Barcoded overexpression screens in gut Bacteroidales identify genes with new roles in carbon utilization and stress resistance

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 Y.Y.H. (yolandahuang@lbl.gov) and A.P.A. (aparkin@lbl.gov)

Abstract

A mechanistic 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 functional discovery, their applications to study gut anaerobes have been limited. For instance, most gain-of-function screens of gut-derived genes have been performed in *Escherichia coli* and assayed in a small number of conditions. To address these challenges, we developed Barcoded Overexpression BActerial shotgun library sequencing (Boba-seq). We demonstrate the power of this approach by assaying genes from diverse gut Bacteroidales overexpressed in *Bacteroides thetaiotaomicron*. From hundreds of experiments, we identified new functions and novel phenotypes for 29 genes involved in carbohydrate metabolism or tolerance to antibiotics or bile salts. Highlights include the discovery of a D-glucosamine kinase, a raffinose transporter, and several routes that increase tolerance to bile salts through lipid biosynthesis. This approach can be readily applied to develop screens in other strains and additional phenotypic assay types.

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 among microbes in this ecosystem. Despite their importance for human health, we lack a detailed mechanistic understanding of microbial functions in the gut.[5,6] The vast majority of bacterial genes have not been experimentally characterized, which in combination with their high functional diversity, has led to a knowledge gap in the field.[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. Moreover, DNA barcoding of individual strains allows for parallel analysis of strain fitness in genome-wide libraries through barcode sequencing (BarSeq).[9] In recent years, randomly barcoded transposon sequencing (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 human 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 *E. coli* as the overexpression strain, which is phylogenetically divergent from members of the abundant gut phyla Bacteroidota (formerly Bacteroidetes) and Bacillota (formerly Firmicutes).[13–18] In particular, promoter recognition of *Bacteroides* genes in *E. coli* can be limiting and *Bacteroides* encode unique ribosomal binding sites (RBSs), which can present barriers for protein expression in *E. coli*.[19–21] Therefore, Bacteroidales serve as more suitable overexpression 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 assays, which has limited these approaches to a small number of conditions. Random DNA barcoding of libraries can drastically increase the scale of overexpression screens. For example, dual barcoding of vectors (Dub-seq) has been used to construct libraries expressed in *E. coli*, but these are currently limited to ~250,000 strains and are restricted to DNA sources with sequence assemblies, which are required for mapping.[24–27]

Here, we exploit advances in long-read sequencing to develop a new workflow termed Boba-seq, or Barcoded Overexpression BActerial shotgun library sequencing, to perform large-scale gain-of-function screens using DNA barcoded vector libraries. We demonstrate this with overexpression in *B. theta*, although this approach is readily applicable to alternative expression strains and is agnostic to the source of DNA being assayed. Bacteroidales are major constituents of the gut microbiota that establish stable, long-term associations with the human host. They play crucial ecological roles in the gut, including metabolism of diverse, host-recalcitrant carbohydrates and modulation of the human immune system.[28]

We constructed Boba-seq libraries from seven Bacteroidales and first validated our method in a complementation screen. We then assayed a pool of six libraries in competitive fitness assays and identified genes involved in the metabolism of 15 carbon substrates and tolerance to seven inhibitory compounds (antibiotics or bile salts). Genes with novel functions or phenotypes included enzymes, transporters, and regulators. In addition to gene function discovery, our work enables large-scale gain-of-function screens in a clade of gut bacteria important for human health.

Results

Overview of workflow to build and assay barcoded overexpression libraries

The Boba-seq workflow utilizes a single DNA barcode per plasmid to increase the throughput of phenotypic assays. This approach consists of five general steps (Fig. 1). First, we construct replicative vectors to encode an inducible promoter, an insertion site for DNA of interest, a terminator, a DNA barcoding site, and an origin of replication for the overexpression host. Empty vectors are barcoded by cloning in random 20 nucleotide barcodes.[29] In the second step, randomly sheared and size-selected DNA fragments from the organism(s) of interest are cloned into the barcoded vectors. Here, fragment sizes can be optimized to capture single genes or larger gene clusters that might encode the target phenotypes. Next, fragments are linked to unique barcodes using long-read sequencing. This simplifies the mapping step to a single PCR amplification step instead of the complex TnSeq-like protocol that was used previously to map dual-barcoded vectors.[24] In step four, libraries are conjugated into the overexpression strain. In the final step, libraries are screened across a panel of conditions where the inoculum (Time0) and samples grown in selective media are harvested for BarSeq. Inserts that lead to increased strain fitness are identified by calculating fitness scores for each barcode, which represents the log2 change in the relative abundance of the strain. Fitness scores are then filtered to identify potential causative genes for the observed fitness gain.

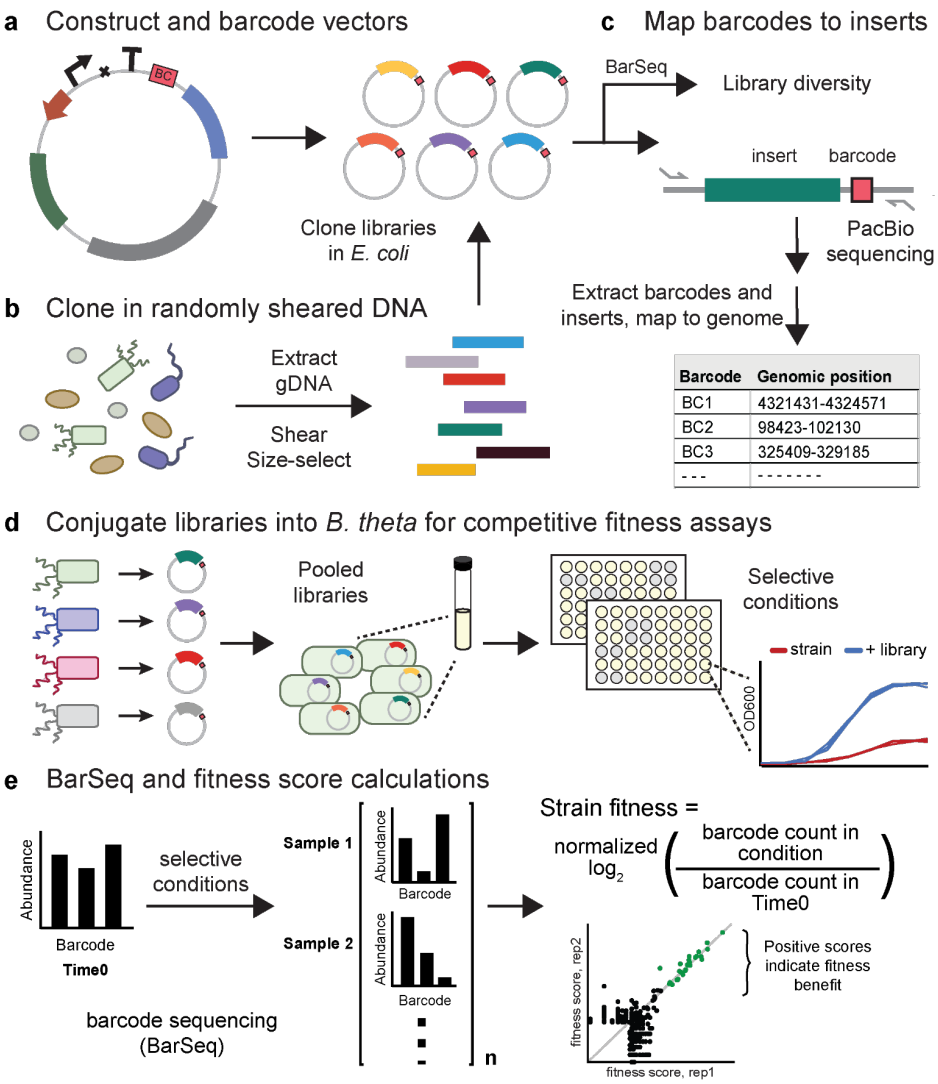


Figure 1. The Boba-seq workflow for high-throughput functional screens. A) Clone DNA barcodes into replicative vectors for the overexpression host. B) Extract DNA, shear, size-select, and clone into barcoded vectors. C) Quantify library diversity using BarSeq. Map barcodes to inserts using long-read sequencing technology and generate a barcode-to-insert mapping table. D) Conjugate mapped libraries into the overexpression strain for fitness screens. Libraries can be pooled to increase the genetic 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 before and after growth on selective conditions using BarSeq. A fitness score is calculated for each barcode or strain. Significantly high fitness scores in both replicates point to putative beneficial genes in the condition.

Construction of DNA barcoded overexpression libraries from Bacteroidales genomic DNA

We generated and characterized new shuttle vectors for *B. theta* VPI-5482 (see Methods).[30–33] This led to three replicative vectors: pNBU2_repA1, pNBU2_repA2, and pNBU2_repA3 (Fig. 2a,

Extended Data Table 1). Our vectors contain an anhydrotetracycline (aTc) inducible promoter, a ribosomal binding site (RBS), and a terminator.[33–35] Inducible gene expression by aTc was verified using NanoLuc reporter protein (Extended Data Fig. 1a). The plasmid copy number for these vectors in *B. theta* was determined to be around 3, 7, and 8, respectively, using droplet digital PCR (Extended Data Fig. 1c). Collectively, these vectors are replicative in 10 of the 13 Bacteroidales tested, including *B. theta* (Extended Data Fig. 1d). The vector pNBU2_repA1 was then barcoded to contain approximately 14 million unique barcodes (see Methods). Insert lengths of ~3 kb were 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*, *Parabacteroides johnsonii*, and *Parabacteroides merdae*, which share a maximum of 92% average nucleotide identity. Phylogenetic relationships between these seven strains are shown in Fig. 2b. 81% of all protein-coding genes in these genomes have at least one protein homolog (≥40% a.a. identity, ≥75% coverage), either in the same genome or in another genome.

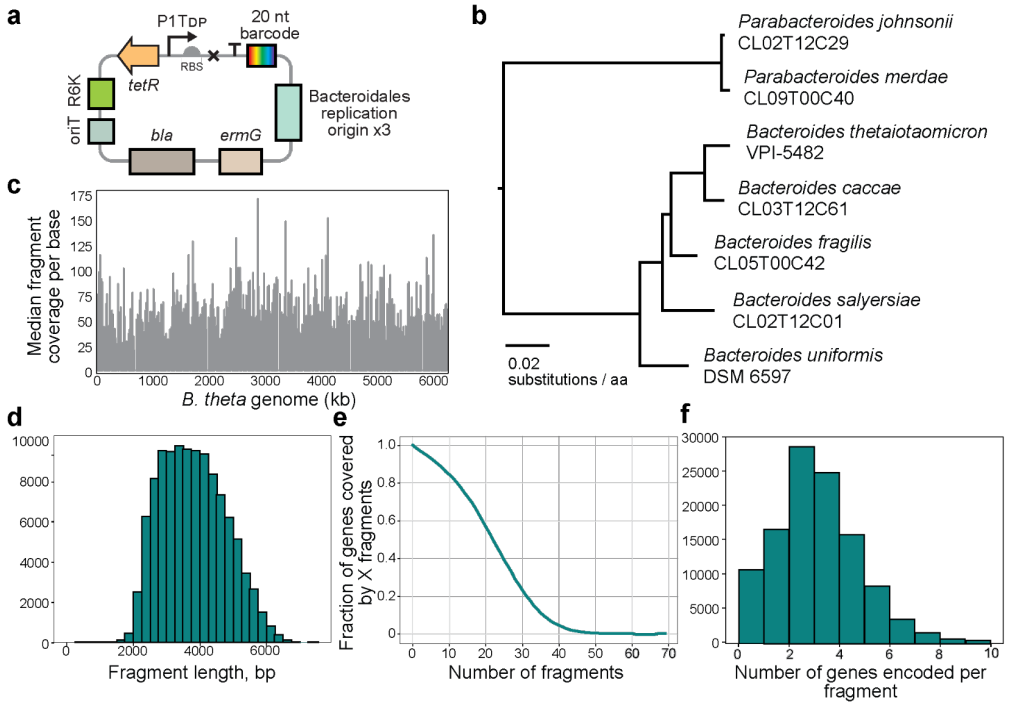


Figure 2. Boba-seq libraries were constructed from gDNA of *Bacteroides* and *Parabacteroides* isolates with high genomic coverage. A) Three replicative vectors were constructed by modifying an integrative vector (pNBU2-ErmG).[33] B) 24 universal bacterial marker genes from the Genome Taxonomy Database (GTDB) were used to generate this phylogenetic tree to show phylogenetic relatedness of the seven isolates used to build Boba-seq libraries.[36] Protein sequences of marker genes were aligned with HMMer 3.3.1 (hmmer.org) and then a tree was inferred with FastTree 2.1.11[37] and midpoint rooted. Coverage of the *B. theta* genomic library: C) fragment coverage per base across the genome (NC_004663.1) 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.

141 Next, we developed a pipeline to link barcodes to fragments on each vector (Extended Data Fig.
142 2). We performed PacBio amplicon sequencing the insert-barcode region to obtain 10–24-fold
143 sequencing depths (Extended Data Table 2). 77–91% of barcodes detected by BarSeq were
144 mapped, hence only a small fraction of each library is not analyzed. On a per genome basis,
145 fragments cover 87–98% of protein-coding genes with 76–477 protein-coding genes unmapped
146 (Extended Data Table 2, Fig. 2c). In all libraries, the missing proteins are generally longer than
147 mapped proteins (median 1,758 vs. 927 bp) and 5% have a nearly-identical copy (≥95% a.a.
148 identity) elsewhere in the genome. Overall, 50% of the missing proteins are over 2 kb in length
149 or are duplicated, which reflects limitations imposed by fragment sizes and difficulties in
150 mapping highly similar regions. After excluding genes of over 1.5 kb or with duplicates, 31% of
151 the remaining genes have a homolog in the same or another library that is also missing
152 (Extended Data Table 3). Conserved missing genes suggest possible toxicity in *E. coli* and
153 include proteins that are resistant to horizontal gene transfer such as ribosomal subunits,
154 although the majority cannot be readily explained based on the annotation.[38]

155 Libraries consist of barcode diversities ranging from 40,911–97,471 and median fragment
156 lengths of 2.7–3.7 kb (Fig. 2d, Extended Data Fig. 3a, Extended Data Table 2). The majority of
157 fragments encode 1–2 complete genes and the median fragment coverage per gene is 12–24
158 (Fig. 2e,f, Extended Data Fig. 3b,c, Extended Data Table 2). Overall, 50,000 fragments of ~3.1
159 kb are sufficient to cover a bacterial genome of approximately 5–6.5 Mbp with a median
160 fragment coverage of >10. A sequencing depth of >10× with long reads can map a majority of
161 barcodes. In summary, we constructed seven libraries with high genome-wide coverages.

162 **Complementation assays as a proof-of-concept**

163 To validate our workflow in a phenotypic screen, we complemented *B. theta* ΔBT2158 and
164 ΔBT3703 with *B. theta, B. fragilis, B. uniformis*, and *P. johnsonii* libraries.[11] BT2158
165 (BT_RS10925) encodes a periplasmic glycoside 3-dehydrogenase required for disaccharide
166 and glucosinolate metabolism.[11,39] ΔBT2158 does not grow on trehalose (glucose-α-1,1-
167 glucose), leucrose (glucose-α-1,5-fructose), or palatinose (glucose-α-1,6-fructose) (Extended
168 Data Fig. 4). The second mutant lacks the α-glucosidase SusB (BT3703/BT_RS18670) part of
169 the starch utilization system and has reduced growth on α-cyclodextrin compared to wild-type.[11]
170 We then inoculated each library in a defined minimal medium (Varel-Bryant or VB) containing a
171 single carbon substrate in the presence of the inducer aTc (Extended Data Table 4).[40] Among
172 these four source genomes, three encode BT2158 or a homolog, and three encode BT3703 or a
173 homolog, which are expected to provide fitness gain in the conditions tested.

174 We computed fitness scores for each fragment using the log2 fold-change of each barcode's
175 relative abundance between the selective condition and the Time0. We focus on high positive
176 scores because Boba-seq is a gain-of-function assay and we established filters to identify
177 biologically consistent hits. First, fragments with a fitness score ≥5 and a z score ≥4 in both
178 replicates are considered statistically significant. Using these thresholds, no significant scores
179 were found in a control comparison of Time0 samples (see Methods). Second, we consider
180 each hit to be biologically consistent if it is confirmed by overlap or by protein similarity.
181 Specifically, a hit is biologically consistent if 1) there is an overlapping fragment with significant

182 high scores or 2) there is a protein homolog (≥40% identity) from a different genomic region
183 (from within the same library or a different library) with a significant fitness score. Finally, we
184 inspect the gene(s) covered by these fragments to identify the causative gene. This led to 16
185 biologically consistent region by condition pairs encoding 15 unique proteins that conferred a
186 fitness gain in leucrose, palatinose, or α-cyclodextrin (Extended Data Table 5).

187 For libraries expressed in ΔBT2158, growth in leucrose or palatinose led to the expected hits
188 encoding BT2158, its nearly-identical paralog BT4448 (BT_RS22435), or a homolog
189 (BACUNI_RS05440) from *B. uniformis* (Fig. 3a). In trehalose, we observed extremely high
190 fitness scores (>12) for strains encoding each of these genes, but each strain had high fitness in
191 just one replicate, so they did not meet our criteria for statistical significance. These strains likely
192 grew quickly early on in one replicate, but were outcompeted by other fit strains in the second
193 replicate. To verify our results, we isolated clones encoding either BT2158, BT4448, or
194 BACUNI_RS05440 and observed improved growth on trehalose (Fig. 3b). Of note, three of six
195 fragments encode genes oriented opposite to the inducible promoter with <50 bp upstream and
196 were likely transcribed from a weak, unexpected promoter located downstream of the vector
197 insert site (Extended Data Table 6, Extended Data Fig. 1b).

