Impact du régime alimentaire sur la dynamique structurale et fonctionnelle du microbiote intestinal humain

Julien Tap

THESE DE DOCTORAT DE
L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité

Physiologie et physiopathologie

Présentée par

M. Julien Tap

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Impact du régime alimentaire sur la dynamique structurale et fonctionnelle du microbiote intestinal humain

soutenue le 16 décembre 2009

devant le jury composé de :

M. Philippe LEBARON, Président du jury

Mme Annick BERNALIER, Rapporteur

M. Jean FIORAMONTI, Rapporteur

Mme Marion LECLERC, Examinateur

Mme Karine CLEMENT, Examinateur

Mme Gabrielle POTOCKI-VERONESE, Examinateur

M. Eric PELLETIER, Examinateur

Université Pierre & Marie Curie - Paris 6
Bureau d'accueil, inscription des doctorants et base de données
Esc G, 2ème étage
15 rue de l'école de médecine
75270-PARIS CEDEX 06

Tél. Secrétariat : 01 42 34 68 35
Fax : 01 42 34 68 40
Tél. pour les étudiants de A à EL : 01 42 34 69 54
Tél. pour les étudiants de EM à MON : 01 42 34 68 41
Tél. pour les étudiants de MOO à Z : 01 42 34 68 51
E-mail : scolarite.doctorat@upmc.fr

« Tous pour un ! Un pour tous ! »

Alexandre Dumas, Les Trois Mousquetaires (1844).

Cette thèse a été effectuée sous la direction de Marion Leclerc à l'INRA au sein de l'unité Ecologie et Physiologie du Système Digestif et financée par le projet ANR AlimIntest :

INRA

Unité Ecologie et de Physiologie du Système Digestif

Centre de Recherche de Jouy

Domaine de Vilvert

78352 Jouy-en-Josas cedex

FRANCE

REMERCIEMENTS

Ce travail de thèse n'a été possible que par l'intermédiaire de multiples collaborations et enrichi par différentes personnes, associant plusieurs compétences allant de la modélisation mathématique à la biologie moléculaire, en passant par la bioinformatique. Il ne fut possible que par l'apport, le soutient moral et scientifique de toutes ces personnes.

Je tiens à remercier :

Philippe Lebaron, Annick Bernalier et Jean Fioramonti d'avoir accepté de faire parti de mon jury de thèse et de me faire l'honneur de juger mon travail.

Marion Leclerc, pour sa confiance absolue en mon travail et son soutien actif de tous les instants. Je souhaite à tous les doctorants de l'avoir comme encadrante.

Joël Doré, pour sa confiance et son soutien dans toutes mes initiatives, mais aussi pour m'avoir rapidement mis sur le chemin du noyau phylogénétique dès mon arrivée à l'INRA.

Stanislas Mondot, pour son énergie et notre travail en synergie en bioinformatique et en statistique, et j'espère que nous continuerons à travailler ensemble dans le futur.

Jean-Pierre Furet, pour son apport technique et son soutien moral, mais aussi pour les multiples aventures que nous avons vécues notamment avec la métatranscriptomique.

Ludovic Legrand et Christophe Caron, pour leur bonne humeur, leur réceptivité, leur capacité à dialoguer avec des biologistes et tout simplement pour RapidOTU. (Je n'oublie pas Clément Gauthey qui en a réalisé la première version).

Eric Pelletier, Edgardo Ugarte et Denis Le Paslier, pour tous leurs coups de pouce, toujours au bon moment, leur intervention a toujours été cruciale pour faire avancer cette thèse.

Florence Levenez, pour son soutien technique qui a été à la base de ce travail de thèse.

Patricia Lepage et Karine Gloux, pour toutes nos conversations enrichissantes.

Rafaël Muñoz Tamayo, pour son ouverture d'esprit et sa modélisation mathématique du côlon humain que j'ai pu utiliser dans ce travail de thèse.

Gérard Corthier, pour son apport dans mon travail de thèse qui a été considérable aussi bien sur le plan technique que scientifique, notamment pour la mise en œuvre de la méta-transcriptomique.

Tout le consortium du projet AlimIntest mais aussi toute l'Unité d'Ecologie et Physiologie du Système Digestif, qui m'ont fourni un cadre idéal pour mener à bien ce travail de thèse.

Gabrielle Veronese et Lena Tasse, en particulier, pour notre collaboration notamment sur l'approche métagénomique fonctionnelle.

Catherine Philippe et Sylvie Rabot, pour leur apport technique dans le dosage des acides gras à chaînes courtes.

Eric Fontaine et Hubert Roth, pour leur rôle dans l'étude clinique AlimIntest.

Toute l'équipe de Karine Clément, pour notre collaboration fructueuse et de m'avoir transmis beaucoup de connaissance sur les maladies métaboliques.

Un remerciement particulier à Omar Lakhdari car cela fait maintenant huit ans que nous travaillons ensemble, et bien évidemment Karine Le Roux mais ça elle sait pourquoi…♥

RESUME

Nutritionnistes et gastroentérologues s'accordent pour admettre que le microbiote intestinal conditionne de nombreuses fonctions de l'hôte et joue un rôle important dans le bien-être digestif. Comprendre comment notre régime alimentaire modifie la structure et les fonctions du microbiote intestinal est essentiel afin de connaître en retour son impact sur notre santé.

Le consortium d'espèces du microbiote intestinal de l'Homme adulte apparaît spécifique de l'individu. Ceci constitue un paradoxe dans la mesure où il existe une grande homogénéité fonctionnelle de l'écosystème intestinal quant à sa fonction physiologique de dégradation des fibres alimentaires. A partir d'un inventaire moléculaire basé sur le gène de l'ARNr 16S à partir de 17 volontaires sains ayant des régimes alimentaires variés, il apparaît en effet que la plupart des phylotypes sont spécifiques de l'individu. Néanmoins, 2% des phylotypes sont partagés par plus de 50% des individus représentant 35,5% des séquences obtenues. Ce petit nombre limité de phylotypes constituerait le noyau phylogénétique du microbiote intestinal et son rôle apparaît critique dans le bien-être digestif.

Dans le cadre d'une étude clinique sur volontaires sains, impliquant deux régimes alimentaires contrôlés variant selon la teneur en fibres, une approche microbiomique a montré que la dynamique structurale et fonctionnelle du microbiote pouvait être modifiée en cinq jours. De plus, il a été montré que la structure du microbiote intestinal restait sous influence du régime alimentaire administré au moins 15 jours auparavant.

Ces travaux ouvrent de nouvelles perspectives pour de futures investigations nutritionnelles et épidémiologiques.

Mots clés : Microbiote, Fibres, Microbiomique, Noyau, Alimentation, Santé

ABSTRACT

Nutritionists and gastroenterologists agree to admit that intestinal microbiota determines many host functions and plays an important role in the digestive well being. Understanding how our diet alters the structure and functions of intestinal microbiota is essential to know in turn its impact on our health.

The species consortium of adult's intestinal microbiota appears specific to the individual. This seems a paradox since there are homogeneous physiological functions of the intestinal ecosystem such as dietary fiber degradation. From a molecular inventory, based on 16S rRNA genes from 17 healthy volunteers with different diets, it indeed appeared that most phylotypes were specific of the individual. However, 2% of the phylotypes were shared by more than 50% of individuals and represented 35.5% of the sequences obtained. This small and limited number of phylotypes constitutes an intestinal microbiota phylogenetic core and its role appears critical for digestive well-being.

As part of a clinical study on healthy volunteers, involving two controlled diets varying according to fiber content, a microbiomics approach showed that the structural and functional dynamics of the microbiota could be modified within five days. Moreover, unexpectedly, the intestinal microbiota structure remained under the influence of the diet for at least 15 days after its administration.

These results open new perspectives for future nutrition and epidemiology investigations.

Keywords : Microbiota, Fiber, Microbiomics, Core, Food, Health

Title : Diets impact on structural and functional dynamic of the human intestinal microbiota

TABLE DES MATIERES

LISTE DE PUBLICATIONS

Article 1 : Furet JP, Firmesse O, Gourmelon M, Bridonneau C, **Tap J**, Mondot S, Doré J, Corthier G. Comparative assessment of human and farm animal faecal microbiota using real-time quantitative PCR. *FEMS Microbiol Ecol.* 2009 Jun; **68**(3):351-62. Epub 2009 Mar 19. PubMed PMID: 19302550.

Article 2 : **Tap J**, Legrand L, Gauthey C, Caron C, Doré J, Le Paslier D, Pelletier E, Leclerc M. RapidOTU: 16S rRNA gene sequences clustering into operational taxonomic units using tetranucleotides frequencies. *PLoS Comp Biol.* 2009 Nov, (Submitted, 09-PLCB-RA-1457).

Article 3 : Tap J, Mondot S, Levenez F, Pelletier E, Caron C, Furet JP, Ugarte E, Muñoz-Tamayo R, Le Paslier D, Nalin R, Dore J, Leclerc M. Towards the human intestinal microbiota phylogenetic core. *Environ Microbiol.* 2009 Oct; **11**(10):2574-84. Epub 2009 Jul 6. PubMed PMID: 19601958.

Article 4 : Furet JP, Kong L, **Tap J**, Poitou C, Basdevant A, Bouillot JL, Mariat D, Corthier G, Doré J, Henegar C, Rizkalla S, Clément K. Differential adaptation of human gut microbiota to bariatric surgery-induced weight loss: links with metabolic and low grade inflammation markers. *PLoS Med.* 2009 Nov, (Submitted, 09-PLME-RA-3135)

Article 5 : Krause L, Moine D, Rytz A, Leclerc M, Doré J, **Tap J**, Arigoni F, Berger B. Profiling microbial communities using multiplex pyrosequencing: a validation study. 2009 Nov, (in prep).

Communications/Posters :

Tap J, Mondot S, Levenez F, Pelletier E, Caron C, Furet JP, Ugarte E, Muñoz-Tamayo R, Nalin R, Le Paslier D, Doré J and Leclerc M. Towards the Healthy Human Intestinal Microbiota Phylogenetic Core. Proceedings of the 2009 Conference on Gastrointestinal Function, Chicago, USA, April 20-22. *Microb Ecol.* (2009) **57**:562-588.

Tap J, Mondot S, Levenez F, Pelletier E, Caron C, Furet JP, Ugarte E, Muñoz-Tamayo R, Nalin R, Le Paslier D, Doré J and Leclerc M. Towards the Human Intestinal Microbiota Phylogenetic Core. Gut Microbiome Symposium 2008. 6th congress INRA Rowett Research Institut. Clermont-Ferrand. 2008 June 17-20th.

Legrand L, **Tap J**, Gauthey C, Doré J, Caron C, Leclerc M. RapidOTU: A fast pipeline to analyze 16S rDNA sequences by alignment or tetranucleotide frequency. Gut Microbiome Symposium 2008. 6th congress INRA Rowett Research Institut. Clermont-Ferrand. 2008 June 17-20th.

TABLE DES ILLUSTRATIONS

1 PREALABLE

Nous naissons 99 % eucaryotes et nous mourons 99 % procaryotes. En effet, avant même notre naissance, par l'intermédiaire du cordon ombilical, nous sommes colonisés par des bactéries (Jimenez et al., 2005). Puis, c'est au cours des premiers mois de notre vie que nous acquérons un partenaire singulier : notre microbiote (Mackie et al., 1999). Le microbiote représente l'ensemble des microorganismes peuplant notre organisme. Ces microorganismes sont principalement des bactéries mais nous pouvons également héberger des archées, des champignons et des virus (principalement sous forme de phages). L'essentiel de notre microbiote se répartit tout du long de notre tractus digestif, et est estimé à plus de 100 mille milliards de bactéries, soit dix fois plus que nos propres cellules humaines. Il atteint une densité maximale dans notre côlon distal avec 10^{11} bactéries pour un gramme de contenu (Holdeman et al., 1976; Savage, 1977).

Nous sommes donc vus comme une niche écologique ambulante, ou plutôt préfère-t-on parler de « super-organisme », composé d'un amalgame de cellules microbiennes et d'*Homo sapiens*. Tous les organismes supérieurs tels que les autres mammifères, les insectes et les poissons, ont leur microbiote spécifique. Dans plusieurs cas, l'information génétique combinée des microorganismes constituant le microbiote dépasse de loin celle de leur hôte.

Par ailleurs, l'estimation actuelle du nombre de gènes dans le génome humain est évaluée aux alentours de 23 000 gènes (Wei and Brent, 2006), tandis que pour le métagénome intestinal, l'ensemble combiné des génomes de notre microbiote intestinal, elle s'établit à plus de neuf millions (Yang et al., 2009).

Aussi bien sur le plan de l'abondance structurelle que génétique, le microbiote intestinal, anciennement appelé « flore intestinale », peut être considéré comme un organe à part entière tant son impact est important sur notre vie. En effet, sans lui, nous ne pourrions pas digérer certains composants de notre nourriture (Sonnenburg et al., 2005; Ley et al., 2008), notre système immunitaire serait immature (Mazmanian et al., 2005) et la paroi de notre intestin serait faiblement développée. A mi-chemin entre le mutualisme et le symbiotisme, nous ne pourrions vivre l'un sans l'autre (Hooper and Gordon, 2001; Backhed et al., 2005; Dethlefsen et al., 2007).

Les avancées technologiques récentes ont permis de redéfinir notre vision de cet organe oublié. Nous avons ainsi pu réévaluer l'importance de la biodiversité du microbiote intestinal humain (Suau et al., 1999; Eckburg et al., 2005) ainsi que l'impact fonctionnel sur notre bien-être et notre santé grâce à de nouvelles approches à l'interface entre l'écologie microbienne, génomique et post-génomique (Zoetendal et al., 2008).

Comme l'ensemble de nos organes, le microbiote intestinal est dynamique fonctionnellement et il s'adapte aux différents facteurs environnementaux de l'écosystème intestinal. Parmi ces facteurs abiotiques, notre alimentation joue un rôle majeur et peut modifier directement ou indirectement l'environnement gastro-intestinal. En effet, comme chez les ruminants, la subsistance du microbiote est assurée principalement par les résidus alimentaires, notamment par la fermentation des fibres alimentaires (Flint et al., 2007). Les maladies métaboliques comme l'obésité, où de fait l'alimentation est une problématique importante, ont aussi établi un lien de causalité avec le microbiote (Ley et al., 2005). Plus généralement, ce sont nos pratiques culturelles, nos styles de vie, nos modes alimentaires à l'échelle locale voire mondiale qui entreraient en jeu dans l'interaction entre le microbiote et la santé. C'est pourquoi, il devient essentiel de comprendre comment notre régime alimentaire modifie notre microbiote intestinal afin de connaître en retour son impact sur notre santé.

2 DIVERSITE DU MICROBIOTE INTESTINAL HUMAIN

2.1 Détecter l'incultivable

À partir du milieu des années 1980, Carl Woese a révolutionné le domaine de la microbiologie grâce à des comparaisons phylogénétiques fondées sur les ARN ribosomaux délimitant les trois branches principales de la vie (Woese, 1979, 1987). Aujourd'hui, les analyses fondées sur le séquençage des ARNr[1] restent une méthode de microbiologie, utilisée non seulement pour étudier la diversité microbienne, mais aussi comme une méthode d'identification et de taxonomie moléculaire des bactéries au jour le jour (Amann et al., 1995). Enfin, la définition du phylotype (ou espèce détectée par outils moléculaires) sur la base de séquences codant pour le gène de l'ARNr 16S a été et demeure une norme pour les études de diversité des microorganismes.

En ce qui concerne le microbiote intestinal humain, les études basées sur l'inventaire moléculaire du gène codant pour la sous-unité 16S de l'ARN ribosomal ont montré que plus de 70 % des séquences étaient issues de bactéries non cultivées (Suau et al., 1999; Eckburg et al., 2005). Près des deux tiers étaient spécifiques de chaque individu. Etonnamment, bien que chacun possède un microbiote qui lui est propre, plus de 95% des séquences sont assignées seulement aux Firmicutes, Bacteroidetes, Actinobacteria et Proteobacteria. Les deux premiers phyla se partagent la grande majorité de l'écosystème (Suau et al., 1999; Eckburg et al., 2005).

2.1.1 Les Firmicutes

Le phylum des Firmicutes (bactéries à Gram[2] positif à faible G+C %) est toujours fortement représenté. Il représente en général les trois quarts des espèces détectées par séquençage et la moitié des bactéries du microbiote intestinal. La très grande majorité des espèces des Firmicutes appartient à la classe des Clostridii tandis que moins de 5 % sont membres des classes des Mollicutes et des Bacillii (Eckburg et al., 2005).

La majorité des Clostridii appartient au groupe Clostridiales XIV dit « *Clostridium coccoides* ». Il comprend des espèces bactériennes appartenant aux genres *Eubacterium*, *Butyrovibrio*, *Roseburia*, *Dorea* et *Lachnospira*. Avec d'autres outils moléculaires que le séquençage, ce groupe peut

[1] L'acide ribonucleique ribosomique ou ARNr est le constituant principal du ribosome, organite cellulaire très conservé au sein d'une même espèce. Les ARNr sont eux-mêmes produits à partir de gènes codés dans l'ADN.

[2] La coloration de Gram permet de mettre en évidence les propriétés de la paroi bactérienne, et d'utiliser ces propriétés pour distinguer et classifier les bactéries.

représenter jusqu'à 30 % des bactéries du microbiote intestinal (Sghir et al., 2000; Rigottier-Gois et al., 2003c). Le phylum des Firmicutes comprend également le groupe Clostridiales XV dit « *Clostridium leptum* », avec notamment les espèces *Faecalibacterium prausnitzii*, *Ruminococcus albus* et *R. flavefaciens*, qui dominent quant à elles très largement le microbiote quand on réalise du séquençage. Néanmoins, avec l'utilisation de sondes moléculaires spécifiques, ce groupe ne représenterait en moyenne que 22 % des bactéries du microbiote (Lay et al., 2004). Les autres Clostridii sont membres des groupes III, IV, IX (Acidaminococcaceae), XI, XIII, (Peptostreptococcus), XV, avec aussi d'autres phylotypes inclassables. (Eckburg et al., 2005)

Par ailleurs, si la taxonomie des grandes divisions semble faire consensus (i.e. phylum et classe), le classement des Clostridii en sous-groupes peut prêter à confusion. De plus en plus, les études utilisent le classement par famille pour illustrer la biodiversité du microbiote, où les Lachnospiraceae, Clostridiaceae et Ruminococacae dominent le phylum des Firmicutes (Frank et al., 2007). Ces changements de taxonomie sont liés à l'utilisation du séquençage, qui apporte une résolution plus importante par rapport aux anciennes classifications phénotypiques.

2.1.2 Les Bacteroidetes

Les Bacteroidetes représentent selon les études de 10 % à 40 % du microbiote, avec toutefois un nombre d'espèces détectées plus restreint comparé aux Firmicutes (Suau et al., 1999; Rigottier-Gois et al., 2003c; Eckburg et al., 2005). Les Bacteroidetes sont représentés par les espèces apparentées aux genres *Bacteroides*, *Prevotella* et *Porphyromonas*. Le tiers des séquences assignées au phylum Bacteroidetes est représenté par *Bacteroides vulgatus*. Le phylum des *Bacteroides* est très variable d'un individu à l'autre en termes d'abondance et de répartition des espèces, notamment dans le genre *Prevotella* (Eckburg et al., 2005). Même si par rapport aux Firmicutes, on dénombre moins de *Bacteroides*, il semble que leur activité métabolique soit très importante (Rigottier-Gois et al., 2003b).

2.1.3 Les Actinobacteries

Quelle que soit la méthode utilisée, le phylum Actinobacteria est moins systématiquement détecté en dominance chez les sujets adultes et représente en séquences moins de 1% des bactéries totales (Rigottier-Gois et al., 2003a; Eckburg et al., 2005). On y trouve les bifidobactéries et les bactéries du groupe *Collinsella–Atopobium*. Bien souvent, les espèces détectées forment des singletons, c'est-à-dire des espèces que l'on ne détecte qu'une seule fois par inventaire moléculaire (Eckburg et al., 2005).

2.1.4 Les Proteobacteries

Le phylum Proteobacteria est plus rarement observé dans le microbiote fécal dominant, si bien que l'espèce *Escherichia coli* est rarement détectée en dominance chez les individus (Eckburg et al., 2005).

Les études basées sur le séquençage du gène codant pour la sous-unité 16S du ribosome ont permis de décrire la diversité du microbiote avec une grande finesse. Les critiques de cette méthode sont liées au fait qu'il existe un nombre différent de copies de ce gène en fonction des espèces, et que certaines espèces peuvent être surestimées par rapport à d'autres. De plus, la diversité nucléotidique existant entre les paralogues questionne aussi cette approche. Néanmoins, la plupart des paralogues dans un génome ont une diversité inférieure à 1% (un nucléotide différent pour cent nucléotides comparés), ce qui rend possible l'utilisation d'un seuil de 2% pour séparer les espèces entre elles (Acinas et al., 2004). D'autre part, parmi les groupes dominants du microbiote intestinal, le nombre de paralogues par espèce est compris entre quatre et sept copies (4,14 en moyenne pour les Bacteroidetes et 6,3 copies en moyenne pour les Firmicutes d'après la base de données *rrnDB*[3]), ce qui conduirait à une surestimation des Firmicutes (Lee et al., 2009).

2.2 Homéostasie et dynamisme du microbiote

En plus de ces études instantanées chez l'adulte sain, il est nécessaire d'analyser la dynamique du microbiote sur le long terme pour comprendre les mécanismes qui entrent en jeu dans l'homéostasie intestinale. En outre, la biodiversité du microbiote diffère selon les individus, ce qui suggère des déplacements de l'équilibre implantation/déclin au cours du temps.

2.2.1 A l'échelle d'une vie

La composition du microbiote en dominance est d'une remarquable stabilité au cours d'une vie (Zoetendal et al., 1998; Matsuki et al., 2004). Zoedental et ses collègues ont montré avec des études électrophorétiques que les profils de migration n'ont pas changé sur une période de six mois. L'établissement du microbiote est un processus dynamique en plusieurs phases qui permet, si elles sont réalisées, l'émergence d'un microbiote stable contribuant à un système immunitaire pleinement fonctionnel.

[3] rrnDB : http://ribosome.mmg.msu.edu/rrndb/

Une étude a également montré, chez la souris, l'existence d'un possible passage de bactéries de la mère à sa progéniture *in utero*. Les bactéries identifiées dans le sang du cordon ombilical appartenaient aux genres *Enterococcus*, *Streptococcus* et *Staphylococcus* (Jimenez et al., 2005).

A compter de la naissance, l'implantation du microbiote chez les nouveaux-nés va s'effectuer très rapidement. La population « source »[4] du microbiote fécal, c'est-à-dire celle qui s'implante en premier, est composée principalement de bactéries anaérobies facultatives comme des entérobactéries, des bifidobactéries et des lactobacilles (Favier et al., 2002). Par ailleurs, des composants bactériens, voire des bactéries viables (*Bifidobacterium*) transitant par l'intermédiaire du lait maternel, permettraient d'éduquer le système immunitaire du bébé (Perez et al., 2007). Comparés à des enfants ayant eu du lait infantile, les enfants nourris au lait maternel auront une implantation tardive de *Clostridium* et de *Bacteroides* (Penders et al., 2006). Cette différence peut être expliquée par la présence de caséine, lactoferrine et défensine dans le lait maternel, mais également de substrats présents pour les bactéries.

Le mode de naissance, par voie vaginale ou par césarienne, peut impacter significativement la composition du microbiote intestinal du bébé. Par voie naturelle, le nourrisson est exposé d'abord au microbiote vaginal de la mère, tandis que par césarienne le nourrisson est exposé en premier lieu à l'air de son environnement. Chez ces derniers, ceci aura notamment pour conséquence une implantation plus tardive des espèces apparentées au genre *Bacteroides*.

Néanmoins, dans tous les cas, dès la diversification alimentaire, les deux principaux phyla Bacteroidetes et Firmicutes surpassent en nombre ainsi qu'en diversité les Actinobacteria et les Proteobacteria implantés précédemment. Les études divergent sur le moment (d'une à quatre années) où le microbiote intestinal du nourrisson peut être considéré comme celui de l'adulte (Tannock, 2007). La mise en place du microbiote s'accompagne aussi de changements métaboliques. En effet, alors que les capacités fermentaires du microbiote conduisent à une production de lactate et d'acétate pendant les premiers mois de la vie, les concentrations de butyrate et de propionate deviennent dominantes et stables dès la deuxième année de la vie.

Quel que soit le processus d'implantation du microbiote, une homéostasie s'installe, mais d'autres études seront nécessaires pour connaître les effets de ce processus sur le long terme dans l'éducation du système immunitaire. C'est peut-être cette fenêtre particulièrement « ouverte », au moment de l'implantation du microbiote, qui offre une opportunité de prévenir des maladies immunitaires (Ley et al., 2006a). Même si la composition du microbiote varie entre les individus,

[4] Population source : population pionnière dans un milieu donné et en pleine expansion

les populations bactériennes dominantes restent relativement stables chez l'adulte sain (Zoetendal et al., 1998).

Alors qu'un nombre important d'études a été effectué sur le microbiote intestinal des bébés et des adultes, les effets du vieillissement sur le microbiote sont mal caractérisés. La population « puits »[5] des bifidobactéries décline chez les personnes âgées au profit des entérobactéries et des clostridii (van Tongeren et al., 2005; Woodmansey, 2007). Parallèlement, la diversité des bifidobacteries décroît et se limite à deux espèces : *Bifidobacterium longum* et *Bifidobacterium adolescentis*. Cette chute de bifidobactéries peut avoir des conséquences sur la santé des personnes âgées tant les bifidobactéries sont impliquées dans le métabolisme du microbiote et la stimulation du système immunitaire.

De plus, une baisse des *Bacteroides* a également été montrée chez des personnes âgées, contribuant à des changements significatifs dans le ratio Firmicutes/Bacteroidetes (Mariat et al., 2009). Les *Bacteroides* possédant des facultés à dégrader les polyosides et à produire des acides gras à chaîne courte (AGCC), leur chute peut impacter la digestion et la capture d'énergie. Les changements de composition du microbiote peuvent être dus à une altération partielle du tractus intestinal et peuvent être à l'origine de la malnutrition des personnes âgées (Guigoz et al., 2008).

2.2.2 A l'échelle de l'évolution

Alors qu'il existe plus d'une cinquantaine de phyla dans le monde bactérien (Handelsman, 2004), comparée au métagénome du sol et des océans, la dominance de quatre phyla chez tous les individus suppose que de fortes contraintes entrent en jeu dans le façonnage du microbiote intestinal. De plus, les espèces observées ont le plus souvent une spécificité humaine, et dans tous les cas, elles sont associées à l'environnement digestif de façon quasi exclusive. Cela indique des phénomènes de coévolution avec l'hôte (Ley et al., 2006a).

D'un autre point de vue, lorsque l'on regarde ces phénomènes de coévolution à l'échelle d'une vie ou de deux générations, les études basées sur le génotype de l'hôte et la transmission verticale du microbiote des parents aux descendants représentent un facteur de confusion. Une étude basée sur des empreintes ADN du microbiote intestinal montre que les jumeaux ont un microbiote plus similaire entre eux que leurs conjoints respectifs (Zoetendal et al., 2001b). Les similitudes observées entre les communautés intestinales des jumeaux monozygotes peuvent être interprétées comme un effet du génotype sur la diversité bactérienne. En réalité, à ce niveau d'observation, une

[5] Population puits : population en déclin suite à la colonisation du milieu par d'autres espèces

autre explication tient au fait que ces similitudes sont dues à la colonisation par une mère partagée. Ainsi, lorsque l'on regarde les microbiotes des jumeaux dizygotes comparés à des jumeaux monozygotes, ils se ressemblent tout autant (Ley et al., 2006a; Turnbaugh et al., 2009). Par ailleurs, l'utilisation de souris axéniques ayant des génotypes différents a permis de montrer qu'il n'y avait pas de différence dans l'expression transcriptomique de *Bacteroides thetaiotaomicron* (Sonnenburg et al., 2006).

L'observation de phyla majeurs du microbiote intestinal nous renseigne en fait sur la mise en place lointaine, du fait des mécanismes de mutations/sélections, de capacités fonctionnelles à coloniser un écosystème anaérobie, soumis à des pressions chimiques comme les sels biliaires, et physiques tel que le péristaltisme par exemple. Autrement dit : coloniser « un intestin » en caricaturant, qu'il soit humain ou de mammifère monogastrique. C'est pour cela que l'on retrouve chez tous les mammifères, en proportions variables, les deux principaux phyla que sont les Bacteroidetes et les Firmicutes, et seulement ces deux-là, comparés à toute la diversité des microorganismes de la planète (Ley et al., 2008).

D'autre part, ces contraintes, du point de vue de la coévolution, forment des forces de convergence entraînant la radiation de quelques phylotypes dominants (Ley et al., 2006b), ces derniers formant un arbre phylogénétique semblable à un bambou (Yang et al., 2009). Ces forces écologiques et d'évolution sont longitudinales et s'opposent à d'autres forces « latérales » qui provoquent le buissonnement de l'arbre phylogénétique. En effet, un contraste est observé entre la grande diversité de souches et d'espèces détectées, au regard de seulement quelques grands groupes bactériens. Cette évolution buissonnante témoigne de la présence de genres et d'espèces qui coexistent. Cette coexistence peut s'expliquer par l'intermédiaire des chaînes trophiques, mais aussi par la présence d'échanges génétiques entre les taxons.

Par ailleurs, ce schéma mêlant variations génétiques élevées au niveau de la souche et lignées profondes, a également été observé dans le microbiote intestinal murin (Ley et al., 2005). Peu profondes, ces larges radiations sont le résultat d'une pression de sélection extrême suivie d'une détente (Figure 1). De même, l'architecture phylogénétique de l'intestin pourrait avoir résulté de la diversification d'une communauté initiale limitée en souches, issue par exemple d'un goulot d'étranglement. En outre, la faible profondeur phylogénétique de la communauté intestinale peut être due à la récente existence d'un habitat que constituerait l'intestin des mammifères (Dethlefsen et al., 2007).

Cette architecture phylogénétique peut être la signature de la fonctionnalité de l'écosystème intestinal. Ainsi, cela laisse penser qu'il existe sur le plan fonctionnel une interchangeabilité entre

espèces avec une structure en guilde[6] (Tschop et al., 2009). Ces guildes partageraient au sein de l'écosystème intestinal une niche écologique commune afin d'y remplir les même fonctions requises par l'hôte. Par ailleurs, ces structures en guilde peuvent être le résultat de la concurrence entre les phylotypes faisant partie d'un même « buisson ». C'est cette forme d'architecture, que l'on pourrait qualifier d'eubiose, qui permettrait d'assurer l'homéostasie de l'écosystème intestinal.

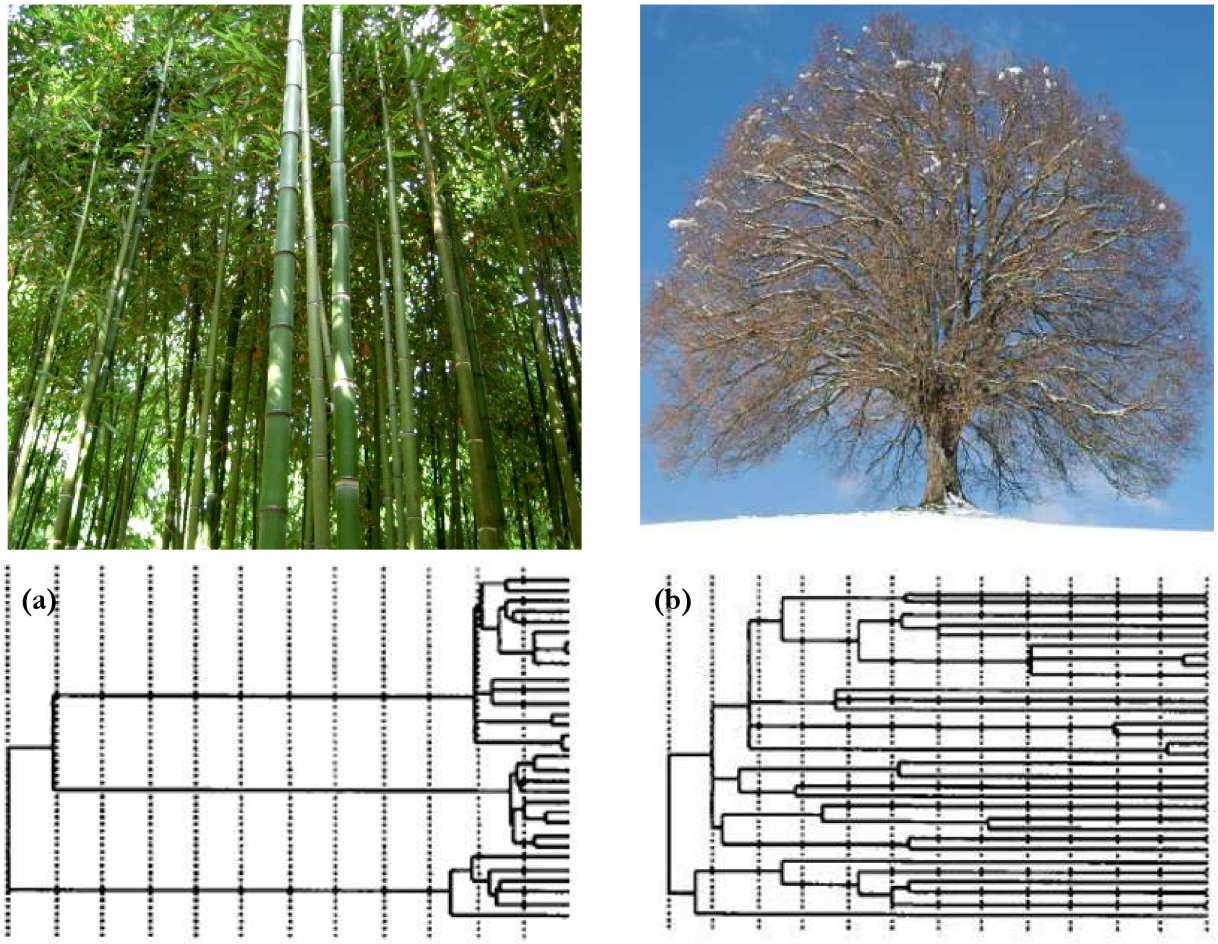


Figure 1 : Conséquence des forces de l'évolution sur la topologie des arbres phylogénétiques. Lorsque que l'on représente les inventaires moléculaires basés sur la séquence de l'ARNr 16S par un dendrogramme, **(a)** la diversité microbienne associée aux mammifères ressemble à la structure d'un bambou avec des lignées profondes suivies d'un accroissement exponentiel de phylotypes génétiquement proches illustrés avec plusieurs feuilles en haut de l'arbre. Cela témoigne de mécanismes récents de balayages sélectifs suivis d'une détente. Ceci s'oppose à un taux constant de renouvellement et d'extinction **(b)**, où la diversité microbienne, associée par exemple à des écosystèmes marins ou de sols, ressemble à un arbre où beaucoup de lignées partent de la racine (Martin, 2002; Dethlefsen et al., 2007).

[6] Guilde : Ensemble d'espèce qui exploitent, d'une façon comparable, la même catégorie de ressources dans un écosystème et appartenant au même groupe taxonomique ou étant apparentées.

2.3 Altération du microbiote

Chaque cellule microbienne est sous une extrême pression de sélection dans l'intestin. Cette pression de sélection permet de fixer des fonctionnalités critiques pour l'hôte, comme l'extraction d'énergie à partir des polyosides ou bien la protection contre les pathogènes. Ces fonctionnalités sont redondantes et sont liées à la robustesse de l'eubiose intestinale.

Par antonymie à l'eubiose, une dysbiose de l'écosystème intestinal serait non seulement associée à des désordres intestinaux mais aussi à des maladies telles que l'obésité (Ley et al., 2005; Turnbaugh et al., 2006), les maladies inflammatoires chroniques intestinales comme la maladie de Crohn (Swidsinski et al., 2002; Manichanh et al., 2006; Frank et al., 2007; Vasquez et al., 2007), les allergies (Macdonald and Monteleone, 2005; Penders et al., 2007b; Penders et al., 2007a) et le cancer colo-rectal (Moore, 1995 ; McGarr, 2005).

D'une manière générale, un écosystème fragilisé par un changement fonctionnel est en dysbiose. Au niveau de l'écosystème intestinal, cette dysbiose peut être expliquée par plusieurs points. Tout d'abord, la nécessité de la présence d'espèces « clé de voûte » pour maintenir le système stable et expliquer cette redondance fonctionnelle partagée par tous les individus. Lorsque l'une serait balayée par des facteurs exogènes comme des substrats issus de l'alimentation ou bien par des facteurs endogènes comme un système immunitaire défaillant, l'écosystème en serait durablement perturbé, entraînant ainsi des maladies. Par opposition à cette hypothèse, l'équipe de J. L. Gordon pense que l'existence d'une redondance fonctionnelle même dissipe la nécessité de telles espèces clés (Ley et al., 2006a; Turnbaugh et al., 2009). En effet, l'absence de l'une d'elles rendrait le système trop fragile et sensible à l'environnement extérieur. D'autre part, la présence facultative de telles espèces s'expliquerait par la présence partagée de familles de gènes dans chacun des génomes des bactéries intestinales.

La dysbiose peut également être expliquée par ce qui constitue la première ligne de dialogue avec notre microbiote, c'est-à-dire le système immunitaire. Le système immunitaire est le premier outil de sélection directe par l'hôte. Bien que le microbiote soit impliqué dans des maladies inflammatoires, aucune espèce seule n'a été jugée totalement responsable. En fait, plusieurs observations démontrent que le système immunitaire répond à un large éventail de marqueurs bactériens. Une étude sur le transcriptome murin démontre que ce sont les gènes impliqués dans le système immunitaire qui sont les plus régulés en présence du microbiote (Mutch et al., 2004). De plus, il a été démontré que *Bateroides fragilis* était capable, par l'intermédiaire de ses polyosides capsulaires, de stimuler une large variété de lymphocytes T (Mazmanian et al., 2005).

La dysbiose peut être aussi caractérisée par un bouleversement complet de l'écosystème ou une mauvaise combinaison de l'abondance et de la diversité d'un groupe bactérien vis-à-vis d'un autre. En effet, chez les patients atteints de la maladie de Crohn, une étude a montré que le groupe « *Clostridium leptum* » était fortement réduit, aussi bien en diversité qu'en abondance (Manichanh et al., 2006). Dans un autre contexte, l'augmentation des *Bacteroides* et la chute des Firmicutes s'accompagneraient d'une faculté du microbiote à stocker plus facilement l'énergie apportée par l'alimentation, ce qui constituerait un facteur de risque pour l'obésité (Backhed et al., 2004; Ley et al., 2006b).

Jusqu'à maintenant, bien que la dysbiose relève d'un changement fonctionnel de l'écosystème, les études sur le microbiote ont constaté cette dysbiose du seul point de vue phylogénétique. Les fonctions d'un écosystème n'étant pas liées spécifiquement aux espèces, il est nécessaire de réaliser des études fonctionnelles de l'écosystème. Bien qu'il soit difficile de définir le sens de la causalité, il est aussi nécessaire d'étudier la dynamique du microbiote pour réaliser des approches métagénomiques intégrées. L'objectif serait dès lors de refaçonner le microbiote avec par exemple une alimentation contrôlée.

3 LES APPROCHES METAGENOMIQUE ET POST-METAGENOMIQUE

3.1 Génome, métagénome et communauté bactérienne

Un génome est la totalité de l'information génétique d'un organisme unique que l'on peut représenter comme une population statistique de gènes. Entre autres, le génome permet aussi de définir une liste de protéines. Comparé à l'ensemble des génomes d'une communauté microbienne, un génome est relativement statique, ce qui rend possible la mise en œuvre d'études comparatives post-génomiques comme la transcriptomique et la protéomique. Une liste de protéines ou d'ARN messagers peut définir un organisme. La transcriptomique et la protéomique permettent d'avoir un point de vue très lié au potentiel fonctionnel d'un organisme.

Un métagénome est la totalité de l'information génétique d'une communauté d'organismes (Handelsman, 2004). Néanmoins, dans le cadre d'une étude d'un écosystème complexe, on ne peut avoir accès à la totalité d'un métagénome. Par conséquent, contrairement à la séquence d'un génome entier, des séquences issues d'une analyse métagénomique ne fournissent pas une population statistique de gènes mais seulement un échantillon. Du fait de la dynamique et de la variation d'une communauté microbienne, il est difficile de mettre en place un référentiel absolu qui permettrait la mise en application d'études comparatives semblables à la post-génomique. Pour l'instant, les études actuelles se limitent à traiter les séquences issues de métagénomique comme une population, avec des outils développés pour la post-génomique (puces à ADN ou interrogation de bases de données issues de la génomique). Les études post-métagénomiques, appelées également « microbiomique »[7], imposent dès lors de nouvelles contraintes qu'il est nécessaire de surmonter.

Du point de vue métagénome, une communauté microbienne peut être définie comme une liste d'organismes et plusieurs stratégies peuvent en découler, comme par exemple la comparaison de communautés. Par ailleurs, la comparaison de communautés est encore effectuée en comparant les séquences du gène ARNr 16S. Selon Schloss (Schloss et al., 2004; Schloss and Handelsman, 2008), il serait utile de s'inspirer de toute cette expérience développée en terme de techniques statistiques et d'intégration des données pour étudier et comparer des échantillons de métagénomes. En outre, en plus de considérer une communauté du point de vue de ses

[7] La microbiomique est un néologisme de plus en plus utilisé pour qualifier cette nouvelle science qui utilise les moyens modernes de la biologie moléculaire visant l'étude d'une communauté microbienne avec comme objectifs de la caractériser et d'évaluer ses fonctions et ses impacts sur son environnement.

organismes, les analyses centrées sur les gènes considèrent une communauté comme une liste de gènes. Les gènes que l'on trouve plus fréquemment dans une communauté sont supposés conférer une fonction bénéfique sur cette communauté (Tringe et al., 2005). La différence entre ces analyses est que les séquences de gènes codant pour l'ARNr 16S sont fonction de la phylogénie tandis que les gènes peuvent être reliés en fonction des voies métaboliques dans lesquelles ils sont impliqués.

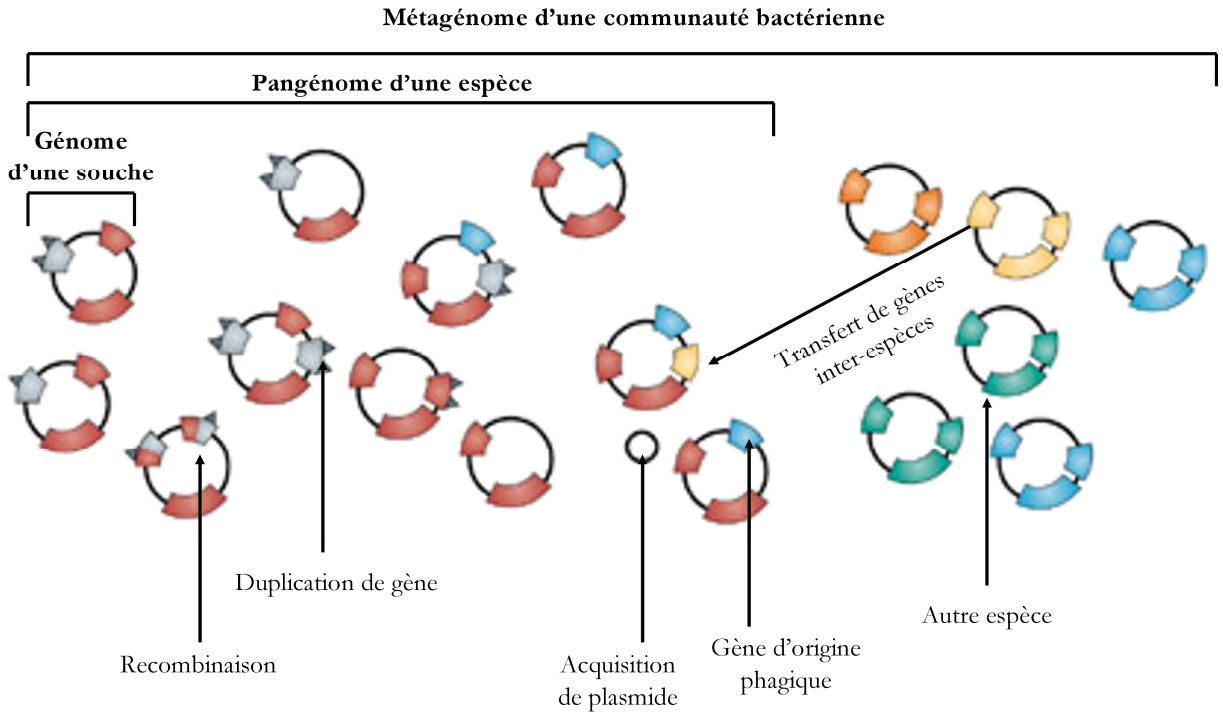


Figure 2 : Dynamique génomique d'une communauté bactérienne. L'écosystème ainsi que la communauté de cet écosystème influent sur la diversité génomique d'une espèce bactérienne. Pour cette raison, un génome d'une souche ne peut pas représenter la diversité pan-génomique d'une espèce bactérienne. Cela explique que des communautés ayant des profils d'organismes similaires ne possèdent pas le même potentiel fonctionnel (Medini et al., 2008).

Avec le séquençage haut débit, une communauté peut être caractérisée par une liste de protéines potentiellement présentes et/ou de gènes transcrits, donnant un aperçu du potentiel fonctionnel de cette communauté. Deux communautés ayant des profils d'organismes similaires peuvent avoir différents potentiels fonctionnels. Par opposition, deux communautés avec le même inventaire de protéines peuvent être très différentes au niveau des organismes. En réalité, les génomes microbiens sont dynamiques et de nombreux mécanismes d'échange d'ADN impactent leur contenu génétique (Figure 2). Chaque espèce ne peut être représentée par un seul génome tant son pan-génome peut être influencé par la pression du microbiome. En effet, le pan-génome décrit la gamme complète de gènes dans une espèce. Il s'agit de l'ensemble de tous les gènes de toutes les souches d'une espèce. Il comprend le génome indispensable à l'espèce, contenu dans

toutes les souches, et le génome « dispensable » spécifique de quelques souches (Medini et al., 2005). Ce dernier est très dynamique et est soumis à des mécanismes tels que la recombinaison, la duplication de gènes et l'acquisition de gènes par transferts latéraux inter-espèces. L'importance du pan-génome se pose dans un contexte évolutif, en particulier en rapport avec la métagénomique. Dès lors, en plus de la génomique et de l'écologie microbienne, la métagénomique doit aussi intégrer la pan-génomique.

3.2 La métagénomique descriptive et intégrative.

La métagénomique est une méthode qui résulte de l'association de l'écologie des communautés et de la génomique. Cela se traduit par l'étude du matériel génétique collecté directement à partir d'échantillons environnementaux (Handelsman, 2004). Alors que la microbiologie traditionnelle et le séquençage de génomes microbiens s'appuient sur des cultures clonales cultivées, la métagénomique permet d'accéder aux organismes difficiles à isoler et à cultiver. Néanmoins, les communautés sont si complexes dans le microbiome[8] intestinal qu'elles ne peuvent qu'être échantillonnées et donc jamais complètement caractérisées. Pouvoir caractériser la biodiversité et le fonctionnement d'une communauté microbienne dépend en grande partie du plan et de l'analyse de l'expérience (Voir la partie « Techniques et méthodes d'analyse », page 43).

Les premières analyses métagénomiques se sont focalisées sur la variété de nouvelles espèces et la communauté formée par celles-ci (Gill et al., 2006). La métagénomique descriptive fournit une vue relativement non biaisée non seulement de la structure d'une communauté, avec son abondance et sa distribution d'espèces, mais aussi de ses fonctions métaboliques potentielles. Par la suite, la métagénomique est devenue « intégrative » en cherchant à identifier un changement fonctionnel microbien en fonction d'un changement de l'environnement (Kurokawa et al., 2007).

L'écologie microbienne se concentre sur les interactions entre les microorganismes et leurs hôtes eucaryotes, sur la compétition et la communication entre microorganismes et sur l'acquisition des substances nutritives, ainsi que sur la production d'énergie (Hugenholtz and Tyson, 2008). Au niveau du tractus gastro-intestinal, l'objectif majeur est d'observer comment les changements fonctionnels impactent la santé humaine.

Par ailleurs, il a été montré que le potentiel fonctionnel d'un microbiote était fonction de son environnement (Tringe et al., 2005). Cependant, il est encore difficile de relier des conditions environnementales distinctes avec des processus biologiques spécifiques. Ainsi, le défi majeur

[8] Microbiome : définit l'habitat , l'aire de vie du microbiote.

consiste à savoir comment l'utilisation de réseaux métaboliques spécifiques reflète l'adaptation de communautés microbiennes à travers des environnements et des habitats (Gianoulis et al., 2009). De plus, l'assignation phylogénétique d'une séquence, qui est importante en vue de relier la fonction à une espèce, demeure très complexe. Par ailleurs, la composition phylogénétique détectée est impactée par la stratégie d'échantillonnage, et la composition fonctionnelle observée dépend du nombre et de la longueur des séquences obtenues (Voir la partie « Séquençage haut débit », page 47).

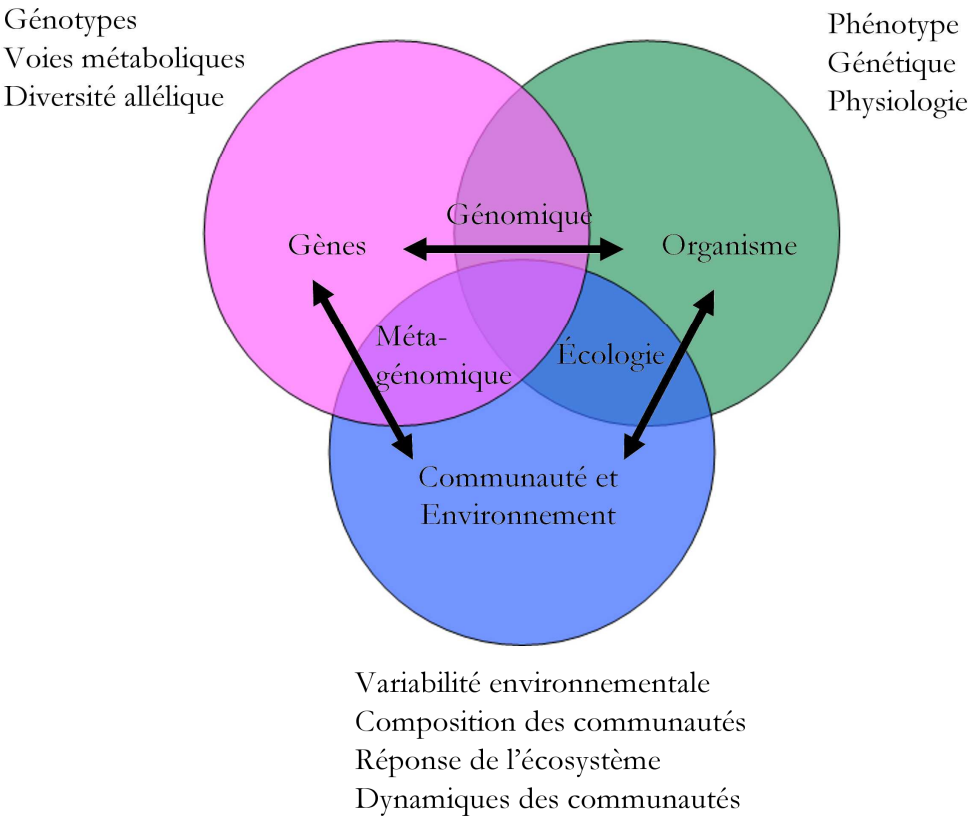


Figure 3 : Intersection de la génomique, de l'écologie et de la métagénomique. Chaque discipline fait le lien entre chaque grande aire d'étude (les gènes, l'organisme et sa communauté). Un effort supplémentaire sera nécessaire pour réaliser la synthèse totale des trois disciplines. (DeLong, 2009).

Malgré ces difficultés, près des trois quarts d'un métagénome peuvent être assignés à une fonction grâce aux stratégies de comparaison sur des bases de référence, et une majorité de gènes peut être assignée à un groupe phylogénétique grâce aux nombreux programmes de séquençage de souches bactériennes cultivées. Après avoir défini cette liste de microorganismes et de fonctions, les outils de bioinformatique devront standardiser l'information obtenue pour réaliser des comparaisons avec d'autres métagénomes (Raes et al., 2007; Field et al., 2008). La standardisation des données participera à l'intégration de l'écologie, la génomique et la métagénomique (Figure 3).

Néanmoins, l'écart entre les protéines bien caractérisées et les protéines détectées dans les métagénomes se creuse à un rythme alarmant. En parallèle des ressources informatiques dont les besoins augmentent exponentiellement, l'accumulation de gènes non caractérisés est susceptible d'être le principal goulet d'étranglement à l'avenir. Cela signifie que notre compréhension des écosystèmes microbiens sera partielle et basée au mieux sur ce que nous pouvons déduire de nos connaissances actuelles de la biochimie (Hugenholtz and Tyson, 2008). Les futures perspectives de la métagénomique seront peut-être la prédiction de changements fonctionnels et structuraux. Après l'intégration : la prédiction?

3.3 Les fonctions du microbiote intestinal révélées par la métagénomique

L'équipe de J. L. Gordon, qui a obtenu près de 78 mégabases (Mb) de séquences métagénomiques des microbiotes intestinaux de deux adultes sains, a comparé l'ensemble des gènes annotés de ces microbiotes intestinaux avec les gènes humains. Cette étude a permis d'identifier un nombre important de gènes bactériens qui ne sont pas codés dans le génome humain (Gill et al., 2006).

Les fonctions codées par ces gènes contribuent largement au métabolisme des glycanes, des acides aminés, des xénobiotiques, et à la biosynthèse des vitamines et des isoprénoïdes, processus indispensables à l'homme. Ces résultats révèlent une relation symbiotique entre le microbiote intestinal et son hôte, appuyant le concept du « super-organisme » et la théorie de l'hologénome[9] (Zilber-Rosenberg and Rosenberg, 2008).

Ensuite, l'équipe de Kurokawa a analysé 13 microbiotes intestinaux comprenant cette fois-ci des adultes, des enfants et des nourrissons non sevrés. Cette étude a permis d'obtenir 479 Mb de séquences métagénomiques (Kurokawa et al., 2007). Etonnamment, plus de la moitié (jusqu'à 90%) des séquences métagénomiques ont été assemblées pour former de longs fragments de séquences dans chaque échantillon, ce qui contraste fortement avec le microbiote du sol dans lequel seulement 1% des séquences a pu être assemblé (Rondon et al., 2000; Tringe et al., 2005).

Ces résultats suggèrent qu'avec environ 50 Mb de données de séquençage en méthode Sanger pour chaque échantillon, on pourrait couvrir à la fois les fonctions et les espèces les plus redondantes du microbiote intestinal. Si l'on considère qu'un génome bactérien possède une taille

[9] L'hologénome est défini comme la somme des informations génétiques de l'hôte et de son microbiote. La théorie de l'hologénome repose sur le principe que l'hôte doit établir des relations symbiotiques avec son microbiote, que le microbiote doit être transmis entre les générations, et que l'association entre l'hôte et son symbiote détermine son adaptation avec son environnement.

moyenne de quatre Mb, alors on aurait l'équivalent métagénome d'une dizaine d'espèces (Kurokawa et al., 2007). Par conséquent, afin d'étudier des fonctions moins représentées et des espèces moins abondantes, il faudra produire un nombre de séquences d'un ordre de grandeur plus important (Figure 4).

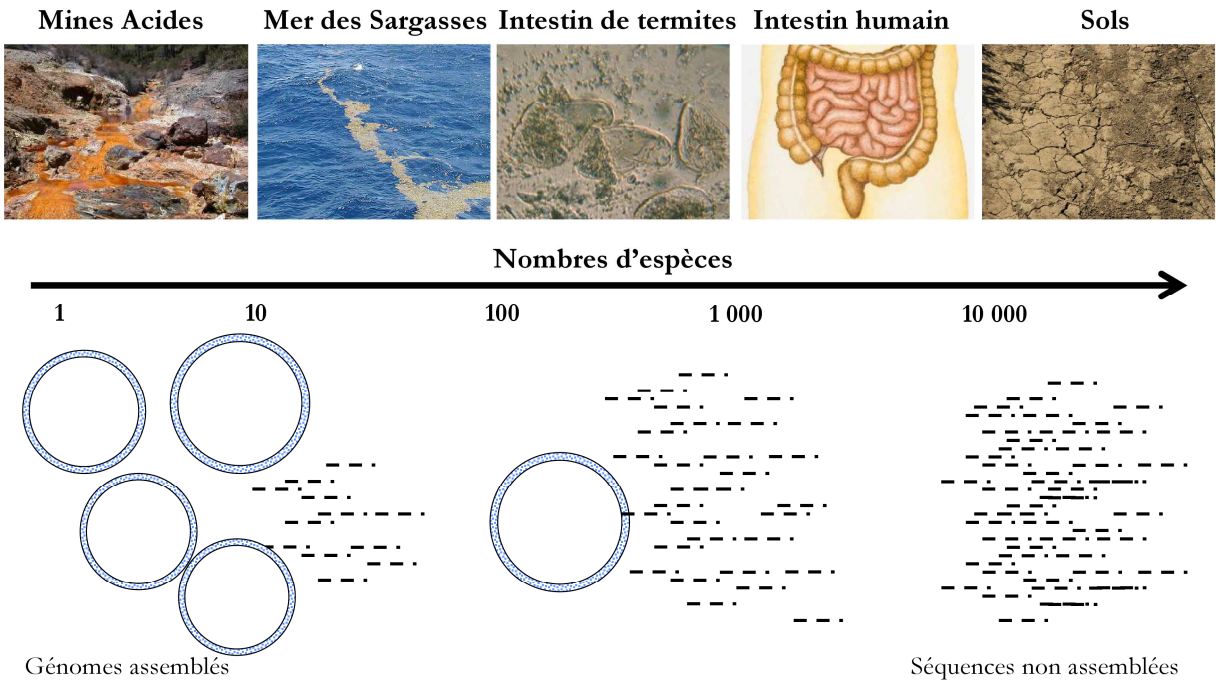


Figure 4 : métagénomique et complexité de l'assemblage en fonction de l'environnement. Divers habitats (microbiomes) ont été étudiés jusqu'à présent. Chaque microbiome possède une diversité et une complexité d'espèces différentes. Plus cette complexité est croissante plus il est difficile d'assembler des génomes entiers. Les efforts d'échantillonnage peuvent différer d'un facteur dix d'un écosystème à un autre. Avec plus de 1000 espèces par individu le microbiote intestinal humain est un écosystème complexe, où il est difficile d'assembler de grands fragments génomiques.

Cette étude a également trouvé 647 familles de gènes spécifiquement enrichies dans le microbiote intestinal, en comparaison avec des gènes présents dans les données métagénomiques d'autres échantillons environnementaux, comme la surface de la mer, la mer profonde et le sol. Ce résultat fut confirmé par une autre méta-analyse des données basée sur les voies métaboliques (Turnbaugh et al., 2007).

Ces gènes ont été assignés respectivement à 237 et 136 groupes de gènes orthologues (COG) pour les microbiotes d'adultes et de nourrissons et partagent 58 COGs pour un total de 315 COGs. Dans les 315 COGs, les fonctions associées aux métabolismes glucidiques sont particulièrement

enrichies, mais les répertoires fonctionnels diffèrent nettement entre les adultes et les nourrissons non sevrés. Le microbiote des adultes est riche en enzymes dégradant les polyosides tandis que celui des nourrissons est riche en transporteurs de sucre. Ces données indiquent que la fonctionnalité de l'écologie microbienne intestinale chez un hôte sain repose largement sur les éléments nutritifs disponibles dans l'alimentation. Par ailleurs, comme attendu, chez les adultes les séquences obtenues sont assignées aux *Bacteroides*, tandis que celles obtenues chez les nourrissons sont assignées aux Bifides et Lactobacilles (Kurokawa et al., 2007).

Par la suite, l'équipe de J.L. Gordon a réalisé l'analyse d'échantillons provenant de 154 individus, conduisant à près de deux millions de séquences codant pour l'ARNr 16S et plus de deux Gigabases (Gb) de métagénome intestinal. Parmi ces individus se trouvaient des jumeaux monozygotes et dizygotes, discordants ou concordants pour l'obésité, ainsi que leur mère. Les résultats révèlent que le microbiote intestinal humain est partagé par les membres d'une même famille, mais que chaque communauté microbienne fluctue en fonction des lignées bactériennes avec un degré de variation comparable entre jumeaux monozygotes et jumeaux dizygotes. Cependant, un large éventail de gènes microbiens est partagé entre les échantillons des individus, comprenant un noyau fonctionnel très étendu. L'obésité est associée à des changements au niveau du phylum, à une réduction au niveau de la diversité bactérienne, et à une altération dans la représentation de certains gènes et de certaines voies métaboliques. La majorité des gènes surreprésentés chez les obèses est assignée aux Actinobactéries (75%) et aux Firmicutes (25%), tandis que chez les sujets sains les gènes prédominants sont assignés aux Bacteroidetes. Ce noyau fonctionnel constitué de gènes partagés par tous les individus sains serait altéré dans le cadre de pathologies comme l'obésité (Turnbaugh et al., 2009).

Ce noyau fonctionnel serait constitué essentiellement de gènes liés à des fonctions métaboliques comme par exemple le métabolisme des hydrates de carbone, des glycanes et des acides aminés. Les gènes les plus variables se retrouvent dans les voies impliquées dans la signalisation et le transport membranaire. Ainsi, les fonctions partagées par tous les individus seraient liées à l'alimentation, tandis que les fonctions de dialogue membranaire seraient plus spécifiques de l'individu.

4 NUTRITION, MICROBIOTE ET SANTE

4.1 Influence du régime alimentaire

Le régime alimentaire est un sujet d'intérêt très important dans les programmes de recherche internationaux en raison de son potentiel de modulation du microbiote intestinal de l'hôte, qu'il soit bénéfique ou néfaste. Les habitudes alimentaires ont un impact important sur la composition du microbiote intestinal, notamment dans les premières années de la vie. Par exemple, la composition du microbiote fécal diffère entre les enfants ayant eu une alimentation par allaitement maternel et ceux ayant eu du lait infantile[10], avec notamment plus de bactéries lactiques et de bifidobactéries chez les bébés allaités.

Tableau 1 : Substrats fermentescibles arrivant dans le côlon (Egert et al., 2006)

Substrats	Composante	apport (g/jour)
Glucides	Amidon résistant	5 – 35
	Polyosides non-digestibles	10 – 25
	Oligosaccharides (i.e. fructo-oligosaccharides, inuline)	2 – 8
	Monosaccharides (i.e.. sucres, alcool)	2 – 5
	Mucines	3 – 5
Protéines	Provenant de l'alimentation	1 – 12
	Origine endogène (i.e. enzymes pancréatiques et autres sécrétions)	4 – 8
	Cellules épithéliales desquamées	30 – 50
Autres	Urée, nitrate	~ 0,5
	Acides organiques, lipides, composés bactériens	inconnu

Cependant, lorsque des régimes plus complexes sont comparés par des approches de culture *in vitro*, comme par exemple le régime à l'occidentale dit « western diet », plus riche en graisses, et le régime à l'orientale plus riche en fibres, peu de genres bactériens du microbiote intestinal

[10] Lait infantile : Lait reconstitué, lait industriel, en anglais « formula feds »

varient. De même, seules quelques différences sont observées entre la composition bactérienne de sujets omnivores et celle de végétariens (Aries et al, 1971).

D'autres études au niveau du côlon distal ont montré des profils de production d'AGCC différents entre les végétariens et les omnivores. Néanmoins à ce niveau, ces différences témoignent plus d'un changement fonctionnel que d'un changement dans la composition du microbiote (Peltonen et al., 1992). Il semble en réalité que le régime alimentaire puisse apporter des changements importants et durables dans la composition du microbiote, davantage au niveau de l'iléon que du côlon, bien que cette supposition s'appuie sur des patients iléostomisés (Booijink et al., 2007).

Dans le cadre des maladies métaboliques, il a été montré que le régime pouvait influer sur l'abondance de grandes divisions bactériennes du microbiote intestinal. Des patients obèses qui ont subi soit un régime restreint en graisses soit un régime restreint en sucres pendant une année ont montré une augmentation prononcée des Bacteroidetes accompagnée d'une chute des Firmicutes (Ley et al., 2006). Cependant, le lien entre ce rapport Firmicutes/Bacteroidetes et l'obésité n'a pas été redémontré dans les études qui ont suivi (Duncan et al., 2008; Schwiertz et al., 2009).

Une autre étude a constaté que les souriceaux avaient classiquement un corps constitué de 40 % de matières grasses en plus, et 47 % de matières grasses gonadiques en plus que les souris sans germe, même s'ils consommaient moins de nourriture que leurs homologues sans germe. Le microbiote du côlon distal de la souris normale a ensuite été transplanté dans les souris sans germe, produisant une augmentation de 60 % de gras corporel dans les deux semaines, sans aucune augmentation de la consommation d'aliments ni de différences évidentes dans les dépenses d'énergie. Ce résultat confirme l'hypothèse que le microbiote intestinal module la quantité d'énergie extraite de l'alimentation. L'augmentation de la masse grasse a été accompagnée d'une résistance à l'insuline, d'une hypertrophie des adipocytes, et d'un niveau accru de diffusion de la leptine et du glucose (Backhed et al., 2004).

Pour élucider les mécanismes potentiels sous-jacents, ces chercheurs ont montré que le microbiote favorisait l'absorption des monosaccharides dans l'intestin et induisait la lipogenèse hépatique chez l'hôte. Enfin, par l'utilisation de souris génétiquement modifiées pour le facteur adipocytaire FIAF, ils ont démontré que le microbiote intestinal pouvait inhiber le facteur FIAF, également connu comme étant une angiopoïétine de type IV. FIAF inhibe l'activité de la lipoprotéine lipase, qui catalyse la libération d'acides gras à partir de lipoprotéines associées aux triglycérides, qui sont ensuite repris par le muscle et le tissu adipeux. Dans l'étude, la protéine

FIAF a entraîné la suppression de l'activité de la lipoprotéine lipase dans les adipocytes et le stockage des calories sous forme de graisse, entraînant l'équipe de J. L. Gordon à postuler que la régulation énergétique par le microbiote intestinal se fait par un certain nombre de mécanismes interdépendants. Ces mécanismes comprennent la fermentation bactérienne des polyosides non-digestibles, l'absorption intestinale des monosaccharides et des AGCC convertis ultérieurement en graisse dans le foie, ainsi que la régulation des gènes de l'hôte favorisant le dépôt de graisses dans les adipocytes (Backhed et al., 2004).

L'apport en fibres alimentaires peut engendrer un bénéfice aux individus ayant des syndromes métaboliques et des désordres gastro-intestinaux très variés. Les avantages d'une prise importante de fibres ont été reportés chez des patients atteints de diabète, d'hypercholestérolémie, d'hypertriglycéridémie, d'obésité ou bien d'hypertension (Anderson, 1986). Il a également été rapporté que les individus ayant un apport important en fibres sont moins sensibles au développement des maladies cardio-vasculaires ou du cancer du côlon (Lupton et al., 1985; Jacobs, 1986).

De même, les régimes avec un apport important en graisses et en protéines, mais de faible teneur en fibres, sont associés à un risque plus important de développer un cancer du côlon, contrairement aux régimes végétariens ou orientaux ayant des apports en fibres plus importants (Hayashi et al., 2002a). De plus, des Japonais qui adopteraient un régime à l'occidentale développeraient plus fréquemment des cancers du côlon (Finegold et al., 1974). Enfin, la production d'AGCC contribue à la prévention du cancer colo-rectal (McIntyre et al., 1993; Pryde et al., 2002).

L'impact du régime alimentaire sur la structure du microbiote n'est pas clair, et les conclusions peuvent différer d'une étude à l'autre, notamment dans le cadre de maladies métaboliques comme l'obésité. Cependant, il semble que la composante « fibre » du régime, qui fut l'objet de nombreuses comparaisons aussi bien du point de vue du microbiote que des désordres gastro-intestinaux, puisse avoir un impact sur la santé. Si son implication dans la structure du microbiote n'est pas totalement établie, c'est peut-être dans le potentiel fonctionnel du microbiote intestinal qu'il faut chercher à établir des relations claires avec l'hôte.

4.2 Les fibres alimentaires

Les fibres alimentaires peuvent être définies comme étant les polyosides des plantes et les lignines résistant aux enzymes digestives humaines. Les fibres alimentaires ne sont pas digérées dans l'intestin grêle et par conséquent entrent dans le côlon en grande partie non-dégradées. Le seul

polyoside des plantes connu comme partiellement hydrolysable par les enzymes humaines est l'amidon. Les lignines sont en fait peu présentes dans l'alimentation humaine. La lignine n'est pas un glucide mais un polymère de phényl-propane et possède des propriétés différentes des polyosides non-amylacés. C'est pourquoi par la suite, les fibres alimentaires que nous considèrerons seront composées de polyosides non-amylacés et de l'amidon résistant à l'hydrolyse de l'α-amylase humaine. Chimiquement, les fibres alimentaires se composent de polyosides non-amylacés, comme la cellulose et bien d'autres composants non-cellulosique tels que les dextrines, l'inuline, les cires, les chitines, les pectines, les bêta-glucanes et les oligosaccharides. Les fibres sont par conséquent une mixture hétérogène aussi bien chimiquement que physiquement et il est difficile de réaliser une généralité au niveau de leurs effets sur l'intestin humain.

Les fibres alimentaires peuvent être solubles dans l'eau ou insoluble. Les fibres solubles, d'ailleurs comme l'ensemble des fibres, ne peuvent pas être digérées (ou en partie seulement pour l'amidon) par les enzymes de l'hôte. Toutefois, lors de leur passage dans le tube digestif, une grande partie est rapidement fermentée par le microbiote intestinal. Le produit de la fermentation des bactéries est alors absorbable sous forme d'AGCC. Les fibres solubles absorbent l'eau pour devenir une substance gélatineuse pendant le transit intestinal.

Quant aux fibres insolubles, elles transitent dans le tractus intestinal tout en restant en grande partie inchangées. Par exemple, une étude a montré que le taux de dégradation de la cellulose est plus faible (15 à 25 %) que celui des polyosides non-cellulosiques (70 à 95 %). La digestibilité de la cellulose peut différer en fonction de son type et des autres fibres composant le régime (Cumming et al, 1980).

Par corollaire, il est aussi possible que les fibres alimentaires affectent en retour les bactéries en changeant leurs activités métaboliques ainsi que leur abondance dans le microbiote intestinal. Connaître l'impact de l'apport en fibres alimentaires sur le microbiote est important car les activités de ce dernier déterminent en grande partie l'environnement physicochimique du système gastro-intestinal.

4.3 La fermentation des fibres alimentaires

L'activité métabolique des bactéries concernées est ici essentiellement celle impliquée dans la fermentation des fibres. Le processus de fermentation est le résultat des actions concertées des espèces présentes dans le microbiote intestinal. La biochimie de cette fermentation essentiellement anaérobie est complexe. Ces larges polymères sont hydrolysés en unités monomériques comme le glucose, le galactose, le xylose, l'arabinose et les acides uroniques.

Via la glycolyse, ces monomères vont être hydrolysés en pyruvate. A partir du pyruvate, plusieurs réactions vont entrer en jeu et vont dépendre des espèces bactériennes présentes. Quelques produits intermédiaires peuvent être trouvés incluant l'éthanol, le méthanol, le formate, le lactate et le succinate. Néanmoins, ceux-ci vont être très rapidement utilisés pour produire des AGCC tels que l'acétate, le propionate et le butyrate, éventuellement accompagnés de gaz tels que l'hydrogène, le dioxyde de carbone et le méthane. La présence d'archées méthanogènes comme *Methanobrevibacter smithii* peut induire la réduction du dioxyde de carbone en méthane en utilisant le dihydrogène.

Les proportions relatives en moles des trois principaux AGCC sont approximativement 60 : 25 : 15 (acétate : propionate : butyrate). Tandis que l'abondance en AGCC augmente en fonction de l'apport en fibres alimentaires, leur proportion relative reste stable.

Après toutes ces considérations, une question demeure : un impact éventuel dû aux fibres alimentaires sur la structure du microbiote peut-il avoir un effet significatif sur l'hôte, notamment du point de vue des syndromes métaboliques ou bien des désordres gastro-intestinaux ?

Pour répondre à cette question, il faut d'abord être certain que les fibres peuvent avoir un impact sur le microbiote, aussi bien au niveau de sa composition que de ses activités métaboliques. Dans les années 1970, plusieurs études ont démontré un effet des fibres sur l'accroissement de la quantité totale des bactéries du microbiote, mais pas sur sa composition. Cependant, les techniques étant basées seulement sur la culture, bien qu'en anaérobiose, elles entraînent tout de même un biais important dans l'analyse car plus de 80 % du microbiote intestinal est incultivable ou incultivé (Suau et al., 1999; Hayashi et al., 2002b).

Peut-être faut-il seulement s'axer sur les activités métaboliques du microbiote et outrepasser l'importance d'énumérer les espèces du microbiote ? Il est vrai que bien des espèces, comme celles faisant partie de la même guilde fonctionnelle, partagent des activités similaires. Cependant, les résultats des différents programmes internationaux incluant du séquençage massif, tels que « MetaHIT » ou bien « the Human Microbiome Project », ne permettront pas de caractériser totalement les capacités métaboliques des différentes espèces composant le microbiote intestinal humain. En outre, deux espèces bactériennes différentes partageant les mêmes enzymes hydrolytiques peuvent ne pas forcément avoir la même efficacité dans un contexte de compétition.

4.4 Ecologie microbienne de la dégradation de la cellulose

Les bactéries colonisant le gros intestin ont accès seulement aux résidus alimentaires qui ont échappé à la digestion par les enzymes de l'hôte dans l'intestin grêle. La quantité et le type de ces glucides « non-digestibles » dans l'alimentation peuvent avoir une influence majeure sur les populations et le métabolisme de différents groupes bactériens du microbiote intestinal (Duncan et al., 2003; Duncan et al., 2007). Des glucides spécifiques comme l'inuline ou bien les fructo-oligossaccharides, aujourd'hui largement utilisés comme additifs alimentaires prébiotiques, ont été conçus pour manipuler le métabolisme intestinal et la biodiversité du microbiote intestinal afin d'être bénéfiques pour la santé (Gibson, 1998; Rowland et al., 1998; Kruse et al., 1999). Le principe des prébiotiques repose sur l'exploitation des différences de préférence de substrats et de capacités de compétition des différents membres de la communauté microbienne intestinale.

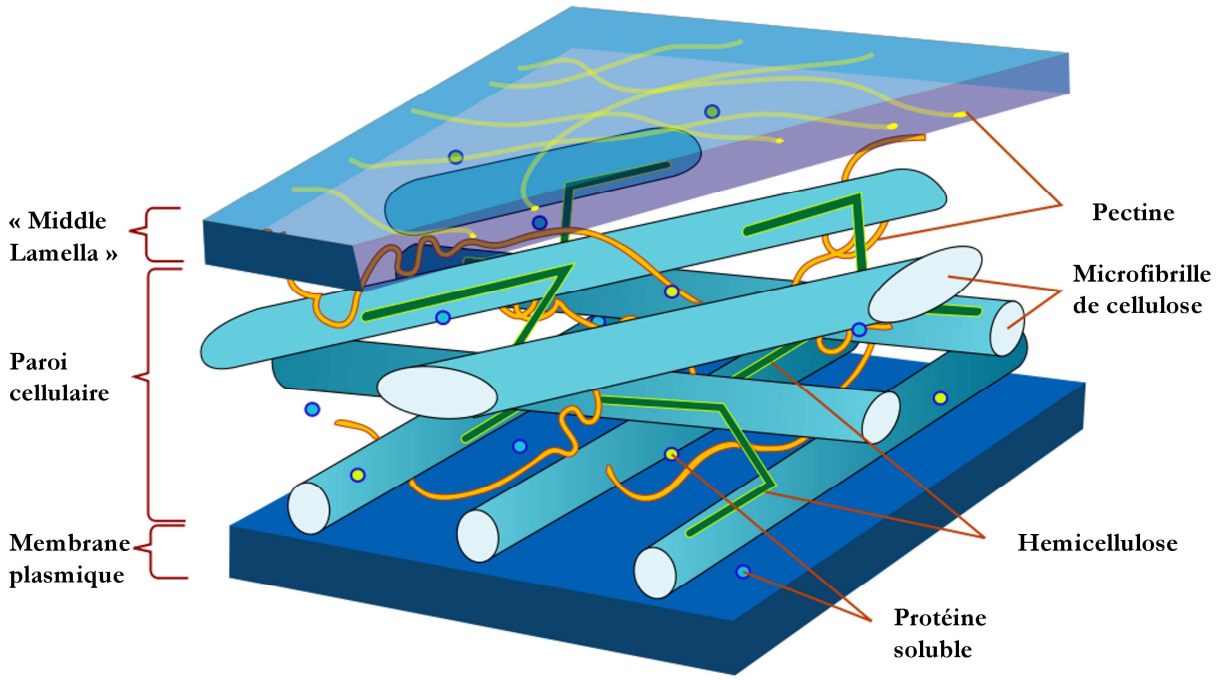


Figure 5 : Représentation schématique de la paroi d'une cellule végétale avec la localisation des principaux polyosides. La première partie est appelée « middle lamella » et est essentiellement composée de pectine. La séparation entre la paroi cellulaire et la « middle lamella » est formée de composés pecto-cellulosiques. La paroi cellulaire est quant à elle formée d'une matrice complexe de protéines solubles, de pectines, de cellulose et d'hémicelluloses (Pérez and Mazeau, 2005).

Les parois des cellules végétales se composent de micro-fibrilles de cellulose, incorporées dans une matrice complexe d'hémicelluloses, de pectines et de protéines (Figure 5). Les hémicelluloses, constituées d'une grande variété de polyosides, forment avec les microfibrilles de cellulose cette

matrice par l'intermédiaire de liaisons hydrogène. Les xyloglucanes, comme les xylanes, sont les constituants majeurs des hémicelluloses (Pérez and Mazeau, 2005).

La capacité à dégrader la cellulose semble être essentielle dans la dégradation de la plupart des structures formant les parois végétales, si bien que les bactéries non-cellulolytiques ont une capacité limitée à solubiliser ce genre de substrat. Par ailleurs *in vitro*, l'apport en protéines et en graisses ne change pas la faculté à dégrader la cellulose par les bactéries cellulolytiques (Firkins et al., 1991), tandis que l'accroissement du pH a un impact négatif sur l'adhérence des bactéries cellulolytiques aux fibres (Mourino et al., 2001).

Ces bactéries cellulolytiques capables de dégrader les xylanes, les mannanes et les pectines, n'utilisent pas forcément leurs produits de dégradation, qui deviennent ainsi disponibles pour les autres membres de la communauté (Coen and Dehority, 1970). Ces chaînes trophiques sont particulièrement mises en évidence lors de l'utilisation d'un substrat unique comme l'inuline ou l'amidon. Néanmoins, c'est l'hydrogène qui est l'élément clé dans les systèmes anaérobies car il est échangé continuellement entre les bactéries pour produire les AGCC à partir des polyosides, ces AGCC étant réduits par la suite en sulfate ou dihydrogène.

Le potentiel oxydo-réducteur de l'écosystème intestinal est essentiel pour que la dégradation de la cellulose soit efficace et rapide, si bien qu'il existerait un lien entre l'abondance des espèces méthanogènes et les bactéries cellulolytiques (Mourino et al., 2001; Robert and Bernalier-Donadille, 2003). La capacité des bactéries à résister à l'acidification de l'écosystème est due à leur potentiel d'abaissement du pH intracellulaire et de maintien d'un gradient de pH relativement faible à travers la membrane cellulaire. Ceci contourne le problème de l'accumulation d'anions acétates toxiques. Cependant, une telle stratégie ne sera couronnée de succès que si la bactérie possède des enzymes intracellulaires capables de résister à un diminution du pH intracellulaire (Russell and Wilson, 1996).

La possibilité d'adhérer au substrat est une autre propriété importante dans le processus de dégradation, et semble être une condition préalable à une dégradation efficace des polyosides provenant de la paroi d'une cellule végétale (Firkins et al., 1991; Weimer, 1996). De plus, ces propriétés d'adhérence peuvent conférer un avantage écologique aux bactéries cellulolytiques. Les études sur le microbiote des ruminants ont apporté le plus d'éléments à ce sujet. Elles ont notamment permis de mettre en évidence l'adhérence des bactéries cellulolytiques à la cellulose par l'intermédiaire d'un cellulosome. Lorsque l'on observe les bactéries en microscopie électronique, on remarque que celles adhérant aux parois végétales développent des protubérances qui « accrochent » les cellules végétales. Ces protubérances forment un cellulosome qui facilite le

processus d'adhérence. Le cellulosome est une structure extracellulaire multienzymatique qui apparaît comme essentielle dans la dégradation des polyosides d'origine végétale. Cet arrangement sous forme de cellulosome fournit un avantage concurrentiel dans l'utilisation directe des produits de l'hydrolyse (Schwarz, 2001).

Chez les ruminants, l'adhésion des bactéries à la cellulose a lieu en plusieurs étapes. Tout d'abord, les bactéries adhérèrent de manière non-spécifique à la matrice végétale. Ensuite, la formation de ligand spécifique avec le substrat va être facilitée par le cellulosome. Enfin, les bactéries ainsi fixées vont pouvoir proliférer sur les fibres végétales potentiellement digestibles en formant un biofilm[11]. Ce processus peut être perturbé par la nature du substrat, la teneur en eau, le pH, la charge ionique mais aussi par la compétition avec les autres microorganismes (Miron et al., 2001).

