Nous présentons la plus grande étude métagénomique explorant l'impact des chirurgies LRGYB et LSG sur le microbiote intestinal. Plus précisément, le but de cette étude est de caractériser les changements taxonomiques et fonctionnels induits par les deux types de chirurgies et mettre en évidence leurs différences. Une meilleure compréhension des effets des chirurgies permettrait de choisir celle la plus appropriée pour les patients et ouvrirait la voie à une médecine personnalisée.

Dans ce but, des selles de patients obèses (89 LRGYB et 108 LSG) de différentes origines géographiques (France, Suisse et USA) ont été prélevées avant l'opération et 6 mois après. Le microbiote a ensuite été caractérisé par séquençage métagénomique shotgun. En utilisant le logiciel METEOR [117], les lectures générées ont été filtrées puis alignées sur le catalogue IGC pour finalement générer une table de comptage des gènes. Dans un premier temps, nous avons effectué deux analyses statistiques indépendantes pour les chirurgies LRYGB et LSG puis nous avons comparé leur impact.

7.2.2 Impact sur la composition en espèces du microbiote

7.2.2.1 *Méthodes*

La table d'abondance des gènes a été normalisée par la longueur des gènes et la nombre de lectures alignées dans chaque échantillon. Ensuite, l'abondance des 1661 MSPs a été estimée à partir l'abondance relative médiane de leurs 30 meilleurs gènes core respectifs (voir 4.1.2). Pour rechercher des MSPs significativement modulées par la chirurgie, un test statistique non paramétrique a été utilisé sur les abondances relatives $\log_{10}$ transformées. Comme les individus sont prélevés avant et après chirurgie, les échantillons sont appariés. Par conséquent, un test des rangs signés de Wilcoxon bilatéral a été utilisé. Les p-valeurs obtenues ont été ajustées avec la procédure de Benjamini-Hochberg [121]. Pour s'assurer que les MSPs étaient bien significatives, le taux d'accroissement $\log_2$-transformé (FC_{log2})

défini comme le rapport entre l'abondance médiane de la MSP chez tous patients avant chirurgie divisé par l'abondance médiane de la MSP chez tous patients après chirurgie a été calculée. Les MSPs avec une p-valeur ajustée inférieure à 0.05 et taux d'accroissement (fold change) supérieur à 1 en valeur absolue ont été considérées comme statistiquement significatives.

7.2.2.2 *Résultats*

51 MSPs ont significativement été impactées après la chirurgie LRYGB (**Figure 60**) et 48 par la chirurgie LSG (**Figure 61**). Parmi ces MSPs, 20 sont communes aux deux interventions et sont pour leur majorité (18/20) enrichies après chirurgie. Remarquablement, l'abondance de la bactérie bénéfique *Akkermansia muciniphila* du phylum *Verrucomicrobia* croit après l'intervention (LRYGB=1,61 FC_{log2} ; LSG=1,82 FC_{log2}), mais aussi celles de bactéries potentiellement pro-inflammatoires du phylum *Proteobacteria* comme *Escherichia coli* (LRYGB=5,22 FC_{log2} ; LSG=1,04 FC_{log2}), *Klebsiella pneumoniae* (LRYGB=4,03 FC_{log2} ; LSG=2,09 FC_{log2}) et *Haemophilus parainfluenzae* (LRYGB=1,61 FC_{log2} ; LSG=1,4 FC_{log2}). On note aussi une baisse de l'espèce anti-inflammatoire *Faecalibacterium prausnitzii* seulement après une chirurgie LRYGB (-1.43 FC_{log2}). 5 espèces généralement détectées dans la cavité orale sont enrichies par les deux interventions dont particulièrement *Veillonela parvula* (LRYGB=2,65 FC_{log2} ; LSG=1,92 FC_{log2}) et *Streptococcus salivarius* (LRYGB=4,07 FC_{log2} ; LSG=1,98 FC_{log2}) mais aussi *Streptococcus gordonii*, *Streptococcus mutans* et *Streptococcus parasanguinis.* Six autres espèces orales (*Fusobacterium nucleatum*, *Streptococcus anginosus*, *Streptococcus oralis*, *Streptococcus vestibularis*, *Veillonela atypica* et *Veillonela sp oral*) sont enrichies uniquement chez les LRYGB. Fait intéressant, deux espèces orales du genre *Bifidobacteria* sont appauvries par les interventions chirurgicales : une par LRYGB (*Bifidobacteria bifidum*) et une par LSG (*Bifidobacteria dentium*). Finalement, quatre MSPs sans annotation au niveau espèce sont enrichies par les deux interventions.

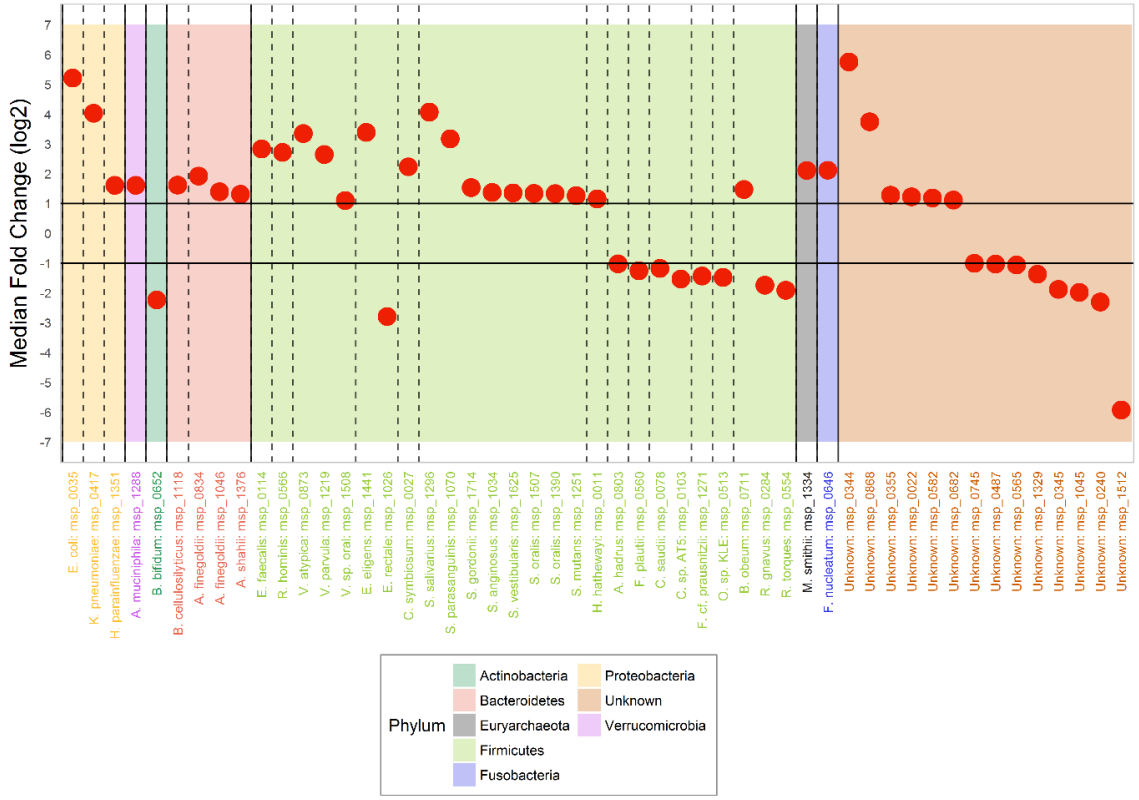


***Figure 60** : Taux d'accroissement des abondances relatives de 51 MSPs significativement impactées par la chirurgie LRYGB.*

Les MSPs sont groupées par phylum puis ordonnées par taux d'accroissement médian décroissant en valeur absolue. Les traits pleins verticaux séparent les phyla et les traits en pointillés les genres. Les traits pleins horizontaux correspondent aux seuils de significativité.

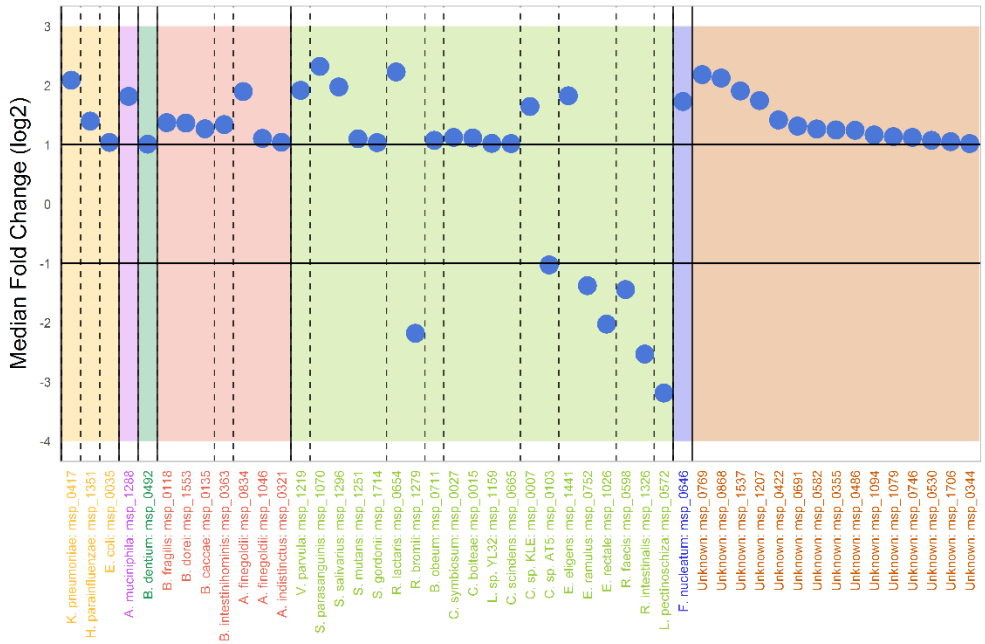


***Figure 61** : Taux d'accroissement des abondances relatives de 49 MSPs significativement impactées par la chirurgie LSG.*

*Les MSPs sont groupées par phylum puis ordonnées par taux d'accroissement médian décroissant en valeur absolue. Les couleurs associées à chaque phylum sont décrites dans la **Figure 60**.*

7.2.3 Impact sur le potentiel fonctionnel du microbiote

7.2.3.1 *Méthodes*

Les séquences nucléotidiques des gènes du catalogue IGC ont été traduites en séquences protéiques avec l'outil sequence-translator [166]. Les gènes traduits ont ensuite été alignés avec blastp contre la base KEGG [92] (version 82 datée d'Avril 2017). Les résultats dont le bitscore était inférieur à 60 ou dont la e-value excédait 0.01 ont été éliminés. Chaque gène a finalement été assigné au groupe fonctionnel (KEGG Ortholog abrégé KO) associé au résultat le plus significatif.

L'abondance relative de chaque KO a été calculée en sommant les abondances relatives des gènes qui lui été assignés. Enfin, les abondances des 132 modules fonctionnels définis dans GOmixer [94] ont été calculés en sommant les abondances de chaque KO les composant.

Pour établir un lien entre les changements fonctionnels et les changements taxonomiques, des corrélations de Spearman entre les profils d'abondance des 13 modules fonctionnels et ceux des 78 MSPs impactées par au moins une des chirurgies ont été calculées indépendamment sur les deux groupes de patients (LRYGB et LSG). Les corrélations supérieures à 0.7 et dont la p-valeur est inférieure à 10^{-5} ont été considérées comme significatives.

7.2.3.2 *Résultats*

13 modules fonctionnels sont significativement plus abondants après une chirurgie LRYGB (**Figure 62**) et 6 après LSG (**Figure 63**). 5 modules appartenant à la famille des transporteurs ABC sont communs aux deux interventions dont ceux impliqués dans le transport de la vitamine B12 (LRYGB=5,23 FC_{log2} ;

LSG=1,92 FC_{log2}), de l'histidine (LRYGB=5,51 FC_{log2} ; LSG=1,18 FC_{log2}), de la lysine/arginine (LRYGB=5,16 FClog2 ; LSG=1,16 FClog2), de la putrescine (LRYGB=4,95 FC_{log2} ; LSG=1,6 FC_{log2}) et du manganèse/zinc (LRYGB=3,30 FC_{log2} ; LSG=2,04 FC_{log2}).

Les transporteurs de la Vitamine B1 (5,23 FC_{log2}), de l'urée (2,01 FC_{log2}) ainsi que les modules de dénitrification (1.0 FC_{log2}) et de production de l'acide gras volatil propanoate (2,53 FC_{log2}) sont enrichis uniquement après une chirurgie LRYGB tandis que le module de dégradation du glutamate est enrichi uniquement après une chirurgie LSG (1,39 FC_{log2}).

67% (4/6) des modules enrichis chez les patients LSG et 77% (10/13) de ceux enrichis chez les patients LRYGB sont liés à la hausse d'*Escherichia coli* ou d'une espèce orale parmi *S. parasanguinis, S.salivarius, S. vestibularis* et *V. parvula*.

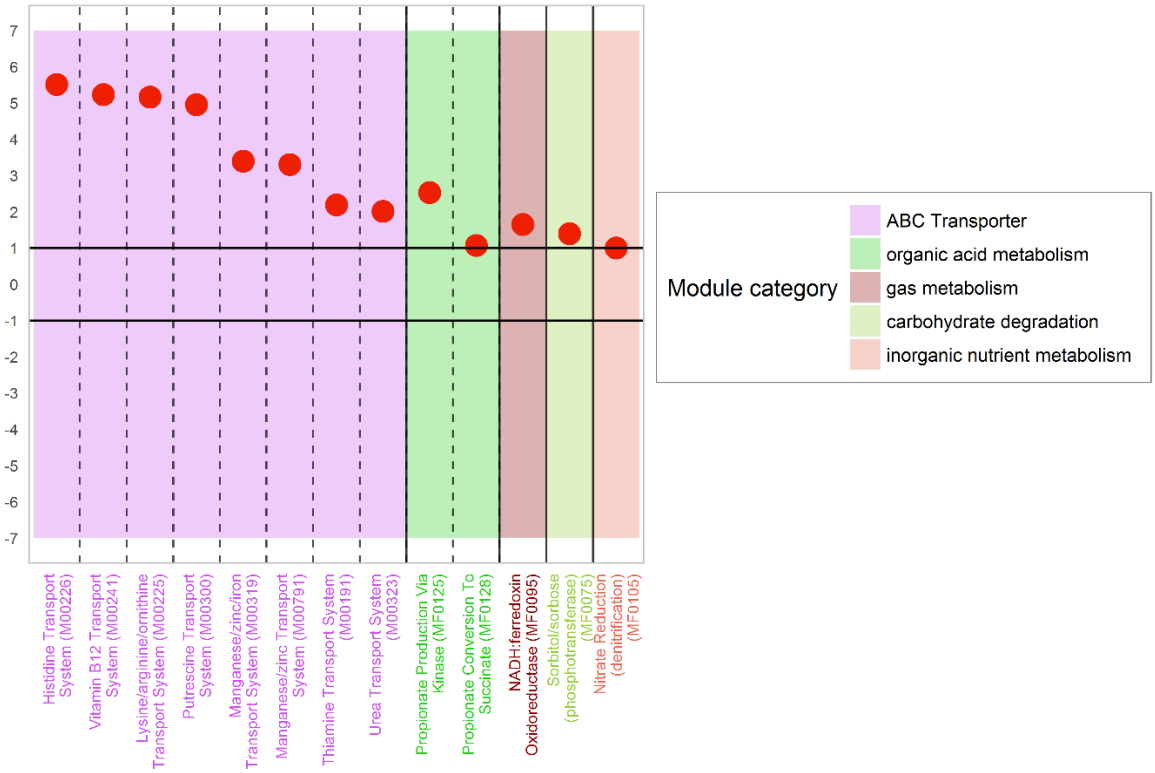


***Figure 62** : Taux d'accroissement des abondances relatives des 13 modules fonctionnels significativement impactées par la chirurgie LRYGB.*

Les modules sont groupés par catégorie puis ordonnés par taux d'accroissement médian.

