Cite as: S. Tang *et al.*, *Science*
10.1126/science.adq0876 (2024).

De novo gene synthesis by an antiviral reverse transcriptase

Stephen Tang[1], Valentin Conte[1]†, Dennis J. Zhang[2]†, Rimantė Žedaveinytė[1], George D. Lampe[1], Tanner Wiegand[1], Lauren C. Tang[2], Megan Wang[1], Matt W.G. Walker[2], Jerrin Thomas George[1]‡, Luke E. Berchowitz[3,4], Marko Jovanovic[2], Samuel H. Sternberg[1]*

[1]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA. [2]Department of Biological Sciences, Columbia University, New York, NY, USA. [3]Department of Genetics and Development, Columbia University, New York, NY, USA. [4]Taub Institute for Research on Alzheimer's and the Aging Brain, New York, NY, USA.

†These authors contributed equally to this work.

‡Present address: Department of Microbiology and Cell Biology, Montana State University, Bozeman, MT, USA.

***Corresponding author. Email: shsternberg@gmail.com**

Defense-associated reverse transcriptase (DRT) systems perform DNA synthesis to protect bacteria against viral infection, but the identities and functions of their DNA products remain largely unknown. Here we show that DRT2 systems encode an unprecedented immune pathway that involves de novo gene synthesis via rolling circle reverse transcription of a non-coding RNA (ncRNA). Programmed template jumping on the ncRNA generates a concatemeric cDNA, which becomes double-stranded upon viral infection. Remarkably, this DNA product constitutes a protein-coding, nearly endless ORF (*neo*) gene whose expression leads to potent cell growth arrest, thereby restricting the viral infection. Our work highlights an elegant expansion of genome coding potential through RNA-templated gene creation, and challenges conventional paradigms of genetic information encoded along the one-dimensional axis of genomic DNA.

Mobile genetic elements (MGEs), including viruses, plasmids, and transposons, act as a major driving force of genome evolution by expressing enzymes that catalyze diverse DNA rearrangements (*1*). Many of these enzymes are encoded by the most abundant gene family in nature (*2*), and over sufficiently long timescales, MGEs can become dominant constituents of their host genomes due to their proliferative properties (*3*). In response, cells harbor arsenals of defense mechanisms to counter the pervasive spread of MGEs, which range in complexity from single-gene modules to the multiorgan vertebrate immune system (*4*). MGEs themselves often provide the source material from which anti-MGE mechanisms are derived (*5*). In vertebrates, for example, domestication of an ancestral transposon led to the evolution of V(D)J recombination, laying the groundwork for protein-based adaptive antiviral immunity by facilitating antigen receptor diversification (*6*). And in bacteria, recurrent exaptation of transposon-encoded genes enabled the emergence of CRISPR–Cas systems, which exploit guide RNA molecules to recognize and cleave foreign targets during nucleic acid-based adaptive antiviral immunity (*7*–*9*). MGEs have thus contributed to the perennial host–parasite arms race as both aggressor and defender, across all domains of life.

Inspired by the innovative molecular mechanisms of immunity born out of MGE co-option, we set out to investigate additional host pathways used by bacteria to defend against genetic invaders, and in particular, bacteriophages (*10*, *11*). We were especially intrigued by recent discoveries of diverse antiphage defense systems that encode reverse transcriptase (RT) enzymes (*12*, *13*). In stark contrast to well-studied phage defense pathways that target and destroy foreign nucleic acids using nucleases, such as restriction-modification (RM) and CRISPR–Cas, these RT-based systems presumably confer immunity via nucleic acid synthesis.

Prokaryotic RTs are thought to all descend from a common retroelement ancestor, the Group II intron, which catalyzes self-splicing and site-specific DNA insertion, and is also thought to be the precursor to the eukaryotic spliceosome (*14*). The antiviral roles of multiple classes of domesticated retroelements have emerged in recent years (*14*), reinforcing the notion that genetic conflicts position MGEs in both offensive and defensive roles. One class, for example, comprises fusions or operonic associations between an RT and the Cas1 integrase, itself a domesticated transposase, to record molecular memories from past infections by writing fragments of viral RNA into the CRISPR DNA (*15*–*17*). Retrons constitute another class that mediate innate antiphage immunity using complex operons that typically consist of an RT domain, a non-coding RNA (ncRNA), and a toxin protein (*12*, *13*, *18*). In uninfected cells, the RT reverse transcribes the ncRNA to form a complementary DNA (cDNA) product, which is thought to maintain the toxin in an inactive state; phage infection then drives cDNA modification, leading to activation of the toxin and initiation of cell suicide (*12*, *18*). This general strategy, known as abortive infection, prevents the invading phage from completing its lytic cycle and provides population-level immunity at the expense of the infected cell.

Defense-associated RT (DRT) systems comprise a third class of retroelements with antiphage function, which originate from a monophyletic clade of RTs termed the Unknown Group (UG) (*13*, *19*). In contrast to RT-Cas1 and retron systems, many DRTs feature single-gene operons (*19*), implying that reverse transcription activity alone could be responsible for providing phage defense. Initial experimental efforts have confirmed the phage defense functions of 9 UG subgroups (DRT1-9), in addition to demonstrating the functional requirement for an intact RT domain (*13, 19*). However, the identities of the DRT cDNA products, and their mechanisms of immunity, have not been studied.

Here we develop a systematic approach to profile the cDNA products of any RT enzyme of interest, and apply it to DRT2-family phage defense systems. This strategy revealed an unprecedented rolling circle reverse transcription activity, leading to the synthesis of long, concatemeric cDNA products. We discover that these cDNAs contain an open reading frame (ORF) that remains in-frame through each repeat, and that promoter sequences formed across the repeat junction result in transcription to mRNA. Expression of the nearly endless ORF (*neo*) gene, which is stringently regulated by the presence or absence of phage, leads to rapid growth arrest and programmed dormancy. Beyond revealing an elegant example of retroelement exaptation for host–MGE defense, we present evidence supporting the broad conservation of this mechanism for RNA-templated creation of extrachromosomal genes.

Results

cDNA synthesis by a defense-associated reverse transcriptase

We focused our attention on DRT2 systems because of their intriguing minimal architecture, consisting of a single ORF and upstream ncRNA described previously (*19*). Unlike most other DRT and retron systems, which typically encode a reverse transcriptase enzyme alongside one or more additional protein domains predicted to function as effectors of the immune response, DRT2 systems lack additional protein-coding genes, and the RT protein lacks domains beyond the predicted RNA-directed DNA polymerase (fig. S1A) (*11*, *19*). We therefore hypothesized that the cDNA product of the RT enzyme likely plays a critical and central role in the DRT2 immune mechanism. To identify this cDNA, we developed a sequencing approach to systematically identify RT-associated cDNA synthesis products based on immunoprecipitation of FLAG-RT fusions, which we termed cDNA immunoprecipitation and sequencing (cDIP-seq) (Materials and Methods), and performed these experiments alongside traditional RNA immunoprecipitation (RIP)-seq to also capture RT-associated RNA substrates (Fig. 1A and fig. S1B). We validated this approach on the well-studied Retron-Eco1 (formerly Ec86) (*20*) after first verifying that the FLAG-tagged retron RT retained phage defense activity comparable to the WT RT (fig. S1C). RIP-seq and cDIP-seq of Retron-Eco1 recapitulated all known features of the multicopy single-stranded DNA (msDNA), including RNase H processing of the RNA component, and the precise 5′ and 3′ ends of the DNA (fig. S1, D and E) (*21*). These results increased our confidence that a similar 'reverse transcriptomics' approach could provide new insights into the DRT2 molecular mechanism, and so we turned our attention to a candidate system from *Klebsiella pneumoniae* (*Kpn*DRT2).

After confirming that fusing the *Kpn*DRT2 RT with a FLAG epitope tag did not affect defense activity against T5 phage (fig. S1C), we performed RIP-seq and cDIP-seq from cells constitutively expressing plasmid-encoded ncRNA and RT from their native promoter, and then performed genome-wide analyses to identify RNA and cDNA molecules enriched by IP. The resulting datasets revealed that the highest enriched RNA and cDNA transcripts mapped to the *Kpn*DRT2 ncRNA locus (Fig. 1B), suggesting that the primary substrate for reverse transcription by *Kpn*DRT2 is encoded in *cis*, similar to retron systems. We ascribed the apparent RIP-seq enrichment of *RT* mRNA to the likely presence of read-through transcripts extending from the ncRNA into the coding sequence, or to co-translational immunoprecipitation of ribosome-bound mRNA. To differentiate genuine cDNA synthesis products from artifactually enriched molecules, we performed control experiments using a *Kpn*DRT2 system mutated in the RT active site (*13*) (hereafter YCAA) that is inactive for phage defense (fig. S1C). These experiments implicated the *Kpn*DRT2 ncRNA locus as the sole specifically enriched cDIP-seq hit (fig. S2A). RIP-seq and cDIP-seq experiments in the presence of T5 phage also revealed a strong and specific enrichment of transcripts derived from the *Kpn*DRT2 ncRNA locus (fig. S2B), indicating that RT substrate choice is largely unchanged during phage infection.

Mapping of RIP-seq and cDIP-seq data onto the *Kpn*DRT2 locus revealed the presence of a large ncRNA and a seemingly well-defined cDNA with the opposite strandedness relative to the ncRNA, as expected for reverse transcription (Fig. 1C). Control experiments with the inactive YCAA RT mutant demonstrated that ncRNA enrichment occurred independently of reverse transcriptase activity, whereas cDNA enrichment from this locus required an intact RT active site (Fig. 1C and fig. S2A). We next leveraged custom RNA-seq library preparation protocols based on dRNA-seq (*22*) and Term-seq (*23*) to demarcate the precise 5′ and 3′ ends of the 281-nucleotide (nt) ncRNA (Fig. 1C), while analyzing start and end coordinates from cDIP-seq alignments to define the 5′ and 3′ ends of the 119-nt cDNA (fig. S2C). dRNA-seq data revealed a single transcription start site (TSS) upstream of the ncRNA, but not the *RT* gene (Fig. 1C), suggesting that the ncRNA and *RT* share an upstream promoter, and are separated into mature transcripts via an unknown processing step. We next used a multiple sequence alignment (MSA) of DRT2 homologs to generate a covariance model of the ncRNA (fig. S2D), onto which we mapped the sequence of the *Kpn*DRT2 homolog to infer its secondary structure (Fig. 1D). The ncRNA features several conserved stem-loop (SL) elements, a template region corresponding to the cDNA product abutted by a short basal stem, and a large 3′ region that we hypothesize serves as a scaffold for sequence and/or structure-guided recruitment of the RT.

We next investigated how cDNA synthesis is altered as cells are actively infected by, and defending against, T5 phage. cDIP-seq data largely recapitulated the same observations made in the absence of phage (fig. S2, B and E), but we were wary of drawing conclusions about reverse transcription output based on an approach that would only quantify cDNAs still bound to the RT after immunoprecipitation. We therefore turned to total DNA sequencing using the input controls from cDIP-seq experiments, and our analyses revealed a strong induction in *Kpn*DRT2 cDNA levels upon phage infection (Fig. 1E and fig. S2F). Surprisingly, while cDNA synthesis products in the absence of phage were predominantly single-stranded, with opposite strandedness to the ncRNA, the presence of phage induced higher levels of both the initial cDNA product and its reverse complement (Fig. 1E). Although these experiments do not reveal the identity of the polymerase necessary for second-strand synthesis, we hypothesize that the RT possesses both RNA-templated and DNA-templated DNA polymerase activity, similar to other well-studied bacterial reverse transcriptases (*24*), and that conversion of single-stranded DNA (ssDNA) to double-stranded DNA (dsDNA) may be a key step within the antiviral defense pathway.

Rolling circle reverse transcription generates concatemeric cDNA products

We next sought to investigate the sequence requirements and potential antiviral function of cDNA synthesis. We began by mutating the sequence of individual SLs throughout the ncRNA in order to eliminate base-pairing, focusing on SL1 at the 5′ end, SL2 at the base of the template region, SL5 within the template region, and SL6 within the scaffold region (Fig. 2A). Mutations to all four regions led to a complete loss of phage defense activity, indicating possible defects in ncRNA binding, cDNA synthesis, or both (Fig. 2B). When we directly interrogated ncRNA binding and cDNA synthesis by the RT using RIP-seq and cDIP-seq, respectively, we found that SL1 and SL6 mutants led to either a partial or complete loss of cDNA synthesis, likely due to disruptions in the positioning of the RT on the ncRNA (Fig. 2C). The SL5 mutant exhibited strong ncRNA and cDNA enrichment, as did an additional mutant in which the region surrounding the cDNA synthesis start site was scrambled (fig. S3, A to C), suggesting that defense activity depends on not only cDNA synthesis, but also on the sequence of the cDNA product itself. The phenotype of the SL2 mutant, however, was puzzling: the sequence of the template region was completely unchanged and cDNA production resembled the WT system, and yet phage defense was completely abolished (Fig. 2, B and C). This apparent discrepancy indicated that, beyond production of cDNA with the appropriate ncRNA-specified sequence, additional features of the cDNA product underlie phage defense activity.

Given that the stem of SL2 borders the template region, we hypothesized that disruption of this structural element might lead to imprecise initiation or termination of cDNA synthesis, which in turn might explain the altered immune function. We inspected the 3′ termini of cDIP-seq reads more closely, and to our surprise, we found that the large majority of reads extended well beyond the boundary defined by the read alignment coverage; these extensions had been soft-clipped from the alignments by conventional mapping algorithms (Fig. 2D). To determine the identity of these soft-clipped extensions, we extracted their sequences and mapped them back to the plasmid and *E. coli* genome. Remarkably, these sequences in fact derived from the 5′ end of the cDNA (Fig. 2D), suggesting a template jumping mechanism whereby the RT proceeds from the end of the template region back to the start, resulting in concatemeric cDNA repeat products. Whereas the concatemeric cDNAs generated by the WT system had a precise and uniform head-to-tail junction, including one additional nucleotide immediately adjacent to SL2, junction sequences for the SL2 mutant were more heterogeneous (Fig. 2D). Indeed, when we quantified the abundance of the expected junction sequence across all tested ncRNA mutants, we found that only the SL5 and cDNA start mutants retained WT levels of the repeat junction, whereas all other SL mutants nearly eliminated the expected template jumping products (fig. S3D).

We next quantified concatemeric cDNA (ccDNA) in total DNA samples from cells +/− T5 phage infection. T5 phage infection triggered a large increase in the abundance of bottom-strand junction-spanning reads, corresponding to the initial products of RNA-templated DNA synthesis (Fig. 2E). This was matched with a concomitant increase in top-strand junction-spanning reads (Fig. 2E), suggesting that single-stranded ccDNA is efficiently converted into dsDNA in a phage and

RT-dependent manner. Similar analyses from cDIP-seq datasets showed a lesser increase in top-strand junction-spanning reads during phage infection (fig. S3E), which we attribute to the RT likely having lower affinity for the dsDNA generated by second-strand cDNA synthesis, such that it releases these products in cells and/or during immunoprecipitation.

Although short-read sequencing enabled accurate determination and quantification of cDNA repeat junctions, we next leveraged long-read Nanopore sequencing to assess the length of ccDNA products, and to obtain orthogonal evidence of template jumping with a PCR-free approach. Remarkably, ccDNA from phage-infected cells spanned a range of repeat lengths from 1–40 (Fig. 2F and fig. S3F), suggesting that reverse transcription by *Kpn*DRT2 may be highly processive and involve many consecutive rounds of template jumping to generate long ccDNA (from 120 to ~5000 bp). Finally, we carefully inspected the sequence and secondary structure of the ncRNA in order to better understand the mechanism of template jumping. Concatenation of cDNA repeats occurs between the sequences directly abutting SL2, and we noticed that the terminal 3-nt of each repeat are templated by a conserved 3′-ACA-5′ (ACA-1) whose sequence perfectly matches the right half of SL2 (ACA-2; Fig. 2G). We therefore hypothesized that the RT may dynamically melt SL2 during each round of reverse transcription, allowing the terminal 5′-TGT-3′ of nascent cDNA transcripts to equilibrate between hybridization to ACA-1 and ACA-2, and thus prime a subsequent round of cDNA synthesis (Fig. 2G). This model was supported by a complete loss of defense activity in ncRNA mutants disrupting homology between the ACA motifs (fig. S3G). We note that the proposed cDNA concatenation mechanism resembles rolling circle DNA replication (*25*), and henceforth refer to the generation of ccDNA as rolling circle reverse transcription (RCRT) (Fig. 2G).

