

Dear Author

Here are the proofs of your article.

- You can submit your corrections **online, or** via **e-mail**.
- For **online** submission please insert your corrections in the online correction form. Always indicate the line number to which the correction refers.
- You can also insert your corrections in the proof PDF and **email** the annotated PDF.
- Remember to note the **journal title, manuscript number, and your name** when sending your response via e-mail.
- **Check** any questions that have arisen during copy editing or typesetting and insert your answers/corrections.
- **Check** that the text is complete and that all figures, tables and their legends are included. Also check the accuracy of special characters, equations, and additional files if applicable. Substantial changes in content, e.g., new results, corrected values, title and authorship are not allowed without the approval of the responsible editor. In such a case, please contact us for further advice.
- If we do not receive your corrections **within 48 hours**, we will send you a reminder.
- The final versions of your article will be published around one week after receipt of your corrected proofs.

RESEARCH

Open Access

Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics

Dapeng Zhang¹, Robson F de Souza^{1,2}, Vivek Anantharaman¹, Lakshminarayan M Iyer¹ and L Aravind^{1*}

Abstract

Background: Proteinaceous toxins are observed across all levels of inter-organismal and intra-genomic conflicts. These include recently discovered prokaryotic polymorphic toxin systems implicated in intra-specific conflict. They are characterized by a remarkable diversity of C-terminal toxin domains generated by recombination with standalone toxin-coding cassettes. Prior analysis revealed a striking diversity of nuclease and deaminase domains among the toxin modules. We systematically investigated polymorphic toxin systems using comparative genomics, sequence and structure analysis.

Results: Polymorphic toxin systems are distributed across all major bacterial lineages and are delivered by at least eight distinct secretory systems. In addition to type-II, these include type-V, VI, VII (ESX), and the poorly characterized "*Photorhabdus* virulence cassettes (PVC)", PrsW-dependent and MuF phage-capsid-like systems. We present evidence that trafficking of these toxins is often accompanied by autoproteolytic processing catalyzed by HINT, ZU5, PrsW, caspase-like, papain-like, and a novel metallopeptidase associated with the PVC system. We identified over 150 distinct toxin domains in these systems. These span an extraordinary catalytic spectrum to include 23 distinct clades of peptidases, numerous previously unrecognized versions of nucleases and deaminases, ADP-ribosyltransferases, ADP ribosyl cyclases, RelA/SpoT-like nucleotidyltransferases, glycosyltransferases and other enzymes predicted to modify lipids and carbohydrates, and a pore-forming toxin domain. Several of these toxin domains are shared with host-directed effectors of pathogenic bacteria. Over 90 families of immunity proteins might neutralize anywhere between a single to at least 27 distinct types of toxin domains. In some organisms multiple tandem immunity genes or immunity protein domains are organized into polyimmunity loci or polyimmunity proteins. Gene-neighborhood-analysis of polymorphic toxin systems predicts the presence of novel trafficking-related components, and also the organizational logic that allows toxin diversification through recombination. Domain architecture and protein-length analysis revealed that these toxins might be deployed as secreted factors, through directed injection, or via inter-cellular contact facilitated by filamentous structures formed by RHS/YD, filamentous hemagglutinin and other repeats. Phyletic pattern and life-style analysis indicate that polymorphic toxins and polyimmunity loci participate in cooperative behavior and facultative 'cheating' in several ecosystems such as the human oral cavity and soil. Multiple domains from these systems have also been repeatedly transferred to eukaryotes and their viruses, such as the nucleo-cytoplasmic large DNA viruses.

Conclusions: Along with a comprehensive inventory of toxins and immunity proteins, we present several testable predictions regarding active sites and catalytic mechanisms of toxins, their processing and trafficking and their role (Continued on next page)

* Correspondence: aravind@ncbi.nlm.nih.gov

¹National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Full list of author information is available at the end of the article

(Continued from previous page)

in intra-specific and inter-specific interactions between bacteria. These systems provide insights regarding the emergence of key systems at different points in eukaryotic evolution, such as ADP ribosylation, interaction of myosin VI with cargo proteins, mediation of apoptosis, hyphal heteroincompatibility, hedgehog signaling, arthropod toxins, cell-cell interaction molecules like teneurins and different signaling messengers.

Reviewers: This article was reviewed by AM, FE and IZ.

Background

Production and deployment of “chemical armaments” is one of the most common strategies in inter-organismal conflict. Such molecules, namely toxins or antibiotics, are observed at practically every level of biological organization ranging from multicellular organisms like animals and plants, through bacteria, all the way down to intra-genomic selfish elements [1-4]. These molecules span an entire biochemical spectrum from diffusible small molecules (e.g. antibiotics) to some of the largest proteins in the biological world (secreted bacterial toxins) [5,6]. Beyond their natural roles, these molecules have considerable significance as biotechnological reagents, biodefense agents, therapeutic targets, and therapeutics against numerous disease-causing agents [1,2,4,6,7]. Traditional toxicology has now been joined by genomics and sequence analysis in uncovering the enormous biochemical diversity across life forms of such molecules and of the systems that synthesize and traffic them. This diversity is seen both in the structure and action of systems involved in synthesis of diffusible antibiotics and proteinaceous toxins [5,6]. It is becoming increasingly clear that proteinaceous toxins are a common feature of biological conflicts at every organizational level [7]: 1) In antagonistic interactions between different multicellular eukaryotes, such as the castor bean ricin, *Aspergillus* sarcin and various snake venom proteins [2,3,8,9]. 2) Action by multicellular organisms against their pathogens (e.g. anti-microbial peptide toxins and defensive RNases such as RNaseA and RNase L [10-13]). 3) Action of pathogenic and symbiotic bacteria directed against their hosts (e.g. the cholera toxin and the shiga toxin [4,14]). 4) Inter-specific conflict in bacteria [15]. 5) Conflict between bacterial sibling strains of the same species, namely contact dependent inhibition systems and related secreted toxins [16-19]. 6) Inter-genomic conflicts between cellular genomes and selfish replicons residing in the same cell (e.g. classical bacteriocins and plasmid addiction toxins [20]). 7) Intra-genomic conflicts between selfish elements and the host genome (restriction-modification systems [21] and genomic toxin-antitoxin systems [22-24]).

Studies in the past decade are pointing to certain unifying themes across the proteinaceous toxins deployed in each of these distinct types of biological conflict. The most prominent theme is the use of enzymatic toxins

that disrupt the flow of biological information by targeting nucleic acids and proteins [7]. Thus, several toxin domains are nucleases targeting genomic DNA, tRNAs and rRNAs, nucleic acid base glycosylases, nucleic acid-modifying enzymes, peptidases that cleave key protein targets, and protein-modifying enzymes that alter the properties of proteins, such as components of the translation apparatus [4,6,7,17,18,25]. A secondary theme seen across toxins from phylogenetically diverse sources is the presence of domains that disrupt cellular integrity by forming pores in cellular membranes [26,27]. Genomic analysis has also revealed that the richest source of proteinaceous toxins is the bacterial superkingdom, wherein several systems involved in most of the levels of biological conflict enumerated above are encountered [4,6,17,18,21,22,25].

It is also becoming apparent that inter- and intra-specific and inter- and intra-genomic conflicts in prokaryotes has resulted in an intense arms race with respect to proteinaceous toxins. There is evidence for multiple episodes of escalation of the conflict in terms of the evolution of immunity proteins, followed by alterations in the toxins to evade the action of the immunity proteins [15,17,18,24,28]. Another major evolutionary theme seen in secreted proteinaceous toxins is the exploration of several alternative secretory mechanisms for their effective trafficking and delivery to potential targets. In particular, bacteria display at least eight distinct secretory mechanisms over and beyond the ancestral Sec (or Type II) system that is shared with the other branches of life (Table 1). Both the T2SS and alternative secretory mechanisms have been repeatedly coopted for trafficking toxins [15,17,18,29,30]. In addition to the T2SS, examples of other widely utilized secretory pathways that have been frequently coopted for trafficking of toxins include three distinct systems dependent on ATPase pumps: 1) ABC ATPase-dependent Type I system, which has been adapted for the delivery of the large RTX toxins [31]; 2) the FtsK-like ATPase-dependent type VII (ESX) system of Gram-positive bacteria, which has been recruited for delivering several toxins, including those frequently deployed in intraspecific conflict [17,32,33]; 3) the plasmid conjugation apparatus-derived type IV system [34], which is also dependent on FtsK-related ATPases [33]. On the other hand some of the other alternative

t1.1 **Table 1 Features of secretion pathways by which polymorphic toxins are exported**

t1.2	Secretion pathway	Signature N-terminal leader domains or pre-toxin-domains	Signature genes in neighborhood	Processing proteases/repeats in toxin proteins	Phyletic patterns	Additional Notes
t1.3 t1.4 t1.5 t1.6 t1.7	T2SS/Sec-dependent system	Signal peptide	-	<i>Proteases:</i> Caspase, HINT, MCF-SHE, subtilisin ³ , ZU5 ⁴ <i>Repeats:</i> ALF, ankyrins, β -propeller, RHS, Sel1 ¹ , TPR ¹ , Tail-fiber ²	In all bacteria	Default pathway for protein export. Might contain MAFB-N (DUF1020), MicroscillaN, APD1, APD2, Inactive transglutaminase
t1.8 t1.9 t1.10 t1.11	T5SS	N-terminal TpsA-like secretion domain (TPSASD) <i>Pre-toxin domains:</i> DUF637(PT637), DUF637-N, PT-VENN	FhaB/CdiB coding for porin-like protein	<i>Proteases:</i> HINT <i>Repeats:</i> FilH	$\alpha, \beta, \gamma, \delta, \epsilon$ -proteobacteria, acidobacteria, bacteroidetes/chlorobi, firmicutes ⁵ , fusobacteria	The TPSASD domain binds the outer-membrane FhaB/CdiB during the export of the toxin domain
t1.12	T6SS	VgrG domain, PAAR domain, Hcp1	ClpV-like AAA + Atpase, MOG1/PspB-like, VgrG, Hcp1, Phage tail/base-plate related proteins	<i>Repeats:</i> RHS	All proteobacteria, acidobacteria, bacteroidetes/chlorobi, firmicutes	Complete T6SS delivered toxins are often typified by a N-terminal PAAR domain
t1.13 t1.14 t1.15	<i>Photorhabdus</i> virulence cassette pathway (PVC)	PVC-Metallopeptidase	CDC48-like AAA + ATPase, VgrG, Phage tail/base-plate related proteins	<i>Proteases:</i> Metallopeptidase, Subtilisin, Caspase, MCF-SHE <i>Repeats:</i> RHS, tail fiber	Euryarchaeota, $\alpha, \beta, \gamma, \delta, \epsilon$ -proteobacteria, acidobacteria, actinobacteria, bacteroidetes, chlorobi, chloroflexi, cyanobacteria, deinococci, firmicutes, nitrospirae, spirochaetes	
t1.16 t1.17	T7SS/ESX/ESAT-6 secretion system	WxG, LxG, LDxD domains	YueA-like FtsK/HerA ATPase, EsaC	<i>Proteases:</i> HINT, Caspase, MCF-SHE <i>Repeats:</i> RHS, Tail-fiber	Firmicutes, actinobacteria, chloroflexi, other bacterial lineages ⁶	Toxins exported by these systems may or may not possess repeat domains
t1.18 t1.19 t1.20 t1.21	TcdB/TcaC	A signal peptide followed by a SpvB domain coupled to a C-terminal integrin-like β -propeller domain	TcdB	Repeats: Integrin-like beta propeller, RHS, tail-fiber Proteases: HINT, Caspase, ZU5	Euryarchaeota, $\alpha, \beta, \gamma, \delta$ -proteobacteria, actinobacteria, bacteroidetes Chloroflexi, fibrobacteres, firmicutes, lentisphaerae, spirochaetes	

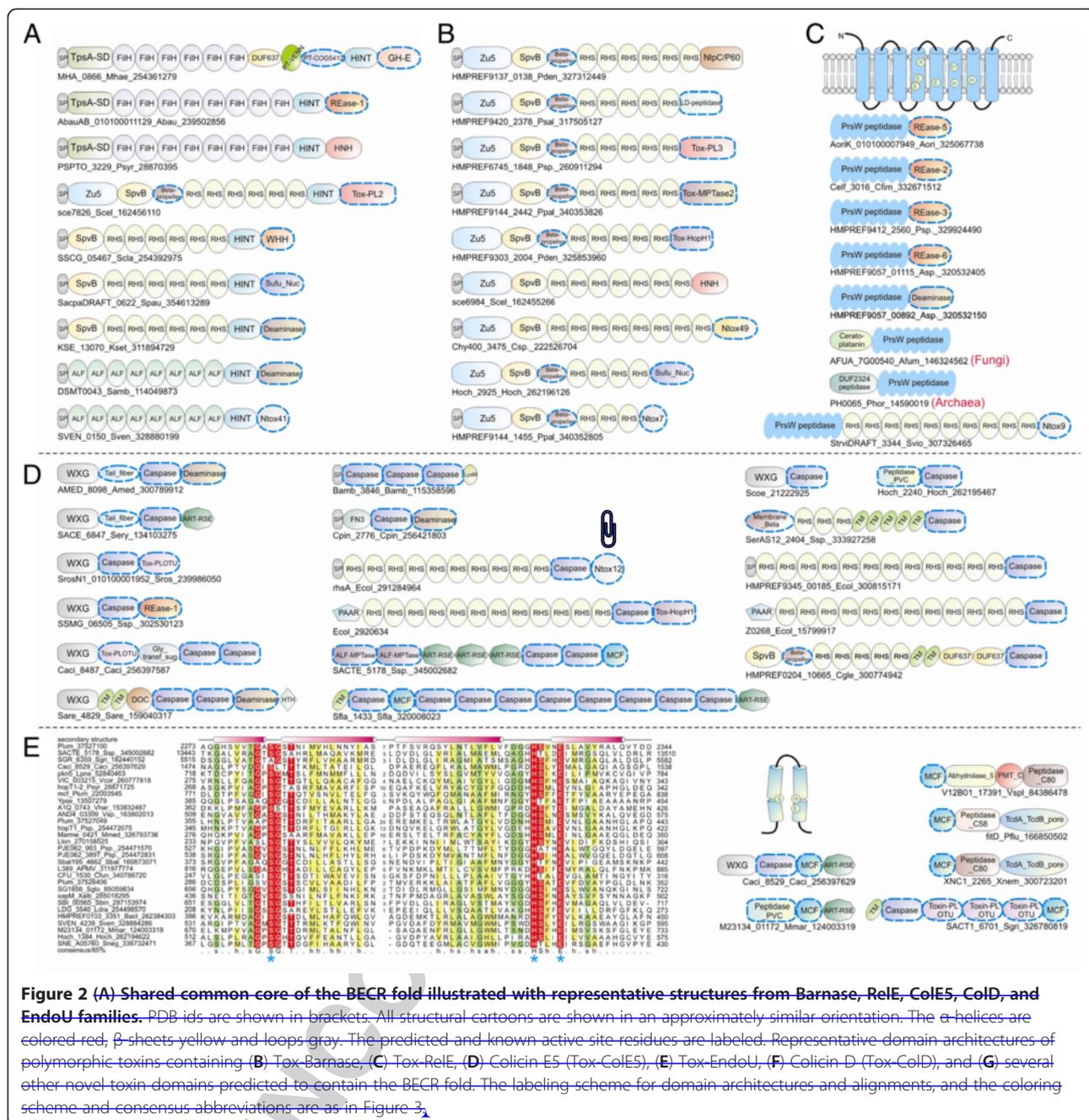
Table 1 Features of secretion pathways by which polymorphic toxins are exported (Continued)

	PrsW-peptidase domain	PrsW	Repeats: RHS	Proteases: PrsW	PrsW is a transmembrane peptidase with several transmembrane helices
t1.23					
t1.24					
t1.25	Phage DNA packaging system	MuF			The toxin is predicted to be packaged into the phage head as in phage transduction systems
t1.26		MuF, large and small subunits of terminase	Papain-like		
t1.27					
t1.28					
t1.29					
t1.30					

1: Note only fused to toxins exported by the SEC-dependent pathway in *Amoebophilus asiaticus*; 2: Note only fused to toxins exported by the SEC-dependent pathway in *Microcilla marini*; 3: Note only fused to toxins exported by the SEC-dependent pathway in *Acerivibrio cellulolyticus*; 4: Note only fused to toxins exported by the SEC-dependent pathway in *Caldicellulosiruptor* species; 5: Note in firmicutes, the export pathway is only present in *Veillonella* and *Selenomonas* species, also referred to as the Negativicutes species; 6: Certain bacterial lineages within the β - γ -proteobacteria, planctomycetes, verrucomicrobia, cyanobacteria and bacteroidetes have solo WXG domains that have a distinct YueA-like ATPase with 3 HerA/FtsK domains of which only the middle one is active. These appear to be mobile versions of T7SS.

secretory mechanisms appear to be primarily utilized in trafficking toxins rather than any other function: 1) The type III system based on the flagellar basal body-like apparatus [35]; 2) the two-partner or Type V system which resembles the porins [36,37]; 2) the type VI [38,39]; 3) *Photorhabdus* virulence cassette (PVC)-type secretory system [40,41]. Both T6SS and the PVC-SS utilize caudate bacteriophage tail-derived proteins as an “injection syringe” and distinct AAA + ATPases to recycle the injection apparatus in an ATP-dependent manner after a single use [39]; 4) TcdB/TcaC-like export pathway [42]; 4) the PrsW-like peptidase-dependent system export system [43]. Depending on the secretory pathway, toxins might either be directly injected into target cells (e.g. T6SS delivered toxins) or diffuse into the surrounding medium (e.g. certain T2SS or T7SS toxins) or be anchored on the surface of producing cells to be delivered upon contact with the target cell (e.g. T5SS and certain T2SS, T6SS and T7SS delivered toxins). Additionally, these prokaryotic toxins might also display further adaptations that allow their processing subsequent to their secretion – these include the presence of “pre-toxin domains” that might be sites for proteolytic processing or in-built peptidase domains that cleave off the toxin domain to facilitate its delivery into the target cell [17,20] (Table 1).

The selective pressures related to the above-described adaptations for trafficking, processing and delivery appear to have been instrumental in shaping the domain architectures of plasmid-encoded bacteriocins and prokaryotic toxins deployed in inter- and intra-specific conflicts [17,20]. Consequently, most toxin proteins have N-terminal domains involved in secretion and/or cell surface anchorage, central domains involved in adhesion or presentation to target cells and C-terminal domains that bear the actual toxin activity (Figure 1, Table 1). These might be occasionally combined with further processing-peptidase or pre-toxin domains [17,18,20]. These stereotypic architectural features strongly distinguish such toxins from those involved in intra-genomic conflicts, such as those from classical toxin-antitoxin systems and restriction-modification systems, even though certain domains with toxin activity might be common across these different systems [17,22,28]. Hence, domain architectural analysis considerably aids in the detection of new toxins involved in inter-organismal conflicts and the delineation of specific domains associated with each of the above-listed trafficking related roles. This has led to an exciting discovery in the past two years, namely the identification and characterization of an extremely widespread system of secreted toxins, primarily involved in intra-specific conflict between related strains of prokaryotes [16-19]. These toxin systems are found in practically all major bacterial lineages and also a small number



220 action, such as protein AMP/UMPylating enzymes,
 221 ADP-ribosyltransferases and peptidases. Interestingly,
 222 we observed that several of the toxin and processing
 223 peptidase domains from polymorphic secreted toxins are
 224 also present in toxin domains of conventional toxins
 225 deployed in inter-specific conflict, such as against
 226 eukaryotic hosts by pathogenic or symbiotic bacteria
 227 [46-54]. In a similar vein, we also observed that both the
 228 polymorphic toxins deployed in intra-specific conflicts
 229 and toxins used in inter-specific conflict often rely on
 230 similar secretory mechanisms, such as the T5SS, T6SS

and T7SS [17,18]. These observations suggested that
 both types of secreted toxins have been “constructed” in
 course of evolution from a common pool of domains
 and consequently possess similarities in their domain
 architectures. We also observed that several domains
 present in secreted prokaryotic toxins and their immunity
 proteins have been transferred to eukaryotes and their
 viruses, and have contributed to the provenance of
 major regulatory molecules in the development of multi-
 cellular animals, RNA editing, DNA mutagenesis and
 virus-host interactions [17,18]. Thus, the evolutionary

242 and functional significance of domains found in prokary-
243 otic toxin systems extends beyond the mechanisms and
244 dynamics of intra-organismal conflict.

245 Our previous studies on the polymorphic toxins fo-
246 cused on identifying and characterizing the diversity of
247 toxin domains that operate on nucleic acids, in particular
248 nucleases and deaminases, and characterizing some of
249 the most prevalent immunity proteins, such as those with
250 the SUKH and SuFu domains. We also reported a pre-
251 liminary characterization of the major secretory systems
252 involved in toxin trafficking and processing peptidases.
253 Here, we build on our previous studies to systematically
254 characterize novel domains in polymorphic toxin sys-
255 tems, with a particular focus on those involved in toxin
256 activity, immunity and maturation of toxins. Conse-
257 quently, we report herein a greatly expanded repertoire
258 of toxin domains and immunity proteins directed against
259 them. Thus, we also considerably extend their structural
260 and mechanistic diversity to include a diverse array of
261 peptidases, ADP ribosyltransferases, glycosyltransferases,
262 kinases, membrane perforators and domains with several
263 other activities. Even in terms of toxin acting on nucleic
264 acids we report numerous previously unrecognized
265 nucleases and deaminases. This expanded repertoire of
266 toxin domains also helps better understand the common-
267 alities between the polymorphic toxin systems and the
268 classical secreted toxins deployed against distantly
269 related organisms. This comprehensive characterization
270 also provides a handle to investigate the ecological sig-
271 nificance of such secreted toxin systems in prokaryotes.
272 Our analysis also uncovered novel features regarding the
273 secretory systems that traffic these toxins. The detailed
274 analysis of these toxins systems and their immunity pro-
275 teins also pointed to several additional examples of
276 domains from them being acquired by eukaryotes and
277 their viruses. Thereby we greatly widen the contributions
278 of components of these systems to the evolution of sev-
279 eral eukaryotic regulatory systems. We present a compre-
280 hensive inventory of intra-specific polymorphic toxin
281 systems and related components from toxin systems
282 deployed in inter-specific conflicts. This database is likely
283 to serve as a useful reference for future studies on this
284 enormously significant group of proteins.

285 Results and discussion

286 Search strategy to identify new toxins and immunity 287 proteins

288 In order to identify novel polymorphic toxins we adopted
289 a strategy of matching diagnostic domain-architecture
290 and gene-neighborhood templates, similar to what we
291 had done earlier to identify novel type II toxin-antitoxin
292 systems [22]. In the case of polymorphic toxins the do-
293 main architecture template is defined by the presence of
294 multi-domain proteins, wherein the C-terminal-most

domain has toxin activity, while the N-terminal-most 295
domains are associated with trafficking (Table 1, Figure 1). 296
The central domains might be involved in adhesion, pres- 297
entation or processing. One of the most common features 298
of this central region is the presence of RHS (Recombina- 299
tion hot spot)/YD or filamentous hemagglutinin (FilH) 300
repeats which form extended fibrous or filamentous 301
structures that help in displaying the C-terminal toxin 302
domain in the cell-surface [17,18,37,45,55,56]. With the 303
above domain-architecture template (Figure 1), we identi- 304
fied an initial set of exemplars, which were used in se- 305
quence similarity searches to identify homologs that were 306
similar over most of their length but differing in their 307
C-terminal-most domains – a hallmark of polymorphic 308
toxins (Figure 1B). This enabled us to precisely define the 309
boundaries of the C-terminal toxin domains and use 310
them as seeds in iterative sequence profile searches with 311
the PSI-BLAST and JACKHMMER programs. These 312
searches allowed us to recover both standalone toxin do- 313
main cassettes and examples where they are combined 314
with other types of N-terminal trafficking, presentation 315
and processing domains, distinct from those found in the 316
starting queries. This process was used transitively to de- 317
tect further toxin domains and full length toxins. As a 318
result, we were able to not only capture other polymorphic 319
toxins but also identify cases where these toxin domains 320
might be used as the active domains of other secreted 321
toxins that are deployed against more distantly related 322
organisms (e.g. T3SS or T4SS delivered host-directed 323
toxins). To further understand the sequence and struc- 324
ture affinities of toxin domains, we also used their mul- 325
tiple alignments in profile-profile comparisons with the 326
HHpred program to recover distant homologs and deter- 327
mine their protein fold. Additionally, detailed domain- 328
architecture analysis of the associated domains in the 329
case of the full length toxins allowed us to delineate the 330
domains involved in the other processes mentioned 331
above. 332

In terms of gene-neighborhood templates (Figure 1), 333
we exploited the fact that the polymorphic toxin genes 334
are accompanied by several solo toxin cassettes and 335
genes for immunity proteins and in some cases genes en- 336
coding trafficking components (e.g. T6SS or PVC-SS). 337
Hence, we systematically extracted the genomic neigh- 338
borhoods for all detected toxin-encoding genes from 339
complete genome sequences or assembled CONTIGs 340
and subjected them to gene-neighborhood analysis. 341
Matches to the above template allowed us to distinguish 342
the classical polymorphic toxins from related toxin sys- 343
tems that are deployed against more distantly related 344
organisms. A combination of the gene-neighborhood 345
analysis with the domain architecture analysis also 346
allowed us to determine the trafficking mechanisms of 347
full-length toxins in the majority of cases. Further, this 348

349 genomic analysis also led to the recovery of the potential
350 immunity proteins associated with the polymorphic tox-
351 ins. The identification of novel immunity proteins uti-
352 lized the fact that the immunity protein gene/s are
353 invariably adjacent to the toxin gene in an operon and
354 typically encode a small single domain protein (Figure 1).
355 We confirmed novel immunity proteins by initiating se-
356 quence searches with them and using the newly detected
357 homologs in gene-neighborhood analysis to check if they
358 showed any co-occurrence with toxin genes. The gene-
359 neighborhood analysis of the newly identified immunity
360 proteins also helped recover any loci that might have
361 been missed in the initial toxin-centric analysis and also
362 pointed to certain novel types of loci comprised primarily
363 of multiple immunity genes (See below).

T3T2

364 As a result of the above searches, we were able to
365 assemble a comprehensive inventory of toxins and im-
366 munity proteins, which we provide as a resource accom-
367 panying this article (Table 2, 3 and Additional File 1).
368 For the sake of systematic nomenclature we adopted the
369 following convention: 1) The toxin domains are labeled
370 'Tox' followed by the name of the superfamily they
371 belong to. Thus, a toxin domain of the restriction endo-
372 nuclease (REase) superfamily would be labeled Tox-
373 REase. 2) The domain might be further distinguished by
374 a numeral if there are multiple distinct toxin families
375 within a given superfamily, e.g. ~~Tox-REase1~~, ~~Tox-REase2~~,
376 and so on. 3) In the case of certain highly divergent fam-
377 ilies, each with their own structurally distinct features,
378 such as those belonging to the HNH/EndoVII nuclease
379 fold, each family of toxin domains might receive a separ-
380 ate label, e.g., ~~Tox-classical~~-HNH, Tox-AHH, Tox-LHH
381 or Tox-NucA that identifies the specific family of
382 nucleases. 4) Novel toxins that could not be unified with
383 any previously known superfamily are labeled as 'Ntox'
384 followed by a number, e.g. Ntox1, Ntox2 etc. (we identi-
385 fied a total of 50 such novel, monophyletic toxin groups
386 in this study). 5) The immunity proteins were similarly
387 named according to their superfamily. Thus, immunity
388 proteins of the SUKH, SuFu and LRR superfamilies are
389 respectively labeled as Imm-SUKH, Imm-SUFU or Imm-
390 LRR. 6) Novel immunity proteins that could not be
391 unified with any known superfamily were labeled as
392 Imm followed by a number, e.g. Imm1, Imm2 etc. (we
393 detected 73 such immunity proteins in this work).

