ARTICLES
https://doi.org/10.1038/s41564-022-01219-4
nature microbiology



OPEN

# A functional selection reveals previously undetected anti-phage defence systems in the *E. coli* pangenome

Christopher N. Vassallo[1], Christopher R. Doering[1], Megan L. Littlehale[1], Gabriella I. C. Teodoro[1] and Michael T. Laub[1,2] ✉

**The ancient, ongoing coevolutionary battle between bacteria and their viruses, bacteriophages, has given rise to sophisticated immune systems including restriction-modification and CRISPR-Cas. Many additional anti-phage systems have been identified using computational approaches based on genomic co-location within defence islands, but these screens may not be exhaustive. Here we developed an experimental selection scheme agnostic to genomic context to identify defence systems in 71 diverse *E. coli* strains. Our results unveil 21 conserved defence systems, none of which were previously detected as enriched in defence islands. Additionally, our work indicates that intact prophages and mobile genetic elements are primary reservoirs and distributors of defence systems in *E. coli*, with defence systems typically carried in specific locations or hotspots. These hotspots encode dozens of additional uncharacterized defence system candidates. Our findings reveal an extended landscape of antiviral immunity in *E. coli* and provide an approach for mapping defence systems in other species.**

Bacteriophages (or simply, phages) are an extraordinarily diverse and ubiquitous class of viruses that pose a nearly constant threat to bacteria. Phages are the most abundant biological entity on the planet, with estimates of $10^{31}$ particles that drive the daily turnover of ~20% of all bacteria in some environments[1,2]. Bacteria and their viral predators are locked in a perpetual coevolutionary battle, leading to the emergence of sophisticated mechanisms by which phage manipulate and exploit their hosts, and an equally diverse set of bacterial immune mechanisms collectively referred to as anti-phage defence systems[3]. These immunity systems include both innate mechanisms, such as restriction-modification systems, and adaptive mechanisms, such as CRISPR-Cas. Recent studies have begun to identify many new defence systems, but the full inventory probably extends well beyond what is currently defined.

Identifying additional anti-phage defence systems promises to provide new insight into the ancient coevolutionary conflict between viruses and their hosts. Recent work has found that many defence systems have homologues with similar function in eukaryotic innate immunity, indicating a potentially ancient origin and cross-kingdom conservation of many immune systems[4–6]. Additionally, previous studies of anti-phage defence have produced precision molecular tools such as CRISPR and restriction enzymes, so the discovery of new immune mechanisms may enable new tools for manipulating cells and genomes. Finally, there is growing interest in using phages to treat antibiotic-resistant bacterial infections and to manipulate microbiomes[7–9]. A more complete understanding of the diverse mechanisms by which bacteria defend themselves may be critical for these endeavours[10].

Multiple groups have previously used computational methods to identify uncharacterized defence systems on the basis of the observation that anti-phage defence systems often cluster in bacterial genomes in high density, forming so-called 'defence islands'[11–13]. However, not all defence systems may be detectable in these guilt-by-association approaches. For systems that are rare or not widely conserved, it may be difficult to detect enrichment within defence islands, and not all defence systems necessarily associate with defence islands. Additionally, candidates identified computationally must be expressed in a model laboratory organism and then tested against a panel of phages. Some systems may not work in a heterologous host or protect against the phages examined, and demonstrating that a given system provides defence in its native context is typically not tested or even possible. We reasoned that an experimental selection scheme to uncover antiviral proteins may reveal new insights into bacterial immunity, including identifying defence systems that remain uncharacterized and revealing the relative frequency of the different genomic contexts of these bacterial immune systems (that is, defence islands, mobile genetic elements (MGEs) or other contexts). To this end, we took a functional metagenomic approach to map the range of defence systems in the *Escherichia coli* pangenome (Fig. 1). In addition to revealing previously uncharacterized phage defence systems in *E. coli*, our work also provides a robust screening methodology that can now be adapted to systematically identify phage defence mechanisms in virtually any bacterial genome or metagenomic sample.

## Results

**Identification of novel anti-phage defence systems.** To sample the immune landscape of *E. coli*, we collected a diverse set of wild isolate strains from the ECOR collection as well as 19 clinical isolates[14] (71 strains in total), all with available draft genome sequences (Extended Data Fig. 1). The ECOR collection is a set of strains curated to span the phylogenetic diversity of the species[15]. Together, the 71 strains collected encode 21,149 unique gene clusters, of which >10,000 exist in only one or two strains (Extended Data Fig. 1). From genomic DNA, we constructed a 100x-coverage library

[1]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA, USA. ✉e-mail: laub@mit.edu

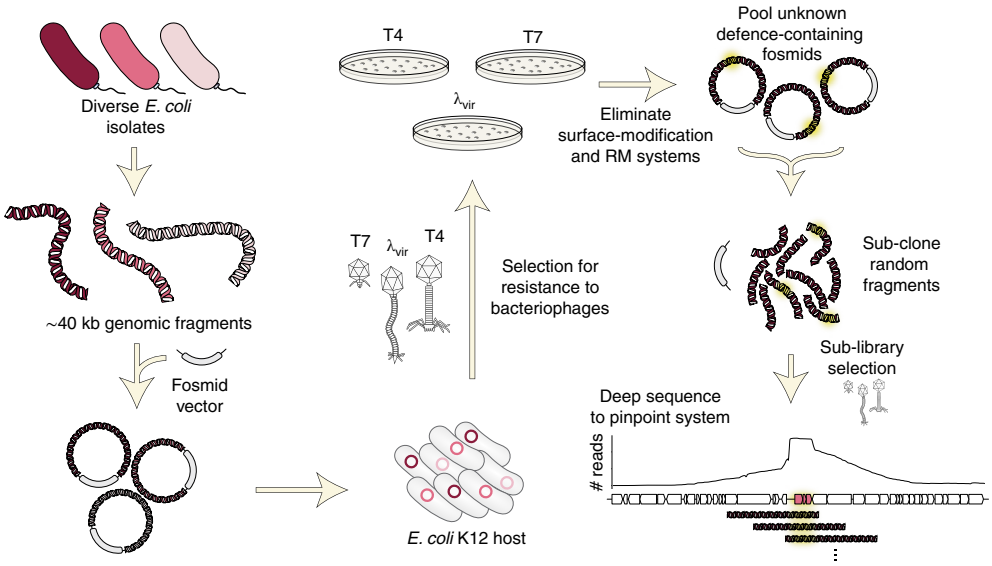



**Fig. 1 | Selection strategy for identifying phage defence systems.** A fosmid library of random ~40 kb fragments of genomic DNA from 71 *E. coli* strains was transformed into an *E. coli* K12 host and then challenged with three different phages. Survivors were isolated and fragments mapped to their genome sequence. After eliminating duplicates, clones affecting adsorption, and clones harbouring restriction-modification or known defence systems, the unique fosmids corresponding to each phage selection were used to construct plasmid libraries, which were subjected to a second selection. Surviving clones were deep-sequenced and candidate defence loci pinpointed by mapping sub-library reads to genome sequences of the original fosmid inserts.

of fosmids, each harbouring a ~40 kb genome fragment, in EPI300, a derivative of *E. coli* K12. We used large-insert fosmids to minimize the size of the library, to include potentially large defence systems, and because the copy number is maintained around one, minimizing false positives due to overexpression.

Many anti-phage defence systems work by an abortive infection (Abi) mechanism in which an infected cell sacrifices its viability to prevent phage replication and thereby protect uninfected cells in a population[16]. Thus, it is impossible to directly select for clones containing an Abi-based defence system because the infected cell dies. Instead, we used a selection strategy, historically known as the 'tab' (T4 abortive) method[17]. In brief, cells harbouring the gDNA library are mixed with phage in a structured medium (soft agar) at varying concentrations of phage (Fig. 2a). At intermediate phage concentrations, individual clones from the library can grow and form small populations before encountering a phage particle. Any such micro-colonies that harbour an Abi defence system can be infected, but the initially infected cell will die without producing progeny phage, enabling the rest of the population to survive and produce a colony. Thus, our screening approach allows the identification of both Abi and conventional defence systems.

Using this general strategy, we challenged cells harbouring the fosmid library with three lytic phages: T4, $\lambda_{vir}$ and T7, each representing a major class of Caudovirales, the tailed bacteriophages (Fig. 1). From each selection, we isolated approximately 90 surviving colonies and then sequenced the ends of the vector insert in each clone to identify the genomic region and strain of origin of each fragment. For each positive clone, we then measured the efficiency of plaquing (EOP)—the ratio of plaques formed by a phage plated on the positive clone to plaques formed by a phage-sensitive control strain (Fig. 2b). In some cases, these EOP assays revealed nearly complete protection even at phage titres of $>10^9$ p.f.u. ml$^{-1}$. Upon initial sequencing, representative clones with this phenotype consistently encoded lipopolysaccharide (LPS) or capsule biosynthesis genes. This phenotype is consistent with a loss of adsorption that results from modification of the core surface properties of the cell (Fig. 2b and Extended Data Fig. 2a). Other positive clones produced protection accompanied by a high rate of escape plaques, a phenotype seen with many restriction-modification (RM) systems (Fig. 2b). Escape plaques can arise with many defence systems but are well known to arise at high frequencies for RM systems due to epigenetic escape[18]. Upon sequencing, the clones exhibiting high-frequency escape plaques indeed encoded RM systems. Thus, clones with these two phenotypes were excluded from further analyses (Methods). From 257 initial clones, 117 and 9 were eliminated as probably resulting from changes in cell surface properties affecting adsorption or RM-mediated defence, respectively. After also accounting for redundancy in the remaining 131 clones, we had 43 clones that we hypothesized to encode novel defence systems.

To pinpoint possible defence systems, we pooled the three sets of remaining fosmids corresponding to each phage and generated random 6–12 kb fragments. These fragments were then sub-cloned into a low-copy plasmid vector to create three high-coverage sub-libraries. Cells harbouring these high-coverage sub-libraries were then selected for resistance to their respective phage, and plasmids from positive clones were sequenced by Nanopore long-read sequencing. The sub-library clones that survived selection each contained the originally selected defence system, flanked by random lengths of adjacent DNA from the original fosmid insert. Thus, when reads were mapped back to the genome fragment in a given fosmid, the coverage maxima typically delineated the boundaries of each candidate defence system (Fig. 2c). In some cases, this identified previously characterized systems, including type III and IV RM systems and an Old-family endonuclease, or non-defence genes that account for phage resistance, such as *malI*, a regulator of the lambda receptor LamB (Supplementary Table 1). After discarding such cases, there were 21 unique candidate defence systems, with 10, 6 and 5 systems from the selections for T4, $\lambda_{vir}$ and T7, respectively

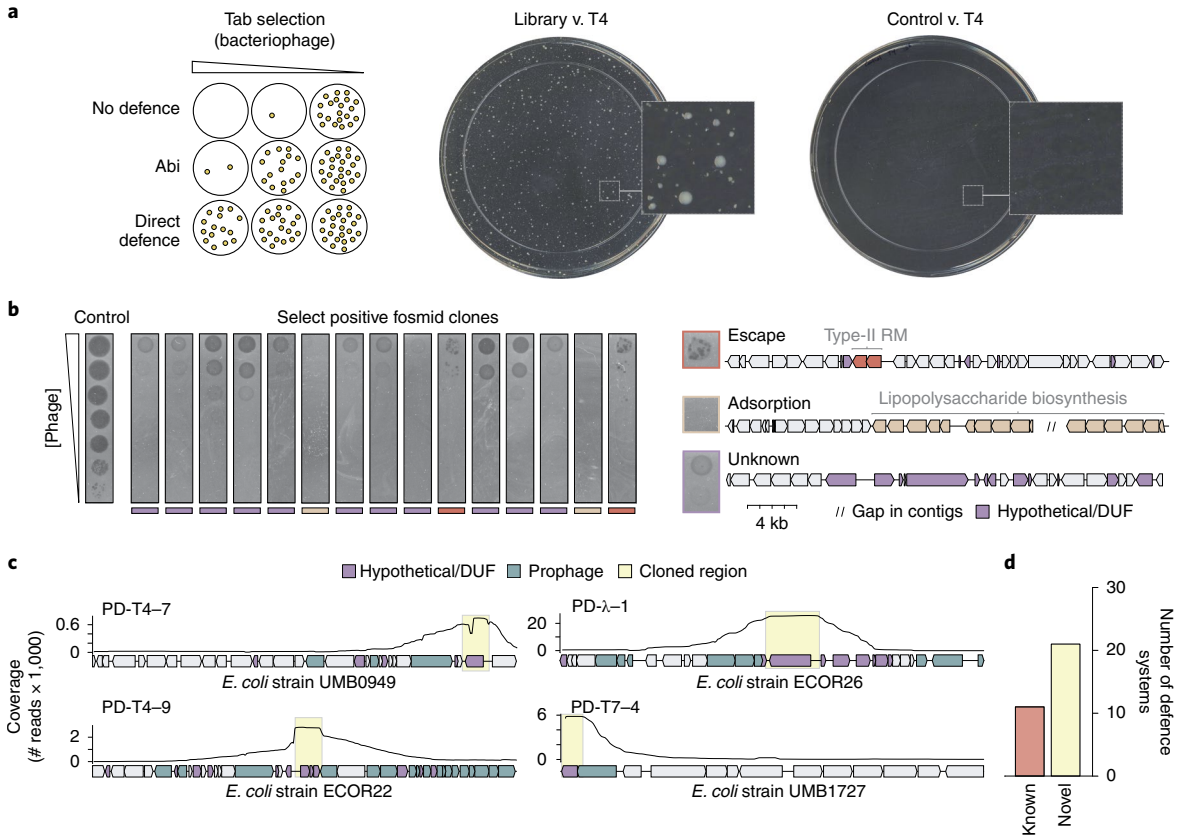



**Fig. 2 | Identification of phage defence systems. a**, Left: schematic of the 'tab' selection method. At intermediate concentrations of phage, tab selection facilitates the survival of cells with either abortive infection or direct defences. Right: examples of T4 selection plates for cells containing the fosmid library or an empty vector control. **b**, Left: tenfold dilutions of $\lambda_{vir}$ phages on lawns of a sample of 15 positive clones from the $\lambda_{vir}$ screen. Multiple phenotypes were observed, including reduction of plaquing with individual escape plaques indicative of a restriction-modification system; no lysis at any concentration of phage typically reflecting a loss of adsorption; or reduction of plaquing, generally indicative of a phage defence system. Right: examples of fosmid inserts corresponding to exemplar phenotypes in **b** (left), with relevant genes coloured. **c**, Examples of read coverage (100 bp moving average) from deep sequencing of sub-libraries generated from positive fosmid clones, with maxima delineating defence system candidates. Genes were coloured or shaded as indicated at the top. **d**, Summary counts of defence systems identified.

