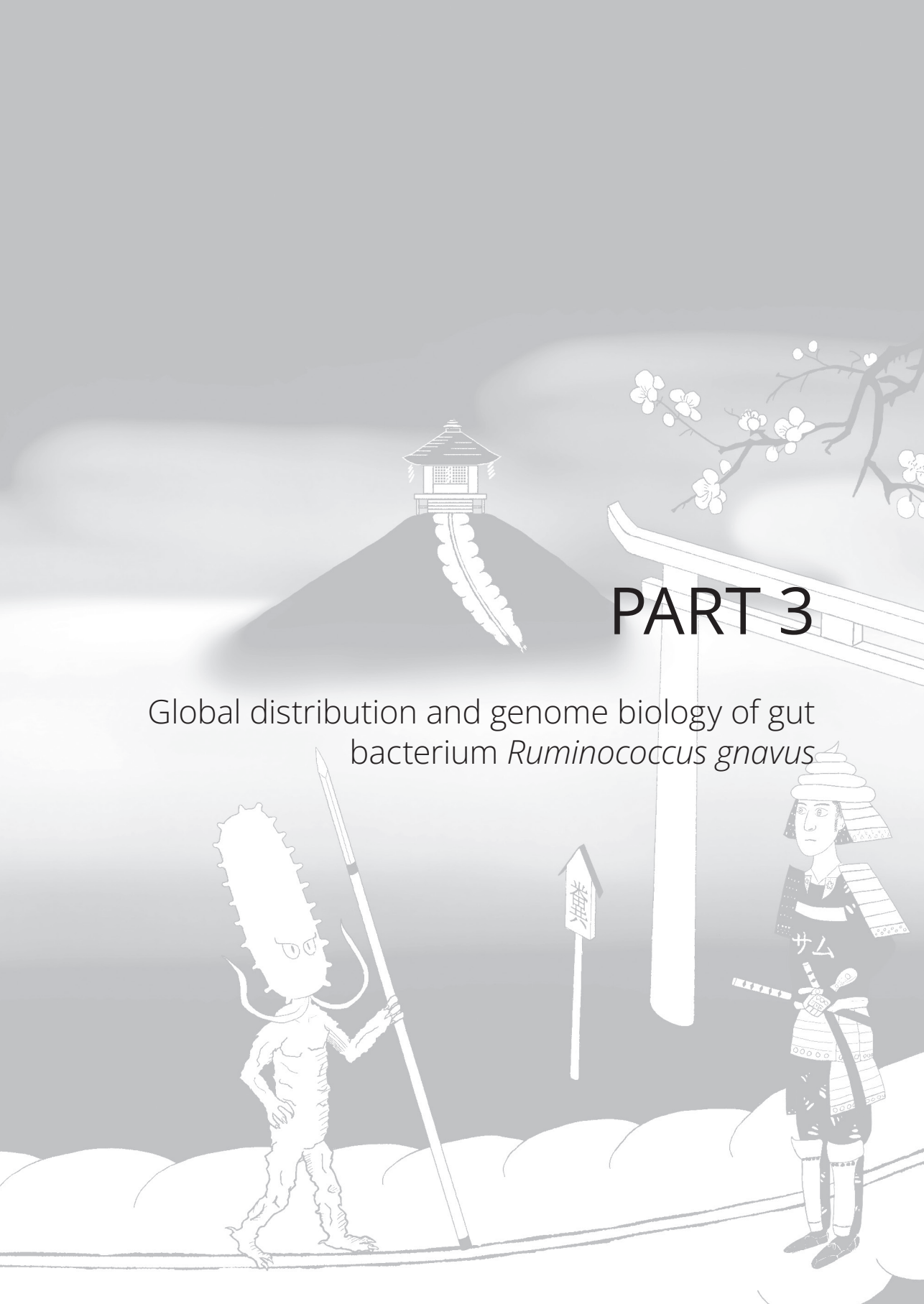
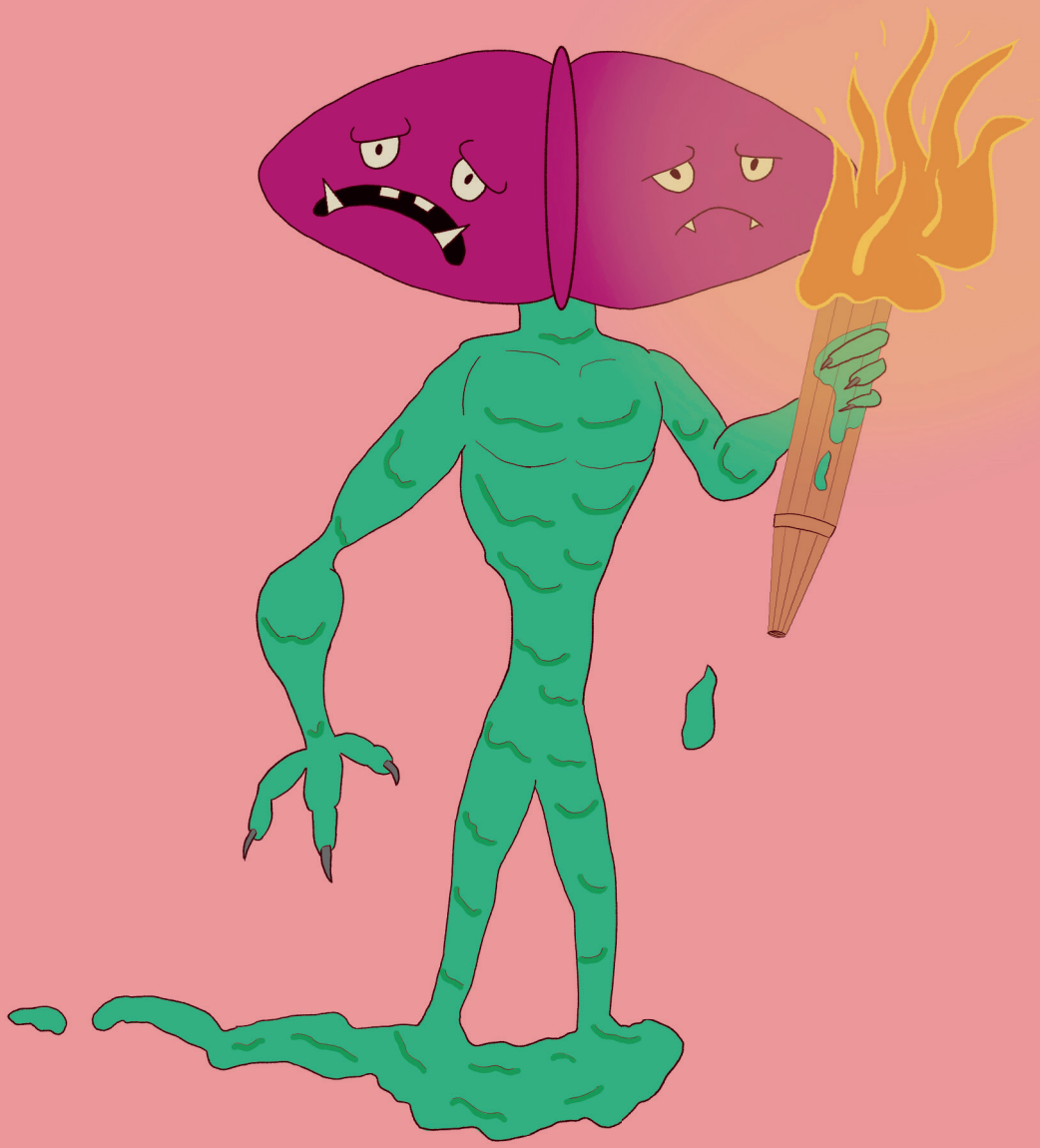
Guardians of the gut: harnessing bioinformatics to study the gut microbiome and faecal microbiota transplantation in intestinal disorders

Nooij, S.



PART 3

Global distribution and genome biology of gut bacterium *Ruminococcus gnavus*



Chapter 7.1

Metagenomic global survey and in-depth genomic analyses of *Ruminococcus gnavus* reveal differences across host lifestyle and health status

S. Nooij[1,2,3], N. Plomp[4], I.M.J.G. Sanders[1], L. Schout[1,2], A.E. van der Meulen[5], E.M. Terveer[1,2,3], J.M. Norman[6], N. Karcher[7], M.F. Larralde[1], R.H.A.M. Vossen[8], S.L. Kloet[8], K.N. Faber[9], H.J.M. Harmsen[4], G.F. Zeller[1,2,7], E.J. Kuijper[1,2], W.K. Smits[1,2], Q.R. Ducarmon[1,2,7]

[1]Leiden University Center of Infectious Diseases (LUCID), Leiden University Medical Center, Leiden, The Netherlands

[2]Center for Microbiome Analyses and Therapeutics, Leiden University Medical Center, Leiden, The Netherlands

[3]Netherlands Donor Feces Bank (NDFB), Leiden University Medical Center, Leiden, the Netherlands

[4]Department of Medical Microbiology and Infection Prevention, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

[5]Department of Gastroenterology and Hepatology, Leiden University Medical Center, Leiden, The Netherlands

[6]Vedanta Biosciences, Inc., Cambridge, Massachusetts, USA

[7]Molecular Systems Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany

[8]Leiden Genome Technology Center, Leiden University Medical Center, Leiden, The Netherlands

[9]Department of Gastroenterology and Hepatology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

Nature Communications ***16***, *1182 (2025). DOI: 10.1038/s41467-025-56449-x*

Summary

Ruminococcus gnavus is a gut bacterium found in >90% of healthy individuals, but its increased abundance is also associated with chronic inflammatory diseases, particularly Crohn's disease. Nevertheless, its global distribution and intraspecies genomic variation remain understudied. By surveying 12,791 gut metagenomes, we recapitulated known associations with metabolic diseases and inflammatory bowel disease. We uncovered a higher prevalence and abundance of *R. gnavus* in Westernized populations and observed bacterial relative abundances up to 83% in newborns. Next, we built a resource of *R. gnavus* isolates (N=45) from healthy individuals and Crohn's disease patients and generated complete *R. gnavus* genomes using PacBio circular consensus sequencing. Analysis of these genomes and publicly available high-quality draft genomes (N=333 genomes) revealed multiple clades which separated Crohn's-derived isolates from healthy-derived isolates. Presumed *R. gnavus* virulence factors could not explain this separation. Bacterial genome-wide association study revealed that Crohn's-derived isolates were enriched in genes related to mobile elements and mucin foraging. Together, we present a large *R. gnavus* resource that will be available to the scientific community and provide novel biological insights into the global distribution and genomic variation of *R. gnavus*.

Introduction

The human gut microbiome is a topic of intense research interest and many bacterial species have been associated with specific diseases[1]. One such species is *Ruminococcus gnavus*, for which associations with human health have been reported in the context of various ailments[2-7]. Officially, its taxonomic status has been revised and *R. gnavus* is now member of the genus *Mediterraneibacter*, but it has also been termed *Faecalicatena gnavus*[8]. Here, we will designate the species as *Ruminococcus gnavus*. *R. gnavus* is a non-spore forming Gram-positive member of the bacterial phylum Bacillota (formerly Firmicutes) and was first described in 1976[9]. It is considered a prevalent member of the human gut microbiome (present in > 90% of healthy European and North-American adults), but can also be found in the gastrointestinal tract of a variety of animal species[10,11]. Its median relative abundance in humans is reported to be approximately 0.1% - 0.3%, although it should be noted that these estimates were based on small and geographically restricted studies[12,13].

In microbiome association studies, increases in *R. gnavus* relative abundance have consistently been linked to diseases including metabolic syndrome, type 2 diabetes mellitus and Crohn's disease (CD, a form of inflammatory bowel disease (IBD))[2,3,14]. Furthermore, its relative abundance increased concomitantly with symptomatic flares in CD, where it reached up to 69.5% of the gut microbiome[2]. While it remains unknown if *R. gnavus* causally contributes to disease development or whether the increased abundance is a result of the changing intestinal environment, several molecular mediators have been identified that potentially contribute to disease. For instance, the cell-surface exposed polysaccharide glucorhamnan has been described as pro-inflammatory, with a strain-dependent effect, depending on whether the *R. gnavus* isolate carried a capsular polysaccharide that promoted a more tolerogenic response[15,16]. However, these observations are limited by the fact that they were made using one or few isolates and strain variation remains underexplored in many gut microbes, including *R. gnavus*.

