AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE

Characterization and Spatial Mapping of the Human Gut Metasecretome

Florencia Velez-Cortes,[a,b] Harris Wang[a,c]

[a]Department of Systems Biology, Columbia University, New York, New York, USA
[b]Integrated Program in Cellular, Molecular, and Biomedical Studies, Columbia University, New York, New York, USA
[c]Department of Pathology and Cell Biology, Columbia University, New York, New York, USA

ABSTRACT Bacterially secreted proteins play an important role in microbial physiology and ecology in many environments, including the mammalian gut. While gut microbes have been extensively studied over the past decades, little is known about the proteins that they secrete into the gastrointestinal tract. In this study, we developed and applied a computational pipeline to a comprehensive catalog of human-associated metagenome-assembled genomes in order to predict and analyze the bacterial metasecretome of the human gut, i.e., the collection of proteins secreted out of the cytoplasm by human gut bacteria. We identified the presence of large and diverse families of secreted carbohydrate-active enzymes and assessed their phylogenetic distributions across different taxonomic groups, which revealed an enrichment in *Bacteroidetes* and *Verrucomicrobia*. By mapping secreted proteins to available metagenomic data from endoscopic sampling of the human gastrointestinal tract, we specifically pinpointed regions in the upper and lower intestinal tract along the lumen and mucosa where specific glycosidases are secreted by resident microbes. The metasecretome analyzed in this study constitutes the most comprehensive list of secreted proteins produced by human gut bacteria reported to date and serves as a useful resource for the microbiome research community.

IMPORTANCE Bacterially secreted proteins are necessary for the proper functioning of bacterial cells and communities. Secreted proteins provide bacterial cells with the ability to harvest resources from the exterior, import these resources into the cell, and signal to other bacteria. In the human gut microbiome, these actions impact host health and allow the maintenance of a healthy gut bacterial community. We utilized computational tools to identify the major components of human gut bacterially secreted proteins and determined their spatial distribution in the gastrointestinal tract. Our analysis of human gut bacterial secreted proteins will allow a better understanding of the impact of gut bacteria on human health and represents a step toward identifying new protein functions with interesting applications in biomedicine and industry.

KEYWORDS gut microbiome, human microbiome, metagenomics

The gut microbiome plays a vital role in human metabolism, and its deviation from homeostasis has increasingly been linked to various diseases (1–3). Our understanding of the healthy equilibrium state of the gut microbiome is complicated by the fact that closely related taxa possess vastly different, often understudied, metabolic abilities (4). In particular, gut microbes harbor huge metabolic capacities for biotransformation and degradation of dietary substrates that are otherwise indigestible by the host (5). For instance, bacterial carbohydrate-active enzymes (CAZymes) that process complex dietary and host-derived polysaccharides are abundantly found in the gut microbiome (3, 6, 7). CAZymes not only help convert and release various sugars into

Editor Naseer Sangwan, Cleveland Clinic

Address correspondence to Harris Wang, hw2429@columbia.edu.

The authors declare a conflict of interest. H.W. is a scientific advisor and equity holder of SNIPR Biome and Kingdom Supercultures. The authors declare no additional competing interests.

Received 1 August 2022
Accepted 11 November 2022

absorbable forms for the host, but they also facilitate interspecies cross-feeding (8). Since gut microbes are heterogeneously distributed along the gastrointestinal (GI) tract, their associated metabolic capacities can impact dietary metabolism from the proximal to the distal regions (9). Unfortunately, the biodistributions of bacterial metabolic enzymes across the GI tract and between the luminal and mucosal areas have not been adequately described to date. A deeper understanding of the spatial geography of microbial metabolism can help elucidate key gut metabolic biotransformation processes with relevance for nutrition and human health.

At a cellular resolution, microbially associated metabolism in the gut occurs either in the intracellular compartments of individual bacteria or in the extracellular milieu along the lumen or mucosal interfaces through bacterial secretion of digestive enzymes and proteins. Bacterial secretion mainly takes place through the general secretory (Sec) pathway, which relies on recognition of a N-terminal signal peptide tag on a target protein for active transport across a SecYEG channel (10) out of the cytoplasm. These Sec-exported proteins remain in the periplasmic space, are embedded into the inner or outer membranes, or are completely secreted extracellularly. Gram-negative gut microbes such as *Bacteroidetes* often contain many glycoside hydrolases and polysaccharide lyases, with some genomes encoding hundreds of such CAZymes (5). These CAZymes can often contain secretion-associated peptide sequences (5), which suggests that they may function in the extracellular compartment with community-wide effects (11, 12). Delineating the microbiome secretome can help elucidate the main modulators of bacterial community structure in the gut.

Past studies of protein secretion relied heavily on low-throughput experimental strategies that required expression, purification, and mass spectrometry analysis of the secreted proteins individually (13). More recently, advances in machine learning and protein structure predictions have led to *in silico* predictions of secreted proteins, and this method has been applied on large swaths of genomic data in bacteria from a variety of different environments (12, 14). One study revealed that host-associated bacteria encoded more extracellular proteins than bacteria from other environments (12). However, that study used inferred protein annotations from mapping 16S rRNA amplification data sets to reference genomes, which is an approach that can be limited when trying to annotate protein repertoires that are less conserved. The recent increase in metagenomic data sets and new assembly and binning pipeline improvements have created a wealth of metagenomically assembled genomes (MAGs) that have increased the number and diversity of available bacterial genomes (4, 15, 16). These advances can help better dissect the gut microbial secretome but have not been implemented to date.

Here, we describe a systematic analysis of the human gut secretome using a combination of *in silico* approaches to predict secreted proteins from MAGs and map their spatial distribution along the gastrointestinal tract. We annotated the function of secreted proteins and cataloged their enrichment in specific bacterial taxa in the gut. Analysis of the biogeography of secreted enzymes revealed interesting patterns of distributions that suggested functional specialization in different GI compartments, especially those belonging to CAZymes. This work represents the first large-scale systematic study of secreted proteins in bacterial MAGs associated with the human gut and provides a foundation to facilitate future efforts in gut microbiome manipulation and engineering.

RESULTS

Establishing a comprehensive gut bacterial secretome. We aimed to generate a comprehensive database of secreted proteins (i.e., the secretome) from the human gut microbiota (Fig. 1a) to elucidate their role in intermicrobial and host interactions and explore their functional significance in human physiology and metabolism. We first amassed and annotated 24,323 publicly available high-quality human gut MAGS (15) using Prodigal v2.6.3 (17) (see Materials and Methods), which resulted in 54 million open reading frames (ORFs) that were then clustered at 95% amino acid identity using

