



OPEN

Prevotella species in the human gut is primarily comprised of *Prevotella copri*, *Prevotella stercorea* and related lineages

Yun Kit Yeoh^{1,2,3}, Yang Sun^{4,5}, Lawrence Yuk Ting Ip^{1,6}, Lan Wang^{4,5}, Francis K. L. Chan^{1,2,6}, Yinglei Miao^{4,5}✉ & Siew C. Ng^{1,2,6}✉

Prevotella species in the human gut microbiome are primarily comprised of *Prevotella copri*, and its diversity and function were recently investigated in detail. Much less is known about other *Prevotella* species in the human gut. Here, we examined the composition of *Prevotella* species in human guts by mapping publicly available gut metagenomes to a dereplicated set of metagenome-assembled genomes (MAGs) representing *Prevotella* lineages found in human guts. In most human cohorts, *P. copri* is the most relatively abundant species (e.g. up to 14.3% relative abundance in Tangshan, China). However, more than half of the metagenome reads in several cohorts mapped to *Prevotella* MAGs representing *P. stercorea* and several other species sister to *P. stercorea* and *P. copri*. Analyses of genes encoded in these genomes indicated that *P. stercorea* and related lineages lacked many hemicellulose degrading enzymes and were thus less likely to metabolise hemicelluloses compared with *P. copri* and *copri*-related lineages. Instead, *P. stercorea* genomes possess several carbohydrate esterases that may be involved in releasing ester modifications from carbohydrates to facilitate their degradation. These findings reveal unexplored *Prevotella* diversity in the human gut and indicate possible niche partitions among these related species.

Several *Prevotella* species inhabit the human gut, among which *Prevotella copri* is the most common with estimates indicating ~40% prevalence in the wider human population and relative abundances that exceed 50% in some individuals¹. This species is more prevalent in non-Western populations likely due to its association with high fibre low fat diets^{2–4}, and is often linked with desirable health measures such as reduced visceral fat and improved glucose metabolism^{5,6}. However, there are conflicting reports that implicate *P. copri* in adverse conditions such as insulin resistance⁷, hypertension⁸ and persistent gut inflammation⁹. Recently, Tett and colleagues demonstrated that the existing *P. copri* lineage is comprised of at least four distinct species-level lineages each sharing less than 95% average nucleotide identity with one another¹. Furthermore, the use of polysaccharides among *P. copri* isolates can markedly vary indicating distinct metabolic patterns within this lineage¹⁰. These inconsistencies between traditional taxonomic designations and species boundaries coupled with strain-specific metabolic diversity and dietary preferences could be reasons behind the conflicting observations related to *P. copri* and human health¹¹, and more broadly underscores the need to characterise microbial diversity in the human microbiome to understand its function.

Other than *P. copri*, the human gut hosts several other *Prevotella* species such as *P. stercorea*¹², *P. rectalis*¹³, *P. histicola*¹⁴ and some only known by genomes recovered from human gut metagenomes^{15–18}. Of these, *P. stercorea* is likely the second most prevalent and relatively abundant species in the healthy human gut after *P. copri* based on healthy human gut metagenomes¹⁹ while much less is known about the distribution of other *Prevotella* species. For example, *P. rectalis*¹³ and *P. ihumii*²⁰ were recently isolated from individuals with no apparent disease and to our knowledge have not been implicated in any health conditions. Furthermore, DNA-based gut microbiome

¹Microbiota I-Center, Shatin, Hong Kong SAR, China. ²Centre for Gut Microbiota Research, Faculty of Medicine, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China. ³Department of Microbiology, Faculty of Medicine, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China. ⁴Department of Gastroenterology, The First Affiliated Hospital of Kunming Medical University, Kunming, Yunnan, China. ⁵Yunnan Province Clinical Research Centre for Digestive Diseases, Kunming, Yunnan, China. ⁶Department of Medicine and Therapeutics, Institute of Digestive Disease, Li Ka Shing Institute of Health Sciences, Faculty of Medicine, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China. ✉email: miaoyinglei@yeah.net; siewchiengng@cuhk.edu.hk

analyses often mention unclassified *Prevotella* but fall short of pinpointing exact species^{21,22} likely due to uncharacterised diversity in this lineage.

Since *P. copri* are important members of the human gut microbiome due to their roles in human health as well as putative associations with several adverse conditions, we hypothesised that uncharacterised *Prevotella* species could also play important functions in the human gut as they are likely to share similar traits with *P. copri*. As such, an assessment of overall *Prevotella* diversity and distribution in the human gut will greatly facilitate understanding of their roles in health. Here, we examined the genomes of *Prevotella* species recovered from human gut metagenomes and show that the most prevalent and relatively abundant species are primarily related to *P. copri* and *P. stercorea*, and these species possess distinct gene repertoires likely reflecting adaptations to metabolic niches. Specifically, *P. stercorea* and related species lack several genes for xylan metabolism but have carbohydrate esterases that suggest they work cooperatively with *P. copri* in metabolising dietary fibre.

Methods

Obtaining *Prevotella* genomes from public repositories and gut metagenome data sets. This study was approved by the Joint Chinese University of Hong Kong New Territories East Cluster Clinical Research Ethics Committee (reference number 2016.707 and 2016.407) and was performed in accordance with relevant guidelines, regulations, and the Declaration of Helsinki. Written informed consent was obtained from all participants. Microbial genomes were downloaded from publicly available sources including the Genome Taxonomy Database (GTDB)²³ R202 which contains RefSeq genomes and from several studies of human microbiome diversity^{15,17,18}. In addition, we obtained *Prevotella* genomes from two previous studies examining gut microbiome diversity in Hong Kong²⁴ and Yunnan²⁵. For these two studies, metagenome sequences were quality filtered using Trimmomatic²⁶ v0.39 to remove adapter sequences and low-quality regions. Quality filtered reads were then de novo assembled using MEGAHIT²⁷ v1.2.9. Metagenome-assembled genomes (MAGs) were reconstructed from each metagenome assembly by mapping quality filtered reads to their respective assemblies using BWA MEM²⁸ v0.7.17 and then using these coverage profiles as input into MetaBAT²⁹ v2.10.2, v2.12.1 and MaxBin³⁰ v2.2.7. The resulting MAGs were dereplicated using DAS Tool³¹ and checked for completeness and contamination using CheckM³² (Parks et al., 2015) v1.1.3. Taxonomy was inferred using GTDB-Tk³³ v1.4.1 with the R95 database.

