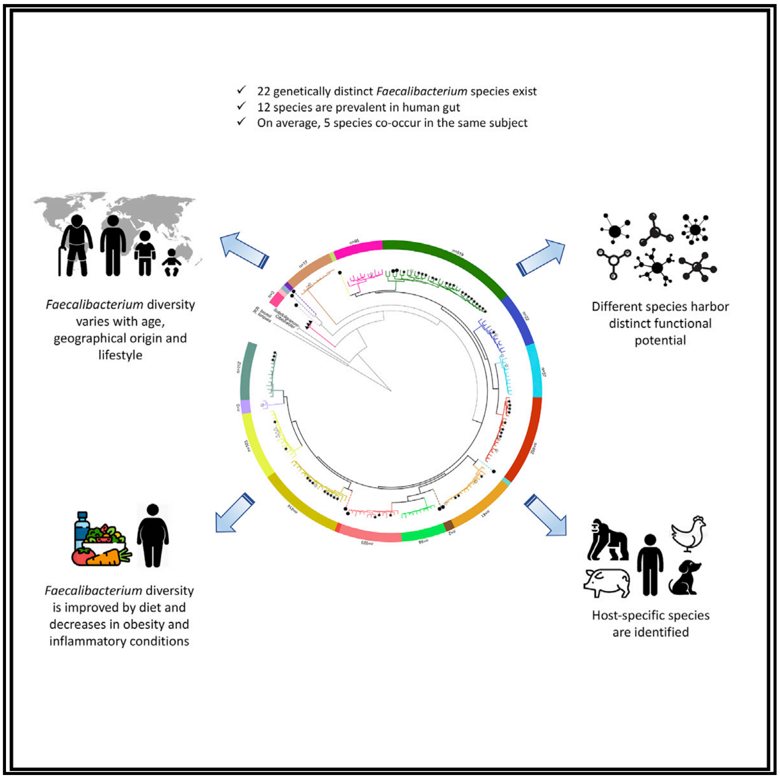
Current Biology

Article

Newly Explored *Faecalibacterium* Diversity Is Connected to Age, Lifestyle, Geography, and Disease

Graphical Abstract


✓ 22 genetically distinct *Faecalibacterium* species exist
✓ 12 species are prevalent in human gut
✓ On average, 5 species co-occur in the same subject
Faecalibacterium diversity varies with age, geographical origin and lifestyle
Different species harbor distinct functional potential
Faecalibacterium diversity is improved by diet and decreases in obesity and inflammatory conditions
Host-specific species are identified

Authors

Francesca De Filippis, Edoardo Pasolli, Danilo Ercolini

Correspondence

ercolini@unina.it

In Brief

Faecalibacterium is one of the most promising taxa for the development of novel probiotics. De Filippis et al. explore *Faecalibacterium* diversity in human and animal guts, highlighting the presence of 22 novel species with different functional potential and diverse associations with diet, health, and disease.

Highlights

- *Faecalibacterium* species and strain diversity in human and animal gut were explored
- Twenty-two new *Faecalibacterium*-like species were identified
- *Faecalibacterium* diversity varies according to age, origin, lifestyle, and disease
- *Faecalibacterium* species show distinct functional potential

De Filippis et al., 2020, Current Biology *30*, 1–12
December 21, 2020 © 2020 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.cub.2020.09.063

CellPress

Please cite this article in press as: De Filippis et al., Newly Explored *Faecalibacterium* Diversity Is Connected to Age, Lifestyle, Geography, and Disease, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.09.063

Current Biology

CellPress
OPEN ACCESS

Article

Newly Explored *Faecalibacterium* Diversity Is Connected to Age, Lifestyle, Geography, and Disease

Francesca De Filippis,[1,2] Edoardo Pasolli,[1,2] and Danilo Ercolini[1,2,3,*]
[1]Department of Agricultural Sciences, University of Naples Federico II, Portici 80055, Italy
[2]Task Force on Microbiome Studies, University of Naples Federico II, Naples 80100, Italy
[3]Lead Contact
*Correspondence: ercolini@unina.it

https://doi.org/10.1016/j.cub.2020.09.063

SUMMARY

Faecalibacterium is prevalent in the human gut and a promising microbe for the development of next-generation probiotics (NGPs) or biotherapeutics. Analyzing reference *Faecalibacterium* genomes and almost 3,000 *Faecalibacterium*-like metagenome-assembled genomes (MAGs) reconstructed from 7,907 human and 203 non-human primate gut metagenomes, we identified the presence of 22 different *Faecalibacterium*-like species-level genome bins (SGBs), some further divided in different strains according to the subject geographical origin. Twelve SGBs are globally spread in the human gut and show different genomic potential in the utilization of complex polysaccharides, suggesting that higher SGB diversity may be related with increased utilization of plant-based foods. Moreover, up to 11 different species may co-occur in the same subject, with lower diversity in Western populations, as well as intestinal inflammatory states and obesity. The newly explored *Faecalibacterium* diversity will be able to support the choice of strains suitable as NGPs, guided by the consideration of the differences existing in their functional potential.

INTRODUCTION

Faecalibacterium is a member of the *Clostridium leptum* group [1] and represents one of the most prevalent species in the gut microbiome of healthy human adults [2–4]. Currently, *F. prausnitzii* is the only known species of the genus *Faecalibacterium*.

F. prausnitzii has received increasing interest in recent years as a biomarker of gut health [5]. Indeed, a decreased abundance of this genus was found in association with inflammatory bowel syndrome (IBS) and inflammatory bowel disease (IBD), as well as with colorectal cancer and diabetes and in frail elderly people [6, 7]. Accordingly, some studies reported the ability of *F. prausnitzii* strains to produce anti-inflammatory metabolites, including butyrate [8], peptides [9], and an extracellular polymeric matrix [10], showing anti-inflammatory activity both *in vitro* and in animal models [11]. For these reasons, *F. prausnitzii* currently represents one of the most promising taxa for the development of next-generation probiotics (NGPs) [11–13].

Faecalibacterium is sensitive to oxygen and difficult to cultivate; therefore, most knowledge about its prevalence and ecology is based on the analysis of metagenomic data [12]. Previous studies suggested the existence of different species and still-unexplored diversity at the subspecies level. Lopez-Siles et al. [14] defined two phylogroups with <97% identity in the 16S rRNA gene sequence among 17 *Faecalibacterium* isolates. It was also suggested that other, less common *F. prausnitzii* phylotypes exist and that prevalence and abundance of *F. prausnitzii* phylotypes differ depending on gut inflammatory conditions [15, 16]. In addition, genomic comparison confirmed the presence of at least two different genomospecies (species that can be differentiated using genomic analysis) among 31 sequenced *F. prausnitzii* genomes from isolates [17]. However, *Faecalibacterium* genomic diversity assessed on the basis of genomes from cultured isolates may not be completely representative of the real diversity [18]. The large number of accessible human metagenomes represents an invaluable source of biological information for a more in-depth exploration of *F. prausnitzii* diversity, ecology, and potential impact. In addition, it is not clear to what extent *F. prausnitzii* genomic diversity at the species or sub-species level may lead to differences in the phenotypes and to the production of a specific pattern of metabolites [7]. Indeed, the production of beneficial or detrimental molecules that are important to define the role of gut microbes may differ between species/strains and have different influences on human health [19–21].

Therefore, an evaluation of *Faecalibacterium* diversity in available gut metagenomes is crucial for defining the diversity existing within this genus and for investigating whether different species/strains and possible specific metabolic functions are equally distributed among populations, according to geographical origin, lifestyle, age, or health status.

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

RESULTS

Faecalibacterium Abundance across Populations Is Highly Variable and Depends on Age, Geographical Origin, and Lifestyle

We performed a large-scale meta-analysis (n = 7,907 human gut metagenomes; STAR Methods) and found that *F. prausnitzii* was highly prevalent and abundant across the global population [2–4] (as detected by MetaphlAn2 [Metagenomic Phylogenetic Analysis], which calculates taxa abundance mapping all non-human reads against species-specific marker genes). This taxon was detected by MetaPhlAn2 in 85% of the gut samples (never detected in additional 1,538 human non-gut metagenomes) with an abundance up to 75% (mean: 6.5% ± 7.6%; median: 4.8%), whose average value was the highest among the species detected in the gut [22]. However, we observed marked differences according to age, geographical origin, and lifestyle (Figure S1). It was almost always identified in the adult population (94% of prevalence; 19–65 years old), but the prevalence decreased to 82% and 18% in children (1–12 years old) and newborns (<1 year old), respectively (Fisher's test; $p < 0.01$). Lower prevalence was also observed in seniors (83% of prevalence in subjects >65 years old; Fisher's test; $p < 0.01$). Such a pattern was confirmed in terms of abundance, and we found *Faecalibacterium* abundance ranging from an average of 0.96% in newborns to 7.6% in adults (Figure S1A). Thus, *F. prausnitzii* increased with age from newborns to adults and decreased again at later ages. Marked variation was also observed in terms of lifestyle. Non-Western (i.e., populations from non-industrialized countries) showed a higher prevalence than Western populations (99.6% versus 93.4% in the adult population, respectively; Fisher's test; $p < 0.05$), a difference that was much more evident in terms of abundance (median: 9.1% versus 5.5%; Wilcoxon test; $p < 2e{-}16$; Figure S1B). Abundance was highly variable according to the country of origin, with the highest levels detected for non-Western countries (e.g., Tanzania and Perù; Figure S1C). Within Western populations (China [CHN], Europe [EU], and North America [NAM] were the three main regions), no differences were observed in terms of prevalence, although a progressively lower abundance was observed from CHN to NAM through EU (Figure S1D).

Genome-wide Analysis of *Faecalibacterium* Reveals 22 *Faecalibacterium*-like Clades and High Sub-species Diversity

We analyzed 2,859 quality-controlled (completeness >80% and contamination <5%; STAR Methods) *Faecalibacterium* genomes that were retrieved from the NCBI (n = 64 after quality filtering and removal of mislabeled genomes and including both genomes from isolates and metagenome-assembled genomes [MAGs]) and from publicly available resources of MAGs from human (n = 2,784) [23] and NHPs (non-human primates) (n = 11) [24] gut metagenomes (Data S1A and S1B). Applying our validated pipeline to delineate SGBs at 5% genomic distance [23, 24] (average nucleotide distance [ANI]; see STAR Methods), we defined 21 *Faecalibacterium*-like clades (Figure 1), in addition to one NCBI genome labeled *Faecalibacterium* sp., which we identified with the recently proposed candidate *Cibiobacter* sp. [23, 25] (clade U). and considered here as an outgroup (Figure 1A). We identified twelve clades (clade A with 923 genomes through clade L with 17 genomes, sorted in descending number of genomes) that were prevalent (>10 reconstructed genomes) in the human gut and that we define here as the human *Faecalibacterium* complex, following a nomenclature recently suggested for *Prevotella copri* [26]. We found an additional rare human clade (clade N; n = 3) and multiple host-specific clades with genomes from NHPs or other animal guts (Figures 1A–1D; Table S1; Data S1C).

