Phenotypic and genomic diversification in complex carbohydrate degrading human gut bacteria

Nicholas A. Pudlo1*, Karthik Urs1*, Ryan Crawford2, Ali Pirani1, Todd Atherly3,4, Roberto Jimenez5, Nicolas Terrapon6,7, Bernard Henrissat6,7,8, Daniel Peterson5,9, Cherie Ziemer3,4, Evan Snitkin1 and Eric C. Martens1

1Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, MI 48109
2Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI 48109
3Iowa State University, Dept. of Animal Science, Ames, IA
4United States Department of Agriculture Agricultural Research Station, Ames, IA
5University of Nebraska, Department of Food Sciences, Lincoln, NE
6Aix Marseille Univ, CNRS, UMR7257 AFMB, Marseille, France
7INRAE, USC1408 AFMB, Marseille, France
8Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia
9Johns Hopkins University, Department of Pathology, Baltimore, MD

*Authors contributed equally to this work

Correspondence to: emartens@umich.edu

Running Title: Bacteroidetes carbohydrate utilization
Abstract

Symbiotic bacteria are responsible for the majority of complex carbohydrate digestion in the human colon. Since the identities and amounts of dietary polysaccharides directly impact the gut microbiota, determining which microorganisms consume specific nutrients is central to defining the relationship between diet and gut microbial ecology. Using a custom phenotyping array, we determined carbohydrate utilization profiles for 354 members of the Bacteroidetes, a dominant saccharolytic phylum. There was wide variation in the numbers and types of substrates degraded by individual bacteria, but phenotype-based clustering grouped members of the same species indicating that each species performs characteristic roles. The ability to utilize dietary polysaccharides and endogenous mucin glycans was negatively correlated, suggesting exclusion between these niches. By analyzing related Bacteroides ovatus/xylanisolvens strains that vary in their ability to utilize mucin glycans, we addressed whether gene clusters that confer this complex, multi-locus trait are being gained or lost in individual strains. Pangenome reconstruction of these strains revealed a remarkably mosaic architecture in which genes involved in polysaccharide metabolism are highly variable and bioinformatics data provide evidence of interspecies gene transfer that might explain this genomic heterogeneity. Global transcriptomic analyses suggest that the ability to utilize mucin has been lost in some lineages of B. ovatus and B. xylanisolvens, which still harbor residual gene clusters that are involved in mucin utilization by strains that still actively express this phenotype. Our data provide insight into the breadth and complexity of carbohydrate metabolism in the microbiome and the underlying genomic events that shape these behaviors.
Introduction

Microbial communities in the distal intestines of humans and other mammals play critical roles in the digestion of dietary polysaccharides (1-3). Unlike proteins, lipids and simple sugars, which can be assimilated in the small intestine, the vast majority of non-starch polysaccharides (fibers) transit undegraded to the distal gut due to a lack of requisite enzymes encoded in the human genome (4). Microbial transformation of dietary fiber polysaccharides into host-absorbable organic and short chain fatty acids is a beneficial process that unlocks otherwise unusable calories from our diet (5), shapes the composition and behavior of the gut microbial community (6-8), provides preferred nutrients directly to the colonic epithelium (9-11) and shapes the development of immune cell populations (12, 13).

The abundance of dietary fiber in the mammalian diet, and the substantial chemical diversity within this class of molecules, provides a prominent selective pressure that drives genome evolution and diversification within symbiotic bacterial populations. The genomes of individual human gut bacteria frequently encode dozens-hundreds more polysaccharide-degrading enzymes than human secrete into the gastrointestinal tract, reflecting gut microbial adaptations to degrade dietary fibers (3, 4). As examples, the genomes of a few well-studied Gram-negative Bacteroides (B. thetaiotaomicron, B. ovatus and B. cellulosylyticus) encode between 250 and over 400 CAZymes that collectively equip them to target nearly all commonly available dietary polysaccharides (14-16). However, none of these three species is by itself capable of degrading all available polysaccharides, a conclusion that was supported by early phenotypic surveys of cultured human gut bacteria that encompassed species from other phyla (17, 18). These findings suggest that individual microbes fill multiple, specific carbohydrate degradation niches and that a diverse community is required to ensure degradation of the entire
repertoire of dietary fibers. Given that hundreds of different microbial species typically coexist in an individual over long time periods (19), it is important to understand how many different polysaccharide metabolism pathways are present within the individual microbial species that compose a community and how these traits are represented across strains and species. If some species possess very similar phenotypic abilities, they may be functional surrogates or compete for similar niches and therefore seldom co-occur.

Members of the Bacteroidetes are often among the most numerous bacteria in the colonic microbiota of people from industrialized countries (19-21). These bacteria are well appreciated for their abilities to degrade a broad range of polysaccharides (16-18, 22, 23) and modify disease states in a bacterial species-specific fashion (24-26). In this study, we empirically measured the abilities of members of 29 different Bacteroidetes species to grow on a custom panel of carbohydrates that span the diversity of plant, animal and microbial polysaccharides. Our results reveal a wide range of metabolic breadth between different species, indicating that some have evolved to be carbohydrate generalists while others have become metabolically specialized to target just one or a few nutrients. Pangenome analysis of several related strains provides insight into the evolutionary events that shape carbohydrate utilization among these important symbionts and reveals a dizzying mosaic of heterogeneity at the level of discrete gene clusters mediating polysaccharide metabolism. Based on analysis of several variable loci, we provide evidence to support a mechanism of lateral gene transfer that may account for this mosaic architecture. Our results provide a glimpse into the metabolic breadth and diversity of an important group of human gut bacteria towards polysaccharide metabolism. Given the large amount of genomic and metagenomic sequence information that has been generated from the human microbiome,
fenotypic studies such as the one presented here represent important next steps in deciphering
the functionality of these organisms in their native gut habitat.

111

112 **Results**

113 Phenotypes are the ultimate measures of biological function. However, large-scale
phenotypic analyses are still uncommon in surveys of the human gut microbiome, which have
instead relied on sequence-based approaches to infer function, often with substantial uncertainty.
This lack of phenotypic information is partly due to a lack of high-density (e.g., strain level)
culture representation for the dominant taxa combined with a lack of defined growth conditions
to measure behavior of these organisms. With the resurgence of gut microbial culturing, both of
these gaps have begun to close (27-30), revealing an urgent need for scalable platforms to define
the actual behavior of these organisms. To address this gap, we assembled a collection of human
and animal gut Bacteroidetes and constructed a custom anaerobic phenotyping platform centered
around carbohydrate metabolism, a key function that symbiotic gut microorganisms contribute to
mammalian digestion (4). This array consists of 45 different carbohydrates (30 polysaccharides
and 15 monosaccharides) that span the repertoire of common sugars and linkages present in
dietary plants and meat, as well as host mucosal secretions and some rare nutrients consumed in
regional populations or as food additives (see Fig. S1 for a summary of polysaccharide
structures).

128 The carbohydrate utilization abilities of 354 different human and animal Bacteroidetes
strains were measured by individually inoculating each into this custom growth array and
automatically monitoring anaerobic growth every 10-20 min for four days (see Materials and
Based on 16S rRNA gene sequence for each strain, this collection encompasses 29 different species based on the requirement that each strain possesses ≥98% 16S rRNA gene identity to a named type strain in a given species (Table S1, note that all but three strains, which were all related to each other and to Bacteroides uniformis, met this criterion). The resulting 31,860 individual growth curves were first inspected manually and then subjected to automated analysis to quantify total growth and growth rate parameters for each substrate (see Materials and Methods). A normalization scheme was employed to compensate for general growth differences in the two different defined medium formulations employed (see Table S1 for a full list of strains assayed and all raw and normalized growth measurements, Fig. S2 for analysis of replicates).

Members of the same species possess similar carbohydrate utilization profiles

Growth results are summarized in Figs. 1, 2 and S3. Whether considered from the perspective of how many species degrade a particular polysaccharide (Fig. 1A), or how many individual polysaccharides are targeted by members of a particular species (Fig. 1B), there was substantial variability in carbohydrate utilization among the organisms surveyed (range, 1-28 polysaccharides degraded per strain; mean, 15.6). Some polysaccharides like soluble starch/glycogen were degraded by a majority of the species tested, yet others like the edible seaweed polysaccharides carrageenan and porphyran were used by just one or two strains.

Given the diversity in observed carbohydrate utilization phenotypes, we wished to address if closely related strains display similar abilities or instead if strains of the same species have diverged from one another. To assist in visualizing the overall trends in carbohydrate utilization across this phylum, we performed unsupervised clustering of the strains based on their carbohydrate utilization profiles. While many species are not deeply represented by multiple
strains, clustering based on a combination of normalized growth and rate measurements largely grouped strains of the same species together (Fig. 2) and, as expected, this was driven mostly by polysaccharide utilization abilities (Fig. S4).

Our data reveal that strains belonging to several individual species possess more similar polysaccharide degrading abilities to each other compared to their more distant relatives, a finding that has importance for interpreting or predicting function based on community sequencing data. As examples, all 56 strains of *B. fragilis* clustered together, reflecting their generally restricted abilities to utilize forms of soluble starch/glycogen, inulin and mucus O-glycans. Likewise, all 36 strains of *B. uniformis*, a species with broader metabolic capacity that includes digestion of plant cell wall hemicelluloses, were also grouped together into a single branch. The inclusivity of these groupings was generally independent of the time period when strains were isolated or whether they were isolated from humans or other mammals (Fig. 2).

Another important feature of the observed species clustering is that the grouping does not mirror the overall phylogeny of the gut Bacteroidetes. Rather, phylogenetically separated species often group adjacent to one another based on similarities in carbohydrate metabolism (e.g., *B. ovatus/xylanisolvens* and *B. cellulosilicus*; and *B. vulgatus/dorei* and *B. fragilis*; see Fig. 3A for a phylogenetic tree based on conserved housekeeping genes) (31, 32). In the latter case, it is interesting to directly compare *B. fragilis* and *B. vulgatus/dorei*, two groups with deep strain representation (Fig. 2). Despite being phylogenetically more distant, these species possess very similar phenotypic patterns that reflect degradation of soluble starch and similar molecules (glycogen, pullulan), inulin and mucin O-glycans. The major distinguishing feature between these groups is the presence of some, often-weak, pectin utilization among strains of *B. vulgatus/dorei*. 
Some polysaccharides, especially those present in the cell walls of dietary plants, occur in the same physical context and presumably traverse the gut together, potentially exerting selective pressure for bacteria to use them simultaneously. To test for co-occurrence of traits, we performed a pairwise correlation analysis to determine the extent to which any two polysaccharides were co-utilized by the same strain (Fig. S5). The presence of two different soluble starches (potato and maize amylopectin) and two starch-like glycans (glycogen and pullulan) provide an internal control since they are essentially identical in their sugar and linkage chemistry but vary in the proportion and placement of branches as well as polymer length, crystallinity and solubility (Fig. S1). These four molecules are utilized through a single degradation/transport system in the type strain of *B. thetaiotaomicron*, which was included in our study (33). As expected, the abilities to use these four polysaccharides were among the strongest positive correlations (between 44-75%); although, there was not a perfect correlation suggesting that some finer adaptation may exist even for different structural forms of a chemically similar molecule.

