Article

Intragenic DNA inversions expand bacterial coding capacity

https://doi.org/10.1038/s41586-024-07970-4

Received: 5 September 2023

Accepted: 20 August 2024

Published online: 25 September 2024

Check for updates

Rachael B. Chanin[1,5], Patrick T. West[1,5], Jakob Wirbel[1], Matthew O. Gill[2], Gabriella Z. M. Green[1], Ryan M. Park[2], Nora Enright[3], Arjun M. Miklos[1], Angela S. Hickey[2], Erin F. Brooks[1], Krystal K. Lum[4], Ileana M. Cristea[4] & Ami S. Bhatt[1,2]✉

Bacterial populations that originate from a single bacterium are not strictly clonal and often contain subgroups with distinct phenotypes[1]. Bacteria can generate heterogeneity through phase variation—a preprogrammed, reversible mechanism that alters gene expression levels across a population[1]. One well-studied type of phase variation involves enzyme-mediated inversion of specific regions of genomic DNA[2]. Frequently, these DNA inversions flip the orientation of promoters, turning transcription of adjacent coding regions on or off[2]. Through this mechanism, inversion can affect fitness, survival or group dynamics[3,4]. Here, we describe the development of PhaVa, a computational tool that identifies DNA inversions using long-read datasets. We also identify 372 'intragenic invertons', a novel class of DNA inversions found entirely within genes, in genomes of bacterial and archaeal isolates. Intragenic invertons allow a gene to encode two or more versions of a protein by flipping a DNA sequence within the coding region, thereby increasing coding capacity without increasing genome size. We validate ten intragenic invertons in the gut commensal *Bacteroides thetaiotaomicron*, and experimentally characterize an intragenic inverton in the thiamine biosynthesis gene *thiC*.

Adaptation is a cornerstone of survival for any species. In complex microenvironments, bacteria experience many stressors including nutritional and niche competition, oxidative and nitrosative stress, or antibiotics. To overcome these challenges, bacteria may activate specific response programmes that alter transcriptional or translational profiles that promote survival under these conditions. Additionally, bacterial daughter cells may acquire mutations, such as single nucleotide variations or small insertions or deletions, within genes. These gene alterations can then promote survival in the right circumstances.

However, beyond mutations and small insertions and deletions, there are only a few known mechanisms for introducing gene variation in bacteria, including: alternative translational start sites or terminators[5,6]; slipped-strand mispairing[7]; encoding small proteins or microproteins within larger proteins[8,9]; and diversity generating retroelements[10]. Outside of these rare gene-varying events, the typical prokaryotic 'one gene, one gene product' rule generally holds.

Bacteria can reversibly adapt through a fairly prevalent process called phase variation. This preprogrammed mechanism generates phenotypic diversity in a clonal population[1]. One type of phase variation occurs through DNA inversion. Site-specific recombinases recognize a pair of inverted repeats in genomic DNA and invert the intervening DNA sequence[2]. In the first described example, DNA inversion of a promoter sequence resulted in the switching of expression from one flagellar antigen (H1) to another (H2) in *Salmonella enterica* serovar Typhimurium[3,4,11]. This DNA inversion, and the many others that have since been discovered, have critical adaptive roles in both commensal and pathogenic bacteria.

For decades, these invertible loci were identified individually. Since then, computational approaches have enabled higher-throughput discovery of these 'invertons' across the genomes of a small subset of specific bacterial species[12,13]. In 2019, Jiang et al. developed a method that facilitated broad-scale identification of 4,686 intergenic invertons (invertons between genes) in 54,875 bacterial reference genomes[14], using short-read mapping as evidence. Similarly, in early 2023, Milman et al.[15] used a computational model to predict more than 11,000 potential invertons that partially overlap with genes (partial intergenic) in more than 35,000 bacterial species. These types of invertons are sometimes referred to as shufflon systems; in many cases, one portion of the gene product remains constant, whereas the other portion can have several 'choices' that can be shuffled in through DNA inversion[16,17]. Once Milman et al. predicted the candidate invertons, they manually inspected publicly available long-read datasets, which led to the validation of 22 of the more than 11,000 predicted invertons[15].

Here we find that in addition to intergenic and partially intergenic invertons, DNA inversions can occur entirely within a gene. These intragenic invertons expand bacterial coding capacity by either recoding protein sequences within the inverted region or introducing premature stop codons. In both cases, intragenic invertons result in a single gene being able to produce two or more different protein products. To aid in discovery and detection of intragenic invertons, we also developed

[1]Department of Medicine, Division of Hematology, Stanford University, Stanford, CA, USA. [2]Department of Genetics, Stanford University, Stanford, CA, USA. [3]Department of Bioengineering, Stanford University, Stanford, CA, USA. [4]Department of Molecular Biology, Princeton University, Princeton, NJ, USA. [5]These authors contributed equally: Rachael B. Chanin, Patrick T. West. ✉e-mail: asbhatt@stanford.edu

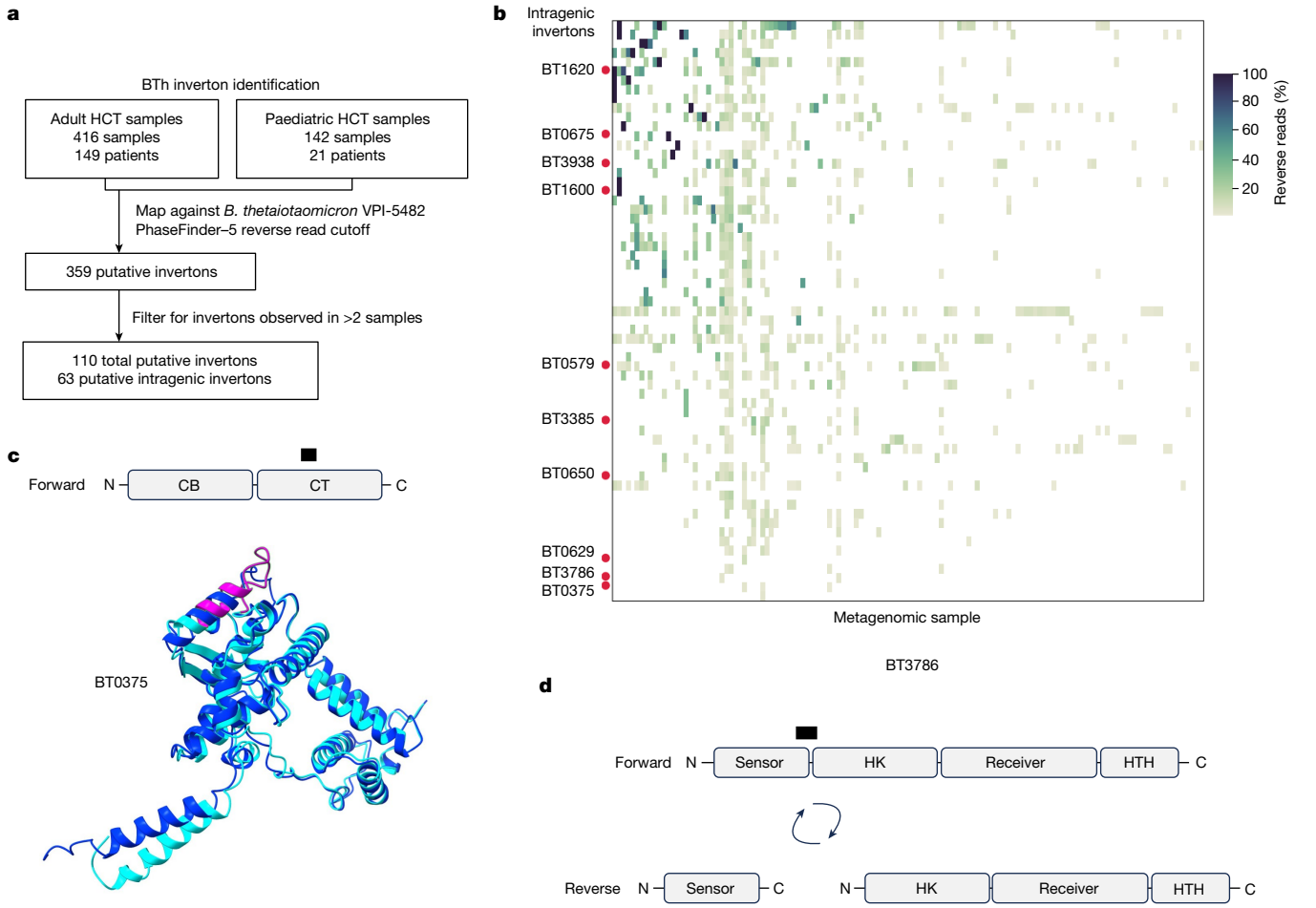


Fig. 1 | Short-read metagenomic datasets reveal intragenic invertons in BTh. **a**, Overview of the analysis pipeline for identifying putative invertons in short-read datasets. **b**, Heat map of the proportion of inversions among intragenic invertons in BTh. Samples with no intragenic invertons were removed. Rows labelled with a gene name represent intragenic invertons with PCR and Sanger sequencing evidence of inversion. **c**,**d**, Protein diagrams for confirmed intragenic invertons in BTh. Grey boxes indicate annotated protein domains. Black bars indicate the region contained within the inverton. **c**, A recoding intragenic inverton in the putative CPS1 invertase BT0375. AlphaFold structures of the forward (dark blue) and reverse (light blue) translation products are shown. Amino acids affected by the flipped inverton are shown in pink. Root mean squared deviation (r.m.s.d.) across all pairs: 2.776 Å; pLDDT forward: 93.2; pLDDT reverse: 88.6. **d**, An intragenic inverton that introduces a premature stop codon in BT3786, a hybrid two-component system. CB, core binding domain; CT, catalytic domain; HCT, haematopoietic cell transplantation; HK, histidine kinase domain; HTH, helix-turn-helix domain.

PhaVa, a software tool that identifies intragenic, intergenic and partial intergenic invertons from long reads. By applying PhaVa to long-read sequencing data for 29,989 bacterial isolates from 4,115 unique species, we find that intragenic invertons occur in many phyla across the prokaryotic tree of life. We focus on *B. thetaiotaomicron*, a model enteric commensal, and validate ten intragenic invertons. We then experimentally characterize the inverton contained within the thiamine biosynthesis protein *thiC*. Finally, we have made the PhaVa software package and all of the identified invertons publicly available.

Intragenic invertons in clinical samples

Most knowledge regarding bacterial genes and their regulation is based on bacteria that are studied in laboratory conditions. Because of this, invertons that provide a fitness advantage in vivo but not in vitro are likely to have been overlooked[18–20]. We hypothesized that there are currently unknown gut-relevant invertons. To test our hypothesis, we chose a metagenomic sequencing dataset from longitudinally collected human stool samples of patients undergoing haematopoietic cell transplantation[21,22]. These samples were selected because bacteria in these patients would encounter many different stressors that might induce or select for inverton flipping, such as chemotherapy, antibiotic treatment, variation in food intake and interactions with the host or other members of the microbial community.

Within these complex samples, we first chose to examine invertons in organisms within the taxon Bacteroidetes, because they are prevalent and abundant in the human gut and they are known to have intergenic invertons[14]. So that we could orthogonally confirm sequencing-based observations in subsequent microbiological and genetic experiments, we focused our analysis on *B. thetaiotaomicron* VPI-5482 (BTh), a genetically tractable strain that is suitable for downstream experimental manipulation. To identify invertons in BTh, we used PhaseFinder[14], a short-read, reference-based inverton detection pipeline (Fig. 1a and Methods). As an internal control to assess whether PhaseFinder could sensitively detect BTh invertons in our metagenomic samples, we examined the capsular polysaccharide (CPS) genes of BTh, a known set of invertible loci. BTh has eight loci that encode different CPS proteins, five of which are controlled by invertible promoters[23–25]. CPS proteins are important mediators of phage susceptibility[26] and can modulate the host immune system[27–29]. Using PhaseFinder on the patient sample datasets, we found read evidence of all five CPS invertons in both the reference and inverted (flipped) orientations (Extended Data Fig. 1a). This demonstrates that PhaseFinder is able to detect known invertons and that these samples

Table 1 | Confirmed intragenic invertons in BTh

Gene	Annotation	Consequence
BT0375	Integrase	Recoding
BT0579	Cys-tRNA(Pro) deacylase	Premature stop codon
BT0629	Mn^{2+} and Fe^{2+} transport protein	Premature stop codon
BT0650	Thiamine biosynthesis protein ThiC	Premature stop codon
BT0675	*N*-acetylglucosamine-6-phosphate deacetylase (NagA)	Recoding
BT1600	BexA, membrane protein	Premature stop codon
BT1620	SusD homologue	Premature stop codon
BT3385	Putative helicase	Premature stop codon
BT3786	Two-component system sensor histidine kinase/response regulator, hybrid (one-component system)	Premature stop codon
BT3938	ATP-dependent DNA helicase RecQ	Premature stop codon

Intragenic invertons confirmed in vitro in BTh. Invertons from short-read datasets were called with PhaseFinder on metagenomic samples. The predicted consequence of inversion is also listed.

have enough *Bacteroides* sequencing depth to identify invertons in these metagenomic samples. Of note, in vitro transcriptional analyses from laboratory conditions show that the majority of invertible CPS loci are not transcriptionally active[30] (their promoters are in the OFF orientation). Finding read evidence of inversion for all invertible CPS loci suggests that the in vivo patient datasets are an ideal environment to detect invertible events that are rare in laboratory-grown bacteria but may be prevalent in bacteria living in more 'natural' ecological settings.

In addition to known intergenic invertons such as those in the CPS loci, we also found read evidence of intragenic invertons in BTh across 132 short-read metagenomic samples (Fig. 1b). We use the term 'intragenic inverton' to describe invertible regions found entirely within single genes. To our knowledge, the only description of invertible DNA sequences entirely within a gene are in isolated cases of very short (7 bp) flips within mitochondrial DNA in certain pathogenic states[31]. These 7-bp mitochondrial DNA flips are postulated to be the consequence of an enzyme-independent event, and thus are different from what we predict here to be an invertase-mediated, preprogrammed inversion. Intragenic invertons were not exclusive to BTh, as we also identified them in *Bacteroides fragilis* (Supplementary Table 1).

To validate the predicted BTh intragenic invertons, we analysed the DNA sequences in these gene regions in vitro. We designed PCR primer sets that enabled us to amplify either the reference or the inverted version (Extended Data Fig. 1b). We tested 59 of the 63 predicted intragenic invertons and confirmed that 10 of them had DNA molecules in both the reference and inverted orientation in a laboratory-grown population of BTh (Fig. 1b and Table 1). The 49 unconfirmed intragenic invertons might be due to the absence of cues or signals in the growth conditions required to flip or select for the inverted orientation and/or false positives from the metagenomic read-based evidence. Of note, we did not detect any differences in taxonomic diversity between metagenomic samples that contained BTh intragenic invertons and those that did not. However, BTh intragenic positive samples, in general, tended to have a higher read coverage of *Bacteroides* (Extended Data Fig. 2).

Among the intragenic invertons that we observed, there were two predicted consequences at the protein level. In some cases, the intragenic inverton resulted in a portion of the protein being 'recoded'. For example, we observed a 57-bp inversion in BT0375, the invertase that is hypothesized to flip the adjacent CPS1 invertible promoter (Fig. 1c). This intragenic inversion changes the amino acid sequence of the flipped region. This change could alter the affinity or specificity of its interaction with the invertible repeats that it targets for flipping or might alter enzyme kinetics[32]. In other cases, the intragenic inverton resulted in the introduction of a 'premature' stop codon, affecting the prediction of protein-coding open reading frames (ORFs). Often, inversion resulted in two predicted ORFs. For example, the inverton in the hybrid two-component system BT3786 occurs between two predicted protein folding domains, and thus might untether the 'sensing' and 'response' elements of this signalling protein (Fig. 1d). Verifying the biochemical consequences of each of these invertons will require further experiments. However, the discovery of intragenic invertons and their consequences suggests that these gene-diversifying processes may have broad effects on bacterial physiology and lifestyle.

PhaVa

Genomic structural variation often involves highly repetitive or low-complexity regions. Because short reads are often not long enough to resolve these types of sequences[33], structural variant callers that use short reads have limited sensitivity. We therefore developed a long-read-based inverton predictor, PhaVa. PhaVa maps long reads against both a forward (identical to reference) and a reverse orientation version of potential invertons (Extended Data Fig. 3a,e). To ensure accurate performance of PhaVa, we optimized read mapping parameters by simulating long-read datasets from 110 bacterial genomes at various sequencing depths using two different read-simulation strategies (Extended Data Figs. 3b,c, 4 and Supplementary Table 2). In general, the false-positive rate was low (0 false positives in 66% of simulated readsets, and a false-positive rate between 0.0001 and 0.005 for the remaining readsets; 'false positive' refers to an identified inversion that was not a simulated true-positive inversion). We observed in our *Bordetella pertussis* simulated readsets that the same false-positive inverton appeared multiple times (Extended Data Fig. 3d). Further investigation revealed that this was due to a putative inverton with inverted repeats longer than 750 bps. In this case, einverted, the computational tool used to detect inverted repeats, only called a small portion of each of these ultra-long invertible repeats, which led to the false-positive call (Extended Data Fig. 5). In summary, our long-read-based inverton predictor PhaVa demonstrates high accuracy in resolving complex genomic structural variations, with only rare instances of false positives.

