Article

TnpB homologues exapted from transposons are RNA-guided transcription factors

https://doi.org/10.1038/s41586-024-07598-4

Received: 27 November 2023

Accepted: 23 May 2024

Published online: 26 June 2024

Check for updates

Tanner Wiegand[1], Florian T. Hoffmann[1], Matt W. G. Walker[2], Stephen Tang[1], Egill Richard[1], Hoang C. Le[1], Chance Meers[1] & Samuel H. Sternberg[1]✉

Transposon-encoded *tnpB* and *iscB* genes encode RNA-guided DNA nucleases that promote their own selfish spread through targeted DNA cleavage and homologous recombination[1–4]. These widespread gene families were repeatedly domesticated over evolutionary timescales, leading to the emergence of diverse CRISPR-associated nucleases including Cas9 and Cas12 (refs. 5,6). We set out to test the hypothesis that TnpB nucleases may have also been repurposed for novel, unexpected functions other than CRISPR–Cas adaptive immunity. Here, using phylogenetics, structural predictions, comparative genomics and functional assays, we uncover multiple independent genesis events of programmable transcription factors, which we name TnpB-like nuclease-dead repressors (TldRs). These proteins use naturally occurring guide RNAs to specifically target conserved promoter regions of the genome, leading to potent gene repression in a mechanism akin to CRISPR interference technologies invented by humans[7]. Focusing on a TldR clade found broadly in *Enterobacteriaceae*, we discover that bacteriophages exploit the combined action of TldR and an adjacently encoded phage gene to alter the expression and composition of the host flagellar assembly, a transformation with the potential to impact motility[8], phage susceptibility[9], and host immunity[10]. Collectively, this work showcases the diverse molecular innovations that were enabled through repeated exaptation of transposon-encoded genes, and reveals the evolutionary trajectory of diverse RNA-guided transcription factors.

Transposons play a central role in driving genome evolution and genome expansion, due to their proliferative nature and capacity for horizontal gene transfer, and the genes responsible for their mobility are among the most abundant genes in all of nature[11]. Although unchecked transposition poses a perennial threat that has spurred the evolution of cellular defence mechanisms, transposons also encode a vast repertoire of diverse enzymes that have been repeatedly repurposed by hosts, leading to the emergence of biological pathways as varied as intron splicing, immunoglobulin gene diversification, genome rearrangement, and genome defence[12,13]. Indeed, some of the most intricate cellular reactions involving nucleic acids have arisen in the genetic conflict, cooperation, and cooption between cells and transposable elements.

The origins of bacterial adaptive immune systems known as CRISPR–Cas can be traced directly to such host–transposon interactions, in which enzymes originally adapted for transposition were exapted[14] for novel roles in viral DNA acquisition and targeting[15,16]. The 'universal' *cas1* gene encodes an integrase responsible for preserving memories of past infections by splicing viral DNA fragments into the CRISPR array, in a biochemical reaction reminiscent of transposon integration[17,18], and ancestral CRISPR-less Cas1 homologues perform similar reactions within the context of transposable elements known as casposons[19,20]. Analogously, recent studies have demonstrated that the biochemical activities of Cas9 and Cas12, which perform RNA-guided DNA binding and cleavage during an immune response, can be traced back to ancestral transposon-encoded nucleases known as IscB and TnpB, respectively[1,2], which perform similar reactions to promote transposon maintenance and spread[3,4]. In turn, nuclease-deficient CRISPR–Cas systems have been repurposed by transposons on at least four independent occasions, to facilitate a novel RNA-guided DNA integration pathway mediated by CRISPR-associated transposases[21–23]. Similar cooption events have also frequently occurred between bacteria and bacteriophages[24,25], highlighting the extensive flux of genetic information between hosts and diverse types of mobile genetic elements[26].

Transposon-encoded TnpB proteins represent a vast reservoir of RNA-guided nucleases that are found in association with diverse transposons/transposases across all three domains of life[27–29]. In bacteria, *tnpB* genes are encoded within IS*200*/IS*605* and IS*607* family transposons, which are minimal selfish genetic elements that are mobilized by tyrosine-family and serine-family transposases (both named TnpA), respectively, but often exist in a non-autonomous form[30]. These transposons have conserved left end (LE) and right end (RE) sequences that define the boundaries of the mobile DNA, and in addition to protein-coding genes, they also encode non-coding RNAs, referred to as ωRNA (or reRNA), that feature a scaffold region spanning the transposon RE and an approximately 16-nt guide derived from the

[1]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA. [2]Department of Biological Sciences, Columbia University, New York, NY, USA.
✉e-mail: shsternberg@gmail.com

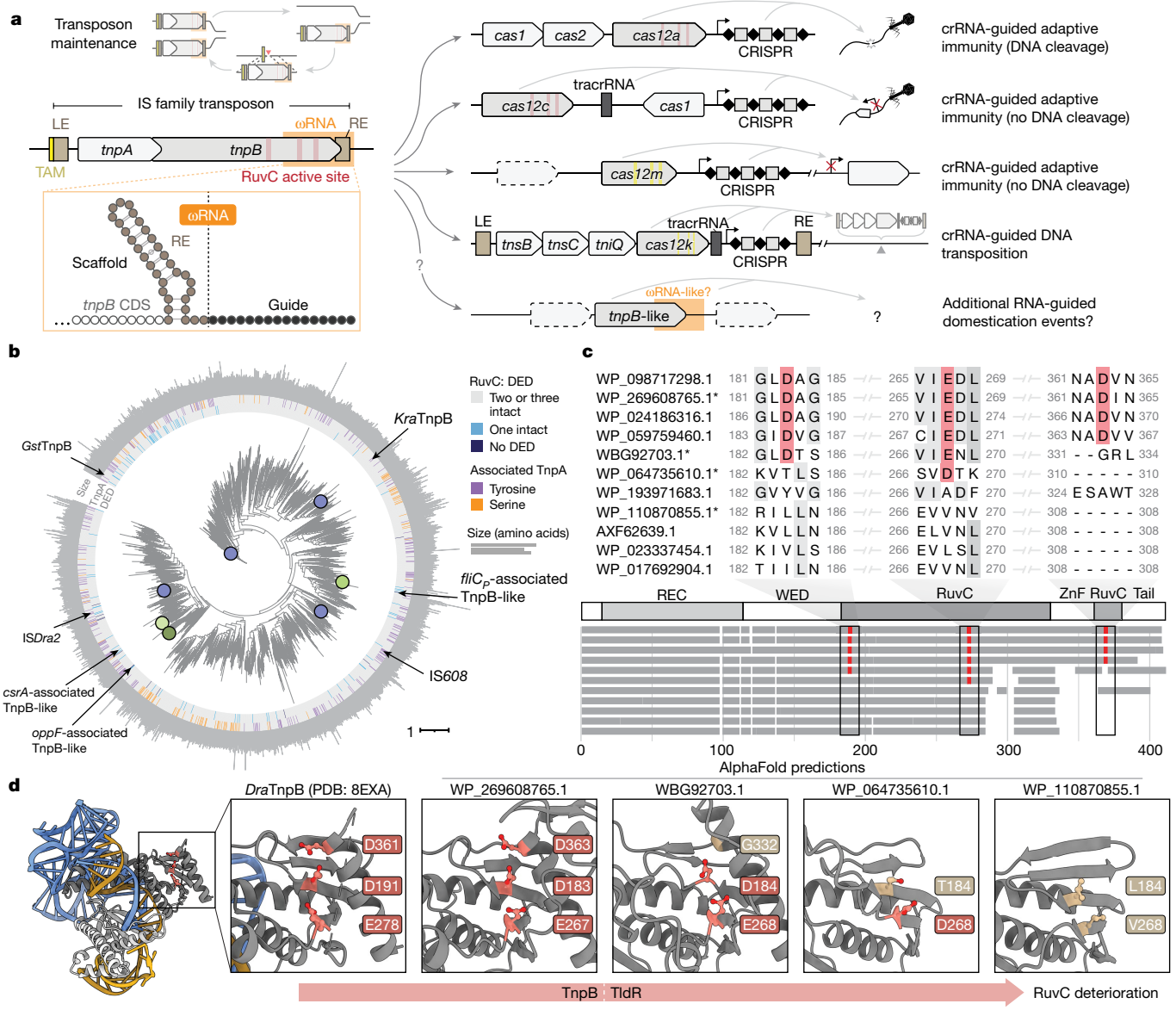


Fig. 1 | Bioinformatic identification of naturally occurring, nuclease-deficient TnpB homologues. a, TnpB proteins are RNA-guided nucleases that preserve bacterial transposons known as IS elements at sites of excision during transposition (left). Domestication of *tnpB* genes led to the evolution of diverse CRISPR-associated *cas12* derivatives, with diverse functions and mechanisms (right). CDS, coding sequence; LE, left end; RE, right end; crRNA, CRISPR RNA; tracrRNA, trans-activating crRNA. **b**, Phylogenetic tree of TnpB proteins, with previously studied homologues (blue) and newly identified TldR (green) proteins highlighted. The rings indicate RuvC DED active site intactness (inner), tyrosine-family or serine-family TnpA transposase association (middle), and protein size (outer). **c**, Multiple sequence alignment of representative TnpB and TldR homologues, highlighting deterioration of RuvC active site motifs and loss of the C-terminal zinc-finger (ZnF)/RuvC domain. Intact active site residues are highlighted in red, and highly conserved residues are shown in grey. **d**, Empirical (*Dra*TnpB) and predicted AlphaFold structures of TnpB and TldR homologues marked with an asterisk in **c**, showing progressive loss of the active site catalytic triad.

transposon-flanking sequence[1,2] (Fig. 1a). This guide sequence directs TnpB cleavage activity to complementary DNA sequences that are flanked by a cognate transposon/target-adjacent motif (TAM)[2–4].

TnpB nucleases have been independently domesticated numerous times over evolutionary timescales, leading to the emergence of dozens of unique CRISPR–Cas12 subtypes that feature diverse guide RNA (gRNA) requirements and PAM specificities[31,32]. In nearly all cases, Cas12 homologues rely on the same RuvC nuclease domain as TnpB for target cleavage. However, recent studies uncovered atypical Cas12 homologues – Cas12c and Cas12m – that have lost the ability to cleave target DNA but instead bind to and repress gene transcription as an alternative mechanism preventing mobile genetic element proliferation[33,34]. Type V-K CRISPR-associated transposases similarly rely on nuclease-inactivated Cas12k homologues that are still active for RNA-guided DNA binding, leading to programmable transposition[22,23,35] (Fig. 1a). Given the sheer abundance of *tnpB* genes and the profound utility of RNA-guided DNA binding – as exemplified in both biology and biotechnology[36] – we hypothesized that TnpB-like proteins may have been domesticated for novel functions, and we set out to test this hypothesis by specifically mining for nuclease-inactivated variants located in diverse genetic neighbourhoods.

Here we report the discovery of a novel family of TnpB-like nuclease-dead repressors (TldRs) that function not for transposition, but for RNA-guided transcriptional control, thus rendering the name 'TnpB (transposon/transposase B)' inapposite. Using a custom bioinformatics pipeline, we identified multiple independent TldR clades that

evolved from transposon-encoded TnpB nucleases via RuvC active site deterioration, coincident with newly acquired, non-transposase gene associations. TldRs function with adjacently encoded non-coding gRNAs to target complementary DNA sequences flanked by a TAM within promoter regions, and target binding downregulates gene expression through competitive exclusion of RNA polymerase. Flagellin (FliC)-associated TldR homologues are exploited by prophages to specifically remodel the host flagellar apparatus, which we discovered using in vivo genetic perturbation experiments in a clinical *Enterobacter* isolate. Collectively, this work reveals a novel evolutionary trajectory of transposon-derived, RNA-guided nucleases, and highlights the molecular opportunities afforded by transposon gene exaptation.

Detection of nuclease-dead TnpB proteins

We developed a bioinformatics pipeline to identify TnpB proteins with inactivating mutations in the RuvC domain, motivated by the hypothesis that these would represent likely gene exaptations for functions beyond transposon proliferation. We clustered a multiple alignment of 95,731 unique TnpB-like sequences, retrieved using a hidden Markov model (HMM) search at 50% sequence identity, and then performed an automatic assessment of the conservation of RuvC active site residues. TnpB nucleases, like Cas12 nucleases, exhibit a catalytic motif consisting of three acidic residues (DED), and mutating any residue in this motif abolishes nuclease activity[2,3,37]. However, recent analyses of TnpBs and eukaryotic TnpB-like proteins (that is, Fanzors) have indicated that one of the catalytic residues can occur at an alternate position in the RuvC domain[28]. Thus, we restricted our initial analysis to TnpB-like proteins with two or more mutations in the RuvC DED motif.

This search, supplemented with additional homologues that were identified in more focused analyses of three specific clades described below, identified 506 unique TnpB-like proteins with conserved mutations that are predicted to inactivate the RuvC nuclease domain (Fig. 1b and Supplementary Table 1). The polyphyletic distribution of these putatively inactivated nucleases suggest that they emerged on multiple occasions independently (Fig. 1b), and on the basis of their predicted roles in transcriptional repression (see below), we refer to them as TldRs. TldRs exhibit a range of deteriorated active sites, with one, two, or all three acidic residues mutated, and many homologues also feature C-terminal domain truncations that ablate RuvC and zinc-finger domains (Fig. 1c and Extended Data Fig. 1). AlphaFold predictions provided further structural support for the sequential deterioration of the RuvC active site, without more extensive degradation in the remainder of the overall TnpB/TldR fold or the RNA-binding interface (Fig. 1d), suggesting the intriguing possibility that RNA-guided DNA targeting functions could be preserved for these inactivated nucleases.

Viral *tldRs* associate with novel genes

Canonical *tnpB* genes in bacteria, alongside their ωRNA guides, are encoded within IS*200*/IS*605* family or IS*607* family transposons that can be straightforwardly identified using both comparative genomics and by defining the transposon ends[1–4]; in addition, a hallmark feature is their frequent association with *tnpA* transposase genes[30,38] (Fig. 2a, top). The genomic context surrounding *tldR* genes consistently lacked *tnpA* and identifiable LE/RE sequences, and instead, we observed strong genetic associations with novel, non-transposon genes that were often clade-specific (Figs. 1b and 2a). One TldR group is consistently associated with five to six genes encoding components of ABC transporter systems[39,40], the last of which is *oppF*, and is mainly present in enterococcal genomes (Supplementary Table 2). A second TldR group is tightly associated with *fliC*, a gene encoding the flagellin subunit of flagellar assemblies that propel bacteria in aqueous environments[10], and is found in diverse *Enterobacteriaceae* (Supplementary Table 3). A third TldR group from clostridial genomes is similarly associated with flagellin genes, in addition to a carbon storage regulator gene (*csrA*) that is involved in flagellar subunit regulation[41] (Supplementary Table 1). In all three cases, we observed loci encoding TldRs and their associated genes in varied genetic contexts, suggesting that they have maintained their associations over long timescales and/or that they have been mobilized in tandem. Strong genetic associations are also often indicative of functional coupling[42], indicating that TldR proteins may be involved in flagellar and ABC transporter expression and/or assembly pathways.

A closer inspection of genomic loci encoding *fliC–tldR* revealed the striking presence of numerous upstream genes with bacteriophage (phage) annotations, suggesting the potential presence of an integrated prophage (Fig. 2a and Supplementary Data Fig. 1a). When we used BLAST to search the NCBI non-redundant and whole-genome shotgun databases, we identified genomes that were highly similar to those encoding *fliC–tldR* but lacked phage genes, enabling us to confidently annotate boundaries of diverse prophages and their *attL*/*attR* recombination sequences that enclose *fliC–tldR* loci (Fig. 2b and Extended Data Fig. 2). Additional BLAST searches revealed two metagenome-assembled phage genomes in the taxon *Caudovirales* that contain *tldR* and its accompanying *fliC* gene (hereafter *fliCp*, for phage-encoded) (Supplementary Data Fig. 1b). Collectively, these data demonstrate that at least one TnpB domestication event involved the loss of nuclease activity, the loss of flanking transposon end sequences, and the gain of an accessory gene possibly linked to a novel function in phage biology (Fig. 2c). Of note, no similar bacteriophage associations were detected for *oppF*-associated or *csrA*-associated TldRs.

RIP-seq reveals mature TldR gRNAs

Transposon-encoded TnpB proteins function together with gRNAs transcribed from within or near the 3′ end of the *tnpB*-coding sequence to perform RNA-guided DNA cleavage[1,2]. Like crRNAs, these gRNAs comprise both an invariant 'scaffold' sequence that is a binding site for TnpB, as well as the 'guide' sequence that extends beyond the transposon RE and specifies target sites through complementary RNA–DNA base pairing. Numerous in silico strategies can be applied for gRNA identification, including comparative genomics, the ISfinder database[43], covariance models of the gRNA structure, and sequence alignments (Fig. 3a). Using these strategies, we identified the LE/RE boundaries and gRNAs associated with nuclease-active TnpB homologues that are closely related to $fliC_P$-associated and *oppF*-associated TldRs (Fig. 3b). Similar analyses also revealed the predicted 3–5-bp TAM sequences recognized by these TnpB homologues during DNA binding and cleavage[1–3] (Fig. 3b), akin to the role of PAM in DNA binding and cleavage by CRISPR–Cas9 and CRISPR–Cas12 (ref. 44).

The absence of identifiable transposon ends flanking *tldR* rendered similar annotations of its gRNA unfeasible, so we used covariance models built from gRNA sequences of related TnpB homologues. We hypothesized that TldR-associated gRNAs would be encoded near the gene, and after scanning a 500-bp window flanking each *tldR* gene with the gRNA covariance models, we identified high-confidence gRNA-like sequences for both $fliC_P$-associated and *oppF*-associated *tldR* loci (Supplementary Data Fig. 2). When we analysed published RNA sequencing (RNA-seq) datasets[45] from organisms with $fliC_P$–*tldR* or *oppF–tldR*, we observed read coverage over the regions identified by our covariance model search, as well as over the TldR open reading frame (Fig. 3c), providing evidence of native gRNA expression from regions flanking *tldR* loci.

To determine whether TldR proteins bind to their associated gRNAs, we cloned a representative FLAG-tagged $fliC_P$-associated TldR (*Eho*TldR) and *oppF*-associated TldR (*Efa1*TldR) into expression plasmids, alongside 240 bp encompassing the putative gRNA scaffold and a 20-bp guide sequence. After performing RNA immunoprecipitation sequencing

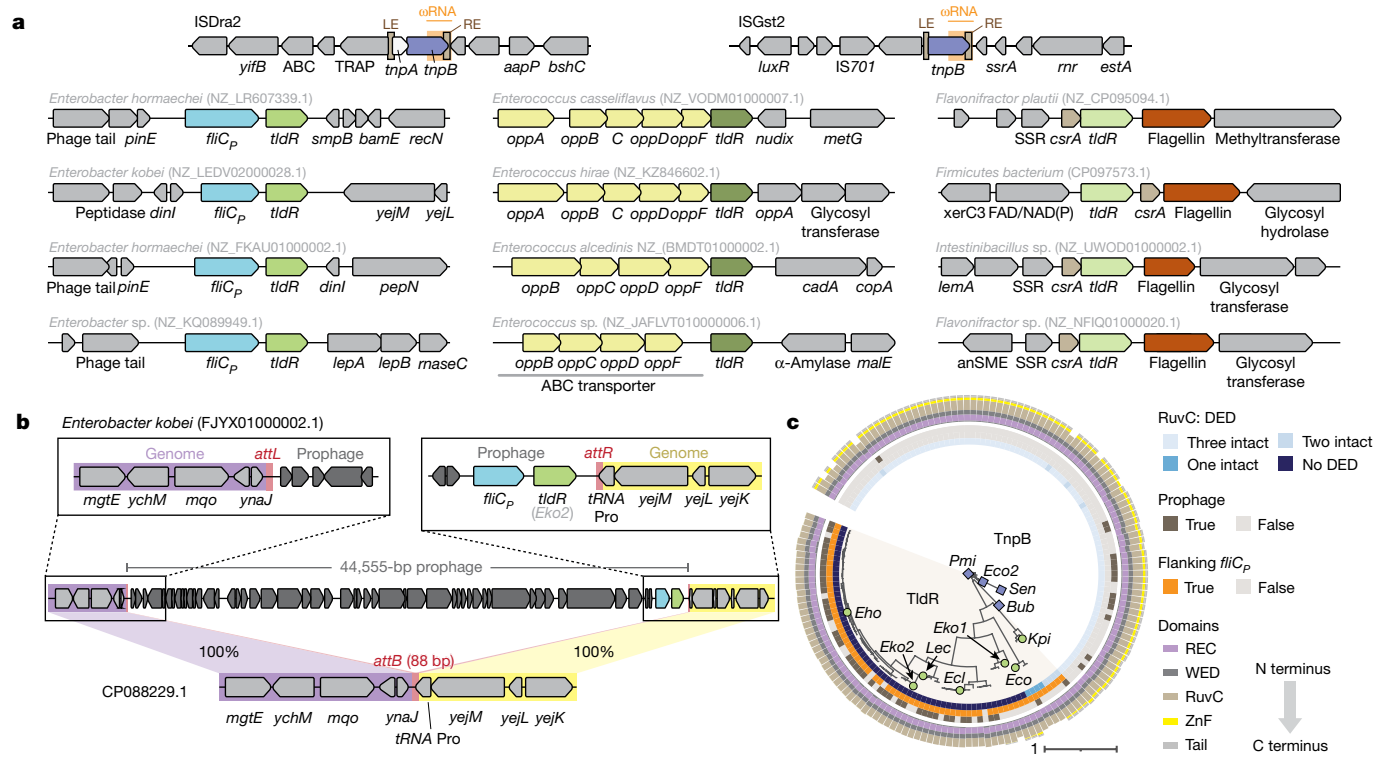


Fig. 2 | *tldR* genes are strongly associated with diverse non-transposon genes and encoded in prophages. a, Genomic architecture of well-studied transposons that encode TnpB (top), and of novel regions that encode TldR (bottom) in association with prophage-encoded *fliC*$_P$ (left), *oppF* and ABC transporter operons (middle), and a transcriptional regulator (*csrA*) of an accompanying flagellin (right). Genomic accessions are shown above the genetic maps. **b**, Comparison of a representative *fliC*$_P$–*tldR* locus with a closely related *Enterobacter kobei* strain reveals that the entire locus is encoded within the boundaries of a prophage element, with identifiable recombination sequences (*attL*/*attR*/*attB*). **c**, Phylogenetic tree of *fliC*$_P$-associated TldR proteins from **a**, together with closely related TnpB proteins that contain intact RuvC active sites. The rings indicate RuvC DED active site intactness (inner), prophage association (middle), *fliC*$_P$ association (middle), and TldR/TnpB domain composition (outer). Assessments of prophage and *fliC*$_P$ associations are described in the Methods. Homologues marked with a blue square (TnpB) or green circle (TldR) were tested in heterologous experiments.