We also observed positive correlations in the ability to use components of two different groups of plant cell wall polysaccharides (pectins and hemicelluloses), as well as animal tissue glycosaminoglycans, despite the fact that the polysaccharides within each of these groups often possess different chemical structures (Fig. S1). In the case of the hemicelluloses, there was even some apparent separation based on dicotyledonous vs. monocotyledonous sources. The predominantly dicot hemicelluloses (Fig. 2, blue labels) and monocot hemicelluloses (Fig. 2, green labels) show some exclusivity with respect to the bacteria that utilize them. Many *B. ovatus/B. xylanisolvens* strains lack the ability to utilize the three dicot hemicelluloses (GalM, GlcM, XyG); whereas the ability to degrade those from monocots (OSX, WAX, BBG) is more
evenly distributed. *B. uniformis* has a partially opposite pattern, preferring substrates from dicots, while only degrading one of the two major monocot structures (BBG) and poorly degrading the two xylans tested (OSX, WAX). Similar observations were also made for pectins and GAGs and could reflect adaptations to simultaneously harvest different nutrients from digesta particles derived from dicot plant cell walls or animal tissue ingested in a carnivorous diet.

**Specialization for mucus O-linked glycans**

The most noteworthy correlation between polysaccharide utilization traits was observed between utilization of host-produced mucin O-glycans and many of the other polysaccharides tested. Growth on a total of 19/30 polysaccharides showed negative correlations with the ability to utilize O-glycans, with the strongest negative correlations being between O-glycans and the seven different hemicelluloses (Fig. S5). This negative correlation is easily observed by comparing the rightmost column in Fig. 2 (O-glycan utilization) with the respective columns for hemicellulose degradation. Because this trend was observed across several species, it suggests that there could be a more general exclusive relationship between the two niches associated with foraging on mucus and hemicellulose. This idea is further supported by experiments described below, which suggest that isolates of *B. ovatus* and *B. xylanisolvens*, both adept hemicellulose consumers, are in the process of losing the ability to degrade O-glycans, relative to an ancestor that contained multiple gene clusters involved in the metabolism of these structures.

Interestingly, the mucin O-glycan mixture was the only substrate for which we observed absolute metabolic specialization among the substrates tested. A single, and only available strain of *Barnesiella intestinihominis* exhibited the ability to exclusively utilize mucin O-glycans, along with a subset of the sugars that are contained in these structures (Table S1). Three strains of
Bacteroides massiliensis exhibited similar behavior with very strong growth on mucin O-glycans and only weak growth on soluble starches and a few other polysaccharides (Table S1). These three B. massiliensis strains were also restricted in the repertoire of simple sugars they could metabolize with this list being limited to those found in mucin and other host glycans (galactose, N-acetylgalactosamine, N-acetylglucosamine, N-acetylneuraminic acid and L-fucose; weak fructose utilization by one strain was the only exception). Members of these two species are poorly represented in culture collections and remain lightly studied. However, their specific adaptations for host mucin glycans may render them important members of the microbiota, potentially thriving at the interface between the gut lumen and host tissue and relying exclusively on the host to be sustained. The continuous supply of mucin in vivo could explain why some species have become specialized for it as a nutrient, whereas dietary fiber degraders may need to be more generalist since the substrates available to them change with the host’s meals.

Pangenome reconstruction reveals extensive genetic diversification among related Bacteroides

With a view of the carbohydrate utilization traits present in our gut Bacteroidetes collection, we next sought to determine if certain variable traits were being gained or lost within strains of certain species and if available genomes provide insight into the mechanisms driving genomic adaptations to particular nutrients. Connections between polysaccharide utilization phenotypes and the underlying genes involved have been systematically explored for a few Bacteroides species (B. thetaiotaomicron, B. ovatus and B. cellulosilyticus) with partial analyses in others (6, 16, 22, 23, 34-37). These studies have revealed that, in essentially all cases, the ability to degrade a particular polysaccharide is conferred by one or more clusters of co-
expressed genes termed polysaccharide utilization loci (PULs) (38). PULs share defining features such as genes encoding homologs of outer membrane TonB-dependent transporters (SusC-like), surface glycan-binding proteins (SGBPs; or SusD- and SusE/F-like), usually an associated sensor/transcriptional regulator and one or more degradative CAZymes (glycoside hydrolase, GH; polysaccharide lyase, PL; carbohydrate esterase, CE), as well as other enzymes like sulfatases or proteases. Since the presence of one or more cognate PULs is required to utilize a given polysaccharide and these genes typically exhibit large increases in gene expression in response to their growth substrate, we rationalized that we could focus on traits that were variable in closely related strains and locate the associated PULs by transcriptomic analysis to gain insight into the basis of their acquisition or loss.

To test this, we focused on members of two closely related species, *B. ovatus* (*Bo*) and *B. xylanisolvens* (*Bx*), for which there is noticeable inter-strain variation in the ability to use mucin O-glycans (Figs. 2, 3). Investigation of these two species also benefits from substantial culture depth and many strains with available sequences. The O-glycans attached to mucins represent a diverse family of over one hundred different structures (39), albeit with common linkage patterns (Fig. S1). Correspondingly, the ability to utilize these glycans is a complex trait, involving simultaneous expression of at least 6-13 different O-glycan inducible PULs in *B. thetaiotaomicron, B. massiliensis, B. fragilis* and *B. caccae* (6, 22, 35). Quantification of O-glycan growth for individual *Bo* and *Bx* strains was widely variable (Fig. 3B). One hypothesis to explain this variability is that some *Bo* and *Bx* strains have gained the ability to utilize O-glycans relative to an ancestor that lacked this phenotype. If so, the PULs they express during O-glycan degradation might be unique to their genomes and may indicate lateral gene transfer (LGT) as has been the case for acquisition of phenotypes such as porphyran, agarose and λ-carrageenan.
utilization in gut Bacteroides, which are all components of integrative conjugative elements or mobilizable plasmids (31, 40). An alternative hypothesis is that some Bo and Bx strains are in the process of losing this ability from a common ancestor. If so, the genomes of non-degraders may still contain some PULs that are homologous to those present in more proficient O-glycan-degrading strains, but these strains may have lost a key step(s) that has eroded their ability to express this phenotype.

To distinguish these hypotheses, we selected seven strains (black arrows in Fig. 3B) that vary in their ability to degrade O-glycans and for which genome sequences exist. Note that three strains that degrade O-glycans were initially chosen because they were among the strongest degraders in our dataset with sequenced genomes when we initiated these experiments. We later identified strains with better O-glycan growth abilities and address one of these (strain H59) separately below. Four of the selected strains were Bo (two positive and two negative for O-glycan degradation); three strains were Bx (one weakly positive and two negative for O-glycan degradation). One of these strains (B. xylanisolvens XB1A) has a finished circular genome and was used as a scaffold to align the remaining six draft genome sequences, with manual curation (see Material and Methods), resulting in a nearly contiguous pangenome sequence that captures the spatial arrangement of homologous and variable genes that are present in these seven strains (Table S2, Fig. S6).

Analysis of the Bo/Bx pangenome revealed remarkable variability in gene content among just the seven strains used. A total of 12,960 different genes were delineated based on ≥90% identity in their translated amino acid sequence (Table S2). Remarkably, only 2,264 (17.5%) of these genes were shared among all seven strains. The largest proportion of genes (7,244; 55.9%) was only present in one of the seven strains. Separating two major classes of core PUL functions,
SusC/D homologs and degradative CAZymes (GH, PL and CE), revealed that these key components of Bacteroidetes polysaccharide metabolism were also heavily represented in the “accessory gene” pool that is not common to all strains (Fig. 4A).

Through informatics-based and manual annotation of gene clusters containing typical PUL functions, we delineated between 180-236 different PULs in the reconstructed pangenome (ambiguity is caused by many PULs occurring adjacent to each other; although in many cases separation of adjacent PULs according to individual genomes allowed us to make more precise delineations, Table S3). Direct comparison of the O-glycan-degrading and non-degrading strains revealed that there was a substantial number of genes (3,351) that were unique to the three O-glycan degrading strains, including genes belonging to 51 PULs (Fig. 4B). However, such a distribution in gene content might be expected given the overall large proportion of non-core genes in these seven strains and there was correspondingly no indication that all three O-glycan-degrading strains shared overlapping PULs with each other: no PULs were common to all three O-glycan degraders and only five PULs were shared by any two strains (Fig. 4C). Considering that there are 51 total PULs that are unique to the mucin-degrading strains, if these strains have gained the ability to degrade O-glycans from an ancestral lineage that lacked this ability it likely occurred by acquisition of separate gene clusters. To more directly distinguish between the two hypotheses given above, we performed transcriptional profiling on all three O-glycan degrading strains to determine if the PUL genes that they express during O-glycan degradation are indeed unique to these strains.

Compared to reference growth in minimal medium containing glucose (MM-glucose), the Bx D22, Bo 3-1-23 and Bo D2 strains activated expression of 196, 227 and 359 total genes more than 10-fold and these gene lists included components of 14, 19, and 42 different PULs,
respectively (Tables S4-6). As expected from studies in other *Bacteroides*, these PULs were scattered throughout the genome (Fig. S7), suggesting that they are autonomously regulated in response to glycan cues present in the *O*-glycan mixture. Strikingly, the majority of PULs that contained *O*-glycan-activated genes (63/75, 84%) were not unique to the *O*-glycan degrading strains (Tables S4-S6, Fig. S7). Moreover, in each of the three strains analyzed, the most highly upregulated PULs were also often shared with non-mucin degrading strains. These observations lend support to the hypothesis that strains of *Bo* and *Bx* are in the process of losing the ability to utilize *O*-glycans relative to a common ancestor that possessed a more expansive gene repertoire to successfully access these nutrients. However, we cannot rule out that individual non-degrading strains are separately acquiring PULs that are associated with mucin degradation and retaining them without the full benefit that presumably occurs with the ability to fully execute this growth phenotype. This latter idea is consistent with inter-species PUL exchange observations elaborated on below.