Invertons exist in many prokaryotes

To find invertons across prokaryotic genomes, we ran PhaVa on 29,989 prokaryotic isolate long-read datasets deposited on the Sequence Read Archive (SRA). We limited our analysis to readsets belonging to bacteria or archaea and with 50 Mb or more of total sequencing, which resulted in our final analysis containing results from 4,115 unique species (Extended Data Fig. 6). The vast majority of these datasets represented bacteria, with only 42 archaeal long-read sequencing datasets. In total, we identified 4,622 unique invertons, 3,468 of which were intergenic (Supplementary Table 3). Of note, compared with Jiang et al.[14], we find invertons at a higher rate per sequencing dataset (0.15 versus 0.07) and per individual genome (1.15 versus 0.09), highlighting the increased sensitivity of long reads for detecting this type of structural variation. Similar to Jiang et al.[14], we found that Bacteroidetes have a relatively large number of intergenic invertons (673; Fig. 2a) and intergenic invertons per genome (2.18; Fig. 2b). Fusobacteria, Gammaproteobacteria and Verrucomicrobia also have high numbers of intergenic invertons per genome (Fig. 2b), with Verrucomicrobia having the highest number per genome overall at 3.13 intergenic invertons per genome. In addition to the intergenic invertons, we also identified 733 partial intergenic invertons (Fig. 2a). Many of these partial intergenic invertons may form shufflon systems, and thus—as expected—these invertons are significantly longer than intergenic or intragenic invertons (Fig. 2c, $P = 7.1 \times 10^{-293}$ and $P = 7.6 \times 10^{-67}$ with a two-tailed

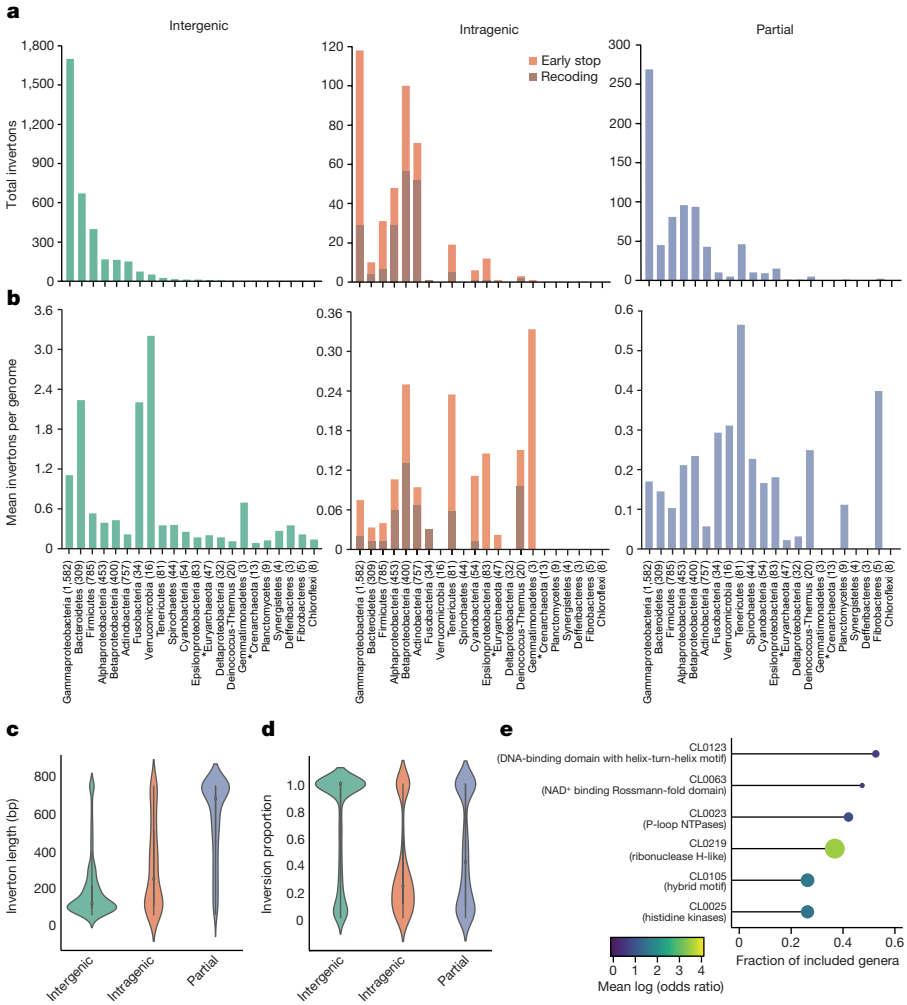


Fig. 2 | PhaVa analysis of long-read sequencing data from isolates reveals that intragenic inversions are prevalent across the bacterial tree of life. **a**, The total number of invertons found within various bacterial phyla from 29,989 publicly available long-read isolate sequencing datasets. Green bars indicate intergenic invertons; orange bars indicate intragenic invertons and blue bars indicate partial intergenic invertons. For intragenic invertons, dark orange indicates recoding invertons and light orange indicates invertons that introduce a premature stop codon. **b**, The mean number of invertons found per genome within a phylum that had at least one inverton. Genomes per phyla are listed in parentheses next to phylum names. A single genome encompasses all long-read datasets that are mapped to that genome. Asterisks denote phyla within archaea. **c**, The distribution of lengths of identified invertons, grouped by inverton type (intergenic, partial and intragenic). Median value is indicated by grey dots. Partial length distribution was found to be significantly different from intergenic ($P < 2.225 \times 10^{-308}$) and intragenic ($P = 1.99 \times 10^{-63}$) with a two-tailed *t*-test. **d**, The distribution of inversion rates of identified invertons, defined as the percentage of reads mapped in the reverse orientation. Median value is indicated by grey dots. **e**, Pfam clan enrichment across several genera. Dot size and fill colour is proportional to the mean log-odds ratio, a measure of effect size for enrichment. The length of the line indicates the fraction of included genera in which an enrichment score for the specific clan could be calculated.

t-test, respectively). This finding of 733 partial intergenic invertons adds to the 22 long-read-validated invertons reported by Milman et al.[17]. Thus, our analysis of prokaryotic isolate long-read datasets from diverse species uncovered both known and novel invertons, shedding light on the remarkable structural variability within prokaryotic genomes and emphasizing the heightened sensitivity of long-read sequencing in this context.

Beyond intergenic and partial intergenic invertons, we also found evidence of intragenic invertons across multiple phyla, including the major gut microbiome-related phyla, Proteobacteria, Firmicutes and Bacteroidetes (Fig. 2a,b). We found the largest number of intragenic invertons, 118, in Gammaproteobacteria, including from organisms such as *Escherichia coli* and *Salmonella*; this is largely owing to the abundance of samples for these organisms in the SRA and our dataset (around 4,000 *E. coli* samples (Extended Data Fig. 6)), given that Gammaproteobacteria have a relatively small number of intragenic invertons detected per genome (0.07; Fig. 2b). Few long-read datasets for Archaea were available, with 36 and 6 for Euryarchaeota and Crenarchaeota, respectively. Despite this, 12 putative archaeal invertons were identified; 10 intergenic, 1 partial intergenic, and 1 intragenic inverton that introduces an early stop in a adenylosuccinate synthase gene in *Salarchaeum sp. JOR-1* (42 total archaeal genomes searched; Fig. 2a,b and Supplementary Table 3). Chromosomal invertons have only been minimally investigated in Archaea. However, our study and a recent computational analysis of phase variable type 1 restriction modification systems by Atack et al.[17] suggest that inverton-mediated phase variation may be an important, yet understudied, regulatory mechanism in this domain. The mean number of intragenic invertons per genome varied greatly between different phyla (Fig. 2a) with Tenericutes, Betaproteobacteria and Actinobacteria having a relatively high number of intragenic invertons detected per genome, at 0.23, 0.25 and 0.1, respectively. The distribution of inversion proportions of individual intragenic invertons was different from that of intergenic invertons (Fig. 2d); intergenic inversions typically appeared to be in

either an ON or OFF state in a given sample—suggesting that all of the organisms within that population shared the same biological state of that inverton. By contrast, intragenic invertons more commonly had inversion proportions somewhere between 0 and 1 (Fig. 2d), indicating the presence of both the forward and reverse orientations within a given 'clonal' sample. Invertons with a 100% or near-100% proportion in the 'reverse' orientation may also represent those that can no longer be flipped, either because of mutations in the invertible repeat or because of loss of the invertase that flips the inverton.

Predicted functional consequences

In total, we identified 372 intragenic invertons, of which 169 are predicted to recode the protein. We find this type of intragenic inverton of particular interest, as recoding invertons would enable a single gene to encode two 'variants' of a protein. We found several notable recoding invertons. First, we identified a recoding intragenic inverton in *slmA*, a nucleoid occlusion factor important for timing cellular division in *Bordetella bronchiseptica* (Extended Data Fig. 7a). The inverton, which affects 128 out of its 191 amino acids, alters the predicted 3D structure of the recoded protein. When queried using the structural alignment software FoldSeek[34], the inverted form of SlmA most closely resembles a Tet repressor. This is likely to result in a change of function of the protein. Next, we identified a recoding intragenic inverton in *barA* from *Aeromonas hydrophila* (Extended Data Fig. 7b). BarA is a two-component system, and although it is not characterized in *A. hydrophila*, these types of proteins are generally important in enabling bacteria to sense and respond to different stimuli, cues or signals in their environment[35]. In *barA*, the intragenic inverton we identified recodes 40 amino acids in the HPt domain, which is responsible for passing signals from the sensing protein (BarA) to its cognate transcription factor. Of note, the inverted form maintains the catalytic amino acids in this domain, and we hypothesize that this could influence either the efficiency with which it transmits its signal or alter the receptor that it interacts with. Finally, we identified recoding inversions in two *hsdS* genes found in *Mycoplasma hominis* (Extended Data Fig. 7c). HsdS is the specificity protein of the type 1 restriction enzyme complex and determines which motifs are targeted for methylation or restriction[36]. For the *hsdS* genes, the inverton recodes a region in the TRD domain which is likely to be important for interacting with DNA. Type 1 restriction enzyme complexes are known for their ability to phase vary[37]. Previously identified mechanisms of phase variation in these genes are via shufflon systems (in which domains are swapped out between homologous genes located in close proximity) and via slipped-strand mispairing, which causes premature stop codons and alters expression of genes. Here we hypothesize that intragenic invertons would expand on this and demonstrate an additional way of generating variation in these genes. Although additional experimental work is needed to fully characterize the effects of these recoding intragenic invertons, they present notable examples of intragenic invertons that may influence key aspects of cellular physiology and behaviour.

We next investigated whether specific gene types or functions were enriched for the presence of intragenic invertons by doing a clade-resolved enrichment analysis. We calculated gene set enrichments per genome, species and genus, combining the genes from all genomes in a specific clade (Fig. 2e and Extended Data Fig. 8). We found six Pfam clans enriched across several genera with the strongest and most consistent enrichments for the Pfam clans CL0123 (helix-turn-helix) and CL0219 (RNase H-like) (Extended Data Fig. 8). This indicates that intragenic invertons occur more frequently than would be expected by chance in genes that have DNA binding, or DNA or RNA modifying activity.

As noted previously, we postulate that inverton orientation likely relates to the environment of a bacterium, and that invertons are more likely to be in the non-reference orientation in organisms that are living in their natural ecological settings. Therefore, we also ran PhaVa on 210 de novo assembled long-read metagenomes from the human gut[38,39], mapping sequencing reads back to their respective metagenomic assemblies (Fig. 3 and Supplementary Table 4). This enabled us to detect invertons that may be absent in isolated bacteria grown in laboratory cultures, but present in vivo. Doing so, we identified more than 3,500 putative invertons, largely from contigs assigned to the phyla Bacteroidetes and Firmicutes (Fig. 3a). In keeping with our model that invertons are more likely to be 'active' in vivo than in vitro, significantly more invertons were identified per species in the metagenomic samples than in the isolate sequencing samples (Fig. 3b) and we find a higher percentage of invertons in the reverse orientation (Fig. 3c). We hypothesize that this is because bacteria grown as isolates in laboratory settings do not experience the wide range of diverse environmental conditions that they do in their natural, polymicrobial habitats. Our analysis of the metagenomic data with PhaVa suggests that bioinformatic analysis of genomes of prokaryotic isolates grown in laboratory conditions probably underestimates the number and range of invertons that exist in microbes. Therefore, the invertons called from the isolate datasets can be thought of as a 'minimal set', as isolate conditions may not be the ideal setting to uncover phase variable regions relative to metagenomic samples or co-cultures.

Intragenic inversion in *thiC*

Both short-read-based and long-read-based analyses of metagenomic datasets revealed that intragenic invertons exist. However, the biological consequences of these invertons are not known. Thus, to evaluate the phenotypic consequences of a particularly prevalent inverton, we focused on an intragenic inverton that introduces a premature stop codon in the BTh BT0650 gene (which encodes the thiamine biosynthesis protein ThiC) (Fig. 4a). Thiamine is an essential cofactor in many cellular biochemical processes and is essential for nearly all organisms. Some organisms, such as humans and certain gut microbes, are fully reliant on dietary, host or other microbial sources for vitamins or their building blocks; others, such as many gut microbes including BTh, have the capacity to biosynthesize thiamine, albeit at a large energetic cost[40,41]. Thus, thiamine availability has been hypothesized to strongly influence microbial community composition[42]. We chose to characterize the intragenic inverton in *thiC*, as this gene has a defined function in thiamine biosynthesis[43,44]. Specifically, the *thiC* gene product, which encodes the enzyme 2-methyl-4-amino-5-hydroxymethylpyrimidine phosphate (HMP-P) kinase, catalyses the conversion of aminoimidazole ribotide (AIR) to 4-amino-5-hydroxymethyl-2-methylpyrimidine (HMP) and forms a key wing in thiamine biosynthesis. In addition to having a defined role, we detected intragenic inversion in both DNA and RNA in our laboratory-grown BTh strain (Fig. 4b). We predicted that the non-reference orientation of the inverton introduces a premature stop codon in the *thiC* mRNA. As noted previously, BTh can grow in the absence of exogenous thiamine as it can synthesize thiamine de novo using a biosynthesis pathway that includes ThiC. We hypothesized that inversion of the invertible locus in *thiC* would interfere with thiamine biosynthesis and would phenocopy a ThiC null mutant. To test the biological consequences of inversion, we generated 'locked' versions of the *thiC* inverton that prevent inversion from occurring within the gene. Traditionally, locking elements in a specific orientation is accomplished by mutating the nucleotides in the inverted repeat regions required for inversion or by deleting the inverted sequences entirely. However, for intragenic invertons, deletion of these sequences or complete mutations would alter the corresponding amino acid sequences and confound interpretation. We therefore exploited the wobble position of the codon to maximize mismatches between the inverted repeats. By mutating these residues, we introduced mismatches in 6 out of 11 positions of the inverted repeat (Extended Data Fig. 9a,b). We created

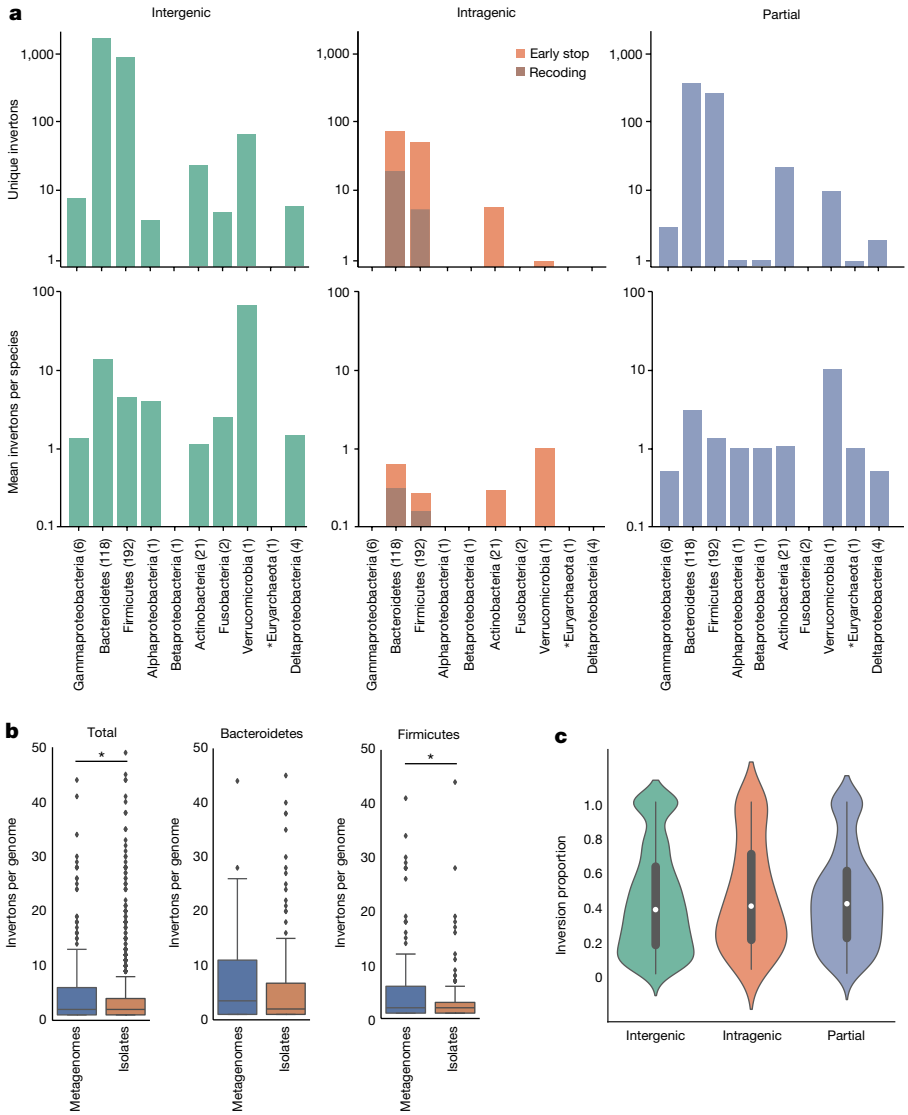


Fig. 3 | PhaVa analysis of 210 long-read metagenomes from human stool. **a**, Counts of invertons identified with PhaVa in 210 stool samples, grouped by phylum and type of inverton. Asterisk denotes a phylum within archaea. For intragenic invertons, dark orange indicates recoding invertons and light orange indicates invertons that introduce a premature stop codon. For the mean number of invertons found per species, a single species encompasses contigs from the 210 stool samples assigned the same taxonomy by Kraken2 (see 'Methods'). The numbers in parentheses next to the phylum names indicate the number of species per phylum with at least one inverton. **b**, Comparisons of the number of invertons found in metagenomic datasets (per genome bin) versus SRA isolate sequencing samples (per genome). Total refers to all invertons identified, regardless of taxonomic classification. The distribution of inverton counts per genome was found to be significantly different between metagenomes and isolate samples in both the total and Firmicutes comparisons ($P = 3.35 \times 10^{-5}$ and $P = 0.005$, respectively) with a Kolmogorov–Smirnov test. Other individual phyla were not compared, owing to small counts of invertons per binned taxon. **c**, The distribution of inversion rates of identified invertons, defined as the percentage of reads mapped in the reverse orientation. Median value is indicated by white dots.

locked-forward (reference orientation) and locked-reverse (flipped intragenic inverton) *thiC* strains. We also generated a *thiC* clean deletion strain.

