Cell Host and Microbe, Volume *6*

Supplemental Data

Identifying Genetic Determinants Needed to Establish a Human Gut Symbiont in Its Habitat

Andrew L. Goodman, Nathan P. McNulty, Yue Zhao, Douglas Leip, Robi D. Mitra, Catherine A. Lozupone, Rob Knight, and Jeffrey I. Gordon

Figure S1. Selective capture of informative DNA fragments by MmeI restriction digestion. (A) Sequence alignment of closely related *mariner* inverted repeat (IR) sequences (Bigot *et al.*, 2005) indicates that a single G-T transversion (red) of a nonconserved nucleotide in the Himar1 IR creates a MmeI Type IIs restriction enzyme site (5′-TCCRAC) at the distal end of the IR. Conserved bases are shown as white characters on black background, non-conserved bases as black characters on white background. **(B)** Short genomic DNA fragments are sufficient to determine transposon location. In this simulation of random transposon insertion (described in Supplemental *Experimental Procedures*), pseudoreads of varying length were generated from, and mapped against, the genomes of a representative gut bacterium (~6.3 Mbp) and a simple eukaryote (~12.5 Mbp). **(C)** Features of pSAM. Marked restriction sites, with the exception of AscI, are unique. Insertion of a species-specific promoter element to direct expression of the transposase, and substitution of the antibiotic resistance marker if necessary, allows the vector to be customized for a recipient species of interest.

Goodman *et al.*, Supplemental Fig. S1

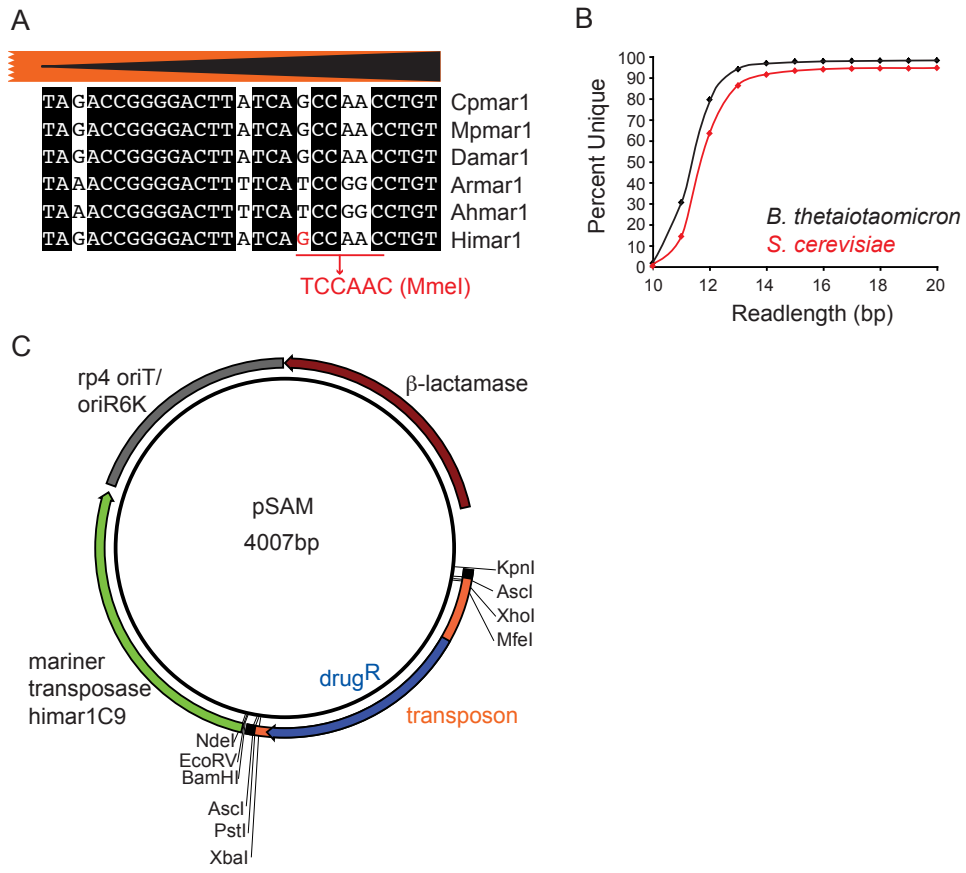


Figure S2. Transposon mutagenesis of *Bacteroides thetaiotaomicron*. (A) Features of pSAM_Bt. The promoter region from BT1311 (the *B. thetaiotaomicron* homolog of the housekeeping sigma factor *rpoD*) drives expression of the *mariner* transposase. Illumina bridge PCR priming sites (P7) and transcriptional terminators (hairpins) are marked. Transposition efficiency was equivalent between MmeI-adapted and traditional *mariner* transposons (data not shown). **(B)** Transposon insertion occurs without integration of the vector backbone and with a high frequency of single insertions per recipient cell. The wild-type parent, 20 randomly chosen transposon insertion strains, and a control strain bearing an Amp^R/Erm^R pGERM (Salyers *et al.*, 2000) vector insertion were tested by PCR for presence of the transposon (measured by amplification of the *ermG* antibiotic resistance cassette) and the vector backbone (measured by amplification of the beta-lactamase gene). Primers specific to a *B. thetaiotaomicron* gene serve as a positive control. To estimate the randomness of integration and frequency of multiple insertions in the same clone, genomic DNA was digested with EcoRV, resolved by electrophoresis and immobilized on a nylon membrane. A dioxygenin-labeled transposon-specific probe was used to identify transposon-containing fragments by Southern blot (bottom of panel **B**). **(C)** To measure the efficiency of MmeI digestion, genomic DNA from two independent transposon insertion strains was digested with MmeI and analyzed by Southern blotting using a transposon-specific probe. **(D)** Uniform capture of size-matched genomic fragments from an MmeI-adapted transposon. Several representative sequencing outputs (from input and cecal output populations, containing a total of ~20 million raw reads) were queried for transposon starting position: transposon sequences beginning at basepair 17/18 reflect a distance of 20/21 basepairs between the MmeI recognition site and the end of the captured genomic fragment.

Goodman *et al.*, Supplemental Fig. S2

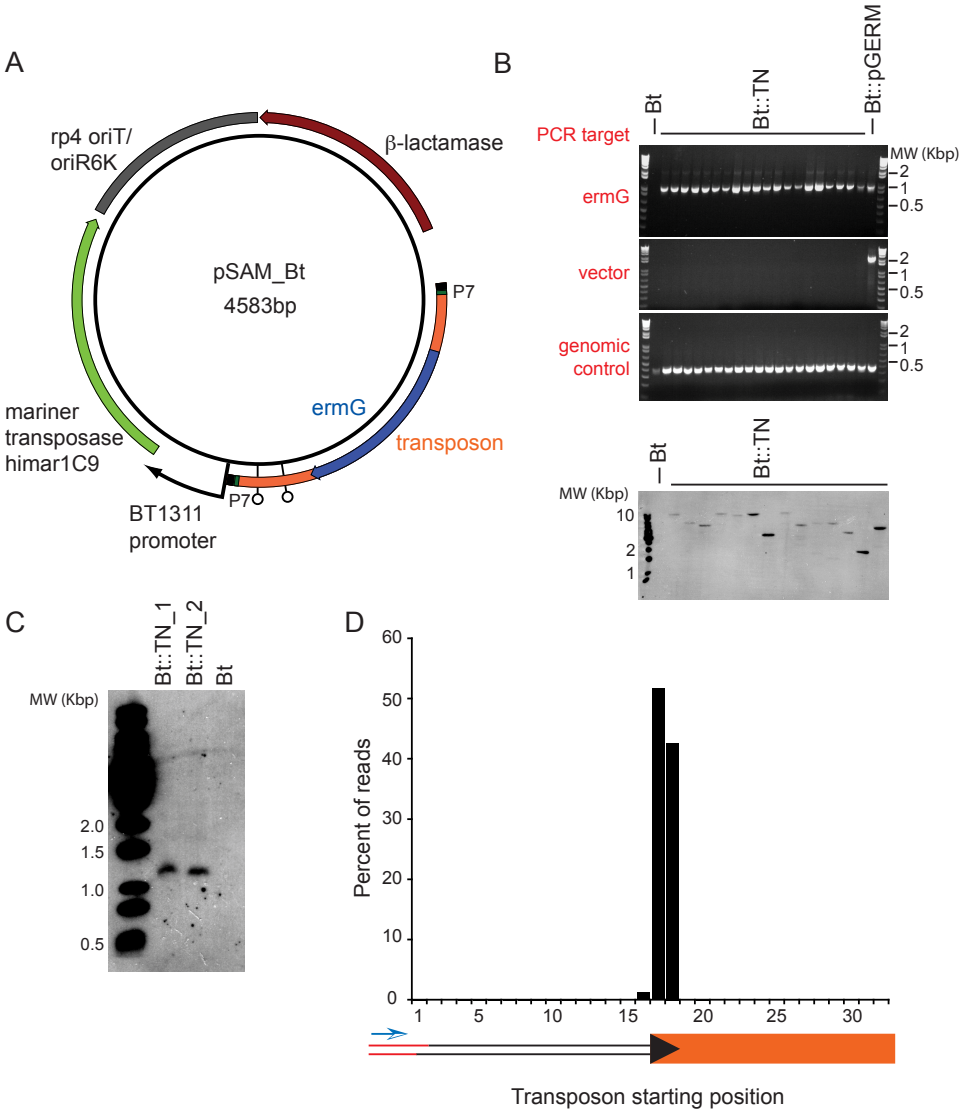


Figure S3. Reproducibility of library preparation and sequencing protocols. (A) Quantification of left- vs. right-hand reads for matched read-pairs in a representative dataset. **(B,C)** Comparison of relative abundance of individual reads (panel B) and of insertions (panel C) in technical replicates (separate libraries prepared and sequenced from one starting population). To adjust for differences in sequencing depth, insertion counts are median-normalized at the level of insertions (panel C). Coefficient of determination (R^2) values on log-transformed abundances are 0.63, 0.79, and 0.73 for panels A, B, and C, respectively.