Inferring *Prevotella* phylogeny. In addition to completeness and contamination estimates from CheckM and taxonomies from GTDB-Tk, all MAGs were scanned using Kraken2³⁴ v2.1.1 with the StandardPlusPF precompiled database (v20210517) to assess the fraction of each genome consistent with taxonomy inferred from GTDB-Tk to avoid including chimeric genome assemblies. Genomes that scored > 90% completeness and < 5% contamination based on CheckM, inferred as *Prevotella* by GTDB-Tk and had > 70% of their genomes classified as *Prevotella* by Kraken2 were selected for phylogenetic inference and genome annotation. Identical genomes were dereplicated based on a threshold of 99% average nucleotide identity (ANI) using CoverM v0.6.0 (<https://github.com/wwood/CoverM>). The genomes were also dereplicated at 95% ANI to obtain a set of unique species. Bootstrapped phylogenetic tree (1000 bootstraps) was inferred using IQ-TREE2³⁵ v2.1.2 based on a concatenated alignment of 120 phylogenetically informative single copy bacterial marker genes generated by GTDB-Tk for both the set of 99% and 95% ANI genomes.

Estimating *Prevotella* prevalence and relative abundance in diverse human populations. To estimate the prevalence and abundance of *Prevotella* in the human population, we downloaded publicly available human gut metagenomes (Table S1) from countries including Austria³⁶, Denmark, Spain³⁷, France³⁸, Sweden³⁹, China^{8,40}, Japan⁴¹, El Salvador, Peru⁴², Fiji⁴³, Israel⁴⁴, Mongolia⁴⁵, Tanzania^{46,47} and USA^{48,49}. These data were quality filtered using Trimmomatic v0.39 and then mapped to the species dereplicated set of *Prevotella* genomes (95% ANI) using CoverM v0.6.0 which produces relative abundance estimates of each genome in each cohort based on coverage and number of mapped reads (see <https://github.com/wwood/CoverM/#calculation-methods>). Prevalence was calculated as the number of metagenomes in which *Prevotella* was detected divided by the total number of metagenomes in each cohort. For surveys with case–control cohorts (e.g. colorectal cancer in Japan⁴¹), only non-disease subjects were included in this study.

Annotating genes. We used Panaroo⁵⁰ v1.2.8 to determine gene orthologues shared among the *Prevotella* genomes. Orthologues were assigned Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthologs (KOs) and Carbohydrate active enzymes (CAZy) families by searching against a database of Hidden Markov Models implemented in EnrichM (<https://github.com/geronimp/enrichM>) v0.5.0. Distribution of KOs and CAZy families among the MAGs were visualised using non-metric multidimensional scaling (NMDS) ordinations on Jaccard distances using the Vegan package⁵¹ v2.5.7 in R v4.0.3.

Results

Phylogenetic diversity of human gut associated *Prevotella* species. To assess the extent of *Prevotella* diversity in the human gut, we first downloaded publicly available collections of MAGs derived from human gut metagenomes^{15,17,18}. Additionally, we obtained MAGs from gut metagenome surveys of Hong Kong²⁴ and Yunnan populations²⁵ as we were interested in the distribution of *Prevotella* in these locales. From these data sets, we selected MAGs classified as members of the *Prevotella* genus based on GTDB taxonomy. After quality filtering and dereplication at 99% ANI to remove identical genomes, 3943 MAGs remained of which more than 60% were comprised of five highly represented species including *P. copri* (1579 genomes), *P. sp003447235*