Next, we grew wild-type BTh, locked-forward, locked-reverse and *thiC*-knockout strains in various concentrations of thiamine (Fig. 4c). The locked-forward strain phenocopied the wild-type strain, as it was able to grow to the same optical density regardless of whether thiamine was added to the media. By contrast, the locked-reverse strain mirrored the *thiC*-knockout strain and was only able to grow to wild-type levels when 0.1 µM or greater thiamine was added to the medium. This finding confirms that the reverse version of the intragenic inverton interferes with ThiC function. Furthermore, we performed quantitative proteomics on the various BTh strains to determine the effect of inversion on the ThiC protein (Extended Data Fig. 10). As the inversion recodes a small portion of ThiC prior to the stop codon, we focused on this unique peptide in our analyses. We found specific evidence of this truncated N-terminal domain of the *thiC* gene product in the reverse BTh strains. Notably, we did not detect this in the forward BTh strains. Although we detect very low levels of *thiC* inversion at the DNA and RNA level in the wild-type BTh laboratory culture, we do not detect the ThiC reverse peptide in these cultures, suggesting that it is expressed below the level of protein detection. As expected, we detect the second 'half' of ThiC (radical SAM domain) in the forward strains. However, we did not identify any of the unique peptides associated with the second domain that begins downstream of the stop codon in any of the *thiC* reverse strains. In addition to querying the specific peptides noted above, we were also able to query many other peptides originating from the forward and reverse *thiC* mutants. This allowed us to detect

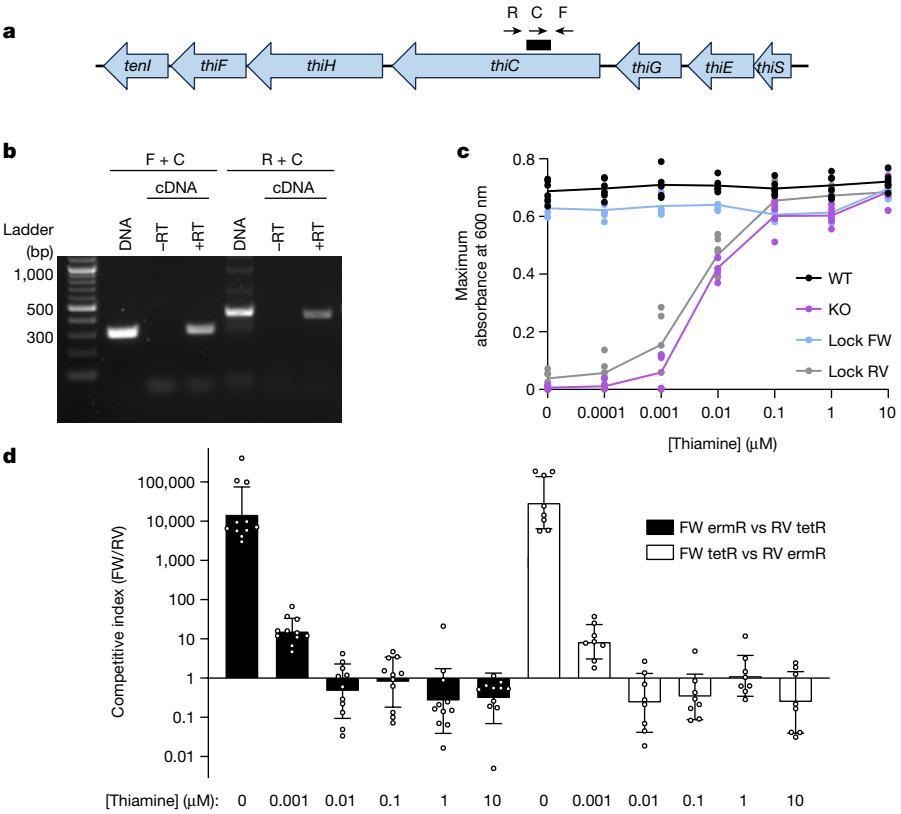


Fig. 4 | Consequences of inversion in thiamine biosynthesis protein. **a**, Schematic showing the location of the *thiC* intragenic inverton (black bar) in the genomic region. Black arrows indicate the binding location and orientation of primers (C, common; F, forward; R, reverse) used in PCR assays to determine inverton orientation. **b**, PCR confirmation of the *thiC* intragenic inverton in genomic DNA and reverse-transcribed RNA (cDNA). Bands were extracted and confirmed with Sanger sequencing. See Supplementary Fig. 1 for the full gel image. **c**, Locked-forward *thiC* (FW), locked-reverse *thiC* (RV), *thiC* clean deletion (KO) and wild-type (WT) BTh strains were grown in VB medium with the indicated concentrations of thiamine. The maximum absorbance reached by each strain was recorded. Each dot represents an individual replicate. The experiment was conducted in biological triplicate and repeated twice. Lines connect means from each concentration. **d**, Locked *thiC* strains were competed against each other in thiamine-containing medium at a 1:1 ratio. After 40 h, the abundance of each strain was determined and the competitive index (the ratio of the recovered locked-forward strain over the locked-reverse strain) was calculated. Black bars indicate that the locked-forward strain is marked with an erythromycin-resistant cassette (ermR) and the locked-reverse strain is marked with a tetracycline-resistant cassette (tetR). White bars indicate that the locked-forward strain is marked with a tetR cassette and the locked-reverse strain is marked with an ermR cassette. Each dot represents an individual replicate. Data are geometric mean ± s.d. Experiments were performed in biological duplicate or triplicate and repeated 4 or 6 times.

an additional peptide that might originate from a new ORF oriented on the antisense strand of inverted *thiC* strains. This ORF falls almost entirely within the inverted region and, except for three amino acids at the end, is identical to the amino acid sequence found within the forward version of the ThiC inverton. Together, in the reverse *thiC* mutants, we find clear evidence of transcription and translation of the predicted truncated ThiC, no evidence of translation of the second domain of the protein and some evidence of translation of a potential novel ORF encoded within the inverton.

Having found that the locked-reverse strain of the *thiC* intragenic inverton phenocopies the null mutant, we explored whether there may be physiological circumstances that favour this mutant over the wild-type or locked-forward strain. A classical approach to assess the relative fitness of two bacterial strains is to perform a competitive growth experiment. Thus, to test whether the inverted form of the *thiC* inverton provides a fitness advantage in different conditions, we competed the locked-forward strain against the locked-reverse strain in an equal proportion in varying concentrations of thiamine. Each strain was chromosomally marked with a different antibiotic resistance cassette. Then we determined the competitive index, which is the ratio of recovered locked-forward bacteria to recovered locked-reverse bacteria (Fig. 4d). To account for any fitness advantages conferred by the antibiotic resistance cassettes, we repeated the competition with the cassette swapped between the two strains. Although results varied slightly between these two complementary versions of the experiment, they were generally concordant. Specifically, we found that as thiamine concentration increases in the media the advantage conferred by the locked-forward version of *thiC* was first diminished and then abolished at concentrations above 0.01 μM. In one version, the locked-reverse strain significantly outcompeted the locked-forward strain at 1 and 10 μM locked-reverse, whereas in the other the version the reverse strain significantly outcompeted the locked-forward strain at 0.01, 0.1 and 10 μM (Extended Data Fig. 9c). Notably, at physiologically relevant thiamine concentrations, in the human intestine[45] (0.02–2 μM), the locked-reverse strain was more fit than the locked-forward strain. This finding complements previous work showing that auxotrophs have a fitness advantage in conditions containing a low level of exogenous metabolites when competing against prototrophic strains[46]. The reversible nature of an enzyme-mediated inversion would allow a subgroup to switch between phenotypes.

On the basis of the fitness advantage of the reverse strain in our competitions, we hypothesized that thiamine availability could regulate flipping of this inverton. To test this hypothesis, we first performed RNA sequencing on wild-type and the *thiC* mutants strains grown in various thiamine-containing conditions (0.001, 0.1 and 10 μM) to determine whether thiamine availability altered expression of any invertases.

We hypothesized that if thiamine was a regulator of the *thiC* intragenic inverton, variations in thiamine concentration would alter expression of one of the 54 BTh invertases. Identification of the invertase responsible for flipping the *thiC* inverton would help characterize its regulation. However, whereas thiamine biosynthesis and uptake genes were transcriptionally regulated by different thiamine concentrations, none of the BTh invertases were (Extended Data Fig. 11 and Supplementary Table 5), suggesting that thiamine may not be a strong positive transcriptional regulator of the invertase responsible for flipping the *thiC* intragenic inverton. Second, we generated 'unlocked' versions of the *thiC* inverton strains (Extended Data Fig. 12) and serially cultured these strains in low (0.001 μM, unlocked reverse) or high thiamine (10 μM, unlocked forward). In contrast to the locked versions, these strains maintain their inverted repeats without any introduced nucleotide substitutions and can therefore be recognized by their cognate invertase. Although we detected flips in both conditions (reverse to forward and forward to reverse) and one of the unlocked-reverse replicates completely flipped to the forward, thiamine concentration did not appear to drive rapid *thiC* flipping. This result, in combination with our RNA-sequencing experiment, supports a model in which thiamine availability does not promote rapid *thiC* inverton flipping and suggests that in these conditions, flipping is probably stochastic and occurs at low frequency.

Discussion

To our knowledge, entirely within-gene DNA inversions have not previously been described in prokaryotes. Although they are not present in every genome, intragenic inversions represent a way in which a single genetic locus can encode multiple genes. Here we developed and benchmarked a long-read inverton-finding pipeline, PhaVa, to more sensitively identify invertons. Using BTh as a model organism, we experimentally validated ten intragenic invertons and characterized the phenotypic effects of an intragenic inverton found in the BTh thiamine biosynthesis gene *thiC*. In addition to known mechanisms of regulation of thiamine acquisition and biosynthesis in BTh[43,47,48], we find that thiamine biosynthesis may also be regulated through inversions. The *thiC* intragenic inverton induces a premature stop codon and we detect this truncated reverse isoform in targeted proteomics. Although the reverse *thiC* strain had impaired growth in thiamine-limited conditions, we found that at physiological concentrations of thiamine found in the human intestinal lumen, organisms encoding a locked-reverse isoform of *thiC* have a competitive growth advantage over the locked-forward isoform. This supports the presence of a novel mechanism of thiamine biosynthesis regulation and suggests a possible ecological explanation for the existence of a 'toggle-able' switch of isoforms.

Although we validated the presence of the intragenic invertons in BTh, we have not experimentally identified the mechanism that flips them. Because these loci resemble classical invertible elements in their size and the characteristics of their inverted repeats, we postulate that flipping is an enzyme-mediated event. We suspect that there is an underlying 'molecular grammar' and that certain invertases recognize and flip specific sequences; specificity of invertases for a given sequence are likely to lie in the inverted repeats, but might also lie within the inverted regions. It is possible that invertases function at a basal level and therefore there is a baseline, low level of inversion that occurs in a small proportion of the population. The data that we present on flipping of the unlocked-forward and reverse *thiC* invertons are consistent with this model. Alternatively, invertases may be expressed in response to specific cues or signals. As BTh encodes 54 annotated invertases, further work is needed to identify which invertase flips particular invertons and under what conditions this occurs. Understanding how these elements are regulated and the consequences of inversion will probably advance our understanding of gene regulation by inversion, and may advance their use in synthetic biology.

In addition to altering protein sequence, intragenic inversions might have transcriptional consequences. For example, the proteomic analysis of ThiC presented here suggests that an additional antisense peptide is found in the reverse strains. More experiments are needed to determine how this peptide arises and what it does, but this finding suggests possible additional consequences of inversion. Furthermore, although we present proteomic data for a single example of an intragenic inverton that introduces a premature stop codon in BTh, we expect that this inverton may not be representative of all intragenic invertons. Future molecular experimentation will be needed to determine the consequences of inversion, particularly in recoding invertons.

Although we find fairly extensive evidence of intragenic invertons using sequencing-based approaches and explore some of them in detail, this work has limitations. First, our analysis of invertons across the prokaryotic tree of life was performed on previously sequenced isolates that were grown in nutrient-replete laboratory conditions; this might not recapitulate physiological conditions in which invertases are active or reverse orientations are favourable. Consequently, we have probably identified a minimal set of invertons in this study and we estimate the full 'invertome' is likely to be substantially larger. The higher frequency of intragenic invertons in long-read metagenome assembled-genomes samples compared to isolates supports the prediction that invertons are more likely to be active in more physiological or mixed microbial conditions. Second, if the invertons that we identified are not representative of the true capacity for inversion, our gene set enrichment analysis may also not adequately identify gene containing invertons. Third, both PhaVa and PhaseFinder are reliant on a reference genome or de novo assembly for read mapping. Detection of invertons is thus restricted to the genomic sequence common between the input sequenced strain and the reference. PhaVa uses relatively strict mapping parameters and if the selected reference is distantly related to the sequenced strain, read mapping quality will decrease and reduce the discovery rate. However, using a de novo assembly instead may result in missing 'fully inverted' invertons relative to reference strains, which may be of interest.

Intragenic invertons represent a novel mechanism for genetic variation and adaptation in bacteria. In this Article, we present a 'roadmap' for in depth investigation of a specific invertible intragenic locus. We expect future niche-specific investigation of inverton-containing organisms to identify additional invertons. Additionally, we anticipate that future studies of intragenic inversion will uncover new layers of bioregulation in prokaryotes and demonstrate many hidden genetic programmes that exist within highly plastic bacterial genomes.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-07970-4.

1. van der Woude, M. W. & Bäumler, A. J. Phase and antigenic variation in bacteria. *Clin. Microbiol. Rev.* **17**, 581–611 (2004).
2. Trzilova, D. & Tamayo, R. Site-specific recombination—how simple DNA inversions produce complex phenotypic heterogeneity in bacterial populations. *Trends Genet.* **37**, 59–72 (2021).
3. Zieg, J., Silverman, M., Hilmen, M. & Simon, M. Recombinational switch for gene expression. *Science* **196**, 170–172 (1977).
4. Stocker, B. A. Measurements of rate of mutation of flagellar antigenic phase in *Salmonella* typhimurium. *J. Hyg.* **47**, 398–413 (1949).
5. Meydan, S., Vázquez-Laslop, N. & Mankin, A. S. Genes within genes in bacterial genomes. *Microbiol. Spectr.* **6**, rwr-0020-2018 (2018).
6. Zhong, A. et al. Toxic antiphage defense proteins inhibited by intragenic antitoxin proteins. *Proc. Natl Acad. Sci. USA* **120**, e2307382120 (2023).
7. Moxon, R., Bayliss, C. & Hood, D. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu. Rev. Genet.* **40**, 307–333 (2006).
8. Sberro, H. et al. Large-scale analyses of human microbiomes reveal thousands of small, novel genes. *Cell* **178**, 1245–1259.e14 (2019).

9. Schlub, T. E. & Holmes, E. C. Properties and abundance of overlapping genes in viruses. *Virus Evol.* **6**, veaa009 (2020).
10. Medhekar, B. & Miller, J. F. Diversity-generating retroelements. *Curr. Opin. Microbiol.* **10**, 388–395 (2007).
11. Andrewes, F. W. Studies in group-agglutination I. The *Salmonella* group and its antigenic structure. *J. Pathol. Bacteriol.* **25**, 505–521 (1922).
12. Goldberg, A., Fridman, O., Ronin, I. & Balaban, N. Q. Systematic identification and quantification of phase variation in commensal and pathogenic *Escherichia coli*. *Genome Med.* **6**, 112 (2014).
13. Sekulovic, O. et al. Genome-wide detection of conservative site-specific recombination in bacteria. *PLoS Genet.* **14**, e1007332 (2018).
14. Jiang, X. et al. Invertible promoters mediate bacterial phase variation, antibiotic resistance, and host adaptation in the gut. *Science* **363**, 181–187 (2019).
15. Milman, O., Yelin, I. & Kishony, R. Systematic identification of gene-altering programmed inversions across the bacterial domain. *Nucleic Acids Res.* **51**, 553–573 (2023).
16. Komano, T. Shufflons: multiple inversion systems and integrons. *Annu. Rev. Genet.* **33**, 171–191 (1999).
17. Atack, J. M., Guo, C., Yang, L., Zhou, Y. & Jennings, M. P. DNA sequence repeats identify numerous type I restriction-modification systems that are potential epigenetic regulators controlling phase-variable regulons; phasevarions. *FASEB J.* **34**, 1038–1051 (2020).
18. Chatzidaki-Livanis, M., Coyne, M. J., Roche-Hakansson, H. & Comstock, L. E. Expression of a uniquely regulated extracellular polysaccharide confers a large-capsule phenotype to *Bacteroides fragilis*. *J. Bacteriol.* **190**, 1020–1026 (2008).
19. Taketani, M., Donia, M. S., Jacobson, A. N., Lambris, J. D. & Fischbach, M. A. A phase-variable surface layer from the gut symbiont *Bacteroides thetaiotaomicron*. *mBio* **6**, e01339-15 (2015).
20. Troy, E. B., Carey, V. J., Kasper, D. L. & Comstock, L. E. Orientations of the *Bacteroides fragilis* capsular polysaccharide biosynthesis locus promoters during symbiosis and infection. *J. Bacteriol.* **192**, 5832–5836 (2010).
21. Severyn, C. J. et al. Microbiota dynamics in a randomized trial of gut decontamination during allogeneic hematopoietic cell transplantation. *JCI Insight* **7**, e154344 (2022).
22. Siranosian, B. A. et al. Rare transmission of commensal and pathogenic bacteria in the gut microbiome of hospitalized adults. *Nat. Commun.* **13**, 586 (2022).
23. Martens, E. C., Chiang, H. C. & Gordon, J. I. Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host Microbe* **4**, 447–457 (2008).
24. Martens, E. C., Roth, R., Heuser, J. E. & Gordon, J. I. Coordinate regulation of glycan degradation and polysaccharide capsule biosynthesis by a prominent human gut symbiont. *J. Biol. Chem.* **284**, 18445–18457 (2009).
25. Krinos, C. M. et al. Extensive surface diversity of a commensal microorganism by multiple DNA inversions. *Nature* **414**, 555–558 (2001).
26. Porter, N. T. et al. Phase-variable capsular polysaccharides and lipoproteins modify bacteriophage susceptibility in *Bacteroides thetaiotaomicron*. *Nat. Microbiol.* **5**, 1170–1181 (2020).
27. Round, J. L. et al. The Toll-like receptor 2 pathway establishes colonization by a commensal of the human microbiota. *Science* **332**, 974–977 (2011).
28. Neff, C. P. et al. Diverse intestinal bacteria contain putative zwitterionic capsular polysaccharides with anti-inflammatory properties. *Cell Host Microbe* **20**, 535–547 (2016).
29. Mazmanian, S. K., Liu, C. H., Tzianabos, A. O. & Kasper, D. L. An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell* **122**, 107–118 (2005).
30. Porter, N. T., Canales, P., Peterson, D. A. & Martens, E. C. A Subset of polysaccharide capsules in the human symbiont *Bacteroides thetaiotaomicron* promote increased competitive fitness in the mouse gut. *Cell Host Microbe* **22**, 494–506.e8 (2017).
31. Musumeci, O. et al. Intragenic inversion of mtDNA: a new type of pathogenic mutation in a patient with mitochondrial myopathy. *Am. J. Hum. Genet.* **66**, 1900–1904 (2000).
32. Smyshlyaev, G., Bateman, A. & Barabas, O. Sequence analysis of tyrosine recombinases allows annotation of mobile genetic elements in prokaryotic genomes. *Mol. Syst. Biol.* **17**, e9880 (2021).
33. West, P. T., Chanin, R. B. & Bhatt, A. S. From genome structure to function: insights into structural variation in microbiology. *Curr. Opin. Microbiol.* **69**, 102192 (2022).
34. van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
35. Casino, P., Rubio, V. & Marina, A. The mechanism of signal transduction by two-component systems. *Curr. Opin. Struct. Biol.* **20**, 763–771 (2010).
36. Loenen, W. A. M., Dryden, D. T. F., Raleigh, E. A. & Wilson, G. G. Type I restriction enzymes and their relatives. *Nucleic Acids Res.* **42**, 20–44 (2014).
37. De Ste Croix, M. et al. Phase-variable methylation and epigenetic regulation by type I restriction-modification systems. *FEMS Microbiol. Rev.* **41**, S3–S15 (2017).
38. Chen, L. et al. Short- and long-read metagenomics expand individualized structural variations in gut microbiomes. *Nat. Commun.* **13**, 3175 (2022).
39. Maghini, D. G. et al. Quantifying bias introduced by sample collection in relative and absolute microbiome measurements. *Nat. Biotechnol.* **42**, 328–338 (2024).
40. Rodionov, D. A. et al. Micronutrient requirements and sharing capabilities of the human gut microbiome. *Front. Microbiol.* **10**, 1316 (2019).
41. Sharma, V. et al. B-vitamin sharing promotes stability of gut microbial communities. *Front. Microbiol.* **10**, 1485 (2019).
42. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
43. Costliow, Z. A. & Degnan, P. H. Thiamine acquisition strategies impact metabolism and competition in the gut microbe *Bacteroides thetaiotaomicron*. *mSystems* **2**, e00116–17 (2017).
44. Martinez-Gomez, N. C. & Downs, D. M. ThiC is an [Fe-S] cluster protein that requires AdoMet to generate the 4-amino-5-hydroxymethyl-2-methylpyrimidine moiety in thiamin synthesis. *Biochemistry* **47**, 9054–9056 (2008).
45. Said, H. M. Intestinal absorption of water-soluble vitamins in health and disease. *Biochem. J.* **437**, 357–372 (2011).
46. D'Souza, G. et al. Less is more: selective advantages can explain the prevalent loss of biosynthetic genes in bacteria. *Evolution* **68**, 2559–2570 (2014).
47. Jurgenson, C. T., Ealick, S. E. & Begley, T. P. Biosynthesis of thiamin pyrophosphate. *EcoSal Plus* https://doi.org/10.1128/ecosalplus.3.6.3.7 (2009).
48. Rodionov, D. A., Vitreschak, A. G., Mironov, A. A. & Gelfand, M. S. Comparative genomics of thiamin biosynthesis in prokaryotes. *J. Biol. Chem.* **277**, 48949–48959 (2002).