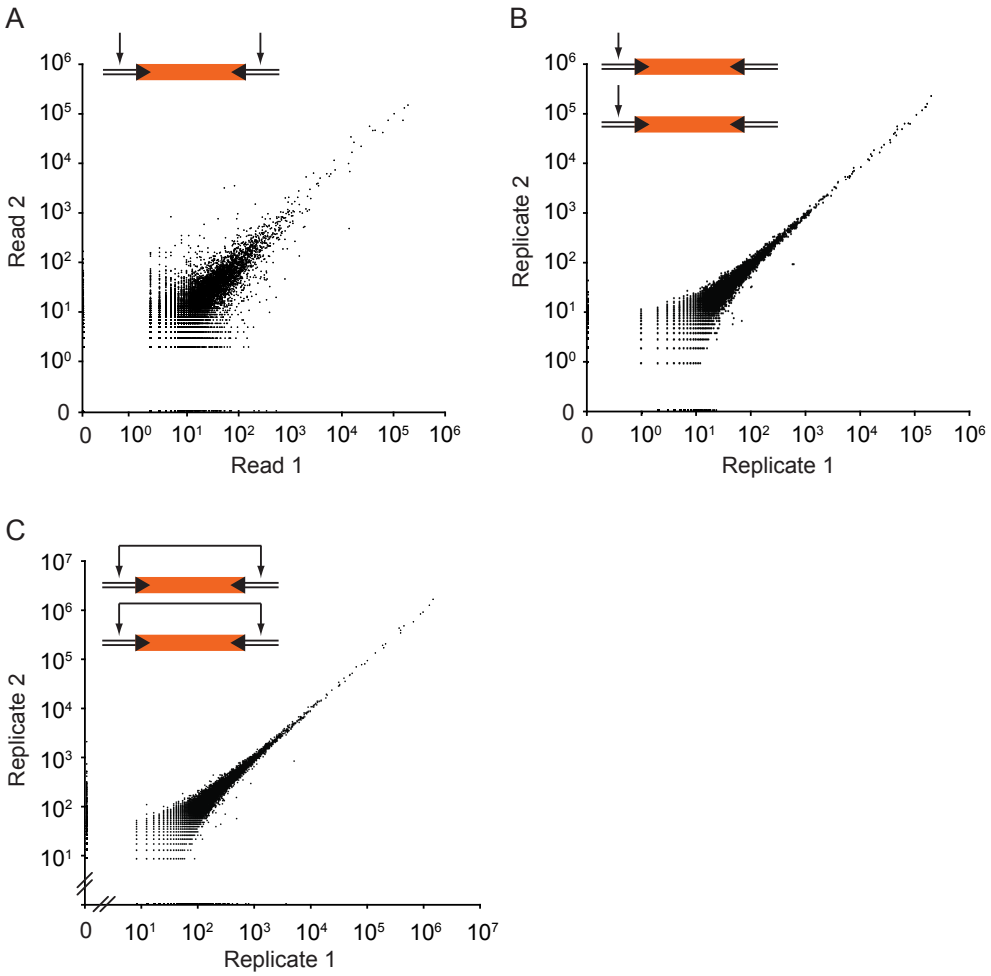

A
Read 2
Read 1
B
Replicate 2
Replicate 1
C
Replicate 2
Replicate 1

Figure S4. Mapping insertion sites in a near-saturated *B. thetaiotomicron* mutant library. (A) As predicted by the *in silico* model (**Figure S1B**), 16bp reads are sufficient to uniquely map the location for 98% of insertions. Simulation shows the average of ten iterations: observed data represent the average of two independently generated and sequenced mutant populations. Error bars indicate one standard deviation. **(B)** Evaluation of nucleotide bias. Reads from a representative dataset were aligned to account for the 1bp variance in MmeI digestion length, and nucleotide frequencies were calculated at each position. A sequence logo (Crooks *et al.*, 2004) reflecting these frequencies shows the known TA insertion requirement for *mariner* transposition (Bryan *et al.*, 1990). No other nucleotide preferences in insertion site choice, library preparation, or sequencing efficiency were detected. **(C)** In addition to its ~6.3 Mb circular chromosome, *B. thetaiotaomicron* possesses an ~30 kb plasmid; insertions segregate between the bacterial chromosome (blue) and plasmid (red) in proportion to their relative size. **(D)** Rarefaction analysis of genes hit directly (3′ 10% of each gene excluded), or by polar effect, in simulations and in observed sequencing results. The cumulative number of genes disrupted (excluding insertions in the 3′ 10% of each gene) was evaluated with every 100 unique insertions added to the collector's curve. Operon structure was taken into account by tracking predicted polar effects of each insertion (including intergenic insertions and those in the 3′ 10% of an ORF). Two genes were predicted to be in the same operon if the pairwise operon probability for every gene between them (inclusive) was at least 90% (Sonnenburg *et al.*, 2005; Westover *et al.*, 2005). For the *in silico* simulation, TA dinucleotide insertions were sampled randomly with replacement. This model suggests that a mutagenized population of ~35,000-strain complexity is approaching saturation; the difference between the simulation and observed results likely reflects essential genes. **(E)** Insertion site density shows some correlation with gene expression (R^2 of log-adjusted values = 0.33, $p<0.0001$). For each gene, the number of observed insertion sites across two independently generated *B.*

thetaiotaomicron mutant populations was adjusted for gene length by correcting for the number of possible insertion sites in that gene ('TA' dinucleotides, excluding sites in the 3' 10% of each ORF, and sites for which paired 16bp reads would not produce an unambiguous localization). These values were log-transformed and plotted against corresponding log-transformed gene expression levels of exponentially growing *B. thetaiotaomicron* cultures in TYG medium [profiled using custom Affymetrix GeneChips (GEO accessions GSM40904 and GSM40905) (Sonnenburg *et al.*, 2005)]. A plot of insertion read count against expression level shows similar results (data not shown).