(RIP-seq) experiments and mapping reads to the *Escherichia coli* genome and expression plasmid (Fig. 3d), we identified a mature, approximately 113-nt gRNA for *Eho*TldR that encompassed a 97-nt scaffold upstream of a 16-nt guide, indicating processing from the initial transcript down to a final mature form (Fig. 3e). Previous work has shown that TnpB proteins catalyse processing of their own guides through a RuvC-dependent mechanism[4,46]; however, the absence of an intact catalytic triad in TldR proteins suggests that the mature gRNA may instead represent the sequence protected from cleavage by cellular ribonucleases.

Unexpectedly, RIP-seq revealed that the *oppF*-associated *Efa1*TldR bound an even shorter gRNA, comprising a 100-nt scaffold and an approximately 9-nt guide (Extended Data Fig. 3a), and a similarly truncated guide (11 nt) was also observed for another homologue from publicly available RNA-seq data[45] (Extended Data Fig. 3b). RIP-seq data from replicates and five additional homologues corroborated the short guide for *Efa1*TldR (Extended Data Fig. 3c) while revealing more heterogeneous processing for diverse homologues (Supplementary Data Fig. 3).

TldR gRNAs target conserved promoters

We reasoned that identifying the putative gRNA substrates of TldRs would provide a major clue to their biological function, and thus made this our next objective. We extracted guide sequences based on our bioinformatics (Fig. 3f) and RIP-seq results, and then used these as queries in BLAST searches to identify potential genomic targets of *fliC*$_P$-associated TldR. The strongest match was in a genomic region that encodes other flagellar components and, strikingly, was specifically located in the intergenic region between *fliD* and a second (host) *fliC* gene distinct from the prophage-encoded *fliC*$_P$ orthologue (Fig. 3g). In *E. coli*, *fliC* expression is regulated by an alternative sigma factor (σ^{28}) also known as FliA[47,48]. The putative targets of multiple homologous TldR-associated gRNAs directly overlapped the *fliA* −10 promoter element, and were flanked by a conserved GTTAT motif that is highly similar to the TAM recognized by TnpB nucleases similar to TldR (Fig. 3h). The identity of these targets immediately suggested a model for *fliC*$_P$–*tldR* function, in which phage-encoded TldR–gRNA complexes could repress expression of the host FliC protein while producing their own FliC$_P$ homologue.

We performed a separate search for native targets of *oppF*-associated TldRs using the shortened 9-nt guide, combined with the predicted TAM recognized by related TnpB nucleases (TTTAA or TTTAT) (Extended Data Fig. 4a). This analysis led to the identification of a conserved target upstream of the start codon of one of the ABC transporter genes (*oppA*) encoded proximally to *tldR* (Extended Data Fig. 4b,c). *tldR*-associated OppA homologues are most similar to substrate-binding proteins that recognize short polypeptides in ABC transport systems[40,49], and the putative TldR targets overlap the predicted promoter[45,50] of *oppA*, suggesting that TldR would repress *oppA* transcription (Extended Data Fig. 4d,e). We also identified other potential gRNA targets in genomes encoding *oppA*–*tldR* loci, raising the possibility that these TldR proteins contribute towards a more complex transcriptional regulatory network than *fliC*$_P$-associated TldR proteins (Extended Data Fig. 5).

Together, these data strongly support the hypothesis that TldR proteins across multiple independent lineages function as RNA-guided transcription factors to regulate gene expression, and that their

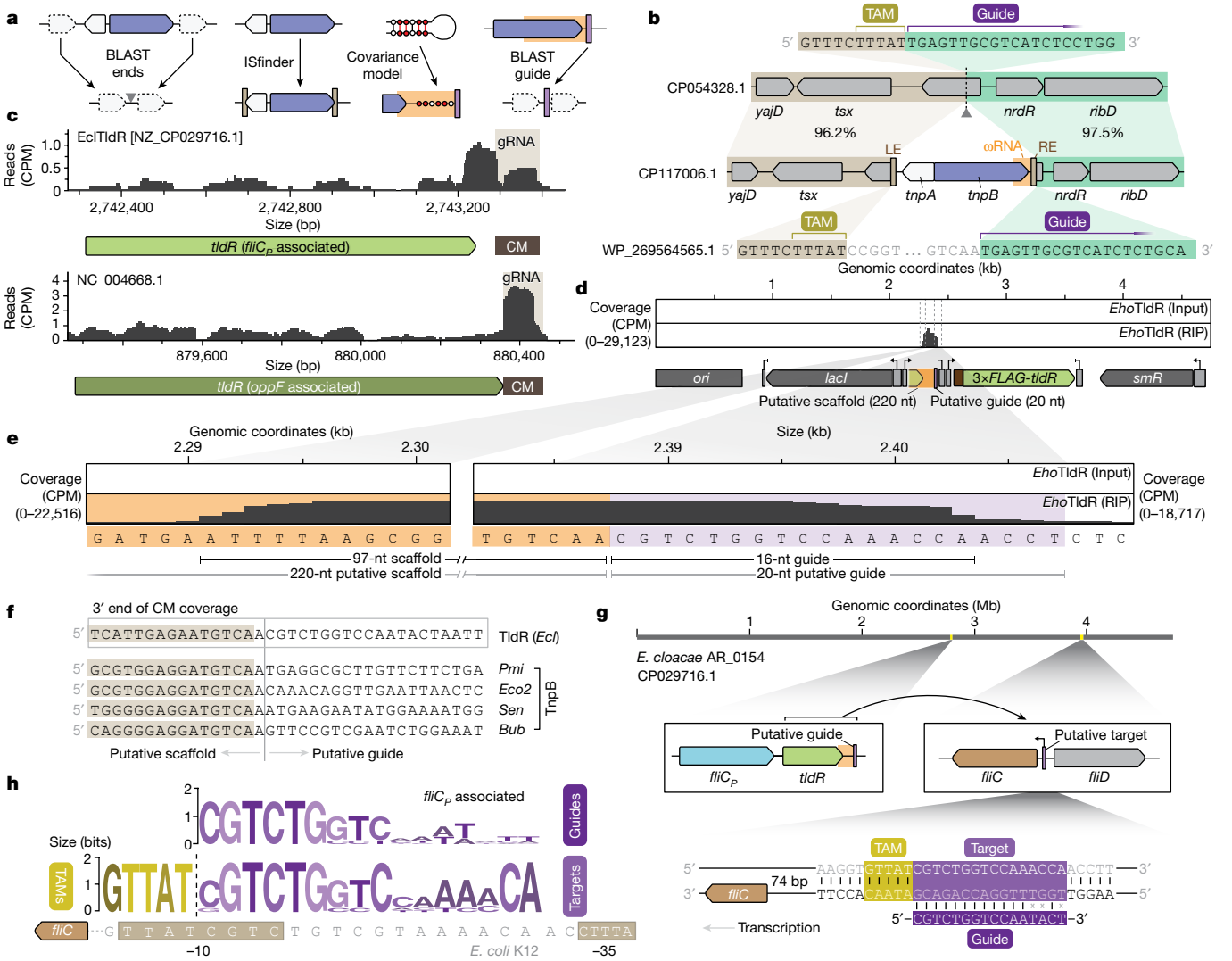


Fig. 3 | TldR proteins are encoded next to gRNAs that target conserved genomic sites. a, Bioinformatic strategies to investigate *tldR*/*tnpB* loci, including comparative genomics, searching within the ISfinder database, gRNA prediction using covariance models, and target prediction using BLAST. **b**, Representative *tnpB* locus and an isogenic locus above that lacks the IS element. Comparison reveals the putative TAM recognized by TnpB, which flanks the transposon LE, and the guide portion of the ωRNA, which flanks the transposon RE. **c**, Published RNA-seq data for *E. cloacae* (top) and *Enterococcus faecalis* (bottom) reveal evidence of native *tldR* and gRNA expression for *fliC*$_P$-associated and *oppF*-associated TldRs, respectively. The predicted gRNAs from covariance model (CM) analyses are indicated, and unique genome-mapping reads are shown as overlays of three replicates. **d**,**e**, RIP-seq data from a *fliC*$_P$-associated TldR homologue from *Enterobacter hormaechei* (*Eho*TldR) reveals the mature gRNA boundaries; an input control is shown. **f**, Alignment of a representative *fliC*$_P$–*tldR* locus from *E. cloacae* (*Ecl*) and related *tnpB* loci at a region near the 3′ end of covariance model coverage (shaded in brown) reveals conservation of 3′ scaffold sequences. **g**, Analysis of the guide sequence from the *Ecl*TldR-associated gRNA in **f** revealed a putative genomic target near the predicted promoter of a distinct (host) copy of *fliC* located approximately 1 Mb away (middle). The magnified schematic (bottom) shows the predicted TAM and gRNA–target DNA base-pairing interactions, relative to the *fliC* coding sequence at left. **h**, WebLogos of predicted guides and genomic targets associated with diverse *fliC*$_P$-associated TldRs from Fig. 2c (top). Targets overlap conserved promoter elements recognized by FliA/σ^{28} in *E. coli* K12 (bottom).

biological targets relate to the accessory genes with which they stably associate.

TldRs are RNA-guided transcription factors

We selected seven *fliC*$_P$-associated (Fig. 2c) and eight *oppF*-associated (Extended Data Fig. 1a) TldR homologues for functional assays, which were chosen to sample the diversity within each clade (Supplementary Data Fig. 4), cloned them into expression vectors alongside their putative gRNAs, and expressed them in an *E. coli* K12 strain containing a genomically integrated target site (see Methods). We profiled genome-wide binding specificity using chromatin immunoprecipitation followed by sequencing (ChIP–seq), and the resulting data revealed strongly enriched peaks corresponding to the expected target site for nearly all homologues tested (Fig. 4a and Supplementary Data Fig. 5). These data demonstrate that TldR proteins retain the ability to perform highly specific, RNA-guided DNA target binding in cells, despite possessing RuvC mutations and C-terminal truncations.

We next analysed prominent off-target peaks in the ChIP–seq dataset (Extended Data Fig. 6). One of these off-target peaks for *fliC*$_P$-associated TldRs corresponded to the intergenic region between *E. coli fliC* and *fliD* (Fig. 4a,b). The guide sequence used in these experiments is complementary to the native *fliC* target from *Enterobacter cloacae* sp. AR_154 but mutated relative to the *E. coli* K12 sequence at five positions (Fig. 4b), suggesting a high tolerance for TldR binding to mismatched targets (Extended Data Fig. 6). These data support the definition of an

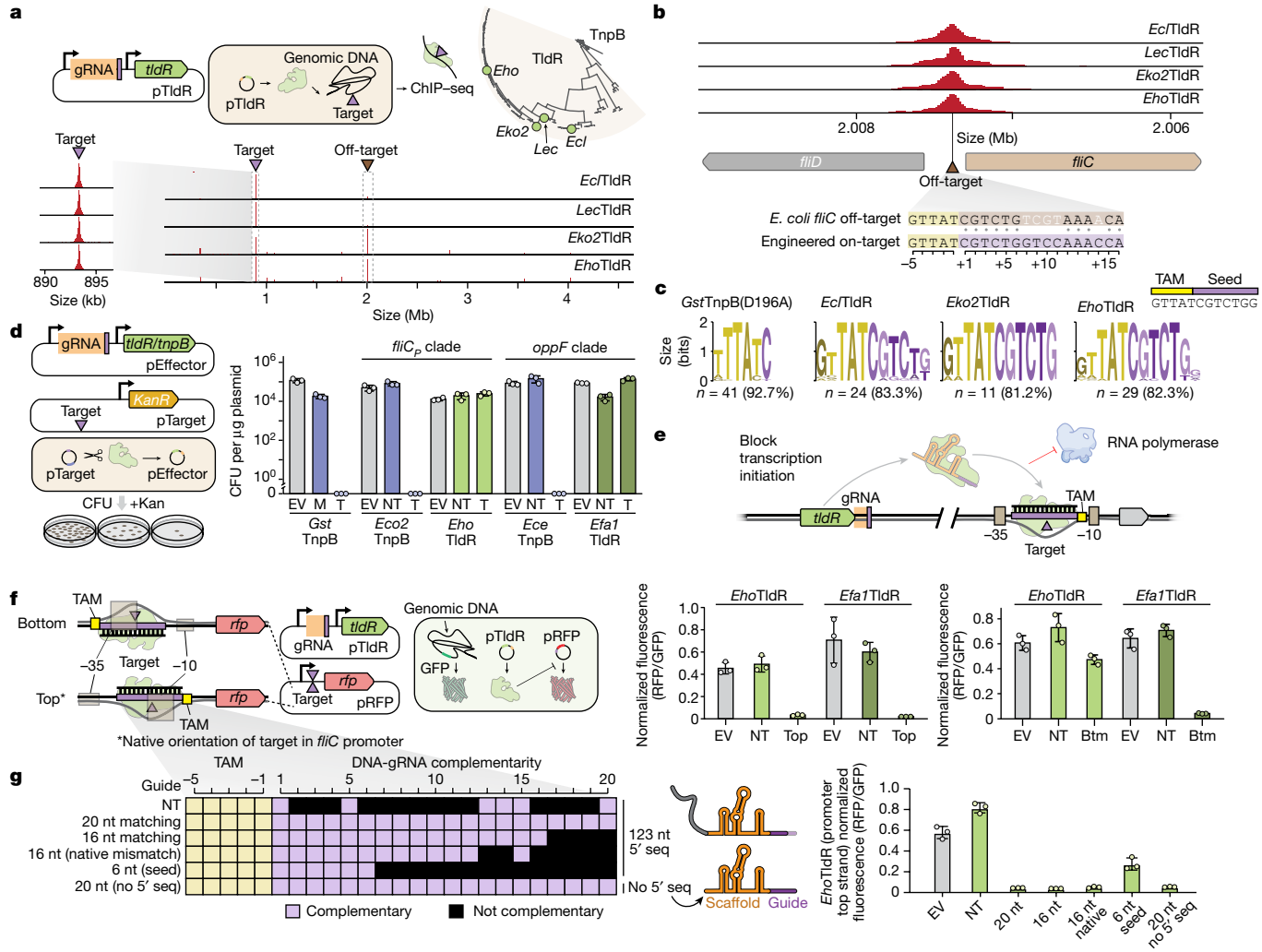


Fig. 4 | TldRs are RNA-guided DNA binding proteins capable of programmable transcriptional repression. a, ChIP–seq approach to investigate RNA-guided DNA binding for TldR candidates (top), and representative ChIP–seq data revealing strong enrichment at the genomic target site and a prominent off-target (bottom). **b**, Magnified view of ChIP–seq peaks at the off-target site in **a**, which corresponds to a TAM and partially matching target sequence at the *E. coli* K12 *fliC* promoter. **c**, Analysis of conserved motifs bound by the indicated TldR homologue using MEME ChIP reveals specificity for the TAM and an approximately 6-nt seed sequence. The number and percentage of total called peaks contributing to each motif are indicated; low-occupancy positions were manually trimmed from the motif 5′ ends. **d**, Schematic of *E. coli*-based plasmid interference assay using pEffector and pTarget (left), and bar graph plotting colony-forming units (CFU) for the indicated conditions and proteins (right). TldR homologues have no effect on cell viability, indicating a lack of DNA cleavage activity. Bars indicate mean ± s.d. ($n = 3$ biological replicates). EV, empty vector; Kan, kanamycin; M, TnpB mutant; NT, non-targeting guide; T, targeting guide. **e**, Model of TldR-mediated transcriptional repression by blocking RNAP transcription initiation (blue). **f**, gRNAs were designed to target the top or bottom strand of a promoter driving *rfp* expression (left). The bar graphs plot normalized RFP fluorescence for the indicated conditions (right). Top, gRNA targeting the top strand; Btm, gRNA targeting the bottom strand. Bars indicate mean ± s.d. ($n = 3$ biological replicates). **g**, RFP repression activity (right) using modified gRNAs exhibiting variable complementarity to the target site (schematized at left). A gRNA was also tested that lacked the extra 5′ sequence, which was absent in RIP-seq reads of mature gRNAs (20 nt no 5′ seq). Bars indicate mean ± s.d. ($n = 3$ biological replicates).

approximately 6-nt TldR seed sequence, consistent with the previously reported 6-nt seed for some Cas12a homologues[51].

ChIP–seq also captures transient interactions due to the crosslinking step, and we reasoned that systematic analysis of all peaks could report on the underlying TAM specificity of select TldR homologues, as we previously showed for TnpB[3]. Using MEME to detect enriched motifs, we found that *fliC*$_P$-associated TldRs were enriched at 5′-GTTAT-3′ motifs, the same pentanucleotide TAM that flanks putative TldR–gRNA targets within *fliC* promoters (Fig. 4c and Supplementary Data Fig. 5). Similarly, *oppF*-associated TldR homologues bound DNA sequences enriched in 5′-TTTAA-3′ motifs, consistent with the bioinformatically predicted TAM specificities for their closely related TnpB relatives (TTTAA and TTTAT) (Supplementary Data Fig. 6). Taken together, these results indicate that TAM sequences for TldR proteins can be accurately predicted via ChIP–seq and in silico motif detection, even without the transposon context clues used for TnpB nucleases[2,3].

To verify that the naturally occurring RuvC mutations in TldR proteins actually abolish nuclease activity, we tested TldR homologues or their related TnpB counterparts in plasmid interference assays. Effector expression plasmids (pEffector) encoding TldR or TnpB and their associated gRNA were used to transform *E. coli* cells, along with a target plasmid (pTarget) bearing a kanamycin resistance cassette and a TAM-flanked target sequence (Fig. 4d). Nuclease activity is expected to eliminate pTarget, resulting in fewer surviving colonies when cells are plated on selective media. A previously studied TnpB homologue[3] (that is, *Gst*TnpB3) and nuclease-active TnpB homologues similar to

TldRs (that is, *Eco2*TnpB and *Ece*TnpB) reduced colony-forming units in a gRNA-specific manner, whereas TldR homologues produced colony counts comparable with empty vector controls (Fig. 4d and Extended Data Fig. 7), confirming that TldR proteins function as RNA-guided DNA binding proteins that lack the ability to cleave DNA.

Finally, we set out to investigate whether target DNA binding by TldR could modulate gene expression, akin to the engineered use of nuclease-dead Cas9 and Cas12 variants in CRISPR interference (CRISPRi) applications[52,53]. To test this, we developed an RFP/GFP reporter assay in which target DNA binding represses *rfp* gene expression relative to a control *gfp* locus[52], and designed gRNAs to either occlude transcription initiation by targeting promoter sequences (Fig. 4e,f), or to block transcription elongation by targeting the 5′ untranslated regions. With promoter-targeting gRNAs, we found that representative *fliC*$_P$-associated (*Eho*) and *oppF*-associated (*Efa1*) TldR homologues robustly repressed RFP fluorescence when targeting the top (that is, sense) strand (Fig. 4f), in agreement with polarity effects previously observed for synthetic nuclease-dead Cas12a variants[53,54]. gRNAs with shorter stretches of RNA–DNA complementarity (down to just 6 nt) produced comparable levels of gene repression as 20-nt guides, regardless of whether naturally occurring guide–target mismatches were present (Fig. 4g and Extended Data Fig. 8a). When we instead targeted the 5′ untranslated region (Extended Data Fig. 8b), select TldRs from both clades only efficiently repressed RFP when targeting the bottom strand, in which the TAM-proximal end was oriented towards the promoter and elongating RNA polymerase (RNAP), at efficiencies slightly below dCas9 and dCas12 (Extended Data Fig. 8c).

Collectively, our results demonstrate that TldRs lack any detectable cellular nuclease activity and instead function as RNA-guided DNA binding proteins, with the potential to potently repress gene expression in a mechanism reminiscent of engineered, nuclease-dead CRISPR–Cas effectors.

Viral TldRs natively repress host *fliC*

Having analysed TldRs in heterologous contexts, we next investigated their native function, focusing specifically on *fliC*$_P$–*tldR* loci. FliC is the major extracellular subunit that polymerizes in tens of thousands of copies to form mature flagellar filaments, enabling bacterial locomotion[8] (Fig. 5a). When we compared host FliC and prophage FliC$_P$ sequences, we found that solvent-exposed regions (domains D2–3)[55,56] were highly variable, whereas inter-protomer contacting regions (domains D0–1)[55,56] were highly conserved (Fig. 5b,c). This suggests that prophage flagellin would probably retain the ability to form flagella together with host components, while nevertheless diversifying the chemical composition of exposed filament surfaces[57,58].

To test the hypothesis that TldRs repress host *fliC* expression to remodel flagella, we obtained and cultured three *Enterobacter* strains that each had a prophage-encoded *fliC*$_P$–*tldR* locus, alongside a closely related control strain that lacked it, and performed total RNA-seq. Each strain with *tldR* exhibited robust gRNA expression, with 5′ and 3′ boundaries that were in excellent agreement with our heterologous RIP-seq data (Extended Data Fig. 9). When we analysed flagellin gene expression relative to the flagellar hook (*fliD*), we found that host *fliC* was nearly undetectable in all three strains that encoded *tldR*, whereas *fliC*$_P$ was strongly expressed (Fig. 5d), consistent with our hypothesis on TldR–gRNA function. By contrast, *fliC* was highly expressed in the control strain that lacked TldR and the prophage (Fig. 5d).

