Functional screens of barcoded expression libraries uncover new gene functions in carbon utilization among gut Bactereroidales

Yolanda Y. Huang[1], Morgan N. Price[1], Allison Hung[2], Omree Gal-Oz[1], Davian Ho[3], Héloïse Carion[3], Adam M. Deutschbauer[1,4], Adam P. Arkin[1,3*]

[1]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[2]Department of Molecular and Cell Biology, University of California-Berkeley, Berkeley, CA 94720, USA
[3]Department of Bioengineering, University of California-Berkeley, Berkeley, CA 94720, USA
[4]Department of Plant and Microbial Biology, University of California-Berkeley, Berkeley, CA 94720, USA

*Correspondence should be addressed to A.P.A. (aparkin@lbl.gov)

Abstract

mechanisticA understanding of host-microbe interactions in the gut microbiome is hindered by poorly annotated bacterial genomes. While functional genomics can generate large gene-to-phenotype datasets to accelerate gene discovery, their applications to study gut anaerobes have been limited. For instance, most gain-of-function screens of gut bacterial genes have been performed in an aerobic host and included a small number of conditions. To address these challenges, we developed a strategy to barcode expression libraries for high-throughput interrogation of gene functions in competitive fitness assays. We demonstrate the power of this approach to uncover novel phenotypes for uncharacterized genes using pooled libraries constructed from a diverse set of gut Bacteroidales expressed in *Bacteroides thetaiotaomicron*. We identified new roles in carbohydrate metabolism for nine proteins, including enzymes, transporters, a regulator, and hypothetical proteins from mobile genetic elements. This approach can be readily applied to other organisms and additional phenotypic assays.

Introduction

The past decade has seen massive sequencing initiatives and -omics studies on the human gut microbiome.[1–4] We now appreciate that there is tremendous complexity in the network of interactions between host and microbes, and within microbes in this ecosystem. Despite their importance for human health, we lack a detailed mechanistic understanding of microbial functions in the gut.[5,6] In particular, it remains difficult to infer function from sequencing data as bacterial genomes encode many genes of unknown function.[7,8] Accurately predicting gene functions often requires both a mechanistic understanding of the protein family and supporting phenotypic data, which can be difficult to obtain.

Functional genomic libraries can rapidly connect genotypes to phenotypes. Specifically, DNA barcoding of individual strains allows for parallel analysis of strain fitness in genome-wide libraries, through the use of barcode sequencing (BarSeq).[9] In recent years, randomly barcoded TnSeq (RB-TnSeq) mutant libraries have enabled high-throughput genetic screens across hundreds of conditions in a cost-effective manner and have accelerated functional assignments of genes in diverse bacteria, including the gut anaerobe *Bacteroides thetaiotaomicron* (*B. theta*).[10–12] Gain-of-function screens are complementary to loss-of-function screens and have the advantage of capturing genes from bacteria that have yet to be cultivated or genetically modified. However, most libraries expressing gut-derived DNA have relied on *Escherichia coli* as the heterologous expression host, which is phylogenetically divergent from members of the gut abundant phyla Bacillota (formerly Firmicutes) and Bacteroidota (formerly Bacteroidetes).[13–18] Promoter recognition of *Bacteroides* genes in *E. coli* can be limiting and *Bacteroides* encode unique ribosomal binding sites (RBSs), which can present barriers for heterologous gene expression in *E. coli*.[19–21] Therefore, Bacteroidales serve as more suitable expression platforms to study functions in the gut, yet few genetic screens have been performed in them.[22,23]

Another challenge in gain-of-function screens is the reliance on deep sequencing or isolation of individual strains to identify beneficial genes.[13–18,22] This is time-consuming and laborious to perform for hundreds of phenotypic conditions, which have limited functional screens to a small number of conditions.[13–18,22] Random DNA barcoding of libraries can drastically increase the scale of phenotypic screens, however, it is technically challenging to link both ends of each insert to a barcode. Dual barcoding of vectors (Dub-seq) was established as a solution and was used to construct an *E. coli* genomic library as proof-of-concept.[24] In this approach, unique barcode pairs are first mapped in the empty vector library and then used to link both ends of an insert fragment in the final genomic library. However, Dub-seq library is limited in size with ~250,000 unique barcode pairs, and the approach is restricted to DNA inserts with reference genome assemblies.

Here, we exploit developments in long-read sequencing to develop a new workflow termed Boba-seq, or barcoded overexpression bacterial shotgun expression library sequencing, to perform large-scale gain-of-function screens using uniquely-barcoded vector libraries. We demonstrate this with *B. theta* as an expression platform to study gene functions from seven Bacteriodales, although this approach is readily applicable to alternative expression strains and

source DNA from any organism of interest. Bacteroidales are major constituents of the gut microbiota that establish stable, long-term associations with the human host. They have important ecological roles in the gut, including metabolism of diverse, host-recalcitrant carbohydrates and modulation of the host immune system.[25] While Bacteroidales are mostly associated with health, some species have been isolated from abscesses and blood upon infiltration of the intestinal epithelial barrier.[25] In addition, this core group of gut bacteria have many cultured isolates and genetic tools available.

Our strategy provides key advantages over previously reported functional genomic libraries and the dual-barcoded *E. coli* Dub-seq library: 1) we constructed a barcoded empty library with high diversity (~14 million barcodes), which allowed multiple libraries to be pooled, 2) we assayed pooled libraries across many conditions using BarSeq, and 3) we apply our tools to a prevalent clade of gut anaerobes with important roles in human health. In addition, by using PacBio sequencing to map barcodes to entire inserts, we can map libraries in one step and use DNA from isolates without a reference genome assembly to generate metagenomic libraries

We constructed barcoded libraries from genomic DNA of seven Bacteroidales and first validated our method in a complementation screen. We combined six libraries, tested this in 53 conditions, and analyzed 41 conditions to enrich genes involved in carbon utilization and stress tolerance. From this data, we uncovered phenotypes for previously uncharacterized genes, including carbohydrate transporters, an α-glucosidase, and prophage hypothetical proteins. Homologous hits from multiple libraries in our parallelized screen provided high confidence in our findings. In addition to gene discovery, our work also expands the genetic toolbox for Bacteroides as expression platforms and provides a complementary approach to high-throughput screens of loss-of-function libraries.

Results

Overview of workflow to build and assay barcoded expression libraries

The Boba-seq workflow utilizes a single DNA barcode per plasmid to increase the throughput of phenotypic assays. To our knowledge, single barcoding of shotgun expression libraries in bacteria has not been previously reported for functional screens. An overview of this approach is summarized in Figure 1 and consists of five general steps.

First, we construct replicative vectors to encode an inducible promoter, a cloning insertion site for DNA of interest, a terminator, a barcoding site, and a replicon for the host strain of interest. Empty vectors are first barcoded by cloning in random 20 nucleotide barcodes using Golden Gate Assembly.[26] In the second step, randomly sheared and size-selected DNA fragments from the organism(s) of interest are cloned into the barcoded vectors. Here, the fragment sizes can be optimized to capture single genes or larger gene clusters that might encode the target phenotypes. We assess library diversity based on estimates from colony counts and then barcode amplicon sequencing, or BarSeq. Third, insert fragments are mapped to unique barcodes using long-read sequencing. By using single barcodes instead of barcode pairs to tag and characterize each vector, we can drastically increase library diversity.[24] We also simplify the

mapping step of linking each barcode to the cloned genomic fragment to a single PCR amplification compared to the more complex TnSeq-like protocol required to map dual-barcoded vectors.[24] In step four, mapped libraries are conjugated into the expression strain individually and conjugation efficiency is assessed using BarSeq. In the final step, libraries are screened across a panel of selective conditions where the inoculum (Time0) and samples after growth are harvested for BarSeq. Fragments that led to increased strain fitness are identified by calculating fitness scores for each barcode, which represents the log2 change in strain abundance. Fitness scores are filtered to identify potential causative genes for the observed fitness gain.

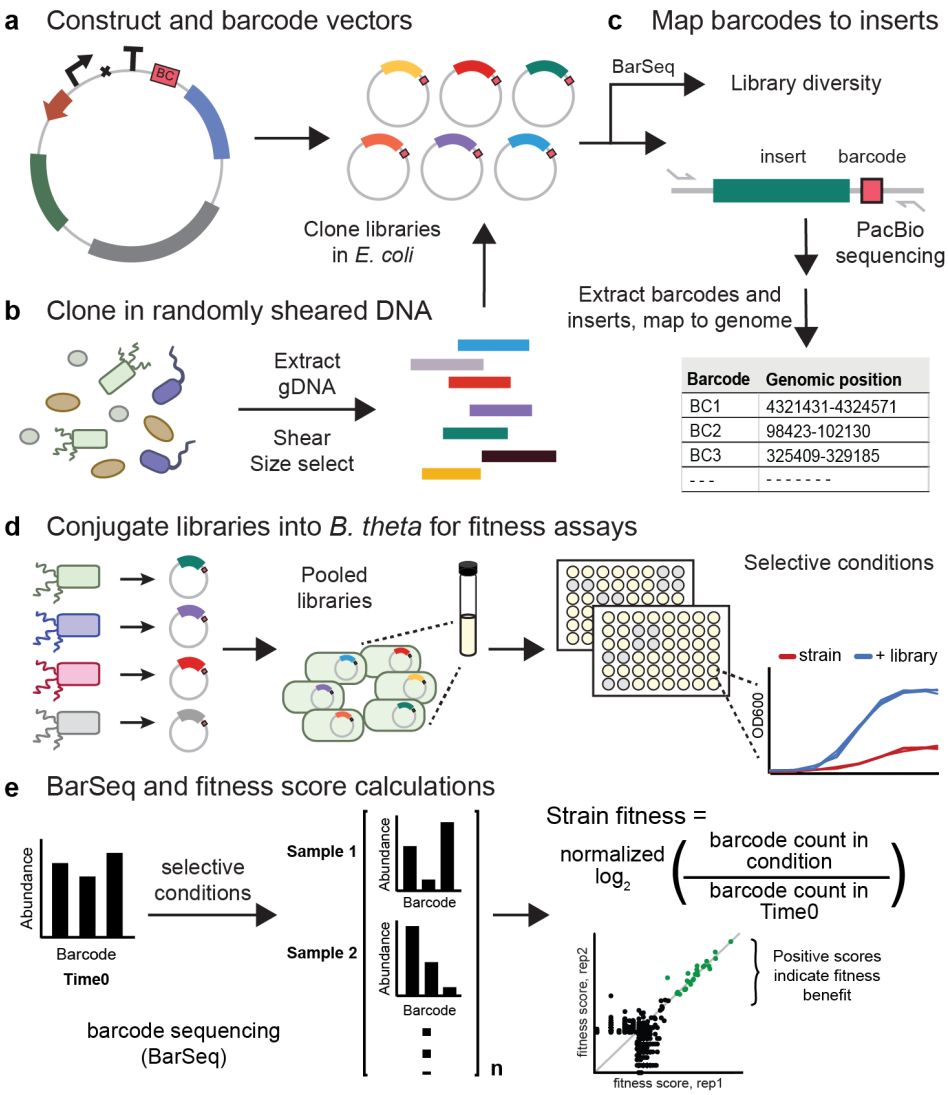

Figure 1. Overview of workflow to build barcoded expression libraries for large-scale phenotypic screens. A) Clone and barcode replicative vectors for expression strain of interest. B) Extract DNA, shear, size-select, and clone into barcoded vectors. C) Quantify library diversity using barcode sequencing (BarSeq). Map barcode to insert using long read sequencing technology and generate a barcode-to-insert mapping table. D) Conjugate mapped libraries into the expression strain for fitness screens. Libraries can be pooled to increase the genomic diversity assayed. A range of selective conditions can be used to determine fitness differences across

strains, such as carbon utilization and stress tolerance. E) Quantify library composition in the inoculum before and after growth using BarSeq. Fitness scores are calculated for each barcode or strain to identify genes potentially conferring a benefit in the condition.

Construction of barcoded expression libraries from Bacteroidales genomic DNA

B. theta VPI-5482 is the best characterized reference strain for gut Bacteroidales and the focus of genetic tool development. Therefore, we sought to develop and characterize replicative vectors by adding different replicons to an integrative vector for *B. theta*.[27] This led to three replicative vectors: pNBU2_repA1, pNBU2_repA2, and pNBU2_repA3 (Fig. 2a, Supplementary Table 1).[28–30] Our vectors contain an anhydrotetracycline (aTc) inducible promoter, a ribosomal binding site (RBS), and a terminator.[27,31,32] Inducible gene expression by aTc was verified using the NanoLuc reporter gene (Supplementary Fig. 1a). The plasmid copy number for these vectors in *B. theta* was determined to be around 2, 7, and 8, respectively, using droplet digital PCR (Supplementary Fig. 2). These replicons led to replication in 9 of the 12 additional Bacteroidales tested (Supplementary Fig. 3). pNBU2_repA1 was barcoded with approximately 14 million unique barcodes, based on estimates from BarSeq data (see Methods). We initially attempted to clone in randomly-sheared ~3 kb genomic fragments from Bacteroidales using blunt-end ligation, but this led to vector rearrangements in the *E. coli* cloning strain. Therefore, most libraries were constructed by cloning fragments into a holding vector followed by Gibson assembly of these inserts into barcoded pNBU2_repA1. The 3 kb fragment size was selected to capture at least one gene on most fragments. In total, seven libraries were generated using gDNA from *B. theta*, *B. caccae*, *B. fragilis*, *B. salyersiae*, *B. uniformis*, *P. johnsonnii*, and *P. merdae*, which share a maximum of 92% average nucleotide identity. Phylogenetic relationships between these seven strains are shown in Fig. 2b. Libraries consist of barcode diversities ranging from 40,911 to 97,471 (Table 1). 81% of all protein-coding genes in these genomes have at least one homolog (≥40% a.a. identity, ≥75% coverage) in the same or a different genome.

