

Figure S1, related to Figure 1. Tryptophanase abundance and taxa correlations.

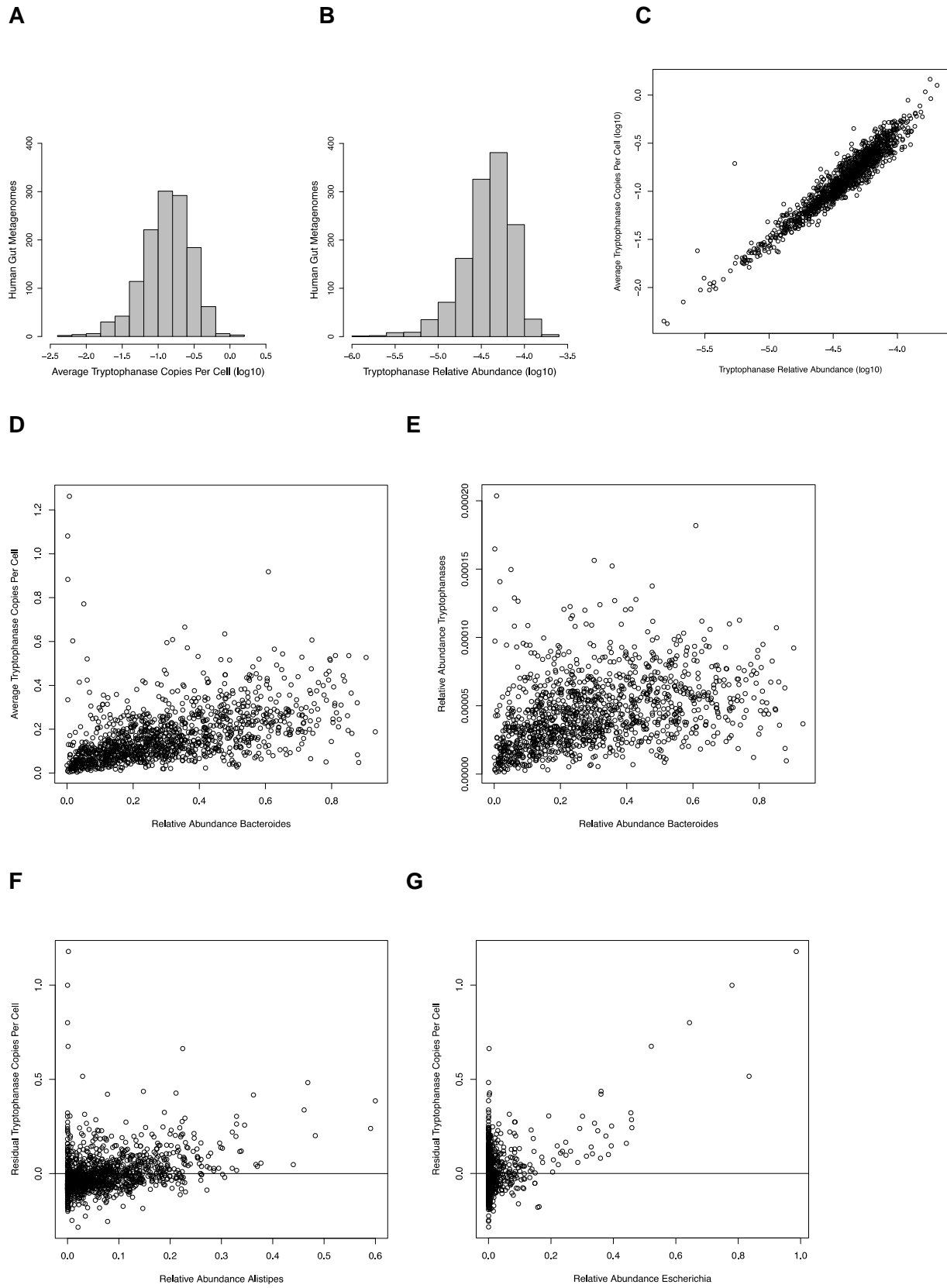


Figure S2, related to Figure 1. Indole production from tryptophan by *Bacteroides* species.