(Fig. 2d and Supplementary Table 2). Each candidate system was provisionally named with a PD prefix, for phage defence, followed by the phage used for selection and a unique number.

**Validation of candidate defence systems.** To validate these novel defence systems, we cloned each candidate open reading frame (ORF) or operon into a low-copy vector under the control of its native promoter in wild-type MG1655 (Supplementary Table 2). We confirmed that each system did not affect phage adsorption (Extended Data Fig. 2b) and then challenged each with a panel of 10 diverse phages (Fig. 3a). Each candidate system was confirmed to substantially reduce the EOP for the phage originally used to select the system, and often others. In some cases, a given defence system did not change the EOP of a phage but instead produced smaller plaques. Although most defence systems were relatively specific, protecting against only a few phages, some systems provided relatively broad protection, such as PD-λ-5, which affected EOP or plaque size for all but one of the phages tested. The 10 systems selected to defend against T4 also generally protected against the other, related T-even phages, T2 and T6. Most systems protected most strongly against the phage originally used to select it, but with some exceptions. For instance, PD-λ-5, PD-T7-1 and PD-T7-3 protected more strongly against the T-even phages than against the $\lambda_{vir}$ and T7 phages used to identify them. The fact that these systems provided robust defence against T4 but were not identified in the T4 selection indicates that our screen was not saturating and that the systems identified represent only a subset of the defence systems in the original 71 *E. coli* isolates used.

We then sought to classify whether each system functioned via Abi or provided direct immunity such that infected cells could survive an infection. Because Abi systems require killing of infected cells, these systems typically only provide defence at a low multiplicity of infection (MOI) in which bacteria outnumber phages, whereas direct immunity provides defence to the infected cell and thus allows comparable, although not identical, rates of growth at MOIs above and below 1.0. We thus tested the growth of strains harbouring each defence system infected at an MOI of 0.05 and 5. Of the 21 systems tested, 9 provided direct inhibition of phage infection, producing comparable protection at both MOIs; the other 12 probably use an Abi mechanism of protection with stronger protection at the lower MOI (Fig. 3b and Extended Data 3a). These results validate the ability of our screening strategy to detect Abi defences and underscore the notion that Abi systems contribute considerably to *E. coli* immunity.

One advantage to our screen is that we have the strains from which these defence systems originated, in contrast to computational screens that have identified many defence systems from species that have sequenced genomes but are not readily available.

We were thus able to delete candidate systems in their originating strains and test whether, in their native context, they protect against phage. Specifically, we tested the native role of PD-T4-2 and PD-T4-9, which originate from ECOR65 and ECOR22, respectively, strains to which T4 can adsorb but not infect. Deleting each system dramatically increased the plaquing efficiency of T4, demonstrating that these systems provide defence in both their original, native context and when introduced into *E. coli* K12 (Fig. 3c). We also asked whether our defence systems work in *E. coli* strains other than MG1655 by testing four candidate defence systems in ECOR13 and *E. coli* C, which are natively susceptible to the three phages. All four systems provided protection, indicating that the function of the systems identified is not strictly dependent on strain background (Extended Data Fig. 3b).

In total, 26 of 32 proteins in the 21 systems identified here were annotated in GenBank as either 'hypothetical protein' or as containing domains of unknown function and had no primary sequence homology to any characterized anti-phage defence system. To more sensitively characterize each protein, we used HHpred to detect even remote similarity to PFAM domains[19]. This did not reveal any homology to a known system, and in the majority of cases, most of each protein remained uncharacterized. We were, however, able to detect potential similarity to some motifs or domains characteristic of defence systems, for example, nucleic-acid binding or cleavage domains (Fig. 3a and Supplementary Table 3). In most cases, this was limited to small regions hinting at enzymatic function, but not a mechanism of activation or specific targets.

Remote homology detection revealed several intriguing features uncharacteristic of known defence systems. These included (1) similarity to a ribosome-dependent ribonuclease (RelE) in conjunction with a phage-sheath-like domain, (2) a three-gene operon encoding an exotoxin A-like domain known to participate in bacterial virulence although not phage defence, an unknown protein and a SecB-like chaperone, (3) a putative membrane-anchored protein with a central coiled-coil domain (DUF4041) and a C-terminal DNA binding/cleavage domain, (4) a beta propeller fused to a DNA binding/cleavage domain, (5) a P4 phage β-like protein, (6) a lipoprotein, (7) a zinc-finger-like domain fused with a C-terminal domain of unknown function that belongs to an extended family of ribonucleases, (8) a CoiA domain and (9) DUF6575.

Several defence systems showed similarity to domains that are less frequently associated with defence systems compared with domains such as nucleases and helicases. These less frequent domains included a peptidase, a eukaryotic-like Ser/Thr kinase[20], a $NAD^+$-binding Sir2 homologue[21] and a GIY-YIG nuclease[22]. Four of the systems identified contained a component with similarity to a toxin of toxin-antitoxin (TA) systems, but either none of these systems were found in existing TA databases or they encoded additional uncharacterized components aside from simply toxin and antitoxin. Collectively, our results reveal a striking diversity of proteins involved in bacterial defence and highlight the vast, unexplored landscape of antiviral immunity in bacteria.

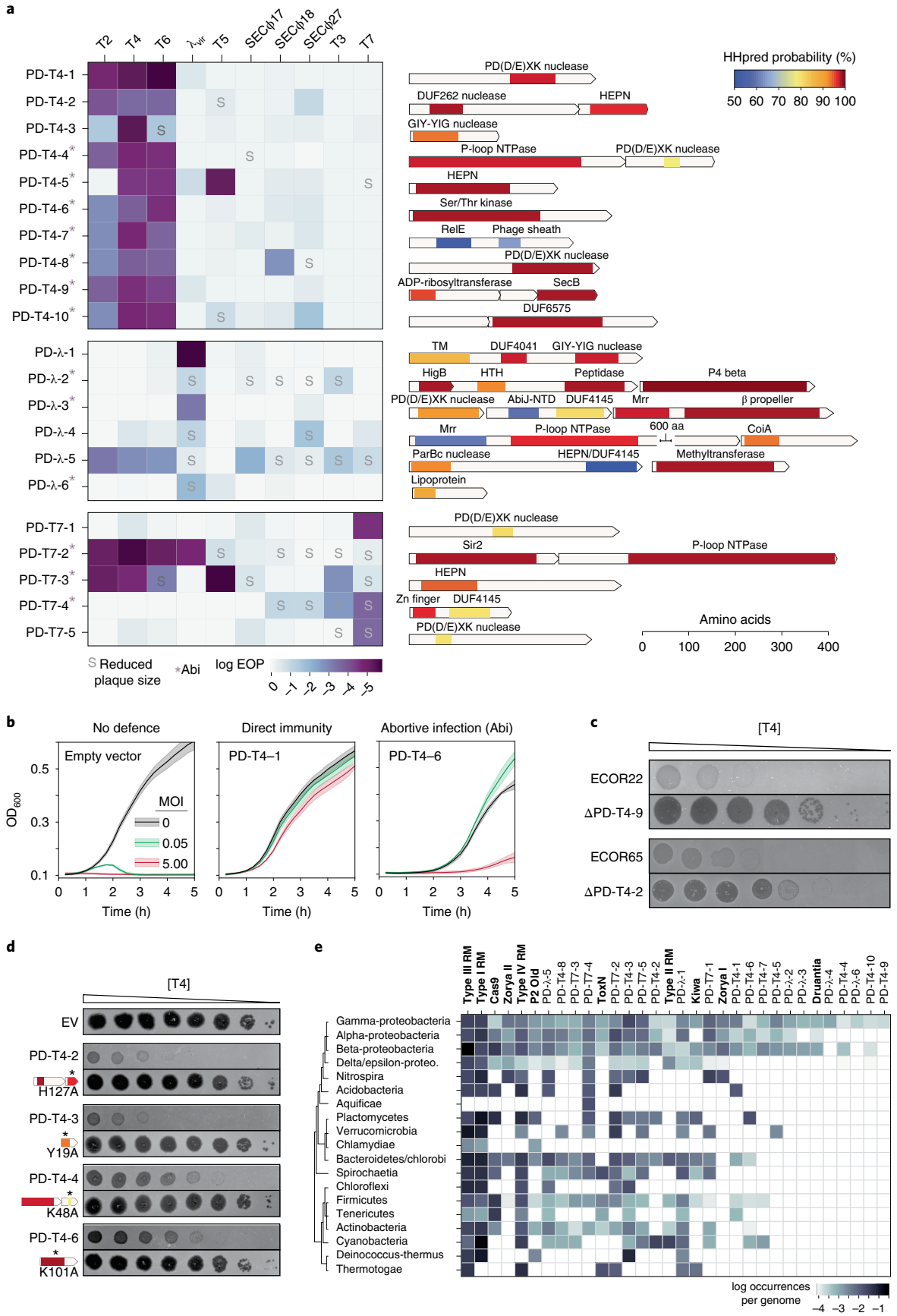
In total, 11 defence systems had a component suggestive of DNA binding or cleavage activity (often including distant similarity to the restriction endonuclease-like fold PD(D/E)XK), although this similarity was, as noted above, typically restricted to a small region or motif. These putative DNases are unlikely to be part of restriction-modification systems as only one included a predicted methylase. Notably, 7 of 11 of these provided direct immunity (not Abi), suggesting non-self, nucleic-acid-targeting activity in potentially novel ways. Remote similarity to HEPN motifs or domains were found in 6 of the 21 systems. These domains are also present in the ribonucleases associated with CRISPR-Cas and toxin-antitoxin systems, supporting the notion that they are common, versatile components of defence systems[23]. In four cases, we mutagenized key conserved residues in the predicted domains and found that they were essential to phage protection, suggesting that the remote domain predictions feature in the function of these systems (Fig. 3d).

To assess the conservation of the 21 systems identified, we investigated the phylogenetic distribution of each (Fig. 3e). Homologues of each system (that is, those that encode all components) were found in other gamma-proteobacteria, with 16 and 18 having homologues in alpha- and beta-proteobacteria, respectively. More than half were also abundant in Firmicutes, Actinobacteria, Bacteroidetes and Spirochaetia, suggesting that many of the systems represent new, widespread classes of anti-phage defence systems. Seven of the systems we identified were restricted to proteobacteria, and three were exclusively found in γ-proteobacteria. Thus, the phage immune landscape of *E. coli* is composed of both widespread and clade-specific systems.

**Mobile elements dominate the defence system landscape in *E. coli*.** Previous searches for proteins enriched in defence islands identified and validated 38 novel defence systems[12,13]. None of the 21 systems identified here are homologues of those systems, and only one component (of PD-T7-2) resembles (32% identity) a protein of a previously validated multi-component system. PD-T4-8 has a DUF4263 domain in common with the Shedu defence system but is not homologous to the validated *Bacillus cereus* system. The computational approach in ref. [13] also identified 7,472 protein families enriched in defence islands that have yet to be validated. Only 14 of 32 proteins identified here have homology to those, and often with <35% identity over limited regions of the proteins (Supplementary Table 4). These observations suggest that our experimental-based selection may uncover different types of defence systems than can be found computationally by searching for enrichment in defence islands.