Goodman *et al.*, Supplemental Fig. S4

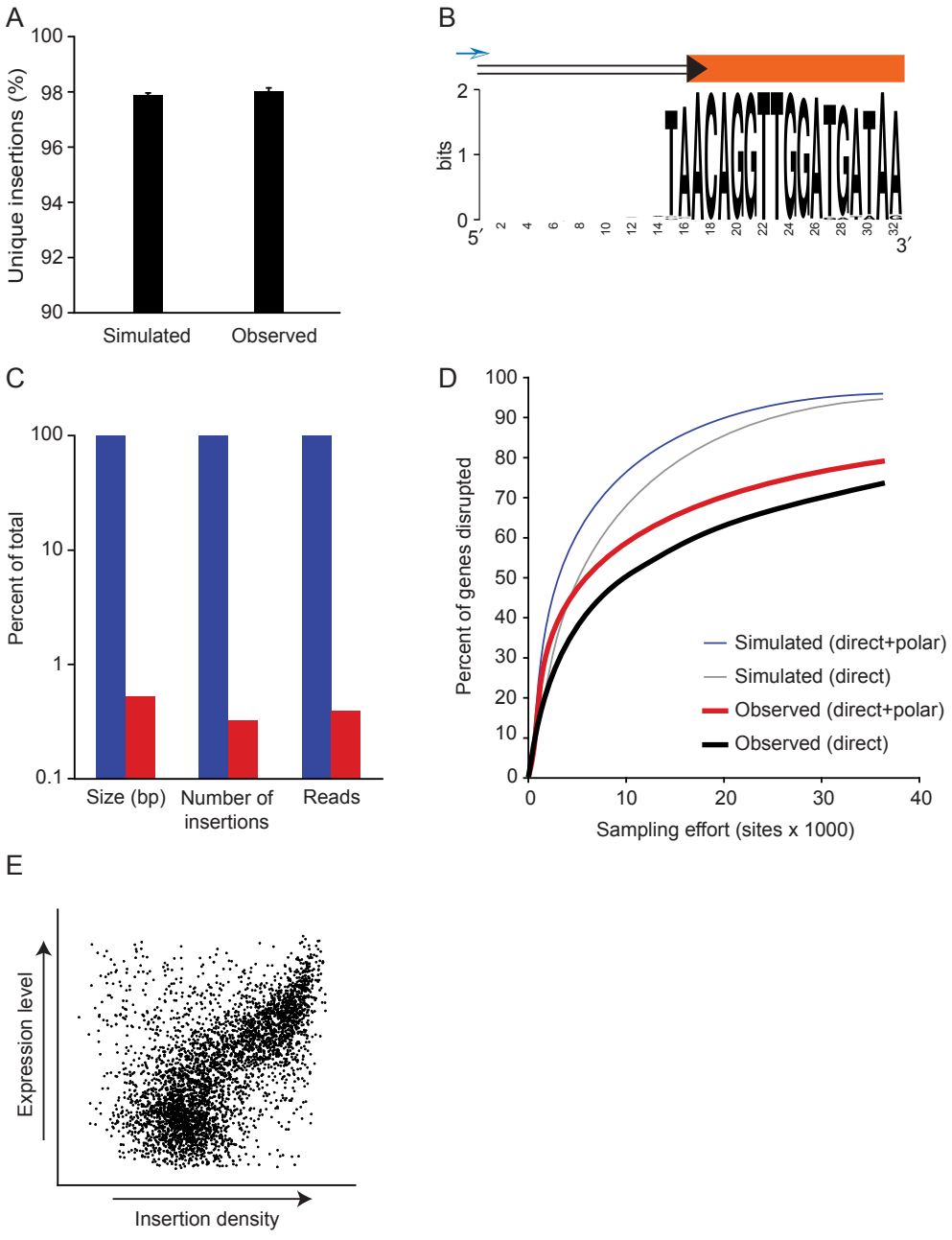


Figure S5. Rationale for arrayed library mapping by combinatorial pooling. (A) 24-bit strings extended from the [14, 23, 5] Wagner series by adding an even parity bit have a minimum Hamming distance of 6. **(B)** Clonality produces a bitwise OR. Multiple archived strains A and B mapping to the same genomic coordinate will produce an incorrect output string for that location; an error will result if this (A or B) string has been assigned to another strain in the archived collection. **(C)** To avoid these errors, we determined the bitwise OR sums for all pairwise combinations of [24,14,6] strings and rank-ordered the strings by the frequency in which they contributed (as terms or sum) to potential clonality errors (A or B = C in which A, B, and C are all [24,14,6] strings). Subsampling the [24,14,6] set in this rank-order (red) prevents these clonality errors until >13,000 strings are sampled (arrow); subsampling these same strings in a random order accumulates clonality errors at the rate shown in black (ten iterations shown). **(D)** To translate the sequencing data into binary strings, the relative abundance of each barcode was normalized across pools and a cutoff was determined based on the mean and standard deviation of the barcode distribution. Hamming distance was used to match these derived binary strings (insertion locations in the genome) to the corresponding input strings (physical locations of the strain in the archive). **(E)** Histogram of mapping errors. Over 80% of the mapped insertions encode an exact match to a pattern assigned a strain in the archived collection; an additional 10% have a single mismatch. Because at least 6 differences separate the patterns assigned to archived strains, a single mismatch from one such pattern is five mismatches away from the next most likely candidate.

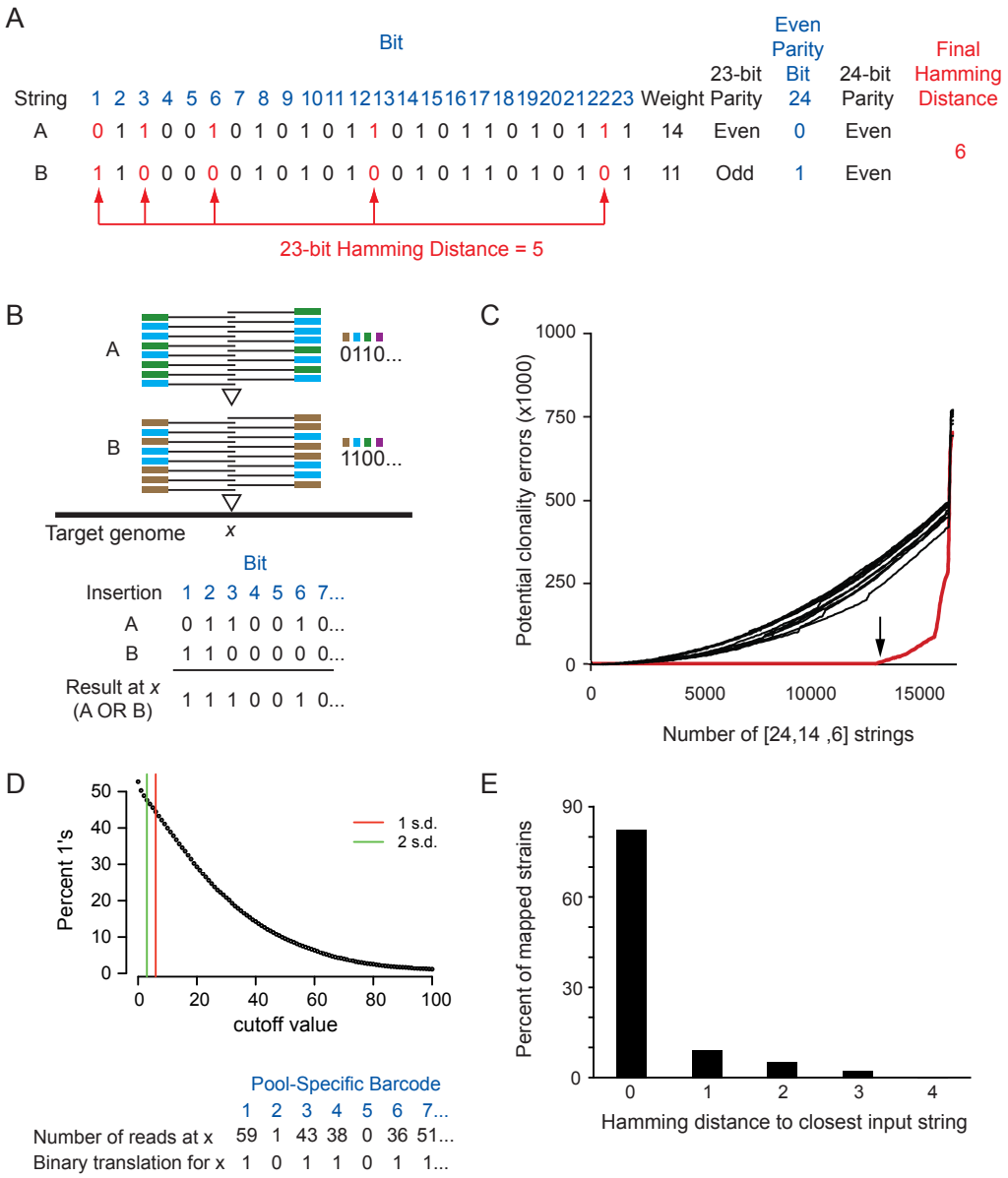

A
Bit
Even Parity Bit 24
Final Hamming Distance
String 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 Weight 23-bit Parity 24-bit Parity
A 0 1 1 0 0 1 0 1 0 1 0 1 1 0 1 0 1 1 0 1 0 1 1 14 Even 0 Even
B 1 1 0 0 0 0 0 1 0 1 0 1 0 0 1 0 1 1 0 1 0 1 0 1 11 Odd 1 Even
6
23-bit Hamming Distance = 5
B
A
0110...
B
1100...
Target genome
x
Bit
Insertion 1 2 3 4 5 6 7...
A 0 1 1 0 0 1 0...
B 1 1 0 0 0 0 0...
Result at x (A OR B) 1 1 1 0 0 1 0...
C
Potential clonality errors (x1000)
1000
750
500
250
0
0
5000
10000
15000
Number of [24,14 ,6] strings
D
Percent 1's
50
40
30
20
10
0
0
20
40
60
80
100
cutoff value
1 s.d.
2 s.d.
Pool-Specific Barcode
1 2 3 4 5 6 7...
Number of reads at x 59 1 43 38 0 36 51...
Binary translation for x 1 0 1 1 0 1 1...
E
Percent of mapped strains
90
60
30
0
0
1
2
3
4
Hamming distance to closest input string

Figure S6. Genes with significantly altered representation after continuous exponential growth *in vitro*. (A) Candidate 'neutral intergenic' regions (black lines) examined as negative controls. The intergenic regions that are flanked by the transcriptional termini of both neighboring genes (i.e., regions unlikely to encode coding or regulatory sequences or to confer a selective advantage or disadvantage to the cell) were assembled into gene-sized concatemers and analyzed with the same statistical criteria used to identify differentially represented genes. Genes are shown as gray arrows. **(B)** The relative abundance of mutations in each gene (points) was compared between input and output populations (mean of 4 independent chemostats maintained in exponential phase in TYG medium). The 477 genes that show statistically significant changes in representation are shown in red; others in gray. As a control, 'neutral' intergenic regions were concatenated into gene-length assemblages and subjected to the same statistical analysis (non-significant in black; significant in green).

A

B

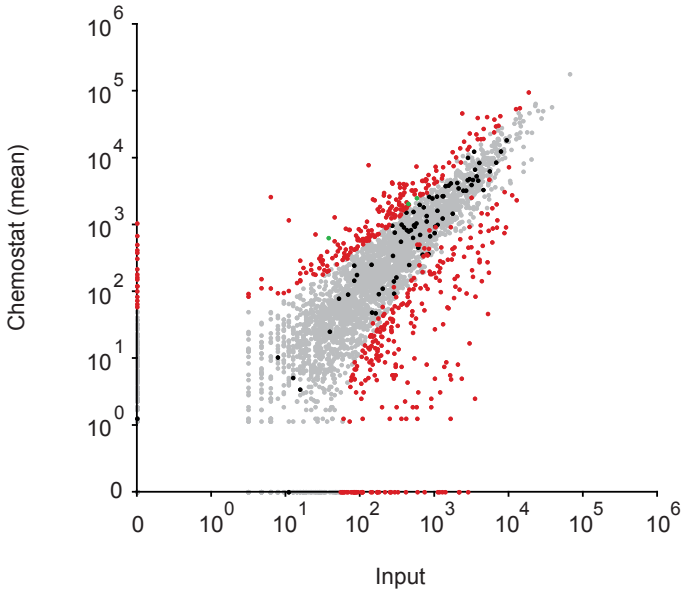


Figure S7. COG category-based classification of genes critical for fitness specifically in amino acid-replete or amino acid–deplete defined minimal medium *in vitro*. Percent representation of COG categories was calculated as number of genes in gene list/number of genes in genome. Significant enrichments in specific COG categories (assessed by comparing these percentages to a null expectation based on the size of the gene list and the representation of a given category in the genome) are marked with asterisks (*Benjamini-Hochberg corrected $p<0.05$; **$p< 0.001$; ***$p<0.0005$).