Methods

Strains and media

The bacterial strains used in this study are listed in Supplementary Table 6. *E. coli* strains were routinely grown in LB Miller medium (Fisher) at 37 °C. When necessary, carbenicillin was added at 100 μg ml^{-1}. BTh was grown anaerobically (90% nitrogen, 5% carbon dioxide, 5% hydrogen) in an anaerobic chamber (Sheldon Manufacturing) in hemin (5 μg ml^{-1}) and L-cysteine (1 mg ml^{-1}) supplemented Brain Heart Infusion (Sigma) media (BHIS) or Varel-Bryant broth (VB)[49]. When necessary, the antibiotics tetracycline (2.5 μg ml^{-1}), erythromycin (25 μg ml^{-1}), or gentamicin (200 μg ml^{-1}) were added to the media. Thiamine HCl (Sigma) was added at the specified concentrations. All media used to grow BTh was preincubated in the anaerobic chamber overnight.

Construction of *thiC* mutant strains

The *thiC* strains were generated via allelic exchange, as previously described[50]. For Δ*thiC*, 600–700-bp flanking regions of the coding region were amplified using Q5 High Fidelity Polymerase (New England Biolabs). For locked strains, plasmid overhangs, flanking regions, locked repeats and intervening forward or inverted sequences were synthesized (Twist Biosciences) (Supplementary Table 6). For unlocked strains, the forward and reverse region was synthesized (Twist Biosciences) and the flanking regions were amplified via PCR. Regions were then assembled into pExchange-*tdk* using the HiFi DNA Assembly Kit (New England Biolab). Plasmid inserts were verified by long-read sequencing from Plasmidsaurus. Sequence confirmed plasmids were propagated in *E. coli* DH5α λ*pir*. *E. coli* S17-1 λ*pir* was used as a donor strain for conjugation into BTh Δ*tdk*. Exconjugants with chromosomal integration of plasmids were recovered on BHIS plates containing gentamicin and erythromycin. Second crossover events were selected by using BHIS FudR (5-fluoro-2-deoxy-uridine; 200 μg ml^{-1}). Deletion, locking, and inverton status was confirmed by PCR. To generate differentially resistant *thiC*-locked strains, the suicide vectors pNBU2_tet and pNBU2_erm were used. Single crossover events were selected by plating on gentamicin plates containing either erythromycin or tetracycline, respectively. Recombinant DNA used in this study is listed in Supplementary Table 6.

Validating inversion in DNA

Intragenic inverton confirmation primers were designed using NCBI Primer Blast under default settings with the addition of a GC clamp. PCR product size was targeted to be between 300 and 600 bp. The common and reverse primer were oriented on the same strand of the reference genome and the forward primer was located on the complementary strand. The common primer was located in between the two inverted repeats (Extended Data Fig. 1b; primers listed in Supplementary Table 6). Four of the predicted invertons were not experimentally tested, as target-specific primers could not be generated within the above constraints.

DNA was isolated from wild-type BTh cultures grown for 18 h in either BHIS or VB media. Cultures were started with a single colony. DNA was isolated using a chemical and enzymatic lysis. Glass beads (0.1 mm) were added to bacterial pellets along with 700 μl of extraction buffer (50 mM Tris pH 7.5, 1 mM EDTA, 100 mM NaCl, 1% (w/v) SDS) and 25 μl of Proteinase K (10 mg ml^{-1}). Pellets were vortexed for 20 s and incubated at 55 °C for 60 min. Seven-hundred microlitres of phenol:chloroform:isoamyl alcohol (25:24:1 by volume) was added to the mixture prior to incubating at room temperature for 5 min. Phases were separated by centrifugation at 10,000 rpm for 5 min. The aqueous upper layer was collected and transferred to a new tube. 5 μl of RNAse A (10 mg ml^{-1}) was added and incubated at 37 °C for 15 min. An equal volume of phenol:chloroform:isoamyl alcohol was added, mixed, and incubated at room temperature for 5 min. Phases were separated as above and the aqueous phase was added to a new tube containing an equal volume of chloroform: isoamyl alcohol (24:1 by volume). Tubes were mixed and incubated at room temperature for 5 min prior to phase separation via centrifugation. The aqueous phase was added to a new tube along with 45 μl of 3 M sodium acetate and 1 ml cold 100% ethanol. DNA was precipitated overnight at −20 °C. Pellets were washed twice with 1 ml of cold 70% ethanol. Dried pellets were resuspended in water.

PCR reactions were performed using Q5 high fidelity polymerase (68 °C annealing temp, 10 s annealing time, and 30 s extension time). PCR reactions were run on an 1.2% agarose gel. If multiple bands were visible, bands of the expected size were gel purified using the QIAquick Gel Extraction kit (Qiagen). If a single band of the expected size was observed, the PCR reaction was purified using the Monarch PCR Cleanup Kit (New England Biolabs). DNA was sent for Sanger sequencing (Elim Bio). Sequencing was compared to the in silico prediction.

Validating *thiC* inversion in RNA

RNA was isolated from wild-type BTh cultures grown for 18 h in BHIS media. 5 ml cultures were quenched using 500 μl phenol/ethanol solution (90% (v/v) ethanol and 10% (v/v) saturated phenol pH 4-5). Pellets were spun down and stored at −80 °C until extraction. Pellets were lysed in 250 μl PBS and 10 μl of lysozyme (10 mg ml^{-1}) at 37 °C for 30 min. 30 μl 20% SDS was added prior to an additional 30 min incubation. Trizol (1.5 ml) was added to the mixture and incubated at room temperature for 10 min. Chloroform (0.5 ml) was added to each sample and inverted vigorously for 15 s. The aqueous phase was taken from centrifuged samples and an equal volume of 100% ethanol was added. RNA was purified using the Zymo RNA clean kit. DNA was removed using Ambion Turbo DNAse. cDNA was made using Taqman Reverse Transcription reagents (Invitrogen) according to the manual. A no-reverse-transcriptase control was performed to ensure that all DNA was removed. PCR was performed to determine orientation of inverton as above. Correctly sized bands were sent for Sanger sequencing.

BTh growth in thiamine concentrations

BTh wild-type, *thiC* locked-forward, *thiC* locked-reverse and *thiC*-knockout strains were grown overnight in BHIS media. Aliquots of each were then washed twice in preincubated PBS containing cysteine (1 mg ml^{-1}). Strains were inoculated at an OD600 of 0.05 in VB media containing the indicated concentration of Thiamine in a 96-well flat bottom plate. Readings were taken in a Stratus plate reader (Cerillo) every ten minutes. Non-inoculated VB media from each time point was used as a blank. The maximum OD600 value achieved per well was determined.

Competitive growth assay

Antibiotic-marked BT0650 locked strains were grown overnight in BHIS with appropriate antibiotics. Strains were washed twice with preincubated PBS containing cysteine (1 mg ml^{-1}). A glass dilution tube containing 3 ml of VB with indicated concentrations of thiamine was inoculated with 1×10^3 colony-forming units (CFU) ml^{-1} of each strain. After 40 h of growth at 37 °C in the anaerobic chamber, CFU ml^{-1} was determined by plating on selective agar. A competitive index was calculated by dividing the recovered CFU ml^{-1} of the locked-forward strain by the CFU ml^{-1} of the locked-reverse strain and then correcting by the inoculum.

RNA-sequencing data generation and analysis

BTh wild-type, *thiC* locked-forward, *thiC* locked-reverse and *thiC*-knockout strains were grown overnight in BHIS media. Aliquots of each were then washed twice in PBS containing cysteine (1 mg ml^{-1}) preincubated in the anaerobic chamber. Strains were then inoculated into 5 ml of VB media containing various concentrations of thiamine (0.001 μM, 0.01 μM, and 10 μM) in a 1:100 dilution. Two biological replicates were performed for each condition. Cultures were grown at 37 °C until approximately mid-exponential phase before 4 ml of each

were quenched with an ice-cold 10% acidic phenol solution, vortexed vigorously, and placed on ice. After being removed from the anaerobic chamber, cell pellets were collected and frozen at −80 °C. For the *thiC* deletion and locked RV strains grown in 0.001 µM thiamine, cultures did not reach approximately equivalent density to the rest of the samples, but still yielded sufficient and high enough quality RNA for sequencing.

RNA was isolated from frozen pellets using a customized version of the Zymo Quick-RNA Fungal/Bacterial Miniprep Kit (Zymo Research) protocol. Specifically, pellets were first resuspended in a pre-lysis buffer consisting of 0.5 mg ml^{-1} lysozyme and 0.1 U $µl^{-1}$ SUPERase-In (Thermo) in PBS and vortexed on a benchtop vortexer for 10 min at room temperature. Samples were then mixed with 600 µl RNA Lysis Buffer in ZR BashingBead tubes and vortexed using a benchtop vortexer at maximum speed. Manufacturer's instructions were followed for the remainder of the isolation, except for two changes: performing on-column DNase treatment with NEB DNase I, and eluting total RNA in DEPC-treated nuclease-free water. Samples were quantified with the Qubit RNA HS assay and qualified with both Nanodrop and the Agilent Bioanalyzer RNA 6000 Nano assay. Total RNAs were further cleaned with the Zymo RNA Clean-&-Concentrator-5 (Zymo Research) kit to improve A260/230 ratios to above 1.8. All samples had RNA integrity numbers (RINs) ≥5.8, with most RINs ≥7.0. All samples were subjected to rRNA depletion with Illumina Ribo-Zero Plus before stranded library prep and 150-bp paired-end sequencing on a NovaSeq X (Novogene). Sequencing reads were processed with a customized version of a prokaryotic RNA-seq Nextflow script (https://github.com/adamd3/BactSeq). In brief, we used TrimGalore to trim adapters and remove low-quality bases. We aligned to the BTh VPI-5482 reference genome with bwa-mem[51], and assigned 'counted' to features in the reference genome using featureCounts[52], requiring at least two nucleotides of overlap to be counted. The raw counts were processed with default DEseq2[53] parameters, including Benjamini–Hochberg *P* value adjustment for multiple hypothesis testing, to assess differential expression of BTh genes across both thiamine concentrations and *thiC* genetic backgrounds. Raw counts were also normalized to reads per kilobase per million mapped reads (RPKM) for visualization purposes, with all counts receiving an artificial +0.1 pseudo-count to remove zeros.

Long-term evolution experiment

Unlocked *thiC* strains were grown overnight in BHIS. Cultures were then diluted 1:100 into 5 ml VB containing either 0.001 µM (unlocked reverse) or 10 µM (unlocked forward) thiamine. Once cultures reached saturation (every 24–48 h), they were subcultured into fresh medium. After every transfer, 1–2 ml of culture was frozen. DNA was extracted from pellets using the DNeasy Blood and Tissue kit (Qiagen) according to the manual for extraction from Gram-negative bacteria. Quantitative PCR was run using the Luna probe qPCR master mix according to the manual. Twenty nanograms of DNA was used as the template per reaction. qPCR was run on a QuantStudio5 (Applied Biosystems) and analysed using Design & Analysis Software 2.6.0 (Thermo). Plasmids containing the forward and reverse sequences of the *thiC* intragenic inverton were used to generate standards for copy number determinations.

Identifying invertons in BTh and *B. fragilis* with PhaseFinder

Two short-read datasets were used for identifying invertons in BTh and *B. fragilis*, 416 samples from 149 adult haematopoietic cell transplantation patients[22] (BioProject PRJNA707487) and 142 samples from 21 paediatric haematopoietic cell transplantation patients[21] (BioProject PRJNA787952). Each individual short-read dataset was analysed with PhaseFinder[14] with the BTh VPI-5482 and *B. fragilis* FDAARGOS_1225 reference genomes and default parameters to identify putative invertons. Invertons were included in further analysis if they had at least five reads mapping to the reverse orientation of the inverton, and had reads mapping to the reverse orientation in at least three different samples. Inverton–gene overlaps and partial overlaps were found using a custom script, now incorporated in PhaVa in the 'Create' step, and the gene annotations from the VPI-5482 and FDAARGOS_1225 genbank files (.gbff).

The PhaVa algorithm

Inverted repeats are identified with einverted, part of the EMBOSS suite[54]. For each putative inverton, two sequences are then created: one where the sequence between identified inverted repeat pairs is inverted (reverse) and one where it is not (forward), along with flanking sequence on either side, similar to PhaseFinder. Long reads are mapped against the created sequence with minimap2[55] and must pass several filters to be included as evidence of inversion. (1) Reads must have a MAPQ score of ≥2 to eliminate multimapping reads. (2) Reads must span the entire length of the inverton and at least 30 bp into the flanking sequence on either side. (3) The mismatch rate along the length of a read must be below a maximum mismatch rate. The mismatch rate is considered separately over the length of an inverton and over flanking sequence, to avoid reads that map well to only one region or the other. An adjustable mismatch rate is used instead of a flat mismatch cutoff to account for both the variable length of long reads and the high sequencing error rate of current long-read sequencing technologies relative to short-read sequencing. After mapping, reads mapped to the inverted and non-inverted sequences are tallied and two optional post-mapping filters are applied before reporting detected invertons: a minimum number of total reads mapped to the 'reverse' sequence, and a minimum proportion of total mapped reads mapped to the 'reverse' sequence.

PhaVa was also tested on DNA isolated from the locked-forward and locked-reverse BTh strains grown BHIS. DNA was extracted using the DNeasy Blood and Tissue Kit (Qiagen) according to the manual for extraction from Gram-negative bacteria. Long-read sequencing was performed using the Oxford Nanopore platform (10.4.1) by Plasmidsaurus.

Simulating long-read datasets for optimizing PhaVa

Benchmarking was performed using two different long-read-simulation strategies: custom-trained models with NanoSim[56] and a pre-trained model with pbsim2[57]. For the NanoSim based approach, ten bacterial species were selected, in part based on the relevance in the human microbiome. For each species, a new long-read-simulation model was generated. To generate the model, a reference genome and reference long-read dataset were obtained from NCBI. Long reads were mapped against their respective reference genome with minimap2 and the mappings were used as input for the 'characterization stage' of NanoSim[56] in genome mode. The resulting NanoSim models were used to generate simulated long-read datasets in the 'simulation stage' in genome mode (Supplementary Table 2). Reads were generated from the unmodified reference genome, and so no evidence of inversion for any inverton is expected, and any invertons identified by PhaVa would be false positives. For each species, five coverage levels and six replicates of each coverage level were generated totaling 300 simulated long-read datasets. We chose to generate 6 replicates so that we could differentiate between stochastic and recurrent false positives. NanoSim does not allow for simulation of reads to a specified coverage, so read count was instead used to obtain similar coverage levels across species. *Streptomyces eurocidicus* and *Enterococcus faecalis* simulated readsets were generated at relatively deeper coverages due to poor read mapping characteristics from the selected reference long-read datasets that were used to generate the read-simulation model resulting in a smaller proportion of reads passing PhaVa read mapping filters. Simulated readsets were then analysed with PhaVa and used to estimate false-positive rates and optimize post-mapping filters.