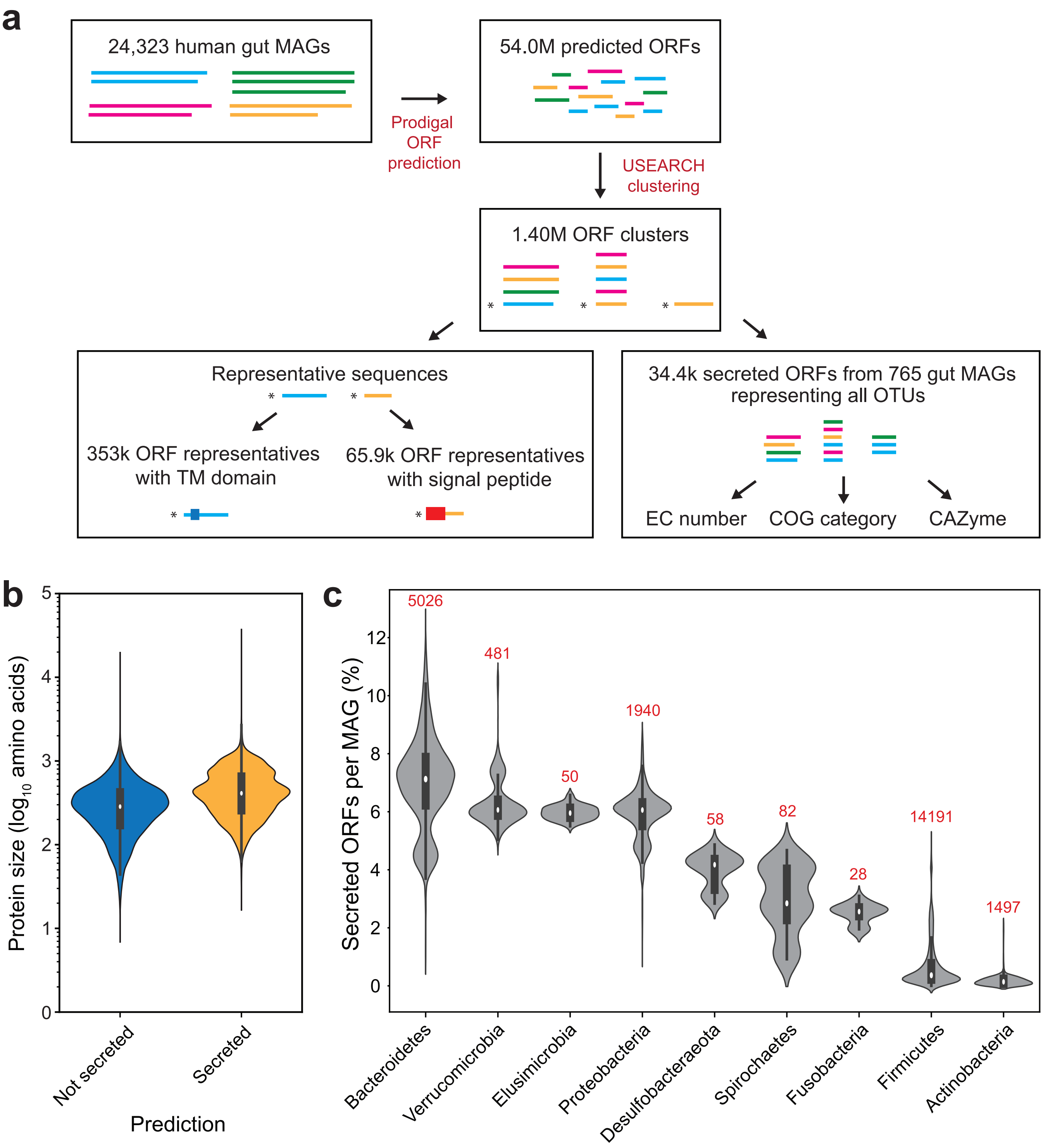

FIG 1 Human gut metasecretome prediction. (a) Illustration of human gut metasecretome prediction pipeline. Centroid sequences are denoted by asterisks. (b) Protein lengths of secreted and nonsecreted proteins. (c) Percentages of secreted ORFs by phylum in HGM MAGs. The numbers of MAGs in each phylum are shown above each violin plot.

USEARCH (18) into 1.40 million ORF clusters, each with at least 5 ORF sequences. We utilized a strategy of clustering sequences at a high sequence identity and annotating only representative centroid sequences to reduce the computational resources necessary to analyze the large HGM data set. Thus, the representative centroid sequence of each ORF cluster derived from USEARCH was then annotated using SignalP 5.0 (19) to identify possible presence of signal peptides. Since we are interested only in extracellularly secreted proteins, we filtered out sequences with lipoprotein signal peptides (a SignalP output) because they are likely to be embedded in the membrane (20), and we only maintained those with Sec1 and Tat pathway signal peptides (21). We further used TMHMM v2.0 (22) to identify and exclude sequences with transmembrane domains. A total of 37,511 representative centroid sequences were collated to yield a set of representative secretome ORFs that contained signal peptides but no transmembrane domains or lipoprotein signal peptides. Secretome designation of members within ORF clusters was assigned based on the representative secretome ORFs, which resulted in 1,627,958 ORFs that are putatively secreted out of the cytoplasm, which we designated the "gut bacterial secretome."

Since the gut is a highly competitive environment, there are likely important evolutionary drivers for optimization of secreted proteins. Secreted proteins are generally encoded with amino acids that are less expensive to produce (23), suggesting that there is a balance between the beneficial and altruistic functions of secreted proteins and their fitness burden on the producing bacteria. We therefore first explored whether there was a correlation between secretion status and protein size. Interestingly, we found that the gut-secreted proteins tend to be larger than nonsecreted proteins from gut bacteria (482 versus 329 amino acids on average, respectively; Mann-Whitney U-test $P < 10^{-3}$) (Fig. 1b). This suggested that, at the global level, the benefits of the secreted protein outweigh any metabolic cost of production for the secreting bacteria. When we analyzed the biosynthetic cost of secreted proteins, we observed that *Bacteroidetes* and verrucomicrobial MAGs tended to have similar biosynthesis costs per residue for secreted and nonsecreted proteins, while other gut phyla had a lower median cost for secreted proteins (Kruskal-Wallis H-test, $P < 10^{-56}$; Mann-Whitney U test with Bonferroni correction, $P < 10^{-4}$) (see Fig. S1 in the supplemental material).

Bacteroidetes strains, as well as some *Verrucomicrobia*, have been predicted to secrete a large proportion of their proteome (12, 24, 25). To examine the contribution of each phylum to the gut metasecretome, we compared the number of encoded proteins with signal peptides in each of the most prominent and diverse phyla in the human gut, selecting phyla that had at least 5 MAGs in the data set (Fig. 1c). *Bacteroidetes* and *Verrucomicrobia* tended to secrete a larger percentage of their proteome compared to other major gut phyla (Kruskal-Wallis H-test, $P < 10^{-3}$; Mann-Whitney *U* test with Bonferroni correction, $P < 10^{-3}$), with *Bacteroidetes* the top secreting phylum in the gut (Mann-Whitney *U* test with Bonferroni correction, $P < 10^{-3}$). Taken together, these results suggest that *Bacteroidetes* and *Verrucomicrobia* invest significantly in their secreted proteome and are key players in the final composition of the gut bacterial metasecretome.

Functional assessment of the gut bacterial secretome. To survey the functions of the secretome, we used the eggNOG protein ortholog database (26) to annotate secreted ORFs from clusters of proteins with more than 5 member sequences (Table S1). We only annotated one MAG per strain-level operational taxonomic unit (OTU) in the Human Gut Metagenomes (HGM) data set (15), so as to reduce redundancy and limit the computational resources required to do this task. The number of ORFs secreted per MAG was averaged over all representative MAGs in each phylum. The main COG categories among predicted secreted proteins in the gut microbiome were the following: (i) cell wall structure and biogenesis and outer membrane, (ii) carbohydrate metabolism and transport, (iii) inorganic ion transport and metabolism, (iv) energy production and conversion, (v) amino acid metabolism and transport, and (vi) secretion, motility, and chemotaxis (Fig. 2a). *Bacteroidetes* and *Verrucomicrobia* encoded the highest average number of secreted proteins per MAG in the carbohydrate metabolism and transport category. Top Enzyme Commission (EC) categories in the gut metasecretome included proteinases involved in cell membrane biogenesis, nucleotidases, and a variety of CAZymes, such as galactosidases and glucosidases (Fig. 2b), which implied that the human gut metasecretome is highly functionally diverse.