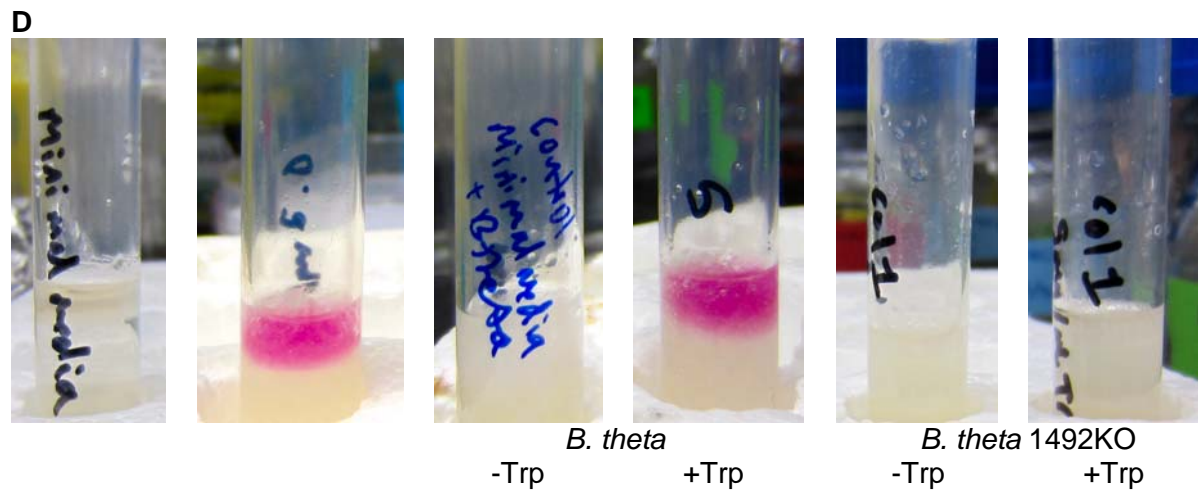
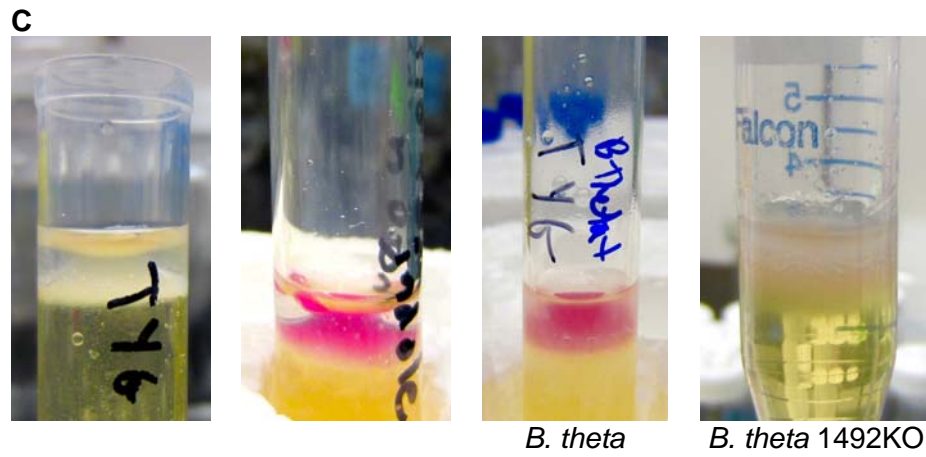
