Supplemental Figures

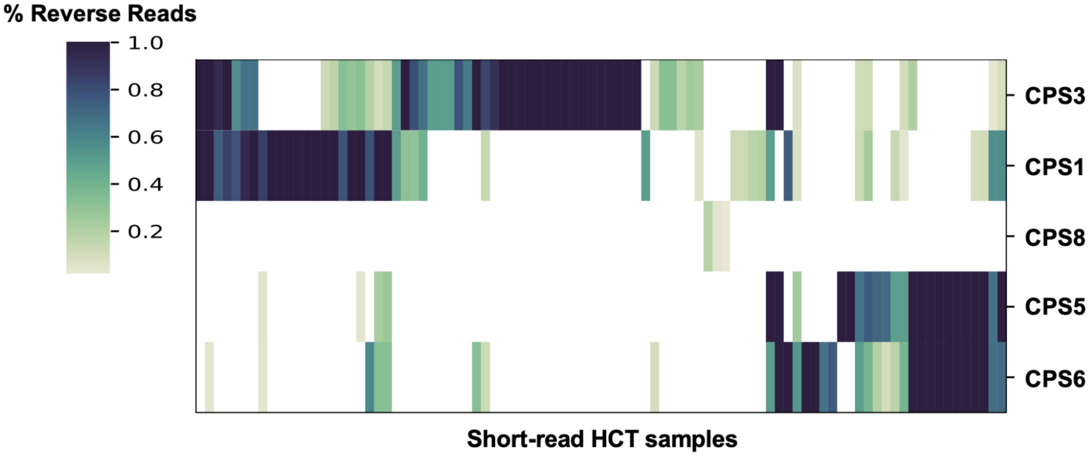


Fig. S1. Inversion proportion of CPS loci invertons in BTh. Inversion proportions of CPS loci invertons in HCT metagenomic samples measured with PhaseFinder. Samples with no inversions in the five CPS invertons were removed.

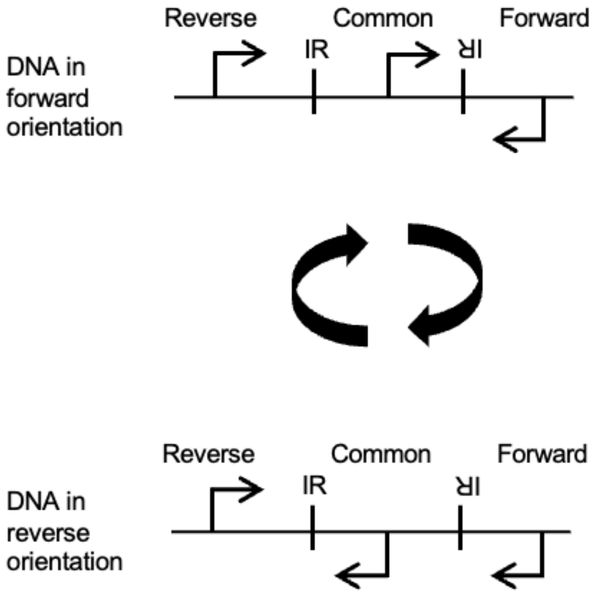


Fig. S2. Inverton confirmation PCR primer design. A Forward and Reverse primer bind to regions of the genome upstream and downstream of the inverton on opposite strands. The Common primer binds the DNA inside of the inverton, between the inverted repeats. When the DNA is in the forward orientation, the Common and Forward primer will generate a PCR product. When the inverton flips, the Common and Reverse primer will generate a PCR product.