To further probe this idea, we analysed the native genomic context of the 21 systems identified here. We found that 12 of the 21 systems were located in intact prophages (Extended Data Fig. 4). Seven of these systems were found within P2-like prophages, with four located in the same position of the P2 genome, directly between the genes encoding the P2 replication endonuclease and

**Fig. 3 | Summary and annotation of 21 previously uncharacterized defence system loci. a**, Left: each defence system was cloned into a low-copy plasmid with its native promoter and the EOP tested for a panel of ten phages. Darker colours indicate a higher level of protection. Systems leading to smaller plaque sizes are noted with an 'S', and systems that protect via an Abi mechanism are indicated with an asterisk. Right: for each defence system identified, the operon structure and predicted domain composition of each component are shown. Shaded regions correspond to domain predictions using HHpred, summarized by association to the PFAM clan, with short descriptions at the top. TM, transmembrane domain; HTH, helix-turn-helix. **b**, Bacterial growth in the presence of phage at MOIs of 0, 0.05 or 5. Robust growth at MOI 0.05 but not at MOI 5 indicates an Abi mechanism. Lines represent the mean of three technical replicates, with shaded regions indicating s.d. (see Extended Data Fig. 3 for extended MOI data). **c**, Plaquing of T4 on *E. coli* isolates ECOR22 and ECOR65 or the isogenic defence system deletions. Dilutions were done on two different plates and images combined for presentation. **d**, Plaquing of T4 on strains harbouring the indicated defence system or isogenic site-directed mutants of predicted domains. Asterisks indicate approximate location of mutations made. **e**, Instances of homologues of defence systems by bacterial class, sorted by number of instances, descending from left to right. Known systems are listed in bold for comparison to newly identified systems.


a
T2
T4
T6
λ-vir
T5
SECϕ17
SECϕ18
SECϕ27
T3
T7
PD-T4-1
PD-T4-2
PD-T4-3
PD-T4-4*
PD-T4-5*
PD-T4-6*
PD-T4-7*
PD-T4-8*
PD-T4-9*
PD-T4-10*
PD-λ-1
PD-λ-2*
PD-λ-3*
PD-λ-4
PD-λ-5
PD-λ-6*
PD-T7-1
PD-T7-2*
PD-T7-3*
PD-T7-4*
PD-T7-5
S Reduced plaque size
*Abi
log EOP
HHpred probability (%)
50 60 70 80 90 100
PD(D/E)XK nuclease
DUF262 nuclease
HEPN
GIY-YIG nuclease
P-loop NTPase
PD(D/E)XK nuclease
HEPN
Ser/Thr kinase
RelE
Phage sheath
PD(D/E)XK nuclease
ADP-ribosyltransferase
SecB
DUF6575
TM
DUF4041
GIY-YIG nuclease
HigB
HTH
Peptidase
P4 beta
PD(D/E)XK nuclease
AbiJ-NTD
DUF4145
Mrr
β propeller
Mrr
P-loop NTPase
600 aa
CoiA
ParBc nuclease
HEPN/DUF4145
Methyltransferase
Lipoprotein
PD(D/E)XK nuclease
Sir2
P-loop NTPase
HEPN
Zn finger
DUF4145
PD(D/E)XK nuclease
Amino acids
0 100 200 300 400
b
No defence
Direct immunity
Abortive infection (Abi)
Empty vector
PD-T4–1
PD-T4–6
MOI
0
0.05
5.00
OD600
Time (h)
c
[T4]
ECOR22
ΔPD-T4-9
ECOR65
ΔPD-T4-2
d
[T4]
EV
PD-T4-2
H127A
PD-T4-3
Y19A
PD-T4-4
K48A
PD-T4-6
K101A
e
Type III RM
Type I RM
Cas9
Zorya II
Type IV RM
P2 Old
PD-λ-5
PD-T4-8
PD-T7-3
PD-T7-4
ToxN
PD-T7-2
PD-T4-3
PD-T7-5
PD-T4-2
Type II RM
PD-λ-1
Kiwa
PD-T7-1
Zorya I
PD-T4-1
PD-T4-6
PD-T4-7
PD-T4-5
PD-λ-2
PD-λ-3
Druantia
PD-λ-4
PD-T4-4
PD-λ-6
PD-T4-10
PD-T4-9
Gamma-proteobacteria
Alpha-proteobacteria
Beta-proteobacteria
Delta/epsilon-proteo.
Nitrospira
Acidobacteria
Aquificae
Planctomycetes
Verrucomicrobia
Chlamydiae
Bacteroidetes/chlorobi
Spirochaetia
Chloroflexi
Firmicutes
Tenericutes
Actinobacteria
Cyanobacteria
Deinococcus-thermus
Thermotogae
log occurrences per genome
−4 −3 −2 −1

portal protein (Fig. 4a). This location has been previously found to harbour the defence systems *tin* and *old*, the former being also identified in our screen against $\lambda_{vir}$[24] (Supplementary Table 1). More recently, this location was found to encode a wide array of previously uncharacterized defence systems[25]. We also observed a second defence-enriched locus in P2-like phages, which contained three of the systems discovered here, and the previously identified defence gene *fun* in the P2 reference genome[24] (Fig. 4a). This hotspot was not noted in the recent analysis of P2 defences. Notably, 47 of our 71 *E. coli* strains together encode 63 P2 portal proteins, with 110 unique proteins found in the adjacent hotspot. Only 13 of these 63 hotspots contain previously known defence systems, and most genes are annotated as hypothetical. These findings suggest that not only do P2 prophages encode a rich diversity of anti-phage proteins[25] but that P2 defence hotspots make up a substantial fraction of the immune landscape in *E. coli* (Fig. 4b).

Five defence systems were found associated with other types of prophages or their remnants, including P4 satellite prophages or related integrases, Mu-like phages and lambdoid phages (Extended Data Fig. 4). We also observed a defence-enriched locus within an integrative and conjugative-like element (ICE), containing two systems we identified—PD-T4-4 and PD-T7-2 (Fig. 4a). This element is not widely distributed in our *E. coli* strain collection, but similar elements in other genomes encode known defence systems at this locus, as well as hypothetical proteins that may also function in phage defence (Fig. 4a). PD-T7-4 and its homologues often overlap an integrase gene, while PD-T4-5 was identified on a plasmid. The remaining 4 systems did not appear to reside within active MGEs, but each had a nearby integrase gene suggesting that they may be part of decaying MGEs (Extended Data Fig. 4).

Our study supports previous findings that in addition to defence islands, prophages and other MGEs are a rich reservoir of defence systems[26]. However, these categories are not mutually exclusive, as some defence islands may be carried on or derived from MGEs[27]. To more systematically document the different genomic contexts for the systems identified here, we collected, for each of the 21 systems, all homologues in a set of 844,603 publicly available bacterial genomes. We classified the genes within 10 kb upstream and downstream of each homologue as either defence-related, prophage-related or neither (Methods). We then tabulated the number of defence- and prophage-related genes flanking each homologue (Fig. 4c,d and Extended Data Fig. 5).

We detected two distinct patterns. For systems that we found outside of prophages in our strains, the homologues were also typically not prophage-associated (Fig. 4c and Extended Data Fig. 5) and were often near several other genes encoding defence-associated domains. Thus, these systems do appear in defence islands, even though they were not previously detected as enriched in them. For the systems we identified within *E. coli* prophages, some of their homologues were also found in prophages, as evidenced by dozens of flanking prophage-related genes (Fig. 4d and Extended Data Fig. 5). These prophage-associated homologues were typically near 1–2 defence-associated genes but rarely more than 2, suggesting that some systems reside in small defence islands or clusters within a prophage, as with the P2 hotspots (Fig. 4d)[25].

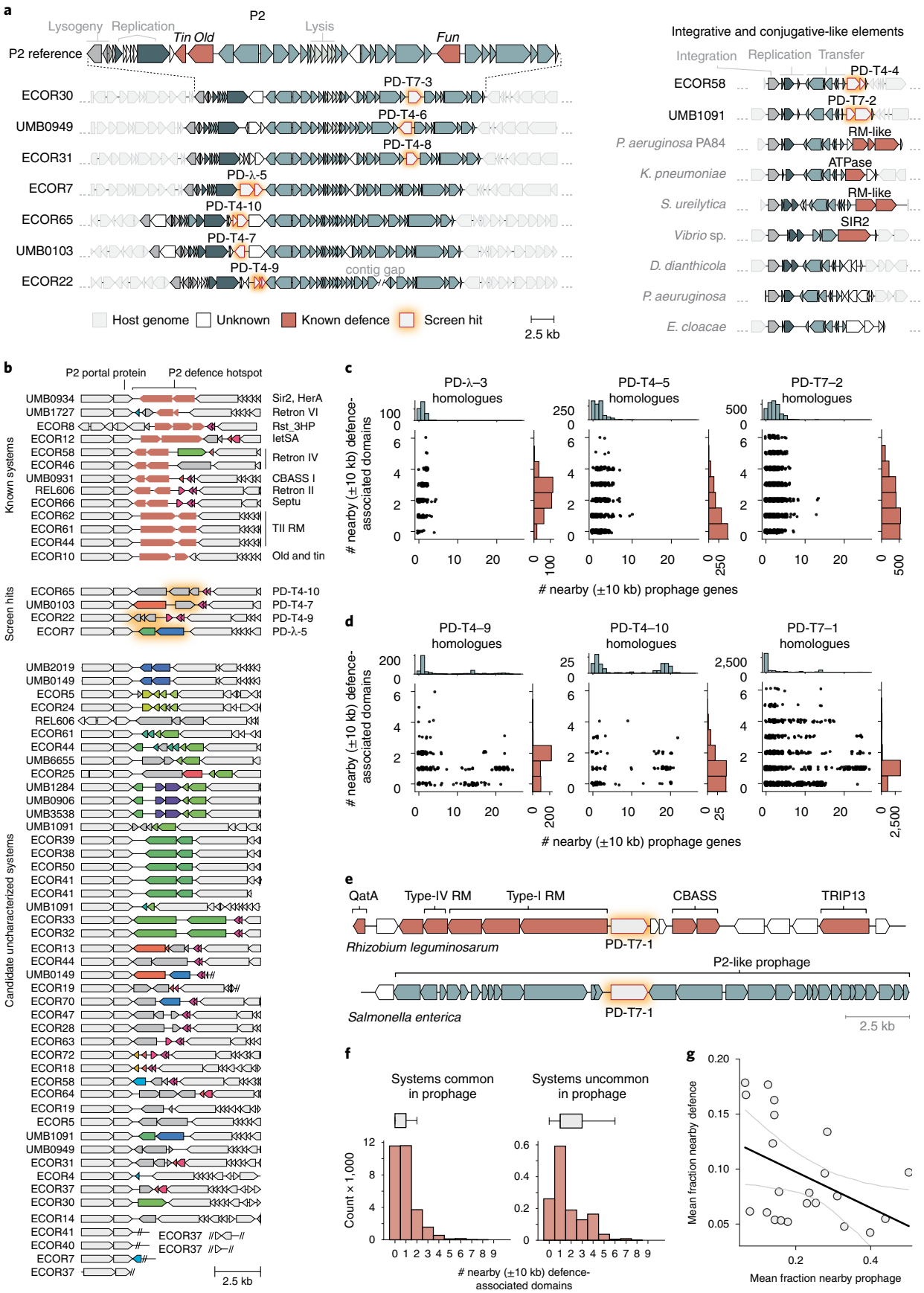
Notably, there are homologues of each system we identified that can be found in defence islands (Fig. 4e), some more rarely than others, indicating that they do not require a prophage context to function. In aggregate, we observed an inverse correlation between the number of neighboring defence-associated genes for homologues of systems commonly found in prophages versus those that are not (Fig. 4f,g). This highlights the constraint on how many defence systems can be carried by prophages, or within a given hotspot, due to size limitations in DNA that can be packaged. We suspect that this is one reason why many of these genes are enriched in prophages as compared with defence islands.

Finally, we compared defence island and prophage enrichment (Methods) between systems discovered here and those previously predicted computationally and validated experimentally[12,13]. We found that our experimentally selected systems were on average less frequently associated with known defence genes but more frequently associated with prophage genes (Extended Data Fig. 6). These analyses suggest that defence-island enrichment methods may be less sensitive in identifying defence systems frequently found in prophage.

**Previously uncharacterized toxin-antitoxin-like systems.** Toxin-antitoxin (TA) systems are typically composed of a protein toxin that can arrest cell growth but is normally neutralized by a cognate, co-expressed antitoxin. TA systems are extremely prevalent in bacterial genomes and MGEs, but their functions remain poorly understood[28]. A handful of TA systems have been found to provide anti-phage defence through an Abi mechanism[29–31]. Our selection yielded four different systems that were recognizable as TA-like in nature. These encoded gene products with sequence similarity to toxic proteins, mostly featured multiple components, and provided Abi defence (PD-T4-5, PD-T4-7, PD-T4-9 and PD-λ-2). A fifth (PD-T4-10) facilitated Abi defence and had two overlapping ORFs, reminiscent of many TA systems. As noted above, none of these systems were previously annotated as TA systems, so we sought to validate the three featuring multiple components: PD-T4-10, PD-λ-2 and PD-T4-9 (Fig. 5a).

For PD-T4-10, neither of the two proteins had a predicted function. We expressed each component from an inducible promoter and found that the second component, PD-T4-10B, was toxic. This toxicity was fully neutralized when PD-T4-10A was co-expressed (Fig. 5b). PD-T4-10A could not be deleted, consistent with it being an antitoxin, whereas deletion of the toxin PD-T4-10B abolished resistance to T4 infection (Fig. 5c). Thus, this system comprises a novel, bona fide TA system that provides strong protection against T-even phages.