198 Previously, it was observed that the ΔBT2158 mutant cannot grow on these disaccharides,
199 suggesting that BT2158 is essential for metabolism and BT4448 cannot rescue its activity.[11] In
200 contrast, our data show that BT4448 can functionally complement ΔBT2158 when expressed
201 from a vector. BT4448 is likely not expressed to sufficient levels in these conditions to rescue
202 growth in ΔBT2158 and that overexpression from our vector led to higher activity.

203 Complemented ΔBT3703 libraries in α-cyclodextrin led to a cysteine desulfidase (CdsB) and a
204 cysteine β-lyase as beneficial hits (Fig. 3c, Extended Data Table 5). Here, the selection for
205 cysteine degrading enzymes appears to be stronger than for BT3703 and its homologs since
206 ΔBT3703 is capable of growth on α-cyclodextrin.[11] Cysteine degradation is likely beneficial due
207 to toxicity of the 8.25 mM cysteine in the minimal medium (Extended Data Fig. 5).[41,42] This is
208 also supported by RB-TnSeq data for *B. theta*, where disruption of CdsB
209 (BT2080/BT_RS10540) led to reduced fitness across all conditions in the same minimal medium
210 (Fig. 3d).[11] Similarly, cysteine β-lyase mutants of *B. theta* exhibited mild fitness defects in many
211 conditions.[11] Of note, the rich medium used for *B. theta* RB-TnSeq fitness assays also contains
212 8.25 mM cysteine, and thus this effect is specific to minimal media conditions (Fig. 3d).[11]
213 Cysteine is commonly added to Bacteroidales growth media as a reductant, but our result
214 cautions against its use at a high concentration in minimal media.

215 In summary, we established a set of stringent filters to identify beneficial hits. In particular,
216 complementation worked well for conditions with strong selective pressure where the host strain
217 exhibits little or no growth.

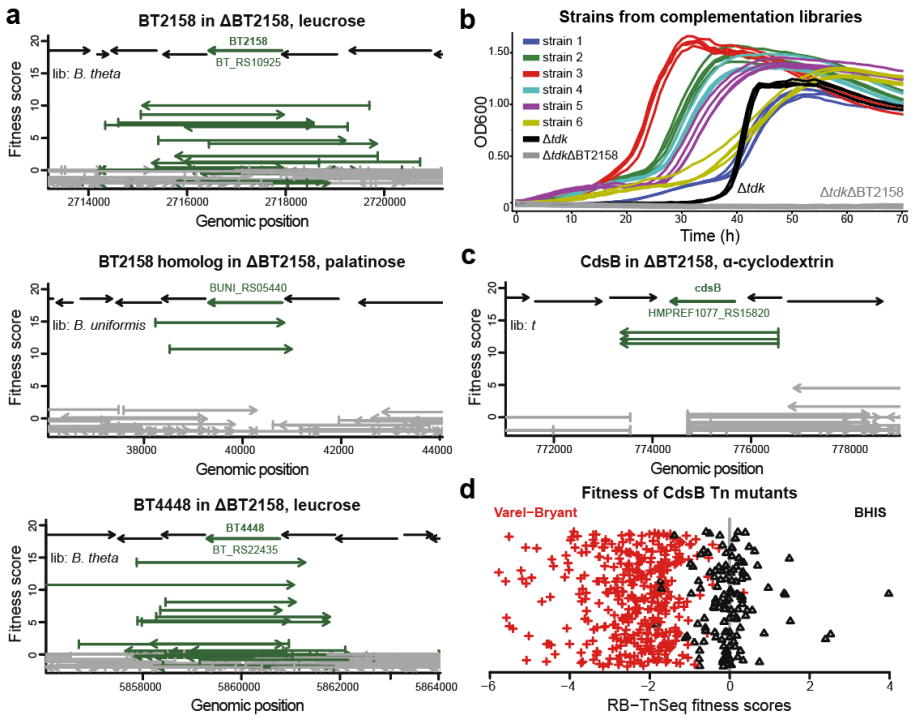


Figure 3. Fitness data from complementation assays. BT2158, its homologs, and cysteine degrading enzymes were enriched in complementation assays. The average fitness score from both replicates is plotted for each fragment. Fragments that cover the complete gene are indicated in green. A) BT2158 and BT4448 in *B. theta*, and a homolog (BACUNI_RS05440), were beneficial for the ΔBT2158 mutant during growth in leucrose and palatinose. B) Growth curves of ΔBT2158 mutant complemented with either BT2158 or a homologous hit. Complemented isolates, the Δ*tdk* parental strain, and the Δ*tdk*ΔBT2158 mutant were cultured in the VB medium with 20 mM trehalose. Strains 1–2 encode BT2158, strains 3–4 encode BT4448, and strains 5–6 encode BACUNI_RS05440. See Extended Data Table 6 for the precise genomic boundaries of cloned fragments. Each strain was grown in replicates of 4. OD600 values are pathlength-corrected and blank-normalized. C) CdsB was enriched in the ΔBT3703 mutant in α-cyclodextrin. D) RB-TnSeq fitness scores for CdsB (BT2080/BT_RS10540) are averages across all 37 CdsB transposon mutants in the *B. theta* library and are log2 fold-changes that were previously reported (x-axis).[11] Scores are highlighted in red for assays in the VB minimal medium and in black for the BHIS rich medium. The y-axis is random scatter 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 using wild-type *B. theta* as an expression host. In total, we performed 222 assays consisting of 42 single carbon sources in a defined minimal medium and 16 inhibitory compounds at 2–5 concentrations in a rich medium (Extended Data Table 4). The vast majority of assays were carried out in the presence of the inducer aTc. Carbon sources included dietary and host-derived complex polysaccharides, gelatin, mucins, and simple carbohydrates. Inhibitory compounds consisted of antibiotics and

243 bile salts. Six libraries (excluding the *B. theta* library) were pooled into a single inoculum (or
244 Time0). Growth was observed in 29 carbon substrates and all stress conditions. 192 samples
245 were processed for BarSeq and included two replicates per condition and two Time0 samples
246 per fitness screen.

247 Of the 305,382 uniquely-mapped barcodes present in the six libraries in the conjugation donor
248 strain, we detected 283,816 barcodes in our fitness assays (having at least 2 reads across all
249 assays and controls). Across all libraries, 21,042 protein-coding genes were assayed. We
250 identified 1,172 barcodes with statistically significant scores using the same criteria as
251 previously detailed for the complementation screen (Fig. 4a). We identified biologically
252 consistent fragments for 120 unique regions across 22 compounds. The majority of beneficial
253 regions were curated to assign a putative causative protein or proteins by inspecting fitness
254 score vs. fragment coverage plots (Extended Data Fig. 6). Experimental evidence from any
255 characterized homolog was assessed using PaperBLAST and taken into consideration as
256 well.[43] Overall, hits across all conditions consist of 76 protein clusters with 30 clusters containing
257 multiple homologs at ≥40% amino acid identity (Extended Data Table 5).

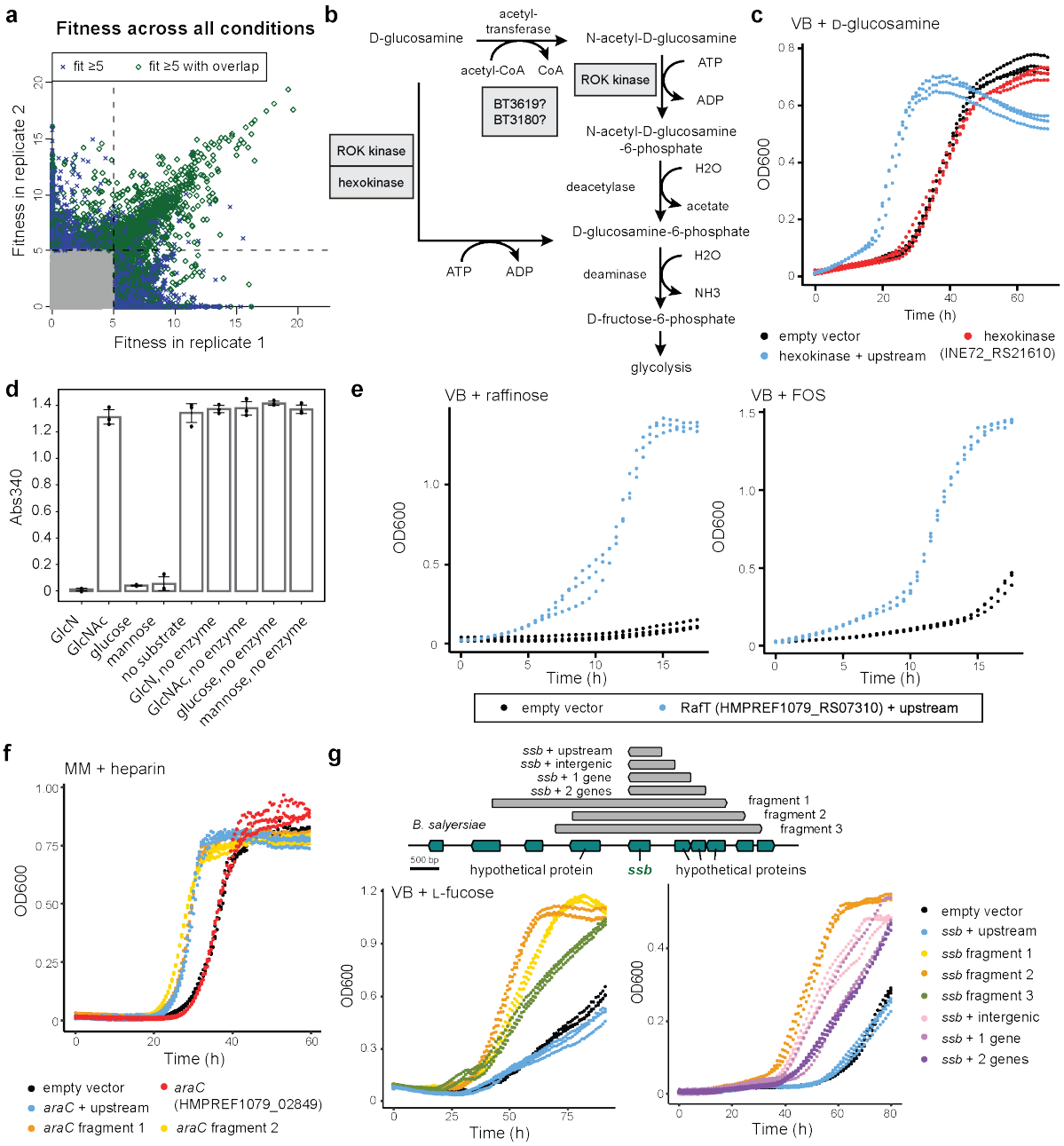


Figure 4. Competitive fitness assays in wild-type *B. theta* uncovered novel functions in carbon utilization. A) Consistency of strain fitness values between two replicates across 93 pairs of experiments with 6 pooled libraries. Grey points represent fitness scores <5. Dashed lines separate fitness scores of ≥5. Barcodes with fitness scores ≥5 whose regions overlap with other high-scoring barcodes are highlighted in green. B) D-Glucosamine and GlcNAc share a metabolic pathway in Bacteroidales. ROK kinase and hexokinase hits from our screen are displayed based on biochemical evidence. C) A hexokinase conferred a growth advantage on D-glucosamine when overexpressed in *B. theta*. D) *In vitro* end-point enzyme assays coupled to NADH depletion were performed for hexokinase to reveal a substrate preference for D-

268 glucosamine (GlcN), glucose, and mannose. E) *B. theta* expressing RafT is more fit growing on
269 raffinose and FOS as the sole carbon source. F) An AraC regulator provides a fitness gain on
270 heparin. G) A genomic region containing a ssDNA-binding protein (HMPREF1071_RS19060) is
271 beneficial on L-fucose. For all growth curves, individual strains were cloned to identify the
272 causative region and constructs with "+ upstream" indicate the inclusion of the upstream 200 bp.
273 The VB minimal medium was used for each condition except for heparin where a different
274 defined minimal medium (MM)[44] was used (Extended Data Table 4). See Extended Data Table
275 6 for the genomic boundaries of library fragments verified by growth assays of individual strains.

276 Consistent with the complementation screen, we detected high fitness scores for CdsB and
277 cysteine β-lyase across 9 conditions containing cysteine (Extended Data Table 5). In addition,
278 21 carbon substrates were tested using an alternative reductant, dithiothreitol (DTT). While most
279 of these assays led to similar hits, after excluding Cys-degrading enzymes, assays with DTT
280 generally led to more hits compared to assays with cysteine.

281 We noticed that a large number of strains did not show a benefit even though they encode a
282 curated causative gene or genes (Extended Data Table 5). Specifically, 31% of strains that
283 contain the putative causative gene(s) show a significant benefit (fitness score ≥5 and z ≥4 in ≥1
284 replicate). This is likely partly due to differences in gene expression caused by variations in
285 fragment boundaries: for strains with the exact same insert sequence where at least one strain
286 has a statistically significant score, 70% have a significant score in at least one replicate. This is
287 much higher than the 31% rate for all strains with causative genes. Our vector's inducible
288 promoter was designed to ameliorate expression issues, but many genes are likely transcribed
289 from their native promoters, which are not always included in fragments. Moreover, most
290 protein-coding genes in Bacteroidales do not use a Shine-Dalgarno sequence for translation
291 initiation, therefore the vector's promoter or RBS could interfere with translation.[21,45]

292 **High-confidence hits uncover metabolic enzymes in carbohydrate utilization**

293 Here, we focus on metabolic enzymes beneficial on single carbon sources. Three substrates led
294 to expected hits based on known functions or annotations (Extended Data Table 5). These
295 include the L-fucose catabolic operon from L-fucose and a β-glucosidase from D-cellobiose and
296 D-salicin (Extended Data Fig. 6).[11,46] We verified that the β-glucosidase confers a growth benefit
297 when expressed in *B. theta* on D-cellobiose and D-salicin (Extended Data Fig. 7a).

298 Additionally, an α-glucosidase (INE72_RS06280) in the glycosyl hydrolase family 31 was
299 beneficial in palatinose in both the pooled library assays and when overexpressed in individual
300 strains (Extended Data Fig. 6, 7b). Since palatinose contains an α-1,6-glycosidic bond, we
301 annotate this enzyme as an α-1,6-glucosidase.

302 An ROK family (repressor, open reading frame, kinase) protein and a hexokinase were both
303 important for growth on D-glucosamine. We first show that strains encoding the ROK kinase or
304 the hexokinase have a growth advantage in D-glucosamine (Fig. 4c, Extended Data Fig. 7c). In
305 support of our data, a ROK kinase homolog in *B. theta* (31% identity) can phosphorylate D-
306 glucosamine and N-acetyl- D-glucosamine (GlcNAc) *in vitro*.[47]

307 In *B. theta*, D-glucosamine and GlcNAc share a metabolic pathway where D-glucosamine can be
308 acetylated to form GlcNAc (Fig. 4b).[48] GlcNAc is then phosphorylated and deacetylated to form
309 D-glucosamine-6-phosphate.[49] While two kinases were beneficial for growth on D-glucosamine,
310 no hits were observed in GlcNAc, which may be a result of lower selective pressure since *B.*
311 *theta* grows better on GlcNAc than D-glucosamine in the VB medium. The hexokinase is
312 important for glucose and mannose metabolism as previously reported using a *B. fragilis*
313 deletion mutant (AY664812, 79% identity).[50] While our fitness data support D-glucosamine as an
314 additional substrate, we cannot exclude the possibility of GlcNAc phosphorylation by this
315 hexokinase since it is an intermediate of the metabolic pathway (Fig. 4b). Therefore, we purified
316 the hexokinase (INE72_RS21610) to establish its activity in a coupled enzyme assay. The
317 hexokinase has a strong substrate preference for glucose, mannose, and D-glucosamine, with
318 little or no activity toward GlcNAc (Fig. 4d). Here, we use both genetics and biochemical
319 approaches to identify a new D-glucosamine kinase and demonstrate its importance in D-
320 glucosamine metabolism in Bacteroidales.

321 **TonB-dependent transporters mediate raffinose uptake**

322 The TonB-dependent transporter (TBDT) family can carry out active transport of vitamins,
323 siderophores, peptides, and glycans across the outer membrane, and are prevalent in the
324 phylum Bacteroidota, with 121 putative TBDTs in *B. theta* alone.[51] Two TBDTs were uncovered
325 to be beneficial for growth in raffinose and fructooligosaccharides (FOS) (Extended Data Table
326 5, Extended Data Fig. 6). To our knowledge, this is the first report of raffinose transporters in
327 Bacteroidales. Raffinose is a trisaccharide part of the raffinose family of oligosaccharides
328 (RFOs) that is widespread in dietary plants and abundant in the standard mouse chow used in
329 microbiome studies.[52,53] A distantly-related TBDT from *B. theta* (BT1763/BT_RS08935, ≤28%
330 identity) is involved in the uptake of FOS with β-2,6 linkages, but the FOS substrate used in our
331 study contains predominantly β-2,1-linkages (Extended Data Table 4).[54]

332 We name these transporters RafT as they share similarity at 49% identity
333 (HMPREF1077_RS18310, HMPREF1079_RS07310), but have distinct gene neighborhoods. An
334 α-galactosidase is present downstream of *Pj*-RafT (from *P. johnsonii*) and can likely release
335 galactose from raffinose.[55] A β-D-fructofuranosidase is genomically adjacent to *Bf*-RafT (from *B.*
336 *fragilis*) and can hydrolyze raffinose.[56] Both nearby enzymes are predicted to be periplasmic,
337 which supports a direct uptake of raffinose by the RafTs. This is also supported by RB-TnSeq
338 data for a RafT homolog in *Phocaeicola vulgatus* (HMPREF1058_RS00615, >47% identity) that
339 is important for raffinose utilization (unpublished data). In addition to raffinose, *Bf*-RafT was also
340 a hit in FOS and in the complementation screen for leucrose. Leucrose, raffinose, and FOS all
341 contain D-fructose as a terminal residue, which might mediate substrate recognition by *Bf*-RafT.
342 We confirmed that *Bf*-RafT is beneficial for *B. theta* grown on raffinose and FOS (Fig. 4e).
343 Collectively, the combination of data from Boba-seq libraries, RB-TnSeq libraries, gene
344 neighborhoods, and growth experiments, support raffinose uptake as a function of these
345 TBDTs.