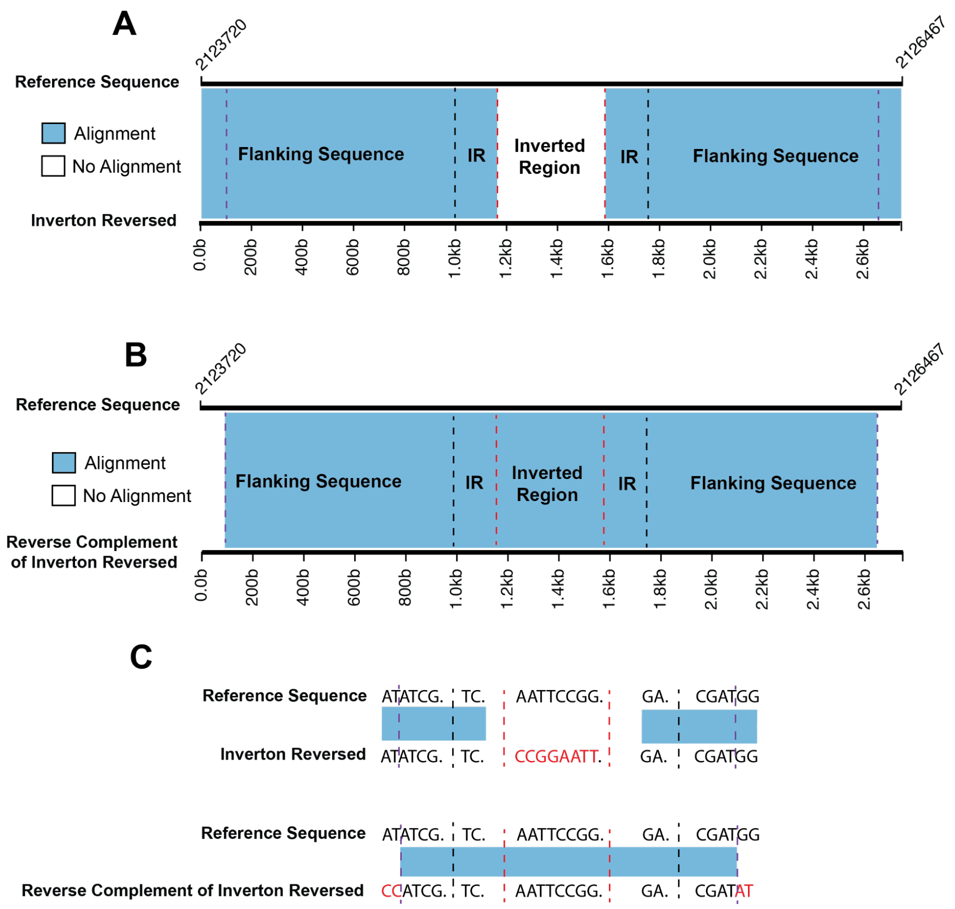


Fig. S3. Very long (>750bp), near perfect, inverted repeats can lead to false positives. (**A**) Alignment of inverton NZ_CP025371.1:2124719-2124870-2125316-2125467, with its invertible sequence inverted, against the *B. pertussis* genome leads to perfect alignment of flanking and IR regions as expected. 'Reference genome' refers to the *B. pertussis* reference genome sequence. 'Inverton reversed' refers to the putative inverton sequence and flanking sequence, with the invertible sequence inverted. Red dashed lines indicate boundaries of the invertible sequence, black dashed lines indicate boundaries of the inverted repeats as detected by einverted, and purple dashed lines indicate the true boundary of inverted repeats. (**B**) Alignment of the reverse complement of the entire inverton NZ_CP025371.1:2124719-2124870-2125316-2125467 with its invertible sequence inverted and flanking sequence, against the *B. pertussis* genome leads to near perfect alignment (6 mismatches) spanning far into the flanking sequence to the true boundary of the inverted repeats, allowing for reads to map regardless of inverton orientation. (**C**) Example with toy nucleotide sequences. Red nucleotides indicate mismatches.