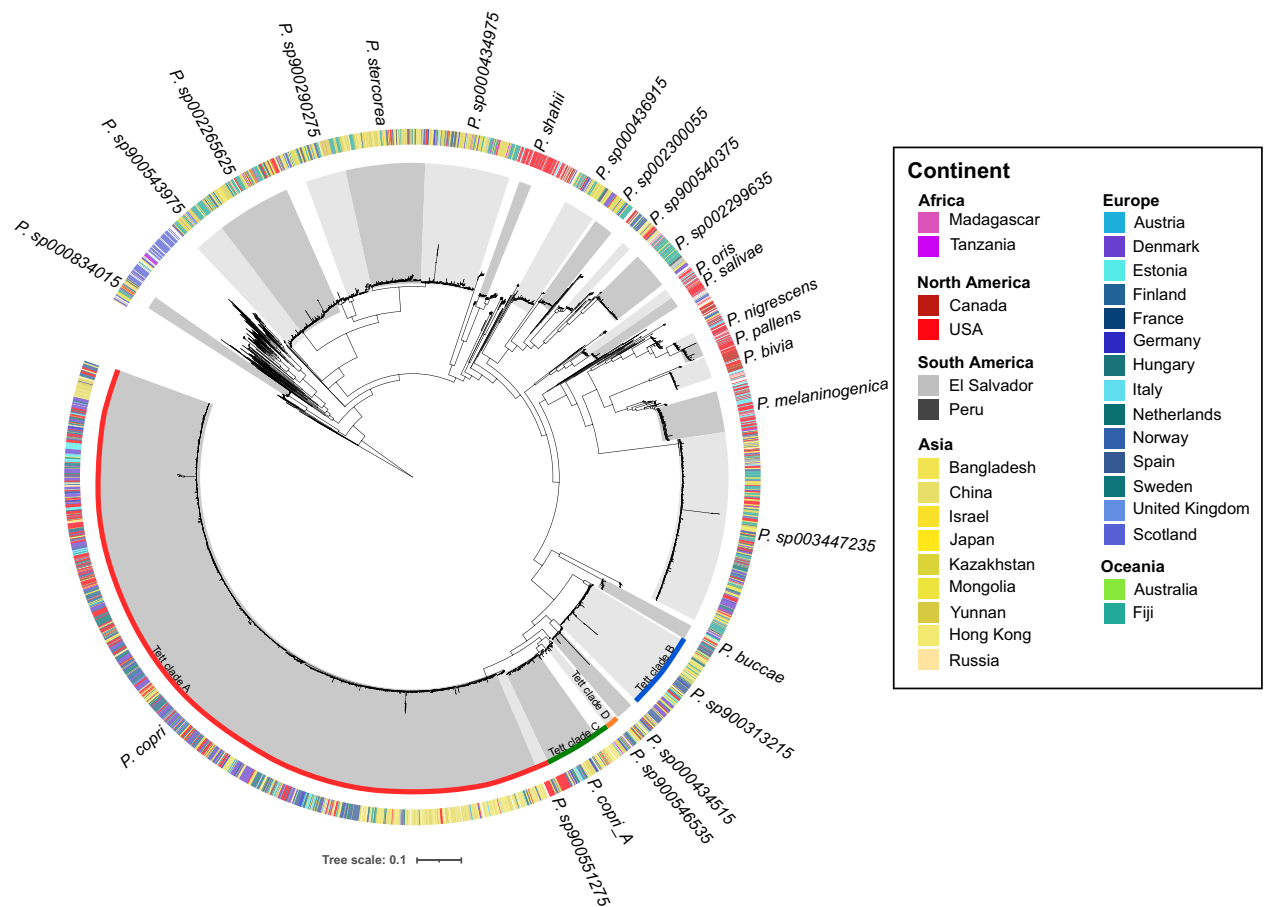


Figure 1. Phylogenetic tree representing phylogenetic relationships among *Prevotella* isolates and metagenome-assembled genomes (MAGs) recovered from human gut metagenomes. Five *Paraprevotella* genomes were used as outgroup to root the tree. MAGs included in this tree were downloaded from publicly available repositories^{15,17,18} and from healthy human gut metagenome surveys of Hong Kong²⁴ and Yunnan populations²⁵. Genomes were dereplicated at 99% average nucleotide identity (ANI) and a concatenated amino acid alignment consisting of 120 phylogenetically informative marker genes (from GTDB) was used to infer phylogenetic trees using IQTREE2. Taxonomy labels from GTDB are shown for species represented by at least 20 genomes (i.e., lower odds of spurious misassemblies or genome binning) (see Table S2 for genome counts); species boundaries are indicated by the alternating grey shades. Note that the GTDB-based taxonomy used here is not comparable to the four *P. copri* clades A–D from Tett et al.¹—these four clades from Tett et al.¹ are indicated separately in the figure. The outer colour strip indicates source country/continent of the respective genomes shaded by continent.

(416 genomes), *P. sp000434975* (185 genomes), *P. stercorea* (174 genomes) and *P. sp002265625* (164 genomes) (Tables S2 and Table S3).

To establish phylogenetic relationships among the 3943 MAGs representing *Prevotella* species detected in the guts of the wider human population, we downloaded an additional 222 NCBI RefSeq *Prevotella* genomes held in the GTDB database and incorporated these with the set of 3943 MAGs. Phylogenetic trees were then inferred based on a concatenated alignment of 120 bacterial single copy marker genes to produce a bootstrapped consensus tree (Fig. 1). Within the lineage consisting of *P. copri* species, MAGs from the Asian continent were phylogenetically closer to one another compared with those from Western geographies (USA, Canada, Europe) possibly reflecting adaptations to local populations. These patterns were also observed in *P. stercorea* in which 128 out of 174 genomes were derived from Asian populations. Conversely, species such as *P. pallens* (24 of 33), *P. salivae* (18 of 23), and *P. bivia* (41 of 47) were largely represented by MAGs from Canada and USA. As for MAGs derived from gut metagenomes representing rural populations (e.g. Tanzania, Fiji, El Salvador), they were primarily classified as several not yet formally characterised species including sp002299635, sp002265625, sp000434975 and sp900543975. These observations indicate that numerous *Prevotella* species are found in human guts (at least 25 when considering those represented by > 20 MAGs for confidence; Table S2), many of which have not been formally characterised especially those from rural populations.

Uncharacterised *Prevotella* species are prevalent and relatively abundant in the human gut. Since many of these *Prevotella* MAGs are directly derived from human gut metagenomes and may not be identified using conventional microbial community profiling tools, we sought to estimate their prevalence

