Exapted CRISPR-Cas12f homologs drive RNA-guided transcription

Florian T. Hoffmann[1], Tanner Wiegand[1,2], Adriana I. Palmieri[1], Juniper Glass-Klaiber[3], Renjian Xiao[4], Stephen Tang[1], Hoang Le[1], Chance Meers[1,2], George D. Lampe[1,2], Leifu Chang[4,5], Samuel H. Sternberg[1,2†]

[1]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA. [2]Howard Hughes Medical Institute, Columbia University, New York, NY, USA. [3]Department of Cellular Physiology and Biophysics, Columbia University, New York, NY, USA. [4]Department of Biological Sciences, Purdue University, West Lafayette, IN, USA. [5]Purdue Institute for Cancer Research, Purdue University, West Lafayette, IN, USA.

[†]Corresponding author

ABSTRACT

Bacterial transcription initiation is a tightly regulated process that canonically relies on sequence-specific promoter recognition by dedicated sigma (σ) factors, leading to functional DNA engagement by RNA polymerase (RNAP)[1]. Although the seven σ factors in *E. coli* have been extensively characterized[2], Bacteroidetes species encode dozens of specialized, extracytoplasmic function σ factors (σ^E) whose precise roles are unknown, pointing to additional layers of regulatory potential[3]. Here we uncover an unprecedented mechanism of RNA-guided gene activation involving the coordinated action of σ^E factor in complex with nuclease-dead Cas12f (dCas12f). We screened a large set of genetically-linked dCas12f and σ^E homologs in *E. coli* using RIP-seq and ChIP-seq experiments, revealing systems that exhibited robust guide RNA enrichment and DNA target binding with a minimal 5′-G target-adjacent motif (TAM). Recruitment of σ^E was dependent on dCas12f and guide RNA (gRNA), suggesting direct protein-protein interactions, and co-expression experiments demonstrated that the dCas12f-gRNA-σ^E ternary complex was competent for programmable recruitment of the RNAP holoenzyme. Remarkably, dCas12f-RNA-σ^E complexes drove potent gene expression in the absence of any requisite promoter motifs, with *de novo* transcription start sites defined exclusively by the relative distance from the dCas12f-mediated R-loop. Our findings highlight a new paradigm of RNA-guided transcription (RGT) that embodies natural features reminiscent of CRISPRa technology developed by humans[4,5].

INTRODUCTION

Canonical CRISPR-Cas systems function as a widespread form of adaptive immunity in bacteria and archaea by eliminating foreign nucleic acids from invading bacteriophages and plasmids[6]. Within this broad diversity, which spans seven unique types, Cas9 and Cas12 represent archetypal single-effector proteins that target and cleave double-stranded DNA (dsDNA) substrates recognized via base-pairing to a guide RNA (gRNA) and interactions with the protospacer-adjacent motif (PAM)[7,8]. Remarkably, both proteins are ancestrally related to homologous, transposon-encoded nucleases within the TnpB superfamily, which similarly use gRNAs to target dsDNA for cleavage based on RNA-DNA complementary and target-adjacent motif (TAM) recognition, thereby promoting selfish transposon spread[9–12]. The recent discovery of RNA-guided transposases encoded by distinct IS110-family transposons[13,14], as well as their striking resemblance to eukaryotic homologs comprising Nop5 domains[15], establishes the pervasive degree to which noncoding RNAs encoded within mobile genetic elements were exapted for essential functions relating to RNA methylation, RNA splicing, and RNA-guided DNA recognition.

TnpB homologs, whose eukaryotic homologs are known as Fanzors[10,16], are one of the most abundant and ubiquitous proteins in prokaryotes[17]. Although they have primarily expanded in copy number via transposon-mediated horizontal gene transfer, numerous TnpB subfamilies exhibit low mobility scores and unusual genomic neighborhoods indicative of novel function[17], and in some cases, their RuvC nuclease domains harbor active site mutations expected to abolish DNA cleavage behavior[17,18]. Building off this observation, we recently reported numerous instances of TnpB-like nuclease dead repressors (TldR), whose RNA-guided DNA binding activities direct potent gene silencing that facilitates flagellar remodeling of the host, among other biological roles (**Fig. 1a**)[18]. This finding reframes CRISPR-Cas systems as just one pathway among countless TnpB domestication trajectories explored across evolutionary timescales, while also highlighting the recurrent co-option of RNA-guided DNA binding proteins for transcriptional reprogramming. Indeed, even within CRISPR-Cas systems, Cas9 nucleases (Type II) are repurposed as RNA-guided gene repressors for both autoregulatory and virulence control[19–21], Cas12 proteins (Type V) sometimes mediate adaptive immunity by blocking transcription rather than cleaving DNA[22,23], and Cascade complexes (Type I) can silence toxin RNAs to promote their selfish retention[24]. Collectively, these naturally occurring examples resemble the CRISPR interference (CRISPRi) technologies that have been developed by researchers for RNA-guided transcriptional silencing[25–27].

Have TnpB or CRISPR-associated proteins also been coopted for natural modes of transcriptional *activation*, akin to CRISPRa technologies developed by humans[4,5,28,29]? In the course of our ongoing efforts to bioinformatically mine sequenced genomes for additional examples of RNA-guided effectors domesticated for novel cellular functions, we were intrigued to uncover unusual Cas12f genes — which again harbored inactivating RuvC mutations — encoded adjacent to sigma (σ) factor genes, suggesting a potential role in transcription initiation[17]. In bacteria, σ factor is an essential subunit of the RNA polymerase (RNAP) holoenzyme, and it imparts sequence-specific protein-DNA interactions that direct RNAP to promoter sequences where transcription begins[1]. Promoter interactions of the core RNAP subunits without σ are insufficient to initiate transcription, underpinning the central role of σ in bacterial gene expression[30,31].

σ factor genes have been comprehensively studied in *E. coli*, which encode seven distinct homologs that are specialized for transcriptional responses associated with normal growth, stress adaptation, nutrient availability and motility control[32,33], but their numbers are massively expanded in other bacteria. Group 4 σ factors in particular, also known as the extracytoplasmic function (ECF) family because of their roles in envelope stress response and metabolite uptake[34–38], are highly abundant in Bacteroidetes[3] and encompass the specific subtype found adjacent to Cas12f genes. We therefore hypothesized that these unusual operons, comprising nuclease-dead Cas12f (dCas12f) and ECF-like sigma (σ^E) factors, might encode novel machinery to perform RNA-guided transcription of target genes.

Here, we report the discovery of a natural pathway for programmable gene activation in bacteria. After bioinformatically exploring hundreds of dCas12f-σ^E systems, we reconstituted their activity in *E. coli*, demonstrating that dCas12f-gRNA complexes recognize target sites through RNA-DNA base-pairing and a minimal TAM. Target binding triggers recruitment of native σ^E-RNAP holoenzyme complexes, potently initiating transcription initiation at a fixed distance downstream of the target site. In native host cells, dCas12f-σ^E systems are themselves likely controlled by associated transcriptional regulators, and they control the expression of conserved transmembrane transport systems implicated in nutrient import and utilization. Collectively, our findings uncover another unexpected evolutionary link between CRISPR-Cas systems and transcriptional gene regulation, expanding the functional repertoire of RNA-guided mechanisms in bacteria.

RESULTS

Genetic association between nuclease-dead Cas12f and σ^E

We developed a bioinformatics pipeline to identify Cas12-like proteins hypothesized to lack DNA cleavage activity based on the presence of characteristic RuvC nuclease domain mutations, building off of our previous work focused on TnpB[18]. This analysis uncovered a large group of inactivated effectors closely related to Cas12f proteins, of which the majority exhibited a lack of CRISPR association but strong genetic association with ECF-like sigma (σ^E) factors[34,39] and helix-turn-helix (HTH) domain proteins (**Fig. 1b**), paralleling recent observations by Zhang and colleagues[17]. Ancestrally related to this clade, we identified a distinct cluster of putative nuclease-dead Cas12f (dCas12f) proteins without an associated σ^E that are encoded in the vicinity of CRISPR arrays, whose predicted gRNAs sometimes target the promoter region of efflux system expression cassettes elsewhere in the genome, suggesting a potential role in transcriptional repression (**Extended Data Fig. 1a,b**). Compared to nuclease-active Cas12f homologs[40], dCas12f proteins are shorter due to C-terminal truncations that eliminate part of the RuvC active site, but AlphaFold 3 structural models suggested they nonetheless maintain a Cas12f-like fold (**Fig. 1c**).

The strong co-localization of genes encoding dCas12f (*dcas12f*) and σ^E (*rpoE*), including their conserved operonic arrangement (**Fig. 1d**), suggested an intriguing potential coordination between RNA-guided DNA binding and transcription initiation. The nearby association with HTH proteins (encoded by *hth*), whose functions also generally involve transcriptional control[41], was more enigmatic, and phylogenetic analyses indicated a greater degree of HTH diversification and/or recombination within *rpoE-dcas12f*

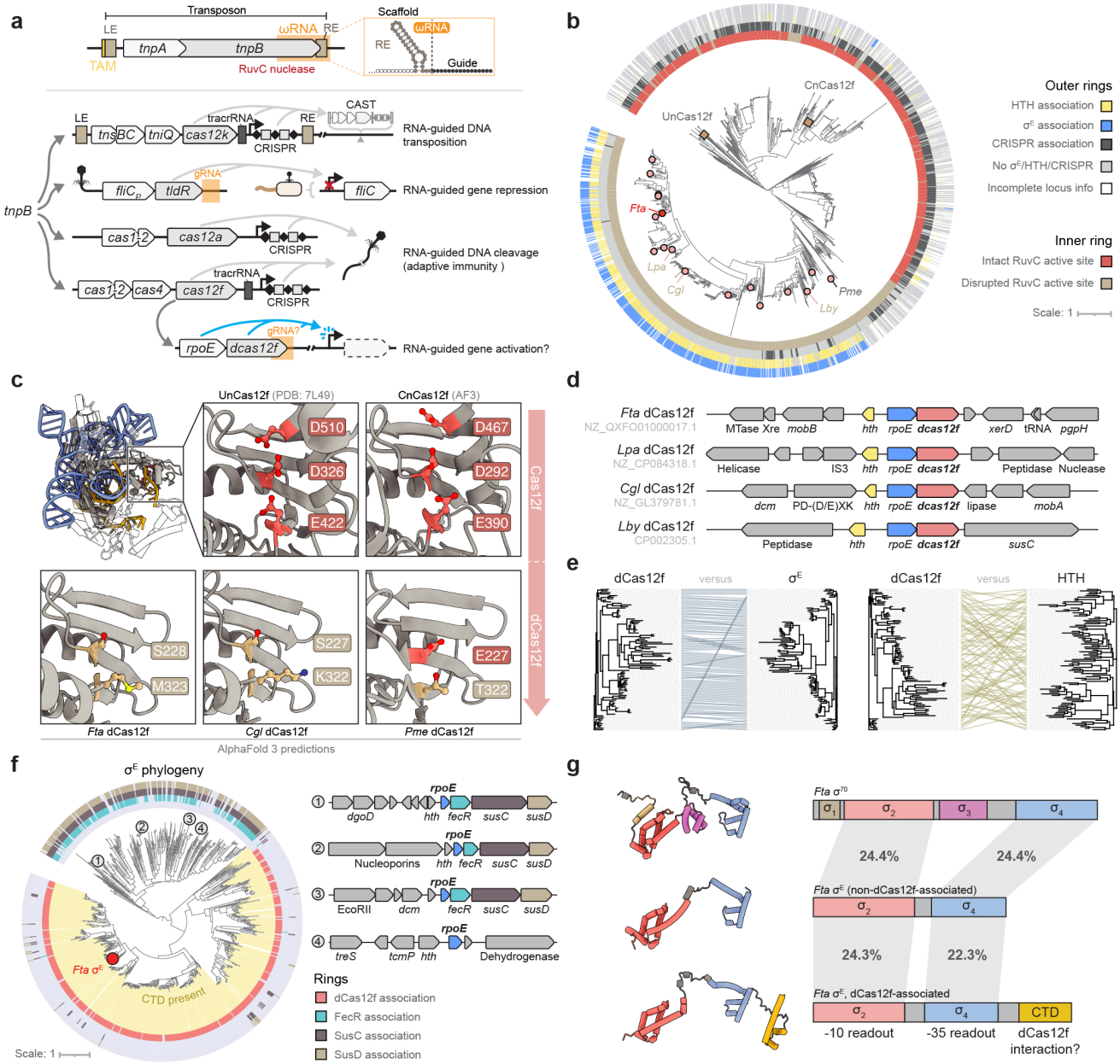

Figure 1 | Naturally occurring, nuclease-dead Cas12f homologs are genetically associated with atypical σ factor genes. a, Schematic depicting recurrent exaptation events of transposon-encoded *tnpB* nuclease genes for novel, RNA-guided functions. Primordial *tnpB* homologs (top) are encoded within the left end (LE) and right end (RE) boundaries of bacterial transposons alongside the *tnpA* transposase, where they function together with ωRNA guides to promote transposon retention. Prior studies indicate that *tnpB* domestication events gave rise to diverse CRISPR-associated effector proteins such as *cas12a*, CRISPR-associated transposon (CAST) targeting proteins such as *cas12k*, and RNA-guided transcriptional repressors such as *tldR* (middle). Here we focus on novel nuclease-dead *cas12f* genes (*dcas12f*, bottom) that function with gRNAs to promote gene activation via interactions with extracytoplasmic sigma factor (σ^E, encoded by *rpoE*). **b,** Phylogenetic tree of Cas12f homologs, showing the progressive inactivation of the RuvC nuclease domain active site (inner ring, red/grey), loss of CRISPR array association, and gain in both HTH and σ^E (*rpoE* gene) association, moving clockwise around the tree. Biochemically characterized Cas12f nucleases are indicated on specific leaves of the tree (red squares), as are candidate nuclease-dead Cas12f homologs (dCas12f) tested in this study (red circles); the homolog from *Flagellimonas taeanensis* (*Fta*, red circle) was selected for detailed characterization. **c,** Magnified view of the RuvC nuclease domain from AlphaFold3 structure predictions of three select dCas12f homologs from the tree in **b** (bottom), compared to biochemically and structurally characterized CRISPR-associated Cas12f nucleases (top). Active site residues and their corresponding naturally occurring mutations are labeled. **d,** Representative genomic neighborhoods of select dCas12f homologs from the tree in **b**, with labeled genomic accession IDs. *dCas12f* genes occur in an operon together with *rpoE* (encoding σ^E), and are proximal to an oppositely oriented *hth* gene. **e,** Comparison of

phylogenetic trees for select dCas12f and σ^E (*rpoE* gene) homologs, and for select dCas12f and HTH homologs, which come from the same genomic neighborhoods. This analysis reveals that *dCas12f* and *rpoE* genes exhibit a strong genetic linkage, as opposed to *dCas12f* and *hth* genes, which do not. **f**, Phylogenetic tree of σ^E (*rpoE* gene) homologs, showing the progressive loss of the FecR and SusCD associations (outer rings) and gain of the dCas12f association (inner ring, red), moving clockwise around the tree; the highlighted yellow sector represents the concomitant gain of the additional C-terminal domain (CTD). Representative genomic neighborhoods for the indicated leaves (1–4) are shown at right, compared to the *Ata* locus in **d**. **g,** AlphaFold3 structure predictions (left) and domain annotations (right) for three σ factors from *F. taeanensis*: σ^{70} (*rpoD* gene), a representative σ^E homolog (*rpoE*) that is non-dCas12f-associated, and the dCas12f-associated σ^E (*rpoE*) from **d**. Each protein shares homologous σ_2 and σ_4 domains that are predicted to recognize the -10 and -35 boxes of transcriptional promoters, but the dCas12f-associated σ^E uniquely exhibits an additional CTD that may interact with dCas12f.