4.5 La dégradation des fibres d'un point de vue enzymatique

Afin de comprendre les mécanismes biochimiques qui entrent en jeu, il est nécessaire de s'intéresser aux enzymes impliquées dans la dégradation des fibres. La plupart de ces enzymes font partie de la famille des glycolyse hydrolases (GH). Leur fonction est d'hydrolyser la liaison glycosidique entre les glucides, entre hydrates de carbone ou bien entre un glucide et une autre molécule. L'hydrolyse d'un glucide permet la formation d'un glucide et d'un autre composant. Le terme « hydrolase » signifie que les liaisons carbone-oxygène, carbone-azote ou carbone-carbone peuvent être rompues durant l'hydrolyse. L'étape d'hydrolyse nécessite généralement une catalyse acide et requiert un donneur de protons (sous forme d'une molécule d'eau par exemple).

La dégradation de la cellulose requiert généralement une batterie de GH incluant des cellulases, des endoglucanases, des exoglucanases et des β-glucosidases qui agissent en synergie pour hydrolyser la fraction non-amylacée des fibres alimentaires. Par ailleurs, il est important de noter que le microbiote intestinal possède également toute une batterie d'enzymes capables d'hydrolyser des liaisons glucosidiques, autres que celles impliquées dans les fibres alimentaires (Tableau 2).

[11] Un biofilm est une communauté de micro-organismes (bactéries, champignons, algues ou protozoaires), adhérant entre eux et à une surface, et marquée par la sécrétion d'une matrice adhésive et protectrice.

Tableau 2 : Exemple de liaisons glycosidiques des polyosides ciblées par les enzymes du microbiote intestinal humain.

Liaison Carbone	Disaccharide correspondant	Structure	Origine	Enzyme ciblant la liaison
(1→4)	Maltose	Glc*p* α 1→4 Glc	Amidon	α -amylase
	Cellobiose	Glc*p* β 1→4 Glc	Cellulose	Cellulase (β-1,4-glucanase)
	Lactose	Gal*p* β 1→4 Glc	Lait	Lactase (β -galactosidase)
	Xylobiose	Xyl*p* β 1→4 Xyl	Xylane	Xylanase (Endo-1,4-β-xylanase)
	Chitobiose	GlcN β 1→4 GlcN	Chitine	Chitinase (1,4-β-poly-N-acetylglucosaminidase)
(1→6)	Isomaltose	Glc*p* α 1→6 Glc	Amylopectine	Amylopectin-1,6-glucosidase
	Gentiobiose	Glc*p* β 1→6 Glc	Gentianose	β-glucosidase
	Melibiose	Gal*p* α 1→6 Glc	Raffinose	Melibiase (α –galactosidase)
(1→3)	Acide hyalobiuronique	GlcUA*p* β 1→3 GlcN	Acide Hyaluronique	Hyaluronidase (Hyaluronoglucuronidase)
(1→2)	Saccharose	Fruc*f* β 1→2 α Glc*p*	Betteraves	β-fructofuranosidase

Glc=gluco-, Xyl=Xylo-, Gal=Galacto-, Fruc=Fructo-, N=amino-, GlcUA= acide glucronique. *p* et *f* indiquent respectivement pyranose et furanose.

Les GH peuvent être classées selon leurs domaines et leur similarité en acides aminés. La base de données CAZy[12] (Carbohydrate Active enZyme) mise en place et gérée par l'équipe de Bernard Henrissat contient les informations sur les GH et leur classification. Cette base a permis de définir 113 familles de GH (Cantarel et al., 2009). Elle comporte aussi 91 familles de glycotransférases, 19 familles de polyoside lyases et 52 familles modules fixant les hydrates de carbone (CBM : carbohydrate-binding module).

L'hydrolyse des substrats amylacés requiert l'intervention d' α-amylases faisant partie de la plus grande famille des GH. Cette famille, la GH 13, est imposante par sa diversité, si bien qu'il a été utile de la subdiviser en 35 sous-classes monofonctionnelles : c'est-à-dire une enzyme correspondant à un seul substrat (Stam et al., 2006).

Le séquençage complet de quelques génomes bactériens présents dans l'intestin des mammifères tels que ceux de *Ruminococcus flavefaciens* (Berg Miller et al., 2009) et *Bacteroides thetaiotaomicron* (Xu et

[12] La base de données CAZy (http://www.cazy.org) décrit les familles d'enzymes issues des domaines Eucaryote, Archea et Bacteria impliquent dans la dégradation, la modification ou la création de liaisons glucosidiques.

al., 2003) a permis d'apporter des informations complémentaires sur la complexité de l'interaction entre les bactéries et les polyosides. *Ruminococcus flavefaciens* produit une large panoplie d'enzymes correspondant à de nombreux substrats qui constituent la paroi végétale. Ces enzymes ont souvent une organisation multi-domaines comprenant des domaines catalytiques et des CBMs. La plupart des enzymes contiennent aussi des modules d'accrochage et de cohésion permettant la formation du cellulosome. Ces protéines enzymatiques sont codées dans le génome par le groupe de gènes *sca*, et leurs interactions permettent l'ancrage de *R. flavefaciens* à travers la paroi végétale (Flint et al., 2008).

Quant au génome de *B. thetaiotaomicron*, il inclurait des gènes codant 236 GHs et 15 polyosides lyases. Son activité hydrolytique ne serait pas extracellulaire mais périplasmique. Sa faculté à dégrader l'amidon serait codée par le groupe de gènes *sus*. Certains gènes de ce groupe participeraient à la fixation du substrat sur la membrane bactérienne, tandis que les autres permettraient d'hydrolyser des amyloses et des amylopectines (Flint et al., 2008). Il faut noter que *B. thetaiotaomicron* n'est pas représentatif du genre *Bacteroides* et la comparaison génomique de quatre espèces du genre *Bacteroides* a montré des potentialités différentes via la composition en GH sur leurs génomes (Xu et al., 2007).

Même si les différences sont évidentes entre une bactérie Gram positif spécifique du microbiote des ruminants et une bactérie Gram négatif isolée du microbiote intestinal humain, il existe une interopérabilité entre les espèces du microbiote. En termes de dynamique des génomes, des événements de duplication et de fusion des domaines enzymatiques sont observés, entraînant une large diversité aussi bien organisationnelle que génétique. Néanmoins, si l'on considère chaque enzyme de dégradation des fibres comme un assemblage de modules basiques, c'est-à-dire des modules catalytiques, de fixation de substrats, de modules transmembranaires ou extra-membranaires ou bien de modules d'accrochage et de cohésion, c'est une convergence du contenu génétique qui est observée notamment chez les bactéries qui partagent le même habitat (Xu et al., 2007). La nécessité de s'adapter à la variété de substrats alimentaires explique cette diversité dans l'organisation des modules. Cependant, l'interopérabilité des espèces est le résultat d'une forte pression de sélection qui a nécessité le partage et l'intégration de ces différents modules par transferts horizontaux de gènes (Lozupone et al., 2008).

L'enjeu est de déterminer quelles espèces du microbiote possèdent cette interopérabilité fonctionnelle et cette faculté à entrainer des flux métaboliques différents dans la chaîne trophique lorsqu'il y a des changements dans l'apport en quantité de ces substrats fermentescibles arrivant dans le côlon (Voir Tableau 1). Ces nouvelles connaissances permettront d'évaluer la robustesse de l'écosystème face à des changements environnementaux.

5 TECHNIQUES ET METHODES D'ANALYSE

5.1 Méthodes d'extraction et de préparation des acides nucléiques

La méthode d'extraction et de purification de l'ADN est une étape extrêmement critique dans les études moléculaires d'un écosystème complexe, notamment celui du microbiote intestinal, composé de 10^{11} bactéries majoritairement anaérobies par gramme de matière fécale. L'extraction d'acides nucléiques bactériens demeure encore plus problématique lorsqu'il s'agit de biopsie ou bien de pièce opératoire.

Le microbiote intestinal étant composé notamment de bactéries Gram positif et Gram négatif, l'accessibilité aux acides nucléiques de certaines espèces est difficile. En effet, la couche de peptidoglycane des bactéries à Gram positif, très épaisse par rapport à celle des bactéries à Gram négatif, nécessite l'utilisation de méthodes appropriées tout en limitant l'action des enzymes dégradant l'ADN des bactéries Gram négatif. De plus, d'autres éléments tels que les acides humiques et composés aromatiques perturbent les étapes de purification qu'il convient d'éliminer avant la précipitation de l'ADN. Par ailleurs, certains d'entre eux empêchent une quantification correcte avec les techniques d'absorbance UV.

Il existe de nombreux protocoles à disposition pour extraire et purifier l'ADN de différents types de matrices. Néanmoins, en ce qui concerne l'étude du microbiote intestinal, essentiellement deux principes ont été utilisés dans les techniques d'extractions utilisées jusqu'à présent : la lyse mécanique grâce à l'agitation en présence de billes de verre ou de zirconium (Matsuki et al., 2004; Gill et al., 2006; Ley et al., 2006b) et la lyse enzymatique (Eckburg et al., 2005; Kurokawa et al., 2007). Ces deux méthodes sont non exclusives et peuvent être utilisées conjointement. L'utilisation de billes est généralement jugée plus appropriée pour assurer l'efficacité de la lyse des microorganismes Gram positif même s'il convient d'optimiser le temps d'agitation pour s'assurer d'une lyse correcte de la paroi cellulaire sans toutefois entraîner la dégradation des acides nucléiques (Zoetendal et al., 2001a). Parfois, il est nécessaire de réaliser à cette étape des réplicats techniques si les échantillons sont récalcitrants et que la quantité d'ADN est insuffisante pour la suite (McOrist et al., 2002; Scupham et al., 2007).

L'extraction ADN est évaluée en fonction de la qualité et la quantité d'ADN obtenu à partir des échantillons fécaux. Cette évaluation est nécessaire pour des analyses en aval comme la PCR et la construction de banque pour le séquençage. En outre, une bonne qualité de l'ADN est d'une

importance critique pour les analyses en aval, ainsi que l'absence d'agent inhibiteur de PCR[13], comme les cycles aromatiques ou les polyphénols, sont fréquents dans des échantillons fécaux. En présence de telles molécules, il est nécessaire de réaliser une dilution telle qu'il sera possible d'effectuer l'amplification des gènes codant pour l'ARNr 16S.

Lorsque que l'on souhaite accéder à l'activité transcriptionnelle du microbiote, il est nécessaire de réaliser une extraction des ARN totaux. Cependant, cette dernière est particulièrement délicate. En effet cette molécule simple brin est sensible aux ribonucléases ubiquitaires et présentes en particulier à la surface de la peau. L'extraction d'ARN repose sur le même principe que l'extraction d'ADN à ceci près qu'elle demande l'utilisation d'une solution phénol-chloroforme à manipuler sous une hotte chimique (Zoetendal et al., 2006). L'utilisation d'une solution de phénol-chloroforme acide (i.e. pH = 5) permet de dénaturer l'ADN qui va se retrouver dans la partie organique pour ne trouver que les ARN dans la partie aqueuse. Le produit d'une extraction d'ARN fournit près de 99 % d'ARN ribosomiques comprenant les sous-unités 23S, 16S, 5S ainsi que les ARN de transfert, le reste formant les ARN messagers (ARNm). Une solution d'ARN doit être manipulée dans la glace pour ralentir l'action d'enzymes potentiellement présentes et avec des gants pour éviter toute contamination par des ribonucléases par l'utilisateur. De plus, pour une utilisation sur le long terme, cette solution doit être stockée à -80°C. Une rétro-transcription suivie d'une polymérisation permet de stabiliser l'ARN simple brin en ADN complémentaire (ADNc). Ce dernier, plus résistant, est plus facilement manipulable qu'une solution d'ARN. Si l'on souhaite étudier l'expression des gènes d'un échantillon, il est nécessaire d'accéder plus facilement aux ARNm. Ceci est rendu possible notamment grâce à l'utilisation de kits d'appauvrissement en ARNr (Voir la partie « Méthodologie pour accéder aux ARN messagers », page 68).

[13] PCR : la « polymerase chain reaction » ou réaction en chaîne par polymérase, permet de copier avec un facteur de l'ordre du milliard une séquence d'ADN.

5.2 Ecologie moléculaire

L'écologie moléculaire consiste à appliquer des techniques de biologie moléculaire comme la PCR quantitative ou bien la génomique, à des questions écologiques comme par exemple l'étude de la composition et de la dynamique d'une communauté en fonction des changements environnementaux. L'étude de la composition du microbiote repose sur l'analyse de l'ADN génomique. En pratique, la première étape consiste à réaliser une amplification par PCR du gène à cibler notamment celui codant pour l'ARNr 16S.

5.2.1 La PCR du gène codant pour l'ARNr 16S

Il existe un certain nombre de biais et de limites associés à la réaction de PCR sur une matrice ADN complexe. Une des limites critiques est la sélection des amorces de PCR. En outre, des amorces considérées comme « universelles » pour le règne des bactéries excluent de fait un grand nombre de séquences issues des bases de données en perpétuelle expansion. La stratégie consiste alors à cibler un groupe de séquences connues de l'écosystème intestinal à étudier, par exemple celles appartenant à un genre, et à situer les amorces sur les régions dites conservées de l'ARNr 16S.

Cependant, cela peut engendrer le sacrifice d'une partie de la séquence au profit de plus de diversité détectée. Par exemple, l'amorce dite « universelle » située la plus en amont (en 5') du gène codant pour l'ARNr 16S, la « Bact-8F », est placée sur la région conservée A (Figure 6), est aspécifique des Actinobactéries avec trois nucléotides polymorphes (Edwards et al., 1989). C'est pourquoi les études privilégiant l'obtention des séquences complètes d'ARNr 16S peuvent entraîner un biais vis-à-vis des Actinobactéries. A contrario, une amorce de PCR placée en amont de la région variable V3 (i.e. environ 300 pb en aval du 5') permet de capter plus de diversité malgré une longueur de séquence finale obtenue inférieure à 1100 pb.

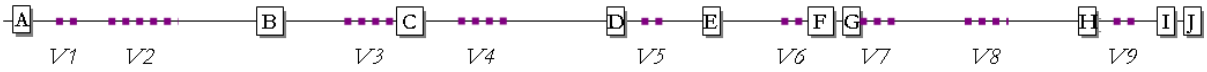


Figure 6 : Représentation schématique de la distribution des régions hypervariables ainsi que des régions conservées du gène de l'ARN 16S d'*Escherichia coli* (orientation 5' – 3'). Les régions hypervariables (notées de V1 à V9) se caractérisent par une diversité nucléotidique très importante et permettent de séparer facilement des espèces voire des souches d'une même espèce. Les régions conservées (notées de A à J) constituent le squelette inamovible de la sous-unité 16S de l'ARN ribosomal si bien qu'elles constituent des cibles idéales pour une amplification universelle par PCR.

Un autre facteur important pouvant intervenir dans la représentativité de la diversité bactérienne est le nombre de cycles de PCR. Il a été montré que plus le nombre de cycles était important plus

la diversité bactérienne détectée était faible (Bonnet et al, 2002). Il est donc nécessaire de réaliser plusieurs PCR avec un nombre de cycles réduit et de regrouper les produits d'amplification avant de réaliser l'étape de séquençage.

5.2.2 La PCR quantitative (qPCR)

En biologie moléculaire, la réaction de polymérase en chaîne en temps réel, également appelée PCR quantitative (qPCR), utilise les principes de la PCR classique afin d'amplifier et de quantifier simultanément une molécule ciblée d'ADN. La quantification repose sur le suivi à chaque cycle de la réaction d'amplification enzymatique au moyen d'une molécule fluorescente utilisée comme marqueur et capable d'émettre dans des conditions bien définies un rayonnement proportionnel à la quantité d'ADN produite (Jung et al ; 2000).

En théorie, à partir d'un brin d'ADN on obtient 2^n brins d'ADN après n cycles de PCR. Néanmoins, cette phase exponentielle, plus ou moins efficace, n'intervient qu'après une phase d'initiation nécessaire à l'obtention d'une quantité suffisante de produits PCR. Une phase plateau, correspondant à une modification du milieu réactionnel, marque la fin de la phase exponentielle.

La quantification d'une molécule d'ADN ciblée par des amorces spécifiques, par exemple le gène de l'ARNr 16S ou un autre gène d'intérêt, est relative à la durée de la phase d'initiation, ce qui conduit à détecter le moment où débute la phase exponentielle : le « treshold cycle » ou Ct. Ce moment est défini comme étant le nombre de cycles nécessaires pour obtenir un signal fluorescent significativement plus élevé que le bruit de fond. Plus le Ct est important, plus le nombre de molécules cibles d'ADN à quantifier est faible.

En écologie microbienne, la qPCR demande des conditions particulières. En effet, l'ADN extrait d'un échantillon fécal peut contenir des molécules inhibitrices de la PCR qu'il faut évaluer avant de réaliser l'analyse. Ces dernières peuvent faire varier l'efficacité de l'amplification d'un échantillon à l'autre et donc fausser l'interprétation. D'autre part, l'évaluation de la spécificité de la qPCR vis-à-vis d'une molécule cible dans un échantillon complexe est très importante. En pratique, si l'on souhaite quantifier un groupe bactérien, la molécule cible sera le gène codant pour la sous-unité 16S de l'ARNr avec des amorces spécifiques de groupe. Un alignement multiple des séquences connues pour ce groupe est comparé aux séquences que l'on ne souhaite pas cibler. C'est ainsi que l'on peut définir sur la séquence, une région de quelques nucléotides spécifique de ce groupe, où l'on dessinera *in silico* une amorce de PCR. Ensuite *in vitro*, on peut utiliser des clones bactériens et des extraits d'ADN de souches bactériennes pour borner le système PCR.

5.2.3 Séquençage haut débit

La méthode de séquençage Sanger a permis d'effectuer les premiers inventaires moléculaires basés sur le séquençage du gène codant l'ARNr 16S (Suau et al., 1999). Néanmoins, l'information obtenue par le clonage et le séquençage des gènes était subordonnée au nombre de clones séquencés. De plus, cette technique prend beaucoup de temps du fait de l'isolement de clones bactériens et présente des coûts relativement élevés.

Auparavant, le dilemme était de choisir entre une analyse en profondeur de quelques échantillons et une analyse avec plus d'échantillons à inclure mais avec une résolution plus faible. Avant d'être définitivement dépassée par les techniques à haut débit de séquençage comme le pyroséquençage, la technique de séquençage de Sanger permet d'obtenir des fragments de séquences plus longs, permettant un accès plus facile à plus de diversité nucléotidique par séquence.

A grande échelle, c'est-à-dire avec un nombre de séquences obtenues supérieur à 10 000 lectures, cette technique a permis de fournir de précieuses informations sur la diversité microbienne jusqu'alors inconnue de différents sites anatomiques du corps humain (Eckburg et al., 2005; Bik et al., 2006). Grâce aux avancées méthodologiques, le séquençage haut débit permet d'obtenir un grand nombre de séquences sur un grand nombre d'échantillons pour des coûts moindres par rapport à la méthode de Sanger (Tableau 3).

Le pyroséquençage inclus dans le « 454 » a permis une élévation de la puissance pour étudier la complexité des communautés microbiennes (Margulies 2005). A chaque utilisation, cette approche fournit généralement plusieurs centaines de milliers de séquences par série, là où la méthode de Sanger est limitée au nombre de puits sur une plaque PCR.

Cette technologie, qui auparavant fournissait des fragments de séquences courts d'environ 50 à 100 paires de bases nucléotidiques, permet avec l'avènement des technologies dites « FLX » puis « Titanium » d'obtenir des fragments d'une longueur supérieure à 400 paires de bases. Ces dernières requièrent encore l'utilisation de la méthode de Sanger pour finaliser le séquençage d'un génome par exemple. Néanmoins, il y a fort à parier que l'avancement technologique mettra définitivement un terme à la méthode de Sanger lorsque la longueur des séquences en haut débit atteindra plus de 1 000 paires de bases (Tableau 3).

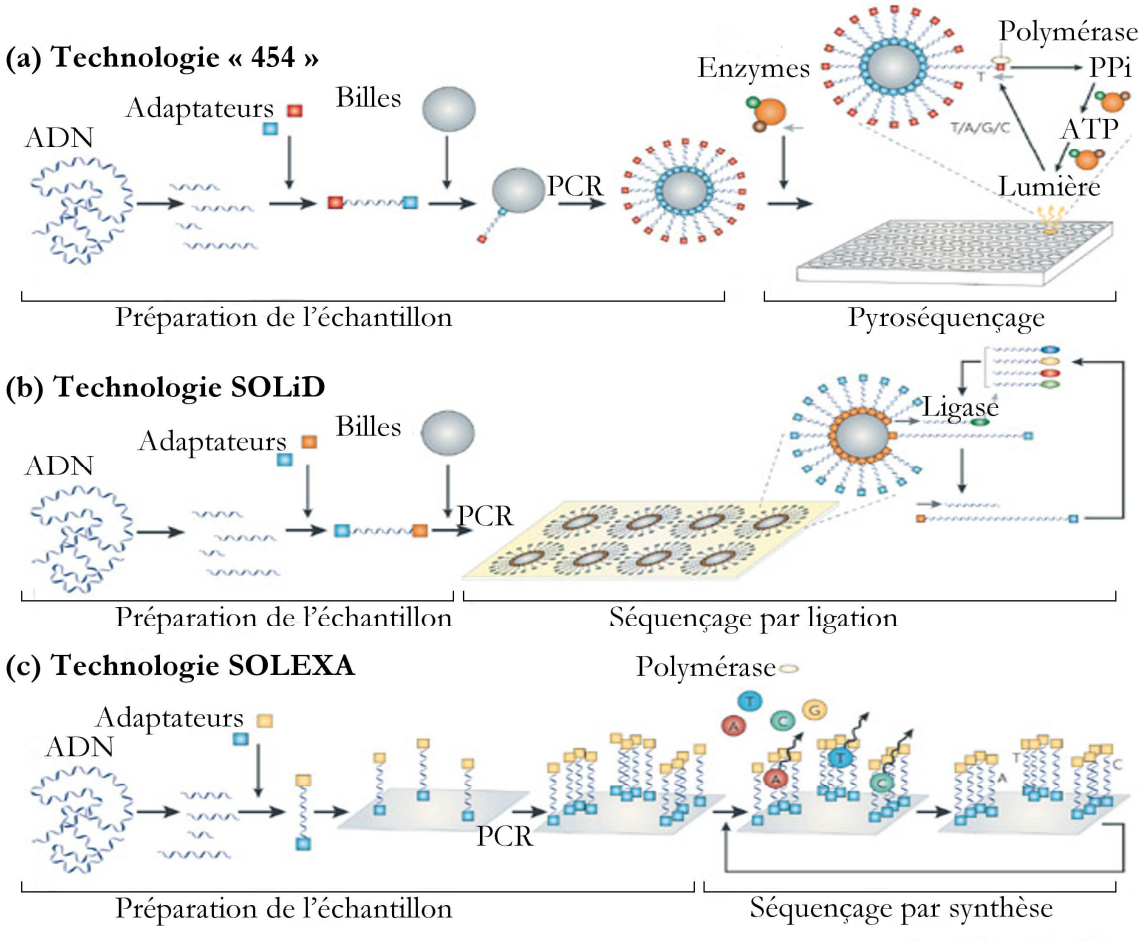


Figure 7 : Illustration des différentes techniques de séquençage à haut débit. (a) la méthode de séquençage « 454 » est une approche en deux étapes. Premièrement l'ADN est nébulisé et des adaptateurs de quelques nucléotides sont attachés. Chaque fragment est attaché à une bille et chaque bille est amplifiée dans une gouttelette d'une PCR en émulsion. Ceci génère des copies multiples d'un même fragment d'ADN sur chaque bille. Deuxièmement, les billes sont capturées sur une plaque avec des puits d'un volume d'un picolitre et le pyroséquençage est réalisé en parallèle sur chaque fragment d'ADN. L'incorporation des nucléotides est détectée par le largage d'un pyrophosphate inorganique (PPi), ce qui conduit à la génération enzymatique de photons (i.e. le PPi est converti en ATP et la luciférase utilise l'ATP pour générer de la lumière). Ce cycle est itérativement répété pour les quatre bases A/T/G/C. **(b)** La technologie SOLiD a une procédure d'amplification similaire au « 454 », mais la stratégie de séquençage est radicalement différente. Les billes sont déposées sur une lame de verre et la séquence est déterminée par une hybridation et une ligation séquentielle d'oligonucléotides quasi aléatoires, avec une paire de bases bien déterminée identifiable par un fluorophore. Après que la couleur ait été enregistrée et l'oligonucléotide ligué enlevé, ce processus est alors répété six à sept fois afin d'obtenir une longueur de séquence d'environ 35pb. **(c)** La première étape du séquençage SOLEXA est basée sur l'amplification de l'ADN sur une surface solide à l'aide d'une PCR avec des amorces ancrées. De multiples cycles d'amplification sont ensuite réalisés pour créer un millier de copies simple brin de chaque fragment d'ADN. Le séquençage est effectué séquentiellement à l'aide d'amorces, de l'ADN polymérase et de quatre nucléotides labellisés par un fluorophore, bloquant réversiblement la PCR. Après l'incorporation d'un nucléotide, l'image est capturée et l'identité de la première base est enregistrée. Les fluorophores sont ensuite retirés et les étapes d'incorporation, de détection et d'identification sont répétées (Medini et al., 2008).

Tableau 3: Comparaison des coûts et des sorties des technologies de séquençage.

Méthodes de séquençage	millions de bases par série	coût par base	longueur de la lecture en paires de bases
Sanger	0,07	0,1	700
454 pyroséquençage	400	0,003	400
SOLiD/SOLEXA	2000	0,0007	35

De plus, l'autre intérêt de la technique de pyroséquençage réside dans l'utilisation de codes-barres que l'on place en aval des adaptateurs, par ligation sur les fragments que l'on souhaite séquencer (Figure 7). Ce code-barres constitué de nucléotides prédéterminés permet, lorsque l'on mélange plusieurs échantillons différents dans la même série, de tracer chaque échantillon individuellement. En utilisant l'approche des codes-barres, plusieurs échantillons peuvent être traités en parallèle sur une plaque. Cela permet entre autres de réaliser des réplicats techniques. Bien que cette approche ne soit pas soumise à des biais dus au clonage, il y a encore des doutes sur les déviations que pourrait introduire la PCR en émulsion. Le pyroséquençage, dont le coût est continuellement en baisse, est devenu une méthode classique dans l'analyse de la structure d'un écosystème complexe. Cette technique a d'ores et déjà été utilisée dans de nombreuses publications, aussi bien pour des inventaires moléculaires du gène codant l'ARNr 16S (Turnbaugh et al., 2009; Zhang et al., 2009) que pour des inventaires fonctionnels de microbiomes basés sur l'ADN génomique ou les ARN messagers (Gilbert et al., 2008; Willner et al., 2009).

Cependant, bien que la couverture d'espèces estimée par inventaire moléculaire du de l'ARNr 16S de l'écosystème intestinal semble être atteinte à plus de 90 % grâce au pyroséquençage, ce dernier redevient une technique exploratrice lorsqu'il s'agit d'étudier le potentiel génétique ou l'activité transcriptionnelle du microbiote. L'effort de séquençage pour couvrir les pan-génomes et transcriptomes de la communauté microbienne est bien plus important lorsqu'il s'agit d'étudier la diversité fonctionnelle d'un écosystème. C'est pour cela que les technologies de séquençage massif comme le SOLiD et SOLEXA ouvrent des perspectives telles, qu'elles sont en passe d'effacer des technologies comme les puces à ADN (Figure 7).

L'évolution de toutes ces techniques a évidemment entraîné en parallèle le fourmillement d'applications bioinformatiques. De plus en plus faciles d'utilisation pour les biologistes non experts, les outils de bioinformatique sont couplés directement à la sortie du séquenceur

permettant par exemple des analyses différentielles très rapides et précises à l'échelle d'une unité taxonomique ou fonctionnelle.

Les méthodologies de séquençage ne sont plus un frein à l'exploration moléculaire d'un écosystème, aussi bien dans sa dynamique structurelle que fonctionnelle. Il appartient maintenant aux biologistes de parfaire leur plan d'expérience afin d'obtenir la puissance statistique nécessaire pour répondre à leurs questions. Il est nécessaire aussi pour le biologiste d'anticiper la quantité de données à traiter, où la séquence est devenue une unité de mesure à la fois qualitative et quantitative.

5.3 Bioinformatique

Avec l'avènement du séquençage haut débit, l'utilisation et le développement d'outils bioinformatique sont devenus encore plus indispensables. Dans une approche métagénomique, le but est aussi de pouvoir caractériser l'inconnu. Ainsi, l'utilisation, de manière systématique, de requêtes sur des bases de données de séquences connues peut engendrer un biais dans l'analyse des séquences obtenues car une partie d'entre elles sont inconnues. Dés lors, lorsque l'on veut effectuer une approche sans *a priori*, la première chose à faire est de comparer toutes les séquences obtenues entre elles. Deux types d'approches ont émergé. L'une, très répandue, est basée sur l'alignement de ces séquences, tandis que l'autre, en cours de développement, réunit les méthodes indépendantes de l'alignement comme l'approche tétranucléotides (Teeling et al., 2004a; Teeling et al., 2004b; Woyke et al., 2006). Ensuite, l'utilisation de ces comparaisons de séquences peut permettre de réaliser d'autres approches sans *a priori* comme le regroupement des séquences selon un critère de similarité ou bien de tester l'existence de ces séquences dans les bases de données relationnelles.

Pour finir, ces séquences peuvent être stockées dans des bases de données relationnelles. Avant de parcourir plus profondément les différentes méthodes de bioinformatiques appliquées à l'analyse de séquences, il est nécessaire de mentionner le problème dû aux séquences chimériques, qui est un problème typique du séquençage massif à partir d'un ADN extrait d'un échantillon complexe.

5.3.1 Les séquences chimériques

Des méta-analyses ont montré que de nombreuses séquences du gène codant pour l'ARNr 16S pouvaient être en fait des artefacts de la PCR. Ainsi, il a été estimé que, globalement, 5 % des inventaires moléculaires seraient susceptibles de contenir des anomalies (Ashelford et al., 2006).

La plupart des anomalies détectées dans les inventaires moléculaires sont constituées de séquences chimériques. Ces anomalies peuvent être de deux types : soit des séquences dites chimériques c'est-à-dire produites à partir de deux ou plusieurs fragments d'ADN phylogénétiquement distincts au cours de l'amplification PCR, soit des erreurs de séquençage lors de l'assemblage, mais celles-ci peuvent être rectifiées par re-séquençage.

La présence de ce grand nombre de séquences chimériques, de 1 à 30 % selon les études (Ashelford et al., 2006), inhérent à la PCR a pour conséquence de surestimer la biodiversité présente dans un écosystème. Par ailleurs, cette présence peut entraîner des relations phylogénétiques improbables, et par conséquent empêcher une identification taxonomique correcte. Le risque de ne pas rechercher systématiquement les chimères dans son jeu de données est de créer de fausses lignées phylogénétiques nouvelles et d'analyser ensuite une diversité inexistante biologiquement. Néanmoins, la PCR en émulsion utilisée dans les nouvelles technologies, qui isole chaque fragment d'ADN individuellement, tend à réduire ce problème de séquences chimériques.

Avec la croissance des inventaires moléculaires aussi bien en nombre qu'en taille, la nécessité de régler le problème des séquences chimériques est passée d'un acte occasionnel possible avec des logiciels comme « Chimera check », à un acte automatisé avec « Mallard » (Ashelford et al., 2006) ou bien « bellerophon 3 » (Huber et al., 2004). Ces outils fonctionnent sur des séquences de gène codant pour l'ARNr 16S et ont été calibrés par rapport à des jeux de données existants. Par conséquence, il est alors difficile de discerner la variabilité biologique (i.e. faux positifs) de celles engendrée par les séquences chimériques.

5.3.2 Comparaison des séquences par alignement

A l'ère de la post-génomique et de la métagénomique, la bioinformatique connaît une véritable révolution grâce à l'émergence des biotechnologies à haut débit. L'enjeu actuel pour les « supercalculateurs » est de soutenir ce flux de données constant, issu du séquençage en masse des acides nucléiques. L'une des applications les plus importantes qui en découlent est la comparaison des séquences afin par exemple de sonder la biodiversité fonctionnelle ou phylogénétique.

D'une manière générale, la comparaison s'effectue à travers l'alignement multiple (global ou local) des séquences nucléiques pour comparer plusieurs longs fragments génomiques (supérieurs à 40 kb) ou bien des dizaines de milliers de petites séquences (inférieures à 50 pb). Pour répondre à ces besoins, plusieurs algorithmes ont été développés, soit pour mettre en évidence des motifs communs au sein de plusieurs séquences, soit pour traiter des séquences de plus en plus distantes.

Depuis 1990, le Blast (Basic Local Alignment Search Tool) puis PSI-Blast de Altschul et ses collègues (Altschul et al., 1997) est certainement devenu l'outil de bioinformatique le plus utilisé par les biologistes, en supplantant FASTA (alignement global) (Pearson et al, 1988), pour réaliser des alignements locaux paires à paires des séquences nucléiques et protéiques. Pour réaliser des alignements multiples globaux, ClustalW, réputé pour sa finesse, est particulièrement utilisé, aussi bien en phylogénie qu'en recherche de motifs conservés (Thompson et al., 1994).

Cependant, avec la hausse de la quantité de séquences à traiter, ClustalW apparaissant comme un algorithme trop gourmand en temps de calcul, d'autres aligneurs multiples bien plus rapides ont émergé comme MUSCLE, utilisant des k-mers, ou bien MAFFT, utilisant des transformations de Fourier. Néanmoins, la rapidité se fait bien souvent au détriment de la qualité des alignements, indispensables par la suite pour en déduire les distances génétiques séparant les séquences. L'exigence de la finesse de l'alignement conduit les biologistes à se tourner vers des aligneurs spécialisés comme NAST conçu, en s'appuyant sur une base de référence, pour aligner uniquement des séquences codant pour la sous-unité ribosomale 16S des procaryotes (DeSantis et al., 2006). Voici une liste non exhaustive de différents aligneurs :

- ClustalW : Le plus utilisé car le plus ancien, un des plus fins, des plus diffusés et accessibles grâce à son interface ClustalX pour les biologistes. Algorithme très gourmand, mais une issue est possible avec sa version MPI[14] (Li, 2003).
- MAFFT : Utilisant la transformation de Fourier et les itérations pour affiner l'alignement multiple, il offre une multitude d'options pour aligner localement et globalement de longs fragments ou des milliers de petites séquences. Il n'existe pas de version MPI disponible pour l'instant.
- MUSCLE : S'appuyant sur le comptage de k-mers, il est très rapide mais moins précis que MAFFT.
- T-coffee : S'appuyant sur des bibliothèques de pré-alignement, il est surtout utilisé pour aligner des séquences protéiques. Il est assez lent.
- Praline : il exploite la structure secondaire des protéines pour réaliser un alignement multiple. Il est très lent.

5.3.3 Comparaison par approche tétranucléotides

Dans une approche métagénomique, l'ADN est directement extrait d'un échantillon environnemental et cloné dans des vecteurs tels que des cosmides, des fosmides ou bien des chromosomes artificiels bactériens (BAC). Les banques métagénomiques obtenues peuvent alors

[14] MPI pour Message Parsing Interface est un protocole de communication utiliser pour programmer des ordinateurs dans une architecture parallélisée. L'utilisation du protocole MPI et d'architectures parallélisées permet de réduire considérablement la temps de calcul.

être criblées pour une fonction donnée et les inserts peuvent être séquencés, permettant l'accès à de nouvelles séquences spécifiques.

Néanmoins, malgré le potentiel de l'approche métagénomique, qui permet d'augmenter considérablement la connaissance de la composition et de la fonction d'une communauté microbienne, plusieurs problèmes méthodologiques doivent être résolus. Un des problèmes majeurs rencontré est l'identification taxonomique de l'origine de l'insert. En effet, seulement 5 à 10 % des fosmides contiennent un marqueur phylogénétique comme l'ADNr 16S ou bien des gènes de ménage (*rpoA*, *recA*) et peuvent alors être assignés à une espèce ou un groupe taxonomique.

C'est pourquoi, il y a un réel besoin de nouveaux outils d'assignation. Le biais dans la composition nucléotidique des génomes procaryotes est le résultat de la pression sélective, et des mécanismes de réparation et de réplication de l'ADN. Ce biais constitue une signature génomique qui peut être exploité pour l'assignation taxonomique. Tout d'abord, on peut apparier deux fragments nucléiques selon leur teneur en G+C %. Puis en complément, on peut utiliser le meilleur « Blast hit » ou l'usage du codon pour évaluer l'origine taxonomique (Danchin, 2002).

Cependant, ces techniques possèdent des biais importants. En effet, le G+C % peut varier considérablement au sein du génome et ne permet pas d'obtenir un signal phylogénétique puissant. Pour un insert métagénomique de 40kb, soit environ 40 gènes, sa requête contre les banques publiques de séquences par Blast peut fournir des résultats non significatifs. Fréquemment, dans un insert donné, plusieurs hits peuvent avoir une origine phylogénétique différente. C'est le cas par exemple, lorsqu'on obtient des séquences de familles de protéines phylogénétiquement non spécifiques. Quant à l'analyse de l'usage du codon, son signal phylogénétique peut être brouillé par des transferts de gènes horizontaux (Teeling et al., 2004a).

Plus que le biais de codons, l'enchaînement des codons lui-même n'est pas aléatoire. C'est pour ces raisons que l'apprentissage de la signature génomique doit s'effectuer sur des mots d'au moins quatre nucléotides (dits aussi 4-mers ou tétranucléotides). Pour un génome entier, dans le cadre d'une détection de gènes, les jeux d'apprentissage utilisés sont formés à partir de mots de cinq ou six nucléotides (dits 5-mers ou 6-mers). Il est raisonnable de penser que pour assigner des fragments métagénomiques de 40 kb, un apprentissage de la signature génomique peut s'effectuer avec des mots de quatre nucléotides (McHardy et al., 2007).

Au l'échelle du microbiome, le métagénome d'une communauté contient une mixture de plusieurs génomes individuels et ne possède pas de signature proprement-dite. Seulement, l'approche métagénomique génère beaucoup de séquences avec peu de similarité avec les séquences connues

dans les bases de données. Avec les techniques à haut débit comme le pyroséquençage, il est difficile d'obtenir à partir d'un écosystème complexe de longs fragments génomiques non chimériques. Pourtant, l'équipe de F. Rohwer a émis l'hypothèse que sous la pression de son environnement, un microbiote devrait avoir sa propre signature. En effet, la composition en dinucléotides de séquences issues du pyroséquençage permet d'expliquer près de 80 % de la variabilité entre différents métagénomes d'écosystèmes très différents tels que le microbiome humain et les mines acides (Willner et al., 2009). Par ailleurs, cette propriété fonctionne aussi avec les métagénomes viraux. Ces signatures dinucléotidiques sont entraînées par la sélection de l'environnement, lequel environnement peut être dominé par quelques espèces très abondantes influençant la fréquence des dinucléotides.

A l'échelle du génome, chaque espèce a sa propre signature génomique. Ce biais génomique peut aussi être exploité pour discriminer des niveaux phylogénétiques plus élevés. L'exploration de ce biais permet de trouver un signal phylogénétique qui peut être utilisé pour l'assignation d'un fragment de génome (Teeling et al., 2004a). L'apprentissage de cette signature génomique peut s'effectuer en comptant la fréquence des 256 combinaisons de tétranucléotides possibles. Les fragments métagénomiques peuvent être comparés les uns par rapport aux autres pour former des groupes taxonomiques auxquels ils peuvent être assignés. Ce biais génomique a pu être comparé à des phylogénies basées sur le gène de l'ARNr 16S. Les similarités observées entre les phylogénies basées sur l'ADNr 16S et celles créées à partir de l'usage des tétranucléotides indiquent que ce dernier contient un signal phylogénétique fort (Teeling et al., 2004b). Plusieurs tentatives ont été publiées en utilisant ce principe mais pour l'instant, malgré des résultats prometteurs, la précision de l'assignement n'est pas compatible avec l'exigence attendue.

En revanche, la méthode des tétranucléotides a pu être utilisée avec succès directement sur les séquences issues d'inventaires moléculaires sur le gène de l'ARNr 16S (Woyke et al., 2006; Rudi et al., 2007). Cette méthode pourrait remplacer à l'avenir les approches dépendantes de l'alignement consommant outrageusement du temps de calcul. De nouveaux algorithmes mathématiques devront être développés dans le futur pour exploiter la signature de fragment génomique.

5.3.4 Les matrices de distance et les regroupements de séquences.

En admettant que l'échantillonnage et le traitement d'analyse des échantillons produisent une image représentative de l'écosystème de départ, aux questions « qui sont-ils » et « que font-ils ? » vient s'ajouter la question « et en quelles proportions ? ». C'est le regroupement de séquences en unités opérationnelles qui permet de répondre à cette dernière. En effet, plus une unité opérationnelle regroupera un nombre important de séquences, plus la proportion de ce taxon ou

de cette famille de gènes (un COG par exemple) aura une importance dans l'écosystème étudié (Tringe et al., 2005).

La manière la plus répandue d'estimer le contenu taxonomique d'une communauté est d'utiliser des marqueurs phylogénétiques comme le gène codant pour l'ARNr 16S. Les séquences sont regroupées entre elles en unités taxonomiques opérationnelles (OTUs) grâce à DOTUR (Schloss and Handelsman, 2005). Une séquence représentative est ensuite utilisée pour assigner les OTUs à un taxon grâce par exemple à une requête sur la base RDP II (Cole et al., 2005).

Le regroupement en OTUs nécessite la génération d'une matrice de distances nucléotidiques entre les séquences, calculée par exemple avec la suite de logiciels Phylip (Felsentein, 1989). Cependant avec l'augmentation exponentielle des données, notamment avec l'arrivée du séquençage haut débit, le temps de calcul augmente dramatiquement et de plus en plus d'algorithmes utilisent la parallélisation des flux de données sur plusieurs processeurs de manière à diminuer le délai d'obtention du résultat (Sun et al., 2009).

Une autre manière d'estimer la diversité taxonomique est de réaliser un sondage plus flexible de la communauté à partir de séquences métagénomiques. Là où un sondage « plat » déduit un rang taxonomique à partir d'un marqueur phylogénétique de confiance, un sondage « flexible » déduit des séquences différentes taxonomies dépendant du niveau de conservation des séquences. Cette approche peut être réalisée avec le logiciel MEGAN (Huson et al., 2007). Cependant, cette méthode nécessite l'utilisation d'une base de référence servant à assigner par Blast chaque séquence à un taxon afin de les trier pour effectuer des approches comparatives. On se retrouve confronté au problème dit de l'ADN « sombre », c'est-à-dire des séquences qu'on ne peut assigner et c'est dans ce sens qu'il faut plus de génomes de référence.

Il est aussi possible de regrouper sans *a priori* les gènes codant pour des protéines en unités opérationnelles en utilisant une distance de dissimilarité plus faible que l'ARNr 16S (Li and Godzik, 2006; Schloss and Handelsman, 2008). Le défi consiste à regrouper entre elles des séquences incomplètes codant pour le même gène mais qui ne s'alignent pas.

5.3.5 Les bases de données relationnelles

Les bases de données relationnelles sont des outils indispensables pour l'écologie moléculaire. De plus, ces bases de données sont relationnelles, elles stockent l'information de manière optimale et donnent aussi des informations sur la nature des échantillons. Reliées entre elles, ces bases de données permettent un gain de temps pour assigner rapidement une fonction ou un taxon à une séquence, mais surtout elles permettent de prendre du recul sur l'information engendrée pour en

retenir les interconnections biologiques. Ces interconnections peuvent être de nature taxonomique (RDP II), métabolique (KEGG) ou bien fonctionnelle (COG, STRING).

Le gène de l'ARNr 16S est très conservé entre tous les microorganismes, d'une longueur convenable (environ 1500 pb) pour une analyse en bioinformatique, et est une excellente molécule pour discerner l'évolution des relations entre les organismes procaryotes. Pour toutes ces raisons, cette molécule a donné lieu à une énorme base de données publique, la « Ribosomal Database Project II » (RDP II). Le 5 octobre 2009, la base RDP version 10 contenait 1 104 383 séquences de gènes d'ARNr 16S. 180 573 proviennent de souches cultivées tandis que 923 810 proviennent d'échantillons environnementaux. 5 534 séquences proviennent de souches types. Ces dernières sont particulièrement importantes car elles permettent de relier taxonomie et phylogénie. Un des nombreux logiciels développés par l'équipe de Cole est le « RDP classifier », très efficace pour assigner jusqu'au genre avec des indices de confiance les séquences à la volée en très peu de temps (Wang et al., 2007). Les fichiers de sortie sont très facilement utilisables et permettent d'avoir un aperçu rapide de la diversité taxonomique de l'échantillon. Cependant, ils ne permettent pas de regrouper les séquences sous forme d'OTUs, démarche indispensable pour aller plus loin en écologie numérique (Voir la partie « Bio-statistique et Ecologie numérique », page 58).

La base de données KEGG pour « Kyoto Encyclopedia of Genes and Genomes » est une base de connaissance pour l'analyse des fonctions des gènes en terme de voies métaboliques (Ogata et al., 1999). Cette base, en plus de maintenir l'effort de collection de nouvelles voies métaboliques et d'intégrer de nouveaux gènes provenant des génomes annotés, développe et fournit des outils pour reconstruire les voies métaboliques en jeu dans un génome. Avec la métagénomique, cette base de données KEGG a pris une autre dimension puisqu'elle permet de synthétiser rapidement l'information métabolique d'un microbiome. De plus, elle permet de réaliser des analyses statistiques centrées sur l'interaction métabolique entre les gènes détectés dans un métagénome (Voir la partie « Bio-statistique et Ecologie numérique », page 58). Conçue pour la génomique couplée à la métabolomique, cette base souffre d'un déficit d'assignation puisqu'une grande partie des métagénomes séquencés, dont près d'un tiers pour le microbiote intestinal, n'est pas utilisable.

Dans ce contexte où la plupart des protéines répertoriées issues du séquençage restent de fonction inconnue, les COGs, répertoriés dans une base du même nom, semblent être un moyen très utile pour la prédiction de fonctions. Actuellement, la construction de ces COGs est basée sur les séquences de 66 génomes complets, dont 50 bactériens, issus de grands groupes phylogénétiques (Tatusov et al., 2001). Tout d'abord, la comparaison de ces séquences par paires a permis de créer un réseau de protéines orthologues ou COG « spécialisé » dans une fonction unique. Ainsi, la méthode des COGs, en regroupant des protéines d'espèces distantes, de fonction connue ou

inconnue, s'appuie sur le haut degré de conservation des séquences protéiques pour réaliser ces prédictions. Comme la base KEGG, les COGs souffrent d'un manque de représentativité au regard des séquences issues de la métagénomique, et près de 20 % des séquences codant potentiellement pour un gène sont répertoriées dans les catégories COGs très peu caractérisées, comme R « fonctions inconnues » et S « Fonction générale de prédiction seulement » (Tableau 4).

Tableau 4 : les grandes catégories de COG et leur description

Code	Catégories	Description
A	Modification et processus des ARN	Processus et stockage de l'information
B	Dynamique et structure de la Chromatine	
J	Traduction	
K	Transcription	
L	Réparation et réplication de l'ADN	
Y	Structure nucléaire	
D	Mitose et contrôle du cycle cellulaire	Processus cellulaires
O	Modification post-traductionnelle, fonction chaperonne	
M	Biogénèse de la membrane et de la paroi cellulaire	
N	Mobilité cellulaire	
P	Métabolisme et transport des ions inorganiques	
T	Transduction du signal	
U	Sécrétion et trafic intracellulaire	
Z	Cytosquelette	
C	Conversion et production d'énergie	Métabolisme
E	Transport et métabolisme des acides aminés	
F	Transport et métabolisme des nucléotides	
G	Transport et métabolisme des glucides	
H	Métabolisme des coenzymes	
I	Métabolisme des Lipides	
Q	Biosynthèse des métabolites secondaires	
R	Fonctions générales prédictives seulement	Très peu caractérisées
S	Fonctions inconnues	

En complément de ces bases de données, la base de données STRING fournit une ressource agrégeant la plupart de l'information disponible sur les interactions entre les protéines (Figure 8). La mise en œuvre des connections entre les protéines tient compte non seulement de leur homologie de séquence ainsi que de leurs occurrence et position dans les génomes séquencés, mais aussi des bases externes de données telles que KEGG, « Gene Ontology » et de l'exploration des données issues des publications. Ainsi, un score de partenariat fonctionnel est établi en

fonction de tous ces paramètres et permet de relier les protéines entre elles. Les informations que l'on retire de cette base peuvent être reliées avec d'autres bases comme ExPASy[15], SMART[16] afin d'affiner l'exploration fonctionnelle d'une protéine particulière.

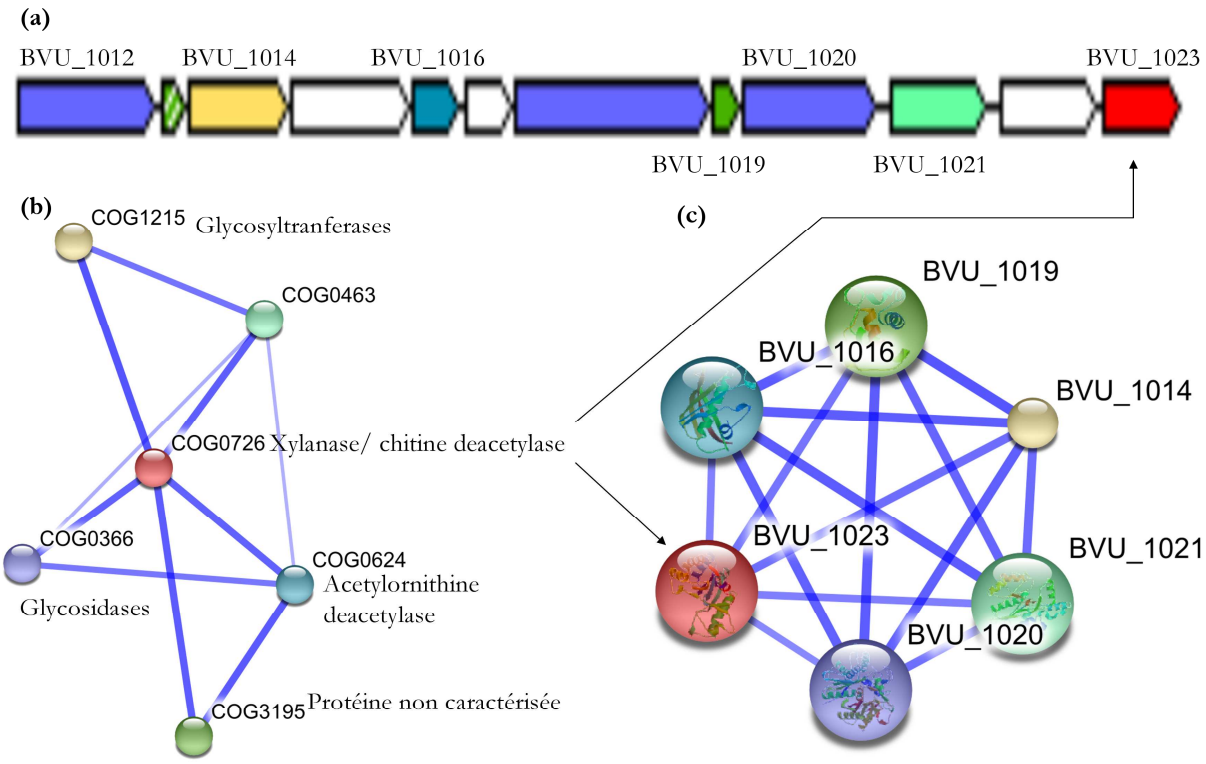


Figure 8 : Exemple d'utilisation de la base de données STRING avec une protéine xylanase / chitine deacetylase et le génome de *Bacteroides vulgatus*. (a) Cette protéine annotée « BVU_1023 » dans le génome de *B. vulgatus* est impliquée dans la dégradation des xylanes, elle permet notamment l'hydrolyse des liaisons carbone-azote. **(b)** Cette protéine fait partie du COG0726 et possède des partenaires fonctionnels tels que des glycotransferases impliquées dans la biogénèse de la membrane cellulaire, et des glycosidases comme par exemple des α-amylases. **(c)** La proximité dans le génome avec d'autres protéines chez *B.vulgatus* laisse supposer une structure en opéron par exemple.

5.4 Bio-statistique et Ecologie numérique

5.4.1 Ecologie numérique « classique »

Le concept théorique actuel en écologie est celui du modèle des habitats. Ce concept voit cet habitat comme un modèle pour les réponses écologiques et traite de la relation entre l'évolution

[15] La base « Expert Protein Analysis Sytem » ou ExPASy est disponible sur : http://www.expasy.ch/

[16] « Simple Modular Architecture Research Tool » ou SMART : http://smart.embl.de/

des espèces et les conditions de cet habitat (Dolédec et al., 1996; Legendre and Legendre, 1998). Ceci suppose que l'habitat (par exemple le microbiome intestinal humain) fournisse des conditions telles que les forces de l'évolution puissent s'exercer sur les traits des espèces de l'écosystème (par exemple sur le microbiote intestinal humain). C'est ici qu'intervient l'écologie numérique, c'est-à-dire à la frontière entre écologie et statistique. L'écologie numérique est un champ de l'écologie quantitative consacré à l'analyse numérique de données écologiques. Le but de l'écologie numérique est de décrire et d'interpréter la structure des données en combinant une large variété d'approches numériques (Legendre and Legendre, 1998). L'écologie numérique diffère de la bio-statistique descriptive dans le sens où cette discipline combine systématiquement des méthodes statistiques multivariées avec des techniques numériques non-statistiques comme les analyses par regroupement (« clustering »).

Par exemple, pour investiguer les relations entre la structure d'une communauté et les changements environnementaux, les écologistes collectent l'abondance des espèces dans un plan d'échantillonnage et enregistrent les variables environnementales dans ce même plan d'échantillonnage. Cela conduit à l'obtention de deux types de données. D'une part, un inventaire des espèces qui contient l'abondance des espèces en fonction de l'échantillon (tableau L), et d'autre part un tableau environnemental incluant des mesures quantitatives ou qualitatives des sites de prélèvement (tableau R). Une des tâches consiste alors à arranger les échantillons et/ou les espèces le long d'un gradient environnemental et d'en tirer un motif qui permet cet arrangement (Dolédec et al., 1996).

Selon la question posée, plusieurs analyses statistiques multivariées sont disponibles. Les analyses canoniques de correspondance se focalisent sur l'occurrence des espèces en fonction de l'environnement, quand la régression PLS (« partial least squares ») cherche à prédire des variables environnementales en fonction de l'assemblage des espèces. Les analyses de co-inertie et les analyses en composantes principales sur variables instrumentales (ACPVI) permettent d'étudier le lien de variation conjointe entre l'abondance des espèces et les variables de l'environnement (Dolédec and Chessel, 1994). Cependant, le concept du modèle des habitats qui prend en compte les forces de l'évolution requiert également de s'intéresser aux variations génétiques des espèces étudiées. La mesure des ces variations génétiques est devenue très résolutive avec le développement de la biologie moléculaire et des technologies de séquençage. Ainsi, un troisième type de données peut donc être pris en compte comme une matrice de distance génétique entre les différents taxons présents dans l'écosystème (tableau Q).

La relation entre la variation génétique des espèces et leur abondance dans un échantillon peut être analysée selon plusieurs critères comme les indices de diversité classique, tels que les indices

de Simpson et de Shannon, qui peuvent être calculés facilement avec DOTUR lorsque qu'on réalise des inventaires moléculaires basés sur le gène de l'ARNr 16S par exemple (Schloss and Handelsman, 2005). L'analyse de raréfaction et l'estimation de richesse par l'indice de Chao permettent entre autres de connaître l'effort d'échantillonnage d'une communauté. L'analyse d'arbres phylogénétiques en fonction des échantillons permet aussi d'observer la pression de sélection sur les différentes lignées (Voir Figure 1 dans la partie « A l'échelle de l'évolution », page 21). Le lien entre la topologie des arbres phylogénétiques et la niche écologique des espèces peut être testé avec un test de permutation (Martin, 2002; Schloss and Handelsman, 2006). L'analyse moléculaire de la variance (ou AMOVA) permet de tester si deux communautés ont une diversité significativement distincte (Chessel, 2004; Pavoine et al., 2004; Schloss, 2008). En complément, l'analyse moléculaire de l'homoscedasticité de la variance permet de connaître si une population est une sous-population par rapport à une autre (Schloss, 2008). Si l'analyse en coordonnées principales (PCoA) permet de visualiser les relations génétiques principales entre les taxons, la double analyse en coordonnées principales (dPCoA) permet quant à elle de relier une PCoA et une table d'abondance des espèces en fonction de l'échantillon (Pavoine et al., 2004; Eckburg et al., 2005).

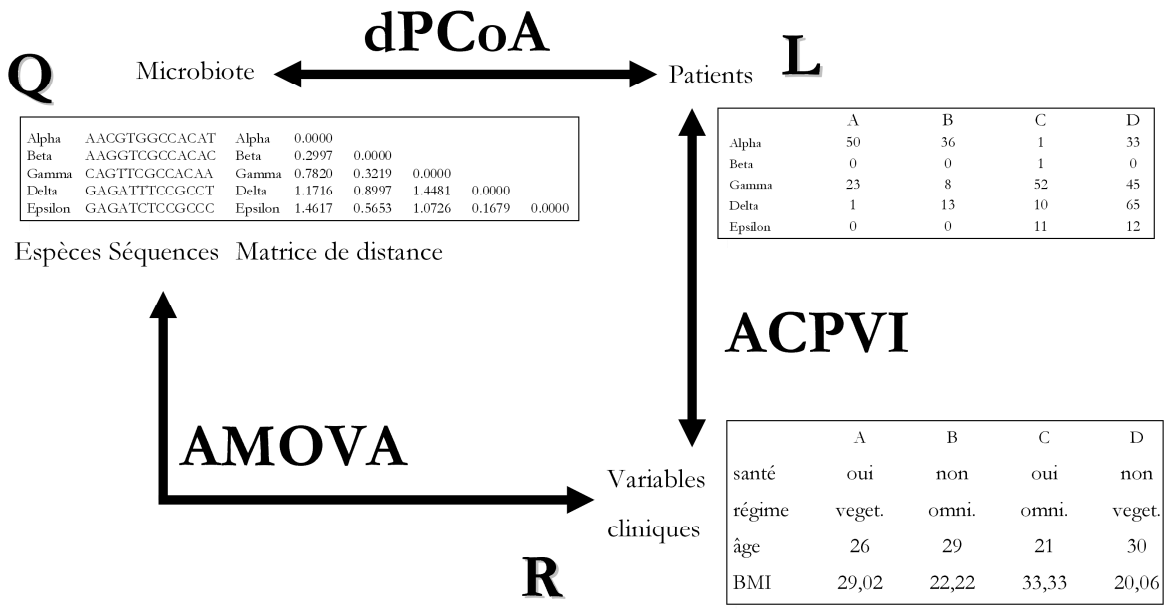


Figure 9 : Schéma d'un exemple d'analyse RLQ entre le microbiote, les patients et les variables cliniques. Le tableau R est un tableau de facteurs environnementaux (variables cliniques). Le tableau L est la composition en espèces et/ou fonctions décrivant un habitat (le microbiome des patients). Les données Q sont les relations génétiques ou fonctionnelles entre les protéines et/ou taxons (le microbiote). Ces différents types de données peuvent être reliés par quelques analyses statistiques comme par exemple l'AMOVA, l'ACPVI et la dPCoA. L'analyse globale RLQ peut être réalisée en effectuant une analyse des inerties des données Q et R reliées par la table de contingence L (Dolédec et al., 1996).

Lorsque l'enjeu est de connaître comment la biodiversité génétique est impactée par l'environnement, ce qui est le cas si l'on veut analyser finement le potentiel génétique du microbiote intestinal humain dans le cadre de la prise alimentaire, il est nécessaire de faire l'analyse conjointe de ces trois types de données. Pour répondre à cette problématique, une analyse RLQ semble très appropriée (Chessel, 1996). Cette technique permet d'incorporer dans une analyse la relation entre l'abondance des espèces, leur environnement, et leurs traits génétiques (Figure 9). Dans une approche métagénomique, on pourrait relier la diversité génétique et fonctionnelle avec l'abondance des espèces et de leur environnement, ou en d'autres termes, des changements fonctionnels microbiens peuvent être reliés par leurs impacts sur l'écosystème intestinal et la santé de l'hôte.

5.4.2 La bio-statistique appliquée à la microbiomique

Si l'écologie numérique est applicable à des sujets tels que l'impact des saisons sur la diversité des oiseaux ou bien la distribution géographique des poissons en fonction des stations d'épuration, il y a encore un effort à fournir lorsqu'il s'agit d'appliquer ces concepts au microbiome humain. Face à l'approche métagénomique, l'écologie numérique se retrouve confrontée à deux problèmes. Premièrement, ces concepts sont basés sur la notion d'espèce dont la définition est claire pour les plantes et les animaux mais qui provoque de grands débats lorsqu'il s'agit de bactéries. En effet, même si l'ARNr 16S constitue une norme pour définir une espèce, la précaution impose désormais de parler de phylotypes car on sait que des mécanismes d'échanges génétiques sont largement utilisés entre les bactéries de genres voire de phyla différents, ce qui est impossible pour les animaux et les plantes. Deuxièmement, l'écologie numérique se retrouve confrontée à des problèmes d'ordre de grandeur en ce qui concerne la mesure de la diversité génétique des espèces comparée par exemple aux variables cliniques ou au nombre d'échantillons. Concrètement, là où pour une étude sur l'abondance des poissons, on étudie une trentaine de sites avec 10 espèces et moins d'une dizaine de variables environnementales (Dolédec and Chessel, 1994), avec la métagénomique le nombre d'espèces à étudier passe à plus d'un millier. Si l'on reprend le schéma d'analyse RLQ, la métagénomique provoque une grosse distorsion sur les données du tableau L.

Avec la microbiomique, plusieurs publications ont commencé à apporter une réflexion pour adapter les bio-statistiques et l'écologie numérique aux technologies à haut débit (Dinsdale et al., 2008; Kristiansson et al., 2009; White et al., 2009). L'équipe de F. Rohwer a permis de montrer que l'analyse canonique discriminante (CDA) est très puissante pour séparer neuf microbiomes différents regroupant 45 microbiotes au total (Dinsdale et al., 2008). La CDA est une analyse factorielle des correspondances (ici les séquences en fonction des métagénomes) sous contrainte

d'une variable qualitative (les métagénomes en fonction de leur microbiome). Cependant, la CDA est une méthode d'identification de variables discriminantes entre les différents groupes, qu'il faut ensuite tester par des analyses de variance plus classiques. Un autre intérêt de la CDA décrit dans cette étude est de pouvoir construire un modèle prédictif pour classer les métagénomes en fonction de leur microbiome.

La CDA a montré son utilité à séparer des microbiomes très différents, du microbiome humain au microbiome du moustique. Néanmoins, identifier des marqueurs fonctionnels ou phylogénétiques du microbiote intestinal humain en fonction d'un effet clinique ou d'une étude nutritionnelle, demande l'utilisation de méthodes d'analyse utilisant des techniques de ré-échantillonnage et de permutation. J.R. White, avec son script R[17] nommé « Metastats »[18], a adapté un test de Student multiple pour évaluer l'effet d'un traitement sur l'abondance d'un taxon ou d'une protéine détectée. Cependant, comme le nombre de tests à réaliser dépend d'un nombre de taxa détectés (plusieurs centaines par échantillon du microbiote intestinal humain), il est nécessaire de faire une correction. C'est pourquoi, il est nécessaire de calculer en parallèle le taux de fausses découvertes qui est défini comme la proportion de faux positifs dans un ensemble de prévisions. Pour cela, des permutations statistiques sont réalisées pour tester la significativité du test de Student (White et al., 2009).

Hugenholtz et ses collègues ont développé une bibliothèque de fonctions[19] utilisables avec le langage R pour effectuer des comparaisons fonctionnelles de métagénomes. Les comparaisons de métagénomes reprennent le principe de « Metastats », excepté que l'analyse est rendue possible en tenant compte par exemple des voies métaboliques basées sur KEGG ou de familles de gènes basées sur les COGs. De plus, de nouvelles fonctions sont basées sur des modèles poissonniens, ce qui permet une flexibilité dans l'analyse de différents plans d'expérience comme des comparaisons par paires ou bien des dynamiques dans le temps.

[17] Le langage R est un langage de programmation et un environnement mathématique utilisés pour le traitement de données et l'analyse statistique. http://www.r-project.org/

[18] Une interface web de Metastats est disponible sur : http://metastats.cbcb.umd.edu/

[19] ShotgunFunctionalizeR disponible en téléchargement sur : http://shotgun.zool.gu.se/

6 RESULTATS ET DISCUSSION DU PROJET DE THESE

Cette thèse s'inscrit dans le projet « AlimIntest » financé par l'Agence Nationale pour la Recherche. Le projet « AlimIntest » a pour objectifs d'une part de développer de nouveaux outils moléculaires pour l'étude du microbiote intestinal et d'autre part de les valider sur une étude clinique nutritionnelle. Cette étude, menée par le centre d'investigation clinique de l'hôpital universitaire de Grenoble, teste l'impact de deux régimes contrôlés variant selon leur teneur en fibres (10 g et 40 g de fibres par jour) sur des volontaires sains. Ces régimes ont été administrés à 20 volontaires, âgés de 18 à 25 ans, en cross-over randomisé et en double aveugle (Figure 10). Les deux phases de régime ont été séparées par une période de deux semaines. Les volontaires ont reçu chaque régime (trois repas par jour) pendant une période de cinq jours[20]. Pour constituer la fécathèque, les échantillons ont été collectés avant et après les deux périodes de régime. Pour réaliser un contrôle, des échantillons ont été récoltés une semaine avant le début des régimes et une semaine après la fin de l'étude. Tous les échantillons ont été étiquetés et stockés immédiatement à -80°C. Afin de tester la répétabilité technique des outils moléculaires, une partie des échantillons a été préparée en double. Par ailleurs, à l'occasion de la collecte, l'eau fécale a été extraite des échantillons par ultracentrifugation pour établir des profils d'acides gras à chaînes courtes.

En parallèle de l'étude clinique, de nouveaux outils ont été développés. Tout d'abord, un référentiel écologique basé sur un inventaire moléculaire du gène de l'ARN 16S a été créé (**Article 3**). Les outils moléculaires comme des systèmes de PCR quantitative (**Article 1**) et une puce phylogénétique ainsi que des outils bioinformatiques (**Article 2**) ont été validés à partir de ce référentiel. En complément, une banque métagénomique de 156 000 clones a été créée et criblée sur plusieurs fonctions hydrolytiques : glucanase, xylanase, pectinase, amylase, galactanase et fructanase. Les résultats de ce criblage ont permis, entre autres, de mettre en évidence de nouveaux modules hydrolytiques et d'utiliser ces modules pour dessiner de nouveaux systèmes qPCR.

Tous ces outils ont été testés sur l'étude clinique « AlimIntest ». Ces travaux de thèse s'appuient principalement sur le projet « AlimIntest » pour évaluer l'impact de l'alimentation sur le microbiote intestinal. Néanmoins, les outils développés au cours de ce projet qui ont un but

[20] La composition des repas à 10 g et 40 g de fibres par jour se situe en annexe.

générique, ont également servi à évaluer l'adaptation du microbiote pendant la perte de poids dans le cadre de maladies métaboliques telles que l'obésité (**Article 4**).

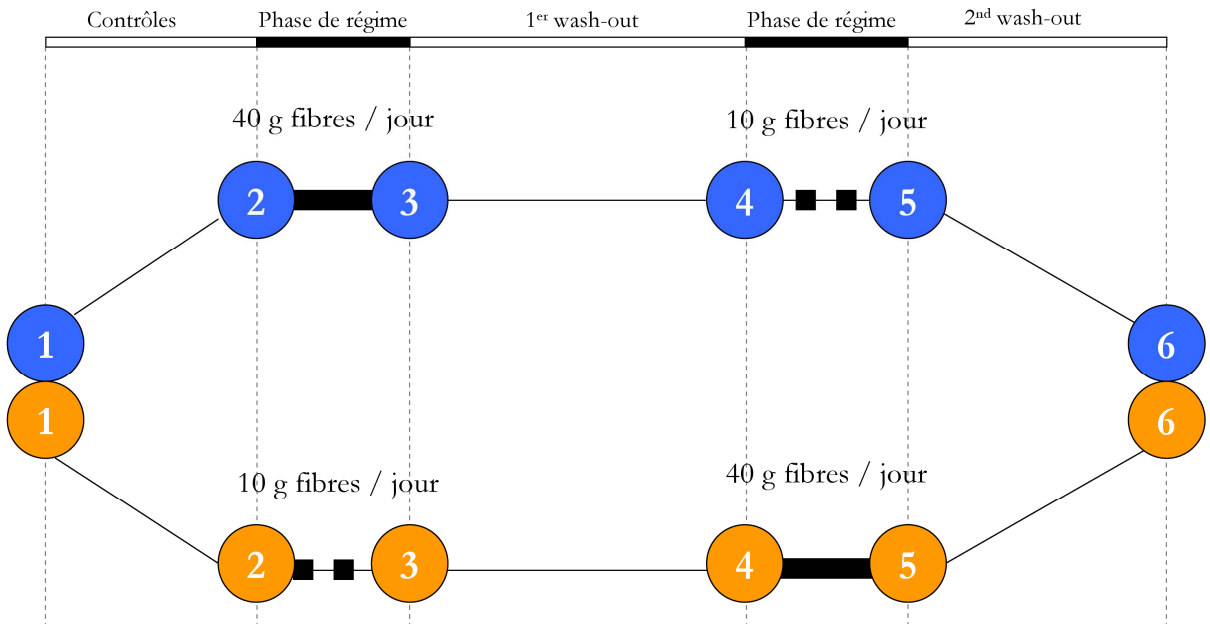


Figure 10 : Schéma de l'intervention clinique du projet AlimIntest. Chaque disque représente un point de collecte. Chaque phase de régime dure 5 jours. Les points n°1 et n°2 (avant le début de la phase clinique) sont séparés d'une semaine ainsi que les points n°5 et n°6. La première période de « wash out » dure 15 jours. Chaque groupe est composé de 10 individus.

Les résultats de ce projet de thèse se repartissent en trois parties. Tout d'abord, le développement de nouveaux outils moléculaires et bioinformatiques a été nécessaire pour répondre aux objectifs du projet « AlimIntest », mais également pour s'adapter à l'évolution des techniques de séquençage haut débit. Ensuite, la caractérisation écologique du microbiote par séquençage a débouché sur la mise en évidence d'un noyau phylogénétique. Pour finir, l'adaptation structurelle et fonctionnelle du microbiote a été évaluée en fonction du régime alimentaire.

6.1 Développement de nouveaux outils moléculaires et bioinformatiques

Au début des années 2000, l'exploration du microbiote s'effectuait essentiellement par des techniques électrophorétiques comme la TTGE et par le séquençage bas débit par méthode Sanger. Ces techniques permettent essentiellement de connaître le profil d'un échantillon d'un point de vue qualitatif. La qPCR sur le gène de l'ARNr 16S permet d'apporter un complément d'informations car elle permet de quantifier les populations bactériennes dans le microbiote. Mes premières contributions dans l'unité d'écologie du système digestif furent d'une part d'apporter un soutien bioinformatique pour l'élaboration des systèmes qPCR, et d'autre part de mettre en place une démarche d'analyse statistique pour en analyser les résultats.

6.1.1 Composition et activité physiologique du microbiote par PCR quantitative

Une autre variable, qui dans certains cas peut influencer la mesure et la comparaison de différents groupes bactériens, est la teneur en eau de chaque échantillon. Une faible teneur en eau pourrait contribuer à une forte concentration bactérienne. Afin de surmonter cette variable, les données sont normalisées en fonction de la population bactérienne totale.

Un objectif supplémentaire a été de développer un moyen pour quantifier l'activité transcriptionnelle du microbiote en effectuant de la qPCR sur les ARN totaux. Pour cela, Jean-Pierre Furet de l'UEPSD a développé une méthode pour extraire les ARN totaux. Après une rétro-transcription, nous utilisons les systèmes qPCR publiés (**Article 1**) pour évaluer la quantité de molécules d'ARNr 16S en fonction des groupes dominants du microbiote. Nous voulons utiliser cette quantification pour calculer le ratio ARN/ADN afin d'évaluer l'activité physiologique des groupes dominants du microbiote.

6.1.2 La méthode basée sur les tétranucléotides pour accélérer la détection des OTUs

Dans les premiers mois suivant mon arrivée, voyant augmenter la charge en séquençage de l'équipe, j'ai ressenti l'envie de connecter les logiciels existants pour détecter les OTUs dans des inventaires moléculaires. En collaboration avec Christophe Caron de l'unité MIG[21], une interface a été créée pour que chacun puisse faire la détection d'OTUs à partir d'un jeu de séquences issues d'inventaires moléculaires, de manière conviviale (Figure 11). Néanmoins, ayant rapidement en charge, avec l'utilisation de ce logiciel baptisé RapidOTU, plus de 20 000 séquences dans le projet AlimIntest et anticipant l'avènement de la technologie « 454 » dans le laboratoire, il fallait trouver une alternative aux algorithmes d'alignement. C'est pour cela que pour la première fois, j'ai décidé de connecter la fréquence des tétranucléotides d'une séquence donnée avec un logiciel de regroupement d'OTUs comme DOTUR. Les premiers résultats furent très encourageants. En plus de comparer les deux algorithmes pour valider cette nouvelle méthode, il a fallu également analyser comment la méthode utilisant les tétranucléotides se comportait avec des séquences plus courtes issues de la technologie « 454 ». Néanmoins, à l'avenir le séquençage haut débit évoluera de telle manière à obtenir des séquences aussi longues que la méthode Sanger (Voir la partie « Séquençage haut débit », page 47).

[21] MIG : Mathématique, Informatique et Génome

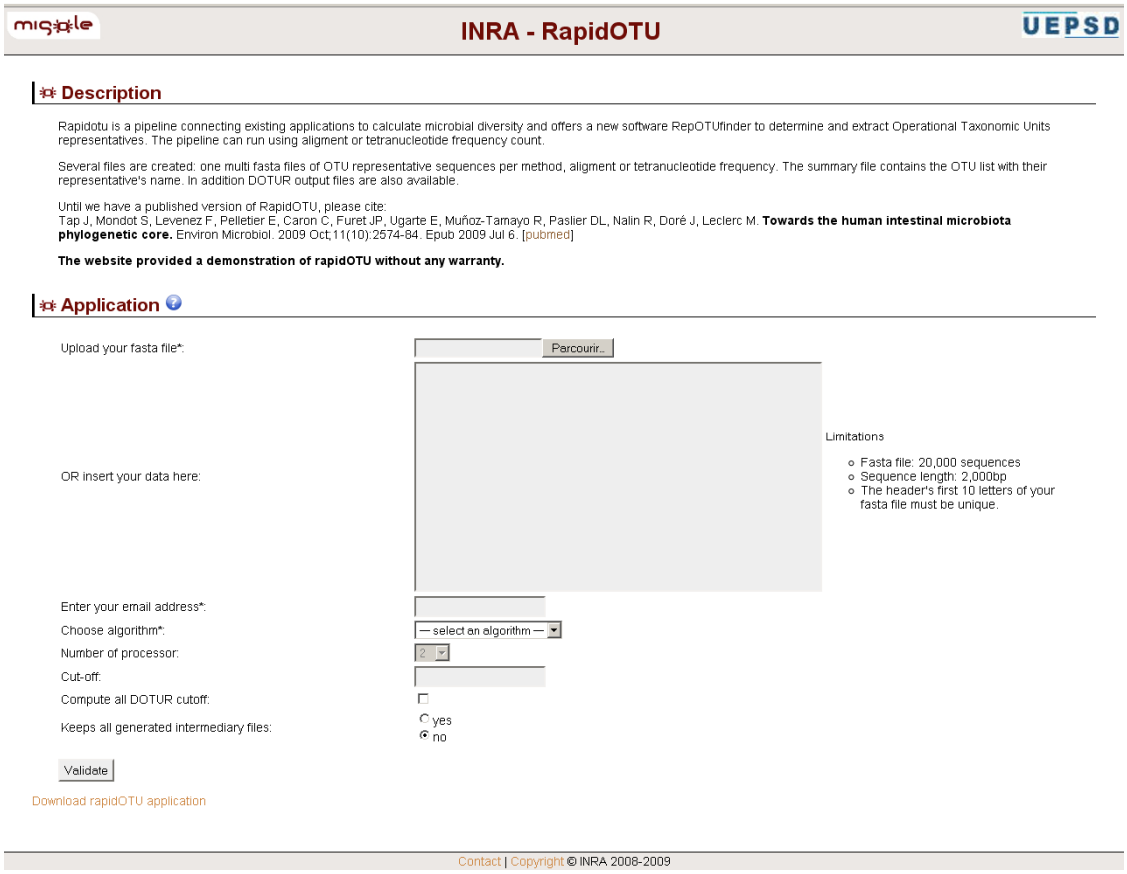


Figure 11 : Interface Web de RapidOTU http://genome.jouy.inra.fr/rapidotu. Via l'interface, l'utilisateur peut téléverser les séquences du gène d'ARNr 16S sur le serveur distant par un simple copier-coller ou en explorant son ordinateur. Les résultats lui seront envoyés par mail. Pour éviter une surcharge de la plateforme de calcul de l'INRA de Jouy, l'utilisateur peut analyser 20 000 séquences à la fois.

En tirant au hasard 5000 séquences dans la base RDP II et en sélectionnant les régions encadrant les parties variables V6-V8, on peut remarquer que la méthode utilisant les tétranucléotides conserve la diversité détectée au sein des 5000 séquences quelle que soit la longueur des séquences, alors que la méthode des alignements multiples sous-estime la diversité quand elle est appliquée à des séquences courtes (Figure 12).

Il était essentiel dès lors de démontrer que le regroupement d'OTUs avec les tétranucléotides était le même qu'avec la méthode utilisant des alignements (**Article 2**). En collaboration avec le Genoscope, nous avons élaboré une stratégie pour évaluer la similarité de regroupement entre l'algorithme basé sur l'alignement et celui basé sur les fréquences des tétranucléotides. La sensibilité et la spécificité de la méthode des tétranucléotides par rapport à celle basée sur l'alignement ont été évaluées. La sensibilité est la faculté de classer deux éléments dans la même catégorie lorsqu'ils le sont vraiment et la spécificité est la faculté de séparer deux éléments quand ils doivent être séparés. L'indice de Rand basé sur la spécificité et la sensibilité apparaît comme un bon indicateur pour juger les deux méthodes.

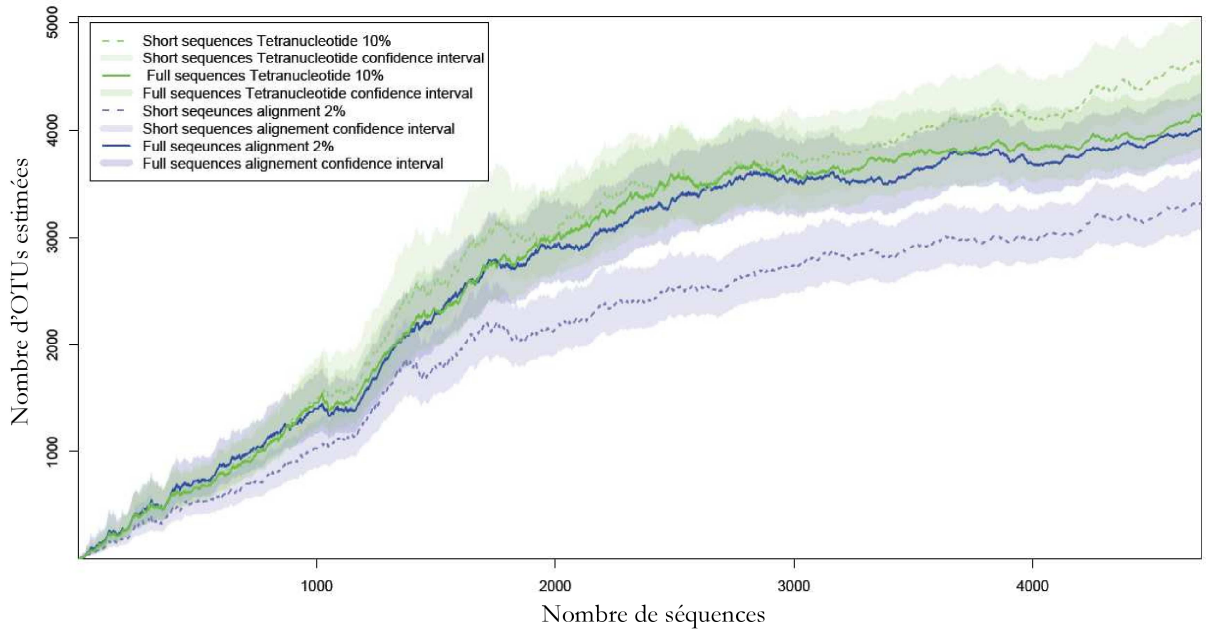


Figure 12 : Comparaison des différents algorithmes en fonction de la richesse estimée en OTUs et de la taille des séquences.

Cependant, cet indice n'avait jamais été testé sur une telle problématique, c'est pourquoi il a été nécessaire de tester sa fiabilité. En effectuant en parallèle des comparaisons d'échantillons indépendants et des comparaisons appariées, on peut voir que les comparaisons appariées donnent toujours un meilleur score que les comparaisons de tirages indépendants. Cela signifie que le score fourni par l'indice de Rand est un bon indicateur de similarité entre deux algorithmes de regroupement et que son score n'est pas dû au hasard (**Article 2**, Figure 3).

