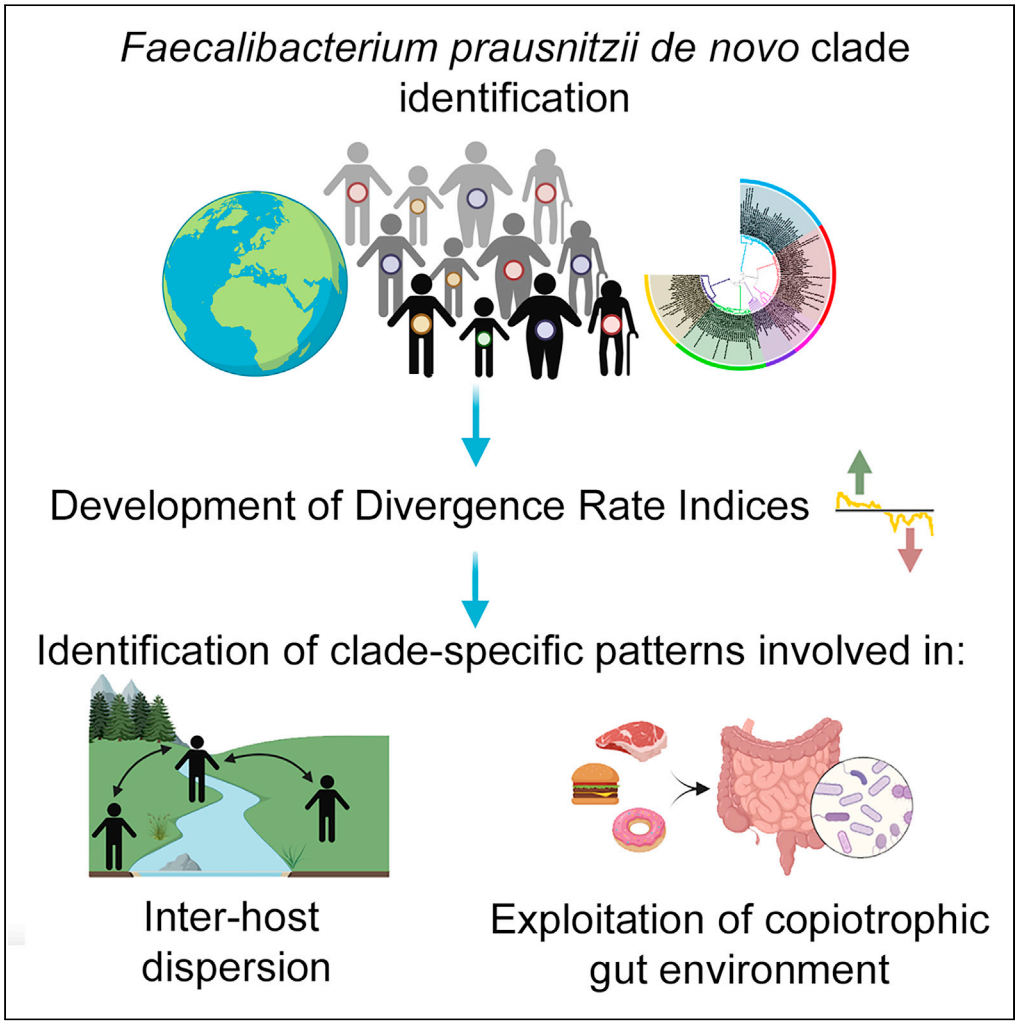
iScience

CellPress
OPEN ACCESS

Article

Exploring clade differentiation of the *Faecalibacterium prausnitzii* complex



Marco Fabbrini, Marco Candela, Silvia Turroni, Patrizia Brigidi, Simone Rampelli

simone.rampelli@unibo.it

Highlights

Five *Faecalibacterium prausnitzii* clades were *de novo* identified

Divergence rate indices were developed at clade level and gene level

Specific patterns of diverging genes were identified for each clade

Patterns concern inter-host dispersion and a better exploitation of the gut environment

Fabbrini et al., iScience *25*, 105533
December 22, 2022
https://doi.org/10.1016/j.isci.2022.105533

iScience

CellPress
OPEN ACCESS

Article

Exploring clade differentiation of the *Faecalibacterium prausnitzii* complex

Marco Fabbrini,[1] Marco Candela,[2] Silvia Turroni,[2] Patrizia Brigidi,[1] and Simone Rampelli[2,*]

SUMMARY

***Faecalibacterium prausnitzii* is one of the most prevalent and abundant polyphyletic health-promoting components of the human gut microbiome with a propensity for dysbiotic decreases. To better understand its biology in the human gut, we specifically explored the divergence pressures acting on *F. prausnitzii* clades on a global scale. Five *F. prausnitzii* clades were *de novo* identified from 55 publicly available genomes and 92 high-quality metagenome assembled genomes. Divergence rate indices were constructed and validated to compare the divergence rates among the different clades and between each of the diverging genes. For each clade we identified specific patterns of diverging functionalities, probably reflecting different ecological propensities, in term of inter-host dispersion capacity or exploitation of different substrates in the gut environment. Finally, we speculate that these differences may explain, at least in part, the observed differences in the overall divergence rates of *F. prausnitzii* clades in human populations.**

INTRODUCTION

Faecalibacterium prausnitzii is one of the most wide-spread and abundant bacteria in the human gut microbiome (GM). It is probably an integral component of our evolutionary history which has populated our lineage for at least 1M years.[1] *F. prausnitzii* has been consistently reported as one of the main health-promoting components found in the intestine,[2] showing a crucial role in host nutrition and immunity, where it acts as an important anti-inflammatory commensal.[3] Indeed, recent studies[4–6] have shown that *F. prausnitzii* can attenuate the severity of inflammation through the release of a panel of anti-inflammatory metabolites, which enhance the intestinal barrier acting on tight junctions, as well as on peroxisome proliferator-activated receptor alpha (PPAR-α), PPAR-γ and PPAR β/δ genes.[7]

Over the last few years an increasing number of studies have reported a depletion of *F. prausnitzii* in GMs associated with multiple diseases, enteric and non-enteric,[8–12] to the point that this bacterium has been proposed as a possible biomarker of dysbiotic shifts. This defines a complex scenario where, on the one hand, *F. prausnitzii* has a crucial role in maintaining gut homeostasis, but on the other hand it is extremely prone to dysbiotic reductions. However, at present, it still remains elusive which biotic and abiotic factors regulate its presence in the gut, the extent of their influence and the mechanisms involved in its retention.

First *16S rRNA* gene-based phylogenetic analyses showed that at least two different *F. prausnitzii* phylogroups can be found in the human GM, whose distribution is different between healthy subjects and patients with gut disorders.[13,14] Most recently, the polytypic nature of *F. prausnitzii* has been confirmed, detecting up to 11 different clades, which show a different prevalence and/or abundance in the human GM depending on age, geographical origin and lifestyle.[15] These authors also confirmed the depletion of this species in inflammatory bowel disease and obesity. Although these findings certainly represent a milestone for a better understanding of *F. prausnitzii* biology in the human gut, there is still no evidence concerning possible selective pressures driving for the observed clades divergences, and it has not yet been investigated why such clades exhibited a markedly different distribution in the human population.

In an attempt to answer these questions, here we explored the dynamics involved in the divergence processes of the clade-specific marker genes in the *F. prausnitzii* complex, dissecting the peculiarities of each clade and providing some glimpses on the putative pressures selectively acting on each of them. Specifically, we reconstructed high-quality *F. prausnitzii* genomes from metagenomes (MAGs) starting from ~750 healthy human gut metagenomes[16–22] and identified *F. prausnitzii* clades by implementing a

[1]Microbiomics Unit, Department of Medical and Surgical Sciences, University of Bologna, Bologna 40138, Italy

[2]Unit of Microbiome Science and Biotechnology, Department of Pharmacy and Biotechnology, University of Bologna, Bologna 40126, Italy

*Correspondence: simone.rampelli@unibo.it

https://doi.org/10.1016/j.isci.2022.105533

previously validated pipeline.[15,23,24] Then, the within-clade genetic diversity have been analyzed, allowing to dissect the putative evolutionary forces acting on each clade. In particular, the divergence dynamics were assessed by accounting for the specific pattern of mutations accumulating in the respective clade-specific genes. Given the high susceptibility of this species to alteration of host homeostasis and environmental stresses, our findings may provide new insights into the determinants responsible for its decrease in disease conditions and help to find solutions for the recovery of this keystone taxon.