Not only mechanistic, but also genomic studies of *R. gnavus* have suffered from a limited scope. One study divided *R. gnavus* into two clades based on genome sequences and noted that one was enriched in IBD patients[2]. However, this study was limited by a low number of draft isolate genomes (N = 11) and a scarcity of knowledge on experimentally verified virulence factors of *R. gnavus* at the time[15-20]. A more recent study based on 152 draft genomes identified three major lineages, but genomes of different host organisms were mixed and this study did not investigate associations of genetic features with metadata[11]. Therefore, an important outstanding question remains whether proposed *R. gnavus* virulence factors are enriched in IBD-derived isolates, or whether different genes and functions could separate IBD-derived *R. gnavus* isolates from controls.

In this work, we surveyed global *R. gnavus* prevalence and abundance across thousands of gut metagenomes to provide a more nuanced picture across human lifespan, different lifestyles, and disease, thereby revealing striking differences. Next, through extensive culturing efforts we established a resource of 45 *R. gnavus* isolates and applied PacBio circular consensus sequencing (CCS) to generate complete genomes. This collection of isolates and their complete genomes provides ample scope for targeted experimental follow-up work and will be available as a community resource for the scientific community. We complemented this unique collection with publicly available (short-read draft) genomes, which allowed us to perform large-scale comparative genomics at both the level of phylogeny and predicted gene functions.

Results

Intestinal colonization with *R. gnavus* is associated with age, health, geography and lifestyle

In order to provide a nuanced view of *R. gnavus* prevalence and abundance across health and disease, geography, and lifestyle, we screened 12,791 publicly available metagenomes from all over the world with manually curated metadata (Fig. 1, Supplementary Data 1; full per-sample metadata are available through https://waldronlab.io/curatedMetagenomicData/)[21]. We observed *R. gnavus* in 50.58% of all included subjects and the prevalence in 9,126 healthy individuals was 43.09% (Fig. 1a). As *R. gnavus* has been robustly associated with disease, especially with metabolic disease and IBD[2,3], we compared *R. gnavus* prevalence and abundance between patients with these diseases and healthy subjects (or asymptomatic control subjects) in a meta-analysis. *R. gnavus* was approximately 1.6 times more prevalent in IBD patients (70.2%; logistic regression, $p < 2.2 \times 10^{-16}$, odds ratio (OR [95% confidence interval]) = 3.1 [2.6-3.7]), 1.3 times more with hypertension (58.0%; p = 0.00127, OR = 1.8 [1.3-2.5]), 1.5 times with type-2 diabetes (T2D; 62.9%; $p = 1.52 \times 10^{-9}$, OR = 2.2 [1.8-2.8]), and 2.2 times with atherosclerotic cardiovascular diseases (ACVD; 96.2%; $p < 2.2 \times 10^{-16}$, OR = 33.4 [17.6-74.1]) compared to healthy subjects. Furthermore, the relative abundance of *R. gnavus* was also higher in these conditions as compared to healthy (Fig. 1b; healthy: median = 0%, 1st-3rd quartile [0-0.08%]; IBD: median = 0.11% [0-1.04%], linear model, $p < 2.2 \times 10^{-16}$; T2D: median = 0.027% [0.0-0.22%], $p = 1.9 \times 10^{-10}$; ACVD: median = 0.78% [0.09-3.14%], $p < 2.2 \times 10^{-16}$), except hypertension (median = 0.01% [0-0.07%], p = 0.399). Together, we thus recapitulated that *R. gnavus* occurs more frequently and in higher abundances in the gut microbiome of patients suffering from IBD, hypertension and T2D. Additionally, our analysis uncovered a striking novel enrichment in ACVD, which had the highest prevalence and abundance of any disease group.

Subsequently, we investigated prevalence and relative abundance of *R. gnavus* across countries (Fig. 1c, 1d and Supplementary Fig. 1a). We show only healthy

individuals to exclude possible confounding by diseases such as IBD and metabolic disease. We observed large differences in prevalence, which ranged between 10-90% across countries (overall median: 41%) and mean relative abundance per country ranged between 0.0078-4.05% (overall mean = 0.67% ±3.20 standard deviation; Supplementary Fig. 1a). This variation could be partly explained by Westernization status; this binary classification of Westernized / non-Westernized lifestyles is based on, among others, access to medical care and pharmaceuticals, livestock exposure and diet[22]. Westernized individuals had higher prevalence and abundance of *R. gnavus* compared to non-Westernized individuals (Fig. 1c-e; prevalence: logistic regression, $p < 2.2 \times 10^{-16}$; abundance: linear model, $p < 2.2 \times 10^{-16}$). As these data were generated in multiple studies, we cannot exclude effects of technical differences (e.g., DNA extraction method). To partially check for this, we investigated sequencing depth and found that higher prevalence and abundance were not the result of higher sequencing depth in Westernized countries as non-Western samples were sequenced deeper (Supplementary Fig. 1b; t-test $p = 6.9 \times 10^{-20}$). These differences hold true for any 10% quantile of sequencing depth (Supplementary Fig. 1c, Methods). We also checked for possible correlations between sequencing depth and *R. gnavus* abundance and found a weakly negative correlation in both Westernized and non-Westernized metagenomes (Supplementary Fig. 1d). In conclusion, *R. gnavus* colonization is vastly different between countries, and Westernization (lifestyle) may be a major factor contributing to these differences.

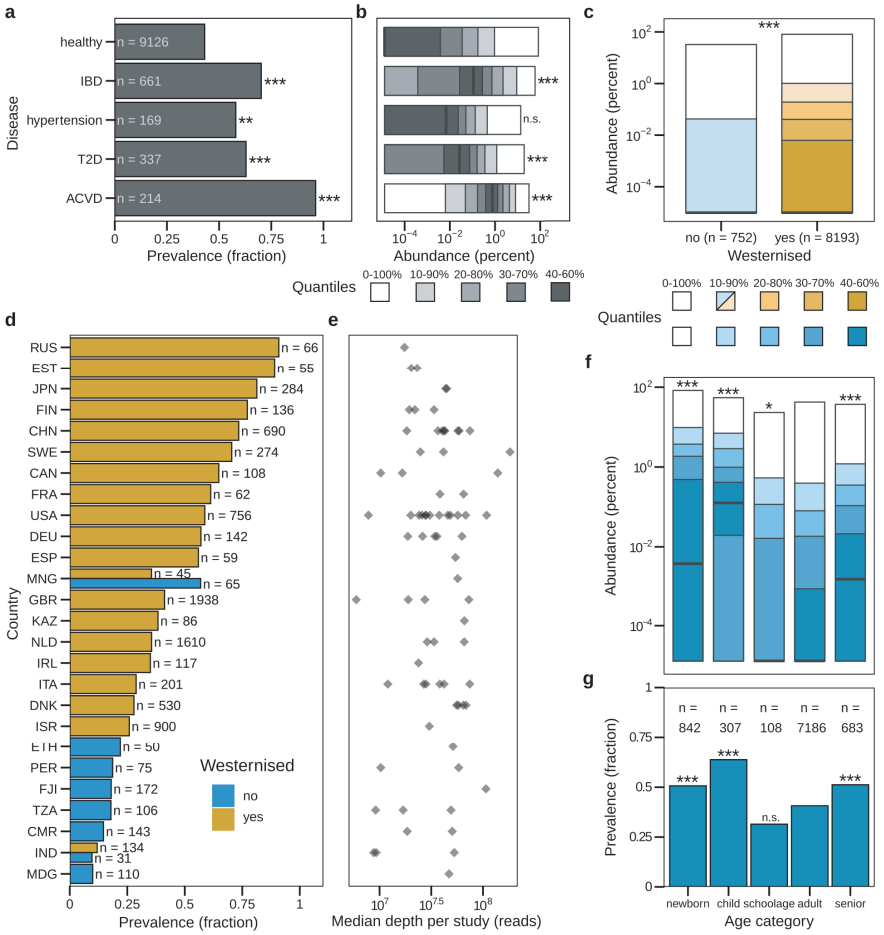


Fig. 1. Intestinal colonization with *R. gnavus* is associated with age, health, geography, and lifestyle. a We queried the public resource curatedMetagenomicData for relative abundances of *R. gnavus* in human stools to conduct a meta-analysis of global prevalence and abundance. Prevalence is shown as fraction of subjects with *R. gnavus* abundance > 0, grouped by selected health conditions. IBD: inflammatory bowel diseases, T2D: type-2 diabetes, ACVD: atherosclerotic cardiovascular diseases. Each disease is compared to healthy. Each disease group is compared to healthy using logistic regression. IBD: $p < 2.2 \times 10^{-16}$; hypertension: $p = 0.00127$; T2D: $p = 1.52 \times 10^{-9}$; ACVD: $p < 2.2 \times 10^{-16}$. **b** Relative abundance of *R. gnavus* in the same groups as **a**, shown as quantile plots, using quantiles ranging from 0 to 100% in increments of 10 with the median shown as a thick black line and quantiles closer to the median shown as darker shades of the same color (see Methods). Each disease is compared to healthy using linear regression. IBD: $p < 2.2 \times 10^{-16}$; hypertension: $p = 0.399$; T2D: 1.9×10^{-10}; ACVD: $p < 2.2 \times 10^{-16}$. **c** Comparison of *R. gnavus* abundance between healthy people from Westernized and non-Westernized societies as quantile plot. $P < 2.2 \times 10^{-16}$, calculated using linear regression. **d** Prevalence of *R. gnavus* grouped per country and colored by Westernization, only showing results from countries from which at least 50 samples were collected. (Countries are abbreviated by ISO 3166-1 alpha-3 codes.) **e** Sequencing depth control per country (same as **d**). Each diamond represents a study that col-

lected samples from the corresponding country. Sequencing depth is shown as median number of reads generated per country in the study. **f** Relative abundance of *R. gnavus* in different age categories (newborn: < 1 year, child: 1-11 years, school age: 12-18 years, adult: 19-65, senior: 65+ years) shown as quantile plots. Age categories are listed in **g**. Each age category is compared to adult using linear regression. Newborn: $p < 2.2 \times 10^{-16}$; child: $p < 2.2 \times 10^{-16}$; schoolage: $p = 0.0164$; senior: $p = 1.37 \times 10^{-6}$. **g** Prevalence of *R. gnavus* among different age categories. Each category is compared to adult using logistic regression. Newborn: $p = 1.14 \times 10^{-6}$; child: $p < 2.2 \times 10^{-16}$; schoolage: $p = 0.0797$; senior: $p = 2.92 \times 10^{-4}$. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, n.s. not significant. In **b**, **c**, and **f**, a pseudocount of 1.3×10^{-5} is added to all abundances to enable visualization on a logarithmic scale. Source data are provided as a Source Data file.

Subsequently, we investigated prevalence and relative abundance of *R. gnavus* across countries (Figures 1C, 1D and S1A). In this analysis, we included only healthy individuals to account for the possibility of confounding by diseases such as IBD and metabolic disease. We observed large differences in prevalence, which ranged between 10-90% across countries (overall median: 41%) and mean relative abundance per country ranged between 0.0078-4.05% (overall mean = 0.67% ±3.20 standard deviation; Figure S1A). This variation could be partly explained by Westernization status; this binary classification of Westernized / non-Westernized lifestyles is based on, among others, access to medical care and pharmaceuticals, livestock exposure and diet[21]. Westernized individuals had higher prevalence and abundance of *R. gnavus* compared to non-Westernized individuals (Figure 1C-E; prevalence: Chi-square $p = 2.6 \times 10^{-47}$). As these data were generated in multiple studies, we cannot exclude effects of technical differences (e.g., DNA extraction method). To partially check for this, we investigated sequencing depth and found that higher prevalence and abundance were not the result of higher sequencing depth in Westernized countries as non-Western samples were sequenced deeper (Figure S1B; t-test $p = 6.9 \times 10^{-20}$). These differences hold true for any 10% quantile of sequencing depth (Figure S1C, Methods). We also checked for possible correlations between sequencing depth and *R. gnavus* abundance and found a weakly negative correlation in both Westernized and non-Westernized metagenomes (Figure S1D). In conclusion, *R. gnavus* colonization is vastly different between countries, and Westernization (lifestyle) may be a major factor contributing to these differences.