Concatemeric cDNAs encode a translated open reading frame (ORF)

Conventional rolling circle amplification during plasmid and phage replication utilizes a circular template (*25*), and thus we sought to rule out the alternative explanation that RCRT occurs as a result of ncRNA circularization at the repeat junction. To investigate this hypothesis, we reanalyzed our RIP-seq input controls, which represent total RNA-seq datasets, for the presence of reads spanning the repeat junction. Strikingly, we detected such reads abundantly in *Kpn*DRT2 samples, but they strictly depended on the presence of phage and an active RT, and more unexpectedly, their strandedness was opposite to that of the ncRNA (Fig. 3A and fig. S4A). This observation raised the intriguing possibility that cDNA second-strand synthesis might generate a template strand for another round of transcription by RNA polymerase (Fig. 3B). The resulting transcript would have opposite strandedness to the initial ncRNA and would contain multiple repeats of the cDNA sequence.

In agreement with this hypothesis, inspection of the cDNA sequence produced by RCRT revealed consensus promoter elements spanning the repeat junction (Fig. 3B), highly reminiscent of transposon promoters that are selectively formed upon DNA circularization during the transposon excision step (*26*, *27*). These observations suggested that phage infection might trigger the production of a high-copy, concatemeric RNA molecule with downstream antiphage function. Consistent with this idea, we found that concatemeric RNA transcripts were strongly induced shortly after phage infection by ~10,000-fold (Fig. 3C). Northern blot analysis from phage-infected cells using a probe selective for the repeat–repeat junction revealed a broad size distribution spanning hundreds to thousands of nucleotides (Fig. 3C), in excellent agreement with the large size of cDNA products observed via Nanopore sequencing (Fig. 2F).

What could be the function of transcribing a repetitive cDNA sequence into RNA during an antiphage immune response? We closely examined the sequence of the cDNA and noticed that if we translated this sequence *in silico*, one out of three reading frames would lack any stop codons (Fig. 3D and fig. S4B). This observation led us to hypothesize that concatemeric RNA produced during phage infection might be translated to generate an antiviral polypeptide. This hypothesis was supported by the presence of a predicted ribosome binding site upstream of the predicted start codon (Fig. 3D and fig. S4C), and by the observation that programmed template jumping adds one additional nucleotide during each round of cDNA synthesis (Fig. 2G). This activity generates a 120-bp cDNA repeat unit comprising exactly 40 sense codons, such that the reading frame would be preserved through each repeat to yield a continuous open reading frame (ORF) (Fig. 3D).

We set out to comprehensively test the hypothesis that translation of the continuous ORF within the concatemeric RNA is necessary for phage defense. First, we identified a region within SL3 that was not strongly conserved in sequence, and introduced single-bp mutations that would generate a synonymous, missense, or nonsense codon (Fig. 3E). While the synonymous (silent) and missense mutations had mild effects on defense activity that we attributed to perturbation of the ncRNA secondary structure, the nonsense mutation completely abolished phage defense despite retaining WT levels of concatemeric RNA production (Fig. 3F and fig. S4D). We also targeted the predicted start codon and found that defense activity was partially preserved with mutation to GUG, a common non-canonical start codon in *E. coli* (*28*), while all other mutations were inactive (fig. S4, E and F). To assess whether translation of multiple contiguous repeats of the

ORF was necessary for phage defense, we tested mutations that would introduce stop codons near the end of one full ORF repeat and found that these, too, led to a loss of defense (fig. S4, E and F). Finally, inspired by classic experiments performed by Crick and Brenner to deduce the triplet nature of the genetic code (*29*), we selected three ncRNA loop regions and designed insertions ranging from 1–9 bp in length. We hypothesized that if translation of the ORF was required for defense, then out-of-frame mutations would lead to a loss of defense, whereas in-frame mutations (i.e., insertions of a multiple of 3 bp) would be partially or completely tolerated. Remarkably, we found that all out-of-frame perturbations across three non-conserved loop regions, including minimal 1-bp insertions, caused a $>10^3$-fold decrease in phage defense, while insertions of 3, 6, or 9 bp maintained near-WT activity levels (Fig. 3G).

Collectively, these experiments provided compelling genetic evidence for the existence and expression of a cryptic gene produced by RNA-templated concatenation of DNA repeats. Intriguingly, this de novo gene exhibits a heterogeneous length distribution and lacks any in-frame stop codons, and thus we refer to it as *neo* (nearly endless ORF).

Neo-*encoded polypeptides induce cell dormancy*

Encouraged by our genetics assays supporting the translation of *neo*, we next sought unambiguous biochemical evidence of Neo protein products. We initially designed a panel of ncRNA variants that would encode epitope-tagged Neo proteins to facilitate protein visualization, but were unable to isolate an epitope-tagging scheme that retained phage defense activity (fig. S5A). We therefore proceeded to mass spectrometry (MS)-based proteomics, and designed a custom protease cocktail based on the predicted amino acid composition of Neo that would yield suitable peptide fragments for MS (fig. S5B). We then extracted proteins from *Kpn*DRT2-expressing cells and performed liquid chromatography with tandem mass spectrometry (LC-MS/MS) analysis (Fig. 4A). Neo-derived peptides were exclusively detected in phage-infected cells that expressed the WT RT enzyme (Fig. 4B), and their abundances were substantial when compared to the rest of the *E. coli* proteome (Fig. 4C). These results provide concrete proof that *neo* mRNAs transcribed from ccDNA are translated into protein.

To gain further insights into the physiological consequences of Neo expression, we performed additional proteomics experiments using a more standard trypsin-based digestion procedure, and analyzed the differential protein abundance between T5 phage-infected cells expressing WT or YCAA-mutant *Kpn*DRT2. Phage proteins were significantly depleted in WT cells, as expected for a protective immune response (Fig. 4D). On the host side, two significantly enriched cellular factors immediately captured our attention—ArfA and RMF—due to their associations with ribosome stress and ribosome hibernation, respectively (Fig. 4D). ArfA (alternative ribosome-rescue factor A) is a translation factor that specifically rescues ribosomes stalled on aberrant mRNAs lacking a stop codon, acting as an alternative to the transfer-messenger RNA (tmRNA) pathway that tags nascent polypeptide chains for degradation (*30*, *31*). ArfA is known to be specifically up-regulated under conditions of tmRNA depletion (*32*), and its ribosome rescue activity in *neo*-expressing cells would elegantly resolve the conundrum of how stop codonless *neo* mRNAs are nonetheless translated into functional proteins (Fig. 4E). Meanwhile, RMF (ribosome modulation factor) is a ribosome-associated protein that directs the assembly of 70S ribosomes into inactive 100S dimers during stationary phase (*33*, *34*), and is activated by the alarmone ppGpp, a known trigger of growth arrest and cellular dormancy (*35*, *36*). We assessed the contributions of ribosome rescue and hibernation pathways to DRT2 defense using a panel of single-gene knockout strains and found that immune activity was only moderately affected (fig. S5C), suggesting the presence of compensatory pathways or pleiotropic effects, or alternatively, that ArfA and RMF up-regulation are byproducts of Neo activation rather than key driving processes. Altogether, ArfA induction indicates that Neo translation is associated with activation of an alternative ribosome rescue pathway, and RMF induction suggests that Neo production is linked to cellular dormancy.

Abortive infection and programmed dormancy have emerged in recent years as common mechanisms by which bacterial immune systems provide population-level immunity against phage infection, as host shutdown of metabolic processes prevents phage replication, and consequently, viral spread (*11*, *37*, *38*). To investigate whether DRT2 uses a similar immune mechanism, we performed phage infection assays in liquid culture at varying multiplicities of infection (MOI). *Kpn*DRT2-expressing cultures survived T5 phage infection at low MOI, but infection at high MOI led to stalling of growth (Fig. 4F). Further analysis of cultures infected at high MOI revealed that *Kpn*DRT2 effectively blocked T5 replication (fig. S6A), and that the growth-arrested cells remained viable (fig. S6, B and C), altogether supporting a mechanism of phage defense via programmed dormancy.

We next sought to test the physiological effects of recombinant Neo expression from coding sequences containing start and stop codons, independent of RT-ncRNA activity, to eliminate any confounding factors from the intricate steps involved in *neo* gene synthesis. We initially attempted to clone various repeat lengths of *neo* onto a standard inducible expression vector and test the hypothesis that *neo* expression would be sufficient to trigger cellular dormancy. Yet repeated attempts to clone expression vectors with more than 2 repeats of WT *neo* proved unsuccessful: the few colonies that emerged consistently exhibited frameshift mutations or lacked the *neo* insert altogether (fig. S6, D and E). In contrast, control sequences encoding the same amino acids in a scrambled order could be cloned with high efficiency (fig. S6E). These qualitative results suggested that Neo may potently arrest cell growth, and that its leaky expression had prevented the isolation of positive clones. Intriguingly, this effect was only observed with Neo repeat lengths of 3 or more.

To circumvent this challenge, we adopted an alternative strategy (*39*), in which *neo* genes were placed on a low-copy vector under the control of a tightly regulated pBAD promoter and theophylline riboswitch (Fig. 4G). This multilayered strategy for control of *neo* expression — which evokes the elaborate regulation of *neo* expression by native DRT2 loci — enabled the isolation of the desired clones. We then transformed cells with expression vectors encoding WT or scrambled *neo* with 1-3 repeats, and monitored cell growth in liquid culture before and after inducing *neo* expression with arabinose and theophylline. We found that only the 3-repeat WT Neo construct exhibited any growth defect compared to an empty vector control (Fig. 4H). To assess whether the growth-arrested cells could recover from dormancy, we plated cells from the final time point of the liquid culture experiment on solid media supplemented with either repressor (glucose) or inducer (arabinose and theophylline). We found that cells expressing 3-repeat WT Neo exhibited a ~10^2-fold increase in colony-forming units when plated on repressor versus inducer (fig. S6F), indicating strong recovery from Neo-induced dormancy.

Considered together, these results suggest that the intricate gene synthesis mechanism encoded by *Kpn*DRT2 may have evolved in order to strictly control the production of an effector protein whose toxicity is too potent to be safely controlled by conventional regulatory strategies.

Neo *gene synthesis and Neo protein toxicity is a broadly conserved phage defense strategy*

Equipped with a wealth of mechanistic information on the production of Neo protein by *Kpn*DRT2, we set out to explore the evolutionary conservation of this gene synthesis strategy for antiviral defense. Starting with a large phylogenetic tree of DRT2 homologs (fig. S7A and table S1), we used covariance models to annotate *RT*-associated ncRNAs and then extracted the putative *neo* gene and Neo protein sequences based on the expected mechanism of template jumping and absence of in-frame stop codons (Fig. 5A and Materials and methods). Our pipeline identified candidate ncRNAs and Neo proteins for the vast majority of DRT2 systems that were related to *Kpn*DRT2 (Fig. 5B), revealing broad conservation of this unique mechanism for concatemeric gene synthesis. Notably, sequence motifs expected to be critical for *neo* gene synthesis and expression, including ACA-1, ACA-2, and repeat junction-flanking promoter elements, were also strongly conserved across diverse homologs (fig. S7B). Iterative generation of additional covariance models also enabled ncRNA prediction for divergent DRT2 clades, but Neo protein annotation was more challenging, suggesting the possibility of alternative mechanisms of RCRT and *neo* gene expression (fig. S7, A and C).

Next, we investigated the amino acid sequence of diverse Neo proteins in more detail. Bioinformatics analyses of multiple sequence alignments failed to identify any functional domains or similarities to known proteins, but they did reveal high-confidence predictions of α-helical secondary structural elements (Fig. 5C). Using a 3-repeat Neo sequence, we predicted the 3D protein fold using multiple independent methods, which yielded a structure reminiscent of HEAT repeats and other alpha solenoids consisting of repeating antiparallel α-helices (*40*, *41*) (Fig. 5D and fig. S7D). To test this model, we introduced helix-breaking proline residues into either the loop connecting two α-helices, or into the helices directly (Fig. 5D), and assessed the effects of these perturbations on cell growth. Consistent with our structural model, insertions into either helix eliminated Neo-induced growth arrest, whereas the loop insertion mutant exhibited a dormancy phenotype similar to the WT Neo sequence (Fig. 5E).

Having found that Neo proteins exhibit conserved α-helical folds, we sought to experimentally test the conservation of additional critical features of Neo production and cellular function. We selected and cloned 5 diverse DRT2 homologs (Fig. 5B and fig. S8A), and performed Nanopore sequencing of total DNA from cells transformed with DRT2 expression vectors to assess the distribution of *neo* cDNA repeat lengths. Remarkably, nearly all of the tested systems exhibited RCRT upon heterologous expression in *E. coli*, and the cDNAs spanned a wide range of abundance levels and repeat lengths (Fig. 5F). We also tested the effect of recombinant expression of the Neo proteins predicted to be encoded by these concatemeric cDNAs. In all cases, Neo homolog expression led to repeat length-dependent growth arrest, further confirming the requirement for cDNA repeat concatenation in the programmed dormancy mechanism to defend against phage infection (Fig. 5G and fig. S8B).

Collectively, these experiments establish the generalizability of the *Kpn*DRT2 gene synthesis and antiphage defense mechanisms across a large swath of related immune systems.

Discussion

Our work reveals an unprecedented mechanism of antiviral immunity mediated by DRT2 defense systems (Fig. 5H). In uninfected cells, the ncRNA and RT enzyme are constitutively expressed from a single promoter, leading to synthesis of a repetitive single-stranded cDNA via precise, programmed template jumping that mediates rolling circle reverse transcription (RCRT). Upon phage infection, second-strand synthesis is triggered, leading to the accumulation of double-stranded, concatemeric cDNA molecules. A promoter created across the junction between adjacent cDNA repeats then leads to expression of abundant, heterogeneously sized mRNAs encoding a stop codon-less, nearly endless ORF (*neo*). It is the Neo protein, we propose, that acts as the effector arm of the immune system by rapidly arresting cell growth and inducing programmed dormancy, thus protecting the larger bacterial population from the spread of phage.

This pathway complicates textbook descriptions of the central dogma of molecular biology by highlighting complex and repeated transitions back and forth between DNA- and RNA-based carriers of genetic information, before translation finally yields a protein product. Furthermore, it challenges the universal paradigm that genes are encoded linearly along the chromosomal axis. Genes across all three domains of life are arranged in a polarized and singular orientation from head to tail, even considering the existence of intron splicing and discontinuous exon joining in eukaryotes. Although *neo* proto-genes are similarly arranged, synthesis of the mature gene form requires RNA-templated concatenation of the tail of one proto-gene to the head of another. When considered alongside other examples of strategies used by mobile elements to compactly encode genetic information, including ribosomal frameshifting, overlapping ORFs, and nested genes, our work adds another layer of complexity to the ways in which protein-coding sequences can be stored in the genome.

Why would such an elaborate immune pathway and gene synthesis mechanism have evolved? One potential explanation is the need for stringent control of Neo expression. Our initial attempts to clone recombinant Neo for functional testing were fraught with challenges, including the rapid selection of loss-of-function mutants due to cellular toxicity, suggesting an extreme fitness cost associated with even low levels of Neo expression. It is likely that genomic encoding of pre-assembled *neo* genes, under standard transcriptional control mechanisms, would pose intolerable autoimmune risk to the host. We therefore hypothesize that gating Neo expression behind multiple layers of regulation, including RNA-templated cDNA synthesis, phage-triggered dsDNA synthesis, and ArfA-mediated ribosome rescue, likely allowed a potent dormancy factor to be stably maintained as part of the immune response.

Perhaps the most mysterious aspect of DRT2 immunity that remains to be elucidated is the structure and molecular function of Neo. Although we detected unambiguous Neo-derived peptides in phage-infected cells via mass spectrometry, the necessary protease digestion steps precluded determination of its size. Our results indicate that a Neo polypeptide of at least 3 repeats in length is necessary and sufficient to induce cell dormancy, but it is worth noting that *neo* mRNAs range in size from 200 to >5,000 nt, and that translation could in theory initiate from within any *neo* repeat, each of which contains its own RBS. Thus, we anticipate that native Neo proteins are similarly heterogeneous in size, potentially spanning hundreds to thousands of amino acids. Mass spectrometry-based proteomics also highlighted ribosome hibernation as a downstream effect of DRT2 immune system function, though additional experiments will be necessary to determine whether RMF activation is a direct consequence of Neo, or, more likely, an indirect consequence of programmed dormancy. The fact that RMF activation is driven by the alarmone ppGpp (*35*), which causes dormancy and growth arrest and is itself synthesized on ribosomes (*42*), raises the possibility that Neo directly induces this cellular pathway. Finally, the conservation of α-helical repeat (αRep) domains fused to reverse transcriptase domains in Class 1 DRT systems (*19*), which strongly resemble the predicted α-helical fold of Neo, tantalizingly suggest a potential unifying theme for the effector functions across all DRT systems.