---

loci, compared to the clear pattern of dCas12f and σ^E co-evolution (**Fig. 1e, Extended Data Fig. 1c**).

A more thorough analysis of σ^E phylogeny provided additional clues supporting its predicted interactions with dCas12f (**Fig. 1f**). Canonical σ^E factors are encoded next to anti-σ factors such as FecR, which reside in the inner membrane and sequester σ^E until sensing of an extracellular stimulus activates FecR cleavage and concomitant σ^E release, leading to transcription of target operons[42]. We observed genetic architectures consistent with this paradigm for σ^E clades that were distantly related to dCas12f-σ^E, with *rpoE-fecR* loci furthermore associated with *susCD* genes involved in transmembrane transport (**Fig. 1f**). In stark contrast, *rpoE-dcas12f* loci consistently lacked any FecR or SusCD association, suggesting distinct modes of regulation and biological function. Intriguingly, when we analyzed the domain composition of σ^E proteins more closely, we noticed that dCas12f-associated homologs possessed a unique C-terminal domain (CTD) extension, which we hypothesized could mediate dCas12f interactions. This domain was absent from nearly all non-dCas12f-associated homologs (**Fig. 1f,g**)[17]. The only notable exception was a comparatively rare clade in which *rpoE* genes were encoded in the same operon as unusual restriction endonuclease–like genes, whose protein products were also predicted to be nuclease inactive (**Extended Data Fig. 1d**).

Collectively, these bioinformatics observations strongly suggested a novel role for dCas12f proteins unrelated to CRISPR-Cas adaptive immunity, and one that likely involves transcriptional regulation by σ^E and/or HTH. We were especially interested in testing the hypothesis that RNA-guided DNA targeting by dCas12f might lead to functional recruitment of σ^E and RNA polymerase (RNAP), but were initially hampered by the inability to computationally predict dCas12f-associated guide RNAs.

dCas12f homologs exhibit RNA-guided DNA target binding

To experimentally investigate dCas12f function, we applied a heterologous expression strategy in *E. coli* that would reveal both its gRNA and target DNA binding motifs via RNA and chromatin immunoprecipitation sequencing (RIP-seq; ChIP-seq), respectively (**Fig. 2a**)[18,43]. We selected 16 phylogenetically diverse dCas12f-σ^E systems (**Fig. 1b, Extended Data Fig. 2**) and cloned both protein-coding genes on a single effector expression plasmid, alongside *hth* and candidate regions that we hypothesized would contain gRNAs (**Fig. 2a**). The resulting RIP-seq data revealed highly enriched gRNAs that were encoded immediately downstream of *dcas12f*, with clearly demarcated 5'/3' ends and average lengths of 85–105 nt

(**Fig. 2b,c, Extended Data Fig. 3a,b**). dCas12f homologs from *Flagellimonas taeanensis* (*Fta*) and *Leadbetterella byssophila* (*Lby*) associated with a single gRNA, whereas homologs from *Leeuwenhoekiella palythoae* (*Lpa*) and *Paenimyroides ummariense* (*Pum*) enriched multiple tandemly encoded gRNAs with similar sequence (**Fig. 2c, Extended Data Fig. 3b,c**).

To further corroborate our heterologous identification of functional gRNAs, we obtained multiple *Fta* strains, developed conditions to culture them (**Methods**), and performed whole genome sequencing (**Extended Data Fig. 4a,b**). The genome for *Fta* strain 4759 encodes two distinct *rpoE-dcas12f* loci (**Extended Data Fig. 4b,c**), and gRNAs from both were highly expressed and exhibited boundaries that were in excellent agreement with RIP-seq results (**Fig. 2d,e**). Intriguingly, we also observed a highly abundant non-coding RNA (ncRNA) within the intergenic region between *hth* and *rpoE*, as well as a similar ncRNA sequence adjacent to an *hth* homolog present elsewhere in the genome (**Fig. 2d,e, Extended Data Fig. 4d**) suggesting potential regulatory roles (see below).

Next, we performed ChIP-seq experiments with the same 16 dCas12f-σ^E systems, hypothesizing that off-target sites resembling the gRNA would be promiscuously bound by dCas12f-RNA complexes, enabling unambiguous identification of the gRNA guide sequence and target adjacent motif (TAM). The resulting data revealed hundreds of highly enriched regions for those dCas12f homologs whose gRNAs were readily identified by RIP-seq (**Fig. 2f, Extended Data Fig. 5a**), and subsequent analyses uncovered conserved motifs common to the majority of peaks within each dataset (**Fig. 2g, Extended Data Fig. 5b**). By integrating these motif results with RIP-seq data and RNA structural analyses, we were able to confidently deduce the TAM and a 14–16-nt guide sequence for 7 distinct dCas12f systems (**Fig. 2h, Extended Data Fig. 3a-c, 5b**). Of note, the predicted TAM for the dCas12f homolog from *Fta* was just a single 5'-G nucleotide.

Our analyses thus far failed to identify tracrRNA-like species, in contrast to related Cas12f1 homologs from Type V-F1 CRISPR-Cas systems[6,44,45]. Instead, comparisons of empirically identified dCas12f gRNAs across multiple systems revealed strong primary sequence conservation and secondary structure predictions of a multiple stem-loop architecture more reminiscent of TnpB-associated ωRNAs, with a guide sequence positioned at the 3' end (**Fig. 2i**)[12,17,18,46]. We therefore concluded that dCas12f homologs would function with single guide RNAs comprising both a scaffold and guide portion, and hypothesized that they might have evolved from an ancestral state more akin to CRISPR arrays. In striking agreement with this notion, we identified an *rpoE-dcas12f* locus from *Sphingobacterium* that appears to encode full-length gRNAs closely resembling the repeat sequences found in CRISPR arrays interspersed throughout the same locus (**Extended Data Fig. 3d**). This observation suggests the intriguing possibility that dCas12f-associated gRNAs emerged from the repeat-spacer sequence of CRISPR arrays themselves.

Bioinformatic identification of genomic sites targeted by dCas12f-gRNA

With knowledge of the native guide sequence, we next sought to identify putative genomic targets of dCas12f, based on the hypothesis that RNA-guided σ^E recruitment might lead to RNAP binding and transcription initiation. Strikingly, when we manually searched the *Fta* genome for potential targets, we

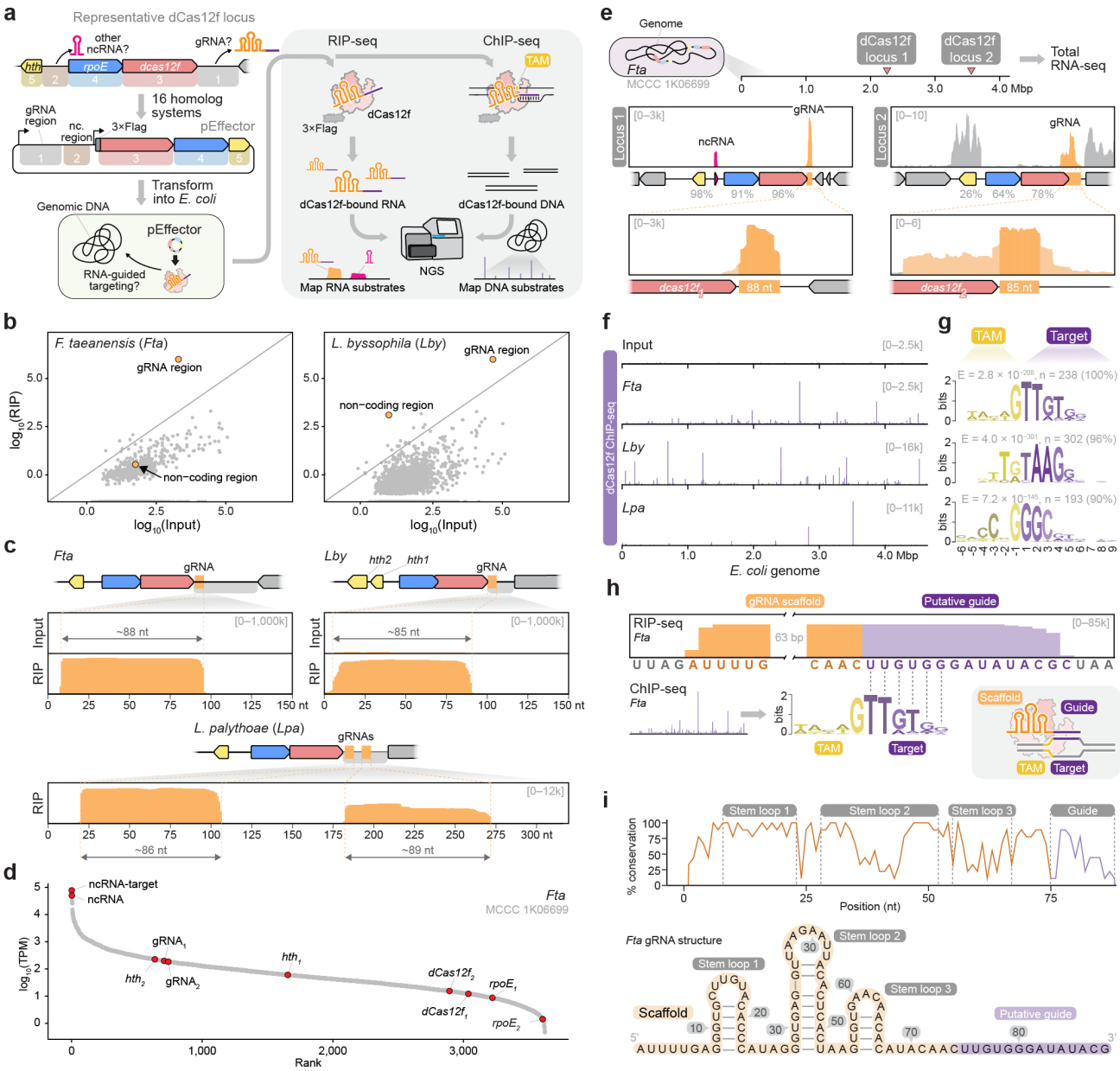

Figure 2 | Experimental discovery of guide RNA and target DNA substrates of dCas12f homologs. a, Vector design (left) and experimental workflow (right) to identify bound gRNAs and target DNA sites by RIP-seq and ChIP-seq, respectively. The indicated genes and non-coding regions from select dCas12f loci (left, top) were cloned into heterologous expression plasmids (left, middle), and *E. coli* cells were then transformed, cultured, and subjected to RIP-seq and ChIP-seq analysis (right), taking advantage of a 3×Flag tag appended to the N-terminus of dCas12f. The resulting next-generation sequencing (NGS) data report on non-coding RNAs (ncRNA) and DNA target sites bound by dCas12f. **b,** Representative RIP-seq data for *Fta* (left) and *Lby* (right) dCas12f , with the scatter plot comparing the log-transformed RNA abundance in the IP fraction (y-axis) versus the input fraction (x-axis). Both homologs strongly enriched a species from the gRNA region (labeled 1 in **a**), and *Lby* also enriched a second species from the non-coding region (labeled 2 in **a**). **c,** RIP-seq coverage plots comparing the input and IP fraction for the three indicated homologs, with the corresponding genomic neighborhoods schematized above each graph. dCas12f associates with a well-defined gRNA that is 85–89 nucleotides (nt) in length, encoded immediately downstream of the *dcas12f* gene. Interestingly, the *Lpa* homolog co-purifies with two gRNAs encoded in tandem (bottom). **d,** Ranked transcript abundance, plotted as the log-transformed transcripts per million (TPM), from RNA-seq data of a native *F. taeanensis* strain that encodes two distinct *rpoE-dcas12f* systems. Key genes and RNAs from each operon are labeled, including the highly abundant gRNAs and *hth*-associated ncRNA. **e,** RNA-seq coverage plots from data shown in **d,** mapped onto the two distinct *dcas12f–rpoE* loci encoded within the genome. These results are in excellent agreement with the *Fta* gRNA identified from heterologous RIP-seq

experiments in *E. coli* shown in **c**, and they furthermore reveal the highly expressed ncRNA (magenta) that is associated with, and expressed upstream from, the *hth* gene. The percent sequence identities between the proteins encoded by *hth*, *rpoE*, and *dcas12f* genes in this native strain, and the corresponding *Fta* homologs tested heterologously in *E. coli*, are shown at top. **f,** Genome-wide representation of ChIP-seq data for the indicated dCas12f homologs analyzed in **c** (purple), compared to the input control (top). Coverage is shown as counts per million (CPM), normalized to the highest peak in the targeting sample. **g,** Binding events were analyzed by MEME-ChIP, which revealed strongly conserved consensus motifs that correspond to the putative target-adjacent motif (TAM) and gRNA-matching target DNA sequence within the seed, for each homolog. E, E-value significance; n, number of peaks contributing to the motif. **h,** Schematic demonstrating how a comparison of the RIP-seq and ChIP-seq motif data for the *Fta* dCas12f homolog results in high-confidence identification of the TAM and target DNA-gRNA base-pairing, consistent with the expected arrangement of the bound R-loop complex (bottom right). **i,** Sequence (top) and secondary structure (bottom) of the *Fta* dCas1f-associated gRNA from RIP-seq data in **b,c**, with labeled stem loop and guide features. The graph at the top shows sequence conservation across an alignment of nine gRNAs identified in RIP- or RNA-seq experiments, at positions occupied by a nucleotide in the *Fta* ncRNA.

---

immediately identified candidate sites matching both dCas12f-associated gRNAs within the large intergenic region upstream of a *susCD* operon implicated in nutrient import and utilization (**Fig. 3a,b**). Both sites were flanked by the expected 5'-G TAM, exhibited 13-14 matching base pairs with the predicted 14-nt guide sequence, and were positioned in an appropriate location to potentially serve promoter-like functions (**Fig. 3b**). Intriguingly, the very same genomic region encoded one of the *hth*-associated ncRNAs noted earlier, and upon closer inspection, we observed that the ncRNA overlapped precisely with the 5' edge of the gRNA and exhibited the same strandedness (**Fig. 3a,b**).