We noted extremely high *R. gnavus* abundance values in healthy people, up to a relative abundance of 83%. Metagenomes with the highest abundances were often samples collected from newborns and children up to age 2, most of whom were recorded not to have received antibiotics. This motivated a further analysis of age-related patterns of *R. gnavus* colonization (Fig. 1f)[21]. *R. gnavus* abundances were higher in newborns (linear model, $p < 2.2 \times 10^{-16}$), children up to 11 years old ($p < 2.2 \times 10^{-16}$), and adolescents between 12 and 18 years old ('schoolage'; $p = 0.0164$) as compared to adults. Abundances were also higher in seniors (65-92 years old) than in adults ($p = 1.37 \times 10^{-6}$). We observed similar patterns regarding *R. gnavus* prevalence (Figure 1G), where newborns (logistic regression, $p = 1.14 \times 10^{-6}$), children aged 1-11 ($p < 2.2 \times 10^{-16}$) and

seniors ($p = 2.92\times 10^{-4}$) were more likely to carry *R. gnavus* than adults. Adolescents and adults did not have different prevalence of *R. gnavus* ($p = 0.0797$).

The high abundances of *R. gnavus* in infants instigated a closer inspection of abundance over age and in correlation to breastfeeding, as breastfeeding was recently reported to have a strong impact on *R. gnavus* colonization[23]. Looking at *R. gnavus* abundance in the first ten years of life (Supplementary Fig. 2a), we see a rapid increase after the first half year, followed by a decline and rebound around 8 years. We found the shift in the first half year to strongly correlate with feeding practice (Chi-square, $p = 1.49 \times 10^{-15}$). Specifically, infants that were breastfed had lower *R. gnavus* abundance than children that received no breastfeeding (linear model; exclusive breastfeeding, p = 6.49 x 10-11; mixed feeding, $p = 4.69 \times 10^{-4}$; Supplementary Fig. 2b). To exclude possible confounding and identify other associated factors, we also tested for associations between *R. gnavus* abundance with feeding practice (n = 184), mode of delivery (n = 170) and antibiotics use (n = 94) in infants of age up to two years, for whom feeding practice data had been recorded, using multivariable linear modelling. This indicated that only feeding practice was significantly associated with *R. gnavus* abundance in infants (Chi-square of total variable effect, $p = 8.93 \times 10^{-6}$). In summary, we find evidence that indicates that breastfeeding delays of *R. gnavus* colonization in infants, corresponding with previous reports[23].

In summary, colonization with *R. gnavus* appears to be dynamic across the lifespan in healthy individuals, with the highest abundances observed in newborns. While these metagenomic analyses provide important insight into the global distribution of *R. gnavus*, in-depth genomic analyses are required to investigate whether genomic content differs across described factors such as disease and geography.

Newly generated complete genomes have superior assembly characteristics and cover phylogenetic diversity

For our large-scale genomic analysis of *R. gnavus*, we first established an isolate collection through extensive culturing efforts and by collecting available isolates, from which we sequenced the genome of 45 isolates using PacBio circular consensus sequencing (CCS) to yield complete, circular genomes and potential extrachromosomal elements (Fig. 2, Methods; Supplementary Data 2). We next complemented these with 208 available MAGs for which sufficient metadata could be retrieved and short-read genome data of an additional 79 isolates (Methods). To obtain assemblies of optimal quality, we tested five long-read *de novo* assemblers and selected the result with the longest contig (Methods, Supplementary Fig. 3, Supplementary Data 3). We also comprehensively analysed methylation patterns for the sequenced isolates (Supplementary Methods, Supplementary Fig. 4, Supplementary Data 4). Comparing the quality of these genomes, we observed that MAGs were worse in every aspect of genome assembly when compared to isolate assemblies (Fig. 2a). While total length

and number of genes were lower for MAGs as expected, GC content clearly differed between MAGs and isolate genomes, suggesting that current MAG binning techniques may fail to capture AT-rich regions. We further observed that isolates that underwent PacBio CCS were often assembled into single circular contigs, in contrast to a mean of 107 (±58.4 standard deviation) contigs per short-read isolate genome. Additionally, we found four circular extrachromosomal elements predicted to be plasmids with 99.9% confidence (Fig. 2a,b, Supplementary Fig. 5), demonstrating the added value of PacBio CCS. These four putative plasmids comprise two different large sequences of 191kb and 164kb, which derived from two distinct isolates from healthy individuals (i.e., QRD006, QRD009 and QRD010 contain one plasmid, QRD011 the other), and have not been described in *R. gnavus* to date. The plasmids are modular and highly related, that is, they are identical except for one gene cluster that is missing from the shorter 164kb plasmid (Supplementary Fig. 5a). They do not contain evident predicted antibiotic resistance or virulence genes (Methods). The plasmids are likely conjugative or mobilizable based on identified putative transposase genes which is consistent with their geographically distinct origins (USA and Japan). The plasmids contain a putative ParABS segregation system, annotated as 'Soj' (ParA) and 'ParB domain containing protein' (ParB). A key feature is a (hypothetical) non-ribosomal protein synthesis (NRPS) cluster with no known homologs (Supplementary Fig. 5b). However, upstream of it we identified with moderate confidence a transcription factor binding site for CatR, an H_2O_2-responsive repressor.

Leveraging our large genome collection, we then investigated the phylogenetic diversity of *R. gnavus* (Fig. 2c). This revealed no continent or genome source-specific clustering, but importantly, demonstrated that our *R. gnavus* isolate collection captures the full breadth of phylogenetic diversity across the tree (Fig. 2c).

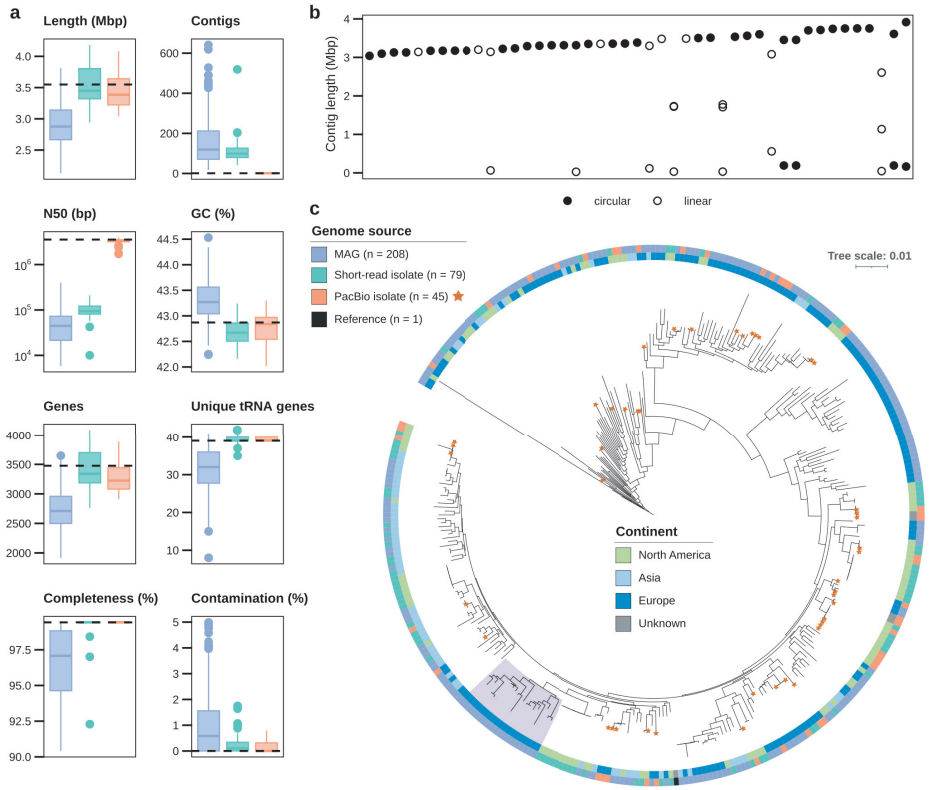


Fig. 2. Newly generated complete genomes have superior assembly characteristics and cover phylogenetic diversity. a We collected both publicly available short-read-based genomes from isolates and metagenome-assembled genomes (MAG), as well as long-read genomes generated from isolates in this study using PacBio HiFi sequencing and compared them to the one reference genome from NCBI GenBank (accession number GCF_009831375.1). Assembly statistics of each group of genomes are compared to the reference genome, shown as dashed line. Thick lines indicate medians, boxes represent first and third quantile and whiskers indicate the rest of the data excluding outliers; outliers are shown as separate dots. Color legend is shared with **c**. **b** Length and circularity of *de novo* assembled contigs from PacBio HiFi reads. **c** Maximum likelihood phylogenetic tree based on concatenated core genes. Each genome is annotated with its corresponding genome source and continent of origin. Stars mark genomes sequenced with PacBio newly added in this work. The gray shaded area marks the infant-associated clade that contains 8/10 MAGs with flagellum genes. Source data are provided as a Source Data file.

R. gnavus motility possibly restricted to infant-derived strains

In order to characterize the functional capacity of *R. gnavus*, we annotated our genomes with functional orthologs, modules and pathways (from KEGG[24]) and used linear modelling to identify associations between microbial functions and metadata. Using this methodology, we observed flagellum biosynthesis exclusively in newborns and infants up to 1 year of age, and this association was also statistically significant ($p = 0.008$). We further investigated flagellum biosynthesis together with chemotaxis, as these are

functionally closely related, and found both pathways in ten out of 333 genomes. These ten genomes are all MAGs originating from newborns and infants up to 1 year of age (Supplementary Fig. 6a) and contained (almost) complete operons (Supplementary Fig. 6b). To ensure this finding was not a technical assembly artefact, we traced the origin of these genomes, which revealed that these MAGs derive from infants sampled in three studies and five geographically separated locations (Estonia, Finland, Italy, Russia, and Sweden). Eight out of ten genomes with flagellum genes belong to a phylogenetic clade that is associated with newborns and infants (17/19 genomes in that clade derive from infants of 1 year old or younger; Fig. 2c, clade highlighted in gray), suggesting that motility might be associated with a specific infant-associated clade of *R. gnavus*. The absence of isolates in this clade precludes experimental verification of flagellum functionality, but strain differences in flagella and motility have been described[9].

We also screened all genomes for antibiotic resistance genes and found that resistance against tetracycline is the most common among *R. gnavus* (75/125 isolates; Supplementary Fig. 7). A minority of genomes contains resistance genes against aminoglycosides (n = 19), chloramphenicol (n = 8), trimethoprim (n = 11), lincosamide/macrolide (n = 24), and one and two genomes contain genes related to beta-lactamase and streptothricin resistance, respectively. For selective culturing of *R. gnavus* we therefore deem tetracycline the most helpful and *in vitro* validation confirmed that at least isolates containing the *tet(O)* and/or *tet(40)* genes, which account for the majority of the observed tetracycline resistance determinants, indeed have increased minimum inhibitory concentrations compared to isolates without *tet* gene (Supplementary Data 5).

Genomic differences between isolates from healthy and Crohn's indicates a Crohn's-specific subspecies

To evaluate whether CD-derived *R. gnavus* isolates genomically differ from healthy-derived isolates, we first placed our genomes into a core genome-based phylogenetic tree (Fig. 3a). As this tree contains practically identical isolates derived from the same person, we also constructed a tree of deduplicated genomes to facilitate statistical testing (Supplementary Fig. 8). This revealed three main clades with a strong enrichment of Crohn's-derived isolates in the two more basal clades (Fisher's exact test, $p = 3.7 \times 10^{-4}$, OR = 12.1 [2.5-69.8]). As our phylogenetic tree was reconstructed from only the core genome, we next performed whole-genome ANI analysis and accessory genome comparisons to also assess differences in the other genomic loci, which resulted in a highly similar clustering (Fig. 3a,b). As all *R. gnavus* genomes included here share at least 95% similarity with one another, which is often considered the species boundary[25,26], we consider that these clades represent subspecies. Together, these results demonstrate that at *R. gnavus* isolates from CD patients are often but not always genomically distinct from isolates from healthy controls based both on their core and accessory genome. The phylogeny indicates that most healthy-derived isolates form a

monophyletic subspecies clade, while the CD isolates appear polyphyletic and may be categorized into multiple groups.