Next, to prove that *fliC* downregulation was a direct consequence of TldR-mediated repression, rather than an indirect effect relating to the complex regulatory network controlling flagellar assembly[59], we generated precise genetic perturbations to the *fliC*$_P$–*tldR* locus in *Enterobacter* sp. BIDMC93 and measured the corresponding effects on host *fliC* expression by quantitative PCR with reverse transcription (RT–qPCR). Deletion of *tldR*, *tldR*–gRNA, the entire *fliC*$_P$–*tldR*–gRNA locus, or the entire prophage all led to an approximately 100-fold increase in host *fliC* expression, and crucially, the same increase was observed after substituting the guide portion of the gRNA with a non-targeting control sequence (Fig. 5e). Repression of *fliC* was rescued by complementing the *tldR*–gRNA deletion mutant with a plasmid-encoded *tldR*–gRNA cassette (Fig. 5e). RNA-seq of three biological replicates revealed clear evidence of host *fliC* de-repression when the genomically encoded guide sequence was mutated (Fig. 5f), as well as strong expression of *fliC*$_P$, *tldR*, and gRNA at similar levels as genes involved in lysogeny maintenance (Extended Data Fig. 10a). Differential gene expression analyses further revealed that *fliC* was the most strongly upregulated (that is, de-repressed) gene transcriptome wide (Fig. 5g), with the only other significant changes arising in genes whose expression has been linked to flagellar gene transcription[60,61].

Enterobacter mutants with deletions of *fliC*$_P$ or the entire prophage remained motile in swimming assays, at levels comparable to or slightly greater than wild type (Extended Data Fig. 10b). Liquid chromatography with tandem mass spectrometry analyses of flagellar samples isolated from these strains further revealed that FliC$_P$ is both expressed and partitions in the flagellar fraction, consistent with transcriptomic data, and that deletion of the prophage (including *tldR* and the gRNA) de-repressed host FliC expression at the protein level (Extended Data Fig. 10c–e). In summary, these data indicate that host flagella are remodelled via the coordinated repression of host *fliC* and coincident incorporation of *fliC*$_P$ gene products into chimeric assemblies with other host flagellar subunits.

Closer inspection of the RNA-seq data lent further support to our conclusion that TldR represses gene expression through competitive binding to promoter elements, as the *fliC* transcription start site agreed with the −35 and −10 promoter annotations informed from FliA/σ^{28} data in *E. coli* K12 (Fig. 5h and Extended Data Fig. 10g,h). This interpretation was also corroborated by comparisons of predicted TldR–gRNA–DNA structures with an experimentally determined RNAP–FliA–DNA holoenzyme structure, which demonstrate that TldR target binding would sterically block FliA access to DNA (Fig. 5i). To determine how prophage-encoded *fliC*$_P$ genes would escape TldR-mediated repression, we applied MEME and Tomtom to identify conserved motifs in the region upstream of the experimentally determined *fliC*$_P$ transcription start site (Extended Data Fig. 10i,j). These analyses revealed that prophages probably recruit the very same host FliA/σ^{28} transcriptional program to produce FliC$_P$, but with highly conserved mutations in both the TAM and the seed sequence that preclude TldR–gRNA recognition (Fig. 5j). Collectively, the *fliC*$_P$–*tldR* locus is therefore elegantly adapted to remodel composition of the flagellar apparatus upon establishment of a lysogen, by selectively repressing host flagellin through RNA-guided DNA targeting while hijacking cellular machinery to express its own homologue substitute (Fig. 5k).

Discussion

Bacterial flagella represent a critical nexus at the host–pathogen interface, and the attendant selective pressures probably contributed to the domestication and emergence of *fliC*-associated *tldR* genes on at least two independent occasions, to sensitively regulate flagellar expression and composition (Fig. 2a). Several bacteriophages, including the well-studied *Salmonella* Chi phage, recognize the flagellar filament as a receptor during cell absorption[9,58,62], and phage-mediated substitution of host flagellin may thus prevent superinfection and/or render the cell invisible to competing flagellotropic phages in the environment (Fig. 5k). FliC, also known as the H antigen in bacterial pathogen serotyping, also functions as a primary antigen that is recognized by both receptors and antibodies in the mammalian innate and adaptive immune systems[63–66], and the pervasive presence of prophage-encoded *fliC*$_P$–*tldR* loci in clinical isolates from humans (Supplementary Table 3) could represent a novel example of lysogenic conversion[67], in which

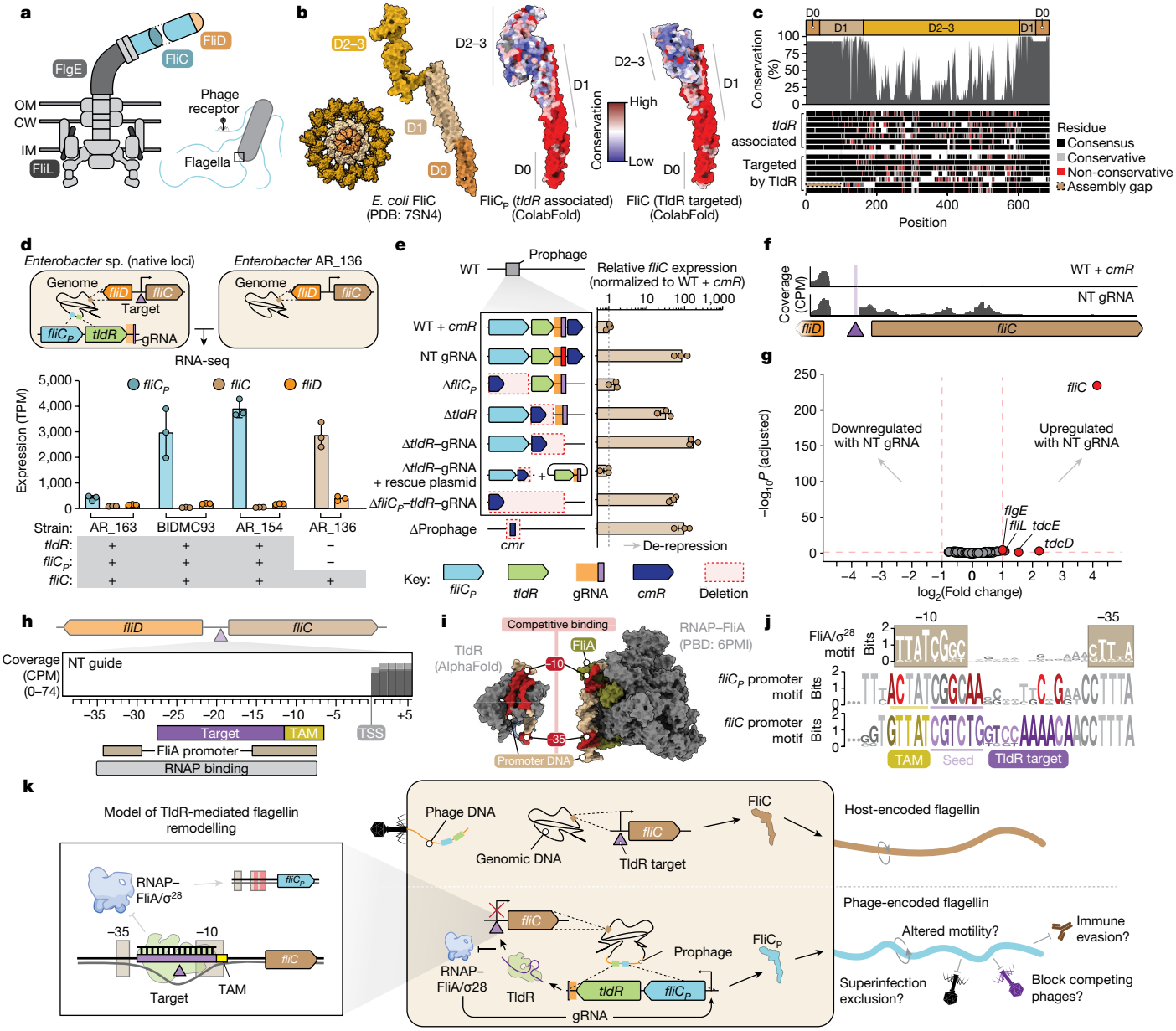


Fig. 5 | Flagellin-associated TldRs repress host flagellin gene expression in native *Enterobacter* strains. a, Flagellar assemblies span the inner membrane (IM), cell wall (CW) and outer membrane (OM). Flagellin (FliC) filaments comprise several thousand subunits and are receptors of flagellotropic phages. **b**, Surface representation of *E. coli* FliC coloured by domains, for both a filament cross-section and a single monomer (left). ColabFold-predicted prophage FliC$_P$ (middle) and host FliC (right) structures from *E. cloacae*, coloured with AL2CO conservation scores calculated from the multiple sequence alignment (MSA) shown in **c**. **c**, MSA of prophage FliC$_P$ and host FliC proteins, showing highly conserved D0–1 domains and hypervariable D2–3 domains. **d**, *Enterobacter* strains selected for RNA-seq analysis (top), and expression data plotted as transcripts per million mapped reads (TPM) for *fliC$_P$* (when present) and host *fliC* and *fliD*. Bars indicate mean ± s.d. ($n = 3$ biological replicates). **e**, *E. cloacae* mutants generated by recombineering (left), and RT–qPCR analysis of host *fliC* expression levels normalized to the wild-type (WT) strain. NT, non-targeting. Bars indicate mean ± s.d. ($n = 3$ biological replicates). **f**, RNA-seq coverage at the host *fliC* locus overlayed for three biological replicates of the indicated strains in **e**. The gRNA-matching target site is represented with a purple triangle. **g**, Differential gene expression analysis for the WT and NT gRNA strains in **f**. Wald test *P* values adjusted for multiple comparisons were calculated using the Benjamini–Hochberg approach. Genes with a $\log_2$(fold change) ≥ 1 and an adjusted $P < 0.05$ are highlighted in red. **h**, Magnified view of data in **f**, showing the TAM/target overlap with the predicted FliA/σ^{28} promoter. TSS, transcription start site. **i**, Predicted AlphaFold structure of TldR bound to target DNA (left; superimposed from PDB 8EXA) compared with a similarly scaled experimental structure of RNAP (grey) and FliA/σ^{28} (green) bound to promoter DNA (right). **j**, Comparison of promoter motifs for host *fliC*, prophage *fliC$_P$*, and the FliA/σ^{28} motif. **k**, Model for the role of TldR in RNA-guided repression of host *fliC*.

phages enable their bacterial hosts to evade an immune response. Finally, flagellar remodelling could modulate motility of bacterial host cells, and thus impact their capabilities for chemotaxis and nutrient acquisition. Resolving how RNA-guided repression of host flagellin gene expression impacts one or more facets of bacterial physiology, and whether flagellar composition is dynamically altered over time in these lysogenic strains, will be a major goal of future research efforts.

The biological purpose of *oppF*-associated TldR homologues that are encoded next to ABC transporter operons (Fig. 2a) is less clear, although our gRNA discovery approach – blending both

bioinformatics predictions and experimental validation via RIP-seq – revealed a likely function in controlling expression of the key periplasmic-binding protein OppA (Extended Data Fig. 4). ABC transporters are ubiquitous membrane-bound protein complexes that move substrates in and out of cells, and are responsible for nutrient uptake, drug resistance, toxin efflux, and virulence factor secretion, among many other roles[40,68]. TldR proteins may provide a mechanism to regulate *oppA* expression in response to external biological cues, although more experiments will be needed to better understand their activities and specificities in vivo, especially given the truncated guide sequences they use.

By integrating knowledge about the biochemical properties of closely related TnpB nucleases, including their TAM and ωRNA requirements, together with systematic biochemical profiling using ChIP–seq and reporter assays, we were able to straightforwardly identify the gRNAs and targets recognized by TldR proteins, providing an important advance beyond descriptive bioinformatic observations of similar loci[31]. It appears certain that additional examples of TnpB domestication will be uncovered with further bioinformatic and experimental mining, for both bacterial and eukaryotic Fanzor homologues[28,29], and future efforts should be broadened to include nuclease-dead and nuclease-active variants that exist in non-transposon and non-mobile genetic element contexts.

It is noteworthy that evolution has repeatedly sampled some of the very same molecular innovations invented by humans during the development of CRISPR-based genome-engineering technologies. A decade ago, to our knowledge, Qi and colleagues developed the first applications of synthetic, nuclease-dead variants of Cas9 (that is, dCas9) for transcriptional modulation[52], and intense efforts ever since have resulted in a multitude of highly effective tools for epigenome editing, typically via engineered fusions of dCas9 to diverse effector domains[69]. The miniature size of TldR proteins (mean of 327 amino acids) renders them appealing platforms for similar engineering applications; however, further experiments are needed to determine the flexibility of TAM recognition and DNA targeting specificity in organisms with larger genomes. Indeed, the strong gene repression that we observed with a 6-bp guide–target duplex and 5-bp TAM suggests that TldR proteins may exhibit increased off-target binding compared with dCas9 proteins, which typically require at least 9 bp of target–guide complementarity and a 2-bp PAM[70–72].

Cas9 already exploits a mechanism of autoregulatory gene expression control using natural long-form tracrRNAs with truncated guides to bind, but not cleave, its own *cas9* promoter sequence in the native bacterial context[73]. Other non-canonical gRNAs similarly program Cas9 for natural gene repression functions as a means of promoting virulence[74], and some CRISPR–Cas subtypes leverage nuclease-dead Cas12 variants for adaptive immune protection, in a mechanism that relies on high-affinity RNA-guided DNA binding without cleavage[33,34]. Whereas these effectors presumably emerged via RuvC inactivation subsequent to the initial cooption of a TnpB family nuclease by CRISPR–Cas systems, our study indicates that TldRs evolved independently without ever passing through a CRISPR-associated intermediate. The recent observation of *cas12f*/*tnpB*-like genes adjacent to sigma factor genes[31] has further suggested the exciting possibility that nuclease-dead, RNA-guided DNA targeting proteins have also been exapted for gene activation. Together with our discoveries of TldR function, these examples reveal that transcriptional downregulation and upregulation via programmable CRISPR interference-like and CRISPR activation-like pathways, respectively, emerged in nature long before humans deciphered the molecular mechanisms of CRISPR–Cas.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-07598-4.

1. Altae-Tran, H. et al. The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases. *Science* **374**, 57–65 (2021).
2. Karvelis, T. et al. Transposon-associated TnpB is a programmable RNA-guided DNA endonuclease. *Nature* **599**, 692–696 (2021).
3. Meers, C. et al. Transposon-encoded nucleases use guide RNAs to promote their selfish spread. *Nature* **622**, 863–871 (2023).
4. Zedaveinyte, R. et al. Antagonistic conflict between transposon-encoded introns and guide RNAs. Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.20.567912 (2023).
5. Kapitonov, V. V., Makarova, K. S. & Koonin, E. V. ISC, a novel group of bacterial and archaeal DNA transposons that encode Cas9 homologs. *J. Bacteriol.* **198**, 797–807 (2015).
6. Chylinski, K., Makarova, K. S., Charpentier, E. & Koonin, E. V. Classification and evolution of type II CRISPR–Cas systems. *Nucleic Acids Res.* **42**, 6091–6105 (2014).
7. Larson, M. H. et al. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.* **8**, 2180–2196 (2013).
8. Nakamura, S. & Minamino, T. Flagella-driven motility of bacteria. *Biomolecules* **9**, 279 (2019).
9. Samuel, A. D. et al. Flagellar determinants of bacterial sensitivity to χ-phage. *Proc. Natl Acad. Sci. USA* **96**, 9863–9866 (1999).
10. Wilson, D. R. & Beveridge, T. J. Bacterial flagellar filaments and their component flagellins. *Can. J. Microbiol.* **39**, 451–472 (1993).
11. Aziz, R. K., Breitbart, M. & Edwards, R. A. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* **38**, 4207–4217 (2010).
12. Cosby, R. L., Chang, N. C. & Feschotte, C. Host–transposon interactions: conflict, cooperation, and cooption. *Genes Dev.* **33**, 1098–1116 (2019).
13. Jangam, D., Feschotte, C. & Betran, E. Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet.* **33**, 817–831 (2017).
14. Gould, S. J. & Vrba, E. S. Exaptation — a missing term in the science of form. *Paleobiology* **8**, 4–15 (1982).
15. Koonin, E. V. & Makarova, K. S. Mobile genetic elements and evolution of CRISPR–Cas systems: all the way there and back. *Genome Biol. Evol.* **9**, 2812–2825 (2017).
16. Koonin, E. V. & Makarova, K. S. Origins and evolution of CRISPR–Cas systems. *Phil. Trans. R. Soc. B* **374**, 20180087 (2019).
17. Nunez, J. K., Lee, A. S., Engelman, A. & Doudna, J. A. Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. *Nature* **519**, 193–198 (2015).
18. McGinn, J. & Marraffini, L. A. Molecular mechanisms of CRISPR–Cas spacer acquisition. *Nat. Rev. Microbiol.* **17**, 7–12 (2019).
19. Krupovic, M., Makarova, K. S., Forterre, P., Prangishvili, D. & Koonin, E. V. Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR–Cas immunity. *BMC Biol.* **12**, 36 (2014).
20. Hickman, A. B., Kailasan, S., Genzor, P., Haase, A. D. & Dyda, F. Casposase structure and the mechanistic link between DNA transposition and spacer acquisition by CRISPR–Cas. *eLife* **9**, e50004 (2020).
21. Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219–225 (2019).
22. Strecker, J. et al. RNA-guided DNA insertion with CRISPR-associated transposases. *Science* **365**, 48–53 (2019).
23. Faure, G. et al. CRISPR–Cas in mobile genetic elements: counter-defence and beyond. *Nat. Rev. Microbiol.* **17**, 513–525 (2019).
24. Seed, K. D., Lazinski, D. W., Calderwood, S. B. & Camilli, A. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* **494**, 489–491 (2013).
25. Al-Shayeb, B. et al. Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425–431 (2020).
26. Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* **3**, 722–732 (2005).
27. Bao, W. & Jurka, J. Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements. *Mob. DNA* **4**, 12 (2013).
28. Jiang, K. et al. Programmable RNA-guided DNA endonucleases are widespread in eukaryotes and their viruses. *Sci. Adv.* **9**, eadk0171 (2023).
29. Saito, M. et al. Fanzor is a eukaryotic programmable RNA-guided endonuclease. *Nature* **620**, 660–668 (2023).
30. Siguier, P., Gourbeyre, E., Varani, A., Ton-Hoang, B. & Chandler, M. Everyman's guide to bacterial insertion sequences. *Microbiol. Spectr.* **3**, MDNA3-0030-2014 (2015).
31. Altae-Tran, H. et al. Diversity, evolution, and classification of the RNA-guided nucleases TnpB and Cas12. *Proc. Natl Acad. Sci. USA* **120**, e2308224120 (2023).
32. Makarova, K. S. et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).
33. Huang, C. J., Adler, B. A. & Doudna, J. A. A naturally DNase-free CRISPR–Cas12c enzyme silences gene expression. *Mol. Cell* **82**, 2148–2160.e4 (2022).
34. Wu, W. Y. et al. The miniature CRISPR–Cas12m effector binds DNA to block transcription. *Mol. Cell* **82**, 4487–4502.e7 (2022).
35. Peters, J. E., Makarova, K. S., Shmakov, S. & Koonin, E. V. Recruitment of CRISPR–Cas systems by Tn7-like transposons. *Proc. Natl Acad. Sci. USA* **114**, E7358–E7366 (2017).
36. Nakamura, M., Gao, Y., Dominguez, A. A. & Qi, L. S. CRISPR technologies for precise epigenome editing. *Nat. Cell Biol.* **23**, 11–22 (2021).
37. Zetsche, B. et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR–Cas system. *Cell* **163**, 759–771 (2015).
38. He, S. et al. The IS200/IS605 family and "peel and paste" single-strand transposition mechanism. *Microbiol. Spectr.* https://doi.org/10.1128/microbiolspec.MDNA3-0039-2014 (2015).

