Title: Intragenic DNA inversions expand bacterial coding capacity

Authors: Rachael B. Chanin[1]†, Patrick T. West[1]†, Ryan M. Park[1], Jakob Wirbel[1], Gabriella Z. M. Green[1], Arjun M. Miklos[1], Matthew O. Gill[2], Angela S. Hickey[2], Erin F. Brooks[1], Ami S. Bhatt[1,2]*

Affiliations:

† These authors contributed equally to this work, listed alphabetically by last name

[1]Department of Medicine (Hematology, Blood and Marrow Transplantation); Stanford, USA.

[2]Department of Genetics, Stanford University; Stanford, USA.

*Corresponding author. Email: asbhatt@stanford.edu

Abstract: Bacterial populations that originate from a single bacterium are not strictly clonal. Often they contain subgroups that have distinct phenotypes. One way that bacteria generate this heterogeneity is through phase variation: enzyme-mediated, reversible inversion of specific intergenic regions of genomic DNA. These DNA inversions can flip the orientation of promoters within otherwise isogenic populations, impacting fitness and survival. We developed and applied bioinformatic approaches that enabled the discovery of thousands of previously undescribed phase-variable regions in prokaryotes using long-read datasets. We identified 'intragenic invertons', a surprising new class of invertible elements found entirely within genes, across the prokaryotic tree of life. Intragenic invertons allow a single gene to encode two or more versions of a protein by flipping a DNA sequence within the gene, thereby increasing coding capacity without increasing genome size. We experimentally characterize specific intragenic invertons in the gut commensal *Bacteroides thetaiotaomicron*, presenting a 'roadmap' for investigating this gene-diversifying phenomenon.

One-Sentence Summary: Intragenic DNA inversions, identified using long-read sequencing datasets, are found across the prokaryotic tree of life.

Main Text:

Adaptation is a cornerstone of survival for any species. Eukaryotes use alternative splicing to produce different transcripts, and therefore different adaptive states, from a single gene. This is an example of how an organism can maximize the potential of its genomic content. In contrast to eukaryotes, prokaryotes are not known to use alternative splicing. Instead, prokaryotes achieve broad variations in phenotype through phase variation (*1–3*). One common type of preprogrammed variation occurs through DNA inversion. Site-specific recombinases recognize a pair of inverted repeats in genomic DNA and invert the intervening DNA sequence (*3*). 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. The change in antigen expression determined whether the bacterium was bound by antiserum, and thus was termed a 'phase-determining event'(*4–6*). This DNA inversion, and the many others that have since been discovered, play critical roles in both commensal and pathogenic bacteria.

For decades, these invertible loci were identified individually. Then, computational approaches enabled higher-throughput discovery of these 'invertons' across the genomes of a small subset of specific bacterial species (*7*, *8*). Most recently, in 2019, Jiang *et al.* developed a method that facilitated broad scale identification of 4,686 intergenic invertons (i.e., invertons between genes) through a search of 54,875 bacterial reference genomes (*9*).

Here, we find that the same mechanisms underlying intergenic invertons can occur entirely within a gene. These intragenic invertons expand or alter 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. We develop PhaVa, a long-read based tool to identify both intragenic and intergenic invertons. By applying PhaVa to ~30,000 bacterial isolates from ~4,000 unique species, we find that intragenic invertons occur across the prokaryotic tree of life. In particular, we focus on *Bacteroides thetaiotaomicron*, a model enteric commensal, and validate 15 intragenic invertons experimentally. Finally, we make both the PhaVa software package and all of the findings of intergenic, partial intergenic (*10*), and intragenic invertons publicly available.

Results

As many bacteria are studied in laboratory conditions, invertons that provide a fitness advantage *in vivo* but may not be advantageous to fitness *in vitro* have likely been overlooked (*11–13*). We therefore hypothesized that there are currently unknown gut-relevant invertons. To test our hypothesis, we first examined Bacteroidetes, a prevalent host gut-associated phylum, which includes organisms known to have many intergenic invertons (*9*). To orthogonally

confirm sequencing-based observations, we focused our analysis on *B. thetaiotaomicron*, a genetically tractable species suitable for downstream experimental manipulation. We used PhaseFinder (*9*), a short-read, reference-based inverton detection pipeline, with *B. thetaiotaomicron* VPI-5482 (BTh) as the reference genome, and relatively relaxed filters to increase sensitivity (see methods, Fig. 1). The input samples were a longitudinal collection of sequenced stool samples from 149 adult and 21 pediatric patients undergoing hematopoietic cell transplantation (*14*, *15*) (Fig. 1A). These samples were selected given the varying and complex environment enteric bacteria would encounter over time, with many different stressors present such as antibiotic treatment and inflammation, which we hypothesized might induce inverton flipping.

To first test if PhaseFinder could detect BTh invertons in our metagenomic samples, we examined BTh's capsular polysaccharide (CPS) genes. BTh has 8 loci that encode different CPS, 5 of which are controlled by invertible promoters (*16–18*). CPS are important mediators of phage susceptibility (*19*) and can modulate the host immune system (*20–22*). Using PhaseFinder on the patient sample datasets, we found read evidence of all 5 CPS invertons in both the reference and inverted orientations (fig. S1), demonstrating that PhaseFinder is able to detect known invertons in these metagenomic samples and that these samples have enough *Bacteroides* sequencing depth to call invertons. Of note, in the reference BTh genome, the invertible promoter for each of these 5 loci is in the 'OFF' state by virtue of it being oriented in the opposite direction of the CPS genes. Similarly, *in vitro* transcriptional analyses support the finding that the majority of invertible CPS loci are in the OFF orientation (*23*), suggesting that in laboratory conditions, these loci are not utilized. Finding read evidence of inversion for all invertible CPS loci suggests that the *in vivo* patient datasets are an ideal environment to detect potentially rare invertible events.

