Published online 9 January 2023

https://doi.org/10.1093/nar/gkac1166

Systematic identification of gene-altering programmed inversions across the bacterial domain

Oren Milman[1], Idan Yelin[1] and Roy Kishony[1,2,3,*]

[1]Faculty of Biology, Technion–Israel Institute of Technology, Haifa, Israel, [2]Faculty of Computer Science, Technion–Israel Institute of Technology, Haifa, Israel and [3]Faculty of Biomedical Engineering, Technion–Israel Institute of Technology, Haifa, Israel

Received August 01, 2022; Revised October 22, 2022; Editorial Decision November 11, 2022; Accepted January 05, 2023

ABSTRACT

Programmed chromosomal inversions allow bacteria to generate intra-population genotypic and functional heterogeneity, a bet-hedging strategy important in changing environments. Some programmed inversions modify coding sequences, producing different alleles in several gene families, most notably in specificity-determining genes such as Type I restriction-modification systems, where systematic searches revealed cross phylum abundance. Yet, a broad, gene-independent, systematic search for gene-altering programmed inversions has been absent, and little is known about their genomic sequence attributes and prevalence across gene families. Here, identifying intra-species variation in genomes of over 35 000 species, we develop a predictive model of gene-altering inversions, revealing key attributes of their genomic sequence attributes, including gene-pseudogene size asymmetry and orientation bias. The model predicted over 11,000 gene-altering loci covering known targeted gene families, as well as novel targeted families including Type II restriction-modification systems, a protein of unknown function, and a fusion-protein containing conjugative-pilus and phage tail domains. Publicly available long-read sequencing datasets validated representatives of these newly predicted inversion-targeted gene families, confirming intra-population genetic heterogeneity. Together, these results reveal gene-altering programmed inversions as a key strategy adopted across the bacterial domain, and highlight programmed inversions that modify Type II restriction-modification systems as a possible new mechanism for maintaining intra-population heterogeneity.

INTRODUCTION

Phase variation is a process that generates intra-population phenotypic heterogeneity in bacteria. As different phenotypes are often better equipped to overcome different challenges, such intra-population heterogeneity might allow the bacterial population a bet-hedging strategy to better survive sudden environmental challenges. Indeed, phase variation was observed in various bacterial processes (1–4) and was shown to be important for survival in major environmental challenges faced by bacteria, including bacteriophages (5,6), antibiotic drugs (7,8) and virulence (9–11).

The underlying mechanism of many phase variation systems is programmed chromosomal inversions (1,12). Programmed inversions are frequent inversions surgically targeted to specific genomic regions flanked by inverted repeats (12). Programmed inversions are catalyzed by recombinase enzymes, usually encoded near or inside the invertible region (12,13). Yet some invertible regions lacking a nearby recombinase gene were also observed (14,15), as well as programmed inversions that were not completely diminished upon deletion of the nearby recombinase gene (16). Programmed inversions often modify regulatory DNA sequences, typically inverting a promoter to switch a gene on or off (12). Conversely, gene-altering programmed inversions target coding sequences (Figure 1A), producing different alleles, encoding for different protein variants (12).

The most studied gene-altering programmed inversion targets are Type I restriction-modification (RM) enzymes, whose phase variants exhibit different sequence specificity for RM activity (10,17–19). This diversity inherently serves as a bet-hedging strategy in the face of multiple potential invading bacteriophage strains, forbidding any one bacteriophage from eliminating the entire bacterial population (6,20). These variants were also observed to differ in opacity (10), virulence in mice (10), as well as human cell infection efficiency and survival in human blood (11), presumably resulting from Type I RM phase variants inducing differential methylation of host DNA (20). Other interesting phase variation systems include a programmed inversion targeting the *pilV* gene (shufflon), leading to phase variants differing

[*]To whom correspondence should be addressed. Tel: +972 4 8293737; Email: rkishony@technion.ac.il

in the specificity of their conjugative pili (21), and a programmed inversion targeting phage tail genes, conferring different host specificity (22).

The genomic signature of inverted repeats flanking programmed inversions, as well as the rapidly growing amount of publicly available DNA sequence data, provide an opportunity for computational identification of programmed inversions. Indeed, multiple methods to identify programmed inversions were developed and successfully deployed (7,8,12,17,18,23–27). Yet, only a few of these methods were applied widely across the bacterial domain, specifically searching for programmed inversions that target promoters (7,8), or searching for gene-altering programmed inversions targeting a specific gene family, the Type I RM specificity subunit HsdS (17,18). A more comprehensive gene-independent search for programmed inversions was developed based on abnormally aligned short-reads and manual curation, and was used to identify programmed inversions, regardless of target, in over 200 genomes (27). The approach though requires short-read data and some manual curation and is difficult thereby to apply more broadly. Thus, a wide gene-family independent search for gene-altering programmed inversions is still lacking.

The lack of a systematic search for gene-altering programmed inversions leaves some fundamental aspects of such programmed inversions underexplored. First, other than the presence of inverted repeats (12) and nearby recombinase genes (12,13), the genomic sequence attributes of gene-altering programmed inversions are uncharacterized. Second, it is unknown which gene families are targeted by programmed inversions, and what their genomic contexts and abundance across species are. Revealing such gene families and genomic contexts might highlight central bacterial pathways and environmental challenges.

Here, seeking to shed light on these aspects, we computationally and systematically scan over 35 000 bacterial species for gene-altering programmed inversions. We start by identifying candidates for gene-altering programmed inversions, using inverted repeats and annotated coding sequence locations. Then, for each such candidate locus, we search for intra-species variation, producing a diverse dataset of 128 putative gene-altering programmed inversions. Next, we identify genomic sequence attributes enriched in these putative programmed inversions, and utilize these genomic signatures to identify a large and diverse dataset of 11 955 predicted programmed inversions, revealing associated gene families. Finally, we find in publicly available long-read genome sequencing data evidence for programmed inversions in selected loci representing different predicted target gene families. For three programmed inversion loci, we further identify in publicly available RNA-seq data evidence for expression of different gene variants. This analysis identifies known programmed inversion loci as well as previously unknown loci, including programmed inversions targeting a gene coding for a protein of unknown function, a presumable PilV and phage tail collar fusion-gene, and various Type II RM genes across multiple phyla, highlighting the Type II RM family as a major target of gene-altering programmed inversions.

MATERIALS AND METHODS

Retrieval of representative genomes for each of 35 366 species

The bacterial NCBI RefSeq assembly summary file (ftp://ftp.ncbi.nih.gov/genomes/refseq/bacteria/assembly_summary.txt, retrieved on 10 January 2022) was filtered to include only assembly entries with *version_status* = 'latest' and *genome_rep* = 'Full'. For each species, genomes were sorted by *assembly_level* ('Complete Genome', 'Chromosome', 'Scaffold' and 'Contig') and then by *refseq_category* ('reference genome', 'representative genome', 'na'), and the first genome was chosen as the species representative. Taxonomy of each species was retrieved from the NCBI Taxonomy Database using Biopython version 1.78 Entrez.efetch. The GenBank file of each representative genome was downloaded using ncbi-genome-download version 0.3.1 (https://github.com/kblin/ncbi-genome-download). A list of the 35 366 species for which representative genomes were retrieved is provided in Supplementary Table S1. Non-continuous GenBank Coding Sequences (CDSs) were discarded if the total distance between parts was >100 bp.

Identification of CDS pairs with inverted repeats

For each scaffold in each representative genome, BLASTN version 2.12 (28) was run locally to find alignments between sequences in the scaffold, using the following arguments:

- *strand minus -ungapped -word_size 20 -evalue 1000 -window_size 0 -dust no*

Alignments were filtered to include only inverted repeats with length ≥22 bp and flanking regions of length ≤15 kb. Inverted repeats strictly contained within other inverted repeats were discarded. Each repeat was then linked to a CDS if it was either strictly contained in one, appeared immediately upstream to it, or overlapped its start codon but did not contain the CDS. Finally, pairs were filtered to include only those in which both repeats were linked to CDSs on opposing strands with at least one of them strictly contained in its linked CDS, producing 196 653 CDS pairs linked to one or more inverted repeat pairs.

Discarding CDS pairs containing repetitive inverted-repeats

To avoid repetitive sequences, such as transposons, for each pair of inverted repeats, one repeat was arbitrarily chosen and blasted against its genome using the following arguments (BLASTN version 2.12):

- *strand both -ungapped -word_size 15 -evalue 1e-05 -window_size 0 -dust no*

Alignments were filtered to include only alignments to sequences in other scaffolds or sequences at least 50kb away from the inverted repeat pair. If any base pair in the repeat was part of three or more alignments (Supplementary Figure S1), the inverted repeat pair was marked as repetitive, and CDS pairs with any CDS strictly containing a repeat of a repetitive inverted repeat pair were discarded. Remaining

120 686 CDS pairs are hereafter referred to as programmed inversion candidates (PICs).

Same-species genome choice and retrieval

The BLAST nt database was downloaded on 17 January 2022 from ftp://ftp.ncbi.nlm.nih.gov/blast/db. For each species with any PIC, BLASTN version 2.12 get_species_taxids was run, and at most 100 of the first returned NCBI Taxonomy IDs, including the species Taxonomy ID, were chosen for subsequent queries of the BLAST nt database.

In addition, for each species, NCBI Nucleotide IDs of at most 500 longest WGS Nucleotide entries (belonging to this species) were retrieved from the NCBI Nucleotide Database using Biopython version 1.78 Entrez.esearch with the following search term:

(txid ⟨ species_taxonomy_id⟩[orgn:exp] AND ‘wgs’[properties] AND (‘40000’[SLEN] :‘100000000’[SLEN])) NOT ‘wgs master’[properties]

NCBI Nucleotide accessions of chosen WGS Nucleotide entries were retrieved from the NCBI Nucleotide Database using Biopython version 1.78 Entrez.esummary. These Nucleotide accessions were then used to download chosen WGS Nucleotide entries using ncbi-acc-download version 0.2.8.

Identification of similar length loci in same-species genomes

To identify within species variation in our collection of PICs, we set up a pipeline for identifying and comparing homologous regions of these PICs in a large dataset of genomes of each species. First, overlapping PICs were merged to form ‘PIC loci’ ($n = 108\ 940$). Second, for each such PIC locus, we define its genomic region as the region that contains its CDS pairs and inverted repeats. Third, for each such PIC locus region, we define the left and right margins as the flanking 200bp genomic regions, thereby defining the genomic context for the locus. PIC loci with partial left or right margins (due to proximity to scaffold edge), or with margins containing bases other than A/C/G/T, were excluded from further searches for intra-species variation. Fourth, for each PIC locus, we blast the left and right margin (BLASTN version 2.12) against same-species genomes both from the BLAST nt database and from WGS Nucleotide entries (see above). For blasting against WGS genomes, we used:

- *strand both -ungapped -word_size 20 -evalue 1e-05 -window_size 0 -dust no*

and for the BLAST nt database search, we used the same arguments as well as *-taxids* to specify chosen taxa (see above). Left and right margin alignments to the same scaffold and strand were paired to form alignment pairs. For each alignment pair, we identified the margin-spanning region, the scaffold region spanning the alignment pair. As homologous sequences should be of similar relative length, we discarded margin-spanning regions whose length differed from the length of the PIC locus by more than 5% (Supplementary Figure S2). This produced a set of similar length loci in same-species genomes for each PIC locus.

Identification of within-species rearrangements, indicating putative programmed inversions

To identify PIC loci with within-species rearrangements, we compared each PIC locus with all of its same-species similar length loci, and searched for cases showing genomic rearrangements. First, for each PIC locus, we discarded same-species similar length loci perfectly identical to the PIC locus. Second, remaining similar length loci were sorted by relative locus length difference (smallest first) and were each (100 at maximum) aligned to the PIC locus using progressiveMauve (build date 13 February 2015) (29). For each such alignment, we defined the match proportion as the proportion of matching base pairs for the longer aligned sequence. Loci whose alignment to the PIC locus had a match proportion smaller than 0.95 were discarded (Supplementary Figure S3). Furthermore, loci such that any sub-alignment (in the alignment to the PIC locus) had a match proportion smaller than 0.95 were also discarded (Supplementary Figure S4). Remaining loci are hereafter referred to as homologous loci (a list of homologous loci is provided in Supplementary Table S2). Out of 120 686 PICs, 6372 had at least one homologous locus.

To identify rearrangements, for each homologous locus, sub-alignments were sorted according to the aligned region location in the PIC locus, and the region between each two consecutive sub-alignments was marked as a breakpoint-containing region in case: (a) the consecutive sub-alignments matched homologous locus regions on different strands; or (b) The consecutive sub-alignments were not consecutive if sorted according to the aligned region location in the homologous locus.

Finally, each pair of inverted repeats was examined to determine whether for any homologous locus, each repeat is at most 10 bp away from a different breakpoint-containing region of that homologous locus (Supplementary Figure S5). 128 PICs linked to any such inverted repeat pair were identified and marked as PICs with intra-species variation (a list of PICs for which intra-species variation was identified is provided in Supplementary Table S3).

Programmed inversion candidate clustering

To avoid counting the same PIC more than once in statistical analyses (described below), CDSs of all PICs were clustered by vsearch version 2.17.1 (30), using the following arguments:

–id 0.95 –iddef 1 –strand plus –minseqlength < shortest_CDS_length> –maxseqlength < longest_CDS_length> –qmask none

PICs whose CDSs belong to the same set of clusters (e.g. the left and right CDSs of PIC A belong to clusters α and β, respectively, and the left and right CDSs of PIC B belong to clusters β and α, respectively) were considered to be of the same PIC cluster.

