Revealing microbial species diversity using sequence capture by hybridization

Sophie Marre, Cyrielle Gasc, Camille Forest, Yacine Lebbaoui, Pascale Mosoni, Pierre Peyret

MICROBIAL GENOMICS

RESEARCH ARTICLE

Marre *et al.*, *Microbial Genomics* 2021;7:000714

DOI 10.1099/mgen.0.000714

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Revealing microbial species diversity using sequence capture by hybridization

Sophie Marre†, Cyrielle Gasc†,‡, Camille Forest, Yacine Lebbaoui, Pascale Mosoni and Pierre Peyret*

Abstract

Targeting small parts of the 16S rDNA phylogenetic marker by metabarcoding reveals microorganisms of interest but cannot achieve a taxonomic resolution at the species level, precluding further precise characterizations. To identify species behind operational taxonomic units (OTUs) of interest, even in the rare biosphere, we developed an innovative strategy using gene capture by hybridization. From three OTU sequences detected upon polyphenol supplementation and belonging to the rare biosphere of the human gut microbiota, we revealed 59 nearly full-length 16S rRNA genes, highlighting high bacterial diversity hidden behind OTUs while evidencing novel taxa. Inside each OTU, revealed 16S rDNA sequences could be highly distant from each other with similarities down to 85%. We identified one new family belonging to the order *Clostridiales*, 39 new genera and 52 novel species. Related bacteria potentially involved in polyphenol degradation have also been identified through genome mining and our results suggest that the human gut microbiota could be much more diverse than previously thought.

INTRODUCTION

Identifying the microbial taxa present in complex biological samples is the most frequently encountered challenge in microbiology. To achieve this objective, amplifying and sequencing the 16S rRNA gene-variable regions, also called metabarcoding or amplicon sequencing, has become the most widely used molecular method to survey and compare microbial communities in a cultivation-independent manner [1]. High-throughput DNA sequencing has made microbial community profiling affordable and easy to perform routinely. This approach has led to the discovery of many unexpected evolutionary lineages [2].

Unfortunately, metabarcoding cannot achieve taxonomic resolution at the species or strain level [3, 4]. Indeed, the short-read length of the most commonly used second-generation sequencing platforms (e.g. Illumina) generally allows sequence assignment at the family level and in some favourable cases at the genus level, thus reducing the accuracy and reliability of characterizing microbial communities [5, 6]. Short reads often result in incorrect or inaccurate taxonomic assignment of amplicons, and only reconstruction of complete sequences of rRNA genes allows taxonomic resolution to be achieved at the species or strain level [7]. Several methodological [8–10] or bioinformatics [11, 12] strategies have been developed to recover complete or near-complete rRNA genes, but all of these strategies suffer from major limitations linked to the difficulties inherent in comprehensively exploring complex microbial diversity. Shotgun reads obtained from metagenomic studies are a source of sequences that are not subject to PCR bias, thereby enhancing phylogenetic assignment [7]. However, shotgun sequencing of metagenomic samples preferentially provides sequences of dominant microorganisms, thus diminishing the phylogenetic description of microbial communities. Even with ultra-deep sequencing, it remains difficult to access subdominant microorganisms and rare biospheres that could play essential roles in the explored environments [13]. Furthermore, managing a large amount of data and conducting bioinformatics analyses to efficiently explore metagenomic samples are not trivial undertakings. Even with the decreasing sequencing costs, such an approach is expensive.

Received 21 June 2021; Accepted 11 October 2021; Published 09 December 2021

Author affiliations: [1]Université Clermont Auvergne, INRAE, MEDIS, F-63000, Clermont-Ferrand, France.

***Correspondence:** Pierre Peyret, pierre.peyret@uca.fr

Keywords: microbial diversity; species identification; gene capture by hybridization; 16S rRNA gene; polyphenol degradation; human gut microbiota; rare biosphere.

Abbreviations: OTU, operational taxonomic unit; SCFA, short-chain fatty acid.

‡Present address: MaaT Pharma, F-69007 LYON, France.

The data supporting the findings of this study are available in the European Nucleotide Archive under study accession number PRJEB43604.

†These authors contributed equally to this work

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Supplementary data are available in the online version of this article.

A recently developed alternative is the use of 'third-generation' long-read sequencing technologies after PCR amplification to obtain full-length 16S rRNA gene sequences. This approach improves taxonomic and phylogenetic resolution by increasing the number of informative sites sequenced while continuing to use universal marker genes. New long-read sequencing technologies, namely the Pacific Biosciences (PacBio) and Oxford Nanopore technologies, can sequence the entire 16S rRNA gene, but high error rates have limited their attractiveness. For now, microbiome analysis pipelines take advantage of PacBio circular consensus sequencing (CCS) technology to sequence and error-correct full-length bacterial 16S rRNA genes [14, 15]. However, comparative analyses have revealed that the PacBio data showed a weaker relationship with the reference whole-metagenome shotgun datasets than profiles generated by short-read sequencing platforms [16]. In addition, the high costs impel most researchers to limit their use of long-read sequencing, and the insufficient sequencing depth that does not allow access to subdominant and rare microorganisms remains a potential issue.

Hybridization capture has proven to be an innovative and efficient tool for targeting and enriching whole genomes, specific DNA regions or biomarkers in complex DNA mixtures [17]. Functional microbial markers have been enriched from various ecosystems, showing that such an approach can be more sensitive than the usual molecular methods for detecting rare sequences [18–20]. More recently, capture by hybridization has also been applied to the gold standard phylogenetic marker 16S rRNA gene for microbiota profiling [21, 22]. In the present study, we applied gene capture by hybridization to gain phylogenetic resolution of metabarcoding-derived operational taxonomic units (OTUs) and precisely identify OTUs at the species level. In a previous study, using V3–V5 metabarcoding, we revealed OTUs from the rare biosphere of the human gut microbiota (under 0.1 %) that are potentially involved in the metabolism of polyphenols and, thus, the production of bioactive compounds linked to beneficial health effects [23]. Unfortunately, these OTUs were identified at the low-resolution of the family *Lachnospiraceae* or were unclassified in the order *Clostridiales*. We applied gene capture by hybridization on metagenomic samples using a reduced set of specific probes targeting these 450 bp long OTU sequences, allowing adjacent sequence enrichment for nearly full-length 16S rRNA gene reconstruction. Using this strategy, we revealed a complex microbial species diversity underlying the OTU sequences. Species description provides new insights for further research, providing a better understanding of the functions of these microorganisms.

Impact Statement

Achieving microbial species-level identification improves the ecological and/or clinical relevance of the results compared to identification at higher taxonomic levels. Exploring polyphenol-degrading bacteria, we revealed an important hidden microbial diversity behind 16S operational taxonomic unit sequences from the human gut microbiota rare biosphere using gene capture by hybridization. Obtaining such precision could not be resolved by current methods, including shotgun metagenomics and profiling by third-generation sequencing.

METHODS

Sequencing library construction

DNA from human faeces was extracted using the QIAamp DNA Stool Mini Kit (Qiagen). DNA purity was checked using a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific). DNA integrity was confirmed by electrophoresis on 0.7% agarose gels, and the DNA was quantified by using the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific).

Sequencing libraries were constructed using the Nextera XT DNA Library Preparation Kit (Illumina) according to the manufacturer's instructions.

Probe design

Sequences of three OTUs of interest (identified in a previous study [23]) were targeted using probes. Thirty-mer probes (Table 1) were designed using KASpOD software [24]. Adaptor sequences were added to the ends of the probes to enable their amplification by PCR, resulting in 'ATCGCACCAGCGTGT-N_X-CACTGCGGCTCCTCA' sequences, with N_X representing OTU-specific capture probes. Biotinylated RNA capture probes were then synthesized as described by Ribière *et al.* [25]. In brief, adaptors containing the T7 promoter were added to 16S rRNA gene-specific capture probes via ligation-mediated PCR, and the final biotinylated RNA probes were obtained after *in vitro* transcription and purification.

