Systematic Discovery of Antimicrobial Polymorphic Toxins

Nimrod Nachmias[1,*], Noam Dotan[1,*], Maor Shalom[1], Arbel Rivitz[1], Naama Shamash-Halevy[1], Yaara Oppenheimer-Shaanan[1], and Asaf Levy[1, &]

[1] The Department of Plant Pathology and Microbiology, The Institute of Environmental Science, The Robert H. Smith Faculty of Agriculture, Food, and Environment, The Hebrew University of Jerusalem, Rehovot, Israel.

[*] These authors contributed equally to the work.

[&] Corresponding Author. alevy@mail.huji.ac.il

Abstract

Microbes employ a large diversity of toxins to kill competing microbes or eukaryotic host cells. Polymorphic toxins are a class of protein toxins abundant in Nature and secreted through various secretion systems. Polymorphic toxins have a modular protein architecture with a toxin domain found at the protein C-terminus. The discovery of microbial toxins is important to improve our understanding of microbial ecology and infectious diseases. Here, we developed a computational approach to discover novel toxin domains of polymorphic toxins and applied it to 105,438 microbial genomes. We validated nine short novel toxins ("PTs") that lead to bacterial cell death; one of them also kills yeast. For four toxins, we also identified a cognate immunity gene ("PIM") that protects the toxin producing cell from toxicity. Additionally, we found that the novel PTs are encoded by ~2.2% of the sequenced bacteria, including numerous Gram-negative and -positive pathogens. We explored two avenues to determine the mechanism of action of these nine PTs. First, we used fluorescence microscopy to observe severe changes in cell size, membrane and chromosome morphology upon expression of the novel toxins. We then identified putative catalytic sites of these toxins, which, upon mutation, abolished their harmful activities. Thus, using this unique hybrid computational-experimental pipeline, we were able to expand the microbial toxins repertoire significantly. These new potent toxins likely play essential roles in inter-microbial competition in Nature and can be utilized in various biotechnological and clinical applications.

Keywords

Bacterial toxins, polymorphic toxins, inter-microbial competition, bacterial genomics

Introduction

Microbes have evolved a plethora of mechanisms to compromise surrounding eukaryotic and prokaryotic cells. One of the common killing or growth inhibition strategies is the injection of protein toxins through various secretion systems, such as the type III-VII bacterial secretion systems (T3SS-T7SS). The outcome of killing competing microbes or host cells can be a shift in the microbial community structure or a disease to the host, respectively. The proteins that are being injected target essential cellular components such as the DNA [1–5], RNA [6,7], cell membrane [8], cell wall [910], translation machinery [6,7,11], ATP [12], among others. Among the antagonistic molecular tools employed by bacteria, polymorphic toxins [13,14] stand out as an intriguing toxin protein class. These are fairly large proteins that are defined by a specific protein domain architecture (Figure S1). The protein starts with an N-terminal domain that serves as a trafficking domain, which associates with a secretion system or serves as a signal peptide translocating the protein to the periplasm. The trafficking domain is followed by repetitive elements that might stretch over significant length, such as Rearrangement Hot-Spot (Rhs) or hemagglutinin repeats. Toward the C-terminal end of the protein there are releasing peptidase or pre-toxin domains. The C-terminus of the protein is a toxin domain that is eventually delivered into and kills target cells by forming pores in the cell envelope, degrading or blocking the synthesis of peptidoglycan and nucleotides or degrading nucleic acid derivatives [13]. Genes encoding polymorphic toxins that are designed to kill microbes are usually followed by immunity genes that protect the toxin producing microbe from self-intoxication (Figure S1) [15]. The name 'polymorphic toxin’ is derived from the modular nature of these systems, which display high polymorphism in all of its constituent domains and elements, including at the toxin position. Namely, polymorphic toxin variants can exist in different genomes, with some protein architectural changes which allow these proteins to be delivered by different secretion systems, to be processed differently, and to kill cells using a variety of mechanisms. This variability likely allows polymorphic toxin diversification through homologous recombination events [9,14].

Polymorphic toxins mostly play important competition roles in various microbes. For example, Tc toxins are known for their insecticidal role. A recent survey of TcC proteins of bacterial Tc toxins identified that these are polymorphic toxins with over 100 different putative toxic domains[16]. Firmicutes in the human gut are armed with LXG PTs which mediate antimicrobial antagonism[10]. *Staphylococcus aureus* employs T7SS to secrete a polymorphic toxin with a nuclease activity to kill rival bacteria [17] and Bacteroidetes likely use T9SS to secrete antibacterial DNase polymorphic toxin[18]. *Burkholderia thailandensis* uses contact-dependent growth inhibition (CDI) and associated delivered polymorphic toxins to antagonize foreign bacteria and to promote biofilm formation in resistant self bacteria [19]. The soil microbe *Serratia marcescens* secretes Rhs-encoding effectors with DNAse activity to mediate intraspecies competition[4]. Pei *et al*. studied a polymorphic toxin that acts as an endonuclease and is secreted by T6SS of *Aeromonas dhakensis*[5]. Myxobacteria use arrays of polymorphic toxins that are transferred through outer membrane exchange (OME) mechanism to discriminate clonal cooperators from non-self and to ensure populations are genetically homogenous[20,21]. Given their highly conserved functions in inter-microbial antagonism, it is clear that discovery of novel polymorphic toxins will provide a better mechanistic understanding of microbial colonization and persistence in various habitats. Discovery of novel toxin domains may lead to elucidation of novel efficient enzymatic mechanisms as these domains likely kill target cells even in low concentration. Furthermore, toxin activities and targets may inspire development of novel therapies for medical use [22] and biotechnological applications such as base editing [23]. Hence, our hypothesis was that a systematic search for functional C-terminal toxin domains within polymorphic toxin genes can yield novel short toxins with unique and efficient toxic enzymatic activity.

Here, we performed a large-scale discovery followed by an in depth characterization of new toxin domains that are abundant in microbial polymorphic toxins. First, we developed a computational framework to predict novel C-terminal toxin protein domains based on their association with polymorphic toxins. To do so, we analyzed 105,438 microbial genomes across the tree of life and predicted conserved toxin domains that were part of polymorphic toxins. The association of these putative toxin domains with different trafficking domains suggest that they are employed as weapons secreted by various bacterial secretion systems. For several toxin domains we also identified adjacent putative cognate immunity genes. We focused on novel toxin domains that lacked amino acid sequence similarity to known toxins. Then, heterologous expression experiments of toxin candidates were performed to test their cellular toxicity. Our experiments uncovered nine new antibacterial toxin domains that efficiently kill *E. coli*, one of which is also an anti-eukaryotic domain that efficiently kills *S. cerevisiae*. For four antibacterial toxins, cognate immunity proteins were identified that rescued the expressing cells from toxicity. We analyzed the taxonomic and ecological distribution of the novel toxins and found that they are encoded by ~2.2% of sequenced bacteria, including by major human, animal, and plant pathogens. We further used fluorescence microscopy and observed various cell morphology damage types that are caused by the activity of toxins with diverse modes of action. Furthermore, we revealed and mutated amino acids that are critical for the cellular toxicity, and are likely located within the toxin catalytic sites. Additionally, we used prediction of 3D folding to simulate the toxin structure and the microenvironment of its putative active site. Hence, this work significantly expands our knowledge of conserved and potent microbial toxins, their cognate immunity proteins that are used in warfare against prokaryotes and eukaryotes, and it uncovers mechanistic details regarding activity of the novel toxins.

Results

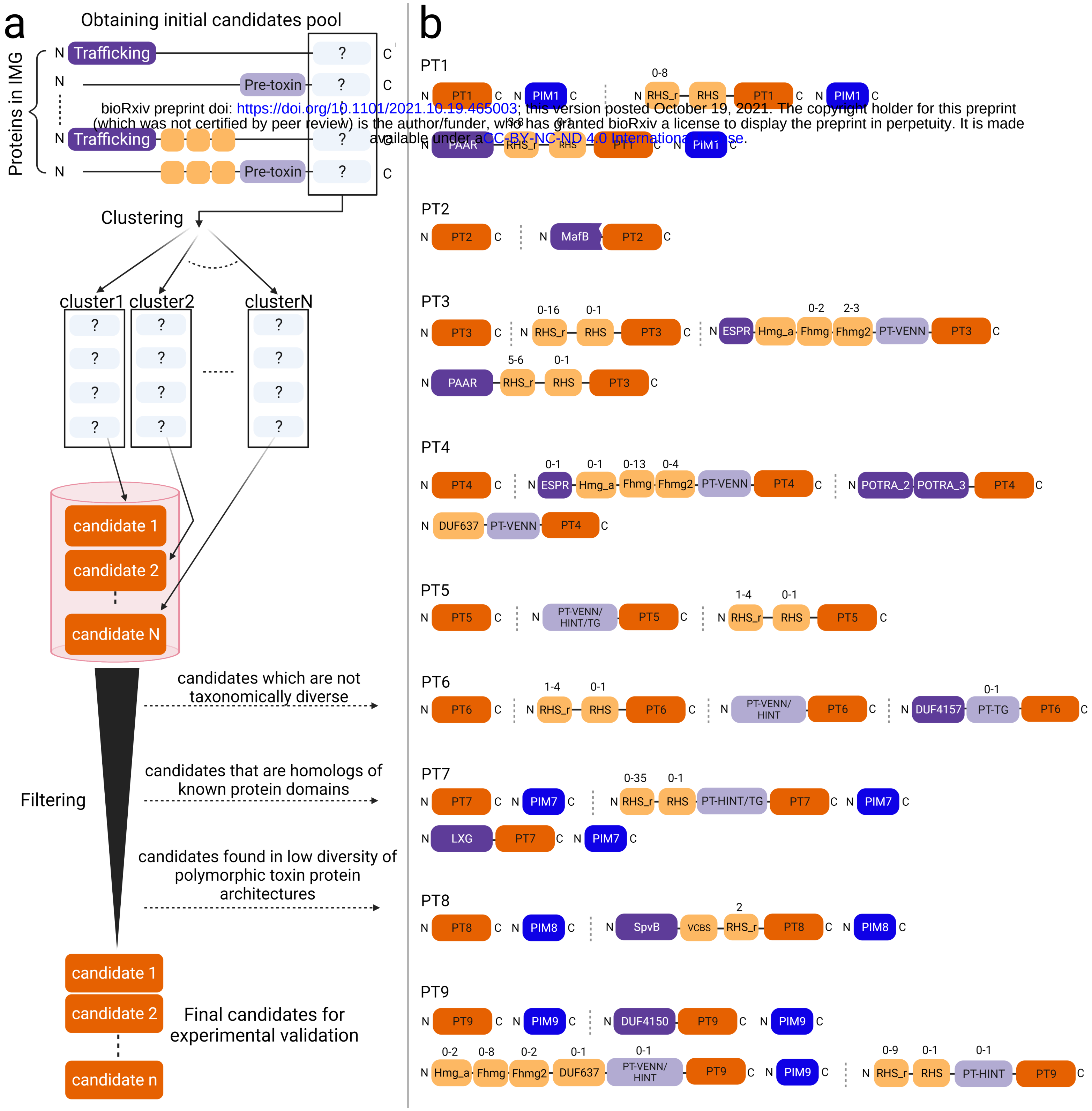
In silico systematic discovery of novel C-terminal toxin domains of polymorphic toxin proteins and their cognate immunity genes.

Our objective was to discover novel toxin domains that are located at the C-terminal end of polymorphic toxin proteins. The algorithm we developed is described in Figure 1a. First, we compiled a list of 217 potential polymorphic toxin domains, defined as trafficking, repeat, pre-toxin or toxin domains, by performing an exhaustive iterative search in the Pfam database (Supplementary Table S1, Methods). Using these domains, we searched for polymorphic toxin genes that encode these domains within 105,438 microbial genomes. From these polymorphic toxin genes, we retrieved their putative C-terminal toxic domains (serving as toxin candidates), defined as the last 133 aa of the protein, which is the median size of a protein domain in Pfam. Next, We clustered these sequences to yield 152,150 clusters representing putative toxin families. To reduce false positive hits, several filters were applied to the clusters (Figure 1a). We kept clusters that had no hits to known domains (to focus on novel toxins), were taxonomically diverse enough (assuming that they are under long purifying selection or may have been horizontally transferred), and were found in high diversity of polymorphic toxin protein architectures (Methods). Finally, we obtained a list of putative C-terminal domains of PTs that very likely serve as potent, novel toxins (Table 1 and Supplementary Table S2).

