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

Research thesis

In partial fulfillment of the requirement for the

degree of Master of Science in Biology

Oren Milman

Submitted to the Senate of the Technion - Israel Institute of Technology

Nisan 5782, Haifa, April 2022

The research thesis was done under the supervision of Prof. Roy Kishony in the Faculty of Biology.

The generous financial help of the Leonard and Diane Sherman Interdisciplinary Graduate School Fellowship, as well as the Israel Science Foundation is gratefully acknowledged.

Table of Contents

List of Figures and Tables

Abstract

Programmed chromosomal inversions allow bacteria to generate intra-population genotypic and functional heterogeneity, a bet-hedging strategy important in changing environmental challenges, such as bacteriophages and antibiotic drugs. Some programmed inversions modify coding sequences, producing different alleles in several gene families, most notably in specificity-determining genes such as phage-tail, conjugative pili and Type I restriction-modification systems, where systematic searches revealed cross phylum abundance. Yet, a broad systematic search for gene-altering programmed inversions, not guided by previously known gene families, has been absent, and little is known about their prevalence across gene families and their common genomic architectures. Here, scanning for 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 architectures, including gene-pseudogene size asymmetry and gene-pseudogene orientation such that the invertible region is shorter. The model predicted over 14,000 gene-altering loci covering known gene families as well as Type II restriction-modification systems previously not characterized for programmed inversions. Publicly available long-read sequencing datasets validated representatives of these predicted inversion-targeted gene families, confirming intra-population genetic heterogeneity, even with multiple co-existing combinatorial variants. 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.

List of symbols and abbreviations

ABC	ATP-binding cassette
bp	base pairs
CDS	coding sequence
DUF	domain of unknown function
IR	inverted repeat
kbp	kilo base pairs
MTase	methyltransferase
PIC	programmed inversion candidate
RM	restriction-modification
SNP	single nucleotide polymorphism
TSS	transcription start site

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 (1). Indeed, phase variation was observed in various bacterial processes (2–5) and was shown to be important for survival in major environmental challenges faced by bacteria, including bacteriophages (6, 7), antibiotic drugs (8, 9) and virulence (10–12).

The underlying mechanism of many phase variation systems is programmed chromosomal inversions (2, 13). These are frequent and surgical inversions of specific genomic regions flanked by inverted repeats (similar DNA sequences appearing on opposite strands) (13). Programmed inversions are catalyzed by recombinase enzymes, usually encoded near or inside the invertible region (13, 14). Yet some invertible regions lacking a nearby recombinase gene were also observed (15, 16), as well as programmed inversions that were not completely diminished upon deletion of the nearby recombinase gene (17). Programmed inversions often modify regulatory DNA sequences, typically inverting a promoter to switch a gene on or off (13). Conversely, gene-altering programmed inversions target coding sequences (**Figure 1A**), producing different alleles that would give rise to different protein variants (13).

The signature of inverted repeats flanking programmed inversions, as well as a 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, published as tools or applied ad hoc (8, 9, 13, 18–24). Only a few of these methods were applied to widely scan the bacterial domain, searching for invertible promoters (9), invertible promoters of antibiotic resistance genes (8), and programmed inversions targeting Type I restriction-modification (RM) specificity subunit HsdS (23, 24). One method was used to identify any programmed inversion, not limiting the search to promoters or specific target gene families, yet it was applied only to 209 genomes and involved a step in which identified putative inversions were scored manually (22). Thus, a systematic and wide search for gene-altering programmed inversions is still lacking.

The lack of a systematic search leaves some fundamental questions about gene-altering programmed inversions unanswered. First, it is unknown which loci are targeted by gene-altering programmed inversions. Systematically Identifying programmed inversions would not only highlight potentially crucial loci, but also provide a diverse dataset enabling application of statistical methods to tease out genomic architectures characterizing gene-altering programmed inversions. Inverted repeats are considered to be the only universal genomic attribute of programmed inversions (13), and recombinase genes often accompany programmed inversions (13, 14), yet, other, perhaps more subtle genomic architectures, are still unknown. Furthermore, identifying such genomic architectures would improve our ability to identify programmed inversions, where other types of evidence are missing or partial. Finally, it is unknown which gene families are targeted by programmed inversions, and in what genomic contexts. Revealing such gene families and genomic contexts might highlight central bacterial pathways and environmental challenges.

Here, seeking to shed light on these questions, 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. We then search for intra-species variation for each candidate, producing a diverse dataset of 128 putative gene-altering programmed inversions. Next, we identify genomic architectures enriched in these putative programmed inversions, and subsequently utilize these architectures to predict more programmed inversions, obtaining a large and diverse dataset of 14,226 predicted programmed inversions, revealing associated gene families. Finally, we find in publicly available long-read sequencing data compelling evidence for gene-altering programmed inversions in selected loci representing different such predicted programmed inversions. We thus identify and validate 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

The bacterial NCBI RefSeq assembly summary file (ftp://ftp.ncbi.nih.gov/genomes/refseq/bacteria/assembly_summary.txt, retrieved January 10th 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). Overall, representative genomes of 35,356 species were retrieved. Non-continuous GenBank Coding Sequences (CDSs) were discarded if the total distance between parts was >100bp.

Inverted repeat identification and CDS assignment

For each scaffold in each representative genome, blastn version 2.12 (25) 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 such that the length of the region flanked by the repeats was >0 and <=15kbp. Inverted repeats strictly contained in other inverted repeats were discarded. Next, inverted repeats with repeats shorter than 22bp 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 196,564 CDS pairs with any linked inverted repeat pair.

Repetitive inverted-repeat filtering

For each pair of inverted repeats, one repeat was arbitrarily chosen and blastn version 2.12 was run locally to find alignments between the repeat and any sequence in the representative genome of the same species, using the following arguments:

-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 50kbp away from the inverted repeat pair. If any base pair in the repeat was part of 3 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 120,639 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 PICs, 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 loci in same-species genomes

For each PIC, we define its region to be the smallest region that contains its CDS pair as well as all inverted repeats linked to its CDS pair. Overlapping PIC regions were merged to form PIC loci.
We define the left and right margins of a PIC locus as the 200bp regions flanking 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 same-species genomes. For each PIC locus, blastn version 2.12 was used to find alignments between the locus margins and chosen same-species genomes in the BLAST nt database and in downloaded WGS Nucleotide entries (see above), using the following arguments:

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

For the BLAST nt database search, *-taxids* was also used, to specify chosen taxa (see above).
For each same-species genome scaffold and each margin, 20 alignments with lowest evalue were retained. Left and right margin alignments to the same scaffold and strand were paired to form alignment pairs, and non-collinear alignment pairs (namely, alignments to different strands or in opposite order) were discarded. For each alignment pair and each scaffold, we define the alignment region to be the shortest region in the scaffold spanning both alignments. If the lengths of the alignment regions in the same-species genome and in the PIC locus are similar, then the alignment region in the same-species genome might be homologous to the PIC locus or to another variant of the PIC locus. We thus define the relative locus length difference to be the absolute difference between the lengths of alignment regions in the PIC locus and in the same-species genome, normalized to the length of the alignment region in the PIC locus. Alignment pairs with a relative locus length difference above 0.05 were discarded (**Supplementary Figure S2**). For each PIC locus, alignment pairs with overlapping alignment regions in the same-species genome were grouped, and for each group, the alignment pair with smallest relative locus length difference was chosen as group representative. This produced a set of similar length loci in same-species genomes for each PIC locus.

Identification of programmed inversion candidate variants

For each PIC locus, similar length loci were sorted by relative locus length difference (smallest first); loci identical to the PIC locus were discarded, and at most first 100 of remaining loci were each aligned to the PIC locus using progressiveMauve build date 13 February 2015 (26). We define the match proportion of an alignment to be 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 similar loci.

For each similar 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 similar locus regions on different strands; and (b) The consecutive sub-alignments were not consecutive if sorted according to the aligned region location in the similar locus.
Finally, each pair of inverted repeats linked to a PIC was examined to determine whether for any similar locus, each repeat is at most 10bp away from a different breakpoint-containing region of that similar locus (**Supplementary Figure S5**). The 128 PICs linked to any such inverted repeat pair were marked as PICs with intra-species variation.

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 (27), 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 7 and 23, respectively, and the left and right CDSs of PIC B belong to clusters 23 and 7, respectively) were considered to be of the same PIC cluster.

Genomic architecture definitions

Transcription units were predicted by grouping consecutive CDSs on the same strand such that the maximal distance between consecutive CDSs was 20bp.
We define o_L (“outer”) to be the length of the part of the left transcription unit left to the leftmost left repeat. Mirroring o_L, we define o_R to be the length of the part of the right transcription unit right to the rightmost right repeat. Similarly, we define i_L (“inner”) to be the length of the part of the left transcription unit right to the rightmost left repeat. Mirroring i_L, we define i_R to be the length of the part of the right transcription unit left to the leftmost right repeat. Furthermore, we define u_L (“upstream”) to be the length of the part of the left transcription unit upstream to the most upstream repeat (of any inverted repeat pair linked to the PIC). u_R is defined similarly for the right transcription unit (**Figure 2A**).

We define four genomic architecture 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 transcription units; (c) *orientation matching* = $\frac{o_L+o_R}{o_L+o_R+i_L+i_R}$, a measure of matching between the orientation of the two transcription units (head-to-head or tail-to-tail) and repeat positions inside transcription units, while we consider a matching to be better (i.e., higher orientation matching) in case switching the transcription unit 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 transcription units upstream to the most upstream repeats (**Supplementary Figure S7**).

Genomic architecture enrichment analysis

Out of PICs with at least one similar locus, for each PIC cluster, one PIC was chosen randomly as the cluster representative. All subsequent steps for identifying gene-altering programmed inversion genomic architectures, 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 architecture measures (*CDS distance*, *repeat length*, *asymmetry*, *orientation matching*), 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 architecture measures, giving rise to four genomic architectures: long repeats (*repeat length* >=44bp), short CDS distance (*CDS distance*<=2588bp), high orientation matching (*orientation matching*>=0.58), and high asymmetry (*asymmetry*>=0.87).

Gene-altering programmed inversion prediction

A logistic regression model was trained on PIC cluster representatives (representatives of PICs with at least one similar locus), using Python's statsmodels.api.Logit.fit. For each PIC, the model was provided with the four genomic architectures (low CDS distance, high repeat length, high asymmetry, and high orientation matching) 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 (trained on all PIC cluster representatives) was used to predict gene-altering programmed inversions for all PICs, using Python's statsmodels.discrete.discrete_model.BinaryResults.predict.

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. 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, PICs with programmed inversion prediction probabilities (see above) >=0.05 were marked as predicted programmed inversions (**Supplementary Figure S8**). Then, 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 test for independence between the PIC variables corresponding to the following questions: "Is it the annotation of any target CDS of the PIC?" and "Was

the PIC predicted to be a programmed inversion?". The tests were performed using scipy.stats.fisher_exact or scipy.stats.chi2_contingency (with *lambda_*="log-likelihood"). 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 test for independence between the PIC locus variables corresponding to the following questions: "Is it the annotation of any neighbor of any target CDS of the PIC?" and "Was the PIC predicted to be a programmed inversion?". As before, a bonferroni correction was used to obtain corrected p-values.

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

Predicted programmed inversions targeting gene families found to be enriched (see above) were scanned semi-manually (i.e., we didn't just look at a table of all genomic contexts. Rather, we examined the results of different calls to pd.DataFrame.value_counts(), each time for a different set of nearby CDS product annotations, etc) for recurring genomic contexts, namely, gene content and organization of the predicted programmed inversion locus. For some loci containing such genomic contexts, the NCBI Nucleotide and SRA Databases were manually searched for Nucleotide entries that both contain similar loci 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 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 sequencing data

For each of these loci (for which NCBI Nucleotide, SRA and Assembly entries were found and retrieved), a region hypothesized to contain the programmed inversion was assigned manually. First, inverted repeats in this region were identified using blastn version 2.12 locally 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 20bp long or longer.

Next, blastn version 2.12 was run locally to find alignments between the region hypothesized to contain the programmed inversion, including margins of at least 10kbp on each side, and extracted long reads, with the following arguments:

-strand both *-max_target_seqs* 100000000 *-word_size* 8 *-evalue* 0.0001
-window_size 0 *-dust* no

Reads were filtered to include only reads with at least 2 kbp of alignment to the genome. Subsequently, 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 (which were identified beforehand, 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.

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 500bp to the left/right of the leftmost/rightmost inverted repeat, such that these alignments cover a continuous region in the read. Of these reads, those with collinear alignments (i.e., alignments such that with more relaxed mismatch and indel thresholds would be identified as a single long alignment) that together completely cover that region were marked as matching the reference variant, namely, they matched the NCBI Nucleotide entry, 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 in the manual examination. 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). blastn version 2.12 was run locally to align these truncated reads to the whole reference genome (that is, the NCBI Assembly entry linked to the Nucleotide entry containing the locus), with the following arguments:

-strand both *-max_target_seqs* 100000000 *-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 beas 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. Remaining reads, with their assignment to the reference variant or another variant (defined by its inverted repeats), were counted and plotted in **Figure 4A**, **Figure 4C** and **Supplementary Figures S9-S29**.

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 progressiveMauve build date 13 February 2015. Finally, matching positions were extracted from the mauve .xmfa file, and were plotted in **Figure 4** 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 product annotations and product annotations of genes in programmed inversion loci (see above), were inspected manually and assigned to broad gene families. 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 (28)), or according to protein sequence similarity (assessed by BLAST). For most of these CDSs, the GenBank product annotation was "hypothetical protein".

Distribution of programmed inversions across phyla

For each programmed inversion locus, blastn version 2.12 was run locally to find alignments between the longest target CDS and any sequence in any representative genome, using the following arguments:

-*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. For each such homologous sequence, the CDS covering the most base pairs of the homologous sequence was linked to it. 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, blastn version 2.12 was run locally to find alignments between each of these homologs and the two 10kbp regions flanking it, using the following arguments:

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

Each homolog with any alignment with evalue<=1e-6 was marked as a homolog potentially targeted by programmed inversions.

Results

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

To identify bacterial gene-altering programmed inversions in a systematic and wide manner, we started by compiling a set of programmed inversion candidates, based on coding sequence (CDS) and inverted repeat positions (**Figure 1B**, top). First, for each of 35,356 bacterial species, we retrieved one representative genome from the NCBI RefSeq database. We then identified inverted repeats (IRs) in each scaffold and sought out those of them that might promote gene-altering programmed inversions. To this end, we 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. These tests revealed 196,564 CDS pairs with at least one pair of IRs linked to them. Finally, we reasoned that many of these IRs are part of mobile elements, rather than programmed inversions; therefore, we next discarded CDS pairs containing repeats with multiple copies in the genome, in addition to copies residing in the IR pair locus (**Supplementary Figure S1**). The remaining 120,639 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 variants of PIC loci that may arise from programmed inversions (**Figure 1B**, bottom). First, we assigned each PIC to a PIC locus, which we define as the smallest locus containing the PIC such that the locus length would be the same in all variants potentially arising from programmed inversions. PIC loci were estimated by grouping overlapping PICs. Then, for each PIC locus, we searched other genomes of the same species for loci flanked by sequences homologous 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 arising from programmed inversions. This search identified at least one similar locus for PIC loci of 6370 PICs, while for 128 of these PICs, at least one similar 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 similar loci exhibiting different variants revealed high sequence similarity (**Figure 1C**), suggesting that observed rearrangements occurred relatively recently.