394 In the initial section we present the results of the
395 above analysis from a domain-centric viewpoint by lay-
396 ing out the main conserved domains we identified in
397 toxins (Table 2), immunity proteins (Table 3) and some
398 novel features associated with trafficking (Table 1). In
399 course of discussing the conserved domain families, we
400 describe key features relating to their domain architec-
401 tures and gene-neighborhoods, and present the relevant
402 functional inferences derived from them. In the

following sections we explore the general features of the 403
domain architecture and gene-neighborhood networks, 404
phyletic distribution, relationships between various pro- 405
teinaceous toxin systems, ecological implications and the 406
evolutionary connections between components of these 407
toxin systems and eukaryotic and viral functional 408
systems. 409

Peptidase domains in polymorphic toxins and 410 related proteins 411

Peptidase domains from these systems can be function- 412
ally categorized into ~~those that are~~: 1) involved primarily 413
in processing toxin proteins; 2) those that function both 414
in processing and as toxins; 3) those that function 415
mainly as toxins. Autoproteolytic processing by diverse 416
peptidases has been long recognized in classical secreted 417
toxins deployed by pathogenic bacteria against their 418
hosts [49,51,54]. For example, the *Vibrio cholera* RTXA 419
peptide ligase toxin, clostridial glucosyltransferase toxins 420
and certain *Yersinia* toxins are autoproteolytically pro- 421
cessed by intrinsic caspase-like thiol peptidase domains, 422
which are induced by small molecules such as GTP and 423
inositol hexakisphosphate in the host cytoplasm 424
[49,52,57]. Similarly, we presented evidence that the 425
HINT autopeptidase domains are likely to be an import- 426
ant player in the autoproteolytic release of several poly- 427
morphic toxins (Figure 3A) [17]. In toxins of several 428
pathogens, peptidase domains have also been character- 429
ized bearing the actual toxin activity. Examples include 430
the *Yersinia pestis* YopT papain-like peptidase domain 431
that triggers actin depolymerization in host cells by 432
cleaving the C-termini of Rho GTPases [50] and the *Ba- 433*
cillus anthracis lethal factor that disrupts signaling cas- 434
cades by cleaving the N-termini of several MAPK kinase 435
[48]. However, to date peptidase domains have not been 436
systematically characterized in classical polymorphic 437
toxin systems. In polymorphic toxins, peptidases acting 438
in either of the above three functional categories can be 439
distinguished mainly based on their location within the 440
polypeptide. Those involved in autoproteolytic proces- 441
sing are mostly located either at the N-terminus or prior 442
to the C-terminal toxin domain in the multi-domain 443
toxin proteins (Figure 1). The toxin versions invariably 444
occur at the C-termini. Those which might occur at both 445
of these locations can be inferred as functioning as ei- 446
ther toxins or processing proteins depending on their 447
position in the polypeptide. In addition to these categor- 448
ies, there are inactive peptidase domains that might 449
serve as peptide-binding modules involved in anchorage 450
and interactions of toxins. We discuss below the previ- 451
ously unrecognized peptidase domains that we identified 452
in polymorphic toxin systems and also discuss their con- 453
nections to related peptidase domains in other toxin sys- 454
tems (Table 2). 455

F3

t2.1 **Table 2 Phyletic distribution, export pathways, and contextually-associated domains and proteins of polymorphic toxin domains**

t2.2	Toxin ¹	Fold; conserved residues or motifs ² and additional notes	Phyletic spread ³	Export pathway ⁴	Immunity proteins	Repeats/processing Proteases
t2.3	DNase toxins					
t2.4	Tox-NucA	HNH/EndoVII fold; GH, N, N, E	Actinobacteria, $\alpha,\beta,\gamma,\delta$ -proteobacteria, bacteroidetes, chloroflexi, firmicutes, spirochaetes, verrucomicrobia	T2SS, T5SS, T6SS, T7SS (WXG,LXG,LDXD), PVC	Imm36, Imm-SUKH, Imm-NTF2	<i>Proteases:</i> PVC-Metallopeptidase, Caspase; <i>Repeats:</i> FilH, RHS, Tail-fiber
t2.5	Tox-ColE7	HNH/EndoVII fold (PDB: 1zns);HH, H, H	Bacteroidetes, $\alpha,\gamma,\delta,\epsilon$ -proteobacteria, firmicutes	T2SS, T5SS, T6SS, T7SS (WXG,LXG), PyocinS	Imm-ColE7, Imm-SUKH	<i>Repeats:</i> FilH, RHS
t2.6	Tox-HNH (including Tox-HNH-CIDE)	HNH/EndoVII fold; A DHxxE characterizes the Tox-HNH-CIDE clade.	Acidobacteria, actinobacteria, bacteroidetes, chlorobi, firmicutes, proteobacteria, Eukaryotes:metazoa	T2SS, T5SS, T7SS (WXG,LXG, LDXD), PVC, TcdB/TcaC	Imm-SUKH, Imm-SuFu, Imm14, Imm18, Imm24, Imm33,	<i>Proteases:</i> PVC-Metallopeptidase, HINT, Tox-PLOTU, ZU5; <i>Repeats:</i> FilH, RHS
t2.7						
t2.8	Tox-AHH	HNH/EndoVII fold; [AG]HH, N, H, H, Y motif and residues	Actinobacteria, $\alpha,\beta,\gamma,\delta,\epsilon$ -proteobacteria, bacteroidetes, cyanobacteria, firmicutes, fusobacteria, lentisphaerae, planctomycetes, spirochaetes, verrucomicrobia, eukaryotes: hexapoda, Viruses: Ostreococcus lucimarinus virus, Bathycoccus sp. RCC1105 virus	T2SS, T5SS, T6SS, T7SS (LXG, WXG, LDxD), TcdB/TcaC	Imm-PA2201, Imm-ank, Imm11, Imm20, Imm23, Imm24, Imm43	<i>Proteases:</i> HINT; <i>Repeats:</i> RHS, FilH
t2.9	Tox-DHNNK	HNH/EndoVII fold; DH, N, N, N, K motif and residues	Acidobacteria, actinobacteria $\alpha,\beta,\gamma,\delta,\epsilon$ -proteobacteria, firmicutes, fusobacteria, planctomycetes, spirochaetes, <i>archaea:</i> euryarchaeota, eukaryotes: fungi(ascmycota, basidiomycota)	T2SS, T6SS, T7SS (LXG, LDXD,WXG), PVC	Imm-SUKH, Imm-SuFu, Imm33	<i>Proteases:</i> PVC-Metallopeptidase, HINT
t2.10	Tox-EHHH	HNH/EndoVII fold; [ED]H, H, H	Actinobacteria, bacteroidetes, β,γ,δ -proteobacteria, firmicutes	T2SS, T5SS T6SS, T7SS (WXG, LxG), TcdB/TcaC	Imm8, Imm50	<i>Repeats:</i> FilH, RHS
t2.11	Tox-GH-E	HNH/EndoVII fold; GH, E, N, E motif and residues	Actinobacteria, bacteroidetes, $\beta,\gamma,\delta,\epsilon$ -proteobacteria, chloroflexi, firmicutes, planctomycete, spirochaetes, <i>archaea:</i> euryarchaeota	T2SS (MafBN), T5SS, T6SS, T7SS (WXG, LxG, LDXD), PVC	Imm-SuFu, Imm-ank	<i>Proteases:</i> HINT, PVC-Metallopeptidase; <i>Repeats:</i> RHS, FilH, Tail Fiber
t2.12	Tox-GHH	HNH/EndoVII fold; WxxE, W, G[HO]H, NixF, [DE]H; Eukaryotic versions lack the conserved histidines and a C-terminal helix	Acidobacteria, bacteroidetes, firmicutes, γ -proteobacteria, planctomycete, eukaryotes: metazoa	T2SS, T6SS, T7SS (LXG), TcdB/TcaC	Imm-SUKH	<i>Repeats:</i> RHS
t2.13	Tox-GHH2	HNH/EndoVII fold; s[AGP]HH, HxxxH	β,γ -proteobacteria, bacteroidetes, firmicutes	T2SS, T6SS	-	<i>Repeats:</i> RHS
t2.14	Tox-HHH	HNH/EndoVII fold; N, s[GD]xxR, HHH, H	Actinobacteria, bacteroidetes, γ -proteobacteria, firmicutes	T2SS, T5SS,T6SS, T7SS (LXG,LDXD), PVC	Imm-SUKH	<i>Proteases:</i> PVC-Metallopeptidase; <i>Repeats:</i> FilH, RHS
t2.15	Tox-LHH	HNH/EndoVII fold; N, LHH, E, H, H, W	Actinobacteria, $\alpha,\beta,\gamma,\delta,\epsilon$ -proteobacteria, bacteroidetes, firmicutes, fusobacteria, planctomycetes	T2SS, T5SS, T6SS, T7SS (WXG,LXG), PVC	Imm-SUKH	<i>Proteases:</i> PVC-Metallopeptidase,

Table 2 Phyletic distribution, export pathways, and contextually-associated domains and proteins of polymorphic toxin domains (Continued)

t2.16	Tox-SHH	HNH/EndoVII fold; N, LHH, E, H, H, R motif and residues	Actinobacteria, $\alpha, \beta, \gamma, \delta$ -proteobacteria, bacteroidetes, cyanobacteria, firmicutes, planctomycetes, eukaryotes: crustacea, viruses: caudovirales	T2SS, T5SS, T6SS, T7SS (LDXD, LXG, WXG)	Imm-SUKH, Imm11, Imm24, Imm30, Imm55	HINT; Repeats: FilH, RHS, Tail-fiber Proteases: HINT Repeats: FilH, RHS, ALF
t2.17	NGO1392-like	HNH/EndoVII fold; CxxC, DH, CXXC, Q	Actinobacteria, $\alpha, \beta, \gamma, \delta$ -proteobacteria, chlorobi, chloroflexi, cyanobacteria, firmicutes, spirochaetes, eukaryotes: alveolata(apicomplexa), choanoflagellida, metazoa, stramenopiles, viridiplantae, Viruses: several Mycobacteriophages, caudovirales	T2SS (MafBN), T5SS, TcdB/TcaC, PVC	Imm-SuFu, Imm13, Imm21, Imm33, Imm38	Proteases: HINT, PVC-Metallopeptidase, ZU5; Repeats: FilH, RHS, Tail fiber
t2.18	(Also known as					
t2.19	Tox-SuFu-Nuc)					
t2.20	Tox-WHH	HNH/EndoVII fold; WHH, L, H, HxG	Actinobacteria, $\alpha, \beta, \gamma, \delta, \epsilon$ -proteobacteria, bacteroidetes, chloroflexi, firmicutes, fusobacteria, planctomycete, synergistetes	T2SS, T5SS, T6SS, T7SS (WXG, LXG, LDXD), PVC, TcdB/TcaC	Imm-SUKH, Imm28, Imm37	Proteases: HINT, PVC-Metallopeptidase; Repeats: RHS, ALF, FilH
t2.21	Tox-REase-1	Restriction endonuclease fold; E, D, ExK, Q	Actinobacteria, bacteroidetes, β, γ, ϵ -proteobacteria, cyanobacteria, fusobacteria, firmicutes, gemmatimonadetes, planctomycetes, eukaryotes: alveolata, heterolobosea	T2SS, T5SS, T6SS, T7S (WXG, LXG), TcdB/TcaC	Imm-PA2201, Imm49	Proteases: HINT, Caspase, ZU5; Repeats: FilH, RHS, Tail-fiber
t2.22	Tox-REase-2	Restriction endonuclease fold; E, DG, [DE]xK, T, W	Actinobacteria	T2SS, T7SS (WXG), PrsW	-	Proteases: PrsW-peptidase
t2.23	Tox-REase-3	Restriction endonuclease fold; [KR]ExD, K, ExQxK	β, γ -proteobacteria, firmicutes	T2SS (MafBN), T6SS, T7SS (WXG), PrsW	Imm-SUKH, Imm7	Proteases: PrsW-peptidase; Repeats: RHS
t2.24	Tox-REase-4	Restriction endonuclease fold; D, ExK	Actinobacteria, $\alpha, \beta, \gamma, \delta$ -proteobacteria, bacteroidetes, cyanobacteria, firmicutes, planctomycetes, spirochaetes, eukaryotes: stramenopiles	T2SS, T5SS, T6SS, T7SS (WXG, LDXD), PrsW	Imm-SUKH, Imm22, Imm54	Proteases: PrsW-peptidase; HINT; Repeats: FilH, RHS, Tail fiber
t2.25	Tox-REase-5	Restriction endonuclease fold; Y, FDG, EAK, Y, Q, W	Actinobacteria, $\alpha, \beta, \gamma, \delta$ -proteobacteria, firmicutes, fusobacteria, Viruses: caudovirales	T2SS, T5SS, T6SS, PrsW	Imm52	Proteases: PrsW-peptidase; Repeats: FilH, RHS
t2.26	Tox-REase-6	Restriction endonuclease fold; E, D, ExK, Q, Y	Actinobacteria, α, β, γ -proteobacteria, bacteroidetes, cyanobacteria, firmicutes, eukaryotes: heterolobosea	T2SS, T5SS, T6SS, T7SS (WXG), PrsW	Imm49	Proteases: PrsW-peptidase; Repeats: RHS, Tail fiber
t2.27	Tox-REase-7	Restriction endonuclease fold; GxxxE, lxD, ExK, Q	Actinobacteria, α, γ, ϵ -proteobacteria, bacteroidetes, cyanobacteria, firmicutes, planctomycetes, verrucomicrobia	T2SS, T5SS, T6SS, T7SS (WXG)	ImmHEAT, Imm23, Imm54	Proteases: HINT; Repeats: FilH, RHS, Tail-fiber
t2.28	Tox-REase-8	Restriction endonuclease fold; GxxxQ, DD, QxK	Actinobacteria, $\alpha, \beta, \gamma, \delta$ -proteobacteria, bacteroidetes, chlorobi, chloroflexi, firmicutes, spirochaetes, verrucomicrobia, eukaryotes: metazoa(crustacea, hexapoda, placoza)	T2SS (APD1)	-	Repeats: Ankyrin repeats, TPR repeats, RHS
t2.29	Tox-Rease-9	Restriction endonuclease fold; GxxxH, E, D, ELKP, YxxE	Actinobacteria, γ -proteobacteria, bacteroidetes, chlamydiae, firmicutes	T2SS, T7SS (LxG)	Imm54	Proteases: HINT; Repeats: RHS

Table 2 Phyletic distribution, export pathways, and contextually-associated domains and proteins of polymorphic toxin domains (Continued)

t2.30	Tox-Rease-10	Restriction endonuclease fold; E, Q, [DE], ExKNY, R, DxRG	β, γ, ϵ -proteobacteria, firmicutes, fusobacteria, spirochaetes	T2SS, T5SS, T7SS (WXG, LXG),	Imm54, Imm70	Repeats: FilH
t2.31	Tox-URI1	URI nuclease fold; Y, YxG, R, [RK]xxE, N	Actinobacteria, $\alpha, \beta, \gamma, \delta$ -proteobacteria, bacteroidetes, chlamydiae, chloroflexi, firmicutes, lentisphaerae, nitrospirae, verrucomicrobia, <i>archaea</i> : euryarchaeota, viruses: <i>Ostreococcus lucimarinus</i> virus, eukaryotes: fungi	T2SS, T5SS, T6SS, TcdB/TcaC	Imm14, Imm26, Imm44, Imm51	Proteases: HINT; Repeats: RHS, FilH, Tail fiber
t2.32	Tox-URI2	URI nuclease fold; Y, KxG, [EQ]	Actinobacteria, α, β, γ -proteobacteria, bacteroidetes, firmicutes	T2SS, T6SS	Imm9, Imm39, Imm12, Imm44	Proteases: HINT; Repeats: RHS, Tail fiber
t2.33	RNase toxins of known fold					
t2.34	Tox-Barnase	Barnase-EndoU-ColicinE5/D-ReLE like nuclease (BECR) fold ($\alpha + \beta$); H, H, [ST], FP, [STD]	Actinobacteria, bacteroidetes, $\beta, \gamma, \delta, \epsilon$ -proteobacteria, chlamydiae, chloroflexi, cyanobacteria, deinococci, fibrobacteres, firmicutes, fusobacteria, nitrospirae, planctomycetes <i>archaea</i> : euryarchaeota	T2SS, T6SS, T7SS (WXG), TcdB/TcaC, MuF, PVC	Imm-Barstar	Proteases: PVC-Metallopeptidase; Repeats: RHS
t2.35	Tox-Colicin D	BECR fold ($\alpha + \beta$); (PDB: 1v74); [KH]K, Hxx[ED], [ST], [TS]xxK; Of the conserved residues in ColicinD (PDB: 1v74), K607, K608, H611, D614, and S677 are essential for activity	β, γ, δ -proteobacteria, chloroflexi, firmicutes, spirochaetes, <i>archaea</i> : euryarchaeota, eukaryotes: fungi (ascomycota)	T2SS, T5SS, Cloacin, TcdB/TcaC, PVC, MuF	ImmD, Imm64; ImmD is the major immunity protein share with plasmid borne colicin systems	Proteases: PVC-Metallopeptidase; Repeats: RHS, FilH
t2.36	Tox-ColicinC/E5	BECR fold ($\alpha + \beta$, PDB: 2dfx); K, W, Y, Y, Q, [RK], W; Of the conserved residues in Colicin E5 (PDB: 2dfx), Y81 and S95 are predicted to be involved in catalysis	β, γ -proteobacteria, firmicutes, Plasmid Cole5-099	T2SS, T5SS, T7SS (LXG), Cloacin/PyocinS, TcdB/TcaC	ImmE5	Repeats: RHS, FilH
t2.37	tRNase					
t2.38	Tox-EndoU	BECR fold ($\alpha + \beta$, PDB: 2c1w); H, H, [SNT],[SNT]; This structural core contains two BECR fold units, where the N-terminal unit has lost strand-4, while the helix in the C-terminal unit has flipped to the opposite end. In 2c1w, H162 and T278 form one pair of catalytic residues and H178 and S229 form the other (Figure 2E). Some members use a Mn ²⁺ probably as a transition state stabilizer	Actinobacteria, α, β, γ -proteobacteria, bacteroidetes, chlamydiae, cyanobacteria, fibrobacteres, firmicutes, fusobacteria, tenericutes, eukaryotes: hemichordata, viridiplantae, stramenopiles, metazoa	T2SS (MafBN), T5SS, T6SS, T7SS (WXG,LXG)	Imm-SUKH, Imm-SuFu, Imm28	Proteases: HINT; Repeats: FilH, RHS
t2.39	(including XendoU)					
t2.40	Tox-ReLE	BECR fold ($\alpha + \beta$); [KR], R; The active site residues in the classical ReLE (PDB: 3kha) correspond to residues R61 and R81	Actinobacteria, $\alpha, \gamma, -$ proteobacteria, bacteroidetes, cyanobacteria, firmicutes, fusobacteria	T2SS	Imm54	Proteases: HINT; Repeats: RHS

Table 2 Phyletic distribution, export pathways, and contextually-associated domains and proteins of polymorphic toxin domains (Continued)

t2.41	Ntox7	Predicted BECR fold ($\alpha + \beta$); DGx + xhR, N motif	Actinobacteria, bacteroidetes, β, γ, δ - proteobacteria, chlamydiae, chloroflexi, firmicutes	T2SS (MafBN), T2SS (APD1), T5SS, T7SS, TcdB/TcaC	Imm8, Imm31, Imm32, Imm-NMB0513, Imm-SuFu; Imm8 is the predominant immunity protein across a wide phyletic range	<i>Proteases:</i> HINT, ZU5; <i>Repeats:</i> FilH, RHS	
t2.42	Ntox19	Predicted BECR fold ($\alpha + \beta$); D,H,DxxxR,E,HxxF; Also found in mimivirus, where it is fused to ankyrin repeats,	β, γ, δ - proteobacteria, firmicutes, fusobacteria, bacteroidetes, Viruses: Acanthamoeba polyphaga mimivirus	T2SS (MafBN), T5SS, T7SS (LxG and WxG), TcdB/TcaC	Imm38, Imm40. These associations are seen across many different bacterial lineages	<i>Repeats:</i> FilH, RHS	
t2.43	Ntox21; Also referred	Predicted BECR fold ($\alpha + \beta$); K, [DS]xDxxxH, K, RxG[ST], RxxD	Actinobacteria, α, β, γ -proteobacteria bacteroidetes, firmicutes	T2SS (MafBN), T5SS, T4SS, T7SS	Imm-Barstar, Imm41	<i>Proteases:</i> HINT; <i>Repeats:</i> RHS, FilH	
t2.44	to as the <i>E. cloacae</i>						
t2.45	CdiAC; Shown to						
t2.46	be a tRNAse						
t2.47	Ntox35	Predicted BECR fold ($\alpha + \beta$); H, KH	Actinobacteria, bacteroidetes, β -proteobacteria, chlamydiae, chloroflexi, firmicutes, planctomycetes	T2SS (MafBN)	-	<i>Repeats:</i> RHS	
t2.48	Ntox36	Predicted BECR fold ($\alpha + \beta$); N, [RY], [DE]	Acidobacteria, actinobacteria, β, γ -proteobacteria, cyanobacteria, elusimicrobia, firmicutes	T2SS, T5SS	-	<i>Proteases:</i> HINT; <i>Repeats:</i> RHS, FilH	
t2.49	Ntox41	Predicted BECR fold ($\alpha + \beta$); [RK]H, [KR], [ST]xxP	Actinobacteria, α, β, γ -proteobacteria, bacteroidetes, firmicutes, planctomycetes	T2SS, T5SS, T7SS (WXG,LXG)	-	<i>Proteases:</i> HINT; <i>Repeats:</i> RHS, FilH, ALF	
t2.50	Ntox47	Predicted BECR fold ($\alpha + \beta$); D, [HRK], RT, E, D, PH, H, [DE], R	β, γ -proteobacteria, firmicutes	T2SS, T6SS, T7SS (LXG,WXG)	-	<i>Proteases:</i> HINT; <i>Repeats:</i> RHS	
t2.51	Ntox48	Predicted BECR fold ($\alpha + \beta$); R, [RK], Q, Q	Acidobacteria, actinobacteria, $\alpha, \beta, \gamma, \delta$ -proteobacteria, bacteroidetes, cyanobacteria, firmicutes, fusobacteria, planctomycetes	T2SS, T5SS, T6SS T7SS (WXG,LXG),	Imm60, Imm62, Imm66, Imm71	<i>Proteases:</i> HINT; <i>Repeats:</i> RHS, FilH	
t2.52	Ntox49	Predicted BECR fold ($\alpha + \beta$); H, [KR]	Actinobacteria, $\alpha, \beta, \gamma, \delta$ -proteobacteria, bacteroidetes, chlamydiae, chloroflexi, cyanobacteria, firmicutes, thermotogae, <i>archaea:</i> euryarchaeota, eukaryotes: stramenopiles, viridiplantae, viruses: caudovirales	T2SS (MafBN), T5SS, T7SS (WXG,LXG), MuF, PVC	Imm22	<i>Proteases:</i> PVC-Metallopeptidase, HINT, ZU5; <i>Repeats:</i> RHS	
t2.53	Ntox50	Predicted BECR fold ($\alpha + \beta$); H, S, K, T, H, K, HxVP	Actinobacteria, β, γ, δ -proteobacteria, chlamydiae, firmicutes, fusobacteria, viruses: caudovirales	T2SS (MafBN), T6SS, T7SS (WXG,LXG), MuF	-	<i>Proteases:</i> HINT	
t2.54	Predicted metal-independent RNase toxins						
t2.55	Tox-CdiAC	All- β ; N, [DSN],E	β, γ, δ -proteobacteria	T2SS, T5SS, T6SS, TcdB/TcaC	Imm-Cdil, Imm5 + Imm36. Imm-Cdil is the most prominent immunity protein to this toxin	<i>Repeats:</i> RHS, FilH	
t2.56	Tox-Cole3	All- β ; Cole3 cytotoxic ribonuclease fold, R, Dxx + [HK], E, H	Actinobacteria, α, β, γ -proteobacteria, bacteroidetes, cyanobacteria, firmicutes, fusobacteria	T2SS (MafBN), T5SS, T7SS (WXG,LXG)	Imm-Cloacin, Imm45	<i>Proteases:</i> HINT; <i>Repeats:</i> RHS, FilH	

Table 2 Phyletic distribution, export pathways, and contextually-associated domains and proteins of polymorphic toxin domains (Continued)

t2.57	Tox-RES; PF08808 in	$\alpha + \beta$; R, R, E, S	Acidobacteria, actinobacteria,	T2SS, T5SS, T6SS	Imm51, Antitoxin-	Repeats: RHS, FilH
t2.58	Pfam. Also found		$\alpha, \beta, \gamma, \delta, \epsilon$ -proteobacteria, bacteroidetes,		DUF2384(in AT system)	
t2.59	in toxin-antitoxin		chlorobi, chloroflexi, cyanobacteria,			
t2.60	systems (see text);		deinococci, firmicutes, nitrospirae, spirochaetes, synergistetes, verrucomicrobia, Viruses: caudovirales			
t2.61	Ntox2	$\alpha + \beta + \alpha$ -helical C-terminus; GEsH motif and conserved E, RE, H and K; Multiple copies in the same gene neighborhood in <i>Microscilla</i>	<i>Microscilla marina</i> (Bacteroidetes)	PVC	-	Proteases: PVC-Metallopeptidase
t2.62	Ntox4	$\alpha + \beta$; Several charged residues	<i>Nitrosococcus</i> , <i>Frankia</i>	PVC	-	Proteases: PVC-Metallopeptidase
t2.63	Ntox5	$\alpha + \beta$; Several charged residues	<i>Streptomyces</i> , <i>Nitrobacter</i>	PVC	-	Proteases: PVC-Metallopeptidase
t2.64	Ntox9	Mostly β ; RxY, E, WxE and H; Catalytic mechanism likely to be similar to that of Colicin-E3	Actinobacteria, α, β, γ -proteobacteria bacteroidetes, chlamydiae, fusobacteria	T2SS (MafBN), T5SS, T6SS	-	Proteases: PrsW peptidase; Repeats: RHS
t2.65	Ntox12	All- β ; D, D, H	Actinobacteria, chlamydiae, firmicutes, α, β, γ - proteobacteria	T2SS, T5SS T6SS, T7SS (WxG and LxG), TcdB/TcaC	Imm32; Note immunity protein also present in intracellular parasite <i>Odysella</i>	Proteases: OUT; Repeats: RHS, FilH
t2.66	Ntox13	β/α , KxxxxxE motif	Firmicutes, β -proteobacteria	T2SS	Imm59	Repeats: RHS
t2.67						Proteases: Transglutaminase
t2.68						
t2.69	Ntox15	Mostly α , HxxD motif	Actinobacteria, firmicutes, α, β, γ - proteobacteria	T2SS, T6SS, T7SS (LDxD and LxG), PVC	Imm-SUKH	Proteases: PVC-Metallopeptidase, HINT
t2.70	Ntox16	α -helical domain; R, [DNE]xxH; part of polytoxin in <i>Xanthomonas fuscans</i>	Cyanobacteria, β, γ, δ proteobacteria, verrucomicrobia	T2SS, T6SS, PVC	-	Proteases: PVC-Metallopeptidase; Repeats: RHS
t2.71	Ntox17	Mostly β ; ExD, H, several charged residues	α, β, γ proteobacteria, firmicutes	T2SS (MafB), TcdB/TcaC, T7SS	Imm31; association widespread several lineages	Repeats: RHS
t2.72	Ntox20	Mostly β ; conserved R	Acidobacteria, $\alpha, \beta, \gamma, \epsilon$ -proteobacteria	T2SS (MafBN), T5SS	Imm-NMB0513, Imm-SUKH Imm28	Repeats: FilH
t2.73	Ntox23	All- β ;	Bacteroidetes	T2SS, TcdB/TcaC	-	Repeats: RHS
t2.74		ND, DxxR, H				
t2.75	Ntox24	All- β ; Y, H, H; Also found in Toxin-Antitoxin systems (see text)	Actinobacteria, α, β, γ -proteobacteria, chlamydiae, chloroflexi, firmicutes, fusobacteria	T2SS, T5SS T7SS (WXG,LXG), MuF	Imm50, Imm53	Proteases: HINT; Repeats: RHS, FilH
t2.76	Ntox25	Mostly β ; FGPY motif	α, γ -proteobacteria, bacteroidetes	T2SS, T5SS	-	Repeats: FilH