**Fig. 4 | Prophages and MGEs are major sources of defence systems. a**, Hotspots of previously uncharacterized defence systems. Left: the native genome context of seven defence systems identified here, showing the boundaries of P2-like prophages in the genome from which they originated. Genes are colour-coded as indicated at the bottom. Right: two defence systems were identified in the accessory region of an ICE-like element within the indicated genome. Homologous elements from other bacterial genomes contain known and putative defence systems in the same location. **b**, All identified instances of P2 defence hotspot #1 in our 73-strain *E. coli* collection. White genes are flanking conserved P2 genes. Each colour of gene within the hotspot represents a protein cluster (30% identity). All grey genes belong to a lone cluster. Double slashes denote the end of a contig. **c**, Number of defence and prophage-associated genes ±10 kb from system homologues. The scatterplots indicate, for each homologous system, the numbers of prophage and defence-associated genes within ±10 kb. Examples in **c** represent systems that were found outside of prophages in our genome collection. For all 21 systems, see Extended Data Fig. 5. **d**, Same as **c** but for systems we found in prophages. **e**, Examples of the ±10 kb context for PD-T7-1 homologues in a defence island or prophage. **f**, Distribution of the number of nearby defence-domain-containing genes in homologues of systems commonly found in prophages (>10% homologue-containing regions with 8+ prophage genes in proximity) or not; $n = 12$ and 9 systems, respectively. Boxes indicate bounds of the distribution as median ± quartiles, and limits exclude outliers. **g**, Linear least squares regression for total nearby prophage gene and defence-domain-containing genes for each system. Pearson $r = -0.442$, $P = 0.045$; error indicates 95% CI.


a
P2
Lysogeny
Replication
Tin
Old
Lysis
Fun
P2 reference
ECOR30
PD-T7-3
UMB0949
PD-T4-6
ECOR31
PD-T4-8
ECOR7
PD-λ-5
ECOR65
PD-T4-10
UMB0103
PD-T4-7
ECOR22
PD-T4-9
contig gap
Host genome
Unknown
Known defence
Screen hit
2.5 kb
Integrative and conjugative-like elements
Integration
Replication
Transfer
ECOR58
PD-T4-4
UMB1091
PD-T7-2
P. aeruginosa PA84
RM-like
K. pneumoniae
ATPase
S. ureilytica
RM-like
Vibrio sp.
SIR2
D. dianthicola
P. aeuruginosa
E. cloacae
b
P2 portal protein
P2 defence hotspot
Known systems
UMB0934
Sir2, HerA
UMB1727
Retron VI
ECOR8
Rst_3HP
ECOR12
IetSA
ECOR58
ECOR46
Retron IV
UMB0931
CBASS I
REL606
Retron II
ECOR66
Septu
ECOR62
ECOR61
TII RM
ECOR44
ECOR10
Old and tin
Screen hits
ECOR65
PD-T4-10
UMB0103
PD-T4-7
ECOR22
PD-T4-9
ECOR7
PD-λ-5
Candidate uncharacterized systems
UMB2019
UMB0149
ECOR5
ECOR24
REL606
ECOR61
ECOR44
UMB6655
ECOR25
UMB1284
UMB0906
UMB3538
UMB1091
ECOR39
ECOR38
ECOR50
ECOR41
ECOR41
UMB1091
ECOR33
ECOR32
ECOR13
ECOR44
UMB0149
ECOR19
ECOR70
ECOR47
ECOR28
ECOR63
ECOR72
ECOR18
ECOR58
ECOR64
ECOR19
ECOR5
UMB1091
UMB0949
ECOR31
ECOR4
ECOR37
ECOR30
ECOR14
ECOR41
ECOR37
ECOR40
ECOR37
ECOR7
ECOR37
2.5 kb
c
PD-λ–3 homologues
PD-T4–5 homologues
PD-T7–2 homologues
# nearby (±10 kb) defence-associated domains
# nearby (±10 kb) prophage genes
d
PD-T4–9 homologues
PD-T4–10 homologues
PD-T7–1 homologues
# nearby (±10 kb) defence-associated domains
# nearby (±10 kb) prophage genes
e
QatA
Type-IV RM
Type-I RM
CBASS
TRIP13
Rhizobium leguminosarum
PD-T7-1
P2-like prophage
Salmonella enterica
PD-T7-1
2.5 kb
f
Systems common in prophage
Systems uncommon in prophage
Count × 1,000
# nearby (±10 kb) defence-associated domains
g
Mean fraction nearby defence
Mean fraction nearby prophage

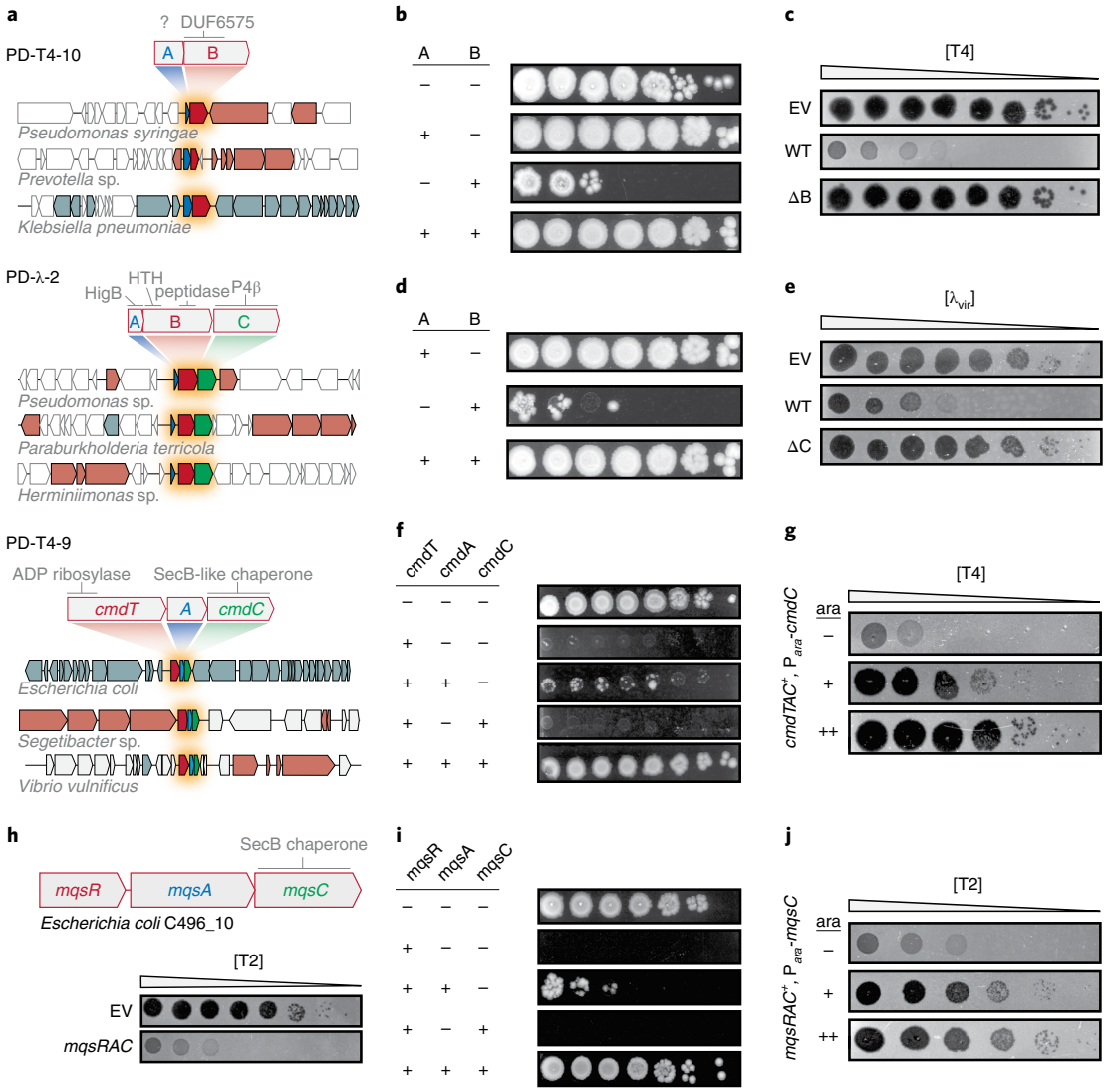



**Fig. 5 | Previously uncharacterized toxin-antitoxin-derived defence systems. a**, Schematics of PD-T4-9, PD-T4-10 and PD-λ-2 defence system operons and their domain predictions. Representative homologues of the systems are shown in their genomic contexts and indicate conservation and order of the system components. Blue, putative antitoxin; red, putative toxin; green, accessory factor. **b**,**d**,**f**,**i**, Each component or pair of components indicated was expressed (+) or not (−) from an inducible promoter and assayed for viable colony-forming units in tenfold serial dilutions. **c**,**e**, Plaquing assays for the phage indicated on cells harbouring an empty vector or a vector containing a given defence system with all components (WT) or lacking the component indicated. **g**,**j**, Plaquing of phages on TAC-containing strains expressing a second copy of the chaperone component to varying levels during infection. **h**, Schematic and T2 plaquing assay of the *mqsRAC* TAC system.

PD-λ-2 features three components. The first is similar to HigB toxins that inhibit translation, the second has a HigA antitoxin-like domain (Xre-family helix-turn-helix) fused to a C-terminal peptidase domain and the third component is related to a P4 phage antitoxin. An Xre-peptidase fusion, co-expressed with an upstream toxin-like gene, has been observed in sequence-based TA searches but the function of these loci is unknown[32]. Unexpectedly, we found that overexpressing the second component, PD-λ-2B, alone was toxic, with toxicity rescued by co-expression with PD-λ-2A (Fig. 5d). We also confirmed that PD-λ-2C, although not required to neutralize the toxin, is required for defence against $\lambda_{vir}$ (Fig. 5e).

PD-T4-9 also contains a third component, a SecB-like chaperone, suggesting that it is related to an enigmatic class of TA systems called toxin-antitoxin-chaperone (TAC) systems, so we renamed this system CmdTAC for chaperone-mediated defence. The antitoxins of TAC systems are typically homologous to canonical antitoxins but feature an unstructured extension of their C termini called the chaperone addiction domain (ChAD)[33]. In the absence of the cognate SecB-like chaperone, the ChAD renders the antitoxin prone to aggregation and proteolytic degradation, thereby freeing the toxin[33]. Inducing the expression of CmdT, a previously uncharacterized toxin that has an ADP-ribosyltransferase-like domain, was toxic. Co-expression with the presumed antitoxin, CmdA, only marginally improved viability. However, co-expression with both the putative antitoxin and chaperone components completely restored viability (Fig. 5f). Together, these results suggest that CmdTAC is a novel TAC system that protects against phage.

To further characterize the role of the chaperone, CmdC, we overproduced it during T4 infection of CmdTAC$^+$ cells. Interestingly, an oversupply of CmdC abolished phage protection by CmdTAC, suggesting that destruction of CmdC, or sequestration of CmdC from the complex may allow this TA system to activate in response to phage infection (Fig. 5g). The chaperone normally promotes neutralization of CmdT by CmdA but, following infection, could be

depleted or sequestered by a phage product, leading to liberation of the toxin and abortive infection. Providing additional CmdC prevents the loss or complete sequestration of chaperone, thereby preventing release of CmdT from CmdA. Additional work is needed to further dissect the mechanistic nature of chaperone-phage interactions.

We hypothesized that TAC systems may be a broad class of phage defence systems. To test whether other TAC systems can protect *E. coli* MG1655 against phage infection, we cloned and tested an MqsRAC system from *E. coli* C496_10. Although completely unrelated in toxin and antitoxin sequence to CmdTAC, MqsRAC is a canonical TAC system, has been characterized in Mycobacteria and includes a SecB-like chaperone homologous to CmdC[33]. This system conferred robust protection against T2 (Fig. 5h) but not T4 as with *cmdTAC*. Similar to CmdTAC, toxicity of MqsR could only be rescued by expressing MqsA and MqsC (Fig. 5i), and inducing additional MqsC in MqsRAC$^{+}$ cells inhibited phage protection (Fig. 5j). Thus, our work indicates that TAC systems may be a widespread and diverse new class of phage defence system.

## Discussion

Our work indicates that a large reservoir of diverse, previously unknown phage defence genes is distributed across the *E. coli* pangenome. Like many bacteria, there is tremendous variability in the 'accessory' genomes of different strains of *E. coli*. Many of these accessory genes are probably associated with phage defence, as recently suggested for marine *Vibrio* species[34]. Although efforts to find new defence systems on the basis of proximity to known systems have proven fruitful, our work reveals that many phage protective systems remain unidentified.

Our functional screening, which is agnostic to the genomic context of defence systems, indicates that many systems are not commonly or detectably enriched among known defence islands. Indeed, none of the 21 systems we identified were previously reported to provide phage defence in previous discovery efforts based on defence island enrichment. Moreover, only three of the systems we identified were within 10 kb of other known defence systems in our *E. coli* genomes. However, notably, 15 of the 21 systems we identified were present in apparently active or recently active MGEs (prophages, ICEs or plasmids) (Extended Data Fig. 4), with the other 6 located in regions suggestive of MGEs in the late stages of decay. Homologues of the systems discovered here are sometimes present in defence islands (Figs. 4d,e and 5a), but these associations are often relatively rare.

It is well established that MGEs contribute to antiviral defence in bacteria[26]. By providing a glimpse into the relative contributions of MGEs and defence islands to immune system context, our results support the notion that active prophages and other MGEs are probably the primary reservoirs of anti-phage defence systems in *E. coli*. This idea is consistent with studies that have identified diverse anti-phage systems in P2 and P4 prophages[24,25] and another recent study revealing how the transfer of defence-containing ICEs drove the emergence of phage resistance in clinical isolates of *V. cholerae*[34]. The defence systems found in functional MGEs probably help these elements to protect themselves and host resources by preventing infection by other phages[27]. However, given genome size and packaging constraints, there is a limit to the number of defence systems carried by a given prophage. Such constraints may help explain why increased prophage association was correlated with lower defence gene association (Fig. 4g). Future investigation should work towards further uncovering the vast reservoir of anti-phage elements carried on MGEs.

Of the 32 proteins in 21 systems found here, 13 feature domains never before reported to function in phage defence. These include a protein with distant similarity to RelE and a region of a phage tail sheath, an Exotoxin A-like domain previously only shown to function in bacterial virulence, a SecB-like chaperone, DUF6575, DUF4041, a CoiA competence-related domain, a β-propeller fold, an ImmA/IrrE peptidase, a HigB toxin and three proteins with no ascertainable similarity to deposited domains. In addition, ten proteins contained large regions with no predicted domains (Fig. 3a).