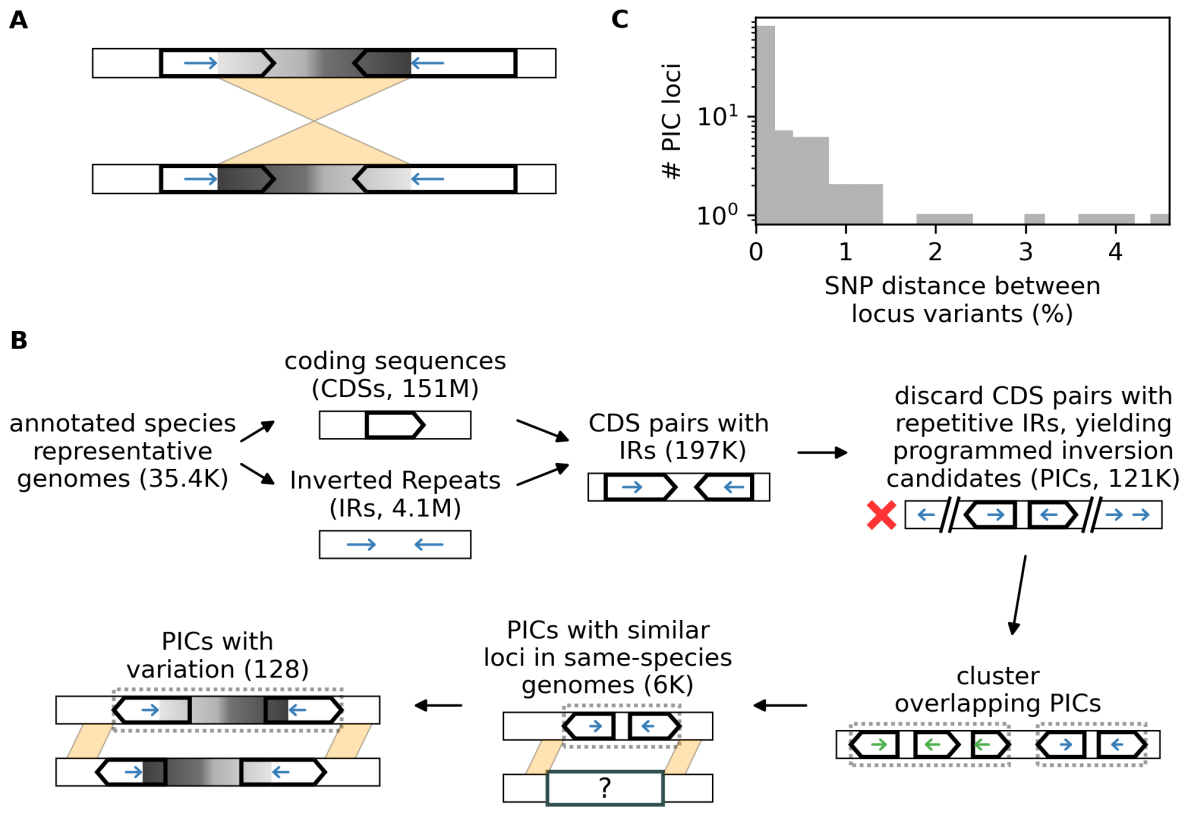


Figure 1. A pipeline for identifying intra-species variation, revealing putative gene-altering programmed inversions.
(**A**) Schematic illustration of two variants of a locus containing a gene-altering programmed inversion. Inverted repeats and coding sequences are indicated as colored stick arrows 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 (given by $100 - percent\ identity$) between PIC loci and similar 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 programmed inversion candidate variants').

Genomic architectures associated with identified programmed inversions

In order to reveal genomic architectures associated with gene-altering programmed inversions, we sought out architectures enriched in PICs with, compared to those without, identified intra-species variation. We identified four quantitative genomic architecture characteristics with statistically significantly different distributions in the putative intra-species varying PICs (**Figure 2A**). First, hypothesizing that programmed inversions require relatively long IRs, we defined the length of the longest repeat linked to a PIC as the "repeat length" measure and compared its distribution between PICs with and without intra-species variation. Indeed, a Mann–Whitney U test revealed statistically significantly longer repeats in PICs with intra-species variation (**Figure 2B**). Next, hypothesizing that programmed inversions are

more efficient for shorter invertible regions, we defined the distance between transcription units (for transcription unit definition and prediction, see **Materials and Methods**, 'Genomic architecture definitions') of PIC CDSs as the "CDS distance" measure and examined its distribution. As expected, a Mann–Whitney U test revealed statistically significantly shorter CDS distances in PICs with intra-species variation (**Figure 2C**).

The third identified genomic architecture is concerned with the orientation of candidate transcription units relative to each other, i.e., head-to-head or tail-to-tail. We noticed that typically, for PICs with transcription units in a tail-to-tail configuration, IRs were situated close to the transcription start site (TSS), such that the shortest region flanked by IRs was relatively short. If these transcription units had instead exhibited a head-to-head configuration, with IRs situated close to the TSS, the shortest region flanked by IRs would have been longer (**Supplementary Figure S6A**). Likewise, for PIC with transcription units oriented head-to-head, IRs were situated close to the termination site. If these transcription units had instead been oriented tail-to-tail, the shortest region flanked by IRs would have been longer (**Supplementary Figure S6B**). We hence defined the "orientation matching" measure, to quantify the extent to which the observed transcription unit orientation matches the position of IRs inside the transcription units, such that IRs flank shorter regions, relative to the opposite hypothetical transcription unit orientation. Indeed, a Mann–Whitney U test showed statistically significantly higher orientation-matching values in PICs with intra-species variation (**Figure 2D**).

The last identified genomic architecture involves an asymmetry in lengths of PIC transcription units. Similar to previous observations in gene-altering programmed inversions (14, 24, 29, 30), we observed, in many PICs with identified intra-species variation, one transcription unit to lack a region containing the TSS. Specifically, the upstream edge of the transcription unit retained part often coincided with the IR closest to the TSS. We thus defined the "asymmetry" measure, to quantify both the length of the transcription unit truncated part and its proximity to the IR closest to the TSS (**Supplementary Figure S7**). A Mann–Whitney U test revealed statistically significantly higher asymmetry values in PICs with intra-species variation (**Figure 2E**).

Next, to estimate the relative capacity of these genomic architectures to predict programmed inversions, and whether they are independent of each other, we used a logistic regression model. PICs for which at least one similar locus was found were provided to the model as inputs. Striving for a simple model, such that relative importance of different predictors could be easily examined, we used only binary predictors. For each PIC, the model received binarized versions of the four genomic architecture measures - long repeats, short CDS distance, high orientation-matching, and high asymmetry - as predictor variables, and whether intra-species variation was identified as the response variable. The model revealed the high asymmetry predictor to have the biggest contribution, but also that each other genomic architecture is not entirely dependent on other genomic architectures (**Figure 2F**).

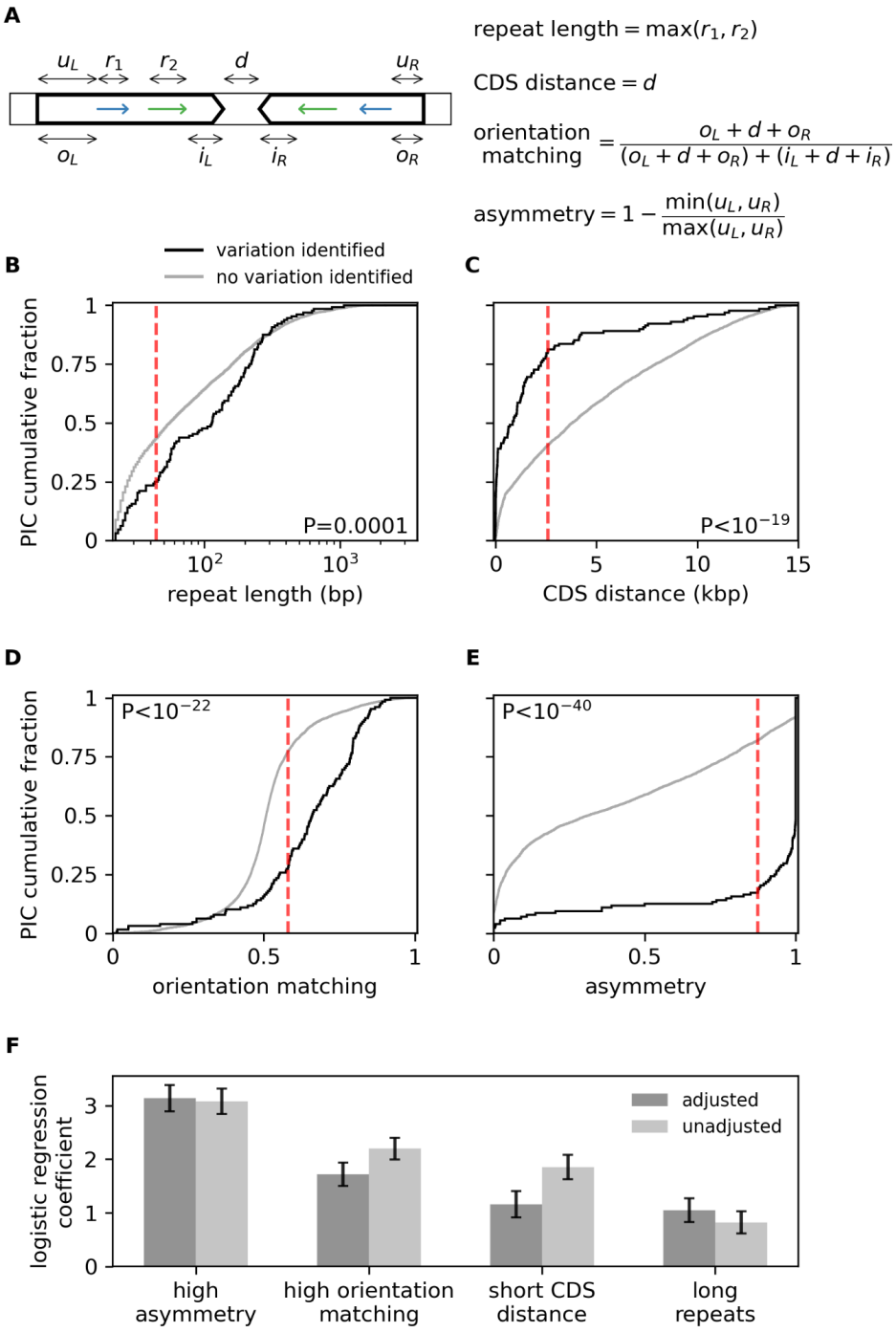


Figure 2. Putative programmed inversions are enriched for several genomic architectures. (**A**) Schematic illustration of genomic architecture measure definitions. d is the distance between coding sequences (or more generally among transcription units containing these coding sequences, not shown); r_1 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 outer regions relative to repeats; i_L and i_R are the lengths of coding sequence inner regions relative to repeats. (**B,C,D,E**) Cumulative distribution functions of each genomic architecture measure, shown for programmed inversion candidates (PICs) with identified intra-species variation (n=128), and for PICs with at least one similar locus but

without identified intra-species variation (n=5964). p-values were produced by two-sided Mann–Whitney U tests. Dashed red lines indicate values corresponding to Kolmogorov–Smirnov statistics, which were used to binarize genomic architecture measures. (**F**) Logistic regression coefficients produced by a model trained to predict for PICs with at least one similar locus whether intra-species variation was identified. "Adjusted" coefficients were produced by a model using all four genomic architecture binarized measures as predictors, while "unadjusted" coefficients were produced by models using a single genomic architecture binarized measure as the only predictor. Error bars indicate coefficient values ± 1SE. Abbreviations: CDS, coding sequence; PIC, programmed inversion candidate.

Gene families associated with predicted programmed inversions

Most PICs lacked similar loci in same-species genomes, such that intra-species variation could not be detected for them; thus, we used identified genomic architectures to predict which of these PICs were programmed inversions. To this end, we utilized the logistic regression model that was trained on PICs with at least one similar locus, obtaining a predicted probability for each PIC. Using 0.05 as a cutoff (**Supplementary Figure S8**), we marked PICs with higher predicted probability as predicted programmed inversions.

Leveraging predicted programmed inversions, we next sought to identify gene families targeted by programmed inversions. We scanned GenBank product annotations of CDS potentially targeted by PICs for annotations appearing in at least 12 PICs. 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 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 (23, 24, 31–34), phage tail (29, 30), TonB-linked outer membrane protein (6, 35, 36), and Shufflon PilV (14, 37, 38), while some annotations belonged to proteins of unknown function. Multiple statistically significantly enriched annotations were characteristic of Type II RM enzymes, suggesting this family to be a major target of programmed inversions. Yet, to the best of our knowledge, only two studies (21, 39) found evidence for programmed inversions targeting genes encoding Type II RM enzymes.

Using the same approach, we attempted to find gene families associated with, but not targeted by, programmed inversions. A two-sided Fisher's exact test or G-test was performed for each GenBank product annotation appearing in at least 12 PICs in CDS neighboring potential PIC target CDS, to find whether the annotation is enriched in neighbor CDS 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 (2, 13, 14, 22, 23), multiple statistically significantly enriched annotations belonged to recombinase families. Curiously, the three depleted annotations with lowest corrected p-values all belonged to the ATP-binding cassette (ABC) transporter family.

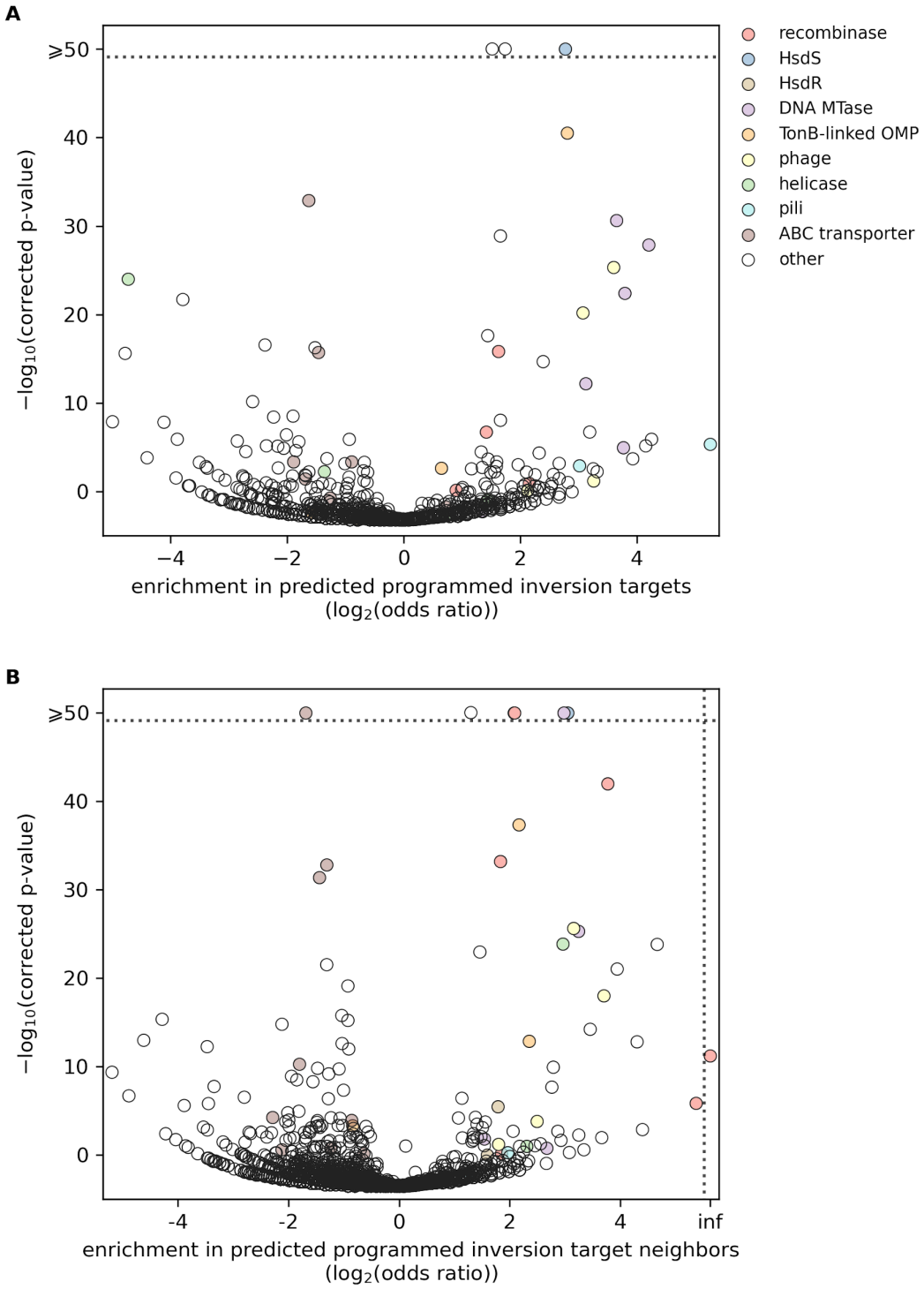


Figure 3. Programmed inversions predicted by identified genomic architectures highlight associated gene families.

