Integrative genomic reconstruction of carbohydrate utilization networks in bifidobacteria: global trends, local variability, and dietary adaptation

Aleksandr A. Arzamasov[1], Dmitry A. Rodionov[1], Matthew C. Hibberd[2,3,4], Janaki L. Guruge[2,3], Marat D. Kazanov[5], Semen A. Leyn[1], James E. Kent[1], Kristija Sejane[6], Lars Bode[6], Michael J. Barratt[2,3,4], Jeffrey I. Gordon[2,3,4], and Andrei L. Osterman[1]

[1]Infectious and Inflammatory Disease Center, Sanford Burnham Prebys Medical Discovery Institute, 10901 North Torrey Pines Rd, La Jolla, CA 92037, USA

[2]Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63110, USA.

[3]Center for Gut Microbiome and Nutrition Research, Washington University School of Medicine, St. Louis, MO 63110, USA.

[4]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110, USA.

[5]Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey, 34956

[6]Department of Pediatrics, Larsson-Rosenquist Foundation Mother-Milk-Infant Center of Research Excellence (MOMI CORE), and the Human Milk Institute (HMI), University of California San Diego, La Jolla, CA 92093, USA

ABSTRACT

Bifidobacteria are among the earliest colonizers of the human gut, conferring numerous health benefits. While multiple *Bifidobacterium* strains are used as probiotics, accumulating evidence suggests that the individual responses to probiotic supplementation may vary, likely due to a variety of factors, including strain type(s), gut community composition, dietary habits of the consumer, and other health/lifestyle conditions. Given the saccharolytic nature of bifidobacteria, the carbohydrate composition of the diet is one of the primary factors dictating the colonization efficiency of *Bifidobacterium* strains. Therefore, a comprehensive understanding of bifidobacterial glycan metabolism at the strain level is necessary to rationally design probiotic or synbiotic formulations that combine bacterial strains with glycans that match their nutrient preferences. In this study, we systematically reconstructed 66 pathways involved in the utilization of mono-, di-, oligo-, and polysaccharides by analyzing the representation of 565 curated metabolic functional roles (catabolic enzymes, transporters, transcriptional regulators) in 2973 non-redundant cultured *Bifidobacterium* isolates and metagenome-assembled genomes (MAGs). Our analysis uncovered substantial heterogeneity in the predicted glycan utilization capabilities at the species and strain level and revealed the presence of a yet undescribed phenotypically distinct subspecies-level clade within the *Bifidobacterium longum* species. We also identified Bangladeshi isolates harboring unique gene clusters tentatively implicated in the breakdown of xyloglucan and human milk oligosaccharides. Predicted carbohydrate utilization phenotypes were experimentally characterized and validated. Our large-scale genomic analysis considerably expands the knowledge of carbohydrate metabolism in bifidobacteria and provides a foundation for rationally designing single- or multi-strain probiotic formulations of a given bifidobacterial species as well as synbiotic combinations of bifidobacterial strains matched with their preferred carbohydrate substrates.

Keywords

bifidobacteria; carbohydrate metabolism; metabolic reconstruction; probiotics; prebiotics; synbiotics; MAG

INTRODUCTION

Bifidobacteria are Gram-positive saccharolytic microorganisms that predominantly inhabit the gastrointestinal tracts of various animals[1]. Multiple *Bifidobacterium* species colonize the human gut throughout our lifespan, with many factors shaping this process, most notably the carbohydrate content of the diet[2,3]. For example, breastfeeding plays a pivotal role in fostering the dominance of specific *Bifidobacterium* species within the neonatal gut microbiota[4] due to their evolutionary adaptation to metabolize milk glycans, such as human milk oligosaccharides (HMOs)[5–9]. Dietary changes throughout early life, namely the cessation of exclusive breastfeeding and the introduction of solid complementary foods, drive a gradual succession of bifidobacterial taxa from those tuned for HMO utilization to those more adapted to foraging dietary glycans (oligo- and polysaccharides) of plant origin[10–12].

Geographic/cultural differences in dietary practices profoundly influence the prevalence and abundance of specific *Bifidobacterium* taxa in gut microbial communities. For example, *Bifidobacterium longum* subsp. *infantis* (*Bl. infantis*), a specialized HMO utilizer[13], constitutes up to 90% of the microbial composition in the microbiota of healthy breastfed infants from non-Westernized populations[14–16]. The neonatal microbiota of individuals from Westernized populations often lack *Bl. infantis* and instead are enriched in less proficient HMO utilizers such as *Bifidobacterium longum* subsp. *longum* (*Bl. longum*), *Bifidobacterium breve*, and *Bifidobacterium pseudocatenulatum*; this appears to be mainly due to a preference for formula use versus breastfeeding and other cultural differences[4,14,17,18].

Despite this variability in taxonomic composition, the overall predominance of *Bifidobacterium* in gut communities is widely associated with multiple health benefits, especially in infancy. For example, the end products of bifidobacterial carbohydrate metabolism, lactate and acetate, can inhibit pathogen colonization[19–21] and serve as a basis for cross-feeding with various community members[22,23]. Additionally, multiple *Bifidobacterium* species produce aromatic lactic acids and other metabolites that modulate the immune system[24–27]. Given these beneficial properties, some bifidobacterial strains are commonly administered as probiotics[28–32]. Complementary strategies, such as prebiotic administration, are employed to facilitate the engraftment of probiotic strains[32–34].

Methods for the rational selection of specific probiotic strains and prebiotic glycans for different populations need to be developed further. For example, the engraftment of a commonly

used probiotic *Bl. infantis* strain may not be sufficiently stable in Bangladeshi children who do not receive enough breast milk and consume supplementary diets with high plant glycan content[32]. Instead, *Bl. infantis* strains isolated from members of the target population may be better adapted to the local diet due to distinctive genomic features related to the metabolism of glycans commonly consumed in that population[32]. Thus, *a priori* knowledge of the metabolic potential to utilize carbohydrates, especially in *Bifidobacterium* strains from understudied populations, may be instrumental in the rational development of locally adapted pro- and prebiotics.

Comparative genomics and related computational approaches have been widely employed to predict the carbohydrate utilization capabilities that underpin the metabolic preferences of bifidobacterial strains. These approaches include analyses of the distribution of Carbohydrate Active Enzymes (CAZymes)[11,35–39], genotype-to-phenotype matching[40–42], and genome-scale metabolic models[43–45]. However, insufficiently detailed expert curation of gene annotations and carbohydrate utilization pathways is a common shortcoming of many previous studies. For example, glycan transporters and transcription factors are frequently overlooked in metabolic reconstructions. Moreover, most analyses have used genomic datasets that did not fully capture the diversity within the *Bifidobacterium* genus, especially at the strain level.

In this work, we reconstructed 66 metabolic pathways defining carbohydrate utilization capabilities (phenotypes) by analyzing the representation of a manually curated set of 565 metabolic functional roles (catabolic enzymes, CAZymes, transporters, and transcriptional regulators) in 263 non-redundant reference genomes of human-associated *Bifidobacterium* strains, including 19 novel isolates from Bangladeshi children. Forty predicted glycan utilization phenotypes were tested and validated *in vitro* using a panel of 16 strains. The obtained genomic compendium was leveraged to develop and apply a machine learning model, which enhanced the ability to predict the distribution of carbohydrate utilization pathways in additional 2710 non-redundant genomes of *Bifidobacterium* isolates or metagenome-assembled genomes (MAGs). Our results uncovered remarkable variability of predicted glycan utilization capabilities at the species and strain levels and revealed the presence of a novel phenotypically distinct subspecies-level clade within the *Bifidobacterium longum* species. Furthermore, we identified several Bangladeshi *Bifidobacterium* strains harboring unique gene clusters enabling the catabolism of plant hemicelluloses and HMO mixtures, suggesting an evolutionary adaptation to maintenance

in the intestinal ecosystems of children who were well into the weaning process. The knowledge database obtained in our analyses enhances our understanding of carbohydrate metabolism and its diversification in bifidobacteria and provides a foundation for rational development of pre-, pro- and synbiotics.

RESULTS

In silico reconstruction of carbohydrate metabolism in reference *Bifidobacterium* strains

Detailed metabolic reconstruction and prediction of carbohydrate utilization phenotypes — Our reconstruction covered a set of 263 non-redundant *Bifidobacterium* genomes of cultured isolates, including 19 novel strains from Bangladeshi infants (**Supplementary Table 1**). We first performed a pangenome analysis to identify 487 core genes shared by all genomes and constructed a maximum-likelihood phylogenetic tree based on the alignment of concatenated sequences of these genes to verify the taxonomic assignments of each strain (**Supplementary Fig. 1**). This phylogenomic analysis, combined with pairwise comparisons of average nucleotide identities (ANI) of the genomes, allowed us to refine taxonomic assignments for 24 strains and delineate subspecies-level clades within the *Bifidobacterium longum* and *Bifidobacterium catenulatum* species (**Extended Data Fig. 2**; **Supplementary Note 1**). Overall, the genomes in the reference set were both taxonomically (19 species and subspecies) and geographically (20 countries) diverse (**Fig. 1a**).

We leveraged a subsystem-based comparative genomics approach[46,47] to reconstruct carbohydrate utilization pathways and predict associated metabolic phenotypes (**Fig. 1b**). To do so, we first performed extensive mining of available literature and databases to identify 425 groups of orthologous carbohydrate utilization genes whose encoded protein products were either experimentally characterized in bifidobacteria (200 metabolic functional roles) or were homologous to proteins whose function had been established in other microbial taxa. The analysis of genomic context combined with *in silico* reconstruction of transcriptional regulons (see *Methods* and **Supplementary Notes 2-3**) allowed us to tentatively functionally annotate 140 additional groups of orthologous genes representing novel metabolic roles involved in glycan metabolism. As a result, we created a curated list of 565 roles that included 226 components of glycan-specific transporters, 188 catabolic CAZymes, 67 downstream catabolic enzymes, and 84 transcription factors (**Supplementary Table 4**). Selected functional roles were used to

reconstruct 21 pathways involved in the utilization of monosaccharides and their derivatives (sugar alcohols and acids), 39 pathways for the utilization of di- and oligosaccharides, and 10 pathways for polysaccharide degradation (**Supplementary Table 3**). The reconstructed pathways captured 81.8% of all catabolic CAZymes [glycoside hydrolases (GHs), carbohydrate esterases (CEs), and polysaccharide lyases (PLs)], which genes were identified in 263 reference genomes (**Extended Data Fig. 1b**).

We designed specific rules associating the gene patterns of each carbohydrate utilization pathway with corresponding metabolic phenotypes (*Methods;* **Extended Data Fig. 1a**; **Supplementary Table 5**). These "phenotype rules" described sets of genes whose presence or absence in a genome distinguished metabolic pathway variants. Pathway variants were translated into binary phenotypes corresponding to utilizers of a specific glycan ("1") and non-utilizers ("0") (*Methods*; **Supplementary Table 5**). The results were combined into a Binary Phenotype Matrix (BPM) that captured a representation of all predicted glycan utilization capabilities across the 263 reference *Bifidobacterium* strains (**Supplementary Table 6**). To initially assess the robustness of this approach, we compared the predicted carbohydrate utilization phenotypes with previously published *in vitro* growth data for six *B. breve* and 19 *B. longum* reference strains grown on 17 and 23 carbon sources[40,48], respectively (**Supplementary Table 12a,b**). The overall prediction accuracy was 95% (F-score 0.95; **Supplementary Table 12c**).

Species-level variability of predicted carbohydrate utilization phenotypes — The obtained BPM was used to analyze the distribution of predicted glycan utilization capabilities across the *Bifidobacterium* genus. Hierarchical clustering of the BPM highlighted the variation in conservation of carbohydrate utilization pathways (**Fig. 2**). Glucose (Glc), galactose (Gal), lactose (Lac), galactoligosaccharide (GOS), and fructose (Fru) utilization pathways were present in >99% of the analyzed reference strains, thus forming the core sugar catabolic capabilities of human-colonizing bifidobacteria. The predicted ability to utilize sucrose (Scr), maltose (Mal), malto- and isomaltooligosaccharides (MOS and IMO), melibiose (Mel), raffinose-family oligosaccharides (RFO), and short-chain fructooligosaccharides (scFOS) was conserved in all *Bifidobacterium* species except *Bifidobacterium bifidum* and *Bifidobacterium gallicum*. Other pathways were more sporadically distributed and associated with distinct glycan preferences of different *Bifidobacterium* species (**Fig. 2**). For example, *B. bifidum, B. breve,* and *Bl. infantis* strains had multiple pathways involved in the utilization of host-derived glycans, such as HMOs

and their constituents, whereas *Bifidobacterium adolescentis*, *Bifidobacterium dentium*, *Bl. longum*, and *B. pseudocatenulatum* had many pathways driving the catabolism of plant oligo- and polysaccharides.

In addition to species-level differences, we observed a high degree of strain-level variability in predicted carbohydrate utilization capabilities. Namely, 55 out of 66 binary phenotypes varied between strains within at least one *Bifidobacterium* species (**Supplementary Table 6**). The genomic differences driving this variability ranged from the presence of individual genes encoding extracellular GHs to genomic loci containing up to 31 genes encoding entire metabolic pathways (**Figs. 4a, 5a**; **Extended Data Figs. 4b,d, 5a,d,g**; **Supplementary Table 7**). This strain-level heterogeneity suggested that a general carbohydrate utilization strategy characteristic for a particular *Bifidobacterium* species can often be tuned at the strain level. For example, *Bl. infantis* Bg064.S07_13.C6 harbored gene clusters encoding xylooligosaccharide (XOS) and long-chain fructooligosaccharide (lcFOS) utilization pathways, making this strain more adapted to metabolize dietary glycans of plant origin compared to most other *Bl. infantis* strains (**Fig. 2**; **Supplementary Table 7**).

The hierarchical clustering by predicted phenotypes did not completely mirror the order of branches on the phylogenetic tree based on core genes (**Fig. 2**; **Supplementary Fig. 1**), indicating an incomplete correspondence between the set of shared phenotypes and phylogenetic relatedness between genomes within the *Bifidobacterium* genus. For instance, strains belonging to the species within the *B. adolescentis* phylogenetic group (*B. adolescentis*, *B. pseudocatenulatum*, *B. catenulatum*, and *B. dentium*)[1] clustered together based on predicted carbohydrate utilization capabilities. In contrast, predicted phenotypic profiles of strains belonging to *Bl. infantis* and *Bl. longum*, two phylogenetically related subspecies within the *B. longum* species, were strikingly different. The representation of glycan utilization pathways in *Bl. infantis*, in fact, more resembled *B. breve* — a more distantly related species also inhabiting the neonatal human gut. Given the importance of *B. longum* in neonatal life, we performed a more focused analysis of pathway distribution within this heterogeneous species.

Subspecies-level diversity of carbohydrate metabolism within the B. longum species — The *B. longum* species includes multiple subspecies-level clades delineated based on specific genetic and phenotypic properties[37,49–53]. Our phylogenomic and ANI-based analyses clustered *B. longum* genomes into three such clades corresponding to *Bl. infantis*, *Bl. longum*,

Bifidobacterium longum subsp. *suis* (*Bl. suis*), and a distinct clade termed hereafter *Bl. nov*. (**Extended Data Fig. 2a**; **Supplementary Fig. 1**; **Supplementary Note 1**). Compared to other *B. longum* subspecies, this distinct clade had fewer predicted host glycan utilization pathways, suggesting a lineage-specific loss of genes and gene clusters (**Fig. 2**; **Supplementary Table 7**; **Supplementary Note 4**). For example, all *Bl. nov* strains lacked the lacto-*N*-biose/galacto-*N*-biose (LNB/GNB) utilization pathway[54,55] conserved in all other *B. longum* strains. Conversely, only *Bl. nov*. strains encoded orthologs of extracellular α-amylase/pullulanase (GH13_14_32) that releases MOS from soluble starch (ST)[56] as well as a pathway driving the utilization of difructose dianhydrides (DFA)[57]. These observations suggest that, unlike other *B. longum* subspecies, *Bl. nov* has a decreased capability to metabolize host-derived glycans but can potentially metabolize ST and related polysaccharides. These differences in carbohydrate utilization capabilities, coupled with the ANI analysis, provide a foundation for the future tentative separation of *Bl. nov.* as a novel subspecies within *B. longum*.

Analysis of the representation of carbohydrate utilization pathways revealed stark differences between *Bl. infantis* and *Bl. longum* (**Fig. 2**). The primary genomic features differentiating *Bl. infantis* strains were gene clusters H1[13] and FL1/2[8] encoding transporters and GHs contributing to the utilization of lacto-*N*-neotetraose (LNnT), 2'/3-fucosyllactoses (FL), lactodifucotetraose (LDFT), lacto-*N*-fucopentaose I (LNFP I), 3'/6'-sialyllactoses (SL), and long-chain HMO species (lcHMOs) (**Fig. 5a**; **Supplementary Table 7**). Certain *Bl. infantis* strains also possessed putative glucuronate (GlcA), inositol (Ino), and gluconate (Gco) utilization pathways not represented in other *B. longum* subspecies (**Extended Data Fig. 5d-h**; **Supplementary Note 3**). The predicted HMO utilization abilities of most *Bl. longum* strains were driven by an extracellular GH136-family lacto-*N*-biosidase that cleaves lacto-*N*-tetraose (LNT) into LNB/Lac and partially degrades LNFP I[58]. Four *Bl. longum* strains, however, harbored a gene cluster driving intracellular FL, LDFT, and LNFP I utilization[59,60] (**Supplementary Table 7**). On the other hand, *Bl. longum* strains had multiple extracellular GHs conferring the ability to (partially) degrade type I and II arabinogalactan (AGI/II)[61,62], arabinan (AR)[63], arabinoxylan (AX)[64,65], and gum arabic (GA)[66,67]. These observations emphasize the general adaptation of *Bl. infantis* to the foraging of milk glycans, whereas *Bl. longum* is adapted to metabolize plant mono-, oligo-, and polysaccharides, with a minority of strains being able to internalize fucosylated HMOs.

We observed substantial heterogeneity of predicted carbohydrate utilization phenotypes within *Bl. suis*. Five *Bl. suis*-like genomes clustered together with *Bl. longum* based on the distribution of carbohydrate utilization pathways (**Fig. 2**). Some subtle features differentiating these *Bl. suis* isolates from the majority of *Bl. longum* strains included the more common occurrence of L-fucose (Fuc)[68] and FL/LDFT utilization pathways, as well as the absence of specific extracellular α-L-arabinofuranosidases involved in the degradation of AX (**Supplementary Table 7**). In contrast, the outlier Bangladeshi *Bl. suis* strain Bg131.S11_17.F6 exhibited a *Bl. infantis*-type pattern of carbohydrate utilization phenotypes and CAZymes (**Fig. 2**; **Extended Data Fig. 1c**). Specifically, this strain harbored the H1 gene cluster involved in the utilization of multiple HMOs but lacked plant glycan utilization pathways (e.g., Ara, AGII) conserved in *Bl. longum* and other *Bl. suis* strains (**Fig. 5a**; **Supplementary Table 7**). These results suggest that Bg131.S11_17.F6 can utilize more HMO species (e.g., LNnT, SL, lcHMOs) than other closely related *Bl. suis* strains but has a decreased ability to metabolize plant glycans. Analysis of metabolic pathway representation in 11 additional *Bl. suis*-like genomes confirmed the uniqueness of the Bg131.S11_17.F6 strain and further highlighted the phenotypic heterogeneity of this group (**Extended Data Fig. 3**; **Supplementary Note 4**).

Together, these results demonstrate the high variability in carbohydrate utilization capabilities within the *B. longum* species, both at subspecies and strain levels. They also point to the unique genomic properties of the Bangladeshi isolates, which we sought to explore further (as described below).

Distinctive carbohydrate utilization features of Bangladeshi strains — Our previous study identified distinctive genomic features of particular Bangladeshi *Bl. infantis* strains pertaining to *N*-glycan and beta-glucoside catabolism[32] and hypothesized that such strains might be better suited to deploy as probiotics in Bangladeshi children consuming diets containing these glycans. Here, we uncovered additional genomic features (e.g., gene clusters) of Bangladeshi isolates that differentiated them from other reference strains of the same species.

One such cluster was identified in *Bifidobacterium catenulatum* subsp. *kashiwanohense* (*Bc. kashiwanohense*) Bg42221_1E1 and *Bifidobacterium catenulatum* subsp. *kashiwanohense_A* Bg42221_1D3, but not in the non-Bangladeshi reference strains belonging to the *B. catenulatum* species (**Supplementary Table 7**). We designated this cluster "*xgl*", as it encoded multiple putative GHs (families 3, 5_4, 29, 31, 42, 43_12), an unclassified CE, β-

glucoside kinase, and an ABC transport system that together we hypothesized catabolize specific plant hemicelluloses, xyloglucans (XGL) (**Figure 4a,b**). Among other species, a similar gene cluster was also identified in all reference *B. dentium* strains. The first step of the pathway reconstructed from the *xgl* cluster was the hydrolysis of the XGL backbone to oligosaccharides by extracellular endo-β-1,4-glucanase Xgl5A (GH5_4). The function of this enzyme was proposed based on its distant homology with xyloglucanase PpXG5 from *Paenibacillus pabuli* XG5[69]. Specifically, catalytic and glycan-binding residues of PpXG5 were conserved in Xgl5A (**Extended Data Fig. 4a**). Released oligosaccharides would be transported into the cell by the associated ABC transport system (XglABC) and degraded sequentially to individual monosaccharides and cellobiose by a coordinated action of GHs and CEs via a mechanism similar to one described in Gram-positive *Ruminiclostridium cellulolyticum*[70]. Transcriptional regulon reconstruction of XglR, a putative transcription factor from the TetR family, suggested that *xgl* cluster genes were potentially co-regulated with the *cbpA* gene encoding cellobiose phosphorylase (GH94), which would act on cellobiose released during XGL degradation (**Fig. 4a-c**; **Supplementary Table 16**).

Another gene cluster unique to *Bc. kashiwanohense* Bg42221_1E1 contained orthologs of genes constituting the H1 cluster in all *Bl. infantis* strains and the "outlier" Bangladeshi *Bl. suis* Bg131.S11_17.F6 strain (**Fig. 5a**). The H1 cluster canonically encodes elements of HMO utilization pathways and is believed to confer the ability to metabolize a multitude of HMO species[5,13,71]. The H1 variant identified in Bg42221_1E1 encoded two ABC transporters, one of which (HmoABC) is likely involved in the uptake of LNnT and other type II HMO structures based on biochemical data[72]. Additionally, this cluster encoded β-*N*-acetylglucosaminidase (GH20), two α-fucosidases (GH29 and GH95), and α-sialidase (GH33) involved in intracellular HMO hydrolysis[73–75], as well as downstream catabolic pathways for *N*-acetylglucosamine (GlcNAc) and fucose. The H1 cluster genes in *Bc. kashiwanohense* Bg42221_1E1 were predicted to be regulated by NagR (**Fig. 5a**; **Supplementary Table 16**), a ROK-family transcriptional repressor that controls HMO utilization in various bifidobacteria and senses GlcNAc and its phosphorylated derivatives[76].

Among other genomic features of Bangladeshi *Bifidobacterium* genomes not represented in other reference genomes was a putative D-galactonate utilization pathway exclusively present in *B. breve* Bg41721_1C11 (**Extended Data Fig. 5g,h**; **Supplementary Note 3**). Additionally,

this strain lacked the *nan* gene cluster that encodes a well-characterized catabolic pathway for sialic acid (Neu5Ac)[77] (**Extended Data Fig. 5a,b**). This absence of *nan* genes in Bg41721_1C11 was remarkable, given their presence in all other reference *B. breve* genomes and their importance for the foraging of Neu5Ac released from sialylated HMOs and mucin *O*-glycans by, for example, *B. bifidum*[77,78]. Overall, these results demonstrate that specific Bangladeshi *Bifidobacterium* strains have a distinct enhanced ability to catabolize dietary plant polysaccharides and HMOs, but also may lack pathways typically conserved in reference strains. These traits may reflect a strain-level adaptation to the lifestyle and diet of Bangladeshi children. We then sought to experimentally assess the key metabolic predictions for a set of *Bifidobacterium* isolates.