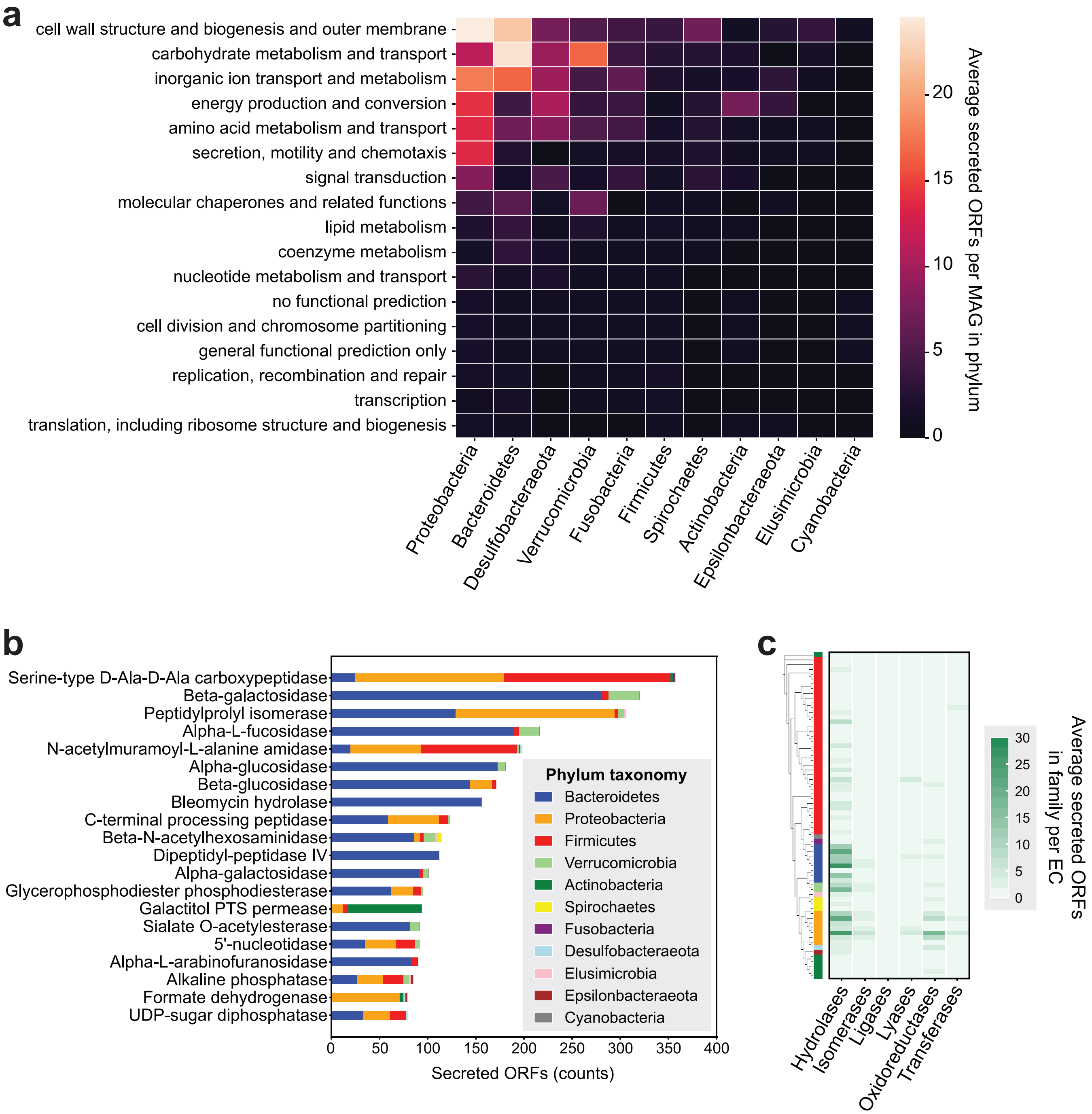

FIG 2 Main functions of the human gut metasecretome. (a) Average number of gut secreted ORFs in each COG category. (b) Top EC annotations of gut secreted proteins. (c) Average numbers of secreted ORFs belonging to each EC category in each MAG, clustered by phylogeny (see Materials and Methods for details).

To understand the differences in secreted protein functions within and across phyla, we examined the average number of encoded secreted enzymes per MAG in each bacterial family represented in the HGM data set. EC annotations were used to identify the distribution of hydrolases, isomerases, transferases, oxidoreductases, lyases, and ligases in gut bacterial families (Fig. 2c). We found that secreted hydrolases were encoded more commonly in the secretomes of *Bacteroidetes*, *Verrucomicrobia*, and certain *Proteobacteria* families. *Bacteroides* strains are known to form complex cross-feeding networks in the mammalian gut via secreted CAZymes (8), while *Verrucomicrobia* like

Akkermansia municiphila are known mucin degraders (27, 28). In line with these observations, *Bacteroidetes* and *Verrucomicrobia* MAGs in our study appeared to secrete more hydrolases than other gut phyla. Conversely, *Proteobacteria* secreted more oxidoreductases than other gut phyla included in this study, which was interesting since there are only a few known oxidoreductases that are extracellular in gut bacteria. One such case is Cgr2, a reductase encoded by *Eggerthella lenta* gut strains, which has the ability to metabolize and inactivate cardiac drug digoxin (29).

Since the gut microbiome contains a wealth of CAZymes and polysaccharide degradation plays a role in determining ecological niches within the gut, we wanted to define which gut bacterial CAZymes were likely to form part of the metasecretome. We used HMMER to annotate 1,911,738 ORFs from representative HGM MAGs (see Materials and Methods) with dbCAN CAZyme families, yielding 43,331 CAZyme annotations, 10.5% of which were secreted (Table S2). The most common CAZyme EC functions in the gut metasecretome included beta-*N*-acetylhexosaminidases, which are involved in degradation of chitin, intestinal mucosal glycans, and milk oligosaccharides (30–32), alpha-glucosidases, which are involved in the degradation of starch (33), and beta-glucosidases, which degrade cellulose (34, 35) (Fig. 3a). These results highlighted the complexity and abundance of gut bacterial enzymes dedicated to degrading human dietary substrates and host glycoproteins. When we compared the number of secreted CAZymes across different phyla, we found that *Bacteroidetes* and *Verrucomicrobia* secreted a higher percentage of their CAZymes than other gut phyla (Kruskal-Wallis H-test, $P < 10^{-3}$; Mann-Whitney *U* test with Bonferroni correction, $P < 10^{-3}$), in addition to encoding a large number of CAZymes (Fig. S2). To compare secretion of CAZyme families in gut bacterial MAGs, we performed principal-coordinates analysis (PCoA) (Fig. 3b). The principal coordinates were computed based on a matrix of the number of ORFs in each MAG that were annotated as part of a CAZyme family and whether these ORFs were annotated as secreted or not. We found that HGM bacterial phyla tended to cluster together, implying that a large part of the CAZyme repertoire is preserved at the phylum level in the human gut microbiome. However, *Verrucomicrobia* and *Bacteroidetes* MAGs tended to cluster closely together, suggesting that these MAGs have similar CAZyme repertoire features. To investigate this further, we examined the CAZyme abundance and secretion of MAGs in this *Bacteroidetes*-*Verrucomicrobia* (BV) cluster (Fig. S3). MAGs in the BV cluster tended to both encode and secrete a larger number of CAZyme families than did MAGs outside of this cluster. Many of the CAZymes unique to this cluster were predicted to be secreted and have been previously known to act upon animal carbohydrates (GH2, GH20, GH29, GH33, GH43, GH84, GH92, GH95, and GH109), plant cell wall carbohydrates (GH2, GH29, GH31, GH51, GH95, and GH127), or starch or glycogen (GH13) (6, 36). While some of these glycoside hydrolases have been previously reported in *Akkermansia municiphila*, other less-well-studied HGM *Verrucomicrobia* included in this study have not been previously reported to encode these CAZyme families. For instance, we identified several ORFs from *Opitutales* and UBA8416 MAGs as secretors of GH109 (data not shown), a CAZyme family with reported mucin-degrading abilities (37). However, these putative GH109 proteins had no close BLASTp matches and thus may represent novel degradative abilities in understudied members of the human gut *Verrucomicrobia*.