(**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 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	38.040	4.32E-06
DUF4393 domain-containing protein	19.024	1.23E-06
BREX-1 system adenine-specific DNA-methyltransferase PglX	18.414	1.26E-28
DUF4965 domain-containing protein	17.754	6.80E-06
DUF1793 domain-containing protein	15.215	1.97E-04
N-6 DNA methylase	13.822	4.07E-23
type II restriction endonuclease	13.588	1.03E-05
Eco57I restriction-modification methylase domain-containing protein	12.552	2.44E-31
tail fiber protein	12.109	4.60E-26
endonuclease domain-containing protein	9.961	5.72E-03
DUF559 domain-containing protein	9.509	2.68E-03
DUF736 domain-containing protein	9.120	1.89E-07
class I SAM-dependent DNA methyltransferase	8.703	6.37E-13
phage tail protein	8.384	6.12E-21
pilin	8.069	1.16E-03
SusC/RagA family TonB-linked outer membrane protein	6.955	3.19E-41
restriction endonuclease subunit S	6.824	3.37E-187
carboxypeptidase-like regulatory domain-containing protein	6.341	6.61E-04
peptidase C14	6.339	3.38E-02
HtaA domain-containing protein	5.917	1.18E-02
IS5/IS1182 family transposase	5.242	2.10E-15
ISAs1 family transposase	4.999	4.40E-05
transposase domain-containing protein	4.756	1.25E-02
Rpn family recombination-promoting nuclease/putative transposase	4.296	6.25E-03
type VI secretion system tip protein VgrG	3.893	9.31E-04
transposase	3.334	1.92E-85
IS630 family transposase	3.154	8.84E-09
IS3 family transposase	3.153	1.33E-29
tyrosine-type recombinase/integrase	3.070	1.38E-16
variable large family protein	3.058	2.98E-02
IS256 family transposase	3.019	4.45E-04

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

GenBank CDS product annotation	enrichment in predicted programmed inversion targets (odds ratio)	corrected p-value
IS21 family transposase	3.006	1.43E-03
IS30 family transposase	2.949	5.87E-03
hypothetical protein	2.859	0.00E+00
transposase family protein	2.811	6.79E-03
IS66 family transposase	2.721	2.13E-04
IS5 family transposase	2.712	2.43E-18
site-specific integrase	2.658	1.74E-07
RHS repeat-associated core domain-containing protein	2.539	1.87E-03
S-layer homology domain-containing protein	2.509	3.35E-05
IS6 family transposase	2.234	2.73E-03
TonB-dependent receptor	1.558	2.16E-03
glycosyltransferase	0.650	6.74E-03
MFS transporter	0.625	5.00E-04
response regulator	0.619	4.50E-03
ABC transporter permease	0.537	4.40E-04
ATP-binding cassette domain-containing protein	0.523	1.25E-06
sigma-70 family RNA polymerase sigma factor	0.521	2.35E-02
LysR family transcriptional regulator	0.496	7.04E-04
EAL domain-containing protein	0.401	1.86E-04
DEAD/DEAH box helicase	0.389	5.66E-03
ABC transporter substrate-binding protein	0.362	1.82E-16
acyl-CoA dehydrogenase family protein	0.348	5.65E-17
extracellular solute-binding protein	0.323	6.00E-03
ABC transporter ATP-binding protein	0.321	1.22E-33
flagellin	0.312	1.62E-02
sugar ABC transporter permease	0.310	3.80E-02
CDP-glycerol glycerophosphotransferase family protein	0.309	1.32E-02
PAS domain S-box protein	0.287	2.39E-06
PAS domain-containing protein	0.277	2.32E-05
sugar ABC transporter ATP-binding protein	0.269	4.25E-04
protein kinase	0.268	3.01E-09
bifunctional glycosyltransferase family 2 protein/CDP-glycerol:glycerophosphate glycerophosphotransferase	0.266	1.68E-02

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

GenBank CDS product annotation	enrichment in predicted programmed inversion targets (odds ratio)	corrected p-value
universal stress protein	0.248	3.87E-07
glycosyltransferase family 4 protein	0.240	1.29E-05
aldehyde dehydrogenase	0.225	2.13E-03
HAMP domain-containing histidine kinase	0.223	7.42E-06
tripartite tricarboxylate transporter substrate binding protein	0.212	3.84E-09
RNA polymerase sigma factor	0.195	6.76E-06
serine/threonine protein kinase	0.192	2.74E-17
glycerol kinase GlpK	0.166	6.95E-11
copper-translocating P-type ATPase	0.153	3.51E-02
ABC-F family ATP-binding cassette domain-containing protein	0.153	3.04E-05
endo-1,4-beta-xylanase	0.148	2.50E-02
TRAP transporter large permease subunit	0.147	1.64E-02
RdlA protein	0.138	1.93E-06
TRAP transporter substrate-binding protein	0.110	1.58E-02
arylsulfatase	0.097	2.28E-03
ATP-dependent Clp protease ATP-binding subunit	0.095	1.52E-03
acetyl-CoA C-acetyltransferase	0.088	4.99E-04
transcription-repair coupling factor	0.072	2.01E-22
heavy metal translocating P-type ATPase	0.068	1.22E-06
cyanophycin synthetase	0.067	2.93E-02
citrate synthase	0.058	1.46E-08
agmatine deiminase family protein	0.047	1.51E-04
ATP-dependent DNA helicase RecG	0.038	9.72E-25
arginine-ornithine antiporter	0.036	2.46E-16
glutamine synthetase	0.031	1.30E-08
PQQ-dependent dehydrogenase, methanol/ethanol family	0.000	4.66E-07
citrate synthase 2	0.000	3.30E-07
PQQ-dependent methanol/ethanol family dehydrogenase	0.000	1.38E-05
Xaa-Pro dipeptidase	0.000	2.41E-03
monovalent cation/H+ antiporter subunit A	0.000	8.19E-05
citrate synthase/methylcitrate synthase	0.000	1.81E-02
putative monovalent cation/H+ antiporter subunit A	0.000	1.81E-02
type I glutamate--ammonia ligase	0.000	2.90E-13

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

GenBank CDS product annotation	enrichment in predicted programmed inversion targets (odds ratio)	corrected p-value
MHS family MFS transporter	0.000	2.76E-02
TonB-dependent siderophore receptor	0.000	1.72E-06

Table 2. Product annotations 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
Tsr0667 family tyrosine-type DNA invertase	inf	5.94E-12
tyrosine-type DNA invertase PsrA	41.213	1.38E-06
BREX system P-loop protein BrxC	25.424	1.56E-24
DUF2612 domain-containing protein	21.129	1.31E-03
BREX-1 system phosphatase PglZ type A	19.748	1.60E-13
DUF1016 family protein	15.389	9.40E-22
master DNA invertase Mpi family serine-type recombinase	13.666	9.92E-43
phage tail protein I	13.012	9.38E-19
cell filamentation protein Fic	12.677	1.06E-02
virulence RhuM family protein	10.971	6.05E-15
DUF417 family protein	9.509	5.47E-03
SAM-dependent DNA methyltransferase	9.477	4.95E-26
tail fiber assembly protein	8.913	2.43E-26
restriction endonuclease subunit S	8.279	9.17E-118
type I restriction-modification system subunit M	7.865	4.55E-90
DEAD/DEAH box helicase family protein	7.793	1.41E-24
DUF2971 domain-containing protein	7.607	2.15E-02
PAAR domain-containing protein	7.316	2.04E-03
alginate lyase family protein	6.911	1.17E-10
glutamate-1-semialdehyde 2,1-aminomutase	6.785	2.10E-08
tail fiber protein	5.612	1.44E-04
SusD/RagB family nutrient-binding outer membrane lipoprotein	5.104	1.36E-13
RagB/SusD family nutrient uptake outer membrane protein	4.486	4.40E-38
tyrosine-type recombinase/integrase	4.248	7.76E-89
site-specific integrase	4.208	2.24E-110
putative DNA binding domain-containing protein	4.174	2.02E-03
recombinase family protein	3.548	6.25E-34
type I restriction endonuclease subunit R	3.437	3.18E-06
Fic family protein	2.946	1.82E-04
N-6 DNA methylase	2.908	1.52E-02
IS6 family transposase	2.844	7.90E-04
ISL3 family transposase	2.795	7.98E-03
IS3 family transposase	2.752	1.12E-23

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

GenBank CDS product annotation	enrichment in predicted programmed inversion target neighbors (odds ratio)	corrected p-value
S-layer homology domain-containing protein	2.661	1.08E-02
IS5/IS1182 family transposase	2.655	3.21E-04
IS66 family insertion sequence element accessory protein TnpB	2.603	5.35E-05
IS21 family transposase	2.539	3.82E-03
transposase family protein	2.539	2.02E-02
IS66 family transposase	2.527	9.66E-03
IS30 family transposase	2.489	2.95E-02
transposase	2.458	5.34E-53
IS256 family transposase	2.210	1.12E-02
IS5 family transposase	2.197	3.75E-07
IS110 family transposase	2.104	6.42E-05
helix-turn-helix transcriptional regulator	0.698	7.97E-03
alpha/beta hydrolase	0.677	2.56E-03
glycosyltransferase	0.660	5.75E-04
SDR family NAD(P)-dependent oxidoreductase	0.629	2.77E-02
FAD-dependent oxidoreductase	0.618	2.42E-02
GntR family transcriptional regulator	0.596	1.03E-02
LacI family DNA-binding transcriptional regulator	0.584	8.32E-03
TetR family transcriptional regulator	0.564	9.81E-04
MarR family transcriptional regulator	0.561	4.07E-03
TonB-dependent receptor	0.556	4.57E-04
ABC transporter permease subunit	0.551	1.08E-04
substrate-binding domain-containing protein	0.543	9.93E-03
amidohydrolase family protein	0.537	2.89E-02
tetratricopeptide repeat protein	0.537	6.22E-03
TetR/AcrR family transcriptional regulator	0.532	1.01E-12
SDR family oxidoreductase	0.526	6.01E-16
MFS transporter	0.526	7.65E-20
extracellular solute-binding protein	0.498	4.65E-08
cupin domain-containing protein	0.493	4.55E-04
response regulator	0.489	2.47E-13

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

GenBank CDS product annotation	enrichment in predicted programmed inversion target neighbors (odds ratio)	corrected p-value
LysR family transcriptional regulator	0.487	1.68E-16
cytochrome P450	0.477	6.96E-05
amidohydrolase	0.474	2.73E-02
ATP-binding cassette domain-containing protein	0.471	1.83E-10
universal stress protein	0.466	4.48E-04
sensor histidine kinase	0.426	6.02E-05
carboxymuconolactone decarboxylase family protein	0.420	2.25E-02
acyl-CoA dehydrogenase family protein	0.414	6.64E-10
IclR family transcriptional regulator	0.412	4.22E-07
sugar ABC transporter substrate-binding protein	0.410	2.93E-02
response regulator transcription factor	0.404	3.04E-22
ABC transporter ATP-binding protein	0.402	1.60E-33
efflux RND transporter periplasmic adaptor subunit	0.379	4.58E-04
FAD-binding oxidoreductase	0.370	7.37E-04
ABC transporter substrate-binding protein	0.368	4.14E-32
NAD(P)-dependent oxidoreductase	0.366	4.65E-04
enoyl-CoA hydratase/isomerase family protein	0.360	1.55E-10
EAL domain-containing protein	0.359	1.27E-04
enoyl-CoA hydratase	0.357	2.01E-04
PAS domain S-box protein	0.354	6.09E-04
CoA transferase	0.339	5.10E-09
FadR family transcriptional regulator	0.332	8.34E-04
Lrp/AsnC family transcriptional regulator	0.327	2.75E-03
aspartate aminotransferase family protein	0.323	2.05E-03
LCP family protein	0.310	1.05E-02
ABC transporter permease	0.309	7.27E-60
acyl-CoA thioesterase	0.308	2.44E-02
sugar phosphate isomerase/epimerase	0.305	2.71E-03
sugar ABC transporter permease	0.286	5.10E-11
acyl-CoA/acyl-ACP dehydrogenase	0.285	1.12E-05
glycosyltransferase family 4 protein	0.277	3.10E-09

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

GenBank CDS product annotation	enrichment in predicted programmed inversion target neighbors (odds ratio)	corrected p-value
S9 family peptidase	0.268	1.07E-02
tripartite tricarboxylate transporter substrate binding protein	0.259	1.23E-09
hydroxymethylglutaryl-CoA lyase	0.254	1.28E-02
aldehyde dehydrogenase	0.252	5.38E-04
RDD family protein	0.251	1.11E-04
pentapeptide repeat-containing protein	0.248	1.65E-05
glucose 1-dehydrogenase	0.242	2.10E-02
aquaporin family protein	0.241	1.60E-04
phosphotransferase	0.230	1.64E-15
amino acid ABC transporter permease	0.205	5.62E-05
efflux transporter outer membrane subunit	0.202	2.99E-03
FecR domain-containing protein	0.190	9.12E-03
RdlA protein	0.179	2.44E-03
TRAP transporter small permease	0.155	3.70E-03
SIS domain-containing protein	0.152	2.55E-03
DNA-3-methyladenine glycosylase 2 family protein	0.145	3.51E-02
glycerol-3-phosphate dehydrogenase/oxidase	0.144	2.88E-07
glutamine amidotransferase	0.110	3.23E-02
acetyl-CoA C-acyltransferase	0.098	1.76E-08
acyl-CoA thioesterase II	0.092	1.51E-06
TRAP transporter small permease subunit	0.090	1.43E-03
chaplin	0.090	5.56E-13
TRAP transporter substrate-binding protein	0.086	6.89E-04
PilZ domain-containing protein	0.068	2.49E-06
BON domain-containing protein	0.061	1.78E-02
pyruvate dehydrogenase (acetyl-transferring) E1 component subunit alpha	0.054	3.84E-03
arginine deiminase	0.051	4.50E-16
bacterial proteasome activator family protein	0.041	1.06E-13
pyridoxamine 5'-phosphate oxidase	0.034	2.00E-07

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

GenBank CDS product annotation	enrichment in predicted programmed inversion target neighbors (odds ratio)	corrected p-value
bifunctional [glutamine synthetase] adenylyltransferase/[glutamine synthetase]-adenylyl-L-tyrosine phosphorylase	0.027	4.32E-10
HpcH/HpaI aldolase/citrate lyase family protein	0.000	7.42E-03
enoyl-ACP reductase FabI	0.000	2.44E-02
succinate dehydrogenase assembly factor 2	0.000	6.89E-29
quinoprotein relay system zinc metallohydrolase 1	0.000	1.70E-02
DUF502 domain-containing protein	0.000	1.02E-06
UPF0149 family protein	0.000	1.60E-02
circularly permuted type 2 ATP-grasp protein	0.000	1.58E-02
cytochrome c-550 PedF	0.000	2.22E-06
Na+/H+ antiporter subunit C	0.000	5.71E-06
NtaA/DmoA family FMN-dependent monooxygenase	0.000	4.24E-05
Fe-S cluster assembly protein HesB	0.000	5.74E-04
flagellar protein FlaG	0.000	3.68E-02

Long-read sequencing data confirm predicted programmed inversion representatives

To obtain more compelling evidence for programmed inversions, we turned to examine long-read sequencing data. We started by manually inspecting predicted programmed inversions potentially targeting identified enriched gene families, looking for recurring genomic contexts. Next, for each 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 sequencing data. Then, for each found locus, we retrieved and searched the sequencing data for reads matching different variants of the locus that might arise from programmed inversions (**Figure 4A**). We thus identified 22 programmed inversion loci, each exhibiting a different genomic context (**Table 3, Figure 4B, Figure 4C, Supplementary Figures S9-S29**). 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**).

Some of the identified programmed inversion loci contain systems known to be targeted by programmed inversions (6, 14, 23, 24, 29–38), but most contain systems that, to the best of our knowledge, were not previously described to be targeted by programmed inversions (40). Most notably, 11 of these programmed inversions target systems encoding Type II restriction-modification enzymes, which can be divided to three classes: (a) Class 1 DISARM (41) and similar systems; (b) systems containing *C. jejuni* Cj0031 (42) homologs; and (c) BREX type 1 (39). 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). Finally, in one locus, programmed inversions target a gene which, according to the NCBI Conserved Domain Database, seems to encode a shufflon PilV (pfam04917) and phage tail collar (pfam07484) fusion-protein.

Protein sequence alignment revealed the identified shufflon PilV and phage tail collar fusion-protein to be only partially homologous to previously described shufflons (14, 37). Alignment (using online blastp) of this fusion protein to previously described shufflons (NCBI Protein accession WP_001389385.1, encoded in *Salmonella enterica* subsp. enterica serovar Typhimurium plasmid R64, Nucleotide accession AP005147.1, 104790-106214, and in *Shigella sonnei* plasmid P9, Nucleotide accession NC_002122.1, 77407-78831; Protein accession WP_010895887.1, encoded in *Escherichia coli* plasmid R721, Nucleotide accession NC_002525.1, 44733-45986; Protein accession AAO71709.1, encoded in *Salmonella enterica* subsp. enterica serovar Typhi Ty2, Nucleotide accession AE014613.1, 4418334-4419641) revealed sequence similarity in the shufflon PilV domain, but not in the phage tail collar domain.

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) seemed biased in favor of one type of variants. Of the 176 reads matching the locus (and satisfying our requirements, see **Materials and Methods**, 'Variant identification in long-read sequencing 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), and 5 matched variants that can be transformed back to the reference variant by a single inversion (1-inversion 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. 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 ($P = 2.9 \cdot 10^{-44}$, 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, $P = 9.6 \cdot 10^{-19}$, two-sided binomial test), suggesting 1-inversion variants are less stable than other variants. Given the observation that a BREX type 1 system in *Bacillus cereus* H3081.97 contains two transcription units: *brxA-brxB-brxC-pglX* and *pglZ-brxL* (39) 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 transcription unit 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 (43).