Genomic sequence attribute definitions

Operons were predicted by grouping consecutive CDSs on the same strand such that the maximal distance between consecutive CDSs was 20bp. In the following definitions, for simplicity, we refer to CDSs not predicted to be part of any operon as belonging to single-CDS operons.

We define o_L ('outer') to be the length of the part of the left operon left to the leftmost repeat. Mirroring o_L, we define o_R to be the length of the part of the right operon right to the rightmost repeat. Similarly, we define i_L ('inner') to be the length of the part of the left operon right to the rightmost left repeat. Mirroring i_L, we define i_R to be the length of the part of the right operon left to the leftmost right repeat. Furthermore, we define u_L ('upstream') to be the length of the part of the left operon upstream to all repeats linked to the PIC. u_R is defined similarly for the right operon (Figure 2A).

We define four genomic sequence attribute measures of a PIC: (a) *repeat length* = length of the longest repeat linked to the PIC; (b) *CDS distance* = length of the region flanked by the PIC operons; (c) *orientation matching* = $\frac{o_L+CDS\ distance+o_R}{(o_L+CDS\ distance+o_R)+(i_L+CDS\ distance+i_R)}$, a measure of matching between the orientation of the two operons (head-to-head or tail-to-tail) and repeat positions inside operons, while we consider a matching to be better (i.e. higher orientation matching) in case switching the operon orientations would result in a locus in which the shortest region flanked by inverted repeats is longer (Supplementary Figure S6); and (d) *asymmetry* $= 1 - \frac{min(u_L,u_R)}{max(u_L,u_R)}$, a measure of asymmetry between the regions in the two operons upstream to all repeats (Supplementary Figure S7).

Genomic sequence attribute enrichment analysis

Out of PICs with at least one homologous locus, for each PIC cluster (see 'Programmed inversion candidate clustering'), one PIC was chosen randomly as the cluster representative. All subsequent steps for identifying gene-altering programmed inversion genomic sequence attributes, as well as training the logistic regression model (detailed below) were performed using only these PIC cluster representatives.

PIC cluster representatives were split into two groups: PICs with and without intra-species variation. For each of the four genomic sequence attribute measures (*repeat length*, *CDS distance*, *orientation matching, asymmetry*), a two-sided Mann–Whitney *U* test was performed using Python's scipy.stats.mannwhitneyu to calculate a *P*-value. Furthermore, the value corresponding to the Kolmogorov-Smirnov-statistic was chosen as a threshold to binarize the genomic sequence attribute measures, giving rise to four genomic sequence attributes: long repeats (*repeat length* $\geq$ 44 bp), short CDS distance (*CDS distance* $\leq$ 2588 bp), high orientation matching (*orientation matching* $\geq$ 0.58), and high asymmetry (*asymmetry* $\geq$ 0.87).

Gene-altering programmed inversion prediction

A logistic regression model was trained on PIC cluster representatives (see 'Programmed inversion candidate clustering'), using Python's statsmodels.api.Logit.fit. For each PIC, the model was provided with the four genomic sequence attributes (high repeat length, low CDS distance, high orientation matching, and high asymmetry) as binary predictors, a constant predictor, and whether intra-species variation was identified as the binary response variable.

To obtain logistic regression coefficients for each predictor alone, namely, unadjusted models, a similar method was used for each predictor: the model was provided with two binary predictors - the predictor of interest and a constant predictor, as well as whether intra-species variation was identified as the binary response variable.

To assess the performance of the adjusted model, cross validation was used (Supplementary Figure S8). In each of 500 simulations, 20% of the PIC cluster representatives were randomly chosen to form a testing set, while the rest 80% formed a training set, which was used to train the model. Then, the trained model was applied to the testing set, and true and false positive rates were calculated. This provided a receiver operating characteristic (ROC) curve, and the area under the ROC curve (AUC) was calculated using Python's scipy.integrate.trapz. AUC values obtained from 500 simulations had a mean of 0.91 and a standard deviation of 0.025.

Finally, the trained adjusted model was used to predict gene-altering programmed inversions for all PICs, using Python's statsmodels.discrete.discrete_model.BinaryResults.predict. PICs with programmed inversion prediction probability >0.05 were marked as predicted programmed inversions (Supplementary Figure S8; a list of all PICs, including genomic sequence attribute measures and predicted probability, is provided in Supplementary Table S4).

Enrichment analysis of gene families associated with predicted programmed inversions

For the following enrichment analyses, a random PIC cluster representative was chosen for each PIC cluster (see 'Programmed inversion candidate clustering'). Product annotations ('/product' qualifier in GenBank file) of CDSs linked to PICs, namely, target genes, as well as their neighbor CDSs (at most four neighbor CDSs per PIC), were extracted from corresponding GenBank files. Next, for each annotation that appears as the product annotation of a target gene in at least 12 PICs, a two-sided Fisher's exact test or G-test was performed to assess enrichment for PICs containing any target CDS with this annotation among predicted programmed inversions. Fisher's exact test or G-test was performed using Python's scipy.stats.fisher_exact or scipy.stats.chi2_contingency (with *lambda_* = 'log-likelihood'), respectively. A G-test was performed in case all expected frequencies were at least 5; otherwise, a Fisher's exact test was performed. A Bonferroni correction was applied in order to obtain corrected p-values. Similarly, for each annotation that appears as the product annotation of a target gene neighbor in at least 12 PICs, a two-sided Fisher's exact test or G-test was performed to assess enrichment for PICs containing any target neighbor CDS (which is not a target CDS) with this annotation among predicted programmed inversions. A Bonferroni correction was used to obtain corrected p-values (for both types of statistical tests, lists of product annotations and their corresponding odds ratio and p-values are provided in Supplementary Table S5 and Supplementary Table S6).

Choice and retrieval of genomic context representatives and corresponding long-read genome sequencing data

Predicted programmed inversions targeting gene families found to be significantly enriched (see above) were scanned manually for recurring genomic contexts, namely, locus gene content and organization. For some loci containing such genomic contexts, the NCBI Nucleotide and SRA Databases were manually searched for Nucleotide entries that both contain similar genomic contexts and have matching long-read SRA entries. Found Nucleotide entries were downloaded using ncbi-acc-download version 0.2.8, and their linked NCBI Assembly entries were downloaded using ncbi-genome-download version 0.3.1. Found SRA entries were downloaded from NCBI (using the download link provided in https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=⟨SRA_entry_id⟩), and fastq-dump version 2.10.0 was run to extract reads into fasta files with the following arguments:

–skip-technical –readids –read-filter pass –dumpbase –split-spot –clip

Variant identification in long-read genome sequencing data

For each locus (for which NCBI Nucleotide, SRA and Assembly entries were found and retrieved), a region hypothesized to contain programmed inversions was assigned manually. Then, inverted repeats in this region were identified using BLASTN version 2.12 with the following arguments:

- *strand minus -ungapped -word_size 7 -evalue 1000 -window_size 0 -dust no*

Found alignments were filtered to include only inverted repeats such that the length of the region flanked by the repeats is greater than zero and repeats are $>=$ 14 bp long.

Next, the region hypothesized to contain programmed inversions, including margins of at least 10kbp on each side, was blasted against extracted long reads with the following arguments (BLASTN version 2.12):

- *strand both -word_size 8 -evalue 0.0001 -window_size 0 -dust no*

Reads were filtered to include only reads with at least 2 kb of alignment to the genome, and alignments strictly contained in other alignments were discarded. Finally, remaining alignments of each read were manually examined for rearrangements that may result from a single or multiple chromosomal inversions, such that inverted regions are flanked by inverted repeats (identified previously, see above). In some cases, nested close pairs of inverted repeats seemed equally suitable to be identified as flanking the identified chromosomal inversions; such inverted repeat pairs were merged to form longer inverted repeats. Thus, for each locus, a list of inverted repeats (presumably) promoting programmed inversions was compiled, and the locus was marked as a programmed inversion locus (a list of programmed inversion loci and corresponding inverted repeats of each variant is provided in Supplementary Table S7).

Distribution of variants in long reads

For each programmed inversion locus, we define the inverted repeat region to be the shortest region that contains all regions flanked by inverted repeats (presumably) promoting programmed inversions. Long reads were now further filtered to include only reads with alignments that together completely cover the inverted repeat region, as well as 500 bp flanking all repeats, namely, 500 bp left to the leftmost repeat or right to the rightmost repeat. It was further required that this region of the programmed inversion locus was covered by alignments of a continuous region in the read. Of these reads, those with collinear alignments that together completely cover that region were marked as matching the reference variant, while the rest of the reads were marked as matching a non-reference variant. These non-reference variant reads were manually assigned to different variants, according to the associated inverted repeats, except for a few reads that were discarded due to an anomaly revealed during manual examination (a list of these discarded reads is provided in Supplementary Table S8). Finally, each remaining read (both reference and non-reference variant reads) was truncated to keep the smallest region in the read that contained all bases that were aligned to the region hypothesized to contain the programmed inversion (including margins). These truncated reads were blasted to the whole reference genome (that is, the NCBI Assembly entry linked to the Nucleotide entry containing the locus), with the following arguments (BLASTN version 2.12):

- *strand both -word_size 8 -evalue 0.0001 -window_size 0 -dust no*

Found alignments overlapping the inverted repeat region were discarded, and for each read, the number of read base pairs covered by the remaining alignments was compared to the number of read base pairs covered by the initial alignment to the region hypothesized to contain the programmed inversion (including margins). Reads were filtered to include only reads such that the initial alignment covered more read base pairs. In other words, reads were discarded if they matched another region in the genome better than they matched the programmed inversion locus. Remaining reads, with their assignment to the reference variant or another variant (defined by its inverted repeats), were counted and plotted in Figure 4A, C and Supplementary Figures S9–S29 (a list of these reads is provided in Supplementary Table S9).

Visualization of variants identified in long reads

For each identified variant of each programmed inversion locus, a representative long-read was chosen manually. Then, either the truncated read (see above) or its reverse complement was aligned to the region hypothesized to contain the programmed inversion, using progressive-Mauve (build date 13 February 2015). Finally, matching positions were extracted from the mauve .xmfa file, and were plotted in Figure 4A and Supplementary Figures S9–S29, after subtracting a constant number from the read position, so that (relative) positions in read would always start from 1.

Gene family assignment

Statistically significantly enriched CDS product annotations and CDS product annotations of genes in programmed inversion loci (see above), were inspected manually and assigned to broad gene families (a list of product annotations with manually assigned gene families is provided in Supplementary Table S10). In addition, CDSs in programmed inversion loci were scanned manually, and some specific CDSs were assigned gene families, either according to protein predicted conserved domains (obtained from NCBI Conserved Domain Database (31)), or according to protein sequence similarity (assessed by BLAST).

Distribution of programmed inversions across phyla

For each programmed inversion locus, the longest target CDS was blasted against each of the 35,366 representative genomes, using the following arguments (BLASTN version 2.12):

- *strand both -word_size 10 -evalue 1e-05 -window_size 0 -dust no*

Alignments covering <0.5 of the longest target CDS were discarded (Supplementary Figure S30), with remaining alignments revealing homologous sequences to the longest target CDS of the programmed inversion locus. Each such homologous sequence was linked to an overlapping CDS; if more than one such CDS existed, the CDS with the longest overlapping region was chosen. Homologous sequences with linked CDS covering <0.9 of their base pairs were discarded (Supplementary Figure S31). CDSs linked to remaining homologous sequences were filtered to include only those that are at least 10kbp away from the scaffold edge on each side. Remaining CDSs were marked as homologs of the longest target CDS of the programmed inversion locus. Next, each of these homologs was blasted against the two 10kbp regions flanking it, using the following arguments (BLASTN version 2.12):

- *strand minus -ungapped -word_size 14 -evalue 0.0001 -window_size 0 -dust no*

Each homolog with any alignment to the opposite strand with evalue ≤1e−6 was marked as a homolog potentially targeted by programmed inversions (a list of homologs is provided in Supplementary Table S11).

Choice and retrieval of RNA sequencing data

For each long-read confirmed programmed inversion locus whose target gene family is not a well-known programmed inversion target (see 'Variant identification in long-read genome sequencing data', and Figure 4B), the NCBI SRA Database was manually searched for paired-read RNA sequencing experiments of the corresponding species. Found SRA entries were downloaded from NCBI (using the download link provided in https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=⟨SRA_entry_id⟩), and fastq-dump version 2.10.0 was run to extract reads into fasta files with the following arguments:

- *skip-technical –readids –read-filter pass –dumpbase –split-3 –clip*

Programmed inversion identification in RNA sequencing data

For each retrieved SRA entry (see above), bowtie2 (32) version 2.2.5 was run to align the reads to the corresponding reference genome, which contains one of the long-read confirmed programmed inversion loci, using the following arguments:

- *end-to-end –no-unal –no-mixed –no-discordant*

Then, for each pair of inverted repeats in the programmed inversion locus (identified by BLASTN, see 'Variant identification in long-read genome sequencing data') with evalue <0.001, the region flanked by inverted repeats was inverted *in silico*, producing an 'alternative reference genome', identical to the publicly available reference genome, except for the inverted region. bowtie2 was used again, using the same arguments, to align the reads to the alternative reference genome. Finally, non-primary alignments were discarded (according to the flag in bowtie2 .sam output), and reads were filtered to keep only those whose alignment scores to the reference and alternative reference differed substantially (a difference of at least 100 in the alignment scores). The alignment score of each read pair was considered to be the sum of alignment scores reported by bowtie2 for the two paired reads (in case no concordant alignment was identified, the score of the read pair was considered to be minus infinity). The high threshold, namely 100, for difference in alignment score was chosen so that identified reads would strongly support one orientation of the region flanked by inverted repeats, over the other orientation. Thus, in case at least one read was found to support each orientation, the identified reads were considered as evidence for a programmed inversion of the region flanked by inverted repeats. As these reads represent transcriptomics data, they also indicate that the different gene variants are expressed.