To understand the differences in secreted CAZyme repertoire in HGM bacterial families, we calculated the percentage of CAZymes secreted by each MAG in each CAZyme family in the HGM (Fig. S4). Most CAZyme families in the HGM are glycoside hydrolases and polysaccharide lyases, so we focused our analysis on these degradative enzymes. MAGs from the same phylum tended to cluster together, with some proteobacterial MAGs and some for *Firmicutes* as the exceptions. Verrucomicrobial MAGs tended to cluster closely with *Bacteroidetes*, suggesting a shared repertoire of CAZymes and thus similar glycan degradative capacities. Several glycoside hydrolases were more likely to be secreted in *Bacteroidetes* MAGs than in other phyla, including plant cell wall hydrolases (GH2, GH3, GH5, GH29, GH31, GH36, GH43, GH51, and GH127), peptidoglycan hydrolases (GH23, GH25, and GH73), sucrose or fructan hydrolases (GH32), and animal

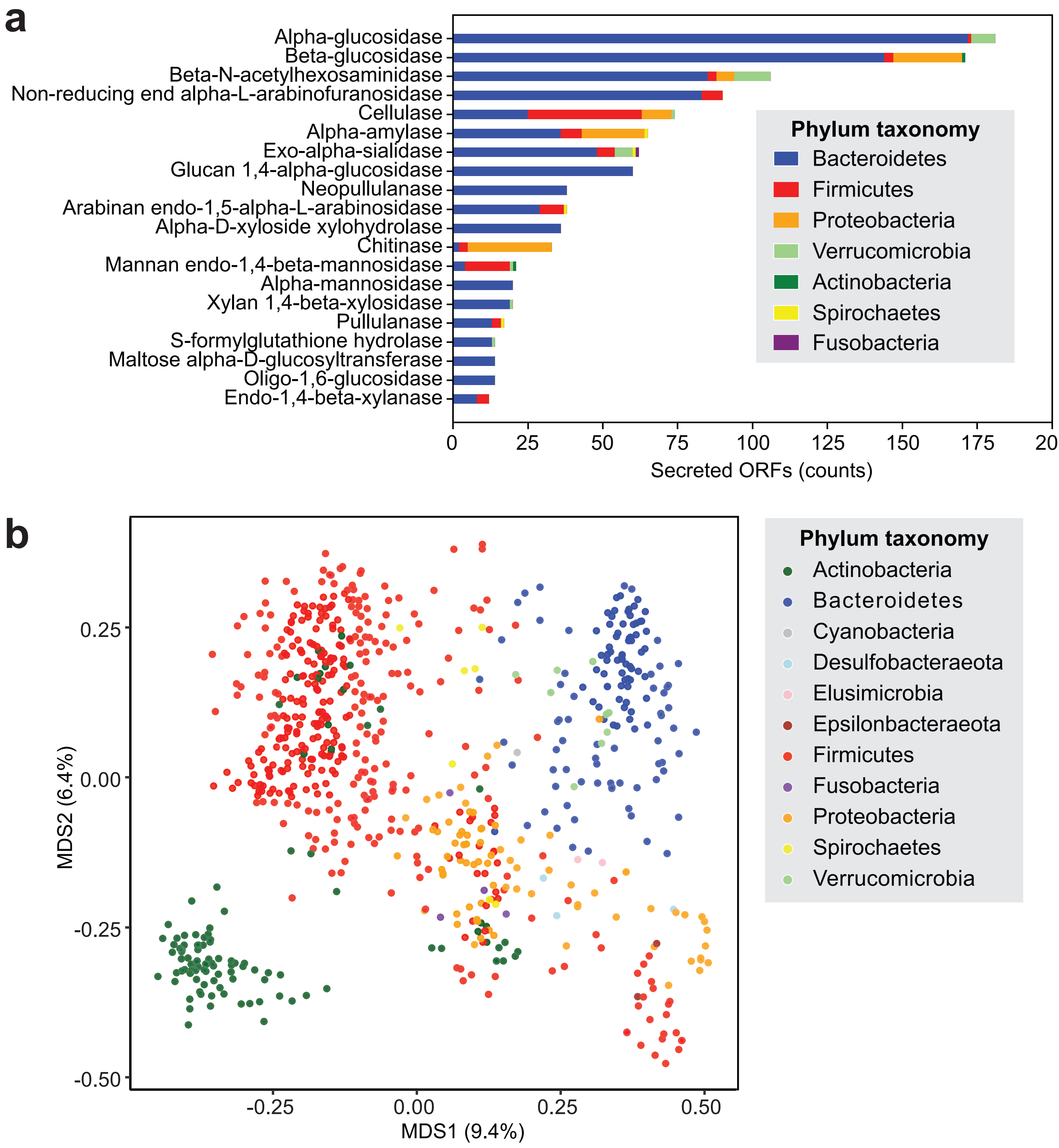

FIG 3 Carbohydrate degradation in the human gut metasecretome. (a) Most abundant secreted CAZyme families in the gut, colored by phylum. (b) Principal-coordinates analysis of representative MAGs secreted and nonsecreted with CAZyme abundance.

carbohydrate hydrolases (GH2, GH3, and GH29). However, these GHs were still encoded in the MAGs of most phyla in the data set, which suggested that they either fulfill other functions inside of the cell, or are exported via another route than the Sec pathway, or contain signal peptides that are not recognized by SignalP. GH1 and GH4 were two CAZymes that were encoded in most gut MAGs and not secreted, but they were also conspicuously not found in *Bacteroidetes* MAGs, despite being found in most other MAGs in the HGM, mainly *Firmicutes*. GH1 is a glycoside hydrolase family with

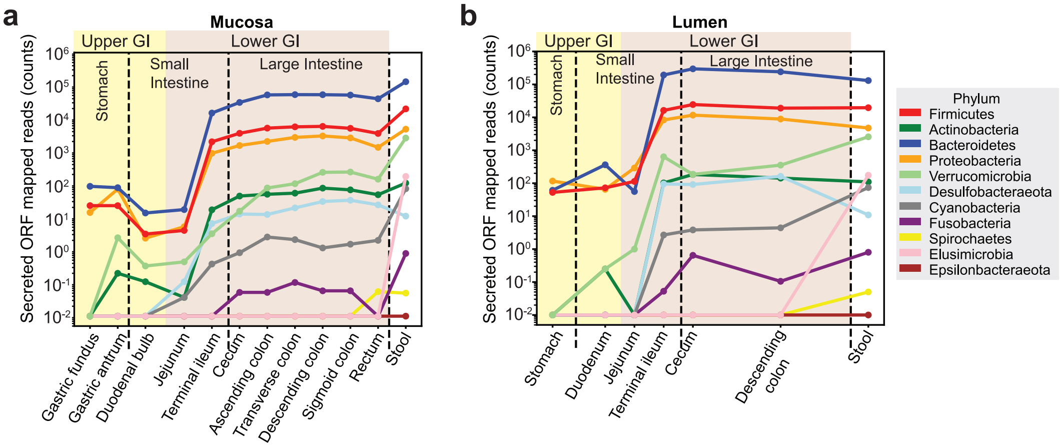

FIG 4 Biogeography of the gut metasecretome. (a) Reads mapping to secreted ORFs at the mucosa along the gastrointestinal tract. (b) Reads mapping to secreted ORFs in the lumen along the gastrointestinal tract.

beta-galactosidase, beta-glucosidase, and mannosidase activities. Some *Bacteroidetes* from nongut environments possess GH1, but thus far the only gut *Bacteroidetes* strain reported to possess GH1 is a ruminal strain of *Bacteroides* (38). GH4 has been mostly characterized in soil bacteria and participates in the degradation of raffinose, a plant-derived polysaccharide (39, 40). One of the few glycoside hydrolases that was more frequently secreted in *Firmicutes* than in *Bacteroidetes* was GH18, a CAZyme family that includes chitinases (EC 3.2.1.14) and endo-β-*N*-acetylglucosaminidases (EC 3.2.1.96). This may be indicative of a specific niche occupied by some *Firmicutes* strains in the gut or an alternative export mechanism for GH18 enzymes in *Bacteroidetes*. Together, our analysis of the human gut metasecretome demonstrated the diversity and abundance of secreted enzymatic functions in the healthy human gut. We showed that gut phyla encode vastly different CAZyme repertoires, with the exception of *Bacteroidetes* and *Verrucomicrobia*, which suggests an overlap in secreted degradative abilities among these two members of the gut microbiome. Finally, we identified novel putative glycosidase hydrolase families in verrucomicrobial MAGs.

Mapping the GI biogeography of secreted proteins. To establish the general biogeography of secreted bacterial proteins in the gut, we aligned publicly available metagenomic reads from endoscopic and stool samples (41) to secreted ORFs predicted from a set of representative HGM MAGs by using Bowtie2 (see Materials and Methods) (Fig. S5a). Quality filtering of the metagenomic reads showed a high, albeit variable, sequencing depth across the human GI (Fig. S5b), and after mapping we were able to obtain reasonable coverage of HGM MAGs (Fig. S5c), which implied that our metasecretome ORFs were detectable in metagenomic GI reads.