In the human *Faecalibacterium* complex, we identified eleven clades (clade A through clade K) as *F. prausnitzii* because containing NCBI genomes were labeled as *F. prausnitzii*. Interestingly, reference genomes retrieved from the NCBI fell in many of these clades (Figures 1A–1D; Data S1A), suggesting that *F. prausnitzii* genomes can indeed belong to different species. We observed bias between the availability of reference genomes across the different clades and their prevalence in the human gut. Half of the available genomes (47%) belonged to clades B and D, although the more prevalent clade A was associated with only three NCBI reference genomes. Moreover, no genomes from isolates were available for three clades (clades G, J, and L), which highlighted the need to combine genomes from both isolate and metagenomic sources to limit possible culturing biases. The twelfth clade L was genetically (Figure 1C; Data S2) and phylogenetically (Figure 1A) distant from the others and probably belonged to a different *Faecalibacterium* species or to an unidentified genus. The twelve clades exhibited a median intra-clade distance ranging from 1.6% (clade L) to 3.6% (clade B), always below 5%, which is usually considered the cutoff for defining different species (Figure 1C; Data S2) [27]. Conversely, inter-clade distances were quite high, with median values ranging from 13.2% to 18.2%, with clade L being the most distant clade in agreement with its phylogenetic placement (Figure 1C; Data S2). Distances between clades were also confirmed considering the gene content (Figure 1D).

Comparing the species diversity found in humans with that found in NHPs is of great interest for exploring human evolution. Although only 11 *Faecalibacterium* MAGs were reconstructed from NHPs [24], they exhibited quite high diversity and fell into nine distinct clades (Figures 1A and 1B). MAGs from wild apes [28, 29] clustered separately from human clades. Interestingly, we found four clades that still fell in proximity to human *F. prausnitzii* clades, therefore likely representing NHP-specific *F. prausnitzii* species not found in human gut. On the other hand, MAGs from macaques in captivity [30] clustered together with human MAGs, suggesting the possible sharing of strains between NHPs and humans, either when living in close association or because of partially overlapping dietary patterns. Finally, clade M included only three genomes from chicken gut isolates and may be considered as a different genus (average genomic distance >18% from all the other clades) common in the chicken gut (see below).

We further explored sub-species diversity by estimating the number of sub-clusters in each of the twelve clades representing the human *Faecalibacterium* complex (STAR Methods). This was achieved by determining the number of clusters that were supported by a prediction strength >0.8 and an average Jaccard similarity >0.85 (Table S2; STAR Methods); at the same time, this was confirmed by genetic and phylogenetic analyses (Figure 2; Data S3). At least seven out of 12 clades clearly separated into different

Please cite this article in press as: De Filippis et al., Newly Explored *Faecalibacterium* Diversity Is Connected to Age, Lifestyle, Geography, and Disease, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.09.063

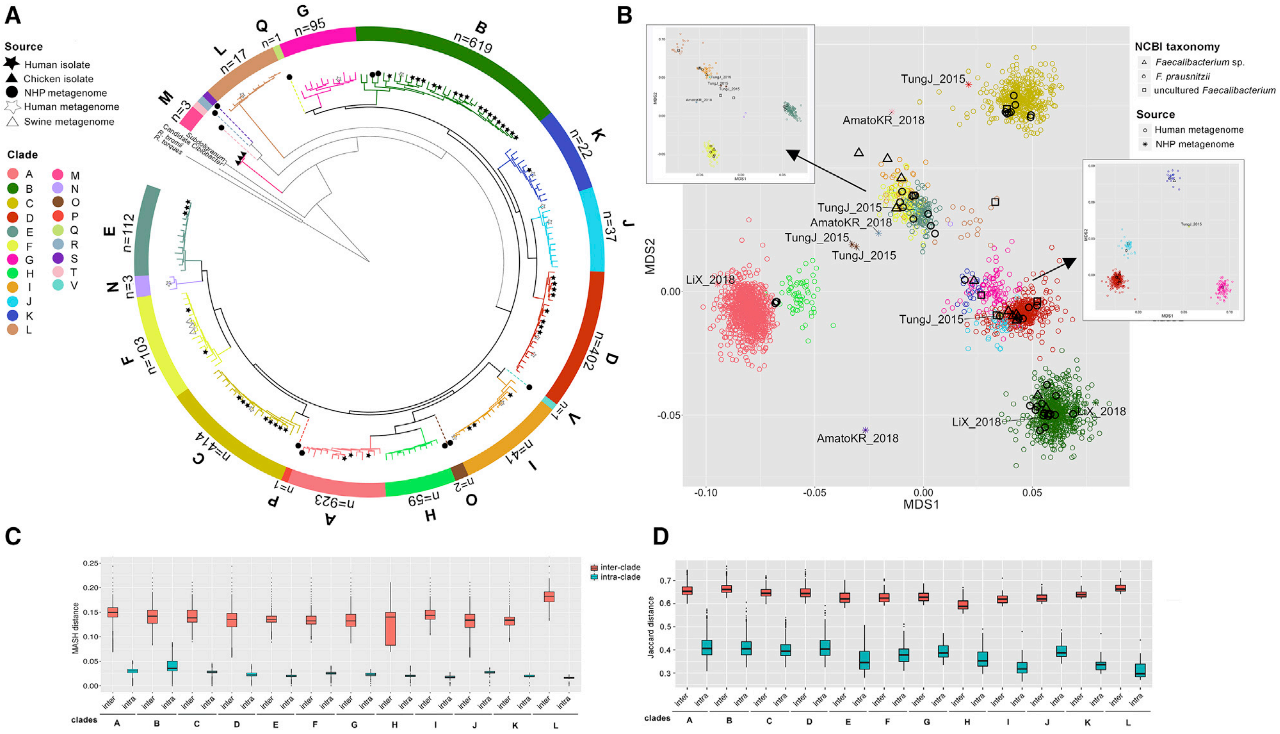


Figure 1. Twenty-Two *Faecalibacterium*-like Clades Can Be Identified

(A) Phylogenetic tree of a subset of the 2,795 human/NHP MAGs and the 64 *Faecalibacterium* NCBI genomes, showing separation in 22 clades. For each clade, ten MAGs and all the NCBI genomes were included. The clade name and the number of MAGs present in each clade are indicated on the outer ring, which is colored according to the clade. Clade U contained one NCBI reference genome incorrectly labeled as *Faecalibacterium* sp., which we identified as the recently proposed candidate *Cibiobacter* sp. Dashed lines indicate NHP-specific SGBs.

(B) Classical multidimensional scaling (MDS) carried out on MinHash distance estimation (MASH) genomic distance matrix of all genomes included in the 22 clades (small boxes show focus on some clades). Clade U contained one NCBI reference genome incorrectly labeled as *Faecalibacterium* sp., which we identified as the recently proposed candidate *Cibiobacter* sp. Labels indicate MAGs from NHPs.

(C and D) MASH genetic distance (C) and Jaccard distance based on pairwise gene content (D) within a clade (intra-clade) or between clades (inter-clade) for the 12 major *Faecalibacterium* clades identified.

See also Figure S1, Table S1, and Data S2.

sub-species, with a maximum of four sub-clusters for clade A and clade B. The strongest associations between sub-species delineation and phenotypic traits were found for geographical origin. We identified five clades (i.e., clades A, B, C, G, and F) with a clear geographical stratification of the strains (Figure 2; Data S3); one subtype was almost exclusive to Chinese subjects, although the other subtypes prevailed in Europeans or North Americans. Reference genomes were widespread in the different sub-species identified here and clustered consistently with the geographical origin of the isolate in clades A, B, and C, which supported the overall validity of our approach (Figure 2; Data S3). A more complex scenario was identified for clade A, where one of the sub-species was particularly prevalent in non-adult subjects (Fisher's test; p < 1e−10), suggesting that adult and non-adult subjects may harbor different strains of this prevalent *F. prausnitzii* species (Figure 2C).

The Prevalence and Diversity of *Faecalibacterium* Clades Vary with Age, Geographical Origin, and Lifestyle and Are Associated with Disease

We took advantage of the newly identified diversity of *Faecalibacterium* to assess the prevalence of each clade in the global population through a mapping-based approach. We employed a pipeline (validated on real and synthetic data; see STAR Methods) able to identify clade-specific marker genes (n = 6,204 distributed among the 12 clades) and use them to map reads in a computationally efficient way.

Applying this mapping-based pipeline to the same set of 7,895 human gut metagenomes considered before (Data S1B), we detected the presence of at least one clade in 83% of the samples. This overall prevalence was mainly driven by age, increasing from 14% of newborns to 92% of adults (Fisher's test; $p < 10^{-5}$) and decreasing in the elderly population (78%; $p < 10^{-5}$). In accordance with the genome reconstruction strategy, clade A was the most prevalent clade (detected in 78% of the subjects with at least one clade), followed by clade D (73%) and clade C (68%). Although clade B ranked second in terms of the number of reconstructed human MAGs, it was detected in only 66% of the subjects. Clade L was confirmed as the rarest clade, with a prevalence of 0.9% (Figure 3A).

In addition, we observed that more prevalent clades (A, B, C, and D) often occurred together in the same subject (Table S3). In particular, when up to three clades co-occurred, A and D were the most prevalent ones, followed by B and C. In addition, clade L mostly occurred when a low number of clades (less than

Please cite this article in press as: De Filippis et al., Newly Explored *Faecalibacterium* Diversity Is Connected to Age, Lifestyle, Geography, and Disease, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.09.063