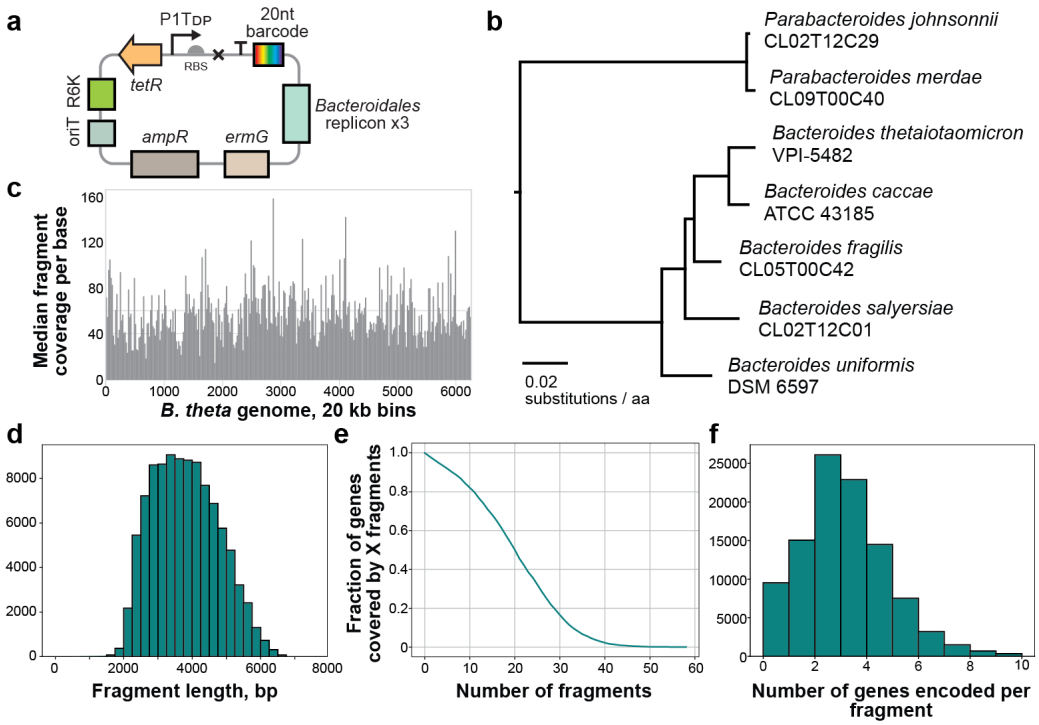

Figure 2. Seven vector libraries were constructed from genomic DNA of *Bacteroides* and *Parabacteroides* strains. A) Replicative vectors constructed by adding a replicon and a barcoding sites to replace the attachment sites on the pNBU2-ErmG-P1T_DP-GH023 integrative vector. Three replicative vectors were generated. B) 24 universal bacterial marker genes from the Genome Taxonomy Database (GTDB) were used to generate this phylogenetic tree.[33] Protein sequences of marker genes were aligned with HMMer 3.3.1 (http://hmmer.org), a tree was inferred with FastTree 2.1.11[34], and the tree was midpoint rooted. Plots for the *B. theta* genomic library were generated as part of the mapping script: C) fragment coverage per base across the genome plotted in 20 kb windows, D) histogram of fragment lengths, E) cumulative gene coverage plot, and F) histogram of number of full-length genes encoded per fragment.

Next, we developed a computational pipeline to connect barcodes to vector inserts using long-read sequencing (Supplementary Fig. 4). We performed PacBio sequencing on amplicons of the insert-barcode region to obtain ~4–23-fold read depths (Supplementary Table 2). 78–91% of barcodes detected by BarSeq were mapped, hence only a small fraction of each library is not interpretable. On a per genome basis, fragments cover 87–98% of protein-coding genes with 73–473 protein-coding genes unmapped (Table 1, Fig. 2c). In all libraries, the missing proteins are generally longer than mapped proteins (median 1,755 vs. 929 bp) and 5% have a nearly-identical copy (≥95% amino acid identity) elsewhere in the genome. Overall, 50% of the missing proteins are over 2 kb in length or are duplicated, which reflects limitations imposed by fragment sizes and difficulties in mapping highly similar regions. After excluding genes of over 1.5 kb or with duplicates, 31% of the remaining genes have a homolog in the same or another

library that is also missing (Supplementary Table 3). Conservation of missing proteins may indicate toxicity in *E. coli* that led to their absence.

Libraries range in median fragment lengths from 2.7–3.8 kb (Table 1, Fig. 2d, Supplementary Fig. 5a). The majority of fragments encode 1–2 complete genes and the median fragment coverage per gene is 12–24 (Table 1, Fig. 2e,f, Supplementary Fig. 5b,c). In summary, we constructed and mapped seven libraries with high genomic coverages.

Table 1. Summary of barcoded genomic libraries constructed in this study. Barcode diversity is reported for each library in *E. coli* EC100D using BarSeq data. The percentage of barcodes mapped to an insert fragment and a genomic location was computed using the PacBio data and the barcode diversity for each library in EC100D.

Genome	Library name	Barcode diversity	Barcodes mapped	Cloning method	# of genes	Genes cover-ed	Median inserts per gene	Median length (nt)
B. thetaiotaomicron VPI-5482	Btheta	70,515	78%	Blunt-end ligation	4,767	98%	20	3,751
B. caccae CL03T12C61	Bcaccae	56,770	91%	Gibson assembly	4,479	91%	15	3,015
B. fragilis CL05T00C42	Bfrag	58,885	85%	Gibson assembly	4,327	93%	20	3,358
B. salyersiae CL02T12C01	Bsa	48,016	89%	Gibson assembly	4,447	91%	12	2,851
B. uniformis ATCC 8492	Buni	53,772	91%	Gibson assembly	3,764	89%	13	2,708
P. johnsonii CL02T12C29	Pjohn	97,471	86%	Gibson assembly	3,671	87%	24	3,080
P. merdae CL09T00C40	Pmerdae	40,911	79%	Gibson assembly	3,570	89%	12	3,115

Complementation assays as a proof-of-concept

To validate these libraries in a phenotypic screen, we selected two *B. theta* deletion mutants as expression strains for complementation assays.[11] BT2158 (BT_RS10925) encodes a periplasmic glucose:fructose oxidoreductase required for disaccharide and glucosinolate metabolism.[11,35] ΔBT2158 does not grow on trehalose (glucose-α-1,1-glucose), leucrose (glucose-α-1,5-fructose), or palatinose (glucose-α-1,6-fructose) (Supplementary Fig. 6). The second mutant lacks α-glucosidase SusB (BT3703 or BT_RS18670, part of the starch utilization system), and has reduced growth on α-cyclodextrin.[11] We conjugated *B. theta, B. fragilis, B. uniformis*, and *P. johnsonnii* libraries individually into ΔBT2158 and ΔBT3703 mutants. We then

inoculated each library in a defined minimal medium (Varel-Bryant) containing a single carbon substrate in the presence of the inducer aTc (Supplementary Table 4).[36] Of these four source genomes, three encode BT2158 or a close homolog, and three encode BT3703 or a homolog (>70% a.a. identity). We expected to observe the enrichment of BT2158 or a homolog on trehalose, leucrose, and palatinose, and the enrichment of BT3703 or a similar protein on α-cyclodextrin.

We computed fitness scores for each fragment using the log2 fold-change of each barcode's relative abundance between the selective condition and the Time0. We focus on high positive scores, since this is a gain-of-function assay, and then filter for biologically consistent hits. First, fragments with a fitness score ≥5 and a z score ≥4 in both replicates are considered statistically significant. Using these thresholds, no significant hits were found in a control comparison of Time0 samples (see Methods). Second, we consider each insert hit to be biologically consistent if 1) there are overlapping fragments with significant high scores or 2) there is a protein homolog (≥40% a.a. identity) from a different genomic region or library with a significant fitness score. This led to 21 biologically consistent region by condition pairs. For each region, we assess which gene or genes are likely responsible for the observed fitness gain by computing the average fitness score from all fragments with complete coverage of each gene. The gene with the highest fitness value of each region is selected as a hit (Supplementary Table 5). Using this method, we detected 12 genes that conferred a fitness gain in leucrose, palatinose, or α-cyclodextrin.

For libraries expressed in ΔBT2158, leucrose and palatinose conditions led to expected results where high fitness scores were obtained for inserts encoding BT2158 (BT_RS10925), its nearly-identical paralog BT4448 (BT_RS22435), or a homolog (BACUNI_RS05440) from *B. uniformis* (Fig. 3A). In trehalose, we observed extremely high fitness scores (>14) for three strains encoding each of these genes, but these strains have a significant score in just one replicate and therefore did not meet our criteria for statistical significance. It is likely that these strains grew quickly early on and outcompeted other strains with a beneficial gene. In addition to the expected hits, a TonB-dependent transporter from *B. fragilis* (HMPREF1079_RS07310) was enriched in leucrose and may be involved in leucrose uptake (Supplementary Table 5).

To verify our results, we also plated out strains from the above assays and identified clones encoding either BT2158, BT4448, or the *B. uniformis* homolog using Sanger sequencing. We selected six strains that collectively cover all three genes and cultured them individually to demonstrate improved growth in trehalose compared to the parental strain (Supplementary Table 6*,* Supplementary Fig. 7). Of note, three of the six fragments encode genes oriented opposite to the inducible promoter with <50 bp upstream, which is unlikely to encode a promoter (Supplementary Table 6). To account for this observation, we verified there is weak, although unexpected, promoter activity on the replicon fragment of the pNBU2_repA1 vector located downstream of the insert site (Supplementary Fig. 1b).

Our data show that BT4448 can functionally complement ΔBT2158 when expressed from a vector, which contrasts with the observation that the ΔBT2158 mutant cannot grow on these disaccharides and thus that BT2158 is required for metabolism.[11] It is probable that BT4448 is

not expressed at sufficient levels under our laboratory conditions in *B. theta* to support ΔBT2158 mutant growth and that expression from the vector enabled growth rescue.

In contrast to the ΔBT2158 libraries, growth of complemented ΔBT3703 libraries in α-cyclodextrin did not enrich for the deleted α-glucosidase nor any homolog but revealed cysteine desulfidase (CdsB) and cysteine β-lyase from all four libraries to be beneficial (Fig. 3b, Supplementary Table 5). Here, the selection for cysteine degrading enzymes appears to be stronger than for BT3703 and homologs, likely because ΔBT3703 is capable of growth on α-cyclodextrin.[11] Both enzymes can catalyze the degradation of cysteine to sulfide, ammonium, and pyruvate, and were likely beneficial as a detoxification response to the high cysteine concentration (8.25 mM) in our minimal medium.[37,38] This is also supported by RB-TnSeq data for *B. theta*, where disruption of CdsB (BT2080 or BT_RS10540) led to reduced fitness across all minimal media conditions (Fig. 3c).[11] Similarly, cysteine β-lyase containing strains in the *B. theta* RB-TnSeq library exhibited mild fitness defects in many conditions.[11] Of note, the rich medium used for RB-TnSeq fitness assays was also supplemented with 8.25 mM cysteine, and thus this effect is specific to the Varel-Bryant minimal medium.[11] Cysteine is commonly added to Bacteroidales growth media as a reducing agent, but our result cautions its use at a high concentration in minimal media.

In summary, complementation worked well for conditions with strong selective pressure where the expression strain has a strong fitness defect. Multiple fragments encoding BT2158 or a homolog were strongly selected when expressed in the ΔBT2158 mutant grown on leucrose and palatinose.

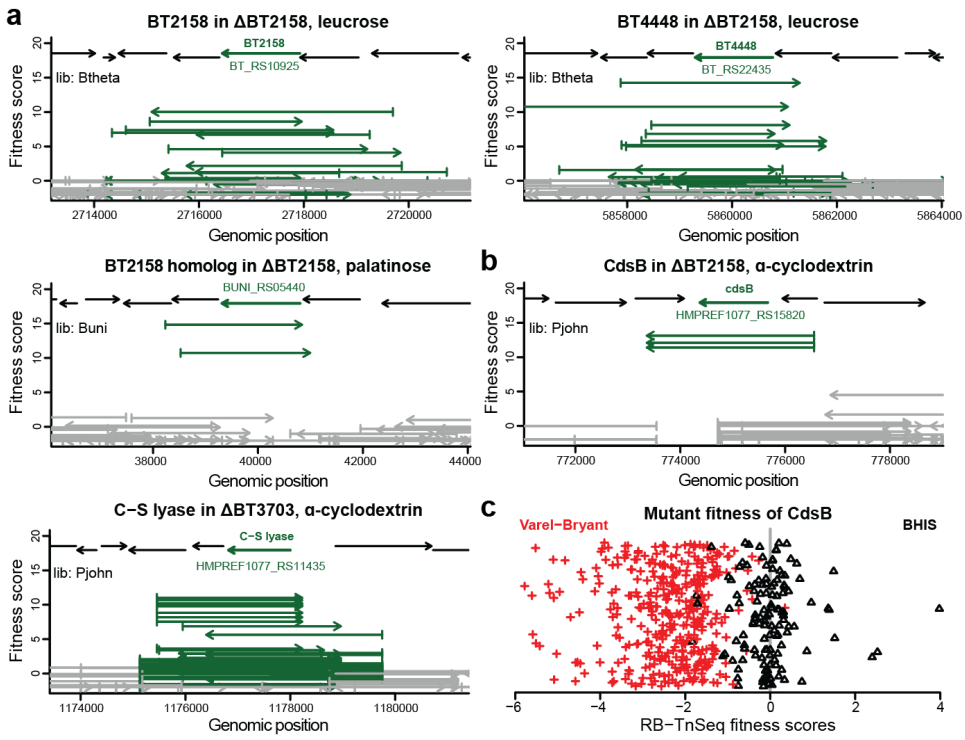

Figure 3. Complementation assays led to enrichment of BT2158 and homologs as well as cysteine degrading enzymes. The average fitness score from both replicates is plotted for each fragment. Fragments that completely cover the gene of interest are indicated in green. A) BT2158 and BT4448 in *B. theta*, and a homolog in *B. uniformis* were beneficial in the ΔBT2158 mutant upon growth in leucrose and palatinose. B) Enrichment of CdsB and cysteine β-lyase in α-cyclodextrin. C) RB-TnSeq fitness values for CdsB (BT2080 or BT_RS10540) are averaged log2 changes in abundance of 37 *B. theta* transposon mutants (x-axis). The y-axis is random for visualization.

Fitness profiling of pooled libraries using BarSeq

After establishing a pipeline for fitness data analysis, we sought to map genotypes to phenotypes in conditions relevant for the gut habitat. We tested 53 conditions in the presence of the inducer aTc consisting of 41 single carbon sources in a defined minimal medium and 5 stressor compounds at 2 to 3 different concentrations in a rich medium (Supplementary Table 4). Carbon sources include dietary and host-derived complex polysaccharides, gelatin, mucins, and simple carbohydrates. Inhibitory compounds include antibiotics and bile salts. Six libraries (excluding the *B. theta* library) were conjugated into *B. theta* and pooled into a single inoculum (or Time0). 0.5% of barcodes were redundant across libraries and were excluded from analysis. Growth was observed among 29 carbon sources and all stress conditions. 84 samples were processed for BarSeq and included two replicates per condition and two Time0 samples.

Of the 303,011 uniquely-mapped barcodes that are present in the six conjugation donor libraries, we detected and computed fitness scores for 260,862 barcodes in our fitness assays. This includes 111,534 distinct fragments that cover 20,744 protein-coding genes. We identified 384 barcodes with statistically significant scores in both replicates of a condition, using the same criteria as previously detailed for complementation assays (Fig. 4a). We identified biologically-consistent fragments for 48 region by condition pairs, including 15 conditions and 26 unique regions (Supplementary Table 5). The likely causative protein for each region's benefit was identified by inspecting fitness score vs. fragment coverage plots (Supplementary Table 5). For example, a few fragments with high fitness scores in raffinose cover both the RafT3 gene (HMPREF1079_RS07310) and the gene upstream (Fig. 4b). However, high fitness scores are also observed in many shorter fragments that only cover RafT3 and not the adjacent gene. Therefore, we selected RafT3 as the likely causative protein for the observed benefit in raffinose. Overall, clustering at 40% amino acid identity threshold led to 13 protein functional groups across all conditions (Supplementary Table 5).