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

Table 3. Gene-altering programmed inversion loci (continued)

NCBI Nucleotide accession	longest target CDS location	longest target CDS product description	locus description
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
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

Table 3. Gene-altering programmed inversion loci (continued)

NCBI Nucleotide accession	longest target CDS location	longest target CDS product description	locus description
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

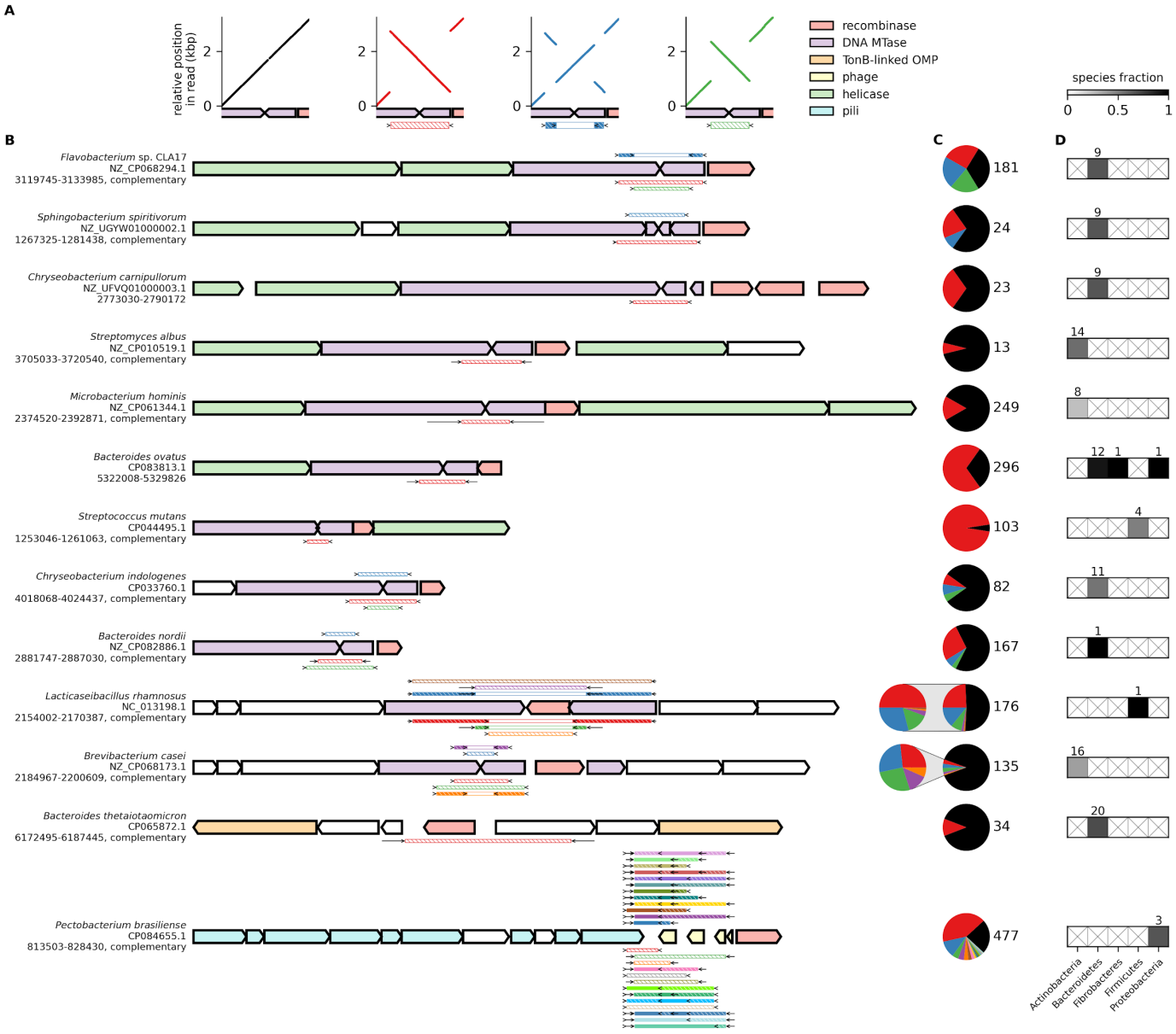


Figure 4. Long read sequencing data show variant coexistence in multiple gene-altering programmed inversion loci.

(**A**) Alignments of reads representing four variants of a programmed inversion locus in *Flavobacterium* sp. CLA17. Each plot shows the read alignment to the reference variant sequence at the varying locus (NZ_CP068294.1, 3120509-3123718). 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 presumably promoting programmed inversions that allow switching between the variants. 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 non-reference variant. Species, NCBI Nucleotide accession and location are specified for each locus (left). For each non-reference variant identified, 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 the illustration of their difference from the reference variant. Total number of reads matching the locus is indicated (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 greyscale. 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.

Discussion

To date, a broad and systematic search for gene-altering programmed inversions has been absent. Two studies systematically and widely searched for programmed inversions targeting genes encoding Type I specificity subunit HsdS (23, 24), and one study systematically searched 203 bacterial genomes for any programmed inversion, regardless of whether it targets genes or regulatory elements (22). Thus, 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 inverted repeats (IRs) overlapping coding sequences (CDSs), and identifying intra-species variation in these loci by comparing them to other loci in same-species genomes.

Analyzing this dataset, we identified four genomic architectures enriched for putative gene-altering programmed inversions, i.e., programmed inversion candidates for which intra-species variation was identified. One genomic architecture is long IRs, while another is a short distance between transcription units targeted by programmed inversions. A more intriguing genomic architecture is the orientation of transcription units targeted by programmed inversions, i.e., head-to-head or tail-to-tail, such that IRs are closer to each other, relative to the opposite orientation. Assuming that typically, a shorter distance between IRs allows for higher frequency of programmed inversions, we hypothesized that throughout evolution of transcription units targeted by programmed inversions, transcription units oriented such that IRs are closer to each other would be favored. The last genomic architecture is the asymmetry of transcription units targeted by programmed inversions, in terms of the length of parts of transcription units upstream to all IRs. In other words, one of the targeted transcription units is missing its part which is upstream to all IRs. Similar to what was previously observed (14, 24, 29, 30), it seems that these target transcription units have a single copy of their upstream part, while programmed inversions switch between different versions of the downstream part of the transcription unit. Maybe this pattern arose from an inverted duplication followed by elimination of transcription of one of the copies, e.g., by a promoter mutation. This would render the non-transcribed region upstream to the most upstream repeat non-functional in all variants, ultimately leading to its deletion.

Using these genomic architectures, 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 inversion, 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 (6, 14, 23, 24, 29–38). Conversely, several statistically significantly enriched gene families code for proteins of unknown function that are not known to be targeted by programmed inversions, though for most of them it is unclear whether they are really targeted by programmed inversions. 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 (21, 39), yet the family is largely considered to not be a target of programmed inversions (11, 44, 45).

Finally, for some recurring genomic contexts of predicted programmed inversions targeting enriched gene families, we found sound evidence for variant coexistence in long-read sequencing experiments,

strongly suggestive of programmed inversions. We thus identified multiple programmed inversion loci, exhibiting different genomic contexts.

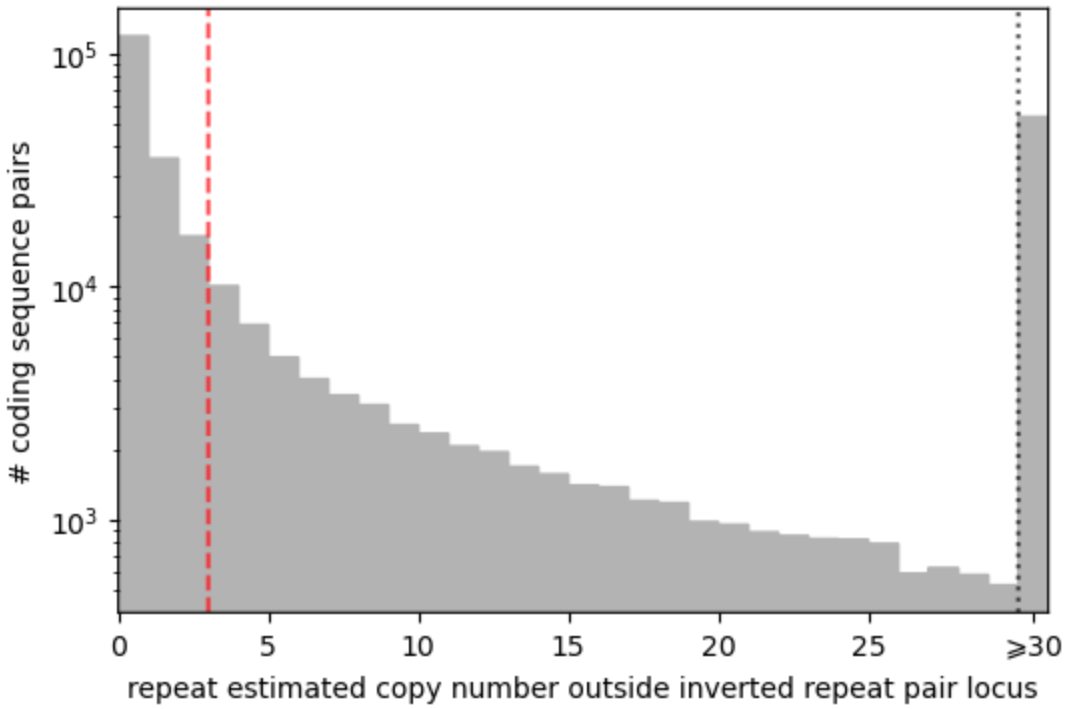
Examining this diverse set of programmed inversion genomic contexts revealed some recurring patterns. One of these patterns was fusion-genes, which appeared in multiple identified programmed inversion loci, encoding for various domains, including helicase YprA, DUF1998, SNF2 helicase, Type II RM, SNF 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 (38). 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 containing the phage defense system Class 1 DISARM (41) or similar systems 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 (46). One of these loci contained a programmed inversion targeting a homolog of *C. jejuni cj0031*, which also phase-varies, but through a hypermutable polyG tract rather than inversions (42), while the other locus contained a Type I restriction-modification 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 transcription unit for *brxA-brxB-brxC-pglX*, as was shown to be the case in *Bacillus cereus* H3081.97 (39), 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 S32**, 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 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. 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* transcription unit. Compared to the *L. rhamnosus* locus, the transcription unit 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 S32**). 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.

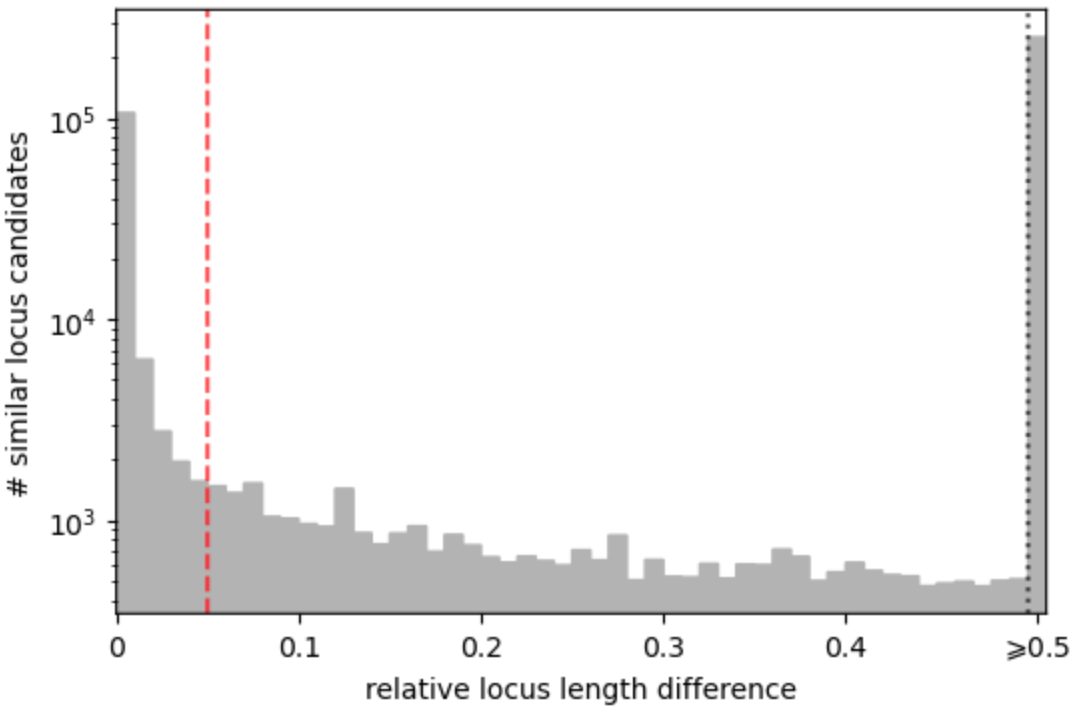
Our approach has several limitations. First, we started our analysis by choosing a single representative genome for each bacterial species. This limited the bias toward more sequenced species, but we probably missed many programmed inversions due to this choice. In addition, we looked for long-read evidence 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**).

In summary, using a computational and systematic approach, we predicted many loci across the bacterial domain to contain gene-altering programmed inversions and identified characteristic genomic architectures 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.

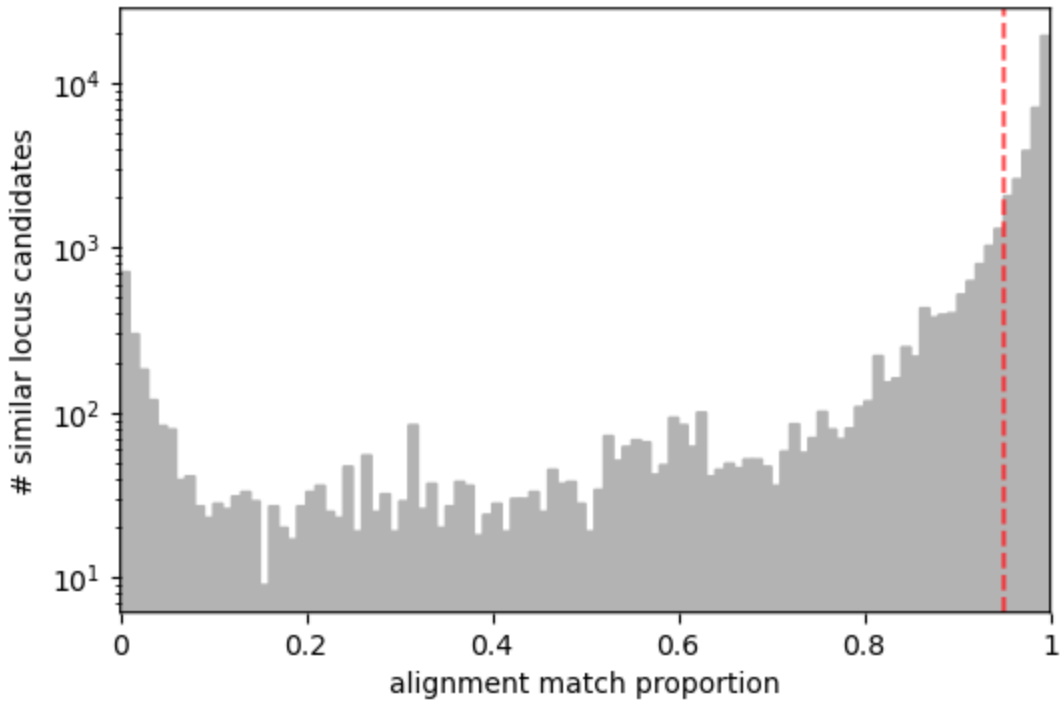
Supplementary Figures



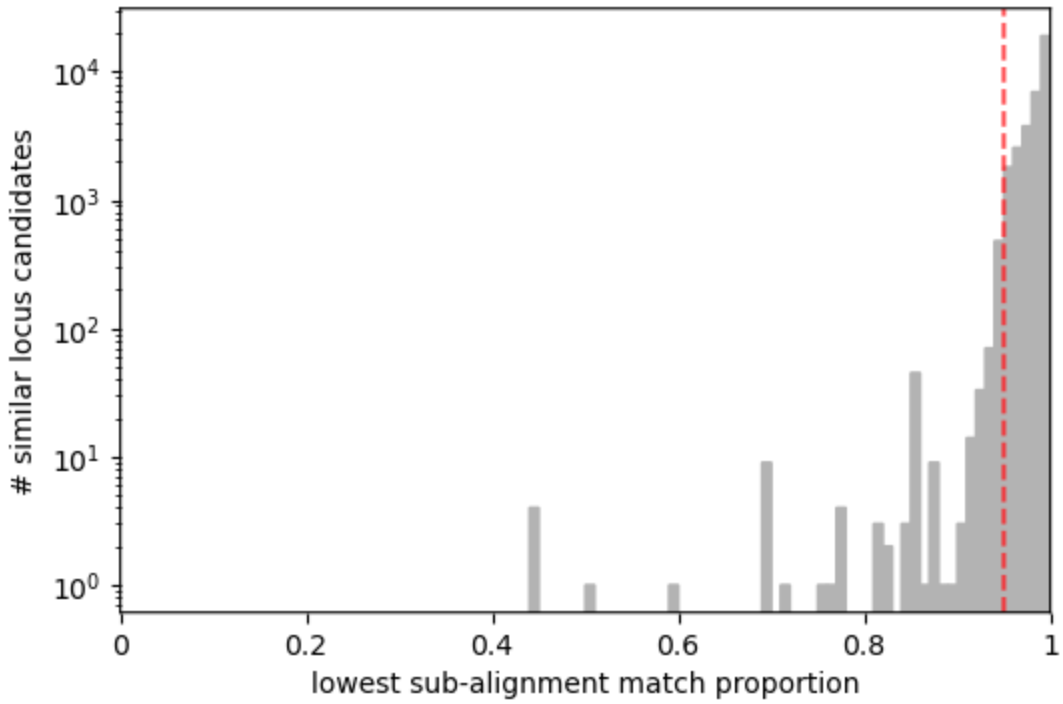
Supplementary Figure S1. Discarding coding sequence pairs with repetitive inverted repeats. Distribution of inverted repeat estimated copy number in their representative genome, excluding the locus in which the inverted repeat pair was found and linked to a coding sequence pair. To avoid loci satisfying the criteria of programmed inversion candidates, but are actually mobile elements which are not targeted by programmed inversions, coding sequence pairs containing inverted repeats with high estimated copy number were excluded (>=3, dashed vertical red line).