Host phenotypes cannot be explained by previously identified putative virulence factors in *R. gnavus*

A previous study established that *R. gnavus* can secrete a glucorhamnan polysaccharide with pro-inflammatory properties[15]. However, another study found the putative gene cluster encoding the production machinery for this polysaccharide varied between strains, but direct comparison was not possible with short-read sequencing data[27]. Such insights into genomic variations may be crucial to understand immunogenicity of different isolates, motivating a more detailed analysis of this gene cluster and other genes with similar putative functions. We therefore tested whether previously suggested *R. gnavus* virulence factors could explain the association with CD (Supplementary Fig. 8). First, we observed that four genes or gene clusters (superantigens, tryptophane decarboxylase, bilirubin reductase and selenium-dependent xanthine dehydrogenase) were present in all complete *R. gnavus* genomes and are therefore part of the core genome (Fig. 3a). While we saw variation in several other gene clusters (glucorhamnan-producing gene cluster, Fisher's exact test, $p = 0.19$; and the *nan* gene cluster, $p = 0.35$), only one, namely the capsular polysaccharide gene (*cps*) cluster was associated with the distinction and was detected exclusively in isolates from the healthy-associated clade ($p = 8 \times 10^{-4}$). In conclusion, only the *cps* cluster, that leads to a more tolerogenic immune response[16], could distinguish host phenotype groups.

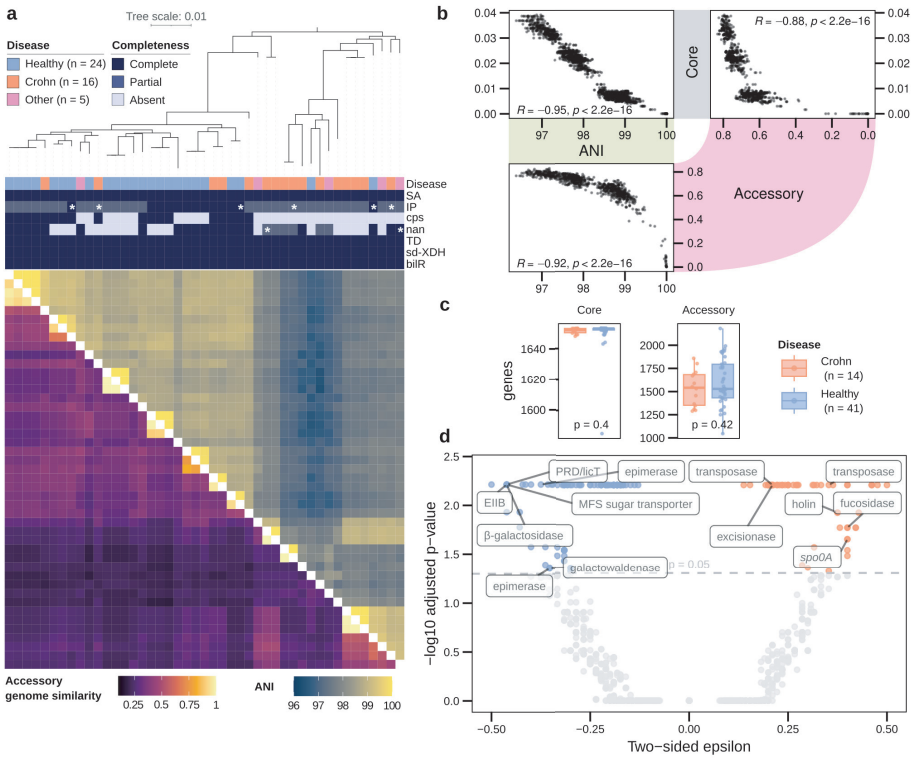


Fig. 3. Genomic differences between isolates from healthy and Crohn's indicates a Crohn's-specific subspecies. a Using our newly generated PacBio genomes, we compared genomes of isolates from healthy people to isolates from CD patients. Maximum likelihood phylogenetic tree of PacBio isolate genomes using concatenated core genes, with annotation of disease status and genes and gene clusters described previously in literature. Asterisks indicate gene clusters from genomes that are highlighted in Supplementary Fig. 9. Below are heatmaps of pairwise average nucleotide identity (ANI) and accessory genome similarity (calculated as 1 / binary distance). SA: superantigen (2 genes), IP: inflammatory polysaccharide (23 genes, 'partial' = 20 or 21 genes), cps: capsular polysaccharide (20 genes), nan: sialic acid metabolic cluster (11 genes, 'partial' = 6 genes), TD: tryptophane decarboxylase (1 gene), sd-XHD: selenium-dependent xanthine dehydrogenase (1 gene), bilR: bilirubin reductase (1 gene). **b** Comparison of genome comparison metrics core genome phylogenetic distance, average nucleotide identity and accessory genome binary distance tested with Spearman correlations. $P < 2.2 \times 10^{-16}$. **c** Comparison of core and accessory genome size between deduplicated isolate genomes with a CD or healthy phenotype, derived from short-read or long-read sequencing. Box plots represent median values with first and third quartile, whiskers indicate the rest of the data excluding outliers, and overlayed dots (jitter) show individual values. P-values were calculated using two-sided Wilcoxon rank-sum test. Core genome: p = 0.4, accessory genome: p = 0.42. **d** We compared accessory genomes of isolates from healthy people and CD patients using a bacterial GWAS to identify genes associated with disease phenotype. Results are expressed as false discovery rate-adjusted p-value (using the Benjamini-Hochberg correction) and epsilon, which is a measure of association strength between phenotype and genotype based on the (maximum likelihood) phylogenetic tree. The gray dashed line indicates a p-value of 0.05, anything above the line is

considered statistically significant. Positive values of epsilon correspond to an enrichment in CD and negative epsilon values are associated with a healthy host phenotype. P- and epsilon-values are adapted from the synchronous GWAS model as implemented in Hogwash. Source data are provided as a Source Data file.

Genomic architecture of gene cluster producing the proinflammatory polysaccharide glucorhamnan reveals genomic variations

Previous studies have highlighted the relevance and genomic architecture of the gene cluster producing inflammatory glucorhamnan based on complete, intermediate, or limited short-read coverage[27]. Here, we re-examined in our diverse collection of complete genomes if these clusters derive from the same genomic locus and are likely to be homologous (Supplementary Fig. 9). Compared to the isolate in which the gene cluster was experimentally verified (QRD039 = RJX 1121)[15], we saw variations in multiple genes, including several glycosyltransferases (Supplementary Fig. 9a). We observed 13 out of 45 long-read genomes to have the complete original cluster as identified in RJX1121, while 30 genomes had 20/23 genes as annotated in NZ_AAYG02000032.1 and two had 21/23 genes (those with 20 or 21 hits are subsequently called 'partially complete')[15,27]. These genomes lacked the same genes: a glycosyltransferase (RUMGNA_03519; present in the two genomes with 21 genes found), a transporter (RUMGNA_03522) and a polyphosphoglycerol synthesis gene (RUMGNA_03523). These partially complete cluster variants lack the genes in positions that were reported to have low coverage and we think they are therefore the same as those described in Sorbara *et al.*, 2020 as 'intermediate coverage'. To elucidate whether these genomes contain a truly different gene cluster at a different genomic location, the flanking genes were determined to map the genomic neighborhood. All investigated genomes had the same neighboring genes, thereby revealing a conserved genomic locus. (The 3' and 5'-flanking genes are annotated as 'HPr family phosphocarrier protein' and 'glutamine-fructose-6-phosphate transaminase'.) By closer inspection of the genomic loci, we found that the operon lacking RUMGNA_03519, RUMGNA_03522 and RUMGNA_03523 had other genes inserted instead (Supplementary Fig. 9a). Moreover, the variability at protein level compared to the reference gene (30-70% identity) suggests that this whole locus may be subject to positive selection or adaptation pressure. Nevertheless, based on similarity in genomic architecture we expect that all these strains still produce polysaccharides, although it remains to be established whether all of them induce pro-inflammatory effects.

A similar comparative genomics analysis for the *nan* gene cluster, responsible for releasing 2,7-anhydro-Neu5Ac from mucin[20], showed some genomes with *nan*-like genes in a different locus (Supplementary Fig. 9b). All these alternative *nan*-like clusters had the same genomic architecture, which importantly lacked the *nanH* (intramolecular trans sialidase) gene, suggesting that this partial cluster does not confer the same function. Together, these data show that strain differences across functionally relevant gene clusters are common, indicating that statements regarding virulence of R. gnavus based on single isolates should be interpreted with caution. Our collection of well-

characterized isolates allows researchers to assess the relevance of strain differences in future experiments.

GWAS reveals genes related to healthy or Crohn's-associated phenotype

In order to find genes that could explain differences in genomic repertoire of Crohn's-derived versus healthy-derived isolates, we conducted a bacterial GWAS using Hogwash, which incorporates genomic relatedness information (Methods). On a technical note, we confirmed high correlation between core and accessory genomes (Fig. 3b), and high pangenome size similarity between the Crohn's-associated and healthy-associated groups (Fig. 3c). We deemed including MAGs for this analysis to be inappropriate, as both the core and accessory genome of MAGs are substantially smaller than that of isolates (Supplementary Fig. 10, $p < 2 \times 10^{-16}$). Thus, their inclusion may increase false negatives or otherwise lead to spurious results.

Our bacterial GWAS analysis revealed 163 genes that were robustly associated with Crohn's isolates (FDR < 0.05, stricter synchronous model) through a high epsilon value, which quantifies the correlation between genotype and phenotype (Fig. 3d)[28]. We visualised and counted the presence of these genes in all *R. gnavus* genomes to better understand their possible correlation with host phenotype (Supplementary Fig. 11,12). Among the genes enriched in Crohn's-derived isolates we found nineteen genes related to mobile genetic elements (transposases and excisionases), a predicted fucosidase which might be involved in cleaving off the terminal fucose residue on mucin, a response regulator that Bakta annotated as '*spo0A*', and a holin gene (Supplementary Data 6). We screened the consensus sequence of this putative fucosidase gene for CAZyme domains (Methods) to gain more functional insight and indeed found a GH29 domain encoding a fucosidase. We also compared fucosidase domains between Crohn's and healthy isolates using CAZyme annotations for GH29 and GH95 (CAZymes with known fucose-cleaving functionality off mucin molecules), but found no significant differences (Wilcoxon rank sum test, $p = 0.098$ and $p = 0.39$, respectively; Supplementary Fig. 13). On the other hand, healthy-derived isolates were especially enriched for galactosidases and other genes involved in sugar metabolism (Fig. 3d, Supplementary Data 5). Taken together, we find novel gene-phenotype associations and provide a set of candidate genes for follow-up research on the role of *R. gnavus* in CD.

Discussion

Host phenotype-microbe association studies are often restricted to single diseases, age groups and geographic regions, which has also been the case for *R. gnavus*[12,13]. In this work we provide a detailed, global image of both the relative abundance and prevalence of *R. gnavus*, while we also investigate genomic variation within *R. gnavus*

isolates in depth. In both aspects, this is to our knowledge the largest investigation to date. Key findings are the remarkably high relative abundance in newborns and young infants (Fig. 1f), which is inversely associated with breastfeeding (Supplementary Fig. 2), and the increased prevalence and abundance of *R. gnavus* in Westernized populations (Fig. 1c,d). Given the robust associations of increased relative abundance of *R. gnavus* with several inflammatory diseases and allergies, many of which have high incidence in high-income countries and have their incidences rapidly increasing in newly industrialized countries[29-32], this begs the question of whether *R. gnavus* can have detrimental immunogenic effects on the host and whether this is strain-dependent. We show extensive genetic variation between strains in immunomodulating gene clusters, and our genetically well-characterized isolate resource can be used for experimental validation of differences in immunogenicity. The high prevalence of *R. gnavus* across both healthy and diseased individuals suggests that the consequences of being colonized with *R. gnavus* per se are unlikely exclusively negative, prompting the question if disease-associations become apparent when distinguishing *R. gnavus* strains. This hypothesis is in line with what we observed in clustering of our isolate genomes (Fig. 3a), where we see that isolates deriving from healthy individuals generally cluster apart from those isolated from Crohn's patients. Indeed, there have also been examples in literature of a positive health influence of *R. gnavus*, for example with healthy weight gain in undernourished children[33]. It would therefore be crucial that future intervention studies using *R. gnavus* determine if the used isolates belong to a healthy-associated or disease-associated clade.

In the past decade MAGs have been increasingly used in large-scale gut bacterial genomics studies[34-38], especially because culturing of specific gut bacteria can be highly laborious and challenging. While these MAGs have led to important biological advances, we show here that even high-quality MAGs (as defined by international standards[39]) remain of substantially worse quality than isolate genomes in multiple aspects (lower genome size and missing genes, higher GC content, amongst others, Figure 2A)[40]. In case of bacterial GWAS analyses, which aims to associate bacterial genes or genomic features with a phenotype of interest, including MAGs may therefore lead to biases and spurious associations caused by (non-)randomly missing genes due to binning and assembly artifacts. Extrachromosomal elements such as plasmids are generally not represented in MAGs, as they cannot be confidently binned, while these may be the most relevant in connection to disease and treatment options[41,42].