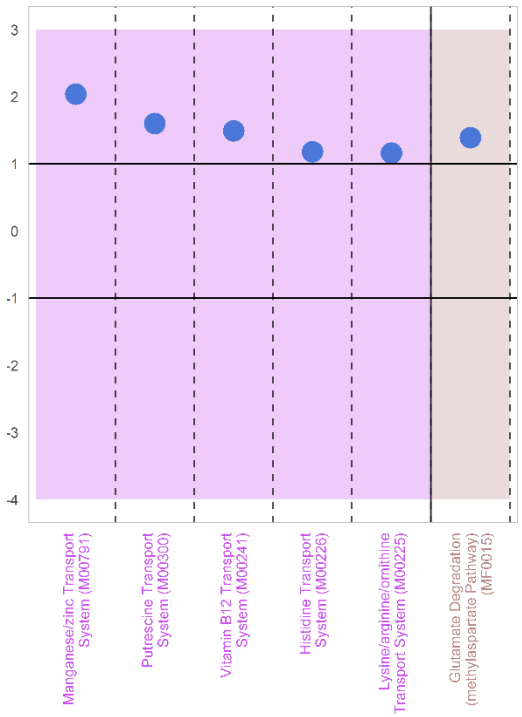


***Figure 63** : Taux d'accroissement des abondances relatives des 6 modules fonctionnels significativement impactées par la chirurgie LSG.*

*Les modules sont groupés par catégorie par catégorie puis ordonnés par taux d'accroissement médian. Les couleurs associées à chaque catégorie fonctionnelle sont décrites dans la **Figure 62**.*

7.2.4 Comparaison de l'effet des chirurgies

Dans un second temps, nous avons comparé l'impact des chirurgies LRYGB et LSG sur le microbiote intestinal. Plus précisément, on recherche des MSPs ou des modules fonctionnels différemment impactées par les deux types de chirurgie.

7.2.4.1 *Méthodes*

Seules les MSPs ou les modules fonctionnels significatifs pour au moins un type de chirurgie ont été considérés. Les abondances relatives post-opération de ces objets ont été divisées par les abondances pré-opération puis $\log_2$ transformées. Ces ratios d'abondance ont été utilisés pour comparer les deux types de chirurgie en effectuant un test Wilcoxon-Mann-Whitney bilatéral. Les p-valeurs obtenues ont ensuite été ajustées avec la procédure de Benjamini-Hochberg. Finalement, les MSPs/modules avec une p-valeur ajustée inférieure à 0.05 ont été considérés comme statistiquement significatifs.

7.2.4.2 *Résultats*

Parmi les 78 MSPs impactées par au moins un type de chirurgie, 24 ont des ratios d'abondance significativement différents entre les groupes LRYB et LSG (**Figure 64**). Deux espèces du phylum *Proteobacteria* (*Escherichia coli* et *Klebsiella pneumoniae*) et 6 espèces du microbiote oral (*Streptococcus oralis*, *Streptococcus parasanguinis*, *Streptococcus salivarius*, *Veillonella atypica, Veillonella parvula* et *Veillonella sp. oral*) sont significativement plus enrichies chez les LRYGB que chez les LSG. Cependant, l'abondance de l'espèce orale *Bifidobacterium bifid*um est plus diminuée chez les LRYGB. Deux espèces du genre *Roseburia* (*R. faecis* et *R. hominis*) sont significativement plus enrichies par la chirurgie LRYGB tout comme *Enterococcus faecalis*. Enfin, 5 MSPs assignées à des espèces de l'ordre *Clostridiales* sont plus enrichies après une chirurgie LSG.

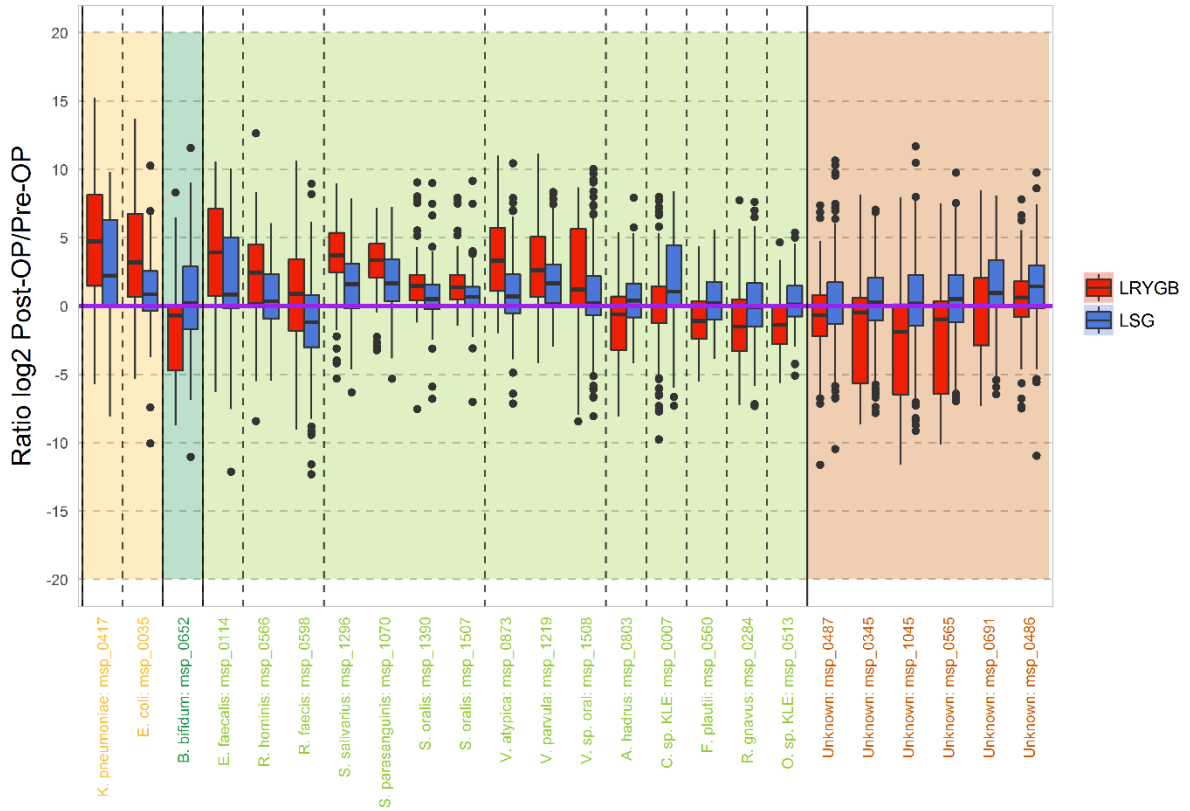


***Figure 64** : Ratios des abondances post-opération/pré-opération de 78 MSPs différemment impactées par les deux types de chirurgie.*

Sur les 14 modules impactés par au moins un type de chirurgie, 13 ont des ratios d'abondance significativement différents entre les groupes LRYB et LSG (**Figure 65**). Remarquablement, 8 modules fonctionnels appartenant à la famille des transporteurs ABC sont significativement plus abondants après une chirurgie LRYGB ainsi que les modules impliqués dans la dénitrification et la production de propionate.

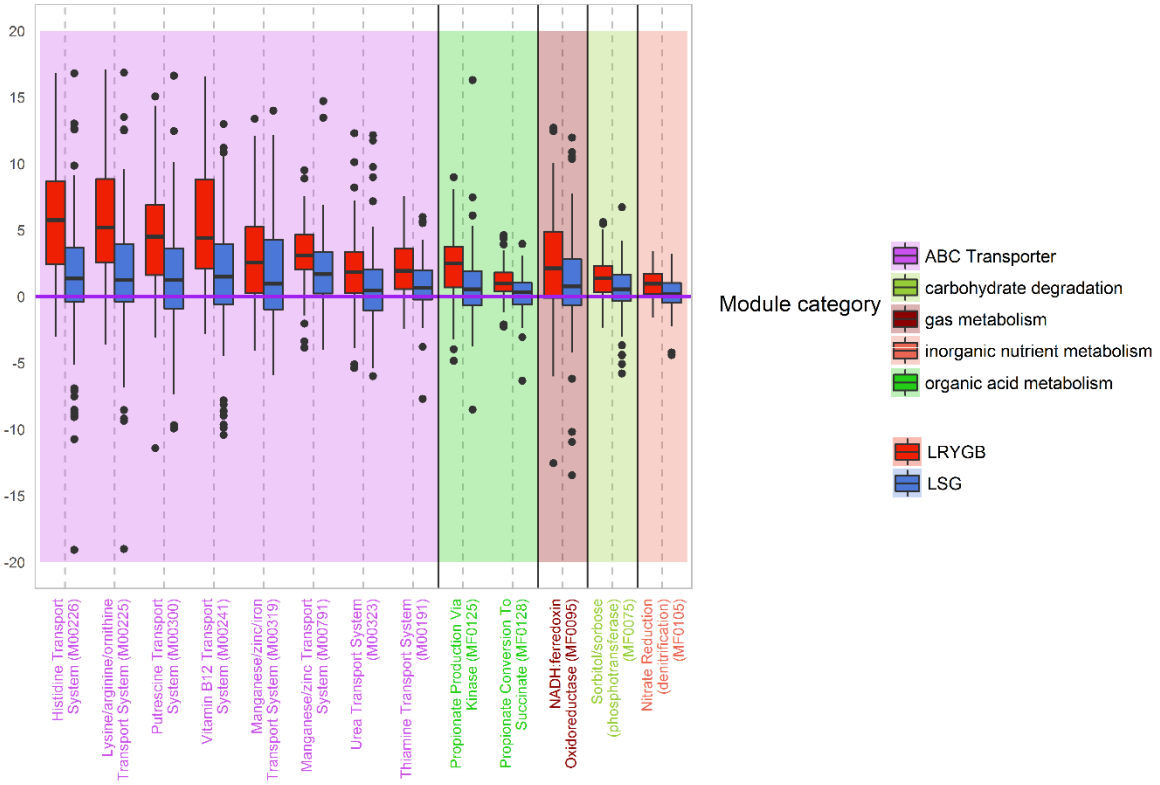


***Figure 65** : Ratios des abondances post-opération/pré-opération des 13 modules fonctionnels différemment impactées par les deux types de chirurgie.*

7.2.5 Discussion et perspectives

Le séquençage métagénomique des selles de 197 patients montre que la chirurgie bariatrique impacte fortement le microbiote intestinal 6 mois après l'intervention. La nature et l'ampleur des changements observés dépendent grandement du type chirurgie. Globalement, la chirurgie LYRGB a un impact plus important sur la composition taxonomique et le potentiel fonctionnel du microbiote intestinal.

Premièrement, la chirurgie LRYGB entraine une hausse plus importante des espèces colonisant la cavité orale (genres *Veillonela* et *Streptococcus*). Il est possible qu'une exposition limitée au suc gastrique favorise la colonisation de l'intestin par des bactéries orales. Cependant, l'appauvrissement des espèces orales du genre *Bifidobacteria* suggère qu'un accès facilité à l'intestin est insuffisant à l'implantation de microorganismes et que d'autres facteurs entre en jeu. Le contournement du duodédum lors de la chirurgie LRYGB introduit du dioxygène dans le tractus digestif ordinairement anaérobie. Le dioxygène inhiberait les espèces anaérobies strictes et promouvrait la croissance des microorganismes aérotolérants ou aérobies dont font partie les espèces orales. Dans notre étude, les espèces anaérobies de l'ordre des *Clostridiales* sont négativement impactées par la chirurgie LRYGB alors que leur abondance augmente par LSG suggérant que l'intestin est toujours un milieu anaérobie après chirurgie. Dans le même ordre d'idée, l'abondance relative du module de la ferrédoxine-NADP+ réductase généralement associée à une respiration aérobie est plus élevée chez les patients LRYGB.

L'autre différence remarquable concerne l'ampleur de la hausse des espèces du phylum *Proteobacteria*. En effet, les espèces *E. coli* et *K. pneumoniae* sont enrichies après chirurgie mais la hausse est bien plus importante après une chirurgie LRYGB. Selon certains auteurs, l'enrichissement en *E. coli* résulterait d'une adaptation de l'hôte et du microbiote intestinal pour maximiser la collecte d'énergie dans un contexte de restriction calorique consécutif à la chirurgie bariatrique. Fait

intéressant, il a récemment été montré que les nitrates accélèrent la croissance d'*E. coli* qui entre en concurrence avec les espèces qui réalisent uniquement de la fermentation aérobie. Or, une hausse du module de réduction des nitrates est observée après une chirurgie LRYGB.

De plus, on observe une hausse importante des modules fonctionnels de la famille des transporteurs ABC et plus particulièrement ceux de la vitamine B12, la vitamine B1 (thiamine) et du fer/manganèse après chirurgie. Comme mentionné dans de précédentes études, cette hausse est plus marquée pour les patients LRYGB. Après chirurgie, les patients sont supplémentés en vitamines, fer et calcium pour compenser une diminution de l'apport alimentaire et une malabsorption des nutriments. Les niveaux élevés des modules fonctionnels liés au transport de la vitamine B12, de la thiamine et du fer (en particulier dans LRYGB) suggèrent que des microorganismes opportunistes utilisent ces nutriments. La corrélation entre l'abondance du module de la vitamine B12 et *E. coli* ainsi l'association positive entre l'abondance du transporteur de fer et celles de *S. salivarius* et *V. parvula* vont dans ce sens. Ces fonctions pourraient faciliter l'implantation des espèces orales et agiraient en synergie avec une baisse de l'exposition aux sucs gastriques et une hausse de la concentration en oxygène.

En conclusion, nos résultats suggèrent que la chirurgie LRYGB a un impact délétère plus important que la LSG sur le microbiote intestinal. Ainsi, la chirurgie LRYGB pourrait conduire le microbiote dans un état dysbiotique. Cependant, notre étude se focalise sur le microbiote 6 mois après l'intervention. Les possibles conséquences à long terme sur la santé de ces altérations restent à établir.

7.3 Contribution des MSPs à la taxonomie

7.3.1 MSPs représentatives de plusieurs espèces

Le processus d'annotation taxonomique a mis en évidence des MSPs groupant des gènes provenant de plusieurs espèces le plus souvent du même genre.

Généralement, on décrète que deux souches appartiennent à la même espèce lorsque leur pourcentage d'identité nucléotidique moyen (ANI) excède 95% [167]. Cependant, l'ANI entre certaines souches excède ce seuil mais elles sont pourtant assignées à des espèces distinctes car elles colonisent des hôtes différents ou possèdent des traits phénotypiques remarquables tels que la pathogénicité (cas *Shigella spp – Escherichia coli*).

Lors de la construction du catalogue IGC, les auteurs ont choisi de regrouper les gènes ayant plus de 95% d'identité nucléotidique sur 90% de leur longueur (suppression de la redondance). De même, lors de la procédure de mapping consistant à projeter des lectures contre le catalogue, seuls les alignements globaux avec plus de 95% d'identité sont considérés. Ainsi, une MSP peut être considérée comme le répertoire de gènes de souches dont l'ANI est supérieur à 95%. Cette définition de l'espèce s'est donc imposée à nous de par les choix techniques effectués en amont.

A titre d'exemple, la msp_0832 groupe des gènes assignés aux espèces *Bifidobacterium gallinarum* et *Bifidobacterium saeculare*. Or, l'ANI entre les génomes représentatifs de ces deux espèces (*B. gallinarum* LMG 11586 et *B. saeculare* DSM 6531) calculé avec l'algorithme OrthoANI [168] est de 96,9%. Lorsqu'elle a été isolée [169], la souche DSM 6531 n'a pas été assignée à l'espèce *B. gallinarum* car le pourcentage d'hybridation avec la souche LMG 11586 était trop faible (62%, seuil à 70%). Aussi, ces deux souches ne dégradent pas les mêmes glucides ce qui a conforté le choix effectué à l'époque.

De même, la msp_0136 contient des gènes assignés à *Megamonas funiformis* et *Megamonas rupellensis* alors que l'ANI entre les génomes représentatifs de ces deux espèces (*M. funiformis* YIT 11815 et *M. rupellensis* DSM 6531) est de 98,1%. Lorsqu'elle a été isolée [170], la souche DSM 6531 n'a pas été assignée à *M. funiformis* car le pourcentage d'identité nucléotidique de son gène codant pour l'ARNr 16s avec celui de la souche YIT 11815 était de 97,6%. Ce pourcentage d'identité proche du

seuil de 97% combiné à un métabolisme des sucres différent ont été jugés suffisants pour proposer un nouveau nom d'espèce.

Un autre cas intéressant concerne la msp_0136 assignée à *Tidjanibacter massiliensis* et *Alistipes inops*. L'ANI entre les génomes représentatifs de ces espèces (*T. massiliensis Marseille-P3084* et *A. inops 627)* est de 99,97% ce qui montre qu'il s'agit clairement de la même espèce. Les génomes de *T. massiliensis Marseille-P3084* et *A. inops 627* ont été déposées sur le NCBI respectivement en Novembre 2017 et en Décembre 2015 sur le site du NCBI. On peut supposer que l'équipe qui a isolé, cultivé et séquencé la souche *Marseille-P3084* pensait avoir découvert une nouvelle espèce et lui a proposé un nom alors que celle-ci était déjà connue. Leur base de données d'espèces n'était probablement pas à jour.

7.3.2 Espèces représentées par plusieurs MSPs

Lors du processus d'annotation taxonomique, nous avons découvert des espèces microbiennes représentées par plusieurs MSPs comme *Enterobacter cloacae* (5 MSPs), *Faecalibacterium prausnitzii* (5 MSPs), *Bacteroides fragilis* (2 MSPs), *Methanobrevibacter smithii* (2 MSPs) ou *Hungatella hathewayi* (2 MSPs). Les MSPs assignées à la même espèce ne sont pas des doublons car elles ne partagent ni gènes core, ni gènes accessoires.