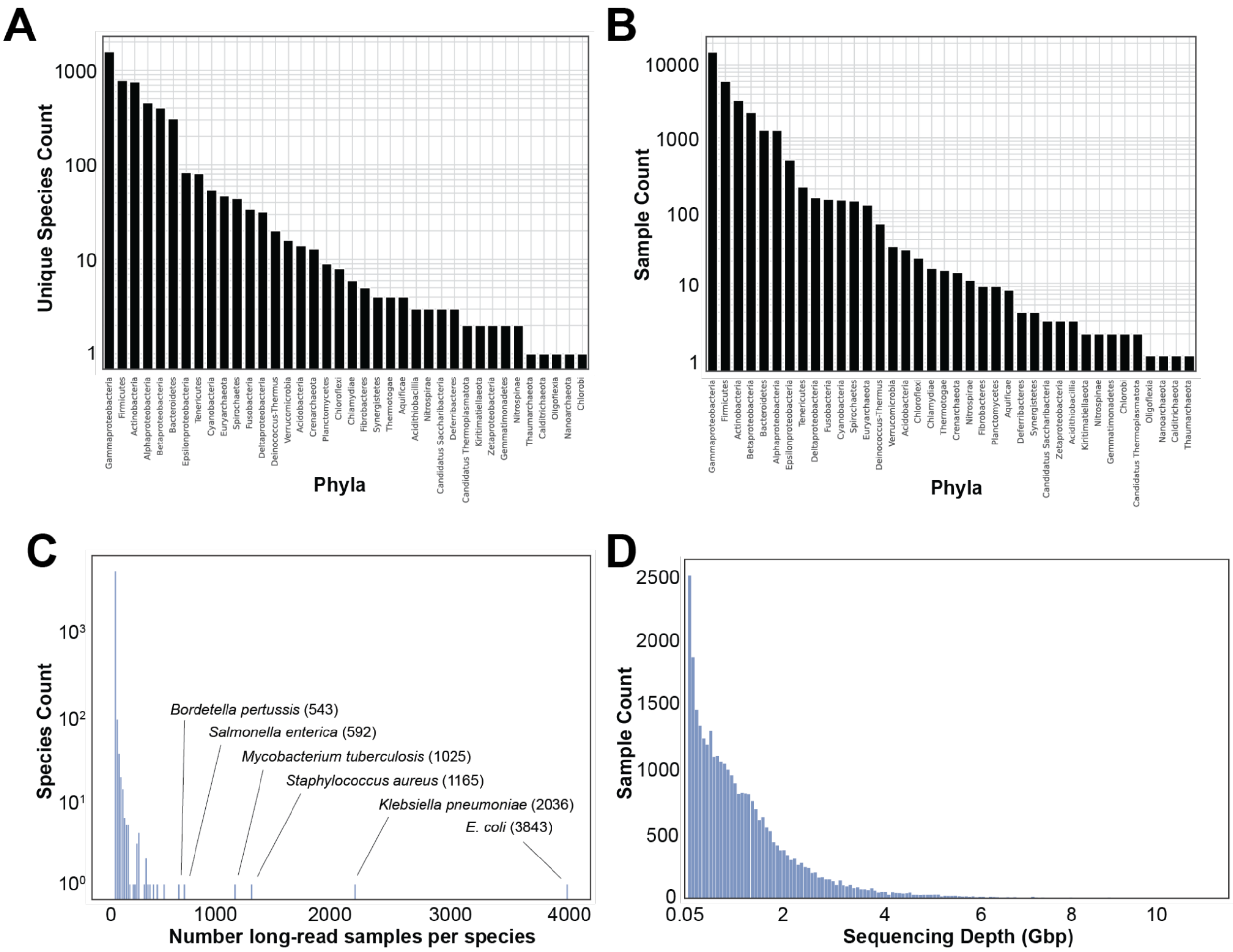


Fig. S4. Overview of SRA long-read isolate sequencing samples analyzed with PhaVa. (**A**) The number of unique species represented in the dataset, grouped by phylum. (**B**) The raw number of sequencing samples, grouped by phylum. (**C**) Histogram of sequencing samples per species. Species with particularly large numbers of samples are labeled. (**D**) A histogram of sequencing depths for all long-read isolate sequencing samples.

