PNAS RESEARCH ARTICLE | MICROBIOLOGY

Toxic antiphage defense proteins inhibited by intragenic antitoxin proteins

Aoshu Zhong[a], Xiaofang Jiang[b], Alison B. Hickman[c], Katherine Klier[a,1], Gabriella I. C. Teodoro[d], Fred Dyda[c], Michael T. Laub[d,e], and Gisela Storz[a,2]

Contributed by Gisela Storz; received May 2, 2023; accepted June 21, 2023; reviewed by Christopher M. Waters and Malcolm F. White

Recombination-promoting nuclease (Rpn) proteins are broadly distributed across bacterial phyla, yet their functions remain unclear. Here, we report that these proteins are toxin–antitoxin systems, comprised of genes-within-genes, that combat phage infection. We show the small, highly variable Rpn *C*-terminal domains (Rpn_S), which are translated separately from the full-length proteins (Rpn_L), directly block the activities of the toxic Rpn_L. The crystal structure of $RpnA_S$ revealed a dimerization interface encompassing α helix that can have four amino acid repeats whose number varies widely among strains of the same species. Consistent with strong selection for the variation, we document that plasmid-encoded $RpnP2_L$ protects *Escherichia coli* against certain phages. We propose that many more intragenic-encoded proteins that serve regulatory roles remain to be discovered in all organisms.

toxin-antitoxin | small protein | Rpn | iTIS

Significance

Here, we document the function of small genes-within-genes, showing they encode antitoxin proteins that block the functions of the toxic DNA endonuclease proteins encoded by the longer *rpn* (Recombination-promoting nuclease) genes. Intriguingly, a sequence present in both long and short proteins shows extensive variation in the number of four amino acid repeats. Consistent with a strong selection for the variation, we provide evidence that the Rpn proteins represent a phage defense system.

The annotation of bacterial genes generally assumes that each coding sequence directs the synthesis of one protein product. However, there are a few examples where two protein products are translated from the same gene, one product corresponding to the full open reading frame, and the second translated from an internal translation initiation start (iTIS) (1). The functions of most products of these genes-within-genes are not known, though for the few that have been characterized, the two products can have complementary or opposing functions. One example of a complementary function is provided by a *Synechocystis* type I-D CRISPR-Cas Cascade system, where the full-length Cas10d protein forms a complex with Cas11d, which is translated from an iTIS (2). The complex is required for specific DNA binding by the type I-D Cascade complex, and without Cas11d, the Cascade complex has little or no DNA binding activity. Recent experiments to examine transcriptome-wide ribosome binding (ribo-seq) in the presence of inhibitors that trap the ribosome on translation initiation sites suggested that there are more iTIS than initially considered (3, 4). The five *rpn* (recombination-promoting nuclease) genes of *Escherichia coli* each have an iTIS that could potentially direct the synthesis of small proteins corresponding to the variable *C*-terminal tail (Fig. 1*A* and *SI Appendix*, Fig. S1*A*).

Rpn proteins, which are members of the diverse PD-(D/E)XK phosphodiesterase superfamily (6), have been proposed to be involved in horizontal gene transfer based on the report that the conserved *N*-terminal domain has homology to transposases (7). Previous studies showed that overexpression of *E. coli* Rpn proteins reduced cell viability in a *recA*$^-$ background and induced the DNA damage (SOS) response in a *recA*$^+$ host (5). Furthermore, the purified RpnA protein was shown to possess DNA endonuclease activity. These observations support the idea that Rpn proteins are DNA-mobilizing enzymes, but the physiological role of this DNA cleavage activity is unknown. Thus, we set out to characterize the long, full-length proteins (Rpn_L) as well as the small *C*-terminal proteins (Rpn_S), which we hypothesized are translated from the iTIS and could act in conjunction with the Rpn_L proteins.

Results

***rpn* Genes Move by Horizontal Gene Transfer but Do Not Encode Conventional Transposases.** Phylogenetic analysis revealed that members of the *rpn* orthologous gene cluster (COG5464) are widely distributed and are found in 34 bacterial and two archaeal phyla (Fig. 1*B*). Their distribution is strongly nonuniform indicating that *rpn* genes are prone to frequent horizontal gene transfer. Additionally, the number of *rpn* genes can vary between different strains of the same species. For example, no *rpn* genes were found in the genome of *Bacillus thuringiensis serovar kurstaki str. T03a001* (assembly accession: GCF_000161575.1) while 26 copies are present in *Bacillus thuringiensis serovar yunnanensis* (assembly accession: GCF_002147825.1).

Author affiliations: [a]Division of Molecular and Cellular Biology, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, MD 20892; [b]Intramural Research Program, National Library of Medicine, NIH, Bethesda, MD 20894; [c]Laboratory of Molecular Biology, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD 20892; [d]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; and [e]HHMI, Massachusetts Institute of Technology, Cambridge, MA 02139

Author contributions: A.Z., X.J., A.B.H., K.K., F.D., M.T.L., and G.S. designed research; A.Z., X.J., A.B.H., K.K., G.I.C.T., and F.D. performed research; A.Z., X.J., A.B.H., K.K., F.D., M.T.L., and G.S. analyzed data; and A.Z., X.J., A.B.H., and G.S. wrote the paper.

Reviewers: C.M.W., Michigan State University; and M.F.W., University of St Andrews.

The authors declare no competing interest.

[1]Present address: Freshwater and Marine Sciences, University of Wisconsin-Madison, Madison, WI 53706.

[2]To whom correspondence may be addressed. Email: storzg@mail.nih.gov.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2307382120/-/DCSupplemental.

Published July 24, 2023.

rpn genes are commonly annotated as encoding a "putative transposase" and genomic synteny among strains differing in *rpn* genes suggests that the *rpn* genes were recently acquired or lost as a single gene cluster, a pattern similar to insertion sequences (8). However, unlike typical insertion sequences, the coding regions of *rpn* genes are not flanked by identifiable inverted terminal repeats or target-site duplications. Furthermore, comparative genomic analysis of closely related genomes with and without *rpn* genes shows that the ends of the genomic region containing the *rpn* gene vary, contrasting with insertion sequences where the ends are generally defined. In genomes where *rpn* gene arrays were found, the arrays appear to be generated by tandem direct duplications of individual genomic regions rather than by transposition of different *rpn* genes into the same genomic locus given the absence of target-site duplications and the presence of different types of duplications (*SI Appendix*, Fig. S1*B*). Although the TnsA protein of the heteromeric Tn7 transposase has a PD-(D/E)XK catalytic domain, there are no known autonomous transposases with this protein fold (9, 10), and we find no evidence that *rpn* genes are capable of autonomous transposition.