Nous prendrons ici l'exemple de *Methanobrevibacter smithii* mais le même raisonnement s'applique aux espèces citées ci-dessus. La comparaison des génomes des 23 souches de *M. smithii* disponibles sur GenBank met en évidence 2 phylogroupes (**Figure 66**). Le pourcentage d'identité entre deux souches représentatives de chaque phylogroupe (ATCC 35061 et TS147A) est en moyenne de 93%. Cependant, elles possèdent plus de 700 gènes qui s'alignent avec au moins 95% d'identité nucléotidique.

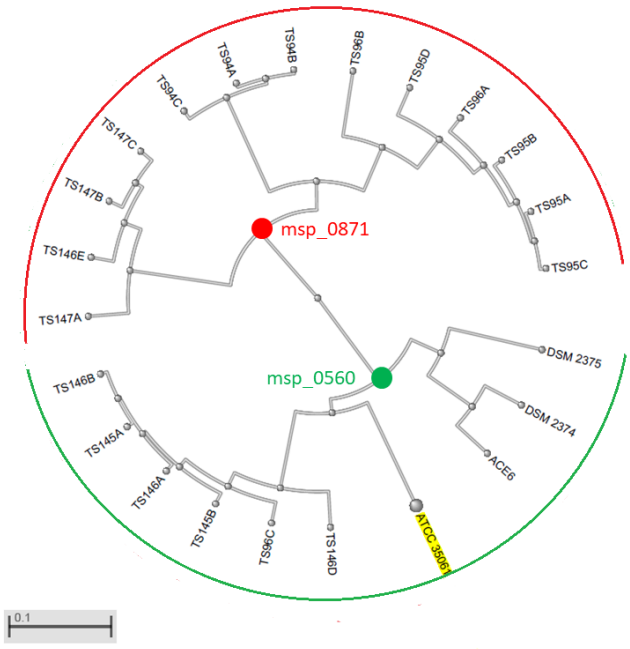


***Figure 66** : Arbre phylogénétique des souches séquencées de l'archée M. smithii téléchargées depuis le NCBI.*

L'arbre met en évidence deux phylogroupes dont l'identité nucléotidique moyenne (ANI) est d'environ 93%. Le phylogroupe vert correspond à la msp_0560 et le phylogroupe rouge à la msp_0871.

2 MSPs sont assignées à *M. smithii* : la msp_0560 (2 307 gènes) et la msp_0871 (1 730 gènes). Si l'on étudie en détail l'annotation taxonomique des gènes de chaque MSP au niveau souche, on remarque que les gènes non partagés de la msp_0560 sont assignés exclusivement aux souches du phylogroupe vert et ceux de la msp_0871 exclusivement au phylogroupe rouge. On en déduit que la msp_0560 regroupe les gènes du phylogroupe vert et la msp_0871 les gènes du phylogroupe rouge.

447 gènes sont communs aux deux MSPs et sont classifiés comme partagés dans chacune d'elles. Ces gènes sont présents dans des souches des deux phylogroupes et sont particulièrement conservés (identité nucléotidique > 95%) contrairement aux gènes spécifiques à chaque MSP qui ont plus divergé. Finalement, les gènes conservés sont représentés par un seul gène dans le catalogue alors que les gènes qui ont plus divergé sont représentés par des gènes distincts. L'abondance des gènes partagés est une combinaison des abondances des gènes core des deux MSPs (**Figure 67** et **Figure 68**). Fait intéressant, plusieurs gènes partagés assurent des fonctions indispensables à la survie du microorganisme (**Tableau 15**). Ces derniers ont probablement été soumis à des contraintes évolutives plus fortes et ont par conséquent accumulé moins de mutations.

Module fonctionnel	Nombre de gènes
méthanogénèse à partir de CO_2	17
méthanogénèse - Méthyl-coenzyme M réductase	3
métabolisme du dihydrogène	2
cycle de Krebs	2

***Tableau 15** : Exemples de fonctions assurées par les gènes partagés par la msp_0560 et la msp_0871.*

Lorsqu'elles ont été isolées et séquencées [171], les souches du phylogroupe rouge ont été assignées à *M. smithii* car leurs gènes codant pour l'ARNr 16s sont très similaires à celui de la souche représentative du phylogroupe vert (> 99% d'identité nucléotidique). De plus, les souches des deux phylogroupes ont des traits fonctionnels semblables (méthanogénèse, consommation de dihydrogène). Néanmoins, il semble pertinent que *M. smithii* soit représentée par deux MSPs car si on appliquait le critère basé sur le pourcentage d'identité nucléotidique chaque phylogroupe devrait être assigné à une espèce différente.

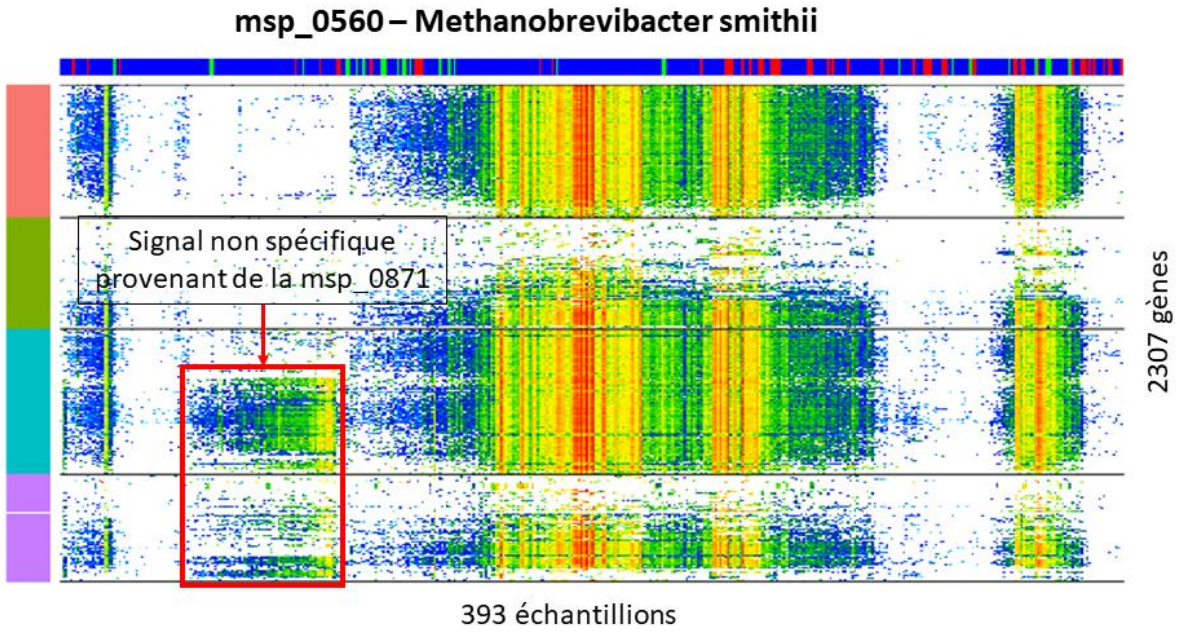


***Figure 67** : Heatmap représentative de l'abondance des 2307 gènes de la msp_0560 (M. smithii, phylogroupe vert) dans 393 échantillons du catalogue IGC.*

*La majorité des gènes cyans et violets sont partagés avec la msp_0871 correspondant au phylogroupe rouge (voir **Figure 68**). Le signal non spécifique provenant de cette MSP est mis en évidence par le rectangle rouge. Pour une description du contenu de la heatmap, se référer à la **Figure 58** (page 96)*

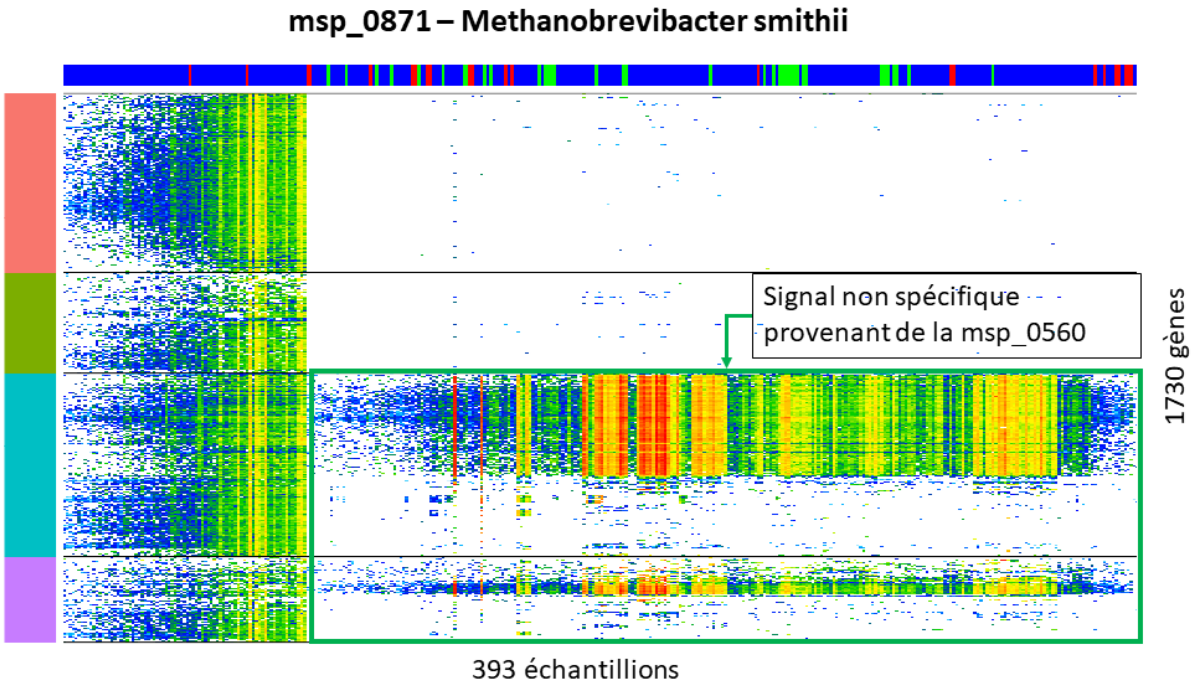


***Figure 68** : Heatmap représentative de l'abondance des 1730 gènes de la msp_0871 (Methanobrevibacter smithii, phylogroupe rouge) dans 393 échantillons du catalogue IGC.*

*La majorité des gènes cyans et violets sont partagés avec la msp_0560 correspondant au phylogroupe vert (voir **Figure 68**). Le signal non spécifique provenant de cette MSP est mis en évidence par le rectangle vert. Pour une description du contenu de la heatmap, se référer à la **Figure 58** (page 96)*

7.3.3 Réannotation de génomes et détection de contaminations

Lorsque les gènes core d'une MSP annotée au niveau espèce sont très proches (pourcentage d'identité ≥ 97.5% et couverture moyenne ≥ 90%) de ceux présents dans un génome assigné à une espèce différente, nous avons proposé une annotation alternative pour ce génome en l'assignant à la même espèce que la MSP.

Cette stratégie a permis de préciser l'annotation de certains génomes (**Tableau 16**) ou même de la corriger (**Tableau 17**).

GenBank accession	Ancienne annotation	Annotation proposée	MSP correspondante
AWUR	Acidaminococcus sp. BV3L6	Acidaminococcus intestini	msp_0216
CAYI	Bacteroides sp. CAG:189	Bacteroides salyersiae	msp_0035
BAIA	Clostridiales bacterium VE202-13	Anaerotruncus colihominis	msp_0503

***Tableau 16** : Exemple de génomes pour lesquels une meilleure annotation est proposée grâce au MSPs.*

GenBank accession	Ancienne annotation	Annotation proposée	MSP correspondante
FKZO	Bacteroides thetaiotaomicron 2789STDY5834846	Bacteroides faecis	msp_0018
CP000964	Klebsiella pneumoniae 342	Klebsiella variicola	msp_0234
CSXB	Mycobacterium abscessus PAP053	Enterococcus faecalis	msp_0479

***Tableau 17** : Exemple de génomes dont l'annotation a été corrigée grâce au MSPs.*

Lorsque que les gènes de plusieurs MSPs représentant des espèces distantes sont alignés sur un même génome, cela indique que ce génome est probablement contaminé. Par exemple, 1 299 gènes core de la msp_0957 (*Intestinimonas massiliensis*) et 1 480 gènes core de la msp_0208 (*Flavonifractor plautii)* sont alignés sur le génome *Flavonifractor plautii* 2789STDY5834906. On peut donc supposer que *F. plautii* 2789STDY583490 est un mélange d'une souche de *F. plautii* et d'une souche de *I. massiliensis*. Le mélange est confirmé en alignant *F. plautii* 2789STDY583490 contre les génomes représentatifs RefSeq *F. plautii* (YL31) et *I. massiliensis* (GD2) (**Tableau 18**).

Genome A	Genome B	Couverture du génome A	Couverture du génome B	ANI
Flavonifractor plautii 2789STDY5834906	Flavonifractor plautii YL31	**31,8%**	61,3%	97,8%
Flavonifractor plautii 2789STDY5834906	Intestinimonas massiliensis GD2	**26,2%**	62,2%	98,2%
Flavonifractor plautii YL31	Intestinimonas massiliensis GD2	**22,5%**	**27,7%**	**76,7%**

***Tableau 18** : Alignement deux à deux des génomes F. plautii 2789STDY583490, F. plautii YL31 et I. massiliensis GD2 en utilisant OrthoANI.*

Les génomes représentatifs de F. plautii et I. massiliensis ont un pourcentage d'identité moyen faible (ANI=76,7%) ce qui montre qu'il s'agit d'espèces différentes.

F. plautii 2789STDY583490 a un pourcentage d'identité moyen élevé avec F. plautii YL31 (ANI = 97,8%= et I. massiliensis (ANI = 98,2%) mais les alignements ne couvrent que 31,8% et 26,2% du génome. Ceci indique que F. plautii 2789STDY583490 est un mélange d'une souche de F. plautii et d'une souche de I. massiliens

8. Discussion

8.1 Limites de la méthode

8.1.1 MSPs chimériques et manquantes

8.1.1.1 *MSPs fusionnant 2 espèces proches*

A la fin du processus d'annotation taxonomique, certaines MSPs sont assignées à deux espèces distinctes généralement du même genre. Contrairement aux cas exposés en 7.3.1, l'ANI entre les souches représentatives de ces espèces est inférieur au seuil de 95%.

Par exemple, la msp_0696 groupe des gènes attribués à *Streptococcus infantarius* et *Streptococcus lutetiensis*. Or, l'ANI entre les génomes représentatifs de ces deux espèces (*S. infantarius* CJ18 et *S. lutetiensis* 033) est égale à 93,5%. Cette valeur nettement inférieure au seuil de 95% justifie que ces souches aient été classées dans deux espèces distinctes mais assignées au même genre bactérien.

Lorsqu'on analyse en détail l'annotation taxonomique de cette MSP, on remarque que certains gènes sont attribués aux deux espèces mais que d'autres sont attribués à une seule. Après suppression de la redondance, les gènes orthologues très conservés (> 95% d'identité nucléotidique) entre *S. infantarius* et *S. lutetiensis* sont représentés dans le catalogue par un seul gène partagé par les deux espèces. Les gènes orthologues plus distants (<95% d'identité) sont quant à eux représentés par deux gènes distincts.

Pour rappel, lorsque plusieurs seeds sont associées, celui contenant le plus de gènes est sélectionné comme core par MSPminer (voir 4.1.5). Or dans la msp_0696, les gènes partagés sont plus nombreux que ceux spécifiques à chaque espèce (**Figure 69**). Ainsi, une seule MSP ayant comme core les gènes partagés a été générée. Or, il aurait fallu créer deux MSPs distinctes ayant pour core respectifs les gènes spécifiques à *S. infantarius* et *S. lutetiensis.*

Il est difficile de recenser les MSPs fusionnant deux espèces proches car la majorité d'entre elles ne sont pas annotées. Néanmoins, celles composées de deux groupes de gènes accessoires observés dans des ensembles distincts d'échantillons et dont la somme des signaux est égale au signal des gènes core (**Figure 69**) sont des cas douteux. Nous avons recensé 20 MSPs (1,2%) potentiellement chimériques ce qui montre que le problème est marginal. Cependant, on pourrait envisager d'améliorer le critère de sélection des gène core en ne se basant plus uniquement sur un critère de taille.