Although some regions of the proteins identified have distant similarity to known nuclease motifs, these domains are found here in new or unusual contexts and associations, which raises fascinating questions for future investigation. For example, one features a 7-bladed beta propeller with a separate N-terminal Mrr-family nuclease domain. The β-propeller fold consists of separate modules that adopt a disc-like, circularly arranged structure with a central channel that can accommodate many substrates including protein and DNA[35]. Determining how this domain aids in activation or target specificity of this system opens many avenues of future discovery. We also uncovered five single-protein systems that exclusively contain a putative DNase domain. Although these domains are found in defence systems such as RM and CRISPR, how orphan DNase-like proteins sense and respond to phage infection, especially to provide direct defence, is unclear. As all but one is encoded without a DNA methylase, they probably do not distinguish phage and host DNA in the same way as RM systems. Additionally, these DNase-containing systems protected against T4 and T7, which are intrinsically resistant to most RM systems[36,37]. Type-IV RM systems are single-component defences known to target the modified cytosine of T-even phages[38], but none of the five proteins discussed here resemble these systems. How those that are specific to T7 might target T7 DNA, which is not modified, is unclear. The existence of unconventional, phage-targeting, nucleic-acid degradation systems underscores a knowledge gap in the molecular mechanisms of viral resistance and self/non-self recognition.

We also identified and validated three TA systems as phage defence elements including *cmdTAC*, prompting the discovery that unrelated TAC systems such as *mqsRAC* also function in antiviral immunity. MqsRA is a well-studied TA system but has no documented role in phage defence; our results suggest that bacteria have co-opted this TA system by addition of a chaperone-dependent antitoxin to activate the toxin in response to infection. Our findings support the notion that TA systems play a central role in phage defence[29].

Finally, we note that four systems identified here probably belong to classes of previously described systems but have diverged quite significantly such that they share little significant sequence homology, reinforcing the extreme divergence and adaptations that typify many immune systems. These divergent systems include PD-T7-2 in which the second protein is similar to HerA of Sir2/HerA[13], PD-T4-8 whose central domain bears distant resemblance to Shedu[12], PD-λ-5 which appears to be a highly compacted prophage version of an RM+Abi system and PD-T4-5—a plasmid-borne gene that is a distant relative of AbiF/D.

Our screening methodology enables the experimental discovery of anti-phage defence genes and has several powerful features. First, we can return to the strain of origin for a given system and demonstrate that it provides defence in its native context. Second, there is a built-in pairing of defence systems to the phage they defend against, whereas with computational studies, the phage(s) a given system defends against must be subsequently identified. Third, in the pipeline described here, the source DNA comes from other *E. coli* strains, which probably minimizes false negatives that can arise from producing candidates in a heterologous host. Finally, our experimental approach is not limited to genes that are detectably enriched in defence islands. As noted, only three of the systems we identified were natively associated with obvious defence islands, and some also do not appear to have many close homologues in defence islands. Some defence systems may not associate with defence islands, while some may have arisen too recently or not be

widespread enough to detect an association. Indeed, some of the systems identified here show a relatively limited phylogenetic distribution. However, the phage defence capabilities of bacteria probably include both broadly conserved and clade-specific systems adapted to the unique biology of a given organism and its phages.

A limitation of our method is that the initial selection does not distinguish between clones that lead to cell surface modifications that block adsorption and those providing intracellular defence systems; these must be distinguished in downstream steps. Further, at present, our method probably does not reach saturation as some systems are strongly overrepresented among the positive clones, which probably hinders the identification of others. The reason for this uneven distribution is unclear, but efforts are underway to increase throughput and sensitivity of the method.

The methodology developed here can be powerfully extended in several ways. First, genomic DNA from other sources, including metagenomic DNA, could be used as input material. From just 71 strains, we identified 21 new defence systems, suggesting that fertile ground remains to be discovered both in and beyond *E. coli*. Second, the panel of phages tested here was limited to three and could easily be expanded, particularly given the enormous diversity of phages. Finally, with only small modifications, any transformable bacterium could be used as the host strain. Further identification and characterization of bacterial immune systems promise to shed new light on the ancient arms race between bacteria and their viral predators, and may also have practical applications, providing the foundation for precise molecular tools and helping to inform future efforts to develop phage as therapeutics.

## Methods

**Bacteria and phage growth and culture conditions.** Cultures were routinely grown at 37 °C in LB medium unless otherwise stated. Phage stocks were propagated on MG1655, filtered through a 0.2 µm filter and stored at 4 °C. Select ECOR strains were obtained from the Thomas S. Whittam STEC Center at Michigan State University, and UMB isolates were obtained from Alan J. Wolfe at Loyola University Chicago (Supplementary Table 5). Other strains, plasmids and primers + synthetic gene fragments are listed in Supplementary Tables 6–8, respectively.

**Library construction.** Genomic DNA was collected from pooled overnight cultures of each *E. coli* isolate using the PureLink genomic DNA mini kit (Invitrogen). From this sample, a fosmid library was constructed by Rx Biosciences using the CopyControl fosmid library production kit (Lucigen) according to the manufacturer's protocol. Plasmid sub-libraries were constructed first by extracting fosmid DNA from select positive clones using the ZR plasmid miniprep kit (Zymo Research). Equimolar pooled fosmids were sheared to an average of 8 kb using g-TUBEs (Covaris). Sheared DNA was end-repaired and 5'-phosphorylated using the End-It DNA end-repair kit (Lucigen) and then purified using the DNA Clean and Concentrator kit (Zymo Research). The plasmid vector was prepared by PCR, and blunt-ended fragments were ligated to the plasmid using T4 DNA Ligase (NEB) for 2 h at room temperature. The ligation reaction was electroporated into MegaX DH10B T1R Electrocomp cells (Thermo Fisher) and selected on LB with 50 µg ml$^{-1}$ kanamycin.

**Defence system selection.** We used a variant of the previously described 'tab' selection procedure to select for fosmids that provide resistance to phage infection[17]. A heavy inoculum (>30 µl) of a high-titre library freezer stock or empty vector freezer stock was inoculated into 5 ml LB containing 20 µg ml$^{-1}$ chloramphenicol and grown to stationary phase at 37 °C (approximately 4 h, optical density (OD)$_{600}$ = 2–3). Cultures were adjusted to OD$_{600}$ = 1.0, and 0.1 ml (~8 × 10$^7$ cells) was pipetted onto one side of 3–6 empty 15 cm Petri dishes. A tenfold dilution series of phage stock was prepared, and 0.1 ml of each dilution was pipetted onto the empty plates containing the bacterial cultures (onto a separate area of the plate, preventing mixing). Molten LB (20 ml) 0.5% agar was added to the plate and briefly mixed to disperse bacteria and phage. Plates were incubated at 37 °C overnight, except for one T4 screen which was conducted at room temperature. Bacterial colonies were picked from the plate containing the phage dilution that produced the largest difference in number of colonies between the control and library samples and then streaked onto fresh plates to isolate single colonies. To test phage resistance phenotypes, single colonies were cultured overnight, and 30 µl of culture were mixed with 5 ml of molten LB 0.5% agar and 30 µg ml$^{-1}$ chloramphenicol in 8-well rectangular dishes. Serial dilutions of phage were spotted onto the solidified culture media and incubated overnight at 37 °C. Fosmid DNA was isolated from positive clones, and the end of each fosmid insert was Sanger sequenced.

With some phage-resistant clones, we observed no lysis even at very high phage concentrations (Fig. 2b). All sequenced clones with this phenotype contained either LPS or capsule biosynthesis genes. In our experience, no intracellular defence system completely prevents visible lysis at extremely high phage concentrations, whereas changes in the phage receptor or cell surface can, so we suspected that these clones disrupted phage adsorption. Similarly, with regard to T7, genes for capsule biosynthesis survived selection but did not display any difference from the control in plaquing efficiency. All clones showing a complete lack of lysis, or no change in EOP and encoding capsule genes, were discarded. Twelve positive clones (all from the $\lambda_{vir}$ selection) produced a high number of discrete escape plaques, consistent with RM systems. Strains that displayed this escape phenotype and whose fosmids contained an identifiable RM system were discarded from further analysis.

Sub-libraries were screened as described above using a variation that allowed bulk collecting of all positive colonies directly from the screening plate (similar to the previously published *gro* screen)[39]. In this variation, instead of molten LB 0.5% agar, bacteria and phage samples were spread on the surface of LB 1.2% agar using glass beads.

**Long-read sequencing and defence system identification.** After cells were collected in bulk from the screening plates, total plasmid DNA was extracted and linearized by digestion with the restriction enzyme NdeI, EagI or FseI. For Oxford Nanopore sequencing, linearized samples were characterized on a FemtoPulse (Agilent Technology) to confirm integrity, and high-quality samples were indexed by native barcoding (ONT kits EXP-NBD104/114) with supporting reagents from New England Biolabs. Libraries were prepared using the LSK-109 chemistry, and samples were run on either a MinION (R9) or PromethION (R9.4) flowcell. Basecalling was done using built-in ONT tools. Processed reads were aligned to public reference genomes of the source organisms, or the relevant portions of the genomes contained in the fosmid inserts, using the Minimap2 plugin within the Geneious Prime 2020.2.4 software suite. Fosmids that could not be mapped to their genomes due to contig gaps were also sequenced identically, and de novo assembly was conducted using Flye[40], also in Geneious. Candidate defence systems were predicted to be any gene or operon residing under the coverage maxima. In the few cases in which the result was ambiguous, candidates were cloned after prioritization by features including domain prediction, location in defence hotspots, hypothetical proteins and by general comparative genomic investigation. Multi-component systems (operons) were predicted by ORF proximity, promoter prediction and gene co-occurrence in homologues.

**Strain construction.** Defence system cloning was performed using Gibson Assembly of PCR products containing predicted defence systems and their predicted promoters into a destination vector lacking an upstream promoter. Assembled plasmids were transformed into MG1655 using the transformation and storage solution (TSS) method[41]. MG1655 with a deletion of the region containing *mrr* was used as the host strain for PD-λ-5. Site-directed mutagenesis was conducted by PCR using outward facing primers containing the desired mutation and with compatible overlapping regions. Amplification of the wild-type template plasmid was cycled 18 times, and the reaction was chemically transformed into DH5α cells. In-frame deletions of defence systems were constructed by transforming a temperature-sensitive plasmid expressing λ-red recombinase into the target strain. Oligos with overlapping regions to the genome targeted for deletion were used to amplify a kan$^R$-resistance marker. The amplicon was then electroporated into the target strain induced to express λ-red, and recombinants were selected on kanamycin. The recombinase plasmid was then cured from the target strain by growth at 37 °C.

**Efficiency of plaquing assays.** 50 µl of overnight culture were mixed with 3 ml LB 0.5% agar and overlayed onto plates containing LB 1.2% agar and appropriate antibiotics. Two microlitres of phage from a tenfold serial dilution of stocks were pipetted onto the surface of the overlay plate. Spots were allowed to dry and incubated at 37 °C until plaques were visible. Plaques were then enumerated and EOP was measured as total plaque-forming units on the experimental strain divided by plaque-forming units on the control WT strain. Often, individual plaques were not distinguishable, that is, no viable phage were produced in an infection, resulting in a lysis zone but no discrete plaques. In such an event, samples were counted as having one plaque on the last dilution that showed lysis. For EOP assays with TAC systems, chaperone expression was titrated by overlaying cultures on media with increasing concentrations of arabinose before spotting phage dilutions.

**Bacteriophage adsorption assay.** The method is adapted from ref. [42]. Overnight bacterial cultures were diluted 1:100 and grown to OD$_{600}$ = 0.5. Cultures were infected at an MOI of 0.1. Samples were then incubated at 37 °C (T4 and $\lambda_{vir}$) or 25 °C (T7) for 15, 25 or 15 min for T4, $\lambda_{vir}$ and T7, respectively. Samples (500 µl) were then added to a tube of ice-cold chloroform and vortexed, and unadsorbed phage were enumerated by the top agar overlay method using a susceptible

indicator strain. Percent adsorption was determined relative to a simultaneous mock control experiment that contained growth medium but no host cells.

**Abortive infection assays.** Overnight cultures were normalized to $OD_{600} = 1.0$ and then diluted 100-fold. Diluted cultures (150 µl) were dispensed in a flat-bottomed 96-well plate. Phage dilutions (10 µl) were added to each well such that the MOI varied from 50 to 0.005. Wells were then overlayed with 20 µl mineral oil, and plates were covered with a breathable membrane. Plates were incubated at 37 °C in a Biotek Synergy H1 microplate reader. $OD_{600}$ was measured every 15 min. Three technical replicates were conducted for each strain.

**Toxicity assays.** Strains containing plasmids with inducible promoters were grown overnight at 37 °C in LB under repressing conditions (LB or LB + 0.2% glucose). Cultures were washed in LB, and tenfold serial dilutions were spotted on LB agar with and without inducer (0.2% arabinose, 200 µg ml$^{-1}$ vanillate, or 100 µg ml$^{-1}$ anhydrotetracycline). Plates were incubated overnight at 37 °C.

**Bioinformatic analyses.** Sanger sequences of fosmid ends were mapped to their strains of origin using BLASTn[43], followed by manual inspection. General remote domain prediction was done using the HHpred online web server (https://toolkit.tuebingen.mpg.de/) or locally (HHblits and HHsearch) against Pfam A domains (v. 35.0)[19,44]. To label domains in Fig. 3, we used the top HHsearch hit for each independent region of the protein. If there were many good matches, the bounds of the predicted domain were taken from the top hit, while the label was chosen on the basis of the Pfam clan to which the top hits belonged. The only exception was PD-λ-5, for which the top hit, 'methyltransferase', was chosen as we deemed it more descriptive than the Pfam clan designation. Investigation of the P2 defence hotspot was conducted by identifying homologues of the P2 portal protein using BLASTp, extracting the surrounding genes, clustering to 30% protein identity and visualizing using Clinker and clustermap.js[45]. Pangenome analyses were performed by annotating Genbank assemblies with Prokka[46], followed by analysis with Roary[47]. The phylogenetic tree was generated using FastTree[48] on the core genome alignment produced by Roary, using a generalized time-reversible model. We used DefenseFinder[49] on our protein sequences to determine whether homology to known systems could be detected. Other software that were instrumental for routine genome analyses were PATRIC web server, DNAFeaturesViewer and Mauve[50–52].