346 **A regulator confers fitness in growth on heparin**

347 In addition to genes involved in metabolic and transport pathways, our results point to regulatory
348 mechanisms as well. An AraC family transcriptional regulator (HMPREF1079_RS07695)
349 conferred a fitness benefit on heparin in both pooled fitness assays and individual strain growth
350 assays (Fig. 4f, Extended Data Fig. 6). This protein has been associated with resistance to DNA
351 damaging agents, but a role in heparin utilization or more broadly, carbohydrate metabolism,
352 had not been reported.[57,58]

353 **A mobile genetic element contributes to competitive growth on L-fucose**

354 L-Fucose is a key nutrient for Bacteroidales in the gut with sources including dietary
355 polysaccharides and host glycoproteins such as mucins. In addition to being a carbon source, L-
356 fucose is a component of bacterial capsular polysaccharides and fucosylated glycoproteins.[59]
357 During nutrient-limiting conditions, *B. theta* can induce fucosylation of host glycans to increase
358 nutrient availability for resident microbes, thus altering gut microbiota composition.[60,61] Given the
359 importance of this molecule for host-microbe interactions, we sought to identify genes that
360 provide a competitive growth advantage on L-fucose.

361 Here, L-fucose catabolic genes were beneficial across all L-fucose conditions (Extended Data
362 Table 5). Unexpectedly, a genomic region encoding a ssDNA-binding protein (SSB) was also
363 detected among all L-fucose conditions (Extended Data Fig. 8a). Additionally, there are three
364 hypothetical proteins upstream of *ssb* in this conserved region, which is part of a ~50 kb
365 putative mobile genetic element (MGE) present in Bacteroidales (Extended Data Fig. 8b). In
366 support of this, a conserved recombinase and resolvase are encoded nearby. First, we verified
367 our fitness data using three high-scoring fragments and observed growth benefits on L-fucose
368 (Fig. 4g). Then, by using truncated fragments, *ssb* and the upstream intergenic region appears
369 sufficient for a partial fitness gain (Fig. 4g).

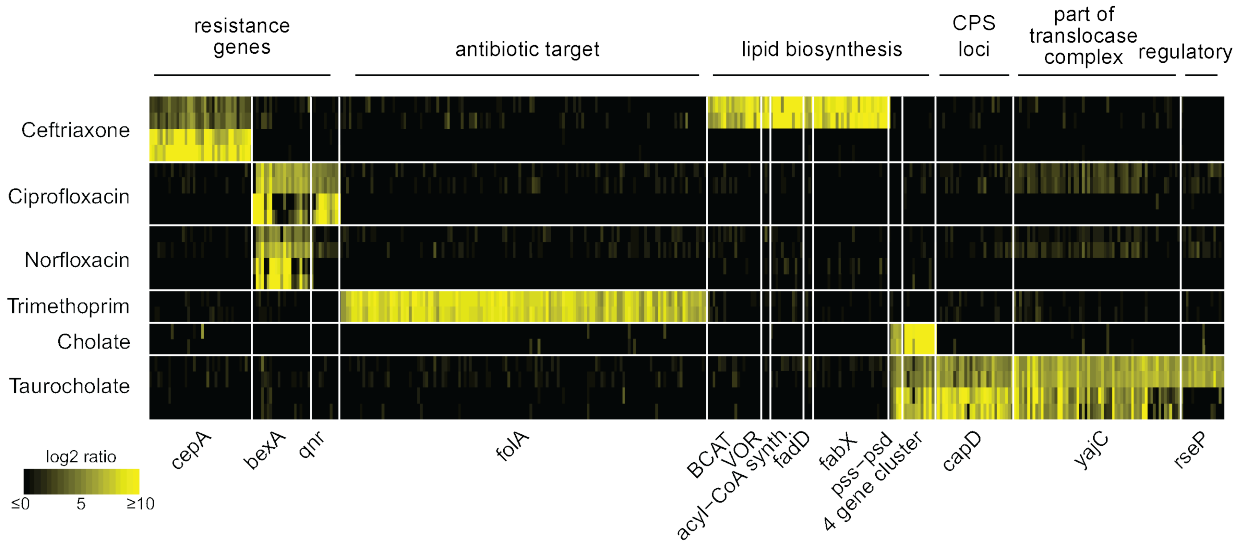
370 SSB proteins play important roles in DNA replication, recombination, and repair, and have
371 multiple protein binding partners.[62] Additionally, they are present on plasmids and play anti-
372 defense roles by inhibiting SOS response and by protecting transferred ssDNA from nuclease-
373 catalyzed degradation.[63] Based on known functions of SSB, we could not readily propose a
374 mechanism to account for the observed growth advantage on L-fucose. We also cannot exclude
375 the possibility that the upstream intergenic region might contribute to the phenotype. Therefore,
376 in an attempt to gain molecular insights, we performed an RNA-seq experiment using the most
377 fit strain encoding a library fragment and a strain with an empty vector (Extended Data Fig. 8c,
378 Extended Data Table 7**,** *p*-value <0.05, Wald test). Transcripts for L-fucose catabolic genes are
379 not differentially abundant between these two strains, which rules out a direct regulatory
380 mechanism by which this region is conferring a benefit.

381 A partial restriction endonuclease subunit S (BT4522/BT_RS22805) was strongly upregulated in
382 the strain expressing the *ssb* fragment (log2 fold-change or log2FC of 5.5). This is caused by
383 phase-variation in the Type I restriction modification system and is likely catalyzed by the
384 adjacent recombinase. Based on transcript read alignments, the partial S subunit gene
385 recombined with an upstream S subunit gene such that it became expressed as part of the new

386 S subunit (Extended Data Fig. 8d). We also observed expression differences in capsular
387 polysaccharide (CPS) biosynthesis, with the CPS3 locus upregulated (average log2FC of 1.46)
388 and the CPS5 locus repressed (average log2FC of -4.6). Both CPS loci are regulated by
389 invertible promoters in addition to the upxY transcriptional anti-terminator and upxZ *trans* locus
390 inhibitor system.[64] Consistent with its known role, we speculate that the SSB affects
391 recombination events and altered phase-variation at these loci. However, we cannot pinpoint a
392 direct mechanism by which these transcriptional changes might impart a benefit on L-fucose.
393 Despite this, we uncovered a novel link between this conserved region within a MGE and a
394 growth advantage on L-fucose.

395 **Multiple strategies confer resistance to antibiotics**

396 The gut microbiota is under dynamic selection pressure from host-secreted factors, dietary
397 compounds, and xenobiotics. Therefore, we examined genes important for fitness in the
398 presence of different classes of antibiotics, bile salts, and an antioxidant (Extended Data Table
399 4). Seven inhibitory compounds led to significant hits with greater fitness gains typically in the
400 highest concentrations (Extended Data Table 5).



401
402 **Figure 5.** Selected gene hits that conferred benefit on antibiotics and bile salts. Fragments
403 encoding these gene hits with significant high fitness scores from all Boba-seq libraries are
404 displayed as fitness values. Replicates are plotted separately for each compound at either one
405 or two concentrations. For compounds with hits at two concentrations, the lower concentration is
406 plotted above the higher concentration. See Extended Data Table 5 for assay conditions, locus
407 tags, and proposed mechanism of tolerance. Causative gene(s) across multiple libraries are
408 plotted within their respective protein clusters.

409 First, we identified three known genes (*cepA*, *bexA*, *folA*) important for antibiotic resistance
410 through antibiotic inactivation, drug efflux, or increased target expression (Fig. 5). Additionally, a
411 pentapeptide repeat-containing protein hit on ciprofloxacin is similar to a quinolone-resistance
412 protein (~30% identity, EF0905) that inhibits quinolone binding to DNA gyrase and
413 topoisomerase.[65]

414 Our data also associated five genes involved in the biosynthesis of branched-chain or
415 unsaturated lipids with ceftriaxone tolerance, a third-generation cephalosporin in clinical use
416 (Fig. 5, Extended Data Fig. 6). Collectively, a branched-chain amino acid aminotransferase
417 (BCAT) and a 2-ketoisovalerate oxidoreductase (VOR) catalyze the conversion of branched-
418 chain amino acids to α-ketocarboxylic acids, which are oxidatively decarboxylated to form
419 branched-chain acyl-CoAs. These can then be converted to acyl-ACPs and undergo elongation.
420 In support of this, the BCAT homolog in *B. fragilis* (BF9343_3671, ~88% identity) is involved in
421 the biosynthesis of branched-chain α-galactosylceramides, a type of immunomodulatory
422 sphingolipids.[66] We identified an acyl-CoA synthetase and a long-chain fatty acid-CoA ligase
423 (FadD) in this condition as well (Extended Data Fig. 6, Fig. 6a). Specifically, biochemical
424 characterization of a homolog of the acyl-CoA synthetase (Q8TLW1, ~58% identity) revealed
425 activity as a 2-methylbutyryl-CoA synthetase.[67] Finally, a beneficial nitronate monooxygenase
426 flavoprotein was identified to be FabX as it is similar to HP0773 (~40% identity) from
427 *Helicobacter pylori*.[68,69] FabX is essential for unsaturated fatty acyl biosynthesis and requires
428 molecular oxygen as the electron acceptor *in vitro*.[68,69] However, a homolog in *Neisseria*
429 *gonorrhoeae* (NGO1024, ~36% identity) is essential for anaerobic unsaturated fatty acyl
430 biosynthesis and relies on an unknown electron acceptor.[70] We cloned out BCAT, VOR, and
431 FabX and confirmed they are beneficial in individual growth assays (Fig. 6b, Extended Data Fig.
432 7d). Our results indicate that upregulating steps in branched-chain or unsaturated fatty acyl
433 biosynthesis increases tolerance to ceftriaxone. Changes in the outer membrane lipid
434 composition, including sphingolipids, is a potential strategy to reduce antibiotic concentrations in
435 the periplasm where β-lactams act. However, the precise mechanism by which this is achieved
436 warrants further work to decipher.

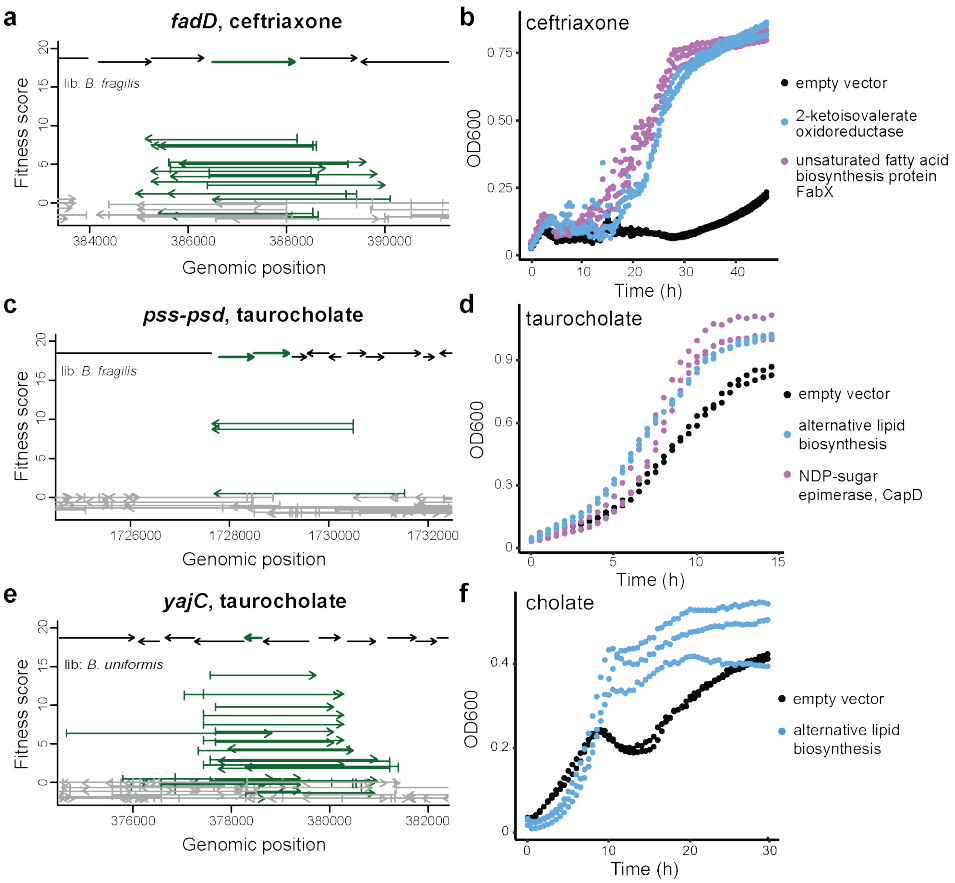


Figure 6. Genes involved in lipid biosynthesis, membrane protein translocation, and capsular polysaccharide biosynthesis increase tolerance to ceftriaxone, taurocholate, or cholate. Fragment plots for A) *fadD* (HMPREF1079_RS19625) in 0.5 mM ceftriaxone, C) phosphoethanolamine biosynthesis genes (HMPREF1079_RS14395, HMPREF1079_RS14390) in 10 mM taurocholate, and E) *yajC* (BACUNI_RS18600) in 10 mM taurocholate. Growth curves of individual hits expressed in *B. theta*: B) 2-ketoisovalerate oxidoreductase (HMPREF1077_RS19335, _RS17900, _RS17905, _RS17910) and *fabX* (HMPREF1077_RS10180) in 1 mM ceftriaxone; D, F) a gene cluster for alternative lipid biosynthesis (HMPREF1077_RS03620 to HMPREF1077_RS03635) in 5 mM cholate or 5 mM taurocholate; and D) a NDP-sugar epimerase *capD* (BACUNI_RS12910) in 5 mM taurocholate. All inserts include 200 bp upstream of the gene(s).

Genes in lipid and surface polysaccharide biosynthesis contribute to bile salt tolerance

Bile salts are host-secreted factors that have antibacterial activities and can modulate host signaling pathways to drive dynamic changes in the gut microbiome. These molecules act as detergents to disrupt and destabilize membrane integrity. In the colon, bacteria may encounter bile salts up to 1 mM and encode various strategies to counter their toxicity.[71] Here, we sought to identify factors that reduce susceptibility to bile stress in these microbes and thus may be important for colonization and competition in the human gut.

456 In both taurocholate and cholate, two beneficial loci are involved in lipid biosynthesis (Fig. 5).
457 First, a two-gene cluster encoding phosphatidylserine synthase (Pss) and phosphatidylserine
458 decarboxylase (Psd) catalyze the formation of phosphatidylserine and
459 phosphatidylethanolamine (Fig. 6c, Extended Data Fig. 6). These genes are essential and
460 cannot be studied using loss-of-function approaches, highlighting an advantage of using Boba-
461 seq. Second, a conserved cluster of four genes (adenylyltransferase/cytidyltransferase, short-
462 chain alcohol dehydrogenase/reductase, CDP-alcohol phosphatidyltransferase,
463 lysophospholipid acyltransferase) is absent in *B. theta* and likely involved in the biosynthesis of
464 an alternative lipid (Extended Data Fig. 6). This operon was enriched from the two
465 *Parabacteroides* libraries and is prevalent among *Parabacteroides*, but not *Bacteroides*. We
466 further confirmed its benefit for growth in taurocholate and cholate (Fig. 6d,f).

467 On the bile salt taurocholate, overexpression of RIP metallopeptidase RseP led to high fitness
468 (Fig. 5). This is likely mediated through sigmaE factor activation by RseP in response to
469 extracytoplasmic stress to induce transcriptional changes in cell surface proteins.[72] Additional
470 hits on taurocholate include the preprotein translocase subunit YajC and a NDP-sugar
471 epimerase CapD (Fig. 5, 6e, Extended Data Fig. 6). YajC is a small inner membrane protein
472 part of the translocase protein complex with SecDF, but it is not essential and its role is not well
473 understood.[73] In support of a role in responding to bile salts, protein levels of YajC in
474 *Campylobacter jejuni* were elevated during exposure to chenodeoxycholate, a bile salt.[74] NDP-
475 sugar epimerase is often part of CPS loci in *B. theta*.[75] The *cps1* locus in *B. theta* was induced
476 in a rich medium by taurocholate but not by cholate[76] and encodes a NDP-sugar epimerase
477 (BT0381, >60% identity) similar to hits in our screen. Here, we further demonstrate that
478 overexpressing the epimerase alone in *B. theta* is sufficient to increase tolerance to bile stress
479 (Fig. 6d). Overall, both hits are supported by published gene expression data and might induce
480 changes in membrane proteins or surface polysaccharides as a strategy to alleviate
481 taurocholate toxicity.