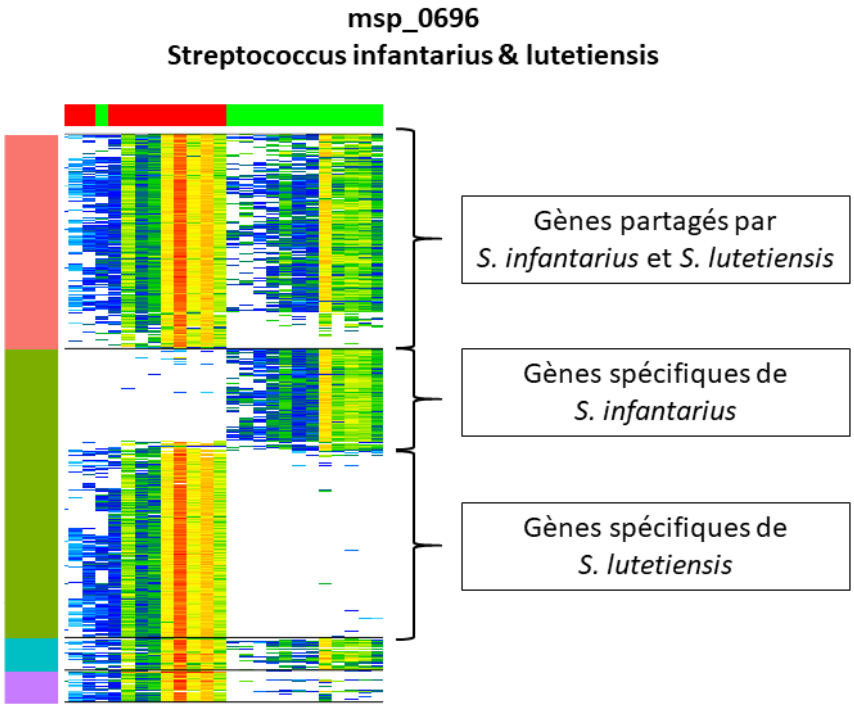


***Figure 69** : Heatmap représentative de l'abondance relative des 2028 gènes de la msp_0696 dans 24 échantillons du catalogue IGC.*

*Les gènes core de la MSP sont partagés par S. infantarius et S. lutetiensis. Les gènes accessoires présents dans les échantillons groupés à droite sont spécifiques à S. infantarius. Ceux détectés dans les échantillons groupés à gauche sont spécifiques à S. lutetiensis. Pour une description du contenu de la heatmap, se référer à la **Figure 58** (page 96).*

8.1.1.2 *MSPs contaminées*

L'annotation taxonomique met aussi en évidence quelques MSPs groupant des gènes provenant de deux espèces distantes d'un point de vue phylogénétique.

Par exemple, la msp_0558 contient des gènes de *Streptococcus anginosus* et de *Lactobacillus salivarius*. Or, ces deux espèces sont phylogénétiquement distantes car l'ANI entre leurs génomes représentatifs (*S. anginosus* C238 et *L. salivarius* UCC118) atteint seulement 67,5%.

Si on étudie en détail l'annotation taxonomique des gènes de cette MSP, on remarque que l'ensemble de ses gènes core sont assignés *S. anginosus* tandis qu'un seul groupe de gènes accessoires est assigné à *L. salivarius* (**Figure 70**).

Dans les différents échantillons où ils sont détectés, les gènes assignés à *S. anginosus* ont des comptages bruts directement proportionnels aux comptages des gènes de *L. salivarius* (**Figure 71.A**). Cependant, les comptages normalisés par la profondeur de séquençage et la longueur des gènes sont toujours proportionnels mais ne sont pas du même ordre de grandeur (**Figure 70** et **Figure 71.B**).

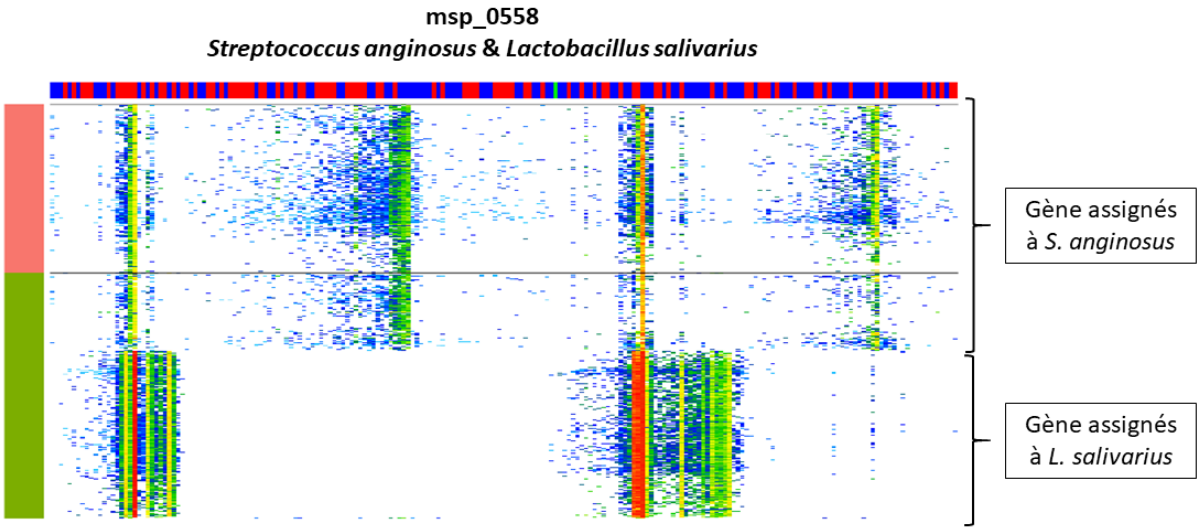


Figure 70 *: Heatmap représentative de l'abondance relative des 2312 gènes de la msp_0558 dans 209 échantillons du catalogue IGC.*

L'abondance relative des gènes assignés à S. anginosus (en rouge) est clairement différente de celles des gènes assignés à L. salivarius (vert kaki) Pour une description du contenu de la heatmap, se référer à la Figure 58 (page 96)

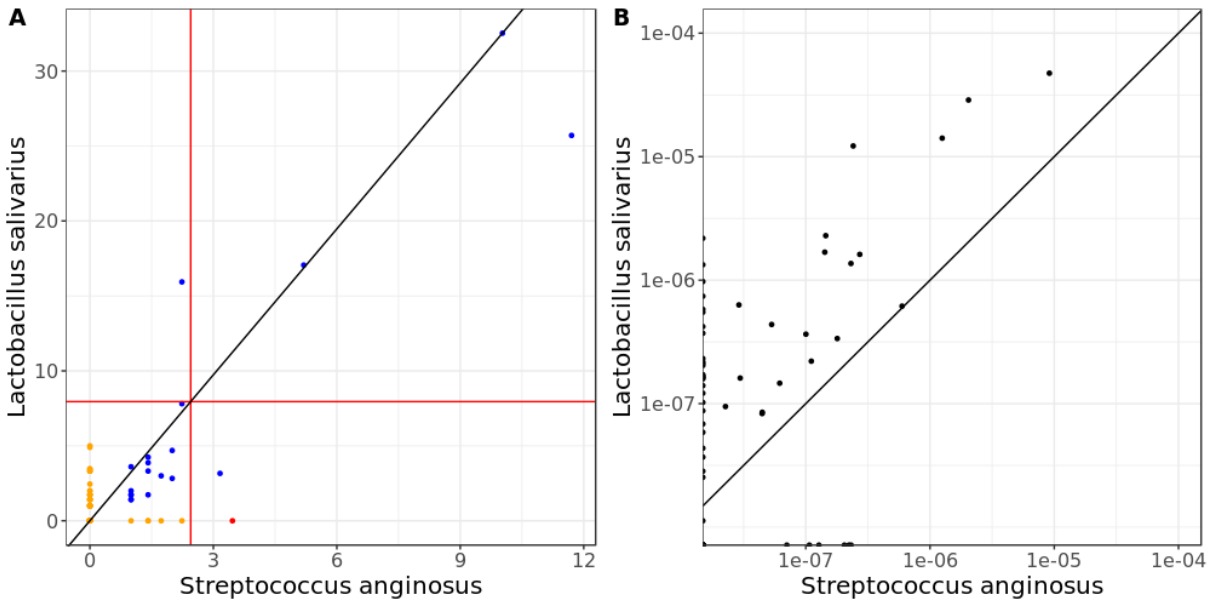


Figure 71 *: Comparaison dans la msp_0558 de l'abondance médiane des gènes associés à S. anginosus (axe des abscisses) et des gènes à S. salivarius (axe des ordonnées)*

A. Comptages bruts en échelle racine carrée. La mesure de la proportionnalité entre les comptages S. anginosus et S. salivarius est de 0.9

B. Comptages normalisés par la longueur du gène et la profondeur de séquençage en échelle log_{10}*. La droite noire a pour équation* $y = x$*. Si les comptages des gènes provenant de chaque espèce étaient du même ordre de grandeur, ils s'aligneraient sur cette droite ce qui n'est pas le cas.*

En conclusion, la msp_0558 correspond à l'espèce *S. anginosus* mais est contaminée par des gènes de *L. salivarius*. Ce cas illustre les limites des mesures de la proportionnalité (p_{nr} et p_r), qui dans certains cas, groupent au sein d'une même MSP des gènes qui n'auraient pas dû l'être. Ici, l'utilisation de comptages normalisés par la longueur des gènes permet de détecter ces faux positifs.

Par la suite, nous avons étiqueté comme contaminées les MSPs possédant un ou plusieurs groupes de gènes accessoires dont les comptages normalisés ne sont pas du même ordre de grandeur que leurs gènes core respectifs. Au final, nous n'avons recensé que 10 MSPs (0,6%) contaminées. Même si les MSPs contaminées sont peu nombreuses, elles peuvent mener à des conclusions biologiques erronées. Dans un premier temps, un nettoyage des MSPs après l'exécution de MSPminer permet de résoudre ce problème. Dans le futur, nous envisageons d'améliorer la mesure de la proportionnalité pour supprimer ces faux positifs.

8.1.1.3 *MSPs manquantes*

La comparaison du clustering généré par MSPminer avec celui produit par Canopy a mis en évidence une vingtaine de CAGs qui n'ont pas d'équivalent chez les MSPs. Parmi ces CAGs, on retrouve des espèces répertoriées comme prévalentes et abondantes dans le microbiote intestinal humain telles que *Bacteroides xylanisolvens* ou *Bacteroides ovatus.* Par conséquent, il est indispensable que l'on ait des MSPs représentatives de ces espèces sans quoi les analyses basées sur les MSPs souffriraient d'un biais trop important.

Pour comprendre pourquoi MSPminer était mis en défaut, nous avons exécuté sur les CAGs correspondant aux espèces manquantes l'algorithme regroupant les gènes co-abondants et co-occurrents (4.1.1) dans des seeds. Dans chaque CAG, nous avons conservé la plus grande seed découverte et nous l'avons étiquetée comme core génome de l'espèce.

Les seed core des espèces manquées sont remarquablement petites. Leur taille est nettement inférieure au seuil de 200 gènes à partir duquel une seed est utilisée pour reconstituer une MSP (**Tableau 19**, 5ème colonne). Par conséquent, ces seeds de petite taille ont été filtrées par MSPminer et les MSPs qui en découlent manquent. Pour le moment, nous n'avons pas expliquer pourquoi ces espèces ont une core si petit.

Par la suite, nous avons forcé la reconstitution des MSPs correspondant aux espèces manquantes en supprimant le critère le critère de filtrage basé la taille des seeds. Les MSPs obtenues regroupaient parfois beaucoup plus de gènes que les CAGs correspondants (**Tableau 19**, 3ème et 5ème colonne). Sur les 21 MSPs reconstituées, le gain médian atteint 113%.

Espèce	Nombre de gènes du CAG correspondant	MSP reconstituée	Nombre de gènes dans la MSP	Nombre de gènes core dans la MSP
Bacteroides ovatus	2 066	msp_a01	9 897	131
Bacteroides xylanisolvens	2 321	msp_a02	7 078	78
Butyricimonas virosa	586	msp_a04	3 862	153
Parasutterella excrementihominis	1 297	msp_a05	3 782	157
Bifidobacterium catenulatum	357	msp_a10	1 757	122

***Tableau 19** : Exemple de MSPs manquantes reconstituées a posteriori*

8.1.2 Limites de la reconstitution de pan-génomes par regroupement des gènes co-abondants

8.1.2.1 *Variabilité du nombre de copies d'un gène*

Dans un génome, un gène est parfois présent en plusieurs copies. De plus, le nombre de copies d'un gène peut varier au sein de souches d'une même espèce. Par exemple, les souches de *Bifidobacterium longuum* possèdent de 1 à 4 copies du gène codant pour l'ARNr 16s [69].

Lorsqu'un gène à nombre de copies variable est comparé au signal du core génome de l'espèce à laquelle il appartient, plusieurs coefficients de proportionnalité coexistent simultanément. Chacun de ces coefficients correspond à un nombre de copies du gène différent.

MSPminer peut détecter de tels cas lorsqu'une majorité d'échantillons (70% par défaut) possède un même nombre de copies du gène. Les gènes avec un nombre de copies différent seront classifiés comme valeurs aberrantes et seul le coefficient de proportionnalité majoritaire sera pris en compte (**Figure 72.B**).

Lorsqu'aucun nombre de copies n'est dominant, MSPminer échoue à détecter la relation. En effet, il existe plusieurs coefficients de proportionnalité alors que notre modèle suppose qu'il n'en existe qu'un seul (**Figure 72.A**).

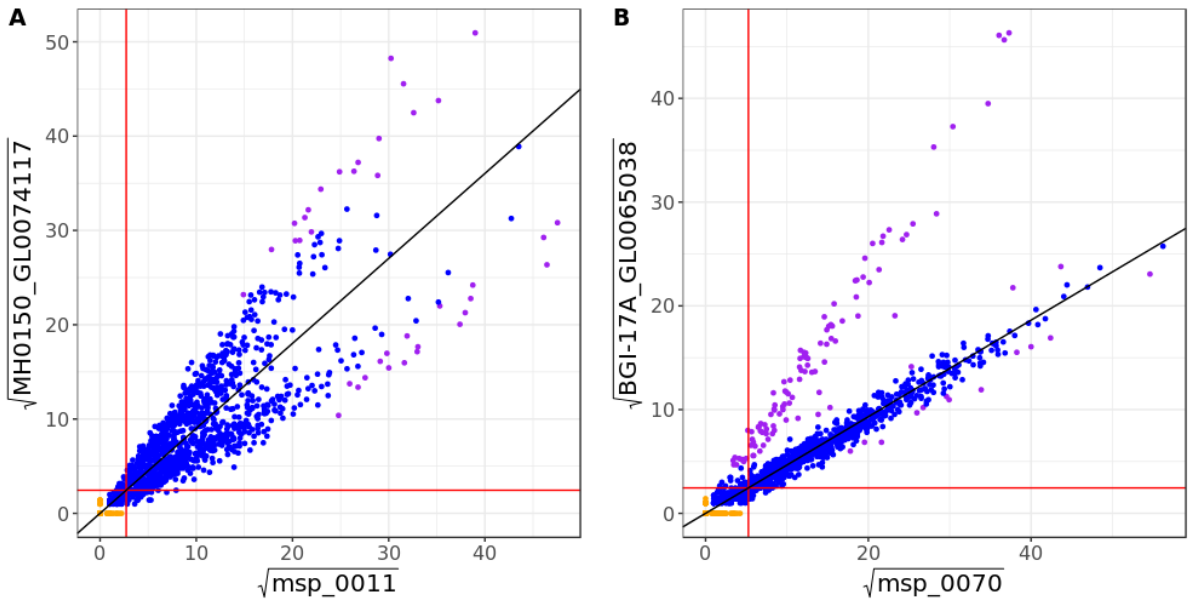


***Figure 72** : Exemple de gènes à nombre de copies variable*

A. Comparaison de l'abondance médiane du core génome de la msp_0011 représentative de Parabacteroides *distasonis (axe des abscisses) à un de ses gènes core (axe des ordonnées) ayant un nombre de copies variables. Ici, le coefficient de proportionnalité estimé est incorrect car il n'existe pas de nombre de copies dominant parmi les échantillons.*

B. Comparaison de l'abondance médiane du core génome de la msp_0070 représentative d'Eubacterium rectale (axe des abscisses) à un de ses gènes core (axe des ordonnées) ayant un nombre de copies variables. Ici, le coefficient de proportionnalité estimé est correct car la majorité des échantillons (points bleus) possède le même nombre de copies du gène. Les échantillons possédant un nombre de copies différents sont classifiés comme valeurs aberrantes (points violets).

Une étude analysant 109 échantillons fèces par métagénomique shotgun montre que la variation du nombre de copies d'un gène est un phénomène courant dans le microbiote intestinal humain ;

certaines souches ayant plus de 20% de leurs gènes impactés [21]. Ainsi, le développement de nouvelles méthodes statistiques est nécessaire pour assigner systématiquement des gènes à nombre de copies variables à leurs MSPs respectives. Cependant, il est probable qu'une majorité de gènes à nombre de copies variable soient correctement assignés à leur MSP car le nombre de copies de ces gènes est globalement constant et ne varie que dans une minorité de souches.