Through bacterial culture combined with PacBio CCS, we have generated high-quality genome data that lead to novel insights into *R. gnavus* biology. Two aspects that highlight this are the identification of large plasmids and a conserved methylated sequence motif. To date, only one 7kb-long plasmid of *Ruminococcus gnavus* is described in GenBank (accession number NZ_CP084015.1)[43]. The two related novel plasmids we identified in the present study are much larger (164kb and 191kb; Supplementary Fig. 5) and likely conjugative,

indicating a diversity of plasmids in *R. gnavus* that is of yet underexplored. The methylated DNA motifs that are identified here are different from those known so far (http://rebase.neb.com/cgi-bin/pacbioget?10929; Supplementary Fig. 4)[44], in line with the high variability in motifs we found per genome. Nevertheless, we find a single m^4C-methylated motif that is almost universally conserved across *R. gnavus* genomes (VNNVNCTGVNCAN). These results are reminiscent of those described for *Clostridioides difficile*[45].

We demonstrated that *R. gnavus* is a polyphyletic species, divided into multiple (genotypically and phenotypically distinct) subspecies clades. Notably, Crohn's-derived isolates were overrepresented in specific phylogenetic groups, while previously suggested virulence factors could not explain this separation. This suggests that these virulence factors may not play a significant role in CD symptomatology. Instead, by bacterial GWAS we identified 163 genes that could be targets for experimental validation of their role in CD development (Fig. 3, Supplementary Data 5). Among these genes are 56 that we find overrepresented in CD. However, we advise further validation of these genes in larger numbers of Crohn's-derived *R. gnavus* genomes before conducting laborious in vitro or in vivo experiments. Validations with the currently available data indicate that some presumable Crohn's-associated genes are also common among *R. gnavus* derived from healthy people. We listed the more noticeable candidates for which functions could be predicted. The most striking candidate is a putative fucosidase gene, as this could be directly involved in relevant cellular processes such as cell adhesion and immune system regulation[46]. Secondly, we hypothesize that genomic rearrangements and horizontal gene transfer may play an important role in the evolution of CD-associated *R. gnavus*, given the enrichment of predicted transposase and excisionase genes. Thirdly, we find a predicted holin gene which, although highly speculative, might play a role in suppressing competing bacteria[47]. A previous study identified 199 IBD-specific genes[2], based on a pangenome of 17 draft genomes. Those draft genomes include multiple IBD-related strains and genomes from the type strain, which we find to be phylogenetically distant in our core genome phylogeny based on a pangenome of 333 genomes. This increase in genome number in the current work particularly expands the accessory genome, where the largest differences in functionality are expected. Both the previous report and our results indicate predicted functional differences in e.g. mobile elements such as transposases and (putative) mucus utilization genes underscoring the robustness of the results and narrowing down the set of target genes for IBD-specific research[2]. Furthermore, IBD research on *R. gnavus* could benefit from considering the host and possible complex host-microbe interplay for the proposed virulence factors. For example, in antibiotic-treated mice the genetic background determined whether *R. gnavus* would ameliorate or exacerbate colitis[48].

In conclusion, we present one of the largest collections of complete genomes and associated extrachromosomal elements of any gut microbe not usually causing acute infection[49], and provide important novel biological insight into the global epidemiology

and genomic variation of *R. gnavus*. *R. gnavus* has an ambiguous relationship with human health[50], and different strains may exert different effects on their host. Our resource of complete genomes and isolates opens promising avenues for experimental validation and further bioinformatic scrutiny, and we expect this to be valuable to the broad gut microbiome research community.

Materials and methods

Assessing prevalence and abundance of *R. gnavus* across human populations
We used the publicly available 'curatedMetagenomicData' (version 3.6.2) resource to screen 21,030 fecal metagenomes from 86 studies on all habitable continents for the prevalence and abundance of *R. gnavus*[21]. We used R (version 4.0.2; https://www.R-project.org/) to interrogate this dataset and calculate statistical parameters. We focused our analyses on metagenomes with a sequencing depth of at least five million reads and retained only the first sample per subject ID, after which 12,791 samples remained. We used the accompanying curated metadata to assess prevalence and abundance among healthy individuals across age, geography, lifestyle, and health states (Supplementary Data 1). Prevalence of *R. gnavus* was compared using logistic regression. Relative abundances were compared after adding a pseudocount of 1.3×10^{-5}, followed by log-transformation and multivariable linear modelling. To identify suitable variables for logistic and linear models, we calculated collinearity between variables using Variance Inflation Factors (VIF) using the 'vif()' function from the 'car' package. VIF values above 2 were excluded by removing age (in years) and country from the models, leaving disease, age category, gender and westernization included as informative variables. Rows with missing values were discarded when building the models. For the final models, the association with each variable to *R. gnavus* prevalence or abundance was tested with Chi-square using the 'drop1()' R function. For infants, linear models were built using the same approach, including the variables feeding_practice, born_method and antibiotics_current_use. Correlation between feeding practice and age under or over half a year were tested using Chi-square. Sequencing depth (number of reads) was also log-transformed and compared using parametric t-test. P-values ≤ 0.05 were considered significant. To compare differences in *R. gnavus* prevalence in relation to sequencing depth, we divided all Westernized and non-Westernized metagenomes in ten equal groups (quantiles) based on sequencing depth (number of reads). Relative abundances of *R. gnavus* are shown as quantiles, as adapted from previous publications[51,52].

Mapping the distribution of *R. gnavus* across environments

To map the spread of *R. gnavus* across different environments, we searched publications and online resources that link the presence of *R. gnavus* to an environment or biome. *R. gnavus* has been described to reside in the intestinal tract of different animals: cats and dogs[10], chickens[53], lambs[54], rodents and pigs[11], and cattle[55]. Furthermore, we have

downloaded and screened the dataset related to the 2022 Microbiome publication by Ruscheweyh and colleagues to visualize prevalence and abundance of *R. gnavus* (Supplementary Fig. 14)[56].

Collection and curation of publicly available genome datasets

To compose a collection of *R. gnavus* metagenome-assembled genomes (MAGs) and isolate genomes, we queried a large, recent collection of gut MAGs[34]. Here, we specifically selected high-quality (HQ) MAGs annotated as *Ruminococcus gnavus* or its synonym *Faecalicatena gnavus* (with completeness > 90% and contamination < 5%)[39]. As the metadata from Almeida *et al.* does not contain curated information on disease status of the individual and this is of prime interest to our study[34], we matched identifiers to those present in the curatedMetagenomicData package. HQ-MAGs were only included if at least both disease status and geographic origin of the original sample could be traced back. This led to a collection of 201 HQ *R. gnavus* MAGs with associated metadata.

In order to obtain additional isolate genomes to complement the MAG collection, we queried the NCBI database in December 2021 and associated metadata to retrieve at least information on disease status and geographic origin of the isolate, like the HQ-MAGs. This yielded an additional 65 *R. gnavus* isolate genomes, which all originated from China or the USA. Furthermore, we included the type strain as reference genome (ATCC 29149, accession number GCA_009831375.1)[2,27,57].

Metagenome-assembled genome generation from fecal metagenomes derived from multiple recurrent *Clostridioides difficile*-infected patients

We used an in-house metagenomic dataset of multiple recurrent *Clostridioides difficile*-infected patients to generate seven additional HQ *R. gnavus* MAGs – the metagenomic data of which are available in the European Nucleotide Archive under project number PRJEB44737[58]. To produce high-quality metagenome-assembled genomes (MAGs), we adapted a previously published protocol[59].

The workflow is available as Snakemake[60] on Zenodo (https://doi.org/10.5281/zenodo.14628195) and works as follows. Raw metagenomics sequencing reads, from which human reads had already been removed, were preprocessed using fastp (version 0.20.1, parameters: '--cut_right --cut_window_size 4 --cut_mean_quality 20 -l 75 --detect_adapter_for_pe -y') to trim low-quality ends, remove reads shorter than 75 bases, remove adapter sequences and remove low-complexity reads[61]. (Note: preprocessing is not part of the workflow as described on Zenodo.) Remaining, high-quality reads were assembled into scaffolds using metaSPAdes (version 3.15.4, parameters: '--only-assembler')[62]. Scaffolds were binned with metaWRAP[63] (version 1.3.2) using three binning tools: MaxBin2[64] (version 2.2. 6), MetaBAT2[65] (version 2.12.1) and CONCOCT[66] (version 1.0.0) using a minimum contig length of 2500bp ('-l' option). Bins were then

refined using metaWRAP's 'bin_refinement' function, which uses CheckM[67] (version 1.0.12) to assess bin quality, setting completeness and contamination cut-offs of 75% and 10%, respectively ('-c' and '-x' options). After refinement, bins were reassembled using metaWRAP's 'reassemble_bins' function with assemblers MEGAHIT[68] (version 1.1.3) and metaSPAdes (version 3.13.0), again setting the minimum completeness to 75% and contamination to 10%, and the minimum length to 2000 ('-l' option). The resulting refined and reassembled bins were classified with the Genome Taxonomy Database toolkit (GTDB-Tk; version 2.1.0)[69]. Bins classified as *Ruminococcus gnavus* with >90% completeness and <5% contamination were included for further analyses.

Culturing of *R. gnavus* from feces of healthy donors and patient material

We ordered *R. gnavus* strain H2_28 (DSM number 108212) from the German Collection of Microorganisms and Cell Cultures (DSMZ, Braunschweig, Germany), resuspended it in Brain Heart Infusion broth (bioMérieux, Marcy-l'Étoile, France) and streaked it on Tryptic Soy agar +5% Sheep blood (TSS; bioMérieux) to isolate pure cultures. Two unique cultures (QRD001-QRD002) were isolated from feces by streaking on Columbia Naladixic acid Agar (bioMérieux; Supplementary Data 2). These were all cultured in an anaerobic cabinet (Whitley A35, Don Whitley Scientific Limited, UK) with an anaerobic gas mixture (10% H_2, 10% CO_2, 80% N_2) at 37°C. These samples were cultured from two different sample collections. First, healthy-derived isolates were obtained from donor faecal samples of Netherlands Donor Feces Bank donors and written informed consent was obtained for using these and clinical data, and approved by the Medical Ethics Committee at Leiden University Medical Center (P15.145). Second, CD-derived isolates from LUMC were obtained from fecal samples of patients aged above 18 years with a planned fistula surgery at LUMC and material was collected between July 2019 and June 2021. The study was approved by the Central Committee on Research involving Human Subjects and the local Medical Ethical Committee of the Leiden University Medical Center (study number P18.069). All patients gave written informed consent.

To further expand our *R. gnavus* genome collection, we cultured fourteen *R. gnavus* isolates from fecal samples of healthy feces donors that were available at Vedanta Biosciences (Supplementary Data 2). Human donor samples were obtained from both university hospitals and commercial sources. In all instances, informed consent language was reviewed and approved by the local ethics and regulatory authorities. Consent for the use of the sample was obtained from each subject. These were isolated and identified as follows: *R. gnavus* strains were isolated from various healthy donor stools by generating spore and non-spore fractions. Briefly, the non-spore fraction was generated by resuspending 1g of fecal material in 10mL sterile, pre-reduced PBS. The spore fraction was generated by adding 100% ethanol to the PBS fecal suspension to achieve a 50% (v/v) ethanol concentration. The fecal ethanol suspension was incubated at 25°C for 1hr while shaking. Following incubation, the fecal ethanol

suspension was centrifuged at 3400×g for 20 minutes and the cell pellet resuspended in 1mL of sterile, reduced PBS. Serial dilutions of the spore and non-spore fraction were plated on either Eggerth-Gagnon + 5% horse blood agar, Brucella Blood Agar (Anaerobe Systems, Inc., Morgan Hill, California, USA), MSAT (Anaerobe Systems), or chocolate agar and incubated at 37°C anaerobically for 72hr. Isolated colonies were identified by Sanger sequencing of the 16S amplicon using 8F and 1492R primers and Illumina shotgun sequencing. Isolated colonies were inoculated into 1.2mL of Peptone Yeast Extract Broth with Glucose (PYG; Anaerobe Systems) in a 96-deep well plate and incubated at 37°C anaerobically for 48hr. After incubation, colony identity was determined by performing PCR from 200µL of the culture using universal 16S primers 8F and 1492R. Selected isolates were then sub-cultured from the 96-deep well plate onto the appropriate agar medium and incubated at 37°C anaerobically for 72hr. An isolated colony from this plate was inoculated into 5mL of PYG and incubated at 37°C anaerobically for 24hr. 1mL of the culture was pelleted by centrifuging at 10000×g for 5 minutes. DNA was extracted from the pellet using the DNeasy blood and tissue kit (Qiagen, Hilden, Germany) following the manufacturer instructions. Colony identity was determined again by Sanger sequencing of the 16S gene amplicon using 8F and 1492R primers and Illumina shotgun sequencing.