39. Doeven, M. K., van den Bogaart, G., Krasnikov, V. & Poolman, B. Probing receptor–translocator interactions in the oligopeptide ABC transporter by fluorescence correlation spectroscopy. *Biophys. J.* **94**, 3956–3965 (2008).
40. Biemans-Oldehinkel, E., Doeven, M. K. & Poolman, B. ABC transporter architecture and regulatory roles of accessory domains. *FEBS Lett.* **580**, 1023–1035 (2006).
41. Mukherjee, S. et al. CsrA–FliW interaction governs flagellin homeostasis and a checkpoint on flagellar morphogenesis in *Bacillus subtilis*. *Mol. Microbiol.* **82**, 447–461 (2011).
42. Lawrence, J. G. Shared strategies in gene organization among prokaryotes and eukaryotes. *Cell* **110**, 407–413 (2002).
43. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32–D36 (2006).
44. Leenay, R. T. & Beisel, C. L. Deciphering, communicating, and engineering the CRISPR PAM. *J. Mol. Biol.* **429**, 177–191 (2017).
45. Michaux, C. et al. Single-nucleotide RNA maps for the two major nosocomial pathogens *Enterococcus faecalis* and *Enterococcus faecium*. *Front. Cell. Infect. Microbiol.* **10**, 600325 (2020).
46. Nety, S. P. et al. The transposon-encoded protein TnpB processes its own mRNA into ωRNA for guided nuclease activity. *CRISPR J.* **6**, 232–242 (2023).
47. Ohnishi, K., Kutsukake, K., Suzuki, H. & Iino, T. Gene fliA encodes an alternative sigma factor specific for flagellar operons in *Salmonella typhimurium*. *Mol. Gen. Genet.* **221**, 139–147 (1990).
48. Ide, N., Ikebe, T. & Kutsukake, K. Reevaluation of the promoter structure of the class 3 flagellar operons of *Escherichia coli* and *Salmonella*. *Genes Genet. Syst.* **74**, 113–116 (1999).
49. Klepsch, M. M. et al. *Escherichia coli* peptide binding protein OppA has a preference for positively charged peptides. *J. Mol. Biol.* **414**, 75–85 (2011).
50. Solovyev, V. A. S. in *Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies* (ed Li, R. W.) 61–78 (Nova Science Publishers, 2011).
51. Swarts, D. C., van der Oost, J. & Jinek, M. Structural basis for guide RNA processing and seed-dependent DNA targeting by CRISPR–Cas12a. *Mol. Cell* **66**, 221–233.e4 (2017).
52. Qi, L. S. et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).
53. Zhang, X. et al. Multiplex gene regulation by CRISPR–ddCpf1. *Cell Discov.* **3**, 17018 (2017).
54. Kim, S. K. et al. Efficient transcriptional gene repression by type V-A CRISPR–Cpf1 from *Eubacterium eligens*. *ACS Synth. Biol.* **6**, 1273–1282 (2017).
55. Samatey, F. A. et al. Structure of the bacterial flagellar protofilament and implications for a switch for supercoiling. *Nature* **410**, 331–337 (2001).
56. Yonekura, K., Maki-Yonekura, S. & Namba, K. Complete atomic model of the bacterial flagellar filament by electron cryomicroscopy. *Nature* **424**, 643–650 (2003).
57. Reid, S. D., Selander, R. K. & Whittam, T. S. Sequence diversity of flagellin (fliC) alleles in pathogenic *Escherichia coli*. *J. Bacteriol.* **181**, 153–160 (1999).
58. Esteves, N. C., Bigham, D. N. & Scharf, B. E. Phages on filaments: a genetic screen elucidates the complex interactions between *Salmonella enterica* flagellin and bacteriophage Chi. *PLoS Pathog.* **19**, e1011537 (2023).
59. Guttenplan, S. B. & Kearns, D. B. Regulation of flagellar motility during biofilm formation. *FEMS Microbiol. Rev.* **37**, 849–871 (2013).
60. Dacquay, L. C. et al. *E. coli* nissle increases transcription of flagella assembly and formate hydrogenlyase genes in response to colitis. *Gut Microbes* **13**, 1994832 (2021).
61. Kim, M. J., Lim, S. & Ryu, S. Molecular analysis of the *Salmonella typhimurium* tdc operon regulation. *J. Microbiol. Biotechnol.* **18**, 1024–1032 (2008).
62. Esteves, N. C. & Scharf, B. E. Flagellotropic bacteriophages: opportunities and challenges for antimicrobial applications. *Int. J. Mol. Sci.* **23**, 7084 (2022).
63. Yoon, S. I. et al. Structural basis of TLR5-flagellin recognition and signaling. *Science* **335**, 859–864 (2012).
64. Tenthorey, J. L. et al. The structural basis of flagellin detection by NAIP5: a strategy to limit pathogen immune evasion. *Science* **358**, 888–893 (2017).
65. Wang, L., Rothemund, D., Curd, H. & Reeves, P. R. Species-wide variation in the *Escherichia coli* flagellin (H-antigen) gene. *J. Bacteriol.* **185**, 2936–2943 (2003).
66. Cullender, T. C. et al. Innate and adaptive immunity interact to quench microbiome flagellar motility in the gut. *Cell Host Microbe* **14**, 571–581 (2013).
67. Brussow, H., Canchaya, C. & Hardt, W. D. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* **68**, 560–602 (2004).
68. Rees, D. C., Johnson, E. & Lewinson, O. ABC transporters: the power to change. *Nat. Rev. Mol. Cell Biol.* **10**, 218–227 (2009).
69. Holtzman, L. & Gersbach, C. A. Editing the epigenome: reshaping the genomic landscape. *Annu. Rev. Genomics Hum. Genet.* **19**, 43–71 (2018).
70. Bikard, D. et al. Programmable repression and activation of bacterial gene expression using an engineered CRISPR–Cas system. *Nucleic Acids Res.* **41**, 7429–7437 (2013).
71. Cui, L. et al. A CRISPRi screen in *E. coli* reveals sequence-specific toxicity of dCas9. *Nat. Commun.* **9**, 1912 (2018).
72. Vigouroux, A., Oldewurtel, E., Cui, L., Bikard, D. & van Teeffelen, S. Tuning dCas9's ability to block transcription enables robust, noiseless knockdown of bacterial genes. *Mol. Syst. Biol.* **14**, e7899 (2018).
73. Workman, R. E. et al. A natural single-guide RNA repurposes Cas9 to autoregulate CRISPR–Cas expression. *Cell* **184**, 675–688.e19 (2021).
74. Ratner, H. K. et al. Catalytically active Cas9 mediates transcriptional interference to facilitate bacterial virulence. *Mol. Cell* **75**, 498–510.e5 (2019).

Methods

Bioinformatic identification of natural TldRs

An initial search of the NCBI non-redundant protein database — queried with TnpB sequences from *Helicobacter pylori* and *Geobacillus stearothermophilus* (WP_078217163.1 and WP_047817673.1, respectively) in Jackhmmer[75] (as described in Meers et al.[3]) — resulted in the identification of 95,731 unique TnpB-like proteins, which were further clustered at 50% amino acid identity (across 50% sequence coverage) via CD-HIT[76] to produce a set of 2,646 representative TnpB sequences. A multiple sequence alignment (MSA) was then constructed with MAFFT[77] (EINSI; four rounds), which was trimmed manually with trimAl[78] (90% gap threshold; v1.4.rev15). The resulting alignment of TnpB/TldR homologues was used to construct a phylogenetic tree in IQTree[79] (WAG model, 1,000 replicates for SH-aLRT, aBayes and ultrafast bootstrap)[3], which was annotated and visualized in ITOL[80].

To assess the conservation of RuvC catalytic residues in each TnpB protein sequence, we compared each sequence in the MSA to structurally characterized orthologues (that is, *Dra*TnpB from ISDra2 and Cas12f; PDB ID 8H1J and 7L48, respectively). This comparison was performed by aligning each candidate, as well as the homologues represented in the closest five tree branches on either side of it, to *Dra*TnpB and *Un*Cas12f using the AlignSeqs function of the DECIPHER package[81] in R. TnpB-like protein sequences with less than two conserved residues of the RuvC DED catalytic motif were extracted using the Biostrings package[82] in R. For each sequence with less than two active site residues identified (defined as TldR), related homologues were retrieved from initial sequence clusters, and additional related homologues were identified via BLASTP searches of the non-redundant protein database ($e < 1 \times 10^{-50}$, query coverage > 80%, maximum target sequences = 50)[83]. Each representative sequence and all of their cluster members were used as queries in these BLASTP searches, and the active sites from BLAST hits were checked by aligning proteins to structurally determined representatives, as described above. This approach resulted in the identification of 494 unique TldR homologues. Genomes encoding each TldR were retrieved from the NCBI using the batch-entrez tool. TldR-encoding loci (that is, *tldR* ± 20 kb) were extracted using the Biostrings package[82] in R, and each *tldR* locus was annotated with Eggnog (-m diamond --evalue 0.001 --score 60 --pident 40 --query_cover 20 --subject_cover 20 --genepred prodigal --go_evidence non-electronic --pfam_realign none)[84]. Annotated *tldR* loci were manually inspected in Geneious.

To assess transposase associations, TnpA was detected using the Pfam entries for Y1_Tnp (PF01797) and serine resolvase (PF00239) via an hmmsearch from the HMMR suite (v3.3.2), with an *e* value threshold of 10^{-4}, as previously described[3]. This search was performed independently on both the curated coding sequences of each TnpB-encoding contig from NCBI and the open reading frames (ORFs) predicted by Prodigal[85] with default settings. The union of these searches was used as the final set of detected TnpA proteins. IS elements that encoded TnpB homologues within 10,000 bp of a detected TnpA are plotted in Fig. 1b.

Bioinformatic analyses of TldR homologues associated with *fliC*$_P$, *oppF* and *csrA*

To further investigate *fliC*-associated TldR homologues, we extracted cluster members for three representative branches in the tree shown in Fig. 1b (WP_193971683.1, WP_064735610.1 and WP_048785942.1). The protein file representing these combined clusters was supplemented with additional homologues identified via BLASTP searches of the non-redundant database[83]. The resulting concatenated protein file included both TldR and related TnpB sequences. To increase the diversity of TnpB proteins represented in this dataset, three additional TnpB homologues (WP_269608765.1, WP_024186316.1 and WP_059759460.1) were identified and manually added to this protein file via web-based BLASTP searches queried with the TnpB protein sequences already present in the dataset ($e < 0.05$). An MSA was constructed from these sequences and *Dra*TnpB using the AlignSeqs function of the DECIPHER package[81] in R to verify the active site composition of each orthologue. To determine which *tldR/tnpB* genes were associated with *fliC*, we analysed Eggnog annotation information for each locus (described above) and extracted TldR/TnpB sequences that were encoded within three ORFs of *fliC*.

A locus was defined as phage associated if it contained four or more gene annotations that contained the word 'Phage', 'phage', 'Viridae' or 'viridae'. TldR/TnpB protein sequences were then de-duplicated via CD-HIT[76] (-c 1.0), and an MSA was built in MAFFT[77] (LINSI) from the resulting set of 160 unique proteins. Protein domain coordinates displayed around the tree in Fig. 2c were inferred by cross-referencing the MSA and predicted structures. The phylogenetic tree shown in Fig. 2c was built from the TldR/TnpB MSA in FastTree[86] (-wag -gamma) and was annotated and visualized in ITOL[80]. Structural models of each candidate shown in Fig. 1d were predicted with AlphaFold[87] (v2.3) and displayed with ChimeraX[88] (v1.6); MSAs were visualized in Jalview[89].

To interrogate *oppF*-associated TldR sequences, we extracted cluster members and additional homologues identified via BLASTP[83] searches of the non-redundant database ($e < 1 \times 10^{-50}$, query coverage > 80%, maximum target sequences = 50) for six branches representing TldR proteins in the tree in Fig. 1b (RBR34854.1, WP_016173224.1, WP_156233666.1, NTQ19983.1, OTP13636.1 and OSH30650.1). We concatenated these sequences with cluster members and additional homologues identified through an identical BLASTP search of one representative TnpB branch (EOH94253.1) that corresponded to the closest branch to the six TldR branches in the tree. To increase the diversity of related TnpB proteins represented in this dataset, three additional TnpB homologues (WP_242450195.1, WP_028983493.1 and WP_277281207.1) were identified and manually added to this protein file via web-based BLASTP searches queried with the TnpB protein sequences already present in the dataset ($e < 0.05$). Genomes encoding TldR/TnpB proteins were downloaded from the NCBI using the Batch-entrez tool, relevant loci (*tldR/tnpB* ± 20 kb) were extracted using the Biostrings package[82] in R, and each locus was annotated with Eggnog (see above)[84]. Each TldR/TnpB protein was individually aligned to *Dra*TnpB using the AlignSeqs function of the DECIPHER package[81] in R to verify its RuvC active site composition. TldR/TnpB sequences were then deduplicated via CD-HIT[76] (-c 1.0), and an MSA was built in MAFFT[77] (LINSI) from the resulting set of 204 unique proteins. An initial phylogenetic tree was constructed in FastTree[86] (-wag -gamma), and this tree was used to guide the selection of eight representative TldRs and four representative TnpBs (shown in Supplementary Fig. 4) that were structurally predicted with ColabFold[90] (v1.5). These 12 predicted structures were used to guide an alignment of TldR/TnpB protein sequences in Promals3D[91], and the resulting MSA was used to build the tree in Extended Data Fig. 1 in FastTree (-wag -gamma). Protein domain coordinates displayed around the tree in Extended Data Fig. 1 were inferred by cross-referencing the MSA and predicted structures. The phylogenetic tree was annotated and visualized in ITOL[80].

To probe *oppF*-associated TldR loci, we extracted cluster members and additional homologues identified via BLASTP[83] searches of the non-redundant database ($e < 1 \times 10^{-50}$, query coverage > 80%, maximum target sequences = 500) for one TldR protein in the tree in Fig. 1b (WP_204886977.1). Genomes encoding TldR/TnpB proteins were downloaded from NCBI using the Batch-entrez tool, relevant loci (*tldR/tnpB* ± 20 kb) were extracted using the Biostrings package[82] in R, and each locus was annotated with Eggnog (see above)[84]. Each TldR/TnpB protein was individually aligned to *Dra*TnpB using the AlignSeqs function of the DECIPHER package[81] in R to verify its RuvC active site composition. TldR/TnpB sequences were then deduplicated via CD-HIT[76] (-c 1.0), resulting in 36 additional unique TldR proteins.

Bioinformatic identification of TldR-associated gRNA sequences

To define the boundaries of gRNA scaffolds in *fliC*$_P$–*tldR* loci, we used a general gRNA covariance model described in previous work[3]. The CMsearch function of Infernal (Inference of RNA alignments; v1.1.2)[92] was used to scan nucleotide sequences of *tldR* and 500-bp flanking windows, resulting in the identification of putative gRNA scaffold sequences. These TldR-associated gRNA scaffold boundaries were confirmed by comparing *fliC*$_P$–*tldR* loci to ωRNAs from confidently predicted annotations of catalytically active TnpB loci. Putative TldR guide sequences could then be retrieved from the 3′ boundary of putative gRNA scaffolds, enabling prediction of native *fliC*$_P$-associated TldR targets. Putative guides are listed in Supplementary Table 3.

An analogous search of *oppF*-associated *tldR* loci with a general gRNA covariance model failed to identify putative gRNA sequences. For this group of *tldR* loci, we instead built a new covariance model from ωRNA sequences associated more closely related TnpB loci. Using the comparative genomics strategy outlined in Fig. 3a, we manually identified the putative transposon RE for one TnpB-encoding IS element (WP_113785139.1 in KZ845747). We then aligned nucleotide sequences for all the related *tnpB* genes and 500 bp of sequence downstream of *tldR* with MAFFT[77] (LINSI). The resulting alignment was trimmed at the 3′ end to the position of the ωRNA scaffold–guide boundary identified for the WP_113785139.1 locus. This putative set of TnpB ωRNA sequences was realigned with LocaRNA[93] (--max-diff-at-am=25 --max-diff=60 --min-prob=0.01 --indel = -50 --indel-opening = -750 --plfold-span=100 --alifold-consensus-dp; v2.0.0), and a covariance model (ABC_gRNA_v1) was built and calibrated with Infernal. The CMsearch function of Infernal was then used to search sequences composed of *tldR/tnpB* and 500 bp of downstream sequence with the ABC_gRNA_v1 covariance model. This search resulted in gRNA identification for some, but not all, *tldR* loci. Thus, a second gRNA covariance model was built by extracting the newly identified TldR/TnpB gRNA sequences from their respective genomes, merging them with the sequences used to construct ABC_gRNA_v1, aligning the prospective gRNA dataset in LocaRNA, and building and calibrating a new covariance model with Infernal (ABC_gRNA_v2). When sequences comprising *tldR/tnpB* and 500 bp downstream were scanned with the ABC_gRNA_v2 covariance model, via CMsearch, putative gRNA sequences were identified for the remaining *tldR* loci (listed in Supplementary Table 2).

Visualization of RNA-seq data from the NCBI Sequence Read Archive and the Gene Expression Omnibus

To assess gRNA expression from a representative *fliC*$_P$–*tldR* locus, an RNA-seq dataset was downloaded from the NCBI Sequence Read Archive (SRA; accession ERR6044061). Reads were aligned to the *E. cloacae* AR_154 genome (CP029716.1) with bwa-mem2 (v2.2.1)[94] in paired-end mode with default parameters, and alignments were converted to BAM files with SAMtools[95]. Bigwig files were generated with the bamCoverage utility in deepTools[96], and unique reads mapping to the forward strand were visualized with the Integrated Genome Viewer (IGV)[97]. Expression of gRNA and *oppA* from an *oppF*–*tldR* locus was assessed by downloading an RNA-seq analysis from the NCBI Gene Expression Omnibus (GEO; accession GSE115009). Normalized coverage files (ID-005241, ID-005244, ID-005245 and ID-005246) for the forward strand were visualized in IGV[97].

Plasmid and *E. coli* strain construction

All strains and plasmids used in this study are described in Supplementary Tables 4 and 5, respectively, and a subset is available from Addgene. In brief, genes encoding candidate TldR and TnpB homologues (Supplementary Table 6), alongside their putative gRNAs, were synthesized by GenScript and subcloned into the PfoI and Bsu36i restriction sites of pCDFDuet-1, to generate pEffector, similar to our previous work[3]. Expression vectors contained constitutive J23105 and J23119 promoters driving expression of *tldR*/*tnpB* and the gRNA, respectively, and *tldR*/*tnpB* genes encoded an appended 3×FLAG-tag at the N terminus. gRNAs for *fliC*$_P$-associated TldRs were designed to target the host *fliC* 5′ untranslated site (UTR) site, whereas gRNAs of *oppF*-associated TldRs were engineered to target the genomic site natively targeted by a *Gst*TnpB3 homologue. Derivatives of these pEffector plasmids, or their associated pTarget plasmids (for plasmid interference assays), were cloned using a combination of methods, including Gibson assembly, restriction digestion–ligation, ligation of hybridized oligonucleotides, and around-the-horn PCR. Plasmids were cloned, propagated in NEB Turbo cells, purified using Miniprep kits (Qiagen), and verified by Sanger sequencing (Genewiz).

A custom *E. coli* K12 MG1655 strain that contained genomically encoded *sfGFP* and *mRFP* genes was constructed by adding three target sites adjacent to bioinformatically predicted TAM sequences upstream of the *mRFP* ORF, in between the constitutive promoter driving RFP expression and the corresponding ribosome-binding site (sSL3580; derivative of GenBank NC_000913.3)[52] (Supplementary Table 4). The original strain (with genomic *sfGFP* and *mRFP*) was a gift from L. S. Qi. The inserted target sites represent 25-bp sequences derived from the 5′ UTR of host *fliC* (*E. cloacae* complex sp. strain AR_0154; GenBank CP029716.1), an ABC transporter gene (*Enterococcus faecium* strain BP657; GenBank CP059816.1), and a *Gst*TnpB3 native target previously used[3].

ChIP–seq and motif analyses of genomic sites bound by TldR

ChIP–seq experiments and data analyses were generally performed as previously described[3,98], except for the use of sSL3580. In brief, *E. coli* MG1655 cells were transformed with pEffector and incubated for 16 h at 37 °C on LB agar plates with antibiotic (200 μg ml^{-1} spectinomycin). pEffector plasmids encoded *tldR/tnpB* under a constitutive promoter (Supplementary Table 5), together with a gRNA sequence complementary to either the *E. coli lacZ* gene (for dGstTnpB), a genomically integrated native TldR target sequence (for *fliC*$_P$-associated TldRs), or a genomically integrated synthetic sequence (for *oppF*-associated TldRs). The latter two targets were integrated at a synthetic locus between *mRFP* and *sfGFP* that disrupts the *E. coli nsfA* gene (coordinates 891,184–891,906 in NC_000913.3). Cells were scraped and resuspended in LB medium. The optical density at 600 nm (OD_{600}) was measured, and approximately 4.0×10^8 cells (equivalent to 1 ml with an OD_{600} of 0.25) were spread onto two LB agar plates containing antibiotic (200 μg ml^{-1} spectinomycin). Plates were incubated at 37 °C for 24 h. All cell material from both plates was then scraped and transferred to a 50-ml conical tube. Crosslinking was performed in LB medium using formaldehyde (37% solution; Thermo Fisher Scientific) and was quenched using glycine, followed by two washes in TBS buffer (20 mM Tris-HCl (pH 7.5) and 0.15 M NaCl). Cells were pelleted and flash-frozen using liquid nitrogen and stored at −80 °C.

ChIP of FLAG-tagged TnpB and TldR proteins was performed using Dynabeads Protein G (Thermo Fisher Scientific) slurry (hereafter, beads or magnetic beads) conjugated to anti-FLAG M2 antibodies (4 μl of 1 mg ml^{-1}) produced in mouse (Sigma-Aldrich). Samples were sonicated on a M220 Focused-Ultrasonicator (Covaris) with the following SonoLab 7.2 settings: minimum temperature of 4 °C; set point of 6 °C; maximum temperature of 8 °C; peak power of 75.0; duty factor of 10; cycles/bursts of 200; and 17.5 min of sonication time. After sonication, a non-immunoprecipitated input control sample was frozen. The remainder of the cleared sonication lysate was incubated overnight with anti-FLAG-conjugated magnetic beads. The next day, beads were washed and protein–DNA complexes were eluted. The non-immunoprecipitated input samples were thawed, and both immunoprecipitated and non-immunoprecipitated controls were incubated at 65 °C overnight to reverse-crosslink proteins and DNA. The next day, samples were treated with RNase A (Thermo Fisher Scientific) followed

by Proteinase K (Thermo Fisher Scientific) and purified using QIAquick spin columns (Qiagen).

ChIP–seq Illumina libraries were prepared for immunoprecipitated and input samples using the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB). Following adapter ligation, Illumina barcodes were added by PCR amplification (12 cycles). Approximately 450-bp DNA fragments were selected using two-sided AMPure XP bead (Beckman Coulter) size selection. DNA concentrations were determined using the DeNovix dsDNA Ultra High Sensitivity Kit and dsDNA High Sensitivity Kit. Illumina libraries were sequenced in paired-end mode on the Illumina NextSeq platform, with automated demultiplexing and adapter trimming (Illumina). More than 2,000,000 raw reads, including genomic-mapping and plasmid-mapping reads, were obtained for each ChIP–seq sample.

Following sequencing, paired-end reads were trimmed and mapped to a custom *E. coli* K12 MG1655 reference genome (derivative of GenBank NC_000913.3). Genomic *lacZ* and *lacI* regions partially identical to plasmid-encoded genes were masked in all alignments (genomic coordinates 366,386–367,588). Mapped reads were sorted and indexed, and multi-mapping reads were excluded. Alignments were normalized by counts per million (CPM) and converted to 1-bp bin bigwig files using the deepTools2 (ref. 96) command bamCoverage, with the following parameters: --normalizeUsing CPM -bs 1. CPM-normalized reads were visualized in IGV[97]. Genome-wide views were generated using plots of maximum read coverage values in 1-kb bins. Peak calling was performed using MACS3 (version 3.0.0a7)[99] using the non-immunoprecipitated control sample of *Eco*TldR as reference. For each peak, 200-bp sequences were extracted from the *E. coli* reference genome using BEDTools[100] (v2.30.0) and sequence motifs were identified using MEME-ChIP[101] (5.4.1). Primers used for Illumina library preparation are listed in Supplementary Table 7, and ChIP–seq read and meta information is listed in Supplementary Table 8.