RESULTS

De novo identification and functional characterization of 5 *F. prausnitzii* clades

We assembled 92 high-quality *F. prausnitzii* MAGs from 740 human gut metagenomes from a corresponding number of healthy subjects from 7 different studies, representing 8 different populations (Germans, Italians, Swedes, North Americans, Japanese, Peruvians and Tanzanian hunter-gatherers) (Figure S1). The obtained MAGs showed >95% completeness and <5% contamination levels.[25] These 92 MAGs were complemented with 55 *F. prausnitzii* genomes directly downloaded from the NCBI RefSeq database (https://www.ncbi.nlm.nih.gov/refseq) (Table S1), for a total of 147 genomes used for the subsequent analyses. By computing the average nucleotide identity (ANI) distances, the Jaccard dissimilarity matrix on genes content and the PhyloPhlAn2[26] phylogenetic grouping, we were able to identify 5 clades (A to E), with the largest (clade C) hosting 39 genomes and the smallest (clade E) comprising 12 genomes (Figure 1). By means of alignments, we noted that the 11 clades previously reported by DeFilippis et al.[15] were represented within ours (Table S2). Arguably, the higher completeness threshold we applied for MAGs assembly explains the lower number of clades we were able to identify in our study. When we sought for functional specificities, we observed considerable functional differences between our clades, in terms of presence/absence of specific KEGG Orthology (KO) functionalities (Figure S2) and carbohydrate-active enzymes (CAZymes) (Figure S3). Most of the differences in KO genes concerned broad cellular processes, such as energy metabolism, ABC transporters and dehydrogenases. As regards carbohydrates metabolism, clade A was the most eclectic, bearing the highest fraction of CAZymes, followed by clades D and E. In contrast, clade B seemed to behave as a specialist, possessing a lower amount of CAZymes showing a particularly underrepresented glycoside hydrolase functional potential.

We next assessed the distribution of the 5 clades in the human population (see STAR Methods). According to our data, the 5 clades we identified are distributed across the entire set of human populations considered, thus all the clades can be regarded as cosmopolitan (Figure S4A). To investigate if these clades were mutually exclusive or able to co-inhabit the bowel, we evaluated the co-presence within the same metagenomic sample. This analysis clearly revealed that the degree of co-presence is variable in the human population, with some subjects harboring all the clades, whilst others not harboring *F. prausnitzii* at all. In particular, we observed that the presence/copresence of the *F. prausnitzii* clades was associated with age, geographical origin and subsistence strategy (Figures S4B–S4D), confirming what previously highlighted in another study (De Filippis et al., 2020[15]). Indeed, *F. prausnitzii* was almost always present in adults (96% contained at least 1 *F prausnitzii* clade, 18–69 years old), but the prevalence considerably decrease in infant (29%, <1 years old, Fisher's test $p<0.01$), and centenarians (40%, >99 years old, $p<0.01$). Lower prevalence was also detected in children (89%, 1–16 years old, $p<0.01$) and elderly people (89%, 70–97 years old, $p<0.01$). Finally, the intra-individual clades diversity was highly variable according to the geographical area and the related lifestyle, with higher levels in non-Western countries (e.g., Tanzania, Wilcoxon test $p=0.0001$), respect to Western countries, that showed a progressively lower prevalence for all clades from Europe to Japan through North America.

Construction and validation of divergence indices

To account for the rate of divergence between the *F. prausnitzii* clades, we developed two Divergence Rate Indices (DRIs), one at the clade level and the other at the gene level. The clade-level DRI (DRIc), was specifically conceived to account for the overall divergence rate of each clade and was computed as the natural logarithm (ln) of the ratio between the median number of single nucleotide polymorphisms (SNPs) of the whole set of clade-specific genes (M_E) – defined as genes present in at least the 95% of the genomes from a given clade and absent in the other clades – and the median SNPs for a basal set of housekeeping reference genes (M_H), *i.e.*, genes showing a little divergence within a clade. The housekeeping references was constituted by a panel of 10 genes (*recA*, *rplS*, *rplI*, *purN*, *mreB*, *maf*, *fmt*, *gyrB*, *rpoB*, *proC*) (Table S3), comprising essential genes we found present in all the genomes and MAGs we analyzed. On the other hand, the gene-level DRI (DRIg) was created to account for the absolute divergence rate for a single clade-specific gene for a given clade and was

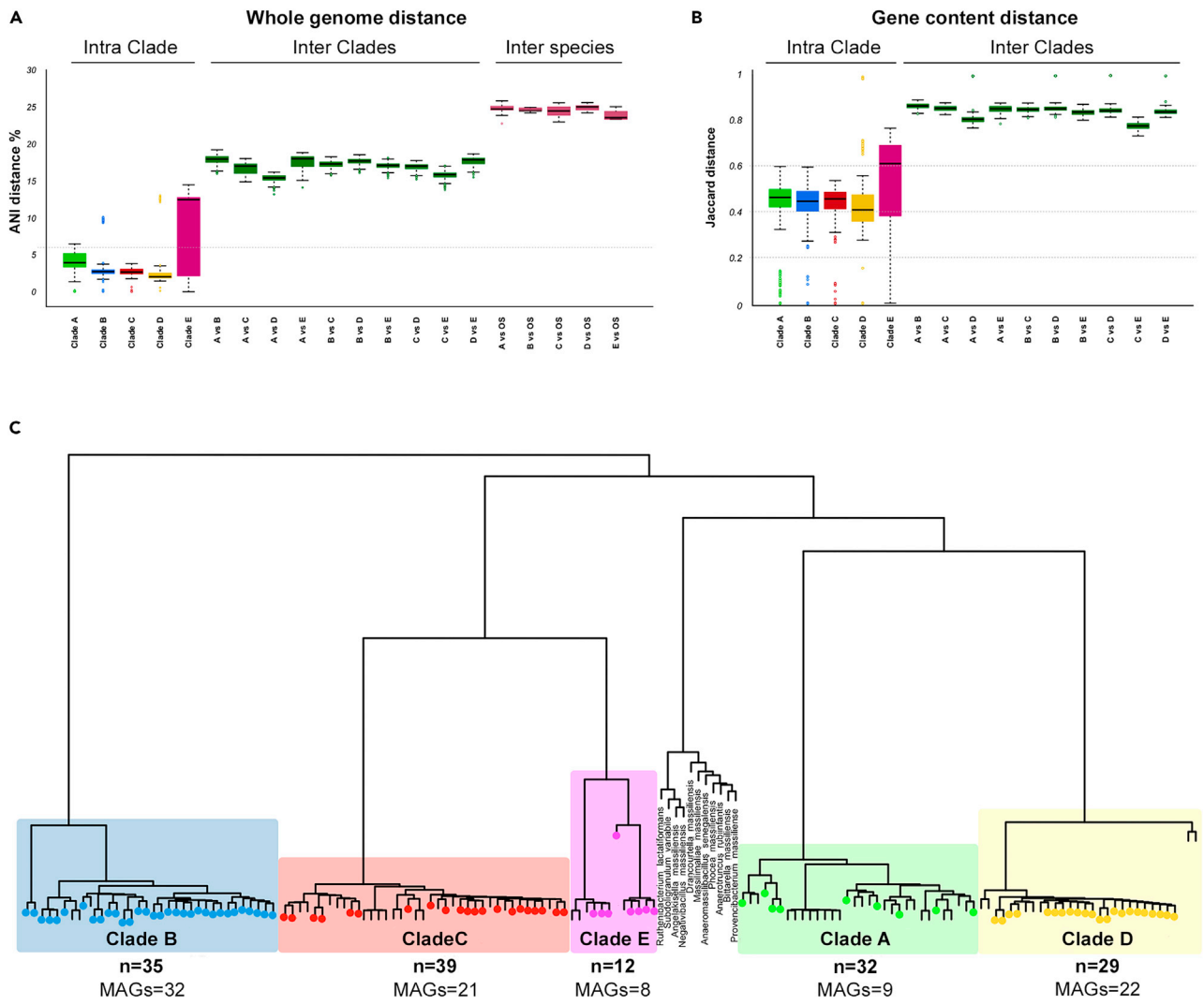


Figure 1. *De novo* identification and functional characterization of 5 *Faecalibacterium prausnitzii* clades

(A) Genetic distances in terms of ANI, within a clade (intra-clade), between clades (inter-clade) and between clades and other species (OS) of the Ruminococcaceae family, used as outgroups (see STAR Methods). Five *F. prausnitzii* clades (A to E) were identified. The dotted line denotes the 6% ANI distance threshold.

(B) Jaccard distance based on gene content between (inter-clade) and within (intra-clade) *F. prausnitzii* clades.