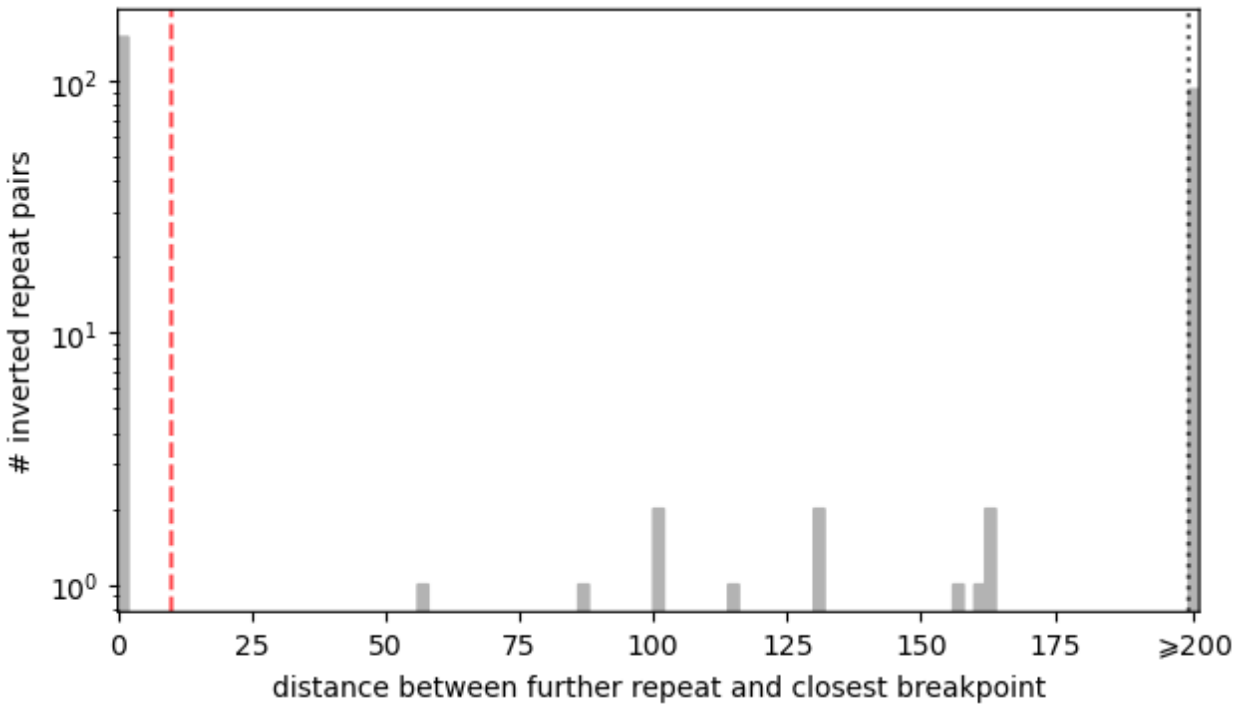
Supplementary Figure S2. Discarding collinear alignment pairs that match PIC loci to loci of substantially different length.
Distribution of absolute difference between lengths of programmed inversion candidate (PIC) loci and their corresponding similar locus candidates, normalized by the length of the PIC locus, namely, relative locus length difference. To avoid similar locus candidates that are flanked by the same regions as the PIC locus, but cannot contain different variants of the PIC locus, as inversions do not change the sequence length, similar locus candidates with a substantially different length were excluded (>0.05, dashed vertical red line).



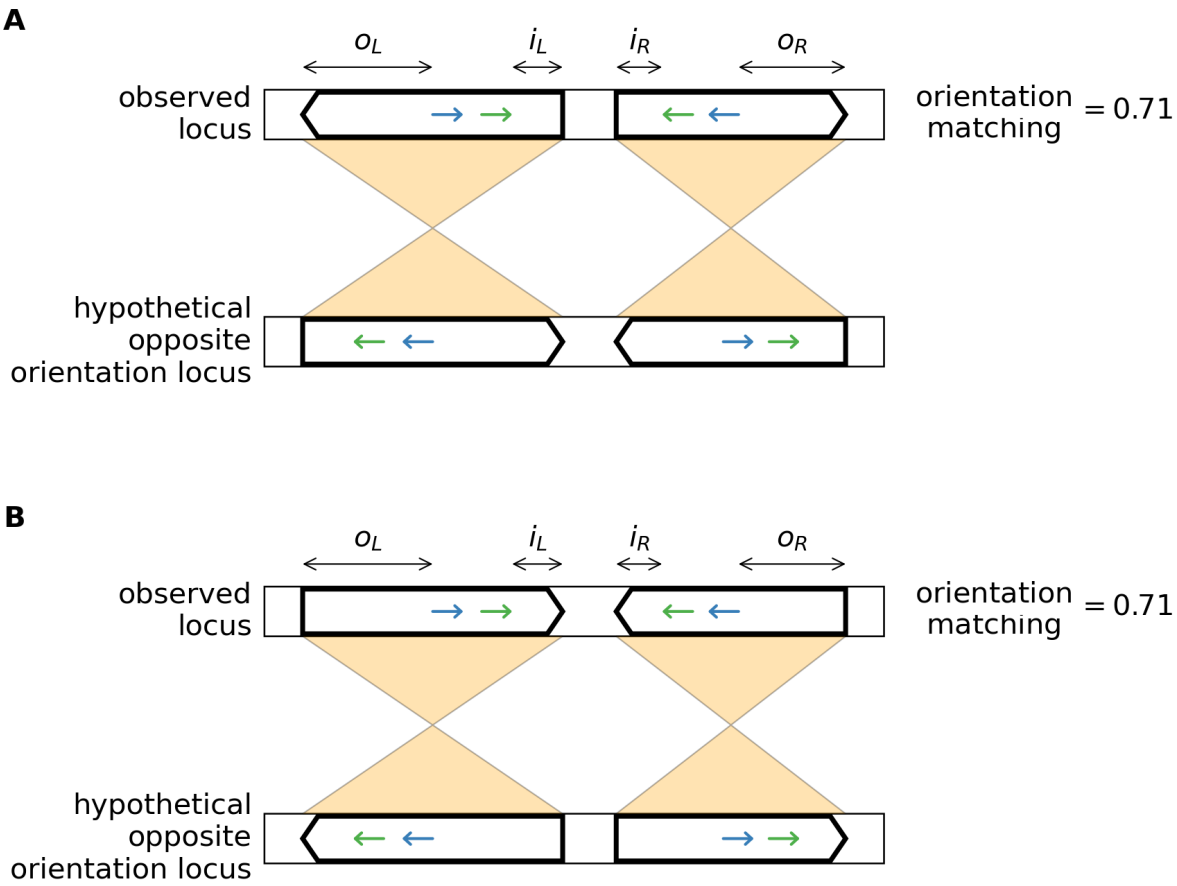
Supplementary Figure S3. Discarding similar length loci with low sequence similarity to their corresponding PIC locus.
Distribution of match proportion in progressiveMauve alignments between PIC loci and their corresponding similar locus candidates. To avoid similar locus candidates with substantial difference in sequence content, such that it is unlikely that they contain different variants of the PIC locus, similar locus candidates with a low match proportion in the alignment to PIC locus were excluded (<0.95, dashed vertical red line).



Supplementary Figure S4. Discarding similar length loci with low sequence similarity to their corresponding PIC locus in any synteny block.
Distribution of lowest sub-alignment match proportion in progressiveMauve alignments between PIC loci and their corresponding similar locus candidates. To avoid similar locus candidates with substantial difference in sequence content, such that it is unlikely that they contain different variants of the PIC locus, similar locus candidates with a low match proportion in any sub-alignment in the alignment to PIC locus were excluded (<0.95, dashed vertical red line).

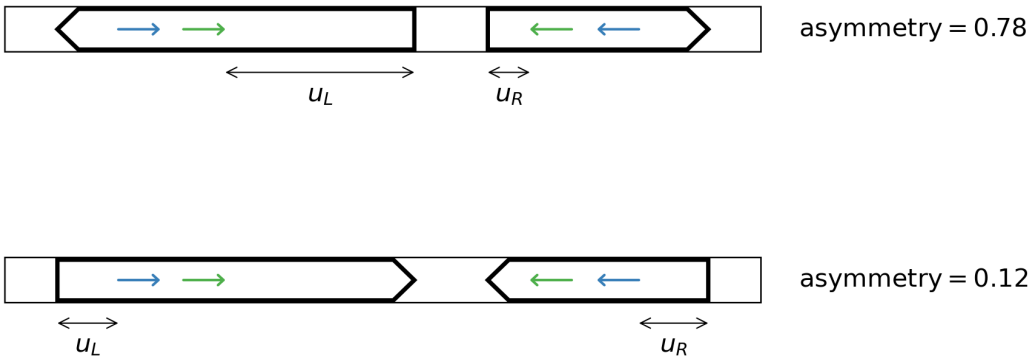


Supplementary Figure S5. Discarding inverted repeat pairs without any similar locus such that both repeats overlap or appear near breakpoints.
Distribution of distances between inverted repeats and closest breakpoints. For each inverted repeat, the shortest distance between the further repeat, i.e., the repeat further from its closest breakpoint, and its closest breakpoint-containing region, was calculated for each similar locus, and the minimal distance was considered. To avoid inverted repeats that appear near breakpoints by chance, for each pair of inverted repeats, similar loci in which any of the repeats was far from any breakpoint were not considered as evidence for intra-species variation (>10, dashed vertical red line).

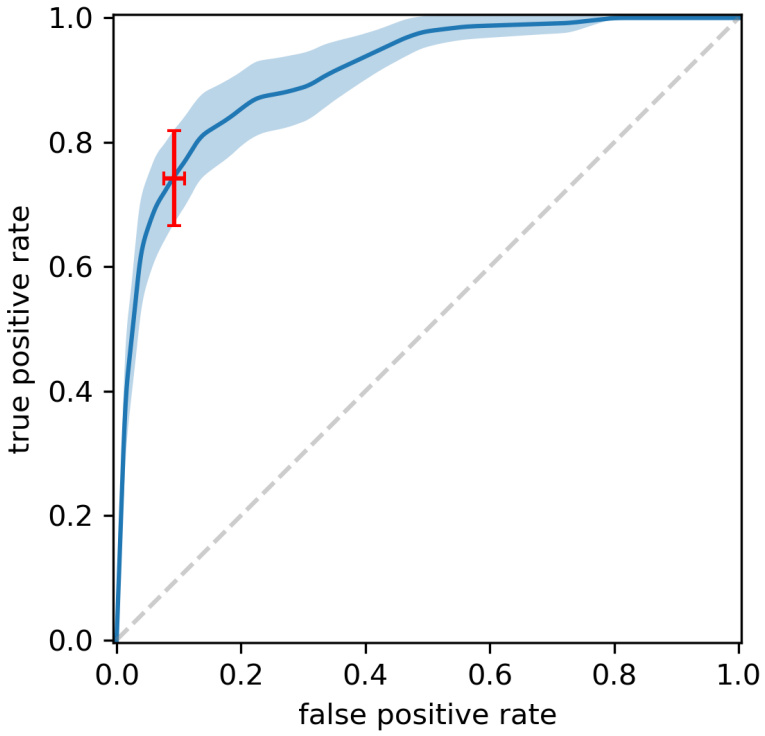


Supplementary Figure S6. Schematic examples of the orientation matching genomic architecture measure.

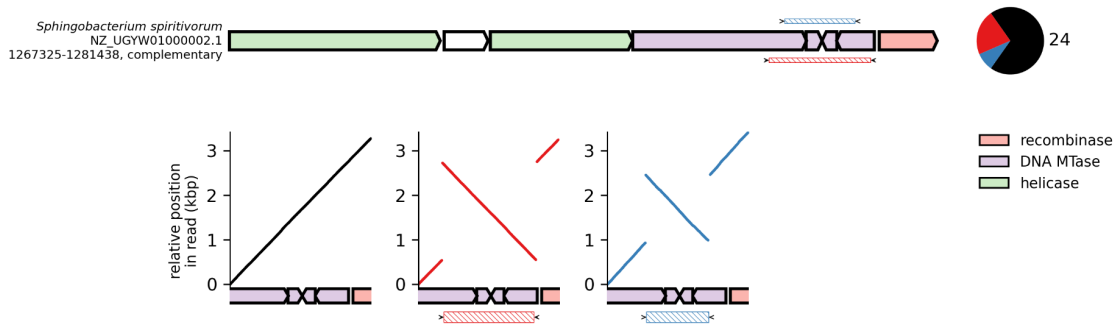
Schematic illustrations to demonstrate the rationale behind the orientation matching genomic architecture measure, defined as $\frac{o_L+o_R}{o_L+o_R+i_L+i_R}$, while o_L and o_R are the lengths of coding sequence (or more generally, transcription units containing these coding sequences, not shown) outer regions relative to repeats, and i_L and i_R are the lengths of coding sequence inner regions relative to repeats. Regardless of whether in the observed locus the coding sequences are oriented tail-to-tail (**A**) or head-to-head (**B**), in both examples the total length of inner regions is smaller than that of outer regions, such that in the observed orientation, the shortest region flanked by inverted repeats is shorter, relative to a hypothetical locus in which the coding sequences exhibit the opposite orientation, namely, head-to-head (**A**) or tail-to-tail (**B**). Inverted repeats and coding sequences are indicated as colored stick arrows and wide black arrows, respectively. Alignments are indicated by light orange projections.



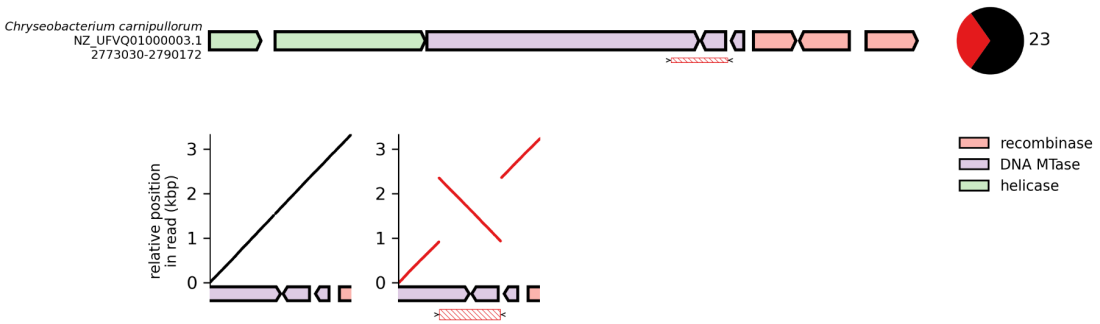
Supplementary Figure S7. Schematic examples of the asymmetry genomic architecture measure. Schematic illustrations to demonstrate the rationale behind the asymmetry genomic architecture measure, defined as $1 - \frac{min(u_L, u_R)}{max(u_L, u_R)}$, while u_L and u_R are the lengths of coding sequence (or more generally, transcription units containing these coding sequences, not shown) regions upstream to all repeats. (**A**) The region upstream to all repeats in the left coding sequence is much longer than the region upstream to all repeats in the right coding sequence. Accordingly, the asymmetry measure value is high. (**B**) The regions upstream to all repeats in the left and right coding sequences are of similar length. Accordingly, the asymmetry measure value is low. Inverted repeats and coding sequences are indicated as colored stick arrows and wide black arrows, respectively.