L'utilisation des fréquences de tétranucléotides pour comparer des séquences a bien évidemment d'autres avantages que la rapidité d'exécution. En effet, cette méthode est très conservatrice en ce sens que la distance qui sépare deux séquences sera toujours la même quel que soit le nombre de séquences à comparer. Ceci constitue un avantage décisif sur les algorithmes d'alignement qui peuvent fluctuer en fonction des séquences à aligner. Ainsi deux séquences n'auront pas la même distance si elles sont incluses dans des jeux de données différents. Avec la comparaison de plusieurs millions de séquences, l'utilisation des tétranucléotides pour comparer les séquences d'ARNr 16S entre elles paraît dorénavant incontournable.

6.1.3 Evaluation technique de l'utilisation du pyroséquençage sur le microbiote

C'est une collaboration avec le centre de recherche et développement de Nestlé, que les premières séquences de pyroséquençage sont arrivées dans l'équipe. J'ai pu tester RapidOTU sur des

réplicats techniques, c'est-à-dire cinq événements de séquençage sur le même échantillon. Alors que la proportion de phyla et de familles ne varie pas au sein des réplicats (**Article 5**), il semble qu'un nombre important d'OTUs ne soit pas détecté dans tous les réplicats.

A partir d'un échantillon fécal, une extraction ADN a été faite puis les régions V1-V2 et V4 du gène de l'ARNr 16S ont été séquencées avec la technologie « 454 » en cinq réplicats techniques. 8617 séquences ont été obtenues pour la région V1-V2 et 10522 séquences pour la région V4. Avec la méthode des tétranucléotides, RapidOTU permet de détecter 687 OTUs pour la région V1-V2 et 719 OTUs pour la région V4. 324 OTUs pour la région V1-V2 et 361 OTUS pour la région V4 ont été trouvées dans un seul réplicat sur les cinq. Ces OTUs dites « réplicats spécifiques » ont une abondance inférieure à 5 séquences quelle que soit la région du gène de l'ARNr 16S étudiée. 132 OTUs et 142 OTUs ont été retrouvées respectivement dans les cinq réplicats pour les régions V1-V2 et V4. De manière surprenante, 14 OTUs pour la région V1-V2 et 7 OTUs pour la région V4 ayant une abondance totale supérieure à 20 séquences n'ont pas été retrouvées dans tous les réplicats.

Ces informations nous renseignent que la répétabilité technique peut engendrer un biais dans l'analyse et que les OTUs détectées dans un seul réplicat constituent un bruit de fond important représentant près de 50 % des OTUs détectées. De plus, seulement environ 20 % ont été retrouvées dans tous les réplicats et près de 2 % des OTUs détectées en abondance ne sont pas détectées dans tous les réplicats. Toutes ces observations sont en faveur de l'utilisation de réplicats techniques pour la technologie du « 454 » pour la réalisation d'inventaires moléculaires. Néanmoins, lorsqu'il n'est pas possible de réaliser des réplicats techniques, il est nécessaire de prendre en compte que près de 50 % des OTUs, généralement peu abondantes et quelle que soit la région du gène de l'ARNr 16S, peuvent être dues à l'aléatoire et non à l'échantillon étudié.

6.1.4 Méthodologie pour accéder aux ARN messagers

La technique d'extraction des ARN totaux mise au point par Jean-Pierre Furet de l'UEPSD permet d'avoir une quantité très importante d'acides nucléiques (jusqu'à 100 µg pour 200 mg d'échantillon fécal). Cependant, l'accès par séquençage aux ARNm qui représentent moins de 5 % des ARN totaux est très difficile. Pour l'instant, c'est le kit d'hybridation soustractive « Microbes express® » qui a été utilisé. Un kit de purification permettant d'enlever les acides faisant moins de 100 pb est utilisé. Son utilisation a aussi pour conséquence d'appauvrir l'échantillon en ARNr 5S.

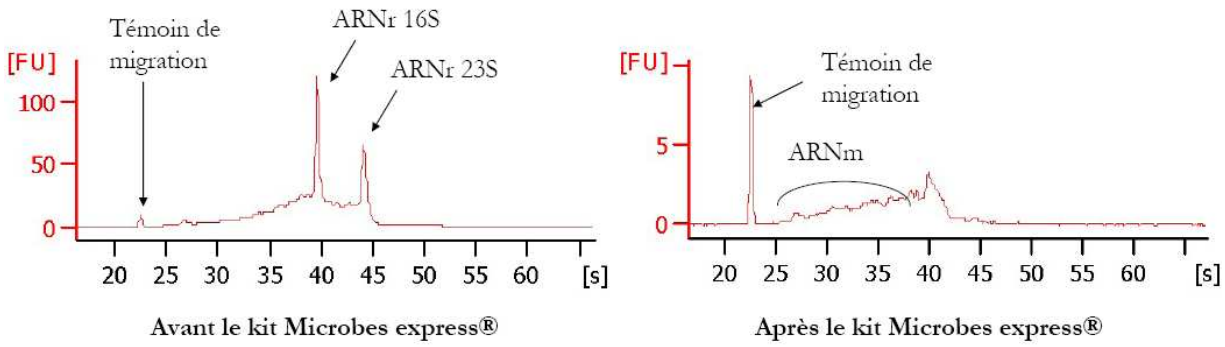


Figure 13 : Visualisation sur un profil de Bioanalyzer d'un échantillon d'ARN d'origine fécale avant et après l'utilisation du kit Microbes express®. Le produit d'une extraction ARN à partir d'un échantillon fécal est appauvri avec le kit d'hybridation soustractive en ARN ribosomaux. L'acide nucléique marqué migre en fonction de sa taille dans un capillaire. Le temps de migration, calibré par un témoin de migration, permet d'évaluer la taille des fragments nucléiques. Les deux pics illustrant la présence des ARNr 16S et 23S ne sont plus retrouvés après l'utilisation du kit.

Nous avons pu vérifier par Bioanalyzer d'Agilent (Figure 13), les profils des acides nucléiques avant et après l'utilisation du kit d'hybridation soustractive. De plus, une évaluation par qPCR des ARNr 16S, a permis de montrer la réduction d'un facteur 100 du nombre de copies d'ARNr 16S dans un échantillon.

6.2 Le microbiote est constitué d'un noyau phylogénétique

Le paradoxe que constitue l'hétérogénéité de la composition du microbiote intestinal et l'homogénéité fonctionnelle permettant le maintien de l'homéostasie chez les individus sains peut être expliqué par la présence d'un petit nombre d'espèces partagées par tous : le noyau phylogénétique du microbiote intestinal humain (**Article 3**). Pour caractériser ce noyau phylogénétique, plus de 10 000 séquences d'ARNr 16S ont été analysées. Elles sont issues d'échantillons provenant de 17 individus ayant des régimes variés, allant des régimes omnivores et à des régimes végétariens stricts. Un petit nombre de phylotypes représentant 2 % du nombre total d'OTUs détectées représente plus du tiers des séquences analysées. De plus, ces phylotypes sont partagés par la moitié des individus. Une OTU assignée à *F. prausnitzii* est partagée par 16 individus sur 17. D'autre part, un inventaire plus important de 5 000 séquences sur l'individu « AT » ayant le moins contribué au noyau phylogénétique d'OTUs a permis de détecter les OTUs du noyau assignées à *F. prausnitzii* et apparentées au genre *Faecalibacterium*.

Un nombre important d'OTUs est proche d'espèces types qui ont été bien caractérisées comme *Roseburia intestinalis* ou bien *Bacteroides vulgatus* par exemple. Ce qui est intéressant, c'est qu'une large variété de fonctions métaboliques comme les chaînes trophiques du métabolisme des glucides allant de l'hydrolyse jusqu'à la production des AGCC, est couverte et peut être attribuée en grande

partie à ces phylotypes. La phylogénie buissonnante du microbiote intestinal est largement attribuée aux espèces du noyau phylogénétique. Le nombre important détecté d'OTUs assignées aux genres *Bacteroides*, *Faecalibacterium*, *Ruminococcus* et *Roseburia* est une indication importante sur la diversité pan-génomique potentielle des espèces affiliées à ces genres. Les quelques génomes séquencés des espèces du noyau phylogénétique indiquent un potentiel adéquat pour la fermentation des fibres alimentaires et la plupart de ces espèces sont de fortes productrices d'AGCC. Il reste néanmoins un nombre d'OTUs très peu caractérisées dont le potentiel génétique reste à découvrir, notamment dans la famille des Lachnospiraceae et Ruminococcaceae.

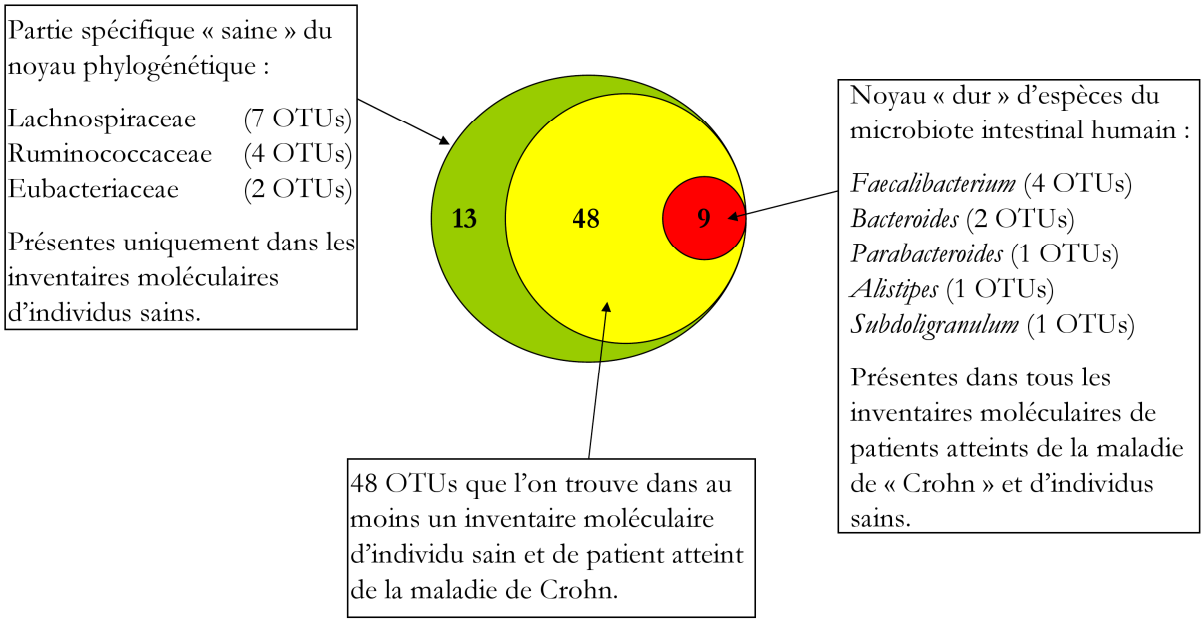


Figure 14 : Comparaison du noyau phylogénétique avec des inventaires de patients atteints de la maladie de Crohn. Les OTUs de trois et quatre inventaires moléculaires de patients atteints de la maladie de Crohn (Lepage et al., 2005; Gophna et al., 2006; Manichanh et al., 2006) et d'individus sains (Eckburg et al., 2005; Gill et al., 2006; Manichanh et al., 2006; Li et al., 2008) ont été comparées par Blast aux espèces du noyau phylogénétique.

Malgré les différentes méthodologies, ces 66 OTUs issues d'individus français et néerlandais (n=17) ont toutes été retrouvées dans les autres inventaires moléculaires issus d'individus sains américains (n=5) (Eckburg et al., 2005; Gill et al., 2006) et chinois (n=5) (Li et al., 2008). Cela supporte le concept du noyau phylogénétique du microbiote intestinal à travers une grande partie de l'humanité. La caractérisation de ce noyau devra être supportée par des analyses à grande échelle aussi bien d'un point de vue géographique que d'un point de vue de la profondeur de séquençage. Puisque le noyau phylogénétique issu d'individus sains supporte la robustesse fonctionnelle du microbiome intestinal humain, il est intéressant de tester sa présence en cas de dysbiose comme c'est le cas dans les maladies inflammatoires telles que la maladie de Crohn (Manichanh et al., 2006).

En effet, lorsque l'on teste la présence des 66 OTUs du noyau phylogénétique dans les inventaires de patients atteints de la maladie de Crohn, 13 OTUs constituent une partie « saine » spécifique de ce noyau par rapport à cette maladie (Figure 14). Ces OTUs sont principalement des Lachnospiraceae du genre *Roseburia* et des Ruminococcaceae comme par exemple l'espèce *Oscillibacter valericigenes*. Par ailleurs, cette OTU apparentée à *O. valericigenes* a été très peu détectée par qPCR dans une cohorte constituée de 16 individus atteints de la maladie de Crohn par rapport aux individus sains (Mondot et al., données non publiées).

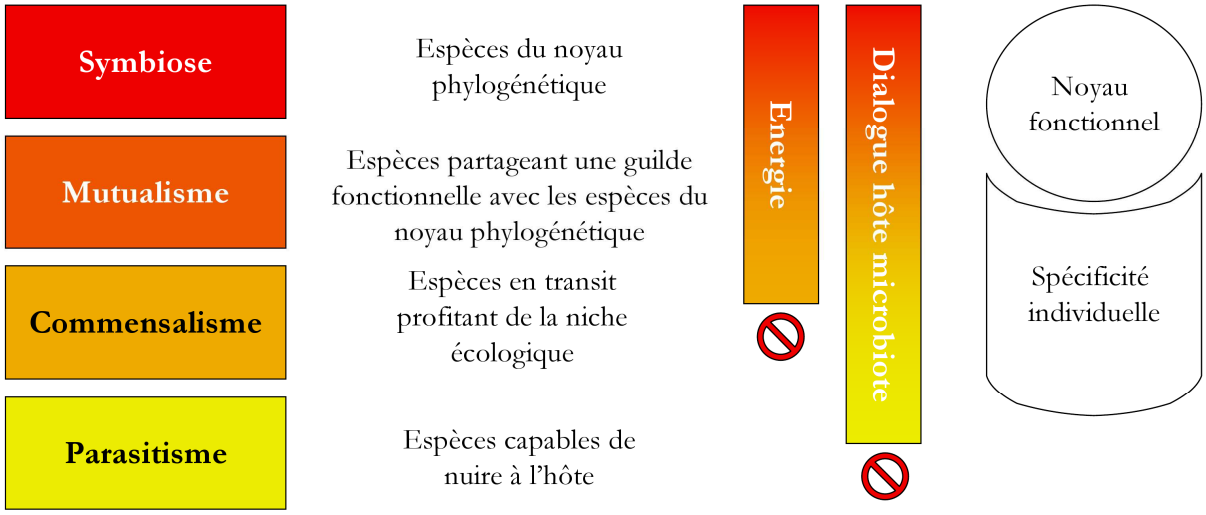


Figure 15 : Schéma de l'organisation structurelle du microbiote intestinal humain. Les espèces du noyau phylogénétique et celles faisant partie de leurs guildes fonctionnelles respectives sont prépondérantes dans le noyau fonctionnel du microbiote intestinal humain. Elles peuvent facilement acquérir de l'énergie et participent pleinement au dialogue hôte microbiote. Les espèces mutualistes qui ne sont pas partagées par tous participent en grande partie aussi à la spécificité individuelle. A contrario, les bactéries commensales qui profitent de la niche écologique sont le plus souvent en transit dans le tractus en ayant une compétitivité plus faible pour acquérir de l'énergie par rapport aux espèces du noyau phylogénétique. Néanmoins, elles peuvent participer au dialogue hôte microbiote et à la spécificité individuelle. Les espèces parasites étant capables de nuire sont expulsées de la niche écologique et ne participent pas au dialogue hôte microbiote.

Le concept du noyau phylogénétique n'est pas incompatible avec le concept du noyau fonctionnel proposé par Turnbaugh et ses collègues. Pour soutenir le principe du noyau fonctionnel, Turnbaugh propose que l'homéostasie du microbiote soit portée par une structure en guilde fonctionnelle avec des espèces interchangeables pour chaque fonction du microbiote. Chaque guilde fonctionnelle serait en mutualisme avec l'hôte. Le concept du noyau phylogénétique va plus loin en proposant des espèces en symbiose avec l'hôte partageant des niches écologiques communes avec ces guildes fonctionnelles. En termes d'évolution, les guildes fonctionnelles dérivent de ces espèces symbiotiques en formant des buissons phylogénétiques. Le noyau phylogénétique participe au maintien du bon fonctionnement de ces guildes fonctionnelles et en conséquence assure les fonctions partagées par tous les individus sains. Si le noyau phylogénétique

du microbiote est altéré, le noyau fonctionnel le sera aussi. La dysbiose fonctionnelle de l'écosystème intestinal va se refléter dans l'altération de ces guildes puis dans la disparition d'espèces du noyau.

Connaître l'impact de l'environnement, et en particulier du régime alimentaire, sur ce noyau d'espèces partagées par tous permettra d'en évaluer les bienfaits.

6.3 Impact des régimes omnivore et végétarien sur le microbiote

Parmi les 17 sujets, neufs se sont déclarés omnivores et huit se sont déclarés végétariens. Lorsque l'on compare les données de qPCR quantitative normalisées par le système « All Bacteria », on observe une différence significative entre les deux groupes au niveau de l'abondance des *Bacteroides* (Figure 16). Les autres systèmes qPCR ne permettent pas de mettre en évidence de différences significatives entre les omnivores et les végétariens. La PCR quantitative révèle également une grande variabilité entre les individus avec parfois des valeurs extrêmes en *F. prausnitzii* pour l'individu AT par exemple. Pour le genre *Bifidobacterium*, on observe des différences de l'ordre d'un facteur 100 entre certains individus du même groupe.

La faiblesse de l'analyse qPCR sur des grands groupes bactériens est que parfois les cibles sont très larges, en particulier pour les groupes *C. coccoides* et *C. leptum* couvrant une diversité bactérienne importante, si bien que les variations « intracibles » ne sont pas observables. Concrètement, le système qPCR ciblant le groupe *C. coccoides* ne permet pas d'observer de différence entre les *Lachnospiraceae* et les *Eubacteriaceae*. C'est pourquoi, l'inventaire moléculaire du gène de l'ARNr 16S semble être un bon outil pour évaluer, avec une résolution se situant au niveau de l'OTU, des différences entre les deux régimes.

L'analyse interclasses, qui est un cas particulier de l'ACPVI, entre les omnivores et les végétariens permet de comparer la fréquence de distribution des OTUs entre les deux régimes. Moins de 5% de la variabilité totale permet de discriminer significativement les deux groupes. Cela peut être mis à profit pour mettre en valeur les OTUs les plus discriminantes entre les deux régimes.

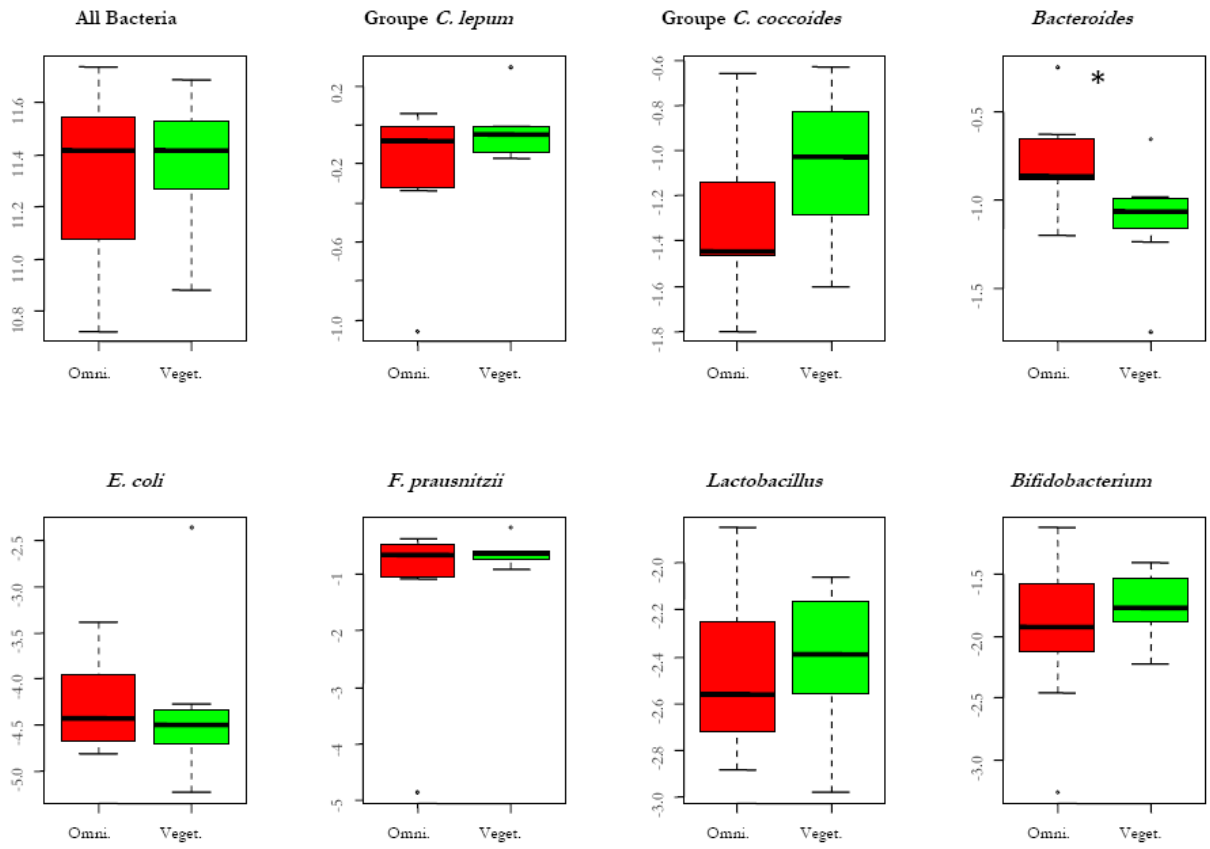


Figure 16 : Comparaison de la composition du microbiote de neuf volontaires sains omnivores et de huit volontaires sains végétariens par PCR quantitative. Les résultats sont normalisés par le système « All bacteria » : les résultats sont exprimés en log et en fonction de la différence entre l'abondance des groupes bactériens ciblés et l'abondance de toutes les bactéries ciblées par le système « All Bacteria ». Le système All Bacteria est exprimé en log équivalent génome d'*E. coli*. *Les omnivores sont enrichis en *Bacteroides* par rapport aux végétariens (p = 0,028).

L'analyse en coordonnées principales permet de séparer les OTUs en fonction de leur distance génétique (Figure 17). La distance représentée alors dans un espace à deux dimensions est euclidienne. Plus la distance entre les OTUs est importante, plus la distance génétique est grande. Combinée à l'analyse interclasses, l'analyse en composantes principales permet de confirmer une distribution en OTUs assignées au phylum Bacteroidetes en faveur des sujets omnivores, confirmant alors les résultats de qPCR. Même si la plupart des espèces du genre *Bacteroides* contiennent l'appareillage génomique pour dégrader certaines fibres alimentaires, supposées plus abondantes chez les végétariens, *B. fragilis* par exemple se cultive sur des milieux riches en peptones. Le régime végétarien profite peut-être aux espèces appartenant aux Firmicutes comme *R. intestinalis* et *R. bromii*, connues pour dégrader les polyosides complexes.

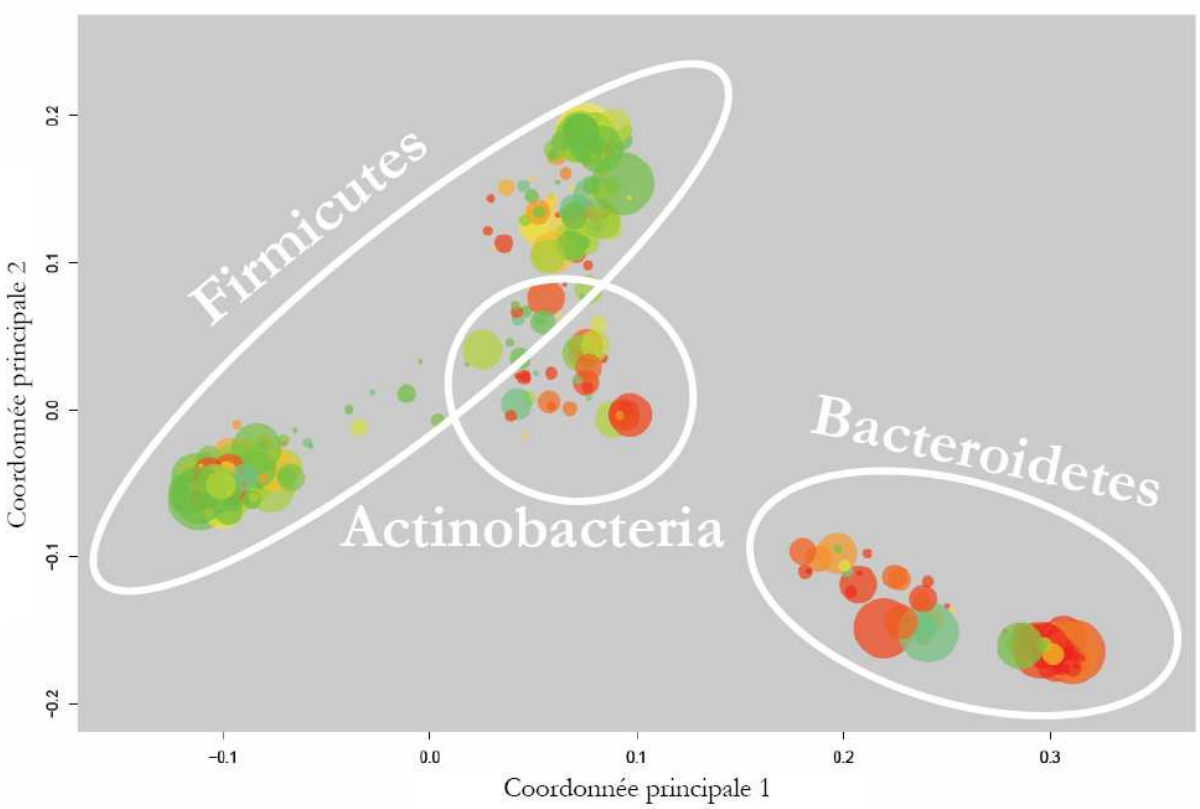


Figure 17 : Analyse en coordonnées principales des OTUs du microbiote fécal de 17 individus sains. L'analyse en coordonnées principales est réalisée à partir de la matrice de distance entre les séquences représentatives de chaque OTUs. Chaque OTU est symbolisée par un disque dont l'aire est proportionnelle au nombre total de séquences. Les couleurs allant du vert au rouge tiennent compte des scores de l'analyse interclasses des OTUs en fonction des deux régimes. La couleur rouge indique une OTU retrouvée plus abondamment dans les microbiotes des sujets omnivores tandis que la couleur verte indique une OTU retrouvée plus abondamment dans les microbiotes des sujets végétariens. Les OTUs représentant une seule séquence n'ont pas été représentées.

Au niveau de l'OTU, on peut faire la distinction entre deux Bifidobactéries, l'une assignée à *Bifidobacterium sp.*, et l'autre assignée à *B. longum*. La première est retrouvée exclusivement chez les omnivores tandis que la deuxième est enrichie chez les végétariens. Cela indique que pour des espèces du même genre, il existe des différences au niveau de certaines potentialités fonctionnelles.

6.4 L'apport en fibres impacte-t-il les fonctions du microbiote ?

Cette partie s'appuie sur l'étude clinique du projet AlimIntest menée par la Pr Eric Fontaine.

6.4.1 Structuration du microbiote par le régime

Avant le début de l'étude clinique, nous n'observons pas de variation significative des biomasses bactériennes lorsque que l'on prend en compte la mesure réalisée par le système de qPCR

« All Bacteria ». Entre le point n°1 et le point n°2, les variations observées à la fois entre les individus et dans le temps ne dépassent pas un facteur dix. Nous observons une variation totale de 11 à 11,8 log de bactéries totales par gramme d'échantillon avant la phase de régime. Les mesures extrêmes pendant les phases de régime ont été mesurées à 10,7 et à 12,3 log de bactéries totales. Les comparaisons appariées des individus ne donnent pas de résultats significativement différents en biomasse bactérienne pendant l'étude clinique.

Lorsque que l'on réalise une analyse en composantes principales en prenant comme variables instrumentales (Figure 18) les points de collecte de la dynamique du régime, un test de Monte Carlo indique que la variation du microbiote est bien structurée en fonction du régime ($p < 0,05$). Cette variation représente près de 14 % de la variation totale observée. Cette analyse révèle de la même manière que le microbiote est aussi structuré significativement en fonction des individus. Cette variation représente près de 50 % de la variation totale ($p < 0,001$). D'autre part, nous observons qu'avant l'intervention clinique, les points n°1 et n°2 sont très proches entre eux, illustrant que le microbiote de chaque individu est resté stable. La différence à l'état initial entre les deux groupes n'est pas significative.

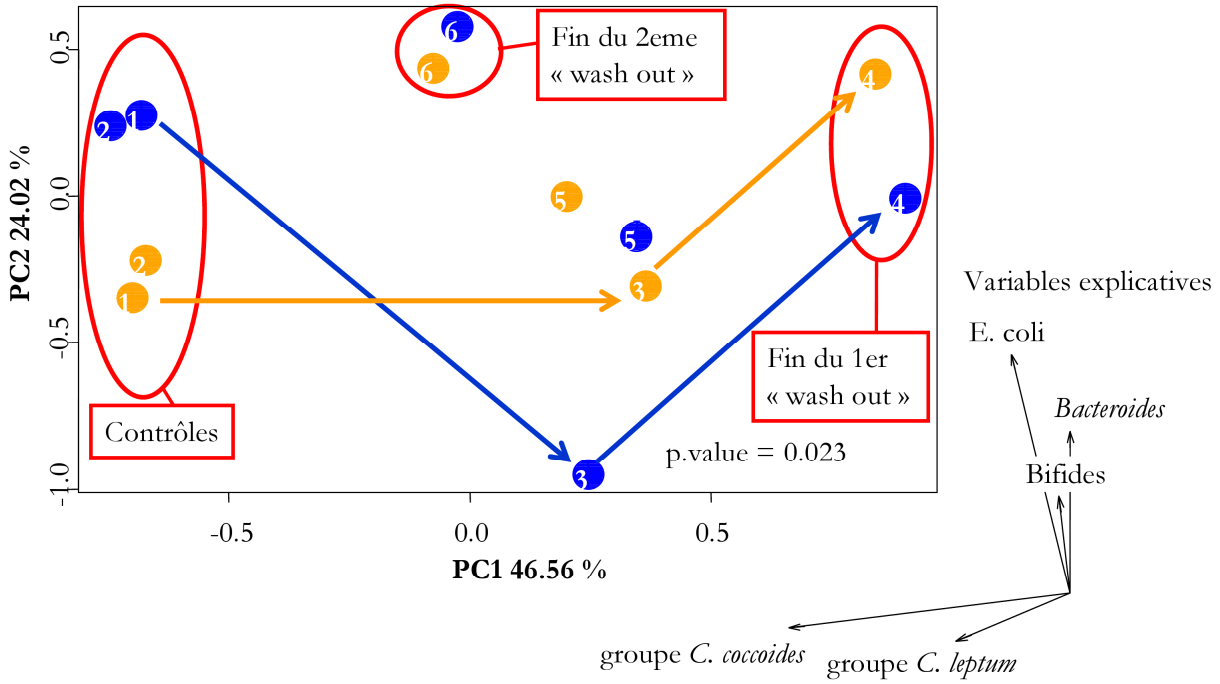


Figure 18 : Analyse en composantes principales avec la dynamique de l'étude clinique comme variables instrumentales. Chaque point correspond à un point de collecte de l'étude illustrée en Figure 10. Les variables explicatives correspondant au plan d'inertie sont illustrées sur la droite. Les points bleus correspondent aux individus ayant pris un régime de 40 g de fibres par jour en premier, tandis que les points oranges correspondent aux individus ayant pris 10 g de fibres par jour.

D'après l'ACPVI, au terme de la période où les individus reprennent pendant 15 jours leurs habitudes alimentaires après la première phase de régime (entre les points n°3 et n°4), le microbiote ne revient pas à l'état initial (points n°1 & 2). Par conséquent, la deuxième phase de régime est directement impactée par la première phase de régime et cette période de 15 jours n'est donc pas suffisante pour réaliser un « Wash out ».

Durant la première phase du régime, le groupe ayant reçu le régime riche en fibres (40 g de fibres par jour), est particulièrement perturbé sur la composante *E. coli* par rapport au groupe ayant reçu le régime à 10 g de fibres par jour. Cette chute d'*E. coli* est significative pour le premier groupe ($p < 0,01$). De manière surprenante, quelle que soit la séquence du régime (i.e. 40-10 ou 10-40), nous observons une baisse significative du groupe *C. coccoides* jusqu'à la deuxième phase du régime, compensée par une augmentation de la proportion de *E. coli* et du groupe *Bacteroides/Prevotella*. L'abondance du groupe *C. coccoides* augmente dès la deuxième phase du régime jusqu'à une semaine après la fin du régime.

La composition du microbiote intestinal est donc structurée significativement en fonction du régime alimentaire. Cependant, cet effet n'est pas observable significativement au niveau de son activité physiologique donnée par RT-PCR quantitative.

6.4.2 Profils des acides gras à chaînes courtes pendant les phases de régime

Il nous a ensuite semblé important de quantifier les acides gras à chaînes courtes afin d'estimer le rendement ou l'activité globale de fermentation. Pour cela, Catherine Philippe de l'UEPSD a dosé l'acétate, le propionate, le butyrate, le valérate, le caproate ainsi que les iso-acides respectifs.

Au temps initial, nous observons une grande hétérogénéité chez les individus. Les profils d'AGCC sont très variables. Les concentrations en AGCC dominants sont pour l'acétate de 5,66 à 60,1 mM, pour le propionate de 1,87 à 15,94 mM et pour le butyrate de 1,11 à 26 mM. Les ratios acétate : propionate : butyrate oscillent entre 62 : 10 : 26 et 44 : 20 : 36. Par ailleurs, des concentrations très distinctes en iso-acides sont également mesurées, avec des concentrations en iso-butyrate et iso-valérate allant jusqu'à 8 mM.

Quelle que soit la séquence du régime appliqué, 10-40 ou 40-10, les profils sont extrêmement variables dans le temps et en fonction des individus. Notamment, avec des réponses aux régimes différentes voire opposées en fonction des régimes et d'un « wash out » qui ne permet pas un retour à l'état initial. Par ailleurs, l'ACPVI ne permet pas de détecter une structuration significative de l'activité métabolique du microbiote en fonction de la dynamique du régime ($p > 0.05$). La variation inter-individus est tellement importante, que près de 55 % de la variabilité totale est

expliquée par l'individu ($p < 0{,}001$). A la fin de l'intervention clinique, les rapports acétate : propionate : butyrate oscillent entre 59 : 30 : 9 et 38 : 20 : 41.

Ces données indiquent également que la représentation en ratio d'AGCC semble montrer une plus grande homogénéité dans le temps, représentant pour certains individus un profil fermentaire stable dans le temps. Par contre, cette représentation marque une dynamique très importante de chaque AGCC (Figure 19).

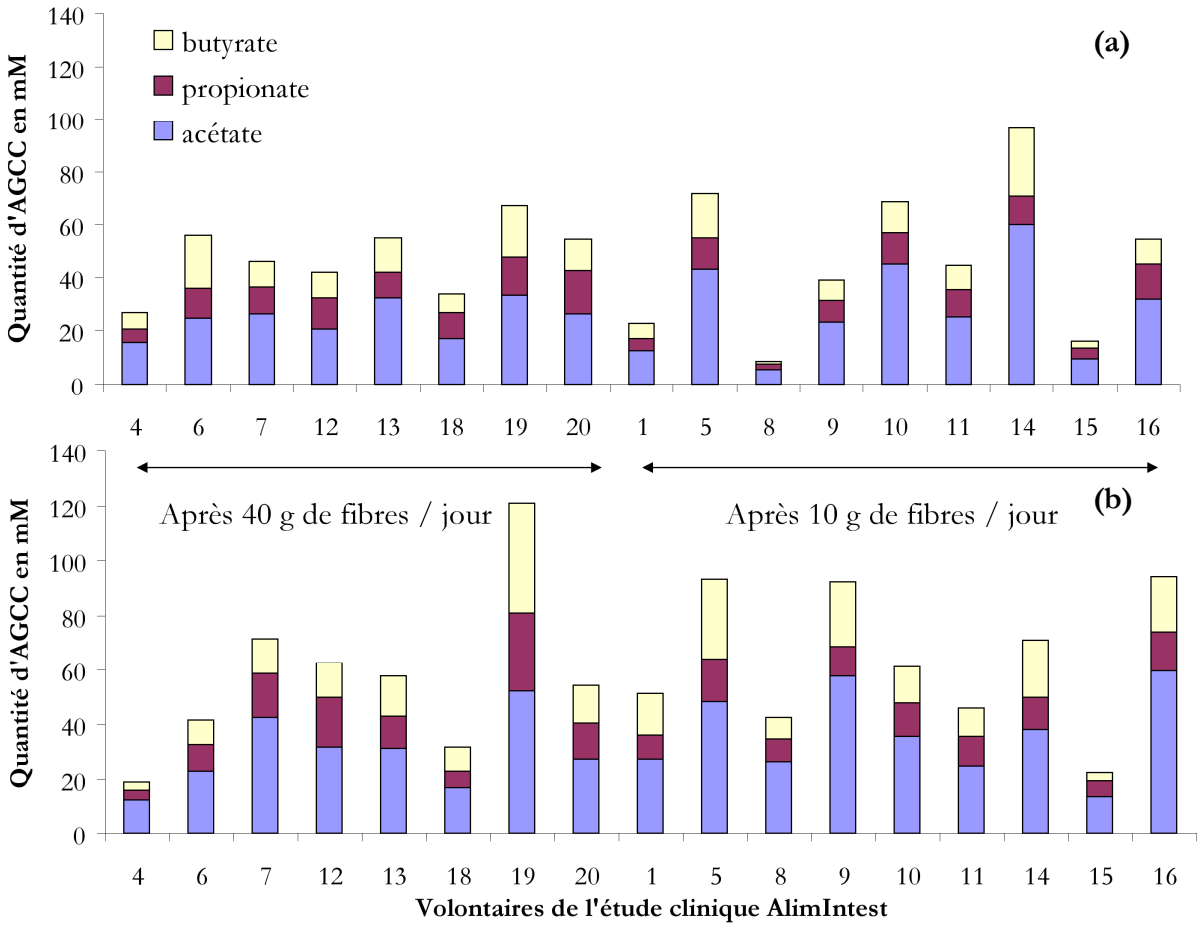


Figure 19 : Profils des AGCC en fonction des individus avant et après la première phase du régime. (a) La quantité totale mesurée en AGCC peut différer d'un facteur 10 entre individus. **(b)** Apres la première phase de régime, les individus répondent très variablement à la quantité de fibres ingérées, si bien que la variabilité inter-individus représente 55 % de la variabilité totale.

Ces observations peuvent résulter d'une modulation des flux de production, de voies métaboliques régulées de manière différente par un apport plus élevé en substrats, de transports d'AGCC dont l'expression et la régulation ne sont pas identiques d'un AGCC à l'autre et en fonction du temps. Ces données ne permettent pas de conclure quant à l'augmentation spécifique d'un AGCC, qui pourrait avoir un effet bénéfique sur la santé de l'épithélium colique et donc de l'hôte. Une alimentation riche en fibres dans le cadre d'un régime normal, basée sur ces données

préliminaires, ne produit pas d'effet « butyrate » comme recherché dans l'administration d'une fibre particulière comme prébiotique.

6.4.3 Corrélation entre les groupes dominants du microbiote et les profils AGCC

Lorsque l'on réalise une analyse de co-inertie en fonction des individus pour comparer les structures des données de qPCR et de dosages des AGCC, nous observons une co-structure significative entre les deux jeux de données (Figure 20). La même analyse en fonction de la dynamique du régime ne permet pas d'observer de structure significative. Cela est attendu puisque l'ACPVI ne donne pas de résultat significatif en fonction du régime pour les AGCC.

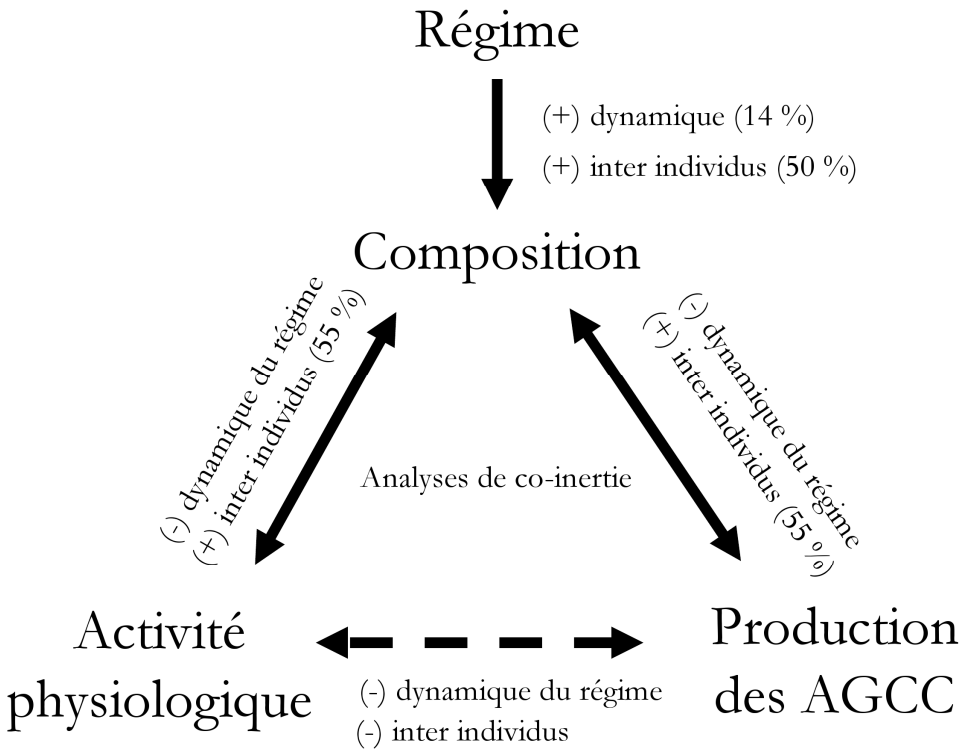


Figure 20 : Décomposition statistique des relations entre la composition du microbiote, son activité physiologique et la production des AGCC en fonction du régime. Malgré un impact inter-individus de 50% sur la variation totale, la dynamique imposée par l'étude clinique impacte la composition du microbiote et est responsable de près de 14 % de la variation totale. La composition du microbiote, son activité physiologique et la production des AGCC ne peuvent être corrélés entre eux que par la spécificité individuelle avec des analyses de co-inertie.

Dès lors, grâce à l'analyse de co-inertie, nous pouvons établir des corrélations fortes entre la présence de certains groupes bactériens et le dosage des AGCC (Figure 21). Nous observons que l'abondance de *E. coli* est très fortement anti-corrélée avec la production des AGCC principaux tels que l'acétate, le propionate est le butyrate. *E. coli* est très peu fermentaire et ne doit donc pas intervenir dans la production d'AGCC.

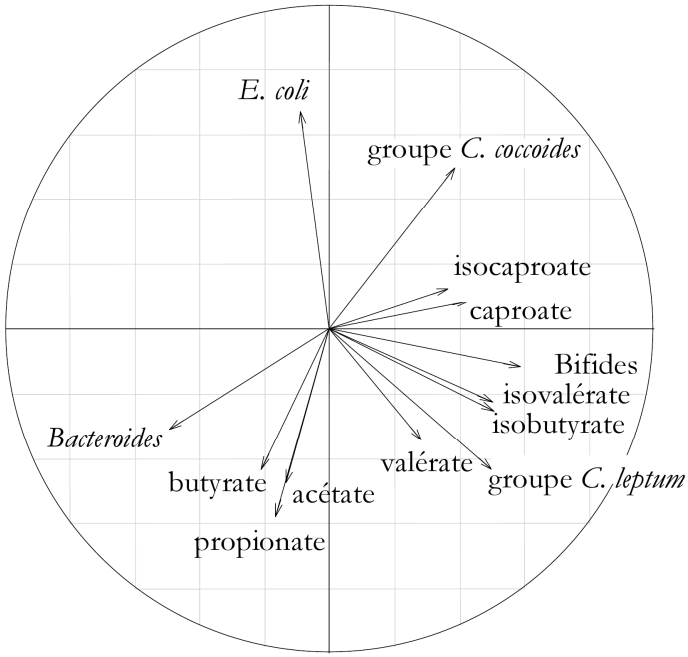


Figure 21 : Cercle de corrélations entre l'abondance des groupes du microbiote et la quantité d'acides gras à chaînes courtes.

L'abondance du groupe *Bacteroides/Prevotella* est bien corrélée avec la présence d'acétate, propionate, et butyrate. De manière surprenante, l'abondance de *C. coccoides* est anti-corrélée avec les trois AGCC principaux. Ceci peut être expliqué par le fait que le temps de croissance connu pour les espèces isolées de ce groupe est plus long.

La production des autres AGCC et des iso-acides varie indépendamment des acides principaux et de l'abondance d'*E. coli*, *Bacteroides* et du groupe *C. coccoides*. La production de ces acides est plutôt bien corrélée avec l'abondance en *C. leptum* et en Bifides.

Si la production de certains AGCC est corrélée significativement avec certaines composantes du microbiote, il est clair que la dynamique de production des AGCC est différente de celle observée avec le microbiote. Le microbiote réagit quantitativement au régime alimentaire avec des temporalités différentes en fonction des groupes bactériens dominants. La production des AGCC est très liée à l'individu.

Il semble qu'il existe un pouvoir tampon non lié au microbiote car les évolutions basées sur la qPCR semblent minimes. En revanche, l'absorption des différents AGCC doit être très différente d'un individu à l'autre. En combinant les effets sur le temps de transit et l'absorption des AGCC, une partie des résultats peut être interprétée comme liée au pouvoir tampon du tractus digestif.

Les variations de la structure du microbiote en fonction du régime ne peuvent être liées ni avec la production des AGCC ni avec l'activité transcriptionnelle basée sur le rapport ARN/ADN.

D'autres mécanismes transcriptionnels sont en jeu, notamment à l'interface hôte/microbiote[22]. Par ailleurs, nous avons montré que bien d'autres facteurs entraient en jeu dans la dynamique structurelle du microbiote et celle de l'hôte. Certaines composantes du microbiote sont corrélées dynamiquement et significativement avec des variables métaboliques et inflammatoires de l'hôte, indépendamment de l'apport calorique (**Article 4**).

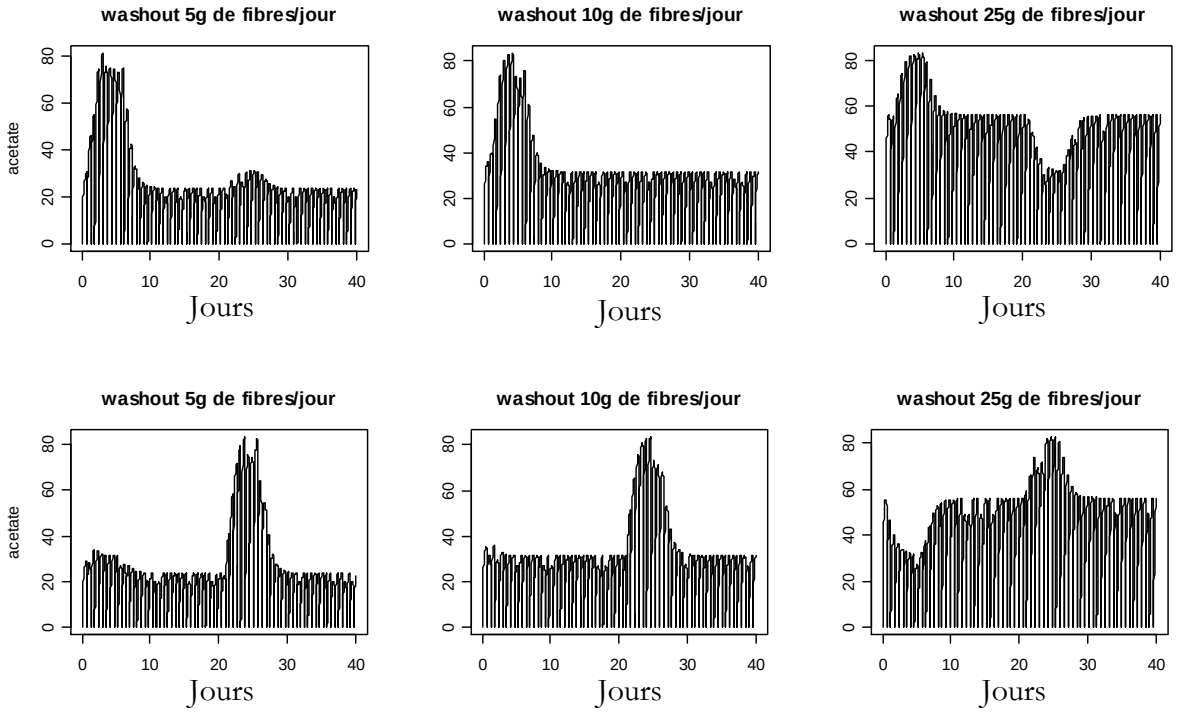


Figure 22 : Simulation de l'étude clinique AlimIntest avec différentes habitudes alimentaires sur la production d'acétate en mM. Trois types d'habitudes alimentaires ont été testées : 5 g de fibres par jour, 10 g de fibres par jour et 25 g de fibres par jour. En haut : séquences de régimes 40-10, en bas : séquences de régimes 10-40.

Un travail du laboratoire réalisé en parallèle, sur la modélisation de la chaîne trophique du côlon humain, permet d'obtenir des simulations théoriques. A partir de données issues de la littérature, le côlon a été modélisé (Muñoz-Tamayo et al., 2007). En réalisant des simulations identiques à l'étude clinique, nous avons mimé les différents régimes de base des individus. Le modèle retenu pour le ratio acétate : propionate : butyrate est 50 : 25 : 25. Les simulations du modèle ont permis de reproduire en théorie le comportement du microbiote des volontaires après 10 puis 40 ou 40 puis 10 g de fibres, en fonction du régime de départ (5 g, 10 g ou 25 g). Malgré les recommandations santé, les individus peuvent avoir un régime de base variant de 5 à 60 g de fibres par jour.

[22] Nous attendons des données métatranscriptomiques.

Le modèle donne une représentation simplifiée des profils d'AGCC attendus en fonction des régimes. Les paramètres d'absorption d'AGCC sont fixes et la réaction enzymatique peu modulée dans ce modèle. Malgré ces simplifications, le modèle permet tout de même d'anticiper des variations observées dans l'étude clinique. Un des problèmes rencontrés lors des études d'intervention nutritionnelle est la caractérisation du régime des volontaires avant l'intervention. Des questionnaires validés sont utilisés mais ne détaillent pas toujours les paramètres susceptibles d'influencer le microbiote, et sont basés sur du « déclaratif ». Dans le cas d'un régime à 25 g de fibres, on peut ainsi anticiper que chez certains individus le régime à 10 g de fibres voit les AGCC diminuer, puis être restitués lors du « wash out ». L'inverse est observé si le régime de départ est à 5 g par jour. Or c'est exactement ce que l'on peut observer sur certaines dynamiques d'AGCC pour certains individus (données non montrées).

Il est clair que de nombreux paramètres, en plus des habitudes alimentaires des patients, doivent être pris en compte pour savoir si le régime alimentaire impacte l'activité du microbiote. La variabilité des individus est très importante ainsi que leur réponse face au régime alimentaire. Si la composition du microbiote semble être impactée par le régime, son activité physiologique, mesurée par le ratio ARN/ADN, et les profils AGCC ne sont pas impactés significativement par le régime alimentaire. C'est peut-être à une autre échelle que le régime alimentaire peut influencer le microbiote, notamment au niveau de son méta-transcriptome. Une approche méta-transcriptomique a objectif d'obtenir une résolution plus fine et offrirait une vision plus large des fonctions du microbiote que régulerait un régime riche en fibres.

6.4.4 Etude de la modulation de l'activité du microbiote par un régime riche en fibres

Une approche méta-transcriptomique a été mise en œuvre pour étudier la variation de l'expression des ARN messagers du microbiote intestinal entre le régime à 10 g de fibres par jour et le régime à 40 g de fibres par jour. Quatre individus participant à l'étude clinique AlimIntest ont été choisis au hasard parmi ceux ayant subi la séquence de régime 10-40. Une extraction d'ARN a été effectuée sur les échantillons des points n°3 et n°5 (Figure 10). Une hybridation soustractive des ARN ribosomiques et une rétro-transcription ont ensuite été réalisées afin d'obtenir une banque d'ADN complémentaires.

Plus de 600 000 séquences ont été obtenues à partir de ces huit banques par pyroséquençage (GS FLX Titanium). Après nettoyage en fonction de la qualité de séquence, et extraction *in silico* des séquences d'ARNr, 118 301 séquences ont été comparées par Blastx aux bases de données NR et KEGG afin d'assigner une fonction aux séquences.

Dans les deux conditions de régime, le même nombre de requêtes (~ 10 000) a été retrouvé dans la base de données NR avec cependant un nombre inférieur de gènes à 40 g de fibres (- 25 %). Ceci pose la question d'une diversité fonctionnelle plus faible à 40 g de fibres qu'à 10 g de fibres et hypothétiquement un resserrement autour du métagénome minimal du microbiote.

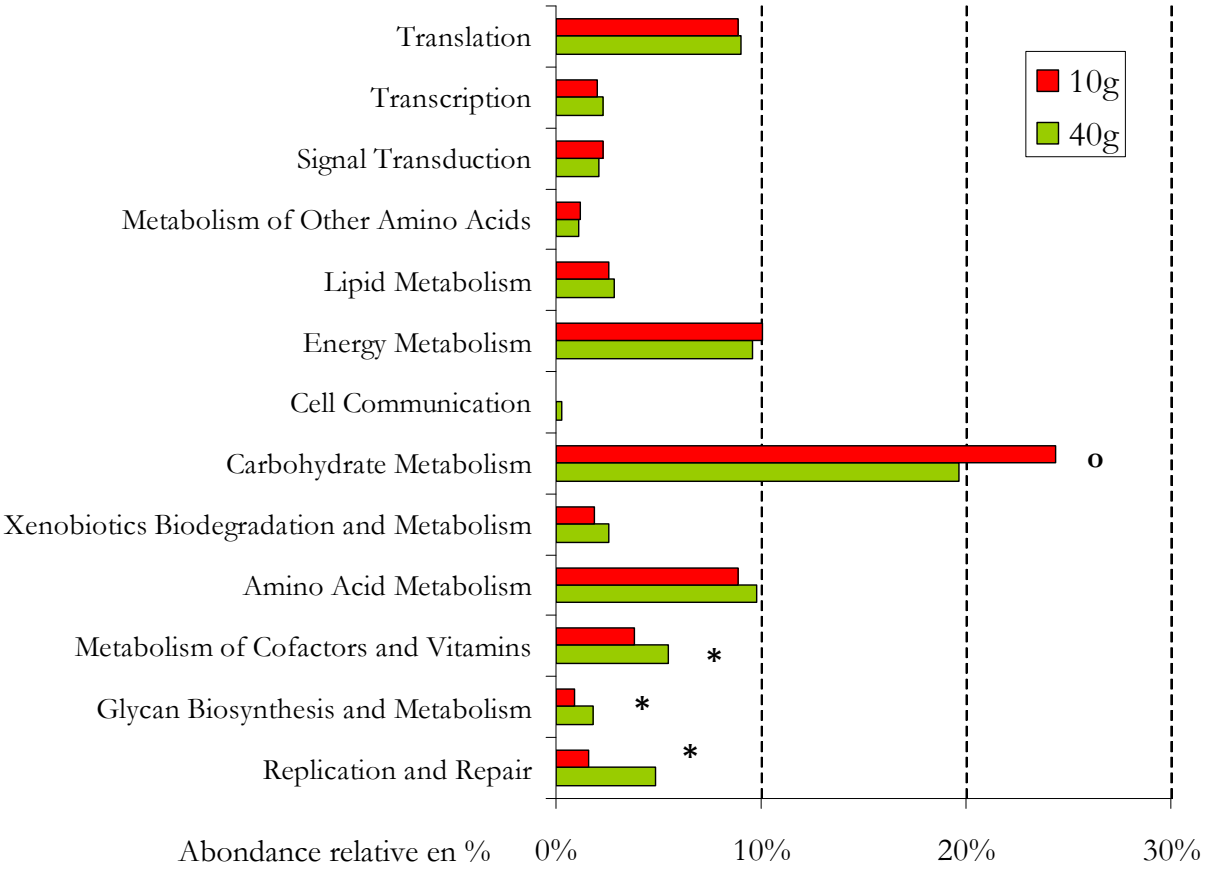


Figure 23 : Abondance relative des ARNm dans les sous-catégories de la base KEGG en fonction de la teneur en fibres du régime. La significativité a été évaluée avec la bibliothèque ShotgunFunctionalizeR avec un modèle poissonien (* : p < 0,05 ; o : p < 0,1).

Effectivement, les catégories fonctionnelles, ayant été décrites comme faisant partie du métagénome minimal comme le métabolisme des glycanes et celui des vitamines, semblent être surexprimées à 40 g de fibres (Figure 23). Néanmoins, ces observations sont à nuancer puisque 90 % des séquences obtenues n'ont pas été retrouvées dans les bases de données, ce qui montre qu'une grande partie de la diversité fonctionnelle du microbiote intestinal est inconnue. Parmi ces séquences, une partie d'entre elles a été retrouvées dans les clones ayant une activité hydrolytique dans le cadre du projet AlimIntest (données non montrées). Ceci suggère que l'approche méta-transcriptomique couplée à la métagénomique fonctionnelle est une approche puissante pour explorer cette diversité fonctionnelle inconnue du microbiote intestinal.

CONCLUSIONS ET PERSPECTIVES

Ces travaux de thèse ont permis de ré-évaluer la relation intime que nous avons avec notre microbiote en mettant en évidence l'existence d'un noyau phylogénétique d'espèces partagées par tous les individus. Ce nombre limité de phylotypes est particulièrement bien adapté à l'écosystème intestinal, si bien qu'on les trouve partagés par la majorité des individus. Il contribuerait à maintenir l'homéostasie intestinale ainsi que les fonctions principales assurées par le microbiote. L'existence de ce noyau proviendrait de la coévolution entre les espèces du microbiote et l'Homme. Il y a potentiellement deux forces qui s'affrontent pour maintenir ce consortium d'espèces à l'espèce humaine. D'une part, une pression de l'hôte oblige le génome de chaque souche de chaque espèce à être spécialement adapté à l'écosystème intestinal et d'autre part, une coopération et un dialogue entre les espèces elles-mêmes s'opèrent sous la forme de chaînes trophiques et de « quorum-sensing » leur permettant d'être résilientes dans le microbiome intestinal. Cela contribue à la persistance d'un « éco-génome » intestinal, à mettre en parallèle avec la notion d'écotype, c'est-à-dire d'une fraction génomique qui permet aux espèces de s'adapter à l'écosystème intestinal.

Cet « éco-génome » peut varier en fonction des niches écologiques dans l'intestin et en fonction du style de vie de l'individu, dont ses habitudes alimentaires. De plus, par l'intermédiaire de transferts horizontaux de gènes, il peut être partagé avec d'autres espèces, leur conférant un aspect mutualiste avec l'hôte. Etant précieux pour la résilience des espèces dans le microbiote, cet « éco-génome » doit être particulièrement transcrit par le microbiote. Il manque cependant des données pour étayer ces hypothèses, comme notamment le suivi métagénomique de l'implantation du microbiote et des espèces du noyau phylogénétique chez les nouveaux-nés. Cela permettrait d'en savoir davantage sur cette fenêtre ouverte où le dialogue entre l'hôte et son microbiote, au niveau immunitaire, est particulièrement intense.

D'un point de vue épidémiologique, il sera intéressant de confronter les séquences du noyau phylogénétique avec le suivi de familles ayant des membres atteints d'une maladie inflammatoire de l'intestin, ou bien de les utiliser comme outils diagnostiques pour anticiper la récidive de la maladie chez les patients après chirurgie. A posteriori, il sera utile d'étudier l'impact de l'absence d'une ou plusieurs espèces du noyau sur l'homéostasie du microbiote intestinal.

Les nouvelles technologies comme le séquençage génomique « single cell » permettront dans un futur très proche, d'avoir accès aux génomes des espèces du noyau qui sont phylogénétiquement loin des souches cultivées, et d'anticiper ou de suggérer leur rôle fonctionnel. D'autre part, le

nano-séquençage permettra de séquencer massivement sans passer par une étape chimique ou enzymatique. Cette technologie permet aussi d'avoir accès directement aux acides nucléiques simple brin sans passer par une étape de rétro-transcription. Cela facilitera bien évidemment les études de méta-transcriptomique. Le développement d'outils bioinformatiques et bio-statistiques devra faire face, plus que jamais, aux évolutions technologiques futures en créant de nouveaux concepts d'analyse. Le besoin de standardiser les méthodes d'analyse est devenu un point très critique pour intégrer les données des autres études. Tant que les méthodes de production de données et d'analyse de résultats ne seront pas standardisées, nous allons être confrontés à un grand nombre d'études dont les messages ou conclusions se contrediront alors que les données ne sont pas si antinomiques.

A travers le projet AlimIntest, ces travaux de thèse ont tenté d'intégrer des concepts mathématiques, microbiologiques, physiologiques et écologiques au service d'une question nutritionnelle et de santé. Malgré la variabilité inter-individus et intra-individus dans le temps, l'homogénéité de la cohorte clinique et le schéma de l'étude en cross-over randomisé, permettent d'avancer de premières conclusions solides sur l'impact des fibres alimentaires sur le microbiote. Celui-ci est directement structuré dans sa composition dans le temps en fonction des régimes. L'étude en cross-over a permis de révéler qu'une période de « wash-out » de 15 jours, pour ce type d'étude, n'était pas suffisante, ce qui permet indirectement aussi d'affirmer que le microbiote est impacté par le régime alimentaire pendant au moins deux semaines. Pour finir, le microbiote est d'abord corrélé à la production des AGCC par la spécificité individuelle et non par l'impact du régime. Cela suggère que les recommandations nutritionnelles futures devront tenir compte de la spécificité de chacun. Pour finir, ces travaux ouvrent ainsi de nouvelles perspectives pour de futures investigations nutritionnelles et épidémiologiques.

RÉFÉRENCES

Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V., and Polz, M.F. (2004) Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J Bacteriol* **186**: 2629-2635.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389-3402.

Amann, R.I., Ludwig, W., and Schleifer, K.H. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews* **59**: 143-169.

Anderson, J.W. (1986) Dietary fiber in nutrition management of diabetes. In *Dietary fiber*. Plenum (ed). New York, pp. 343-360.

Ashelford, K.E., Chuzhanova, N.A., Fry, J.C., Jones, A.J., and Weightman, A.J. (2006) New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. *Applied and Environmental Microbiology* **72**: 5734-5741.

Backhed, F., Ley, R.E., Sonnenburg, J.L., Peterson, D.A., and Gordon, J.I. (2005) Host-Bacterial Mutualism in the Human Intestine. *Science* **307**: 1915-1920.

Backhed, F., Ding, H., Wang, T., Hooper, L.V., Koh, G.Y., Nagy, A. et al. (2004) The gut microbiota as an environmental factor that regulates fat storage. *Proc Natl Acad Sci U S A* **101**: 15718-15723.

Berg Miller, M.E., Antonopoulos, D.A., Rincon, M.T., Band, M., Bari, A., Akraiko, T. et al. (2009) Diversity and strain specificity of plant cell wall degrading enzymes revealed by the draft genome of Ruminococcus flavefaciens FD-1. *PLoS ONE* **4**: e6650.

Bik, E.M., Eckburg, P.B., Gill, S.R., Nelson, K.E., Purdom, E.A., Francois, F. et al. (2006) Molecular analysis of the bacterial microbiota in the human stomach. *Proceedings of the National Academy of Sciences* **103**: 732-737.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* **37**: D233-238.

Chessel, D., Dufour, A.- B. and Thioulouse, J. (2004) The ade4 package-I- One-table methods. *R News* **4**: 5 - 10.

Coen, J.A., and Dehority, B.A. (1970) Degradation and utilization of hemicellulose from intact forages by pure cultures of rumen bacteria. *Appl Microbiol* **20**: 362-368.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam, S.A., McGarrell, D.M. et al. (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Research* **33**: D294-296.

Danchin, A. (2002) Génomes et évolution. *Annales de l'Institut Pasteur* **11**: 9-18.

DeLong, E.F. (2009) The microbial ocean from genomes to biomes. *Nature* **459**: 200-206.

DeSantis, T.Z., Jr., Hugenholtz, P., Keller, K., Brodie, E.L., Larsen, N., Piceno, Y.M. et al. (2006) NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res* **34**: W394-399.

Dethlefsen, L., McFall-Ngai, M., and Relman, D.A. (2007) An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* **449**: 811-818.

Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., Brulc, J.M. et al. (2008) Functional metagenomic profiling of nine biomes. *Nature* **452**: 629-632.

Dolédec, S., and Chessel, D. (1994) Co-inertia analysis: an alternative method for stidying species - environement relationships. *Freshwater Biology* **31**: 277-294.

Dolédec, S., Chessel, D., Ter Braak, C.J.F., and Champely, S. (1996) Matching species traits to environmental variables: a new three-table ordination method. *Environmental and Ecological Statistics* **3**: 143-166.

Duncan, S., Belenguer, A., Holtrop, G., Johnstone, A., Flint, H., and Lobley, G. (2007) Reduced dietary intake of carbohydrates by obese subjects results in decreased concentrations of butyrate and butyrate-producing bacteria in feces. *Applied and Environmental Microbiology* **73**: 1073-1078.

Duncan, S.H., Scott, K.P., Ramsay, A.G., Harmsen, H.J.M., Welling, G.W., Stewart, C.S., and Flint, H.J. (2003) Effects of Alternative Dietary Substrates on Competition between Human Colonic Bacteria in an Anaerobic Fermentor System. In, pp. 1136-1142.

Duncan, S.H., Lobley, G.E., Holtrop, G., Ince, J., Johnstone, A.M., Louis, P., and Flint, H.J. (2008) Human colonic microbiota associated with diet, obesity and weight loss. *Int J Obes (Lond)* **32**: 1720-1724.

Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M. et al. (2005) Diversity of the Human Intestinal Microbial Flora. *Science* **308**: 1635-1638.

Edwards, U., Rogall, T., Blocker, H., Emde, M., and Bottger, E.C. (1989) Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. *Nucleic Acids Res* **17**: 7843-7853.

Egert, M., de Graaf, A.A., Smidt, H., de Vos, W.M., and Venema, K. (2006) Beyond diversity: functional microbiomics of the human colon. *Trends Microbiol* **14**: 86-91.

Favier, C.F., Vaughan, E.E., De Vos, W.M., and Akkermans, A.D. (2002) Molecular monitoring of succession of bacterial communities in human neonates. *Applied and Environmental Microbiology* **68**: 219-226.

Felsentein, J. (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5**: 164-166.

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P. et al. (2008) The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* **26**: 541-547.

Finegold, S.M., Attebery, H.R., and Sutter, V.L. (1974) Effect of diet on human fecal flora: comparison of Japanese and American diets. *Am J Clin Nutr* **27**: 1456-1469.

Firkins, J.L., Bowman, J.G., Weiss, W.P., and Naderer, J. (1991) Effects of protein, carbohydrate, and fat sources on bacterial colonization degradation of fiber in vitro. *J Dairy Sci* **74**: 4273-4283.

Flint, H.J., Duncan, S.H., Scott, K.P., and Louis, P. (2007) Interactions and competition within the microbial community of the human colon: links between diet and health. In, pp. 1101-1111.

Flint, H.J., Bayer, E.A., Rincon, M.T., Lamed, R., and White, B.A. (2008) Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nature Reviews. Microbiology* **6**: 121-131.

Frank, D.N., St Amand, A.L., Feldman, R.A., Boedeker, E.C., Harpaz, N., and Pace, N.R. (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci U S A* **104**: 13780-13785.

Gianoulis, T.A., Raes, J., Patel, P.V., Bjornson, R., Korbel, J.O., Letunic, I. et al. (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci U S A* **106**: 1374-1379.

Gibson, G.R. (1998) Dietary modulation of the human gut microflora using prebiotics. *British Journal of Nutrition* **80**: S209-212.

Gilbert, J.A., Field, D., Huang, Y., Edwards, R., Li, W., Gilna, P., and Joint, I. (2008) Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One* **3**: e3042.

Gill, S.R., Pop, M., DeBoy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S. et al. (2006) Metagenomic Analysis of the Human Distal Gut Microbiome. *Science* **312**: 1355-1359.

Gophna, U., Sommerfeld, K., Gophna, S., Doolittle, W.F., and Veldhuyzen van Zanten, S.J.O. (2006) Differences between Tissue-Associated Intestinal Microfloras of Patients with Crohn's Disease and Ulcerative Colitisâ–¿â€¡. *J Clin Microbiol* **44**: 4136-4141.

Guigoz, Y., Dore, J., and Schiffrin, E.J. (2008) The inflammatory status of old age can be nurtured from the intestinal environment. *Curr Opin Clin Nutr Metab Care* **11**: 13-20.

Handelsman, J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* **68**: 669-685.

Hayashi, H., Sakamoto, M., and Benno, Y. (2002a) Fecal microbial diversity in a strict vegetarian as determined by molecular analysis and cultivation. *Microbiology and Immunology* **46**: 819-831.

Hayashi, H., Sakamoto, M., and Benno, Y. (2002b) Phylogenetic analysis of the human gut microbiota using 16S rDNA clone libraries and strictly anaerobic culture-based methods. *Microbiology and Immunology* **46**: 535-548.

Holdeman, L.V., Good, I.J., and Moore, W.E.C. (1976) Human fecal flora : variation in bacterial composition within individuals and a possible effect of emotional stess. *Applied and Environmental Microbiology* **31**: 359-375.

Hooper, L.V., and Gordon, J.I. (2001) Commensal Host-Bacterial Relationships in the Gut. *Science* **292**: 1115-1118.

Huber, T., Faulkner, G., and Hugenholtz, P. (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20**: 2317-2319.

Hugenholtz, P., and Tyson, G.W. (2008) Microbiology: metagenomics. *Nature* **455**: 481-483.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res* **17**: 377-386.

Jacobs, L.R. (1986) Dietary fiber and gastrointestinal epithelial cell proliferation. In *Dietary fiber*. Plenum (ed). New york, pp. 211-228.

Jimenez, E., Fernandez, L., Marin, M.L., Martin, R., Odriozola, J.M., Nueno-Palop, C. et al. (2005) Isolation of commensal bacteria from umbilical cord blood of healthy neonates born by cesarean section. *Curr Microbiol* **51**: 270-274.

Kristiansson, E., Hugenholtz, P., and Dalevi, D. (2009) ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics* **25**: 2737-2738.

Kruse, H.P., Kleessen, B., and Blaut, M. (1999) Effects of inulin on faecal bifidobacteria in human subjects. *Br J Nutr* **82**: 375-382.

Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A. et al. (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Research* **14**: 169-181.

Lay, C., Sutren, M., Rochet, V., Saunier, K., Doré, J., and Rigottier-Gois, L. (2004) Design and validation of 16S rRNA probes to enumerate members of the Clostridium leptum subgroup in human faecal microbiota. *Environmental Microbiology* **in press**.

Lee, Z.M., Bussema, C., 3rd, and Schmidt, T.M. (2009) rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. *Nucleic Acids Res* **37**: D489-493.

Legendre, P., and Legendre, L. (1998) *Numerical ecology. Second english edition*. Amsterdam: Elsevier.

Lepage, P., Seksik, P., Sutren, M., Cochetière, M.-F.d.l., Jian, R., Marteau, P., and Doré, J. (2005) Biodiversity of the mucosa-associated microbiota is stable along the distal digestive tract in healthy individuals and patients with IBD. *Inflammatory Bowel Diseases* **11**: 473-480.

Ley, R.E., Peterson, D.A., and Gordon, J.I. (2006a) Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124**: 837-848.

Ley, R.E., Turnbaugh, P.J., Klein, S., and Gordon, J.I. (2006b) Microbial ecology: Human gut microbes associated with obesity. *Nature* **444**: 1022.

Ley, R.E., Backhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D., and Gordon, J.I. (2005) Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences* **102**: 11070-11075.

Ley, R.E., Hamady, M., Lozupone, C., Turnbaugh, P.J., Ramey, R.R., Bircher, J.S. et al. (2008) Evolution of mammals and their gut microbes. *Science* **320**: 1647-1651.

Li, K.B. (2003) ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics* **19**: 1585-1586.

Li, M., Wang, B., Zhang, M., Rantalainen, M., Wang, S., Zhou, H. et al. (2008) Symbiotic gut microbes modulate human metabolic phenotypes. *Proceedings of the National Academy of Sciences of the United States of America* **105**: 2117-2122.

Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658-1659.

Lozupone, C.A., Hamady, M., Cantarel, B.L., Coutinho, P.M., Henrissat, B., Gordon, J.I., and Knight, R. (2008) The convergence of carbohydrate active gene repertoires in human gut microbes. *Proc Natl Acad Sci U S A* **105**: 15076-15081.

Lupton, J.R., Coder, D.M., and Jacobs, L.R. (1985) Influence of luminal pH on rat large bowel epithelial cell cycle. *American Journal of Physiology* **249**: G382-G388.

Macdonald, T.T., and Monteleone, G. (2005) Immunity, inflammation, and allergy in the gut. *Science* **307**: 1920-1925.

Mackie, R., Sghir, A., and Gaskins, H.R. (1999) Developmental microbial ecology of the neonatal gastrointestinal tract. *American Journal of Clinical Nutrition* **69**: 1035S-1045S.

Manichanh, C., Rigottier-Gois, L., Bonnaud, E., Gloux, K., Pelletier, E., Frangeul, L. et al. (2006) Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55**: 205-211.

Mariat, D., Firmesse, O., Levenez, F., Guimaraes, V., Sokol, H., Dore, J. et al. (2009) The Firmicutes/Bacteroidetes ratio of the human microbiota changes with age. *BMC Microbiol* **9**: 123.

Martin, A.P. (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Applied and Environmental Microbiology* **68**: 3673-3682.

Matsuki, T., Watanabe, K., Fujimoto, J., Kado, Y., Takada, T., Matsumoto, K., and Tanaka, R. (2004) Quantitative PCR with 16S rRNA-gene-targeted species-specific primers for analysis of human intestinal bifidobacteria. *Applied and Environmental Microbiology* **70**: 167-173.

Mazmanian, S.K., Liu, C.H., Tzianabos, A.O., and Kasper, D.L. (2005) An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell* **122**: 107-118.

McHardy, A.C., Martin, H.G., Tsirigos, A., Hugenholtz, P., and Rigoutsos, I. (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* **4**: 63-72.

McIntyre, A., Gibson, P.R., and Young, G.P. (1993) Butyrate production from dietary fibre and protection against large bowel cancer in a rat model. *Gut* **34**: 386-391.

Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005) The microbial pan-genome. *Curr Opin Genet Dev* **15**: 589-594.

Medini, D., Serruto, D., Parkhill, J., Relman, D.A., Donati, C., Moxon, R. et al. (2008) Microbiology in the post-genomic era. *Nat Rev Microbiol* **6**: 419-430.

Miron, J., Ben-Ghedalia, D., and Morrison, M. (2001) Invited review: adhesion mechanisms of rumen cellulolytic bacteria. *J Dairy Sci* **84**: 1294-1309.

Mourino, F., Akkarawongsa, R., and Weimer, P.J. (2001) Initial pH as a determinant of cellulose digestion rate by mixed ruminal microorganisms in vitro. *J Dairy Sci* **84**: 848-859.

Muñoz-Tamayo, R., Steyer, J.P., Laroche, B., and Leclerc, M. (2007) Human colon: a complex bioreactor. conceptual modelling for the anaerobic digestion of the functional trophic chain. *Proc. 11th World Congress Anaerobic Digestion Bioenergy for our Future, Brisbane, Australia.*

Mutch, D.M., Simmering, R., Donnicola, D., Fotopoulos, G., Holzwarth, J.A., Williamson, G., and Corthesy-Theulaz, I. (2004) Impact of commensal microbiota on murine gastrointestinal tract gene ontologies. *Physiol Genomics* **19**: 22-31.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **27**: 29-34.

Pavoine, S., Dufour, A.B., and Chessel, D. (2004) From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *Journal of Theoretical Biology* **228**: 523-537.

Penders, J., Thijs, C., Vink, C., Stelma, F.F., Snijders, B., Kummeling, I. et al. (2006) Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics* **118**: 511-521.

Penders, J., Thijs, C., van den Brandt, P.A., Kummeling, I., Snijders, B., Stelma, F. et al. (2007a) Gut microbiota composition and development of atopic manifestations in infancy: the KOALA Birth Cohort Study. *Gut* **56**: 661-667.

Penders, J., Stobberingh, E.E., van den Brandt, P.A., Thijs, C., Penders, J., Thijs, C. et al. (2007b) The role of the intestinal microbiota in the development of atopic disorders. *Allergy* **62**: 1223-1236.

Perez, P.F., Dore, J., Leclerc, M., Levenez, F., Benyacoub, J., Serrant, P. et al. (2007) Bacterial Imprinting of the Neonatal Immune System: Lessons From Maternal Cells? *Pediatrics* **119**: e724-732.

Pérez, S., and Mazeau, K. (2005) Conformation, Structures, and Morphologies of Celluloses. In *Polysaccharides: structural diversity and functional versatility*. New York: CRC; 2 edition, pp. 41-68.

Pryde, S.E., Duncan, S.H., Hold, G.L., Stewart, C.S., and Flint, H.J. (2002) The microbiology of butyrate formation in the human colon. *FEMS Microbiology Letters* **217**: 133-139.

Raes, J., Foerstner, K.U., and Bork, P. (2007) Get the most out of your metagenome: computational analysis of environmental sequence data. *Current Opinion in Microbiology* **10**: 490-498.

Rigottier-Gois, L., Le Bourhis, A.-G., Gramet, G., Rochet, V., and Doré, J. (2003a) Fluorescent hybridisation combined with flow cytometry and hybridisation of total RNA to analyse the composition of microbial communities in human faeces using 16S rRNA probes. *FEMS Microbiology Ecology* **43**: 237-245.

Rigottier-Gois, L., Rochet, V., Garrec, N., Suau, A., and Dore, J. (2003b) Enumeration of Bacteroides species in human faeces by fluorescent in situ hybridisation combined with flow cytometry using 16S rRNA probes. *Systematic and Applied Microbiology* **26**: 110-118.

Rigottier-Gois, L., Le Bourhis, A.-G., Gramet, G., Rochet, V., and Dore, J. (2003c) Fluorescent hybridisation combined with flow cytometry and hybridisation of total RNA to analyse the composition of microbial communities in human faeces using 16S rRNA probes. *FEMS Microbiology Ecology* **43**: 237-245.

Robert, C., and Bernalier-Donadille, A. (2003) The cellulolytic microflora of the human colon: evidence of microcrystalline cellulose-degrading bacteria in methane-excreting subjects. *FEMS Microbiology Ecology* **46**: 81-89.

Rondon, M.R., August, P.R., Bettermann, A.D., Brady, S.F., Grossman, T.H., Liles, M.R. et al. (2000) Cloning the Soil Metagenome: a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms. *Applied and Environmental Microbiology* **66**: 2541-2547.

Rowland, I.R., Rumney, C.J., Coutts, J.T., and Lievense, L.C. (1998) Effect of Bifidobacterium longum and inulin on gut bacterial metabolism and carcinogen-induced aberrant crypt foci in rats. *Carcinogenesis* **19**: 281-285.

Rudi, K., Zimonja, M., Kvenshagen, B., Rugtveit, J., Midtvedt, T., and Eggesbo, M. (2007) Alignment-independent comparisons of human gastrointestinal tract microbial communities in a multidimensional 16S rRNA gene evolutionary space. *Applied and Environmental Microbiology* **73**: 2727-2734.

Russell, J.B., and Wilson, D.B. (1996) Why are ruminal cellulolytic bacteria unable to digest cellulose at low pH? *J Dairy Sci* **79**: 1503-1509.

Savage, D.C. (1977) Microbial ecology of the gastrointestinal tract. *Ann. Rev. Microbiol.* **31**: 107-133.

Schloss, P.D. (2008) Evaluating different approaches that test whether microbial communities have the same structure. *Isme J* **2**: 265-275.

Schloss, P.D., and Handelsman, J. (2005) Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. *Applied and Environmental Microbiology* **71**: 1501-1506.

Schloss, P.D., and Handelsman, J. (2006) Introducing TreeClimber, a Test To Compare Microbial Community Structures. *Applied and Environmental Microbiology* **72**: 2379-2384.

Schloss, P.D., and Handelsman, J. (2008) A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *BMC Bioinformatics* **9**: 34.

Schloss, P.D., Larget, B.R., and Handelsman, J. (2004) Integration of Microbial Ecology and Statistics: a Test To Compare Gene Libraries. *Applied and Environmental Microbiology* **70**: 5485-5492.

Schwarz, W.H. (2001) The cellulosome and cellulose degradation by anaerobic bacteria. *Appl Microbiol Biotechnol* **56**: 634-649.