Furthermore, fourteen isolates were cultured and collected at the University Medical Center Groningen as follows. Brucella blood agar medium (Mediaproducts BV, Groningen, The Netherlands) was used to cultivate the *R. gnavus* strains QRD024, QRD025 and QRD028 from human clinical specimens (Supplementary Data 2). QRD024, QRD025 and QRD028 were obtained from clinical samples and isolated bacteria were used for research purposes as no objections were raised by patients and no patient data was used. The plates were transferred to an anaerobic workstation (Whitley A45) after inoculation and incubated for one to three days at 37°C. The anaerobic medium YCFA supplemented with either apple pectin or porcine mucin type III (4.5 g/l) was used for the isolation of QRD026, QRD027, and QRD029-QRD031 as described earlier[70].. Fecal samples of healthy volunteers were used for inoculation on pre-reduced medium and the plates were incubated at 37^{o}C in an anaerobic chamber (Whitley A35 Workstation) with an anaerobic gas mixture (10% H_2, 10% CO_2, 80% N_2). The strains QRD032-QRD037 were isolated from fecal samples of IBD patients on either phenylethyl alcohol agar (Mediaproducts BV, Groningen, The Netherlands), brain heart infusion agar (Oxoid Limited, Cheshire, UK) supplemented with yeast (2,5 g/l), hemin (0,001% w/v) and cysteine (1 g/l) or YCFA medium supplemented with glucose (4.5 g/l). Ethical approval for collecting and using biological material was obtained as previously described for QRD026, QRD027 and QRD029-QRD037 (local ethics committee of the University Medical Center Groningen METc2014.236 and METc2014.291, respectively)[71]. Additional details on logistics and sample collection can be found in Plomp et al for QRD026, QRD027 and QRD029-QRD031[72], and in von Martels et al. (study was registered on ClinicalTrials.gov under NCT02538354) for QRD032-QRD037[71].

Moreover, isolates as cultured in their respective publications were obtained from the Broad Institute[15], and Sanger Institute[73]. All cultures from outside the Leiden University Medical Center (LUMC) were sent to the LUMC as frozen glycerol stocks and anaerobically cultured on TSS. After obtaining pure colonies, all isolates were independently confirmed to be *R. gnavus* in our laboratory using matrix-assisted laser desorption/ionization coupled to a time-of-flight mass spectrometer (MALDI-TOF; Bruker Daltonics GmbH, Bremen, Germany). All isolates were able to grow on TSS, CNA and Chocolate agar PolyViteX (bioMérieux) and the colony morphology appeared on plates as round, glassy white colonies with a bright white center. Sometimes colonies displayed concentric circles, reminiscent of checker game pieces.

Data processing of Illumina-sequenced *R. gnavus* isolates

The fourteen isolates cultured at Vedanta Biosciences were sequenced on the Illumina NextSeq platform using 150bp paired-end reads. These data were included with the isolate short-read-based genomes, increasing the number to 79 short-read isolates. Raw Illumina sequence data was cleaned and trimmed using fastp (v0.23.2) and sequence quality was inspected using Fastqc (v0.11.9; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and Multiqc[74] (v1.8). Cleaned reads were assembled by first using SKESA[75] (v2.4.0) and subsequently SPAdes (v3.15.3) with "--untrusted-contigs" and "--isolate" parameters.

Quality control and annotation of short-read-based genome collection

We have collected a total of 287 short-read-based genomes of *R. gnavus*, consisting of 79 assembled whole-genome sequences from cultured isolates and 208 metagenome-assembled genomes (MAGs). We also added the one available reference sequence in our analyses (NCBI GenBank accession number GCF_009831375.1). We filtered out contigs shorter than 1,000 bp using BBtools' reformat.sh (version 37.62; https://sourceforge.net/projects/bbmap/). We estimated completeness and contamination of all genomes using CheckM (version 1.0.13) and verified that all genomes taxonomically classify as *R. gnavus* using GTDB-Tk (version 2.1.0). Assembly length statistics were determined using QUAST[76] (version 5.0.2). Finally, genomes were annotated using Bakta[77] (version 1.6.1), which also provides the number of open reading frames, or predicted genes, per genome.

DNA isolation of *R. gnavus* isolates and generation of complete genomes using PacBio circular consensus sequencing

To generate complete genomes, 45 isolates were subjected to long read sequencing on the Pacific Biosciences (PacBio, Menlo Park, California, USA) Sequel IIe platform at the Leiden Genome Technology Center. To prepare high molecular weight total DNA, isolates were cultured anaerobically overnight in 10 mL BHI at 37°C. Cells from 5 mL of culture were pelleted and processed using the Qiagen Genomic-tip 100/G, according

to the manufacturer's instructions. SMRTbell® libraries were generated as follows. Genomic DNA was sheared with the Megaruptor 3 system (Diagenode LLC, Denville, New Jersey, USA) using 35 cycles. Libraries were generated according to the following manufacturer's procedure and checklist: Preparing whole genome and metagenome libraries using SMRTbell® prep kit 3.0 (PN 102-166-600 REV02 MAR2023), thereby using barcoded adapters. Size-selection was performed on library sub-pools using either diluted AMPure PBbeads (PacBio, 35% beads, 3.1x v/v ratio) or Blue Pippin (Sage Science, Beverly, Massachusetts, USA), depending on the insert-size of the libraries. The libraries were sequenced on a PacBio Sequel IIe platform with a 30 hour movie time using Sequel II Binding Kit 3.2 and Sequel II sequencing kit 2.0.

Long-read assembler mini-benchmark

Given the relative infancy of assembly algorithms for PacBio CCS data of microbial genomes, we performed a mini-benchmark of five long-read *de novo* assemblers: Canu[78,79] (version 2.2), Flye[80] (version 2.9.2), Raven[81] (version 1.8.1), Hifiasm[82] (version 0.19.6-r595) and IPA (version 1.8.0; https://github.com/PacificBiosciences/pbipa). In this benchmark, each assembler was provided 8 processor threads on the Shark high-performance computing cluster of the Leiden University Medical Center. Shark runs on Rocky Linux 8.7, with SLURM version 23.02.7. The available processors include Intel Xeon E5-2697, E5-2690 and E5-4650. Each assembler was provided as much memory as it needed to complete the assembly. The tools exhibited clear differences in number of contigs generated, processing time and memory use (Supplementary Fig. 3). Note that sample QRD034 was sequenced much deeper than the rest and subsampled to 30% of reads (= 277X coverage) to facilitate assembly. Contigs were taxonomically classified using the Contig Annotation Tool (CAT version 5.2.3)[83] to verify if they derived from *R. gnavus*. Canu, Flye, Hifiasm and IPA report if assembled contigs are linear or circular. From the different assemblies, we selected the assembly that yielded the longest contig and the longest total assembly length (all exceeding 3 Mb), giving Flye precedence as it provides the most extensive statistics (Supplementary Data 3). This resulted in 38 assemblies from Flye, 3 from Hifiasm, and 2 each from IPA and Raven. All contigs from selected assemblies were reoriented using dnaapler[84] (version 0.3.0) to start at the *dnaA*, *repA* or *terL* gene for chromosomes, plasmids and bacteriophages, respectively. Raven and Hifiasm produce assembly graphs, which were viewed to assess if contigs were linear or circular. Assemblies with a smaller secondary circular contig were analyzed with geNomad[85] (version 1.7.4) to predict the probability of it being a plasmid, using the built-in score calibration module with aggregated results from both the marker-based and neural net-based classifications.

We included two isolates derived from the strain DSMZ 108212, of which one we obtained directly from the DSMZ (QRD005) and the other was cultured at the Sanger Institute (QRD022). Assembly with Hifiasm yielded a 3.3Mb contig and a 28kb contig for QRD022, while QRD005 could not be resolved to less than three contigs, with

the longest being 2.4Mbp. These two assemblies were not completely identical and we decided to use a reference-based assembly of the unresolved one against the 3.3Mb contig using minimap2[86] (version 2.29) and samtools[87] consensus (version 1.19; parameters: '--min-MQ 5 --min-depth 10') to generate an improved assembly of QRD005. This resulted in two contigs of 3.3Mb and 178bp. We manually removed the 178bp fragment and use the single 3.3Mb contig assembly as representative of the 'DSMZ-108212' = QRD005 isolate (Supplementary Data 3).

Final genome assemblies were annotated with DNA methylation information from the PacBio SMRT Link Microbial Genome Analysis platform.

Antibiotic resistance screening of isolate genomes

To assess the genotypic antibiotic resistances in isolate genomes, we screened 79 short-read genome sequences of isolates, the 45 newly generated long-read genomes, and the one reference genome for the presence of antibiotic resistance genes using ABRicate (version 0.8.13; https://www.github.com/tseemann/abricate) with NCBI's AMRFinderPlus database (downloaded 11 November 2022, containing 5,735 sequences)[88]. Genes were assumed present if at least 95% of the gene matched with at least 95% identity to the gene in the database. For *in vitro* validation, ten isolates – five with *tet* tetracycline resistance genes and five without – were assessed for tetracycline minimum inhibitory concentrations (MIC at 48h) using an ETEST (bioMérieux) on TSS medium at 37°C in a Whitley A35 anaerobic cabinet. However, since we managed to isolate *R. gnavus* without the use of antibiotic selection and tetracycline resistance is also common among other human gut commensals, we did not pursue this further.

Search for previously described inflammatory factors of *R. gnavus*

Several *R. gnavus* genes have previously been associated with intestinal inflammation. We screened our collection of genomes for the presence of two superantigen genes (accession numbers WP_105084811.1 and WP_105084812.1)[17], 23 genes encoding the machinery to produce a proinflammatory (glucarhamnan) polysaccharide (NZ_AAYG02000032.1)[15], one tryptophane decarboxylase gene (RUMGNA_01526 from UniProt)[18], and 20 genes encoding a capsule polysaccharide (RUMGNA_02411 – RUMGNA_02392 from UniProt)[16]. We used protein BLAST[89] (blastp; version 2.13.0) to screen the genomes for the presence of each of these genes. Only hits that covered at least half of the gene of interest ('-qcov_hsp_perc 50') with an E-value of 1×10^{-20} or smaller ('-evalue 1e-20') were considered for further analysis. Gene clusters were considered present when all the genes were detected.

Using the same method, we also screened genomes for the presence of the bilirubin reductase gene (*bilR,* WP_009244284.1)[90], selenium-dependent xanthine dehydrogenase (*sd-XDH*, QHB24869.1)[19], and the *nan* cluster for sialic acid metabolism

(RUMGNA_02691 through RUMGNA_02701 from UniProt)[20]. Gene operons were visualized using clinker[91].

Annotation of functional pathway genes

We annotated carbohydrate-active enzymes (CAZymes) by comparing the genomes to dbCAN[92] (version 10) using HMMer[93] (version 3.3.2). Within the CAZyme families, we focused on two glycosyl hydrolase families that include fucosidases, GH29 and GH95, which have been described as important for mucus utilization[94], a main feature of *R. gnavus*. Genomes were also annotated using KEGG-Decoder[95]. Pathways for chemotaxis and flagellum biosynthesis were annotated using the KOALA definitions available online[24]. Moreover, genomes were screened for the presence of annotated biosynthetic gene clusters (BGC) using antiSMASH[96] (version 6.1.1).