For three SRA entries, such reads, indicating expression of different gene variants, were identified in the long-read confirmed programmed inversion locus (a list of examined inverted repeats including aggregate summary of identified reads is provided in Supplementary Table S13, and a comprehensive list of identified reads for each pair of inverted repeats is provided in Supplementary Table S14).

RESULTS

Identification of putative gene-altering programmed inversions based on Intra-species variation

We started by compiling a set of candidates for gene-altering programmed inversions, by searching for inverted repeats such that inversion of their flanked region would modify CDSs (Figure 1B, top). First, for each of 35,366 bacterial species, we retrieved one representative genome from the NCBI RefSeq database. Second, we identified inverted repeats (IRs) in each scaffold and used annotated CDS positions to test for each pair of IRs whether inversion of the region flanked by the IRs would modify a pair of CDSs

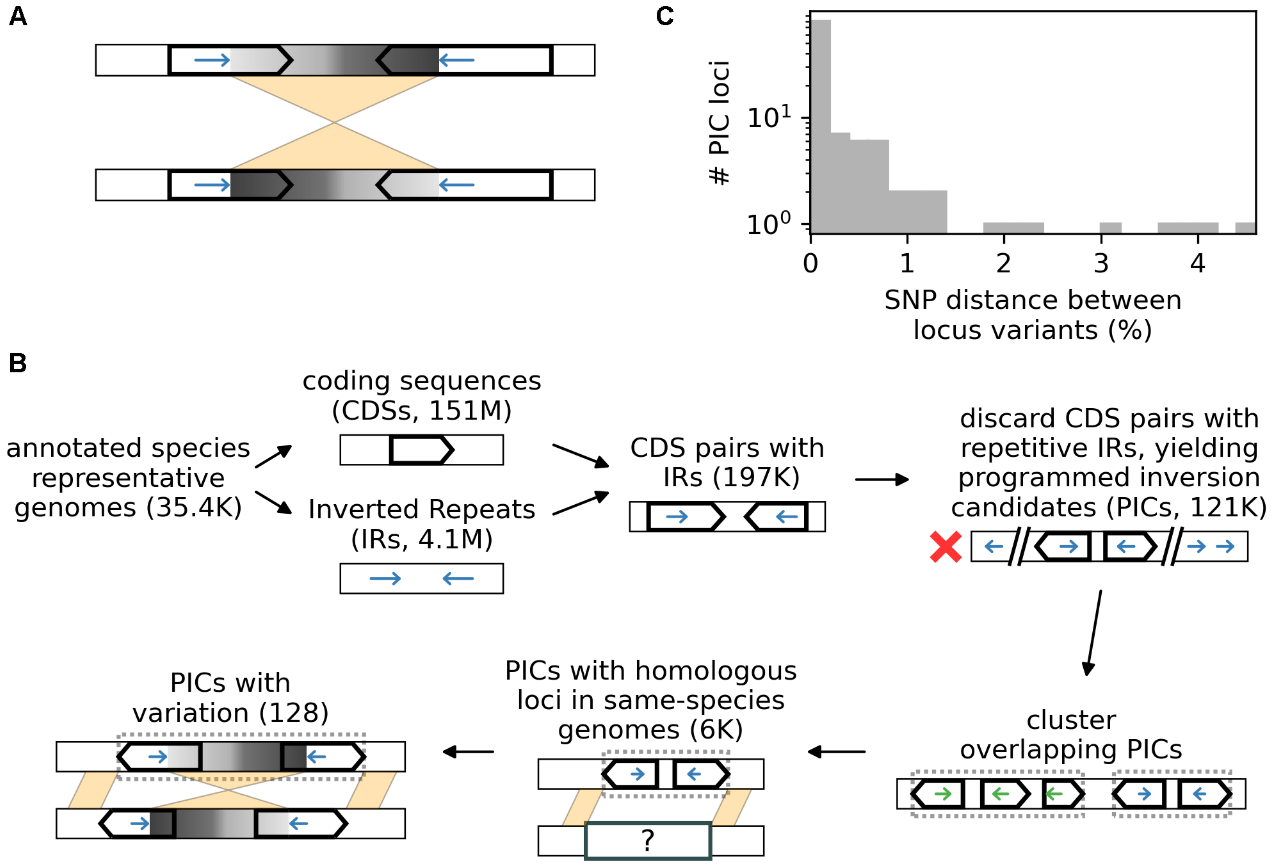
A
C
# PIC loci
10^1
10^0
0
1
2
3
4
SNP distance between locus variants (%)
B
annotated species representative genomes (35.4K)
coding sequences (CDSs, 151M)
Inverted Repeats (IRs, 4.1M)
CDS pairs with IRs (197K)
discard CDS pairs with repetitive IRs, yielding programmed inversion candidates (PICs, 121K)
cluster overlapping PICs
PICs with homologous loci in same-species genomes (6K)
?
PICs with variation (128)

Figure 1. A pipeline for identifying intra-species variation, revealing putative gene-altering programmed inversions. (**A**) Schematic illustration of two variants of a locus containing a gene-altering programmed inversion. Inverted repeats and coding sequences are indicated as colored stick arrows and wide black arrows, respectively. The sequence flanked by inverted repeats is indicated by a grayscale gradient. Alignments are indicated by light orange projections. (**B**) Main steps of the pipeline to identify programmed inversion candidates (PICs) and intra-species variation. See Materials and Methods and main text for an explanation of each step. (**C**) Distribution of SNP distance between PIC loci and homologous loci exhibiting different variants. For data points to be independent, SNP distance between each PIC locus (n = 113) and its closest homologous locus exhibiting a different variant was used. SNP distance was obtained from a progressiveMauve alignment (see Materials and Methods, 'Identification of within-species rearrangements, indicating putative programmed inversions').

(revealing 196 653 CDS pairs with at least one such corresponding IR pair). Finally, reasoning that many of these IRs are part of mobile elements, rather than programmed inversions, we discarded CDS pairs containing repeats with multiple additional copies in the genome (Supplementary Figure S1). The remaining 120 686 CDS pairs were used to define programmed inversion candidates (PICs), each defined by the CDS pair it might modify by inversion of regions flanked by IRs.

Next, we scanned the PICs for intra-species variation, namely, same-species genomes containing different rearrangement variants of PIC loci (Figure 1B, bottom). First, we grouped overlapping PICs, forming PIC loci. Then, for each PIC locus, we searched other genomes of the same species for loci flanked by sequences similar to those flanking the PIC locus. We further filtered found loci to obtain only loci that are either homologous to the PIC locus or to another variant of it potentially arising from programmed inversions. This search identified at least one homologous locus for PIC loci of 6372 PICs. For 128 of these PICs, at least one homologous locus was estimated to contain another variant of the PIC, i.e. intra-species variation was identified for that PIC. Lastly, SNP distance between PIC loci and corresponding homologous loci exhibiting different variants revealed high sequence similarity, suggesting that observed rearrangements occurred relatively recently (Figure 1C).

Genomic sequence attributes associated with identified programmed inversions

In order to reveal genomic sequence attributes associated with gene-altering programmed inversions, we sought out attributes enriched in PICs with identified intra-species variation (compared to those not showing intra-species variation). We identified four quantitative genomic sequence attributes enriched in the intra-species varying PICs (Figure 2A). First, hypothesizing that the length of the IRs might be important for programmed inversions, we compared the distribution of IR length in the PICs with and without intra-species variation (for PICs with more than one IR, the longest repeat was used). This analysis revealed significantly longer repeats in PICs with intra-species variation ($P = 0.0001$, two-sided Mann–Whitney U test; Figure 2B).

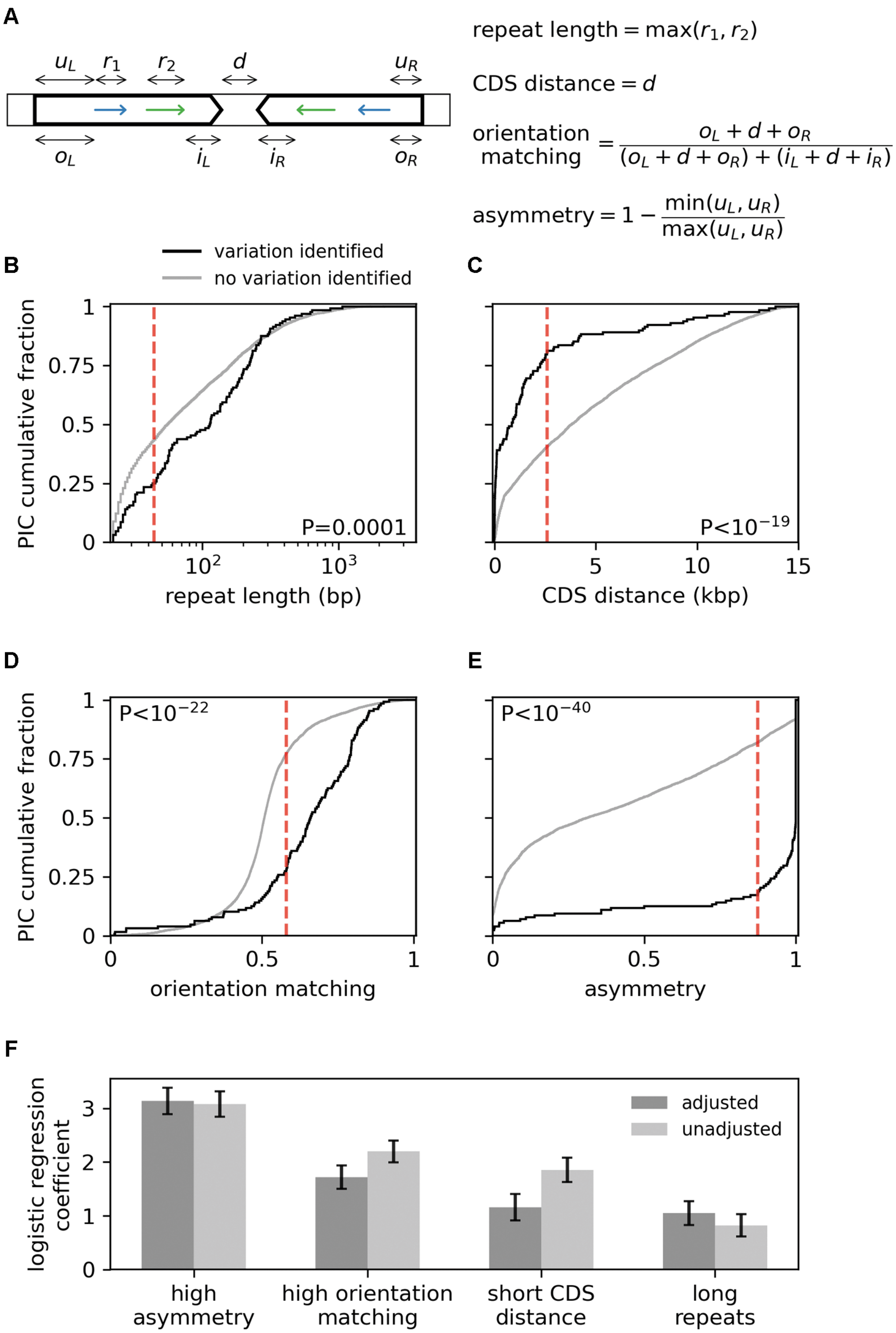

Figure 2. Putative programmed inversions are enriched for several genomic sequence attributes. (**A**) Schematic illustration of genomic sequence attribute measure definitions. d is the distance between coding sequences (or more generally among operons containing these coding sequences, not shown); r_1 and r_2 are the lengths of inverted repeats; u_L and u_R are the lengths of coding sequence regions upstream to all repeats; o_L and o_R are the lengths of coding sequence regions flanking the repeats; i_L and i_R are the lengths of coding sequence regions flanked by the repeats. (**B–E**) Cumulative distribution functions of each genomic sequence attribute measure, shown for programmed inversion candidates (PICs) with identified intra-species variation (n = 128), and for PICs with at least one homologous locus but with no identified intra-species variation ($n = 5964$). P-values represent two-sided Mann–Whitney U tests. Vertical dashed red lines indicate values corresponding to Kolmogorov-Smirnov-statistics. (**F**) Logistic regression coefficients of a model trained to predict whether intra-species variation was identified for each PIC with at least one homologous locus. Features for the model are the four genomic sequence attributes of panels B–E binarized based on Kolmogorov–Smirnov-statistics as thresholds. 'Adjusted' coefficients were produced by a model using all four binarized genomic sequence attribute measures, while 'unadjusted' coefficients were produced by models using each single binarized genomic sequence attribute as the only predictor. Error bars indicate coefficient values ±1SE. Abbreviations: CDS, coding sequence; PIC, programmed inversion candidate.