We noticed a high level of false negatives in our data where some strains that mapped to a biologically-consistent hit did not confer a benefit. Only 41% of strains that contain the putative causative protein show a significant benefit (fitness ≥5 and z ≥4) in at least one of the two replicates. For example, 12 strains have fragments that completely cover RafT3, but only a subset of these, 8 strains, have a statistically significant fitness score in raffinose. We speculate that this is partly due to variations in gene expression, which is dependent on the exact boundaries of the fragment. In support of this, for strains with the exact same insert where at

least one strain has a statistically significant score (in both replicates), 80% have a statistically significant score in at least one replicate.

Among strains with fitness scores, 39% encode a gene oriented in the same direction as the aTc inducible promoter. This bias is likely caused by gene expression from the holding vector during cloning in *E. coli* that contributed to gene toxicity or lower fitness (see Methods). Among strains with a biologically-consistent benefit, this value is significantly higher at 48% (P=0.001, chi-squared test of proportions), indicating possible contributions to gene expression from the inducible promoter.

In stress conditions, only trimethoprim at the highest concentration tested yielded hits (Supplementary Table 5). As expected, overexpressing the target of this inhibitor, dihydrofolate reductase, from 5 of 6 libraries was beneficial. The *B. uniformis* library does not have an insert covering this gene, accounting for its absence from our results. We expect higher concentrations of inhibitory compounds would uncover a greater number of hits. Consistent with the complementation screen, we also detected high fitness scores for CdsB and cysteine β-lyase across 9 conditions (Supplementary Table 5).

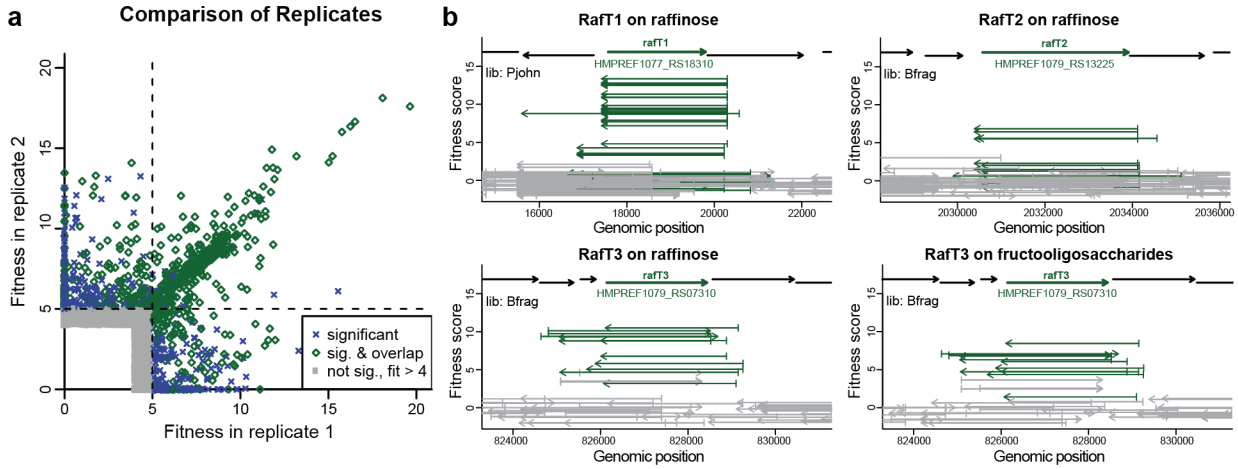

Figure 4. Fitness scores are consistent across replicates in the fitness screen. TonB-dependent transporters were enriched in carbohydrate conditions. A) Consistency of strain fitness values between replicate experiments and for 41 conditions. Strains are only included if they had a fitness value of 4 or higher in at least one replicate. Dashed lines separate significant fitness scores with values ≥5. Barcodes that have significant fitness scores and cover regions with overlapping fragments are highlighted in green. B) Genomic plots of average fitness scores per fragment for three TonB-dependent receptors in raffinose and fructooligosaccharides (FOS).

High-confidence hits uncover metabolic enzymes in carbohydrate utilization

Here, we focus on metabolic enzyme hits uncovered on single carbon sources. Three substrates enriched for genes expected to confer a benefit based on known functions or annotations (Supplementary Table 5). These include fuculose kinase and fucose permease in

fucose, and β-glucosidase in D-cellobiose and D-salicin, which are both β-glucosides (Supplementary Fig. 8a,c).[11,39] We verified that the β-glucosidase gene conferred a growth benefit in individual strains in D-cellobiose and D-salicin (Supplementary Fig. 9a,b).

Additionally, an α-glucosidase (INE72_RS06280) in the glycosyl hydrolase family 31 (GH31) was beneficial in palatinose and has sequence similarity to an α-1,3-glucosidase (36% a.a. identity, Cphy_1877, *Lactococcus lactis*) (Supplementary Fig. 8b).[40] Since palatinose contains an α-1,6-glycosidic bond, we propose this enzyme has α-1,6-glucosidase activity.

A ROK family (repressor, open reading frame, kinase) protein and a hexokinase were both important for growth on D-glucosamine (Supplementary Fig. 8d). We verified that individual strains encoding the ROK kinase displayed a growth advantage in D-glucosamine (Supplementary Fig. 9c). Canonical N-acetyl-D-glucosamine (GlcNAc) metabolism involves phosphorylation and then de-acetylation to D-glucosamine-6-phosphate.[41] A *B. fragilis* homolog of the ROK protein was shown to phosphorylate several substrates *in vitro*, including GlcNAc and mannose.[42] Unfortunately, D-glucosamine was not tested as a substrate for this ROK kinase.[42] Encouragingly, RB-TnSeq mutants of an ROK homolog in *Pedobacter* sp. (CA265_RS11300, 30% a.a. identity) exhibited reduced fitness on D-glucosamine and GlcNAc.[10] Oddly, no high-confidence hits were observed in GlcNAc in our screen, which may be due to differences in gene regulation and/or selective pressure. A homolog of the hexokinase hit (AY664812, 79% a.a. identity) was shown to contribute to growth in glucose and mannose in a *B. fragilis* deletion mutant.[42] Hexokinases have broad substrate scopes and can likely phosphorylate additional substrates, including D-glucosamine.[43] Alternatively, D-glucosamine metabolism could involve acetylation to GlcNAc by a glucosamine acetyltransferase (potentially BT3619 or BT3180) and then phosphorylation by these kinases.[44] Overall, further experiments are needed to determine whether these two kinases can directly phosphorylate D-glucosamine.

TonB-dependent transporters are beneficial for growth on raffinose and fructooligosaccharides

Three TonB-dependent transporters (TBDTs) were found to be important in D-raffinose and fructooligosaccharides (FOS) (Fig. 4b). This large transporter family can carry out active transport of vitamins, siderophores, peptides, and glycans across the outer membrane, and are prevalent in the phylum Bacteroidota, with 121 putative TBDTs in *B. theta* alone.[45] To our knowledge, there is no reported system for raffinose uptake in Bacteroidales. A distantly-related TBDT from *B. theta* (BT1763 or BT_RS08935, ≤28% a.a. identity to our hits) is involved in the uptake of FOS with β-2,6 linkages, but the FOS substrate used in our study contains predominantly β-2,1-linkages (Supplementary Table 4).[46]

Specifically, our data identified two putative raffinose transporters, which we call *Pj*-RafT1 and *Bf*-RafT2, from *P. johnsonnii (*HMPREF1077_RS18310) and *B. fragilis* (HMPREF1079_RS13225), respectively (Supplementary Table 5). In addition, we detected a transporter likely to uptake both raffinose and FOS, which we denote as *Bf*-RafT3 (HMPREF1079_RS07310 from *B. fragilis*). Notably, *Bf*-RafT3 was a hit in the complementation screen for leucrose as well (Supplementary Table 5). Leucrose, raffinose, and FOS all contain

D-fructose as a terminal residue, which might mediate substrate recognition by *Bf*-RafT3. *Pj*-RafT1 and *Bf*-RafT3 are similar (49% a.a. identity), but have distinct gene neighborhoods. An α-galactosidase is present downstream of *Pj*-RafT1 and can likely release galactose from raffinose.[47] A β-fructofuranosidase is genomically adjacent to *Bf*-RafT3, which is important for raffinose metabolism according to RB-TnSeq data for *Phocaeicola vulgatus* CL09T03C04 (unpublished data). Both nearby enzymes are predicted to be periplasmic, which supports a direct uptake of raffinose by the transporters. This is also supported by RB-TnSeq data for a TBDT homolog in *P. vulgatus* (HMPREF1058_RS00615, 47% a.a. identity) that is important for raffinose utilization (unpublished data). Collectively, the combination of data from gain-of-function libraries, transposon mutant libraries, and gene neighborhoods support raffinose as a substrate for these transporters.

Most *Bacteroides* TBDTs are SusC-like and co-localize with a partner lipoprotein-encoding *susD*. SusDs form complexes with SusCs and are thought to be essential for substrate capture and transport by SusC-like TBDTs.[45,48] Strikingly, only *Bf*-RafT2 among our hits is co-localized with a *susD*, however, this partner SusD is not required for the fitness benefit observed in raffinose (Fig. 4b). It is possible that *Bf*-RafT2 retains activity alone or that a *B. theta* SusD can complement the endogenous SusD activity. Similar to our observation, a *B. fragilis* SusC-like sialic acid transporter (NanO) maintained activity in *E. coli* in the absence of its SusD partner.[49] *Bf*-RafT2 belongs to the SusC/RagA family and contains N-terminal extension (NTE), plug, and β-barrel domains, which is the canonical domain structure for *Bacteroides* TBDTs (Supplementary Table 5).[45,50] In contrast, *Pj*-RafT1 and *Bf*-RafT3 lack an adjacent *susD* gene and are likely not SusD-dependent (Supplementary Table 5). The lack of a nearby *susD* for two of the three TBDT hits may be a result of shorter fragment lengths in our libraries. This likely biased hits toward TBDTs that do not require a partner SusD for activity as coverage of both genes would require fragments longer than ~5 kb.

In summary, we uncovered three novel TBDTs involved in raffinose and FOS uptake in our fitness assays. One hit is co-localized with a *susD*, which was not required for the observed fitness gain.

A regulator confers fitness in carbohydrate metabolism

In addition to gene hits involved in metabolic and transport pathways, our results point to regulatory mechanisms as well. An AraC family regulatory protein (HMPREF1079_RS07695), previously associated with increased resistance to DNA damaging agents, conferred a fitness benefit on heparin from porcine intestinal mucosa (Fig. 5a).[51,52] A role for this regulator on heparin or more broadly, carbohydrate utilization, had not been reported.

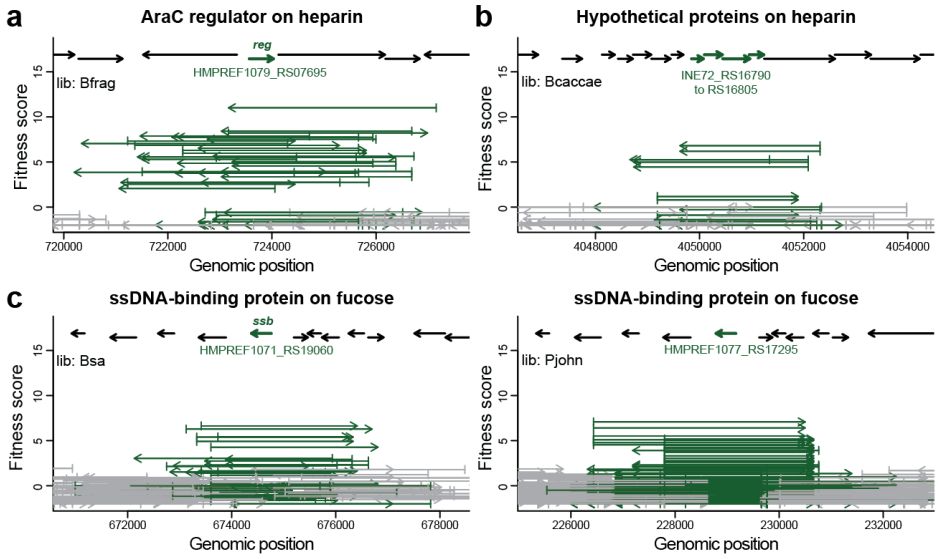

Figure 5. Unexpected gene hits that conferred a growth advantage on heparin and fucose. These include A) an AraC family regulator, B) a cluster of prophage hypothetical proteins, and C) an ssDNA-binding protein.

Mobile genetic elements contribute to competitive growth on fucose and heparin

In fucose, an uncharacterized protein containing an ssDNA-binding domain (*ssb*, HMPREF1077_RS17295 and HMPREF1071_RS19060) was detected from both *P. johnsonnii* and *B. salyersiae* libraries (Supplementary Table 5). The beneficial fragments also encode three annotated hypothetical proteins upstream of ssb and therefore, it is possible that multiple proteins or one of the hypothetical proteins conferred a fitness gain. The *ssb* appears to be part of a ~50 kb mobile genetic element present in several Bacteroidales, with a putative recombinase and resolvase encoded downstream (Supplementary Fig. 10a). The enrichment of this hit was unexpected and presents a novel link between this gene and a phenotype.

In heparin, a 1.4 kb region encoding four small hypothetical proteins was highly enriched from *B. caccae*. Three of the four genes encode transmembrane domains with one encoding signal peptide II as well. A recently identified surface export signal in Bacteroidota appears to be absent in this lipoprotein, therefore it may be localized in the periplasm, although we cannot exclude the possibility of an unknown export signal.[53] Intriguingly, these genes are part of a putative prophage (~45 kb), which is also present in the *B. salyersiae* strain (Supplementary Fig. 10b). A tBLASTn search of the lipoprotein against the NCBI database points to a narrow distribution among *Bacteroides* prophages and phage metagenome-assembled genomes (MAGs) (Supplementary Table 7).

Discussion

We present Boba-seq as a new strategy to construct barcoded expression libraries for scalable and quantitative phenotypic screens, using an important clade of gut bacteria as a case study.