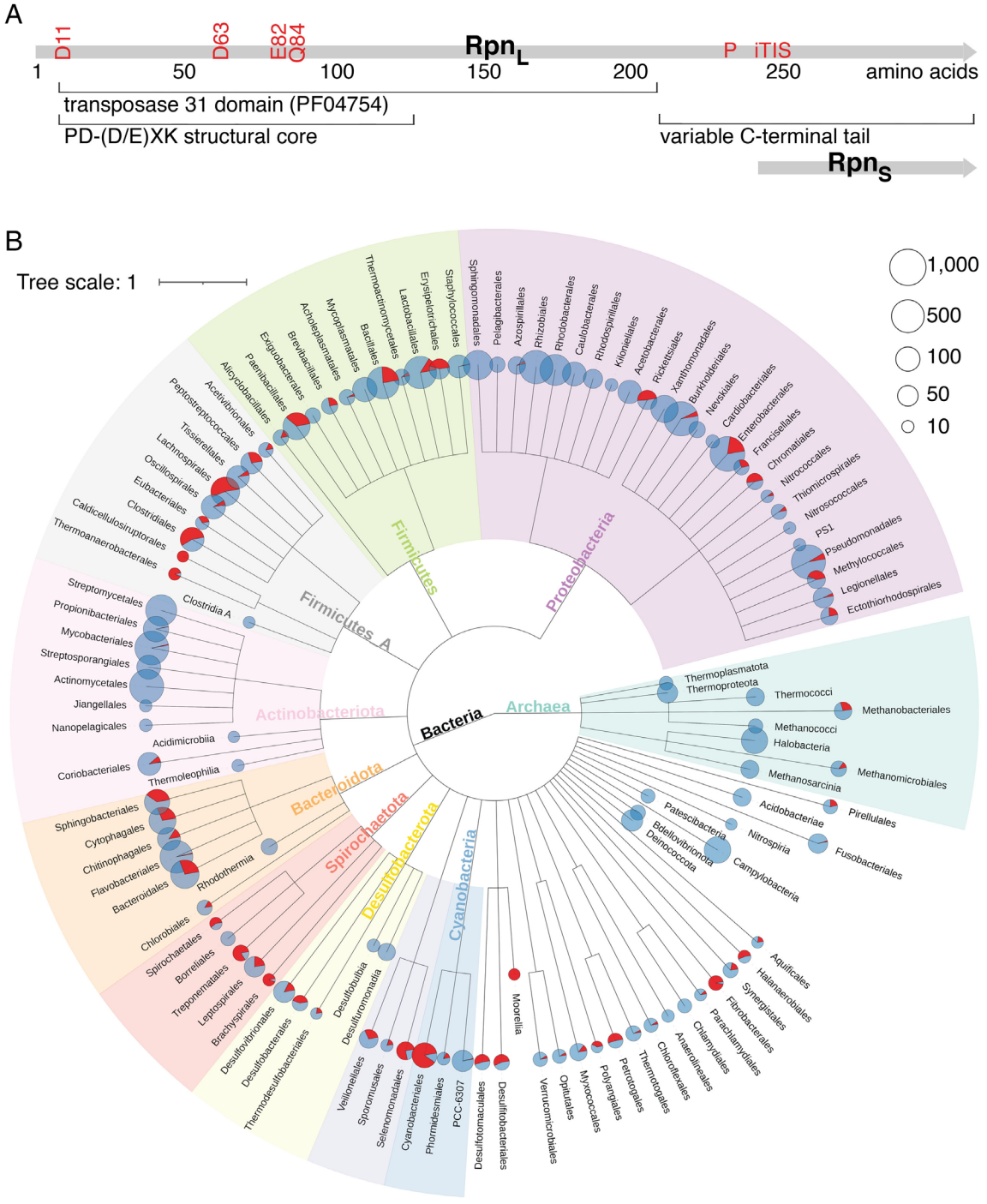

Fig. 1. Broadly-distributed *rpn* genes move by horizontal gene transfer. (*A*) Diagram of *rpn* genes adapted from ref. 5. Positions of catalytic residues for RpnA and internal promoter (P) and iTIS are indicated in red font. (*B*) Species distribution of *rpn* orthologs. Clades with more than 15 children are displayed. The pie charts represent the percentage of species with (red) and without (blue) a *rpn* ortholog. The size of the pie chart represents the number of species in that clade.

We therefore think that it is highly unlikely that Rpn proteins function as transposases.

Small Proteins Are Translated from Within *rpn* Genes. We next wanted to determine the role of the iTIS (Fig. 2*A*) and ribosome profiling signal detected for all five *E. coli rpn* genes after treatment with the translation inhibitors Onc112 or retapamulin (3, 4) (Fig. 2 *B, C, D*, and *E* and *SI Appendix*, Fig. S2*A*). To see whether small proteins corresponding to the *C*-terminal tails are synthesized, we introduced a sequential peptide affinity (SPA) tag on the chromosome upstream of each *rpn* stop codon, permitting detection via immunoblot analysis. In cells grown to exponential ("exp") or stationary ("stat") phase, all five small proteins corresponding to the Rpn_S were observed at different levels (Fig. 2 *F, G*, and *H* and *SI Appendix*, Fig. S2*B*). Among the Rpn_L, only $RpnB_L$, $RpnC_L$, and $RpnE_L$ were detected under the growth conditions examined (Fig. 2 *G* and *H*). The higher levels of the Rpn_S proteins are consistent with previous differential RNA-seq (dRNA-seq) data (11), which shows a transcription start site within all *rpn* genes for a short transcript expressed at higher levels than the full-length mRNA. The introduction of mutations in the ribosome binding site (RBS) and predicted iTIS with minimal change to the *rpnA* coding sequence ($RpnA_{K240R+E241D+M244L}$, hereafter denoted $RpnA_{L*}$) on the *E. coli* chromosome (Fig. 2*A*) abolished $RpnA_S$ expression (Fig. 2*F*), confirming that the $RpnA_S$ protein is translated from the predicted iTIS.

We also wanted to determine whether an Rpn_S protein is expressed from *rpn* genes from other organisms and thus cloned an *rpn* gene from the Gram$^+$ bacterium *Clostridioides difficile* into a plasmid that we introduced into *E. coli*. For this construct, we also observed expression of an Rpn_S protein, which again was eliminated by the introduction of mutations of the predicted RBS and iTIS (*SI Appendix*, Fig. S2 *C* and *D*). This observation suggests coexpression of Rpn_L and Rpn_S proteins is broadly conserved.

Varied Expression of Small Proteins. In an analysis of the *rpn* genes present in diverse *E. coli* strains (*SI Appendix*, Fig. S3 *A* and *B*), we observed that while most of the genes are encoded on the chromosome, some are encoded on plasmids. Intriguingly, on both the chromosome and plasmids, the Rpn_S protein can be expressed from either within an *rpn* gene or as a separate gene, annotated as encoding proteins with the DUF4351 domain of