Finally, because we subsequently identified a *B. ovatus* strain (NLAE-zl-H59, red arrow in Fig. 3B.) with a substantially higher ability to use *O*-glycans relative to the strains used for pangenome construction, we performed additional RNA-seq analysis on this strain. Compared to a glucose reference, this strain activated 373 total genes in response to *O*-glycans, including genes from 30 different PULs (Table S7). Among these, 26 activated PULs were also present in one of the seven strains in our pangenome and 24 were homologous to PULs in strains that did not degrade *O*-glycans. However, this strain did activate expression of genes within four PULs that were completely unique to its genome compared to the seven strains used for pangenome reconstruction, suggesting that it could possess additional genes that augment its ability to grow on mucin *O*-glycans. This increased PUL expression could be responsible for the enhanced
Evidence that intergenomic recombination has driven Bacteroides pangenome evolution

Similar to other bacteria, we observed that many accessory genes in the Bo and Bx pangenome are located in contiguous clusters or “islands” often involving PULs or capsular polysaccharide synthesis gene cluster (Fig. S6, Table S2). In contrast to previously identified Bacteroides PULs that have more obviously been subjects of lateral transfer (31, 40, 41) and are associated with integrative and conjugative elements (ICEs), most of the variable genomic regions that we identified were not associated with functions indicative of mobile DNA. Instead, these regions are often precisely located in between one or more core genes (i.e., those common to all seven strains; herein referred to as “genomic nodes”) that flank each side of the variable gene segment (Fig. 5A,B).

Several intergenomic transfer mechanisms might cause the observed mosaic structure of the Bo-Bx pangenome. The first possibility is movement of variable genes into a recipient genome by direct conjugation of individual, mobile ICEs. While such events would be expected to leave behind residual genes involved in mobilization and transfer, which were not observed, these DNA vehicles are known to target a subset of core genes, such as tRNAs (41), and may have undergone subsequent genomic deletion events that eliminated the mobile DNA. Two other known mechanisms of bacterial LGT are natural competence and phage transduction, neither of which has been observed in members of the Bacteroidetes.

A final potential mechanism is direct conjugation of the chromosome from a donor bacterium into a related recipient, followed by subsequent homologous recombination between
flanking nodes to add or delete intervening DNA in the recipient genome (Fig. 5C). A mechanism that is conceptually similar to high-frequency recombination (Hfr) transfer in *E. coli* has been described for *B. thetaiotaomicron* and *B. fragilis* and involves chromosomal ICEs, which may have lost their ability to transfer autonomously by circularizing from the genome and instead act as transfer initiation points to conjugate a donor genome into a recipient (42-44). If such a mechanism was more broadly active in LGT between *Bacteroides*, we would expect that some of the core/node genes involved would reflect sequence identities that were more similar to the donor and this difference would be more easily detectable if the transfer was between members of different species like *Bo* and *Bx*. Moreover, such transfer events could either result in introduction of new genes into the recipient or elimination of genes depending on the genetic content in between recombination nodes from the donor chromosome.

To test this hypothesis, we took a bioinformatics approach aimed at first identifying high-confidence examples of inter-species recombination involving core genes and then assessed whether those genes were associated with co-transfer of adjacent accessory genes (Fig. 6A). We collected a dataset of 33 *Bo* and *Bx* genomes, which represent a subsample of the isolates for which we generated phenotypic data. We identified a set of 1,384 core genes—expectedly smaller than the core genome of the seven strains used above due to additional strains being added—that are present as a single copy in all members of both species. To identify cases of putative inter-species LGT via homologous recombination at core genes, we searched for instances in which a core gene sequence was more similar to the corresponding sequence from the other species than to sequences of the species to which a strain belonged. To this end, for each allele of each core gene, we calculated the median distance to all other alleles of both species (Fig. 6B). We identified instances where the median distance to the same species was
high and median distance to the opposite species was low and used these genes as markers for putative LGT core loci. To identify additional accessory genes that may have been simultaneously transferred, we searched for instances in which genes were perfectly syntenic and collinear between each genome with a putative LGT core gene and genomes of the opposite species. Among these candidate LGT loci, we then investigated if any of these transfer events have resulted in pan-genome diversification, which we defined as the presence of any accessory gene(s) that was only observed adjacent to a core gene with evidence of LGT based on the above criteria.

In total, we identified 29 different loci at which exchange of core genes appeared to have occurred and LGT accessory genes were identified, including seven that appeared to involve transfer of PULs (Fig. 6C, Fig. S8). Similar numbers of potentially transferred loci were identified for each species, with 16 loci in Bx and 13 loci in Bo. Within the identified HGT events, variable numbers of HGT accessory genes were found within the loci ranging from one to thirteen genes (Fig. 6C, Fig. S8). More genes (57 total) appeared to be transferred into Bo than into Bx (36 total).

Finally, we determined if any of the identified LGT events could explain differential phenotypes measured by our high throughput growth assay by modifying the complement of PULs in individual genomes. As a specific example, we focused on a PUL that was previously associated with β-mannan degradation (23, 45) that was among our candidate loci with evidence of transfer from a Bx ancestor into two Bo strains. The presence of this PUL (PUL-A in Fig. 6C, Fig. S9A) was observed in all strains with the ability to grow on the β-mannan galactomannan (GalM), including two strains of Bo (ATCC8483 and CL02T12C04) for which the flanking node regions were more similar to Bx. We previously showed that deletion of this PUL from B. ovatus
ATCC8483 eliminated growth on GalM and glucomannan (GluM) (45), suggesting that it was both acquired from a Bx strain and conferred growth on these two β-mannans. However, the presence of this PUL was not perfectly correlated with growth on GalM and several strains that lacked PUL-A still exhibited robust growth. Thus, we searched for other PULs that harbor GH26 family enzymes and determined that all of the other strains that grow on GalM, but lack PUL-A, harbor another candidate GalM PUL (PUL-B, Fig. S9A) at a different genomic location and some strains possess both (Fig. 6C). Gene expression analysis by qPCR revealed that PUL-B was highly expressed in strains that lacked PUL-A during growth in GalM (Fig. S9B) and every strain that grew robustly on GalM had at least one of these two PULs. While we had previously shown that PUL-A was required for GluM growth in B. ovatus ATCC8483, there were a number of other strains (red “+” symbols in Fig. 6C) that displayed some ability to grow only on GlcM, while lacking both of the GalM-associated PULs, suggesting the presence of additional, partially orthogonal PULs that confer the ability to grow on variant β-mannans. Such a presence of multiple orthologous PULs that confer the same or similar functions, and some which may be moving between genomes of related species by the putative LGT mechanisms noted above, complicates the process of understanding the genotype-phenotype relationships in human gut Bacteroidetes, but will need to be resolved to make better functional predictions from sequence-based data.

Discussion

In this study we leveraged a scalable, high-throughput quantitative growth platform to characterize the phenotypic abilities that are present in a sample of hundreds of Bacteroidetes strains from the human and animal gut. Our anaerobic screening technique is directly applicable
to other bacterial phyla from the human gut and other environments. Moreover, it can be adapted
to include new polysaccharides or to focus on different nutrient utilization or chemical resistance
phenotypes. The current study, in concert with future applications of phenotypic screening, will
help close the gap between our largely sequence-based view of the human gut microbiota and the
functions that its members provide. However, instances like the ones investigated here for mucin
glycan and β-mannan utilization by Bacteroides serve as a warning that the presence or absence
of genes that are experimentally associated with a particular function do not always indicate that
the phenotype is expressed or not.

Pangenome reconstruction for Bo and Bx revealed extensive variability between strains of
these closely related species, which is not unexpected for bacteria that engage in LGT. However,
the lack of mobile DNA signatures for the majority of accessory genes and evidence of inter-
genomic recombination between species at core genes provides new insight into what may be a
prominent mechanism of genome diversification in members of this phylum. The previously
described intergenomic transfer mechanisms in B. thetaiotaomicron and B. fragilis required the
presence of active or inactive ICEs, highlighting the potential roles for these mobile elements in
not just shaping genomes directly but also indirectly through their ability to catalyze exchange of
broader genomic segments. In B. thetaiotaomicron, genome transfer was determined to initiate at
genomically-integrated ICEs of which there are four in the type strain of B. thetaiotaomicron
(VPI-5482). These have not been shown to be fully functional for circularization and
mobilization. However, introduction and activation of an additional, excision-proficient
conjugative transposon (either cTnDOT or cTnERL) (42), which share common features with the
genomic ICEs, catalyzed expression of genes in the genomic ICEs and transfer of parts of the
genome in a manner that requires recA and homologous DNA to be present in the recipient (42).
An additional study in *B. fragilis* showed that conjugation from a strain with multiple genomic ICEs, with one or more presumably retaining transfer activity, results in transfer of up to 435Kb of chromosome into a recipient that initiates near genomic ICEs, with individual transfer events being of variable size. The latter observation suggests that intergenomic recombination could then occurs at different homologous regions (*i.e.*, the core gene nodes observed in the pangenome), which could depend on the amount of genomic DNA transferred and the length/homology of available recombination sites. Given that the number of ICEs in individual genomes is variable, and their ability to be activated by functional conjugative transposons that are circulating in the ecosystem may also vary, it will be interesting to determine in future work if there are hotspots for genome transfer or if certain strains/species are dominant genome donors that could play a disproportionate role.

The phenotypic similarity between members of the same species (*e.g.*, *Bo* and *Bx*) and the large amount of gene diversity, including genes involved in carbohydrate metabolism, presents a paradox and raises the question of why the genome diversification observed in strains of *Bo* and *Bx* has not pushed members of these species to behave more differently and cluster based on phenotype with members of other species. One answer may be the apparent exclusion of some traits, such as mucin O-glycan/hemicellulose metabolism, which may limit the fitness advantage associated with acquiring new phenotypes. A second emerges from the proposed genome-exchange mechanism for which we offer new experimental support. Since this intergenomic exchange relies on homologous recombination, its frequency should decrease between genomes that are more divergent. Thus, this strategy may be one mechanism through which only closely related bacteria can share traits that are advantageous with other close relatives. The presence of orthologous PULs that confer the same function (*e.g.*, GluM and GalM utilization), some of
which appear to be subjected to LGT, further complicates interpretations of genotype-to-
phenotype relationships in these bacteria. Notably, the genome transfer mechanism proposed
here does not account for how new genes can be incorporated between conserved nodes. Rather,
this variability must pre-exist among different strains and therefore be created by different inter-
and intragenomic diversification mechanisms. Nevertheless, the data that we report here
underscore the notion that individual gut symbiont genomes are not just highly variable, but also
dynamically so.

Materials and methods

Bacterial strains and growth conditions

A total of 354 human and animal gut Bacteroidetes were included in this study. A
complete list is provided in Table S1, along with species designation based on 16S rRNA gene
sequencing and associated meta-data. Dr. Abigail Salyers (University of Illinois, Urbana-
Champagne) kindly provided many of the strains and two large portions of this collection were
isolated over several decades: 99 strains with “WH” designations were collected from fecal
samples of healthy human volunteers as part of the Woods Hole Summer Course on Microbial
Diversity in the late 1990s; 95 additional strains with “VPI” designations were collected from
human samples at the Virginia Polytechnic Institute in the 1960s-1970s. Species classifications
were made based on alignment of a minimum of 734 bp of 16S rRNA gene sequence to a
database containing the type strains of >29 named human gut Bacteroidetes species using the
classify.seqs command with Bayesian settings in the program mothur (46); assignment for each
strain was also manually checked by Blast (47). Isolates with ≥98% 16 rDNA gene sequence
identity to the type strain of a named species were labeled with that species designation. This
classification strategy included all except for three of the 354 strains examined, which ranged
between 96.6 to 96.7% sequence identity to the *B. uniformis* ATCC type strains and based on
sequential isolate numbers might be clones from the same individual (see WH15, WH16, WH17
entries in Table S1). Because of the small number of strains that did not satisfy our 98% cutoff,
we grouped these unclassified strains with their nearest relative and label them as more divergent
in Table S1; although, in most cases the carbohydrate phenotypes of these strains were very
similar to other members of the *B. uniformis* group.