Unexpectedly, we also found read evidence of intragenic invertons, invertons found entirely within single genes, across many samples (Fig. 1B). To date, the only description of invertible DNA sequences entirely within a gene are in isolated cases of very short (7bp) flips within mitochondrial DNA in certain pathogenic states (*24*). These 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. In the intragenic invertons that we observed, there were two predicted consequences. In some cases, the intragenic inverton resulted in a portion of the protein being "re-coded" (Fig. 1C). For example, we observed a 57 bp inversion in BT0375, the invertase that is believed to flip the adjacent CPS1 invertible promoter. Intragenic inversion could alter the binding specificity of the invertase, possibly changing the invertible repeats (IRs) that it targets for flipping or the binding affinity of the invertase for its cognate IRs. 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 (called with Prodigal (*25*)). For example, the inverton in the hybrid two-component BT3786 occurs between domains potentially untethering "sensing" and "response" elements (Fig. 1D). However, we also observed intragenic inversions that

resulted in zero, one, three, or more ORFs .We refer to these two different types of intragenic inversions as those that are 'recoding' and those that cause a 'premature stop', respectively.

To validate these predicted intragenic inversions, we analyzed the DNA sequences in these gene regions *in vitro*. We extracted DNA from wild-type BTh grown in either rich or defined media and designed PCR primer sets that enabled us to amplify either the reference or the inverted version of several invertons (fig. S2). We tested 64 of the 67 predicted intragenic invertons and confirmed that 11 of them had DNA molecules in both the reference and inverted orientation in our laboratory-grown population of BTh (Fig. 1B, Table 1). The 53 unconfirmed intragenic invertons may be due to missing cues in the growth conditions and/or false positives in the read based evidence.

Genomic structural variation often involves highly repetitive or low complexity regions. Short-reads are often not long enough to resolve these sequences (*26*). We, therefore, developed a long-read based inverton predictor, PhaVa. PhaVa maps long reads against both a forward (identical to reference) and reverse orientation version of potential invertons (Fig. 2A). PhaVa's accuracy is high because it requires long reads that span the entire length of a given inverton in order to make a 'call' about the orientation of a given inverton. To ensure accurate performance of PhaVa, we optimized read mapping parameters by simulating long-read datasets from ten bacterial genomes at various sequencing depths (Fig. 2B-C). The reads were generated from a reference genome, and thus no invertons are expected and any detected would be false positives. In general, the false positive rate was very low (mean false positive count per simulated sample of <0.1 in 9/10 species), with the exception of reads simulated from the *Bordetella pertussis* genome (Fig. 2C). Further investigation revealed the false positives detected in *B. pertussis* were due to a putative inverton with inverted repeats longer than 750 bps, of which only a smaller portion were detected (fig. S3).

To find invertons are across prokaryotic genomes, we ran PhaVa on ~30,000 isolate long-read datasets deposited on SRA, limiting to readsets belonging to Bacteria or Archaea and with 50 Mbp or more of total sequencing, spanning ~4,000 unique species in total (fig. S4). Like Jiang *et al. (9)*, we found that Bacteroidetes have a large number of intergenic invertons (Fig. 3A) and intergenic invertons per genome (Fig. 3B). Fusobacteria, Gammaproteobacteria, and Verrucomicrobia also have high numbers of intergenic invertons per genome (Fig. 3B), with Verrucomicrobia having the highest number per genome overall. In our dataset, Verrucomicrobia is composed of only*Akkermansia* strains. As increases in *Akkermansia* abundance correlate with protection against metabolic disease (*27*, *28*), there is interest in its use as a probiotic. However, *Akkermansia* strains exhibit broad phenotypic diversity and differential gut colonization ability (*29*). The role of invertible elements in beneficial outcomes has yet to be explored. We also found a large number of invertons partially overlapping genes, so-called 'partial intergenic invertons', where the overlap occurs either at the start or stop of a gene and the intergenic sequence (Fig. 3A). The partial intergenic invertons are significantly longer than intergenic or intragenic invertons (Fig. 3C, p=0.0 and p=4.5e-146 with a t-test, respectively). Some of these

partial intergenic invertons occur in “shufflon” systems and have been described in previous studies (*30*, *31*).

We also found widespread evidence of intragenic invertons across multiple phyla, including the major gut microbiome-related phyla, Proteobacteria, Firmicutes and Bacteroidetes (Fig. 3A-B). A small number of intragenic invertons were found in Archaea, as well (Fig 3A-B). We found the largest number of intragenic invertons in Gammaproteobacteria, including from organisms such as *Escherichia coli* and *Salmonella*; this is largely due to the abundance of samples for these organisms in SRA and our dataset (~4,000 *E. coli* samples (fig. S4)) given that Gammaproteobacteria have a relatively small number of intragenic invertons detected per genome (0.21, Fig 3B).

Interestingly, the mean number of intragenic invertons per genome varied greatly between different phyla (Fig. 3B) with Tenericutes, Betaproteobacteria, and Actinobacteria having a relatively high number of intragenic invertons detected per genome, at 0.47, 0.69, and 0.48 respectively. The distribution of inversion proportions of individual intragenic invertons was different from that of intergenic invertons (Fig. 3D); 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. 3D), indicating 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 due to mutations in the IR or loss of the invertase that flips the inverton.

Next, we looked for shared functions across genes that contain intragenic invertons by doing a clade-resolved enrichment analysis. We calculated gene set enrichments (using Pfam clan definitions as gene sets) per genome, species, and genus, combining the genes from all genomes in a specific clade (Fig. 3E, S5). 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) (fig. S5). Taken together, we found that intragenic invertons occur more frequently in genes that have DNA binding or DNA/RNA modifying activity.

As inverton orientation likely relates to the environment of a bacterium, we also ran PhaVa on 210 *de novo* assembled long-read metagenomes from the human gut (*32*, *33*), mapping sequencing reads back to their respective assembly. This enabled us to detect invertons that may be absent in isolate bacterial cultures, but present *in vivo.* We identified over 3,500 putative invertons, largely from Bacteroidetes and Firmicutes (fig. S6A). 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. S6B). Taken together, analysis of the metagenomic data with PhaVa suggests that bioinformatic analysis of isolate genomes grown in laboratory conditions likely underestimates the number and range of invertons that exist in microbes. This is likely 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 analyses provide strong evidence that intragenic invertons are present across multiple phyla of bacteria and also in selected archaea. To evaluate the phenotypic consequences of inversion, we focused on the intragenic inverton found in the BTh BT0650 gene (thiamine biosynthesis protein ThiC) as a test case (Fig. 4). 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, like many gut microbes, including BTh, have the capacity to biosynthesize thiamine, albeit at a large energetic cost (*34*, *35*). Thus, thiamine availability has been hypothesized to strongly influence microbial community composition (*36*). We chose to characterize the intragenic inverton in *thiC* as this gene has a defined function in thiamine biosynthesis (*37*, *38*). Specifically, ThiC catalyzes 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). Furthermore, we predicted this inverton introduce a premature stop codon in the *thiC* mRNA, which would result in a truncated protein (Fig. 4A). BTh can grow in the absence of exogenous thiamine as it can synthesize thiamine *de novo,* however, strains that lack ThiC lose this ability. We hypothesized that inversion of the invertible locus in *thiC* would interfere with thiamine biosynthesis.