Next, we consider genomic attributes affecting the length of the inverted region, which we hypothesize could affect inversion efficiency. Assuming two CDSs containing specific IRs at fixed locations within the genes, the distance between these repeats will be determined by the distance between the CDSs and by their relative orientation (head-to-head versus tail-to-tail).

Defining the 'CDS distance' as the distance between the targeted CDSs (or between their operons, see Materials and Methods, 'Genomic sequence attribute definitions'), we find that this distance is significantly shorter in PICs showing intra-species variation than in those not showing intra-species variations ($P < 10^{-19}$, two-sided Mann–Whitney *U* test; Figure 2C). Next, focusing on the orientation of these targeted CDS, we find that they are typically oriented such that their inverted repeats are closer: CDS of PICs showing intra-species variation were typically oriented tail-to-tail when their inverted repeats were situated closer to the gene (or operon) 5′ end (Supplementary Figure S6A), yet head-to-head when their inverted repeats were situated closer to the gene (or operon) 3′ end (Supplementary Figure S6B). Indeed, defining an 'orientation matching' measure, to quantify the extent to which the observed CDS orientation matches the position of IRs inside the CDSs, we found significantly higher orientation-matching values in PICs with intra-species variation ($P < 10^{-22}$, two-sided Mann–Whitney *U* test; Figure 2D).

The last identified genomic sequence attribute involves an asymmetry in lengths of PIC CDSs (or operons). Similar to previous observations in gene-altering programmed inversions (13,18,22,33), we observed in many PICs with identified intra-species variation, that one of the two targeted CDSs (or its operon) lacks the region containing the TSS up to the first IR. We thus defined the 'asymmetry' measure, to quantify both the length of the CDS (or operon) truncated part and its proximity to the IR closest to the TSS (Supplementary Figure S7). We found substantially higher asymmetry values in PICs with intra-species variation ($P < 10^{-40}$, two-sided Mann–Whitney *U* test; Figure 2E).

As these four genomic attributes, enriched in PICs with intra-species variation, are not necessarily statistically independent, we sought to separate their combined and individual contributions. To this end, we used a logistic regression model, for all PICs for which at least one intra-species homologous locus was found. For each such PIC, the model received four binarized features indicating for each of the four genomic sequence attributes, whether its value is above or below the Kolmogorov–Smirnov-statistic value. The model showed that the asymmetry attribute has the largest individual contribution for predicting programmed inversions, yet that all attributes are important for prediction even when the other attributes are considered (Figure 2F).

Gene families associated with predicted programmed inversions

The regression model above allowed us to predict programmed inversions for the entire set of PICs, even for those which lacked homologous loci in same-species genomes. To this end, we utilized the logistic regression model that was trained on PICs with at least one homologous locus, obtaining a predicted probability for each PIC and revealing a large set of potential gene-altering programmed inversion loci (Supplementary Table S4). Using 0.05 as a cutoff (Supplementary Figure S8), we flagged PICs with higher predicted probability as 'predicted programmed inversions'. Leveraging predicted programmed inversions, we next sought to identify gene families enriched in programmed inversion targets as well as genes with close genomic proximity to targeted genes.

Predicted programmed inversion target genes were enriched for several key gene families, including major well known programmed inversion targets, but also Type II restriction-modification (RM) systems and other families. We scanned GenBank product annotations of CDSs potentially targeted by PICs for annotations appearing in multiple PICs (≥12). For each such annotation, we performed a two-sided Fisher's exact test or G-test to test whether it is enriched in predicted programmed inversion targets (Figure 3A). Product annotations with a Bonferroni corrected *P*-value ≤0.05 are listed in Table 1. Multiple statistically significantly enriched annotations belonged to protein families previously described to be targeted by programmed inversions, most notably Type I RM specificity subunit HsdS (17,18,34–37), phage tail (22,33), TonB-linked outer membrane protein (5,38,39), and Shufflon PilV (13,40,41). Yet some other annotations belonged to proteins of unknown function. Notably, multiple statistically significantly enriched annotations belonged to Type II RM enzymes, suggesting this family to be a major target of programmed inversions. To the best of our knowledge, only two prior studies (26,42) found evidence for programmed inversions targeting genes encoding Type II RM enzymes.

Using a similar approach, we attempted to find gene families frequently appearing in close proximity to genes targeted by programmed inversions (yet not themselves targeted). A two-sided Fisher's exact test or G-test was performed for each GenBank product annotation appearing in multiple PICs (≥12) in CDS neighboring potential PIC target CDS, to find whether the annotation is enriched in neighbor CDSs of predicted programmed inversion targets (Figure 3B). Product annotations with a corrected *P*-value ≤0.05 are listed in Table 2. Agreeing with previous observations that gene-altering programmed inversion loci often contain recombinase genes (1,12,13,17,27), multiple statistically significantly enriched annotations belonged to recombinase families.

Long-read genome sequencing data confirm predicted programmed inversion representatives

To directly test our predictions of programmed inversions, we turned to examine long-read genome sequencing data. We started by manually inspecting predicted programmed inversions targeting identified enriched gene families, and identified classes of recurring genomic contexts. Next, for each such recurring genomic context, we manually searched the NCBI Nucleotide and SRA Databases for loci that both contain the genomic context and also reside in genomes with publicly available long-read genome sequencing data.

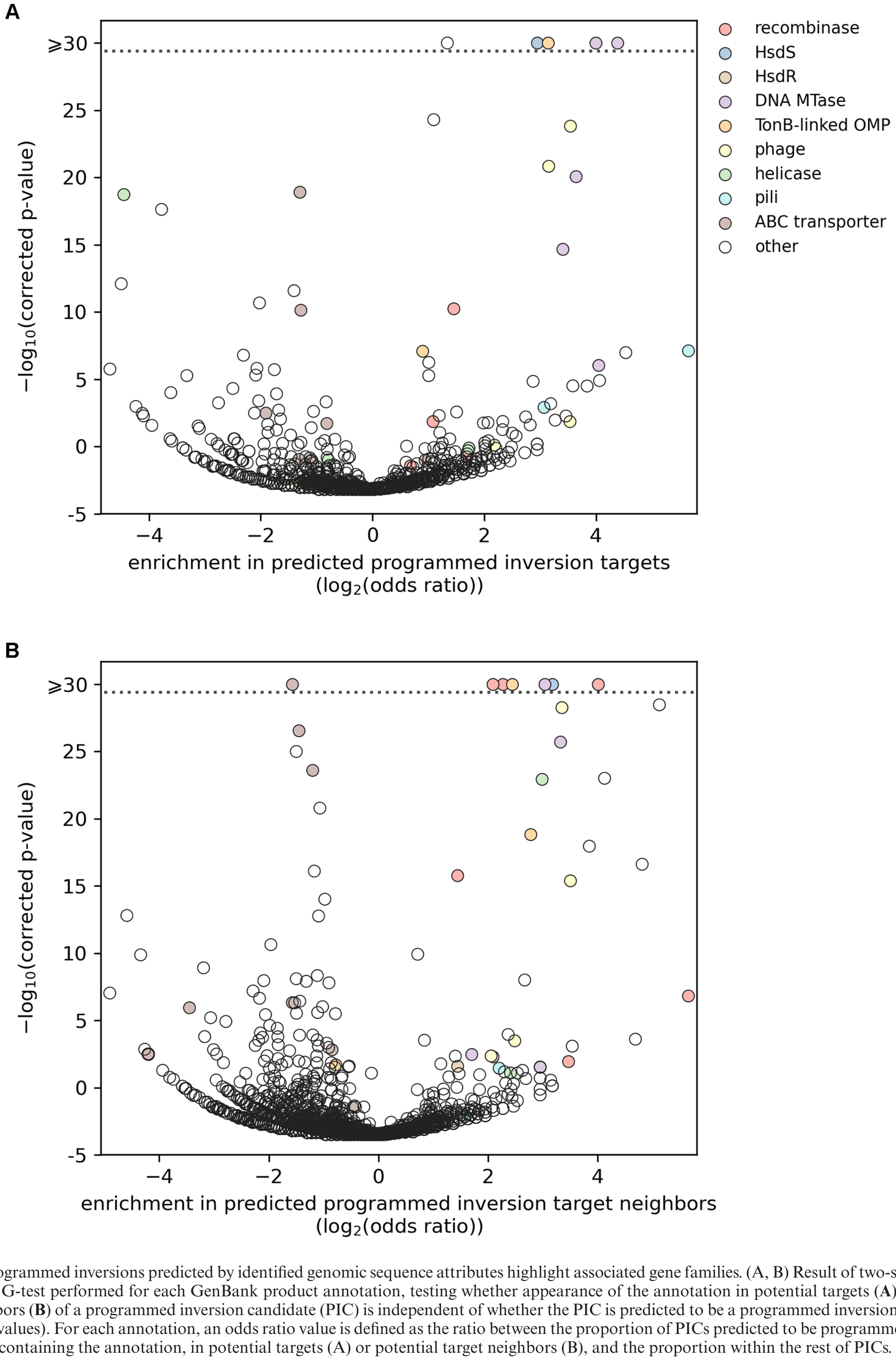

Figure 3. Programmed inversions predicted by identified genomic sequence attributes highlight associated gene families. (A, B) Result of two-sided Fisher's exact test or G-test performed for each GenBank product annotation, testing whether appearance of the annotation in potential targets (**A**) or potential target neighbors (**B**) of a programmed inversion candidate (PIC) is independent of whether the PIC is predicted to be a programmed inversion (Bonferroni corrected *P*-values). For each annotation, an odds ratio value is defined as the ratio between the proportion of PICs predicted to be programmed inversions within PICs containing the annotation, in potential targets (A) or potential target neighbors (B), and the proportion within the rest of PICs. Annotations with an odds ratio of zero are not shown. Abbreviations: ABC, ATP-binding cassette; MTase, methyltransferase; OMP, outer membrane protein.

Table 1. Product annotations statistically significantly enriched or depleted for being targeted by predicted programmed inversions

GenBank CDS product annotation	Enrichment in predicted programmed inversion targets (odds ratio)	Corrected *P*-value
Shufflon system plasmid conjugative transfer pilus tip adhesin PilV	50.191	7.58E-08
DUF4393 domain-containing protein	23.169	1.04E-07
BREX-1 system adenine-specific DNA-methyltransferase PglX	20.928	3.37E-31
DUF4965 domain-containing protein	16.730	1.25E-05
Type II restriction endonuclease	16.549	9.15E-07
Eco57I restriction-modification methylase domain-containing protein	15.980	2.57E-37
DUF559 domain-containing protein	14.339	3.19E-05
N-6 DNA methylase	12.445	8.05E-21
Chemotaxis-specific protein-glutamate methyltransferase CheB	12.012	3.07E-05
Tail fiber protein	11.624	1.44E-24
Tail fiber domain-containing protein	11.577	1.36E-02
DUF1793 domain-containing protein	11.027	5.41E-03
Class I SAM-dependent DNA methyltransferase	10.603	2.15E-15
Endonuclease domain-containing protein	9.649	1.09E-02
Non-ribosomal peptide synthase/polyketide synthase	9.125	6.57E-04
Phage tail protein	8.879	1.41E-21
SusC/RagA family TonB-linked outer membrane protein	8.851	8.50E-52
Pilin	8.364	1.22E-03
Restriction endonuclease subunit S	7.712	6.96E-204
DUF736 domain-containing protein	7.338	1.42E-05
PPE family protein	7.076	2.47E-02
SH3 domain-containing protein	6.691	6.30E-03
Dihydrodipicolinate reductase	5.036	1.37E-02
CpaF family protein	4.192	1.65E-02
ISAs1 family transposase	3.961	1.79E-02
IS5/IS1182 family transposase	2.839	2.70E-03
Tyrosine-type recombinase/integrase	2.731	5.56E-11
Hypothetical protein	2.531	0.00E+00
S-layer homology domain-containing protein	2.288	5.03E-03
Transposase	2.138	5.02E-25
Site-specific integrase	2.120	1.33E-02
IS3 family transposase	2.011	5.58E-07
IS5 family transposase	2.009	5.31E-06
TonB-dependent receptor	1.857	7.92E-08
ABC transporter permease	0.567	1.84E-02
ATP-binding cassette domain-containing protein	0.562	4.77E-04
LysR family transcriptional regulator	0.481	2.42E-03
EAL domain-containing protein	0.468	3.96E-02
ABC transporter substrate-binding protein	0.411	7.36E-11
ABC transporter ATP-binding protein	0.405	1.23E-19
Acyl-CoA dehydrogenase family protein	0.378	2.60E-12
Glycosyltransferase family 4 protein	0.319	8.98E-03
PAS domain-containing protein	0.312	1.96E-03
PAS domain S-box protein	0.304	1.21E-04
Protein kinase	0.296	1.92E-06
Flagellin	0.272	2.23E-02
Sugar ABC transporter ATP-binding protein	0.265	3.28E-03
Serine/threonine protein kinase	0.246	2.11E-11
HAMP domain-containing histidine kinase	0.244	4.02E-04
Tripartite tricarboxylate transporter substrate binding protein	0.238	1.53E-06
Universal stress protein	0.235	4.95E-06
Efflux RND transporter permease subunit	0.231	3.25E-03
Glycerol kinase GlpK	0.202	1.62E-07
RNA polymerase sigma factor	0.177	4.84E-05
ABC-F family ATP-binding cassette domain-containing protein	0.148	5.65E-04
Chaplin	0.117	4.50E-02
ATP-dependent Clp protease ATP-binding subunit	0.115	2.95E-02
RdlA protein	0.100	5.33E-06
Heavy metal translocating P-type ATPase	0.082	9.83E-05
Transcription-repair coupling factor	0.073	2.34E-18
Monovalent cation/H + antiporter subunit A	0.065	2.68E-02
Srylsulfatase	0.058	5.28E-03
Agmatine deiminase family protein	0.057	3.40E-03