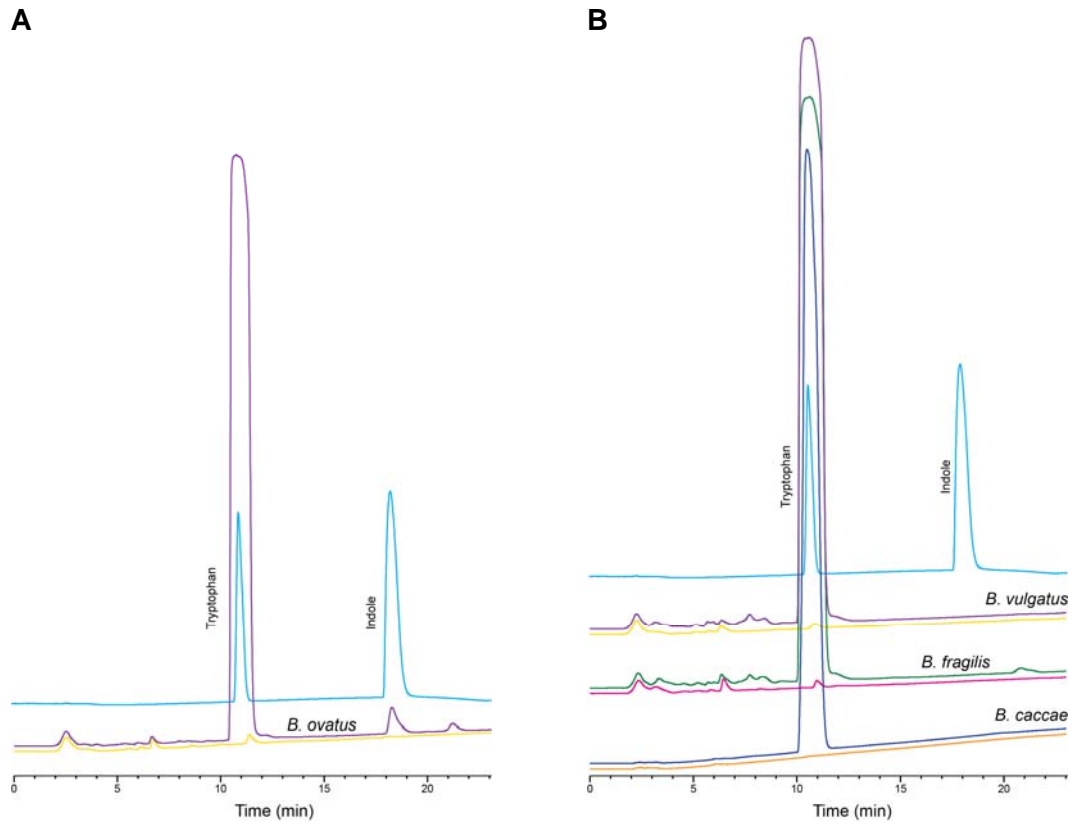
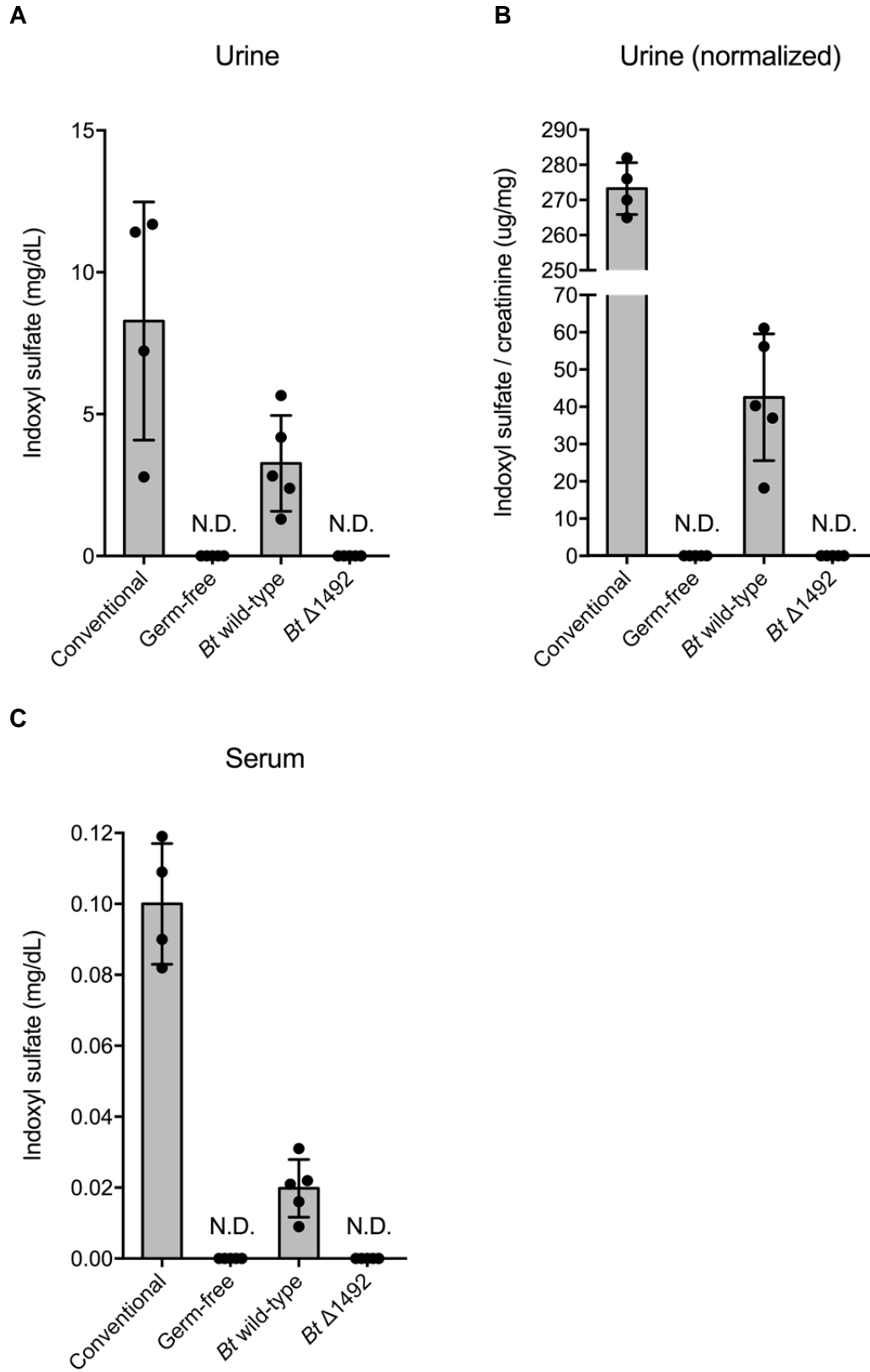