To explore dCas12f targeting more comprehensively, we systematized the search by building covariance models for the gRNA and *hth*-associated ncRNA (**Fig. 3c**), and then globally interrogated genomes with *rpoE-dcas12f* loci to identify their putative genomic targets (**Extended Data Fig. 6a,b**). This search revealed four distinct categories of predicted dCas12f function, all of which were consistent with potential roles in transcription initiation. Most of the queried genomes harbored RNA-guided DNA targets upstream of *susCD* operons involved in transmembrane trafficking (category I) (**Fig. 3d, Extended Data Fig. 6c**), similar to the *Fta* genome (**Fig. 3a,b**). We identified other prominent targets upstream of loci encoding two component systems, with roles in cell signaling (category II) (**Fig. 3d, Extended Data Fig. 6c**). Surprisingly, we also observed gRNA-matching target sites upstream of *rpoE-dcas12f* loci themselves, with gRNAs encoded either immediately downstream of the locus, suggesting positive auto-regulatory feedback *in cis* (category III), or elsewhere in the genome, suggesting a more complex *trans*-acting regulatory network (category IV) (**Fig. 3d**). In some cases, we identified gRNAs that putatively target up to seven distinct loci spanning multiple functional categories (**Extended Data Fig. 6d**), suggesting that *rpoE-dcas12f* loci are responsible for the coordinated regulation of entire gene programs spread across the genome.

The conspicuous and conserved presence of gRNA-matching target sites within intergenic regions that fall upstream of protein-coding genes strongly corroborated our hypothesis that R-loop formation by dCas12f-gRNA complexes could result in recruitment of σ^E and RNAP, thereby leading to transcription initiation. Furthermore, the frequent presence of target sites upstream of TonB-linked SusCD transport-

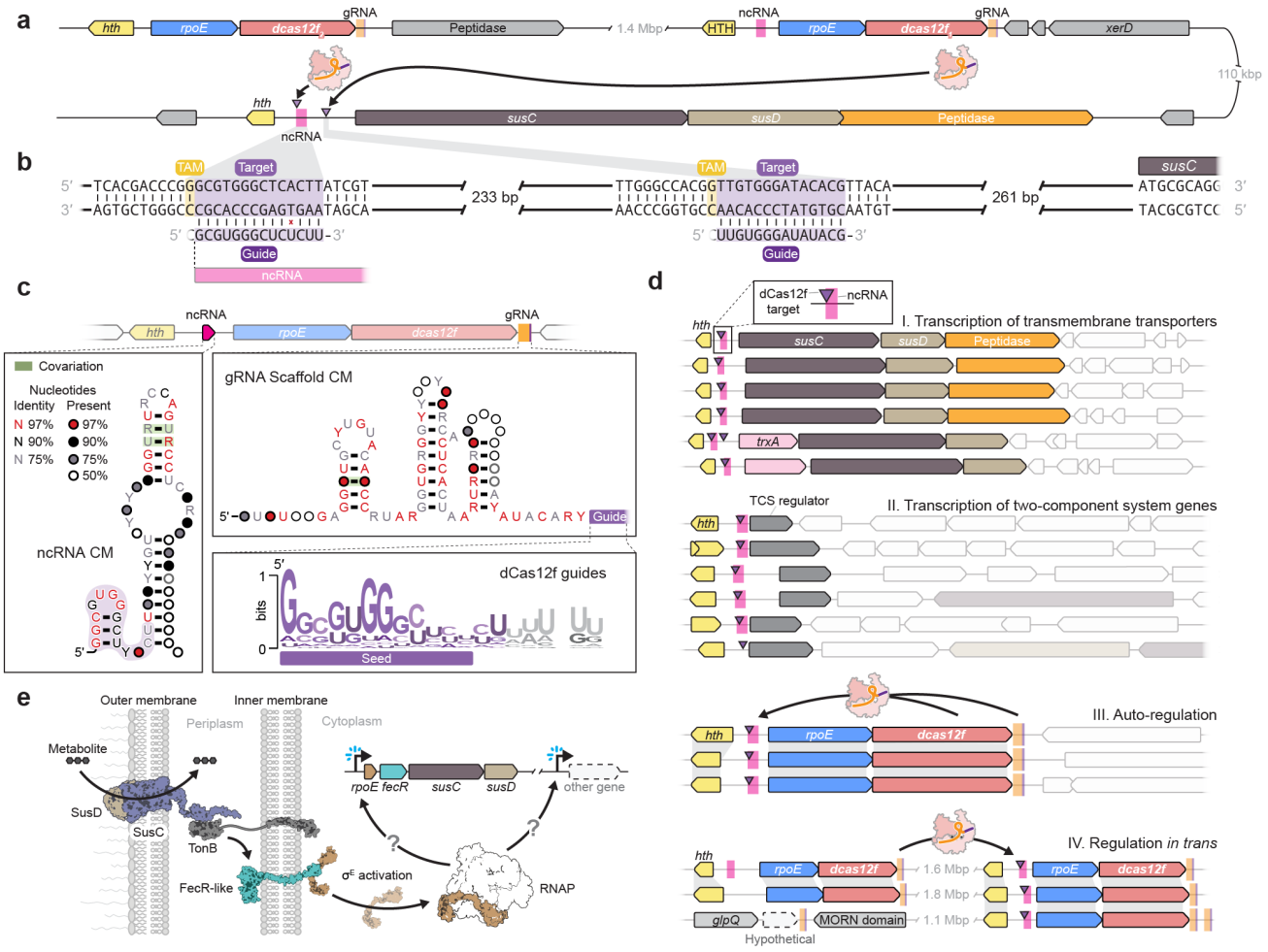

Figure 3 | dCas12f-associated gRNAs target conserved ncRNA loci and regulate diverse gene expression programs. **a,** Schematic of genomic sites (1, 2) targeted by gRNAs associated with two distinct *dcas12f* operons, in the native *Fta* strain from **Fig. 2d**. Intriguingly, the target site matching gRNA.2 precisely overlaps the site of an HTH-associated ncRNA (magenta arrow), which itself closely resembles a similar ncRNA associated with the *hth* gene in the dCas12f.1 operon. **b,** Schematic of gRNA-target DNA complementarity for targets 1 and 2 (left and right, respectively), with the TAM and guide sequences labeled accordingly; base-pairing is indicated with black connecting lines. The guide sequence at target 1 overlaps the 5′ end of the HTH-associated ncRNA, schematized below. The distance between each RNA-DNA match and the *susC* start codon is indicated. **c,** Covariance models (CM) of the *hth*-associated ncRNA (left) and gRNA scaffold (right), with conserved and covarying positions represented as shown in the legend, shown below a representative *dcas12f* locus, The WebLogo at bottom right depicts conservation of the guide sequence, whose 5′ seed sequence closely resembles the 5′ end of the *hth*-associated ncRNA (shaded purple region). Purple bases in the WebLogo correspond to the 14-nucleotide guide length of *Fta*. **d**, Schematic showing four major categories of genetic loci targeted by dCas12f-associated gRNAs, from a systematic bioinformatics analysis of sequenced bacterial genomes; predicted gRNA-matching target sites are indicated with purple triangles, and predicted *hth*-associated ncRNAs are indicated with magenta triangles. The most frequent targets putatively regulate the transcription of transmembrane transporters (I) and two-component system (TCS) regulator genes (II), but gRNAs also likely control expression of the *rpoE-dcas12f* operon itself via auto-regulation or regulation *in trans* (III and IV). The final example in class IV is a locus encoding an orphan gRNA that putatively controls *rpoE-dcas12f* expression *in trans.* Note that dCas12f-associated gRNAs in some cases regulate multiple loci within the same genome. **e,** Model of canonical σ^E-mediated transcriptional activation of *susCD* and/or other genes to regulate outer membrane transport systems. SusC and SusD mediate the uptake of metabolites across the outer membrane in coordination with TonB, whose activities are signalled through a FecR-like protein to activate σ^E, an extracytoplasmic function (ECF) sigma factor encoded by *rpoE* that recruits RNA polymerase (RNAP) to drive transcription at the indicated promoters. The distinct genomic neighborhoods of homologous *rpoE* genes shown in **d** represent likely novel regulatory functions.

ers strongly implicates these systems in regulating uptake of specific nutrients, and indeed, TonB-linked transporters are canonically regulated by σ^E factors, which increase expression when the extracellular nutrient concentration escalates (**Fig. 3e**)[47,48]. Our observation that some *rpoE-dcas12f* loci are natively encoded adjacent to *susCD* operons (**Fig. 1d**), and that non-dCas12f-associated *rpoE* genes are typically encoded proximally to similar *susCD* gene clusters (**Fig. 1f**), strongly supports this functional link. We even detected instances in which rudimentary CRISPR arrays encoded adjacent to *rpoE-dcas12f* loci harbor conventional spacers that also target *susCD* operons, consistent with the hypothesis that *susCD*-specific spacers may have been integrated via spacer acquisition prior to the emergence of chimeric single gRNAs (**Extended Data Fig. 6e**).

Remarkably, we consistently predicted *hth*-associated ncRNAs that precisely overlap with gRNA-matching target sites across each of these major categories (**Fig. 3c,d, Extended Data Fig. 6b-d**). Moreover, these gRNAs and ncRNAs are encoded on the same strand, ruling out a role for sense-antisense transcriptional interference. Although the biological roles of the ncRNA themselves remain unknown, as do the promoter elements controlling their expression, we hypothesize that the *hth*-associated ncRNAs might function as negative regulators of the same genes being acted upon by *rpoE-dcas12f* systems, similar to the role of anti-σ factors regulating canonical σ^E.

The dCas12f-gRNA complex guides σ^E to programmable target sites

Having identified the native guides and TAMs of dCas12f-σ^E systems, we investigated targeting by reprogramming the gRNA guide sequence and performing ChIP-seq. Focusing on the *Fta* system, we designed five gRNAs to target sites tiled across the *E. coli* genome, with no sequence constraints except the requirement of a flanking 5'-G TAM (**Fig. 4a**). Strong enrichment at all five target sites confirmed that this minimal TAM and guide-target pairing were sufficient for efficient dCas12f binding (**Fig. 4a**), and additional dCas12f homologs exhibited similar activities (**Extended Data Fig. 7a**). To determine whether σ^E is recruited alongside dCas12f, we relocated the 3×Flag tag to σ^E and performed similar ChIP-seq experiments, which revealed σ^E occupancy at the very same target sites (**Fig. 4b**). Control experiments using ChIP-qPCR further demonstrated that dCas12f and its associated gRNA were necessary and sufficient for target-specific binding of dCas12f, but that σ^E, HTH, and the *hth*-associated ncRNA were dispensable (**Fig. 4c**).

Although phylogenetic information disfavored the hypothesis that the HTH co-evolved with dCas12f and σ^E to form a larger complex, we nevertheless performed experiments in which HTH was directly immunoprecipitated. The resulting ChIP-seq data revealed widespread HTH binding across the *E. coli* genome, in stark contrast to the discrete RNA-guided binding profiles of dCas12f and σ^E (**Fig. 4d**). Motif detection uncovered a conserved DNA binding motif that precisely matched a highly enriched sequence found within the intergenic sequence immediately upstream of the *hth* coding sequence, which we cloned into the expression plasmid (**Fig. 4d–f**). This region encompasses inverted repeats that align with the MEME-derived motif and are signature constituents of HTH binding sites[41,49]. Intriguingly, the HTH binding site is situated 27 bp from the *hth* start codon, strongly implicating HTH-DNA binding in gene

repression, and a broader analysis revealed inverted repeats within 50 bp of most *dcas12f*-associated *hth* genes (**Extended Data Fig. 7b**), suggesting that this mode of auto-regulation is common in *dcas12f-rpoE* loci.

Collectively, these results demonstrate direct recruitment of σ^E to genomic sites bound by dCas12f-gRNA complexes, independent of HTH. To test whether such interactions are sufficient to drive

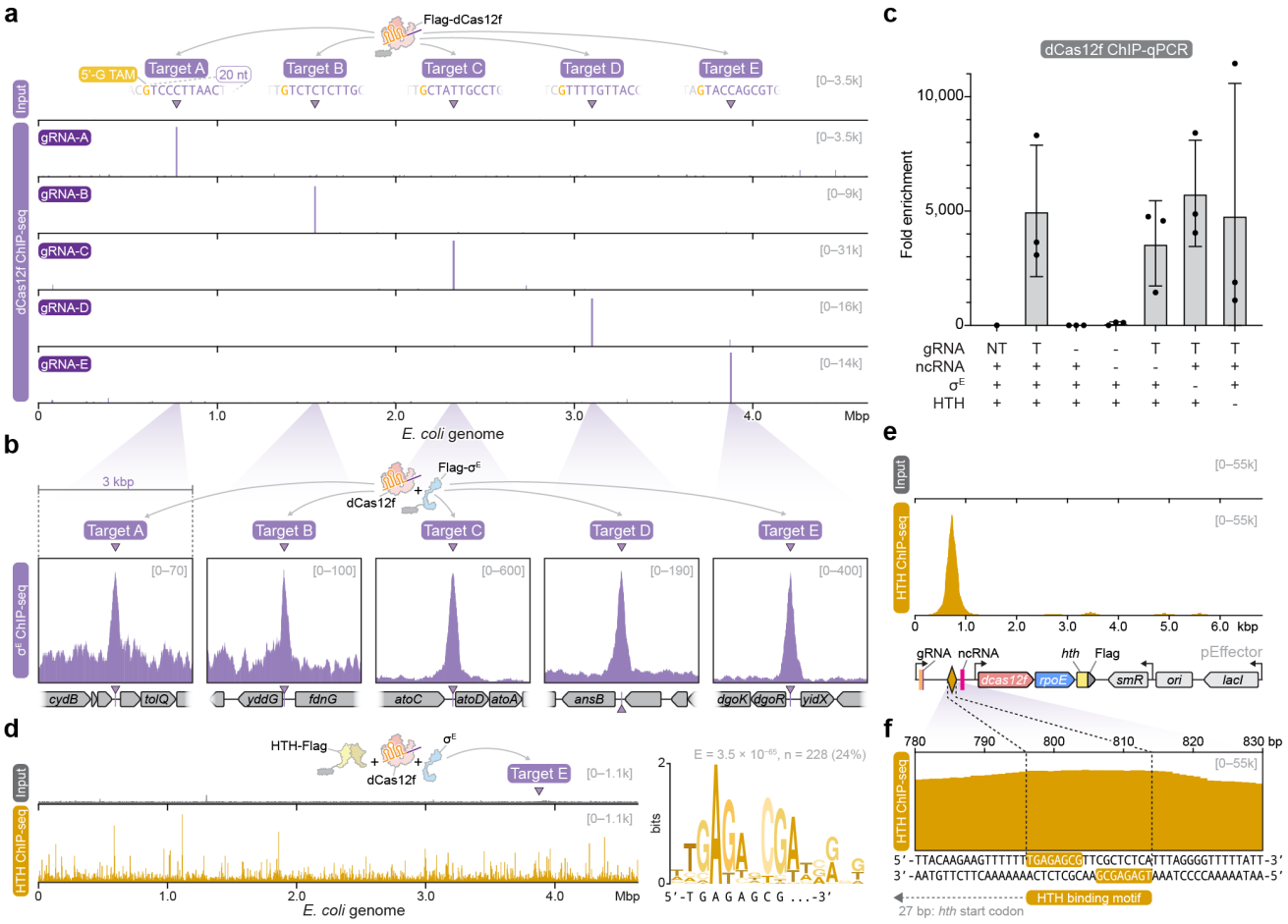

Figure 4 | RNA-guided dCas12f recruits σ^E, but not HTH, to bound genomic target sites. a, Genome-wide representation of ChIP-seq data for *Fta* Flag-dCas12f programmed by gRNAs targeting one of five *E. coli* genomic target sites (A–E, marked with purple triangle) flanked by a 5′-G TAM, alongside the input control (top). Coverage is shown as counts per million (CPM), normalized to the highest peak in the targeting sample. **b,** Magnified view of ChIP-seq data for targets A–E in **a**, but from experiments expressing Flag-σ^E and untagged dCas12f, demonstrating that σ^E is efficiently recruited to RNA-guided DNA target sites. Genomic regions are annotated below each graph, and CPM coverage is normalized to the highest peak. **c,** Bar graph showing ChIP-qPCR data from experiments in which *Fta* Flag-dCas12f was tested alongside the indicated conditions. Data are shown as a fold enrichment, calculated from $\Delta\Delta C_q$. T, targeting gRNA; NT, non-targeting gRNA; data are mean ± s.d. for $n = 3$ biologically independent samples. **d,** Genome-wide representation of ChIP-seq data (left) for HTH-Flag co-expressed with σ^E, dCas12f, and a target E-specific gRNA, alongside the input control (top left); coverage is shown as in **a**. Binding events were analyzed by MEME-ChIP (right), which revealed a strongly conserved consensus motif corresponding to the putative native HTH binding site (see **e,f**). E, E-value significance; n, number of peaks contributing to the motif. **e,** pEffector plasmid-wide representation of ChIP-seq data for HTH-Flag co-expressed with σ^E, dCas12f, and a target E-specific gRNA, alongside the input control (top left). Coverage is shown as in **a**, and plasmid annotations are shown below the graph. **f**, Magnified view of the peak center from **e**, shown about the underlying DNA sequence. Comparison to the motif from MEME analysis shown in **c** reveals the likely HTH binding motif comprising two perfectly inverted repeats (yellow boxes, bottom).