RIP-seq of RNA bound by TldR

Cells harvested for RIP-seq were cultured as described for ChIP–seq using an *E. coli* K12 MG1655 strain expressing *sfGFP* and *mRFP* (sSL3580). Colonies from a single plate were scraped and resuspended in 1 ml of TBS buffer (20 mM Tris-HCl (pH 7.5) and 0.15 M NaCl). Next, the OD_{600} was measured for a 1:20 mixture of the cell suspension and TBS buffer, and a standardized amount of cell material equivalent to 20 ml of OD_{600} = 0.5 was aliquoted. Cells were pelleted by centrifugation at 4,000*g* and 4 °C for 5 min. The supernatant was discarded and pellets were stored at −80 °C.

Antibodies for immunoprecipitation were conjugated to magnetic beads as follows: for each sample, 60 µl Dynabeads Protein G (Thermo Fisher Scientific) were washed three times in 1 ml RIP lysis buffer (20 mM Tris-HCl (pH 7.5), 150 mM KCl, 1 mM $MgCl_2$ and 0.2% Triton X-100), resuspended in 1 ml RIP lysis buffer, combined with 20 µl of 1 mg ml^{-1} anti-FLAG M2 antibody (Sigma-Aldrich), and rotated for more than 3 h at 4 °C. Antibody–bead complexes were washed three times to remove unconjugated antibodies and resuspended in 60 µl RIP lysis buffer per sample.

Flash-frozen cell pellets were resuspended in 1.2 ml RIP lysis buffer supplemented with cOmplete Protease Inhibitor Cocktail (Roche) and SUPERase•In RNase Inhibitor (Thermo Fisher Scientific). Cells were then sonicated for 1.5 min total (2 s on, 5 s off) at 20% amplitude. Lysates were centrifuged for 15 min at 4 °C at 21,000*g* to pellet cell debris and insoluble material, and the supernatant was transferred to a new tube. At this point, a small volume of each sample (24 µl or 2%) was set aside as the 'input' starting material and stored at −80 °C.

For immunoprecipitation, each sample was combined with 60 µl antibody–bead complex and rotated overnight at 4 °C. Next, each sample was washed three times with ice-cold RIP wash buffer (20 mM Tris-HCl, 150 mM KCl and 1 mM $MgCl_2$). After the last wash, beads were resuspended in 1 ml TRIzol (Thermo Fisher Scientific) and RNA was eluted from the beads by incubating at room temperature for 5 min. A magnetic rack was used to separate beads from the supernatant, which was transferred to a new tube and combined with 200 µl chloroform. Each sample was mixed vigorously by inversion, incubated at room temperature for 3 min, and centrifuged for 15 min at 4 °C at 12,000*g*. RNA was isolated from the upper aqueous phase using the RNA Clean & Concentrator-5 kit (Zymo Research). RNA from input samples was isolated in the same manner using TRIzol and column purification. High-throughput sequencing library preparation was performed as described below for total RNA-seq of *Enterobacter* strains. Libraries were sequenced on an Illumina NextSeq 550 in paired-end mode with 75 cycles per end.

Adapter trimming, quality trimming and read length filtering of RIP-seq reads was performed as described below for total RNA-seq experiments. Trimmed and filtered reads were mapped to a reference containing both the MG1655 genome (NC_000913.3) and the plasmid sequences using bwa-mem2 v2.2.1, with default parameters. Mapped reads were sorted, indexed, and converted into coverage tracks as described below for total RNA-seq experiments.

Plasmid cleavage assays

Plasmid interference assays were generally performed as previously described[3]. *E. coli* K12 MG1655 (sSL0810) cells were transformed with pTarget plasmids (vector sequences are listed in Supplementary Table 5), and single-colony isolates were selected to prepare chemically competent cells. Next, cells were transformed with 400 ng of pEffector plasmid or empty vector. After 3 h of recovery at 37 °C, cells were pelleted by centrifugation at 4,000*g* for 5 min and resuspended in 100 µl of H_2O. Cells were then serially diluted (10×), plated as 8-µl spots onto LB agar supplemented with spectinomycin (200 µg ml^{-1}) and kanamycin (50 µg ml^{-1}), and grown for 16 h at 37 °C. Plate images were taken using a Bio-Rad Gel Doc XR+ imager.

Plasmid interference assays were quantified by determining the number of CFU following transformation. Experiments were performed as described above; however, for each experiment, 30 µl of a tenfold dilution was plated onto a full LB agar plate containing spectinomycin (200 µg ml^{-1}) and kanamycin (50 µg ml^{-1}). CFUs were counted following 16 h of growth at 37 °C and reported as CFUs per microgram of transformed pEffector plasmid.

RFP repression assays

The RFP repression assay protocol was adapted from our previous study[3,98]. An *E. coli* strain expressing a genomically integrated *sfGFP* (sSL3761), derived from a strain provided by L. S. Qi[52], was co-transformed with 200 ng of pEffector and pTarget (vector sequences listed in Supplementary Table 5). Protein components and guide RNAs (gRNA, sgRNA or crRNA) were constitutively expressed from pEffector. pTargets were cloned to encode an *mRFP* gene under the control of a constitutive promoter. For RFP repression assays shown in Fig. 4f,g and Extended Data Fig. 8c, gRNAs were designed to target the constitutive *RFP* promoter on either strand (Fig. 4f) or the top strand only (Fig. 4g and Extended Data Fig. 8c), and 5-bp TAM sequences were inserted 5′ of each target site. For RFP repression assays shown in Extended Data Fig. 8b, 25-bp sequences containing the TAM/PAM and target site in either orientation were inserted in between the *mRFP* promoter and ribosome-binding site.

Transformed cells were plated on LB agar with antibiotic selection, and at least three of the resulting colonies on each plate were used to inoculate overnight liquid cultures. For each sample, 1 µl of the overnight culture was used to inoculate 200 µl of LB medium on a 96-well optical-bottom plate. The fluorescence signals for sfGFP and mRFP were measured alongside the OD_{600} using a Synergy Neo2 microplate reader (Biotek), while shaking at 37 °C for 16 h. For all samples, the fluorescence intensities at OD_{600} = 1.0 were used to determine the fold repression for each TldR or Cas-targeting complex, and the data were

normalized to the non-repressed signal for sSL3761. Background GFP and RFP fluorescence intensities at OD_{600} = 1.0 were determined using an *E. coli* K12 MG1655 strain (sSL0810) lacking *sfGFP* and *mRFP* genes, and were subtracted from all RFP and GFP fluorescence measurements. Repression activity for dCas9 exhibited less strand orientation bias than previously described[52] (Extended Data Fig. 8b), which may be explained by the fact that gRNA targets tested in this study were located in the 5′ UTR and thus much closer to the transcription start site than to gRNA targets previously tested directly within the RFP ORF[52].

Total RNA-seq of *Enterobacter* strains

E. cloacae strains (sSL3710, sSL3711, and sSL3712) were obtained from a CDC isolate panel (Enterobacterales Carbapenemase Diversity; CRE in ARIsolateBank), and an *Enterobacter* sp. BIDMC93 (sSL3690) was provided by A. M. Earl at the Broad Institute; strain information is listed in Supplementary Table 4. Biological replicates were obtained by isolating three individual clones of each *Enterobacter* strain on LB agar plates and using these to inoculate overnight cultures in liquid LB medium. All strains were grown at 37 °C without antibiotics and with agitation when in liquid medium (240 rpm) in a BSL-2 environment. For total RNA-seq library preparation, RNA was purified from 2 ml of exponentially growing cultures of sSL3690, sSL3710, sSL3711, and sSL3712, as RT–qPCR analyses of *fliC* expression showed that the TldR-mediated repression was more robust in the exponential than in the stationary phase. RNA was extracted using TRIzol and column purification (NEB Monarch RNA cleanup kit), and samples were then individually diluted in NEBuffer 2 (NEB) and fragmented by incubating at 92 °C for 1.5 min. The fragmented RNA was simultaneously treated with RppH (NEB) and TURBO DNase (Thermo Fisher Scientific) in the presence of SUPERase•In RNase Inhibitor (Thermo Fisher Scientific), to remove DNA and 5′ pyrophosphate. For further end repair to enable downstream adapter ligation, the RNA was treated with T4 PNK (NEB) in 1× T4 DNA ligase buffer (NEB). Samples were column purified using RNA Clean & Concentrator-5 (Zymo Research), and the concentration was determined using the DeNovix RNA Assay (DeNovix). Illumina adapter ligation and cDNA synthesis were performed using the NEBNext Small RNA Library Prep kit, using 100 ng RNA per sample. High-throughput sequencing was performed on an Illumina NextSeq 550 in paired-end mode with 75 cycles per end.

RNA-seq reads were processed using cutadapt[102] (v4.2) to remove adapter sequences, trim low-quality ends from reads, and exclude reads shorter than 15 bp. Trimmed and filtered reads were aligned to reference genomes (accessions listed in Supplementary Table 4) using bwa-mem2 (v2.2.1)[94] in paired-end mode with default parameters. SAMtools[95] (v1.17) was used to filter for uniquely mapping reads using a MAPQ (mapping quality) score threshold of 1, and to sort and index the unique reads. Coverage tracks were generated using bamCoverage[96] (v3.5.1) with a bin size of 1, read extension to fragment size and normalization by CPM with exact scaling. Coverage tracks were visualized using IGV[97]. For transcript-level quantification, the number of read pairs mapping to annotated transcripts was determined using featureCounts[103] (v2.0.2). The resulting count values were converted to TPM by normalizing for transcript length and sequencing depth. For differential expression analysis between genetically engineered *Enterobacter* strains, the count matrix was first filtered to remove rows with fewer than 10 reads for at least 3 samples. The filtered matrix was then processed by DESeq2 (v1.40.2)[104] to determine the $\log_2$(fold change) for each transcript between the experimental conditions, as well as the Wald test *P* value adjusted for multiple comparisons using the Benjamini–Hochberg approach. Significantly differentially expressed genes were determined by applying thresholds of $|\log_2(\text{fold change})| > 1$ and adjusted $P < 0.05$.

Construction of *Enterobacter* sp. BIDMC93 mutants

E. cloacae strains AR_154 and AR_163 (sSL3711 and sSL3712; respectively) are both resistant to the antibiotics commonly used for colony selection following plasmid transformation, so we proceeded with recombineering in *Enterobacter* sp. BIDMC93. Genomic mutants (listed in Supplementary Table 4) were generated using Lambda Red recombineering as previously described[105]. Mutants were designed to introduce a chloramphenicol resistance cassette at each disrupted locus. The chloramphenicol resistance cassette was amplified by PCR with Q5 High Fidelity DNA Polymerase (NEB), using primers that contained at least 50-bp of homology to the disrupted locus. Amplified products were resolved on a 1% agarose gel and purified by gel extraction (Qiagen). Electrocompetent *Enterobacter* sp. BIDMC93 cells were prepared containing a temperature-sensitive plasmid encoding Lambda Red components under a temperature-sensitive promoter (pSIM6). Immediately before preparing electrocompetent cells, Lambda Red protein expression was induced by incubating cells at 42 °C for 25 min. Of each insert, 200–500 ng was used to transform cells via electroporation (2 kV, 200 Ω and 25 μF). Cells were recovered by shaking in 1 ml of LB medium at 37 °C overnight. After recovery, cells were spread on 100-mm plates with 25 μg ml^{-1} chloramphenicol and grown at 37 °C. Chloramphenicol-resistant colonies were genotyped by Sanger sequencing (Genewiz) to confirm the desired genomic mutation.

RT–qPCR to assess host *fliC* transcription in *Enterobacter* sp. BIDMC93

Of the purified total RNA, 200 ng was used as an input for the reverse transcription reaction. First, total RNA was treated with 1 μl dsDNase (Thermo Fisher Scientific) in 1X dsDNase reaction buffer in a final volume of 10 μl and incubated at 37 °C for 20 min. Then, 1 μl of 10 mM dNTP, 1 μl of 2 μM oSL14254, and 1 μl of 2 μM oSL14280 were added for gene-specific priming (*rrsA* and *fliC*, respectively) and reactions were heated at 65 °C for 5 min; oligonucleotide sequences are listed in Supplementary Table 7. Reactions were then placed directly on ice, followed by addition of 4 μl of SSIV buffer, 1 μl 100 mM dithiothreitol, 1 μl SUPERase•In (Thermo Fisher Scientific), and 1 μl of SuperScript IV Reverse Transcriptase (200 U μl^{-1}; Thermo Fisher Scientific), followed by incubation at 53 °C for 10 min and then incubation at 80 °C for 10 min. qPCR was performed in 10 μl reaction containing 5 μl SsoAdvanced Universal SYBR Green Supermix (Bio-Rad), 1 μl H_20, 2 μl of primer pair at 2.5 μM concentration and 2 μl of 100-fold diluted room temperature product. Two primer pairs were used: oSL14254–oSL14255 was used to amplify *rrsA* cDNA and oSL14279–oSL14280 was used to amplify host *fliC* cDNA. Reactions were prepared in 384-well clear/white PCR plates (Bio-Rad) and measurements were performed on a CFX384 RealTime PCR Detection System (Bio-Rad) using the following thermal cycling parameters: polymerase activation and DNA denaturation (98 °C for 2.5 min), 35 cycles of amplification (98 °C for 10 s and 62 °C for 20 s). For each sample, Cq values were normalized to that of *rrsA* (reference housekeeping gene). Then, the normalized Cq values were compared with the normalized Cq value of *fliC* in the control strain (sSL3868, knock-in of *cmR* downstream of *tldR* in strain BIDMC93) to obtain relative expression levels, such that a value of 1 is equal to that of the control and higher values indicate higher expression levels. Data were plotted in Prism (v10.1.1).

Bacterial motility assays

Motility assays were performed by the soft agar method, essentially as previously described[106], with minor variations. Overnight cultures were diluted 1:100 in LB medium supplemented with the appropriate antibiotic, then grown to OD_{600} = 0.6 at 30 °C. Of culture, 2 μl at OD_{600} = 0.6 was pipetted on the centre of semisolid agar plates (2.5% Miller's LB broth and 0.25% Bacto agar) and then incubated at 30 °C for 14 h before imaging. Images were captured in a Bio-Rad Gel Doc XR Imaging System, using epi-illumination and automatic exposure settings. Colony diameter was measured in ImageJ by taking the average of the vertical and horizontal diameter measurements for three replicates.

Flagellar filament isolation

Flagellar filaments were isolated by mechanical shearing and centrifugation, essentially as previously described[107], with minor modifications. Overnight cultures of each strain were diluted 1:100 in LB medium with 25 μg ml^{-1} chloramphenicol, then incubated at 37 °C until mid-log phase (OD_{600} = 0.4). Cells were centrifuged at 4,000*g* for 5 min, and the pellet was resuspended in deflagellation buffer (1 M Tris-Cl (pH 6.5) and 100 mM NaCl). Flagellar filaments were sheared off cell bodies by passing cells through a 27-gauge needle 15–20 times, and filaments were then separated from cell bodies by centrifuging at 10,000*g* for 15 min. The supernatant containing flagellar filaments was removed and concentrated by acetone precipitation; 5 volumes of cold acetone was added to each filament sample, mixed by vortexing, incubated at −20 °C for 1 h, and then centrifuged at 14,000*g* for 10 min at 4 °C to pellet the concentrated filaments. After decanting acetone with a pipet, the pellet was air-dried for 15 min and resuspended in 2X SDS loading dye (100 mM Tris-HCl (pH 6.8), 4% (s/v) SDS, 0.07% (w/v) bromophenol blue and 30% (v/v) glycerol) with 10 mM dithiothreitol.

Flagellar in-gel digestion for mass spectrometry

Samples were boiled at 95 °C for 10 min, and 8 μl of each sample underwent separation on a 4–12% gradient SDS–PAGE gel (Mini-PROTEAN TGX, Bio-Rad) which was stained with SimplyBlue (Thermo Fisher Scientific). The gel area encompassing the FliC, $FliC_P$, and $FliC_2$ (a second host flagellin gene copy encoded at an alternate flagellar assembly locus in *Enterobacter* sp. BIDMC93 that is not targeted by TldR and commonly absent in other *Enterobacter* strains) bands in each lane was excised and in-gel digestion was performed following a previously described protocol[108], with minor modifications. Gel slices were washed with a solution of 1:1 acetonitrile and 100 mM ammonium bicarbonate for 30 min, followed by dehydration with 100% acetonitrile for 10 min until shrinkage. Excess acetonitrile was removed and the slices were dried in a speed vacuum at room temperature for 10 min.

The gel slices were reduced with 5 mM dithiothreitol for 30 min at 56 °C, cooled to room temperature, and then alkylated with 11 mM iodoacetamide for 30 min in the dark. The slices were subsequently washed with 100 mM ammonium bicarbonate and 100% acetonitrile for 10 min each. After removal of excess acetonitrile, the slices were dried in a speed vacuum for 10 min at room temperature. Gel slices were rehydrated in a solution of 25 ng μl^{-1} trypsin in 50 mM ammonium bicarbonate for 30 min on ice and then digested overnight at 37 °C. Digested peptides were collected and extracted from the gel slices in an extraction buffer (1:2 ratio by volume of 5% formic acid:acetonitrile) at high speed, shaking in an air thermostat. The supernatants from both extractions were combined and dried in a speed vacuum. Peptides were dissolved in 1% trifluoroacetic acid, vortexed and subjected to StageTip clean-up via SDB-RPS[109], followed by drying in a speed vacuum. Finally, peptides were resuspended in 10 μl LC buffer (3% acetonitrile/0.1% formic acid). Peptide concentrations were determined using NanoDrop, and 200 ng of each sample was utilized for PASEF analysis on timsTOFPro2.

Liquid chromatography with tandem mass spectrometry

Peptides were separated within 65 min at a flow rate of 400 nl min^{-1} on a reversed-phase C18 column with an integrated CaptiveSpray Emitter (25 cm × 75 μm, 1.6 μm, IonOpticks). Mobile phases A and B were with 0.1% formic acid in water and 0.1% formic acid in acetonitrile. The fraction of B was linearly increased from 2% to 25% within 35 min, followed by an increase to 40% within 10 min, and a further increase to 95% before re-equilibration. The timsTOF Pro2 was operated in PASEF mode[110] with the following settings: mass range of 100–1,700 m/z, 1/K0 start of 0.6 Vs cm^{-2}, end of 1.4 Vs cm^{-2}, ramp time of 100 ms, lock duty cycle to 100%, capillary voltage of 1.600 V, dry gas of 3 l min^{-1}, and dry temperature of 200 °C; with PASEF (parallel accumulation serial fragmentation) settings: 10 MS/MS frames (1.17 s duty cycle), charge range of 0–5, an active exclusion for 0.4 min, target intensity of 20,000, intensity threshold of 2,500, and CID (collision-induced dissociation) collision energy of 59 eV. A polygon filter was applied to the m/z and ion mobility plane to select features most likely representing peptide precursors rather than singly charged background ions.

Liquid chromatography with tandem mass spectrometry data analysis. Acquired PASEF raw files were analysed using the MaxQuant environment v2.4.13.0 and Andromeda[111] for database searches at default settings with a few modifications. The default is used for the first search tolerance and main search tolerance (20 ppm and 4.5 ppm, respectively). MaxQuant was set up to search with the reference *Enterobacter* sp. BIDMC93 proteome database downloaded from UniProt (proteome accession UP000036586). The protein sequences for FliC, $FliC_P$ or $FliC_2$ were modified to only include their variable D2–3 regions. MaxQuant performed the search trypsin digestion with up to two missed cleavages. Peptide, site, and protein false discovery rates were all set to 1% with a minimum of one peptide needed for identification; label-free quantitation was performed with a minimum ratio count of 1. The following modifications were used for protein identification and quantification: carbamidomethylation of cysteine residues (+57.021 Da) was set as static modifications, whereas the oxidation of methionine residues (+15.995 Da) and deamidation (+0.984) on asparagine were set as a variable modification. For results obtained from MaxQuant, protein groups tables were further used for data analysis.

Reporting summary

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

Data availability

Next-generation sequencing data generated in this study were deposited in the NCBI SRA (BioProject accession PRJNA1029663) and GEO (GSE245749). The published genome used for ChIP–seq analyses was obtained from NCBI (GenBank NC_000913.3). Publicly available RNA-seq data analysed for TldR–gRNA expression are in the NCBI SRA (ERR6044061) and GEO (GSE115009) databases. The published genomes used for bioinformatics analyses were obtained from NCBI (Supplementary Table 4). The ISfinder database can be accessed at https://www-is.biotoul.fr/index.php.

Code availability

Custom scripts used for bioinformatics, TAM library analyses, and ChIP–seq data analyses are available on request. The R script describing initial steps to discover TldRs is available at https://github.com/sternberglab/Wiegand_etal_2024.

75. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
76. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).
77. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).
78. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
79. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
80. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
81. Wright, E. S. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics* **16**, 322 (2015).
82. Biostrings: String objects representing biological sequences, and matching algorithms. R package version 2.70.1 (Pagès, H. A. P., Gentleman, R. & DebRoy, S., 2023).
83. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

84. Cantalapiedra, C. P., Hernandez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
85. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
86. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 — approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
87. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
88. Goddard, T. D. et al. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).
89. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview version 2 — a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
90. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
91. Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295–2300 (2008).
92. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
93. Will, S., Joshi, T., Hofacker, I. L., Stadler, P. F. & Backofen, R. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA* **18**, 900–914 (2012).
94. Vasimuddin M., Misra, S., Li, H. & Aluru, S. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* https://doi.org/10.1109/IPDPS.2019.00041 (IEEE, 2019).
95. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
96. Ramirez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
97. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
98. Hoffmann, F. T. et al. Selective TnsC recruitment enhances the fidelity of RNA-guided transposition. *Nature* **609**, 384–393 (2022).
99. Zhang, Y. et al. Model-based analysis of ChIP–seq (MACS). *Genome Biol.* **9**, R137 (2008).
100. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
101. Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
102. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* https://doi.org/10.14806/ej.17.1.200 (2011).
103. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
104. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
105. Sharan, S. K., Thomason, L. C., Kuznetsov, S. G. & Court, D. L. Recombineering: a homologous recombination-based method of genetic engineering. *Nat. Protoc.* **4**, 206–223 (2009).
106. Luo, G. et al. flrA, flrB and flrC regulate adhesion by controlling the expression of critical virulence genes in *Vibrio alginolyticus*. *Emerg. Microbes Infect.* **5**, e85 (2016).
107. Kreutzberger, M. A. B. et al. Flagellin outer domain dimerization modulates motility in pathogenic and soil bacteria from viscous environments. *Nat. Commun.* **13**, 1422 (2022).
108. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1**, 2856–2860 (2006).
109. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).
110. Meier, F. et al. Online parallel accumulation-serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. *Mol. Cell. Proteomics* **17**, 2534–2545 (2018).
111. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).