Table 1. Continued

GenBank CDS product annotation	Enrichment in predicted programmed inversion targets (odds ratio)	Corrected *P*-value
Acetyl-CoA C-acetyltransferase	0.053	1.05E-03
ATP-dependent DNA helicase RecG	0.046	1.83E-19
Arginine-ornithine antiporter	0.044	7.92E-13
Glutamine synthetase	0.038	1.74E-06
PQQ-dependent dehydrogenase, methanol/ethanol family	0.000	1.98E-05
NAD-dependent succinate-semialdehyde dehydrogenase	0.000	1.90E-03
Citrate synthase	0.000	5.18E-09
Citrate synthase 2	0.000	1.33E-05
Xaa-Pro dipeptidase	0.000	2.41E-02
PQQ-dependent methanol/ethanol family dehydrogenase	0.000	3.58E-04
Type I glutamate–ammonia ligase	0.000	1.25E-10
TonB-dependent siderophore receptor	0.000	7.59E-05

Then, for each found locus, we retrieved the long-read genome sequencing data and searched it for reads spanning the potential programmed inversion loci. While most of these reads have typically matched the reference genome, reads not matching the reference were manually examined, revealing different variants of the locus that might arise from programmed inversion facilitated by inverted repeats (Figure 4A). This analysis, demonstrating coexistence of different programmed inversion variants, provided direct evidence for programmed inversions in 22 different loci, representing different genes and genomic contexts (Table 3, Figure 4B, C, Supplementary Figures S9–S29). Some loci even exhibited more than two variants, corresponding to multiple programmed inversions at the same locus (Figure 4C).

The long-read confirmed programmed inversions appeared in multiple different phyla: Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria. To estimate the distribution of each programmed inversion across bacterial phyla, we searched for homologs of target CDSs, and then searched for inverted repeats overlapping each homolog, as an indication that the homolog might be targeted by programmed inversions (Figure 4D). While confirmed programmed inversion locus homologs were typically found in a single phylum, for one *Bacteroides ovatus* programmed inversion (Nucleotide accession CP083813.1, 5322008–5329826), homologs were found in three different phyla.

Some of the long-read confirmed programmed inversion loci contain systems known to be targeted by programmed inversions (5,13,17,18,22,33–41), but most contain systems that, to the best of our knowledge, were not previously described to be targeted by programmed inversions (43). Most notably, 11 of these programmed inversions target systems encoding Type II RM enzymes, which can be divided to three classes: (a) the anti-phage defense system Class 1 DISARM (44) and similar systems; (b) systems containing homologs of *C. jejuni* Cj0031, which also phase-varies, but through a hypermutable polyG tract rather than inversions (45); and (c) the anti-phage defense system BREX type 1 (42). In addition, one programmed inversion targets a gene encoding a protein of unknown function, containing four domains of unknown function (DUFs): DUF4964 (pfam16334), DUF5127 (pfam17168), DUF4965 (pfam16335) and DUF1793 (pfam08760), resembling a family of fungal glutaminases (46). Manual examination of predicted programmed inversions targeting this alleged glutaminase, revealed that in almost all cases, these target genes appear upstream, potentially in the same operon, to genes coding for TonB-linked outer membrane protein, another target of programmed inversions (5,38,39), perhaps suggesting some similarity in their function or in the function of their inversions. Finally, in one locus, programmed inversions target a gene encoding a shufflon PilV (pfam04917) and phage tail collar (pfam07484) fusion-protein (based on the NCBI Conserved Domain Database). Several PilV genes were described to be targeted by programmed inversions and named shufflons (13,40). However, protein sequence alignment (using online BLASTP) revealed the identified shufflon PilV and phage tail collar fusion-protein (Protein accession UCP91044.1) to be homologous to previously described shufflon proteins (Supplementary Table S12) in the shufflon PilV domain, but not in the phage tail collar domain.

Next, considering novel programmed inversion target genes, we asked whether programmed inversion variants not only coexist, but also express their different gene variants. To this end, we manually scanned the NCBI SRA Database for transcriptomics datasets of our long-read confirmed programmed inversion loci. As these datasets are composed of short reads, they generally cannot directly identify programmed inversion variants, yet they can support specific programmed inversions. Indeed, for each of three confirmed programmed inversion loci, we identified an RNA-seq dataset supporting expression of different programmed inversion gene variants (Supplementary Figure S32). This analysis further supports the hypothesis that in these loci, programmed inversions generate not only genetic heterogeneity, but also expression differences and therefore possibly phenotypic heterogeneity.

The distribution of long-reads across variants of a *Lacticaseibacillus rhamnosus* locus containing a BREX type 1 system (Nucleotide accession NC_013198.1, 2154002–2170387, Figure 4B) was highly biased against variants with a single inversion, favoring instead the reference and variants with two inversions. Of the 176 reads matching the locus (and satisfying our requirements, see Materials and Methods, 'Variant identification in long-read genome se-

Table 2. Product annotations statistically significantly enriched or depleted for appearing adjacent to genes targeted by predicted programmed inversions

GenBank CDS product annotation	Enrichment in predicted programmed inversion target neighbors (odds ratio)	Corrected *P*-value
Tyrosine-type DNA invertase PsrA	50.191	1.54E-07
BREX system P-loop protein BrxC	34.853	3.39E-29
BREX-1 system phosphatase PglZ type A	28.033	2.45E-17
DUF2612 domain-containing protein	25.731	2.50E-04
DUF1016 family protein	17.424	9.89E-24
Master DNA invertase Mpi family serine-type recombinase	16.093	8.20E-48
Virulence RhuM family protein	14.373	1.11E-18
DUF417 family protein	11.581	8.26E-04
Phage tail protein I	11.310	4.00E-16
Tsr0667 family tyrosine-type DNA invertase	11.027	1.10E-02
Tail fiber assembly protein	10.146	5.26E-29
SAM-dependent DNA methyltransferase	9.971	1.91E-26
Restriction endonuclease subunit S	8.990	7.43E-123
type I restriction-modification system subunit M	8.153	2.56E-89
DEAD/DEAH box helicase family protein	7.876	1.19E-23
Eco57I restriction-modification methylase domain-containing protein	7.719	2.92E-02
Mycothione reductase	7.719	2.92E-02
SusD/RagB family nutrient-binding outer membrane lipoprotein	6.839	1.49E-19
Alginate lyase family protein	6.345	9.93E-09
Tail fiber protein	5.594	3.22E-04
RagB/SusD family nutrient uptake outer membrane protein	5.420	3.39E-47
Glutamate-1-semialdehyde 2,1-aminomutase	5.151	1.12E-04
Site-specific integrase	4.803	2.55E-125
Prepilin peptidase	4.576	3.43E-02
HsdR family type I site-specific deoxyribonuclease	4.249	5.12E-03
Tyrosine-type recombinase/integrase	4.232	4.81E-81
Phage tail protein	4.132	4.16E-03
N-6 DNA methylase	3.238	3.45E-03
Recombinase family protein	2.718	1.71E-16
Type I restriction endonuclease subunit R	2.709	2.52E-02
Non-ribosomal peptide synthetase	2.642	4.59E-03
Amino acid adenylation domain-containing protein	2.197	1.71E-02
IS3 family transposase	1.787	3.00E-04
Transposase	1.642	1.21E-10
Helix-turn-helix transcriptional regulator	0.690	2.52E-02
Glycosyltransferase	0.684	2.86E-02
TonB-dependent receptor	0.582	1.99E-02
Response regulator	0.580	3.19E-06
LacI family DNA-binding transcriptional regulator	0.576	3.04E-02
ABC transporter permease subunit	0.555	1.52E-03
MarR family transcriptional regulator	0.551	1.41E-02
Tetratricopeptide repeat protein	0.539	4.04E-02
Alpha/beta hydrolase	0.534	1.62E-08
TetR family transcriptional regulator	0.532	1.04E-03
GntR family transcriptional regulator	0.519	4.96E-04
SDR family oxidoreductase	0.507	9.71E-15
Extracellular solute-binding protein	0.493	9.66E-07
MFS transporter	0.477	1.63E-21
Response regulator transcription factor	0.469	1.72E-13
Substrate-binding domain-containing protein	0.467	8.05E-04
FAD-dependent oxidoreductase	0.463	2.58E-06
ATP-binding cassette domain-containing protein	0.461	4.65E-09
Cytochrome P450	0.460	2.83E-04
Aldehyde dehydrogenase family protein	0.458	1.45E-02
Sensor histidine kinase	0.458	4.43E-03
TetR/AcrR family transcriptional regulator	0.444	7.87E-17
Cupin domain-containing protein	0.436	1.22E-04
ABC transporter ATP-binding protein	0.433	2.47E-24
Acyl-CoA dehydrogenase family protein	0.401	1.24E-08
Enoyl-CoA hydratase	0.393	1.29E-02
IclR family transcriptional regulator	0.369	3.72E-07
Glycosyltransferase family 4 protein	0.369	1.26E-04
Acyltransferase	0.365	1.05E-02
ABC transporter substrate-binding protein	0.364	2.79E-27
Universal stress protein	0.354	2.96E-06
LysR family transcriptional regulator	0.354	1.00E-25

Table 2. Continued

GenBank CDS product annotation	Enrichment in predicted programmed inversion target neighbors (odds ratio)	Corrected *P*-value
Enoyl-CoA hydratase/isomerase family protein	0.353	8.00E-09
CoA transferase	0.347	4.96E-07
ABC transporter permease	0.337	4.85E-45
Sugar ABC transporter permease	0.335	4.56E-07
Acyl-CoA/acyl-ACP dehydrogenase	0.325	1.82E-03
Lrp/AsnC family transcriptional regulator	0.314	1.20E-02
Aquaporin family protein	0.294	2.00E-02
FadR family transcriptional regulator	0.281	6.46E-04
Sugar phosphate isomerase/epimerase	0.281	7.00E-03
carboxymuconolactone decarboxylase family protein	0.280	1.65E-04
NAD(P)-dependent oxidoreductase	0.278	3.88E-05
Acyl-CoA thioesterase	0.268	3.16E-02
Phosphotransferase	0.256	2.32E-11
PAS domain S-box protein	0.253	2.36E-05
S9 family peptidase	0.252	3.87E-02
LCP family protein	0.244	4.25E-03
FAD-binding oxidoreductase	0.239	3.75E-06
Oxidoreductase	0.236	1.06E-02
Tripartite tricarboxylate transporter substrate binding protein	0.234	1.11E-08
Aspartate aminotransferase family protein	0.231	1.84E-04
Pentapeptide repeat-containing protein	0.225	8.65E-05
EAL domain-containing protein	0.222	2.25E-07
RDD family protein	0.220	3.06E-04
Aldehyde dehydrogenase	0.210	7.36E-04
Efflux RND transporter periplasmic adaptor subunit	0.205	6.47E-08
Glycerol-3-phosphate dehydrogenase/oxidase	0.145	1.21E-05
TRAP transporter small permease	0.141	1.39E-02
RdlA protein	0.129	3.18E-03
PilZ domain-containing protein	0.124	1.50E-03
Acetyl-CoA C-acyltransferase	0.120	6.33E-06
Acyl-CoA thioesterase II	0.111	1.61E-04
Chaplin	0.109	1.24E-09
Amino acid ABC transporter permease	0.091	1.11E-06
TRAP transporter small permease subunit	0.055	3.12E-03
Na+/H + antiporter subunit C	0.054	3.21E-03
sn-glycerol-3-phosphate ABC transporter ATP-binding protein UgpC	0.054	3.21E-03
TRAP transporter substrate-binding protein	0.052	1.41E-03
Bacterial proteasome activator family protein	0.049	1.34E-10
Arginine deiminase	0.041	1.59E-13
Bifunctional [glutamine synthetase] adenylyltransferase/[glutamine synthetase]-adenylyl-L-Tyrosine phosphorylase	0.033	9.29E-08
Na+/H+ antiporter subunit B	0.000	1.32E-02
Succinate dehydrogenase assembly factor 2	0.000	1.37E-23
NtaA/DmoA family FMN-dependent monooxygenase	0.000	1.10E-03
DUF502 domain-containing protein	0.000	4.03E-05
Pyridoxamine 5'-phosphate oxidase	0.000	6.73E-07
Fe-S cluster assembly protein HesB	0.000	9.58E-03
Cytochrome *c*-550 PedF	0.000	9.42E-05

quencing data' and 'Distribution of variants in long reads'), 90 matched the reference variant, 85 matched variants that can be transformed back to the reference variant by two inversions (2-inversion variants), yet only five matched variants that can be transformed back to the reference variant by a single inversion (1-inversion variants). As 5/176 of the reads matched 1-inversion variants, it is highly unlikely that the proportion of 1-inversion variants in the sequenced sample was 0.5 ($P = 2.9 \times 10^{-44}$, two-sided binomial test). Also when only considering non-reference variants, it is highly unlikely that the proportion of 1-inversion variants and 2-inversion variants in the sequenced sample were identical (5/86 of the reads, $P = 9.6 \times 10^{-19}$, two-sided binomial test), suggesting 1-inversion variants are less stable than other variants. Focusing on the PglX CDS immediately downstream to the BrxC CDS (located at 2165561–2169193) brings to light some differences between 1-inversion variants and the rest of the variants. In 2-inversion variants and in the reference variant, this PglX can be divided to three regions: (a) a region upstream to all repeats, encoding the N-terminus of the protein; (b) a variable region overlapping repeats; and (c) a region downstream to all repeats, encoding the C-terminus of the protein. Conversely, in 1-inversion variants, only the first two regions appear in this PglX, with the recombinase gene (located at 2160833–2161915 in the reference genome) replacing the third region. Given the observation that a BREX type 1 system in *Bacillus cereus* H3081.97 contains two operons:

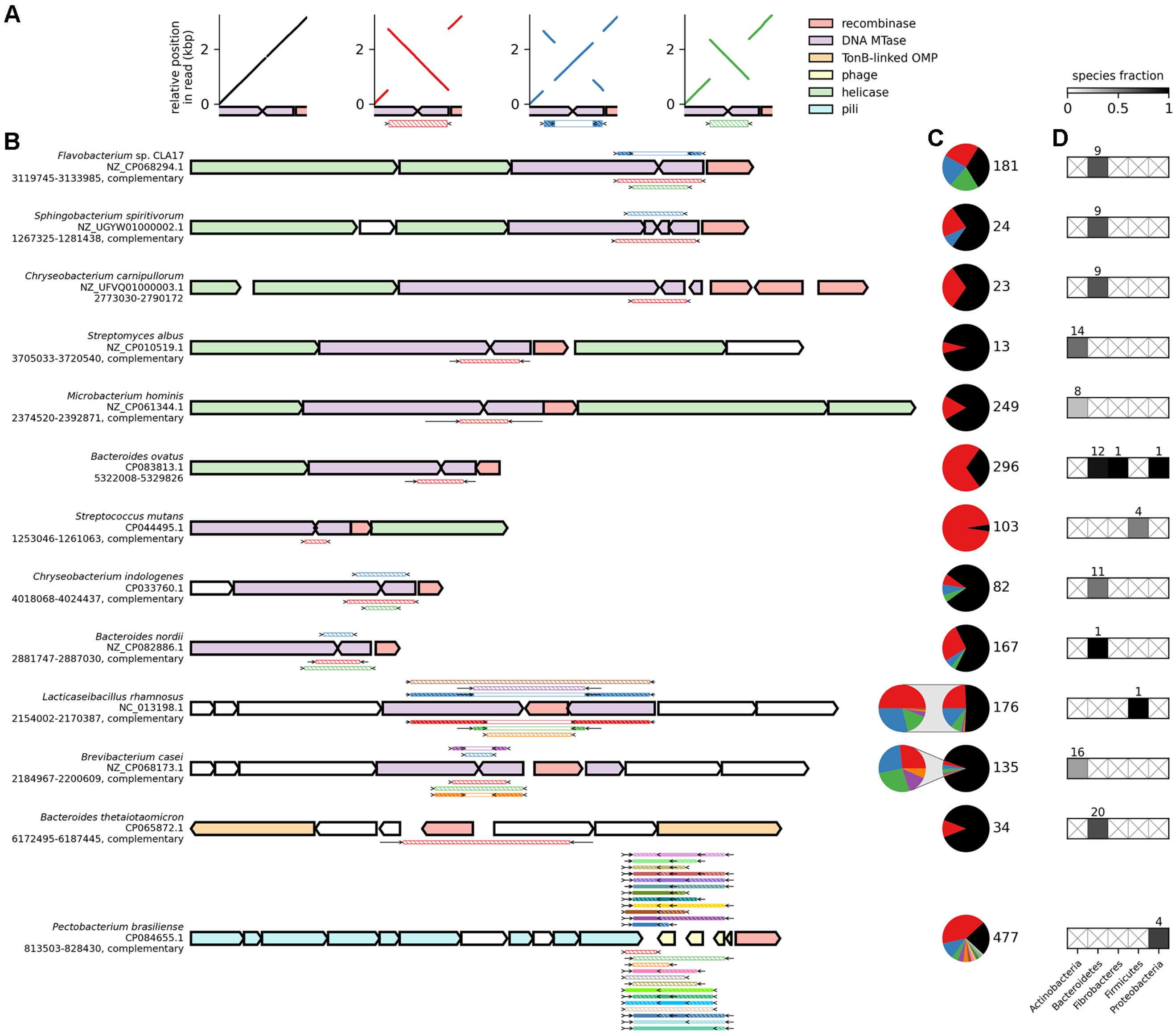
A
relative position in read (kbp)
2
0
recombinase
DNA MTase
TonB-linked OMP
phage
helicase
pili
species fraction
0
0.5
1
B
C
D
Flavobacterium sp. CLA17
NZ_CP068294.1
3119745-3133985, complementary
181
9
Sphingobacterium spiritivorum
NZ_UGYW01000002.1
1267325-1281438, complementary
24
9
Chryseobacterium carnipullorum
NZ_UFVQ01000003.1
2773030-2790172
23
9
Streptomyces albus
NZ_CP010519.1
3705033-3720540, complementary
13
14
Microbacterium hominis
NZ_CP061344.1
2374520-2392871, complementary
249
8
Bacteroides ovatus
CP083813.1
5322008-5329826
296
12
1
1
Streptococcus mutans
CP044495.1
1253046-1261063, complementary
103
4
Chryseobacterium indologenes
CP033760.1
4018068-4024437, complementary
82
11
Bacteroides nordii
NZ_CP082886.1
2881747-2887030, complementary
167
1
Lacticaseibacillus rhamnosus
NC_013198.1
2154002-2170387, complementary
176
1
Brevibacterium casei
NZ_CP068173.1
2184967-2200609, complementary
135
16
Bacteroides thetaiotaomicron
CP065872.1
6172495-6187445, complementary
34
20
Pectobacterium brasiliense
CP084655.1
813503-828430, complementary
477
4
Actinobacteria
Bacteroidetes
Fibrobacteres
Firmicutes
Proteobacteria

Figure 4. Long-read genomic sequencing data show coexistence of different programmed inversion variants in multiple gene-altering programmed inversion loci. (**A**) Example alignments of representative reads of four variants of a programmed inversion locus in *Flavobacterium* sp. CLA17. Each plot shows the read alignment to the reference variant sequence at the programmed inversion locus (NZ_CP068294.1, 3119745–3133985). For each read, the variant architecture supported by the alignment is illustrated as difference from the reference variant (bottom), with black arrows indicating inverted repeats at locations matching the observed inversions. Sub-regions of the region differentiating between the variants are shown as rectangles with different fill and stripe patterns: white-filled if the sub-regions appear in the same strand and location in both variants; white-filled with colored diagonal stripes if the sub-regions appear in the same location but on opposite strands; color-filled if the sub-regions appear in the same strand but in different locations; and color-filled with white diagonal stripes if the sub-regions appear on opposite strands and in different locations. (**B**) Genomic contexts of programmed inversion loci as appearing in the reference genome, and differences from each long-read non-reference variant. Species, NCBI Nucleotide accession and location are specified for each locus (left). For each non-reference variant observed in long reads, sub-regions of the region differentiating between the variant and the reference variant are shown as rectangles with different fill and stripe patterns, as in (A). Coding sequences are colored if they belong to a manually compiled set of gene families, see Materials and Methods, ‘Gene family assignment’. In case the longest target coding sequence in the locus appears on the reverse strand in the reference genome, a mirror image of the locus is shown. (**C**) Identified variant distribution across reads that match the locus (see Materials and Methods, ‘Distribution of variants in long reads’). The reference variant is colored black, while other variants are colored to match their illustration in (B) of their difference from the reference variant. Total number of reads matching the locus is indicated (pie chart, right). (**D**) Distribution of homologs of programmed inversion targets across bacterial phyla. For each programmed inversion and for each phylum, the fraction of this phylum species in which any homolog contains inverted repeats, marking it as potentially targeted by programmed inversions, is indicated in grayscale. The number of this phylum species in which homologs were found is indicated, or the phylum rectangle is marked with an X in case no homologs were found. Abbreviations: MTase, methyltransferase; OMP, outer membrane protein.

Table 3. Gene-altering programmed inversion loci

NCBI Nucleotide accession	Longest target CDS location	Longest target CDS product description	Locus description
NZ_CP068294.1	3122130–3125861	Type II restriction-modification enzyme, homologous to Class 1 DISARM DrmMI	Similar to Class 1 DISARM in terms of some encoded protein domains, but with a different gene order, and with a single CDS encoding two short YprA (COG1205, helicase) domains and a DUF1998 (pfam09369) domain
NZ_UGYW01000002.1	1269931–1273395	Type II restriction-modification enzyme, homologous to Class 1 DISARM DrmMI	Similar to Class 1 DISARM in terms of some encoded protein domains, but with a different gene order
NZ_UFVQ01000003.1	2778297–2784881	Type II restriction-modification enzyme with an SNF2 (COG0553, helicase) domain, with its C-terminus part homologous to Class 1 DISARM DrmMI	Similar to Class 1 DISARM in terms of some encoded protein domains, but with a different gene order, and with a single CDS encoding a short YprA (COG1205, helicase) domain and a DUF1998 (pfam09369) domain
NZ_CP010519.1	3712964–3717292	Class 1 DISARM DrmMI	Class 1 DISARM
NZ_CP061344.1	2385445–2390031	Type II restriction-modification enzyme, partially homologous to Class 1 DISARM DrmMI	Similar to Class 1 DISARM in terms of some encoded protein domains and gene order, and with a single CDS encoding a long YprA (COG1205, helicase) domain and a DUF1998 (pfam09369) domain
CP083813.1	5324991–5328359	Type II restriction-modification enzyme, homologous to Cj0031 of C. jejuni	Type II restriction-modification CDS with an immediately upstream CDS encoding a phospholipase D (cd09178) domain and a SNF (cd18793, helicase) domain
CP044495.1	1257875–1261063	Type II restriction-modification enzyme, homologous to Cj0031 of C. jejuni	Type II restriction-modification CDS with a downstream CDS encoding a phospholipase D (cd09178) domain and a SNF (cd18793, helicase) domain
CP033760.1	4019614–4023351	Type II restriction-modification enzyme, homologous to Cj0031 of C. jejuni	Type II restriction-modification CDS with an immediately upstream CDS encoding a DUF1016 (pfam06250) domain
NZ_CP082886.1	2883305–2887030	Type II restriction-modification enzyme, homologous to Cj0031 of C. jejuni	Solitary Type II restriction-modification CDS
NC_013198.1	2161973–2165527	BREX type 1 PglX	BREX type 1
NZ_CP068173.1	2193315–2195903	N-terminus and middle parts of BREX type 1 PglX	BREX type 1, with one CDS encoding the N-terminus and middle parts of PglX (targeted by programmed inversions), and a downstream CDS encoding the C-terminus part of PglX (not targeted by programmed inversions)
CP065872.1	6177248–6179761	A protein of unknown function containing 4 domains of unknown function (DUFs): DUF4964 (pfam16334), DUF5127 (pfam17168), DUF4965 (pfam16335) and DUF1793 (pfam08760)	A protein of unknown function (targeted by programmed inversions) with a downstream presumed operon (according to short distances between CDSs) containing, among others, CDSs encoding outer membrane proteins SusC and SusD
CP084655.1	816997–818571	A protein containing a Shufflon PilV N-terminus (pfam04917) domain and a phage tail collar (pfam07484) domain	Two adjacent presumed operons (according to short distances between CDSs) containing multiple pilus associated CDSs
NZ_CP022464.2	6453804–6455096	Type I restriction-modification HsdS	Type I restriction-modification, with core CDS order HsdR-HsdM-HsdS
CP046428.1	3459162–3460715	Type I restriction-modification HsdS	Type I restriction-modification, with core CDS order HsdR-HsdM-HsdS, and with a CDS encoding a DUF1016 (pfam06250) domain between core CDSs

Table 3. Continued

NCBI Nucleotide accession	Longest target CDS location	Longest target CDS product description	Locus description
CP081899.1	2601241–2602485	Type I restriction-modification HsdS	Type I restriction-modification, with core CDS order HsdR-HsdM-HsdS, and with a CDS encoding a dinD (PRK11525) domain and a RhuM (pfam13310) domain, as well as a CDS encoding a GIY-YIG nuclease (cl15257) domain, between core CDSs
NZ_CP082886.1	4369547–4371175	Type I restriction-modification HsdS	Type I restriction-modification, with core CDS order HsdR-HsdM-HsdS, and with a CDS encoding a hypothetical protein, as well as a CDS encoding a Fic/DOC (pfam02661) domain, between core CDSs
NZ_CP059830.1	15345–16556	Type I restriction-modification HsdS	Type I restriction-modification, with core CDS order HsdM-HsdS-HsdR
CP056267.1	5092989–5094518	Phage tail fiber (COG5301) domain-containing protein	Prophage
CP076386.1	93499–94659	Phage tail fiber (COG5301) domain-containing protein	Prophage, with a CDS encoding DNA endonuclease SmrA, a CDS encoding a MFS transporter domain, and a CDS encoding a Phytase (pfam13449) domain
CP066032.1	4023024–4024490	Phage tail fiber protein, homologous to the variable tail fiber protein of phage Mu (NP_050653.1)	Prophage
NZ_CP012938.1	2847275–2850391	SusC	A presumed operon (according to short distances between CDSs) composed of CDSs encoding outer membrane proteins SusC and SusD

brxA-*brxB*-*brxC*-*pglX* and *pglZ*-*brxL* (42) we hypothesized that this is also the case in *L. rhamnosus*. This might suggest two causes for 1-inversion variant alleged instability: (a) Lack of C-terminus in PglX someway promotes inversions; and/or (b) the recombinase gene is transcribed as part of the first operon in 1-inversion variants, making the recombinase more active in these variants and leading to rapid switching to other variants, somewhat similar to increased levels of FimE recombinase in *fim* ON variants (47).