Table 2 Phyletic distribution, export pathways, and contextually-associated domains and proteins of polymorphic toxin domains (Continued)

t2.77	Ntox27	$\alpha + \beta$; D, E, RxW	Actinobacteria, bacteroidetes, fusobacteria	T2SS, T7SS (WXG)	-	<i>Proteases</i> : HINT; <i>Repeats</i> : ALF, RHS
t2.78	Ntox28	All- α ; D,K[DE], [DN]HxxE, E	Actinobacteria, α,γ -proteobacteria, firmicutes	T2SS, T5SS T7SS (WXG)	-	<i>Repeats</i> : FilH
t2.79	Ntox31	$\alpha + \beta$; K, E, E	Actinobacteria, α,γ -proteobacteria, bacteroidetes, firmicutes, eukaryotes: ciliophora	T2SS, T5SS, T6SS, T7SS (WXG, LXG)	Imm62	<i>Repeats</i> : RHS, FilH
t2.80	Ntox32	All- α ; H, [KR], [ED], [DE]	Bacteroidetes, α,γ -proteobacteria, firmicutes, eukaryotes: insects	T2SS	-	<i>Proteases</i> : Peptidase S8 (Subtilisin family); <i>Repeats</i> : RHS
t2.81	Ntox34	All- α ; GNxxD, K, C, C, K, WxCxH and other charged residues	γ,δ,ϵ -proteobacteria, firmicutes	T2SS, T6SS	Imm-HEAT	<i>Repeats</i> : RHS
t2.82	Ntox37	All- β ; E, [KR] Hx[DH]	Actinobacteria, γ -proteobacteria, chlamydiae, chloroflexi, firmicutes	T2SS, T7SS(WXG)	Imm32	<i>Proteases</i> : Tox-PLOTU; <i>Repeats</i> : RHS
t2.83	Ntox39	All- β ; Several basic residues	Firmicutes	T2SS	-	<i>Repeats</i> : RHS
t2.84	Ntox40	All- β ; DRxxG, R, Y	Acidobacteria, actinobacteria, $\alpha,\beta,\gamma,\epsilon$ -proteobacteria, bacteroidetes, firmicutes, planctomycetes, synergistetes, eukaryotes: fungi	T2SS, T5SS, T6SS, T7SS (WXG,LXG,LDXD), TcdB/TcaC	Imm35, Imm36, Imm59, Imm60, Imm61, Imm63	<i>Repeats</i> : RHS, FilH
t2.85	Ntox42	$\alpha + \beta$; GK, ExxxH, DxYxF[ED]	Firmicutes (negativicutes)	T5SS	-	<i>Repeats</i> : FilH
t2.86	Ntox44	All- α ; DxK, GNxxxG, and DxxxD.	Actinobacteria, $\alpha,\beta,\gamma,\delta$ -proteobacteria, bacteroidetes, chloroflexi, firmicutes, proteobacteria, spirochaetes, eukaryotes: fungi (microsporidia)	T2SS, T6SS, T7SS(WXG,LXG)	-	<i>Proteases</i> : Papain-like protease; <i>Repeats</i> : RHS, ALF
t2.87	Predicted RNase toxins with two conserved histidine residues					
t2.88	Tox-EDA39C	$\alpha + \beta$; H, Sx[HS]Y; Present in a wide range of eukaryotes where it might be a defensive RNase	Acidobacteria, actinobacteria, $\alpha,\beta,\gamma,\delta$ -proteobacteria, bacteroidetes, chlamydiae, chloroflexi, firmicutes, gemmatimonadetes, planctomycetes, verrucomicrobia, eukaryotes: plants, chlorophytes, fungi, dictyosteliida, stramenopiles	T2SS, T5SS, T6SS, T7SS (LXG)	Imm-SuFu	<i>Proteases</i> : HINT; <i>Repeats</i> : RHS
t2.89	Ntox18	α/β ; H, S, H	α,β,γ - proteobacteria, bacteroidetes, chloroflexi, cyanobacteria, firmicutes, eukaryotes: metazoan: Lateral transfer to Branchiostoma	T2SS (MafBN), T2SS	Imm29, Imm42; Imm29 association is widespread across bacteria	<i>Proteases</i> : HINT; <i>Repeats</i> : RHS, FilH
t2.90	Ntox22	Mostly β , D, D, H, E, H	<i>Ralstonia</i> , <i>Burkholderia phymatum</i>	T5SS	-	<i>Repeats</i> : FilH
t2.91	Ntox26	$\alpha + \beta$; KHxx[DE], Q, W, H	Actinobacteria, α,β,γ -proteobacteria, firmicutes, fusobacteria	T2SS, T5SS T7SS (LXG)	-	<i>Proteases</i> : HINT; <i>Repeats</i> : RHS, FilH, Tail fiber
t2.92	Ntox30	All- β ; RxH, R THIP	Actinobacteria, bacteroidetes, α,γ -proteobacteria, firmicutes, spirochaetes	T2SS, T6SS, T7SS (WXG, LXG), TcdB/TcaC	-	<i>Repeats</i> : RHS
		$\alpha + \beta$; with two conserved H		T2SS, TcdB/TcaC	-	<i>Repeats</i> : RHS

Table 2 Phyletic distribution, export pathways, and contextually-associated domains and proteins of polymorphic toxin domains (Continued)

t2.93	Ntox43; <i>Pseudomonas</i>		Actinobacteria, γ , δ -proteobacteria,			
t2.94	RhsT-C belongs to		firmicutes, verrucomicrobia			
t2.95	this clade					
t2.96	Tox-JAB1	Deaminase fold ($\alpha + \beta$); NxxxE, HxH, S, D	Bacteroidetes	T2SS	Imm65	Repeats: RHS
t2.97	Tox-JAB2 (DUF4329	Deaminase fold ($\alpha + \beta$);	α , γ , δ -proteobacteria bacteroidetes,	T2SS, T6SS, T7SS	Imm-NTF2 family 2	Repeats: RHS
t2.98	in Pfam)	E, H[ST]H, S, D	cyanobacteria, firmicutes, eukaryotes: fungi (ascomycota), viruses: caudovirales	(WXG), TcdB/TcaC		
t2.99	Tox-ComI	$\alpha + \beta$ fold; DE motif	Actinobacteria, α , β , γ -proteobacteria, bacteroidetes, firmicutes, verrucomicrobia, eukaryotes: dictyosteliida, fungi (ascomycota, basidiomycota), viruses: Bacillus phage SP10	T2SS, T6SS	Imm-ComJ, Imm-SUKH	Proteases: HINT; Repeats: RHS
t2.100	Tox-HET-C	All- α ; H, [DE], HxD, HxxxDxxxH, Nxx[DE], [ST]G; We predict that the Het-C domain is related to phospholipase C and the S1-P1 nuclease and shares a common active site and fold (see text)	Actinobacteria, cyanobacteria, γ , δ -proteobacteria, dictyoglomi, eukaryotes: fungi (ascomycota, basidiomycota), metazoa	T2SS, T6SS, PVC	-	Proteases: PVC-Metallopeptidase
t2.101	Ntox29	All- β ; D,D, HxE, D, K, R residues	β , γ -proteobacteria, firmicutes	T2SS, T5SS,T7SS (LXG)	Imm41	Proteases: HINT; Repeats: RHS, FilH
t2.102	Predicted RNase toxins with uncertain metal dependence					
t2.103	Ntox1	$\alpha + \beta$ fold; C, C, H, E	Acidobacteria, α -proteobacteria	PVC		Proteases: PVC-Metallopeptidase
t2.104	Ntox3	All- β ; several charged residues including as D, R, H, C; associated with Annexin domain in <i>Haliangium</i>	<i>Haliangium</i> (δ -proteobacteria), <i>Microscilla</i> (Bacteroidetes)	PVC	-	Proteases: PVC- Metallopeptidase; Repeats: Annexin
t2.105	Ntox6	$\alpha + \beta$; several charged residues;	<i>Microcoleus</i> (Cyanobacteria), <i>Haliangium</i> (δ -proteobacteria)	PVC	-	Proteases: PVC- Metallopeptidase
t2.106	Ntox8	$\alpha + \beta$ fold; HxR and HxxxH motif	β -proteobacteria, bacteroidetes, firmicutes, eukaryotes: dictyosteliida	T2SS, T6SS	Imm16	Repeats: RHS
t2.107	Ntox10	$\alpha + \beta$; Several charged residues	Bacteroidetes, verrucomicrobia	T2SS	Imm27, Imm53; Imm27	Repeats: RHS
t2.108					primary immunity	Proteases:
t2.109					protein across most lineages	Transglutaminase
t2.110	Ntox11	α/β followed by β rich C-terminus; N-terminal GxR, RxxxoH motif, C-terminal domain has H, GxE, GxxH and an acidic residues; <i>Naegleria</i> possibly secreted	Actinobacteria, cyanobacteria, firmicutes α , δ , γ -proteobacteria, eukaryotes: Trichoplax, Naegleria	PVC	-	Proteases: PVC- Metallopeptidase
t2.111	Ntox14	$\alpha + \beta$; Several charged residues	<i>Desulfobacca</i> , <i>Pelobacter</i> (δ -proteobacteria)	PVC	Imm22	Proteases: PVC-Metallopeptidase

Table 2 Phyletic distribution, export pathways, and contextually-associated domains and proteins of polymorphic toxin domains (Continued)

t2.112	Ntox33	$\alpha + \beta$; [DN]xHxxK, DxxxD	Actinobacteria, cyanobacteria, firmicutes, γ -proteobacteria, verrucomicrobia	T2SS	-	-
t2.113	Ntox45	$\alpha + \beta$; Dx D motif	Actinobacteria, α -proteobacteria, bacteroidetes	T2SS	-	<i>Proteases:</i> HINT; <i>Repeats:</i> RHS
t2.114	Other toxins that act on nucleic acids					
t2.115	Tox-Deaminase	Deaminase fold ($\alpha + \beta$); [HCD]xE, CxxC; As	Acidobacteria, actinobacteria, bacteroidetes, chlorobi, cyanobacteria, firmicutes, $\alpha, \beta, \gamma, \delta$ -proteobacteria	T2SS (MafBN), T5SS, T6SS, T7SS (WXG, LDxD, LXG), PVC, TcdB/TcaC	Imm1, Imm2, Imm3, Imm4, Imm5, Imm6, Imm10, Imm18, Imm-SUKH, Imm-ank	<i>Proteases:</i> PVC-Metallopeptidase, HINT, CPD, PrsW peptidase, Caspase; <i>Repeats:</i> RHS, FilH, ALF, PPR
t2.116		previously reported, nine distinct families of deaminase belonging to two distinct clades are present in polymorphic toxin systems as toxins. We report two additional families below	Eukaryotes: See text and previous publication			
t2.117						
t2.118	Tox-Deaminase (sce3516-like)	Deaminase fold ($\alpha + \beta$); H[occasionally D]xE, CxxC; Toxins of this family belong to the strand-hairpin clade of deaminases	Actinobacteria, β, γ, δ -proteobacteria	T2SS, T5SS, T6SS, T7SS, TcdB/TcaC	Imm-SUKH	<i>Proteases:</i> HINT <i>Repeats:</i> RHS, FilH
t2.119						
t2.120						
t2.121	Tox-Deaminase (WD0512-like)	Deaminase fold ($\alpha + \beta$); CxE, CxxC; Toxins of this family belong to the Helix-4 clade of deaminases. These proteins additionally contain a C-terminal toxin, the Tox-Latrotoxin-CTD	α -proteobacteria (<i>Wolbachia</i>)	T2SS	-	<i>Repeats:</i> RHS
t2.122						
t2.123	Tox-ParB	ParB fold ($\alpha + \beta$); R	Actinobacteria, $\alpha, \beta, \gamma, \delta$ -proteobacteria, bacteroidetes, firmicutes	T2SS (MafBN), T5SS, T6SS, T7SS (WXG), PVC	Imm20, Imm27, Imm-SuFu	<i>Proteases:</i> PVC-Metallopeptidase, HINT; <i>Repeats:</i> RHS, FilH
t2.124	Tox-ParBL1	Predicted ParB fold ($\alpha + \beta$); [ST], [NT][RT][RT]; note the latter two residues of this motif are mostly R	Actinobacteria, α, β, γ -proteobacteria, firmicutes, euryarchaea, eukaryotes: stramenopiles, viridiplantae, ascomycota, chlorophyta, choanoflagellida, metazoa, ciliophora, kinetoplastida	T2SS (MafBN), T5SS, T6SS, T7SS (WXG, LXG)	Imm-SUKH, Imm44	<i>Proteases:</i> HINT; <i>Repeats:</i> FilH, RHS
t2.125	Tox-HTH	HTH fold; RxxY, R, [ST]	Acidobacteria, actinobacteria, $\alpha, \beta, \gamma, \delta, \epsilon$ -proteobacteria, bacteroidetes, cyanobacteria, firmicutes, proteobacteria, archaea, eukaryotes: ascomycota, viridiplantae,	T2SS, T5SS, T6SS, T7SS (LXG, WXG, LDxD), PVC, MuF	-	<i>Proteases:</i> PVC-Metallopeptidase; <i>Repeats:</i> FilH
t2.126	Peptidase toxins					
t2.127	Tox-	metallopeptidase fold	Actinobacteria, bacteroidetes,	PVC, T2SS	Imm-SuFu	<i>Proteases:</i> PVC-Metallopeptidase
t2.128	ALFMetallopeptidase(Anthrax lethal factor)	($\alpha + \beta$); HExxH	δ -proteobacteria, firmicutes, fibrobacteres			<i>Repeats:</i> FilH
t2.129						
t2.130						
t2.131	Tox-HopH1	metallopeptidase fold ($\alpha + \beta$); HExxH, [DE]N	Actinobacteria, α, β, γ -proteobacteria, bacteroidetes, planctomycetes	T2SS, T5SS, T6SS, T7SS (WXG), PVC, TcdB/TcaC	-	<i>Proteases:</i> PVC-Metallopeptidase,

Table 2 Phyletic distribution, export pathways, and contextually-associated domains and proteins of polymorphic toxin domains (Continued)

t2.132	Tox-Metallopeptidase1	metallopeptidase fold (α + β); HExxH	Actinobacteria, α,β,γ,δ-proteobacteria, bacteroidetes, chlorobi, cyanobacteria, deinococci, planctomycetes, spirochaetes, thermotogae	T2SS,T7SS (WXG), TcdB/TcaC	-	ZU5, caspase; <i>Repeats</i> : RHS <i>Repeats</i> : RHS
t2.133	Tox-Metallopeptidase2	metallopeptidase fold (α + β); Y, HExxH,	Bacteroidetes	TcdB/TcaC	-	<i>Proteases</i> : ZU5; <i>Repeats</i> : RHS
t2.134	Tox-Metallopeptidase3	metallopeptidase fold (α + β); K, HExxH, F[DE]	α-proteobacteria, bacteroidetes	T2SS, PVC	-	<i>Proteases</i> : PVC-Metallopeptidase; <i>Repeats</i> : RHS
t2.135	Tox-Metallopeptidase4	metallopeptidase fold (α + β); F[DN], [RK], HExxH	γ-proteobacteria, fusobacteria, firmicutes, planctomycetes	T2SS, T6SS, T7SS (WXG,LDXD,LXG)	-	<i>Repeats</i> : RHS
t2.136	Tox-Metallopeptidase5	metallopeptidase fold (α + β); HEELH	Actinobacteria, γ-proteobacteria	T2SS	-	<i>Repeats</i> : RHS
t2.137	PVC-Metallopeptidase	metallopeptidase fold (α + β); HExxH; Most versions of this domain are releasing peptidases in polymorphic toxins. However, some versions, often present at the C-terminal end of polymorphic toxins, are likely to additionally function as toxins	Acidobacteria, actinobacteria, α,β,γ,δ-proteobacteria, bacteroidetes, chlorobi, chloroflexi, cyanobacteria, deinococci, firmicutes, nitrospirae, verrucomicrobia, <i>archaea</i> : euryarchaeota, eukaryotes: fungi(ascomycota)	PVC	-	<i>Proteases</i> : PVC-Metallopeptidase; <i>Repeats</i> : RHS
t2.138	Tox-MCF-SHE	All-α; S, T, HSxxE	Actinobacteria, α,β,γ,δ-proteobacteria, bacteroidetes, chlamydiae, viruses: Acanthamoeba polyphaga mimivirus	T2SS, T7SS(WXG), PVC	-	<i>Proteases</i> : PVC-Metallopeptidase, Caspase, Tox-PLOTU
t2.139	Tox-SerPeptidase	α + β; H, R, R	Actinobacteria, α,β,γ,δ,ε-proteobacteria	T2SS, T7SS (WXG)	-	<i>Proteases</i> : Tox-PLOTU
t2.140	Tox-YabG	α + β; HxD, Y, E, [DE], GHD, Y, R	Bacteroidetes, firmicutes	PVC	DUF1021(antitoxin in toxin-antitoxin systems), Imm-SUKH	<i>Proteases</i> : PVC-Metallopeptidase
t2.141	Tox-LD-peptidase	LD-peptidase (PDB: 1ZAT); H, S, C	Actinobacteria, bacteroidetes, β,γ,δ-proteobacteria, chloroflexi, firmicutes	T2SS,T6SS, TcdB/TcaC	Imm16, Imm57	<i>Proteases</i> : ZU5; <i>Repeats</i> : RHS
t2.142	Tox-Caspase	Caspase-like fold (α/β); H, C; Most versions of this domain are releasing peptidases in polymorphic toxins. However, some versions, often present at the C-terminal end of polymorphic toxins, are likely to additionally function as toxins	Actinobacteria, α,β,γ,δ,ε-proteobacteria, bacteroidetes, chloroflexi, cyanobacteria, firmicutes, viruses: caudovirales	T2SS,T6SS, T7SS (WXG,PPE), PVC	Imm36	<i>Proteases</i> : PVC-Metallopeptidase; <i>Repeats</i> : RHS
t2.143	Tox-HDC	α + β; H, D, C	β,γ-proteobacteria, viruses: caudovirales	T2SS	-	<i>Proteases</i> : Caspase; <i>Repeats</i> : RHS
t2.144	Tox-NLPC/P60	Papain-like peptidase fold (α + β); C, H, D	Bacteroidetes, δ-proteobacteria	T6SS, PVC, TcdB/TcaC	-	

Table 2 Phyletic distribution, export pathways, and contextually-associated domains and proteins of polymorphic toxin domains (Continued)

t2.145 Tox-PL1	Papain-like peptidase fold ($\alpha + \beta$); NC, H, D; Most versions of this domain are toxins in polymorphic toxins. However, some versions are, additionally, likely to be releasing peptidases	Actinobacteria, bacteroidetes, γ, δ -proteobacteria, firmicutes, fusobacteria, gemmatimonadetes	T2SS, T6SS, T7SS (WXG), MuF	-	<i>Proteases:</i> PVC-Metallopeptidase, ZU5; <i>Repeats:</i> RHS <i>Proteases:</i> Tox-Caspase, HINT; <i>Repeats:</i> RHS
t2.146 Tox-PL-2	Papain-like peptidase fold ($\alpha + \beta$); C, NxxH, DN	β, δ -proteobacteria, cyanobacteria, firmicutes	T2SS, TcdB/TcaC	Imm73	<i>Proteases:</i> HINT, PLOTU, ZU5; <i>Repeats:</i> RHS
t2.147 Tox-PL3	Papain-like peptidase fold ($\alpha + \beta$); C, [DE]H, [DE], R	Bacteroidetes, fibrobacteres, δ, ϵ -proteobacteria	T2SS, TcdB/TcaC	-	<i>Proteases:</i> ZU5; <i>Repeats:</i> RHS
t2.148 Tox-PLOTU	Papain-like peptidase fold ($\alpha + \beta$); C, H, D; Most versions of this domain are releasing peptidases in polymorphic toxins. However, some versions, often present at the C-terminal end of polymorphic toxins, are likely to additionally function as toxins	Actinobacteria, α, γ -proteobacteria, bacteroidetes, chlamydiae, eukaryotes: fungi (ascomycota), metazoa, viridiplantae, viruses: Invertebrate iridescent virus 3, Wiseana iridescent virus	T2SS (APD1), T7SS (WXG)	-	<i>Repeats:</i> Ankyrin, Sel1, FilH
t2.149 Tox-PLC39	Papain-like peptidase fold ($\alpha + \beta$); C, H, D	Bacteroidetes, chloroflexi, firmicutes	T2SS, T6SS, PVC	-	<i>Proteases:</i> PVC-Metallopeptidase; <i>Repeats:</i> RHS
t2.150 Tox-PLDMTX	Papain-like peptidase fold ($\alpha + \beta$); C, W, H, D, Q	α, β, γ -proteobacteria	T2SS	-	-
t2.151 Tox-TGase	Papain-like fold ($\alpha + \beta$); C, H, D	β, γ, δ -proteobacteria, bacteroidetes, cyanobacteria	T2SS, PVC	-	<i>Proteases:</i> PVC-Metallopeptidase
t2.152 Tox-UCH	Papain-like fold ($\alpha + \beta$); C, H, D	β -proteobacteria	PVC	-	<i>Proteases:</i> PVC-Metallopeptidase
t2.153 Tox-OmpA	$\alpha + \beta$;	α, β, γ -proteobacteria, cyanobacteria	PVC	-	<i>Proteases:</i> PVC-Metallopeptidase
t2.154 Protein-modifying toxins					
t2.155 Tox-ART-RSE;	ADP-ribosyltransferase fold ($\alpha + \beta$); RxDxR, S, [DN]xN, E	Actinobacteria, $\alpha, \beta, \gamma, \delta$ -proteobacteria, bacteroidetes, chloroflexi, firmicutes, planctomycetes, spirochaetes, tenericutes, eukaryotes: fungi (ascomycota, basidiomycota), metazoan (hexapoda, mollusca), viridiplantae, viruses: Vibrio phage CTX	T2SS, T6SS, T7SS (WXG, LXG, LDXD)	Imm41, Imm-ADP-RGHD (ADP-ribosylglycohydrolase)	<i>Proteases:</i> HINT, Caspase, MCF-SHE; <i>Repeats:</i> RHS, Tail-fiber
t2.156 Tox-ART-PARP	ADP-ribosyltransferase fold ($\alpha + \beta$); HG[ST], Y, K, E	Actinobacteria	PVC	-	<i>Proteases:</i> PVC-Metallopeptidase

Table 2 Phyletic distribution, export pathways, and contextually-associated domains and proteins of polymorphic toxin domains (Continued)

t2.157	Tox-ART-HYE1	ADP-ribosyltransferase fold (α + β); H, Y, E	γ-proteobacteria	TcdB/TcaC?	-	Repeats: RHS
t2.158	Tox-ART-HYD1	ADP-ribosyltransferase fold (α + β); H,[RK], [FY], [DE]	Actinobacteria, β,γ-proteobacteria, bacteroidetes, firmicutes	T2SS, T6SS, T7SS	Imm-My6CBD;	Proteases: HINT; Repeats: RHS
t2.159	Tox-ART-HYD2	ADP-ribosyltransferase fold (α + β); H, D, GFY, W, R	Actinobacteria, bacteroidetes, deinococci, fibrobacteres, firmicutes, fusobacteria, γ-proteobacteria, lentisphaerae, spirochaetes, synergistetes, eukaryotes: choanoflagellida, filasterea, fungi, cnidaria	T2SS, PVC	-	Proteases: HINT, PVC-Metallopeptidase; Repeats: RHS, Tail-Fiber
t2.160	Tox-Arc	Flavodoxin fold (α/β); [ST]	Actinobacteria, bacteroidetes, cyanobacteria, firmicutes, β, γ- proteobacteria, spirochaetes	T2SS, T5SS, T6SS, T7SS (LXG, WXG)	Imm74, Imm63; Imm74 is the primary immunity protein across wide phyletic range	Repeats: RHS, FilH
t2.161	(ADP-Ribosyl cyclase)	[DE], S, E	eukaryotes: fungi (ascomycota, basidiomycota), Capsaspora, choanoflagellida, metazoa; This domain appears to have independently been acquired by the fungi and the animals from the bacteria.			
t2.162	Tox-Doc	Doc/Fic fold (PDB: 2f6s, All-α); HxFx[DE]GNxR; (See Pfam PF02661)	Actinobacteria, γ-proteobacteria	T5SS, T7SS (WXG)	Imm23, Imm-SUKH, Imm13	Proteases: Caspase; Repeats: FilH
t2.163	Tox-CNF (Cytotoxic necrotizing factor)	CNF1/YfiH fold (α + β, PDB: 1hzg); D, C, H; See Pfam PF05785	γ-proteobacteria	T6SS	-	Repeats: RHS
t2.165	Tox-Glycosyltransferase	Nucleotide diphospho-sugar transferase fold (α/β); [DNE]xxR, YxDxD; See Pfam PF04488	Actinobacteria	T7SS (WXG), PVC	-	Proteases: PVC-Metallopeptidase
t2.166	Tox-Peptide Kinase	α + β; DxH, YKP[KR], DxHxEN, DxE, S, R; Related to the kinase domain found in lantibiotic synthetases	Firmicutes	PVC	-	Proteases: PVC-Metallopeptidase
t2.167	Pore-forming toxins					
t2.168	Tox-WTIP	Two membrane spanning α-helices; RxxR, Wx[ST]IP	α,β,γ-proteobacteria	T2SS, PVC	-	Proteases: PVC-Metallopeptidase; Repeats: RHS
t2.169	Toxins that act on carbohydrates					
t2.170	Tox-Aldo-ketoreductase	Rossmann (α/β);	Bacteroidetes, cyanobacteria	PVC	-	Proteases: PVC-Metallopeptidases;
t2.171	Tox-Glucosaminidase	Lysozyme-like fold (α + β); E, N, Y (See Pfam PF01832)	Firmicutes	T6SS, PVC	-	Proteases: PVC-Metallopeptidase
t2.172	Toxins that act on lipids					
t2.173	Tox-AB hydrolase1	α/β hydrolase (α/β); DG, [ST]N,	Acidobacteria, actinobacteria,	T2SS, T6SS	-	Repeats: RHS
t2.174	(Pfam DUF2235)	[KR], D, ExE, GxHxD	α,β,γ,δ,ε-proteobacteria, bacteroidetes, cyanobacteria, nitrospirae planctomycetes, verrucomicrobia, eukaryotes:			

Table 2 Phyletic distribution, export pathways, and contextually-associated domains and proteins of polymorphic toxin domains (Continued)

		fungi(ascomycota, basidiomycota), rhodophyta, viridiplantae			
t2.175	Tox- AB hydrolase3	α/β hydrolase (α/β); G[ST], GHSxG	Actinobacteria, α,β,γ -proteobacteria, bacteroidetes, firmicutes	T2SS, T6SS,T7SS (WXG), TcdB/TcaC	Imm66, Imm69 <i>Repeats:</i> RHS, FilH
t2.176	Tox-PLA2	Phospholipase A2 fold (All- α , PDB: 1kp4); DxC[ST], CxxHxxxYxN, C	Actinobacteria, $\alpha,\beta,\gamma,\delta$ -proteobacteria, aquificae, bacteroidetes, chlorobi, chloroflexi, cyanobacteria, deinococci, firmicutes, fusobacteria, nitrospirae, planctomycetes, spirochaetes, eukaryotes: fungi(ascomycota), heterolobosea, metazoa, stramenopiles, viridiplantae, Viruses: Campylobacter phage	T2SS	- <i>Repeats:</i> RHS, ALF
t2.177	Tox-CDP-alcohol	All- α ; DxxDGxxxR, DxxxD; See Pfam PF01066	β -proteobacteria (mainly Neisseria species)	PVC	- <i>Proteases:</i> PVC-Metallopeptidase
t2.178	phosphatidyltransferase				
t2.179	Tox-Glycerophosphoryl diester phosphodiesterase	TIM Barrel (PDB: 1VD6; α/β); HRG, E, ExD, D, H; See Pfam PF03009	<i>Cyanotheca</i> sp. (Cyanobacteria)	PVC	- <i>Proteases:</i> PVC-Metallopeptidase
t2.181	(GDPD)				
t2.182	Miscellaneous toxins				
t2.183	Tox-AB hydrolase2	α/β hydrolase superfamily (α/β); NG, [DE], [KR], HSxG, D, H	acidobacteria, $\alpha,\beta,\gamma,\delta,\epsilon$ -proteobacteria, chlamydiae, fusobacteria, verrucomicrobia, eukaryotes: fungi(ascomycota, basidiomycota), stramenopiles	T2SS, T5SS, T6SS	Imm-SUKH <i>Repeats:</i> FilH, RHS
t2.184	Tox-ODYAM1	All- α ; Several charged residues	α -proteobacteria, bacteroidetes	T2SS (APD1)	- <i>Proteases:</i> Tox-PLOTU; <i>Repeats:</i> Sel1
t2.185	Tox-LatrotoxinCTD	Two conserved α -helices; D, [ST], Y, E	α,γ -proteobacteria, eukaryotes: metazoa (Latrodectus hasseltii, Latrodectus tredecimguttatus)	T2SS	- <i>Proteases:</i> Tox-PLOTU; <i>Repeats:</i> ankyrin
t2.186	Tox-SGS (salivary gland	$\alpha + \beta$; C, C, C, C, [DE]xx[ND]	Eukaryotes: metazoan (crustacea, hexapoda)	T2SS	- <i>Repeats:</i> RHS
t2.187	secreted protein)				
t2.188	Ntox38	All- β ; PXhhG and several hydrophobic residues	Actinobacteria	T2SS, T7SS (WXG)	Imm56 <i>Proteases:</i> Mycosin (Subtilisin)-like protease in the neighborhood
t2.189	Ntox46	$\alpha + \beta$; [KR]STxxPxxDxx[ST], Q	α,γ,δ -proteobacteria, bacteroidetes	T2SS, T6SS	- <i>Repeats:</i> RHS, FilH

t2.190 1. Toxins are grouped and arranged based on the similarity of their known or predicted biochemical functions.

t2.191 2. Where possible, known or predicted folds are described. The folds are further classified as All- α (composed entirely of α -helices), All- β (composed entirely of β -strands), $\alpha + \beta$ (Containing α -helices and β -strands) or

t2.192 α/β (comprising repeated α -helix- β -strand units) depending on the arrangement of their structural elements. Individual conserved residues and motifs are separated by commas. Alternative residues are enclosed in

t2.193 square brackets; 'x' denotes any residue, 'h' indicates a hydrophobic residue (LIYVFMCW).