Figure S3, related to Figure 2. BT1492 status determines presence or absence of circulating IS in mono-colonized germ-free mice.



Supplemental Figure and Table Legends

Figure S1, related to Figure 1. Estimated abundance of known and putative tryptophanases across 1,267 human gut metagenomes and correlations with bacterial taxa. (A-B) On average, tryptophanases are found once per 6 microbial cells (A) or once per 22,000 genes (B). (A) Estimated average tryptophanase copies per cell. Axis is on a log₁₀ scale. A value of log₁₀(1.0) indicates that on average each genome contained one gene copy; a value of log₁₀(1e⁻²) indicates that 1 out of every 100 genomes contained the gene. (B) Estimated tryptophanase gene relative abundance. Axis is on a log₁₀ scale. Relative abundance is defined as the proportion of total genes that are tryptophanases. (C) Correlation of (A) and (B). (D-E) Tryptophanase average copy number (D) and relative abundance (E) versus *Bacteroides* relative abundance across 1,267 gut metagenomes ($R^2=0.22$, $p<1e-16$ and $R^2=0.12$, $p<1e-16$, respectively). (F-G) Residual tryptophanase average copy number versus the relative abundance of the genera *Alistipes* (F) and *Escherichia* (G). A positive value on the y-axis indicates that a sample had a higher level of tryptophanase than was expected given the level of *Bacteroides*.

Figure S2, related to Figure 1. Indole production from tryptophan (Trp) by *Bacteroides* species. (A-B) HPLC analysis of indole production by *Bacteroides* species. Clarified supernatant analyzed at t=6h. (A) *B. ovatus* is an indole producer. Yellow: *B. ovatus* plus minimal media only. Purple: *B. ovatus* plus 5 mg/mL Trp. Light blue: Trp and indole standards. (B) *B. caccae*, *B. fragilis*, and *B. vulgatus* are non-producers. Orange: *B. caccae* plus minimal media only. Dark blue: *B. caccae* plus 5 mg/mL Trp. Pink: *B. fragilis* plus minimal media only. Green: *B. fragilis* plus 5 mg/mL Trp. Yellow: *B. vulgatus* plus minimal media only. Purple: *B. vulgatus* plus 5 mg/mL Trp. Light blue: Trp and indole standards. (C-D) Ehrlich's test for indole production from Trp by *Bacteroides* thetaiotaomicron and *B. theta* 1492KO. *B. theta* wild type produces indole, whereas *B. theta* 1492KO does not. Pink color at the organic/aqueous interface indicates the presence of indole. Note that TYG medium contains Trp. (C) Test in TYG medium. *Left to right*: medium only (negative control); + 0.2 mg/mL indole (positive control); + *B. theta* (overnight culture); + *B. theta* 1492KO (overnight culture). (D) Test in minimal medium, t=6h. *Left to right*: medium only (negative control); + 0.2 mg/mL indole (positive control); + *B. theta*; +5 mg/mL Trp + *B. theta*; + *B. theta* 1492KO; +5 mg/mL Trp + *B. theta* 1492KO.

Figure S3, related to Figure 2. Comparison of urinary and serum IS concentrations in mono-colonized germ-free mice. Germ-free mice were monocolonized with either *Bt* wild-type or *Bt* Δ 1492, and serum and urine samples were collected one week post-colonization. Conventional and germ-free mouse groups were included as controls. In contrast to conventional and *Bt* wild-type-colonized mice, the Δ 1492-colonized mice had no detectable (N.D.) IS in urine (A-B) or serum (C) (detection limit 0.003 mg/dL in serum). The relative levels of IS in urine mirrored those in serum. These data suggest that differences observed in urinary IS excretion were due to increased circulating IS, not differences in kidney excretion rate. Data shown are means from four (conventional) or five (*Bt* wild-type, *Bt* Δ 1492, germ-free) biological replicates. Error bars represent standard deviation. Urine data are presented as normalized to creatinine excretion in (B) to account for differences in hydration level across mice.