Hybridization capture targeting the OTU sequences and sequencing

To perform hybridization capture, 2.5 µg of salmon sperm DNA (Ambion) and 500 ng of Illumina libraries were mixed, denatured for 5 min at 95 °C and incubated for 5 min at 65 °C before adding 13 µl of prewarmed (65 °C) 2× hybridization buffer (10× SSPE, 10× Denhardt's solution, 10 mM EDTA and 0.2% SDS) and 500 ng of prewarmed (65 °C) biotinylated RNA probes. After hybridization at 65 °C for 24 h, the probe/target heteroduplexes were captured using 500 ng of washed streptavidin-coated paramagnetic beads (Dynabeads M-280 Streptavidin; Invitrogen). The beads were collected using a magnetic stand (Ambion) and washed once at room temperature with 500 µl of 1× SSC/0.1% SDS and three times at 65 °C with 500 µl of prewarmed 0.1× SSC/0.1% SDS. The captured fragments were eluted with 50 µl of 0.1 M NaOH. After magnetic bead collection, the DNA supernatant was transferred to a sterile tube containing 70 µl of 1 M Tris-HCl (pH 7.5) and PCR-amplified using primers complementary to the library adapters (TS-PCR Oligo 1, 5′-AATGATAC GGCGACCACCGAGA-3′; and TS-PCR Oligo 2, 5′-CAAG

Table 1. Probe sequences targeting OTUs 146, 393 and 1761

Targeted OTUs	Probe name	Probe sequence
146	146 R1_57–86	ACGCCGCGTGAGTGAAGAAGTATTTCGGTA
	146 R1_90–119	AAAGCTCTATCAGCAGGGAAGAAGAAATGA
	146 R1_121–150	GGTACCTGACTAAGAAGCCCCGGCTAACTA
	146 R1_221–250	GACGGTGAAGCAAGTCTGAAGTGAAAGGTT
	146 R2_1–30	AAGGCGGCTTACTGGACTGTAACTGACGTT
	146 R2_71–100	TGGTAGTCCACGCCGTAAACGATGATTACT
	146 R2_107–136	TGGTGGATATGGATCCATCGGTGCCGCAGC
	393 R1_15–44	CAGTGGGGAATATTGCACAATGGAGGAAAC
393	393 R1_71–100	AAGAAGTAATTCGTTATGTAAAGCTCTATC
	393 R1_110–139	GATAGTGACGGTACCTGACTAAGAAGCTCC
	393 R1_221–250	TGGCAAGGCAAGTCAGATGTGAAAGCCCGG
	393 R2_1–30	GAAGGCGGCTTACTGGACTGTAACTGACAC
	393 R2_113–142	CCCACAGGGCTTCGGTGCCGCAGCAAACGC
393	1761 R1_55–84	CGACGCCGCGTGAGCGAAGAAGTATTTCGG
	1761 R1_104–133	AGGGAAGATAATGACGGTACCTGACTAAGA
	1761 R1_221–250	CGGGATATCAAGTCAGAAGTGAAAATTACG
	1761 R2_1–30	GAAGGCGGCTTGCTGGGCTTTTACTGACGC
	1761 R2_60–89	GATGAGATACCCTGGTAGTCCACGCCGTAA
	1761 R2_112–141	GGATTGACCCCTTCCGTGCCGGAGTAAACA
	1761 R2_171–200	CGCAAGATTGAAACTTAAATGAATTGACGG

CAGAAGACGGCATACGAG-3′). To increase the enrichment efficiency, a second round of hybridization capture was performed using the first-round capture products. The enriched DNA was then sequenced using Illumina MiSeq 2×250 bp runs. Reads were deposited in the European Nucleotide Archive under study accession number PRJEB43604.

PCR experiments and Sanger sequencing of amplicons

Specific primers targeting nearly full-length reconstructed 16S rRNA genes were designed using the KASpOD algorithm [24]. The final 25 µl PCR mixture consisted of 5 ng DNA (M1 to M5), 1 µM dNTPs, 1 µM of the corresponding primers and 2 units of GoTaq DNA polymerase (Promega). The PCR conditions were as follows: 5 min at 94 °C, followed by 30 cycles of 94 °C for 15 s, the annealing temperature (50–55 °C) for 15 s and 72 °C for 90 s. After the final cycle, the temperature was maintained at 72 °C for 7 min to allow completion of synthesis of the amplified products. For very-low-abundance 16S rRNA genes, a first PCR using the universal primers 27F and R1492 was followed by a second nested PCR using specific primers. The PCR products were then visualized on ethidium bromide-stained agarose gels (1%). DNA fragments were purified with the QIAquick Gel Extraction kit (Qiagen) and cloned into pCR II-TOPO (Invitrogen). Five clones for each PCR product were then Sanger-sequenced.

Bioinformatic and phylogenetic analyses

Reads were scanned for library adaptors and quality-filtered using the PRINSEQ-lite PERL script [26] prior to analysis. 16S rRNA gene reconstruction and OTU clustering from the five samples were performed using EMIRGE 0.60 [11]. Taxonomic classification of the sequences was then performed with RDP Classifier [27] using the Silva [28] database 119 release with the confidence cut-off set at 0.5. The pipeline is available on GitHub (https://github.com/SoMarre/CaptOTU). Phylogenetic analysis was conducted using a phylogeny analysis pipeline [29]. The candidate 16S rRNA gene sequences were first submitted in fasta format; sequences were aligned with MUSCLE [30], and the aligned sequences were curated with Gblocks [31]; the phylogenetic tree was reconstructed with PhyML by using the maximum-likelihood method [32]; and the reconstructed phylogenetic tree was visualized and rendered by FigTree v1.4.3 [33]. A similarity search was conducted using the BLAST algorithm [34] with default parameters in the GenBank database. We used the identity thresholds defined by Yarza *et al.* [7] to evaluate novel taxa (i.e. 97% for species, 94.5% for genus, 86.5% for family,