t2.194 3. By default most lineages are bacterial unless stated otherwise. Eukaryotes and viruses are shown in bold.

t2.195 4. T2SS: Type 2 secretion system; T5SS: Type 5 secretion system, T6SS: Type 6 secretion system, T7SS: Type 7 secretion system. The secretory domains for T7SS are shown next to it in parentheses.

t3.1 **Table 3 Phyletic distribution and associated toxins of Immunity proteins associated with polymorphic toxin systems**

t3.2	Immunity protein	Fold; Conservation ¹	Associated toxins ²	Phyletic distribution	Additional Notes
t3.3	Imm-SUKH	$\alpha + \beta$ (PDB: 3D5P); Several hydrophobic residues and family-specific differences. Refer to previous paper for details	<i>HNH fold families</i> : Tox-SHH, Tox-HNH, Tox-HNH-CIDE, Tox-WHH, Tox-DHNNK, Tox-LHH, Tox-GHH, Tox-HHH, Tox-NucA, Tox-ColE7;	Acidobacteria, actinobacteria, $\alpha\beta\gamma\delta\epsilon$ -proteobacteria, bacteroidetes,, chloroflexi, cyanobacteria, deinococci, firmicutes, fusobacteria, planctomycetes, spirochaetes, synergistetes, verrucomicrobia	This superfamily comprises 5 major families (SUKH1-5), which have been combined in this study; Shows fusions on occasions to toxins and immunity domains; For e.g. fusions to Tox-GHH, Imm-SuFu, Imm33, Imm37, Imm66, Imm67, Imm68, Imm69. Found in homogeneous and heterogenous polyimmunity loci
t3.4			<i>Restriction endonuclease fold families</i> : Tox-REase-4, Tox-REase-3;	Eukaryotes: Giardia, ciliophora, choanoflagellida, fungi, Naegleria, metazoa, stramenopiles, viridiplantae, chlorophyta, eukaryotic viruses	
t3.5			<i>Deaminase families</i> : YwqJ, XOO2897, BURPS668_1122		
t3.6			Proteases: YabG, Tox-PL1;		
t3.7			<i>Other toxins</i> : Tox-EndoU, Tox-DOC, Caspase, Tox-ParBL1, Tox-ComI, Ntox15, Ntox20, Tox-ABhydrolase2, Tox-ABhydrolase3		
t3.8					
t3.9					
t3.10					
t3.11					
t3.12					
t3.13					
t3.14	Imm-SuFu	$\alpha + \beta$ (PDB: 1M1L); GxS, E, E, DxxR	NGO1392-like Tox-HNH fold domain ^a (SuFu-associated nuclease), Tox-GHE ^b , Tox-ParB ^c , Tox-DHNNK ^d , Tox-AHH ^e , Tox-HNH ^f , Tox-EndoU ^g , Tox-EDA39C ^h , Tox-PL-C39-like peptidase ⁱ , Tox-ALF-MPTase ^j , Ntox7 ^k	Acidobacteria, actinobacteria ^{ab,d} , $\alpha, \beta^{a,b,c, f, \gamma, c, d, \delta, e, h, \epsilon}$ -proteobacteria, bacteroidetes ^{b,l,k} , chloroflexi, firmicutes ^{b,e,g} , fusobacteria, planctomycetes, spirochaetes ⁱ , tenericutes verrucomicrobia.	Fused to members of the SUKH family, ankyrin repeats, Imm5, Imm11, Imm33, Imm36, Imm66, Imm67, Imm68, Imm69, PsbP/MOG1. Found in homo- and heterogeneous polyimmunity loci. See Pfam PF05076
t3.15				Eukaryotes: chlorophyta, ascomycota, choanoflagellida, metazoa	
t3.16					
t3.17	Imm-SuFu- family 2	$\alpha + \beta$; [ST]xxG, [DE]	Tox-ColE7 ^a , Tox-DHNNK, Tox-HNH fold ^b , Tox-ALFMPTase ^c , Tox-GDPD ^d	actinobacteria $\alpha^d, \beta, \gamma, \delta, \epsilon$ -proteobacteria, bacteroidetes, cyanobacteria, fibrobacteres, firmicutes ^{a,b} , fusobacteria, gammaproteobacteria, planctomycetes, proteobacteria, spirochaetes ^c , verrucomicrobia	Fused to Imm34, Imm33, Imm66, Imm67, Imm68, Imm69; Found in heterogeneous polyimmunity loci
t3.18	Imm-Cloacin	FKBP-like $\alpha + \beta$; EYSxD, NxG	Tox-ColE3 ^a	Plasmid ^a , ColE6-CT14 ^a , γ -proteobacteria ^a	
t3.19	HEAT repeats	All α ;	Tox-REase-7 ^a	Actinobacteria ^a , bacteroidetes, cyanobacteria, γ -proteobacteria, planctomycetes ^a , verrucomicrobia ^a	
t3.20	Ankyrin repeats	All α ;	Tox-AHH ^a	Firmicutes ^a , planctomycetes ^a , γ -proteobacteria ^a	Fused to SuFu-like immunity domains in firmicutes and found in heterogeneous polyimmunity loci
t3.21	(Imm-ank)				
t3.22	LRR-repeats	α/β ;	Next to T5SS ^a toxins	actinobacteria, bacteria, $\beta, \gamma, \delta, \epsilon$ -proteobacteria, firmicutes, tenericutes	Found in heterogeneous polyimmunity loci
t3.23	Imm-CdiI	Two transmembrane helices; several hydrophobic residues	CdiAC	γ -proteobacteria	
t3.24	Imm-NTF2	NTF2 fold ($\alpha + \beta$); W, W, W	Tox-NucA ^a	Bacteroidetes, $\beta, \gamma^a, \epsilon$ -proteobacteria, firmicutes, fusobacteria, verrucomicrobia	Fused to ankyrin repeats and Imm13 in some proteins
t3.25	Imm-NTF2-2	NTF2 fold ($\alpha + \beta$); Y, W	Tox-JAB-2	γ -proteobacteria (<i>E. coli</i> only)	Although related in structure to Imm-NTF2, the sequences are quite divergent from each other

Table 3 Phyletic distribution and associated toxins of Immunity proteins associated with polymorphic toxin systems (Continued)

t3.26	Imm-PA2201	Two all- α domains(PDB: 2FEF); D, W,GxWxxE, D, YPxD	Tox-REase-1 ^a , Tox-AHH ^b	Bacteroidetes ^a , β ^a , γ ^{a,b} , ϵ ^a -proteobacteria, firmicutes ^a	See Pfam DUF1910 + DUF1911
t3.27	Imm-Barstar	α/β (PDB: 1BRS); DxxxD and several hydrophobic residues	Tox-Barnase-like ribonuclease ^a	Acidobacteria, actinobacteria ^a , α , β^a , γ^a , δ^a , ϵ^a -proteobacteria bacteroidetes ^a , chlamydiae ^a , chloroflexi ^a , cyanobacteria ^a , deinococci ^a , elusimicrobia, firmicutes ^a ,fusobacteria ^a , nitrospirae ^a , planctomycetes ^a , verrucomicrobia, Archaea: euryarchaea ^a , Eukaryotes:dictyosteliida, Naegleria, chlorophyta	See Pfam PF01337
t3.28	Imm-ADP-RGHD; ADP	All- α ; (PDB: 1t5j); D, D[DE], [RK], H	Tox-ART-RSE ^a	acidobacteria, β , γ^a -proteobacteria, firmicutes ^a	See Pfam Pf03747; an example of an enzymatic immunity protein
t3.29	ribosyl glycohydrolase				
t3.30	Imm-NMB0513	wHTH fold ($\alpha + \beta$, PDB: 2O5H); W, W	Ntox20 ^a , Ntox7 ^b	betaproteobacteria ^{a,b} gammaproteobacteria ^a	Corresponds to Pfam DUF596
t3.31	Imm-ComJ	Mostly β ; W, F[DE], PF, Y, Y	Tox-ComI-like competence nuclease ^a	$\alpha^a\beta^a\gamma^a$ -proteobacteria, bacteroidetes ^a , cyanobacteria, firmicutes ^a , Eukaryotes: viridiplantae	
t3.32					
t3.33	Imm-VC0424	$\alpha + \beta$; $\alpha + \beta$ RRM fold, W at C-terminus	-	Firmicutes, fusobacteria, α,β,γ -proteobacteria	Also known as DUF1260 in the Pfam database. Only a subset of members is found in polymorphic toxin systems as potential immunity proteins. These species are listed in column 3
t3.34	Imm-My6CBD	$\alpha + \beta$; E, R, F, W	Tox-ART-HYD1 ^a	actinobacteria ^a , bacteroidetes ^a , firmicutes ^a , fusobacteria, β^a,γ^a -proteobacteria, Eukaryotes: Metazoa	The type VI myosin cargo-binding domain of metazoa appears to have been acquired by lateral transfer from a bacterial version
t3.35	Imm1	$\alpha + \beta$; aromatic and W at C-terminus	SCP1.201 deaminases ^a	Actinobacteria ^a , bacteroidetes, cyanobacteria, firmicutes, planctomycetes α,β,γ -proteobacteria, verrucomicrobia	
t3.36	Imm2	All α ; acidic and hydrophobic residues	BURPS668_1122 deaminases	β , γ - proteobacteria	
t3.37	Imm3	All α ; charged, V	BURPS668_1122 deaminases	Firmicutes	found in heterogeneous polyimmunity loci
t3.38	Imm4	$\alpha + \beta$	SCP1.201 deaminases	<i>Burkholderia pseudomallei</i>	
t3.39	Imm5	Mostly α ; R, D	DYW deaminases ^a , CdiAC ^b	Actinobacteria ^a , bacteroidetes ^a , firmicutes ^a , α,β,γ^a , ^b ,proteobacteria	Fused to Imm36 on occasions
t3.40	Imm6	Mostly α ; P, [DE]	YwqJ deaminases ^a	Actinobacteria ^a , α -proteobacteria, firmicutes ^a	Found in homo and heterogeneous polyimmunity loci
t3.41	Imm7	$\alpha + \beta$; GxaG	Tox-REase-3 ^a	actinobacteria, firmicutes ^a , planctomycetes	
t3.42	Imm8	$\alpha + \beta$; WEa (a:aromatic) at C-terminus	Ntox7 ^a	Acidobacteria, actinobacteria, bacteroidetes ^a , firmicutes ^a , α , β^a , γ^a , δ -proteobacteria	
t3.43	Imm9		Tox-URI2	Bacteroidetes, γ -proteobacteria	

Table 3 Phyletic distribution and associated toxins of Immunity proteins associated with polymorphic toxin systems (Continued)

		α + β; K and several conserved acidic residues			Found in heterogeneous polyimmunity loci
t3.44	Imm10	Mostly β; R and several hydrophobic residues	Pput_2613 deaminase ^a	actinobacteria bacteroidetes chloroflexi firmicutes β, γ ^a , δ, ε-proteobacteria; Eukaryotes: ascomycetes	Lateral transfer to fungi, found in heterogeneous polyimmunity loci
t3.45	Imm11	α + β; several conserved hydrophobic residues	Tox-AHH ^a , Tox-HNH ^b , Tox-SHH ^c	Bacteroidetes ^a , chloroflexi, cyanobacteria, firmicutes ^a , planctomycetes ^a , α, β ^a , γ ^a , δ ^{a, b, c} , ε ^a -proteobacteria spirochaetes ^a verrucomicrobia ^a	Listed in the Pfam database as DUF1629. Fused to SuFu on occasions. Found in heterogeneous and homogeneous polyimmunity loci.
t3.46	Imm12	α + β; several conserved charged and hydrophobic residues	Tox-URI2 ^a	Bacteroidetes ^a , spirochaetes	Found in heterogeneous polyimmunity loci
t3.47	Imm13	α + β; D, D, D, D	Tox-DOC ^a	Actinobacteria, bacteroidetes cyanobacteria, firmicutes, fusobacteria ^a , spirochaetes, verrucomicrobia, α, β, γ, δ-proteobacteria,	Note lateral transfer to eukaryotes. Found in heterogeneous polyimmunity loci. Fused to Imm33 in some instances
t3.48				Eukaryotes: Naegleria	
t3.49	Imm14	Mostly β; several hydrophobic residues	Tox-URI1 ^a , Tox-HNH ^b	Actinobacteria ^a , α, β ^a , γ ^a , δ ^b -proteobacteria, bacteroidetes ^a , chlamydiae ^a , chloroflexi ^a , cyanobacteria, firmicutes ^a , fusobacteria, spirochaetes, verrucomicrobia	Found in heterogeneous polyimmunity loci; Fused to Imm51 in one instance
t3.50	Imm15	α + β; several polar and hydrophobic residues		Bacteroidetes, firmicutes, synergistetes	Found in heterogeneous polyimmunity loci
t3.51	Imm16	α + β; several hydrophobic residues including a highly conserved W	Ntox8 ^a	Actinobacteria, bacteroidetes ^a , β ^a , γ, δ-proteobacteria, firmicutes ^a , planctomycetes, spirochaetes, verrucomicrobia	Also known as DUF2750
t3.52	Imm17	Two TM helices; WxW and a R in the region between them		Bacteroidetes, firmicutes, fusobacteria, spirochaetes	Found in heterogeneous polyimmunity loci
t3.53	Imm18	Mostly β; highly conserved D	Tox-HNH ^a	Actinobacteria ^a , α, β ^a , γ ^a , δ ^a -proteobacteria, bacteroidetes ^a , firmicutes	Found in heterogeneous polyimmunity loci
t3.54	Imm19	α + β; HxxRN motif and several conserved hydrophobic residues	-	Bacteroidetes	Found in heterogeneous polyimmunity loci
t3.55	Imm20	α + β; several conserved hydrophobic residues	Tox-AHH ^a , Tox-ParB ^b	Acidobacteria, actinobacteria, bacteroidetes, β ^{a, b} , γ ^a , δ-proteobacteria, cyanobacterium firmicutes ^a , fusobacteria, planctomycetes, spirochaetes, verrucomicrobia, Eukaryotes: ascomycota	Found in heterogeneous polyimmunity loci. Note presence in ascomycetes
t3.56	Imm21	α + β; absolutely conserved WxG, YxxC and several hydrophobic residues	NGO1392-like HNH fold ^a	Actinobacteria, α, δ-proteobacteria, bacteroidetes, firmicutes ^a , verrucomicrobia	Found in heterogeneous polyimmunity loci
t3.57	Imm22				

Table 3 Phyletic distribution and associated toxins of Immunity proteins associated with polymorphic toxin systems (Continued)

		$\alpha + \beta$; W, Y, and an acidic residue (mostly D)	Cold/E5 fold ^a , Tox-REase-4 ^b , Ntox49 ^c , Ntox14 ^d	Actinobacteria, bacteroidetes ^{a,c} , β, γ -proteobacteria, firmicutes ^{b,d} , fusobacteria, planctomycetes, verrucomicrobia, Eukaryotes: ascomycota	Previously known as SNCF1. Found in heterogeneous polyimmunity loci across a wide range of bacteria
t3.58	Imm23	$\alpha + \beta$; several hydrophobic residues including a WxW motif	Tox-AHH ^a , Tox-REase-7 ^b	bacteroidetes ^a cyanobacteria ^b , firmicutes γ -proteobacteria verrucomicrobia	Some versions fused to Imm11; found in heterogeneous polyimmunity loci
t3.59	Imm24	Mostly α -helical with C-terminal β -hairpin; several hydrophobics including a PxG motif (where x is mostly C)	Tox-AHH ^a , Tox-SHH ^b	Bacteroidetes ^c , $\beta^a, \gamma^a, \epsilon$ -proteobacteria, firmicutes ^{a,b} , verrucomicrobia	found in heterogeneous polyimmunity loci
t3.60	Imm25	$\alpha + \beta$; highly conserved in limited sequences	-	Bacteroidetes	Potential immunity protein found in heterogeneous polyimmunity loci, and a limited phyletic presence
t3.61	Imm26	Mostly α ; R and D and several hydrophobic residues	Tox-URI1 ^a	Actinobacteria, bacteroidetes ^a , β, γ^a, δ -proteobacteria, firmicutes, fusobacteria, planctomycetes, spirochaetes, Eukaryotes: Ascomycota	Note presence in ascomycetes, present in heterogeneous polyimmunity loci
t3.62	Imm27	$\alpha + \beta$; D, GGxP	Ntox10 ^a , Tox-ParB ^b	Actinobacteria, bacteroidetes ^a , β, δ^b -proteobacteria, verrucomicrobia ^a	Wide distribution but sporadic numbers
t3.63	Imm28	Mostly α ; acidic, P,G, R	Tox-WHH ^a , Tox-EndoU ^b , Ntox20 ^c	Actinobacteria, $\alpha^a, \beta^b, \gamma^a$ -proteobacteria	Note presence in <i>Odysella</i> , present in heterogeneous polyimmunity loci
t3.64	Imm29	Mostly α ; R and acidic and several hydrophobic residues	Ntox18 ^a	Actinobacteria, $\alpha^a, \beta^a, \gamma^a$ -proteobacteria, bacteroidetes, firmicutes, fusobacteria	Note presence in <i>Odysella</i> , present in heterogeneous polyimmunity loci
t3.65	Imm30	Mostly α ; Several conserved hydrophobics and DxG motif	Tox-SHH ^a	$\alpha^a, \beta, \gamma^a$ -proteobacteria	Note presence in <i>Odysella</i> . Limited number of hits, present in heterogeneous polyimmunity loci
t3.66	Imm31	All- β ; GxS, [R]	Ntox17 ^a , Ntox7 ^b	$\alpha^a, \beta^b, \gamma^a, \delta$ -proteobacteria, cyanobacteria	Note presence in <i>Odysella</i> . Limited distribution
t3.67	Imm32	$\alpha + \beta$; H, and several conserved residues	Ntox12 ^a , Ntox37 ^b , Ntox7 ^c	$\alpha^a, \beta, \gamma^a, \delta$ -proteobacteria, chlamydiae, bacteroidetes ^b , firmicutes ^a , verrucomicrobia	Note presence in <i>Odysella</i> , chlamydiae. Limited distribution
t3.68	Imm33	Mostly β ; W	Tox-HNH ^a , Tox-DHNNK ^b ; NGO1392-like- HNH ^c	Acidobacteria, actinobacteria, $\alpha, \beta^a, \gamma^a, \delta^c$ -proteobacteria, bacteroidetes, chloroflexi, firmicutes ^b , fusobacteria, planctomycetes, Eukaryotes: dictyosteliida	Also known as DUF2185 in the Pfam database, fused to Imm- SUKH, Imm13, Imm34 and Imm-SuFu, Note presence in dictyosteliida where it is fused to Imm34, present in homo and heterogeneous polyimmunity loci
t3.69	Imm34	Mostly β ; ExxW, C-terminal D	-	Actinobacteria, $\alpha, \beta, \gamma, \delta, \epsilon$ -proteobacteria, bacteroidetes, firmicutes, fusobacteria, planctomycetes, spirochaetes, verrucomicrobia, Eukaryotes: dictyosteliida, heterolobosea, cnidaria	Also known as DUF2314. Fused to Imm-SuFu family 2, Imm33, ankyrin repeats, TM helices, fusion to Imm33 appears to have occurred on multiple occasions independently, present in heterogeneous polyimmunity loci.

Table 3 Phyletic distribution and associated toxins of Immunity proteins associated with polymorphic toxin systems (Continued)

t3.70	Imm35	$\alpha + \beta$; W, [ST]	Tox-PL1 ^a , Ntox40 ^b	Actinobacteria ^{a, b} , bacteroidetes ^a , β, γ -proteobacteria, planctomycetes	Note presence in <i>Naegleria</i> , dictyosteliida and cnidarians. In dictyostellids, it is fused to Imm33
t3.71	Imm36	BH3703-like fold ($\alpha + \beta$); W, W	Tox-NucA ^a , DYW-Deaminase ^b , Ntox40 ^c , Tox-CdiAC ^d , Tox-Caspase ^e	Actinobacteria ^{a, c, e} , $\alpha, \beta, \gamma, \delta$ -proteobacteria, bacteroidetes ^{a, b} , firmicutes ^a , fusobacteria, spirochaetes ^a	Fused to Papain-like toxin and ADP-ribosyl glycohydrolase and Peptidase S8, in some instances. Possible protease inhibitor
t3.72	Imm37	$\alpha + \beta$; ExG	Tox-WHH ^a	Acidobacteria, actinobacteria, $\alpha\beta\gamma\epsilon$ -proteobacteria, bacteroidetes, chloroflexi, cyanobacteria, deinococci, firmicutes ^a , fusobacteria ^a , planctomycetes, verrucomicrobia	Also known as DUF600, fused to Tox-NucA, Imm-SuFu, Imm5, on occasions. Tox-NucA appears to be the primary toxin association. One of the large families. Found in homo and heterogeneous poly-immunity loci. Profile-profile analysis predicts a BH3703-like fold
t3.73	Imm38	Mostly α ; W at N and aromatic residue at C	Ntox19 ^a , NGO1392-like- HNH ^b	Actinobacteria, bacteroidetes ^a , β, γ, δ -proteobacteria, firmicutes ^a , fusobacteria ^a , nitrospirae	Previously known as SNCF2, fused to SUKH in some instances. Found in heterogeneous polyimmunity loci
t3.74	Imm39	$\alpha + \beta$; GR, GxK and several polar and hydrophobic residues	Tox-URI2 ^a	α, γ -proteobacteria	Also known as DUF2247. Found in heterogeneous polyimmunity loci
t3.75	Imm40	$\alpha + \beta$; GGD, F, W	Ntox19 ^a	bacteroidetes ^a , chloroflexi firmicutes, β, ϵ, γ -proteobacteria	Limited distribution
t3.76	Imm41	$\alpha + \beta$; SF, W and several hydrophobic residues	Ntox21 ^a , Ntox29 ^b , Tox-ART-RSE ^c	Actinobacteria, $\beta, \epsilon, \gamma, \delta$ -proteobacteria, firmicutes, planctomycetes	Found in homo- and heterogeneous polyimmunity loci
t3.77	Imm42	$\alpha + \beta$; Several conserved hydrophobic residues	Ntox18 ^a	α, β, γ -proteobacteria, firmicutes ^a	
t3.78	Imm43	α/β ; W, P, D, S, R	Tox-AHH ^a	Bacteroidetes ^a , β -proteobacteria ^a , firmicutes	Found in heterogeneous polyimmunity loci
t3.79	Imm44	$\alpha + \beta$; Multiple polar and hydrophobic residues	Tox-URI1 ^a , Tox-URI2 ^b , Tox-ParBL1 ^c	Bacteroidetes, β -proteobacteria ^{a, b} , firmicutes ^c	Limited phyletic distribution; Found in heterogeneous polyimmunity loci that show variations in structure even between closely related strains
t3.80	Imm45	$\alpha + \beta$; C-terminal W	Tox-ColE3 ^a	bacteroidetes, β, γ, ϵ -proteobacteria, firmicutes	
t3.81	Imm46	$\alpha + \beta$; E, W, E	-	Bacteroidetes, β -proteobacteria	Limited phyletic distribution. Found next to a predicted toxin
t3.82	Imm47	$\alpha + \beta$; KxGDxxK	-	β -proteobacteria, firmicutes	Found in heterogeneous polyimmunity loci
t3.83	Imm48	All- α ; HRG	-	Firmicutes, verrucomicrobia	Found in heterogeneous polyimmunity loci

Table 3 Phyletic distribution and associated toxins of Immunity proteins associated with polymorphic toxin systems (Continued)

t3.84	Imm49	All α ; Hydrophobic residues, P	REase-1 ^a , REase-6 ^b	Actinobacteria ^b , Bacteroidetes ^{a,b} , cyanobacteria ^b , firmicutes ^a , fusobacteria ^a , planctomycetes, β ^{a,b} , δ , γ ^{a,b} -proteobacteria	Also known as DUF556
t3.85	Imm50	Mostly β ; Several hydrophobic residues	Tox-HHH ^a , Ntox24 ^b	actinobacteria, bacteroidetes ^a , firmicutes ^a , planctomycetes, α , β ^{a,b} , γ ^a -proteobacteria, verrucomicrobia	
t3.86	Imm51	$\alpha + \beta$; W, Dx[DE] and several hydrophobic residues	Tox-RES ^a , Tox-URI1 ^b	Actinobacteria, bacteroidetes ^a , β , γ -proteobacteria, cyanobacteria, firmicutes ^b , fusobacteria, spirochaetes	Fused to Imm14 on one occasion, Found in polyimmunity loci
t3.87	Imm52	$\alpha + \beta$; W, GT, F	Tox-REase-5 ^a	Caudoviruses ^a , α , β ^a , γ ^a , δ ^a -proteobacteria	
t3.88	Imm53	$\alpha + \beta$ (Central β -sheet with flanking α -helices); W, WE, PGW, W	Ntox24 ^a , Ntox10 ^b	Acidobacteria, actinobacteria, α , β , γ , δ , ϵ -proteobacteria, bacteroidetes, chlamydiae ^b , cyanobacteria, firmicutes ^a , spirochaetes, verrucomicrobia	
t3.89	Imm54	$\alpha + \beta$; GF, Q	Tox-REase-9 ^a , Tox-RelE ^b , Tox-URI ^c , Tox-REase-4 ^d , Tox-REase-7 ^e , Tox-REase-10 ^f	actinobacteria, bacteroidetes ^{a, c, d} , chlamydiae ^a , firmicutes ^{a, c, d, e} , fusobacteria ^{b, f} , planctomycetes, α , β ^{c, \gamma} ^a , δ , ϵ -proteobacteria, spirochaetes, verrucomicrobia	Found in heterogeneous polyimmunity loci
t3.90	Imm55	$\alpha + \beta$; G and several hydrophobic residues	Tox-SHH ^a	actinobacteria, bacteroidetes ^a , cyanobacteria ^a , firmicutes ^a , lentisphaerae, planctomycetes, α , β , γ ^a -proteobacteria, synergistetes, verrucomicrobia	
t3.91	Imm56	$\alpha + \beta$; D, GR	Ntox38 ^a , Tox-HNH ^b	Actinobacteria ^{a, b} , chloroflexi ^a	
t3.92					
t3.93	Imm57	Mostly α ; D, SE, C	Δ D-peptidase ^a , Tox-Caspase ^b	β ^a , γ ^{a, b} -proteobacteria	
t3.94	Imm58	$\alpha + \beta$; YxxxD, WxG, KxxxE	Unknown toxins with RHS repeats	β , δ -proteobacteria	Limited distribution
t3.95	Imm59	$\alpha + \beta$ (Central β -sheet with flanking α -helices); [DE]R motif	Ntox13 ^a , Ntox40 ^b	firmicutes ^{a, b}	Fused to Imm63 on some instances
t3.96	Imm60	Mostly β ; N, W	Ntox40 ^a , Ntox48 ^b	bacteroidetes	Found in heterogeneous polyimmunity loci
t3.97				firmicutes ^a , fusobacteria,	
t3.98				α ^b , γ ^b -proteobacteria,	
t3.99				euryarchaea	
t3.100	Imm61	$\alpha + \beta$; R	Ntox40 ^a	actinobacteria ^a	
t3.101	Imm62	$\alpha + \beta$; -(mostly E), W	Ntox31 ^a , Ntox48 ^b	Firmicutes ^{a, b} ,	Found in heterogeneous polyimmunity loci
t3.102				γ -proteobacteria	
t3.103	Imm63	$\alpha + \beta$; E + G, -(mostly E)xxY	Ntox40 ^a , Tox-CdiAC ^b , Tox-Arc ^c	actinobacteria ^{a, c}	Found in polyimmunity loci
t3.104				bacteroidetes	
t3.105				firmicutes ^a , β , γ ^{a, b} -proteobacteria	

Table 3 Phyletic distribution and associated toxins of Immunity proteins associated with polymorphic toxin systems (Continued)

t3.106	Imm64	$\alpha + \beta$; DxEA, R motifs	Tox-ColD ^a	Euryarchaea ^a , firmicutes ^a , ϵ -proteobacteria	
t3.107	Imm65	$\alpha + \beta$; YxC, and several charged residues	Tox-JAB1	Bacteroidetes	Contains a signal peptide and a lipbox
t3.108	Imm66	Mostly α ; D, W, F, Y,W	Tox-ABHYDROLASE3 ^a , Ntox48 ^b	Actinobacteria, bacteroidetes, cyanobacteria, firmicutes	Fused to one or more immunity domains such as Imm68, SUKH, Imm-SuFu- family 2, Imm33, Imm69, Imm67, Imm-SuFu, Imm66, and TPR repeats. Some proteins in firmicutes have up to 10 immunity domains
t3.109	t3.110	t3.111		Fusobacteria, $\alpha, \beta, \gamma, \epsilon$ -proteobacteria, spirochaetes, verrucomicrobia, Eukaryotes: Ascomycota, viridiplantae	
t3.112	Imm67	$\alpha + \beta$; W, E, W	-	actinobacteria, bacteroidetes, chloroflexi, cyanobacteria, firmicutes, fusobacteria, planctomycetes, $\alpha, \beta, \gamma, \delta, \epsilon$ -proteobacteria, spirochaetes, verrucomicrobia	Fused to one or more immunity domains such as Imm68, Imm33, Imm-SUKH, Imm-SuFu-family 2, Imm69, Imm-SuFu, Imm66, Imm67, TPR and ankyrin repeats. Some proteins in firmicutes have up to 10 immunity domains
t3.113	Imm68	$\alpha + \beta$; E	-	actinobacteria, bacteroidetes, firmicutes, spirochaetes	Fused to one or more immunity domains such as Imm-SUKH, Imm-SuFu, Imm67, Imm66, Imm-SuFu-family 2, Imm69, Imm33, Imm68 and TPR repeats. Some proteins in firmicutes have up to 10 immunity domains
t3.114	Imm69	$\alpha + \beta$; W,hGE(h: hydrophobic)	Tox-ABhydrolase3 ^a	Actinobacteria, bacteroidetes, firmicutes ^a , fusobacteria, planctomycetes, $\alpha, \beta, \gamma, \epsilon$ -proteobacteria, spirochaetes, verrucomicrobia	Fused to one or more immunity domains such as Imm68, Imm-SUKH, Imm33, Imm-SuFu-family 2, Imm-SuFu, Imm67, Imm66, SP, Imm69 and TPR repeats. Some proteins in firmicutes have up to 10 immunity domains
t3.115	Imm70	$\alpha + \beta$; Y,W	Tox-REase-10 ^a	Acidobacteria, actinobacteria, bacteroidetes, firmicutes ^a , β, γ, ϵ -proteobacteria, spirochaetes ^a , verrucomicrobia	
t3.116	Imm71	Mostly α ; R,F, R	Ntox48 ^a	acidobacteria ^a , β, γ ^a -proteobacteria	Often fused to Imm72
t3.117				Eukaryotes: viridiplantae	
t3.118	Imm72	All- β ; GxxE, WxDxRY, E	Ntox48 ^a	acidobacteria ^a , β, γ ^a -proteobacteria	Often fused to Imm71
t3.119	Imm73	All- α ; Several hydrophobic residues	Tox-PL-2 ^a , Tox-HNH ^b	acidobacteria, actinobacteria ^b , bacteroidetes, cyanobacteria ^a , firmicutes ^a , fusobacteria, β, γ, δ ^a -proteobacteria, verrucomicrobia	Sometimes found in 2–3 tandem copies in a polypeptide
t3.120	Imm74	$\alpha + \beta$; G[DE], [DE]	Tox-Arc ^a	bacteroidetes ^a , firmicutes ^a , planctomycetes, $\alpha, \beta, \gamma, \delta$ -proteobacteria,	Found in heterogeneous polyimmunity loci

t3.121 1. Where possible, known or predicted folds are described. The folds are further classified as All- α (composed entirely of α -helices), All- β (composed entirely of β -strands), $\alpha + \beta$ (Containing α -helices and β -strands) or α/β (comprising repeated α -helix- β -strand units) depending on the arrangement of their structural elements. Individual conserved residues and motifs are separated by commas. Alternative residues are enclosed in square brackets; 'x' denotes any residue.

t3.124 2. Each toxin in column3 that is present in a gene neighborhood along with the corresponding immunity protein in column 1 in the toxin-immunity gene order is marked by a superscript letter, so as to identify the phyletic pattern of this association in column 4.