8.1.2.2 *Gènes partagés par plusieurs espèces*

Nous appelons gènes partagés des gènes homologues avec une forte identité nucléotidique (≥ 95%) qui ne sont pas spécifiques à une espèce. Il peut s'agit de gènes orthologues conservés ou de gènes sujets à un transfert horizontal inter-espèces [172].

Lorsqu'un gène est partagé, son abondance est une combinaison linéaire des abondances des MSPs qui le portent. Ainsi, une relation de proportionnalité directe entre un gène partagé et celle du core génome des MSPs est détectée lorsque les MSPs sont présentes dans des sous-ensembles d'échantillons quasiment distincts. Par exemple, les gènes partagés par les deux sous-espèces de *Methanobrevibacter smithii* (voir 7.3.1) sont groupés dans leurs MSPs respectives car les deux sous-espèces sont très rarement détectées dans un même échantillon (**Figure 73**). En effet, la mesure robuste de la proportionnalité (ρ_r) écarte les échantillons où les autres MSPs sont présentes ce qui permet de détecter des associations partielles.

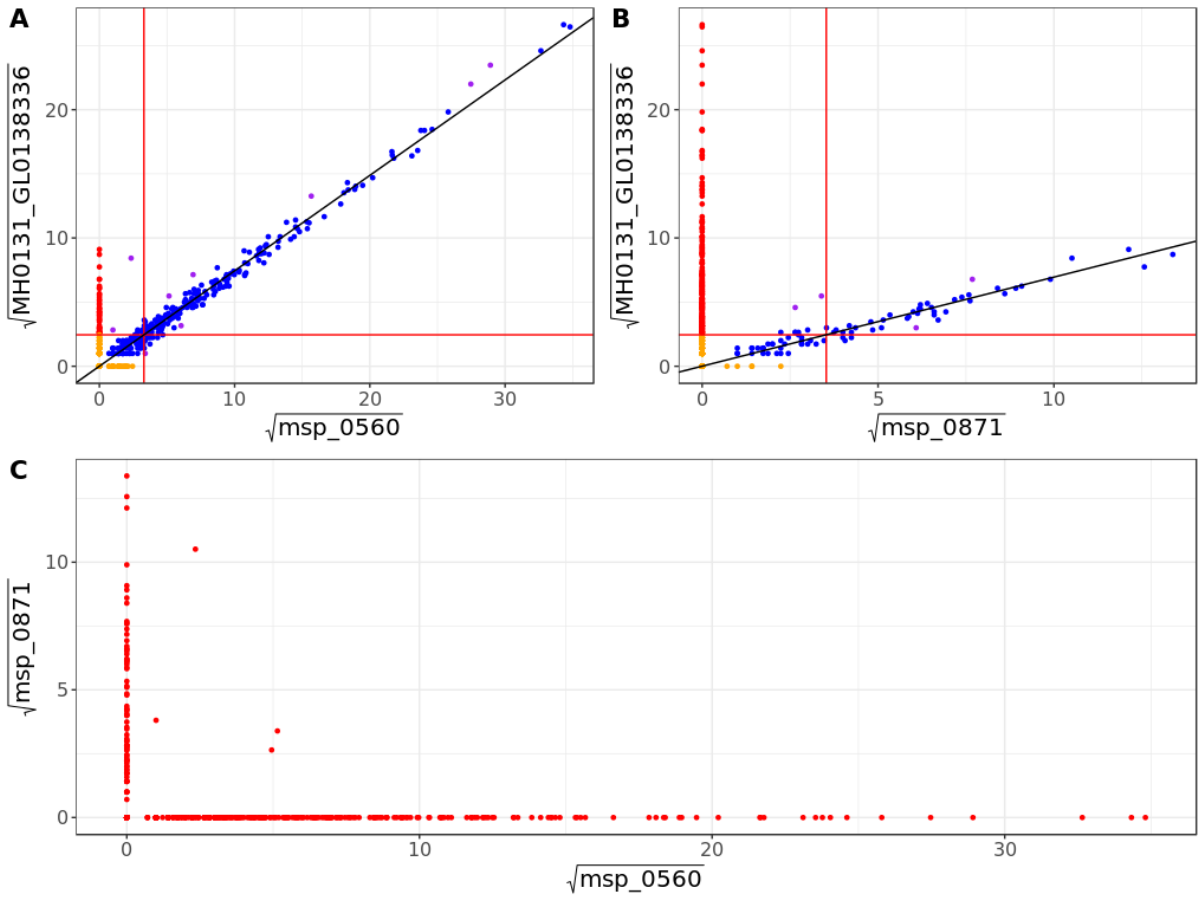


***Figure 73** : Exemple de gène partagé par 2 MSPs*

Le gène MH0131_GL0138336 est classifié en tant que gène core partagé dans les 2 MSPs assigné à Methanobrevibacter smithii *(msp_0560 et msp_0871).*

A. Comparaison de l'abondance médiane du core génome de la msp_0560 et du gène MH0131_GL0138336. Un lien de proportionnalité direct est observé dans tous les échantillons bleus. On remarque que le gène est également détecté dans des échantillons où la msp_0560 est absente (échantillons rouges). Ce signal inattendu provient de la msp_0871 et n'est pas pris en compte lors de la mesure de proportionnalité.

B. Comparaison de l'abondance médiane du core génome de la msp_0871 et du gène MH0131_GL0138336. Un lien de proportionnalité direct est observé dans tous les échantillons bleus. On remarque que le gène est également détecté dans des échantillons où la msp_0871 est absente (échantillons rouges). Ce signal inattendu provient de la msp_0560 et n'est pas pris en compte lors de la mesure de proportionnalité.

C. Comparaison de l'abondance médiane du core génome de la msp_0560 et de la msp_0871. Bien qu'elles partagent le gène MH0131_GL0138336, les MSPs ont des signaux différents et sont détectés dans des ensembles d'échantillons distincts.

Lorsque les MSPs partageant un gène sont détectées simultanément dans un grand nombre d'échantillons, le lien de proportionnalité du gène avec leurs core respectifs est trop fortement dégradé pour être détecté par la mesure robuste. Finalement, ces gènes ne seront pas regroupés dans les MSPs.

Dans le futur, nous envisageons d'implémenter dans MSPminer des algorithmes de déconvolution [173] qui pourraient détecter des relations plus complexes entre un gène partagé et les différentes MSPs qui contribuent à son signal. A partir des vecteurs d'abondance du core génome des MSPs, ces algorithmes résolvent un système d'équations linéaires qui estime pour chaque gène les MSPs auxquelles il doit être attaché.

8.1.2.3 *Mélange de plusieurs souches dans un échantillon*

En utilisant des données simulées, nous avons montré que la coexistence de plusieurs de souches de la même espèce dans des échantillons métagénomiques a un impact négatif sur la complétude des MSPs (4.3.2). Lorsque le mélange de souches concerne plus de 50% des échantillons où une espèce est détectée, une proportion importante de ses gènes accessoires de prévalence faible ou intermédiaire ne sont pas assignés à sa MSP représentative.

Or, une étude caractérisant 125 espèces microbiennes sur plus de 1500 métagénomes du microbiote intestinal humain indique qu'un mélange de souches est détecté dans 64% des paires échantillon-espèce [98]. Ainsi, la performance de MSPminer pourrait être fortement dégradée par ce phénomène.

Dans la simulation présentée en 4.3.2, nous avons considéré que la présence d'un gène accessoire dans une souche était indépendante de sa présence dans l'autre. Or, la faible divergence nucléotidique fréquemment observées entre les souches présentes simultanément dans un échantillon de fèces suggère qu'elles pourraient avoir un contenu en gènes similaire. Ainsi, la relation de proportionnalité entre un gène accessoire et le core de la MSP pourrait être beaucoup moins dégradée que ce que montre notre simulation.

8.1.2.4 *Impact de la croissance des populations microbiennes sur la couverture de séquençage*

A quelques exceptions près, les bactéries possèdent un seul chromosome qui lors de la division cellulaire se réplique dans les deux sens à partir d'une seule origine jusqu'à une seule terminaison. Lors de la réplication, les régions formant l'œil auront deux copies tandis que celles en dehors en auront une seule.

Dans les fèces, certaines espèces microbiennes sont en phase de croissance. En alignant les lectures produites lors d'un séquençage métagénomique sur le chromosome d'une de ces espèces, on constate que la couverture est maximale au niveau de l'origine de réplication puis décroit progressivement jusqu'à la terminaison [174]. Remarquablement, la couverture ne change pas brutalement suivant une fonction étagée car les cellules ne sont pas au même stade de réplication. Le rapport entre la couverture à l'origine et à la terminaison de la réplication est proportionnel à la vitesse moyenne de réplication. Ce rapport peut être supérieur à deux car les cellules avec un fort taux de croissance ont plusieurs yeux de réplications.

Finalement, la croissance des populations microbiennes pourrait perturber le lien de proportionnalité entre les vecteurs d'abondance de gènes de la même espèce car la vitesse de réplication n'est pas la même d'un échantillon à l'autre.

8.2 Facteurs impactant la qualité des MSPs

8.2.1 Nombre et diversité des échantillons métagénomique

Le nombre croissant d'échantillons métagénomiques disponibles sur les bases de données publiques (**Figure 74**) ainsi que la grande diversité phénotypique des individus dont ils proviennent (âge, origine géographique, état de santé, régime alimentaire etc.) améliorera la complétion des MSPs ainsi que leur qualité.

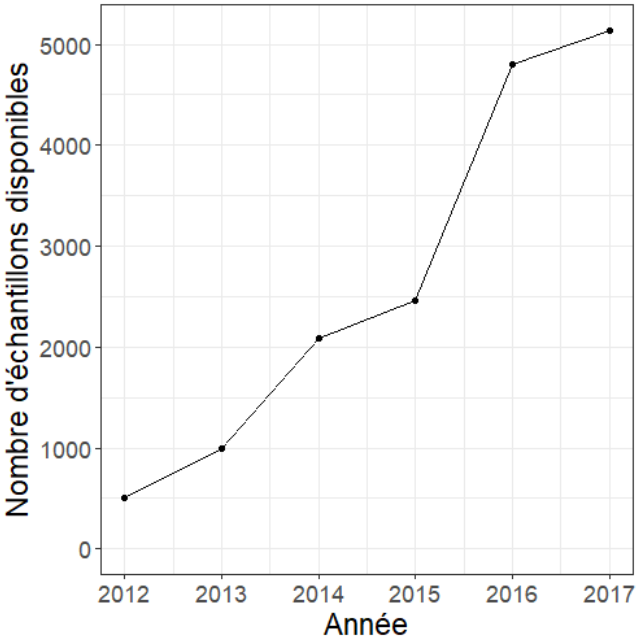


Figure 74 : Nombre total d'échantillons métagénomiques disponibles sur la base de données NCBI SRA par année.

Ces données sont fournies par l'outil curatedMetagenomicData [175]. Le nombre d'échantillon disponibles en 2017 est sous-estimé car tous les projets n'ont pas encore été indexés.

A mesure que le nombre d'échantillons augmente, MSPminer identifiera des espèces rares et assignera des gènes accessoires rares (cloud genes) à leurs MSPs respectives. De plus, les gènes accessoires hautement prévalent seront reclassifiés de core à accessoires comme observé lors du séquençage d'un nombre croissant de souches d'une espèce [28].

8.2.2 Séquençage

Les technologies de séquençage ainsi que des paramètres tels que le nombre de lectures générées, leur longueur et leur type (single-end, paired-end) impactent la qualité des MSPs et leur quantification.

8.2.2.1 *Profondeur de séquençage*

L'augmentation de la profondeur de séquençage consistant à générer plus de lectures par échantillon permet d'obtenir des assemblages métagénomiques de meilleure qualité composés de contigs plus longs et moins fragmentés. Par conséquent, les MSPs grouperont plus de gènes dont une proportion importante de gènes complets. Les espèces sous-dominantes seront détectées et partiellement assemblées ce qui permettra éventuellement de générer des MSPs leur correspondant. Finalement, l'estimation de l'abondance des MSPs sera plus précise car les comptages des gènes seront plus forts.

8.2.2.2 *Longueur des lectures*

Actuellement des lectures courtes d'environ 100 paires de bases sont utilisées pour l'assemblage métagénomique et le comptage des gènes. La génération de lectures plus longues permettrait d'améliorer la qualité des gènes regroupés dans les MSPs. En effet, les assemblages seraient moins fragmentés car on lèverait des ambiguïtés liées aux régions répétées dans les génomes. Ainsi, on réduirait la proportion de gènes incomplets et on estimerait plus précisément l'abondance des gènes car le nombre d'alignements ambigus où une lecture est assignée à plusieurs gènes diminuerait. De plus, on augmenterait la proportion de comptages uniques et on diminuerait le nombre de gènes détectés à tort (faux positifs).

Dans le même ordre d'idée, le séquençage pairé (paired-end) consistant à séquencer les deux extrémités d'un fragment d'ADN génère virtuellement des lectures plus longues. La production de telles lectures améliore la qualité des assemblages et la spécificité des alignements.

8.2.3 Construction du catalogue

Chaque étape de la création d'un catalogue de gènes nécessite une expertise bioinformatique pour choisir les stratégies, outils et paramètres les plus appropriés.

8.2.3.1 *Assemblage*

L'assemblage métagénomique demeure une tâche complexe malgré le développement d'algorithmes et de logiciels dédiés [106]. En particulier, la présence dans un même échantillon d'espèces apparentées (même genre) peut amener à générer des contigs chimériques contenant des gènes sans réalité biologique.

8.2.3.2 *Prédiction des gènes*

Pour prédire des gènes sur les contigs, on utilise généralement des logiciels entraînés sur des génomes procaryotes comme Prodigal [176] ou MetaGeneMark [110]. L'utilisation de ces outils sur des génomes eucaryotes génère des résultats incohérents car les gènes y ont une structure différente (présence d'introns et d'exons). Par conséquent, les MSPs d'espèces eucaryotes contiennent actuellement un nombre de gènes étonnamment grand dont la plupart sont fragmentés. Par exemple, la msp_0002 représentative de l'espèce *Blastocystis sp. subtype 1* est composée de 13 372 gènes dont 96,5% sont classifiés comme core. Or, le génome représentatif de cette espèce disponible sur Genbank (*Blastocystis sp. ATCC 50177/Nand II*) ne possède que 6 544 gènes. 96% (12 857/13 372) des gènes de la msp_0002 sont alignés sur toute leur longueur sur le génome (pourcentage d'identité ≥ 95%) mais seulement 2,9% (187/6 544) des gènes du génome sont complets dans la MSP. Pour éviter ce problème, on pourrait annoter les contigs obtenus après assemblage métagénomique et utiliser un prédicteur de gènes pour eucaryotes ou procaryotes suivant le domaine auquel ils sont assignés.

En outre, ces outils peuvent parfois prédire des gènes chimériques fusionnant plusieurs gènes constitutifs. Les comptages des gènes chimériques seront incohérents s'ils agglomèrent des gènes qui ne sont pas systématiquement cooccurrents.

8.2.3.3 *Suppression de la redondance*

Pour rappel, la suppression de la redondance consiste à regrouper les gènes ayant un fort degré d'homologie. Les gènes regroupés sont finalement représentés par un seul gène dans le catalogue non redondant. Pour réaliser cette tâche complexe, on s'appuie sur des outils comme CD-HIT [111] qui pour produire des résultats en un temps raisonnable, choisissent systématiquement le gène le plus long comme représentant d'un cluster.

Ainsi, les gènes chimériques fusionnant plusieurs gènes ont une forte de chance de devenir représentant d'un cluster dans le catalogue final. Pour éviter ce problème, on pourrait étudier la longueur des gènes au sein d'un cluster et éliminer ceux ayant une longueur bien supérieure aux autres.