transcription, we next developed a sensitive reporter assay to investigate RNA-guided gene expression.

dCas12f-gRNA and σ^E recruit RNAP for RNA-guided transcription

We utilized our dCas12f-σ^E effector plasmid for a fluorescence reporter assay, in which we cloned the native intergenic *susC* sequence upstream of a fluorescent reporter gene on a second plasmid (pTarget) (**Fig. 5a**). We anticipated that, in the presence of a gRNA with a native guide sequence, dCas12f-mediated DNA binding would recruit σ^E, thereby driving transcription and stimulating red fluorescence. Disappointingly, however, initial experiments resulted in no detectable RFP expression above background levels (**Fig. 5b**). Reasoning that *Fta* σ^E might be unable to productively engage *Eco*RNAP, we cloned an additional expression plasmid encoding all four *Fta*RNAP subunits (α, β, β', ω) and repeated the assay (**Fig. 5a**). Remarkably, co-expression of dCas12f, a targeting gRNA, σ^E, and cognate *Fta*RNAP was sufficient to robustly activate RFP expression in *E. coli* (**Fig. 5b**). Control experiments demonstrated the necessity of all components for this transcriptional activation, with the exception of HTH and the *hth*-associated ncRNA, whose absence or presence had no impact on RFP expression (**Fig. 5b**). Furthermore, RNA sequencing confirmed strong upregulation of *mRFP* transcription compared to all other *E. coli* genes in presence of a targeting gRNA (**Extended Data Fig. 8a**), and revealed the emergence of a well-defined, novel transcription start site (TSS) located 46 bp downstream of the TAM (**Fig. 5c**). Additional controls demonstrated that RFP fluorescence was not simply the result of transcriptional readthrough from plasmid-encoded *E. coli* promoters, and that bulk fluorescence measurements agreed closely with parallel single-cell measurements using flow cytometry (**Extended Data Fig. 8b–d, Supplementary Fig. 1**).

We were immediately curious as to whether the native target site contained additional –10 and –35 promoter motifs upstream of the TSS recognized by *Fta* σ^E, or whether gRNA-DNA base pairing via dCas12f provided the sole sequence specificity determinants. Consistent with this latter scenario, comprehensive bioinformatics analysis of all dCas12f gRNA-matching target sites showed an absence of any conserved promoter motifs beyond the presumed RNA-DNA heteroduplex site (**Fig. 5d**), apart from a weak thymine preference that may facilitate DNA bending. These clues suggested the intriguing possibility that promoter specificity in dCas12f-σ^E systems may rely exclusively on guide RNA-target DNA complementarity, with little or no dependence on direct σ^E-DNA sequence recognition. To test this, we systematically mutated the region downstream of the guide-target match — including any putative –10 and –35 elements, and the TSS (**Fig. 5e**) — and then repeated reporter assays. Strikingly, RFP expression remained essentially unchanged in the presence of these sequence perturbations, including mutation 8 that scrambled the entirety of the sequence following the RNA-matching target site (**Fig. 5f**). These results indicate that sequence-specific σ^E-DNA interactions are unlikely to be involved in transcription initiation by dCas12f-σ^E systems. Structural data from a parallel study indeed suggests that dCas12f-associated σ^E engages DNA through non-canonical interactions, explaining reduced requirements for the –10 and –35 promoter motifs compared to typical σ factors[50].

Given the strong stimulation of RNA-guided transcription by *Fta*RNAP, we were curious to repeat Flag-σ^E ChIP-seq assays in its presence. Interestingly, inclusion of *Fta*RNAP substantially increased

levels of σ^E enrichment at dCas12f-bound target sites, and the sequencing coverage extended well beyond the RNA-matching target site (**Extended Data Fig. 8e**). This result underscores the critical interactions observed between σ^E and its cognate RNAP within the structure of the transcription initiation complex[50],

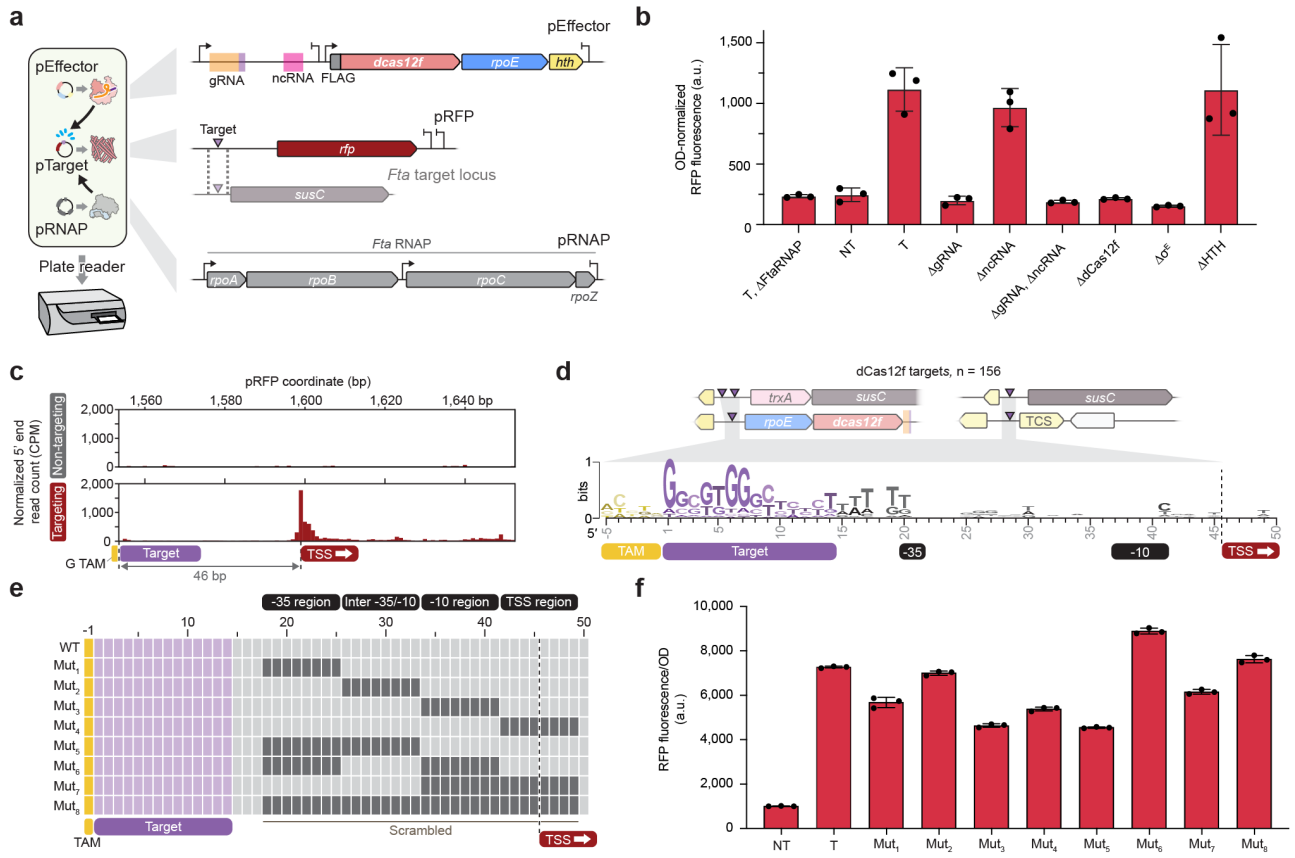

Figure 5 | dCas12f and σ^E promote RNA-guided transcription independent of conserved promoter motifs. a, Schematic of transcriptional reporter assay (left) in which protein and RNA components encoded by pEffector and pRNAP drive gene expression of a plasmid-encoded RFP (pTarget), as measured by a fluorescence plate reader. The architecture of each vector is shown in the expanded view at right; *rfp* was cloned downstream of the *Fta* intergenic region that encompasses the predicted native target site. **b,** OD-normalized RFP fluorescence for the indicated experimental conditions, using the transcriptional reporter assay shown in **a**. T, targeting gRNA; NT, non-targeting gRNA. RNA-guided transcription requires dCas12f, targeting gRNA, σ^E, and cognate *Fta*RNAP, but not HTH or the HTH-associated ncRNA. **c,** Magnified view of RNA-seq data mapped to the indicated pTarget coordinates for targeting (T) and non-targeting (NT) samples in **b**, shown as the 5′ end read counts per million (CPM); the TAM and target site are shown below the graph. Transcription initiation is tightly confined to a transcription start site (TSS) positioned 46 bp downstream of the TAM. **d,** Consensus WebLogo (bottom) of dCas12f gRNA-matching target sites from 156 aligned genomic loci (schematic, top right), demonstrating conservation within the TAM, the target region, and a short T-rich stretch immediately adjacent to the target — but an absence of substantial conservation at either the expected -35 and -10 boxes, or the TSS itself. This observation suggests that RNA-guided transcription mediated by dCas12f and σ^E may proceed largely without DNA sequence preferences imposed by either protein component. **e**, Panel of mutations in the intergenic region in between the RNA-matching target site and empirically determined TSS, which were tested in the transcription reporter assay shown in **a**. The nucleotides highlighted in dark grey were mutated and corresponding putative promoter sequence or TSS regions are highlighted in black boxes at the top. **f,** OD-normalized RFP fluorescence for the indicated reporter DNA constructs shown in **e**, using the transcriptional reporter assay shown in **a**. WT, targeting gRNA with WT (unmutated) intergenic region; NT, non-targeting gRNA control. RNA-guided transcription does not depend on any critical sequence features other than RNA-DNA complementarity, since scrambling of various sequence features downstream of the target site has little to no impact on RFP expression. Data in **b** and **f** are shown as mean ± s.d. for $n = 3$ biologically independent samples.

while further reflecting their likely co-migration during the initiation and elongation phases of transcription, prior to σ^E being released from the RNAP complex, as has been observed for other well-studied σ factors[51].

Finally, we sought to determine the optimal and minimal guide length required for efficient RNA-guided transcription, by incrementally reducing RNA-DNA complementarity in 2-bp steps (**Extended Data Fig. 8f**). We observed strong fluorescence levels for all guide lengths greater than 10 nt, and a 14-nt guide elicited the most efficient RFP expression (**Extended Data Fig. 8g**), in excellent agreement with the *Fta* guide length determined by RIP-seq and native RNA-seq (**Fig. 2c,e,h**).

dCas12f-σ^E systems can be reprogrammed to transcribe user-defined regions of interest

Given the absence of promoter requirements for transcription by dCas12f-σ^E, other than an RNA-DNA match, we hypothesized that gRNAs could be designed to initiate gene expression at arbitrary loci in the *E. coli* genome, with single nucleotide TSS precision. We focused on a transcriptionally silent intergenic site where ChIP-seq revealed strong dCas12f and σ^E binding (**Fig. 4a,b**), and then assessed RNA-guided transcription in the presence of *Fta*RNAP by RNA-seq (**Fig. 6a**). Remarkably, our results again revealed precise transcription initiation 46 bp downstream of the TAM, dependent only on dCas12f, σ^E, *Fta*RNAP, a targeting gRNA, and a 5'-G TAM (**Fig. 6b**). Transcriptome-wide analyses confirmed the high specificity of gene activation (**Fig. 6c**), and RT-qPCR corroborated an approximately 140-fold increase in *yidX*-derived transcripts compared to non-targeting controls (**Fig. 6c, Extended Data Fig. 9a**). Interestingly, we also detected significant upregulation of two additional genes — *tyrS* and *pdxY* — but closer inspection revealed a likely off-target site upstream of both genes that exhibited substantial similarity to the guide sequence (**Fig. 6c, Extended Data Fig. 9b**). This observation highlights the potency of RNA-guided transcription, its ability to activate multiple genes across operons, and the potential for off-target activity in a TAM- and seed sequence-dependent manner.

Next, building off our hypothesis that RNA-guided transcription was dependent only on RNA-DNA complementarity, and not on additional promoter motifs, we envisioned that transcription start sites could be predictably shifted within any desired region of interest, simply by redesigning the guide sequence. To test this, we generated a panel of seven gRNAs tiled across the *dgoR-yidX* locus, with the sole criteria involving a 5'-G TAM and 20 bp of RNA-DNA complementarity. The resulting RNA-seq data strikingly revealed a concomitant and predictable shift in the TSS for each tested gRNA (**Fig. 6d, Extended Data Fig. 9c**), with only Target 3 exhibiting intractability, likely due to competing transcriptional regulators binding within this region (**Extended Data Fig. 9d**). RNA-guided transcription with the dCas12f-σ^E system strictly required a 5'-G TAM, as activity was lost with gRNAs shifted sequentially by 1 bp to sample each potential TAM at Target 6 (**Extended Data Fig. 9e**). Of note, RNA-guided transcription functioned robustly even when targeting within the *dgoR* and *yidX* gene bodies, where pre-existing promoter motifs would surely be absent, and we were easily able to drive antisense transcription by designing gRNAs to target the opposite DNA strand (**Fig. 6e**).