Table S1A, related to Figure 1. Tryptophanase homologs are highly prevalent in the gut microbiome. We called a reference gene a tryptophanases homolog if had >X% AA identity to one of the 36 query genes, where X is indicated by the 1st column. Both genes needed to be covered by >70%. A gene was called present in a sample if its estimated abundance was at least X, where X is indicated by the second column. Count Samples is number of samples where gene was detected (out of 1,267). Fraction Samples is the fraction of samples where gene was detected (out of 1,267). Mean Copy Number is the average abundance of the tryptophanase homologs across 1,267 samples. **Conclusions:** Tryptophanases (>95% AA ID to known and putative genes) comprise 51% of the abundance across samples. More distantly related genes (between 70 and 95% AA ID) comprise 44% of the abundance. Divergent genes (<70% AA ID) comprise <5% of the abundance. Tryptophanase homologs are extremely prevalent, regardless of the %ID to known and putative genes and their abundance levels.

Table S1B, related to Figure 1. Tryptophanase gene abundances across 1,267 human gut metagenomes. Tryptophanase genes were defined at 50% AAID homologs to one of the 36 known and putative tryptophanases found in reference genomes. Average copy number is defined as the average gene copies per cell. Relative abundance is defined as the proportion of total genes.

Table S1C, related to Figure 1. Abundance of tryptophanase homologs that derive from different genera. Tryptophanases identified from the integrated gene catalog were annotated at the genus level based on a megablast nucleotide alignment against NCBI RefSeq with >85% DNA identity. Tryptophanases were also annotated at the phylum level based on a BLASTp protein alignment against NCBI RefSeq database with >50% amino acid identity. Tryptophanases that did not meet these criteria were labeled as "unknown". The column "% of tryptophanase abundance across metagenomes" indicates the percentage of total tryptophanase abundance across all metagenomes comprised by tryptophanases from a particular genus. The column "Average % of tryptophanase abundance per metagenome" indicates the average percentage of total tryptophanase abundance per metagenome comprised by

tryptophanases from a particular genus. **Conclusions:** The most abundant known tryptophanase genes derive from abundant *Bacteroides* species. Many novel tryptophanases likely derive from unknown genera within Bacteroidetes and Firmicutes.

Table S1D, related to Figure 1. Taxa correlated with tryptophanase gene abundance. Linear regressions were performed between tryptophanase gene abundance and MetaPhlAn taxonomic abundances across 1,267 human gut metagenomes. Tryptophanase genes were defined at 50% AAID homologs to one of the 36 known and putative tryptophanases found in reference genomes.

Supplementary Experimental Procedures

All bacterial strains used in this study are listed in the Supplemental Experimental Procedures. All *Bacteroides* strains were grown at 37 °C in brain heart infusion (BHI) agar supplemented with 10% horse blood, tryptone-yeast extract-glucose (TYG) medium, or *Bacteroides* minimal medium (MM) (Holdeman et al., 1977; Martens et al., 2008). *Bifidobacterium longum* NCC2705 was grown in Man-Rogosa-Sharpe (MRS) medium, *Clostridium scindens* ATCC 35704 was grown in reinforced clostridium medium (RCM), *Edwardsiella tarda* ATCC 23685 and *Providencia stuartii* ATCC 25827 were grown in Luria broth (LB), and *Enterococcus faecalis* ATCC 27276 was grown in brain heart infusion (BHI) broth. All aforementioned bacterial strains were cultured under anaerobic conditions at 37°C, using an anaerobic chamber (Coy Laboratory Products) with a 5% H₂, 20% CO₂, and N₂ (balance) gas mix. *Escherichia coli* strains were grown aerobically at 37 °C in LB medium supplemented with ampicillin to select for either the pExchange-tdk plasmid or the pNBU2-*bla-ermG*b plasmid. All in vivo experiments were performed using male mice.