482 In summary, our results connect genes involved in lipid biosynthesis, a regulatory pathway,
483 membrane protein translocation, and surface polysaccharide biosynthesis to bile salt tolerance.
484 While a few hits have supporting gene expression data from other studies, most genes are
485 newly associated with this phenotype. Additionally, most studies quantify differentially abundant
486 transcripts or proteins to uncover pathways involved in response to bile stress.[74,76,77] Here, we
487 leveraged high-throughput genetics to directly demonstrate that additional gene copies or the
488 gain of a new gene cluster is sufficient to confer protection against bile salts.

489 **Discussion**

490 Boba-seq is a new platform for constructing barcoded overexpression libraries for scalable and
491 quantitative phenotypic screens. Here, we generated seven libraries using genomic DNA from
492 *Bacteroides* and *Parabacteroides* species, an important clade of human gut bacteria. First, we
493 validated our method in a complementation screen. Next, using wild-type *B. theta* as the
494 expression host, we identified novel phenotypes in carbon utilization and stress tolerance
495 across 120 unique regions and 22 compounds (Extended Data Table 5). By assaying DNA from

496 multiple bacteria, we identified 30 hits supported by multiple protein homologs, which provided
497 further confidence in our results.

498 Overall, we improved annotations and/or obtained new phenotypes for 29 protein clusters. In
499 particular, we identified an outer membrane transporter for the uptake of raffinose, a plant-
500 derived carbohydrate that is non-digestible by humans. This marks the first report of raffinose
501 transporters in Bacteroidales. Using biochemistry and genetics, we identified a hexokinase
502 critical for D-glucosamine metabolism. A region encoding a ssDNA-binding protein within a MGE
503 led to a genomic recombination event and to a growth advantage on L-fucose, a key nutrient in
504 the gut, although the exact mechanism remains unclear. Furthermore, we discovered genes that
505 enhance tolerance toward antibiotics and bile salts, prevalent stressors encountered in the gut.
506 Among these, genes involved in lipid biosynthesis, including a gene cluster absent in *B. theta*,
507 increased tolerance to a cephalosporin and to bile salts. An epimerase that may alter surface
508 polysaccharides was also beneficial in bile salts.

509 The selection of a physiologically related genetic background is critical for phenotypic assays.
510 For instance, the kinase hits likely would not have been identified in an *E. coli* overexpression
511 strain. *E. coli* uses the phosphotransferase system for D-glucosamine uptake where the
512 substrate is phosphorylated during uptake, thus rendering the kinase activities redundant.[78]

513 We demonstrate the benefit of combining overexpression data with loss-of-function data to
514 strengthen gene-to-phenotype mappings. 10 protein clusters have RB-TnSeq fitness data that
515 are consistent with our Boba-seq data (Extended Data Table 5).[10,11] The scalability of both loss-
516 of-function and gain-of-function libraries using barcoding techniques for Bacteroidales is an
517 exciting development. By constructing RB-TnSeq and Boba-seq libraries for a large number of
518 isolates and screening across identical conditions, we can rapidly generate specific functional
519 hypotheses for novel genes.

520 A few variables can be adjusted in our workflow. Current libraries have fragment lengths of ~3
521 kb and may not capture pathways with several proteins, therefore insert size can be increased
522 to cover larger gene clusters. Additionally, the nature of competitive growth assays can lead to
523 "jackpot" effects where the most fit strains outcompete other strains with a beneficial fragment
524 and reduce the number of hits. This effect can be reduced by increasing the number of
525 replicates, as BarSeq only costs a few dollars per sample, or by growing libraries on agar plates
526 or in microfluidic droplets to spatially separate strains and prevent competition.[79,80]

527 Boba-seq can be readily scaled to a larger number of libraries and assays. Our empty vector
528 library has ~14 million barcodes and although we multiplexed six libraries with ~300,000 total
529 strains, this can be increased. Libraries can be built using DNA from complex samples to
530 access greater genetic diversity and uncultivated taxa. Moreover, our vectors replicate in at
531 least nine additional Bacteroidales besides *B. theta*, which expands the diversity of functions
532 that can be screened. Additionally, loss-of-function mutant strains can be used for high-
533 throughput complementation screens. Finally, Boba-seq libraries can be tested in additional
534 assay formats, including host colonization, phage infections, suppressors of toxic genes, and
535 adhesion.[12,25,26,53,81,82]

536 **Online Methods**

537 **Strains and cultivation conditions**

538 Strains, plasmids, and primers used in this study are listed in Extended Data Table 1. *E. coli*
539 strains were routinely cultured in Luria-Bertani Lennox (LB) with the appropriate selection using
540 50 µg/mL carbenicillin (Carb50) or 50 µg/mL kanamycin (Kan50) at 37 °C with shaking at 200
541 rpm. *E. coli* WM3064 was grown with 300 µM diaminopimelic acid (DAP) supplement. Species
542 within the order Bacteroidales were grown at 37 °C in an anaerobic chamber (Coy) maintained
543 with a gas mix of 10% H_2, 5% CO_2, 85% N_2. Media recipes and commercial sources of
544 substrates are provided in Extended Data Table 4. Routine culturing was carried out in
545 supplemented Brain Heart Infusion (BHIS) where BD Bacto Brain Heart Infusion was freshly
546 supplemented with 0.5 mg/L hemin and 2.5 µL/L vitamin K_1. The Varel-Bryant (VB) defined
547 minimal medium was used for the majority of complementation and fitness assays as well as
548 validation growth curves.[40] B12 vitamin was excluded and instead, methionine was used in our
549 medium. Cysteine was substituted with dithiothreitol (DTT) as the reducing agent for certain
550 carbon conditions. An exception to this was the heparin validation growth curves, which were
551 obtained in another defined minimal medium (MM).[44] Media were made anaerobic by pre-
552 reducing in the anaerobic chamber at least one day prior to use. Conjugation mixes were
553 selected on 200 µg/mL gentamicin (Gm200) and 5 µg/mL erythromycin (Erm5) to obtain the
554 correct mutant.

555 **Construction of replicative vectors for *B. theta***

556 The pNBU2_erm-TetR-P1T_DP-GH023 integrative vector was modified to generate replicative
557 vectors. This vector contains a previously reported anhydrotetracycline (aTc) inducible promoter
558 and RBS (P1T_DP-GH023).[33] A PmlI restriction site, the synthetic terminator L3S2P21, and
559 BsmBI sites used for barcoding were inserted downstream of the inducible promoter-RBS
560 cassette.[29,34,35] First, a BsmBI restriction site was removed from the pNBU2_erm-TetR-P1T_DP-
561 GH023 vector using site-directed mutagenesis. The pNBU2 vector with no BsmBI site was
562 linearized by PCR to exclude the attachment sites. Fragments containing replicative machinery
563 and origin of replication from pTIO-1, pKSH24, and pFD340 were PCR amplified and purified
564 using GFX DNA and Gel Band Purification kit.[31,32,83] The third gene block consisted of an
565 insertion site, a terminator, and a barcoding site.[10,34,35] Three-part Gibson Assembly reactions
566 were performed to generate pNBU2_repA1, pNBU2_repA2, and pNBU2_repA3 containing
567 portions from pTIO-1, pKSH24, and pFD340, respectively. Gibson reactions were carried out
568 according to manufacturer's recommendations (NEB, Gibson Assembly master mix). This was
569 transformed into electrocompetent *E. coli* EC100D *pir+* and plated onto LB-Carb50 to obtain the
570 correct construct. All vectors were confirmed by Sanger sequencing.

571 **Detection of inducible expression using Luciferase assays**

572 To quantify gene expression on the three pNBU2_repA vectors, luciferase assays were
573 performed as previously reported, using Promega's Nano-Glo Luciferase Assay System Kit.[33]
574 Briefly, the NanoLuc gene was inserted downstream of the aTc-inducible promoter via Gibson
575 assembly. Sequence-verified vectors were conjugated into *B. theta*. These *B. theta* strains were

576 inoculated into BHIS-Erm5 and grown anaerobically at 37 °C for 16–18 h. Starter cultures were
577 subcultured into BHIS-Erm5 and varying concentrations of aTc at a 1:100 dilution to mid-
578 exponential phase (OD of 0.4-0.6, 3–5 hours). 500 µL of each culture were then pelleted by
579 centrifugation at 5,000 × *g*, and cells were lysed with 50 µL of 1× BugBuster Protein Extraction
580 Reagent (Sigma). Lysates clarified at 21,130 × *g* for 5 minutes, and 10 µL of the supernatant
581 was mixed with an equal volume of the NanoLuc reaction buffer containing the NanoLuc
582 substrate. Luminescence was immediately measured using a BioTek Synergy H1 microplate
583 reader (Agilent), and normalized to OD600 (optical density at 600 nm) of the culture at time of
584 lysis.

585 **Determination of plasmid copy number using droplet digital PCR**

586 Droplet digital PCR (ddPCR) was performed following a previously reported procedure to
587 estimate the plasmid copy number of the pNBU2_repA vectors in *B. theta*.[84] Briefly, *B. theta*
588 containing each vector was grown in BHIS-Erm5 for 16–18 h at 37 °C, followed by a 1:100
589 dilution and subculture to mid-exponential phase (OD of 0.7–0.9, 3–5 h). 500 µL of each culture
590 were then harvested for DNA extraction using Qiagen's DNeasy blood & tissue kit, after which 1
591 µg of DNA was digested using NdeI (NEB). ddPCR was performed using approximately 0.01,
592 0.1, or 1 ng of the digested DNA as template for each 40 µL reaction. PCR was performed using
593 the Evagreen supermix and primers p-AH204-p-AH207 (Extended Data Table 1), following
594 BioRad's protocols for multiplexing with Evagreen and using an annealing temperature of 60 °C.
595 Samples were analyzed using Bio-Rad's QX200 droplet digital PCR system. The resulting data
596 points were manually gated, and reported concentrations for each amplicon were used in
597 determining the plasmid copy number.

598 **Barcoding of the pNBU2_repA1 vector**

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

616 **Construction and quantification of barcoded genomic libraries**

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

625 The *B. theta* library was constructed using blunt-end ligation, however vector rearrangement
626 issues were observed by Sanger sequencing of individual *E. coli* clones. Cloning steps taken
627 are similar to the previously reported Dub-seq library.[24] Briefly, sheared and size-selected gDNA
628 from *B. theta* was end-repaired using End-IT DNA End-Repair kit (Lucigen) in a 50 µL reaction
629 consisting of 2 µL End-It Enz mix, 5 µL dNTP mix, 5 µL 10 mM ATP, 5 µL buffer, and 4 µg
630 gDNA. This was incubated at room temperature for 45 min and deactivated at 70 °C for 20 min.
631 This was gel-purified and size-selected for 3–6 kb fragments. The end-repaired, sheared DNA
632 fragments were then ligated to linearized and dephosphorylated barcoded vectors. A 20 µL
633 ligation reaction was performed with 10 µL Blunt/TA ligase master mix (NEB), 100 ng of vector
634 backbone, and 4-fold molar excess of end-repaired DNA fragments, and incubated at room
635 temperature for 20 min. Reaction was purified using Monarch PCR & DNA cleanup kit and
636 eluted twice with 10 µL H_2O. 20 ng of cleaned ligation reaction was transformed into 50 µL
637 aliquots of commercial *E.coli* EC100D that was diluted 4-fold with 10% glycerol. A total of 8
638 transformations were performed and each recovered in 1 mL of S.O.C. medium at 37 °C for 1.5
639 h with shaking at 200 rpm. An aliquot was removed for CFU counting to estimate library
640 diversity. Recovered cultures were combined and inoculated into 500 mL of LB-Carb50 for
641 selection. The final library was purified for BarSeq to obtain an estimated diversity. The initial *B.*
642 *theta* library consists of ~440,000 barcodes and was diluted to obtain the final library with
643 ~70,000 barcodes.

644 An alternative strategy was applied to the remaining six libraries, which did not lead to any
645 observable vector rearrangement. They were constructed by cloning DNA fragments into the
646 pCR™Blunt II-TOPO™ vector using Zero Blunt™ TOPO™ PCR Cloning Kit (Invitrogen),
647 followed by PCR amplification of the inserts and Gibson assembly into barcoded pNBU2_repA1.
648 In this protocol, sheared gDNA is first end-repaired in 20 µL reactions using Quick CIP (NEB)
649 and following the manufacturer's protocol. End-repaired inserts are column-purified. Ligations
650 were carried out in 6 µL TOPO reactions that contained 0.5 or 1 µL of pCR II-Blunt-TOPO
651 vector, 1 µL of salt solution, and 20 ng of end-repaired inserts. Reactions were incubated at
652 room temperature for 30 min and cleaned using SPRI beads (Beckman Coulter). Briefly, 0.9
653 equivalent volume of bead solution was mixed with each TOPO reaction and incubated for 15
654 min. Beads were washed three times with 80% EtOH and allowed to dry briefly. DNA was eluted
655 with 5 µL of H_2O heated to 50 °C and transformed into 50 µL of electrocompetent 3-fold diluted
656 *E. coli* EPI300 (Lucigen). An aliquot was taken out for CFU counting on LB-Kan50. Recovered
657 cultures were stored as glycerol stocks at −80 °C. Multiple TOPO reactions and transformations

658 were carried out until an estimate of ~40,000 to 100,000 CFUs was obtained for each library.
659 Glycerol stocks of recovered cells from the same library were combined and inoculated into 200
660 mL of LB-Kan50 for selection. Inserts were PCR amplified from 150 ng of purified TOPO vectors
661 using Q5 polymerase with an annealing temperature of 67 °C, 2 minute extension, and 8 cycles.
662 Template vectors were digested using DpnI and the 3–6 kb size range was gel-purified. Insert
663 amplicons were ligated into barcoded pNBU2_repA1 in 20 µL of Gibson reactions using the
664 manufacturer's protocol. 20 ng of cleaned reaction was transformed into 50 µL of 4-fold diluted
665 electrocompetent *E. coli* EC100D. Cells were recovered and stored as glycerol stocks before
666 selection. Aliquots of the recovered cells were plated onto LB-Carb50 to estimate CFUs.
667 Libraries were serially diluted to <100,000 CFUs and selected in liquid LB-Carb50 to make final
668 glycerol stocks and perform BarSeq. All libraries were transformed into *E. coli* WM3064 for
669 conjugation into *B. theta*.

670 Differences between libraries constructed using each cloning workflow were observed. The
671 ligated *B. theta* library has the highest percentage of genes covered (98%) and the lowest
672 strand bias. The PCR required for Gibson Assembly, though with a low cycling number,
673 introduced biases that likely led to lower genome coverage values (Table 1). Gibson-assembled
674 libraries exhibited a slight strand bias toward fragments oriented opposite to the promoter (ratio
675 of genes oriented opposite over genes oriented the same direction as vector promoter: 1.43–
676 1.88). In contrast, the number of genes on either strand is very similar in the *B. theta* library,
677 with a ratio of 0.98. This is likely a result of gene toxicity in *E. coli* caused by the lac promoter on
678 the TOPO vector. This strand bias did not increase upon library conjugation into *B. theta*, which
679 is consistent with the low basal expression of the aTc inducible promoter.

680 **PacBio sequencing to map libraries**

681 Barcoded libraries purified from EC100D were used as templates for amplicon sequencing.
682 Each sample was dual indexed using 5'-phosphorylated PCR primers containing 16 nt
683 barcodes.[85] Barcode and insert regions were amplified in 50 µL PCRs with Q5 polymerase, 150
684 ng of vector template, 57 °C annealing temperature, 2 min extension time, and 6 or 8 cycles. 8–
685 10 reactions per library were pooled, digested with DpnI twice (NEB), and column-concentrated.
686 Amplicons were size-selected for >800 bp lengths using 0.55 equivalent volume of AMPure PB
687 beads (PacBio). 200–600 ng of this was used as sample input for the manufacturer's protocol:
688 Preparing SMRTbell libraries using PacBio Barcoded Overhang Adapters for Multiplexing
689 Amplicons. The universal adapter was used during the ligation step instead of a barcoded
690 adapter since amplicons are dual indexed. Adapter-ligated amplicons were digested with
691 exonuclease using SMRTbell Enzyme Clean Up Kit 2.0 and purified using 0.55 equivalent
692 volume of AMPure PB beads. Final amplicon solutions were quantified using Qubit BR Assay
693 and sizes were confirmed on the Bioanalyzer HS DNA chip (Agilent). Samples were sequenced
694 on the PacBio Sequel II instrument using SMRT cell 8M (QB3 Genomics, UC Berkeley). See
695 Extended Data Table 2 for the number of demultiplexed CCS reads used to map each library. In
696 general, a good run on one PacBio SMRT cell 8M was sufficient to map 3–4 libraries.

697 **Computational pipeline used to map barcodes to genes**

698 The Boba-seq script used to map barcodes to genes and detailed explanations can be found at
699 https://github.com/OGalOz/Boba-seq and as a part of the deposited data. A summary is
700 provided as Extended Data Figure 4. Input data files are required and consist of: a reference
701 assembly file (.fasta), a genome annotation file (.gff/.gff3), and long reads data (.fastq). The user
702 also needs to provide a fasta file containing 4 short oligo sequences that flank the barcode and
703 the insert fragment on each amplicon. A configuration .JSON file is used to set parameters for
704 each step. Briefly, reads were first demultiplexed using PacBio's lima tool. Barcode and insert
705 sequences are extracted using 15bp flanking oligos using usearch (www.drive5.com/usearch/).
706 Filters are applied to exclude reads with concatemers and incorrect barcode lengths. Inserts
707 with an expected error of >10 are excluded using vsearch (https://github.com/torognes/vsearch).
708 Inserts are mapped to the reference genome using minimap2 and results are kept for hits with
709 high percent coverage and percent match values.[86] The majority of barcodes mapped to a
710 single genomic location. For barcodes that are mapped to multiple locations, most are due to
711 sequencing errors and are within a few nt of each other. For barcodes mapping to overlapping
712 locations, the location with the top read count is kept. Barcodes mapped to distant genomic
713 locations are kept in the mapping tables, but are excluded from our fitness analysis. Finally,
714 genomic positions are mapped to protein-coding genes to generate mapping tables, which are
715 used to calculate library statistics.