(C) Whole-genome phylogenetic tree derived from PhyloPhlAn2, representing the genome panel (n = 158, of which 92 MAGs, 55 reference genomes and 11 OS) clustered into the 5 identified clades. Colored circles indicate the genomes we assembled from metagenomes (MAGs).

computed as the ln of the ratio between the SNPs of the selected clade-specific gene (M_G) and M_H (the median SNPs of the basal set of housekeeping reference genes).

Further, to fully assess divergence pressures, the aforementioned indices were implemented by considering the ratio of non-synonymous to synonymous substitutions (dN/dS).[27] Consistently, 2 non-synonymous divergence rate indices (NDRIs) were developed: (1) The clade-level NDRI (NDRIc), which considers the ln of the ratio between the mean dN/dS values for the whole set of clade-specific genes (μ_E) and the mean dN/dS for the basal set of housekeeping reference genes (μ_H), (2) the gene-level NDRI (NDRIg), as ln of the ratio between the dN/dS value of the selected clade-specific gene (μ_G) and μ_H.

Generally, for a given clade, a positive value for the DRIc and NDRIc indices points out that the corresponding set of clade-specific genes are accumulating SNPs and non-synonymous SNPs faster than

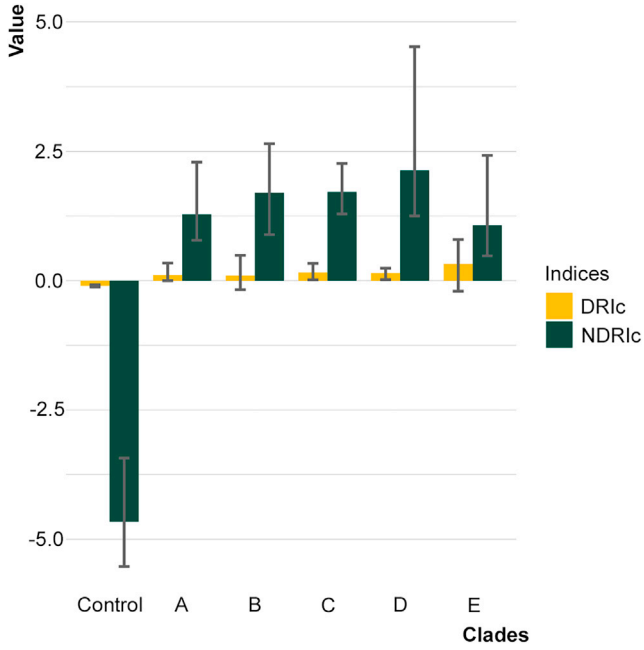


Figure 2. Clade-specific marker genes showed higher divergence indices than core genes

For each metagenomic sample we computed the DRIc and NDRIc values for the 5 *F prausnitzii* clades detected (A to E). Median values among the 740 samples investigated are shown with whiskers ranging from the 25th to the 75th percentiles. Control refers to 500 core genes taken from the pan-genome of the *F. prausnitzii* complex as detected by the ROARY[28] pipeline. For further information concerning DRI and NDRI calculation and marker-genes/core-gene identification, consult the STAR Methods.

housekeeping references; the higher the index values, the greater the divergence rate for the specific clade. Analogously, for a given clade-specific gene, a positive value for the DRIg and NDRIg indices indicates that the gene is accumulating SNPs and non-synonymous SNPs faster than housekeeping references; the higher the index values, the greater the divergence rate for the given clade-specific gene.

Because the DRIg and NDRIg indices were first necessarily computed at the level of the single metagenomes, to be then extrapolated at the population and metapopulation levels, and to verify any bias due to the sequencing yields, for each clade we performed Pearson's correlation tests between M_H and μ_H values and metagenome lengths and the computed *F. prausnitzii* abundances. Correlations were also sought between gene prevalence and DRIg/NDRIg indices, to assess the presence of biases due to sequencing coverage on specific genes. According to our findings, no significant correlations were found ($p>0.05$).

Divergence dynamics: Each clade shows a distinctive profile

Once defined and validated, we utilized our indices to study the divergence of the *F. prausnitzii* complex in the human population. First, we assessed the divergence of the clades in the human population by calculating the global DRIc and NDRIc indices (Figure 2) as the median of all the DRIc and NDRIc indices computed for the single metagenomic samples. For each clade, both global DRIc and NDRIc indices showed positive values, in contrast to the global indices for 500 randomly picked core genes (see STAR Methods), which resulted in negative values. This confirms that clade-specific marker genes are globally accountable for the divergence of the clades; hence, investigating their function may provide new glimpses over the selective pressure driving clades divergence. In particular, clade D showed the highest NDRIc values - with relatively high values for the DRIc index as well – resulting in the most rapidly diverging clade in the human population.

Next, to highlight for each clade the most diverging clade-specific genes, the clade-specific patterns of DRIg and NDRIg indices were computed (Figure 3 and Table S4). For each clade, gene-level divergence indices showed positive values for a multitude of clade-specific genes, indicating an overwhelming

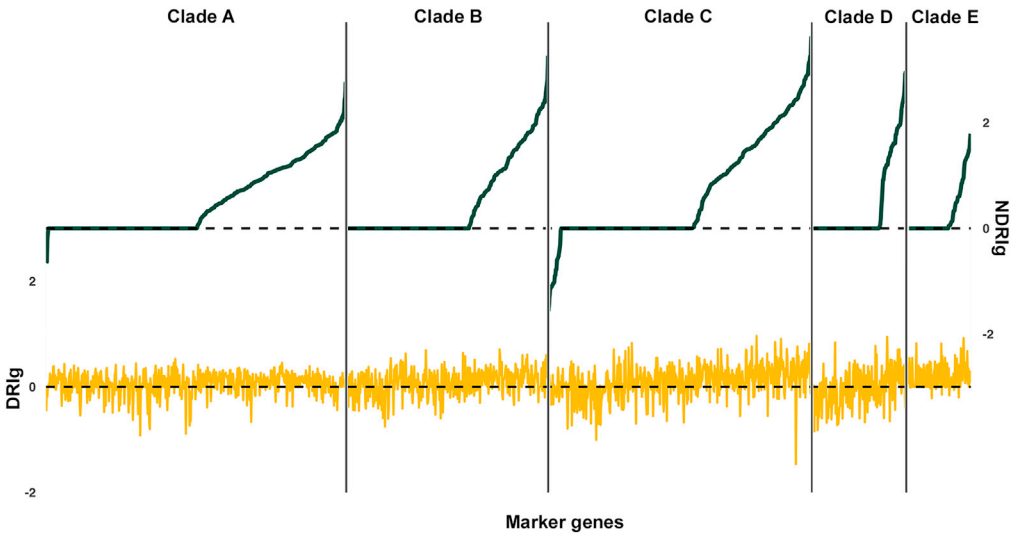


Figure 3. Clade-specific marker genes show different values of DRIg and NDRIg indices

Curves represent the median values of the DRIg (yellow, bottom) and NDRIg (green, top) indices across the 740 metagenomic samples for each clade-specific marker gene. Genes are in increasing order with respect to the NDRIg values. See also Table S4 for the DRIg and NDRIg values for each clade-specific marker gene and Table S7 for the number of marker genes for each clade.

divergence rate in the human population that far exceeds that characteristic of housekeeping genes, as representative of a basal divergence.

For each clade, marker genes were then filtered, keeping only those with both global DRIg and NDRIg positive values. We interpreted the combination of higher mutation rates and more impactful mutations as a signature of active divergence of those regions, therefore investigating the function of such sequences may provide new glimpses over the selective pressures acting globally on *F. prausnitzii*.

Clade-specific marker genes show genetic signatures of purifying selection and selective sweeps

To confirm that clade-specific marker genes are evolving under a non-neutral process, we added Tajima's D[29] to our approach. This parameter allows one to identify sequences that do not fit the neutral theory model at equilibrium between mutation and genetic drift. Computing Tajima's D for *F. prausnitzii* on synonymous sites, to reduce the effects of selection, we observed negative values for all 5 clades (mean −1.6), with clade A showing the lowest value (−2.1) and clade D the highest (−1.3). Looking at the single gene contributions, we found that clade-specific marker genes contributed more to the negative values than core genes, indicating strong level of purifying selection with an excess of rare polymorphisms (Figure 4). Also, together with the evidence from our indices, these estimates suggest that the higher values of the dN/dS ratio of the marker genes are probably caused by recent mutations, capturing a current selection still in progress, acting immediately after or in a context of selective sweeps.