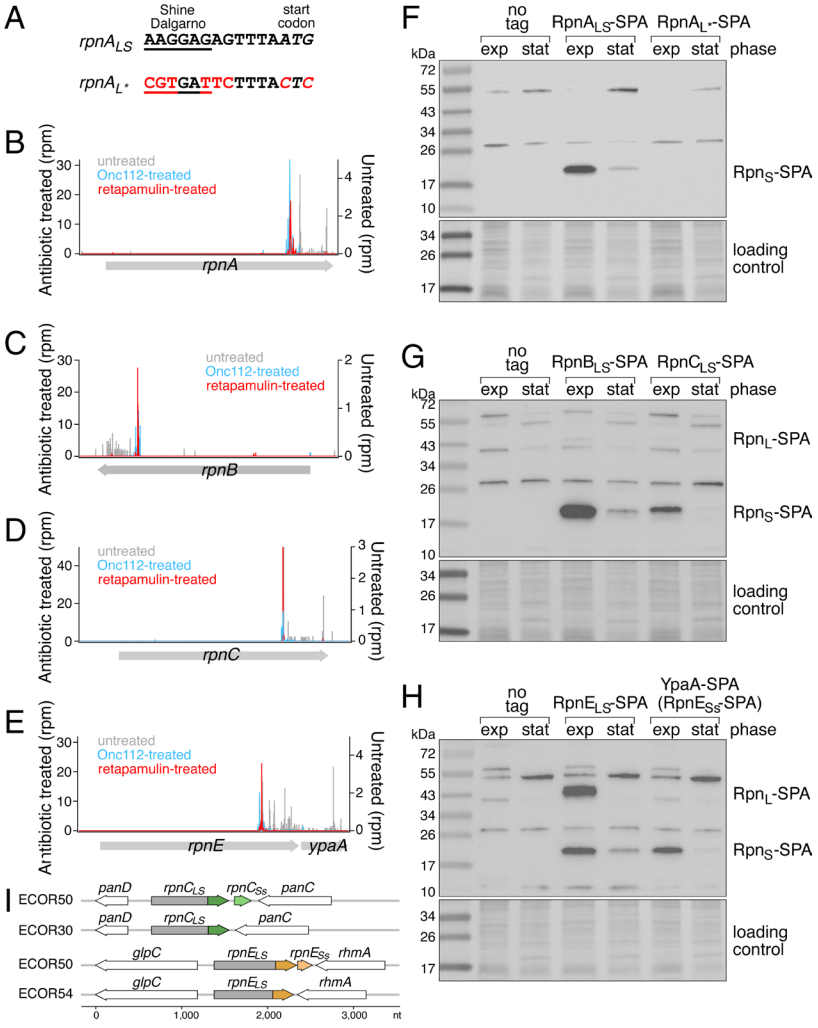

Fig. 2. The *rpn* genes encode smaller proteins (Rpn_S). (*A*) Sequence of $RpnA_S$ RBS, and mutations to eliminate $rpnA_S$ ribosome binding and start codon. Browser images of ribosome profiling data for *rpnA* (*B*), *rpnB* (*C*), *rpnC* (*D*), *rpnE-ypaA*/$rpnE_{Ss}$ (*E*). Ribosome density for an untreated control (gray) and cells treated with Onc112 (blue) (4) or retapamulin (red) (3) are shown. Immunoblot analysis of the levels of SPA-tagged $RpnA_S$ (*F*) and $RpnB_S$, and $RpnC_S$ (*G*) $RpnE_S$ and YpaA/$RpnE_{Ss}$ (*H*). *E. coli* MG1655 strains were grown to exponential (exp) and stationary (stat) phase in LB. The SPA tag was detected with monoclonal anti-FLAG M2-peroxidase (HRP) antibody. Ponceau S staining of membranes served as loading controls. (*I*) Diagram showing an example of insertion of a $rpnC_{Ss}$ gene (light green) and a $rpnE_{Ss}$ gene (light mustard) downstream of the *rpnC* and *rpnE* genes, respectively, among strains of *E. coli*.

unknown function (PF14261) in some species (Dataset S1). For example, we noted that a second copy of $RpnE_S$ was encoded by the *ypaA* gene (denoted $rpnE_{Ss}$ here) downstream of the *rpnE* gene. $RpnE_{Ss}$ was detected upon SPA tagging indicating the protein is synthesized (Fig. 2 *H* and *I*). A second copy of $rpnC_S$ is similarly found downstream of *rpnC* in some *E. coli* strains (Fig. 2*I*). Phylogenetic analysis, as demonstrated in *SI Appendix*, Fig. S3*C*, revealed that the downstream small protein homologs $RpnC_{Ss}$ and $RpnE_{Ss}$ cluster together with their respective upstream small protein homologs, $RpnC_S$ and $RpnE_S$, rather than clustering with each other. This observation suggests that the origin of the downstream small proteins occurred independently, with $RpnC_{Ss}$ originating from RpnC and $RpnE_{Ss}$ originating from RpnE. The rpn_{Ss} gene also can be lost through the deletion of the encoding region or fusion of the *rpn N*-terminal domain and rpn_{Ss} encoding region.

For further analysis of Rpn protein function, we expressed both RpnA and RpnB proteins from a rhamnose-inducible promoter as done previously (here denoted as expressing $RpnA_{LS}$ or $RpnB_{LS}$) (5) or from their own promoters on a low-copy plasmid. We also expressed plasmid-encoded $RpnP2_{LS}$ from its own native promoter on a low-copy plasmid. To first compare the relative levels of the proteins expressed from these constructs, we again integrated an SPA tag upstream of the stop codon. Intriguingly, the relative levels of Rpn_L and Rpn_S varied widely with higher levels of Rpn_L than Rpn_S from the rhamnose-inducible promoter than the native promoter, except for $RpnP2_L$ (*SI Appendix*, Fig. S2 *E* and *F*). The reasons for the differential expression are not known but suggest additional regulation.

Rpn_S Proteins Block Rpn_L-Dependent Growth Inhibition. Although previous studies showed that Rpn proteins expressed from the rhamnose promoter reduce cell viability when overexpressed (5), Rpn_S likely was unknowingly coexpressed from these constructs, complicating the interpretation of these results. To deconvolute the effects of the two proteins, we constructed plasmids carrying *rpnA* or *rpnB* with mutated iTIS and RBS sequences ($RpnA_{L^*}$ or $RpnB_{L^*}$) to fully abolish the expression of $RpnA_S$ or $RpnB_S$. When assessed by cell density measurements in liquid culture, there was no effect on *E. coli* ER2170 growth (Fig. 3*A*) with the empty plasmid (black) or $RpnA_{LS}$ overexpression (green), while a defect was observed when expressing $RpnA_{L^*}$ (red). For the RpnB constructs (Fig. 3*B*), we observed some growth defect for cells expressing $RpnB_{LS}$ (green), but the effect was even stronger for overexpression of $RpnB_{L^*}$ alone (red). For RpnP2 (Fig. 3*C*), all constructs for $RpnP2_{L^*}$ expressed from its native promoter had additional inactivating site mutations indicating strong selection against expression of $RpnP2_L$ alone. We were only able to obtain the intact $RpnP2_{L^*}$ construct when the *rpnP2* gene was cloned behind the P_{BAD} promoter, which is repressed by glucose and activated by arabinose. In the presence of glucose, the $RpnP2_{L^*}$ cells showed a slight growth defect (top panel, red). In contrast, growth was strongly inhibited by $RpnP2_{L^*}$ upon the addition of arabinose (bottom panel, red), while $RpnP2_{LS}$ cells (green), which also expressed the $RpnP2_S$ protein, grew like vector control cells (black) under both conditions. Thus, Rpn proteins have the properties of toxin–antitoxin systems (12) with Rpn_L proteins inhibiting growth and Rpn_S proteins serving as the cognate antitoxin.

Rpn_S Proteins Block Rpn_L-Dependent SOS Induction. To test whether the Rpn_S proteins can block the DNA damage caused by Rpn_{LS} overexpression reported previously (5), we employed a two-plasmid system to separately overexpress Rpn_S. We observed that, when the proteins are overexpressed from the rhamnose-inducible P_{rhaB} promoter, $RpnA_{LS}$ and $RpnB_{LS}$ both induce the SOS response (*SI Appendix*, Fig. S4*A*), as monitored by the chromosomally encoded, DNA damage-inducible P_{dinD}-*lacZ* reporter (5). For these experiments, induction of $RpnA_{LS}$ and $RpnB_{LS}$ was titrated to allow for overexpression without significantly impacting growth. Coexpression of $RpnA_S$ or $RpnB_S$ reduced SOS induction in the corresponding Rpn_{LS} strain (Fig. 3 *D* and *E*). This block is no longer observed when a stop codon was introduced in the $RpnA_S$ (at E25) or $RpnB_S$ (at D29) coding sequence indicating that the repressive effect is due to the protein and not due to overexpression of the RNA. The repressive effect of each small protein is specific to the Rpn_L protein encoded by the same sequence as overexpression of $RpnA_S$ did not block $RpnB_{LS}$-dependent *dinD-lacZ* induction while $RpnB_S$ did not block $RpnA_{LS}$-dependent induction. These observations further support the conclusion that $RpnA_S$ and $RpnB_S$ specifically block the activities of the corresponding larger protein.