For the pbsim2[57]-based approach, we tested four different aspects of PhaVa. First, we simulated reads across a broad panel of bacterial genomes (Supplementary Table 2). 100 genomes were randomly

selected from the 4,115 isolate reference genomes with the GNU 'shuf' command. For each genome, true positives were simulated by inverting 300 invertons prior to read simulation with PhaVa's '–mockGenome' option. Readsets were then simulated with pbsim2 at 100x coverage for each genome, with three replicates per genome. False-positive rate was measured as the number of false positives divided by the number of true negatives and false positives. To measure the effect of coverage on the false negative rate, readsets were simulated with pbsim2 for both *E. coli* and BTh at 10x, 50x, 100x, 500x and 1,000x coverage levels, with 3 replicates at each coverage. To measure the effect of read length on the false-positive rate, readsets were simulated for *E. coli* with pbsim2, varying the option '--length-mean' by 100-bp increments from 100–1,000 bp. To measure the effect of taxonomic distance on PhaVa output, 1,000 complete *E. coli* genomes were clustered with the dRep[58] compare workflow, using the '--S_algorithm fastANI' option. Eight genomes with varying ANI to a single reference were then used to simulate reads with pbsim2 at 100x coverage, with 3 replicates for each genome.

Identifying putative invertons from public long-read sequencing data with PhaVa

Candidate isolate long-read sequencing datasets were identified on NCBI with the following search criteria: "(Bacteria[Organism] OR Archaea[Organism]) AND ("pacbio smrt"[Platform] OR "oxford nanopore"[Platform]) AND genomic[Source]". Datasets were further filtered by removing datasets with the "amplicon" flag, and removing datasets with less than 50 Mb of sequencing in total (Supplementary Table 3). Individual read datasets were downloaded with fastq-dump, a part of the SRA Toolkit (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software). Nanostat[59] was run on each remaining readset to measure dataset characteristics. For each unique taxid represented in the resulting readsets, a reference genome and paired genbank file (.gbff) were selected by identifying a genome with the highest level of completion for that species, and the least number of contigs. In the case of reference genomes with equal quality based on these parameters, the first identified was selected. Long-read datasets were then analysed with PhaVa (v0.1.0) with default parameters. Gene overlaps and partial gene overlaps were identified by comparing coordinates of genbank file annotations with inverton coordinates, a function available for use in PhaVa.

Recoding intragenic inverton structural prediction

Structural predictions of the amino acid sequences for the forward and reverse orientations of the recoding intragenic invertons were generated using ColabFold[60]. The highest-confidence structures were directly compared using the Matchmaker function in the UCSF ChimeraX tool[61]. Domain structure changes were identified using the InterProScan tool[62] to annotate protein domains. Forward and reverse structures for SlmA were analysed using Foldseek[34] to identify closest structural analogues. See Supplementary Table 6 for sequences of recoding intragenic invertons.

Gene set enrichment analysis

In order to assess which functional groups were enriched for genes harbouring intragenic invertons, we performed a clade-resolved gene set enrichment analysis. We first annotated genes with KEGG KOs using the kofamKOALA tool[63] and with Pfam domains by running HMMER3[64] with the Pfam domain database. KEGG pathways and modules were filtered for those that were present in bacterial genomes and Pfam clan definitions were downloaded from the Interpro website[65]. We then calculated enrichments per genome, per species, and per genus (for those combining the genes from all genomes in a specific clade). At each level, we filtered out groups with fewer than 5 intragenic invertons, resulting in 10 genomes, 12 species, and 19 genera being included for downstream analysis. We also considered genes with both intragenic or partial intergenic invertons, resulting in 47 genomes, 52 species, and 54 genera being tested. In each group, we tested if the genes annotated with this pathway were enriched for those carrying invertons by using a one-sided Fisher test. Enrichment analysis was only performed for pathways containing genes with at least one inverton. Multiple testing correction was performed with the Benjamini–Hochberg procedure[66].

Identifying putative invertons from long-read metagenomes

Two-hundred hybrid short-read and long-read human stool metagenomic datasets were accessed from BioProject PRJNA820119[38]. Each hybrid dataset was assembled using SPAdes[67] with the '-meta' flag and long reads provided with the '--nanopore' option. An additional ten nanopore long-read human stool microbiome metagenomic datasets from BioProject PRJNA940499[39] were assembled with Flye[68]. Assembled contigs are deposited at https://doi.org/10.5281/zenodo.7662825. Gene annotations for assemblies were obtained with Prodigal[69] using the '-meta' flag. Contig taxonomic classifications were obtained with Kraken2 using a database constructed using the Genbank reference[70]. Each long-read dataset was then analysed with PhaVa (v0.1.0) with default parameters, using its respective de novo assembly as its reference assembly. Resulting inverton calls were dereplicated by clustering the inverton with 1000 bp flanking sequence upstream and downstream at 99% average nucleotide identity with CD-HIT[71].

Sample processing for mass spectrometry analysis

Cultured bacteria were pelleted and lysed in 5% SDS, 50 mM triethylammonium bicarbonate (TEAB), pH 8.5, and 1x HALT Protease and Phosphatase Inhibitor (Thermo Fisher Scientific). Unlocked-reverse and unlocked-forward samples were taken from transfer 12 of the long-term thiamine exposure experiment. The locked-reverse sample was grown in low-thiamine-containing conditions (0.001 μM) and the locked-forward and wild-type samples were grown in BHIS. Samples were sonicated in a cup horn sonicator with 20 bursts of 1 s pulses and cell debris was pelleted at 5,000 rpm for 5 min. Protein concentration was determined by BCA assay (Pierce). Ten micrograms of protein per sample were reduced and alkylated with the incorporation of tris(2-carboxyethyl)phosphine and 2-chloroacetamide at final concentrations of 25 mM and 50 mM, respectively, followed by incubation at 70 °C for 20 min. Aqueous phosphoric acid was added to samples at a final concentration of 2.5% and loaded onto S-Trap micro spin columns (Protifi) with binding buffer (100 mM TEAB, pH 7.55, and 90% methanol). Following the manufacturer's protocol, S-Trap columns were used for desalting, digestion, and peptide extraction. In brief, samples were washed on the S-Trap columns for a total of 10 rounds. Trypsin digestion was performed at a 1:25 enzyme:protein ratio (w/w) with incubation at 37 °C overnight. Peptides were eluted, dried down in a SpeedVac, and resuspended in 0.1% formic acid, 4% acetonitrile, and 0.015% *n*-dodecyl β-D-maltoside at a final concentration of 0.15 μg μl^{-1} using UHPLC-MS water. A total of 0.15 μg (1 μl) was loaded on column for liquid chromatography–mass spectrometry (LC–MS/MS) analysis.

LC–MS/MS analysis

Peptides were analysed by LC–MS/MS using a nanoElute 2 coupled to a timsTOF Ultra mass spectrometer (Bruker Daltonics). Peptides were separated using a one column method with a PepSep ULTRA C18 HPLC column (25 cm × 75 μm × 1.5 μm; Bruker) heated at 50 °C, and a 10-μm emitter (Bruker) attached to a CaptiveSpray Ultra source. For peptide separation, a linear gradient was run consisting of 3 to 34% buffer B over 30 min at a flow rate of 200 nl min^{-1}. The mobile phases were 0.1% formic acid in UHPLC-MS water (buffer A) and 0.1% formic acid/99.9% UHPLC-MS acetonitrile (buffer B).

A test mix solution of each sample at an equivalent ratio was run in data-dependent acquisition (DDA) parallel accumulation serial fragmentation (PASEF) mode to optimize a data-independent acquisition-PASEF (dia-PASEF) method. The dia-PASEF method was

used for subsequent analysis of each sample independently. First, DDA-PASEF mode was run with five PASEF ramps and an MS1 scan range of 100 to 1700 m/z with positive ion polarity. TIMS settings included an ion mobility $1/K_0$ range starting at 0.64 and ending at 1.45 V s cm^{-2}, ramp time of 50 ms, and 100% duty cycle. Singly charged precursors were excluded based on their position in the PASEF m/z–ion mobility plane. Precursor intensities over a threshold of 500 arbitrary units were selected for MS/MS fragmentation while those over a target intensity of 20,000 arbitrary units were actively excluded for 0.4 min. Second, for analysis in dia-PASEF mode, the mass spectrometry scan range was set to 100 to 1,700 m/z with positive ion polarity. TIMS settings included an ion mobility $1/K_0$ range from 0.65 to 1.46 V s cm^{-2}, ramp time of 50 ms, and 100% duty cycle. Using a spectral library generated from the test mix sample, the py_diAID tool[72] was applied to optimize the dia-PASEF method, which included 16 dia-PASEF scans separated into three ion mobility windows per scan (Supplementary Table 7).

Mass spectrometry data analysis

To generate the spectral library from the DDA-PASEF test mix run for py_diAID assessment of the precursor ion density distribution in the m/z–ion mobility plane, raw files (.d) were analysed in FragPipe (v21.1)[73] using the default workflow. MSFragger database searching was performed against a Uniprot Reference proteome that included *B. thetaiotaomicron* VPI-5482 (FASTA downloaded February 2024) customized by appending the three determined ThiC intragenic inverted ORFs. In the database search, variable modifications of methionine oxidation and protein N-terminal acetylation were specified. DIA data was analysed in DIA-NN[74] using an in silico spectral library generated (Bruker) with the custom BTh FASTA. Precursor ion generation included prediction of deep learning-based spectra, retention times, and ion mobilities, in silico digestion with trypsin, a maximum of one missed cleavage, N-terminal methionine excision, fixed cysteine carbamidomethylation, precursor charge states of 1 to 4, and a fragment ion m/z range of 200 to 1,800. From the DIA runs, a spectral library was generated and the resulting DIA identifications at the peptide and protein levels were normalized by MS1 intensity. Identified ThiC peptides were validated for quantitative comparisons based on reliable detection in at least one sample, as assessed in Skyline[75,76], and MS/MS spectra and extracted ion chromatograms from resulting peptides were visualized.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Short-read adult stool sequencing data was previously published and is available under NCBI BioProject ID PRJNA707487. Short-read paediatric stool sequencing data were previously published and are available under NCBI BioProject ID PRJNA787952. Long-read metagenomic sequencing data were previously published and are available under BioProject PRJNA820119 and BioProject PRJNA940499. Assembled metagenomic contigs are available at https://doi.org/10.5281/zenodo.7662825. A list of accession numbers for long-read isolate sequencing data are available in Supplementary Table 3. Mass spectrometry raw files (.d) generated in this study have been deposited to the ProteomeXchange Consortium through the PRIDE partner repository[77] (project accession PXD054577). Long-read sequencing data for the locked *thiC* intragenic inverton strains and RNA-sequencing data are available under NCBI BioProject ID PRJNA1118344. Accession codes for long-read datasets are listed in Supplementary Table 3. The reference genome for *B. thetaiotaomicron* VPI-5482 is the NCBI reference sequence AE015928.1. The reference genome for *B. fragilis* FDAARGOS_1225 is the NCBI reference sequence NZ_CP069563.1. Source data are provided with this paper.

Code availability

PhaVa is available at https://github.com/patrickwest/PhaVa. Long-read datasets were analysed with PhaVa (v0.1.0) with default parameters.

49. Bacic, M. K. & Smith, C. J. Laboratory maintenance and cultivation of bacteroides species. *Curr. Protoc. Microbiol.* https://doi.org/10.1002/9780471729259.mc13c01s9 (2008).
50. Zhu, W. et al. Xenosiderophore utilization promotes *Bacteroides thetaiotaomicron* resilience during colitis. *Cell Host Microbe* **27**, 376–388.e8 (2020).
51. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
52. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
53. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
54. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
55. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
56. Yang, C., Chu, J., Warren, R. L. & Birol, I. NanoSim: nanopore sequence read simulator based on statistical characterization. *Gigascience* **6**, gix010 (2017).
57. Ono, Y., Asai, K. & Hamada, M. PBSIM2: a simulator for long-read sequencers with a novel generative model of quality scores. *Bioinformatics* **37**, 589–595 (2021).
58. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
59. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
60. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
61. Meng, E. C. et al. UCSF ChimeraX: tools for structure building and analysis. *Protein Sci.* **32**, e4792 (2023).
62. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
63. Aramaki, T. et al. KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
64. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
65. Paysan-Lafosse, T. et al. InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427 (2023).
66. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
67. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes de novo assembler. *Curr. Protoc. Bioinformatics* **70**, e102 (2020).
68. Lin, Y. et al. Assembly of long error-prone reads using de Bruijn graphs. *Proc. Natl Acad. Sci. USA* **113**, E8396–E8405 (2016).
69. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
70. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
71. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
72. Skowronek, P. et al. Rapid and in-depth coverage of the (phospho-)proteome with deep libraries and optimal window design for dia-PASEF. *Mol. Cell. Proteomics* **21**, 100279 (2022).
73. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).
74. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41–44 (2020).
75. MacLean, B. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).
76. Pino, L. K. et al. The Skyline ecosystem: informatics for quantitative mass spectrometry proteomics. *Mass Spectrom. Rev.* **39**, 229–244 (2020).
77. Perez-Riverol, Y. et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **50**, D543–D552 (2022).

Acknowledgements The authors thank D. Schmidtke, A. Natarajan, J. D. Shanahan, D. Maghini, M. Dvorak, A. Han, M. Chakraborty, X. Jin, E. Martens and Bhatt laboratory members for helpful conversations and scientific advice regarding this project; W. Zhu for plasmids (pNBU2_tet and pNBU2_erm); S. Winter for the DH5α strain; and D. Haft and F. Thibaud-Nissen for helpful discussion about accessing SRA long-read datasets. Funding was provided as follows: National Institutes of Health R01 AI148623 (A.S.B.), National Institutes of Health R01 AI143757 (A.S.B.), Stand Up 2 Cancer Foundation (A.S.B. and I.M.C.), National Institutes of Health R01 AI174515 (I.M.C.), National Institutes of Health T32 training Grant HG000044 (R.B.C.), The AP Giannini Foundation (R.B.C.), National Institutes of Health T32 training Grant HL120824 (P.T.W.), National Science Foundation Graduate Research Fellowship (M.O.G., A.S.H. and N.E.), Stanford DARE fellowship (N.E.), and National Institutes of Health TL1 training Award TL1TR003019 (K.K.L.). Computing costs were supported, in part, by an NIH S10 Shared Instrumentation grant (1S10OD02014101).

Author contributions R.B.C., P.T.W. and A.S.B. conceived and designed the study. Data acquisition and processing was performed by R.B.C., P.T.W., J.W., R.M.P., G.Z.M.G., A.S.H., M.O.G., E.F.B., A.M.M., N.E. and K.K.L. Data were visualized by R.B.C., P.T.W., J.W., R.M.P., K.K.L.

and M.O.G. Data interpretation was done by R.B.C., P.T.W., J.W., M.O.G., K.K.L. and A.S.B. Funding acquisition was done by A.S.B. and I.M.C. Writing of the original draft was performed by R.B.C., P.T.W. and A.S.B. All authors contributed to the review and editing of this manuscript.

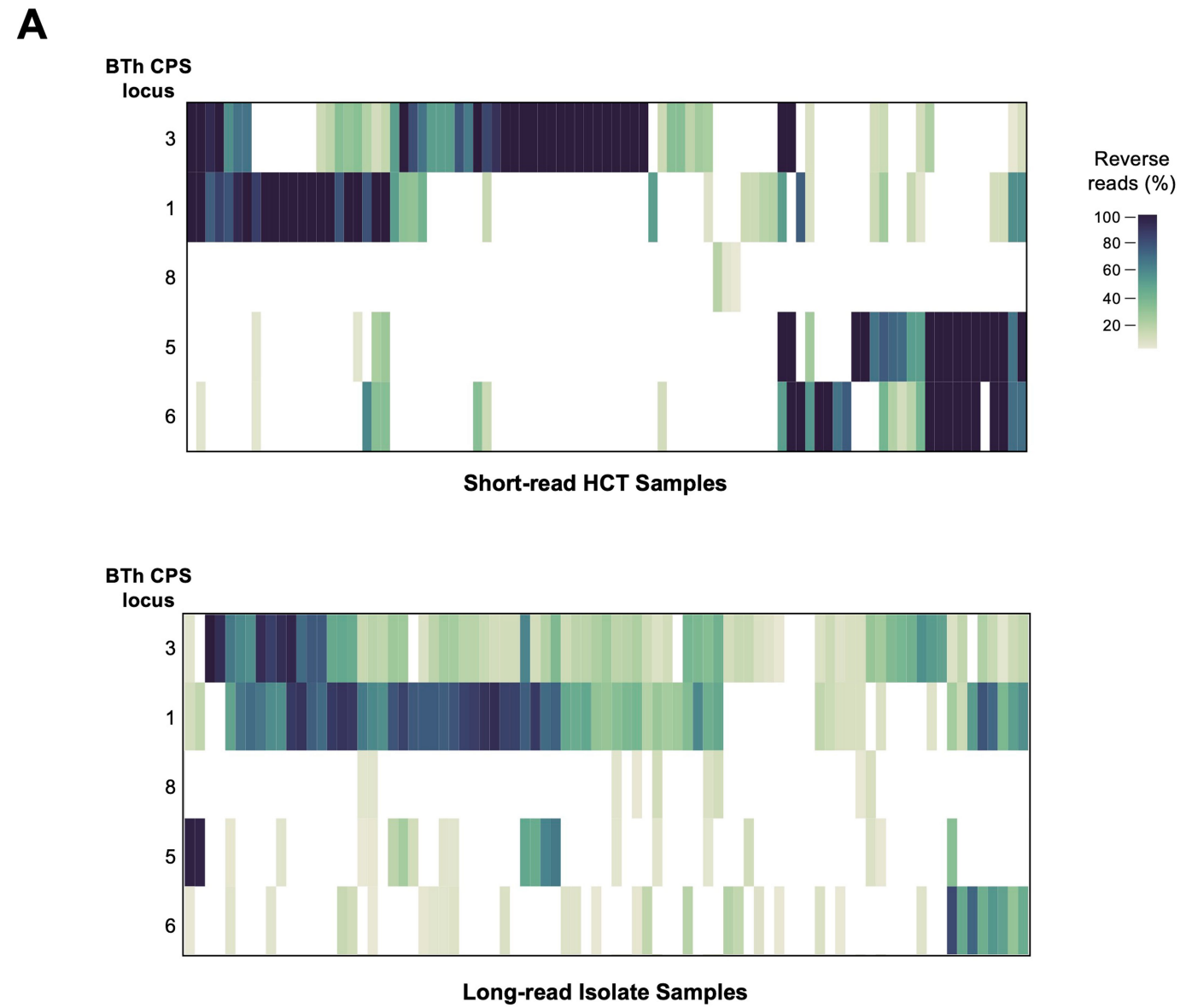
Competing interests P.T.W. is a contract bioinformatician at Oxford Nanopore Technologies; this position started during the review process. The other authors declare no competing interests.