Different clades show different functions of the clade-specific marker genes under divergence pressure

To investigate the function of clade-specific marker genes filtered according to the combination of DRIg and NDRg indices, KEGG Orthology[30] was used, allowing one to take into account the possible functional redundancy among the different markers. Thus, for each clade, we were able to obtain a profile of KOs corresponding to the most diverging clade-specific genes, *i.e.*, those showing positive DRIg and NDRIg values. As expected, several KOs were specific to each single clade, whilst others were shared by two to four clades. No common functions to all clades were identified (Figure 5 and Table S5).

In particular, clade A showed 64 distinctive KOs, including many genes related to sporulation, DNA repair, microbial resistance mechanisms (*e.g.*, antibiotic biosynthesis, xenobiotic degradation, CRISPR proteins,

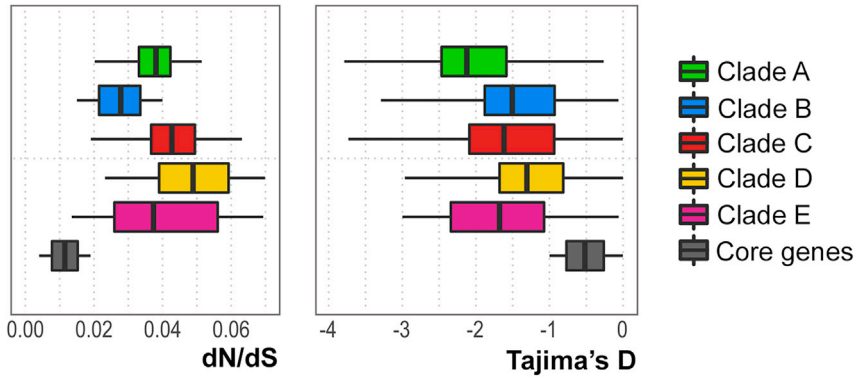


Figure 4. Clade-specific marker genes show genetic signatures of purifying selection and selective sweeps

dN/dS and Tajima's D estimates were computed on clade-specific marker genes and 500 core genes, the same defined in Figure 2. The marker genes showed higher level of purifying selection estimated through the ratio of non-synonymous to synonymous nucleotide substitutions (dN/dS) and Tajima's D values respect to core genes.

CAMP-resistance) and several transporters and transcription factors. As for clade B, we identified 43 unique KOs, mainly concerning the two-component system, antibiotic resistance genes, membrane transporters, as well as DNA repair and one carbon pool by folate. Clade C presented 39 selective KOs involved in DNA repair, sporulation, antimicrobial resistance, beta-lactam resistance, xenobiotic degradation, as well as several efflux proteins, transcription factors, genes involved in tRNA biogenesis, ribosome biogenesis and aminoacyl-tRNA biosynthesis. In addition to these functions, Clade C was the only clade that showed the anti-inflammatory MAM (microbiota anti-inflammatory molecule) protein within the filtered marker genes. Clade D and clade E exhibited 11 specific KOs, with the first particularly enriched in inorganic ion transporters and functions related to amino acids metabolism and transport, and the second in carbohydrate and lipid transporters (Table S6).

Finally, for each clade, we explored the variation in clade-level divergence rates in different human populations. According to our findings, all clades, with the exception of clade C, showed a heterogeneous pattern of DRIc and NDRIc in the human populations considered (Figure 6). In particular, quite opposite trends were found for clades A and D, with the former showing the highest divergence rates in hunter-gatherers and rural communities, and the latter diverging most actively in industrial urban populations.

DISCUSSION

Starting from previous evidences[12,14,15] that *F. prausnitzii* is a polytypic species, we performed a *de novo* clade identification process and then took a step forward to gauge possible determinants of clades divergence. In particular, by analyzing a panel of 92 *F prausnitzii* MAGs assembled from 740 human metagenomes and 55 available genomes from NCBI, we were able to define 5 distinct clades of the *F. prausnitzii* complex, on which we based our further research. Four divergence rate indices (DRIc, NDRIc, DRIg and NDRIg) were constructed and validated, which, combined with Tajima's D estimation, allowed for a curated assessment of the non-neutral divergence rate of each clade down to the gene level. Of interest, the exploitation of gene-level indices to identify the most rapidly diverging clade-specific marker genes allowed us to dissect the signatures of the possible selective pressures acting over these clades. In particular, for the clades A, B and C, the most rapidly diverging genes corresponded to functionalities that may allow to better cope with environmental changes, as well as to increase the inter-host dispersion capacity. Indeed, clade A was found to rapidly diverge in genes involved in several stages of the sporulation process, DNA repair and microbial resistance mechanisms, all of which are important factors for a prokaryotic cell to withstand and counteract environmental stresses. Similarly, clade B revealed a propensity to diverge functionalities related to the two-component system, mRNA expression regulation, aminoacyl-tRNA biosynthesis, transporters and membrane proteins, which may allow for a better metabolic flexibility in response to environmental stimuli. Finally, clade C combined a certain resistance potential, attributable to DNA repair, sporulation and resistance genes, with functional adaptability, as evidenced by several genes encoding transcription factors and transporters, and involved in the expression regulation. As a distinctive feature of clade C, the most rapidly diverging clade-specific genes also included functions related to the modulation of the immune response, such as the MAM protein, which has been shown to exert anti-inflammatory activities primarily via NF-κB pathway inibition.[31] In contrast, clade D and E showed a different pattern of

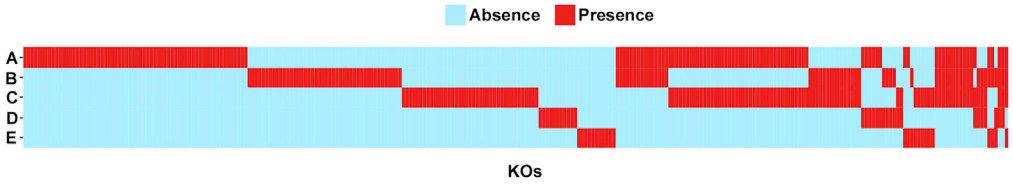


Figure 5. Different clades show different functions of the clade-specific marker genes under divergence pressure
Five hundred and fifty-five clade-specific marker genes with NDRIg and DRIg >0 were classified in the KEGG database and visualized for their *F. prausnitzii* clade (A to E)-specific presence. Red, presence; cyan, absence. See also Table S5 for the complete list of KOs under divergence pressure for each clade.

diverging functionalities, probably related to the exploitation of a copiotrophic gut environment, as indicated by the distinctive presence of genes coding for amino acid transport and metabolism, as well as carbohydrates and lipids transporters.

In our study, we found differences in the clade-level divergence rates between different human populations. In particular, clade A is diverging faster in hunter-gathering and rural populations, whereas clade D showed an opposite trend. Taken together, these observations might suggest that *F. prausnitzii* clades – or at least some of them - are evolving characteristic functional specializations that are better suited to the context of a specific host subsistence strategy which, in turn, would favor a more rapid divergence rate. For instance, clade A – which is evolving functionalities to survive outside host – showed a better fit in traditional populations, where inter-host dispersion of GM components is still an active process as it is not compromised by the sanitization practices typical of Western populations.[32,33] Conversely, clade D, which is evolving adaptations for efficient exploitation of different substrates within the gut environment, showed a better fit and faster adaptive evolution in industrial urban populations, who are well known to consume high-fat/high-protein diets enriched in simple carbohydrates.[34] Future studies including the isolation and cultivation of different *F. prausnitzii* strains representing each clade should be crucial to better identify the specific selective pressures driving clade differentiation.

Overall, our findings may provide new insights into the possible factors driving to the differentiation of the observed subspecies groups in the *F. prausnitzii* taxon. This information may be helpful for better understanding the evolutionary propensity of this health-promoting GM component allowing, from our side, to adopt sustainable dietary and lifestyle practices to favor its retention in the human gut. This is particularly important for industrial urban populations, where a decrease in *F. prausnitzii* diversity and prevalence has been observed.[15] Possibly, the excess of sanitization practices typical of these populations is just facilitating the reduction of the *F. prausnitzii* clades A-C, which are evolving for better outside-host survival as a strategic factor for increasing their colonization of the human population.

Finally, the procedure we developed and implemented in this work can be virtually applied to every polytypic species of bacteria and, assuming the use of a sufficient number of genomes and metagenomes, could provide new ecological insights over the evolutionary forces shaping the microbial world around and within us.