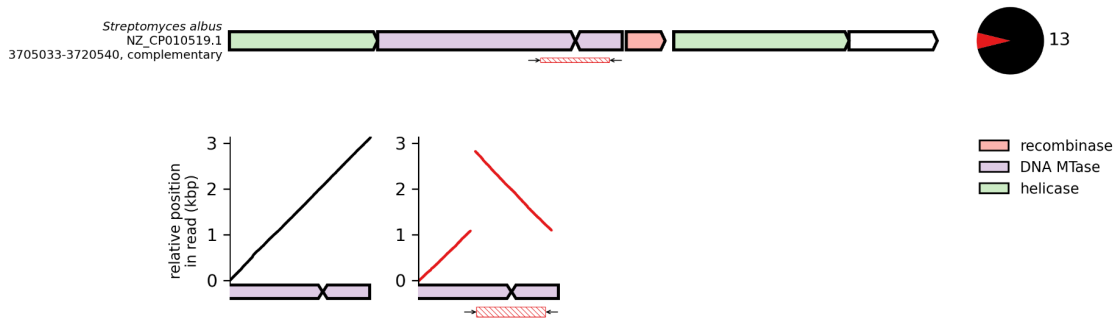
Supplementary Figure S8. Gene-altering programmed inversion prediction model assessment. Model performance was assessed using 500 cross validation simulations (see **Materials and Methods**, 'Gene-altering programmed inversion prediction'), producing receiver operating characteristic (ROC) curves. Mean±1SD ROC curve is shown (blue and shaded areas). Area under the ROC curve (AUC) values had a mean of 0.91 and a standard deviation of 0.025. To binarize predictions, 0.05 was used as a cutoff, namely, programmed inversion candidates with predicted probability >0.05 were marked as predicted programmed inversions. Mean±1SD of false positive rate and true positive rate for the cutoff 0.05, obtained from the simulations, are indicated (red). y=x is indicated (dashed gray line).



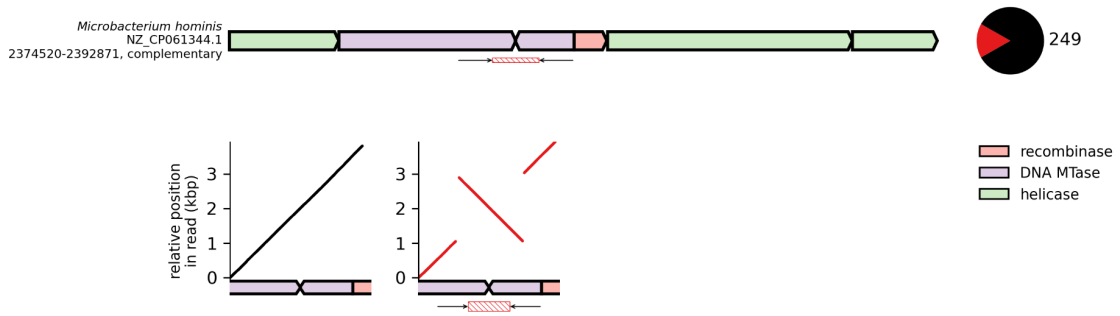
Supplementary Figure S9. Long read sequencing data show variant coexistence in a gene-altering programmed inversion locus similar to Class 1 DISARM.
Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Sphingobacterium spiritivorum* (NZ_UGYW01000002.1, 1267325-1281438), as in **Figure 4**. Abbreviations: MTase, methyltransferase.



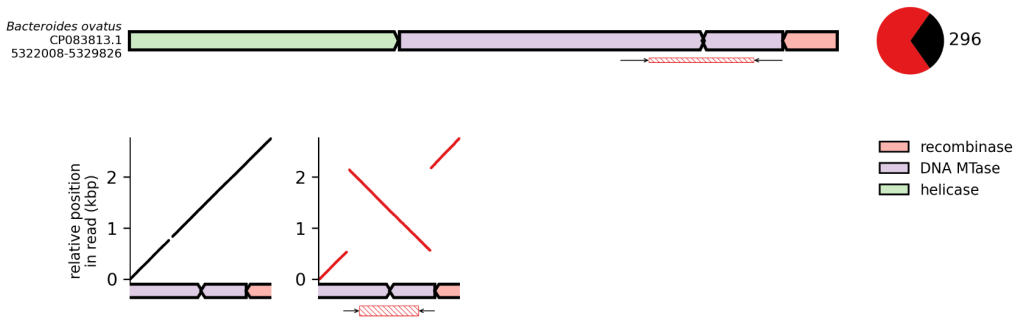
Supplementary Figure S10. Long read sequencing data show variant coexistence in a gene-altering programmed inversion locus similar to Class 1 DISARM.
Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Chryseobacterium carnipullorum* (NZ_UFVQ01000003.1, 2773030-2790172), as in **Figure 4**. Abbreviations: MTase, methyltransferase.



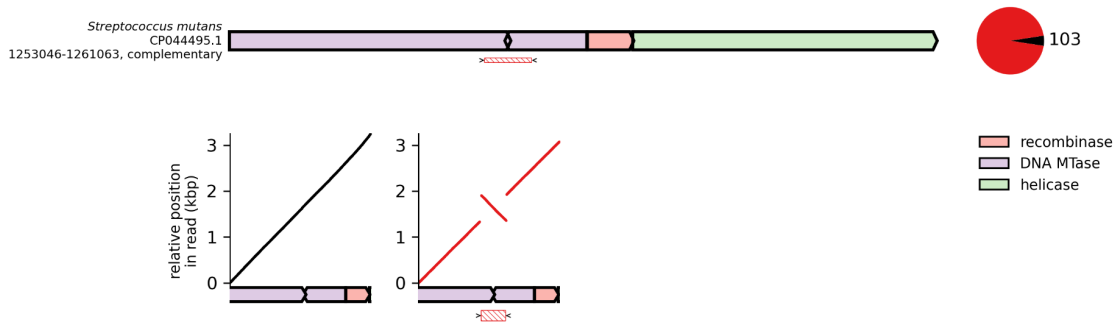
Supplementary Figure S11. Long read sequencing data show variant coexistence in a gene-altering programmed inversion locus similar to Class 1 DISARM.
Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Streptomyces albus* (NZ_CP010519.1, 3705033-3720540), as in **Figure 4**. Abbreviations: MTase, methyltransferase.



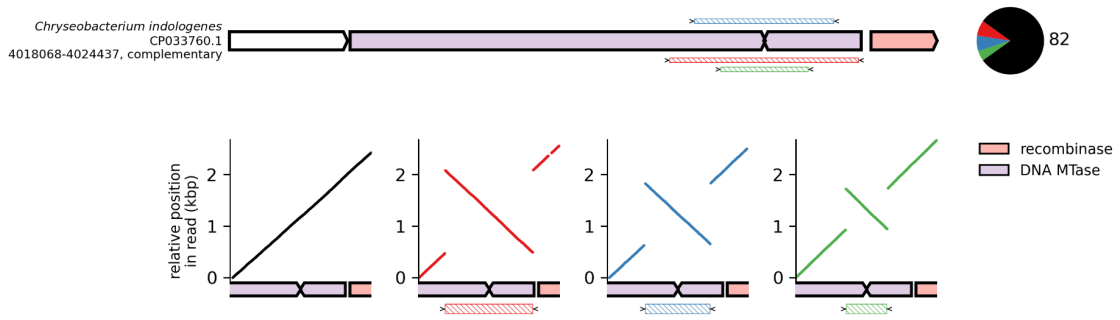
Supplementary Figure S12. Long read sequencing data show variant coexistence in a gene-altering programmed inversion locus similar to Class 1 DISARM.
Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Microbacterium hominis* (NZ_CP061344.1, 2374520-2392871), as in **Figure 4**. Abbreviations: MTase, methyltransferase.



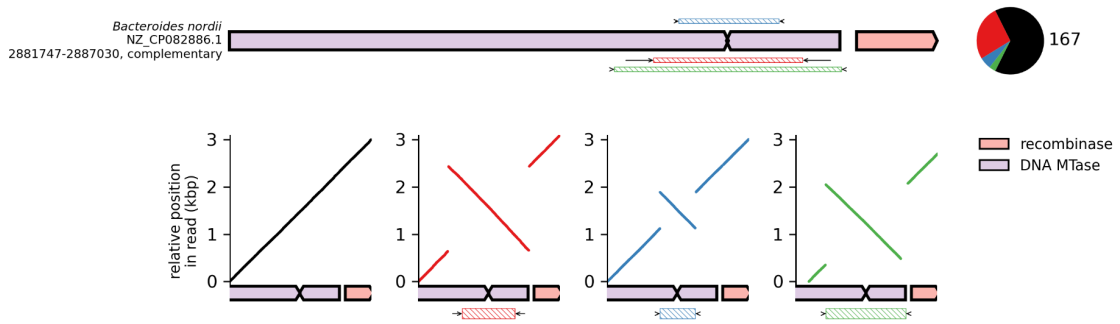
Supplementary Figure S13. Long read sequencing data show variant coexistence for a gene-altering programmed inversion targeting a C. jejuni Cj0031 homolog.
Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Bacteroides ovatus* (CP083813.1, 5322008-5329826), as in **Figure 4**. Abbreviations: MTase, methyltransferase.



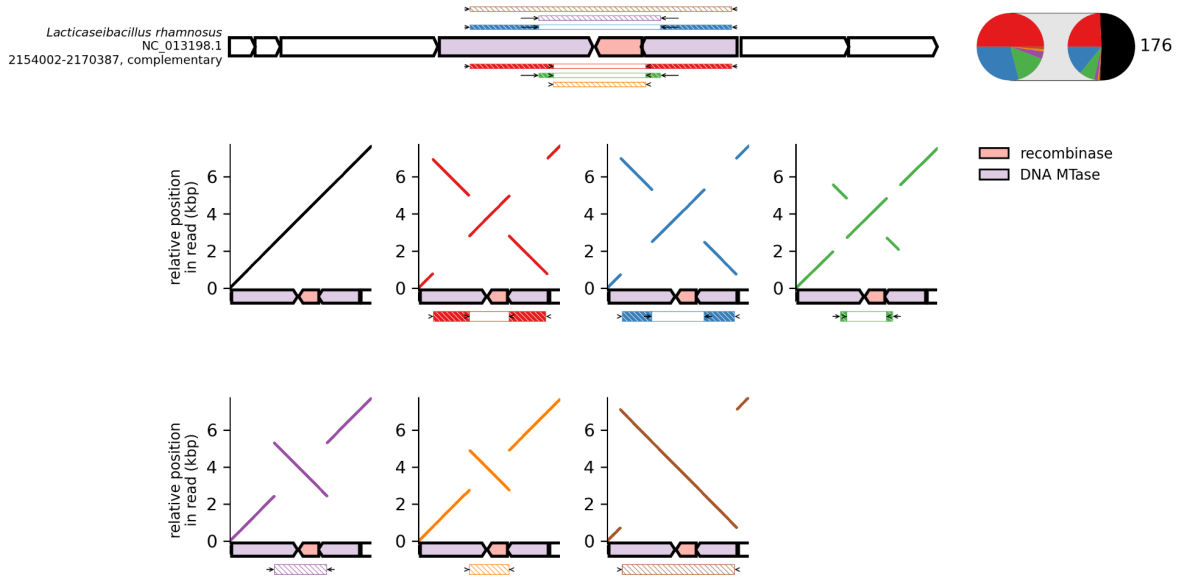
Supplementary Figure S14. Long read sequencing data show variant coexistence for a gene-altering programmed inversion targeting a C. jejuni Cj0031 homolog.
Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Streptococcus mutans* (CP044495.1, 1253046-1261063), as in **Figure 4**. Abbreviations: MTase, methyltransferase.



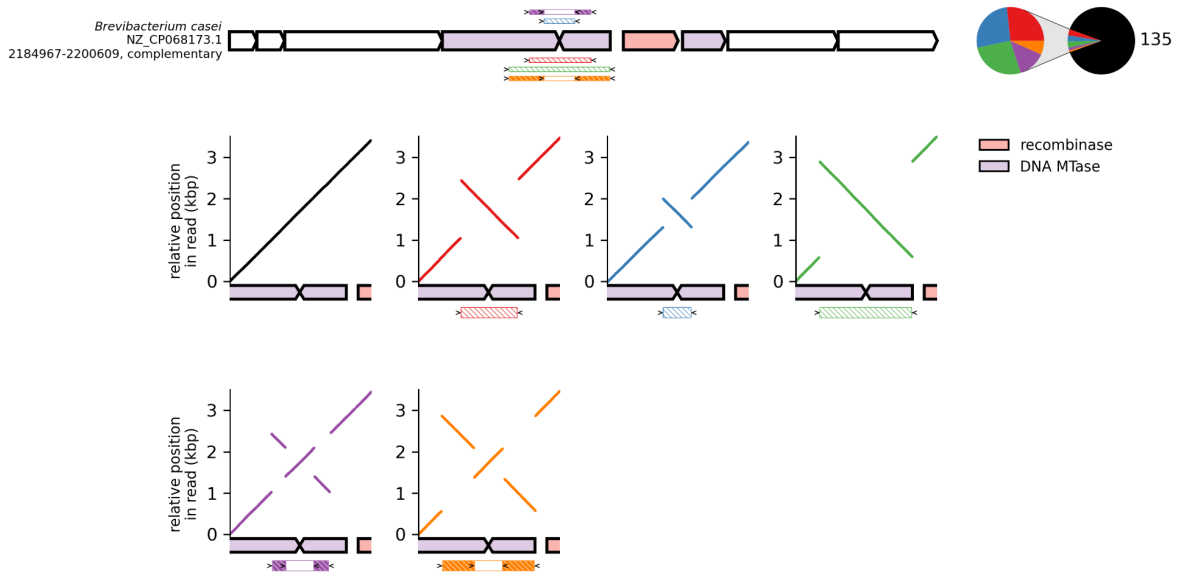
Supplementary Figure S15. Long read sequencing data show variant coexistence for a gene-altering programmed inversion targeting a C. jejuni Cj0031 homolog.
Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Chryseobacterium indologenes* (CP033760.1, 4018068-4024437), as in **Figure 4**. Abbreviations: MTase, methyltransferase.



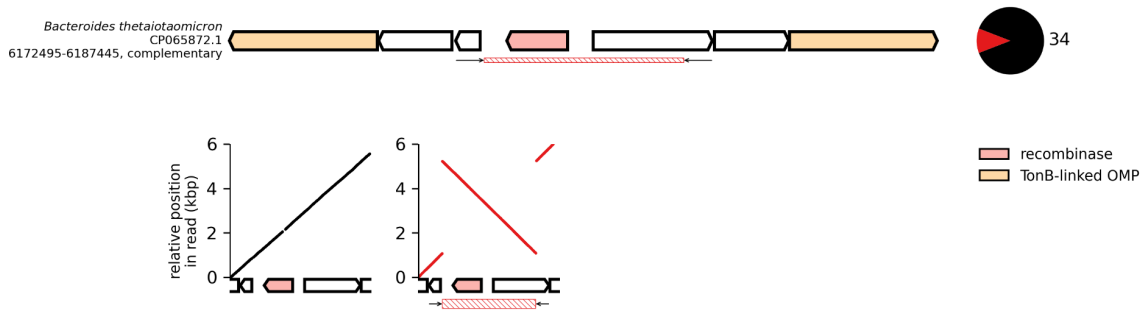
Supplementary Figure S16. Long read sequencing data show variant coexistence for a gene-altering programmed inversion targeting a C. jejuni Cj0031 homolog.
Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Bacteroides nordii* (NZ_CP082886.1, 2881747-2887030), as in **Figure 4**. Abbreviations: MTase, methyltransferase.