Computational identification of indole-producing genes

We used the approach implemented in MetaQuery (Nayfach et al., 2015) to identify and quantify the abundance of tryptophanases across publicly available human gut metagenomes. Specifically, we used BLAST (Altschul et al., 1990) to identify 237 homologs of 36 known and putative tryptophanases in the integrated catalog of reference genes in the human gut microbiome (IGC), which is a comprehensive set of 9,879,896 genes collected from reference genomes and metagenomic assemblies. The 36 query sequences consisted of both experimentally characterized tryptophanases and putative tryptophanases identified through homology searches (Table S1A). A reference gene was considered a homolog if it aligned to one of the 36 query genes with $\geq 50\%$ amino acid identity (AAID) over $\geq 70\%$ of each gene's length. Next, we obtained the relative abundance of the tryptophanase homologs across 1,267 gut microbiome samples from American, European, and Chinese individuals, which were computed previously and obtained from GigaDB (Li et al., 2014). We also estimated the average copy number of tryptophanase genes per genome, which was obtained by normalizing gene relative abundances by the relative abundance of 30 universal single copy genes. (Wu et al., 2013). We estimated the prevalence of tryptophanases at various abundance and homology thresholds (Table S1A). We annotated the 237 tryptophanase homologs at the genus level based on a megablast nucleotide alignment of these genes against the NCBI RefSeq database (Pruitt et al., 2007). A tryptophanase was annotated according to its top-hit if it aligned with $>85\%$ DNA identity. We annotated the homologs at the phylum level based on a BLASTp protein alignment against the NCBI RefSeq database with $>50\%$ amino acid identity. Tryptophanases that did not meet these criteria were labeled as “unknown” at the genus or phylum levels. We used MetaPhlan (Segata et al., 2012) to estimate the abundance of taxonomic groups and performed linear regression to identify taxa that were predictive of tryptophanase levels.

Construction of *Bacteroides thetaiotaomicron* mutants

All plasmids and primers are listed in the two tables at the end of this section. All mutants were created in the *B. thetaiotaomicron* VPI-5482 Δtdk background. The *Bt* $\Delta 1492$ mutant was constructed using a previously described counterselectable allele exchange method (Koropatkin et al., 2008). Briefly, ~ 1 kb fragments upstream and downstream of the BT1492 gene were cloned and fused using primer pairs BT1492KO SalI-UF/UR and DF/XbaI-DR, respectively, and ligated into the suicide vector pExchange-*tdk*. The resulting vector was electroporated into *Escherichia coli* S17-1 λ *pir* and then conjugated into *B. thetaiotaomicron*. Single-crossover integrants were selected on BHI-blood agar plates containing 200 μ g/ml gentamicin and 25 μ g/mL erythromycin, cultured in TYG medium overnight, and then plated onto BHI-blood agar plates containing 200 μ g/ml 5-fluoro-2-deoxyuridine (FUdR). Candidate BT1492 deletions were screened by PCR using the diagnostic primers listed at the end of this section and confirmed by DNA sequencing to identify isolates that had lost the gene.

Bt wild type + ErmR and *Bt* 1492KO + ErmR mutants were constructed by incorporating an erythromycin resistance marker (ErmR) into *B. theta* VPI-5482 Δtdk and *Bt* 1492KO, respectively, using NBU genomic integration (Wang et al., 2000). Briefly, the pNBU2-*bla-ermG*b plasmid was electroporated into *Escherichia coli* S17-1 λ *pir* and then conjugated into *B. theta* VPI-5482 Δtdk and *B. theta* 1492KO. Single recombinants were selected as described above and cultured in TYG containing 25 μ g/mL erythromycin.

In vitro assays for indole production from tryptophan by *Bacteroides* strains

Bacteroides strains were cultured in 4 mL of TYG medium overnight. *B. theta* wild type + ErmR and *B. theta* 1492KO + ErmR mutants were grown in TYG containing 25 μ g/mL erythromycin. The resultant cultures were either removed from the chamber and tested directly with Ehrlich's reagent (vide infra) or centrifuged, resuspended in 4 mL of minimal medium only or minimal media containing 5 mg/mL Trp, and incubated anaerobically for 6 hours. Cultures in minimal medium were then divided in half and analyzed either by HPLC (10 μ L of clarified supernatant) or with Ehrlich's reagent. *Indole test using Ehrlich's reagent*: To 2 mL of *Bacteroides* culture was added 500 μ L of a 1:1 mixture of *o*- and *p*-xylenes. The tube was shaken several times and allowed to settle. Ehrlich's reagent was added (~ 6 drops). A bright pink color at the organic/aqueous interface indicated the presence of indole (Dowell and Hawkins, 1974; Lombard and Dowell, 1983).