Goodman *et al.*, Supplemental Fig. S7

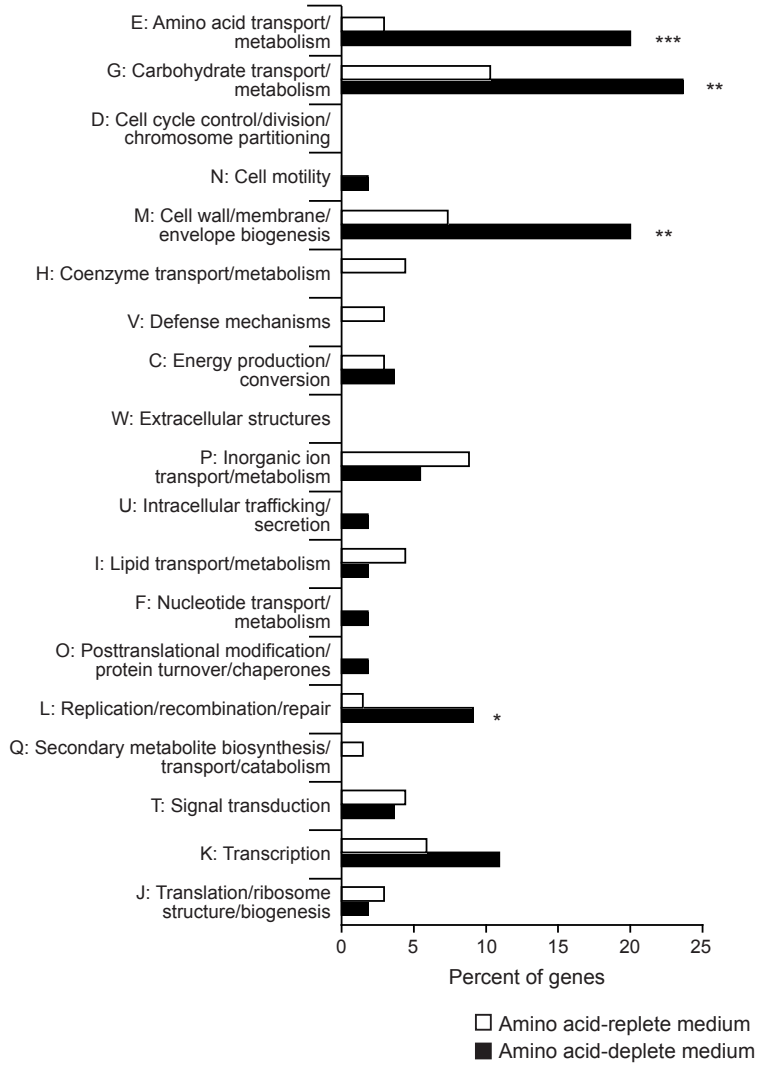


Figure S8. Example of a genetic locus that is specifically required *in vivo*. The abundance of mutants in the *rnf*-like Na^+ NADH:ubiquinone oxidoreductase locus BT0616-22 are shown after *in vivo* monoassociation. Median read counts and individual insertion locations (open arrowheads) in a representative gene in this locus (BT0618, green arrow) are presented at the top. Input counts are in black and output in orange. Gene output to input ratios for individual samples are below; asterisks indicate the fdr-corrected *p*-value (*q*) in each experimental cohort (chemostat in red, monoassociated wild-type mice in tan, brown, and gray bars, $n = 5$ mice/cohort): *$q<0.05$; **$q<0.01$.

Goodman *et al.*, Supplemental Fig. S8

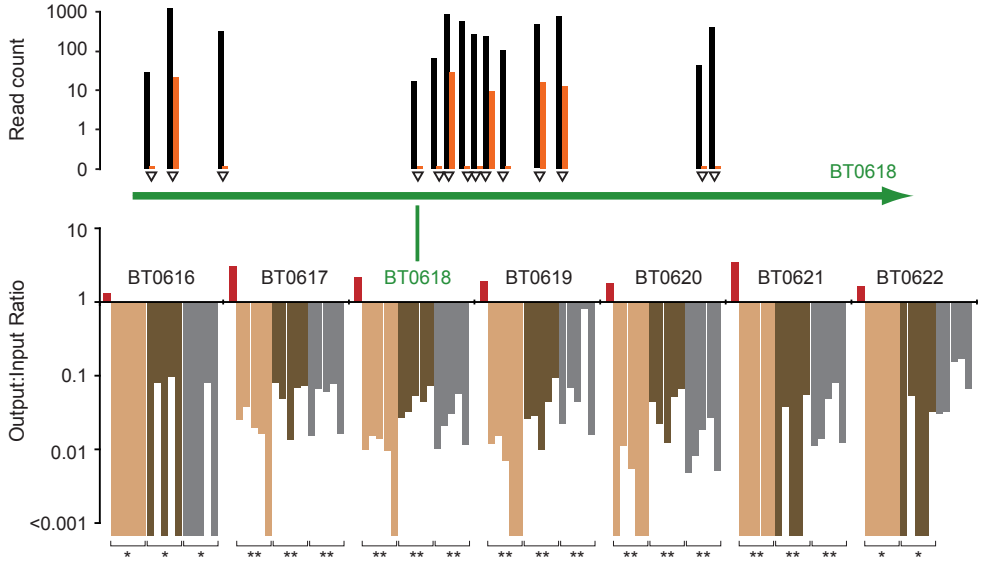


Figure S9. Insertions in the *CPS4* locus do not show a fitness defect during exponential growth *in vitro*. *CPS4* gene insertions (red) show an equal, or slightly increased, relative abundance after prolonged culturing of the mutant population in exponential phase. Non-*CPS4* genes are shown in gray.

Goodman *et al.*, Supplemental Fig. S9

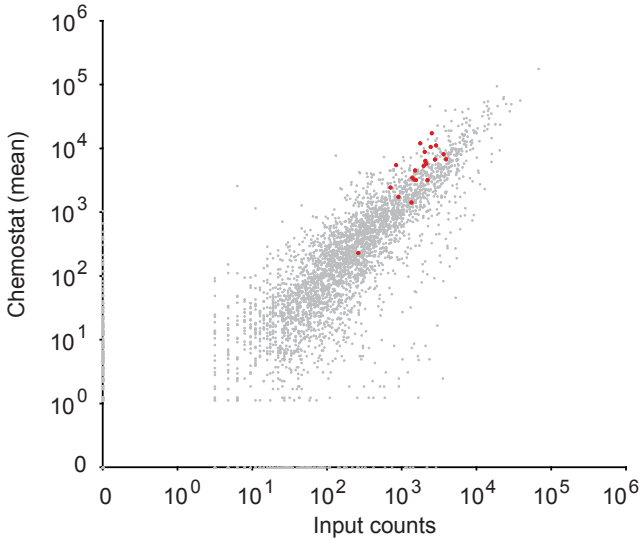
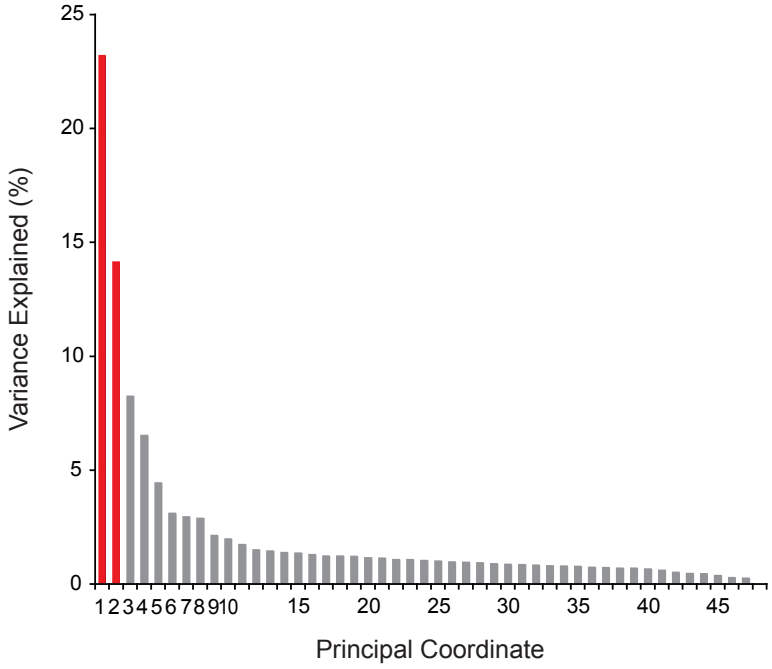


Figure S10. A scree plot of variance explained by principal coordinates analysis suggests that a large portion of the non-noise variance is captured in the first two principal coordinates. PCoA was performed on log-transformed (output/input) ratio data using a Euclidian distance metric.