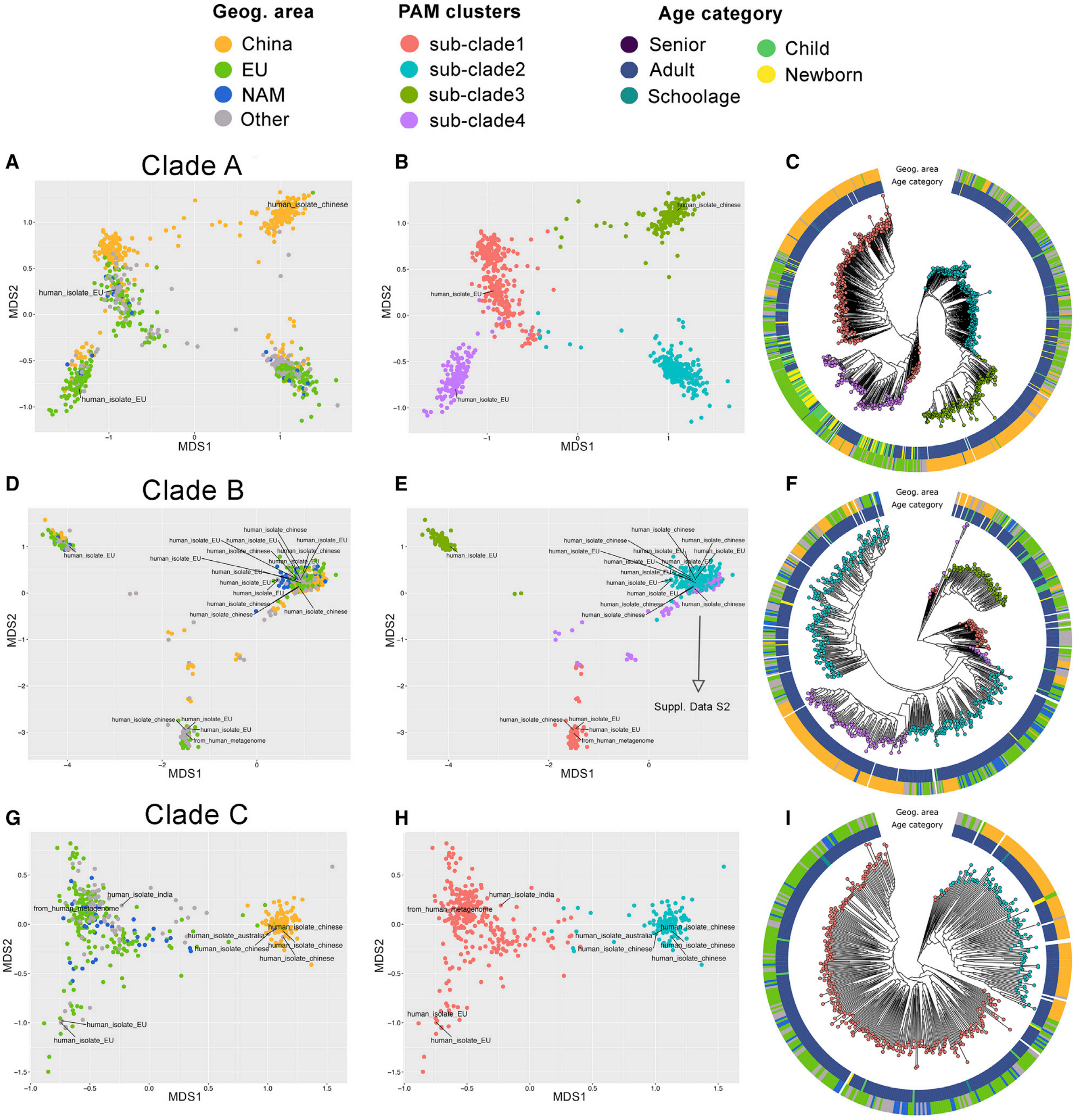


Figure 2. ***Faecalibacterium*** **Strain-Level Diversity**

Representative classical multidimensional scaling (MDS) carried out on ANI genetic distance matrix (A, B, D, E, G, and H) and phylogenetic trees based on clade-specific marker genes (C, F, and I). MAGs in MDS are colored according to the geographical origin (A, D, and G) or to the sub-cluster identified by the combined prediction strength and cluster stability criterion (B, E, and H). In phylogenetic trees, leaves are colored according to the partitioning around medoid (PAM) sub-cluster, although rings are colored by geographical origin and age category (C, F, and I). Only the three major clades A–C are reported. See also Table S2 and Data S3 for MDS and phylogenetic trees for other clades and a focus on sub-clades 3 and 4 within clade B. Geog. area: EU, Europe; NAM, North America.

three) were present and was never found in subjects with eight or more clades (Table S3).

High inter-personal variability in the clade occurrence pattern was found, with up to 11 clades co-occurring in the same subject (mean in the whole population: 4.9; Figures 3A–3C). First, clade diversity (considered here as the number of clades co-occurring in the same sample) was higher in adult age compared to school age, child, and newborn (Figures 3B and 3C). More than 80% of newborns and 60% of children showed the co-presence of only one or two clades, although this percentage decreased to 18% in school-aged children (12–18 years old) and similarly in adults (Figure 3C). Indeed, *Faecalibacterium* clade diversity was

Please cite this article in press as: De Filippis et al., Newly Explored *Faecalibacterium* Diversity Is Connected to Age, Lifestyle, Geography, and Disease, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.09.063

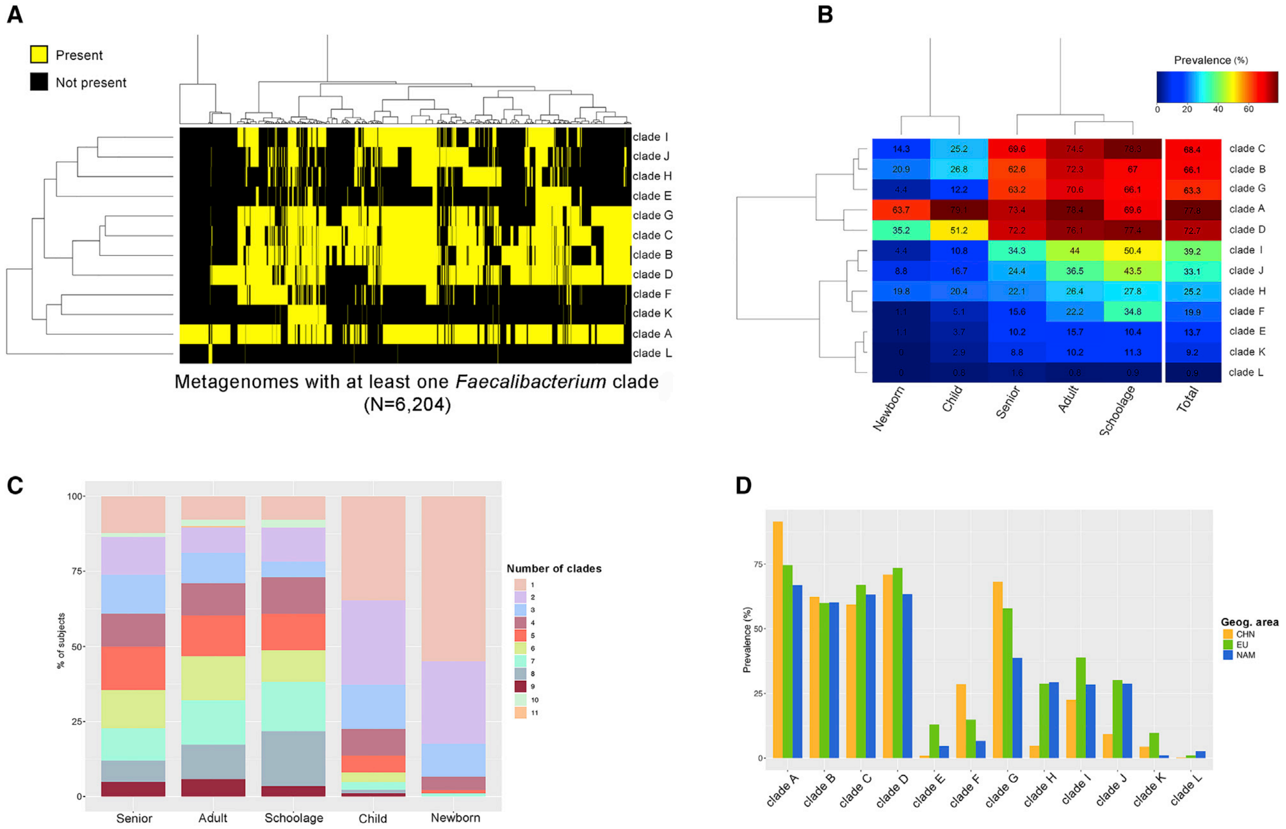


Figure 3. *Faecalibacterium* Clade Pattern in Human Gut Metagenomes

(A) Presence-absence of the 12 major *Faecalibacterium* clades based on mapping of 7,895 human gut metagenomes against 6,204 clade-specific marker genes.

(B and C) Average prevalence of the 12 major *Faecalibacterium* clades in different age groups (B) and number of different clades found in the same subject for each age group (C; senior, >65 years old; adults, 19–65 years old; school age, 13–18 years old; child, 1–12 years old; newborn, <1 year old).

(D) Prevalence of the different *Faecalibacterium* clades in subjects grouped by geographical origin (CHN, China; EU, Europe; NAM, North America).

See also Table S3 and Data S4.

positively correlated with the relative abundance in all age groups (Figures S1E and S1F), which indicated that an increase in the abundance was associated with an increase in the number of clades. Starting from school age, five main clades were the most prevalent ones (i.e., clades A, B, C, D, and G). School-aged children showed a pattern similar to that of adults (only clade F was significantly enriched in adults; Fisher's test; $p < 0.05$; Data S4C), although several clades were depleted in seniors compared to adults (Fisher's test; $p < 0.05$; Data S4D; Figure 3B). In contrast, newborns and children were mainly colonized by only two clades (i.e., clade A and clade D were the most prevalent clades). Therefore, clade A was quite prevalent during the entire lifespan, although different sub-species may occur with progression from childhood to adulthood, as previously discussed (Figure 2C).

Clade prevalence also varied according to geographical origin. Considering the three main geographical areas covering most of our samples (EU, CHN, and NAM), a higher prevalence was observed in CHN (88%), followed by EU (79%) and NAM (65%). Clades A, F, and G were more prevalent in CHN than in both EU and NAM ($p < 0.01$), although clades E, H, I, J, and L showed higher prevalence in EU and NAM (Fisher's test; $p < 0.01$; Data S4K–S4M; Figure 3D).

Additionally, we considered differences associated with lifestyle by comparing the 452 non-Western subjects from six different datasets and spanning different geographical areas with the 2,708 Western subjects (Data S1B). Non-Western populations showed a higher prevalence and diversity of *Faecalibacterium* clades. At least one clade was detected in 95% of non-Western adult subjects, a value that decreased to 82% (Fisher's test; $p < 10^{-5}$) in westernized populations. Clade diversity was also significantly affected by westernization (median: 7.3 and 5.0 in non-Western and Western subjects, respectively; $p < 2 \times 10^{-16}$; Figure 4A), and we detected the co-occurrence of >7 clades in 60% of NW versus 25% of W subjects (Figure S2). Nine clades were enriched in non-Western, although only clades A, D, and L had higher prevalence in Western subjects (Fisher's test; $p < 0.05$; Data S4N; Figure 4B). However, we cannot exclude a possible bias due to different number of available subjects in Western and non-Western cohorts.

We validated such findings by applying the same approach to ten independent cohorts (including 680 Western and 353 non-Western subjects; Data S1D) that were not employed until this stage. We confirmed a higher prevalence (93% and 88% in non-Western and Western subjects, respectively; $p < 10^{-5}$) and clade diversity in non-Western (median: 7.3 and 5.1 in

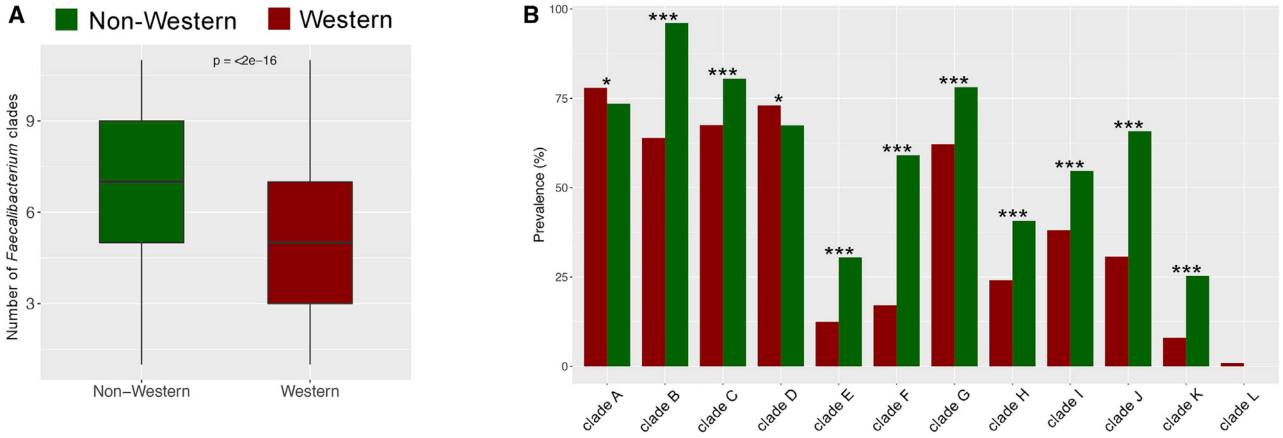


Figure 4. *Faecalibacterium* Clade Pattern Is Associated with Westernization Level

(A) Number of different clades co-occurring in the same subject in Western and non-Western populations (p values were computed using Wilcoxon's test).