To demonstrate the application of this workflow, we generated seven libraries using genomic DNA from *Bacteroides* and *Parabacteroides* species. First, we validated our approach in a complementation screen using two *B. theta* deletion mutants as expression strains. Here, we observed that high selection pressure was important to enrich fragments encoding genes that complement the missing function. Enrichment of cysteine metabolic enzymes was also observed as a response to cysteine toxicity in our minimal medium. To identify novel phenotypes for genes, we multiplexed 6 libraries and screened across 53 conditions that span single carbon sources and inhibitory compounds. In total, we detected high-confidence fitness scores for 26 regions in 15 conditions, including hits for genes expected to confer a fitness benefit (Supplementary Table 5, Supplementary Fig. 8). Nine hits had no prior association with the observed phenotype (α-glucosidase, ROK kinase, hexokinase, TonB-dependent receptors, AraC regulator, ssDNA-binding protein, hypothetical proteins) (Supplementary Table 5). From our data, we were able to assign putative functions for an α-glucosidase and three TonB-dependent receptors.

We demonstrate the power of using complementary loss-of-function data from RB-TnSeq libraries to strengthen gene-to-phenotype mappings identified from our work. Here, five homologs from RB-TnSeq libraries (*B. theta, P. vulgatus*) have fitness phenotypes consistent with our gain-of-function data.[11] The scalability of both loss-of-function and gain-of-function libraries using barcoding techniques for Bacteroidales is an exciting development. By constructing RB-TnSeq and Boba-seq libraries for a large number of Bacteroidales and screening across identical conditions, we can rapidly generate specific functional hypotheses for novel genes. This work expands the genetic capabilities for gain-of-function screens in gut Bacteroidales and is critical for realizing this goal.

Our workflow uses single barcoding, which offers key advantages over the previously reported dual-barcoded libraries in *E. coli*.[24] Our protocol circumvents the step required to define all barcode pairs in the empty vector library. We generated a highly diverse empty library that is >50-fold greater in barcodes than the empty Dub-seq library (14 million vs. 250,000).[24] This enabled us to multiplex 6 genomic libraries to a total diversity of ~300,000 strains for fitness assays, drastically increasing the throughput of our genetic screens. Notably, the high barcode diversity will allow us to expand beyond lab isolates and capture highly diverse genetic materials from metagenomic samples. Here, we determined that ~50,000 fragments are sufficient to cover a bacterial genome with a median fragment coverage of >10 and that a sequencing depth of ~10× can map most of the barcodes (Supplementary Table 2).

There are a few variables that can be further optimized in our workflow. For example, our current libraries have fragment sizes of ~3.1 kb, which do not capture pathways that require several proteins. Fragment sizes can be increased to cover functions that require multiple genes such as polysaccharide utilization loci (PULs). Although we tried to prioritize carbon sources where *B. theta* does not grow or grow well on, better curation of such conditions should increase the success rate of obtaining high fitness scores. Finally, we found cysteine to be inhibitory for *B. theta* in a minimal medium, which is a common component of Bacteroidales media (Supplementary Fig. 11). Therefore, an optimized defined minimal medium should uncover additional gene hits that are not cysteine related.

We constructed our libraries with sufficient genomic coverage with overlapping fragments and also multiplexed libraries to obtain homologous hits, both of which increased confidence in our results. On average, 41% of strains encoding a beneficial gene exhibited statistically significant high fitness scores with the remaining strains representing false negatives in our data. We expect this is due to differences in gene expression, selection pressure, and/or competition between strains. For example, some fragments do not encode a native promoter. The inducible promoter on our vector was designed to reduce expression issues. We did observe an increase of ~9% in genes on the same strand as the promoter among beneficial fragments compared to all fragments. However, this effect is weak and we suspect that most genes are transcribed from their native promoters. In comparison, fitness data from the *E. coli* Dub-seq library had low rates of false negatives.[24] This highlights the gap in genetic tool advancements in the best characterized model bacterium and gut Bacteroidales, and encourages further development of vectors and genetic parts in *B. theta.* Finally, it can be difficult to predict functions even with strong phenotypic data, especially for hypothetical proteins. However, experimental evidence obtained from genetic screens is an important first step toward functional elucidation.

This approach can be readily used to scale functional screens to a larger number of libraries and assay conditions. Libraries can be built using different DNA sources, such as metagenomic samples and viral-like particle fractions. In addition, at least nine additional Bacteroidales strains can serve as expression strains using our vectors. This provides alternative genetic backgrounds to *B. theta*, which may be crucial when screening for specific functions and should uncover additional genes with new phenotypes. Deletion or individual transposon mutants can also be used as expression strains for high-throughput complementation assays.[54] In addition to the conditions reported here, our libraries can be used for other assays, including host colonization, phage infections, toxin suppression, and adhesion.[55–59]

With rapid generation of sequencing data, functional characterization of the growing genomic diversity lags behind. Here, we developed tools for scalable gain-of-function screens in a prominent clade of gut bacteria. With decreasing sequencing costs, we expect quantitative phenotypic screens will uncover many novel gene-to-phenotype associations to accelerate gene discovery in the gut microbiota. This has important applications toward identifying microbial factors that drive diseases, predictive modeling of microbiomes, and developing strategies to manipulate the gut microbiota for health applications.

Methods

Strains and cultivation conditions

All strains, plasmids, and primers used in this study are listed in Supplementary Tables 4 and 8. *E. coli* strains were routinely cultured in Luria-Bertani Lennox (LB) with the appropriate selection using 50 µg/mL carbenicillin (Carb50) or 50 µg/mL kanamycin (Kan50) at 37 °C with shaking at 200 rpm. *E. coli* WM3064 was grown with 300 µM diaminopimelic acid (DAP) supplement. Species within the order Bacteroidales were grown at 37 °C in an anaerobic chamber (Coy) maintained with a gas mix of 10% H_2, 5% CO_2, 85% N_2. Routine culturing was carried out in supplemented Brain Heart Infusion (BHIS) where BD Bacto Brain Heart Infusion was freshly

supplemented with 0.5 mg/L hemin and 2.5 μL/L vitamin K_1. The Varel-Bryant (VB) defined medium was used for all complementation and fitness assays as well as validation growth curves.[36] Media recipes and commercial sources of substrates are provided in Supplementary Table 4. Media were made anaerobic by pre-reducing in the anaerobic chamber at least one day prior to use. Conjugation mixes were selected on 200 μg/mL gentamicin (Gm200) and 5 μg/mL erythromycin (Erm5) to obtain the correct mutant.

Construction of replicative vectors for *B. theta*

See Supplementary Table 8 for all primers and genetic parts used for cloning. The pNBU2_erm integrative vector was modified to generate replicative vectors. This vector contains a previously reported anhydrotetracycline (aTc) inducible promoter and RBS combination (P1T_DP-GH023).[27] A PmlI restriction site, the synthetic terminator L3S2P21, and BsmBI sites used for barcoding were inserted downstream of the inducible promoter-RBS cassette.[26,31,32] First, a BsmBI restriction site was removed from the pNBU2_erm-TetR-P1T_DP-GH023 vector using site-directed mutagenesis. The pNBU2 vector with no BsmBI site was linearized by PCR to exclude the attachment sites. Fragments containing replicons from pTIO-1, pKSH24, and pFD340 were PCR amplified and purified using GFX DNA and Gel Band Purification kit. The third gene block consisted of an insertion site, a terminator, and a barcoding site.[10,31,32] Three-part Gibson Assembly reactions were performed to generate pNBU2_repA1, pNBU2_repA2, and pNBU2_repA3 containing replicon fragments from pTIO-1, pKSH24, and pFD340, respectively. Gibson reactions were carried out according to manufacturer's recommendations (NEB, Gibson Assembly master mix). This was transformed into electrocompetent *E. coli* EC100D *pir+* and plated onto LB-Carb50 to obtain the correct construct. All vectors were confirmed by Sanger sequencing.

Detection of inducible expression using Luciferase assays

To quantify gene expression on the three pNBU2_repA vectors, luciferase assays were performed as previously reported, using Promega's Nano-Glo Luciferase Assay System Kit.[27] Briefly, the NanoLuc gene was inserted downstream of the aTc-inducible promoter via Gibson assembly. Sequence-verified vectors were conjugated into *B. theta*. These *B. theta* strains were inoculated into BHIS-Erm5 and grown anaerobically at 37 °C for 16–18 h. Starter cultures were subcultured into BHIS-Erm5 and varying concentrations of aTc at a 1:100 dilution to mid-exponential phase (OD of 0.4-0.6, 3–5 hours). 500 μL of each culture were then pelleted by centrifugation at 5,000 × *g*, and cells were lysed with 50 μL of 1× BugBuster Protein Extraction Reagent (Sigma). Lysates clarified at 21,130 × *g* for 5 minutes, and 10 μL of the supernatant was mixed with an equal volume of the NanoLuc reaction buffer containing the NanoLuc substrate. Luminescence was immediately measured using a BioTek Synergy H1 microplate reader (Agilent), and normalized to OD600 of the culture at time of lysis.

Determination of plasmid copy number using droplet digital PCR

Droplet digital PCR (ddPCR) was performed following a previously reported procedure to estimate the plasmid copy number of the pNBU2_repA vectors in *B. theta*.[60] Briefly, *B. theta* containing each vector was grown in BHIS-Erm5 for 16–18 h at 37 °C, followed by a 1:100

dilution and subculture to mid-exponential phase (OD of 0.7–0.9, 3–5 h). 500 µL of each culture were then harvested for DNA extraction using Qiagen's DNeasy blood & tissue kit, after which 1 µg of DNA was digested using NdeI (NEB). ddPCR was performed using approximately 0.01, 0.1, or 1 ng of the digested DNA as template for each 40 µL reaction. PCR was performed using the Evagreen supermix and primers p-AH204-p-AH207 (Supplementary table 8), following BioRad's protocols for multiplexing with Evagreen and using an annealing temperature of 60 °C. Samples were analyzed using Bio-Rad's QX200 droplet digital PCR system. The resulting data points were manually gated, and reported concentrations for each amplicon were used in determining the plasmid copy number.

Barcoding of the pNBU2_repA1 vector

Barcodes were amplified from synthesized DNA using Phusion polymerase (NEB) at an annealing temperature of 58 °C, elongation time of 60 s, and 6 cycles.[9] pNBU2_repA1 vector was barcoded with random 20 nucleotides using Golden Gate Assembly. Each 20 µL reaction consisted of 1 µg of plasmid, 35 ng of barcode amplicons, 2 µL of T4 DNA ligase buffer, 1 µL of T4 DNA ligase (NEB), and 1 µL of BsmBI/Esp3I enzyme (Thermo). This resulted in a 2:1 molar ratio of insert to vector. Six reactions were incubated at 37 °C for 5 min and 16 °C for 10 min, which were repeated 10 cycles, followed by 37 °C for 30 min. All reactions were pooled, cleaned across two columns from the DNA clean and concentrator kit (Zymo), and eluted twice with 10 µL H_2O. Uncut vectors were removed with additional incubations with BsmBI/Esp3I in 25 µL reactions according to the manufacturer's protocol. Reactions were incubated at 37 °C for 2 h, after which 1 µL of BsmBI/Esp3I was added for incubation at 37 °C for 16 h. Each reaction was cleaned using DNA clean and concentrator kit (Zymo) and eluted twice with 8 µL H_2O. 7 µL of cleaned reaction was transformed into 50 µL of electrocompetent *E. coli* WM3064 for a total of 4 transformations. All aliquots were recovered in 7 mL S.O.C. medium with 300 µM DAP for 1.5 h. An aliquot was removed for CFU counting and confirmed high transformation efficiency of 10^{6-7} CFUs per mL. Recovered cells were inoculated into 1 L of LB (50 µg/mL Carb, 300 µM DAP) for selection. Each vector library was purified for BarSeq to determine barcode diversity.

Construction and quantification of barcoded genomic libraries

gDNA was purified from Bacteroidales using a DNeasy UltraClean Microbial Kit or DNeasy Blood & Tissue Kit (Qiagen). Up to 2 µg in 200 µL TE buffer was sheared by ultrasonication to an average of 3–5 kb with a Covaris S220 focused ultrasonicator using red miniTUBEs. Sheared gDNA was then gel-purified to size-selected by cutting out the 3–6 kb range. 4 µg of barcoded pNBU2_repA1 was digested in 50 µL reactions with 4.5 µL of PmlI. This was incubated at 37 °C for 30 minutes. Linearized vectors were dephosphorylated with 1 µL of rSAP (NEB) at 37 °C for 10 min and then deactivated at 65 °C for 5 min. Digested vectors were gel-purified using the Monarch DNA Gel extraction kit.

The *B. theta* library was constructed using blunt-end ligation and steps similar to the previously reported Dub-seq library.[24] Sheared and size-selected gDNA from *B. theta* was end-repaired using End-IT DNA End-Repair kit (Lucigen) in a 50 µL reaction consisting of 2 µL End-It Enz mix, 5 µL dNTP mix, 5 µL 10 mM ATP, 5 µL buffer, and 4 µg gDNA. This was incubated at room temperature for 45 min and deactivated at 70 °C for 20 min. This was gel-purified and

size-selected for 3–5 kb fragments again, and quantified using Qubit dsDNA BR assay (Invitrogen). The end-repaired, sheared DNA fragments were then ligated to linearized and dephosphorylated barcoded vectors. A 20 µL ligation reaction was performed with 10 µL Blunt/TA ligase master mix (NEB), 100 ng of vector backbone, and 4-fold molar excess of end-repaired DNA fragments, and incubated at room temperature for 20 min. Reaction was purified using Monarch PCR & DNA cleanup kit and eluted twice with 10 µL H_2O. 20 ng of cleaned ligation reaction was transformed into 50 µL aliquots of commercial *E.coli* EC100D that was diluted 4-fold with 10% glycerol. A total of 8 transformations were performed and each recovered in 1 mL of S.O.C. medium at 37 °C for 1.5 h with shaking at 200 rpm. An aliquot was removed for CFU counting to estimate library diversity. Recovered cultures were combined and inoculated into 500 mL of LB-Carb50 for selection. Vector library was purified for BarSeq to obtain an estimated diversity. The initial *B. theta* library consists of ~440,000 barcodes and was further diluted to obtain the final library with ~70,000 barcodes.