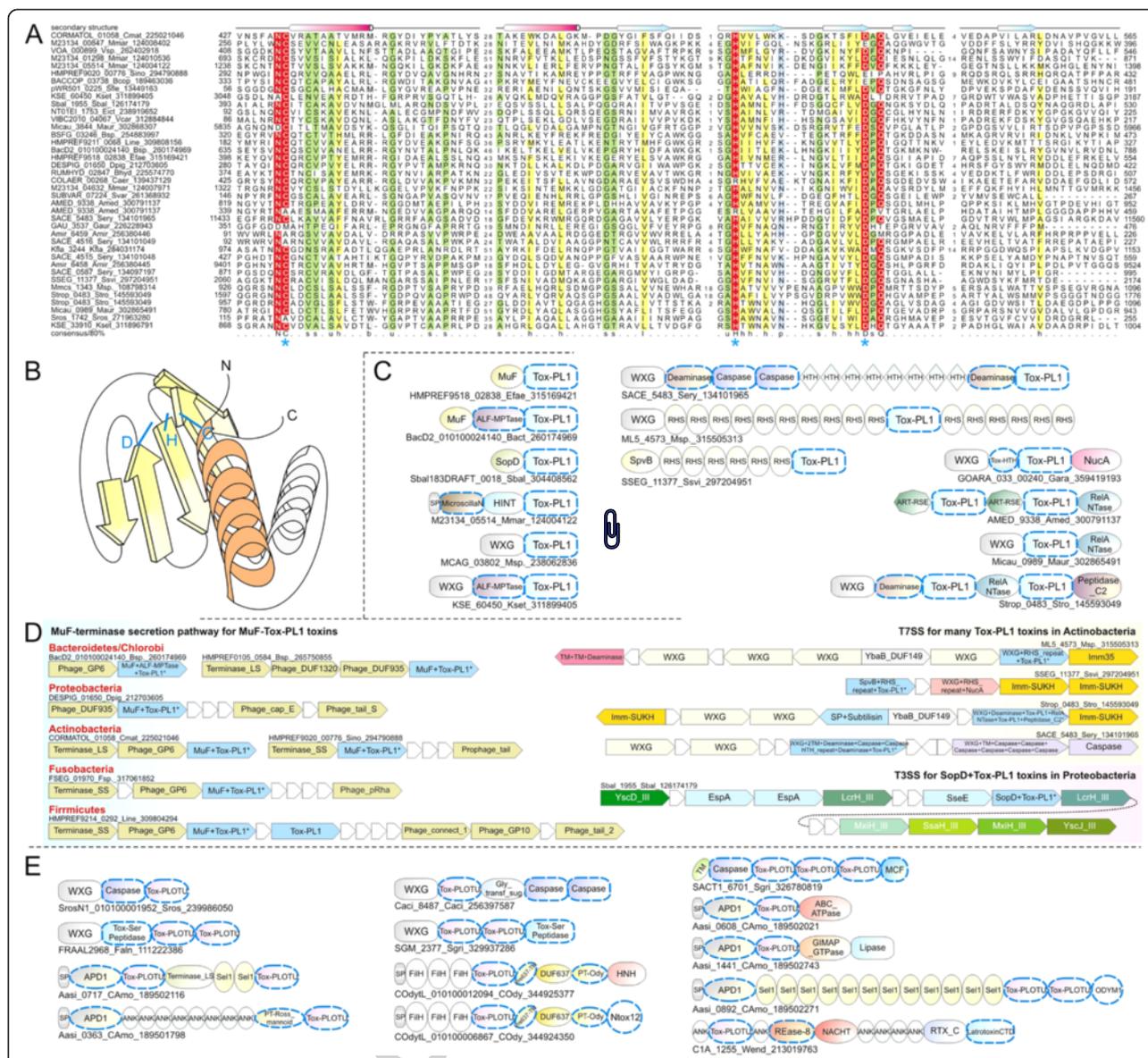


Figure 3 Domain architectures of selected examples of polymorphic toxins containing distinct releasing peptidases: (A) HINT, (B) ZU5, (C) PrsW peptidase, (D) Caspase peptidase, (E) MCF-SHE-like predicted peptidase. The alignment of MCF-SHE domain is shown with predicted catalytic residues marked with blue asterisks. For all alignments in this study, proteins are denoted by their gene name, species abbreviations and GI (Genbank Index) numbers separated by underscores. Secondary structure assignments are shown above the alignment, where the blue arrow represents the β -strand and the red cylinder the α -helix. Poorly conserved inserts are excluded in the alignment and replaced by the length of the insert. Columns in the alignment are colored based on their amino acid conservation at consensus shown below the alignment. The coloring scheme and consensus abbreviations are as follows: h, hydrophobic (ACFILMWWY), l, aliphatic (LIV) and a, aromatic (FWY) residues shaded yellow; b, big residues (LYERFQKMW), shaded gray; s, small residues (AGSVCNDN) and u, tiny residues (GAS), shaded green; p, polar residues (STEDKRNQHC) shaded blue; and c, charged residues (DEHKR) shaded magenta. Absolutely conserved residues are shaded red.

456 **Domains identified as being primarily auto-processing**
 457 **peptidases**
 458 **ZU5 superfamily domains functions as processing**
 459 **autopeptidase in toxins**
 460 The ZU5 (Zona pellucida 5) domain was first identified
 461 as an autoproteolytic domain in the PIDD protein which
 462 forms the core of the PIDDosome, a protein complex in
 463 animals providing a platform for recognizing molecular

patterns that are associated with loss of genomic integ- 464
 rity and genotoxic stress [58]. It is a major player in 465
 p53-induced apoptosis and activation of NF- κ B pathway 466
 in response to DNA damage and its assembly involves 467
 multiple autoproteolytic cleavages mediated by its two 468
 ZU5 domains [59]. Our structural comparisons with the 469
 DALIite program and sequence profile searches 470
 revealed that the ZU5 domain is homologous to the GPS 471

472 domain involved in autoproteolytic cleavage of the
473 polycystin-1 and certain G-protein-couple receptors
474 [60], and the autoproteolytic domain of the nuclear pore
475 Nup96/98 proteins [61]. All these domains are charac-
476 terized by the presence of a C-terminal CxH motif which
477 forms their thiol autopeptidase active site (Additional
478 File 1). Accordingly, we include all these domains in the
479 ZU5 superfamily. Our iterative sequence searches identi-
480 fied ZU5 domains in several potential polymorphic tox-
481 ins: They are typically located at the N-terminus of large
482 proteins with central RHS repeats (Figure 3B). In poly-
483 morphic toxins, the ZU5 domain is most frequently
484 associated with the SpvB and β -propeller domains sug-
485 gesting that it might be functionally coupled to the
486 TcdB/TcaC-like export pathway [42,62]. Its N-terminal
487 location is notably different from the previously
488 observed HINT autopeptidase domains of polymorphic
489 toxins which are instead found at the C-terminus close
490 to the toxin domain [17] (Figure 3B). This suggests that
491 the autoproteolytic activity of the two peptidases have
492 distinct functions – the ZU5 autopeptidase most likely
493 cleaves the toxin at the base of the filamentous structure
494 in order to release it at the cell surface during its extru-
495 sion by the TcdB/TcaC system. In contrast, the C-
496 terminally located HINT autopeptidase is likely to be
497 critical for the release of just the toxin domain, probably
498 upon contact with the target cell. In the classical poly-
499 morphic toxins ZU5 autopeptidases are found in associ-
500 ation with a diverse array of nuclease and peptidase
501 toxin domains (Figure 3B). Related ZU5 domains are
502 also found in several other large bacterial cell surface
503 proteins, which additionally contain diverse adhesion
504 modules and other enzymatic domains, such as glycohy-
505 drolases, lipases and phosphodiesterases (Additional File
506 1). Thus, ZU5 autoproteolytic processing might be a
507 more general feature among bacterial surface proteins
508 that are deployed for the degradation or remodeling of
509 extracellular biopolymers and matrices.

510 *PrsW* peptidase family defines a novel secretion pathway to 511 release C-terminal toxin domains

512 The PrsW family of membrane-embedded peptidases is
513 prototyped by the enzyme catalyzing site-1 cleavage of
514 anti- σ^W factor RsiW in *Bacillus subtilis* [43]. Most
515 representatives bear eight transmembrane helices and
516 four conserved motifs (Figure 3), which show distant re-
517 lationship to several other peptidase families like CPBP
518 and APH-1 [63]. Given that the active site of the PrsW
519 is located within the membrane-spanning helices
520 (Figure 3C), it is likely that they also form a transmem-
521 brane conduit for the simultaneous extrusion and pro-
522 cessing of the toxin. We first recognized the PrsW
523 domain as being a potential processing peptidase in
524 polymorphic toxins on account of its N-terminal fusion

with a novel deaminase toxin domain of the DYW clade 525
(gi: 320532150) [18]. Further analysis revealed that N- 526
terminal PrsW domains are associated with a diverse 527
array of toxin domains, including several distinct ver- 528
sions of the restriction endonuclease superfamily 529
(Figure 3C), mainly in Gram-positive bacteria. These 530
toxin domains are typically connected by a short linker 531
to the core membrane-spanning PrsW domain. How- 532
ever, in certain cases the toxin domain might be con- 533
nected via a long filamentous structure formed by RHS 534
repeats to the N-terminal PrsW domain (e.g. in a *Strep-* 535
tomyces violaceus protein with a novel toxin domain 536
(Ntox9; gi: 307326465). Thus, the PrsW domain might 537
be used to autoproteolytically process polymorphic tox- 538
ins both of the soluble secreted type (one with short lin- 539
kers) and of the filamentous contact dependent type 540
(with RHS repeats). In archaea (e.g. *Pyrococcus horikoshi* 541
PH0065) and fungi (e.g. *Aspergillus fumigatus*; gi: 542
146324562), the PrsW peptidase domains are respect- 543
ively fused at their N-termini to another PrsW-like pep- 544
tidase (DUF2324 in PFAM), or a ceratoplatanin domain 545
that is found in secreted phytotoxic virulence factors of 546
fungal pathogens [64]. It is conceivable that in these 547
examples the PrsW domain has been recruited for the 548
processing of potential N-terminal toxins that are used 549
against more distantly related organisms or plant hosts. 550
In several bacteria the PrsW domain is fused to intracel- 551
lular signaling domains such as the PilZ domain which 552
recognizes cyclic diguanylate, cyclic nucleotide binding 553
domains, phosphopeptide-binding FHA domains and 554
Zn-ribbon domains [65] (Additional file 1). These ver- 555
sions can be clearly distinguished both in terms of their 556
sequence relationships and domain architectures from 557
those associated with toxin domains. These are more 558
likely to function as signaling peptidases that cleave pro- 559
teins in conjunction with signals sensed by the asso- 560
ciated domains. 561

562 **Peptidase domains that function both in auto-processing** 563 **and as toxins**

564 *Caspase-like peptidases*

565 As noted above, peptidases of the caspase-like superfam- 566
ily [66] (also known as “clan CD” [67]) were originally 567
identified as processing peptidases of diverse host- 568
directed toxins (e.g. RTX toxins) of pathogenic bacteria 569
[49,57]. Likewise, some of these domains were identified 570
in certain large bacterial surface proteins where they 571
might function as autoproteolytic processing domains 572
[52]. Other secreted bacterial members of this fold, such 573
as the clostripains have been implicated in proteolytic 574
processing of surface proteins, whereas the gingipains 575
act as virulence factors that cleave host proteins [47]. In 576
this study we obtained evidence based on domain archi- 577
tectures and gene neighborhoods that the caspase-like

578 peptidase domains occur both as potential processing
579 peptidases (typically internal domains) and as toxin
580 domains (the C-terminal-most domain) in polymorphic
581 toxins from bacterial lineages such as bacteroidetes,
582 gammaproteobacteria and actinobacteria (Figure 3C).
583 Architectural analysis clearly shows that the caspase do-
584 main toxins might be delivered via the T7SS, PVC-SS,
585 TcdB/TcaC-like export pathway, in addition to the T2SS
586 (Figure 3C). Versions of the caspase-like domain that
587 are likely to function as processing peptidases of poly-
588 morphic toxins usually occur just upstream of a distinct
589 C-terminal toxin domain, in a position similar to the
590 HINT autopeptidase domains in other polymorphic tox-
591 ins (Figure 3), suggesting that they might similarly aid
592 in the autoproteolytic release of the toxin domain.
593 Architectural analysis suggests that the caspase-like
594 peptidase might be nearly as prevalent as the HINT
595 peptidase in proteolytic processing of polymorphic tox-
596 ins (Additional File 1). Certain other toxin proteins have
597 an array of repeats of the caspase-like domain upstream
598 of the C-terminal toxin domain (e.g. a protein from
599 *Streptomyces flavogriseus* with ADP-ribosyltransferase
600 and MCF peptidase toxin domains; gi: 357410654; see
601 below) (Figure 3C), suggesting that their processing
602 might involve multiple autoproteolytic events to release
603 multiple cleavage products. Some of the caspase domain
604 repeats in these proteins lack the catalytic residues and
605 might merely play a structural or peptide-binding role.

606 **Papain-like peptidases**

607 Papain-like peptidase domains, which constitute the
608 most diverse and widespread superfamily of thiol pepti-
609 dases, have been previously recorded as the toxin
610 domains of both exotoxins and those delivered into the
611 host cells by various pathogenic bacteria. Examples of
612 the former include the *Streptococcus pyogenes* exotoxin
613 SpeB, while those of the latter include the *Pseudomonas*
614 *syringae* AvrPphB toxin, which cleaves the plant serine/
615 threonine kinase PBS1, and the *Pasturella multocida*
616 toxin PMT [68-70]. We found evidence for domains
617 belonging to multiple distinct clades of the papain-like
618 superfamily in polymorphic toxin polypeptides.

619 The first of these, the Tox-PL1 (Tox-papain-like-1)
620 family was recovered as a previously unknown conserved
621 domain in several predicted polymorphic toxins, usually
622 secreted by way of the T7SS (i.e. with N-terminal WxG
623 domains) and TcdB/TcaC-like system (N-terminal SpvB
624 domain) in actinobacteria, and bacteroidetes. Examin-
625 ation of its multiple alignment revealed a conserved NC-
F4 626 H-DxQ signature (Figure 4A), which is reminiscent of
627 the conservation pattern seen in papain-like peptidases
628 [53,71,72]. This relationship was confirmed via profile-
629 profile comparisons with the HHpred program that sig-
630 nificantly recovered papain-like peptidases ($p = 10^{-5}$; 95%

probability). In a subset of the predicted polymorphic 631
toxins Tox-PL1 is the only catalytic domain, and occurs 632
at the extreme C-terminus of the toxin polypeptide, sug- 633
gesting that it is the toxin domain (Figure 4C). In other 634
cases it occurs in internal positions in polypeptides bear- 635
ing a diverse set of toxin domains [18], or in the middle 636
of an array of filament-forming RHS repeats (Figure 4C). 637
In these cases it is likely to function as an auto- 638
processing peptidase that releases associated toxin 639
domains comparable to the HINT and caspase-like pep- 640
tidases [17]. In *Shewanella* we observed a protein combin- 641
ing a SopD domain [73] with a C-terminal Tox-PL1 642
domain, which is encoded by a gene embedded within a 643
T3SS operon. Given that *Shewanella* is known to sup- 644
press the growth of competing distantly related bacteria 645
and infect eukaryotic hosts [74], it is possible that this 646
protein might be used as a toxin delivered by the T3SS 647
in such conflicts. In diverse bacteria we observed a dis- 648
tinctive architecture of Tox-PL1, wherein it is fused to 649
the MuF domain (Figure 4C), which we had previously 650
characterized as a DNA-packaging protein of bacterio- 651
phages utilizing the portal-terminal system [75]. Gene- 652
neighborhood analysis indicated that these are encoded 653
by prophage remnants that also include the terminase, 654
portal protein and capsid protein genes (Figure 4D). 655
Additionally, several of these neighborhoods might en- 656
code proteins with previously noted *bona fide* toxin 657
domains that operate on nucleic acids (e.g. the HNH nu- 658
clease; Figure 4)[17,18]. Hence, we propose that these 659
gene neighborhoods represent a novel phage-derived 660
secretory mechanism, distinct from the previously iden- 661
tified T6SS and PVC-SS that utilizes a capsid packaging- 662
like mechanism. It is conceivable that in these systems 663
the toxins encoded by associated genes are loaded into a 664
capsid-like structure that is then delivered to target cells. 665
Here, the Tox-PL1 domain might be involved in proces- 666
sing proteins either during the assembly of the secretory 667
structure or the release of toxins into target cells. 668

The second major family of papain-like peptidases 669
with potential processing as well as toxin functions are 670
those belonging to the OTU family [53,76] (Figure 4E). 671
These enzymes have been studied mainly in eukaryotes, 672
where they function as deubiquitinating enzymes (DUBs) 673
[77]. We found evidence for a diverse set of OTU pep- 674
tidase domains in potential polymorphic toxins delivered 675
by the T7SS (with N-terminal WxG domains) in actino- 676
bacteria and via T2SS in the *Acanthamoeba* endosymbi- 677
ont *Odyssella thessalonicensis* [78]. In these bacterial 678
lineages they occupy positions suggestive of both proces- 679
sing and toxin functions (Figure 4E). Additionally, we 680
found related OTU-like peptidases in large proteins re- 681
sembling polymorphic toxins in several endo- symbiotic/ 682
parasitic bacteria of animals and amoebozoans, such as 683
Amoebophilus, *Waddlia* and *Wolbachia*. However, in 684

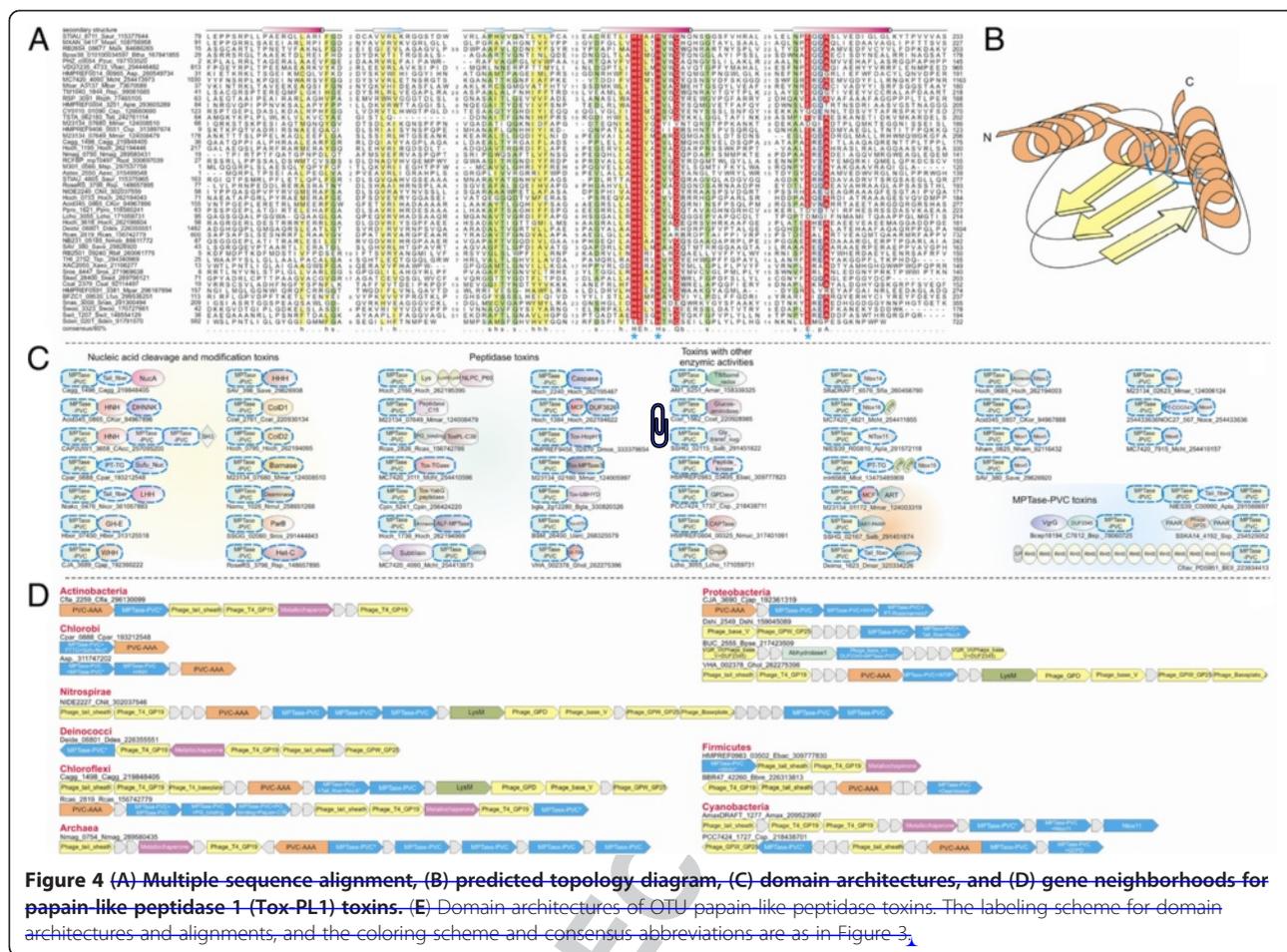


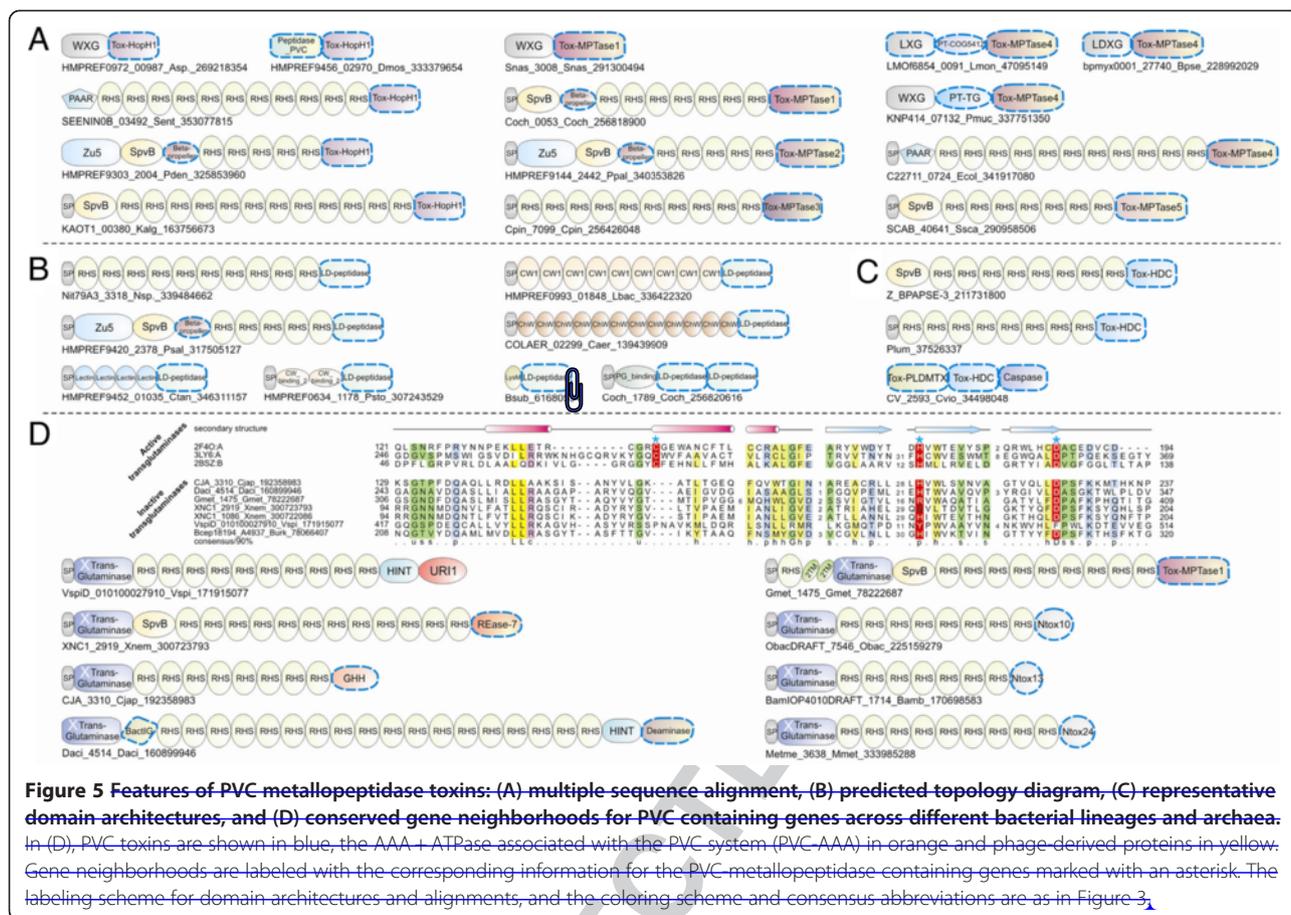
Figure 4 (A) Multiple sequence alignment, (B) predicted topology diagram, (C) domain architectures, and (D) gene neighborhoods for papain-like peptidase 1 (Tox-PL1) toxins. (E) Domain architectures of OTU papain like peptidase toxins. The labeling scheme for domain architectures and alignments, and the coloring scheme and consensus abbreviations are as in Figure 3.