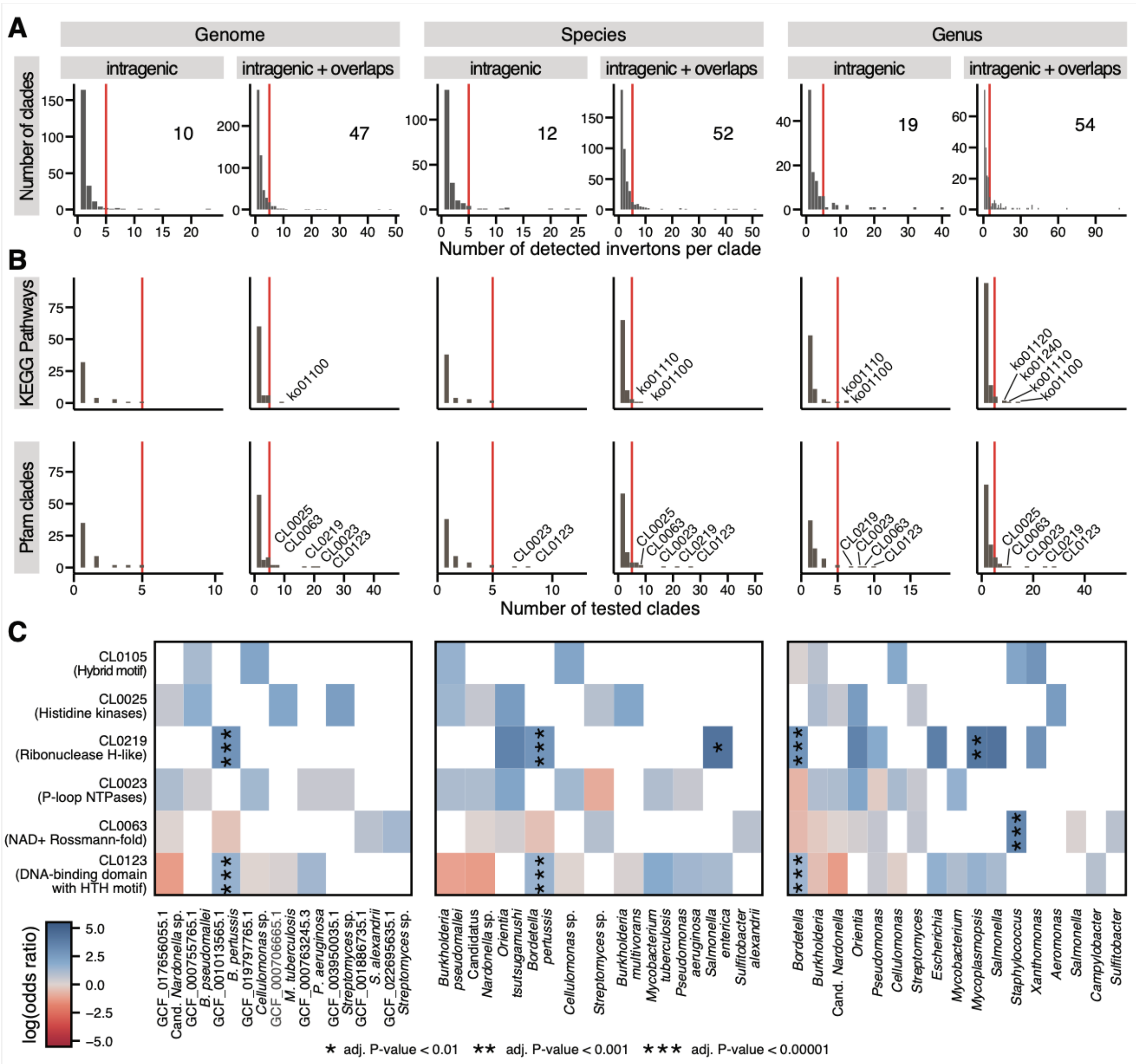


Fig. S5. Intragenic invertons are rare across genomes yet consistently enriched in some Pfam clans. (A) Histograms showing the number of clades (genomes, species, or genera) at various numbers of invertons indicate that invertons are rare, as only one to three invertons can be detected in the majority of clades. Only clades with at least five invertons (red line; number of clades is indicated in the top-right corner of each subplot) were included for the subsequent enrichment analysis. **(B)** KEGG pathways and Pfam clans were tested for enrichment of intragenic (or partial intergenic) invertons in included clades, using a one-sided Fisher's exact test per clade (see Methods). Enrichment was only calculated for sets with at least five invertons associated with genes in the set. Histograms show the number of sets with enrichment score at the number of included clades, showing that most enrichments could be calculated for single

clades only. For example, all KEGG pathways associated with enough intragenic invertons for an enrichment analysis on genome-level were specific for each genome. Sets with enrichment scores across at least five clades (red line) are labeled with their corresponding identifiers. **(C)** Heatmap showing the log-odds ratio (effect size for the enrichment of intragenic invertons) across included clades for the six Pfam clans that have enrichment scores on genus-level (see panel B). Stars indicate significance of the enrichment as calculated by Fisher's exact test and corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure.