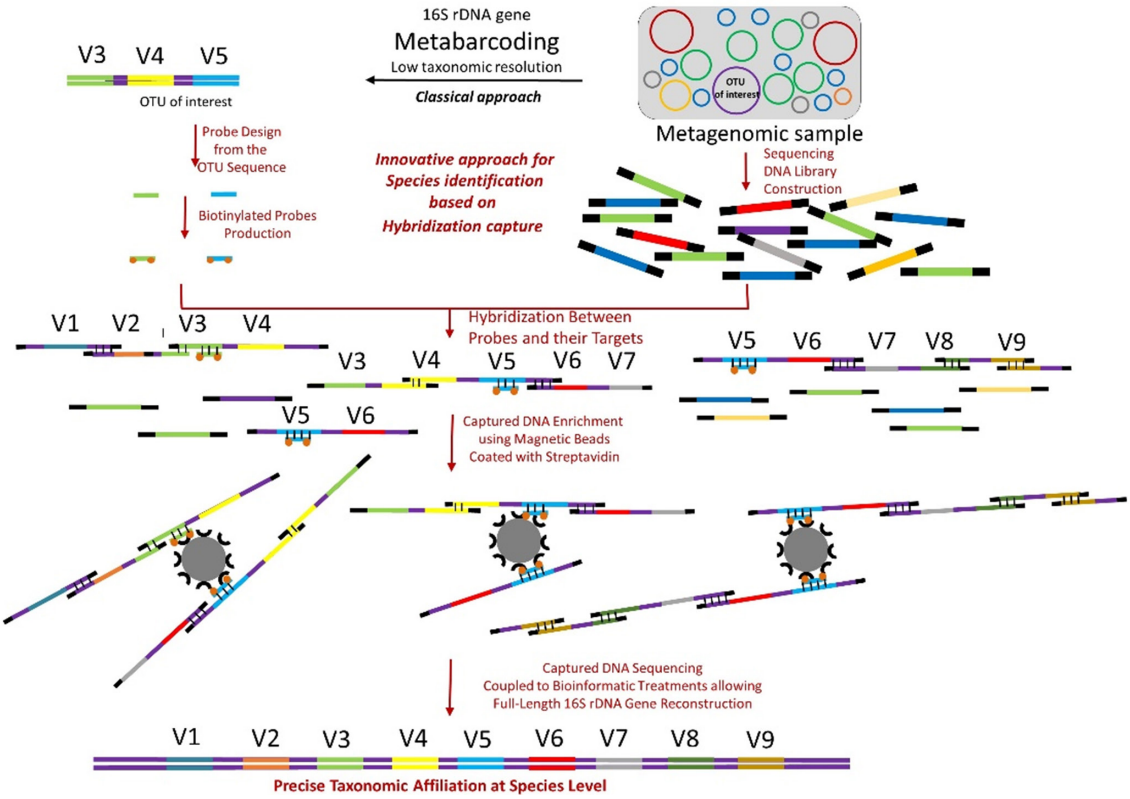

Fig. 1. Experimental scheme to reconstruct full-length 16S rRNA genes from selected short metabarcoding OTU sequences. The principle of the method involves several steps: specific probes are first designed to target OTUs of interest. In the present study, we selected three OTUs that were previously obtained by a metabarcoding approach targeting the V3–V5 region of the 16S rRNA gene [23]. Biotinylated probes are then hybridized to a sequencing library (Illumina library in this study) constructed from the explored metagenomic sample. Probe–target duplexes are enriched using magnetic beads coated with streptavidin, allowing interaction with the biotin incorporated in probes. DNA fragments harbouring OTU sequences targeted by the probes also act as probes targeting adjacent unknown flanking DNA regions. By this process, the unknown DNA regions can be enriched even by using short DNA fragments from the Illumina sequencing library. The enriched DNA fragments are then sequenced, allowing full-length 16S rRNA gene reconstruction for species-level assignment.

82.0% for order) confirmed by phylogenetic tree reconstruction. Text mining of annotated genomes was used to search for enzyme names associated with polyphenol metabolic pathways [35].

RESULTS

Innovative strategy for efficient full-length 16S rRNA gene reconstruction

We developed an innovative hybridization capture strategy aimed at reconstructing the full-length 16S rRNA gene from short metabarcoding OTU sequences (Fig. 1). Based on the ability of sequence capture to provide information beyond the target DNA regions, we used hybridization capture to study the unknown flanking regions of short metabarcoding sequences. In this situation, specifically designed capture probes hybridized to DNA fragments harbouring the known targeted sequence. These DNA fragments also acted simultaneously as probes for enrichment of the next adjacent DNA fragments as previously described [17, 20]. Indeed, even with small DNA fragments (500–600 bp) used for second-generation sequencing library construction, such a method applied to metagenomic samples captured flanking regions that exceeded kilobase pairs. This was particularly well adapted to our experiment with a phylogenetic marker that was approximately 1500 bp long. After the enrichment process, the captured DNA fragments from metagenomic second-generation sequencing libraries were directly sequenced. Bioinformatic analyses based on sequencing data allowed full-length 16S rRNA gene reconstruction and precise taxonomic affiliation at the species level.

Full-length 16S rRNA gene recovery from short metabarcoding sequences

Three OTUs (146, 393 and 1761) derived from the 16S rDNA V3–V5 region were identified in a previous *in vitro* study [23] involving four microbial faecal communities that were incubated with a mixture of purified apple polyphenols and polysaccharides (Data S1, available in the online version of this

Table 2. Enrichment efficiency using OTU sequence capture by hybridization

'Amplicon' results were obtained from a previous study using a metabarcoding approach [23]. 'Capture' indicates the hybridization-based innovative strategy developed in this study.

	M1		M2		M3		M4		M5
	Amplicon	Capture	Amplicon	Capture	Amplicon	Capture	Amplicon	Capture	Capture
Lachnospiraceae (%)	30.3	78	27.3	62.4	12.1	54.5	8.1	81.4	54
OTU 146 (%)	0.08	0.84	0.003	0.09	0.001	0	0.001	1.55	1.19
OTU 393 (%)	0.8	9.45	0.3	1.77	0.2	2.13	0.2	8.95	18.50
OTU 1761 (%)	0.2	0.08	0.6	0.74	0.08	0.58	0.08	0	0

paper). These OTUs were selected because their abundances increased during the fermentation process, suggesting their potential role in metabolizing one or both apple components, including polyphenols. Depending on the faecal sample, the greatest increase was from 0.02 to 0.6% (30-fold increase) for OTU 1761, 0.14% to 0.82% (5.9-fold increase) for OTU 393 and 0.015 to 0.083% (5.5-fold increase) for OTU 146. At such low relative abundances, these OTUs could be considered subdominant or rare microbial members of the community. We checked by similarity searches in the GenBank database that, as previously described, OTUs 146 and 393 were identified as members of the family *Lachnospiraceae* and OTU 1761 was only assigned at the order level as *Clostridiales*.

Using our KASpOD algorithm, we identified the 20 most specific probes (seven, six and seven probes targeting OTUs 146, 393 and 1761, respectively), allowing specific and efficient gene capture from the previously obtained 16S rDNA V3–V5 sequences (Table 1). We selected probes dispersed over the V3–V5 sequences with limited cross-hybridizations to improve enrichment efficiency and specificity. Five Illumina sequencing libraries were generated from DNA extracted from the faecal samples of five healthy individuals (M1 to M5: four faecal samples were obtained from a previous study [23]) and were subjected to *in vitro* incubation (48 h) with apple polyphenols and polysaccharides, and one faecal sample without any prior treatment was added (M5). The five sequencing libraries were subjected to enrichment of the three targeted OTU 16S rDNA sequences using the 20 designed probes. In a single gene capture by hybridization experiment, we efficiently enriched 16S rRNA gene sequences representing 25.4–44.2% of the total sequences, while total 16S rDNA

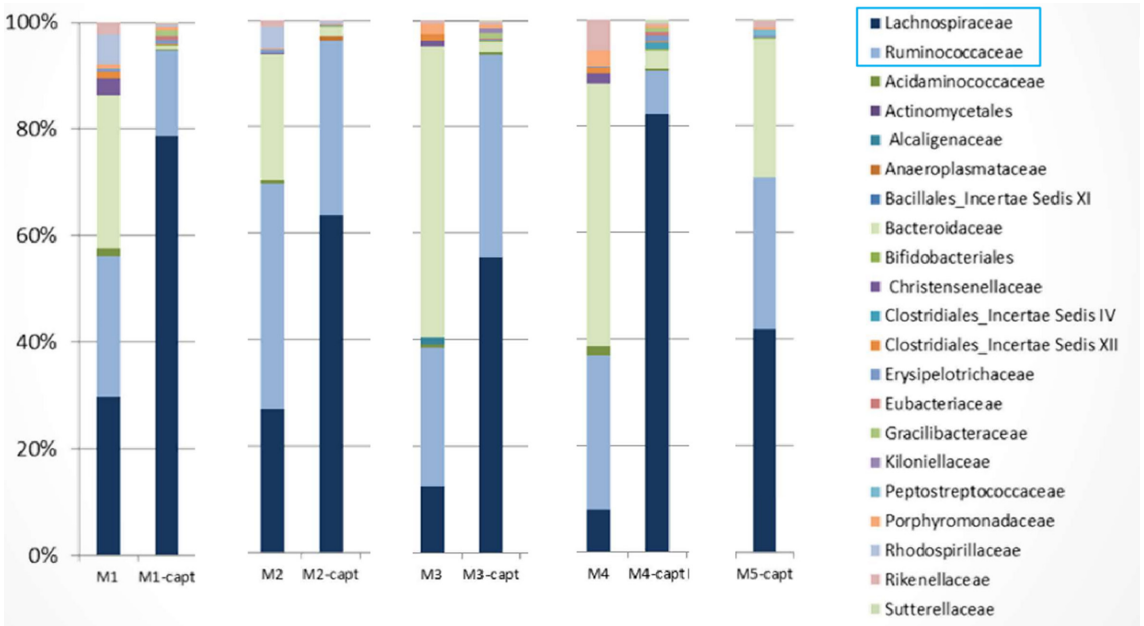

Fig. 2. Microbial community structures at the family level. M1 amplicon to M4 amplicon: results from a previous study [23] obtained by a V3–V5 rDNA region metabarcoding experiment for four subjects. M1 capture to M5 capture: gene capture by hybridization allowing nearly full-length 16S rRNA gene reconstruction applied to five metagenomic samples, including subjects M1–M4 from a previous study and a new subject, M5.