685 these organisms their gene-neighborhoods suggest that
 686 they are unlikely to be polymorphic toxins used in intra-
 687 specific conflicts; rather, they are likely to be used against
 688 their host. In several cases, the OTU-like domains of
 689 these intracellular bacteria occur at the extreme C-
 690 terminus of large proteins with several domains, includ-
 691 ing repeats forming extended structures such as the Sell1,
 692 ankyrin and TPR repeats (Figure 4E). This suggests that
 693 they might be deployed similar to the classical poly-
 694 morphic toxin, but within the host cell. In other proteins
 695 from the same group of bacteria they might occur as in-
 696 ternal domains accompanied by several other potential
 697 toxin domains (Figure 4E), such as GIMAP GTPase, lip-
 698 ase, latrotoxin-C and Tox-MCF1-SHE (see below). The
 699 preponderance of these OTU-like peptidase domains in
 700 intracellular bacteria suggests that they might function as
 701 toxins that suppress the Ub-dependent anti-pathogen
 702 mechanisms of their eukaryotic hosts due to DUB activ-
 703 ity [79,80]. Indeed, a comparable role was originally pro-
 704 posed for the OTU-like peptidases in chlamydiae [53,76].
 705 However, their presence in free-living bacteria (e.g. di-
 706 verse actinobacteria) indicates that a subset of these
 707 OTU-like peptidase proteins might function as either as

processing-peptidases that autoproteolytically process 708
 polypeptides or as conventional toxin domains that 709
 cleave proteins in rival cells. 710

PVC secretory system-type metallopeptidase domains 711

The “*Photorhabdus virulence cassette*” or PVC-SS was 712
 originally identified as a prophage-derived secretory sys- 713
 tem in *Serratia entomophila*, where it delivers toxins that 714
 confer strong anti-feeding activity against the infected 715
 grass grub beetle larvae [41] and in *Photorhabdus*, where 716
 it extrudes toxins that destroy insect hemocytes by induc- 717
 ing actin condensation [40]. This system is typified by 718
 several caudate phage-derived gene products, such as the 719
 tail sheath protein and gp19 (these two form the tail tu- 720
 bule), gp25 (forms the baseplate), and a distinct clade of 721
 AAA + ATPases that are related to CDC48 [81]. Thus, 722
 the PVC-SS parallels the T6SS in being derived from the 723
 tails of prophages, but differs from it in terms of the 724
 associated AAA + ATPase, which in the case of T6SS is a 725
 member of the ClpB clade of AAA + ATPases (ClpV) 726
 [39,81,82]. Hence, these two systems represent independ- 727
 ent prophage-based innovations that have recruited dis- 728
 tinct sets of AAA + ATPases to facilitate recycling of the 729



730 injection apparatus after it has been deployed. We
 731 observed in our recent studies that several toxin
 732 domains closely related to those found in polymorphic
 733 toxins are secreted via the PVC-SS across most major
 734 bacterial lineages and certain euryarchaea (Figure 5).
 735 Our preliminary analysis of these toxin proteins secreted
 736 via the PVC-SS revealed that they contained a conserved
 737 metallopeptidase domain that occurred N-terminal to
 738 the toxin domain [17,18]. A more detailed analysis in
 739 course of this study indicated that this metallopeptidase
 740 domain is a pervasive feature of the PVC-SS and pro-
 741 vides an excellent marker to identify novel toxins
 742 secreted via this system. Accordingly, we term it the
 743 PVC-metallopeptidase (Figure 5). This domain is charac-
 744 terized by a highly conserved HExxHxxQ-E signature
 745 and profile-profile comparisons using HHpred recovered
 746 several zincin-like metallopeptidases as the best hits (e.g.
 747 PDB: 2vqx, 1u4g, 3cqb; $p < 10^{-5}$; >90% probability). A
 748 multiple alignment based on these hits suggests that the
 749 PVC-metallopeptidase adopts a similar structure with
 750 three beta-strands and three alpha helices, with the con-
 751 served histidines on the second helix and glutamate on
 752 the third helix forming the Zn-dependent active site [83]
 753 (Figure 5A, B).

Our analysis of the domain architectures of PVC-
 metallopeptidase proteins affirmed their general resem-
 blance to the classical polymorphic toxins: the strongly
 conserved metallopeptidase domain occupied the N-
 terminal region, followed in each protein by highly vari-
 able C-termini, each of which usually corresponded to a
 different family of toxin domains. Thus, they appear to
 have evolved through a recombination process compar-
 able to that of the polymorphic toxins, which combined a
 “constant” N-terminal peptidase with variable C-terminal
 toxin domains (Figure 5C). This positional polarity of
 the PVC-metallopeptidase domains with respect to the
 associated toxin domains resembles that of the HINT,
 PrsW, caspase-like and papain-like peptidases, indicat-
 ing that they are likely to act as autoprolytic domains
 that release the toxin after or during its export by the
 PVC-SS [17,18]. The C-terminal toxin domains asso-
 ciated with the PVC metallopeptidases span an extraor-
 dinary diversity and include numerous, structurally
 unrelated nucleases, nucleic acid deaminases, peptidases,
 pore-forming domains and several other enzymatic
 domains (Figure 5C). There are multiple toxins with the
 PVC architecture in several bacteria and archaea (e.g.
Halogeometricum borinquense; Additional File 1), with a

778 high diversity of C-terminal toxin domains similar to
779 those found in conventional polymorphic toxins. Our
780 analysis also showed that the PVC toxins are not limited
781 to pathogenic or symbiotic bacteria but are abundant in
782 several free-living bacteria (e.g. the cyanobacterium
783 *Microcoleus chthonoplastes* and *Nitrosococcus oceani*)
784 and archaea (e.g. *Halogeometricum borinquense*). This
785 suggests that the PVC-SS toxins are not exclusively used
786 against host but might also be used in inter-bacterial
787 conflicts, just like the T6SS [15,30,39]. However, a not-
788 able proportion of the PVC-SS dependent systems, un-
789 like conventional polymorphic toxin systems, lack
790 adjacent genes encoding immunity proteins (Figure 5C).
791 This might imply the activity of PVC toxins is primarily
792 directed against distantly related organisms.

793 In addition to the above cases, we observed instances
794 where a second PVC-metallopeptidase domain occurred
795 at the extreme C-termini of proteins in a position com-
796 parable to the toxin domain (Figure 5C). Consistent with
797 this, domain architecture and gene-neighborhood analysis
798 showed that the PVC-metallopeptidase indeed also occurs
799 as a toxin domain of certain polymorphic toxins, pre-
800 ceded by an array of RHS repeats (e.g. a protein from the
801 verrucomicrobium *Pedospira parvula*; gi: 223934413;
802 Figure 5C). Similarly, the PVC-metallopeptidase domain
803 might occur as a C-terminal domain fused to a T6SS
804 phage base-plate/tail polypeptide (e.g. *Burkholderia* sp.;
805 gi: 78060725) (Figure 4). These examples suggest that in
806 addition to its predominant role in autoproteolytically
807 processing PVC toxins, this metallopeptidase might take
808 on the role of a peptidase toxin in several cases.

809 **The MCF1-SHE domain: A possible novel serine peptidase** 810 **shared by polymorphic toxins and secreted effectors?**

811 We initially identified this domain as a conserved region
812 shared by certain predicted polymorphic toxins (e.g.
813 Caci_8529 from the actinobacterium *Catenulispora acid-*
814 *iphila*) and PVC-SS toxins (e.g. Hoch_1384 *Haliangium*
815 *ochraceum*). Iterative sequence profile searches with the
816 PSI-BLAST program recovered homologous regions in
817 proteins from a diverse group of bacteria and the mimi-
818 virus (L389, gi: 311977774) prior to convergence. These
819 proteins include the MCF1 (makes caterpillars floppy)
820 [84] and FitD entomotoxins, respectively from *Photo-*
821 *rhabdus luminescens* and *Pseudomonas fluorescens* [85-
822 87], and the phytotoxin of *Pseudomonas syringae*
823 HopT1-1 which is secreted via the T3SS [88,89]. A mul-
824 tiple alignment of this domain revealed that its core com-
825 prises of two kinked helices, predicted to form a hairpin
826 (Figure 3E). The predicted kinks in the two helices are re-
827 spectively associated with a conserved serine and a
828 HxxxE motif and are likely to face each other. Accord-
829 ingly, we named this domain the MCF1-SHE domain for
830 the first characterized protein that bears it and the

conserved triad of residues. While this domain does not
resemble any previously known domain, the above cata-
lytic triad suggests that it could potentially function as a
novel serine peptidase. In several cases its occurrence at
the extreme C-termini of polymorphic toxin proteins
points to a potential toxin function for the MCF1-SHE
domain (Figure 3E). Consistent with this, it is also found
in several secreted proteins of both extracellular patho-
gens such as *Edwardsiella* and *Xenorhabdus*, and intra-
cellular bacterial and viral pathogens such as *Legionella*,
Coxiella burnetii and *Yersinia pseudotuberculosis* and the
mimivirus (Figure 3E). In particular it appears to have
expanded in legionellae, where up to four distinct MCF1-
SHE toxin paralogs might be present per organism. This
phyletic pattern suggests that MCF1-SHE proteins might
be both toxins in intra-specific conflict and also import-
ant effectors that have dispersed through lateral transfer
across phylogenetically diverse pathogens. Certain do-
main architectures of the MCF1-SHE domain are consis-
tent with the predicted peptidase role, although in a
different capacity. It often occurs just upstream of several
toxin domains, such as the ADP ribosyltransferase
domains related to those found in the *Pseudomonas syr-*
ingae HopU1 phytotoxin (Figure 3E). In these cases, it
could function as a potential processing peptidase that
releases the C-terminal toxin. Similarly, in actinobacteria,
it is embedded in gigantic proteins (>10,000 amino acids
in length) with other peptidase domains such as the
anthrax-lethal factor metallopeptidase, caspase-like and
OTU domains (e.g. gis: 345002682, 326780819).

831 **Other peptidases that function predominantly as toxin** 832 **domains of polymorphic toxin proteins**

833 Besides the above discussed domains, we uncovered sev-
834 eral other peptidase domains that are clearly predicted
835 to function as toxin domains rather than as processing
836 peptidases on the basis of their domain architectures
837 (Table 2). In addition to classical polymorphic toxin sys-
838 tems and PVC-SS delivered toxins, these peptidase toxin
839 domains are also found in several host-directed effectors
840 of pathogenic bacteria. However, it should be noted that
841 outside of these toxin systems, related peptidase
842 domains might perform other unrelated functions.

843 **Papain-like peptidases**

844 Several of the peptidases predicted to function as the
845 toxin domains of classical polymorphic and PVC-SS
846 delivered toxins belong to a number of distinct clades
847 from the papain-like superfamily (Figure 3, 5): 1) The
848 NlpC/P60 clade – peptidases of this clade were first
849 recognized as enzymes that cleaved peptide bonds in
850 peptidoglycan and are nearly universally distributed
851 across bacteria and also found in several bacteriophages
852 [71]. We recovered such peptidase toxins in proteins

883 such as Hoch_2166 from the myxobacterium *Halian-*
884 *gium* (gi: 262195395, [Figure 5C](#)); by analogy to other
885 members of the NlpC/P60 clade they are predicted to
886 function by degrading cell-walls of target cells. 2) The
887 Tox-transglutaminase domain (Tox-TGase) – In
888 addition to toxins from free-living bacteria, this transglu-
889 taminase domain is also found in toxins delivered by dif-
890 ferent secretory systems of parasitic bacteria, where they
891 appear to be directed against the host cells. In particular,
892 it is the toxin domain of T3SS effectors directed against
893 plants, such as AvrPphE *Pseudomonas syringae* (gi:
894 30231092) and related effectors of *Ralstonia*, *Xanthomonas*
895 and *Acidovorax*, in RTX toxins directed against animal
896 hosts (e.g. *Vibrio caribbenthicus* RtxA; gi:
897 312885249) and in a novel secreted effector of *Legionella*
898 *pneumophila* (lpg2408; gi: 52842617). These enzymes
899 might either catalyze a conventional thiol peptidase reac-
900 tion or act as transglutaminases that mediate crosslink-
901 ing of proteins via a transglutaminase reaction [53].
902 Alternatively, they could catalyze polyamination of target
903 glutamine, as has been observed in the case of the *Bor-*
904 *datella pertussis* transglutaminase that modifies the
905 mammalian RhoA GTPase [90]. 3) The Tox-PL-C39 do-
906 main – these peptidase domains are related to the C39/
907 ComA-like peptidase domains that cleave the leader-
908 peptides of certain proteins secreted by ABC transpor-
909 ters such as the bacteriocins ([Figure 5C](#)) [91,92]. 4)
910 Papain-like peptidases Tox-PL2 and Tox-PL3 – these
911 are novel peptidase domains that we identified in this
912 study and the former is prototyped by the toxin domain
913 of a polymorphic toxin from *Sorangium cellulosum* (gi:
914 162456110, [Figure 3A](#)) and the latter by a polymorphic
915 toxin from *Prevotella sp.* (gi: 260911294, [Figure 3B](#)).
916 Thus far, such peptidase domains are not found outside
917 of polymorphic toxin systems and are typified by a C-H-
918 D catalytic triad. 5) We also detected a toxin domain
919 with a papain-like peptidase belonging to the classical
920 ubiquitin C-terminal hydrolase (UBCH/UBHYD) clade
921 associated with the PVC-SS in the plant pathogenic bac-
922 terium *Burkholderia gladioli* (gi: 330820326, [Figure 5C](#)).
923 Similar UBCH domains are also found in potential tox-
924 ins secreted by a variety of other bacterial endosym-
925 bionts of amoebae such as *Simkania negevensis*,
926 *Waddlia chondrophila*, *Amoebophilus asiaticus* and *Pro-*
927 *tochlamydia amoebophila* and giant nucleocytoplasmic
928 DNA viruses that infect them (Additional File 1). These
929 predicted toxins display no associated immunity proteins
930 suggesting that like the OTU domains of pathogens and
931 endosymbionts, they are likely to function as DUBs that
932 deubiquitinate eukaryotic target proteins [79].

933 **Metallopeptidases**

934 Beyond the *toxin versions* (as opposed to autoproteolytic
935 processing versions) of the PVC-metallopeptidase

domain described above, we recovered several other dis- 936
tinct clades of the Zincin-like metallopeptidase super- 937
family that are predicted to function solely as toxin 938
domains in classical polymorphic and PVC-SS toxin pro- 939
teins ([Figure 6](#)). These include: 1) The anthrax lethal 940
factor-like metallopeptidase (ALF-MPTase) domains [48] 941
that are found primarily among PVC-SS delivered toxins 942
(e.g. Hoch_1736 from *Haliangium*; gi: 262194969, [Fig-](#) 943
[ure 5C](#)). 2) The HopH1-like metallopeptidase domain 944
([Figure 6A](#))—this domain is also found in several plant- 945
directed T3SS-delivered effectors, such as *Pseudomonas* 946
syringae HopH1 (gi: 28867816), and the animal-directed 947
T3SS effectors such as *Citrobacter rodentium* and enter- 948
opathogenic and enterohemorrhagic *Escherichia coli* 949
NleD that blocks apoptosis of mammalian cells [93,94]. 950
3) We also identified five smaller families of previously 951
unknown zincin-like metallopeptidases (Tox-MPTase1- 952
5) that are exclusively found in polymorphic toxins from 953
phylogenetically diverse of bacteria ([Figure 6A](#)). In gen- 954
eral terms they are similar in size and distantly related 955
to the Wss1-like desumoylating metallopeptidase of 956
eukaryotes [95]. All of these are typically associated with 957
N-terminal RHS repeats and at least in the case of a 958
polymorphic toxin with a Tox-MPTase4 domain from *E.* 959
coli, it might be delivered via the T6SS. 960

961 **Other miscellaneous peptidases**

Beyond these, we also recovered domains in PVC-SS and 962
polymorphic toxins belonging to the L,D-peptidase, 963
pyroglutamyl-peptidase [96] and YabG peptidase families 964
[97]. Of these, the L,D peptidase domain is a distinct thiol 965
peptidase domain with a β -barrel catalytic domain that is 966
unrelated to the papain-like peptidases ([Figure 6B](#)) [98,99]. 967
It has been shown that the classical cell-wall associated 968
LD-peptidase domain catalyzes a transpeptidase reaction 969
that cleaves the peptide bond between L-Lys3-D-Ala4 in 970
peptidoglycan while concomitantly forming a crosslinking 971
peptide bond between the COOH group of L-Lys3 and the 972
NH2 group of the D-isoasparagine linked to the ϵ - 973
NH2 group of Lys3 from an adjacent chain [98]. Cell-wall 974
associated L,D-peptidases are found in most major 975
lineages of bacteria and are likely to play a role in the re- 976
modeling of peptidoglycan especially in face of antibiotics 977
that inhibit cross-linking. Polymorphic toxins with L,D- 978
peptidase domain are distinguished from the typical 979
cell-wall associated L,D peptidases by their distinct archi- 980
tecture with RHS repeats and genomic organization with 981
linked immunity proteins. It is likely that the toxin L,D- 982
peptidases act by hydrolyzing L-Lys3 crosslinks with D- 983
amino acids, thereby compromising the integrity of the 984
cell-wall. 985

The bacteriophage APSE of the endosymbiont *Hamil-* 986
tonella defensa, which protects aphids and other sap- 987
feeding insects against parasitoid wasps, encodes several 988

999 likely to be similar to the Rossmannoid three-layered
1000 sandwich adopted by the caspases and the flavodoxin-
1001 like fold. The absolutely conserved H, D/N and C are
1002 predicted to lie at the ends of the three successive
1003 strands of this structure and are likely to comprise the
1004 catalytic triad of the peptidase active site. Accordingly
1005 we named this domain Tox-HDC and predict that it
1006 might function as a thiol peptidase or a transglutami-
1007 nase. Proteins bearing this predicted toxin domain are
1008 particularly common in both intracellular (e.g. *Coxiella*
1009 *burnetii*) and extracellular (e.g. *Xenorhabdus nemato-*
1010 *phila* and *Photorhabdus luminescens*) pathogens and
1011 typically lack associated genes coding for immunity pro-
1012 teins. Thus, these toxins appear to be primarily directed
1013 against distantly related targets such as eukaryotes.

1014 In conclusion, at least 23 distinct clades of peptidases
1015 belonging to several structurally unrelated superfamilies
1016 have been recruited as toxins, and are often shared be-
1017 tween polymorphic toxins and host-directed effectors
1018 from diverse plant and animal pathogens. This suggests
1019 that several of these peptidase domains have evolved
1020 considerable substrate flexibility in targeting both
1021 eukaryotic and bacterial proteins.

1022 Inactive transglutaminase domains in polymorphic toxins

1023 In course of the current study we observed that several
1024 polymorphic toxin proteins with several distinct types of
1025 C-terminal toxin domains displayed a N-terminal trans-
1026 glutaminase domain (Figure 6D). However, closer examin-
1027 ation of the multiple alignment of these transglutaminase
1028 domains revealed that one or more of the conserved resi-
1029 dues (a C, H, and D), which constitute the catalytic triad
1030 of their papain-like peptidase active site, were lost [53]
1031 (Figure 6D). This suggests that they lack peptidase activ-
1032 ity. Domain architectural analysis showed that these in-
1033 active transglutaminase domains are always located
1034 immediately after a N-terminal signal peptide or TM
1035 helix and are followed by an array of RHS repeats that
1036 constitute the filamentous part of the toxin. Occasionally,
1037 they might be adjacent to domains of the immunoglobu-
1038 lin superfamily (the so called “bacterial Ig” type domains;
1039 Figure 6D). This position suggests that, unlike the above-
1040 described active peptidase domains, these inactive trans-
1041 glutaminases have no role in toxin or processing activity.
1042 Instead, they might simply serve in anchoring the toxin
1043 on the cell surface by binding peptides.

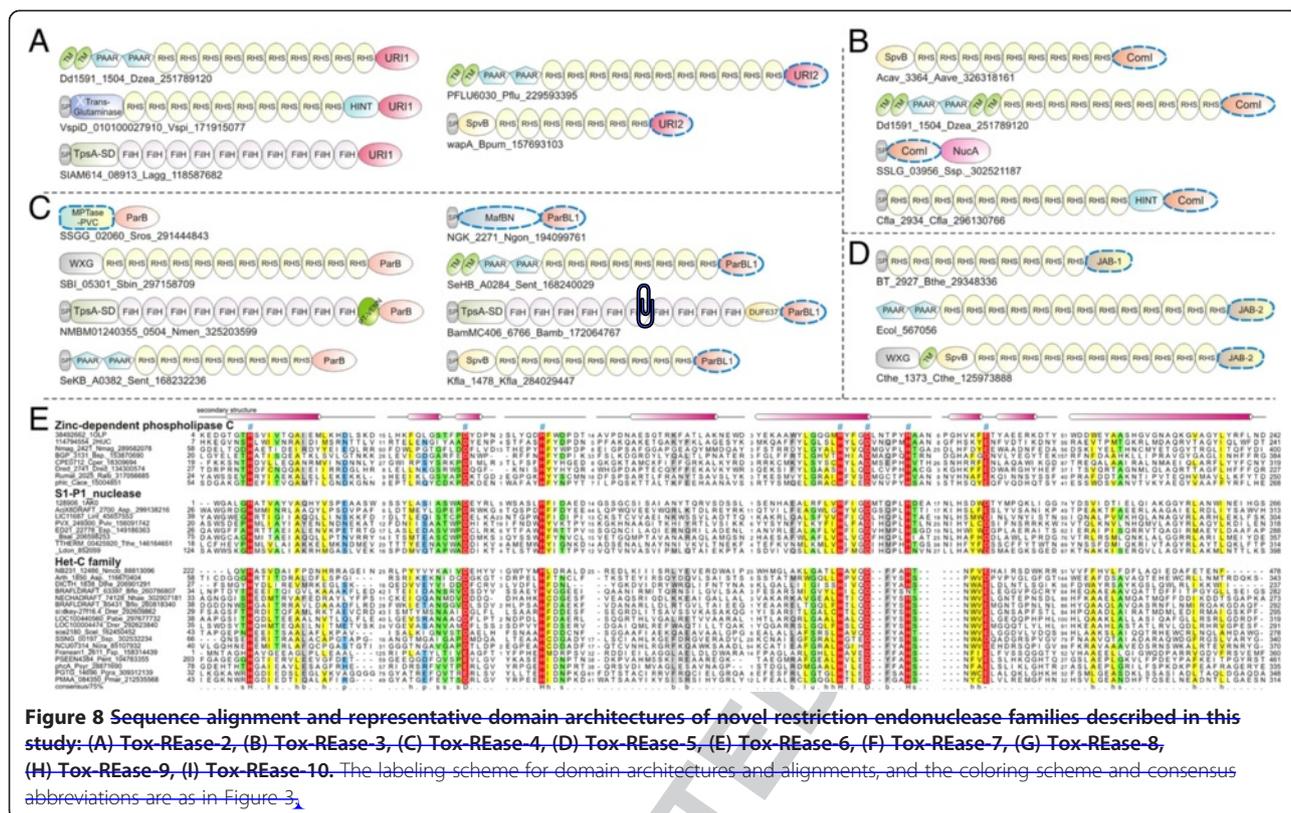
1044 Identification of further toxin domains in polymorphic 1045 toxins and related proteins that operate on nucleic acids

1046 In our earlier study we had shown that majority of toxin
1047 domains in polymorphic toxin systems operate on nu-
1048 cleic acids – nucleases and base deaminases [17,18]. In
1049 this study we were able to further extend the diversity of
1050 toxin domains that act on nucleic acids via the discovery

of additional nucleases and deaminases that were not
previously recognized (Figures 7, 8, 9, 2). We observed
that the divalent cation-dependent nucleases among
polymorphic toxins are frequently drawn from ancient
nuclease folds, namely the HNH/EndoVII, REase and
URI endonuclease folds [102-107]. Additionally, we
present evidence below that representatives of few other
potential cation-dependent enzymatic domains might
function as nuclease domains in polymorphic toxins.
Interestingly, the PIN domains, which are major divalent
cation-dependent nucleases in the toxin-antitoxin sys-
tems [22,108], do not appear to be utilized in the poly-
morphic toxins and related systems. Toxin nucleases that
utilize divalent cations can catalyze the direct hydrolysis
of the phosphodiester bond and as a result attack both
DNA and RNA. However, the metal-independent
nucleases can only act as RNases as their endonucleolytic
action involves the formation of a cyclic 2'-3' phosphate
that does not require metal-dependent direction of a
hydrolytic attack [107]. Such RNases belong to many dis-
tinct folds, several of which appear to have emerged only
in course of the diversification of toxin domains of poly-
morphic toxins, bacteriocins and classical toxin-antitoxin
systems [17,22,28,107,109,110]. While we were able to
unify several of the metal-independent RNases, which
were previously considered to be unrelated, into a single
monophyletic assemblage, there are still several distinct
toxin domains that likely to represent novel metal-
independent RNases (see below; novel toxins). This
structural diversity of metal-independent RNases and the
repeated emergence of several such nuclease domains
among different toxin systems suggest that there are
some fundamental constraints in the evolutionary
innovation of nuclease domains. It appears that the inde-
pendent emergence of multiple residues for metal-
chelation and acid-base catalysis to constitute an active
site that can support hydrolytic cleavage of nucleic acids
is a far less likely event than the emergence of a metal-
independent active site that utilizes the innate reactivity
of RNA to facilitate an internal attack with the formation
of 2'-3' cyclic phosphates. We briefly describe below the
newly recovered toxin domains that act on nucleic acids.

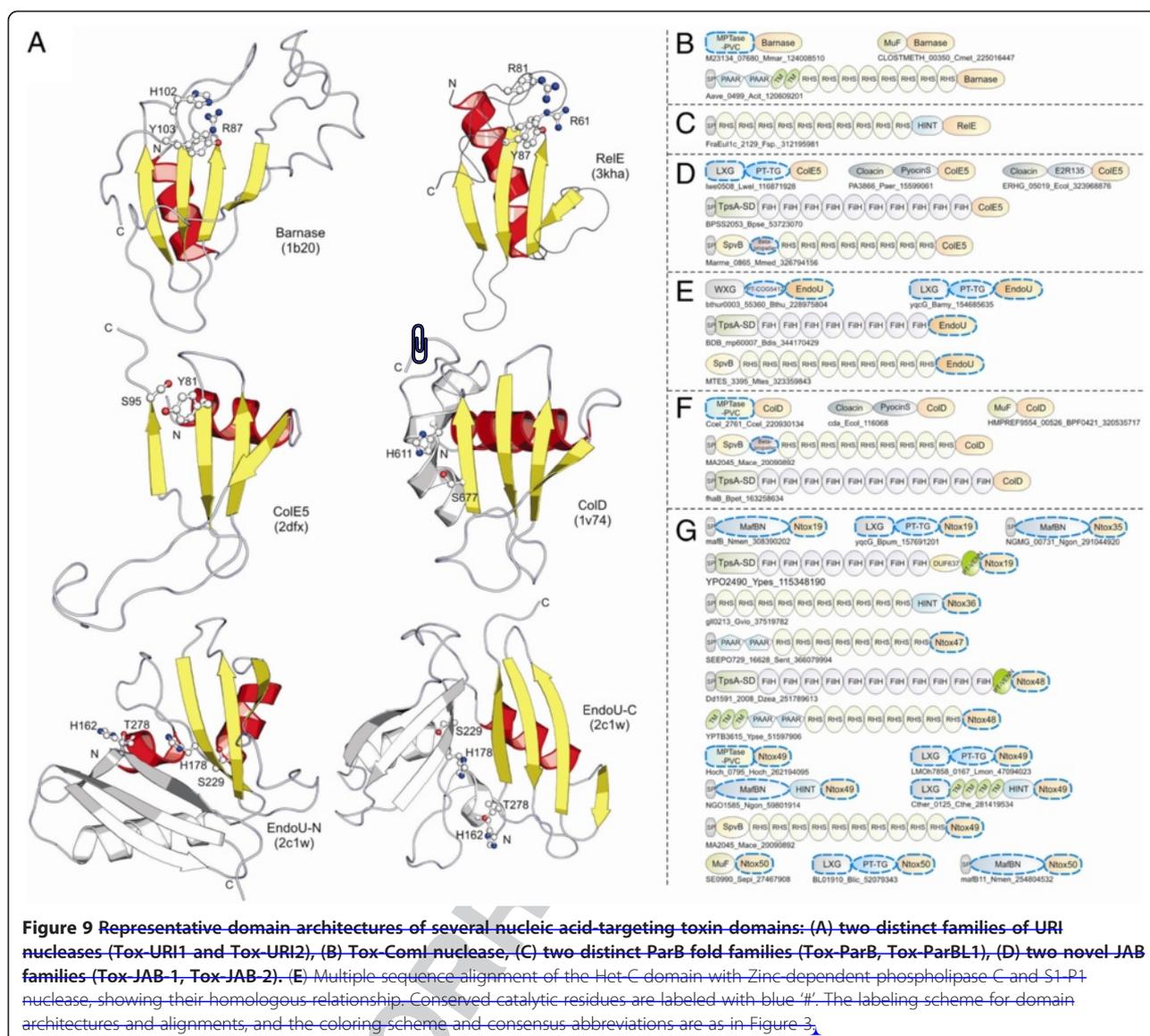
1093 Novel toxins with the HNH/EndoVII nuclease domain

1094 In our earlier studies we found nuclease toxin domains
1095 belonging to eight distinct clades of the HNH/EndoVII
1096 fold among the polymorphic toxin systems [17,18]. Of
1097 these, nucleases belonging to the classical HNH and
1098 NucA clades widely occur beyond the polymorphic toxins
1099 across diverse sub-cellular systems, such as, DNA repair/
1100 recombination, restriction-modification (R-M) and envir-
1101 onmental nucleic acid degradation systems [103,106,111].
1102 In contrast, the GH-E, DHNNK, WHH, LHH and AHH
1103 domains appear to have arisen in and remained largely



1104 restricted to polymorphic toxin systems. The NGO1392
 1105 clade appears to have arisen in the bacterial polymorphic
 1106 toxin systems, but was transferred to eukaryotes where it
 1107 might have assumed a role in DNA repair [17]. In this
 1108 study we recovered six more clades of HNH domain
 1109 nucleases that appear to have primarily diversified among
 1110 bacterial polymorphic and related PVC-SS-associated
 1111 toxins. Keeping with the earlier nomenclatural system,
 1112 we named five of these novel clades on the basis of the
 1113 conserved motifs that characterized them as the SHH,
 1114 HHH, GHH, GHH-2 and EHHH clades of HNH domains
 1115 (Figure 7). The sixth of these is related to the version of
 1116 the HNH domains found in the restriction enzyme SphI
 1117 [112] and the animal CIDE (CAD/DF40) protein involved
 1118 in nucleolytic DNA fragmentation during apoptosis [113],
 1119 and is termed HNH-CIDE (Table 2). Architectural analysis
 1120 indicated that the novel HNH clades occur both as poten-
 1121 tial diffusible toxins (mainly in Gram-positive bacteria)
 1122 and as contact-dependent toxins borne at the tip of
 1123 long filamentous structures (proteobacteria, bacteroi-
 1124 detes, planctomycetes and certain Gram-positive bacteria;
 1125 Figure 7). Representatives of the SHH clade have been
 1126 transferred to crustacean (e.g. *Daphnia*; gi: 321474287)
 1127 and tailed bacteriophages (e.g. *Bacillus* phage SPbeta; gi:
 1128 9630134). The former transfer is consistent with occur-
 1129 rence of an effector with a SHH nuclease domain in the
 1130 eukaryotic endosymbiont, *Simkania* (gi: 338732338).