716 **Conjugation into *B. theta* and other Bacteroidales**

717 Single vectors were first transformed into conjugation donor strains *E. coli* S17-1 or WM3064.
718 40 mL of Bacteroidales recipient at OD ~0.1 was mixed with 4 mL of *E. coli* donor at OD ~1
719 outside of the anaerobic chamber. Mixed cultures were pelleted by centrifugation and
720 resuspended in 250 µL BHIS. Each cell mixture was then pooled onto BHIS or BHIS-DAP (for
721 WM3064) plates and incubated at 37 °C overnight aerobically for >16 h. The entire cell mass
722 was then scraped and resuspended in 1 mL of BHIS, from which 20–100 µL was plated onto
723 BHIS-Gm200-Erm5 agar. Antibiotic plates were incubated at 37 °C anaerobically for 48 h or
724 longer until distinct colonies formed. Single colonies were restreaked onto BHIS-Gm200-Erm5
725 once, from which a single colony was then inoculated into BHIS-Gm200-Erm5. Bacteroidales
726 cultures were inoculated into LB and incubated aerobically at 37 °C to ensure a lack of *E. coli*
727 contamination. Glycerol stocks were checked by Sanger sequencing of plasmids and 16S
728 rRNA. Using this protocol, pKSH24 and pTIO-1 were conjugated into *B. theta* to test replication
729 in a new host. pNBU2_repA1, pNBU2_repA2, and pNBU2_repA3 with NanoLuc were
730 conjugated into 13 different *Bacteroides*, *Parabacteroides*, and *Phocaeicola* species.

731 **Complementation libraries and assays**

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

738 For complementation assays, the starter culture of each library was diluted to an OD of 0.5 in
739 BHIS-Gm200-Erm5 with 40 ng/mL aTc and incubated at 37 °C for 3.5 h to induce gene
740 expression. Cells were pelleted and washed with VB minimal medium twice; these served as
741 Time0 samples. Each library was inoculated at a starting OD of 0.02 in 1 mL VB-Erm5-aTc40
742 and 20 mM of either leucrose, palatinose, trehalose, or α-cyclodextrin. Cultures were incubated
743 at 37 °C and harvested when turbid. Each assay was also streaked onto BHIS-Gm200-Erm5
744 agar plates. Colony PCR and Sanger sequencing were used to identify clones containing
745 BT2158, BT4448, or BACUNI_RS05440.

746 **Growth assays of complemented mutants**

747 Mutants complemented with BT2158, BT4448, or BACUNI_RS05440 (Extended Data Table 6),
748 *B. theta* Δ*tdk*ΔBT2158, and *B. theta* Δ*tdk* parental strain were inoculated into BHIS-Erm5 or
749 BHIS and grew overnight at 37 °C. Cells were pelleted and washed once with VB minimal
750 medium. Each mutant was inoculated at a starting OD of 0.05 in 200 µL VB-Erm5 and 20 mM
751 trehalose. *B. theta* Δ*tdk* and *B. theta* Δ*tdk*ΔBT2158 were inoculated in the same manner, except
752 without Erm. Growth measurements were obtained in a 96-well plate using an Epoch 2
753 Microplate Spectrophotometer (Agilent) inside of an anaerobic chamber. OD600 values shown
754 on plots are pathlength-corrected and blank-normalized.

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

756 *B. caccae, B. fragilis*, *B. salyersiae, B. uniformis, P. johnsonii,* and *P. merdae* libraries were
757 conjugated into wild-type *B. theta* as described for complementation libraries. For fitness
758 assays, 1 mL of glycerol stock was pelleted and resuspended in 10 mL of BHIS-Erm5 with 40
759 ng/mL of aTc as starter cultures to grow for 16–18 h. These were then pelleted and washed with
760 VB-Erm5 twice. Libraries were pooled equally based on OD to a final OD of 5, which served as
761 Time0 samples. For carbon utilization assays, pooled libraries were inoculated at a starting OD
762 of 0.04 in 1 mL of VB-Erm5 with 40 ng/mL aTc, and a single carbon substrate at the
763 concentration listed in Extended Data Table 4. For stress tolerance assays, pooled libraries
764 were inoculated at a starting OD of 0.1 or 0.04 in 0.5 mL of BHIS-Erm5, 40 ng/mL aTc, and an
765 inhibitory compound at a range of concentrations listed on Extended Data Table 4. Assays were
766 carried out in 48-well plates and incubated at 37 °C. Each condition was set up in replicates of
767 two on the same day. A few conditions were repeated on a second day of experimentation
768 where new conditions were tested as well. Cultures were harvested when turbid for BarSeq.

769 **DNA barcode sequencing (BarSeq) and analysis**

770 Vector libraries were purified from *E. coli* using E.Z.N.A. Plasmid DNA Mini Kit (Omega Bio-tek)
771 or Monarch Plasmid Miniprep kit (NEB). However, this led to low yields and high levels of
772 contaminants when extracting from Bacteroidales. DNA purified from the same set of
773 Bacteroidales samples using plasmid purification kits and genomic DNA purification kits were
774 tested for barcode PCR amplification. BarSeq results from both purified plasmids and gDNA
775 were comparable and therefore, DNeasy Blood & Tissue Kit or QiaAmp DNA QIAcube HT Kit
776 (Qiagen) for 96-well sample plates were used to extract DNA from *B. theta* cultures for BarSeq
777 for all fitness assays.

778 Barcodes were amplified using the previous-published “BarSeq_V3” primers or using new
779 “BarSeq_V4” primers (Extended Data Table 1).[9,87] The BarSeq version 4 primers are similar to
780 the BarSeq version 3 primers but use a 10-bp index sequence for P7 primers and a 8-bp index
781 for P5 (instead of 6 bp for both in version 3) and allow multiplexing up to 768 samples (instead
782 of 96). Briefly, BarSeq PCRs were performed as 50 µL reactions with Q5 polymerase (NEB) and
783 200 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
784 for 30 s, and 72 °C for 5 min. For both BarSeq3 and BarSeq4 designs, the P7 (or P2) BarSeq
785 primers allowed demultiplexing by Illumina software,[9] and the P5 (or P1) BarSeq primers
786 contained an additional index checked by the MultiCodes.pl script.[87] All PCRs were pooled
787 equally in volume and column-cleaned. Up to 96 samples (BarSeq3) or 152 samples (BarSeq4)
788 were multiplexed for sequencing on full lanes of Illumina HiSeq400 SE50 (QB3 Genomics, UC
789 Berkeley), partial lanes of NovaSeq6000 PE150 (Novogene), or full lanes of MiSeq (v2 reagent
790 kit). 11,000–94,000 reads per sample were obtained for complementation assays sequenced on
791 the MiSeq and 0.3–5.3 million reads were obtained for samples sequenced on the HiSeq or
792 NovaSeq.

793 BarSeq reads were converted to counts per barcode using the MultiCodes.pl script in the feba
794 code base (https://bitbucket.org/berkeleylab/feba/), with the -minQuality 0 option and either -bs3
795 or -bs4 for each version of BarSeq primers. To estimate the diversity of barcoded libraries, only
796 reads that had a quality score of ≥30 at each position (the -minQuality 30 option) were used,
797 which corresponds to an error rate for barcodes of at most 0.001 * 20 nt = 2%. Furthermore, any
798 barcodes that were off-by-1 errors from a more common barcode were eliminated. Barcodes
799 observed in the PacBio dataset are compared to barcode sequences from the BarSeq data to
800 determine the percentage of barcodes mapped.

801 For the highly diverse empty library, the Chao1 estimator was used to estimate the barcode
802 diversity from 2.99 million high-quality reads,[88] The Chao1 estimator uses the number of
803 barcodes seen just once, but some may be due to sequencing errors. To correct for this,
804 barcodes that were off-by-1 nucleotide from a more abundant barcode were ignored, and the
805 number of singletons was reduced by an assumed error rate ranging from 0% to 2%. This
806 yielded estimates ranging from 13.5–14.1 million barcodes for the empty pNBU2_repA1 library.

807 **Calculation of fitness scores for each fragment**

808 Only barcodes that mapped to a single location in one library and that were present in the
809 conjugation donor (*E. coli* WM3064) libraries were considered. 2% of barcodes mapped to more
810 than one library and another 5% of mapped barcodes were not detected in the BarSeq data for
811 the conjugation donor libraries. For complementation assays, we computed fitness scores for all
812 of these fragments. For fitness assays in the wild-type *B. theta* background with six pooled
813 libraries, 280,036 of the 305,382 barcodes (that were present in WM3064) were detected in the
814 Time0 samples. Also, barcodes that were not detected in Time0 samples occasionally had
815 statistically significant benefits. This is likely due to positive selection that led mutants to gain in
816 abundance in the assay condition despite their low initial abundance. Therefore, fitness scores
817 were computed for all mapped barcodes that were present in WM3064. The script used to
818 calculate fitness scores can be found at https://github.com/morgannprice/BobaseqFitness.

819 The fitness of each strain was defined as:

820 $$\log_2\{ (n + \psi)/N \} - \log_2\{ (n_0 + 1/\psi)/N_0 \}$$

821 $$\psi = \mathrm{sqrt}(N/N_0)$$

822 where n is the number of reads for this barcode in the experimental sample, n_0 is the number of
823 reads for this barcode across the Time0 samples, and N and N_0 are the total number of reads
824 for all considered barcodes in the experimental or Time0 samples, respectively. The
825 pseudocount ψ is added to the counts to ensure that the log ratio is not undefined if either count
826 is zero. This value of ψ also ensures that if both counts (n and n_0) are 0, then the fitness will be
827 zero (corresponding to no evidence of a change in relative abundance). If the experimental and
828 Time0 samples have the same total number of reads, then ψ = 1.

829 **Determination of statistically significant hits**

830 The minimum fitness value threshold for statistical significance was selected based on a control
831 comparison between the two Time0 samples collected on the same day for the large fitness
832 screen. Each sample contained a mixture of six genomic libraries and had 2.9 million or 3.0
833 million total reads. Of the 305,382 barcodes considered, the fitness values ranged from -3.9 to
834 +4.1. Therefore, a minimum fitness value of +5 was selected for a statistically-significant hit or
835 barcode.

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

839 $$\sigma = \mathrm{sqrt}\{ 1/(n + \psi) + 1/(n_0 + 1/\psi) \} / \ln(2)$$

840 The z score is then fitness / σ and z ≥ 4 is selected as the cutoff. For the standard normal
841 distribution and 305,382 barcodes, 10 values would be expected above this threshold.

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

843 We focused on barcodes and inserts that had significantly high fitness scores in both replicate
844 experiments conducted in the same condition and on the same day. We considered these
845 barcodes to have a biologically consistent phenotype if either an overlapping fragment or a
846 fragment encoding a similar protein had statistically significant benefits. Specifically, an insert
847 was confirmed by overlap if another insert from the same region of the genome was significant
848 in both replicates. An insert was confirmed by similarity if the likely causative protein for the
849 region (see below) was similar (at least 40% identity and 75% coverage) to a protein that was
850 found within another insert that had significantly high fitness in two replicates for the same
851 condition. For experiments in the wild-type *B. theta* background where libraries were pooled, the
852 confirmation would occur in the same experiments. For complementation assays where libraries
853 were assayed separately, the criterion might be fulfilled from different assays in the same
854 condition.

855 For each region with inserts that were statistically significant in both replicates, the insert with
856 the highest fitness score (after averaging replicates) was selected first. For each protein
857 encoded within this insert, we computed several scores. First, we counted the number of
858 significantly-beneficial inserts that contain this protein (with each replicate counted separately).
859 Second, we counted the number of similar proteins that were in high-scoring regions. Third, we
860 computed the average fitness score for all inserts containing this gene. To select the likely
861 causative protein for each region, we chose the one with the highest number of beneficial
862 inserts, or the highest number of similar proteins, or the highest average fitness. We used the
863 additional scores to break ties.

864 For each of the putative causative proteins, we visually inspected the plot of fitness scores
865 versus insert fragments. When available, we also considered the plots for similar experiments
866 (such as the same carbon source with a different base medium, or a different concentration of
867 the stress, or the same condition conducted on a different day) with similar protein hits. In 20
868 cases, we determined that more than one protein is necessary to confer a strong benefit. In 6
869 cases, we identified a different protein as conferring the benefit. In 28 cases, we were not sure
870 which protein(s) conferred the benefit and we left these hits uncurated in Extended Data Table
871 5. For the remaining 162 cases, the automatically-selected protein was deemed correct.

872 **Validation of gene hits using individual strains**

873 Selected gene hits from the fitness screens were PCR amplified from the corresponding gDNA,
874 cloned into the pNBU2_repA1 backbone via Gibson Assembly, and transformed into *E. coli*
875 S17-1 to generate validation vectors. Mutants with high fitness scores were selected for
876 validations and the exact fragment boundaries for these were extracted from library mapping
877 tables to design primers (Extended Data Table 6). Additionally, amplicons of the precise gene or
878 genes were obtained. Validation vectors were conjugated into *B. theta*, and each strain was
879 grown in the same condition as carried out for the fitness assay with AraC hits being an
880 exception. Here, we switched to a different defined minimal medium (MM) where *B. theta* grew
881 more quickly in heparin (see Extended Data Table 4 for composition).[44] Each strain was
882 cultured in 200 µL on a 96-well plate in replicates of 3. In a few cases, the empty vector control
883 was grown up in replicates of 2. Growth measurements were obtained using an Epoch 2
884 Microplate Spectrophotometer (Agilent). OD values are pathlength-corrected and blank-
885 normalized.

886 **Overexpression and purification of hexokinase**

887 Hexokinase (INE72_RS21610) was amplified from *B. caccae* CL03T12C61 and cloned into
888 pET29b vector for protein overexpression and purification. 500 mL of *E. coli* BL21(DE3)
889 transformed with pET29b-hexokinase was induced with 0.5 mM IPTG at OD ~0.6. Pellet was
890 harvested and then resuspended in lysis buffer (50 mM Tris HCl pH 7, 100 mM NaCl, 1 mM
891 PMSF, 0.2 mg/ml lysozyme, 1X SIGMAFAST protease inhibitor, 1% streptomycin sulfate, 5 mM
892 BME). Resuspended cells were lysed using a sonicator probe. The clarified supernatant was
893 loaded onto equilibrated Ni Sepharose 6 Fast Flow resin (Cytiva) and washed with a buffer
894 containing 50 mM Tris HCl pH 7, 100 mM NaCl, and 5 mM BME. Protein was eluted using the
895 wash buffer with 10, 20, 50, 100, or 200 mM imidazole and fractions were collected at each

896 concentration. All fractions were visualized on SDS-PAGE and the appropriate fractions were
897 pooled and concentrated down using 30K MWCO concentrators. Concentrated protein eluate
898 was desalted using a PD-10 column loaded with Sephadex G-25 resin (Cytiva). Protein
899 concentrations were calculated using Abs280 measurements from Nanodrop and a molar
900 extinction coefficient of 44,350 M^{-1} cm^{-1}. In total, 4.2 mg of protein was purified to a
901 concentration of 105 μM. Protein aliquots were stored at –80°C.

902 **Hexokinase enzyme assays**

903 *In vitro* coupled enzyme assays were performed to test substrate preference of hexokinase.
904 Assay components include 50 mM Tris-Cl pH 7, 50 mM NaCl, 2 mM $MgSO_4$, 1 mM NADH, 2
905 mM phospho(enol)pyruvic acid tri(cyclohexylammonium) salt (PEP), 1 mM ATP, and 1 mM
906 substrate. Glucose, mannose, D-glucosamine, and N-acetyl- D-glucosamine were tested as
907 substrates. Enzymes were added to concentrations of 10 U/mL lactate dehydrogenase from
908 rabbit muscle (L2500-10KU, Sigma), 10 U/mL pyruvate kinase from rabbit muscle (P1506-1KU,
909 Sigma), and 5 μM purified hexokinase. Lactate dehydrogenase and pyruvate kinase
910 concentrations were also tested at a 2-fold increase to ensure they were in excess and not the
911 rate limiting step. Assays were initiated by adding hexokinase and incubated for 5 minutes at
912 room temperature. Assays were performed in volumes of 40 uL in triplicates in a 384-well plate.
913 NADH consumption was observed through pathlength-corrected absorbance at 340 nm
914 measured using a BioTek Synergy HTX plate reader.

915 **RNA-seq analysis of *ssb* region in L-fucose**

916 For RNA-seq experiments, a *B. theta* strain expressing the *ssb* fragment 2 (Extended Data
917 Table 6) and a strain containing the empty vector were grown in replicates of 3 in BHIS-Erm5-
918 aTc40 to mid-exponential phase (OD600 = 0.4–0.8) and harvested using RNAprotect bacterial
919 agent (Qiagen) and the RNeasy Mini kit (Qiagen) combined with on-column DNA depletion
920 using the RNase-free DNase set (Qiagen). ~500 ng of total RNA from each sample was
921 submitted to Novogene for further processing; this included rRNA depletion followed by library
922 preparation with the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB).
923 Library quality was checked using a 2100 Bioanalyzer (Agilent).