Our discovery of highly efficient rolling circle reverse transcription activity represents a unique biochemical behavior that produces concatemeric, repetitive cDNA molecules with precise junction sequences. Although many other characterized reverse transcriptases exhibit intermolecular template switching activity in vitro (*43*, *44*), the intramolecular template jumping mechanism we describe here is a striking example of such an activity creating de novo protein-coding genes in vivo, with direct implications for biological function. Additional work will be needed to determine the specific adaptations in the RT enzyme that facilitate this activity, but our bioinformatics analyses and experimental results point to the critical importance of conserved ncRNA structure and sequence motifs — in particular, the ACA motifs found abutting and within SL2. These motifs likely guide reannealing of the nascent cDNA transcript after one round of cDNA synthesis to a second template immediately upstream of the cDNA start site, thereby initiating a new round of cDNA synthesis. When considered alongside other well-studied examples of programmed template switching — such as the synthesis of subgenomic RNAs by coronaviral RNA-dependent RNA polymerases (*45*, *46*), and the synthesis of full-length genomic cDNAs by retroviral reverse transcriptases (*47*) — our work expands the diversity of products that can be generated by a single polymerase enzyme from its substrate.

Our results demonstrate that DRT2 systems constitutively produce single-stranded concatemeric cDNA molecules in the cell, but intriguingly, phage infection drives production of a double-stranded form that is necessary for ensuing transcription and translation. A critical area of future investigation will be identifying both the phage trigger(s) of second-strand DNA synthesis, as well as the responsible polymerase. Based in part on prior reports of group II intron-encoded RT enzymes possessing DNA-dependent DNA polymerase activity (*24*), we hypothesize that the DRT2-encoded RT itself generates dsDNA, but detailed biochemical experiments will be needed to thoroughly investigate this possibility. Collectively, the unique properties of DRT2-encoded enzymes offer considerable potential for biotechnology applications that leverage templated DNA production in vivo (*48–50*), but with the added advantage of programmed amplification. Notably, our data demonstrate that rolling circle reverse transcription is maintained with mutation of SL5 or the reverse transcription start site, suggesting that DRT2 could be harnessed to produce high-copy concatemeric cDNAs with user-defined sequences.

The identification of Neo proteins disrupts conventionally held notions of features that define protein-coding genes, as well as our broader understanding of genome composition. Current genome annotation algorithms generally rely on the definition of an ORF as a translated sequence of at least 30 amino acids bounded by a start and stop codon (*51*). However, only 26 codons of *neo* are identifiable from a linear view of the *K. pneumoniae* genome, and the proto-gene lacks a stop codon. Thus, *neo* genes are hidden in regions of genomes previously thought to be exclusively non-coding, suggesting that alternative bioinformatics approaches will be needed in order to discover similar genes that elude standard methods of ORF prediction. These findings seem especially important when considering the large proportion of non-coding DNA in higher eukaryotes. For example, only ~1.5% of the human genome is thought to encode proteins (*52*). While much of the remaining ~98.5% of genome content encodes RNAs with known or predicted gene regulatory functions, we posit that additional examples of Neo-like, non-canonical protein coding genes likely await future discovery in our own genomes.

Materials and methods

Plasmid and E. coli *strain construction*

All strains and plasmids used in this study are described in tables S2 and S3, respectively. Briefly, plasmids were cloned using a combination of methods, including Gibson assembly, restriction digestion-ligation, ligation of hybridized oligonucleotides, Golden Gate Assembly, and around-the-horn PCR. Plasmids were cloned and propagated in *E. coli* strain NEB Turbo (sSL0410), and all experiments were performed in *E. coli* str. K-12 substr. MG1655 (sSL0810). Clones were verified by Sanger sequencing or whole plasmid sequencing. pLG007 (Retron-Eco1) and pLG010 (DRT type 2) were gifts from Feng Zhang (Addgene plasmids # 157885, # 157888) (*13*). Single-gene knockout strains from the Keio collection (*53*) were gifts from M. Gottesman. Substitution and insertion mutations to the *Kpn*DRT2 ncRNA are numbered relative to the first nucleotide of the ncRNA, with insertion numbering referencing the nucleotide immediately upstream of the inserted bases (for instance, A48G indicates mutation of A at position 48 to G, and 53insG indicates insertion of G downstream of position 53). Insertion mutations to Neo are numbered relative to the first amino acid of each Neo repeat (for instance, N2insP indicates insertion of P downstream of residue N at position 2 within each repeat). For recombinant expression of *neo*, coding sequences were bounded by start and stop codons, and codon optimization was performed to minimize nucleotide-level sequence identity between each repeat.

Phage amplification and plaque assays

Phage T5 (a gift from Michael Laub) was amplified in liquid culture by diluting an overnight culture of MG1655 cells 1:100 in 10 mL fresh LB media, adding 50 uL of phage, and incubating at 37°C for 3-4 hours. Chloroform was added to a final concentration of 5% to facilitate complete bacterial lysis, after which the lysate was centrifuged at 4,000 x *g* for 10 min to pellet cell debris. The supernatant was passed through a sterile 0.22 μm filter, and the phage-containing filtrate was stored at 4°C.

Small-drop plaque assays were performed as follows: *E. coli* str. K-12 substr. MG1655 (sSL0810) or the indicated single-gene deletion strains (see table S2 for strain descriptions and genotypes), were transformed with the indicated plasmid construct (see table S3 for plasmid descriptions and sequences) and plated on solid LB media. Single colonies were inoculated in liquid LB media containing the appropriate antibiotic and grown overnight at 37°C with shaking. The next day, 100 μL of overnight culture were mixed with 4 mL freshly prepared molten soft agar (0.5% agar in LB media containing the appropriate antibiotic) at 42°C and poured over solid bottom agar (1.5% agar in LB media containing the appropriate antibiotic) in a 10 cm Petri dish. The soft agar was allowed to solidify for 15 min at RT, during which 10× serial dilutions of phage T5 in LB were prepared. For plating, 3 μL of each phage dilution were spotted onto the surface of the soft agar lawn and were allowed to dry uncovered for 10 min under a laminar flow hood. Plates were incubated at 37°C for 8-16 hours to allow the formation of plaques. After selecting a phage dilution with clearly distinguishable plaques, plaque forming units (PFU) mL^{-1} were calculated using the following formula: $\frac{\text{number of plaques}}{0.003\text{ml} \times \text{dilution factor}}$. Phage defense activity was assessed by calculating the fold reduction in efficiency of plating (EOP), which was determined by dividing the PFU mL^{-1} obtained on a lawn of empty vector (EV) control cells by the PFU mL^{-1} obtained on a lawn of defense system-expressing cells.

RNA and cDNA immunoprecipitation and sequencing (RIP-seq and cDIP-seq)

E. coli str. K-12 substr. MG1655 (sSL0810) was transformed with plasmids encoding C-terminally 3×FLAG-tagged Retron-Eco1 or N-terminally 3×FLAG-tagged *Kpn*DRT2 (WT or RT-inactive YCAA mutant), as well as their native flanking sequences (see table S3 for plasmid sequences). Individual colonies were inoculated in liquid LB with chloramphenicol (25 µg mL^{-1}) and grown at 37°C to OD$_{600}$ of 0.5. For experiments +/− phage infection, 40 mL cultures were split in half, and phage T5 was added to one half at a multiplicity of infection (MOI) of 5, which was calculated as the ratio of phage PFU to bacterial colony forming units (CFU), assuming 8×10^8 CFU in 1 mL culture at OD$_{600}$ of 1.0. Uninfected and infected cultures were grown for 1 hour at 37°C. For experiments without phage infection, 20 mL cultures were grown to OD$_{600}$ of 0.5 and directly harvested. Cells were harvested by centrifugation at 4,000 x *g* for 10 min at 4°C, and the supernatant was removed. The pellet was washed with 5 mL of cold TBS (20 mM Tris-HCl, pH 7.5 at 25°C, 150 mM NaCl) and spun down again as before. The supernatant was removed, and the pellet was washed with 1 mL of cold TBS before centrifugation at 10,000 x *g* for 5 min at 4°C. The supernatant was removed, and the pellet was flash-frozen in liquid nitrogen and stored at -80°C.

Antibodies for immunoprecipitation were conjugated to magnetic beads as follows: for each sample, 60 µL Dynabeads Protein G (Thermo Fisher Scientific) were washed 3× in 1 mL 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), resuspended in 1 mL IP lysis buffer, combined with 20 µL anti-FLAG M2 antibody (Sigma-Aldrich, F3165), and rotated for > 3 hours at 4°C. Antibody-bead complexes were washed 2× to remove unconjugated antibodies and resuspended in 60 µL of IP lysis buffer per sample.

Flash-frozen pellets were thawed on ice and resuspended in 1.2 mL IP lysis buffer supplemented with 1× cOmplete Protease Inhibitor Cocktail (Roche) and 0.1 U µL^{-1} SUPERase•In RNase Inhibitor (Thermo Fisher Scientific). To lyse cells, samples were sonicated using a 1/8″ sonicator probe for 1.5 min total (2 s ON, 5 s OFF) at 20% amplitude. To clear cell debris and insoluble material, lysates were centrifuged at 21,000 x *g* for 15 min at 4 °C, and 1 mL supernatant was transferred to a new tube. At this point, two small volumes of each sample (10 µL for RIP-seq and 10 µL for cDIP-seq) were set aside as "input" starting material and stored at -80°C. For immunoprecipitation, each sample was combined with 60 µL antibody-bead complex and rotated overnight at 4°C. The next day, each sample was washed 3× with 1 mL ice-cold IP wash buffer (20 mM Tris-HCl, pH 7.5 at 25°C, 150 mM KCl, 1 mM MgCl$_2$), using a magnetic rack to immobilize the beads in between each wash. During the final wash, each sample was separated into two separate 500 µL volumes for downstream RIP or cDIP processing.

For RIP elution, the supernatant was removed, and beads were resuspended in 750 µL TRIzol (Thermo Fisher Scientific). After 5 min incubation at RT, the supernatant containing eluted RNA was transferred to a new tube and combined with 150 µL chloroform. Samples were mixed vigorously by inversion, incubated at RT for 3 min, and centrifuged at 12,000 x *g* for 15 min at 4°C. RNA was isolated from the upper aqueous phase using the RNA Clean & Concentrator-5 kit (Zymo Research) and eluted in 15 µL RNase-free water. RNA from input samples was isolated in the same manner using TRIzol and column purification. Purified RNA was stored at -80°C before proceeding to library preparation.

For cDIP elution, the supernatant was removed, and beads were resuspended in 90 µL IP wash buffer and treated with 5 µg RNase A (Thermo Fisher Scientific) for 30 min at 37°C. Input samples were adjusted to 90 µL with IP wash buffer and treated with RNase A in parallel. SDS was added to IP and input samples to a final concentration of 1%, and samples were treated with 25 µg Proteinase K (Thermo Fisher Scientific) for 30 min at 55°C. Beads were immobilized using a magnetic rack, and the supernatant containing eluted DNA was transferred to a new tube. DNA was isolated using the Monarch PCR and DNA Cleanup kit (NEB), following the Oligonucleotide Cleanup protocol and eluting in 15 µL DNase-free water. For Retron-Eco1 samples, DNA was treated with DBR1 (Origene) in reactions containing 2 µL DNA, 0.5 µL DBR1, 1× rCutSmart in 10 µL total volume, in order to cleave the 2′-5′ phosphodiester linkage between the RNA and DNA components of msDNA. Reactions were cleaned up using the Monarch PCR and DNA Cleanup kit (NEB), with elution in 15 µL DNase-free water. Purified DNA was stored at -80°C before proceeding to library preparation.

For RIP-seq library preparation (input and RIP eluates), RNA was fragmented by random hydrolysis by combining 7 µL RNA, 6 µL water, and 2 µL NEBuffer 2, and heating to 92°C for 2 min. To remove DNA and prepare RNA ends for adapter ligation, samples were 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. This was followed by treatment with 1 µL T4 PNK (NEB) in 1× T4 DNA ligase buffer (NEB) for 30 min at 37°C. Reactions were column-purified using the Zymo RNA Clean & Concentrator-5 kit and eluted in 10.5 µL RNase-free water. RNA concentrations were quantified using the DeNovix RNA Assay. Illumina sequencing libraries were prepared using the NEBNext Small RNA Library Prep kit, and libraries were sequenced on an Illumina NextSeq 500 in paired-end mode with 150 cycles per end.

For cDIP-seq library preparation, 2 µL of each input sample and 10 µL of each IP eluate were diluted to 15 µL with

DNase-free water. Samples were denatured by heating at 95°C for 2 min, and then immediately placed on ice. Ligation of Illumina adapters and conversion of ssDNA to dsDNA were performed using the xGen ssDNA & Low-Input DNA Library Prep Kit (IDT), and libraries were sequenced on an Illumina NextSeq 500 in paired-end mode with 150 cycles per end.

RIP-seq and total RNA-seq analyses

RIP-seq and corresponding input datasets were processed using cutadapt (*54*) (v4.2) to remove Illumina adapter sequences, trim low-quality ends from reads, and filter out reads shorter than 15 bp. Reads were mapped to combined reference files containing the MG1655 genome (NC_000913.3) and relevant plasmid sequence, as well as the T5 genome (NC_005859.1) for +/− infection experiments, using bwa-mem2 (*55*) (v2.2.1) with default parameters. SAMtools (*56*) (v1.17) was used to sort and index alignments. Coverage tracks were generated using bamCoverage (*57*) (v3.5.1) with a bin size of 1, separation of top and bottom strand alignments, and scaling of coverage according to sequencing depth (based on the total number of reads passing initial trimming and length filtering). Coverage tracks were visualized in IGV (*58*).

For transcriptome-wide analyses of RNAs enriched by RIP-seq, aligned reads were assigned to annotated transcriptome features using featureCounts (*59*) (v2.0.2) with -s 1 for strandedness. The resulting counts matrices were passed to DESeq2 (*60*) to calculate fold-change and FDR (using the Benjamini-Hochberg procedure) between input and IP for each annotated transcript. Comparisons were visualized using ggplot2, plotting the “baseMean” (mean normalized counts across all conditions) against $\log_2$(fold change). All comparisons included three independent biological replicates.

For counting of *neo* repeat junction-spanning reads in RIP input (i.e., total RNA) samples, a custom reference sequence was made which consisted of two concatenated *neo* cDNA repeats. A 20-bp feature annotation was added, centered at the repeat–repeat junction. Reads were aligned to the custom reference sequence using bwa-mem2, and featureCounts was used to count alignments spanning the junction annotation. The resulting counts were normalized for sequencing depth.

cDIP-seq and total DNA sequencing analyses

Adapter trimming, quality trimming, and length filtering of cDIP-seq and corresponding input datasets were performed as described above for RIP-seq experiments. Trimmed and filtered reads were mapped to combined reference files, sorted, indexed, and plotted onto coverage tracks as described above. Alignments over annotated transcriptome features were counted using featureCounts with -s 2 for strandedness. The resulting counts matrices were processed by DESeq2 and plotted as described above. All transcriptome-wide comparisons were performed using three independent biological replicates.

In order to plot cDNA 5′ and 3′ ends over the *Kpn*DRT2 ncRNA locus, cDIP-seq alignment coordinates were extracted using the bamtobed utility from bedtools (*61*) (v2.31.0). The 5′ boundary of each read pair was determined as the start coordinate of read 1, for transcripts on the top strand, or the end coordinate of read 1, for transcripts on the bottom strand. Meanwhile, the 3′ boundary of each read pair was determined as the end coordinate of read 2, for transcripts on the top strand, or the start coordinate of read 2, for transcripts on the bottom strand. The boundary coordinates thus defined for each read pair were plotted as a histogram over the *Kpn*DRT2 ncRNA locus.

For counting of reads mapping to the *Kpn*DRT2 cDNA, a custom annotation file was created which defined the DRT2 cDNA feature based on the coverage boundaries from cDIP-seq of *Kpn*DRT2. Alignments over this feature were counted using featureCounts with -s 2 for strandedness and–minOverlap 60. Counting of *neo* repeat–repeat junction-spanning reads was performed as described above for RIP input samples. The proportion of junction-spanning versus non-junction-spanning cDNA alignments was calculated by dividing the junction-spanning read counts by the total number of reads mapped to the custom concatenated reference sequence.