$RpnA_S$ Blocks $RpnA_L$-Dependent DNA Cleavage In Vitro. To directly examine the effect of the Rpn_L and Rpn_S proteins on DNA cleavage, RpnA derivatives, which were easier to express given that they were the least toxic, were overexpressed as *C*-terminally tagged intein fusion proteins and purified. When the WT *rpnA* gene was cloned into an overexpression vector, the *C*-terminally tagged $RpnA_L$ and $RpnA_S$ (61 kDa and 33 kDa, respectively) were both detected (Fig. 3*F*) and purified together (Fig. 3*G*). Thus, we also overexpressed and purified the $RpnA_{L^*}$ and $RpnA_S$ proteins separately (Fig. 3 *F* and *G*). In assays for DNA cleavage activity using both double-stranded (Fig. 3*H*) and single-stranded (*SI Appendix*, Fig. S4*B*) DNA substrates, the $RpnA_{L^*}$ protein had strong DNA cleavage activity (lane 2). Consistent with the conclusion that $RpnA_S$ blocks $RpnA_L$ activity, we found that $RpnA_{LS}$ has significantly less endonuclease activity (lane 3), and the addition of $RpnA_S$ to $RpnA_{L^*}$ completely blocks the cleavage (lanes 5 and 6). Interestingly, the DNA cleavage activities of both $RpnA_{L^*}$ and $RpnA_{LS}$ are higher at a more alkaline pH (Fig. 3*I*) as was previously observed for $RpnA_{LS}$ (5), suggestive of a role for the cleavage activity at higher pH or other specific growth condition with similar consequences.

$RpnA_S$ Forms a Complex with $RpnA_L$. To examine whether $RpnA_S$ is inhibiting $RpnA_L$ through a direct interaction, the oligomerization states of the purified proteins were assessed by size exclusion chromatography (SEC) (Fig. 4*A*). The individually purified $RpnA_{L^*}$ (33.3 kDa) and $RpnA_S$ (5.4 kDa) proteins eluted as distinct peaks. However, when $RpnA_{L^*}$ and $RpnA_S$ were mixed, a stable $RpnA_{L^*}$–$RpnA_S$ complex was observed that eluted at the same volume (arrow) as the native $RpnA_{LS}$ complex. SDS/PAGE analysis of the peak fractions confirmed that both $RpnA_{L^*}$ and $RpnA_S$ are in the corresponding fractions (*SI Appendix*, Fig. S5*A*). In a multiangle light scattering coupled with SEC (SEC-MALS) experiment, the molecular weight of $RpnA_S$ was determined to be 10.2 ± 0.2 kDa, consistent with a dimer. The molecular weight of the $RpnA_{LS}$ complex was determined to be 74.1 ± 0.5 kDa by SEC-MALS, consistent with a tetramer of two $RpnA_S$ subunits and two $RpnA_L$ subunits. Intriguingly, the $RpnA_{LS}$ complex is stable at pH 7.0 and 8.0, yet the subunits dissociate at pH 9.0 and 10.0 (*SI Appendix*, Fig. S5*B*), paralleling the increase in DNA cleavage activity observed for $RpnA_{LS}$ (Fig. 3*I*).

Rpn_S Vary by Four Amino Acid Repeats and Comprise an Oligomerization Domain. A striking feature of the amino acid sequence present in both the Rpn_L and Rpn_S proteins, which is observed from the alignment of homologs from different strains of