All strains were routinely grown in an anaerobic chamber (Coy Lab Products, Grass
Lake, MI) at 37°C under an atmosphere of 5% H₂, 5% CO₂, and 90% N₂ on brain-heart infusion
(BHI, Beckton Dickinson) agar that included 10% defibrinated horse blood (Colorado Serum
Co.) and gentamicin (200 µg/ml). A single colony was picked into either tryptone-yeast extract-
glucose (TYG) media (48) or modified chopped-meat carbohydrate broth (Table S8) and then
sub-cultured into a minimal medium (MM) formulation that contained a mixture of
monosaccharides, vitamins, nucleotides, amino acids and trace minerals (Table S2 provides
components and a complete recipe).

Carbohydrate growth array setup and data collection

Two different minimal medium formulations were used in the carbohydrate growth
arrays (Table S1 lists the formulation used for each isolate). The simpler of the two formulations
(medium 1) was identical to the above MM, except that no carbohydrates were included and the
medium was prepared at 2X concentration. The second minimal medium formulation (medium
2) was identical to medium 1, but included beef extract (0.5% w/v final concentration) as an
additional supplement. We initially attempted to cultivate all of the species tested using only
medium 1, but determined that beef extract was specifically required to allow growth of some
species, especially *Parabacteroides* spp., *Barnesiella intestinohominis*, *Odoribacter splanchnicus*
and the branch of *Bacteroides* that includes *B. plebeius* and *B. massiliensis*. Growth in the
absence of an added carbohydrate source was generally not observed or very low, except with
*Parabacteroides* that may be able to grow to a low level on the added 0.5% beef extract. The
corresponding negative control wells for each strain assayed were averaged and this value
subtracted from the total growth calculation of the corresponding to strain on other carbohydrates
tested. Despite several attempts to supplement minimal media with different components or
employ more stringent anaerobic methods, we were unable to cultivate several common
Bacteroidetes genera/species (*Prevotella* spp., *Paraprevotella* spp., *Alistipes* spp., and
*Bacteroides coprocola* and *Bacteroides coprophilus*) in these two MM formulations and therefore did not include them in this study. All of these isolates readily grew in rich medium,
suggesting that they have specific nutritional requirements that were not met in the MM
formulations used.

Carbohydrate growth arrays were run as described previously (23) using a list of 45
carbohydrates (*Table S9*) that were present in duplicate, non-adjacent wells of a 96-well plate;
two additional wells contained no carbohydrate and served as negative controls. Each MM was
prepared as a 2X concentrated stock without carbohydrates (MM-no carb). An aliquot of each
strain was taken from a MM-monosaccharides culture (grown for 16-20 h) and was centrifuged
to pellet cells. Bacteria were resuspended in the same volume of 2X MM-no carb and then
centrifuged again prior to suspension in a volume of 2X MM-no carb that was equal to the
original volume. These washed bacterial cells were then inoculated at a 1:50 ratio into 2X MM-
no carb and the suspension was added in equal volume (100μl/well) to the 96 wells of the
carbohydrate growth array. Each well of the carbohydrate growth array contained 100μl of 2X carbohydrate stock (10-20mg/ml); thus, when diluted 2-fold resulted in 1X MM containing a unique carbohydrate and a bacterial inoculum that was identical to other wells. Growth arrays were monitored at kinetic intervals of 10-20 minutes using a microplate stacking device and coupled absorbance reader (Biotek Instruments; Winooski, VT) and data recorded for 4 d (variable kinetic interval times reflect variations in the number of microtiter plates present in a given batch).

**Carbohydrate growth array data processing**

Growth data were processed according to the following workflow: 1. data for each strain were exported from Gen5 software (Biotek Instruments; Winooski, VT) into Microsoft Excel and a previously described automated script was employed to call the points at which growth began (min) and ended (max) (23); 2. Each file was manually checked to validate that appropriate calls were made and the min and max values edited if needed (generally, only due to obvious baselining artifacts or erroneously high calls caused by temporary bubbles or precipitation); 3. “total growth” \( (A_{600} \text{ max} - A_{600} \text{ min}) \) and “growth rate” \( [(A_{600} \text{ max} - A_{600} \text{ min}) / (t \text{ max} - t \text{ min})] \) were calculated for each strain on each substrate \( (A_{600} \) is the absorbance value at 600 nm that corresponds to each min and max point; \( t \) is the corresponding time values in minutes); 4. Individual cultures in which total growth was \( \leq 0.1 \) were scored as “no growth” and their \( A_{600} \) values converted to 0. Only assays in which both replicates showed an increase in \( A_{600} \geq 0.1 \) were considered as growth; if the two replicate assays were discordant (one positive, one negative), then both values were converted to zero.
To normalize the results for each strain, the substrate(s) that provided maximum total growth and growth rate values were determined and these were set to 1.0. All other growth values for a given strain were normalized to this maximum value, providing a range of values between 0 and 1.0. We next normalized growth ability across individual substrates using the previously normalized values for each individual strain: the strain with the maximum total growth and growth rate values were identified (many of these were already set to 1.0). Then, the corresponding values for each other species on that particular substrate were calculated as a fraction of the maximum value for that substrate, yielding a range of values between 0 and 1.0 for each substrate. These values were used to create the heat map shown in Figs. 2 and S3 and all raw and normalized values are provided in Table S1.

Data clustering and statistics

Heatmaps and corresponding dendrograms were generated using the “heatmap” function in the “stats” package of R (version 3.4.0) which employs unsupervised hierarchical clustering (complete linkage method) to group similar carbohydrate growth profiles. Pearson Correlation was used to calculate co-occurrence of the ability to grow on each pair of different substrates. The normalized growth value for each substrate was compared to the corresponding growth values on all other substrates using the Pearson correlation test in R and these values are displayed in the Pearson correlation plot in Figure S5.

Pangenome reconstruction for B. ovatus and B. xylanisolvens strains

Since one of the seven strains used for pangenome reconstruction (B. xylanisolvens XB1A) was assembled into a single circular chromosome, we used this genome as a scaffold for
the contigs representing the remaining six strains. Contigs from the six unfinished strains were aligned against the XB1A genome using a combination of Mauve (50), to align and orient larger contigs, and reciprocal best Blast-hit analysis using \( \geq 90\% \) amino acid identity to identify likely homologs, to provide finer resolution. Contigs from draft genome assemblies or Bx XB1A were broken as needed to accommodate the inclusion of unique accessory genes, but only in circumstances where genes on both sides of the break could be aligned to homologs in one or more genomes with a contig that spanned that break point. After constructing a preliminary assembly, we analyzed the size distribution of putative homologous ORFs as a measure of assembly accuracy and to identify variations in genetic organization that might be attributable to real genetic differences such as frame shifts, which would result in two homologous gene calls of smaller size in the genome containing the frameshift. Any variation \( > 50\% \) of homologous ORF size was inspected manually using the “orthologous neighborhood viewer, by best Blast hit” function in the U. S. Dept. of Energy Integrated Microbial Genomes (IMG) website. Introduced contig breaks are documented in Table S2 and Fig. S6. GenVision software (DNAstar, Madison, WI) was used to visualize and label selected functions in the pan-genome assembly and also display RNAseq data as a function of shared and unique PULs.

**RNAseq analysis**

For RNAseq, *B. xylanisolvens* and *B. ovatus* cells were grown to mid-exponential phase on either purified mucin O-linked glycans (purified in house from Sigma Type III porcine gastric mucin) or glucose as a reference as previously described (22). Total RNA was extracted using an RNeasy kit (Qiagen), treated with Turbo DNase I (Ambion), and mRNA was enriched using the Bacterial Ribo-Zero rRNA removal kit (Epicentre). Residual mRNA was converted to
sequencing libraries using TruSeq barcoded adaptors (Illumina) and sequenced at the University of Michigan Sequencing Core in an Illumina HiSeq instrument with 24 samples multiplexed per lane. Bar-coded data were demultiplexed and analyzed using the Arraystar software package with Qseq (DNAstar). All RNAseq data are publicly available from the National Institutes of Health Gene Expression Omnibus Database under accession numbers GSM4714867-GSM4714890.

Core gene determination and detection of LGT events between Bo and Bx strains

The core gene alignment was generated with cognac (51). The alignment was then partitioned into the individual component genes and approximate maximum likelihood gene trees were generated with fastTree (52). Co-phylogenetic distances were calculated with APE (53). A distance threshold of greater than 0.1 to the same species and less than 0.1 to the opposite species was used to identify alleles bearing signatures of HGT. All analyses were performed in R (version 3.6.3) (54). All code developed for this project are available at https://github.com/rdcrawford/bacteroides_hgt.

Acknowledgements

This work is dedicated to the memory of Dr. Cherie Ziemer. We thank Thomas Schmidt (University of Michigan) for helpful advice on developing the phenotype clustering score. We thank Abigail Salyers (University of Illinois, Urbana-Champaign) who kindly provided a large portion of the strains used in this work. Additional support with strain culture and resources was provided by Nadja Shoemaker (University of Illinois, Urbana-Champaign), Emma Allen-Vercoe (University of Guelph), Laurie Comstock (Harvard University), Jin-Woo Bae (Kyung-Hee University, Korea), Tomomi Kuwahara (Kagawa University, Japan) and Jeffrey Gordon.
This work was supported by funds from the U.S. National Institutes of Health (DK084214, DK118024, DK125445), the University of Michigan Biological Sciences Scholars Program and pilot/feasibility grants from the University of Michigan Center for Gastrointestinal Research (NIDDK DK034933).

**Competing Interests:** The authors declare no competing interests.

**Figure legends**

**Figure 1. Glycan degradation abilities among gut Bacteroidetes.** (A.) The number of species out of 29 tested that degrade each polysaccharide is listed in order of decreasing degradation frequency from left to right. Since not all strains within a given species necessarily have the metabolic potential to utilize each polysaccharide, colors illustrate the percentage of strains within each degrading species that possess the indicated ability. (B.) The number of polysaccharides that a given species degrades in decreasing order. The number of strains tested for each species is listed in parentheses and colors represent the percentage of strains in each indicated species that degrade each glycan counted towards the total.