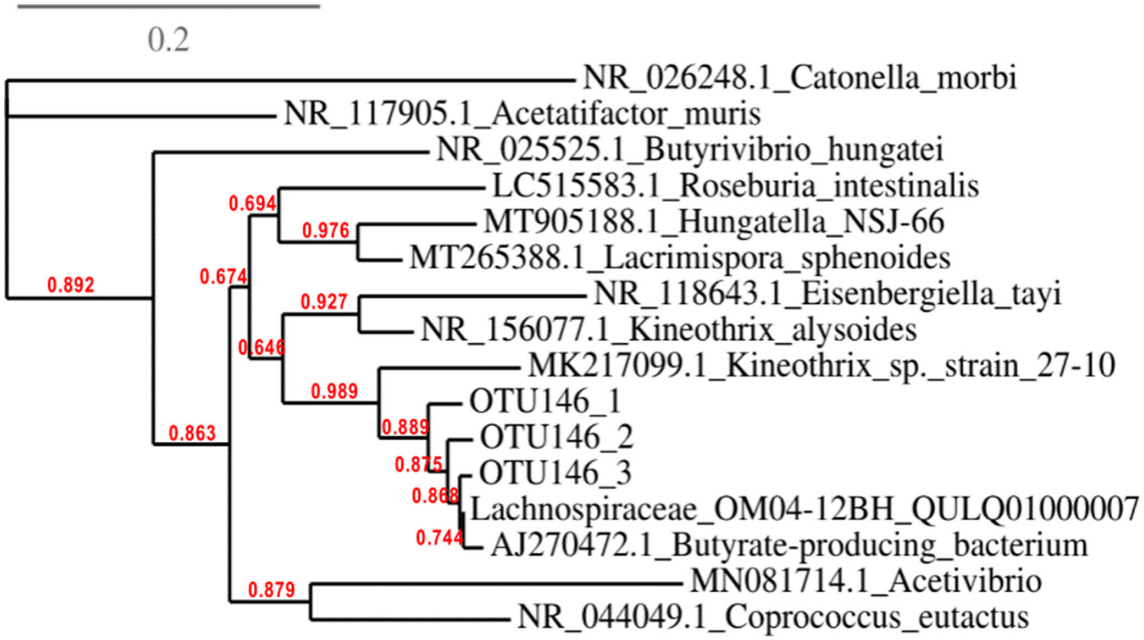

Fig. 3. Nearly full-length OTU 146 reconstruction (OTU146_1 to 3) positions in a 16S rDNA maximum-likelihood tree. The names for the representative species and their accession numbers are given. Numbers at nodes indicate branch support calculated with the Shimodaira–Hasegawa test. Bar, 0.2 nucleotide sequence divergence.

reads were usually found at less than 1% in the shotgun metagenomic data. After gene capture by hybridization, *Lachnospiraceae* sequences accounted for 54–81.4% of the total 16S rDNA sequences, in contrast to our previous metabarcoding study, where they represented 8.1–30.3% of the total sequences (Table 2). Other sequences representing a minority in terms of relative abundance were distributed in a few other families, as shown in Fig. 2. From the captured sequences, we were able to reconstruct 709, 533, 519, 468 and 527 nearly full-length 16S rDNA genes for the five metagenomic samples (M1 to M5, respectively). We identified 59 (i.e. 15, 36 and eight) nearly complete 16S rDNA sequences that shared ≥97% identity to OTUs 146, 393 and 1761, respectively (Data S2). These 59 16S sequences accounted for 35% of all the reconstructed 16S rDNA genes generated from the five metagenomic samples, confirming the targeted enrichment efficiency of this innovative approach. This strategy was estimated to enrich the sequencing library in the targeted OTU sequences by 44- to 422-fold depending on the starting metagenomic library.

Microbial diversity hidden behind OTU sequences

The 59 reconstructed sequences were positioned in phylogenetic trees to obtain more precise assignments.

Regarding OTU 146, three reconstructed sequences (OTU146_1 to OTU146_3, with lengths of 1124, 1133 and 1332 bp, respectively) showed 100% identity with the V3–V5 OTU 146 sequence and were close to each other, with 98–99% identity. This indicates that the three sequences probably belong to the same species. They could potentially represent three different strains, but we cannot exclude the possibility that a strain could harbour several variant copies of the 16S rDNA gene. These three sequences were assigned to the family *Lachnospiraceae* as previously suggested by the amplicon sequence analysis. However, the sequences are distant from known genera of this family, indicating that they belong to a new genus, the closest relative genus being *Kineothrix* (Fig. 3). Nevertheless, we identified very close sequences in the GenBank database showing 98% identity with our reconstructed sequences (identity based on the complete length). All the retrieved sequences originated from human gut samples obtained through different studies. Although most of them were annotated as 'uncultured', two 16S rDNA sequences originated from isolated strains: that is, strain T2-145 from a study focused on gut butyrate-producing bacteria (AJ270472.1) and the *Lachnospiraceae* bacterium OM04-12BH (QULQ01000007.1) isolated from Chinese samples, the latter appearing the closest to the reconstructed sequences. The 12 other sequences (OTU 146_4 to OTU 146_15, with lengths ranging from 1025 to 1367 bp) showed between 97.5 and 99% identity with the V3–V5 OTU 146 sequence. They were also close to the OTU 146_1, 146_2 and 146_3 sequences, with 89–94.5% identity. A similarity search identified close sequences annotated as 'uncultured' bacteria (97–98.6% identity between DQ793421, DQ806641, EF403928 and EU766058 and OTU146_9, _10, _11 and _14–15, respectively), all from the human gut, confirming the existence of these sequences in biological samples. Matrix distances among the 15 sequences for phylogenetic tree reconstruction highlighted