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. Unfortunately, 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 (fig. S7). Using this method, we created a locked forward (reference orientation) and a locked reverse (flipped intragenic inverton) *thiC* strain.

Next, we grew wild-type BTh, locked forward, locked reverse, and a *thiC* knockout strain 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 abundance regardless of whether thiamine was added to the media. By contrast, the locked reverse strain mirrored the *thiC* knock-out strain and was only able to grow to wild-type levels when 0.1 μM or greater thiamine was added to the media. This finding confirms our expectation that the reverse version of the intragenic inverton interferes with ThiC function. Additionally, we competed the locked forward against the locked reverse in an equal ratio in varying concentrations of thiamine and determined the competitive index (CI), the ratio of the recovered locked forward strain over the reverse strain (Fig. 4D). In the absence of thiamine, the locked forward strain outcompeted the reverse (CI of 51,703). When

0.001 μM thiamine was added to the media, the competitive advantage decreased (CI of 7.1). However, at concentrations above 0.01 μM, the fitness advantage of the locked forward was completely abolished. In fact, at 0.01, 0.1, and 10 μM the reverse strain had a slight growth advantage over the forward strain (CIs of 0.23, 0.27, and 0.20 corresponding to a 4.3, 3.7, and 5.0 fold advantage for the reverse orientation, respectively). In the intestine, thiamine concentration varies between 0.02-2uM (*39*). This suggests that at physiologically relevant thiamine concentrations, the strain with the *thiC* inverted is more fit than the forward strains in these conditions. Taken together, we find evidence of a physiologically relevant condition in which intragenic inversion within the *thiC* gene may provide an energetic or other form of competitive advantage, and thus might be adaptive (*40*).

Intragenic invertons are a potential mechanism for genetic variation and adaptation in bacteria. However, it remains to be determined how intragenic inversions are regulated (*41*). It is possible that some invertases are expressed at a basal level and therefore there is a baseline, low level of inversion that occurs in a small proportion of the population. Alternatively, some invertases may be expressed in response to specific stimuli and then can result in coordinated flipping of invertons. While the flipping is reversible, presumably the fitness benefits that result from an inversion, or similar factors, may cause selective pressures to then favor an organism with one orientation of the inverton versus another.

In conclusion, we present evidence of widespread intragenic invertons across the prokaryotic tree of life. This finding reshapes a fundamental tenet of microbiology, that one bacterial gene only leads to one protein. Here we highlight how genes containing these elements may produce two or more distinct proteins. Furthermore, we created PhaVa, a long-read sequencing analysis workflow, to identify additional intragenic invertons. In this manuscript, we present a 'roadmap' for more in depth investigation of a specific invertible intragenic locus. Future studies that apply methods such as transcriptomics, ribosome profiling, and proteomics will facilitate this. We anticipate future research in this area to have profound effects on synthetic biology, microbe-microbe interactions, and microbial physiology. Furthermore, we anticipate that future studies of intragenic inverton regulation will uncover new layers of bioregulation in prokaryotes.

References

1. K. P. Obergfell, H. S. Seifert, Mobile DNA in the pathogenic Neisseria. *Microbiol Spectr*. **3** (2015), doi:10.1128/microbiolspec.MDNA3-0015-2014.

2. R. Moxon, C. Bayliss, D. Hood, Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu. Rev. Genet.* **40**, 307–333 (2006).

3. D. Trzilova, R. Tamayo, Site-Specific Recombination - How Simple DNA Inversions Produce Complex Phenotypic Heterogeneity in Bacterial Populations. *Trends Genet.* **37**, 59–72 (2021).

4. J. Zieg, M. Silverman, M. Hilmen, M. Simon, Recombinational switch for gene expression. *Science*. **196**, 170–172 (1977).

5. F. W. Andrewes, Studies in group-agglutination I. The salmonella group and its antigenic structure. *J. Pathol. Bacteriol.* **25**, 505–521 (1922).

6. B. A. Stocker, Measurements of rate of mutation of flagellar antigenic phase in Salmonella typhi-murium. *J. Hyg.* . **47**, 398–413 (1949).

7. A. Goldberg, O. Fridman, I. Ronin, N. Q. Balaban, Systematic identification and quantification of phase variation in commensal and pathogenic Escherichia coli. *Genome Med.* **6**, 112 (2014).

8. O. Sekulovic, E. Mathias Garrett, J. Bourgeois, R. Tamayo, A. Shen, A. Camilli, Genome-wide detection of conservative site-specific recombination in bacteria. *PLoS Genet.* **14**, e1007332 (2018).

9. X. Jiang, A. B. Hall, T. D. Arthur, D. R. Plichta, C. T. Covington, M. Poyet, J. Crothers, P. L. Moses, A. C. Tolonen, H. Vlamakis, E. J. Alm, R. J. Xavier, Invertible promoters mediate bacterial phase variation, antibiotic resistance, and host adaptation in the gut. *Science*. **363**, 181–187 (2019).

10. T. Komano, Shufflons: multiple inversion systems and integrons. *Annu. Rev. Genet.* **33**, 171–191 (1999).

11. M. Chatzidaki-Livanis, M. J. Coyne, H. Roche-Hakansson, L. E. Comstock, Expression of a uniquely regulated extracellular polysaccharide confers a large-capsule phenotype to Bacteroides fragilis. *J. Bacteriol.* **190**, 1020–1026 (2008).

12. M. Taketani, M. S. Donia, A. N. Jacobson, J. D. Lambris, M. A. Fischbach, A Phase-Variable Surface Layer from the Gut Symbiont Bacteroides thetaiotaomicron. *MBio*. **6**, e01339–15 (2015).