(B) Prevalence of the different *Faecalibacterium* clades in subjects grouped by lifestyle. *p < 0.05; ***p < 0.0001 as defined by Fisher's test.

See also Figures S2 and S3 and Data S4.

non-Western and Western subjects, respectively; $p < 0.001$), with the same clades that were enriched in one of the two categories as previously reported (Figure S3A; Data S4O). Interestingly, a higher diversity in non-Western subjects was also observed when the subjects were stratified by age category (Figure S3B).

We finally tested whether *Faecalibacterium* diversity was associated with disease conditions. We considered the ten case-control studies available in our dataset (Data S1B) and spanning metabolic (diabetes types 1 and 2, metabolic syndrome, and impaired glucose control; three datasets) and intestinal diseases (IBD, colorectal cancer [CRC], and intestinal adenoma; seven datasets). In addition, we considered healthy adults (1,186 subjects) belonging to seven datasets for which body mass index (BMI) was available and stratified subjects into normal weight (NW) (with BMI: 18–25 kg/m^2), overweight (OW) (BMI: 25.1–30 kg/m^2), and obese (OB) (BMI: $\geq$30.1 kg/m^2) groups. No differences were found for multiple metabolic diseases, although lower *Faecalibacterium* clade diversity was specifically found for intestinal diseases and obesity (Figure 5). We further investigated whether specific clades were responsible for this relationship. Interestingly, no clades were significantly different between healthy and CRC subjects (Data S4T), although a depletion of specific clades was observed for obesity and IBD (Data S4P and S4S). More specifically, clades A, D, F, G, I, and J were enriched in normal weight compared with obese subjects (no differences were found between obese and overweight subjects; Data S4Q), although clades B, D, F, G, and K were depleted in IBD (Data S4S). Interestingly, four of the clades (B, F, G, and K) were previously found to be depleted in Western subjects in comparison with non-Western populations.

We further applied our mapping-based strategy to a cohort we recently investigated, including obese individuals undergoing a 2-month isocaloric dietary intervention with a Mediterranean diet or following a control diet [31]. Consistently with the findings obtained for Western/non-Western cohorts, we found that, probably owing to the increase in consumption of dietary fiber, increasing Mediterranean diet adherence determined a prompt increase in *Faecalibacterium* diversity already after 4 weeks (Figure 5D), although no specific clade was enriched upon the intervention.

Faecalibacterium Clades Have Distinct Functional Potentials

We considered the same set of *Faecalibacterium* genomes (n = 2,859) previously investigated to evaluate potential functional diversity across the different clades. After gene calling, we obtained a total of 5.0 M genes, which were recapitulated into 35,448 distinct UniRef annotations (STAR Methods).

We observed considerable functional diversity among clades, with separation that resembled that previously described based on genomic distance and on phylogenetic placement (Figure 6A; Data S6). Clade L appeared to be the most different from the others, although clades A/H, E/F, and B/D/G/J/K grouped according to the functional potential (Figure 6A). We found 11,160 genes that were differentially prevalent between clades (Fisher's test; $p < 0.05$; Data S5A). Different clades showed specific carbohydrate utilization patterns. For example, alpha-xylosidase (involved in xyloglucan degradation) was enriched in clades I and L, and acetylxylan esterase (responsible for xylan catabolism) was enriched in clades B, D, and J. Clade B also showed a higher prevalence of endo-polygalacturonases (degradation of pectins). In contrast, clades G, H, J, and L were depleted in cellobiose phosphorylase, which is involved in the degradation of cellulose.

Functional differences were also highlighted when comparing subjects according to their lifestyle (Figure 6B), with 109 genes that showed differential prevalence between Western and non-Western populations ($p < 0.05$; Data S5B). Genes related to complex carbohydrate degradation (such as pectinesterases, pectate-disaccharide lyase, endo-1,4-beta-xylanase, and beta-glucosidase) were enriched in non-Western subjects, although genes coding for phosphotransferase system (PTS) transport of simple sugars and resistance to antibiotics were more prevalent in Western subjects. Finally, methionine aminotransferase (involved in isothiocyanate biosynthesis

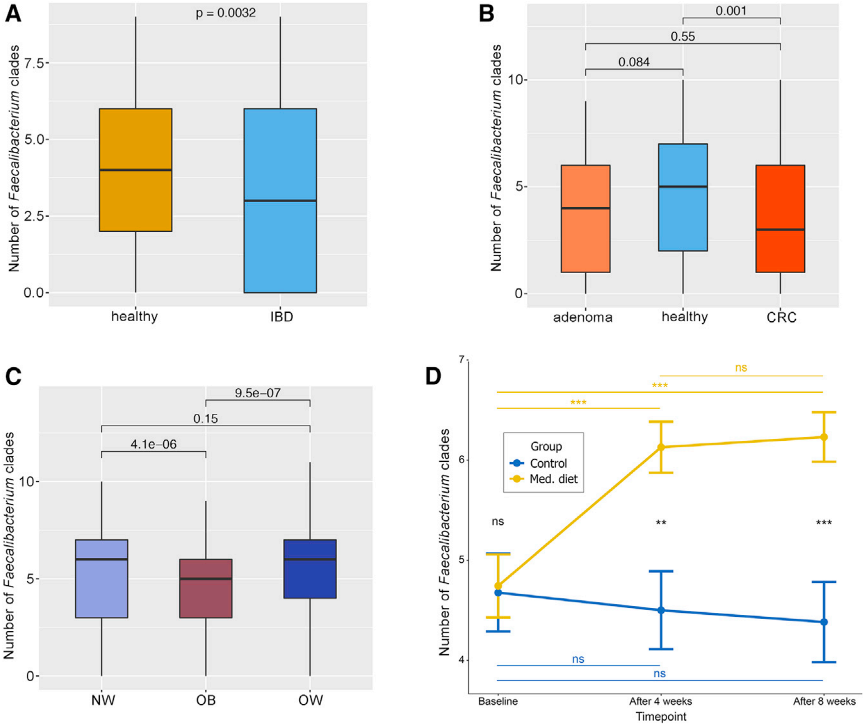


Figure 5. Loss of *Faecalibacterium* Diversity Is Associated with Inflammatory Conditions

(A–C) Number of different clades co-occurring in the same subject in healthy or diseased populations. (A) Healthy and inflammatory bowel disease (IBD); (B) healthy, colorectal cancer (CRC) and intestinal adenoma; and (C) normal weight (NW), overweight (OW), and obese (OB).

(D) Variation in *Faecalibacterium* diversity in obese subjects consuming a Mediterranean (Med. diet) or a control diet for 8 weeks. **p < 0.01; ***p < 0.001. Colored p values indicate significance within Med or control group, between time points; black p values indicate significance between groups at the same time point. Significance was computed by Wilcoxon's tests.

from glucosinolates) and 3-isopropylmalate dehydrogenases (related to branched-chain amino acids production) were depleted and enriched in Western subjects, respectively (Figure 6B; Data S5B). We also observed differences within westernized populations. Indeed, subjects from EU and NAM showed a higher prevalence of genes related to protein degradation and lactose metabolism than Chinese subjects. Moreover, subjects from EU and NAM showed the highest and lowest prevalence of genes involved in the degradation of complex polysaccharides, respectively. In contrast, subjects from CHN were enriched in genes involved in starch degradation and simple sugar transport, as well as in antibiotic resistance (Figure 6C; Data S5C–S5E).

A Remarkable *Faecalibacterium* Diversity in the Animal Gut

We finally evaluated *Faecalibacterium* genomic diversity in different animal species by applying the same genome assembly and mapping pipelines (Data S1E). We considered a total of 2,672 publicly available gut metagenomes from NHPs (n = 203; wild and living in captivity; already considered in Figure 1), dogs (n = 129), swine (n = 695), chickens (n = 1,350), and ruminants (n = 295, including cattle, sheep, and deer). Taxonomic profiles showed high variability in the occurrence and relative abundance of *Faecalibacterium* across species, as well as high inter-individual differences among animals of the same species. *Faecalibacterium* was never found in ruminants, although its average relative abundance was 2.7% ± 4.2%, 0.56% ± 0.69%, and 0.14% ± 0.56% in swine, dogs, and chickens, respectively (Figure S4A). As for NHPs, abundance and prevalence were lifestyle dependent. Indeed, it was more abundant in wild NHPs (average relative abundance of 10.8% ± 14.7%), although animals in captivity showed a lower abundance (2.8% ± 1.6%; Figure S4A).

Accordingly, when we used the mapping pipeline to detect the occurrence of the twelve clades belonging to the human *Faecalibacterium* complex, they were never found in ruminants. Clade L, rarely occurring in humans, was instead the only clade found in dogs, with a prevalence close to 100%, although chickens and wild NHPs showed a very low prevalence (Figures 7A and S4B). Conversely, a higher prevalence of human clades was obtained in NHPs living in captivity and swine (Figure 7A). We therefore suggest that the high relative abundance of *Faecalibacterium* detected in wild NHPs by MetaPhlAn2 taxonomic profiling was probably due to the presence of different non-human, host-specific *Faecalibacterium* clades.

We further explored the possibility of detecting host-specific species by assembling animal metagenomes. In addition to the 2,985 human and NHP MAGs already presented in Figure 1 [24], we generated an additional 7,602 MAGs (>80% completeness; <5% contamination) from the other animal metagenomes (STAR Methods). This resulted in a total of 153 *Faecalibacterium*-like animal MAGs (Figure 7B). In addition to NHPs (with MAGs from NHPs in captivity overlapping with human clades and MAGs from wild NHPs associated with distinct host-specific clades; Figure 1), only dogs and swine harbored human-associated clades. Indeed, 40 MAGs from dogs were associated with clade L (Figure 7B), according to its high prevalence estimated by raw-read mapping (Figure 7A). Strain-level analysis revealed two different strains, likely host-associated (Figures 7C and 7D). However, three dog MAGs overlapped with human MAGs, highlighting possible pet-human strain-transmission events (Figures 7C and 7D). We extracted 59 *Faecalibacterium*-like MAGs from swine that fell into four human clades. Fourteen MAGs were associated with clade F, and also in this case, we observed host adaptation of the strains (Figures 7E and 7F). In addition, three and two swine MAGs clustered with clades J and B, respectively (Figure 7B). Forty swine MAGs were finally associated with clade U, which contained a single reference genome incorrectly labeled *Faecalibacterium* sp. and instead belonging

Please cite this article in press as: De Filippis et al., Newly Explored *Faecalibacterium* Diversity Is Connected to Age, Lifestyle, Geography, and Disease, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.09.063

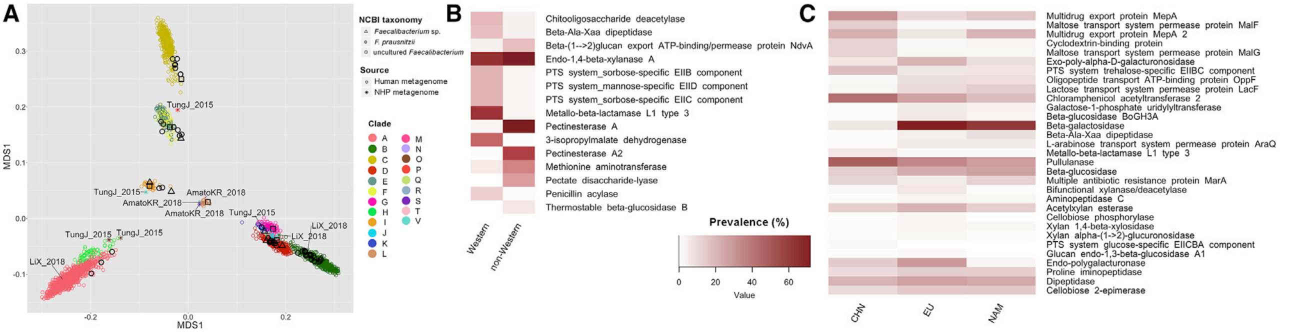


Figure 6. *Faecalibacterium* Clades Show Different Functional Potential

(A) Classical MDS carried out on Jaccard distance matrix based on the presence/absence pattern of UniRef genes.