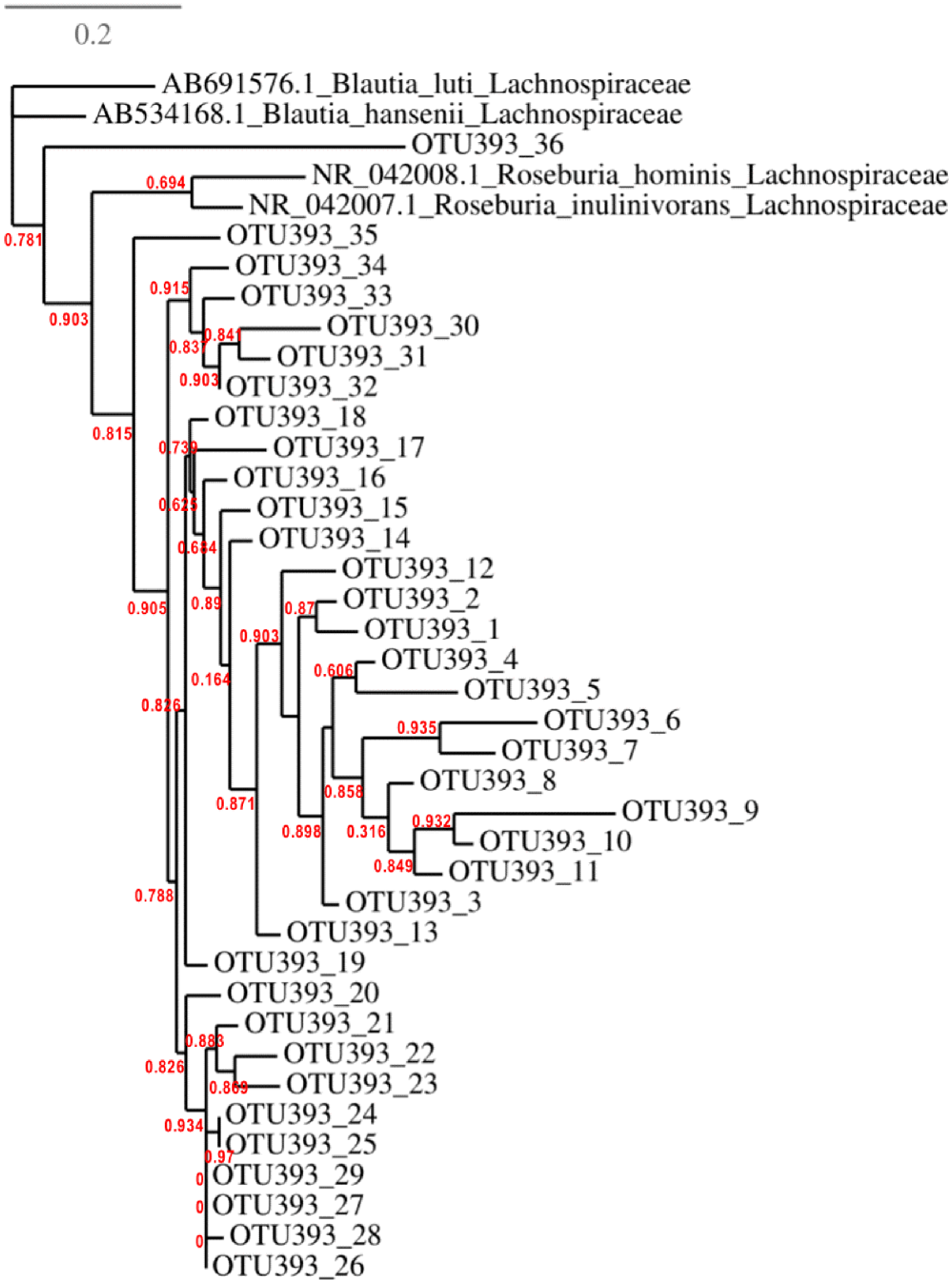

Fig. 4. Nearly full-length OTU 393 reconstruction (OTU393_1 to 36) positions in a 16S rDNA maximum-likelihood tree. The names for the representative species and their accession numbers are given. Numbers at nodes indicate branch support calculated with the Shimodaira–Hasegawa test. Bar, 0.2 nucleotide sequence divergence.

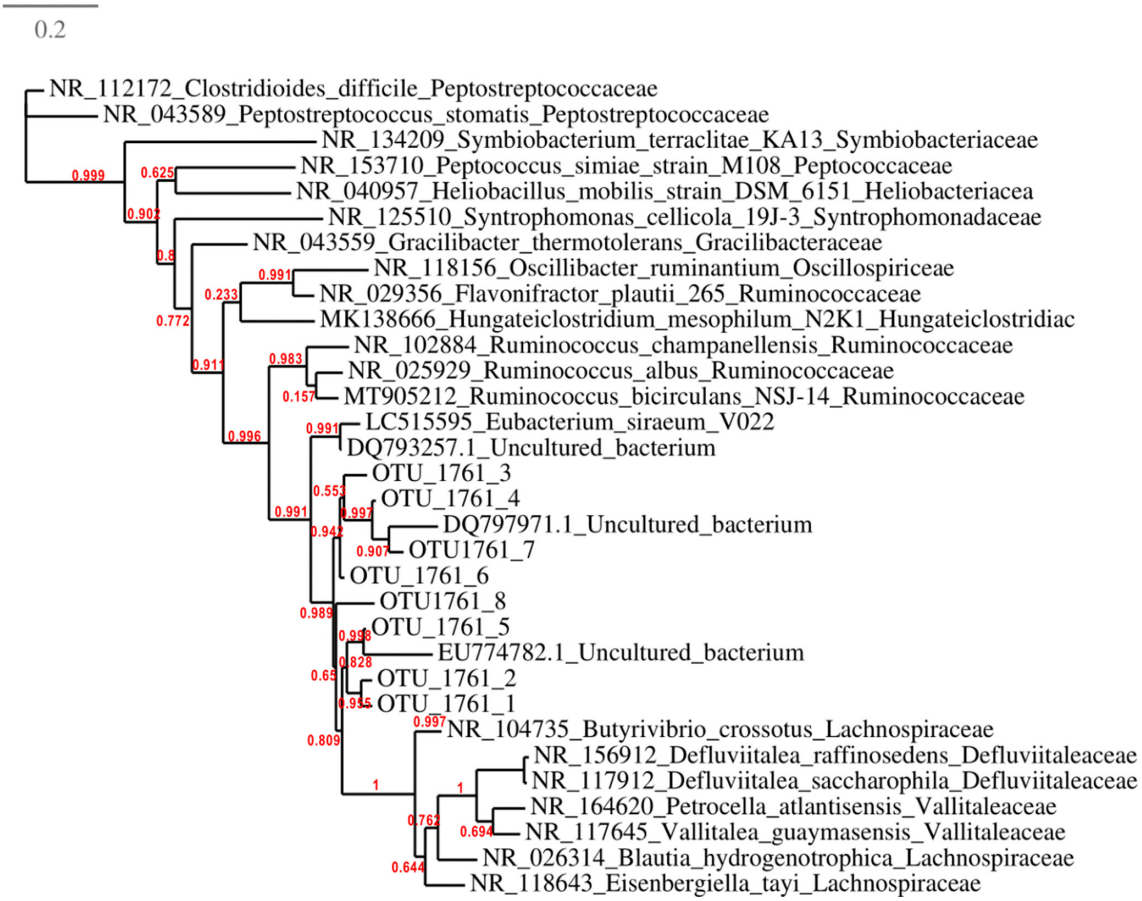

Fig. 5. Nearly full-length OTU 1761 reconstruction (OTU1761_1 to 8) positions in a 16S rDNA maximum-likelihood tree. The names for the representative species, their family affiliation and their accession numbers are given. Numbers at nodes indicate branch support calculated with the Shimodaira–Hasegawa test. Bar, 0.2 nucleotide sequence divergence.

11 new genera (Data S3). The OTU 146_14 and 146_15 sequences belonged to the same species with 97% identity. OTU 146_11 was very close to *Lachnoclostridium* and could represent a new species of this genus. We conclude that hidden behind the original 146 OTU sequence, 12 genera (11 of which are novel) and 14 novel species, all from the family *Lachnospiraceae*, were identified.