13. E. B. Troy, V. J. Carey, D. L. Kasper, L. E. Comstock, Orientations of the Bacteroides fragilis capsular polysaccharide biosynthesis locus promoters during symbiosis and infection. *J. Bacteriol.* **192**, 5832–5836 (2010).

14. C. J. Severyn, B. A. Siranosian, S. T.-J. Kong, A. Moreno, M. M. Li, N. Chen, C. N. Duncan, S. P. Margossian, L. E. Lehmann, S. Sun, T. M. Andermann, O. Birbrayer, S. Silverstein, C. G. Reynolds, S. Kim, N. Banaei, J. Ritz, A. A. Fodor, W. B. London, A. S. Bhatt, J. S. Whangbo, Microbiota dynamics in a randomized trial of gut decontamination during allogeneic hematopoietic cell transplantation. *JCI Insight*. **7** (2022), doi:10.1172/jci.insight.154344.

15. B. A. Siranosian, E. F. Brooks, T. Andermann, A. R. Rezvani, N. Banaei, H. Tang, A. S. Bhatt, Rare transmission of commensal and pathogenic bacteria in the gut microbiome of hospitalized adults. *Nat. Commun.* **13**, 586 (2022).

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

17. E. C. Martens, R. Roth, J. E. Heuser, J. I. Gordon, Coordinate regulation of glycan degradation and polysaccharide capsule biosynthesis by a prominent human gut symbiont. *J. Biol. Chem.* **284**, 18445–18457 (2009).

18. C. M. Krinos, M. J. Coyne, K. G. Weinacht, A. O. Tzianabos, D. L. Kasper, L. E. Comstock, Extensive surface diversity of a commensal microorganism by multiple DNA inversions. *Nature*. **414**, 555–558 (2001).

19. N. T. Porter, A. J. Hryckowian, B. D. Merrill, J. J. Fuentes, J. O. Gardner, R. W. P. Glowacki, S. Singh, R. D. Crawford, E. S. Snitkin, J. L. Sonnenburg, E. C. Martens, Phase-variable capsular polysaccharides and lipoproteins modify bacteriophage susceptibility in Bacteroides thetaiotaomicron. *Nat Microbiol*. **5**, 1170–1181 (2020).

20. J. L. Round, S. M. Lee, J. Li, G. Tran, B. Jabri, T. A. Chatila, S. K. Mazmanian, The Toll-like receptor 2 pathway establishes colonization by a commensal of the human microbiota. *Science*. **332**, 974–977 (2011).

21. C. P. Neff, M. E. Rhodes, K. L. Arnolds, C. B. Collins, J. Donnelly, N. Nusbacher, P. Jedlicka, J. M. Schneider, M. D. McCarter, M. Shaffer, S. K. Mazmanian, B. E. Palmer, C. A. Lozupone, Diverse Intestinal Bacteria Contain Putative Zwitterionic Capsular Polysaccharides with Anti-inflammatory Properties. *Cell Host Microbe*. **20**, 535–547 (2016).

22. S. K. Mazmanian, C. H. Liu, A. O. Tzianabos, D. L. Kasper, An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell*. **122**, 107–118 (2005).

23. N. T. Porter, P. Canales, D. A. Peterson, E. C. Martens, 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).

24. O. Musumeci, A. L. Andreu, S. Shanske, N. Bresolin, G. P. Comi, R. Rothstein, E. A. Schon, S. DiMauro, Intragenic inversion of mtDNA: a new type of pathogenic mutation in a

patient with mitochondrial myopathy. *Am. J. Hum. Genet.* **66**, 1900–1904 (2000).

25. D. Hyatt, G.-L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, L. J. Hauser, Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. **11**, 119 (2010).

26. P. T. West, R. B. Chanin, A. S. Bhatt, From genome structure to function: insights into structural variation in microbiology. *Curr. Opin. Microbiol.* **69**, 102192 (2022).

27. M. C. Dao, A. Everard, J. Aron-Wisnewsky, N. Sokolovska, E. Prifti, E. O. Verger, B. D. Kayser, F. Levenez, J. Chilloux, L. Hoyles, MICRO-Obes Consortium, M.-E. Dumas, S. W. Rizkalla, J. Doré, P. D. Cani, K. Clément, Akkermansia muciniphila and improved metabolic health during a dietary intervention in obesity: relationship with gut microbiome richness and ecology. *Gut*. **65**, 426–436 (2016).

28. A. L. Greninger, K. Messacar, T. Dunnebacke, S. N. Naccache, S. Federman, J. Bouquet, D. Mirsky, Y. Nomura, S. Yagi, C. Glaser, M. Vollmer, C. A. Press, B. K. Kleinschmidt-DeMasters, S. R. Dominguez, C. Y. Chiu, Clinical metagenomic identification of Balamuthia mandrillaris encephalitis and assembly of the draft genome: the continuing case for reference genome sequencing. *Genome Med.* **7**, 113 (2015).

29. B. Becken, L. Davey, D. R. Middleton, K. D. Mueller, A. Sharma, Z. C. Holmes, E. Dallow, B. Remick, G. M. Barton, L. A. David, J. R. McCann, S. C. Armstrong, P. Malkus, R. H. Valdivia, Genotypic and Phenotypic Diversity among Human Isolates of Akkermansia muciniphila. *MBio*. **12** (2021), doi:10.1128/mBio.00478-21.

30. O. Milman, I. Yelin, R. Kishony, Systematic identification of gene-altering programmed inversions across the bacterial domain. *Nucleic Acids Res.* **51**, 553–573 (2023).

31. J. M. Atack, C. Guo, L. Yang, Y. Zhou, M. P. Jennings, 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).

32. L. Chen, N. Zhao, J. Cao, X. Liu, J. Xu, Y. Ma, Y. Yu, X. Zhang, W. Zhang, X. Guan, X. Yu, Z. Liu, Y. Fan, Y. Wang, F. Liang, D. Wang, L. Zhao, M. Song, J. Wang, Short- and long-read metagenomics expand individualized structural variations in gut microbiomes. *Nat. Commun.* **13**, 3175 (2022).

33. D. Maghini, M. Dvorak, A. Dahlen, M. Roos, S. Kuersten, A. S. Bhatt, Achieving quantitative and accurate measurement of the human gut microbiome. *bioRxiv* (2022), p. 2022.09.28.509972.