(B and C) Prevalence of selected genes discussed in the text in different lifestyles (Western/non-Western; B) or geographical areas (C). CHN, China; EU, Europe; NAM, North America.

See also Data S5 and Data S6.

to a more distant and recently proposed candidate, *Cibiobacter* sp. [23, 25]. We finally extracted 42 MAGs from the chicken gut that were recapitulated in seven *Faecalibacterium*-like, chicken-specific SGBs that also incorporated the previously identified non-human clade M, which included only three NCBI genomes from chicken isolates (Figure 7B).

DISCUSSION

In this work, we carried out a deep investigation of the diversity within the *Faecalibacterium* genus. *Faecalibacterium* abundance was previously suggested to be associated with a diet rich in plant-based foods [31–33] and with the production of butyrate from fiber fermentation [8, 34], although its depletion was reported in association with different inflammatory conditions [35, 36]. Although *Faecalibacterium* is one of the most abundant and recurrent genera in the human gut [2–4]; to date, *F. prausnitzii* is the only species that was successfully isolated. Previous studies based on 16S rRNA analysis [15] or whole-genome comparison of a few genomes from isolates [17, 37] suggested the presence of an unexplored diversity within this genus, with the presence of at least two phylotypes within *F. prausnitzii*. However, if these phylotypes were functionally different, their spread in the human gut and their possible links with health or diseases were still unexplored. *F. prausnitzii* is regarded as one of the most promising taxa for the development of NGPs or biotherapeutics [13, 37]. In relation to this aim, understanding the diversity existing within this genus, whether different species exist and how this diversity is associated with health or disease, is of paramount importance for pushing the use of this microbe on the market. Previous work based on comparison of genomes from isolates suggested the presence of two *F. prausnitzii* phylotypes [17]. Analyzing almost 3,000 *Faecalibacterium*-like MAGs reconstructed from human and NHP gut metagenomes spanning different age groups, geographical origins and lifestyles, we highlight the presence of 22 different *Faecalibacterium*-like, species-level clades (SGBs). These results emphasize the underestimated diversity within *Faecalibacterium*, which likely includes at least two separate genera and 22 species, only partially represented by the reference genomes available from the NCBI. Indeed, we highlight that genomes from isolates currently available in NCBI may be biased toward one clade (clade B; Data S1A–S1F), although only three genomes are available for the most prevalent clade (clade A), most likely more difficult to cultivate with usual methods. Consistently, we found that all reference genomes of the phylotype 1 identified by Fitzgerald and colleagues [17] belong to clade B, although the phylotype 2 was sub-divided in at least seven clades.

Twelve of the identified *Faecalibacterium* SGBs are globally spread, although we highlight that the differential occurrence of some of them varies with age, geographical location, and lifestyle. *Faecalibacterium* SGBs also show different functional potentials. Genes involved in the metabolism of complex polysaccharides differ among SGBs, endowing each one with a specific carbohydrate utilization pattern and making clade co-occurrence and *Faecalibacterium* diversity avenues of diverse complex carbohydrate fermentation genes. This can be considered a relevant source of added value in plant food utilization potential. The observed diversity may also explain the contrasting results reported in clinical trials for the response of *F. prausnitzii* to prebiotic fiber interventions in clinical trials [33]. Therefore, we can speculate that higher *Faecalibacterium* SGB diversity may be associated with increased potential in the metabolism of complex polysaccharides and that this diversity may be promoted by a diet rich in fiber, such as that typical of non-Western populations, which were found to harbor higher diversity than Western subjects of similar age, as we previously also suggested for *Prevotella copri* [19, 26]. Consistently, we also observed that a dietary intervention with a Mediterranean diet, characterized by high intake of food products rich in fiber (fruit, vegetables, whole grains, and legumes), can boost *Faecalibacterium* diversity in obese subjects, opening new possibilities in dietary strategies for the restoration of gut microbiome diversity. Accordingly, *Faecalibacterium* SGB diversity increases with age, when more complex carbohydrates are introduced into the diet. Within Western populations, differences according to geography, diet, and lifestyle were also highlighted. *Faecalibacterium* SGBs enriched in Chinese subjects showed a higher prevalence of genes involved in starch degradation, although genes related to lactose and protein metabolism were

Please cite this article in press as: De Filippis et al., Newly Explored *Faecalibacterium* Diversity Is Connected to Age, Lifestyle, Geography, and Disease, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.09.063

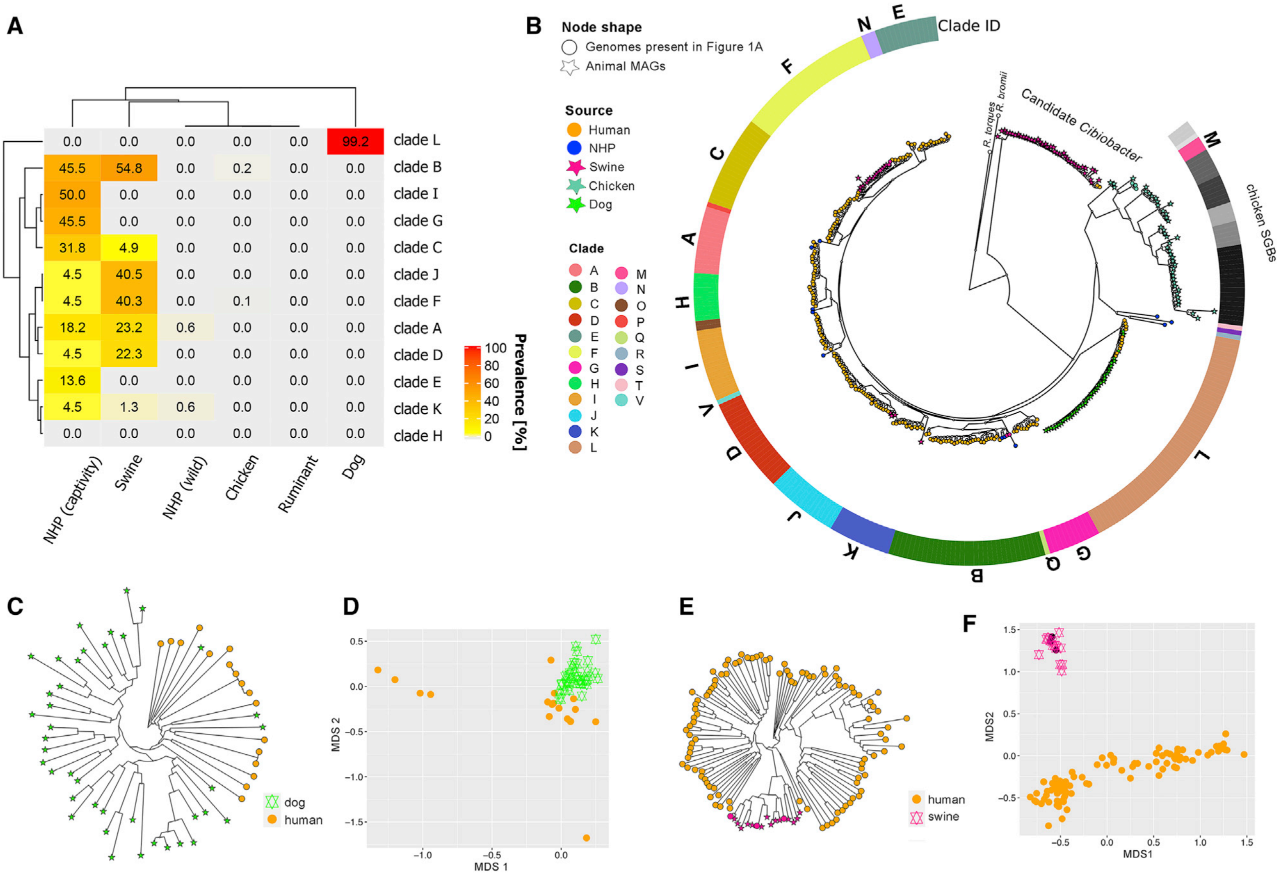


Figure 7. *Faecalibacterium* Diversity across Different Animal Hosts

(A) Average prevalence of the 12 major human-associated *Faecalibacterium* clades in different animal species.

(B) Phylogenetic tree, including all human and NHP MAGs reported in Figure 1, as well as all *Faecalibacterium*-like MAGs retrieved from different animal metagenomes. Outer ring is colored according to the clade. Chicken *Faecalibacterium*-like MAGs are colored in gray scale. Clade names are indicated on the top.

(C–F) Phylogenetic trees (C and E) and MDS (D and F) of MAGs from humans and animals belonging to *Faecalibacterium* clades L (C and D) and F (E and F).

See also Figure S4.

depleted, reflecting a diet richer in rice and poorer in milk and proteins compared with those of European and North American subjects [38, 39]. Moreover, enrichment in genes related to antibiotic resistance was observed in the *Faecalibacterium* pangenome of Western subjects as well as in Chinese subjects compared with subjects from other Western countries. This result may be a consequence of the substantial increase in antibiotic consumption registered in Asia in recent decades [40].

Strain-level diversity was also detected in some of the major *Faecalibacterium* SGBs, where up to four different strains were found. In some cases, the distribution of the strains was found to be related to geographical origin, highlighting that a selection process may also occur at the strain level, according to different diets or lifestyles. In addition, although some *Faecalibacterium* species/strains seemed to be host adapted, some strains were shared between humans and other mammals living in close association (swine, dogs, and NHPs living in captivity) and most likely sharing environments and foods, highlighting the ability of these strains to adapt to different hosts.

Finally, we show that the current association of the whole *Faecalibacterium* genus with positive health effects might be an oversimplification that ignores the high level of diversity existing within this genus. Indeed, we highlight that IBD and CRC are associated with lower *Faecalibacterium* diversity, as previously suggested [15, 35], although we were able to identify much higher diversity than previous studies. Moreover, our results support the hypothesis that this decrease in diversity may be generalized to other inflammatory states, such as obesity; the decrease in *F. prausnitzii* often reported in the literature for such conditions is likely due only to the declines in specific SGBs; and recovering of such diversity through diet modulation is possible.