To explore the mechanism by which RNA-guided DNA targeting by dCas12f leads to functional σ^E-

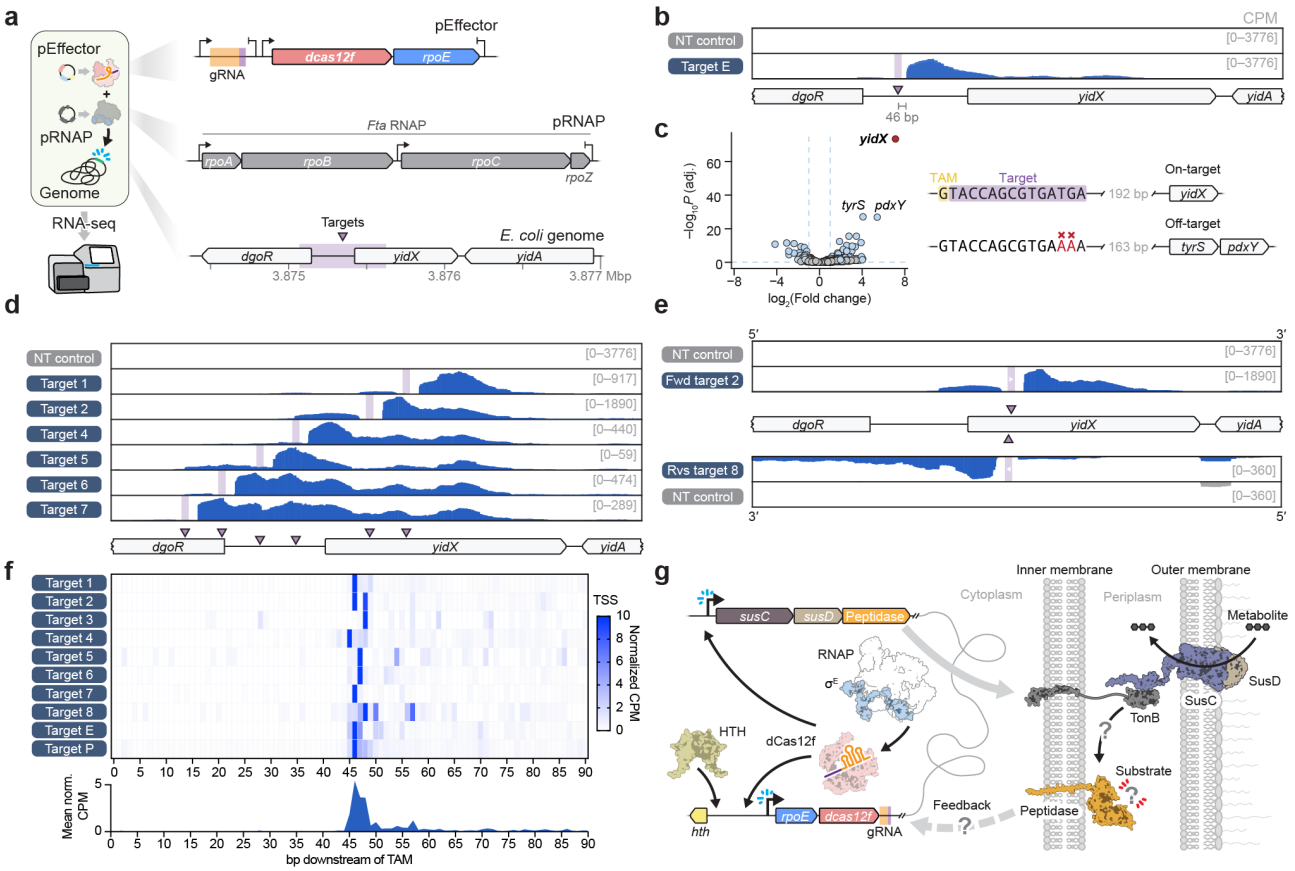

Figure 6 | dCas12f directs programmable, RNA-guided transcription with single-base pair resolution. a, Schematic of RNA-guided transcription assay (left) in which protein and RNA components encoded by pEffector and pRNAP drive RNA synthesis from gRNA-matching target loci, as measured by RNA-seq. The architecture of each vector is shown in the expanded view at right; gRNAs were designed to target within an untranscribed region (purple box/triangle) near *dgoR* and *yidX* genes. Genomic coordinates correspond to *E. coli* K12 substrain MG1655 (NCBI: NC_000913.1) **b,** RNA-seq coverage at the *yidX* locus for three biological replicates targeting the indicated region (purple triangle), with non-targeting (NT) controls shown at top. Coverage is shown for the forward strand. Robust transcription initiates at a narrowly defined distance downstream of the target site. **c,** Differential gene expression analysis for the targeting and non-targeting gRNA experiments in **b**, demonstrating that *yidX* is the most significantly up-regulated gene, transcriptome-wide. Other up-regulated genes (labeled) are encoded downstream of an identifiable off-target site. Wald test P values adjusted for multiple comparisons were calculated using the Benjamini–Hochberg approach. **d**, RNA-seq coverage at the *dgoR-yidX* locus with a panel of tiled gRNAs targeting sites 1–7 (shaded purple rectangles, and purple triangles at bottom), with non-targeting (NT) control shown at top; the genomic neighborhood is shown below the graph. Coverage is shown for the forward strand. Transcription start sites can be seamlessly relocated using distinct gRNAs, while also leading to noticeable RNA accumulation immediately upstream of the expected site of dCas12f R-loop formation. **e,** RNA-seq coverage for two gRNAs targeting opposite strands within the *yidX* gene body (shaded purple rectangles and purple triangles), alongside matched non-targeting controls. These results demonstrate that dCas12f can drive transcription in either polarity, depending only on the orientation of the gRNA. Coverage is shown for the forward strand (top), depicting the same data as in **Fig. 6d** Target 2, and reverse strand (bottom). **f,** Heat map showing the distance distribution between the TAM and transcription start site (TSS) for each of 10 tested gRNAs and matching target sites, as indicated on the left. The aggregated data is presented in a histogram at the bottom, demonstrating that the TSS occurs 46 bp downstream of the TAM. **g,** Model for RNA-guided transcriptional control driven by dCas12f and σ^E (*rpoE* gene). gRNAs encoded downstream of *dcas12f* often serve to both auto-regulate *rpoE-dCas12f* expression while also driving transcription of transporter proteins like SusCD, in a mechanism that involves cognate protein-protein interactions between σ^E, RNAP, and the DNA-bound dCas12f-RNA complex. This mechanism likely evolved to respond to, and regulate, metabolite import during growth in dynamically variable environmental conditions.

RNAP recruitment, we systematically analyzed TSS distributions across all of the tested gRNAs. We observed that transcription initiation consistently occurred within a narrow window 45–48 bp downstream of the TAM (**Fig. 6f**), suggesting a tight physical and structural coupling between dCas12f-mediated R-loop formation and subsequent formation of the transcription bubble, as was confirmed by the corresponding cryo-EM structure of the RNA-guided transcription initiation complex[50]. Although the predominant TSS was often sharply defined, the presence of some variability indicates that additional sequence determinants may cause genomic context-specific patterns. RNA-guided transcription also consistently produced a slight accumulation of positive-sense RNA-seq reads upstream of the DNA target site (**Fig. 6b,d,e**), which we hypothesize results from occasional loading of *Fta*RNAP onto a nearby chromosomal site in 3D that is upstream of DNA-bound dCas12f, leading to low-level transcription up until RNAP-dCas12f collisions occur.

Collectively, our results demonstrate that the dCas12f-σ^E system drives RNA-guided transcription at sites that can be easily reprogrammed, without any requirement for pre-existing promoter motifs beyond a 5'-G TAM, leading to targeted gene expression of any user-defined locus of interest.

DISCUSSION

In their foundational work on the *lac* repressor, Jacob and Monod originally theorized that transcription factors (TFs) might be RNAs that could control gene expression through base-pairing interactions[52]. Since then, however, extensive studies of gene regulatory mechanisms have established the prevailing dogma in the field, which primarily ascribes gene regulatory function to protein-based TFs[53], particularly in the context of transcription initiation and repression. Our work[18] now demonstrates that diverse gene expression pathways in bacteria — both repressive and activating — leverage guide RNAs to specify genomic targets through RNA–DNA base-pairing, rather than exclusively relying on protein-DNA recognition (**Extended Data Fig. 10**). Diverse regulatory RNAs have previously been shown to modulate gene expression post-transcriptionally by influencing mRNA secondary structure[54] or stability[55], but programmable, RNA-guided control of transcription itself represents a novel category of gene regulation (**Extended Data Fig. 10**). Our finding that dCas12f-σ^E systems drive RNA-guided transcription without any additional promoter motifs beyond RNA-DNA complementarity further defines a unique mechanism of transcription initiation, which we speculate could enable more rapid gene evolution via the *de novo* transcription of previously non-coding regions of the bacterial genome.

Based on our analyses thus far, we demonstrated that transmembrane transport operons are frequent native targets of dCas12f-σ^E (**Fig. 3d**). These *susCD*-containing loci are typically regulated by a proximally encoded ECF σ factor along with its corresponding anti-σ factor, to fine-tune iron uptake, peptide transport or carbohydrate import as part of polysaccharide utilization loci (PUL)[42,56,57]. Operonically linked nutrient-metabolizing hydrolases often reflect the specific substrate recognized by SusCD transporters[58,59]. Here, the *susCD*-adjacent gene is a peptidase (**Fig. 3d**). PULs can be further regulated by two-component systems (TCS)[60], which we also identified as common RNA-guided targets (**Fig. 3d**). We therefore hypothesize that dCas12f-σ^E systems emerged to regulate peptide, carbohydrate or iron acqui-

sition pathways, in which metabolite sensing triggers expression of dCas12f-σ^E loci, thereby leading to upregulation of genes involved in mediating substrate uptake (**Fig. 6g**). Future efforts will be necessary to identify the specific metabolite(s) affected by dCas12f-σ^E regulation, including experiments dedicated to determining the physiological conditions that induce their native expression, as well as how the HTH protein and its associated ncRNA are involved. Addressing these biological questions remains a challenge since Bacteroidetes genomes harbor dozens of PULs, each responsible for the import and utilization of a distinct glycan[61]. Intriguingly, we also predicted frequent examples of dCas12f-σ^E systems that likely stimulate their own expression (**Fig. 3d**), similarly to known autoregulatory mechanisms for non-dCas12f ECF σ factors[62], suggesting a positive feedback loop for amplifying transcription signals.

A fundamental question persists: what advantages do dCas12f-σ^E systems provide over the existing set of specialized σ factor groups, beyond introducing regulatory flexibility? We hypothesize that their promoter-independent behavior would allow a small set of target genes to possess their own privatized σ^E, without risking low-level promoter cross-recognition by other σ factors — thus ultimately minimizing basal transcription. This property may be physiologically advantageous by tightly regulating energetically costly pathways like transmembrane transport that requires dedicated channel proteins and metabolic enzymes only needed in the presence of select metabolites[63]. Controlling transcription via an RNA-guided mechanism would also facilitate seamless multiplexing and rapid rates of target diversification, given the relative ease with which specificity can be altered through co-varying base-pair mutations.

It is remarkable that TnpB — a transposon-encoded RNA-guided nuclease that was itself exapted during the emergence of RNA-guided transcriptional repressors[18] — became domesticated by host adaptive immune systems in the form of CRISPR-Cas12f[6,17], before being exapted yet again during the emergence of RNA-guided transcriptional activators (**Fig. 1b**). In the course of this evolutionary trajectory, the transposon-encoded single-guide ωRNA (TnpB)[10,46] presumably split into a dual-guide CRISPR RNA and tracrRNA substrate (Cas12f)[44], before collapsing again into the single chimeric gRNAs used by dCas12f. Further evidence for this sequential transformation is featured in the characteristic dimeric architecture[40,64] of dCas12f bound to the σ^E-RNAP holoenzyme within the context of a full R-loop transcription bubble, reported in an accompanying article[50]. This cryo-EM structure elegantly corroborates our observation of highly focused transcription initiation occurring 45–48 bp downstream of the TAM recognized by dCas12f, while further revealing the unique interactions mediated between dCas12f and the extended CTD present in dCas12f-associated σ^E homologs (**Fig. 1f,g**). We identified a narrow clade of related homologs that unexpectedly shares this domain but is instead associated with restriction endonuclease (RE)-like genes (**Extended Data Fig. 1c**), suggesting that transcription initiation pathways may have co-opted inactivated nucleases from both adaptive and innate immune systems to provide novel modes of DNA specificity.

Numerous CRISPR-based technologies exist to programmably inhibit or activate gene expression, respectively referred to as CRISPRi and CRISPRa[4,5,25]. Strategies developed in mammalian cells rely on fusions of dCas9 to potent repressor or activator domains like KRAB[26] or VP64[5], which function by either adding repressive epigenetic marks (CRISPRi) or remodelling chromatin to facilitate RNA polymerase recruitment[4,5]. In bacteria, simple promoter occlusion by dCas9-RNA complexes functions robustly for

CRISPRi, but CRISPRa methods have been hampered by intrinsic constraints in the ability to selectively recruit RNAP to pre-existing promoter motifs[29], which limit applications in both well-characterized model organisms and challenging biological contexts where promoter architectures are unknown. By contrast, the dCas12f-σ^E systems described here circumvent these restrictions by directly coupling RNA-guided DNA recognition to *de novo* transcription initiation, independent of promoters or host transcriptional machinery. We envision that this promoter-agnostic mechanism could enable precise and flexible control of gene expression from virtually any genomic locus of interest that meets minimal TAM requirements, offering new opportunities to activate silent gene clusters, explore functional genomics in non-model bacteria, and rewire regulatory networks with unprecedented resolution. Moreover, the ability to sensitively define custom transcription start sites may further expand the utility of this system for synthetic biology and metabolic engineering applications.

More broadly, our work adds to a growing body of evidence that bacterial RNA-guided proteins — once thought to function primarily in DNA cleavage during antiviral defense — have been repeatedly exapted to serve diverse roles in cell biology. From transposon proliferation to adaptive immunity, and now to transcriptional repression and activation, these systems illustrate the remarkable evolutionary plasticity of RNA-guided mechanisms and highlight their untapped potential for genome control.

METHODS

Candidate system choice and phylogenetic analyses of σ^E-dCas12f systems. An initial set of σ^E-dCas12f systems were identified by via a BLASTp[65] search (-evalue 1e-50 -qcov_hsp_perc 80 -max_hsps 50) of the NR database with a dCas12f homolog (WP_076354695.1) observed during manual analyses of Cas12 loci. The resulting 499 Cas12f-like proteins were then aligned to the structurally characterized *Cn*Cas12f homolog (WP_120361969.1, PDB: 8HR5), and the dCas12f homolog used to query the NR database, confirming that all 499 homologs from the initial search were predicted to have mutations in their RuvC nuclease active sites. These sequences were then used as queries in a subsequent BLASTp[65] search of the NR database, which identified 161 additional homologs. This joint dataset was de-duplicated, resulting in 660 unique dCas12f homologs, which were merged with σ^E-associated dCas12f homologs identified in a recent bioinformatic survey of TnpB/Cas12 proteins[17]. Homologs encoded in a genomic contig that precluded extraction of a dCas12f locus at least 30 kbp in length were then removed from this dataset of 869 unique dCas12f proteins. The final dataset of 707 unique dCas12f proteins were aligned with MAFFT[66] (LINSI option) and a phylogenetic tree was built with FastTree[67] (-wag -gamma). Sixteen dCas12f homologs were then chosen to sample the phylogenetic diversity of these systems.