Acknowledgements We thank S. R. Pesari and Z. Akhtar for laboratory support; G. D. Lampe for suggesting the TldR moniker; A. Bernheim for helpful discussions; F. Tesson, A. Bernheim, A. M. Earl and D. Gray for sharing *E. coli* and *Enterobacter* strains; C. Lu for Covaris sonicator access; R. K. Soni for mass spectrometry support; L. F. Landweber for qPCR instrument access; and the JP Sulzberger Columbia Genome Center for next-generation sequencing support. S.T. was supported by a Medical Scientist Training Program grant (5T32GM145440-02) from the NIH. M.W.G.W. was supported by a National Science Foundation Graduate Research Fellowship. C.M. was supported by the NIH Postdoctoral Fellowship F32 GM143924-01A1. S.H.S. was supported by the NSF Faculty Early Career Development Program (CAREER) Award 2239685, a Pew Biomedical Scholarship, an Irma T. Hirschl Career Scientist Award, and a startup package from the Columbia University Irving Medical Center Dean's Office and the Vagelos Precision Medicine Fund.

Author contributions T.W., C.M., and S.H.S. conceived and designed the project. T.W. performed all of the bioinformatics experiments and aided in the design of the experimental assays. F.T.H. performed plasmid interference, ChIP–seq, and the RFP repression assays. M.W.G.W. designed and generated the *E. coli* strains and plasmids for the RFP repression assays, fragments for *Enterobacter* recombineering, conducted the motility assays and isolated flagella for liquid chromatography with tandem mass spectrometry. S.T. performed and analysed the RNA-seq and RIP-seq experiments. E.R. cultured *Enterobacter* strains, extracted RNA for RNA-seq, and performed the RT–qPCR and recombineering experiments. C.M. performed the preliminary TnpB bioinformatics and neighbourhood analyses, together with H.C.L., and helped design the ChIP–seq and RFP repression assays. T.W. and S.H.S. discussed the data and wrote the manuscript, with input from all authors.

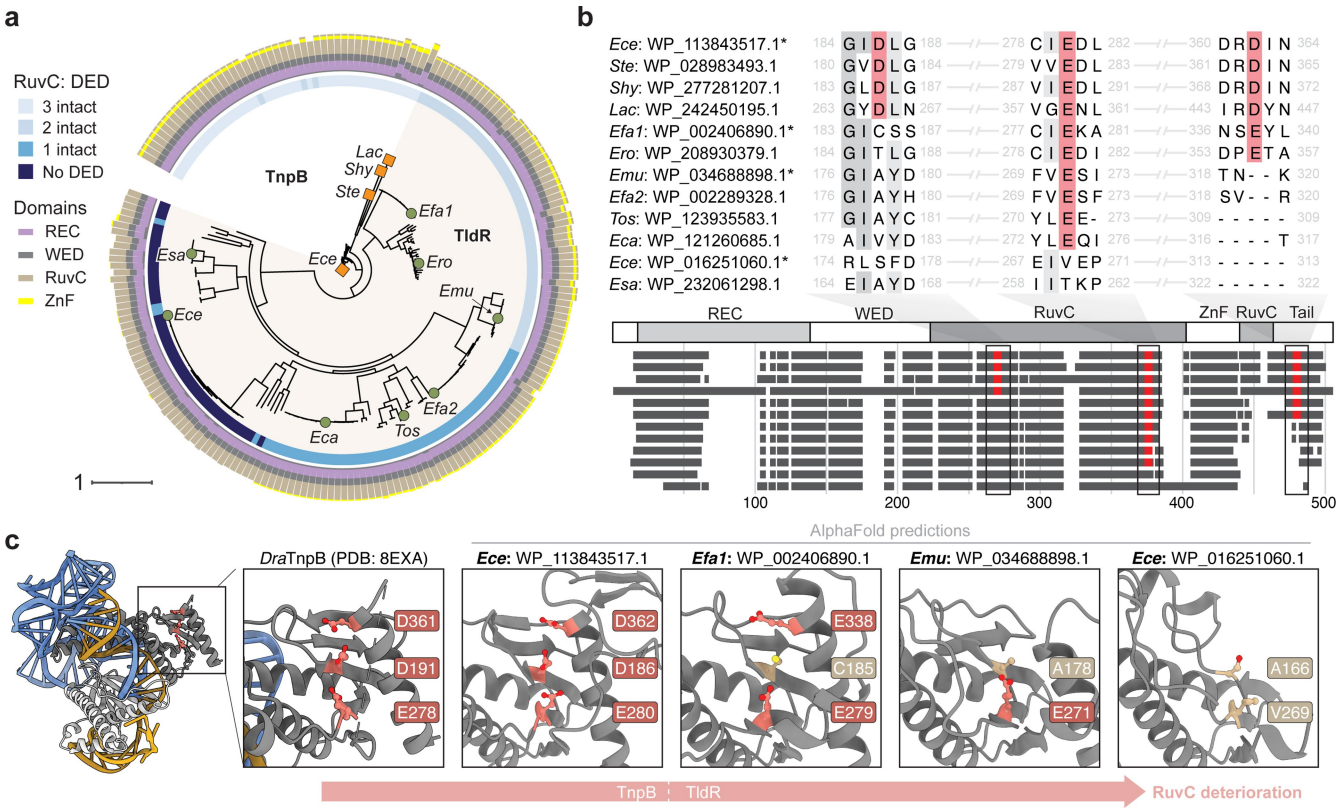
Competing interests Columbia University has filed a patent application related to this work. M.W.G.W. is a co-founder of Can9 Bioengineering. S.H.S. is a co-founder and scientific advisor to Dahlia Biosciences, a scientific advisor to CrisprBits and Prime Medicine, and an equity holder in Dahlia Biosciences and CrisprBits. All other authors declare no competing interests.

Additional information

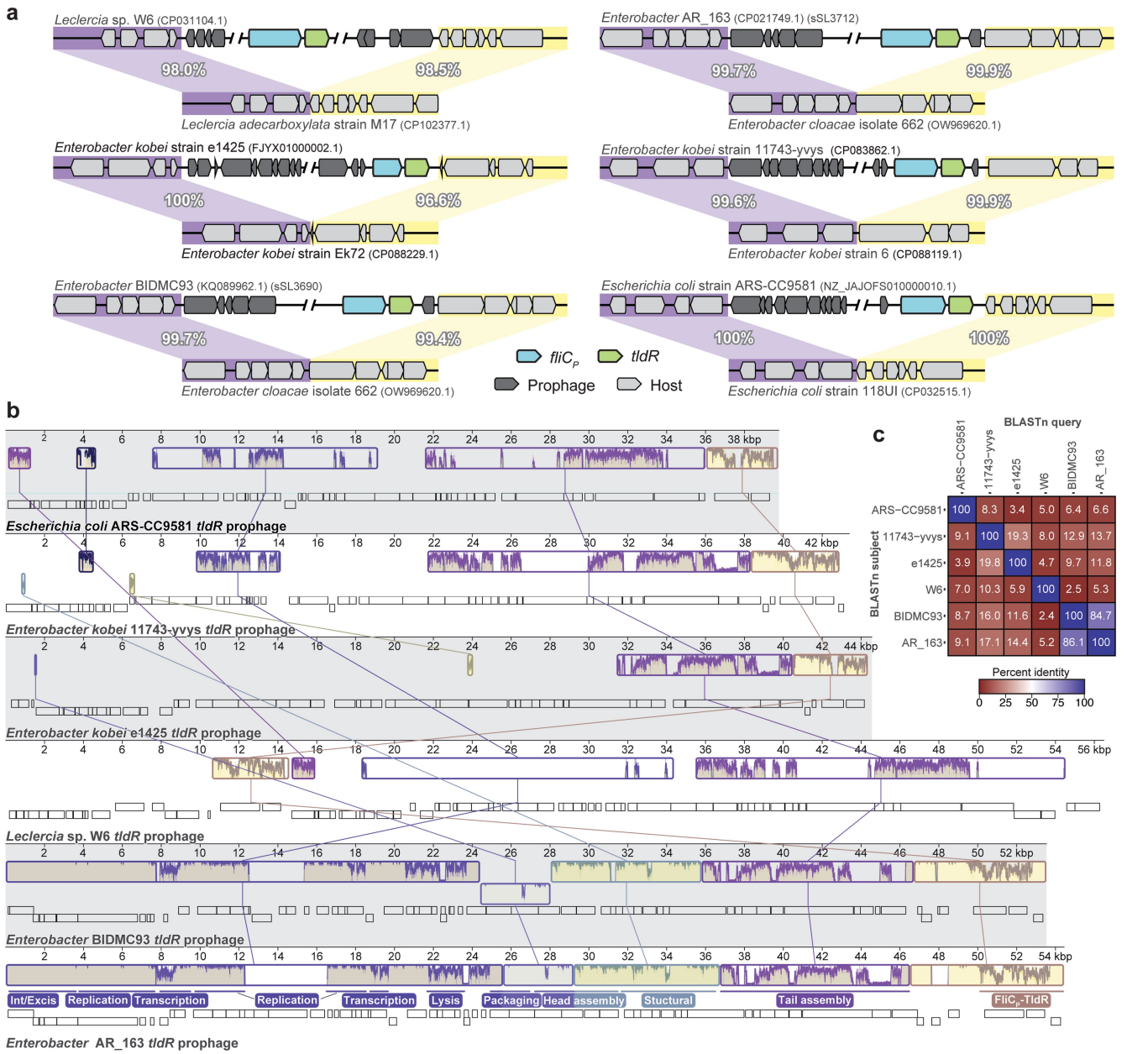
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-07598-4.

Correspondence and requests for materials should be addressed to Samuel H. Sternberg.

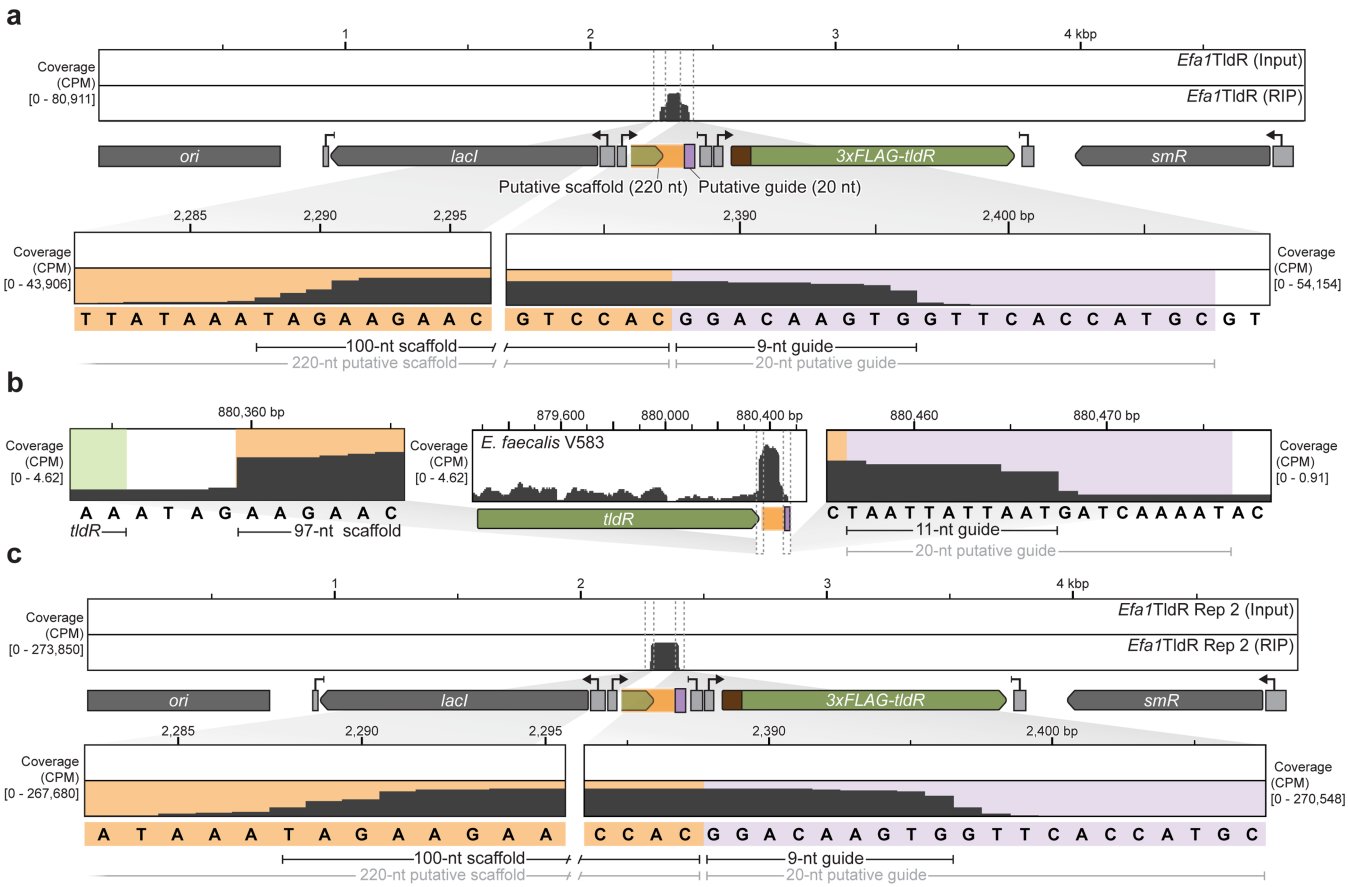
Peer review information *Nature* thanks Wen Wu and the other, anonymous reviewer(s) for their contribution to the peer review of this work.



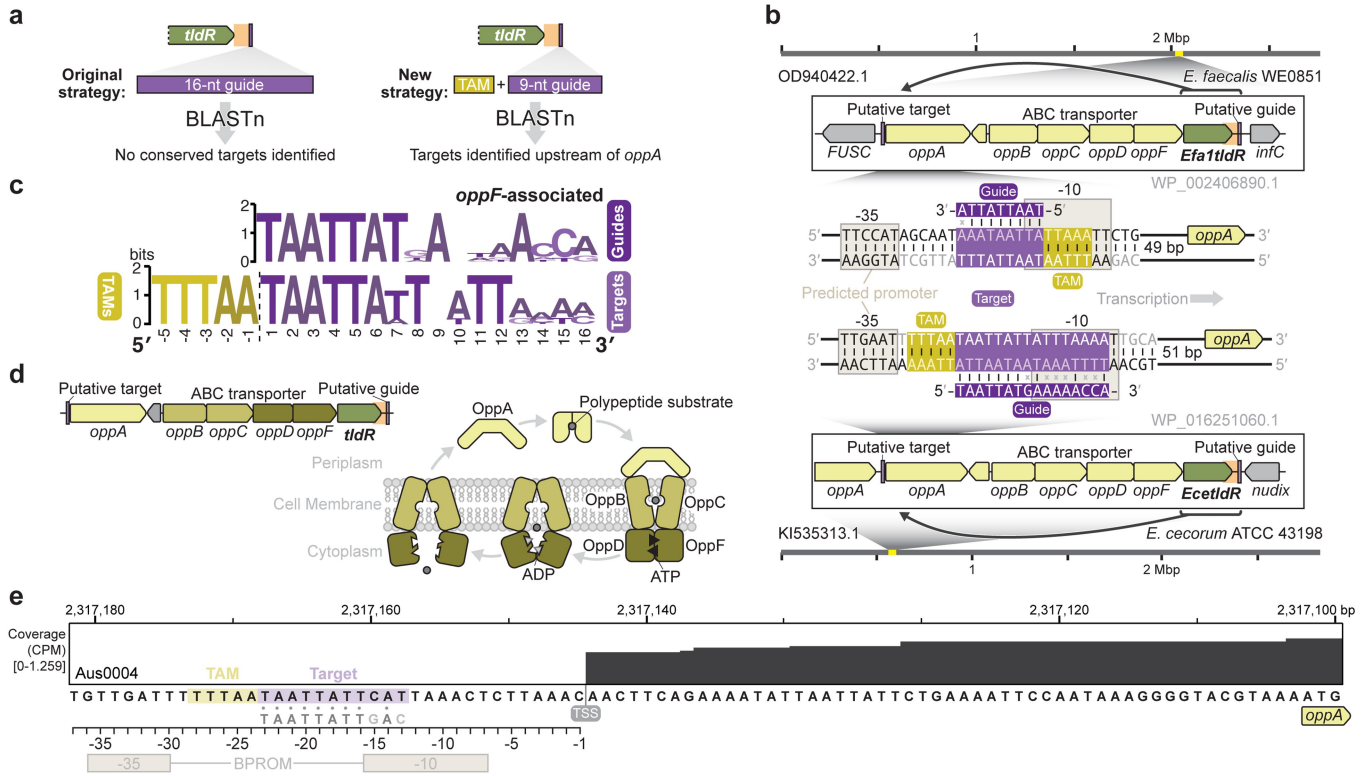
Extended Data Fig. 1 | Phylogeny and RuvC nuclease domain analysis of *oppF*-associated TldRs. a, Phylogenetic tree of *oppF*-associated TldR proteins from Fig. 2a, together with closely related TnpB proteins that contain intact RuvC active sites. The rings indicate RuvC DED active site intactness (inner) and TldR/TnpB domain composition (outer). Homologs marked with an orange square (TnpB) or green circle (TldR) were tested in heterologous experiments. **b**, Multiple sequence alignment of representative TnpB and TldR sequences from **a**, highlighting deterioration of RuvC active site motifs (shaded in red) and loss of the C-terminal zinc-finger (ZnF)/RuvC domain. Highly conserved residues are shaded in grey. **c**, Empirical (*Dra*TnpB) and predicted AlphaFold structures of TnpB and TldR homologs marked with an asterisk in **b**, showing progressive loss of the active site catalytic triad.



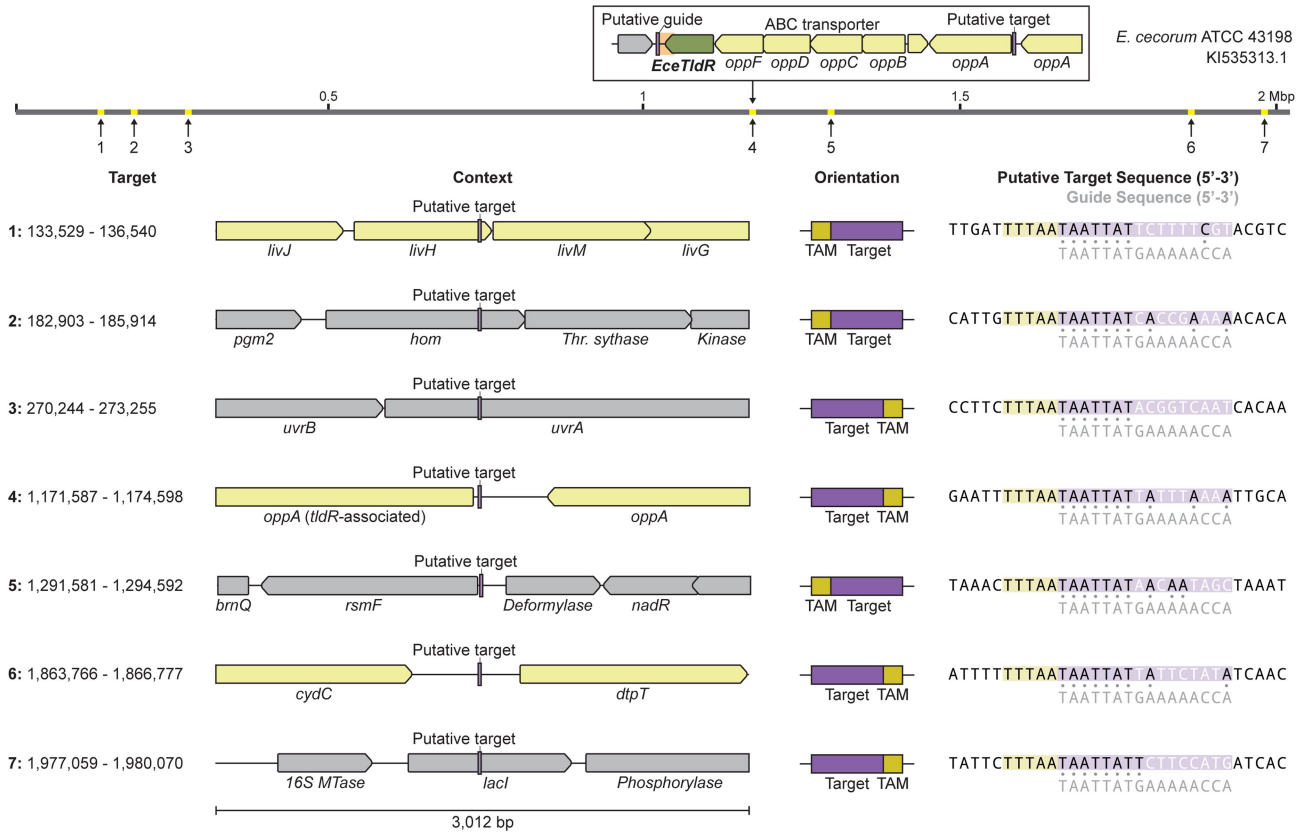
Extended Data Fig. 2 | Diverse prophages encode *fliC*$_P$-associated *tldR* genes. a, Genomic architecture of representative prophage elements whose boundaries could be identified by comparison to closely related isogenic strains. In each example, the prophage-containing strain is shown above the prophage-lacking strain, with species/strain names and NCBI genomic accession IDs indicated. Sequences flanking the left (5′) and right (3′) ends are highlighted in purple and yellow, respectively, together with their percentage sequence identities calculated using BLASTn. **b**, Alignment of distinct prophage elements, constructed using Mauve. Empty boxes represent open reading frames, and windows show sequence conservation for regions compared between prophage genomes with lines. Putative gene functions are shown below sequence conservation windows for the *fliC*$_P$*-tldR*-encoding prophage from *Enterobacter* AR_163 (bottom). **c**, DNA sequence identities between the prophages in **a**, calculated with BLASTn. Identities were calculated as total matching nucleotides across the two genomes being compared, divided by the length of the query prophage genome.