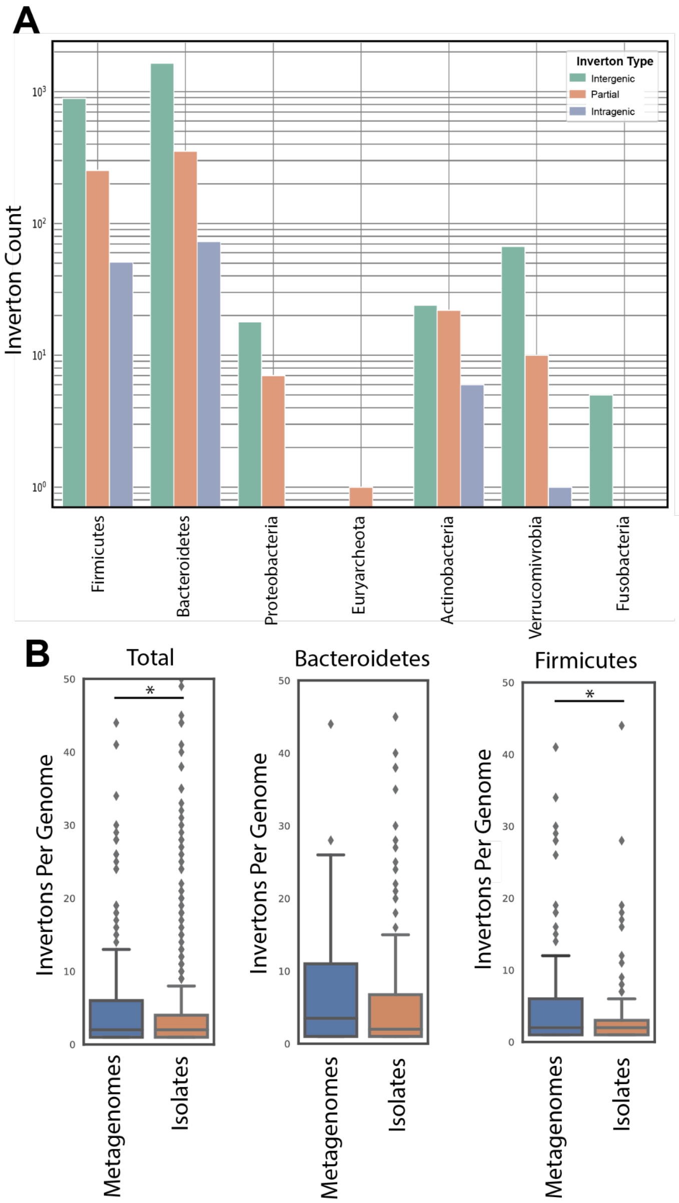


Fig. S6. PhaVa analysis of 210 long-read metagenomes from human stool. (**A**) Counts of invertons identified with PhaVa in 210 stool samples, grouped by phylum and the type of inverton. (**B**) Comparisons of the number of invertons (per genome) found in metagenomic datasets vs. SRA isolate sequencing samples. Total refers to all invertons identified, regardless of taxonomic classification. The distribution of inverton counts per species were found to be significantly different between metagenomes and isolate samples in both the Total and Firmicutes comparisons (p=3.35e-05 and p=0.005 respectively) with a Kolmogorov–Smirnov test. Other individual phyla were not compared due to small species counts with invertons in metagenomic samples.