To construct a phylogenetic tree of dCas12f proteins and their most closely related nuclease-active relatives, known σ^E-associated dCas12f proteins were plotted on a phylogenetic tree from a recently published bioinformatic survey of Cas12/TnpB[17] proteins, and a clade consisting of 99 tips was extracted from this tree. This subtree was made up of Cas12f and dCas12f representative proteins that had been clustered at 50% amino acid identity. To expand this tree, we fetched 1,487 homologs represented by the branches of this tree, aligned them with MAFFT[66] (LINSI option), and built a new tree with FastTree[67] (-wag -gamma). RuvC active site intactness and domain composition were annotated for each homolog via comparison to a homolog in the tree that is nearly identical (99.8% aa identity) to the structurally characterized *Un*Cas12f homolog (PDB: 7L48). CRISPR associations were extracted from the published metadata accompanying the previous dCas12f bioinformatic survey by Altae-Tran *et al*.[17] Associations for *rpoE* and HTH genes were annotated via BLASTp[65] searches of ORFs occurring upstream of dCas12f, queried with σ^E and HTH protein sequences from the 16 systems selected for experimental characterization.

To obtain a set of σ^E sequences for phylogenetic analysis, we conducted a BLASTp[65] search of a local copy of the NCBI non-redundant protein database (-evalue 0.01 -max_target_seqs 10000), queried with the *Fta* σ^E sequence (RIV51358.1). The resulting 2,709 sequences were then clustered with the easy-linclust option in MMseqs[68] at a threshold of 50% amino acid identity and 80% alignment coverage. Next, cluster representatives (n = 872) were aligned with MAFFT[66] (LINSI option), the alignment was trimmed of columns composed of >50% gaps with Trimal[69], and a phylogenetic tree was built from the trimmed MSA with FastTree[67] (-wag -gamma). This tree was used to guide the selection of 247 cluster representatives that formed a monophyletic clade with dCas12f-associated σ^Es. The 823 members of clusters these sequences represented were then retrieved, aligned (MAFFT[66] – LINSI), and used to build a tree (FastTree[67] -wag -gamma). Genomic assemblies encoding each *rpoE* homolog were fetched with

the rentrez[70] package in R, and *rpoE* loci (comprising 20 kbp of genomic sequence) were annotated with Eggnog[71] mapper to assess the genetic context of each homolog. dCas12f associations were additionally verified using a custom profile hidden Markov model built from dCas12f homologs identified in the analyses described above.

Cloning and strain generation. All proteins, strains and plasmids used in this study are described in **Supplementary Tables 4-6**. Initially, genes encoding the σ^E, dCas12f, and HTH protein were synthesized by Genscript and cloned into pCDFDuet-1 vectors (pEffector) along with intergenic (non-coding) regions flanking dCas12f and σ^E-coding genes using PfoI and Bsu36I restriction sites (**Fig. 2a**). Three systems encoded two distinct HTH proteins. For those, both HTH-coding genes were cloned onto the same plasmid. A 3×Flag epitope tag was appended to σ^E, the N-terminus of dCas12f, or the C-terminus of HTH. Expression of these proteins was driven by a constitutive J23105 promoter. Transcription of non-coding regions was driven by a separate constitutive J23119 promoter. The *F. taeanensis* (GCA_003584105.1) RNA polymerase subunits encoded by *rpoA*, *rpoB*, *rpoC* and *rpoZ* were cloned onto a pACYCDuet vector (p*Fta*RNAP) driven by two T7 promoters (**Fig. 5a**). Codon optimization was performed for all genes. Subsequent cloning used a variety of methods, including inverse (around-the-horn) PCR, Gibson assembly, restriction-digestion ligation, and ligation of hybridized oligonucleotides. Q5 DNA polymerase was used for all PCR steps. Plasmids were cloned, grown in NEB Turbo cells, purified with Qiagen miniprep kits, and then verified by Sanger Sequencing (Genewiz).

RNA immunoprecipitation and sequencing (RIP). RIP-seq sample preparation was generally performed as described in our previous studies[18,72] with differences in culturing. Individual *E. coli* K-12 MG1655 (sSL0810) colonies harboring the pEffector plasmid, with either dCas12f or HTH carrying a 3×Flag tag, were inoculated into 25 ml of LB media supplemented with spectinomycin (100 $\mu g\,ml^{-1}$). Liquid cultures were grown at 37 °C for approximately 16 h. A volume equivalent to 20 ml of $OD_{600\,nm} = 0.5$ was aliquoted. Cells were harvested by centrifugation at 4,000 g for 5 min, followed by a wash with 1 ml of cold TBS buffer. The supernatant was removed, and the pellet was flash-frozen in liquid nitrogen and stored at -80°C.

For immunoprecipitation, monoclonal anti-Flag M2 antibodies produced in mouse (Sigma-Aldrich, F1804) were conjugated to Dynabeads Protein G (Thermo Fisher Scientific) magnetic beads. For each sample, 60 µl of magnetic beads were washed three times in IP lysis buffer (20 mM Tris-HCl, pH 7.5 at 25°C, 150 mM KCl, 1 mM $MgCl_2$, 0.2% Triton X-100) and resuspended in 1 ml IP lysis buffer. Then, 20 µl anti-Flag antibody was added and incubated under rotation for >3 h at 4 °C. Antibody-bead-mixtures were washed two times and resuspended in 60 µl × n samples of IP lysis buffer.

Cell pellets stored at -80°C were thawed on ice and resuspended in 1.2 ml IP lysis buffer supplemented with 1× cOmplete Protease Inhibitor Cocktail (Roche) and 1 µl SUPERase•In RNase Inhibitor (Thermo Fisher Scientific). Cells were lysed by sonication using a 1/8″ sonicator probe with the following settings: 1.5 min total sonication time (2 s on, 5s off), 20% amplitude. Lysates were centrifuged at 21,000 g for 15 min at 4 °C and the supernatant was transferred to a new tube. 10 µl of samples were set aside as

non-immunoprecipitated input controls and stored at -80 °C. 60 µl of conjugated antibody-beads slurry was added to the remainder of the supernatant and rotated overnight at 4 °C. After incubation, samples were washed three times with 1 ml of cold IP wash buffer (20 mM Tris-HCl, pH 7.5 at 25°C, 150 mM KCl, 1 mM $MgCl_2$) on a magnetic rack. To elute immunoprecipitated RNA, magnetic beads were resuspended in 1 ml TRIzol (Thermo Fisher Scientific) and incubated at RT for 5 min. The supernatant containing eluted RNA was transferred to a new tube and 200 µl of chloroform was added, followed by vigorous mixing. The samples were incubated at RT for 3 min and centrifuged for 15 min at 12,000 g at 4 °C. Then, the RNA Clean & Concentrator-5 kit (Zymo Research) was used to isolate RNA from the upper aqueous layer. RNA was eluted in 15 µl nuclease-free water. RNA from input samples was isolated alongside the RIP samples using TRIzol followed by column purification. Purified RNA was stored at -80 °C prior to next-generation sequencing library preparation.

For RIP-seq library preparation, RNA from input and RIP samples was fragmented by random hydrolysis of 7 µl RNA, 6 µl water, and 2 µl NEBuffer 2 at 92 °C for 2 min. Samples were then treated with 2 µl TURBO DNase (Thermo Fisher Scientific) and 2 µl RppH (NEB) in the presence of 1 µl SUPERase•In RNase Inhibitor for 30 min at 37°C. Next, samples were treated with 1 µl T4 PNK (NEB) in 1× T4 DNA ligase buffer (NEB) for 30 min at 37°C. RNA was column purified using the Zymo RNA Clean and Concentrator-5 kit and eluted in 10.5 µl nuclease-free water. RNA concentrations were determined using the DeNovix RNA Assay Kit. Sequencing libraries were prepared using the NEBNext Small RNA Library Prep Kit. Oligonucleotide and barcode primer sequences used for library generation are listed in **Supplementary Table 7**. Illumina libraries were sequenced in 75×75 and 150×150 paired-end mode on the Illumina NextSeq 500 and Element Biosciences Aviti Cloudbreak Freestyle platforms.

RIP-seq analyses. Sequencing analyses were generally performed as described previously[72]. Raw and processed sequencing files are described in **Supplementary Table 8**. RIP-seq and input datasets were processed using cutadapt[73] (v4.2) to trim Illumina adapter sequences, remove low-quality ends, and eliminate reads shorter than 15 bp. Read mapping was performed using a combined reference genome file containing the *E. coli* K-12 MG1655 genome (NC_000913.3) and relevant plasmid sequences using bwa-mem2[74] (v2.2.1) with default parameters. Alignments were sorted and indexed using SAMtools[75] (v1.17). deepTools2[76] (v3.5.4) bamCoverage was used to generate counts per million (CPM)-normalized coverage tracks with a bin size of 1, with combined top and bottom strand alignments, scaled according to sequencing depth. RIP-seq coverage tracks were visualized in IGV[77] (v2.14.1). Transcriptome-wide analyses of RIP-seq enrichment at annotated features was performed using featureCounts[78] (v2.0.2) with -s 1 for strandedness, and further processed using DESeq2[79] to calculate fold-change and FDR (using the Benjamini-Hochberg approach) between input and RIP sample for each annotated feature. Visualizations were generated using ggplot2.

Chromatin immunoprecipitation (ChIP). Sample generation for ChIP was generally performed as previously described[80], with changes in culturing and sonication conditions. Individual *E. coli* K-12 MG1655

(sSL0810) colonies harboring the pEffector plasmid, with either dCas12f, σ^E or HTH carrying a 3×Flag tag, were inoculated into 25 ml of LB media supplemented with spectinomycin (100 μg ml^{-1}). For one sample, p*Fta*RNAP was co-transformed with pEffector (**Extended Data Fig. 8e**) and LB media was supplemented with additional chloramphenicol (25 μg ml^{-1}). Liquid cultures were grown at 37 °C for approximately 16 h, before fresh LB was added to a final volume of 40 ml. Cells were fixed promptly by addition of 1 ml of 37% (v/v) formaldehyde (Thermo Fisher Scientific) followed by vigorous vortexing and crosslinking continued for 20 min at RT while gently shaking. Crosslinking was quenched by addition of 4.6 ml of 2.5 M glycine, followed by 10 min incubation with gentle shaking. To pellet cells, samples were centrifuged at 4,000 g for 8 min at 4 °C. The next steps were performed on ice. The supernatant was discarded and pellets were resuspended in 40 ml TBS buffer (20 mM Tris-HCl pH 7.5, 0.15 M NaCl). The TBS repeat was repeated. Cells equivalent to 40 ml of $OD_{600\ nm}$ = 0.6 were aliquoted into a new tube. Cells were pelleted again by centrifugation and transferred into an Eppendorf tube, flash-frozen in liquid nitrogen and stored at -80 °C or kept on ice prior to the subsequent steps.

Immunoprecipitation was performed as follows: bovine serum albumin (GoldBio) was dissolved in PBS buffer (Gibco) for a final concentration of 5 mg/ml BSA. 25 µl of Dynabeads Protein G (Thermo Fisher) slurry (hereafter referred to as 'beads' or 'magnetic beads') were prepared for each sample. At the maximum, 250 µl of the initial slurry were processed in a single tube. All bead washes were performed at RT using a magnetic rack to retain the beads. First the initial supernatant was removed. Then, beads were washed four times using 1 ml of BSA solution. Beads were resuspended in 25 µl × n samples of BSA solution, followed by addition of 4 µl × n samples of monoclonal anti-Flag M2 antibody produced in mouse (Sigma-Aldrich, F1804). Antibody and magnetic beads were conjugated for >3 h at 4 °C while rotating. In the meantime, crosslinked cell pellets were thawed on ice and resuspended in FA lysis buffer 150 (50 mM HEPES-KOH pH 7.5, 0.1% (w/v) sodium deoxycholate, 0.1% (w/v) SDS, 1 mM EDTA, 1% (v/v) Triton X-100, 150 mM NaCl), supplemented with protease inhibitor cocktail (Sigma-Aldrich) and transferred to a 1 ml milliTUBE AFA Fiber (Covaris). Next, samples were sonicated on a LE220 Focused-ultrasonicator (Covaris) using a 24 place milliTUBE rack (Covaris) with the following SonoLab 7.3 settings: min. temp. 4 °C, set point 6 °C, max. temp. 8 °C, peak power 420, duty factor 30, cycles/bursts 200, 17.5 min sonication time. Sonicated samples were centrifuged at 20,000 g and 4 °C for 20 min. The supernatant was transferred into a new tube and kept on ice. 10 µl of the supernatant was then set aside as a non-immunoprecipitated input control, flash-frozen in liquid nitrogen, and stored at -80 °C. After >3 h of conjugating antibodies to magnetic beads, the slurry was washed four times with 1 ml of BSA solution, as described earlier. Then, magnetic beads were resuspended in 30 µl × n samples of FA lysis buffer 150 with protease inhibitor, and 31 µl of resuspended antibody-conjugated beads were added to each sonicated cell lysate sample, followed by overnight incubation at 4 °C while rotating to immunoprecipitate the Flag-tagged proteins. Then next day, beads were washed using 1 ml of buffers, as follows: two washes with FA lysis buffer 150, one wash with FA lysis buffer 500 (50 mM HEPES-KOH pH 7.5, 0.1% (w/v) sodium deoxycholate, 0.1% (w/v) SDS, 1 mM EDTA, 1% (v/v) Triton X-100, 500 mM NaCl), one wash with ChIP wash buffer (10 mM Tris-HCl pH 8.0, 250 mM LiCl, 0.5% (w/v) sodium deoxycholate, 0.1% (w/v) SDS, 1 mM

EDTA, 1% (v/v) Triton X-100, 500 mM NaCl), and two washes with TE Buffer 10/1 (10 mM Tris-HCl pH 8.0, 1 mM EDTA). Finally, immunoprecipitated protein-DNA complexes were eluted by addition 200 µl of freshly made ChIP elution buffer (1% (w/v) SDS, 0.1 M $NaHCO_3$) and incubation for 1.25 h at 65 °C with vortexing every 15 min. In the meantime, input samples were thawed and 190 µl of ChIP elution buffer was added, followed by the addition of 10 µl of 5 M NaCl. After the 1.25 h incubation, magnetic beads were discarded and the supernatant was transferred to a new tube, and 9.75 µl of 5 M NaCl were added. ChIP and input samples were incubated at 65°C overnight to de-crosslink protein-DNA complexes. Then, RNA was degraded by addition of 1 µl of 10 mg/ml RNase A (Thermo Fisher) and incubation for 1 h at 37 °C, followed by protein degradation by addition of 2.8 µl of 20 mg/ml Proteinase K (Fisher Scientific) and 1 h incubation at 55 °C. 1 ml of buffer PB (QIAGEN) was added to all ChIP and input samples, followed by column purification and elution in 40 ul TE Buffer 10/0.1 (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA).