We experimentally validated nine novel toxin domains and termed “PT1-9” (a short for Polymorphic Toxin C-terminal domains) and we describe these in the next sections (Figure 1b). Notably, we observed that the novel PTs are part of varied large polymorphic toxin proteins that have a large diversity of N-terminal trafficking domains. The trafficking domains can lead to the PT secretion through various secretion systems. Therefore, we predict that PT1, PT3, and PT9 are secreted through the T6SS as they are fused to PAAR[24] and DUF4150[25] domains, PT6 is secreted through the extracellular contractile injection system (eCIS) as it is fused to DUF4157 domain [26], and PT4 is secreted via the T5SS as it is fused through ESPR domain[26,27]. In addition, PT7 is likely secreted via the T7SS as it is fused to LXG domain[10], and PT4 is likely secreted through the Omp85/TPS β-barrel proteins as it fused to POTRA_2 and POTRA_3 domains[28]. Importantly, for certain PTs we identified a cognate downstream gene that we predicted to serve as an immunity gene (Figure 1b). We termed these genes Polymorphic IMmunity genes (PIMs).

The PT-PIM pairs were predicted based on conserved synteny of the two gene families across numerous genomes.

Figure 1. Computational prediction of novel toxin domains (PTs) and cognate immunity genes (PIMs) in polymorphic toxin proteins. **(a) A pipeline for prediction of novel toxic protein domains (PTs).** C-terminal sequences of predicted polymorphic toxin proteins having one or more trafficking/pre-toxin/repeat domains were considered potential toxin domains (PTs). These C-termini were clustered to obtain an initial pool of toxin candidates. Candidates were filtered out in multiple stages, keeping ones which are sufficiently taxonomically diverse, novel, and found in high diversity of polymorphic toxin protein architectures. This pipeline led to a final list of candidates that very likely bear toxic activity. For each final candidate, we obtained a representative by searching for a single domain gene which is a homolog of the cluster members. The representatives were used in the experimental validation step. **(b) Most abundant protein architectures that include our novel PTs.** The protein domains were annotated with Pfam domains using hmmscan (RHS_r, Hmg_a, Fhmg, and Fhmg2 stands for RHS_repeat, Haemagg_act, Fil_haemagg, and Fil_haemagg_2, respectively). The toxin domains (PTs) are shown in orange, cognate adjacent immunity genes are in blue, repeat domains are in beige, pre-toxin domains are in light purple, and trafficking domains are in dark purple. The number above the domain is the range of possible consecutive occurrences of the domain in this type of architecture. Exact numbers of occurrences of the architectures are shown in Supplementary Table S5.


a
Obtaining initial candidates pool
Proteins in IMG
N Trafficking ? C
N Pre-toxin ? C
N Trafficking ? C
N Pre-toxin ? C
Clustering
cluster1
cluster2
clusterN
?
......
candidate 1
candidate 2
candidate N
Filtering
candidates which are not taxonomically diverse
candidates that are homologs of known protein domains
candidates found in low diversity of polymorphic toxin protein architectures
candidate 1
candidate 2
candidate n
Final candidates for experimental validation
b
PT1
N PT1 C N PIM1 C
0-8
N RHS_r RHS PT1 C N PIM1 C
3-8 0-1
N PAAR RHS_r RHS PT1 C N PIM1 C
PT2
N PT2 C
N MafB PT2 C
PT3
N PT3 C
0-16 0-1
N RHS_r RHS PT3 C
0-2 2-3
N ESPR Hmg_a Fhmg Fhmg2 PT-VENN PT3 C
5-6 0-1
N PAAR RHS_r RHS PT3 C
PT4
N PT4 C
0-1 0-1 0-13 0-4
N ESPR Hmg_a Fhmg Fhmg2 PT-VENN PT4 C
N POTRA_2 POTRA_3 PT4 C
N DUF637 PT-VENN PT4 C
PT5
N PT5 C
N PT-VENN/HINT/TG PT5 C
1-4 0-1
N RHS_r RHS PT5 C
PT6
N PT6 C
1-4 0-1
N RHS_r RHS PT6 C
N PT-VENN/HINT PT6 C
0-1
N DUF4157 PT-TG PT6 C
PT7
N PT7 C N PIM7 C
0-35 0-1
N RHS_r RHS PT-HINT/TG PT7 C N PIM7 C
N LXG PT7 C N PIM7 C
PT8
N PT8 C N PIM8 C
2
N SpvB VCBS RHS_r PT8 C N PIM8 C
PT9
N PT9 C N PIM9 C
N DUF4150 PT9 C N PIM9 C
0-2 0-8 0-2 0-1 0-1
N Hmg_a Fhmg Fhmg2 DUF637 PT-VENN/HINT PT9 C N PIM9 C
0-9 0-1 0-1
N RHS_r RHS PT-HINT PT9 C

Experimental validation of novel toxins (PTs) show that they efficiently kill bacteria or yeast

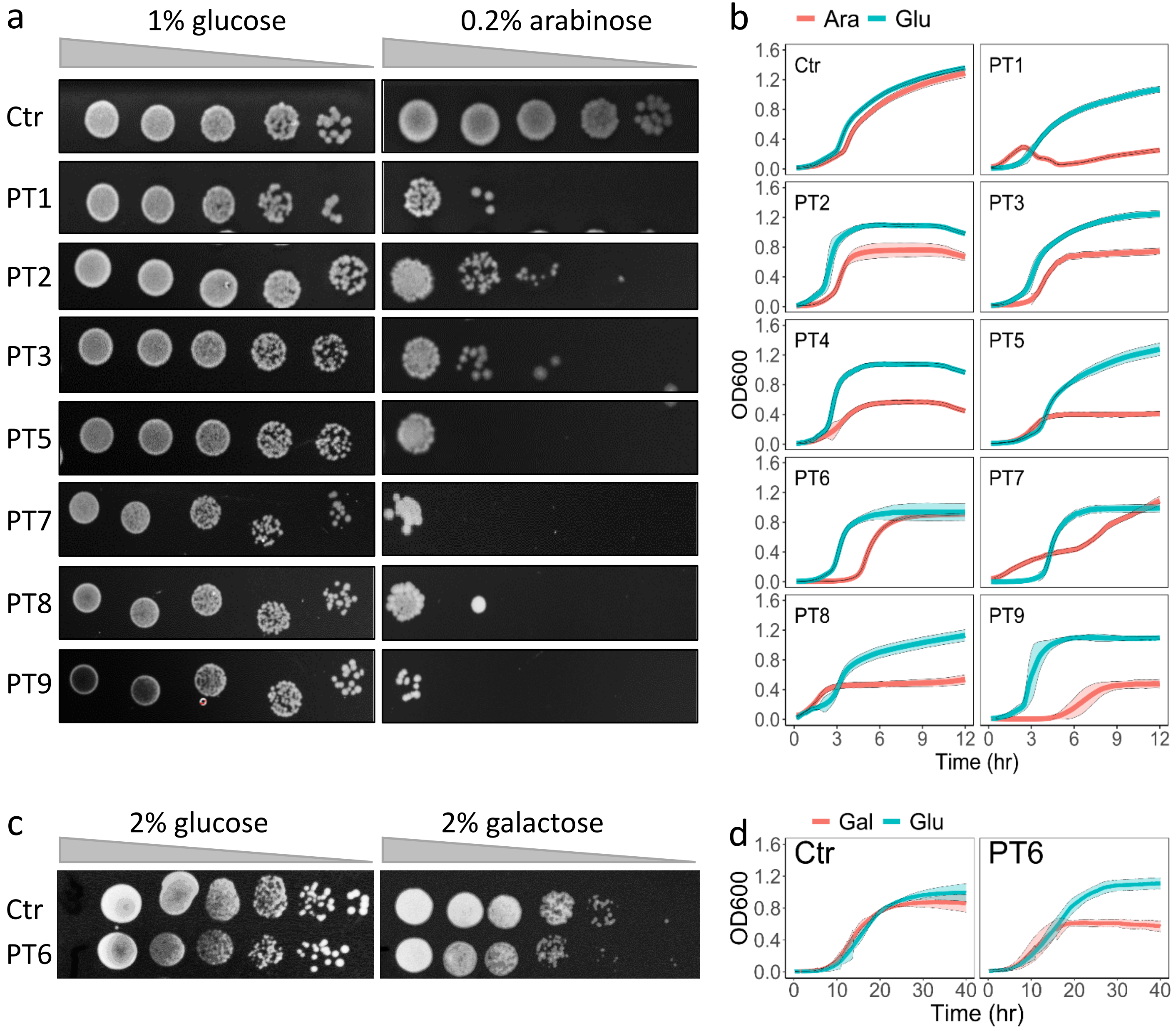
We asked whether the predicted toxic domains (PTs) will exert a toxic phenotype in bacterial or eukaryotic cells. Therefore, we chose an *in-vivo* approach and used heterologous expression assays to test toxicity of the fourteen novel toxins in expressing cells. Since we did not have the precise PT domain borders we searched for single domain genes that include the PT (Table 1, Supplementary Table S2). The selected short genes were taken from various Gram-positive and -negative genomes, and the genes encoded putative toxin domains shorter than 150 aa that were not mapped to any known Pfam domain. Our selection of PTs's target cell in *E. coli* or *S. cerevisiae*, was partially based on presence or absence of cognate immunity genes next to the PTs, as we expected immunity proteins are necessary to prevent prokaryotic death. We cloned and expressed the putative antibacterial PTs in pBAD24 plasmid in *E. coli* with arabinose induction and the anti-eukaryotic PTs in pESC plasmid in *S. cerevisiae* with galactose induction. Strikingly, already in the synthesis and cloning stage we noticed that three genes, despite being synthesizable *de novo,* were toxic and recalcitrant to cloning and propagation in plasmids. Namely, these genes were toxic to the hosting microbe even without induction as even leaky expression was sufficient for cell killing. Genes that were successfully cloned in *E. coli* and *S. cerevisiae* were plated on agar plates supplemented with inducer (arabinose or galactose) or a repressor (glucose). Importantly, nine genes (PT1-9) exhibited high toxicity to *E. coli* of up to four orders of magnitude reduction in CFU*,* two of which required a periplasmic translocation signal (twin-arginine translocation signal) (PT4 and PT6) (Figure 2a, 2c and Supplementary Figure S2b). Moreover, the PT activity of nearly all toxins inhibited growth in liquid media. We monitored a rapid growth arrest resulting in a range of 4-10 fold longer lag time or approximately two-fold decrease in exponential growth bacterial concentrations (as measured by optical density) of PTs expressing strains compared to the control (Figure 2b). Remarkably, PT6 was toxic to both yeast cells and bacterial cells when expressed in the bacteria periplasm (Figure 2b-d, Supplementary Figure S2c). These findings demonstrate that the nine novel toxins we predicted using our computational algorithm are highly toxic to recipient cells.

Table 1. Novel PTs and PIMs identified in this study. Protein ID and existing protein annotation lay on the NCBI database.