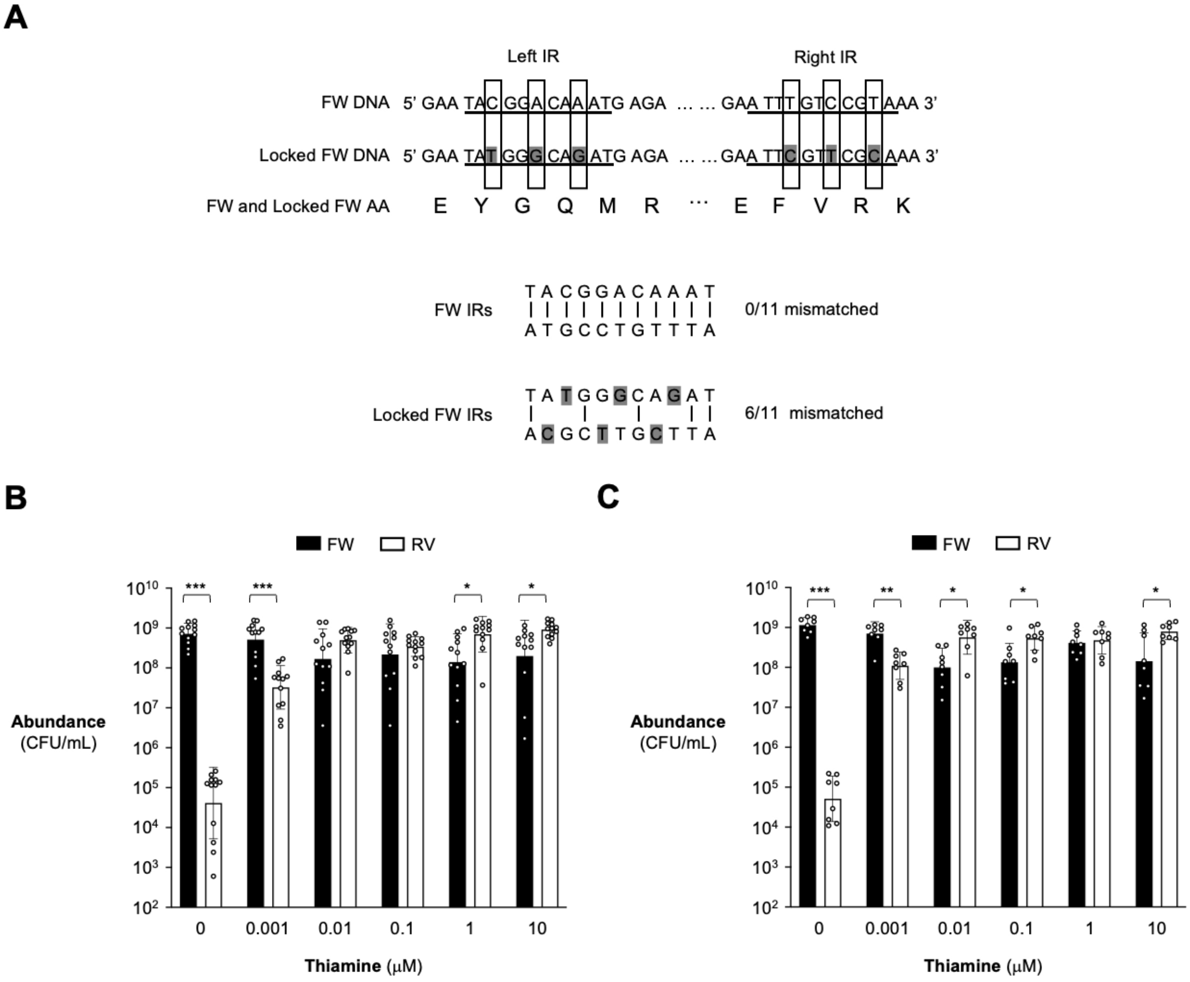
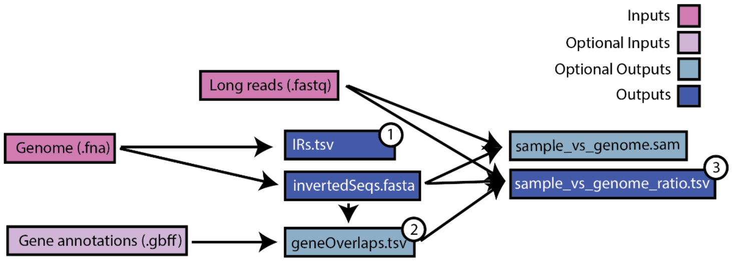