Six libraries were constructed by first cloning DNA fragments into the pCR™Blunt II-TOPO™ vector using Zero Blunt™ TOPO™ PCR Cloning Kit (Invitrogen), followed by PCR amplification of the inserts and Gibson assembly into barcoded pNBU2_repA1. In this protocol, sheared gDNA is first end-repaired in 20 µL reactions using Quick CIP (NEB) and following the manufacturer's protocol. End-repaired inserts are column-purified. Ligations were carried out in 6 µL TOPO reactions that contained 0.5 or 1 µL of pCR II-Blunt-TOPO vector, 1 µL of salt solution, and 20 ng of end-repaired inserts. Reactions were incubated at room temperature for 30 min and cleaned using SPRI beads (Beckman Coulter). Briefly, 0.9 equivalent volume of bead solution was mixed with each TOPO reaction and incubated for 15 min. Beads were washed three times with 80% EtOH and allowed to dry briefly. DNA was eluted with 5 µL of H_2O heated to 50 °C and transformed into 50 µL of electrocompetent 3-fold diluted *E. coli* EPI300 (Lucigen). An aliquot was taken out for CFU counting on LB-Kan50. Recovered cultures were stored as glycerol stocks at −80 °C. Multiple TOPO reactions and transformations were carried out until an estimate of ~40,000 to 100,000 CFUs was obtained for each library. Glycerol stocks of recovered cells from the same library were combined and inoculated into 200 mL of LB-Kan50 for selection. Inserts were PCR amplified from 150 ng of purified TOPO vectors using Q5 polymerase with an annealing temperature of 67 °C, 2 minute extension, and 8 cycles. Template vectors were digested using DpnI and the 3–5 kb size range was gel-purified. Insert amplicons were ligated into barcoded pNBU2_repA1 in 20 µL of Gibson reactions using the manufacturer's protocol. 20 ng of cleaned reaction was transformed into 50 µL of 4-fold diluted electrocompetent *E. coli* EC100D. Cells were recovered and stored as glycerol stocks before selection. Aliquots of the recovered cells were plated onto LB-Carb50 to estimate CFUs. Libraries were serially diluted to <100,000 CFUs and selected in liquid LB-Carb50 to make final glycerol stocks and perform BarSeq. All libraries were transformed into *E. coli* WM3064 for conjugation into *B. theta*.

Differences between libraries constructed using each cloning workflow were observed. The ligated *B. theta* library has the highest percentage of genes covered (98%) and the lowest strand bias. The PCR required for Gibson Assembly, though with a low cycling number, introduced biases that likely led to lower genome coverage values (Table 1). Gibson-assembled libraries exhibited a slight strand bias toward fragments oriented opposite to the promoter (ratio

of genes oriented opposite over genes oriented the same direction as vector promoter: 1.43–1.88). In contrast, the number of genes on either strand is very similar in the *B. theta* library, with a ratio of 0.98. This is likely a result of gene toxicity in *E. coli* caused by the lac promoter on the TOPO vector. This strand bias did not increase upon library conjugation into *B. theta*, which is supported by the low basal expression of the aTc inducible promoter.

PacBio sequencing to map libraries

Barcoded libraries purified from EC100D were used as templates for amplicon sequencing. Each sample was dual indexed using 5'-phosphorylated PCR primers containing 16 nt barcodes.[61] Barcode and insert regions were amplified in 50 µL PCRs with Q5 polymerase, 150 ng of vector template, 57 °C annealing temperature, 2 min extension time, and 6 or 8 cycles. 8–10 reactions per library were pooled, digested with DpnI twice (NEB), and column-concentrated. Amplicons were size-selected for >800 bp lengths using 0.55 equivalent volume of AMPure PB beads (PacBio). 200–600 ng of this was used as sample input for the manufacturer's protocol: Preparing SMRTbell libraries using PacBio Barcoded Overhang Adapters for Multiplexing Amplicons. The universal adapter was used during the ligation step instead of a barcoded adapter since amplicons are dual indexed. Adapter-ligated amplicons were digested with exonuclease using SMRTbell Enzyme Clean Up Kit 2.0 and purified using 0.55 equivalent volume of AMPure PB beads. Final amplicon solutions were quantified using Qubit BR Assay and sizes confirmed on the Bioanalyzer HS DNA chip (Agilent). Samples were sequenced on the PacBio Sequel II instrument using SMRT cell 8M (QB3 Genomics, UC Berkeley). See Supplementary Table 2 for the number of PacBio reads used to map each library. In general, one PacBio SMRT cell 8M was sufficient to map 3–4 libraries.

Computational pipeline used to map barcodes to genes

The Boba-seq script used to map barcodes to genes and detailed explanations can be found at https://github.com/OGalOz/Boba-seq and as a part of the deposited data. A summary is provided as Supplementary Figure 4. Input data files are required and consist of: a reference assembly file (.fasta), a genome annotation file (.gff), and long reads data (.fastq). The user also needs to provide a fasta file containing 4 short oligo sequences that flank the barcode and the insert fragment on each amplicon. A configuration .JSON file is used to set parameters for each step. Briefly, reads were first demultiplexed using PacBio's lima tool. Barcode and insert sequences are extracted using 15bp flanking oligos using usearch (www.drive5.com/usearch/). Filters are applied to exclude reads with concatemers and incorrect barcode lengths. Inserts with an expected error of >10 are excluded using vsearch (https://github.com/torognes/vsearch). Inserts are mapped to the reference genome using minimap2 and results kept for hits with high percent coverage and percent match values.[62] The majority of barcodes mapped to a single genomic location. For barcodes that are mapped to multiple locations, most are due to sequencing errors and are within a few nt of each other. For barcodes mapping to overlapping locations, the location with the top read count is kept. Barcodes mapped to distant genomic locations are kept in the mapping tables, but are excluded from our fitness analysis. Finally, genomic positions are mapped to protein-coding genes to generate mapping tables, which are used to calculate library statistics.

Conjugation into *B. theta* and other Bacteroidales

Single vectors and barcoded libraries were first transformed into conjugation donor strains *E. coli* S17-1 or WM3064, respectively. 40 mL of Bacteroidales recipient at OD ~0.1 was mixed with 4 mL of *E. coli* donor at OD ~1 outside of the anaerobic chamber. Mixed cultures were pelleted by centrifugation and resuspended in 250 µL BHIS. Each cell mixture was then pooled onto BHIS or BHIS-DAP (for WM3064) plates and incubated at 37 °C overnight aerobically for >16 h. The entire cell mass was then scraped and resuspended in 1 mL of BHIS, from which 20–100 µL was plated onto BHIS-Gm200-Erm5 agar. Antibiotic plates were incubated at 37 °C anaerobically for 48 h or longer until distinct colonies formed. Single colonies were restreaked onto BHIS-Gm200-Erm5 once, from which a single colony was then inoculated into BHIS-Gm200-Erm5. Bacteroidales cultures were inoculated into LB and incubated aerobically at 37 °C to ensure a lack of *E. coli* contamination. Glycerol stocks were checked by Sanger sequencing of plasmids and 16S amplicons. Using this protocol, pKSH24 and pTIO-1 were conjugated into *B. theta* to test replication in a new host. pNBU2_repA1, pNBU2_repA2, and pNBU2_repA3 with NanoLuc were conjugated into 13 different *Bacteroides*, *Parabacteroides*, and *Phocaeicola* species.

Complementation libraries and assays

B. theta, B. uniformis, B. fragilis, and *P. johnsonii* libraries were conjugated into *B. theta* Δ*tdk*ΔBT2158 or Δ*tdk*ΔBT3703 using *E. coli* WM3064 as the conjugation strain. 200 µL of conjugation mix in 1 mL BHIS was inoculated into 100 mL of pre-reduced BHIS-Gm200-Erm5 in the anaerobic chamber. An aliquot of the conjugation mix was serially diluted for CFU counting to ensure that ~10^6 transconjugants were inoculated for selection. Cultures at OD >1 were pelleted and resuspended in BHIS with 0.15% Cys and 15% glycerol to make glycerol stocks.

For complementation assays, the starter culture of each library was diluted to OD of 0.5 in BHIS-Gm200-Erm5 with 40 ng/mL aTc and incubated at 37 °C for 3.5 h to induce gene expression. Cells were pelleted and washed with VB minimal medium twice, this served as Time0 samples. Each library was inoculated at a starting OD of 0.02 in 1 mL VB medium, 5 µg/mL Erm, 40 ng/mL aTc, and 20 mM of either leucrose, palatinose, trehalose, or α-cyclodextrin. Cultures were incubated at 37 °C and harvested when turbid. Each assay was also streaked onto BHIS-Gm200-Erm5 agar plates. Colony PCR and Sanger sequencing were used to identify clones containing BT2158, BT4448, or BACUNI_RS05440.

Growth assays of complemented mutants

Mutants complemented with BT2158, BT4448, or BACUNI_RS05440 (Supplementary Table 6), *B. theta* Δ*tdk*ΔBT2158, and *B. theta* Δ*tdk* parental strain were inoculated into BHIS-Erm5 or BHIS and grew overnight at 37 °C. Cells were pelleted and washed once with VB minimal medium. Each mutant was inoculated at a starting OD of 0.05 in 200 µL VB-Erm5 and 20 mM trehalose. *B. theta* Δ*tdk* and *B. theta* Δ*tdk*ΔBT2158 were inoculated in the same manner, except without Erm. Growth measurements were obtained in a 96-well plate using an Epoch 2 Microplate Spectrophotometer (Agilent) inside of an anaerobic chamber. OD600 was measured with pathlength corrected to 1 cm.

Fitness screen of pooled libraries in wild-type *B. theta*

B. caccae, B. fragilis, *B. salyersiae, B. uniformis, P. johnsonii,* and *P. merdae* libraries were conjugated into wild-type *B. theta* similarly as described for complementation libraries. For fitness assays, 1 mL of glycerol stock was pelleted and resuspended in 10 mL of BHIS-Erm5 with 40 ng/mL of aTc as starter cultures to grow for 16–18 h. These were then pelleted and washed with VB-Erm5 twice. Libraries were pooled equally based on OD to a final OD of 5, which served as Time0 samples. For carbon utilization assays, pooled libraries were inoculated at a starting OD of 0.04 in 1 mL of VB-Erm5 with 40 ng/mL aTc, and a single carbon substrate at the concentration (5 mg/mL, 10 mg/mL, or 20 mM) listed in Supplementary Table 4. For stress tolerance assays, pooled libraries were inoculated at a starting OD of 0.1 in 0.5 mL of BHIS-Erm5, 40 ng/mL aTc, and a range of concentrations of an inhibitory compound listed on Supplementary Table 4. Assays were carried out in 48-well plates and incubated at 37 °C. Cultures were harvested when turbid for BarSeq.

DNA barcode sequencing (BarSeq) and analysis

Vector libraries were purified from *E. coli* cultures using E.Z.N.A. Plasmid DNA Mini Kit (Omega Bio-tek) or Monarch Plasmid Miniprep kit. However, these plasmid purification kits lead to low yields and high levels of contaminants for Bacteroidales. DNA purified from the same set of Bacteroidales cultures using plasmid purification kits and genomic DNA kits were used as template for barcode PCR amplification. The BarSeq results were comparable and therefore, DNeasy Blood & Tissue Kit (Qiagen) or QiaAmp DNA QIAcube HT Kit for 96-well sample plates were used to extract DNA from *B. theta* cultures for BarSeq.

Barcodes are amplified using previous published primers and methods (Supplementary Table 8).[9,63] Briefly, BarSeq PCRs were performed as 50 µL reactions with Q5 polymerase (NEB) and 200 ng of template DNA: 98 °C for 4 min, 25 cycles of 98 °C for 30 s, 55 °C for 30 s, and 72 °C for 30 s, and 72 °C for 5 min. The P2 BarSeq primers allowed demultiplexing by Illumina software.[9] The P1 BarSeq primers contained an additional index checked by the MultiCodes.pl script.[63] All PCRs were pooled equally in volume and column-cleaned. Up to 96 samples were multiplexed for sequencing on full lanes of Illumina HiSeq400 SE50 (QB3 Genomics, UC Berkeley), partial lanes of NovaSeq6000 PE150 (Novogene), or full lanes of MiSeq (v2 reagent kit). 11,000–94,000 reads per sample were obtained for complementation assays sequenced on the MiSeq and 0.8–3.9 million reads were obtained for samples sequenced on the HiSeq or NovaSeq.

BarSeq reads were converted to counts per barcode using the MultiCodes.pl script in the feba code base (https://bitbucket.org/berkeleylab/feba/), with the -bs3 and -minQuality 0 options. To estimate the diversity of barcoded libraries, only reads that had a quality score of ≥30 at each position (the -minQuality 30 option) were used, which corresponds to an error rate for barcodes of at most 0.001 * 20 nt = 2%. Furthermore, any barcodes that were off-by-1 errors from a more common barcode were eliminated. Barcodes observed in the PacBio dataset are compared to barcode sequences from the BarSeq data to determine the percentage of barcodes mapped. From these mapped barcodes, 0.5% are shared among the 6 libraries used in fitness assays.

For the highly diverse empty library, the Chao1 estimator was used to estimate the barcode diversity from 2.99 million high-quality reads,[64] The Chao1 estimator uses the number of barcodes seen just once, but some may be due to sequencing errors. To correct for this, barcodes that were off-by-1 nucleotide from a more abundant barcode were ignored, and the number of singletons was reduced by an assumed error rate ranging from 0% to 2%. This yielded estimates ranging from 13.5–14.1 million barcodes for the empty pNBU2_repA1 library.

Calculation of fitness scores for each fragment

Only barcodes that were mapped to a single location and were present in the conjugation donor (*E. coli* WM3064) libraries were considered. For complementation assays, we computed fitness scores for all of these fragments.

For fitness assays in the wild-type *B. theta* background with six pooled libraries, only 238,787 of the 303,011 barcodes (that were present in WM3064) were detected in the Time0 samples. But barcodes that were not detected in Time0 samples occasionally had statistically-significant benefits. This is likely due to positive selection that led mutants to gain in abundance in the assay condition despite their low initial abundance. Therefore, fitness scores were computed for all barcodes with a reasonable total read count across all Time0 and fitness assay samples. A bimodal distribution was observed for total reads, with peaks at 0 (7% of strains) and at around 300. 73% of strains had between 10 and 10,000 total reads. Strains with less than 10 total reads (14%) are considered to be at low abundance and excluded from analysis. Fitness scores were computed for the remaining 260,862 barcodes.

The fitness of each strain was defined as:

$$\log_2\{ (n + \psi)/N \} - \log_2\{ (n_0 + 1/\psi)/N_0 \}$$

$$\psi = \mathrm{sqrt}(N/N_0)$$

where n is the number of reads for this barcode in the experimental sample, n_0 is the number of reads for this barcode across the Time0 samples, and N and N_0 are the total number of reads for all considered barcodes in the experimental or Time0 samples, respectively. The pseudocount ψ is added to the counts to ensure that the log ratio is not undefined if either count is zero. This value of ψ also ensures that if both counts (n and n_0) are 0, then the fitness will be zero (corresponding to no evidence of a change in relative abundance). If the experimental and Time0 samples have the same total number of reads, then ψ = 1. The script for fitness score calculations are available at: https://doi.org/10.6084/m9.figshare.21304941.v1.

Determination of statistically significant hits

The minimum fitness value threshold for statistical significance was selected based on a control comparison between the two Time0 samples from the fitness screen. Each sample contained a mixture of six genomic libraries and had 2.9 million or 3.0 million total reads. Of the 260,862 barcodes considered, the fitness values ranged from -4.1 to +4.0. Therefore, a minimum fitness value of +5 was selected for a statistically-significant hit or barcode.