Comparison of whole genomes to find clusters of genomic variants

Whole genomes were compared to one another using average nucleotide identity (ANI) with fastANI (version 1.33)[26]. Furthermore, genomes were subjected to a pangenome analysis using Panaroo (version 1.3.0; parameters '--clean-mode strict -a core --aligner mafft --core_threshold 0.95')[97]. For the pangenome, we considered genes that occur in at least 95% of genomes core genes as recommended when including MAGs[98]. The core genes were concatenated and using MAFFT[99] (version 7.505) a core genome multiple sequence alignment was generated, which was automatically trimmed using trimAl[100] (version 1.4.1). A maximum likelihood phylogeny was inferred from the trimmed multiple alignment using IQ-tree[101] (version 2.2.0.3), including ModelFinder Plus[102] to automatically select the best fitting evolutionary model and ultrafast bootstrap (1000 replicates) to calculate branch support[103]. The selected models were: short-read genomes GTR+F+I+I+R9; long-read genomes GTR+F+R7; all genomes GTR+F+R10. Trees were visualized in iTOL[104].

Bacterial genome-wide association study (GWAS)

To identify genes that are putatively associated with CD, we subjected genomes of *R. gnavus* isolates to a bacterial genome-wide association study using Hogwash (version 1.2.6; parameters: 'fdr = 0.05, bootstrap = 0.875, grouping_method = "post-ar" ')[28]. Hogwash implements a more stringent version of the homoplasy-based PhyC method introduced in 2013[105]. Hogwash reconstructs the evolutionary history of the genomes of interest using a phylogenetic tree and predicts where genotype and phenotype transitions occurred to assess where genotype and phenotype transitions coincide. We made use of the high correlation between core and accessory genome to use these two as input, together with phenotype of either CD or healthy. Genomes were assigned healthy or CD phenotype based on available metadata on health status from the person from whom the *R. gnavus* isolate was cultured. We included short-read sequencing isolate draft genomes as well as our in-house generated PacBio complete genomes. If multiple sequences of the same isolate existed, we deduplicated based

on ANI > 99.9%. Of these duplicates, we picked the first based on alphabetic order as representative, and we preferentially select long-read-based genomes when available. This resulted in fourteen *R. gnavus* isolate genomes derived from CD patients and 41 from healthy people (total N = 55). We used a matrix of (accessory) gene presence and absence generated by Panaroo as input for Hogwash. As phylogenetic tree, we pruned the tree of all *R. gnavus* genomes inferred by IQ-tree to include only this set of 55 deduplicated genomes and midpoint rooted the tree. Associations between genotype and phenotype are evaluated both by p-value indicating statistical significance, and epsilon value, which calculates the strength of genotype-phenotype association on a 0-1 scale (Supplementary Data 6).

To further validate the genes found to be significantly associated with either a healthy or Crohn's host phenotype, we counted the prevalence of each group of genes in both healthy-derived (n = 123) and IBD-derived MAGs (Crohn's n=8; ulcerative colitis n = 1; Supplementary Fig. 11). Furthermore, we visualised the prevalence of these genes among genomes, annotated by their host disease phenotype, as a heatmap to visually inspect the predicted gene associations (Supplementary Fig. 12).

Statistical analyses

All tools were run with default parameters unless stated otherwise. Statistical analyses and visualization were done in R (version 4.0.2) using RStudio (https://posit.co/). A p-value of 0.05 or smaller was considered significant. Data were visualised using the R package ggplot2 (version 3.5.0)[106], with the publication theme from ggembl (version 0.1.2; https://git.embl.de/grp-zeller/ggembl). Figures were polished manually using Inkscape (version 0.92.5; https://inkscape.org/).

Data availability

The long-read whole-genome sequencing data generated in this study and corresponding assemblies of isolates presented in this study are available from the European Nucleotide Archive under accession number PRJEB76407 (https://www.ebi.ac.uk/ena/browser/view/PRJEB76407). Raw metagenomic data used for additional MAG building, and the MAGs themselves, are available under accession number PRJEB44737 (https://www.ebi.ac.uk/ena/browser/view/PRJEB44737). A complete set of short and long-read genomes together with metadata, along with processed data, is available through Zenodo, https://doi.org/10.5281/zenodo.13907031. Isolates will be made available upon request to the corresponding author (q.r.ducarmon@lumc.nl). The use of biological materials for research purposes generated in this study by Vedanta Biosciences can be made available under a material transfer agreement. Correspondence should be sent to jnorman@vedantabio.com and legal@vedantabio.com and will addressed within 2 weeks. Source data are provided with this paper.

Code Availability

Scripts of both the whole-genome annotation and comparative genomics analyses, as well as further downstream and statistical analyses are available on Zenodo, https://doi.org/10.5281/zenodo.14628203. The code used to generate MAGs is also available on Zenodo (https://doi.org/10.5281/zenodo.14628195).

Acknowledgements

We thank all members of the scientific community that generously provided us with *R. gnavus* isolates. This study was supported by the Leiden University Fund / Dr. F.F. Hofman Fonds, (www.luf.nl) to QD. Further funding was provided by the LUMC (LUMC Fellowship to G.Z.), the Health + Life Science Alliance Heidelberg Mannheim through state funds approved by the State Parliament of Baden-Württemberg (postdoctoral fellowships to Q.D. and N.K.) and EMBO postdoctoral fellowship (ALTF 1030-2022 to Q.D.). The Graduate School of the Medical Sciences of the University of Groningen provided a grant to N.P. The NDFB (S.N. and E.M.T.) received an unrestricted research grant from Vedanta Biosciences.

Author contributions statement

Conceptualization, EK, WKS and QD; methodology, NP, IS, LS, AvdM, ET, JN, NK, RV, SK, HH, KF; investigation, SN, NP, RV, IS, LS, ML and QD; formal analysis, SN and QD; writing – original draft, SN and QD; writing – review & editing, all authors; supervision, SK, GZ, EK, WKS and QD; funding acquisition: QD

Competing interest statement

JN is an employee of Vedanta Biosciences Inc. The other authors report no competing interests.

Supplementary Material

Supplementary figures, methods and data are available online on the publisher's website:

https://www.nature.com/articles/s41467-025-56449-x#Sec28

References

1. VanEvery, H., Franzosa, E. A., Nguyen, L. H. & Huttenhower, C. Microbiome epidemiology and association studies in human health. *Nature Reviews Genetics* **24**, 109-124 (2023). https://doi.org:10.1038/s41576-022-00529-x

2. Hall, A. B. *et al.* A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Medicine* **9**, 103 (2017). https://doi.org:10.1186/s13073-017-0490-5

3. Grahnemo, L. *et al.* Cross-sectional associations between the gut microbe Ruminococcus gnavus and features of the metabolic syndrome. *The Lancet Diabetes & Endocrinology* **10**, 481-483 (2022). https://doi.org:10.1016/S2213-8587(22)00113-9

4. De Filippis, F. *et al.* Specific gut microbiome signatures and the associated pro-inflamatory functions are linked to pediatric allergy and acquisition of immune tolerance. *Nature Communications* **12**, 5958 (2021). https://doi.org:10.1038/s41467-021-26266-z

5. Wirbel, J., Essex, M., Forslund, S. K. & Zeller, G. Evaluation of microbiome association models under realistic and confounded conditions. *bioRxiv*, 2022.2005.2009.491139 (2022). https://doi.org:10.1101/2022.05.09.491139

6. Berland, M. *et al.* Both Disease Activity and HLA-B27 Status Are Associated With Gut Microbiome Dysbiosis in Spondyloarthritis Patients. *Arthritis & Rheumatology* **75**, 41-52 (2023). https://doi.org:10.1002/art.42289

7. Watanabe, N. *et al.* Clinical and microbiological characteristics of Ruminococcus gnavus bacteremia and intra-abdominal infection. *Anaerobe* **85**, 102818 (2024). https://doi.org:10.1016/j.anaerobe.2024.102818

8. Togo, A. H. *et al.* Description of Mediterraneibacter massiliensis, gen. nov., sp. nov., a new genus isolated from the gut microbiota of an obese patient and reclassification of Ruminococcus faecis, Ruminococcus lactaris, Ruminococcus torques, Ruminococcus gnavus and Clostridium glycyrrhizinilyticum as Mediterraneibacter faecis comb. nov., Mediterraneibacter lactaris comb. nov., Mediterraneibacter torques comb. nov., Mediterraneibacter gnavus comb. nov. and Mediterraneibacter glycyrrhizinilyticus comb. nov. *Antonie Van Leeuwenhoek* **111**, 2107-2128 (2018). https://doi.org:10.1007/s10482-018-1104-y

9. Moore, W. E. C., Johnson, J. L. & Holdeman, L. V. Emendation of Bacteroidaceae and Butyrivibrio and Descriptions of Desulfomonas gen. nov. and Ten New Species in the Genera Desulfomonas, Butyrivibrio, Eubacterium, Clostridium, and Ruminococcus. *International Journal of Systematic and Evolutionary Microbiology* **26**, 238-252 (1976). https://doi.org:https://doi.org/10.1099/00207713-26-2-238

10. Branck, T. *et al.* Comprehensive profile of the companion animal gut microbiome integrating reference-based and reference-free methods. *ISME Journal* (2024). https://doi.org:10.1093/ismejo/wrae201

11. Abdugheni, R. *et al.* Comparative genomics reveals extensive intra-species genetic divergence of the prevalent gut commensal Ruminococcus gnavus. *Microbial Genomics* **9** (2023). https://doi.org:10.1099/mgen.0.001071

12. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65 (2010). https://doi.org:10.1038/nature08821

13. Kraal, L., Abubucker, S., Kota, K., Fischbach, M. A. & Mitreva, M. The prevalence of species and strains in the human microbiome: a resource for experimental efforts. *PLoS One* **9**, e97279 (2014). https://doi.org:10.1371/journal.pone.0097279

14. Xu, T. *et al.* Microbiome Features Differentiating Unsupervised-Stratification-Based Clusters of Patients with Abnormal Glycometabolism. *mBio* **14**, e0348722 (2023). https://doi.org:10.1128/mbio.03487-22

15. Henke, M. T. *et al.* Ruminococcus gnavus, a member of the human gut microbiome associated with Crohn's disease, produces an inflammatory polysaccharide. *Proceedings of the National Academy of Sciences* **116**, 12672-12677 (2019). https://doi.org:10.1073/pnas.1904099116

16. Henke, M. T. *et al.* Capsular polysaccharide correlates with immune response to the human gut microbe Ruminococcus gnavus. *Proceedings of the National Academy of Sciences* **118** (2021). https://doi.org:10.1073/pnas.2007595118

17. Bunker, J. J. *et al.* B cell superantigens in the human intestinal microbiota. *Science Translational Medicine* **11** (2019). https://doi.org:10.1126/scitranslmed.aau9356

18. Zhai, L. *et al.* Ruminococcus gnavus plays a pathogenic role in diarrhea-predominant irritable bowel syndrome by increasing serotonin biosynthesis. *Cell Host & Microbe* **31**, 33-44 e35 (2023). https://doi.org:10.1016/j.chom.2022.11.006

19. Yan, Y. *et al.* Commensal bacteria promote azathioprine therapy failure in inflammatory bowel disease via decreasing 6-mercaptopurine bioavailability. *Cell Reports Medicine* **4**, 101153 (2023). https://doi.org:10.1016/j.xcrm.2023.101153

20. Bell, A. *et al.* Elucidation of a sialic acid metabolism pathway in mucus-foraging Ruminococcus gnavus unravels mechanisms of bacterial adaptation to the gut. *Nature Microbiology* **4**, 2393-2404 (2019). https://doi.org:10.1038/s41564-019-0590-7

21. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nature Methods* **14**, 1023-1024 (2017). https://doi.org:10.1038/nmeth.4468

22. Valles-Colomer, M. *et al.* The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* **614**, 125-135 (2023). https://doi.org:10.1038/s41586-022-05620-1

23. Shenhav, L. *et al.* Microbial colonization programs are structured by breastfeeding and guide healthy respiratory development. *Cell* **187**, 5431-5452 e5420 (2024). https://doi.org:10.1016/j.cell.2024.07.022

24. Tully, B. J. *KEGGDecoder KOALA definitions*, <https://github.com/bjtully/BioData/blob/master/KEGGDecoder/KOALA_definitions.txt> (2021).

25. Goris, J. *et al.* DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology* **57**, 81-91 (2007). https://doi.org:10.1099/ijs.0.64483-0

26. Jain, C., Rodriguez, R. L., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **9**, 5114 (2018). https://doi.org:10.1038/s41467-018-07641-9

27. Sorbara, M. T. *et al.* Functional and Genomic Variation between Human-Derived Isolates of Lachnospiraceae Reveals Inter- and Intra-Species Diversity. *Cell Host & Microbe* **28**, 134-146 e134 (2020). https://doi.org:10.1016/j.chom.2020.05.005

28. Saund, K. & Snitkin, E. S. Hogwash: three methods for genome-wide association studies in bacteria. *Microbial Genomics* **6** (2020). https://doi.org:10.1099/mgen.0.000469

29. Collaborators, G. B. D. I. B. D. The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet Gastroenterology and Hepatology* **5**, 17-30 (2020). https://doi.org:10.1016/S2468-1253(19)30333-4

30. Fogarty, A. W. What have studies of non-industrialized countries told us about the cause of allergic disease? *Clinical & Experimental Allergy* **45**, 87-93 (2015). https://doi.org:10.1111/cea.12339

31. Tian, J., Zhang, D., Yao, X., Huang, Y. & Lu, Q. Global epidemiology of systemic lupus erythematosus: a comprehensive systematic analysis and modelling study. *Annals of the Rheumatic Diseases* **82**, 351-356 (2023). https://doi.org:10.1136/ard-2022-223035

32. Gacesa, R. *et al.* Environmental factors shaping the gut microbiome in a Dutch population. *Nature* **604**, 732-739 (2022). https://doi.org:10.1038/s41586-022-04567-7

33. Blanton, L. V. *et al.* Gut bacteria that prevent growth impairments transmitted by microbiota from malnourished children. *Science* **351** (2016). https://doi.org:10.1126/science.aad3311

34. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology* **39**, 105-114 (2021). https://doi.org:10.1038/s41587-020-0603-3

35. Karcher, N. *et al*. Analysis of 1321 Eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. *Genome Biology* **21**, 138 (2020). https://doi.org:10.1186/s13059-020-02042-y

36. Karcher, N. *et al.* Genomic diversity and ecology of human-associated Akkermansia species in the gut microbiome revealed by extensive metagenomic assembly. *Genome Biology* **22**, 209 (2021). https://doi.org:10.1186/s13059-021-02427-7

37. Tett, A. *et al.* The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. *Cell Host & Microbe* **26**, 666-679 e667 (2019). https://doi.org:10.1016/j.chom.2019.08.018

38. Blanco-Miguez, A. *et al.* Extension of the Segatella copri complex to 13 species with distinct large extrachromosomal elements and associations with host conditions. *Cell Host & Microbe* **31**, 1804-1819 e1809 (2023). https://doi.org:10.1016/j.chom.2023.09.013

39. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* **35**, 725-731 (2017). https://doi.org:10.1038/nbt.3893

40. Meziti, A. *et al.* The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample. *Applied Environmental Microbiology* **87** (2021). https://doi.org:10.1128/AEM.02593-20

41. Castañeda-Barba, S., Top, E. M. & Stalder, T. Plasmids, a molecular cornerstone of antimicrobial resistance in the One Health era. *Nature Reviews Microbiology* **22**, 18-32 (2024). https://doi.org:10.1038/s41579-023-00926-x

42. Zorea, A. *et al.* Plasmids in the human gut reveal neutral dispersal and recombination that is overpowered by inflammatory diseases. *Nature Communications* **15**, 3147 (2024). https://doi.org:10.1038/s41467-024-47272-x

43. Tourlousse, D. M. *et al.* Characterization and Demonstration of Mock Communities as Control Reagents for Accurate Human Microbiome Community Measurements. *Microbiology Spectrum* **10**, e0191521 (2022). https://doi.org:10.1128/spectrum.01915-21

44. Blow, M. J. *et al.* The Epigenomic Landscape of Prokaryotes. *PLoS Genetics* **12**, e1005854 (2016). https://doi.org:10.1371/journal.pgen.1005854

45. Oliveira, P. H. *et al.* Epigenomic characterization of Clostridioides difficile finds a conserved DNA methyltransferase that mediates sporulation and pathogenesis. *Nature Microbiology* **5**, 166-180 (2020). https://doi.org:10.1038/s41564-019-0613-4

46. Li, J., Hsu, H. C., Mountz, J. D. & Allen, J. G. Unmasking Fucosylation: from Cell Adhesion to Immune System Regulation and Diseases. *Cell Chemical Biology* **25**, 499-512 (2018). https://doi.org:10.1016/j.chembiol.2018.02.005

47. Backman, T. *et al.* A phage tail-like bacteriocin suppresses competitors in metapopulations of pathogenic bacteria. *Science* **384**, eado0713 (2024). https://doi.org:10.1126/science.ado0713

48. Yu, S. *et al.* Paneth Cell-Derived Lysozyme Defines the Composition of Mucolytic Microbiota and the Inflammatory Tone of the Intestine. *Immunity* **53**, 398-416 e398 (2020). https://doi.org:10.1016/j.immuni.2020.07.010

49. Bartlett, A., Padfield, D., Lear, L., Bendall, R. & Vos, M. A comprehensive list of bacterial pathogens infecting humans. *Microbiology* **168** (2022). https://doi.org:10.1099/mic.0.001269

50. Crost, E. H., Coletto, E., Bell, A. & Juge, N. Ruminococcus gnavus: friend or foe for human health. *FEMS Microbiology Reviews* **47** (2023). https://doi.org:10.1093/femsre/fuad014

51. Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature Medicine* **25**, 679-689 (2019). https://doi.org:10.1038/s41591-019-0406-6

52. Wirbel, J. *et al.* Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biology* **22**, 93 (2021). https://doi.org:10.1186/s13059-021-02306-1

53. Li, Z. *et al.* Effects of herbal dregs supplementation of Salvia miltiorrhiza and Isatidis Radix residues improved production performance and gut microbiota abundance in late-phase laying hens. *Frontiers in Veterinary Science* **11**, 1381226 (2024). https://doi.org:10.3389/fvets.2024.1381226

54. Xiao, H. *et al.* The effect of early colonized gut microbiota on the growth performance of suckling lambs. *Frontiers in Microbiology* **14**, 1273444 (2023). https://doi.org:10.3389/fmicb.2023.1273444

55. Chen, Z. *et al.* Differences in meat quality between Angus cattle and Xinjiang brown cattle in association with gut microbiota and its lipid metabolism. *Frontiers in Microbiology* **13**, 988984 (2022). https://doi.org:10.3389/fmicb.2022.988984

56. Ruscheweyh, H. J. *et al.* Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. *Microbiome* **10**, 212 (2022). https://doi.org:10.1186/s40168-022-01410-z

57. Zou, Y. *et al.* 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nature Biotechnology* **37**, 179-185 (2019). https://doi.org:10.1038/s41587-018-0008-8

58. Nooij, S. *et al.* Fecal Microbiota Transplantation Influences Procarcinogenic Escherichia coli in Recipient Recurrent Clostridioides difficile Patients. *Gastroenterology* **161**, 1218-1228 e1215 (2021). https://doi.org:10.1053/j.gastro.2021.06.009

59. Saheb Kashaf, S., Almeida, A., Segre, J. A. & Finn, R. D. Recovering prokaryotic genomes from host-associated, short-read shotgun metagenomic sequencing data. *Nature Protocols* **16**, 2520-2541 (2021). https://doi.org:10.1038/s41596-021-00508-2

60. Molder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10**, 33 (2021). https://doi.org:10.12688/f1000research.29032.2

61. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890 (2018). https://doi.org:10.1093/bioinformatics/bty560

62. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Research* **27**, 824-834 (2017). https://doi.org:10.1101/gr.213959.116

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

64. 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). https://doi.org:10.1093/bioinformatics/btv638

65. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015). https://doi.org:10.7717/peerj.1165

66. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nature Methods* **11**, 1144-1146 (2014). https://doi.org:10.1038/nmeth.3103

67. 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 Research* **25**, 1043-1055 (2015). https://doi.org:10.1101/gr.186072.114

68. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3-11 (2016). https://doi.org:10.1016/j.ymeth.2016.02.020

69. 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 (2019). https://doi.org:10.1093/bioinformatics/btz848

70. Lopez-Siles, M. *et al.* Cultured representatives of two major phylogroups of human colonic Faecalibacterium prausnitzii can utilize pectin, uronic acids, and host-derived substrates for growth. *Applied Environmental Microbiology* **78**, 420-428 (2012). https://doi.org:10.1128/AEM.06858-11

71. von Martels, J. Z. H. *et al.* Riboflavin Supplementation in Patients with Crohn's Disease [the RISE-UP study]. *J Crohns Colitis* **14**, 595-607 (2020). https://doi.org:10.1093/ecco-jcc/jjz208

72. Plomp, N. *et al.* A convenient and versatile culturomics platform to expand the human gut culturome of Lachnospiraceae and Oscillospiraceae. *Benef Microbes*, 1-16 (2024). https://doi.org:10.1163/18762891-bja00042

73. Forster, S. C. *et al.* A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nature Biotechnology* **37**, 186-192 (2019). https://doi.org:10.1038/s41587-018-0009-7

74. Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-3048 (2016). https://doi.org:10.1093/bioinformatics/btw354

75. Souvorov, A., Agarwala, R. & Lipman, D. J. SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biology* **19**, 153 (2018). https://doi.org:10.1186/s13059-018-1540-z

76. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075 (2013). https://doi.org:10.1093/bioinformatics/btt086

77. Schwengers, O. *et al.* Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial Genomics* **7** (2021). https://doi.org:10.1099/mgen.0.000685

78. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* **27**, 722-736 (2017). https://doi.org:10.1101/gr.215087.116

79. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Research* **30**, 1291-1305 (2020). https://doi.org:10.1101/gr.263566.120

80. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **37**, 540-546 (2019). https://doi.org:10.1038/s41587-019-0072-8

81. Vaser, R. & Sikic, M. Time- and memory-efficient genome assembly with Raven. *Nature Computational Science* **1**, 332-336 (2021). https://doi.org:10.1038/s43588-021-00073-4

82. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170-175 (2021). https://doi.org:10.1038/s41592-020-01056-5

83. von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H. & Dutilh, B. E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biology* **20**, 217 (2019). https://doi.org:10.1186/s13059-019-1817-x

84. Bouras, G., Grigson., S., Papudeshi., B., Mallawaarachchi V., Roach, M. J. *Dnaapler: A tool to reorient circular microbial genomes* <https://github.com/gbouras13/dnaapler> (2023).

85. Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nature Biotechnology* (2023). https://doi.org:10.1038/s41587-023-01953-y

86. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018). https://doi.org:10.1093/bioinformatics/bty191

87. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009). https://doi.org:10.1093/bioinformatics/btp352

88. Feldgarden, M. *et al.* AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Scientific Reports* **11**, 12728 (2021). https://doi.org:10.1038/s41598-021-91456-0

89. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). https://doi.org:10.1186/1471-2105-10-421

90. Hall, B. *et al.* BilR is a gut microbial enzyme that reduces bilirubin to urobilinogen. *Nature Microbiology* **9**, 173-184 (2024). https://doi.org:10.1038/s41564-023-01549-x

91. Gilchrist, C. L. M. & Chooi, Y. H. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* **37**, 2473-2475 (2021). https://doi.org:10.1093/bioinformatics/btab007

92. Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research* **46**, W95-W101 (2018). https://doi.org:10.1093/nar/gky418

93. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Computational Biology* **7**, e1002195 (2011). https://doi.org:10.1371/journal.pcbi.1002195

94. Berkhout, M. D., Plugge, C. M. & Belzer, C. How microbial glycosyl hydrolase activity in the gut mucosa initiates microbial cross-feeding. *Glycobiology* **32**, 182-200 (2022). https://doi.org:10.1093/glycob/cwab105

95. Graham, E. D., Heidelberg, J. F. & Tully, B. J. Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME Journal* **12**, 1861-1866 (2018). https://doi.org:10.1038/s41396-018-0091-3

96. Blin, K. *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research* **49**, W29-W35 (2021). https://doi.org:10.1093/nar/gkab335

97. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology* **21**, 180 (2020). https://doi.org:10.1186/s13059-020-02090-4

98. Li, T. & Yin, Y. Critical assessment of pan-genomic analysis of metagenome-assembled genomes. *Briefings in Bioinformatics* **23** (2022). https://doi.org:10.1093/bib/bbac413

99. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059-3066 (2002). https://doi.org:10.1093/nar/gkf436

100. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009). https://doi.org:10.1093/bioinformatics/btp348

101. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530-1534 (2020). https://doi.org:10.1093/molbev/msaa015

102. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**, 587-589 (2017). https://doi.org:10.1038/nmeth.4285