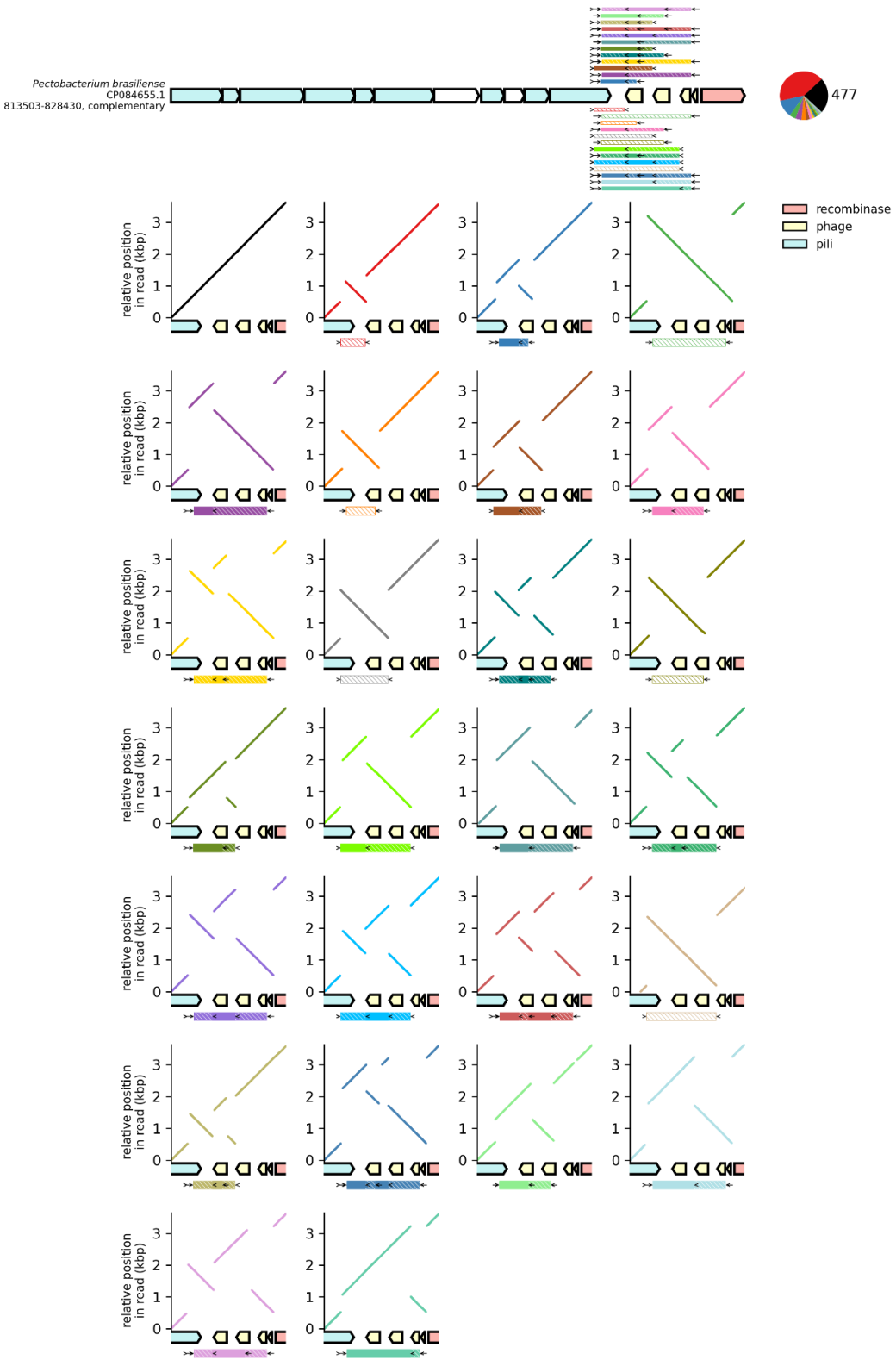
Supplementary Figure S17. Long read sequencing data show variant coexistence in a gene-altering programmed inversion locus containing a BREX type 1 system.
Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Lacticaseibacillus rhamnosus* (NC_013198.1, 2154002-2170387), as in **Figure 4**. Abbreviations: MTase, methyltransferase.



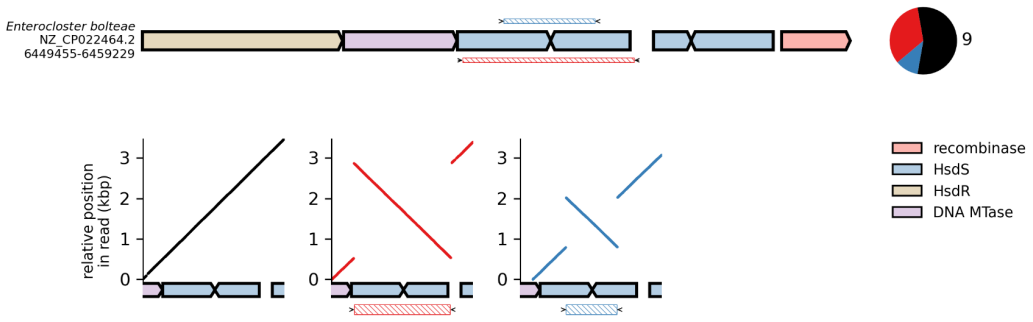
Supplementary Figure S18. Long read sequencing data show variant coexistence in a gene-altering programmed inversion locus containing a BREX type 1 system.
Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Brevibacterium casei* (NZ_CP068173.1, 2184967-2200609), as in **Figure 4**. Abbreviations: MTase, methyltransferase.



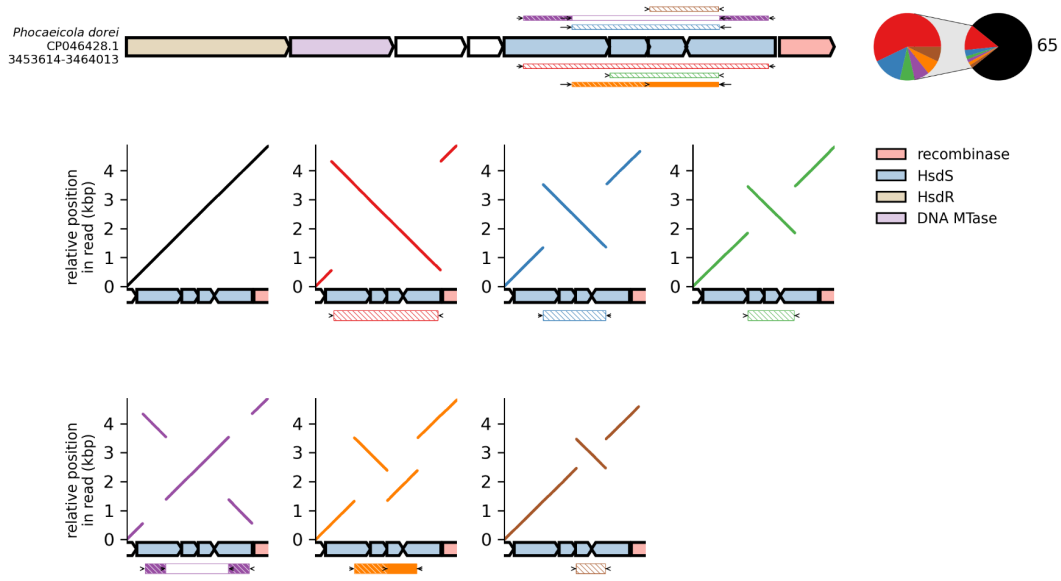
Supplementary Figure S19. Long read sequencing data show variant coexistence for a gene-altering programmed inversion targeting a protein of unknown function.
Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Bacteroides thetaiotaomicron* (CP065872.1, 6172495-6187445), as in **Figure 4**. Abbreviations: OMP, outer membrane protein.



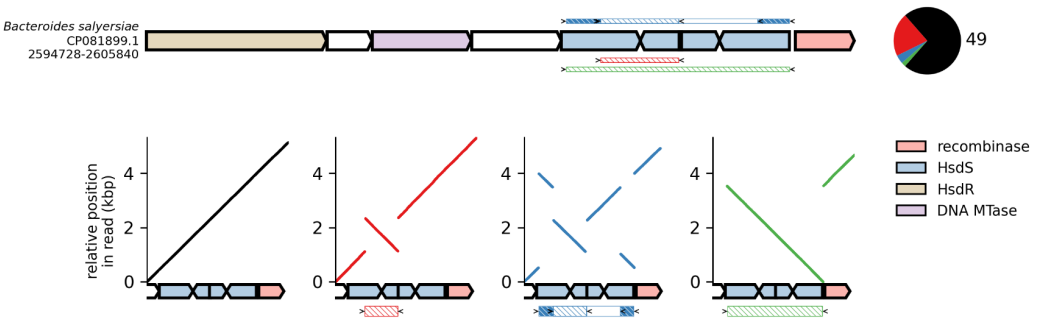
Supplementary Figure S20. Long read sequencing data show variant coexistence for a gene-altering programmed inversion targeting a shufflon PilV and phage tail collar fusion-protein. Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Pectobacterium brasiliense* (CP084655.1, 813503-828430), as in **Figure 4**.



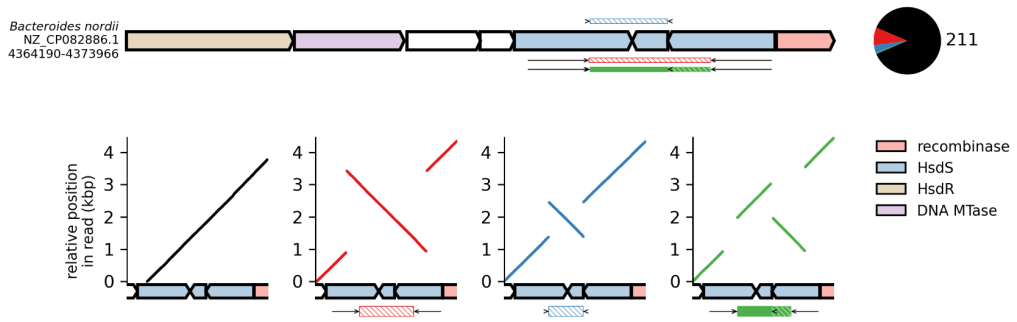
Supplementary Figure S21. Long read sequencing data show variant coexistence in a gene-altering programmed inversion locus containing a Type I restriction-modification system. Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Enterocloster bolteae* (NZ_CP022464.2, 6449455-6459229), as in **Figure 4**. Abbreviations: MTase, methyltransferase.



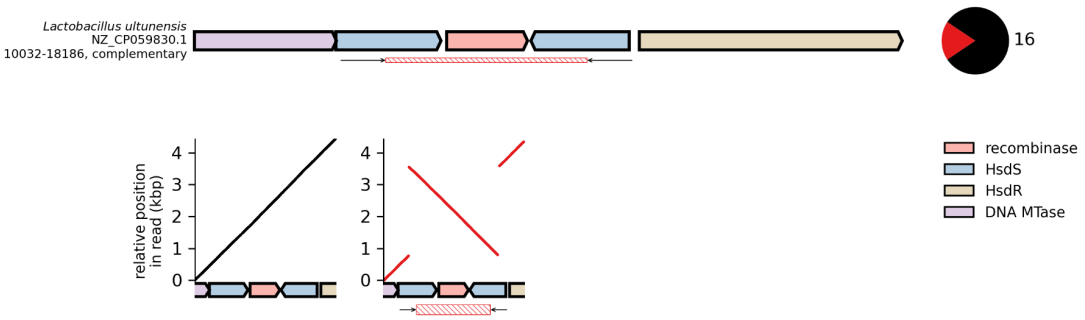
Supplementary Figure S22. Long read sequencing data show variant coexistence in a gene-altering programmed inversion locus containing a Type I restriction-modification system. Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Phocaeicola dorei* (CP046428.1, 3453614-3464013), as in **Figure 4**. Abbreviations: MTase, methyltransferase.



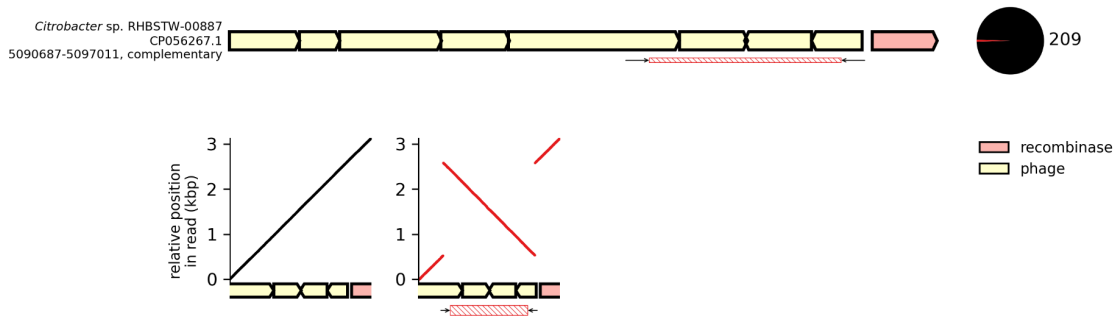
Supplementary Figure S23. Long read sequencing data show variant coexistence in a gene-altering programmed inversion locus containing a Type I restriction-modification system. Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Bacteroides salyersiae* (CP081899.1, 2594728-2605840), as in **Figure 4**. Abbreviations: MTase, methyltransferase.



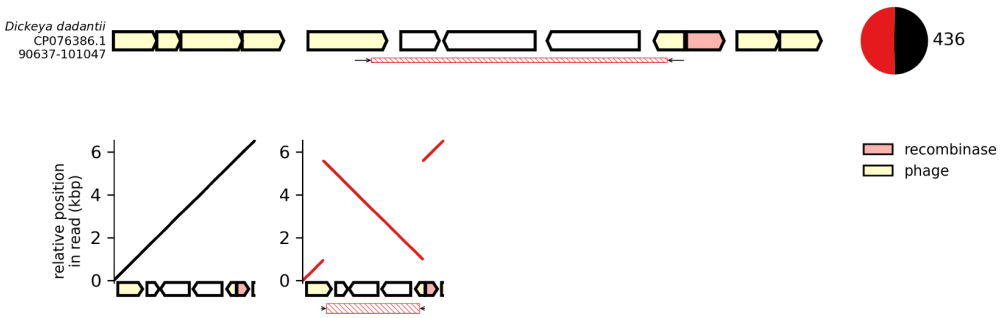
Supplementary Figure S24. Long read sequencing data show variant coexistence in a gene-altering programmed inversion locus containing a Type I restriction-modification system. Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Bacteroides nordii* (NZ_CP082886.1, 4364190-4373966), as in **Figure 4**. Abbreviations: MTase, methyltransferase.



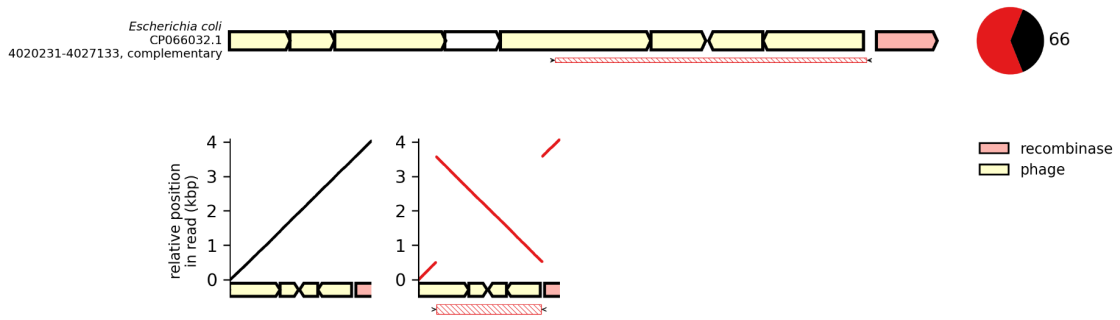
Supplementary Figure S25. Long read sequencing data show variant coexistence in a gene-altering programmed inversion locus containing a Type I restriction-modification system. Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Lactobacillus ultunensis* (NZ_CP059830.1, 10032-18186), as in **Figure 4**. Abbreviations: MTase, methyltransferase.



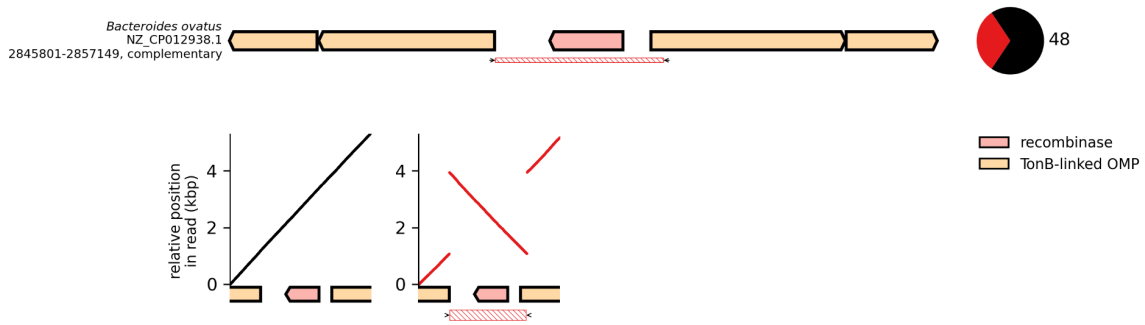
Supplementary Figure S26. Long read sequencing data show variant coexistence for a gene-altering programmed inversion targeting a phage tail protein.
Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Citrobacter* sp. RHBSTW-00887 (CP056267.1, 5090687-5097011), as in **Figure 4**.



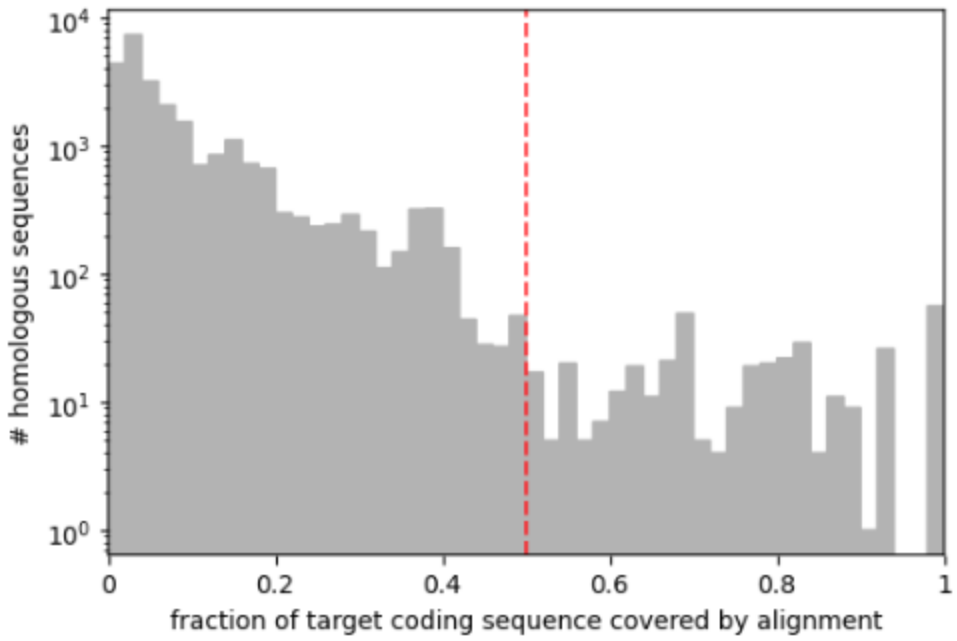
Supplementary Figure S27. Long read sequencing data show variant coexistence for a gene-altering programmed inversion targeting a phage tail protein.
Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Dickeya dadantii* (CP076386.1, 90637-101047), as in **Figure 4**.