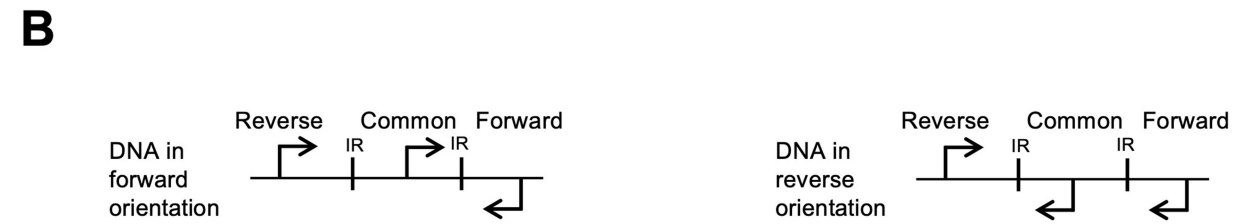
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-07970-4.

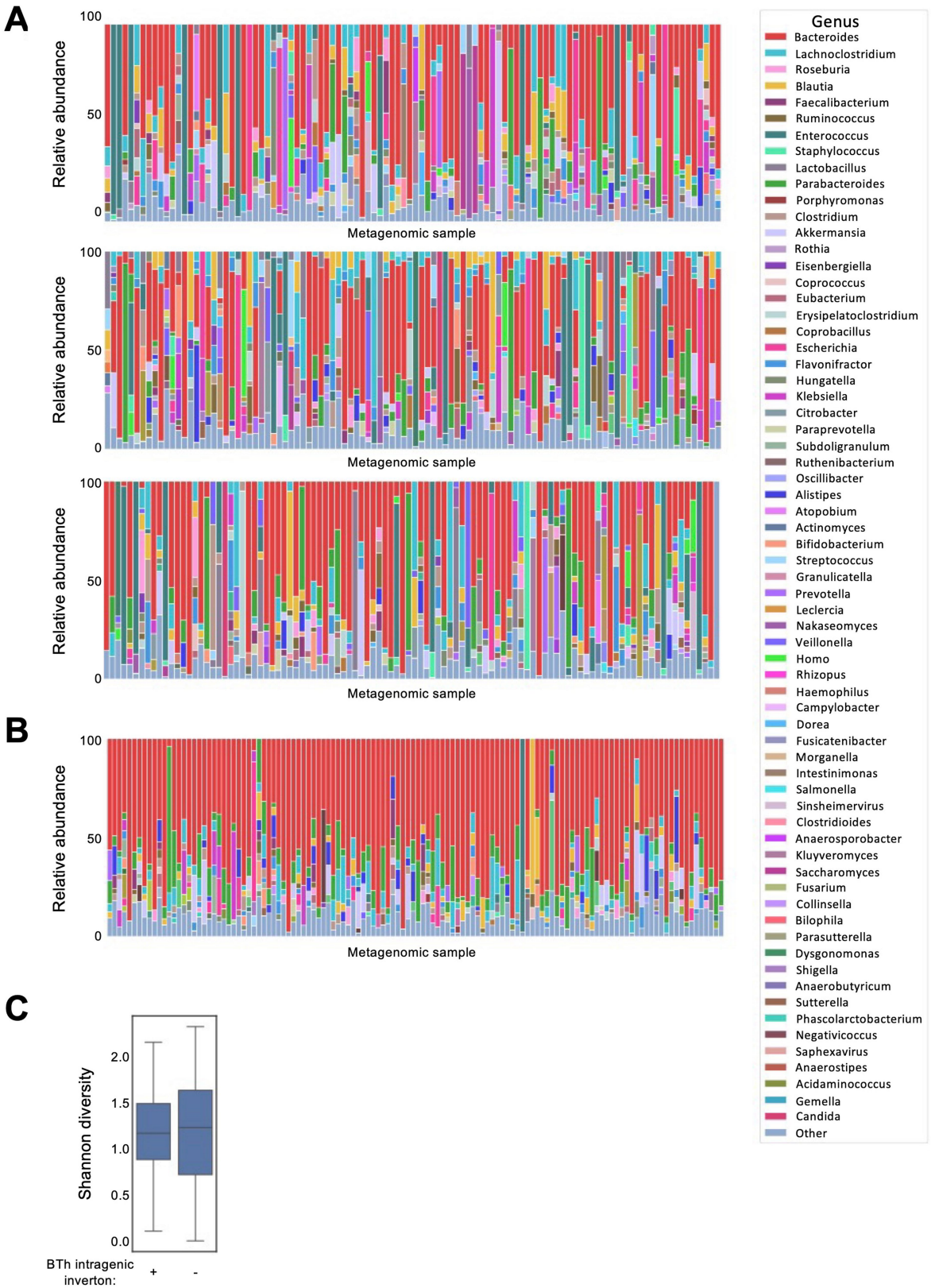
Correspondence and requests for materials should be addressed to Ami S. Bhatt.

Peer review information *Nature* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

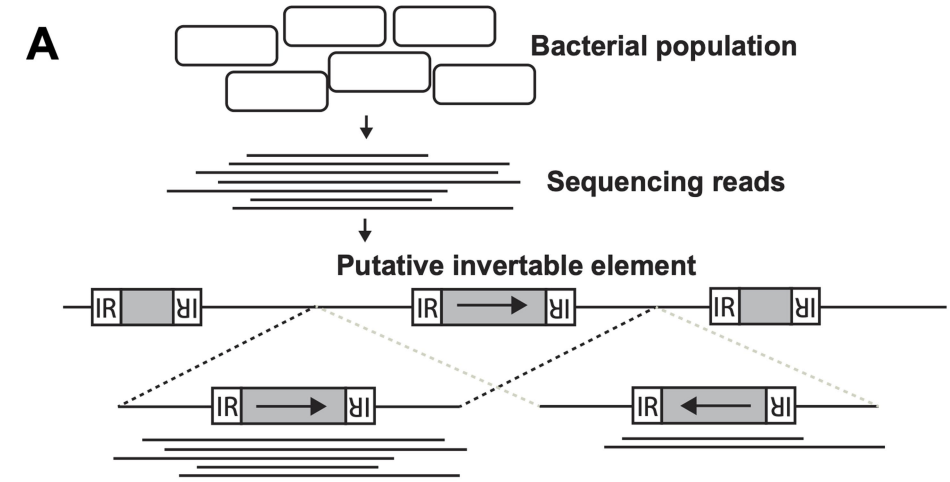


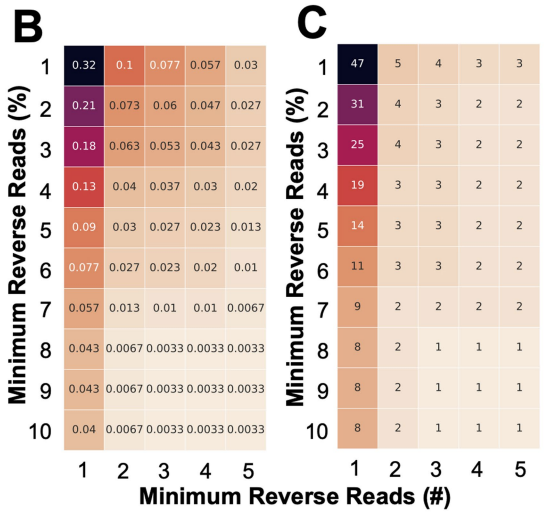


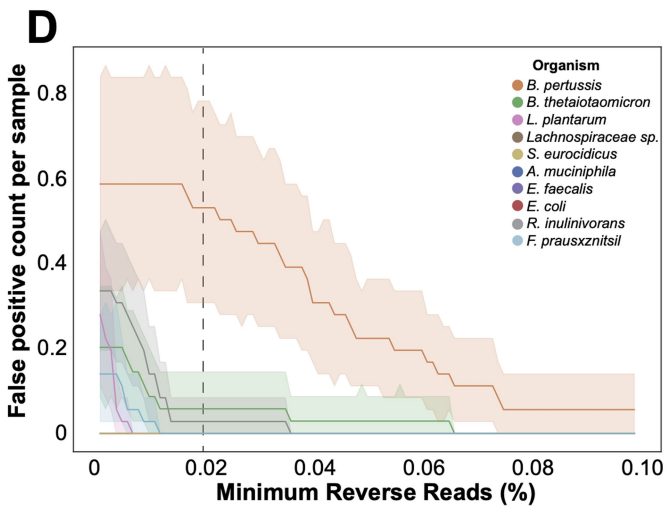
Extended Data Fig. 1 | Inverton detection and confirmation in BTh. (**A**) Inversion proportions of CPS loci in metagenomic samples measured with PhaseFinder (Top) and PhaVa (Bottom). Samples with no inversions in the five CPS invertons were removed. (**B**) Schematic of PCR confirmation. Forward and Reverse primers 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 (left), the Common and Forward primer will generate a PCR product. When the inverton flips, the Common and Reverse primer will generate a PCR product (right). HCT, hematopoietic cell transplantation.

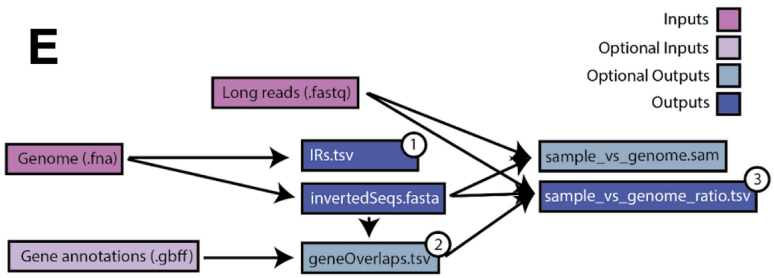


Extended Data Fig. 2 | Taxonomic composition of short-read samples from Siranosian et al. 2020. (**A-B**) Taxonomic distribution for samples at the genus level. Individual reads were taxonomically classified with Kraken2 using a Genbank reference set. Relative abundances were estimated with Bracken. Genera that represented less than 2% estimated relative abundance in a given sample were collapsed into 'other' for plotting. (**A**) Samples without detected BTh intragenic inversions and (**B**) samples with detected BTh intragenic inversions are shown. (**C**) The distribution of genus level Shannon diversity calculated for individual samples. Samples are grouped by presence or absence of BTh intragenic inversions. Center line represents the mean Shannon diversity of grouped samples, boxes represent quartiles, and whiskers extend to the furthest datapoint in either direction within 1.5 times the interquartile range.









(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	AT...CG	TACGGA...TATTAG	CG...AT

(2) geneOverlaps.tsv

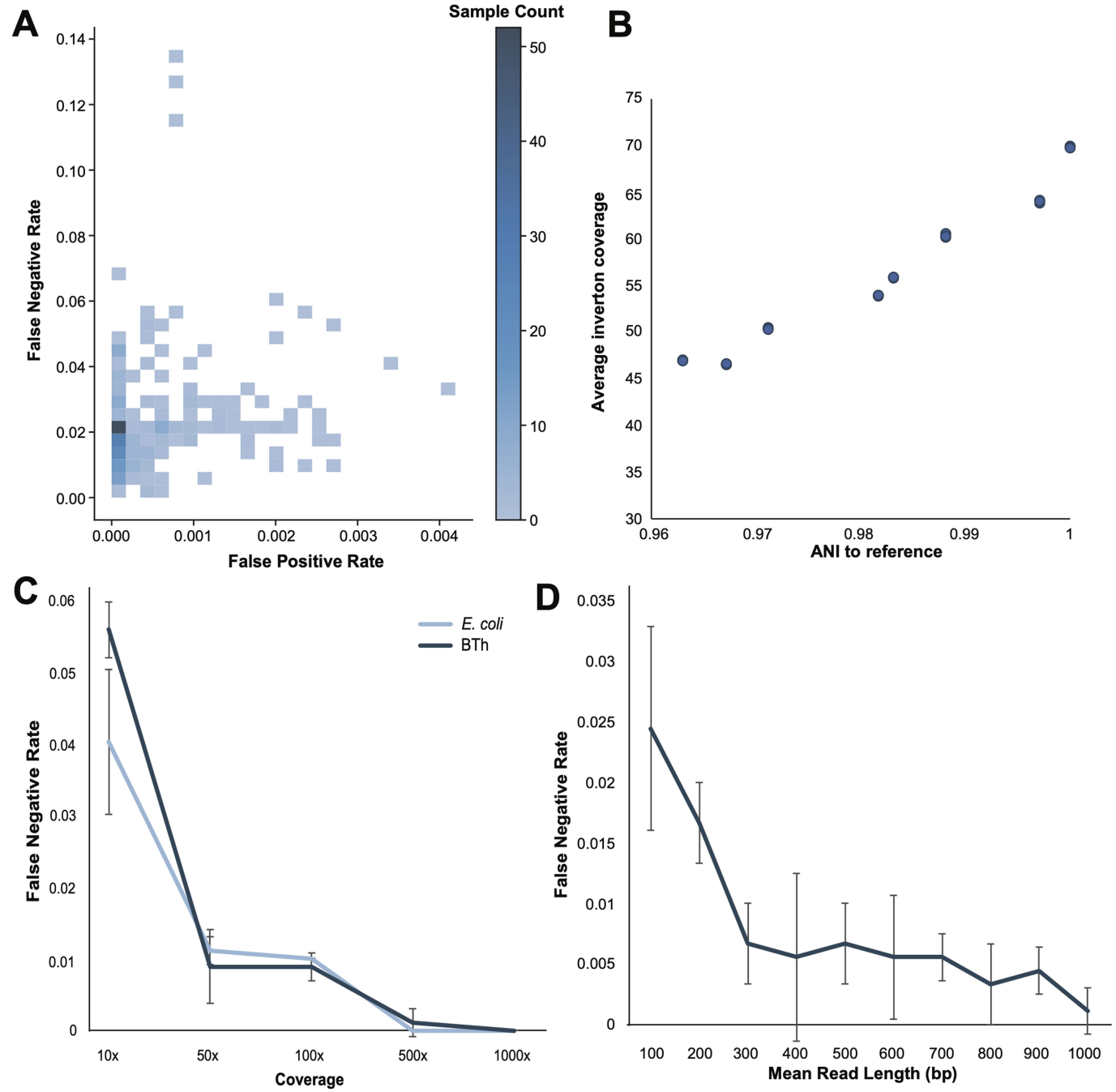
Inverton	gene overlaps	Upstream Gene	Upstream Gene Strand	Upstream Gene Distance	Downstream Gene	Downstream Gene Strand	Downstream Gene Distance
inv1	intragenic BT04	BT03	+	100	BT05	-	150

(3) sample_vs_genome_ratio.tsv

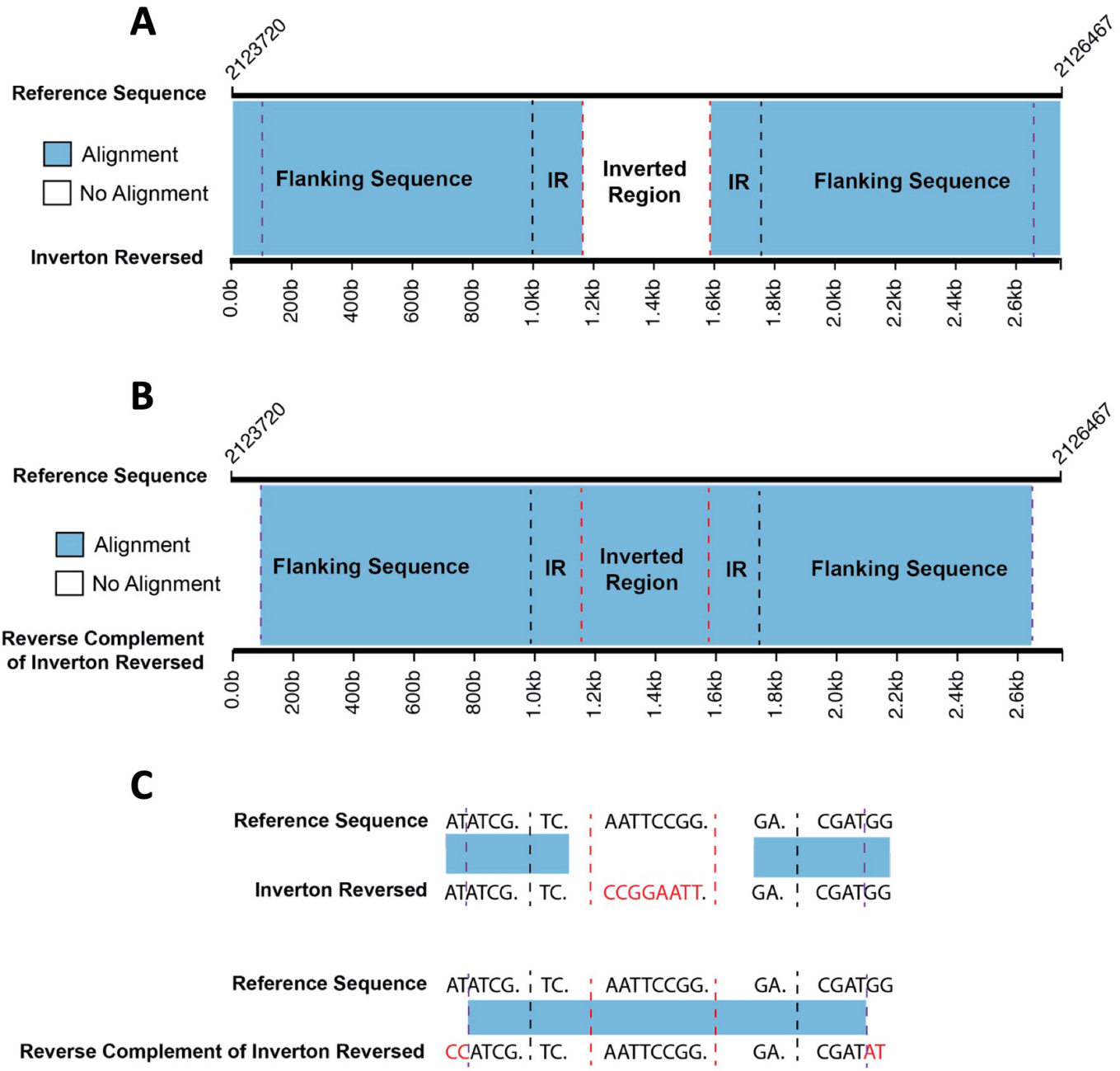
Inverton	gene overlaps	forward read #	reverse read #	reverse ratio	sample	Upstream Gene	Upstream Gene Strand	Upstream Gene Distance	Downstream Gene	Downstream Gene Strand	Downstream Gene Distance
inv1	intragenic BT04	15	5	0.25	SRR123	BT03	+	100	BT05	-	150

Extended Data Fig. 3 | See next page for caption.