Protein Name (this work)	**Locus Tag**	**Protein ID**	**Organism**	**Protein Length (AA)**	**Existing Protein Annotation**	**Domain Annotation**
PT1	C4A13_00009	RDR26550.1	*Escherichia marmotae*	119	hypothetical protein	Unannotated
PIM1	C4A13_00008	RDR26549.1	*Escherichia marmotae*	122	hypothetical protein	Unannotated
PT2	NMEN3081_2264	EJU75496.1	*Neisseria meningitidis* NM3081	276 (the toxin should be ≤ 146 aa)	hypothetical protein	MafB at the N-terminus, the C-terminus is unannotated
PT3	JAW84_RS16460	WP_080727551.1	*Ralstonia solanacearum*	125	hypothetical protein	Unannotated
PT4	EGH62_13680	RSW81050.1	*Klebsiella aerogenes*	149	adhesin	Unannotated
PT5	Gene is no longer available	WP_084153669.1	*Rheinheimera baltica* DSM 14885	119	hypothetical protein	Unannotated
PT6	B0180_RS04825	WP_078255899.1	*Moraxella canis* strain CCUG 8415A	85	type IV secretion protein Rhs	Unannotated
PT7	IE3_05452	EJQ03417.1	*Bacillus cereus* BAG3X2-1	78	hypothetical protein	Unannotated

PIM7	IE3_05453	EJQ03418.1	*Bacillus cereus* BAG3X2-1	117	hypothetical protein	Unannotated
PT8	LEP1GSC009_0601	EKO85988.1	*Leptospira interrogans* serovar Grippotyphosa str. Andaman	118	hypothetical protein	Unannotated
PIM8	LEP1GSC009_0600	EKO85992.1	*Leptospira interrogans* serovar Grippotyphosa str. Andaman	116	hypothetical protein	Unannotated
PT9	LH23_RS24110	WP_156108067.1	*Cedecea neteri* strain M006	88	hypothetical protein	Unannotated
PIM9	LH23_RS23680	WP_071842756.1	*Cedecea neteri* strain M006	147	DUF4279 domain-containing protein	DUF4279

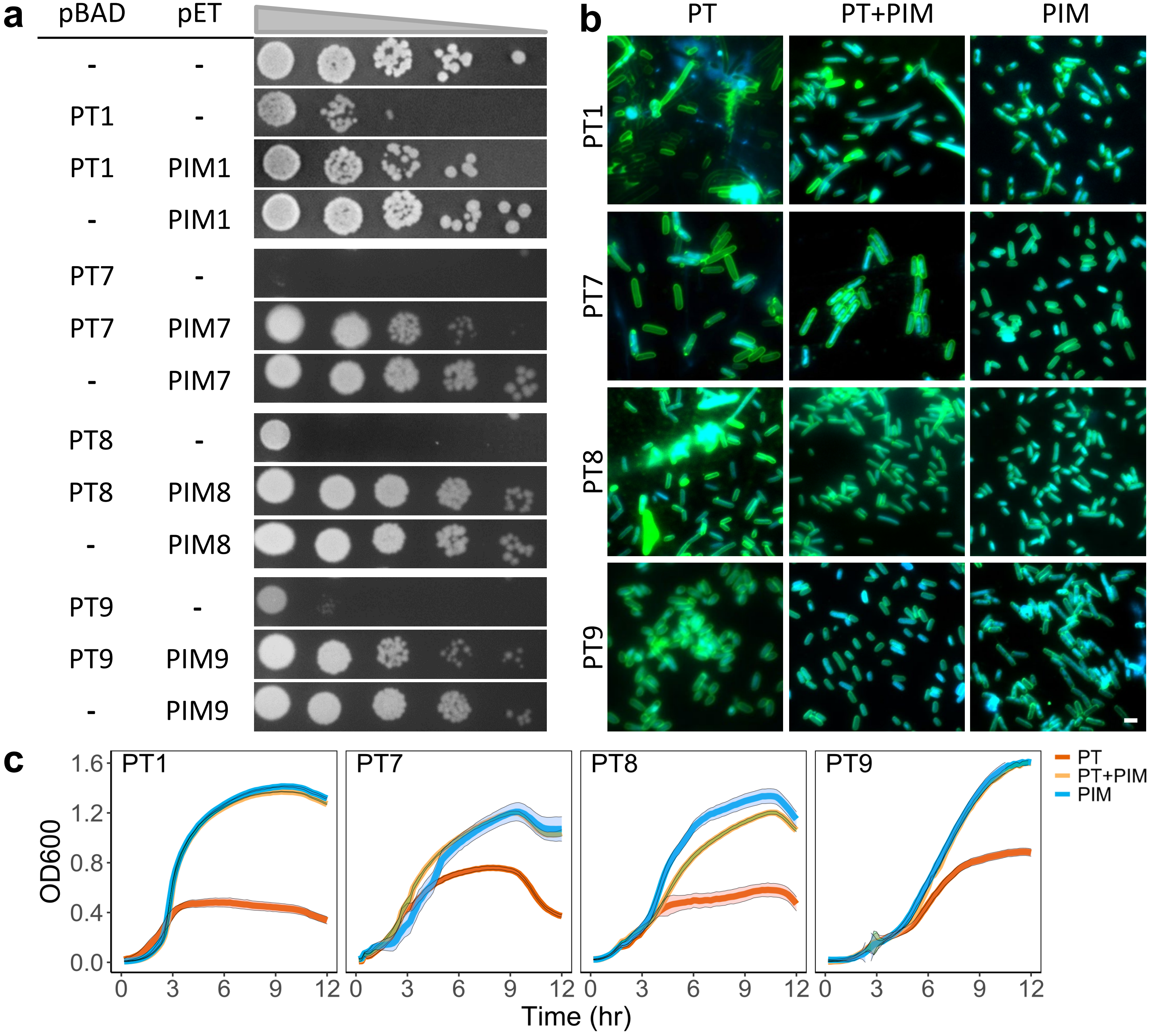
Figure 2. The novel PTs efficiently kill bacteria and/or yeast. The toxins were cloned into *E. coli* BL21 (DE3) under the control of an arabinose-inducible promoter. As control, cells that lack the toxin (empty vector) were cloned (Ctr) **(a)** Bacteria were plated in 10-fold serial dilution on LB-agar in conditions that repress gene expression (1% glucose) or induce expression (0.2% arabinose). **(b)** Growth curves of *E. coli* BL21 (DE3) in liquid culture. Cells were grown in LB-liquid media supplemented with 1% glucose or 0.2% arabinose. OD600, optical density at a wavelength of 600 nm. (*n*=3, ± SD for values). **(c-d)** PT6 toxin was cloned into *Saccharomyces cerevisiae* BY4742 under the control of a galactose-inducible promoter. As control, cells that lack the toxin (empty vector) were cloned (Ctr) **(c)** Yeast were plated in 10-fold serial dilution on SD-agar in conditions that repress gene expression (1% glucose) or induce expression (1% galactose). **(d)** Growth curves demonstrate the toxic effect of PT6 toxin on *S. cerevisiae* BY4742 growth for 40 hours. Cells were grown in SD-liquid media supplemented with 2% glucose or 2% galactose. OD600 of three technical replicates were tracked. (*n*=3, ±SD for values).


a
1% glucose
0.2% arabinose
Ctr
PT1
PT2
PT3
PT5
PT7
PT8
PT9
b
Ara
Glu
Ctr
PT1
PT2
PT3
PT4
PT5
PT6
PT7
PT8
PT9
OD600
1.6
1.2
0.8
0.4
0.0
0
3
6
9
12
Time (hr)
c
2% glucose
2% galactose
Ctr
PT6
d
Gal
Glu
Ctr
PT6
OD600
0
10
20
30
40
Time (hr)

The toxicity of four novel PTs is rescued by novel cognate immunity genes

Antibacterial toxins encoded by bacteria are often escorted by adjacent immunity genes, as known for T6SS, toxin-antitoxin, and contact-dependent inhibition systems, for instance [15,29–31]. We speculated that some of the novel PTs had adjacent cognate PIM genes (Figure 1b). We hypothesized that PIM genes have the ability to protect host bacteria from self-killing. To this end, we tested the immunity function of these genes by co-transformation and co-induction of the putative PT-PIM pairs carried on different expression vectors, pBAD24 and pET28a for PT and PIM, respectively. As control, cells were co-transformed with empty vectors along with the PT or the PIM. Bacteria harboring the PIM grew similarly to the control strain on solid and liquid media and exhibited normal cellular morphology (Figure 3). These results point out that four PIMs are non-toxic to *E. coli*. As predicted, these PIMs successfully rescue the bacteria from self-killing by the cognate PT to a large degree (Figure 3). The rescue by the four PIMs was clearly noted when the bacteria were grown on solid agar and in liquid media (Figure 3a and 3c,). Moreover, using fluorescence microscopy we observed that the PT-PIM pair or the PIM led to normal cells, whereas the PT expression led to cellular aberrations at the single cell level (Figure 3b). These cellular aberrations will be discussed in further detail in a later section. We noted that the novel PIMs lack sequence similarity to known immunity genes, except for PIM9 which is mapped to an unknown function DUF4279 domain.

Figure 3. PIMs provide protection against toxicity of cognate PTs. The toxins (PT) were cloned under the control of an arabinose-inducible promoter (pBAD24) and the protein immunity protein (PIM) were cloned under the control of an inducible promoter system that is active only in the presence of IPTG (pET28) into *E. coli* BL21 (DE3). **(a)** Bacteria were plated in a 10-fold serial dilution on LB-agar in conditions that induce expression (0.2% arabinose and/or 0.01mM IPTG). As control, cells that lack the toxin (empty vector) or PIM (empty vector) were plated. **(b)**. Fluorescence microscopy images of membrane stain (green) with DNA stain (blue) overlay captured at 40 min after induction with 0.2% arabinose and 0.01 mM IPTG. Each row shows cells containing one novel toxin with PIM. Representative images from a single replicate out of three independent replicates are shown. Scale bar corresponds to 2 µm. **(c)**. Cells were grown in LB liquid media supplemented with 0.2% arabinose and 0.01m mM IPTG. OD600 = optical density at a wavelength of 600 nm. (*n*=3 , ± SD for values).


a
pBAD
pET
-
-
PT1
-
PT1
PIM1
-
PIM1
PT7
-
PT7
PIM7
-
PIM7
PT8
-
PT8
PIM8
-
PIM8
PT9
-
PT9
PIM9
-
PIM9
b
PT
PT+PIM
PIM
PT1
PT7
PT8
PT9
c
PT1
PT7
PT8
PT9
OD600
1.6
1.2
0.8
0.4
0.0
0
3
6
9
12
Time (hr)
PT
PT+PIM
PIM

Distribution of the novel toxins across bacterial taxa, ecosystems, and in bacterial pathogens

To provide a better understanding of the broad function of the novel PTs we analyzed their taxonomic distribution (Figure 4) and their presence in various pathogens (Table 2). PT1 is found in Proteobacteria, Firmicutes and Bacteroidetes phyla including in the human pathogens *Clostridium botulinum*, *Burkholderia cenocepacia*, *Cronobacter sakazakii*, and the plant pathogen *Dickeya zeae*. PT2 is Neisseriaceae-specific, including in the human pathogens *Neisseria gonorrhoeae* and *Neisseria meningitidis*. PT3 is widely distributed along different phyla (Firmicutes, Actinobacteria, Fusobacteria, Bacteroidetes, Proteobacteria). It is found in the human pathogens *Pseudomonas aeruginosa* (Supplementary Table S3), *Klebsiella aerogenes*, *Salmonella enterica*, *Acinetobacter baumannii*, and *Cronobacter sakazakii* and the plant pathogen *Ralstonia solanacearum*. The PT3 representative we validated from *Ralstonia solanacearum* shares weak amino acid sequence similarity (41% identity) with the C-terminal of the T6SS effector RhsP2 from *Pseudomonas aeruginosa* [32] and likely these toxins belong to the same toxin superfamily. PT4 is found only in the Proteobacteria phylum including in the human pathogens *Klebsiella aerogenes*, *Klebsiella pneumoniae*, *Klebsiella variicola*, *Yersinia pestis*, *Yersinia pseudotuberculosis*, *Serratia marcescens*, *Salmonella enterica*, *Cronobacter sakazakii* and in the animal and plant pathogens *Edwardsiella anguillarum*, *Edwardsiella ictaluri*, *Photorhabdus luminescens*, and *Pantoea agglomerans*. PT5 is found in Actinobacteria, Cyanobacteria and Proteobacteria phyla, including in the human pathogen *Vibrio vulnificus*. PT6 is found in Firmicutes, Actinobacteria, Spirochaetes, Bacteroidetes and Proteobacteria phyla but is absent from pathogens. PT7 is found in Firmicutes and Proteobacteria phyla, including the human pathogens *Bacillus cereus*, *Acinetobacter baumannii*, *Listeria monocytogenes* and in the insect pathogen *Bacillus thuringiensis*. PT8 is specific to the Leptospiraceae family, almost only in the human pathogen for Leptospirosis, *Leptospira interrogans*. PT9 is found in Cyanobacteria, Bacteroidetes and Proteobacteria phyla. It is found in the human pathogen *Bordetella genomosp* and the insect pathogen *Xenorhabdus cabanillasii*. Proteobacteria phylum encoded all novel PTs, except for PT8, with Enterobacteriaceae, Morganellaceae, Pseudomonadaceae, and Neisseriaceae families encoding four PTs each. Interestingly, although Cyanobacteria is a relatively sequenced phylum with thousands of available genomes we identified a single new PT family (PT5) in a single family, Microcoleaceae.