For OTU 393, 36 sequences were reconstructed, with a mean length of 1302 bp, comprising potentially 21 new genera (Fig. 4). In the phylogenetic tree, all the sequences were placed in the family *Lachnospiraceae*. Six sequences (OTU393_24 to OTU393_29) were close to each other, with a similarity percentage above 97%, indicating that they probably belonged to the same species. They could represent six different strains of the same species, but we cannot exclude the presence of 16S rDNA gene variants in one or several strains. A search in the GenBank Whole-Genome Shotgun database highlighted the high percentages of identity (97.07%–99.91%) between these reconstructed sequences and two cultivated species isolated from the human gut in the same study, namely *Clostridium* sp. AF36-4 (QTVH01000004.1) and *Clostridium* sp. AF37-5 (QUDR01000049.1). Numerous 16S rDNA sequences from the human gut annotated as 'uncultivated' also showed high identity with these sequences, reaching 100% identity between OTU393_29 and GQ898152 (Data S4). One sequence annotated *Eubacterium* sp. M5 (MT905187.1) also showed 97% to nearly 100% identity with these six reconstructed sequences. OTU393_11 also showed nearly 97% identity with 10 sequenced genomes, comprising *Eubacterium* sp. (QRVH01000128.1; QSEU01000032), *Lachnospira eligens* (QSHM01000002; QSBA01000007) and *Eubacterium eligens* (CP001104), but also numerous 'uncultivated' bacteria from the gut. OTU393_8, _14, _21, _23 and _33 were also very close to 'uncultured' bacteria from the human gut, corresponding to the GQ89683, DQ801296, DQ793227, EF403072 and DQ798951 sequences, respectively. The other reconstructed sequences were more divergent and might correspond to novel taxa. The closest sequences annotated as 'uncultivated' originated from the human gut and showed between 94.5 and 96.5% similarity. By targeting the reconstructed sequences with specific primers designed against these targets using a PCR and cloning-based approach, we obtained amplicon

sequences by Sanger sequencing that were very close to OTU393_26 (99,8%), _19 (98.5%), _27 (98%), _3 (98%) and _11 (98%), validating the efficiency of our strategy (Data S5). OTU393_3 is one of the sequences that showed nearly 100% identity with the original OTU 393 V3–V5 sequence. Sequence variability could be observed for OTU393_26 through five different PCR products obtained after cloning (Data S6). This suggests a larger microbial diversity at the strain level within species. However, we could not exclude PCR or sequencing errors even by using proofreading DNA polymerase and Sanger sequencing. Variation in the copy number of the 16S rRNA gene within species could also be an explanation. In summary, hidden behind the OTU 393 sequence, we identified 21 new putative genera, including 30 novel species, indicating that one OTU sequence can hide very diverse and distant sequences.

Finally, analysis of the eight reconstructed sequences (mean length 1317 bp) after targeting the metabarcoding OTU 1761 sequence initially assigned to the order *Clostridiales* allowed us to discover a new family (Fig. 5). Eight 16S rDNA sequences (OTU1761_1 to OTU1761_8) represented seven new genera in this new family (86.5–95.05% identity between sequences). OTU1761_4 and OTU1761_7 represented two distinct species in the same genus. The eight sequences showed similarities (92.26–96.25%) with sequences from GenBank annotated as 'uncultured' bacteria from the human gut. OTU1761_1 showed the closest proximity (96.25%) to 'uncultured' bacteria (DQ793257) from the human gut. Using the PCR approach with primers designed from our reconstructed sequences, we amplified two DNA fragments from mixed initial metagenomic samples. Sanger sequencing (OTU1761_PCR1 and OTU1762_PCR2; Data S7) of these amplicons showed 97% identity with OTU1761_1 and 97.5% identity with OTU_1761_8, demonstrating that these species were actually present in the faecal microbiome. The most proximal GenBank-assigned sequence [*Eubacterium*] *siraeum* (LC515595) showed 94% identity with the OTU1761_1 and OTU1761_PCR1 sequences. A draft genome (FP929059) similarly assigned by the MetaHIT consortium showed the same proximities. These sequences could be bacterial representatives of a newly discovered family. Another small contig (667 bp) from the human gut metagenome (QRFC01107314) showed very high identity (99%) with our PCR product. To conclude, hidden behind the OTU 1761 sequence, we identified a new family from the order *Clostridiales*, including seven new genera and eight novel species.

Inter-individual microbial diversity

From three OTU sequences, we eventually described a high diversity of bacterial sequences with 59 nearly full-length 16S rRNA genes. We observed a high inter-individual distribution of these sequences (Data S8). The 15 identified sequences linked to OTU 146 were not all present in the five explored human faecal samples, with zero (in M3) to six (in M1) of the sequences observed in individual samples. The abundances of OTU 146 sequences after gene capture varied between 0.0094 and 1.48%. As these OTU 146 microorganisms seem to be part of the rare biosphere, caution must be taken in interpreting these results. Indeed, we cannot exclude the presence of different species in each sample at such low levels that they would not be detected by our approach, even though it is very sensitive. We observed similar results with OTU 393, for which 36 nearly full-length sequences were revealed. The M1 sample harboured the highest bacterial diversity of this OTU, with 13 sequences, while the M2 and M3 samples showed the lowest diversity, with only three sequences. The abundances of OTU 393-related sequences after gene capture varied greatly, from 0 (M1–M5) to 15% (in the case of OTU393_27 in M5). OTU393_24 to _29 were considered to belong to the same species due to the high similarity of these six sequences. This species appeared to be dominant in all the samples (M1 OTU393_24, 8.75%; M2 OTU393_26, 1.73%; M3 OTU393_29, 2.02%; M4 OTU393_25 6.9%; M5 OTU393_27 and _28, 15.07% and 1.54%, respectively). Finally, the M4 and M5 samples showed no sequences linked to OTU 1761. In contrast, the M3 sample harboured five sequences among the eight identified sequences. The abundances of these 1761 OTU-related sequences varied from 0.008 to 0.69% after gene capture. Overall, the DNA samples originating from five individuals did not share the same OTU-reconstructed sequences. The sequence pattern was specific for each sample and confirmed the important inter-individual microbial diversity of the targeted microorganisms, even within one species.

Microbial genome mining for the discovery of genes encoding polyphenol- and polysaccharide-degrading enzymes

Some of the reconstructed 16S rDNA sequences showed proximity to bacteria whose genomes have been sequenced and annotated. We explored these genomes to identify genes related to polyphenol or polysaccharide degradation. The genome of the *Lachnospiraceae* bacterium OM04-12BH (QULQ01000007.1, isolated from Chinese human faeces), which was the closest to OTU 146, appeared well adapted to the gut environment. We identified 22 genes encoding glycosyl hydrolases (GHs) belonging to GH families 2, 3, 5, 13, 25, 31, 32, 125 and 127 participating in polysaccharide degradation and two genes encoding a butyrate kinase (RHV49074.1) and an acetate kinase (RHV52498.1) involved in the production of short-chain fatty acids (SCFAs), terminal products of anaerobic fermentation. In addition, from this genome, we selected three genes encoding enzymes annotated as NAD(P)-dependent oxidoreductase (RHV45720.1, RHV51714.1, RHV48663.1), showing low identity (nearly 30%) with dihydrodaidzein reductase, which is involved in polyphenol bioconversion [35]. RHV45720.1 also showed a very high similarity (99%) with an enzyme annotated as bile acid 7-dehydroxylase (SCH75146) from an 'uncultured *Clostridium* sp.' (FMEV01000006.1) isolated from human faeces. RHV48663.1 was 99% identical to an enzyme

annotated as '3-oxoacyl-[acyl-carrier-protein] reductase FabG' (SCI13795) from an uncultured *Clostridium* sp. (BioSample: SAMEA3545292) from human faeces. The genome also harboured eight genes encoding FAD oxidoreductase, three genes encoding β-glucosidase and three genes encoding a glycosidase not annotated as GH, enzymes that can also participate in polyphenol and carbohydrate degradation. Mining for the other genomes that showed proximity to our 16S rDNA reconstructed sequences (see above) did not allow us to identify other enzymes potentially involved in polyphenol or polysaccharide degradation.

DISCUSSION

The human gut microbiome has been implicated in important phenotypes related to human health and disease [36, 37]. Our understanding of the microbial communities that inhabit the human body and other environments has greatly improved due to sequencing and computational advances in metagenomic exploration [38, 39]. Studies have massively expanded the known species repertoire of the body-wide human microbiome, making unprecedented numbers of new cultured and uncultured genomes available [40, 41]. Recently, the Unified Human Gastrointestinal Genome (UHGG) collection identified 204 938 non-redundant genomes encoding more than 170 million protein sequences from 4644 gut prokaryotes [42]. However, incomplete reference data that lack sufficient microbial diversity hamper our understanding of the roles of individual microbiome species as well as their functions and interactions. Low-abundance taxa, which are usually missed by sequencing techniques due to the difficulty in accessing their genetic material, could play important roles in the functioning of ecosystems. New microbial species that are part of the set of unknown microorganisms referred to as 'microbial dark matter' remain inaccessible [43]. Targeting 16S rDNA variable regions with short-read sequencing platforms is largely used to reveal microbial diversity but cannot achieve the taxonomic resolution afforded by sequencing the entire gene [4]. Nonetheless, amplicon sequencing is an easier and lower-cost way to detect rare species in complex communities than shotgun sequencing methods [44].