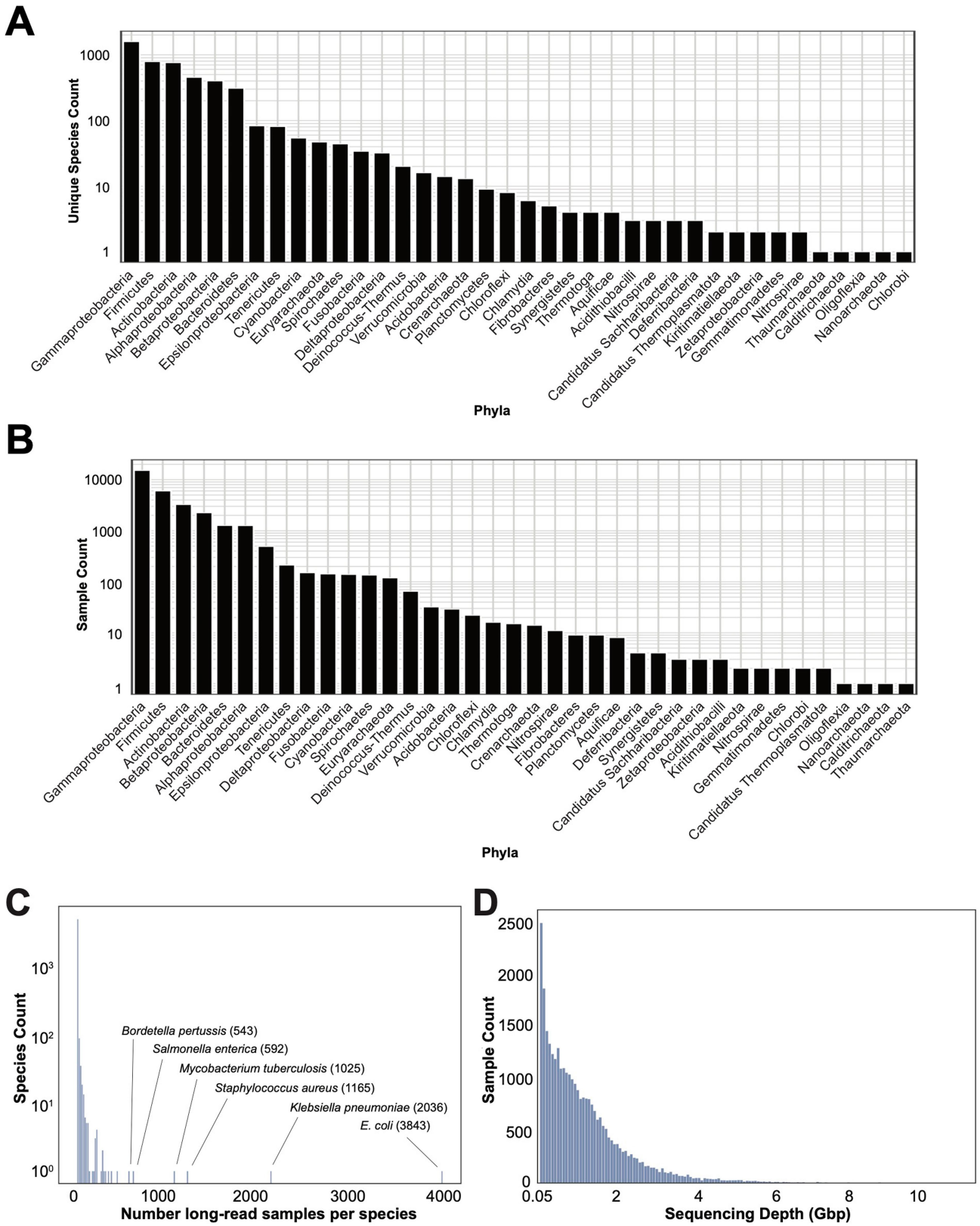
Extended Data Fig. 3 | Developing and optimizing PhaVa, a long-read based, accurate inverton caller. (**A**) Schematic of the PhaVa workflow. Putative invertons are identified and long-reads are mapped to both a forward (highlighted by the black dashed lines) and reverse orientation (highlighted by the grey dashed lines) version of the inverton and surrounding genomic sequence. Reads that do not map across the entire inverton and into the flanking sequence on either side or have poor mapping characteristics are removed. See methods for details. (**B**-**C**) Optimizing cutoffs for the minimum number of reverse reads, as both a raw number and percentage of all reads, to reduce false positive inverton calls with simulated reads. Cell color and number represent (**B**) the false positive rate per simulated readset and (**C**) the total number of unique false positives across all simulated datasets. (**D**) False positives in simulated data plotted per species. All measurements were made with a minimum of three reverse reads cutoff and varying the percentage of minimum reverse reads cutoff. Dashed line indicates the minimum reverse reads percent cutoff used for isolate and metagenomic datasets. Solid lines indicate sample mean while colored bands indicate 95% confidence interval. (**E**) Output tables of particular interest are labeled and shown below the diagram with example output.



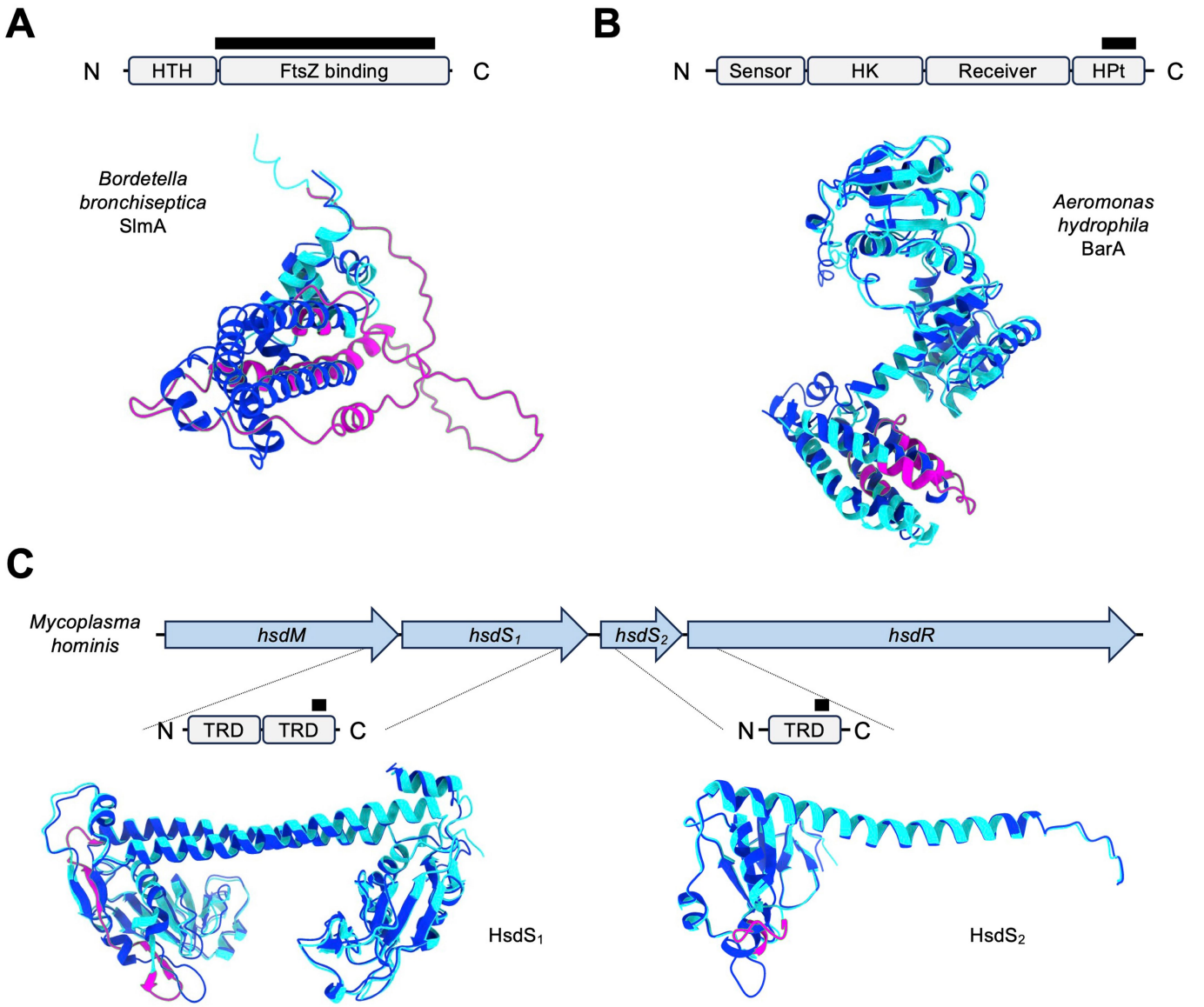
Extended Data Fig. 4 | Benchmarking PhaVa with pbsim2 simulated reads. (**A**) Density plot of false positive rate vs false negative rate for 100 bacterial species. Three 100x replicates were generated per species. (**B**) Scatterplot of simulated reads from 8 different *E. coli* genomes, with varying ANI to a singular reference *E. coli* genome, showing reduced coverage (and thus reduced detection) of invertons when mapping to a distant reference. Three 100x replicates were generated per genome. (**C**) False positive rate for read sets simulated for both *E. coli* and BTh and varying coverage levels. Three replicates were generated per coverage level. Error bars represent standard deviation between replicates. (**D**) False positive rates in readsets simulated from *E. coli* with varying mean read length. Three 100x replicates were generated per mean read length. Error bars represent standard deviation between replicates.



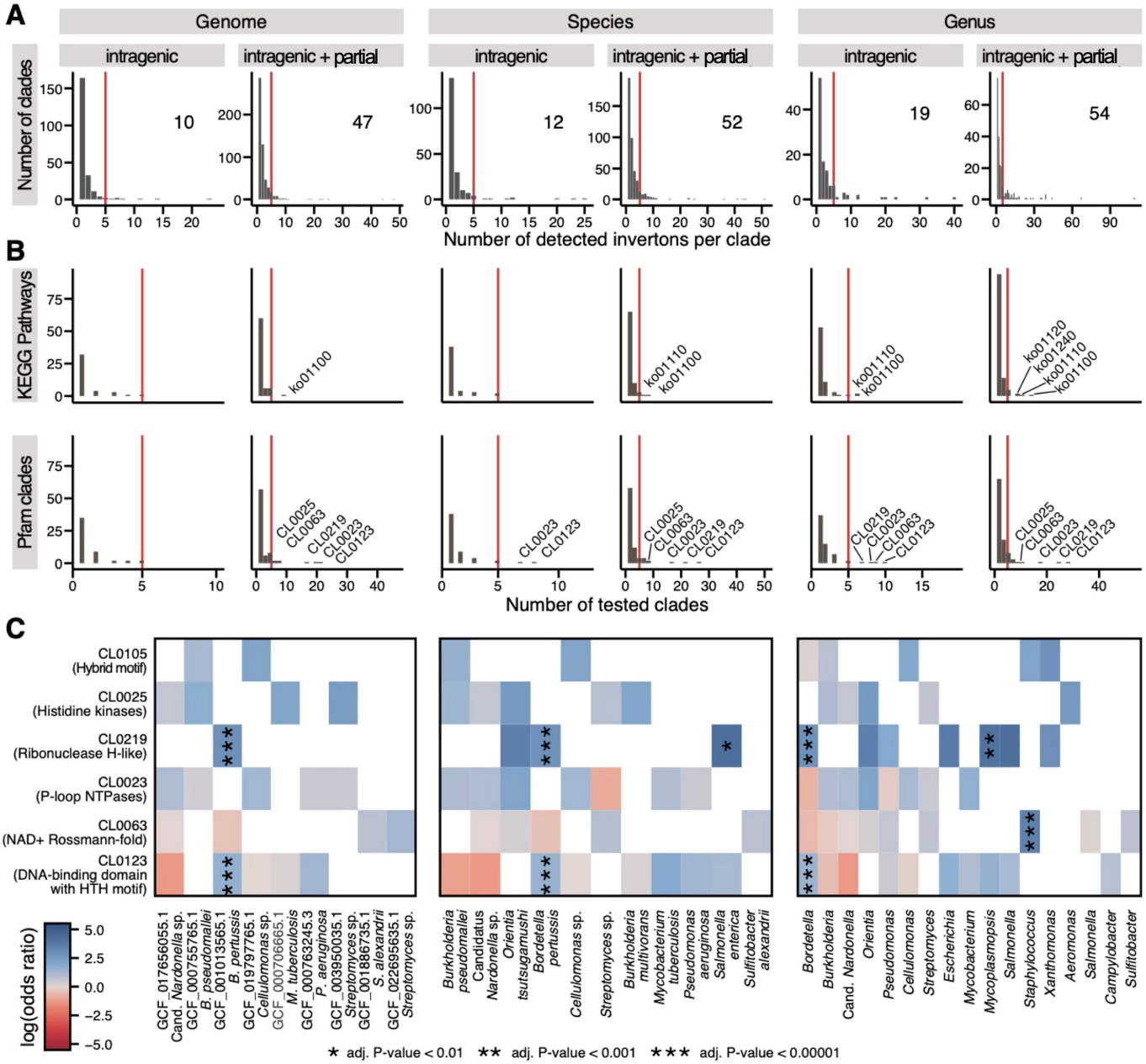
Extended Data Fig. 5 | Very long (>750 bp), near perfect, inverted repeats can lead to false positives. (**A**) Alignment of inverton NZ_CP025371.1:2124719-2124870-2125316-2125467, with its invertible sequence flipped, against the *B. pertussis* genome leads to perfect alignment of flanking and IR regions 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.



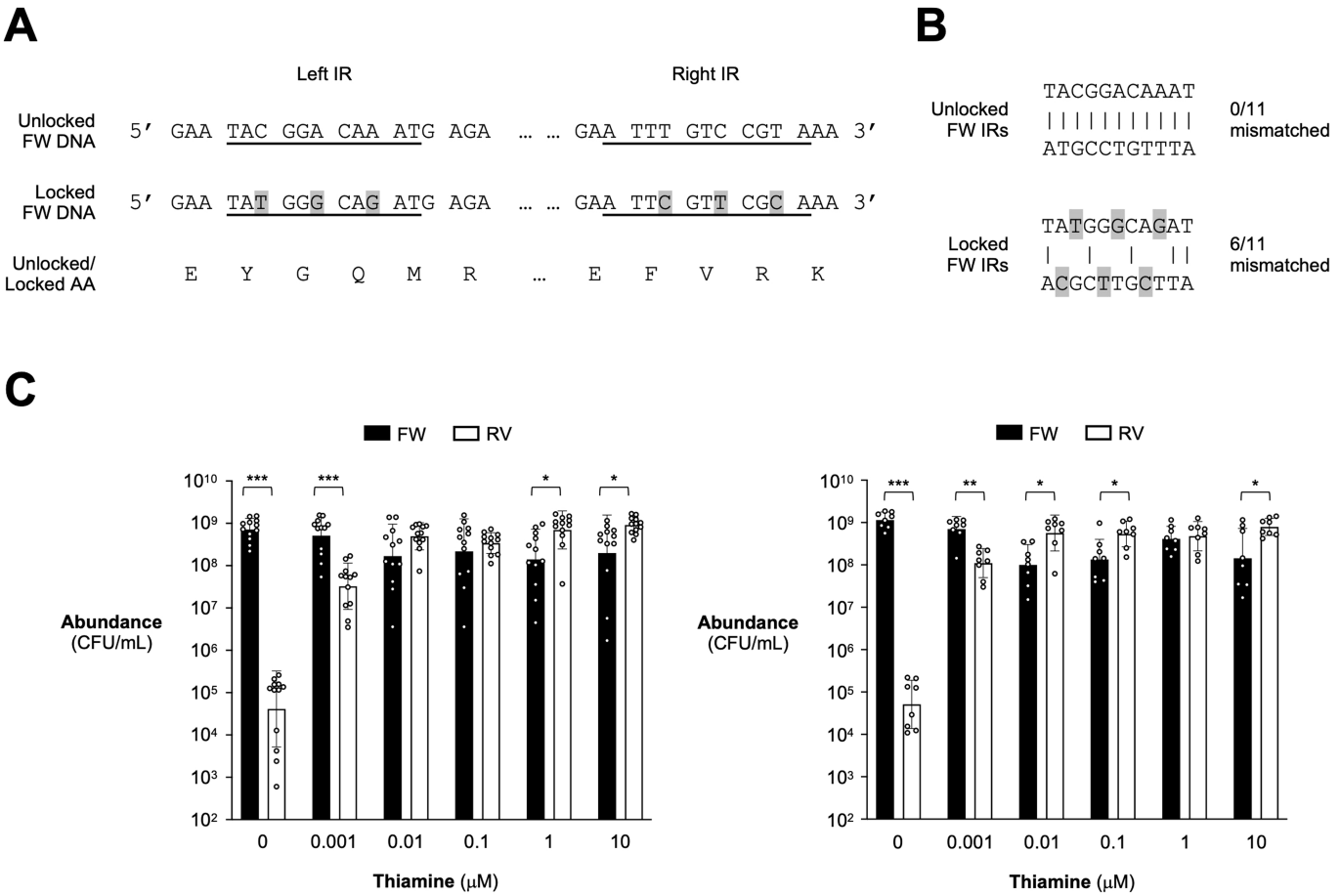
Extended Data Fig. 6 | Overview of SRA long-read isolate sequencing samples analyzed with PhaVa. (**A**) The number of unique species represented in the dataset, grouped by phylum. (**B**) The raw number of sequencing samples, grouped by phylum. (**C**) Histogram of sequencing samples per species. Species with large numbers of samples are labeled. (**D**) A histogram of sequencing depths for all long-read isolate sequencing samples.