Goodman *et al.*, Supplemental Fig. S10



Supplemental Experimental Procedures

An *in silico* model of random transposon insertion.

To determine the length of adjacent genomic sequence necessary to unambiguously localize an inserted element, pseudoreads from both sides of 50,000 randomly chosen insertion sites (TA dinucleotides) were generated from, and mapped against, the query genome (10 iterations/readlength). Unique reads occurred only once in the target genome, or only once at a location for which an adjacent read (generated from the other side of the transposon) was also generated in the 100,000-pseudoread set.

Construction of mutagenesis vector pSAM (Sequencing-Adapted Mariner).

The *E. coli* ß-lactamase gene was amplified by PCR from pUC19 (New England Biolabs) using primers MluI Ap 5′ and MluI Ap 3′ (see **Table S19** for a list of all primers used in this study), and subcloned as an MluI fragment into pKNOCK_Cm (Alexeyev, 1999). Platinum Pfx DNA polymerase (Invitrogen) was used for PCR. The *mariner* transposon was constructed by PCR amplification of the *ermG* fragment from pGERM (Salyers *et al*., 2000) using primers ErmGm 5′ and ErmGm 3′, and introduced into pKNOCK_Ap as a KpnI/NotI fragment to create pMar1m. An alternate construct pMar1, lacking MmeI sites in the transposon inverted repeats, was constructed by PCR amplification of the *ermG* fragment from pGERM with primers ErmG 5′ and ErmG 3′. The gene encoding Himar1C9 transposase (Lampe *et al*., 1999) was cloned from pBADC9 (kindly provided by J. Mougous, University of Washington), using primers NdeI Himar1C9 5′ and NotI Himar1C9 3′, as an NdeI/Not fragment to create pSAM. The 300bp region upstream of the *Bacteroides thetaiotaomicron* VPI-5482 *rpoD* gene (BT1311) was PCR-amplified using primers Bt1311 prom 5′ and Bt1311 prom 3′ and mobilized as an NdeI/BamHI fragment to create pMar3m_Bt1311. To introduce the Illumina P7 sequencing adapters into the transposon, the *ermG* fragment of pMar3m_Bt1311 was

PCR-amplified with P7 5′ and P7 3′ and re-cloned into the vector backbone. An *E. coli rrnB* T1/T2 transcriptional terminator cassette was amplified from pFW11 (Whipple, 1998) (kindly provided by Simon Dove, Harvard Medical School) using primers XbaI Term 5′ and PstI Term 3′ and cloned as an XbaI/PstI fragment to create pSAM_Bt. Intermediate and final clones were verified by bi-directional sequencing.

Transposon mutagenesis of *Bacteroides thetaiotaomicron.*

B. thetaiotaomicron VPI-5482 was mutagenized by conjugation. 100mL of an exponential-phase culture (OD_{600}=0.5) of *B. thetaiotaomicron* and 25mL of an exponential phase culture (OD_{600}=0.5) of *E. coli* pSAM_Bt were each pelleted, and resuspended in 1mL TYG medium. These concentrated cultures were then combined and plated as twenty 100μL puddles on brain-heart-infusion (BHI; Becton Dickinson) agar supplemented with 10% horse blood (Colorado Serum Co.). After 5 h of aerobic incubation at 37ºC, conjugation reactions were pooled and resuspended in a total of 50mL phosphate-buffered saline (PBS). 1mL aliquots were then plated on TYG agar BioAssay plates (Nunc) supplemented with gentamicin and erythromycin. Plates were incubated anaerobically at 37ºC for 24h. Approximately 40,000 colonies were pooled in TYG supplemented with 20% glycerol, and adjusted to an OD_{600} of 2.0 prior to storage at -80ºC in 1mL aliquots. These aliquots were used directly for gavage of gnotobiotic mice.

Growth of *B. thetaiotaomicron* transposon mutant populations in continuous-flow chemostats.

Mutant populations were grown at 37ºC with agitation (100 rpm) in BioFlo-110 fermentors (New Brunswick Scientific) in 1.3L vessels containing 400mL of TYG medium pre-sparged with 80% N_2 /20% CO_2. Cultures were held in exponential phase ($OD_{600} \leq 0.4$) by input pumps delivering pre-sparged fresh medium and rate-matched output pumps.

Growth of *B. thetaiotaomicron* mutant populations in minimal defined medium.

To assess the impact of exogenous amino acids on COG category representation among genes required for fitness *in vitro*, ~10^7 CFU of the *B. thetaiotaomicron* mutant population were inoculated into quadruplicate 1L batch cultures containing anaerobic minimal glucose medium (Martens *et al.*, 2008) in the presence or absence of the 20 standard amino acids (0.01% w/v each). Cultures were incubated anaerobically at 37°C and harvested in mid-exponential phase (OD_{600} ~0.5). INSeq libraries were prepared as above using barcoded dsDNA adapters (**Table S2**), pooled at equimolar concentration, and sequenced using 3 lanes of an Illumina GA2 instrument. The output data were parsed by barcode before filtering, normalization, and mapping as above.

Growth rates of individual strains were measured by inoculating single colonies from BHI blood agar plates into 10mL TYG medium and incubating these cultures anaerobically at 37°C. Stationary phase cultures were washed 3 times in PBS, quantified by A_{600}, and inoculated into triplicate 10mL cultures at a starting OD_{600} of 0.001 for subsequent measurement of growth at 37°C under anaerobic conditions.

Additional details about the colonization of gnotobiotic mice.

Wild-type C57BL/6J mice or isogenic mutant strains homozygous for null alleles of the *Rag1* or *Myd88* genes were maintained in flexible plastic gnotobiotic isolators under a strict 12h light cycle (Hooper *et al.*, 2002). Animals were fasted for 4h before oral gavage with 0.5mL of medium containing the 35,000-member *B. thetaiotaomicron* transposon mutant population (~10^8 CFU) alone (monoassociation) or with the mutant population plus a consortium of other sequenced human gut-derived microbes: 6 other Bacteroidetes species, 8 human gut Firmicutes/Actinobacteria, or all 14 of these bacterial species. Differential plating (± erythromycin; aerobic/anaerobic) of fresh cecal contents, and PCR analysis of selected colonies, confirmed that the transposon mutant population

did not harbor donor *E. coli* or unmutagenized *B. thetaiotaomicron* (data not shown). For assembly of multispecies consortia, cultures of each species were grown in TYG_s (described below) in an anaerobic chamber to late-exponential/early stationary phase, concentrated by centrifugation, and supplemented with 50% glycerol (20% final concentration). Cell aliquots were stored at -80°C in 1.8ml glass E-Z vials (Wheaton, Millville, NJ) to protect strict anaerobes from oxygen exposure during storage. The viability of these species-specific aliquots was evaluated prior to pooling by plating on BHI blood agar. Multi-species input pools were assembled 1-2h prior to host inoculation under anaerobic conditions using freshly thawed aliquots of each species. Input volumes were normalized using viability estimates such that each member was equally represented (with the exception of the *B. thetaiotaomicron* transposon mutant population, which comprised 15-20% of the total cells in each multi-species input community). Low recovery rates were observed for frozen aliquots of two Firmicutes (*E. rectale* and *R. torques*). For *E. rectale,* we used cells from an overnight culture in TYG_s medium at early stationary phase. *R. torques* was included in mice 1-5 of the Bacteroidetes+Firmicutes community only.

Mice colonized with multiple Bacteroidetes species or a complete (15-species) community received one dose of $8x10^7$ CFU, while mice colonized with multiple Firmicutes received $1x10^8$ CFU split into two doses over two days. For the latter, *B. thetaiotaomicron* transposon mutants were only included in the first gavage. For all experiments, DNA was isolated from frozen cecal contents by bead beating plus phenol-chloroform extraction as previously described (Ley *et al.*, 2005).

TYG_smedium used to grow community members prior to their inoculation into germ-free mice.

Supplemented TYG medium (TYG_s) consisted of TYG (Holdeman, 1977) supplemented with D-(+)-Cellobiose (0.1% w/v; Sigma), D-(+)-Maltose (0.1% w/v;

Sigma), D-(-)-Fructose (0.1% w/v; Sigma), Tween 80 (0.05% v/v; Sigma), Meat Extract (0.5% w/v; Sigma), ATCC Trace Mineral Supplement (1% v/v), ATCC Vitamin Supplement (1% v/v), N-butyric acid (4mM), propionic acid (8mM), isovaleric acid (1mM), and acetic acid (30mM).

Quantifying bacterial community composition by qPCR.

Primer design. Species-specific primers (**Table S19**) were designed against each of the species in this study using Primer3Plus with a common set of optimal thermal profile parameters (22bp; Tm, 65ºC; G/C content, 54-60%). All primers were tested *in silico* by BLAST for homology to second sites within the target genome and to non-specific sites within the other genomes.

Calculation of a standard curve. Genomic DNA isolated from *in vitro* cultures of each species included in this study was diluted to approximately 1ng/μL and quantified using the Quant-iT dsDNA HS assay kit and a Qubit fluorometer (Invitrogen). Four-point standard curves were prepared from ten-fold serial dilutions of this starting material using 2μL per well in duplicate reactions (i.e. dynamic range of ~2ng-2pg). In every case, the standard curves demonstrated good linearity over four orders of magnitude (R^2 of 0.999-1.000) and good amplification efficiency (93.7-102.8%). Limits of detection ranged from about 13 (*B. caccae*) to 38 (*C. aerofaciens*) genome equivalents.