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
- METHOD DETAILS
  - Constructing a *F. prausnitzii* genome panel with additional curated genomes from metagenomes
  - Metagenomic assembly to MAGs
  - Average nucleotide and genetic distances within the *F. prausnitzii* complex and between the complex and related species
  - Phylogenetic analysis of the *F. prausnitzii* genomes included in the genome panel

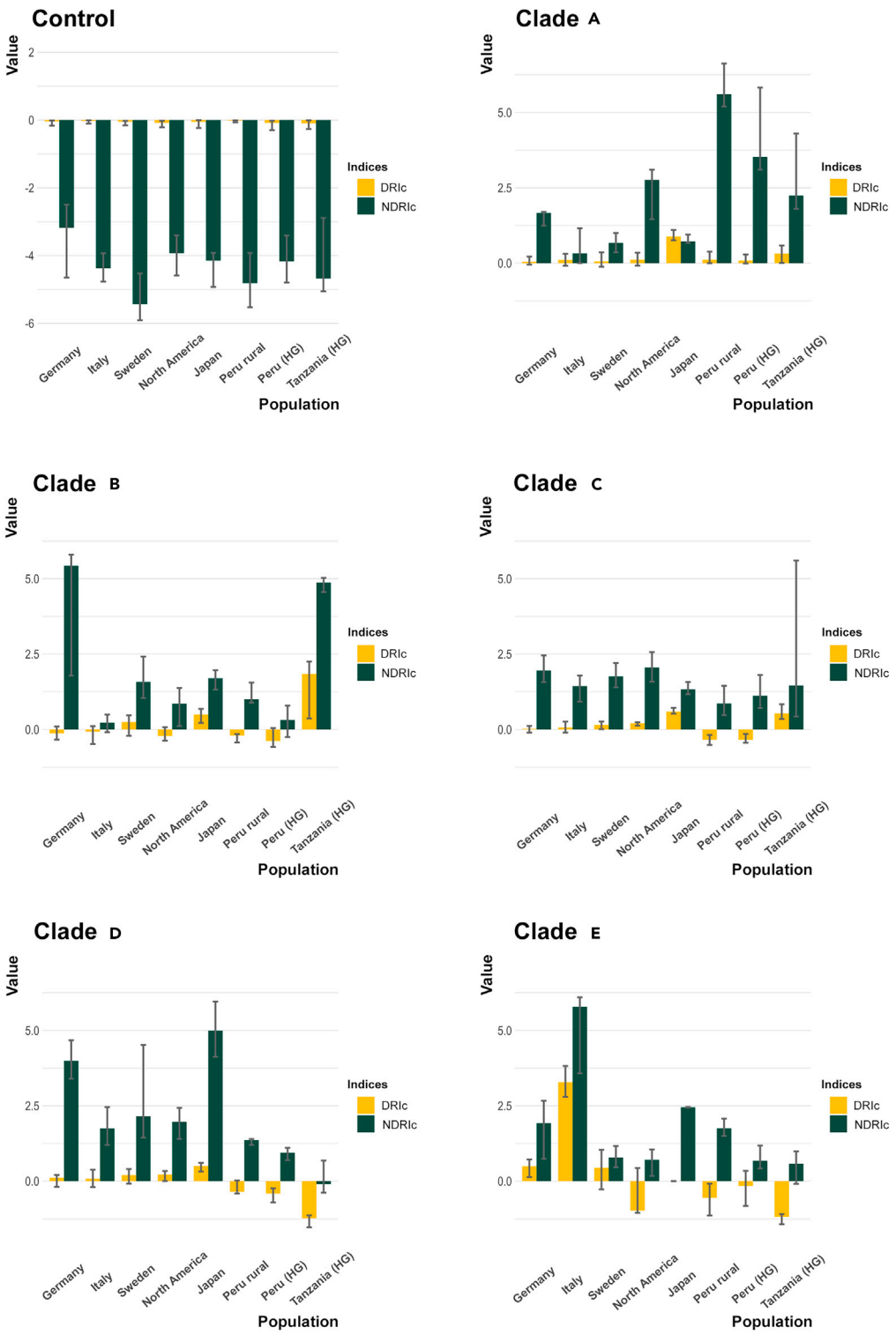


Figure 6. Clade-level indices show differences among the considered human populations

DRIc and NDRIc indices were computed at the single population level as the median of the individual DRIc and NDRIc values among the subjects belonging to that population. For each clade (A–E), divergence rates showed sign of active divergence compared to the housekeeping Control. Clade C resulted particularly consistent across all populations, whilst clade A and D showed opposite trends, being respectively highly divergent in rural

Figure 6. *Continued*
communities and industrial urban populations. The 25th and 75th percentiles are shown with whiskers. Control refers to the 500 core genes as in Figure 2. The following populations were considered: industrial urbans from Germany, Italy, Sweden, North America and Japan, rural inhabitants from Peru, and hunter-gatherers (HG) from Peru and Tanzania.

- Identification of clade-specific marker genes and abundance analysis
- Functional annotation
- SNP calling procedure and estimation of dN/dS and Tajima's D values in metagenomic samples
- Implementation of divergence rate indices (DRIs) and Non-synonymous divergence rate indices (NDRIs)

- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105533.

ACKNOWLEDGMENTS

This work was supported by the "Controlling Microbiomes Circulations for Better Food Systems" (CIRCLES) project, funded by the European Union's Horizon 2020 research and innovation program under grant agreement no. 818290.

AUTHOR CONTRIBUTIONS

Conceptualization, M.C. and S.R.; Methodology, M.F. and S.R.; Formal Analysis, M.F.; Investigation, M.F.; Writing - Original Draft, M.F., M.C. and S.R., Writing - Review & Editing, M.C., S.T. and S.R.; Visualization, M.F. and S.R., Supervision M.C., P.B. and S.R.; Funding Acquisition, M.C. and P.B.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 20, 2021
Revised: March 8, 2022
Accepted: November 4, 2022
Published: December 22, 2022

REFERENCES

1. Rampelli, S., Turroni, S., Mallol, C., Hernandez, C., Galván, B., Sistiaga, A., Biagi, E., Astolfi, A., Brigidi, P., Benazzi, S., et al. (2021). Components of a Neanderthal gut microbiome recovered from fecal sediments from El Salt. Commun. Biol. *4*, 169. https://doi.org/10.1038/s42003-021-01689-y.

2. Parsaei, M., Sarafraz, N., Moaddab, S.Y., and Ebrahimzadeh Leylabadlo, H. (2021). The importance of Faecalibacterium prausnitzii in human health and diseases. New Microbes New Infect. *43*, 100928. https://doi.org/10.1016/J.NMNI.2021.100928.

3. Breyner, N.M., Michon, C., de Sousa, C.S., Vilas Boas, P.B., Chain, F., Azevedo, V.A., Langella, P., and Chatel, J.M. (2017). Microbial anti-inflammatory molecule (MAM) from Faecalibacterium prausnitzii shows a protective effect on DNBS and DSS-induced colitis model in mice through inhibition of NF-κB pathway. Front. Microbiol. *8*, 114. https://doi.org/10.3389/FMICB.2017.00114.

4. Carlsson, A.H., Yakymenko, O., Olivier, I., Håkansson, F., Postma, E., Keita, Å.V., and Söderholm, J.D. (2013). Faecalibacterium prausnitzii supernatant improves intestinal barrier function in mice DSS colitis. Scand. J. Gastroenterol. *48*, 1136–1144. https://doi.org/10.3109/00365521.2013.828773.

5. Martín, R., Miquel, S., Chain, F., Natividad, J.M., Jury, J., Lu, J., Sokol, H., Theodorou, V., Bercik, P., Verdu, E.F., et al. (2015). Faecalibacterium prausnitzii prevents physiological damages in a chronic low-grade inflammation murine model. BMC Microbiol. *15*, 67. https://doi.org/10.1186/s12866-015-0400-1.

6. Lenoir, M., Martín, R., Torres-Maravilla, E., Chadi, S., González-Dávila, P., Sokol, H., Langella, P., Chain, F., and Bermúdez-Humarán, L.G. (2020). Butyrate mediates anti-inflammatory effects of Faecalibacterium prausnitzii in intestinal epithelial cells through Dact3. Gut Microbes *12*, 1–16. https://doi.org/10.1080/19490976.2020.1826748.

7. Moosavi, S.M., Akhavan Sepahi, A., Mousavi, S.F., Vaziri, F., and Siadat, S.D. (2020). The effect of Faecalibacterium prausnitzii and its extracellular vesicles on the permeability of intestinal epithelial cells and expression of PPARs and ANGPTL4 in the Caco-2 cell culture model. J. Diabetes Metab. Disord. *19*, 1061–1069. https://doi.org/10.1007/s40200-020-00605-1.

8. Machiels, K., Joossens, M., Sabino, J., De Preter, V., Arijs, I., Eeckhaut, V., Ballet, V., Claes, K., Van Immerseel, F., Verbeke, K., et al. (2014). A decrease of the butyrate-producing species roseburia hominis and faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis. Gut *63*, 1275–1283. https://doi.org/10.1136/gutjnl-2013-304833.