By mapping metagenomic reads from endoscopic and stool samples to secreted ORFs from each phylum, we were able to estimate the contribution of each phylum to the metasecretome. We found that luminal samples mirrored stool samples more closely than mucosal samples and that samples taken at distal sites were more similar to stool samples than those from proximal sites (Fig. 4). This was similar to what was observed in the KEGG Orthology functional distribution of upper and lower GI in the original study (41). To some extent, this discrepancy arises from the varying taxonomic composition found across the GI tract, which reflects the need for different resource-harvesting abilities at each habitat within the GI tract. Across the entire GI tract, *Bacteroidetes*, *Firmicutes*, and *Proteobacteria* tend to encode the highest number of secretome ORFs of all gut phyla included in this analysis, with *Bacteroidetes* accounting for the highest number of mapped reads to secreted ORFs in the lower GI. *Bacteroidetes* tended to have more reads mapping to secreted ORFs in luminal samples than in mucosal samples, especially in the distal small intestine and proximal large intestine. This finding could be due to the diversity of dietary substrates available for degradation present in the luminal compartment. Based on the number of secretome reads mapped against the GI metagenomic samples, we determined that (i) the predicted metasecretome proteins were detectable throughout the GI tract, (ii) both mucosal and luminal compartments were abundant in secreted proteins, particularly in the lower GI, and (iii) secreted proteins from *Bacteroidetes* dominated the lower GI.

To determine which secreted protein functions were most abundant across different sublocalizations of the GI tract, we mapped GI endoscopic reads to ORFs from each EC function in the metasecretome (Table S3). We calculated a modified reads per kilobase per million reads (RPKM) metric for each EC function based on the mean nucleotide length of ORFs annotated with the EC function (see Materials and Methods). Most of the upper GI tract samples had little to no signal, which was likely due to (i) lower sequencing depth of upper GI tract samples (Fig. S5b) and (ii) lower overlap between HGM MAGs (15) derived from stool bacteria and bacteria residing in the upper GI tract (Fig. S5c). We focused on EC functions with pseudo-RPKM values higher than 50 in at least one GI site in order to reduce effects from noise and analyze more-prevalent ECs. We found that two EC functions, beta-galactosidase (EC 3.2.1.23) and alpha-L-fucosidase (EC 3.2.1.51), had the highest relative abundances in luminal and mucosal lower gastrointestinal tract samples (Fig. S6). While nearly 70% of the population is lactose intolerant (42), generally the Western human diet is high in lactose, which is a substrate for beta-galactosidases (3), and previous studies have noted the presence of this enzyme in *Bifidobacteria* in the gut microbiome (43).

To compare the abundance of each secreted enzyme across sites, we normalized the RPKM values of each GI site against the highest RPKM for that enzyme, generating a *z*-score for each EC and GI site. Most of the secreted proteins that we were able to map were more abundant in lower GI samples than in stool, including numerous glycosidases (EC 3.2.1.x) and 2-dehydropantoate 2-reductase (EC 1.1.1.169) (Fig. 5a and Fig. S7a). The most abundant secreted EC functions in the upper GI tract were 5′-nucleotidases and 3′-nucleotidases. There were also several secreted glycosidases, such as cellulose and mannan endo-1,4-beta-mannosidase, as well as a serine endopeptidase, that were overrepresented in stool samples relative to the rest of the GI tract. Generally, the lower GI tract was enriched in secreted proteins, which is expected since *Bacteroidetes* tend to reside in the colon (9) and are major contributors to the metasecretome.

The luminal and mucosal compartments of the GI contain differing substrates that microbes can feed on, with higher concentrations of mucin and other host glycans at the mucosa and fiber and starches from digesta in the lumen. We posited that these differences in substrates would result in secreted degradative proteins at different parts of the intestine. To identify secreted EC functions that were specific to mucosal or luminal sites of the GI tract, we took the ratio of RPKM in mucosal sites and luminal sites (Fig. 5b and Fig. S7b). We identified glycosidases that were overrepresented in the stomach mucosa, in particular, xylan 1,4-beta-xylosidase, sialase *O*-acetlyesterase, neopullulanase, alpha-L-rhamnosidase, alpha-galactosidase, alpha-L-fucosidase, dipeptidyl-peptidase IV, beta-glucosidase, and beta-galactosidase. In general, most putative secreted EC functions were enriched in the luminal areas of the GI tract. However, the distribution of secreted EC functions became more evenly distributed between luminal and mucosal compartments as we approached the distal end of the GI tract. Since the mucus layer becomes thicker in the colon (9), the mucosa can harbor more bacterial metabolic activity in the form of mucin degradation (44) and may support subsequent cycles of cross-feeding interactions.

DISCUSSION

Here, we have presented an extensive data set of human gut secreted proteins and an analysis of their main functions and distribution in the gut. When we analyzed the phylogenetic distribution of secreted proteins, we found that the number of secreted proteins encoded in different phyla varied widely, with *Bacteroidetes* being the main

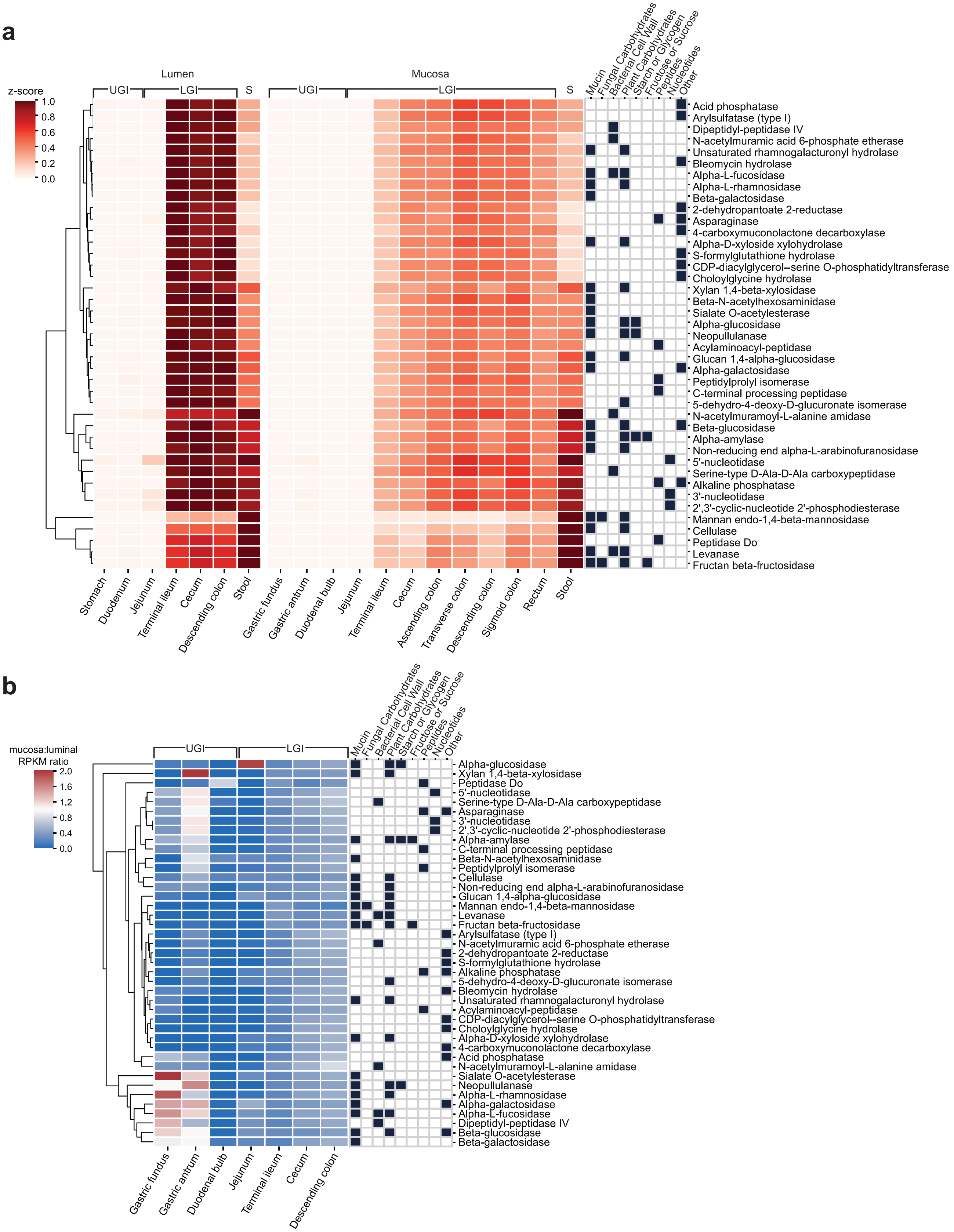

FIG 5 Biogeography of gut microbiome secreted functions. (a) *Z*-score of RPKM abundance of EC categories across gastrointestinal luminal and mucosal samples, with stool sample repeated for ease of comparison. (b) Ratio of mucosal to luminal RPKM abundance of EC categories across the gastrointestinal tract. Sample RPKM labels: UGI, upper gastrointestinal tract; LGI, lower gastrointestinal tract; S, stool.