924 Sequencing reads were trimmed and aligned to the *B. theta* VPI-5482 genome
925 (GCF_000011065.1) using HISAT2 (2.2.1) with the strand-specific option for paired-end reads,
926 and default parameters otherwise.[89] Mapped reads were then assigned to genome features
927 using featureCounts with the strand specific setting.[90] Due to concern over repeat regions,
928 counts were generated with and without multimapping using featureCounts, which led to similar
929 results. Final counts were generated with multimapping. These counts were normalized and
930 analyzed using the DESeq2 package in R, resulting in log2 fold-changes along with test
931 statistics and adjusted *p*-values generated using the Wald test (Extended Data Table 7).[91]

932 **Materials and data availability**

933 The script for mapping Boba-seq libraries is available at https://github.com/OGalOz/Boba-seq.
934 The script for calculating fitness scores is available at

935 https://github.com/morgannprice/BobaseqFitness. Plasmid maps, a snapshot of the python
936 script used for library mapping, mapping tables for each library, a snapshot of the R script used
937 for fitness score calculations along with the input files, output files from fitness analysis, and raw
938 reads from the RNA-seq experiment are deposited at:
939 https://doi.org/10.6084/m9.figshare.24195054. All other data are available from the authors
940 upon request. Please direct plasmids and strain requests to Y.Y.H. and A.P.A.

941 **Acknowledgements**

942 We thank Dr. Naoya Ohara (Okayama University) for pTIO-1, Dr. Andrei Shkoporov (University
943 of College Cork) for pKSH24, and Dr. Laurie Comstock (University of Chicago) for pFD340.
944 pNBU2_erm-TetR-P1T_DP-GH023 and pNBU2_erm-TetR-P1T_DP-GH023-NanoLuc were gifts
945 from Dr. Andrew Goodman at Yale University (Addgene plasmids #90324 and #117728). We
946 thank Surya Tripathi for access to unpublished RB-TnSeq fitness data, and Bradley Biggs,
947 Shekhar Mishra, Vivek Mutalik, and Valentine Trotter for helpful feedback on the manuscript.
948 Y.Y.H. is an Astellas Pharmaceuticals Awardee of the Life Sciences Research Foundation,
949 which funded part of this work. A.H. is an awardee of NSF GRFP grant DGE-2146752. We are
950 grateful for the support of the Dean A. Richard Newton Memorial Professor Chair Funds. This
951 work was supported in part by ENIGMA – Ecosystems and Networks Integrated with Genes and
952 Molecular Assemblies (http://enigma.lbl.gov), a Science Focus Area Program at Lawrence
953 Berkeley National Laboratory, supported by the U.S. Department of Energy, Office of Science,
954 Office of Biological & Environmental Research under contract number DE-AC02-05CH11231.
955 This work was partly supported by the NIH grant RM1 GM135102 (to A.M.D.). We acknowledge
956 QB3 Genomics, UC Berkeley, Berkeley, CA for sequencing support through
957 RRID:SCR_022170. This work was supported by NIH S10 OD018174 Instrumentation Grant.
958 Several Bacteroidales used in this study were obtained through BEI Resources, NIAID, NIH as
959 part of the Human Microbiome Project.

960 **Author contributions**

961 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.
962 carried out experimental work and collected data. Y.Y.H., M.N.P., A.H., and O.G. performed
963 computational analyses. Y.Y.H. and M.N.P. wrote the original draft. Y.Y.H., M.N.P., A.H.,
964 A.M.D., and A.P.A. revised and edited the paper.

965 **Competing interests**

966 The authors have no competing interests to declare.

967 **References**

968 1. Human Microbiome Project Consortium. Structure, function and diversity of the healthy
969 human microbiome. *Nature* **486**, 207–214 (2012).
970 2. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome
971 Project. *Nature* **550**, 61–66 (2017).
972 3. Wu, G. D. *et al.* Comparative metabolomics in vegans and omnivores reveal constraints on

973 diet-dependent gut microbiota metabolite production. *Gut* **65**, 63–72 (2016).
974 4. Franzosa, E. A. *et al.* Relating the metatranscriptome and metagenome of the human gut.
975 *Proc. Natl. Acad. Sci.* **111**, E2329–E2338 (2014).
976 5. Vujkovic-Cvijin, I. *et al.* Host variables confound gut microbiota studies of human disease.
977 *Nature* **587**, 448–454 (2020).
978 6. Ni, J., Wu, G. D., Albenberg, L. & Tomov, V. T. Gut microbiota and IBD: causation or
979 correlation? *Nat. Rev. Gastroenterol. Hepatol.* **14**, 573–584 (2017).
980 7. Lobb, B., Tremblay, B. J.-M., Moreno-Hagelsieb, G. & Doxey, A. C. An assessment of
981 genome annotation coverage across the bacterial tree of life. *Microb. Genomics* **6**, e000341
982 (2020).
983 8. Vanni, C. *et al.* Unifying the known and unknown microbial coding sequence space. *eLife*
984 **11**, e67667 (2022).
985 9. Wetmore, K. M. *et al.* Rapid Quantification of Mutant Fitness in Diverse Bacteria by
986 Sequencing Randomly Bar-Coded Transposons. *mBio* **6**, e00306-15 (2015).
987 10. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function.
988 *Nature* **557**, 503–509 (2018).
989 11. Liu, H. *et al.* Functional genetics of human gut commensal Bacteroides thetaiotaomicron
990 reveals metabolic requirements for growth across environments. *Cell Rep.* **34**, 108789
991 (2021).
992 12. Adler, B. A. *et al.* The genetic basis of phage susceptibility, cross-resistance and host-range
993 in Salmonella. *Microbiology* **167**, 001126 (2021).
994 13. Yaung, S. J. *et al.* Improving microbial fitness in the mammalian gut by in vivo temporal
995 functional metagenomics. *Mol. Syst. Biol.* **11**, 788 (2015).
996 14. Joglekar, P. *et al.* Intestinal IgA Regulates Expression of a Fructan Polysaccharide
997 Utilization Locus in Colonizing Gut Commensal Bacteroides thetaiotaomicron. *mBio* **10**,
998 e02324-19 (2019).
999 15. Tasse, L. *et al.* Functional metagenomics to mine the human gut microbiome for dietary
1000 fiber catabolic enzymes. *Genome Res.* **20**, 1605–1612 (2010).
1001 16. Yoon, M. Y. *et al.* Functional screening of a metagenomic library reveals operons
1002 responsible for enhanced intestinal colonization by gut commensal microbes. *Appl. Environ.*
1003 *Microbiol.* **79**, 3829–3838 (2013).
1004 17. Crofts, T. S., Mcfarland, A. G. & Hartmann, E. M. Mosaic Ends Tagmentation (METa)
1005 assembly for extremely efficient 1 construction of functional metagenomic libraries.
1006 *mSystems* **6**, e00524-21 (2021).
1007 18. Forsberg, K. J. *et al.* Functional metagenomics-guided discovery of potent Cas9 inhibitors in
1008 the human microbiome. *eLife* **8**, e46540 (2019).
1009 19. Mastropaolo, M. D., Thorson, M. L. & Stevens, A. M. Comparison of Bacteroides
1010 thetaiotaomicron and Escherichia coli 16S rRNA gene expression signals. *Microbiology* **155**,
1011 2683–2693 (2009).
1012 20. Vingadassalom, D. *et al.* An unusual primary sigma factor in the Bacteroidetes phylum. *Mol.*
1013 *Microbiol.* **56**, 888–902 (2005).
1014 21. Wegmann, U., Horn, N. & Carding, S. R. Defining the Bacteroides ribosomal binding site.
1015 *Appl. Environ. Microbiol.* **79**, 1980–1989 (2013).
1016 22. Lee, S. M. *et al.* Bacterial colonization factors control specificity and stability of the gut
1017 microbiota. *Nature* **501**, 426–429 (2013).
1018 23. Lam, K. N., Martens, E. C. & Charles, T. C. Developing a Bacteroides System for Function-
1019 Based Screening of DNA from the Human Gut Microbiome. *mSystems* **3**, e00195-17 (2018).
1020 24. Mutalik, V. K. *et al.* Dual-barcoded shotgun expression library sequencing for high-
1021 throughput characterization of functional traits in bacteria. *Nat. Commun.* **10**, 1–14 (2019).
1022 25. Mutalik, V. *et al.* High-throughput mapping of the phage resistance landscape in E. coli |
1023 PLOS Biology. *PLoS Biol* **18**, e3000877 (2020).

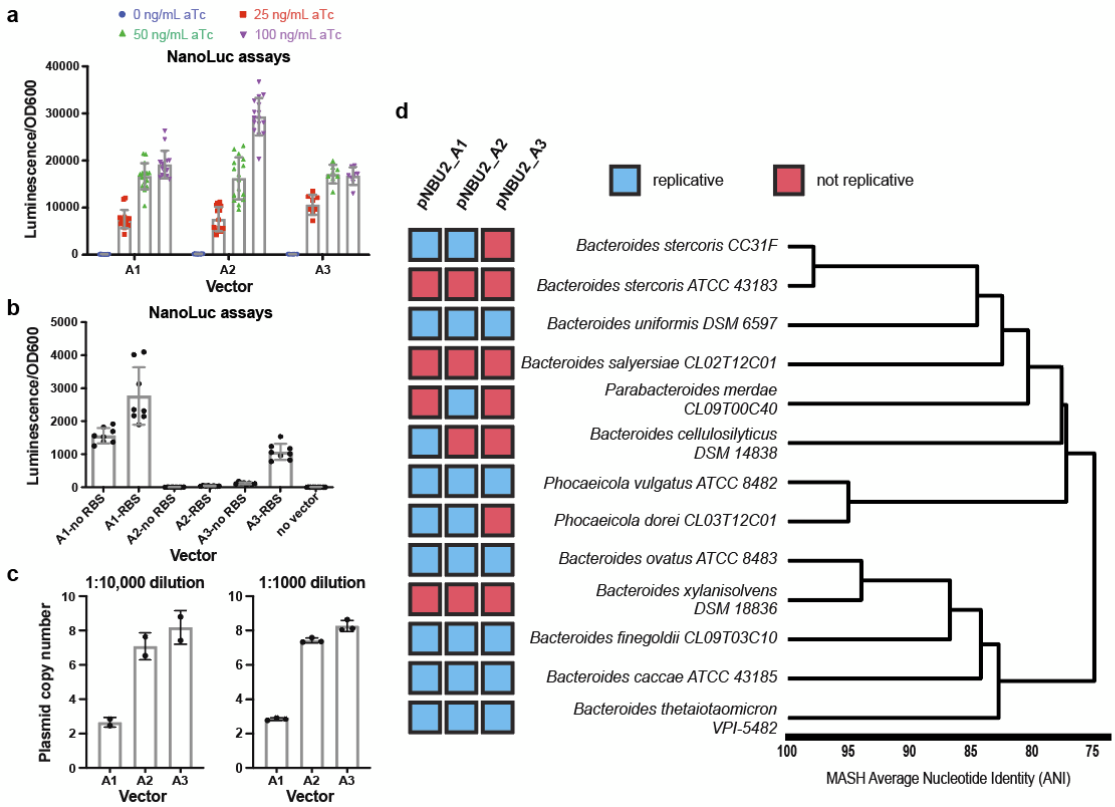
1024 26. Adler, B. A. *et al.* Multicopy suppressor screens reveal convergent evolution of single-gene
1025 lysis proteins. *Nat. Chem. Biol.* **19**, 759–766 (2023).
1026 27. Apjok, G. *et al.* Characterization of antibiotic resistomes by reprogrammed bacteriophage-
1027 enabled functional metagenomics in clinical strains. *Nat. Microbiol.* **8**, 410–423 (2023).
1028 28. Wexler, A. G., Goodman, A. L., Voigt, C. A., Lu, T. K. & Hennet, T. An insider's perspective:
1029 Bacteroides as a window into the microbiome. *Nat. Microbiol.* **2**, 17026 (2017).
1030 29. Liu, H. *et al.* Magic Pools: Parallel Assessment of Transposon Delivery Vectors in Bacteria.
1031 *mSystems* **3**, e00143-17 (2018).
1032 30. Tagawa, J. *et al.* Development of a novel plasmid vector pTIO-1 adapted for
1033 electrotransformation of Porphyromonas gingivalis. *J. Microbiol. Methods* **105**, 174–179
1034 (2014).
1035 31. Shkoporov, A. N. *et al.* Plasmid Analysis of a novel 8.9 kb cryptic plasmid from Bacteroides
1036 uniformis , its long-term stability and spread within human microbiota. *Plasmid* **69**, 146–159
1037 (2013).
1038 32. Smith, C. J., Rogers, M. B. & Mckee, M. L. Heterologous Gene Expression in Bacteroides
1039 fragilis. *Plasmid* **27**, 141–154 (1992).
1040 33. Lim, B., Zimmermann, M., Barry, N. A. & Goodman, A. L. Engineered Regulatory Systems
1041 Modulate Gene Expression of Human Commensals in the Gut. *Cell* **169**, 547-558.e15
1042 (2017).
1043 34. Chen, Y. J. *et al.* Characterization of 582 natural and synthetic terminators and
1044 quantification of their design constraints. *Nat. Methods* **10**, 659–664 (2013).
1045 35. Taketani, M. *et al.* Genetic circuit design automation for the gut resident species
1046 Bacteroides thetaiotaomicron. *Nat. Biotechnol.* **38**, 962–969 (2020).
1047 36. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny
1048 substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
1049 37. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood
1050 Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).
1051 38. Sorek, R. *et al.* Genome-Wide Experimental Determination of Barriers to Horizontal Gene
1052 Transfer. *Science* **318**, 1449–1452 (2007).
1053 39. Liou, C. S. *et al.* A Metabolic Pathway for Activation of Dietary Glucosinolates by a Human
1054 Gut Symbiont. *Cell* **180**, 717-728.e19 (2020).
1055 40. Varel, V. H. & Bryant, M. P. Nutritional Features of Bacteroides fragilis subsp. fragilis. *Appl.*
1056 *Microbiol.* **28**, 251–257 (1974).
1057 41. Auger, S., Gomez, M. P., Danchin, A. & Martin-Verstraete, I. The PatB protein of Bacillus
1058 subtilis is a C-S-lyase. *Biochimie* **87**, 231–238 (2005).
1059 42. Méndez, J. *et al.* A Novel cdsAB Operon Is Involved in the Uptake of l-Cysteine and
1060 Participates in the Pathogenesis of Yersinia ruckeri. *J. Bacteriol.* **193**, 944–951 (2011).
1061 43. Price, M. N. & Arkin, A. P. PaperBLAST: Text Mining Papers for Information about
1062 Homologs. *mSystems* **2**, 10.1128/msystems.00039-17 (2017).
1063 44. Martens, E. C., Chiang, H. C. & Gordon, J. I. Mucosal Glycan Foraging Enhances Fitness
1064 and Transmission of a Saccharolytic Human Gut Bacterial Symbiont. *Cell Host Microbe* **4**,
1065 447–457 (2008).
1066 45. Accetto, T. & Avguštin, G. Inability of Prevotella bryantii to Form a Functional Shine-
1067 Dalgarno Interaction Reflects Unique Evolution of Ribosome Binding Sites in Bacteroidetes.
1068 *PLoS ONE* **6**, (2011).
1069 46. Larsbrink, J. *et al.* A discrete genetic locus confers xyloglucan metabolism in select human
1070 gut Bacteroidetes. *Nature* **506**, 498–502 (2014).
1071 47. Cartmell, A. *et al.* How members of the human gut microbiota overcome the sulfation
1072 problem posed by glycosaminoglycans. *Proc. Natl. Acad. Sci.* **114**, 7037–7042 (2017).
1073 48. Price, M. N., Deutschbauer, A. M. & Arkin, A. P. Filling gaps in bacterial catabolic pathways
1074 with computation and high-throughput genetics. *PLOS Genet.* **18**, e1010156 (2022).