Gnotobiotic mouse experiments

Swiss-Webster germ free (SWGf) mice were originally obtained from Taconic Biosciences (Hudson, NY) and colonies were maintained in gnotobiotic isolators in accordance with A-PLAC, the Stanford IACUC. For the 28-day mono-colonization experiment, 6-10 week old SWGF mice (5 per group) were colonized with either *B. thetaiotaomicron* wild type or the Δ 1492 mutant by oral gavage of overnight cultures as previously described (Marcobal et al., 2011). The mice were maintained on a standard diet (LabDiet 5k67; 0.2% Trp) for two weeks followed by switching to a high protein diet (Teklad TD.90018; 40% protein, 0.5% Trp) for another two weeks. Urine samples were collected on days 0, 5, 9, 14, 19, 23, and 28 post-colonization and frozen at -20 °C prior to analysis. For the 7-day monocolonization experiment, 6-10 week old SWGF mice (5 per group) were colonized as above or maintained germ-free. A fourth group of 6-10 week old conventional SW mice (4 mice) was used as an additional control group. All of the mice were maintained on a standard diet for one week, at which point urine and serum were collected.

For the defined community experiments, 6-10 week old SWGF mice (4 mice per group) were colonized as previously described (Lichtman et al., 2013). Equal volumes of overnight cultures for each organism were combined and mice were inoculated by oral gavage with 200 μ l of the corresponding bacterial mix. Mice were maintained on standard chow (LabDiet 5K67) for four weeks, then were switched to a polysaccharide deficient (PD) diet (BioServ AIN-93G custom diet with 68% glucose) or a diet enriched in fructo-oligosaccharides (FOS) (BioServ AIN-93G custom diet with 10% FOS, 58% glucose) for an additional two weeks. The concentration of tryptophan in the standard, PD, and FOS diets was equivalent at 0.2% w/w. Urine and fecal samples were collected at days 14, 28, 35, and 42 and frozen at -20 °C prior to analysis. IS values were not normalized to creatinine in these experiments (GF and defined community) to avoid the potentially confounding effects of diet on creatinine levels (Lew and Bosch, 1991).

Indoxyl sulfate determination

Indoxyl sulfate levels were measured by LC/MS/MS with isotopic dilution as previously described (Sirich et al., 2013).

Conventional mouse reprogramming experiment

Swiss-Webster restricted flora (RF) mice (6-10 weeks old) were obtained from Taconic Biosciences (Hudson, NY) and maintained on standard chow and allowed to acclimate to the mouse facility for ~one week after receipt. On day zero of the experiment, the mice were given an antibiotic cocktail in their drinking water consisting of kanamycin (0.4 mg/mL), gentamycin (0.035 mg/mL), colistin (0.057 mg/mL), metronidazole (0.215 mg/mL), vancomycin (0.045 mg/mL), and erythromycin (0.01 mg/mL). The antibiotic cocktail was administered for a total of 3 days, then on day 3, the water was replaced with water containing only erythromycin (0.01 mg/mL). Mice were inoculated with 200 μ L of an overnight culture of either wild-type *Bt* or the Δ 1492 mutant by oral gavage beginning on day 3 and continuing until the end of the experiment (day 9). Urine was collected daily (including day zero before antibiotic administration), and stored at -20 °C prior to analysis by LC-MS/MS. This experiment was performed twice with two different groups of Swiss-Webster RF mice obtained from Taconic and the data were combined for analysis.

References

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Dowell, V.R., and Hawkins, T.M. (1974). *Laboratory Methods in Anaerobic Bacteriology*, CDC Laboratory Manual.
- Holdeman, L.V., Cato, E.P., and Moore, W. (1977). *Anaerobe Laboratory Manual* (Blacksburg, VA: Anaerobe Laboratory Virginia Polytechnic Institute and State University).
- Koropatkin, N.M., Martens, E.C., Gordon, J.I., and Smith, T.J. (2008). Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices. *Structure* 16, 1105–1115.
- Lew, S.Q., and Bosch, J.P. (1991) Effect of diet on creatinine clearance and excretion in young and elderly healthy subjects and in patients with renal disease. *J. Am. Soc. Nephrol.* 2, 856-865.
- Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J.R., Prifti, E., Nielsen, T., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* 32, 834–841.
- Lichtman, J.S., Marcobal, A., Sonnenburg, J.L., and Elias, J.E. (2013). Host-centric proteomics of stool: A novel strategy focused on intestinal responses to the gut microbiota. *Mol. Cell. Proteomics* 12, 3310–3318.