To assess whether defence systems were potential toxin-antitoxin systems, we used WU-BLAST 2.0 to search against TADB v2.0 (https://bioinfo-mml.sjtu.edu.cn/cgi-bin/TADB2/nph-blast-TADB.pl)[53].

**Identification of defence system homologues and genomic context analysis.** For each defence system, we searched for homologues of each individual component using blastp against all bacterial proteins in the NCBI non-redundant (nr) protein database using the following parameters: -evalue 0.00001 -qcov_hsp_perc 80 for single-gene systems and -evalue 0.00001 for multi-gene systems. The NCBI nr database was downloaded for local use in March 2021. All instances of the homologues identified in the nr search were then located within all full bacterial genomes ($n = 844{,}603$) downloaded from Genbank and RefSeq (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/) in April 2021. For multi-gene systems, the system was only considered present in a given genome if all components of the system were present in the same genomic region. The only exception was CmdTAC in which otherwise clearly homologous systems were widely variable in the antitoxin sequence. For this system, homologues were required to have a CmdT homologue, a second downstream ORF and a SecB-like chaperone as the third component.

The local genomic context of a defence system homologue was defined as all coding sequences ±10 kb of the system (or to the end of a contig if less than 10 kb). All coding sequences within this local context were searched for defence-related and prophage-related domains using HMMER3 hmmscan[54] with *E*-value cut-offs of $10^{-5}$ and $10^{-15}$ for the defence-related and prophage-related domain searches, respectively. For the defence-related domain search, sequences were searched against defence-related pfam and COG domains identified and used in refs. [11–13]. We considered each gene flanking a given homologue separately, even if multiple adjacent genes were part of a single multi-component defence system. For the prophage-related domain search, sequences were searched against all pVOG[55] domains available as of May 2021 when the pVOG database was downloaded. For scatterplots and marginal histograms in Fig. 4 and Extended Data Fig. 5, any regions with <10 coding sequences (that is, located on short contigs) were excluded. In native context schematics, prophage genes were predicted by annotation, pVOG analysis and BLASTp against the ACLAME database[56].

To compare defence and prophage context between our systems and those that were identified computationally and subsequently validated in refs. [12,13] (Extended Data Fig. 6), we identified homologues of each system and their flanking genes as described above. The collected flanking proteins were then clustered using the function cluster within MMseqs2[57] with the following parameters: –cluster-mode 1–min-seq-id 0.9. Each resulting cluster was called as defence- or prophage-related if at least 90% of the proteins within the cluster contained the same defence- or prophage-related domain(s), respectively. This clustering helps to control for overrepresentation of closely related sequences. Defence and prophage enrichment for a given system was then calculated as the number of defence- or prophage-domain-containing clusters divided by the total number of clusters.

**Taxonomy analysis.** The taxonomic distribution of each system was defined by the system's presence across the downloaded bacterial genomes with the same parameters as described above. For a given genome with a defence system present, the NCBI taxid was extracted and translated to major bacterial classes using taxon kit[58]. For comparison, we also examined the taxonomic distribution of the following known systems: type I-IV RM systems, EcoKI, EcoRI, EcoPI and McrBC, respectively; P2 old, AAD03309.1; Cas9, WP_032462936.1; Zorya I, system containing BV17222.1; Zorya II, system containing ACA79490.1; ToxN, WP_000675353.1; Kiwa, system containing AEZ43441.1; Druantia, system containing ERA40829.1.

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

## Data availability

All data generated or analysed during this study are included in the published article (and its Supplementary Information).

## Code availability

Code used in this study is available at https://github.com/chrisdoering8197/phagedefense.

Received: 12 May 2022; Accepted: 28 July 2022;
Published online: 19 September 2022

## References

1. Mushegian, A. R. Are there $10^{31}$ virus particles on Earth, or more, or fewer? *J. Bacteriol.* https://doi.org/10.1128/JB.00052-20 (2020).
2. Suttle, C. A. The significance of viruses to mortality in aquatic microbial communities. *Microb. Ecol.* **28**, 237–243 (1994).
3. Rostøl, J. T. & Marraffini, L. Ph)ighting phages: how bacteria resist their parasites. *Cell Host Microbe* **25**, 184–194 (2019).
4. Cohen, D. et al. Cyclic GMP–AMP signalling protects bacteria against viral infection. *Nature* **574**, 691–695 (2019).
5. Burroughs, A. M. & Aravind, L. Identification of uncharacterized components of prokaryotic immune systems and their diverse eukaryotic reformulations. *J. Bacteriol.* https://doi.org/10.1128/JB.00365-20 (2020).
6. Ye, Q. et al. HORMA domain proteins and a Trip13-like ATPase regulate bacterial cGAS-like enzymes to mediate bacteriophage immunity. *Mol. Cell* **77**, 709–722.e7 (2020).
7. Hsu, B. B. et al. Dynamic modulation of the gut microbiota and metabolome by bacteriophages in a mouse model. *Cell Host Microbe* **25**, 803–814.e5 (2019).
8. Hesse, S. & Adhya, S. Phage therapy in the twenty-first century: facing the decline of the antibiotic era; is it finally time for the age of the phage? *Annu. Rev. Microbiol.* **73**, 155–174 (2019).
9. Mirzaei, M. K. & Deng, L. New technologies for developing phage-based tools to manipulate the human microbiome. *Trends Microbiol.* **30**, 131–142 (2022).
10. Javaudin, F., Latour, C., Debarbieux, L. & Lamy-Besnier, Q. Intestinal bacteriophage therapy: looking for optimal efficacy. *Clin. Microbiol. Rev.* https://doi.org/10.1128/CMR.00136-21 (2021).
11. Makarova, K. S., Wolf, Y. I., Snir, S. & Koonin, E. V. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J. Bacteriol.* https://doi.org/10.1128/JB.05535-11 (2011).
12. Doron, S. et al. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* https://doi.org/10.1126/science.aar4120 (2018).
13. Gao, L. et al. Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science* https://doi.org/10.1126/science.aba0372 (2020).
14. Garretto, A. et al. Genomic survey of *E. coli* from the bladders of women with and without lower urinary tract symptoms. *Front. Microbiol.* **11**, 2094 (2020).
15. Ochman, H. & Selander, R. K. Standard reference strains of *Escherichia coli* from natural populations. *J. Bacteriol.* https://doi.org/10.1128/jb.157.2.690-693.1984 (1984).
16. Chopin, M.-C., Chopin, A. & Bidnenko, E. Phage abortive infection in *Lactococci*: variations on a theme. *Curr. Opin. Microbiol.* **8**, 473–479 (2005).
17. Takahashi, H., Coppo, A., Manzi, A., Martire, G. & Pulitzer, J. F. Design of a system of conditional lethal mutations (tab/k/com) affecting protein-protein interactions in bacteriophage T4-infected *Escherichia coli*. *J. Mol. Biol.* **96**, 563–578 (1975).
18. Luria, S. E. & Human, M. L. A nonhereditary, host-induced variation of bacterial viruses. *J. Bacteriol.* **64**, 557–569 (1952).

19. Zimmermann, L. et al. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.* **430**, 2237–2243 (2018).
20. Depardieu, F. et al. A eukaryotic-like serine/threonine kinase protects staphylococci against phages. *Cell Host Microbe* **20**, 471–481 (2016).
21. Garb, J. et al. Multiple phage resistance systems inhibit infection via SIR2-dependent NAD+ depletion. Preprint at *bioRxiv* http://biorxiv.org/lookup/doi/10.1101/2021.12.14.472415 (2021).
22. Barth, Z. K., Nguyen, M. H. & Seed, K. D. A chimeric nuclease substitutes a phage CRISPR-Cas system to provide sequence-specific immunity against subviral parasites. *eLife* **10**, e68339 (2021).
23. Anantharaman, V., Makarova, K. S., Burroughs, A. M., Koonin, E. V. & Aravind, L. Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. *Biol. Direct* **8**, 15 (2013).
24. Odegrip, R., Nilsson, A. S. & Haggård-Ljungquist, E. Identification of a gene encoding a functional reverse transcriptase within a highly variable locus in the P2-like coliphages. *J. Bacteriol.* https://doi.org/10.1128/JB.188.4.1643-1647.2006 (2006).
25. Rousset, F. et al. Phages and their satellites encode hotspots of antiviral systems. *Cell Host Microbe* https://doi.org/10.1016/j.chom.2022.02.018 (2022).
26. Rocha, E. P. C. & Bikard, D. Microbial defenses against mobile genetic elements and viruses: who defends whom from what? *PLoS Biol.* **20**, e3001514 (2022).
27. Koonin, E. V., Makarova, K. S., Wolf, Y. I. & Krupovic, M. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nat. Rev. Genet.* **21**, 119–131 (2020).
28. Van Melderen, L. Toxin–antitoxin systems: why so many, what for? *Curr. Opin. Microbiol.* **13**, 781–785 (2010).
29. LeRoux, M. et al. The DarTG toxin-antitoxin system provides phage defence by ADP-ribosylating viral DNA. *Nat. Microbiol.* **7**, 1028–1040 (2022).
30. Dy, R. L., Przybilski, R., Semeijn, K., Salmond, G. P. C. & Fineran, P. C. A widespread bacteriophage abortive infection system functions through a Type IV toxin–antitoxin mechanism. *Nucleic Acids Res.* **42**, 4590–4605 (2014).
31. Koga, M., Otsuka, Y., Lemire, S. & Yonesaki, T. *Escherichia coli* rnlA and rnlB compose a novel toxin–antitoxin system. *Genetics* **187**, 123–130 (2011).
32. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Comprehensive comparative-genomic analysis of Type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. *Biol. Direct* **4**, 19 (2009).
33. Bordes, P. et al. Chaperone addiction of toxin–antitoxin systems. *Nat. Commun.* **7**, 13339 (2016).
34. Hussain, F. A. et al. Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science* https://doi.org/10.1126/science.abb1083 (2021).
35. Chen, C. K.-M., Chan, N.-L. & Wang, A. H.-J. The many blades of the β-propeller proteins: conserved but versatile. *Trends Biochem. Sci.* **36**, 553–561 (2011).
36. Berkner, K. L. & Folk, W. R. The effects of substituted pyrimidines in DNAs on cleavage by sequence-specific endonucleases. *J. Biol. Chem.* **254**, 2551–2560 (1979).
37. Krüger, D. H., Schroeder, C., Hansen, S. & Rosenthal, H. A. Active protection by bacteriophages T3 and T7 against *E. coli* B-and K-specific restriction of their DNA. *Mol. Gen. Genet.* **153**, 99–106 (1977).
38. Loenen, W. A. M. & Raleigh, E. A. The other face of restriction: modification-dependent enzymes. *Nucleic Acids Res.* **42**, 56–69 (2014).
39. Georgopoulos, C. P. Bacterial mutants in which the gene N function of bacteriophage lambda is blocked have an altered RNA polymerase. *Proc. Natl Acad. Sci.* **68**, 2977–2981 (1971).
40. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
41. Chung, C. T., Niemela, S. L. & Miller, R. H. One-step preparation of competent *Escherichia coli*: transformation and storage of bacterial cells in the same solution. *Proc. Natl Acad. Sci.* **86**, 2172–2175 (1989).
42. Clokie, M. R. J. & Kropinski, A. M. (eds) *Bacteriophages: Methods and Protocols* Vol. 1 (Springer, 2009).
43. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
44. Mistry, J. et al. Pfam: the protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
45. Gilchrist, C. L. M. & Chooi, Y.-H. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* **37**, 2473–2475 (2021).
46. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
47. Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
48. Price, M. N., Dehal, P. S. & Arkin, A. P. Fasttree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
49. Tesson, F. et al. Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat. Commun.* **13**, 2561 (2022).
50. Davis, J. J. et al. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res.* **48**, D606–D612 (2020).
51. Zulkower, V. & Rosser, S. DNA Features Viewer: a sequence annotation formatting and plotting library for Python. *Bioinformatics* **36**, 4350–4352 (2020).
52. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
53. Xie, Y. et al. TADB 2.0: an updated database of bacterial type II toxin–antitoxin loci. *Nucleic Acids Res.* **46**, D749–D753 (2018).
54. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
55. Grazziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* **45**, D491–D498 (2017).
56. Leplae, R., Lima-Mendez, G. & Toussaint, A. ACLAME: A CLAssification of Mobile genetic Elements. *Nucleic Acids Res.* **38**, D57–D61 (2010).
57. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
58. Shen, W. & Ren, H. TaxonKit: a practical and efficient NCBI taxonomy toolkit. *J. Genet. Genomics* **48**, 844–850 (2021).


## Acknowledgements

We thank I. Frumkin, C. Guegler, M. LeRoux, S. Srikant and T. Zhang for comments on the manuscript. This work was supported by a National Institutes of Health, NIGMS grant 5F32 GM139231-02 to C.N.V and a Gilliam Fellowship from the Howard Hughes Medical Institute to C.R.D. This work was also funded by an MIT-Skoltech grant to M.T.L., who is an Investigator of the Howard Hughes Medical Institute. This article is subject to HHMI's Open Access to Publications policy. HHMI lab heads have previously granted a non-exclusive CC BY 4.0 license to the public and a sublicensable license to HHMI in their research articles. Pursuant to those licenses, the author-accepted manuscript of this article can be made freely available under a CC BY 4.0 license immediately upon publication.

## Author contributions

M.T.L and C.N.V. designed the experiments and wrote the manuscript. C.N.V designed and implemented the screen. C.R.D. wrote code to collect homologue data. C.N.V. and C.R.D. analysed data. C.N.V., C.R.D., G.I.C.T. and M.L.L. contributed to routine experiments and cloning.

## Competing interests

The authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at https://doi.org/10.1038/s41564-022-01219-4.

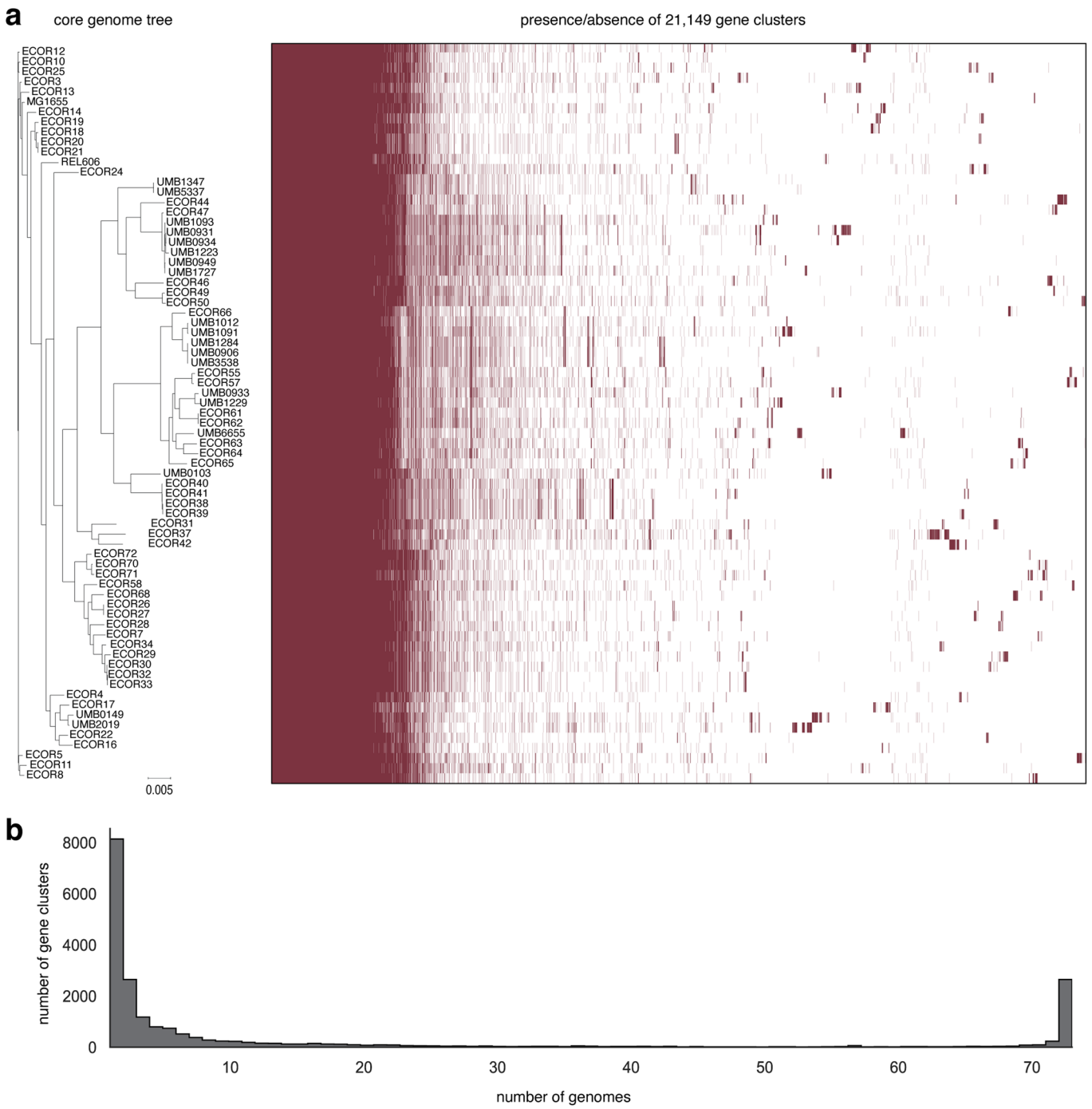
**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41564-022-01219-4.

**Correspondence and requests for materials** should be addressed to Michael T. Laub.

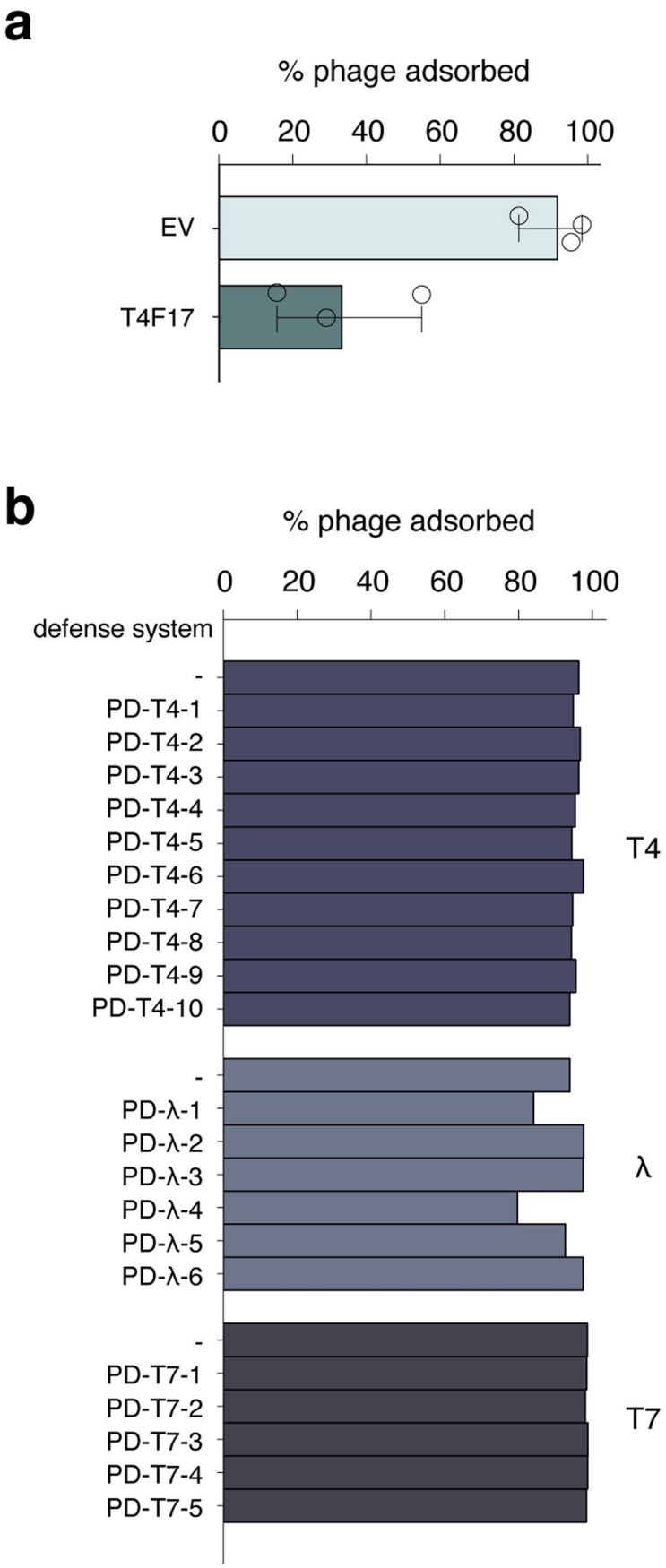
**Peer review information** *Nature Microbiology* thanks Karen Maxwell and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at www.nature.com/reprints.

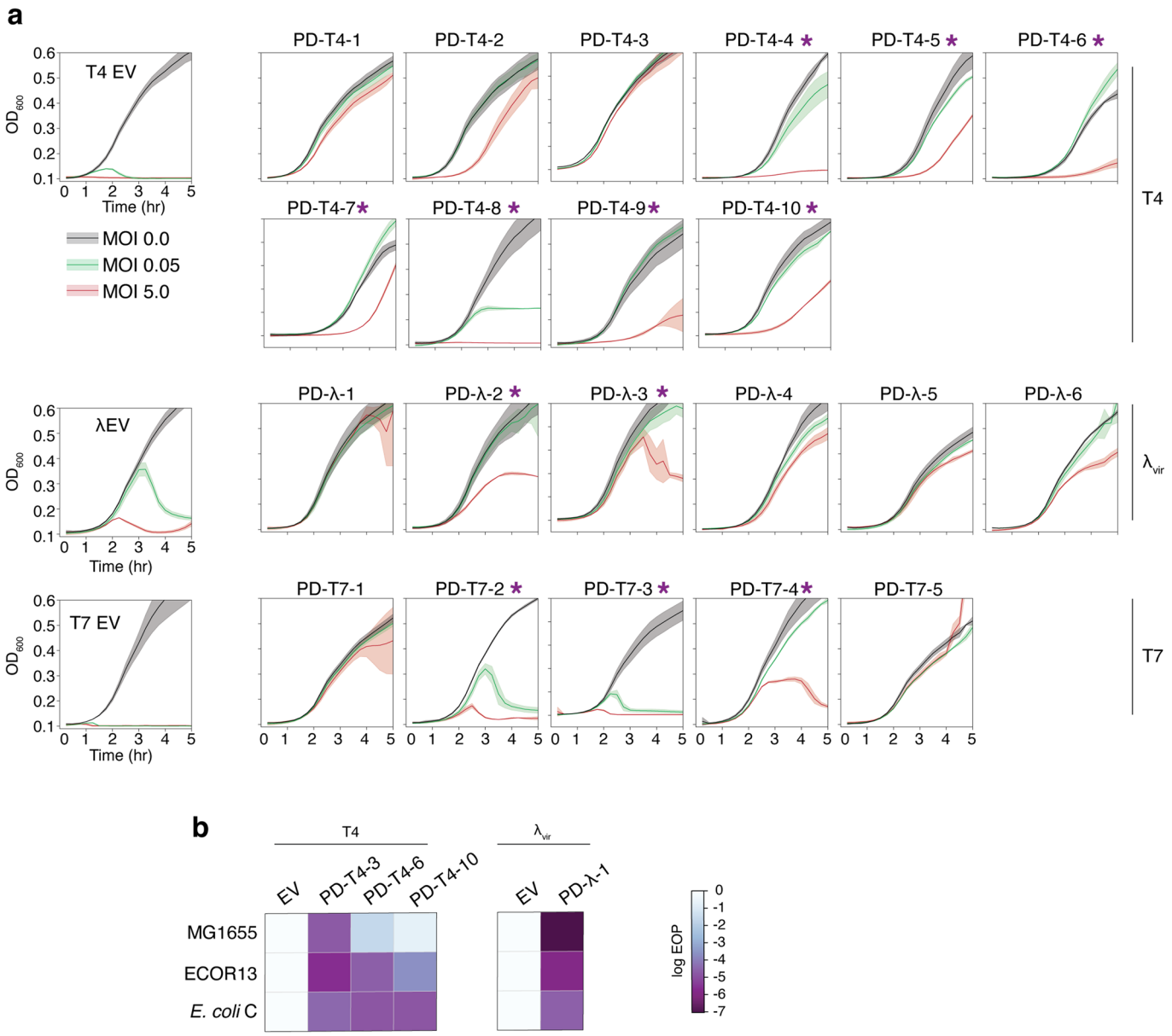
**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