34. D. A. Rodionov, A. A. Arzamasov, M. S. Khoroshkin, S. N. Iablokov, S. A. Leyn, S. N. Peterson, P. S. Novichkov, A. L. Osterman, Micronutrient Requirements and Sharing Capabilities of the Human Gut Microbiome. *Front. Microbiol.* **10**, 1316 (2019).

35. V. Sharma, D. A. Rodionov, S. A. Leyn, D. Tran, S. N. Iablokov, H. Ding, D. A. Peterson, A. L. Osterman, S. N. Peterson, B-Vitamin Sharing Promotes Stability of Gut Microbial

Communities. *Front. Microbiol.* **10**, 1485 (2019).

36. T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. A. Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight, J. I. Gordon, Human gut microbiome viewed across age and geography. *Nature*. **486**, 222–227 (2012).

37. Z. A. Costliow, P. H. Degnan, Thiamine Acquisition Strategies Impact Metabolism and Competition in the Gut Microbe Bacteroides thetaiotaomicron. *mSystems*. **2** (2017), doi:10.1128/mSystems.00116-17.

38. N. C. Martinez-Gomez, D. M. Downs, 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).

39. H. M. Said, Intestinal absorption of water-soluble vitamins in health and disease. *Biochem. J*. **437**, 357–372 (2011).

40. G. D'Souza, S. Waschina, S. Pande, K. Bohl, C. Kaleta, C. Kost, Less is more: selective advantages can explain the prevalent loss of biosynthetic genes in bacteria. *Evolution*. **68**, 2559–2570 (2014).

41. E. Kussell, S. Leibler, Phenotypic diversity, population growth, and information in fluctuating environments. *Science*. **309**, 2075–2078 (2005).

42. M. K. Bacic, C. J. Smith, *Curr. Protoc. Microbiol.*, in press.

43. W. Zhu, M. G. Winter, L. Spiga, E. R. Hughes, R. Chanin, A. Mulgaonkar, J. Pennington, M. Maas, C. L. Behrendt, J. Kim, X. Sun, D. P. Beiting, L. V. Hooper, S. E. Winter, Xenosiderophore Utilization Promotes Bacteroides thetaiotaomicron Resilience during Colitis. *Cell Host Microbe*. **27**, 376–388.e8 (2020).

44. P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).

45. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. **34**, 3094–3100 (2018).

46. C. Yang, J. Chu, R. L. Warren, I. Birol, NanoSim: nanopore sequence read simulator based on statistical characterization. *Gigascience*. **6**, 1–6 (2017).

47. W. De Coster, S. D'Hert, D. T. Schultz, M. Cruts, C. Van Broeckhoven, NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*. **34**, 2666–2669 (2018).

48. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A.

Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature*. **596**, 583–589 (2021).

49. T. Aramaki, R. Blanc-Mathieu, H. Endo, K. Ohkubo, M. Kanehisa, S. Goto, H. Ogata, KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*. **36**, 2251–2252 (2020).

50. S. R. Eddy, Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

51. T. Paysan-Lafosse, M. Blum, S. Chuguransky, T. Grego, B. L. Pinto, G. A. Salazar, M. L. Bileschi, P. Bork, A. Bridge, L. Colwell, J. Gough, D. H. Haft, I. Letunić, A. Marchler-Bauer, H. Mi, D. A. Natale, C. A. Orengo, A. P. Pandurangan, C. Rivoire, C. J. A. Sigrist, I. Sillitoe, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, A. Bateman, InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427 (2023).

52. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).

53. A. Prjibelski, D. Antipov, D. Meleshko, A. Lapidus, A. Korobeynikov, Using SPAdes De Novo Assembler. *Curr. Protoc. Bioinformatics*. **70**, e102 (2020).

54. Y. Lin, J. Yuan, M. Kolmogorov, M. W. Shen, M. Chaisson, P. A. Pevzner, Assembly of long error-prone reads using de Bruijn graphs. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E8396–E8405 (2016).

55. D. E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

56. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. **22**, 1658–1659 (2006).

57. N. M. Koropatkin, E. C. Martens, J. I. Gordon, T. J. Smith, Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices. *Structure*. **16**, 1105–1115 (2008).

58. R. Simon, U. Priefer, A. Pühler, A Broad Host Range Mobilization System for In Vivo Genetic Engineering: Transposon Mutagenesis in Gram Negative Bacteria. *Biotechnology*. **1**, 784–791 (1983).

59. D. Pal, T. Venkova-Canova, P. Srivastava, D. K. Chattoraj, Multipartite regulation of rctB, the replication initiator gene of Vibrio cholerae chromosome II. *J. Bacteriol.* **187**, 7167–7175 (2005).

Acknowledgments: We thank Nora Enright, Danica Schmidtke, Aravind Natarajan, Jack Diaz Shanahan, Dylan Maghini, Mai Dvorak, Alvin Han, Meena Chakraborty, and Bhatt Lab members for helpful conversations and scientific advice regarding this project. We also thank Wenhan Zhu for plasmids (pNBU2_tet and pNBU2_erm) and Sebastian Winter for the strain (DH5α) that were used in the context of this project. We also thank Daniel Haft and Francoise Thibaud-Nissen at NCBI for helpful discussion about accessing SRA long-read datasets.

Funding:

National Institutes of Health R01 AI148623 (ASB)

National Institutes of Health R01 AI143757 (ASB)

Stand Up 2 Cancer Foundation

National Institutes of Health T32 training Grant HG000044 (RBC)

National Institutes of Health T32 training Grant HL120824 (PTW)

Author Contributions:

Conceptualization: RBC, PTW, ASB

Methodology: RBC, PTW, JW

Investigation: RBC, PTW, JW, RMP, GZMG, ASH, MOG, EFB, AMM

Visualization: RBC, PTW, JW, RMP

Funding acquisition: ASB

Project administration: RBC, PTW, ASB

Supervision: RBC, PTW, ASB

Writing – original draft: RBC, PTW, ASB

Writing – review & editing: RBC, PTW, JW, RMP, GZMG, ASH, MOG, EFB, AMM, ASB

Competing Interests: Authors declare that they have no competing interests.

Data and materials availability: PhaVa is available at (https://github.com/patrickwest/PhaVa). 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 will be made available after publication 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 S3.