To make progress in uncovering hidden microbiome diversity at the species level, we developed an efficient capture-based hybridization method targeting OTU sequences, allowing us to reconstruct full-length 16S rRNA genes. Amplicon sequence variants (ASVs) could also be used to improve specific probe design. By targeting three OTU sequences (16S rDNA V3–V5 region) previously identified as being potentially involved in polyphenol and/or polysaccharide degradation, we revealed the microbial diversity hidden behind these short sequences. Behind these three OTUs, we identified one new family belonging to the order *Clostridiales*, 39 new genera (seven from the new family and 32 belonging to the family *Lachnospiraceae*) and 52 novel species (44 from genera belonging to the family *Lachnospiraceae* and eight from genera belonging to the new family). The family *Lachnospiraceae* (comprising 58 genera and several unclassified strains) is a phylogenetically and morphologically heterogeneous taxon belonging to clostridial cluster XIVa of the phylum *Firmicutes* [45]. Our results confirm this important diversity and suggest the presence of a large fraction of still-unexplored diversity within this phylum. The current estimations that indicate that humans have several hundred microbial species in their gut are likely to be underestimates. At the strain level, the gap between known and true diversity must be much higher. We still need to continue exploring microbial diversity, including rare biospheres, despite the existing technical issues. It is also important to note that studies using sequencing approaches demonstrate that microbial composition is also highly variable among individuals. In our study, we also demonstrated such inter-individual variability, although the study was performed with only five volunteers. Each nearly full-length reconstructed 16S rDNA sequence was specific to one individual, highlighting individual strain patterns related to each OTU. Surprisingly, even in the same species (OTU393_24–29), we observed individual-specific patterns. We cannot exclude that a part of this diversity originates from artificial diversity created *in silico* during full-length sequence reconstruction, even though we validated the efficiency of this step by detecting similar sequences in the GenBank database and during PCR experiments followed by cloning and Sanger sequencing. As indicated by EMIRGE developers, occasional presence of small indel errors in the reconstructed sequence could occur but, in practice, these rare indels have little effect on taxonomic results [11]. Other tools for 16S reconstruction could also be used [46].

In this study, we characterized rare OTUs from the human gut that were previously detected by metabarcoding at abundances between 0.001 and 0.8%, confirming that our approach is highly sensitive. It has been suggested that rare taxa are not necessarily important for the comparison and analysis of microbial community profiles [47]. Since the discovery that most microbial communities comprise a large percentage of rare bacterial taxa, also called the 'rare biosphere' [2], rare taxa have frequently been shown to contribute to a variety of ecosystemic functions. However, most frequently, studies on the human gut microbiota largely focus on the dominant bacterial phyla *Firmicutes* and *Bacteroidetes* and ignore the large low-abundance communities present in the human gut. These rare taxa are phylogenetically diverse and could independently or collectively participate in diverse metabolic functions important to human health. For instance, *Oxalobacter formigenes* is one of the rare taxa present in the human gut [48], and yet this bacterium has been found to be able to reduce oxalate levels and could become an important probiotic species for controlling hyperoxaluria and associated disorders [49].

The biological context of this study was also considered. Although dietary polyphenols are generally not recognized as essential components of the diet, epidemiological data suggest a positive relationship between dietary exposure to polyphenols and health [50]. The beneficial effect of polyphenols is due to their antioxidant properties, among other biological activities. Because polyphenols are poorly bioavailable and

reach the lower gut (colon) undegraded, the hypothesis that the commensal microbiota could participate in the health benefits of polyphenols has been proposed [51]. Decades ago, the same hypothesis was made for dietary fibres, and it has now been proven that microbial metabolism of polysaccharides plays an important role in human health, in part through the production of SCFAs [52]. In a recent *in vitro* study, we showed that the co-metabolism by the gut microbiota of these two complex moieties (in this case purified from apple) could generate an anti-inflammatory metabolome [23]. Unfortunately, the compositional microbial data obtained by metabarcoding did not allow us to obtain enough information on the microbial players potentially involved in the pathways leading to anti-inflammatory metabolites. In particular, we found OTUs that were significantly enriched and for which we had no definitive assignment that would have allowed us to further investigate the potential activities of these OTUs. Consequently, in addition to precisely assigning the microorganisms corresponding to the OTUs of interest, we sought to obtain information on their potential enzyme activity in relation to polyphenol or polysaccharide metabolism.

We identified microorganisms that could be closely related to the reconstructed 16S rDNA sequences of OTU 146, and one of them had a sequenced genome. By genome-driven analysis, we identified some interesting potential metabolic capacities that may be related to the metabolization of polyphenols and polysaccharides. For instance, we found ten genome-encoded proteins that were automatically annotated as NAD- or FAD-dependent oxidoreductases. Thus far, very few enzymes involved in polyphenol bioconversion have been identified. The most well-studied pathway corresponding to the bioconversion of daidzein to equol [53] involves three reductases, including dihydrodaidzein reductase, which shared low similarity with three of the genome-encoded proteins. Although this information is valuable, we cannot conclude whether OTU 146-related microorganisms actually harbour metabolic activities against polyphenols without isolating the microorganisms and assaying their enzyme activity. We also showed that the genome closest to OTU 146 harboured at least 22 genes encoding glycoside hydrolases, which generally play a major role in polysaccharide hydrolysis and carbohydrate metabolism in the human gut [54], such as family 5, 13 and 32 glycoside hydrolases, which are active against cellulose, starch and fructans, respectively. Furthermore, the presence of genes encoding butyrate kinases suggests that this bacterium produces butyrate as an end product of carbohydrate fermentation. In summary, this genome-driven analysis of the bacterium closest to OTU 146, via reconstructed sequences, showed that it might metabolize apple polyphenols and polysaccharides and produce butyrate, a well-known anti-inflammatory metabolite produced by the intestinal microbiota [55]. Ultimately, all the information obtained in this study will be extremely valuable for isolation and identification of these OTUs, for example by using culturomic approaches coupled to 16S rDNA sequencing and/or metabolic screening.

In conclusion, the relationship between dietary polyphenols and the intestinal microbiota remains unclear and unexplored. Although the relationship between dietary polysaccharides and the gut microbiota is much better documented, there are key microorganisms and metabolic pathways that have not yet been discovered. The intestinal microbiota is quite diverse among individuals. Using our innovative approach for species identification, we detected candidate microorganisms that could act as direct or indirect players in such complex metabolic pathways. Characterizing such microorganisms using culturomics will be helpful in the elucidation of polyphenol microbial degradation and its role in health. The rare biosphere could play a determinant role in such metabolic pathways. Finally, our strategy revealed important hidden microbial diversity behind OTU sequences. Our results suggest that the gut microbiota could be much more diverse and have much greater inter-individual variability than previously thought.

Funding information
This research was supported by INRAE Metaprogramme MEM.

Acknowledgements
We are grateful to the Mésocentre Clermont Auvergne University and AuBi platform for providing computing resources.

Author contributions
C.G., P.M. and P.P. designed the study. S.M., C.G. and P.P. performed sequencing data analyses. C.G., C.F. and Y.L. performed biological experiments. S.M. and P.P. wrote the manuscript, and all authors contributed to manuscript revisions.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Ethical statement
This was a non-interventional study with no additions to usual clinical care. According to French Health Public Law (CSP Art L 1121-1.1), such a protocol does not require approval of an ethics committee. Five healthy donors (two males and three females; M1–M4 from a previous study [23] and M5 from this study) were informed of the study aims and procedures and provided written consent for their faecal matter to be used for the experiments.