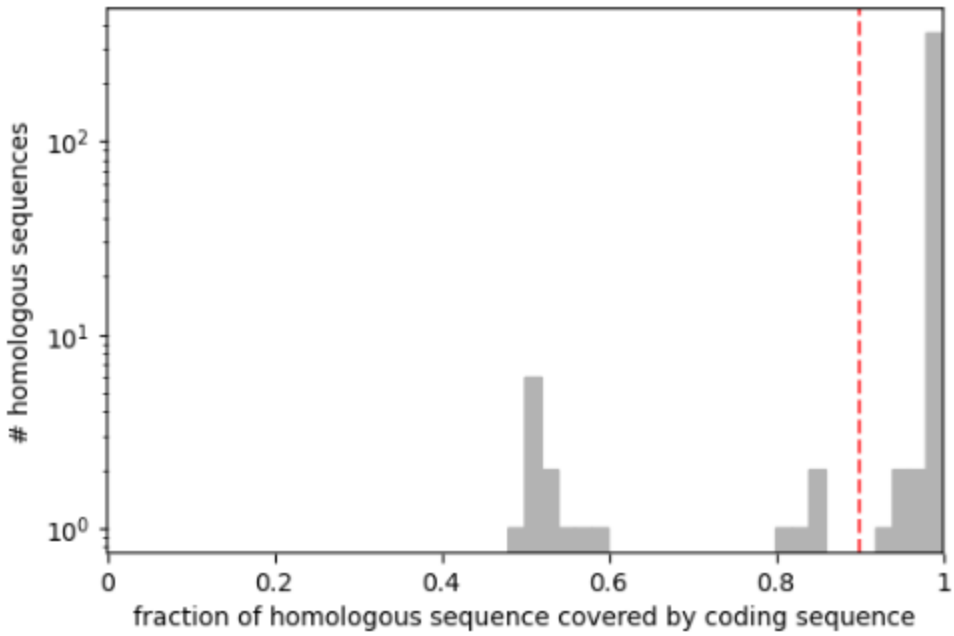
Supplementary Figure S28. Long read sequencing data show variant coexistence for a gene-altering programmed inversion targeting a phage tail protein.
Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Escherichia coli* (CP066032.1, 4020231-4027133), as in **Figure 4**.



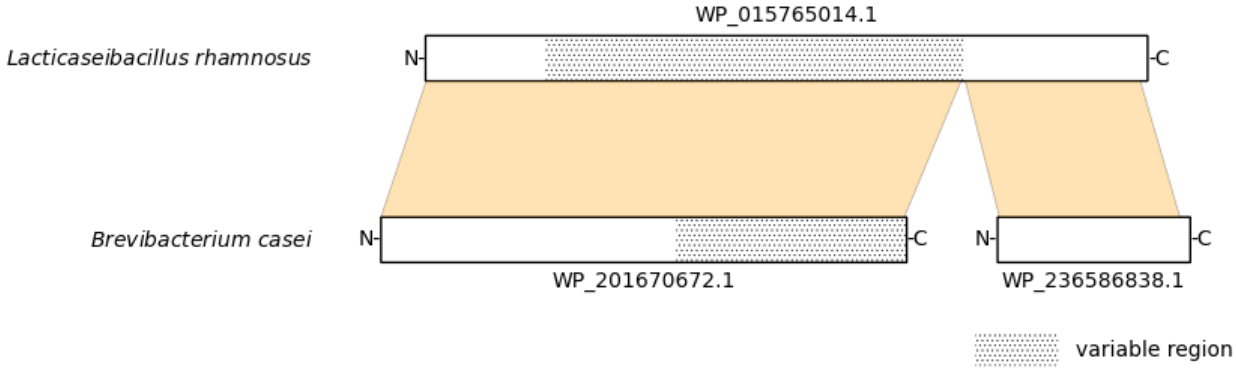
Supplementary Figure S29. Long read sequencing data show variant coexistence for a gene-altering programmed inversion targeting a TonB-linked outer membrane protein. Genomic context, distribution of reads across variants, differences between the reference variant and each non-reference variant, and alignments of reads representing identified variants, are shown for a programmed inversion locus in *Bacteroides ovatus* (NZ_CP012938.1, 2845801-2857149), as in **Figure 4**. Abbreviations: OMP, outer membrane protein.



Supplementary Figure S30. Discarding sequences that are only partially homologous to programmed inversion target coding sequence.
Distribution of fraction of target coding sequence covered by alignment to a homologous sequence, across all homologous sequences. For each homologous sequence identified, that is, a sequence homologous to the longest target coding sequence of a programmed inversion, the fraction of the target coding sequence covered by the alignment between the target coding sequence and the homologous sequence was considered. To avoid sequences that are only partially homologous to the target coding sequence, sequences such that only a short part of the target coding sequence was covered by the alignment were excluded (<0.5, dashed vertical red line).



Supplementary Figure S31. Discarding homologous sequences that are only partially covered by a coding sequence.
Distribution of fraction of a homologous sequence covered by a coding sequence, across all homologous sequences. For each homologous sequence identified, that is, a sequence homologous to the longest target coding sequence of a programmed inversion, the fraction of the homologous sequence covered by a coding sequence was considered. To avoid homologous sequences whose annotated coding sequence is shorter than expected, homologous sequences not mostly covered by a coding sequence were excluded (<0.9, dashed vertical red line).



Supplementary Figure S32. In *Brevibacterium casei*, PglX is split to two proteins, with a variable C-terminus in the first.
Alignment between PglX proteins encoded in programmed inversion loci in *Lacticaseibacillus rhamnosus* and *Brevibacterium casei* (NC_013198.1, 2154002-2170387, and NZ_CP068173.1, 2184967-2200609, respectively, **Figure 4B**), and relation to variable region locations. Products of the PglX coding sequence downstream to the BrxC coding sequence in *L. rhamnosus* and *B. casei* are shown (top, Protein accession WP_015765014.1, encoded in Nucleotide accession NC_013198.1, 2161973-2165527, and bottom left, Protein accession WP_201670672.1, encoded in Nucleotide accession NZ_CP068173.1, 2193315-2195903, respectively), as well as the PglX coding sequence in *B. casei* that does not contain any inverted repeat (bottom right, Protein accession WP_236586838.1, encoded in Nucleotide accession NZ_CP068173.1, 2189649-2190596). In each coding sequence that contains any inverted repeats, non-variable regions were considered to be regions upstream to all repeats or downstream to all repeats. Non-variable regions in proteins were considered to be regions such that their corresponding codons are non-variable. "N-" and "-C" indicate protein N and C termini, respectively. Alignments are indicated by light orange projections

Bibliography

1. Reyes Ruiz,L.M., Williams,C.L. and Tamayo,R. (2020) Enhancing bacterial survival through phenotypic heterogeneity. *PLoS Pathog.*, **16**, e1008439.

2. Woude,M.W. van der, van der Woude,M.W. and Bäumler,A.J. (2004) Phase and Antigenic Variation in Bacteria. *Clinical Microbiology Reviews*, **17**, 581–611.

3. 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.

4. 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.

5. 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.

6. 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.

7. 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**.

8. 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.

9. 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.

10. Phillips,Z.N., Trappetti,C., Van Den Bergh,A., Martin,G., Calcutt,A., Ozberk,V., Guillon,P., Pandey,M., von Itzstein,M., Edward Swords,W., *et al.* (2022) A phasevarion controls multiple virulence traits, including expression of vaccine candidates, in Streptococcus pneumoniae. *bioRxiv*, 10.1101/2022.02.08.479631.

11. 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.

12. 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**, 1–9.

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

14. Komano,T. (1999) Shufflons: multiple inversion systems and integrons. *Annu. Rev. Genet.*, **33**,

171–191.

15. Sitaraman,R., Denison,A.M. and Dybvig,K. (2002) A unique, bifunctional site-specific DNA recombinase from Mycoplasma pulmonis. *Mol. Microbiol.*, **46**, 1033–1040.

16. 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.

17. Li,J.-W., Li,J., Wang,J., Li,C. and Zhang,J.-R. (2019) Molecular Mechanisms of Inversions in the Locus of Streptococcus pneumoniae. *J. Bacteriol.*, **201**.

18. Zabelkin,A., Yakovleva,Y., Bochkareva,O. and Alexeev,N. (2021) PaReBrick: PArallel REarrangements and BReaks identification toolkit. *Bioinformatics*, 10.1093/bioinformatics/btab691.

19. 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.

20. 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. U. S. A.*, **101**, 14919–14924.

21. 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.

22. 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.

23. 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 Research*, **48**, 11468–11485.

24. 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 Specificity Genes That Can Switch System Specificity by Recombination. *mSystems*, **5**.

25. 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.

26. Darling,A.E., Mau,B. and Perna,N.T. (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*, **5**, e11147.

27. Rognes,T., Flouri,T., Nichols,B., Quince,C. and Mahé,F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2584.

28. 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**, D192–6.

29. Grundy,F.J. and Howe,M.M. (1984) Involvement of the invertible G segment in bacteriophage mu tail

fiber biosynthesis. *Virology*, **134**, 296–317.

30. 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**.

31. Dybvig,K. and Yu,H. (1994) Regulation of a restriction and modification system via DNA inversion in Mycoplasma pulmonis. *Mol. Microbiol.*, **12**, 547–560.

32. 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.

33. Atack,J.M., Weinert,L.A., Tucker,A.W., Husna,A.U., Wileman,T.M., F Hadjirin,N., 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.

34. 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.

35. 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.

36. 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.

37. 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.

38. 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.

39. 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.

40. Price,M.N. and Arkin,A.P. (2017) PaperBLAST: Text Mining Papers for Information about Homologs. *mSystems*, **2**.

41. 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.

42. 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.

43. 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.

44. Anton,B.P. and Roberts,R.J. (2021) Beyond Restriction Modification: Epigenomic Roles of DNA Methylation in Prokaryotes. *Annu. Rev. Microbiol.*, **75**, 129–149.

45. 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.

46. 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.

תקציר

היפוכים כרומוזומליים מתוכננתים הם היפוכים מדויקים של אזורים גנומיים ספציפיים, אשר מתרחשים בתדירות גבוהה. ההיפוכים מזורזים על ידי אנזימי רקומבינציה, אשר מביאים להיפוכים של אזורים גנומיים שמשני צידיהם רצפים חוזרים הפוכים, כלומר רצפים חוזרים שמופיעים על גדילים שונים של הגנום. היפוכים מתוכננתים מאפשרים לחיידקים לייצר שונות גנטית ופונקציונלית בתוך האוכלוסייה, כאשר וריאנטים שונים מותאמים לסביבות ולתרחישים שונים. הודות לכך, היפוכים מתוכננתים מהווים אסטרטגיית גידור סיכונים שמשפרת את סיכויי ההישרדות של אוכלוסיית החיידקים במקרה של הופעה פתאומית של גורמי סיכון סביבתיים, דוגמת בקטריופאג'ים ותרופות אנטיביוטיות. חלק מההיפוכים המתוכננתים משנים רצפי בקרה, ובכך משפיעים על ביטוי גנים. לרוב מדובר בהיפוך של פרומוטר שגורם להשתקה או להפעלה של גן. לעומת זאת, חלק מההיפוכים המתוכננתים משנים רצפים מקודדים של גנים, ובכך מייצרים אללים שונים של הגן, שיובילו לייצור וריאנטים שונים של חלבון. היפוכים מתוכננתים שמשנים גנים נצפו בעיקר במשפחות גנים בעלות תפקיד מרכזי בתהליכי זיהוי, כגון זנב בקטריופאג', שעריות קוניוגציה, ומערכות אנזימי הגבלה מסוג I. עד כה, בוצעו רק שני חיפושים סיסטמטיים רחבים אחר היפוכים מתוכננתים שמשנים גנים. עם זאת, שני החיפושים הללו הוגבלו להיפוכים מתוכננתים שמשנים גנים שמקודדים למערכות אנזימי הגבלה מסוג I. לפיכך, חיפוש רחב וסיסטמי של היפוכים מתוכננתים שמשנים גנים, שאינו מוגבל למשפחות גנים שידועות ככאלו שהיפוכים מתוכננתים משנים אותן, לא נעשה עדיין. בהתאם, מעט ידוע כיום בנוגע לשכיחות היפוכים מתוכננתים בקרב משפחות גנים שונות, וכן בנוגע לארכיטקטורות גנומיות משותפות להיפוכים מתוכננתים. בעבודה זו סרקנו למעלה מ-35,000 מינים של חיידקים במטרה למצוא היפוכים מתוכננתים שמשנים גנים. תחילה, זיהינו כ-120,000 אזורים גנומיים כהיפוכים מתוכננתים פוטנציאלים. זיהוי זה נעשה על ידי חיפוש של רצפים חוזרים הפוכים אשר ביניהם רצפים שהיפוך שלהם יביא לשינוי של גנים. לאחר מכן השוונו אזורים גנומיים אלה לאזורים גנומיים דומים שהופיעו בגנומים של חיידקים מאותו זן, וזיהינו שונות גנטית שסביר שנבעה מהיפוכים מתוכננתים. על ידי השוואה בין היפוכים מתוכננתים פוטנציאלים עבורם מצאנו שונות לבין כאלה שלא מצאנו בהם שונות, חשפנו ארבע ארכיטקטורות גנומיות שמועשרות בהיפוכים מתוכננתים, ביניהן אסימטריות באורך של הגנים שמשנים ההיפוכים. בשלב הבא, פיתחנו מודל חיזוי של היפוכים מתוכננתים, שמתבסס על הארכיטקטורות הגנומיות שמצאנו. כעת, הפעלנו את המודל עבור כל ההיפוכים המתוכננתים הפוטנציאלים שזיהינו מוקדם יותר, והמודל חזה כי למעלה מ-14,000 מהם הם היפוכים מתוכננתים. על ידי השוואה בין היפוכים מתוכננתים פוטנציאלים שהמודל חזה כהיפוכים מתוכננתים לבין כאלה שהמודל לא חזה כהיפוכים מתוכננתים, מצאנו העשרה עבור גנים ממשפחות שידוע שהיפוכים מתוכננתים משנים אותן, וכן מערכות אנזימי הגבלה מסוג II, שלא היו ידועות עד כה ככאלה שהיפוכים מתוכננתים משנים אותן. בנוסף, סרקנו באופן ידני אזורים גנומיים שהמודל חזה ככאלה שמכילים היפוכים מתוכננתים שמשנים גנים שנמצאו כמועשרים. עבור מערכות שהופיעו באזורים גנומיים שונים, סרקנו נתוני ריצוף פומביים שהתקבלו ממכונות ריצוף של קריאות ארוכות, ומצאנו ראיות חזקות לכך שוריאנטים שונים הופיעו באותה דגימה שרוצפה, מה שמצביע על כך שאכן מדובר בהיפוכים מתוכננתים. באופן זה חשפנו היפוכים מתוכננתים שונים שמשנים גנים של מערכות אנזימי הגבלה מסוג II, היפוך מתוכנת שמשנה גן שמקודד לחלבון שהפונקציה שלו אינה ידועה, וכן היפוכים מתוכננתים שמשנים גן-היתוך שמקודד לחלבון שמכיל אזור של שעריות קוניוגציה וגם אזור של זנב בקטריופאג'. יחד, התוצאות שלנו חושפות ארכיטקטורות גנומיות אופייניות להיפוכים מתוכננתים, וכן מסמנות את משפחת מערכות אנזימי הגבלה מסוג II כמשפחה מרכזית בקרב היפוכים מתוכננתים שמשנים גנים.

המחקר נעשה בהנחיית פרופסור רועי קישוני בפקולטה לביולוגיה.

אני מודה לקרן ע"ש לאונרד ודיאן שרמן וכן לקרן הלאומית למדע על התמיכה הכספית הנדיבה בהשתלמותי.

זיהוי סיסטמטי של היפוכים מתוכנתים שמשנים גנים בחיידקים

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר מגיסטר למדעים בביולוגיה

אורן מילמן

הוגש לסנט הטכניון - מכון טכנולוגי לישראל

ניסן תשפ"ב, חיפה, אפריל 2022