To analyze cDIP-seq reads with soft-clipped extensions beyond the DRT2 cDNA coverage boundary, cutadapt was used to extract reads containing the full-length *Kpn*DRT2 cDNA and then trim the cDNA repeat sequence from the 5′ end of the read. This step produced trimmed reads containing only the portion of the read extending beyond the coverage boundary. The extensions were subsequently mapped back to the combined MG1655 genome, T5 phage, and DRT2 plasmid reference using bwa-mem2. Coverage tracks of the alignments were generated using bamCoverage.

dRNA-seq

To precisely map the transcription start site of the *Kpn*DRT2 ncRNA, a custom RNA-seq library preparation protocol was used to enrich primary transcripts from the total RNA pool, as previously described (*22*). *E. coli* MG1655 cells transformed with a plasmid encoding *Kpn*DRT2 were grown to exponential phase, and total RNA was extracted using TRIzol (Thermo Fisher Scientific). 1 µg of total RNA was fragmented in 1× NEBuffer 2 by heating at 92°C for 1.5 min. DNase treatment was performed with 1 µL TURBO DNase (Thermo Fisher Scientific) in the presence of 1 µL SUPERase•In RNase Inhibitor (Thermo Fisher Scientific) for 10 min at 37°C. Samples were treated with 1 µL T4 PNK in 1× T4 DNA ligase buffer (NEB) at 37°C for 30 min and purified using the Zymo RNA Clean & Concentrator-5 kit. Samples were split in two and either treated to enrich primary transcripts or used as untreated controls. To enrich primary transcripts with tri-phosphorylated 5′ ends, samples were treated with 1 µL of Terminator Exonuclease (Biosearch Technologies) in 1× Terminator Reaction Buffer A (Biosearch Technologies) supplemented with 0.5 µL SUPERase•In RNase Inhibitor (Thermo Fisher Scientific). Reactions were incubated at 30°C for 1 hour and stopped by adding EDTA to a final concentration of 5 mM. Samples were purified using the Zymo RNA Clean & Concentrator-5 kit, and then treated with 2 µL RppH (NEB) in 1× NEBuffer 2 supplemented with 1 µL SUPERase•In RNase Inhibitor (Thermo Fisher Scientific). Reactions were incubated at 37°C for 30 min and purified using the Zymo RNA Clean & Concentrator-5 kit. Illumina sequencing libraries were prepared using the NEBNext Small RNA Library Prep kit, and libraries were sequenced on an Illumina NextSeq 500 in single-end mode with 75 cycles per end.

Adapter trimming, quality trimming, and length filtering of dRNA-seq reads were performed as described above for RIP-seq experiments. Trimmed and filtered reads were mapped to reference files using bowtie2 (*62*) (v2.4.5) with default parameters. Alignments were sorted and indexed as described above. Alignment coordinates were extracted from read 1 using the bamtobed utility from bedtools (v2.31.0). The 5′ boundary of each read was determined as the start coordinate of read 1, for transcripts on the top strand, or the end coordinate of read 1, for transcripts on the bottom strand. The 5′ boundary thus defined for each read was plotted as a histogram over the KpnDRT2 ncRNA locus. Transcription start sites were evaluated based on enrichment of a given position in the Terminator Exonuclease-treated sample compared to the untreated control.

Term-seq

Term-seq was performed to enrich the 3′ ends of transcripts, as previously described (*23*), using the same RNA sample as used for dRNA-seq. 1 µg of total RNA was treated with 1 µL TURBO DNase in 1× TURBO DNase Buffer (Thermo Fisher Scientific) supplemented with 1 µL SUPERase•In RNase Inhibitor (Thermo Fisher Scientific) for 10 min at 37°C, followed by cleanup using the Zymo RNA Clean & Concentrator-5 kit. Ligation of an i7 Illumina adapter to RNA 3′ ends was performed using the NEBNext Small RNA Library Prep kit, followed by cleanup using the Zymo RNA Clean & Concentrator-5 kit. Samples were fragmented in 1× NEBuffer 2 by heating at 92°C for 1.5 min, then treated with 2 µL RppH (NEB) in the presence of 1 µL SUPERase•In RNase Inhibitor (Thermo Fisher Scientific) for 30 min at 37°C. This was followed by treatment with 1 µL T4 PNK in 1× T4 DNA ligase buffer (NEB) at 37°C for 30 min and cleanup using the Zymo RNA Clean & Concentrator-5 kit. Illumina library preparation continued with the remainder of the NEBNext Small RNA Library Prep protocol after the initial i7 adapter ligation step. Libraries were sequenced on an Illumina NextSeq 500 in single-end mode with 75 cycles per end.

Adapter trimming, quality trimming, and length filtering of Term-seq reads were performed as described above for RIP-seq experiments. Trimmed and filtered reads were mapped to reference files using bowtie2 (*62*) (v2.4.5) with default parameters. Alignments were sorted and indexed as described above. Alignment coordinates were extracted from read 2 using the bamtobed utility from bedtools (v2.31.0). The 3′ boundary of each read was determined as the end coordinate of read 2, for transcripts on the top strand, or the start coordinate of read 2, for transcripts on the bottom strand. RNA 3′ ends were plotted as a histogram over the *Kpn*DRT2 ncRNA locus.

Long-read DNA sequencing

Total DNA was extracted from *E. coli* str. K-12 substr. MG1655 (sSL0810) cells transformed with the indicated DRT2 expression vectors, using the Wizard Genomic DNA purification kit (Promega). For experiments performed in the absence of phage infection, single-stranded DNA was converted to double-stranded DNA using the Adaptase and Extension modules of the xGen ssDNA & Low-Input DNA Library Prep Kit (IDT). DNA was then purified using 1.2× AMPure XP beads (Beckman Coulter). This dsDNA conversion step was omitted for experiments performed in the presence of phage, as the Adaptase reaction is biased toward short ssDNA fragments (see user manual), and because phage infection is expected to trigger the in vivo conversion of single-stranded DRT2 cDNA to double-stranded DNA. DNA samples were prepared for long-read sequencing using the Native Barcoding Kit (Oxford Nanopore), following the manufacturer's protocol. Sequencing using an ONT MinION was performed with real time basecalling, barcode balancing, minimum read length of 200 bp, read splitting on, and minimum Q score of 8.

Adapter trimming and barcode trimming were performed with guppy barcoder (v6.5.7). To filter out non-cDNA reads, minimap2 (*63*) (v2.26) was used to align reads to plasmid reference sequences in which the expected cDNA region had been removed, as well as to the *E. coli* genome. Unmapped reads were then extracted for downstream analysis using SAMtools. The number of cDNA repeats detected in each sequencing read was determined using countPattern from the Biostrings package (v2.70.3) in R, and counts were normalized to the total number of sequenced reads for each sample. For visualization of concatemeric cDNAs from the phage-infected *Kpn*DRT2 sample, reads were aligned to an artificial reference sequence using the built-in aligner in Geneious with medium sensitivity and an iteration of up to five times. The artificial reference sequence was created by concatenating up to 50 repeats of the cDNA template. To ensure that reads were aligned to the start of the cDNA concatemer, and not stochastically across the repeated sequence, an 'anchor' sequence was appended to the 5′ end of the first strand in all filtered sequences and the beginning of the cDNA concatemer sequence, thereby enforcing synchronous alignment starting at the 5′ end of the cDNA. Coverage over the reference cDNA concatemer was then exported for visualization.

Liquid chromatography with tandem mass spectrometry

E. coli str. K-12 substr. MG1655 (sSL0810) cells transformed with the indicated DRT2 expression vectors were grown at 37°C in 50 mL LB with chloramphenicol (25 µg mL^{-1}) to OD_{600} of 0.5. Phage T5 was added at MOI 5 and cultures were infected for 1 hour. Cells were harvested by centrifugation at 4,000 x *g* for 10 min at 4°C, and the supernatant was discarded. The pellet was washed with 5 mL cold TBS (20 mM Tris-HCl, pH 7.5 at 25°C, 150 mM NaCl) and spun down again as before. The supernatant was removed, and the pellet was washed with 1 mL of cold TBS before centrifugation at 20,000 x *g* for 5 min at 4°C. The supernatant was removed, and the pellet was flash-frozen in liquid nitrogen and stored at -80°C.

Flash-frozen pellets were thawed on ice and resuspended in 1 mL lysis buffer (100 mM ammonium bicarbonate, 2% sodium deoxycholate). Cells were sonicated using a 1/8″ sonicator probe for 1.5 min total (5 s ON, 10 s OFF) at 20% amplitude. Lysates were heated to 95°C for 10 min. Protein concentrations were assessed using the Pierce BCA assay (Thermo Fisher Scientific). 50 µg of each sample were subjected to reduction by DTT and alkylation by IAA before being precipitated onto SP3 beads as previously described (*64*). The beads were washed and then the samples were split in two, to be digested under different digestion conditions. In one condition, proteins underwent on-bead digestion by trypsin, glu-c, and chymotrypsin; this protease mixture was specifically chosen to generate peptides from the *Kpn* Neo protein in an amino acid length range suitable for detection by LC-MS/MS (fig. S5A). In the other condition, proteins underwent on-bead digestion by trypsin alone. This more conventional digestion approach was adopted to facilitate the analysis of global proteomic changes that occurred under the different experimental conditions. Each of the proteases was added in a 1:50 enzyme:substrate ratio for overnight digestion at room temperature.

Whole proteome, label-free MS analyses were performed by data-independent acquisition (DIA). Approximately 1 µg of total peptides was analyzed on a Waters M-Class UPLC using a 15 cm IonOpticks Aurora Elite column (75 µm inner diameter; 1.7 µm particle size; heated to 45°C) coupled to a benchtop Thermo Fisher Scientific Orbitrap Q Exactive HF mass spectrometer. Peptides were separated at a flow rate of 400 nL/min with a 150 min gradient, including sample loading and column equilibration times. Data were acquired in data-independent mode using Xcalibur 4.5 software. MS1 Spectra were measured with a resolution of 120,000, an AGC target of 3 × 10^6 and a mass range from 350 to 1600 m/z. Per MS1, 29 equally distanced, sequential segments were triggered at a resolution of 30,000, an AGC target of 3 × 10^6, a segment width of 43 m/z, and a fixed first mass of 200 m/z. The stepped collision energies were set to 22.5, 25, and 27.

Two separate searches were conducted for the two digestion conditions. All DIA data were analyzed with Spectronaut software (*65*) (v18.6) using directDIA analysis methodology against a combined reference database including the *E. coli* proteome (NCBI RefSeq assembly GCF_000005845.2), T5 phage proteome (NCBI RefSeq assembly GCF_000858785.1), and the *Kpn*DRT2 RT and Neo (5 repeat) sequences. Cysteine carbamidomethylation was set as a fixed modification, and methionine oxidation and N-terminal acetylation were set as variable modifications. For the Neo-targeted experiment, trypsin, glu-c, and chymotrypsin were set as the digestion enzymes. For the global proteomics experiment, trypsin was set as the digestion enzyme. Normalization was performed using 'automatic normalization' in Spectronaut. Imputation was performed using 'global imputation' in Spectronaut for the global proteomics experiment, and was not performed for the Neo-targeted experiment. For differential protein abundance analysis, calculation of $\log_2$(fold change) and q-value was performed by Spectronaut using three independent biological replicates for each condition.

Concatemeric RNA production time course

For time course experiments measuring concatemeric RNA production during T5 infection of DRT2-expressing cells, *E. coli* MG1655 cells were transformed with plasmids encoding WT or catalytically inactive (YCAA) *Kpn*DRT2 and grown to mid-log phase. An aliquot of each pre-infection culture was taken as the t = 0 time point. Infections were initiated by adding phage at an MOI of 5. Over the course of 2 hours of incubation with shaking at 37°C, aliquots of each culture were taken for RNA extraction and processed as described below.

RT-qPCR

Samples for RT-qPCR analysis were prepared with three independent biological replicates. At each timepoint, 1 mL of culture was removed and centrifuged at 3,000 x *g* for 3 min. The supernatant was discarded, and the pellet was resuspended in 750 µL of TRIzol and incubated at room temperature for 5 min. 150 µl of chloroform were added, and samples were mixed by shaking and centrifuged at 12,000 x *g* for 15 min at 4°C. The upper aqueous phase was transferred to a new tube and mixed with an equal volume of absolute ethanol. Total RNA was purified using the Monarch RNA Cleanup Kit (NEB) and stored at -80°C.

cDNA synthesis was performed using 500 ng of total RNA as the input, which was first treated with 1 µl of dsDNase (Thermo Fisher Scientific) in 1× dsDNase reaction buffer in a final volume of 10 µL, and incubated at 37°C for 2 min. Reactions were stopped by adding DTT to a final concentration of 10 mM and heating to 55°C for 5 min. Reverse transcription was performed using the iScript cDNA Synthesis Kit (BioRad) following the manufacturer's instructions. The samples were stored at –20°C.

Quantitative PCR was performed in 10 µL reactions containing 5 µL SsoAdvanced Universal SYBR Green Supermix (BioRad), 0.5 µL of each primer pair at 10 µM concentration, and 4 µL of 25-fold diluted cDNA. Primers were designed to span the cDNA repeat junction. For normalization, primer pairs that anneal to the reference gene *rrsA* were used. Reactions were prepared in 384-well PCR plates (BioRad), and measurements were performed on a CFX384 RealTime PCR Detection System (BioRad) using the following thermal cycling parameters: polymerase activation and DNA denaturation (98°C for 2.5 min), 40 cycles of amplification (98°C for 10 s, 62°C for 20 s), and terminal melt-curve analysis (decrease from 95°C to 65°C in 0.5°C/5 s increments). Values are plotted as abundance of concatemeric RNA, relative to *rrsA*, relative to the WT sample at t = 0 ($2^{-\Delta\Delta Cq}$). All primer sequences are provided in table S4.

Northern blotting

RNA samples collected for RT-qPCR analysis, described above, were also used for Northern blotting analysis. After RNA purification by TRIzol and the Monarch RNA Cleanup Kit, samples were treated with TURBO DNase in TURBO DNase buffer (Thermo Fisher Scientific) for 30 min at 37°C. Reactions were cleaned up using the Monarch RNA Cleanup Kit, and RNA concentrations were measured using the DeNovix RNA Assay.

Northern blotting was performed as previously described (*66*), with modifications. In brief, equal amounts of RNA (1.2 µg) were adjusted to 8 µL total volume with water and combined with 22 µL denaturing mix (15 µL formamide, 5.5 µL formaldehyde, and 1.5 µL 10× MOPS). Samples were heated at 55°C for 15 min prior to separation on a denaturing agarose gel (1% agarose, 3.7% formaldehyde, 1× MOPS buffer) for 2.5 hours at 80 V. RNA was transferred to a Hybond-N+ membrane (GE Healthcare) by upward capillary transfer in 10× SSC (1.5 M NaCl, 0.15 M trisodium citrate dihydrate, pH 7). The next day, RNA was crosslinked to the membrane using a UV crosslinker, and the membrane was pre-hybridized in ULTRAhyb-Oligo buffer (Thermo Fisher Scientific) for 1 hour at 42°C. A biotinylated oligonucleotide probe specific for the concatemeric RNA repeat–repeat junction was added to the hybridization buffer at a final concentration of 5 nM and hybridization was performed overnight at 42°C. The next day, the membrane was washed twice with Wash Buffer 1 (2× SSC with 0.1% SDS) and twice with Wash Buffer 2 (0.1× SSC with 0.1% SDS). The membrane was developed using the Chemiluminescent Nucleic Acid Detection Module Kit (Thermo Fisher Scientific) and imaged with an Amersham Imager 600 (GE Healthcare). The membrane was then stripped using boiling 0.1% SDS, pre-hybridized with ULTRAhyb-Oligo buffer, and reprobed using a biotinylated oligonucleotide probe specific for 16S rRNA. Hybridization, washes, and imaging were done as before. All probe sequences are provided in table S4.

Infection response growth curves

Overnight cultures of *E. coli* MG1655 cells transformed with either empty vector (EV) or WT DRT2 expression vector were diluted 1:100 in LB with chloramphenicol (25 µg mL^{-1}), grown to exponential phase, and normalized to OD$_{600}$ of 0.2. 180 µL of cell culture were transferred into wells of a 96-well optical plate containing 20 µL of T5 lysate diluted to result in a final MOI of 5 or 0.05, or 20 µL of LB for the uninfected condition. The plate was incubated for 5 hours at 37°C with shaking. OD$_{600}$ values were recorded every 10 min using a Synergy Neo2 microplate reader (Biotek).