References

1. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* 1998;95:6578–6583.
2. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, *et al*. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A* 2006;103:12115–12120.
3. Fuks G, Elgart M, Amir A, Zeisel A, Turnbaugh PJ, *et al*. Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome* 2018;6:17.
4. Johnson JS, Spakowicz DJ, Hong B-Y, Petersen LM, Demkowicz P, *et al*. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun* 2019;10:5029.
5. Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR, *et al*. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Res* 2010;38:22.
6. Schloss PD. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput Biol* 2010;6:e1000844.
7. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, *et al*. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* 2014;12:635–645.

8. Singer E, Bushnell B, Coleman-Derr D, Bowman B, Bowers RM, *et al*. High-resolution phylogenetic microbial community profiling. *ISME J* 2016;10:2020–2032.
9. Zhang Y, Ji P, Wang J, Zhao F. RiboFR-Seq: a novel approach to linking 16S rRNA amplicon profiles to metagenomes. *Nucleic Acids Res* 2016;44:e99.
10. Burke CM, Darling AE. A method for high precision sequencing of near full-length 16S rRNA genes on an Illumina MiSeq. *PeerJ* 2016;4:e2492.
11. Miller CS, Baker BJ, Thomas BC, Singer SW, Banfield JF. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol* 2011;12:R44.
12. Yuan C, Lei J, Cole J, Sun Y. Reconstructing 16S rRNA genes in metagenomic data. *Bioinformatics* 2015;31:i35-43.
13. Jousset A, Bienhold C, Chatzinotas A, Gallien L, Gobet A, *et al*. Where less may be more: how the rare biosphere pulls ecosystems strings. *ISME J* 2017;11:853–862.
14. Earl JP, Adappa ND, Krol J, Bhat AS, Balashov S, *et al*. Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes. *Microbiome* 2018;6:190.
15. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, *et al*. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Res* 2019;47:18.
16. Whon TW, Chung W-H, Lim MY, Song E-J, Kim PS, *et al*. The effects of sequencing platforms on phylogenetic resolution in 16 S rRNA gene profiling of human feces. *Sci Data* 2018;5:180068.
17. Gasc C, Peyretaillade E, Peyret P. Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. *Nucleic Acids Res* 2016;44:4504–4518.
18. Denonfoux J, Parisot N, Dugat-Bony E, Biderre-Petit C, Boucher D, *et al*. Gene capture coupled to high-throughput sequencing as a strategy for targeted metagenome exploration. *DNA Res* 2013;20:185–196.
19. Ranchou-Peyruse M, Gasc C, Guignard M, Aüllo T, Dequidt D, *et al*. The sequence capture by hybridization: a new approach for revealing the potential of mono-aromatic hydrocarbons bioattenuation in a deep oligotrophic aquifer. *Microb Biotechnol* 2017;10:469–479.
20. Gasc C, Peyret P. Revealing large metagenomic regions through long DNA fragment hybridization capture. *Microbiome* 2017;5:33.
21. Gasc C, Peyret P. Hybridization capture reveals microbial diversity missed using current profiling methods. *Microbiome* 2018;6:61.
22. Rassoulian Barrett S, Hoffman NG, Rosenthal C, Bryan A, Marshall DA, *et al*. Sensitive Identification of Bacterial DNA in Clinical Specimens by Broad-Range 16S rRNA Gene Enrichment. *J Clin Microbiol* 2020;58:12.
23. Le Bourvellec C, Bagano Vilas Boas P, Lepercq P, Comtet-Marre S, Auffret P, *et al*. Procyanidin-cell wall interactions within apple matrices decrease the metabolization of procyanidins by the human gut microbiota and the anti-inflammatory effect of the resulting microbial metabolome in vitro. *Nutrients* 2019;11:E664.
24. Parisot N, Denonfoux J, Dugat-Bony E, Peyret P, Peyretaillade E. KASpOD--a web service for highly specific and explorative oligonucleotide design. *Bioinformatics* 2012;28:3161–3162.
25. Ribière C, Beugnot R, Parisot N, Gasc C, Defois C, *et al*. Targeted gene capture by hybridization to illuminate ecosystem functioning. *Methods Mol Biol* 2016;1399:167–182.
26. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011;27:863–864.
27. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 2007;73:5261–5267.
28. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, *et al*. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013;41:D590-6.
29. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, *et al*. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* 2008;36:W465-9.
30. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004;5:113.
31. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000;17:540–552.
32. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 2003;52:696–704.
33. Chevenet F, Brun C, Bañuls A-L, Jacq B, Christen R. TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics* 2006;7:439.
34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
35. Stevens JF, Maier CS. The chemistry of gut microbial metabolism of polyphenols. *Phytochem Rev* 2016;15:425–444.
36. Qin J, Li Y, Cai Z, Li S, Zhu J, *et al*. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 2012;490:55–60.
37. Lynch SV, Pedersen O. The human intestinal microbiome in health and disease. *N Engl J Med* 2016;375:2369–2379.
38. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* 2017;35:833–844.
39. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, *et al*. A genomic catalog of Earth's microbiomes. *Nat Biotechnol* 2021;39:499–509.
40. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, *et al*. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* 2019;176:649–662.
41. Zou Y, Xue W, Luo G, Deng Z, Qin P, *et al*. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat Biotechnol* 2019;37:179–185.
42. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, *et al*. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* 2021;39:105–114.
43. Thomas AM, Segata N. Multiple levels of the unknown in microbiome research. *BMC Biol* 2019;17:48.
44. Tessler M, Neumann JS, Afshinnekoo E, Pineda M, Hersch R, *et al*. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Sci Rep* 2017;7:6589.
45. Vacca M, Celano G, Calabrese FM, Portincasa P, Gobbetti M, *et al*. The controversial role of human gut lachnospiraceae. *Microorganisms* 2020;8:E573.
46. Gruber-Vodicka HR, Seah BKB, Pruesse E. phyloFlash: Rapid small-subunit rrna profiling and targeted assembly from metagenomesmall-Subunit rRNA Profiling and Targeted Assembly from Metagenomes. *mSystems* 2020;5.
47. Gobet A, Quince C, Ramette A. Multivariate cutoff level analysis (MultiCoLA) of large community data sets. *Nucleic Acids Res* 2010;38:e155.
48. Barnett C, Nazzal L, Goldfarb DS, Blaser MJ. The presence of oxalobacter formigenes in the microbiome of healthy young adults. *J Urol* 2016;195:499–506.
49. Jairath A, Parekh N, Otano N, Mishra S, Ganpule A, *et al*. Oxalobacter formigenes: Opening the door to probiotic therapy for the treatment of hyperoxaluria. *Scand J Urol* 2015;49:334–337.
50. Scalbert A, Manach C, Morand C, Rémésy C, Jiménez L. Dietary polyphenols and the prevention of diseases. *Crit Rev Food Sci Nutr* 2005;45:287–306.
51. Fraga CG , Croft KD , Kennedy DO , Tomás-Barberán FA . The effects of polyphenols and other bioactives on human health. *Food Funct* 2019;10:514–528.

52. Koh A, De Vadder F, Kovatcheva-Datchary P, Bäckhed F. From dietary fiber to host physiology: short-chain fatty acids as key bacterial metabolites. *Cell* 2016;165:1332–1345.
53. Mayo B, Vázquez L, Flórez AB. Equol: A bacterial metabolite from the daidzein isoflavone and its presumed beneficial health effects. *Nutrients* 2019;11:2231.
54. El Kaoutari A, Armougom F, Raoult D, Henrissat B. [Gut microbiota and digestion of polysaccharides]. *Med Sci* 2014;30:259–265.
55. Liu H, Wang J, He T, Becker S, Zhang G, *et al*. Butyrate: A double-edged sword for health? *Adv Nutr* 2018;9:21–29.