Schwiertz, A., Taras, D., Schafer, K., Beijer, S., Bos, N.A., Donus, C., and Hardt, P.D. (2009) Microbiota and SCFA in Lean and Overweight Healthy Subjects. *Obesity (Silver Spring)* **4**: 4.

Sghir, A., Gramet, G., Suau, A., Rochet, V., Pochart, P., and Dore, J. (2000) Quantification of Bacterial Groups within Human Fecal Flora by Oligonucleotide Probe Hybridization. *Applied and Environmental Microbiology* **66**: 2263-2266.

Sonnenburg, J.L., Chen, C.T., and Gordon, J.I. (2006) Genomic and metabolic studies of the impact of probiotics on a model gut symbiont and host. *PLoS Biol* **4**: e413.

Sonnenburg, J.L., Xu, J., Leip, D.D., Chen, C.-H., Westover, B.P., Weatherford, J. et al. (2005) Glycan Foraging in Vivo by an Intestine-Adapted Bacterial Symbiont. *Science* **307**: 1955-1959.

Stam, M.R., Danchin, E.G., Rancurel, C., Coutinho, P.M., and Henrissat, B. (2006) Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins. *Protein Engineering, Design and Selection* **19**: 555-562.

Suau, A., Bonnet, R., Sutren, M., Godon, J.J., Gibson, G.R., Collins, M.D., and Dore, J. (1999) Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Applied and Environmental Microbiology* **65**: 4799-4807.

Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M.L., McKendree, W., and Farmerie, W. (2009) ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Research* **37**: e76.

Swidsinski, A., Ladhoff, A., Pernthaler, A., Swidsinski, S., Loening-Baucke, V., Ortner, M. et al. (2002) Mucosal flora in inflammatory bowel disease. *Gastroenterology* **122**: 44-54.

Tannock, G.W. (2007) What immunologists should know about bacterial communities of the human bowel. *Semin Immunol* **19**: 94-105.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S. et al. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29**: 22-28.

Teeling, H., Meyerdierks, A., Bauer, M., Amann, R., and Glockner, F.O. (2004a) Application of tetranucleotide frequencies for the assignment of genomic fragments. In, pp. 938-947.

Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., and Glockner, F.O. (2004b) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5**: 163.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**: 4673-4680.

Tringe, S.G., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W. et al. (2005) Comparative Metagenomics of Microbial Communities. *Science* **308**: 554-557.

Tschop, M.H., Hugenholtz, P., and Karp, C.L. (2009) Getting to the core of the gut microbiome. *Nat Biotechnol* **27**: 344-346.

Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R., and Gordon, J.I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**: 1027.

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. (2007) The human microbiome project. *Nature* **449**: 804-810.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E. et al. (2009) A core gut microbiome in obese and lean twins. *Nature* **457**: 480-484.

van Tongeren, S.P., Slaets, J.P., Harmsen, H.J., and Welling, G.W. (2005) Fecal microbiota composition and frailty. *Appl Environ Microbiol* **71**: 6438-6442.

Vasquez, N., Mangin, I., Lepage, P., Seksik, P., Duong, J.-P., Blum, S. et al. (2007) Patchy distribution of mucosal lesions in ileal Crohn's disease is not linked to differences in the dominant mucosa-associated bacteria: A study using fluorescence in situ hybridization and temporal temperature gradient gel electrophoresis. *Inflammatory Bowel Diseases* **13**: 684-692.

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**: 5261-5267.

Wei, C., and Brent, M.R. (2006) Using ESTs to improve the accuracy of de novo gene prediction. *BMC Bioinformatics* **7**: 327.

Weimer, P.J. (1996) Why don't ruminal bacteria digest cellulose faster? *J Dairy Sci* **79**: 1496-1502.

White, J.R., Nagarajan, N., and Pop, M. (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* **5**: e1000352.

Willner, D., Thurber, R.V., and Rohwer, F. (2009) Metagenomic signatures of 86 microbial and viral metagenomes. *Environ Microbiol* **18**: 18.

Woese, C.R. (1979) A proposal concerning the origin of life on the planet earth. *Journal of Molecular Evolution* **13**: 95-101.

Woese, C.R. (1987) Bacterial evolution. *Microbiological Reviews* **51**: 221-271.

Woodmansey, E.J. (2007) Intestinal bacteria and ageing. *J Appl Microbiol* **102**: 1178-1186.

Woyke, T., Teeling, H., Ivanova, N.N., Huntemann, M., Richter, M., Gloeckner, F.O. et al. (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443**: 950.

Xu, J., Bjursell, M.K., Himrod, J., Deng, S., Carmichael, L.K., Chiang, H.C. et al. (2003) A Genomic View of the Human-Bacteroides thetaiotaomicron Symbiosis. *Science* **299**: 2074-2076.

Xu, J., Mahowald, M.A., Ley, R.E., Lozupone, C.A., Hamady, M., Martens, E.C. et al. (2007) Evolution of Symbiotic Bacteria in the Distal Human Intestine. *PLoS Biology* **5**: e156.

Yang, X., Xie, L., Li, Y., and Wei, C. (2009) More than 9,000,000 unique genes in human gut bacterial community: estimating gene numbers inside a human body. *PLoS One* **4**: e6074.

Zhang, H., DiBaise, J.K., Zuccolo, A., Kudrna, D., Braidotti, M., Yu, Y. et al. (2009) Human gut microbiota in obesity and after gastric bypass. *Proc Natl Acad Sci U S A* **106**: 2365-2370.

Zilber-Rosenberg, I., and Rosenberg, E. (2008) Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. *FEMS Microbiol Rev* **32**: 723-735.

Zoetendal, E.G., Akkermans, A.D., and De Vos, W.M. (1998) Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria. *Applied and Environmental Microbiology* **64**: 3854-3859.

Zoetendal, E.G., Rajilic-Stojanovic, M., and de Vos, W.M. (2008) High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. *Gut* **57**: 1605-1615.

Zoetendal, E.G., Ben-Amor, K., Akkermans, A.D., Abee, T., and de Vos, W.M. (2001a) DNA isolation protocols affect the detection limit of PCR approaches of bacteria in samples from the human gastrointestinal tract. *Systematic and Applied Microbiology* **24**: 405-410.

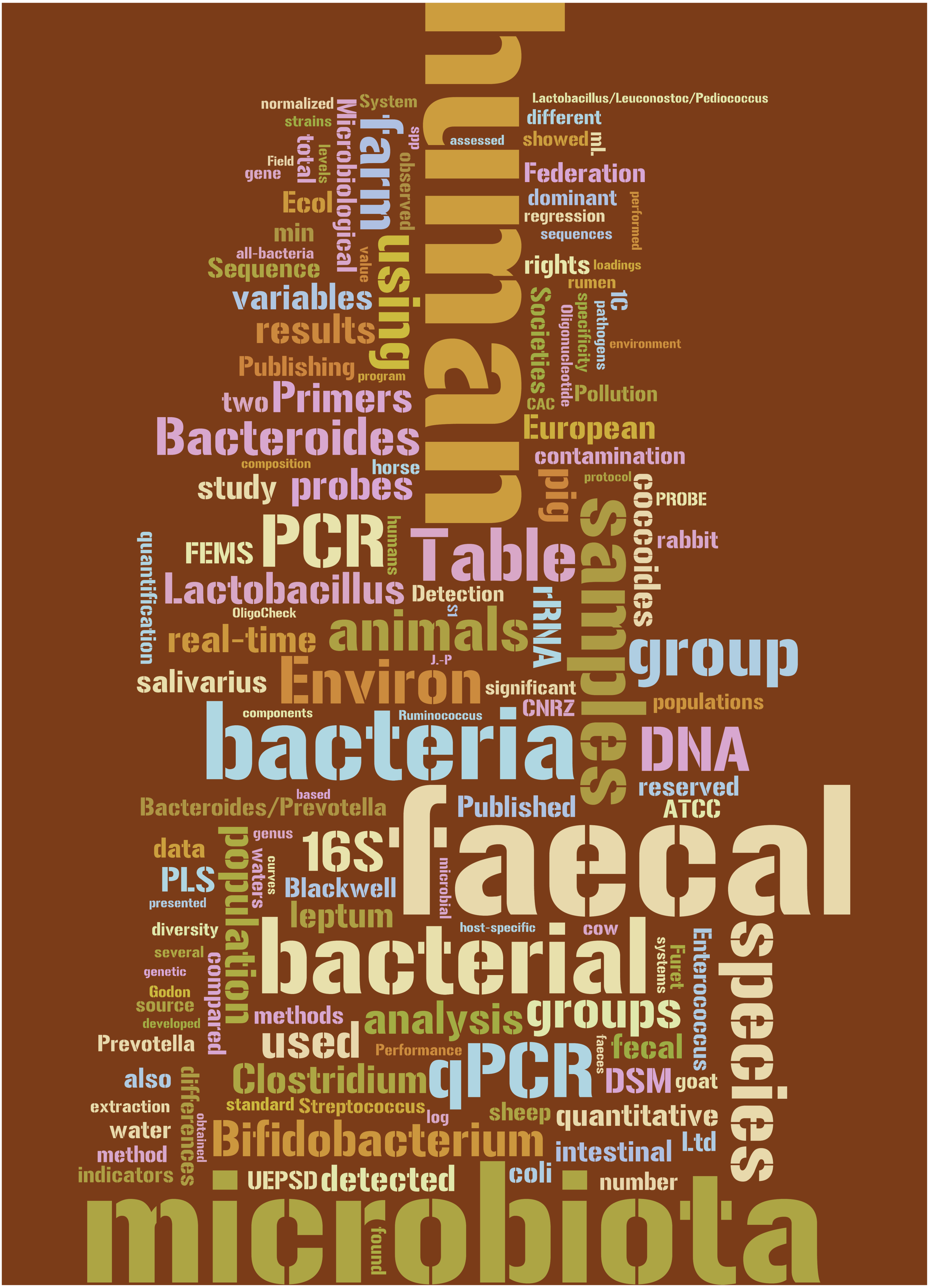
Zoetendal, E.G., Akkermans, A.D.L., Akkermans-van Vliet, W.M., De Visser, J.A.G.M., and De Vos, W.M. (2001b) The Host Genotype Affects the Bacterial Community in the Human Gastronintestinal Tract. *Microbial Ecology in Health and Disease* **13**: 129 - 134.

Zoetendal, E.G., Booijink, C.C., Klaassens, E.S., Heilig, H.G., Kleerebezem, M., Smidt, H., and de Vos, W.M. (2006) Isolation of RNA from bacterial samples of the human gastrointestinal tract. *Nat Protoc* **1**: 954-959.

PUBLICATIONS

L'**article 1** intitulé « Comparative assessment of human and farm animal faecal microbiota using real-time quantitative PCR » publié dans *FEMS Microbiology Ecology* a permis de dessiner et de valider de nouveaux systèmes qPCR pour l'étude du microbiote ainsi qu'une démarche statistique. L'**article 2** intitulé « RapidOTU: 16S rRNA gene sequences clustering into operational taxonomic units using tetranucleotides frequencies » soumis à *PLoS Computational Biology* est un article méthodologique qui propose une alternative pour l'analyse de séquences issues d'inventaires moléculaires du gène de l'ARNr 16S. L'**article 3** intitulé « Towards the human intestinal microbiota phylogenetic core » publié dans *Environmental Microbiology* confirme l'existence d'un noyau phylogénétique partagé par tous, dont les espèces qui le composent devront attirer une attention particulière pour les futures études épidémiologiques et nutritionnelles. L'**article 4** intitulé « Differential adaptation of human gut microbiota to bariatric surgery-induced weight loss: links with metabolic and low grade inflammation markers. » soumis à *PLoS medecine* montre que certains grands groupes bactériens sont corrélés à des paramètres inflammatoires, métaboliques et nutritionnels chez les obèses ayant subi un by-pass gastrique. L'**article 5**, en préparation, intitulé « Profiling microbial communities using multiplex pyrosequencing: a validation study » est un article méthodologique qui évalue l'utilisation de la technologie « 454 » pour réaliser des inventaires moléculaires sur le gène de l'ARN 16S en fonction de la région ciblée.

Tous les articles sont mis les uns à la suite des autres dans les pages suivantes.


human
faecal
microbiota
bacterial
bacteria
samples
species
animals
group
Environ
Table
PCR
qPCR
Bacteroides
Lactobacillus
Bifidobacterium
Clostridium
groups
analysis
DNA
farm
using
Microbiological
Primers
probes
results
variables
real-time
16S
population
used
detected
Published
Lactobacillus/Leuconostoc/Pediococcus
normalized
System
strains
total
levels
spp
observed
Field
gene
Ecol
min
all-bacteria
Sequence
value
Publishing
program
two
composition
horse
study
different
assessed
showed
ml
Federation
dominant
performed
regression
sequences
rights
loadings
rumen
Societies
specificity
pathogens
IC
Oligonucleotide
environment
Pollution
CAC
European
contamination
protocol
pig
coccoides
PROBE
rabbit
humans
FEMS
quantification
Detection
OligoCheck
rRNA
S1
J.-P
salivarius
significant
CNRZ
populations
components
Ruminococcus
reserved
based
Bacteroides/Prevotella
ATCC
genus
data
curves
waters
PLS
Blackwell
microbial
presented
leptum
diversity
host-specific
cow
several
systems
Furet
Enterococcus
genetic
Godon
source
compared
developed
methods
Prevotella
Performance
fecal
faeces
also
differences
DSM
goat
extraction
standard
Streptococcus
log
sheep
quantitative
water
obtained
intestinal
Ltd
method
indicators
UEPSD
coli
number
found

RESEARCH ARTICLE

Comparative assessment of human and farm animal faecal microbiota using real-time quantitative PCR

Jean-Pierre Furet[1], Olivier Firmesse[1], Michèle Gourmelon[2], Chantal Bridonneau[1], Julien Tap[1], Stanislas Mondot[1], Joël Doré[1] & Gérard Corthier[1]

[1]INRA, U910, Unité d'Ecologie et de Physiologie du Système Digestif, Jouy-en-Josas, France; and [2]IFREMER, Laboratoire de Microbiologie EMP/MIC, Plouzané, France

Correspondence: Jean-Pierre Furet, INRA, U910, Unité d'Ecologie et de Physiologie du Système Digestif, 78350 Jouy-en-Josas, France. Tel.: +33 1 34 65 29 29; fax: +33 1 34 65 24 62; e-mail: jean-pierre.furet@jouy.inra.fr

Received 21 February 2008; revised 13 February 2009; accepted 19 February 2009.

DOI:10.1111/j.1574-6941.2009.00671.x

Editor: Julian Marchesi

Keywords
quantitative PCR; faecal microbiota; human; farm animals.

Abstract

Pollution of the environment by human and animal faecal pollution affects the safety of shellfish, drinking water and recreational beaches. To pinpoint the origin of contaminations, it is essential to define the differences between human microbiota and that of farm animals. A strategy based on real-time quantitative PCR (qPCR) assays was therefore developed and applied to compare the composition of intestinal microbiota of these two groups. Primers were designed to quantify the 16S rRNA gene from dominant and subdominant bacterial groups. TaqMan® probes were defined for the qPCR technique used for dominant microbiota. Human faecal microbiota was compared with that of farm animals using faecal samples collected from rabbits, goats, horses, pigs, sheep and cows. Three dominant bacterial groups (*Bacteroides/Prevotella*, *Clostridium coccoides* and *Bifidobacterium*) of the human microbiota showed differential population levels in animal species. The *Clostridium leptum* group showed the lowest differences among human and farm animal species. Human subdominant bacterial groups were highly variable in animal species. Partial least squares regression indicated that the human microbiota could be distinguished from all farm animals studied. This culture-independent comparative assessment of the faecal microbiota between humans and farm animals will prove useful in identifying biomarkers of human and animal faecal contaminations that can be applied to microbial source tracking methods.

Introduction

Faecal pollution in coastal or fresh waters leads to human disease and economic losses such as closure of commercial shellfish harvesting and recreational and bathing areas. Recent incidents include the isolation of human enteric viruses and bacteria such as norovirus, hepatitis A virus, and *Salmonella* from coastal waters and shellfish, which were implicated in shellfish-borne outbreaks after oyster consumption (Potasman *et al.*, 2002; Martinez-Urtaza *et al.*, 2004). In light of this risk to health and safety, it is important to identify the source of faecal contamination to better facilitate resource management and remediation.

Faecal contamination of water resources is currently evaluated by employing culturing methods to detect and enumerate living facultative-anaerobic bacteria, such as *Escherichia coli*, enterococci, or faecal coliforms. Samples are normally obtained from shellfish or directly from bathing waters (Directives 2006/113/CE; 2006/7/CE). The species traditionally used as faecal indicators, however, have limitations owing to several factors, including (1) their short survival time in an open-water environment, (2) their ability to proliferate in soil, sand or sediments absent in any point-source faecal contamination, (3) the low levels of correlation with the actual presence of pathogens, (4) the underestimation of true bacterial presence through omission of non-cultivable bacteria, (5) their inability to track the source of faecal contamination because coliforms and enterococci are common to all mammalian hosts (Roszak & Colwell, 1987; Pommepuy *et al.*, 1996; Gordon & Cowling, 2003; Wheeler *et al.*, 2003; Hörman *et al.*, 2004; Savichtcheva & Okabe, 2006). In order to overcome these shortcomings, alternative

methods and indicators must be developed. Potential alternative indicators of faecal contamination could be anaerobic bacteria such as *Bacteroides* and *Bifidobacterium* that are more abundant in the faeces of warm-blooded animals than *E. coli* (Fiksdal *et al.*, 1985; Suau *et al.*, 1999). Importantly, these species have been shown to exhibit host-specific adaptation on the genetic level (Dick *et al.*, 2005). While these bacteria are fastidious to enumerate with conventional culture techniques, they can nonetheless be easily detected using current molecular methods. Because uncultivated bacteria represent 70–80% of the total human microbiota, culture-independent methods of analysis based on 16S rRNA gene have been developed (Suau *et al.*, 1999; Eckburg *et al.*, 2005). These studies showed that the most highly represented bacterial groups in human stools were the *Clostridium leptum* and the *Clostridium coccoides* groups of the *Firmicutes* followed by the *Bacteroides/Prevotella* group and the *Bifidobacterium* genus (Harmsen *et al.*, 2002; Lay *et al.*, 2005a). Studies involving domestic animal microbiota are less numerous and are mainly focused on the phylogenetic diversity of the intestinal bacterial community in pigs, cattle and chicken (Lan *et al.*, 2002; Leser *et al.*, 2002; Ozutsumi *et al.*, 2005). Recently, specific quantitative PCR (qPCR) approaches were used to estimate a limited number of bacterial species or groups of faecal microbiota (Matsuki *et al.*, 2004; Seurinck *et al.*, 2005; Reischer *et al.*, 2006).

The work presented here seeks to establish a more comprehensive dataset in comparing human and farm animal microbiota. To this end, we developed and optimized a qPCR-based approach, which was subsequently applied to analyse faecal samples collected from humans and farm animals. Using such molecular techniques, we overcome the limits of traditional faecal indicators, including culturing methods, which consistently underestimate faecal population. The development and application of our qPCR systems quantifies faecal bacteria groups in human and animal faecal samples and provides essential information concerning potential alternative faecal indicators and host-specific bacterial groups.

Materials and methods

DNA extraction from faecal samples

The DNA extracts from faecal samples of 21 human stools were prepared as described previously (Godon *et al.*, 1997; Lay *et al.*, 2005b). Faecal samples from five individual animals were collected for each of six farm species (rabbit, goat, horse, pig, sheep and cow) and stored at $-80\,^{\circ}C$ immediately after sampling. Total cellular DNA was extracted from 0.2 g of animal faecal material using the G'NOME® kit (BIO 101, La Jolla, CA) with modifications. Faecal samples were homogenized in the supplied cell suspension solution. Cell lysis/denaturing solution was then added and the samples incubated at $55\,^{\circ}C$ for 2 h. To improve cellular lysis, 750 μL of 0.1-mm-diameter silica beads were added, and agitation carried out at maximum speed for 10 min in a Beadbeater (Biospec, Bartlesville, OK). Polyvinylpolypyrrolidone (15 mg) was added to ensure removal of polyphenol contamination that could inhibit subsequent qPCR reactions. Samples were vortexed and centrifuged at 20 000 ***g*** for 3 min and the supernatant was recovered. The remaining pellet was washed with 400 μL of TENP [50 mM Tris (pH 8), 20 mM EDTA (pH 8), 100 mM NaCl, 1% polyvinylpolypyrrolidone] and centrifuged at 20 000 ***g*** for 3 min. The washing step was repeated once more and the resulting supernatants pooled. Nucleic acids were precipitated by addition of one volume isopropanol, storage at $-20\,^{\circ}C$ for 20 min, and centrifugation at 20 000 ***g*** for 10 min. The pellet was resuspended in 400 μL of distilled water plus 100 μL of salt-out mixture and incubated at $4\,^{\circ}C$ for 10 min. Samples were spun for 10 min at maximum speed, and the supernatant containing the DNA was transferred to a clean 1.5-mL microcentrifuge tube. DNA was precipitated with two volumes of 100% ethanol at room temperature for 5 min followed by centrifugation at 16 000 ***g*** for 5 min. DNA was resuspended in 150 μL of TE buffer. DNA solutions were stored at $-20\,^{\circ}C$ for later analysis.

Validation of the G'NOME DNA extraction method

We compared our DNA extraction method with our former reference (Godon *et al.*, 1997). Two series of DNA extracts from 12 human faecal samples were prepared by each method. The all-bacteria primers (Table 1) were used to perform PCR to compare both DNA extraction protocols and to validate our method.

Performance of the real-time qPCR protocol in artificial mixtures

To validate the performance of our modified G'NOME DNA extraction protocol and to facilitate real-time qPCR methods, we employed an approach whereby individual samples were spiked with a measured quantity of a known bacterial species. Briefly, several tubes (1 mL) of pure culture *Lactococcus lactis* were centrifuged. Pelleted cells were either stored pure at $-80\,^{\circ}C$ or used to spike otherwise lactococci-free faecal samples before storage. Total bacterial DNA from six pellets and 12 spiked faecal samples was extracted. The resulting levels of *L. lactis* were assessed by real-time qPCR using species-specific 16S rRNA gene primers (Llac05-F: AGCAGTAGGGAATCTTCGGCA and Llac02-R: GGGTAGTTACCGTCACTTGATGAG). The quantitative results from bacterial pellets and spiked faecal samples were compared to validate the performance of our protocol.

Table 1. Group and species-specific 16S rRNA gene-targeted primers and probes used in this study

Target organism	Primers and probes	Sequence 5′–3′	Sources or references
All bacteria*	F_Bact 1369	CGG TGA ATA CGT TCC CGG	Suzuki *et al*. (2000)
	R_Prok1492	TAC GG**C** TAC CTT GTT ACG ACT T	
	P_TM1389F	**6FAM-CTT GTA CAC ACC GCC CGT C**	
C. leptum	F_Clept 09	CCT TCC GTG CCG SAG TTA	This study
	R_Clept 08	GAA TTA AAC CAC ATA CTC CAC TGC TT	
	P-Clep 01	**6FAM-CAC AAT AAG TAA TCC ACC**	
Bifidobacterium	F_Bifid 09c	CGG GTG AGT AAT GCG TGA CC	This study
	R_Bifid 06	TGA TAG GAC GCG ACC CCA	
	P_Bifid	**6FAM-CTC CTG GAA ACG GGT G**	
C. coccoides	F_Ccoc 07	GAC GCC GCG TGA AGG A	This study
	R_Ccoc 14	AGC CCC AGC CTT TCA CAT C	
	P_Erec482*	**VIC-CGG TAC CTG ACT AAG AAG**	Franks *et al*. (1998)
Bacteroides/	F_Bacter 11	CCT WCG ATG GAT AGG GGT T	This study
Prevotella	R_Bacter 08	CAC GCT ACT TGG CTG GTT CAG	
	P_Bac303*	**VIC-AAG GTC CCC CAC ATT G**	Manz *et al*. (1996)
E. coli	E.coli F	CAT GCC GCG TGT ATG AAG AA	Huijsdens *et al*. (2002)
	E.coli R	CGG GTA ACG TCA ATG AGC AAA	
Lactobacillus/Leuconostoc/Pediococcus	F_Lacto 05	AGC AGT AGG GAA TCT TCC A	This study
	R_Lacto 04	CGC CAC TGG TGT TCY TCC ATA TA	
S. salivarius	Stherm 03	TTA TTT GAA AGG GGC AAT TGC T	Furet *et al*. (2004)
	Stherm 08	GTG AAC TTT CCA CTC TCA CAC	
Enterococcus	F_Entero	CCC TTA TTG TTA GTT GCC ATC ATT	Rinttilä *et al*. (2004)
	R_Entero	ACT CGT TGT ACT TCC CAT TGT	

Probe sequences are in bold.
*Modified from reference.

Oligonucleotide primers and probes

TaqMan® qPCR was adapted to quantify total bacteria population in addition to the dominant (> 1% of faecal bacteria population) bacterial species *C. coccoides*, *C. leptum*, *Bacteroides/Prevotella* and *Bifidobacterium*. Quantitative PCR using SYBR-Green® was performed for the subdominant bacterial species *E. coli*, *Streptococcus salivarius*, for the previously described *Enterococcus* group, and for the *Lactobacillus/Leuconostoc/Pediococcus* group. Primers and probes used in this study (Table 1) were designed based on 16S rRNA gene sequences (EMBL database) aligned with the program CLUSTAL W (Thompson *et al*., 1994). Primer design was carried out using PRIMER-EXPRESS version 2.0 (Applied-Biosystems). The specificity of the primers and probes was tested by submitting the sequences to the PROBE MATCH program (Ribosomal Database Project II; Maidak *et al*., 2001). Before laboratory testing, OligoCheck (http://www.bioinformatics-toolkit.org/Dandelion/index.html) was used to examine the *in silico* performance of the PCR systems against 5127 sequences of 16S rRNA gene from type strains of intestinal bacteria. The TaqMan® probes were synthesized by Applied-Biosystems Applera-France. Primers were purchased from MWG (MWG-Biotech AG, Ebersberg, Germany). Primer and probe specificities were further assessed using the real-time qPCR protocol against a series of selected cultured strains (Table 3).

Real-time qPCR

Real-time qPCR was performed using an ABI 7000 Sequence Detection System with software version 1.2.3 (Applied-Biosystems). Amplification and detection were carried out in 96-well plates with TaqMan® Universal PCR 2 × Master Mix (Applied-Biosystems) or with SYBR-Green® PCR 2 × Master Mix (Applied-Biosystems). Each reaction was run in duplicate in a final volume of 25 μL with 0.2 μM final concentration of each primer, 0.25 μM final concentration of each probe and 10 μL of appropriate dilutions of DNA samples. Amplifications were carried out using the following ramping profile: 1 cycle at 95 °C for 10 min, followed by 40 cycles of 95 °C for 30 s, 60 °C for 1 min. For SYBR-Green® amplifications, a melting step was added to improve amplification specificity.

Bacterial strains and growth conditions

The various bacterial strains used to control for the specificity of the primers and probes in this study are shown in Table 3. Bacterial strains were either available in our laboratory collection or were otherwise obtained from the German Collection of Microorganisms and Cell Cultures (DSMZ). Bacteria were cultured aerobically or anaerobically on selective broth as recommended by DSMZ. For each culture, the total number of bacteria, in terms of CFU, was

determined by plating. Aliquots of 1 mL of culture were centrifuged at 12 000 ***g*** for 3 min and the bacterial pellets were stored at − 80 °C before use.

Bacterial DNA extraction, standard curves and quantification

Bacterial genomic DNA used to generate standard curves was extracted twice with the Wizard Genomic DNA Purification Kit (Promega) following the manufacturer's instructions. For the quantification of bacterial species and groups, standard curves were generated from serial dilutions of a known concentration of genomic DNA from each species or group. Standard curves were generated by plotting threshold cycles (C_t) vs. bacterial quantity (CFU). The total number of bacteria (CFU) was interpolated from the averaged standard curves as described previously (Lyons *et al.*, 2000). When PCR was performed on unknown faecal samples, we used these standard curves to quantify each bacterial population. The lower limit for detection for bacterial enumeration with good precision is 10^6 bacteria per gram of stool.

Normalization of qPCR results

In human and animal microbiota, all-bacteria results are presented as the mean of the $\log_{10}$ value ± SEM. To overcome the fact that faecal samples may contain more or less water, we have normalized the data for each faecal sample. The level for each bacterial species or group was subtracted by the level of all-bacteria content. The data are given as the log number of bacteria per gram of faecal sample.

Statistics

On comparing the human microbiota with those of animals, a one-way ANOVA test was performed using JMP® software (Abacus Concepts, Berkeley, CA). When ANOVA indicated a significant result, values were subsequently compared using nonparametric tests (Wilcoxon). Statistical significance was accepted at $P < 0.05$ (P value adjustment method, Holm). Partial least squares (PLS) regression was also used (Moulin-Schouleur *et al.*, 2006) to assess the differences between human and farm animal microbiota (variables Y) on the basis of the qPCR results (variables X). PLS-predictive models using PLS regression were established using the SIMCA software, version 8.1 (Umetri, Umeå, Sweden). The PLS regression between variables X and variables Y yielded the PLS components. These components described the variables X and explained the variables Y. The number of useful PLS components was determined by cross-validation (SIMCA-P 9.0, 2001). The X loadings and the Y loadings were noted as w^* and c, respectively. Groups of strains were presented as situated on a plane defined by the PLS components. The predictive quality of the model was evaluated using the R^2Y coefficient, which corresponded to the proportion of the variance of variables Y explained by variables X.

Results

Validation and performance of DNA extraction

Total bacteria counts, as measured by qPCR, performed on DNA extractions obtained using the former reference method of Godon *et al.* (1997) and our modified G'NOME method were highly similar. Total bacteria levels in the two series of DNA preparations were 11.55 ± 0.1 and 11.44 ± 0.1 logs of enumerated bacterial for the Godon and G'NOME methods, respectively, with no statistical difference. This result indicates that the performance of our technique is equivalent to that of Godon *et al.* (1997).

Performance of the real-time qPCR protocols in artificial mixtures

Population levels of *L. lactis* determined using qPCR on *L. lactis* bacterial pellets and spiked faecal samples were 9.31 ± 0.35 and 9.05 ± 0.39 logs of bacteria, respectively. No significant difference between the two was observed. This result further confirmed the robust nature of the real-time qPCR assay coupled with our DNA extraction method for quantification of bacterial population levels in faecal samples.

Validation of primers and probes

The specificity of all PCR systems (Table 1) was tested by submitting each oligonucleotide sequence to the PROBE MATCH program (Ribosomal Database Project II) (Maidak *et al.*, 2001). This program identifies the target species, if any, matching each PCR system (Table 2). The results from a complementary program, OLIGOCHECK details the number and position of any mismatches (Table 2; positions of mismatches are provided in Supporting Information, Table S1).

We tested the resulting PCR systems specificity against DNA extracted from pure cultures of 48 different strains of bacterial (Table 3). All positive and negative PCR assay results corroborated our *in silico* predictions. For the *Lactobacillus* group, it was not possible to design genus-specific primers because *Leuconostoc* was also detected by the PCR system (Table 3).

Composition of human faecal microbiota assessed by qPCR

For the different targeted bacterial groups, qPCR systems were validated using genomic DNA extracted from the faecal microbiota of healthy human subjects. These results defined a 'standard' profile for dominant and subdominant groups present in the human intestinal microbiota. Dominant

Table 2. Bacterial target species for group or species-specific primers

PCR systems	Target species*
C. leptum group	*Clostridium leptum*† (1), *C. methylpentosum* (2), *C. sporosphaeroides* (2), *Faecalibacterium prausnitzii*† (1) *Ruminococcus albus*† (0), *R. callidus* (0), *R. flavefaciens* (0), *R. bromii* (1) Others: see Table S1
C. coccoides group	*Clostridium coccoides*† (0), *C. aerotolerans* (3), *C. indolis* (4), *C. algidixylanolyticum* (4), *C. aminophilum* (2), *C. aminovalericum* (5), *C. amygdalium* (4), *C. bolteae* (5), *C. celerecrescens* (4), *C. clostridioforme* (2), *C. hathewayi* (3), *C. herbivorans* (2), *C. hylemonae* (2), *C. jejuense* (2), *C. lentocellum* (5), *C. nexile* (2), *C. oroticum* (7), *C. populeti* (2), *C. proteoclasticum* (2), *C. scindens* (2), *C. saccharolyticum* (4), *C. sphenoides* (4), *C. symbiosum* (2), *C. xylanolyticum* (4), *C. xylanovorans* (2) *Eubacterium rectale*† (2), *E. hallii* (3), *E. ruminantium* (2), *E. cellulosolvens* (3), *E. contortum* (3), *E. eligens* (4), *E. ramulus* (4), *E. xylanophilum* (3) *Ruminococcus gnavus*† (2), *R. hansenii*† (0), *R. luti* (0), *R. obeum* (2), *R. hydrogenotrophicus* (3), *R. lactaris* (2), *R. schinkii* (2), *R. torques* (3) Others: see Table S1
Bacteroides/Prevotella group	*Bacteroides fragilis*† (0), *B. vulgatus*† (1), *B. uniformis*† (2), *B. eggerthii*† (2), *B. ovatus*† (1), *B. thetaiotaomicron*† (0), *B. caccae*† (1), *B. acidifaciens* (2), *B. stercoris* (0), *B. plebeius* (0), *B. splanchnicus* (5), *B. salyersiae* (0), *B. nordii* (0), *B. plebeius* (0), *B. coprocola* (0), *B. massiliensis* (1), *B. intestinalis* (2), *B. finegoldii* (0), *B. dorei* (2), *Parabacteroides distasomis* (1) *Prevotella albensis*† (4), *P. bivia* (5), *P. bryantii* (4), *P. buccalis* (5), *P. denticola* (5), *P. disiens* (5), *P. enoeca* (5), *P. heparinolytica* (0), *P. intermedia* (4), *P. melaninogenica* (5), *P. multiformis* (4), *P. nigrescens* (5), *P. oris* (6), *P. oulorum* (5), *P. pallens* (5), *P. salivae* (5), *P. tannerae* (1), *P. veroralis* (5), *P. zoogleoformans* (0)
Bifidobacterium genus	*Bifidobacterium adolescentis*† (0), *B. longum XX* bv. *infantis*† (0), *B. animalis* (0), *B. breve*† (1), *B. choerinum* (0), *B. gallicum* (0), *B. thermacidophilum* (0), *B. boum* (0), *B. merycicum* (0), *B. ruminantium* (0), *B. angulatum* (0), *B. pseudocatenulatum* (0), *B. dentium* (0), *B. gallinarum* (0), *B. saeculare* (0), *B. pullorum* (0), *B. longum* (0), *B. pseudolongum* (0), *B. indicum* (1), *B. bifidum* (1), *B. catenulatum* (2), *B. asteroides* (1), *B. coryneforme* (0), *B. cuniculi* (1), *B. minimum* (0), *B. scardovii* (0), *B. psychraerophilum* (2), *B. subtile* (0) Others: see Table S1
Lactobacillus/Leuconostoc/ Pediococcus group	*Lactobacillus acidophilus*† (0), *L. casei*† (0), *L. paracasei*† (0), *L. delbrueckii*† (0), *L. fermentum*† (0), *L. helveticus*† (0), *L. johnsonii*† (0), *L. plantarum*† (0), *L. rhamnosus*† (0), *L. crispatus*† (0), *L. salivarius*† (0), *L. gasseri*† (0), *L. mucosae*† (0), *L. acetotolerans* (0), *L. acidifarinae* (0), *L. acidipiscis* (0), *L. agilis* (0), *L. alimentarius* (0), *L. amylophilus* (0), *L. amylovorus* (0), *L. antri* (0), *L. aviarius* (0), *L. bifermentans* (0), *L. brevis* (0), *L. buchneri* (0), *L. coleohominis* (0), *L. collinoides* (0), *L. concavus* (0), *L. coryniformis* (0), *L. curvatus* (0), *L. durianis* (0), *L. equi* (0), *L. farciminis* (0), *L. fornicalis* (0), *L. fructivorans* (0), *L. frumenti* (0), *L. fuchuensis* (0), *L. gallinarum* (2), *L. gastricus* (0), *L. graminis* (0), *L. hammesii* (0), *L. harbinensis* (0), *L. hilgardii* (0), *L. homohiochii* (1), *L. ingluviei* (0), *L. intestinalis* (0), *L. jensenii* (0), *L. kalixensis* (0), *L. keferi* (0), *L. kefiranofaciens* (0), *L. kimchii* (0), *L. kitasatonis* (0), *L. kunkeei* (0), *L. lindneri* (2), *L. malefermentans* (0), *L. mali* (0), *L. manihotivorans* (0), *L. mindensis* (0), *L. murinus* (0), *L. pontis* (0) *L. oligofermentans* (0), *L. oris* (0), *L. panis* (0), *L. pantheris* (0), *L. parabrevis* (0), *L. parabuchneri* (0), *L. paracollinoides* (0), *L. parakefiri* (0), *L. paralimentarius* (0), *L. paraplantarum* (0), *L. pentosus* (0), *L. perolens* (0), *L. rennini* (0), *L. reuteri* (0), *L. pseudomesenteroides* (0), *L. rossii* (0), *L. ruminis* (0), *L. sakei* (0), *L. saerimneri* (0), *L. salivarius* (0), *L. sanfranciscensis* (2), *L. vini* (0), *L. satsumensis* (0), *L. sharpeae* (0), *L. siligionis* (0), *L. sobrius* (0), *L. spicheri* (0), *L. suebicus* (0), *L. vaccinostercus* (0), *L. vaginalis* (1), *L. versmoldensis* (0), *L. zeae* (0) *Leuconostoc mesenteroides*† (0), *L. pseudomesenteroides*† (1), *L. durionis* (1), *L. fructosum* (1), *L. ficulneum* (1), *L. gelidum* (1), *L. gasicomitatum* (1), *L. inhae* (1), *L. gelidum* (1), *L. kimchii* (1), *L. lactis* (0), *L. pseudoficulneum* (1), *L. fallax* (1) *Pediococcus inopinatus* (0), *P. parvulus* (0), *P. celliocola* (0), *P. acidilactici* (0), *P. pentosaceus* (0), *P. claussenii* (0), *P. stilesii* (0), *P. dextrinicus* (0)

*Target species were obtained by using PROBE MATCH program (Ribosomal Database Project II) (Maidak *et al*., 2001) by checking each probe and primers with the following data set options: strain, type; source, isolates; size, $\geq$1200 and $<$ 1200 nt; quality, good.
Homology of the TaqMan probe was absolute as described previously (Holland *et al*., 1991). OLIGOCHECK v. 1 (http://www.cf.ac.uk/biosi/research/biosoft) was used to assist in primer design and to confirm the specificity of primers and probes. The maximum mismatch number determined by OLIGOCHECK for the type-strain sequences is shown in parentheses. The positions of mismatches are shown in Table S1.
†Species tested as control in real-time qPCR (c.f. Table 3).

species or groups are defined as those found to represent 1% (-2.0 log no. of bacteria) or more of the faecal bacteria population. *Clostridium leptum*, *C. coccoides* and *Bacteroides/Prevotella* groups are dominant populations (Table 4). Thus, the *Bifidobacterium* population, having a value of -2.4, suggests a subdominant population of human microbiota. This microbiota profile was subsequently used in comparisons against that of farm animals.

Table 3. Specificity of oligonucleotide primers and probes in real-time PCR assessed using pure bacterial culture DNA

		PCR results with each primer set								
Strain	Origin*	Bacteria	*C. leptum*	*C. coccoides*	*Bacteroides/ Prevotella*	*Bifidobacterium*	*E. coli*	*S. salivarius*	*Lactobacillus/Leuconostoc/ Pediococcus*	*Enterococcus*
Clostridium leptum	ATCC 29065	+	+	−	−	−	−	−	−	−
Faecalibacterium prausnitzii	UEPSD L43	+	+	−	−	−	−	−	−	−
Ruminococcus albus	UEPSD M30	+	+	−	−	−	−	−	−	−
Clostridium coccoides	ATCC 29236	+	−	+	−	−	−	−	−	−
Ruminococcus gnavus	ATCC 29149	+	−	+	−	−	−	−	−	−
Ruminococcus hansenii	DSM 20583^T	+	−	+	−	−	−	−	−	−
Eubacterium rectale	UEPSD A4	+	−	+	−	−	−	−	−	−
Bacteroides fragilis	ATCC43185	+	−	−	+	−	−	−	−	−
Bacteroides ovatus	ATCC 8483	+	−	−	+	−	−	−	−	−
Bacteroides thetaiotaomicron	ATCC 29148	+	−	−	+	−	−	−	−	−
Bacteroides uniformis	ATCC 8492	+	−	−	+	−	−	−	−	−
Bacteroides vulgatus	ATCC 8482	+	−	−	+	−	−	−	−	−
Bacteroides caccae	ATCC 43185	+	−	−	+	−	−	−	−	−
Bacteroides eggerthii	UEPSD L78	+	−	−	+	−	−	−	−	−
Prevotella oralis	DSM 20702^T	+	−	−	+	−	−	−	−	−
Prevotella buccae	DSM 20615	+	−	−	+	−	−	−	−	−
Prevotella albensis	DSM 11730^T	+	−	−	+	−	−	−	−	−
Bifidobacterium adolescentis	ATCC15703	+	−	−	−	+	−	−	−	−
Bifidobacterium breve	ATCC15700	+	−	−	−	+	−	−	−	−
Bifidobacterium infantis	ATCC 15697	+	−	−	−	+	−	−	−	−
Escherichia coli	UEPSD S123	+	−	−	−	−	+	−	−	−
Streptococcus salivarius	DSM 20067	+	−	−	−	−	−	+	−	−
Streptococcus thermophilus	DSM 20259	+	−	−	−	−	−	+	−	−
Streptococcus vestibularis	DSM 5636^T	+	−	−	−	−	−	+	−	−
Lactobacillus acidophilus	UEPSD R52	+	−	−	−	−	−	−	+	−
Lactobacillus casei	CNRZ	+	−	−	−	−	−	−	+	−
Lactobacillus paracasei	CNRZ	+	−	−	−	−	−	−	+	−
Lactobacillus delbrueckii	CNRZ	+	−	−	−	−	−	−	+	−
Lactobacillus fermentum	CNRZ	+	−	−	−	−	−	−	+	−
Lactobacillus johnsonii	CNRZ	+	−	−	−	−	−	−	+	−
Lactobacillus plantarum	CNRZ	+	−	−	−	−	−	−	+	−
Lactobacillus rhamnosus	UEPSD R11	+	−	−	−	−	−	−	+	−
Lactobacillus helveticus	CNRZ	+	−	−	−	−	−	−	+	−

Lactobacillus crispatus	DSM 20584^T	+	−	−	−	−	−	−	+	−
Lactobacillus salivarius	DSM 20555^T	+	−	−	−	−	−	−	+	−
Lactobacillus gasseri	DSM 20243^T	+	−	−	−	−	−	−	+	−
Lactobacillus mucosae	DSM 13345^T	+	−	−	−	−	−	−	+	−
Leuconostoc mesenteroides	CNRZ	+	−	−	−	−	−	−	+	−
L. pseudomesenteroides	CNRZ	+	−	−	−	−	−	−	+	−
Enterococcus faecium	UEPSD L99	+	−	−	−	−	−	−	−	+
Enterococcus faecalis	UEPSD L98	+	−	−	−	−	−	−	−	+
Clostridium perfringens	ATCC 13124	+	−	−	−	−	−	−	−	−
Clostridium sordelii	VPI 9048	+	−	−	−	−	−	−	−	−
Atopobium parvulum	UEPSD B69	+	−	−	−	−	−	−	−	−
Atopobium vaginae	DSM 15829^T	+	−	−	−	−	−	−	−	−
Atopobium rimae	DSM 7090^T	+	−	−	−	−	−	−	−	−
Clostridium bifermentans	NCTC 506	+	−	−	−	−	−	−	−	−
Streptococcus gordonii	DSM 6777^T	+	−	−	−	−	−	−	−	−

*ATCC, DSM, VPI and NCTC referred to the strain names in commercial collections. UEPSD and CNRZ corresponded to two INRA collections in Jouy-en-Josas.
+, positive; −, negative.

Table 4. Composition of human faecal microbiota compared with farm animal microbiota

		TaqMan detection					SYBR-Green detection			
	n	All-bacteria*	*C. leptum* group†	*C. coccoides* group†	*Bacteroides/Prevotella* group†	*Bifidobacterium* genus†	*Lactobacillus/Leuconostoc/ Pediococcus* group†	*E. coli*†	*S. salivarius* species†	*Enterococcus* genus†
Human	21	11.5 ± 0.1	− 0.7 ± 0.05 (A)	− 1.3 ± 0.08 (A)	− 1.5 ± 0.06 (A)	− 2.4 ± 0.33 (A)	− 3.9 ± 0.13 (A)	− 3.8 ± 0.34 (A)	− 3.1 ± 0.12 (A)	− 5.0 ± 0.15
Horse	5	11.5 ± 0.1	− 1.5 ± 0.06 (B)	− 2.2 ± 0.18 (B)	− 2.3 ± 0.04 (B)	− 4.8 ± 0.13 (B)	− 2.4 ± 0.82 (A)	− 5.0 ± 0.03 (B)	− 5.2 ± 0.20 (B)	Not detected
Cow	5	11.4 ± 0.1	− 1.0 ± 0.03 (B)	− 2.6 ± 0.03 (B)	− 2.3 ± 0.01 (B)	− 3.6 ± 0.37 (B)	− 3.1 ± 0.06 (A)	− 5.0 ± 0.31 (B)	− 5.0 ± 0.04 (B)	Not detected
Goat	5	12.0 ± 0.1	− 1.0 ± 0.07 (B)	− 2.2 ± 0.11 (B)	− 2.4 ± 0.19 (B)	− 1.8 ± 0.26 (A)	− 3.2 ± 0.78 (A)	− 4.5 ± 0.48 (A)	− 4.3 ± 0.43 (B)	Not detected
Rabbit	5	11.7 ± 0.1	− 0.7 ± 0.03 (A)	− 1.9 ± 0.03 (B)	− 1.2 ± 0.09 (A)	− 1.6 ± 0.07 (A)	− 5.1 ± 0.59 (A)	Not detected	Not detected	Not detected
Sheep	5	11.9 ± 0.1	− 1.0 ± 0.05 (B)	− 2.7 ± 0.08 (B)	− 2.4 ± 0.08 (B)	− 4.2 ± 0.11 (B)	− 5.3 ± 0.60 (A)	− 4.1 ± 0.52 (A)	Not detected	Not detected
Pig	5	11.9 ± 0.1	− 1.2 ± 0.11 (B)	− 1.7 ± 0.35 (A)	− 1.9 ± 0.17 (A)	− 3.4 ± 0.69 (B)	− 1.2 ± 0.54 (B)	− 2.7 ± 0.06 (B)	Not detected	Not detected

n represents the numbers of studied samples.
The reference for the statistics is with human faecal samples. The nonparametric Wilcoxon test was performed if the one-way ANOVA for the bacterial group was significant.
Data not sharing the same letter within a column are significantly different to the human population, at $P < 0.05$.
*All-bacteria results obtained by qPCR were expressed as the mean of the $\log_{10}$ value ± SEM.
†Results were expressed as the mean of the $\log_{10}$ value ± SEM of normalized data, calculated as the log no. of targeted bacteria minus the log of all-bacteria number.

Comparison of bacterial populations in stools from human and farm animals

Differences in the bacterial composition of animal stool samples compared with those found in the human faecal microbiota were assessed using qPCR (Table 4). Global one-way ANOVA testing showed significant differences in bacterial compositions between the two groups.

The nonparametric Wilcoxon test was used to reveal whether each qPCR system allows for discrimination of the bacterial population of humans and animals. This statistical test can also show how animal microbiota differs from human. The *C. leptum* qPCR system revealed several significant differences between human and horse, cow, goat, and sheep microbiota (Table 4). When comparing results between human and rabbit microbiota for the *C. leptum* group, no significant difference was observed (Table 4).

Although unable to distinguish between the microbiota of human and pig, the *C. coccoides* group qPCR system produced significantly different results for all other animals, with values being higher than that of human (Table 4).

The *Bacteroides*/*Prevotella* group displayed the same type of enrichment as *C. coccoides* for horse, cow, goat and sheep microbiota. Two exceptions were noted, however, in rabbit and pig, where no statistical difference with respect to human samples was observed (Table 4).

We also found the *Bifidobacterium* genus to vary significantly in the faeces of horse, cow, sheep and pig compared with human (Table 4). The *Bifidobacterium* population in goat and rabbit faeces were similar in relation to human and showed the lowest normalized data (Table 4).

The *Lactobacillus*/*Leuconostoc*/*Pediococcus* group failed to discriminate the microbiota of animals and human, with the sole exception being for pig samples. It is important to note that the targeted lactobacilli population in pig microbiota showed the lowest normalized result (Table 4).

The *E. coli* species qPCR system could distinguish human and animal microbiota except in the cases of goat and sheep. Our study showed that the *E. coli* value in pig microbiota is the lowest (-2.7 log no. of bacteria) when compared with those of animals and humans, and was not detected in the faecal samples of rabbit (Table 4). *Streptococcus salivarius* species was also not detected in faecal samples of rabbit, in addition to being absent from both sheep and pig. Nevertheless, the results show that *S. salivarius* can be used to distinguish the human microbiota from those of horse, cow and goat (Table 4). *Streptococcus salivarius* was more abundant in human faecal samples than in the other faecal samples. The *Enterococcus* species could not be detected in any animal faecal sample in contrast to its presence in human samples (Table 4).

PLS regression analysis based on faecal microbiota composition assessed using real-time qPCR confirmed that the human faecal microbiota could be clearly differentiated from that of farm animals in the 95% probability region (Fig. 1a). The first two components of the PLS model explained 85% of the variation of the *Y*-matrix, indicating a good separation of the human group compared with the groups of farm animals. The *X* loadings (w^*) corresponding to faecal microbiota quantifications and the *Y* loadings (c) corresponding to the human and farm animal groups are presented in Fig. 1b. PLS regression analysis demonstrated that the *C. coccoides* group, *Enterococcus* genus and *S. salivarius* species characterize the human faecal microbiota and *Lactobacillus/Leuconostoc/Pediococcus* characterize the pig faecal microbiota.

Discussion

Pollution by human and animal faeces harbouring potential human pathogens represents a serious environmental threat that affects many natural waters. Waters contaminated with human faeces, in particular, are generally considered to represent a greater risk for human health as they contain human-specific enteric pathogens (Baudart *et al.*, 2000; Koopmans & Duizer, 2004; Godfree & Farrell, 2005). Animals can also serve as reservoirs for numerous enteric pathogens (Hancock *et al.*, 2001; Brown *et al.*, 2004; Cox *et al.*, 2005). Given this complex situation, the ability to accurately track faecal contamination in the environment and identify its origin is of great importance. The key points of such a technique are the choice of reliable and differential faecal indicators and the development of quantitative microbial source tracking methods.

To address these requirements, a robust and reproducible protocol is required to quantify bacterial species and groups in faecal samples originating from different possible contamination sources. Matsuki *et al.* (2004) were the first to apply qPCR, based on 16S rRNA gene quantification, to analyse the diversity of human intestinal *Bifidobacterium*. In our work, employing an optimized protocol, we quantified equivalent numbers of *Bifidobacterium* in human samples, compared with Matsuki and colleagues. This corroborative result gave us confidence in expanding the use of the qPCR technique to compare the whole human faecal microbiota with that of animals.

One additional variable that, in some cases, could influence the measurement and comparison of different groups of bacteria is the water content of each faecal sample. Low water content, for example, could contribute to the high bacterial concentration observed in goat and sheep samples. To overcome this potential variable, we normalized our data using all-bacteria populations.

As discussed below, our data are consistent with a number of smaller-scale investigations which focused on individual

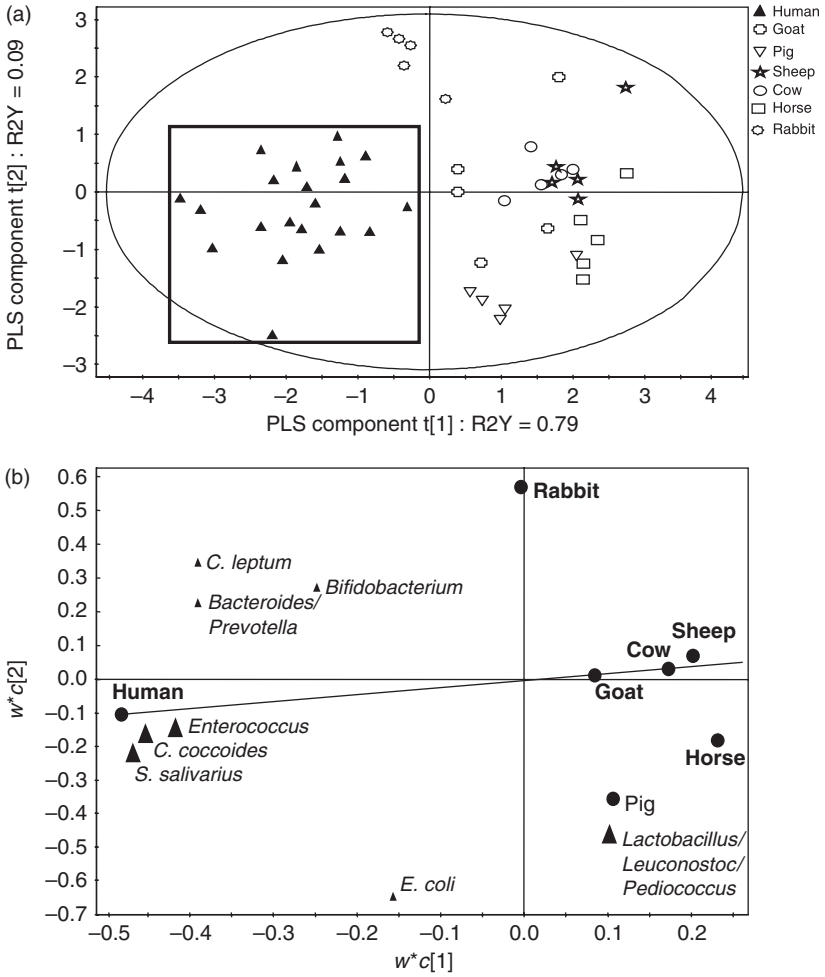


Fig. 1. PLS discrimination between microbiota of human and farm animals. (a) Relationship between faecal microbiota (variables *X*) and human or farm animals (variables *Y*) using PLS regression. The cross-validation led to two components represented here as *t*(1) and *t*(2). The corresponding PLS model explains 80.0% of the variation of the *Y*-matrix. The 95% probability region defined by the model is delimited by the ellipse. The human (▲) group ($n = 21$) can be distinguished and is delimited by the black square. ○, cow ($n = 5$); □, horse ($n = 5$); ▽, pig ($n = 5$); ◌, rabbit ($n = 5$); ☆, sheep ($n = 5$); ⬡, goat ($n = 5$). (b) The window shows the *X* loadings (w^*) of the *X* variables (faecal microbiota quantifications) and the *Y* loadings (*c*) of the *Y* variables (human and animal groups), and thereby shows the correlation between *X* and *Y*. The *X* (black triangles) and *Y* (black circles) variables combine in the projections, and the *X* variables relate to the *Y* variables, as shown in the figure. The *Clostridium coccoides* group, *Streptococcus salivarius* species and *Enterococcus* genus, significant for the discrimination of human and farm animals, and the *Lactobacillus/Leuconostoc/Pediococcus* group, characterizing the pig microbiota, are denoted by large black triangles (small black triangles represent less significant *X* variables).

farm species or targeted groups of bacteria. In our study, we observed that the pig faecal microbiota is characterized by a population of *Lactobacillus/Leuconostoc/Pediococcus* higher than that found in other animals or humans. Given the value of -1.2 log no. of bacteria, this population could be considered dominate in pig microbiota. These data are in agreement with the observation by Castillo *et al.* (2006) showing a high level of *Lactobacillus* in the upper gastrointestinal tract of pig. These results, combined with those obtained for *E. coli*, suggest that both populations can be considered important in pig microbiota.

Canzi *et al.* (2000) enumerated *Bacteroides* and *Clostridium* in rabbit faeces. We also found the same range of populations for the *Bacteroides/Prevotella* group. However, for *Clostridium* populations, our study indicated higher colonization levels (about 6 logs higher) than those observed by these authors. This discrepancy could be due to methodological differences as Canzi and colleagues used spore enumerations for their *Clostridia* estimation. The fact that our technique enumerates vegetative cells as well as non-cultivable bacteria is the most likely explanation for the higher concentration observed. Moreover, our PCR system also detected *Eubacteria* and *ruminococci* species which are part of the *Clostridium* group.

For equine microbiota, our results are consistent with a previous study (Daly & Shirazi-Beechey, 2003) where the authors used oligonucleotide probes in hybridization assays. Daly and Shirazi-Beechey found no *Bifidobacterium* and observed that the *Eubacterium rectale–C. coccoides* group, combined with *Spirochaetaceae* and the *Cytophaga–Flexibacter–Bacteroides* assemblage, represented the largest colonized populations (10–30%). The authors further noted that the *Bacillus–Lactobacillus–Streptococcus* group with *Fibrobacter* constituted 1–10% of the total microbiota in horse samples.

It is likely that the bacterial biodiversity of the equine microbiota compared with human contributes to the significant differences in bacterial quantification. Quantitative PCR developed to detect intestinal bacteria in human samples further highlight the species specificity of our protocols and the fact that the bacterial biodiversity of the equine microbiota is notably different from that of human.

Several studies have also reported on the bovine intestinal microbiota. Stahl *et al.* (1988) used species- and group-specific 16S rRNA gene-targeted probes for enumeration of two species (*Fibrobacter succinogenes* and *Lachnospira ruminicola*) in the rumen of animals treated with antibiotics. Tajima *et al.* (2001) used qPCR to quantify several *Prevotella* and some *Ruminococcus*, *Fibrobacter*, and *Eubacterium* species in the rumen. In 2005, An *et al.* estimated the prokaryote diversity in the rumen of yak (*Bos grunniens*) and Jinnan cattle (*Bos taurus*) by 16S rRNA gene sequence homology analysis. Their results showed a prevalence of *Bacteroides*; however, no sequence was related to *Ruminococcus albus* (a species of the *C. leptum* group) in the yak and cow rumen. In our study, the level of *Bacteroides/Prevotella* population presents a normalized difference of -2.3 log number of bacteria and cannot be regarded as a dominant population, while *C. leptum* group shows only -1.0 log number of bacteria and is part of the dominant population. Whitford *et al.* (1998) and Ozutsumi *et al.* (2005) presented a phylogenetic analysis of rumen bacteria by comparative sequence analysis of cloned 16S rRNA gene. Approximately 30% of the sequences were related to bacteria of the *Bacteroides/Prevotella* group, most of which clustered with *Prevotella ruminicola*. The remaining sequences clustered with members of the *Clostridium* genus. The differences observed with our findings are likely due to different technical approaches and/or diversity of microbiota among bovine herds.

To our knowledge, no previous study has used qPCR techniques to describe and compare the intestinal microbiota between animal and human. Our qPCR systems, checked *in silico* by OLIGOCHECK against RDP databases, were successfully able to discriminate different intestinal microbiota.

Our global comparison between human and farm animal microbiota provides data to select host-specific bacterial groups and alternative faecal indicators from all hosts considered.

Our PLS regression analysis showed that the *C. coccoides* group, *Enterococcus* genus and *S. salivarius* species could be considered as specific markers for human faecal microbiota and that *Lactobacillus/Leuconostoc/Pediococcus* can be used as a specific marker of pig microbiota.

The *C. leptum* group was found to have the lowest normalized data in humans and animals and thus represents a promising candidate for use as a reliable faecal indicator. It is largely distributed among animal species and in humans and has also been linked with diseases (Manichanh *et al.*, 2006; Sokol *et al.*, 2006). Our study also shows high concentrations of *Bacteroides/Prevotella* and *Bifidobacterium* in all host faecal samples tested. Such anaerobic bacteria do not persist for long periods of time in aerobic waters and are generally unable to multiply in such conditions (Fiksdal *et al.*, 1985; Kreader, 1998). These inherent physiological characteristics make the *Bacteroides* and *Bifidobacterium* excellent candidates for detecting faecal contamination in the environment. Integrated within these two dominant bacterial groups are several species that were found to be host-specific in several studies (Bernhard & Field, 2000a; Bonjoch *et al.*, 2004; Dick *et al.*, 2005). Host-specific *Bacteroides* markers were developed (Bernhard & Field, 2000b; Dick *et al.*, 2005) and applied in a watershed in the United States (Shanks *et al.*, 2006). They were also validated on French faecal and environmental samples (Gourmelon *et al.*, 2007). Quantitative PCR assays are currently in progress and some results have already been published for human and bovine-specific *Bacteroides* (Seurinck *et al.*, 2005; Reischer *et al.*, 2006).

Among the teams who have studied the microbiota of animals over the last decade none, up to now, has presented a global comparison of the faecal microbiota composition of humans and animals. Our results are thus promising in advancing the goal to define a discrete set of host-specific faecal microbiota biomarkers. Additional investigations are continuing to refine a set of comprehensive, reliable, and predictive host-specific markers.

Acknowledgement

The authors thank Valeria Dellaretti Guimarães and Sean P. Kennedy for critical reading of this manuscript.

References

An D, Dong X & Dong Z (2005) Prokaryote diversity in the rumen of yak (*Bos grunniens*) and Jinnan cattle (*Bos taurus*) estimated by 16S rDNA homology analyses. *Anaerobe* **4**: 207–215.

Baudart J, Gradulos J, Barusseau JP & Lebaron P (2000) *Salmonella* spp. and fecal coliform loads in coastal waters from a point vs. nonpoint source of pollution. *J Environ Qual* **29**: 241–250.

Bernhard AE & Field KG (2000a) Identification of nonpoint sources of fecal pollution in coastal waters by using host-specific 16S ribosomal DNA genetic markers from fecal anaerobes. *Appl Environ Microbiol* **66**: 1587–1594.

Bernhard AE & Field KG (2000b) A PCR assay to discriminate human and ruminant feces on the basis of host differences in *Bacteroides–Prevotella* genes encoding 16S rRNA. *Appl Environ Microbiol* **66**: 4571–4574.

Bonjoch X, Ballesté E & Blanch AR (2004) Multiplex PCR with 16S rRNA gene-targeted primers of *Bifidobacterium* spp. to identify sources of fecal pollution. *Appl Environ Microbiol* **70**: 3171–3175.

Brown PE, Christensen OF, Clough HE *et al.* (2004) Frequency and spatial distribution of environmental *Campylobacter* spp. *Appl Environ Microbiol* **70**: 6501–6511.

Canzi E, Zanchi R, Camaschella P, Cresci A, Greppi GF, Orpianesi C, Serrantoni M & Ferrari A (2000) Modulation by lactic-acid bacteria of the intestinal ecosystem and plasma cholesterol in rabbit fed a casein diet. *Nutr Res* **22**: 1329–1340.

Castillo M, Martin-Orue SM, Manzanilla EG, Badiola I, Martin M & Gasa J (2006) Quantification of total bacteria, enterobacteria and lactobacilli populations in pig digesta by real-time PCR. *Vet Microbiol* **114**: 165–170.

Cox P, Griffith M, Angles M, Deere D & Ferguson C (2005) Concentrations of pathogens and indicators in animal feces in the Sydney watershed. *Appl Environ Microbiol* **71**: 5929–5934.

Daly K & Shirazi-Beechey SP (2003) Design and evaluation of group-specific oligonucleotide probes for quantitative analysis of intestinal ecosystems: their application to assessment of equine colonic microflora. *FEMS Microbiol Ecol* **44**: 243–252.

Dick LK, Bernhard AE, Brodeur TJ, Santo Domingo JW, Simpson JM, Walters SP & Field KG (2005) Host distributions of uncultivated fecal *Bacteroidales* bacteria reveal genetic markers for fecal source identification. *Appl Environ Microbiol* **71**: 3184–3191.

Directive 2006/7/CE of the European Parliament and of the Council of 15 February 2006 concerning the management of bathing water quality and repealing Directive 76/160/EEC. *Off J Eur Union* **L64**: 37–51.

Directive 2006/113/CE of the European Parliament and of the Council of 12 December 2006 on the quality required of shellfish waters. 27/12/2006. *Off J Eur Union* **L376**: 14–20.

Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE & Relman DA (2005) Diversity of the human intestinal microbial flora. *Science* **308**: 1635–1638.

Fiksdal L, Maki JS, LaCroix SJ & Staley JT (1985) Survival and detection of *Bacteroides* spp., prospective indicator bacteria. *Appl Environ Microbiol* **49**: 148–150.

Franks AH, Harmsen HJ, Raangs GC, Jansen GJ, Schut F & Welling GW (1998) Variations of bacterial populations in human feces measured by fluorescent *in situ* hybridization with group-specific 16S rRNA-targeted oligonucleotide probes. *Appl Environ Microbiol* **64**: 3336–3345.

Furet JP, Quenee P & Tailliez P (2004) Molecular quantification of lactic acid bacteria in fermented milk products using real-time quantitative PCR. *Int J Food Microbiol* **2**: 197–207.

Godfree A & Farrell J (2005) Processes for managing pathogens. *J Environ Qual* **34**: 105–113.

Godon JJ, Zumstein E, Dabert P, Habouzit F & Moletta R (1997) Molecular microbial diversity of an anaerobic digestor as determined by small-subunit rDNA sequence analysis. *Appl Environ Microbiol* **63**: 2802–2813.

Gordon DM & Cowling A (2003) The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. *Microbiology* **149**: 3575–3586.

Gourmelon M, Caprais MP, Ségura R, Le Mennec C, Lozach S, Piriou JP & Rincé A (2007) Evaluation of two library-independent microbial source tracking methods to identify sources of fecal contamination in French estuaries. *Appl Environ Microbiol* **73**: 4857–4866.

Hancock D, Besser T, Lejeunes J, Davis M & Rice D (2001) The control of VTEC in the animal reservoir. *Int J Food Microbiol* **66**: 71–78.

Harmsen HJ, Raangs GC, He T, Degener JE & Welling GW (2002) Extensive set of 16S rRNA-based probes for detection of bacteria in human feces. *Appl Environ Microbiol* **6**: 2982–2990.

Holland PM, Abramson RD, Watson R & Gelfand DH (1991) Detection of specific polymerase chain reaction product by utilizing the 5′–3′ exonuclease activity of Thermus aquaticus DNA polymerase. *P Natl Acad Sci USA* **15**: 7276–7280.

Huijsdens XW, Linkens RK, Mak M, Neuwissen SG, Vanderbroucke-Grauls CM & Savelkoul PH (2002) Quantification of bacteria adherent to gastrointestinal mucosa by real-time PCR. *J Clin Micobiol* **40**: 4423–4427.

Hörman A, Rimhanen-Finne R, Maunula L, von Bonsdorff CH, Torvela H, Heikinheimo A & Hänninen ML (2004) *Campylobacter* spp., *Giardia* spp., *Cryptosporidium* spp., noroviruses, and indicator organisms in surface water in southwestern Finland, 2000–2001. *Appl Environ Microbiol* **70**: 87–95.

Koopmans M & Duizer E (2004) Foodborne viruses: an emerging problem. *Int J Food Microbiol* **90**: 23–41.

Kreader CA (1998) Persistence of PCR-detectable *Bacteroides distasonis* from human feces in river water. *Appl Environ Microbiol* **64**: 4103–4105.

Lan PTN, Hayashi H, Sakamoto M & Benno Y (2002) Phylogenetic analysis of cecal microbiota in chicken by the use of 16S rDNA clone libraries. *Microbiol Immunol* **46**: 371–382.

Lay C, Rigottier-Gois L, Holmstrom K *et al.* (2005b) Colonic microbiota signatures across five northern European countries. *Appl Environ Microbiol* **7**: 4153–4155.

Lay C, Sutren M, Rochet V, Saunier K, Dore J & Rigottier-Gois L (2005a) Design and validation of 16S rRNA probes to enumerate members of the *Clostridium leptum* subgroup in human faecal microbiota. *Environ Microbiol* **7**: 933–946.

Leser TD, Amenuvor JZ, Jensen TK, Lindecrona RH, Boye M & Moller K (2002) Culture-independent analysis of gut bacteria: the pig gastrointestinal tract microbiota revisited. *Appl Environ Microbiol* **68**: 673–690.

Lyons SR, Griffen AL & Leys EJ (2000) Quantitative real-time PCR for *Porphyromonas gingivalis* and total bacteria. *J Clin Microbiol* **6**: 2362–2365.

Maidak BL, Cole JR, Lilburn TG, Parker CT Jr, Saxman PR, Farris RJ, Garrity GM, Olsen GJ, Schmidt TM & Tiedje JM (2001) The RDP-II (Ribosomal Database Project). *Nucleic Acids Res* **1**: 173–174.

Manichanh C, Rigottier-Gois L, Bonnaud E *et al.* (2006) Reduced diversity of fecal microbiota in Crohn's disease revealed by a métagenomic approach. *Gut* **55**: 2005–2011.

Manz W, Amann R, Ludwig W, Vancanneyt M & Schleifer KH (1996) Application of a suite of 16S rRNA-specific oligonucleotide probes designed to investigate bacteria of the

phylum *Cytophaga–Flavobacter–Bacteroides* in the natural environment. *Microbiology* **142**: 1097–1106.

Martinez-Urtaza J, Saco M, de Novoa J, Perez-Pineiro P, Peiteado J, Lozano-Leon A & Garcia-Martin O (2004) Influence of environmental factors and human activity on the presence of *Salmonella* serovars in a marine environment. *Appl Environ Microbiol* **70**: 2089–2097.

Matsuki T, Watanabe K, Fujimoto J, Kado Y, Takada T, Matsumoto K & Tanaka R (2004) Quantitative PCR with 16S rRNA-gene-targeted species-specific primers for analysis of human intestinal bifidobacteria. *Appl Environ Microbiol* **70**: 167–173.

Moulin-Schouleur M, Schouler C, Tailliez P, Kao MR, Bree A, Germon P, Oswald E, Mainil J, Blanco M & Blanco J (2006) Common virulence factors and genetic relationships between O18:K1:H7 *Escherichia coli* isolates of human and avian origin. *J Clin Microbiol* **10**: 3484–3492.

Ozutsumi Y, Hayashi H, Sakamoto M, Itabashi H & Benno Y (2005) Culture-independent analysis of fecal microbiota in cattle. *Biosci Biotech Bioch* **9**: 1793–1797.

Pommepuy M, Butin M, Derrien A, Gourmelon M, Colwell RR & Cormier M (1996) Retention of enteropathogenicity by viable but nonculturable *Escherichia coli* exposed to seawater and sunlight. *Appl Environ Microbiol* **62**: 4621–4626.

Potasman I, Paz A & Odeh M (2002) Infectious outbreaks associated with bivalve shellfish consumption: a worldwide perspective. *Clin Infect Dis* **35**: 921–928.

Reischer GH, Kasper DC, Steinborn R, Mach RL & Farnleitner AH (2006) Quantitative PCR method for sensitive detection of ruminant fecal pollution in freshwater and evaluation of this method in Alpine karstic regions. *Appl Environ Microbiol* **72**: 5610–5614.

Rinttilä T, Kassinen A, Malinen E, Krogius L & Palva A (2004) Development of an extensive set of 16S rDNA-targeted primers for quantification of pathogenic and indigenous bacteria in faecal samples by real-time PCR. *J Appl Microbiol* **97**: 1166–1177.

Roszak DB & Colwell RR (1987) Metabolic activity of bacterial cells enumerated by direct viable count. *Appl Environ Microbiol* **53**: 2889–2893.

Savichtcheva O & Okabe S (2006) Alternative indicators of fecal pollution: relations with pathogens and conventional indicators, current methodologies for direct pathogen monitoring and future application perspectives. *Water Res* **40**: 2463–2476.

Seurinck S, Defoirdt T, Verstraete W & Siciliano SD (2005) Detection and quantification of the human-specific HF183 *Bacteroides* 16S rRNA genetic marker with real-time PCR for assessment of human faecal pollution in freshwater. *Environ Microbiol* **7**: 249–259.

Shanks OC, Nietch C, Simonich M, Younger M, Reynolds D & Field KG (2006) Basin-wide analysis of the dynamics of fecal contamination and fecal source identification in Tillamook Bay, Oregon. *Appl Environ Microbiol* **72**: 5537–5546.

SIMCA-P9.0 (2001) A new standard in multivariate data analysis. *User's Guide and Tutorial*, pp. 122. Umetrics, Umeå, Sweden.

Sokol H, Seksik P, Rigottier-Gois L, Lay C, Lepage P, Podglajen I, Marteau P & Dore J (2006) Specificities of the fecal microbiota in inflammatory Bowel disease. *Inflamm Bowel Dis* **12**: 106–111.

Stahl DA, Flesher B, Mansfield HR & Montgomery L (1988) Use of phylogenetically based hybridization probes for studies of ruminal microbial ecology. *Appl Environ Microbiol* **5**: 1079–1084.

Suau A, Bonnet R, Sutren M, Godon JJ, Gibson GR, Collins MD & Dore J (1999) Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Appl Environ Microbiol* **65**: 4799–4807.

Suzuki MT, Taylor LT & DeLong EF (2000) Quantitative analysis of small-subunit rRNA genes in mixed microbial populations via 5′-nuclease assays. *Appl Environ Microbiol* **11**: 4605–4614.

Tajima K, Aminov RI, Nagamine T, Matsui H, Nakamura M & Benno Y (2001) Diet-dependent shifts in the bacterial population of the rumen revealed with real-time PCR. *Appl Environ Microbiol* **6**: 2766–2774.

Thompson JD, Higgins DG & Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673–4680.

Wheeler AE, Burke J & Spain A (2003) Fecal indicator bacteria are abundant in wet and at freshwater beaches. *Water Res* **37**: 3978–3982.

Whitford MF, Foster RJ, Beard CE, Gong J & Teather RM (1998) Phylogenetic analysis of rumen bacteria by comparative sequence analysis of cloned 16S rRNA genes. *Anaerobe* **4**: 153–163.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Sequence alignment of the species targeted by OLIGOCHECK software showing sequence differences.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Table S1: Sequence alignment of the species targeted by OligoCheck software showing sequence differences. Only those nucleotides that are different from the target sequence for each organism are shown.

[(1)]. As indicated, there were two exceptions to the rule of absolute homology for the TaqMan probes based on this silico assessment. The type strain *P. oralis* DSM20702^{T} and the strain *P. buccae* DSM20615 have one and two C/T mismatches, respectively, and are positive in PCR assays (Table 3).