We also correlated the toxin encoding bacteria with the bacterial habitat, life-style, and physiology suggesting that toxin presence might be associated with these biological features (Table 3). This can allow us to infer important ecological features that likely pertain to the toxin activity. We statistically corrected this analysis for biases in the genomic data as the PTs can be found in phylogenetically related genomes which are intensively sequenced (Brynildsrud et al. 2016) (Methods). Three PTs, PT3, PT4, and PT9, were enriched in host-associated bacteria. PT4 is enriched in motile bacteria isolated from humans, fish, insects, and nematodes. PT3 is enriched in specific human-associated niches where bacteria dwell including sputum, the excretory system and the respiratory system. Interestingly, PT5 is enriched in environmental bacteria, such as the aquatic *Rheinheimera baltica* and *Cupriavidus pauculus,* and the plant-associated *Cupriavidus plantarum*. Overall, the nine novel PTs are widespread across the microbial world, and are encoded in microbes isolated from various ecosystems, of different life-styles, including in major human, animal, and plant pathogens.

Figure 4. **Taxonomic distribution of the novel toxins.** The phylogenetic tree was constructed using marker genes from representatives from each taxonomic family. Therefore, each leaf in the phylogenetic tree represents a bacterial family (labeled in the perimeter). Leaf color represents the phylum of the family and outer rings represent the percentage of genomes in the matching family that contain at least one occurrence of the toxin, according to the IMG database (shades of red).

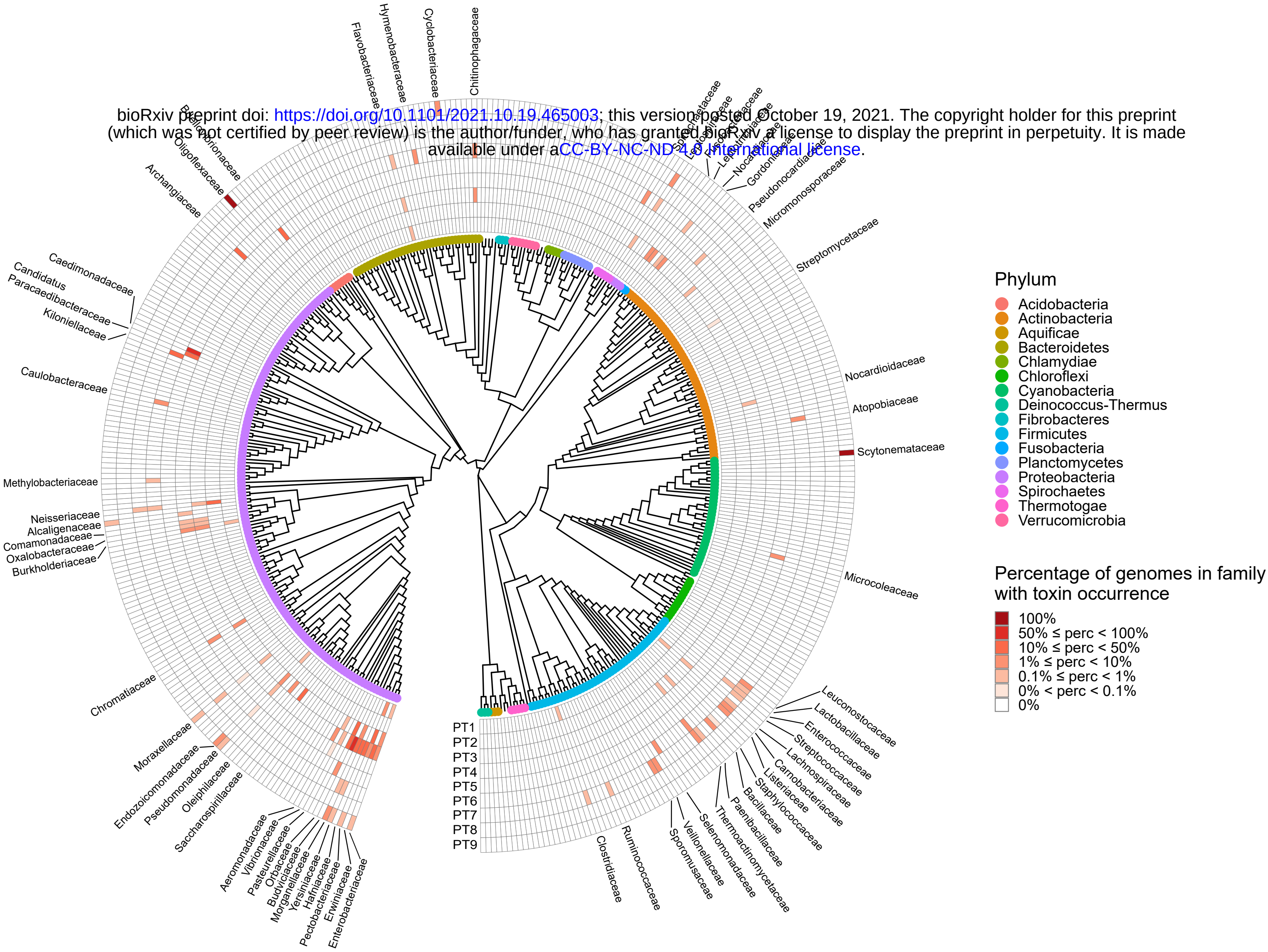

Flavobacteriaceae
Hymenobacteraceae
Cyclobacteriaceae
Chitinophagaceae
Spirochaetaceae
Leptospiraceae
Fusobacteriaceae
Leptotrichiaceae
Nocardiaceae
Gordoniaceae
Pseudonocardiaceae
Micromonosporaceae
Streptomycetaceae
Nocardioidaceae
Atopobiaceae
Scytonemataceae
Microcoleaceae
Leuconostocaceae
Lactobacillaceae
Enterococcaceae
Streptococcaceae
Lachnospiraceae
Carnobacteriaceae
Listeriaceae
Staphylococcaceae
Bacillaceae
Paenibacillaceae
Thermoactinomycetaceae
Selenomonadaceae
Veillonellaceae
Sporomusaceae
Ruminococcaceae
Clostridiaceae
PT1
PT2
PT3
PT4
PT5
PT6
PT7
PT8
PT9
Enterobacteriaceae
Erwiniaceae
Pectobacteriaceae
Hafniaceae
Yersiniaceae
Morganellaceae
Budviciaceae
Orbaceae
Pasteurellaceae
Vibrionaceae
Aeromonadaceae
Saccharospirillaceae
Oleiphilaceae
Pseudomonadaceae
Endozoicomonadaceae
Moraxellaceae
Chromatiaceae
Burkholderiaceae
Oxalobacteraceae
Comamonadaceae
Alcaligenaceae
Neisseriaceae
Methylobacteriaceae
Caulobacteraceae
Kiloniellaceae
Paracaedibacteraceae
Candidatus
Caedimonadaceae
Archangiaceae
Oligoflexaceae
Bdellovibrionaceae
Phylum
● Acidobacteria
● Actinobacteria
● Aquificae
● Bacteroidetes
● Chlamydiae
● Chloroflexi
● Cyanobacteria
● Deinococcus-Thermus
● Fibrobacteres
● Firmicutes
● Fusobacteria
● Planctomycetes
● Proteobacteria
● Spirochaetes
● Thermotogae
● Verrucomicrobia
Percentage of genomes in family
with toxin occurrence
■ 100%
■ 50% ≤ perc < 100%
■ 10% ≤ perc < 50%
■ 1% ≤ perc < 10%
■ 0.1% ≤ perc < 1%
■ 0% < perc < 0.1%
□ 0%

Table 2. Bacterial pathogens that encode the novel PTs. Human pathogens are colored in red, non-human animal pathogens are in purple, plant pathogens are in green.

Species	Toxins encoded
Acinetobacter baumannii	PT3, PT7
Actinobacillus equuli	PT4
Aeromonas eucrenophila	PT4
Bacillus cereus	PT7
Bacillus mycoides	PT7
Bacillus thuringiensis	PT7
Brenneria alni	PT4
Brenneria nigrifluens	PT4
Burkholderia cenocepacia	PT1
Burkholderia multivorans	PT1
Clostridium botulinum	PT1
Cronobacter sakazakii	PT1,PT3,PT4
Dickeya zeae	PT1,PT4
Edwardsiella anguillarum	PT4
Edwardsiella ictaluri	PT4
Klebsiella aerogenes	PT3,PT4
Klebsiella pneumoniae	PT4
Klebsiella variicola	PT4
Leptospira interrogans	PT8
Listeria monocytogenes	PT7
Neisseria gonorrhoeae	PT2
Neisseria meningitidis	PT2

Pantoea agglomerans	PT4
Photorhabdus luminescens	PT4
Pluralibacter gergoviae	PT1
Pseudomonas aeruginosa	PT3
Ralstonia solanacearum	PT3
Salmonella enterica	PT3,PT4
Serratia marcescens	PT4
Vibrio vulnificus	PT5
Xenorhabdus cabanillasii	PT9
Yersinia pestis	PT4
Yersinia pseudotuberculosis	PT4
Yersinia ruckeri	PT4

Table 3. Correlation of bacteria encoding the novel toxins with habitats and bacterial life-styles

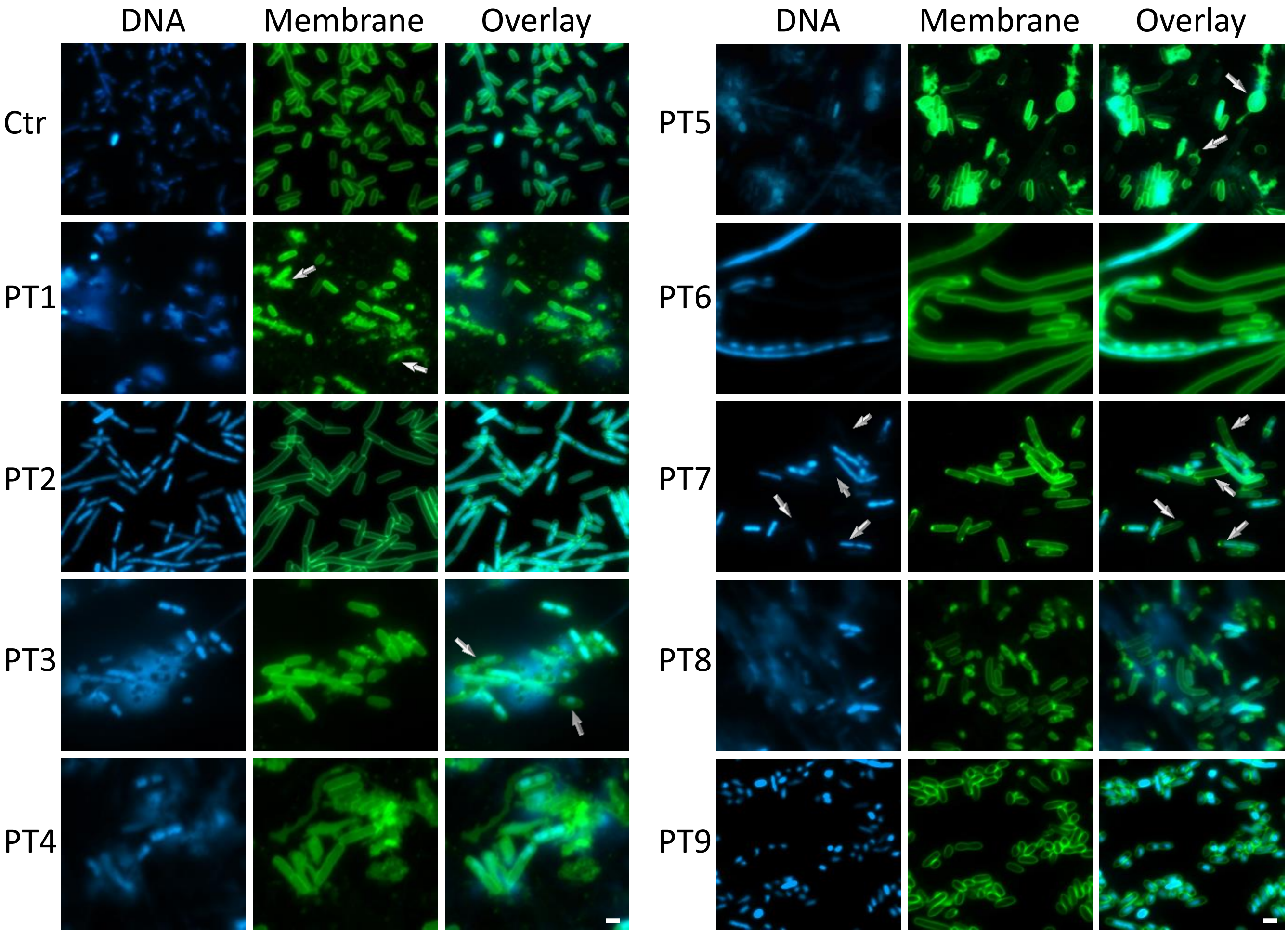
Toxin	Category	Trait	Number of toxin encoding genomes with trait	Fisher Test Odds ratio (enrichment)	Fisher test q-value (-log10)	Scoary q-value (-log10)
PT3	Ecosystem	Host associated	135	5.085	15.886	1.814
	Ecosystem subtype	Sputum	19	15.416	13.498	4.47
	Gram staining	Gram negative	44	10.008	13.801	2.143
	Habitat	Human airways	3	22.644	1.461	1.458
	Ecosystem type	Excretory system	5	5.496	1.708	5.111
	Ecosystem type	Respiratory system	20	5.763	6.947	5.476
PT4	Gram staining	Gram negative	136	inf	70.182	3.311
	Ecosystem	Host associated	277	6.537	38.291	6.77
	Ecosystem category	Human	156	2.161	9.622	1.713

	Ecosystem category	Insecta	9	3.666	2.136	1.55
	Ecosystem category	Fish	7	3.912	1.84	1.615
	Ecosystem type	Nematoda	3	8.332	1.391	2.164
	Motility	Motile	50	9.455	9.223	1.716
PT5	Ecosystem	Environmental	10	inf	3.66	1.505
PT9	Ecosystem	Host associated	33	9.482	5.273	2.651
	Gram staining	Gram negative	49	inf	25.219	1.806

The effect of the novel toxin activities on cell size, cell membrane and DNA morphology

To investigate the mechanism of action of the toxins, we observed the damage caused by the novel PTs at a single cell resolution. PT-expressing cells were followed by fluorescence microscopy at approximately 40 minutes post-induction (Figure 5). Some of the toxins likely cause damage to the cell DNA. PT3 and PT7 cells maintained their rod shape and membrane architecture, characteristics of wild-type cells, but the chromosome appears to be contracted and in some cells has vanished. Namely, the signal from 4′,6-diamidino-2-phenylindole (DAPI) staining was hardly detectable. Thus, PT3 and PT7 may degrade the chromosome. Remarkably, PT9 expression led to mini-cells and chromosomal condensation and DNA foci at the cell center suggestive a DNA stress response. Other toxins likely target the cell membrane or the cell wall. PT1 displayed membrane damage manifested by membrane perturbation and leakage of the DNA outside the cells. This finding suggests that PT1 may cause membrane perforation. Interestingly, we observed that PT7 causes membrane accumulation in the cell poles with elongated cells, which may indicate a damage to membrane integrity. When we expressed PT4 in cells a membrane disintegration was caused and the cells were swollen and lost their rod shape. Induction of PT5 led to round cells suggesting that this toxin damages the peptidoglycan layer and thereby leading to loss of the rod-like structure. In addition we observed large cell aggregates. Two other PTs led to cell division inhibition. PT2 induction led to cell elongation and long interconnected cells. Expression of PT6 with an N-terminal *Tat* sequence led to massive cell filamentation with multiple nuclei which indicates a strong cell division inhibition. However, when PT6 was expressed in yeast cells, we observed some cells with an abnormal membrane (Supplementary Figure S3). PT8 expressing cells had odd shapes as well as membrane protrusions that indicate probably perforation of the membrane or inappropriate cell divisions. In line with those views, the different morphological abnormalities may provide directions to study more thoroughly the PT mechanisms of cellular toxicity.

Figure 5. Toxins lead to cell death in various manners. *E. coli* BL21 (DE3) cells harboring toxin genes (PT) or empty vector as control (Ctr) were grown in LB media in conditions that induce expression (0.2% arabinose). After 40 min of incubation, samples were stained with membrane and DNA stains and visualized by fluorescence microscopy. Shown are DNA (DAPI) (blue), membrane (FM1-43; green) and overlay images of the bacterial cells with toxins as indicated. Ctr is the control of wild-type *E. coli* BL21 transformed with an empty vector. Arrows point to abnormal cells in the specific toxin expreation. Representative images from a single replicate out of three independent replicates are shown. Scale bar corresponds to 2 µm.

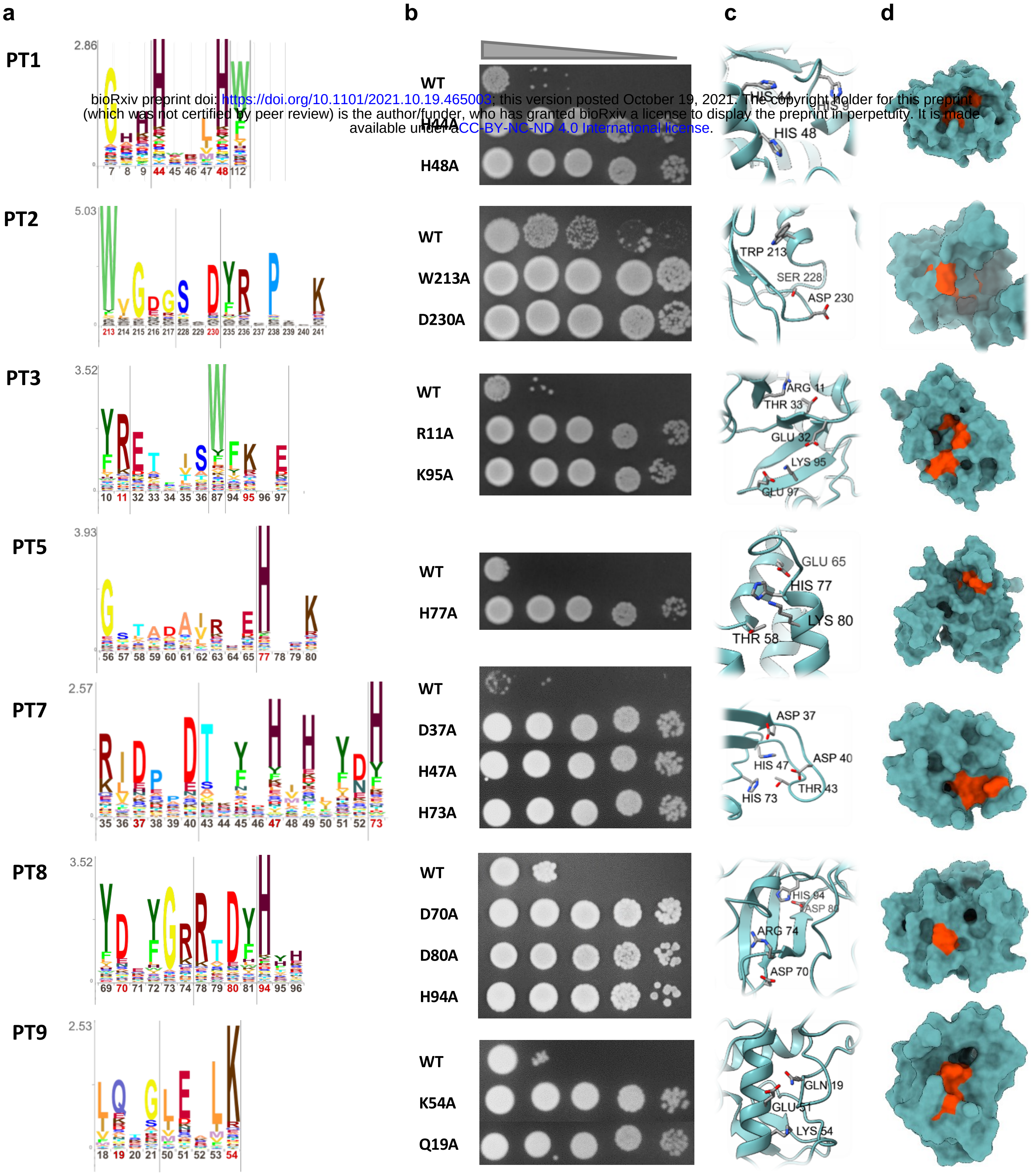

DNA
Membrane
Overlay
DNA
Membrane
Overlay
Ctr
PT1
PT2
PT3
PT4
PT5
PT6
PT7
PT8
PT9

Conserved amino acids and fold prediction of protein structures suggest involvement of the toxins in diverse enzymatic activities

Bacteria employ highly active enzymes as cytotoxic effectors as a well-established strategy in bacterial competition as well as in virulence in eukaryotes [14,33–35]. We sought to explore whether newly discovered toxins possess unique folds and conserved residues to execute their lethal functions. To explore the folding of the proteins, we predicted the PT 3D structures using roseTTA Fold [36] and compared the structures to known structures using DALI [37]. Our results suggest that most of the PTs we discovered do not resemble the structure of any known protein (Supplementary Table S4). However, PT3 predicted structure resembles the Diphtheria and Cholix toxins which act as ADP-ribosyltransferase of EF-2 protein [38,39] , and PT4 resembles proteins that act as NADH dehydrogenase. The other PTs presented weaker similarities to known protein (lower Dali Z-score) or inconsistent hits. Next, we searched for conserved residues that are critical for toxin activity and may serve in potential enzymatic activity of the PTs. We used two approaches to select conserved residues to mutate. First we generated HMM logos based on alignments of all the toxin homologs that are present in over 100,000 bacterial genomes and selected the most conserved residues (Figure 6a and Supplementary Figure S5). Second, we applied Jensen-Shannon Divergence scoring (Methods) to each position to infer functionally important residues from sequence conservation [40]. Among the high scoring positions, we looked for residues with high propensity to serve in catalytic sites of enzymes, i.e. residues with basic, acidic or nucleophilic nature [41]. Using these features, we selected 1-3 residues to mutate per toxin. We substituted the conserved residue with alanine via directed mutagenesis of the pBAD24 plasmid harboring the toxins. Strikingly, point mutations of selected PTs led to complete abolishment of the toxic phenotype, restoring normal growth when bacterial cultures were grown on solid (Figure 6b) and liquid media (Supplementary Figure S2), and we observed normal cellular morphology at the single cell level (Supplementary Figure S6).

Next, we mapped the residues that were critical to PT function to the folded protein. Surprisingly, highly conserved residues with high propensity to participate in catalysis were spatially clustered and directed towards each other on the simulated atomic microenvironment (Figure 6c and Supplementary Figure S7). Surface display of the model shows that these residues generally form a binding pocket-like shape on the protein surface (Figure 6d). Taken together, these data suggest that the new PTs serve as a diverse group of enzymatic toxins. For instance, PT1 harbors two highly conserved histidines spaced by three variable positions and is predicted to be located along an α-helix. In the predicted fold, these residues are in proximity with histidine in position 9. This motif may serve in binding of divalent ions similarly to other zinc-binding metallopeptidases that harbor HEXXH motifs [42] and synthetic 6X histidine tags that chelate Nickel ions [42,43]. The conserved simulated microenvironment of PTs 3,5,7 and 9 show variable combinations of a base, an acid, and a nucleophile residue, suggesting a catalytic-triad type mechanism of action that is characteristic of many hydrolases and transferases [44,45]. To conclude, the novel PTs toxins have a protein fold that resembles several enzymes with critical amino acids in the predicted catalytic sites. The exact enzymatic activities are yet to be elucidated.