**Figure 2. Heat map of individual polysaccharide utilization traits.** Species are clustered by glycan utilization phenotype based on normalized total growth level (**Fig. S4B**). The magnitude of growth is indicated by the heatmap scale at the bottom right. Columns at the left indicate the source (human, animal) and time period of isolation. The cladogram at the far left shows the results of unsupervised clustering of the data based the normalized growth data shown. The
species designations at the right are the results of 16S rRNA gene sequencing (>98% identity to the species type strain was used to assign species). All raw and normalized growth and rate data for individual strains may be found in Table S1 see Fig. S3 for an expanded heatmap with monosaccharide data and individual strain names labels.

Figure 3. Host mucin O-glycan metabolism within the Bacteroides. (A.) A phylogenetic tree based on housekeeping genes that compares mucin O-glycan utilization across species. The diameter of the black circles represents the number of strains tested within each species (sample depth), whereas the size of the overlaid red circle corresponds to the number of strains exhibiting O-glycan metabolism. Note that some species have either full or no penetrance of this phenotypic trait yet others like B. ovatus/B. xylanisolvens have more extensive variability among strains. (B.) Strains of B. ovatus (blue) and B. xylanisolvens (green) that show variable growth abilities on mucin O-glycan (n=2 growth assays per bar, error bars are range between values). Gray histogram bars are total growth controls on an aggregate of the monosaccharides that all strains of these two species grow on (Table S1) and are provided as a reference for overall growth ability on a non O-glycan substrate. Data from two established O-glycan degraders, B. massiliensis and B. thetaiotaomicron, are also shown for reference. Species with black arrows were used for pangenome analyses to compare genetic traits associated with mucin O-glycan metabolism. We performed RNA-seq on three strains included in this pangenome analyses (black boxes) positive for O-glycan utilization and an additional strain, B. ovatus NLAE-zl-H59 (red arrow, box), to see if there were unique genes/PULs present in strains that have the ability to grow on mucin O-glycans.
Figure 4. Distribution of all genes as well as core polysaccharide utilization functions in the Bo/Bx pangenome. (A.) The left panel shows the number of core genes (i.e., those present in all seven strains used for pangenome construction) compared to genes present in 2-7 of the individual strains. The right panel shows the same distribution of genes assigned to PULs or particular degradative CAZyme families (GH, PL, CE, see Tables S2 and S3 for more detailed assignments. (B.) The distribution of genes between mucin-degrading (n=3) and non-degrading (n=4) strains used to construct the pangenome. Top numbers indicate total genes, while numbers in parentheses indicate the number of PULs (not individual PUL genes) in each category. (C.) Distribution of the genes that are unique to the three mucin-degrading strains within each genome. Genes/PULs are numbered as described for B. Note that no PULs are shared by all three strains.

Figure 5. (A.) A higher-resolution view of a region of the Bo/Bx pangenome shows the variable presence of at least six different PULs occurring between three genomic nodes (nodes 33-35 in this quarter of the total pangenome. Segment 2 of the physical pangenome map was selected because the first segment initiated with numerous small contigs and this segment contained previously validated genes for xyloglucan metabolism (49). Node genes are colored red, while susC-like and susD-like genes are colored purple and orange, respectively, and glycoside hydrolase genes in light blue. GH family numbers are given below select PULs starting from the top to indicate potential specificity and new numbers are only added going down the schematic if the family assignments are different, indicating a different PUL. A well-studied B. ovatus PUL for xyloglucan degradation (49) is shown in the center and occurs variably between two nodes and also has variable gene content. The two bottom genomes are from different species, Bacteroides finegoldii (Bfin) and Bacteroides fragilis (Bfra) and show less complex genome
architecture with the Bfra region possessing no PULs. (B.) A broader view of the genome region shown in A. showing that the same mosaic pattern is common across the pangenome. Only PULs are illustrated, although many other genes were also variable in these regions. The numbers at the bottom delineate the presence of 35 different core gene nodes (as in A. some nodes contain multiple core genes) in this section of the genome and the presence of homologous or unique PULs is illustrated according to the color code at right (see Fig. S6 for high resolution physical maps of the pan-genome with PUL annotations). Note that in some cases up to five different PULs were located at one location (C.) A schematic showing the proposed mechanism of genome exchange based on previous studies (42-44) and observations presented here. Genomic ICEs that are either partially active (excision deficient, but capable of initiating DNA strand breakage and conjugation) or activated in trans by the presence of an exogenous conjugative transposon, initiate genome mobilization from a donor into a recipient. If sufficient homology between node genes exists in the recipient, homologous recombination between two nodes can replace a section of the recipient with a segment from the donor. Note that genomic regions are shown as linear fragments for simplicity, but would be circular.

Figure 6. (A.) Schematic of the workflow to identify putative LGT core genes: align genes and build corresponding trees for each core gene, determine the median substitution distances for each allele of a core gene in a given strain to both species, and identify loci with an identical conserved structure between isolates of opposite species. (B.) Plot of median distances for all core genes identified in the 33 genomes analyzed. The boxes show the regions containing genes for which the median distance was ≥ 0.1 to the assigned species for a given strain and ≤ 0.1 for the opposite species to which a strain is assigned. These genes were determined to be high-confidence examples of core/node genes that had been replaced by an
allele from the other species. (C.) A region of the Bo/Bx pan-genome that contains a PUL involved in galactomannan (GalM) and glucomannan (GluM) degradation. This PUL is present in six strains of Bx and two strains of Bo and in the latter cases flanking node genes exhibit signatures of being derived from LGT with a Bx donor (the yellow box highlights potential recombination region). The columns at the left indicate the growth of each strain on GalM or GluM. The ability to grow on GalM is fully correlated with the presence of one of two different PULs, or both, that are transcriptionally activated during growth on this substrate (Fig. S9) (23). Notably, some strains (red “+”) are able to grow weakly on GluM but do not possess either of the identified PULs, suggesting that additional, partially orthologous PULs exist that confer the ability to use only GluM.

Supplementary Figure and Table Legends

Figure S1. Schematics of the polysaccharides used in this study with sugar composition and linkages schematized according to the “Symbol nomenclature for glycans” standard format and based on the symbol key provided at the right. Linkages are labeled as α or β and the number provided represents the carbon position in the recipient sugar. The carbon in the donor sugar is carbon-1 in all cases except N-acetyl neuraminic acid and is not shown. Note that pectic galactan (potato and lupin), xylan (oat spelt and wheat arabinoxylan) and amylopectin (potato and maize) can have variable structures based on plant source. Abbreviations for several polysaccharides are provided in parentheses and used throughout the text and figures.
**Figure S2.** Correlation of replicate growth and rate measurements. Two replicate measurements were made for each of the two parameters recorded, total growth (**A.**) and growth rate (**B.**) for each species on each carbohydrate substrate. Data points are color-coded based on whether the two replicates exhibited variation between 0-5% (black), 5-10% (blue), 10-20% (green), >20% (orange) or growth in one assay and no growth in the other (red). (**C.**) A linear function was fitted (with red points omitted) to calculate an $r^2$ value for the data set associated with utilization of each individual substrate. Measurements on some substrates were more variable than on others due, at least in part, to the tendency of these substrates to partially precipitate or retrograde during growth, which yielded variable levels of background absorbance.

**Figure S3.** A heatmap identical to the one shown in **Fig. 2** main text, except that monosaccharide growth data is included. Strain names are also noted at the far right (best viewed in electronic PDF form with magnification) and animal strains are labeled in red font.

**Figure S4.** (**A.**) A scheme for evaluating which aspects of growth phenotype data are most influential for clustering strains that belong to the same species using hypothetical B. theta data as an illustrative example. A quantitative index was used in which the number of strains tested is divided by the minimum number of branches needed to encompass all of the strains for that species, with a perfect score being “1” (*e.g.*, eight B. theta strains divided by the minimum of eight branches needed to encompass all strains in the top example). (**B.**) Actual clustering index data for the raw and normalized growth and rate data gathered for 354 different Bacteroidetes strains. M and P stand for “monosaccharide” and “polysaccharide” growth, respectively. One of the two most optimal conditions, which incorporates normalized growth data on polysaccharides only, was used to construct **Figs. 2** and **S3**.
Figure S5. A Pearson correlation plot to determine if individual growth abilities co-occur in the same strains. Positive or negative correlations that are $\geq 0.40$ are shown in the colors indicated.

Figure S6. High-resolution maps of the entire reconstructed pangenome. These maps are provided in four separate parts due to their large size (labeled as Fig. S6a, b, c, d) and correspond to the data table provided in Table S2. Note that a 5th file is provided with information about the gene, locus and strand breaking legend data.

Figure S7. Circular pangenome and corresponding mucin O-glycan transcriptomics from Bx D22, Bo 3_1_23 and Bo D2.

Figure S8. Individual maps of high-confidence inter-genomic exchange events between Bo and Bx strains.

Figure S9. (A.) Schematics of PUL-A and PUL-B associated with GalM and GlcM utilization. In Bo ATCC8384, elimination of PUL-A eliminates both of these growth abilities. (B.). Expression analysis by qPCR of two sentinel genes from PUL-B in Bo strain D2 that lacks PUL-A but still exhibits robust growth on GalM.

Table S1. Strain designations, growth levels, growth rates, host species, isolation periods, growth media, 16S rRNA similarities and, if applicable, public genome sequence references for all Bacteroidetes strains used in this study.

Table S2. Data table of the reconstructed pangenome of seven Bo/Bx strains. Additional notes are provided directly on the table.
Table S3. PULs that were delineated in the seven strain pan-genome with annotations based on whether they were unique to mucin non-degrading strains, unique to mucin-degraders or shared between strains in both categories. Additional notes are provided directly on the table.

Table S4. Gene expression changes detected using whole-genome transcriptional profiling by RNA-seq of *B. xylanisolvens* D22 grown on mucin *O*-glycan as a sole carbon source compared to glucose reference. Additional notes are provided directly on the table.

Table S5. Gene expression changes detected using whole-genome transcriptional profiling by RNA-seq of *B. ovatus* 3-1-23 grown on mucin *O*-glycan as a sole carbon source compared to glucose reference. Additional notes are provided directly on the table.

Table S6. Gene expression changes detected using whole-genome transcriptional profiling by RNA-seq of *B. ovatus* D2 grown on mucin *O*-glycan as a sole carbon source compared to glucose reference. Additional notes are provided directly on the table.

Table S7. Gene expression changes detected using whole-genome transcriptional profiling by RNA-seq of *B. ovatus* NLAE-zl-H59 grown on mucin *O*-glycan as a sole carbon source compared to glucose reference. Additional notes are provided directly on the table.

Table S8. Liquid media recipes (sheet A) and components (sheet B) for growing the Bacteroidetes used in this study.
Table S9. Mono- and polysaccharides used in the phenotypic growth arrays and corresponding supplier or purification details.
References

1. Koropatkin NM, Cameron EA, Martens EC. How glycan metabolism shapes the human gut microbiota. Nat Rev Microbiol. 2012; 10: 323-35.