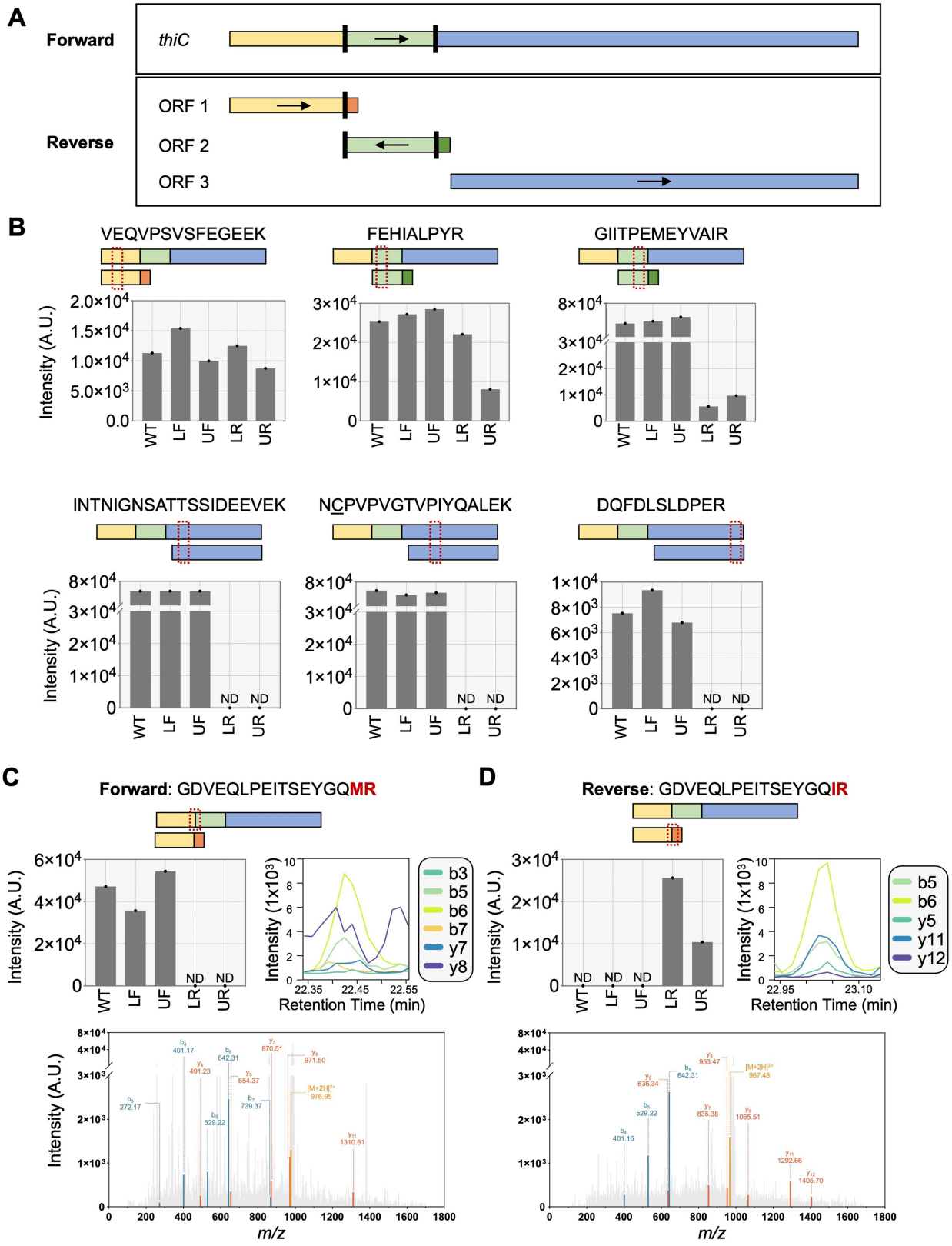
Extended Data Fig. 7 | Intragenic invertons that recode proteins are identified in long-read isolate datasets. (**A**-**C**) Genome diagrams for recoding intragenic invertons are shown. Grey boxes indicate annotated protein domains. Black lines indicate the region contained within the inverton. AlphaFold structure of the forward (dark blue) and reverse (light blue) are shown. Amino acids affected by the inverton are shown in pink. (**A**) *slmA* nucleoid occlusion factor *Bordetella bronchiseptica*. RMSD 26.287 angstroms across all pairs. pLDDT forward: 94.5. pLDDT reverse: 51.3. (**B**) *barA* two-component sensor histidine kinase in *Aeromonas hydrophila*. The Receiver and HPt domain are shown. RMSD 5.492 angstroms across all pairs. pLDDT forward: 83.4. pLDDT reverse: 76.2. (**C**) Type I restriction enzyme S subunit $hsdS_1$ and $hsdS_2$ in *Mycoplasma hominis*. RMSD 1.809 and 4.167 angstroms across all pairs, respectively. pLDDT $hsdS_1$ forward 90.8, reverse 87. pLDDT $hsdS_2$ forward 92.4, reverse 84.2. HTH, helix-turn-helix; HK, histidine kinase; HPt, histidine phosphotransfer domain; TRD, target recognition domains.



Extended Data Fig. 8 | 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.

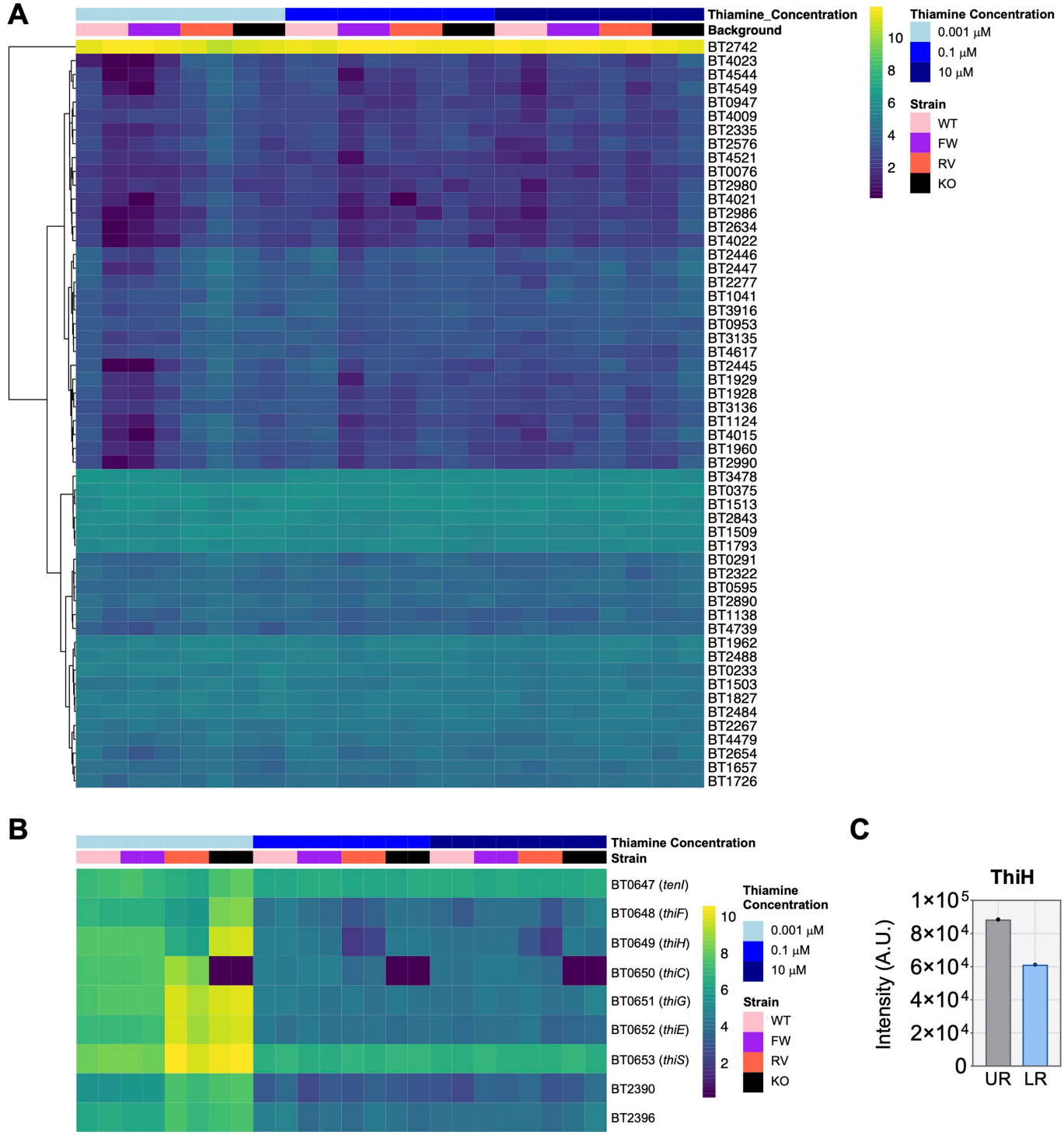


Extended Data Fig. 9 | Locked *thiC* intragenic inverton construction and growth competition. (**A**) Generation of locked intragenic invertons. The forward and locked forward *thiC* inverted repeat (IR) nucleotide sequences are shown. When possible, the wobble position of each codon corresponding to the IR was mutated to increase mismatches between the two palindromic sequences while maintaining the amino acid sequence. (**B**) Mutated nucleotides are highlighted in grey. (**C**) Locked *thiC* strains were competed against each other in thiamine-containing media in a 1:1 ratio. After 40 h, 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. Left - the locked forward strain is marked with an erythromycin resistant cassette and the locked reverse strain is marked with a tetracycline resistant cassette. Right - 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. Each dot represents an individual replicate. Experiments were done in biological duplicate or triplicate and repeated 4 or 6 times. A two-tailed ratio paired t test was performed on the locked forward and locked reverse abundances. ***, $p < 0.001$; **, $p < 0.01$; *, $p < 0.05$.

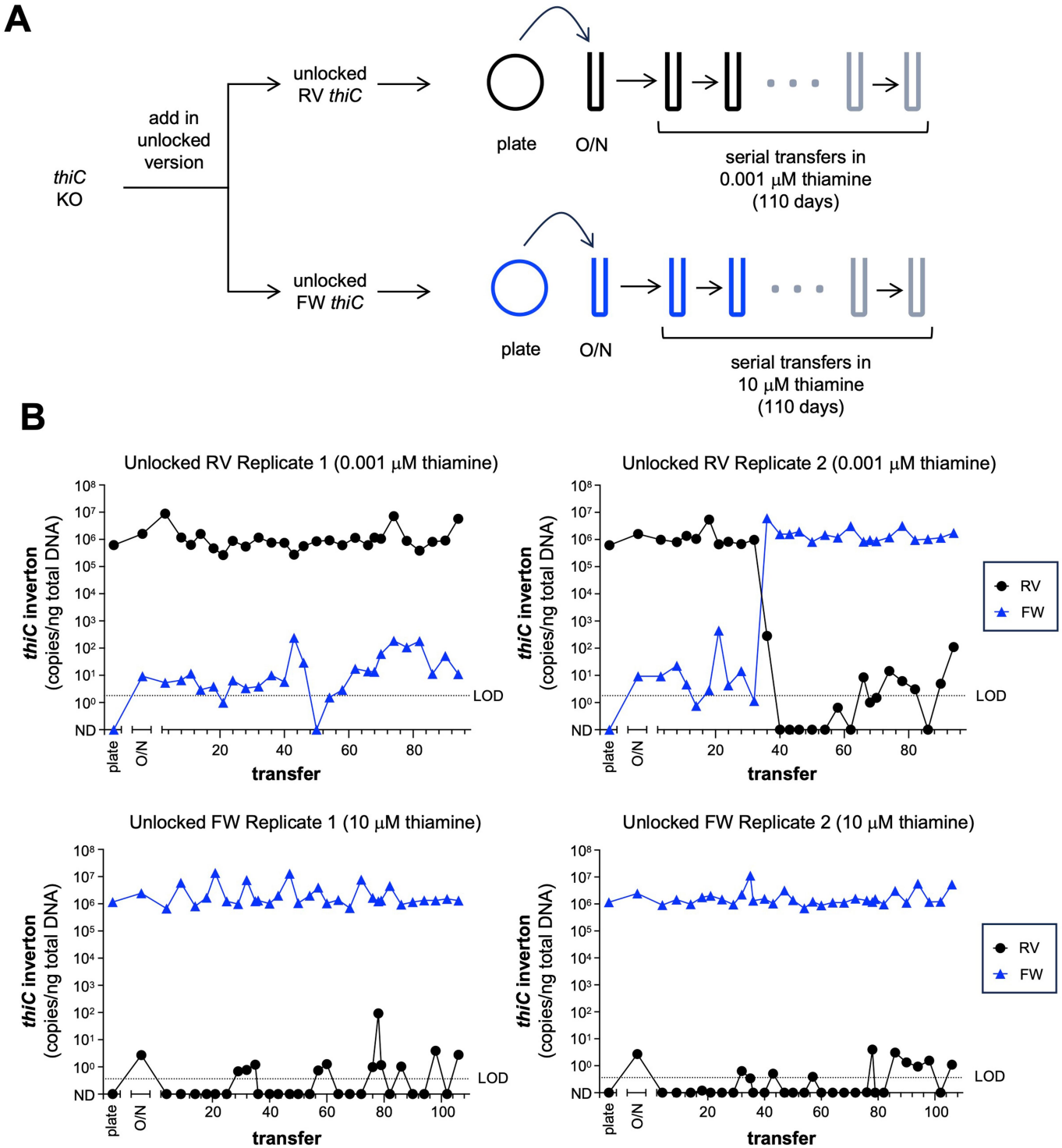


Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Detection of ThiC unique peptides in the forward and reverse orientation. (**A**) Schematic showing the ThiC protein sequence in the forward and predicted reverse open reading frames (ORF1, ORF2, ORF3). Arrows indicate the direction of transcription. Shared colors between the forward and reverse ORFs indicate identical amino acids. The dark orange box in ORF1 is the ThiC reverse specific peptide. (**B**) Mass spectrometry quantification, by data-independent acquisition, of ThiC tryptic peptides aligned with each ORF. NCPVPVGTVPIYQALEK includes cysteine carbamidomethylation. (**C**) Quantification of the unique ThiC tryptic peptides that align exclusively to the forward or reverse sequences. Representative extracted ion chromatograms show the identified fragment ions applied in the quantification of the forward and reverse peptides, as detected in WT and LR, respectively. Representative MS/MS spectra of the unique ThiC forward peptide GDVEQLPEITSEYGQMR detected in WT and unique ThiC predicted inverton peptide GDVEQLPEITSEYGQIR detected in LR. Spectra include the MS1 precursor ion (orange), as well as *b*- and *y*- fragment ions (blue and red, respectively). ND, not detected; WT, wild-type; LF, locked forward *thiC* intragenic inverton; UF, unlocked forward *thiC* intragenic inverton; LR, locked reverse *thiC* intragenic inverton; UR, unlocked reverse *thiC* intragenic inverton.



Extended Data Fig. 11 | Effect of thiamine on BTh invertases and thiamine biosynthesis and uptake loci. (**A**) Log_2RPKM values for all annotated invertases in BTh are shown across the *thiC* mutant backgrounds and thiamine concentrations. Axis is clustered with pheatmap's default parameters (Euclidean). (**B**) Log_2RPKM values for transcript levels of all genes in the thiamine biosynthesis pathway (BT0647-0653) and genes involved in uptake of thiamine (BT2390, BT2396) are denoted. (**C**) Intensity of ThiH protein as determined by mass spectrometry using data-independent acquisition. UR, unlocked *thiC* reverse; LR, locked *thiC* reverse.



Extended Data Fig. 12 | Thiamine is not a strong driver of *thiC* intragenic inverton flipping. (**A**) Schematic denoting the generation of unlocked *thiC* strains and the experimental setup for the long-term thiamine exposure assay. (**B**) Results of the long-term thiamine exposure assay. Top – two replicates of the unlocked reverse *thiC* strain that were serially cultured in low thiamine (0.001 μM). Bottom – two replicates of the unlocked forward *thiC* strain that were serially cultured in high thiamine (10 μM). LOD, limit of detection; O/N, overnight culture; ND, not detected.

nature portfolio

Corresponding author(s): Bhatt

Last updated by author(s): Jul 23, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Candidate isolate long-read sequencing datasets were identified on NCBI with the following search criteria: "(Bacteria[Organism] OR Archaea[Organism]) AND ("pacbio smrt"[Platform] OR "oxford nanopore"[Platform]) AND genomic[Source]". Datasets were further filtered by removing datasets with the "amplicon" flag, and removing datasets with less than 50 Mbp of sequencing in total. Individual read datasets were downloaded with fastq-dump, a part of the sratoolkit (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software). Nanostat was run on each remaining readset to measure dataset characteristics.
Data analysis	Long-read datasets were then analyzed with PhaVa (v0.1.0) with default parameters. PhaVa is available at https://github.com/patrickwest/PhaVa

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Short-read adult HCT stool sequencing data was previously published and is available at (NCBI BioProject ID PRJNA707487). Short-read pediatric HCT stool sequencing data was previously published and is available at (NCBI BioProject ID PRJNA787952). Long-read metagenomic sequencing data was previously published and is available at BioProject PRJNA820119 and BioProject PRJNA940499. Assembled metagenomic contigs are available at https://doi.org/10.5281/zenodo.7662825. A list of accession numbers for long-read isolate sequencing data is available in supplementary file Data S5. MS raw files (.d) generated in this study have been deposited to the public repository, ProteomeXchange Consortium, through the PRIDE partner repository (project accession is PXD054577). Long-read sequencing data for the locked thiC intragenic inverton strains and RNA sequencing data are available NCBI BioProject ID PRJNA1118344. Accession codes for long read datasets are listed in the extended data section. The reference genome for B. thetaiotaomicron VPI-5482 is the NCBI Reference Sequence AE015928.1. The reference genome for B. fragilis FDAARGOS_1225 is the NCBI Reference Sequence NZ_CP069563.1.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	NA
Reporting on race, ethnicity, or other socially relevant groupings	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample-size calculation was performed. For our in vitro assays with significance testing, we followed standard laboratory procedure to repeat experiments in biological duplicate or triplicate on at least three separate occasions. For in vitro assays, this has historically been sufficient to identify statistical differences between groups using routine statistical tests.
Data exclusions	Candidate isolate long-read sequencing datasets were chosen by a predetermined set of criteria described in detail in the methods section. No in vitro data was excluded.
Replication	For the in vitro competitive growth assays, antibiotic markers were flipped between strains to control for any fitness advantages/ disadvantages these markers would provide. Experiments replicated as expected
Randomization	No randomization was performed for in vitro assays as inclusion in a group was determined by the bacterial genotype. We controlled for confounding variables by treating all samples in exactly the same way and with the same media.
Blinding	Blinding was not done as there was no opportunity for this to occur.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging