The extracellular contractile injection system is enriched in environmental microbes and associates with numerous toxins

Alexander Martin Geller^{1,*}, Inbal Pollin^{1,*}, David Zlotkin¹, Aleks Danov¹, Nimrod Nachmias¹, William B Andreopoulos², Keren Shemesh¹, and Asaf Levy^{1,♀}

¹ Department of Plant Pathology and Microbiology, the Robert H. Smith Faculty of Food and Environment, the Hebrew University of Jerusalem, Rehovot, Israel.

² Department of Computer Science, San Jose State University, California, USA

* These authors contributed equally.

♀ Corresponding author, alevy@mail.huji.ac.il

Abstract

Bacteria employ toxin delivery systems to exclude bacterial competitors and to infect host cells. Characterization of these systems and the toxins they secrete is important for understanding microbial interactions and virulence in different ecosystems. The extracellular Contractile Injection System (eCIS) is a toxin delivery particle that evolved from a bacteriophage tail. Four known eCIS systems have been shown to mediate interactions between bacteria and their invertebrate hosts, but the broad ecological function of these systems remains unknown. Here, we identify eCIS loci in 1,249 prokaryotic genomes and reveal a striking enrichment of these loci in environmental microbes and absence from mammalian pathogens. We uncovered 13 toxin genes that associate with eCIS from diverse microbes and show that they can inhibit growth of bacteria, yeast or both. We also found immunity genes that protect bacteria from self-intoxication, supporting an antibacterial role for eCIS. Furthermore, we identified multiple new eCIS core genes including a conserved eCIS transcriptional regulator. Finally, we present our data through eCISem; an extensive eCIS repository. Our findings define eCIS as a widespread environmental prokaryotic toxin delivery system that likely mediates antagonistic interactions with eukaryotes and prokaryotes. Future understanding of eCIS functions can be leveraged for the development of new biological control systems, antimicrobials, and cell-free protein delivery tools.

Introduction

The extracellular contractile injection system (eCIS; previously termed “PLTS”, phage-like protein-translocation structures) is a cell-free protein delivery system that is prevalent in bacteria and archaea, but its biological function is poorly understood. The eCIS particle resembles the contractile tail of a T4 bacteriophage and is mostly encoded by an operon of 15-28 genes ¹. This protein complex is 110-120 nm long and includes a baseplate, a sheathed hollow tube that has a needle-like tip (spike) on one side and a cap on the other side,

and tail fibers that likely serve to adhere to target cells (Figure S1)^{2,3}. eCIS contraction propels the tube out of the sheath, likely enabling the sharp tip to perforate the target cell membrane. Effectors, which are proteins that are injected by the particle eCIS into target cells upon contraction. These effectors are usually encoded at the 3' end of the operon⁴⁻⁶. The effectors that have been studied were shown to perform enzymatic activities in the target eukaryotic cell, most of which lead to cell toxicity. eCIS shares structural similarity with other contractile nano weapons such as type VI secretion system (T6SS) and R-type pyocins but differs from these by being extracellular and by injecting effectors into the target cell instead of just perforating it, respectively. Despite the prevalence of eCISs across the microbial world, only four eCIS loci, Afp, AfpX, PVC, and MACs, have been experimentally studied.

The Antifeeding Prophage (Afp) is encoded by *Serratia entomophila* and it is sufficient to cause feeding cessation and death to the New Zealand grass grub pest⁷⁻¹⁰. A homologous Afp structure, termed AfpX, was described in *Serratia proteamaculans* and demonstrated insecticidal activity against the larvae of two insects⁶. Effectors of Afp and AfpX have not been confirmed. Photorhabdus Virulence Cassettes (PVC) was characterized in *Photorhabdus* spp. and has an injectable insecticidal activity against wax moth larvae⁵. Four effectors are associated with PVC and demonstrate cytotoxicity against eukaryotic cells: Plu1690⁵, Pnf⁴, RRSP_{pa}¹¹, and SepC-like⁵. The high-resolution structures of Afp and PVC particles were recently solved using cryo-EM and provided excellent maps of the protein interactions within these particles^{2,3}. Unlike the insecticidal activities of Afp, AfpX, and PVC, the metamorphosis-associated contractile structures (MACs) encoded by *Pseudoalteromonas luteoviolacea* plays a mutualistic role. MACs loci produce arrays of ~100 eCIS structures that trigger metamorphosis of the marine tubeworm *Hydroides elegans* hosting the bacteria¹². MACs particles inject a nuclease effector that kills eukaryotic cell lines but this activity is not essential for metamorphosis¹³. Mif1 protein was shown as a cargo inside the MACs tube lumen but was not shown to act as a toxin¹⁴. To summarize, eCIS particles interact with different invertebrates and inject effectors that are toxic to eukaryotic cells. To the best of our knowledge, only seven eCIS effectors were experimentally validated as toxic to recipient cells or are known to be injected by eCIS. All known effectors are encoded in gammaproteobacteria.

Here we provide a systematic characterization of eCIS in the microbial world. We identified 1,425 eCIS loci encoded within 1,249 prokaryotic genomes, corresponding to 1.9% of the analyzed microbial genomes. Strikingly, eCIS loci are strongly enriched in environmental microbial taxa from different ecosystems and in microbiomes of specific hosts (plants, specific animals, protists), and are depleted from mammalian pathogens that have been extensively cultured and sequenced. We analyzed the proteins and protein domains of all eCIS loci and identified new core protein domains, including a putative metallopeptidase that associates with toxins and a putative master transcriptional regulator. We further bioinformatically analyzed the fibre proteins of eCIS

that confer cell specificity and identified the first group of eCIS particles likely targeting bacteria. We identified a large set of putative toxins (effectors) that are genetically associated with eCIS operons, the majority of which are specific to certain eCIS loci. Through heterologous expression experiments we showed that 13 toxins were capable of killing *E. coli* and/or *S. cerevisiae* cells. In some cases, we were able to experimentally identify immunity genes against that rescued bacteria from self-intoxication by their cognate toxins, supporting the toxin intended activity against bacteria. Finally, we developed eCISem; an online resource to present our findings that will serve as a valuable resource for the research community. We thus provide here an in-depth ecological and functional characterization of a poorly studied toxin delivery system and its associated toxins.

Results

eCIS are encoded by 1.9% and 1.2% of sequenced bacteria and archaea, respectively, with a highly biased taxonomic distribution

First, we were interested in identifying all eCIS loci in a large genomic dataset. We compiled a set of 64,756 microbial isolate genomes retrieved from Integrated Microbial Genomes ¹⁵. To identify core component homologs from known systems, we searched for genes with known eCIS-associated pfam annotations. To supplement this, we also annotated homologous genes ourselves by searching using Hidden Markov Model (HMM) profiles from a recent publication^{1,16}. We defined putative eCIS operons as gene cassettes that included these multiple eCIS core genes in close proximity, and were not bacteriophage, T6SS, or R-type pyocins (Materials and Methods). Overall, we identified eCIS operons encoded in 1,230 (1.9%) bacteria and 19 (1.2%) archaea from our genomic repository (Table S1-S2). We identified two core genes, *Afp8* and *Afp11*, that are present together across 98.7% all eCIS loci and used their protein sequences to construct an eCIS phylogenetic gene tree (Figure 1a). eCIS is scattered across the prokaryotic diversity with presence in 14 bacterial phyla and in one archaeal phylum. The incongruence between this tree and the genomic phylogeny suggests that eCIS undergo HGT frequently, as was proposed before ^{1,16}. The previously experimentally characterized eCIS are located within a narrow clade on the eCIS tree, pointing to the possibility that other eCIS particles may play more diverse ecological roles (Figure 1a).

Figure 1. Taxonomic Distribution of eCIS-encoding microbes. **A.** A phylogenetic tree of eCIS across the microbial world. eCIS core genes Afp8 and Afp11 from each operon were concatenated, aligned, and used to construct the phylogenetic tree. The Domain and Phylum corresponding to each leaf are indicated in the inner and outer rings, respectively. Scaffolds encoding eCIS were predicted to be plasmids using Deeplasmid and marked with a triangle. Previously experimentally investigated eCIS are marked on their respective leaves (2 o'clock). **B.** eCIS distribution in different genera. We calculated the eCIS distribution across genera using a Fisher exact test. The Odds Ratio represents the enrichment or depletion magnitude, with redder colors representing enrichment, and bluer colors representing depletion. Calculated p values were corrected for multiple testing using FDR to yield minus log10 q values, shown in shades of grey.

Next, we looked for genetic mechanisms that may mediate the eCIS HGT. Using Deeplasmid, a new plasmid prediction tool that we developed, we identified that 7.6% of eCIS are likely plasmid-borne (Figures 1a and S2, Table S3, Materials and Methods). In other cases, we found a clear signature of eCIS operon integration into a specific bacterial chromosome (Figure S3). For example, we identified a likely homologous recombination event between identical tRNA genes, a classical integration site¹⁷ (Figure S3b). These genomic integration events and the plasmid-borne eCIS operons shed light on the mechanisms through which eCIS loci have been horizontally propagated in microbial genomes.

eCIS displays a highly biased taxonomic distribution

Given the propensity of eCIS to transfer between microbes as phylogenetically distant as bacteria and archaea, we were surprised by its scarcity in microbial genomes. We tested if eCIS loci are homogeneously distributed across microbial taxa and found that eCIS are mostly constrained to particular taxa (Figure 1b, Table S4). Strikingly, we found that it is present in 100% (18/18) of *Photorhabdus* genomes in our dataset with 2-5 eCIS operon copies per genome (Fisher exact test, odds ratio = infinity, q value = $2.97e^{-28}$), 89% of sequenced

Chitinophaga (odds ratio = 276, q value = 1.69e^{-35}), 86% of sequenced *Dickeya* (odds ratio = 211, q value = 3.78e^{-52}), and 69% of sequenced *Algoriphagus* (odds ratio = 73, q value = 1.99e^{-24}). These genera are known as environmental microbes; *Photorhabdus* is a commensal of entomopathogenic nematodes¹⁸, *Chitinophaga* is a soil microbe and a fungal endosymbiont¹⁹, *Dickeya* is a plant and pea aphid pathogen^{20,21}, and *Algoriphagus* is an aquatic or terrestrial microbe^{22–26}. In contrast, eCIS is strongly depleted from the most cultured and sequenced genera of Gram- positive and negative human pathogens, including *Staphylococcus*, *Escherichia*, *Salmonella*, *Streptococcus*, *Acinetobacter*, and *Klebsiella*. Strikingly, within these genera, for which our repository had 18,355 genomes, eCIS was totally absent (odds ratio = 0, q value $\leq 3.86\text{E-}16$ for each one of these genera), suggesting a very potent purifying selection acting against eCIS integration into these microbial genomes, despite the eCIS operon's tendency for extensive lateral transfer and its presence in other host-associated systems.

eCIS presence is highly correlated with specific ecosystems, microbial lifestyles, and microbial hosts

Given the strong eCIS taxonomic bias we identified, we were curious to know if we could further associate eCIS with specific ecological features. To this end, we retrieved metadata available for all sequenced genomes in our repository (Materials and Methods). These traits include the microbial isolation site, ecosystem and habitat, microbial lifestyle and physiology, and the organisms hosting the microbes (Table S5). We calculated the correlation of eCIS presence with certain microbial traits to identify significant enrichments and depletion patterns. This was done using a naïve enrichment test (Fisher exact test) together with a phylogeny-aware test, Scoary²⁷, which is used to correct for the phylogenetic bias of the isolate genomes. Using this test we quantify to what extent the eCIS presence in a genome correlates with a certain trait, independently of the microbial phylogeny (Figure 2). Notably, eCIS is positively correlated with terrestrial and aquatic environments, such as soil, sediments, lakes, and coasts, but is depleted from built environments and food production venues. In terms of microbial lifestyle and physiology, eCISs are enriched in environmental microbes, mostly symbiotic, and are depleted from pathogens (the vast majority of which were isolated from humans). eCISs are enriched in aerobic microbes that dwell in mild and cold temperatures. In general, the eCIS-encoding microbes tend to associate with terrestrial hosts including insects, nematodes, annelids, protists, fungi, and plants, and in aquatic hosts such as fish, sponges, and molluscs. Intriguingly, we detected a strong depletion from bacteria that were isolated from birds and mammals, including humans. Specifically, the operon is depleted from all tissues in which the human microbiome is abundant: oral and digestive systems, skin, and the urogenital tract. However, we detected a mild eCIS enrichment in the human gut commensal *Bacteroides* (Figure 1B) and *Parabacteroidetes* genera. *Bacteroides* was recently reported by the Shikuma group as being eCIS-rich²⁸.

Given these findings, we conclude that eCIS is a strongly enriched environmental microbial secretion system. Moreover, to the best of our knowledge, this is the first secretion system that exhibits such a strong genomic enrichment in environmental microbes and a depletion from mammalian and avian microbiomes. Therefore, we hypothesize that this environmental association points to the eCIS biological function and target organism specificity.

Figure 2. eCIS-encoding microbes' lifestyle and isolation. A Fisher exact test combined with a modified version of Scoary was used to perform a population-aware analysis of eCIS-encoding microbes' metadata. The Odds Ratio represents the enrichment or depletion magnitude, with redder colors representing enrichment, and bluer colors representing depletion. The negative log₁₀ of the q-values, shown in shades of grey, are corrected for multiple hypothesis testing. One q-value corresponds to the statistical significance of the Fisher exact test, and the other represents the same for the Scoary pairwise comparison test.

eCIS tail fibres provide information about target specificity and suggest targeting of bacteria

All previously studied eCIS target invertebrate cells. One of the proteins that likely mediates specific eCIS binding to target cells prior to effector injection is the tail fibre protein that is bound to the baseplate²⁹. This protein is encoded by the *Afp13* gene and has homologues across the different eCIS operons. The *Afp13* from *Serratia* spp. was shown to share amino acid sequence similarity with the tail fibre protein of adenovirus, an eukaryotic-targeting virus^{6,10}. We sought to classify the *Afp13* homologues in our dataset according to the fiber protein they are most similar to, whether from eukaryote-targeting viruses or from bacteria-targeting viruses (phages) to assign potential eCIS target cells. To do so we compiled a database of all proteins from

192,410 eukaryote-targeting viruses and 12,508 tailed phages and used protein BLAST with 629 Afp13 protein sequences as the query (Table S6, Materials and Methods). We identified Afp13 that have strong matches to eukaryote-targeting viruses, including fibre proteins from Equine Adenovirus 1, Bat mastadenovirus B, and Rhesus adenovirus 60. Interestingly, hits were also detected against the algae-infecting Organic Lake Phycodnavirus, as well as against amoeba-infecting Yasminevirus (Figure S4a,b), suggesting new eCIS eukaryotic targets beyond invertebrates. Our data showed that 99 Afp13 sequences matched eukaryotic-targeting viruses better than they matched phages. However, 16 Afp13 sequences aligned best with phage fibres (Figure S4c-f, Table S6). We therefore propose that in these cases, the eCIS particles likely target bacteria. This notion is supported by the fact that highly-related contractile tail particles, R-type pyocins, have analogous tail fibers that indeed specify host range by binding to specific strains of bacteria^{30,31}.

Notably, only 18% of the Afp13 protein sequences tested share a clear sequence similarity to a known tail fibre from phages or from eukaryotic-targeting viruses (but not to both groups). Therefore, in the majority of the cases, it is challenging to predict the kingdom to which the eCIS target organism belongs solely based on tail fibre protein sequence.

Characterization of the eCIS known and new core and shell protein domains

Hundreds of eCIS operons were previously defined^{1,16} but a systematic analysis of the genes that compose the eCIS operons has not been carried out yet. To comprehensively characterize the eCIS repertoire of genes, we compiled the list of Protein Family (pfam) domains across the 1,425 eCIS operons in our database (Materials and Methods). We then performed an enrichment test to identify protein domains that are statistically enriched in eCIS operons in comparison to all other genes in our 64,756 genome dataset (Figure 3a, Table S7). As expected, most of the known core eCIS protein domains are highly enriched and are present in nearly all eCIS loci (Figure 3a and b). These include the domains encoding eCIS tail tube, spike complex, sheath, baseplate, tail terminator protein, and ATP supply (Figure 3a). To identify yet unknown core domains we searched for protein domains that are present in eCIS from at least 10 microbial phyla. By doing so we identified a novel highly conserved eCIS core domain of unknown function 4157 (DUF4157). DUF4157 is found in 13 phyla and has high eCIS specificity (odd ratio = 2614, q value = 0). It contains the Zn-binding motif HExxH which characterizes many metalloproteases³²⁻³⁵. This domain, previously termed PVC-metalloprotease, was identified as a marker of eCIS based on limited genome analyses^{36,37} but its role has never been studied in a broad and functional context. By analyzing the eCIS operons and proteins we identified that DUF4157 is strongly associated with a high number of putative eCIS toxins, either as a gene that is adjacent to toxins or in the form of a multidomain protein, in which DUF4157 is found at the N-terminus and the toxin domain is found at the C-terminus (Figure S6). Interestingly, in many cases there is a very long linker sequence between

DUF4157 and the various toxin domains, occasionally more than 1000 amino acids long (Figure S6). The role of the DUF4157 domain in eCIS remains to be studied. We hypothesize that it is related to toxin loading, release, maturation, or trafficking inside the target cell.

Next, we defined ‘shell’ domains as those that are present in eCIS from 4-10 microbial phyla and likely produce specific eCIS subtypes or are responsible for regulation (Figure 3, Table S7). These shell domains include, for example, LysM (PF01476), Tail_P2_I (PF09684), and the bacterial transcriptional activator domain (BTAD, PF03704). An example of a novel eCIS shell domain is the GerE transcriptional regulator (PF00196) which is a LuxR-type DNA-binding domain. We find GerE located next to 25.8% of the eCIS loci, thereby defining it as the first conserved eCIS transcriptional regulator (Figure 3a and b). In 9% of the eCIS operons containing the GerE transcriptional regulator, a histidine kinase gene was found upstream to the GerE genes (Figure S8). These two genes might constitute a two-component system regulating the eCIS operon. Another novel shell domain is the Peptidase S74 (PF13884). The Peptidase S74 is a known domain of the intramolecular chaperone of the bacteriophage T5 tail fibre³⁸ (Figure S9). eCIS operons containing Peptidase S74 genes did not have annotated tail fibres (*Afp13*), suggesting they might have similar functions.

Figure 3. Protein Family (Pfam) domains that are enriched in eCIS. A. Volcano plot of pfam domain enrichment in eCIS operons. Fisher Exact test corrected with Benjamini Hochberg procedure. X axis is Log2 of odds ratio, Y axis is negative log10 of corrected P-value. The shape of each point is represented by the number of Phyla containing the Pfam in eCIS operon; >13 is defined “Core”, 4-10 is “Shell”, <4 is “Cloud”. The color of each point represents a functional context of the Pfam domain. The known eCIS core domains encode the tail tube (Phage_T4_gp19), spike complex (PAAR_motif, Phage_GPD), sheath (Phage_sheath_1), baseplate (GPW_gp25, Baseplate_J), tail terminator protein (DUF4255), and ATP supply (AAA). DUF4157 is a novel eCIS core domain., **B. Phylogenetic distribution of eCIS-associated pfams.** At the 12 o’clock and 6 o’clock positions of the tree are labels for select eCIS-associated pfams. Next to each pfam domain name at 12 o’clock is a heatmap quantifying how many phyla a given pfam is found in, along with a ring corresponding to each label (surrounding the circumference of the tree) with colors corresponding to A.

A large collection of toxin domains is encoded within eCIS operon genes

We identified numerous toxin domains located next to eCIS core components. Intriguingly, unlike the core and shell domains, most putative toxins are specific to certain eCIS loci (Figure 3b), suggesting rapid evolution and diversification of these toxins as part of the arms race against different target cells. Some of these domains are likely antibacterial toxins, such as: a DNase with an Ntox15 domain (PF15604)³⁹ and pesticin (PF16754)⁴⁰, which are encoded by the eCIS of *Lysobacter antibioticus* and *Pseudoalteromonas luteoviolacea*, respectively, supporting the idea that some eCIS are released to target bacteria. Other toxins, as expected, seem to target eukaryotes, e.g.: cycle inhibiting factor (PF16374) that arrest cell cycle⁴¹, and Von Willebrand factor type A domain, which is a glycoprotein found in blood plasma (PF00092). In addition, we predict that other eCIS domains serve as eCIS anti-eukaryotic toxin effectors as they are mostly found in eukaryotic proteins (Figure S7), such as hemopexin that is found in plasma proteins (PF00045)⁴², Annexin (PF00191), and the animal-specific prominin domain (PF05478)⁴³. Eukaryotic domains are common in effectors translocated into eukaryotic cells by various secretion systems⁴⁴⁻⁴⁷. Overall, we identified at least 71 protein domains that are likely toxins and are found next to eCIS core genes.

Experimental validation of 12 novel eCIS-associated toxins that target bacteria

Next, we decided to test our computational toxin gene predictions. We tested killing of *E. coli* as we predict that certain eCIS particles inject toxins into bacteria. To identify new toxins we first clustered all eCIS proteins based on sequence similarity and annotated core genes (Materials and Methods, Table S8). We selected 20 candidate eCIS toxins based on several features, such as small size, encoding domains (e.g. proteins with DUF4157), and location at the 3' end of their respective operon⁴⁻⁶. We synthesized and cloned genes of interest under an inducible T7 promoter in pET vectors and transformed these into *E. coli*. We used a known antibacterial toxin, Tse2⁴⁸, as a positive control, and two negative controls: a monomeric enhanced GFP gene⁴⁹ and an eCIS phage tail protein that is not expected to be toxic (Ga0070536_101452). Following serial dilutions of the colonies we could quantify the expressed protein toxicity levels. Out of the 20 genes tested, 12 were found to be toxic to *E. coli* (Figure 4a). We designate these as eCIS-associated toxins ("EATs") 1-12. Based on sensitive sequence similarity search we identified that the EATs likely encode various molecular functions, such as ADP ribosylation, peptidase, NAD⁺ phosphorylase, endonuclease, peptidoglycan hydrolysis, and deaminase (Figure 4b, Table 1). Most importantly, these are the first EATs demonstrating antibacterial activity. Seven EATs include DUF4157 domains as part of the protein or in a nearby gene (Table 1).

Figure 4. EATs are toxic to *E. coli*. EATs were cloned into pET28/29, transformed into *E. coli*, and were serially diluted and plated onto agar in induction conditions (right panel) or non-inducing conditions (left panel). tse2, a known Type VI Secretion System effector, was used as a positive control. Two negative controls used were meGFP, and an eCIS gene (phage tail). On the right side of each experiment is a cartoon representing the operon of the eCIS which the EAT was cloned from (purple gene with a red border); below is a legend explaining each component's color. Image shows representative results of one experiment out of 2 experiments with similar results. **B.** Schematic representation of EAT protein and the domains within them. Size in amino acids is indicated in parentheses.

Table 1. New eCIS-associated toxins (EATs) experimentally validated in the current study.

Toxin Name	Gene ID (IMG gene id)	Organism (taxon)	Activity	Predicted Molecular Function	DUF4147 gene/protein association
EAT1	Ga0102068_1 13144 (2644141250)	<i>Rhizobacter</i> sp. Root29 (Betaproteobacteria)	Antibact., anti-euk.	ADP-ribosyl transferase (SpvB) and adhesion (VCBS)	Absent
EAT2	Ga0070133_0 156 (2616189889)	<i>Flavobacterium</i> sp. FV08 (Bacteroidetes)	Antibact.	Unknown. Encodes ankyrin repeats.	One and two genes upstream to EAT2
EAT3	Dd1591_0669 (644851057)	<i>Dickeya zeae</i> Ech1591 (Gammaproteo.)	Antibact., anti-euk.	Chitinase?	Absent
EAT4	Ga0066759_1 0824 (2609591113)	<i>Mucilaginibacter</i> sp. OK098 (Bacteroidetes)	Antibact.	peptidyl-Lys metalloendo peptidase	Within protein center
EAT5	EX18DRAFT_0 3240 (2587734256)	<i>Methylibium</i> sp. CF468 (Betaproteobacteria)	Antibact.	NAD+ Phosphorylas e	Two genes upstream to EAT5.
EAT6	Ga0131261_1 13100 (2695001213)	<i>Parafilimonas terrae</i> DSM 28286 (Bacteroidetes)	Antibact.	L,D-transpep tidase	Within protein
EAT7	Ga0059271_0 2137 (2602024578)	<i>Janthinobacterium</i> sp. 551a (Betaproteobacteria)	Antibact.	Unknown	Within protein (N terminus)
EAT8	Ga0079878_1 06272 (2668039404)	<i>Paenibacillus jilunlii</i> CGMCC 1.10239 (Firmicutes)	Antibact.	Unknown	Absent

EAT9	Ga0081190_1 15160 (2631153083)	<i>Moorea producens</i> PAL 15AUG08-1 (Cyanobacteria)	Antibact.	Restriction endonucleas e	Three genes upstream to EAT9.
EAT10	Ga0176038_1 12447 (2720476818)	<i>Pseudoalteromonas</i> <i>luteoviolacea</i> S4054249 (Gammaproteo.)	Antibact.	Peptidoglyca n hydrolysis	Absent
EAT11	AmyniDRAFT_ 7433 (2515141031)	<i>Amycolatopsis</i> <i>nigrescens</i> CSC17Ta-90, DSM 44992 (Actinobacteria)	Antibact., anti-euk.	Deaminase	Within protein (N terminus)
EAT12	Ccel_1518 (643608442)	<i>Ruminiclostridium</i> <i>cellulolyticum</i> H10 (Firmicutes)	Antibact.	Nucleoporin?	Absent
EAT13	Ga0111076_1 23119 (2653242303)	<i>Candidatus</i> <i>Udaeobacter</i> <i>copiosus</i> (Verrucomicrobia)	Anti-euk.	Actin cross-linking	Flanked by genes with DUF4157 domain

Antibact. - toxins killing the bacterium *E. coli*. Anti-euk. - toxins killing the yeast *S. cerevisiae*.

Several eCIS-associated toxins are genetically linked to antitoxins

We identified several cases of toxin-immunity gene pairs next to eCIS operons supporting the antibacterial role of these eCIS loci as they require immunity genes (antitoxins) to protect from self-intoxication (Figure 5A). In one case, The toxin, EAT5, resembles RES (Figure 4b), an NADase that is accompanied by an antitoxin called Xre^{50,51}. We identified a gene, EX18DRAFT_03239, encoding an unknown protein located upstream of the toxin. We predicted that it serves as a new cognate immunity gene for the RES family, as the gene pair organization is conserved in phylogenetically diverse bacterial genomes (Figure S5). Expression of the RES-like EAT5 toxin was indeed toxic to *E. coli* while the putative immunity gene was not (Figure 5B). Co-expression of the toxin-immunity pair was sufficient to rescue bacteria from toxicity (Figure 5B). Finally, we showed that an intact immunity protein is required for the rescue phenotype, since deletion of 17 amino acids which encompass the C terminal alpha helix domain from the antitoxin abolished the immunity function (Figure 5B). Interestingly, the tail fibre protein (encoded by gene EX18DRAFT_03233) of the associated eCIS resembles a protein from Thermus phage YS40_Isch (see gene 2587734249 at Table S6), providing additional evidence that this eCIS targets bacteria and is therefore accompanied by an immunity protein.

Putting the data together, we present here the first 12 antibacterial EATs and the first EAT-immunity gene pairs.

Figure 5. Toxin-immunity pairs associated with eCIS **A.** Pairs of known antibacterial toxins found in eCIS. eCIS operons containing known or putative antibacterial toxic or immunity domains. **B. XRE-like Antitoxin Protects From RES-like Toxin.** Induction by IPTG leads to expression of RES-like toxin (IMG locus tag EX18DRAFT_03240). Glucose represses XRE-like antitoxin (IMG locus tag EX18DRAFT_03239) or its truncated version while Arabinose induces expression of XRE-like antitoxin (or its truncated version). Each section (separated by blue lines) represents three biological repeats.

Experimental validation of novel four eCIS toxins that target eukaryotic cells

Following our extensive bioinformatic analysis we were also interested in discovery of novel eCIS anti-eukaryotic toxins. We predicted new EATs that target eukaryotes based on their presence in the 3' end of the eCIS operon, presence of eukaryotic domains within the encoded proteins, predicted enzymatic activity on eukaryote-specific molecules (such as actin), and similarity to known virulence factors. We cloned 14 genes into pESC-leu galactose-inducible vectors and transformed them into *S. cerevisiae* cells, serving as models for eukaryotic cells (Materials and Methods). Upon expression induction we identified four EATs that efficiently killed yeast (Figures 6, S9). Three EATs, EAT1, EAT3, and EAT11, also killed *E. coli*. EAT13 gene was tested only in yeast. These are predicted to act as an ADP-ribosyl transferase, a chitinase, a deaminase, and an actin cross-linker (Table 1). We think that yeast is probably not the native target for many other anti-eukaryotic EATs and therefore, it is likely that some of the toxins we predicted may be active when expressed in different organisms found in the native habitat of eCIS-encoding microbes.

Figure 6. EATs are toxic to *S. cerevisiae*. EATs were cloned into pESC -leu Galactose inducible plasmids that were then transformed into *Saccharomyces Cerevisiae* BY4742 strain. Overnight cultures of the strains harboring the vectors of interest were grown in SD -leu media. The cultures OD was normalized and then washed once with water and split into two: one part was grown overnight in repressive conditions (SD -leu + 2% glucose) and the other part was grown in inductive conditions (SD -leu + 2% galactose). Dilutions were spotted on SD -leu plates containing glucose or galactose and the plates were incubated two nights at 30C. Negative controls: empty vector (ev) and non-toxin (yeGFP gene). Three biological replicates of this experiment are presented in Figure S10.

Development of a comprehensive eCIS database

Finally, we developed a new repository called eCISstem that contains 1,425 operons from 1249 bacterial and archaeal genomes. The repository is publicly available at: <http://ecistem.pythonanywhere.com>. eCISstem provides visualization of all eCIS operons and includes gene annotation, an eCIS operon search option based on taxonomy, pfam domains, and gene clusters, and links to additional gene and protein data from the IMG database⁵². eCISstem also contains ecological and life-style metadata of the eCIS encoding microbe. This valuable information can help researchers to study eCIS with regard to its specific ecological functions and to deduce the eCIS target organism. Figure 7 presents some of the information held in eCISstem. eCISstem is a significant expansion of the recently developed dbEIS which includes 631 operons¹⁶. We also expanded dbEIS functionality by providing protein domain data for all eCIS genes, encoding organism metadata, and a full database download option which will facilitate future bioinformatic analysis of our dataset by the entire

research

community.

Figure 7. eCIS database provides information on 1425 eCIS operons and their encoding genomes. Blue arrows represent links from page to page. A. A view of eCIS gene operons. B. An operon entry with information of each gene in the operon with operon, gene, protein, and protein domain information. C. Examples of eCIS protein cluster (clustered by 40% amino acids sequence identity). D. Examples of pfam domains entry for a specific gene. The information held in eCIS is freely available to users from academic non-profit institutions.

Discussion

In this study we use bioinformatic, phylogenetic, and molecular methods in order to thoroughly characterize the eCIS system across the microbial world. Based on our results we provide an improved model for eCIS function that enriches our understanding of eCIS ecological and taxonomic distribution within microbes, potential eCIS regulation and cell attachment mechanisms. Moreover, we expanded the known repertoire of

eCIS targets. We show that, most likely, eCISs not only target eukaryotic cells, as previously reported, but also bacteria (Figure 8). Our extensive analysis uncovered new conserved eCIS core and shell components that encode the eCIS structure and regulation and numerous EATs genetically associated with eCIS loci. The high number and variety of EATs suggest that eCIS loci can serve as a treasure trove for toxin gene discovery. Previous works uncovered seven EATs, all from Gammaproteobacteria class, whereas here we added 13 novel EATs from six different bacterial phyla. We examine the occurrence of eCIS operons within a large-scale genome set and provide a strong support for eCIS enrichment in environmental microbes from soil and aquatic ecosystems and hosts, and a depletion from mammalian and avian microbiomes; with a specific depletion from their pathogens.

Figure 8. A revised model for eCIS function and ecology. eCIS is encoded by operons that are enriched in environmental microbes. These operons are regulated by adjacent genes such as RfaH⁷⁷ and genes that carry GerE and BTAD domains. The regulators are occasionally activated by kinases and are activated/repressed by

yet unknown signals. Effectors (EATs) tend to be encoded in the 3' end of the operons. EATs often carry DUF4157 domains in their N-terminus or found adjacent to genes with this domain. The eCIS particle is assembled with the toxins inside the tube ¹⁴ and is released following cell lysis ¹². The particles carry tail fibres that mediate attachment to environmental eukaryotic or prokaryotic target cells. The tail fibres attach, contract, and release the EATs inside the target cells, mostly leading to its death or growth arrest. The EATs function intracellularly in diverse ways, enzymatically acting on different cellular components that were identified in this work and by other groups.

We identified that some eCIS tail fibres that can help predict whether eCIS target bacteria or eukaryotes. Interestingly we identified eCIS tail fibres that share high sequence similarity with similar proteins of *Yasminevirus* giant virus and *Phycodnavirus*, which infect amoebae ⁵³ and algae ⁵⁴, respectively (Figure S4). Based on this finding we propose that eCIS from these bacteria could target similar hosts. The fact that most eCIS tail fibers matched both phage and eukaryotic-targeting viruses suggests they may have broad spectrum binding activity, and therefore broad targeting activity.

One intriguing mystery is related to the antibacterial EATs and their lack of immunity genes in most cases. Nearly all antibacterial toxins produced by bacteria are accompanied by adjacent immunity genes ^{36,39,55–61}, which we could not bioinformatically identify next to most of the newly verified EAT genes. We do not know how these antibacterial EATs are being produced and loaded onto the eCIS system without killing the producer cell prior to eCIS release. One possibility is that these are produced right before the cell is lysed to release the eCIS, and therefore the cell does not require protection at this time, as it is already “fated” to die, and cellular activity shifts to releasing eCIS particles. It is possible that EATs even purposely contribute to lysis. Another option is related to the yet mysterious function of the DUF4157 domain which accompanies many of the EATs. This domain may confer some immunity to the eCIS producers through temporal toxin inactivation.

Shedding light on eCIS biological role in nature can reveal new means by which microbes affect the health and development of their hosts and shape neighboring eukaryotic and prokaryotic microbial communities.

Materials and Methods

Identification of eCIS operons within microbial genomes

The starting database for identification of eCIS was 64,756 publicly available genomes from the Integrated Microbial Genomes database ⁵², which includes all protein-coding genes along with their pfam annotations, as per the IMG pipeline ⁶². Genomes were searched for the presence of genes annotated as containing pfam

domains⁶³ associated with eCIS (Pfam IDs: 00004, 04865, 04965, 04984, 05488, 05954, 06841, 14065). Genes encoding proteins with the aforementioned pfam annotations were grouped by physical linkage (encoded within 12,000 base pairs of one another). Only those groups of linked genes with at least four members, with at least three of them having different pfam annotations, were considered putative eCIS operons. Because of the evolutionary similarity between eCIS and other contractile machinery, we removed any putative operons with genes annotated with phage-related pfams (Pfam IDs: 04860, 03237, 17289, 05929, 05125, 05944, 05926, 04550, 05840, 04606, 05065, 07068, 16855, 06810, 03864, 02924, 10124), such as capsid genes, within 10,000 base pairs of any genes in our putative operons. Furthermore, we removed any putative operons with genes annotated as COGs⁶⁴ associated with Type VI Secretion Systems (COG IDs: 0542, 3521, 3523) within 10,000 base pairs of any genes in our putative operons. We believe this is a very stringent filtration protocol that was also added to filtration of pyocins (see below). The boundaries of these putative operons were expanded by 10 genes up- and downstream of the extrema of the putative operons to identify eCIS accessory (non-core) genes. Only protein coding sequences were kept, and tRNA and other genes were ignored. Using previously published HMM profiles of eCIS core elements¹⁶, and the hmmsearch tool (a part of the HMMER package;⁶⁵), the expanded putative operons were surveyed and their genes were annotated as core elements. Only the top half of the hmmsearch results were used (according to bitscore), in order to minimize false positive hits. Furthermore, genes annotated by IMG as having pfams that match core elements were also annotated as core genes. The core-annotated putative operons were filtered for those with at least 10 core genes, and with at least one eCIS-specific core (AFP 12, 13, 14, or 16;¹⁶), in order to remove any possible R-type pyocins from the dataset. The final database of eCIS operons contains 1,249 genomes, with 1,425 distinct operons.

Phylogenetic eCIS gene tree construction

Amino acid sequences for Afp11 and Afp8 homologs were retrieved from each eCIS operon. In the case of operons with more than one copy of either the Afp8 or Afp11 homologs, the genes with the larger sequence were chosen. In the minority of cases (15/1425) where one or both of the genes were missing, they were not included in the tree. The Afp11 and Afp8 sequences were separately aligned with clustal omega⁶⁶ using standard settings. The alignments were trimmed using TrimAl⁶⁷ with the “automated1” flag selected. The Afp11 and Afp8 alignments from corresponding operons were concatenated. The concatenated alignments were inputted into FastTree using standard settings⁶⁸, and visualized and annotated using interactive Tree Of Life (iTOL;⁶⁹).

Systematic tail fibre (Afp13) sequence analysis

629 Afp13 amino acid sequences, as defined by genes labeled by hmmsearch as afp13 within the top 25th percentile of bitscores, were inputted as queries for two searches using BLASTP against the NCBI nr database using standard parameters⁷⁰. In the first search, the nr database was filtered by taxon IDs corresponding to tailed bacteriophages (NCBI taxid: 28883). In the second search, the nr database was filtered by taxon IDs corresponding to eukaryotic-targeting viruses, by manually removing prokaryotic viruses from taxid 10239 (see Table S6 for full taxid list). The results from each search were filtered for those hits with p-value < 0.001, leaving 296 unique Afp13 hits. Each query Afp13 gene was checked for the hit with the top bitscore from the first and second searches to determine if a given gene better aligned with tailed phages, or with eukaryotic-targeting viruses (Table S6). Those Afp13 with bitscore difference greater than 15 were marked as having a preference for one dataset over another; within 15 was considered inconclusive.

Detection of eCIS within plasmids

Together with our collaborator Dr. William Anreopolous (San Jose State University) we developed Deeplasmid (<https://sourceforge.net/p/deeplasmid/>). We ran Deeplasmid on all scaffolds carrying eCIS. Deeplasmid uses a deep learning algorithm that classifies scaffolds as being chromosomes or plasmids based on different features, such as GC content, characteristic genes, and coding density. The algorithm achieves an AUC-ROC of over 93% based on our tests. The score cutoff of ≥ 0.7 was chosen as it correctly identifies the pADAP plasmid that encodes the Antifeeding Prophage (AFP). Other known plasmids carrying eCIS received Deeplasmid scores higher than 0.95 (Table S3).

eCIS association with genera

The taxonomic makeup of the eCIS database was compared to the number of genera in publicly available genomes of Bacteria, Archaea, and plasmids of bacteria from IMG using a custom python script. A Fisher exact test was calculated for each genus, which returned an odds ratio and an associated p-value. Multiple hypothesis testing correction was performed using the Benjamini Hochberg procedure. Those genera with a q-value of less than 0.001 and at least 10 entries in the publicly available IMG database are displayed in Figure 1.

eCIS association with ecological and physiological features

To correct for taxonomic bias that affects enrichment analysis, a population-aware enrichment analysis based on Scoary²⁷ was used. Scoary accepts input of a tree to analyze Fisher exact test enrichments in the context of phylogeny. To create the tree, 16s sequences for each genome with eCIS was queried. In the case of more than one 16s gene per genome, one was chosen randomly. These sequences were inputted into clustal omega⁶⁶ to generate a UPGMA guide tree. The inputs to Scoary were (1) the guide tree, (2) a presence-absence file,

indicating if a genome has or lacks an eCIS operon, (3) metadata files that indicate for each genome if they possess a certain characteristic, e.g. whether it was isolated from soil. This program outputs an odds ratio, a q-value for the Fisher exact test that produced the odds ratio, and a q-value based on how many contrasting pairs (instances not from the same clade) support the enrichment, i.e. how many examples of the enrichment are independent of taxonomy.

Protein Domain (Pfam) Enrichment Analysis

We collected the pfam annotations for the 64,756 proteins encoded by publicly available genomes from IMG. Based on our eCIS operon database, we marked all the pfam domains that were found in our eCIS database as “pfam domains within eCIS operons”. Duplicated domains within a single gene (usually of repeat domains) were dropped, in order to avoid inflation of enrichment. We then counted all the occurrences for each pfam - within eCIS operons, outside eCIS operons, and in total. After counting all pfam occurrences we performed a Fisher Exact Test for pfam enrichment in eCIS operons, compared to the rest of genomes in the analysis. Multiple hypothesis testing correction was performed using the Benjamini Hochberg procedure. The adjusted P value (q value) of the Core Component was zero, so in order to plot it we changed it to $1e^{-250}$. Then we plotted the data using R ⁷¹. We counted the number of Phyla each Pfam domain appeared in eCIS. Pfam domains appearing in >10 Phyla were marked as “Core Domains”, domains found in 4-10 Phyla were marked as “Shell Domains”, and domains found in <4 were marked as “Cloud Domains”. These terms were borrowed from pangenomics.

Toxins and accessory gene clustering

To identify accessory genes, i.e. genes that are not annotated as homologs of Afp proteins, the operons were expanded from their extrema by four protein coding genes up- and downstream. All genes lacking an annotation of an Afp homolog were clustered. This included the genes up and downstream of extrema, as well as genes in between annotated Afp homologs. Using CD-HIT ⁷², the amino acid sequences of these unmarked genes were clustered with a threshold of at least 40% identity over at least 80% of the length of both query and subject. Each gene was inputted into CD-Search ⁷³ with standard settings to determine if the gene encoded for conserved domains. Each 40%-identity-protein-cluster was then manually examined for domains associated with known biological function, e.g. known toxin domains, known transcription factor domains, etc.

Heterologous expression of putative EATs in bacteria

Candidate gene sequences were retrieved from IMG ⁵² and were synthesized, following codon adaptation, and cloned into either pET28, pET29, or pBAD24 plasmids by Twist Bioscience. The plasmids were then transformed into *E. coli* BL21 (DE3) pLysS strain. Overnight cultures of the strains harboring the vectors of

interest were grown in LB containing the proper selection: Kan or Amp for pET28/29 or pBAD24, respectively. The cultures were normalized by OD to be equivalent and subsequently serially diluted. Dilutions were spotted on LB agar containing the proper selection and inducer (500 μ M IPTG, 0.1% arabinose) or repressor (1% glucose) and the plates were incubated overnight at 37C.

The toxin-antitoxin assay was done as followed: E. coli BL21 (DE3) pLysS strain was transformed with the toxin (EAT 5) on a pET28 plasmid and its predicted antitoxin (IMG gene id 2587734255) on a pBAD24. Two control strains were made, an empty vector (pBAD24) and a truncated antitoxin. We removed the last 51 nucleotides (17 amino acids) from the gene (which was synthesized by TWIST bioscience) using inverse PCR, then it was re-ligated with NEbuilder (NEB) kit. All three strains were tested in a drop assay with three biological replicates. Induction of toxin and antitoxin was performed with 100 μ M IPTG and 0.1% arabinose respectively.

Heterologous expression of putative toxins in yeast

Candidate sequences were retrieved from IMG ⁵². They were synthesized, following codon adaptation to yeast by Twist Bioscience and were cloned into pESC -leu Galactose inducible plasmids. The plasmids were then transformed into *Saccharomyces Cerevisiae* BY4742 strain ⁷⁴. Overnight cultures of the strains harboring the vectors of interest were grown in SD -leu media. The cultures OD was normalized and then washed once with water and split into two: one part was grown overnight in SD-leu 2% glucose media and the other part was grown in SD-leu 2% galactose media. Dilutions were spotted on SD-leu plates containing glucose or galactose and the plates were incubated for 48 hours at 30C.

Construction of eCIStem repository

The ecis database was constructed based on the eCIS data from our research (Tables S1, S2, S7, S8). The information is divided into four tables: Operons Table, Gene Table, Organism Table and Pfam Table. The eCIStem database scheme is described in Figure S11. The Django framework was used to represent the data and its relationships and deploy it for public usage.

S74 Peptidase (PF13884) secondary structure analysis

We clustered all 84 genes containing PF13884, using CD-HIT with 40% similarity ⁷². Used two cluster representatives for structural prediction in Phyre2 ⁷⁵, and secondary structure alignment using Praline ⁷⁶.

distribution among both prokaryotic domains: a comparative genomic and phylogenetic study. *Genome Biol. Evol.* **6**, 1739–1747 (2014).

2. Desfosses, A. *et al.* Atomic structures of an entire contractile injection system in both the extended and contracted states. *Nat Microbiol* **4**, 1885–1894 (2019).
3. Jiang, F. *et al.* Cryo-EM Structure and Assembly of an Extracellular Contractile Injection System. *Cell* **177**, 370–383.e15 (2019).
4. Vlisidou, I. *et al.* The *Photorhabdus asymbiotica* virulence cassettes deliver protein effectors directly into target eukaryotic cells. *Elife* **8**, (2019).
5. Yang, G., Dowling, A. J., Gerike, U., French-Constant, R. H. & Waterfield, N. R. *Photorhabdus* virulence cassettes confer injectable insecticidal activity against the wax moth. *J. Bacteriol.* **188**, 2254–2261 (2006).
6. Hurst, M. R. H. *et al.* *Serratia proteamaculans* Strain AGR96X Encodes an Antifeeding Prophage (Tailocin) with Activity against Grass Grub (*Costelytra giveni*) and Manuka Beetle (*Pyronota* Species) Larvae. *Appl. Environ. Microbiol.* **84**, (2018).
7. Rybakova, D., Schramm, P., Mitra, A. K. & Hurst, M. R. H. Afp14 is involved in regulating the length of Anti-feeding prophage (Afp). *Mol. Microbiol.* **96**, 815–826 (2015).
8. Heymann, J. B. *et al.* Three-dimensional structure of the toxin-delivery particle antifeeding prophage of *Serratia entomophila*. *J. Biol. Chem.* **288**, 25276–25284 (2013).
9. Sen, A., Rybakova, D., Hurst, M. R. H. & Mitra, A. K. Structural study of the *Serratia entomophila* antifeeding prophage: three-dimensional structure of the helical sheath. *J. Bacteriol.* **192**, 4522–4525 (2010).
10. Hurst, M. R. H., Glare, T. R. & Jackson, T. A. Cloning *Serratia entomophila* antifeeding genes—a putative defective prophage active against the grass grub *Costelytra zealandica*. *J. Bacteriol.* **186**, 5116–5128 (2004).
11. Wang, X., Shen, J., Jiang, F. & Jin, Q. The *Photorhabdus* Virulence Cassettes RRSP-Like Effector Interacts With Cyclin-Dependent Kinase 1 and Causes Mitotic Defects in Mammalian Cells. *Front. Microbiol.* **11**, 366

(2020).

12. Shikuma, N. J. *et al.* Marine tubeworm metamorphosis induced by arrays of bacterial phage tail-like structures. *Science* **343**, 529–533 (2014).
13. Rocchi, I. *et al.* A Bacterial Phage Tail-like Structure Kills Eukaryotic Cells by Injecting a Nuclease Effector. *Cell Rep.* **28**, 295–301.e4 (2019).
14. Ericson, C. F. *et al.* A contractile injection system stimulates tubeworm metamorphosis by translocating a proteinaceous effector. *Elife* **8**, (2019).
15. Markowitz, V. M. *et al.* IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res.* **40**, D115–22 (2012).
16. Chen, L. *et al.* Genome-wide Identification and Characterization of a Superfamily of Bacterial Extracellular Contractile Injection Systems. *Cell Rep.* **29**, 511–521.e2 (2019).
17. Williams, K. P. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res.* **30**, 866–875 (2002).
18. Forst, S., Dowds, B., Boemare, N. & Stackebrandt, E. Xenorhabdus and Photorhabdus spp.: bugs that kill bugs. *Annu. Rev. Microbiol.* **51**, 47–72 (1997).
19. Shaffer, J. P., U'Ren, J. M., Gallery, R. E., Baltrus, D. A. & Arnold, A. E. An Endohyphal Bacterium (Chitinophaga, Bacteroidetes) Alters Carbon Source Use by Fusarium keratoplasticum (F. solani Species Complex, Nectriaceae). *Front. Microbiol.* **8**, 350 (2017).
20. Pritchard, L. *et al.* Draft Genome Sequences of Four Dickeya dianthicola and Four Dickeya solani Strains. *Genome Announc.* **1**, (2013).
21. Grenier, A.-M., Duport, G., Pagès, S., Condemine, G. & Rahbé, Y. The phytopathogen Dickeya dadantii (Erwinia chrysanthemi 3937) is a pathogen of the pea aphid. *Appl. Environ. Microbiol.* **72**, 1956–1965 (2006).
22. Van Trappen, S., Vandecandelaere, I., Mergaert, J. & Swings, J. Algoriphagus antarcticus sp. nov., a novel psychrophile from microbial mats in Antarctic lakes. *Int. J. Syst. Evol. Microbiol.* **54**, 1969–1973 (2004).

23. Yoon, J.-H., Kang, S.-J., Jung, S.-Y., Lee, C.-H. & Oh, T.-K. *Algoriphagus yeomjeoni* sp. nov., isolated from a marine solar saltern in the Yellow Sea, Korea. *Int. J. Syst. Evol. Microbiol.* **55**, 865–870 (2005).
24. Yoon, J.-H., Lee, M.-H., Kang, S.-J. & Oh, T.-K. *Algoriphagus terrigena* sp. nov., isolated from soil. *Int. J. Syst. Evol. Microbiol.* **56**, 777–780 (2006).
25. Young, C.-C. *et al.* *Algoriphagus olei* sp. nov., isolated from oil-contaminated soil. *Int. J. Syst. Evol. Microbiol.* **59**, 2909–2915 (2009).
26. Li, Y. *et al.* *Algoriphagus faecimaris* sp. nov., isolated from coastal sediment. *Int. J. Syst. Evol. Microbiol.* **61**, 2856–2860 (2011).
27. Brynildsrud, O., Bohlin, J., Scheffer, L. & Eldholm, V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* **17**, 238 (2016).
28. Rojas, M. I. *et al.* A Distinct Contractile Injection System Found in a Majority of Adult Human Microbiomes. 865204 (2019) doi:10.1101/865204.
29. Patz, S. *et al.* Phage tail-like particles are versatile bacterial nanomachines - A mini-review. *J. Advert. Res.* **19**, 75–84 (2019).
30. Köhler, T., Donner, V. & van Delden, C. Lipopolysaccharide as shield and receptor for R-pyocin-mediated killing in *Pseudomonas aeruginosa*. *J. Bacteriol.* **192**, 1921–1928 (2010).
31. Buth, S. A., Shneider, M. M., Scholl, D. & Leiman, P. G. Structure and Analysis of R1 and R2 Pyocin Receptor-Binding Fibers. *Viruses* **10**, (2018).
32. Menach, E., Hashida, Y., Yasukawa, K. & Inouye, K. Effects of conversion of the zinc-binding motif sequence of thermolysin, HEXXH, to that of dipeptidyl peptidase III, HEXXXH, on the activity and stability of thermolysin. *Biosci. Biotechnol. Biochem.* **77**, 1901–1906 (2013).
33. Franco, A. A., Buckwold, S. L., Shin, J. W., Ascon, M. & Sears, C. L. Mutation of the zinc-binding metalloprotease motif affects *Bacteroides fragilis* toxin activity but does not affect propeptide processing. *Infect. Immun.* **73**, 5273–5277 (2005).
34. Grasso, G. *et al.* The copper(II) and zinc(II) coordination mode of HExxH and HxxEH motif in small

- peptides: the role of carboxylate location and hydrogen bonding network. *J. Inorg. Biochem.* **130**, 92–102 (2014).
35. Lenart, A., Dudkiewicz, M., Grynberg, M. & Pawłowski, K. CLCAs - a family of metalloproteases of intriguing phylogenetic distribution and with cases of substituted catalytic sites. *PLoS One* **8**, e62272 (2013).
 36. Zhang, D., de Souza, R. F., Anantharaman, V., Iyer, L. M. & Aravind, L. Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. *Biol. Direct* **7**, 18 (2012).
 37. Makarova, K. S. *et al.* Antimicrobial Peptides, Polymorphic Toxins, and Self-Nonself Recognition Systems in Archaea: an Untapped Armory for Intermicrobial Conflicts. *MBio* **10**, (2019).
 38. Garcia-Doval, C. *et al.* Structure of the Receptor-Binding Carboxy-Terminal Domain of the Bacteriophage T5 L-Shaped Tail Fibre with and without Its Intra-Molecular Chaperone. *Viruses* **7**, 6424–6440 (2015).
 39. Ma, L.-S., Hachani, A., Lin, J.-S., Filloux, A. & Lai, E.-M. *Agrobacterium tumefaciens* deploys a superfamily of type VI secretion DNase effectors as weapons for interbacterial competition in planta. *Cell Host Microbe* **16**, 94–104 (2014).
 40. Patzer, S. I., Albrecht, R., Braun, V. & Zeth, K. Structural and mechanistic studies of pesticin, a bacterial homolog of phage lysozymes. *J. Biol. Chem.* **287**, 23381–23396 (2012).
 41. Taieb, F., Nougayrède, J.-P. & Oswald, E. Cycle inhibiting factors (cifs): cyclomodulins that usurp the ubiquitin-dependent degradation pathway of host cells. *Toxins* **3**, 356–368 (2011).
 42. Tolosano, E. & Altruda, F. Hemopexin: structure, function, and regulation. *DNA Cell Biol.* **21**, 297–306 (2002).
 43. Corbeil, D., Röper, K., Fargeas, C. A., Joester, A. & Huttner, W. B. Prominin: a story of cholesterol, plasma membrane protrusions and human pathology. *Traffic* **2**, 82–91 (2001).
 44. Burstein, D. *et al.* Genome-scale identification of *Legionella pneumophila* effectors using a machine learning approach. *PLoS Pathog.* **5**, e1000508 (2009).

45. Price, C. T. D., Al-Khodori, S., Al-Quadan, T. & Abu Kwaik, Y. Indispensable role for the eukaryotic-like ankyrin domains of the ankyrin B effector of *Legionella pneumophila* within macrophages and amoebae. *Infect. Immun.* **78**, 2079–2088 (2010).
46. Gomez-Valero, L. *et al.* More than 18,000 effectors in the *Legionella* genus genome provide multiple, independent combinations for replication in human cells. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 2265–2273 (2019).
47. Jehl, M.-A., Arnold, R. & Rattei, T. Effective--a database of predicted secreted bacterial proteins. *Nucleic Acids Res.* **39**, D591–5 (2011).
48. Hood, R. D. *et al.* A type VI secretion system of *Pseudomonas aeruginosa* targets a toxin to bacteria. *Cell Host Microbe* **7**, 25–37 (2010).
49. Cormack, B. P., Valdivia, R. H. & Falkow, S. FACS-optimized mutants of the green fluorescent protein (GFP). *Gene* **173**, 33–38 (1996).
50. Skjærning, R. B., Senissar, M., Winther, K. S., Gerdes, K. & Brodersen, D. E. The RES domain toxins of RES-Xre toxin-antitoxin modules induce cell stasis by degrading NAD⁺. *Mol. Microbiol.* **111**, 221–236 (2019).
51. Freire, D. M. *et al.* An NAD⁺ Phosphorylase Toxin Triggers *Mycobacterium tuberculosis* Cell Death. *Mol. Cell* **73**, 1282–1291.e8 (2019).
52. Chen, I.-M. A. *et al.* IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–D677 (2019).
53. Bajrai, L. H. *et al.* Isolation of Yasminevirus, the First Member of Klosneuvirinae Isolated in Coculture with *Vermamoeba vermiformis*, Demonstrates an Extended Arsenal of Translational Apparatus Components. *J. Virol.* **94**, (2019).
54. Van Etten, J. L., Graves, M. V., Müller, D. G., Boland, W. & Delaroque, N. Phycodnaviridae--large DNA algal viruses. *Arch. Virol.* **147**, 1479–1516 (2002).
55. 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).
56. Jana, B., Fridman, C. M., Bosis, E. & Salomon, D. A modular effector with a DNase domain and a marker for T6SS substrates. *Nat. Commun.* **10**, 3595 (2019).
 57. Aoki, S. K. *et al.* A widespread family of polymorphic contact-dependent toxin delivery systems in bacteria. *Nature* **468**, 439–442 (2010).
 58. Anderson, M. S., Garcia, E. C. & Cotter, P. A. Kind discrimination and competitive exclusion mediated by contact-dependent growth inhibition systems shape biofilm community structure. *PLoS Pathog.* **10**, e1004076 (2014).
 59. Ahmad, S. *et al.* An interbacterial toxin inhibits target cell growth by synthesizing (p)ppApp. *Nature* **575**, 674–678 (2019).
 60. Verster, A. J. *et al.* The Landscape of Type VI Secretion across Human Gut Microbiomes Reveals Its Role in Community Composition. *Cell Host Microbe* **22**, 411–419.e4 (2017).
 61. Engelberg-Kulka, H., Amitai, S., Kolodkin-Gal, I. & Hazan, R. Bacterial programmed cell death and multicellular behavior in bacteria. *PLoS Genet.* **2**, e135 (2006).
 62. Huntemann, M. *et al.* The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4). *Stand. Genomic Sci.* **10**, 86 (2015).
 63. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–30 (2014).
 64. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, D261–9 (2015).
 65. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
 66. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **27**, 135–145 (2018).
 67. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

68. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
69. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
70. 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).
71. Blighe, K. *EnhancedVolcano*. (Bioconductor, 2018). doi:10.18129/B9.BIOC.ENHANCEDVOLCANO.
72. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
73. Marchler-Bauer, A. & Bryant, S. H. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* **32**, W327–31 (2004).
74. Jungnickel, H. *et al.* Application of TOF-SIMS with Chemometrics To Discriminate between Four Different Yeast Strains from the Species *Candida glabrata* and *Saccharomyces cerevisiae*. *Anal. Chem.* **77**, 1740–1745 (2005).
75. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
76. Simossis, V. A. & Heringa, J. PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res.* **33**, W289–94 (2005).
77. Hurst, M. R. H., Beard, S. S., Jackson, T. A. & Jones, S. M. Isolation and characterization of the *Serratia entomophila* antifeeding prophage. *FEMS Microbiol. Lett.* **270**, 42–48 (2007).
78. Bawono, P. & Heringa, J. PRALINE: a versatile multiple sequence alignment toolkit. *Methods Mol. Biol.* **1079**, 245–262 (2014).

Acknowledgments

AL is generously supported by the Israeli Science Foundation (Grant #1535/20) and Alon Fellowship of the Israeli council of higher education. AMG is generously supported by the Kaete Klausner Scholarship and a scholarship from the Israeli Ministry of Aliyah and Integration. We thank Prof. Maya Schuldiner, Prof. Yechiel Shay, and Prof. Saul Burdman for providing different plasmids and cells. We thank Dr. Omri Finkel, Dr. Hila Sberro, Prof. Saul Burdman, Prof. Rotem Sorek, Dr. Erez Mills, Dr. Gal Ofir, and Dr. Lianet Noda-Garcia for careful evaluation of the manuscript.

Supplementary Figures

Figure S1. eCIS structure. eCIS particles are encoded by an operon of 15-28 genes, a typical operon is shown, with cartoons showing corresponding structural components (right side). The core proteins encode a sheath that has a cap on one side and a baseplate on the other side. Inside the sheath there is a tube that ends with a sharp spike. The particle likely binds to receptors located on the target cell through tail fibres. Once the contact is established, the eCIS contracts, injects the tube, and eCIS-associated toxins (EATs) are released into the target cell (left side). Some of the toxins were reported to be located inside the hollow tube. The illustration is based on Jiang et al.³

Scaffold: ga0248456_123
Niveispirillum sp. th1-14
IMG ID: 2773857929

Scaffold: ga0058932_11
Burkholderia sp. b14
IMG ID: 2603880219

Scaffolds predicted to be eCIS containing plasmids

- ★ Transposase related genes
- ★ integrase related genes

Figure S2. Four examples of predicted plasmid-borne eCIS operons. Genes that play a role in plasmidic activities are marked.

Figure S3. Evidence for local eCIS locus integration events into bacterial genomes. a. A local eCIS integration into specific strains of *Paenibacillus* strain. **b.** A local eCIS insertion into an *Enterobacter* strain. The integration was likely mediated via homologous recombination using the flanking eCIS tRNA-Ser genes and an external eCIS copy that carries this gene.

Query AFP 13 homolog -- IMG gene ID, genome name

Subject Organism name of BLAST hit in NCBI; Accession

Query Afp13 homolog = 637464290

(*Photorhabdus laumondii* TTO1)

Subject = Organic Lake phycodnavirus; ADX06690.1

putative phage tail collar domain family protein [Organic Lake phycodnavirus]

Sequence ID: [ADX06690.1](#) Length: 303 Number of Matches: 3

Range 1: 107 to 303 [GenPept](#) [Graphics](#) [Next Match](#)

Score	Expect	Method	Identities	Positives	Gaps
79.7 bits(195)	4e-21	Compositional matrix adjust.	72/227(32%)	103/227(45%)	55/227(24%)
Query 217	GLSLKNGVKVAGN	-----GIKVDK-----VSID-----PKNVLPGRMIVMFSG			
Sbjct 107	GL + K + AGN + K + D G V + ID N + PRG I + M + SG				
Query 259	SSVPEGWALCDGKDNRPNLIDRFIM-GGTTQNIIGKSSDSFSGAKDNKFTFISEQTVR				
Sbjct 167	+PE W + LCDG + P + L + RFI + G + NIG SG S + T + T				
Query 318	ISG-STDGHLTADENGPHQHEQGETLNRQKCHNGYTEDNDRDWDGRRGPEPP---				
Sbjct 212	+S + H T + + G H H Q GY + D + + + G G PP				
Query 374	-NYRAYTFPSGKGN-PHSHPINLTSGGHNSNNVTVPYILAFIIMKL 418				
Sbjct 257	GNVRTSTYIDNTNSSHQHFTTNSGTSFDIRPPYALAFIMKL 303				

Query Afp13 homolog = 2558033937

(*Photorhabdus temperata* J3)

Subject = Yasminevirus sp. GU-2018; VBB19014.1

hypothetical protein YASMINEVIRUS_1546 [Yasminevirus sp. GU-2018]

Sequence ID: [VBB19014.1](#) Length: 226 Number of Matches: 2

Range 1: 98 to 226 [GenPept](#) [Graphics](#) [Next Match](#)

Score	Expect	Method	Identities	Positives	Gaps
74.3 bits(181)	6e-20	Compositional matrix adjust.	54/176(31%)	78/176(44%)	51/176(28%)
Query 243	VLPRGMIVMFGSSVPIGWALCNGENDTPNLIDRFIMGGTQDIGGSSNSFSGSKDKK				
Sbjct 98	VLPRGIIVAWSGSTIPNGWALCDGNSGTPNLIDRFILGSRIN-----AKATGGSS				
Query 303	FTFTSENQTVHISGKTGHLTADENGKHTHEQGETLNEKGMCHNKYTTDSSRDVWNGG				
Sbjct 149	T + E Q H H + + G + C K + + + GG				
Query 363	SSGSNPPKYRPFYTFPSGE----GKPHSHSINLTSGHSHKNNVTVPYILAFIMKL 414				
Sbjct 192	+S YR + P + E G + P S + P + Y LAFIM + L 226				

Query Afp13 homolog = 2523973452

(*Aquimarina latercula* DSM 2041)

Subject = Burkholderia virus Bcep221; NP_944294.2

NP_944294.2 phage tail fiber protein [Burkholderia virus Bcep221]

Sequence ID: [Query_58581](#) Length: 292 Number of Matches: 3

Range 1: 118 to 290 [Graphics](#) [Next Match](#)

Score	Expect	Method	Identities	Positives	Gaps
96.7 bits(239)	2e-27	Compositional matrix adjust.	68/180(38%)	86/180(47%)	28/180(15%)
Query 210	GLISMWSGQ---DIPK---GWVLCDDGTNTTDLRSKGFIVGMEKNNPEYQIGKGGKAKEVT				
Sbjct 118	G + + W + G DIP GW L DG N T DL RFIV + Y G GGA V				
Query 264	LTEAQMPSHTL---DKGHNHNITDPGHNNHNSGFNGLVRFTRDRTDMDNPNAGKEFA				
Sbjct 175	LT AQMP H H D GH H + + DP H H S G + T + + G G F				
Query 322	LNNF---GTIQSNKTGVSIKTAKANL-----QSTGNGKAHENRPPYFLAYIMKL				
Sbjct 231	+N + GTI + + TG + I A + + Q + G AHENRPPY + LA I +				

Query Afp13 homolog = 2599844965

(*Thiothrix caldifontis* DSM 21228)

Subject = Staphylococcus phage vB_SauS_IMEP5; ANM47011.1

tail fiber protein [Staphylococcus phage vB_SauS_IMEP5]

Sequence ID: [ANM47011.1](#) Length: 603 Number of Matches: 3

Range 1: 396 to 601 [GenPept](#) [Graphics](#) [Next Match](#)

Score	Expect	Method	Identities	Positives	Gaps
72.4 bits(176)	8e-18	Compositional matrix adjust.	64/215(30%)	95/215(44%)	25/215(11%)
Query 228	WSSLSFNAHNNNNNAWVFPDNTTAVTLEMDHSTG-LNQAREFDVWTTTQSKTAWALRL				
Sbjct 296	+S S + H A + D T T TL D S G N F + S A + R				
Query 287	RILGDSGNV---GIGVEQPSKELEVNGNIK-----INGIKTNFTIQEDWIAPTL				
Sbjct 451	G + + IG V + + + NG I + + GG + + T + D T				
Query 334	LNGWVYGHVWNTAGYFKDSFGIVHLKGLIKNGS--TGQPIFTLPAGYRPAQQLHAAC				
Sbjct 508	NGW Y + + G VHL GLIK + G + G FTLP GY P + +				
Query 392	NSNTIGRIDIRANGEVIMESGSNAWISLDGITFRA 426				
Sbjct 567	+ + IG + I I + G + ES + + ISL GI + F A 601				

Query Afp13 homolog = 2640816185

(*Photorhabdus thracensis*:DSM 15199)

Subject = Phage 5P_1; AZF90167.1

hypothetical protein BPH43C_44 [Phage 5P_1]

Sequence ID: [AZF90167.1](#) Length: 237 Number of Matches: 4

Range 1: 86 to 237 [GenPept](#) [Graphics](#) [Next Match](#)

Score	Expect	Method	Identities	Positives	Gaps
70.5 bits(171)	1e-18	Compositional matrix adjust.	59/184(32%)	84/184(45%)	35/184(19%)
Query 224	LPGKMIVMFGSSAPTGWAFCDGKTYNGVTPDLRSRFVMCGETISDTGKSSNKAAGSS				
Sbjct 86	+P + G + I + + + SG + P + GWA CDG NG PDLR + RFV + G + G				
Query 284	AKNFFRDTPKTTVSVSVNQNTTLTEAQIPSHKHIGGMPYCWGTGMKYIGFSD---TQTO				
Sbjct 142	----LQYIKP-TATVTIGGHSLLTAEV-MHTHSYDYDTP-----TGNTLGAYGSRRTTST				
Query 341	YQIDNTYNSSIRKHSQSLYACTSNTGGGQHNNHATASSPSHNHSDVIPPYYLLAF				
Sbjct 192	Y + T YAC + + GG H H AT + + DV PPY LA +				
Query 401	IMKL 404				
Sbjct 234	IMKL 237				

Query Afp13 homolog = 2609602034

(*Chitinophaga* sp. CF418)

Subject = uncultured Mediterranean phage uvMED; BAR33157.1

virion structural protein [uncultured Mediterranean phage uvMED]

Sequence ID: [BAR33157.1](#) Length: 1620 Number of Matches: 13

Range 1: 1514 to 1620 [GenPept](#) [Graphics](#) [Next Match](#)

Score	Expect	Method	Identities	Positives	Gaps
75.1 bits(183)	9e-18	Compositional matrix adjust.	45/113(40%)	66/113(58%)	7/113(6%)
Query 452	KGKARAFGLYLD-TDVRKVSNPVKLGNDVVKVKNLPVTFTFNQKANCPPAEQIGFLPH				
Sbjct 1514	+ + A G + + D + DV + V + + D VKKL P T + P E QIGF +				
Query 511	QVEEFFPELVNTDGDGTQTLAYANMVAULTKAIQEQEETIAALQKRLEALEGK 563				
Sbjct 1569	+ + E PE + V + + DG + + + AY + AVLT KAI QEQE I LQ + + L GK				

Figure S4. Examples of high identity BLAST protein sequence alignments between Afp13 sequence queries and tail fibre proteins from bacteriophages and viruses as the targets.

Figure S5. RES-Xre Gene pair synteny. A gene pair containing the gene for EAT5 is conserved through various strains from proteobacteria and actinobacteria. All occurrences except for the top one are outside eCIS genomic context (IMG locus tag: Ajs_1583; RhiBR816DRAFT_0322 ;Swit_5317; Ga0056099_04803 ;Ga0056090_4110). Based on this synteny we predicted that the purple gene serves as an antitoxin.

Figure S6. Examples for putative eCIS-associated toxins (EATs) carrying DUF4157 at the N terminus. The upper number indicates IMG gene ID. The lower number indicates the locus tag (from IMG website).

Hemopexin (PF00045)

Annexin (PF00191)

MORN (PF02493)

Prominin (PF05478)

GRP (PF07172)

PAAR_motif (PF05488)

Figure S7. Several eCIS toxins carry protein domains that are mostly found in eukaryotic annotated genomes.

The sunburst plots, representing species distribution and taken from Pfam website <https://pfam.xfam.org/>, represent the distribution of protein domains in available annotated genomes. The color coding for each plot appears below. As a reference we added the PAAR motif found in the eCIS spike which is mostly a bacterial domain.

Figure S8. Kinase domains are found upstream to the putative GerE eCIS transcription regulator. An illustration of the different pfam domain architectures and their occurrences within the kinase genes found upstream to the GerE gene that is found in the eCIS operons.

Figure S10. Biological replicates of EATs killing *S. cerevisiae*. EATs were cloned into pESC -leu Galactose inducible plasmids that were then transformed into *Saccharomyces cerevisiae* BY4742 strain. Overnight cultures of the strains harboring the vectors of interest were grown in SD -leu media. The cultures OD was normalized and then washed once with water and split into two: one part was grown overnight in repressive conditions (SD -leu + 2% glucose) and the other part was grown in inductive conditions (SD -leu + 2%

galactose). Dilutions were spotted on SD -leu plates containing glucose or galactose and the plates were incubated two nights at 30C. Negative controls: empty vector (ev) and non-toxin (yeGFP gene).

Figure S11. eCIStem database scheme.

Supplementary Tables

Table S1. eCIS operons. Information about the 1426 eCIS operons we identified and their encoding organisms.

Table S2. Genes included in eCIS operons.

Table S3. eCIS predicted to be carried on plasmids. We ran Deeplasmid tool on all contigs/scaffolds carrying eCIS operon and annotated plasmids based on Deeplasmid score >0.7. Deeplasmid is a deep learning based classifier that identifies plasmids based on presence of plasmidic genes and other features such as coding density, and GC content.

Table S4. Correlation between eCIS and genera. For each genus in our genomic database we performed a fisher exact test by counting the number of sequenced genomes within the genus encoding eCIS, those that lack eCIS, genomes from all other genera that encode eCIS, and enomes from all other genera that lack eCIS.

Table S5. eCIS correlation with microbial ecological and physiological features. All data was downloaded from IMG database.

Table S6. BLAST results of Apf13 proteins against proteins from viruses and phages. The best hits are presented.

Table S7. Pfam enrichment in eCIS operons. Results of Fisher exact test for enrichment of the different Pfam domains within proteins encoded by eCIS operons.

Table S8. eCIS Accessory Gene Clusters. This table shows all accessory proteins (i.e. non-AFP homologs), clustered by at least 40% identity over at least 80% of each of their lengths. Cluster ID = a unique nominal identifier for each 40% identity group. Blank entries in Conserved Domain columns represent no hit from the NCBI CDD database. Widespreadness refers to each cluster, and how many of the indicated taxa are represented in each cluster. Operon ID eCISem Database is a unique identifier for the eCISem database (see full text for more information).

Table S9. Core Gene Clusters. This table shows clusters of core genes (i.e. AFP homologs). Each cluster is a unique nominal identifier for genes with at least 40% ID over at least 80% of clusters.