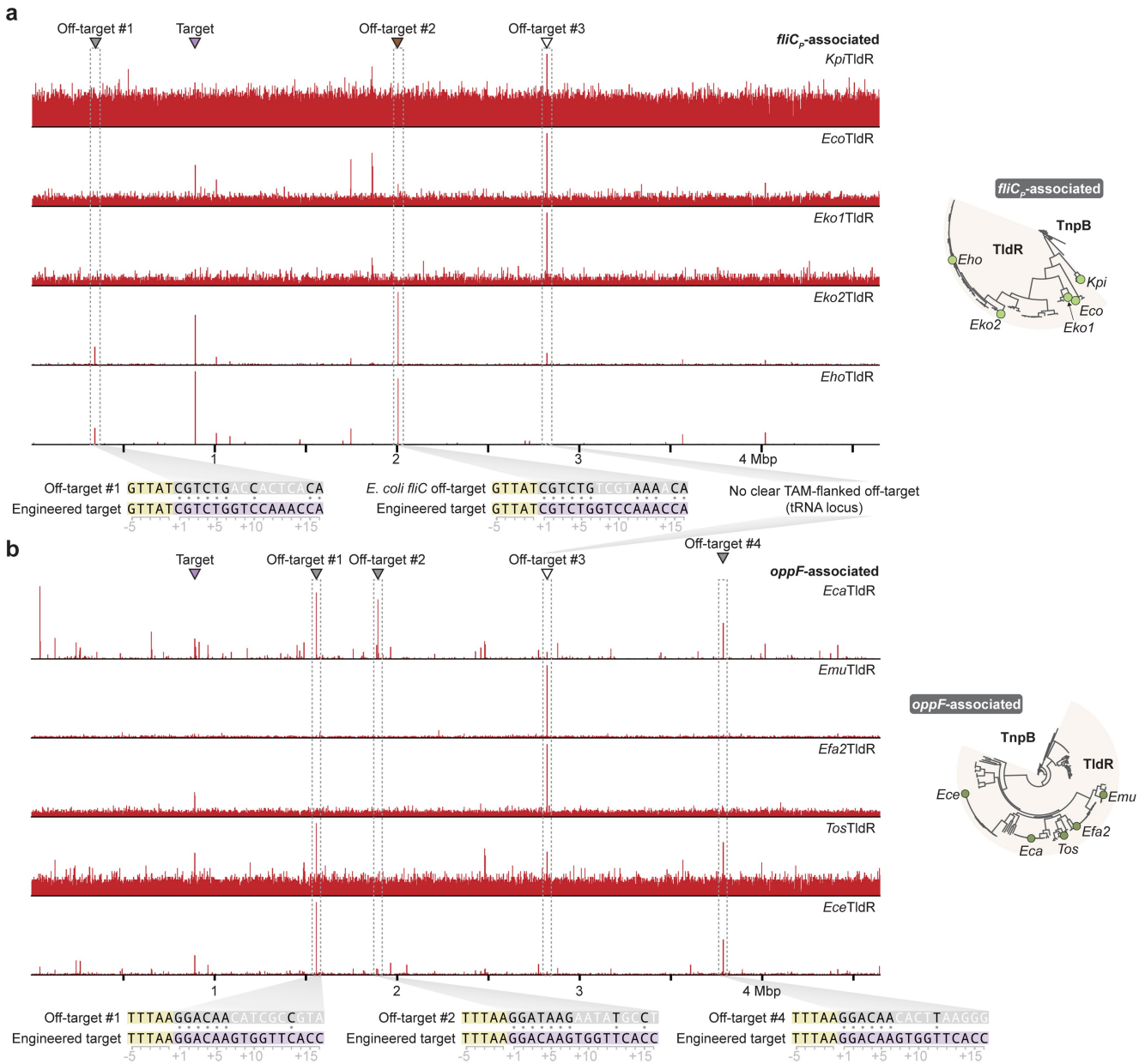
Extended Data Fig. 3 | RIP-seq reveals that some *oppF*-associated TldR proteins use short, 9–11-nt guides. a, RNA immunoprecipitation sequencing (RIP-seq) data for an *oppF*-associated TldR homolog from *Enterococcus faecalis* (*Efa1*TldR) reveals the boundaries of a mature gRNA containing a 9-nt guide sequence. Reads were mapped to the TldR-gRNA expression plasmid; an input control is shown. **b**, Published RNA-seq data for *Enterococcus faecalis* V583 reveals similar gRNA boundaries, including an approximately 11-nt guide. **c**, RIP-seq data as in **a** for a second biological replicate of *Efa1*TldR, further corroborating the observed 9–11-nt guide length.



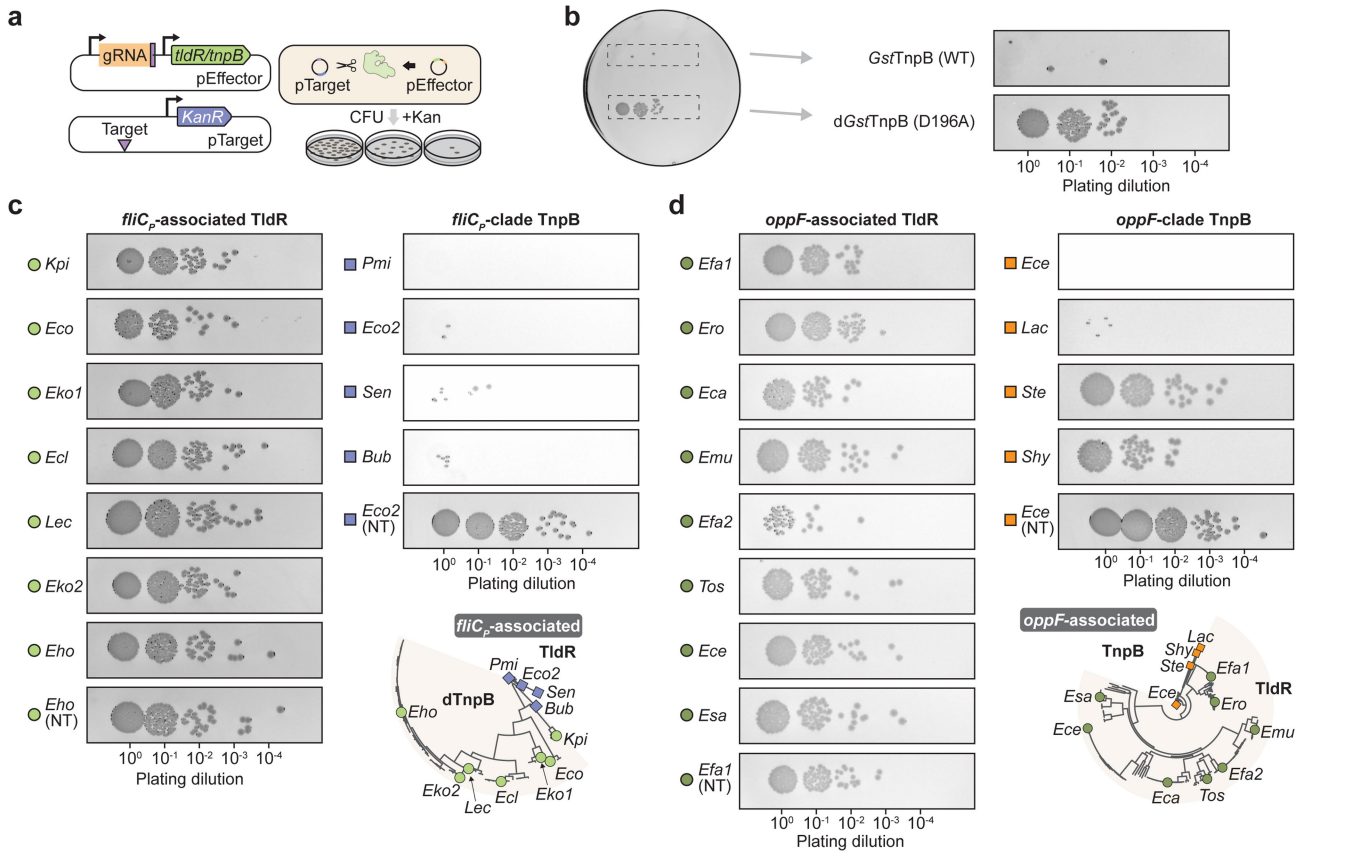
Extended Data Fig. 4 | *oppF*-associated TldRs target conserved genomic sequences that overlap with promoter elements driving *oppA* expression. **a**, Schematic of original (left) and improved (right) search strategy to identify putative targets of gRNAs used by *oppF*-associated TldRs. Key insights resulted from the use of TAM and a shorter, 9-nt guide. **b**, Analysis of the guide sequence from the *Efa1*TldR-associated gRNA in Extended Data Fig. 3 revealed a putative genomic target near the predicted promoter of *oppA* encoded within the same ABC transporter operon immediately adjacent to the *tldR* gene. The magnified schematics at the bottom show the predicted TAM and gRNA-target DNA base-pairing interactions for two representatives (*Efa1*TldR and *Ece*TldR), in which the gRNAs target opposite strands. Promoter elements predicted with BPROM are shown as brown squares. **c**, WebLogos of predicted guides and genomic targets associated with diverse *oppF*-associated TldRs highlighted in Extended Data Fig. 1. **d**, Schematic of the *oppF-tldR* genomic locus (left) alongside the predicted function of OppA as a solute binding protein that facilitates transport of polypeptide substrates from the periplasm to the cytoplasm, in complex with the remainder of the ABC transporter apparatus. **e**, Published RNA-seq data for *Enterococcus faecium* AUS0004[45], highlighting the *oppA* transcription start site (TSS). The predicted gRNA guide sequence (grey) is shown beneath the putative TAM (yellow) and target (purple) sequences, with guide-target complementarity represented by grey circles.



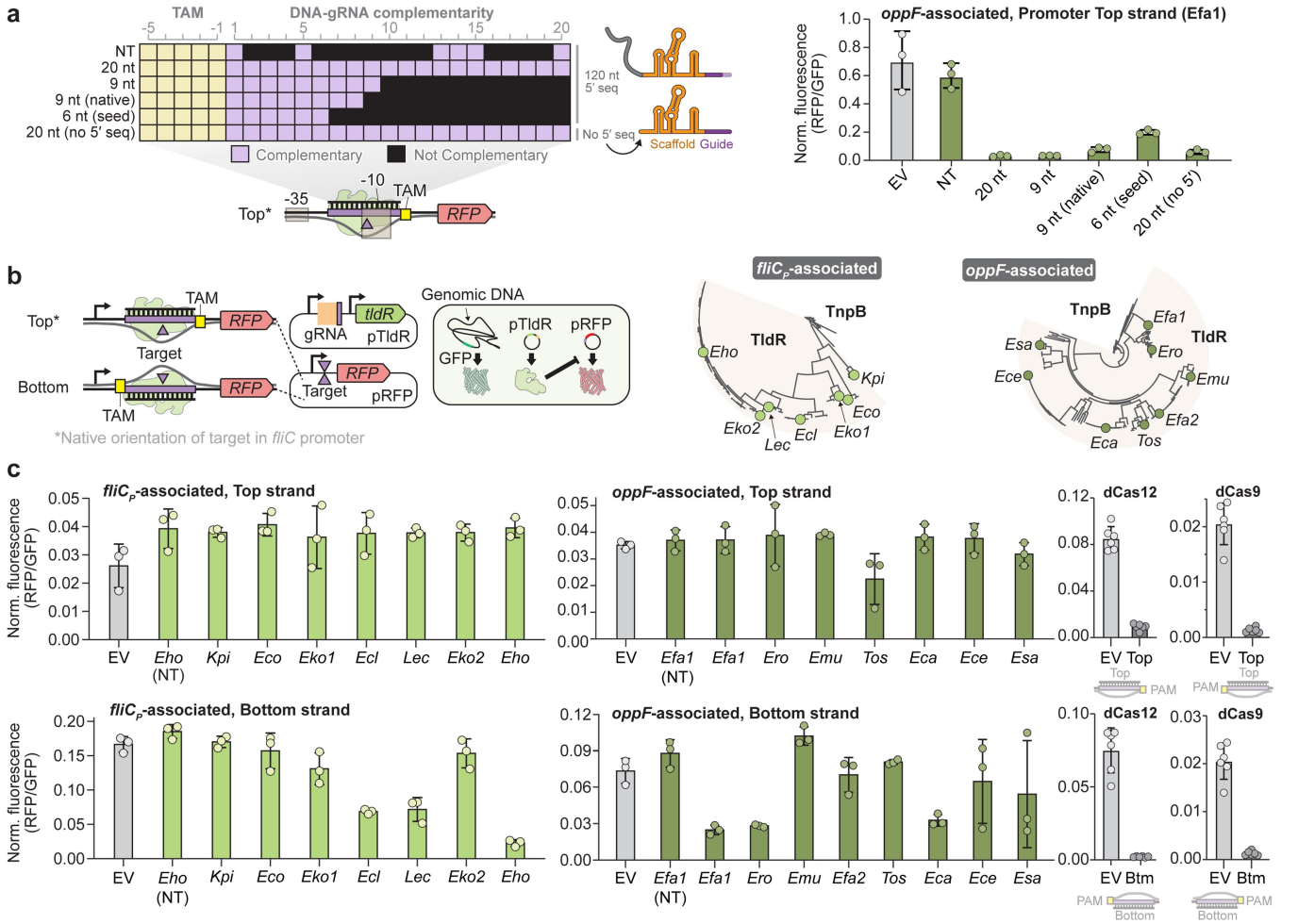
Extended Data Fig. 5 | *oppF*-associated TldR homologs may target additional sites across the genome. Schematic of *Enterococcus cecorum* genome and inset showing the *oppF-tldR* locus (top), with additional putative targets of the gRNA, other than the *oppA* promoter, numbered and highlighted in yellow along the genomic coordinate. A magnified view for each numbered target is shown below, with TAMs in yellow, prospective targets in purple, and TldR gRNA guide sequences in grey. Grey circles (right) represent positions of expected guide-target complementarity.



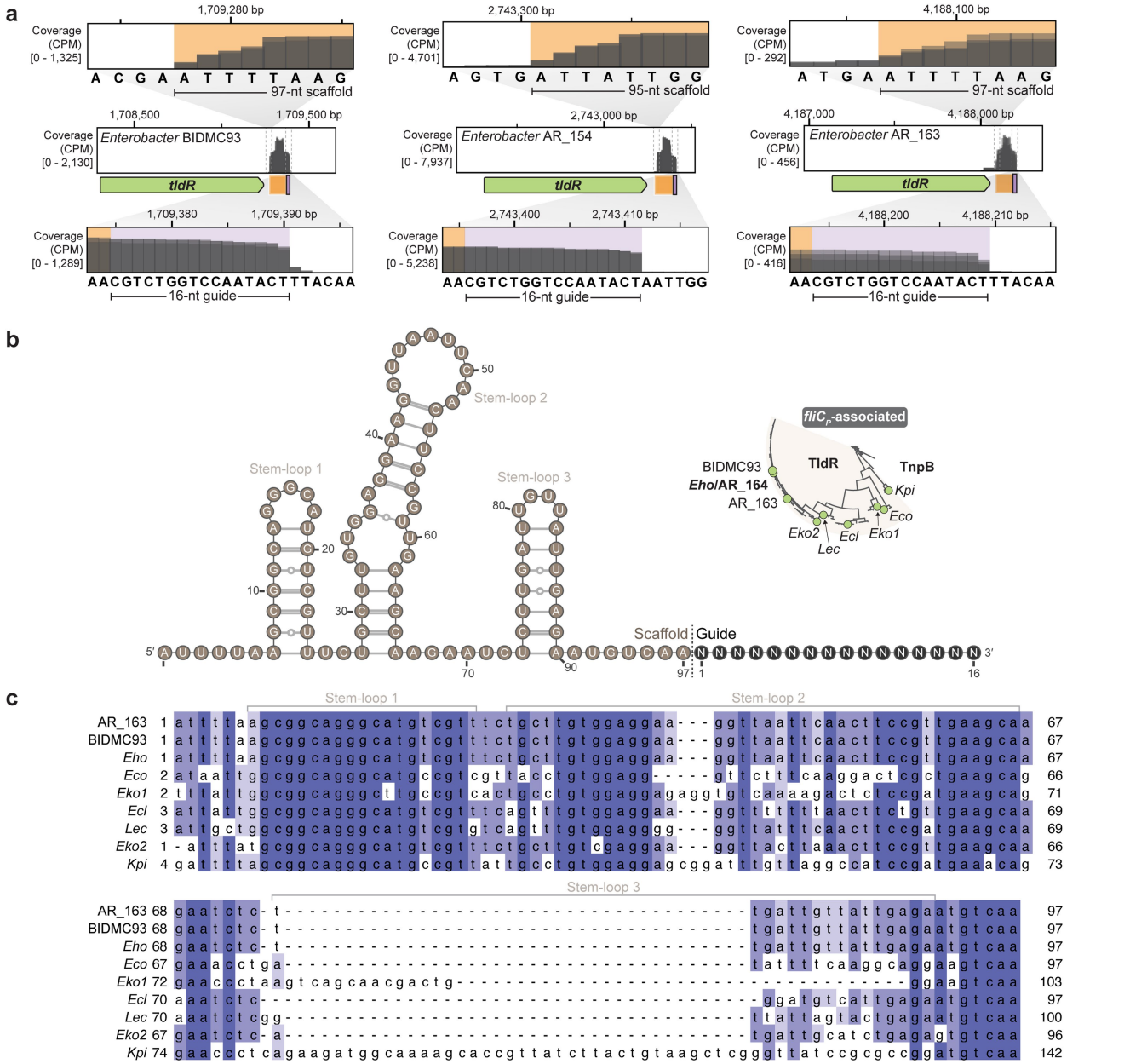
Extended Data Fig. 6 | Genome-wide binding data from ChIP-seq experiments suggest a high mismatch tolerance for some TldR homologs. **a**, Genome-wide ChIP–seq profiles for the indicated *fliC*$_P$-associated TldR homologs, normalized to the highest peak within each dataset. The magnified insets at the bottom show the off-target sequences (grey) compared to the intended (engineered) on-target sequence (purple), with TAMs in yellow. Off-target #3 has no clear TAM-flanked off-target sequence but is intriguingly located at a tRNA locus, and binding was observed for diverse *fliC*$_P$- and *oppF*-associated TldRs that recognized distinct TAMs. The phylogenetic tree at right indicates the relatedness of the tested and labeled homologs. **b**, Results for the indicated *oppF*-associated TldR homologs, shown as in **a**.



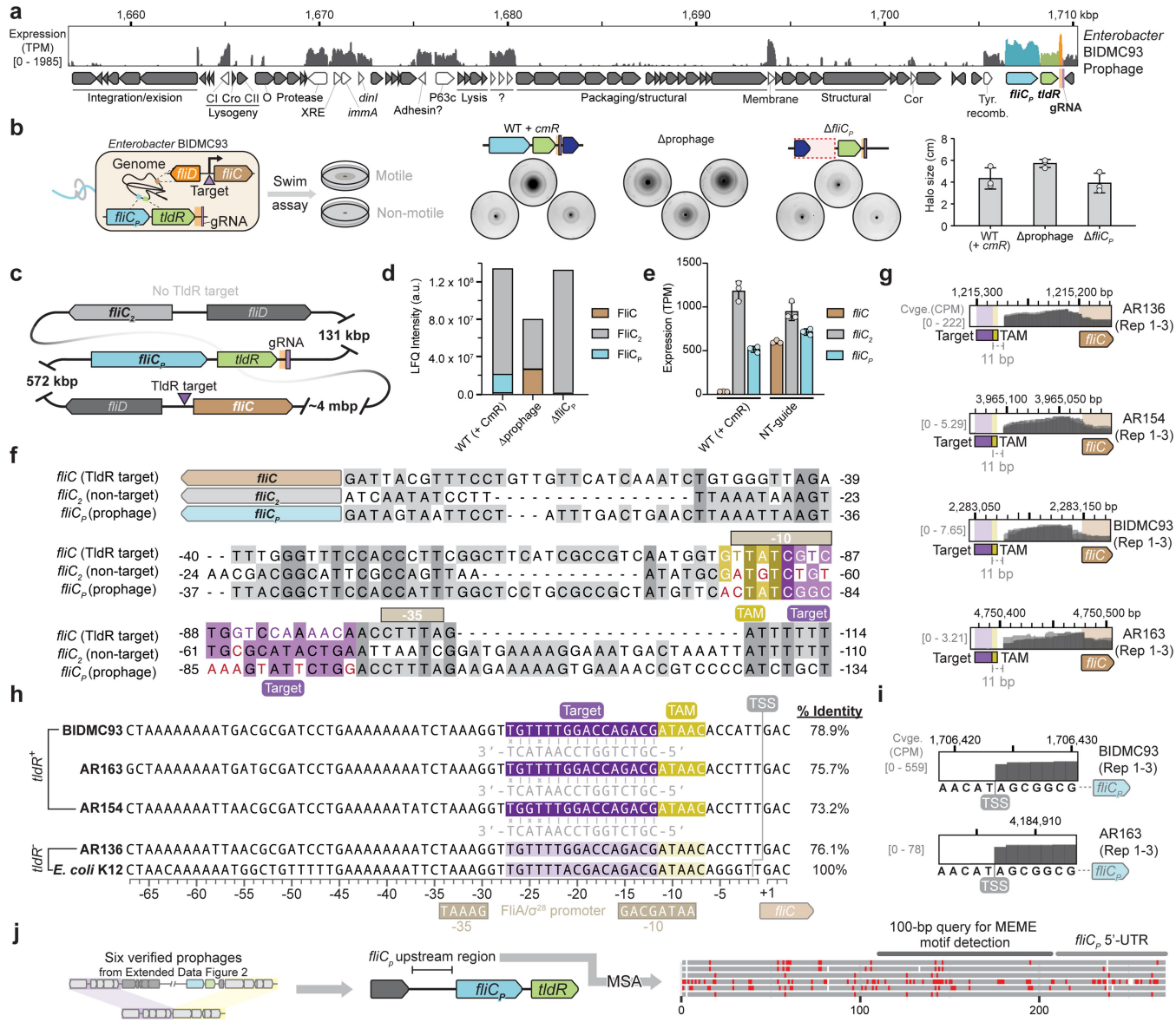
Extended Data Fig. 7 | Plasmid interference assays confirm that TldR homologs lack detectable nuclease activity. a, Schematic of *E. coli*-based plasmid interference assay using pEffector and pTarget. **b**, Representative dilution spot assays for *Gst*TnpB3 and synthetically inactivated RuvC mutant (D196A), showing the entire plate (left) and the magnified area of plating. Transformants were serially diluted, plated on selective media, and cultured at 37 °C for 16 h. Colony visibility was enhanced by inverting the colors and increasing contrast/brightness. **c**, Dilution spot assays for the indicated *fliC*-associated TldR homologs (left) and closely related TnpB homologs (right). Non-targeting (NT) gRNA controls are shown at the bottom, and the phylogenetic tree indicates the relatedness of the tested proteins. **d**, Results for the indicated *oppF*-associated TldR and TnpB homologs, shown as in **c**.