contributor to the gut metasecretome. Moreover, *Verrucomicrobia* also appeared to secrete a notable portion of its proteome. We performed the most extensive comparative analysis of secreted CAZymes in the gut to date and showed that *Verrucomicrobia* and *Bacteroidetes* encode similar secreted CAZyme repertoires, which hints at similar glycan degradation abilities. Further study on cultured isolates is necessary to verify the degradative abilities that are encoded by these predicted secreted CAZymes. A previous study found evidence of *Bacteroidetes*-*Verrucomicrobia* horizontal gene transfer in the human gut microbiome, and CAZymes have been shown to be part of the mobilome (45), but further research is needed to determine whether secreted CAZymes are being shared between *Verrucomicrobia* and *Bacteroidetes*.

In this work, we found that *Bacteroidetes* and *Verrucomicrobia* tend to export a substantial fraction of their proteome. A large proportion of the secretome of these two phyla is dedicated to CAZymes that participate in the breakdown of complex carbohydrates, which has been shown previously in certain members of these phyla in the gut (3, 5). Our observations reinforced a view that places most *Bacteroidetes* and *Verrucomicrobia* in the highest trophic level of the gut, where they can utilize dietary fibers and mucin directly, while other phyla benefit from the breakdown products derived from these primary degradation reactions (46, 47). We also observed that secreted animal carbohydrate hydrolases were present in higher numbers in *Akkermansiaceae* and *Bacteroidaceae* among other *Verrucomicrobia* and *Bacteroidetes* families compared to other gut phyla (data not shown), which further suggested that strains from these families are specialized in host glycan degradation. Development of methods for the study of complex communities (48, 49) as well as more high-throughput assays that can characterize CAZyme substrate specificity will be helpful in elucidating the ecological interactions and dynamics of gut bacteria.

We have taken care to impose several filters on our secreted protein prediction; however, signal peptide prediction is not always accurate, although algorithms have improved over the last decade. The genomes used to train these algorithms reflect the availability of current data, which is biased toward more well-researched phyla, such as *Proteobacteria*. We also expect that a portion of the predicted secreted proteins are periplasmic. Since some of these periplasmic proteins may become public goods via outer membrane vesicles (50), we decided to not filter them. Because of computational resource limitations, we focused our functional analysis of secreted proteins on one MAG per strain-level OTU. We also clustered protein sequences and removed proteins that did not cluster with at least 4 other proteins in the data set. This underestimated the true diversity of the secretome, which likely varied highly at the strain level. However, we expect this approach still leaves us with a comprehensive catalog of the gut microbiome metasecretome. Finally, while many CAZymes we found are associated with breakdown of mucin or dietary fiber or other sources of nourishment for the gut microbiome, some CAZymes we identified in the gut are not unique to the gut and are involved with energy production (GH1, GH13, GH31, GH32, and GH38) or peptidoglycan breakdown (GH23, GH25, and GH73) (6).

To our knowledge, this is the first study to map the bacterial secretome across the GI tract, in which we validated the prevalence of the predicted metasecretome proteins and identified functional enrichment in different gut habitats. We observed that many secreted glycosidases are enriched in the luminal lower GI tract and are underrepresented in stool samples. This underscores the importance of expanding our studies beyond stool isolates and ensure the representation of bacteria unique to the GI tract in culturomics efforts. Given that the gut presents a unique habitat for cooperativity and competition among bacteria (12) with opportunities for the evolution of secreted proteins with undiscovered degradative abilities (32), the present study shows the vastness of secreted proteins in the gut has barely been uncovered and represents an opportunity to discover new interactions among bacteria and between humans and their gut microbiome.

MATERIALS AND METHODS

MAGs and ORF annotation. We used the high-quality Human Gut Metagenome (HGM) MAGs reported by Nayfach et al. (15). We selected MAGs that were in OTUs that were classified as Bacterial by

Nayfach using GTDB-Tk (51). We annotated each MAG using Prodigal version 2.6.3 (17) to identify ORFs by using modified settings that allowed for smaller ORF discovery (52). Briefly, we modified prodigal source code so that we could find smaller genes; specifically, we changed the MIN_GENE parameter in Prodigal-2.6.3/node·h to 15, so that we could identify ORFs that had 15 nucleotides or more. To ensure the MAGs we were using were of sufficient quality, we required a minimum of 482 ORFs per MAG, since that is the number of genes in the smallest known bacterial species, *Mycoplasma genitalium* (53). We selected representative gut MAGs by taking one MAG at random from each species-level OTU in the HGM high-quality MAG data set, resulting in 765 representative gut bacterial MAGs.

ORFs from representative MAGs were annotated with HMMER (54) using the dbCAN CAZyme database (55) and eggNOG v5.0 database (26). HMMER search criteria for CAZyme identification included 0.35 minimum coverage of the CAZyme and a minimum e-value of 1e−15, which were enforced using the hmmscan-parser tool from dbCAN. eggNOG search results were required to have a minimum e-value of 0.001. The CAZyme heatmap was clustered using seaborn with Euclidean metric and single method. CAZyme families were removed from the heatmap if present in only one representative MAG. Broad substrates were obtained using the supplementary table from Cantarel et al. (6), in addition to literature searches using CAZydb.

Phylogenetic analysis. To construct a phylogenetic tree of all the families of interest in the Nayfach data, we used the same method used by Nayfach but on a select set of representative MAGs, one from each bacterial OTU, totaling 765 MAGs. In brief, GTDB-Tk was used with the classify_wf workflow to call the marker protein sequences from MAGs using Prodigal and HMMER, aligned the concatenated marker sequences, and used pplacer to construct a maximum-likelihood tree. This tree was plotted using iqtree version 1.6.12 with the following server command: iqtree -s gtdbtk.bac120.user_msa.fasta -st AA -m MFP -nt 4.

Clustering and SignalP secretion tag prediction. We clustered human gut metagenome ORFs with USEARCH (18) at 95% amino acid sequence identity. Representative sequences of clusters with more than 5 sequences were annotated using SignalP 5.0 (19) to predict secretion tags. Sequences with a tag corresponding to type 1 secretion or TAT secretion were considered "secreted," and those for lipoprotein or no secretion tag were considered "not secreted." Protein sequences were further annotated using TMHMM (22) to determine whether they contained transmembrane domains. Proteins that contained transmembrane domains were classified as not secreted.

Principal-coordinates analysis of CAZymes in representative human gut MAGs. An in-house R script (56) was used to create a matrix of counts of ORFs present in each CAZyme family from CAZydb annotations of the representative human gut MAGs data set. If a CAZyme family had ORFs that were secreted and nonsecreted, we counted these as two separate CAZyme families. We then calculated a distance matrix based on the Spearman correlation of the CAZyme family counts.

Mapping of secreted proteins to GI tract. We used bowtie2 version 2.4.2 (57) to align Elinav metagenomics reads from endoscopic and stool samples to a set of representative gut MAGs. First, we applied a similar quality-filtering method to that used by Elinav et al. (41) for their samples. That is, we used trimmomatic version 0.39 (58) to perform adapter trimming using the following command: trimmomatic PE -validatePairs -threads 2 -phred33 input_readsQC/eachSample_R1_001.fastq.gz input_readsQC/eachSample_R2_001.fastq.gz eachSample_R1_paired.fq.gz eachSample_R1_unpaired.fq.gz eachSample_R2_paired.fq.gz eachSample_R2_unpaired.fq.gz ILLUMINACLIP: TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 MINLEN:50.