2. Flint HJ, Scott KP, Duncan SH, Louis P, Forano E. Microbial degradation of complex carbohydrates in the gut. Gut microbes. 2012; 3:289-306.

3. Porter NT, Martens EC. The Critical Roles of Polysaccharides in Gut Microbial Ecology and Physiology. Annual review of microbiology. 2017; 71: 349-69.

4. El Kaoutari A, Armougom F, Gordon JI, Raoult D, Henrissat B. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. Nature reviews Microbiology. 2013; 11: 497-504.

5. McNeil NI. The contribution of the large intestine to energy supplies in man. The American journal of clinical nutrition. 1984; 39: 338-42.

6. Desai MS, Seekatz AM, Koropatkin NM, Kamada N, Hickey CA, Wolter M, et al. A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility. Cell. 2016; 167: 1339-53.

7. Sonnenburg ED, Smits SA, Tikhonov M, Higginbottom SK, Wingreen NS, Sonnenburg JL. Diet-induced extinctions in the gut microbiota compound over generations. Nature. 2016; 529: 212-5.

8. Sonnenburg ED, Sonnenburg JL. Starving our microbial self: the deleterious consequences of a diet deficient in microbiota-accessible carbohydrates. Cell metabolism. 2014; 20: 779-86.

9. Wrzosek L, Miquel S, Noordine ML, Bouet S, Joncquel Chevalier-Curt M, Robert V, et al. Bacteroides thetaiotaomicron and Faecalibacterium prausnitzii influence the production of mucus glycans and the development of goblet cells in the colonic epithelium of a gnotobiotic model rodent. BMC Biol. 2013; 11: 61.

10. Ganapathy V, Thangaraju M, Prasad PD, Martin PM, Singh N. Transporters and receptors for short-chain fatty acids as the molecular link between colonic bacteria and the host. Current opinion in pharmacology. 2013; 13: 869-74.

11. Cook SI, Sellin JH. Review article: short chain fatty acids in health and disease. Aliment Pharmacol Ther. 1998; 12: 499-507.

12. Smith PM, Howitt MR, Panikov N, Michaud M, Gallini CA, Bohlooly YM, et al. The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis. Science. 2013; 341: 569-73.
13. Kim M, Qie Y, Park J, Kim CH. Gut Microbial Metabolites Fuel Host Antibody Responses. *Cell Host Microbe.* 2016; 20: 202-14.

14. Xu J, Mahowald MA, Ley RE, Lozupone CA, Hamady M, Martens EC, et al. Evolution of Symbiotic Bacteria in the Distal Human Intestine. *Plos Biol.* 2007; 5: e156.

15. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic acids research.* 2009; 37: D233-D8.

16. McNulty NP, Wu M, Erickson AR, Pan C, Erickson BK, Martens EC, et al. Effects of diet on resource utilization by a model human gut microbiota containing *Bacteroides cellulosilyticus* WH2, a symbiont with an extensive glycobiome. *Plos Biol.* 2013; 11: e1001637.

17. Salyers AA, Vercellotti JR, West SE, Wilkins TD. Fermentation of mucin and plant polysaccharides by strains of *Bacteroides* from the human colon. *Appl Environ Microbiol.* 1977; 33: 319-22.

18. Salyers AA, West SE, Vercellotti JR, Wilkins TD. Fermentation of mucins and plant polysaccharides by anaerobic bacteria from the human colon. *Appl Environ Microbiol.* 1977; 34: 529-33.

19. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature.* 2010; 464: 59-65.

20. Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, et al. Evolution of mammals and their gut microbes. *Science.* 2008; 320: 1647-51.

21. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, et al. Diversity of the human intestinal microbial flora. *Science.* 2005; 308: 1635-8.

22. Martens EC, Chiang HC, Gordon JI. Mucosal Glycan Foraging Enhances Fitness and Transmission of a Saccharolytic Human Gut Bacterial Symbiont. *Cell Host Microbe.* 2008; 4: 447-57.

23. Martens EC, Lowe EC, Chiang H, Pudlo NA, Wu M, McNulty NP, et al. Recognition and Degradation of Plant Cell Wall Polysaccharides by Two Human Gut Symbionts. *Plos Biol.* 2011; 9: e1001221.

24. Bloom SM, Bijanki VN, Nava GM, Sun L, Malvin NP, Donermeyer DL, et al. Commensal *Bacteroides* species induce colitis in host-genotype-specific fashion in a mouse model of inflammatory bowel disease. *Cell Host Microbe.* 2011; 9: 390-403.
25. Hickey CA, Kuhn KA, Donermeyer DL, Porter NT, Jin C, Cameron EA, et al. Colitogenic Bacteroides thetaiotaomicron Antigens Access Host Immune Cells in a Sulfatase-Dependent Manner via Outer Membrane Vesicles. *Cell Host Microbe.* 2015; 17: 672-80.

26. Mazmanian SK, Round JL, Kasper DL. A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature.* 2008; 453: 620-5.

27. Lagkouvardos I, Pukall R, Abt B, Foesel BU, Meier-Kolthoff JP, Kumar N, et al. The Mouse Intestinal Bacterial Collection (miBC) provides host-specific insight into cultured diversity and functional potential of the gut microbiota. *Nat Microbiol.* 2016; 1: 16203.

28. Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, Stares MD, et al. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature.* 2016; 533: 543-6.

29. Lagier JC, Khelaifia S, Alou MT, Ndongo S, Dione N, Hugon P, et al. Culture of previously uncultured members of the human gut microbiota by culturomics. *Nat Microbiol.* 2016; 1: 16203.

30. Tramontano M, Andrejev S, Pruteanu M, Klunemann M, Kuhn M, Galardini M, et al. Nutritional preferences of human gut bacteria reveal their metabolic idiosyncrasies. *Nat Microbiol.* 2018; 3: 514-22.

31. Hehemann JH, Kelly AG, Pudlo NA, Martens EC, Boraston AB. Bacteria of the human gut microbiome catabolize red seaweed glycans with carbohydrate-active enzyme updates from extrinsic microbes. *Proceedings of the National Academy of Sciences of the United States of America.* 2012; 109: 19786-91.

32. Tamura K, Hemsworth GR, Dejean G, Rogers TE, Pudlo NA, Urs K, et al. Molecular mechanism by which prominent human gut Bacteroidetes utilize mixed-linkage β-glucans, major health-promoting cereal polysaccharides. Cell Reports. 2017; 21: 417-430.

33. Shipman JA, Berleman JE, Salyers AA. Characterization of four outer membrane proteins involved in binding starch to the cell surface of Bacteroides thetaiotaomicron. *J Bacteriol.* 2000; 182: 5365-72.

34. Déjean G, Tamura K, Cabrera A, Jain N, Pudlo N, Holm Viborg A, et al. Synergy between cell-surface glycosidases and glycan-binding proteins dictates the utilization of specific beta(1,3)-glucans by human gut *Bacteroides.* *mBio.* 2020; 11: e00095-20.

35. Pudlo NA, Urs K, Kumar SS, German JB, Mills DA, Martens EC. Symbiotic Human Gut Bacteria with Variable Metabolic Priorities for Host Mucosal Glycans. *mBio.* 2015; 6: e01282-15.
36. Despres J, Forano E, Lepercq P, Comtet-Marre S, Jubelin G, Chambon C, et al. Xylan degradation by the human gut Bacteroides xylanisolvens XB1A(T) involves two distinct gene clusters that are linked at the transcriptional level. *BMC genomics*. 2016; 17: 326.

37. Sonnenburg ED, Zheng H, Joglekar P, Higginbottom SK, Firbank SJ, Bolam DN, et al. Specificity of polysaccharide use in intestinal Bacteroides species determines diet-induced microbiota alterations. *Cell*. 2010; 141: 1241-52.

38. Terrapon N, Lombard V, Drula E, Lapebie P, Al-Masaudi S, Gilbert HJ, et al. PULDB: the expanded database of Polysaccharide Utilization Loci. *Nucleic Acids Res*. 2018; 46:D677-D83.

39. Johansson ME, Holmen Larsson JM, Hansson GC. Microbes and Health Sackler Colloquium: The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host-microbial interactions. *Proceedings of the National Academy of Sciences of the United States of America*. 2011; 108 Suppl 1: 4659-65.

40. Pudlo NA, Vasconcelos-Pereira G, Parnami J, Cid M, Markert S, Tingley JP, et al. Extensive transfer of genes for edible seaweed digestion from marine to human gut bacteria. *Preprint*. 2021; BIORXIV/2020/142968.

41. Martens EC, Kelly AG, Tauzin AS, Brumer H. The devil lies in the details: how variations in polysaccharide fine-structure impact the physiology and evolution of gut microbes. *Journal of molecular biology*. 2014; 426: 3851-65.

42. Moon K, Sonnenburg J, Salyers AA. Unexpected effect of a Bacteroides conjugative transposon, CTnDOT, on chromosomal gene expression in its bacterial host. *Mol Microbiol*. 2007; 64: 1562-71.

43. Husain F, Tang K, Veeranagouda Y, Boente R, Patrick S, Blakely G, et al. Novel large-scale chromosomal transfer in Bacteroides fragilis contributes to its pan-genome and rapid environmental adaptation. *Microb Genom*. 2017; 3:e000136.

44. Whittle G, Hamburger N, Shoemaker NB, Salyers AA. A Bacteroides conjugative transposon, CTnERL, can transfer a portion of itself by conjugation without excising from the chromosome. *J Bacteriol*. 2006; 188: 1169-74.

45. Reddy SK, Bagenholm V, Pudlo NA, Bouraoui H, Koropatkin NM, Martens EC, et al. A beta-mannan utilization locus in Bacteroides ovatus involves a GH36 alpha-galactosidase active on galactomannans. *FEBS Lett*. 2016; 590: 2106-18.

46. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*. 2009; 75: 7537-41.
47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. *J Mol Biol.* 1990; 215: 403-10.

48. Holdeman LV, Cato ED, Moore WEC. Anaerobe Laboratory Manual. Moore WEC, editor. Blacksburg, Va.: Virginia Polytechnic Institute and State University Anaerobe Laboratory; 1977.

49. Larsbrink J, Rogers TE, Hemsworth GR, McKee LS, Tauzin AS, Spadiut O, et al. A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. *Nature.* 2014; 506(7489):498-502.

50. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 2004; 14: 1394-403.

51. Crawford RD, Snitkin ES. cognac: rapid generation of concatenated gene alignments for phylogenetic inference from large, bacterial whole genome sequencing datasets. *BMC Bioinformatics.* 2021; 22: 70.

52. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol.* 2009; 26: 1641-50.