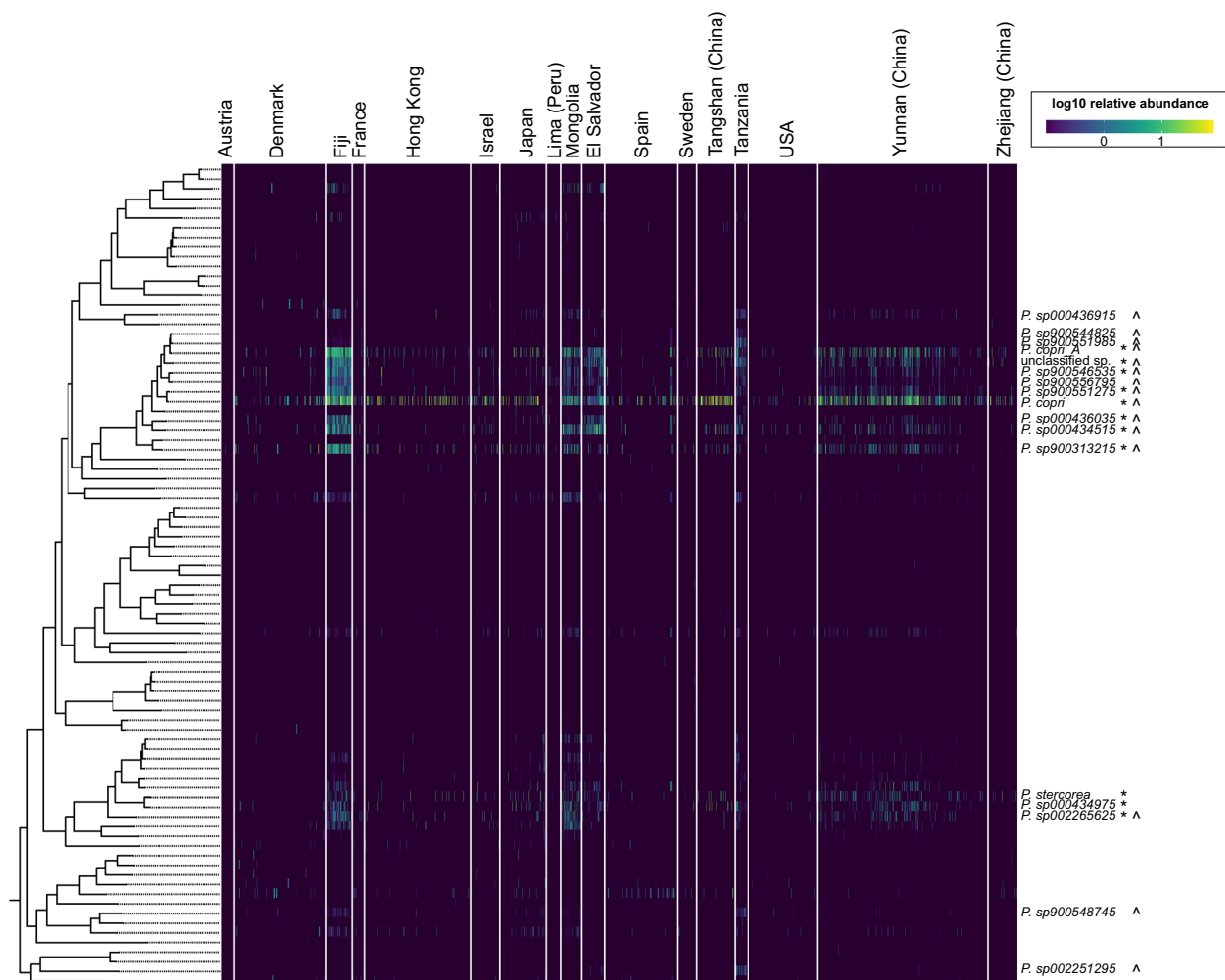


Figure 2. Estimated *Prevotella* relative abundances across several human gut metagenome data sets. Phylogenetic tree indicates relationships of the *Prevotella* genomes dereplicated at the species level (>95% average nucleotide identity). Only genomes with reads mapped are included in this figure. Taxonomy labels are shown for species with at least 1% relative abundance (indicated with *) or at least 50% prevalence (indicated with ^) in any one data set. See Table S4 for relative abundance and prevalence values.

and relative abundances in the human population by mapping metagenome reads. In total, we downloaded 4095 human gut metagenome data sets from non-disease cohorts representing populations from China, Japan, USA, Austria, Denmark, France, Spain, Israel, Sweden, El Salvador, Peru, Fiji, Mongolia, and Tanzania, and mapped quality filtered reads to the set of *Prevotella* MAGs dereplicated at the species level (95% ANI). As expected, total *Prevotella* relative abundance was highest in Fiji, a rural population data set (average 36.1% relative abundance), followed by Tangshan (China; 20.5%) and Mongolia (19.6%) (Fig. 2) (Table 1). They were also more widespread in rural communities such as Fiji (97.8% prevalence), El Salvador (94.7%) and Mongolia (91.8%) compared with urban cohorts in which prevalence ranged between 11 and 58% (Table 1). Among the *Prevotella* species identified above, *P. copri* was the most prevalent (average 30.7%) and relatively abundant overall (average 3.3%) (Table 2, Table S4). Other prevalent and relatively abundant species were primarily from two distinct lineages, one more closely related to *P. copri* and the second to *P. stercorea* (hereafter referred to as copri and stercorea groups) (Fig. 2) (Table S4). Collectively, these species other than *P. copri* account for an average 4.3% relative abundance which is comparable to the 3.2% for *P. copri* (Table S5). Like *P. copri*, most of them are more prevalent and relatively abundant in rural populations such as Fiji, Mongolia and El Salvador, possibly indicating a similar association with non-Western lifestyles and/or high fibre diets. In addition, we detected a few lineages at lower prevalence and relative abundances that are possibly more restricted in their distribution such as *P. sp001275135* and *P. sp003447235* (Table S4). For example, *P. sp001275135* has an average prevalence of only 3.7% across all data sets included in this study, but its prevalence in France and Spain is 14.8% and 17.5%, respectively. Taken together, these observations indicate that the most common *Prevotella* species in the human gut are primarily comprised of *P. copri*, *P. stercorea* and several closely related lineages, although some species with more limited distributions may be prevalent in certain biogeographies.

Data set	Relative abundance \pm standard deviation (%)	Prevalence (%)	Number of metagenomes
Austria	0.09 \pm 0.35	11.11	63
Denmark	2.42 \pm 6.19	41.35	474
Fiji	36.08 \pm 17.64	97.83	138
France	2.26 \pm 5.99	52.46	61
Hong Kong	6.16 \pm 13.21	33.27	547
Israel	5.82 \pm 10.34	58.00	150
Japan	8.15 \pm 14.75	41.95	236
Lima (Peru)	0.37 \pm 1.56	19.48	77
Mongolia	19.61 \pm 16.05	91.82	110
El Salvador	10.24 \pm 9.70	94.74	114
Spain	3.38 \pm 9.69	39.68	378
Sweden	2.58 \pm 9.87	37.00	100
Tangshan	20.48 \pm 24.84	56.63	196
Tanzania	7.36 \pm 8.85	79.10	67
USA	1.12 \pm 4.85	17.60	358
Yunnan rural	8.22 \pm 12.32	69.32	427
Yunnan urban	8.70 \pm 13.67	56.39	454
Zhejiang	4.87 \pm 10.66	37.93	145

Table 1. Distribution of total *Prevotella* in the human gut.