9. Duvallet, C., Gibbons, S.M., Gurry, T., Irizarry, R.A., and Alm, E.J. (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat. Commun. *8*, 1784. https://doi.org/10.1038/s41467-017-01973-8.

10. Demirci, M., Tokman, H.B., Uysal, H.K., Demiryas, S., Karakullukcu, A., Saribas, S., Cokugras, H., and Kocazeybek, B.S. (2019). Reduced Akkermansia muciniphila and Faecalibacterium prausnitzii levels in the gut microbiota of children with allergic asthma.

Allergol. Immunopathol. *47*, 365–371. https://doi.org/10.1016/j.aller.2018.12.009.

11. Gurung, M., Li, Z., You, H., Rodrigues, R., Jump, D.B., Morgun, A., and Shulzhenko, N. (2020). Role of gut microbiota in type 2 diabetes pathophysiology. EBioMedicine *51*, 102590. https://doi.org/10.1016/J.EBIOM.2019.11.051.

12. Zhao, H., Xu, H., Chen, S., He, J., Zhou, Y., and Nie, Y. (2021). Systematic review and meta-analysis of the role of Faecalibacterium prausnitzii alteration in inflammatory bowel disease. J. Gastroenterol. Hepatol. *36*, 320–328. https://doi.org/10.1111/jgh.15222.

13. 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. https://doi.org/10.1128/AEM.02006-15.

14. 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). 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. https://doi.org/10.1097/MIB.0000000000000590.

15. De Filippis, F., Pasolli, E., and Ercolini, D. (2020). Newly explored Faecalibacterium diversity is connected to age, lifestyle, geography, and disease. Curr. Biol. *30*, 4932–4943.e4. https://doi.org/10.1016/j.cub.2020.09.063.

16. Asnicar, F., Manara, S., Zolfo, M., Truong, D.T., Scholz, M., Armanini, F., Ferretti, P., Gorfer, V., Pedrotti, A., Tett, A., et al. (2017). Studying vertical microbiome transmission from mothers to infants by strain-level metagenomic profiling. mSystems *2*. e00164001644-16. https://doi.org/10.1128/msystems.00164-16.

17. 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. https://doi.org/10.15252/msb.20177589.

18. Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., et al. (2015). Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe *17*, 690–703. https://doi.org/10.1016/j.chom.2015.04.004.

19. Biagi, E., Franceschi, C., Rampelli, S., Severgnini, M., Ostan, R., Turroni, S., Consolandi, C., Quercia, S., Scurti, M., Monti, D., et al. (2016). Gut microbiota and extreme longevity. Curr. Biol. *26*, 1480–1485. https://doi.org/10.1016/j.cub.2016.04.016.

20. Nishijima, S., Suda, W., Oshima, K., Kim, S.W., Hirose, Y., Morita, H., and Hattori, M. (2016). The gut microbiome of healthy Japanese and its microbial and functional uniqueness. DNA Res. *23*, 125–133. https://doi.org/10.1093/dnares/dsw002.

21. Obregon-Tito, A.J., Tito, R.Y., Metcalf, J., Sankaranarayanan, K., Clemente, J.C., Ursell, L.K., Zech Xu, Z., Van Treuren, W., Knight, R., Gaffney, P.M., et al. (2015). Subsistence strategies in traditional societies distinguish gut microbiomes. Nat. Commun. *6*, 6505. https://doi.org/10.1038/ncomms7505.

22. Rampelli, S., Schnorr, S.L., Consolandi, C., Turroni, S., Severgnini, M., Peano, C., Brigidi, P., Crittenden, A.N., Henry, A.G., and Candela, M. (2015). Metagenome sequencing of the Hadza hunter-gatherer gut microbiota. Curr. Biol. *25*, 1682–1693. https://doi.org/10.1016/j.cub.2015.04.055.

23. 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. https://doi.org/10.1186/s13059-019-1923-9.

24. 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. https://doi.org/10.1016/j.cell.2019.01.001.

25. Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. *35*, 725–731. https://doi.org/10.1038/nbt.3893.

26. Segata, N., Börnigen, D., Morgan, X.C., and Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat. Commun. *4*, 2304. https://doi.org/10.1038/ncomms3304.

27. Zhao, S., Lieberman, T.D., Poyet, M., Kauffman, K.M., Gibbons, S.M., Groussin, M., Xavier, R.J., and Alm, E.J. (2019). Adaptive evolution within gut microbiomes of healthy people article adaptive evolution within gut microbiomes of healthy people. Cell Host Microbe *25*, 656–667.e8. https://doi.org/10.1016/j.chom.2019.03.007.

28. 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. https://doi.org/10.1093/bioinformatics/btv421.

29. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595.

30. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.

31. 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. https://doi.org/10.1136/GUTJNL-2014-307649.

32. Martínez, I., Stegen, J.C., Maldonado-Gómez, M.X., Eren, A.M., Siba, P.M., Greenhill, A.R., and Walter, J. (2015). The gut microbiota of rural Papua New Guineans: composition, diversity patterns, and ecological processes. Cell Rep. *11*, 527–538. https://doi.org/10.1016/j.celrep.2015.03.049.

33. Ayeni, F.A., Biagi, E., Rampelli, S., Fiori, J., Soverini, M., Audu, H.J., Cristino, S., Caporali, L., Schnorr, S.L., Carelli, V., et al. (2018). Infant and adult gut microbiome and metabolome in rural Bassa and urban settlers from Nigeria. Cell Rep. *23*, 3056–3067. https://doi.org/10.1016/j.celrep.2018.05.018.

34. Sonnenburg, J.L., and Sonnenburg, E.D. (2019). Vulnerability of the industrialized microbiota. Science *366*, eaaw9255. https://doi.org/10.1126/SCIENCE.AAW9255.

35. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data.

36. McIver, L.J., Abu-Ali, G., Franzosa, E.A., Schwager, R., Morgan, X.C., Waldron, L., Segata, N., and Huttenhower, C. (2018). BioBakery: a meta'omic analysis environment. Bioinformatics *34*, 1235–1237. https://doi.org/10.1093/bioinformatics/btx754.

37. Uritskiy, G.V., DiRuggiero, J., and Taylor, J. (2018). MetaWRAP - a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome *6*, 158. https://doi.org/10.1186/s40168-018-0541-1.

38. 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. https://doi.org/10.1038/nmeth.3589.

39. 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. https://doi.org/10.1093/bioinformatics/btv033.

40. 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. https://doi.org/10.7717/peerj.1165.

41. Wu, Y.W., Tang, Y.H., Tringe, S.G., Simmons, B.A., and Singer, S.W. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome *2*, 26. https://doi.org/10.1186/2049-2618-2-26.

42. 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. https://doi.org/10.1101/GR.186072.114.

43. 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. https://doi.org/10.1038/s41467-020-16366-7.

44. Pritchard, L., Glover, R.H., Humphris, S., Elphinstone, J.G., and Toth, I.K. (2016). Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Anal. Methods *8*, 12–24. https://doi.org/10.1039/c5ay02550h.

45. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

46. Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. Methods Mol. Biol. *1079*, 155–170. https://doi.org/10.1007/978-1-62703-646-7_10.

47. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60. https://doi.org/10.1038/nmeth.3176.

48. 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.

49. 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.

50. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One *5*, e9490. https://doi.org/10.1371/journal.pone.0009490.

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

52. Jensen, L.J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., and Bork, P. (2008). eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res. *36*, 250–254. https://doi.org/10.1093/nar/gkm796.

53. Eddy, S.R. (2011). Accelerated profile HMM searches. PLoS Comput. Biol. *7*, 1002195. https://doi.org/10.1371/journal.pcbi.1002195.

54. 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. https://doi.org/10.1016/S0022-2836(05)80360-2.

55. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

56. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, et al.. (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

57. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509.

58. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

59. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158. https://doi.org/10.1093/BIOINFORMATICS/BTR330.

60. Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. Trends Genet. *16*, 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2.

61. Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680. https://doi.org/10.1093/nar/22.22.4673.

62. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., Mcgettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and clustal X version 2.0. Bioinformatics *23*, 2947–2948. https://doi.org/10.1093/bioinformatics/btm404.

63. Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. *34*, W609–W612. https://doi.org/10.1093/nar/gkl315.

64. Yang, Z. (1997). Paml: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. *13*, 555–556. https://doi.org/10.1093/bioinformatics/13.5.555.

65. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591. https://doi.org/10.1093/molbev/msm088.

66. Leinonen, R., Sugawara, H., and Shumway, M.; International Nucleotide Sequence Database Collaboration (2011). The sequence read archive. Nucleic Acids Res. *39*, D19–D21. https://doi.org/10.1093/nar/gkq1019.