Supplementary Materials

Materials and Methods

Figs. S1 to S7

Tables S1 to S2

Data S1 to S6

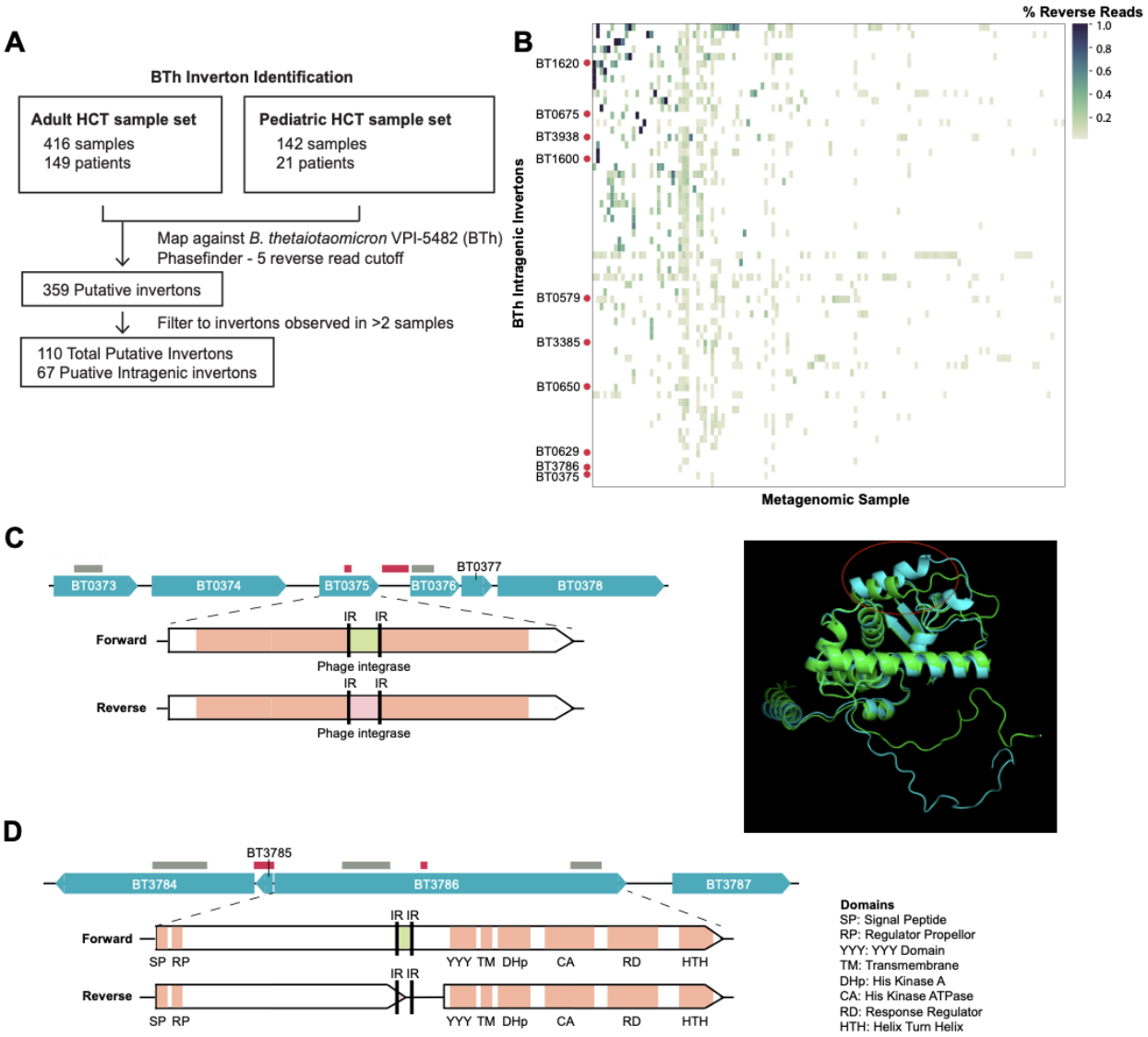

Fig. 1. Short-read metagenomic datasets reveal intragenic invertons in BTh. (**A**) An overview of the analysis pipeline for identifying putative invertons in short-read datasets. (**B**) A heatmap of the inversion proportion of intragenic invertons in BTh. Samples with no intragenic invertons were removed. Rows labeled with a gene name represent intragenic invertons with PCR and sanger sequencing evidence of inversion. (**C-D**) Genome diagrams for confirmed intragenic invertons in BTh. Gray bars indicate putative invertons without sequencing support. Red bars indicate invertons with sequencing evidence. (**C**) Left - Genome diagram of the region surrounding the BT0375 recoding intragenic inverton, and a domain diagram of the BT0375 gene with the location of the inverton IRs indicated. Right - AlphaFold overlay of the BT0375 forward (blue) and reverse (green). The region that is recoded is circled in red. (**D**) Genome diagram of the region surrounding the BT3786 premature stop codon intragenic inverton, and a

domain diagram of the gene. The consequence of the inversion and resulting two predicted ORFs
500 are indicated in the domain diagram.

Source	Gene	Annotation	Consequence
Short-read	BT0375	integrase	recoding
Short-read	BT0579	putative transcription regulator	premature stop codon
Short-read	BT0629	Mn2+ and Fe2+ transport protein	premature stop codon
Short-read	BT0650	thiamine biosynthesis protein ThiC	premature stop codon
Short-read	BT0675	N-acetylglucosamine-6-phosphate deacetylase	recoding
Short-read	BT1600	BexA, membrane protein	premature stop codon
Short-read	BT1620	SusD homolog	premature stop codon
Long-read	BT1654	polysialic acid transport protein kpsD	recoding
Long-read	BT2981	hypothetical protein	premature stop codon
Short-read	BT3385	putative helicase	premature stop codon
Long-read	BT3415	hypothetical protein	premature stop codon
Short-read	BT3786	two-component system sensor histidine kinase/response regulator, hybrid (one-component system)	premature stop codon
Long-read	BT3818	hypothetical protein	recoding
Short-read	BT3843	ATP-dependent protease; ATP-binding subunit ClpX	recoding
Short-read	BT3938	ATP-dependent DNA helicase recQ	premature stop codon

Table 1. Confirmed intragenic invertons BTh. Intragenic invertons confirmed *in vitro* in BTh are listed. Invertons from short-read datasets were called with PhaseFinder on metagenomic samples (see Fig.1). Invertons from long-read datasets were called with PhaVa using isolate sequences. The predicted consequence of inversion is also listed.