Fig. S7. Locked *thiC* intragenic inverton construction and growth competition. (**A**) Generation of locked intragenic invertons. The forward and locked forward *thiC* IR nucleotide sequences are shown. When possible, the wobble position of each codon corresponding to the IR was mutated to increase mismatches between the two palindromic sequences while maintaining the amino acid sequence. Nucleotides that were mutated are highlighted in gray. (**B-C**) Locked *thiC* strains were competed against each other in thiamine-containing media in a 1:1 ratio. After 40 hours, the abundance of each strain was enumerated using selective agar. Black bars indicate the locked forward strain and white bars indicate the locked reverse strain. Recovered abundances shown here correspond with the competitive index shown in Fig. 4D. In (**B**) the locked forward strain is marked with an erythromycin resistant cassette and the locked reverse strain is marked with a tetracycline resistant cassette. In (**C**) the locked forward strain is marked with a tetracycline resistant cassette and the locked reverse strain is marked with an erythromycin resistant cassette. Geometric mean and geometric standard deviation are shown for replicates conducted across 4-6 independent experiments. For each thiamine concentration a ratio paired t test was performed on the locked forward and locked reverse abundances. ***, p < 0.001; **, p < 0.01; *, p < 0.05.



① IRs.tsv

chromosome	left IR start	left IR stop	right IR start	right IR stop	left IR sequence	invertible sequence	right IR sequence
chr1	1032	1046	1146	1160	ATCG	TACGGATATTACG	CGAT

② geneOverlaps.tsv

Inverton	gene overlaps	Upstream Gene	Upstream Gene Strand	Upstream Gene Distance	Downstream Gene	Downstream Gene Strand	Downstream Gene Distance
inv1	intragenic BT04	BT03	+	100	BT05	-	150

③ sample_vs_genome_ratio.tsv

Inverton	gene overlaps	forward read #	reverse read #	reverse ratio	sample	Upstream Gene	Upstream Gene Strand	Upstream Gene Distance	Downstream Gene	Downstream Gene Strand	Downstream Gene Distance
inv1	intragenic BT04	15	5	0.25	SRR123	BT03	+	100	BT05	-	150

Figure S8: Inputs and outputs of a variation_wf PhaVa run. Output tables of particular interest are labeled and shown below the diagram with example output.

Strain name	**Source**	**Identifier**
Bacteroides thetaiotaomicron VPI-5482 Δ*tdk*	[79]	WT
Bacteroides thetaiotaomicron VPI-5482 Δ*tdk* ΔBT0650	this study	RC131
Bacteroides thetaiotaomicron VPI-5482 Δ*tdk* BT0650 locked RV	this study	RC149
Bacteroides thetaiotaomicron VPI-5482 Δ*tdk* BT0650 locked FW	this study	RC134
Bacteroides thetaiotaomicron VPI-5482 Δ*tdk* BT0650 locked FW *NBU2::NBU2_tet*	this study	RC165
Bacteroides thetaiotaomicron VPI-5482 Δ*tdk* BT0650 locked FW *NBU2::NBU2_erm*	this study	RC 166
Bacteroides thetaiotaomicron VPI-5482 Δ*tdk* BT0650 locked RV *NBU2::NBU2_erm*	this study	RC164
Bacteroides thetaiotaomicron VPI-5482 Δ*tdk* BT0650 locked RV *NBU2::NBU2_tet*	this study	RC163
E. coli S17-1 λ*pir*; *zxx::*RP4 2-(Tetr::Mu) (Kanr::Tn7) λ*pir*	[80]	S17-1 λpir
E. coli DH5α *λpir; F- endA1 hsdR17 (r-m+) supE44 thi-1 recA1 gyrA relA1 Δ(lacZYA-argF)U189 φ80lacZΔM15 λpir*	[81]	DH5α λpir

Table S1. Strains used in this study

Recombinant DNA	**Identifier**	**Source**
pKNOCK-*bla-ermGb*::*tdk*	pExchange	79
pExchange BT0650 KO	pRBC20	this study
pExchange BT0650 locked FW	pRBC21	this study
pExchange BT0650 locked RV	pRBC22	this study
pNBU2_tet	tetR	24
pNBU2_erm	ermR	24

Table S2. Recombinant DNA used in this study