The diversity found here should be taken into account in the development of novel, personalized strategies for the modulation of specific *F. prausnitzii* strains. The results will surely be of value for future interpretations of diet-driven features and changes in the gut microbiome or for the evaluation of diseases associated with dysbiosis. In addition, the newly explored *F. prausnitzii* diversity will be able to support the choice of strains suitable for NGPs, guided by the consideration of strain differences in functional potential.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead Contact
  - ○ Materials Availability
  - ○ Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Human metagenomes
  - ○ Animal metagenomes
- METHOD DETAILS
  - ○ Faecalibacterium genomes from isolates and metagenome-assembled genomes (MAGs) from publicly available databases
  - ○ Species- and sub-species level genome clustering
  - ○ Reconstruction of phylogenies
  - ○ Species-specific marker gene identification and metagenome mapping
  - ○ Validation of mapping-based pipeline on real and synthetic metagenomes
  - ○ Faecalibacterium mapping and MAG reconstruction from animal gut metagenomes
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Statistical analysis

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cub.2020.09.063.

ACKNOWLEDGMENTS

The study has received support by the JPI HDHL-INTIMIC-Knowledge Platform of Food, Diet, Intestinal Microbiomics and Human Health, granted by the Italian Ministry of Agricultural, Food and Forestry and Tourism Policies with ID 790. The study was also developed within the project MASTER (microbiome applications for sustainable food systems through technologies and enterprise), receiving funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 818368. This manuscript reflects only the authors' views, and the European Commission is not responsible for any use that may be made of the information it contains.

AUTHOR CONTRIBUTIONS

Conceptualization, F.D.F., E.P., and D.E.; Methodology, F.D.F. and E.P.; Formal Analysis, F.D.F. and E.P.; Investigation, F.D.F. and E.P.; Resources, D.E.; Data Curation, F.D.F. and E.P.; Writing – Original Draft, F.D.F.; Writing – Review & Editing, D.E. and E.P.; Funding Acquisition, D.E.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 6, 2020
Revised: September 2, 2020
Accepted: September 18, 2020
Published: October 15, 2020

REFERENCES

1. Benevides, L., Burman, S., Martin, R., Robert, V., Thomas, M., Miquel, S., Chain, F., Sokol, H., Bermudez-Humaran, L.G., Morrison, M., et al. (2017). New insights into the diversity of the genus *Faecalibacterium*. Front. Microbiol. *8*, 1790.
2. Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., Kurilshikov, A., Bonder, M.J., Valles-Colomer, M., Vandeputte, D., et al. (2016). Population-level analysis of gut microbiome variation. Science *352*, 560–564.
3. Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., Costea, P.I., Godneva, A., Kalka, I.N., Bar, N., et al. (2018). Environment dominates over host genetics in shaping human gut microbiota. Nature *555*, 210–215.
4. Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A.T., Creasy, H.H., Earl, A.M., FitzGerald, M.G., Fulton, R.S., et al.; Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214.
5. Leylabadlo, H.E., Ghotaslou, R., Feizabadi, M.M., Farajnia, S., Moaddab, S.Y., Ganbarov, K., Khodadadi, E., Tanomand, A., Sheykhsaran, E., Yousefi, B., and Kafil, H.S. (2020). The critical role of *Faecalibacterium prausnitzii* in human health: An overview. Microb. Pathog. *149*, 104344.
6. Ferreira-Halder, C.V., Faria, A.V.S., and Andrade, S.S. (2017). Action and function of *Faecalibacterium prausnitzii* in health and disease. Best Pract. Res. Clin. Gastroenterol. *31*, 643–648.
7. Lopez-Siles, M., Duncan, S.H., Garcia-Gil, L.J., and Martinez-Medina, M. (2017). *Faecalibacterium prausnitzii*: from microbiology to diagnostics and prognostics. ISME J. *11*, 841–852.
8. Zhou, L., Zhang, M., Wang, Y., Dorfman, R.G., Liu, H., Yu, T., Chen, X., Tang, D., Xu, L., Yin, Y., et al. (2018). *Faecalibacterium prausnitzii* produces butyrate to maintain Th17/Treg balance and to ameliorate colorectal colitis by inhibiting histone deacetylase 1. Inflamm. Bowel Dis. *24*, 1926–1940.
9. Quévrain, E., Maubert, M.A., Michon, C., Chain, F., Marquant, R., Tailhades, J., Miquel, S., Carlier, L., Bermúdez-Humarán, L.G., Pigneur, B., et al. (2016). Identification of an anti-inflammatory protein from *Faecalibacterium prausnitzii*, a commensal bacterium deficient in Crohn's disease. Gut *65*, 415–425.
10. Rossi, O., Khan, M.T., Schwarzer, M., Hudcovic, T., Srutkova, D., Duncan, S.H., Stolte, E.H., Kozakova, H., Flint, H.J., Samsom, J.N., et al. (2015). *Faecalibacterium prausnitzii* strain HTF-F and its extracellular polymeric matrix attenuate clinical parameters in DSS-induced colitis. PLoS ONE *10*, e0123013.
11. Martín, R., Bermúdez-Humarán, L.G., and Langella, P. (2018). Searching for the bacterial effector: the example of the multi-skilled commensal bacterium *Faecalibacterium prausnitzii*. Front. Microbiol. *9*, 346.
12. Miquel, S., Martín, R., Rossi, O., Bermúdez-Humarán, L.G., Chatel, J.M., Sokol, H., Thomas, M., Wells, J.M., and Langella, P. (2013). *Faecalibacterium prausnitzii* and human intestinal health. Curr. Opin. Microbiol. *16*, 255–261.
13. O'Toole, P.W., Marchesi, J.R., and Hill, C. (2017). Next-generation probiotics: the spectrum from probiotics to live biotherapeutics. Nat. Microbiol. *2*, 17057.
14. Lopez-Siles, M., Khan, T.M., Duncan, S.H., Harmsen, H.J.M., Garcia-Gil, L.J., and Flint, H.J. (2012). Cultured representatives of two major phylogroups of human colonic *Faecalibacterium prausnitzii* can utilize pectin, uronic acids, and host-derived substrates for growth. Appl. Environ. Microbiol. *78*, 420–428.
15. Lopez-Siles, M., Martinez-Medina, M., Abellà, C., Busquets, D., Sabat-Mir, M., Duncan, S.H., Aldeguer, X., Flint, H.J., and Garcia-Gil, L.J. (2015). Mucosa-associated *Faecalibacterium prausnitzii* phylotype richness is reduced in patients with inflammatory bowel disease. Appl. Environ. Microbiol. *81*, 7582–7592.
16. Lopez-Siles, M., Martinez-Medina, M., Surís-Valls, R., Aldeguer, X., Sabat-Mir, M., Duncan, S.H., Flint, H.J., and Garcia-Gil, L.J. (2016).

Please cite this article in press as: De Filippis et al., Newly Explored *Faecalibacterium* Diversity Is Connected to Age, Lifestyle, Geography, and Disease, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.09.063

Changes in the abundance of *Faecalibacterium prausnitzii* phylogroups I and II in the intestinal mucosa of inflammatory bowel disease and patients with colorectal cancer. Inflamm. Bowel Dis. *22*, 28–41.

17. Fitzgerald, C.B., Shkoporov, A.N., Sutton, T.D.S., Chaplin, A.V., Velayudhan, V., Ross, R.P., and Hill, C. (2018). Comparative analysis of *Faecalibacterium prausnitzii* genomes shows a high level of genome plasticity and warrants separation into new species-level taxa. BMC Genomics *19*, 931.
18. Van Rossum, T., Ferretti, P., Maistrenko, O.M., and Bork, P. (2020). Diversity within species: interpreting strains in microbiomes. Nat. Rev. Microbiol. *18*, 491–506.
19. De Filippis, F., Pasolli, E., Tett, A., Tarallo, S., Naccarati, A., De Angelis, M., Neviani, E., Cocolin, L., Gobbetti, M., Segata, N., and Ercolini, D. (2019). Distinct genetic and functional traits of human intestinal *Prevotella copri* strains are associated with different habitual diets. Cell Host Microbe *25*, 444–453.e3.
20. De Filippis, F., Vitaglione, P., Cuomo, R., Berni Canani, R., and Ercolini, D. (2018). Dietary interventions to modulate the gut microbiome - how far away are we from precision medicine. Inflamm. Bowel Dis. *24*, 2142–2154.
21. Roager, H.M., and Dragsted, L.O. (2019). Diet-derived microbial metabolites in health and disease. Nutr. Bull. *44*, 216–227.
22. Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D.T., Beghini, F., Malik, F., Ramos, M., Dowd, J.B., et al. (2017). Accessible, curated metagenomic data through ExperimentHub. Nat. Methods *14*, 1023–1024.
23. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell *176*, 649–662.e20.
24. Manara, S., Asnicar, F., Beghini, F., Bazzani, D., Cumbo, F., Zolfo, M., Nigro, E., Karcher, N., Manghi, P., Metzger, M.I., et al. (2019). Microbial genomes from non-human primate gut metagenomes expand the primate-associated bacterial tree of life with over 1000 novel species. Genome Biol. *20*, 299.
25. Moss, E.L., Maghini, D.G., and Bhatt, A.S. (2020). Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nat. Biotechnol. *38*, 701–707.
26. Tett, A., Huang, K.D., Asnicar, F., Fehlner-Peach, H., Pasolli, E., Karcher, N., Armanini, F., Manghi, P., Bonham, K., Zolfo, M., et al. (2019). The *Prevotella copri* complex comprises four distinct clades underrepresented in Westernized populations. Cell Host Microbe *26*, 666–679.e7.
27. Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. *9*, 5114.
28. Amato, K.R., G Sanders, J., Song, S.J., Nute, M., Metcalf, J.L., Thompson, L.R., Morton, J.T., Amir, A., J McKenzie, V., Humphrey, G., et al. (2019). Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes. ISME J. *13*, 576–587.
29. Tung, J., Barreiro, L.B., Burns, M.B., Grenier, J.-C., Lynch, J., Grieneisen, L.E., Altmann, J., Alberts, S.C., Blekhman, R., and Archie, E.A. (2015). Social networks predict gut microbiome composition in wild baboons. eLife *4*, e05224.
30. Li, X., Liang, S., Xia, Z., Qu, J., Liu, H., Liu, C., Yang, H., Wang, J., Madsen, L., Hou, Y., et al. (2018). Establishment of a *Macaca fascicularis* gut microbiome gene catalog and comparison with the human, pig, and mouse gut microbiomes. Gigascience *7*, giy100.
31. Meslier, V., Laiola, M., Roager, H.M., De Filippis, F., Roume, H., Quinquis, B., Giacco, R., Mennella, I., Ferracane, R., Pons, N., et al. (2020). Mediterranean diet intervention in overweight and obese subjects lowers plasma cholesterol and causes changes in the gut microbiome and metabolome independently of energy intake. Gut *69*, 1258–1268.
32. Ghosh, T.S., Rampelli, S., Jeffery, I.B., Santoro, A., Neto, M., Capri, M., Giampieri, E., Jennings, A., Candela, M., Turroni, S., et al. (2020). Mediterranean diet intervention alters the gut microbiome in older people reducing frailty and improving health status: the NU-AGE 1-year dietary intervention across five European countries. Gut *69*, 1218–1228.
33. Verhoog, S., Taneri, P.E., Roa Díaz, Z.M., Marques-Vidal, P., Troup, J.P., Bally, L., Franco, O.H., Glisic, M., and Muka, T. (2019). Dietary factors and modulation of bacteria strains of *Akkermansia muciniphila* and *Faecalibacterium prausnitzii*: a systematic review. Nutrients *11*, 1565.
34. Zhang, M., Zhou, L., Wang, Y., Dorfman, R.G., Tang, D., Xu, L., Pan, Y., Zhou, Q., Li, Y., Yin, Y., et al. (2019). *Faecalibacterium prausnitzii* produces butyrate to decrease c-Myc-related metabolism and Th17 differentiation by inhibiting histone deacetylase 3. Int. Immunol. *31*, 499–514.
35. Lopez-Siles, M., Enrich-Capó, N., Aldeguer, X., Sabat-Mir, M., Duncan, S.H., Garcia-Gil, L.J., and Martinez-Medina, M. (2018). Alterations in the abundance and co-occurrence of *Akkermansia muciniphila* and *Faecalibacterium prausnitzii* in the colonic mucosa of inflammatory bowel disease subjects. Front. Cell. Infect. Microbiol. *8*, 281.
36. Thingholm, L.B., Rühlemann, M.C., Koch, M., Fuqua, B., Laucke, G., Boehm, R., Bang, C., Franzosa, E.A., Hübenthal, M., Rahnavard, A., et al. (2019). Obese individuals with and without Type 2 diabetes show different gut microbial functional capacity and composition. Cell Host Microbe *26*, 252–264.e10.
37. Martín, R., Miquel, S., Benevides, L., Bridonneau, C., Robert, V., Hudault, S., Chain, F., Berteau, O., Azevedo, V., Chatel, J.M., et al. (2017). Functional characterization of novel *Faecalibacterium prausnitzii* strains isolated from healthy volunteers: a step forward in the use of *F. prausnitzii* as a next-generation probiotic. Front. Microbiol. *8*, 1226.
38. Afshin, A., Sur, P.J., Fay, K.A., Cornaby, L., Ferrara, G., Salama, J.S., Mullany, E.C., Abate, K.H., Abbafati, C., Abebe, Z., et al.; GBD 2017 Diet Collaborators (2019). Health effects of dietary risks in 195 countries, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet *393*, 1958–1972.
39. Zhang, R., Wang, Z., Fei, Y., Zhou, B., Zheng, S., Wang, L., Huang, L., Jiang, S., Liu, Z., Jiang, J., and Yu, Y. (2015). The difference in nutrient intakes between Chinese and Mediterranean, Japanese and American diets. Nutrients *7*, 4661–4688.
40. Klein, E.Y., Van Boeckel, T.P., Martinez, E.M., Pant, S., Gandra, S., Levin, S.A., Goossens, H., and Laxminarayan, R. (2018). Global increase and geographic convergence in antibiotic consumption between 2000 and 2015. Proc. Natl. Acad. Sci. USA *115*, E3463–E3470.
41. Poyet, M., Groussin, M., Gibbons, S.M., Avila-Pacheco, J., Jiang, X., Kearney, S.M., Perrotta, A.R., Berdy, B., Zhao, S., Lieberman, T.D., et al. (2019). A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. Nat. Med. *25*, 1442–1452.
42. Coelho, L.P., Kultima, J.R., Costea, P.I., Fournier, C., Pan, Y., Czarnecki-Maulden, G., Hayward, M.R., Forslund, S.K., Schmidt, T.S.B., Descombes, P., et al. (2018). Similarity of the dog and human gut microbiomes in gene content and response to diet. Microbiome *6*, 72.
43. Glendinning, L., Stewart, R.D., Pallen, M.J., Watson, K.A., and Watson, M. (2020). Assembly of hundreds of novel bacterial genomes from the chicken caecum. Genome Biol. *21*, 34.
44. Hou, Q., Kwok, L.-Y., Zheng, Y., Wang, L., Guo, Z., Zhang, J., Huang, W., Wang, Y., Leng, L., Li, H., and Zhang, H. (2016). Differential fecal microbiota are retained in broiler chicken lines divergently selected for fatness traits. Sci. Rep. *6*, 37376.
45. Huang, P., Zhang, Y., Xiao, K., Jiang, F., Wang, H., Tang, D., Liu, D., Liu, B., Liu, Y., He, X., et al. (2018). The chicken gut metagenome and the modulatory effects of plant-derived benzylisoquinoline alkaloids. Microbiome *6*, 211.
46. Munk, P., Knudsen, B.E., Lukjancenko, O., Duarte, A.S.R., Van Gompel, L., Luiken, R.E.C., Smit, L.A.M., Schmitt, H., Garcia, A.D., Hansen, R.B., et al.; EFFORT Group (2018). Abundance and diversity of the faecal resistome in slaughter pigs and broilers in nine European countries. Nat. Microbiol. *3*, 898–908.