Plaque and colony formation time course

For time course experiments measuring plaque forming units (PFU) and colony forming units (CFU) during T5 infection of DRT2-expressing cells, *E. coli* MG1655 cells were transformed with plasmids encoding WT or catalytically inactive (YCAA) *Kpn*DRT2, grown to mid-log phase, and divided into uninfected and infected conditions. An aliquot of each culture was taken as the t = 0 time point for CFU counting, and an aliquot of pure phage lysate was taken as the t = 0 time point for PFU counting. Infections were initiated by adding phage at an MOI of 10, and uninfected control cultures were grown in parallel. Over the course of 2 hours of incubation with shaking at 37°C, aliquots of each culture were taken for phage titer measurements or colony counting as described below.

Phage titer measurements

Aliquots of each infected culture were removed at the indicated time points and immediately treated with chloroform (4% final concentration) in order to lyse cells and terminate infections. Lysates were centrifuged at 3,000 x *g* for 3 min in order to pellet cell debris. Supernatants containing phages were serially diluted in LB, and plaque forming units per milliliter (PFU/mL) were enumerated using the plaque assay protocol described above.

Colony counting

Aliquots of each culture were removed at the indicated time points. Cells were pelleted at 3,000 x *g* for 3 min, washed with 1 mL LB media, and resuspended in fresh LB in order to remove residual phages. Serial dilutions of each culture were prepared and 7.5 µL of each dilution were spot-plated on LB agar supplemented with chloramphenicol (25 µg mL^{-1}). Plates were incubated overnight at 37°C, and colonies were counted the next day. Colony forming units per milliliter (CFU/mL) were calculated using the following formula: $\frac{\text{number of colonies}}{0.0075\,\text{ml} \times \text{dilutiuon factor}}$.

Resazurin cell viability assays

Cell viability was evaluated with the resazurin-based reagent alamarBlue HS (Thermo Fisher Scientific). 200 µL cultures were prepared as described above for cell growth experiments performed at varying MOIs. Infections proceeded for 3 hours before 180 µL of cell culture were mixed with 20 µL of alamarBlue HS and incubated with shaking at 37°C. During incubation, fluorescence was measured in relative fluorescence units every 10 min according to the manufacturer's guidelines, using a Synergy Neo2 microplate reader (Biotek) with a monochromator module set to a fixed gain setting of 75. The fluorescence of blank LB controls was subtracted as background from all other measured values.

Neo induction experiments

Cellular growth curves

E. coli MG1655 cells were transformed with plasmids encoding various repeat lengths of WT or mutant Neo. Individual colonies were inoculated in LB supplemented with kanamycin (50 µg mL^{-1}) and glucose (2%) and grown until cells reached OD_{600} 0.8-1.0. The cells were then pelleted and resuspended in LB media with kanamycin (50 µg mL^{-1}). For each sample, OD_{600} density was normalized to 0.1, and 200 µL of cell suspension were transferred to a 96-well clear-bottom plate. The OD_{600} was measured using a Synergy Neo2 microplate reader (Biotek) while shaking at 37 °C for 50 min (until OD_{600} reached ~0.3). Neo expression was then induced by the addition of arabinose (final concentration 0.5%) and theophylline (final concentration 0.5 mM), and cell growth was monitored for another 2 hours. For experiments testing the induction of diverse Neo homologs, growth rates were calculated using the formula $\mu = \frac{\ln(OD600_t) - \ln(OD600_0)}{t - t_0}$. The time window of 30 min to 80 min after induction was used to calculate the growth rate for each condition.

Spot assays and CFU counting after Neo induction

To assess cell viability after *Kpn*Neo induction, a small volume from each well of the growth curve experiment described above was taken for plating on LB agar. 10× serial dilutions of each culture were prepared and spot-plated on LB agar supplemented with either kanamycin (50 µg mL^{-1}) and glucose (2%), or kanamycin (50 µg mL^{-1}), arabinose (0.5%), and theophylline (0.5 mM). Plates were incubated overnight at 37°C, and colony forming units per milliliter (CFU/mL) were counted the next day.

Protein secondary structure prediction

Six Neo protein sequences were aligned with MAFFT (*67*) (LINSI option; v7.520), and the resulting alignment was visualized with Jalview (*68*) (v2.11.3.2). Secondary structural elements were predicted by submitting the alignment to Ali2D (*69*). The consensus predicted structure annotations and mean confidence values are plotted above the alignment in Fig. 5C.

Protein tertiary structure prediction

The Neo 3D structure was modeled using three independent prediction tools. The primary amino acid sequence of *Kpn*Neo was used as input for AlphaFold2 using MMseqs2 (via ColabFold), (*70*, *71*) and the same sequence was used as input for ESMFold (*72*). A multiple sequence alignment (MSA) of the Neo homologs shown in Fig. 5C was used as input for trRosetta (*73*). All predictions were based on 3 concatenated repeats of Neo.

Start codon prediction

The *Kpn neo* start codon was predicted using the RBS Calculator tool (*74*), using one cDNA repeat unit as the input sequence and specifying *K. pneumoniae* as the host organism.

Sequence identity matrices

Pairwise sequence identity matrices were generated in Geneious from MAFFT alignments of ncRNA and cDNA nucleic acid sequences, or of RT and Neo amino acid sequences, using default settings. Accession numbers for RT proteins are listed in table S5.

ncRNA covariance modeling

Homologs of *Kpn*DRT2 were identified using the RT amino acid sequence (WP_012737279.1) as the seed query in a BLASTP search of the NR protein database (max target sequences = 100). Nucleotide sequences 1 kb upstream and downstream of *RT* genes were retrieved, clustered at 99.9% sequence identity to remove replicates using CD-HIT (*75*) (v4.8.1), and aligned using MAFFT (*76*) (v7.505). The resulting alignment was trimmed at the 5′ and 3′ ends to the exact boundaries of the ncRNA, as determined by RIP-seq experiments with *Kpn*DRT2. These putative ncRNA sequences were clustered at 95% sequence identity using CD-HIT and realigned using mLocARNA (*77*) (v1.9.1) with default parameters. The resulting structure-based multiple sequence alignment was used to build and calibrate a covariance model (CM) using the Infernal suite (*78*) (v1.1.4). The CMsearch function of Infernal was then used to scan through nucleotide sequences of additional *drt2* loci and 1-kb flanking regions, generated by an expanded BLASTP search (max target sequences = 5000) queried on *Kpn*DRT2 and clustered at 85% sequence identity using CD-HIT. The final hits (n = 303 DRT2 loci, including *Kpn*DRT2) from the CM used to identify *Kpn*DRT2-like ncRNAs were evaluated for statistically significant covarying base pairs with R-scape (*79*) at an *E*-value threshold of 0.05 (fig. S2D).

RNA secondary structure prediction

To mitigate potential variability in results from free energy-based secondary structure predictions, evolutionary information was incorporated to inform structure inference. Covariance modeling (see above and fig. S2D) indicated that the DRT2 ncRNA is segmented into three major regions: 5′ scaffold, reverse transcription template, and 3′ scaffold. Each region was folded separately using RNAfold (*80*) and visualized using RNAcanvas (*81*) prior to reconstruction into a single structural prediction.

Phylogenetic analyses

An initial set of DRT2 sequences was identified by querying the *Kpn*RT protein sequence (WP_012737279.1) against the NR database with PSI-BLAST (3 iterations; default settings) (*82*). The top 500 results from this search did not produce any clusters at a threshold of 80% amino acid identity, so this diverse set of homologs was used for an additional BLASTP (-evalue 0.01 -max_target_seqs 1000000) search of a local copy of the NCBI NR database (downloaded on April 4, 2023). The resulting hits were further restricted to an e-value cutoff of 1 × 10^{-30}, resulting in a set of 3,056 protein accessions, for which identical protein group (IPG) information was pulled from NCBI with the Batch Entrez tool. Where possible, two genomes encoding each unique DRT2 homolog were randomly sampled from the IPG information, and these genomic sequences were retrieved from NCBI with the Batch Entrez. DRT2 homologs for which we were unable to retrieve IPG information or genomic sequences were removed from the analysis, resulting in a final dataset of 2,116 DRT2 homologs (table S1). 616 protein sequences in this final DRT2 dataset were also identified as DRT2 homologs in a previous analysis of reverse transcriptases (*19*), while no other non-DRT2 homologs from that previous analysis were present in our dataset. Finally, this set of DRT2 sequences was aligned with MAFFT (LINSI option; v7.520) and a phylogenetic tree was constructed with FastTree (*83*) (-wag -gamma; 2.1.11), before being visualized with iTOL (*84*). A subtree of KpnDRT2-like sequences was constructed by manually subsetting the tree in fig. S6A to include only a monophyletic clade encompassing KpnDRT2. Eight sequences with unexpectedly long branch lengths were manually pruned from this subtree, resulting in a phylogeny of 539 KpnDRT2-like sequences (Fig. 5B).

Systematic ncRNA and Neo prediction

An updated *Kpn*DRT2-like CM (v2) was built by retrieving genomes for the top 500 DRT2 hits from the PSI-BLAST search described above, extracting the DRT2 loci (*drt2* +/− 1 kb), searching each locus with the CMsearch function of Infernal (v1.1.5; default parameters), aligning hits that met the inclusion threshold (n = 287) with LocARNA (v2.0.0; mlocarna option with default parameters), and building a CM with the CMbuild function of Infernal (v1.1.5; default parameters).

DRT2 loci, corresponding to the *RT* gene and 1 kb of upstream and downstream sequence, were then extracted from genomes encoding the 2,116 DRT2 homologs described above, using coordinates in the IPG dataset. To identify putative ncRNA sequences, these loci were queried with the *Kpn*DRT2-like ncRNA CM (v2) using the CMsearch function of Infernal (v1.1.5), with default parameters. Hits that met the inclusion threshold (e-value < 0.01) were extracted using the coordinates in the CMsearch output, and these putative ncRNA sequences were de-duplicated, prior to alignment of the sequences with the DECIPHER package (*85*) (v2.30.0) in R. The *Kpn*DRT2 locus was used as a reference to extract likely cDNA regions from the resulting alignment.

To predict Neo sequences in homologous DRT2 loci, the reverse complements of putative cDNA sequences were assessed in all three possible reading frames to determine which frame contained the fewest stop codons; these were assumed to be the *neo* open reading frames (ORFs). Start codons (ATG, GTG, TTG) were then probed in the resulting putative *neo* ORFs, and the start codon of *neo* was presumed to occur after the first ten amino acids of the putative reading frame translation, consistent with the *Kpn*DRT2 locus. Putative Neo sequences were then constructed by concatenating the translation of this downstream start codon in the putative *neo* ORF through the end of the putative cDNA (which represents the first unit of cDNA produced by the putative DRT2 rolling circle mechanism; repeat 1), to a translation of the full length putative *neo* ORF (which represents cDNA units produced by successive rounds of the DRT2 rolling circle mechanism; repeat 1 + *n*). Putative *neo* sequences that did not contain an internal stop codon, were then checked to determine if the final ten amino acids of the Neo sequence (i.e., translated from repeat 1+*n*) were identical to the final ten amino acids of Neo translated from the first unit of cDNA synthesis (i.e., translated from repeat 1) (Fig. 5A). Putative Neo sequences that met this criterion were predicted to represent bona fide Neo protein products of DRT2 immune systems.

Putative ncRNA sequences identified with the *Kpn*DRT2-like ncRNA CM (v2) were primarily restricted to a monophyletic clade that included *Kpn*DRT2 (Fig. 5B). An alignment of these ncRNA sequences was built with the DECIPHER package in R, and sequence logos (fig. S6B) were generated with the web-based version of WebLogo (*86*). Logos of cDNA sequences with identified Neo proteins (fig. S6B) were similarly built from an MSA generated with DECIPHER. To identify ncRNAs in other regions of the larger phylogenetic tree presented in fig. S6A, additional CMs were constructed via the same approach described above (i.e., mlocarna with default settings, CMbuild in Infernal) by manually selecting regions of the tree that had CM hits for at least three closely related DRT2 sequences. These CMs were used to search the DRT2 loci, and then new CMs were built from the resulting hits; this process was iterated until ncRNAs had been identified across most DRT2 systems. Exemplary ncRNA CMs generated in this process are shown in fig. S6C. Finally, Neo sequences were predicted in these newly identified putative ncRNAs using the same approach described above.

REFERENCES AND NOTES

1. L. S. Frost, R. Leplae, A. O. Summers, A. Toussaint, Mobile genetic elements: The agents of open source evolution. *Nat. Rev. Microbiol.* **3**, 722–732 (2005). doi:10.1038/nrmicro1235 Medline
2. R. K. Aziz, M. Breitbart, R. A. Edwards, Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* **38**, 4207–4217 (2010). doi:10.1093/nar/gkq140 Medline
3. A. Canapa, M. Barucca, M. A. Biscotti, M. Forconi, E. Olmo, Transposons, Genome Size, and Evolutionary Insights in Animals. *Cytogenet. Genome Res.* **147**, 217–239 (2015). doi:10.1159/000444429 Medline
4. E. V. Koonin, M. Krupovic, Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat. Rev. Genet.* **16**, 184–192 (2015). doi:10.1038/nrg3859 Medline
5. E. V. Koonin, K. S. Makarova, Y. I. Wolf, M. Krupovic, Evolutionary entanglement of mobile genetic elements and host defence systems: Guns for hire. *Nat. Rev. Genet.* **21**, 119–131 (2020). doi:10.1038/s41576-019-0172-9 Medline
6. C. Liu, Y. Zhang, C. C. Liu, D. G. Schatz, Structural insights into the evolution of the RAG recombinase. *Nat. Rev. Immunol.* **22**, 353–370 (2022). doi:10.1038/s41577-021-00628-6 Medline
7. T. Karvelis, G. Druteika, G. Bigelyte, K. Budre, R. Zedaveinyte, A. Silanskas, D. Kazlauskas, Č. Venclovas, V. Siksnys, Transposon-associated TnpB is a programmable RNA-guided DNA endonuclease. *Nature* **599**, 692–696 (2021). doi:10.1038/s41586-021-04058-1 Medline
8. H. Altae-Tran, S. Kannan, F. E. Demircioglu, R. Oshiro, S. P. Nety, L. J. McKay, M. Dlakić, W. P. Inskeep, K. S. Makarova, R. K. Macrae, E. V. Koonin, F. Zhang, The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases. *Science* **374**, 57–65 (2021). doi:10.1126/science.abj6856 Medline
9. C. Meers, H. C. Le, S. R. Pesari, F. T. Hoffmann, M. W. G. Walker, J. Gezelle, S. Tang, S. H. Sternberg, Transposon-encoded nucleases use guide RNAs to promote their selfish spread. *Nature* **622**, 863–871 (2023). doi:10.1038/s41586-023-06597-1 Medline
10. D. Mayo-Muñoz, R. Pinilla-Redondo, N. Birkholz, P. C. Fineran, A host of armor: Prokaryotic immune strategies against mobile genetic elements. *Cell Rep.* **42**, 112672 (2023). doi:10.1016/j.celrep.2023.112672 Medline
11. H. Georjon, A. Bernheim, The highly diverse antiphage defence systems of bacteria. *Nat. Rev. Microbiol.* **21**, 686–700 (2023). doi:10.1038/s41579-023-00934-x Medline
12. A. Millman, A. Bernheim, A. Stokar-Avihail, T. Fedorenko, M. Voichek, A. Leavitt, Y. Oppenheimer-Shaanan, R. Sorek, Bacterial Retrons Function In Anti-Phage Defense. *Cell* **183**, 1551–1561.e12 (2020). doi:10.1016/j.cell.2020.09.065 Medline
13. L. Gao, H. Altae-Tran, F. Böhning, K. S. Makarova, M. Segel, J. L. Schmid-Burgk, J. Koob, Y. I. Wolf, E. V. Koonin, F. Zhang, Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science* **369**, 1077–1084 (2020). doi:10.1126/science.aba0372 Medline
14. A. González-Delgado, M. R. Mestre, F. Martínez-Abarca, N. Toro, Prokaryotic reverse transcriptases: From retroelements to specialized defense systems. *FEMS Microbiol. Rev.* **45**, fuab025 (2021). doi:10.1093/femsre/fuab025 Medline
15. S. Silas, G. Mohr, D. J. Sidote, L. M. Markham, A. Sanchez-Amat, D. Bhaya, A. M. Lambowitz, A. Z. Fire, Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science* **351**, aad4234 (2016). doi:10.1126/science.aad4234 Medline
16. F. Schmidt, M. Y. Cherepkova, R. J. Platt, Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* **562**, 380–385 (2018). doi:10.1038/s41586-018-0569-1 Medline
17. A. González-Delgado, M. R. Mestre, F. Martínez-Abarca, N. Toro, Spacer acquisition from RNA mediated by a natural reverse transcriptase-Cas1 fusion protein associated with a type III-D CRISPR-Cas system in Vibrio vulnificus. *Nucleic Acids Res.* **47**, 10202–10211 (2019). doi:10.1093/nar/gkz746 Medline
18. J. Bobonis, K. Mitosch, A. Mateus, N. Karcher, G. Kritikos, J. Selkrig, M. Zietek, V. Monzon, B. Pfalz, S. Garcia-Santamarina, M. Galardini, A. Sueki, C. Kobayashi, F. Stein, A. Bateman, G. Zeller, M. M. Savitski, J. R. Elfenbein, H. L. Andrews-Polymenis, A. Typas, Bacterial retrons encode phage-defending tripartite toxin-antitoxin systems. *Nature* **609**, 144–150 (2022). doi:10.1038/s41586-022-05091-4 Medline
19. M. R. Mestre, L. A. Gao, S. A. Shah, A. López-Beltrán, A. González-Delgado, F. Martínez-Abarca, J. Iranzo, M. Redrejo-Rodríguez, F. Zhang, N. Toro, UG/Abi: A highly diverse family of prokaryotic reverse transcriptases associated with defense functions. *Nucleic Acids Res.* **50**, 6084–6101 (2022). doi:10.1093/nar/gkac467 Medline
20. A. J. Simon, A. D. Ellington, I. J. Finkelstein, Retrons and their applications in genome engineering. *Nucleic Acids Res.* **47**, 11007–11019 (2019). doi:10.1093/nar/gkz865 Medline
21. Y. Wang, Z. Guan, C. Wang, Y. Nie, Y. Chen, Z. Qian, Y. Cui, H. Xu, Q. Wang, F. Zhao, D. Zhang, P. Tao, M. Sun, P. Yin, S. Jin, S. Wu, T. Zou, Cryo-EM structures of Escherichia coli Ec86 retron complexes reveal architecture and defence mechanism. *Nat. Microbiol.* **7**, 1480–1489 (2022). doi:10.1038/s41564-022-01197-7 Medline
22. C. M. Sharma, J. Vogel, Differential RNA-seq: The approach behind and the biological insight gained. *Curr. Opin. Microbiol.* **19**, 97–105 (2014). doi:10.1016/j.mib.2014.06.010 Medline
23. D. Dar, M. Shamir, J. R. Mellin, M. Koutero, N. Stern-Ginossar, P. Cossart, R. Sorek, Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. *Science* **352**, aad9822 (2016). doi:10.1126/science.aad9822 Medline
24. S. K. Park, G. Mohr, J. Yao, R. Russell, A. M. Lambowitz, Group II intron-like reverse transcriptases function in double-strand break repair. *Cell* **185**, 3671–3688.e23 (2022). doi:10.1016/j.cell.2022.08.014 Medline
25. P. Wawrzyniak, G. Płucienniczak, D. Bartosik, The Different Faces of Rolling-Circle Replication and Its Multifunctional Initiator Proteins. *Front. Microbiol.* **8**, 2353 (2017). doi:10.3389/fmicb.2017.02353 Medline
26. B. Ton-Hoang, M. Bétermier, P. Polard, M. Chandler, Assembly of a strong promoter following IS911 circularization and the role of circles in transposition. *EMBO J.* **16**, 3357–3371 (1997). doi:10.1093/emboj/16.11.3357 Medline
27. D. Perkins-Balding, G. Duval-Valentin, A. C. Glasgow, Excision of IS492 requires flanking target sequences and results in circle formation in Pseudoalteromonas atlantica. *J. Bacteriol.* **181**, 4937–4948 (1999). doi:10.1128/JB.181.16.4937-4948.1999 Medline