Then, we performed filtering of human reads from the samples using bowtie2 against the bowtie2 index for the human genome reference Hg19. We counted unique reads mapping to secreted ORFs from each phylum and for each Enzyme Commission functional annotation using an in-house script. We calculated a pseudo-RPKM value for each sample using the mean nucleotide lengths of ORFs in representative HGM MAGs that were annotated with a particular EC function. Samples taken from the same GI site were averaged. We included in the heatmaps only EC categories that had a minimum value greater than 50 RPKM for at least one GI site. To normalize all EC function RPKMs, we divided RPKM values in an EC function by the maximum RPKM for that EC function across all GI tract and stool samples.

Statistical analysis. Python package SciPy (59) was used to perform a Mann-Whitney *U* test to determine *P* values for the differences in protein length between secreted and nonsecreted ORFs. The same package was used to perform a Kruskal-Wallis test and *post hoc* Mann-Whitney *U* tests with Bonferroni corrections to determine differences in biosynthetic costs from each major phylum and differences in percentages of proteins secreted by each major phylum and of CAZymes secreted by each major phylum. MAGs were considered to be major gut phyla if there were over 50 MAGs in the data set.

Code availability. Metasecretome prediction scripts can be accessed at https://github.com/fgv2104/gut_metasecretome.

Data availability. Predicted HGM secreted protein sequences are available for download as HGM_secreted_orfs.faa.gz from https://github.com/fgv2104/gut_metasecretome.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TABLE S1, XLSX file, 4 MB.

TABLE S2, XLSX file, 3.9 MB.

TABLE S3, XLSX file, 0.1 MB.

FIG S1, EPS file, 0.8 MB.

FIG S2, EPS file, 0.9 MB.

FIG S3, EPS file, 2.5 MB.

FIG S4, TIF file, 2.6 MB.
FIG S5, TIF file, 2.7 MB.
FIG S6, EPS file, 2.7 MB.
FIG S7, EPS file, 1.8 MB.

ACKNOWLEDGMENTS

We thank the Wang lab for feedback and helpful discussions regarding this project, especially G. Urtecho, C. Mavros, and Y. Huang. H.W. acknowledges funding support from DOE (47879/SCW1710), NSF (MCB-2032259, MCB-2025515), NIH (1R01AI132403, 1R01DK118044, 1R21AI146817), the Burroughs Wellcome Fund (1016691), and the Hirshl Trust. F.V.-C. is supported by an NSF Graduate Research Fellowship (DGE-1644869).

H.W. is a scientific advisor and equity holder of SNIPR Biome and Kingdom Supercultures. We declare no other competing interests.

REFERENCES

1. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI. 2009. A core gut microbiome in obese and lean twins. Nature 457:480–484. https://doi.org/10.1038/nature07540.
2. Desai MS, Seekatz AM, Koropatkin NM, Kamada N, Hickey CA, Wolter M, Pudlo NA, Kitamoto S, Terrapon N, Muller A, Young VB, Henrissat B, Wilmes P, Stappenbeck TS, Núñez G, Martens EC. 2016. A dietary fiber-deprived gut microbiota degrades the colonic mucus barrier and enhances pathogen susceptibility. Cell 167:1339–1353.e21. https://doi.org/10.1016/j.cell.2016.10.043.
3. Cockburn DW, Koropatkin NM. 2016. Polysaccharide degradation by the intestinal microbiota and its influence on human health and disease. J Mol Biol 428:3230–3252. https://doi.org/10.1016/j.jmb.2016.06.021.
4. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, Segata N, Kyrpides NC, Finn RD. 2021. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol 39:105–114. https://doi.org/10.1038/s41587-020-0603-3.
5. El Kaoutari A, Armougom F, Gordon JI, Raoult D, Henrissat B. 2013. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. Nat Rev Microbiol 11:497–504. https://doi.org/10.1038/nrmicro3050.
6. Cantarel BL, Lombard V, Henrissat B. 2012. Complex carbohydrate utilization by the healthy human microbiome. PLoS One 7:e28742. https://doi.org/10.1371/journal.pone.0028742.
7. Wardman JF, Bains RK, Rahfeld P, Withers SG. 2022. Carbohydrate-active enzymes (CAZymes) in the gut microbiome. Nat Rev Microbiol 20:542–556. https://doi.org/10.1038/s41579-022-00712-1.
8. Rakoff-Nahoum S, Coyne MJ, Comstock LE. 2014. An ecological network of polysaccharide utilization among human intestinal symbionts. Curr Biol 24:40–49. https://doi.org/10.1016/j.cub.2013.10.077.
9. Donaldson GP, Lee SM, Mazmanian SK. 2016. Gut biogeography of the bacterial microbiota. Nat Rev Microbiol 14:20–32. https://doi.org/10.1038/nrmicro3552.
10. Tsirigotaki A, De Geyter J, Šoštaric N, Economou A, Karamanou S. 2017. Protein export through the bacterial Sec pathway. Nat Rev Microbiol 15:21–36. https://doi.org/10.1038/nrmicro.2016.161.
11. Rakoff-Nahoum S, Foster KR, Comstock LE. 2016. The evolution of cooperation within the gut microbiota. Nature 533:255–259. https://doi.org/10.1038/nature17626.
12. Garcia-Garcera M, Rocha EPC. 2020. Community diversity and habitat structure shape the repertoire of extracellular proteins in bacteria. Nat Commun 11. https://doi.org/10.1038/s41467-020-14572-x.
13. Xiong W, Abraham PE, Li Z, Pan C, Hettich RL. 2015. Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota. Proteomics 15:3424–3438. https://doi.org/10.1002/pmic.201400571.
14. Orsi WD, Richards TA, Francis WR. 2018. Predicted microbial secretomes and their target substrates in marine sediment. Nat Microbiol 3:32–37. https://doi.org/10.1038/s41564-017-0047-9.
15. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. 2019. New insights from uncultivated genomes of the global human gut microbiome. Nature 568:505–510. https://doi.org/10.1038/s41586-019-1058-x.
16. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol 31:533–538. https://doi.org/10.1038/nbt.2579.
17. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119.
18. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461. https://doi.org/10.1093/bioinformatics/btq461.
19. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 37:420–423. https://doi.org/10.1038/s41587-019-0036-z.
20. Okuda S, Tokuda H. 2011. Lipoprotein sorting in bacteria. Annu Rev Microbiol 65:239–259. https://doi.org/10.1146/annurev-micro-090110-102859.
21. Green ER, Mecsas J. 2016. Bacterial secretion systems: an overview. Microbiol Spectr 4. https://doi.org/10.1128/microbiolspec.VMBF-0012-2015.
22. Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580. https://doi.org/10.1006/jmbi.2000.4315.
23. Nogueira T, Touchon M, Rocha EPC. 2012. Rapid evolution of the sequences and gene repertoires of secreted proteins in bacteria. PLoS One 7:e49403. https://doi.org/10.1371/journal.pone.0049403.
24. Bendtsen JD, Binnewies TT, Hallin PF, Sicheritz-Pontén T, Ussery DW. 2005. Genome update: prediction of secreted proteins in 225 bacterial proteomes. Microbiology (Reading) 151:1725–1727. https://doi.org/10.1099/mic.0.28029-0.
25. Wilson MM, Anderson DE, Bernstein HD. 2015. Analysis of the outer membrane proteome and secretome of Bacteroides fragilis reveals a multiplicity of secretion mechanisms. PLoS One 10:e0117732. https://doi.org/10.1371/journal.pone.0117732.
26. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol 34:2115–2122. https://doi.org/10.1093/molbev/msx148.
27. Tailford LE, Crost EH, Kavanaugh D, Juge N. 2015. Mucin glycan foraging in the human gut microbiome. Front Genet 6:81. https://doi.org/10.3389/fgene.2015.00081.
28. Derrien M, Vaughan EE, Plugge CM, de Vos WM. 2004. Akkermansia muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. Int J Syst Evol Microbiol 54:1469–1476. https://doi.org/10.1099/ijs.0.02873-0.
29. Koppel N, Bisanz JE, Pandelia ME, Turnbaugh PJ, Balskus EP. 2018. Discovery and characterization of a prevalent human gut bacterial enzyme sufficient for the inactivation of a family of plant toxins. Elife 7. https://doi.org/10.7554/eLife.33953.
30. Bode L. 2012. Human milk oligosaccharides: every baby needs a sugar mama. Glycobiology 22:1147–1162. https://doi.org/10.1093/glycob/cws074.
31. Chen HC, Chang CC, Mau WJ, Yen LS. 2002. Evaluation of N-acetylchitooligosaccharides as the main carbon sources for the growth of intestinal

bacteria. FEMS Microbiol Lett 209:53–56. https://doi.org/10.1111/j.1574-6968.2002.tb11108.x.