Extended Data Fig. 8 | RFP repression assays reveal variable abilities of TldR homologs to block transcription elongation. a, RFP repression activity was measured (right) as in Fig. 4f,g using modified gRNAs exhibiting variable complementarity to the target site, as schematized in the grid (left). A gRNA was also tested that lacked the extra 5′ sequence which was absent in RIP-seq reads of mature gRNAs (20 nt no 5′ seq). Bars indicate mean ± s.d. (n = 3 biological replicates). **b**, Schematic of RFP repression assay in which gRNAs were designed to target either the top or bottom strand within the 5′ UTR of *RFP*, downstream of the promoter. The phylogenetic trees (right) indicate the relatedness of the tested and labeled homologs. **c**, Bar graphs plotting normalized RFP fluorescence for the indicated conditions and TldR homologs. EV, empty vector; NT, non-targeting guide. Results with nuclease-dead dCas12 and dCas9 are shown for comparison. Bars indicate mean ± s.d. (n = 3 biological replicates for TldR; n = 6 biological replicates for dCas12/dCas9).



Extended Data Fig. 9 | *Enterobacter* RNA-seq data confirm the native expression of gRNAs from *fliC*$_P$-*tldR* loci. a, RNA-seq read coverage from three *Enterobacter* strains that natively encode *fliC*$_P$-*tldR* loci, revealing clear peaks associated with mature gRNAs containing ~95–97-nt scaffolds and 16-nt guides. Data from three biological replicates are overlaid. **b**, Predicted secondary structure and sequence of the gRNA associated with *Eho*TldR. **c**, Multiple sequence alignment of the DNA encoding gRNA scaffold sequences for representative *fliC*$_P$-associated TldRs, with conserved positions colored in darker blue.



Extended Data Fig. 10 | FliC$_P$ is expressed and incorporated into *Enterobacter* flagella, concomitantly with host FliC repression. a, RNA-seq read coverage across the *tldR*-encoding prophage of *Enterobacter* sp. BIDMC93, demonstrating strong expression of *fliC$_P$*, *tldR*, and the gRNA, alongside other genes involved in lysogeny maintenance (e.g. CI). **b**, Motility assays (left) with wild-type (WT) and *Enterobacter* deletion strains reveal similar motility phenotypes, as visualized with LB-agar plate images (middle) and a bar graph quantifying motility via halo size (right). Plate images and bar graphs represent three biological replicates; bars indicate mean ± s.d. **c**, Schematic representation of FliC/FliC$_P$ homologs encoded by *Enterobacter* sp. BIDMC93, with relative genomic positions indicated. FliC$_2$ is a second host flagellin gene copy encoded at an alternate flagellar assembly locus within this strain, which is not targeted by TldR and not commonly present in other *Enterobacter* strains. **d**, Results from liquid chromatography with tandem mass spectrometry (LC–MS/MS) analyses performed on digested peptides from purified flagellar filaments, isolated from the three indicated *Enterobacter* sp. BIDMC93 strains. The WT (+CmR) strain encodes the *cmR* gene downstream of the *tldR-gRNA* locus (as in Fig. 5e). Data represent the label free quantification (LFQ) intensities reflecting the variable D2-3 regions of FliC, FliC$_P$, or FliC$_2$. Although the FliC$_2$ appears to be the most dominant flagellin component, the relevant amounts of host FliC and FliC$_P$ demonstrate that prophage-encoded FliC$_P$ readily assembles into extracellular flagellar filaments, and that host FliC production is de-repressed upon prophage deletion. **e**, Quantification of changes in the expression profiles of *Enterobacter* FliC homologs, measured from RNA-seq data of three biological replicates depicted in Fig. 5f,g. TPM, transcripts per million. **f**, Alignment of *fliC*/*fliC$_P$*/*fliC$_2$* promoters indicates that guide RNA-target DNA mismatches prevent TldR-targeting of *fliC$_2$* and *fliC$_P$* in *Enterobacter* sp. BIDMC93. **g**, RNA-seq read coverage in the host *fliC* promoter/5′-UTR region overlayed for three biological replicates of four *Enterobacter* strains, with labeled TAM and target sequences highlighted upstream of the TSS. Strain AR136 (top) does not encode a *fliC$_P$-tldR* locus; note the distinct expression levels, measured via relative counts per million (CPM). **h**, Alignment of host *fliC* promoter regions for the strains shown in **g** compared to *E. coli* K12, with percent sequence identities indicated on the right. Reported FliA/σ^{28} promoter elements from *E. coli* K12 are shown below the alignment. **i**, RNA-seq read coverage in the prophage-encoded *fliC$_P$* promoter/5′-UTR region overlayed for three biological replicates of two representative *Enterobacter* strains, confirming the predicted TSS. **j**, Schematic of multiple sequence alignment of the promoter region driving *fliC$_P$* gene expression, across six verified prophages described in Extended Data Fig. 2, highlighting the region that was queried for MEME motif detection.

nature portfolio

Corresponding author(s): Samuel H. Sternberg

Last updated by author(s): Apr 17, 2024

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	Next-generation sequencing data utilized the Illumina platform (BaseSpace), including automated de-multiplexing and adapter trimming. RT-qPCR measurements were taken on a CFX384 RealTime PCR Detection System (BioRad).
Data analysis	Next-generation sequencing data were analyzed and visualized using custom scripts and IGV (version 2.8.13). Structural figures were generated with ChimeraX (v1.6.1). Protein structures were predicted with AlphaFold (v2.3) and ColabFold (v1.5). RNA sequencing data was processed with cutadapt (v4.2), aligned with bwa-mem2 (v2.2.1), filtered and converted to BAM with SAMtools (v1.17), and converted to bigwig files with deeptools (v3.5.1). Transcript counts were calculated with featureCounts (v2.0.2) and log2 f(old changes) were calculated with DESeq2 (v1.40.2). For gRNA searches, covariance models were built with LocaRNA (v2.0.0) and searches were executed with Infernal (v1.1.2). For analysis of TldR loci, MAFFT (v7.511) and Promals3d (online) used for alignments, FastTree (v2.1.11) was used to build phylogenies, and iTOL (online) was used to visualize trees. Sequence analyses in R utilized the Biostrings (v2.70.1), DECIPHER (v2.3.0), tidyverse (v.2.0.0), ggplot2 (v3.4.4), and biomartr (v1.0.6) packages. TldR loci maps were rendered in Snapgene Viewer (v7.2) or Geneious (v2024.0.3), and RNAseq/RT-qPCR data was plotted in Prism (v10.1.1). Custom code used to analyzed sequences is available at: https://github.com/sternberglab/Wiegand_etal_2024.

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

Data

Policy information about availability of data

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

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

Next-generation sequencing data generated in this study were deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (BioProject Accession: PRJNA1029663) and the Gene Expression Omnibus (GSE245749). The published genome used for ChIP-seq analyses was obtained from NCBI (GenBank: NC_000913.3). Publicly available RNA-seq data analyzed for TldR gRNA expression are in the NCBI SRA (ERR6044061) and GEO (GSE115009) databases. The published genomes used for bioinformatics analyses were obtained from NCBI (Supplementary Table 4). The ISfinder database can be accessed at https://www-is.biotoul.fr/index.php.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	Sample sizes are reported in the figure legends and generally encompassed three biological replicates, as in previous studies of TnpB proteins. This sample size is sufficient to reflect the degree of uncertainty and experimental differences between measurements.
Data exclusions	No data were excluded.
Replication	All data could be reproduced, and most experiments and analyses presented were the result of two to three independent biological replicates.
Randomization	Samples were not randomized as it was not applicable for the design of this study (this study did not involve selecting samples from a larger population).
Blinding	Investigators were not blinded as it was not applicable for the design of this study and experiments included necessary controls.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Antibodies

Antibodies used	Monoclonal ANTI-FLAG M2 antibody produced in mouse, Sigma Aldrich, Catalogue number: F1804, clone M2
Validation	According to the manufacturer, the "monoclonal ANTI-FLAG® M2 may be used in IP [immunoprecipitation] procedures when used in conjunction with an insoluble carrier matrix, such as a Protein G resin" (https://www.sigmaaldrich.com/deepweb/assets/sigmaaldrich/product/documents/175/747/f1804bul-ms.pdf). As suggested, the ANTI-FLAG M2 antibody was used together with

Dynabeads Protein G resin (Thermo Fisher) in this study. Furthermore, ANTI-FLAG M2 antibody was used in a ChIP-seq study by Partridge et al., Nature (2020), titled "Occupancy maps of 208 chromatin-associated proteins in one human cell type."

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

Raw sequencing reads, processed sequencing reads and MACS3 peak call files (all files listed below) were uploaded to GEO (accession: GSE245749). The GEO reviewer accession token for this submission is: arynyuoonbanhgl. All data deposited in GEO is publicly accessible.

Files in database submission

Raw sequencing files:

Eca_dTnpB_ChIP-seq_paired_raw_1.fastq.gz
Eca_dTnpB_ChIP-seq_paired_raw_2.fastq.gz
Ece_dTnpB_ChIP-seq_paired_raw_1.fastq.gz
Ece_dTnpB_ChIP-seq_paired_raw_2.fastq.gz
Ecl_dTnpB_ChIP-seq_paired_raw_1.fastq.gz
Ecl_dTnpB_ChIP-seq_paired_raw_2.fastq.gz
Eco_dTnpB_ChIP-seq_paired_raw_1.fastq.gz
Eco_dTnpB_ChIP-seq_paired_raw_2.fastq.gz
Efa1_dTnpB_ChIP-seq_paired_raw_1.fastq.gz
Efa1_dTnpB_ChIP-seq_paired_raw_2.fastq.gz
Efa2_dTnpB_ChIP-seq_paired_raw_1.fastq.gz
Efa2_dTnpB_ChIP-seq_paired_raw_2.fastq.gz
Eho_dTnpB_ChIP-seq_paired_raw_1.fastq.gz
Eho_dTnpB_ChIP-seq_paired_raw_2.fastq.gz
Eko1_dTnpB_ChIP-seq_paired_raw_1.fastq.gz
Eko1_dTnpB_ChIP-seq_paired_raw_2.fastq.gz
Eko2_dTnpB_ChIP-seq_paired_raw_1.fastq.gz
Eko2_dTnpB_ChIP-seq_paired_raw_2.fastq.gz
Emu_dTnpB_ChIP-seq_paired_raw_1.fastq.gz
Emu_dTnpB_ChIP-seq_paired_raw_2.fastq.gz
Ero_dTnpB_ChIP-seq_paired_raw_1.fastq.gz
Ero_dTnpB_ChIP-seq_paired_raw_2.fastq.gz
Esa_dTnpB_ChIP-seq_paired_raw_1.fastq.gz
Esa_dTnpB_ChIP-seq_paired_raw_2.fastq.gz
Gst_TnpB2_ChIP-seq_paired_raw_1.fastq.gz
Gst_TnpB2_ChIP-seq_paired_raw_2.fastq.gz
Input_Eco_dTnpB_ChIP-seq_paired_raw_1.fastq.gz
Input_Eco_dTnpB_ChIP-seq_paired_raw_2.fastq.gz
Kpi_dTnpB_ChIP-seq_paired_raw_1.fastq.gz
Kpi_dTnpB_ChIP-seq_paired_raw_2.fastq.gz
Lec_dTnpB_ChIP-seq_paired_raw_1.fastq.gz
Lec_dTnpB_ChIP-seq_paired_raw_2.fastq.gz
Tos_dTnpB_ChIP-seq_paired_raw_1.fastq.gz
Tos_dTnpB_ChIP-seq_paired_raw_2.fastq.gz

Processed sequencing files (bigWig):

Eca_dTnpB_ChIP-seq_paired.bw
Ece_dTnpB_ChIP-seq_paired.bw
Ecl_dTnpB_ChIP-seq_paired.bw
Eco_dTnpB_ChIP-seq_paired.bw
Efa1_dTnpB_ChIP-seq_paired.bw
Efa2_dTnpB_ChIP-seq_paired.bw
Eho_dTnpB_ChIP-seq_paired.bw
Eko1_dTnpB_ChIP-seq_paired.bw
Eko2_dTnpB_ChIP-seq_paired.bw
Emu_dTnpB_ChIP-seq_paired.bw
Ero_dTnpB_ChIP-seq_paired.bw
Esa_dTnpB_ChIP-seq_paired.bw
Gst_TnpB2_ChIP-seq_paired.bw
Input_Eco_dTnpB_ChIP-seq_paired.bw
Kpi_dTnpB_ChIP-seq_paired.bw
Lec_dTnpB_ChIP-seq_paired.bw
Tos_dTnpB_ChIP-seq_paired.bw
Eca_dTnpB_ChIP-seq_paired_max-value-1kb-windows.bw
Ece_dTnpB_ChIP-seq_paired_max-value-1kb-windows.bw

Ecl_dTnpB_ChIP-seq_paired_max-value-1kb-windows.bw
Eco_dTnpB_ChIP-seq_paired_max-value-1kb-windows.bw
Efa1_dTnpB_ChIP-seq_paired_max-value-1kb-windows.bw
Efa2_dTnpB_ChIP-seq_paired_max-value-1kb-windows.bw
Eho_dTnpB_ChIP-seq_paired_max-value-1kb-windows.bw
Eko1_dTnpB_ChIP-seq_paired_max-value-1kb-windows.bw
Eko2_dTnpB_ChIP-seq_paired_max-value-1kb-windows.bw
Emu_dTnpB_ChIP-seq_paired_max-value-1kb-windows.bw
Ero_dTnpB_ChIP-seq_paired_max-value-1kb-windows.bw
Esa_dTnpB_ChIP-seq_paired_max-value-1kb-windows.bw
Gst_TnpB2_ChIP-seq_paired_max-value-1kb-windows.bw
Input_Eco_dTnpB_ChIP-seq_paired_max-value-1kb-windows.bw
Kpi_dTnpB_ChIP-seq_paired_max-value-1kb-windows.bw
Lec_dTnpB_ChIP-seq_paired_max-value-1kb-windows.bw
Tos_dTnpB_ChIP-seq_paired_max-value-1kb-windows.bw
Eca_dTnpB_ChIP-seq_CPM_paired.bw
Ece_dTnpB_ChIP-seq_CPM_paired.bw
Ecl_dTnpB_ChIP-seq_CPM_paired.bw
Eco_dTnpB_ChIP-seq_CPM_paired.bw
Efa1_dTnpB_ChIP-seq_CPM_paired.bw
Efa2_dTnpB_ChIP-seq_CPM_paired.bw
Eho_dTnpB_ChIP-seq_CPM_paired.bw
Eko1_dTnpB_ChIP-seq_CPM_paired.bw
Eko2_dTnpB_ChIP-seq_CPM_paired.bw
Emu_dTnpB_ChIP-seq_CPM_paired.bw
Ero_dTnpB_ChIP-seq_CPM_paired.bw
Esa_dTnpB_ChIP-seq_CPM_paired.bw
Gst_TnpB2_ChIP-seq_CPM_paired.bw
Input_Eco_dTnpB_ChIP-seq_CPM_paired.bw
Kpi_dTnpB_ChIP-seq_CPM_paired.bw
Lec_dTnpB_ChIP-seq_CPM_paired.bw
Tos_dTnpB_ChIP-seq_CPM_paired.bw
Eca_dTnpB_ChIP-seq_CPM_paired_max-value-1kb-windows.bw
Ece_dTnpB_ChIP-seq_CPM_paired_max-value-1kb-windows.bw
Ecl_dTnpB_ChIP-seq_CPM_paired_max-value-1kb-windows.bw
Eco_dTnpB_ChIP-seq_CPM_paired_max-value-1kb-windows.bw
Efa1_dTnpB_ChIP-seq_CPM_paired_max-value-1kb-windows.bw
Efa2_dTnpB_ChIP-seq_CPM_paired_max-value-1kb-windows.bw
Eho_dTnpB_ChIP-seq_CPM_paired_max-value-1kb-windows.bw
Eko1_dTnpB_ChIP-seq_CPM_paired_max-value-1kb-windows.bw
Eko2_dTnpB_ChIP-seq_CPM_paired_max-value-1kb-windows.bw
Emu_dTnpB_ChIP-seq_CPM_paired_max-value-1kb-windows.bw
Ero_dTnpB_ChIP-seq_CPM_paired_max-value-1kb-windows.bw
Esa_dTnpB_ChIP-seq_CPM_paired_max-value-1kb-windows.bw
Gst_TnpB2_ChIP-seq_CPM_paired_max-value-1kb-windows.bw
Input_Eco_dTnpB_ChIP-seq_CPM_paired_max-value-1kb-windows.bw
Kpi_dTnpB_ChIP-seq_CPM_paired_max-value-1kb-windows.bw
Lec_dTnpB_ChIP-seq_CPM_paired_max-value-1kb-windows.bw
Tos_dTnpB_ChIP-seq_CPM_paired_max-value-1kb-windows.bw

Processed sequencing files (bed files containing MAC3 peak calls):

Eca_dTnpB_macs3_summits.bed
Ece_dTnpB_macs3_summits.bed
Ecl_dTnpB_macs3_summits.bed
Eco_dTnpB_macs3_summits.bed
Efa1_dTnpB_macs3_summits.bed
Efa2_dTnpB_macs3_summits.bed
Eho_dTnpB_macs3_summits.bed
Eko1_dTnpB_macs3_summits.bed
Eko2_dTnpB_macs3_summits.bed
Emu_dTnpB_macs3_summits.bed
Ero_dTnpB_macs3_summits.bed
Esa_dTnpB_macs3_summits.bed
Gst_TnpB2_macs3_summits.bed
Kpi_dTnpB_macs3_summits.bed
Lec_dTnpB_macs3_summits.bed
Tos_dTnpB_macs3_summits.bed

The Meta table uploaded to GEO contains the same read count information as Supplementary Table 5: Table_Supplement.xlsx

Genome browser session
(e.g. UCSC)

Modified reference genomes (accessions listed in Supplementary Table 1) were used. Normalized bigWig files can be visualized in the Integrative Genomics Viewer (IGV) using the bigWig (.bw) file of choice (normalized using bamCoverage) provided in GEO together with the respective reference genome file used for read mapping. To navigate which .bw file and reference genome to use, use Supplementary Table 5.

Methodology

Replicates	One biological replicate was used for ChIP-seq samples.
Sequencing depth	Number of raw reads, uniquely mapped reads, length of reads and paired- or single-end nature are provided in Supplementary Table 5.
Antibodies	Monoclonal ANTI-FLAG M2 antibody produced in mouse, Sigma Aldrich, Catalogue number: F1804, clone M2
Peak calling parameters	#Trim reads using fastp fastp -i "input_read1.fastq.gz" -I "input_read2.fastq.gz" -o "trimmed_output_read1.fastq.gz" -O "trimmed_output_read2.fastq.gz" -j "log".json -h "log".html #Map reads using bowtie2 (creates a .sam output file) #Reads were mapped to a modified E. coli K12 reference genome (derivative of GenBank: NC_000913.3) bowtie2 -x “directory_to_E-coli_reference_genome_file” -1 “trimmed_output_read1.fastq.gz” -2 “trimmed_output_read1.fastq.gz” -S “output.sam” #Convert .sam into .bam file using samtools samtools -view -b “input.sam” > “output_directory” #Sort the .bam files using samtools samtools sort -o output_directory “input.bam” #Index the aligned and sorted .bam files using samtools samtools index -b “input.bam” “output.bam.bai” #Eliminate multi-mapping reads using samtools (retains only uniquely mapping reads); #uses a MAPQ score of 10 as a cutoff samtools view -bq 10 “input.bam” > “output_directory” #Create index files for the trimmed, aligned, sorted and uniquely mapping reads using samtools samtools index -b “input.bam” “output.bam.bai” #Normalize reads using deepTools2 bamCoverage with the option "RPKM" or "CPM" bamCoverage --normalizeUsing CPM -bs 1 -b "not_normalized.bam" -o "normalized.bw" #MACS3 peak calling macs3 callpeak -t target.bam -c input.bam -n "dTnpB" -g 4500000 --nomodel --extsize 400 -q 0.05 -B --outdir "dTnpB_peak_calls" Control file (input file): Input_Eco_dTnpB_ChIP-seq_paired_raw All index files generated using "samtools faidx <E_coli_K12_reference_genome>", and will be made available on GitHub upon publication. Reference genome files as described in Supplementary Figure 5 will be made publicly available on GitHub.
Data quality	Data quality: Bowtie2 used default read quality parameters for mapping. Multi-mapping reads with a MAPQ score <10 were eliminated. All peaks called are below FDR 5%, as per MACS3 standard output. Peaks above 5-fold enrichment: Kpi_dTnpB: 0 Eko1_dTnpB: 2 Eko2_dTnpB: 11 Ero_dTnpB: 17 Tos_dTnpB: 4 GstTnpB2: 1 Ece_dTnpB: 14 Eco_dTnpB: 6 Ecl_dTnpB: 20 Eho_dTnpB: 28 Eca_dTnpB: 170 Efa1_dTnpB: 10 Emu_dTnpB: 2 Esa_dTnpB: 0 Lec_dTnpB: 12 Efa2_dTnpB: 2
Software	Illumina BaseSpace was used for automated read demultiplexing and adaptor trimming. All custom code for the ChIP-seq analysis is available upon request.