PCR assays: positive: positive PCR, negative: negative PCR, nt: no tested

Clostridium leptum group

Species	EMBL Access Number	F_Clept09 primer 5′ 3′ CCTTCCGTGCCGSAGTTA	P_Clept01 probe 5′ 3′ CACAATAAGTAATCCACC	R_Clept08 primer 5′ 3′ GAATTAAACCACATACTCCACTGCTT	PCR assay
Clostridium					
C. leptum	AJ305238	--C----------	----------	----------	positive
C. methylpentosum	Y18181	---T------G-----	----------	----------	nt
C. sporosphaeroides	M59116	-N----------G-----	----------	----------	nt
F. prausnitzii	AJ413954	-----A----------	----------	----------	positive
Ruminococcus					
R. callidus	L76596	----------	----------	----------	nt
R. flavefaciens	L76603	----------	----------	----------	nt
R. albus	L76598	----------	----------	----------	positive
R. bromii	L76600	----------G-----	----------	----------	nt
Eubacterium					
Eub. siraeum	L34625	----------	----------	----------	nt
Other genera					
Ac. elongatum	AY487928	----------	----------	----------	nt
Anaerotruncus colihominis	AJ315980	----------	----------	----------	nt
An. pentosovorans	X97852	-T----------	----------	----------	nt
An. Agile	X98011	-TC----------	----------	----------	nt

Bifidobacterium genus

Species	EMBL Access Number	F_Bifid 09c primer 5′ 3′ CGGGTGAGTAATGCGTGACC	P_Bifid probe 5′ 3′ CTCCTGGAAACGGGTG	R_Bifid 06 primer 5′ 3′ TGATAGGACGCGACCCCA	PCR assay
Bifidobacterium					
B. infantis	D86184	--------------------	----------------	----------------	positive
B. animalis	X89513	--------------------	----------------	----------------	nt
B. choerinum	D86186	--------------------	----------------	----------------	nt
B. gallicum	D86189	--------------------	----------------	----------------	nt
B. thermacidophilum	AY148470	--------------------	----------------	----------------	nt
B. boum	D86190	--------------------	----------------	----------------	nt
B. merycicum	D86192	--------------------	----------------	----------------	nt
B. ruminantium	D86197	--------------------	----------------	----------------	nt
B. angulatum	D86182	--------------------	----------------	----------------	nt
B. pseudocatenulatum	D86187	--------------------	----------------	----------------	nt
B. dentium	D86183	--------------------	----------------	----------------	nt
B. gallinarum	D86191	--------------------	----------------	----------------	nt
B. saeculare	D89328	--------------------	----------------	----------------	nt
B. pullorum	D86196	--------------------	----------------	----------------	nt
B. longum	M58739	--------------------	----------------	----------------	nt
B. pseudolongum	M58742	--------------------	----------------	---------N-------	nt
B. indicum	M58737	-----N--------------	----------------	---------N-------	nt
B. breve	M58731	----N---------------	----------------	----------------	positive
B. bifidum	M38018	--------------------	----------------	---------N-------	nt
B. adolescentis	M58729	--------------------	----------------	----------------	positive
B. catenulatum	M58732	---NN---------------	----------------	---------N-------	nt
B. asteroides	M58730	--------------------	----------------	----------------N	nt
B. coryneforme	M58733	--------------------	----------------	----------------	nt
B. cuniculi	M58734	--------------------	----------------	----------------	nt
B. minimum	M58741	--------------------	----------------	----------------	nt
B. scardovii	AJ307005	--------------------	----------------	----------------	nt
B. psychraerophilum	AY174108	------------A------T	----------------	----------------	nt
B. subtile	D89378	--------------------	----------------	----------------	nt
Others					
Brevundimonas diminuta	M59064	-----------CA----G-A	----------------	A---A------GG---GC	nt
Mycoplana bullata	D12785	-----------CA----G-A	----------------	A---A------GG---GC	nt
Parascardovia denticolens	D89331	-------------------T	----------------	----------------	nt
Scardoviaino pinata	AB029087	-------------------T	----------------	----------------	nt

Clostridium coccoides group

Species	EMBL Access Number	F_Ccocc07 primer 5′ 3′ GACGCCGCGTGAAGGA	P_Erec482 probe 5′ 3′ CGGTACCTGACTAAGAAG	R_Ccocc14 primer 5′ 3′ AGCCCCAGCCTTTCACATC	PCR assay
Clostridium					
C. coccoides	M59090	----------	----------N-----	--N----------	positive
C. aerotolerans	X76163	----------GT--	----------	--------G-------TC-	nt
C. algidixylanolyticum	AF092549	----------CT--	----------	------G-G-------TC-	nt
C. aminophilum	L04165	----------GC--	----------	------G-G----------	nt
C. aminovalericum	X73436	----------GT--	----------	------G--NN-----AC-	nt
C. amygdalinum	AY353957	----------CT--	----------	------G-G-------TC-	nt
C. bolteae	AJ508452	----------GT--	----------	-----TG-GT------T--	nt
C. celerecrescens	X71848	----------GT--	----------	------G-G-------TC-	nt
C. clostridioforme	M59089	----------GT--	----------	------G-G----------	nt
C. hathewayi	AJ311620	----------GT--	----------	------G-G-------T--	nt
C. herbivorans	L34418	----------GT--	----------	--A---------C------	nt
C. hylemonae	AB023973	----------	----------	------G-G----------	nt
C. indolis	Y18184	----------GT--	----------	------G-G-------TC-	nt
C. jejuense	AY494606	----------GT--	----------	------G-G----------	nt
C. lentocellum	X71851	----------	----------	------TAAA-------T-	nt
C. nexile	X73443	----------GC--	----------	------G-G----------	nt
C. oroticum	M59109	----------	----------	---NNNNNNG---------	nt
C. populeti	X71853	----------GT--	----------	------G-G----------	nt
C. proteoclasticum	U37378	----------GT--	----------	------G-G----------	nt
C. saccharolyticum	Y18185	----------GT--	----------	------G-G-------TC-	nt
C. scindens	AF262238	----------	----------	------G-G----------	nt
C. sphenoides	X73449	----------GT--	----------	------G-G-------TC-	nt
C. symbiosum	M59112	----------GT--	----------	------G-G----------	nt
C. xylanolyticum	X71855	----------GT--	----------	------G-G-------TC-	nt
C. xylanovorans	AF116920	----------GT--	----------	------G-G----------	nt
Eubacterium					
E.hallii	L34621	A----------GT--	----------	------G----------	nt
E.ruminantium	AB008552	----------GT--	----------	------G--A---------	nt
E. cellulosolvens	X71860	----------GT--	----------	------A-GT------T--	nt
E. contortum	L34615	----------	----------	------G-GT---------	nt
E. eligens	L34420	----------GT--	----------	------N-AT------T--	nt
E. ramulus	L34623	----------GC--	----------	---N--G-G-------A--	nt
E. rectale	L34627	----------GC--	----------	------G-G----------	positive
E. xylanophilum	L34628	----------GT--	----------	------GN-T---------	nt
Ruminococcus					
R.gnavus	L76597	----------GC--	----------	N-N----------	positive
R.hansenii	M59114	----------	----------	----------	positive
R.hydrogenotrophicus	X95624	----------	----------	------TA--------A--	nt
R.lactaris	L76602	----------GC--	----------	------G-G----------	nt
R.luti	AJ133124	----------	----------	----------	nt
R.obeum	L76601	----------	----------	------C---------A--	nt
R.schinkii	X94965	----------	----------	------T---------A--	nt
R.torques	D14137	----------GC--	----------	------G-GT---------	nt
Others					
Coprococcus catus	AB038359	----------	----------	------G--NN-----T--	nt
Coprococcus eutactus	D14148	G—C---------GT--	----------	------G--A------T--	nt
Desulfotomaculum guttoideum	Y11568	----------GT--	----------	------G-G-------TC-	nt
Dorea formicigenerans	L34619	----------	----------	-----AT---------T--	nt
Dorea longicatena	AJ132842	----------	----------	-----GG-CT---------	nt
Hespellia porcina	AF445239	----------	----------	------G-G----------	nt
Hespellia stercorisuis	AF445264	----------	----------	------G-G----------	nt
Lachnobacterium bovis	AF298663	----------AC--	----------	------G-G----------	nt
Lachnospira pectinoschiza	L14675	----------GT--	----------	------G-G----------	nt
Pseudobutyrivibrio ruminis	X95893	----------GC--	----------	------G-G----------	nt
Roseburia cecicola	L14676	----------GC--	----------	------G-N----------	nt
Roseburia intestinalis	AJ312385	----------GC--	----------	------G-G----------	nt
Syntrophococcus sucromutans	Y18191	----------GT--	----------	------G-G----------	nt
Acetitomaculum ruminis	M59083	A-T---------AG--	----------	G-----N-GA---------	nt
Anaerostipes caccae	AJ270487	----------GT--	----------	------G-G-------T--	nt
Catonella morbi	X87151	----------GT--	----------	------CTG---C---TCT	nt

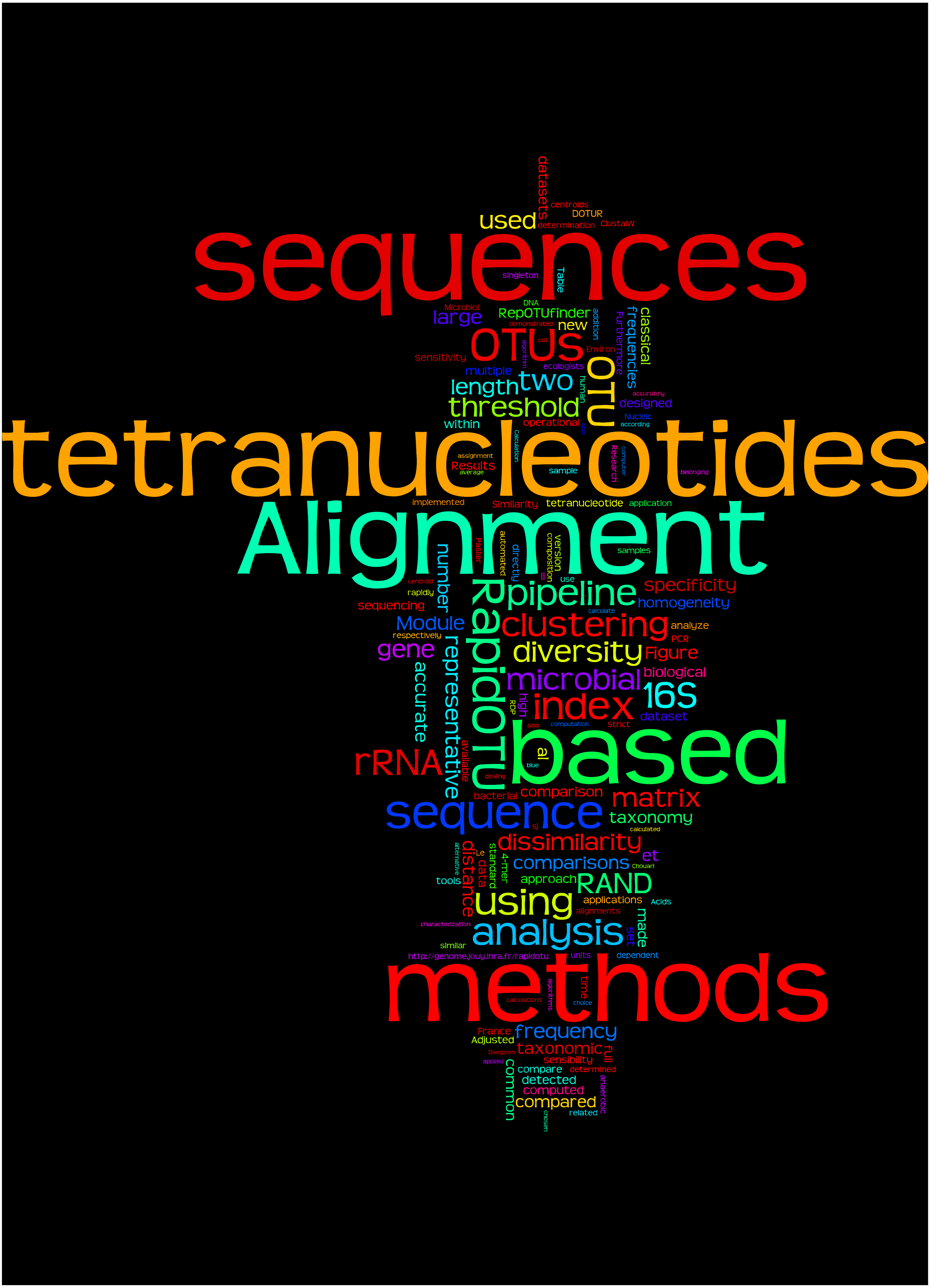
Bacteroides / *Prevotella* group

Species	EMBL Access Number	F_Bacter 11 primer 5′ 3′ CCTWCGATGGATAGGGGTT	P_Bact303 probe 5′ 3′ AAGGTCCCCCACATTG	R_Bacter 08 primer 5′ 3′ CACGCTACTTGGCTGGTTCAG	PCR assay
Bacteroides					
B. acidifaciens	AB021164	A-A----------	----------	----------	positive
B. caccae	X83951	A----------	----------	----------	positive
B. coprocola	AB200224	----------	----------	----------	nt
B. dorei	AB242142	T-A----------	----------	----------	nt
B. eggerthii	L16485	T-A----------	----------	----------	positive
B. finegoldii	AB222699	----------	----------	----------	nt
B. fragilis	X83935	----------	----------	----------	positive
B. intestinalis	AB214328	A-A----------	----------	----------	nt
B. massiliensis	AY126616	T----------	----------	----------	nt
B. nordii	AY608697	----------	----------	----------	nt
B. ovatus	X83952	A----------	----------	----------	positive
B. plebeius	AB200217	----------	----------	----------	nt
B. salyersiae	AY608696	----------	----------	----------	nt
B. splanchnicus	L16496	A-G-T---A-G--------	----------	-----A----------	nt
B. stercoris	X83953	----------	----------	----------	nt
B. thetaiotaomicron	L16489	----------	----------	----------	positive
B. uniformis	L16486	---N--------N------	----------	-----N----------	positive
B. vulgatus	M58762	T----------	----------	----------	positive
Parab. distasonis	M86695	--G----------	----------	-----A----------	nt
Prevotella					
Prevotella albensis	AJ011683	--G-----CAG--------	----------	----------	positive
P. bivia	L16475	G-A-----CAG--------	----------	----------	nt
P. bryantii	AJ006457	--G-----CAG--------	----------	----------	nt
P. buccae ***(1)***	L16478	--------C-G--------	----C--------C--	----------	positive
P. buccalis	L16476	G-G-----CAG--------	----------	----------	nt
P. denticola	L16467	G-G-----CAG--------	----------	--N----------	nt
P. disiens	L16483	G-A-----CAG--------	----------	----------	nt
P. enoeca	AJ005635	G-A-----CAG--------	----------	----------	nt
P. heparinolytica	L16487	----------	----------	----------	nt
P. intermedia	X73965	G-------CAG--------	----------	----------	nt
P. melaninogenica	L16469	G-A-----CAG--------	----------	----------	nt
P. multiformis	AB182483	G-------CAG--------	----------	----------	nt
P. nigrescens	L16471	G-G-----CAG--------	----------	----------	nt
P. oralis ***(1)***	L16480	G-A-----C-G--------	----------C--	----------	positive
P. oris	L16474	G-GN----CAG--------	----------	----------	nt
P. oulorum	L16472	G-A-----CAG--------	----------	----------	nt
P. pallens	Y13105	G-N-----CAG--------	----------	----------	nt
P. salivae	AB108826	G-A-----CAG--------	----------	----------	nt
P. tannerae	AJ005634	----------G--------	----------	----------	nt
P. veroralis	L16473	G-G-----CAG--------	----------	----------	nt
P. zoogleoformans	L16488	----------	----------	----------	nt

Lactobacillus / Leuconostoc / Pediococcus group

Species	EMBL Access number	F_lacto05 primer 5′ 3′ AGCAGTAGGGAATCTTCCA	R_LactoO4 Primer 5′ 3′ CGCCACTGGTGTTCYTCCATATA	PCR assay
Lactobacillus				
L. acetotolerans	M58801	----------	----------	nt
L. acidifarinae	AJ632158	----------	----------	nt
L. acidipiscis	AB023836	----------	----------	nt
L. acidophilus	m58802	----N-----	----------	positive
L. agilis	M58803	----------	----------	nt
L. algidus	AB033209	----------	----------A-------	nt
L. alimentarius	M58804	----------	----------	nt
L. amylophilus	M58806	----------	----------	nt
L. amylovorus	m58805	----------	----------	nt
L. animalis	M58807	----------N-------N	----------	nt
L. antri	AY253659	----------	----------	nt
L. aviarius	M58808	----------	----------	nt
L. bifermentans	M58809	----------	----------	nt
L. brevis	ab070611	----------	----------	nt
L. buchneri	M58811	----------	----------	nt
L. casei	D16548	----------	----------	positive
L. coleohominis	AM113776	----------	----------	nt
L. collinoides	AB005893	----------	----------	nt
L. concavus	AY683322	----------	----------	nt
L. coryniformis	AJ575741	----------	----------	nt
L. crispatus	y17362	----------	----------	positive
L. curvatus	AJ270951	----------	----------	nt
L. delbrueckii	x52654	----------	----------	positive
L. durianis	AJ315640	----------	----------	nt
L. equi	AB048833	----------	----------	nt
L. farciminis	M58817	----------	----------	nt
L. fermentum	af522394	----------	----------	positive
L. fornicalis	Y18654	----------	----------	nt
L. fructivorans	m58818	----------	----------	nt
L. frumenti	AJ250074	----------	----------	nt
L. fuchuensis	AB063479	----------	----------	nt
L. gallinarum	AJ242968	-A----------C	----------	nt
L. gasseri	m58820	----------	----------	positive
L. gastricus	AY253658	----------	----------	nt
L. graminis	AM113778	----------	----------	nt
L. hammesii	AJ632219	----------	----------	nt
L. harbinensis	AB196123	----------	----------	nt
L. helveticus	ay369116	----------	----------	positive
L. hilgardii	M58821	----------	----------	nf
L. homohiochii	AM113780	----------	----------A-------	nf
L. ingluviei	AF317702	----------	----------	nf
L. intestinalis	aj306299	----------	----------	nf
L. jensenii	AF243176	----------	----------	nf
L. johnsonii	aj002515	----------	----------	positive
L. kalixensis	AY253657	----------	----------	nt
L. keferi	AJ621553	----------	----------	nt
L. kefiranofaciens	AM113781	----------	----------A-------	nt
L. kimchii	AF183558	----------	----------	nt
L. kitasatonis	AB107638	----------	----------	nt
L. kunkeei	Y11374	----------	----------	nt
L. lindneri	X95421	----------	----------C-T------	nt
L. malefermentans	AM113783	----------	----------	nt
L. mali	M58824	----------	----------	nt
L. manihotivorans	AF000162	----------	----------	nt
L. mésenteroides	m23035	T----------	----------	nt
L. mindensis	AJ313530	----------	----------	nt
L. mucosae	AF126738	----------	----------	positive
L. murinus	M58826	----------	----------	nt
L. oligofermentans	AY733084	----------	----------	nt
L. oris	X94229	----------	----------	nt
L. panis	X94230	----------	----------	nt
L. pantheris	AF413523	----------	----------	nt
L. parabrevis	AM158249	----------	----------	nt
L. parabuchneri	AY026751	----------	----------	nt
L. paracasei	d79212	----------	----------	positive
L. paracollinoides	AJ786665	----------	----------	nt
L. parakefiri	AY026750	----------	----------	nt
L. paralimentarius	AJ417500	----------	----------	nf
L. paraplantarum	AJ306297	----------	----------	nt
L. pentosus	D79211	----------	----------	nt

L. perolens	Y19167	----------	----------	nt
L. plantarum	D79210	----------	----------	positive
L. pontis	AJ422032	----------	----------	nt
L. pseudomesenteoides	ab023237	----------	----------	nt
L. rennini	AJ576007	----------	----------	nt
L. reuteri	l23507	----------	----------	nt
L. rhamnosus	m58815	----------	----------	positive
L. rossii	AJ564009	----------	----------	nt
L. ruminis	m58828	----------	----------	nt
L. saerimneri	AY255802	----------	----------	nt
L. sakei	ay204897	----------	----------	nt
L. salivarius	af089108	----------	----------	positive
L. sanfranciscensis	X76327	----------	----------C-T------	nt
L. satsumensis	AB154519	----------	----------	nt
L. sharpeae	m58831	----------	----------	nt
L. siligionis	DQ168027	----------	----------	nt
L. sobrius	AY700063	----------	----------	nt
L. spicheri	AJ534844	----------	----------	nt
L. suebicus	AJ306403	----------	----------	nt
L. vaccinostercus	AM113786	----------	----------	nt
L. vaginalis	x61136	----------	----------N-	nt
L. versmoldensis	AJ496791	----------	----------	nt
L. vini	AJ576009	----------	----------	nt
L. zeae	D86516	----------	----------	nt
Leuconostoc				
Leuco mesenteroides	m23035	----------	----------	positive
Leuco pseudomesenteroides	ab023237	T----------	----------	positive
Pediococcus				
P.inopinatus	AJ271383	----------	----------	nt
P.parvulus	D88528	----------	----------	nt
P.cellicola	AY956788	----------	----------	nt
P.acidilactici	M58833	----------	----------	nt
P.pentosaceus	M58834	----------	----------	nt
P.claussenii	AJ621555	----------	----------	nt
P.stilesii	AJ973157	----------	----------	nt
P.dextrinicus	D87679	----------	----------	nt
Enterococcus				
E.aquimarinus	AJ877015	----------GG	----------	nt
E.asini	Y11621	----------GG	----------	nt
E.avium	AF133535	----------GC	----------	nt
E.caccae	AY943820	----------GC	----------	nt
E.canintestini	AJ888906	----------GC	----------	nt
E.casseliflavus	Y18161	----------GC	----------	nt
E.cecorum	AF061009	----------GC	----------	nt
E.devriesei	AJ891167	----------GC	----------	nt
E.dispar	AF061007	----------GC	----------	nt
E.durans	AJ420801	----------GC	----------	nt
E.faecalis	af515223	----------GC	----------C------	negative
E.faecium	AJ276355	----------GC	----------	negative
E.gilvus	AY033814	----------GC	----------	nt
E.hermanniensis	AY396047	----------GC	----------	nt
E.hirae	AJ276356	----------GC	----------	nt
E.italicus	AJ582753	----------GC	----------	nt
E.malodoratus	AF061012	----------GC	----------	nt
E.moraviensis	AF286831	----------GC	----------	nt
E.mundtii	AF061013	----------GC	----------	nt
E.pallens	AY033815	----------GC	----------	nt
E.phoeniculicola	AY028437	----------GC	----------	nt
E.pseudoavium	AF061002	----------GC	----------	nt
E.raffinosus	Y18296	----------GC	----------	nt
E.ratti	AF326472	----------GC	----------	nt
E.silesiacus	AM039966	----------GC	----------	nt
E.sulfureus	X55133	----------GC	----------	nt
E.termitis	AM039968	----------GC	----------	nt
E.villorum	AF335596	----------GC	----------	nt
Others				
Lactoccocus lactis	M58836	----------GC	------C-------C--------	nt
Streptococcus thermophilus	x68418	----------GC	------C---------C------	negative



sequences
tetranucleotides
Alignment
based
methods
RapidOTU
OTUs
OTU
threshold
two
length
pipeline
clustering
diversity
microbial
index
16S
rRNA
sequence
matrix
dissimilarity
using
analysis
RAND
representative
accurate
number
gene
Module
frequency
taxonomic
compared
common
comparisons
distance
taxonomy
comparison
specificity
homogeneity
Figure
biological
dataset
large
used
datasets
RepOTUfinder
new
classical
frequencies
Furthermore
designed
operational
within
multiple
sensitivity
Results
similarity
tetranucleotide
implemented
application
samples
sequencing
analyze
respectively
approach
applications
made
set
tools
http://genome.jouy.inra.fr/rapidotu
units
dependent
time
France
Adjusted
sensibility
compare
determined
detected
computed
related
anaerobic
full
et
al
bacterial
available
high
RDP
Strict
computation
4-mer
standard
data
alternative
similar
characterization
alignments
Acids
Chouari
calculated
sj
dealing
blue
also
PCR
calculate
use
ii
centroid
rapidly
composition
version
directly
automated
Pasilier
sample
average
assignment
Calculation
Research
computer
belonging
Nucleic
according
human
accurately
ecologists
Environ
cut
algorithm
demonstrated
Microbiol
DNA
addition
singleton
Table
centroids
DOTUR
ClustalW
determination
calculations
algorithms
choice
Daegelen
applied
chosen

RapidOTU: 16S rRNA gene sequences clustering into operational taxonomic units using tetranucleotides frequencies

J. Tap[1¤], L. Legrand[2¤], C. Gauthey[2], C. Caron[2], J. Doré[1], D. Le Paslier [3,4], E. Pelletier[3,4] and M. Leclerc[*1]

[1]INRA, UEPSD, UR910, Domaine de Vilvert, 78350 Jouy-en-Josas, France.

[2]INRA, MIG, UR1077, Domaine de Vilvert, 78350 Jouy-en-Josas, France.

[3]CEA-Genoscope, 91057 Evry cedex, France.

[4]CNRS, UMR 8030, 91057 Evry cedex, France

[¤] These authors equally contributed to the work

ABSTRACT (300 words max)

Background

Recent advances in high-throughput sequencing have made it possible to produce very large datasets, requiring powerful automated and accurate computer tools for analysis. For ecologists dealing with 16S rRNA gene sequences obtained by metagenomic or PCR approach, the standard methods based on alignments of large 16S rRNA sequence libraries prevents a correct assessment of Operational Taxonomic Units (OTU) or phylotypes. Alignment free approaches have been developed but never connected directly with phylotypes clustering. No integrated tools are easy to use for biologists wishing to get accurate and automatic analysis of their datasets by alignment or tetranucleotide frequency directly connected to OTU clustering.

Results

We developed a new pipeline, RapidOTU, which connects existing and new applications to calculate microbial diversity from large data sets. Tetranucleotides frequency method connected to OTUs clustering could be used to rapidly and accurately analyze 16S rRNA gene sequences, from Sanger or 454 pyrosequencing. In addition, a newly designed method, RepOTUfinder, automatically calculated and extracted a representative sequence for each OTU. We also showed, using 289,052 sequences from RDPII that connecting tetranucleotides frequencies directly to a clustering algorithm gave similar results and accelerated the analysis

with excellent specificity and sensibility compared to alignment based methods. Finally, tetranucleotide based method was validated using a case study from a highly diverse biological ecosystem.

Conclusions

Through a user friendly web interface, RapidOTU provides biologists with a flexible pipeline for fast and accurate estimation of diversity from large sequence datasets with tetranucleotides based method as an alternative of alignment dependent approach. In addition, RepOTUfinder, a new method, included in RapidOTU, calculates the representative sequence for each OTU. RapidOTU pipeline outputs were compatible with all downstream connections, other assignments or/and phylogenetic tools. The pipeline is flexible enough to allow the implementation of future analysis tools. RapidOTU is available at http://genome.jouy.inra.fr/rapidotu.

Contact

marion.leclerc@jouy.inra.fr

AUTHORS SUMMARY

The RapidOTU pipeline was implemented to facilitate the analysis of multiple sequences files by ecologists dealing with metagenomic data and 16S rRNA sequences. RapidOTU connects existing applications to calculate microbial diversity. Alignment based method and, for the first time, alignment free tetranucleotides composition can be used to rapidly analyze sequence data. In addition, a newly designed method, RepOTUfinder automatically calculates and extracts a representative sequence for each Operational Taxonomic Unit.

Tetranucleotides based method implemented in the RapidOTU pipeline did speed up by more than 30 fold the analysis, compared to the classical alignment based methods using ClustalW MPI version. In term of taxonomy homogeneity, specificity, sensibility and clustering, tetranucleotides based methods gave similar results than classical alignment based methods. RapidOTU is freely available at http://genome.jouy.inra.fr/rapidotu.

INTRODUCTION

Understanding microbial processes relies on the accurate determination of microbial species to measure and compare microbial diversity. 16S rRNA gene sequencing and analysis has been recognized as a powerful method to determine microbial diversity [1]. Several programs are currently available to analyze bacterial sequence datasets [2,3]. However, recent advances in high-throughput sequencing have made it possible to produce very large datasets, requiring powerful automated and accurate computer tools for analysis. Furthermore, new sequencing technologies, 454 pyrosequencing, now being widely used in microbial ecology, provide shorter length sequences but larger datasets.

Calculation time required for precise alignments of large 16S rRNA sequence libraries prevents a correct assessment of Operational Taxonomic Units (OTUs). In addition, the sequences representative of OTU are often subjectively chosen. A rational determination is of importance to compare datasets from different studies and avoid a biased representation of diversity. A new method, RepOTUfinder, was designed to accurately compute and retrieve representative sequences from distance matrix algorithms.

Finally, apart from alignment of sequences, the use of tetranucleotides frequencies has been reported as an accurate tool for clustering sequences based on their taxonomy [4,5,6]. Still, this alignment-free approach has never been integrated into a fully automated web based analysis pipeline, available to the entire microbiologist community. Tetranucleotides approach was only used to override comparison between genetically distant sequences [7]. Furthermore, tetranucleotides method has never been accurately compared with alignment dependent methods for OTUs calculations.

We designed and implemented a pipeline named RapidOTU that combines existing applications and newly designed ones, to rapidly determine microbial diversity using alignment or tetranucleotides frequencies-count methods. The aims of RapidOTU development were to:

(i) Compare alignment dependent methods with tetranucleotides methods and challenge them on a biological application.

(ii) Provide biologists with a web-based flexible analysis pipeline for molecular inventories dealing with large number of 16S rRNA sequences with alignment dependent and alignment independent methods.

(iii) Offer a new simple method to choose a representative sequence for each OTU.

Calculation times were drastically decreased by parallelizing computation and optimizing algorithms, in order to analyze several thousands of sequences within hours.

The pipeline was validated on RDP II sequences with full length, and has already been used to characterize the bacterial diversity from human gastro intestinal tract [8] and from pigs fecal samples [9].

METHODS

Implementation

The RapidOTU pipeline was designed with three modules (Figure 1).

Tetranucleotide Module

This first module based on tetranucleotides frequency method offered a different analysis approach compared to the standard alignment methods: The 4-mer composition of each sequence is counted using OCOUNT software [6], and, accounting for sequence length, normalized into a frequency dataset. The pairwise comparison of 4-mer composition is computed using Pearson's correlation and modified to obtain a full dissimilarity matrix corresponding for instance to a "dnadist-like matrix" generated from the alignment method.

To compare to the tetranucleotides based method, a classical alignment based module was created to assess validity of the tetranucleotides based method. Sequences are aligned by default using the full dynamic programming algorithm of ClustalW [3] or by the MAFFT software [10]. Then, a pairwise distance matrix is built using the Phylip dnadist [11] application.

Clustering module

The clustering module of the pipeline processes distance matrix data originating from the two methods described above. This module is used for the detection of OTUs and for calculating diversity indices. The DOTUR [12] threshold for 4-mer frequency method was set to 11% in order to match the 2% dissimilarity OTU determination from the alignment based method (Figure 2). However, DOTUR threshold for both methods can easily be set up by users.

RepOTUfinder

In order to normalize the choice of OTUs representative sequence, we implemented a method based on the centroid concept.

For the OTU j with s_j sequences, a square matrix M_j of dimension s_j x s_j is built. The element M_j (i, k) represents the distance between the sequence i and the sequence k.

Calling d the vector such that the element $d(i)$ was the sum of the elements belonging to the ith row of the matrix M_j given by:

$$d(i) = \sum_{k=1}^{s_j} M_j(i,k)$$

The representative sequence rs_j corresponds to the sequence of index l ($l \in [1, s_j]$), such that d(l) = *min*(d). The estimation of these centroid sequences is wrapped in a Perl application which can be run standalone.

This fully modular architecture was chosen to allow the future integration of new methods or algorithms plugging.

Comparison between alignment and tetranucleotides based methods

OTUs cut off clustering

289,052 sequences belonging to the Bacteria domain were downloaded from the RDP II (RDP Release 10, Update 11) on May 2009. Hundred samples consisting of 1,000 randomly chosen sequences were analyzed with RapidOTU using the two methods with OTU dissimilarity thresholds ranging from 0 to 20 % with 1 % increment.

Clustering Similarity by RAND index

From the OTU clustering , we determined *(i)* the number of pairs of related sequences in both clustering, *(ii)* the number of pairs of sequences not related, *(iii)* the number of related sequences in tetranucleotides module but not in alignment module, *(iv)* the number of related sequences in alignment module but not in tetranucleotides module. These data were used to calculate the Rand index R [13] which compares the similarity between two methods. Furthermore, the Adjusted Rand index *AR* [14] which is the Adjusted-for-chance form of the Rand index was also calculated.

To evaluate the robustness of this index, within comparisons (n= 4,950) were made by tetranucleotides and alignment based methods. For each sample of 1,000 sequences, comparisons were made between tetranucleotides and the multiple alignment methods (n = 100) by the Rand and the Adjusted Rand indexes.

Sensibility and specificity of tetranucleotides frequency method

As alignments based methods are commonly used by microbial ecologists, it was defined as the reference method (Figure 1) to which we compared the tetranucleotides based method. Using counts from *i*, *ii*, *iii* and *iv* defined above, the specificity (Sp) and sensitivity (Se) of

tetranucleotides based method regarding the classical alignment based method were calculated.

A sensitivity of 100% indicated that tetranucleotides method grouped two sequences in the same OTU as did the alignment method. A specificity of 100% indicated that the tetranucleotides method dissociated two sequences that were not in the same OTU as did the alignment method.

Biological case study: Microbial Diversity of laboratory scale bioreactors

The goal of the project was to link the effect of a thermal stress on the microbial diversity of anaerobic bioreactors. Samples originated from laboratory scale anaerobic digesters collected between March and July 2008 (grant ANR-DIGUE). Genomic DNA was extracted as previously described [15]. 16S rRNA gene was amplified by PCR (25 cycles) using Bacteria domain primers (8F 5'-TGAGCCAGGATCAAACTCT-3' and 1390R 5'-GACGGGCGGTGTGTACAA-3'). Triplicate PCR reactions were pooled, PCR products were ligated into a pGEM®-T vector which was then inserted into competent *E. coli* DH10B™ by electroporation. The nucleotide sequence of plasmid inserts was determined by classical automated Sanger sequencing. The 16S rRNA gene sequences for each clone were assembled by Phrap (www.phrap.org). Only good quality 16S rRNA gene sequences (longer than 1200 bp and with Phred qualities of above 15 for each base) were selected for further analysis.

RESULTS & DISCUSSION

Tetranucleotides based method gives results highly similar to classical alignment based method

Each boxplot (Figure 2) represents the variations of OTU numbers detected in 100 subsets of 1,000 randomly sampled sequences from the 289,052 RDP II sequences. Each sequences sample was processed by RapidOTU using the two methods. A 2% dissimilarity threshold for alignment method gave no significant difference compared to an 11% threshold for the tetranucleotides method (p.value > 0.8). At these thresholds the standard deviations are similar between the two methods. Noticeably, the standard deviation increased according to the dissimilarity threshold with the tetranucleotides approach, while it decreased with the alignment method. The variation of sensitivity between the 2% dissimilarity threshold for alignment method and the 11% dissimilarity threshold tetranucleotides based methods is shown on Figure 2. For further steps of the analysis, comparisons between alignment and tetranucleotides based methods were made with 2% and 11% dissimilarity threshold respectively.

Whatever the index used, “between methods” tests showed a higher similarity clustering index than “within methods” tests (Figure 3). Due to the sampling of a large database, singleton OTUs were detected as highly represented in each 1,000 sequences sample. This phenomenon had a strong impact on Adjusted Rand index and may explain the gap observed between the two indices.

Compared to the alignment method considered as the reference procedure, tetranucleotides frequency method demonstrated a high sensibility (Se = 73.9 % ± 4.8) and a high specificity (Sp = 99.981 % ± 6 x 10^{-03}). Nevertheless, the high specificity was also affected by singleton OTUs present in these randomly sampled sets of sequences. Therefore, a biological dataset, leading to less singletons, was analyzed to evaluate the pipeline on a more realistic ecologic dataset.

Tetranucleotides based method fastens analysis compared to classical alignment based method

The tetranucleotides frequencies determination of more than 5,000 sequences on one processor (8Gb RAM) was four times faster than a ClustalW MPI version deployed on 16

processors Intel QuadCore 2.33GHz (Table S1). Furthermore, 50,000 sequences could be computed in 133 hours with tetranucleotides based method deployed on 64 processors.

Since the size of distance matrix quadratically increases, it was important to make a parallelisation. The distance matrix with an algorithm of $O(n^2)$ was reduced to sub matrices generated for simultaneous computation. Furthermore, clustering calculations time rapidly increased according the number of sequences. This is linked to the problem of RAM usage by DOTUR and RepOTUfinder.

Biological validation

A set of 10,295 16S rRNA bacterial gene sequences was computed by the RapidOTU pipeline. A total of 1,526 OTUs were detected with alignment methods at 2% dissimilarity threshold and 1,382 OTUs with tetranucleotides at 11% dissimilarity threshold. 818 and 695 singletons sequences were detected with alignment and tetranucleotides base methods, respectively (Table S2). 317 OTUs constituted by at least two sequences were found strictly identical between the two methods and 682 singletons sequences were common between the two methods. 833 centroids sequences were identical between the two methods.

Very high Rand index (R = 0.999) and Ajusted Rand index (AR = 0.930) were observed. As expected, OTU singleton scarcity in this type of assay reduced the gap between the Rand index and Adjusted Rand index. Meanwhile, specificity and sensitivity of the tetranucleotides method remained above 99%. This suggests that for taxonomic assignment, the tetranucleotides frequency method is as accurate as alignment method, as confirmed by the homogeneity of blast based taxonomic assignment of OTU members (Table 1).

CONCLUSION

Owing to an optimization of memory usage and parallel computation, the pipeline RapidOTU allowed the characterization of microbial diversity on large sequences dataset with an accurate choice of representative sequences. The use of tetranucleotides frequencies gave the opportunity to easily and efficiently analyze sets of thousands of sequences.

Comparison between alignment dependent and tetranucleotides approaches based on Rand index demonstrated that clustering similarity was very high. Furthermore, tetranucleotides method high specificity and sensibility set it as a fast alternative method to the reference methods. Tetranucleotides method could not only be used to override the comparison of genetically distant sequences but directly to compare and cluster sequences into OTUs. Tetranucleotides based method have indeed been demonstrated as successful for the analysis of biological cases.

Furthermore, if sequencing technologies such as the "GS FLX Titanium Series" produce sequences with an average size of 500 bp, they will soon reach 1000 bp. This will open the possibility to analyze the almost entire 16S rRNA genes and very accurately perform microbial diversity analysis. Using 454 pyrosequencing, metagenomic datasets have indeed rapidly been generated within the last years. Numerous projects, including clinical studies, include or rely on deep sequencing of 16S rDNA genes to asses the microbial diversity of an ecosystem or to investigate the link between micro-organisms and clinical or environmental parameters. Finally, the tetranucleotide approach is not restricted to rRNA genes but could be applied to other genes highly conserved or of interest when one study a microbial community.

RapidOTU is a fast accurate and convenient web-based tool for studying microbial communities and provides ecologists with diversity characterization and a rational choice of representative sequences. The output files can directly be used for taxonomic characterization or phylogeny analysis.

RapidOTU source files and web interface are available at http://genome.jouy.inra.fr/rapidotu.

AVAILABILITY AND SYSTEM REQUIREMENTS

Text Project name: RapidOTU

Project home page: http://genome.jouy.inra.fr/rapidotu

Operating system(s): Any

Programming language: Perl

License: CeCILL GNU GPL (http://genome.jouy.inra.fr/rapidotu/html/Licence_CeCILL_V2-en.html)

Any restrictions to use by non-academics: None.

FUNDING

J. Tap is supported by a PhD fellowship from the ANR French National Agency for Research, ANR/DEDD/PNRA/PROJ/200206-01-01, within the AlimIntest program.

AKNOWLEDGEMENT

We are grateful to Patricia Lepage (INRA UEPSD, France) for helpful comments on our work and on the manuscript.

REFERENCES

1. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, et al. (2005) Diversity of the Human Intestinal Microbial Flora. Science 308: 1635-1638.
2. DeSantis TZ, Jr., Hugenholtz P, Keller K, Brodie EL, Larsen N, et al. (2006) NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res 34: W394-399.
3. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22: 4673-4680.
4. Rudi K, Zimonja M, Kvenshagen B, Rugtveit J, Midtvedt T, et al. (2007) Alignment-independent comparisons of human gastrointestinal tract microbial communities in a multidimensional 16S rRNA gene evolutionary space. Applied and Environmental Microbiology 73: 2727-2734.
5. Teeling H, Meyerdierks A, Bauer M, Amann R, Glockner FO (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. pp. 938-947.
6. Teeling H, Waldmann J, Lombardot T, Bauer M, Glockner FO (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics 5: 163.
7. Sun Y, Cai Y, Liu L, Yu F, Farrell ML, et al. (2009) ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. Nucleic Acids Research 37: e76.

8. Tap J, Mondot S, Levenez F, Pelletier E, Caron C, et al. (2009) Towards the human intestinal microbiota phylogenetic core. Environ Microbiol 11: 2574-2584.
9. Mieszkin S, Furet JP, Corthier G, Gourmelon M (2009) Estimation of pig fecal contamination in a river catchment by real-time PCR using two pig-specific Bacteroidales 16S rRNA genetic markers. Applied and Environmental Microbiology 75: 3045-3054.
10. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Research 33: 511-518.
11. Felsentein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166.
12. Schloss PD, Handelsman J (2005) Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. Appl Environ Microbiol 71: 1501-1506.
13. Rand WM (1971) Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66: 846 – 850.
14. Hubert L, Arabie P (1985) Comparing partitions. Journal of Classification 2: 193–218.
15. Chouari R, Le Paslier D, Daegelen P, Ginestet P, Weissenbach J, et al. (2003) Molecular evidence for novel planctomycete diversity in a municipal wastewater treatment plant. Appl Environ Microbiol 69: 7354-7363.

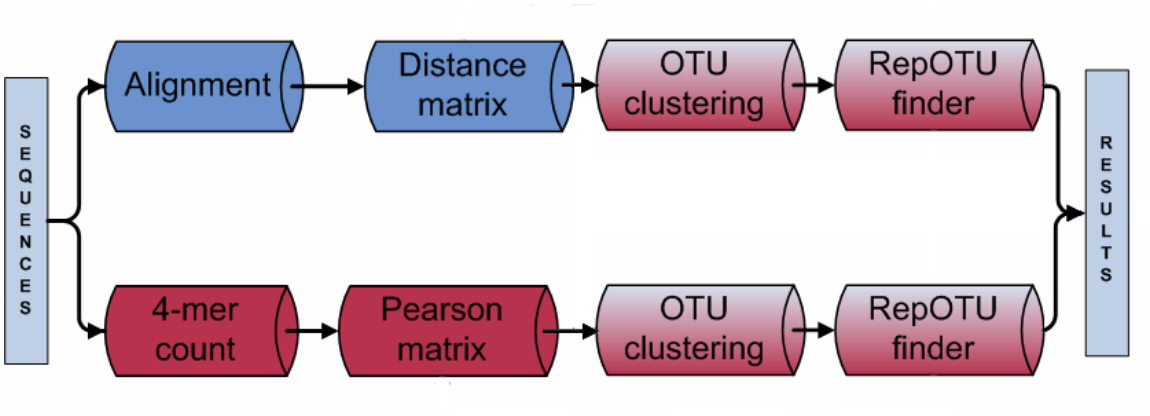


Figure 1: RapidOTU pipeline organization. The RapidOTU pipeline is based on two different analytical methods, with all applications linked through Perl components. The 4-mer composition of each sequence is counted using OCOUNT software and, accounting for sequence length, normalized into a frequency dataset. Pairwise comparisons are computed using Pearson's correlation and modified to obtain a distance matrix (in red). Alternatively (in blue), multiple alignments are performed with ClustalW-MPI or with MAFFT and distance matrix is computed by fdnadist. In both methods, OTUs are determined with DOTUR, based on a distance threshold of 2% and 11% for the alignment and the 4-mer frequency count method respectively. According to these thresholds, from computed matrix, representative sequences are determined by RepOTUfinder as the OTUs centroid.

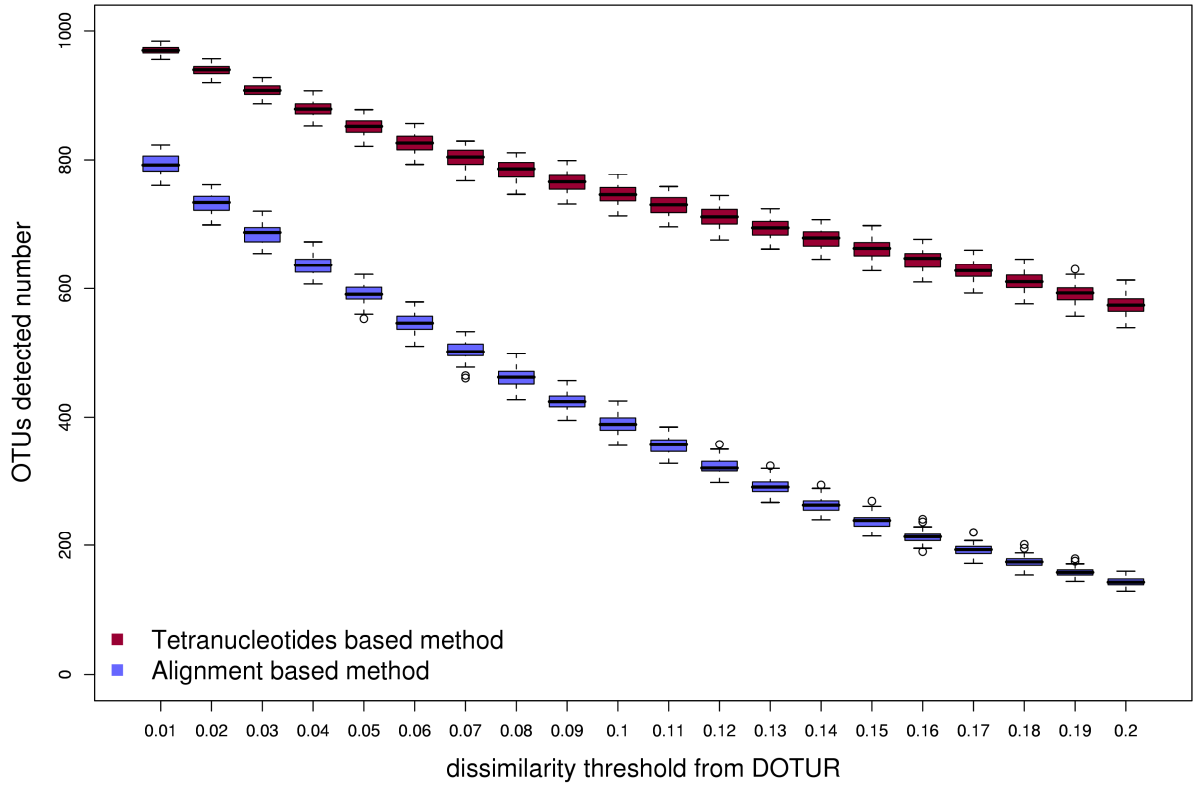


Figure 2: Bootstrap between tetranucleotides and alignment methods. Each boxplot represented variations in the number of OTU detected in 100 samples of 1,000 sequences sampled. Each sample of 1,000 sequences was analyzed by RapidOTU with the two methods (red: tetranucleotides method; blue: alignment method). X axis showed the dissimilarity threshold settled to group sequences in OTUs. There was no significant difference between a 0.002 dissimilarity threshold with the alignment method and a 0.11 dissimilarity threshold for the tetranucleotides frequencies method (p.value> 0.8).

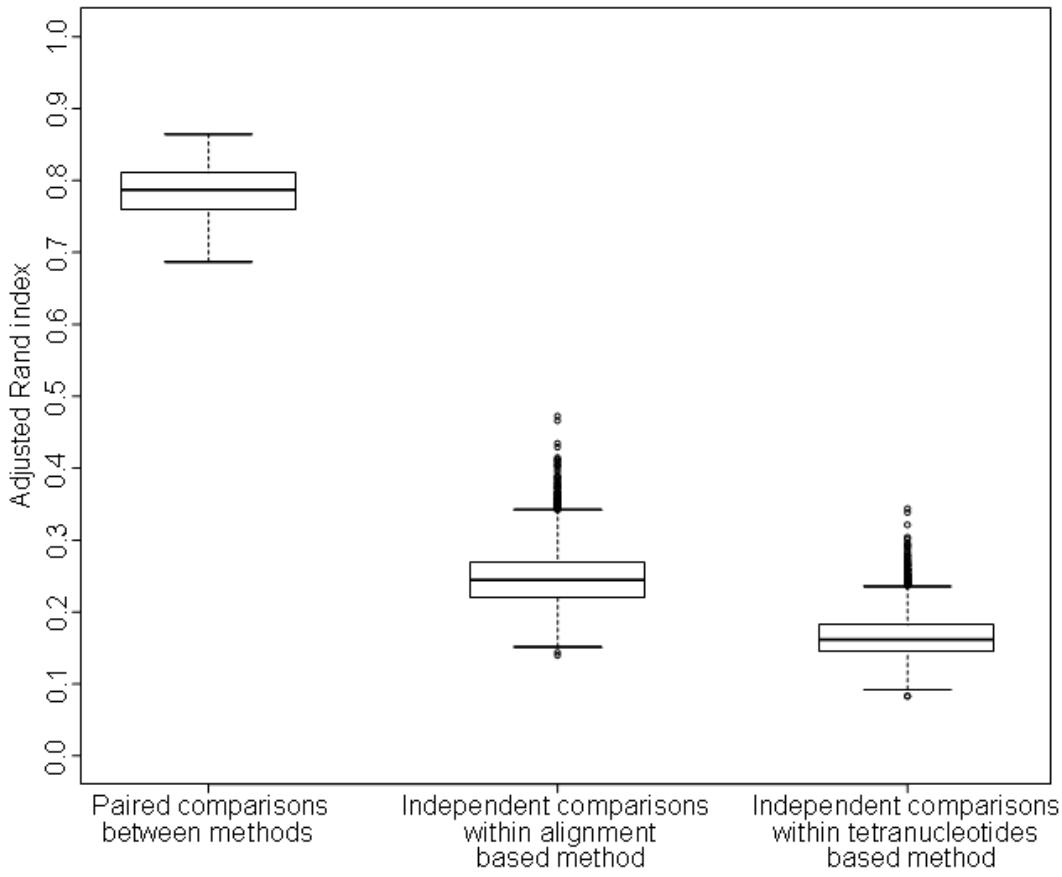


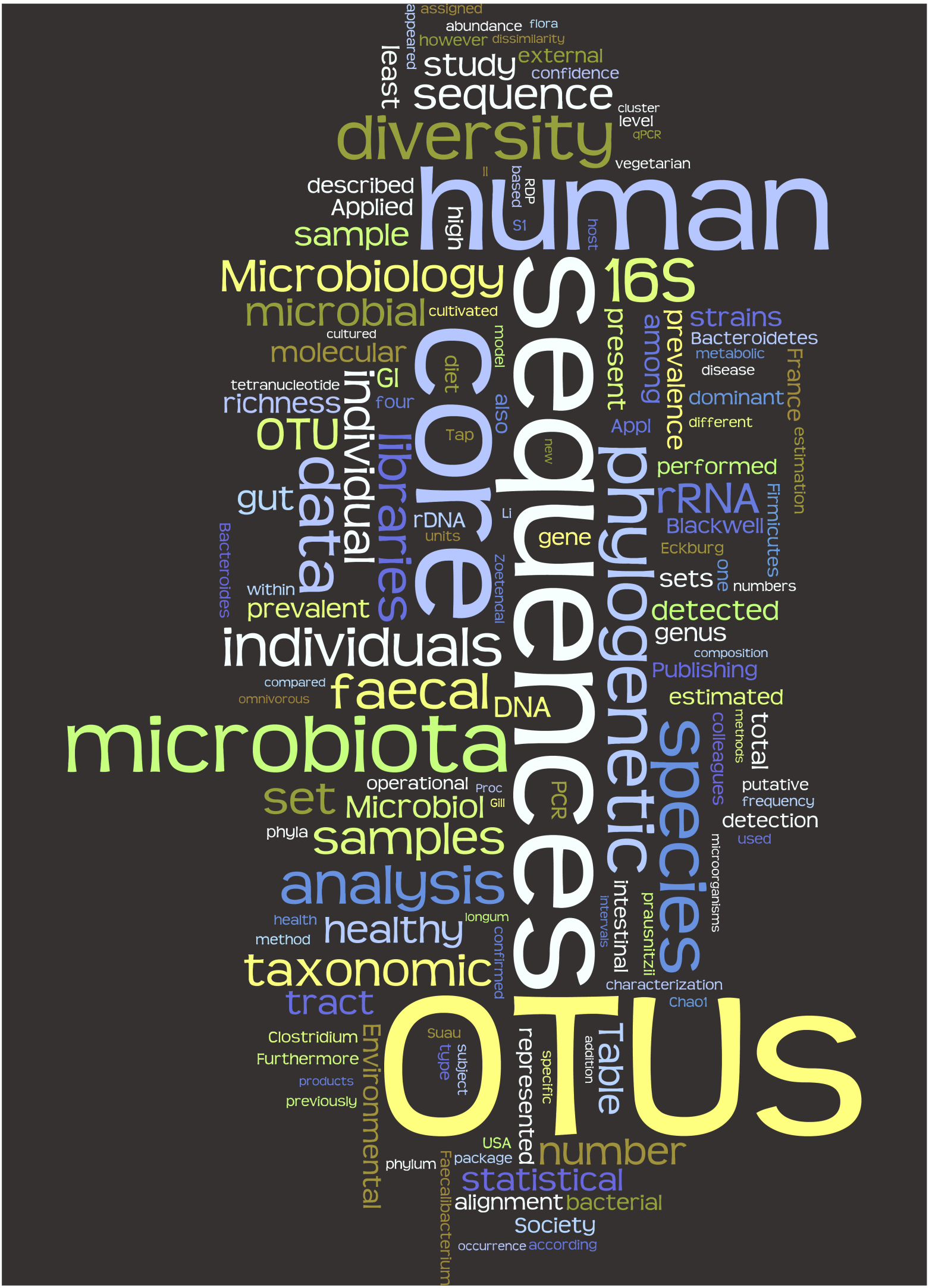
Figure 3: Similarity evaluations between tetranucleotides and alignment methods with Adjusted Rand index. 100 paired comparisons were made between tetranucleotides (cut off 11%) and alignment (cut off 2%) methods. 4,950 independents comparisons were made between the results from tetranucleotides and from alignment methods. Whatever the index used, paired comparisons between methods showed a higher Rand index than the "within methods" comparisons.

Table 1: Comparison of OTUs detected from the biological sequences dataset by alignment or tetranucleotides method. Strict taxonomy homogeneity was tested by Blastn on Greengenes "HT" taxonomy (Hugenholtz taxonomy). Strict taxonomy homogeneity was found when sequences of a given OTU belonged to exactly the same taxon. When sequences belonging to « Unclassified; otu_2389 » were removed, homogeneity of 98.8% and 98.9% was respectively found for tetranucleotides and alignment based methods.

Method	Alignment	vs	tetranucleotides
Total number OTUs	1,526		1,382
OTUs with strict "HT" taxonomy homogeneity	1,433 (93.9%)		1,269 (91.8%)
Common OTUs		999	
Common centroids		833	
Nb OTUs singletons	818		695
Common OTUs		682	
Common centroids		682	
Nb OTUs (n > 2 sequences)	708		687
OTUs with strict "HT" taxonomy homogeneity[*]	615 (86.9%)		574 (83.6%)
Common OTUs		317	
Common centroids		151	

Table S1: Benchmark of RapidOTU pipeline calculation times with short and full length sequences. 5,000 full length 16S rRNA gene sequences were randomly downloaded from RDP II database (average length 1472bp). Short length sequences dataset resulted from the extraction of the V6-V8 region of the 16S rRNA gene (average length 385bp). Alignments were computed with ClustalW (MPI version).

Dataset	short length sequences		full length sequences	
Method	Alignment	tetranucleotides	Alignment	tetranucleotides
Execution time 1 processor	24h 20min	08h 35min	undetermined	08h 34min
16 processors	04h 28min	01h 08min	35h 33min	01h 16min
OTUs detected	1,780	2,164	2,078	2,018
Singleton OTUs	986	1,348	1,218	1,207


assigned
abundance
flora
however
dissimilarity
appeared
least
study
external
confidence
sequence
cluster
level
qPCR
diversity
vegetarian
described
Applied
RDP
based
high
human
sample
S1
host
Microbiology
16S
microbial
cultivated
cultured
strains
Bacteroidetes
metabolic
disease
France
estimation
molecular
model
among
prevalence
present
core
sequences
phylogenetic
tetranucleotide
GI
diet
richness
four
also
dominant
different
Appl
OTU
Tap
new
performed
individual
libraries
data
rRNA
gut
rDNA
Li
Firmicutes
Blackwell
gene
units
Eckburg
Bacteroides
Zoetendal
sets
one
numbers
within
prevalent
detected
genus
individuals
composition
Publishing
compared
omnivorous
faecal
DNA
estimated
species
microbiota
colleagues
total
methods
operational
Proc
PCR
putative
set
Microbiol
Gill
frequency
detection
phyla
samples
used
microorganisms
analysis
intestinal
prausnitzii
intervals
health
longum
healthy
method
confirmed
taxonomic
characterization
tract
Chao1
OTUs
Clostridium
Suau
Table
Environmental
Furthermore
subject
type
represented
specific
addition
products
previously
USA
number
phylum
package
statistical
alignment
bacterial
Society
occurrence
according
Faecalibacterium

Environmental Microbiology (2009) doi:10.1111/j.1462-2920.2009.01982.x

Towards the human intestinal microbiota phylogenetic core

Julien Tap,[1] Stanislas Mondot,[1] Florence Levenez,[1] Eric Pelletier,[2,3] Christophe Caron,[4] Jean-Pierre Furet,[1] Edgardo Ugarte,[2,3] Rafael Muñoz-Tamayo,[1,5,6] Denis L. E. Paslier,[2,3] Renaud Nalin,[7] Joel Dore[1] and Marion Leclerc[1]*

[1]*INRA, UEPSD, UR910, 78350 Jouy en Josas, France.*
[2]*CEA, DSV, IG, Genoscope, 91057 Evry, France.*
[3]*CNRS UMR 8030, 91057 Evry, France.*
[4]*INRA, MIG, UR1077, 78350 Jouy en Josas, France.*
[5]*INRA, MIA, UR341, 78350 Jouy en Josas, France.*
[6]*L2S, UMR8506, Univ. Paris Sud-CNRS-SUPÉLEC, 91190 Gif sur Yvette, France.*
[7]*Libragen, 31400 Toulouse, France.*

Summary

The paradox of a host specificity of the human faecal microbiota otherwise acknowledged as characterized by global functionalities conserved between humans led us to explore the existence of a phylogenetic core. We investigated the presence of a set of bacterial molecular species that would be altogether dominant and prevalent within the faecal microbiota of healthy humans. A total of 10 456 non-chimeric bacterial 16S rRNA sequences were obtained after cloning of PCR-amplified rDNA from 17 human faecal DNA samples. Using alignment or tetranucleotide frequency-based methods, 3180 operational taxonomic units (OTUs) were detected. The 16S rRNA sequences mainly belonged to the phyla *Firmicutes* (79.4%), *Bacteroidetes* (16.9%), *Actinobacteria* (2.5%), *Proteobacteria* (1%) and *Verrumicrobia* (0.1%). Interestingly, while most of OTUs appeared individual-specific, 2.1% were present in more than 50% of the samples and accounted for 35.8% of the total sequences. These 66 dominant and prevalent OTUs included members of the genera *Faecalibacterium*, *Ruminococcus*, *Eubacterium*, *Dorea*, *Bacteroides*, *Alistipes* and *Bifidobacterium*. Furthermore, 24 OTUs had cultured type strains representatives which should be subjected to genome sequence with a high degree of priority. Strikingly, 52 of these 66 OTUs were detected in at least three out of four recently published human faecal microbiota data sets, obtained with very different experimental procedures. A statistical model confirmed these OTUs prevalence. Despite the species richness and a high individual specificity, a limited number of OTUs is shared among individuals and might represent the phylogenetic core of the human intestinal microbiota. Its role in human health deserves further study.

Received 5 November, 2008; accepted 28 May, 2009. *For correspondence. E-mail marion.leclerc@jouy.inra.fr; Tel. (+33) 1 34 65 23 06; Fax (+33) 1 34 65 24 92.

Introduction

The human gut microbiota is a complex ecosystem, which is now recognized as a key component in gastrointestinal tract (GI tract) homeostasis. Its involvement in immune diseases has recently been demonstrated and bacterial imbalance or so-called 'dysbiosis' has been associated with pathologies such as inflammatory bowel disease and obesity (Marteau *et al*., 2004; Ley *et al*., 2005; 2006; Swidsinski *et al*., 2005). These observations have stirred a renewed interest into the mechanisms underlying such imbalances and a search for biomarkers of healthy versus diseased GI tract microbiota.

Culture-based methods initially provided a basic knowledge on numbers and diversity of culturable microorganisms from human GI tract. Bacterial diversity was estimated to exceed 400 culturable species and two archaeal methanogenic species were isolated from human faecal samples (Savage, 1977; Miller *et al*., 1982; Finegold *et al*., 1983). Molecular analysis based on rDNA gene structure (Woese *et al*., 1975; 1990), by targeting both cultured and uncultured microorganisms, shed light on microbial diversity (Amann *et al*., 1995). In human GI tract, depending on the method, 10–50% microbial population was reported uncultured (Amann *et al*., 1995; Zoetendal *et al*., 2004; Ley *et al*., 2006).

The very first 16S rDNA molecular inventories of healthy human faecal microbiota (Wilson *et al*., 1997; Suau *et al*., 1999) had demonstrated the high diversity of this ecosystem and pointed to the important number of molecular species that did not correspond to any cultured strains from available collections. Improved technical performances have since led to higher numbers of clones investigated in studied data sets (Eckburg *et al*., 2005). Furthermore, within the last few years, metagenomics, thanks to PCR-free identification, has been offering a new

insight into microbial diversity of the dominant microorganisms (Gill *et al*., 2006; Manichanh *et al*., 2006). Hence revisited, the human GI tract microbiota appeared dominated by very few phyla when compared with other complex ecosystems such as soils and oceans (Cole *et al*., 2005), but nonetheless highly diverse and complex at the level of 'phylotypes'.

Profiling techniques targeting 16S rRNA genes indicated that the human GI tract microbiota was stable over time through adulthood (Zoetendal *et al*., 1998; Sutren *et al*., 2000) and resilient to antibiotic treatment (De La Cochetiere *et al*., 2005). Most importantly, it showed an important subject specificity in composition and species diversity (Zoetendal *et al*., 1998).

At a macroscopic level, however, the microbiota supports a common set of metabolic pathways assembled in a trophic chain common to all healthy individuals (Macfarlane and Gibson, 1994), with fermentation of dietary compounds and endogenous substrates, followed by host absorption and excretion of SCFA (acetate, propionate, butyrate) and gas. Although the microbiota composition seems to be host specific, the high degree of conservation in its expressed functions and metabolites between humans should translate into conserved features of the environmental metabolome and proteome, derived from redundancies in the GI tract microbiota transcriptome and genome. We hypothesized that this should be supported by the existence of a bacterial 'phylogenetic core' in healthy adult faecal microbiota, consisting of a set of dominant and prevalent microbial species. Extensive molecular inventories of 16S rRNA genes were generated for the faecal microbiota of 17 healthy individuals. Candidate core species present in more than 50% of individuals in the studied cohort were identified and further validated against recently published 16S rDNA sequence data sets of human faecal microbiota from other countries. This observation should have major implications in human GI tract microbiomics.

Results

Richness and diversity of human adult faecal microbiota

From the global analysis of the 10 456 sequences, 3180 operational taxonomic units (OTUs) were obtained for the 17 subjects (Table S1). The total number of OTUs differed by less than 4% according to the analysis software, from 3180 to 3186 with CLUSTALW and MAFFT respectively. Furthermore, when tetranucleotide frequency method was used instead of alignment, 3097 OTUs were obtained (Table S2).

The Chao1 estimation of total richness for the whole sequences set, whatever the alignment or clustering method, led to very similar curves (Fig. 1). The cumulative number of OTUs linearly increased, up to 8000 analysed. For more than 8000 clones, a plateau seemed to be reached, indicating that the sampling effort from this data set allowed the estimation of dominant bacterial richness. From this analysis, the faecal microbiota of 17 healthy adults would at least reach 9940 OTUs.

When each subject data set was considered separately, the average OTUs number per subject was 259, ranging from 159 to 383 (Table 1). There was no correlation between OTUs numbers and the number of sequences

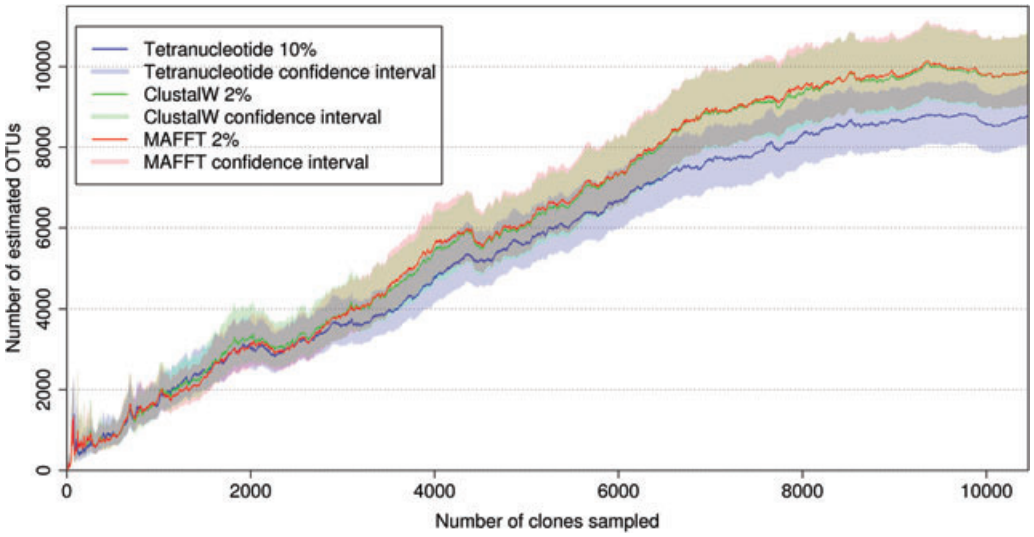


Fig. 1. Chao1 estimates of human gut bacterial richness as a function of sample size. Sequences analysis methods: blue, tetranucleotide frequency; green, alignment with CLUSTALW; red, alignment with MAFFT. Ninety-five per cent confidence intervals were computed with DOTUR. Given the OUT definition, the total bacterial richness estimated by Chao1 did not significantly differ according to the sequence analysis methods, because the confidence intervals overlapped at the significance level of 0.05.

Table 1. Characteristics of human fecal samples, and sequence data. Fecal samples were from 17 healthy adult individuals, eight males and nine females, between 28 and 54 years old, living in France or in the Netherlands. Eight individuals followed a vegetarian diet, with various daily intakes regarding protein sources, dairy products, fibers, from vegetarian to vegan. The others were omnivorous, with also differences in diet. Diets, country, DNA concentration, chimera checked sequences, sequence accession numbers are detailed in Table S1.

Sample	Sex	Age	Number of unambiguous sequences	Number of OTUs (2%)	Estimated richness (Chao1)	Estimated diversity (Simpson; 1-D)
AA	M	39	636	256	886.4	0.9773
AB	F	39	468	236	819.4	0.9695
AC	M	45	679	276	948.5	0.9876
AD	F	34	633	235	580.4	0.9795
AF	F	41	619	245	1110.3	0.9802
AG	M	33	500	234	532.4	0.9894
AH	M	36	426	195	931.3	0.9658
AI	F	28	625	285	954.6	0.9841
AL	M	54	603	326	1651.1	0.9864
AM	F	41	573	254	901.5	0.9881
AN	F	31	491	278	1478.0	0.9894
AP	F	49	653	383	1294.0	0.9942
AQ	M	33	655	271	992.0	0.9449
AR	F	31	607	297	797.7	0.9885
AS	F	32	550	296	1008.5	0.9908
AT	M	37	839	175	343.1	0.9257
AV	M	29	899	159	288.0	0.9136

obtained per individual ($r^2 = 0.00056$, $P = 0.7754$, Spearman method). Unambiguous sequences per individual ranged from 426 to 899 (Table S1). Rarefaction curves did not show any plateau except for samples AT and AV (Fig. S1). In addition, diet did not have a statistically significant impact on diversity, since the diversity detected within the microbiota associated to vegetarian or omnivorous diet did not statistically differ from the overall diversity (AMOVA calculations, Table S3). The estimated richness averaged 943 OTUs per subject, and drastically differed between individuals, ranging from 288 to 1651. At the subject level, the Chao1 estimated richness did not reach saturation except for the two samples AT and AV for which both Chao and Simpson indexes indicated a lower diversity (Table 1).

Taxonomic description of global and individual libraries

The taxonomic affiliation of the 10 456 sequences 16S rRNA gene sequences confirmed that the dominant human faecal microbiota belonged to five phyla, with 79.4% *Firmicutes*; 16.9% *Bacteroidetes*; 2.5% *Actinobacteria*; 1% *Proteobacteria*; 0.1% *Verrucomicrobia*; and 0.1% others (data not shown). Differences were observed in the taxonomic make-up of the 17 individual libraries. The proportions of the three major phyla varied, from one sample with only few sequences related to the *Clostridium leptum* cluster, to another sample with only one OTU belonging to the *Bacteroidetes* phylum (assigned to the genus *Alistipes*). It was noticeable that for most of the genera, OTUs were not evenly distributed: most OTUs gathered only few sequences and, conversely, few OTUs gathered most of the sequences found in the corresponding genus.

Quantitative PCR (qPCR) results were consistent with molecular inventories data and confirmed this taxonomic composition of the libraries. The same average composition of taxonomic groups was obtained when qPCR data versus cloning-based sequencing were compared. Indeed, the *Firmicutes* members dominated, with *C. leptum* cluster IV, *Clostridium coccoides* cluster XIV and *Bacteroides*/*Prevotella* as the most prevalent groups (Table S4). When few sequences were assigned to a group, the qPCR results demonstrated the same trend. At a subdominant species level, molecular inventories and qPCR were also consistent for *Escherichia coli* determination. However, the qPCR results and the molecular inventory taxonomic assignment of the sequences from the genera *Lactobacillus* and *Bifidobacterium* were not in agreement.

A set of OTUs shared among individuals

Among the 3180 OTUs detected, 2500 OTUs were present in only one sample, which represented 78.6% of subject specificity (Fig. 2). All the 680 remaining OTUs (21.4%) were common to at least two samples. However, none of the OTUs could be detected in all samples. The prevalence curve followed an increase towards a limited number of OTUs detected in more than half of the samples (Fig. 2). Interestingly, 66 OTUs, representing 2.1% of the total detected OTUs, were present in more than 50% of the individuals of the study. In addition, they represented 35.8% of the sequences (3740 sequences).

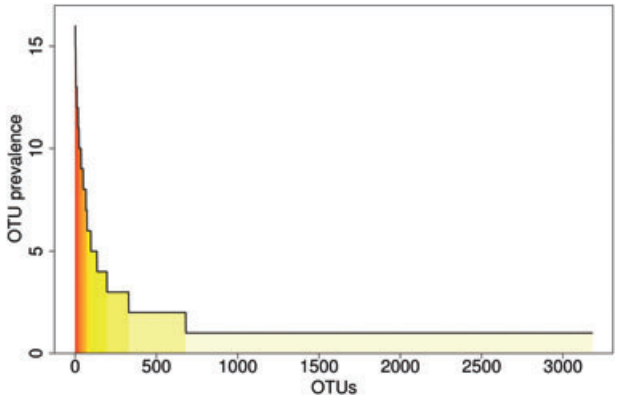


Fig. 2. Distribution of OTUs as a function of their prevalence in the 17 individuals. Operational taxonomic units were ranked from the most prevalent (present in 16/17 individuals) to the least prevalent ones (individual specific). Most prevalent OTUs, present in 8 out of 17 individuals or more, corresponded to 2.1% of all OTUs ($n = 66$) but represented 35.8% of all sequences ($n = 3740$).

These 66 OTUs appeared at the same time more frequently shared among individuals and accounting for more sequences, indicating that they might represent a phylogenetic core.

Taxonomic distribution of phylogenetic core OTUs

The diversity originating from the 17 faecal microbiota was mapped using principal component analysis (PCoA) (Fig. 3). The core OTUs were not restricted to a specific genus or even phylum, but fell into distinct phyla and families, with the prevalent and dominant members of *Bacteroides vulgatus*, *Roseburia intestinalis*, *Ruminococcus bromii*, *Eubacterium rectale*, *Coprobacillus* sp., *Bifidobacterium longum* (Fig. 3). The OTU with the highest prevalence, 16 out of 17 individuals, belonged to *Faecalibacterium prausnitzii*. At the opposite, some OTUs from the core represented by few sequences appeared less visible, such as an OTU classified as a *Lachnospiraceae*, shared by eight subjects but only represented by 11 sequences. At the same time, one OTU specific to AT sample was represented by more than 150 sequences. These observations suggest that abundance was not invariably related to frequency of observation.

The phylogenetic core of healthy humans' faecal microbiota herein described exhibited representatives of the main phyla, and the 66 OTUs belonged to 18 genera (Fig. 4). However, compared with the whole data set, the *Firmicutes* phylum was highly represented in the core (57/66 OTUs), while the *Bacteroidetes* phylum only accounted for seven OTUs.

Each individual microbiota contributed to the phylogenetic core and harboured an average of 40 OTUs from the phylogenetic core, ranging from 20 to 49 OTUs (Fig. 4). AT sample with a lesser diversity [Chao1 = 343.115 and Simpson (1-D) = 0.9257] also provided a lesser contribution to the phylogenetic core. There was, however, no correlation between the contribution to the core and the total number of OTUs, per sample ($r^2 = 0.1196$, $P = 0.1739$). Each sample harboured core OTUs from the two main phyla *Bacteroidetes*, *Firmicutes* and 14 out of 17 from the *Actinobacteria*. A similar trend was observed at the genus level. For instance, except for two of them, all samples exhibited at least four OTUs assigned to the genus *Faecalibacterium*. Similarly, all samples harboured at least one OTU assigned to the genus *Roseburia* and to the *Bacteroides* (except subject AL).

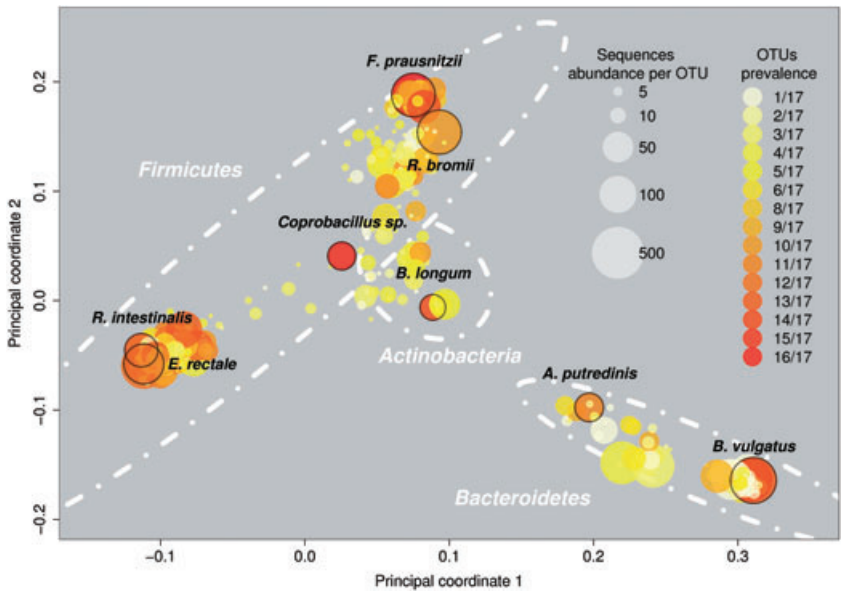


Fig. 3. Principal coordinate analysis of OTUs from the faecal microbiota of 17 healthy human individuals. A principal coordinate analysis was performed using the full distance matrix. Each OTU was pictured as a disk whose area was proportional to the number of sequences and the heat colours accounted for the prevalence among the 17 individuals. Operational taxonomic units represented by a unique sequence (singleton) were not plotted.

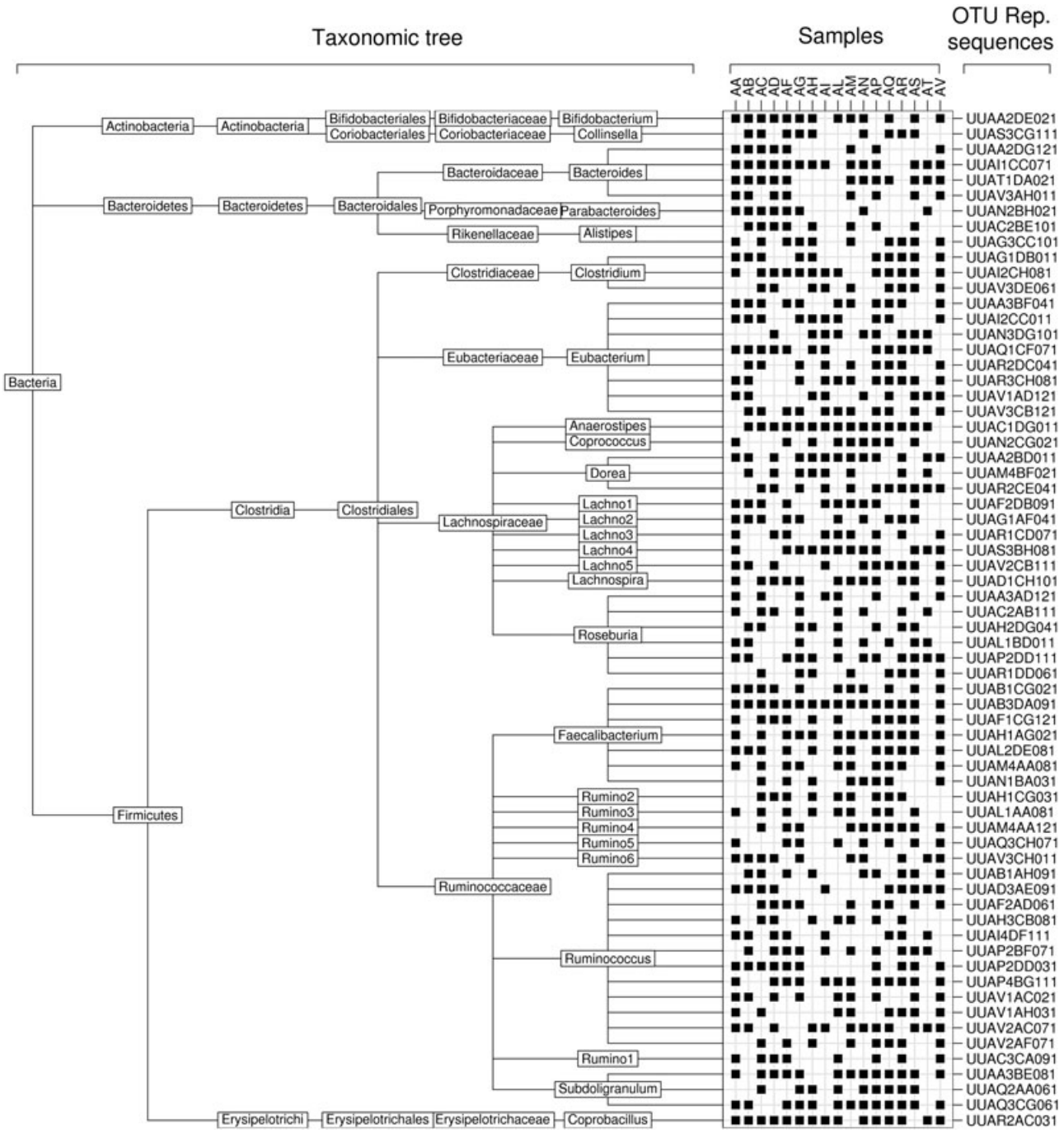


Fig. 4. Taxonomic and prevalence characterization of the phylogenetic core. Sixty-six OTUs present in at least 8 individuals out of 17 were shown, the black dot representing their detection in a given individual. The taxonomic assignment of the 66 OTUs was obtained using classifier (RDP II release 9.61). The tree was built using ade4 package in R. 'Rumino' and 'Lachno' indicated OTUs whose taxonomic affiliation could only reach the family levels, *Lachnospiraceae* and *Ruminococcaceae* respectively.

In addition, when compared with the cultivated type strains from RDP II, 38 OTUs (58%) were similar to a cultivated species, with a 2% sequence dissimilarity threshold (Table S5). Among the *Bacteroidetes*, the species were *Bacteroides stercoris*, *B. vulgatus*, *B. massiliensis*, *Parabacteroides distasonis*, *Alistipes putredinis*, *Alistipes shahii*, and among the *Firmicutes*, the species were *F. prausnitzii*, *Ruminococccus obeum*, *R. bromii*, *E. rectale*, *E. halii*, *E. eligens*, *Dorea longicatena*. Only two cultured strains from the *Actinobacteria* were represented, *B. longum* biovar *longum* and *Colinsella aerofaciens*. At the opposite, among the 42% not assigned to a

species, 14 OTUs, from the *Firmicutes* and to a lesser extent from the *Bacteroidetes* phylum, were distant by more than 5% sequence divergence from the closest cultivated type strains.

Statistical characterization of the phylogenetic core

Based on the statistical model and the chosen criterion (50% of individuals), a subset of 49 OTUs (on a total of 3180 OTUs) was selected as the putative core. These 49 OTUs were the most prevalent among the 66 previously selected. All core OTUs were described with their corresponding probability estimates, within a 95% confidence interval and their normalized abundance *pj* in the core (Table S6). The calculation of confidence intervals attached to the probabilities estimation, enabled to evaluate the uncertainty of this assessment of the core. According to the confidence intervals, the 10 most frequent OTUs, very likely to be part of the core with respect to the 50% threshold, were related to the following species: *F. prausnitzii*; *Anaerostipes caccae*; *Clostridium spiroforme*; *Bacteroides uniformis*; *D. longicatena*; *B. longum* biovar *longum*; *Clostridium* sp. BI-114; *Clostridium bolteae*. Furthermore, in order to take into account the number of sequences per OTU in the core set, the normalized abundance of the OTUs was calculated and varied from 0.5% to 9%. Ten OTUs with the highest normalized abundance would have an important contribution to the core, and were affiliated to their closest isolated type strain from RDP II database (Fig. S2).

Core OTUs presence in external data sets

A systematic comparison of the sequences originating from this data set against the published libraries was performed, in order to get a broader estimation of OTU redundancy (i.e. recovery of the same OTUs in four libraries from other international studies), while taking into account biases associated with experimental procedures. From the whole data set, 17% of OTUs were present in other 16S rRNA libraries, and 83% (3780 sequences) were specific to this study (Fig. S3).

Strikingly, the 66 OTUs demonstrated a higher prevalence in public data sets (Fig. 5). All of them were detected at least once in the four external libraries, and 78.8% of them (52 OTUs) were detected in at least three of these four libraries. When the core OTUs highlighted by the statistical model were subjected to the same analysis, this occurrence in at least three libraries reached 81.6%.

When the presence in all data sets was the criterion, 24 core OTUs were retrieved. They all belonged to the *Firmicutes*, and, for example, the OTUs assigned to the genus *Faecalibacterium* were all detected in the four

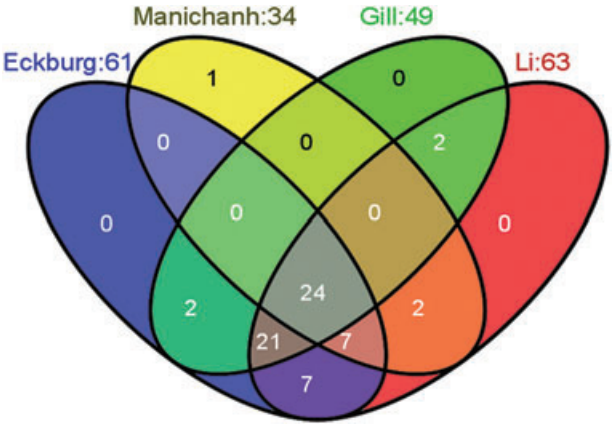


Fig. 5. Venn diagram representation of 66 putative core OTUs hits against external libraries. The occurrence of the 66 prevalent OTUs was assessed in the publicly available 16S rRNA libraries. Sequences originating from healthy individual faecal samples only were downloaded from GenBank from four external libraries: Eckburg and colleagues (2005) (2339 sequences); Gill and colleagues (2006) (2062 sequences); Manichanh and colleagues (2006) (539 sequences); Li and colleagues (2008) (5413 sequences). BLASTN algorithm was used to determine the OTU occurrence in external libraries with a minimum coverage of 900 bases pairs and a minimum pairwise identity of 98%. Four-way Venn diagrams were plotted with VENNY (http://bioinfogp.cnb.csic.es/tools/venny/index.html).

external libraries. Conversely, the representation of OTU from other phyla was different: one OTU was only found in this study and Manichanh and colleagues (2006) and shared more than 99% of similarity with the species *B. longum* (NCC2705 strain). Seven OTUs assigned to the phylum *Bacteroidetes* were not found in Gill and colleagues (2006) library but at least twice in the other libraries.

Overall, the criterion chosen for phylogenetic core determination seemed robust. From the biological data obtained in this study and in the so far published data sets, which were confirmed by statistical models, a set of approximately 50 bacterial species may represent part of the healthy human phylogenetic core.

Discussion

The goal of this study was to assess the existence of a phylogenetic core, consisting of a set of dominant species prevalent among healthy adults. Because of the recent demonstrations of strong links between phylogenetic dysbiosis and health impairment or diseases, such a group of microorganisms are expected to play a preponderant role in gut homeostasis and human health.

A precise quantification of the extent of human GI tract diversity has indeed been a critical ecological question for more than 30 years. The estimate of 400 cultivated species (Savage, 1977; Finegold *et al.*, 1983) was eclipsed by 16S rRNA-targeted molecular studies and

numbers from several thousands (Eckburg *et al*., 2005) up to 40 000 of species have been estimated (Frank *et al*., 2007). It remains critical to circumscribe the GI tract microbial diversity inherent to humans. From this data set, Chao estimates indicated that the human gut microbiota richness could reach a saturation corresponding to at least 10 000 OTUs, which is much higher than previously reported (Eckburg *et al*., 2005). Taxonomic make-up of the libraries was consistent with previous study, even though in Eckburg and colleagues (2005), the estimated richness per individual was lower than the least diverse sample from this study.

In this study, 3180 OTUs were observed and this appeared as the highest diversity ever obtained with PCR-based method, and for the first time 17 individuals were investigated. Furthermore, the OTUs sequences were all more similar to human GI tract species than to any other clone sequences from the databases. This suggests a larger trend in microbial evolution that faecal microbiota communities of same species (conspecific) appeared more similar to each other than to those of different host species.

Core OTUs were first chosen as present in more than 8 out of 17 of individuals. The further comparison with publicly available human data sets strongly confirmed the prevalence of these core OTUs. Strikingly, these experiments sampled the same core OTUs, even though they were performed worldwide with very different protocols (sample handling, DNA extraction, Eubacteria-Universal PCR primers, chimera detection procedure) known to lead to different pictures of microbial diversity (Suau *et al*., 1999; Kurokawa *et al*., 2007; Li *et al*., 2008). Most of them were present in three out of the four available sequences data sets on healthy human faecal samples, obtained in Japan or in the USA. The only differences were the underrepresentation in other libraries of core sequences related to *Bacteroides* and *Bifidobacterium* genera, whose occurrences have already been discussed by Kurokawa and colleagues (2007) and Suau and colleagues (1999).

In addition to the biological investigations, the probability estimates from the binomial distribution of OTUs enabled to model, as the core set, the 49 most prevalent OTUs from the primary selection of 66. The calculation of confidence intervals attached to the probabilities estimation, enabled to evaluate the uncertainty of the assessment of the core. In this way, according to the chosen criterion (> 50% of individuals), the first 10 OTUs with the highest probabilities were statistically considered to be part of the core. Additional data would improve the estimation and the narrowing of the confidence intervals because the uncertainty of the probability estimates is still high, due to small sample size ($n = 17$). In addition, in the statistical analyses, no distinction was made between the sample of OTUs experimentally detected and the real microbiota. As a consequence, one may expect an underestimation of OTUs present at a low abundance level, close to detection threshold.