DISCUSSION

Our study addresses an unmet need for a broad and systematic search for gene-altering programmed inversions. Two studies systematically and widely searched for programmed inversions targeting genes encoding Type I RM specificity subunit HsdS (17,18), and one study systematically searched 203 bacterial genomes for any programmed inversion, regardless of whether it targets genes or regulatory elements (27). Yet, potential insights that might be gained by analyzing a large set of gene-altering programmed inversions targeting diverse gene families, have been out of reach. Here, we compiled such a diverse set by scanning representative genomes of over 35 000 species for loci containing IRs overlapping CDSs, and identifying intra-species variation in these loci by comparing them to other loci in same-species genomes. Finally, key predictions of this analysis were confirmed using long-read genome sequencing data revealing within-sample variant coexistence.

Analyzing this dataset, we identified four genomic sequence attributes enriched for putative gene-altering programmed inversions, i.e. programmed inversion candidates for which intra-species variation was identified. Programmed inversions were enriched in genomic sequence attributes with long IRs or with a short distance between the targeted CDSs (or their operons). A perhaps less expected genomic sequence attribute is the orientation of CDSs targeted by programmed inversions, i.e. head-to-head or tail-to-tail, which tend to orient such that IRs are closer to each other. Assuming that typically, a shorter distance between IRs allows for higher frequency of programmed inversions, we hypothesized that throughout evolution of CDSs targeted by programmed inversions, CDSs oriented such that IRs are closer to each other would be favored. The last genomic sequence attribute is the asymmetry of CDSs (or their operons) targeted by programmed inversions, in terms of the length of parts of CDSs upstream to all IRs. In other words, one of the targeted CDSs is missing its part which is upstream to all IRs. Similar to what was previously observed (13,18,22,33), it seems that these target CDSs have a single copy of their upstream part, while programmed inversions switch between different versions of the downstream part of the CDS. It is possible that this pattern arose from an inverted duplication followed by elimination of transcription of one of the copies, e.g. by a promoter mutation.

Such a process would render the non-transcribed region upstream to all IRs non-functional in all variants, ultimately leading to its deletion.

Using these genomic sequence attributes, we predicted many more gene-altering programmed inversions and identified gene families associated with predicted programmed inversions. Leveraging programmed inversion predictions, we searched for gene families enriched for appearing in CDSs targeted by predicted programmed inversions, relative to CDSs targeted by programmed inversion candidates not predicted to be programmed inversions. Unsurprisingly, multiple statistically significantly enriched gene families were previously described to be targeted by programmed inversions (5,13,17,18,22,33–41). Intriguingly, the Type II RM family was one of the most prominent statistically significantly enriched gene families that came up in this analysis. Two studies found evidence for Type II RM genes being targeted by programmed inversions (26,42), yet the family is largely considered to not be a target of programmed inversions (9,48,49). Additionally, several significantly enriched gene families code for proteins of unknown function and transposases, which are not known to be targeted by programmed inversions. Long-read evidence of programmed inversion variants was obtained for representatives of some of these families. For others, including transposases, long-read evidence supporting programmed inversions is absent; thus, it is still unclear whether they are indeed targets of programmed inversions.

Finally, we have used same-sample long-read variations to directly confirm coexisting programmed inversion variants in recurring genomic contexts of predicted programmed inversions targeting enriched gene families. This analysis provided sound evidence for variant coexistence in long-read genome sequencing experiments, strongly suggestive of programmed inversions. We thus identified multiple programmed inversion loci, exhibiting different genomic contexts.

Examining this diverse set of programmed inversion genomic contexts revealed some recurring patterns. One of these patterns was fusion-genes, which appeared in multiple identified programmed inversion loci, encoding for various domains, including helicase YprA, DUF1998, SNF2 helicase, Type II RM, SNF helicase, phospholipase D, Shufflon PilV, and phage tail collar (Table 3). With regard to the identified Shufflon PilV and phage tail collar fusion-protein, we noted that previously, variation in shufflon PilV conferring variable specificity in pilus binding to recipient lipopolysaccharide was compared to variation in Bacteriophage Mu tail fiber conferring variable specificity in phage tail binding to lipopolysaccharide (41). This comparison leads to the hypothesis that the fusion-gene we identified, appearing with multiple pilus genes, constitutes a repurposing of the phage tail domain for recipient specificity in bacterial conjugation. Moreover, three loci reminiscent of the anti-phage defense system Class 1 DISARM (44) were very similar in terms of order of encoded protein domains, but differed in terms of whether protein domains were encoded together or by different genes (NZ_CP068294.1: 3119745–3133985, NZ_UGYW01000002.1: 1267325–1281438, NZ_UFVQ01000003.1: 2773030–2790172, Figure 4B). In addition, two loci containing different systems exhibited a gene encoding a DUF1016 domain, which was previously predicted to have endonuclease activity (50). One of these loci contained a programmed inversion targeting a homolog of *C. jejuni cj0031* (45), while the other locus contained a Type I RM system (CP033760.1: 4018068–4024437; CP046428.1: 3453614–3464013, Figure 4B, Supplementary Figure S22).

Another curious pattern we noticed was the seeming two different approaches to generate variation in a protein while keeping its N- and C-termini constant. One approach can be seen in a *Lacticaseibacillus rhamnosus* locus containing a BREX type 1 system (NC_013198.1, 2154002–2170387, Figure 4B). Our results, combined with the assumption of a single operon for *brxA*-*brxB*-*brxC*-*pglX*, as was shown to be the case in *Bacillus cereus* H3081.97 (42), and the assumption that a PglX missing its C-terminus is not fully-functional, suggest that programmed inversions in this locus are meant to switch between different versions of a middle part of PglX, keeping its N- and C-termini constant (Supplementary Figure S33, top). Moreover, as two nested inversions are required to switch some part in the middle of PglX, switching between two functional variants necessitates first switching to an intermediate variant. Our data indicate that such intermediate variants, which we termed 1-inversion variants, are much less stable than fully-functional variants, suggesting active regulation. Thus, it seems that the approach used in this *L. rhamnosus* locus requires active regulation to lower levels of intermediate variants, possibly by an architecture which up-regulates the recombinase in the 1-inversion state. Another approach can be seen in a *Brevibacterium casei* locus, which also contains a BREX type 1 system (NZ_CP068173.1, 2184967–2200609, Figure 4B), but with a disrupted *brxA*-*brxB*-*brxC*-*pglX* operon. Compared to the *L. rhamnosus* locus, the operon in *B. casei* is split to two, with the interruption in protein product coinciding with the C-terminus end of the variable region in PglX (Supplementary Figure S33). Thus, this *B. casei* locus seems to demonstrate another approach to generate variation in a protein while keeping its N- and C-termini constant: splitting the protein into two proteins where the variable region ends, and then using programmed inversions to modify the C-terminus of the first protein. One advantage of this approach is the lack of non-functional intermediate variants.

Considering the functions of gene families and pathways that are targeted by gene-altering programmed inversions, specificity determining pathways and genes involved in phage-bacteria interaction stand out. Programmed inversions targeting Type I RM systems were shown to alter DNA motifs affected by RM-activity (10,17–19). Due to differential distribution of DNA motifs across phage strains, it seems plausible that different variants are more efficient at restricting different phage strains (6). Similarly, perhaps programmed inversions targeting Type II RM systems also provide different phase variants with differential resistance to phage strains. Moreover, one programmed inversion-targeted TonB-dependent transporter was hypothesized to be a phage receptor (5), possibly altering phage-susceptibility. Additionally, some phage tail genes are also targeted by programmed inversions, conferring different host specificity (22). Finally, shufflon programmed inversions were shown to provide different conjugative pilus specificities (21).

It is unclear why bacteria use such sophisticated programmed inversion systems making each cell express a single gene variant, rather than encoding and expressing multiple gene variants in each cell. In the case of Type I RM systems, perhaps having multiple HsdS variants in the same cell substantially increases the risk of autoimmunity (51). Additionally, if the functional complex contains multiple copies of the protein, as was shown for some phage tail fibers (trimer) (52,53) and TonB-linked outer membrane protein SusC (dimer) (54), then perhaps multiple protein variants in the same cell would lead to nonfunctional hetero-complexes. Furthermore, in the case of a phage tail fiber, multiple such proteins need to come together to form a single structure, namely a virion possessing multiple tail fibers. It is possible that a virion having tail fibers of different host-specificities would exhibit impaired infectivity. As bacteria were observed to be connected through multiple pili (55), similar reasoning might lead to an analogous hypothesis for the shufflon PilV protein, which was shown to determine conjugation specificity (21).

Our approach has several limitations. First, we chose to scan only bacterial species, as most reported gene-altering programmed inversions were found in bacteria. Our analysis, therefore, does not identify and characterize programmed inversions in viruses, archaea and eukaryotes. Second, we analyzed only a single representative genome for each bacterial species. This limited the bias toward more sequenced species, but we probably missed many programmed inversions due to this choice. In addition, we looked for long-read evidence for programmed inversion only for manually chosen loci. Performing this search systematically and computationally would probably uncover more programmed inversions. Finally, our search heavily relied on annotation of CDS locations. This seems especially problematic, as often there is high asymmetry between target CDSs, with one very short target CDS. Very short CDSs might not be identified by annotation software, as indeed seems to be the case for some of the targeted very short CDSs in the programmed inversion locus we identified in *Pectobacterium brasiliense* (CP084655.1, 813503–828430, Figure 4B).

Our methodology as well as the large set of programmed inversion candidates and predictions reported in this study (Supplementary Table S4) open some interesting directions for further investigation. First, a similar method might be applied in order to characterize and reveal gene-altering programmed inversions in archaea and viruses. Second, comparing species with predicted programmed inversions to those lacking them, might uncover ecological niches in which such systems are more beneficial. Third, recombinase genes adjacent to predicted programmed inversions may be analyzed, along with inverted repeat sequences, to identify recombinase target motifs characteristic of gene-altering programmed inversions, as well as associated recombinase families (17). Finally, a similar approach might be applied to identify genomic sequence attributes of other types of phase variation systems, such as invertible promoters (8), slipped-strand mispairing, and transposition-mediated phase variation (1).

In summary, using a systematic computational approach, we predicted many loci across the bacterial domain to contain gene-altering programmed inversions and identified characteristic genomic sequence attributes and associated gene families. Furthermore, we found programmed inversions targeting a protein of unknown function, as well as a presumable PilV and phage tail collar fusion-gene. Most importantly, we revealed Type II restriction-modification genes to be major targets of programmed inversions. Gene-altering programmed inversions seem widespread, providing a rapid diversification mechanism across phyla and gene functions.

DATA AVAILABILITY

All Python scripts required to produce our results are available at https://github.com/Technion-Kishony-lab/gene_altering_programmed_inversions. No experimental data was produced in this study.

All scanned species accessions, identified gene-altering programmed inversion loci including programmed inversion candidates, enriched gene families, as well as long reads validating programmed inversions, are available in the supplementary tables.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Rotem Sorek, Debbie Lindell and members of the Kishony Lab for fruitful discussions.

FUNDING

Israel Science Foundation [3055/19], within the Israel Precision Medicine Partnership program; Israel Science Foundation – Broad Institute Joint Program [2790/19]. Funding for open access charge: Israel Science Foundation [3055/19], within the Israel Precision Medicine Partnership program.

Conflict of interest statement. None declared.

REFERENCES

1. Woude,M.W., van der Woude,M.W. and Bäumler,A.J. (2004) Phase and antigenic variation in bacteria. *Clin. Microbiol. Rev.*, **17**, 581–611.
2. Moxon,R., Bayliss,C. and Hood,D. (2006) Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu. Rev. Genet.*, **40**, 307–333.
3. Zhou,K., Aertsen,A. and Michiels,C.W. (2014) The role of variable DNA tandem repeats in bacterial adaptation. *FEMS Microbiol. Rev.*, **38**, 119–141.
4. Srikhanta,Y.N., Fox,K.L. and Jennings,M.P. (2010) The phasevarion: phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes. *Nat. Rev. Microbiol.*, **8**, 196–206.
5. Porter,N.T., Hryckowian,A.J., Merrill,B.D., Fuentes,J.J., Gardner,J.O., Glowacki,R.W.P., Singh,S., Crawford,R.D., Snitkin,E.S., Sonnenburg,J.L. *et al.* (2020) Phase-variable capsular polysaccharides and lipoproteins modify bacteriophage susceptibility in Bacteroides thetaiotaomicron. *Nat Microbiol*, **5**, 1170–1181.