Figure 6. Folded structures and critical residues in the novel PTs. (**a**) HMM logo of conserved boxes in the novel toxins (names appear from top to bottom on the left side of the panel). Highly conserved regions in logos generated by skylign were cropped and are numbered by position with respect to each experimental candidate. Position numbers marked in red note experimentally tested residues. **(b)** The toxins (WT) and generated toxin mutants were cloned into *E. coli* BL21 (DE3) under the control of an arabinose-inducible promoter. Bacteria were plated in 10-fold serial dilution on LB-agar in conditions that induce expression (0.2% arabinose). (**c**) Microenvironment 3D modeling of roseTTA Fold predictions for candidates (from top to bottom; % confidence): PT1 (72%), PT2 (52%), PT3 (84%), PT5 (79%), PT7 (85%), PT8 (50%) and PT9 (66%). Backbone is displayed as a cartoon model colored in cadet blue. Residues with high conservation ratios and propensity for involvement in catalysis are displayed as sticks without hydrogen, residue three letter code and position number are marked. **(d)** Surface displays of the predicted models, colored in cadet blue. Residues that were marked in **(b)** are colored in orange red.


a
b
c
d
PT1
2.86
7 8 9 44 45 46 47 48 112
WT
H44A
H48A
HIS 44
HIS 9
HIS 48
PT2
5.03
213 214 215 216 217 228 229 230 235 236 237 238 239 240 241
WT
W213A
D230A
TRP 213
SER 228
ASP 230
PT3
3.52
10 11 32 33 34 35 36 87 94 95 96 97
WT
R11A
K95A
ARG 11
THR 33
GLU 32
LYS 95
GLU 97
PT5
3.93
56 57 58 59 60 61 62 63 64 65 77 78 79 80
WT
H77A
GLU 65
HIS 77
LYS 80
THR 58
PT7
2.57
35 36 37 38 39 40 43 44 45 46 47 48 49 50 51 52 73
WT
D37A
H47A
H73A
ASP 37
ASP 40
HIS 47
HIS 73
THR 43
PT8
3.52
69 70 71 72 73 74 78 79 80 81 94 95 96
WT
D70A
D80A
H94A
HIS 94
ASP 80
ARG 74
ASP 70
PT9
2.53
18 19 20 21 50 51 52 53 54
WT
K54A
Q19A
GLN 19
GLU 51
LYS 54

Discussion

Polymorphic toxins are multi-domain proteins involved primarily in inter-bacterial competition that were discovered nearly a decade ago [14]. In recent years it is becoming clear that polymorphic toxins play important roles in various bacteria, archaea and temperate phages, and their toxins can serve as DNAses, RNAses, rRNA ADP-ribosylating and membrane depolarization enzymes, metallopeptidase, and cell wall biosynthesis inhibitors [9–11,13,15,18,20,46–52]. Here, we developed a new algorithm to predict novel toxin domains (PTs) located at the C-terminus of polymorphic toxin proteins across over 100,000 genomes from the microbial world. We also predicted the cognate immunity genes (PIMs) of the new PTs. This *in silico* analysis was experimentally confirmed in heterologous expression assays in *E. coli* and *S. cerevisiae* and led to discovery of nine classes of potent PTs that are toxic to bacteria or yeast and four cognate PIMs that rescue cells from the toxin activities.

The novel PTs are part of various polymorphic toxins that based on the N-terminal trafficking domains genetically associate with various microbial secretion systems, such as T5SS, T6SS, T7SS, and eCIS (Figure 1b). We analyzed the ecological and phylogenetic properties of the PTs and identified that they are encoded by microbes from seven different bacterial phyla, including in major human, animal, and plant pathogens. Some of the PTs are associated with certain ecological niches. For example, PT4 is enriched in animal microbiomes and PT3 is enriched in the microbiome of specific human tissues. This finding supports the existing hypothesis arguing that proteins involved in antagonistic activity are commonly shared between bacterial phyla and are probably genetically associated with the delivery apparatus via horizontal gene transfer and recombination events.

We shed light on the basic yet versatile mechanism of action of the toxin through imaging of PT expressing cells using fluorescence microscopy. We can indicate potential cellular targets of toxin activity such as DNA, cell wall, cell membrane damaging, or the cell division process. We then predicted the protein folding of the novel toxins and experimentally confirmed critical residues, buried in predicted enzymatic active sites, that are likely participating in substrate catalysis. Overall, this work significantly expands the space of toxin domains of PTs and uncovers in the macro-scale the broad role of these toxins through their correlation with different phylogenetic and ecological data and potential target organisms, and in the micro-scale the PT cellular function within the attacked cells. Using our approach we confirmed that the modular nature of these systems serves as a good predictor for novel PTs. We strongly suggest that additional PT domains and cognate PIMs are still hiding in microbial genomes and some fine-tuning of the parameters we used in computational prediction can expand these gene functions much further. Identification of novel efficient enzymes are of great interest for the research community as well as the biotechnological and medical industries.

Some predicted toxins failed to show toxicity in our systematic experimental pipeline. It is plausible that some of our false negative predictions kill bacterial taxa that we have not tested such as Gram-positive bacteria which are well-known targets of polymorphic toxins[10,46,51,52]. Similarly, we are aware that the heterologous expression method we used ironically fails on some of the most potent toxins that are unclonable due to toxicity via leaky expression in *E. coli* (Supplementary Table S2). In some cases we tried transforming the predicted PT into *E. coli* cells expressing the predicted immunity gene but we still got no clones (data not shown). Thus, expanding the host range for heterologous expression in other model organisms might yield favorable results in future studies.

Additional work is required in order to decipher the complete mechanism of action of the novel PTs. For example, beyond correlation with known trafficking domains of specific secretion systems it is important to demonstrate that the PTs are actually being secreted by these secretion systems and that they confer a fitness advantage to the attacking cell against the prey cell. It is important to elucidate the exact biochemical functions of the PTs and decipher how the cognate PIMs inhibit the PT activity. To address these important mechanistic questions there is a need to employ extensive biochemical assays that can reveal the toxin substrates and their precise enzymatic activity. Nevertheless, the mutants generated in this study may aid in paving the way for elucidation of mechanisms of action via *in vitro* biochemical assays with purified proteins. Another aspect that requires more attention is the ecological role of the toxin in colonization and inter-microbial competition in different niches and their plausible role in virulence against eukaryotic cells. Interestingly, the short PT-encoding genes we experimentally validated as toxins are by definition not part of long multi-domains polymorphic toxins. This finding may suggest that these specific genes either function similarly to classical toxin-antitoxin systems. Namely, bacteria use toxins that are abundant in polymorphic toxin proteins for their own growth regulation in response to stress, such as phage attack. In their original paper describing polymorphic toxins Zhang *et al.* described some minor overlap between the toxin-antitoxin systems and polymorphic toxins [14], and here we show that this overlap is more prominent. Alternatively, these short toxins might be loaded into other bacterial secretion systems using adaptor proteins, a phenomenon which is known in many secretion systems [5,53–59].

Materials and Methods

Collection of known domains found in polymorphic toxins

We collected a list of Pfam domains [60] and one TIGRFAM domain that are common in polymorphic toxin proteins. These served as anchors for our search for toxin C-terminal domain. The list included trafficking domains, repeat domains, pre-toxins, and toxins. Trafficking domains included: VgrG, PAAR [3], DUF4150, HCP, LXG [10], DUF4157 [26,50], MafB [61], Wxg100_esat6 , Phage_min_cap2 [14], SpvB [14], PrsW-protease, DUF4280 [3], and TANFOR[18]. Repeat domains included: Haemagg_act [61], filamentous hemagglutinin repeats [61], RHS, RHS repeat [14], TcdB_toxin_midN, DUF637, and ALF. Pre-toxins included PT-HINT [14], PT-TG [14,61], and PT-VENN. Toxins domains included: RNAse_A_bac, AHH, Ntox21, Ntox27, Colicin_D, Ribonuclease, XOO_2897-deam, SCP1201-deam, HNH, Tox-REase-7. We then expanded this list using the next procedure several times, each time treating the newly expanded list as the input for the procedure, until the list was no longer expanded. For each known PT domain ('anchor domain'), we iterated over the Pfam architectures that contain it [60] and examined the domains in each architecture that are yet unknown PT domains (denoted as 'potential PT domain'). If a potential domain in some architecture containing some anchor domain met one of the conditions according to the schema in Supplementary Figure S4, it was added to the expanded list with an annotation of a certain type (trafficking, repeat, pre-toxin, toxin). Assessment of domain type was done using its Pfam name and description. Each type was assigned with some name-keywords and description-keywords that were manually extracted from Pfam description of known PT domains (e.g. "tox" as name-keyword for toxin domains, and "bind" for trafficking). Each presence in the name of some name-keyword contributed 1.5 to the score, and each presence in the description of some description keyword contributed 0.5. If name or description contained the string "polymorphic toxins" it contributed 1.5 to all PT domain types. If the PT domain type with maximum score had a score equal or higher than 1.5 the domain was assigned with this type, otherwise it was annotated as a domain whose type could not be assessed and it was discarded. Putative domains that their assessed type didn't make sense were omitted manually. The final list of 217 domains is listed in Supplementary Table S1 and includes for instance VCBS as a repeat domain and PG_binding_1 as a trafficking domain.

Obtaining initial toxin domain candidates

We used 105,438 microbial genomes that were downloaded from the IMG database in 2018 [62]. We first defined candidate polymorphic toxin genes. Namely, those genes that contain multiple domains found in polymorphic toxins. Gene domain annotations were obtained from IMG data. We calculated that the median domain size in pfam is 133 aa (as of 2020) and therefore we filtered in only genes encoding proteins of at least 2.5 domains (133 aa x 2.5) that contained at least one domain that is a known\putative trafficking, repeat or pre-toxin domain according to the expanded list from the previous analysis. We defined the C-terminus as the last 133 amino acids of the protein. The initial list included 2,633,634 C-terminal sequences from 82,253 genomes. These were clustered using cd-hit version 4.8.1 (40% identity) [63] to obtain 152,150 clusters. These clusters represent unique C-terminal domains that are candidate toxins from polymorphic toxins. However, we were aware that this list included many false positive predictions and in the next steps we worked to increase the signal to noise ratio in predictions.

Filtering putative toxin domains for taxonomic diversity

We took each candidate's cluster and obtained the genus of each member from the IMG genome metadata. We omitted candidates which had a total of less than three unique genera among their cluster members. This left us with a total of 11,196 candidates. The purpose of this step was to maintain only domains that are more likely to be functional and therefore selected during microbial evolution or horizontally transferred between multiple genera.

Filtering putative toxin domains for novelty

To represent the sequence space of each cluster we calculated the multiple sequence alignment of the cluster sequences with Clustal Omega version 1.2.4[64] and then, using the MSA, we constructed a profile hidden markov model (HMM) using HMMER's [65] hmmbuild program. We used the hmmsearch program to search all of the candidate hmms against the Pfam-A fasta file (e-value <= 0.001). Each candidate with some hit to a known domain was omitted, to give us a total of 1,986 candidates that had no Pfam hits. We performed a more sensitive novelty check against additional databases only for top-scored candidates (see below: "Re-filtering toxins for novelty using a more sensitive sequence-based search").

Scoring putative toxin domains for modularity

We used the modularity feature of polymorphic toxins to verify that the novel domains appear in various architectures of polymorphic toxins, thereby supporting their polymorphism. Namely, it is more likely that a real toxin domain (PT) will be fused to multiple polymorphic trafficking and repeat domains that can therefore be secreted by various secretion systems. For each cluster, we searched with DIAMOND [66] (e-value <= 0.001) the cluster representative sequence (decided by cd-hit) against NCBI nr, taking only the subject proteins that had a hit at their C-terminal (alignments that end maximum of 103 aa from the C-terminus), as we were interested in proteins that manifest the polymorphic toxins template architecture. Then, we annotated all these protein domains using hmmscan [65,66] (e-value <= 0.001) against Pfam and created the domain architecture for each of the proteins. For each architecture, we kept a domain annotation only if it was in the expanded list of PT domains, and only if it wasn't at the C-terminus (ended more upstream than 103 aa from the C-terminus). Additionally, if a given architecture had consecutive occurrences of a repetitive domain, we treated it as a single domain occurrence. Finally, we counted all unique architectures, and defined this number as the modularity score of the cluster (namely, the candidate toxin). We treated all candidates with a modularity score of at least four as top-scored candidates. We manually omitted candidates that were only derived from proteins of many different combinations of some repetitive domains, without trafficking or pre-toxin domains for instance.