32. Luis AS, Martens EC. 2018. Interrogating gut bacterial genomes for discovery of novel carbohydrate degrading enzymes. Curr Opin Chem Biol 47:126–133. https://doi.org/10.1016/j.cbpa.2018.09.012.
33. D'Elia JN, Salyers AA. 1996. Contribution of a neopullulanase, a pullulanase, and an α-glucosidase to growth of Bacteroides thetaiotaomicron on starch. J Bacteriol 178:7173–7179. https://doi.org/10.1128/jb.178.24.7173-7179.1996.
34. Moraïs S, Ben David Y, Bensoussan L, Duncan SH, Koropatkin NM, Martens EC, Flint HJ, Bayer EA. 2016. Enzymatic profiling of cellulosomal enzymes from the human gut bacterium, Ruminococcus champanellensis, reveals a fine-tuned system for cohesin-dockerin recognition. Environ Microbiol 18:542–556. https://doi.org/10.1111/1462-2920.13047.
35. Singhania RR, Patel AK, Sukumaran RK, Larroche C, Pandey A. 2013. Role and significance of beta-glucosidases in the hydrolysis of cellulose for bioethanol production. Bioresour Technol 127:500–507. https://doi.org/10.1016/j.biortech.2012.09.012.
36. Drula E, Garron ML, Dogan S, Lombard V, Henrissat B, Terrapon N. 2022. The carbohydrate-active enzyme database: functions and literature. Nucleic Acids Res 50:D571–D577. https://doi.org/10.1093/nar/gkab1045.
37. Teze D, Shuoker B, Chaberski EK, Kunstmann S, Fredslund F, Nielsen TS, Stender EGP, Peters GHJ, Karlsson EN, Welner DH, Hachem MA. 2020. The catalytic acid-base in GH109 resides in a conserved GGHGG loop and allows for comparable α-retaining and β-inverting activity in an N-acetylgalactosaminidase from Akkermansia muciniphila. ACS Catal 10:3809–3816. https://doi.org/10.1021/acscatal.9b04474.
38. Whitehead TR. 1993. Analyses of the gene and amino acid sequence of the Prevotella (bacteroides) ruminicola 23 xylanase reveals unexpected homology with endoglucanases from other genera of bacteria. Curr Microbiol 27:27–33. https://doi.org/10.1007/BF01576830.
39. Thompson J, Hess S, Pikis A. 2004. Genes malh and pagl of Clostridium acetobutylicum ATCC 824 encode NAD+- and Mn2+-dependent phospho-α-glucosidase(s). J Biol Chem 279:1553–1561. https://doi.org/10.1074/jbc.M310733200.
40. Heravi KM, Watzlawick H, Altenbuchner J. 2019. The melREDCA operon encodes a utilization system for the raffinose family of oligosaccharides in bacillus subtilis. J Bacteriol 201. https://doi.org/10.1128/JB.00109-19.
41. Zmora N, Zilberman-Schapira G, Suez J, Mor U, Dori-Bachash M, Bashiardes S, Kotler E, Zur M, Regev-Lehavi D, Brik RBZ, Federici S, Cohen Y, Linevsky R, Rothschild D, Moor AE, Ben-Moshe S, Harmelin A, Itzkovitz S, Maharshak N, Shibolet O, Shapiro H, Pevsner-Fischer M, Sharon I, Halpern Z, Segal E, Elinav E. 2018. Personalized gut mucosal colonization resistance to empiric probiotics is associated with unique host and microbiome features. Cell 174:1388–1405.e21. https://doi.org/10.1016/j.cell.2018.08.041.
42. Storhaug CL, Fosse SK, Fadnes LT. 2017. Country, regional, and global estimates for lactose malabsorption in adults: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol 2:738–746. https://doi.org/10.1016/S2468-1253(17)30154-1.
43. Viborg AH, Katayama T, Arakawa XT, Hachem XMA, Leggio XL, Lo Kitaoka XM, Svensson B, Fushinobu XS. 2017. Discovery of α-L-arabinopyranosidases from human gut microbiome expands the diversity within glycoside hydrolase family 42. J Biol Chem 292:21092–21101. https://doi.org/10.1074/jbc.M117.792598.
44. Berry D, Stecher B, Schintlmeister A, Reichert J, Brugiroux S, Wild B, Wanek W, Richter A, Rauch I, Decker T, Loy A, Wagner M. 2013. Host-compound foraging by intestinal microbiota revealed by single-cell stable isotope probing. Proc Natl Acad Sci U S A 110:4720–4725. https://doi.org/10.1073/pnas.1219247110.
45. Groussin M, Poyet M, Sistiaga A, Kearney SM, Moniz K, Noel M, Hooker J, Gibbons SM, Segurel L, Froment A, Mohamed RS, Fezeu A, Juimo VA, Lafosse S, Tabe FE, Girard C, Iqaluk D, Nguyen LTT, Shapiro BJ, Lehtimäki J, Ruokolainen L, Kettunen PP, Vatanen T, Sigwazi S, Mabulla A, Domínguez-Rodrigo M, Nartey YA, Agyei-Nkansah A, Duah A, Awuku YA, Valles KA, Asibey SO, Afihene MY, Roberts LR, Plymoth A, Onyekwere CA, Summons RE, Xavier RJ, Alm EJ. 2021. Elevated rates of horizontal gene transfer in the industrialized human microbiome. Cell 184:2053–2067.e18. https://doi.org/10.1016/j.cell.2021.02.052.
46. Chung WSF, Walker AW, Vermeiren J, Sheridan PO, Bosscher D, Garcia-Campayo V, Parkhill J, Flint HJ, Duncan SH. 2018. Impact of carbohydrate substrate complexity on the diversity of the human colonic microbiota. FEMS Microbiol Ecol 95:fiy201. https://doi.org/10.1093/femsec/fiy201.
47. Flint HJ, Scott KP, Duncan SH, Louis P, Forano E. 2012. Microbial degradation of complex carbohydrates in the gut. Gut Microbes 3:289–306. https://doi.org/10.4161/gmic.19897.
48. Vrancken G, Gregory AC, Huys GRB, Faust K, Raes J. 2019. Synthetic ecology of the human gut microbiota. Nat Rev Microbiol 17:754–763. https://doi.org/10.1038/s41579-019-0264-8.
49. Shetty SA, Kuipers B, Atashgahi S, Aalvink S, Smidt H, de Vos WM. 2022. Inter-species metabolic interactions in an in-vitro minimal human gut microbiome of core bacteria. NPJ Biofilms Microbiomes 8. https://doi.org/10.1038/s41522-022-00275-2.
50. Elhenawy W, Debelyy MO, Feldman MF. 2014. Preferential packing of acidic glycosidases and proteases into Bacteroides outer membrane vesicles. mBio 5. https://doi.org/10.1128/mBio.00909-14.
51. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. Bioinformatics 36. https://doi.org/10.1093/bioinformatics/btz848.
52. Durrant MG, Bhatt AS. 2021. Automated prediction and annotation of small open reading frames in microbial genomes. Cell Host Microbe 29:121–131.e4. https://doi.org/10.1016/j.chom.2020.11.002.
53. Goffeau A. 1995. Life with 482 genes. Science 270:445–446. https://doi.org/10.1126/science.270.5235.445.
54. Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
55. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. 2012. DbCAN: A web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 40:W445–W451. https://doi.org/10.1093/nar/gks479.
56. R Core Team. 2016. R: a language and environment for statistical computing. R Found Stat Comput Version 3:3503.
57. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.
58. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
59. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, Vijaykumar A, Bardelli A, Pietro Rothberg A, Hilboll A, Kloeckner A, Scopatz A, Lee A, Rokem A, Woods CN, Fulton C, Masson C, Häggström C, Fitzgerald C, Nicholson DA, Hagen DR, Pasechnik DV., SciPy 1.0 Contributors. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2.