Because a few complementation assays were sequenced at much lower coverage, a z score cutoff was also applied. The standard error of the fitness value was estimated as the square root of the variance given Poisson noise in counts[9], or

$$\sigma = \mathrm{sqrt}\{ 1/(n + \psi) + 1/(n_0 + 1/\psi) \} / \ln(2)$$

The z score is then fitness / σ and z ≥4 is selected as the cutoff. For the standard normal distribution and 260,862 barcodes, 8 values would be expected above this threshold.

Selection of biologically consistent hits and the likely causative protein for each region

For each barcode and insert that had significantly high fitness scores in both replicates, the phenotype is biologically consistent if either an overlapping fragment or a fragment encoding a similar protein have statistically significant fitness. Specifically, an insert was confirmed by overlap if another insert from the same region of the genome was significant in both replicates. An insert was confirmed by similarity if the top protein for the region (see below) was similar (at least 40% identity) to a protein that was found within another insert that had significantly high fitness in two replicates for the same condition. For experiments in the wild-type *B. theta* background where libraries were pooled, the confirmation would occur in the same experiments. For complementation assays where libraries were assayed separately, the criterion might be fulfilled from different assays in the same condition.

For each region with inserts that were statistically significant in both replicates, the insert with the highest fitness score (after averaging replicates) was selected first. For each protein encoded within this insert, the average fitness score for all inserts containing this gene was computed. The gene with the highest average fitness is selected as the top-scoring protein for this region. The top-scoring protein of each region was used to identify cases of confirmation by similarity, and also served as a first guess for the causative protein. In a few cases, visual inspection of 'fitness scores versus inserts' coverage plots led to a different gene being selected as the causative factor in the observed phenotype.

Experimental validation of gene hits from fitness assays

Selected gene hits from the fitness screens were PCR amplified from the corresponding gDNA, cloned into the pNBU2_repA1 backbone via Gibson Assembly, and transformed into *E. coli* S17-1 to generate validation vectors. Mutants with high fitness scores were selected for validations and the exact fragment boundaries for these were extracted from library mapping tables to design primers (Supplementary Table 9). Additionally, amplicons of the precise gene were obtained. Validation vectors were conjugated into *B. theta*, and each strain was grown in the same condition as carried out for the fitness assay. Each strain was cultured in 200 µL on a 96-well plate in replicates of 3. Growth measurements were obtained using an Epoch 2 Microplate Spectrophotometer (Agilent) inside of an anaerobic chamber.

Prophage predictions and gene annotations

The hypothetical proteins enriched in heparin were determined to be within a prophage region using various tools. VirSorter 2 identified 2 hallmark genes and a score of 0.96 for a dsDNA

phage.[65] Gene annotations from VirSorter 2 and Galaxy CPT phage annotation workflow were combined to generate Supplementary Figure 9.[65,66] Phage MAGs from gut metagenomes part of the BLAST results included additional gene annotations, which were incorporated as well.[67] tBLASTn results using INE72_RS16800 as query against the NCBI nr/nt database (accessed July, 2022) are used to generate Supplementary Table 7.

Materials and data availability

The Boba-seq script used to map libraries is available at https://github.com/OGalOz/Boba-seq. Plasmid maps, a snapshot of the python script used for library mapping, mapping tables for each library, R script for fitness score calculations, and output files from fitness analysis are deposited at: https://doi.org/10.6084/m9.figshare.21304941.v1. All other data are available from the authors upon request. Please direct plasmids and strain requests to Y.Y.H. and A.P.A.

Acknowledgements

We thank Dr. Naoya Ohara (Okayama University) for pTIO-1, Dr. Andrei Shkoporov (University of College Cork) for pKSH24, and Dr. Laurie Comstock (University of Chicago) for pFD340. pNBU2_erm-TetR-P1T_DP-GH023 and pNBU2_erm-TetR-P1T_DP-GH023-NanoLuc were gifts from Dr. Andrew Goodman at Yale University (Addgene plasmids #90324 and #117728). We thank Surya Tripathi for access to unpublished RB-TnSeq fitness data. We thank Dr. Bradley Biggs for helpful comments on the manuscript. Y.Y.H. is an Astellas Pharmaceuticals Awardee of the Life Sciences Research Foundation. A.H. is an awardee of NSF GRFP grant DGE-2146752. We are grateful for the support of the Dean A. Richard Newton Memorial Professor Chair Funds. This work was supported in part by ENIGMA – Ecosystems and Networks Integrated with Genes and Molecular Assemblies (http://enigma.lbl.gov), a Science Focus Area Program at Lawrence Berkeley National Laboratory is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research under contract number DE-AC02-05CH11231. This work was partly supported by the NIH grant RM1 GM135102 (to A.M.D.).

Author contributions

Y.Y.H., M.N.P., A.M.D., and A.P.A. conceived the project. Y.Y.H., A.H., A.M.D., D.H., and H.C. carried out experimental work and collected data. Y.Y.H., M.N.P., and O.G. performed computational analyses. Y.Y.H. and M.N.P. wrote the original draft. Y.Y.H., M.N.P., A.H., A.M.D., and A.P.A. revised and edited the paper.

Competing interests

The authors have no competing interests to declare.

References

1. Human Microbiome Project Consortium. Structure, function and diversity of the healthy

human microbiome. *Nature* **486**, 207–214 (2012).
2. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017).
3. Wu, G. D. *et al.* Comparative metabolomics in vegans and omnivores reveal constraints on diet-dependent gut microbiota metabolite production. *Gut* **65**, 63–72 (2016).
4. Franzosa, E. A. *et al.* Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci.* **111**, E2329–E2338 (2014).
5. Vujkovic-Cvijin, I. *et al.* Host variables confound gut microbiota studies of human disease. *Nature* **587**, 448–454 (2020).
6. Ni, J., Wu, G. D., Albenberg, L. & Tomov, V. T. Gut microbiota and IBD: causation or correlation? *Nat. Rev. Gastroenterol. Hepatol.* **14**, 573–584 (2017).
7. Lobb, B., Tremblay, B. J.-M., Moreno-Hagelsieb, G. & Doxey, A. C. An assessment of genome annotation coverage across the bacterial tree of life. *Microb. Genomics* **6**, e000341 (2020).
8. Vanni, C. *et al.* Unifying the known and unknown microbial coding sequence space. *eLife* **11**, e67667 (2022).
9. Wetmore, K. M. *et al.* Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons. *mBio* **6**, e00306-15 (2015).
10. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **557**, 503–509 (2018).
11. Liu, H. *et al.* Functional genetics of human gut commensal Bacteroides thetaiotaomicron reveals metabolic requirements for growth across environments. *Cell Rep.* **34**, 108789 (2021).
12. Adler, B. A. *et al.* Systematic Discovery of Salmonella Phage-Host Interactions via High-Throughput Genome-Wide Screens. *bioRxiv* (2020) doi:10.1101/2020.04.27.058388.
13. Yaung, S. J. *et al.* Improving microbial fitness in the mammalian gut by in vivo temporal functional metagenomics. *Mol. Syst. Biol.* **11**, 788 (2015).
14. Joglekar, P. *et al.* Intestinal IgA Regulates Expression of a Fructan Polysaccharide Utilization Locus in Colonizing Gut Commensal Bacteroides thetaiotaomicron. *mBio* **10**, e02324-19 (2019).
15. Tasse, L. *et al.* Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. *Genome Res.* **20**, 1605–1612 (2010).
16. Yoon, M. Y. *et al.* Functional screening of a metagenomic library reveals operons responsible for enhanced intestinal colonization by gut commensal microbes. *Appl. Environ. Microbiol.* **79**, 3829–3838 (2013).
17. Crofts, T. S., Mcfarland, A. G. & Hartmann, E. M. Mosaic Ends Tagmentation (METa) assembly for extremely efficient 1 construction of functional metagenomic libraries. *mSystems* **6**, e00524-21 (2021).
18. Forsberg, K. J. *et al.* Functional metagenomics-guided discovery of potent Cas9 inhibitors in the human microbiome. *eLife* **8**, e46540 (2019).
19. Mastropaolo, M. D., Thorson, M. L. & Stevens, A. M. Y. 2009. Comparison of Bacteroides thetaiotaomicron and Escherichia coli 16S rRNA gene expression signals. *Microbiology* **155**, 2683–2693 (2009).
20. Vingadassalom, D. *et al.* An unusual primary sigma factor in the Bacteroidetes phylum. *Mol. Microbiol.* **56**, 888–902 (2005).
21. Wegmann, U., Horn, N. & Carding, S. R. Defining the Bacteroides ribosomal binding site. *Appl. Environ. Microbiol.* **79**, 1980–1989 (2013).
22. Lee, S. M. *et al.* Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature* **501**, 426–429 (2013).
23. Lam, K. N., Martens, E. C. & Charles, T. C. Developing a Bacteroides System for Function-Based Screening of DNA from the Human Gut Microbiome. *mSystems* **3**,

e00195-17 (2018).
24. Mutalik, V. K. *et al.* Dual-barcoded shotgun expression library sequencing for high-throughput characterization of functional traits in bacteria. *Nat. Commun.* **10**, 1–14 (2019).
25. Wexler, A. G., Goodman, A. L., Voigt, C. A., Lu, T. K. & Hennet, T. An insider's perspective: Bacteroides as a window into the microbiome. *Nat. Microbiol.* **2**, 17026 (2017).
26. Liu, H. *et al.* Magic Pools: Parallel Assessment of Transposon Delivery Vectors in Bacteria. *mSystems* **3**, e00143-17 (2018).
27. Lim, B., Zimmermann, M., Barry, N. A. & Goodman, A. L. Engineered Regulatory Systems Modulate Gene Expression of Human Commensals in the Gut. *Cell* **169**, 547-558.e15 (2017).
28. Tagawa, J. *et al.* Development of a novel plasmid vector pTIO-1 adapted for electrotransformation of Porphyromonas gingivalis. *J. Microbiol. Methods* **105**, 174–179 (2014).
29. Shkoporov, A. N. *et al.* Plasmid Analysis of a novel 8 . 9 kb cryptic plasmid from Bacteroides uniformis , its long-term stability and spread within human microbiota. *Plasmid* **69**, 146–159 (2013).
30. Smith, C. J., Rogers, M. B. & Mckee, M. L. Heterologous Gene Expression in Bacteroides fragilis. *Plasmid* **27**, 141–154 (1992).
31. Chen, Y. J. *et al.* Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nat. Methods* **10**, 659–664 (2013).
32. Taketani, M. *et al.* Genetic circuit design automation for the gut resident species Bacteroides thetaiotaomicron. *Nat. Biotechnol.* **38**, 962–969 (2020).
33. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
34. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).
35. Liou, C. S. *et al.* A Metabolic Pathway for Activation of Dietary Glucosinolates by a Human Gut Symbiont. *Cell* **180**, 717-728.e19 (2020).
36. Varel, V. H. & Bryant, M. P. Nutritional Features of Bacteroides fragilis subsp. fragilis. *Appl. Microbiol.* **28**, 251–257 (1974).
37. Auger, S., Gomez, M. P., Danchin, A. & Martin-Verstraete, I. The PatB protein of Bacillus subtilis is a C-S-lyase. *Biochimie* **87**, 231–238 (2005).
38. Méndez, J. *et al.* A Novel cdsAB Operon Is Involved in the Uptake of l-Cysteine and Participates in the Pathogenesis of Yersinia ruckeri. *J. Bacteriol.* **193**, 944–951 (2011).
39. Larsbrink, J. *et al.* A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. *Nature* **506**, 498–502 (2014).
40. Ikegaya, M. *et al.* Structural basis of the strict specificity of a bacterial GH31 α-1,3-glucosidase for nigerooligosaccharides. *J. Biol. Chem.* **298**, 101827 (2022).
41. Álvarez-Añorve, L. I., Calcagno, M. L. & Plumbridge, J. Why Does Escherichia coli Grow More Slowly on Glucosamine than on N-Acetylglucosamine? Effects of Enzyme Levels and Allosteric Activation of GlcN6P Deaminase (NagB) on Growth Rates. *J. Bacteriol.* **187**, 2974–2982 (2005).
42. Brigham, C. J. & Malamy, M. H. Characterization of the RokA and HexA broad-substrate-specificity hexokinases from Bacteroides fragilis and their role in hexose and N-acetylglucosamine utilization. *J. Bacteriol.* **187**, 890–901 (2005).
43. Romero-Rodríguez, A., Ruiz-Villafán, B., Rocha-Mendoza, D., Manzo-Ruiz, M. & Sánchez, S. Biochemistry and regulatory functions of bacterial glucose kinases. *Arch. Biochem. Biophys.* **577–578**, 1–10 (2015).
44. Price, M. N., Deutschbauer, A. M. & Arkin, A. P. Filling gaps in bacterial catabolic pathways with computation and high-throughput genetics. *PLOS Genet.* **18**, e1010156 (2022).