Re-filtering toxins for novelty using a more sensitive sequence-based search

We took the multiple sequence alignment of each top-scoring candidate and validated their novelty with HHpred server [67] against PDB, SMART, NCBI CDD, Pfam-A, omitting candidates with significant hits (probability > 95%, e-value <= 0.001). Additionally, we searched each candidate's profile hmm with hmmsearch against Swiss-Prot, omitting every candidate with a significant hit of a toxin or an enzyme (e-value <= 0.001, query cover >= 50%).

Selecting genes for synthesis

One of the challenges we faced was to identify the toxin domain borders so we can test our computational predictions in molecular biology experiments. The strategy we employed was to select single-domain protein-coding genes that contain the candidate toxins domains. For each candidate, we searched its cluster representative sequence using BLASTP [68] (e-value <= 0.01) against nr and DIAMOND [66] (e-value <= 0.01, 'very sensitive' option) against nr and IMG databases and manually extracted sequences of significant hits that had long alignment coverage of the subject sequence and did not have significant hits when searched in HHpred (against Pfam-A, NCBI CDD, SMART, PDB databases, probability > 95%, subject cover >= 50%). Seven of the final chosen sequences had length that was abundant among the hits of the cluster sequences. Here we assumed that a high frequency of specific gene length will be the most conserved functional homolog of our candidate toxin.

Obtaining candidates for being immunity genes

For seven of our final candidates we predicted an immunity protein as well by taking the gene downstream to the single domain PT gene in the genome, and manually validating that it was in synteny with the single domain PT gene.

Architectures analysis

Novel toxins' sequences were searched against IMG with DIAMOND (very sensitive, evalue <= 0.001, query cover >= 80%). For each toxin, we analysed the domain architecture of the proteins that had hit of the toxin using hmmscan (c-e-value <= 0.001, query cover >= 50%) against a profile HMM database containing Pfam and TANFOR domain[69] for TIGRFAM[70]. We considered an overlap of domain annotation to be an overlap of >= 50 amino acids of two domain annotations, and dealt with it by keeping the annotation with the lower c-e-value. Hits of validated immunity proteins were found using DIAMOND as above. We extracted all hits of immunity proteins which were immediately downstream of some occurrence of their matching validated toxin using IMG gene IDs, and showed it in Figure 1b as well (Supplementary Table S5).

Phylogenetic analysis

In order to build a phylogenetic tree that spanned our bacterial genomes database, we chose a representative genome from each taxonomic Family as the resolution for our dataset, which yielded a tree in which each leaf represents a Family. As a basis for comparison, we used universal marker genes. Specifically, we used 29 COGs out of 102 COGs that correspond to ribosomal proteins [71]. We aligned each COG from each representative genome using Clustal Omega version 1.2.4 [64]. We then concatenated all the COGs alignments for each genome, filling missing COGs with gaps. We used this 29 marker gene concatenated alignment as input to FastTree2 with default parameters [72] to create the phylogenetic tree. To exhibit the distribution of occurrences of novel toxins, we first identified orthologs of the tested genes by searching for them in Bacterial genomes in IMG using DIAMOND with query cover >=80%, e-value <= 0.001 and 'very sensitive' option. R's ggtree v2.4.2 and ggtreeExtra v1.0.4 packages were used to plot the tree as well as the percentage of genomes in the family having at least one occurrence of the toxin according to our DIAMOND homology results. We present in the tree phylum annotation only for families that encode at least one of the novel toxins, or families that belong to a phylum that contains more than two families according to our tree.

Ecosystem enrichment analysis

The enrichment analysis assigns an odds ratio that quantifies the enrichment of metadata labels using a Fisher exact test. To correct for taxonomic bias , a population-aware enrichment analysis based on Scoary[73] version 1.6.16 was used. The inputs to Scoary were (1) a guide tree, (2) a presence-absence file based on occurrences of toxins (see phylogenetic analysis), indicating if a genome encodes for the toxin, (3) metadata files that indicate for each genome if they possess a certain characteristic, e.g. whether it was isolated from soil. One guide tree was created per metadata category, based on universal marker genes (see phylogenetic analysis) and only those entries with at least 25 instances in the database were included. Statistical significance for both the naive Fisher exact test and for the phylogeny-aware Scoary test were calculated, displayed as -log10 of the q-values. Correlations with q-value <= 0.05 for both Fisher exact test and Scoary test were considered significant. q-values (False discovery Rates) were obtained using the Benjamini-Hochberg procedure [74].

Conserved amino acids analysis

For each toxin we obtained hits in IMG database using DIAMOND (e-value <= 0.001, query-cover >=80%, ‘very sensitive’ settings). For PT2, PT8 we had a small number of hits so we used DIAMOND to search the obtained hits again in the IMG database and used these hits. For each toxin we clustered the hit alignments using cd-hit version 4.8.1 (90% identity) and calculated the MSA of the clusters’ representatives using Clustal Omega version 1.2.4. We then used each of these MSAs to generate HMM logo for each toxin with skylign.org [75] using the following parameters: Create HMM - keep all columns, Alignment sequences are full length, Information Content – All. The same logos, but generated with ‘Create HMM - remove mostly-empty columns’ are shown in Supplementary Figure S7. Additionally, we scored each position using Jensen-Shannon Divergence [40]. Based on this information, we manually chose each toxin top-scored positions with amino acids that have high propensity to be active in catalysis in enzymes [41].

Structural analysis

Predicted structures of the experimentally validated toxins were obtained using RobbeTTAFold [36,76]. Generated 3D models were visualized using ChimeraX [77]. In order to demonstrate co-localization in the microenvironment of catalytic amino acids, we chose cartoon mode and highlighted specific residues (without their hydrogens). Visualisation of the surface with coloring of chosen residues was used to demonstrate catalytic pockets. We used Dali server [78] in order to find proteins with similar structure, and used the ‘Matches against PDB25’ results.

Calculation of percentage of sequenced bacteria encoding the novel toxins

Homologs of novel toxins obtained from the IMG microbial genome database that was downloaded in 2018 using DIAMOND [66] with the following parameters: e-value <= 0.001, query-cover >= 80%, ‘very sensitive’ option). The percentage of sequenced bacteria encoding the novel toxins was calculated based on the number of unique genomes encoding these homologs (n=1339) and total number of IMG genomes with ‘Bacteria’ value in the ‘Domain’ field according to the IMG metadata (n=62075).

Bacterial strains and strain construction

Candidate toxin and immunity gene sequences were retrieved from IMG, synthesized (codon optimized for *E. coli*), and cloned into either pBAD24 (Thermo Fisher Scientific, 43001) or pET28a plasmids by Twist Bioscience, respectively. Growth control for the experiments was an empty vector. The plasmids were then transformed using the previously described TSS method [79] into *E. coli* BL21 (DE3) strain. For Toxin-immunity protection assay, *E. coli* BL21 (DE3) cells harboring pBAD24 plasmids with toxic genes were co-transformed with pET28 vectors harboring cognate immunity genes or with empty vectors as control (Supplementary Table S6 and S7).

Site-directed mutagenesis

Chosen residues were substituted with alanine using Q5® Site-Directed Mutagenesis Kit (NEB) according to supplier protocol. In brief, primers for inverse PCR were generated using automated software offered by NEB. pBAD vectors harboring the genes were linearized using the primers in inverse PCR and then incubated in 1X KLD mix provided with the kit for 5-10 minutes. The mixture was transformed into *E. coli* DH5α strain (NEB), screened by colony PCR, and validated by Sanger sequencing. Plasmids from positive clones were prepared using a mini prep kit and re-transformed into *E. coli* BL21 (DE3) strain. Residues that failed this method were substituted by re-cloning the gene as two overlapping fragments using NEBuilder® HiFi DNA Assembly Cloning Kit (NEB) according to supplier protocol and recommendations. In brief, target genes were PCR amplified as two overlapping fragments having the point of mutation designed in the midst of the overlap. pBAD24 vector linearized with NcoI was incubated with fragment and NEbuilder enzyme mix for 30-45 minutes. The mix was cloned into DH5α strain and validated as mentioned above (Supplementary Table S6 and S7).

Drop assays and Growth curves

For bacterial drop assays and growth curves, overnight cultures of the strains harboring the vectors were grown in LB supplemented with ampicillin, 100 mg/liter (Tivan biotech) or also with kanamycin 50 mg/ml (Tivan biotech) for toxin-immunity cells with shaking (200 rpm). For drop assay, cultures were normalized to OD600 = 0.3 and then serially diluted by a factor of ten. Dilutions were spotted as three biological replicates on LB agar containing ampicillin (100mg/ml) and 0.2% arabinose or 1% glucose for cells containing toxin or mutant toxin only. For toxin-immunity experiments, dilutions were spotted as three biological replicates on LB agar supplemented with both ampicillin (100mg/ml), kanamycin (50 mg/ml), 0.2% arabinose and 0.01mM IPTG respectively. The plates were incubated overnight at 37 °C. Results were documented using amersham ImageQuant 800. To construct growth curves cultures were normalized to OD600 = 0.03 then incubated at 37 °C with shaking at 150 rpm for up to 12 h in a microplate reader (Synergy H1, BioTeck). OD600 were measured continuously over 12 h. Data analysis was performed with R (version 4.0.5) software, ‘ggplot2’ and ‘growthrates’ packages [80,81].

Yeast gene synthesis, heterologous expression and drop assay and growth curves

Heterologous expression of putative toxin in yeast candidate sequences were retrieved from IMG. 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 (Supplementary Table S7). Overnight cultures of the strains harboring the vectors of interest were grown in SD-leu media. The cultures OD was normalized OD600 = 0.3 ,washed once with water and then serially diluted by a factor of ten. Dilutions were spotted as three biological replicates on SD-leu agar containing 2% glucose or 2% galactose and plates were incubated for 48 hours at 30°C. Results were documented using amersham ImageQuant 800. To construct growth curves cultures were normalized to OD600 = 0.03 then incubated at 30 °C with shaking at 150 rpm for up to 40 h in a microplate reader (Synergy H1, Bio Teck). OD600 were measured continuously over 40 h. Data analysis was performed with R (version 4.0.5) software, ‘ggplot2’ and ‘growthrates’ packages [80,81].

Fluorescence microscopy

Microscopy of induced cells *E. coli* BL21 cells that contain the wild-type PTs, the mutated PTs or PT-PIM pairs were grown in LB media at 37 °C. When growth reached an OD600 of 0.3, bacteria were induced with 0.2% arabinose for PTs or 0.2% arabinose and 0.01 Mm IPTG for PT- PIM's strains for 40 minutes. 300 µl of the induced samples were centrifuged at 8,000xg for 2 minutes at 25 °C and resuspended in 5 µl of Dulbecco's phosphate-buffered saline (DPBS) (Thermo Fisher Scientific 14200075), supplemented with 1 mg/ml membrane stain FM1-43 for bacteria and FM4-64 for yeast (Thermo Fisher Scientific T35356 and T13320) and 2 µg/ml DNA stain 4,6-diamidino2-phenylindole (DAPI) (Sigma-Aldrich D9542-5MG). Cells were visualized and photographed using an Axioplan2 microscope (ZEISS) equipped with ORCA Flash 4.0 camera (HAMAMATSU). System control and image processing were carried out using Zen software version 2.0 (Zeiss).

Acknowledgements

AL is generously supported by the Israeli Science Foundation (Grants #1535/20, #3300/20), Alon Fellowship of the Israeli council of higher education, The Hebrew University - University of Illinois Urbana-Champaign seed grant, the Israeli Ministry of Agriculture (Grant 12-12-0002), and ICA in Israel. NN is supported by the Israeli Science Foundation (Grant #1535/20) and a PhD scholarship from Faculty of Agriculture, Food, and Environment, Hebrew University. We thank Asaf Levy's Lab members for support, advice and helpful discussions throughout. We especially thank Alex Geller and Dr Netanel Tzarem (The Hebrew Univeristy of Jerusalem) for critical reading and important discussions.