Species (GTDB taxonomy)	Average relative abundance \pm standard deviation (%)	Average prevalence (%)
<i>Prevotella copri</i>	3.27 \pm 3.62	30.73 \pm 26.43
<i>Prevotella copri_A</i>	1.10 \pm 1.59	20.11 \pm 25.05
<i>Prevotella sp000434515</i>	0.59 \pm 0.82	16.89 \pm 25.17
<i>Prevotella sp900313215</i>	0.58 \pm 0.92	14.44 \pm 20.81
Unclassified <i>Prevotella</i>	0.37 \pm 0.61	15.53 \pm 25.16
<i>Prevotella sp000434975</i>	0.28 \pm 0.43	10.22 \pm 14.77
<i>Prevotella stercorea</i>	0.25 \pm 0.29	8.59 \pm 8.51
<i>Prevotella sp900546535</i>	0.25 \pm 0.36	14.18 \pm 22.96
<i>Prevotella sp002265625</i>	0.20 \pm 0.38	10.48 \pm 18.11
<i>Prevotella sp900551275</i>	0.19 \pm 0.36	14.47 \pm 23.20
<i>Prevotella sp000436035</i>	0.16 \pm 0.37	8.05 \pm 16.36
<i>Prevotella sp900290275</i>	0.11 \pm 0.16	7.31 \pm 13.58
<i>Prevotella sp900556795</i>	0.10 \pm 0.21	14.75 \pm 25.02
<i>Prevotella sp000436915</i>	0.09 \pm 0.15	8.83 \pm 16.88
<i>Prevotella sp900543975</i>	0.09 \pm 0.21	7.43 \pm 15.23

Table 2. Most relatively abundant and prevalent *Prevotella* species in the human gut.

Human gut *Prevotella* species occupy distinct functional niches. *Prevotella copri* in the human gut are associated with high fibre non-Western diets as they possess extensive repertoires of carbohydrate active enzymes that allow this species to metabolise complex polysaccharides. We thus hypothesised that other gut *Prevotella* species are also adapted to carbohydrate-related metabolism although their exact gene content likely differs compared with *P. copri*, thus allowing them to utilise distinct fractions of dietary fibre. To identify differences in gene content, we first predicted protein coding genes in each genome and then compared predicted genes to the KEGG Orthology (KO) database to infer metabolic potential. NMDS ordinations of KO counts indicated that genomes of prevalent and relatively abundant species belonging to the two *copri* and *stercorea* groups identified above significantly differed in their gene content (Fig. S1), and this could be attributed to several pectin and hemicellulolytic enzymes detected in the *copri* group but missing from *stercorea* (Table S6). Hemicellulose is a polysaccharide polymer in plant cell walls and typically consists of sugars such as xylose, arabinose, mannose, galactose and rhamnose. The *stercorea* group either lacks or has fewer copies of genes such as xylanase (K01181), xylosidase (K01198), xylose symporters (K08138), mannosidase (K01218), pectinesterase (K01051), rhamnosidase (K05989) and endoglucanase (K01179) suggesting that *stercorea* species have reduced ability to utilise hemicelluloses compared with *copri*-related species. Instead, *stercorea* species may participate in breaking down oligosaccharides and glycoproteins based on detection of genes such as N-acetylglucosamine-6-phosphate deacetylase (K01443) and endo-beta-N-acetylglucosaminidase (K01227). In addition, *stercorea*