28. A. Hecht, J. Glasgow, P. R. Jaschke, L. A. Bawazer, M. S. Munson, J. R. Cochran, D. Endy, M. Salit, Measurements of translation initiation from all 64 codons in E. coli. *Nucleic Acids Res.* **45**, 3615–3626 (2017). doi:10.1093/nar/gkx070 Medline

29. F. H. C. Crick, L. Barnett, S. Brenner, R. J. Watts-Tobin, General nature of the genetic code for proteins. *Nature* **192**, 1227–1232 (1961). doi:10.1038/1921227a0 Medline

30. Y. Chadani, K. Ono, S. Ozawa, Y. Takahashi, K. Takai, H. Nanamiya, Y. Tozawa, K. Kutsukake, T. Abo, Ribosome rescue by Escherichia coli ArfA (YhdL) in the absence of trans-translation system. *Mol. Microbiol.* **78**, 796–808 (2010). doi:10.1111/j.1365-2958.2010.07375.x Medline

31. K. C. Keiler, Mechanisms of ribosome rescue in bacteria. *Nat. Rev. Microbiol.* **13**, 285–297 (2015). doi:10.1038/nrmicro3438 Medline

32. F. Garza-Sánchez, R. E. Schaub, B. D. Janssen, C. S. Hayes, tmRNA regulates synthesis of the ArfA ribosome rescue factor. *Mol. Microbiol.* **80**, 1204–1219 (2011). doi:10.1111/j.1365-2958.2011.07638.x Medline

33. A. Wada, Y. Yamazaki, N. Fujita, A. Ishihama, Structure and probable genetic location of a "ribosome modulation factor" associated with 100S ribosomes in stationary-phase Escherichia coli cells. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 2657–2661 (1990). doi:10.1073/pnas.87.7.2657 Medline

34. T. Prossliner, K. Skovbo Winther, M. A. Sørensen, K. Gerdes, Ribosome Hibernation. *Annu. Rev. Genet.* **52**, 321–348 (2018). doi:10.1146/annurev-genet-120215-035130 Medline

35. K. Izutsu, A. Wada, C. Wada, Expression of ribosome modulation factor (RMF) in Escherichia coli requires ppGpp. *Genes Cells* **6**, 665–676 (2001). doi:10.1046/j.1365-2443.2001.00457.x Medline

36. T. K. Wood, S. Song, Forming and waking dormant cells: The ppGpp ribosome dimerization persister model. *Biofilm* **2**, 100018 (2020). doi:10.1016/j.bioflm.2019.100018 Medline

37. A. J. Meeske, S. Nakandakari-Higa, L. A. Marraffini, Cas13-induced cellular dormancy prevents the rise of CRISPR-resistant bacteriophage. *Nature* **570**, 241–245 (2019). doi:10.1038/s41586-019-1257-5 Medline

38. L. Fernández-García, S. Song, J. Kirigo, M. E. Battisti, M. E. Petersen, M. Tomás, T. K. Wood, Toxin/antitoxin systems induce persistence and work in concert with restriction/modification systems to inhibit phage. *Microbiol. Spectr.* **12**, e0338823–e23 (2024). Medline

39. J. Vandierendonck, Y. Girardin, P. De Bruyn, H. De Greve, R. Loris, A Multi-Layer-Controlled Strategy for Cloning and Expression of Toxin Genes in *Escherichia coli*. *Toxins (Basel)* **15**, 508 (2023). doi:10.3390/toxins15080508 Medline

40. M. A. Andrade, C. Perez-Iratxeta, C. P. Ponting, Protein repeats: Structures, functions, and evolution. *J. Struct. Biol.* **134**, 117–131 (2001). doi:10.1006/jsbi.2001.4392 Medline

41. S. H. Yoshimura, T. Hirano, HEAT repeats - versatile arrays of amphiphilic helices working in crowded environments? *J. Cell Sci.* **129**, 3963–3970 (2016). doi:10.1242/jcs.185710 Medline

42. S. Ronneau, R. Hallez, Make and break the alarmone: Regulation of (p)ppGpp synthetase/hydrolase enzymes in bacteria. *FEMS Microbiol. Rev.* **43**, 389–400 (2019). doi:10.1093/femsre/fuz009 Medline

43. M. G. Wulf, S. Maguire, P. Humbert, N. Dai, Y. Bei, N. M. Nichols, I. R. Corrêa Jr., S. Guan, Non-templated addition and template switching by Moloney murine leukemia virus (MMLV)-based reverse transcriptases co-occur and compete with each other. *J. Biol. Chem.* **294**, 18220–18231 (2019). doi:10.1074/jbc.RA119.010676 Medline

44. A. M. Lentzsch, J. L. Stamos, J. Yao, R. Russell, A. M. Lambowitz, Structural basis for template switching by a group II intron-encoded non-LTR-retroelement reverse transcriptase. *J. Biol. Chem.* **297**, 100971 (2021). doi:10.1016/j.jbc.2021.100971 Medline

45. S. Zúñiga, I. Sola, S. Alonso, L. Enjuanes, Sequence motifs involved in the regulation of discontinuous coronavirus subgenomic RNA synthesis. *J. Virol.* **78**, 980–994 (2004). doi:10.1128/JVI.78.2.980-994.2004 Medline

46. I. Sola, F. Almazán, S. Zúñiga, L. Enjuanes, Continuous and Discontinuous RNA Synthesis in Coronaviruses. *Annu. Rev. Virol.* **2**, 265–288 (2015). doi:10.1146/annurev-virology-100114-055218 Medline

47. K. Delviks-Frankenberry, A. Galli, O. Nikolaitchik, H. Mens, V. K. Pathak, W.-S. Hu, Mechanisms and factors that influence high frequency retroviral recombination. *Viruses* **3**, 1650–1680 (2011). doi:10.3390/v3091650 Medline

48. A. V. Anzalone, P. B. Randolph, J. R. Davis, A. A. Sousa, L. W. Koblan, J. M. Levy, P. J. Chen, C. Wilson, G. A. Newby, A. Raguram, D. R. Liu, Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019). doi:10.1038/s41586-019-1711-4 Medline

49. S. C. Lopez, K. D. Crawford, S. K. Lear, S. Bhattarai-Kline, S. L. Shipman, Precise genome editing across kingdoms of life using retron-derived DNA. *Nat. Chem. Biol.* **18**, 199–206 (2022). doi:10.1038/s41589-021-00927-y Medline

50. S. Tang, S. H. Sternberg, Genome editing with retroelements. *Science* **382**, 370–371 (2023). doi:10.1126/science.adi3183 Medline

51. D. Hyatt, G.-L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, L. J. Hauser, Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010). doi:10.1186/1471-2105-11-119 Medline

52. E. S. Lander, Initial impact of the sequencing of the human genome. *Nature* **470**, 187–197 (2011). doi:10.1038/nature09792 Medline

53. T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, H. Mori, Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. *Mol. Syst. Biol.* **2**, 0008 (2006). doi:10.1038/msb4100050 Medline

54. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011). doi:10.14806/ej.17.1.200

55. Md. Vasimuddin, S. Misra, H. Li, S. Aluru, "Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems" in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (2019; https://ieeexplore.ieee.org/document/8820962), pp. 314–324.

56. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi:10.1093/bioinformatics/btp352 Medline

57. F. Ramírez, D. P. Ryan, B. Grüning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dündar, T. Manke, deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44** (W1), W160-5 (2016). doi:10.1093/nar/gkw257 Medline

58. J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P. Mesirov, Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011). doi:10.1038/nbt.1754 Medline

59. Y. Liao, G. K. Smyth, W. Shi, featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014). doi:10.1093/bioinformatics/btt656 Medline

60. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014). doi:10.1186/s13059-014-0550-8 Medline

61. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). doi:10.1093/bioinformatics/btq033 Medline

62. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012). doi:10.1038/nmeth.1923 Medline

63. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018). doi:10.1093/bioinformatics/bty191 Medline

64. C. S. Hughes, S. Moggridge, T. Müller, P. H. Sorensen, G. B. Morin, J. Krijgsveld, Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat. Protoc.* **14**, 68–85 (2019). doi:10.1038/s41596-018-0082-x Medline

65. R. Bruderer, O. M. Bernhardt, T. Gandhi, S. M. Miladinović, L.-Y. Cheng, S. Messner, T. Ehrenberger, V. Zanotelli, Y. Butscheid, C. Escher, O. Vitek, O. Rinner, L. Reiter, Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteomics* **14**, 1400–1410 (2015). doi:10.1074/mcp.M114.044305 Medline

66. K. M. McKenney, R. P. Connacher, E. B. Dunshee, A. C. Goldstrohm, Chemi-Northern: A versatile chemiluminescent northern blot method for analysis and quantitation of RNA molecules. *RNA* **30**, 448–462 (2024). doi:10.1261/rna.079880.123 Medline

67. K. Katoh, K. Kuma, H. Toh, T. Miyata, MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005). doi:10.1093/nar/gki198 Medline

68. A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, G. J. Barton, Jalview Version 2—A multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009). doi:10.1093/bioinformatics/btp033 Medline

69. F. Gabler, S.-Z. Nam, S. Till, M. Mirdita, M. Steinegger, J. Söding, A. N. Lupas, V. Alva, Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Curr. Protoc. Bioinformatics* **72**, e108 (2020). doi:10.1002/cpbi.108 Medline

70. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). doi:10.1038/s41586-021-03819-2 Medline

71. M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold: Making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022). doi:10.1038/s41592-022-01488-1 Medline

72. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. Dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023). doi:10.1126/science.ade2574 Medline

73. Z. Du, H. Su, W. Wang, L. Ye, H. Wei, Z. Peng, I. Anishchenko, D. Baker, J. Yang, The trRosetta server for fast and accurate protein structure prediction. *Nat. Protoc.* **16**, 5634–5651 (2021). doi:10.1038/s41596-021-00628-9 Medline

74. H. M. Salis, E. A. Mirsky, C. A. Voigt, Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.* **27**, 946–950 (2009). doi:10.1038/nbt.1568 Medline

75. W. Li, A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006). doi:10.1093/bioinformatics/btl158 Medline

76. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). doi:10.1093/molbev/mst010 Medline

77. S. Will, T. Joshi, I. L. Hofacker, P. F. Stadler, R. Backofen, LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs. *RNA* **18**, 900–914 (2012). doi:10.1261/rna.029041.111 Medline

78. E. P. Nawrocki, S. R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013). doi:10.1093/bioinformatics/btt509 Medline

79. E. Rivas, RNA structure prediction using positive and negative evolutionary information. *PLOS Comput. Biol.* **16**, e1008387 (2020). doi:10.1371/journal.pcbi.1008387 Medline

80. R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, I. L. Hofacker, ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011). doi:10.1186/1748-7188-6-26 Medline

81. P. Z. Johnson, A. E. Simon, RNAcanvas: Interactive drawing and exploration of nucleic acid structures. *Nucleic Acids Res.* **51** (W1), W501–W508 (2023). doi:10.1093/nar/gkad302 Medline

82. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). doi:10.1186/1471-2105-10-421 Medline

83. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5**, e9490 (2010). doi:10.1371/journal.pone.0009490 Medline

84. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49** (W1), W293–W296 (2021). doi:10.1093/nar/gkab301 Medline

85. E. S. Wright, DECIPHER: Harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics* **16**, 322 (2015). doi:10.1186/s12859-015-0749-z Medline

86. G. E. Crooks, G. Hon, J.-M. Chandonia, S. E. Brenner, WebLogo: A sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004). doi:10.1101/gr.849004 Medline