53. Popescu AA, Huber KT, Paradis E. ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics.* 2012; 28: 1536-7.

54. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2017; URL https://www.R-project.org/.
1 - 25% of strains in given species degrade substrate
26 - 50% of strains in given species degrade substrate
51 - 75% of strains in given species degrade substrate
76 - 99% of strains in given species degrade substrate
100% of strains in given species degrade substrate

Figure 1

A. No. of species that degrade polysaccharides

B. No. of polysaccharides that each species degrades

Polysaccharide

Species

B. ovatus (34)
B. thetaiotaomicron (58)
B. xylanisolvens (45)
B. cellulosilyticus (10)
P. distasonis (16)
B. finegoldii (3)
B. uniformis (35)
B. nordii (4)
B. oleiciplenus (1)
B. plebius (1)
B. salyersiae (5)
B. stercoris (7)
B. dorei (14)
B. intestinalis (1)
B. vulgatus (33)
D. gadei (1)
P. johnsonii (1)
P. merdae (3)
B. caccae (12)
B. massiliensis (3)
B. fluxus (1)
B. fragilis (56)
D. mossii (1)
B. clarus (1)
B. eggerthii (3)
P. goldsteinii (2)
P. gordonii (1)
O. splanchnicus (1)
B. intestihominis (1)
Figure 2

Clustering by phenotype

Species

- B. vulgatus/
- B. dorei
- B. nordii
- Parabacteroides sp.
- B. caccae
- B. fragilis
- B. theta iotaomicron/
- B. finegoldii
- B. ovatus/
- B. xylanisolvens
- B. cellulosilyticus
- B. stercoris
- B. eggerthii
- B. salyersiae
- B. uniformis
Figure 3

A.

B. xylanisolvens
B. ovatus
B. finegoldii
B. caccae
B. thetaiotaomicron
B. salyersiae
B. nordii
B. fragilis
B. intestinalis
B. cellulosilyticus
B. oleiciplenus
B. uniformis
B. flux
B. eggerthii
B. stercoris
B. clarus
B. plebeius
B. massiliensis
B. vulgatus
D. gadei
D. mossii
O. splanchnicus
Bar. intestinihominis
Pr. copri
Pa. distasonis
Pa. merdae
Pa. johnsonii
Pa. gordonii
Pa. goldsteinii
Pr. copri

No. of strains tested:
58 29 14 1

No. of mucin O-glycan degrading strains:

B. xylanisolvens
B. ovatus
B. massiliensis
B. thetaiotaomicron
D2 (2_1_39)
3_1_23
3_8_47FAA
ATCC8483
D22 (1_2_8)
D1
XB1A

B. ovatus
B. xylanisolvens

B. massiliensis
B. thetaiotaomicron

"no growth" cutoff

Increase in absorbance (600 nm) on O-glycans

strain tested

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2
Figure 4

A. Total pangenome

No. of genes

Core (all strains)

Accessory (all strains)

CAZymes

SustC/D-like

Genomes with gene present:

1

2

3

4

5

6

B. non mucin O-glycan degraders (n=4) mucin O-glycan degraders (n=3)

4,387 (48)

5,222 (137)

3,351 (51)

C. Bo D2

Bo 3_1_23

Bx D22

1,230 (29)

71 (3)

19 (1)

197 (5)

71 (1)

25 (1)

908 (12)
Figure 5

A. Sequential order of 53 PULs in section 2 of 4 of the B. ovatus-B. xylanisolvens pangenome

B. Common genomic node region: 1 variable PUL region 1 variable PUL region 2 no PUL present PUL variant 1 PUL variant 2 PUL variant 3 PUL variant 4 PUL variant 5 no assembly

C. Donor Bacteroides genome transfer & conjugation from ICE donator genome genomic ICE Recipient Bacteroides homologous recombination within nodes genomic nodes recipient genome Recipient Bacteroides with modified genome genomic nodes recipient genome
1) Identify core genes and create alignments
2) Create individual gene trees
3) Identify gene trees that demonstrate patterns of LGT
4) Identify shared accessory genes that occur at LGT loci

B. xylanisolvens allele
B. ovatus allele

growth Genes
Gal M Glc M PUL- A PUL- B

Figure 6
glycosaminoglycans

fructans

starches

pectin

glycosaminoglycans

hemicellulose

microbial & marine

O-glycans

(representatives of ~10^2 structures)
Figure S2

A. Scatterplot of total growth replicates $R^2 = 0.95$

B. Scatterplot of growth rate replicates $R^2 = 0.86$

**Color codes:**
- ● 0-5% variation
- □ 5-10% variation
- ▲ 10-20% variation
- ■ >20% variation
- ▲ no growth in one replicate

C. **Polysaccharides:**

| Substrate | $R^2$ between growth values | $R^2$ between rate values |
|-----------|-----------------------------|---------------------------|
| AG        | 0.99                        | 0.89                      |
| alg       | 0.96                        | 0.96                      |
| α-mann    | 0.93                        | 0.76                      |
| APm       | 0.94                        | 0.92                      |
| APpo      | 0.96                        | 0.91                      |
| arab      | 0.98                        | 0.96                      |
| BBG       | 0.95                        | 0.70                      |
| carr      | 0.93                        | 0.96                      |
| Cell      | 0.96                        | 0.81                      |
| CS        | 0.96                        | 0.91                      |
| dex       | 0.96                        | 0.87                      |
| GalM      | 0.96                        | 0.98                      |
| GlcM      | 0.93                        | 0.85                      |
| glyc      | 0.96                        | 0.94                      |
| hep       | 0.96                        | 0.85                      |
| hya       | 0.91                        | 0.88                      |
| inulin    | 0.92                        | 0.89                      |
| lam       | 0.96                        | 0.96                      |
| levan     | 0.96                        | 0.88                      |
| lich      | 0.80                        | 0.45                      |
| MOG       | 0.98                        | 0.97                      |
| OSX       | 0.93                        | 0.83                      |
| PGA       | 0.97                        | 0.89                      |
| PGI       | 0.95                        | 0.96                      |
| PGp       | 0.92                        | 0.92                      |
| por       | 0.85                        | 0.85                      |
| pull      | 0.84                        | 0.78                      |
| RGI       | 0.96                        | 0.98                      |
| WAX       | 0.97                        | 0.42                      |
| XyG       | 0.92                        | 0.72                      |

D. **Monosaccharides:**

| Substrate | $R^2$ between growth values |
|-----------|-----------------------------|
| Ara       | 0.90                        |
| Fru       | 0.90                        |
| Fuc       | 0.93                        |
| Gal       | 0.60                        |
| GalA      | 0.86                        |
| GalNAc    | 0.93                        |
| Glc       | 0.69                        |
| GlcA      | 0.87                        |
| GlcNAc    | 0.72                        |
| GlcNH3    | 0.93                        |
| Man       | 0.88                        |
| NeuNAc    | 0.86                        |
| Rha       | 0.96                        |
| Rib       | 0.94                        |
| Xyl       | 0.85                        |
Figure S3

| Strain                      | Species                                      |
|-----------------------------|----------------------------------------------|
| B. vulgatus/                | B. dorei                                     |
| B. nordii                   | Parabacteroides sp.                         |
| B. caccae                   | B. caccae                                    |
| B. fragilis                 | B. fragilis                                  |
| B. thetaiotaomicron/        | B. thetaiotaomicron/                        |
| B. ovatus/                  | B. xylanisolvens                             |
| B. xylanisolvens            |                                             |
| B. cellulosilyticus         |                                             |
| B. stercoris                |                                             |
| B. thetaiotaomicron         |                                             |
| B. ovatus                   |                                             |
| B. salyersiae               |                                             |
| B. uniformis                |                                             |

**Normalized Growth**

- 1: High growth
- 0.75-0.99
- 0.50-0.74
- 0.25-0.49
- 0.01-0.24
- No growth

**Source**

- human
- animal
- pre-1980 VPI
- 1995-99 Woods Hole
- post 2000
- unknown

**Poly saccharide substrate**

- pull
- xyloglucan
- pectin
- GAGs
- starch & fructan

**Monosaccharide substrate**

- GlicNAc
- GlcNAc
- Glic
- GalNAc
- Glic
- Man
- Gal
- Fru
- Rha
- NeuNAc

**Strain**

- B. salyersiae
- B. stercoris
- B. cellulosilyticus
- B. fragilis
- B. thetaiotaomicron
- B. ovatus
- B. caccae
- B. uniformis
Figure S4

A. Example of Cluster scoring scheme

B. Normalized data

C. Unnormalized (raw) or binary (growth/no growth) data

P = polysaccharide data
M = monosaccharide data
| Starches | Fructans | GAGs | Pectins | Hemicelluloses | Microbial & Marine |
|----------|----------|------|---------|---------------|-------------------|
| Pull     | Glyc     | Apo  | Inulin  | Levan         | Hep               |
| 1.00     | 0.60     | 0.43 | 0.51    | 0.10          | 0.11              |
| Glyc     | 0.60     | 1.00 | 0.62    | 0.04          | 0.30              |
| App      | 0.43     | 0.52 | 1.00    | 0.37          | 0.37              |
| Apm      | 0.51     | 0.68 | 0.56    | 0.08          | 0.08              |
| Inulin   | 0.10     | 0.04 | 0.02    | -0.04         | 1.00              |
| Levan    | 0.11     | 0.39 | 0.35    | -0.03         | 0.09              |
| Hep      | 0.14     | 0.47 | 0.30    | 0.09          | 1.00              |
| HyA      | -0.03    | 0.35 | 0.30    | 0.16          | 0.65              |
| CS       | -0.06    | 0.38 | 0.37    | 0.06          | 0.53              |
| PGA      | -0.04    | 0.37 | 0.30    | 0.08          | 0.64              |
| RGI      | 0.10     | 0.41 | 0.37    | 0.39          | 0.50              |
| PGp      | -0.07    | 0.26 | 0.17    | -0.07         | 0.44              |
| PGl      | -0.12    | 0.23 | 0.14    | -0.01         | 0.42              |
| AG       | 0.03     | 0.33 | 0.28    | -0.02         | 0.53              |
| Arab     | 0.07     | 0.25 | 0.27    | -0.23         | 0.37              |
| GalM     | 0.05     | 0.21 | 0.07    | -0.01         | 0.08              |
| Gum      | 0.04     | 0.22 | 0.06    | 0.00          | 0.17              |
| XyG      | 0.04     | 0.23 | 0.03    | -0.02         | 0.14              |
| OSX      | 0.07     | 0.38 | 0.23    | 0.06          | 0.31              |
| WAX      | 0.02     | 0.31 | 0.17    | 0.09          | 0.29              |
| BBG      | 0.10     | 0.32 | 0.28    | 0.02          | 0.20              |
| Cell     | -0.12    | 0.14 | 0.01    | 0.07          | 0.10              |
| Lamm     | -0.14    | 0.04 | -0.03   | 0.00          | 0.12              |
| Lich     | 0.02     | 0.14 | 0.01    | 0.11          | 0.08              |
| DeX      | 0.12     | 0.50 | 0.30    | 0.05          | 0.85              |
| ummann   | 0.08     | 0.32 | 0.28    | 0.29          | 0.03              |
| Alg      | 0.09     | 0.18 | 0.19    | 0.09          | 0.33              |
| Carr     | 0.02     | 0.03 | 0.05    | 0.01          | 0.10              |
| Porph    | -0.12    | -0.09 | -0.07  | -0.08         | 0.07              |
| MOG      | -0.05    | -0.16 | -0.05  | -0.02         | -0.07             |

**Color key:**
- perfect correlation (1.0)
- +0.7 to 1.0
- +0.40 to 0.7
- less than -0.40
Please note that Figure S6 is provided as a zipped folder containing 4 separate quarters of the pan genome assembly along with a corresponding legend that explains the color coding scheme. The four maps correspond to Table S2.