The high prevalence of OTUs was also an indication of the species persistence in the human GI tract, and several ecological factors could account for it. In terms of conditions linked to the ecosystem, attachment to food particles, resistance to stress such as pH or mechanical forces of peristaltic movement, would prevent the species from a wash-out phenomenon. From a metabolic point of view, an inference to the putative role of the core species could be attempted from the close strains that are already sequenced or characterized. Their known metabolic functions in anaerobic degradation of food polymers or their immunological properties in relation to the host epithelium would add critical information on the core putative proteins and metabolites pool. 24 OTUs from the core were closely related to cultivated type strains from the species *E. rectale*, *R. bromii*, *F. prausnitzii*, *Clostridium* sp. BI-114, *B. stercoris*, *B. vulgatus*, *P. distasonis*, *A. putredinis*, *R. obeum*, *E. halii*, *D. longicatena*.

Interestingly, a large range of metabolic functions regarding the carbohydrate catabolism trophic chain were covered since hydrolytic, fermentative, hydrogenotrophic properties, and butyrate, lactate or acetate production could be inferred from OTUs phylogenetic position. Whether the core OTUs represent a set of species sufficient for anaerobic degradation of dietary fibres remains to be determined. A large proportion cannot be cultured; it has, however, been recently shown that assignation of several metabolic signatures to uncultured microbial population was possible (Li *et al*., 2008). This robustness has indeed been described to be related to the functional redundancy of a microbial ecosystem.

From these data, however, the diversity structure appeared to interestingly depend on the genus considered. Furthermore, the diversity structure at different taxonomic levels can indeed be seen as a way to investigate the impact of host on community composition. Even though a 16S rRNA sequence dissimilarity of 3% had been used for molecular species characterization (Stackebrandt and Goebel, 1994) dissimilarity cut-off varied in recent reports on human GI tract microbiota (Suau *et al*., 1999; Eckburg *et al*., 2005; Gill *et al*., 2006). Interestingly, in this study, the same shape of rarefaction curves was obtained when the dissimilarity cut-off ranged from 1% to 5%. Furthermore, tetranucleotide frequency count (Teeling *et al*., 2004) also showed the same trend and this work confirmed that this non-alignment-based method enabled a fast and accurate phylogenetic assignation. A similar approach had been previously described, including the human GI tract (Rudi *et al*., 2007).

One interesting outcome of the large number of sequences per individual performed in this study con-

cerned *Faecalibacterium* genus diversity. *Faecalibacterium prausnitzii*-related sequences have been repeatedly recovered among the most prevalent species, and described as dominant in healthy individuals and underrepresented in patients with inflammatory bowel disease (Manichanh *et al*., 2006). Originally described for its butyrate production (Duncan *et al*., 2002), its anti-inflammatory properties have very recently been described (Sokol *et al*., 2008). Based on the seven distinct OTUs identified in more than 50% of the individuals of this study, we hypothesized a more important phylogenetic and functional diversity in this genus, which would be consistent with the connection of *F. prausnitzii*-related sequences to different metabolites (Li *et al*., 2008).

When diversity was specifically observed at an individual level, a strong host adaptation could be emphasized. For example, the low number of core OTUs from the *Bacteroidetes* phylum may not only be linked to technical differences between the studies or to lower sequence number. Recently, the compositional complexity of this genus was highlighted in human gut metagenomes (Kurokawa *et al*., 2007) and similarly, among the 17 individuals of this study, the individual variability among the *Bacteroides* genus was particularly high.

As another evidence supporting the core concept, a very high individual variability was observed, consistent with earlier works using Ribotyping methods (Zoetendal *et al*., 1998; Sutren *et al*., 2000). Sequence data demonstrated that 78.6% OTUs were specific of a given individual. As a confirmation, when these OTUs were compared with external databases, the prevalence was not high. Quantitative PCR data revealed the same high variability, particularly for the *Actinobacteria* quantity. Furthermore, when the diversity according to age, country of origin, diet was tested with AMOVA, the individual variability, which could be partly random, explained most of the difference.

It meant that the dietary habits (vegetarian versus omnivorous) did not explain much of the genetic diversity. In addition, clone frequencies distribution between vegetarians and omnivorous, statistically compared using discriminant analysis, only explained 5% of variability. More samples and time series, together with genomic characterization, are required to assess how diet shapes the human gut microbiota.

A number of core OTUs were present in all checked databases, pointing as an outcome of this work to give high priority for the sequencing of those strains. Reference genomes are required for the characterization of human gut microbiome and cultured representatives 'have to be selected based on comprehensive 16S rDNA gene based survey' (Turnbaugh *et al*., 2007). Twenty-four OTUs from the core were close to cultivated type strains, with some of them already being sequenced. However, the numerous OTUs far from cultivated strains should also be targeted using cell-sorting strategies and new single-cell sequencing technologies.

Metagenomic data sets have already started to shed light on the functional redundancy between healthy individuals (Gill *et al*., 2006; Kurokawa *et al*., 2007). Future studies on larger individual cohorts will enable to explore the link between gene redundancy and the prevalence of members of the putative phylogenetic core. Statistical models, as developed in this study, are also required in a broader perspective, to estimate sampling depth and number of individuals needed to characterize the 'full' human microbiome.

It is now recognized that microbial groups' imbalance can be linked to diseases. This work, together with others, leads towards a set of species important for human health. If confirmed, the main outcomes of this work will be the design and application of a fast screening of the phylogenetic core as a diagnostic tool. The next step for a better understanding will be to assess how the transformation of human lifestyle influences the microorganisms evolution and thereby health and predisposition to various diseases.

Experimental procedures

Subjects and sampling

The 17 study subjects were healthy adults between 29 and 54 years old, male and female, living in France or in the Netherlands (Table 1). Eight subjects followed a vegetarian diet, with various daily intakes regarding protein sources, dairy products, fibres, constituting a panel from vegetarian to vegan diet. The nine other subjects were omnivorous, with also differences in diet. Faecal samples were stored in sterile Sarstedt tube at –80°C until further processing. None of the volunteers had received antibiotic treatment 6 months prior to sampling.

Extraction of genomic DNA

Total DNA was extracted from 0.2 g of faecal samples, using a bead-beating method as previously described (Godon *et al*., 1997). The DNA preparation for AV sample was performed as previously described (Courtois *et al*., 2003). DNA concentration and purity was estimated by gel electrophoresis and spectrometry (NanoDrop).

Bacterial 16S rRNA amplification

The 16S rDNA genes were amplified from extracted DNA using bacterial primers U-350f (5′-CTCCTACGGGAGG CAGCAGT-3′) (Amann *et al*., 1990) and P-1392r (5′-GCGGTGTGTACAAGACCC-3′) (Kane *et al*., 1993). PCR reactions were run as previously described (Suau *et al*., 1999), using AmpliTaq Gold DNA Polymerase (Applied Biosystems) and a PTC 100 Thermocycler (MJ Research).

Three PCR products from each extracted DNA sample were pooled and purified using Qiaquick PCR purification kit columns (Qiagen), checked and stored at –20°C.

Cloning and sequencing

Cloning and sequencing were performed at the national sequencing centre CEA-Genoscope (Evry, France). Purified PCR products were ligated into pCR-4TOPO TA vectors and electroporated into *E. coli* DH10B-T1 cells, according to the manufacturer's recommendation (Invitrogen). A total of 1500 colonies from each transformation were randomly picked. Bidirectional Sanger sequence reads were trimmed and assembled by PHRED-PHRAP (http://www.phrap.org/phredphrapconsed.html). Sequences orientation were checked using BLASTN (Altschul *et al*., 1997) against the RDP II database. One per cent ambiguous nucleotide was tolerated for sequences with 900 bp length cut-off.

Sequences analysis and OTU representative sequences detection

Chimera check was performed using MALLARD software (Ashelford *et al*., 2006). From 15 532, a strict elimination led to 10 456 unambiguous sequences, which were then analysed using RapidOTU (Legrand *et al*., 2008). RapidOTU, freely available at http://genome.jouy.inra.fr/rapidotu/, and offering up to 64 processors upon request, is a perl-script written pipeline, connecting software for automatic analysis of 16S rRNA genes libraries. Multiple alignment was obtained with CLUSTALW (Thompson *et al*., 1994; Li, 2003) or MAFFT algorithm (Katoh *et al*., 2005). The computing of a precise alignment of the 10 456 sequences on 1317 gapped base pairs was possible by using a perl-script program enabling the parallelization of CLUSTALW. The distance matrixes (F84 model) were computed by fdnadist (PHYLIP package: http://evolution.genetics.washington.edu/phylip.html) (Felsentein, 1989). Tetranucleotide frequency count using OCOUNT (Teeling *et al*., 2004), implemented within the RapidOTU pipeline, was also used to cluster the sequences, and Pearson matrixes were built and converted into distance matrixes. Operational taxonomic units were detected using DOTUR (Schloss and Handelsman, 2005) with a default 2% sequence dissimilarity cut-off. RepOTUfinder, a newly designed tool implemented in RapidOTU, automatically selected and extracted a representative sequence for each OTU by calculating the central sequence, the ones with the lowest distance with all the other OTUs sequences. The 10 456 sequences have been submitted to DDBJ/EMBL/GenBank databases under the accession numbers (FP074904 to FP085359).

Ecology analysis and core phylogenic detection

Ecology analyses were performed on the individual and on the complete 16S rDNA data set. DOTUR files were used to map rarefaction curves and to compute Chao1 estimated OTU richness profiles. Simpson indices (1-D) of variability between samples were obtained from the phylotypes abundances. To assess diet impact on genetic diversity, AMOVA was computed using ade4 statistical package (Chessel *et al*., 2004).

Genetic diversity of the whole data set was represented by a PCoA analysis, computed using R software (http://pbil.univ-lyon1.fr/ADE-4/). The distance matrix of the 3180 OTUs representative sequences was computed using the SeqinR package (Charif and Lobry, 2007) and transformed into an Euclidean matrix before the PCoA analysis.

Operational taxonomic unit prevalence was determined as the sum of their occurrence in the 17 individual 16S rRNA gene libraries. Taxonomic characterization of the OTUs was performed using the RDP II Classifier program (RDP II Release 9.58) and diagram computation with the ade4 statistical package (Chessel *et al*., 2004). The similarity between core OTUs sequences and isolated type strains was obtained by BLASTN against the 5171 isolated type strains 16S rDNA sequences from RDP II.

16S rRNA gene qPCR

Quantitative PCR was performed on 16 of the faecal DNAs using probes and settings previously described (Furet *et al*., 2009). Quantitative PCR systems targeted Eubacteria, and within the *Firmicutes C. leptum* group (*Clostridium* cluster IV), *C. coccoides* group (*Clostridium* cluster XIV), *Bacteroides–Prevotella*, *E. coli*, *F. prausnitzii* (Sokol *et al*., 2008), *Lactobacillus–Leuconostoc* and *Bifidobacterium*.

Statistical detection of a putative phylogenetic core

Assuming that there was not dependence between individuals, a statistical model was used to define a putative phylogenetic core. The presence/absence of the OTUs was represented as a binomial distribution based on the prevalence, where γj denoted the probability that the OTU j is detected in an individual (details in Appendix S1) (Wilson, 1927; Agresti and Coull, 1998). The parameter γj did not provide information about the abundance of the OTUs in the global data set. In order to also have a representation of the abundance, the numbers of sequences of each OTU were averaged on the subset of individuals where the OTU was detected. Afterwards, the average abundances were normalized to have a unitary representation of the core.

Detection of core OTUs in external data sets

From the four published studies on human microbiota, the 16S rRNA gene sequences linked to healthy adult faecal samples were selected and downloaded from GenBank. Comparisons of the 3180 OTUs or the 66 core OTUs were performed using BLASTN with 98% identity threshold and a 900 bases minimum coverage for a given pairwise aligned sequences. Results were shown in a four-way Venn diagram plotted with VENNY (http://bioinfogp.cnb.csic.es/tools/venny/index.html).

Acknowledgements

We are very grateful to Dr E. Zoetendal (Laboratory of Microbiology, Wageningen University, the Netherlands) for provid-

ing us with samples and nutritional information; to Dr K. Kiêu (MIA, INRA, France) for helpful discussions on the statistical approach. J. Tap's PhD and this project are supported by the French National Agency for Research, ANR/DEDD/PNRA/PROJ/200206-01-01, within the AlimIntest program.

References

Agresti, A., and Coull, B.A. (1998) Approximate is better than exact for interval estimation of binomial proportions. *Am Statistician* **52:** 119–125.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402.

Amann, R.I., Binder, B.J., Olson, R.J., Chisholm, S.W., Devereux, R., and Stahl, D.A. (1990) Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Appl Environ Microbiol* **56:** 1919–1925.

Amann, R.I., Ludwig, W., and Schleifer, K.H. (1995) Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol Rev* **59:** 143–169.

Ashelford, K.E., Chuzhanova, N.A., Fry, J.C., Jones, A.J., and Weightman, A.J. (2006) New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. *Appl Environ Microbiol* **72:** 5734–5741.

Charif, D., and Lobry, J.R. (2007) *SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis*. New York, USA: Springer Verlag.

Chessel, D., Dufour, A.-B., and Thioulouse, J. (2004) The ade4 package-I – One-table methods. *R News* **4:** 5–10.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam, S.A., McGarrell, D.M., *et al.* (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res* **33:** D294–D296.

Courtois, S., Cappellano, C.M., Ball, M., Francou, F.X., Normand, P., Helynck, G., *et al.* (2003) Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. *Appl Environ Microbiol* **69:** 49–55.

De La Cochetiere, M.F., Durand, T., Lepage, P., Bourreille, A., Galmiche, J.P., and Dore, J. (2005) Resilience of the dominant human fecal microbiota upon short-course antibiotic challenge. *J Clin Microbiol* **43:** 5588–5592.

Duncan, S.H., Hold, G.L., Harmsen, H., Stewart, C.S., and Flint, H.J. (2002) Growth requirements and fermentation products of *Fusobacterium prausnitzii*, and a proposal to reclassify it as *Faecalibacterium prausnitzii* gen. nov., comb. nov. *Int J Syst Evol Microbiol* **52:** 2141–2146.

Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., *et al.* (2005) Diversity of the human intestinal microbial flora. *Science* **308:** 1635–1638.

Felsentein, J. (1989) PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics* **5:** 164–166.

Finegold, S.M., Sutter, V.L., and Mathisen, G.E. (1983) Normal indigenous intestinal flora. In *Human Intestinal Microflora in Health and Disease*. Hentges, D.J. (ed.). New York, USA: Academic Press, pp. 3–31.

Frank, D.N., St. Amand, A.L., Feldman, R.A., Boedeker, E.C., Harpaz, N., and Pace, N.R. (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci USA* **104:** 13780–13785.

Furet, J.P., Firmesse, O., Gourmelon, M., Bridonneau, C., Tap, J., Mondot, S., *et al.* (2009) Comparative assessment of human and farm animal faecal microbiota using real-time quantitative PCR. *FEMS Microbiol Ecol* **19:** 19.

Gill, S.R., Pop, M., DeBoy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., *et al.* (2006) Metagenomic analysis of the human distal gut microbiome. *Science* **312:** 1355–1359.

Godon, J.J., Zumstein, E., Dabert, P., Habouzit, F., and Moletta, R. (1997) Molecular microbial diversity of an anaerobic digestor as determined by small-subunit rDNA sequence analysis. *Appl Environ Microbiol* **63:** 2802–2813.

Kane, M.D., Poulsen, L.K., and Stahl, D.A. (1993) Monitoring the enrichment and isolation of sulfate-reducing bacteria by using oligonucleotide hybridization probes designed from environmentally derived 16S rRNA sequences. *Appl Environ Microbiol* **59:** 682–686.

Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33:** 511–518.

Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., *et al.* (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* **14:** 169–181.

Legrand, L., Tap, J., Gauthey, C., Doré, J., Caron, C., and Leclerc, M. (2008) Rapid OTU: a fast pipeline to analyze 16S rDNA sequences by alignment or tetranucieotide frequency. *Proc. Gut Microbiome Symp. 2008 6th Congr. INRA Rowett Res. Inst.*, poster 26, pp. 35.

Ley, R.E., Backhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D., and Gordon, J.I. (2005) Obesity alters gut microbial ecology. *Proc Natl Acad Sci USA* **102:** 11070–11075.

Ley, R.E., Turnbaugh, P.J., Klein, S., and Gordon, J.I. (2006) Microbial ecology: Human gut microbes associated with obesity. *Nature* **444:** 1022.

Li, K.B. (2003) ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics* **19:** 1585–1586.

Li, M., Wang, B., Zhang, M., Rantalainen, M., Wang, S., Zhou, H., *et al.* (2008) Symbiotic gut microbes modulate human metabolic phenotypes. *Proc Natl Acad Sci USA* **105:** 2117–2122.

Macfarlane, G.T., and Gibson, G.R. (1994) Metabolic activities of normal colonic flora. In *Human Health: The Contribution of Microorganisms*. Gibson, S.A.W. (ed.). London, UK: Springer Verlag, pp. 17–52.

Manichanh, C., Rigottier-Gois, L., Bonnaud, E., Gloux, K., Pelletier, E., Frangeul, L., *et al.* (2006) Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55:** 205–211.

Marteau, P., Lepage, P., Mangin, I., Suau, A., Dore, J., Pochart, P., and Seksik, P. (2004) Gut flora and inflammatory bowel disease. *Aliment Pharmacol Ther* **20** (Suppl. 4)**:** 18–23.

Miller, T.L., Wolin, M.J., de Macario, E.C., and Macario, A.J.

(1982) Isolation of *Methanobrevibacter smithii* from human feces. *Appl Environ Microbiol* **43:** 227–232.
Rudi, K., Zimonja, M., Kvenshagen, B., Rugtveit, J., Midtvedt, T., and Eggesbo, M. (2007) Alignment-independent comparisons of human gastrointestinal tract microbial communities in a multidimensional 16S rRNA gene evolutionary space. *Appl Environ Microbiol* **73:** 2727–2734.
Savage, D.C. (1977) Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol* **31:** 107–133.
Schloss, P.D., and Handelsman, J. (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* **71:** 1501–1506.
Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermudez-Humaran, L.G., Gratadoux, J.J., *et al.* (2008) *Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc Natl Acad Sci USA* **20:** 20.
Stackebrandt, E., and Goebel, B.M. (1994) Taxonomic note: a place for DNA–DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Bacteriol* **44:** 846–849.
Suau, A., Bonnet, R., Sutren, M., Godon, J.J., Gibson, G.R., Collins, M.D., and Dore, J. (1999) Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Appl Environ Microbiol* **65:** 4799–4807.
Sutren, M., Michel, C., de la Cochetière, M.F., Bernalier, A., Wils, D., Saniez, M.H., and Doré, J. (2000) Temporal temperature gradient gel electrophoresis (TTGE) is an appropriate tool to assess dynamics of species diversity of the human fecal flora. *Reprod Nutr Dev* **40:** 176.
Swidsinski, A., Weber, J., Loening-Baucke, V., Hale, L.P., and Lochs, H. (2005) Spatial organization and composition of the mucosal flora in patients with inflammatory bowel disease. *J Clin Microbiol* **43:** 3380–3389.
Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., and Glockner, F.O. (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5:** 163.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673–4680.
Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. (2007) The human microbiome project. *Nature* **449:** 804–810.
Wilson, E.B. (1927) Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc* **22:** 209–212.
Wilson, K.H., Ikeda, J.S., and Blitchington, R.B. (1997) Phylogenetic placement of community members of human colonic biota. *Clin Infect Dis* **25:** S114–S116.
Woese, C.R., Fox, G.E., Zablen, L., Uchida, T., Bonen, L., Pechman, K., *et al.* (1975) Conservation of primary structure in 16S ribosomal RNA. *Nature* **254:** 83–86.
Woese, C.R., Kandler, O., and Wheelis, M.L. (1990) Towards a natural system of organisms: proposal for the domains *Archaea*, *Bacteria*, and *Eucarya*. *Proc Natl Acad Sci USA* **87:** 4576–4579.
Zoetendal, E.G., Akkermans, A.D., and De Vos, W.M. (1998) Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria. *Appl Environ Microbiol* **64:** 3854–3859.
Zoetendal, E.G., Collier, C.T., Koike, S., Mackie, R.I., and Gaskins, H.R. (2004) Molecular ecological analysis of the gastrointestinal microbiota: a review. *J Nutr* **134:** 465–472.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Rarefaction curves of operational taxonomic unit (OTU) detection per sample. Operational taxonomic units were defined with 2% dissimilarity cut-off, for homogeneous sequences of 1042 bases from nucleotides 350–1392 (*E. coli* 16S rRNA gene numbering) and fully aligned on 1317 bases including gaps.
Fig. S2. Phylogenetic core based on statistical model. Each fraction corresponded to an OTU that is part (%) of the phylogenetic core. Ten OTUs were highlighted because of their occurrence in the phylogenetic core.
Fig. S3. Venn diagram representation of 10 456 sequences set (A) and the 3180 OTUs (B) hits against external libraries. Four-way Venn diagrams were plotted with VENNY (http://bioinfogp.cnb.csic.es/tools/venny/index.html). BLASTN algorithm was used to determine the OTU occurrence in external libraries with a minimum coverage of 900 bases pairs and a minimum pairwise identity of 98%. A total of 550 OTUs (6676 sequences) were found in other 16S rRNA libraries; 2630 OTUs (3780 sequences) were specific to this study.
Table S1. Characteristics of human faecal samples studied, DNA concentration, total sequences, unambiguous sequences and sequences accession number per individual.
Table S2. Number of OTUs and estimated richness assessed on the complete sequences data set according to the alignment or tetranucleotide frequency algorithms.
Table S3. Analysis of molecular variance (AMOVA) between omnivorous and vegetarian diets.
Table S4. Quantitative PCR assays on 16 healthy human faecal samples.
Table S5. 16S rDNA sequence similarity between core OTU representative and sequences from isolated strains.
Table S6. Probability estimation and confidence interval for each OTU in the core to be part of the microbiota.
Appendix S1. Statistical detection of a putative phylogenetic core.

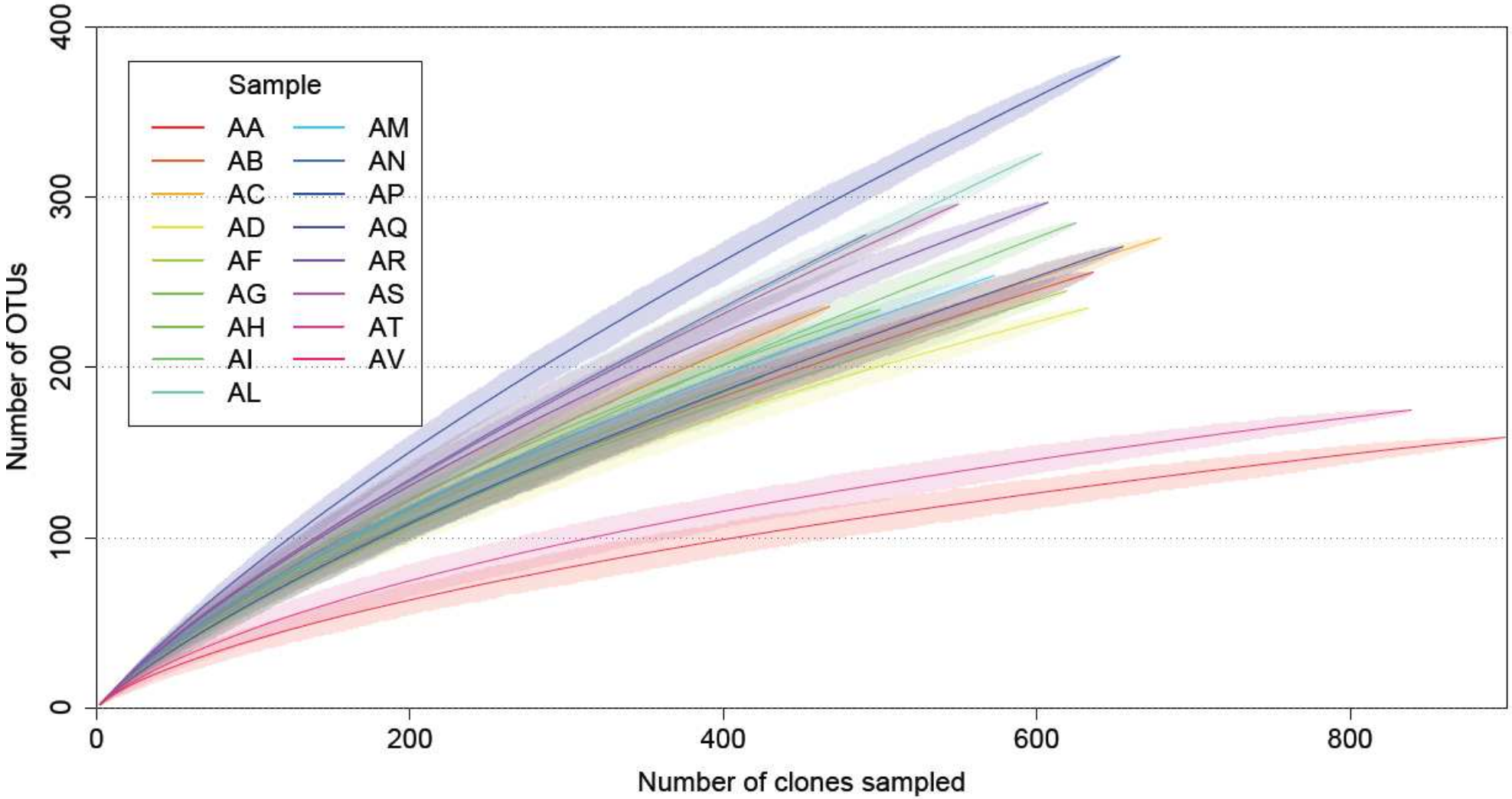
Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

SUPPLEMENTARY TABLES AND FIGURES

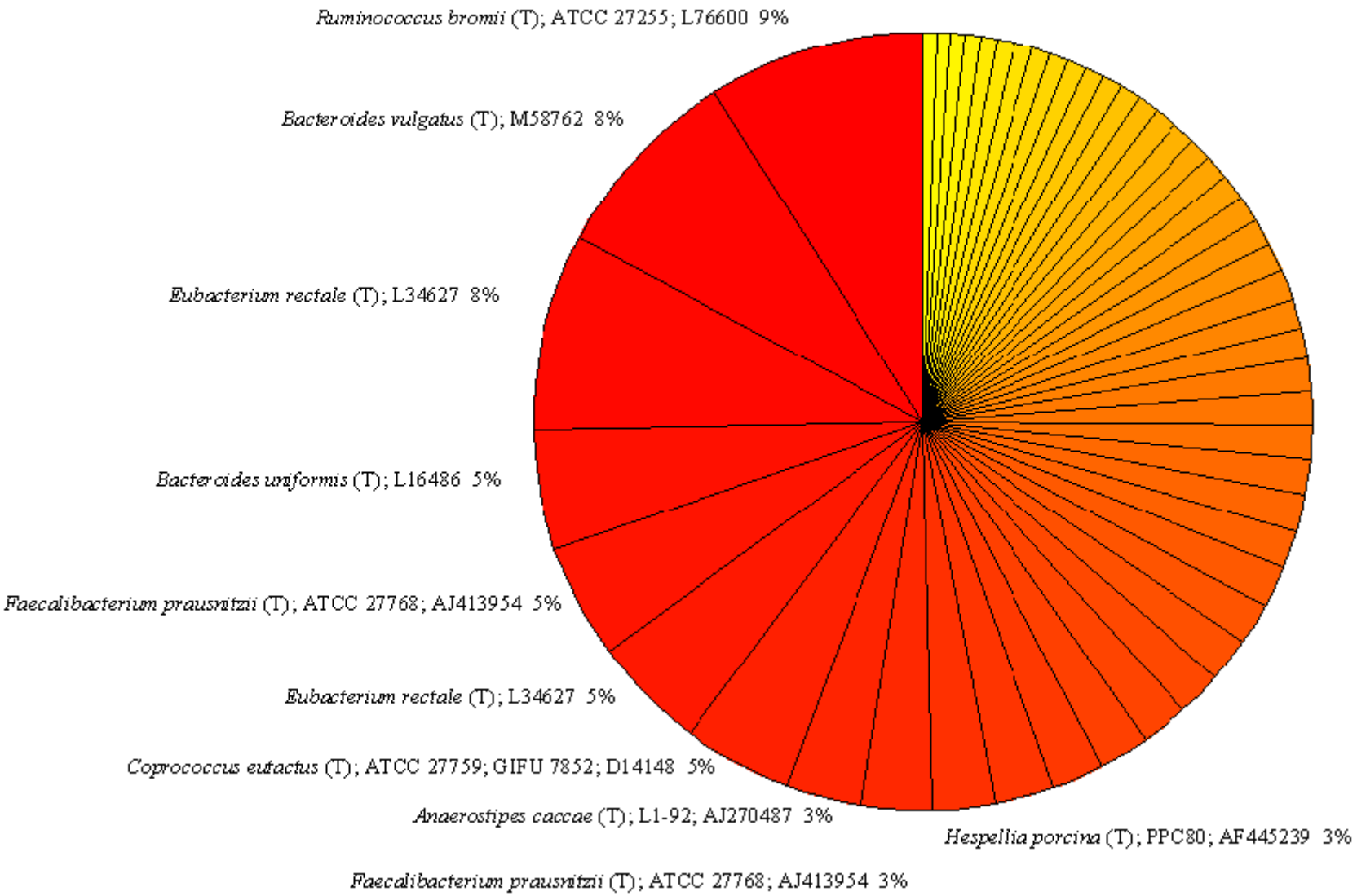
Towards the Human intestinal microbiota phylogenetic core

J. TAP[1], S. MONDOT[1], F. LEVENEZ[1], E. PELLETIER[2,3], C. CARON[4], J.-P. FURET[1], E. UGARTE[2,3], R. MUÑOZ-TAMAYO[1, 5, 7], D. LE PASLIER[2,3], R. NALIN[6], J. DORE[1#] and M. LECLERC[1*#]

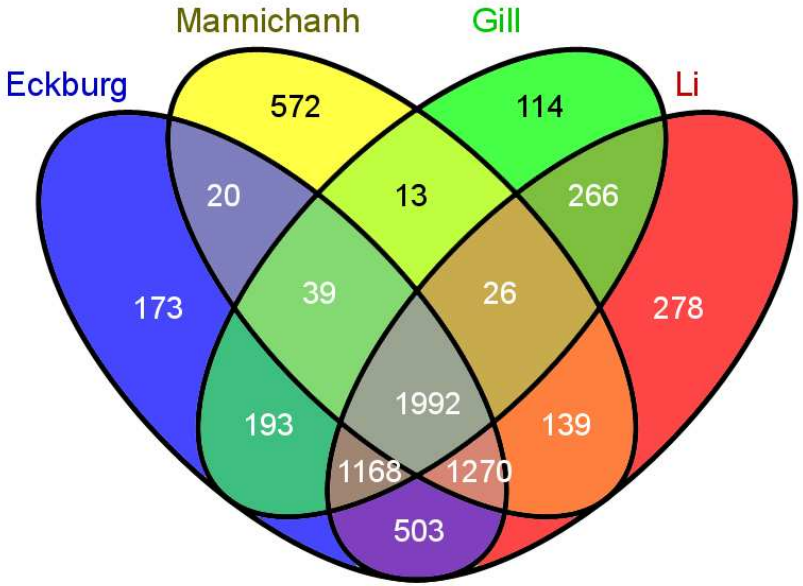
[1]INRA, UEPSD, UR910, 78350 Jouy en Josas, France, [2]CEA, DSV, IG, Genoscope, 91057 Evry, France, [3]CNRS UMR 8030, 91057 Evry, France, [4]INRA, MIG, UR1077, 78350 Jouy en Josas, France, [5]INRA, MIA, UR341, 78350 Jouy en Josas, France, [6]Libragen 31400 Toulouse, France, [7]UMR8506 Univ Paris Sud-CNRS-SUPÉLEC, L2S, 91190 Gif sur Yvette, France



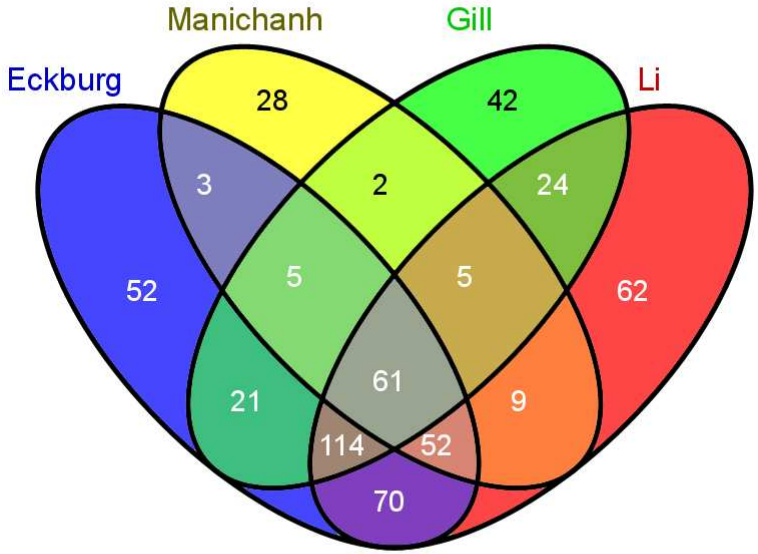
Supplementary Figure S1. Rarefaction curves of OTU detection per sample.
OTUs were defined with 2% dissimilarity cut off, for homogeneous sequences of 1,042 bases from nucleotides 350 to 1,392 (*E. coli* 16S rRNA gene numbering) and fully aligned on 1,317 bases including gaps.



Supplementary Figure S2. Phylogenetic core based on statistical model.
Each fraction corresponded to an OTU that is part (%) of the phylogenetic core. Ten OTUs were highlighted because of their occurrence in the phylogenetic core.



(a) Occurrence of 10,456 total sequences in external libraries



(b) Occurrence of 3,180 total OTUs in external libraries

Supplementary Figure S3. Venn diagram representation of 10,456 sequences set (a) and the 3,180 OTUs (b) hits against external libraries.
4 way Venn diagrams were plotted with VENNY (http://bioinfogp.cnb.csic.es/tools/venny/index.html). BLASTN algorithm was used to determine the OTU occurrence in external libraries with a minimum coverage of 900 bases pairs and a minimum pairwise identity of 98%. 550 OTUs (6,676 sequences) were found in other 16S rRNA libraries, 2,630 OTUs (3,780 sequences) were specific to this study.

Supplementary Table S1. Characteristics of human fecal samples studied, DNA concentration, total sequences, unambiguous sequences and sequences accession number per individual.

Sample	*Diet orientation*	*Country*	*DNA (ng/µL)*	*Number of sequences*	*Number of unambiguous sequences*	*Sequences accession number*
AA	Omnivorous	France	648.71	925	636	FP079445: FP080080
AB	Omnivorous	France	1,009.01	769	468	FP084892: FP085359
AC	Omnivorous	France	418.42	972	679	FP084213: FP084891
AD	Omnivorous	France	350.71	905	633	FP078812: FP079444
AF	Omnivorous	France	418.14	943	619	FP078193: FP078811
AG	Omnivorous	France	693.74	823	500	FP077693: FP078192
AH	Omnivorous	France	495.12	711	426	FP077267: FP077692
AI	Omnivorous	France	504.05	903	625	FP076642: FP077266
AL	Vegetarian	Netherlands	854.98	1,050	603	FP083610: FP084212
AM	Vegetarian	Netherlands	540.27	942	573	FP083037: FP083609
AN	Vegetarian	Netherlands	571.29	844	491	FP082546: FP083036
AP	Vegetarian	France	566.11	1,000	653	FP081893: FP082545
AQ	Vegetarian	Netherlands	661.31	997	655	FP081238: FP081892
AR	Vegetarian	Netherlands	810.19	842	607	FP080631: FP081237
AS	Vegetarian	Netherlands	723.04	915	550	FP080081: FP080630
AT	Omnivorous	France	320.82	990	839	FP075803: FP076641
AV	Vegetarian	France	nd	1,001	899	FP074904: FP075802

Compared to non PCR metagenomic datasets, the number of PCR linked chimera in this study was high, and consistent with data from 16S rRNA mammal fecal microbiota (Ley et al., 2008) using the same tools (Huber et al., 2004; Ashelford et al., 2006)

Supplementary Table S2. Number of OTUs and estimated richness assessed on the complete sequences dataset according to the alignment or tetranucleotide frequency algorithms.

	Algorithms (OTU cut off)		
	MAFFT (2%)	*ClustalW (2%)*	*Tetranucleotide (10%)*
Number of OTUs	3,186	3,180	3,097
Estimated richness (Chao1)	9,912.5	9,940.9	8,776.2
95% Confidence intervals	(9,089.5 – 10,850.3)	(9,111.7 – 10,885.8)	(8,075.7 – 9,575.3)

Supplementary Table S3. Analysis of molecular variance (AMOVA) between omnivorous and vegetarian diets.
AMOVA was computed with ade-4 package in R according to individuals diet, both OTUs diversity and abundance within individuals.
Most of the diversity in term of phylotypes was found within each individual microbiota (97.7%). Few phylotypes diversity separated diets (omnivorous and vegetarian). The difference between the microbiota associated to a diet was low (2.085%). All components of the variance were supported by p-value <0.05. This means that the genetic diversity measured within the microbiota associated to the vegetarian and omnivorous diet did not significantly differ from the total diversity.

Variance component	*Standard deviation*	*Total*	*% total*	*p-values*	*phi-statistic*
Between diets	0.0006134376	0.001797386	0.017%	0.03	0.001797386
Between samples within diet	0.0071172981	0.020853842	2.085%	0.01	0.020891392
Within each sample	0.3335635951	0.977348772	97.734%	<0.0001	0.022651228

Supplementary Table S4. Quantitative PCR assays on 16 healthy human fecal samples.
Results were expressed in log (10) equivalent genome per gram of stool. All q-PCR assays were performed according to the methods previously described in "Furet et al., Comparative assessment of human and farm animal faecal microbiota using real-time quantitative PCR". FEMS Microbial Ecology, under review.

Sample	*All bacteria*	*C. leptum* group	*C. coccoides* group	*Bacteroides Prevotella*	*E. coli*	*F. prausnitzii*	*Lactobacillus Leuconostoc*	*Bifidobacterium*
AA	11.66	11.39	10.12	10.78	7.02	11.00	9.13	9.20
AB	11.74	11.73	10.78	10.94	6.92	11.35	8.86	9.81
AC	11.42	11.10	9.95	10.80	7.10	10.92	8.71	9.30
AF	11.08	11.00	9.84	9.88	6.33	10.03	8.52	9.95
AD	11.35	11.41	10.69	10.70	7.39	10.82	8.49	9.23
AG	11.55	11.21	10.08	10.58	7.12	11.11	9.42	10.00
AH	11.51	11.50	10.37	10.65	6.83	10.68	9.66	9.93
AI	10.90	10.84	9.45	10.04	7.52	9.81	8.18	7.64
AL	11.42	11.41	9.92	10.42	9.06	10.73	9.03	9.20
AM	10.88	11.18	10.25	10.23	6.08	10.69	8.77	9.42
AN	11.55	11.54	9.95	10.49	7.27	10.91	9.49	9.67
AP	11.51	11.34	10.76	10.28	6.89	10.70	8.53	9.74
AQ	11.32	11.18	10.25	10.23	6.08	10.69	8.77	9.42
AR	11.22	11.09	10.19	9.48	6.80	10.29	9.00	9.60
AS	11.69	11.64	10.78	10.71	7.19	11.09	9.12	10.28
AT	10.72	9.66	8.92	10.48	7.23	5.86	8.47	9.03

Supplementary Table S5: 16S rDNA sequence similarity between core OTU representative and sequences from isolated strains.

Core sequences accession number	*Isolated type strain name and GenBank accession number*	*% Identity*
UUAA3AD121	*Roseburia intestinalis ; AJ312385*	99.9
UUAA2BD011	*Dorea longicatena ; AJ132842*	99.8
UUAA2DG121	*Bacteroides massiliensis ; AY126616*	99.8
UUAC2BE101	*Alistipes shahii ; AY974072*	99.6
UUAQ3CG061	*Clostridium sp. BI-114 ; AJ518869*	99.5
UUAM4AA081	*Faecalibacterium prausnitzii ; AJ413954*	99.3
UUAP4BG111	*Ruminococcus obeum ; L76601*	99.1
UUAT1DA021	*Bacteroides vulgatus ; M58762*	99.1
UUAM4BF021	*Dorea longicatena ; AJ132842*	99.0
UUAA2DE021	*Bifidobacterium longum biovar Longum ; M58739*	99.0
UUAN1BA031	*Faecalibacterium prausnitzii ; AJ413954*	98.8
UUAS3CG111	*Collinsella aerofaciens ; AB011816*	98.8
UUAV1AD121	*Eubacterium eligens ; L34420*	98.7
UUAR2CE041	*Dorea formicigenerans ; L34619*	98.7
UUAN2BH021	*Parabacteroides distasonis ; M86695*	98.7
UUAB1AH091	*Ruminococcus obeum ; L76601*	98.6
UUAA3BF041	*Eubacterium rectale ; L34627*	98.6
UUAV3CB121	*Eubacterium rectale ; L34627*	98.6
UUAQ1CF071	*Eubacterium hallii ; L34621*	98.6
UUAV1AC021	*Ruminococcus bromii ; L76600*	98.5
UUAR3CH081	*Eubacterium rectale ; L34627*	98.4
UUAG3CC101	*Alistipes putredinis ; L16497*	98.2
UUAV3AH011	*Bacteroides stercoris ; X83953*	98.1
UUAN3DG101	*Eubacterium hallii ; L34621*	97.9
UUAH1AG021	*Faecalibacterium prausnitzii ; AJ413954*	97.9
UUAV1AH031	*Ruminococcus obeum ; L76601*	97.8
UUAB1CG021	*Faecalibacterium prausnitzii ; AJ413954*	97.7
UUAI2CC011	*Eubacterium ramulus ; L34623*	97.7
UUAP2DD111	*Roseburia intestinalis ; AJ312385*	97.5
UUAF1CG121	*Faecalibacterium prausnitzii ; AJ413954*	97.4
UUAB3DA091	*Faecalibacterium prausnitzii ; AJ413954*	97.3
UUAL2DE081	*Faecalibacterium prausnitzii ; AJ413954*	97.3
UUAA3BE081	*Clostridium sp. BI-114; AJ518869*	97.2
UUAR2DC041	*Eubacterium rectale ; L34627*	97.2
UUAL1BD011	*Roseburia intestinalis ; AJ312385*	97.1
UUAP2DD031	*Ruminococcus luti ; AJ133124*	97.0
UUAH2DG041	*Roseburia intestinalis ; AJ312385*	97.0
UUAF2AD061	*Ruminococcus luti ; AJ133124*	96.7
UUAG1DB011	*Clostridium nexile ; X73443*	96.6
UUAV2AF071	*Ruminococcus luti ; AJ133124*	96.5
UUAI1CC071	*Bacteroides uniformis ; L16486*	96.5
UUAD3AE091	*Ruminococcus schinkii ; X94965*	96.2
UUAD1CH101	*Lachnospira pectinoschiza ; L14675*	96.1
UUAN2CG021	*Coprococcus eutactus ; D14148*	95.8
UUAV2AC071	*Ruminococcus obeum ; L76601*	95.8
UUAQ2AA061	*Clostridium sp. BI-114; AJ518869*	95.6
UUAP2BF071	*Ruminococcus obeum; L76601*	95.6
UUAH3CB081	*Ruminococcus lactaris ; L76602*	95.6
UUAC1DG011	*Anaerostipes caccae ; AJ270487*	95.5
UUAI2CH081	*Clostridium bolteae ; AJ508452*	95.3
UUAC2AB111	*Roseburia intestinalis ; AJ312385*	95.3
UUAV3DE061	*Clostridium bolteae ; AJ508452*	95.0
UUAI4DF111	*Ruminococcus obeum ; L76601*	94.9
UUAR1CD071	*Clostridium xylanolyticum ; X71855*	94.8
UUAR1DD061	*Roseburia intestinalis ; AJ312385*	94.8
UUAF2DB091	*Roseburia intestinalis ; AJ312385*	94.4
UUAQ3CH071	*Sporobacter termitidis ; Z49863*	94.0
UUAG1AF041	*Clostridium clostridioforme ; M59089*	93.5
UUAH1CG031	*Bacteroides capillosus ; AY136666*	93.4
UUAM4AA121	*Bacteroides capillosus ; AY136666*	93.2
UUAR2AC031	*Clostridium spiroforme ; X75908*	93.2
UUAS3BH081	*Hespellia porcina ; AF445239*	92.6
UUAL1AA081	*Clostridium orbiscindens ; Y18187*	92.4
UUAV2CB111	*Clostridium amygdalinum ; AY353957*	91.8
UUAV3CH011	*Clostridium straminisolvens ; AB125279*	89.9
UUAC3CA091	*Dehalobacter restrictus ; U84497*	87.9

Supplementary Table S6. Probability estimation and confidence interval for each OTU in the core to be part of the microbiota

Summary results of the probability estimation and core representation with respect to the normalized abundance. The phylogenetic core was defined as the subset of OTUs found in at least 50 % of the individuals. The OTUs are organized in descending order with respect to the estimated probability to be present in the individuals. Confidence intervals of the probability estimates and the normalized abundances of the OTUs are shown.

The table for each OTU of the phylogenetic core describes their proportion p_j in the core and the probability $\hat{\gamma}_j$ with confidence interval to being part of the microbiota.

OTU	p_j (%)	$\hat{\gamma}_j$	Confidence interval γ_j	OTU	p_j (%)	$\hat{\gamma}_j$	Interval confidence γ_j
UUAB3DA091	4.76	0.94	[0.73 ; 0.99]	UUAF2DB091	0.80	0.59	[0.36 ; 0.79]
UUAC1DG011	3.11	0.88	[0.65 ; 0.97]	UUAG1AF041	1.84	0.59	[0.36 ; 0.79]
UUAR2AC031	0.95	0.88	[0.65 ; 0.97]	UUAG1DB011	1.39	0.59	[0.36 ; 0.79]
UUAI1CC071	4.97	0.82	[0.59 ; 0.94]	UUAG3CC101	1.53	0.59	[0.36 ; 0.79]
UUAA2BD011	1.23	0.76	[0.52 ; 0.90]	UUAI2CC011	0.77	0.59	[0.36 ; 0.79]
UUAA2DE021	0.86	0.76	[0.52 ; 0.90]	UUAM4AA081	1.15	0.59	[0.36 ; 0.79]
UUAA3BE081	2.17	0.76	[0.52 ; 0.90]	UUAM4AA121	1.22	0.59	[0.36 ; 0.79]
UUAH1AG021	1.95	0.76	[0.52 ; 0.90]	UUAP2BF071	0.80	0.59	[0.36 ; 0.79]
UUAI2CH081	2.41	0.76	[0.52 ; 0.90]	UUAP2DD031	1.11	0.59	[0.36 ; 0.79]
UUAQ3CG061	1.90	0.76	[0.52 ; 0.90]	UUAV2CB111	0.56	0.59	[0.36 ; 0.79]
UUAD1CH101	1.22	0.71	[0.47 ; 0.87]	UUAV3CH011	0.90	0.59	[0.36 ; 0.79]
UUAL2DE081	2.96	0.71	[0.47 ; 0.87]	UUAB1AH091	0.73	0.53	[0.31 ; 0.74]
UUAP2DD111	2.12	0.71	[0.47 ; 0.87]	UUAF2AD061	1.70	0.53	[0.31 ; 0.74]
UUAP4BG111	0.55	0.71	[0.47 ; 0.87]	UUAH1CG031	1.04	0.53	[0.31 ; 0.74]
UUAQ1CF071	1.39	0.71	[0.47 ; 0.87]	UUAL1AA081	0.58	0.53	[0.31 ; 0.74]
UUAS3BH081	2.61	0.71	[0.47 ; 0.87]	UUAN2CG021	4.52	0.53	[0.31 ; 0.74]
UUAT1DA021	8.18	0.71	[0.47 ; 0.87]	UUAN3DG101	0.93	0.53	[0.31 ; 0.74]
UUAV2AC071	1.04	0.71	[0.47 ; 0.87]	UUAQ2AA061	0.97	0.53	[0.31 ; 0.74]
UUAA3BF041	8.16	0.65	[0.42 ; 0.83]	UUAR1CD071	0.77	0.53	[0.31 ; 0.74]
UUAB1CG021	1.58	0.65	[0.42 ; 0.83]	UUAR2DC041	0.66	0.53	[0.31 ; 0.74]
UUAF1CG121	1.42	0.65	[0.42 ; 0.83]	UUAS3CG111	0.73	0.53	[0.31 ; 0.74]
UUAR2CE041	0.85	0.65	[0.42 ; 0.83]	UUAV1AC021	9.01	0.53	[0.31 ; 0.74]
UUAR3CH081	1.23	0.65	[0.42 ; 0.83]	UUAV1AD121	1.47	0.53	[0.31 ; 0.74]
UUAV3CB121	4.74	0.65	[0.42 ; 0.83]	UUAV3DE061	0.70	0.53	[0.31 ; 0.74]
UUAD3AE091	1.74	0.59	[0.36 ; 0.79]				

Supporting text. Statistical detection of a putative phylogenetic core.

Under the hypothesis that there was not dependence between the individuals, a statistical model was used to define a putative phylogenetic core. The way to represent the presence/absence of the OTUs was to set a binomial distribution on the prevalence.

Let γ_j denote the probability that the OTU *j* is detected in an individual. The probability that OTU *j* is detected in *k* individuals (*k=0,...,n,* where *n*=17 in the case of study) is equal to:

$$\binom{n}{k} \gamma_j^{\,k} \left(1-\gamma_j\right)^{n-k}. \tag{1}$$

The probability γ_j was estimated by the proportion of individuals where the OTU *j* was detected:

$$\hat{\gamma}_j = \frac{k_j}{n}, \tag{2}$$

where k_j is the number of individuals where OTU *j* was detected. In order to assess the uncertainty related to the sample size (17 individuals), confidence intervals were computed by the method proposed by Edwin B. Wilson (1927) and discussed by Agresti and Coull (1998).
The confidence interval has the form:

$$\frac{\hat{\gamma}_j + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2}\sqrt{\left[\hat{\gamma}_j\left(1-\hat{\gamma}_j\right) + z_{\alpha/2}^2/4n\right]/n}}{1+z_{\alpha/2}^2/n} \tag{3}$$

where $z_{\alpha/2}$ denote the quantile of the standard normal distribution. It was set to a value of 1.96 to correspond to a 95% interval confidence.

The core was defined as the subset of OTUs found in at least 50 % of the individuals.

The parameters γ_j do not provide any information about the abundance of the OTUs in the global core. In order to have a representation of the core in terms of abundance, the numbers of sequences of each OTU were averaged on the subset of individuals where the OTU was detected. Afterwards, the average abundances were normalized to have a unitary representation of the core. In such a way, if p_j is the normalized abundance of the OTU *j* in the core set, then $\sum p_j = 1$.

References

Agresti, A., and Coull, B.A. (1998) Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician* **52**: 119-125.
Wilson, E.B. (1927) Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**: 209-212.


changes
intake
surgery
groups
performed
composition
mean
markers
Statistical
proportions
fat
lean
Bacteroides/Prevotella
energy
gastric
qPCR
increased
al
counts
compared
model
levels
association
values
Paris
de
species
biological
corpulence
concentrations
samples
BMI
time
role
leptin
bacterial
faecal
different
RS
coli
showed
associations
considered
Multivariate
HOMA-IR
inflammation
calorie
independent
state
human
lower
weight
found
RYGB
diabetic
Serum
bacteria
fasting
correlated
Bacteroidetes
France
relationships
type
correlation
improvement
glucose
parameters
clinical
LME
non-diabetic
relationship
proportion
inflammatory
food
post-RYGB
total
homeostasis
Figure
Firmicutes
mice
assessment
obese
gut
M6
prausnitzii
SEM
PCR
also
significant
metabolic
OB/nD
individuals
related
control
observed
subjects
level
using
log
blood
sensitivity
supplementary
M0
population
data
body
analysis
diabetes
resistance
number
hsCRP
group
M3
IL-6
loss
positive
bypass
mmol/l
microbiota
major
NS
months
OB/D
negative
Bifidobacterium
plasma
adiponectin
expressed
mass
Obesity
Table
shown
increase
patients
results
leptum
controls
associated
studies
adipose
orosomucoid
reduction
decreased
significantly
genus
low-grade
populations
insulin
study

Differential adaptation of human gut microbiota to bariatric surgery-induced weight loss: links with metabolic and low grade inflammation markers

Jean-Pierre Furet [1a*], Lingchun Kong [2,3a], Julien Tap [1], Christine Poitou-Berner [2,3], Arnaud Basdevant [2,3], Jean-Luc Bouillot [4], Denis Mariat [1], Gérard Corthier [1], Joël Doré [1], Corneliu Henegar [2], Salwa Rizkalla [2,3], Karine Clément [2,3*].

1 INRA, U910, Unité d'Ecologie et de Physiologie du Système Digestif, F-78350 Jouy-en-Josas, France

2 Assistance Publique-Hôpitaux de Paris, Hôpital Pitié-Salpêtrière, Département de Nutrition et d'Endocrinologie, Paris, F-75013 France; Centre de Recherche Nutrition Humaine, Ile de France, Paris, F-75013 France.

3 INSERM, U872, équipe 7 Nutriomique, Paris, F-75006 France; Université Pierre et Marie Curie-Paris6, Centre de Recherche des Cordeliers, UMR S 872, Paris, F-75006 France.

4 Assistance Publique Hôpitaux de Paris, Département de chirurgie, Hôpital Hôtel-Dieu, Paris, F-75004 France.

[a] authors equally contributed.

* Corresponding authors

Karine Clément: Endocrinology department and INSERM U872 nutriomic team, Pitié-Salpêtrière hospital, boulevard de l'hôpital 75013, Paris, France. Tel 33 (0) 1 4234 7919. Mail : karine.clement@psl.aphp.fr

Jean-Pierre Furet: Unité d'Ecologie et de Physiologie du Système Digestif, 78350 Jouy-en-Josas, France. Tel 33 (0) 1 3465 2929. Mail: jean-pierre.furet@jouy.inra.fr

Running Title: Gut microbiota profile in RYGB induced weight loss

Abbreviations

BMI: body mass index, DXA: Dual-energy X-ray Absorptiometry, HOMA-IR: homeostasis model assessment of insulin resistance, HOMA-B%: homeostasis model assessment to estimate pancreatic insulin secretion, HOMA-S%: homeostasis model assessment to estimate

insulin sensitivity, hsCRP: high sensitive C-reactive protein, IL-6: interleukin-6, LME: linear mixed-effects, OB/nD: non-diabetic obese subjects, OB/D: obese diabetic subjects, PLS-DA: partial least squares discriminant analysis, qPCR: real-time quantitative PCR, REE: resting energy expenditure, RYGB: Roux-en-Y gastric bypass.

Keywords

Obesity, bariatric surgery Roux-en-Y gastric bypass, faecal microbiota, inflammation, insulin resistance, diabetes.

Abstract

Background: Obesity alters gut microbiota ecology and is associated with low-grade inflammation in humans. Diet induced weight loss result in changes in gut microbiota. Roux-en-Y gastric bypass (RYGB) surgery is one of the most efficient procedures for the treatment of morbid obesity resulting in drastic weight loss and improvement of metabolic and inflammatory status. The aim of this study was to analyze the impact of RYGB on the gut microbiota signature and to identify putative links between changes in gut bacteria, metabolic and inflammatory parameters associated with this procedure.

Methods: Gut microbiota was profiled from faecal samples by qPCR in 13 lean controls and in 30 obese individuals (7 presenting with type 2 diabetes) measured before RYGB and 3 and 6 months after RYGB. Microbiota profile changes were compared against the dynamic evolution of a series of clinical phenotypes: body weight, fat mass and fat free mass, resting energy expenditure and blood parameters including adipokines (leptin, adiponectin), estimates for blood glucose homeostasis (glucose and insulin) and inflammatory markers (hsCRP, interleukin-6 and orosomucoid) .

Findings: As expected, surgery resulted in major weight loss, reduction in food intake and improvement in insulin sensitivity and inflammation. Analysis highlighted four major findings: (i) the level of the *Bacteroides*/*Prevotella* group was lower in obese subjects than in controls at baseline and increased 3 months after RYGB. This was negatively correlated with obesity-related parameters, and was highly dependent on caloric intake, (ii) *Escherichia coli* species levels increased 3 months after RYGB and were inversely correlated with fat mass and leptin levels, but were independently of food intake, iii) lactic acid bacteria including the *Lactobacillus/Leuconostoc/Pediococcus* group and *Bifidobacterium* genus decreased 3 months after RYGB, (iv) *Faecalibacterium prausnitzii* species was lower in diabetics and

strongly associated with inflammatory markers at baseline and throughout the follow-up after surgery, independent of changes in food intake.

Conclusions: Several dominant components of the gut microbiota adapt rapidly to the starvation-like situation induced by RYGB while the species *F. prausnitzii* appears linked to the improvement in low grade inflammation state in obesity and diabetes, independent of calorie intake. This study paves the way for future investigations exploring the contribution of *F. prausnitzii* in low-grade inflammation and insulin resistance.

Introduction

Obesity is characterized by increased fat mass accumulation leading to adverse health consequences via the development of a myriad of co-morbidities including metabolic and cardiovascular diseases. Even though some but not all environmental factors have been elucidated, the increasing epidemic of obesity appears virtually impossible to control and the mechanisms associated with fat mass expansion need to be urgently identified. Obesity is considered as a low-grade inflammatory disease with the associated adipose tissue contributing to this state via the secretion of inflammatory molecules capable of altering metabolic homeostasis [1, 2]. Numerous environmental, behavioral, genetic and biological factors interact to cause human obesity. A novel actor in human obesity, and its associated metabolic risks, was recently to be the commensal microbiota of the intestine [3, 4].

A role for the intestinal microbiota in harvesting energy from food [5] and in the regulation of body fat storage [6] was proposed in rodents where germ-free mice colonized by microbiota increase their body-fat and developed insulin-resistance in spite of a 30% decrease in food intake. These changes were associated with a dysbiosis in the obese mice, faecal microbiota characterized by an increased representation of the Firmicutes phylum and a reduced representation of the Bacteroidetes phylum [7]. Other reports have proposed a contribution of the gut microbiota-produced lipopolysaccharides (LPS) to inflammation and the development of metabolic syndrome in rodents [8-10]**.** Interestingly, population studies have shown that increased endotoxemia (*i.e.* circulating LPS) is associated with increased fat consumption contained in a meal [11]. Moreover the comparison of the microbiota composition between lean and obese human subjects and its evolution in obese patients losing weight throughout low calorie diets indicated a diminished proportion of Bacteroidetes and increased Firmicutes in obese subjects that trended to that of lean controls at the end of the dietary intervention [12]. The rebalancing of this Firmicutes/Bacteroidetes ratio is known to improve

inflammation and mitigate metabolic risks. However modification of the Firmicutes/Bacteroidetes ratio observed in obese individuals was not confirmed in other human studies [13]. The dynamics of the microbiota composition and the bacterial fermentation products produced in obese patients upon low calorie intervention showed a reduced concentration of butyrate. This was associated, in terms of microbiota composition, with lower proportions of the *Roseburia-Eubacterium rectale* subgroup (Firmicutes phylum) and the *Bifidobacterium* genus as obese patients were losing weight [14]. These bacterial changes were not associated with the improvement of metabolic or inflammatory phenotypes associated with weight modification over time.

These observations raise the question of the relation between the gut microbiota and the low-grade inflammation status in obesity. Bariatric gastric surgery (RYGB) is an increasingly effective model to study in this context. RYGB leads to major improvements in blood glucose, lipids, insulin levels, insulin sensitivity and hormonal responses [15], as well as in the decrease in inflammatory markers such as C-reactive protein, and interleukin-6 [16]. This procedure is one model to understand the molecular adaptations underlying the observed health benefits and the potential role of calorie restriction induced shift in gut microbiota pattern.

Our present work analyzed the microbiota profiles in the faeces of morbidly obese subjects before and after RYGB. We examined the association between gut microbiota changes and a range of body composition, metabolic and inflammatory markers. These results provide new insight regarding the microbiota changes in obese patients post-RYGB and highlight some bacterial groups as possible factors associated with changes in nutritional status and others with insulin sensitivity and inflammatory parameters.

Materials and Methods

Subjects

Thirty obese subjects (27 women and 3 men) enrolled in a bariatric surgery program were prospectively and consecutively recruited in 2008 in the Department of Nutrition, Center of Reference for Medical and Surgical Care of Obesity (Hôtel-Dieu and Pitié-Salpêtrière hospitals, Paris, France). The patients accepted the criteria for obesity surgery: Body Mass Index BMI $\geq$ 40 kg/m^2 or $\geq$ 35 kg/m^2 with at least two co-morbidities (hypertension, type-2 diabetes, dyslipidemia or obstructive sleep apnea syndrome). Preoperative evaluation included detailed medical history, including physical, nutritional, metabolic, cardiopulmonary, and psychological assessments.

The subject's weight was stable (variation of less than ± 2kg) for at least 3 months prior to surgery. Subjects did not demonstrate evidence of acute or chronic inflammatory diseases, infectious diseases, viral infection, cancer and/or known alcohol consumption (> 20g per day). No antibiotics were taken by the subjects before surgery or during the post-surgery follow-ups. For all patients, clinical and biological parameters were assessed prior to Roux-en-Y surgery (*i.e.* basal or M0) and at 3 months and 6 months post-surgery (M3 and M6, respectively). Oral glucose tolerance test (OGTT) was performed in the 23 non-diabetic subjects (referred to as the OB/nD group). All had a glycemia < 11 mM two hours after 75 g oral glucose. 7 subjects had type 2 diabetes (referred to as the OB/D subgroup), with a fasting glycemia over 7 mM and/or the use of an anti-diabetic drug. Two of these individuals necessitated insulin therapy while the other 5 subjects were treated with metformin and hypolipidemic drugs (either fibrate or statins). In addition, 13 normal- weight healthy women volunteers living in the same area as the obese subjects were recruited as a control group using the same clinical protocol. The Ethics Committee of the Hôtel-Dieu Hospital approved

the clinical protocol both for obese and non-obese individuals. All subjects gave a written informed consent.

Dietary Assessment

At each visit (M0, M3 and M6), caloric intake (kcal per day) and macronutrient portions were evaluated by a registered dietician during a one hour questioning. Multivitamins and iron supplements were provided to avoid deficiencies, which is a well-known secondary effect of bariatric surgery [17]. Serum iron, ferritin, the coefficient of saturation of iron in transferrin, vitamins (A, D, E, B1, B12 and B9), micronutrients (selenium and zinc), and calcium were measured using routine bio-clinical tests. The serum measurement of these parameters revealed them to be in the normal range at all time points in all subjects (data not shown).

Body composition, metabolic and inflammatory parameters

Adiposity markers were determined for all individuals before and after the surgery. BMI was calculated from the body weight and height. Fat-free mass and fat mass were determined by DXA (GE Lunar Prodigy Corporation, Madison, WI, USA). Except for the lean control group, resting energy expenditure (REE) was measured by indirect calorimetry after 12 hours fasting (Deltatrac, Datex, France). Fat and fat-free mass were presented as percentage of measured body weight (Table 1).

Periumbilical surgical biopsies of subcutaneous adipose tissues were obtained before gastric surgery from all subjects and the adipocyte diameter was measured as previously described [18].

Blood samples were obtained at each time point in all subjects after 12 hours fasting and stored at -20°C for later analysis. Biological parameters include: lipid (total cholesterol, HDL-cholesterol and triglycerides), insulin and glucose values, leptin, adiponectin and inflammatory markers (high sensitive C-reactive protein (hsCRP), interleukin-6 (IL-6) and orosomucoid). HOMA insulin resistance (homeostasis model assessment HOMA-IR) was

determined by a mathematical transformation of fasting blood glucose and insulin measurements [19]. HOMA-IR is a useful index when studying morbidly obese individuals in whom the evaluation of insulin sensitivity using the clamp technique has technical limitations due to extreme BMI.

Faecal samples

Faecal samples from 30 morbidly obese subjects (including 7 diabetics) and from 13 healthy controls were recovered at the time of inclusion (M0). During follow-ups, faecal samples were obtained for 26 subjects (including 6 diabetics) at M3 and for 15 subjects (including 5 diabetics) at M6. Finally we obtained a complete course of stool samples (M0, M3 and M6) for 10 individuals. Whole stools were collected in sterile boxes and stored at -20°C within 4 hours. Samples were frozen at -20°C as 200 mg aliquots and stored for further analysis.

DNA extraction from faecal samples

DNA was extracted from 200 mg aliquots of faeces as previously described [20]. After the final precipitation, DNA was resuspended in 150 mL of TE buffer and stored at -20°C prior to further analysis.

Oligonucleotide primers and probes

Primers and probes used in this study are presented in supplementary Table 1. TaqMan® qPCR was adapted to quantify total bacteria population in addition to the dominant (> 1% of faecal bacteria population) bacterial species *Clostridium leptum (C. leptum)*, *Clostridium coccoides (C. coccoides)*, *Bacteroides/Prevotella* and *Bifidobacterium*. Quantitative PCR using SYBR-Green® was performed for the *Lactobacillus/Leuconostoc/Pediococcus* group and for the sub-dominant bacterial species: *Escherichia Coli* (*E. coli)* [21] as well as for the *Faecalibacterium prausnitzii (F. prausnitzii)* [22]. The TaqMan® probes were synthesized by Applied-Biosystems Applera-France. Primers were purchased from MWG (MWG-Biotech AG, Ebersberg, Germany).

Real-time qPCR

Real-time qPCR was performed using an ABI 7000 Sequence Detection System with software version 1.2.3 (Applied-Biosystems, Foster City, Ca, USA). Amplification and detection were carried out in 96-well plates with TaqMan® Universal PCR 2× MasterMix (Applied-Biosystems) or with SYBR-Green® PCR 2× Master Mix (Applied-Biosystems). Each reaction was run in duplicate in a final volume of 25 mL with 0.2 mM final concentration of each primer, 0.25 mM final concentration of each probe and 10 µL of appropriately diluted DNA samples. Amplifications were carried out using the following ramping profile: 1 cycle at 95°C for 10 min, followed by 40 cycles of 95°C for 30 s, 60°C for 1 min. For SYBR-Green® amplifications, a melting step was added to improve amplification specificity. Total numbers of bacteria were inferred from averaged standard curves as described [23].

Normalization of quantitative PCR data

In microbiota, all-bacteria results were presented as the mean of the log10 value ± standard errors of means (SEM). To overcome the fact that faecal samples may contain more or less water, we have normalized the data for each faecal sample as previously described [20]. The level for each bacterial species or group was subtracted by the level of all-bacteria content. The data were given as the log number of bacteria per gram of stool.

Statistical Analysis

Clinical and biological data

Data are expressed as mean ± SEM. Clinical and biological values not normally distributed were log transformed. Analysis of variance (Anova) was used to assess the difference of clinical and biological parameters at baseline between the different complete groups. The progression of the different parameters in the obese subjects before and after RYGB was evaluated by Multivariate Analysis of Variance (Manova). Insulin resistance (HOMA-IR),

insulin sensitivity (HOMA-S%) and beta-cell function (HOMA-B%) provided in supplementary Table 2 were estimated using the method described in [24].

Gut microbiota signatures

Wilcoxon Rank Sum tests were used to assess the statistical significance of differences in bacterial groups between lean controls, OB/nD and OB/D subjects at baseline. Paired rank tests were performed to analyze changes in bacteria faecal counts between various time points (M3 vs. M0 and M6 vs. M0).

Principal component analysis (PCA) combined with co-inertia analysis was used to explore complex and potentially redundant relationships involving a relatively large number of clinical, biological and microbiological variables at baseline, and following RYGB. Co-inertia analysis is a coupling method for comparing different types of parameters presenting different variances. The significance of dynamic variations between various time points after surgery (i.e. M0, M3 and M6), associating microbiota, biological and clinical parameters, was evaluated by Monte Carlo tests. The results of these analyses were visualized by a circle of correlations. Significant associations between the analyzed variables were tested by computing Spearman correlation coefficients.

The significance of the strongest associations, among those identified by PCA and co-inertia analysis, relating variations of clinical-biological parameters and of microbiota counts after surgery, was further evaluated by building linear mixed-effects models (LME) to test for inter-variables redundancies and to adjust for potential confounding factors. All LME models were fit by maximizing the restricted log-likelihood (REML) of their estimated coefficients. All statistical analyses were performed using the R software (http://www.r-project.org). PCA and co-inertia analyses were performed with ADE-4 package [25]**.** LME modeling was performed by relying on functions available in the nlme package [26]. All statistical

computations were considered significant when resulting p-values were smaller than the conventional 0.05 threshold.

Results

Clinical and biological characteristics before RYGB

Clinical characteristics of lean controls and obese subjects, diabetic (OB/D) or non-diabetic (OB/nD), are presented in Table 1. While mean age between controls and OB/nD subjects were not statistically different, OB/D subjects were older. As expected, most clinical and biological parameters were found to be significantly diverged between the control and the two obese groups. Obese subjects had higher leptin, fasting glucose, insulin and triglyceride serum concentrations and lower adiponectin and HDL cholesterol serum concentrations compared with values found in the control group. Inflammatory markers were higher in the OB/D group compared to the OB/nD group, but the difference was not significant (supplementary table 2).

Clinical, metabolic and inflammatory changes after RYGB

Along with the drastic reduction in food consumption, RYGB resulted in significant changes in body weight, BMI and fat mass from M0 to M3 and M6. The progression of clinical parameters related to body composition, metabolic and inflammatory parameters in all obese subjects are presented in Table 2. For the majority of parameters, major changes occurred rapidly in the first three months. At M6, the subjects had lost 22 ± 0.01% of their initial weight ($p < 0.01$). Fat mass decreased and the percentage of fat-free mass increased significantly. Resting energy expenditure (REE) reduced following RYGB, in agreement with the reduction in fat-free mass (data not shown). These changes were associated with a significant decrease in adipocyte cell diameter ($p < 0.05$) and in serum concentrations of leptin ($p < 0.01$). Serum concentrations of cholesterol, triglycerides and of inflammatory parameters (hsCRP, orosomucoid, IL-6) decreased post-surgery while plasma adiponectin levels

increased significantly as expected. These improvements were observed in both groups (OB/nD and OB/D) when considered separately (supplementary Table 2).

In all subjects, plasma glucose, insulin levels and glycosylated hemoglobin (HbA1C) decreased significantly post-RYGB. The change in HOMA-IR (Table 2) was borderline significant owing to the combination of OB/nD and OB/D groups. Supplementary Table 2 shows the significant improvement in insulin sensitivity of the OB/nD group as well as the sub-significant improvement of blood glucose tolerance in the 7 diabetic subjects. Anti-diabetic drugs were stopped in all diabetic subjects as well as hypolipidemic treatment in all obese individuals.

Comparison of bacterial populations in stools from lean controls and obese subjects before RYGB

Microbiota analysis values obtained for the 13 lean subjects, recruited at the same time and from the same geographical area as the obese subjects, were compared to the previously described results of from 21 healthy adults [20]. No significant difference in the composition of the microbiota was observed between these two groups (data not shown). The 13 lean subjects served as the control group.

Average counts for each bacterial group are presented in Figure 1. We assessed the main groups of faecal bacteria in lean and morbidly obese subjects by qPCR. The amounts of *C. leptum*, *C. coccoides*, *Bacteroides/Prevotella*, *Lactobacillus/Leuconostoc/Pediococcus* groups, *Bifidobacterium* genus, *E. coli* and *F. prausnitzii* species were evaluated. Significant changes were observed primarily for *F. prausnitzii* species and *Bacteroides/Prevotella* group.

As indicated in Table 3, the comparison of microbial populations in obese individuals (OB/nD and OB/D) to those of the control group did not show statistically significant differences in *C. leptum, C. coccoides, Lactobacillus/Leuconostoc/Pediococcus* groups, *Bifidobacterium* genus

or *E. coli* species. However, while the population of *C. leptum* was higher in the controls' microbiota compared to that of obese subjects, the differences did not reach statistical significance probably due to the high inter-individual variability in this bacterial population subgroup.

Statistical differences were shown in the *Bacteroides/Prevotella* group with lower amounts recorded in obese subjects (OB/nD: -1.61 ± 0.1, p=0.039 and OB/D: -1.61 ± 0.2, p=0.038) compared to the control group (-1.11 ± 0.1).

Interestingly, the *F. prausnitzii* species qPCR system could reliably distinguish between the control and OB/D microbiota. This study showed that *F. prausnitzii* counts in the OB/D microbiota (-2.79 ± 0.5) were lower when compared with those of control group (-1.06 ± 0.2, $p< 0.01$) and OB/nD subjects (-1.45 ± 0.13). These results suggested that while the obese state leads to modification in the amount of *Bacteroides/Prevotella* group in the faeces of these severely obese subjects, the diabetic condition might influence the abundance of faecal *F. prausnitzii* as illustrated in Figure 2.

Bacterial changes after RYGB in the obese subjects

Gastric bypass drastically improved both metabolic and inflammatory parameters. We also examined the changes in bacterial composition during RYGB-induced weight loss.

Significant changes of faecal bacteria amounts were observed in the obese group as a whole after surgery but with a different pattern depending on the bacterial group (see Figure 1). Supplementary Table 3 illustrates the progression of all bacterial populations within the microbiota before (M0) and after RYGB (M3 and M6) in each obese group, separated by the diabetic status. In the OB/D subjects, a similar pattern of changes as the one characterizing the OB/nD was observed, but changes for certain bacterial groups did not reach statistical significance probably due to the small size of the samples. Post-RYGB,

Bacteroides/Prevotella and *E. coli* populations increased, the *Bifidobacterium* genus and the *Lactobacillus/Leuconostoc/Pediococcus* group decreased.

The *Bacteroides/Prevotella* population, whose level was lower in obese subjects before RYGB, increased at M3 and remained stable until M6 (figure 1, supplementary Table 3) at a level close to that observed in faecal samples of the controls. Importantly, the obese subjects remained obese at M6 (BMI 37.1 ± 1.3 vs. 21.1 ± 0.4 for obese and lean subjects, respectively). At M3, *E. coli* species showed a rapid and significant increase reaching a level higher than that of the controls. An opposite pattern was observed for both the *Bifidobacterium* genus and *Lactobacillus/Leuconostoc/Pediococcus* group. Levels of both populations decreased significantly at M3 and M6 and reached, in the case of *Bifidobacterium*, a level lower than that measured in controls (Figure 1 and supplementary Table 3).

The level of the *F. prausnitzii* population in OB/D subjects was significantly lower compared to OB/nD individuals before RYGB, but increased at M3 and remained stable at M6 (supplementary Table 3). Analysis also showed that the populations of Clostridia (*C. leptum* and *C. coccoides*) were stable post-RYGB.

Microbiota composition and clinical phenotypes before surgery

Bacterial populations were associated with parameters related to body composition, glucose and lipid metabolism as well as inflammation before the surgery. Importantly, no significant association was correlated with age for any analysis.

In OB/nD and OB/D subjects, we observed significant relationships between the amount of *F. prausnitzii*, *E coli* and *Bacteroides/Prevotella* and some metabolic and inflammatory parameters (data not shown). The strongest associations were found for the amount of *F. prausnitzii* which was negatively correlated with serum concentrations of inflammatory circulating markers (hsCRP Rs -0.54, $p< 0.01$ and IL-6 Rs -0.65, $p< 0.001$). We also found a

negative correlation between *F. prausnitzii* and parameters related to blood glucose homeostasis (HbA1C Rs -0.39, $p< 0.05$, fasting glucose Rs -0.47, $p=0.01$ and HOMA-IR Rs -0.47, $p< 0.01$). In the OB/nD subjects, only the negative correlation between *F. prausnitzii* and inflammatory parameters remained consistently significant (hsCRP Rs -0.58, $p< 0.01$, IL-6 Rs-0.60, $p< 0.01$, orosomucoid Rs-0.39, $p< 0.05$). Thus the amount of *F. prausnitzii* was correlated with the low-grade inflammatory state in obese subjects independent of the diabetic state.

Time dependant associations between corpulence, metabolism, calorie intake and bacterial gut populations

We further examined the dynamic relationships between changes in bacterial composition and metabolic parameters. Statistical linear mixed-effects models (LME) were used to distinguish within-subject from between-subject sources of variation, and to describe how trajectories in clinical and bacterial population mean responses showed related changes over time. Analyses first included the entire population of obese subjects, regardless of their diabetic status, and secondly in the OB/nD group or OB/D alone. We observed significant associations between corpulence parameters and the development of the populations of faecal bacteria post-RYGB. Some of these associations were noted to be highly depended on calorie intake.

Bacteroides/Prevotella counts, which increased with weight loss, were negatively correlated with the changes in body weight (Rs -0.33, $p< 0.05$), BMI (Rs -0.35, $p< 0.05$) and body fat mass (Rs -0.32, $p< 0.01$). A strong and negative relation was also observed with leptin serum concentrations (Rs -0.43, $p< 0.01$). Additionally, we found a positive correlation with fat-free mass changes (Rs 0.31, $p< 0.01$). The exclusion of OB/D subjects did not change the negative association found for adiposity-related parameters (data not shown). In the OB/nD group, *Bacteroides/Prevotella* counts correlated negatively with calorie intake (Rs -0.39, $p< 0.01$) which drastically changed after the bypass. Multivariate analysis, performed in the OB/nD

group, associating calorie intake and each of the adiposity-related parameters as fixed-effects in a combined LME model, confirmed the negative relationship between *Bacteroides/Prevotella* counts and the decrease of food consumption post-RYGB. This result was independent of corpulence ($p < 0.05$). However, the combined model could not demonstrate significant independent relationships with any of the adiposity-related parameters, thus indicating that variations in *Bacteroides/Prevotella* population after surgery are related mostly to calorie intake in this cohort.

Negative correlations were also observed between adiposity-related parameters and the changes of *E. coli* counts in faecal samples. *E. coli* changes showed negative correlations with regard to changes in body weight (Rs -0.42, $p < 0.0001$), BMI (Rs -0.47, $p < 0.0001$), fat mass (Rs -0.41, $p < 0.0001$), and calorie intake (Rs -0.47, $p < 0.010$). As observed with the *Bacteroides/Prevotella* population, a strong negative correlation was found with leptin serum concentrations (Rs -0.53, $p < 0.001$). Multivariate analysis, setting calorie intake and each of the adiposity-related parameters as fixed-effects in a combined LME model, confirmed the negative associations with variations in *E. coli* counts independent of food consumption (body weight $p < 0.01$, BMI $p < 0.001$, fat mass $p < 0.001$, leptin $p < 0.05$). Interestingly, unlike the *Bacteroides/Prevotella* population, the relationship between calorie intake and *E. coli* counts lost statistical significance in the combined model. This suggests that *E. coli* could be considered as a marker of corpulence variation after surgery, independent of energy intake. The relationships between the faecal microbiota and these clinical parameters, explored through PCA, is illustrated in Figure 3a, which displays the strong negative correlation between *E. coli* counts and leptin serum concentration. This correlation is further reinforced in Figure 3b, which concomitantly illustrates the kinetic evolution between *E. coli* population and leptin with as a mirror image. In addition, in the OB/nD group, the increase in *E. coli* population post-RYGB is also associated with improvements of blood glucose and insulin

sensitivity with a significant negative association observed with the changes in fasting glucose (Rs -0.35, p=0.005), HbA1C (Rs -0.22, p=0.048), HOMA-IR (Rs -0.33, p=0.034) and a positive association with insulin sensitivity HOMA-S%. These correlations, however, are not significant after adjustment for energy intake in a multivariate LME model.

The *Lactobacillus/Leuconostoc/Pediococcus* and *Bifidobacterium* group demonstrated an inverse pattern of variation as compared to *E.coli* and the *Bacteroides/Prevotella*. *Bifidobacterium* changes showed positive correlations with changes in body weight (Rs 0.19, $p< 0.01$), BMI (Rs 0.17, $p< 0.01$), calorie intake (Rs 0.28, $p< 0.05$), and insulin levels (Rs 0.30, $p< 0.05$). A positive association was found with leptin serum concentrations (Rs 0.34, $p< 0.001$) while adiponectin serum concentration associated negatively with this bacterial population (Rs -0.18, $p< 0.01$). Significant associations were also found with lipid values; positive with the change in triglycerides concentration (Rs 0.12, $p< 0.05$) and negative with HDL-cholesterol (Rs 0.18, $p< 0.05$). Similar patterns of association with corpulence related parameters were observed when excluding the diabetic subgroup (data not shown). Among these parameters, only body weight and BMI demonstrated positive relationships with *Bifidobacterium* population after adjustment for energy intake ($p< 0.01$ and $p< 0.05$, respectively, in a multivariate LME model setting associating calorie intake as a fixed effect with each of these parameters). In addition, multivariate analysis also highlighted the relationship of this bacterial population with food consumption after surgery, regardless of the corpulence level ($p< 0.01$ in a combined LME model setting calorie intake and BMI as fixed effects). For the *Lactobacillus* group, the association with corpulence related parameters was not significant. Positive associations were nevertheless found with changes in total calorie intake in obese subjects (Rs 0.29, $p< 0.01$), or in OB/nD analyzed alone (Rs 0.30, $p< 0.01$).