47. Thomas, M., Wongkuna, S., Ghimire, S., Kumar, R., Antony, L., Doerner, K.C., Singery, A., Nelson, E., Woyengo, T., Chankhamhaengdecha, S., et al. (2019). Gut microbial dynamics during conventionalization of germfree chicken. MSphere *4*, e00035-19.

48. Tan, Z., Yang, T., Wang, Y., Xing, K., Zhang, F., Zhao, X., Ao, H., Chen, S., Liu, J., and Wang, C. (2017). Metagenomic analysis of cecal microbiome identified microbiota and functional capacities associated with feed efficiency in landrace finishing pigs. Front. Microbiol. *8*, 1546.

49. Wang, C., Li, P., Yan, Q., Chen, L., Li, T., Zhang, W., Li, H., Chen, C., Han, X., Zhang, S., et al. (2019). Characterization of the pig gut microbiome and antibiotic resistome in industrialized feedlots in China. mSystems *4*, e00206–e00219.

50. Wang, W., Hu, H., Zijlstra, R.T., Zheng, J., and Gänzle, M.G. (2019). Metagenomic reconstructions of gut microbial metabolism in weanling pigs. Microbiome *7*, 48.

51. Xiao, L., Estellé, J., Kiilerich, P., Ramayo-Caldas, Y., Xia, Z., Feng, Q., Liang, S., Pedersen, A.Ø., Kjeldsen, N.J., Liu, C., et al. (2016). A reference gene catalogue of the pig gut microbiome. Nat. Microbiol. *1*, 16161.

52. Yang, H., Huang, X., Fang, S., Xin, W., Huang, L., and Chen, C. (2016). Uncovering the composition of microbial community structure and metagenomics among three gut locations in pigs with distinct fatness. Sci. Rep. *6*, 27427.

53. Stewart, R.D., Auffret, M.D., Warr, A., Walker, A.W., Roehe, R., and Watson, M. (2019). Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat. Biotechnol. *37*, 953–961.

54. Glendinning, L., Genç, B., Wallace, R.J., and Watson, M. (2020). Metagenomic analysis of the cow, sheep, reindeer and red deer rumen. bioRxiv. https://doi.org/10.1101/2020.02.12.945139.

55. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

56. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055.

57. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069.

58. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.

59. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132.

60. Asnicar, F., Thomas, A.M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., Zhu, Q., Bolzan, M., Cumbo, F., May, U., et al. (2020). Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nat. Commun. *11*, 2500.

61. Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., and Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics *31*, 3691–3693.

62. Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods *12*, 902–903.

63. Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics *31*, 1674–1676.

64. Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. *47* (W1), W256–W259.

65. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

66. Müllner, D. (2013). fastcluster: fast hierarchical, agglomerative clustering routines for R and Python. J. Stat. Softw. *53*, 1–18.

67. Costea, P.I., Coelho, L.P., Sunagawa, S., Munch, R., Huerta-Cepas, J., Forslund, K., Hildebrand, F., Kushugulova, A., Zeller, G., and Bork, P. (2017). Subspecies in the global human gut microbiome. Mol. Syst. Biol. *13*, 960.

68. Hennig, C. (2007). Cluster-wise assessment of cluster stability. Comput. Stat. Data Anal. *52*, 258–271.

69. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

70. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.

71. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE *5*, e9490.

72. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

73. Asnicar, F., Weingart, G., Tickle, T.L., Huttenhower, C., and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ *3*, e1029.

74. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al.; MetaHIT Consortium (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

75. Fritz, A., Hofmann, P., Majda, S., Dahms, E., Dröge, J., Fiedler, J., Lesker, T.R., Belmann, P., DeMaere, M.Z., Darling, A.E., et al. (2019). CAMISIM: simulating metagenomes and microbial communities. Microbiome *7*, 17.

76. Kang, D.D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ *3*, e1165.

Please cite this article in press as: De Filippis et al., Newly Explored *Faecalibacterium* Diversity Is Connected to Age, Lifestyle, Geography, and Disease, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.09.063

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Faecalibacterium-like MAGs from human gut metagenomes	[23]	http://segatalab.cibio.unitn.it/data/Pasolli_et_al.html
Faecalibacterium-like MAGs from non-human primate gut metagenomes	[24]	http://segatalab.cibio.unitn.it/data/Manara_et_al.html
Faecalibacterium-like MAGs from swine, chicken and dog gut metagenomes	This study	Mendeley Data, https://doi.org/10.17632/t74rxwrd6z.1
NCBI *Faecalibacterium* reference genomes	NCBI	Accession numbers reported in Suppl. Data S1A
Faecalibacterium genomes – raw reads	[41]	NCBI: PRJNA544527
Faecalibacterium marker genes database	This study	Mendeley Data, https://doi.org/10.17632/r5rknt2sbx.1
Dog gut metagenomes	[42]	European Nucleotide Archive: PRJEB20308
Chicken gut metagenomes	[43]	European Nucleotide Archive: PRJEB33338
Chicken gut metagenomes	[44]	NCBI: SRP083441
Chicken gut metagenomes	[45]	NCBI: PRJNA417359
Chicken and swine gut metagenomes	[46]	European Nucleotide Archive: PRJEB22062
Chicken gut metagenomes	NCBI	NCBI: PRJEB21076
Chicken gut metagenomes	NCBI	NCBI: PRJEB23356
Chicken gut metagenomes	NCBI	NCBI: PRJNA298736
Chicken gut metagenomes	[47]	NCBI: PRJNA415593
Swine gut metagenomes	[48]	NCBI: SRP108960
Swine gut metagenomes	[49]	NCBI: PRJEB31742
Swine gut metagenomes	[50]	https://doi.org/10.1186/s40168-019-0662-1
Swine gut metagenomes	[51]	European Nucleotide Archive: PRJEB11755
Swine gut metagenomes	[52]	https://doi.org/10.1038/srep27427
Cow rumen metagenomes	[53]	NCBI: PRJEB31266, PRJEB21624
Ruminant gut metagenomes	[54]	European Nucleotide Archive: PRJEB34458
Software and Algorithms		
Blast 2.6	[55]	https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download
CheckM 1.1.2	[56]	https://ecogenomics.github.io/CheckM/
Prokka 1.11	[57]	https://github.com/tseemann/prokka
diamond 0.9.9.110	[58]	http://github.com/bbuchfink/diamond
Mash 2.0	[59]	https://github.com/marbl/Mash/releases
fastANI 1.0	[27]	https://omictools.com/fastani-tool
PhyloPhlAn 3.0	[60]	https://github.com/biobakery/phylophlan
Roary 3.11.2	[61]	https://github.com/sanger-pathogens/Roary/blob/master/README.md
MetaPhlAn2	[62]	https://github.com/biobakery/MetaPhlAn
MEGAHIT 1.2.9	[63]	https://github.com/voutcn/megahit
GraPhlAn 1.1	[60]	https://bitbucket.org/nsegata/graphlan/wiki/Home
iTOL 5.5.1	[64]	https://itol.embl.de
CMSeq	[23]	https://github.com/SegataLab/cmseq
Bowtie2 2.4	[65]	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

Please cite this article in press as: De Filippis et al., Newly Explored *Faecalibacterium* Diversity Is Connected to Age, Lifestyle, Geography, and Disease, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.09.063

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Danilo Ercolini (ercolini@unina.it).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

All human and animal datasets used in this study are available in public repositories (see Key Resources Table and Data S1 for Accession Numbers). The taxonomic profiles with associated metadata from the human metagenomes are available in the *curatedMetagenomicData* package [22]. The MAGs from human metagenomes used in this study are available at http://segatalab.cibio.unitn.it/data/Pasolli_et_al.html. The MAGs from NHP metagenomes used in this study are available at http://segatalab.cibio.unitn.it/data/Manara_et_al.html. The *Faecalibacterium* MAGs from animal gut metagenomes generated in this study and the *Faecalibacterium*-complex marker gene database are available on Mendeley Data (https://doi.org/10.17632/t74rxwrd6z.1 and http://doi.org/10.17632/r5rknt2sbx.1, respectively).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human metagenomes

Human metagenome datasets used in this study are from 47 previously published studies and are available in public repositories (see Key Resources Table and Data S1 for Accession Numbers) and included 7,907 subjects, spanning different ages, countries, health status and lifestyles. Metadata collection and *Faecalibacterium* MAGs reconstruction was previously carried out [23].