Actuellement, le pourcentage d'identité minimal pour regrouper des gènes dans un cluster est fixé à 95% d'identité nucléotidique. En effet, il est généralement admis qu'un pourcentage d'identité nucléotidique moyen (ANI) de 95% est une limite basse pour assigner deux souches à la même espèce [177]. Cependant, utiliser un seuil unique pour traiter tous les gènes n'est probablement pas la meilleure stratégie. Deux souches d'espèces différentes peuvent avoir un ANI inférieur à 95% mais posséder de nombreux gènes orthologues conservés ayant un pourcentage d'identité nucléotidique supérieur à ce seuil. Par exemple, les souches représentatives de *Bacteroides ovatus* (ATCC 8483) et *Bacteroides xylanisolvens* (CL03T12C04) ont un ANI de 92,5% mais possèdent environ 1500 gènes avec pourcentage d'identité nucléotidique supérieur à 95%. Après suppression de la redondance, ces gènes conservés seront regroupés dans un même cluster et leur signal sera une combinaison linéaire de l'abondance de *B. ovatus* et *B. xylanisolvens*. Dans la cohorte du catalogue IGC, *B. ovatus* et *B. xylanisolvens* sont observées simultanément dans 58% des échantillons où au moins l'une de ces espèces est présente. Par conséquent, le signal des gènes conservés ne sera pas co-abondant avec les core génomes de ces espèces. Au final, ces gènes ne seront regroupés ni dans la MSP représentative de *B. ovatus*, ni dans celle représentative de *B. xylanisolvens* même en incluant les catégories « core partagé » et « accessoire partagé ». Pour règler ce problème, on pourrait utiliser un pourcentage d'identité adaptatif lors de la suppression de la redondance qui prendrait en compte le degré de conservation des orthologues au sein d'espèces proches.

8.2.4 Alignement et quantification des gènes

Lors de la construction d'un catalogue, on conserve en général les gènes de plus de 100 paires de bases ce qui correspond approximativement à la longueur des lectures produites par les séquenceurs modernes. A cause d'effets de bord, les lectures provenant de gènes présents dans le catalogue pourraient ne pas être alignées. En effet, il est possible qu'une lecture s'aligne à l'extrémité d'un gène ou qu'elle en chevauche deux (**Figure 75**). Ce phénomène est d'autant plus probable que les gènes sont courts. Dans de tels cas, l'alignement bout à bout (end-to-end) échoue ce qui augmente la dispersion des comptages. La mise en œuvre de nouvelles stratégies telles que l'alignement local ou le découpage des lectures en plusieurs fragments de même longueur permettraient de résoudre ce problème.

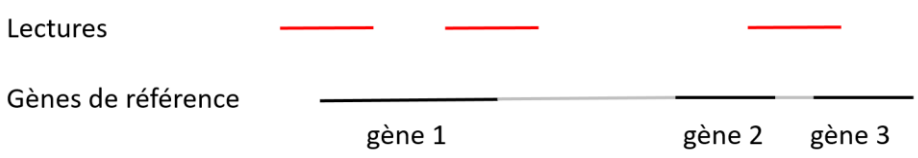


***Figure 75** : Exemples de cas où l'alignement bout à bout est mis en échec.*

On considère 3 gènes (traits noirs) séparés par des régions intergéniques (traits gris) ainsi que 3 lectures (traits rouges). L'alignement bout à bout des deux lectures de gauche échoue pour car elles s'alignent aux extrémités du gène1. L'alignement de la lecture de droite échoue car elle chevauche les gènes 2 et 3.

Après alignement des lectures sur un catalogue, plusieurs stratégies de quantification des gènes sont envisageables. Le comptage unique considère uniquement les lectures s'alignant sur un seul gène du catalogue. Ainsi, on sous-estime l'abondance des gènes constitués de régions conservées partagées avec plusieurs autres gènes. Le comptage partagé incrémente de $\frac{1}{n}$ les n gènes sur lesquels s'aligne une lecture. Cette stratégie génère des faux comptages qui compromettent la détection de la proportionnalité malgré l'utilisation d'une mesure robuste. Enfin, le comptage partagé « intelligent » distribue une lecture alignée sur plusieurs gènes proportionnellement aux comptages uniques de ces derniers. On maximise la sensibilité en considérant toutes les lectures alignées sur le catalogue tout en minimisant le nombre de faux positifs. Ainsi, l'estimation de l'abondance des gènes est plus précise ce qui permet d'obtenir des MSPs de meilleure qualité.

9. Perspectives

9.1 Améliorations et optimisations du logiciel

9.1.1 Alternatives à la médiane pour le calcul du représentant des MSPs

A ce jour, une MSP est représentée par un gène dont l'abondance correspond à la médiane des 30 meilleurs gènes core. En pratique, ces gènes sont les plus longs : leurs comptages sont forts et faiblement dispersés.

Contrairement à la moyenne (**Figure 76.B**) ou la somme (**Figure 76.C**), la médiane (**Figure 76.A**) est une statistique robuste. Seuls les échantillons où au moins la moitié des gènes représentatifs sont détectés auront un comptage non nul. Ceci permet de filtrer les signaux parasites qui génèrent des valeurs aberrantes perturbant la mesure de proportionnalité.

Néanmoins, la somme est plus sensible car elle cumule les comptages de plusieurs gènes. Elle permet d'estimer plus précisément l'abondance de la MSP en particulier dans les échantillons où sa couverture est faible. Lorsque l'on compare le gène représentant une MSP à un de ses gènes et que le coefficient de proportionnalité est strictement supérieur à 1, ceci signifie que des échantillons potentiellement exploitables ne sont pas considérés. Avec la somme, on s'assure que le coefficient de proportionnalité est inférieur à 1 quel que soit le gène et que tous les échantillons avec suffisamment de signal sont considérés. Dans l'exemple ci-dessous, le coefficient de proportionnalité passe de 2.64 avec la médiane **Figure 76.A**) à 0.47 avec la somme (**Figure 76.C**). Quant au nombre d'échantillons où les deux gènes sont détectés sans ambigüité (comptage supérieur à 5), il passe respectivement de 367 à 463. Finalement, la somme présente un avantage conceptuel par rapport à la médiane. L'usage de la médiane est pertinent sur des comptages normalisés mais l'utiliser sur des données brutes est difficilement justifiable. En effet, lors du calcul de la médiane on compare et trie des comptages potentiellement très différents car les 30 gènes n'ont pas la même longueur. En pratique, la médiane fonctionne car les 30 gènes représentants ont un signal très cohérent. Par conséquent, leur vecteur médian est à peu près égal au signal du 15ème gène ayant le coefficient de proportionnalité le plus fort.

Pour utiliser la somme tout en limitant les faux positifs, on envisage de ne considérer que les échantillons où au moins la moitié de gènes représentatifs de la MSP ont un comptage non nul (**Figure 76.D**). Cette stratégie semble cumuler à la fois la robustesse de la médiane et la sensibilité de la somme.

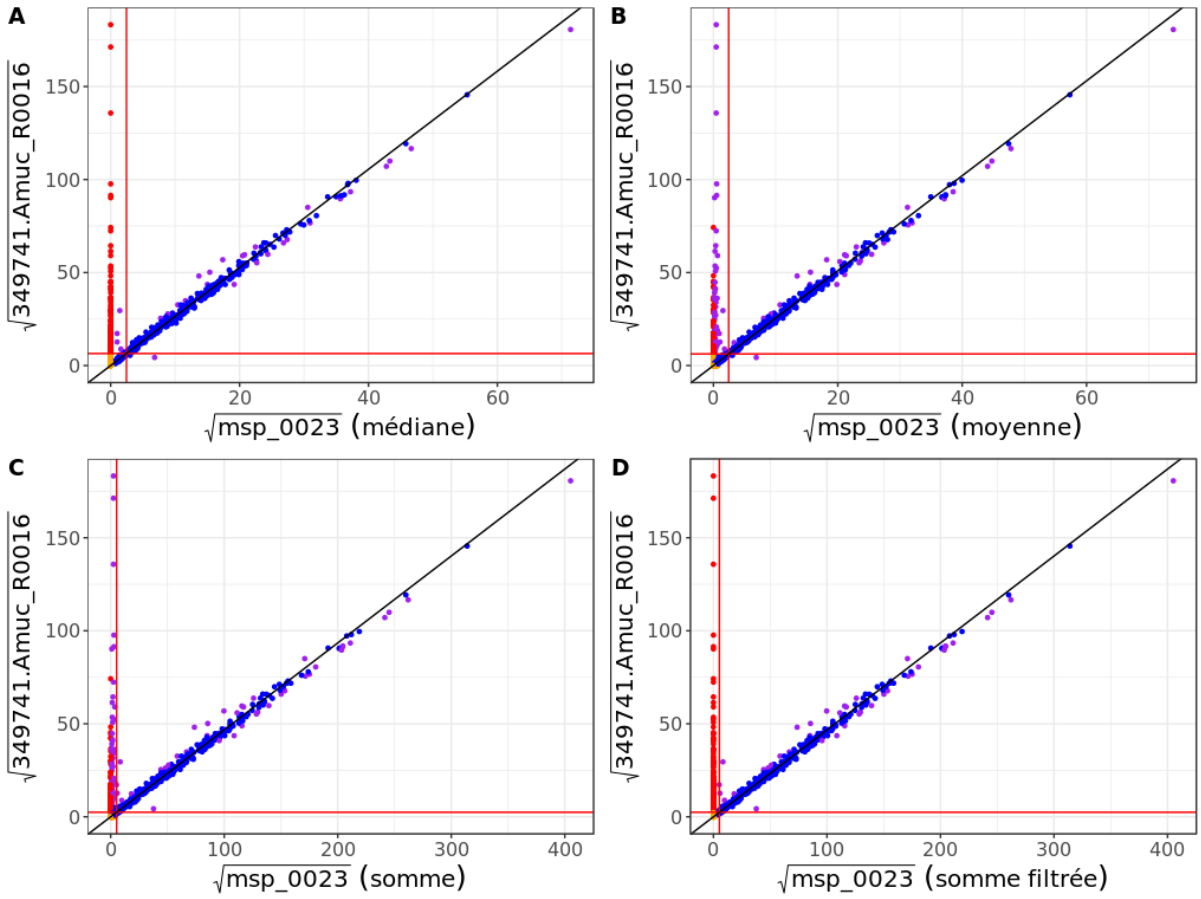


***Figure 76** : Comparaison de l'abondance du gène représentant de la msp_0023 (Akkermansia muciniphila) et un de ses gènes core partagés (349741.Amuc_R0016) en utilisant différentes statistiques.*

A. Le gène représentant correspond à l'abondance médiane des 30 gènes core représentatifs de la msp_0023.

B. Le gène représentant correspond à l'abondance moyenne des 30 gènes core représentatifs de la msp_0023. Contrairement à la médiane, on note à gauche la présence de comptages faibles (points violets) correspondant vraisemblablement à des faux positifs.

C. Le gène représentant correspond à la somme des abondances des 30 gènes core représentatifs de la msp_0023. Tout comme la moyenne, on note à gauche la présence de comptages faibles (points violets) correspondant vraisemblablement à des faux positifs.

D. Le gène représentant correspond à la somme filtrée des abondances des 30 gènes core représentatifs de la msp_0023. Contrairement à la moyenne et à la médiane, les faux positifs ont été filtrés.

9.1.2 Estimation du coefficient de proportionnalité par régression linéaire robuste

Actuellement, le coefficient de proportionnalité entre deux gènes co-abondants est estimé par la médiane du rapport de leurs comptages non nuls. La médiane présente plusieurs avantages : elle est rapide à calculer et est relativement robuste car elle tolère une certaine proportion de valeurs aberrantes.

Néanmoins, la régression linéaire robuste [178] est une candidate de choix pour la remplacer car elle permettrait d'accroitre la robustesse et la précision de l'estimation en contrepartie d'une hausse raisonnable du temps de calcul. Brièvement, la régression linéaire robuste est une méthode itérative

consistant à attribuer aux points un poids compris entre 0 et 1, le poids des points baissant d'autant plus qu'ils s'éloignent de la tendance observée. Seuls les points à poids fort auront un impact lors de l'estimation du coefficient de proportionnalité. Le poids attribué aux points étant une valeur continue, on dépassera les limites de classification binaire actuelle dans laquelle un point est une valeur aberrante ou ne l'est pas.

9.1.3 Utilisation de matrices creuses

Les matrices de comptage traitées sont composées d'une majorité de zéros (parfois plus de 90%) : il s'agit de matrices creuses. Ces matrices devraient être stockées dans une structure de données adaptée [179] car énormément de mémoire vive est perdue pour stocker des zéros.

Cette baisse sensible de consommation de mémoire vive permettra de passer à l'échelle pour faire face au nombre croissant d'échantillons métagénomiques disponibles. Aussi, on pourrait éventuellement traiter les matrices actuelles sur un ordinateur de bureau moderne. Néanmoins, l'utilisation de telles structures de données nécessitera des algorithmes adaptés plus complexes. Ainsi, la baisse de consommation de mémoire vive s'accompagnera vraisemblablement d'une baisse des performances.

9.2 Utilisation des MSPs pour les analyses fonctionnelles

Jusqu'ici l'abondance des modules fonctionnels est calculée en sommant (ou en moyennant) les abondances des gènes assignés aux KOs (familles de gènes orthologues) composant le module. Cette approche décloisonnée considérant le métagénome comme une « soupe » de gènes permet difficilement d'établir un lien entre le potentiel fonctionnel et la composition taxonomique du microbiote. Certains ont identifié les espèces contribuant à un module fonctionnel en calculant des corrélations entre l'abondance du module et l'abondance des espèces [180]. Toutefois, cette approche ne met en évidence que les espèces contribuant majoritairement à l'abondance du module et fonctionne mal quand plusieurs espèces y participent équitablement. De plus, l'approche décloisonnée est sensible aux erreurs d'annotation fonctionnelle. En effet, un gène abondant associé à tort à un module fonctionnel biaise fortement son abondance ce qui peut mener à des conclusions erronées.

Nous proposons une autre approche dite cloisonnée qui consiste à calculer l'abondance d'un module fonctionnel à partir de l'abondance des MSPs le possédant. Cette méthode présente l'avantage d'établir explicitement la liste des MSPs contribuant à l'abondance du module, y compris celles ayant un apport minoritaire. L'approche cloisonnée considère qu'un module fonctionnel est présent seulement si la majorité des KOs composant le module sont retrouvés dans la MSP. Ainsi, les gènes associés à tort à un KO biaiseront moins les résultats car ils ne seront pas considérés lors de l'estimation des abondances des modules fonctionnels.

Dans l'approche cloisonnée, l'abondance d'un module fonctionnel peut être calculée de deux façons. La première stratégie considère que les modules fonctionnels présents dans une MSP sont systématiquement présents dans les échantillons où le core de cette MSP est détecté. Ainsi, l'abondance du module dans la MSP est égale à l'abondance du core de la MSP. La deuxième stratégie prend en compte le fait que les individus sont porteurs de souches différentes de la même espèce et que ces souches peuvent avoir un potentiel fonctionnel différent. Dans ce cas, l'abondance d'un module dans la MSP est égale à l'abondance du core de la MSP que si ce module est effectivement présent dans l'échantillon considéré. Cette stratégie tire pleinement partie des MSPs car celles-ci groupent non seulement les gènes assurant des fonctions core mais aussi des gènes accessoires assurant des fonctions optionnelles.

Néanmoins, l'approche cloisonnée ne peut être mise en œuvre qu'avec des MSPs suffisamment complètes, c'est-à-dire des MSPs contenant des gènes représentatifs du potentiel fonctionnel de l'espèce et dans l'idéal de ses différentes souches. Autrement, les modules fonctionnels auront une abondance estimée bien inférieure à la réalité.

9.3 Reconstitution de génomes par assemblage métagénomique

Lorsque la couverture de séquençage est suffisante, les génomes des microorganismes présents dans un échantillon métagénomique peuvent directement être reconstruits en procédant à un assemblage *de novo* suivi du regroupement des contigs constitués de gènes provenant de la même MSP. Cette stratégie déjà mise en œuvre en utilisant les objets créés par la méthode Canopy [112,181] s'apparente à celles implémentées dans les outils de regroupement de contigs par taxonomie comme MEGAN [182]. Cependant, l'utilisation des MSPs permet l'assemblage de souches correspondant à des espèces sans génome de référence disponible. En excluant les gènes partagés des MSPs, on ne considère que les gènes spécifiques des espèces et on limite ainsi le risque de créer des chimères.

Pour obtenir des assemblages les plus contigus possibles, on effectue en amont un profilage taxonomique basé les MSPs pour lister les échantillons où les espèces d'intérêt sont les mieux couvertes. En outre, on vérifiera grâce à l'arbre phylogénétique qu'aucune espèce proche n'est présente pour éviter la création de contigs chimériques. Néanmoins, cette stratégie n'est efficace que si la MSP représentative de l'espèce à assembler est suffisamment complète. En effet, certains de ses contigs pourraient être manqués s'ils ne contiennent que des gènes non regroupés dans la MSP.

Cette approche présente l'avantage de capturer les fragments de petite taille dès lors qu'ils contiennent un gène spécifique d'une MSP, là où les outils regroupant les contigs par couverture et/ou composition tétranucléotidique ne traitent que les contigs d'au moins 2 500 paires de bases [162]. De plus, les outils de binning de contigs supposent que la structure d'un contig est conservée dans les échantillons où il est quantifié. Or, les gènes accessoires composant le contig ne seront pas nécessairement présents dans tous les échantillons porteurs de l'espèce d'intérêt car les souches sont différentes de celle dont dérive le contig. Ainsi, on peut supposer que la couverture des contigs provenant du même microorganisme sera inférieure à celle attendue dans certains échantillons.