Negative and more marginal associations were found between *F. prausnitzii* population changes and modifications in corpulence related parameters (body weight Rs -0.15 $p< 0.05$,

BMI Rs -0.07, p=0.07, fat-mass Rs -0.03, p< 0.012 and serum leptin Rs -0.14, p< 0.05) in the whole group of obese subjects. These relations could not be confirmed when considering only the OB/nD group in the analysis. Negative associations were found with the improvement of fasting blood glucose (p< 0.0001 for fasting glucose Rs -0.22 and HbA1C Rs -0.17) and HOMA-IR (Rs -0.22, p< 0.001) but these associations could not be confirmed when the analysis was limited to OB/nD subjects. Multivariate analysis confirmed the significance of the negative relation between *F. prausnitzii* population and the improvement of blood glucose tolerance in diabetic subjects, independent of calorie intake level (p< 0.001 for fasting blood glucose and HbA1c and p=0.002 for HOMA-IR in respective LME models associating each of these parameters with calorie intake).

Time dependant associations between inflammatory parameters and faecal bacteria population changes; importance of *F. prausnitzii*

In contrast to other bacterial populations, *F. prausnitzii* showed a consistent correlation with inflammation.markers These circulating inflammatory parameters improved after the surgery. *F. prausnitzii* variation was strongly and negatively correlated with changes of hsCRP (Rs -0.39, p< 0.0001), IL-6 (Rs -0.35, p< 0.0001) and orosomucoid serum levels (Rs -0.32, p< 0.05) when non-diabetic and diabetic obese subjects were grouped together. The correlations with hsCRP and IL-6 were maintained when considering only the OB/nD group, (Rs -0.37, p< 0.0001 and Rs -0.34, p< 0.01, respectively). These relationships were independent of calorie intake.

Discussion

Unraveling the rapid somatic benefits observed after RYGB has proven challenging predominantly because this procedure associates with a myriad of factors such as caloric

restriction, diminished nutrient absorption, reduced adipose mass, modified gut hormone signaling and changes in whole-body glucose metabolism. RYGB is a procedure which significantly diminishes the gastric pouch and reduces proximal small bowel exposure to acid, a phenomenon which, together with modified bile acid fluxes, might also be expected to modify the gut microbiota composition. Another important consequence of RYGB is the improvement of low-grade inflammation which is suggested to contribute to drastic improvement of obesity comorbidities like diabetes and cardiovascular risks.

Our study performed in carefully monitored subjects not only confirmed that obese individuals exhibit distinct intestinal communities compared to lean ones, but also, showed that RYGB leads to dynamic changes in the composition of the faecal microbiota associated with improvements of metabolic and inflammatory markers.

In agreement with the landmark studies of Ley *et al*. [12] and Turnbaugh *et al*. [27], we showed a lower proportion of *Bacteroidetes* in severely obese subjects, but no significant difference between microbiota of lean and obese groups in major components of Firmicutes, *i.e. Clostridia* and *Lactobacillus* groups. Nevertheless proportions of the *C. leptum* group tended to be more than two fold lower in obese diabetic and non-diabetic subjects than in controls. The difference was found to be not statistically significant illustrating the high inter-individual variability in faecal microbiota. Another remarkable result was that compared to lean and non-diabetic obese subjects, type 2 diabetics showed lower proportion of the species *F. prausnitzii*, a major component of *C. leptum group*. This could possibly explain the trend towards decreased *C. leptum* in obesity.

The analysis of the dynamic changes post-RYGB provided important information regarding the potential associations between gut microbiota composition, food intake, metabolic adaptations and inflammation. While some groups of gut bacteria correlated well with energy

intake, body corpulence and metabolic changes, other groups could be associated with changes in the inflammatory state.

Some faecal bacteria populations were closely linked with the changes in body corpulence parameters as illustrated by the decreased proportions of *Bacteroides/Prevotella* in obesity and their rapid increased proportions after the surgery. Correlative studies in LME kinetic models showed that these populations of bacteria were strongly associated with a series of body composition and metabolic parameters. After weight loss, the higher the increase in the faecal proportions of *Bacteroides/Prevotella***,** the better the reduction in body fat mass and leptin circulating levels and the better the improvement of insulin sensitivity. These relationships were also found in the non-diabetic group considered alone. Importantly, we observed that most of the associations between the *Bacteroides/Prevotella* group and these parameters were highly dependent on energy intake. Other groups have observed modifications of *Bacteroides/Prevotella* group after weight loss induced by dietary challenge, but did not analyze the potential correlations with metabolic and/or inflammatory changes. Ley RE *et al* showed an increase in the Bacteroidetes phylum but with a concomitant and significant reduction of Firmicutes by studying 12 adults after one year fat or carbohydrate restricted hypocaloric diet [12]. The estimated Firmicutes/Bacteroidetes ratio diminished substantially, an observation also in our study mostly due to the increase in *Bacteroides*/*Prevotella* group (data not shown). However the other published data has raised a certain degree of controversy with regards to this result. While some studies showed decreased total numbers of bacteria, no changes in the proportion of Bacteroidetes or of the Firmicutes/Bacteroidetes ratio were observed by analyzing the gut microbiota of 23 obese subjects after an 8 weeks diet [13] and of 39 adolescents with limited weight loss (< 2kg) after 10 weeks hypocaloric diet [28]. These apparent discrepancies could be attributed to substantial differences inherent with calorie restriction leading to varying levels of fat mass

loss and in the duration of the diet. In this regard, adolescents who lost more than 4 kg after a 10 weeks diet showed an increase of the proportion in Bacteroidetes and a reduction in *C. coccoides*, a component of Firmicutes phylum [28]. RYGB can be considered as a unique model in this respect, inducing drastic calorie restriction and subsequent weight loss which can be clinically followed over time in the same individual. A first study performed in three adults after RYGB showed a significant decrease in Firmicutes together with an increase in Gamma-Proteobacteria (members of the *Enterobacteriaceae* family) by using pyrosequencing [29]. In this study the individual fecal samples before and after weight loss were not paired and were analyzed in only 3 subjects at 8 and 15 months after RYGB. Weight loss stopped in 1 of the 3 subjects, whereas the 2 others continued to lose weight with no precision provided about the amount lost and about other potential phenotypic changes.

In the present study, patients were followed for 6 months; RYGB resulted in an efficient decrease in body weight (average 22% of the initial weight) and a reduction in percent fat mass by 14%. The adipocyte secreted hormone leptin dropped dramatically at M3 to half of its serum concentrations before the surgery. These changes were related to a marked reduction in mean food intake from 1933 kcal per day to 1080 kcal at M3 and 1355 kcal at M6. While the evaluation of food consumption was performed by an experienced dietetician, it is important to note that there is a well known uncertainty and possible underestimation of food intake in obese subjects [30]. After RYGB, patient's food intake is profoundly modified. After a short fasting period (1 to 3 days), the subjects start increasing their intake principally composed of liquid or semi liquid food for one week. Starch-based food (vegetable soup, mashed potatoes etc.) is the principal food intake during 3 months after RYGB with solid foods being progressively reintroduced (data not shown). Knowing the uncertainty in food intake evaluation, we nevertheless observed that several of the changes in bacterial groups were correlated with rapid modification in food intake. Thus changes in the microbiota

composition could reflect an adaptation to drastic reduction in energy intake as suggested by the multivariate analysis showing that several associations were found to depend on total calorie intake.

The amount of *Bacteroides*/*Prevotella* and *E. coli* negatively associated with leptin variation while the associations were positive with *bifidobacteria* and *lactobacilli.* Leptin is a pivotal adipose tissue secreted hormone with a multitude of physiological functions and a major role in the initiation of adaptation responses to starvation [31]. Leptin levels fall rapidly with the onset of energy deprivation disproportionably to changes in adipose tissue mass. This observation is illustrated by the severe drop of leptin at M3 and relative stabilization at M6 while BMI and fat mass continued to shrink. This phenomenon is recognized to contribute to signaling the shift between sufficient and insufficient body energy. Thus in agreement with *Bajzer et al.* [32], some changes in gut microbiota post- RYGB could be linked to maximizing energy harvest as a host adaptation to the starvation-like and drastic nutrient deprivation situation. The fact that for most gut bacteria, changes were observed mostly at M3 and remained stable at M6 (supplementary Table3 and supplementary figure1) while corpulence and metabolic factors continued to improve favors this particular interpretation. A recent study showed that, compared with germ free animals, hepatic ketogenesis is enhanced during 24-h fasting in mice (CONV-D) after a microbiota transplant from the distal gut of conventionally raised lean counterparts fed with carbohydrate. The CONV-D mice showed an increase of short chain fatty acids and the proportion of Bacteroidetes switched from 20.6% at the fed state to 42.3% at the fasted state. A reduction in the proportion of Firmicutes from 77.1% to 52.6%, respectively, at the fed to fasted state was also found. Previous studies had demonstrated that starved germfree (GF) mice died more quickly from starvation consequences than their conventionally raised counterparts despite losing weight at approximately the same rate [33]. Strikingly, the authors showed that the contamination of GF

animals by the single *E. coli* strain could prolong the animal's survival. Whether the change in gut microbiota in a RYGB model resulted in a substantial increase in production of short chain fatty acids as an additional source of energy for the body is unknown and awaits further evaluation.

On the other hand, we found that the relationship between some gut microbiota changes and corpulence and metabolic parameters was not fully dependent on dietary changes using statistical adjustments as evidenced in the case of *E. coli* and *Bifidobacterium*. This is an indirect indication that microbiota components could intervene in some of the metabolic changes associated with this surgical procedure. The signal molecules mechanistically involved in driving these links need to be more fully understood.

The RYGB procedure *per se* could also contribute to changes in gut microbiota composition. RYGB creates a small (15-30 milliliters) gastric pouch and the distal stomach and proximal small intestine are bypassed by attaching the distal end of the mid-jejunum to the proximal gastric pouch (creating the Roux limb). The bile and pancreatic limb is attached along the Roux limb. Gastric acidity is bypassed and results in a reduction of chloride acid flux in the gut. The resulting pH increase, together with the downstream delivery of bile acids, could contribute to the modification of faecal bacteria population. In *in vitro* culture studies, the growth of *Bacteroidetes* was progressively inhibited at reduced pH values below 6.5 and quite poor at pH 5.5. Growth of *E. coli* which was also shown to be facilitated by increased pH [34]. We were not able to measure pH in our subjects' faecal samples. However, the decreased acidity in gut after RYGB could favor the increase of *Bacteroidetes* and *E. coli* counts. Another consequence could be the decrease in *Lactobacillus* and *Bifidobacterium* [35, 36]. While these two bacterial groups were similar in lean and obese subjects, their proportion significantly decreased after RYGB. This is not consistent with studies in mice suggesting that a beneficial effect of *Bifidobacterium* species in the improvement of obesity-related

metabolic and inflammatory condition [8]. The *Bifidobacterium* genus, however, is complex. In adolescents losing moderately amounts of weight, Santacruz A *et al* found that counts of *B. bifidum* and of *B. breve* diminished while *these of B. catenulatum* increased [37].

More studies are needed to explore, in the very short term (immediately after the surgery) and longer terms (until weight stabilization), the dynamics of subjects undergoing bypass surgery with attention given to food intake behavior, measures of metabolic mediators (such as SCFA and non esterified fatty NEFA), measures of faecal pH to explore the dependency between changes in food intake and gut bacterial groups as well as the influence of the surgery itself. A comparison with patients only subjected to a restrictive surgical procedure (*i.e*. gastroplasty) would be useful in this respect.

The other important information revealed in our current study is that faecal proportions of the species *F. prausnitzii* were consistently associated with inflammatory parameters like hsCRP, orosomucoid and IL-6, an important obesity-related cytokine. These associations were found at the basal state prior to the surgery and after the surgery. *F. prausnitzii* has been identified as a highly conserved and dominant species of the human faecal microbiota of healthy individuals [38]. The reduction of *F. prausnitzii* has been described in inflammatory bowel disease and in infectious colitis patients [22]. It was suggested that *F. prausnitzii* might play an important role in preventing local bowel inflammation and infection in acute inflammatory disease. Our study suggests that *F. prausnitzii* could also play a role in low-grade inflammation pathologies like obesity and diabetes. As in other chronic diseases, human obesity presents with low-grade systemic inflammation and is characterized by an increased production of inflammatory mediators, notably with the activation of pro-inflammatory signaling pathways [39, 40]. Importantly, adipose tissue, characterized by the accumulation of macrophages in obese subjects, contributes to the production of inflammatory mediators in circulation. Low-grade inflammation in obesity is considered as one of the pivotal

mechanisms linking obesity with its co-morbidities such as metabolic diseases (insulin resistance, type 2 diabetes, liver disease) and cardiovascular diseases. We and others observed a moderate increase of inflammatory mediators such as plasma levels of hsCRP, tumor necrosis factor-α (TNF-α) and IL-6 [41-44]. Here we showed a relationship between *F. prausnitzii* and the inflammatory state both in OB/nD and in OB/D patients. In addition, the proportions of *F. prausnitzii* were lower in type 2 diabetics displaying a worsening of their low-grade inflammation [45] and higher insulin resistance. RYGB significantly improved low-grade inflammation together with the improvement of insulin sensitivity.

In correlative studies we observed a negative relationship between the reduction of circulating inflammatory mediators and the proportions of *F. prausnitzii.* This relationship remained significant after adjustments for body corpulence parameters, in non-diabetic obese subjects and was not dependant upon calorie intake. An association was also seen between *F. prausnitzii* and the improvement in insulin sensitivity but with an effect explained by the amelioration of glucose metabolism in the diabetic group. The *F. prausnitzii* population variation was associated with modulation of at least eight urinary metabolites of diverse structure indicating that this species is a highly active member of the microbiome, influencing numerous host pathways [46]. It was recently suggested that *F. prausnitzii* exhibits anti-inflammatory effects, partly due to secreted metabolites able to block Nuclear factor kappa B (NF-κB) activation and the secretion of proinflammatory mediators [47]. Moreover the oral administration of *F. prausnitzii* or of supernatant from *F. prausnitzii* cultures increased the production of IL-10 by blood mononuclear cells and reduced the production of the proinflammatory mediator like IL-12 in the colon. The modulation of NF-κB by pharmacological agents such as statins (pravastatin) or salicylates has been proposed as a tool to improve insulin sensitivity in type 2 diabetes patients [48-50]. Our study raises the question regarding the role of *F. prausnitzii* as a mediator of low-grade inflammation in obesity and

diabetes and open avenues for future investigation exploring its contribution in insulin resistance. Indeed, because of an increasing interest in treating type 2 diabetes with gastric bypass surgery, the rapid and long-term improvement of insulin sensitivity and the reduction of diabetes in subjects post surgery has become a primary axis of interest [51]. Whether *F. prausnitzii* could be considered as a valuable therapeutic tool for the improvement of inflammation, blood glucose tolerance and insulin sensitivity calls for more investigation in the future.

Acknowledgments

We thank the support from Assistance Publique-Hôpitaux de Paris (APHP) and Direction of Clinical research (CRC) which promoted and supported the clinical investigation. Lingchun Kong received a support from DANONE (France) and Corneliu Henegar a support from Sanofi/AFERO (French Association for the Research on Obesity). We thank Mme Christine Baudouin, Dr Florence Marchelli, and Mme Patricia Ancel involved in patient's recruitment, data collection and sampling at the Center of Research on Human Nutrition, Paris Pitié-Salpêtrière Hospital. We thank Dr Sean P Kennedy for critical reading of the manuscript.

Bibliography

1. Clement K, Langin D (2007) Regulation of inflammation-related genes in human adipose tissue. J Intern Med 262: 422-430.
2. Pradhan A (2007) Obesity, metabolic syndrome, and type 2 diabetes: inflammatory basis of glucose metabolic disorders. Nutr Rev 65: S152-156.
3. DiBaise JK, Zhang H, Crowell MD, Krajmalnik-Brown R, Decker GA, et al. (2008) Gut microbiota and its possible relationship with obesity. Mayo Clin Proc 83: 460-469.
4. Yazigi A, Gaborit B, Nogueira JP, Butiler ME, Andreelli F (2008) [Role of intestinal flora in insulin resistance and obesity]. Presse Med 37: 1427-1430.
5. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, et al. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444: 1027-1031.
6. Backhed F, Ding H, Wang T, Hooper LV, Koh GY, et al. (2004) The gut microbiota as an environmental factor that regulates fat storage. Proc Natl Acad Sci U S A 101: 15718-15723.

7. Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, et al. (2005) Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A 102: 11070-11075.
8. Cani PD, Neyrinck AM, Fava F, Knauf C, Burcelin RG, et al. (2007) Selective increases of bifidobacteria in gut microflora improve high-fat-diet-induced diabetes in mice through a mechanism associated with endotoxaemia. Diabetologia 50: 2374-2383.
9. Cani PD, Delzenne NM (2007) Gut microflora as a target for energy and metabolic homeostasis. Curr Opin Clin Nutr Metab Care 10: 729-734.
10. Cani PD, Bibiloni R, Knauf C, Waget A, Neyrinck AM, et al. (2008) Changes in gut microbiota control metabolic endotoxemia-induced inflammation in high-fat diet-induced obesity and diabetes in mice. Diabetes 57: 1470-1481.
11. Amar J, Burcelin R, Ruidavets JB, Cani PD, Fauvel J, et al. (2008) Energy intake is associated with endotoxemia in apparently healthy men. Am J Clin Nutr 87: 1219-1223.
12. Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. Nature 444: 1022-1023.
13. Duncan SH, Lobley GE, Holtrop G, Ince J, Johnstone AM, et al. (2008) Human colonic microbiota associated with diet, obesity and weight loss. Int J Obes 32: 1720-1724.
14. Duncan SH, Louis P, Flint HJ (2007) Cultivable bacterial diversity from the human colon. Lett Appl Microbiol 44: 343-350.
15. Buchwald H, Avidor Y, Braunwald E, Jensen MD, Pories W, et al. (2004) Bariatric surgery: a systematic review and meta-analysis. JAMA 292: 1724-1737.
16. Poitou C, Lacorte JM, Coupaye M, Bertrais S, Bedel JF, et al. (2005) Relationship between single nucleotide polymorphisms in leptin, IL6 and adiponectin genes and their circulating product in morbidly obese subjects before and after gastric banding surgery. Obes Surg 15: 11-23.
17. Kushner RF, Noble CA (2006) Long-term outcome of bariatric surgery: an interim analysis. Mayo Clin Proc 81: S46-51.
18. Clement K, Vega N, Laville M, Pelloux V, Guy-Grand B, et al. (2002) Adipose tissue gene expression in patients with a loss of function mutation in the leptin receptor. Int J Obes Relat Metab Disord 26: 1533-1538.
19. Yokoyama H, Emoto M, Fujiwara S, Motoyama K, Morioka T, et al. (2003) Quantitative insulin sensitivity check index and the reciprocal index of homeostasis model assessment in normal range weight and moderately obese type 2 diabetic patients. Diabetes Care 26: 2426-2432.
20. Furet JP, Firmesse O, Gourmelon M, Bridonneau C, Tap J, et al. (2009) Comparative assessment of human and farm animal faecal microbiota using real-time quantitative PCR. FEMS Microbiol Ecol 68: 351-362.
21. Huijsdens XW, Linskens RK, Mak M, Meuwissen SG, Vandenbroucke-Grauls CM, et al. (2002) Quantification of bacteria adherent to gastrointestinal mucosa by real-time PCR. J Clin Microbiol 40: 4423-4427.
22. Sokol H, Seksik P, Furet JP, Firmesse O, Nion-Larmurier I, et al. (2009) Low counts of Faecalibacterium prausnitzii in colitis microbiota. Inflamm Bowel Dis 15: 1183-1189.
23. Lyons SR, Griffen AL, Leys EJ (2000) Quantitative real-time PCR for Porphyromonas gingivalis and total bacteria. J Clin Microbiol 38: 2362-2365.
24. Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, et al. (1985) Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia 28: 412-419.
25. Chessel D, Dufour AB, Dray S (2009) Multivariate data analysis and graphical display. R package version 14-11.

26. Pinheiro J, Bates B, DebRoy S, Sarkar D, team. atRC (2009) Linear and Nonlinear Mixed Effects Models, R package version 3.1-93.
27. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. (2009) A core gut microbiome in obese and lean twins. Nature 457: 480-484.
28. Nadal I, Santacruz A, Marcos A, Warnberg J, Garagorri M, et al. (2008) Shifts in clostridia, bacteroides and immunoglobulin-coating fecal bacteria associated with weight loss in obese adolescents. Int J Obes (Lond).
29. Zhang H, DiBaise JK, Zuccolo A, Kudrna D, Braidotti M, et al. (2009) Human gut microbiota in obesity and after gastric bypass. Proc Natl Acad Sci U S A 106: 2365-2370.
30. Heitmann BL, Lissner L (1995) Dietary underreporting by obese individuals--is it specific or non-specific? BMJ 311: 986-989.
31. Ahima RS, Prabakaran D, Mantzoros C, Qu D, Lowell B, et al. (1996) Role of leptin in the neuroendocrine response to fasting. Nature 382: 250-252.
32. Bajzer M, Seeley RJ (2006) Physiology: obesity and gut flora. Nature 444: 1009-1010.
33. Tennant B, Malm OJ, Horowitz RE, Levenson SM (1968) Response of germfree, conventional, conventionalized and E. coli monocontaminated mice to starvation. J Nutr 94: 151-160.
34. Duncan SH, Louis P, Thomson JM, Flint HJ (2009) The role of pH in determining the species composition of the human colonic microbiota. Environ Microbiol.
35. Mason EE, Munns JR, Kealey GP, Wangler R, Clarke WR, et al. (1976) Effect of gastric bypass on gastric secretion. Am J Surg 131: 162-168.
36. Mason EE, Munns JR, Kealey GP, Wangler R, Clarke WR, et al. (2005) Effect of gastric bypass on gastric secretion. 1977. Surg Obes Relat Dis 1: 155-160; discussion 161-152.
37. Santacruz A, Marcos A, Warnberg J, Marti A, Martin-Matillas M, et al. (2009) Interplay between weight loss and gut microbiota composition in overweight adolescents. Obesity (Silver Spring) 17: 1906-1915.
38. Tap J, Mondot S, Levenez F, Pelletier E, Caron C, et al. (2009) Towards the human intestinal microbiota phylogenetic core. Environ Microbiol 11: 2574-2584.
39. Hotamisligil GS, Shargill NS, Spiegelman BM (1993) Adipose expression of tumor necrosis factor-alpha: direct role in obesity-linked insulin resistance. Science 259: 87-91.
40. Sartipy P, Loskutoff DJ (2003) Monocyte chemoattractant protein 1 in obesity and insulin resistance. Proc Natl Acad Sci U S A 100: 7265-7270.
41. Poitou C, Coupaye M, Laaban JP, Coussieu C, Bedel JF, et al. (2006) Serum amyloid A and obstructive sleep apnea syndrome before and after surgically-induced weight loss in morbidly obese subjects. Obes Surg 16: 1475-1481.
42. Das UN (2002) Obesity, metabolic syndrome X, and inflammation. Nutrition 18: 430-432.
43. Engstrom G, Stavenow L, Hedblad B, Lind P, Eriksson KF, et al. (2003) Inflammation-sensitive plasma proteins, diabetes, and mortality and incidence of myocardial infarction and stroke: a population-based study. Diabetes 52: 442-447.
44. Maachi M, Pieroni L, Bruckert E, Jardel C, Fellahi S, et al. (2004) Systemic low-grade inflammation is related to both circulating and adipose tissue TNFalpha, leptin and IL-6 levels in obese women. Int J Obes Relat Metab Disord 28: 993-997.
45. Akbay E, Yetkin I, Ersoy R, Kulaksizoglu S, Toruner F, et al. (2004) The relationship between levels of alpha1-acid glycoprotein and metabolic parameters of diabetes mellitus. Diabetes Nutr Metab 17: 331-335.
46. Hooper LV, Midtvedt T, Gordon JI (2002) How host-microbial interactions shape the nutrient environment of the mammalian intestine. Annu Rev Nutr 22: 283-307.

47. Sokol H, Pigneur B, Watterlot L, Lakhdari O, Bermudez-Humaran LG, et al. (2008) Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proc Natl Acad Sci U S A 105: 16731-16736.
48. Shi X, Ding M, Dong Z, Chen F, Ye J, et al. (1999) Antioxidant properties of aspirin: characterization of the ability of aspirin to inhibit silica-induced lipid peroxidation, DNA damage, NF-kappaB activation, and TNF-alpha production. Mol Cell Biochem 199: 93-102.
49. Weitz-Schmidt G (2002) Statins as anti-inflammatory agents. Trends Pharmacol Sci 23: 482-486.
50. Fleischman A, Shoelson SE, Bernier R, Goldfine AB (2008) Salsalate improves glycemia and inflammatory parameters in obese young adults. Diabetes Care 31: 289-294.
51. Couzin J (2008) Medicine. Bypassing medicine to treat diabetes. Science 320: 438-440.

Legends of figures

Figure 1: Quantifications of faecal microbiota in lean controls and obese subjects before (M0) and after surgery (M3 and M6).

The qPCR results were plotted as a boxes and whiskers graph. The boxes (containing 50% of all values) show the median (horizontal line across the middle of the box) and interquartile range, while whiskers represent the 10th and 90th percentiles. The extreme data points are indicated as circles.

Data not sharing the same letter in parenthesis within a horizontal line are significantly different ($P<0.05$).

Figure 2: Dynamic evolution of *F. prausnitzii* population levels from M0 to M3 and M6 after gastric surgery in obese diabetic (OB/D) and non-diabetic (OB/nD) subjects.

OB/nD ◆ and OB/D ● subjects.

■ Average *F. prausnitzii* population levels in control subjects.

Results were expressed as mean ± SEM of the Δlog10 value of normalized data, calculated as the log number of targeted bacteria minus the log number of all-bacteria.

Figure 3: Relationship between changes in faecal microbiota composition and clinical parameters in obese patients following RYGB surgery.

Real-time qPCR quantifications were used to determine the faecal microbiota composition for the bacterial groups indicated in supplementary Table 1. Clinical parameters included calorie intake, BMI, HOMA-IR, leptin, orosomucoid and adipocyte cell size.

3a. Principal Component Analysis (between class analysis). Bold arrows indicate the marked inverse relationship between changes in *E. coli* population and leptin serum concentrations.

3b. Dynamics of *E. coli* population evolution and leptin concentration during the study. *E. coli* population levels are expressed as mean ± SEM of the Δlog10 value of normalized data, calculated as the log number of targeted bacteria minus the log number of all-bacteria. Leptin results were expressed as mean ± SEM of serum concentrations.

Table 1. Clinical and biological characteristics of the lean controls, obese diabetic and non diabetic subjects before gastric surgery.

	Control n = 13	**OB/nD n = 23**	**OB/D n = 7**	***P*(1)**
Age (years)	36 ± 3 (A)	42 ± 2 (A, B)	49 ± 5 (B)	*0.048*
Adiposity				
Body Weight (kg)	59.0 ± 1.9 (A)	127.4 ± 4.3 (B)	121.5 ± 11.6 (B)	*<0.0001*
BMI (kg/m²)	21.7 ± 0.4 (A)	48.3 ± 1.6 (B)	45.4 ± 3.5 (B)	*<0.0001*
Adipocyte diameter (µm)	nd	115.6 ± 1.7 (A)	120.1 ± 3.4 (B)	*<0.0001*
REE (kcal)	nd	1830.3 ± 65.2 (A)	1762.1 ± 85.9 (A)	NS
Fat mass%	26.5 ± 1.4 (A)	48.2 ± 1.1 (B)	46.6 ± 2.1 (B)	*<0.0001*
Fat free mass%	69.5 ± 1.3 (A)	49.7 ± 1.1 (B)	50.9 ± 1.9 (B)	*<0.0001*
Leptin (ng/ml)	9.24 ± 1.46 (A)	50.5 ± 4.1 (B)	52.2 ± 8.3 (B)	*<0.0001*
Plasma glucose homeostasis and insulin sensitivity				
Glycaemia (mmol/l)	4.3 ± 0.2 (A)	5.4 ± 0.2 (A)	9.4 ± 1.4 (B)	*<0.0001*
Insulinemia (µU/ml)	3.8 ± 0.3 (A)	16.6 ± 1.8 (B)	19.4 ± 10.9 (B)	*<0.0001*
Adiponectin (µg/ml)	13 ± 1.43 (A)	6.8 ± 0.6 (B)	5.4 ± 0.8 (B)	*<0.0001*
Plasma lipid homeostasis				
Total cholesterol (mmol/l)	4.8 ± 0.3 (A)	4.5 ± 0.2 (A)	4.6 ± 0.3 (A)	NS
Total triglycerides (mmol/l)	0.8 ± 0.1 (A)	1.5 ± 0.2 (B)	1.8 ± 0.5 (B)	*0.004*
HDL cholesterol (mmol/l)	1.8 ± 0.1 (A)	1.2 ± 0.1 (B)	1.3 ± 0.1 (B)	*<0.0001*
Inflammatory markers				
Plasma hsCRP (mg/dl)	nd	2.7 ± 0.8 (A)	4.4 ± 2.4 (A)	NS
Plasma IL-6 (pg/ml)	nd	4 ± 0.4 (A)	5.8 ± 1.3 (A)	NS
Plasma orosomucoid (g/l)	nd	1.05 ± 0.05 (A)	0.95 ± 0.08 (A)	NS

All values are expressed as mean ± SEM; nd: not determined. (1) Anova for changes between the groups. Data not sharing the same letter within a horizontal line are significantly different ($P<0.05$).

Table 2. Clinical and biological characteristics of obese subjects before, and 3 and 6 months after gastric surgery.

		After by pass		
	Before by pass	**3 months**	**6 months**	***P*(1)**
Food intake				
Food intake (kcal)	1933 ± 101 (A)	1080 ± 87 (B)	1355 ± 54 (C)	*0.046*
Adiposity markers				
Body Weight (kg)	126 ± 4.2 (A)	107 ± 3.9 (B)	98 ± 3.8 (C)	*<0.0001*
BMI (kg/m²)	47.6 ± 1.5 (A)	40.6 ± 1.3 (B)	37.1 ± 1.3 (C)	*<0.0001*
Adipocyte diameter (µm)	116.7 ± 1.5 (A)	114.7 ± 2.2 (A)	103.3 ± 3.2 (B)	*0.011*
REE (kcal)	1814.4 ± 54.8 (A)	1865.3 ± 55.3 (B)	1551.1 ± 42.9 (C)	*0.004*
Fat mass %	47.9 ± 1.0 (A)	44.5 ± 1.0 (B)	41.3 ± 1.2 (C)	*0.004*
Lean mass %	50.0 ± 1.0 (A)	53.0 ± 0.9 (B)	55.9 ± 1.1 (C)	*0.028*
Leptin (ng/ml)	50.8 ± 3.7 (A)	25.6 ± 2.5 (B)	24.9 ± 2.8 (B)	*0.004*
Plasma Glucose homeostasis and insulin sensitivity				
Glycaemia (mmol/l)	6.4 ± 0.5 (A)	5.1 ± 0.2 (B)	4.8 ± 0.1 (B)	*0.038*
HbA1C (%)	6.4 ± 0.3 (A)	5.7 ± 0.1 (B)	5.8 ± 0.1 (B)	*0.018*
Insulinemia (µU/ml)	15.1 ± 1.6 (A)	9.3 ± 0.9 (B)	6.2 ± 0.7 (C)	*0.05*
HOMA-IR	0.88 ± 0.09(A)	0.64 ± 0.03(A)	0.78 ± 0.09(A)	NS
Adiponectin (µg/ml)	6.4 ± 0.5 (A)	7.8 ± 0.7 (A)	8.3 ± 0.7 (B)	*0.013*
Plasma lipid homeostasis				
Total cholesterol (mmol/l)	4.54 ± 0.16(A)	4.23 ± 0.16(A)	4.34 ± 0.15(A)	NS
Triglycerides (mmol/l)	1.57 ± 0.19(A)	1.54 ± 0.17(A)	1.48 ± 0.17(A)	NS
HDL-cholesterol (mmol/l)	1.22 ± 0.05(A)	1.17 ± 0.06(A)	1.30 ± 0.06(B)	*<0.001*
Inflammatory markers				
Plasma hsCRP (mg/dl)	3.1 ± 0.8(A)	2.5 ± 0.9(B)	2.7 ± 0.8(B)	*0.023*
Plasma IL-6 (pg/ml)	4.4 ± 0.4(A)	4.2 ± 0.4(A)	3.4 ± 0.4(A)	NS
Plasma Orosomucoid (g/l)	1.02 ± 0.04(A-B)	0.94 ± 0.04(A)	0.86 ± 0.03(B)	*0.008*

Values are expressed as mean ± SEM; (n=30)**.** Fat mass%, Lean mass%: values expressed as a percentage of body weight. (1) Manova stands for Multivariate Analysis of Variance. Data not sharing the same letter within a horizontal line are significantly different (P<0.05).

Table 3: Composition of microbiota compared in lean controls, obese diabetic (OB/D) and non-diabetic (OB/nD) subjects before gastric surgery.

			Firmicutes					***Bacteroidetes***	
	n	All Bacteria [a]	*Clostridium Coccoides* group [b]	*Lactobacillus/ Leuconostoc/ Pediococcus* group [b]	*Clostridium leptum* group [b]	*Fœcalibacterium prausnitzii* species [b c]	*Bifidobacterium* genus [b]	*Bacteroides/ Prevotella* group [b]	*E. coli* species [b]
Control	*13*	11.74 ± 0.1	- 1.58 ± 0.1 (A)	- 3.46 ± 0.2 (A)	- 0.31 ± 0.1 (A)	- 1.06 ± 0.2 (A)	- 2.47 ± 0.4 (A)	- 1.11 ± 0.1 (A)	- 3.43 ± 0.3 (A)
OB/nD	*23*	11.29 ± 0.1	- 1.58 ± 0.2 (A)	- 2.75 ± 0.3 (A)	- 0.86 ± 0.3 (A)	-1.45 ± 0.2 (A)	- 2.37 ± 0.2 (A)	- 1.61 ± 0.1 (B)	- 3.42 ± 0.3 (A)
OB/D	*7*	11.17 ± 0.1	- 1.46 ± 0.4 (A)	- 2.62 ± 0.5 (A)	- 1.63 ± 0.8 (A)	-2.79 ± 0.5 (B)	- 2.22 ± 0.4 (A)	- 1.61 ± 0.2 (B)	- 2.49 ± 0.3 (A)

n : represents the numbers of studied samples.

Data not sharing the same letter within a column are significantly different ($P<0.05$).

[a] : All-bacteria results obtained by qPCR were expressed as mean of the log10 value ± SEM.

[b] : Results were expressed as the mean of the log10 value ± SEM of normalized data, calculated as the log number of targeted bacteria minus the log number of all-bacteria.

[c] : *Faecalibacterium prausnitzii* is the major component of the *Clostridium leptum* group.

Figure 1 :

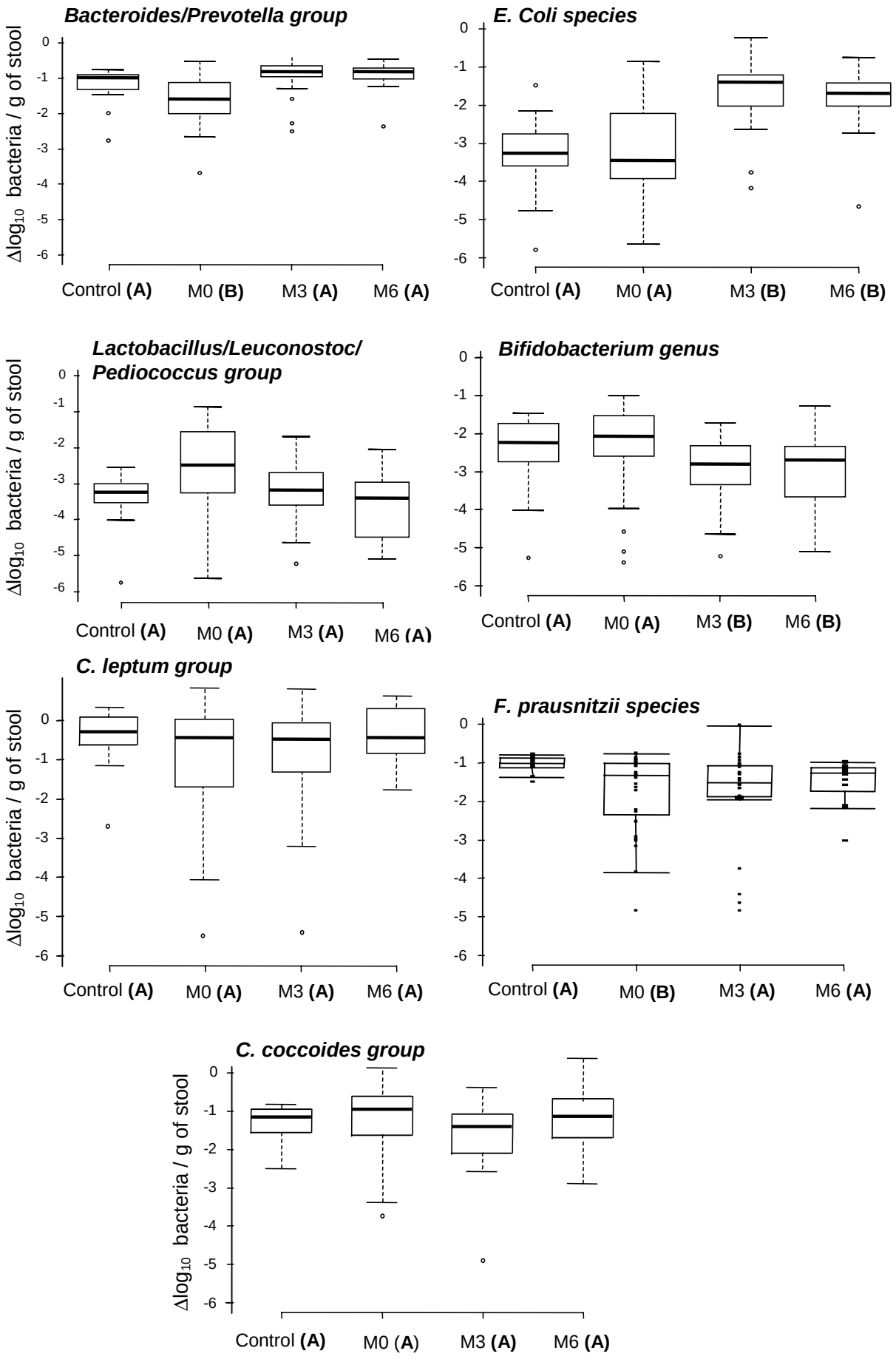

Bacteroides/Prevotella group
E. Coli species
Lactobacillus/Leuconostoc/ Pediococcus group
Bifidobacterium genus
C. leptum group
F. prausnitzii species
C. coccoides group
Δlog10 bacteria / g of stool
Control (A)
M0 (B)
M3 (A)
M6 (A)
Control (A)
M0 (A)
M3 (B)
M6 (B)
Control (A)
M0 (A)
M3 (A)
M6 (A)
Control (A)
M0 (A)
M3 (B)
M6 (B)
Control (A)
M0 (A)
M3 (A)
M6 (A)
Control (A)
M0 (B)
M3 (A)
M6 (A)
Control (A)
M0 (A)
M3 (A)
M6 (A)

Figure 2:

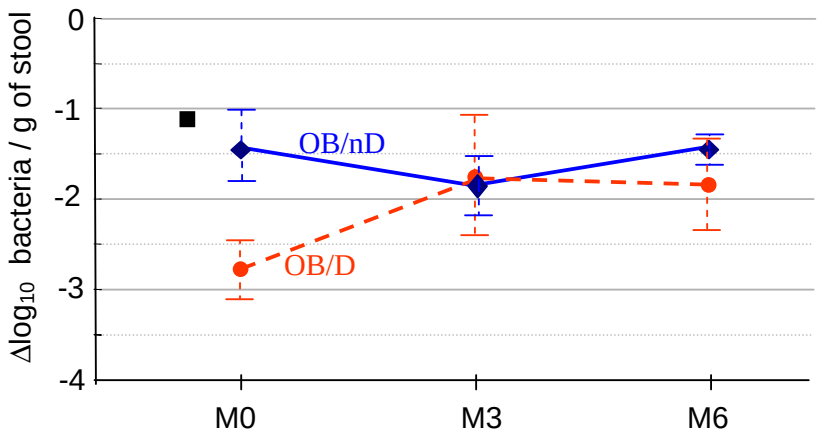


Figure 3a:

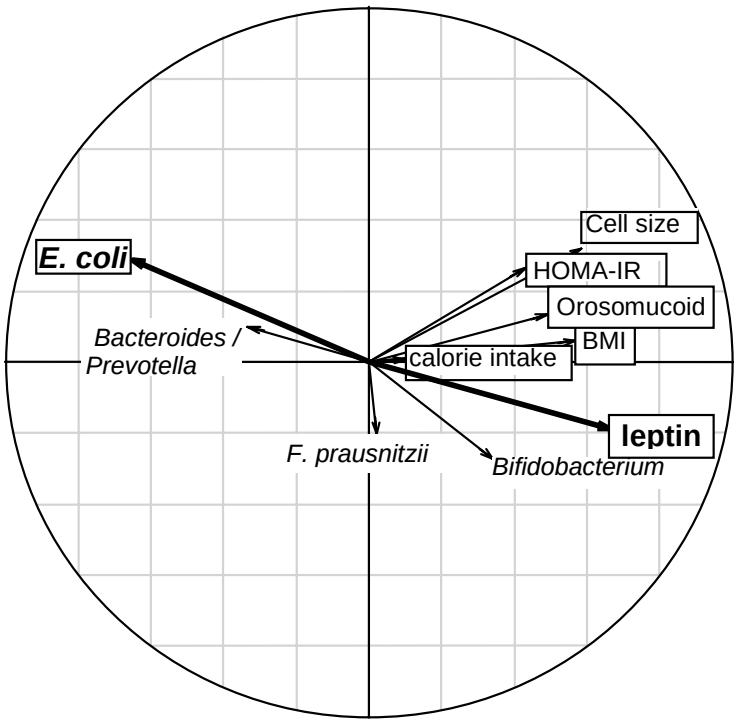
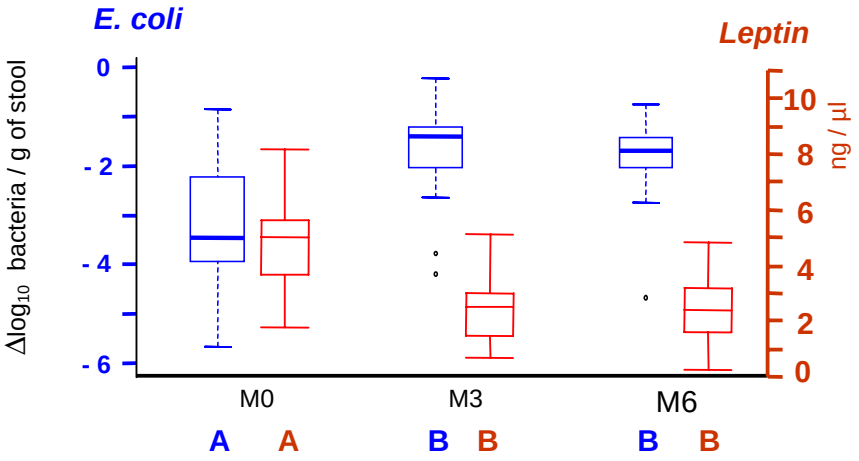


Figure 3b:



Supplementary Table 1: **Group and species-specific 16S rRNA-targeted primers and probes.**

Target organism	*Primer and probe*	*Sequence 5'- 3'*
All-bacteria **(*)**	F_Bact 1369	CGG TGA ATA CGT TCC CGG
	R_Prok1492	TAC GG**C** TAC CTT GTT ACG ACT T
	P_TM1389F	**6FAM-CTT GTA CAC ACC GCC CGT C**
C. leptum	F_Clept 09	CCT TCC GTG CCG SAG TTA
	R_Clept 08	GAA TTA AAC CAC ATA CTC CAC TGC TT
	P-Clep 01	**6FAM-CAC AAT AAG TAA TCC ACC**
Bifidobacterium	F_Bifid 09c	CGG GTG AGT AAT GCG TGA CC
	R_Bifid 06	TGA TAG GAC GCG ACC CCA
	P_Bifid	**6FAM-CTC CTG GAA ACG GGT G**
C. coccoides	F_Ccoc 07	GAC GCC GCG TGA AGG A
	R_Ccoc 14	AGC CCC AGC CTT TCA CAT C
	P_Erec482(*)	**VIC-CGG TAC CTG ACT AAG AAG**
Bacteroides/	F_Bacter 11	CCT WCG ATG GAT AGG GGT T
Prevotella	R_Bacter 08	CAC GCT ACT TGG CTG GTT CAG
	P_Bac303(*)	**VIC-AAG GTC CCC CAC ATT G**
E. coli	E.coli F	CAT GCC GCG TGT ATG AAG AA
	E.coli R	CGG GTA ACG TCA ATG AGC AAA
Lactobacillus/	F_Lacto 05	AGC AGT AGG GAA TCT TCC A
Leuconostoc/ Pediococcus	R_Lacto 04	CGC CAC TGG TGT TCY TCC ATA TA
F. prausnitzii	Fprau 07	CCA TGA ATT GCC TTC AAA ACT GTT
	Fprau 02	GAG CCT CAG CGT CAG TTG GT

(*) modified from reference

Primers and probes used in this study are as previously described [20, 22]. Probe sequences are in bold.

Supplementary Table 2: Evolution of the clinical and biological characteristics of diabetic (OB/D) and non-diabetic (OB/nD) subjects before surgery, 3 months and 6 months after surgery

	OB/nD (n = 23)				OB/D (n = 7)			
	Basal	**3 months**	**6 months**	***P*(1)**	**Basal**	**3 months**	**6 months**	***P*(1)**
Food intake								
Food intake (kcal)	2118 ± 111 (A)	1171 ± 99(A)	1406 ± 55 (B)	*<0.0001*	1469 ± 69 (A)	784 ± 94 (B)	1192 ± 131 (C)	*0.003*
Adiposity markers								
Body Weight (kg)	127.4 ± 4.3 (A)	108.0 ± 4.0 (B)	98.0 ± 3.8 (C)	*<0.0001*	121.5 ± 11.6 (A)	105.3 ± 11.0 (B)	99.3 ± 10.7 (C)	*<0.0001*
BMI (kg/m²)	48.3 ± 1.6 (A)	40.9 ± 1.4 (B)	37.2 ± 1.4 (C)	*<0.0001*	45.4 ± 3.5 (A)	39.3 ± 3.4 (B)	36.9 ± 3.2 (C)	*<0.0001*
Adipocyte diameter (µm)	115.6 ± 1.7 (A)	117.8 ± 2.7 (A)	101.6 ± 3.9 (B)	*0.002*	120.1 ± 3.4(A)	107.4 ± 1.3(A)	109.3 ± 3.7(A)	NS
REE (kcal)	1830.3 ± 65.2 (A)	1859.5 ± 58.2 (B)	1566.4 ± 49.3 (C)	*<0.0001*	1762.1 ± 85.9 (A)	1887.3 ± 152.1 (B)	1492.2 ± 80.1 (C)	*0.014*
Fat mass%	48.2 ± 1.1 (A)	44.9 ± 1.1 (B)	41.3 ± 1.4 (C)	*<0.0001*	46.6 ± 2.1 (A)	43.3 ± 2.5 (B)	41.4 ± 2.3 (C)	*<0.001*
Lean mass%	49.7 ± 1.1 (A)	52.7 ± 1.0 (B)	56.0 ± 1.3 (C)	*<0.0001*	50.9 ± 1.9 (A)	53.9 ± 2.3 (B)	55.6 ± 2.0 (C)	*0.001*
Leptin (ng/ml)	50.5 ± 4.1 (A)	24.1 ± 2.5 (B)	23.0 ± 2.6 (B)	*<0.0001*	52.2 ± 8.3 (A)	30.0 ± 6.4 (B)	30.4 ± 8.6 (B)	*0.041*
Plasma Glucose homeostasis and insulin sensitivity								
Glycaemia (mmol/l)	5.4 ± 0.2 (A)	4.8 ± 0.1 (B)	4.6 ± 0.1 (B)	*<0.001*	9.6 ± 1.7(A)	5.6 ± 0.6(A)	5.5 ± 0.6(A)	NS (2)
HbA1C (%)	5.8 ± 0.1 (A)	5.6 ± 0.1 (B)	5.6 ± 0.1 (B)	*0.021*	8.5 ± 1.3(A)	6.4 ± 0.3(A)	6.4 ± 0.2(A)	NS (2)

Insulinemia (mmol/l)	16.6 ± 1.8 (A)	9.5 ± 1.0 (B)	6.3 ± 0.8 (C)	*<0.0001*	8.4 ± 1.7(A)	8.0 ± 1.4(A)	5.6 ± 0.7(A)	NS (2)
HOMA-B%	67.0 ± 2.8 (A)	77.3 ± 2.6 (B)	89.8 ± 6.0 (C)	*<0.001*	40.3 ± 8.4(A)	65.8 ± 7.7(A)	74.2 ± 15.0(A)	NS (2)
HOMA-S%	145.4 ± 6.3 (A)	168.6 ± 5.7 (B)	180.3 ± 21.3 (B)	*<0.001*	81.5 ± 20.7(A)	142.7 ± 17.3(A)	148.7 ± 25.7(A)	NS (2)
HOMA-IR	0.7 ± 0.04 (A)	0.6 ± 0.02 (B)	0.8 ± 0.11 (A)	*<0.001*	1.6 ± 0.40(A)	0.8 ± 0.10(A)	0.7 ± 0.10(A)	NS (2)
Adiponectin (µg/ml)	6.8 ± 0.6 (A)	7.5 ± 0.7 (B)	8.3 ± 0.7 (B)	*0.015*	5.4 ± 0.8(A)	8.7 ± 1.9(A)	8.4 ± 2.3(A)	NS
Plasma lipid homeostasis								
Total cholesterol (mmol/l)	4.53 ± 0.18(A)	4.23 ± 0.19(A)	4.31 ± 0.17(A)	NS	4.59 ± 0.31 (AB)	4.23 ± 0.23 (A)	4.47 ± 0.26 (B)	*0.037*
Triglycerides (mmol/l)	1.50 ± 0.19(A)	1.47 ± 0.18(A)	1.42 ± 0.16(A)	NS	1.81 ± 0.52(A)	1.77 ± 0.47(A)	1.70 ± 0.55(A)	NS
HDL cholesterol (mmol/l)	1.21±0.06 (AB)	1.15 ± 0.06 (A)	1.28 ± 0.06 (B)	*<0.001*	1.27 ± 0.1(A)	1.23 ± 0.18(A)	1.36 ± 0.2(A)	NS
Inflammatory markers								
hsCRP (mg/dl)	2.7 ± 0.8(A)	2.9 ± 1.2(A)	2.9 ± 1.0(A)	NS	4.4 ± 2.4(A)	1.2 ± 0.6(A)	1.9 ± 0.7(A)	NS
IL-6 (pg/ml)	4 ± 0.4(A)	4.3 ± 0.4(A)	3.7 ± 0.4(A)	NS	5.8 ± 1.3(A)	3.8 ± 0.6(A)	2.7 ± 0.5(A)	NS
Orosomucoide (g/l)	1.05 ± 0.05 (A)	0.94 ± 0.04 (AB)	0.88 ± 0.04 (B)	*0.015*	0.95 ± 0.08(A)	0.92 ± 0.09(A)	0.81 ± 0.05(A)	NS

All values are expressed as mean ± SEM;
REE: (Resting Energy Expenditure).
Fat mass%, Lean mass%: values expressed as a percentage of body weight.
HOMA: homeostasis model assessment, HOMA-B%: estimate the pancreatic β-cell function; HOMA-S%: evaluate the insulin sensitivity and HOMA-IR: represent the insulin resistance.
(1) Manova stands for Multivariate Analysis of Variance.
(2) Comparison in 5 diabetic subjects without insulin injection due to the influence of the insulin treatment for these parameters.
Data not sharing the same letter within a horizontal line are significantly different (P<0.05).

Supplementary Table 3: **Composition of microbiota compared in diabetic (OB/D) and non-diabetic (OB/nD) subjects before surgery, and 3 months and 6 months after surgery.**

		Firmicutes					*Bacteroidetes*	
	All bacteria [a]	*Clostridium coccoides* group [b]	*Lactobacillus/ Leuconostoc/ Pediococcus* group [b]	*Clostridium leptum* group [b]	*Fecalibacterium prausnitzii* species [b, c]	*Bifido-bacterium* genus [b]	*Bacteroides / Prevotella* group [b]	*E. coli* species [b]
OB/nD								
M0 (23)	11.29 ± 0.1	- 1.58 ± 0.2 (A)	- 2.75 ± 0.3 (A)	- 0.86 ± 0.3 (A)	- 1.45 ± 0.2 (A)	- 2.37 ± 0.2 (A)	- 1.61 ± 0.1 (A)	- 3.42 ± 0.3 (A)
M3 (20)	11.00 ± 0.2	- 1.90 ± 0.1 (A)	- 3.26 ± 0.2 (B)	- 1.04 ± 0.3 (A)	- 1.85 ± 0.2 (B)	- 3.10 ± 0.2 (B)	- 0.91 ± 0.1 (B)	- 1.76 ± 0.2 (B)
M6 (11)	11.09 ± 0.2	- 1.50 ± 0.3 (A)	- 3.57 ± 0.3 (B)	- 0.37 ± 0.2 (A)	- 1.40 ± 0.1 (A)	- 2.85 ± 0.3 (B)	- 1.02 ± 0.2 (B)	- 2.13 ± 0.3 (A)
OB/D								
M0 (7)	11.17 ± 0.1	- 1.46 ± 0.4 (A)	- 2.62 ± 0.5 (A)	- 1.63 ± 0.8 (A)	- 2.79 ± 0.5 (B)	- 2.22 ± 0.4 (A)	- 1.61 ± 0.2 (A)	- 2.49 ± 0.3 (A)
M3 (6)	10.73 ± 0.4	- 1.92 ± 0.7 (A)	- 2.81 ± 0.3 (A)	- 0.69 ± 0.5 (A)	- 1.78 ± 0.7 (B)	- 2.45 ± 0.2 (A)	- 0.86 ± 0.4 (A)	- 1.18 ± 0.3 (A)
M6 (4)	10.99 ± 0.2	- 1.70 ± 0.2 (A)	- 3.63 ± 0.4 (A)	- 0.57 ± 0.3 (A)	- 1.82 ± 0.4 (B)	- 3.06 ± 0.7 (A)	- 0.75 ± 0.1 (A)	- 1.27 ± 0.2 (A)

Number in parentheses represents the numbers of studied samples.

Control: healthy female volunteers, OB/nD: obese subjects without diabetes, OB/D: obese subjects with diabetes,

Data not sharing the same letter within a column are significantly different ($P<0.05$) to the human population.

[a] All-bacteria results obtained by qPCR were expressed as mean of the log10 value ± SEM.

[b] Results were expressed as mean of the log10 value ± SEM of normalized data, calculated as the log no. of targeted bacteria minus the log of all-bacteria number.

[c] : *Faecalibacterium prausnitzii* is the major component of the *Clostridium leptum* group.


16S
rRNA
rarefaction
length
index
richness
composition
cut-off
OTUs
assigned
taxa
regions
obtained
pair
optimal
sequence-identity
structure
order
Tree
region
respective
used
Bacteroidetes
RDP-classifier
Firmicutes
proportions
strongly
amplify
Shannon
profiles
reads
median
genus
class
However
gene
sample
bacterial
pV12A
e.g
directly
Unknown
Figure
variable
community
families
replicates
quality
genes
gut
V4
based
Ley
observed
pV6
Sogin
may
applied
shown
significantly
MICROBIAL
different
Supplementary
Lachnospiraceae
allows
primer
phylogenetic
pairs
Taxonomic
similar
low
remaining
conservation
PCR
number
abundant
Wang
full-length
five
Liu
SAMPLES
family
taxon
pV4
Huse
analysis
underrepresented
USING
Hamady
belonging
hence
generated
amplified
COMMUNITIES
taxonomy
depicted
average
rank
sequencing
high
DNA
diversity
groups
known
one
sequences
primers
ribosomal
annealing
Knight
Actinobacteria
silico
microbiota
wide
bacteria
assessed
OTU
fraction
pV12B
well
eight
HUMAN
PYROSEQUENCING
Microbiome
indicate
Table
approach
sequence
bias
mixture
pV3B
range
higher
Results
two
ranks
1A4B
also
reference
amplicons
Turnbaugh
phylum
software
amplification
short
compared
PROFILING

Title

PROFILING MICROBIAL COMMUNITIES USING MULTIPLEX PYROSEQUENCING: A VALIDATION STUDY ON HUMAN GI-TRACT SAMPLES

Lutz Krause[1], Deborah Moine[1], Andreas Rytz[1], Marion Leclerc[2], Joel Doré[2], Julien Tap[2], Fabrizio Arigoni[1], Bernard Berger[1]*

[1]Nestlé Research Center, Nestec Ltd., 1000 Lausanne 26, Switzerland

[2]INRA, UEPSD, UR910, 78350 Jouy en Josas, France

*Corresponding author

Abstract

Introduction
Using high-throughput pyrosequencing of 16S ribosomal genes, the community structure of microbial samples can be characterized to an extent that one would never have dreamed of a decade ago. The multiplexing of pyrosequencing by DNA-barcoding techniques allows the simultaneous characterization of hundreds of samples at low cost. However, despite the wide application of this technique, there is no consensus about the "optimal" primer pairs for the unbiased amplification of the 16S DNA gene and the "optimal" variable region to be targeted.

Results
A validation study of multiplex pyrosequencing of 16S ribosomal genes has been undertaken. We have analyzed 35 'universal' primers for the amplification of 16S DNA and selected and optimized the eight most promising ones. Eight setups are presented addressing the question of the targeted variable region of the 16S gene, the so-called 'universality' of PCR primers as well as practical aspects linked to the preparation of PCR samples. We have also investigated if sequence-identity cut-offs that are widely used for grouping 16S DNA into Operational Taxonomic Units (OTUs) are also adequate for the grouping short 16S amplicons.

Conclusions
The choice of primers and targeted 16S region can strongly affect sequence-based community analysis. We recommend the usage of two primer pairs, one targeting the V1 and V2 variable regions and one targeting the V4 region. Our results further indicate that community structure profiles and diversity measures cannot be directly compared, if different PCR primers are used for the amplification of the 16S gene. Moreover, despite that OTUs are valuable for analyzing the structure of microbial communities, OTUs generated from different variable regions cannot be directly compared and generated OTUs do not generally correspond to existing taxonomies.

Author Summary

Introduction

Prodigious advances in sequencing technologies allow the development of new approaches for studying microbial ecology. A new era of exploring microbial diversity was initiated in 1977 when Carl Woese and colleagues assessed the evolutionary relationships of organisms based on the sequence analysis of small-subunit ribosomal RNA (16s RNA and 18s RNA genes) (Woese and Fox, 1977). Today, 16S DNA-based approaches are the gold standard for analyzing phylogenetic relationships between bacteria. Since the pioneering work of Norman Pace and colleagues in the mid 1980s, our knowledge about the diversity of microbes has largely been increased by directly isolating RNA genes from the environment, followed by their phylogenetic characterization (Hugenholtz and Goebel, 1998;Hugenholtz, 2002;Pace et al., 1985). In particular, this approach allows the detection of microbial groups that resist cultivation and hence provides a window into the uncultured majority.

Microbial community profiling based on the sequencing of 16S ribosomal genes is currently undergoing a renaissance (Tringe and Hugenholtz, 2008) owing to the high-throughput power and affordable price of 454 pyrosequencing. Compared to traditional Sanger sequencing, pyrosequencing is at least an order of magnitude cheaper and faster, and does not require the laborious step of preparing a clone library. The higher throughput allows a much deeper coverage of microbial communities and hence also the detection of rare organisms. For 16S DNA pyrosequencing, at present mainly the Genome Sequencer FLX (GS FLX) is employed, which provides sequence reads with averaging length of 250bp. In contrast, the traditional Sanger technique produces sequence reads with an average length of 600bp. However, Liu et al. and Wang et al. (Liu et al., 2007;Liu et al., 2008;Wang et al., 2007) have demonstrated that the same conclusions can be drawn from 250bp as from full-length sequences and that the benefit of large numbers of short reads outweights the drawback of short sequence length (Liu et al., 2007).

Using barcoding techniques, multiplex pyrosequencing allows the characterization of hundreds of samples in parallel. In this approach, one of the PCR primers used to amplify 16S DNA is labeled with a short, sample-specific sequence key of few bp, called barcode (Binladen et al., 2007;Hoffmann et al., 2007). After the PCR amplification, 16S DNA from multiple samples is pooled in equal amounts and simultaneously sequenced in a single pyrosequencing run. The original sample of each resulting sequence read is determined based on its specific barcode using simple computer programs. Error corrected barcodes can be employed that allow the correction of up to three sequencing errors (Fierer et al., 2008;Hamady et al., 2008).

Multiplex pyrosequencing of 16S DNA is providing striking insights into the composition of microbial communities from disparate environments, including the deap sea (Huber et al., 2007;Sogin et al., 2006), soil (Jones et al., 2009;Lauber et al., 2009), air (Bowers et al., 2009), human mouth (Keijser et al., 2008) or skin (Fierer et al., 2008;Grice et al., 2008;Grice et al., 2009). In particular, during the last five years, the scientific community had an increasing interest in studying the composition of the human gut microbiota (Andersson et al., 2008;Antonopoulos et al., 2009;Chakravorty

et al., 2007;Dethlefsen et al., 2008;Ley et al., 2008;McKenna et al., 2008;Tap et al., 2009;Turnbaugh et al., 2009), which is also reflected by the launch of the NIH Human Microbiome Project (Turnbaugh et al., 2007), MetaHIT (EU and China), MicrOBES (INRA France), Meta-GUT (China) and the Canadian Microbiome Initiative (Canada). Key questions addressed using 16S DNA sequencing are: What is the composition of the human gut microbiota and its impact on our well-being, health and disease (e.g. in the context of nutrition, obesity, diabetes, IBD, or cancer)? What is a 'normal' composition? Are changes in relative abundance of the human gut microbiota generally important (Hamady and Knight, 2009)? Do we have a core gut microbiome (Tap et al., 2009;Tschop et al., 2009;Turnbaugh et al., 2009)? How does the gut microbiota change and evolve over time (Ley et al., 2008)? What are the (long-term) effects of antibiotics (Antonopoulos et al., 2009;Dethlefsen et al., 2008)? For example, recent experiments conducted by Jeffrey Gordon's group have suggested that the human gut microbiome plays a role in the development of obesity (Ley et al., 2005;Ley et al., 2006;Turnbaugh et al., 2006;Turnbaugh et al., 2009). Furthermore, 16S DNA studies have revealed that the human gut-microbiota is highly diverse, and composed of mainly rare organisms (Turnbaugh et al., 2009). Consequently, deep sequencing techniques are required to detect the majority of underrepresented microbial groups.

Despite the wide application of 16S DNA pyrosequencing for profiling microbial communities, there is no consensus about the "optimal" PCR primers and "optimal" 16S DNA region and different groups use different primers targeting different variable regions. In this study, we have selected and optimized PCR primers for the amplification of 16S DNA surrounding different variable regions and we have investigated the effects of different primer pairs on the obtained community profiles. Due to the wide interest in studying the human gut microbiome, these questions were addressed on a mixture of fecal samples from one human adult and four babies, which in the following is named *1A4B mixture*. Adult and four baby fecal samples were pooled together to account for age depending differences and the high variability of the infant gut microbiota (Palmer et al., 2007). We have further investigated if the community structure profiles obtained by 16S DNA pyrosequencing are comparable to the gold standard (clone library sequenced using the Sanger technique). Finally, we have investigated if sequence-identity cut-offs used to group full-length 16S DNAs into so called Operational Taxonomic Units (OTUs) are also appropriate for the grouping of short 16S amplicons. This question was addressed *in silico* on full-length 16S gene sequences downloaded from the ARB database.

Results

Selection of primers for the amplification of 16S variable regions

Over the last decades, a number of primers for the amplification of ribosomal 16S genes have been published. Primers were designed to target specific bacterial taxa, whereas others were referred to as 'universal' since their sequences matched most if not all known 16S sequences. However, the number of 16S sequences in the databases is on the increase, and the once-'universal' primers may miss the newly discovered taxa.

In order to pre-select good primers to comprehensively describe the complex human gut microbiota, we analysed *in silico* the annealing properties of 35 primers targeting the most conserved regions of the 16S gene (Table 1). Using the Probe Match software (Kim et al., 2009) with a tolerance of zero to two mismatches per primer, we assessed the level of bacterial community coverage based on the type strains of RDP (Cole et al., 2003). Solely the type strains were considered to prevent bias due to over-representation of certain species in the database. Despite of a generally good matching of all 35 primers, the percentage of species showing a perfect match varied considerably. As the detrimental effect of a single mismatch on the annealing efficiency highly depends on its position within the primer (Bru et al., 2008;Hongoh et al., 2003;Sipos et al., 2007), only the most stringent condition (zero mismatch) allowed a sound ranking of these primers for the 'universal' amplification of the 16S gene. Interestingly, amongst the most universal primers were the longest ones showing no or a limited number of degenerated positions (e.g. BSF517, Universal R500), which would favour efficient and specific PCR amplification.

In addition to the evaluation of annealing properties, the characteristics of the 16S variable regions amplified by the primers were considered. *In silico* analyses reported that sequences spanning the variable regions V2 and V4 were the most accurate for taxonomical assignment, with satisfactory results for V3 and V7+V8 (Liu et al., 2008;Wang et al., 2007). Although sequences encompassing V6 or V5+V6 have frequently been used in pyrosequencing (Andersson et al., 2008;Antonopoulos et al., 2009;Dethlefsen et al., 2008;Sogin et al., 2006;Turnbaugh et al., 2009), it has been documented that they provide only a limited taxonomic resolution (Huse et al., 2008;Liu et al., 2007;Turnbaugh et al., 2009;Wang et al., 2007). For community clustering and species richness estimates, the fragments encompassing V2 or V4 (Liu et al., 2007) and V4, V5+V6, or V6+V7 (Youssef et al., 2009) were recommended.

All comparative studies quoted above were only based on *in silico* simulations of pyrosequencing datasets, except (Huse et al., 2008) and (Turnbaugh et al., 2009) comparing amplicons of V3 versus V6, and V2 versus V6, respectively. As some biases of primers may not be detected by the computational approaches, we did not restrict the targeted regions to the ones that were most promising according to *in silico* analysis (V2 and V4). The variable regions V3 and V7+V8 were also included, as well as an extended V6 region of twice the size of the highly variable region generally considered. The primers were selected based on table 1 and paired such that amplicons of about 250 bp spanning V1+V2, V3, V4, V6, and V7+V8 were amplified (Figure 1). For V1+V2, three mixes of primers were tested. The pair pV12A was composed of the widely used forward primer 8F and the reverse primer BSR357. Since the 8F primer is limited in the amplification of Bifidobacteria (Suau et al., 1999), pV12B included a cocktail of oligos in order to accommodate the vast majority of 8F/27f variations, as proposed by (Frank et al., 2008). In total, eight primer pairs were tested (Figure 1 and Suppl. Table 1).

Optimization of PCR conditions

The nucleotide sequence of the chosen primers is of primary importance for the non-biased amplification of 16S genes present in a bacterial community. However, to increase the diversity of annealed sequences, the stringency of non-perfectly matching

primers can be modified by lowering the annealing temperature. For each primer pair, a gradient of PCR annealing temperatures was tested (data not shown) and the least stringent temperature which still produced a sufficient amount of product was determined (Suppl. Table 1) and used for subsequent work.

Sequencing results

Variable 16S regions were amplified from the 1A4B DNA mixture using eight different primer pairs (pV12A, pV12B, pV12C, pV3A, pV3B, pV4, pV6, pV78). For each primer pair, 16S DNA was amplified from the 1A4B mixture in five separate polymerase chain reactions (PCRs) and hence for each primer pair five replicates were obtained. In total, 40 PCRs were conducted (8 primer pairs x 5 replicates). The resulting PCR samples were sequenced in a single GS FLX pyrosequencer run in parallel with 262 PCR samples extracted in other experiments. The single pyrosequencing run yielded 568,796 sequence reads with averaging length of 250bp. To avoid wrong assignments of 16S sequences during the taxonomic classification step, 61,479 (11%) low quality reads that likely contain sequencing errors were excluded from further analysis. As a result, in average each of the 302 PCR samples was represented by 1,680 high quality reads (Supplementary figure 5). Only 30 out of the 302 PCR samples were represented by less than 1,000 reads. The 16S DNA of all these problematic samples was amplified using one of the following primer pairs: pV3A, pV3B or pV78.

Different primer pairs may lead to different biological conclusions

All PCR primers are more or less biased owing to differential annealing which leads to the over- or under-representation of specific microbial groups, and some taxa can be entirely missed (Wintzingerode 19997 and Kanagawa 2003). The effects of different PCR primers on the profiling of microbial communities using 16S DNA pyrosequencing was assessed for eight primer pairs (pV12A, pV12B, pV12C, pV3A, pV3B, pV4, pV6, pV78) surrounding seven different variable 16S regions (V1, V2, V3, V4, V6, V7 and V8). This question was addressed on a mixture of fecal samples from one human adult and four babies (1A4B mixture). The results are summarized in Figure 2, Figure 5 and Supplementary figure 1.

In general, reproducible community structure profiles were obtained for the five replicates of each primer pair (Figure 2 and Supplementary figure 1). At high taxonomic rank (phylum), the community structure profiles obtained for the eight different primer pairs were also similar, with the exception of pV6. 16S sequences were assigned to only seven different phyla, the most abundant ones were Firmicutes, Bacteroidetes and Proteobacteria. Only with pV6 Bacteroidetes were almost entirely missed. Also at lower taxonomic rank similar community profiles were obtained for the eight different primer pairs, again with the exception of pV6 (Figure 2 and Supplementary figure 1).

Despite that in general the community structure profiles for seven of the eight primer pairs were similar, the statistical analysis revealed that different primer pairs may lead to significantly different results. The main differences were in the proportions of Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria (Figure 2, Figure 3 and Figure 5). Proportions of Firmicutes ranged from 75% to 43%, proportions of

Bacteroidetes from 41% to ~1%, proportions of Proteobacteria from 18% to 8% and proportions of Actinobacteria from 8% to 0.9%.

The community profiles were also significantly different at lower taxonomic ranks (Figure 2, Figure 4 and Supplementary figure 1). At family rank, the most pronounced differences were observed for the "Unknown" category (18% to 2% of 16S sequences), Lachnospiraceae (25%-4%), Prevotellaceae (20%-1%), Enterococcaceae (8%-1%), Bacteroidaceae (14%-0.3%), Erysipelotrichaceae (30%-13%), Porphyromonadaceae (10%-0.3%), Streptococcaceae (9%-0.2%), and Bifidobacteriaceae (8%-1%).

These results demonstrate that sequence-based profiling of microbial communities may provide significantly different results, depending on the choice of primer pairs used to amplify the 16S gene.

pV12 primer pairs

Community profiles obtained for pV12A, pV12B and pV12C were strongly enriched for Bacteroidetes, but therefore proportions of Firmicutes were low (particularly for pV12C) (Figure 3). Additionally, Actinobacteria were strongly underrepresented for pV12A and Proteobacteria for both pV12B and pV12C.

At family rank, several taxonomic groups belonging to Bacteroidetes were significantly more prevalent, but taxa belonging to Firmicutes were generally underrepresented (Figure 4). In particular, Bifidobacteriaceae were strongly underrepresented for pV12A and Streptococcaceae for pV12C. pV12B had the smallest differences from the median over all primer pairs, when compared to pV12A and pV12B. Overall, these results suggest that the pV12 primer pairs are modestly biased in the amplification of 16S DNA.

pV3 primer pairs

In general, pV3A and pV3B had the smallest difference from the median over all eight primer pairs. At phylum rank, only two categories (Actinobacteria and "Unknown") were significantly enriched. The overrepresentation of the "Unknown" category indicates that a high proportion of 16S DNA from so far uncharacterized Bacteria was amplified. On the other hand, no phylum was significantly underrepresented. At family rank, only two taxa were significantly enriched. For pV3A only two families were significantly underrepresented and only one for pV3B.

pV4 primer pair

For pV4 only small differences from the median over all eight primer pairs were observed, suggesting that also pV4 has a low amplification bias. Three phyla (Actinobacteria, Firmicutes and Proteobacteria) were significantly overrepresented; proportions of two phyla (Bacteroidetes and "Unknown") were significantly low. At family rank, several taxa belonging to the Firmicutes were significantly overrepresented and only two taxa were significantly underrepresented.

pV6 primer pair

The pV6 community structure profiles were strongly enriched for Firmicutes (76% of sequences), on the cost of low proportions of Bacteroidetes. The median for

Bacteroidetes over all eight primer pairs was 25%, but only ~1% of the pV6 sequences was assigned to this phylum. Also Proteobacteria were strongly overrepresented. At family rank, a high number of taxa (6 out of 31) was significantly more abundant, in particular those belonging to Firmicutes. Strikingly, a high fraction of sequences (~19%) was assigned to the "Unknown" category, while at family rank the median for the "Unknown" over all eight primer pairs was 4%. Seven families were significantly underrepresented. Conversely, the results indicate that pV6 is strongly biased but therefore may allow the amplification of 16S DNA from so far uncharacterized Bacteria. A phylogenetic analysis revealed that the pV6 amplicons assigned to the "Unknown" category represent different phylogenetic groups of bacteria (Supplementary figure 2).

pV78

The pV78 community profiles were strongly enriched for Firmicutes (18% higher than the median over all primer pairs). Actinobacteria and Bacteroidetes were significantly underrepresented. Several families belonging to the Firmicutes were significantly more abundant, but therefore families belonging to the Bacteroidetes were generally underrepresented.

Highest number of phylogenetic groups covered by pV12B and pV4

The universality of the eight studied primer pairs, i.e. their ability to efficiently amplify 16S DNA from a wide range of bacterial groups, was also assessed on the 1A4B mixture. We regarded the family rank as a good tradeoff between resolution and proportion of sequences that could be assigned to known taxa. In total, 31 different families were observed in the community profiles of all 8 primer pairs. The number of families (= family richness) that would be observed for different sample sizes was estimated in a rarefaction analysis (Figure 6). The number of observed taxa strongly depends on the used PCR primers. The highest number of families was estimated for pV3B and pV4, followed by pV12A and pV12B. Similar observations were made in a rarefaction analysis of the Shannon index (measuring the overall diversity of the underlying community) (Supplementary figure 3). The highest Shannon index was measured for pV3B and pV4, followed by pV12B and pV12A. Richness and Shannon index measures for the remaining ranks (phylum, class, order, and genus) were in accordance (data not shown). These results indicate that with pV3B, pV4, and pV12B efficiently amplify 16S DNA from a wide phylogenetic range of bacteria.

Similar results obtained from short 16S amplicons as from full length sequences

We further assessed if the same conclusions can be drawn from short 16S amplicons as from full length sequences. This was addressed on a fecal microbial sample from one human adult, which is described in more detail in (Tap et al., 2009). For this sample, 679 (nearly) full-length 16S sequences were available, which were obtained using the "gold standard", i.e. by sequencing a clone library using the Sanger technique (Tap et al., 2009). In this experiment, only pV4 and pV12B were considered, as these were the most promising candidates according to the results presented above. For each primer pair, five separate PCR extractions were done and hence five replicates were available. For each PCR sample between 1,389 and 2,203

16S sequences were obtained, with an average of 1,877 sequences per sample. Reproducible results were obtained for the five replicates (Supplementary figure 7).

In general, the pV4 and pV12B community profiles were in accordance with the ones obtained for the full-length sequences (Figure 8). At ranks phylum to family, a high correlation coefficient was measured in the range between 0.91 and 0.99 (Supplementary table 2). Only at rank genus, the correlation coefficient was modest (0.70-0.74 for pV12B and 0.80-0.87 for pV4).

The modest correlation coefficient at rank genus can be explained by the varying number of sequences assigned to the genus *Lachnospiraceae Incertae Sedsi*, which belongs to the *Lachnospiraceae* family. If *Lachnospiraceae Incertae Sedis* was excluded, the correlation coefficient at rank genus increased considerably to 0.87-0.93 for pV12B and 0.88-0.91 for pV4. At rank family, a large fraction of sequences was assigned to *Lachnospiraceae* (35% of full-length sequences and in average 38% of the pV12B and 39% of the pV4 sequences) (Supplementary figure 6). At rank genus, 62% of the full-length sequences that were assigned to *Lachnospiraceae* at rank family were assigned to *Lachnospiraceae Incertae Sedis*. But the majority of the pV12B and pV4 amplicons that were assigned to *Lachnospiraceae* at rank family were assigned to "Unclassified" at rank genus (87% for pV12B and 82% for pV4). Presumably, owing to the short sequence length, the RDP-classifier was unable to infer the genus of most of the pV12B and pV4 amplicons belonging to *Lachnospiraceae Incertae Sedis*.

Variable regions strongly differ in sequence conservation

The grouping of 16S sequences into Operational Taxonomic Units (OTUs) is widely used to analyze the composition and diversity of microbial communities. In this approach, similar 16S sequences, i.e. those with a high pair-wise sequence identity, are grouped into one OTU. The main rational behind this approach is that each generated OTU represents one phylogenetic group of the original community. Different sequence identity cut-offs are used to defined OTUs, such as 80% at phylum. A cut-off of 97% is widely applied to estimate the number of phylotypes present in the original community (Antonopoulos et al., 2009;Fierer et al., 2008;Grice et al., 2008;Huber et al., 2007;McKenna et al., 2008;Tap et al., 2009;Turnbaugh et al., 2009). Here, we investigated if these widely applied sequence-identity cut-offs are also adequate for the grouping of short 16S amplicons.

The sequence conservation of variable 16S regions as defined by the eight primer pairs varies considerably (Figure 7). For example, 80% of the pV78-sequence-pairs that originate from the same phylum have a sequence-identify above 20%. Contrastingly, only 25% of the pV12C-sequence-pairs that originate from the same phylum had a sequence-identity above 20%. These results indicate a much higher conservation of the pV78-region when compared to the region amplified by the pV12C primer pair. Similar observations were made at lower taxonomic ranks (Figure 7). Moreover, the amplified regions show a taxa-dependent sequence conservation (Supplementary figure 4). For example, 90% of the pV12A-sequence-pairs that originated from Actinobacteria had a pair-wise sequence identity above 90%. But only 10% of the pV12A-sequence-pairs that originated from Bacteroidetes had a pair-wise sequence identity above 20%.

These results indicate that a general sequence-identity cut-off does not exist to group variable 16S regions into OTUs, such that each OTU corresponds to a taxon of an established taxonomy, e.g. the NCBI taxonomy. Furthermore, OTU-based results cannot be directly compared if they were obtained using PCR primers targeting different variable 16S regions. Overall, an OTU analysis is very valuable for assessing the diversity and composition of microbial communities, but the nonuniform sequence conservation of the 16 gene has to be taken into account.

Discussions and Outlook

High-throughput pyrosequencing of the 16S DNA gene allows the rapid profiling of microbial communities at low cost. If the PCR primers are carefully selected, the same conclusions can be drawn as from full length sequences, as demonstrated herein as well as by Liu et al. 2007 and Wang et al. 2008. Moreover, compared to the traditional clone library approach, pyrosequencing of the 16S DNA gene is several folds faster and cheaper, while at the same time avoiding a cloning bias. The drawback of short read length can be bypassed by selecting primer pairs targeting an optimal region of the 16S DNA gene which can accurately be assigned to a taxonomy (Liu et al., 2007;Liu et al., 2008;Wang et al., 2007). Moreover, a high number of short 16S amplicons outweights the short read length when studying the composition of microbial (Liu et al., 2007;Liu et al., 2008;Wang et al., 2007). One remaining drawback is that many existing computational tools for the analysis of 16S DNA are not suited for the characterization of vast numbers of short 16S amplicons. However, recently several computational tools have explicitly been devised for this task, e.g. (Cole et al., 2003;Hamady et al., 2009;Sun et al., 2009;Wang et al., 2007) and presumably most of the computational challenges will soon be resolved.

We have selected and optimized eight universal PCR primer pairs and investigated there usability for the profiling of microbial communities using high-throughput 16S DNA sequencing. Our results demonstrate that different primers can lead to considerably different community profiles and hence to different biological conclusions. In particular, the relative abundance of taxa, the number of detected taxa and hence also the measured biodiversity and richness are affected. Our results further demonstrate that community profiles and biodiversity measures obtained using different PCR primer pairs cannot be directly compared. A careful primer selection is crucial and may depend on the taxa of interest and ecological niche under study.

We recommend the usage of both the pV12B and pV4 primer pairs in parallel. First, the community profiles obtained for these primer pairs were only moderately different from the mean over all primer pairs, indicating a low amplification bias. Second, the rarefaction analysis revealed that pV12B and pV4 well amplify the 16S DNA gene from a broad range of bacterial groups, indicating a high universality. This is inline with *in silico* predictions obtained using the Probe Match software. Third, with these primer pairs we have achieved high amplification rates and did not encounter problems of dimer formation. Fourth, both primer pairs are well suited for the amplification of Bifidobacteria, which are of particular interested when analyzing the human gut microbiota and which are missed by some of the widely used primers (Suau et al., 1999). Fifth, the pV12B and pV4 community profiles were in accordance with the gold standard, i.e. the full-length sequencing of the 16S DNA gene using a clone-library approach. Finally, the comparison to full-length sequences indicated that pV12B and pV4 amplicons are accurately and consistently assigned to a taxonomy by

the RDP-Classifier. This is in accordance with Wang 2007 and Liu 2008 who reported that the V2 and V4 variable regions give the highest accuracy when assigning taxonomy (Liu et al., 2008;Wang et al., 2007).