ACKNOWLEDGMENTS

We thank S. Pesari and Z. Akhtar for laboratory support, A. Bernheim and J. Bondy-Denomy for helpful guidance on phage infection experiments, E. Semenova for helpful advice on ssDNA sequencing library preparation, M. Annavajhala for helpful guidance on Nanopore sequencing, A.W.P. Fitzpatrick for insightful discussions on Neo function, B. Wiedenheft for suggesting the Neo acronym, M. Laub for generously providing T5 phage, L.F. Landweber for qPCR instrument access, and the JP Sulzberger Columbia Genome Center for NGS support. **Funding:** S.T. was supported by an NIH Medical Scientist Training Program grant (T32GM145440) and an NIH Ruth L. Kirchstein Individual Predoctoral Fellowship (F30AI183830). L.C.T. and M.W.G.W. were supported by a National Science Foundation Graduate Research Fellowship. J.T.G. was supported by a Human Frontier Science Program postdoctoral fellowship (LT001117/2021-C). L.E.B. was supported by the Schaefer Research Scholars Program, the Hirschl Family Trust, and NIH grant R35GM124633. M.J. was supported by the NIH (R01AG071869 and R01HG012216), NSF (Award 2224211), and a Columbia University start-up package. This research was supported by a Pew Biomedical Scholarship, an Irma T. Hirschl Career Scientist Award, and a generous start-up package from the Columbia University Irving Medical Center Dean's Office and the Vagelos Precision Medicine Fund (to S.H.S.). **Author contributions:** S.T. and S.H.S. conceived of and designed the project. S.T. performed most experiments and experimental analyses. V.C. performed cell growth and plaque assays, time course experiments, and helped with cloning. D.J.Z. performed initial phylogenetic analyses and selected DRT2 homologs, generated the ncRNA covariance model, performed cell growth and plaque assays, and helped with cloning. R.Z. performed RT-qPCR and cell growth assays, and helped with cloning. G.D.L. performed and helped with the analysis of Nanopore sequencing experiments. T.W. performed phylogenetic analyses and bioinformatically identified Neo homologs. L.C.T. prepared samples and performed mass spectrometry experiments. M.W. performed cell growth assays. M.W.G.W. helped with the design and analysis of Neo protein expression data. J.T.G. contributed to the interpretation of rolling circle reverse transcription results. L.E.B. advised on the design and interpretation of Northern blotting experiments. M.J. advised on the design and interpretation of mass spectrometry experiments. S.T. and S.H.S. discussed the data and wrote the manuscript, with input from all authors. **Competing interests:** Columbia University has filed a patent application related to this work. 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. **Data and materials availability:** Next-generation sequencing data are available in the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO; GSE270164). Mass spectrometry data are available via the Mass Spectrometry Interactive Virtual Environment (MassIVE; MSV000095071). Materials and reagents generated in the current study are available from the corresponding author upon reasonable request.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adq0876
Figs. S1 to S8
Tables S1 to S5
MDAR Reproduicibility Checklist

Submitted 25 April 2024; accepted 17 July 2024
Published online 8 August 2024
10.1126/science.adq0876

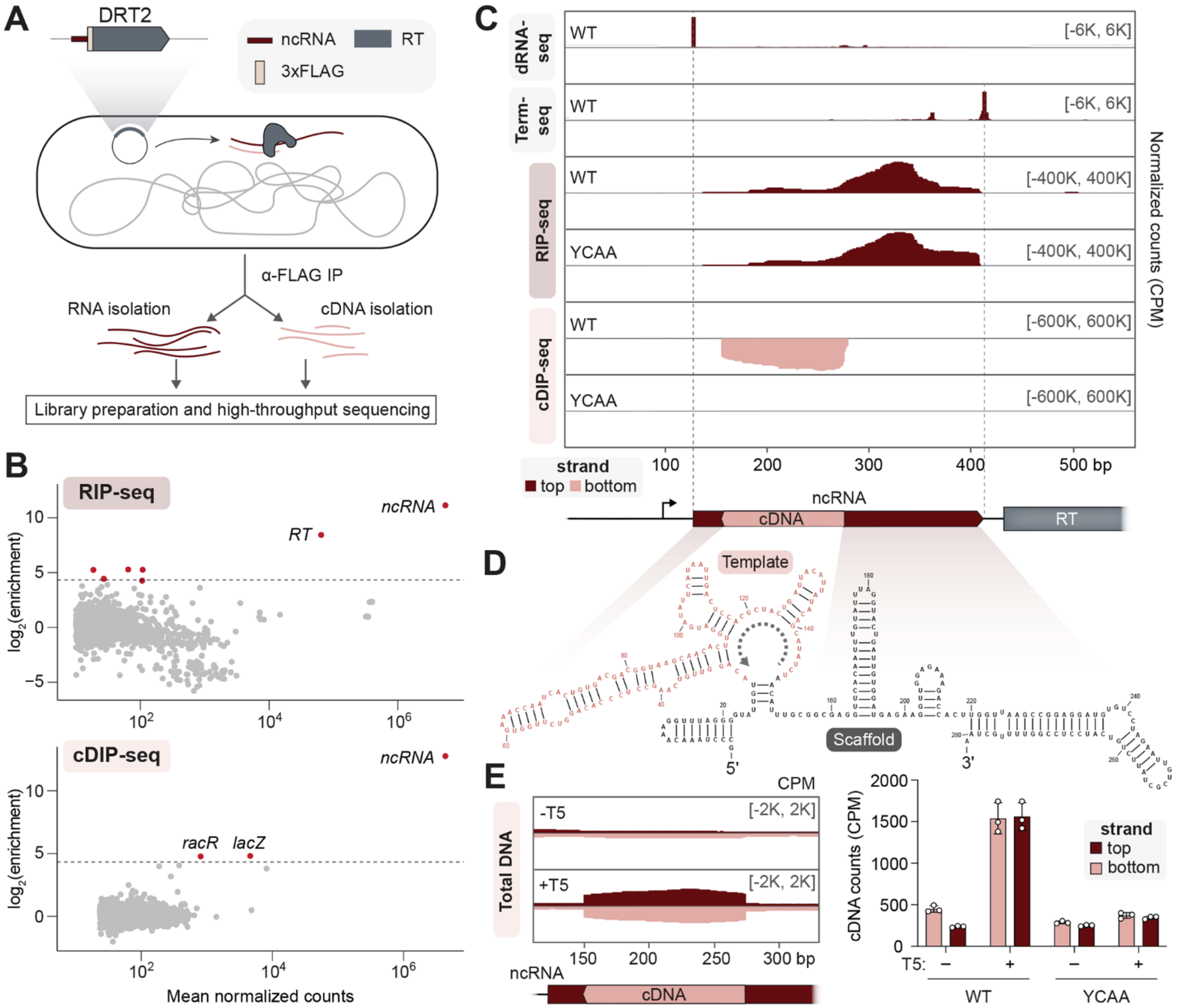

Fig. 1. Systematic discovery of DRT2 reverse transcription substrates and products in vivo. (**A**) Schematic of RNA immunoprecipitation (RIP) and cDNA immunoprecipitation (cDIP) sequencing approaches to identify nucleic acid substrates of FLAG-tagged reverse transcriptase (RT) from *Kpn*DRT2. The plasmid-encoded immune system is schematized top left. (**B**) MA plots showing the RT-mediated enrichment of RNA (top) and DNA (bottom) loci from RIP-seq and cDIP-seq experiments, relative to input controls. Each dot represents a transcript, and red dots denote transcripts with > 20-fold enrichment and false discovery rate (FDR) < 0.05. (**C**) dRNA-seq, Term-seq, RIP-seq, and cDIP-seq coverage tracks, from top to bottom, for either WT RT or a catalytically inactive RT mutant (YCAA). dRNA-seq and Term-seq enrich RNA 5′ and 3′ ends, respectively, whereas RIP-seq and cDIP-seq identify RT-associated RNA and DNA ligands. Red and pink denote top and bottom strands, respectively, and the *Kpn*DRT2 locus is shown at bottom; coordinates are numbered from the beginning of the *K. pneumoniae*-derived sequence on the expression plasmid. Data are normalized for sequencing depth and plotted as counts per million reads (CPM). (**D**) Predicted secondary structure of the *Kpn*DRT2 ncRNA. The cDNA template region is colored in pink, and the gray dotted line denotes the direction of reverse transcription. (**E**) Coverage over the *Kpn*DRT2 ncRNA locus from total DNA sequencing of cells +/– T5 phage infection (left), and bar graph of cDNA counts for the same samples alongside the YCAA mutant (right). Red and pink denote top and bottom strands, respectively; data are mean ± s.d. (*n* = 3 biological replicates).

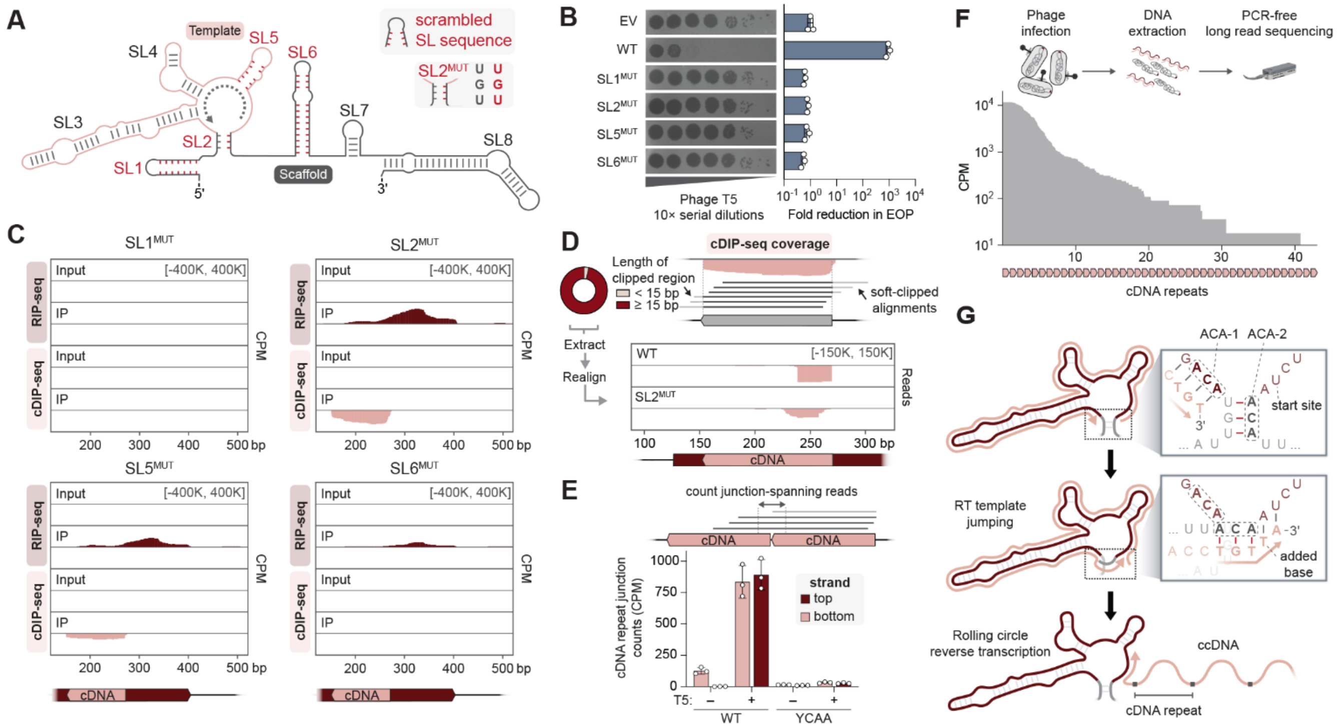

Fig. 2. Rolling circle reverse transcription generates a concatemeric cDNA product. (**A**) Schematic of *Kpn*DRT2 ncRNA secondary structure, with stem-loops (SL) numbered 1–8 and selected perturbations highlighted in red. SL1^{MUT}, SL5^{MUT}, and SL6^{MUT} correspond to ncRNA mutants in which the SL bases were scrambled, resulting in the elimination of sequence motifs and secondary structure. SL2^{MUT} abolishes base pairing within the SL2 stem by mutating the right side of the stem to its complement. Sequences of all mutants are presented in table S3. (**B**) Plaque assay showing loss of phage defense activity for all SL mutants from A (left), and bar graph quantifying the reduction in efficiency of plating (EOP, right); data are mean ± s.d. (*n* = 3 technical replicates). (**C**) RIP-seq and cDIP-seq coverage tracks for the indicated SL mutants alongside input controls, revealing a range of defects in either RNA binding, cDNA synthesis/binding, or both. (**D**) (Top) Schematic of terminal portions of cDIP-seq reads (light gray) failing to align to the cDNA reference, resulting in soft clipping and exclusion from coverage plots. A donut plot reporting the proportion of cDNA-mapping reads with the indicated lengths of 3′-clipped sequences is shown at left for *Kpn*DRT2 WT cDIP-seq. (Bottom) Mapping of 3′-soft-clipped sequences from cDIP-seq experiments back to the *Kpn*DRT2 locus, demonstrating that they derive from the cDNA 5′ end. SL2^{MUT} exhibits an aberrant pattern relative to WT. The consistent ~30-nt length of the re-mapped sequences represents the expected overhang from alignment of 150-nt sequencing reads to a ~120-nt cDNA locus. For C and D, coordinates are numbered from the beginning of the *K. pneumoniae*-derived sequence on the expression plasmid. (**E**) Schematic of sequencing reads that map across the cDNA repeat–repeat junction (top), and bar graph quantifying the abundance of junction-spanning reads from sequencing of total DNA in the indicated conditions (bottom). Red and pink denote top and bottom strands, respectively; data are mean ± s.d. (*n* = 3 biological replicates). (**F**) Schematic of long-read Nanopore sequencing workflow with DNA from phage-infected cells expressing WT *Kpn*DRT2 (top), and Nanopore read coverage over a reference sequence containing concatenated repeats of the *Kpn*DRT2 cDNA sequence (bottom). For (C), (E), and (F), data are normalized for sequencing depth and plotted as counts per million reads (CPM). (**G**) Inferred mechanism of rolling circle reverse transcription (RCRT) mediated by sequence and structural features of SL2. After synthesis of 5′-TGT-3′ templated by ACA-1 at the end of one cDNA repeat (top), the nascent DNA strand dissociates from its template and reanneals with the complementary ACA-2 following SL2 melting (middle). Template jumping initiates a subsequent round of reverse transcription, with concatenation of one cDNA repeat to the next and incorporation of one additional base at the repeat junction, ultimately leading to long concatemeric cDNA (ccDNA) products (bottom).

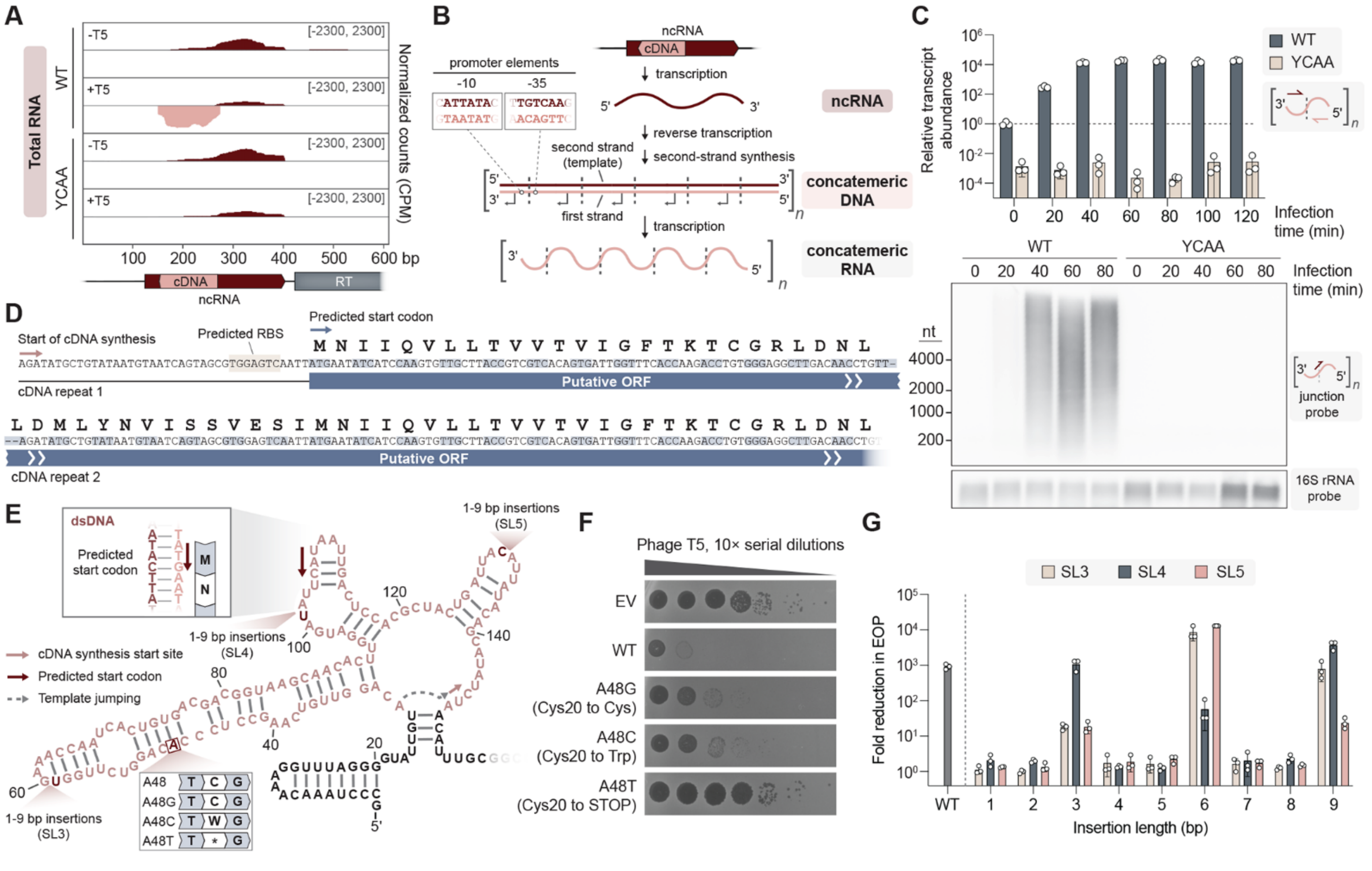

Fig. 3. The concatemeric cDNA product contains a nearly endless ORF (*neo*). (**A**) RNA-seq coverage over the *Kpn*DRT2 ncRNA locus from cells in the absence or presence of phage T5. (**B**) Model showing the consecutive production of ncRNA, concatemeric cDNA, and concatemeric RNA, all encoded by the *Kpn*DRT2 locus. Dashed lines indicate repeat–repeat junctions resulting from rolling circle reverse transcription, and the inset (top left) shows the consensus promoter formed across each junction. (**C**) Bar graph quantifying relative concatemeric RNA abundance in a phage infection time course experiment using RT-qPCR with repeat junction primers (top), and Northern blot of concatemeric RNA using a junction-spanning probe (bottom). RT-qPCR data are normalized to WT uninfected cells (t = 0); data are mean ± s.d. (n = 3 biological replicates). (**D**) Putative open reading frame (ORF) encoded by concatemeric RNA. The cDNA synthesis start site and predicted start codon are indicated (pink and blue arrows, respectively), and the predicted RBS is shaded in beige. A leucine codon spans the repeat–repeat junction. (**E**) Schematic of the cDNA template region (pink), with the predicted start codon and experimentally tested mutations indicated. (**F**) Plaque assay showing that phage defense activity is eliminated with a single-bp substitution that introduces an in-frame stop codon, but is only modestly affected by synonymous or missense mutations. EV, empty vector. (**G**) Bar graph quantifying phage defense activity for insertions within SL3, SL4, or SL5, of the indicated length. Reduction in EOP is calculated relative to an EV control; data are mean ± s.d. (n = 3 technical replicates). The only mutants that retain phage defense activity have insertion lengths of a multiple of 3 bp.