1075 49. Nihira, T. *et al.* Discovery of β-1,4-d-Mannosyl-N-acetyl-d-glucosamine Phosphorylase
1076 Involved in the Metabolism of N-Glycans. *J. Biol. Chem.* **288**, 27366–27374 (2013).
1077 50. Brigham, C. J. & Malamy, M. H. Characterization of the RokA and HexA broad-substrate-
1078 specificity hexokinases from Bacteroides fragilis and their role in hexose and N-
1079 acetylglucosamine utilization. *J. Bacteriol.* **187**, 890–901 (2005).
1080 51. Pollet, R., Marin, L. & Koropatkin, N. TonB-dependent transporters in the Bacteroidetes:
1081 Unique domain structures and potential functions. *Mol. Microbiol.* **115**, 490–501 (2021).
1082 52. Elango, D. *et al.* Raffinose Family Oligosaccharides: Friend or Foe for Human and Plant
1083 Health? *Front. Plant Sci.* **13**, 829118 (2022).
1084 53. Kennedy, M. S. *et al.* Dynamic genetic adaptation of Bacteroides thetaiotaomicron during
1085 murine gut colonization. *Cell Rep.* **42**, 113009 (2023).
1086 54. Gray, D. A. *et al.* Insights into SusCD-mediated glycan import by a prominent gut symbiont.
1087 *Nat. Commun.* **12**, 44 (2021).
1088 55. Zhou, J. *et al.* A new α-galactosidase from symbiotic Flavobacterium sp. TN17 reveals four
1089 residues essential for α-galactosidase activity of gastrointestinal bacteria. *Appl. Microbiol.*
1090 *Biotechnol.* **88**, 1297–1309 (2010).
1091 56. Blatch, G. L. & Woods, D. R. Molecular characterization of a fructanase produced by
1092 Bacteroides fragilis BF-1. *J. Bacteriol.* **175**, 3058–3066 (1993).
1093 57. Casanueva, A. I., Paul, L., Patrick, S. & Abratt, V. R. An AraC/XylS family transcriptional
1094 regulator homologue from Bacteroides fragilis is associated with cell survival following DNA
1095 damage. *FEMS Microbiol. Lett.* **278**, 249–256 (2008).
1096 58. de Freitas, M. C. R. *et al.* Exploratory Investigation of Bacteroides fragilis Transcriptional
1097 Response during In vitro Exposure to Subinhibitory Concentration of Metronidazole. *Front.*
1098 *Microbiol.* **7**, 1465 (2016).
1099 59. Tian, X. *et al.* Comparative Proteomic Analysis of Fucosylated Glycoproteins Produced by
1100 Bacteroides thetaiotaomicron Under Different Polysaccharide Nutrition Conditions. *Front.*
1101 *Microbiol.* **13**, 826942 (2022).
1102 60. Hooper, L. V., Xu, J., Falk, P. G., Midtvedt, T. & Gordon, J. I. A molecular sensor that allows
1103 a gut commensal to control its nutrient foundation in a competitive ecosystem. *Proc. Natl.*
1104 *Acad. Sci. U. S. A.* **96**, 9833–9838 (1999).
1105 61. Pickard, J. M. *et al.* Rapid fucosylation of intestinal epithelium sustains host-commensal
1106 symbiosis in sickness. *Nature* **514**, 638–641 (2014).
1107 62. Shereda, R. D., Kozlov, A. G., Lohman, T. M., Cox, M. M. & Keck, J. L. SSB as an
1108 organizer/mobilizer of genome maintenance complexes. *Crit. Rev. Biochem. Mol. Biol.* **43**,
1109 289–318 (2008).
1110 63. Samuel, B. & Burstein, D. A diverse repertoire of anti-defense systems is encoded in the
1111 leading region of plasmids. 2023.02.15.528439 Preprint at
1112 https://doi.org/10.1101/2023.02.15.528439 (2023).
1113 64. Chatzidaki-Livanis, M., Coyne, M. J. & Comstock, L. E. A Family of Transcriptional
1114 Antitermination Factors Necessary for Synthesis of the Capsular Polysaccharides of
1115 Bacteroides fragilis. *J. Bacteriol.* **191**, 7288–7295 (2009).
1116 65. Arsène, S. & Leclercq, R. Role of a qnr-Like Gene in the Intrinsic Resistance of
1117 Enterococcus faecalis to Fluoroquinolones. *Antimicrob. Agents Chemother.* **51**, 3254–3258
1118 (2007).
1119 66. Oh, S. F. *et al.* Host immunomodulatory lipids created by symbionts from dietary amino
1120 acids. *Nature* **600**, 302–307 (2021).
1121 67. Meng, Y., Ingram-Smith, C., Cooper, L. L. & Smith, K. S. Characterization of an Archaeal
1122 Medium-Chain Acyl Coenzyme A Synthetase from Methanosarcina acetivorans. *J. Bacteriol.*
1123 **192**, 5982–5990 (2010).
1124 68. Bi, H., Zhu, L., Jia, J., Zeng, L. & Cronan, J. E. Unsaturated fatty acid synthesis in the
1125 gastric pathogen Helicobacter pylori proceeds via a backtracking mechanism. *Cell Chem.*

1126 *Biol.* **23**, 1480–1489 (2016).
1127 69. Zhou, J. *et al.* Helicobacter pylori FabX contains a [4Fe-4S] cluster essential for unsaturated
1128 fatty acid synthesis. *Nat. Commun.* **12**, 6932 (2021).
1129 70. Isabella, V. M. & Clark, V. L. Identification of a conserved protein involved in anaerobic
1130 unsaturated fatty acid synthesis in Neiserria gonorrhoeae: implications for facultative and
1131 obligate anaerobes that lack FabA. *Mol. Microbiol.* **82**, 489–501 (2011).
1132 71. Hamilton, J. P. *et al.* Human cecal bile acids: concentration and spectrum. *Am. J. Physiol.-*
1133 *Gastrointest. Liver Physiol.* **293**, G256–G263 (2007).
1134 72. Akiyama, Y., Kanehara, K. & Ito, K. RseP (YaeL), an Escherichia coli RIP protease, cleaves
1135 transmembrane sequences. *EMBO J.* **23**, 4434–4442 (2004).
1136 73. Lycklama a Nijeholt, J. A. & Driessen, A. J. M. The bacterial Sec-translocase: structure and
1137 mechanism. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 1016–1028 (2012).
1138 74. Masanta, W. O. *et al.* Proteome Profiling by Label-Free Mass Spectrometry Reveals
1139 Differentiated Response of Campylobacter jejuni to Sublethal Concentrations of Bile Acids.
1140 *Proteomics Clin. Appl.* **13**, 1800083 (2019).
1141 75. Xu, J. *et al.* A Genomic View of the Human-Bacteroides thetaiotaomicron Symbiosis.
1142 *Science* **299**, 2074–2076 (2003).
1143 76. McMillan, A. S., Foley, M. H., Perkins, C. E. & Theriot, C. M. Loss of Bacteroides
1144 thetaiotaomicron bile acid altering enzymes impact bacterial fitness and the global metabolic
1145 transcriptome. 2023.06.27.546749 Preprint at https://doi.org/10.1101/2023.06.27.546749
1146 (2023).
1147 77. Sun, Z., Wang, Y., Su, X., Yang, X. & Luo, Q. Proteomic Characterization of Human Gut
1148 Habitual *Bacteroides intestinalis* against Common Intestinal Bile Acid Stress. *Adv. Gut*
1149 *Microbiome Res.* **2023**, e8395946 (2023).
1150 78. Álvarez-Añorve, L. I., Calcagno, M. L. & Plumbridge, J. Why Does Escherichia coli Grow
1151 More Slowly on Glucosamine than on N-Acetylglucosamine? Effects of Enzyme Levels and
1152 Allosteric Activation of GlcN6P Deaminase (NagB) on Growth Rates. *J. Bacteriol.* **187**,
1153 2974–2982 (2005).
1154 79. Thibault, D. *et al.* Droplet Tn-Seq combines microfluidics with Tn-Seq for identifying complex
1155 single-cell phenotypes. *Nat. Commun.* **10**, 5729 (2019).
1156 80. Tauzin, A. S. *et al.* Investigating host-microbiome interactions by droplet based
1157 microfluidics. *Microbiome* **8**, 1–20 (2020).
1158 81. Helmann, T. C., Deutschbauer, A. M. & Lindow, S. E. Genome-wide identification of
1159 Pseudomonas syringae genes required for fitness during colonization of the leaf surface
1160 and apoplast. *Proc. Natl. Acad. Sci.* **116**, 18900–18910 (2019).
1161 82. Vassallo, C. N., Doering, C. R., Littlehale, M. L., Teodoro, G. I. C. & Laub, M. T. A functional
1162 selection reveals previously undetected anti-phage defence systems in the E. coli
1163 pangenome. *Nat. Microbiol.* **7**, 1568–1579 (2022).
1164 83. Tagawa, J. *et al.* Development of a novel plasmid vector pTIO-1 adapted for
1165 electrotransformation of Porphyromonas gingivalis. *J. Microbiol. Methods* **105**, 174–179
1166 (2014).
1167 84. Plotka, M., Wozniak, M. & Kaczorowski, T. Quantification of Plasmid Copy Number with
1168 Single Colour Droplet Digital PCR. *PLoS ONE* **12**, e0169846 (2017).
1169 85. Callahan, B. J. *et al.* High-throughput amplicon sequencing of the full-length 16S rRNA gene
1170 with single-nucleotide resolution. *Nucleic Acids Res.* **47**, e103 (2019).
1171 86. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinforma. Oxf. Engl.* **34**,
1172 3094–3100 (2018).
1173 87. Morgan N. Price, a Jayashree Ray, a Anthony T. Iavarone, B., Hans K. Carlson, a Elizabeth
1174 M. Ryan, c Rex R. Malmstrom, C., Adam P. Arkin, a, D. & Deutschbauer, A. M. Oxidative
1175 Pathways of Deoxyribose and Deoxyribonate Catabolism. *mSystems* **4**, e00297-18 (2019).
1176 88. Chao, A. Nonparametric Estimation of the Number of Classes in a Population. *Scand. J.*

1177 *Stat.* **11**, 265–270 (1984).
1178 89. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome
1179 alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915
1180 (2019).
1181 90. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for
1182 assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
1183 91. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for
1184 RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).
1185 92. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate
1186 genomic comparisons that enables improved genome recovery from metagenomes through
1187 de-replication. *ISME J.* **11**, 2864–2868 (2017).
1188 93. Gilchrist, C. L. M. & Chooi, Y.-H. clinker & clustermap.js: automatic generation of gene
1189 cluster comparison figures. *Bioinformatics* **37**, 2473–2475 (2021).
1190

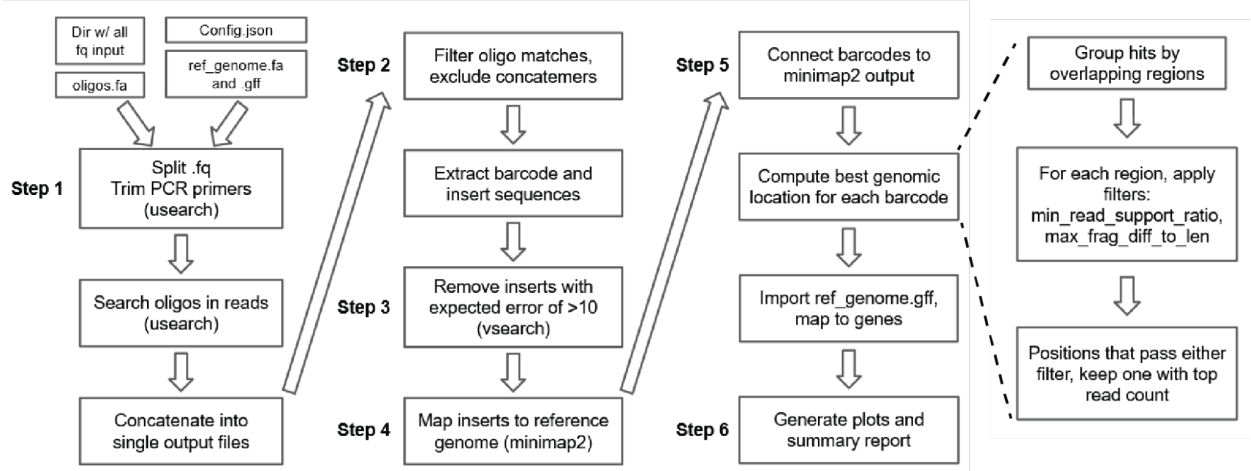
1191 **Extended Data Figures**



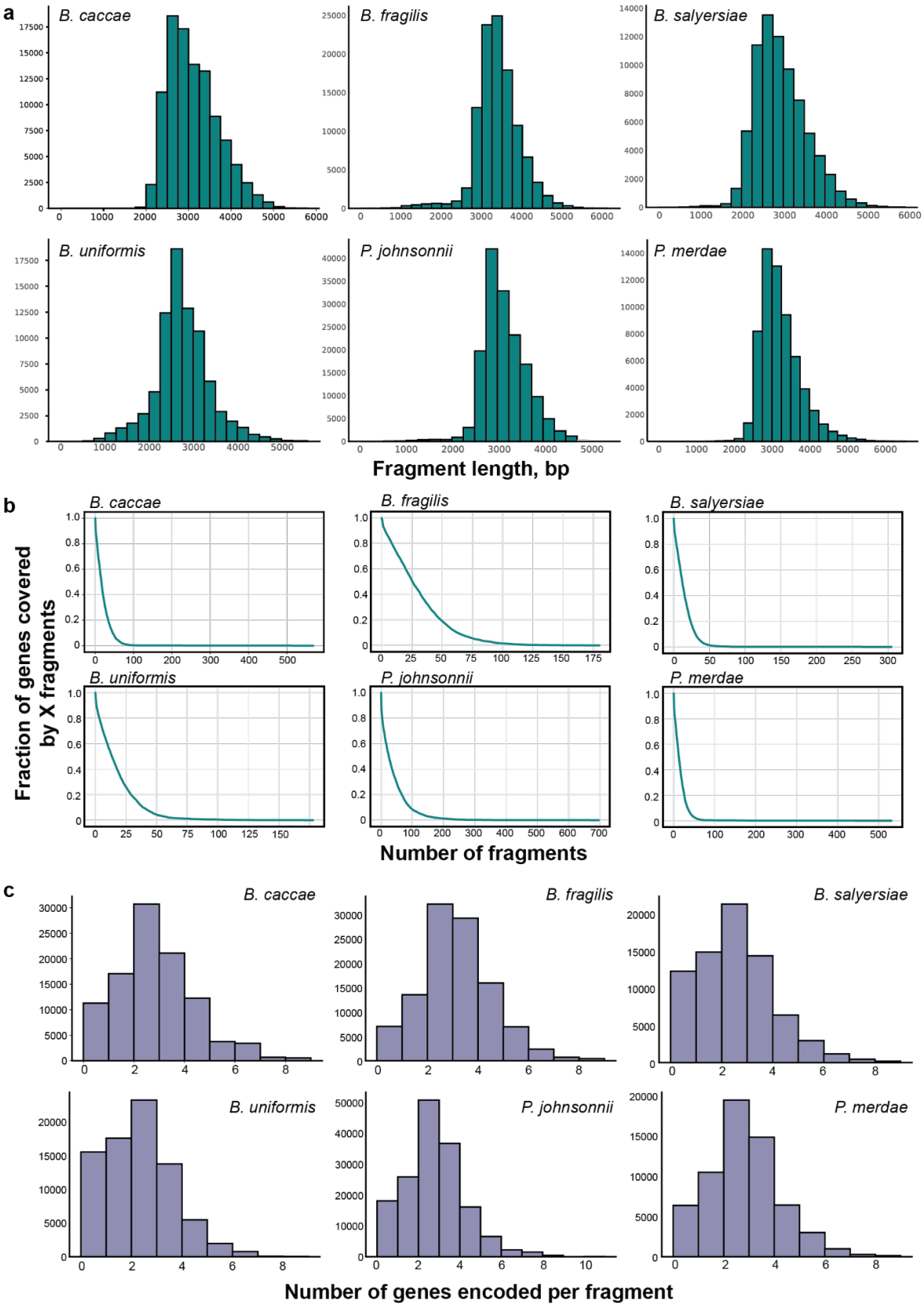
1192

1193 **Extended Data Figure 1.** Anhydrotetracycline (aTc) induces expression of NanoLuc on pNBU2
1194 replicative vectors in B. theta. A) A range of aTc concentrations (0, 25, 50, 100 ng/mL) was
1195 tested for three vectors with NanoLuc inserted downstream of the inducible promoter and an
1196 RBS. A total of two or four experiments were performed per vector with each experiment
1197 consisting of four replicates. B) NanoLuc was inserted into the pNBU2_repA1 vector in the
1198 opposite orientation to the aTc inducible promoter to test for constitutive promoter activity
1199 downstream of the insert site. NanoLuc gene with and without the upstream RBS (GH023) were
1200 tested in the absence of aTc. Two experiments were carried out with each experiment
1201 consisting of four replicates. Luminescence values were normalized by OD600 for all assays.
1202 Standard deviation is shown as error bars for both plots. C) Plasmid copy numbers in B. theta
1203 determined using droplet digital PCR (ddPCR). Cultures in the late exponential phase were
1204 harvested for this assay. Replicates from two dilutions were used to calculate plasmid copy
1205 numbers. D) Host range of pNBU2-based vectors across 13 Bacteroidales. Genome distance
1206 estimation and visualization were performed using dRep.[92]

1207



Extended Data Figure 2. 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, genomic regions. 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 they are reported by the mapping script. Two parameters are used to compute the best genomic location for barcodes that map to very similar positions. First, the ratio of highest read count to the second highest read count (min_BC_support_ratio value) is set at a threshold of 3 to identify the position supported by more reads. Second, 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. This is set to a threshold of 0.25 in our current pipeline. All filters and parameters can be readily adjusted in the configuration .JSON input file.

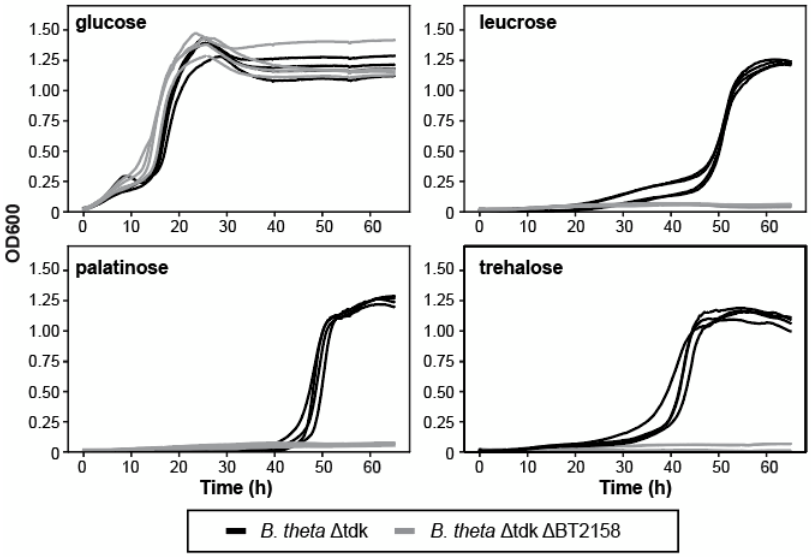

a
B. caccae
B. fragilis
B. salyersiae
B. uniformis
P. johnsonnii
P. merdae
Fragment length, bp
b
B. caccae
B. fragilis
B. salyersiae
B. uniformis
P. johnsonnii
P. merdae
Fraction of genes covered by X fragments
Number of fragments
c
B. caccae
B. fragilis
B. salyersiae
B. uniformis
P. johnsonnii
P. merdae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

1227

1228 **Extended Data Figure 3.** Summary plots for six Boba-seq genomic libraries. Plots are
1229 generated as part of the mapping script. A) Histograms of fragment sizes. B) Cumulative gene
1230 coverage plots. C) Histograms of the number of full-length genes encoded per fragment.