Experimental assessment of predicted carbohydrate utilization capabilities in bifidobacterial isolates

Growth-based assays — To experimentally test some of our *in silico* carbohydrate utilization predictions, we conducted a series of *in vitro* growth experiments for 16 *Bifidobacterium* strains, including 10 Bangladeshi, five US, and one Malawian isolate. We used a panel of carbon sources corresponding to 40 predicted glycan utilization phenotypes: 17 monosaccharides and their derivatives, 18 di- and oligosaccharides, and 5 polysaccharides (**Supplementary Table 12d**). These strains were cultured in the MRS-AC medium containing a test carbon source (5-10 mg/mL), and the growth was defined based on strain-specific OD_{600} thresholds (*Methods*; **Supplementary Fig. 2**).

Consistent with phenotypic predictions, the ability of bifidobacteria to grow on fructooligosaccharides of different chain lengths (phenotypes scFOS and lcFOS) was species and strain-dependent (**Fig 3a**). Among five tested *Bl. infantis* strains, one did not grow on any FOS, two grew only on scFOS, and two grew both on scFOS and lcFOS (**Supplementary Table 12d**). The strain-dependent ability of *Bl. infantis* to grow on GlcA and XOS was in perfect agreement with our predictions (**Extended Data Fig. 5d-f**; **Supplementary Table 12d**). Similarly, certain *B. breve* strains grew on mannotriose (phenotype bMnOS); however, none of them grew in MRS-AC supplemented with konjac glucomannan (phenotype bMAN) unlike *B. dentium* STL_TW14.1_LFYP24, a strain harboring a GH26-family extracellular endo-β-1,4-mannanase (**Extended Data Fig. 4f-e**; **Supplementary Note 2**). Among the 16 tested strains, only *Bc.*

kashiwanohense Bg42221_1E1 and *B. dentium* STL_TW14.1_LFYP24 harboring the *xgl* cluster and, thus, predicted to degrade XGL, could grow in the MRS-AC medium supplemented with tamarind XGL (**Fig. 4d**). We also confirmed that, unlike other *B. breve* strains, the Bangladeshi strain Bg41721_1C11 lacking the Neu5Ac utilization pathway, did not grow in MRS-AC-Neu5Ac (**Extended Data Fig. 5c**).

The overall prediction accuracy was 94% (F-score 0.93; **Fig. 3b**; **Supplementary Table 12e**). The prediction accuracy was lower for glycans in the monosaccharides and derivatives group (90%). Approximately 2/3 of incorrect predictions in that group were false negative, i.e., when a strain had been predicted not to utilize a particular monosaccharide but grew in the MRS-AC medium supplemented with it. It is important to note that strains with false negative predictions did harbor downstream catabolic pathways for the respective monosaccharides but were classified as predicted non-utilizers due to the absence of genes encoding known transporter(s). This observation points to the existence of some yet uncharacterized monosaccharide uptake mechanisms in bifidobacteria. Similar false negative predictions were noted for several HMO utilization phenotypes, where we, for example, underpredicted the ability of *Bl. suis* Bg131.S11_17.F6 to metabolize FL because its genome lacked the genes encoding characterized FL transporters[8,9] (**Supplementary Table 7**). Some false positive predictions (i.e., when a strain had been predicted to utilize a particular glycan but did not grow on it) could be due to the mutations in genes encoding carbohydrate utilization pathways. For example, the *gltA* gene encoding the substrate-specific component of the LNT transporter in *B. pseudocatenulatum* STL_TW14.1_LFYP29 was disrupted by a premature stop codon, which may explain the absence of growth on LNT.

Glycoprofiling of HMO utilization — As noted above, the H1 cluster in *Bl. infantis* is believed to confer the ability to metabolize multiple HMOs[13,79]; however, the exact list of utilized HMO species cannot be confidently predicted due to insufficient knowledge of individual gene (transporter) functions. Given the presence of homologous H1 clusters in *Bl. suis* and *Bc. kashiwanohense* isolates of Bangladeshi origin (**Fig. 5a**), we sought to compare their HMO utilization capabilities with *Bl. infantis* (predicted robust utilizer), *Bl. longum* (predicted weak utilizer), and *B. pseudocatenulatum* (predicted non-utilizer) strains. We first observed that only the strains harboring the H1 cluster could efficiently grow in MRS-AC supplemented with an HMO mixture isolated from pooled human milk (**Fig. 5b**). HPLC-based glycoprofiling of

culture supernatants at two time points (8 and 24 h) revealed that H1 cluster-positive *Bifidobacterium* strains indeed consumed multiple fucosylated and sialylated HMO species compared to H1-negative strains (**Fig. 5d**; **Supplementary Table 13**). The supernatants of H1 cluster-negative *Bl. longum* Bg115.S08_3A11 had reduced quantities of LNT, LNnT, LNFP I, LNFP III, LNH, and FLNH. Since this strain lacked transporters for these HMOs, their depletion was likely not due to intracellular utilization and rather to (partial) extracellular degradation via β-galactosidase and lacto-*N*-biosidase[58] activities, with only LNT degradation products, LNB and Lac, being completely utilized. Supporting this notion, Bg115.S08_3A11 grew in MRS-AC supplemented with LNT but not LNnT (**Supplementary Fig. 2**). The *B. pseudocatenulatum* strain did not utilize tested HMOs, including LNT, consistent with the disruption of the *gltA* gene. Overall, these data provide the first experimental evidence that particular *Bc. kashiwanohense* and *Bl. suis* strains have HMO utilization capability comparable to *Bl. infantis* due to the presence of H1 cluster genes.

RNA-seq analysis of carbohydrate utilization genes in Bc. kashiwanohense — We used RNA-seq to test regulon predictions and functional gene/pathway assignments in the *Bc. kashiwanohense* Bg42221_1E1 strain harboring unique *xgl* and H1 gene clusters. We first compared the transcriptomes of this strain grown in MRS-AC-XGL vs. MRS-AC-Lac. The *xgl* cluster genes (except *xglK*) were the most significantly upregulated (100-550 fold) together with the *cbpA* and *xglE2a* genes (**Fig. 4e**; **Supplementary Table 15a**). This coordinated induction of predicted XGL catabolism genes was consistent with the structure of the reconstructed XglR regulon (**Fig. 4a**). We also observed the moderate induction (4-30 fold) of genes involved in xylose, XOS/AXOS, and xylan/arabinoxylan metabolism (**Fig. 4e**; **Supplementary Table 15a**)[12,80,81]. The upregulation of genes not directly involved in upstream XGL pathway was likely due to the intracellular release of xylose residues acting as a potential transcriptional effector of XylR (ROK family) and XosR (LacI family), transcription factors predicted to repress these genes (**Extended Data Fig. 4b,c**; **Supplementary Table 16; Supplementary Note 2**)[82]. Overall, the transcriptomic profile of *Bc. kashiwanohense* Bg42221_1E1 during growth in MRS-AC-XGL provided additional evidence that the *xgl* cluster confers the ability to catabolize XGL. It also suggests that transcriptional regulatory mechanisms of this strain are adapted to foraging mixtures of various plant oligo- and polysaccharides.

We next compared the transcriptomes of *Bc. kashiwanohense* Bg42221_1E1 grown in MRS-AC-LNT and MRS-AC-Lac. Most H1 cluster genes were significantly upregulated in the presence of LNT (**Fig. 5c; Supplementary Table 15b**). The upregulation of these genes was consistent with the proposed regulation by NagR (**Fig. 5a**), a transcriptional repressor that controls HMO utilization in various bifidobacteria and senses GlcNAc (a structural component of LNT) and its phosphorylated derivatives[76]. These data also suggested that HmoABC or HmoA2B2C2 might be responsible for LNT uptake, given that no other transporter genes were significantly induced.

Overall, the experimental data provided confidence in predicted carbohydrate utilization phenotypes, a justification for their projection over a multitude of newly sequenced *Bifidobacterium* isolates and metagenome-assembled genomes (MAGs). We reasoned that such a projection would enable a comparative analysis of bifidobacterial strains from different geographic areas reflecting different lifestyles and local dietary conditions.

Automated inference of carbohydrate utilization pathways across large collections of genomes and MAGs

We utilized curated metabolic reconstructions obtained for 263 reference *Bifidobacterium* genomes to analyze the representation of glycan utilization pathways in a set of 2710 non-redundant *Bifidobacterium* genomes. This set included 306 genomes of isolates and 2,404 MAGs of high quality (completeness ≥ 97 % and contamination ≤ 3%). Ten of these genomes corresponded to taxa not represented in the reference collection, namely *Bifidobacterium animalis* subsp. *animalis*, *Bifidobacterium pullorum*, *Bifidobacterium tsurumiense*, and *Bifidobacterium ruminantium*. We used the annotated protein sequences of reference strains to assign putative functions to a subset of 400,267 protein sequences in 2710 genomes (**Fig. 1b**). These sequences formed the basis for predicting which of 66 carbohydrate utilization pathways were encoded in each genome via a random forest model (*Methods*; **Supplementary Tables 8**,**9**). To assess the metabolic pathway variability beyond carbohydrate utilization, we analyzed the representation of 29 pathways encompassing B vitamin/amino acid biosynthesis[47] and urea utilization (an important species-specific phenotype) using a previously described pipeline[83] (**Supplementary Table 17**).

Non-metric multidimensional scaling (NMDS) of the Hamming distance matrix calculated from the merged BPM for 2973 genomes, revealed that most genomes of the same species/subspecies formed distinct clusters (**Fig. 6a**). Taxonomy accounted for 90% of the variation in the data (PERMANOVA R^2=0.9, p=0.001), indicating significant differences between the species/subspecies. Furthermore, the test for homogeneity of multivariate dispersions (F=94.42, p=0.001) highlighted the influence of dispersion effects on the observed clustering patterns, indicating significant differences in the strain-level variability of carbohydrate utilization capabilities between taxa. To explore these concepts further, we calculated predicted phenotypic richness (total number of predicted carbohydrate utilization pathways) for each genome (**Fig. 6b**) and the percent of predicted utilizers/degraders within each taxon (**Fig. 6c**). We found significant differences in predicted richness even between phylogenetically close subspecies-level clades. For example, *Bl. nov.* exhibited significantly lower phenotypic richness compared to all other *B. longum* subspecies (generalized linear model, p<0.001, Bonferroni-corrected post-hoc test, **Fig. 6b**), driven by the loss of conserved host glycan utilization pathways (**Fig. 6b**; **Supplementary Note 4**). Similarly, *Bc. catenulatum* had significantly lower phenotypic richness than *Bc. kashiwanohense* due to the absence of fucosylated HMO (FL, LDFT, LNFP I) and plant oligosaccharide (lcFOS, bMnOS, GmOS) utilization pathways in most studied strains (**Extended Data Fig. 6b**; **Supplementary Note 5**).

The species/subspecies-level differences in predicted carbohydrate utilization capacities, in part, reflected distinct ecological strategies with respect to the metabolism of dietary glycans of different origin and structure (**Fig. 6c**). *B. bifidum* showed a notable specialization toward the metabolism of host glycans (mucin *O*-glycans, HMOs, and their constituents), and most strains of this species completely lacked pathways involved in the utilization of plant di-/oligo- and polysaccharides. The predicted glycan utilization capabilities of other *Bifidobacterium* species were more versatile, although the ability to metabolize polysaccharides of plant origin was less prevalent compared to their mono-, di- and oligosaccharide constituents (**Fig. 6c**).

The strain-level variability in the distribution of carbohydrate utilization pathways in 2973 genomes was more pronounced compared to 263 reference genomes, with 64 out of 66 pathways varying between strains in at least one taxonomic group (**Fig. 6c**). Notable examples of strain-level hetergeneity included reference Bangladeshi strains *Bl. suis* Bg131.S11_17.F6 and *Bc. kashiwanohense* Bg42221_1E1; the analysis of additional 2710 genomes did not reveal any

other *Bl. suis* and *Bc. kashiwanohense* genomes harboring similar unique genomic features (e.g., the presence of H1 and *xgl* cluster genes; **Fig. 6a**; **Extended Data Fig. 6b**). We also identified that three *B. adolescentis* genomes (AF96-10M2bTA, B37.CMAG_2, UN03-88) encoded LNB/GNB and *N*-glycan utilization pathways that previously were not reported in this species (**Supplementary Tables 8,9**). The substantial variability of predicted carbohydrate utilization capabilities at the strain level contrasted with the conservation of most pathways responsible for the biosynthesis of essential metabolites such as B vitamins and amino acids (**Extended Data Fig. 6a**).

We next investigated whether specific carbohydrate utilization pathways are enriched in bifidobacterial genomes isolated from different human populations using two categories: (i) host age associated with microbiota maturation (age < 3 years representing infant or transitional microbiota and age ≥ 3 years representing adult-like microbiota), and (ii) host lifestyle as defined in Pasolli et al.[84] ("Westernized" vs. "non-Westernized"). We observed the enrichment of host glycan utilization pathways in genomes recovered from the "age < 3" group samples and the enrichment of plant glycan utilization pathways in genomes recovered from "age ≥ 3" samples in both Westernized and non-Westernized cohorts (Fisher's exact test, $P_{adj} \leq 0.01$; **Extended Data Fig. 7a,b**). When comparing genomes recovered from the same age group but different lifestyles, 11 pathways were enriched in genomes from Westernized child samples, and 16 pathways were enriched in genomes from non-Westernized child samples; no pathways were significantly enriched across genomes recovered from adult samples (Fisher's exact test, $P_{adj} \leq 0.01$; **Extended Data Fig. 7c**). These results are consistent with the uneven taxonomic distribution of *Bifidobacterium* genomes recovered from Westernized and non-Westernized microbiota. For example, pathways enriched in the non-Westernized children group were associated with *Bl. infantis*, a subspecies that is more prevalent in the non-Westernized group (Odds ratio = 5.35, Fisher's exact test, $P_{adj} \leq 0.01$).

Among taxon-specific enrichments, pathways for sorbitol, mannitol (Stl and Mtl; **Supplementary Note 3**), and lcFOS utilization were significantly enriched in *Bl. infantis* genomes of Western origin (Fisher's exact test, $P_{adj} \leq 0.01$; **Extended Data Fig. 7d**). The Stl utilization pathway was enriched in Westernized *B. adolescentis* genomes, as was the Neu5Ac and *N*-acetylglucosamine 6-sulfate utilization pathways in Westernized *B. breve* genomes. Our large-scale analysis pinpoints ecological differences between glycan foraging strategies within

Bifidobacterium that reflect species/subspecies-level evolutionary adaption to different habitats and diets (e.g., infant gut vs. adult gut). The results also emphasize a substantial strain-level variability that may have arisen due to adaptation to lifestyle and local diets.

DISCUSSION

The ability to metabolize dietary carbohydrates is one of the hallmarks of bifidobacterial fitness in the competitive environment of the human gut[85]. The saccharolytic nature of bifidobacteria is reflected in their genomic content, with 8-12 % of all genes dedicated to carbohydrate metabolism[86]. Therefore, detailed knowledge of this complex metabolic process has not only fundamental importance but also practical implications for the rational selection and development of pro- and prebiotics. While previous studies have uncovered many species-level differences of carbohydrate utilization capabilities within the *Bifidobacterium* genus[35,85,87,88], our knowledge of the strain-level variability remained limited, especially when taking into account samples collected in low- and middle-income countries[89] A recent expansion in the number of publicly available *Bifidobacterium* genomes, driven by advances in culturomics[90–93] and metagenomics[94–98], provides an opportunity to address this problem through genomics-based metabolic reconstruction. That said, the accurate prediction of carbohydrate utilization capabilities from genomic data remains a substantial bioinformatic challenge given incomplete and imprecise functional annotations of genes (especially glycan transporters) and metabolic pathways in public databases, as well as generally insufficient incorporation of published experimental data into reconstructions[99].

We have used an *in silico* comparative-genomic approach based on rigorous manual curation of gene annotations and incorporation of literature data (144 publications) to reconstruct 66 carbohydrate utilization pathways encoded in 2973 *Bifidobacterium* genomes and MAGs. The analysis of reconstructed pathway representation revealed major differences in glycan foraging strategies among *Bifidobacterium* species, as well as considerable strain-level variability. The major differences are likely associated with the evolutionary adaptation of various taxa to specific ecological niches, such as the infant (breastfed or weaning) and adult gut. While similar generalizations have been reported previously[35,85,87], the curated metabolic reconstruction allowed us to pinpoint the exact glycan utilization pathways differentiating *Bifidobacterium*

species and subspecies. This analysis also indicated the presence of a previously undescribed subspecies-level clade within the *B. longum* species, termed *Bl. nov*. Despite being closely related to *Bl. longum* phylogenetically, *Bl. nov* strains had several conserved genomic features distinct from those of other *B. longum* subspecies, namely the lack of the lacto-*N*-biose/galacto-*N*-biose pathway[55] and the presence of soluble starch[56] and difructose dianhydride utilization pathways[57]. Given that most *Bl. nov* genomes were recovered from infants from Westernized populations, the unique predicted carbohydrate capabilities of this putative subspecies might indicate an adaptation to Westernized diets, characterized by decreased breast milk and increased formula intake.

The high degree of strain-level variability in the representation of predicted carbohydrate utilization pathways was one of the most important findings. This variability, which stemmed from the differences in the accessory genome ranging from individual genes to gene cassettes encoding complete metabolic pathways, reflects the adaptive versatility of bifidobacteria at the strain level. However, some observed differences between closely related strains were more substantial and comparable to differences observed between species and subspecies. The prime example is the group of *Bl. suis*-like isolates of Bangladeshi origin. Similar *Bl. suis*-like strains isolated from weaning Bangladeshi children (recently proposed to be separated as *Bifidobacterium longum* subsp. *iuvenis*) were predicted to utilize HMOs and plant fibers based on the analysis of the distribution of CAZyme genes[11,53]. Our detailed reconstruction of complete carbohydrate utilization pathways reveals a more nuanced picture of intra-strain variability within this subspecies-level clade. Specifically, we distinguish between two phenotypic subgroups within *Bl. suis*-like strains. The first, represented by Bg131.S11_17.F6, resembles *Bl. infantis* in its ability to forage mixtures of various type I and II HMOs but not polysaccharides of plant origin. The second subgroup is phenotypically comparable to *Bl. longum* strains SC596 or MC10007[59,60], which utilize a more limited set of HMOs (LNT, 2'FL, 3FL, LDFT, and LNFP I) but feature extended plant glycan (e.g., type II arabinogalactan) utilization capabilities. Such considerable strain variability complicates the explicit delineation of certain species/subspecies-level clades within *Bifidobacterium*, which historically heavily relied on specific glycan utilization phenotypes[49,53,88].

The presence of unique gene clusters and metabolic pathways in Bangladeshi strains compared to the reference genomes was notable. One such cluster, *xgl*, was initially found in

certain Bangladeshi strains, such as *Bc. kashiwanohense* Bg42221_1E1. A previous report indicated the presence of a homologous gene cluster in *Bifidobacterium* species isolated from captive common marmosets[100]; however, its function was not tested experimentally. Our growth and transcriptomic data strongly suggest that *xgl* is involved in the degradation of xyloglucan calling for additional studies on the mechanistic details of this process. Another unique gene cluster identified in *Bc. kashiwanohense* Bg42221_1E1 is homologous to the H1 cluster of *Bl. infantis*. This result was unexpected since the H1 cluster was previously thought of as a unique feature of *Bl. infantis*[79], though also more recently identified in certain Bangladeshi *Bl. suis*-like strains such as Bg131.S11_17.F6[11,32]. While some functional roles implemented by H1 remain to be characterized, our growth and HMO consumption profiling data confirm that its presence enables the utilization of multiple HMO species. *Bc. kashiwanohense* strains were previously shown to metabolize fucosylated HMOs (2'FL and LNFP I) and plant hemicelluloses (xylan and arabinoxylan)[12]. Our analysis demonstrates that the Bangladeshi *Bc. kashiwanohense* strain can also degrade xyloglucan and utilize a broader range of HMOs. These results indicate an evolutionary adaptation of Bangladeshi strains to thrive in the microbiota of weaning Bangladeshi children, whose diet, in addition to breast milk, contains many foods rich in plant polysaccharides.

Our reference genomic dataset was heavily skewed towards samples from high-income countries, and non-Westernized strains were mainly represented by isolates from Bangladeshi children. There is a possibility that genomes recovered from other low- and middle-income countries may encode unique metabolic functional roles or novel carbohydrate utilization pathways not described in this work. This points to the need for initiatives that seek to expand collections of *Bifidobacterium* stains and genomes from under-represented populations in an ethical and culturally-sensitive way[89].

Our growth studies showed agreement (accuracy > 90%) between predicted carbohydrate utilization capabilities and observed growth phenotypes *in vitro*. Nevertheless, a small subset of false negative predictions for certain monosaccharides and HMOs point to the existence of yet unknown transport mechanisms not captured in our reconstructions. More work is needed to identify and characterize the structure, substrate specificities, and expression of transporters in different *Bifidobacterium* species and strains. Additionally, approximately 18% of GHs and CEs

were not assigned to any of the 66 carbohydrate utilization pathways, suggesting yet uncharacterized catabolic potential.

Another limitation of our work is that we did not consider the complementary functions pertaining to glycan metabolism provided by the gut microbial community as a whole. The lower prevalence of catabolic pathways for polysaccharides compared to their low-molecular constituents indicates that many *Bifidobacterium* strains suggest syntrophic interactions within the genus and potentially with keystone polysaccharide degraders such as *Bacteroides* and *Prevotella* (*Segatella*)[83,101]. The metabolic reconstruction approach described in this work presents a framework for the detailed expert curation of glycan utilization in other microbial taxa in the future. These observations provide directions for future improvements of the genomic encyclopedia of bifidobacterial carbohydrate glycan pathways reported here, which, in its current form, represents the most detailed characterization of bifidobacterial carbohydrate metabolism to date. It may serve multiple purposes, from more precise and accurate functional annotation of genomic and metagenomic data to facilitating the development of novel pre- and probiotic formulations, including consortia comprised of multiple strains of a given bifidobacteria species that may allow broader use of that species as a probiotic across different human populations.

METHODS

Collection of *Bifidobacterium* genomes

Reference set of isolate genomes for in silico metabolic reconstruction. Bifidobacterium genomes were retrieved from PATRIC[102] and IMG[103] databases as of October 2020. We selected 335 genomes based on the following criteria: (i) human or probiotic product-derived *Bifidobacterium spp.* isolates, (ii) number of contigs ≤ 200, (iii) CheckM[104] completeness ≥ 97 % and contamination ≤ 3%. Additionally, we sequenced 31 genomes of strains isolated from fecal samples obtained from Bangladeshi children. The resulting reference genomic set of 366 isolate genomes was clustered using a 99.95% ANI threshold via dRep (v3.4.2)[105] with the following command "dRep dereplicate -pa 0.9 -sa 0.9995 -nc 0.35 --S_algorithm ANImf" to create a set of 263 non-redundant genomes (**Supplementary Table 1**).

Additional set of genomes for large-scale projection of metabolic phenotypes. Additional human gut-derived *Bifidobacterium* genomes were collected from previous studies and included genomes of isolates from three collections (HBC[90], BIO-ML[92], and CGR[91]) and MAGs from five collections (UHGG[94], KIJ[97], ELGG[95], SPMP[98], and IMGG[96]). We used the same selection criteria for both isolate genomes and MAGs: (i) *Bifidobacterium* spp. based on Genome Taxonomy Database (GTDB) taxonomy[106] except *Bifidobacterium leopoldii* and *Bifidobacterium vaginale*, (ii) number of contigs ≤ 200, (iii) CheckM completeness ≥ 97 % and contamination ≤ 3%. The resulting 4835 genomes were clustered via dRep using the following command "dRep dereplicate -pa 0.9 -sa 0.999 -nc 0.35 --S_algorithm fastANI" to create a set of additional 2710 non-redundant genomes (**Supplementary Table 2**).

Gene prediction and annotation. Protein-coding sequences for all genomes were predicted and annotated with Prokka (v1.14.6)[107] using default settings. The 263 reference genomes were additionally annotated via RAST[108].