Reaction setup and run conditions. Duplicate 25μL qPCR reactions were run containing 12.5μL Brilliant® II SYBR Green QPCR master mix (Stratagene), 9μL nuclease-free water, 1.5μL of primers at a 0.3μM [final] each, and 2.0μL of template DNA. Real-time amplification was carried out on a MX3000P system (Stratagene) running MxPro software using the manufacturer's recommended 2-step cycling protocol: 1 cycle at 95ºC (10 min), 40 cycles at 95ºC (30 sec), 60ºC (1 min). Dissociation curve analysis followed (1 min at 95ºC, ramp down to 55ºC and ramp up to 95ºC at default 0.2ºC/sec) to confirm the amplification of a single product in each reaction.

Data analysis. qPCR-derived estimates of the mass of DNA contributed by each species to the total mass of DNA in each sample were converted to genome equivalents using the estimated molecular weight of each species' genome. In cases where a finished genome was not available, the aggregate length of all contigs in the deep-draft assembly was used to estimate molecular weight. Proportional representation was calculated as the number of genome equivalents contributed by a species to the total number of genome equivalents in the sample.

Mapping and quantification of transposon insertion junctions.

MapSAM, a Perl-based software package for mapping and quantifying tranposon insertion locations, is available for download as Supplemental Data. The software requires three types of input files: .fasta- and .ptt-format file(s) for each DNA template (chromosome or plasmid) and SCARF-format sequence output files produced by the Genome Analyzer Pipeline Software (Illumina). Reads are filtered for the presence of a specified sequence tag (in this case, the inverted repeat sequence of the SAM transposon), and clustered into sets that share the same chromosomal sequence. Insertions identified by paired reads (one from each side of the transposon) are retained in the output files. A variety of output formats are provided, including gene-by-gene, insertion-by-insertion, and compilations of the raw reads assigned to each insertion.

Statistical analysis of transposon mutant populations.

Identifying genes with altered relative abundance in input and output populations. To assign a probability that the representation of mutants in a gene differs between input and output populations, mapped reads from each data set were median-normalized (**Table S18**) and a sum read count was calculated for each gene. Insertions in the distal (3') 10% of each gene, or with a left-right read-pair ratio >10 were discarded. Because a true abundance value could not be assigned for genes undetected in a given sample, we

examined median-normalized technical replicates to estimate sampling error. The mean read count for genes detected in a single technical replicate was 19; we added this number to each gene abundance value before calculating an output:input ratio for each gene in each sample. A z-test was used to identify genes whose log-transformed output:input ratios were significantly different from the overall distribution. After applying a multiple hypothesis testing correction (Storey and Tibshirani, 2003), genes with a q-value of < 0.05 were considered significantly altered from the input. Data filtering, normalization, mapping, and statistical analysis were conducted in Perl and MatLab.

Identification of enriched functional categories. A binomial test was used to identify functional (COG) category enrichment among the essential genes, and genes that showed a significantly altered representation in input and output populations. For the list of essential genes, the fraction of genes in each COG category was compared to their fraction in the genome. For the output populations, the COG category fractions were instead compared to the fraction in the input population (i.e., the genes that had a count above zero in at least one sample). P-values resulting from the binomial test were corrected for multiple hypothesis testing based on the number of functional categories queried (Benjamini and Hochberg, 1995). To establish that the methods used were robust to the precise significance cutoff, we repeated the analysis at $p<0.1$ and $p<0.01$ with similar results (data not shown).

Evaluating relationships between transposon mutant populations. Unsupervised hierarchical clustering (neighbor-joining) and principal coordinates analysis (PCoA) was performed on log-ratio data using a Euclidian distance metric and python code from the PyCogent toolkit (Knight *et al.*, 2007). The confidence in the cluster nodes was assessed by bootstrapping. Specifically, for 1000 replicates, the genes in the *B. thetaiotaomicron* genome were randomly sampled with replacement so that the same number of genes were in the data table, but some were present multiple times and some not at all. A

neighbor-joining tree was constructed using a matrix of Euclidean distances between the log ratios of output/input abundance of this sampling of genes across all mutant populations. Support for each node was calculated as the fraction of bootstrapped clusters that recovered the node.

Random forest analysis. To identify the genes responsible for the clustering of mutant populations from monoassociated mice apart from those co-colonized with other Bacteroidetes, we used a random forest classifier (Breiman, 2001) as implemented in the R project for statistical computing (Liaw and Wiener, 2002). The samples were divided into two classes based on the clustering pattern observed after unsupervised hierarchical clustering and PCoA analysis: (1) "monoassociation", transposon mutant output populations from wild-type, *Rag1*-/-, and *Myd88*-/- mice that were colonized only with the *B. thetaiotaomicron* mutant population (n=24); and (2) "community", outputs from mice that were co-colonized with either the Bacteroidetes/Firmicutes/Actinobacteria community or the Bacteroidetes-only community (n = 15). *In vitro* (chemostat) samples and those from Firmicute/Actinobacteria-only co-colonized mice were excluded. The input data table contained the median-normalized log-ratios of the output versus input mutant counts. The variable importance was estimated for each gene using the mean decrease in accuracy measurement (Breiman, 2001). The 220 genes with positive scores (**Table S14**) were divided into two groups: (1) those that had lower average log-ratios in the monoassociations, indicating that these genes were more important when other Bacteroidetes were not present (144 genes); and (2) those that had lower average log-ratios in the community samples, indicating that these genes had increased importance when other Bacteroidetes were present (76 genes). Enrichment in predicted functional categories was determined by binomial test, as described above.

Construction and mapping of an arrayed mutant strain collection.

B. thetaiotaomicron strain VPI-5482 was mutagenized with pSAM_Bt as above. To avoid picking sister clones, ~10% of the mutagenesis reaction was plated on TYG agar supplemented with gentamicin and erythromycin and the remainder discarded. Approximately 11,000 single colonies were picked into 96-well culture trays containing 250μL TYG supplemented with gentamicin and erythromycin. Trays were incubated for 2d at 37ºC in anaerobic jars (GasPak Plus Anaerobic System, BD Biosciences). 30μL of each culture was added to duplicate 96-well storage trays containing 30μL TYG 40% glycerol. These storage trays were immediately sealed with foil lids and frozen at -80ºC; culture trays were stored at 4ºC for pooling. An EpMotion 5075 Automated Pipetting System (Eppendorf) was used to place each strain in a unique subset of 24 pools. To this end, each strain was assigned a unique 24-bit binary string (**Table S3**; rationale described below). To circumvent the EpMotion graphical interface (which would require a point-and-click action for each of the ~120,000 strain-into-pool allocation commands), we wrote a custom Perl script to directly translate the 24-bit strings into the .dws-format, German-language code used by the EpMotion hardware. Each .dws file matched a specific set of 5 library trays (480 strains) and instructed the EpMotion to transfer 3μL of culture into the appropriate 2mL pool tubes. To avoid cross-contamination, aliquots were dispensed from a fixed height above the pools, and tips were discarded between strains. Pools were stored at -20ºC after allocation. Sequencing libraries were prepared from each of the pools as described above except each was constructed at half scale (25μg input DNA) and pool-specific dsDNA adapters bearing unique 4bp barcodes (**Table S2**) were used in the ligation step. Final libraries were adjusted to 10nM and combined at equimolar concentration into 2 collections of 12 pools each. Each collection was sequenced at 1.5pM on a single lane of an Illumina Genome Analyzer I. The strategy for mapping sequencing data to binary strings (and specific strains in the archived collection) is provided below.

*Rationale for string assignment***.** Each strain in the archived collection was assigned a unique 24-bit binary string that instructs the allocation of that strain across 24 pools ("1" = strain present in pool; "0" = strain absent in pool). In this way, the number of errors required to mistake one strain for another is equivalent to the Hamming distance (number of differences) between the strings associated with the two strains. Consequently, one criterion for string assignment is to maximize the Hamming distance between each string and its most similar neighbors (**Figure S5A**). To this end, we began with the quasi-perfect [23, 14, 5] Wagner code, which uses 23 bits to encode 2^{14} 14-bit source strings with a minimum Hamming distance of 5 between strings (Simonis, 2000; Wagner, 1966). Because this minimum Hamming distance is an odd number, strings A and B with the most similarity will have different parity [the sum of the bits (weight) of one string will be odd and the weight of the other will be even]. Further, adding a 24^{th} even-parity bit to each string (1 if the 23-bit weight is odd, 0 if even) will increase the Hamming distance between strings A and B to 6 because this parity bit will differ for the odd- and even-weight strings. Calculation of the Hamming distances between all pairwise combinations of the 16384 (2^{14}) 24-bit strings produced by this [24, 14, 6] code confirmed this result (data not shown).

We predicted that a second type of error could result from clonality: if identical strains (arising by chance or originating from the same parental transposon insertion) are placed into multiple wells in the archived collection, the reads that map to the shared genomic insertion location will be coupled with the pool-specific barcodes associated with multiple archived strains (**Figure S5B**). In this way, clonality produces a bitwise OR function: if any of the source strings (terms) encodes a "1" at a certain bit, the output string (sum) will have a "1" at that bit. To avoid mistaking this sum for another string (which would incorrectly pair an insertion location with the wrong archived strain), we calculated the OR-sums between all pairwise combinations of the [24, 14, 6] set of 24-bit strings. For each pairwise OR, we determined whether the sum was also present in the

16384 strings (*i.e.*, was also a product of [24, 14, 6] encoding). By rank-ordering the strings by the frequency at which they contributed to these potential clonality errors (as a term or sum), we were able to select a 13,000-string subset for which no pairwise OR produces another string in the set (**Figure S5C**).