References

1. Iyer, L. M., Zhang, D., Rogozin, I. B. & Aravind, L. Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. *Nucleic Acids Res.* **39**, 9473–9497 (2011).
2. Koskiniemi, S. *et al.* Rhs proteins from diverse bacteria mediate intercellular competition. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 7032–7037 (2013).
3. 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).
4. Alcoforado Diniz, J. & Coulthurst, S. J. Intraspecies Competition in Serratia marcescens Is Mediated by Type VI-Secreted Rhs Effectors and a Conserved Effector-Associated Accessory Protein. *J. Bacteriol.* **197**, 2350–2360 (2015).
5. Pei, T.-T. *et al.* Intramolecular chaperone-mediated secretion of an Rhs effector toxin by a type VI secretion system. *Nat. Commun.* **11**, 1865 (2020).
6. Morse, R. P. *et al.* Structural basis of toxicity and immunity in contact-dependent growth inhibition (CDI) systems. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 21480–21485 (2012).
7. Michalska, K. *et al.* Structure of a novel antibacterial toxin that exploits elongation factor Tu to cleave specific transfer RNAs. *Nucleic Acids Res.* **45**, 10306–10320 (2017).
8. Russell, A. B. *et al.* Diverse type VI secretion phospholipases are functionally plastic antibacterial effectors. *Nature* **496**, 508–512 (2013).
9. Ghequire, M. G. K., Buchanan, S. K. & De Mot, R. The ColM Family, Polymorphic Toxins Breaching the Bacterial Cell Wall. *MBio* **9**, (2018).
10. Whitney, J. C. *et al.* A broadly distributed toxin family mediates contact-dependent antagonism between gram-positive bacteria. *Elife* **6**, (2017).

11. Jurėnas, D. *et al.* Photorhabdus antibacterial Rhs polymorphic toxin inhibits translation through ADP-ribosylation of 23S ribosomal RNA. *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkab608.

12. Ahmad, S. *et al.* An interbacterial toxin inhibits target cell growth by synthesizing (p)ppApp. *Nature* **575**, 674–678 (2019).

13. Ruhe, Z. C., Low, D. A. & Hayes, C. S. Polymorphic Toxins and Their Immunity Proteins: Diversity, Evolution, and Mechanisms of Delivery. *Annu. Rev. Microbiol.* **74**, 497–520 (2020).

14. 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).

15. Aoki, S. K. *et al.* A widespread family of polymorphic contact-dependent toxin delivery systems in bacteria. *Nature* **468**, 439–442 (2010).

16. Song, N. *et al.* Genome-wide dissection reveals diverse pathogenic roles of bacterial Tc toxins. *PLoS Pathog.* **17**, e1009102 (2021).

17. Cao, Z., Casabona, M. G., Kneuper, H., Chalmers, J. D. & Palmer, T. The type VII secretion system of Staphylococcus aureus secretes a nuclease toxin that targets competitor bacteria. *Nat Microbiol* **2**, 16183 (2016).

18. Jana, B., Salomon, D. & Bosis, E. A novel class of polymorphic toxins in Bacteroidetes. *Life Sci Alliance* **3**, (2020).

19. Garcia, E. C., Perault, A. I., Marlatt, S. A. & Cotter, P. A. Interbacterial signaling via Burkholderia contact-dependent growth inhibition system proteins. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 8296–8301 (2016).

20. Vassallo, C. N. & Wall, D. Self-identity barcodes encoded by six expansive polymorphic toxin families discriminate kin in myxobacteria. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 24808–24818 (2019).

21. Vassallo, C. N. *et al.* Infectious polymorphic toxins delivered by outer membrane exchange discriminate kin in myxobacteria. *Elife* **6**, (2017).

22. Frankel, A. E. *et al.* Diphtheria toxin fused to human interleukin-3 is toxic to blasts from patients with myeloid leukemias. *Leukemia* **14**, 576–585 (2000).

23. Mok, B. Y. *et al.* A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. *Nature* **583**, 631–637 (2020).

24. Shneider, M. M. *et al.* PAAR-repeat proteins sharpen and diversify the type VI secretion system spike. *Nature* **500**, 350–353 (2013).

25. Bondage, D. D., Lin, J.-S., Ma, L.-S., Kuo, C.-H. & Lai, E.-M. VgrG C terminus confers the type VI effector transport specificity and is required for binding with PAAR and adaptor–effector complex. *Proceedings of the National Academy of Sciences* vol. 113 E3931–E3940 (2016).

26. Geller, A. M. *et al.* The extracellular contractile injection system is enriched in environmental microbes and associates with numerous toxins. *Nat. Commun.* **12**, 3743 (2021).

27. Desvaux, M. *et al.* A conserved extended signal peptide region directs posttranslational protein translocation via a novel mechanism. *Microbiology* **153**, 59–70 (2007).

28. Simmerman, R. F., Dave, A. M. & Bruce, B. D. Structure and function of POTRA domains of Omp85/TPS superfamily. *Int. Rev. Cell Mol. Biol.* **308**, 1–34 (2014).

29. Yamaguchi, Y., Park, J.-H. & Inouye, M. Toxin-antitoxin systems in bacteria and archaea.

Annu. Rev. Genet. **45**, 61–79 (2011).

30. 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).
31. Salomon, D. *et al.* Marker for type VI secretion system effectors. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 9271–9276 (2014).
32. Hachani, A., Allsopp, L. P., Oduko, Y. & Filloux, A. The VgrG Proteins Are 'à la Carte' Delivery Systems for Bacterial Type VI Effectors. *Journal of Biological Chemistry* vol. 289 17872–17884 (2014).
33. Durand, E., Cambillau, C., Cascales, E. & Journet, L. VgrG, Tae, Tle, and beyond: the versatile arsenal of Type VI secretion effectors. *Trends Microbiol.* **22**, 498–507 (2014).
34. Alcoforado Diniz, J., Liu, Y.-C. & Coulthurst, S. J. Molecular weaponry: diverse effectors delivered by the Type VI secretion system. *Cell. Microbiol.* **17**, 1742–1751 (2015).
35. de Moraes, M. H. *et al.* An interbacterial DNA deaminase toxin directly mutagenizes surviving target populations. *Elife* **10**, (2021).
36. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
37. Holm, L. & Laakso, L. M. Dali server update. *Nucleic Acids Res.* **44**, W351–5 (2016).
38. Van Ness, B. G., Howard, J. B. & Bodley, J. W. ADP-ribosylation of elongation factor 2 by diphtheria toxin. Isolation and properties of the novel ribosyl-amino acid and its hydrolysis products. *J. Biol. Chem.* **255**, 10717–10720 (1980).
39. Yahiro, K. *et al.* Cholix toxin, an eukaryotic elongation factor 2 ADP-ribosyltransferase, interacts with Prohibitins and induces apoptosis with mitochondrial dysfunction in human hepatocytes. *Cell. Microbiol.* **21**, e13033 (2019).

40. Capra, J. A. & Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875–1882 (2007).

41. Holliday, G. L., Mitchell, J. B. O. & Thornton, J. M. Understanding the Functional Roles of Amino Acid Residues in Enzyme Catalysis. *Journal of Molecular Biology* vol. 390 560–577 (2009).

42. Hooper, N. M. Families of zinc metalloproteases. *FEBS Lett.* **354**, 1–6 (1994).

43. Bornhorst, J. A. & Falke, J. J. Purification of proteins using polyhistidine affinity tags. *Methods Enzymol.* **326**, 245–254 (2000).

44. Dodson, G. Catalytic triads and their relatives. *Trends in Biochemical Sciences* vol. 23 347–352 (1998).

45. Rauwerdink, A. & Kazlauskas, R. J. How the Same Core Catalytic Machinery Catalyzes 17 Different Reactions: the Serine-Histidine-Aspartate Catalytic Triad of α/β-Hydrolase Fold Enzymes. *ACS Catal.* **5**, 6153–6176 (2015).

46. Kaundal, S., Deep, A., Kaur, G. & Thakur, K. G. Molecular and Biochemical Characterization of YeeF/YezG, a Polymorphic Toxin-Immunity Protein Pair From. *Front. Microbiol.* **11**, 95 (2020).

47. Jamet, A. *et al.* A widespread family of polymorphic toxins encoded by temperate phages. *BMC Biol.* **15**, 75 (2017).

48. Jamet, A. *et al.* Characterization of the Maf family of polymorphic toxins in pathogenic Neisseria species. *Microbial Cell* vol. 2 88–90 (2015).

49. Poole, S. J. *et al.* Identification of Functional Toxin/Immunity Genes Linked to Contact-Dependent Growth Inhibition (CDI) and Rearrangement Hotspot (Rhs) Systems. *PLoS Genetics* vol. 7 e1002217 (2011).

50. 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).

51. Kobayashi, K. Diverse LXG toxin and antitoxin systems specifically mediate intraspecies competition in Bacillus subtilis biofilms. *PLOS Genetics* vol. 17 e1009682 (2021).

52. Ulhuq, F. R. *et al.* A membrane-depolarizing toxin substrate of the type VII secretion system mediates intraspecies competition. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 20836–20847 (2020).

53. Manera, K., Kamal, F., Burkinshaw, B. & Dong, T. G. Essential functions of chaperones and adaptors of protein secretion systems in Gram-negative bacteria. *FEBS J.* (2021) doi:10.1111/febs.16056.

54. Unterweger, D., Kostiuk, B. & Pukatzki, S. Adaptor Proteins of Type VI Secretion System Effectors. *Trends Microbiol.* **25**, 8–10 (2017).

55. Unterweger, D. *et al.* Chimeric adaptor proteins translocate diverse type VI secretion system effectors in Vibrio cholerae. *EMBO J.* **34**, (2015).

56. Ahmad, S. *et al.* Structural basis for effector transmembrane domain recognition by type VI secretion system chaperones. (2020) doi:10.7554/eLife.62816.

57. Wagner, S. *et al.* Bacterial type III secretion systems: a complex device for the delivery of bacterial effector proteins into eukaryotic host cells. *FEMS Microbiol. Lett.* **365**, (2018).

58. Trang H Phan, E. N. G. H. Bacterial secretion chaperones: the mycobacterial type VII case. *FEMS Microbiol. Lett.* **365**, (2018).

59. Castiblanco, L. F., Triplett, L. R. & Sundin, G. W. Regulation of Effector Delivery by Type III Secretion Chaperone Proteins in Erwinia amylovora. *Front. Microbiol.* **0**, (2018).

60. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

61. Jamet, A. & Nassif, X. New players in the toxin field: polymorphic toxin systems in bacteria. *MBio* **6**, e00285–15 (2015).

62. Chen, I.-M. A. *et al.* The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. *Nucleic Acids Research* vol. 49 D751–D763 (2021).

63. 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).

64. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

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

66. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).

67. Gabler, F. *et al.* Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Curr. Protoc. Bioinformatics* **72**, e108 (2020).

68. 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).

69. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).

70. Li, W. *et al.* RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res.* **49**, D1020–D1028 (2021).

71. Puigbò, P., Wolf, Y. I. & Koonin, E. V. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. *J. Biol.* **8**, 59 (2009).

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

73. Brynildsrud, O., Bohlin, J., Scheffer, L. & Eldholm, V. Erratum to: Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* **17**, 262 (2016).

74. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* vol. 57 289–300 (1995).

75. Wheeler, T. J., Clements, J. & Finn, R. D. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics* **15**, 7 (2014).

76. AAAS. https://doi.org/10.1126/science.abj8754.

77. Goddard, T. D. *et al.* UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).

78. Holm, L. & Rosenström, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38**, W545–9 (2010).

79. 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. U. S. A.* **86**, 2172–2175 (1989).

80. Petzoldt, T. Estimate Growth Rates from Experimental Data [R package growthrates version 0.8.2]. (2020).

81. Tollefson, M. Graphics with the ggplot2 Package: An Introduction. *Visualizing Data in R 4* 281–293 (2021) doi:10.1007/978-1-4842-6831-5_7.