67. Sayers, E.W., Agarwala, R., Bolton, E.E., Brister, J.R., Canese, K., Clark, K., Connor, R., Fiorini, N., Funk, K., Hefferon, T., et al. (2019). Database resources of the national center for biotechnology information. Nucleic Acids Res. *47*, D23–D28. https://doi.org/10.1093/nar/gky1069.

68. Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., et al. (2020). Vegan: Community Ecology Package.

69. R Core Team (2020). R: A Language and Environment for Statistical Computing.

70. Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinf. *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

71. 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. https://doi.org/10.1016/j.chom.2019.08.018.

72. Scholz, M., Ward, D.V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., Truong, D.T., Tett, A., Morrow, A.L., and Segata, N. (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat. Methods *13*, 435–438. https://doi.org/10.1038/nmeth.3802.

73. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. *42*, D490–D495. https://doi.org/10.1093/nar/gkt1178.

74. Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., and Xu, Y. (2012). dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. *40*, W445–W451. https://doi.org/10.1093/NAR/GKS479.

75. Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., et al. (2020). gplots: Various R Programming Tools for Plotting Data.

76. Benjamini, Y. (2010). Discovering the false discovery rate. J. R. Stat. Soc. Ser. B Stat. Methodol. *72*, 405–416. https://doi.org/10.1111/J.1467-9868.2010.00746.X.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Faecalibacterium prausnitzii reference genomes	NCBI	Accession numbers reported in Table S1
Human gut metagenomes	Asnicar et al.[16]	Accession numbers reported in Table S1
Human gut metagenomes	Backhed et al.[18]	Accession numbers reported in Table S1
Human gut metagenomes	Biagi et al.[19]	Accession numbers reported in Table S1
Human gut metagenomes	Costea et al.[17]	Accession numbers reported in Table S1
Human gut metagenomes	Nishijima et al.[20]	Accession numbers reported in Table S1
Human gut metagenomes	Obregon Tito et al.[21]	Accession numbers reported in Table S1
Human gut metagenomes	Rampelli et al.[22]	Accession numbers reported in Table S1
Rumiococcaceae reference genomes (here termed "Other Species - OS")	NCBI	NCBI: PRJNA224116
Software and algorithms		
SRA toolkit 2.8.0	Leinonen, Sugawara and Shumway, 2011	https://github.com/ncbi/sra-tools
FastQC 0.11.8	Andrews,[35]	http://www.bioinformatics.babraham.ac.uk/projects/fastqc
KneadData 0.7.2	McIver et al.[36]	https://github.com/biobakery/kneaddata
MetaWRAP 1.0.2	Uritskiy et al.[37]	https://github.com/bxlab/metaWRAP
MetaPhlAn2 2.7.5	Truong et al.[38]	https://github.com/biobakery/MetaPhlAn
MegaHIT 1.1.2	Li et al.[39]	https://github.com/voutcn/megahit
MetaBAT2 2.12.1	Kang et al.[40]	https://bitbucket.org/berkeleylab/metabat
MaxBin2 2.2.5	Wu et al.[41]	https://sourceforge.net/projects/maxbin2/
CheckM 1.0.7	Parks et al.[42]	https://github.com/Ecogenomics/CheckM/wiki
PhyloPhlAn3 0.30	Asnicar et al.[43]	https://github.com/biobakery/phylophlan
Pyani 0.2.6	Pritchard et al.[44]	https://pypi.org/project/pyani/
Prokka 1.14.6	Seeman,[45]	https://github.com/tseemann/prokka
ROARY 3.13.0	Page et al.[28]	https://github.com/sanger-pathogens/Roary
PRANK v.170427	Löytynoja,[46]	http://wasabiapp.org/software/prank/
Diamond 0.9.9.110	Buchfink et al.[47]	https://github.com/bbuchfink/diamond
MAFFT 7.310	Standley and Katoh,[48]	https://mafft.cbrc.jp/alignment/server/
trimAl 1.2rev59	Capella-Gutiérrez et al.[49]	http://trimal.cgenomics.org/
FastTree 2.1.10	Price et al.[50]	https://bio.tools/fasttree
RAxML 8.1.15	Stamatakis,[51]	https://cme.h-its.org/exelixis/web/software/raxml/
EggNOG mapper 1.0.3	Jensen et al.[52]	https://github.com/eggnogdb/eggnog-mapper
HMMER 3.1b2	Eddy,[53]	http://hmmer.org/
Blast 2.2.31+	Altschul et al.[54]	https://blast.ncbi.nlm.nih.gov/
Bowtie2 2.3.5	Langmead and Salzberg,[55]	http://bowtie-bio.sourceforge.net/bowtie2
SAMtools 1.9	Li et al., 2009, 2011[56,57]	http://www.htslib.org/
Bcftools 1.9	Danecek et al.2011,2021[58,59]	https://samtools.github.io/bcftools/bcftools.html
Vcftools 0.1.16	Danecek et al., 2011,2021[58,59]	http://vcftools.sourceforge.net/
EMBOSS transeq 6.6.0	Rice et al.[60]	https://www.ebi.ac.uk/Tools/st/emboss_transeq/
ClustalW 2.1	Thompson et al.,[61] Larkin et al.[62]	http://www.clustal.org/clustal2/
PAL2NAL v14	Suyama et al.[63]	https://bio.tools/pal2nal
PAML 4.9j	Yang, 1997, 2007[64,65]	http://abacus.gene.ucl.ac.uk/software/paml.html

RESOURCE AVAILABILITY

Lead contact

Further information and request for resources and reagents should be directed to and will be fulfilled by the lead contact, Simone Rampelli (simone.rampelli@unibo.it).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All human gut metagenomic sequences used in this study are available in public repositories (see Table S1 for accession numbers).
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human metagenomes

Human metagenome datasets used in this study are from 7 previously published studies, are available in public repositories (see Table S1 for accession numbers), and included 747 subjects spanning different countries (North America, Peru, Sweden, Germany, Italy, Tanzania and Japan) and lifestyles (industrial urban populations, hunter-gatherers and rural communities).

METHOD DETAILS

Constructing a *F. prausnitzii* genome panel with additional curated genomes from metagenomes

A panel of 147 *F. prausnitzii* genomes comprising the entire set of available genomes through the NCBI RefSeq Genome repository (55 genomes, https://www.ncbi.nlm.nih.gov/refseq), and 92 manually curated MAGs (see the paragraph below "metagenomic assembly to MAGs") were collected for performing the analysis. Metagenomic samples from reference studies[16–22] were downloaded via Sequences Read Archive (SRA).[66] We included gut microbiome samples from individuals from different geographical regions and lifestyles for taking into consideration different aspects of gut microbiome variation. In particular, considered regions were: North America (urbans), Peru (rural inhabitants and hunter-gatherers), Sweden (urbans), Germany (urbans), Italy (urbans), Tanzania (hunter-gatherers) and Japan (urbans). Sequences were quality-checked with FastQC v.0.11.8[35] and filtered for human reads using KneadData v0.7.2,[36] in case of single-end reads, and the MetaWRAP command "read_qc" (v1.0.2)[37] for paired-end reads. The panel was complemented with further 11 genomes from species of the *Ruminococcaceae* family that are considered as outgroup for clade definition in the subsequent analyses. Accession numbers of the *F. prausnitzii* NCBI genomes, metagenomic samples and OS reference genomes included in the study are provided in Table S1.

Metagenomic assembly to MAGs

To profile the microbial community composition contained in each quality-filtered sample, shotgun metagenomic sequencing data were analysed with MetaPhlAn2.[38] Reads from samples containing at least 1% *F. prausnitzii* were assembled using MegaHIT.[39] The minimum contig length considered for further analyses was set by default to 1kb. MetaBAT 2[40] and MaxBin 2[41] algorithms were used for the binning procedure, followed by quality analysis with CheckM.[42] Only genome bins with >95% bin completeness and <5% bin contamination were retained and taxonomically classified using PhyloPhlAn 3.0[43] (database*SGB.Dec19*) and MetaWRAP with the NCBI nucleotide and taxonomy databases.[67] Ninety-two high-quality MAGs classified at species level for *F. prausnitzii* were included within the genome panel.

Average nucleotide and genetic distances within the *F. prausnitzii* complex and between the complex and related species

The average nucleotide identity (ANI) pairwise distances were computed using pyani (version 0.2.6; option '-m ANIb')[44] for all the *F. prausnitzii* genomes and 11 publicly available reference genomes from other

species of the *Ruminococcaceae* family included in our panel. Percentage identity was converted into a distance measure, and distances scores were filtered to include only the pairwise comparisons where alignment lengths exceeded 500,000 bp.