Mapping sequencing data onto binary strings. To match an insertion location with a binary string (and its corresponding archived strain), we first retrieved all 32-bp raw sequences bearing the left- or right- hand reads corresponding to a given insertion site. From these full-length sequences, we calculated the number of reads bearing each of the 24 pool-specific barcodes, normalized the counts across barcodes, and translated this distribution into a binary string (0 if the number of reads containing a certain barcode was two standard deviations below the mean abundance for that barcode; 1 otherwise) (**Figure S5D**). For each of these derived insertion-associated output strings, we determined the minimal Hamming distance to a corresponding archive-associated input string: non-unique matches (output strings matching multiple input strings at equal distance) were discarded (**Figure S5E**).

Quantification of BT1954 and BT1956 gene expression by qRT-PCR.

Quadruplicate 10mL cultures were grown anaerobically at 37ºC in defined minimal glucose medium (Martens *et al.*, 2008) with varying concentrations of vitamin B_{12} and harvested at mid-log phase (OD_{600} ~ 0.4). qRT-PCR targets were prepared as described (Sonnenburg *et al.*, 2005) using primer pairs BT1954 qPCR 5/BT1954 qPCR 3 and BT1956 qPCR 5/BT1956 qPCR 3 (**Table S19**). Real-time amplification was carried out as described (Bjursell *et al.*, 2006) on a MX3000P system (Stratagene) running MxPro software. Reaction conditions were 1 cycle at 95ºC (15 min), 40 cycles at 95ºC (15 sec), 55ºC (45 sec), 72ºC (30 sec), fluorescence measurement at 82ºC. Dissociation curve analysis followed (1 min at 95ºC, ramp down to 70ºC and ramp up to 95ºC at default 0.2ºC/sec) to confirm the amplification of a single product in each reaction. Expression

levels were normalized based on parallel reactions with primer pairs specific to *B. thetaiotaomicron* 16S rRNA (Sonnenburg *et al.*, 2005) with a collection temperature of 84ºC.

Assessment of vitamin B_{12} auxotrophy.

Single colonies from BHI blood agar plates were re-streaked on rich defined MOPS medium agar (Neidhardt *et al.*, 1974) (methionine and vitamin B_{12} omitted) supplemented with L-cysteine HCl (0.05% w/v), D-(+)-Cellobiose (0.1% w/v; Sigma), D-(+)-Maltose (0.1% w/v; Sigma), D-(-)-Fructose (0.1% w/v; Sigma), Tween 80 (0.05% v/v; Sigma), (N-butyric acid (4mM), propionic acid (8mM), isovaleric acid (1mM), acetic acid (30mM), and Resazurin (0.1% w/v, Sigma). Trace minerals and vitamins were replaced by the Trace Mineral Supplement (1% v/v; ATCC) and Vitamin Supplement (1% v/v, ATCC formulation with vitamin B_{12} omitted). Filter-sterilized 2x medium (supplemented with the desired concentration of vitamin B_{12}) was combined with autoclaved 2x Noble Agar (final concentration 1.5% w/v, Difco) in an anerobic chamber and plates were poured immediately. Plates were stored anaerobically for 2d before use and incubated anaerobically at 37ºC for 4d after streaking. Colonies appearing on B_{12}-negative plates were restruck on $+B_{12}$ and $-B_{12}$ to confirm viability.

Supplemental Protocol: Preparation of a sequencing library from a bacterial mutant population.

Materials needed
Phosphate-buffered saline (PBS)
Screw-cap tubes, 2mL
Extraction buffer (200mM NaCl, 200mM Tris pH 8, 20mM EDTA)
425-600μm acid-washed glass beads (Sigma)
20% SDS
Phenol:chloroform:isoamyl alcohol (PCI) (25:24:1, pH 7.9) (Ambion)
Phase Lock Gel Light tubes, 2mL (5Prime, Gaithersburg, MD)
3M NaOAc (pH 5.2)
Microfuge tubes, 0.5, 1.5, and 2mL
100% Isopropanol
100% Ethanol
70% Ethanol
TE (10mM Tris, 1mM EDTA, pH 8.0)
DNeasy Blood and Tissue Kit (Qiagen)
Glycogen (Roche)
MmeI (New England Biolabs)
PCR Purification Kit (Qiagen)
5x TBE running buffer (450mM Tris, 450mM Boric acid, 10mM EDTA)
5x PAGE sample buffer (5mL 5x TBE running buffer, 7.5g Ficoll 400, 0.05g Bromophenol Blue, 0.05g Xylene Cyanol, H_20 to 50mL final)
100bp DNA ladder (New England Biolabs)
6% TBE acrylamide gels (1.0 mm, 10 wells) and apparatus (Invitrogen)
21-gauge needles
Siliconized (nonstick) tubes, 1.5mL (Ambion)
LoTE buffer (3mM Tris, 0.2mM EDTA, pH 7.5)
7.5M NH_4OAc
Spin-X tubes (Sigma)
LIB_AdaptT adapter oligonucleotide, HPLC purified
LIB_AdaptB adapter oligonucleotide, HPLC purified
T4 DNA ligase (2000000 U/mL) (New England Biolabs)
PEG-8000
LIB_PCR_5 primer
LIB_PCR_3 primer
Platinum Pfx DNA polymerase (Invitrogen)
10bp DNA ladder (Invitrogen)
Tween 20 (Sigma)

Equipment needed
Bead-beater (BioSpec Products, Bartlesville, OK) *(for disruption of cecal material)*
SpeedVac concentrator (Thermo Scientific)
Qubit fluorometer (Invitrogen) *(or alternative for DNA quantification)*
PCR machine
Illumina Genome Analyzer *(or alternative for high-throughput sequencing)*

Protocol overview
1. Extraction of crude DNA
2. Purification of crude DNA
3. MmeI digestion and library purification
4. Creation of dsDNA adapter molecule

5. Ligation of dsDNA adapter molecule
6. PCR amplification

Procedure 1 - Extraction of DNA from cecal contents
For cecal contents, add 500 μL PBS per gram to allow vortexing
Vortex and spin gently (100 x g, 0.5 min, 25ºC)
To each screw-cap tube (6/sample), add:
- 0.75g glass beads
- 500 μL extraction buffer
- 100 μL sample
- 210 μL 20% SDS
- 500 μL PCI

Disrupt cells by bead beating (maximum setting, 2 min) and place samples on ice
Spin tubes (7000 x g, 3 min, 4ºC)
Transfer aqueous phase (~600 μL) to Phase Lock Gel tube
Add 1 volume PCI, mix by inversion
Spin tubes (16,000 x g, 5 min, 25ºC)
Transfer aqueous phase (~600 μL) to 2 mL microfuge tube
Add 0.1 volumes (~ 60 μL) NaOAc
Add 1 volume (~660 μL) 100% isopropanol
Mix by inversion
Precipitate DNA at -80ºC for 1 h or -20ºC overnight
Spin tubes (22,000 x g, 30 min, 4ºC)
Wash pellet twice with 1 mL 70% ethanol
Dry samples in SpeedVac concentrator (1-3 min, 25ºC)
Resuspend in 200 μL TE
Incubate at 50ºC until resuspended

Procedure 2 – Further purification of bacterial DNA
Move forward with half of material (6 x 100 μL):
Add 300 μL buffer ATL, 3μL RNase A (100 mg/mL stock)
Incubate 2 min at 25ºC
Add 40 μL Proteinase K, 400 μL buffer AL
Incubate 30 min at 70ºC
Add 40 μL NaOAc
Confirm that pH is below 7.0
Add 440 μL 100% ethanol, mix by inversion
Add 750 μL of sample to DNeasy column, spin 6,000 x g for 1 min, repeat
Add 500 μL buffer AW1, spin 6,000 x g 1 min
Add 500 μL buffer AW2, spin 22,000 x g 1 min
Replace collection tube, spin 22,000 x g 1 min
Remove any excess buffer from inner beveled rim of column, spin 22,000 x g 3 min
Elute DNA: Add 200 μL buffer AE, incubate 1 min 25 ºC, spin 22,000 x g 1 min, repeat
Pool two eluates (400 μL total/column)
Ethanol precipitate:
Add 0.1 volume (40 μL) NaOAc
Add 2 μL glycogen
Add 2.5 volumes (1100 μL) 100% ethanol
Precipitate DNA at -80ºC for 1 h or -20ºC overnight
Spin tubes (22,000 x g, 30 min, 4 ºC)
Wash twice with 1 mL 70% ethanol
Dry samples in SpeedVac concentrator (1-3 min, 25ºC)
Resuspend in 10 μL TE per column.

Pool like samples (60 μL total) and quantify DNA.