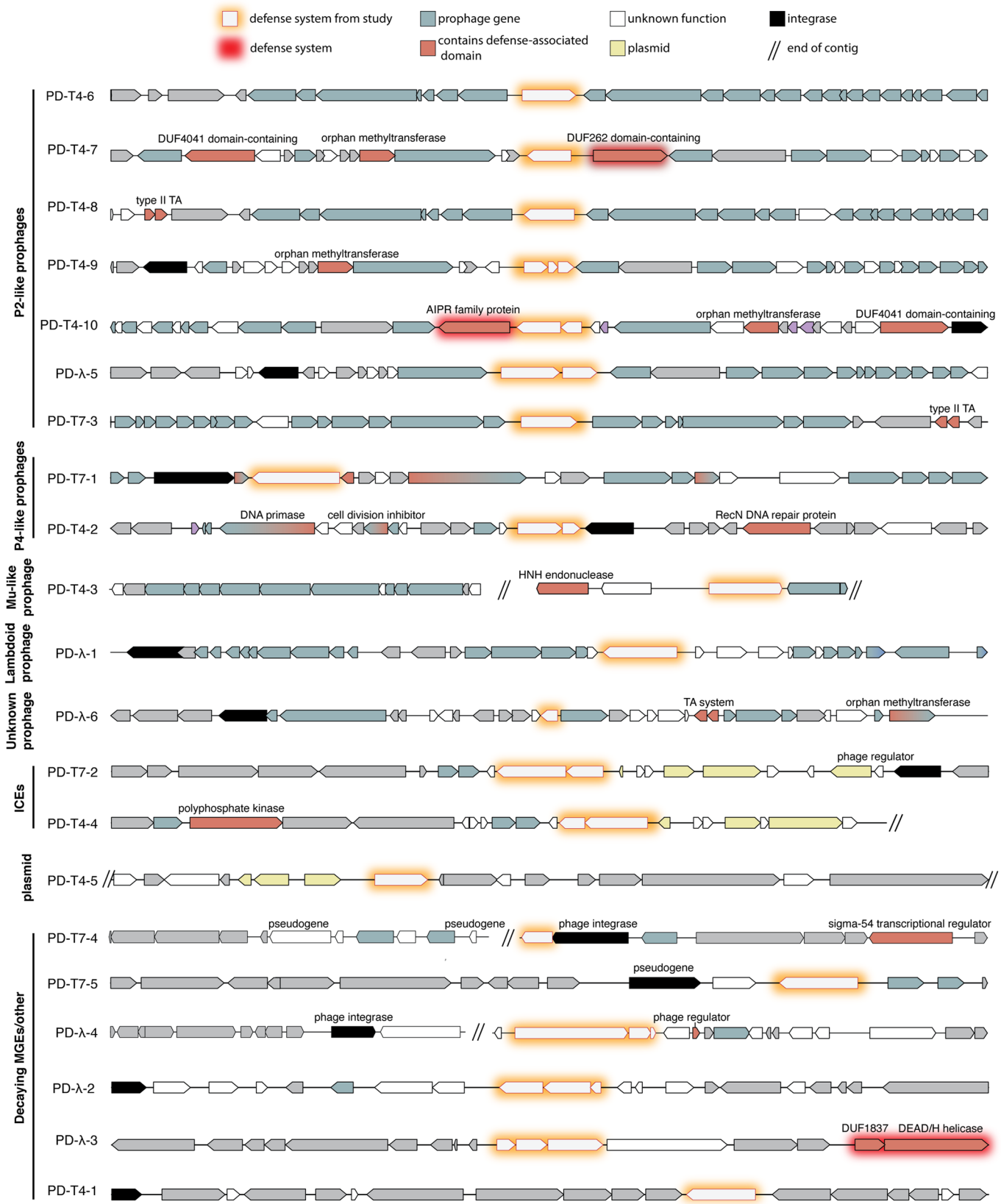

**Extended Data Fig. 1 | Diversity of the *E. coli* isolate pangenome used in this study.** (**a**) (*Left*) Phylogenetic tree of *E. coli* strain collection used to construct the genomic library screened. *E. coli* K-12 (MG1655) and B (REL606) are also included. (*Right*) Bars indicate presence/absence (red/white) of individual gene clusters (95% identity threshold). (**b**) Plot of the number of gene clusters versus the number of strains they are found in, *for example* ~8,000 clusters are each found in only one genome. These sparsely conserved clusters represent the accessory genome, whereas ~3,000 clusters are found in all 73 genomes and represent the *E. coli* core genome.



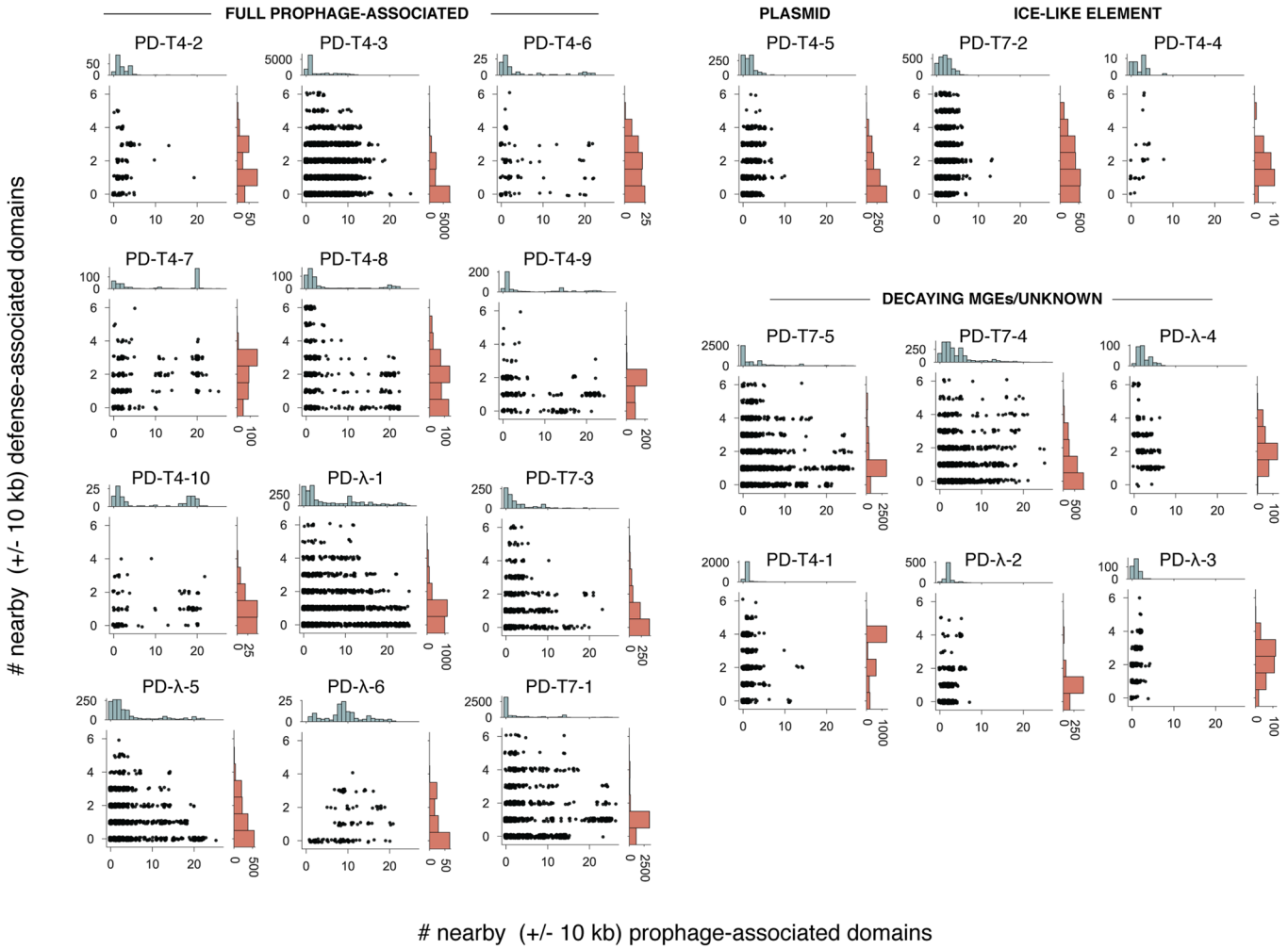

**Extended Data Fig. 2 | Adsorption of bacteriophage on various strains.** (**a**) Mean adsorption of T4 on control (EV) and LPS-containing fosmid strains (T4F17). Error bars represent standard deviation of three biological replicates. (**b**) Adsorption of T4, $\lambda_{vir}$, or T7 on strains expressing defense systems against their respective phages.



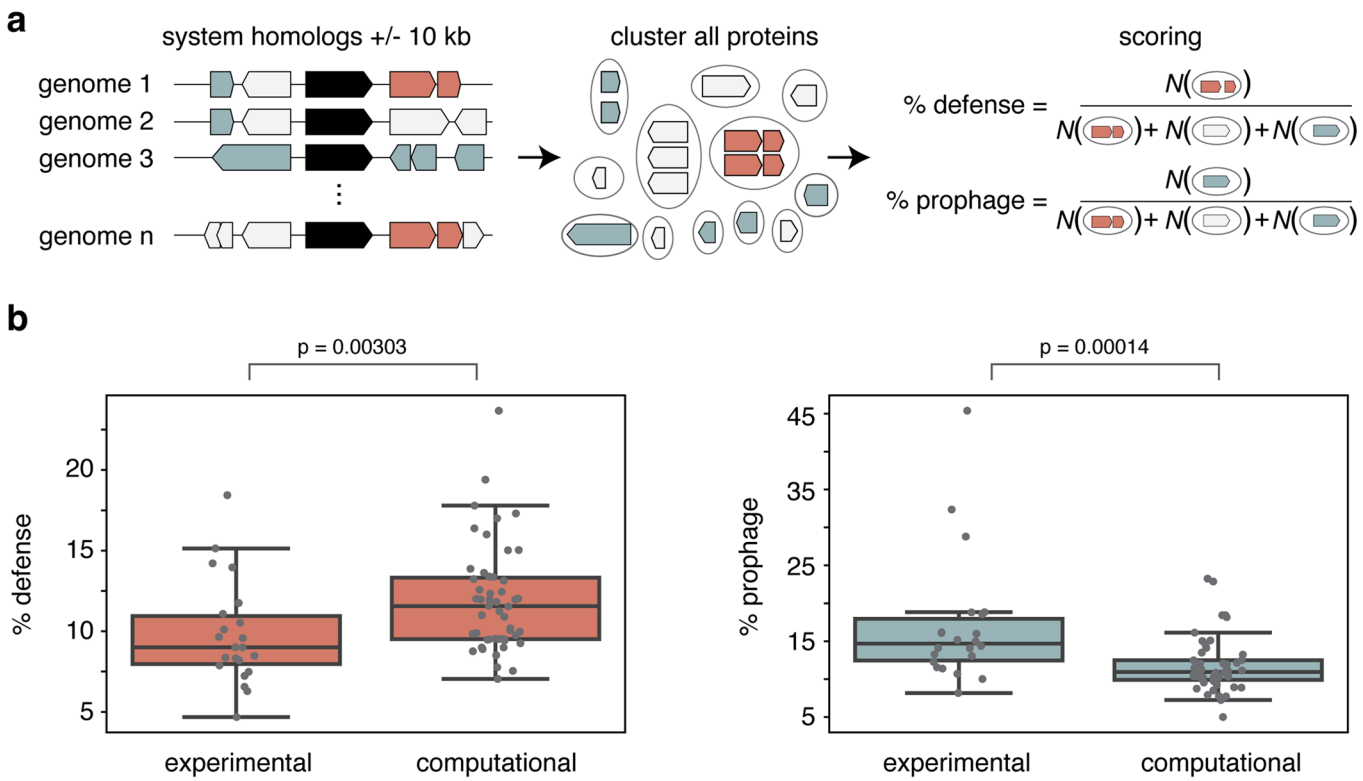

**Extended Data Fig. 3 | Mechanisms of defense and defense system function in other strains.** (**a**) Growth of control (EV) or defense system-expressing strains with phage MOIs of 0, 0.05 or 5. Phages used in each experiment are shown on the right. Asterisks indicate systems not showing direct immunity and likely representing abortive infection mechanisms. Lines represent the mean of three technical replicates with shaded regions indicating standard deviation. (**b**) EOP measurements for T4 and $\lambda_{vir}$ on *E. coli* strains MG1655, ECOR13, or C expressing various defense systems indicated.




**Extended Data Fig. 4 | Native genomic neighborhoods of newly identified defense systems.** Genome maps of native locations of the defense systems, showing the flanking 10 kb regions, unless interrupted by the end of a contig. Prophage and defense-domain containing genes were called as described in the Methods.




**Extended Data Fig. 5 | Genome context of defense system homologs.** Data as in Fig. 4c, extended to all systems discovered here and sorted by the MGE context in which they were found.




**Extended Data Fig. 6 | Comparison of defense and prophage enrichments between experimentally and computationally discovered systems.** (**a**) Overview of method. Genes +/- 10 kb of homologs of defense systems were predicted as defense- or prophage-associated. To minimize the effects of sequence/genome redundancy, proteins were clustered to 95% identity. % scores were calculated as the number of prophage or defense-associated clusters over total clusters. (**b**) Boxplots of % prophage- and defense-associated genes near the experimentally discovered systems (this study) or computationally predicted and validated systems[12,13]. *p* values indicate significance from two-sided Mann-Whitney U test. Boxes indicate bounds of the distribution as median +/- quartiles, and limits exclude outliers.

nature portfolio

Corresponding author(s): Michael Laub

Last updated by author(s): July 25, 2022

# Reporting Summary

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

## Statistics

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

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

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

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | BLAST v. 2.11.0+ for homolog searches. HMMER3 v. 3.1b2 for domain searches. Taxonkit v. 0.8.0 to gather taxonomy information. MMseqs2 v. 13-45111 for protein sequence clustering. Custom code was generated, using these tools, to detect homologous proteins and their genome context from NCBI genomes. This code is publicly available at https://github.com/chrisdoering8197/phagedefense |
| Data analysis | HHsuite v. 3.1.0 for remote homology detection **(includes HHblits and HHsearch)**. Prokka v. 1.14.6 for genome annotation. Roary v. 3.12.0 for pangenome analysis. FastTree v. 2.1.10 for phylogenetic tree construction. **DefenseFinder v1.0 for prediction of known defense systems.** |

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

## Data

Policy information about availability of data

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

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

All data generated or analyzed during this study are included in the published article (and its supplementary information files).

# Field-specific reporting

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

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

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

# Life sciences study design

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

| | |
|---|---|
| Sample size | **No sample size calculation was performed.** Given large effect sizes, we chose to replicate experiments in triplicate **as is standard** to indicate reproducibility. |
| Data exclusions | Only representative sub-library peaks were included in Fig. 2C because the final identification of genes and their functions is confirmed in the subsequent figures. |
| Replication | All experimental findings were repeated at least twice. All reported results were successfully reproduced. |
| Randomization | **No experimental groups or control groups were subjectively chosen and there are no covariates to control for as experiments were done in isogenic strains.** No experiments required randomization. |
| Blinding | Blinding was not relevant because all data were **obtained objectively and had** strong effect sizes and raw data is reported in the manuscript. **All comparative analyses were performed in isogenic strain backgrounds.** |

# Reporting for specific materials, systems and methods

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

## Materials & experimental systems

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

## Methods

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