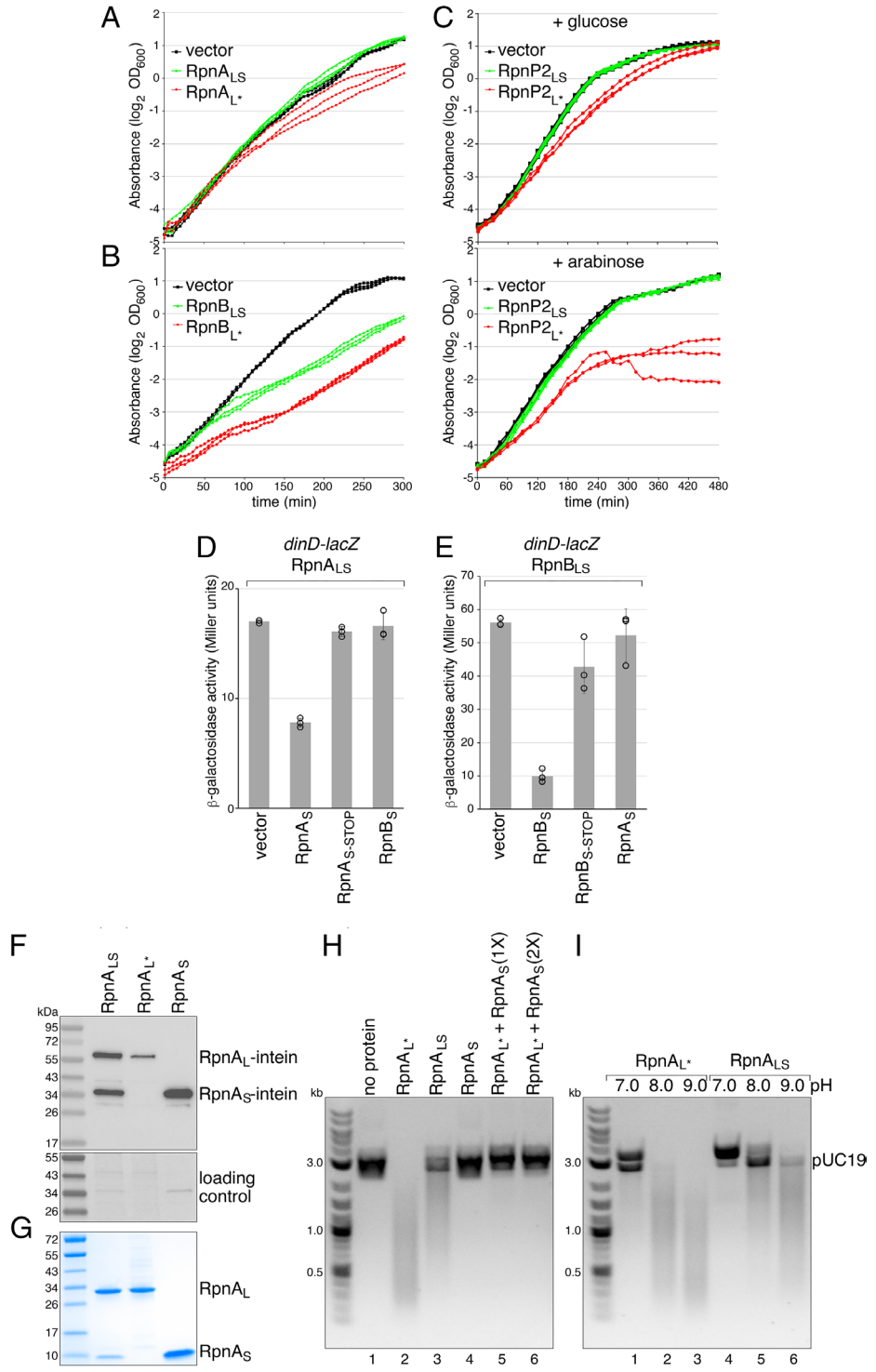

Fig. 3. Rpn_S proteins function as antitoxins. $RpnA_L$ (*A*) and $RpnB_L$ (*B*) inhibit *E. coli* growth. *E. coli* ER2170 cells harboring indicated plasmids were grown in LB. (*C*) $RpnP2_L$ inhibits *E. coli* growth. *E. coli* MG1655 cells harboring indicated plasmids were grown in LB with added glucose or arabinose. For *A*–*C*, three biological replicates are shown. (*D*) $RpnA_S$, but not $RpnB_S$, blocks $RpnA_L$ induction, and (*E*) $RpnB_S$, but not $RpnA_S$, blocks $RpnB_L$ induction of the *dinD-lacZ* reporter of the SOS DNA damage response. $RpnA_{LS}$ and $RpnB_{LS}$ are overexpressed from the rhamnose-inducible P_{rhaB} promoter. $RpnA_S$ and $RpnB_S$ are overexpressed from the arabinose-inducible P_{BAD} promoter. For *D* and *E*, average of three independent biological repeats is given with SD. For *A*–*E*, the ER2170 or MG1655 strain backgrounds were WT for the *rpn* genes. (*F*) Immunoblot analysis of intein-tagged $RpnA_{LS}$, $RpnA_{L*}$, and $RpnA_S$ overexpression. Ponceau S staining of membrane served as loading control. (*G*) Coomassie-stained Tris-glycine SDS-PAGE gel of purified $RpnA_{LS}$, $RpnA_{L*}$, and $RpnA_S$. (*H*) Purified $RpnA_S$ blocks dsDNA endonuclease activity of $RpnA_{L*}$. Indicated purified proteins were incubated with pUC19. (*I*) $RpnA_L$ activity is pH-dependent. Purified $RpnA_{L*}$ (lanes 1 to 3) or $RpnA_{LS}$ (lanes 4 to 6) were incubated with pUC19 at pH 7.0, pH 8.0, or pH 9.0. Images in *H* and *I* are inverted from the original. Each of the nuclease assays was repeated at least twice.

the same species, is variation by four amino acid repeats generally composed of one hydrophobic amino acid and three charged residues or glycine (Fig. 4*B* and *SI Appendix*, Fig. S6). This is seen for all the *E. coli* Rpn_S proteins including Rpn_{Ss} proteins as well as those found in distantly related bacteria such as *Bacillus cereus*, *Bacteroides fragilis*, *Clostridioides difficile*, and *Leptospira interrogans*.

Given the interaction between $RpnA_L$ and $RpnA_S$, we hypothesized that this hypervariable four amino acid repeat region present

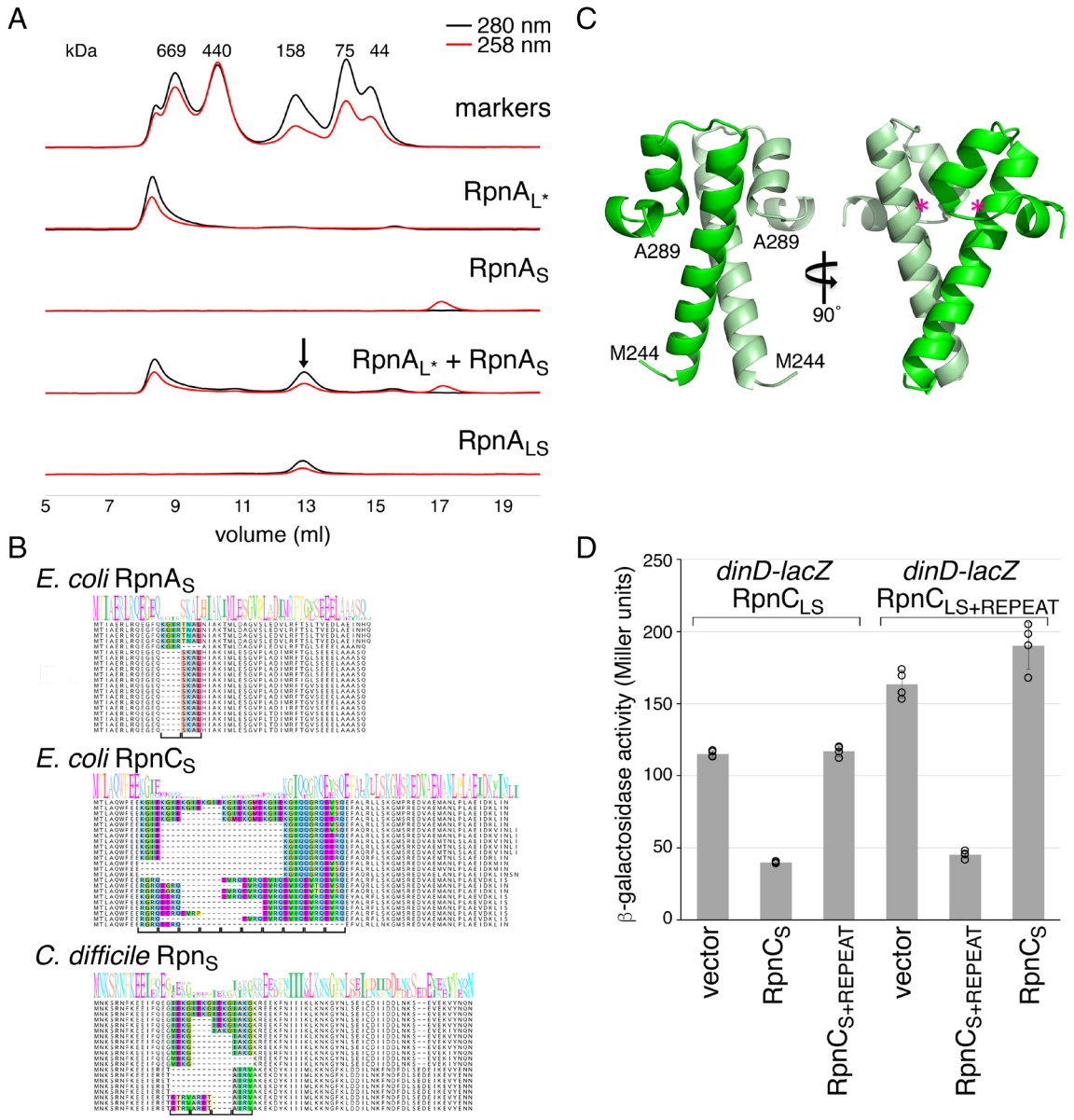

Fig. 4. RpnA$_L$ and RpnA$_S$ form a complex. (*A*) Size exclusion chromatograms of indicated proteins on a Superdex 200 column. Although the RpnA$_{LS}$ peak migrates similarly to the 158 kDa marker peak, SEC-MALS analysis, which is less influenced by protein shape, indicates that the complex has a molecular weight of 74.1 ± 0.5 kDa. The RpnA$_{L*}$ protein migrates as a large molecular weight complex indicative of an aggregate, but we do not observe precipitates of this protein. (*B*) Sequence alignments for *E. coli* RpnA$_S$ and RpnC$_S$ as well as *Clostridioides difficile* Rpn$_S$. The sequence logo was built based on all the sequences for each group though only representative individual sequences are shown. The four amino acid repeat that varies between strains is colored in the alignment, with a bracket indicating the number of repeats. The alignments were manually curated. (*C*) Crystal structure of RpnA$_S$. Pink asterisks represents approximate position where four amino acid insertions occur (H261 in RpnA$_L$). (*D*) Only RpnC$_S$ with same number of four amino acid repeats blocks RpnC$_L$ induction of *dinD-lacZ*. RpnC$_{LS}$ and RpnC$_{LS+REPEAT}$ are expressed from the rhamnose-inducible P$_{rhaB}$ promoter, while RpnC$_S$ and RpnC$_{S+REPEAT}$ are expressed from the arabinose-inducible P$_{BAD}$ promoter. For the REPEAT derivatives, a four amino acid repeat "KGIE" is inserted in the same position. Average of four independent biological repeats is given with SD.