Animal metagenomes

Previously published animal metagenomes from 23 different studies were retrieved (see Key Resources Table and Data S1 for Accession Numbers). Data included non-human primate (n = 203), dog (n = 129), chicken (n = 1350), swine (n = 695) and ruminant (n = 295) metagenomes. MAGs were retrieved with the same pipeline used for human MAGs, as reported below.

METHOD DETAILS

Faecalibacterium genomes from isolates and metagenome-assembled genomes (MAGs) from publicly available databases

We downloaded all the genomes (n = 98) labeled as *Faecalibacterium* from the NCBI GenBank database (May 2019). We enlarged this set by considering MAGs of *Faecalibacterium* spp. that we retrieved from recently published catalogs of the gut microbiome from human (n= 7,907 from 47 studies) and non-human primate (NHP; n = 203 from 6 studies) metagenomes (Data S1B and S1C) [23, 24]. No *Faecalibacterium* MAGs were retrieved from 1,538 human metagenomes from body sites other than the gut (mouth, skin, airways, and vagina) [23]. Only MAGs and isolate genomes with > 80% completeness and < 5% contamination as estimated by CheckM [56] were retained for further analyses. This filtering led to a total of 2,859 genomes that could be categorized as follows: 64 NCBI reference genomes (49 genomes from isolates and 15 MAGs), 2,784 human MAGs (from [23]), and 11 NHP MAGs (from [24]) (see Data S1A–S1C). Protein-coding genes were inferred for each genome with Prokka (default parameters) [57] and functionally annotated with UniRef90 using DIAMOND v0.9.9.110 [58].

Species- and sub-species level genome clustering

Pairwise genetic distances between genomes were calculated using Mash (version 2.0; option "-s 10000" for sketching) [59], and genomes were clustered through hierarchical clustering (package *fastcluster* single-linkage) [66]. Species-level genome bins (SGBs) were obtained by cutting the dendrogram with a 5% distance threshold, which generated a total of twenty-two distinct SGBs, collectively named as *Faecalibacterium*-complex (as previously suggested for *P. copri* [26];). We named these SGBs as Clade A through Clade V, sorting them in descending order by the number of genomes (Table S1). To define sub-species, pairwise average nucleotide identities (ANI) were estimated between genomes falling in the same SGB through FastANI [27] and clustered using the Partitioning Around Medoid (PAM) algorithm. The number of sub-species was estimated by identifying the optimal number of clusters for each SGB. This was performed by considering a more stringent criterion than previously employed. We required the number of clusters to be supported by a prediction strength (*prediction.strength* R function) > 0.8 [26, 67], in addition to an average Jaccard similarity (*clusterboot* R function with bootstrap resampling [68],) > 0.85 (Table S2).

Reconstruction of phylogenies

Phylogenetic trees were inferred with PhyloPhlAn 3.0 [60]. The trees spanning the entire *Faecalibacterium* genus (Figures 1 and 7) were based on the set of *Faecalibacterium*-specific marker genes that can be retrieved with the command phylophlan2_setup_database.py. Each SGB-specific phylogeny was built by using the SGB-specific markers that we extracted as described in the next section. This departed from the default option of using the 400 universal markers and guaranteed a higher phylogenetic resolution. The parameters were set as follows: "–diversity low–fast–min_num_marker 50." External tools embedded in PhyloPhlAn 3.0 were run with their specific options as follows:

- BLASTN (version 2.6.0+ [55];) with the parameters "-outfmt 6 -max_target_seqs 1000000";
- MAFFT (version 7.310 [69];) using the "L-INS-i" algorithm and with the parameters "–anysymbol–auto";
- trimAl (version 1.2rev59 [70];) with the parameter "-gappyout";
- FastTree (version 2.1.9 [71];) with the parameters "-mlacc 2 -slownni -spr 4 -fastest -mlnni 4 -no2nd -gtr -nt";
- RAxML (version 8.1.15 [72];) with the parameters "-p 1989 -m GTRCAT -t <phylogenetic tree computed by FastTree>."

Phylogenetic trees were visualized using GraPhlAn [73], except for the tree in Figure 1, which was visualized in iTOL 5.5.1 [64]. Classical Multidimensional Scaling (MDS, *cmdscale* R function) was carried out on the ANI distance matrix and the Mash distance matrix for the SGB-specific and the overall plot, respectively, which were visualized using the *ggplot2* R package.

Species-specific marker gene identification and metagenome mapping

We identified marker genes specific to each species for the 12 major clades (> 10 genomes) found in the human gut. This was achieved by running Prokka (default parameters [57];) on each genome and then finding core genes with Roary (-i 95 [61];). A gene was defined as core if present in at least 95% of the genomes of the considered SGB (the computational load was limited by randomly selecting 125 genomes for SGBs including > 125 genomes). Such core genes were used to build the SGB-specific phylogenies as described in the previous section. We further moved from core to marker genes by removing the genes occurring in more than one SGB and those < 100 bp, as previously suggested [74]. More specifically, core genes were mapped against all collected genomes using BLASTN and removed if found (> 95% identity over > 50% of the gene length) in > 5% of the genomes of any other SGB. This produced a total of 6,201 unique marker genes, distributed as follows: 395 for clade A, 171 for clade B, 760 for clade C, 340 for clade D, 653 for clade E, 542 for clade F, 444 for clade G, 410 for clade H, 678 for clade I, 408 for clade J, 815 for clade K, and 585 for clade L.

Marker genes were employed to estimate the occurrence of the SGB through a raw read mapping approach based on multiple steps devoted to the minimization of false positives. First, raw reads were mapped against marker genes using Bowtie2 ("–sensitive," [65]). Then, we built the consensus sequences using CMSeq (https://github.com/SegataLab/cmseq [23];) and labeled a gene as present only if it had > 95% identity over > 50% of the gene length. Finally, a clade was considered present in a given metagenome if at least 10% of the markers were hit.

We validated our pipeline by considering a compendium of synthetic and real datasets (see following paragraph) and used this mapping-based pipeline to estimate clades' occurrence in each of the 9,445 human metagenomes [23] already considered for MAG reconstruction. We further applied the pipeline to an additional set of 1,033 human metagenomes coming from 10 datasets (Data S1D) that we retrieved specifically in this work to validate findings using independent cohorts. For the same set of metagenomes, we used MetaPhlAn2 (default parameters [62];) to estimate the relative abundance of *Faecalibacterium* at the genus level (since all *Faecalibacterium* genomes used to build the MetaPhlAn2 database were collapsed into a single species-level taxon named *F. prausnitzii* and no other *Faecalibacterium* species were available). Finally, we applied the mapping-based strategy on a recently published cohort, including obese subjects consuming a Mediterranean diet for 2 months [31].

Validation of mapping-based pipeline on real and synthetic metagenomes

We validated the mapping pipeline by considering a compendium of synthetic and real datasets from both isolate and metagenomic samples (results summarized in Figure S5). We downloaded raw reads for all the isolate genomes (n = 31) sequenced with Illumina and labeled as *F. prausnitzii* from NCBI Sequence Read Archive (SRA) (list of the accession numbers in Data S1F). For 30 out of 31 cases we clearly identified a single clade associated with the sample (Figure S5A). The percentage of hit markers were well above and below the threshold we set to define a clade present or absent in a sample, respectively. Interestingly, most of the samples came from clade B, only few of them were associated with clades A, C, D, and E, while none was assigned to the other seven clades. We didn't detect any clades for sample SRR7969831 (not reported in Figure S5A), which however resulted to be a mislabeling case, since it was found to be close to the reference genome of *Ruminococcus* sp. DSM 100440 (98.7% percent identity) after genome assembling and taxonomic assignment [23]. We further validated the pipeline on metagenomes (Figure S5B). First, we considered the ten synthetic metagenomes coming from the Critical Assessment of Metagenome Annotation (CAMI) challenge [75] and belonging to the gastrointestinal tract. Gold standards associated with raw reads reported the lack of *Faecalibacterium* in all the samples, which was confirmed by our results (Figure S5B). Then, we considered the metagenome of a mock community (ZymoBIOMICS Gut Microbiome Standard; https://files.zymoresearch.com/protocols/_d6331_zymobiomics_gut_microbiome_standard.pdf), that comprised 21 different strains including one of *F. prausnitzii*, that was identified by our approach as clade F (Figure S5B). We finally considered ten synthetic metagenomes. Each of them was built by randomly choosing a real human metagenome having an abundance of *F. prausnitzii* estimated by MetaPhlAn2 equal to 0.0 and adding raw reads of multiple genomes of *F. prausnitzii* isolates (we varied the number of clades between three and five). In all the cases, the approach was able to identify correctly the clades present in the samples.

Faecalibacterium mapping and MAG reconstruction from animal gut metagenomes

The same procedure applied to human metagenomes was employed to conduct a large-scale analysis of animal gut microbiomes. We considered a total of 2,672 animal metagenomes by integrating the 203 NHP samples [24] with other metagenomes from chickens (n = 1,350), dogs (n = 129), ruminants (n = 295), and swine (n = 695) that we retrieved specifically for this study and downloaded from the NCBI SRA. A list of the newly collected animal gut datasets is provided in Data S1E. The developed mapping pipeline was considered to estimate the occurrence of the 12 major human clades, along with the generation of taxonomic profiles with MetaPhlAn2 (default parameters [62];). Samples showing a *Faecalibacterium* abundance > 1% were processed with the same pipeline used for the human and NHP datasets [23, 24] to extract MAGs. Briefly, metagenomes were assembled independently using MEGAHIT [63] contigs > 1,000 bp were binned using MetaBAT2 [76], and MAG quality was estimated with CheckM [56]. Only MAGs with > 80% completeness and < 5% contamination and identified as *Faecalibacterium* spp. (N = 142) were retained for further analyses.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis

Statistical significance of gene or SGB prevalence was verified through Fisher’s test with multiple-hypothesis testing corrections via the false discovery rate (FDR). Comparisons of *Faecalibacterium* abundance and diversity were carried out using pairwise Wilcoxon tests.