45. Pollet, R., Marin, L. & Koropatkin, N. TonB-dependent transporters in the Bacteroidetes: Unique domain structures and potential functions. *Mol. Microbiol.* **115**, 490–501 (2021).
46. Gray, D. A. *et al.* Insights into SusCD-mediated glycan import by a prominent gut symbiont. *Nat. Commun.* **12**, 44 (2021).
47. Zhou, J. *et al.* A new α-galactosidase from symbiotic Flavobacterium sp. TN17 reveals four residues essential for α-galactosidase activity of gastrointestinal bacteria. *Appl. Microbiol. Biotechnol.* **88**, 1297–1309 (2010).
48. Glenwright, A. J. *et al.* Structural basis for nutrient acquisition by dominant members of the human gut microbiota. *Nature* **541**, 407–411 (2017).
49. Phansopa, C. *et al.* Structural and functional characterization of NanU, a novel high-affinity sialic acid-inducible binding protein of oral and gut-dwelling Bacteroidetes species. *Biochem. J.* **458**, 499–511 (2014).
50. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
51. Casanueva, A. I., Paul, L., Patrick, S. & Abratt, V. R. An AraC/XylS family transcriptional regulator homologue from Bacteroides fragilis is associated with cell survival following DNA damage. *FEMS Microbiol. Lett.* **278**, 249–256 (2008).
52. de Freitas, M. C. R. *et al.* Exploratory Investigation of Bacteroides fragilis Transcriptional Response during In vitro Exposure to Subinhibitory Concentration of Metronidazole. *Front. Microbiol.* **7**, 1465 (2016).
53. Lauber, F., Cornelis, G. R. & Renzi, F. Identification of a New Lipoprotein Export Signal in Gram-Negative Bacteria. *mBio* **7**, e01232-16.
54. Shiver, A. L., Culver, R., Deutschbauer, A. M. & Huang, K. C. Rapid ordering of barcoded transposon insertion libraries of anaerobic bacteria. *Nat. Protoc.* **16**, 3049–3071 (2021).
55. Mutalik, V. *et al.* High-throughput mapping of the phage resistance landscape in E. coli | PLOS Biology. *PLoS Biol* **18**, e3000877 (2020).
56. Helmann, T. C., Deutschbauer, A. M. & Lindow, S. E. Genome-wide identification of Pseudomonas syringae genes required for fitness during colonization of the leaf surface and apoplast. *Proc. Natl. Acad. Sci.* **116**, 18900–18910 (2019).
57. Adler, B. A. *et al.* Parallel multicopy-suppressor screens reveal convergent evolution of phage-encoded single gene lysis proteins. *bioRxiv* 2022.01.20.477139 (2022) doi:10.1101/2022.01.20.477139.
58. Vassallo, C. N., Doering, C. R., Littlehale, M. L., Teodoro, G. I. C. & Laub, M. T. A functional selection reveals previously undetected anti-phage defence systems in the E. coli pangenome. *Nat. Microbiol.* **7**, 1568–1579 (2022).
59. Zhang, M. *et al.* Dynamic genetic adaptation of Bacteroides thetaiotaomicron murine gut colonization. *bioRxiv* 2022.02.23.481734 (2022) doi:10.1101/2022.02.23.481734.
60. Plotka, M., Wozniak, M. & Kaczorowski, T. Quantification of Plasmid Copy Number with Single Colour Droplet Digital PCR. *PLoS ONE* **12**, e0169846 (2017).
61. Callahan, B. J. *et al.* High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Res.* **47**, e103 (2019).
62. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinforma. Oxf. Engl.* **34**, 3094–3100 (2018).
63. Morgan N. Price, a Jayashree Ray, a Anthony T. Iavarone, B., Hans K. Carlson, a Elizabeth M. Ryan, c Rex R. Malmstrom, C., Adam P. Arkin, a, D. & Deutschbauer, A. M. Oxidative Pathways of Deoxyribose and Deoxyribonate Catabolism. *mSystems* **4**, e00297-18 (2019).
64. Chao, A. Nonparametric Estimation of the Number of Classes in a Population. *Scand. J. Stat.* **11**, 265–270 (1984).
65. Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 1–13 (2021).
66. Ramsey, J. *et al.* Galaxy and Apollo as a biologist-friendly interface for high-quality

cooperative phage genome annotation. *PLOS Comput. Biol.* **16**, e1008214 (2020).
67. Tisza, M. J. & Buck, C. B. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2023202118 (2021).
68. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
69. Gilchrist, C. L. M. & Chooi, Y.-H. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* **37**, 2473–2475 (2021).

Supplementary figures

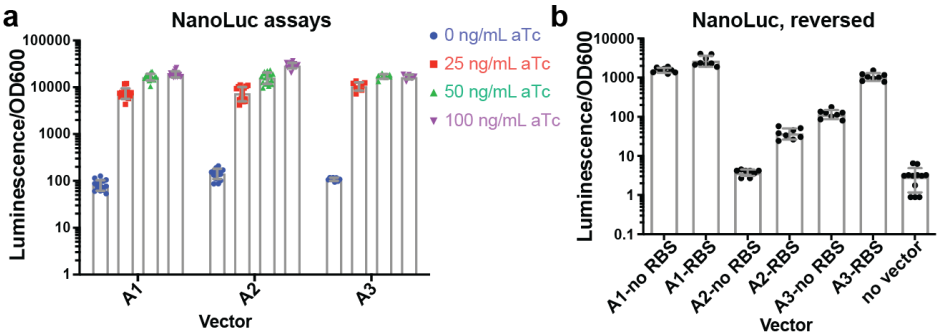

Figure S1. Anhydrotetracycline (aTc) induces expression of NanoLuc on pNBU2 replicative vectors in *B. theta*. A) A range of aTc concentrations (0, 25, 50, 100 ng/mL) was tested for three vectors. Two or four experiments were performed for each vector with a total replicate of 8 or 16. B) NanoLuc was inserted into the pNBU2_repA1 vector in the opposite direction to the aTc inducible promoter to test for constitutive promoter activity from the downstream replicon fragment. NanoLuc gene with and without the upstream RBS (GH023) were tested in the absence of aTc. Two experiments were carried out with n = 4. Luminescence values were normalized by OD600 for all assays. Standard deviation is shown as error bars for both plots.

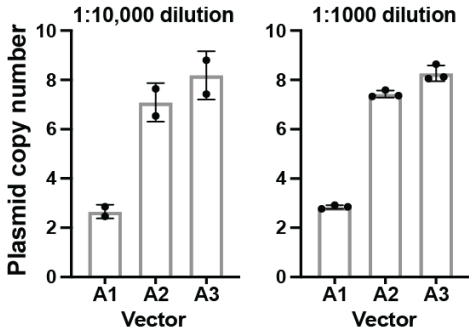

Figure S2. Plasmid copy numbers in *B. theta* determined using droplet digital PCR (ddPCR). Cultures in the late exponential phase were harvested for this assay. Replicates from two dilutions were used to calculate plasmid copy numbers.

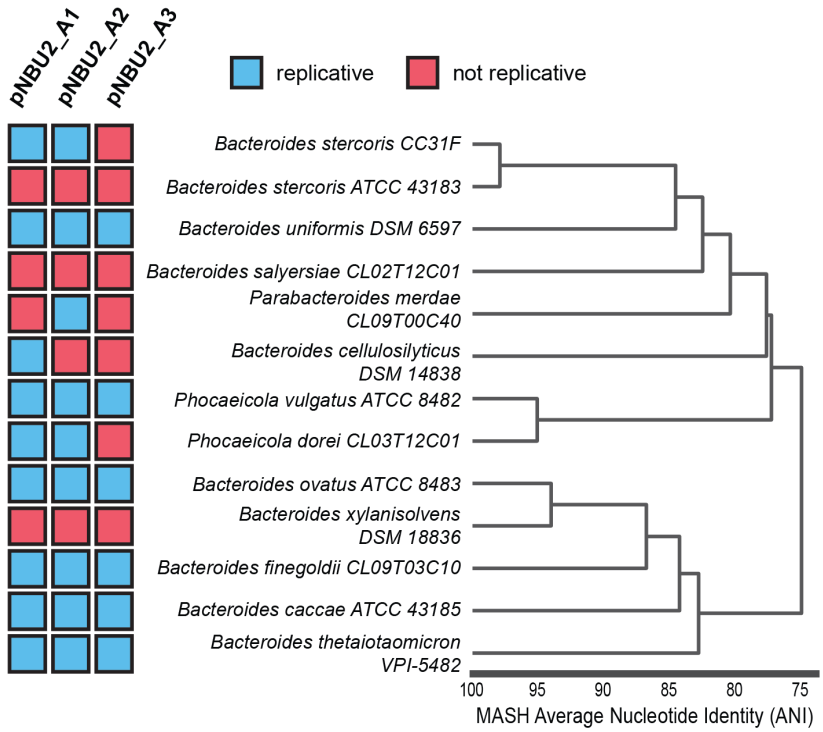

Figure S3. Replication of three pNBU2-based vectors across 13 Bacteroidales isolates. Genome distance estimation and visualization were performed using dRep.[68]

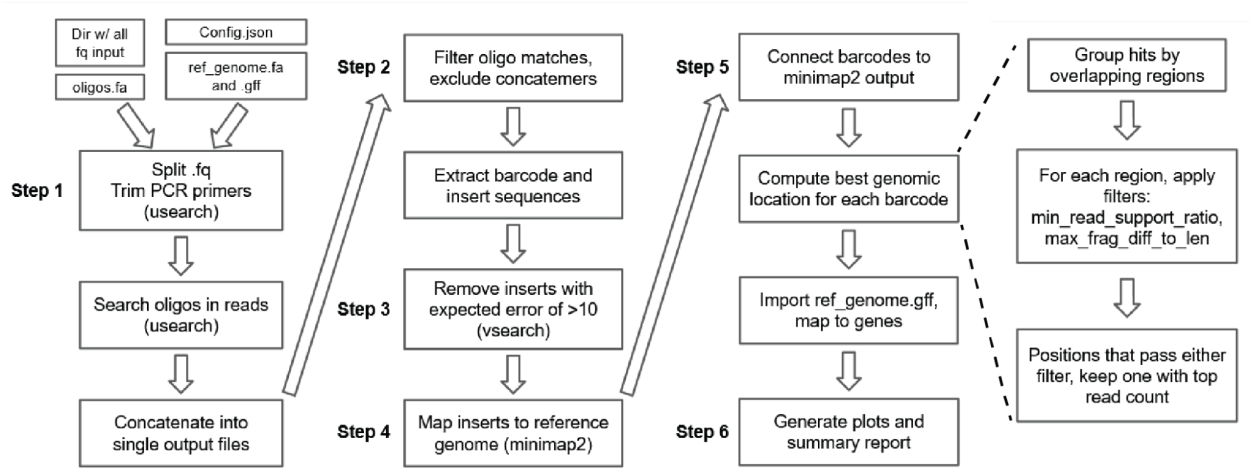

Figure S4. Overview of the Boba-seq mapping script. Input files, tools, and all steps are listed. Additional details for step 5 are listed for barcodes that map to multiple, non-overlapping genomic regions with high fragment length coverage. Although the empty vector library has a high barcode diversity, it is possible to clone multiple inserts into vectors sharing the same barcode. We observed this in a very small fraction of barcodes and later removed them from our fitness score analysis, but kept this functionality in our mapping script. An additional step is used to compute the best genomic location for these barcodes. The min_read_support_ratio value is the ratio of highest read count to the second highest read count. The max_frag_diff_to_len value is the fraction of the maximum distance between start and end positions of 5' of mapped locations plus the maximum distance between start and end positions of 3' of mapped locations within each region, over the average insert length. These two parameters are used as thresholds to select the best genomic location for each barcode.

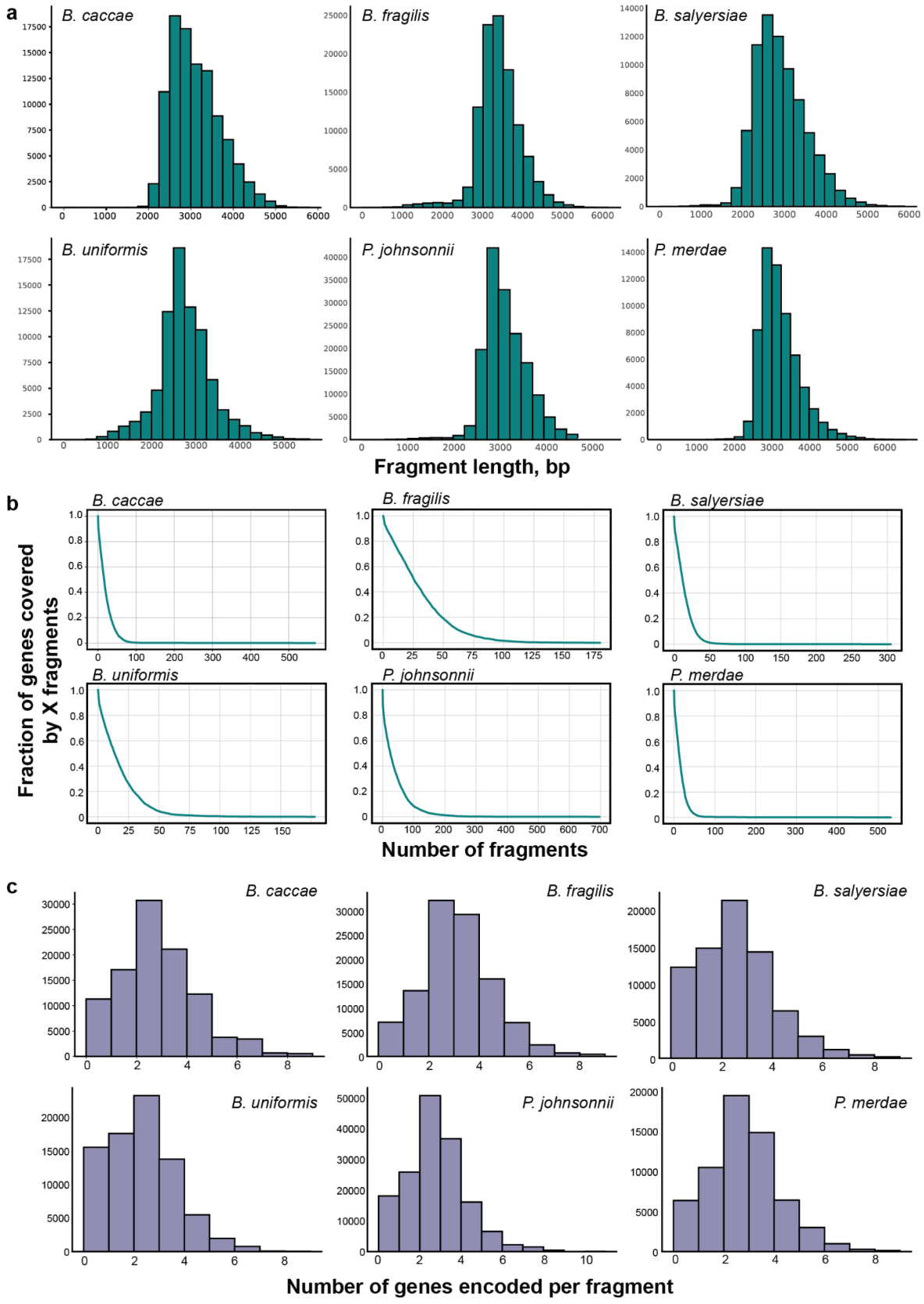
a
B. caccae
B. fragilis
B. salyersiae
B. uniformis
P. johnsonnii
P. merdae
Fragment length, bp
b
B. caccae
B. fragilis
B. salyersiae
B. uniformis
P. johnsonnii
P. merdae
Fraction of genes covered by X fragments
Number of fragments
c
B. caccae
B. fragilis
B. salyersiae
B. uniformis
P. johnsonnii
P. merdae
Number of genes encoded per fragment

Figure S5. Summary plots for six genomic libraries constructed using Gibson Assembly workflow. Plots are generated as part of the Boba-seq mapping script. A) Histogram of fragment sizes. B) Cumulative gene coverage plot. C) Histogram of the number of full-length genes encoded per fragment.