in both forms of the Rpn proteins might comprise an oligomerization domain. To gain insights into the domain, we solved the structure of RpnA$_S$ at a resolution of 1.9 Å using X-ray crystallography (Fig. 4*C* and *SI Appendix*, Table S1). Consistent with the SEC-MALS data, RpnA$_S$ is a dimer in the crystals in which the two monomers each fold as a compact three α-helix bundle. The long first helix of each monomer interacts with the other, burying a total surface area of 976 Å^2. In all Rpn$_S$ proteins where they have been inserted, the four amino acid repeats begin 8 to 20 residues after the initiating methionine (*SI Appendix*, Fig. S6), which places the insertion point for the hypervariable region within the first α-helix of Rpn$_S$. Based on three-dimensional structure predictions by AlphaFold2 (13), the inserted repeats most likely increase the length of the α-helix rather than disrupt it (*SI Appendix*, Fig. S7).

To examine the consequences of changing the number of four amino acid repeats, we generated derivatives of the RpnC$_{LS}$ and RpnC$_S$ proteins, which have the highest number of repeats among *E. coli* strains, by adding one repeat. Interestingly, when assessed in the SOS response assay, the RpnC$_{S+REPEAT}$ can no longer counteract RpnC$_{LS}$ (Fig. 4*D*). Conversely, RpnC$_S$ cannot counteract RpnC$_{LS+REPEAT}$, but repression is restored when the two proteins with the extra repeat are expressed together. These observations are consistent with the repeat regions affecting the binding between the two Rpn$_S$ monomers and providing specificity for the interaction between the Rpn$_L$ and Rpn$_S$ proteins.

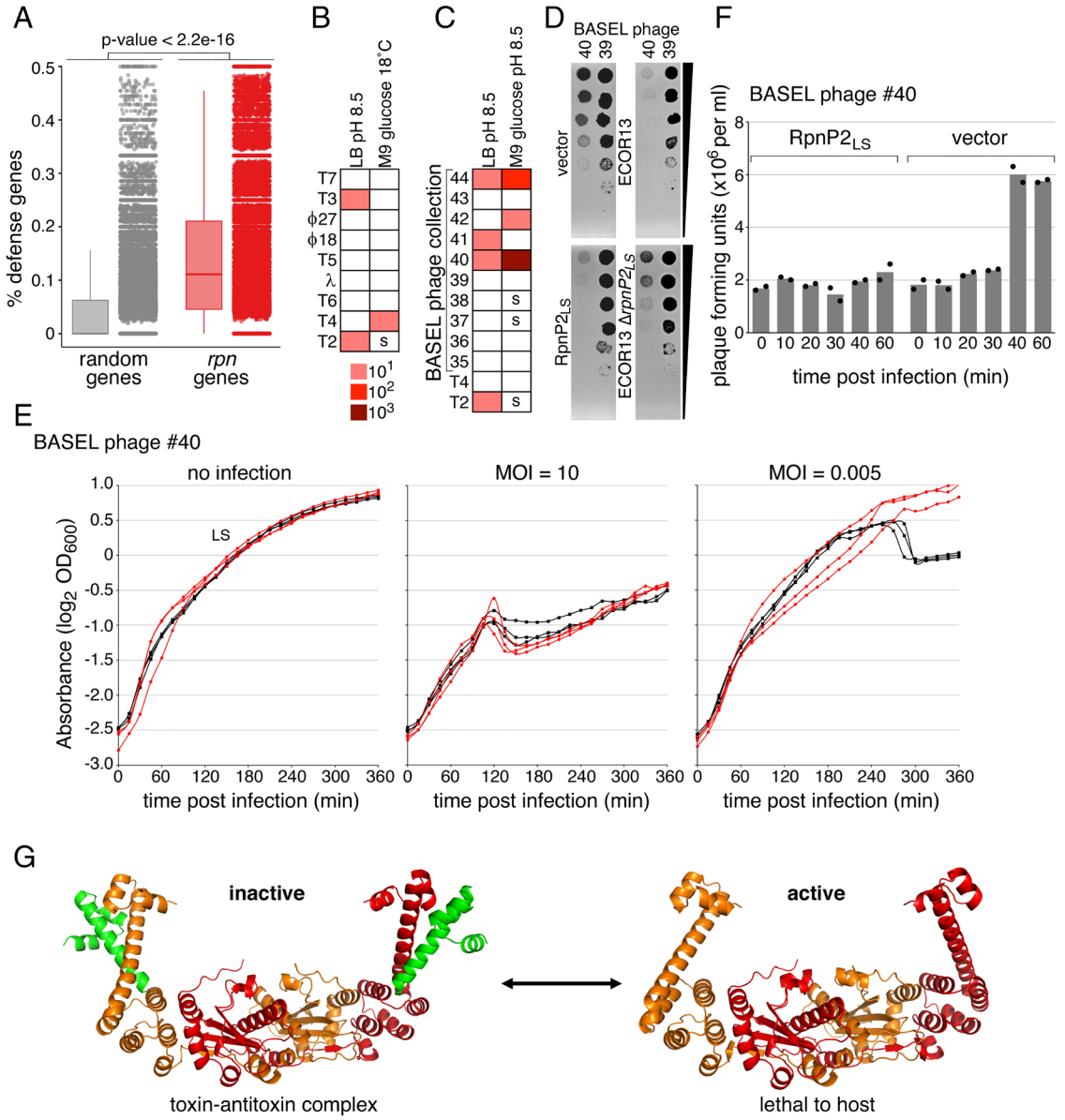

Fig. 5. Rpn proteins block phage infection. (*A*) Enrichment of phage defense genes in the vicinity of *rpn* genes compared to random genes. Box plots (*Left*) and dot plots (*Right*) display the percentage and distribution of defense-associated genes in proximity to these genes, with *P*-values indicating the significance levels determined by the Mann–Whitney *U* Test. (*B*) Efficiency of plaquing (EOP) for the indicated phages infecting cells producing $RpnP2_{LS}$ grown at 30 °C on LB medium buffered to pH 8.5 and 18 °C on M9 glucose medium pH 7.1. (*C*) EOP for the indicated phages infecting cells producing $RpnP2_{LS}$ grown in LB and M9 glucose, pH 8.5. For *B* and *C*, "s" that denotes smaller plaques were observed. (*D*) Images of BASEL phage #40 and #39 plaques for MG1655 cells carrying vector control or producing $RpnP2_{LS}$ (*Left*) or ECOR13 or ECOR13 Δ*rpnP2*$_{LS}$ cells (*Right*), from *SI Appendix*, Fig. S8. The back wedge denotes the 10-fold dilution series. (*E*) Growth of vector control or $RpnP2_{LS}$ expressing cells infected with BASEL phage #40 (at indicated MOIs) in LB pH 8.5. Three biological replicates are presented. The partial regrowth of the strains infected at a MOI of 10 could be due to appearance of suppressors. For *A*–*E*, the MG1655 strain background was WT for the *rpn* genes. (*F*) Plaque-forming units (PFU) per mL of BASEL phage #40 (MOI = 0.005) used to infect vector control or $RpnP2_{LS}$ expressing cells at 0, 10, 20, 30, 40, and 60 min postinfection. Individual data points and average of two biological replicates are shown. (*G*) Model for functions of Rpn_L and Rpn_S proteins. $RpnA_L$ dimer (orange and red, *Right* side) and $RpnA_{LS}$ tetramer (*Left* side, toxin-antitoxin complex) structures were predicted with AlphaFold-Multimer.