The CIDE protein was previously known only from
 metazoans with no known representatives from other
 eukaryotes; hence, its origin remained mysterious [114].
 The identification of the HNH-CIDE toxin domains sug-
 gests that this nuclease domain first arose in context of
 bacterial conflicts and was laterally transferred to animals
 early in their evolution. In animals, its innate cytotoxic
 action appears to have been channelized as an effector of
 apoptosis. Our searches also showed that the C-terminal
 domain of teneurin and Odd Oz proteins from the animal
 lineage (metazoans + choanoflagellates) contain an in-
 active version of a HNH domain belonging to the GHH
 clade (Figure 7E). While presence of RHS repeats in these
 proteins related to those in bacterial RHS proteins has
 been previously recognized [115], the relationship of their
 C-terminal domain to a specific bacterial toxin domain
 has not been hitherto reported. Teneurin/Odd Oz pro-
 teins function as developmental regulators with a poten-
 tial role in cell-surface adhesion in diverse processes such
 as cell migration, neuronal path finding and fasciculation,
 gonad development, and basement membrane integrity
 [115-117]. The region of these proteins spanning the in-
 active GHH nuclease domain has been described as being
 cleaved off and amidated at the C-terminus in vertebrates
 to give rise to a peptide with possible neuromodulatory
 activity [118]. This region in teneurin-2 is also the ligand
 for latrophilin-1, which is also the receptor for another



1158 molecule, latrotoxin, whose origins also lie among the
 1159 bacterial toxins (see below) [116]. Hence, it is conceivable
 1160 that the RHS portion of these proteins participates in cel-
 1161 lular adhesion, while the cleaved off inactive GHH do-
 1162 main act as a diffusible signal. It would be of interest to
 1163 investigate if this inactive GHH domain might bind nu-
 1164 cleic acids upon being taken up by target cells. Our detec-
 1165 tion of the GHH domain in the Teneurin/Odd Oz
 1166 proteins establishes that they have emerged from the single
 1167 transfer of a specific type of a complete bacterial poly-
 1168 morphic toxin gene followed by its fusion to EGF repeats
 1169 of animal provenance (Figure 7E).

1170 **Novel restriction endonuclease fold domains in**
 1171 **polymorphic toxins**

1172 In our earlier study we had identified toxin domains
 1173 in polymorphic toxins belonging to a previously

1174 uncharacterized clade of the REase fold (REase-1) [17].
 1175 Further analysis revealed that there are nine additional,
 1176 previously unknown clades of the REase fold that are
 1177 present exclusively as toxin domains of a diverse group
 1178 of polymorphic toxins (Figure 8; numbered serially
 1179 REase-2-REase-10). Their domain architectures and
 1180 gene-neighborhoods indicate that they are secreted by
 1181 means of the T2SS, T5SS, T7SS, TcdB/TcaC and the
 1182 PrsW-type peptidase-dependent system in different bac-
 1183 terial lineages. Of these, at least four distinct versions,
 1184 namely REase-2, REase-3, REase-5 and REase-6 are
 1185 coupled with a PrsW peptidase, suggesting that a nota-
 1186 ble diversification of these nucleases appears to have
 1187 happened in the context of these systems (Figure 8).
 1188 Many of the REase toxins secreted via the other systems
 1189 have central RHS repeats (e.g. REase-9; Figure 8). These
 1190 architectures suggest that REases might function both as

1191 diffusible and contact-dependent toxins. Tox-REase-8 is
1192 primarily found in the arthropod endosymbiont *Wolba-*
1193 *chia* and the *Acanthamoeba* endosymbiont *Amoebophi-*
1194 *lus* and is usually associated with arrays of ankyrin
1195 repeats (Figure 8C). These lack associated genes for im-
1196 munity proteins and are likely to be deployed against tar-
1197 gets in the host cells – this represents the first instance
1198 of a REase domain effector being used by endosymbionts
1199 of eukaryotes. Representatives of Tox-REase-8 are found
1200 in the genomes of arthropods, such as the crustacean
1201 *Daphnia*, several mosquitoes, ants and beetles, and the
1202 placozoan *Trichoplax*. This suggests that Tox-REase-8
1203 has been repeatedly transferred to diverse animals from
1204 their *Wolbachia*-like endosymbionts. Beyond conven-
1205 tional polymorphic toxin systems, REase-9 is also found
1206 in a *Parachlamydia* effector (PUV_01770, gi: 338174171)
1207 that might target nucleic acids in its host *Acanthamoeba*.
1208 All ten clades of REase toxins have an active site that
1209 closely conforms to the classical REase active site with a
1210 D-[EQ]XK signature in the core strands that constitute
1211 the metal-chelating site [103]. The majority of character-
1212 ized members of this fold act on DNA targets; hence, it is
1213 conceivable that these toxins also attack the genome of
1214 the target cells through endonucleolytic cleavage.

1215 **URI domain nuclease toxins**

1216 The URI domain was first identified as a conserved
1217 metal-dependent endonuclease domain catalyzing the
1218 cleavage of the 3' side of a damaged DNA base during
1219 nucleotide excision repair by UvrC, and mediating site-
1220 specific insertion of certain introns [102,119]. Similar
1221 nuclease domains have also been found in certain
1222 REases, such as R. Eco29kI, and the transposase module
1223 of Penelope-like non-LTR retroelements [104]. In this
1224 work we identified, for the first time, URI domain
1225 nucleases in polymorphic toxins that are present in bac-
1226 teria from most major bacterial lineages (Figure 9A,
1227 Table 2) that are usually secreted via T2SS, T5SS, TcdB/
1228 TcaC and T6SS. The Tox-URI domains can be divided
1229 into two major clades, with the second clade being par-
1230 ticularly divergent (Additional File 1). A version of the
1231 Tox-URI domain belonging to the first clade has also
1232 been transferred to fungi, where it occurs as an intracel-
1233 lular domain fused to an ABC ATPase transporter (e.g.
1234 *Neurospora crassa* NCU06946; gi: 164424641; Additional
1235 File 1). Given this architecture, it is conceivable that they
1236 function in degradation of nucleic acids taken up by
1237 these fungi. Interestingly, certain URI domain toxins
1238 belonging to the second clade are present in distantly
1239 related intracellular symbionts/pathogens of *Acanth-*
1240 *amoeba*, such as the *Simkania negevensis* (gi:
1241 338731950), *Odysseella* (gi: 344925485) and *Rickettsia*
1242 *belli* (gi: 91206213). Analysis of the gene-neighborhoods
1243 of these toxins suggests that they have adjacent genes

1244 encoding immunity proteins (Additional File 1), suggest-
1245 ing that these toxins are likely to be used in intra-
1246 conflict rather than being directed against the host.
1247 Along with the above-described Otu peptidase toxins
1248 from *Odysseella*, these URI domain toxins represent rela-
1249 tively rare examples of polymorphic toxins deployed in
1250 intraspecific conflict by endo-symbiotic/parasitic bac-
1251 teria. Other than the versions from intracellular bacteria,
1252 the URI domain toxins are typically associated with fila-
1253 mentous RHS repeats.

1254 All the above metal-dependent nuclease domains are
1255 shared by polymorphic toxin systems with R-M systems,
1256 but are apparently absent among classical toxin-
1257 antitoxin systems [22,28]. However, the versions found
1258 in the polymorphic toxins differ from those in classical
1259 R-M systems in lacking a complex array of associated
1260 DNA-binding domains [120]. Hence, we suspect that the
1261 versions of these nuclease domains deployed by the
1262 polymorphic toxin systems might have lower target se-
1263 quence specificity than those deployed in R-M systems.
1264 Further, those from the former systems are under selec-
1265 tion imposed by the physical interactions with cognate
1266 immunity proteins. It appears that these factors might
1267 eminently disallow exchange of nuclease domains be-
1268 tween polymorphic toxin and R-M systems.

1269 **The competence nuclease (ComI) domain**

1270 This nuclease domain is prototyped by the secreted 17
1271 kDa competence nuclease ComI of *Bacillus subtilis*,
1272 which is a major determinant of DNA uptake when the
1273 bacterium becomes capable of transformation prior to
1274 stationary phase [121]. We recovered related nucleases
1275 as toxin domains of polymorphic toxins from actinobac-
1276 teria (e.g. gi: 296130766 from *Cellulomonas flavigena*)
1277 and proteobacteria (e.g. gi: 326318161 from *Acidovorax*
1278 *avenae*; Figure 9B). This domain could not be unified
1279 with any previously known fold observed among
1280 nucleases. A multiple alignment of this domain showed
1281 that it contained a central dyad of two acidic residues
1282 (usually a DE motif) followed by a third conserved acidic
1283 residue a few positions downstream (Additional File 1).
1284 These residues could potentially form a divalent cation-
1285 chelating site, suggesting that the ComI nuclease is likely
1286 to be the fourth metal-dependent nuclease superfamily
1287 among the toxin domains. Interestingly, the *B.subtilis*
1288 competence nuclease is physically associated with the 18
1289 kDa product of the adjacent ComJ gene, which acts as
1290 its inhibitor – the interplay between the ComI nuclease
1291 and its inhibitor ComJ has been suggested to be import-
1292 ant for optimal digestion of incoming DNA, so as to fa-
1293 cilitate transformation [121]. The structure of this
1294 operon with a nuclease followed by its inhibitor is rem-
1295 iniscent of the polymorphic toxin systems with the toxin
1296 gene followed by the immunity protein. Consistent with

1297 this, ComJ homologs occurs as an immunity protein for
1298 polymorphic toxins with the ComI nuclease domain in
1299 several proteobacteria. Hence, it is possible that these
1300 key components of the *Bacillus* DNA uptake system
1301 have evolved from a toxin-immunity gene pair.

1302 **ParB domain toxins**

1303 We recovered several polymorphic toxins with N-
1304 terminal filamentous regions formed by RHS or fila-
1305 mentous haemagglutinin repeats and C-terminal ParB
1306 toxin domains (Figure 9C). The ParB domain is the sub-
1307 ject of much confusion: based on a study, which claimed
1308 to demonstrate both endo- and exo- DNase activity in
1309 the ParB protein [122], required for maintenance of the
1310 plasmid RK2, the domain was labeled as a nuclease do-
1311 main. However, it should be noted that this study was
1312 based on entirely erroneous assumptions that the RK2
1313 ParB domain was related to nucleases such as the
1314 staphylococcal nuclease and RuvC [122]. In contrast,
1315 other members of the ParB superfamily, such as sulfire-
1316 doxin, have been convincingly demonstrated to possess
1317 metal-dependent phosphotransferase activity that utilizes
1318 ATP to form a phosphoryl ester of sulfinate generated
1319 from the active site cysteine of the peroxiredoxins [123].
1320 Through sequence profile searches we were able to dem-
1321 onstrate that DndB is a member of the ParB superfamily.
1322 DndB negatively regulates the formation of the unusual
1323 DNA phosphorothioate modification, in which the non-
1324 bridging oxygen in the phosphodiester linkage of DNA
1325 is replaced by a sulfur atom in a sequence-specific man-
1326 ner [124]. Hence, it appears that even this member of
1327 the ParB superfamily, comparable to sulfiredoxin might
1328 hydrolyze a phosphoryl ester linked to a sulfur center.
1329 The convincingly inferred metal-dependent phospho-
1330 transfer activity of the ParB superfamily implies that in
1331 principle certain representatives might also be able to
1332 catalyze nuclease activity through a comparable hydroly-
1333 sis of a phosphodiester bond. Hence, it is conceivable
1334 that, even though the ParB domain was considered a nu-
1335 clease for the wrong reasons, this activity might be still
1336 valid for some representatives of the superfamily. This is
1337 also consonant with the earlier recovery of ParB
1338 domains in nucleases encoded by certain R-M like sys-
1339 tems [103,125]. The predominance of nuclease domains
1340 among the toxin domains of polymorphic toxin systems
1341 also supports a potential nuclease function for the ParB
1342 toxin domains. Examination of the multiple alignment of
1343 the ParB domains from polymorphic toxins suggests that
1344 they possess a strongly conserved DGHHR motif that is
1345 predicted to form part of their highly conserved metal-
1346 binding active site (Additional File 1). In addition to the
1347 classical ParB toxin domains, we recovered a second
1348 large group of toxin domains typified by that found in
1349 *Neisseria gonorrhoeae* NGK_2271 (gi: 194099761), which

could be united using profile-profile comparisons with 1350
the ParB domain (HHpred probability 93%; $p = 2 \times 10^{-6}$ 1351
match to 1vz0 *Thermus* ParB). While being rather diver- 1352
gent from the classical ParB domains, they display a 1353
motif with a conserved arginine that is equivalent to the 1354
DGHHR motif in the former. Additionally, they display 1355
a conserved N-terminal serine that is absent in the clas- 1356
sical ParB domains. Hence, we termed this distinct fam- 1357
ily of ParB-related domains as Tox-ParBL1 (Figure 9). In 1358
addition to the bacterial polymorphic toxins, Tox- 1359
ParBL1 domains are also found in several eukaryotes 1360
such as kinetoplastids, and several metazoans, fungi, 1361
plants, stramenopiles and ciliates (Table 2 and Addi- 1362
tional File 1). Thus, this example represents an inde- 1363
pendent acquisition by eukaryotes of a ParB-related 1364
domain from the polymorphic toxin systems, distinct 1365
from the sulfiredoxins. 1366

1367 **The JAB domain**

1368 We detected two distinct clades of the JAB domain 1369
superfamily as the potential toxin domain of several 1370
classical polymorphic toxins (Figure 9D). The JAB do- 1371
main has been previously shown to be a peptidase that 1372
specifically targets the C-termini of ubiquitin-like pro- 1373
teins (UBLs) either as a DUB or as a processing enzyme 1374
[126-128]. All previously identified prokaryotic JAB 1375
domains are intracellular proteins. Most representatives 1376
of them are components of systems utilizing UBLs in 1377
biosynthetic pathways or protein modification. As these 1378
toxin genes are accompanied by immunity proteins they 1379
are likely to be used in intraspecific conflict rather than 1380
against eukaryotic targets. Hence, the presence of the 1381
JAB domain among the toxin modules of classical poly- 1382
morphic toxins was unexpected, because most of the 1383
bacteria in which they are present lack systems with 1384
conjugated or processed ubiquitin-like proteins [126]. 1385
However, based on contextual information from domain 1386
architectural analysis it was recently proposed that a 1387
subset of the JAB domains (i.e. those belonging to the 1388
RadC clade) are more likely to function as nucleases that 1389
cleave DNA, rather than as peptidases [18]. The two 1390
clades of JAB domains found among the polymorphic 1391
toxins, like RadC, are rather divergent with respect to 1392
those that act on UBLs, and do not conserve the resi- 1393
dues lining the tunnel that accommodates the UBL tail 1394
in the peptidase versions (Additional File 1). This sug- 1395
gests that, as previously proposed for RadC, the toxin 1396
JAB domains might function as nucleases rather than as 1397
peptidases. Of the two clades Tox-JAB-1 is found in only 1398
in the bacteroidetes lineage associated with N-terminal 1399
RHS repeats (Figure 9D). Tox-JAB-2 is more widely dis- 1400
tributed across proteobacteria, bacteroidetes and few fir- 1401
micutes which partly overlaps with the “domain of 1402
unknown function”, DUF4329 from the PFAM database

1403 (Figure 9D). Versions of Tox-JAB-2 are also present in
1404 several NCLDVs, such as iridoviruses, mimiviruses and
1405 algal viruses, and *Xanthomonas* phages (e.g. phage
1406 OP1). These latter versions are secreted proteins and
1407 could potentially function as phage-encoded virulence
1408 factors.

1409 **The Het-C hydrolase domain**

1410 The Het-C domain was first identified as a major player in
1411 the phenomenon of fungal vegetative incompatibility
1412 [129], wherein it mediates programmed cell death upon
1413 interaction with incompatible hyphae. Subsequently, a
1414 version of the Het-C domain encoded by *Pseudomonas*
1415 *syringae* was shown to be required for the infection of
1416 fungal hyphae by this bacterium, by exploiting the mech-
1417 anism of hetero-incompatibility [130]. In our analysis we
1418 recovered Het-C domains in systems related to the poly-
1419 morphic toxins that utilize PVC-SS (e.g. gi: 148657895
1420 from *Roseiflexus*; Figure 5C). Profile-profile comparisons
1421 using an alignment of the Het-C domain (Figure 9E)
1422 revealed hits with borderline significance ($p = .001$; 50%
1423 probability) to a group of α -helical hydrolases sharing a
1424 common a fold, including zinc-dependent phospholipase
1425 C [131] and the S1-P1 nucleases [132]. The predicted
1426 secondary structure for the Het-C domain was also com-
1427 patible with the α -helical fold seen in those hydrolases
1428 and examination of the multiple alignments revealed
1429 that the two possessed a comparable set of conserved ac-
1430 tive site residues (Figure 9E). This includes four con-
1431 served histidines and 3 acidic residues (D/E) suggesting
1432 that the Het-C domain possess a metal-dependent active
1433 site similar to that seen in the phospholipase and S1-
1434 P1-like nucleases. Indeed, secreted versions of this do-
1435 main with both phospholipase and nuclease activity are
1436 known from different bacteria [132]. This suggests that
1437 the Het-C domain might also possess either metal-
1438 dependent nuclease or phospholipase activity, and that
1439 this activity is likely to be critical for the apoptotic and
1440 toxin action of this domain in fungi and bacteria.

1441 **Barnase-EndoU-colicin E5/colicin D-RelE like nuclease fold:**

1442 **A large assemblage of metal-independent RNases**

1443 In our earlier study we had recovered the EndoU do-
1444 main as a metal-independent RNase frequently found in
1445 polymorphic toxin systems. We had further shown that
1446 the EndoU fold is marked by a potential duplication of a
1447 core helix- β -sheet element that constitutes its active site
1448 [17]. In another earlier study we had unified the colicin
1449 E5 and colicin D RNase domains with the RNase do-
1450 main of the RelE toxin that is found in classical toxin-
1451 antitoxin systems [133]. A comparison showed that the
1452 core structural element in EndoU, Colicin E5, colicin D
1453 and RelE is a similar strand- β -sheet unit (Figure 2A).
1454 Transitive structure-comparison searches using the

DALLite program confirmed that these RNase domains 1455
are indeed related as they preferentially recovered each 1456
other (with $Z > 3.5$). Further, these DALLite searches 1457
showed that they could be united with several other 1458
metal-independent RNase domains, namely the RNase 1459
toxins and other secreted RNases from fungi, such as 1460
sarcin, RNaseT and RNase U2, and the bacterial RNases 1461
prototyped by barnase ($Z > 3.5$; Figure 2A; this latter 1462
group is described as the microbial RNase fold in the 1463
SCOP database [134]). We term the common structural 1464
unit shared by all the representatives of the above- 1465
unified assemblage the BECR (Barnase-EndoU-Colicin 1466
E5/D-RelE) fold. The common structural unit, which 1467
constitutes the catalytic domain of the BECR fold 1468
RNases contains a N-terminal helical segment that is fol- 1469
lowed by a sheet formed by 4-stranded meander 1470
(Figure 2A). In several cases the 4th strand is followed by 1471
an additional short 5th strand that is differentially posi- 1472
tioned in various versions of this fold. Furthermore, the 1473
location of the active site residues is often comparable 1474
across these enzymes and our sequence analysis revealed 1475
that many of these RNases (including EndoU, colicin 1476
E5/D and some clades of RelE) share a conserved alco- 1477
holic residue (S/T) in the 4th strand that contributes 1478
to the active site (Figure 2A). 1479

In addition to the EndoU clade, our sequence compar- 1480
isons indicated that several of the newly recovered BECR 1481
fold toxin domains from polymorphic toxin systems be- 1482
long to other previously defined clades in this fold, such 1483
as barnase, colicin E5, and colicin D clades (Figure 2B-F). 1484
While the classical RelE endoRNase domain is common 1485
in type-II toxin-antitoxin systems, we observed only a 1486
single instance of it being used as a toxin domain in the 1487
polymorphic toxins (gi: 357015358 from *Paenibacillus* 1488
elgii). However, using secondary structure prediction 1489
combined with profile-profile comparisons we also dis- 1490
covered distinct, previously unrecognized clades of 1491
RNases displaying the BECR fold (Figure 2G): these in- 1492
clude the clades 1) Ntox7 (e.g. γ 1701, gi: 22125595 from 1493
Yersinia pestis); 2) Ntox19 (NMW_1482, gi: 254673263 1494
in *Neisseria meningitidis*); 3) Ntox35 (typified by 1495
NGMG_00731; gi: 291044920 from *Neisseria gonor-* 1496
rhoae); 4) Ntox36 (typified by the toxin domain of 1497
gll0213; gi: 37519782 from *Gloebacter violaceus*); 5) 1498
Ntox47 (typified by the toxin of rhs2; gi 366079994 from 1499
Salmonella enterica); 5) Ntox48 (e.g. gi:251789613 from 1500
Dickeya zeae); 6) Ntox49 (gi:59801914 in *Neisseria gonor-* 1501
rhoae; 7) Ntox50 (gi: 254804532 in *Neisseria meningiti-* 1502
dis). Together with previously characterized clades, these 1503
seven novel clades are extensively represented among the 1504
toxin domains of classical polymorphic toxins and in 1505
some cases related toxins delivered by the PVC-SS 1506
(Figures 5 and 9). This observation suggests that the BECR 1507
fold has supplied one of the most extensive radiations of 1508

1509 RNase toxins, which cuts across mechanistically distinct
1510 systems – the polymorphic and related secreted toxins
1511 and the classical toxin-antitoxin systems. Examination of
1512 the predicted active site residues among the newly char-
1513 acterized clades pointed to each clade acquiring their
1514 own unique features. For example, Ntox35 has acquired
1515 two conserved N-terminal histidines in addition to the
1516 conserved S/T from the C-terminal strand. Ntox50 and
1517 Ntox19 instead have a single N-terminal histidine, simi-
1518 lar to one observed in several members of the colicin
1519 E5/D clade [110], accompanied by a second C-terminal
1520 histidine found at the position usually occupied by the
1521 conserved S/T of the BECR fold (Additional File 1). The
1522 presence of two histidines in the above three clades is
1523 reminiscent, though not equivalent in terms of second-
1524 ary structure context, to those seen in the EndoU clade,
1525 suggesting a comparable reaction mechanism in all these
1526 versions of the fold. In contrast, Ntox36 lacks any con-
1527 served histidine; instead it displays other clade-specific
1528 conserved residues; e.g. an asparagine in the N-terminal
1529 region. Most of these enzymes, especially those with
1530 two conserved histidines are likely to utilize a metal-
1531 independent mechanism similar to that observed in
1532 RNaseA (see below) [107]. This is supported by the gen-
1533 eration of cleavage products with 2'-3' cyclic phosphate
1534 termini in several biochemically characterized members
1535 of these RNases (e.g. XendoU). Some members of the
1536 EndoU clade have been shown to require Mn^{2+} for ef-
1537 fective catalysis of RNA cleavage [135]; however, given
1538 that they still produce 2'-3' cyclic phosphates, it is likely
1539 that this metal is required for stabilization of the hyperch-
1540 arged transition state rather than the actual phosphoes-
1541 terase activity.

1542 Interestingly, we observed that one RNase of the BECR
1543 fold related to the colicin E5/D clade is also found con-
1544 sistently associated with the flagellar operon across fir-
1545 micutes (e.g. gi: 28211324 from *Clostridium tetani*;
1546 Additional file 1). It would be of interest to investigate if
1547 this RNase is delivered by the flagellar system or alterna-
1548 tively functions to regulate flagellar gene expression as a
1549 RNA-processing enzyme. RNases of the Ntox50 clade
1550 have also been acquired by bacteriophages such as *Clos-*
1551 *tridium* phage phiC2 (gi: 134287339) and might be used
1552 in conflicts with the host or other phages. Likewise
1553 Ntox19 has been acquired by the giant *Acanthamoeba*-
1554 infecting mimivirus and is also found in potential effec-
1555 tors secreted by the *Acanthamoeba* endosymbionts
1556 *Parachlamydia* and *Odyssella*.