The pairwise genetic distances between the same genomes compared above were calculated using a pipeline that included Prokka,[45] ROARY[28] and the package "vegan" of the R software.[68,69] In brief, each genome was first analysed by Prokka with the '–fast' flag, to identify open reading frames.[70] The core genome alignments were produced utilizing PRANK[46] included within the ROARY pipeline. For this step we set the minimum percentage identity for gene clustering to 90% and the minimum required presence for defining core genes to 90% of genomes. The pangenome information obtained, comprising a binary table with gene presence/absence, was used for building a genome-based Jaccard dissimilarity pairwise distance matrix in R using the "vegdist" command.[68]

Clades were finally defined by hierarchical Ward-linkage clustering using both distance matrices. Permutational multivariate analysis of variance was used to verify whether the clades were significantly different from each other in terms of ANI and gene contents (*FDR*< 0.001).

Phylogenetic analysis of the *F. prausnitzii* genomes included in the genome panel

A phylogenetic tree was built using the genome panel and PhyloPhlAn 2[26]. The configuration file was customized as by Tett at al.,[71] using Diamond v0.9.9.110[47] for the mapping step, MAFFT v7.310[48] for the multiple sequence alignment, trimAl version 1.2rev59[49] for trimming, FastTree v2.1.10[50] for the first tree generated and RAxML v8.1.15[51] for the final tree. In addition to the customized configuration file, the parameters used were '–diversity low –fast'.

Identification of clade-specific marker genes and abundance analysis

Marker genes for each clade were identified by analysing the *F. prausnitzii* pangenome obtained with the Prokka and ROARY pipelines (see the "average nucleotide and genetic distances within the *F. prausnitzii* complex and between the complex and related species" paragraph above for further information). In particular, we defined as "marker genes" for a given clade, the genes present in at least 95% of the genomes of that specific clade and completely absent in all the others (see Table S7 for the number of marker genes identified for each clade). Nucleotide sequences for each pool of marker genes were used for building clade-specific databases with bowtie2-build.[55] To determine if a given clade was present in a metagenomic sample, the reads were mapped to the clade-specific markers using Bowtie2[55] and then processed to evaluate the marker genes coverage.[72] A marker was scored present if it had $\geq$0.5X coverage and a clade present if at least 50% of its clade-specific markers were hit. Finally, clade relative abundances for each metagenomic sample were calculated as the mean clade marker coverage multiplied by the *F. prausnitzii* genome size (bp) and divided by the metagenome size (bp).

Functional annotation

The functional annotation step was performed using the EggNOG mapper (version 1.0.3)[52] on the protein sequences identified by Prokka with the '-d bact'database option. The KEGG Brite Hierarchy was used to screen the EggNOG annotations. Fisher's exact test with Bonferroni's correction was used to identify significant differences ($p<0.01$) in gene content between clades.

We also sought for differences in the level of CAZymes.[73] Gene sequences were identified with HMMSEARCH[53] against the dbCAN HMMs v6 database,[74] using default parameters and applying post-processing stringency cut-offs as suggested by the authors (if alignment length >80 AA, E-value is filtered for values < 1e-5, otherwise for values < 1e-3; then a cut-off is applied based on the covered fraction of HMM >0.3).[74] Only CAZy families that were significantly different in at least one clade (Bonferroni-corrected Fisher's exact test, $p<0.01$) were retained and graphically represented using the R package "gplots".[75]

Finally, the genes encoding the MAM protein of *F. prausnitzii* were detected by aligning the protein sequence[31] against the full set of genes from the *F. prausnitzii* pangenome using protein-protein BLAST (v2.2.31+).[54] For a complete list of marker genes with annotated function, refer to Table S8.

SNP calling procedure and estimation of dN/dS and Tajima's D values in metagenomic samples

SNP calling procedure was performed for the clade-specific marker genes, 10 selected housekeeping genes (*recA, rplS, rplI, purN, mreB, maf, fmt, gyrB, rpoB, proC*) (Table S3), and 500 randomly selected *F. prausnitzii* core genes as genes present in at least 95% of genomes within our panel. Metagenomic samples showing at least 1% *F. prausnitzii*, ensuing from the previous MetaPhlAn 2[39] analysis, were aligned against the databases with Bowtie2[55] using the '–end-to-end' and '–very-sensitive' parameters and then sorted using SAMtools.[56,57] Candidate SNPs were identified using BCFtools mpileup,[58] with the '–ploidy' parameter set to 1, to extract all the variants in vcf format. VCFutils varFilter was then used to filter the minimum depth to 10 reads and the QUAL score >200. For each position, only one point mutation was considered, and the SNP-per-base values were calculated for each gene, dividing the total number of identified SNPs in a gene sequence by its length (bp).

Consensus sequences retrieved from the metagenome alignment and reference sequences were then translated into proteins using EMBOSS transeq 6.6.0[60] and the proteins were aligned using ClustalW 2.1.[61,62] Protein alignment was converted into codon-aligned PAML alignment using PAL2NAL v14[63] and analyzed using the CODEML program of the PAML phylogenetic analysis package (v4.9j),[64,65] to compute dN/dS. Codon frequencies were set to '3 × 4' and no phylogenetic tree was submitted. The outputs of the pairwise comparison between reconstructed consensus genes from metagenomes and reference genes were considered and filtered for $0.01 < dS < 2$, because values of $dS \leq 0.01$ or ≥ 2 entail unreliable estimate of dN/dS since the sequences are too similar or too divergent.

Tajima's D values were computed with vcftools 0.1.16[59] over each gene sequence starting from previously identified and quality-filtered polymorphisms. Both population genetic parameters (dN/dS and Tajima's D) for the *F. prausnitzii* clades were calculated for the same set of marker genes and 500 core genes used for the SNP calling procedure. The parameters were calculated separately for each gene, then the median values were used to represent the parameters for each specific clade.

Implementation of divergence rate indices (DRIs) and Non-synonymous divergence rate indices (NDRIs)

In this study we introduced Divergence Rate Indices (DRIs) and Non-Synonymous Divergence Rate Indices (NDRIs), as clade- or gene-specific indices to assess sequence divergence.

DRI indices were estimated using the SNP-per-base values previously computed. For each metagenomic sample we calculated the DRI for a specific gene of interest (DRIg), using the number of SNP-per-base detected for that specific gene of interest (M_G), the median number of SNP-per-base detected for 10 housekeeping genes (M_H), and calculating the ln of the ratio between the two values. Analogously, we defined the clade-level DRI (DRIc) by considering the median number of SNP-per-base for the entire set of clade-specific genes (M_E), the median value of SNP-per-base for the set of housekeeping genes (M_H) and calculating the ln of the ratio between the two values.

On the other hand, NDRI indices were estimated using the dN/dS values previously computed. For each metagenomic sample we calculated the NDRI for a specific gene of interest (NDRIg), using the value of dN/dS ratio detected for that specific gene of interest (μ_G), the mean value of dN/dS ratio detected for the 10 housekeeping genes (μ_H), and calculating the ln of the ratio between the two values. Analogously, we defined the clade-level NDRI (NDRIc) by considering the mean value of dN/dS ratio for the entire set of clade-specific genes (μ_E), the mean value of dN/dS ratio for the set of housekeeping genes (μ_H) and calculating the ln of the ratio between the two obtained values.

$$DRIc = ln\frac{M_E}{M_H} \quad DRIg = ln\frac{M_G}{M_H}$$

$$NDRIc = ln\frac{\mu_E}{\mu_H} \quad NDRIg = ln\frac{\mu_G}{\mu_H}$$

Higher values for all indices indicate a higher number of SNPs or non-synonymous substitutions in the specific gene/group of genes compared to housekeeping genes.

When M_H or μ_H values were equal to zero, we substituted the value with the lowest M_H or μ_H detected in the global human population. Furthermore, when both dividend and divisor were equal to zero, we set the indices to zero. These corrections had no effect on our results since we only focused on positive values to determine the divergence.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis was performed using R software v4.0.3. The pairwise genetic distances between the same genomes of the panel obtained via the Prokka/ROARY pipeline were analysed with the R package "vegan". The Jaccard dissimilarity pairwise distance matrix was built using the "vegdist" command. Permutational multivariate analysis of variance was used to verify whether the clades were significantly different from each other in terms of ANI and gene contents (p value corrected for multiple testing applying Benjamini-Hochberg false discovery rate,[76] $FDR < 0.001$). Fisher's exact test with Bonferroni's correction was used to identify significant differences ($p < 0.01$) in gene content and CAZymes counts between clades. Graphical representations were made using the R packages "gplots", "ggplot2".