Rpn Proteins Contribute to Antiphage Defense. The *rpn* genes share characteristics with known antiphage defense systems such as restriction–modification enzymes and toxin–antitoxin complexes, including the ability to quickly diversify and extensive horizontal mobility. The genes also show a tendency to cluster with other defense genes (Fig. 5*A*), including systems with dCas12a or Type I-E Cascade (14). Like most antiphage defense systems (15, 16), the *rpn* genes are not autonomously mobile but likely rely on other mechanisms such as recombination or "hitchhiking" with mobile elements to move. Given these similarities and the rapid change in the numbers of four amino acid repeats among Rpn_S proteins, which we surmised must be due to strong selective pressure acting on the *rpn* genes to diversify the *C*-terminal domain, we hypothesized that Rpn proteins might defend against phage infection.

To test whether Rpn proteins constitute an antiphage defense, we examined the ability of various phages to form plaques on *E. coli* MG1655 strains expressing plasmid-derived $RpnP2_{LS}$ from its native promoter on a low-copy plasmid (Fig. 5 *B* and *C* and *SI Appendix*, Fig. S8). Considering the high levels of the $RpnP2_S$ protein for cells grown in LB (Luria broth) pH 7 (*SI Appendix*, Fig. S2*F*) could inhibit $RpnP2_L$, we wanted to assay cells grown under conditions where the $RpnP2_S$–$RpnP2_L$ interaction might change. Given that $RpnA_{L^*}$ has higher activity at high pH and $RpnA_{LS}$ complex tends to dissociate at high pH, we first grew cells in LB pH 8.5.

Under these conditions, cells expressing $RpnP2_{LS}$ showed 10-fold reduced plaquing by phage T2. We observed smaller T2 plaques and 10-fold reduced plaquing by T4 for $RpnP2_{LS}$ expressing cells grown in minimal glucose medium pH 7.1 at low temperature (Fig. 5*B* and *SI Appendix*, Fig. S8*B*). We also examined the ability of other T-even phages in the BASEL phage collection (17) to plaque on this strain grown in both LB and minimal glucose media pH 8.5 (Fig. 5*C* and *SI Appendix*, Fig. S8 *C* and *D*). $RpnP2_{LS}$ protected against many of the T-even phages though resistance against BASEL phage #40 was strongest (Fig. 5*D* and *SI Appendix*, Fig. S8*D*). This protection was mostly eliminated by active site mutations. Additionally, we deleted *rpnP2* in the ECOR13 strain. This deletion strain showed reduced protection against BASEL phage #40 compared to the ECOR13 wt strain (Fig. 5*D* and *SI Appendix*, Fig. S8*E*).

To further test for a phage defense function, we challenged $RpnP2_{LS}$ expressing cells with BASEL #40 in liquid media at multiplicity of infections (MOIs) of either 10 or 0.005. Although both the vector control and $RpnP2_{LS}$ cells were equally affected at high MOI, $RpnP2_{LS}$ protects the cells at low MOI (Fig. 5*E*). A one-step growth curve also showed $RpnP2_{LS}$ reduced the BASEL #40 burst size following a single round of infection (Fig. 5*F*). Under these BASEL #40 infection conditions, the levels of the $RpnP2_S$ protein are reduced slightly at later time points (*SI Appendix*, Fig. S9). Together, our data suggest Rpn_{LS} toxin–antitoxin systems can provide protection against specific phage.

Discussion

Our data revealed that Rpn_S proteins are expressed together with Rpn_L proteins in *E. coli* as well as in *C. difficile* and serve as antitoxins for the toxic Rpn_L proteins (Fig. 5*G*). It is possible that other toxin–antitoxin systems are composed of additional components translated from iTIS as, for example, a clear intragenic ribosome profiling signal can be detected for *prlF* of the *yhaV-prlF* toxin–antitoxin system of *E. coli* (3, 4). Based on ribosome profiling data, genes-within-genes likely are far more prevalent than previously appreciated and not restricted to toxin–antitoxin genes.

Although the physiological roles of many toxin–antitoxin systems remain under debate, increasing numbers have been shown to constitute antiphage defenses (12, 18–24). Our bioinformatics data showing *rpn* gene enrichment in defense islands as well as the variation in the four amino acid repeats in Rpn_S across different bacteria species were the initial hints that Rpn proteins also might protect against phage infection. Indeed, we demonstrated that a representative system, plasmid-encoded $RpnP2_{LS}$, reduced plaque formation and infection by specific phages (Fig. 5 *B*–*F*). However, it is possible that some Rpn proteins have housekeeping roles in addition to antiphage defense.

The intriguing variation in four amino acid repeats in Rpn_L and Rpn_S proteins is observed across a wide range of bacterial species. Changes in four amino acid repeats have been observed in the context of the EcoR124I and EcoR124II endonucleases (25). EcoR124I, with two repeats of the sequence "TAEL", is specific for a sequence with a 6 bp gap ($GAAN_6RTCG$), while EcoR124II, with three TAEL repeats, is specific for a sequence with a 7 bp gap ($GAAN_7RTCG$). Our structural studies showed that $RpnA_S$ is composed of a three α-helix bundle. Additionally, HHpred (26) predicts relatedness to helix–turn–helix (HTH) motifs. In this context, it is noteworthy that many antitoxins have HTH motifs and adopt a strongly helical structure. In some examples, the toxin binding site overlaps these motifs (27, 28). The affinity between Rpn_S and Rpn_L is clearly tuned by the number of four amino acid repeats located in the long α-helix (Fig. 4*D*).

The overlapping rpn_L-rpn_S gene arrangement means there is obligatory coevolution of the two proteins. The mechanism of the expansion and possibly contraction of these dodecamer repeats warrant further study. Given the extensive variation within a species, we suspect the changes ultimately are driven by a phage factor. It is possible that the Rpn_L proteins themselves are involved in increasing the number of four amino acid repeats as well as in initiating the recombination events that allow the generation of the small protein homologs encoded downstream of some homologs and in generating the duplicated *rpn* genes.

How Rpn_L activity is released, a key unresolved issue for other toxin-antitoxin systems as well, is an important direction for future work. The dissociation of $RpnA_L$ and $RpnA_S$ and increased activity of $RpnA_L$ at high pH led us to wonder whether Rpn proteins are activated at high pH or another specific physiological condition with similar effects. We found that $RpnP2_{LS}$ provided more protection against different phages at alkaline pH. Since $RpnA_S$ is stable at high pH, it is likely Rpn_S is released rather than degraded. In this context, we were intrigued to observe very different Rpn_L:Rpn_S ratios depending on how these genes were expressed suggesting regulated expression is another factor in modulating Rpn_L activity. Other important questions are whether Rpn_L proteins have preferred substrates, whether the Rpn_L nucleases specifically recognize and cleave phage DNA, and what domains comprise the DNA-recognition determinants of the proteins. We expect that further mechanistic understanding of the actions of the Rpn_L and Rpn_S proteins will allow the exploitation of these unique toxin–antitoxin systems.