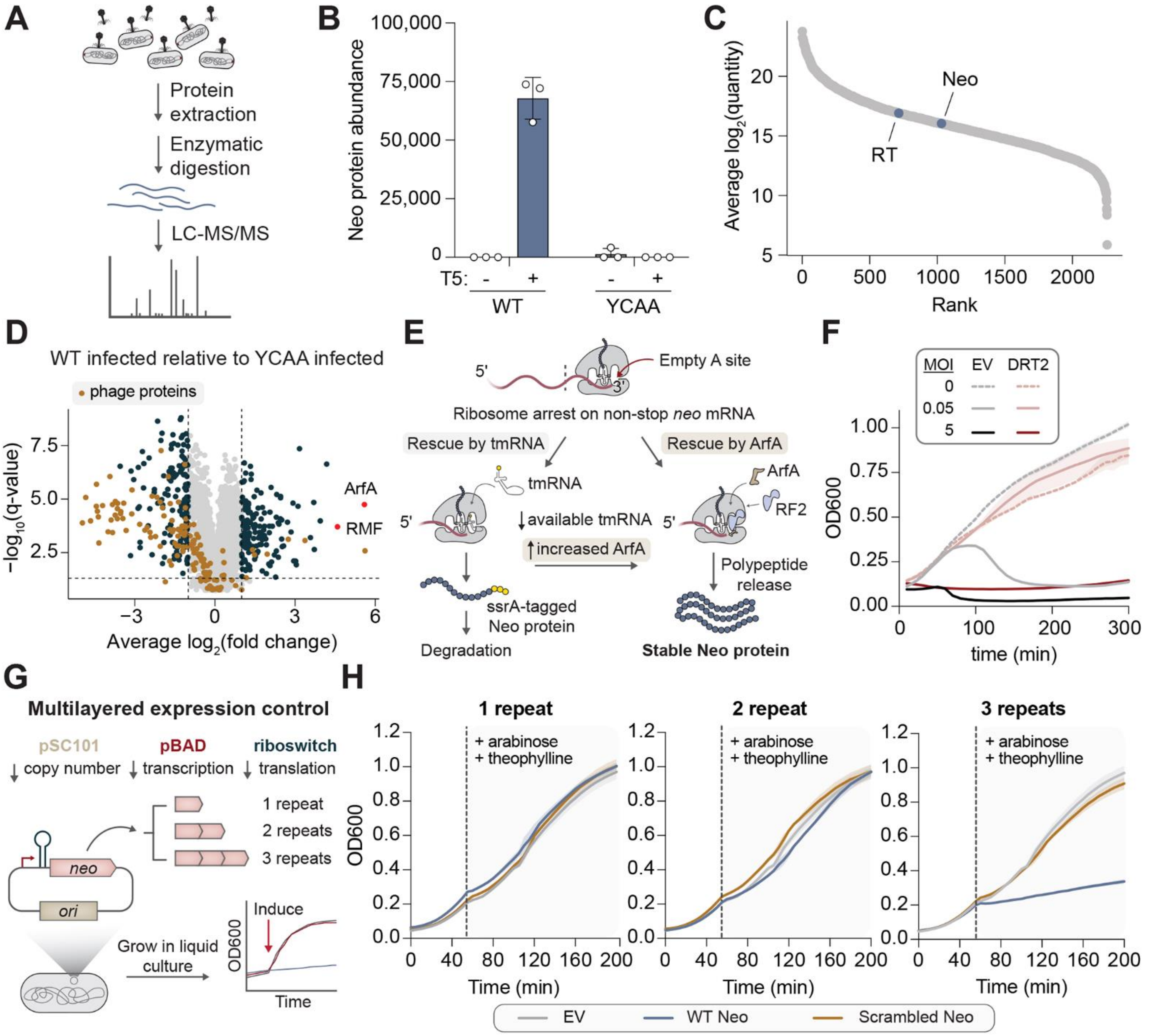

Fig. 4. Neo proteins induce programmed cellular dormancy. (**A**) Schematic of experimental approach to detect Neo in phage-infected cells by liquid chromatography with tandem mass spectrometry (LC-MS/MS). (**B**) Bar graph quantifying Neo protein quantity from cells tested in the indicated conditions. Data are mean ± s.d. (n = 3 biological replicates). (**C**) Abundance of RT and Neo proteins relative to the *E. coli* proteome in phage-infected cells expressing WT *Kpn*DRT2. (**D**) Differential protein abundance in T5-infected cells expressing *Kpn*DRT2 WT or YCAA. Phage proteins are colored in brown, and ArfA and RMF are colored in red and labeled. All other differentially abundant proteins (fold change > 2 and FDR < 0.05) are colored in dark blue. (**E**) Schematic of alternative ribosome rescue pathway mediated by ArfA, which would release Neo proteins from ribosomes stalled on non-stop *neo* mRNAs without targeting them for degradation (right), unlike the tmRNA pathway (left). (**F**) Growth curves of strains transformed with empty vector (EV) or the WT *Kpn*DRT2 system, +/− T5 phage at the indicated multiplicity of infection (MOI). Shaded regions indicate the standard deviation across independent biological replicates (*n* = 3). (**G**) Schematic of cloning and inducible expression strategy to monitor the physiological effects of Neo polypeptides of variable repeat length. (**H**) Growth curves of strains transformed with WT or scrambled Neo sequences of the indicated repeat lengths, alongside an empty vector (EV) control. The dashed line indicates the point of induction with arabinose (0.5%) and theophylline (0.5 mM). Shaded regions indicate the standard deviation across independent biological replicates (*n* = 3).

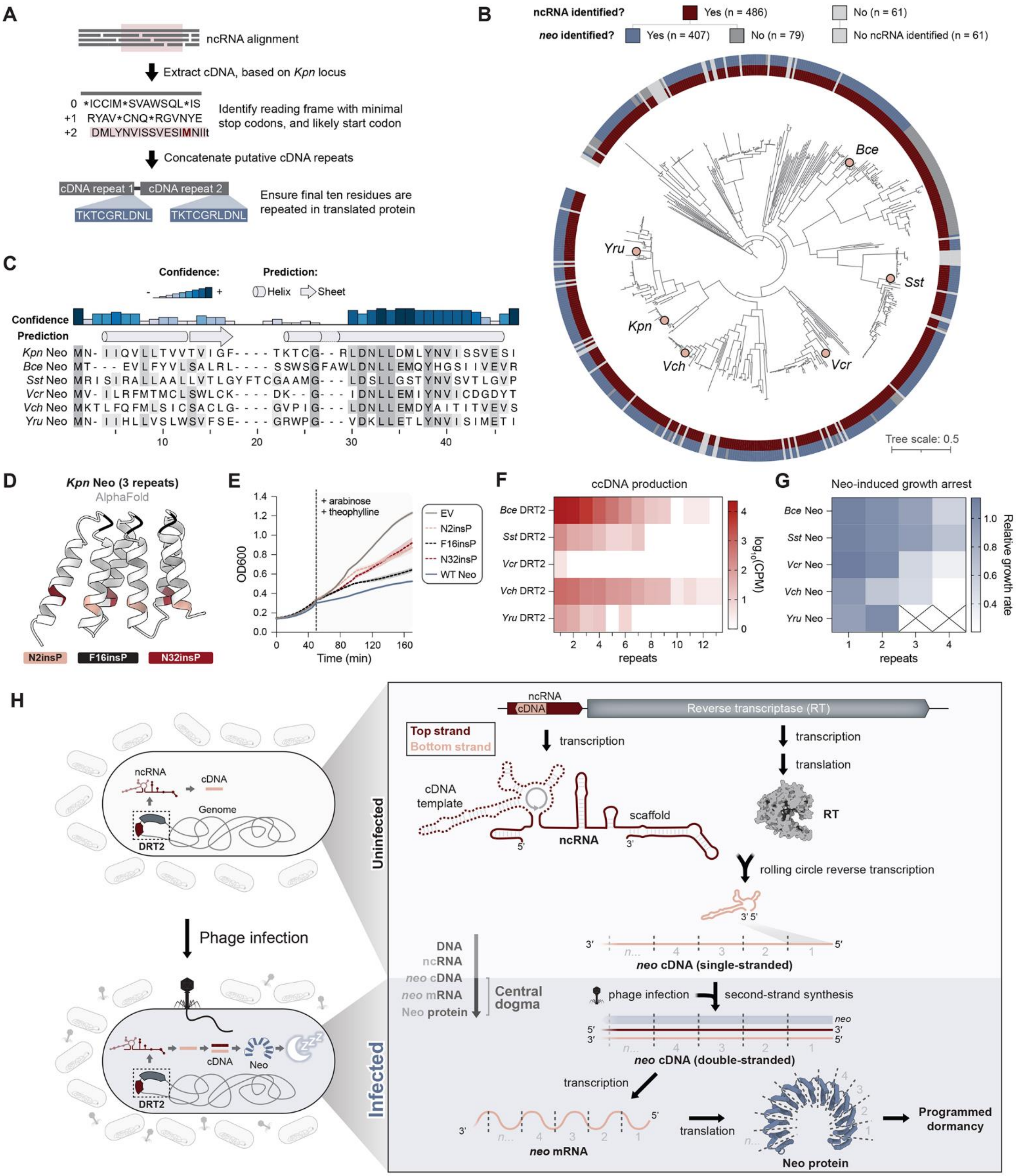
A
ncRNA alignment
Extract cDNA, based on Kpn locus
0 *ICCIM*SVAWSQL*IS
+1 RYAV*CNQ*RGVNYE
+2 DMLYNVISSVESIMNIIt
Identify reading frame with minimal stop codons, and likely start codon
Concatenate putative cDNA repeats
cDNA repeat 1
cDNA repeat 2
TKTCGRLDNL
TKTCGRLDNL
Ensure final ten residues are repeated in translated protein
B
ncRNA identified?
Yes (n = 486)
No (n = 61)
neo identified?
Yes (n = 407)
No (n = 79)
No ncRNA identified (n = 61)
Bce
Sst
Yru
Kpn
Vch
Vcr
Tree scale: 0.5
C
Confidence:
- +
Prediction:
Helix
Sheet
Confidence
Prediction
Kpn Neo MN- IIQVLLTVVTVIGF----TKTCG--RLDNLLDMLYNVISSVESI
Bce Neo MT---EVLFYVLSALRL----SSWSGFAWLDNLLEMQYHGSIIVEVR
Sst Neo MRISIRALLAALLVTLGYFTCGAAMG---LDSLLGSTYNVSVTLGVP
Vcr Neo MV-ILRFMTMCLSWLCK----DK--G---IDNLLEMIYNVICDGDYT
Vch Neo MKTLFQFMLSICSACLG----GVPIG---LDNLLEMDYAITITVEVS
Yru Neo MN-IIHLLVSLWSVFSE----GRWPG---VDKLLETLYNVISIMETI
10
20
30
40
D
Kpn Neo (3 repeats)
AlphaFold
N2insP
F16insP
N32insP
E
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
OD600
+ arabinose
+ theophylline
EV
N2insP
F16insP
N32insP
WT Neo
0
40
80
120
160
Time (min)
F
ccDNA production
Bce DRT2
Sst DRT2
Vcr DRT2
Vch DRT2
Yru DRT2
2
4
6
8
10
12
repeats
4
3
2
1
0
$\log_{10}$(CPM)
G
Neo-induced growth arrest
Bce Neo
Sst Neo
Vcr Neo
Vch Neo
Yru Neo
1
2
3
4
repeats
1.0
0.8
0.6
0.4
Relative growth rate
H
ncRNA
cDNA
Genome
DRT2
Phage infection
cDNA
Neo
DRT2
Uninfected
Infected
ncRNA
cDNA
Reverse transcriptase (RT)
Top strand
Bottom strand
transcription
transcription
translation
cDNA template
scaffold
ncRNA
5′
3′
RT
rolling circle reverse transcription
3′ 5′
3′
5′
n...
4
3
2
1
neo cDNA (single-stranded)
DNA
ncRNA
neo cDNA
neo mRNA
Neo protein
Central dogma
phage infection
second-strand synthesis
neo
5′
3′
3′
5′
neo cDNA (double-stranded)
transcription
3′
5′
neo mRNA
translation
Neo protein
Programmed dormancy

Fig. 5. Concatemeric *neo* genes and programmed dormancy are a broadly conserved phage defense mechanism. (**A**) Schematic for the automated detection of putative Neo proteins in homologous DRT2 operons. (**B**) Phylogenetic tree of DRT2 homologs, with outer rings showing the widespread presence of RT-associated ncRNAs and putative Neo proteins. Homologs selected for experimental testing are indicated with pink circles. (**C**) Multiple sequence alignment (MSA) and secondary structure prediction of Neo proteins identified in B. A single Neo repeat is shown for all homologs; shading indicates amino acid conservation. (**D**) AlphaFold prediction of a 3-repeat Neo polypeptide, showing the sites of proline mutagenesis tested in (E). Prolines were inserted C-terminal to the indicated residues within each of 3 concatenated repeats. (**E**) Growth curves of strains transformed with 3-repeat Neo constructs containing the indicated proline insertions, alongside an empty vector (EV) control. The dashed line indicates the point of induction with arabinose (0.5%) and theophylline (0.5 mM). Shaded regions indicate the standard deviation across independent biological replicates ($n = 3$). (**F**) Heat map showing the distribution of ccDNA repeat lengths in cells expressing the indicated DRT2 homologs. Data are plotted as $\log_{10}$(CPM) from Nanopore sequencing of total DNA. (**G**) Heat map showing the growth rates of cells expressing Neo homologs with the indicated repeat lengths. Growth rates are normalized to an EV control and represent the mean of independent biological replicates ($n = 3$). Empty cells with X indicate Neo expression constructs that could not be successfully cloned, presumably due to toxicity. (**H**) Model for the antiphage defense mechanism of DRT2 systems. RT enzymes bind the scaffold portion of associated ncRNAs and constitutively produce concatemeric cDNA via rolling circle reverse transcription. Phage infection triggers second-strand synthesis, yielding a double-stranded DNA molecule that is transcribed into stop codon-less, nearly endless ORF (*neo*) mRNA. Translation produces Neo proteins that potently arrest cell growth, protecting the larger bacterial population from the spread of phage.