For ChIP-qPCR, pairs of genomic qPCR primers (**Supplementary Table 7**) were designed for the *yidX* target site (tSL0679) and a genomic reference locus (*rssA*). 1 µl of ChIP or input sample was diluted with 14 µl of TE buffer 10/0.1. Parallel qPCR reactions were performed using both genomic target and reference locus primer pairs and the fold change between ChIP and input samples was calculated using the $\Delta\Delta C_q$ method: $A = C_q$(immunoprecipitated sample at target locus) $- C_q$(immunoprecipitated sample at reference locus); $B = C_q$(input sample at target locus) $- C_q$(input sample at reference locus); $\Delta\Delta C_q = 2^{-(A-B)}$.

For ChIP-seq, libraries were generated using the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB). Initial DNA concentrations were quantified using the DeNovix dsDNA Ultra High Sensitivity Kit. DNA amounts were standardized across ChIP and input samples prior to library preparation. Adapters were ligated and 12 to 15 cycles of PCR amplification were performed to add Illumina barcodes (**Supplementary Table 7**). Then, two-sided size selection using AMPure XP beads (Beckman Coulter) was performed to purify DNA fragments of around 450 bp length, as described previously[80]. For sample pooling, final DNA concentrations were quantified using the DeNovix dsDNA High Sensitivity Kit. Illumina libraries were sequenced in 75×75 and 150×150 paired-end mode on the Illumina NextSeq 500 and Element Biosciences Aviti Cloudbreak Freestyle platforms.

ChIP-seq analyses. Sequencing analyses were performed as described previously[80]. Raw and processed sequencing files are described in **Supplementary Table 8**. ChIP-seq reads were trimmed using fastp[81] (v0.23.2) and mapped to a custom *E. coli* K-12 MG1655 reference genome using Bowtie2[82] (v2.2.5) with default parameters. A genomic *lacZ*/*lacI* region and flanking sequences partially identical to plasmid-encoded sequences were masked in all alignments (genomic coordinates 366,386-367,588). Mapped read sorting and indexing was performed using SAMtools[75] (v1.6). Read coverage was normalized by CPM (Counts Per Million) and binned in 1-bp windows using deepTools2[76] (v3.5.4) bamCoverage. Normalized reads were visualized in IGV (v2.14.1)[77]. Genome-wide views showing maximum enrichment values per 1-kbp window, as described previously[83], were generated using bedtools[84] (v2.26.0) and the 'bedtools makewindows -w 1000 command'. Motif analysis was performed as described in our previous study[80]. ChIP-seq peaks were called on read alignments prior to normalization using MACS3[85] (v3.0.2), relative

to non-immunoprecipitated input control samples. Peak summit coordinates were extracted together with 100 bp sequences flanking on either side. MEME ChIP[86] (v5.5.7) using default parameters was used to determine protein-bound sequence motifs.

Whole genome sequencing (WGS). One *M. rigui* and five *F. taeanensis* strains were obtained from the China General Microbiological Culture Collection Center (CGMCC) and the Marine Culture Collection of China (MCCC), respectively (**Supplementary Table 5**). Freeze-dried *M. rigui* cells were resuspended in liquid PCA media (5 g tryptone/l, 2.5 g yeast extract/l, 1 g glucose/l) and grown in 5 ml of PCA media at 25 °C for 4 days without antibiotic selection. A volume equivalent to 10 ml of $OD_{600\,nm}$ = 1.0 was aliquoted for harvesting. Cells were pelleted by centrifugation at 4,000 g for 5 min, followed by a wash with 1 ml of PBS buffer and another 5 min centrifugation step. Cell pellets were resuspended in 0.5 ml of 1X DNA/RNA shield buffer (Zymo Research). WGS and gene annotation was performed by Plasmidsaurus.

Freeze-dried *F. taeanensis* strains were resuspended in Difco Marine Broth 2216 and streaked out onto Difco Marine Broth 2216 agar plates. Agar plates were incubated at 28 °C for 3 days. Next, 20 ml of liquid Marine Broth was inoculated from a single colony and grown for another 3 days at 28 °C without antibiotic selection. Further sample preparation was performed as described for *M. rigui*.

Total RNA sequencing (RNA-seq). Culturing of *F. taeanensis* was performed as described for WGS. For *E. coli* RNA-seq, BL21(DE3) cells with a genomically encoded *sfGFP* harboring a pTarget plasmid and the *Fta*RNAP plasmid, an empty vector or no plasmid, were transformed with a pEffector plasmid encoding the native *Fta* gRNA or gRNAs with re-programmed guide portions to target sites in the *E. coli* genome. *E. coli* cells were grown in 25 ml of LB media supplemented with antibiotics (100 $\mu g\,ml^{-1}$ spectinomycin; 100 $\mu g\,ml^{-1}$ carbenicillin; 25 $\mu g\,ml^{-1}$ chloramphenicol) and 0.01 mM IPTG for *Fta*RNAP induction. After incubating for 16-20 h at 28 °C, cells equivalent to 2 ml of $OD_{600\,nm}$ = 0.5 were harvested by centrifugation at 4,000 g and 4 °C for 5 min. Cell pellets were washed with 1 ml of cold TBS buffer. The supernatant was removed, and the pellet was flash-frozen in liquid nitrogen and stored at -80°C.

Total RNA-seq sample preparation was performed as described previosuly[72]. Cell pellets stored at -80°C were thawed on ice, resuspended in 1 ml TRIzol (Thermo Fisher Scientific), and incubated at RT for 5 min. Then, 200 µl of chloroform was added, followed by vigorous mixing. The samples were incubated at RT for 3 min and centrifuged for 15 min at 12,000 g at 4 °C. Then, the RNA Clean & Concentrator-5 kit (Zymo Research) was used to isolate RNA from the upper aqueous layer. RNA was eluted in 15 µl nuclease-free water. Purified RNA was stored at -80 °C prior to next-generation sequencing library preparation.

For RNA-seq library preparation, RNA was fragmented by random hydrolysis of 7 µl RNA, 6 µl water, and 2 µl NEBuffer 2 at 92 °C for 2 min. Samples were then treated with 2 µl TURBO DNase (Thermo Fisher Scientific) and 2 µl RppH (NEB) in the presence of 1 µl SUPERase•In RNase Inhibitor for 30 min at 37°C. Next, samples were treated with 1 µl T4 PNK (NEB) in 1× T4 DNA ligase buffer (NEB) for 30 min at 37°C. RNA was column purified using the Zymo RNA Clean and Concentrator-5 kit and eluted in 10.5 µl nuclease-free water. RNA concentrations were determined using the DeNovix RNA Assay Kit.

Sequencing libraries were prepared using the NEBNext Small RNA Library Prep Kit. Illumina libraries were sequenced in 75×75 and 150×150 paired-end mode on the Illumina NextSeq 500 and Element Biosciences Aviti Cloudbreak Freestyle platforms.

RNA-seq analyses. Sequencing analyses and visualization were performed as described for RIP-seq analyses, with the exception for read mapping and normalization. Raw and processed sequencing files are described in **Supplementary Table 8**. Read mapping using bwa-mem2[74] (v2.2.1) required a seed sequence of 25 nt for read mapping (-k 25) to avoid gRNA guide sequence mapping to genomic or plasmid target sites. The following *E. coli* BL21(DE3) reference genome was used for read mapping: CP001509.3. Counts per million (CPM)-normalized coverage tracks were generated using deepTools2[76] (v3.5.4) bamCoverage with a bin size of 1, with combined top and bottom strand alignments, or separation of top and bottom strand alignments. Coverage was scaled according to sequencing depth. DESeq2[79] was used to calculate fold-change and FDR (using the Benjamini-Hochberg approach) between NT and *yidX* targeting samples for each annotated feature. Differential expression analyses shown in **Figure 6c** employed four biological replicates, and the genomic coordinates for *yidX* were extended to encompass the 5'-UTR observed in RNA-seq data shown in **Figure 6b**, 161 bp upstream of the *yidX* start codon. Normalized CPM in **Figure 6f** was calculated as follows: normalized CPM = (CPM at coordinate *n*) / ((max. CPM in 90-bp TSS window) / 10). Mean normalized CPM was calculated by taking the mean at every coordinate in the same 90-bp window around the TSS, across the ten RNA-seq samples shown in **Figure 6f**.

RT-qPCR. Purified RNA (pre-library preparation) samples for RNA-seq were used to perform RT-qPCR. The procedure was generally performed as described previously[72]. RNA was treated with 1 µl of dsDNase (Thermo Fisher Scientific) in 1× dsDNase buffer, with a final reaction volume of 10 µl. Samples were incubated at 37 °C for 2 min, followed by adding 1.1 µl of 100 mM DTT and 5 min incubation at 55 °C to stop the reaction. The iScript cDNA Synthesis Kit (BioRad) was used for reverse transcription. cDNA samples were stored at -20 °C. For qPCR, cDNA samples were diluted 10-fold in nuclease-free water. qPCR was performed using 10 µl reactions containing the following components: 5 µl of SsoAdvanced Universal SYBR Green Supermix (BioRad), 2 µl of 2.5 µM primer pair, 1 µl of nuclease-free water, and 2 µl of 10-fold diluted cDNA. Primer pairs were designed to span the intergenic sequence upstream of the *E. coli yidX* gene (**Supplementary Table 7**). Primers designed for the genomic *rssA* gene served as a reference pair for normalization (**Supplementary Table 7**). Reactions were set up in 384-well PCR plates (BioRad). Measurements were obtained using a CFX384 RealTime PCR Detection System (BioRad). The abundance of RNA expressed from the *yidX* locus was calculated relative to the *rssA* locus using the $\Delta\Delta C_q$ method (see ChIP-qPCR).

RFP fluorescence assay. *E. coli* BL21(DE3) cells with a genomically-encoded *sfGFP* harboring a pTarget plasmid in addition to the *Fta*RNAP plasmid, an empty vector or no plasmid, were transformed with a pEffector plasmid encoding the native *Fta* gRNA or gRNAs with re-programmed guide portions to target

sites in the *E. coli* genome. pTarget was designed to encode a *mRFP1* gene downstream of 500 bp of the native dCas12f target region from the *F. taeanensis* genome (GCA_003584105.1) followed by a strong synthetic ribosome binding site (RBS). *E. coli* cultures were grown in triplicates by inoculating 5 ml of LB in 24-well plates or 0.5 ml of LB in 96-well plates with single colonies. *Fta*RNAP expression was induced with IPTG (0.01 mM). Cultures were grown at 28 °C for 20-24 h. After incubation, 200 µl of culture was transferred into a 96-well optical bottom plate (Thermo Scientific) for endpoint measurements. The $OD_{600\ nm}$ was measured in parallel with mRFP fluorescence signal using a Synergy Neo2 microplate reader (Biotek).

Flow cytometry analysis. *E. coli* BL21(DE3) cultures for flow cytometry analyses were grown as described for the RFP fluorescence assay. After 20-24 h incubation, cells equivalent to 100 µl of $OD_{600\ nm}$ = 1.0 were aliquoted into a conical bottom 96-well plate. Cells were pelleted by centrifugation at 4,000 g for 5 min and resuspended in 100 µl flow cytometry buffer (1× PBS, 1 mM EDTA pH 8.0, 1% (w/v) BSA). Then, cells were pelleted again by centrifugation and resuspended in 200 µl flow cytometry buffer. Resuspended cells were then transferred into a round bottom 96-well plate. Flow cytometry analysis was performed using a Novocyte Penteon instrument. Bacteria were gated on FSC-A and FSC-H first to remove debris, and then on FSC-A and SSC-A to analyze singlets (**Supplementary Fig. 1**). Gating was optimized using GFP-only, RFP-only, GFP+RFP, and no GFP/RFP control samples. Then, mean fluorescence intensity was taken for RFP. Flow cytometry data was analyzed using the NovoExpress software.

Covariation model building and native target identification. To build a covariance model (CM) of dCas12f gRNA scaffold sequences, scaffolds observed in RIP-seq from seven diverse dCas12f systems (*Fta*, *Lby*, *Cgl*, *Lpa*, *Mri*, *Pum*, *Smi*) were aligned with LocaRNA[87], then Infernal[88] was used to build a CM (Scaffold CM) (**Extended Data Fig. 6a**). A separate CM was also built from sequences similar to a ncRNA observed in RNA-sequencing data from sSL4759 (*Fta* strain), which encodes a locus that is closely related to the *Fta* dCas12f locus (95.6% nucleotide identity). The ncRNA sequence from sSL4759 was queried against dCas12f loci, resulting in the identification of 65 similar sequences. These hits were used to build an HMM of ncRNA sequences[89], which when searched against the same dCas12f resulted in the identification of 140 ncRNA-like sequences. Finally, these sequences were aligned with LocaRNA, and a covariance model (ncRNA-like CM) was built with Infernal (**Extended Data Fig. 6a**).

To identify dCas12f gRNA sequences and their expected targets (**Extended Data Fig. 6b**), a subset of 106 dCas12f encoding genomes were downloaded from NCBI. These genomes were selected because they represented "complete" or "chromosome" level assemblies, and each genome was re-annotated with BAKTA[90]. The Scaffold CM was then used to identify gRNA scaffold sequences, and 14 nucleotide sequences abutting the 3'-end of these scaffolds were extracted as putative gRNA guide sequences, consistent with the length of the *Fta* guide observed in RIP-seq. BLASTn[65] (-evalue 100 -word_size 7 -reward 1 -penalty -3 -gapopen 5 -gapextend 2) was then used to match each guide to potential targets in its originating genome. The pool of potential targets was filtered to remove sequences with one or more

mismatches over the theoretical duplex formed by six nucleotides of TAM-proximal gRNA guide and the DNA target sequence (i.e., a seed sequence). Potential targets were also filtered to remove BLASTn hits that overlapped with dCas12f guide coordinates, that did not occur in an intergenic region, or that were upstream of a gene improperly oriented — relative to the dCas12f target — to enable sense-stranded transcription of the gene. The genomic contexts of the remaining 626 potential dCas12f targets were then analyzed by extracting 500 nucleotides of sequence 5' of the putative target, and 10 kilobases of sequence 3' of the target. The ncRNA-like CM was used to annotate putative ncRNAs in these loci, and loci were manually scored for similarity, based on the identity of genes immediately downstream of the putative dCas12f gRNA-target duplex.

***In silico* prediction of HTH binding sites.** To identify putative HTH binding sites, intergenic regions comprising less than 1 kbp between the HTH-encoding gene and ncRNAs identified with the ncRNA-like CM described above were extracted from dCas12f loci. An initial nHMMer[89] search of these 349 loci with the sequence of the *Fta* HTH site (identified via ChIP-seq) did not produce any significant hits outside of the *Fta* locus and genetically identical loci. Therefore, we used the palindrome tool from the EMBOSS[91] suite to identify inverted repeats in each intergenic region (-minpallen 7 -maxpallen 9 -gaplimit 2 -num-mismatches 1). The starting position of palindromes that were similar in length to the *Fta* HTH site were then plotted in R using the ggplot2[92] package.

Statistics analyses and figure plotting. All sequencing coverage plots were generated in IGV. Plots showing genome-wide expression and differential expression analyses were generated using custom R scripts. Transcription start site (TSS) plots and heatmaps were generated using a custom python script mapping 5' or 3' end coordinates of sequencing reads.