- Lombard, G.L., and Dowell, V.R. (1983). Comparison of three reagents for detecting indole production by anaerobic bacteria in microtest systems. *J. Clin. Microbiol.* *18*, 609–613.
- Marcobal, A., Barboza, M., Sonnenburg, E.D., Pudlo, N., Martens, E.C., Desai, P., Lebrilla, C.B., Weimer, B.C., Mills, D.A., German, J.B., et al. (2011). Bacteroides in the infant gut consume milk oligosaccharides via mucus-utilization pathways. *Cell Host & Microbe* *10*, 507–514.
- 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.
- Nayfach, S., Fischbach, M.A., and Pollard, K.S. (2015). MetaQuery: a web server for rapid annotation and quantitative analysis of specific genes in the human gut microbiome. *Bioinformatics* *31*, 3368–3370.
- Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* *35* (Database issue), D61–D65.
- Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* *9*, 811–814.
- Sirich, T.L., Funk, B.A., Plummer, N.S., Hostetter, T.H., and Meyer, T.W. (2013). Prominent accumulation in hemodialysis patients of solutes normally cleared by tubular secretion. *J. Am. Soc. Nephrol.* *25*, 615–622.
- Wang, J., Shoemaker, N.B., Wang, G.R., and Salyers, A.A. (2000). Characterization of a Bacteroides mobilizable transposon, NBU2, which carries a functional lincomycin resistance gene. *J. Bacteriol.* *182*, 3559–3571.
- Wu, D., Jospin, G., and Eisen, J.A. (2013). Systematic identification of gene families for use as “markers” for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. *Plos One* *8*, e77033.

Bacterial strains and plasmids

Strain	Plasmid	Source
<i>Escherichia coli</i> S17-1 λ <i>pir</i>	N/A	Prof. Justin Sonnenburg
<i>Escherichia coli</i> S17-1 λ <i>pir</i>	pExchange-tdk	Prof. Justin Sonnenburg
<i>Escherichia coli</i> S17-1 λ <i>pir</i>	Δ BT1492 in pExchange-tdk	This work
<i>Escherichia coli</i> S17-1 λ <i>pir</i>	pNBU2- <i>bla-ermGb</i>	Prof. Justin Sonnenburg
<i>B. thetaiotaomicron</i> VPI-5482 Δ <i>tdk</i>	N/A	Prof. Justin Sonnenburg
<i>B. thetaiotaomicron</i> VPI-5482 Δ <i>tdk</i> Δ 1492	N/A	This work
<i>B. thetaiotaomicron</i> VPI-5482 Δ <i>tdk</i> + ErmR	N/A	This work
<i>B. thetaiotaomicron</i> VPI-5482 Δ <i>tdk</i> Δ 1492 + ErmR	N/A	This work
<i>B. thetaiotaomicron</i> VPI-5482	N/A	ATCC (American Type Culture Collection)
<i>Bacteroides ovatus</i> ATCC 8483	N/A	ATCC
<i>Bacteroides caccae</i> ATCC 43185	N/A	ATCC
<i>Bacteroides fragilis</i> ATCC 25285	N/A	ATCC
<i>Bacteroides vulgatus</i> ATCC 8482	N/A	ATCC
<i>Bifidobacterium longum</i> NCC2705	N/A	ATCC
<i>Clostridium scindens</i> ATCC 35704	N/A	ATCC
<i>Edwardsiella tarda</i> ATCC 23685	N/A	ATCC
<i>Providencia stuartii</i> ATCC 25827	N/A	ATCC
<i>Enterococcus faecalis</i> ATCC 27276	N/A	ATCC

Primers for construction and sequencing of *Bacteroides thetaiotaomicron* 1492 mutant

Use	Name	Sequence
Plasmid Construction	<u>S</u> allI-UF UR DF <u>X</u> baI-DR	GAA AGA AGA TAA CAT TCG <u>AGT CGA CTT</u> CGC TTT TTG TGG ATG TAG TG CAT TGT CTT CGA TTT TAG ATT TTT ATC GAT AAA AAT CTA AAA TCG AAG ACA ATG TAG CAG GTA GAC AAG ACA AAT G CCA CCG CGG TGG CGG CCG <u>CTC TAG</u> <u>ATG</u> GAC AGC GTG TGA AAA GAG
PCR screening for deletion and sequencing	Diagnostic UF Diagnostic DR	ACG TAG GTT GGA TGG GAG AA GGA AAG CCG CAT TTG ACA GA
Sequencing	Sequencing MidF Sequencing MidR	AAT GGT TTG ATT GGA GCG GT CAT GCA ATG GTT AGT TCG CC