6. Furi,L., Crawford,L.A., Rangel-Pineros,G., Manso,A.S., De Ste Croix,M., Haigh,R.D., Kwun,M.J., Engelsen Fjelland,K., Gilfillan,G.D., Bentley,S.D. *et al.* (2019) Methylation warfare: Interaction of pneumococcal bacteriophages with their host. *J. Bacteriol.*, **201**, e00370-19.
7. Yan,W., Hall,A.B. and Jiang,X. (2022) Bacteroidales species in the human gut are a reservoir of antibiotic resistance genes regulated by invertible promoters. *NPJ Biofilms Microbiomes*, **8**, 1.
8. Jiang,X., Hall,A.B., Arthur,T.D., Plichta,D.R., Covington,C.T., Poyet,M., Crothers,J., Moses,P.L., Tolonen,A.C., Vlamakis,H. *et al.* (2019) Invertible promoters mediate bacterial phase variation, antibiotic resistance, and host adaptation in the gut. *Science*, **363**, 181–187.
9. Seib,K.L., Srikhanta,Y.N., Atack,J.M. and Jennings,M.P. (2020) Epigenetic regulation of virulence and immunoevasion by phase-Variable restriction-Modification systems in bacterial pathogens. *Annu. Rev. Microbiol.*, **74**, 655–671.
10. Manso,A.S., Chai,M.H., Atack,J.M., Furi,L., De Ste Croix,M., Haigh,R., Trappetti,C., Ogunniyi,A.D., Shewell,L.K., Boitano,M. *et al.* (2014) A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. *Nat. Commun.*, **5**, 5055.
11. Phillips,Z.N., Trappetti,C., Van Den Bergh,A., Martin,G., Calcutt,A., Ozberk,V., Guillon,P., Pandey,M., von Itzstein,M., Swords,W.E. *et al.* (2022) Pneumococcal phasevarions control multiple virulence traits, including vaccine candidate expression. *Microbiol. Spectr.*, **10**, e00916-22.
12. Trzilova,D. and Tamayo,R. (2021) Site-Specific recombination – How simple DNA inversions produce complex phenotypic heterogeneity in bacterial populations. *Trends Genet.*, **37**, 59–72.
13. Komano,T. (1999) Shufflons: multiple inversion systems and integrons. *Annu. Rev. Genet.*, **33**, 171–191.
14. Sitaraman,R., Denison,A.M. and Dybvig,K. (2002) A unique, bifunctional site-specific DNA recombinase from Mycoplasma pulmonis. *Mol. Microbiol.*, **46**, 1033–1040.
15. Chambaud,I., Heilig,R., Ferris,S., Barbe,V., Samson,D., Galisson,F., Moszer,I., Dybvig,K., Wróblewski,H., Viari,A. *et al.* (2001) The complete genome sequence of the murine respiratory pathogen Mycoplasma pulmonis. *Nucleic Acids Res.*, **29**, 2145–2153.
16. Li,J.-W., Li,J., Wang,J., Li,C. and Zhang,J.-R. (2019) Molecular mechanisms of HsdSinversions in thecod locus of Streptococcus pneumoniae. *J. Bacteriol.*, **201**, e00581-18.
17. Huang,X., Wang,J., Li,J., Liu,Y., Liu,X., Li,Z., Kurniyati,K., Deng,Y., Wang,G., Ralph,J.D. *et al.* (2020) Prevalence of phase variable epigenetic invertons among host-associated bacteria. *Nucleic Acids Res.*, **48**, 11468–11485.
18. Atack,J.M., Guo,C., Litfin,T., Yang,L., Blackall,P.J., Zhou,Y. and Jennings,M.P. (2020) Systematic analysis of REBASE identifies numerous type I restriction-Modification systems with duplicated, distinct HsdSspecificity genes that can switch system specificity by recombination. *Msystems*, **5**, e00497-20.
19. Dybvig,K., Sitaraman,R. and French,C.T. (1998) A family of phase-variable restriction enzymes with differing specificities generated by high-frequency gene rearrangements. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 13923–13928.
20. Phillips,Z.N., Husna,A.-U., Jennings,M.P., Seib,K.L. and Atack,J.M. (2019) Phasevarions of bacterial pathogens - phase-variable epigenetic regulators evolving from restriction-modification systems. *Microbiology*, **165**, 917–928.
21. Allard,N., Neil,K., Grenier,F. and Rodrigue,S. (2022) The type IV Pilus of Plasmid TP114 displays adhesins conferring conjugation specificity and is important for DNA transfer in the mouse gut microbiota. *Microbiol. Spectr*, **10**, e0230321.
22. Grundy,F.J. and Howe,M.M. (1984) Involvement of the invertible G segment in bacteriophage mu tail fiber biosynthesis. *Virology*, **134**, 296–317.
23. Zabelkin,A., Yakovleva,Y., Bochkareva,O. and Alexeev,N. (2021) PaReBrick: PArallel rearrangements and BReaks identification toolkit. *Bioinformatics*, **38**, 357–363.
24. Goldberg,A., Fridman,O., Ronin,I. and Balaban,N.Q. (2014) Systematic identification and quantification of phase variation in commensal and pathogenic Escherichia coli. *Genome Med.*, **6**, 112.
25. Kuwahara,T., Yamashita,A., Hirakawa,H., Nakayama,H., Toh,H., Okada,N., Kuhara,S., Hattori,M., Hayashi,T. and Ohnishi,Y. (2004) Genomic analysis of Bacteroides fragilis reveals extensive DNA inversions regulating cell surface adaptation. *Proc. Natl. Acad. Sci. USA*, **101**, 14919–14924.
26. Shkoporov,A.N., Khokhlova,E.V., Stephens,N., Hueston,C., Seymour,S., Hryckowian,A.J., Scholz,D., Ross,R.P. and Hill,C. (2021) Long-term persistence of crAss-like phage crAss001 is associated with phase variation in Bacteroides intestinalis. *BMC Biol.*, **19**, 163.
27. Sekulovic,O., Mathias Garrett,E., Bourgeois,J., Tamayo,R., Shen,A. and Camilli,A. (2018) Genome-wide detection of conservative site-specific recombination in bacteria. *PLos Genet.*, **14**, e1007332.
28. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
29. Darling,A.E., Mau,B. and Perna,N.T. (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*, **5**, e11147.
30. Rognes,T., Flouri,T., Nichols,B., Quince,C. and Mahé,F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2584.
31. Marchler-Bauer,A., Anderson,J.B., Cherukuri,P.F., DeWeese-Scott,C., Geer,L.Y., Gwadz,M., He,S., Hurwitz,D.I., Jackson,J.D., Ke,Z. *et al.* (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res.*, **33**, D192–D196.
32. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
33. Zaworski,J., Guichard,A., Fomenkov,A., Morgan,R.D. and Raleigh,E.A. (2021) Complete annotated genome sequence of the Salmonella enterica serovar typhimurium LT7 strain STK003, historically used in gene transfer studies. *Microbiol. Resour. Announc.*, **10**, e01217-20.
34. Dybvig,K. and Yu,H. (1994) Regulation of a restriction and modification system via DNA inversion in Mycoplasma pulmonis. *Mol. Microbiol.*, **12**, 547–560.
35. Tettelin,H., Nelson,K.E., Paulsen,I.T., Eisen,J.A., Read,T.D., Peterson,S., Heidelberg,J., DeBoy,R.T., Haft,D.H., Dodson,R.J. *et al.* (2001) Complete genome sequence of a virulent isolate of Streptococcus pneumoniae. *Science*, **293**, 498–506.
36. Atack,J.M., Weinert,L.A., Tucker,A.W., Husna,A.U., Wileman,T.M., Hadjirin,FN., Hoa,N.T., Parkhill,J., Maskell,D.J., Blackall,P.J. *et al.* (2018) Streptococcus suis contains multiple phase-variable methyltransferases that show a discrete lineage distribution. *Nucleic Acids Res.*, **46**, 11466–11476.
37. Ben-Assa,N., Coyne,M.J., Fomenkov,A., Livny,J., Robins,W.P., Muniesa,M., Carey,V., Carasso,S., Gefen,T., Jofre,J. *et al.* (2020) Analysis of a phase-variable restriction modification system of the human gut symbiont Bacteroides fragilis. *Nucleic Acids Res.*, **48**, 11040–11053.
38. Cerdeño-Tárraga,A.M., Patrick,S., Crossman,L.C., Blakely,G., Abratt,V., Lennard,N., Poxton,I., Duerden,B., Harris,B., Quail,M.A. *et al.* (2005) Extensive DNA inversions in the *B. fragilis* genome control variable gene expression. *Science*, **307**, 1463–1465.
39. Nakayama-Imaohji,H., Hirakawa,H., Ichimura,M., Wakimoto,S., Kuhara,S., Hayashi,T. and Kuwahara,T. (2009) Identification of the site-specific DNA invertase responsible for the phase variation of SusC/SusD family outer membrane proteins in Bacteroides fragilis. *J. Bacteriol.*, **191**, 6003–6011.
40. Sekizuka,T., Kawanishi,M., Ohnishi,M., Shima,A., Kato,K., Yamashita,A., Matsui,M., Suzuki,S. and Kuroda,M. (2017) Elucidation of quantitative structural diversity of remarkable rearrangement regions, shufflons, in IncI2 plasmids. *Sci. Rep.*, **7**, 928.
41. Ishiwa,A. and Komano,T. (2000) The lipopolysaccharide of recipient cells is a specific receptor for PilV proteins, selected by shufflon DNA rearrangement, in liquid matings with donors bearing the R64 plasmid. *Mol. Gen. Genet.*, **263**, 159–164.
42. Goldfarb,T., Sberro,H., Weinstock,E., Cohen,O., Doron,S., Charpak-Amikam,Y., Afik,S., Ofir,G. and Sorek,R. (2015) BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J.*, **34**, 169–183.
43. Price,M.N. and Arkin,A.P. (2017) PaperBLAST: Text mining papers for information about homologs. *Msystems*, **2**, e00039-17.
44. Ofir,G., Melamed,S., Sberro,H., Mukamel,Z., Silverman,S., Yaakov,G., Doron,S. and Sorek,R. (2018) DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nat Microbiol*, **3**, 90–98.

45. Anjum,A., Brathwaite,K.J., Aidley,J., Connerton,P.L., Cummings,N.J., Parkhill,J., Connerton,I. and Bayliss,C.D. (2016) Phase variation of a Type IIG restriction-modification enzyme alters site-specific methylation patterns and gene expression in Campylobacter jejuni strain NCTC11168. *Nucleic Acids Res.*, **44**, 4581–4594.
46. Amobonye,A., Singh,S., Mukherjee,K., Jobichen,C., Qureshi,I.A. and Pillai,S. (2022) Structural and functional insights into fungal glutaminase using a computational approach. *Process Biochem.*, **117**, 76–89.
47. Joyce,S.A. and Dorman,C.J. (2002) A rho-dependent phase-variable transcription terminator controls expression of the FimE recombinase in Escherichia coli. *Mol. Microbiol.*, **45**, 1107–1117.
48. Anton,B.P. and Roberts,R.J. (2021) Beyond restriction modification: Epigenomic roles of DNA methylation in prokaryotes. *Annu. Rev. Microbiol.*, **75**, 129–149.
49. Atack,J.M., Tan,A., Bakaletz,L.O., Jennings,M.P. and Seib,K.L. (2018) Phasevarions of bacterial pathogens: Methylomics sheds new light on old enemies. *Trends Microbiol.*, **26**, 715–726.
50. Kinch,L.N., Ginalski,K., Rychlewski,L. and Grishin,N.V. (2005) Identification of novel restriction endonuclease-like fold families among hypothetical proteins. *Nucleic Acids Res.*, **33**, 3598–3605.
51. Pleška,M., Qian,L., Okura,R., Bergmiller,T., Wakamoto,Y., Kussell,E. and Guet,C.C. (2016) Bacterial autoimmunity due to a restriction-Modification system. *Curr. Biol.*, **26**, 404–409.
52. Islam,M.Z., Fokine,A., Mahalingam,M., Zhang,Z., Garcia-Doval,C., van Raaij,M.J., Rossmann,M.G. and Rao,V.B. (2019) Molecular anatomy of the receptor binding module of a bacteriophage long tail fiber. *PLoS Pathog.*, **15**, e1008193.
53. North,O.I., Sakai,K., Yamashita,E., Nakagawa,A., Iwazaki,T., Büttner,C.R., Takeda,S. and Davidson,A.R. (2019) Phage tail fibre assembly proteins employ a modular structure to drive the correct folding of diverse fibres. *Nat. Microbiol.*, **4**, 1645–1653.
54. Gray,D.A., White,J.B.R., Oluwole,A.O., Rath,P., Glenwright,A.J., Mazur,A., Zahn,M., Baslé,A., Morland,C., Evans,S.L. *et al.* (2021) Insights into SusCD-mediated glycan import by a prominent gut symbiont. *Nat. Commun.*, **12**, 44.
55. Darphorn,T.S., Koenders-van Sintanneland,B.B., Grootemaat,A.E., van der Wel,N.N., Brul,S. and Ter Kuile,B.H. (2022) Transfer dynamics of multi-resistance plasmids in Escherichia coli isolated from meat. *PLoS One*, **17**, e0270205.