Procedure 3 - MmeI digestion and library purification
Combine in a 1.5 mL tube on ice:
50 μg DNA
120 μL 10x NEBuffer 4
50 U MmeI
1.875 uL 32 mM SAM (50 mM final)
H_20 to 1.2mL
Incubate at 37ºC for 1 h followed by 80ºC for 20 min
Split reaction into 6 x 200 μL
Run each aliquot over a PCR purification column as per manufacturer's instructions
Elute in 40 μL Buffer EB and pool like samples (240 μL total)
Add 0.2 volumes PAGE sample buffer and load on 8 central lanes of a 6% TBE-acrylamide gel (~35 μL/lane). Load 100bp DNA ladder in lanes 1 and 10 of the gel.
Run gel at 200V for 1 h
Stain gel with ethidium bromide and rinse twice with TBE
With a new razor blade, excise the region of each lane corresponding to 1.2-1.5kb.
Extract DNA by crush/spin:
Pierce the bottom of 0.5 mL microfuge tubes with a 21-gauge needle (pierce tubes from inside-out)
Place each pierced tube in a siliconized 2 mL microfuge tube
Add two gel fragments to each 0.5 mL tube
Spin tubes (22,000 x g, 4 min, 25ºC)
Discard 0.5 mL tubes, add 250 μL LoTE and 50 μL NH_4OAc to gel slurry
(Samples can be stored at 4ºC overnight if necessary)
Incubate at 65ºC for 4 h, vortexing occasionally
Place 5 μL LoTE on the membrane of each Spin-X tube (1 Spin-X tube/slurry tube)
Add gel slurry to SpinX tube
It may be necessary to cut the tip off a 1mL pipet to accommodate gel fragments
Remaining gel fragments can be transferred using a disposable plastic loop
Spin tubes (22,000 x g, 5 min)
Add 50 μL LoTE to Spin-X tube
Spin tubes (22,000 x g, 5 min)
Transfer flowthrough to 1.5 mL microfuge tube and discard Spin-X tubes
Add 2 μL glycogen
Add 1/3 volume 7.5M NH_4OAc
Add 2.5 volumes 100% ethanol
Precipitate DNA at -80ºC for 1 h or -20ºC overnight
Spin tubes (22,000 x g, 30 min, 4ºC)
Wash pellet twice with 1 mL 70% ethanol
Dry samples in SpeedVac concentrator (1-3 min, 25ºC)
Resuspend each tube in 5 μL LoTE, incubate at 37 ºC for 10 min
Pool like samples (20 μL total)
Quantify samples before proceeding to ligation step.

Procedure 4 - Creation of dsDNA adapter molecule
Dilute LIB_AdaptT, LIB AdaptB to 100 μM in EB
Combine in 1.5 mL microfuge tube:
15 μL 100 μM LIB_AdaptT
15 μL 100 μM LIB_AdaptB
1.5 μL 1M NaCl (50mM final)
Heat 95 ºC for 5 min, cool slowly to 4 ºC at 0.1 deg/sec

Aliquot before storing 50 μM aliquots at -20ºC
Thaw aliquots on ice before use.

Procedure 5 - Ligation of dsDNA adapter molecule
Mix on ice in order:
- H_20 (to 50 mL final volume)
- 5 μL 10x T4 DNA ligase buffer
- 1 μg library DNA
- 4.4 μL 5μM dsDNA adapter (dilute in 1x T4 DNA ligase buffer)
- 6.25 μL 40% PEG
- 5 μL T4 DNA ligase

Incubate at 16ºC for 12-16 h
Heat inactivate (65ºC, 10 min)
Add 10 μL 5x PAGE sample buffer
Load over two lanes of a 6% TBE acrylamide gel (use 100bp DNA ladder for size standard and separate different samples by multiple blank lanes to avoid cross-contamination)
Run gel at 200V for 1 h
Stain gel with ethidium bromide and rinse twice with TBE
With a new razor blade, excise the region of each lane corresponding to 1.2-1.5kb.
Extract DNA by crush/spin as above
After ethanol precipitation, resuspend DNA in 10 μL LoTE.

Procedure 6 - PCR amplification
Combine PCR reagents on ice:
- H_20 to 60 μL final volume
- 12 μL 10x Pfx Buffer (final concentration is 2x)
- 2 μL 10 mM dNTPs (300 μM final)
- 0.6 μL 50 mM $MgCl_2$ (0.5 mM final)
- 1.2 μL 20 μM LIB_PCR_5 primer (400 nM final)
- 1.2 μL 20 μM LIB_PCR_3 primer (400 nM final)
- 4 μL library DNA (~100 ng)
- 2 μL Pfx DNA polymerase (2.5 U)

Split reaction into 8x7.5 μL
Run PCR as follows:
- 94ºC 2 min
- 18 cycles of:
  - 94ºC 15 sec
  - 60ºC 1 min
  - 68ºC 2 min
- 68ºC 4 min

Pool like samples
Add 12 μL 5x PAGE sample buffer
Load over two lanes of a 6% TBE acrylamide gel (use 100bp and 10bp DNA ladders for size standard and separate different samples by multiple blank lanes to avoid cross-contamination)
Run gel at 200V for 25 min
Stain gel with ethidium bromide and rinse twice with TBE
With a new razor blade, excise the region of each lane corresponding to 125 bp
Extract DNA by crush/spin as above, except incubate gel slurry at 50ºC for 1 h (instead of the 65ºC/4 h incubation)
After ethanol precipitation, resuspend DNA in 10 μL EB.

Quantify carefully. Supplement with EB and 1% Tween (in EB) to a final DNA concentration of 10nM and a final Tween concentration of 0.1%.
Store at -20ºC and adjust to 1-2 pM immediately prior to sequencing.

Supplemental References

Alexeyev, M.F. (1999). The pKNOCK series of broad-host-range mobilizable suicide vectors for gene knockout and targeted DNA insertion into the chromosome of gram-negative bacteria. Biotechniques *26*, 824-826, 828.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological *57*, 289-300.

Bigot, Y., Brillet, B., and Auge-Gouillou, C. (2005). Conservation of Palindromic and Mirror Motifs within Inverted Terminal Repeats of mariner-like Elements. J Mol Biol *351*, 108-116.

Bjursell, M.K., Martens, E.C., and Gordon, J.I. (2006). Functional genomic and metabolic studies of the adaptations of a prominent adult human gut symbiont, Bacteroides thetaiotaomicron, to the suckling period. J Biol Chem *281*, 36269-36279.

Breiman, L. (2001). Random Forests. Machine Learning *45*, 5-32.

Bryan, G., Garza, D., and Hartl, D. (1990). Insertion and excision of the transposable element mariner in Drosophila. Genetics *125*, 103-114.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188-1190.

Holdeman, L.V., Cato, E.D., Moore, W.E.C. (1977). Anaerobe Laboratory Manual (Blacksburg, VA, Virginia Polytechnic Institute and State University Anaerobe Laboratory).

Hooper, L.V., Mills, J.C., Roth, K.A., Stappenbeck, T.S., Wong, M.H., and Gordon, J.I. (2002). Combining gnotobiotic mouse models with functional genomics to define the impact of the microflora on host physiology. Methods in Microbiology *31*, 559-589.

Knight, R., Maxwell, P., Birmingham, A., Carnes, J., Caporaso, J.G., Easton, B.C., Eaton, M., Hamady, M., Lindsay, H., Liu, Z., *et al.* (2007). PyCogent: a toolkit for making sense from sequence. Genome Biol *8*, R171.

Lamichhane, G., Zignol, M., Blades, N.J., Geiman, D.E., Dougherty, A., Grosset, J., Broman, K.W., and Bishai, W.R. (2003). A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to Mycobacterium tuberculosis. Proc Natl Acad Sci U S A *100*, 7213-7218.

Lampe, D.J., Akerley, B.J., Rubin, E.J., Mekalanos, J.J., and Robertson, H.M. (1999). Hyperactive transposase mutants of the Himar1 mariner transposon. Proc Natl Acad Sci U S A *96*, 11428-11433.

Ley, R.E., Backhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D., and Gordon, J.I. (2005). Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A *102*, 11070-11075.

Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. R News *2/3*, 18.

Martens, E.C., Chiang, H.C., and Gordon, J.I. (2008). Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. Cell Host Microbe *4*, 447-457.

Neidhardt, F.C., Bloch, P.L., and Smith, D.F. (1974). Culture medium for enterobacteria. J Bacteriol *119*, 736-747.

Peterson, D.A., McNulty, N.P., Guruge, J.L., and Gordon, J.I. (2007). IgA response to symbiotic bacteria as a mediator of gut homeostasis. Cell Host Microbe *2*, 328-339.

Sonnenburg, J.L., Xu, J., Leip, D.D., Chen, C.H., Westover, B.P., Weatherford, J., Buhler, J.D., and Gordon, J.I. (2005). Glycan foraging in vivo by an intestine-adapted bacterial symbiont. Science *307*, 1955-1959.Salyers, A.A., Bonheyo, G., and Shoemaker, N.B. (2000). Starting a new genetic system: lessons from bacteroides. Methods *20*, 35-46.

Simonis, J. (2000). The [23, 14, 5] Wagner code is unique. Discrete Mathematics *213*, 269-282.

Sonnenburg, J.L., Xu, J., Leip, D.D., Chen, C.H., Westover, B.P., Weatherford, J., Buhler, J.D., and Gordon, J.I. (2005). Glycan foraging in vivo by an intestine-adapted bacterial symbiont. Science *307*, 1955-1959.
Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc Natl Acad Sci U S A *100*, 9440-9445.
Wagner, T. (1966). A search technique for quasi-perfect codes. Information and Control *9*, 94-99.
Westover, B.P., Buhler, J.D., Sonnenburg, J.L., and Gordon, J.I. (2005). Operon prediction without a training set. Bioinformatics *21*, 880-888.
Whipple, F.W. (1998). Genetic analysis of prokaryotic and eukaryotic DNA-binding proteins in Escherichia coli. Nucleic Acids Res *26*, 3700-3706.