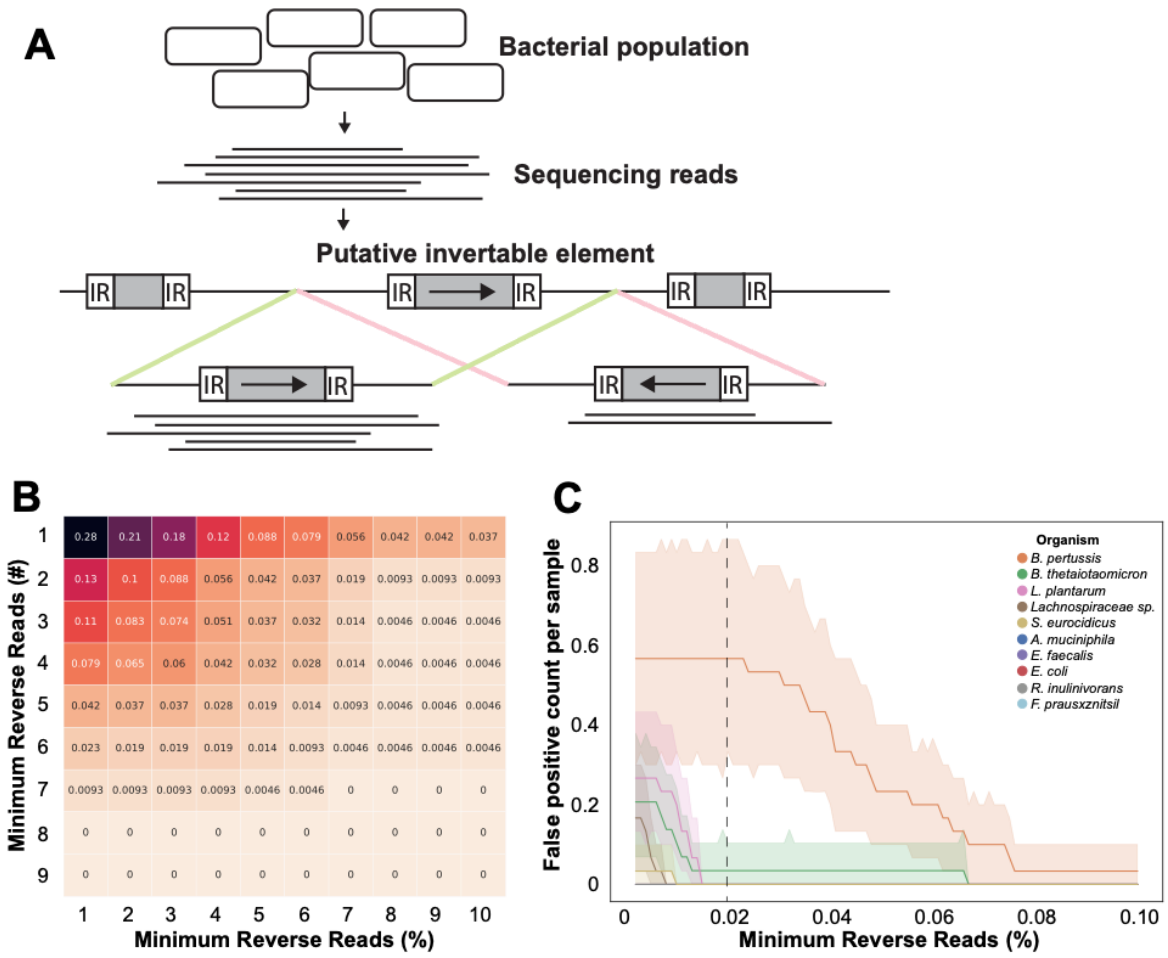

Fig. 2. Developing and optimizing PhaVa, a long-read based, accurate inverton caller. (**A**) Schematic of PhaVa's workflow. Putative invertons are identified, and long reads are mapped to both a forward (highlighted by the green lines) and reverse orientation (highlighted by the pink lines) version of the inverton and surrounding genomic sequence, similar to PhaseFinder. 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**) 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 the false positive rate per simulated readset. (**C**) 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.