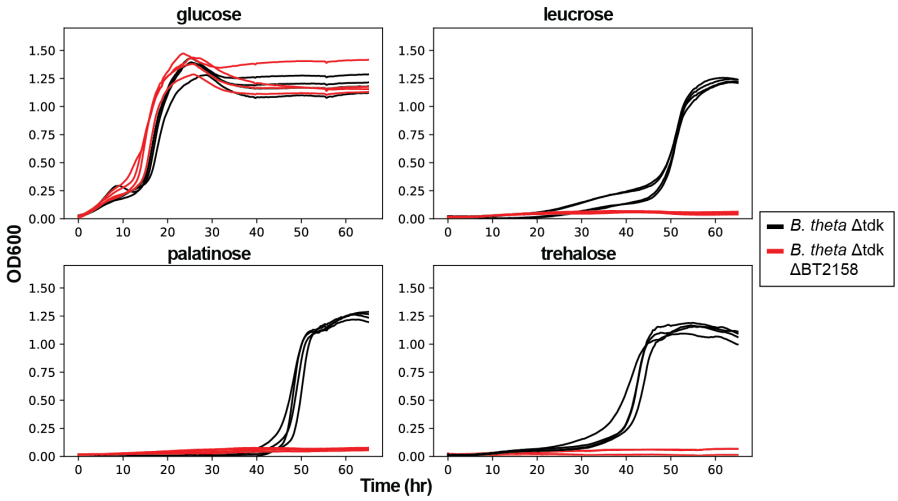

Figure S6. Growth curves of *B. theta* Δ*tdk* parental strain and Δ*tdk*ΔBT2158 mutant on 20 mM glucose, leucrose, palatinose, and trehalose in a minimal medium. OD values are pathlength-corrected and blank-subtracted. Each strain was grown in replicates of 4 per condition.

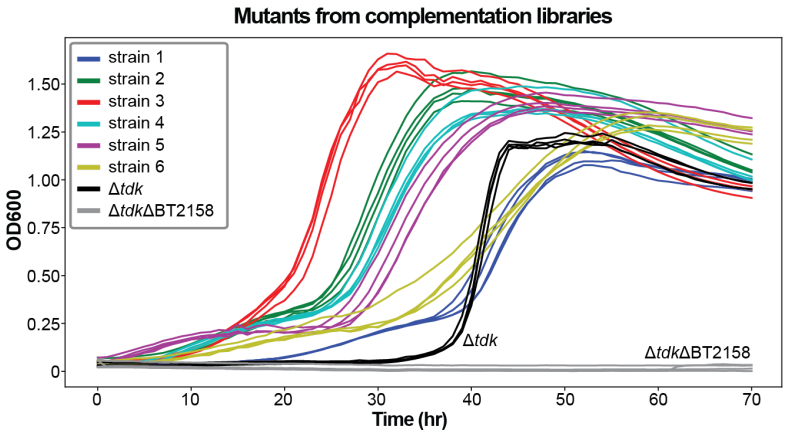

Figure S7. Growth curves of ΔBT2158 mutant complemented with a plasmid containing either BT2158 or a homolog. Complemented isolates, the Δ*tdk* parental strain, and the Δ*tdk*ΔBT2158 mutant were cultured in minimal media with 20 mM trehalose. Strains 1–2 encode BT2158, strains 3–4 encode BT4448, and strains 5–6 encode a homology from *B. uniformis.* See Supplementary Table 6 for the precise genomic fragments encoded in each strain. OD values are pathlength-corrected and blank-normalized. Each strain was grown in replicates of 4.

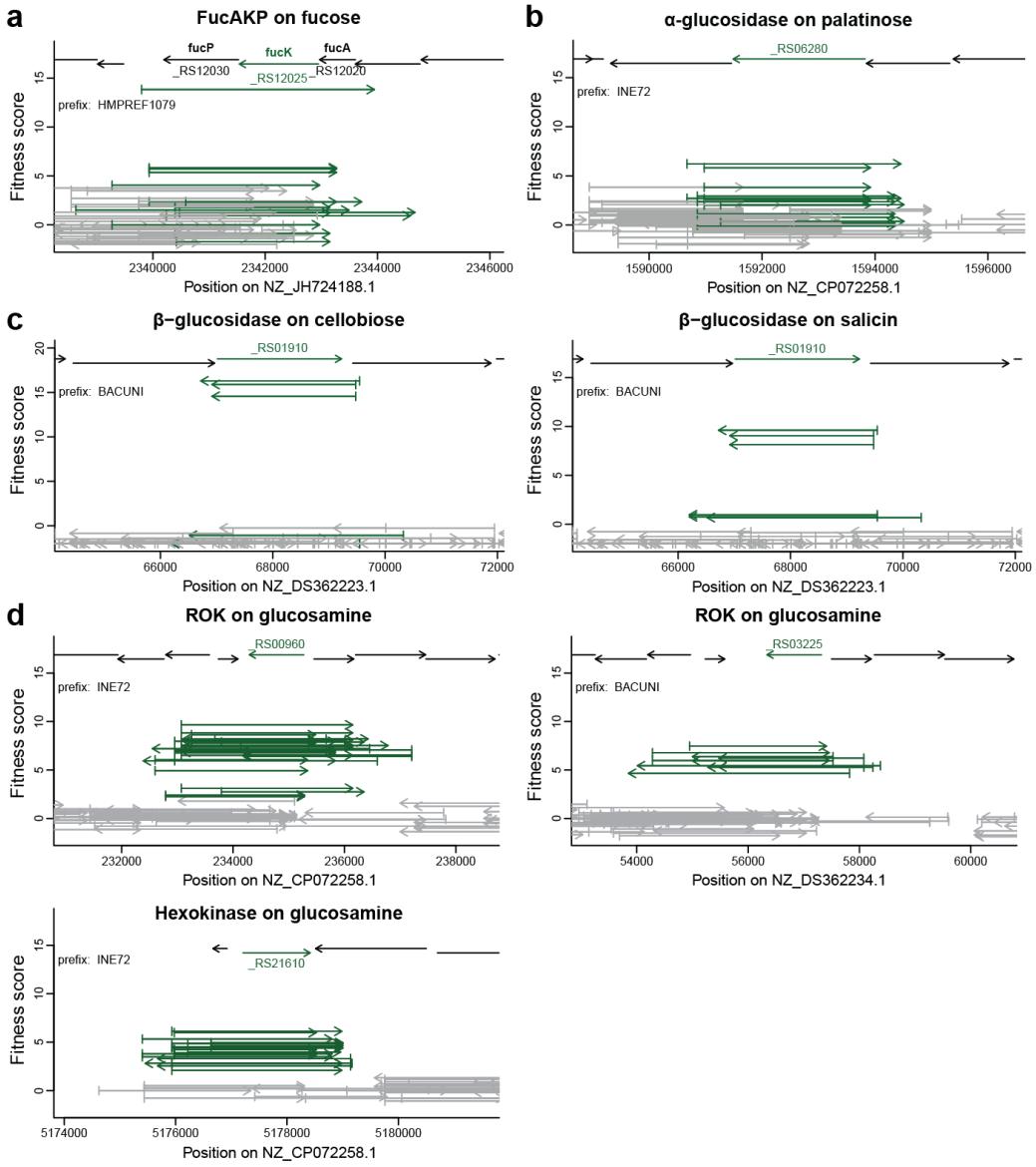

Figure S8. Average fitness scores for mutants containing metabolic gene hits are displayed as genomic plots. A) Fucose kinase and permease were enriched in L-fucose. B) α-Glucosidase (INE72_RS06280) was enriched in palatinose. C) β-Glucosidase (BACUNI_RS01910) was enriched in both D-salicin and D-cellobiose. D) Both an ROK kinase and a hexokinase were enriched in D-glucosamine.

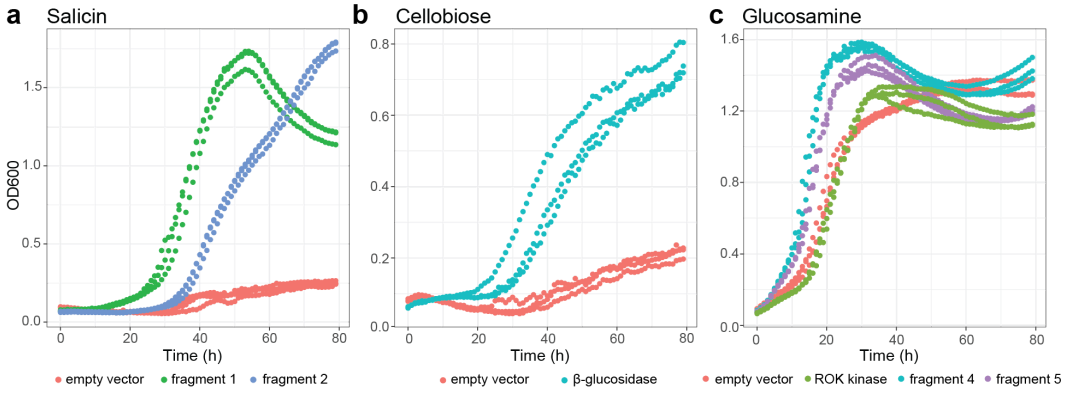

Figure S9. Growth of individual *B. theta* strains with increased fitness conferred by a gene hit. Each strain has a vector encoding either the exact insert fragment that conferred a fitness benefit in our screen, just the predicted causative gene, or an empty vector. See Supplementary Table 9 for precise genomic boundaries, orientation, and fitness scores for each fragment. Strains with β-glucosidase (BACUNI_RS01910) were grown in (A) D-salicin and (B) D-cellobiose. Strains with ROK kinase (INE72_RS00960) were grown in D-glucosamine. All strains were cultures in the VB defined minimal medium with 20 mM carbon substrate. OD values are pathlength-corrected and blank-normalized. Each strain was grown in replicates of 3.

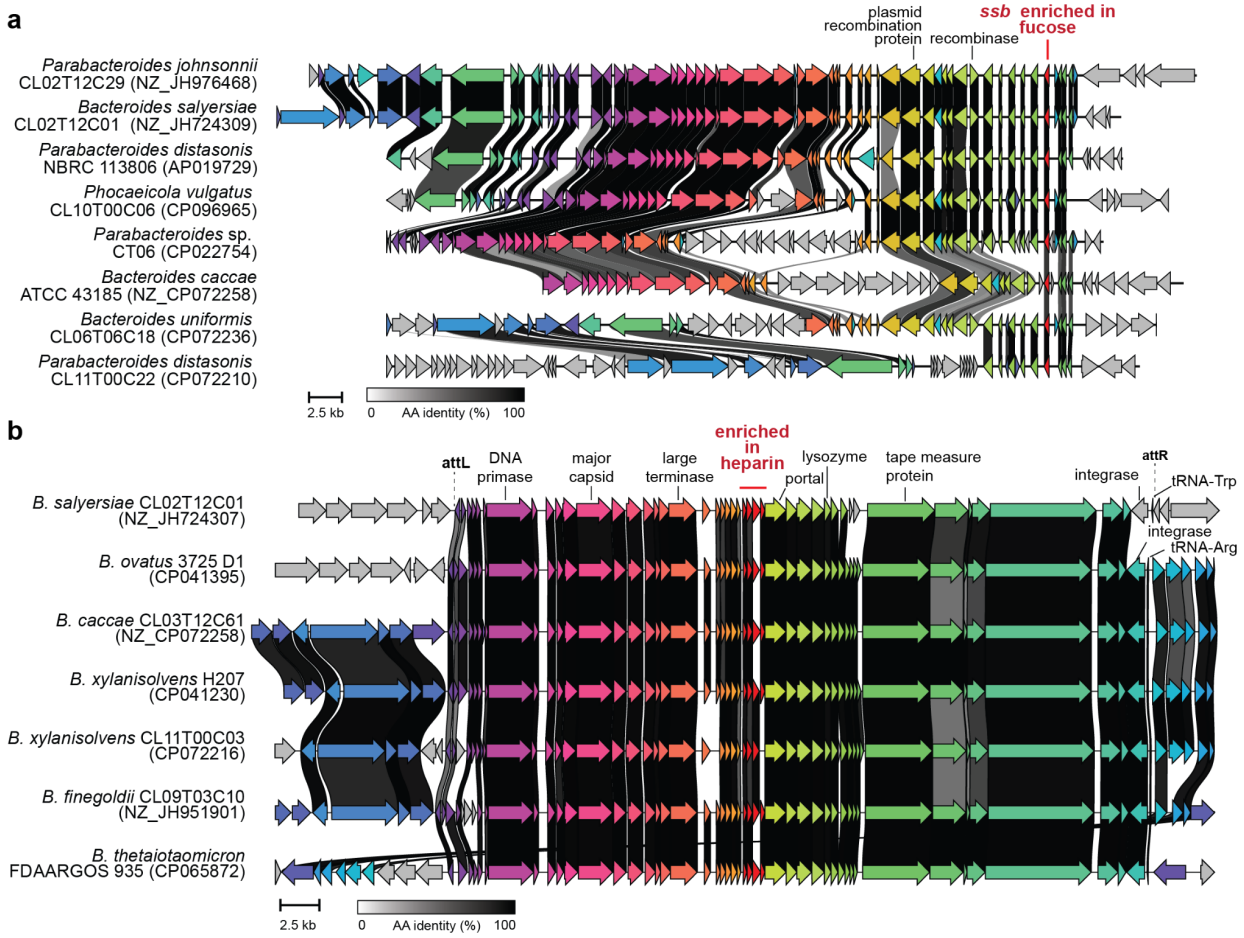

Figure S10. Annotation of mobile genetic elements containing gene hits identified in fucose and heparin conditions. A) Gene neighborhoods of an ssDNA-binding protein (*ssb*) enriched in fucose in Bacteroidales. Insertions and different rearrangements are observed downstream of *ssb*. B) A prophage encoded in *Bacteroides* species. Regions were identified via tBLASTn of gene hits encoding hypothetical proteins from heparin condition. Genomic comparisons were visualized using Clinker with an amino acid identity cutoff of 40%.[69]

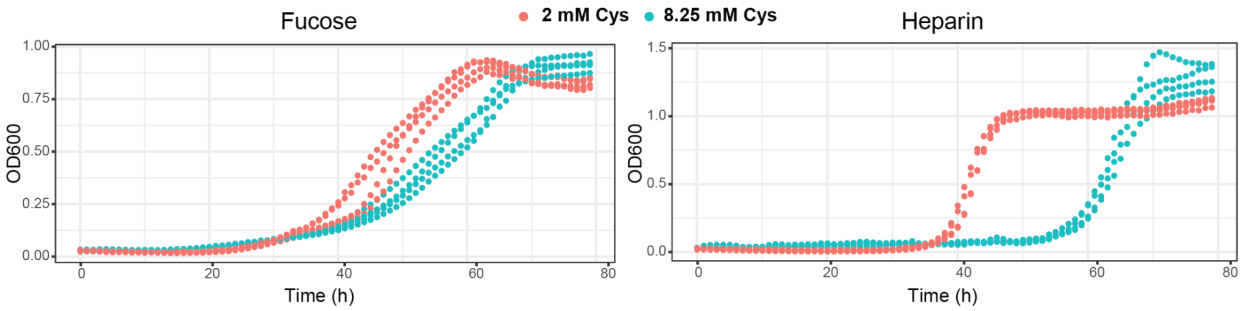

Figure S11. Growth of *B. theta* wild-type strain is inhibited by the high cysteine concentration (8.25 mM) in the VB defined minimal medium. Examples are shown for 20 mM fucose and 10 mg/mL heparin. A lower cysteine concentration (2 mM) resulted in improved growth. Each condition has replicates of 4.