9.4 Les MSPs : un tremplin vers l'isolation et la culture de microorganismes d'intérêt

Même si les MSPs permettent l'identification d'espèces inconnues, la culture microbienne demeure indispensable pour déclarer une nouvelle espèce auprès du *Code International de la Nomenclature Bactérienne*. En effet, en plus de montrer des différences génétiques (ARNr 16S, gènes de ménage), morphologiques, métaboliques (glucides et acides aminés) et biochimiques (composition de la membrane de cellulaire) avec les espèces déjà décrites, la souche type de la nouvelle espèce doit être disponible à l'achat dans des collections de microorganismes (DSMZ, ATCC).

Le profilage de milliers d'échantillons métagénomiques avec les MSPs pourrait guider les méthodes de caractérisation du microbiote intestinal basées sur la culture microbienne en fournissant une liste mise à jour des espèces les plus recherchées (*most wanted species*) [183]. Il s'agit soit d'espèces prévalentes associées à un phénotype d'intérêt pour lesquelles aucun génome séquencé n'est disponible à ce jour ou avec des génomes de référence distants des souches présentes dans les échantillons analysés.

Le potentiel métabolique de ces espèces d'intérêt pourrait être inféré de l'annotation fonctionnelle des génomes reconstruits à partir de l'assemblage de métagénomes. Ainsi, la mise en œuvre de milieux de culture enrichis et/ou sélectifs prenant en compte les spécificités de l'espèce ciblée pourraient faciliter leur isolement et leur culture [184]. Enfin les gènes core des MSPs sont des candidats de choix pour concevoir des sondes (PCR, FISH etc.) spécifiques d'une espèce qui faciliteraient son isolation.

9.5 Application à d'autres écosystèmes microbiens

MSPminer est un outil flexible qui peut s'appliquer à d'autres écosystèmes microbiens même si lors de cette thèse, nous nous sommes restreints à l'étude du microbiote intestinal humain. Dès lors que l'on dispose d'une table quantifiant les gènes représentatifs de l'écosystème étudié dans un ensemble d'échantillons, on peut reconstituer des MSPs pour recenser les espèces présentes et appréhender la variation de leur contenu en gènes.

A ce jour nous avons utilisé MSPminer pour étudier le microbiote intestinal d'animaux d'élevage dont celui du rumen bovin, du porc et du poulet ainsi que le microbiote de la cavité orale chez l'homme.

Bibliographie

1. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol. 2016;14:1–14.

2. Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen T, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science (80-) [Internet]. 2016;352:565–9. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.aad3369

3. Gaci N, Borrel G, Tottey W, O'Toole PW, Brugère JF. Archaea and the human gut: New beginning of an old story. World J Gastroenterol. 2014;20:16062–78.

4. Lukeš J, Stensvold CR, Jirků-Pomajbíková K, Wegener Parfrey L. Are Human Intestinal Eukaryotes Beneficial or Commensals? PLoS Pathog. 2015;11:7–12.

5. Scanlan PD, Marchesi JR. Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces. ISME J [Internet]. 2008;2:1183–93. Available from: http://www.nature.com/doifinder/10.1038/ismej.2008.76

6. Reyes A, Semenkovich NP, Whiteson K, Rohwer F, Gordon JI. Going viral: next-generation sequencing applied to phage populations in the human gut. Nat Rev Microbiol [Internet]. Nature Publishing Group; 2012;10:607–17. Available from: http://www.nature.com/doifinder/10.1038/nrmicro2853

7. Bilen M, Dufour J-C, Lagier J-C, Cadoret F, Daoud Z, Dubourg G, et al. The contribution of culturomics to the repertoire of isolated human bacterial and archaeal species. Microbiome 2018 61 [Internet]. Microbiome; 2018;6:94. Available from: https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0485-5%0Ahttps://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0485-5

8. Rodríguez JM, Murphy K, Stanton C, Ross RP, Kober OI, Juge N, et al. The composition of the gut microbiota throughout life, with an emphasis on early life. Microb Ecol Heal Dis [Internet]. 2015;26:1–17. Available from: http://www.microbecolhealthdis.net/index.php/mehd/article/view/26050

9. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature [Internet]. Nature Publishing Group; 2014;505:559–63. Available from: http://dx.doi.org/10.1038/nature12820

10. Armand-Lefèvre L, Andremont A, Ruppé E. Voyages et acquisition d'entérobactéries multirésistantes. Med Mal Infect [Internet]. Elsevier Masson SAS; 2018; Available from: https://doi.org/10.1016/j.medmal.2018.02.005

11. Maier L, Pruteanu M, Kuhn M, Zeller G, Telzerow A, Anderson EE, et al. Extensive impact of non-antibiotic drugs on human gut bacteria. Nature [Internet]. Nature Publishing Group; 2018;555:623–8. Available from: http://dx.doi.org/10.1038/nature25979

12. Francino MP. Antibiotics and the human gut microbiome: Dysbioses and accumulation of resistances. Front Microbiol. 2016;6:1–11.

13. Kumar M, Babaei P, Ji B, Nielsen J. Human gut microbiota and healthy aging: Recent developments and future prospective. Nutr Heal Aging [Internet]. 2016;4:3–16. Available from: http://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/NHA-150002

14. De Weirdt R, Van de Wiele T. Micromanagement in the gut: microenvironmental factors govern colon mucosal biofilm structure and functionality. npj Biofilms Microbiomes [Internet]. Nature Publishing Group; 2015;1:15026. Available from: http://www.nature.com/articles/npjbiofilms201526

15. Qin J, Li R, Raes J, Arumugam M, Burgdorf S, Manichanh C, et al. A human gut microbial gene catalog established by metagenomic sequencing. Nature. 2010;464:59–65.

16. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature [Internet]. 2011;474:666–666. Available from: http://www.nature.com/doifinder/10.1038/nature10187

17. Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, et al. Environment dominates over host genetics in shaping human gut microbiota. Nature [Internet]. Nature Publishing Group; 2018;555:210–5. Available from: http://dx.doi.org/10.1038/nature25973

18. Gupta VK, Paul S, Dutta C. Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. Front Microbiol. 2017;8.

19. Zhu A, Sunagawa S, Mende DR, Bork P. Inter-individual differences in the gene content of human gut bacterial species. Genome Biol [Internet]. 2015;16:82. Available from: http://genomebiology.com/2015/16/1/82

20. Scholz M, Ward D V, Pasolli E, Tolio T, Zolfo M, Asnicar F, et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods [Internet]. 2016;13:435–8. Available from: http://www.nature.com/doifinder/10.1038/nmeth.3802

21. Greenblum S, Carr R, Borenstein E. Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species. Cell [Internet]. 2015;160:583–94. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0092867415000136

22. Costea PI, Coelho LP, Sunagawa S, Munch R, Huerta-Cepas J, Forslund K, et al. Subspecies in the global human gut microbiome. Mol Syst Biol. 2017;

23. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science (80-). 2011;

24. Lukjancenko O, Wassenaar TM, Ussery DW. Comparison of 61 Sequenced Escherichia coli Genomes. Microb. Ecol. 2010.

25. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr. Opin. Genet. Dev. 2005. p. 589–94.

26. Carlos Guimaraes L, Benevides de Jesus L, Vinicius Canario Viana M, Silva A, Thiago Juca Ramos R, de Castro Soares S, et al. Inside the Pan-genome - Methods and Software Overview. Curr Genomics. 2015;

27. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. Nat. Microbiol. 2017.

28. Touchon M, Hoede C, Tenaillon O, Barbe VVV, Baeriswyl S, Bidet P, et al. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. Casades?s J, editor. PLoS Genet [Internet]. 2009;5:e1000344. Available from: http://dx.plos.org/10.1371/journal.pgen.1000344

29. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli S V, et al. Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. Genome Biol [Internet]. 2010;11:R107. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r107

30. Koonin E V., Wolf YI. Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 2008;36:6688–719.

31. Treangen TJ, Rocha EPC. Horizontal transfer, not duplication, drives the expansion of protein

families in prokaryotes. PLoS Genet. 2011;

32. Marteau P, Doré J. Le microbiote intestinal: un organe à part entière. Ed John Libbey. 2017;

33. Rowland I, Gibson G, Heinken A, Scott K, Swann J, Thiele I, et al. Gut microbiota functions: metabolism of nutrients and other food components. Eur J Nutr. Springer Berlin Heidelberg; 2018;57:1–24.

34. Scott KP, Gratz SW, Sheridan PO, Flint HJ, Duncan SH. The influence of diet on the gut microbiota. Pharmacol Res [Internet]. Elsevier Ltd; 2013;69:52–60. Available from: http://dx.doi.org/10.1016/j.phrs.2012.10.020

35. Arora T, Sharma R, Frost G. Propionate. Anti-obesity and satiety enhancing factor? Appetite [Internet]. Elsevier Ltd; 2011;56:511–5. Available from: http://dx.doi.org/10.1016/j.appet.2011.01.016

36. Canani RB, Costanzo M Di, Leone L, Pedata M, Meli R, Calignano A. Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. World J Gastroenterol. 2011;17:1519–28.

37. LeBlanc JG, Milani C, de Giori GS, Sesma F, van Sinderen D, Ventura M. Bacteria as vitamin suppliers to their host: A gut microbiota perspective. Curr Opin Biotechnol [Internet]. Elsevier Ltd; 2013;24:160–8. Available from: http://dx.doi.org/10.1016/j.copbio.2012.08.005

38. Natividad JMM, Verdu EF. Modulation of intestinal barrier by intestinal microbiota: Pathological and therapeutic implications. Pharmacol. Res. 2013.

39. Kelly JR, Kennedy PJ, Cryan JF, Dinan TG, Clarke G, Hyland NP. Breaking down the barriers: the gut microbiome, intestinal permeability and stress-related psychiatric disorders. Front Cell Neurosci. 2015;

40. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature [Internet]. 2014;513:59–64. Available from: http://www.nature.com/doifinder/10.1038/nature13568

41. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, et al. Richness of human gut microbiome correlates with metabolic markers. Nature [Internet]. 2013;500:541–6. Available from: http://www.nature.com/doifinder/10.1038/nature12506

42. 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 [Internet]. 2012;490:55–60. Available from: http://www.nature.com/doifinder/10.1038/nature11450

43. Loomba R, Seguritan V, Li W, Long T, Klitgord N, Bhatt A, et al. Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. Cell Metab. 2017;

44. Jie Z, Xia H, Zhong S-L, Feng Q, Li S, Liang S, et al. The gut microbiome in atherosclerotic cardiovascular disease. Nat Commun [Internet]. 2017;8:845. Available from: http://www.nature.com/articles/s41467-017-00900-1

45. Manichanh C, Borruel N, Casellas F, Guarner F. The gut microbiota in IBD. Nat. Rev. Gastroenterol. Hepatol. 2012.

46. Wen C, Zheng Z, Shao T, Liu L, Xie Z, Le Chatelier E, et al. Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. Genome Biol. 2017;

47. Kostic AD, Gevers D, Siljander H, Vatanen T, Hyötyläinen T, Hämäläinen A-M, et al. The Dynamics of the Human Infant Gut Microbiome in Development and in Progression toward Type 1 Diabetes. Cell Host Microbe. 2015;

48. Marasco G, Di Biase AR, Schiumerini R, Eusebi LH, Iughetti L, Ravaioli F, et al. Gut Microbiota and Celiac Disease. Dig. Dis. Sci. 2016.

49. Fujimura KE, Lynch S V. Microbiota in allergy and asthma and the emerging relationship with the gut microbiome. Cell Host Microbe. 2015.

50. Bedarf JR, Hildebrand F, Coelho LP, Sunagawa S, Bahram M, Goeser F, et al. Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naïve Parkinson's disease patients. Genome Med. 2017;

51. Pistollato F, Cano SS, Elio I, Vergara MM, Giampieri F, Battino M. Role of gut microbiota and nutrients in amyloid formation and pathogenesis of Alzheimer disease. Nutr Rev. 2016;

52. Cryan JF, Dinan TG. Mind-altering microorganisms: The impact of the gut microbiota on brain and behaviour. Nat. Rev. Neurosci. 2012.

53. Li Q, Han Y, Dy ABC, Hagerman RJ. The Gut Microbiota and Autism Spectrum Disorders. Front Cell Neurosci. 2017;

54. 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;

55. Collins SM. A role for the gut microbiota in IBS. Nat Rev Gastroenterol Hepatol. 2014;

56. Group JFW, Group JFW, others. Guidelines for the evaluation of probiotics in food. London World Heal Organ ON, Canada Food Agric Organ. 2002;

57. Routy B, Le Chatelier E, Derosa L, Duong CPM, Alou MT, Daillère R, et al. Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. Science (80-). 2018;

58. Haiser HJ, Gootenberg DB, Chatman K, Sirasani G, Balskus EP, Turnbaugh PJ. Predicting and manipulating cardiac drug inactivation by the human gut bacterium Eggerthella lenta. Science (80-). 2013;

59. Barnich N, Boudeau J, Claret L, Darfeuille-Michaud A. Regulatory and functional co-operation of flagella and type 1 pill in adhesive and invasive abilities of AIEC strain LF82 isolated from a patient with Crohn's disease. Mol Microbiol. 2003;

60. Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res [Internet]. 2016;26:1612–25. Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.201863.115

61. Alain K, Querellou J, Joe KAÆ. Cultivating the uncultured: Limits, advances and future challenges. Extremophiles. 2009;13:583–94.

62. Wieser A, Schneider L, Jung J, Schubert S. MALDI-TOF MS in microbiological diagnostics-identification of microorganisms and beyond (mini review). Appl Microbiol Biotechnol. 2012;93:965–74.

63. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol [Internet]. Nature Publishing Group; 2014;12:635–45. Available from: http://www.nature.com/doifinder/10.1038/nrmicro3330

64. Maidak BL, Olsen GJ, Larsen N, Overbeek R, McCaughey MJ, Woese CR. The RDP (Ribosomal Database Project). Nucleic Acids Res. 1997;

65. 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. 2013;

66. 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;

67. Meyerhans A, Vartanian JP, Wain-Hobson S. DNA recombination during PCR. Nucleic Acids Res. 1990;18:1687–91.

68. Mathieu-Daudé F, Welsh J, Vogt T, McClelland M. DNA rehybridization during PCR: the "Cot effect" and its consequences. Nucleic Acids Res [Internet]. 1996;24:2080–6. Available from: http://nar.oxfordjournals.org/content/24/11/2080.short%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=145907&tool=pmcentrez&rendertype=abstract

69. Větrovský T, Baldrian P. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. PLoS One. 2013;8.

70. Wagner J, Coupland P, Browne HP, Lawley TD, Francis SC, Parkhill J, et al. Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. BMC Microbiol [Internet]. BMC Microbiology; 2016;16:274. Available from: http://bmcmicrobiol.biomedcentral.com/articles/10.1186/s12866-016-0891-4

71. Kembel SW, Wu M, Eisen JA, Green JL. Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. PLoS Comput Biol. 2012;

72. Langille M, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes J, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013;

73. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun [Internet]. Elsevier Ltd; 2016;469:967–77. Available from: http://dx.doi.org/10.1016/j.bbrc.2015.12.083

74. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, et al. Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol. 2017;

75. Handelsman J, Tiedje J, National Research Council (US) Committee on Metagenomics: Challenges and Functional, Applications. THE NEW SCIENCE OF METAGENOMICS: Revealing the Secrets of Our Microbial Planet. Natl. Acad. Press. 2007.

76. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;

77. Maillet N, Lemaitre C, Chikhi R, Lavenier D, Peterlongo P. Compareads: comparing huge metagenomic experiments. BMC Bioinformatics [Internet]. 2012;13:S10. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-S19-S10

78. Maillet N, Collet G, Vannier T, Lavenier D, Peterlongo P. Commet: Comparing and combining multiple metagenomic datasets. Proc - 2014 IEEE Int Conf Bioinforma Biomed IEEE BIBM 2014. 2014;94–8.

79. Benoit G, Peterlongo P, Mariadassou M, Drezen E, Schbath S, Lavenier D, et al. Multiple Comparative Metagenomics using Multiset k-mer Counting. 2016;1–25. Available from: http://arxiv.org/abs/1604.02412

80. Deorowicz S, Kokot M, Grabowski S, Debudaj-Grabysz A. KMC 2: Fast and resource-frugal k-mer counting. Bioinformatics. 2015;

81. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol [Internet]. Genome Biology;

2016;17:132. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0997-x

82. Plaza Onate F, Batto JM, Juste C, Fadlallah J, Fougeroux C, Gouas D, et al. Quality control of microbiota metagenomics by k-mer analysis. BMC Genomics. 2015;

83. Brady A, Salzberg SL. Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated Markov models. Nat Methods [Internet]. Nature Publishing Group; 2009;6:673–6. Available from: http://dx.doi.org/10.1038/nmeth.1358

84. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic\nsequences. Genome Res. 2016;gr.210641.116.

85. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol [Internet]. 2014;15:R46. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-3-r46

86. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics [Internet]. 2015;16:236. Available from: http://www.biomedcentral.com/1471-2164/16/236

87. Ciccarelli FD, Doerks T, Mering C Von, Christopher J. Toward of a Automatic Reconstruction Tree of Life Highly Resolved. Science [Internet]. 2014;311:1283–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16513982

88. Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics [Internet]. 2011;12:S4. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-12-S2-S4

89. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012;9:811–4.

90. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods [Internet]. Nature Publishing Group; 2015;12:902–3. Available from: http://www.nature.com/doifinder/10.1038/nmeth.3589

91. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;

92. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 1999. p. 29–34.

93. Davis C. mBLAST: Keeping up with the Sequencing Explosion for (Meta) Genome Analysis. J Data Mining Genomics Proteomics. 2013;

94. Darzi Y, Falony G, Vieira-Silva S, Raes J. Towards biome-specific analysis of meta-omics data. ISME J. 2016.

95. Li W. Analysis and comparison of very large metagenomes with fast clustering and functional annotation. BMC Bioinformatics. 2009;

96. Nayfach S, Bradley PH, Wyman SK, Laurent TJ, Williams A, Eisen JA, et al. Automated and Accurate Estimation of Gene Family Abundance from Shotgun Metagenomes. PLoS Comput Biol. 2015;

97. Manor O, Borenstein E. Systematic Characterization and Analysis of the Taxonomic Drivers of Functional Shifts in the Human Microbiome. Cell Host Microbe. 2017;

98. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure & genetic diversity from metagenomes. Genome Res. 2017;27:626–38.

99. Costea PI, Munch R, Coelho LP, Paoli L, Sunagawa S, Bork P. metaSNV: A tool for metagenomic strain level analysis. PLoS One. 2017;12:1–9.

100. Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, Fedi S, et al. Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host Microbe. 2018;

101. Albanese D, Donati C. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. Nat Commun. 2017;

102. Ward D V., Scholz M, Zolfo M, Taft DH, Schibler KR, Tett A, et al. Metagenomic Sequencing with Strain-Level Resolution Implicates Uropathogenic E. coli in Necrotizing Enterocolitis and Mortality in Preterm Infants. Cell Rep. 2016;

103. Li D, Liu C-MM, Luo R, Sadakane K, Lam T-WW. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics [Internet]. 2015;31:1674–6. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv033

104. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. MetaSPAdes: A new versatile metagenomic assembler. Genome Res. 2017;

105. van der Walt AJ, van Goethem MW, Ramond JB, Makhalanyane TP, Reva O, Cowan DA. Assembling metagenomes, one community at a time. BMC Genomics. 2017;

106. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, et al. Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. Nat Methods. 2017;14:1063–71.

107. Pop M. Genome assembly reborn: Recent computational challenges. Brief Bioinform. 2009;10:354–66.

108. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature [Internet]. 2010;464:59–65. Available from: http://www.nature.com/doifinder/10.1038/nature08821

109. Hyatt D, Locascio PF, Hauser LJ, Uberbacher EC. Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics. 2012;28:2223–30.

110. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. 2010;

111. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.

112. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol [Internet]. 2014 [cited 2014 Jul 9];32:822–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24997787

113. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol [Internet]. 2014;32:834–41. Available from: http://www.nature.com/doifinder/10.1038/nbt.2942

114. Wang J, Jia H. Metagenome-wide association studies: fine-mining the microbiome. Nat Rev Microbiol [Internet]. 2016;14:508–22. Available from: http://www.nature.com/doifinder/10.1038/nrmicro.2016.83

115. Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, et al. MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit. PLoS One. 2012;7.

116. Coelho LP, Alves R, Monteiro P, Huerta-Cepas J, Freitas AT, Bork P. NG-meta-profiler: fast processing of metagenomes using NGLess, a domain-specific language. bioRxiv. 2018;

117. Pons N, Batto J-M, Kennedy S, Almeida M, Boumezbeur F, Moumen B, et al. METEOR -a plateform for quantitative metagenomic profiling of complex ecosystems. 2010.

118. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;

119. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;

120. Jonsson V, Österlund T, Nerman O, Kristiansson E. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. BMC Genomics. 2016;

121. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing [Internet]. J. R. Stat. Soc. B. 1995. p. 289–300. Available from: http://www.stat.purdue.edu/~doerge/BIOINFORM.D/FALL06/Benjamini and Y FDR.pdf%5Cnhttp://engr.case.edu/ray_soumya/mlrg/controlling_fdr_benjamini95.pdf

122. Schwartzman A, Lin X. The effect of correlation in false discovery rate estimation. Biometrika [Internet]. 2011;98:199–214. Available from: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/asq075

123. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods [Internet]. 2013;10:1196–9. Available from: http://www.nature.com/doifinder/10.1038/nmeth.2693

124. Tello D, Boumezbeur F, Arslan V, Ducrot V, Leonard P, Moumen B, et al. Optimizations to compute large correlation matrix onto GPU system of hybrid HPC clusters. 2012.

125. OpenGPU [Internet]. Available from: http://opengpu.net/

126. Huang S, Xiao S, Feng W. On the energy efficiency of graphics processing units for scientific computing. Parallel Distrib Process 2009 IPDPS 2009 IEEE Int Symp. 2009;

127. Karypis G, Han EH, Kumar V. Chameleon: Hierarchical clustering using dynamic modeling. Computer (Long Beach Calif) [Internet]. 2002;32:68–75. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=781637

128. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res [Internet]. 2002;30:1575–84. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=101833&tool=pmcentrez&rendertype=abstract%5Cnciteulike-article-id:159967%5Cnhttp://dx.doi.org/10.1093/nar/30.7.1575

129. McCallum A, Nigam K, Ungar LH. Efficient clustering of high-dimensional data sets with application to reference matching. Proc sixth ACM SIGKDD Int Conf Knowl Discov data Min - KDD '00 [Internet]. 2000;169–78. Available from: http://portal.acm.org/citation.cfm?doid=347090.347123

130. Almeida M, Pop M, Le Chatelier E, Prifti E, Pons N, Ghozlane A, et al. Capturing the most wanted taxa through cross-sample correlations. ISME J [Internet]. 2016;10:2459–67. Available from: http://www.nature.com/doifinder/10.1038/ismej.2016.35

131. Bland JM, Altman DG. Statistics Notes: Transforming data. Bmj [Internet]. 1996;312:770–770. Available from: http://www.bmj.com/cgi/doi/10.1136/bmj.312.7033.770

132. McMurdie PJ, Holmes S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Comput Biol. 2014;10.

133. O'Hara RB, Kotze DJ. Do not log-transform count data. Methods Ecol Evol [Internet]. 2010;1:118–22. Available from: http://doi.wiley.com/10.1111/j.2041-210X.2010.00021.x

134. Ridout M, Demetrio CG., Hinde J. Models for count data with many zeros. Int Biometric Conf. 1998;1–13.

135. Huson LW. Performance of Some Correlation Coefficients When Applied to Zero-Clustered Data. J Mod Appl Stat Methods [Internet]. 2007;6:530–6. Available from: http://digitalcommons.wayne.edu/jmasm%5Cnhttp://digitalcommons.wayne.edu/jmasm/vol6/iss2/17

136. Kowalski CJ. On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient. Appl Stat [Internet]. 1972;21:1. Available from: http://www.jstor.org/stable/10.2307/2346598?origin=crossref

137. Osborne CJ. Best Practices in Quantitative Methods [Internet]. Soc. Sci. 2008. Available from: http://srmo.sagepub.com/view/best-practices-in-quantitative-methods/SAGE.xml

138. Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. J Exp Soc Psychol [Internet]. Elsevier Inc.; 2013;49:764–6. Available from: http://dx.doi.org/10.1016/j.jesp.2013.03.013

139. Bromiley PA, Thacker NA. The Effects of a Square Root Transform on a Poisson Distributed Quantity. Tech Rep. 2001;1–5.

140. Joanes DN, Gill CA. Comparing measures of sample skewness and kurtosis. J R Stat Soc Ser D (The Stat [Internet]. 1998;47:183–9. Available from: http://doi.wiley.com/10.1111/1467-9884.00122

141. Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, B??hler J. Proportionality: A Valid Alternative to Correlation for Relative Data. PLoS Comput Biol. 2015;11.

142. Mohri M, Roark B. Structural zeros versus sampling zeros. 2005;1–7. Available from: https://www.cslu.ogi.edu/people/roark/zero.pdf%5Cnhttp://www.cs.nyu.edu/~mohri/pub/zero.pdf

143. Lin LI-K. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics [Internet]. 1989;45:255. Available from: http://www.jstor.org/stable/2532051?origin=crossref

144. Wilcox RR. Inferences based on a skipped correlation coefficient. J Appl Stat. 2004;31:131–43.

145. Tukey JW. Exploratory Data Analysis. Analysis. 1977.

146. Agresti A. A Survey of Exact Inference for Contingency Tables. Stat Sci [Internet]. 1992;7:131–53. Available from: http://projecteuclid.org/euclid.ss/1177011454

147. Dagum L, Menon R. OpenMP: an industry standard API for shared-memory programming. IEEE Comput Sci Eng. IEEE; 1998;5:46–55.

148. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. Supporting data for the paper: "An integrated catalog of reference genes in the human gut microbiome" [Internet]. GigaScience Database; 2014. Available from: http://gigadb.org/dataset/100064

149. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol [Internet]. 1990;215:403–10. Available from: http://www.sciencedirect.com/science/article/pii/S0022283605803602

150. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic

Acids Res. 2004;32:1792–7.

151. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.

152. Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5.

153. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44:W242–5.

154. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal Escherichia coli. Nat Rev Microbiol [Internet]. 2010;8:207–17. Available from: http://www.nature.com/doifinder/10.1038/nrmicro2298

155. Dehoux P, Marvaud JC, Abouelleil A, Earl AM, Lambert T, Dauga C. Comparative genomics of Clostridium bolteae and Clostridium clostridioforme reveals species-specific genomic properties and numerous putative antibiotic resistance determinants. BMC Genomics [Internet]. 2016;17:819. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-3152-x

156. Guo X, Li S, Zhang J, Wu F, Li X, Wu D, et al. Genome sequencing of 39 Akkermansia muciniphila isolates reveals its population structure, genomic and functional diverisity, and global distribution in mammalian gut microbiotas. BMC Genomics. BMC Genomics; 2017;18:1–12.

157. Ruppe E, Ghozlane A, Tap J, Pons N, Alvarez A-S, Maziers N, et al. Prediction of the intestinal resistome by a novel 3D-based method. bioRxiv [Internet]. 2017; Available from: http://biorxiv.org/content/early/2017/09/29/196014.abstract

158. Xu J, Mahowald MA, Ley RE, Lozupone CA, Hamady M, Martens EC, et al. Evolution of Symbiotic Bacteria in the Distal Human Intestine. PLoS Biol. 2007;5:1574–86.

159. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res [Internet]. 2009;19:1639–45. Available from: http://genome.cshlp.org/cgi/doi/10.1101/gr.092759.109

160. Plaza Oñate F. fetch-cscg [Internet]. Available from: https://github.com/fplaza/fetch-cscg

161. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature [Internet]. Nature Publishing Group; 2013;499:431–7. Available from: http://www.nature.com/doifinder/10.1038/nature12352

162. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ [Internet]. 2015;3:e1165. Available from: https://peerj.com/articles/1165

163. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. Nat Methods [Internet]. 2014;11:1144–6. Available from: http://www.nature.com/doifinder/10.1038/nmeth.3103

164. Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO. Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol. 2004;

165. mondiale de la Santé A. Stratégie mondiale pour l'alimentation, l'exercice physique et la santé. 2004.

166. Plaza Oñate F. sequence-translator [Internet]. Available from: https://github.com/fplaza/sequence-translator

167. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for

prokaryotes. Proc Natl Acad Sci [Internet]. 2005;102:2567–72. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0409727102

168. Yoon SH, Ha S min, Lim J, Kwon S, Chun J. A large-scale evaluation of algorithms to calculate average nucleotide identity. Antonie van Leeuwenhoek, Int J Gen Mol Microbiol. 2017;110:1281–6.

169. Biavati B, Mattarelli P, Crociani F. Bifidobacterium saeculare: a New Species Isolated from Feces of Rabbit. Syst Appl Microbiol [Internet]. Gustav Fischer Verlag, Stuttgart · New York; 1991;14:389–92. Available from: http://dx.doi.org/10.1016/S0723-2020(11)80315-2

170. Chevrot R, Carlotti A, Sopena V, Marchand P, Rosenfeld E. Megamonas rupellensis sp. nov., an anaerobe isolated from the caecum of a duck. Int J Syst Evol Microbiol. 2008;58:2921–4.

171. Hansen EE, Lozupone CA, Rey FE, Wu M, Guruge JL, Narra A, et al. Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins. Proc Natl Acad Sci [Internet]. 2011;108:4599–606. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1000071108

172. Dagan T, Artzy-Randrup Y, Martin W. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci [Internet]. 2008;105:10039–44. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0800679105

173. Carr R, Shen-Orr SS, Borenstein E. Reconstructing the Genomic Content of Microbiome Taxa through Shotgun Metagenomic Deconvolution. PLoS Comput Biol. 2013;9.

174. Korem T, Zeevi D, Suez J, Weinberger A, Avnit-Sagi T, Pompan-Lotan M, et al. Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. Science (80-) [Internet]. 2015;349:1101–6. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.aac4812

175. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, et al. Accessible, curated metagenomic data through ExperimentHub. Nat. Methods. 2017.

176. 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;11:119.

177. Kim M, Oh HS, Park SC, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol. 2014;

178. Maronna RA, Martin RD, Yohai VJ. Robust Statistics: Theory and Methods. Ann Stat. 2006;30:17–23.

179. Duff IS. A Survey of Sparse Matrix Research. Proc IEEE. 1977;65:500–35.

180. Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, et al. Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe. 2015;

181. Jeraldo P, Hernandez A, Nielsen HB, Chen X, White BA, Goldenfeld N, et al. Capturing one of the human gut microbiome's most wanted: Reconstructing the genome of a novel butyrate-producing, clostridial scavenger from metagenomic sequence data. Front Microbiol. 2016;

182. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput Biol. 2016;

183. Fodor AA, DeSantis TZ, Wylie KM, Badger JH, Ye Y, Hepburn T, et al. The "most wanted" taxa from the human microbiome for whole genome sequencing. PLoS One. 2012;7.

184. Oberhardt MA, Zarecki R, Gronow S, Lang E, Klenk HP, Gophna U, et al. Harnessing the landscape of microbial culture media to predict new organism-media pairings. Nat Commun. 2015;

Communications scientifiques

Articles

1. **Plaza Oñate F**, Le Chatelier E, Almeida M, Cervino ACL, Gauthier F, Magoulès F, Ehrlich SD, Pichaud M. MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. Bioinformatics. 2018.
2. Farrin W, **Plaza Oñate F**, Plassais J., Bonny C, Beglinger C, Woelnerhanssen B, Nocca D, Magoulès F, Le Chatelier E, Pons N, Cervino ACL, Ehrlich SD. Laparoscopic Roux-en-Y gastric bypass profoundly changes gut microbiota compared to laparoscopic sleeve gastrectomy: a metagenomic comparative analysis. Prépublication disponible sur BioRxiv. 2018.

Présentations orales

1. **Plaza Oñate F**. Abundance-based reconstitution of microbial pan-genomes from whole-metagenome shotgun sequencing data. Conférence Paris Metagenomic Analysis Group, Ecole CentraleSupélec, Gif-Sur-Yvette (France), 2 Février 2018
2. **Plaza Oñate F.** Abundance-based reconstitution of microbial pan-genomes from whole-metagenome shotgun sequencing data. Conférence Recent Computational Advances in Metagenomics (RCAM), Institut Pasteur, Paris (France), 9-10 Octobre 2017

Posters

1. **Plaza Oñate F**, *et al.* Abundance-based reconstitution of microbial pan-genomes from whole-metagenome shotgun sequencing data. Conférence HiTSeq 2017, Prague (République Tchèque), 21-25 Juin 2017.
2. **Plaza Oñate F**, *et al.* Abundance-based reconstitution of microbial pan-genomes from whole-metagenome shotgun sequencing data. 2nd Annual European Microbiome Congress, Londres (Royaume-Uni), 30 Novembre et 1er Décembre 2016
3. **Plaza Oñate F**, *et al.* Abundance-based reconstitution of microbial pan-genomes from whole-metagenome shotgun sequencing data. International Human Microbiota Congress 2016, Texas (Etats-Unis), 9-11 Novembre 2016