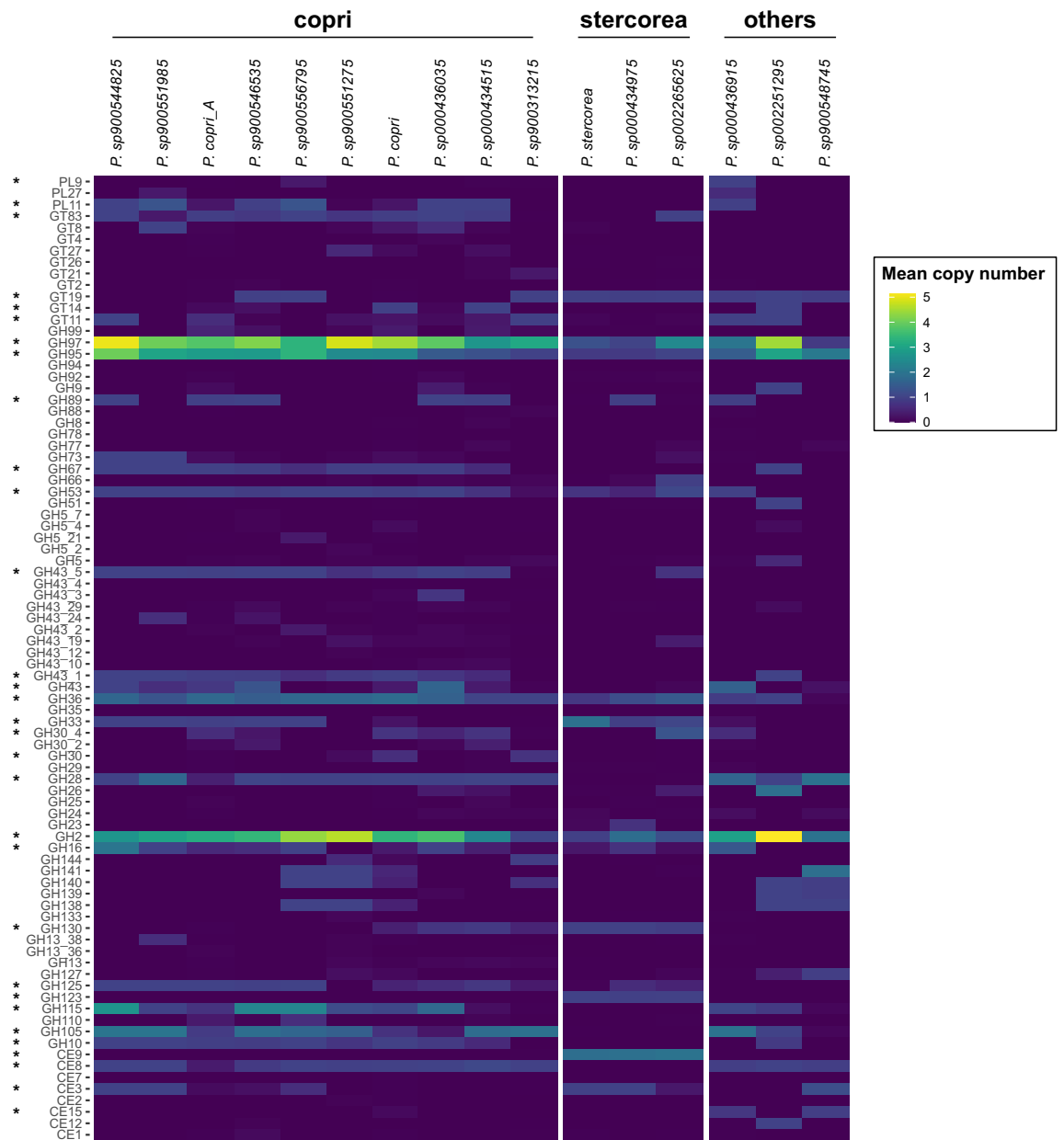


Figure 3. Distribution of carbohydrate active enzyme (CAZy) families in human gut *Prevotella* (average copies per genome). Only genomes with >1% relative abundance or >50% prevalence as indicated in Fig. 2 are included. CAZy families were identified in the genomes using Hidden Markov Models. Asterisks indicate families with statistically different copy counts (FDR < 0.001 Kruskal Wallis test with false discovery rate adjustment) and with at least one copy in any one of the three groups (copri, stercorea or other). See Table S7 for copy numbers. CE carbohydrate esterase, GH glycoside hydrolase, GT glycosyl transferase, PL polysaccharide lyase.

species possess components of the glycine cleavage system and histidine utilisation system (hut) which could indicate a more extensive amino acid metabolism compared with the copri group. These carbohydrate utilisation patterns are also reflected in CAZy profiles indicating the stercorea group as non-hemicellulose degraders. CAZy families involved in hemicellulose degradation such as CE8 (pectin methylesterase), GH10 (β -1,4-xylanase), GH28 (polygalacturonase), GH30 (endo-xylanase), GH43_1 (β -xylosidase, α -L-arabinofuranosidase), GH67 (α -glucuronidase), GH115 (xylan α -1,2-glucuronidase) and GH141 (α -L-fucosidase, xylanase) were absent in nearly all stercorea group genomes, whereas the presence of CE3 (acetyl xylan esterase) and CE9 (N-acetylglucosamine 6-phosphate deacetylase) suggest a possible role in assisting with removal of ester modifications to facilitate carbohydrate degradation⁵² by *P. copri* or other gut bacteria with the necessary gene repertoire (Table S7) (Fig. 3). The presence of GH33 (sialidases) and GH123 (N-acetylgalactosaminidases) further suggest that the stercorea species could metabolise sialic acids derived from dietary (e.g. glycolylneuraminic acid in red meat⁵³) and/or human glycans (e.g. acetylneuraminic acid) and are likely to utilise these substrates in

addition to or instead of dietary fibre. Other *Prevotella* species such as sp001275135 prevalent in France and Spain possess hemicellulolytic CAZy families akin to copri, whereas sp002299635 and sp003447235 are more similar with stercorea in that they lack hemicellulolytic families. These similarities in CAZy content among species are supported by an NMDS of total CAZy counts showing two broad types of CAZy profiles in human gut *Prevotella*, one representing the copri group and the other the stercorea group with reduced capacity to utilise hemicelluloses (Fig. S2).

To examine whether diet may influence the distribution of gut *Prevotella* species, we assessed dietary intake information collected as part of the Yunnan gut microbiome study²⁵ together with abundance values obtained from metagenome read mapping (see section above). The Yunnan survey was comprised of 881 individuals representing six ethnic groups in Yunnan province; 427 of the 881 were recruited from rural counties and another 454 from urban Kunming (capital of Yunnan province). A total of 676 individuals provided their past month dietary intake through a binary response questionnaire. We first established that diet primarily differed among the ethnic groups followed by rural/urban status ($p < 0.05$, PERMANOVA) (Table S8). While the Zang (Tibetan) diet was most distinct compared with other ethnic groups (Fig. S3A), total *Prevotella* relative abundance in the Zang was comparable to several ethnicities (11.7% vs. 10.9–15.5%) whereas the Dai (8.0%) and Han (3.9%) had significantly lower relative abundances (Fig. S3B). Across all ethnicities, only coffee consumption was associated with an increased *Prevotella* relative abundance ($p < 0.05$, generalised linear model with FDR adjustment) (Table S9) although coffee consumption was reported in only 19 individuals. Other foods in the questionnaire were not correlated with either total *Prevotella* relative abundance or with abundances of the two stercorea and copri lineages identified above. Similarly, the low relative abundances of *Prevotella* in the Dai and Han could not be attributed to any specific foods included in this questionnaire. These observations indicate that non-dietary influences (e.g. genetics, lifestyle) could play a role in gut *Prevotella* associations in the Yunnan cohort, although more in-depth and quantitative dietary information in addition to existing data provided in that Yunnan study may be required to resolve exact correlations between these factors and diet.