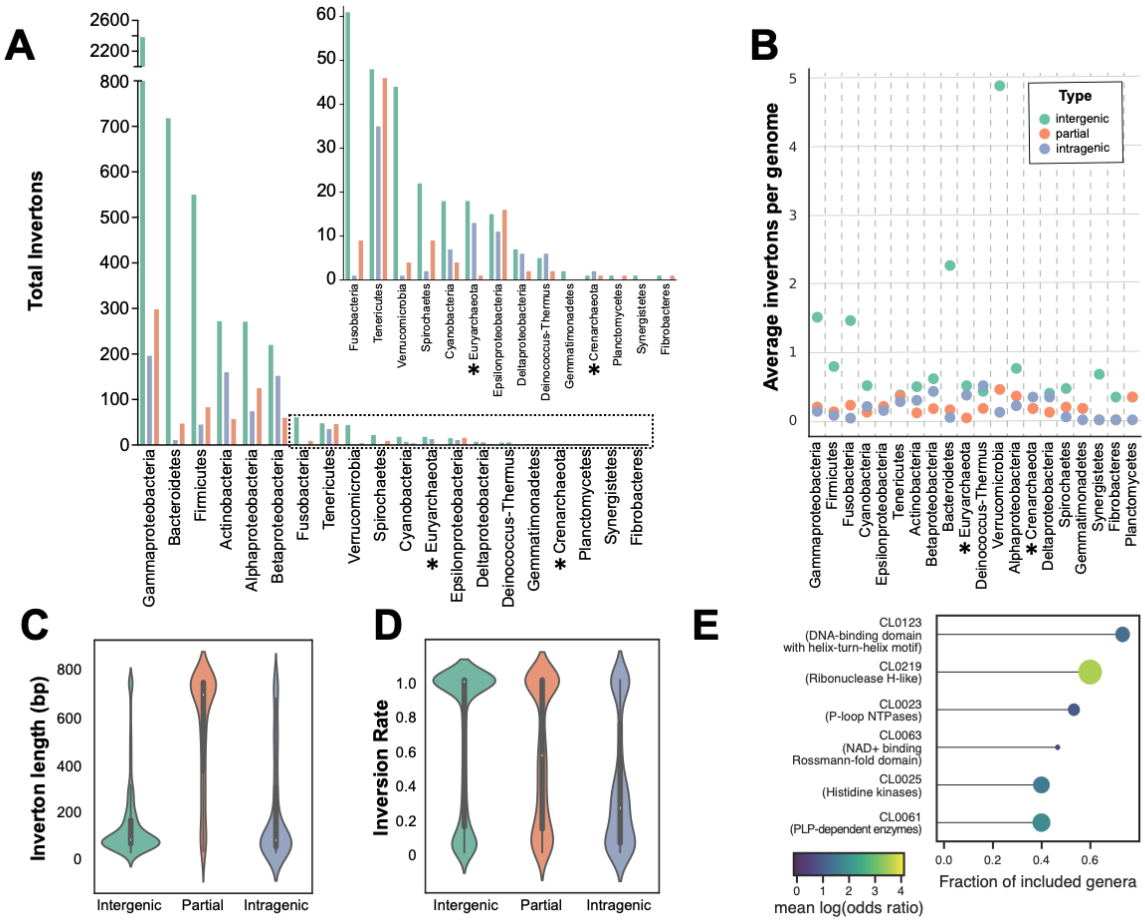

Fig. 3. Public isolate sequencing data reveals intragenic inversions are prevalent across the bacterial tree of life. (**A**) The total number of invertons found within various bacterial phyla from 29,989 long-read isolate sequencing samples. Green bars refer to intergenic invertons. Orange bars refer to partial intergenic invertons. Blue bars refer to intragenic invertons. Asterisks denote phyla within Archaea. Inset corresponds to the portion of the bar graph outlined in dotted lines. (**B**) The mean number of invertons found per genome within a phylum, of genomes that had at least one inverton. Asterisks denote phyla within Archaea. (**C**) The distribution of lengths of identified invertons, grouped by inverton type (intergenic, partial intergenic - denoted 'partial', and intragenic). Median value is indicated by white dots. Partial length distribution was found to be significantly different from intergenic (p=0.0) and intragenic (p=4.5e-146) with a 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 white dots. (**E**) Pfam clan enrichment across several genera. Dot size and fill color is proportional to the mean log-odds ratio, an effect size measure for the enrichment, and the length of the line indicates the fraction of included genera in which an enrichment score for the specific clan could be calculated.

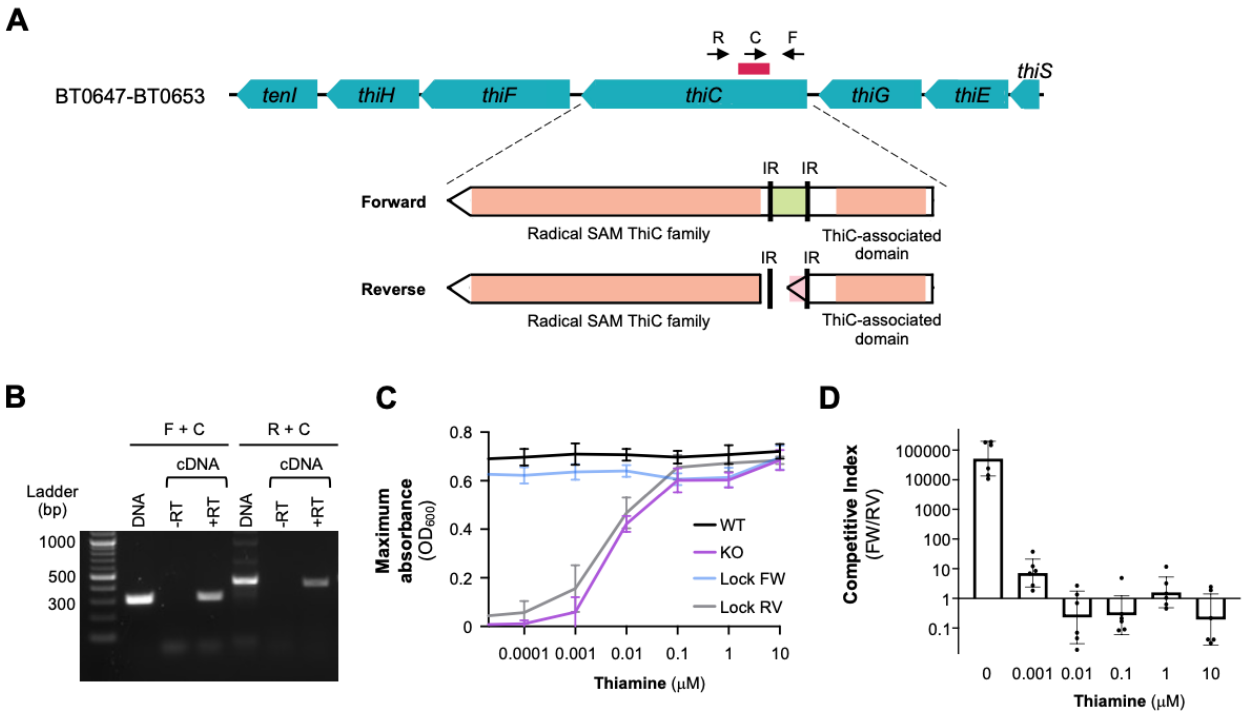

Fig. 4. Consequences of inversion in thiamine biosynthesis protein (*thiC*). (**A**) Schematic showing the location of the *thiC* intragenic inverton (red bar). Inverton flipping results in a premature stop codon located between domains. Black arrows indicate the binding location of primers used to determine the orientation of inverton. (**B**) PCR confirmation of the *thiC* intragenic inverton in both genomic DNA and reverse transcribed RNA (cDNA). Bands were extracted and confirmed with Sanger sequencing. (**C**) Locked forward (blue line), locked reverse (gray line), Δ*thiC* (purple line), and wild-type (black line). BTh strains were grown in defined media with various concentrations of thiamine. The maximum optical density each strain reached was recorded. Each point represents the average of six replicates conducted across two separate experiments. Mean and standard deviation are shown. (**D**) Locked strains were competed against each other in thiamine-containing media. The competitive index was determined. Geometric mean and geometric standard deviation are shown for 6 replicates across 3 independent experiments.