Isolation and sequencing of *Bifidobacterium* strains from Bangladeshi donors

The 31 strains were isolated from fecal samples obtained from 6-24-month-old Bangladeshi children from the cohorts described in previous studies[32,109]. The details about genome sequencing and assembly are given in *Supplementary Methods.*

Taxonomy inference

Initial taxonomic assignments of 263 reference *Bifidobacterium* genomes were retrieved from NCBI and PATRIC for publicly available genomes or identified based on 16S rRNA gene sequencing for Bangladeshi isolates. To check and refine the taxonomy, we built a pangenome via Panaroo (v1.3.2)[110]. Concatenated nucleotide sequences of 487 identified core genes were aligned using MAFFT (v7.515)[111]. A maximum-likelihood phylogenetic tree was built in IQ-TREE (v2.2.0.3)[112] and visualized via iTOL[113]. Pairwise ANI values (ANIb) were calculated using pyani (v0.2.12)[114]. The exact commands and parameters used to run the tools are described in **Supplementary Code File 1**. The topology of the constructed tree and ANI matrices were manually inspected to check (and correct if needed) taxonomic assignments of genomes based on their co-clustering with branches corresponding to the type or well-characterized strains of various *Bifidobacterium* species and subspecies. GTDB-based taxonomic annotation of 2710 additional genomes was retrieved from original publications and further refined using a Mash-based[115] comparison with 263 reference genomes. Taxonomies were assigned at the species level if the Mash distance to the closest genome from the reference set was $\leq$ 0.05 and at the subspecies level if the distance was $\leq$ 0.03.

Subsystems-based annotation and *in silico* metabolic reconstruction of glycan utilization pathways

We used a subsystems-based approach implemented in the SEED platform[46,116] for the reconstruction of carbohydrate utilization pathways across the reference set of 263 *Bifidobacterium* strains. RAST-annotated genomes were uploaded to mcSEED (microbial community SEED), a clone of the SEED annotation environment. We created 27 subsystems that captured the representation of genes that implement *functional roles* (glycan transporters, CAZymes, downstream catabolic enzymes, and transcriptional regulators) constituting a specific carbohydrate utilization pathway (**Supplementary Table 4**). Information about functional roles was collected via extensive literature search[117] as well as exporting data from Transporter Classification (TCDB)[118], Carbohydrate Active Enzyme (CAZy)[119], Kyoto Encyclopedia of Genes (KEGG)[120], and RegPrecise[121] databases. Additional functional gene annotation (i.e., prediction of novel functional roles) was based on three genome context techniques: (i) clustering of genes on the chromosome (operons), (ii) co-regulation of genes by a common regulator, and (iii) co-occurrence of genes in a set of related genomes. Novel regulons of

transcription factors potentially involved in regulating carbohydrate metabolism were reconstructed using a previously established approach[76].

The representation of the curated set of functional roles in 263 genomes was analyzed using homology-based methods implemented in mcSEED, and 70 carbohydrate utilization pathways were reconstructed (**Supplementary Tables 3,5**). Many of these pathways contain alternative biochemical modules (routes) implemented by different subsets of catabolic enzymes and diverse glycan transporters. For all identified pathway variants, we established genomic signatures represented by a subset of functional roles that are required for the utilization or degradation of a specific carbohydrate (**Supplementary Table 5**). *Utilization* was defined as a process when a glycan molecule (mono- or oligosaccharide) is transported into the cell and then catabolized using a combination of CAZymes and downstream catabolic enzymes. *Degradation* was defined as a process when a glycan molecule is partially degraded via extracellular CAZymes, and the degradation products, such as mono- and oligosaccharides, are then utilized. Genomic signatures were further translated to a set of rules underlying the assignment of detailed pathway variants: (i) transporter + catabolic pathway ("U"), (ii) catabolic pathway without transporter ("P"), (iii) no catabolic pathway ("-1") (**Supplementary Tables 5,7**). Detailed pathway variants were then converted to simplified predicted carbohydrate utilization phenotypes: "U" to a numeric binary phenotype "1" (utilization or degradation), "P" and "-1" to a numeric binary phenotype "0" (no utilization or degradation). The resulting 70 predicted carbohydrate utilization phenotypes for 263 strains comprised a reference Binary Phenotype Matrix (BPM) (**Supplementary Table 6**). Four out of 70 pathways (GalNAc, ManAc, Man, and GalA), for which all genomes in the reference set were assigned variants "P/-1" and phenotype "0", respectively, were kept in the BPM for comparison with experimental data but were not used in the phenotype projection pipeline.

The representation of CAZymes in 263 genomes was analyzed using dbCAN (v4.0.0)[122]. HMMER, dbCAN-sub, and DIAMOND-based searches against the CAZy database were used for CAZyme identification. Only GHs, CEs, and PLs predicted by ≥2 approaches were retained.

Phenotype projection pipeline

The workflow used to predict the carbohydrate utilization pathways encoded in a set of additional 2710 *Bifidobacterium* genomes is schematically illustrated in **Fig. 1b**. We first

constructed a reference database containing (i) 39618 functionally annotated protein sequences from 27 curated mcSEED subsystems for 263 reference genomes, (ii) additional 52989 protein sequences ("outgroup" not included in these subsystems), clustered at 95% amino acid identity and 95% coverage using MMSeqs2 (v14.7e284)[123]. We then mapped the protein sequences of 2710 genomes to the constructed reference database via DIAMOND (v2.1.4)[124]. The following method was implemented to annotate multidomain proteins. For each query protein sequence, we used the top 50 hits based on the bitscore and clustered the start and end position coordinates of the corresponding alignments using DBSCAN (scikit-learn v1.2.1)[125]. The centers of each cluster were considered as potential domain boundary coordinates, which were used to split query protein sequences into domains (with database hits attributed to each individual domain). Next, for each domain with ≥ 35 amino acids, we used the Gaussian kernel density modeling (*KernelDensity* function from *sklearn.neighbors*) of the sequence identity distribution of each set of hits to that domain. The highest local minimum (*argrelextrema* function from *sklearn.signal*) was employed as a threshold to remove low-confidence hits. Finally, functional annotations were applied from the reference database to each query protein or domain by a majority rule within each set of high-scoring, domain-specific reference hits. High-identity hits to proteins from the outgroup of the reference database were used as criteria to vote "against" applying annotation to each query. This procedure yielded 400,267 annotated protein sequences, representing 563 unique functional roles (**Supplementary Table 9**).

We integrated the results of gene-level functional annotation into *in silico* predictions denoting the presence ("1") or absence ("0") of metabolic pathways corresponding to predicted carbohydrate utilization phenotypes using a semi-automated process based on a machine learning (ML)-approach. We compared >30 ML methods implemented in *Caret* (v6.0.86)[126], using a "leave-one-out" cross-validation approach, in which we removed a single reference genome from the set of 263 genomes, trained ML models on the remaining genomes, then applied the models to the "test" genome to predict phenotypes. This procedure was then repeated for each genome and each carbohydrate utilization pathway. The results of this analysis identified random forest as the best-performing method (i.e., it produced the greatest number of correctly predicted phenotypes in our reference training dataset). We then built random forest models based on the reference dataset for each pathway except GalNAc, ManAc, Man, and GalA, for which all genomes in the reference set were assigned the binary phenotype "0". The list of functional roles

used as predictors for random forest models was manually curated to match previously designed rules (**Supplementary Table 5**). The model parameters were additionally optimized using a grid search. The constructed models were used to predict the presence/absence (1/0) of 66 carbohydrate utilization pathways in 2710 genomes (**Supplementary Table 8**).

The representation of additional metabolic pathways (e.g., the biosynthesis of B vitamins and amino acids) in 263 reference and 2710 additional genomes was inferred via a similar approach described in detail previously[83].

Visualization and statistical analysis

Principal component analysis (PCA) was used for ordination of the table with CAZyme subfamily representation. Non-metric multidimensional scaling (NMDS) was used for the ordination of the Hamming distance matrix calculated from the merged BPM for 2973 genomes. PERMANOVA and a test for homogeneity of multivariate dispersions were performed using the *adonis2* and *betadisper* functions, respectively, in the *vegan* R package[127]. Differences in predicted phenotypic richness were assessed via a generalized linear model (GLM) that assumes Poisson distribution; post-hoc tests with Bonferroni correction were conducted using the *emmeans* R package[128]. Fisher's exact test was used for the enrichment analysis of pathway representation. For each out of 66 carbohydrate utilization pathways, contingency tables were calculated. Rows represented the absence/presence of predicted pathways ("0" or 1"), whereas columns represented a combination of metadata categories: age group (age < 3 years vs. age ≥ 3) and host lifestyle (Westernized vs. non-Westernized)[84]. The p-values from the Fisher's test were corrected for multiple testing using the Benjamini-Hochberg procedure[129], and $P_{adj} \leq 0.01$ was used as a significance cut-off. Odds ratios were calculated using the *fisher.test* function. A set of R packages was used for the data visualization: *ggplot2*[130] for most plots, *ComplexHeatmap*[131] for heatmaps,. Gene clusters were visualized via clinker (v0.0.27)[132]. The code used for all analyses can be found in **Supplementary Code File 1**.

Growth of *Bifidobacterium* strains on selected carbohydrates *in vitro*

Bifidobacterium strains were routinely grown in Lactobacilli de Man, Rogosa & Sharpe broth without dextrose (Alpha Biosciences) supplemented with 0.34% (wt/vol) sodium ascorbate, 0.029% (wt/vol) L-cysteine-HCl monohydrate, and 10 mg/mL lactose (MRS-AC-Lac) in a

chamber maintained with a gas mix of 10% H_2, 10% CO_2, and 80% N_2 (Coy Laboratory Products). The ability of 16 *Bifidobacterium* strains to grow on 42 carbohydrates (**Supplementary Table 10**) was measured using a custom carbohydrate array constructed in flat-bottom half-area 96-well plates (Costar). Wells were filled with 55 µl of a sterilized carbohydrate stock at 2× concentration (10 to 20 mg/mL) and transferred to the chamber 48 h before the experiment. Each substrate was represented in triplicate. Triplicate wells with sterile water and lactose were included as negative and positive controls, respectively. Cultures for assay inoculations were grown overnight from a frozen glycerol stock in MRS-AC-Lac at 37°C. Each culture was then diluted 1:100 into a fresh MRS-AC-Lac medium and grown at 37°C for 8 h, after which the optical density at 600 nm (OD_{600}) was adjusted to 0.6. A 100 µL aliquot from each culture was harvested at 5,000 × g for 5 min and resuspended in 6 ml of 2× MRS-AC without added carbohydrate. Each well in the carbohydrate array was loaded with 55 µL of the inoculated 2× medium to make individual 110 µL cultures. Growth was monitored for 40 h by measuring OD_{600} every 15 or 30 minutes under constant linear shaking using Synergy H1 Plate Reader (BioTek Instruments). Raw OD_{600} growth data were exported using Gen5 software (BioTek Instruments) and analyzed in R (v4.2.1). Growth on a particular glycan for each strain was defined as follows: (i) no growth ("−") if OD_{600} at any point did not exceed 10% of maximum OD_{600} observed for this strain, (ii) weak growth ("w") if OD_{600} did not exceed 25% of maximum OD_{600}, (iii) growth ("+") if OD_{600} was above 25% of maximum OD_{600} (**Supplementary Table 11; Supplementary Fig. 2**).

Comparison of predicted phenotypes with growth data

Predicted carbohydrate utilization phenotypes were compared with growth phenotypes obtained in this work or published previously[40,48]. Cases when a predicted binary phenotype "1" matched with growth phenotypes "+/w" or "−" were considered true positives (TP) and false positives (FP), respectively. Cases when a predicted binary phenotype "0" matched with growth phenotypes "−" or "+/w" were considered true negatives (TN) and false negatives (FN), respectively. For the phenotype Bgl, growth ("+/w") on any of the three tested carbohydrates (gentiobiose, laminaritriose, sophorose) was considered as a match for the predicted binary phenotype "1". Standard metrics for evaluating two-class confusion matrices were calculated based on inferred TP, TN, FP, and FN counts.

Profiling of HMO consumption

Overnight cultures of selected *Bifidobacterium* strains grown in MRS-AC-Lac, were inoculated (0.5-1% [vol/vol]) in a fresh MRS-AC-Lac medium and grown to OD_{600}=0.6. Cells were harvested at 5,000 × g for 5 min, washed with sugar-free MRS-AC, and used to inoculate 200 µl of MRS-AC supplemented with an HMO mixture (10 mg/mL, isolated from pooled human milk from at least 25 different donors) in 96-well plates (Costar) at OD_{600}=0.005. Samples of culture supernatant (100 µl, triplicates) were collected at 8 and 24 h time points, filtered through Spin-X Centrifuge Tube Filter (0.22 µm, Costar), snap-frozen in liquid nitrogen and stored at −80°C until further use. Analysis of HMOs in supernatants was performed by HPLC with fluorescence detection as previously described for human milk[133] with slight modifications. 10 µL of media was spiked with maltose as an internal standard, lyophilized, and directly labeled with the fluorophore 2-aminobenzamide (2AB). 2AB-labeled HMOs were separated on a TSKgel Amide-80 column (2.0 mm ID x 15 cm, 3 µm, Tosoh Bioscience) and detected at 360 nm excitation and 425 nm emission. HMO peaks were annotated based on standard retention times and quantified in reference to the internal standard. HMO utilization at selected time points was calculated in reference to HMO concentrations in the medium control (without added cells) at the beginning of each experiment (0 h).

Transcriptional profiling (RNA-seq)

An overnight culture of *Bc. kashiwanohense* Bg42221_1E1 grown in MRS-AC-Glc was harvested at 5,000 × g for 5 min, washed in sugar-free MRS-AC, and used to inoculate MRS-AC supplemented with either 10 mg/mL of glucose (Glc), lactose (Lac), lacto-*N*-tetraose (LNT) or 5 mg/mL of tamarind xyloglucan (XGL) at OD_{600}=0.01. Samples (2 mL, biological triplicates) were collected at the early exponential phase (OD_{600}=0.35) and immediately pelleted in a prechilled centrifuge at 4,800 × g for 5 min. Cell pellets were snap-frozen in liquid nitrogen and stored at −80°C until further use. RNA was extracted as described previously[76]; the detailed protocol is given in *Supplementary Methods*. Ribosomal RNA was depleted with the NEBNext rRNA depletion kit for bacteria (New England Biolabs) and a set of 20 pooled sequence-specific probes for *Bc. kashiwanohense* Bg42221_1E1 designed with NEBNext Custom RNA Depletion Design Tool v1.0 (**Supplementary Table 14**). Barcoded libraries were made with NEBNext

Ultra II directional RNA library prep kit for Illumina (New England Biolabs). Libraries were pooled and sequenced (single-end 75-bp reads) on Illumina NextSeq 500 using the High Output V2 kit (Illumina). Sequencing data were analyzed as described previously[76] (additional details are provided in *Supplementary Methods*).

Data availability

Genome sequences of Bangladeshi *Bifidobacterium* isolates have been deposited under accession PRJNA1126848. Nucleotide FASTA and mcSEED-annotated protein FASTA files of 263 reference genomes are available on Figshare under accession doi.org/10.6084/m9.figshare.26053936. The RNA-seq data set has been deposited in the Gene Expression Omnibus under accession no. GSE239955.

Code availability

Code detailing the data analysis steps is available on GitHub (https://github.com/Arzamasov/compendium_manuscript) and **Supplementary Code File 1**.

Acknowledgments

We thank Robert Olson (Hack Biology LLC, Frankfort, IL) for helping with mcSEED maintenance, Kang Liu (SBP Genomics Core) for library preparation and sequencing, Marinela Elane (SBP) and Swetha Nakshatri (WUSTL) for technical assistance, and Glycom A/S and DSM for generously providing individual HMOs. Provision of the Bangladeshi strains used in this work was made possible by a long-standing collaborative research program between Washington University and members of International Center for Diarrheal Disease Research, Bangladesh (icddr,b) led by Dr. Tahmeed Ahmed. An MTA was established for the transfer of these strains to SBP Medical Discovery Institute.

Funding

This work was supported by NIH grant (RO1 DK30292) to A. O and J. I. G. and by the Bill & Melinda Gates Foundation (OPP1134649 to J.I.G.).

Main text figures

Fig. 1. ***In silico* metabolic reconstruction of carbohydrate metabolism in bifidobacteria. a**, 263 reference *Bifidobacterium* genomes stratified by taxa and the country of isolation. **b,** Overview of the metabolic reconstruction approach. We analyzed the representation of 565 metabolic functional roles (glycan transporters, CAZymes, downstream catabolic enzymes, transcription factors) and manually reconstructed 66 carbohydrate utilization pathways encoded in 263 reference *Bifidobacterium* genomes. Each corresponding strain was assigned a detailed pathway variant and a predicted binary glycan utilization phenotype based on designed rules (see example in Extended Data Fig. 1a). Binary phenotypes were compared with *in vitro* growth data to test the predictions. The obtained collection of reconstructed carbohydrate utilization pathways and predicted binary phenotypes were used in the phenotype projection pipeline. The pipeline produced two outputs: (i) a complete set of annotated proteins contributing to reconstructed glycan utilization pathways identified in additional 2710 *Bifidobacterium* genomes (Supplementary Table 9) and (ii) a Binary Phenotype Matrix (BPM) reflecting the inferred presence or absence of carbohydrate utilization pathways in each of the 2710 genomes (Supplementary Table 8).

Fig. 2. **Variability of predicted carbohydrate utilization phenotypes across a reference collection of *Bifidobacterium* strains**. The heatmap represents the hierarchical clustering of BPM for 66 carbohydrate utilization phenotypes (columns) predicted in 263 reference *Bifidobacterium* strains (rows). Full names of phenotype abbreviations are provided in Supplementary Table 3. Row annotation shows taxonomy; column annotations show phenotype classification (glycan type and origin). Lines point to the Bangladeshi strains *Bc. kashiwanohense* Bg42221_1E1, *Bl. infantis* Bg064.S07_13.C6, and *Bl. suis* Bg131.S11_17.F6 that are discussed in the text.

Fig. 3. **Comparison of predicted carbohydrate utilization phenotypes with *in vitro* growth data**. **a**, Growth curves of selected *Bifidobacterium* strains in MRS-AC supplemented with short-chain or long-chain fructooligosaccharides (scFOS and lcFOS, respectively) compared with predicted binary utilization phenotypes. Data represent the mean ± sd of three replicates. All growth curves are provided in Supplementary Fig. 2. **b**, Table showing the comparison between

40 predicted carbohydrate utilization phenotypes and *in vitro* growth data for 16 *Bifidobacterium* strains. Matching predictions (true positives [TP] and true negatives [TN]) are in green; non-matching (false positives [FP] and false negatives [FN]) are in yellow. The column annotation shows phenotype classification (glycan type). Summary of the data is provided in Supplementary Tables 12e; full names of phenotype abbreviations are provided in Supplementary Table 3; the list of carbohydrates used in growth experiments is provided Supplementary Table 10.

Fig. 4. **Genomics-based reconstruction and experimental validation of xyloglucan degradation pathway and its transcriptional regulation in *Bc. kashiwanohense* Bg42221_1E1**. **a**, Gene clusters potentially driving xyloglucan degradation (phenotype XGL) by *Bc. kashiwanohense* Bg42221_1E1. **b,** Reconstructed xyloglucan degradation pathway in *Bc. kashiwanohense* Bg42221_1E1: (i) xyloglucan chains are potentially cleaved by extracellular endo-β-1,4-glucanase Xgl5A, (ii) released oligosaccharides are transported and catabolized inside the cell by a coordinated action of CAZymes and downstream catabolic enzymes. **c,** Predicted binding motif of the transcription factor XglR potentially controlling the expression of xyloglucan metabolism genes. **d,** Growth curves of selected *Bifidobacterium* strains in MRS-AC supplemented with 0.5% tamarind xyloglucan. Data represent the mean ± sd of three replicates (two for *B. dentium* STL_TW14.1_LFYP24). **e,** Volcano plot depicting the $\log_2$FC of gene expression versus the $-\log_{10}P_{adj}$ for *Bc. kashiwanohense* Bg42221_1E1 grown in MRS-AC-XGL vs. MRS-AC-Lac. Criteria for calling differentially expressed genes were as follows: $P_{adj} < 0.01$ and $|\log_2FC| > 2$. Genes constituting the reconstructed XglR, XylR, and XosR regulons, as well as *xyn* cluster genes are colored.

Fig. 5. **Comparative genomic and experimental profiling of HMO utilization capabilities of select *Bifidobacterium* strains**. **a**, Schematic representation of HMO utilization genes in *Bc. kashiwanohense* Bg42221_1E1 and their homology to H1 cluster genes from *Bl. infantis* and *Bl. suis*. **c,** Growth curves of *Bifidobacterium* strains in MRS-AC-pHMO. Growth was monitored by measuring OD_{600} every 15 min. Data represent the mean ± sd of three replicates. **c,** Volcano plot depicting the $\log_2$FC of gene expression versus the $-\log_{10}P_{adj}$ for *Bc. kashiwanohense* Bg42221_1E1 grown in MRS-AC-LNT vs. MRS-AC-Lac. Criteria for calling differentially expressed genes were as follows: $P_{adj} < 0.01$ and $\log_2FC > 2$. Genes in the reconstructed NagR

regulon are colored. **d,** Profiling of HMO consumption (determined by HPLC) of select bifidobacterial strains at 8h and 24h time points. Data represent the percentage of utilized HMO (mean of three replicates). Total = total HMO utilized.

Fig. 6. **Analysis of variability of carbohydrate utilization pathways in 2973 *Bifidobacterium* genomes**. **a,** Non-metric multidimensional scaling (NMDS) of a Hamming distance matrix derived from the representation of metabolic pathways corresponding to 66 carbohydrate utilization phenotypes in 569 isolate plus 2404 metagenome-assembled *Bifidobacterium* genomes. Each data point represents a genome. Lines point to the outlier Bangladeshi strains *Bl. suis* Bg131.S11_17.F6 and *Bc. kashiwanohense* Bg42221_1E1 discussed in the text. **b**, Variability of predicted phenotypic richness (total number of different predicted phenotypes "1") at species and strain levels. Asterisks indicate significant ($p<0.001$) pairwise differences between the means of richness values of specific taxa; ns = non-significant (generalized linear model followed by a post-hoc test with Bonferroni correction). **c,** Heatmap depicting the representation of metabolic pathways corresponding to 66 predicted carbohydrate utilization phenotypes at species/subspecies levels. c within each taxon that encode the respective carbohydrate utilization pathway.

Extended data figures

Extended Data Fig. 1. ***In silico* subsystem-based annotation and prediction of carbohydrate utilization phenotypes**. **a**, (1) A set of reference *Bifidobacterium* genomes (Supplementary Table 1) and a list of curated functional roles (Supplementary Table 4) were used to populate mcSEED subsystems capturing catabolic pathways for various glycans. (2) Carbohydrate utilization pathways were reconstructed based on functional gene annotation using homology-based methods and three genome context techniques: (i) clustering of genes on the chromosome (operons), (ii) co-regulation of genes by a common transcriptional regulator, and (iii) co-occurrence of genes in a set of related genomes. Example: reconstruction of xylose (Xyl) and xylooligosaccharide (XOS) utilization pathways *in Bl. infantis* strains. (3) Each strain was assigned a pathway variant based on designed rules (Supplementary Table 5). All pathway variants were then converted to predicted binary phenotypes (1 or 0) to create a Binary Phenotype Matrix (BPM). **b**, Percentage of selected CAZyme families (GHs, CEs, PLs) captured in the *in silico* metabolic reconstructions for each of the 263 reference *Bifidobacterium* strains. Each data point represents a strain, **c**, Principal component analysis of CAZyme representation (at subfamily level) in 263 reference *Bifidobacterium* strains. Lines indicate specific Bangladeshi strains discussed in the text.

Extended Data Fig. 2. **Additional phylogenetic analysis of reference *B. longum* and *B. catenulatum* genomes**. **a,** Pairwise ANI indices calculated for a set of 26 *B. longum* genomes. For comparative purposes, this set contained 15 genomes from the reference collection (names are in bold) and 11 additional genomes, mostly of non-human origin, including the type strains of *Bl. suis* and *Bl. suillum*. Black lines indicate the taxonomic grouping used in this work, whereas blue lines indicate additional subspecies-level delineation proposed in another study[53]. **b,** Pairwise ANI indices calculated for 10 *B. catenulatum* and *B. sp002742445* genomes. Black lines indicate the taxonomic grouping used in this work. NCBI and GTDB taxonomy assignments are in blue.

Extended Data Fig. 3. **Representation of metabolic pathways in 26 *B. longum* genomes**. For comparative purposes, this set contains 15 genomes from the reference collection (names are in

bold) and 11 additional genomes, mostly of non-human origin, including type strains of *Bl. suis* and *Bl. suillum*. Column annotations show phenotype classification. Row annotations show taxonomy; the left column of the row annotation shows the taxonomic grouping used in this work, whereas the right column shows additional subspecies-level delineation proposed by Modesto et al[53]. **a,** Representation of 66 carbohydrate utilization pathways. Full names of phenotype abbreviations are provided in Supplementary Table 3. **b,** Representation of 29 additional pathways, including biosynthesis of amino acids and B vitamins, as well as urea utilization.

Extended Data Fig. 4. **Genomics-based reconstruction of pathways involved in (arabino)xylan and β-mannan degradation and the utilization of released oligosaccharides.** **a**, Alignment of the amino acid sequences of select endo-β-1,4-glucanases (xyloglucanases): PpXG5 from *Paenibacillus pabuli* XG5 (characterized), Xgl5A and Xgl5B from *Bc. kashiwanohense* Bg42221_1E1 (putative). Conserved catalytic amino acid residues are noted in bold red, and conserved residues forming contacts with the glycan chain are in bold. **b,** Gene clusters potentially driving the degradation of (arabino)xylan and the utilization of (arabino)xylooligosaccharides by *Bc. kashiwanohense* Bg42221_1E1. **c,** Predicted binding motifs of predicted TFs controlling the expression of the gene clusters in **b**. Motifs were built based on operator sequences in the RegPrecise database[121] and identified in this work (Supplementary Table 16). **d**, Gene clusters potentially driving β-mannose oligosaccharide utilization (phenotype bMnOS) by *B. breve* Bg41721_1C11 and β-mannan degradation (phenotype bMAN) by *B. dentium* STL_TW14.1_LFYP24. Orthologous genes are linked, and the link color intensity represents the amino acid identity between corresponding protein products. **e**, Reconstructed β-mannose oligosaccharide utilization and β-mannan degradation pathways. Names of previously characterized pathway elements are in bold. **f**, Growth curves of selected *Bifidobacterium* strains in MRS-AC supplemented with 0.5% mannotriose or 0.5% konjac glucomannan. Data represent the mean ± sd of three replicates. **g,** Predicted MnbR binding motif based on operator sequences identified in this work (Supplementary Table 16).

Extended Data Fig. 5. **Genomics-based reconstruction of pathways involved in the utilization of monosaccharide derivatives**. **a,** Gene cluster (*nan*) driving sialic acid utilization

by *B. breve* strains (phenotype Neu5Ac). **b,** Sialic acid utilization pathway in *B. breve*. **c**, Growth curves of selected *B. breve* strains in MRS-AC supplemented with 1% *N*-acetylneuraminic acid. Data represent the mean ± sd of three replicates. **d,** Gene clusters potentially driving D-glucuronate utilization by *Bl. infantis* JG_Bg463 and glucuronide utilization by *B. breve* Bg41721_1C11 (phenotypes GlcA and GluA, respectively). Orthologous genes are linked, and the link color represents the amino acid identity between corresponding protein products. **e,** Reconstructed D-glucuronate and glucuronide utilization pathways. **f**, Growth curves of selected *Bifidobacterium* strains in MRS-AC supplemented with 1% D-glucuronic acid. Data represent the mean ± sd of three replicates. **g**, Gene clusters potentially driving the utilization of (i) sugar alcohols D-mannitol, D-sorbitol, and *myo*-inositol (phenotypes Mtl, Stl, and Ino), (ii) sugar acids D-gluconate and D-galactonate (phenotypes Gco and Glt), (iii) L-ascorbate (phenotype Asc) by various bifidobacteria. **h,** Reconstructed utilization pathways for sugar alcohols and acids.

Extended Data Fig. 6. The analysis of variability of metabolic pathways extended to 2973 *Bifidobacterium* genomes. **a,** Heatmap depicting the representation of metabolic pathways corresponding to (i) B vitamin and amino acid biosynthesis, (ii) urea utilization (Urea_d) at species/subspecies levels. The coloring shows the percentage of genomes harboring a specific pathway (corresponding to a binary phenotype "1") within each taxon. **b**, Hierarchical clustering of BPM with the representation of metabolic pathways corresponding to 66 carbohydrate utilization phenotypes in 95 *B. catenulatum* and *B. sp002742445* genomes. Full names of phenotype abbreviations are provided in Supplementary Table 3.

Extended Data Fig. 7. Pathway enrichment analysis for 2973 *Bifidobacterium* genomes. Pathways significantly enriched (Fisher's exact test, $P_{adj} \leq 0.01$) in specified groups are shown. Certain pathways have infinite odds ratios because they are found only in one of the groups. **a**, Comparison between all genomes recovered from "Westernized (age < 3)" vs. "Westernized (age ≥ 3)" groups. **b**, Comparison between all genomes recovered from "non-Westernized (age < 3) " vs. "non-Westernized (age ≥ 3)" groups. **c**, Comparison between all genomes recovered from "Westernized (age < 3)" vs. "non-Westernized (age < 3)" groups. **d**, Comparison between genomes from selected taxa recovered from "Westernized" vs. "non-Westernized" groups.

Supplementary figures

Supplementary Fig. 1. **Phylogenetic tree of 263 reference *Bifidobacterium* genomes**. The tree is based on the alignment of concatenated nucleotide sequences of 487 core genes.

Supplementary Fig. 2. **Growth curves of 16 *Bifidobacterium* strains in MRS-AC supplemented with various carbohydrates**. Data are mean and 95% CI of three replicates. Labels represent short names of tested phenotypes; the exact substrates used in the growth assay are provided in Supplementary Table 10. "No sugar" represents wells containing MRS-AC without any added carbohydrate (negative control). Red lines mark the "no growth" threshold (10% of maximum OD_{600} observed for a particular strain). Orange lines mark the "weak growth" threshold (25% of maximum OD_{600} observed for this strain).

Supplementary tables

Supplementary Table 1. List of reference *Bifidobacterium* genomes used in metabolic reconstruction.

Supplementary Table 2. List of additional 2710 *Bifidobacterium* genomes used in the phenotype projection pipeline.

Supplementary Table 3. Glossary for predicted carbohydrate utilization phenotypes.

Supplementary Table 4. Curated list of functional roles of carbohydrate utilization pathways.

Supplementary Table 5. Rules used for the assignment of predicted carbohydrate utilization phenotypes.

Supplementary Table 6. Binary Phenotype Matrix (BPM) depicting the representation of metabolic pathways corresponding to 70 predicted carbohydrate utilization phenotypes in 263 reference *Bifidobacterium* genomes.

Supplementary Table 7. Representation of functional roles comprising carbohydrate utilization pathways in 263 reference *Bifidobacterium* genomes

Supplementary Table 8. Binary Phenotype Matrix (BPM) depicting the presence/absence of metabolic pathways corresponding to 66 predicted carbohydrate utilization phenotypes in additional 2710 *Bifidobacterium* genomes.

Supplementary Table 9. Representation of functional roles comprising carbohydrate utilization pathways in additional 2710 *Bifidobacterium* genomes.

Supplementary Table 10. Carbohydrates used in *in vitro* growth experiments.

Supplementary Table 11. Raw *in vitro* growth data for 16 *Bifidobacterium* strains.

Supplementary Table 12. Comparison between predicted carbohydrate utilization phenotypes and *in vitro* growth data.

Supplementary Table 13. Human milk oligosaccharide (HMO) consumption by select *Bifidobacterium* strains.

Supplementary Table 14. Custom probes used for the depletion of *Bc. kashiwanohense* Bg42221_1E1 rRNA.

Supplementary Table 15. Differential gene expression in *Bc. kashiwanohense* Bg42221_1E1 grown in MRS-AC supplemented with various carbon sources.

Supplementary Table 16. Reconstructed regulons and predicted operators of select transcription factors.

Supplementary Table 17. Binary Phenotype Matrix (BPM) depicting the presence/absence of metabolic pathways corresponding to the biosynthesis of essential metabolites (B vitamins and amino acids) and urea utilization in 2973 *Bifidobacterium* genomes.

References

1. Alessandri, G., van Sinderen, D. & Ventura, M. The genus bifidobacterium: From genomics to functionality of an important component of the mammalian gut microbiota running title: Bifidobacterial adaptation to and interaction with the host. *Comput Struct Biotechnol J* **19**, 1472–1487 (2021).
2. Odamaki, T. *et al.* Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. *BMC Microbiol.* **16**, 90 (2016).
3. Arboleya, S., Watkins, C., Stanton, C. & Ross, R. P. Gut Bifidobacteria Populations in Human Health and Aging. *Front Microbiol* **7**, 1204 (2016).
4. Stewart, C. J. *et al.* Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583–588 (2018).
5. Ward, R. E., Niñonuevo, M., Mills, D. A., Lebrilla, C. B. & German, J. B. In vitro fermentability of human milk oligosaccharides by several strains of bifidobacteria. *Mol Nutr Food Res* **51**, 1398–1405 (2007).
6. Asakuma, S. *et al.* Physiology of consumption of human milk oligosaccharides by infant gut-associated bifidobacteria. *J. Biol. Chem.* **286**, 34583–34592 (2011).
7. James, K., Motherway, M. O., Bottacini, F. & van Sinderen, D. Bifidobacterium breve UCC2003 metabolises the human milk oligosaccharides lacto-N-tetraose and lacto-N-neo-tetraose through overlapping, yet distinct pathways. *Sci Rep* **6**, (2016).
8. Sakanaka, M. *et al.* Evolutionary adaptation in fucosyllactose uptake systems supports bifidobacteria-infant symbiosis. *Sci Adv* **5**, eaaw7696 (2019).
9. Arzamasov, A. A. & Osterman, A. L. Milk glycan metabolism by intestinal bifidobacteria: insights from comparative genomics. *Crit Rev Biochem Mol Biol* 1–23 (2023) doi:10.1080/10409238.2023.2182272.
10. Kujawska, M. *et al.* Succession of Bifidobacterium longum Strains in Response to a Changing Early Life Nutritional Environment Reveals Dietary Substrate Adaptations. *iScience* **23**, 101368 (2020).
11. Vatanen, T. *et al.* A distinct clade of Bifidobacterium longum in the gut of Bangladeshi children thrives during weaning. *Cell* (2022) doi:10.1016/j.cell.2022.10.011.

12. Orihara, K. *et al.* Characterization of Bifidobacterium kashiwanohense that utilizes both milk- and plant-derived oligosaccharides. *Gut Microbes* **15**, 2207455 (2023).
13. Sela, D. A. *et al.* The genome sequence of Bifidobacterium longum subsp. infantis reveals adaptations for milk utilization within the infant microbiome. *Proc Natl Acad Sci U S A* **105**, 18964–18969 (2008).
14. Taft, D. H. *et al.* Bifidobacterium Species Colonization in Infancy: A Global Cross-Sectional Comparison by Population History of Breastfeeding. *Nutrients* **14**, 1423 (2022).
15. Olm, M. R. *et al.* Robust variation in infant gut microbiome assembly across a spectrum of lifestyles. *Science* **376**, 1220–1223 (2022).
16. Derrien, M. *et al.* Gut microbiome function and composition in infants from rural Kenya and association with human milk oligosaccharides. *Gut Microbes* **15**, 2178793 (2023).
17. Tannock, G. W. *et al.* Comparison of the compositions of the stool microbiotas of infants fed goat milk formula, cow milk-based formula, or breast milk. *Appl. Environ. Microbiol.* **79**, 3040–3048 (2013).
18. Casaburi, G. *et al.* Metagenomic insights of the infant microbiome community structure and function across multiple sites in the United States. *Sci Rep* **11**, 1472 (2021).
19. Fukuda, S. *et al.* Bifidobacteria can protect from enteropathogenic infection through production of acetate. *Nature* **469**, 543–547 (2011).
20. Casaburi, G. *et al.* Early-life gut microbiome modulation reduces the abundance of antibiotic-resistant bacteria. *Antimicrob Resist Infect Control* **8**, 131 (2019).
21. Hirano, R. *et al.* Next-generation prebiotic promotes selective growth of bifidobacteria, suppressing Clostridioides difficile. *Gut Microbes* **13**, 1973835 (2021).
22. Belenguer, A. *et al.* Two routes of metabolic cross-feeding between Bifidobacterium adolescentis and butyrate-producing anaerobes from the human gut. *Appl. Environ. Microbiol.* **72**, 3593–3599 (2006).
23. Rios-Covian, D., Gueimonde, M., Duncan, S. H., Flint, H. J. & de los Reyes-Gavilan, C. G. Enhanced butyrate formation by cross-feeding between Faecalibacterium prausnitzii and Bifidobacterium adolescentis. *FEMS Microbiol. Lett.* **362**, (2015).
24. Meng, D. *et al.* Indole-3-lactic acid, a metabolite of tryptophan, secreted by Bifidobacterium longum subspecies infantis is anti-inflammatory in the immature intestine. *Pediatr Res* **88**, 209–217 (2020).

25. Henrick, B. M. *et al.* Bifidobacteria-mediated immune system imprinting early in life. *Cell* **184**, 3884-3898.e11 (2021).
26. Laursen, M. F. *et al.* Bifidobacterium species associated with breastfeeding produce aromatic lactic acids in the infant gut. *Nat Microbiol* **6**, 1367–1382 (2021).
27. Shiver, A. L. *et al.* A mutant fitness compendium in Bifidobacteria reveals molecular determinants of colonization and host-microbe interactions. 2023.08.29.555234 Preprint at https://doi.org/10.1101/2023.08.29.555234 (2023).
28. Maldonado-Gómez, M. X. *et al.* Stable Engraftment of Bifidobacterium longum AH1206 in the Human Gut Depends on Individualized Features of the Resident Microbiome. *Cell Host Microbe* **20**, 515–526 (2016).
29. Frese, S. A. *et al.* Persistence of Supplemented Bifidobacterium longum subsp. infantis EVC001 in Breastfed Infants. *mSphere* **2**, e00501-17 (2017).
30. Alcon-Giner, C. *et al.* Microbiota Supplementation with Bifidobacterium and Lactobacillus Modifies the Preterm Infant Gut Microbiota and Metabolome: An Observational Study. *Cell Rep Med* **1**, 100077 (2020).
31. Beck, L. C. *et al.* Strain-specific impacts of probiotics are a significant driver of gut microbiome development in very preterm infants. *Nat Microbiol* **7**, 1525–1535 (2022).
32. Barratt, M. J. *et al.* Bifidobacterium infantis treatment promotes weight gain in Bangladeshi infants with severe acute malnutrition. *Sci Transl Med* **14**, eabk1107 (2022).
33. Button, J. E. *et al.* Dosing a synbiotic of human milk oligosaccharides and B. infantis leads to reversible engraftment in healthy adult microbiomes without antibiotics. *Cell Host Microbe* **30**, 712-725.e7 (2022).
34. Button, J. E. *et al.* Precision modulation of dysbiotic adult microbiomes with a human-milk-derived synbiotic reshapes gut microbial composition and metabolites. *Cell Host Microbe* **31**, 1523-1538.e10 (2023).
35. Milani, C. *et al.* Bifidobacteria exhibit social behavior through carbohydrate resource sharing in the gut. *Sci Rep* **5**, 15782 (2015).
36. Duranti, S. *et al.* Evaluation of genetic diversity among strains of the human gut commensal Bifidobacterium adolescentis. *Sci Rep* **6**, 23971 (2016).
37. Albert, K., Rani, A. & Sela, D. A. Comparative Pangenomics of the Mammalian Gut Commensal Bifidobacterium longum. *Microorganisms* **8**, (2019).

38. Lugli, G. A. *et al.* Decoding the Genomic Variability among Members of the Bifidobacteriumdentium Species. *Microorganisms* **8**, (2020).
39. Liu, J., Li, W., Yao, C., Yu, J. & Zhang, H. Comparative genomic analysis revealed genetic divergence between Bifidobacterium catenulatum subspecies present in infant versus adult guts. *BMC Microbiol* **22**, 158 (2022).
40. Arboleya, S. *et al.* Gene-trait matching across the Bifidobacterium longum pan-genome reveals considerable diversity in carbohydrate catabolism among human infant strains. *BMC Genomics* **19**, 33 (2018).
41. Bottacini, F. *et al.* Comparative genomics and genotype-phenotype associations in Bifidobacterium breve. *Sci Rep* **8**, (2018).
42. Liu, S. *et al.* Gene-Phenotype Associations Involving Human-Residential Bifidobacteria (HRB) Reveal Significant Species- and Strain-Specificity in Carbohydrate Catabolism. *Microorganisms* **9**, 883 (2021).
43. Magnúsdóttir, S. *et al.* Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat Biotechnol* **35**, 81–89 (2017).
44. Devika, N. T. & Raman, K. Deciphering the metabolic capabilities of Bifidobacteria using genome-scale metabolic models. *Sci Rep* **9**, 18222 (2019).
45. Schöpping, M., Gaspar, P., Neves, A. R., Franzén, C. J. & Zeidan, A. A. Identifying the essential nutritional requirements of the probiotic bacteria Bifidobacterium animalis and Bifidobacterium longum through genome-scale modeling. *NPJ Syst Biol Appl* **7**, 47 (2021).
46. Overbeek, R. *et al.* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research* **33**, 5691–5702 (2005).
47. Rodionov, D. A. *et al.* Micronutrient Requirements and Sharing Capabilities of the Human Gut Microbiome. *Front Microbiol* **10**, 1316 (2019).
48. Bottacini, F. *et al.* Comparative genomics of the Bifidobacterium breve taxon. *BMC Genomics* **15**, 170 (2014).
49. Mattarelli, P., Bonaparte, C., Pot, B. & Biavati, B. Proposal to reclassify the three biotypes of Bifidobacterium longum as three subspecies: Bifidobacterium longum subsp. longum subsp. nov., Bifidobacterium longum subsp. infantis comb. nov. and Bifidobacterium longum subsp. suis comb. nov. *Int J Syst Evol Microbiol* **58**, 767–772 (2008).

50. Chaplin, A. V. *et al.* Intraspecies Genomic Diversity and Long-Term Persistence of Bifidobacterium longum. *PLoS One* **10**, (2015).
51. Yanokura, E. *et al.* Subspeciation of Bifidobacterium longum by multilocus approaches and amplified fragment length polymorphism: Description of B. longum subsp. suillum subsp. nov., isolated from the faeces of piglets. *Syst Appl Microbiol* **38**, 305–314 (2015).
52. Tarracchini, C. *et al.* Phylogenomic disentangling of the Bifidobacterium longum subsp. infantis taxon. *Microb Genom* **7**, (2021).
53. Modesto, M. *et al.* Bifidobacterium longum subsp. iuvenis subsp. nov., a novel subspecies isolated from the faeces of weaning infants. *Int J Syst Evol Microbiol* **73**, (2023).
54. Kitaoka, M., Tian, J. & Nishimoto, M. Novel Putative Galactose Operon Involving Lacto-N-Biose Phosphorylase in Bifidobacterium longum. *Appl Environ Microbiol* **71**, 3158–3162 (2005).
55. Nishimoto, M. & Kitaoka, M. Identification of N-Acetylhexosamine 1-Kinase in the Complete Lacto-N-Biose I/Galacto-N-Biose Metabolic Pathway in Bifidobacterium longum. *Appl Environ Microbiol* **73**, 6444–6449 (2007).
56. O'Connell Motherway, M. *et al*. Characterization of ApuB, an extracellular type II amylopullulanase from Bifidobacterium breve UCC2003. *Appl. Environ. Microbiol.* **74**, 6271–6279 (2008).
57. Kashima, T. *et al.* Identification of difructose dianhydride I synthase/hydrolase from an oral bacterium establishes a novel glycoside hydrolase family. *J Biol Chem* **297**, 101324 (2021).
58. Sakurama, H. *et al*. Lacto-N-biosidase encoded by a novel gene of Bifidobacterium longum subspecies longum shows unique substrate specificity and requires a designated chaperone for its active expression. *J. Biol. Chem.* **288**, 25194–25206 (2013).
59. Garrido, D. *et al.* A novel gene cluster allows preferential utilization of fucosylated milk oligosaccharides in Bifidobacterium longum subsp. longum SC596. *Sci Rep* **6**, (2016).
60. Ojima, M. N. *et al.* Priority effects shape the structure of infant-type Bifidobacterium communities on human milk oligosaccharides. *ISME J* (2022) doi:10.1038/s41396-022-01270-3.
61. Fujita, K., Sakaguchi, T., Sakamoto, A., Shimokawa, M. & Kitahara, K. Bifidobacterium longum subsp. longum Exo-β-1,3-Galactanase, an enzyme for the degradation of type II arabinogalactan. *Appl. Environ. Microbiol.* **80**, 4577–4584 (2014).

62. Fujita, K. *et al.* Degradative enzymes for type II arabinogalactan side chains in Bifidobacterium longum subsp. longum. *Appl. Microbiol. Biotechnol.* **103**, 1299–1310 (2019).
63. Komeno, M., Hayamizu, H., Fujita, K. & Ashida, H. Two Novel α-l-Arabinofuranosidases from Bifidobacterium longum subsp. longum Belonging to Glycoside Hydrolase Family 43 Cooperatively Degrade Arabinan. *Appl. Environ. Microbiol.* (2019) doi:10.1128/AEM.02582-18.
64. Komeno, M. *et al.* Two α-L-arabinofuranosidases from Bifidobacterium longum subsp. longum are involved in arabinoxylan utilization. *Appl Microbiol Biotechnol* **106**, 1957–1965 (2022).
65. Friess, L. *et al.* Two extracellular α-arabinofuranosidases are required for cereal-derived arabinoxylan metabolism by Bifidobacterium longum subsp. longum. *Gut Microbes* **16**, 2353229 (2024).
66. Sasaki, Y. *et al.* Characterization of a novel 3-O-α-d-galactosyl-α-l-arabinofuranosidase for the assimilation of gum arabic AGP in Bifidobacterium longum subsp. longum. *Appl Environ Microbiol* (2021) doi:10.1128/AEM.02690-20.
67. Sasaki, Y. *et al.* Mechanism of Cooperative Degradation of Gum Arabic Arabinogalactan Protein by Bifidobacterium longum Surface Enzymes. *Appl Environ Microbiol* **88**, e0218721 (2022).
68. James, K. *et al.* Metabolism of the predominant human milk oligosaccharide fucosyllactose by an infant gut commensal. *Sci Rep* **9**, 15427 (2019).
69. Tm, G. *et al.* Characterization and three-dimensional structures of two distinct bacterial xyloglucanases from families GH5 and GH12. *The Journal of biological chemistry* **282**, (2007).
70. J, R. *et al.* Mechanisms involved in xyloglucan catabolism by the cellulosome-producing bacterium Ruminiclostridium cellulolyticum. *Scientific reports* **6**, (2016).
71. LoCascio, R. G. *et al.* Glycoprofiling of bifidobacterial consumption of human milk oligosaccharides demonstrates strain specific, preferential consumption of small chain glycans secreted in early human lactation. *J Agric Food Chem* **55**, 8914–8919 (2007).

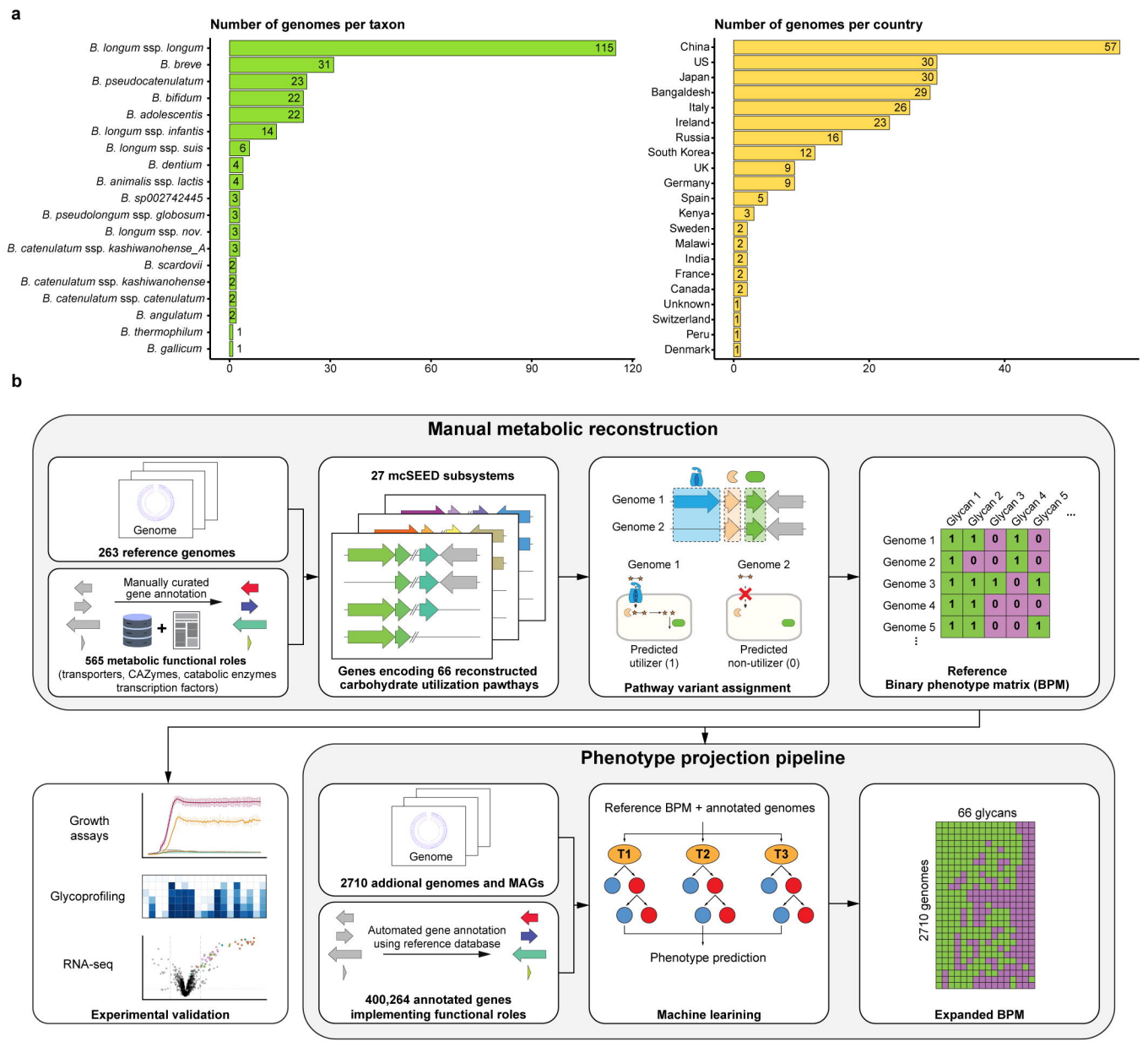
72. Garrido, D., Kim, J. H., German, J. B., Raybould, H. E. & Mills, D. A. Oligosaccharide Binding Proteins from Bifidobacterium longum subsp. infantis Reveal a Preference for Host Glycans. *PLoS One* **6**, (2011).
73. Sela, D. A. *et al.* An infant-associated bacterial commensal utilizes breast milk sialyloligosaccharides. *J. Biol. Chem.* **286**, 11909–11918 (2011).
74. Sela, D. A. *et al.* Bifidobacterium longum subsp. infantis ATCC 15697 α-Fucosidases Are Active on Fucosylated Human Milk Oligosaccharides. *Appl Environ Microbiol* **78**, 795–803 (2012).
75. Garrido, D., Ruiz-Moyano, S. & Mills, D. A. Release and utilization of N-acetyl-d-glucosamine from human milk oligosaccharides by Bifidobacterium longum subsp. infantis. *Anaerobe* **18**, 430–435 (2012).
76. Arzamasov, A. A. *et al.* Human Milk Oligosaccharide Utilization in Intestinal Bifidobacteria Is Governed by Global Transcriptional Regulator NagR. *mSystems* e0034322 (2022) doi:10.1128/msystems.00343-22.
77. Egan, M., O'Connell Motherway, M., Ventura, M. & van Sinderen, D. Metabolism of Sialic Acid by Bifidobacterium breve UCC2003. *Appl Environ Microbiol* **80**, 4414–4426 (2014).
78. Egan, M. *et al.* Cross-feeding by Bifidobacterium breve UCC2003 during co-cultivation with Bifidobacterium bifidum PRL2010 in a mucin-based medium. *BMC Microbiol* **14**, (2014).
79. LoCascio, R. G., Desai, P., Sela, D. A., Weimer, B. & Mills, D. A. Broad conservation of milk utilization genes in Bifidobacterium longum subsp. infantis as revealed by comparative genomic hybridization. *Appl Environ Microbiol* **76**, 7373–7381 (2010).
80. Saito, Y. *et al.* Multiple Transporters and Glycoside Hydrolases Are Involved in Arabinoxylan-Derived Oligosaccharide Utilization in Bifidobacterium pseudocatenulatum. *Appl Environ Microbiol* **86**, (2020).
81. Watanabe, Y. *et al.* Xylan utilisation promotes adaptation of Bifidobacterium pseudocatenulatum to the human gastrointestinal tract. *ISME COMMUN.* **1**, 1–11 (2021).
82. Khoroshkin, M. S., Leyn, S. A., Van Sinderen, D. & Rodionov, D. A. Transcriptional Regulation of Carbohydrate Utilization Pathways in the Bifidobacterium Genus. *Front Microbiol* **7**, (2016).
83. Hibberd, M. C. *et al.* Bioactive glycans in a microbiome-directed food for children with malnutrition. *Nature* **625**, 157–165 (2024).

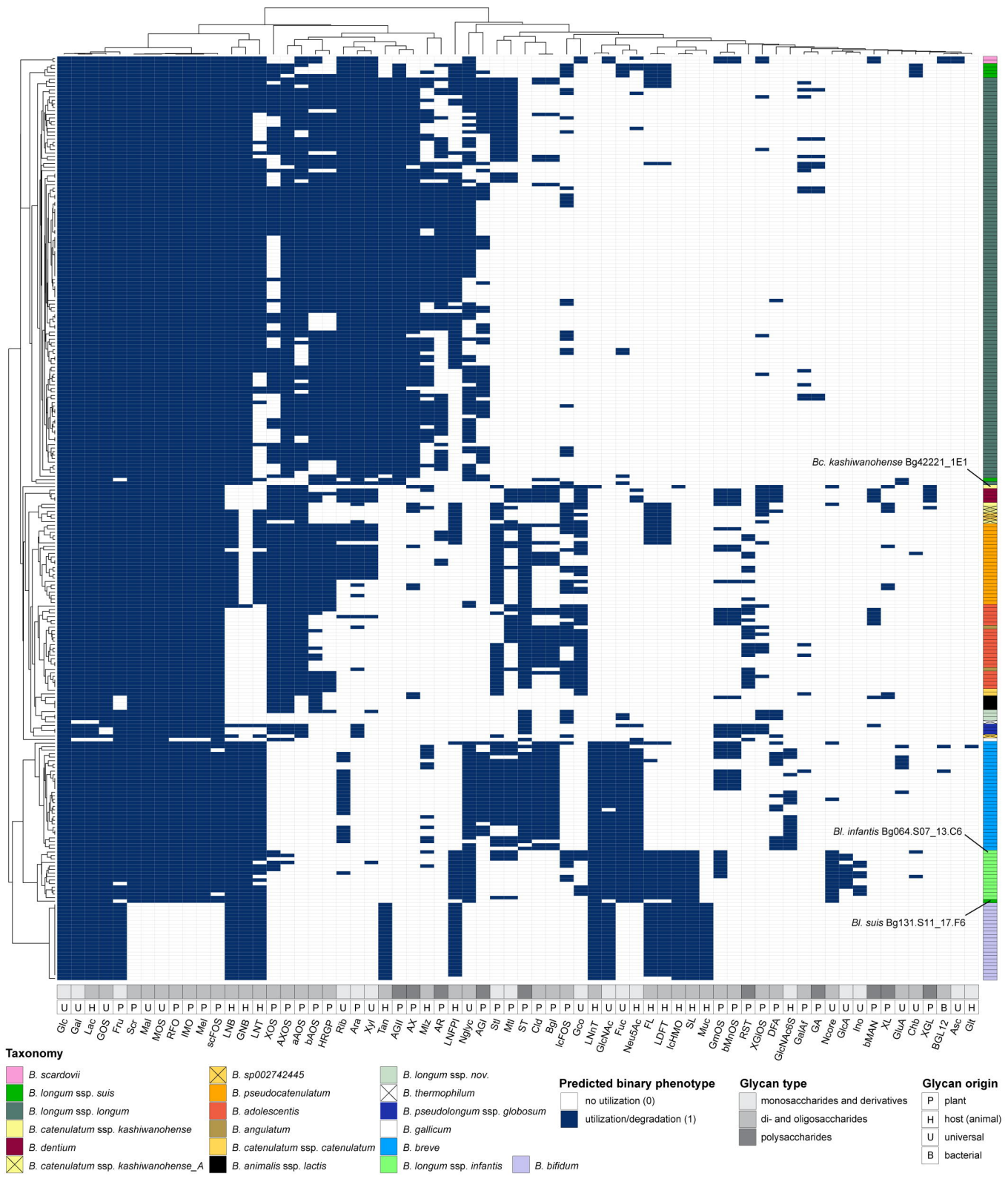
84. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662.e20 (2019).
85. Pokusaeva, K., Fitzgerald, G. F. & van Sinderen, D. Carbohydrate metabolism in Bifidobacteria. *Genes Nutr* **6**, 285–306 (2011).
86. Milani, C. *et al.* Genomic encyclopedia of type strains of the genus Bifidobacterium. *Appl Environ Microbiol* **80**, 6290–6302 (2014).
87. Odamaki, T. *et al.* Comparative Genomics Revealed Genetic Diversity and Species/Strain-Level Differences in Carbohydrate Metabolism of Three Probiotic Bifidobacterial Species. *Int J Genomics* **2015**, 567809 (2015).
88. Crociani, F., Alessandrini, A., Mucci, M. M. & Biavati, B. Degradation of complex carbohydrates by Bifidobacterium spp. *Int J Food Microbiol* **24**, 199–210 (1994).
89. Browne, H. P. *et al.* Boosting microbiome science worldwide could save millions of children's lives. *Nature* **625**, 237–240 (2024).
90. Forster, S. C. *et al.* A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat Biotechnol* **37**, 186–192 (2019).
91. Lin, X. *et al.* The genomic landscape of reference genomes of cultivated human gut bacteria. *Nat Commun* **14**, 1663 (2023).
92. Poyet, M. *et al.* A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat Med* **25**, 1442–1452 (2019).
93. Zou, Y. *et al.* 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat Biotechnol* **37**, 179–185 (2019).
94. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* **39**, 105–114 (2021).
95. Zeng, S. *et al.* A compendium of 32,277 metagenome-assembled genomes and over 80 million genes from the early-life human gut microbiome. *Nat Commun* **13**, 5139 (2022).
96. Jin, H. *et al.* A high-quality genome compendium of the human gut microbiome of Inner Mongolians. *Nat Microbiol* **8**, 150–161 (2023).
97. Kim, C. Y. *et al.* Human reference gut microbiome catalog including newly assembled genomes from under-represented Asian metagenomes. *Genome Med* **13**, 134 (2021).

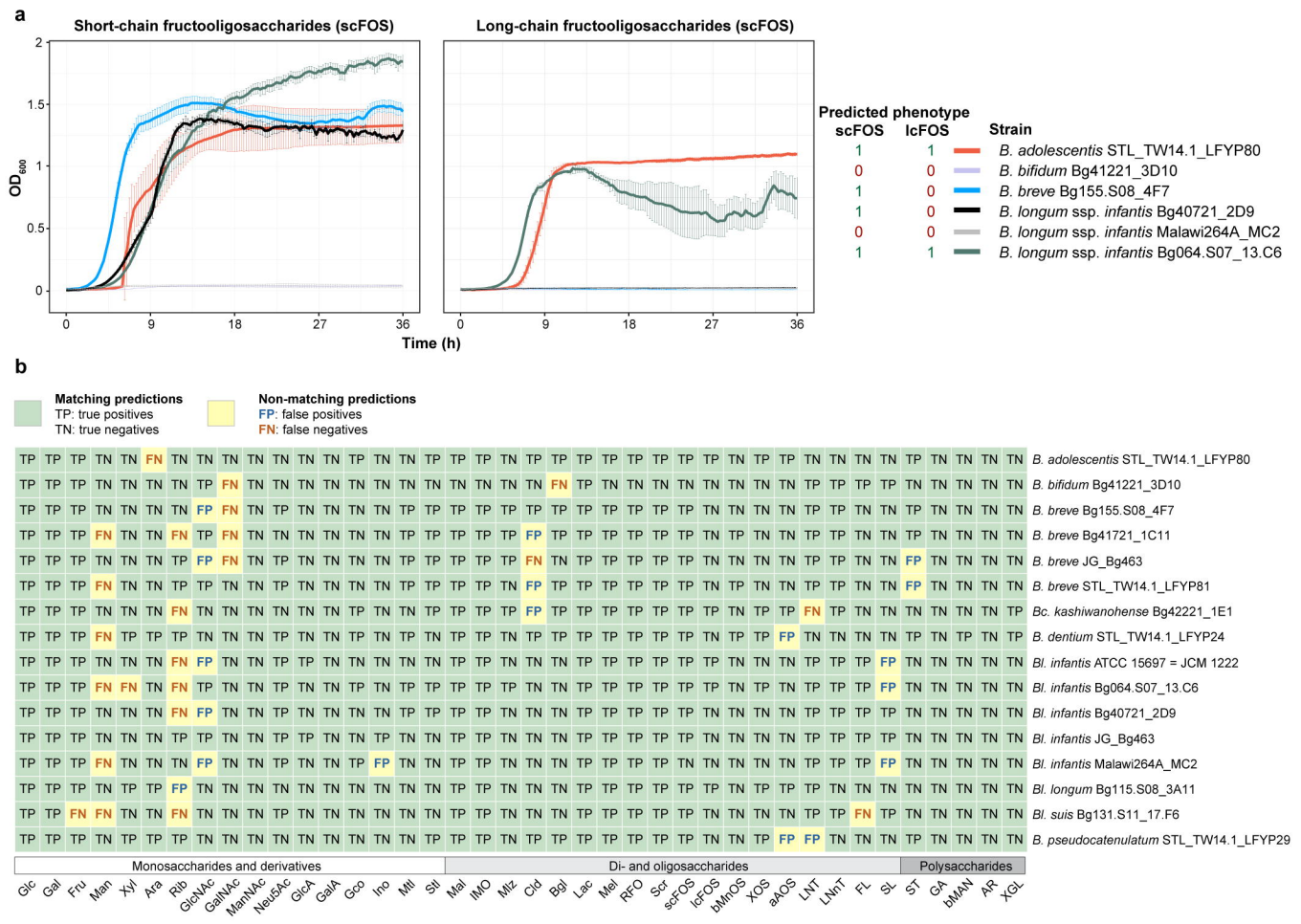
98. Gounot, J.-S. *et al.* Genome-centric analysis of short and long read metagenomes reveals uncharacterized microbiome diversity in Southeast Asians. *Nat Commun* **13**, 6044 (2022).
99. Price, M. N. & Arkin, A. P. Interactive Tools for Functional Annotation of Bacterial Genomes. 2024.04.15.589591 Preprint at https://doi.org/10.1101/2024.04.15.589591 (2024).
100. Zhu, L. *et al.* Captive Common Marmosets (Callithrix jacchus) Are Colonized throughout Their Lives by a Community of Bifidobacterium Species with Species-Specific Genomic Content That Can Support Adaptation to Distinct Metabolic Niches. *mBio* e0115321 (2021) doi:10.1128/mBio.01153-21.
101. Pudlo, N. A. *et al.* Phenotypic and Genomic Diversification in Complex Carbohydrate-Degrading Human Gut Bacteria. *mSystems* **7**, e0094721 (2022).
102. Wattam, A. R. *et al.* Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res* **45**, D535–D542 (2017).
103. Markowitz, V. M. *et al.* IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res* **40**, D115-122 (2012).
104. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043–1055 (2015).
105. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* **11**, 2864–2868 (2017).
106. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* **50**, D785–D794 (2022).
107. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
108. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).
109. Gehrig, J. L. *et al.* Effects of microbiota-directed foods in gnotobiotic animals and undernourished children. *Science* **365**, (2019).

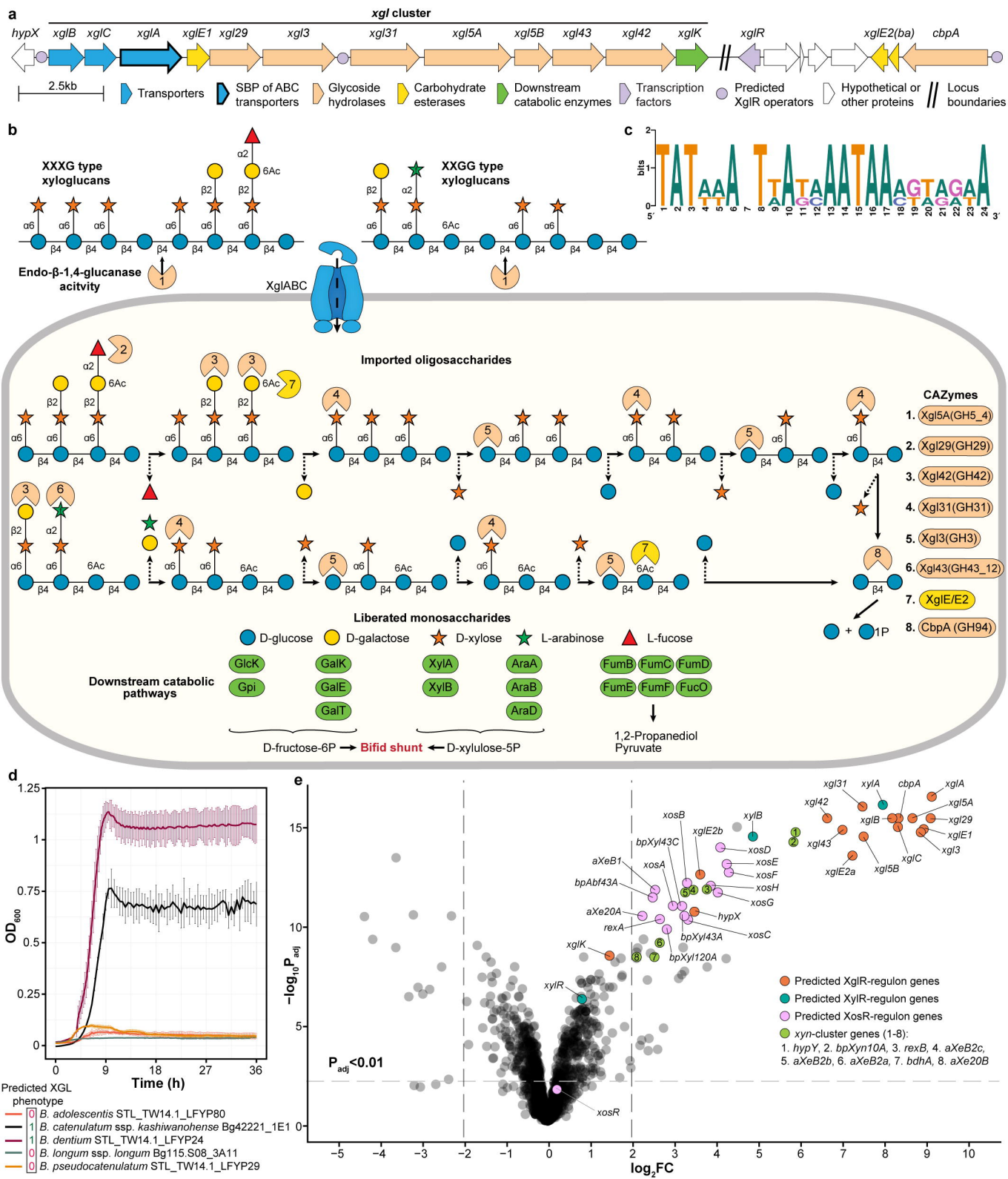
110. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* **21**, 180 (2020).
111. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**, 772–780 (2013).
112. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
113. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* **49**, W293–W296 (2021).
114. Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G. & Toth, I. K. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal. Methods* **8**, 12–24 (2015).
115. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132 (2016).
116. Overbeek, R. *et al.* The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **42**, D206-214 (2014).
117. Price, M. N. & Arkin, A. P. PaperBLAST: Text Mining Papers for Information about Homologs. *mSystems* **2**, e00039-17 (2017).
118. Saier, M. H. *et al.* The Transporter Classification Database (TCDB): 2021 update. *Nucleic Acids Res* **49**, D461–D467 (2021).
119. Drula, E. *et al.* The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res* **50**, D571–D577 (2022).
120. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457-462 (2016).
121. Novichkov, P. S. *et al.* RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. *Nucleic Acids Res* **38**, D111–D118 (2010).
122. Zheng, J. *et al.* dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Res* gkad328 (2023) doi:10.1093/nar/gkad328.
123. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).

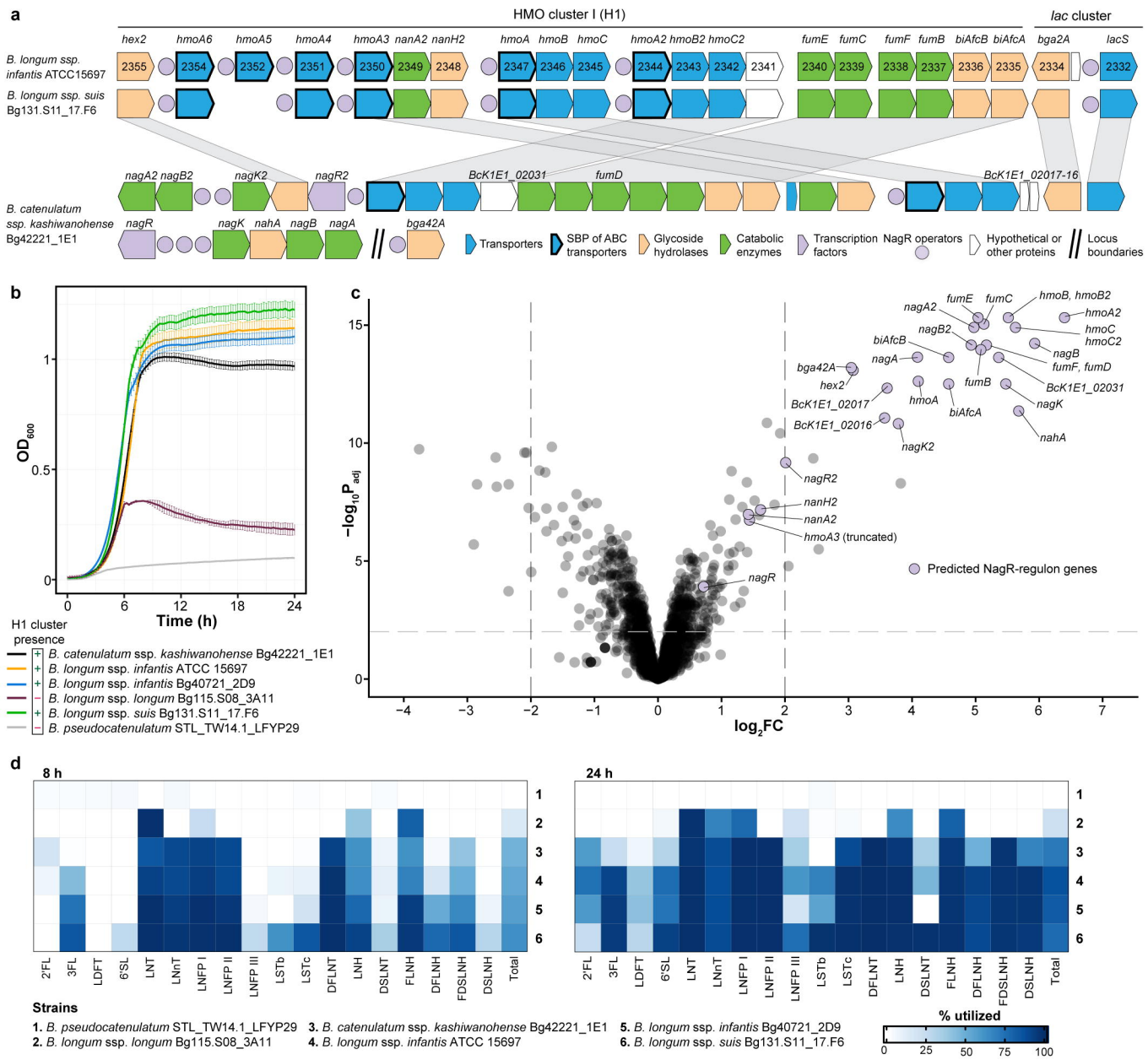
124. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
125. Abraham, A. *et al*. Machine learning for neuroimaging with scikit-learn. *Front Neuroinform* **8**, 14 (2014).
126. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **028**, (2008).
127. Oksanen, J. *et al.* vegan: Community Ecology Package. (2022).
128. Lenth, R. V. *et al.* emmeans: Estimated Marginal Means, aka Least-Squares Means. (2022).
129. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300 (1995).
130. Wickham, Hadley. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag, New York, 2016).
131. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
132. Gilchrist, C. L. M. & Chooi, Y.-H. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* **37**, 2473–2475 (2021).
133. Berger, P. K. *et al.* Stability of Human-Milk Oligosaccharide Concentrations Over 1 Week of Lactation and Over 6 Hours Following a Standard Meal. *J Nutr* **152**, 2727–2733 (2023).

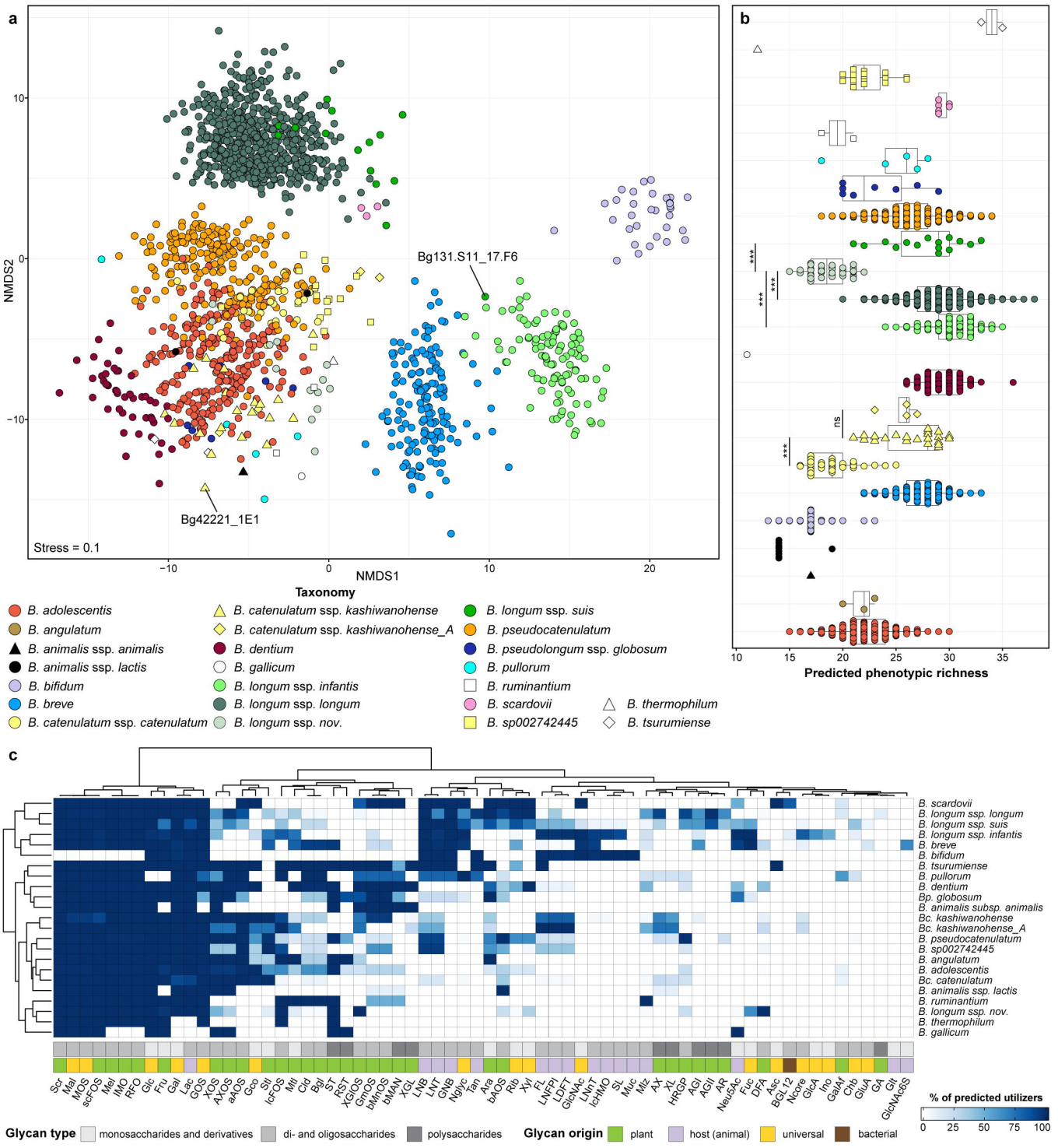

a
Number of genomes per taxon
B. longum ssp. longum 115
B. breve 31
B. pseudocatenulatum 23
B. bifidum 22
B. adolescentis 22
B. longum ssp. infantis 14
B. longum ssp. suis 6
B. dentium 4
B. animalis ssp. lactis 4
B. sp002742445 3
B. pseudolongum ssp. globosum 3
B. longum ssp. nov. 3
B. catenulatum ssp. kashiwanohense_A 3
B. scardovii 2
B. catenulatum ssp. kashiwanohense 2
B. catenulatum ssp. catenulatum 2
B. angulatum 2
B. thermophilum 1
B. gallicum 1
0 30 60 90 120
Number of genomes per country
China 57
US 30
Japan 30
Bangaldesh 29
Italy 26
Ireland 23
Russia 16
South Korea 12
UK 9
Germany 9
Spain 5
Kenya 3
Sweden 2
Malawi 2
India 2
France 2
Canada 2
Unknown 1
Switzerland 1
Peru 1
Denmark 1
0 20 40
b
Manual metabolic reconstruction
Genome
263 reference genomes
Manually curated gene annotation
565 metabolic functional roles
(transporters, CAZymes, catabolic enzymes transcription factors)
27 mcSEED subsystems
Genes encoding 66 reconstructed carbohydrate utilization pawthays
Genome 1
Genome 2
Predicted utilizer (1)
Predicted non-utilizer (0)
Pathway variant assignment
Glycan 1 Glycan 2 Glycan 3 Glycan 4 Glycan 5 ...
Genome 1 1 1 0 1 0
Genome 2 1 0 0 1 0
Genome 3 1 1 1 0 1
Genome 4 1 1 0 0 0
Genome 5 1 1 0 0 1
Reference Binary phenotype matrix (BPM)
Growth assays
Glycoprofiling
RNA-seq
Experimental validation
Phenotype projection pipeline
Genome
2710 addional genomes and MAGs
Automated gene annotation using reference database
400,264 annotated genes implementing functional roles
Reference BPM + annotated genomes
T1
T2
T3
Phenotype prediction
Machine learining
66 glycans
2710 genomes
Expanded BPM


Bc. kashiwanohense Bg42221_1E1
Bl. infantis Bg064.S07_13.C6
Bl. suis Bg131.S11_17.F6
Glc Gal Lac GOS Fru Scr Mal MOS RFO IMO Mel scFOS LNB GNB LNT XOS AXOS aAOS bAOS HRGP Rib Ara Xyl Tan AGII AX Mtz AR LNFPI Nglyc AGI Stl Mtl ST Cld Bgl lcFOS Gco LNnT GlcNAc Fuc Neu5Ac FL LDFT lctMO SL Muc GmOS bMnOS RST XGlOS DFA GlcNAc6S GalAf GA Ncore GlcA Ino bMAN XL GluA Crib XGL BGL12 Asc Glt
U U H U P P U U P P P P H H H P P P P P U P U H P P H P H U P P P P P P P U H U U H H H H H H P P P P P H P P U U U P P U U P B U H
Taxonomy
B. scardovii
B. longum ssp. suis
B. longum ssp. longum
B. catenulatum ssp. kashiwanohense
B. dentium
B. catenulatum ssp. kashiwanohense_A
B. sp002742445
B. pseudocatenulatum
B. adolescentis
B. angulatum
B. catenulatum ssp. catenulatum
B. animalis ssp. lactis
B. longum ssp. nov.
B. thermophilum
B. pseudolongum ssp. globosum
B. gallicum
B. breve
B. longum ssp. infantis
B. bifidum
Predicted binary phenotype
no utilization (0)
utilization/degradation (1)
Glycan type
monosaccharides and derivatives
di- and oligosaccharides
polysaccharides
Glycan origin
P plant
H host (animal)
U universal
B bacterial


a
Short-chain fructooligosaccharides (scFOS)
Long-chain fructooligosaccharides (scFOS)
OD600
Time (h)
Predicted phenotype
scFOS lcFOS Strain
1 1 B. adolescentis STL_TW14.1_LFYP80
0 0 B. bifidum Bg41221_3D10
1 0 B. breve Bg155.S08_4F7
1 0 B. longum ssp. infantis Bg40721_2D9
0 0 B. longum ssp. infantis Malawi264A_MC2
1 1 B. longum ssp. infantis Bg064.S07_13.C6
b
Matching predictions
TP: true positives
TN: true negatives
Non-matching predictions
FP: false positives
FN: false negatives
B. adolescentis STL_TW14.1_LFYP80
B. bifidum Bg41221_3D10
B. breve Bg155.S08_4F7
B. breve Bg41721_1C11
B. breve JG_Bg463
B. breve STL_TW14.1_LFYP81
Bc. kashiwanohense Bg42221_1E1
B. dentium STL_TW14.1_LFYP24
Bl. infantis ATCC 15697 = JCM 1222
Bl. infantis Bg064.S07_13.C6
Bl. infantis Bg40721_2D9
Bl. infantis JG_Bg463
Bl. infantis Malawi264A_MC2
Bl. longum Bg115.S08_3A11
Bl. suis Bg131.S11_17.F6
B. pseudocatenulatum STL_TW14.1_LFYP29
Monosaccharides and derivatives
Di- and oligosaccharides
Polysaccharides
Glc Gal Fru Man Xyl Ara Rib GlcNAc GalNAc ManNAc Neu5Ac GlcA GalA Gco Ino Mtl Stl Mal IMO Mlz Cld Bgl Lac Mel RFO Scr scFOS lcFOS bMnOS XOS aAOS LNT LNnT FL SL ST GA bMAN AR XGL


a
xgl cluster
hypX
xglB
xglC
xglA
xglE1
xgl29
xgl3
xgl31
xgl5A
xgl5B
xgl43
xgl42
xglK
xglR
xglE2(ba)
cbpA
2.5kb
Transporters
SBP of ABC transporters
Glycoside hydrolases
Carbohydrate esterases
Downstream catabolic enzymes
Transcription factors
Predicted XglR operators
Hypothetical or other proteins
Locus boundaries
b
XXXG type xyloglucans
XXGG type xyloglucans
Endo-β-1,4-glucanase acitvity
XglABC
Imported oligosaccharides
CAZymes
1. Xgl5A(GH5_4)
2. Xgl29(GH29)
3. Xgl42(GH42)
4. Xgl31(GH31)
5. Xgl3(GH3)
6. Xgl43(GH43_12)
7. XglE/E2
8. CbpA (GH94)
Liberated monosaccharides
D-glucose
D-galactose
D-xylose
L-arabinose
L-fucose
Downstream catabolic pathways
GlcK
Gpi
GalK
GalE
GalT
XylA
XylB
AraA
AraB
AraD
FumB
FumC
FumD
FumE
FumF
FucO
D-fructose-6P
Bifid shunt
D-xylulose-5P
1,2-Propanediol
Pyruvate
c
d
OD600
Time (h)
Predicted XGL phenotype
B. adolescentis STL_TW14.1_LFYP80
B. catenulatum ssp. kashiwanohense Bg42221_1E1
B. dentium STL_TW14.1_LFYP24
B. longum ssp. longum Bg115.S08_3A11
B. pseudocatenulatum STL_TW14.1_LFYP29
e
−log10Padj
log2FC
Padj<0.01
Predicted XglR-regulon genes
Predicted XylR-regulon genes
Predicted XosR-regulon genes
xyn-cluster genes (1-8):
1. hypY, 2. bpXyn10A, 3. rexB, 4. aXeB2c, 5. aXeB2b, 6. aXeB2a, 7. bdhA, 8. aXe20B


a
HMO cluster I (H1)
lac cluster
hex2 hmoA6 hmoA5 hmoA4 hmoA3 nanA2 nanH2 hmoA hmoB hmoC hmoA2 hmoB2 hmoC2 fumE fumC fumF fumB biAfcB biAfcA bga2A lacS
B. longum ssp. infantis ATCC15697
B. longum ssp. suis Bg131.S11_17.F6
2355 2354 2352 2351 2350 2349 2348 2347 2346 2345 2344 2343 2342 2341 2340 2339 2338 2337 2336 2335 2334 2332
nagA2 nagB2 nagK2 nagR2 BcK1E1_02031 fumD BcK1E1_02017-16
B. catenulatum ssp. kashiwanohense Bg42221_1E1
nagR nagK nahA nagB nagA bga42A
Transporters
SBP of ABC transporters
Glycoside hydrolases
Catabolic enzymes
Transcription factors
NagR operators
Hypothetical or other proteins
Locus boundaries
b
OD600
Time (h)
H1 cluster presence
+ B. catenulatum ssp. kashiwanohense Bg42221_1E1
+ B. longum ssp. infantis ATCC 15697
+ B. longum ssp. infantis Bg40721_2D9
− B. longum ssp. longum Bg115.S08_3A11
+ B. longum ssp. suis Bg131.S11_17.F6
− B. pseudocatenulatum STL_TW14.1_LFYP29
c
−log10Padj
log2FC
fumE fumC hmoB, hmoB2 nagA2 hmoA2 nagB2 hmoC hmoC2 biAfcB nagB nagA fumF, fumD bga42A fumB BcK1E1_02031 hex2 hmoA nagK BcK1E1_02017 biAfcA BcK1E1_02016 nahA nagK2 nagR2 nanH2 nanA2 hmoA3 (truncated) nagR
Predicted NagR-regulon genes
d
8 h
24 h
2'FL 3FL LDFT 6'SL LNT LNnT LNFP I LNFP II LNFP III LSTb LSTc DFLNT LNH DSLNT FLNH DFLNH FDSLNH DSLNH Total
Strains
1. B. pseudocatenulatum STL_TW14.1_LFYP29
2. B. longum ssp. longum Bg115.S08_3A11
3. B. catenulatum ssp. kashiwanohense Bg42221_1E1
4. B. longum ssp. infantis ATCC 15697
5. B. longum ssp. infantis Bg40721_2D9
6. B. longum ssp. suis Bg131.S11_17.F6
% utilized


a
NMDS2
NMDS1
Bg131.S11_17.F6
Bg42221_1E1
Stress = 0.1
Taxonomy
B. adolescentis
B. angulatum
B. animalis ssp. animalis
B. animalis ssp. lactis
B. bifidum
B. breve
B. catenulatum ssp. catenulatum
B. catenulatum ssp. kashiwanohense
B. catenulatum ssp. kashiwanohense_A
B. dentium
B. gallicum
B. longum ssp. infantis
B. longum ssp. longum
B. longum ssp. nov.
B. longum ssp. suis
B. pseudocatenulatum
B. pseudolongum ssp. globosum
B. pullorum
B. ruminantium
B. scardovii
B. sp002742445
B. thermophilum
B. tsurumiense
b
Predicted phenotypic richness
c
B. scardovii
B. longum ssp. longum
B. longum ssp. suis
B. longum ssp. infantis
B. breve
B. bifidum
B. tsurumiense
B. pullorum
B. dentium
Bp. globosum
B. animalis subsp. animalis
Bc. kashiwanohense
Bc. kashiwanohense_A
B. pseudocatenulatum
B. sp002742445
B. angulatum
B. adolescentis
Bc. catenulatum
B. animalis ssp. lactis
B. ruminantium
B. longum ssp. nov.
B. thermophilum
B. gallicum
Scr Mal MOS scFOS Mel IMO RFO Glc Fru Gal Lac GOS XOS AXOS aAOS Geo Stl lcFOS Mtl Cid Bgl 5T RST XGlOS GmOS bMnOS bMaN XGL LNB LNT GNB Ngyc Tan Ara bAOS Rib Xyl FL LNFPI LDFT GlcNAc LNnT lcHMO SL Muc Mtz AX XL HRGP AGI AGII AR Neu5Ac Fuc DFA Asc BGL12 Nacre GlcA Ino GalAf Chb GluA GA Glt GlcNAc6S
% of predicted utilizers
0 25 50 75 100
Glycan type
monosaccharides and derivatives
di- and oligosaccharides
polysaccharides
Glycan origin
plant
host (animal)
universal
bacterial

a

(1) Creating a genomic database of *Bifidobacterium* strains and metabolic functional roles

263 reference *Bifidobacterium* genomes
- 19 species and subspecies
- 19 new genomes of Bangladeshi isolates

27 mcSEED subsystems encompassing catabolic pathways for 66 glycans

Manually curated list of 565 functional roles of metabolic pathways
- 200 characterized in bifidobacteria (from 144 primary publications)
- 225 predicted (TCDB, CAZy, KEGG, RegPrecise, PATRIC)
- 140 predicted in this work

Glycan transporters (226) | Glycoside hydrolases and other CAZymes (188) | Downstream catabolic enzymes (67) | Transcription factors (84)

(2) Subsystem-based analysis of genomic context and reconstruction of carbohydrate utilization pathways

Example: reconstruction of pathways for xylose (Xyl) and xylooligosaccharide (XOS) utilization

***Bl. infantis* R0033**
Gene clusters and regulons
① *xylA* — *xosR* ② *xosA* *xosB* *xosC* *bpXyl43C*
① *xylF* *xylG* *xylH* — *xylB* ① *xylR*

***Bl. infantis* Bg064.S07_13.C6**
Gene clusters and regulons
① *xylA* — *xosR* ② *xosA* *xosB* *xosC* *bpXyl43C*
xylB ① *xylR*

***Bl. infantis* Malawi264A_MC2**
Gene clusters and regulons
xylB (truncated) ① *xylR*

Transporters | Glycoside hydrolases | Downstream catabolic enzymes | Transcription factors | ① ② Predicted XylR and XosR operators

XylFGH | XosABC | BpXyl43C | β4 | D-xylose | XylA | D-xylulose | XylB | D-xylulose-5P | Bifid shunt

(3) Assignment of pathway variants and binary carbohydrate utilization phenotypes based on the presence of signature genes

U = transporter + catabolic pathway
Xyl: *xylF* + *xylA* + *xylB*
XOS: *xosA* + *bpXyl43C* + *xylA* + *xylB*

P = catabolic pathway only
Xyl: *xylA* + *xylB*
XOS: *bpXyl43C* + *xylA* + *xylB*

N = no catabolic pathway

Pathway variants

Strain	Xyl	XOS
Bl. infantis R0033	U	U
Bl. infantis Bg064.S07_13.C6	P	U
Bl. infantis Malawi264A_MC2	N	N

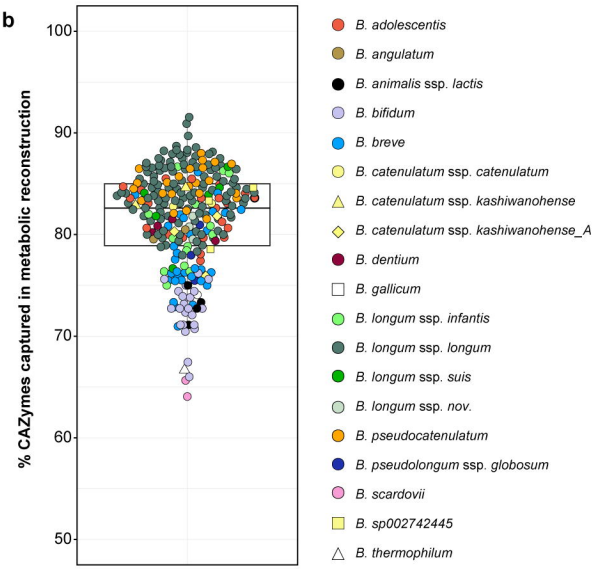
Conversion to binary phenotypes
U --> 1 (utilizer)
P --> 0 (non-utilizer)
N --> 0 (non-utilizer)

Binary phenotypes

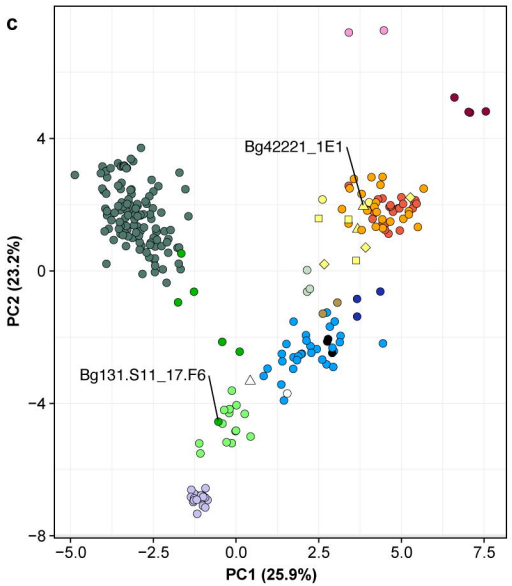
Strain	Xyl	XOS
Bl. infantis R0033	1	1
Bl. infantis Bg064.S07_13.C6	0	1
Bl. infantis Malawi264A_MC2	0	0

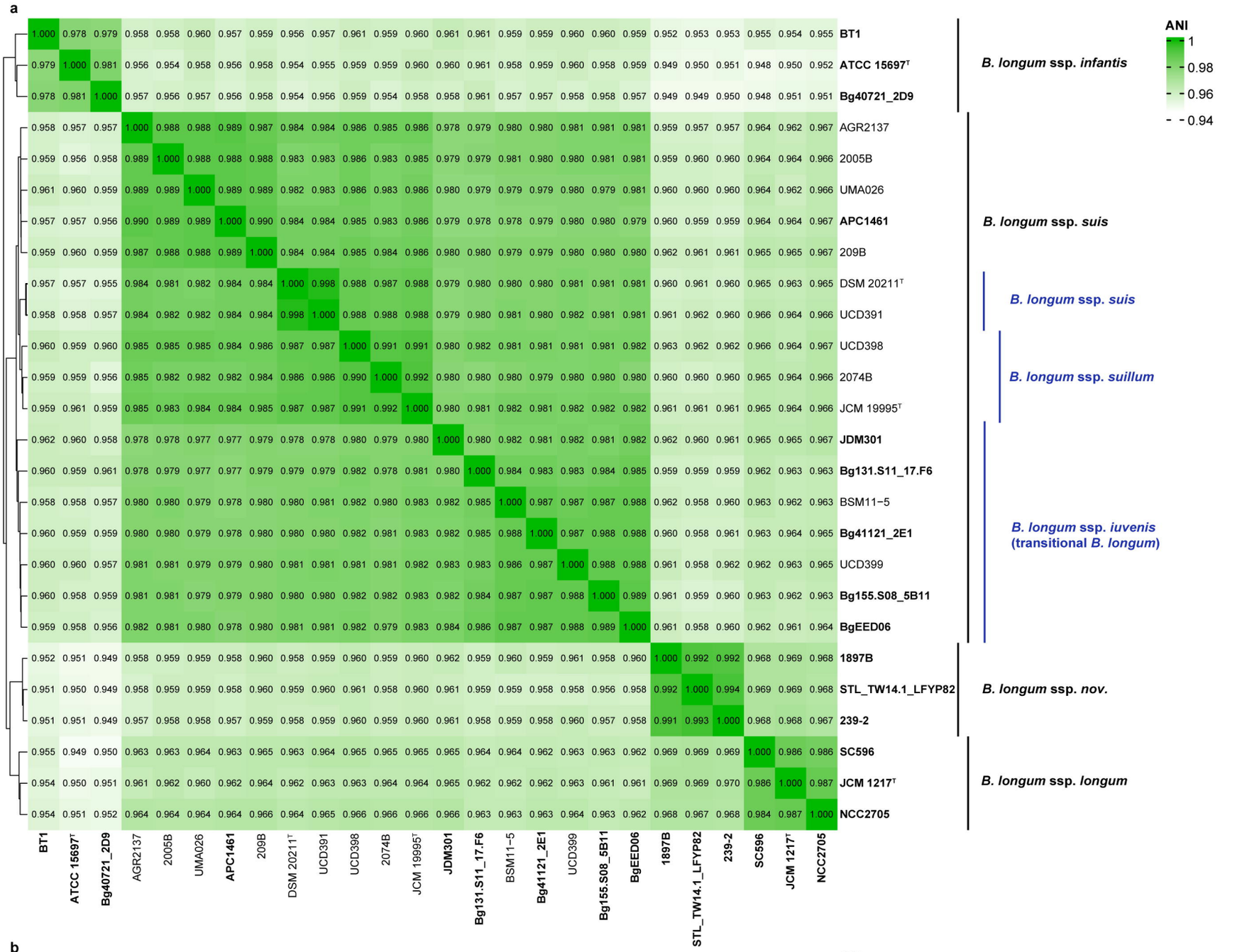
Binary Phenotype Matrix (BPM)
263 genomes
66 binary phenotypes

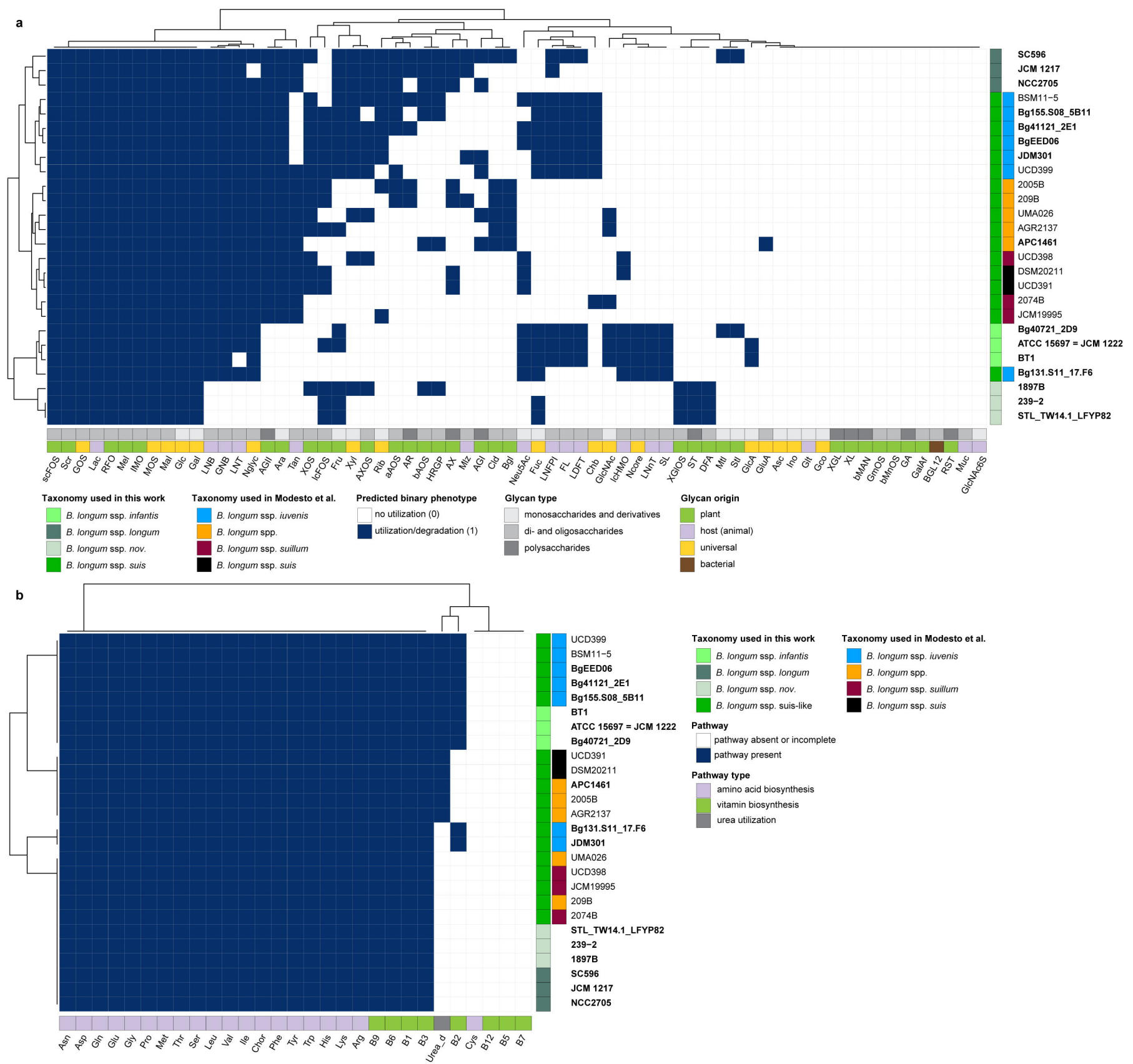
b

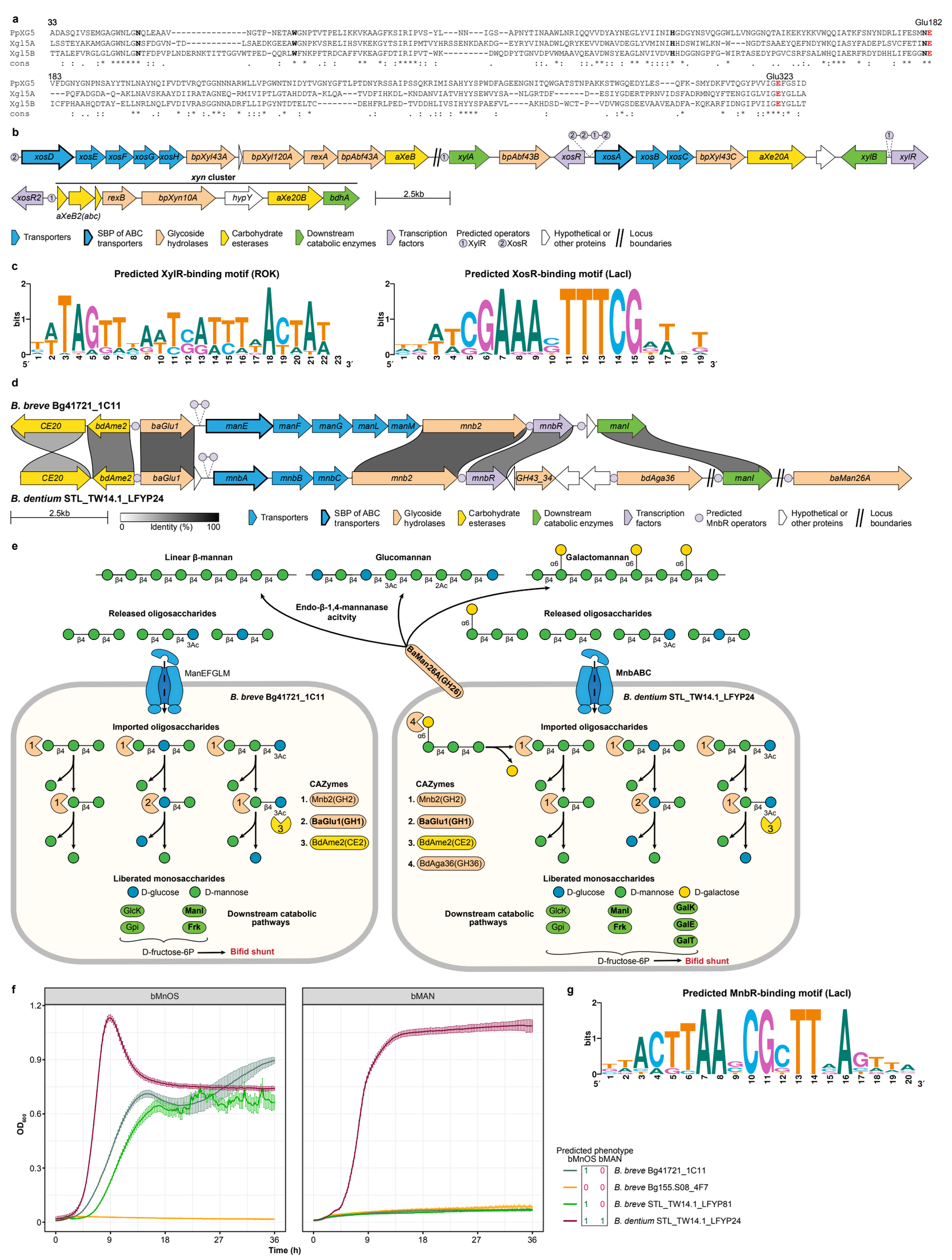


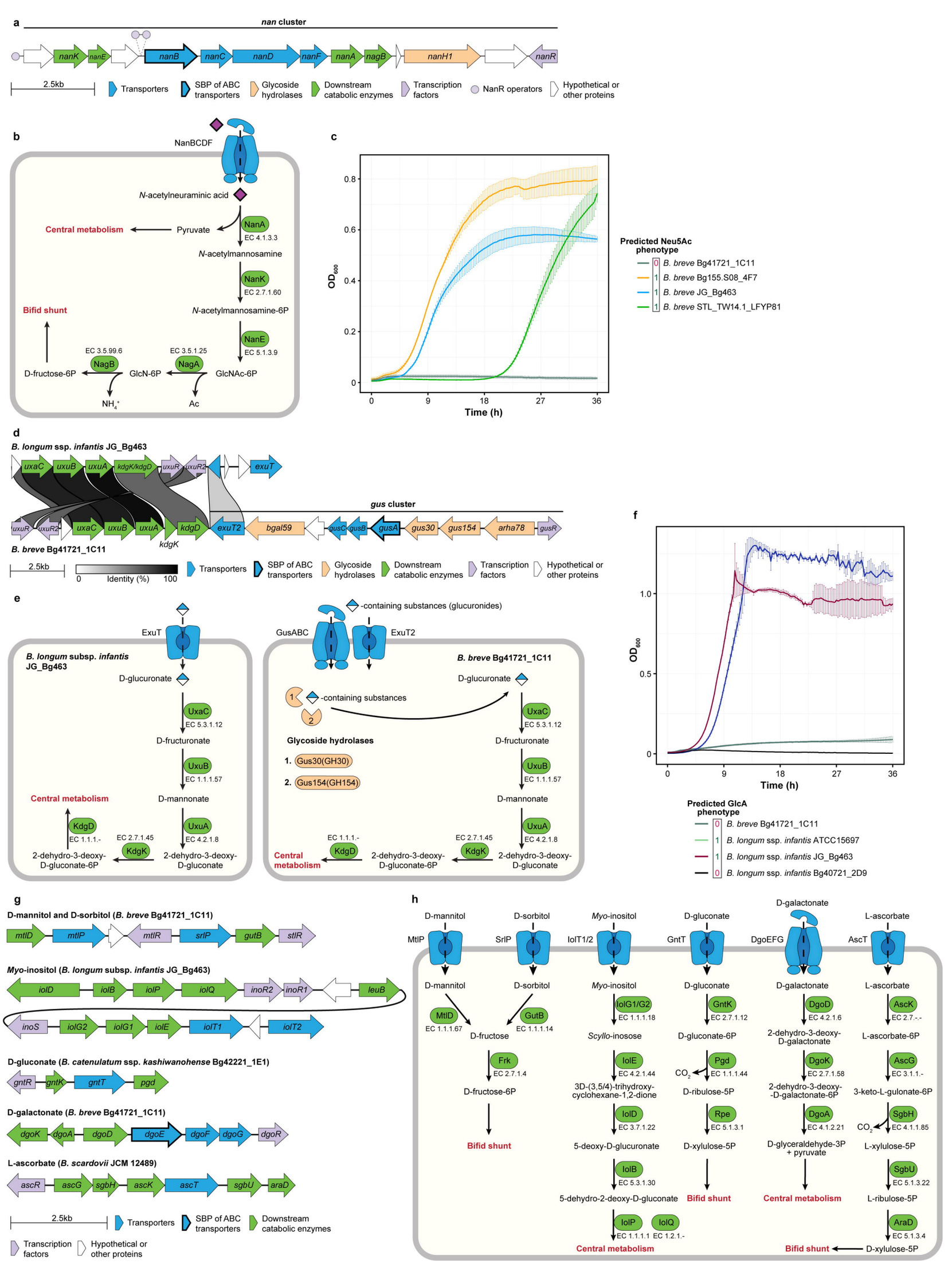
c

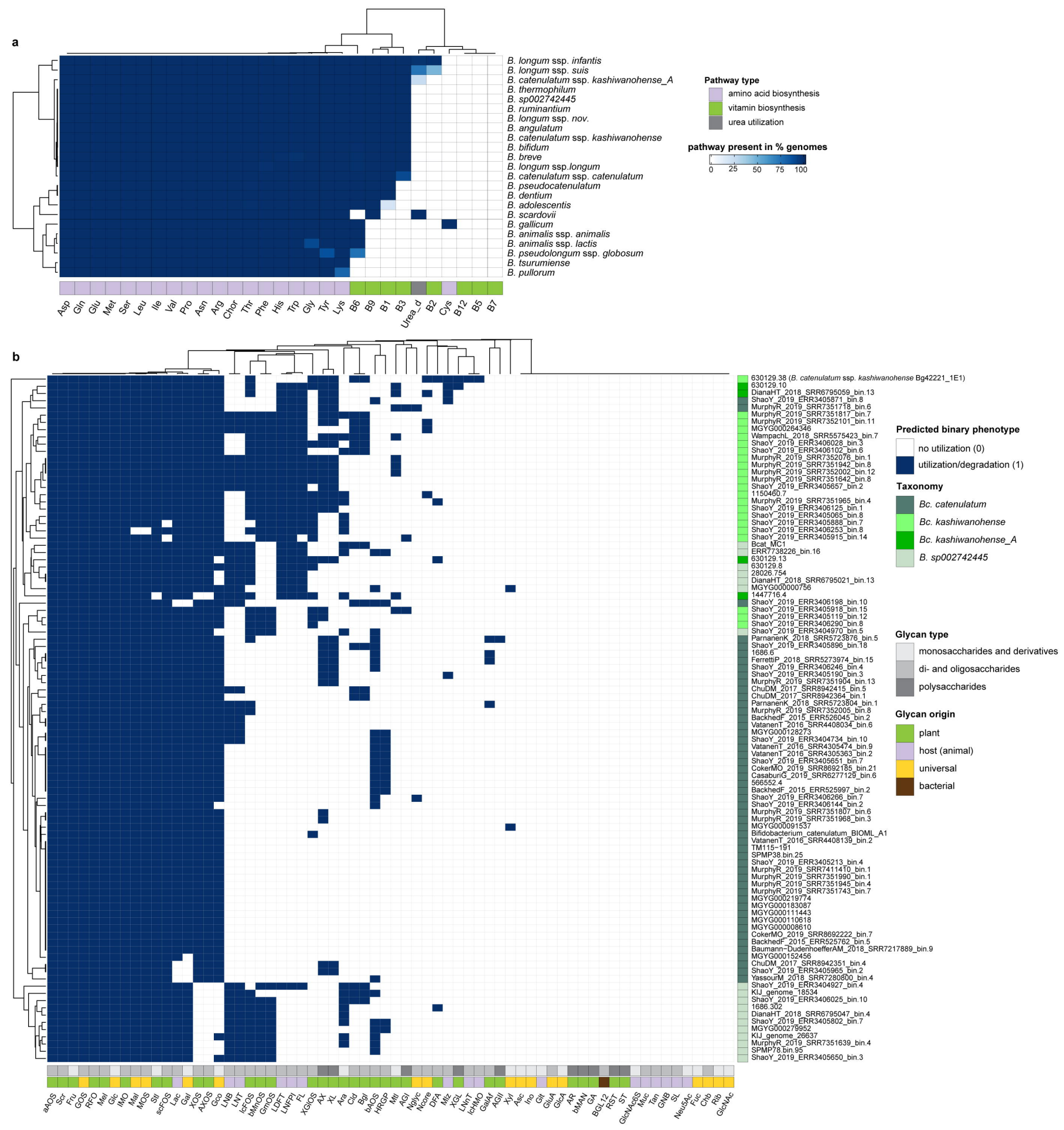


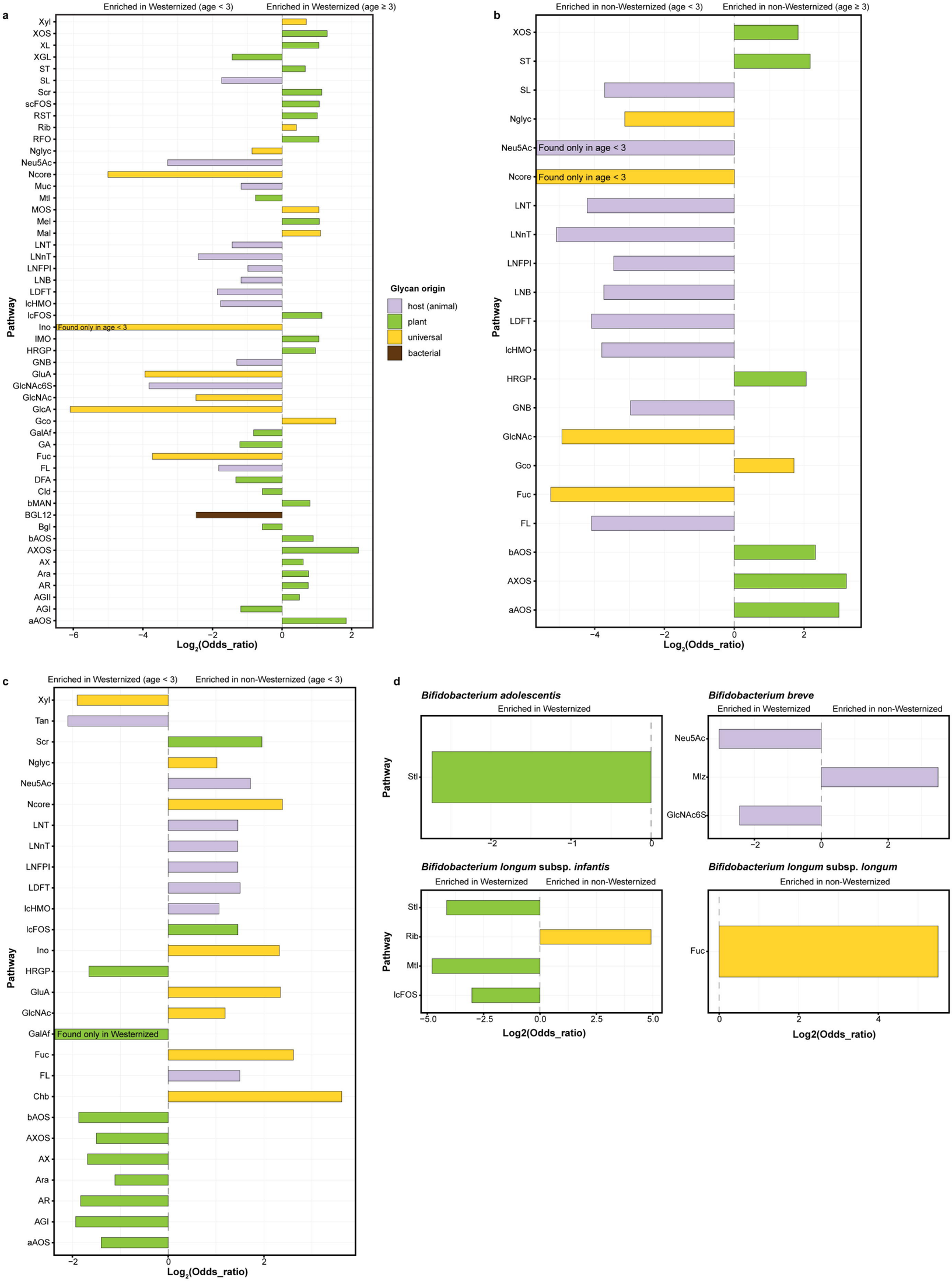

a
ANI
1
0.98
0.96
0.94
BT1 1.000 0.978 0.979 0.958 0.958 0.960 0.957 0.959 0.956 0.957 0.961 0.959 0.960 0.961 0.961 0.959 0.959 0.960 0.960 0.959 0.952 0.953 0.953 0.955 0.954 0.955
ATCC 15697T 0.979 1.000 0.981 0.956 0.954 0.958 0.956 0.958 0.954 0.955 0.959 0.959 0.960 0.960 0.961 0.958 0.959 0.960 0.958 0.959 0.949 0.950 0.951 0.948 0.950 0.952
Bg40721_2D9 0.978 0.981 1.000 0.957 0.956 0.957 0.956 0.958 0.954 0.956 0.959 0.954 0.958 0.958 0.961 0.957 0.957 0.958 0.958 0.957 0.949 0.949 0.950 0.948 0.951 0.951
AGR2137 0.958 0.957 0.957 1.000 0.988 0.988 0.989 0.987 0.984 0.984 0.986 0.985 0.986 0.978 0.979 0.980 0.980 0.981 0.981 0.981 0.959 0.957 0.957 0.964 0.962 0.967
2005B 0.959 0.956 0.958 0.989 1.000 0.988 0.988 0.988 0.983 0.983 0.986 0.983 0.985 0.979 0.979 0.981 0.980 0.980 0.981 0.981 0.959 0.960 0.960 0.964 0.964 0.966
UMA026 0.961 0.960 0.959 0.989 0.989 1.000 0.989 0.989 0.982 0.983 0.986 0.983 0.986 0.980 0.979 0.979 0.979 0.980 0.979 0.981 0.960 0.960 0.960 0.964 0.962 0.966
APC1461 0.957 0.957 0.956 0.990 0.989 0.989 1.000 0.990 0.984 0.984 0.985 0.983 0.986 0.979 0.978 0.978 0.979 0.980 0.980 0.979 0.960 0.959 0.959 0.964 0.964 0.967
209B 0.959 0.960 0.959 0.987 0.988 0.988 0.989 1.000 0.984 0.984 0.985 0.984 0.986 0.980 0.980 0.979 0.979 0.980 0.980 0.980 0.962 0.961 0.961 0.965 0.965 0.967
DSM 20211T 0.957 0.957 0.955 0.984 0.981 0.982 0.984 0.984 1.000 0.998 0.988 0.987 0.988 0.979 0.980 0.980 0.980 0.981 0.981 0.981 0.960 0.961 0.960 0.965 0.963 0.965
UCD391 0.958 0.958 0.957 0.984 0.982 0.982 0.984 0.984 0.998 1.000 0.988 0.988 0.988 0.979 0.980 0.981 0.980 0.982 0.981 0.981 0.961 0.962 0.960 0.966 0.964 0.966
UCD398 0.960 0.959 0.960 0.985 0.985 0.985 0.984 0.986 0.987 0.987 1.000 0.991 0.991 0.980 0.982 0.981 0.981 0.981 0.981 0.982 0.963 0.962 0.962 0.966 0.964 0.967
2074B 0.959 0.959 0.956 0.985 0.982 0.982 0.982 0.984 0.986 0.986 0.990 1.000 0.992 0.980 0.980 0.980 0.979 0.980 0.980 0.980 0.960 0.960 0.960 0.965 0.964 0.966
JCM 19995T 0.959 0.961 0.959 0.985 0.983 0.984 0.984 0.985 0.987 0.987 0.991 0.992 1.000 0.980 0.981 0.982 0.981 0.982 0.982 0.982 0.961 0.961 0.961 0.965 0.964 0.966
JDM301 0.962 0.960 0.958 0.978 0.978 0.977 0.977 0.979 0.978 0.978 0.980 0.979 0.980 1.000 0.980 0.982 0.981 0.982 0.981 0.982 0.962 0.960 0.961 0.965 0.965 0.967
Bg131.S11_17.F6 0.960 0.959 0.961 0.978 0.979 0.977 0.977 0.979 0.979 0.979 0.982 0.978 0.981 0.980 1.000 0.984 0.983 0.983 0.984 0.985 0.959 0.959 0.959 0.962 0.963 0.963
BSM11−5 0.958 0.958 0.957 0.980 0.980 0.979 0.978 0.980 0.980 0.981 0.982 0.980 0.983 0.982 0.985 1.000 0.987 0.987 0.987 0.988 0.962 0.958 0.960 0.963 0.962 0.963
Bg41121_2E1 0.960 0.959 0.959 0.980 0.980 0.979 0.978 0.980 0.980 0.980 0.982 0.981 0.983 0.982 0.985 0.988 1.000 0.987 0.988 0.988 0.960 0.958 0.961 0.963 0.964 0.965
UCD399 0.960 0.960 0.957 0.981 0.981 0.979 0.979 0.980 0.981 0.981 0.981 0.981 0.982 0.983 0.983 0.986 0.987 1.000 0.988 0.988 0.961 0.958 0.962 0.962 0.963 0.965
Bg155.S08_5B11 0.960 0.958 0.959 0.981 0.981 0.979 0.979 0.980 0.980 0.980 0.982 0.982 0.983 0.982 0.984 0.987 0.987 0.988 1.000 0.989 0.961 0.959 0.960 0.963 0.962 0.963
BgEED06 0.959 0.958 0.956 0.982 0.981 0.980 0.978 0.980 0.981 0.981 0.982 0.979 0.983 0.984 0.986 0.987 0.987 0.988 0.989 1.000 0.961 0.958 0.960 0.962 0.961 0.964
1897B 0.952 0.951 0.949 0.958 0.959 0.959 0.958 0.960 0.958 0.959 0.960 0.959 0.960 0.962 0.959 0.960 0.959 0.961 0.958 0.960 1.000 0.992 0.992 0.968 0.969 0.968
STL_TW14.1_LFYP82 0.951 0.950 0.949 0.958 0.959 0.959 0.958 0.960 0.959 0.960 0.961 0.958 0.960 0.961 0.959 0.959 0.958 0.958 0.956 0.958 0.992 1.000 0.994 0.969 0.969 0.968
239-2 0.951 0.951 0.949 0.957 0.958 0.958 0.957 0.959 0.958 0.959 0.960 0.959 0.960 0.961 0.958 0.959 0.958 0.960 0.957 0.958 0.991 0.993 1.000 0.968 0.968 0.967
SC596 0.955 0.949 0.950 0.963 0.963 0.964 0.963 0.965 0.963 0.964 0.965 0.965 0.965 0.965 0.964 0.964 0.962 0.963 0.963 0.962 0.969 0.969 0.969 1.000 0.986 0.986
JCM 1217T 0.954 0.950 0.951 0.961 0.962 0.960 0.962 0.964 0.962 0.963 0.963 0.964 0.964 0.965 0.962 0.962 0.962 0.963 0.961 0.961 0.969 0.969 0.970 0.986 1.000 0.987
NCC2705 0.954 0.951 0.952 0.964 0.964 0.964 0.964 0.966 0.964 0.965 0.966 0.966 0.966 0.966 0.963 0.963 0.963 0.964 0.963 0.962 0.968 0.967 0.968 0.984 0.987 1.000
BT1 ATCC 15697T Bg40721_2D9 AGR2137 2005B UMA026 APC1461 209B DSM 20211T UCD391 UCD398 2074B JCM 19995T JDM301 Bg131.S11_17.F6 BSM11−5 Bg41121_2E1 UCD399 Bg155.S08_5B11 BgEED06 1897B STL_TW14.1_LFYP82 239-2 SC596 JCM 1217T NCC2705
B. longum ssp. infantis
B. longum ssp. suis
B. longum ssp. suis
B. longum ssp. suillum
B. longum ssp. iuvenis (transitional B. longum)
B. longum ssp. nov.
B. longum ssp. longum
b
ANI
1
0.98
0.96
0.94
0.92
JG_Bg468 1.000 0.971 0.970 0.934 0.938 0.935 0.937 0.941 0.940 0.940
SS_Bg39 0.971 1.000 0.986 0.940 0.939 0.945 0.945 0.948 0.946 0.944
Bg41721_1F7 0.972 0.987 1.000 0.941 0.940 0.945 0.943 0.949 0.946 0.945
1899B 0.936 0.941 0.941 1.000 0.983 0.967 0.961 0.948 0.952 0.951
DSM16992 0.939 0.940 0.940 0.982 1.000 0.965 0.960 0.949 0.951 0.950
JCM15439 0.936 0.945 0.945 0.967 0.966 1.000 0.974 0.953 0.956 0.956
Bg42221_1E1 0.938 0.945 0.943 0.961 0.960 0.974 1.000 0.955 0.961 0.965
PV20−2 0.942 0.947 0.948 0.947 0.948 0.953 0.955 1.000 0.962 0.962
Bg41121_2E9 0.941 0.946 0.946 0.951 0.951 0.956 0.962 0.962 1.000 0.980
Bg42221_1D3 0.941 0.944 0.944 0.951 0.951 0.956 0.966 0.963 0.980 1.000
JG_Bg468 SS_Bg39 Bg41721_1F7 1899B DSM16992 JCM15439 Bg42221_1E1 PV20−2 Bg41121_2E9 Bg42221_1D3
Cluster I: B. sp002742445
NCBI: Bifidobacterium catenulatum
GTDB: Bifidobacterium sp002742445
Cluster II: B. catenulatum ssp. catenulatum
NCBI: Bifidobacterium catenulatum ssp. catenulatum
GTDB: Bifidobacterium catenulatum
Cluster III: B. catenulatum ssp. kashiwanohense
NCBI: Bifidobacterium catenulatum ssp. kashiwanohense
GTDB: Bifidobacterium kashiwanohense
Cluster IV: B. catenulatum ssp. kashiwanohense_A
NCBI: Bifidobacterium catenulatum ssp. kashiwanohense
GTDB: Bifidobacterium kashiwanohense_A


a
SC596
JCM 1217
NCC2705
BSM11−5
Bg155.S08_5B11
Bg41121_2E1
BgEED06
JDM301
UCD399
2005B
209B
UMA026
AGR2137
APC1461
UCD398
DSM20211
UCD391
2074B
JCM19995
Bg40721_2D9
ATCC 15697 = JCM 1222
BT1
Bg131.S11_17.F6
1897B
239−2
STL_TW14.1_LFYP82
scFOS Scr GOS Lac RFO Mel IMO MOS Mal Glc Gal LNB GNB LNT Nglyc AGII Ara Tan XOS IcFOS Fru Xyl AXOS Rib aAOS AR bAOS HRGP AX Mlz AGI Cid Bgl Neu5Ac Fuc LNFPI FL LDFT Chb GlcNAc IcHMO Ncore LNnT SL XGlOS ST DFA Mtl Stl GlcA GluA Asc Ino Glt Gco XGL XL bMAN GmOS bMnOS GA GalAr BGL12 RST Muc GlcNAc6S
Taxonomy used in this work
B. longum ssp. infantis
B. longum ssp. longum
B. longum ssp. nov.
B. longum ssp. suis
Taxonomy used in Modesto et al.
B. longum ssp. iuvenis
B. longum spp.
B. longum ssp. suillum
B. longum ssp. suis
Predicted binary phenotype
no utilization (0)
utilization/degradation (1)
Glycan type
monosaccharides and derivatives
di- and oligosaccharides
polysaccharides
Glycan origin
plant
host (animal)
universal
bacterial
b
UCD399
BSM11−5
BgEED06
Bg41121_2E1
Bg155.S08_5B11
BT1
ATCC 15697 = JCM 1222
Bg40721_2D9
UCD391
DSM20211
APC1461
2005B
AGR2137
Bg131.S11_17.F6
JDM301
UMA026
UCD398
JCM19995
209B
2074B
STL_TW14.1_LFYP82
239−2
1897B
SC596
JCM 1217
NCC2705
Asn Asp Gln Glu Gly Pro Met Thr Ser Leu Val Ile Chor Phe Tyr Trp His Lys Arg B9 B6 B1 B3 Urea_d B2 Cys B12 B5 B7
Taxonomy used in this work
B. longum ssp. infantis
B. longum ssp. longum
B. longum ssp. nov.
B. longum ssp. suis-like
Taxonomy used in Modesto et al.
B. longum ssp. iuvenis
B. longum spp.
B. longum ssp. suillum
B. longum ssp. suis
Pathway
pathway absent or incomplete
pathway present
Pathway type
amino acid biosynthesis
vitamin biosynthesis
urea utilization


a
33
Glu182
PpXG5 ADASQIVSEMGAGWNLGNQLEAAV----------NGTP-NETAWGNPTVTPELIKKVKAAGFKSIRIPVS-YL---NN---IGS--APNYTINAAWLNRIQQVVDYAYNEGLYVIINIHGDGYNSVQGGWLLVNGGNQTAIKEKYKKVWQQIATKFSNYNDRLIFESMNE
Xgl5A LSSTEYAKAMGAGWNLGNSFDGVN-TD----------LSAEDKGEEAWGNPKVSRELIHSVKEKGYTSIRIPMTVYHRSSENKDAKDG--EYRYVINADWLQRYKEVVDWAVDEGLYVMINIHHDSWIWLKN-W---NGDTSAAEYQEFNDYWKQIASYFADEPLSVCFETINE
Xgl5B TTALEFVRGLGLGWNLGNTFDPVPLNDERNKTITTGGVWTPED-QQRLWFNKPFTRDCAFFVKEAGFDTIRIPVT-WK---------DWQDPYTGNVDPVWMAAVQEAVDWSYEAGLNVIVDVHHDGGNGPFG-WIRTASEDYPGVCSRFSALWHQIAERFRDYDHHLIFEGGNE
cons : : . :* ****** :: . . : * * .: : ** *: :****:: : . ::. *: ::.**:: : ** *::::* *. . * . .. ..:. *:*** * : : ** **
183
Glu323
PpXG5 VFDGNYGNPNSAYYTNLNAYNQIFVDTVRQTGGNNNARWLLVPGWNTNIDYTVGNYGFTLPTDNYRSSAIPSSQKRIMISAHYYSPWDFAGEENGNITQWGATSTNPAKKSTWGQEDYLES---QFK-SMYDKFVTQGYPVVIGEEFGSID
Xgl5A ---PQFADGDA-Q-AKLNAVSKAAYDIIRATAGNEQ-RMIVIPTLGTAHDDTA-KLQA---TVDFIHKDL-KNDANVIATVHYYSEWVYSA--NLGRTDF------D--ESIYGDERTPRNVIDSFADRMNQYFTENGIGLVIGEYGLLA
Xgl5B ICFPHAAHQDTAY-ELLNRLNQLFVDIVRASGGNNADRFLLIPGYNTDTELTC--------DEHFRLPED-TVDDHLIVSIHYYSPAEFVL----AKHDSD-WCT-P--VDVWGSDEEVAAVEADFA-KQKARFIDNGIPVIIGEYGLLT
cons : .. :: ** .: * :* :.**: * :::* .* : * .: . : .:: : **** : : . :*.: .* * :* ::***:* :
b
xosD xosE xosF xosG xosH bpXyl43A bpXyl120A rexA bpAbf43A aXeB xylA bpAbf43B xosR xosA xosB xosC bpXyl43C aXe20A xylB xylR
xyn cluster
xosR2 aXeB2(abc) rexB bpXyn10A hypY aXe20B bdhA
2.5kb
Transporters
SBP of ABC transporters
Glycoside hydrolases
Carbohydrate esterases
Downstream catabolic enzymes
Transcription factors
Predicted operators ①XylR ②XosR
Hypothetical or other proteins
Locus boundaries
c
Predicted XylR-binding motif (ROK)
Predicted XosR-binding motif (LacI)
bits
5′
3′
d
B. breve Bg41721_1C11
CE20 bdAme2 baGlu1 manE manF manG manL manM mnb2 mnbR manI
B. dentium STL_TW14.1_LFYP24
CE20 bdAme2 baGlu1 mnbA mnbB mnbC mnb2 mnbR GH43_34 bdAga36 manI baMan26A
2.5kb
0 Identity (%) 100
Transporters
SBP of ABC transporters
Glycoside hydrolases
Carbohydrate esterases
Downstream catabolic enzymes
Transcription factors
Predicted MnbR operators
Hypothetical or other proteins
Locus boundaries
e
Linear β-mannan
Glucomannan
Galactomannan
Endo-β-1,4-mannanase acitvity
Released oligosaccharides
Released oligosaccharides
BaMan26A(GH26)
ManEFGLM
MnbABC
B. breve Bg41721_1C11
B. dentium STL_TW14.1_LFYP24
Imported oligosaccharides
Imported oligosaccharides
CAZymes
1. Mnb2(GH2)
2. BaGlu1(GH1)
3. BdAme2(CE2)
CAZymes
1. Mnb2(GH2)
2. BaGlu1(GH1)
3. BdAme2(CE2)
4. BdAga36(GH36)
Liberated monosaccharides
D-glucose D-mannose
GlcK Gpi ManI Frk
Downstream catabolic pathways
D-fructose-6P Bifid shunt
Liberated monosaccharides
D-glucose D-mannose D-galactose
GlcK Gpi ManI Frk GalK GalE GalT
Downstream catabolic pathways
D-fructose-6P Bifid shunt
f
bMnOS
bMAN
OD600
Time (h)
g
Predicted MnbR-binding motif (LacI)
bits
5′
3′
Predicted phenotype
bMnOS bMAN
1 0 B. breve Bg41721_1C11
0 0 B. breve Bg155.S08_4F7
1 0 B. breve STL_TW14.1_LFYP81
1 1 B. dentium STL_TW14.1_LFYP24


a
nan cluster
nanK
nanE
nanB
nanC
nanD
nanF
nanA
nagB
nanH1
nanR
2.5kb
Transporters
SBP of ABC transporters
Glycoside hydrolases
Downstream catabolic enzymes
Transcription factors
NanR operators
Hypothetical or other proteins
b
NanBCDF
N-acetylneuraminic acid
Central metabolism
Pyruvate
NanA
EC 4.1.3.3
N-acetylmannosamine
NanK
EC 2.7.1.60
N-acetylmannosamine-6P
Bifid shunt
NanE
EC 5.1.3.9
EC 3.5.99.6
NagB
EC 3.5.1.25
NagA
D-fructose-6P
GlcN-6P
GlcNAc-6P
NH4+
Ac
c
OD600
Time (h)
Predicted Neu5Ac phenotype
0 B. breve Bg41721_1C11
1 B. breve Bg155.S08_4F7
1 B. breve JG_Bg463
1 B. breve STL_TW14.1_LFYP81
d
B. longum ssp. infantis JG_Bg463
uxaC
uxuB
uxuA
kdgK/kdgD
uxuR
uxuR2
exuT
gus cluster
kdgD
exuT2
bgal59
gusC
gusB
gusA
gus30
gus154
arha78
gusR
kdgK
B. breve Bg41721_1C11
2.5kb
0
Identity (%)
100
Transporters
SBP of ABC transporters
Glycoside hydrolases
Downstream catabolic enzymes
Transcription factors
Hypothetical or other proteins
e
ExuT
B. longum subsp. infantis JG_Bg463
D-glucuronate
UxaC
EC 5.3.1.12
D-fructuronate
UxuB
EC 1.1.1.57
D-mannonate
Central metabolism
KdgD
EC 1.1.1.-
EC 2.7.1.45
KdgK
UxuA
EC 4.2.1.8
2-dehydro-3-deoxy-D-gluconate-6P
2-dehydro-3-deoxy-D-gluconate
-containing substances (glucuronides)
GusABC
ExuT2
B. breve Bg41721_1C11
-containing substances
Glycoside hydrolases
1. Gus30(GH30)
2. Gus154(GH154)
f
OD600
Time (h)
Predicted GlcA phenotype
0 B. breve Bg41721_1C11
1 B. longum ssp. infantis ATCC15697
1 B. longum ssp. infantis JG_Bg463
0 B. longum ssp. infantis Bg40721_2D9
g
D-mannitol and D-sorbitol (B. breve Bg41721_1C11)
mtlD
mtlP
mtlR
srlP
gutB
stlR
Myo-inositol (B. longum subsp. infantis JG_Bg463)
iolD
iolB
iolP
iolQ
inoR2
inoR1
leuB
inoS
iolG2
iolG1
iolE
iolT1
iolT2
D-gluconate (B. catenulatum ssp. kashiwanohense Bg42221_1E1)
gntR
gntK
gntT
pgd
D-galactonate (B. breve Bg41721_1C11)
dgoK
dgoA
dgoD
dgoE
dgoF
dgoG
dgoR
L-ascorbate (B. scardovii JCM 12489)
ascR
ascG
sgbH
ascK
ascT
sgbU
araD
2.5kb
Transporters
SBP of ABC transporters
Downstream catabolic enzymes
Transcription factors
Hypothetical or other proteins
h
D-mannitol
D-sorbitol
Myo-inositol
D-gluconate
D-galactonate
L-ascorbate
MtlP
SrlP
IolT1/2
GntT
DgoEFG
AscT
MtlD
EC 1.1.1.67
GutB
EC 1.1.1.14
D-fructose
Frk
EC 2.7.1.4
D-fructose-6P
Bifid shunt
IolG1/G2
EC 1.1.1.18
Scyllo-inosose
IolE
EC 4.2.1.44
3D-(3,5/4)-trihydroxy-cyclohexane-1,2-dione
IolD
EC 3.7.1.22
5-deoxy-D-glucuronate
IolB
EC 5.3.1.30
5-dehydro-2-deoxy-D-gluconate
IolP
IolQ
EC 1.1.1.1
EC 1.2.1.-
Central metabolism
GntK
EC 2.7.1.12
D-gluconate-6P
Pgd
EC 1.1.1.44
CO2
D-ribulose-5P
Rpe
EC 5.1.3.1
D-xylulose-5P
Bifid shunt
DgoD
EC 4.2.1.6
2-dehydro-3-deoxy-D-galactonate
DgoK
EC 2.7.1.58
2-dehydro-3-deoxy-D-galactonate-6P
DgoA
EC 4.1.2.21
D-glyceraldehyde-3P + pyruvate
Central metabolism
AscK
EC 2.7.-.-
L-ascorbate-6P
AscG
EC 3.1.1.-
3-keto-L-gulonate-6P
SgbH
EC 4.1.1.85
CO2
L-xylulose-5P
SgbU
EC 5.1.3.22
L-ribulose-5P
AraD
EC 5.1.3.4
D-xylulose-5P
Bifid shunt


a
B. longum ssp. infantis
B. longum ssp. suis
B. catenulatum ssp. kashiwanohense_A
B. thermophilum
B. sp002742445
B. ruminantium
B. longum ssp. nov.
B. angulatum
B. catenulatum ssp. kashiwanohense
B. bifidum
B. breve
B. longum ssp.longum
B. catenulatum ssp. catenulatum
B. pseudocatenulatum
B. dentium
B. adolescentis
B. scardovii
B. gallicum
B. animalis ssp. animalis
B. animalis ssp. lactis
B. pseudolongum ssp. globosum
B. tsurumiense
B. pullorum
Asp Gln Glu Met Ser Leu Ile Val Pro Asn Arg Chor Thr Phe His Trp Gly Tyr Lys B6 B9 B1 B3 Urea_d B2 Cys B12 B5 B7
Pathway type
amino acid biosynthesis
vitamin biosynthesis
urea utilization
pathway present in % genomes
0 25 50 75 100
b
630129.38 (B. catenulatum ssp. kashiwanohense Bg42221_1E1)
Predicted binary phenotype
no utilization (0)
utilization/degradation (1)
Taxonomy
Bc. catenulatum
Bc. kashiwanohense
Bc. kashiwanohense_A
B. sp002742445
Glycan type
monosaccharides and derivatives
di- and oligosaccharides
polysaccharides
Glycan origin
plant
host (animal)
universal
bacterial
aAOS Scr Fru GOS RFO Mel Glc IMO Mal MOS Sil scFOS Lac Gal XOS AXOS Gco LNB LNT lcFOS bMnOS GmOS LDFT LNFPI FL XGlcOS AX XL Ara Cid Bgl bAOS HRGP Mtl AGl Nglyc Ncore DFA Mlz XGL LNnT lcHMO GalAt AGlt Xyl Asc Ino Glt GluA GlcA AR bMAN GA BGL12 RST ST GlcNAc6S Muc Tan GNB SL Neu5Ac Fuc Crib Rib GlcNAc


a
Enriched in Westernized (age < 3)
Enriched in Westernized (age ≥ 3)
Pathway
Xyl
XOS
XL
XGL
ST
SL
Scr
scFOS
RST
Rib
RFO
Nglyc
Neu5Ac
Ncore
Muc
Mtl
MOS
Mel
Mal
LNT
LNnT
LNFPI
LNB
LDFT
lcHMO
lcFOS
Ino
Found only in age < 3
IMO
HRGP
GNB
GluA
GlcNAc6S
GlcNAc
GlcA
Gco
GalAf
GA
Fuc
FL
DFA
Cld
bMAN
BGL12
Bgl
bAOS
AXOS
AX
Ara
AR
AGII
AGI
aAOS
−6
−4
−2
0
2
Log2(Odds_ratio)
Glycan origin
host (animal)
plant
universal
bacterial
b
Enriched in non-Westernized (age < 3)
Enriched in non-Westernized (age ≥ 3)
Pathway
XOS
ST
SL
Nglyc
Neu5Ac
Found only in age < 3
Ncore
Found only in age < 3
LNT
LNnT
LNFPI
LNB
LDFT
lcHMO
HRGP
GNB
GlcNAc
Gco
Fuc
FL
bAOS
AXOS
aAOS
−4
−2
0
2
Log2(Odds_ratio)
c
Enriched in Westernized (age < 3)
Enriched in non-Westernized (age < 3)
Pathway
Xyl
Tan
Scr
Nglyc
Neu5Ac
Ncore
LNT
LNnT
LNFPI
LDFT
lcHMO
lcFOS
Ino
HRGP
GluA
GlcNAc
GalAf
Found only in Westernized
Fuc
FL
Chb
bAOS
AXOS
AX
Ara
AR
AGI
aAOS
−2
0
2
Log2(Odds_ratio)
d
Bifidobacterium adolescentis
Enriched in Westernized
Pathway
Stl
−2
−1
0
Bifidobacterium breve
Enriched in Westernized
Enriched in non-Westernized
Neu5Ac
Mlz
GlcNAc6S
−2
0
2
Bifidobacterium longum subsp. infantis
Enriched in Westernized
Enriched in non-Westernized
Pathway
Stl
Rib
Mtl
lcFOS
−5.0
−2.5
0.0
2.5
5.0
Log2(Odds_ratio)
Bifidobacterium longum subsp. longum
Enriched in non-Westernized
Fuc
0
2
4
Log2(Odds_ratio)