Data availability. Next-generation sequencing data will be made available in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (BioProject accession: PRJNA1247282) and the Gene Expression Omnibus (GSE293889) at the time of publication. The published genomes used for bioinformatics analyses were obtained from NCBI (**Supplementary Table 4**). Datasets generated and analyzed in the current study are available from the corresponding authors on reasonable request.

Code availability. Custom scripts used for bioinformatics, RIP-seq, ChIP-seq, and RNA-seq data analyses are available upon request.

ACKNOWLEDGMENTS

We thank A.J. Robinson, S. Kang., and J.L. Ramirez for laboratory support; E.A. Campbell and S.A. Darst for helpful discussion; Z. Hua for custom scripts; Z. Quan at Fudan University for providing *Fta* strains; the JP Sulzberger Columbia Genome Center for NGS support; and L.F. Landweber for qPCR and gel imager instrument access. S.T. was supported by a Ruth L. Kirchstein Individual Predoctoral Fellowship

(F30AI183830) from the NIH. L.C. was supported by NIH grant R01GM138675 and by the National Science Foundation (NSF) Faculty Early Career Development Program (CAREER) Award 2339799. S.H.S. was supported by NSF Faculty Early Career Development Program (CAREER) Award 2239685, a Pew Biomedical Scholarship, an Irma T. Hirschl Career Scientist Award, the Howard Hughes Medical Institute Investigator Program, and a generous startup package from the Columbia University Irving Medical Center Dean's Office and the Vagelos Precision Medicine Fund.

AUTHOR CONTRIBUTIONS

F.T.H. and S.H.S. conceived the project. F.T.H. designed and performed most experiments. T.W. performed most bioinformatics analyses. F.T.H., A.I.P., and J.G.-K. performed cloning and RFP activation assays. R.X. cloned RNAP expression plasmids and contributed to data interpretation. F.T.H. and S.T. analyzed RNA-seq and RIP-seq data. H.L. and C.M. performed initial phylogenetics and bioinformatics analyses. G.D.L. and F.T.H. performed flow cytometry assays. L.C. contributed to data interpretation. S.H.S. oversaw the project. F.T.H., T.W., and S.H.S. discussed the data and wrote the manuscript, with input from all authors.

COMPETING INTERESTS

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. S.H.S., F.T.H., and T.W. are inventors on patents related to CRISPR-Cas-like systems and uses thereof. The other authors declare no competing interests.

Correspondence and requests for materials should be addressed to S.H.S. (shsternberg@gmail.com).

SUPPLEMENTARY TABLES

Supplementary Table 1 | Cas12f homologs in Fig. 1b phylogenetic tree.

Supplementary Table 2 | dCas12f, HTH, and σ^E homologs in Fig. 1e phylogenetic trees.

Supplementary Table 3 | σ^E homologs in Fig. 1f phylogenetic tree.

Supplementary Table 4 | List of dCas12f-σ^E systems and proteins tested in this study.

Supplementary Table 5 | Strains used in this study.

Supplementary Table 6 | Description and sequence of plasmids used in this study.

Supplementary Table 7 | Oligonucleotides used in this study.

Supplementary Table 8 | Sequencing files and information for this study.

REFERENCES

1. Feklistov, A., Sharon, B. D., Darst, S. A. & Gross, C. A. Bacterial Sigma Factors: A Historical, Structural, and Genomic Perspective. *Annu. Rev. Microbiol.* **68**, 1–20 (2014).

2. Sharma, U. K. & Chatterji, D. Transcriptional switching in Escherichia coli during stress and starvation by modulation of σ70 activity. *FEMS Microbiol. Rev.* **34**, 646–657 (2010).

3. Casas-Pastor, D. *et al.* Expansion and re-classification of the extracytoplasmic function (ECF) σ factor family. *Nucleic Acids Res.* **49**, 986–1005 (2021).

4. 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).

5. Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647–661 (2014).

6. 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).

7. Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**, 816–821 (2012).

8. Zetsche, B. *et al.* Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* **163**, 759–771 (2015).

9. Shmakov, S. *et al.* Diversity and evolution of class 2 CRISPR–Cas systems. *Nat. Rev. Microbiol.* **15**, 169–182 (2017).

10. Altae-Tran, H. *et al.* The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases. *Science* **374**, 57–65 (2021).

11. Sasnauskas, G. *et al.* TnpB structure reveals minimal functional core of Cas12 nuclease family. *Nature* **616**, 384–389 (2023).

12. Meers, C. *et al.* Transposon-encoded nucleases use guide RNAs to promote their selfish spread. *Nature* **622**, 863–871 (2023).

13. Durrant, M. G. *et al.* Bridge RNAs direct programmable recombination of target and donor DNA. *Nature* **630**, 984–993 (2024).

14. Siddiquee, R., Pong, C. H., Hall, R. M. & Ataide, S. F. A programmable seekRNA guides target selection by IS1111 and IS110 type insertion sequences. *Nat. Commun.* **15**, 5235 (2024).

15. Vaysset, H., Meers, C., Cury, J., Bernheim, A. & Sternberg, S. H. Evolutionary origins of archaeal and eukaryotic RNA-guided RNA modification in bacterial IS110 transposons. *Nat. Microbiol.* **10**, 20–27 (2025).

16. Saito, M. *et al.* Fanzor is a eukaryotic programmable RNA-guided endonuclease. *Nature* **620**, 660–668 (2023).

17. Altae-Tran, H. *et al.* Diversity, evolution, and classification of the RNA-guided nucleases TnpB and Cas12. *Proc. Natl. Acad. Sci.* **120**, e2308224120 (2023).

18. Wiegand, T. *et al.* TnpB homologues exapted from transposons are RNA-guided transcription factors. *Nature* **631**, 439–448 (2024).

19. Workman, R. E. *et al.* A natural single-guide RNA repurposes Cas9 to autoregulate CRISPR-Cas expression. *Cell* **184**, 675-688.e19 (2021).

20. Sampson, T. R., Saroj, S. D., Llewellyn, A. C., Tzeng, Y.-L. & Weiss, D. S. A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. *Nature* **497**, 254–257 (2013).

21. Ratner, H. K. *et al.* Catalytically Active Cas9 Mediates Transcriptional Interference to Facilitate Bacterial Virulence. *Mol. Cell* **75**, 498-510.e5 (2019).

22. Wu, W. Y. *et al.* The miniature CRISPR-Cas12m effector binds DNA to block transcription. *Mol. Cell* **82**, 4487-4502.e7 (2022).

23. 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).

24. Li, M. *et al.* Toxin-antitoxin RNA pairs safeguard CRISPR-Cas systems. *Science* **372**, (2021).

25. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* **152**, 1173–1183 (2013).

26. Gilbert, L. A. *et al.* CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* **154**, 442–451 (2013).

27. Zhang, X. *et al.* Multiplex gene regulation by CRISPR-ddCpf1. *Cell Discov.* **3**, 17018 (2017).

28. Zalatan, J. G. *et al.* Engineering Complex Synthetic Transcriptional Programs with CRISPR RNA Scaffolds. *Cell* **160**, 339–350 (2015).

29. Fontana, J. *et al.* Effective CRISPRa-mediated control of gene expression in bacteria must overcome strict target site requirements. *Nat. Commun.* **11**, 1618 (2020).

30. Burgess, R. R., Travers, A. A., Dunn, J. J. & Bautz, E. K. F. Factor Stimulating Transcription by RNA Polymerase. *Nature* **221**, 43–46 (1969).

31. Ross, W. *et al.* A Third Recognition Element in Bacterial Promoters: DNA Binding by the α Subunit of RNA Polymerase. *Science* **262**, 1407–1413 (1993).

32. Ishihama, A. Functional modulation of Escherichia coli RNA polymerase. *Annu. Rev. Microbiol.* **54**, 499–518 (2000).

33. Paget, M. S. Bacterial Sigma Factors and Anti-Sigma Factors: Structure, Function and Distribution. *Biomolecules* **5**, 1245–1265 (2015).

34. Erickson, J. W. & Gross, C. A. Identification of the sigma E subunit of Escherichia coli RNA polymerase: a second alternate sigma factor involved in high-temperature gene expression. *Genes Dev.* **3**, 1462–1471 (1989).

35. Mecsas, J., Rouviere, P. E., Erickson, J. W., Donohue, T. J. & Gross, C. A. The activity of sigma E, an Escherichia coli heat-inducible sigma-factor, is modulated by expression of outer membrane proteins. *Genes Dev.* **7**, 2618–2628 (1993).

36. Hove, B. V., Staudenmaier, H. & Braun, V. Novel two-component transmembrane transcription control: regulation of iron dicitrate transport in Escherichia coli K-12. *J. Bacteriol.* **172**, 6749–6758 (1990).

37. Pinto, D. & Fonseca, R. R. da. Evolution of the extracytoplasmic function σ factor protein family. *NAR Genom. Bioinform.* **2**, lqz026 (2020).

38. Martens, E. C., Roth, R., Heuser, J. E. & Gordon, J. I. Coordinate Regulation of Glycan Degradation and Polysaccharide Capsule Biosynthesis by a Prominent Human Gut Symbiont. *J. Biol. Chem.* **284**, 18445–18457 (2009).

39. Ades, S. E. Regulation by destruction: design of the σE envelope stress response. *Curr. Opin. Microbiol.* **11**, 535–540 (2008).

40. Xiao, R., Li, Z., Wang, S., Han, R. & Chang, L. Structural basis for substrate recognition and cleavage by the dimerization-dependent CRISPR–Cas12f nuclease. *Nucleic Acids Res.* **49**, 4120–4128 (2021).

41. Aravind, L., Anantharaman, V., Balaji, S., Babu, M. & Iyer, L. The many faces of the helix-turn-helix domain: Transcription regulation and beyond. *FEMS Microbiol. Rev.* **29**, 231–262 (2005).

42. Yokoyama, T. *et al.* The Escherichia coli S2P intramembrane protease RseP regulates ferric citrate uptake by cleaving the sigma factor regulator FecR. *J. Biol. Chem.* **296**, 100673 (2021).

43. Hoffmann, F. T. *et al.* Selective TnsC recruitment enhances the fidelity of RNA-guided transposition. *Nature* **609**, 384–393 (2022).

44. Harrington, L. B. *et al.* Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. *Science* **362**, 839–842 (2018).

45. Karvelis, T. *et al.* PAM recognition by miniature CRISPR–Cas12f nucleases triggers programmable double-stranded DNA target cleavage. *Nucleic Acids Res.* **48**, 5016–5023 (2020).

46. Karvelis, T. *et al.* Transposon-associated TnpB is a programmable RNA-guided DNA endonuclease. *Nature* **599**, 692–696 (2021).

47. Brooks, B. E. & Buchanan, S. K. Signaling mechanisms for activation of extracytoplasmic function (ECF) sigma factors. *Biochim. Biophys. Acta (BBA) - Biomembr.* **1778**, 1930–1945 (2008).

48. Noinaj, N., Guillier, M., Barnard, T. J. & Buchanan, S. K. TonB-Dependent Transporters: Regulation, Structure, and Function. *Microbiology* **64**, 43–60 (2010).

49. Birkholz, N. *et al.* Phage anti-CRISPR control by an RNA- and DNA-binding helix–turn–helix protein. *Nature* **631**, 670–677 (2024).

50. Xiao *et al.* Structural basis of RNA-guided transcription initiation by a dCas12f-RpoE-RNAP complex. *in preparation*.

51. Harden, T. T. *et al.* Bacterial RNA polymerase can retain σ70 throughout transcription. *Proc. Natl. Acad. Sci.* **113**, 602–607 (2016).

52. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).

53. Balleza, E. *et al.* Regulation by transcription factors in bacteria: beyond description. *FEMS Microbiol. Rev.* **33**, 133–151 (2009).

54. Bak, G., Han, K., Kim, D. & Lee, Y. Roles of rpoS activating small RNAs in pathways leading to acid resistance of Escherichia coli. *MicrobiologyOpen* **3**, 15–28 (2014).

55. Massé, E. & Gottesman, S. A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. *Proc. Natl. Acad. Sci.* **99**, 4620–4625 (2002).

56. Madej, M. *et al.* Structural and functional insights into oligopeptide acquisition by the RagAB transporter from Porphyromonas gingivalis. *Nat. Microbiol.* **5**, 1016–1025 (2020).

57. Grondin, J. M., Tamura, K., Déjean, G., Abbott, D. W. & Brumer, H. Polysaccharide Utilization Loci: Fueling Microbial Communities. *J. Bacteriol.* **199**, 10.1128/jb.00860-16 (2017).

58. Lapébie, P., Lombard, V., Drula, E., Terrapon, N. & Henrissat, B. Bacteroidetes use thousands of enzyme combinations to break down glycans. *Nat. Commun.* **10**, 2043 (2019).

59. Tong, M. *et al.* A highly conserved SusCD transporter determines the import and species-specific antagonism of Bacteroides ubiquitin homologues. *Nat. Commun.* **15**, 8794 (2024).

60. Martens, E. C., Chiang, H. C. & Gordon, J. I. Mucosal Glycan Foraging Enhances Fitness and Transmission of a Saccharolytic Human Gut Bacterial Symbiont. *Cell Host Microbe* **4**, 447–457 (2008).

61. Feng, J. *et al.* Polysaccharide utilization loci in Bacteroides determine population fitness and community-level interactions. *Cell Host Microbe* **30**, 200-215.e12 (2022).

62. Todor, H. *et al.* Rewiring the specificity of extracytoplasmic function sigma factors. *Proc. Natl. Acad. Sci.* **117**, 33496–33506 (2020).

63. Gray, D. A. *et al.* Insights into SusCD-mediated glycan import by a prominent gut symbiont. *Nat. Commun.* **12**, 44 (2021).

64. Takeda, S. N. *et al.* Structure of the miniature type V-F CRISPR-Cas effector enzyme. *Mol. Cell* **81**, 558-570.e3 (2021).

65. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).

66. 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).

67. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).

68. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

69. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–3 (2009).

70. Winter, D. J. rentrez: An R package for the NCBI eUtils API. *The R Journal* **9**, 520–526 (2017).

71. Cantalapiedra, C. P., Hernández-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).

72. Tang, S. *et al.* De novo gene synthesis by an antiviral reverse transcriptase. *Science* **386**, eadq0876 (2024).

73. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetJ.* **17**, 10–12 (2011).

74. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. in vol. 00 314–324 (IEEE Computer Society, 2019).

75. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

76. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160-5 (2016).

77. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24–26 (2011).

78. 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).

79. 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).

80. Hoffmann, F. T. *et al.* Selective TnsC recruitment enhances the fidelity of RNA-guided transposition. *Nature* **609**, 384–393 (2022).

81. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

82. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).

83. Cooper, L. A., Stringer, A. M. & Wade, J. T. Determining the Specificity of Cascade Binding, Interference, and Primed Adaptation In Vivo in the Escherichia coli Type I-E CRISPR-Cas System. *Mbio* **9**, e02100-17 (2018).

84. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–2 (2010).

85. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

86. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202-8 (2009).

87. 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**, 900–14 (2012).

88. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).

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

90. Schwengers, O. *et al.* Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb. Genom.* **7**, 000685 (2021).

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

92. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer, 2016).