Each map contains 5 vertically stacked panes of pan genome map starting in the upper left. Each horizontal pane has 8 rows with the top row representing the pan genome and the corresponding 7 individual genome regions shown below.

The example below shows a small region of pan genome section 2 in which Bo 3-1-23 is missing an ECF-σ regulated PUL. The small text above genes in individual genomes correspond to the contig and the green dashed line represents a region that was broken to accommodate accessory genes.
Figure S7

- Shared PUL genes between at least one mucin degrader and one non-degrader
- Unique PUL genes in mucin degraders
- Non PUL genes
- Transcriptome
- Genome/Pangenome
Figure S8. Bacteroides LGT Loci

- HGT gene
- Present
- Absent
- Allele more like B. ovatus
- Allele more like B. xylanisolvens
PUL LGT Events

B. ovatus PUL LGT Event 1

- TonB dependent receptor
- fec operon regulator FecR
- ECF RNA polymerase sigma factor SigK
- Lipoprotein-releasing system transmembrane protein LoIE
- Ribosome-binding factor A
- Putative O-methyltransferase/MSMEI_4947
- Pyruvate kinase
- 3-dehydroquinate dehydratase
- Tyrosine reductase YrdD
- Tetraaracine repeat protein
- Thiol-disulfide oxidoreductase ResA
- Competence protein ComM
- Aerobic respiration control sensor protein ArcB
- Hypothetical protein
- SusD family protein
- TonB dependent receptor
- IPT/TIG domain protein
- Alpha-xylulose isomerase
- Extracellular exo-alpha-(1,4)-arabinofuranosidase ArbA precursor
- Hypothetical protein
- Hypothetical protein
- SusD family protein
- TonB dependent receptor
- Hypothetical protein
- Hypothetical protein
- Kojibiose phosphorylase
- Hypothetical protein
- Sensor histidine kinase TodS
### non-PUL LGT Events

**B. ovatus non-PUL LGT Event 1**

| Species | Genes |
|---------|-------|
| Exoenzyme S synthesis regulatory protein ExsA |
| TonB dependent receptor |
| SulS family protein |
| Glycosyl hydrolase family 92 |
| Plant Basic Secretory Protein |
| Glycosyl hydrolase family 92 |
| Retaining sigma-galactosidase precursor |
| Sensory transduction protein LysR |
| Sensor histidine kinase YehU |
| Multidrug resistance protein MdrN |
| Cobalt-zinc-cadmium resistance protein CzoA |
| Hypothetical protein Glutathione peroxidase homolog BsaA |
| Isopentenyl-diphosphate Delta-isomerase |
| S-adenosylmethionine:RNA ribosyltransferase-isomerase |
| Arabinosyl operon regulatory protein |
| 3-dehydroquinate synthase |
| Hypothetical protein |
| Major carotenoid synthase CisA |
| Ribosomal RNA small subunit methyltransferase D |
| Hypothetical protein |
| ATP-dependent RecD-like DNA helicase |
| Hypothetical protein Bacteroidetes-Associated Carbohydrate-binding Often N-terminal |
| Endo-beta-N-acetylglucosaminidase F1 precursor |
| Hypothetical protein |
| SulG and Rags outer membrane lipoprotein |
| TonB-dependent Receptor Plug Domain protein |
| Hypothetical protein |
| Endo-beta-N-acetylglucosaminidase F1 precursor |
| P5 beta type C domain protein |
| Glycosyl hydrolase family 92 |
| Glycosyl hydrolase family 92 |
| Ec regulator regulator FeoR |
| ECF RNA polymerase sigma factor SigW |
| Glycosyl hydrolase family 92 |
| Murein DD-endopeptidase MepM |
| HTH-type transcriptional repressor YcgE |
| Guanosine-3’-5’-bis(diphosphate) 3’-pyrophosphohydrolase |
| Membrane-bound lytic murein transglycosylase U precursor |
| Hypothetical protein |
| Guanine chromosome-partitioning protein ParB |
| Sporulation initiation inhibitor protein Sjo |
| Hypothetical protein |
| 5’-nucleotidase SurE |
| Lpd-A disaccharide synthase |
| NgD-like protein |
| Phosphohosphate cytidylyltransferase |
| ATP-dependent zinc metalloprotease PssH |
| Ribosomal silencing factor RatS |
| Hypothetical protein |
| Magnesium transporter MglE |
| Ribosomal RNA small subunit methyltransferase A |
| Hypothetical protein |
| Cytochrome non-specific dipeptidase |
### B. xylanisolvens non-PUL LGT Event 7

| species | hypothetical protein |
|---------|----------------------|
|         | TonB-dependent Receptor Plug Domain protein |
|         | fec operon regulator FecR |
|         | hypothetical protein |
|         | ECF RNA polymerase sigma factor SigE |
|         | Oxygen-independent coproporphyrinogen-III oxidase 1 |
|         | Elongation factor G |
|         | Alkaline phosphatase synthesis sensor protein PhoR |
|         | hypothetical protein |
|         | Organic hydroperoxide resistance transcriptional regulator |
|         | 30S ribosomal protein S6 |
|         | 30S ribosomal protein S18 |
|         | 50S ribosomal protein L9 |
|         | hypothetical protein |
|         | Inositol 2-dehydrogenase |
|         | Inositol 2-dehydrogenase |
|         | hypothetical protein |
|         | Inosine dehydratase |
|         | hypothetical protein |
|         | hypothetical protein |
|         | Long-chain-fatty-acid–CoA ligase FadD15 |
|         | Threonylcarbamoyladenine tRNA methyltransferase MtaB |
|         | Lipid A biosynthesis lauril acyltransferase |
|         | N-acetylglucosaminyl-diphospho-decaprenol L-rhamnosyltransferase Methylyglyoxal synthase |
|         | Dihydroneopterin aldolase |
|         | Acryltransferase family protein |
|         | 4-alpha-glucanotransferase |
|         | Ribonucleoside-diphosphate reductase NrdZ |
|         | FMN reductase [NAD(P)H] |
|         | Chromosomal replication initiator protein DnaA |
|         | mce related protein |
|         | N-acetylmuramoyl-L-alanine amidase AmiA precursor |
|         | hypothetical protein |

**Species**: Bx, Bo
B. xylanisolvens non-PUL LGT Event 8

species

- NADH pyrophosphatase
- hypothetical protein
- Phosphoglucomutase
- Dipeptidase
- Glycyl-glycine endopeptidase ALE-1 precursor
- hypothetical protein
- TonB-dependent Receptor Plug Domain protein
- SusD family protein
- hypothetical protein
- KWG Leptospira
- Alanine dehydrogenase
- Nitrate monooxygenase
- hypothetical protein
- hypothetical protein
- ECF RNA polymerase sigma factor SigW
- putative nicotinate-nucleotide pyrophosphorylase [carboxylating]
- hypothetical protein
- Ribosomal RNA large subunit methyltransferase H
- hypothetical protein
- hypothetical protein
- putative hydrolase
- NADPH-dependent 7-cyano-7-deazaguanine reductase
- 7-cyano-7-deazaguanine synthase
- hypothetical protein
- hypothetical protein
- hypothetical protein
- ECF RNA polymerase sigma factor SigR
- hypothetical protein
- hypothetical protein
- hypothetical protein
- Outer membrane protein transport protein (OMPP1/FadL/TodX)
### B. xylanisolvens non-PUL LGT Event 10

| Species                                           | Bx | Bo |
|---------------------------------------------------|----|----|
| Unsaturated chondroitin disaccharide hydrolase    |    |    |
| Arylsulfatase                                     |    |    |
| Chondroitin sulfate ABC exoxygen precursor        |    |    |
| Cysteine--tRNA ligase                             |    |    |
| Sugar phosphatase YdA                             |    |    |
| Hypothetical protein                              |    |    |
| Hypothetical protein                              |    |    |
| ATP-dependent helicase HepA                       |    |    |
| tRNA-specific 2-thiouridylate MmmA                 |    |    |
| GH3 auxin-responsive promoter                     |    |    |
| 6-phosphofructokinase                             |    |    |
| Ribonuclease 3                                    |    |    |
| 3-oxoacyl-[acyl-carrier-protein] synthase 2        |    |    |
| Acyl carrier protein                              |    |    |
| Phosphoribosylglycinamide formyltransferase       |    |    |
| Erythronate 4-phosphate dehydrogenase             |    |    |
| Hypothetical protein                              |    |    |
| Lipopolysaccharide core biosynthesis protein      |    |    |
| D-inositol-3-phosphate glycosyltransferase        |    |    |
| Lipopolysaccharide core heptosyltransferase RtaQ  |    |    |
| 3-deoxy-D-manno-octulosonic acid kinase            |    |    |
| Hypothetical protein                              |    |    |
| Alpha-D-knosaminyltransferase                     |    |    |
| LicD family protein                               |    |    |
| Bifunctional IPC transferase and DIPP synthase    |    |    |
| Heptidinol-phosphate aminotransferase             |    |    |
| Chondroitin synthase                              |    |    |
| Chondroitin synthase                              |    |    |

---

*Note: The table above lists genes that have been transferred from other species to B. xylanisolvens as a result of non-PUL LGT events.*
xylanisolvensis non-PUL LGT Event 12

- bifunctional 3-demethylubiquinone-9-3-methyltransferase/2-octaprenyl-6-hydroxy phenol methylase
- Cell division protein PtsX
- hypothetical protein
- Undecaprenyl-diphosphatase
- tRNA pseudouridine synthase B
- S-adenosylmethionine:tRNA ribosyltransferase-isomerase
- Bifunctional folate synthesis protein
- hypothetical protein
- S-adenosylmethionine synthase
- hypothetical protein
- hypothetical protein
- hypothetical protein
- hypothetical protein
- hypothetical protein
Figure S9

A. GalM, GluM PUL-A

B. PUL-B expression during GalM growth

GalM, GluM PUL-B

sensor/regulator

susC/D-like binding & transport

GH26
GH36
GH26
GH130

susC/D-like binding & transport

GH130

B. ovatus D2

Fold difference in expression compared to glucose

SuC (BB5G 2492)  SuD (BB5G 2493)

10000
1000
100
10
1