Discussion

The *Prevotella* genus is a lineage of interest as its member species are a common feature of the human gut microbiome and have been repeatedly implicated in health and disease. *P. copri* is often the primary focus of the *Prevotella* genus, but there is increasing incentive to catalogue and characterise other gut *Prevotella* species. For example, a gut microbiome survey of competitive cyclists revealed enrichment of *P. copri* and several unidentified *Prevotella* species correlated with exercise duration²¹. Another study of gut microbiome development in malnourished children reported that children who received a food formula designed to stimulate development of a healthy gut microbiota⁵⁴ showed greater enrichment of *P. copri* and other unnamed *Prevotella* species compared with existing nutrition regimes²². To capture gut *Prevotella* diversity without relying on known marker genes (e.g. MetaPhlAn) or genomes deposited in RefSeq (Kraken2), our workflow in this study classifies metagenome reads against a large collection of human gut-derived *Prevotella* MAGs to better resolve and estimate abundances of *Prevotella* species in the human gut^{15,17,18}.

We show that the most numerically abundant *Prevotella* species in the human gut are largely comprised of two distinct lineages, one closely related to and including *P. copri* while the other to *P. stercorea*. Both the copri and stercorea groups were more relatively abundant in rural non-Westernised human populations, suggesting that the presence of stercorea-related lineages is also driven by diet and lifestyle factors akin to *P. copri*^{2–4}. Using the *P. copri* phylogenetic framework inferred by Tett et al.¹, the copri group of species we identified here largely corresponded with the four clades circumscribed in that study and the potential for carbohydrate metabolism detected in this group also broadly mirrors that reported by Tett et al.¹. While *P. copri* are known hemicellulose degraders^{55,56}, compared with *P. copri* the stercorea group occupies a separate functional niche categorised by a lack of hemicellulolytic gene families. In addition, the stercorea group possesses carbohydrate esterase families that were comparatively much less common in the copri group. From these features, we posit that a possible role of the stercorea group is to remove ester-based modifications from dietary carbohydrates to facilitate hydrolysis by gut species with the necessary hydrolytic capacity⁵². Additionally, the stercorea group lineages possess sialidases. Sialic acids have been demonstrated to play roles in processes such as mediating interaction between immune cells, pathogen binding to human cells⁵⁷ and inflammation in the gut⁵⁸. The presence of sialidases in the stercorea group therefore suggests that stercorea-related bacteria have important functions that could impact human health.

The main limitation of this work is the lack of detailed lifestyle and dietary information needed to investigate whether distribution of the several gut *Prevotella* identified here is influenced by lifestyle and/or diet. Existing evidence indicate that *P. copri* is associated with dietary fibre intake⁵⁹, but we did not observe such associations with the limited dietary data (yes or no responses) provided in the Yunnan study²⁵. In addition, our metabolic inferences of the various *Prevotella* species are based on genome sequences and thus need to be validated by isolation and laboratory culture experiments. Nevertheless, since low fibre diets can result in poor gastrointestinal health⁶⁰ and *Prevotella* are one of the major fibrolytic bacteria in human guts⁶¹, we foresee that there will be increased interest in exploiting *Prevotella* to potentially improve human health and nutrition. Our findings of distinct *Prevotella* species in the human population underscores the need to classify the human gut microbiome in greater resolution to understand and exploit the metabolic potential of gut microorganisms.

Data availability

Metagenome-assembled genomes used in this study can be downloaded from <https://drive.google.com/file/d/1hLaPiR1qHKBeDsK0A3WLzk3ADgCFkrNH/view?usp=sharing>. Codes used to generate genomes are provided in supplementary file 1.