In our opinion the usage of pV12B and pV4 in parallel has several key advantages. Any primer pair is more or less biased towards certain taxa. While the pV12B community profiles were enriched for Bacteroidetes, pV4 profiles were enriched for Firmicutes. The application of two independent primer pairs provides two different, complementary views on the composition of microbial communities and hence amplification biases are compensated. Taxa missed by one primer pair may be detected by the other. Moreover, 16S DNA pyrosequencing is mainly employed to compare the composition of microbial communities in two ore more different conditions (e.g. gut microbial communities from obese and lean subjects). If the same trends are observed for different primer pairs, the results can be treated with high confidence, otherwise with caution.

The 'true' community composition of the microbes of the A14B mixture was unknown, as any method for profiling microbial communities is biased to some unknown extend. Therefore, the amplification bias of eight primer pairs was estimated based on a comparison versus the median over all primer pairs. Despite having a general low amplification bias, an optimal primer pair should also well amplify 16S genes from a wide phylogenetic range of bacteria, particularly those of rare organisms. When comparing gut microbial communities of different groups, e.g. obese versus lean subjects, in our opinion, this factor is even more critical than a low amplification bias. The 'universality' or power to detect a wide range of taxa was determined by a rarefaction analysis of the family richness and overall diversity.

Herein we have focused on selecting optimal primer pairs for studying the human gut microbiome. However, owing to their strong advantages (no dimer formation, high amplification rate, broad range of covered taxa, high classification accuracy) the pV12B and pV4 primer pairs also appear to be well suited for profiling microbial communities not associated to the human intestine.

The sequence conservation of variable 16S regions differs considerably and it may further depend on the taxonomic origin of the 16S gene. Consequently, Operational Taxonomic Units (OTUs) generated from variable regions cannot be directly linked to an established taxonomy, such that each OTU corresponds to one taxon. Furthermore, OTU-based results of sequences representing different variable 16S regions cannot be directly compared. However, the grouping of 16S DNAs into OTUs is a valuable approach for assessing the diversity and composition of microbial communities and particularly for characterizing unknown bacterial groups.

It has been reported that the high variability of the V1+V2 region may lead to the overestimation of community richness (Youssef et al., 2009). However, this drawback can easily be circumvented by using a more restrictive sequence-identity cut-off when generating OTUs or by normalizing the number of observed OTUs (Youssef et al., 2009). In our opinion, when community richness is estimated from generated OTUs, the main problem stems from the uneven variability of the 16S gene in different taxa, at least when focusing on short sub-regions.

Eight primer pairs were compared herein using taxonomic classifications derived by the RDP-Classifier, instead of applying an OTU-analysis. The taxa-driven approach had the limitation that only known bacterial groups were considered. Therefore, family was the lowest taxonomic rank included in our analysis. Rank family provides an appropriate tradeoff between resolution and fraction of sequences that is assigned to a known taxon. However, as variable 16S regions differ in sequence conservation, as demonstrated herein, OTU-based results could not be directly used to compare the eight primer pairs.

Our results demonstrate that *in silico* analysis are valuable for the selection of PCR primers, but also wet-lab validations are essential. For example, our initial *in silico* analysis using Probe Match suggested that pV3A and pV3B would well amplify 16S DNA from a wide range of different taxa and hence were promising candidates for the amplification of 16S variable regions. This was confirmed by our wet-lab experiments. But in practice, with pV3A and pV3B only a low amplification rate was achieved, resulting from the formation of dimers.

Outlook

At present, the throughput of the GS FLX sequencer is sufficient to obtain an overview of the composition of complex microbial communities at higher taxonomic rank. Deeper sequencing methods are required to cover entire microbial communities at lower taxonomic rank, such as family, genus or species. The next generation of the pyrosequencing system (Titanium) provides even higher throughput and increased read length of ~400bp. The higher throughput allows a further reduction in sequencing cost and a deeper coverage of microbial communities. The increased read length will lead to an improved accuracy when assigning taxonomy. For Titanium, combinations of primers studied herein can be used.

For the computational analysis of thousands of short 16S DNA fragments, several tasks still need to be resolved, such as generating high-quality alignments or the reconstruction of phylogenetic trees.

Materials and Methods

Sample preparation and sequencing

DNA was extracted from fecal samples using the GNOME kit (BIO 101, La Jolla, CA, USA) as previously described (Firmesse et al., 2008). Samples came from a healthy adult and four healthy one month old babies. DNA extracts were combined with a ratio 4:1:1:1:1; to balance the contribution from adult and babies. Primers were designed as previously proposed (Hamady et al., 2008) and coupled as reported in Supplementary table 1. No special precaution was taken for the purification of the primers. Truncations frequently happen during synthesis (from 3' to 5') of long primers, but essentially affect the 5' extremity, i.e. the 454 adapters. Consequently, PCR products obtained with these truncated primers were not sequenced during the GS FLX run. Hence truncated primers did not need to be removed in a prior filtering step. For each couple of primers, two 50 µl PCRs were prepared in quintuplet, containing 1X Expand Long Template buffer 1, 50 µM of each dNTP (Roche Applied Science, Basel, Switzerland), 20 pmole of each primers (Microsynth, Balgach,

Switzerland) and 2.5 U Expand Long Template Enzyme mix (Roche Applied Science, Basel, Switzerland). To each reaction a minimum of 2 ng of DNA template was added; a quantity sufficient to limit the sequencing of duplicates of the same molecule (Frank et al., 2008). For pV3 and pV6, 100 µl PCR were prepared. Standard PCR amplifications were performed in GeneAmp PCR System 9700 (Applied Biosystems Inc, Foster City, CA, USA). Gradients PCR were performed with a Biometra T Gradient (Biometra biomedizinische Analytik GmbH, Goettingen, Germany). The PCR parameters were 94°C for 5 min, 25 cycles of 94°C for 30 sec, annealing temperature (see Supplementary table 1) for 30 sec and 72°C for 30 sec, followed by 72°C for 7 min. PCR cycles were restricted to 25 to prevent the saturation of the most abundant sequences (Frank et al., 2008). After pooling the two PCRs of each quintuplet, 10 µl of PCR product were visualized on agarose gel (1.2% in TBE buffer) stained with SYBR Safe (Invitrogen, Eugene, Oregon, USA). For PCR with pV3 and pV6, the 200 µl of product were gel-purified with Qiagen mini-elute kit (Qiagen, Hilden, Germany). Then, PCR products were sent to Beckman Coulter Genomics (Grenoble, France) where equal amounts of each were pooled and sequenced by the GS FLX System (Roche).

Sequence quality control

Errors in unassembled pyrosequencing reads occur with a rate of about 0.5% (Huse et al., 2007). Filtering for low quality reads is particularly important when short pyrosquencing reads are directly taxonomically characterized, e.g. using the RDP-Classifier (Wang et al., 2007). In this case, sequencing errors cannot be resolved in a prior assembly step or based on a comparison to known reference sequences.

Low quality reads were identified using criteria adapted from Huse *et al.* 2007: A) Undefined sequence key; these reads could not be unambiguously assigned to any of the PCR samples. B) Lack of recognizable 5' primer sequence. C) More than one error in 5' primer sequence. D) Average quality score below 25. Huse *et al.* 2007 have demonstrated that the number of errors per sequence correlates with its average quality score. The number of errors per sequence is very low for reads with an average quality score above 25, but considerably higher for reads with an average score below 25. E) Presence of ambiguous characters ('N'). Ambiguous characters are introduced by the Roche pyrosequencing software if bases cannot be unambiguously determined. A strong correlation between the presence of ambiguous characters and other sequence errors has been observed (Huse et al., 2007). All reads containing ambiguous characters were discarded. F) Sequence reads not matching the expected length were excluded, since these are likely to have considerably increased error rates when compared to reads with an optimal read length (Huse et al., 2007). First, primer sequences were removed using the vectorstrip software (Rice et al., 2000) (-besthits 1 –mismatch 30 –besthits parameters). All reads with less than 100 bp were excluded. Subsequently, the median length and the Median Absolute Deviation (MAD) of the remaining reads were calculated separately for each of the 8 different primer pairs. All reads shorter than the median – 5 * MAD were excluded. G) Reads without a BLAST (Altschul et al., 1990) hit to the ARB (Ludwig et al., 2004) 16S DNA database (Evalue cut-off: 10^{-10}) were discarded, since these likely do not represent regions of 16S DNA genes.

Taxonomic characterization and diversity analysis

All high-quality reads were classified into Bergey's taxonomy using the RDP-Classifier (80% confidence cut-off). Amplicons that could not be assigned to any known taxa at a certain rank (phylum, class, order, family, or genus) were classified as "Unknown" at that rank. Conversely, the "Unknown" category was separately defined for each rank.

Rarefaction analysis was conducted separately for each taxonomic rank using Analytic Rarefaction 1.3. The software was used to estimate the number of taxa that would be observed if samples of different sizes were taken. For each primer pair, the amplicons obtained for all five replicates were joined. The estimates were based on the total number of taxa observed for the joined set of amplicons. Rarefaction curves were obtained by plotting the estimated number of observed taxonomic groups versus the sample size. Notably, in the rarefaction analysis only those amplicon sequences were included that could be assigned to a taxon on the respective rank.

Additionally, the Shannon diversity index was applied to estimate the overall microbial diversity of the 1A4B mixture. The Shannon index includes microbial richness and evenness, where evenness measures if taxa are equally abundant in a sample or if dominant/underrepresented taxa exist. An optimal primer pair should provide both high richness and evenness measures and hence a high overall diversity (Shannon index). For each primer pair, 16S DNAs were randomly sampled from the joined set of sequences, taxa were assigned with the RDP-Classifier. The Shannon diversity was subsequently measured for each sub-sample at rank phylum, class, order, family and genus. In the context of this work, for a taxonomic rank *r*, the Shannon diversity index is defined as:

$$H' = -\sum p_i \ln p_i \ ,$$

where p_i is the proportion of 16S sequences assigned to the *i*-th taxon of rank *r* (sequences that could not be assigned to any taxon at rank *r* were excluded). Diversity rarefaction curves were obtained for each rank by plotting the Shannon diversity index against the number of sequences sampled.

Statistical analysis

Significantly over- or underrepresented taxa were identified by an ANOVA and the Fishers Least Significant Difference (LSD) test. In both cases, the significance level p was set to 0.05. The tests were performed on relative abundance values (normalized taxonomic counts): For each PCR sample *s* and for each taxonomic group *t*, the number of 16S amplicons from *s* that were assigned to *t* were divided by the total number of amplicons obtained for *s*. Taxonomy was assigned with the RDP-Classifier, as described above. An ANOVA was applied for each taxon under the null hypothesis that there is no difference between the means obtained for each primer pair. For those taxa for which the null hypothesis was rejected the least significant distance was calculated with the LSD test. The median for each taxon was computed over all 40 PCR samples (8 primer pairs x 5 replicates).

A principal component analysis (PCA) was conducted on the centered matrix of relative taxonomic abundances using the svd() function of the R statistical package.

Measuring the sequence-conservation of variable 16S DNA regions

The sequence conservation of different variable 16S regions (as defined by the eight studied primer pairs) was assessed *in silico* on a reference set of 57,054 full length 16S sequences (length above 1,200 bp and below 1,900 bp) downloaded from RDP v9 (Cole et al., 2003). For each of the eight primer pairs, the respective 16S region between the two primers was extracted from the reference sequences using the vectorstrip software (Rice et al., 2000) (-besthits 1 -mismatch 30 parameters). Extracted sequences containing ambiguous characters (e.g. 'N') were excluded from further analysis. For each primer pair, all extracted sequences were aligned with the muscle software using -diags and -maxiters 2 parameters. For each taxonomic rank (phylum, class, order, family, genus) and each taxon *t* at that rank, the simple pair-wise sequence identity of all sequences originating from *t* was computed. Gaps were not taken into account.

Acknowledgements

Mireille Moser

Author Contributions:
Study concept and design: BB
Acquisition of data: DM, BB
Analysis and interpretation of data: LK, BB
Drafting and writing of the manuscript: LK, BB
Statistical analysis: AR, LK
Overall supervision of the study: BB

References

Altschul,SF, W Gish, W Miller, E W Myers, 1990, Basic local alignment search tool: J Mol Biol, v. 215 %6, p. 403-410.

Andersson,AF, M Lindberg, H Jakobsson, F Backhed, P Nyren, L Engstrand, 2008, Comparative analysis of human gut microbiota by barcoded pyrosequencing: PLoS.One., v. 3, p. e2836.

Antonopoulos,DA, S M Huse, H G Morrison, T M Schmidt, M L Sogin, V B Young, 2009, Reproducible community dynamics of the gastrointestinal microbiota following antibiotic perturbation: Infect.Immun., v. 77, p. 2367-2375.

Binladen,J, M T Gilbert, J P Bollback, F Panitz, C Bendixen, R Nielsen, E Willerslev, 2007, The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing: PLoS.ONE., v. 2, p. e197.

Bowers,RM, C L Lauber, C Wiedinmyer, M Hamady, A G Hallar, R Fall, R Knight, N Fierer, 2009, Characterization of airborne microbial communities at a high-elevation site and their potential to act as atmospheric ice nuclei: Appl.Environ.Microbiol., v. 75, p. 5121-5130.

Bru,D, F Martin-Laurent, L Philippot, 2008, Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example: Appl.Environ.Microbiol., v. 74, p. 1660-1663.

Chakravorty,S, D Helb, M Burday, N Connell, D Alland, 2007, A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria: J.Microbiol.Methods, v. 69, p. 330-339.

Cole,JR, B Chai, T L Marsh, R J Farris, Q Wang, S A Kulam, S Chandra, D M McGarrell, T M Schmidt, G M Garrity, J M Tiedje, 2003, The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy: Nucleic Acids Res, v. 31 %6, p. 442-443.

Dethlefsen,L, S Huse, M L Sogin, D A Relman, 2008, The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing: PLoS.Biol., v. 6, p. e280.

Felsenstein,J, 1989, PHYLIP: Phylogeny Inference Package (Version 3.2): Cladistics, v. 5, p. 164-166.

Fierer,N, M Hamady, C L Lauber, R Knight, 2008, The influence of sex, handedness, and washing on the diversity of hand surface bacteria: Proc.Natl.Acad.Sci.U.S.A, v. 105, p. 17994-17999.

Firmesse,O, A Mogenet, J L Bresson, G Corthier, J P Furet, 2008, Lactobacillus rhamnosus R11 consumed in a food supplement survived human digestive transit without modifying microbiota equilibrium as assessed by real-time polymerase chain reaction: Journal of Molecular Microbiology and Biotechnology, v. 14, p. 90-99.

Frank,JA, C I Reich, S Sharma, J S Weisbaum, B A Wilson, G J Olsen, 2008, Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes: Appl.Environ.Microbiol., v. 74, p. 2461-2470.

Grice,EA, H H Kong, S Conlan, C B Deming, J Davis, A C Young, G G Bouffard, R W Blakesley, P R Murray, E D Green, M L Turner, J A Segre, 2009, Topographical and temporal diversity of the human skin microbiome: Science, v. 324, p. 1190-1192.

Grice,EA, H H Kong, G Renaud, A C Young, G G Bouffard, R W Blakesley, T G Wolfsberg, M L Turner, J A Segre, 2008, A diversity profile of the human skin microbiota: Genome Res., v. 18, p. 1043-1050.

Hamady,M, R Knight, 2009, Microbial community profiling for human microbiome projects: Tools, techniques, and challenges: Genome Res., v. 19, p. 1141-1152.

Hamady,M, C Lozupone, R Knight, 2009, Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data: ISME.J..

Hamady,M, J J Walker, J K Harris, N J Gold, R Knight, 2008, Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex: Nat.Methods, v. 5, p. 235-237.

Hoffmann,C, N Minkah, J Leipzig, G Wang, M Q Arens, P Tebas, F D Bushman, 2007, DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations: Nucleic Acids Research, v. 35.

Hongoh,Y, H Yuzawa, M Ohkuma, T Kudo, 2003, Evaluation of primers and PCR conditions for the analysis of 16S rRNA genes from a natural environment: Fems Microbiology Letters, v. 221, p. 299-304.

Huber,JA, D B Mark Welch, H G Morrison, S M Huse, P R Neal, D A Butterfield, M L Sogin, 2007, Microbial population structures in the deep marine biosphere: Science, v. 318, p. 97-100.

Hugenholtz,P, B M Goebel, 1998, Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity: J Bacteriol, v. 180 %6, p. 4765-4774.

Hugenholtz,P, 2002, Exploring prokaryotic diversity in the genomic era: Genome Biol, v. 3 %6, p. REVIEWS0003.

Huse,SM, L Dethlefsen, J A Huber, D M Welch, D A Relman, M L Sogin, 2008, Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing: PLoS.Genet., v. 4, p. e1000255.

Huse,SM, J A Huber, H G Morrison, M L Sogin, D M Welch, 2007, Accuracy and quality of massively parallel DNA pyrosequencing: Genome Biol., v. 8, p. R143.

Jones,RT, M S Robeson, C L Lauber, M Hamady, R Knight, N Fierer, 2009, A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses: ISME.J., v. 3, p. 442-453.

Keijser,BJ, E Zaura, S M Huse, J M van der Vossen, F H Schuren, R C Montijn, J M ten Cate, W Crielaard, 2008, Pyrosequencing analysis of the oral microflora of healthy adults: J.Dent.Res., v. 87, p. 1016-1020.

Kim,YJ, N Teletia, V Ruotti, C A Maher, A M Chinnaiyan, R Stewart, J A Thomson, J M Patel, 2009, ProbeMatch: rapid alignment of oligonucleotides to genome allowing both gaps and mismatches: Bioinformatics., v. 25, p. 1424-1425.

Lauber,CL, M Hamady, R Knight, N Fierer, 2009, Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale: Appl.Environ.Microbiol., v. 75, p. 5111-5120.

Letunic,I, P Bork, 2007, Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation: Bioinformatics., v. 23, p. 127-128.

Ley,RE, F Backhed, P Turnbaugh, C A Lozupone, R D Knight, J I Gordon, 2005, Obesity alters gut microbial ecology: Proc.Natl.Acad.Sci.U.S.A, v. 102, p. 11070-11075.

Ley,RE, M Hamady, C Lozupone, P J Turnbaugh, R R Ramey, J S Bircher, M L Schlegel, T A Tucker, M D Schrenzel, R Knight, J I Gordon, 2008, Evolution of mammals and their gut microbes: Science, v. 320, p. 1647-1651.

Ley,RE, P J Turnbaugh, S Klein, J I Gordon, 2006, Microbial ecology: human gut microbes associated with obesity: Nature, v. 444, p. 1022-1023.

Liu,Z, T Z DeSantis, G L Andersen, R Knight, 2008, Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers: Nucleic Acids Res., v. 36, p. e120.

Liu,Z, C Lozupone, M Hamady, F D Bushman, R Knight, 2007, Short pyrosequencing reads suffice for accurate microbial community analysis: Nucleic Acids Res., v. 35, p. e120.

Ludwig,W, O Strunk, R Westram, L Richter, H Meier, Yadhukumar, A Buchner, T Lai, S Steppi, G Jobb, W Forster, I Brettske, S Gerber, A W Ginhart, O Gross, S Grumann, S Hermann, R Jost, A Konig, T Liss, R Lussmann, M May, B Nonhoff, B Reichel, R Strehlow, A Stamatakis, N Stuckmann, A Vilbig, M Lenke, T Ludwig, A Bode, K H Schleifer, 2004, ARB: a software environment for sequence data: Nucleic Acids Res., v. 32, p. 1363-1371.

McKenna,P, C Hoffmann, N Minkah, P P Aye, A Lackner, Z Liu, C A Lozupone, M Hamady, R Knight, F D Bushman, 2008, The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis: PLoS.Pathog., v. 4, p. e20.

Pace,NR, D A Stahl, D J Lane, 1985, Analyzing natural microbial populations by rRNA sequences: ASM News, v. 51 %6, p. 4-12.

Palmer,C, E M Bik, D B DiGiulio, D A Relman, P O Brown, 2007, Development of the human infant intestinal microbiota: PLoS.Biol., v. 5, p. e177.

Rice,P, I Longden, A Bleasby, 2000, EMBOSS: the European Molecular Biology Open Software Suite: Trends Genet., v. 16, p. 276-277.

Sipos,R, A J Szekely, M Palatinszky, S Revesz, K Marialigeti, M Nikolausz, 2007, Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis: Fems Microbiology Ecology, v. 60, p. 341-350.

Sogin,ML, H G Morrison, J A Huber, W D Mark, S M Huse, P R Neal, J M Arrieta, G J Herndl, 2006, Microbial diversity in the deep sea and the underexplored "rare biosphere": Proc.Natl.Acad.Sci.U.S.A, v. 103, p. 12115-12120.

Suau,A, R Bonnet, M Sutren, J J Godon, G R Gibson, M D Collins, J Dore, 1999, Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut: Appl.Environ.Microbiol., v. 65, p. 4799-4807.

Sun,Y, Y Cai, L Liu, F Yu, M L Farrell, W McKendree, W Farmerie, 2009, ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences: Nucleic Acids Res., v. 37, p. e76.

Tap,J, S Mondot, F Levenez, E Pelletier, C Caron, J P Furet, E Ugarte, R Munoz-Tamayo, D L Paslier, R Nalin, J Dore, M Leclerc, 2009, Towards the human intestinal microbiota phylogenetic core: Environ.Microbiol..

Tringe,SG, P Hugenholtz, 2008, A renaissance for the pioneering 16S rRNA gene: Curr.Opin.Microbiol., v. 11, p. 442-446.

Tschop,MH, P Hugenholtz, C L Karp, 2009, Getting to the core of the gut microbiome: Nat.Biotechnol., v. 27, p. 344-346.

Turnbaugh,PJ, M Hamady, T Yatsunenko, B L Cantarel, A Duncan, R E Ley, M L Sogin, W J Jones, B A Roe, J P Affourtit, M Egholm, B Henrissat, A C Heath, R Knight, J I Gordon, 2009, A core gut microbiome in obese and lean twins: Nature, v. 457, p. 480-484.

Turnbaugh,PJ, R E Ley, M Hamady, C M Fraser-Liggett, R Knight, J I Gordon, 2007, The human microbiome project: Nature, v. 449, p. 804-810.

Turnbaugh,PJ, R E Ley, M A Mahowald, V Magrini, E R Mardis, J I Gordon, 2006, An obesity-associated gut microbiome with increased capacity for energy harvest: Nature, v. 444, p. 1027-1031.

Wang,Q, G M Garrity, J M Tiedje, 2007, Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy: Appl Environ Microbiol, v. 73 %6, p. 5261-5267.

Woese,CR, G E Fox, 1977, Phylogenetic structure of the prokaryotic domain: the primary kingdoms: Proc.Natl.Acad.Sci.U.S.A, v. 74, p. 5088-5090.

Youssef,N, C S Sheik, L R Krumholz, F Z Najar, B A Roe, M S Elshahed, 2009, Comparison of species richness estimates obtained using nearly complete fragments

and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys: Appl.Environ.Microbiol., v. 75, p. 5227-5236.

Figures and Tables

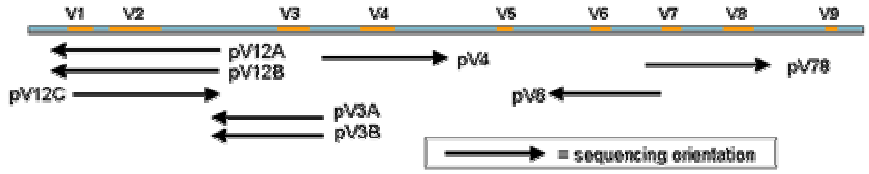


Figure 1
Variable 16S regions amplified by eight different primer pairs compared herein.

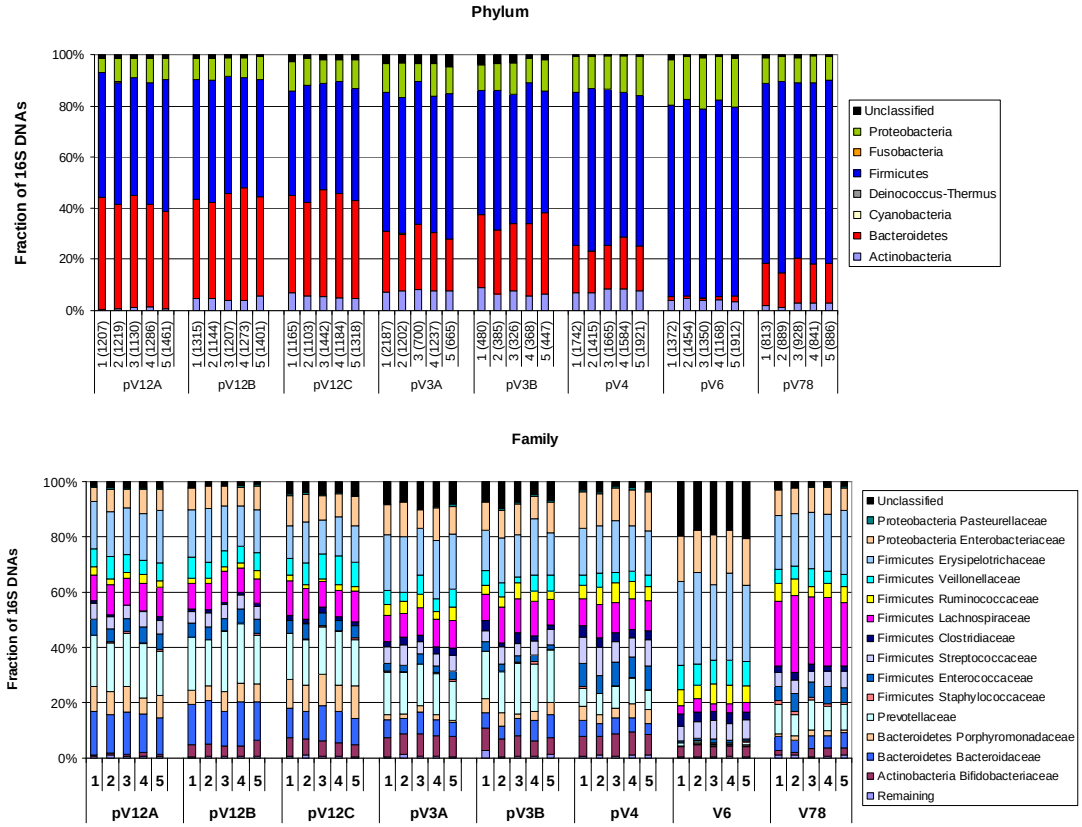


Figure 2
Community structure profiles at ranks phylum and family. For each primer pair, community profiles for five replicates (1 to 5) are shown. Total number of amplicons is given for each PCR sample in brackets. Unknown depicts fraction of 16S amplicons that could not be assigned to any known taxon. At rank family, only the 14 most abundant taxa are represented in detail. Fractions of amplicons assigned to the remaining families are depicted as “Remaining”.

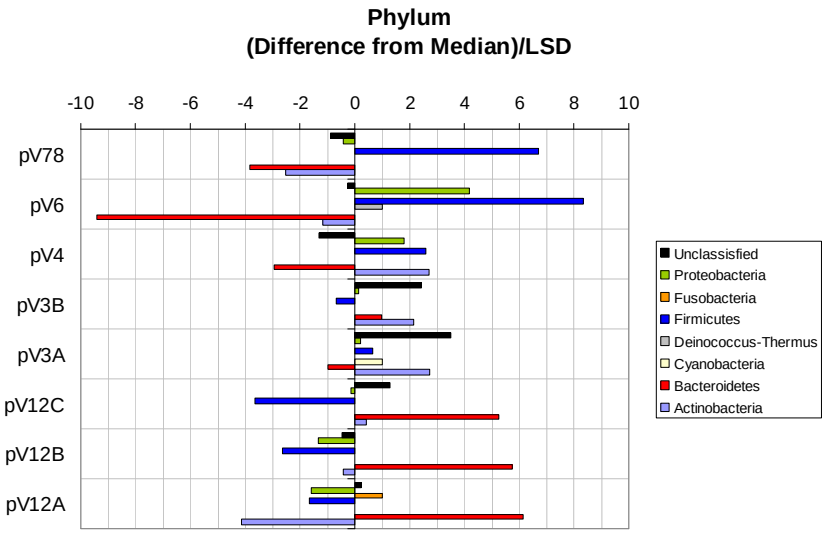


Figure 3
Amplification bias of eight primer pairs at phylum rank. Depicted is the significance of over-/underrepresented phyla. Bars above +1 indicate that the respective phylum is significantly enriched in respect to the median ($p < 0.05$). Bars below -1 indicate that the phylum is significantly underrepresented. Bars represent the (difference from median)/LSD for each primer pair and phylum, where LSD is the Fisher's Least Significant Difference.

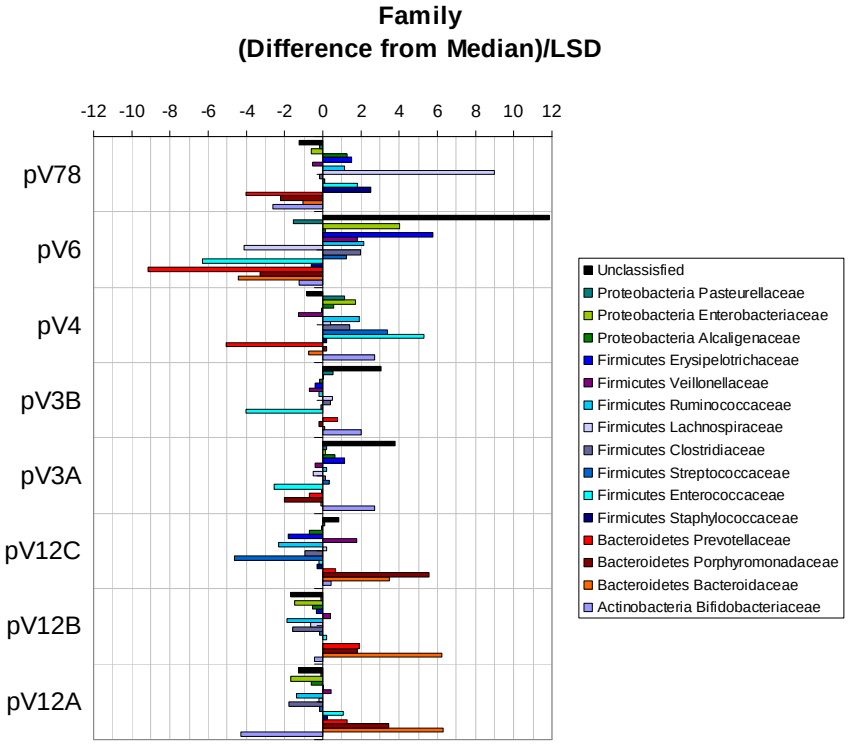


Figure 4
Amplification bias of eight primer pairs at family rank. Depicted is significance of over-/underrepresented families in the 16S amplicons obtained for each primer pair. Bars above +1 indicate that the respective family is significantly enriched in respect to

the median ($p < 0.05$). Bars below -1 indicate that the family is significantly underrepresented. Bars represent the (difference from median)/LSD for each primer pair and family, where LSD is the Fisher's Least Significant Difference. Shown are only those families that were statistically different according to an ANOVA ($p < 0.05$).

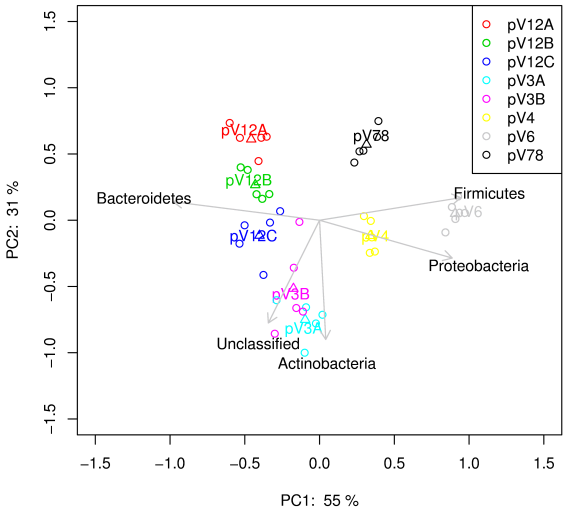


Figure 5

Principle Component Analysis of community profiles at rank phylum. Each PCR sample is represented by a circle, the median for each primer pair is depicted by a rectangle. Right-angle projection of sample points on arrows approximates abundance of the respective phylum in that sample.

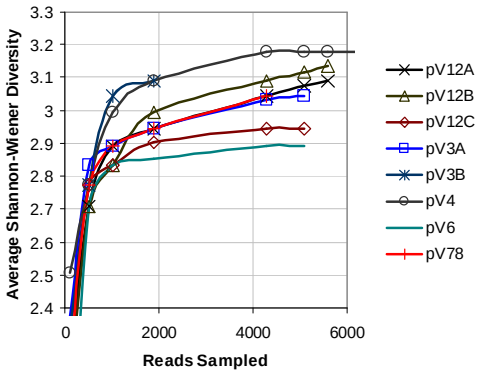


Figure 6

Rarefaction analysis of the estimated number of families observed for different sample sizes. High values indicate that the respective primer pair well amplifies variable 16S regions from a wide range of bacterial groups.

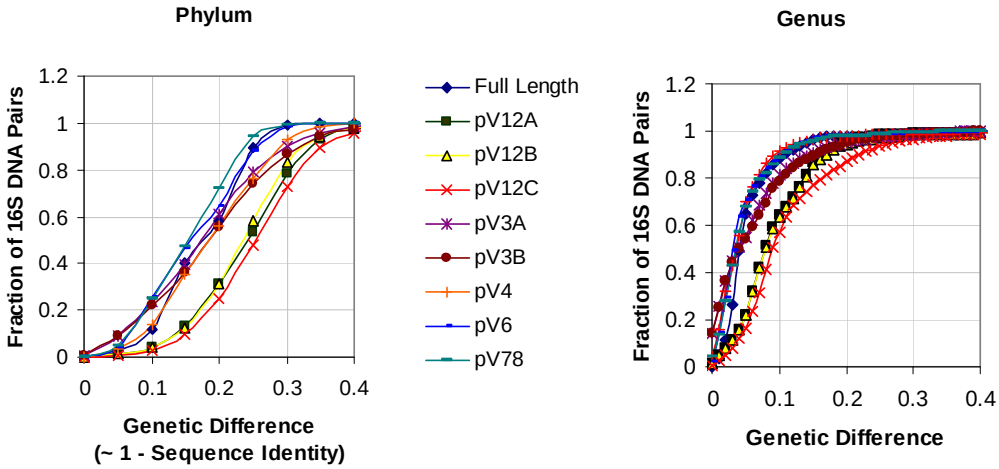


Figure 7
Sequence conservation of variable 16S regions as defined by eight different primer pairs. Shown is the fraction of 16S sequence pairs that have a pair-wise sequence identity above a certain genetic differences. Only those sequence pairs originating from the same phylum or genera, respectively, were taken into account. Higher values indicate that the respective 16S region is more conserved.

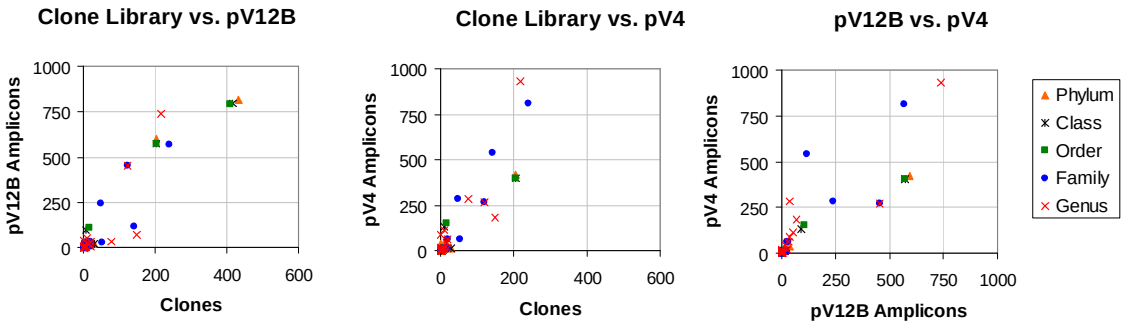


Figure 8
Comparison to “gold standard”. Results obtained for full-length 16S sequences (clone library), pV12B and pV4 are compared. For each taxon at rank phylum to genus, the fraction of sequences assigned to that taxon is plotted. Colors represent the ranks of the respective taxa (Phylum: Orange, Class: Black, Order: Green, Family: Blue, Genus: Red).

Supplementary methods

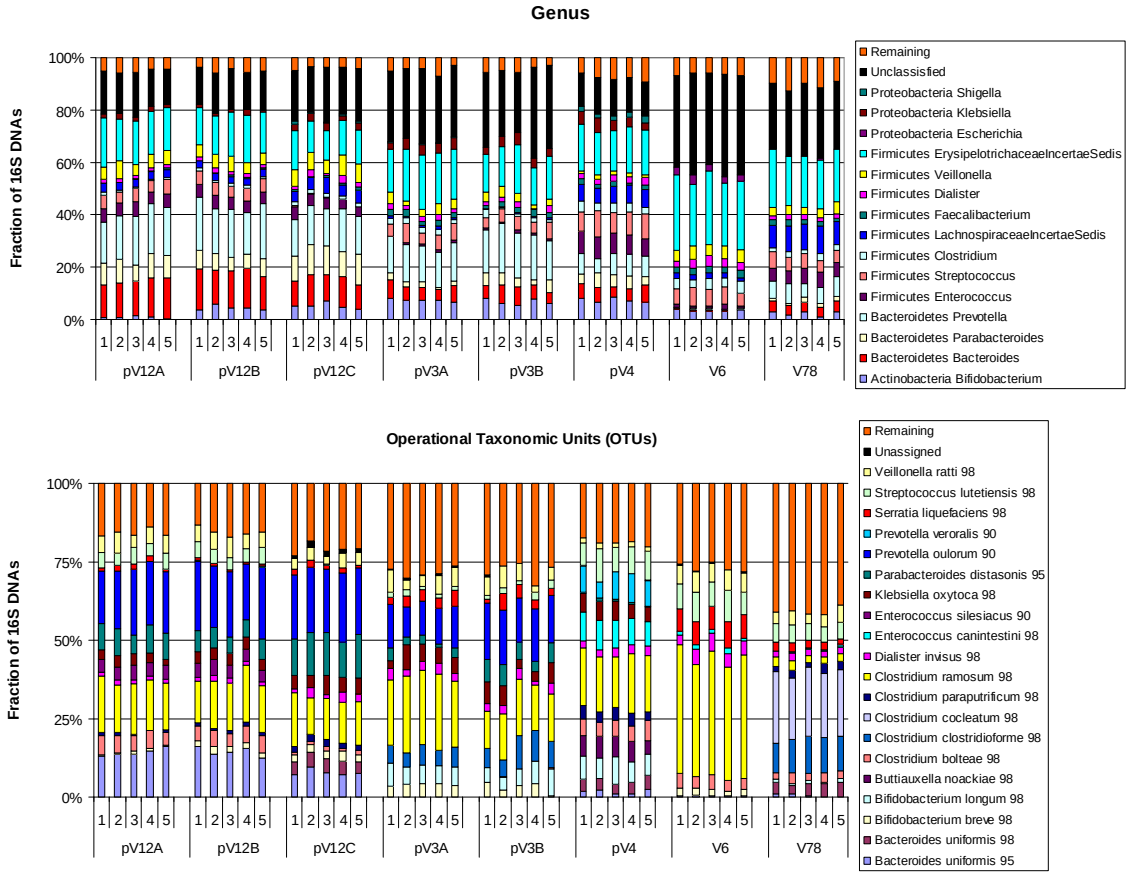
Reconstruction of phylogenetic tree

A phylogenetic tree was reconstructed from all pV6 amplicons that were obtained from the 1A4B mixture: pV6 amplicons were compared to the RDP 16S DNA database using blastN. For each amplicon its best matching reference sequence from the RDP database was downloaded in aligned format. Subsequently, a profile hidden Markov model (HMM) was computed from the multiple alignment of all best matching RDP reference sequences using hmmbuild from the HMMER package. The computed profile HMM was subsequently employed to add all amplicons to the multiple alignment of the best matching reference sequences (with hmmalign using the --withali option). The pair-wise sequence-identity of all amplicon sequences was calculated based on the resulting alignment, gaps were not taken into account. Finally, a phylogenetic tree was reconstructed using the neighbor-joining algorithm (with the neighbor program from the PHYLIP package (Felsenstein, 1989)). The tree was visualized with the iTol software (Letunic and Bork, 2007).

Grouping of 16S amplicons into OTUs

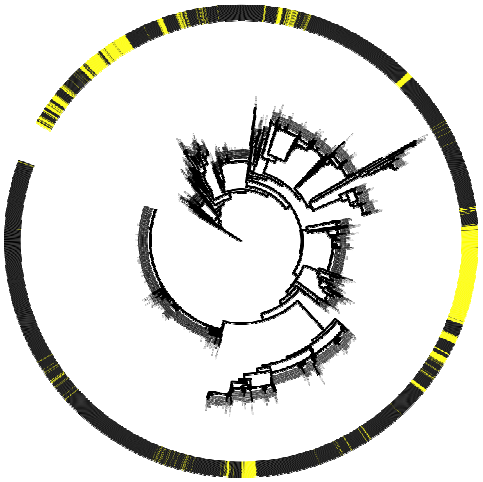
16S amplicons were grouped into Operational Taxonomic Groups (OTUs) based on their best BLAST hit to full-length reference ribosomal RNAs. First, reads were compared to all type strain 16S ribosomal RNAs from the RDP database using blastN. Second, all amplicons with a best match (sequence-identity > 98%) to the same reference RNA were grouped into one OTU. In an analogous manner, all remaining amplicons were iteratively grouped into OTUs using sequence-identity cut-offs of 95%, 90%, 80% and 70%. Each generated OTU was labelled by the type-strain of the respective best-matching reference ribosomal RNA and the used sequence-identity cut-off.

Supplementary figures

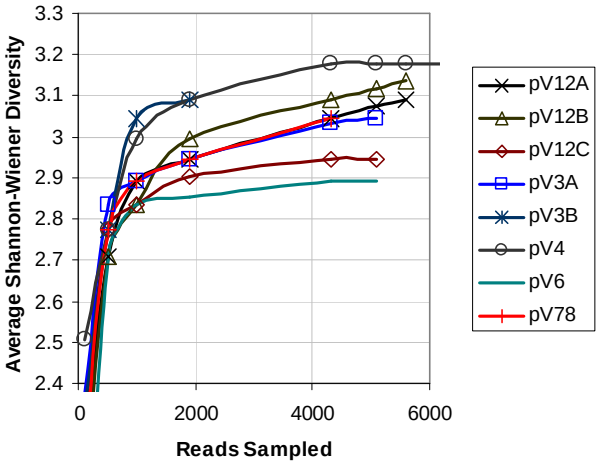


Supplementary figure 1

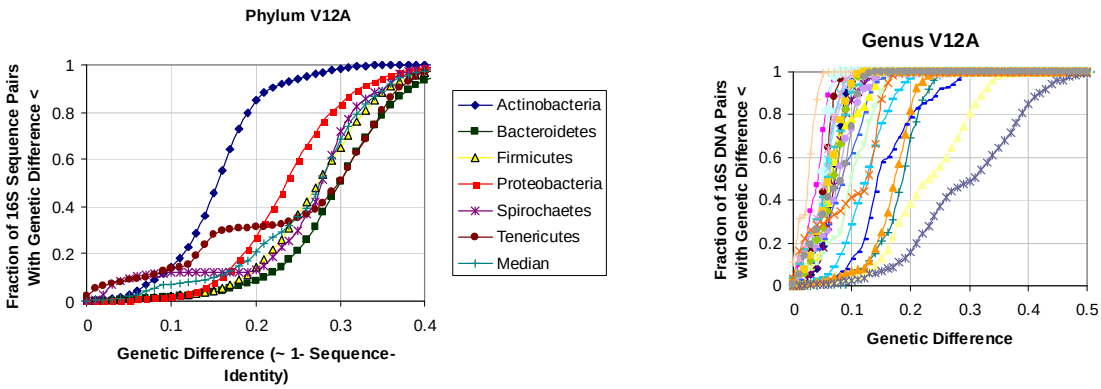
Community structure profiles at rank genus and generated OTUs. For each primer pair community profiles for five replicates (1 to 5) are shown. The community composition at rank genus is shown in the upper chart. Generated OTUs are displayed in the lower chart. For each OTU, the most similar type strain and the applied sequence-identity cut-off are depicted (e.g. *Bacteroides uniformis* 98, where 98 is the sequence-identity cut-off in %). Only the 15 most prevalent genera and 20 most prevalent OTUs are represented in detail. Fractions of amplicons assigned to the remaining genera/OTUs are depicted as “Remaining”. Fractions of amplicons not assigned to any genus or OTU, respectively, are depicted as “Unassigned”.



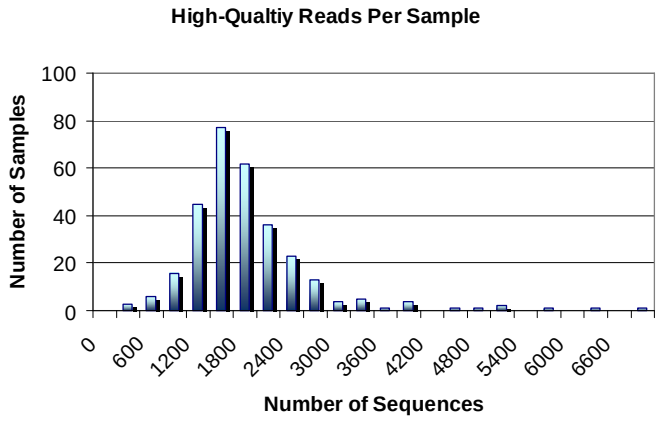
Supplementary figure 2
Phylogenetic tree of pV6 amplicons. The tree was reconstructed from all pV6 amplicons that were obtained from the 1A4B mixture. Amplicons not assigned to any known family are depicted in yellow; the remaining amplicons are marked in black. Family assignments were derived by the RDP-Classifier.



Supplementary figure 3
Average Shannon diversity. Reads were randomly drawn from each PCR-sample. For each drawn sub-set the Shannon diversity index was measured. Depicted is the average Shannon diversity for each primer-pair versus the number of drawn sequence reads.

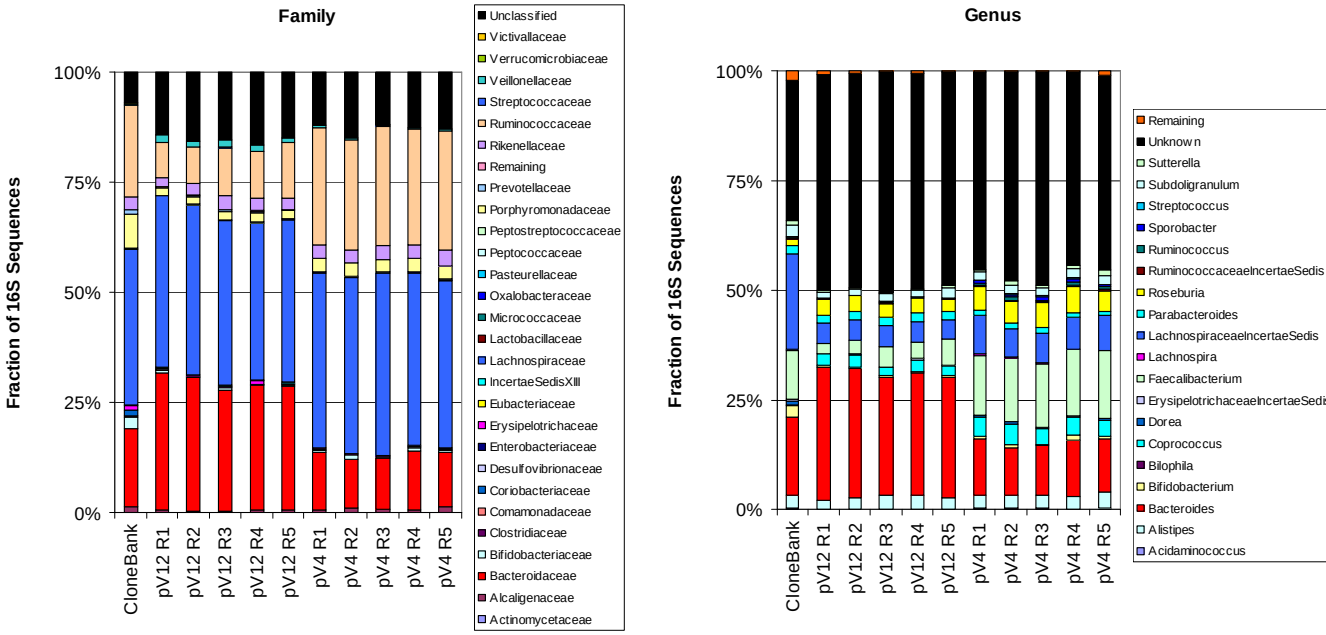


Supplementary figure 4
Phylum and genus dependent sequence conservation of the V12 region (as defined by the pV12 primer). Shown is the fraction of 16S sequence pairs that have a pair-wise sequence identity above a certain genetic differences. Higher values indicate that the respective V12A region is more conserved in 16S DNAs of the respective phylum.

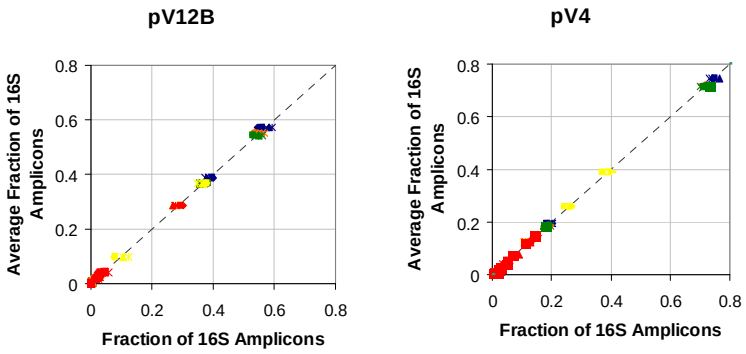


Supplementary figure 5

Number of high-quality sequence reads for each PCR sample.



Supplementary figure 6
Comparison of clone-bank and pyrosequencing results at rank family (left side) and genus (right side). Only the 19 most abundant genera are displayed. The remaining genera are summarized as “Remaining”.



Supplementary figure 7

Reproducibility of extraction. With each of the pV12B and pV4 primer pairs, 16S DNA was amplified from the same human fecal sample in 5 separate PCR reactions. For each PCR sample, and for each taxon at ranks phylum to genus, the fraction of sequences assigned to that taxon is plotted versus the average fraction of sequences assigned to that taxon. The average was computed over the 5 PCR samples. Colors represent the ranks of the respective taxa (Blue: Phylum, Orange: Class, Green: Order, Yellow: Family, Red: Genus).

version with V5 and new re Table 1: in silico evaluation of individual primers for 16S sequence annealing

						x mismatches(f)					
						x=0	x=1	x=2			
Primer	Sequence(a)	Orientation(b)	Positions(c)	Region(d)	Nb of type strains(e)	%	%	%	Flanking(g)	Pair(h)	Reference
8F	AGAGTTTGATCCTGGCTCAG	F	8-27	6-40	1459	56	83	94	V1	pV12-A	Eckburg Relman / turnbaugh gordon
27f-YM	AGAGTTTGATYMTGGCTCAG	F	8-27	6-40	1459	70	88	96	V1	pV12-B	franck2008
TET-63F-primer	CAGGCCTAACACATGCAAGTC	F	43-63	40-65	4451	14	29	90	V1	-	Sipos 2007
–63f	GCCTAACACATGCAAGTC	F	46-63	40-65	4451	24	88	96	V1	-	yu morrison 2004
Bacteria 52f	TAAYACATGCAAGTCG	F	49-64	40-65	4451	83	97	98	V1	pV12-C	Ghani Sghir 'designer' ???
338R	TGCTGCCTCCCGTAGGAGT	R	338-356	300-370	5164	91	97	99	V2	-	turnbaugh gordon
BSR357	CTGCTGCCTYCCGTA	R	343-357	300-370	5164	94	99	100	V2	pV12-A-B-C	ERRdb
env330	ACGGTCCAGACTCCTACGGG	F	329-348	300-370	5164	16	79	93	V3	-	Duthoit et al. 2003
Bacteria F333	GGYCCAGACTCCTACGGGA	F	331-349	300-370	5164	72	92	97	V3	-	Godon et col
Bacteria F333-BB2	GGYCCARACTCCTACGGRA	F	331-349	300-370	5164	79	94	98	V3	pV3-A	this work
338F	ACTCCTACGGGAGGCAGCAG	F	338-357	300-370	5164	91	97	99	V3	-	Huse
U-350f	CTCCTACGGGAGGCAGCAGT	F	339-358	300-370	5164	91	97	99	V3	-	Tap
BSF343	TACGGRAGGCAGCAG	F	343-357	300-370	5164	94	99	100	V3	pV3-B	ERRdb
BSF349	AGGCAGCAGTGGGGAAT	F	349-365	300-370	5164	66	94	98	V3	-	ERRdb
Universal R500 / 533R	TTACCGCGGCTGCTGGCAC	R	515-533	500-550	5166	93	98	99	V3	pV3-A-B	godon et col / Huse
BSR534	ATTACCGCGGCTGCTGGC	R	517-534	500-550	5166	88	97	99	V3	-	ERRdb
BSF517	GCCAGCAGCCGCGGTAA	F	517-533	500-550	5166	93	98	99	V4	pV4	ERRdb
BSR798	GGGGTATCTAATCCC	R	785-797	770-820	5165	4	4	97	V4	-	ERRdb
Bacteria 785-803r	CTACCAGGGTATCTAATCC	R	785-803	770-820	5165	89	99	100	V4	-	Stahl et al ???
806R	GGACTACCAGGGTATCTAAT	R	787-806	770-820	5165	89	99	100	V4	pV4	Bik Relman
784F	AGGATTAGATACCCTGGTA	F	784-801	770-820	5165	89	94	99	V5	-	Andersson
BSF784	RGGATTAGATACCCC	F	784-797	770-820	5165	4	95	100	V5	-	ERRdb
BSR926	CCGTCAATTYYTTTRAGTTT	R	907-926	900-1000	5162	82	99	100	V5	-	ERRdb
Bacteria 927-942r	ACCGCTTGTGCGGGCCC	R	927-942	900-1000	5162	40	82	96	V5	-	Stahl et al ???
Bacteria 930f	AGGAATTGRCGGGGGC	F	915-930	900-1000	5166	58	93	99	V6	-	Stahl et al ???
BSF917	GAATTGACGGGGRCCC	F	917-932	900-1000	5162	80	96	99	V6	pV6	ERRdb
–954f	GCACAAGCGGTGGAGCATGTGG	F	933-954	900-1000	5162	48	71	97	V6	-	yu morrison 2004
F-968	AACGCGAAGAACCTTAC	F	968-984	900-1000	5162	66	87	98	V6	-	yu morrison 2004
1061R	CRRCACGAGCTGACGAC	R	1061-1077	1040-1120	5159	97	100	100	V6	-	Andersson
BSR1114	GGGTTGCGCTCGTTRC	R	1099-1114	1040-1120	5159	90	99	100	V6	pV6	ERRdb
1070f	ATGGCTGTCGTCAGCT	F	1055-1070	1040-1120	5159	56	96	99	V7	-	yu morrison 2004
BSF1099	GYAACGAGCGCAACCC	F	1099-1114	1040-1120	5159	90	99	100	V7	pV78	ERRdb
1369r	GCCCGGGAACGTATTCACCG	R	1369-1388	1360-1420	4914	67	91	99	V8	-	yu morrison 2004
P-1392r/1402r	GCGGTGTGTACAAGRCCC	R	1385-1402	1360-1420	4914	88	97	99	V8	-	Tap3
P-1392r	GCGGTGTGTACAAGACCC	R	1385-1402	1360-1420	4914	16	92	98	V8	-	Tap
P-1392r	GCGGTGTGTRCAAGACCC	R	1385-1402	1360-1420	4914	16	92	98	V8	-	Tap2
Univ-1406R	ACGGGCGGTGWGTRCAA	R	1390-1406	1360-1420	4914	94	98	99	V8	-	Eckburg 2005
1391R	GACGGGCGGTGWGTRCA	R	1391-1407	1360-1420	4914	94	98	99	V8	pV78	Eckburg 2005
Universal 1392r	ACGGGCGGTGTGTRC	R	1392-1406	1360-1420	4914	93	99	99	V8	-	Stahl et al ???
BSR1407	GACGGGCGGTGTGTRC	R	1392-1407	1360-1420	4914	93	99	99	V8	-	ERRdb

a) 5' to 3'

b) F, forward; R, reverse

c) positions on 16S sequence of E.coli (gbk J01695.2)

d) in order to eliminate the sequences whose amplification primers were not trimmed before deposit in RDP database, Probe Match searches were restricted to 16S sequences extending beyong the matching positions

e) total number of type strains with 16S sequences covering this region

f) number of tolerated mismatches during the alignement with the 16S of type strains

g) flanking variable region

h) component of the tested primers pair

Supplementary table 1
35 studied primer pairs.

	Phylum	Class	Order	Family	Genus
pV12B R1	0.97	0.98	0.98	0.91	0.70
pV12B R2	0.98	0.98	0.98	0.91	0.70
pV12B R3	0.99	0.99	0.99	0.94	0.74
pV12B R4	0.98	0.99	0.99	0.93	0.72
pV12B R5	0.99	0.99	0.99	0.95	0.74
pV4 R1	0.98	0.98	0.98	0.98	0.87
pV4 R2	0.97	0.97	0.97	0.98	0.80
pV4 R3	0.97	0.97	0.97	0.97	0.81
pV4 R4	0.98	0.98	0.98	0.98	0.82
pV4 R5	0.98	0.98	0.98	0.97	0.84

Supplementary table 2

Correlation coefficient between community structure profiles obtained by high-throughput 16S DNA sequencing and 16S clone library. For each primer pair (pV12B and pV4), five PCR samples were amplified from the same original fecal sample (replicates R1 to R5).

ANNEXES

Menu 10 g de Fibres par Jour

J1	J2	J3	J4	J5
Midi Laitue (+ vinaigrette) Emincé de Poulet à la tomate Polenta Camembert Kiwis	**Midi** Laitue (+ vinaigrette) blanquette de veau pâtes Vache-qui-rit pomme	**Midi** lentilles (froides avec vinaigrette ou chaudes) sauté de porc aux nouilles Chèvre banane	**Midi** Laitue (+ vinaigrette) Hachis Parmentier Camembert Kiwis	**Midi** lentilles (froides avec vinaigrette ou chaudes) Bœuf sauce tomate coquillettes Vache qui rit pomme
Soir Bouillon vermicelles Spaghetti aux crevettes Yaourt nature Pomme	**Soir** Bouillon vermicelles Parmentier de poisson fromage blanc 40 % banane	**Soir** Bouillon vermicelles Lasagnes Petits suisses Kiwis	**Soir** Bouillon vermicelles Crevettes au gingembre polente yaourt sucré pomme	**Soir** Bouillon vermicelles saumon à la tomate Blé pilaf fromage blanc banane

Menu 40 g de Fibres par Jour

J1	J2	J3	J4	J5
Midi endives (+ vinaigrette) bœuf bourguignon carottes riz pilaf Camembert ananas sirop	**Midi** betteraves (+ vinaigrette) brandade de morue épinards pommes de terre Chèvre Pamplemousse (+ sucre)	**Midi** endives (+ vinaigrette) Chili con carne Vache qui rit ananas sirop	**Midi** betteraves (+ vinaigrette) spaghettis bolognaise haricots verts Camembert Fruits secs (abricots, figues)	**Midi** endives (+ vinaigrette) poulet courgettes Purée Chèvre poire
Soir Potage légumes couscous Semoule Yaourt poire	**Soir** Potage légumes sauté de veau courgettes + petits pois fromage blanc (+ sucre) Pomme	**Soir** Potage légumes sauté de dinde pommes de terre petits-suisses (+ sucre) pamplemousse	**Soir** Potage légumes Poisson épinards Riz yaourt (+ sucre) poire	**Soir** Potage légumes crevettes riz cantonnais petits pois fromage blanc (+ sucre) Fruits secs (abricots, figues)

Dosage des acides gras à chaînes courtes en mM

sample	acétate	propionate	isobutyrate	butyrate	isovalérate	valérate	isocaproate	Caproate	collecte	patient	regime
1J0	12,69	4,69	0,65	5,5	0,75	0,88	0,01	0,57	point_0	1	sequences 10-40
1J5	27,53	8,53	1,01	15,32	0,78	1,56	0,01	0,7	point_aprés_1er_regime	1	sequences 10-40
1J21	19,08	6,55	1,06	6,49	0,74	1,06	0,02	0,62	point_avant_2eme_regime	1	sequences 10-40
1J25	21,86	5,82	0,99	6,55	0,63	0,74	0,04	0,5	point_aprés_2eme_regime	1	sequences 10-40
3J5	11,39	4,36	0,74	4,7	0,63	0,5	0,01	0,05	point_aprés_1er_regime	3	sequences 40-10
3J21	8,02	4,27	0,72	3,84	0,71	0,44	0	0,02	point_avant_2eme_regime	3	sequences 40-10
3J25	15,7	7,52	0,78	6,97	0,72	0,69	0,04	0,03	point_aprés_2eme_regime	3	sequences 40-10
4J0	15,44	5,41	1,31	5,72	1,22	0,97	0	0,38	point_0	4	sequences 40-10
4J5	12,08	4,14	1,15	2,82	1,04	0,33	0	0,15	point_aprés_1er_regime	4	sequences 40-10
4J21	15,14	5,5	1,47	4,72	1,44	0,95	0	0,25	point_avant_2eme_regime	4	sequences 40-10
4J25	20,97	6,79	1,35	7,25	1,15	1,26	0,06	1,09	point_aprés_2eme_regime	4	sequences 40-10
5J0	42,71	12,63	1,5	16,66	1,53	2,39	0,09	2,65	point_0	5	sequences 10-40
5J5	48,47	15,83	1,17	28,5	0,99	2,43	0,12	2,92	point_aprés_1er_regime	5	sequences 10-40
5J21	35,72	13,33	1,6	15,18	1,75	2,69	0,11	2,73	point_avant_2eme_regime	5	sequences 10-40
5J25	37,69	9,37	1,29	22,73	1,12	1,57	0,17	2,81	point_aprés_2eme_regime	5	sequences 10-40
6J0	24,83	10,82	3,19	20,36	4,81	3,11	0,4	1,54	point_0	6	sequences 40-10
6J5	23,25	9,43	1,37	9,12	2,17	1,55	0,22	0,86	point_aprés_1er_regime	6	sequences 40-10
6J22	23,39	7,77	1,5	7,69	1,71	1,48	0,03	1,1	point_avant_2eme_regime	6	sequences 40-10
6J26	27,32	14,31	2,91	15,65	2,8	2,63	0,04	1,34	point_aprés_2eme_regime	6	sequences 40-10
7J0	26,21	10,41	1,37	9,51	1,47	2,1	0,07	1,14	point_0	7	sequences 40-10
7J4	42,8	15,73	3,07	13,08	3,47	2,24	0,05	1,28	point_aprés_1er_regime	7	sequences 40-10
7J22	22,64	7,37	1,8	7,46	1,67	1,26	0,01	1,13	point_avant_2eme_regime	7	sequences 40-10
7J26	24,34	7,6	2,32	7,93	1,7	1,41	0,08	1,05	point_aprés_2eme_regime	7	sequences 40-10
8J0	5,66	1,87	0,57	1,11	0,55	0,37	0,12	0,12	point_0	8	sequences 10-40
8J5	26,23	8,6	2,41	7,79	2,64	1,79	0,09	0,79	point_aprés_1er_regime	8	sequences 10-40
8J22	77,9	21,39	6,61	11,8	8,49	5,83	0,68	1,4	point_avant_2eme_regime	8	sequences 10-40
8J26	36,72	13,25	3,41	14,3	3,94	2,78	0,21	1,15	point_aprés_2eme_regime	8	sequences 10-40
9J0	23,26	8,1	0,77	7,72	1,05	1,05	0,03	0,48	point_0	9	sequences 10-40
9J5	57,71	10,6	0,53	23,93	0,47	1,26	0,04	2,1	point_aprés_1er_regime	9	sequences 10-40
9J21	32,19	12,39	0,83	13,21	1,03	1,7	0,03	0,49	point_avant_2eme_regime	9	sequences 10-40
9J26	50,45	17,94	0,66	32,04	0,79	1,18	0,08	1,2	point_aprés_2eme_regime	9	sequences 10-40
10J0	45	12,32	2,78	11,27	4,16	2,61	0,07	0,93	point_0	10	sequences 10-40
10J5	35,52	12,57	1,41	12,9	1,93	1,68	0,03	0,14	point_aprés_1er_regime	10	sequences 10-40
10J22	35,75	9,41	0,45	21,94	0,35	1,18	0,06	0,04	point_avant_2eme_regime	10	sequences 10-40
10J26	49,3	17,88	1,02	15,73	1,07	1,54	0,02	0,08	point_aprés_2eme_regime	10	sequences 10-40
11J0	25,12	10,1	2,16	9,12	3,68	2,63	0,32	1,17	point_0	11	sequences 10-40
11J5	24,82	10,96	1,91	10,23	3,38	2,5	0,11	0,77	point_aprés_1er_regime	11	sequences 10-40
11J22	22,23	9,1	1,98	11,06	3,45	2,11	0,14	0,75	point_avant_2eme_regime	11	sequences 10-40
11J26	34,47	12,18	1,83	15,54	2,36	2,42	0,06	1,49	point_aprés_2eme_regime	11	sequences 10-40
12J0	20,97	11,39	1,42	9,84	2,37	1,96	0,03	0,53	point_0	12	sequences 40-10
12J5	31,67	18,33	1,21	12,53	1,94	2,64	0,13	0,83	point_aprés_1er_regime	12	sequences 40-10
12J21	18,9	10,47	1,35	8,6	2,39	1,86	0,05	0,39	point_avant_2eme_regime	12	sequences 40-10
12J25	26,11	13,94	1,52	13,42	2,34	2,52	0	0,97	point_aprés_2eme_regime	12	sequences 40-10
13J0	32,17	9,62	1,09	13,14	1,52	1,41	0,85	0,2	point_0	13	sequences 40-10
13J5	31,5	11,46	1,5	14,79	2,14	1,55	0,33	0,08	point_aprés_1er_regime	13	sequences 40-10
13J22	33,37	11,45	2	31,19	2,77	1,88	0,1	0,1	point_avant_2eme_regime	13	sequences 40-10
13J26	55,32	15,69	0,66	42,72	0,75	1,63	0,11	0,32	point_aprés_2eme_regime	13	sequences 40-10
14J0	60,31	10,35	0,38	26	0,45	1,25	0,19	0,62	point_0	14	sequences 10-40
14J5	38,36	11,55	1,51	20,88	1,8	1,7	0,13	0,53	point_aprés_1er_regime	14	sequences 10-40
14J21	27,36	12,15	0,73	9,94	0,94	1,35	0,15	0,31	point_avant_2eme_regime	14	sequences 10-40
14J25	12,56	6,51	0,53	2,06	0,65	0,58	0,03	0,08	point_aprés_2eme_regime	14	sequences 10-40
15J0	9,4	4,02	0,61	2,77	0,74	0,55	0,08	0,17	point_0	15	sequences 10-40
15J5	13,65	5,76	0,64	2,91	0,85	0,73	0,07	0,09	point_aprés_1er_regime	15	sequences 10-40
15J21	16,52	7,84	0,69	5,87	1,01	0,82	0,17	0,19	point_avant_2eme_regime	15	sequences 10-40
15J25	35,59	15,23	0,58	11,9	0,68	1,75	0,06	0,2	point_aprés_2eme_regime	15	sequences 10-40
16J0	31,9	13	1,28	9,84	1,64	1,78	0,06	0,07	point_0	16	sequences 10-40
16J5	59,9	13,95	0,5	20,14	0,43	1,39	0,01	0,61	point_aprés_1er_regime	16	sequences 10-40
16J21	31,05	13,1	1,08	8,73	1,41	1,29	0,06	0,49	point_avant_2eme_regime	16	sequences 10-40
16J25	16,55	7,51	1,97	9,1	2,35	1,5	0,11	0,69	point_aprés_2eme_regime	16	sequences 10-40
17J0	22,9	7,85	1	7,87	1,49	1,37	0,05	0,81	point_0	17	sequences 10-40
17J5	14,25	6,19	1,42	5,68	1,94	1,32	0,04	0,44	point_aprés_1er_regime	17	sequences 10-40
17J21	54,75	11,24	0,32	38,23	0,43	0,91	0,06	1,81	point_avant_2eme_regime	17	sequences 10-40
18J0	16,95	10,08	0,93	6,76	1,23	1,12	0,03	0,08	point_0	18	sequences 40-10
18J5	17,16	5,62	1,01	8,96	1,11	0,73	0,05	0,04	point_aprés_1er_regime	18	sequences 40-10
18J21	27,71	12,68	1,38	13,84	1,24	1,48	0,01	0,15	point_avant_2eme_regime	18	sequences 40-10
18J25	24,16	8,3	1,04	12,51	1,09	1,19	0,02	0,04	point_aprés_2eme_regime	18	sequences 40-10
19J0	33,4	14,21	1,25	19,84	1,51	1,52	0,03	0,05	point_0	19	sequences 40-10
19J5	52,4	28,13	1,42	40,23	1,65	2,28	0,05	0,06	point_aprés_1er_regime	19	sequences 40-10
19J21	45,75	24,26	1,6	49,57	1,98	2,38	0,02	0,06	point_avant_2eme_regime	19	sequences 40-10
19J25	33,05	15,96	1,75	26,46	2,36	1,77	0,05	0,05	point_aprés_2eme_regime	19	sequences 40-10
20J0	26,38	15,94	1,88	12,38	2,64	2,05	0,06	0,68	point_0	20	sequences 40-10
20J5	27,19	13,37	1,35	13,56	1,84	1,46	0,07	0,5	point_aprés_1er_regime	20	sequences 40-10
20J21	17,5	6,39	1,75	8,21	2,65	1,43	0,06	0,27	point_avant_2eme_regime	20	sequences 40-10
20J25	20,76	9,2	1,85	11,21	2,84	1,84	0,02	0,46	point_aprés_2eme_regime	20	sequences 40-10

Dosage des acides gras à chaînes courtes en %

sample	acétate	propionate	isobutyrate	butyrate	isovalérate	valérate	isocaproate	Caproate	point_PCR	patient	regime
1J0	49,3%	18,2%	2,5%	21,4%	2,9%	3,4%	0,0%	2,2%	point_0	1	sequences 10-40
1J5	49,7%	15,4%	1,8%	27,6%	1,4%	2,8%	0,0%	1,3%	point_aprés_1er_regime	1	sequences 10-40
1J21	53,6%	18,4%	3,0%	18,2%	2,1%	3,0%	0,1%	1,7%	point_avant_2eme_regime	1	sequences 10-40
1J25	58,9%	15,7%	2,7%	17,6%	1,7%	2,0%	0,1%	1,3%	point_aprés_2eme_regime	1	sequences 10-40
3J5	50,9%	19,5%	3,3%	21,0%	2,8%	2,2%	0,0%	0,2%	point_aprés_1er_regime	3	sequences 40-10
3J21	44,5%	23,7%	4,0%	21,3%	3,9%	2,4%	0,0%	0,1%	point_avant_2eme_regime	3	sequences 40-10
3J25	48,4%	23,2%	2,4%	21,5%	2,2%	2,1%	0,1%	0,1%	point_aprés_2eme_regime	3	sequences 40-10
4J0	50,7%	17,8%	4,3%	18,8%	4,0%	3,2%	0,0%	1,2%	point_0	4	sequences 40-10
4J5	55,6%	19,1%	5,3%	13,0%	4,8%	1,5%	0,0%	0,7%	point_aprés_1er_regime	4	sequences 40-10
4J21	51,4%	18,7%	5,0%	16,0%	4,9%	3,2%	0,0%	0,8%	point_avant_2eme_regime	4	sequences 40-10
4J25	52,5%	17,0%	3,4%	18,2%	2,9%	3,2%	0,2%	2,7%	point_aprés_2eme_regime	4	sequences 40-10
5J0	53,3%	15,8%	1,9%	20,8%	1,9%	3,0%	0,1%	3,3%	point_0	5	sequences 10-40
5J5	48,3%	15,8%	1,2%	28,4%	1,0%	2,4%	0,1%	2,9%	point_aprés_1er_regime	5	sequences 10-40
5J21	48,9%	18,2%	2,2%	20,8%	2,4%	3,7%	0,2%	3,7%	point_avant_2eme_regime	5	sequences 10-40
5J25	49,1%	12,2%	1,7%	29,6%	1,5%	2,0%	0,2%	3,7%	point_aprés_2eme_regime	5	sequences 10-40
6J0	36,0%	15,7%	4,6%	29,5%	7,0%	4,5%	0,6%	2,2%	point_0	6	sequences 40-10
6J5	48,5%	19,7%	2,9%	19,0%	4,5%	3,2%	0,5%	1,8%	point_aprés_1er_regime	6	sequences 40-10
6J22	52,4%	17,4%	3,4%	17,2%	3,8%	3,3%	0,1%	2,5%	point_avant_2eme_regime	6	sequences 40-10
6J26	40,8%	21,4%	4,3%	23,4%	4,2%	3,9%	0,1%	2,0%	point_aprés_2eme_regime	6	sequences 40-10
7J0	50,1%	19,9%	2,6%	18,2%	2,8%	4,0%	0,1%	2,2%	point_0	7	sequences 40-10
7J4	52,4%	19,2%	3,8%	16,0%	4,2%	2,7%	0,1%	1,6%	point_aprés_1er_regime	7	sequences 40-10
7J22	52,2%	17,0%	4,2%	17,2%	3,9%	2,9%	0,0%	2,6%	point_avant_2eme_regime	7	sequences 40-10
7J26	52,4%	16,4%	5,0%	17,1%	3,7%	3,0%	0,2%	2,3%	point_aprés_2eme_regime	7	sequences 40-10
8J0	54,6%	18,0%	5,5%	10,7%	5,3%	3,6%	1,2%	1,2%	point_0	8	sequences 10-40
8J5	52,1%	17,1%	4,8%	15,5%	5,2%	3,6%	0,2%	1,6%	point_aprés_1er_regime	8	sequences 10-40
8J22	58,1%	16,0%	4,9%	8,8%	6,3%	4,3%	0,5%	1,0%	point_avant_2eme_regime	8	sequences 10-40
8J26	48,5%	17,5%	4,5%	18,9%	5,2%	3,7%	0,3%	1,5%	point_aprés_2eme_regime	8	sequences 10-40
9J0	54,8%	19,1%	1,8%	18,2%	2,5%	2,5%	0,1%	1,1%	point_0	9	sequences 10-40
9J5	59,7%	11,0%	0,5%	24,8%	0,5%	1,3%	0,0%	2,2%	point_aprés_1er_regime	9	sequences 10-40
9J21	52,0%	20,0%	1,3%	21,4%	1,7%	2,7%	0,0%	0,8%	point_avant_2eme_regime	9	sequences 10-40
9J26	48,4%	17,2%	0,6%	30,7%	0,8%	1,1%	0,1%	1,2%	point_aprés_2eme_regime	9	sequences 10-40
10J0	56,9%	15,6%	3,5%	14,2%	5,3%	3,3%	0,1%	1,2%	point_0	10	sequences 10-40
10J5	53,7%	19,0%	2,1%	19,5%	2,9%	2,5%	0,0%	0,2%	point_aprés_1er_regime	10	sequences 10-40
10J22	51,7%	13,6%	0,7%	31,7%	0,5%	1,7%	0,1%	0,1%	point_avant_2eme_regime	10	sequences 10-40
10J26	56,9%	20,6%	1,2%	18,2%	1,2%	1,8%	0,0%	0,1%	point_aprés_2eme_regime	10	sequences 10-40
11J0	46,3%	18,6%	4,0%	16,8%	6,8%	4,8%	0,6%	2,2%	point_0	11	sequences 10-40
11J5	45,4%	20,0%	3,5%	18,7%	6,2%	4,6%	0,2%	1,4%	point_aprés_1er_regime	11	sequences 10-40
11J22	43,7%	17,9%	3,9%	21,8%	6,8%	4,2%	0,3%	1,5%	point_avant_2eme_regime	11	sequences 10-40
11J26	49,0%	17,3%	2,6%	22,1%	3,4%	3,4%	0,1%	2,1%	point_aprés_2eme_regime	11	sequences 10-40
12J0	43,2%	23,5%	2,9%	20,3%	4,9%	4,0%	0,1%	1,1%	point_0	12	sequences 40-10
12J5	45,7%	26,5%	1,7%	18,1%	2,8%	3,8%	0,2%	1,2%	point_aprés_1er_regime	12	sequences 40-10
12J21	42,9%	23,8%	3,1%	19,5%	5,4%	4,2%	0,1%	0,9%	point_avant_2eme_regime	12	sequences 40-10
12J25	42,9%	22,9%	2,5%	22,1%	3,8%	4,1%	0,0%	1,6%	point_aprés_2eme_regime	12	sequences 40-10
13J0	53,6%	16,0%	1,8%	21,9%	2,5%	2,4%	1,4%	0,3%	point_0	13	sequences 40-10
13J5	49,7%	18,1%	2,4%	23,3%	3,4%	2,4%	0,5%	0,1%	point_aprés_1er_regime	13	sequences 40-10
13J22	40,3%	13,8%	2,4%	37,6%	3,3%	2,3%	0,1%	0,1%	point_avant_2eme_regime	13	sequences 40-10
13J26	47,2%	13,4%	0,6%	36,5%	0,6%	1,4%	0,1%	0,3%	point_aprés_2eme_regime	13	sequences 40-10
14J0	60,6%	10,4%	0,4%	26,1%	0,5%	1,3%	0,2%	0,6%	point_0	14	sequences 10-40
14J5	50,2%	15,1%	2,0%	27,3%	2,4%	2,2%	0,2%	0,7%	point_aprés_1er_regime	14	sequences 10-40
14J21	51,7%	23,0%	1,4%	18,8%	1,8%	2,6%	0,3%	0,6%	point_avant_2eme_regime	14	sequences 10-40
14J25	54,6%	28,3%	2,3%	9,0%	2,8%	2,5%	0,1%	0,3%	point_aprés_2eme_regime	14	sequences 10-40
15J0	51,3%	21,9%	3,3%	15,1%	4,0%	3,0%	0,4%	0,9%	point_0	15	sequences 10-40
15J5	55,3%	23,3%	2,6%	11,8%	3,4%	3,0%	0,3%	0,4%	point_aprés_1er_regime	15	sequences 10-40
15J21	49,9%	23,7%	2,1%	17,7%	3,1%	2,5%	0,5%	0,6%	point_avant_2eme_regime	15	sequences 10-40
15J25	53,9%	23,1%	0,9%	18,0%	1,0%	2,7%	0,1%	0,3%	point_aprés_2eme_regime	15	sequences 10-40
16J0	53,6%	21,8%	2,1%	16,5%	2,8%	3,0%	0,1%	0,1%	point_0	16	sequences 10-40
16J5	61,8%	14,4%	0,5%	20,8%	0,4%	1,4%	0,0%	0,6%	point_aprés_1er_regime	16	sequences 10-40
16J21	54,3%	22,9%	1,9%	15,3%	2,5%	2,3%	0,1%	0,9%	point_avant_2eme_regime	16	sequences 10-40
16J25	41,6%	18,9%	5,0%	22,9%	5,9%	3,8%	0,3%	1,7%	point_aprés_2eme_regime	16	sequences 10-40
17J0	52,8%	18,1%	2,3%	18,2%	3,4%	3,2%	0,1%	1,9%	point_0	17	sequences 10-40
17J5	45,6%	19,8%	4,5%	18,2%	6,2%	4,2%	0,1%	1,4%	point_aprés_1er_regime	17	sequences 10-40
17J21	50,8%	10,4%	0,3%	35,5%	0,4%	0,8%	0,1%	1,7%	point_avant_2eme_regime	17	sequences 10-40
18J0	45,6%	27,1%	2,5%	18,2%	3,3%	3,0%	0,1%	0,2%	point_0	18	sequences 40-10
18J5	49,5%	16,2%	2,9%	25,8%	3,2%	2,1%	0,1%	0,1%	point_aprés_1er_regime	18	sequences 40-10
18J21	47,4%	21,7%	2,4%	23,7%	2,1%	2,5%	0,0%	0,3%	point_avant_2eme_regime	18	sequences 40-10
18J25	50,0%	17,2%	2,2%	25,9%	2,3%	2,5%	0,0%	0,1%	point_aprés_2eme_regime	18	sequences 40-10
19J0	46,5%	19,8%	1,7%	27,6%	2,1%	2,1%	0,0%	0,1%	point_0	19	sequences 40-10
19J5	41,5%	22,3%	1,1%	31,9%	1,3%	1,8%	0,0%	0,0%	point_aprés_1er_regime	19	sequences 40-10
19J21	36,4%	19,3%	1,3%	39,5%	1,6%	1,9%	0,0%	0,0%	point_avant_2eme_regime	19	sequences 40-10
19J25	40,6%	19,6%	2,1%	32,5%	2,9%	2,2%	0,1%	0,1%	point_aprés_2eme_regime	19	sequences 40-10
20J0	42,5%	25,7%	3,0%	20,0%	4,3%	3,3%	0,1%	1,1%	point_0	20	sequences 40-10
20J5	45,8%	22,5%	2,3%	22,9%	3,1%	2,5%	0,1%	0,8%	point_aprés_1er_regime	20	sequences 40-10
20J21	45,7%	16,7%	4,6%	21,5%	6,9%	3,7%	0,2%	0,7%	point_avant_2eme_regime	20	sequences 40-10
20J25	43,1%	19,1%	3,8%	23,3%	5,9%	3,8%	0,0%	1,0%	point_aprés_2eme_regime	20	sequences 40-10