1231

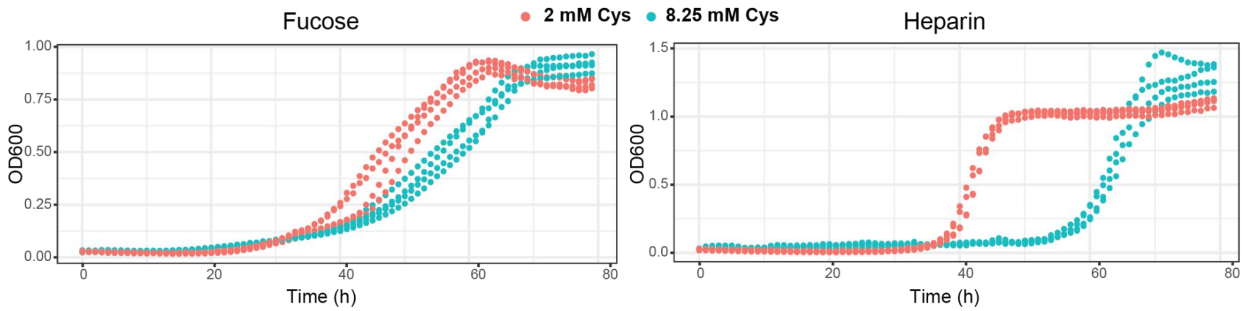
1232



1233

1234 **Extended Data Figure 4.** Growth curves of *B. theta* Δ*tdk* parental strain and Δ*tdk*ΔBT2158
1235 mutant on 20 mM glucose, leucrose, palatinose, or trehalose in the VB minimal medium. Each
1236 strain was grown in replicates of 4 per condition.

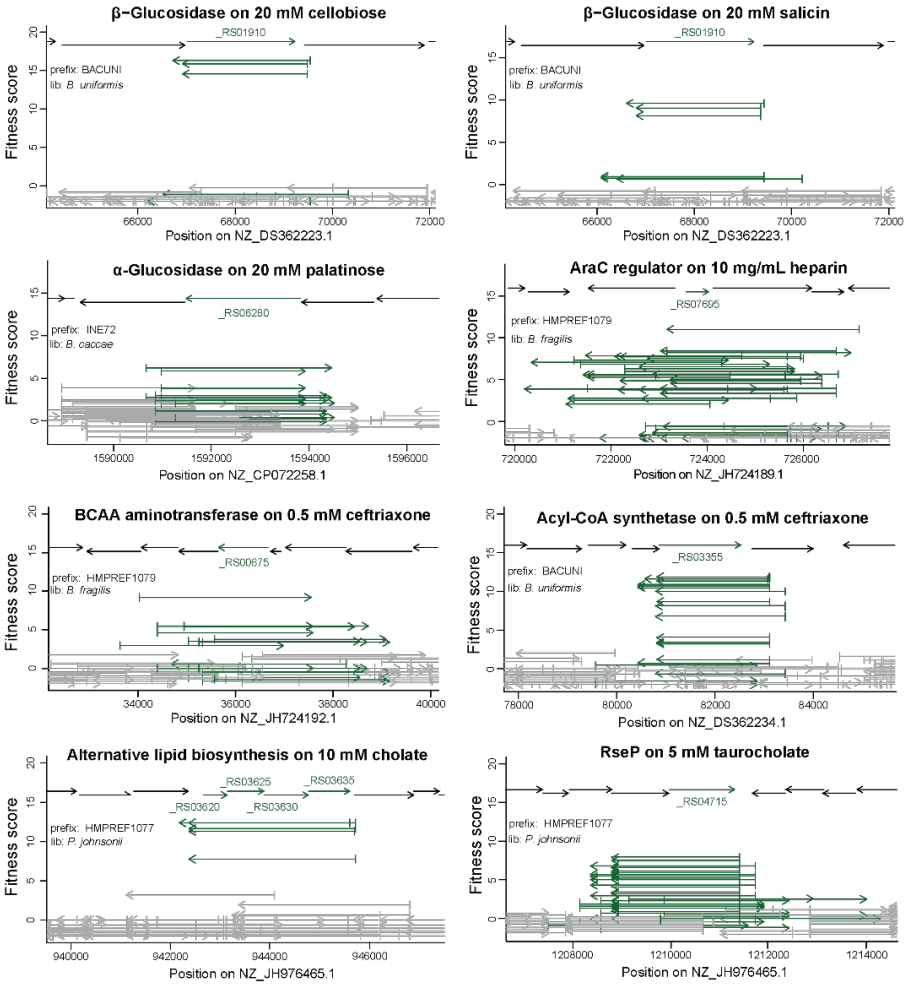
1237



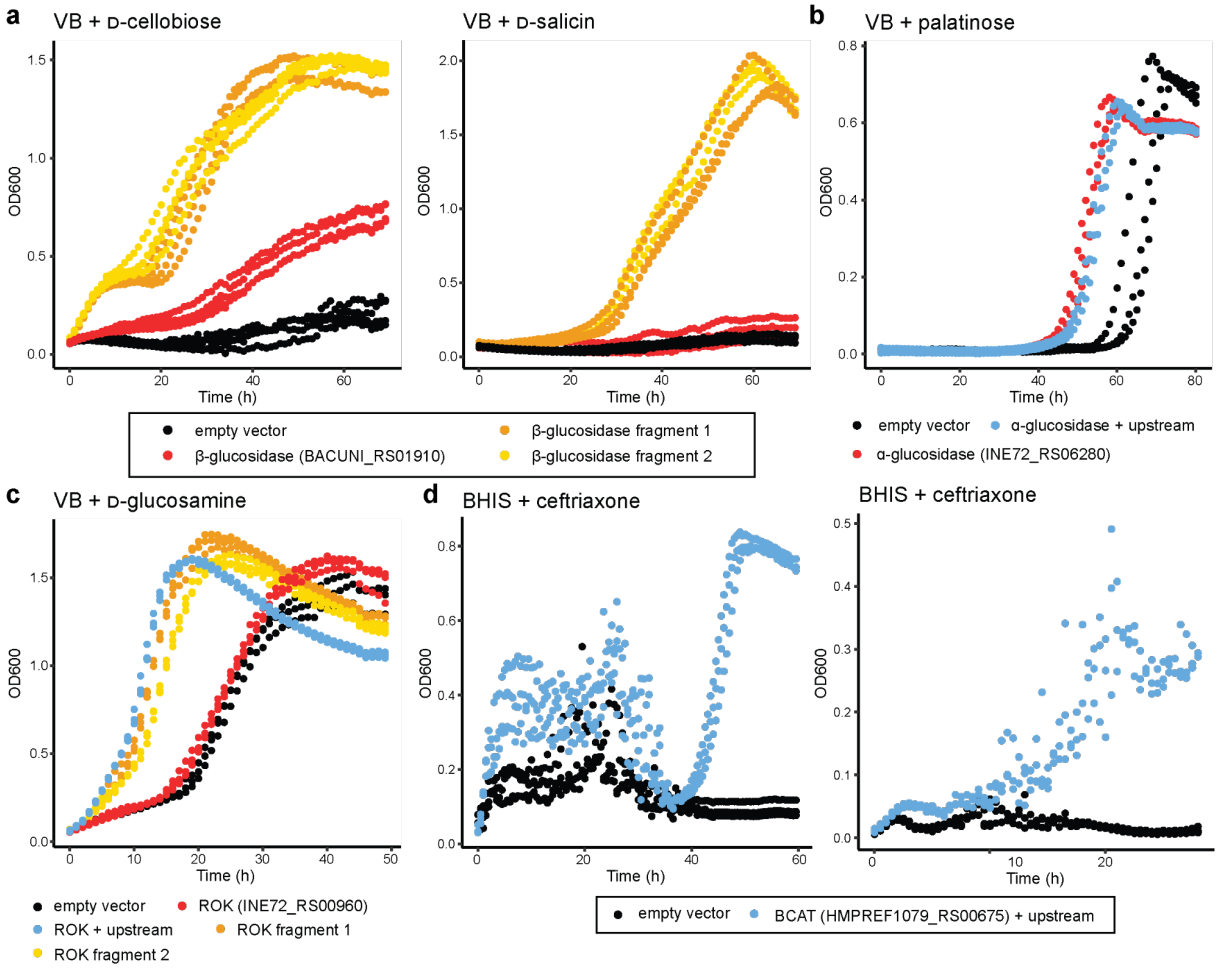
1238

1239 **Extended Data Figure 5.** Growth of wild-type *B. theta* is inhibited by the high cysteine
1240 concentration (8.25 mM) in the VB minimal medium. Examples are shown for 20 mM L-fucose
1241 and 10 mg/mL heparin in replicates of 4. A lower cysteine concentration (2 mM) resulted in
1242 improved growth.

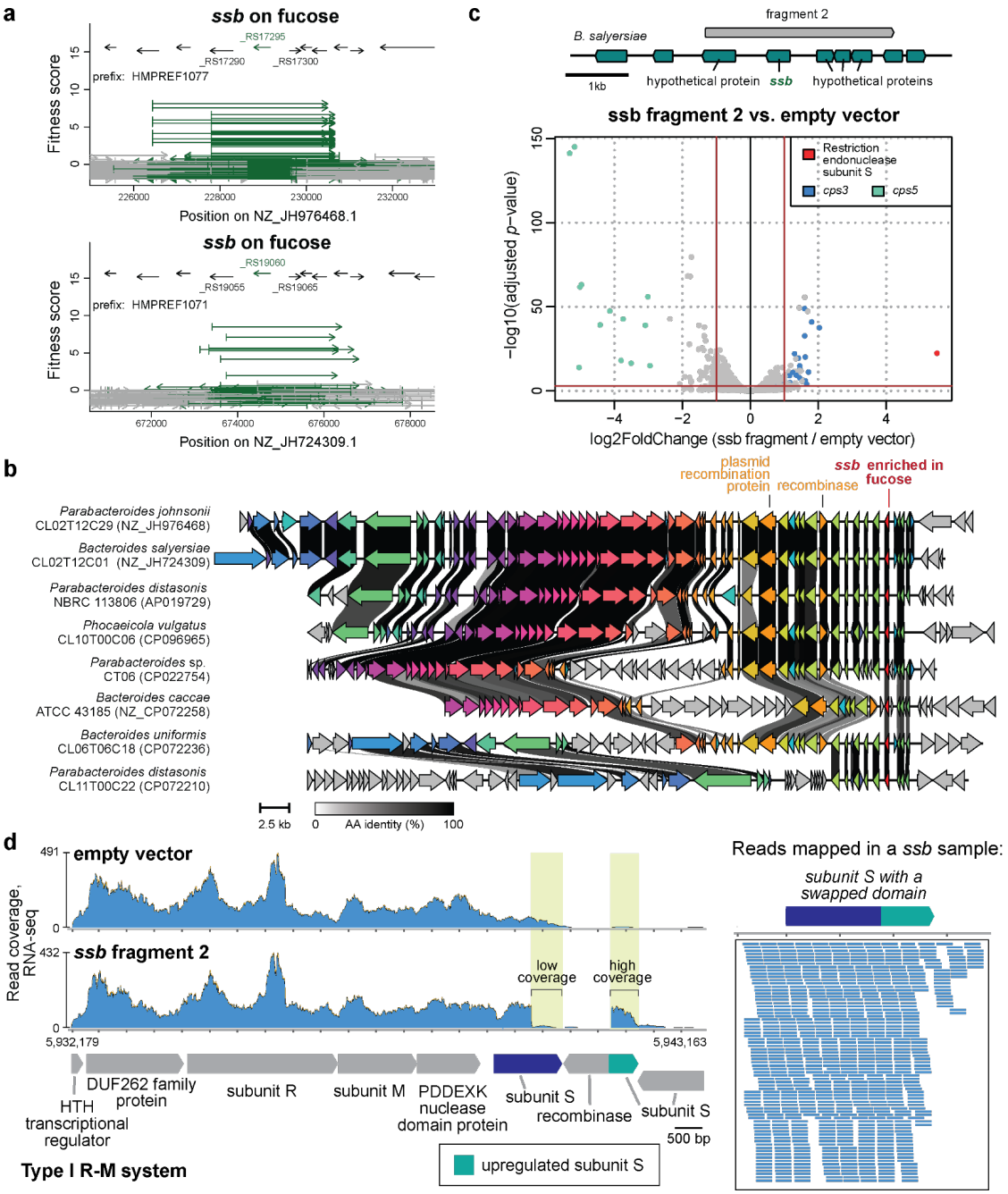
1243



Extended Data Figure 6. Fitness versus fragment plots of gene hits across carbon utilization and stress assays. Each arrow represents a mapped fragment in the corresponding Boba-seq library. The beneficial gene(s) and all inserts covering the full-length gene(s) are highlighted in green. The genomic positions of the inserts within the source genome are displayed on the x-axis. The average fitness score of each fragment from two experiments are shown.



Extended Data Figure 7. Growth benefits conferred by gene hits in individual *B. theta* strains. Each strain is conjugated with either a vector encoding the causative gene, the causative gene and the upstream 200 bp (or 100 bp for BCAT), a complete insert fragment from the library, or an empty vector. See Extended Data Table 6 for precise genomic boundaries, orientations, and fitness scores for each library fragment. Strains were constructed to confirm hits from A) D-cellobiose and D-salicin, B) palatinose, and C) D-glucosamine carbon substrates (20 mM) in the VB defined minimal medium. D) A branched-chain amino acid aminotransferase gene was confirmed to provide a benefit in BHIS with 1 mM ceftriaxone in two experiments. OD values are pathlength-corrected and blank-normalized. Each strain was grown in replicates of 3.


a
ssb on fucose
Fitness score
prefix: HMPREF1077
Position on NZ_JH976468.1
ssb on fucose
prefix: HMPREF1071
Position on NZ_JH724309.1
c
fragment 2
B. salyersiae
hypothetical protein
ssb
hypothetical proteins
1kb
ssb fragment 2 vs. empty vector
Restriction endonuclease subunit S
cps3
cps5
−log10(adjusted p−value)
log2FoldChange (ssb fragment / empty vector)
b
plasmid recombination protein
recombinase
ssb enriched in fucose
Parabacteroides johnsonii CL02T12C29 (NZ_JH976468)
Bacteroides salyersiae CL02T12C01 (NZ_JH724309)
Parabacteroides distasonis NBRC 113806 (AP019729)
Phocaeicola vulgatus CL10T00C06 (CP096965)
Parabacteroides sp. CT06 (CP022754)
Bacteroides caccae ATCC 43185 (NZ_CP072258)
Bacteroides uniformis CL06T06C18 (CP072236)
Parabacteroides distasonis CL11T00C22 (CP072210)
2.5 kb
AA identity (%)
d
empty vector
ssb fragment 2
Read coverage, RNA-seq
low coverage
high coverage
5,932,179
5,943,163
HTH transcriptional regulator
DUF262 family protein
subunit R
subunit M
PDDEXK nuclease domain protein
subunit S
recombinase
subunit S
500 bp
Type I R-M system
upregulated subunit S
Reads mapped in ssb sample:
subunit S with a swapped domain

1261

1262 **Extended Data Figure 8.** A genomic region in two different isolates (*B. salyersiae* and *P.*
1263 *johnsonii*) encoding an SSB protein conferred a strong benefit for growth on 20 mM L-fucose. A)
1264 The fitness versus fragment plots for both regions. The fitness scores displayed are averages
1265 across 8 experiments, including assays with different reductants (Cys or DTT) and in the
1266 presence or absence of aTc. B) Gene neighborhoods of the *ssb* region among Bacteroidales
1267 suggest this region is part of a mobile genetic element inserted at different genomic locations.
1268 Insertions and rearrangements are observed downstream of the *ssb* gene along with a
1269 conserved plasmid recombination protein and a recombinase. Genomic comparisons were
1270 visualized using Clinker with an amino acid identity cutoff of 40%.[93] C) Log2 ratios of fold-
1271 change and adjusted *p*-values (Wald test) from the RNA-seq experiment of *B. theta* encoding
1272 the ssb fragment 2 compared to an empty vector. Selected hits are labeled by color. D)
1273 Coverage plots of sense strand reads from RNA-seq data point to a genomic rearrangement
1274 that led to the upregulation of a subunit S (BT_RS22805/BT4522) part of a type I R-M system.
1275 Alignments were made using HISAT2 with default splice parameters on the strand-specific
1276 setting. Representative plots were generated using sense strand reads from empty_1 and
1277 ssb_1 samples. Sense strand reads from ssb_1 sample (*B. theta* encoding the *ssb* fragment 2)
1278 map across the subunit S gene formed after recombination.