Materials and Methods

Strains, Plasmids, and Oligonucleotides. Strains, plasmids, and oligonucleotides used in this study are listed in Dataset S2. Details about strain and plasmid construction are provided in *SI Appendix*.

Bacterial Growth. Bacterial growth in LB-rich medium or M9 minimal medium was carried out at indicated pH, indicated temperature, and indicated supplements as described in detail in *SI Appendix*.

Immunoblot Analysis. Specific tagged proteins were detected by immunoblot analysis as described in detail in *SI Appendix*.

β-Galactosidase Activity Assays. Induction of a *dinD-lacZ* reporter was assayed as described in detail in *SI Appendix*.

Biochemical Characterization. RpnA protein purification and characterization by in vitro DNA cleavage assays, SEC analysis, and SEC-MALS analysis were carried out as described in detail in *SI Appendix*.

Structure Determination and Prediction. The structure of $RpnA_S$ was determined as described in detail in *SI Appendix*. All Rpn proteins structures were predicted with AlphaFold 2.2.0 (13, 29) using NIH's Biowulf cluster.

Bioinformatic Analysis. Species distribution, phylogenetic analysis, and gene context analysis were carried out as described in detail in *SI Appendix*.

Bacteriophage Assays. Plaque assays and efficiency of plaquing measurements, growth curves following phage infection, and one-step growth curves to measure burst size were carried out as described in detail in *SI Appendix*.

Data, Materials, and Software Availability. The data are available in the manuscript, or protein structure data have been deposited in the Protein Data Bank (www.wwpdb.org) (PDB 7TH0) (30).

ACKNOWLEDGMENTS. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). We thank members of the Storz lab, A.S. Mankin, and E. Raleigh for comments and A. Buskirk for generating browser images. Work by A.Z. was supported by an NICHD Early Career Award. This work was supported by the Intramural Programs of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (A.Z., K.K., and G.S.), National Library of Medicine (X.J.), and National Institute of Diabetes and Digestive and Kidney Diseases (A.B.H. and F.D.), NIH. M.T.L. is an Investigator of the Howard Hughes Medical Institute.

1. S. Meydan, N. Vázquez-Laslop, A. S. Mankin, Genes within genes in bacterial genomes. *Microbiol. Spectr.* **6**, RWR-0020-2018 (2018).
2. T. M. McBride *et al.*, Diverse CRISPR-Cas complexes require independent translation of small and large subunits from a single gene. *Mol. Cell* **80**, 971–979 (2020).
3. S. Meydan *et al.*, Retapamulin-assisted ribosome profiling reveals the alternative bacterial proteome. *Mol. Cell* **74**, 481–493 (2019).
4. J. Weaver, F. Mohammad, A. R. Buskirk, G. Storz, Identifying small proteins by ribosome profiling with stalled initiation complexes. *mBio* **10**, e02819-18 (2019).
5. A. W. Kingston, C. Ponkratz, E. A. Raleigh, Rpn (YhgA-Like) proteins of Escherichia coli K-12 and their contribution to RecA-independent horizontal transfer. *J. Bacteriol.* **199**, e00787-16 (2017).
6. K. Steczkiewicz, A. Muszewska, L. Knizewski, L. Rychlewski, K. Ginalski, Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily. *Nucleic Acid Res.* **40**, 7016–7045 (2012).
7. D. H. Haft *et al.*, TIGRFAMs: A protein family resource for the functional identification of proteins. *Nucleic Acid Res.* **29**, 41–43 (2001).
8. J. Mahillon, M. Chandler, Insertion sequences. *Microbiol. Mol. Biol. Rev.* **62**, 725–774 (1998).
9. A. B. Hickman *et al.*, Unexpected structural diversity in DNA recombination: The restriction endonuclease connection. *Mol. Cell* **5**, 1025–1034 (2000).
10. J. E. Peters, Targeted transposition with Tn7 elements: Safe sites, mobile plasmids, CRISPR/Cas and beyond. *Mol. Microbiol.* **112**, 1635–1644 (2019).
11. M. K. Thomason *et al.*, Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in *Escherichia coli*. *J. Bacteriol.* **197**, 18–28 (2015).
12. M. LeRoux, M. T. Laub, Toxin-antitoxin systems as phage defense elements. *Annu. Rev. Microbiol.* **76**, 21–43 (2022).
13. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
14. J. R. Rybarski, K. Hu, A. M. Hill, C. O. Wilke, I. J. Finkelstein, Metagenomic discovery of CRISPR-associated transposons. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2112279118 (2021).
15. H. G. Hampton, B. N. J. Watson, P. C. Fineran, The arms race between bacteria and their phage foes. *Nature* **577**, 327–336 (2020).
16. A. Bernheim, R. Sorek, The pan-immune system of bacteria: Antiviral defence as a community resource. *Nat. Rev. Microbiol.* **18**, 113–119 (2020).
17. E. Maffei *et al.*, Systematic exploration of *Escherichia coli* phage-host interactions with the BASEL phage collection. *PloS Biol.* **19**, e3001424 (2021).
18. C. K. Guegler, M. T. Laub, Shutoff of host transcription triggers a toxin-antitoxin system to cleave phage RNA and abort infection. *Mol. Cell* **81**, 2361–2373 (2021).
19. D. Jurenas, N. Fraikin, F. Goormaghtigh, L. Van Melderen, Biology and evolution of bacterial toxin-antitoxin systems. *Nat. Rev. Microbiol.* **20**, 335–350 (2022).
20. C. N. Vassallo, C. R. Doering, M. L. Littlehale, G. I. C. Teodoro, M. T. Laub, A functional selection reveals previously undetected anti-phage defence systems in the *E. coli* pangenome. *Nat. Microbiol.* **7**, 1568–1579 (2022).
21. J. Bobonis *et al.*, Bacterial retrons encode phage-defending tripartite toxin-antitoxin systems. *Nature* **609**, 144–150 (2022).
22. T. Zhang *et al.*, Direct activation of a bacterial innate immune system by a viral capsid protein. *Nature* **612**, 132–140 (2022).
23. M. LeRoux *et al.*, The DarTG toxin-antitoxin system provides phage defence by ADP-ribosylating viral DNA. *Nat. Microbiol.* **7**, 1028–1040 (2022).
24. B. Y. Hsueh *et al.*, Phage defence by deaminase-mediated depletion of deoxynucleotides in bacteria. *Nat. Microbiol.* **7**, 1210–1220 (2022).
25. W. A. M. Loenen, D. T. F. Dryden, E. A. Raleigh, G. G. Wilson, Type I restriction enzymes and their relatives. *Nucleic Acid Res.* **42**, 20–44 (2014).
26. L. Zimmermann *et al.*, A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.* **430**, 2237–2243 (2018).
27. M. A. Schureck *et al.*, Structure of the *Proteus vulgaris* HigB-(HigA)2-HigB toxin-antitoxin complex. *J. Biol. Chem.* **289**, 1060–1070 (2014).
28. P. De Bruyn, Y. Girardin, R. Loris, Prokaryote toxin-antitoxin modules: Complex regulation of an unclear function. *Protein Sci.* **30**, 1103–1113 (2021).
29. R. Evans *et al.*, Protein complex prediction with AlphaFold-multimer. bioRxiv [Preprint] (2022), https://doi.org/10.1101/2021.10.04.463034 (Accessed 4 February 2023).
30. A. Zhong, A. B. Hickman, G. Storz, F. Dyda, *Escherichia coli* RpnA-S. Worldwide Protein Data Bank (wwPDB). https://doi.org/10.2210/pdb7th0/pdb. Deposited 10 January 2022.