Received: 19 October 2021; Accepted: 15 April 2022

Published online: 31 May 2022

References

- Tett, A. *et al.* The *Prevotella copri* complex comprises four distinct clades underrepresented in westernized populations. *Cell Host Microbe* **26**, 666–679 (2019).
- Wu, G. D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105 (2011).
- David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).
- De Filippis, F. *et al.* High-level adherence to a Mediterranean diet beneficially impacts the gut microbiota and associated metabolome. *Gut* **65**, 1812–1821 (2016).
- Kovatcheva-Datchary, P. *et al.* Dietary fiber-induced improvement in glucose metabolism is associated with increased abundance of *Prevotella*. *Cell Metab.* **22**, 971–982 (2015).
- Asnicar, F. *et al.* Microbiome connections with host metabolism and habitual diet from 1098 deeply phenotyped individuals. *Nat. Med.* **27**, 321–332 (2021).
- Pedersen, H. K. *et al.* Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* **535**, 376–381 (2016).
- Li, J. *et al.* Gut microbiota dysbiosis contributes to the development of hypertension. *Microbiome* **5**, 14 (2017).
- Dillon, S. M. *et al.* Gut dendritic cell activation links an altered colonic microbiome to mucosal and systemic T-cell activation in untreated HIV-1 infection. *Mucosal Immunol.* **9**, 24–37 (2016).
- Fehlner-Peach, H. *et al.* Distinct polysaccharide utilization profiles of human intestinal *Prevotella copri* isolates. *Cell Host Microbe* **26**, 680–690 (2019).
- De Filippis, F. *et al.* Distinct genetic and functional traits of human intestinal *Prevotella copri* strains are associated with different habitual diets. *Cell Host Microbe* **25**, 444–453 (2019).
- Hayashi, H., Shibata, K., Sakamoto, M., Tomita, S. & Benno, Y. *Prevotella copri* sp. nov. and *Prevotella stercorea* sp. nov., isolated from human faeces. *Int. J. Syst. Evol. Microbiol.* **57**, 941–946 (2007).
- Belkacemi, S. *et al.* Description of *Prevotella rectalis* sp. nov., a new bacterium isolated from human rectum. *New Microbes New Infect.* **36**, 100703 (2020).
- Shahi, S. K. *et al.* *Prevotella histicola*, a human gut commensal, is as potent as COPAXONE® in an animal model of multiple sclerosis. *Front. Immunol.* **10**, 462 (2019).
- Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
- Tett, A., Pasolli, E., Masetti, G., Ercolini, D. & Segata, N. *Prevotella* diversity, niches and interactions with the human host. *Nat. Rev. Microbiol.* **19**, 585–599 (2021).
- Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).
- Pasolli, E. *et al.* Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662 (2019).
- Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).
- Anani, H. *et al.* *Prevotella ihumii* sp. nov., a new bacterium isolated from a stool specimen of a healthy woman. *New Microbes New Infect.* **32**, 100607 (2019).
- Petersen, L. M. *et al.* Community characteristics of the gut microbiomes of competitive cyclists. *Microbiome* **5**, 98 (2017).
- Chen, R. Y. *et al.* A microbiota-directed food intervention for undernourished children. *N. Engl. J. Med.* **384**, 1517–1528 (2021).
- Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).
- Yeoh, Y. K. *et al.* Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor. *Gut* **69**, 1998–2007 (2020).
- Sun, Y. *et al.* Population-level configurations of gut mycobiome across 6 ethnicities in urban and rural China. *Gastroenterology* **160**, 272–286 (2021).
- Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- Kang, D. D. *et al.* MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
- Wu, Y. W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
- Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
- Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
- Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
- Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
- Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
- Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.* **6**, 6528 (2015).
- Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
- Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
- Bäckhed, F. *et al.* Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* **17**, 690–703 (2015).
- Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–64 (2014).
- Yachida, S. *et al.* Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.* **25**, 968–976 (2019).
- Pehrsson, E. C. *et al.* Interconnected microbiomes and resistomes in low-income human habitats. *Nature* **533**, 212–216 (2016).
- Brito, I. L. *et al.* Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535**, 435–439 (2016).
- Zeevi, D. *et al.* Personalized nutrition by prediction of glycemic responses. *Cell* **163**, 1079–1094 (2015).
- Liu, W. *et al.* Unique features of ethnic mongolian gut microbiome revealed by metagenomic analysis. *Sci. Rep.* **6**, 34826 (2016).
- Rampelli, S. *et al.* Metagenome sequencing of the Hadza hunter-gatherer gut microbiota. *Curr. Biol.* **25**, 1682–1693 (2015).
- Smits, S. A. *et al.* Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science* **357**, 802–806 (2017).
- Mehta, R. S. *et al.* Stability of the human faecal microbiome in a cohort of adult men. *Nat. Microbiol.* **3**, 347–355 (2018).

49. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
50. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).
51. Oksanen, J. *et al.* Package 'vegan'. Community ecology package, version 2.5-7 (2020).
52. Nakamura, A. M., Nascimento, A. S. & Polikarpov, I. Structural diversity of carbohydrate esterases. *Biotechnol. Res. Innov.* **1**, 35–51 (2017).
53. Zaramela, L. S. *et al.* Gut bacteria responding to dietary change encode sialidases that exhibit preference for red meat-associated carbohydrates. *Nat. Microbiol.* **4**, 2082–2089 (2019).
54. Gehrig, J. L. *et al.* Effects of microbiota-directed foods in gnotobiotic animals and undernourished children. *Science* **365**, 139 (2019).
55. Accetto, T. & Avguštin, G. The diverse and extensive plant polysaccharide degradative apparatuses of the rumen and hindgut *Prevotella* species: A factor in their ubiquity?. *Syst. Appl. Microbiol.* **42**, 107–116 (2019).
56. Emerson, E. L. & Weimer, P. J. Fermentation of model hemicelluloses by *Prevotella* strains and *Butyrivibrio fibrisolvens* in pure culture and in ruminal enrichment cultures. *Appl. Microbiol. Biotechnol.* **101**, 4269–4278 (2017).
57. Varki, A. Sialic acids in human health and disease. *Trends Mol. Med.* **14**, 351–360 (2008).
58. Huang, Y. L., Chassard, C., Hausmann, M., von Itzstein, M. & Hennet, T. Sialic acid catabolism drives intestinal inflammation and microbial dysbiosis in mice. *Nat. Commun.* **6**, 8141 (2015).
59. Vangay, P. *et al.* US immigration westernizes the human gut microbiome. *Cell* **175**, 962–972 (2018).
60. Desai, M. S. *et al.* A dietary fiber-deprived gut microbiota degrades the colonic mucus barrier and enhances pathogen susceptibility. *Cell* **167**, 1339–1353 (2016).
61. Flint, H. J., Scott, K. P., Duncan, S. H., Louis, P. & Forano, E. Microbial degradation of complex carbohydrates in the gut. *Gut Microbes* **3**, 289–306 (2012).

Acknowledgements

This work was supported by InnoHK (The Government of Hong Kong, Special Administrative Region of the People's Republic of China), the National Natural Science Foundation of China (U1802282, 82060107), Applied Basic Research Projects of Yunnan Province, China (202001AY070001-202), and a seed fund from the Centre for Gut Microbiota Research (Faculty of Medicine, The Chinese University of Hong Kong).

Author contributions

Y.K.Y. conceived the project. Data analyses was performed by Y.K.Y. and L.Y.T.I. Manuscript was written by Y.K.Y. with input from Y.S., L.Y.T.I., L.W., F.K.L.C., Y.M., and S.C.N.

Competing interests

S.C.N. and F.K.L.C. are the scientific co-founders and sit on the board of Directors of GenieBiome Ltd. S.C.N. has served as an advisory board member for Pfizer, Ferring, Janssen, and Abbvie and a speaker for Ferring, Tillotts, Menarini, Janssen, Abbvie, and Takeda. S.C.N. has received research grants from Olympus, Ferring, Janssen and Abbvie. F.K.L.C. has served as an advisor and lecture speaker for Eisai Co. Ltd., AstraZeneca, Pfizer Inc., Takeda Pharmaceutical Co., and Takeda (China) Holdings Co. Ltd.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-12721-4>.

Correspondence and requests for materials should be addressed to Y.M. or S.C.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022