1557 **Novel toxin domains which are likely to function as** 1558 **nucleases**

1559 Our systematic analysis of the polymorphic toxin sys-
1560 tems recovered a total 43 distinct novel toxin domains
1561 that could not be unified with any previously known

domain (Table 2; Additional file 1). Only a small minor- 1562
ity of these domains contain at least one experimentally 1563
characterized member. Their sequence conservation pat- 1564
terns, together with the preponderance of nucleases 1565
among polymorphic toxins, suggest that most of these 1566
novel toxin domains are likely to be nucleases. Indeed, 1567
their conservation patterns suggest that these novel 1568
toxin domains include both potential metal-dependent 1569
and independent enzymes (Table 2; Additional file 1). 1570
The C-terminal toxin domain of the originally character- 1571
ized contact-dependent inhibitor protein CdiA from 1572
Escherichia coli was demonstrated to possess RNase ac- 1573
tivity [44]. We observed that the *E.coli* CdiA-C domain 1574
is widely distributed across polymorphic toxins from di- 1575
verse bacteria. We also uncovered this domain in the 1576
Photorhabdus PalA protein, which lacks an associated 1577
immunity protein but is encoded in a pathogenicity is- 1578
land adjacent to the Mcf gene whose product is a toxin 1579
directed against the caterpillar host [87]. In light of this, 1580
it is possible that *E.coli*-CdiA-C domain in PalA might 1581
be directed against the host as an accessory toxin. Exam- 1582
ination of the *E.coli*-CdiA-C domain shows that it pos- 1583
sesses an all β fold that lacks any conserved residues 1584
typical of metal-dependent nucleases. Hence, it is likely 1585
to be a metal-independent RNase and probably defines a 1586
novel structural theme among them. 1587

We uncovered an uncharacterized toxin domain that 1588
is found in polymorphic toxin systems from a wide 1589
range of bacteria and several potential effectors delivered 1590
by endo-symbiotic/parasitic bacteria (e.g. *Wolbachia*, 1591
Ehrlichia, *Odyssella*, *Rickettsia* and *Legionella*). It is also 1592
found at the C-terminus of a group of eukaryotic pro- 1593
teins typified by the plant protein EDA39 and we ac- 1594
cordingly call it the Tox-EDA39C domain (Additional 1595
File 1). This domain is characterized by two highly con- 1596
served histidines respectively in the N- and C-terminal 1597
halves of the proteins that are likely to comprise its ac- 1598
tive site. This conservation pattern is reminiscent of the 1599
catalytic residues seen in the RNase A domain [136], 1600
and might represent a novel metal-independent RNase 1601
that catalyzes a reaction similar to that of RNase A. The 1602
presence of this domain in several eukaryotic lineages, 1603
such as plants, fungi, oomycetes and *Dictyostelium*, sug- 1604
gests that it might have been acquired by eukaryotes 1605
from bacterial endosymbionts and could have been 1606
recruited as a potential RNase used in anti-pathogen 1607
defense. Ntox43 is typified by the toxin domain of the 1608
recently described RhsT from *Pseudomonas aeruginosa*, 1609
which has been shown to translocate to the host cyto- 1610
plasm and mediate an inflammatory response [46]. This 1611
toxin, like Tox-EDA39C, has two conserved histidines 1612
suggesting that it might also function as a RNase A-like 1613
metal-independent nuclease (Additional File 1). Hence, 1614
we predict that RhsT is likely to activate the 1615

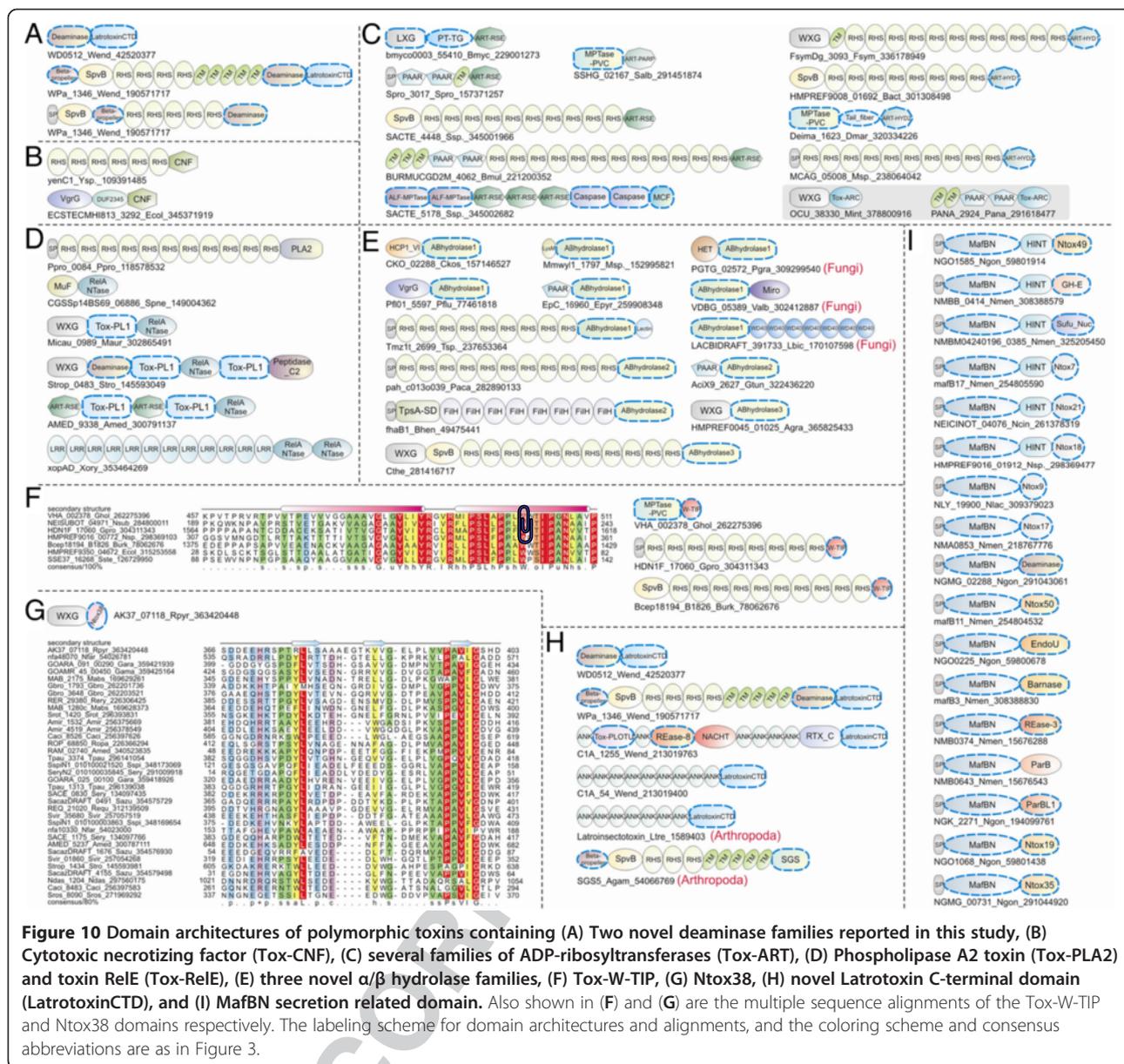
1616 inflammosome via cleavage of specific RNAs. Although
1617 proteins with Ntox43 display architectures are similar to
1618 classical polymorphic toxins, none of them are asso-
1619 ciated with adjacent genes for immunity proteins. This
1620 suggests that they are likely to be used primarily against
1621 eukaryotic hosts. At least four other toxin domains iden-
1622 tified by us (Ntox18, Ntox19, Ntox22, Ntox26, Ntox30)
1623 are likely to be novel metal-independent endo-RNases
1624 that utilize a two histidine-dependent mechanism to
1625 catalyze transesterification and formation of a 2'-3' cyclic
1626 phosphate like RNase A (Table 2).

1627 We observed that the RES domain (PFAM: PF08808),
1628 whose function was previously unknown, is another toxin
1629 domain that is found in polymorphic toxin systems.
1630 Interestingly, it is also found in classical toxin-antitoxin
1631 systems, where it is typically paired with a distinctive
1632 antitoxin (previously labeled as a domain of unknown
1633 function, DUF2384 in the PFAM database). Hence, we
1634 predict that the RES domain is likely to be a novel RNase
1635 domain shared by different toxin systems. Examination
1636 of the alignment of the RES domain revealed two con-
1637 served arginines, a glutamate and a serine – this config-
1638 uration does not appear likely to support a metal-binding
1639 active site; however, these residues are suitable for
1640 catalyzing a distinct metal-independent RNase reaction.
1641 Ntox24 is characterized by a single conserved histidine,
1642 and, like the RES domain, versions of this toxin do-
1643 main are additionally found in what appear to be novel
1644 type-II toxin-antitoxin systems associated with a previously
1645 uncharacterized family of antitoxins (e.g. gi: 139439131).
1646 The toxin domain from the CdiA protein from *Entero-*
1647 *bacter cloacae* (Ntox21) shows universally conserved
1648 residues, including a single histidine and two aspartates,
1649 but could not be unified with any other known domain.
1650 It is conceivable that Ntox24 and Ntox21 act as metal-
1651 independent endoRNases comparable to the Colicin E3
1652 nuclease domain [137], which is also found in poly-
1653 morphic toxin systems (Tox-ColE3)[17]. Our detection
1654 of Tox-ColE3 in these systems also helped in emending
1655 the proposed active site of these RNases. Based on struc-
1656 tural analysis it was previously proposed that the active
1657 site of these enzymes corresponds to D55, H58 and E62
1658 in the structure of colicin E3 (PDB:2xfz) [137]. However,
1659 our analysis indicated that H58 is not conserved across
1660 all members; instead we found that a second histidine,
1661 corresponding to H72 in Colicin E3, is conserved
1662 throughout the fold. Thus, it is possible that the above
1663 types of RNases use a single histidine in conjunction
1664 with an acidic residue that initiates cleavage by inducing
1665 the 2'OH to attack the phosphodiester backbone of
1666 RNA [137]. In contrast, examination of the multiple
1667 alignments of the novel toxins revealed potential metal-
1668 chelating sites in Ntox29 (conserved histidines and aspar-
1669 tates); hence, they could potentially function as novel

metal-dependent nuclease. For the remaining Ntox 1670
domains, while the active site residues could be identified 1671
based on conservation, the nature of catalysis remains 1672
unclear. 1673

1674 Deaminases

1675 Other than the nuclease domains, deaminases are the
1676 most common toxin domains that operate on nucleic
1677 acids in polymorphic toxin systems. As we had exten-
1678 sively characterized the toxin deaminases from these sys-
1679 tems in our earlier study [18], we do not consider them
1680 in detail here. However, in this study we recovered two
1681 additional clades of deaminases that were not previously
1682 detected (Figure 10A). The first of these was found in 1682 **F10**
1683 giant proteins with a toxin-like architecture from the
1684 alphaproteobacterial endosymbionts of the genus *Wol-*
1685 *bachia*, which reside in the cells of two dipterans,
1686 namely *Culex* (gi: 190571717; WPa_1346) and *Drosoph-*
1687 *ila* (gi: 42520377, WD0512). These proteins contain two
1688 toxins at their C-termini, of which the Latrotoxin-CTD
1689 (see below) is the terminal toxin and the deaminase N-
1690 terminal to it (Figure 10). An examination of their gene
1691 neighborhoods revealed that they lacked accompanying
1692 genes encoding immunity proteins. Hence, it appears
1693 that these proteins, while resembling the classical poly-
1694 morphic toxins, are primarily directed against host nu-
1695 cleic acids. The deaminase domains from these proteins
1696 are extremely divergent, but structure prediction based
1697 on a multiple alignment with a comprehensive set of de-
1698 aminase domains showed that they belong to the “Helix-4
1699 division” of the deaminase superfamily in which the
1700 5intervening 4th helix of the core domain causes strands
1701 4 and 5 to be parallel to each other [18]. Thus, they are
1702 united with other deaminases of this division such as
1703 TadA/Tad2, ADAR/TAD1 and the AID/APOBEC-like
1704 deaminases. However, unlike most members of this div-
1705 ision the newly characterized deaminase domains have a
1706 CXE signature in their first active site motif, as opposed
1707 to usual HXE seen in this division (Additional File 1).
1708 These newly detected versions add to the earlier iden-
1709 tified deaminases belonging to the Helix-4 division
1710 among host-directed toxins of alphaproteobacterial endo-
1711 symbionts/parasites, such as those from the *Wolbachia*
1712 endosymbiont of the lepidopteran *Cadre cautella* and
1713 from the *Orientia* and *Rickettsia* species infecting diverse
1714 eukaryotes[18]. This suggests that modification of nucleic
1715 acids by these fast-evolving deaminase toxins related to
1716 the eukaryotic AID/APOBEC-like proteins might be a
1717 widely used strategy by endosymbionts to alter host
1718 physiology. In particular, the presence of such highly
1719 divergent versions of deaminases in *Wolbachia* infecting
1720 diverse arthropods hints that they could be attractive
1721 candidates for mediating failure of paternal chromosome
1722 condensation via its mutagenic action [138]. The second



1723 novel clade of deaminases are toxin domains of classical
 1724 polymorphic toxins from proteobacteria and actinobac-
 1725 teria, which might be delivered via a diverse secretory
 1726 mechanisms such the T2SS, T5SS, T6SS, T7SS and the
 1727 TcdB/TcaC system (prototyped by gi: 162451789,
 1728 sce3516 from *Sorangium cellulosum*; Figure 10A and
 1729 Additional File 1). These deaminases usually have a
 1730 HAE signature in their first active site motif but belong
 1731 to the “C-terminal hairpin” division of the deaminase
 1732 superfamily, which is characterized by a C-terminal β -
 1733 hairpin following the 3rd-helix of the conserved core.
 1734 Given their predominance in free-living bacteria, unlike
 1735 the former deaminases, are likely to be deployed in
 1736 intraspecific conflict rather than against eukaryotic
 1737 hosts.

Other catalytic toxin domains in polymorphic toxin systems

1738 Other than the peptidase and nucleic acid cleaving or
 1739 modifying toxins we uncovered several other less fre-
 1740 quent catalytic domains that function as toxins in poly-
 1741 morphic and related secreted toxin systems (Table 2).
 1742 These display a wide range of activities and are likely to
 1743 elicit their cytotoxic activity by attacking several inde-
 1744 pendent aspects of cellular function. We briefly outline
 1745 these toxin domains and their possible modes of action.
 1746

Domains catalyzing modifications of proteins

1748 The previously characterized DOC domain, which has
 1749 been observed in several host-directed effectors (e.g.
 1750 *Xanthomonas AvrAC*), is found in several polymorphic
 1751

1752 toxins [22,139,140] (Figure 9D). This is a protein-
1753 modifying toxin domain, which transfers AMP or UMP
1754 from nucleotide triphosphates to serines or threonines
1755 on target proteins [139,140]. Another toxin domain that
1756 we recovered in polymorphic-toxin- related systems util-
1757 izing the PVC-SS showed a specific relationship to the
1758 serine/threonine kinase domain found in lantibiotic
1759 synthetases [141] (Figure 5C). The “eukaryote-type” kin-
1760 ase domain in the lantibiotic synthetases phosphorylates
1761 serine/threonine residues in the lantibiotic precursors
1762 to prime them for the generation of the thioether lin-
1763 kages. Lantibiotic synthetase-type kinase domains have
1764 been shown to possess generic S/T kinase activity
1765 [142], suggesting that the toxin versions might carry
1766 out their action by phosphorylation of proteins on S/T
1767 residues in target cells. A comparable protein-modifying
1768 toxin domain (gi: 291451822, from *Streptomyces albus*,
1769 Figure 5C) is a glycosyltransferase, related to the *Clos-*
1770 *tridium difficile* toxin B, which has been shown to gly-
1771 cosylate the hydroxyl group of threonine 37 in the
1772 switch I region of the small GTPase RhoA [143]. Given
1773 the conservation of the Mg²⁺-binding DXD signature,
1774 which is critical for catalyzing the transfer of UDP-
1775 linked sugars, in versions of this domain found in toxin
1776 polypeptides detected in our study, it is likely that it
1777 functions in a similar fashion by glycosylating serines or
1778 threonines in specific proteins in target cells. In
1779 addition to its presence in classical polymorphic toxins
1780 with N-terminal RHS repeats and PVC-SS delivered
1781 toxins, we observed that related glycosyltransferase
1782 domains are also found in effector proteins delivered by
1783 various intracellular bacteria. In the endoparasite *Le-*
1784 *gionella pneumophila* it is present in a toxin delivered
1785 via the T4SS (gi: 307610704) and in the aphid endo-
1786 symbiont *Hamiltonella defensa* (gi: 238899322) it might
1787 be deployed as a toxin against the parasitoid wasps that
1788 attack the host aphids [144]. A distinct protein-
1789 modifying toxin domain is typified by the CNF domain
1790 of the uropathogenic *E. coli* cytotoxic necrotizing factors
1791 1 and 2 and the dermonecrotic toxins of *Bordetella*.
1792 These domains display a 4-layered sandwich fold, with
1793 an active site histidine and cysteine, and catalyze the
1794 deamidation or transglutamination of a specific active
1795 site glutamine in the small GTPases, like RhoA, Rac and
1796 CDC42, in the cells of their eukaryotic host [140]. We
1797 recovered CNF domains in potential proteobacterial
1798 polymorphic toxins (Figure 10B) with N-terminal fila-
1799 mentous regions (*Yersinia* sp. yenC1, gi: 109391485) as
1800 well as those fused to phage-tail VgrG domains of the
1801 T6SS (e.g. 345371919 from *E.coli*).

1802 We also encountered several distinct clades of ADP
1803 ribosyltransferases (ARTs) among the toxin domains of
1804 polymorphic and related toxin systems (Figure 10C)
1805 [145]. The ART superfamily can be divided into two

major clades depending on the conservation pattern of 1806
the three key active site residues associated with the three 1807
conserved motifs, respectively from the N-terminus, cen- 1808
tral region and C-terminus of the domain. These are the 1809
R-S-E clade and the H-Y-E clade, named after their re- 1810
spective conserved active site residues [146-148]. Protein- 1811
modifying ART domains have been extensively studied in 1812
the context of the host-directed toxins of diverse bacteria. 1813
Members from the R-S-E clade include the cholera toxin, 1814
which modifies a specific arginine in a mammalian G α 1815
subunit, the *Bordetella pertussis* toxin which modifies 1816
cysteine, the *Clostridium botulinum* C3 toxin that modi- 1817
fies asparagine, and the *Photobacterium luminescence* 1818
toxin which modifies glutamine in target proteins 1819
[145,148]. The H-Y-E clade includes the *Corynebacterium* 1820
diphtheria, *Vibrio cholix* and *Pseudomonas aeruginosa* 1821
exotoxin A toxins, which modify diphthamide in the 1822
translation GTPase eEF-2, and the polyADP ribosyl 1823
transferases (PARP/PARTs) [146,149,150]. We found 1824
multiple R-S-E clade ART domains in classical poly- 1825
morphic toxin systems. One type of R-S-E clade ART 1826
toxin domains, observed in certain polymorphic toxins 1827
(e.g. gi: 221200352 from *Burkholderia multivorans*), are 1828
also seen in the T3SS effectors of *Pseudomonas syringae*, 1829
namely hopO1-1/2/3, a *Legionella pneumophila* T4SS ef- 1830
fector (gi: 307611385), a novel *Protochlamydia amoeboph-*
1831 *ila* effector (pc1346; gi: 46446980), and *Pseudomonas*
1832 *aeruginosa* exoT (gi: 347302423). Such ART toxin
1833 domains are also found in a remarkable group of giant
1834 proteins from actinobacteria (e.g. 345002682; *Strepto-*
1835 *myces* sp.; Figure 10), which combine several toxin
1836 domains such as two anthrax lethal factor-like metallo-
1837 peptidase, two caspase, three ART and one MCF-SHE
1838 domains (Figure 10). A second distinct type of R-S-E
1839 clade ART domains, which is found in similar actinobac-
1840 terial toxins (e.g., gi: 320008023 from *Streptomyces flavo-*
1841 *griseus*), is closely related to the lepidopteran ARTs, such
1842 as pierisin, which ADP-ribosylates the N2 atom of guan-
1843 ine in DNA to induce apoptosis and the insecticidal
1844 toxin of *Bacillus sphaericus* [151]. Interestingly, the close
1845 relationship of the lepidopteran pierisin-like ARTs to the
1846 bacterial insecticidal toxins suggests that they were prob-
1847 ably a late lateral transfer into these insects from a bac-
1848 terial symbiont or parasite, followed by their reuse as an
1849 apoptotic effector. In this study we found novel toxins of
1850 the H-Y-E clade from actinobacteria, which are closely
1851 related to the eukaryotic PARPs (Tox-ART-PARP), and
1852 are associated with the PVC-SS from (e.g. gi: 291451874
1853 from *Streptomyces albus*). We also identified related
1854 toxin domain among the toxins secreted by the intracel-
1855 lular pathogen *Legionella drancourtii* (e.g. LDG_5757; gi:
1856 374260808). Additionally, we also found three distinct
1857 families of toxin ARTs belonging to the H-Y-E clade. The
1858 first of these is an extremely divergent version, which is
1859

1860 typified by a protein with an architecture similar to a
1861 classical polymorphic toxin from *Shewanella baltica* (gi:
1862 152999126), but without associated immunity proteins
1863 and might be directed against eukaryotic hosts. The two
1864 other families (Tox-ART-HYD1 and 2 prototyped by gi:
1865 336178949 and gi: 238064042 respectively) are widely
1866 distributed in free-living bacteria and are associated with
1867 distinct immunity proteins suggesting that they might
1868 be mainly deployed in intraspecific conflict like the clas-
1869 sical polymorphic toxins. Nevertheless, versions of Tox-
1870 ART-HYD2 appear to have been transferred to several
1871 eukaryotes such as fungi and choanoflagellates (e.g. gi:
1872 331216471 from *Puccinia graminis*). The above observa-
1873 tions suggest that the use of ARTs to modify proteins,
1874 and in some cases DNA, appears to be yet another strat-
1875 egy that is common to effectors deployed in both intra-
1876 bacterial and bacterio-eukaryotic conflicts.

1877 **Lipid-modifying toxin domains**

1878 Three distinct lipid-modifying enzymes are represented
1879 among the toxin domains of classical polymorphic toxins
1880 and related PVC-SS-delivered toxins. Two of these
1881 namely the glycerophosphoryldiester phosphodiesterase
1882 (GPDase, gi: 218438711 from *Cyanothece*) and the CDP-
1883 alcohol phosphatidyltransferase (CAPTase, gi: 317401091
1884 from *Neisseria mucosa*) domains are found exclusively in
1885 PVC-SS toxins (Figure 5C). In contrast, phospholipase
1886 A2 (PLA2) is found in classical polymorphic toxins with
1887 filamentous N-terminal regions (e.g. gi: 118578532 from
1888 *Pelobacter propionicus*), which might be secreted via dif-
1889 ferent mechanisms, including the T6SS (Figure 10D). Of
1890 these the GPDase can catalyze the hydrolysis of glycer-
1891 ophospholipid head groups by releasing alcohols linked to
1892 glycerol 3-phosphate via a phosphodiester linkage [152].
1893 On the other hand, phospholipase A2 can hydrolyze
1894 lipids by releasing of one of the fatty acid tails from gly-
1895 cerol 3-phosphate [153]. Closely related homologs of the
1896 Tox-phospholipase A2 domains (Tox-PLA2) are also
1897 found in secreted proteins from fungi and oomycetes
1898 (Table 2, Additional File 1). More generally, phospholip-
1899 ase A2 domains are also found in animal toxins from
1900 reptilian venom and from mammalian immune systems
1901 [152], suggesting that use of this domain as a toxin is a
1902 prevalent strategy throughout evolution. Intriguingly,
1903 members of the CAPTase superfamily are membrane-
1904 embedded enzymes catalyzing the reverse reaction (lipid
1905 synthesis) using cytidine-diphosphate-linked alcohols as
1906 substrates, e.g. phosphatidylserine, phosphatidylcholine,
1907 phosphatidylglycerolphosphate, phosphatidylinositol and
1908 cardiolipin synthetases [154]. It is conceivable that a
1909 novel lipid synthesized by this toxin domain creates dis-
1910 continuities in lipid bilayers, as has been observed with
1911 cardiolipin [155]. Thus, all three of these enzymes could
1912 potentially mediate their cytotoxicity by damaging the

cell membrane of target cells, either through hydrolysis
of lipids or disruption of the bilayer.

A toxin domain was uncovered in several classical
polymorphic toxins (e.g. Tmz1t_2699 from *Thauera* sp.;
gi: 237653364) that partly overlapped with a “domain of
unknown function” (DUF2235 in the PFAM database).
Sequence profile searches with the PSI-BLAST program
recovered significant hits to α/β hydrolases ($e = 10^{-5}$ - 10^{-7} ;
iteration 3 in a search initiated with the domain from
the above *Thauera* protein). While α/β hydrolase super-
family encompasses hydrolases with several distinct ac-
tivities, such as lipases, peptidases and thioesterases,
profile-profile comparisons with the HHpred program
suggested that these α/β hydrolases (Tox-ABhydrolase-1)
are closest to lipases (e.g. the recovery of triacylglycerol
lipases; PDB: 1tgl). In most cases this α/β hydrolase do-
main is either found fused to N-terminal phage base-
plate modules (e.g. gi: 77461818 from *Pseudomonas*
fluorescens) or encoded by a gene adjacent to a gene
coding for such modules (Figure 10E). This suggests
that Tox-ABhydrolase-1 might be a toxin that is mainly
delivered via T6SS. These α/β hydrolase domains also
appear to have been transferred to fungi prior to the di-
vergence of the ascomycetes and the basidiomycetes and
are present in most fungal lineages. We recovered two
more distinct, previously uncharacterized α/β hydrolase
families that are potential toxin domains that are asso-
ciated with numerous classical polymorphic toxins
(Tox-ABhydrolase-2 and 3, Figure 10E). Profile-profile
searches with ABhydrolase-3 recovers the lipases (e.g.
pdb: 1lgy; $p = 10^{-12}$; probability 95%) as the best hit to
the exclusion of other ABhydrolases. Hence, it is con-
ceivable that Tox-ABhydrolase-1 and Tox-ABhydrolase-
3 are further toxins that might disrupt cell-membranes
of target cells via their action on lipids. ABhydrolase-2
is primarily present in proteobacteria and has also been
transferred to ascomycete fungi. It is also found in the
endosymbiont *Parachlamydia amoebophilus* independ-
ently of an immunity protein and might be deployed
against host molecules. However, Tox-ABhydrolase-2 did
not show any specific relationship to previously charac-
terized lipases. Given, that the ABhydrolase superfamily
includes hydrolases with a very diverse array of activities,
it is not clear if Tox-ABhydrolase-2 might also act on
lipids or target some other cellular component.

1958 **Carbohydrate-related toxin domains**

1959 We detected two enzymatic domains, which are pre-
1960 dicted to act on carbohydrate substrates, as toxin
1961 domains of polymorphic and PVC-SS-delivered toxins.
1962 The first of these belongs to a superfamily of glycohy-
1963 drolases, typified by bacterial proteins, such as FlgJ and
1964 the N-acetylmuramoyl-L-alanine amidase (gi: 220928985
1965 from *Clostridium cellulolyticum*), which cleave the

glycopeptide linkages in peptidoglycan or endo-
glycosidic linkages in oligosaccharides [156,157]. Hence,
it is likely that these toxin domains act by hydrolyzing
linkages in the peptidoglycan of the target cells. These
might be compared to the recently described amidase
toxins from *Pseudomonas aeruginosa* that are believed
to act on peptidoglycan [15]. The second toxin domain
in this group is an oxidoreductase with a TIM barrel fold
catalytic domain (gi: 158339325 from *Acaryochloris mar-*
ina) [158]. Within this superfamily, the toxin domains
are most closely related to the aldo-keto reductases,
such as 2,5-didehydrogluconate reductase, suggesting
that they are likely to act on sugar substrates. However,
the exact mode of action of this toxin remains unclear –
it could either act on carbohydrates in the peptidoglycan
or within target cells.

1982 **Toxin domains related to nucleotide signaling**

1983 The RelA/SpoT-like toxin domain is found in classical
1984 polymorphic toxins from Gram-positive bacteria deliv-
1985 ered by the ESX/T7SS (e.g. 302865491; Micau_0989
1986 from *Micromonospora aurantiaca*; Figure 10D). A
1987 related toxin domain is also found in the T3SS-delivered
1988 effectors directed against plant hosts by several plant
1989 pathogens, such as *Xanthomonas* (e.g. gi: 353464269; the
1990 XopAD effector), *Ralstonia solanacearum* and *Pseudo-*
1991 *monas syringae*. These proteins typically contain two
1992 copies of the RelA/SpoT domain. Further, in several bac-
1993 teria (e.g. gi: 149004362 from *Streptococcus pneumoniae*
1994 and gi: 254362874 from *Mannheimia haemolytica*)
1995 RelA/SpoT toxin domain is found fused to the MuF do-
1996 main of prophages and is thereby predicted to be deliv-
1997 ered via this distinct phage-derived system. The RelA/
1998 SpoT is a nucleotide-binding domain related to the
1999 DNA polymerase β -type nucleotidyltransferase fold
2000 [159] that synthesizes the alarmone (p)ppGpp [160]. It
2001 has been observed that high levels of (p)ppGpp in non-
2002 starvation conditions rapidly inhibits growth and protein
2003 synthesis [160]. Hence, it is conceivable that this toxin
2004 acts as an unregulated alarmone synthetase in target
2005 cells to shut down their protein synthesis. Its widespread
2006 presence in several phylogenetically distant plant patho-
2007 gens is consistent with the presence of a (p)ppGpp-
2008 dependent signaling pathway in plants, similar to that
2009 seen in bacteria [160]. In light of this, it appears likely
2010 that the MuF-fused versions found in the animal patho-
2011 gens such as *Streptococcus pneumoniae* and *Mannhei-*
2012 *mia haemolytica* might be deployed in intra-bacterial
2013 conflict similar to the classical polymorphic toxins, ra-
2014 ther than against the animal hosts.

2015 Another distinct nucleotide generating enzymatic do-
2016 main, which we found in several polymorphic toxins
2017 from several major bacterial lineages (Figure 10C), is the
2018 ADP-ribosyl cyclase (Tox-ARC) domain. These toxins

are coupled to various delivery systems including T5SS, 2019
T6SS and T7SS. This domain has previously only been 2020
characterized in animals and generates two distinct 2021
metabolites, namely cyclic ADP ribose (cADPr) and 2022
nicotinic acid adenine dinucleotide phosphate (NAADP), 2023
respectively from NAD and NADP [161]. The former two 2024
nucleotides have been shown to function as potent indu- 2025
cers of calcium influx via the ryanodine receptors [162]. 2026
At the same time by channeling NAD it can also affect 2027
protein deacylation by Sirtuins and other processes re- 2028
quiring NAD [163]. Given that polymorphic toxins with 2029
Tox-ARC domains occur in free-living bacteria, and are 2030
typically coupled with the genes for the immunity protein 2031
Imm74, it is likely that they are used in intra-specific con- 2032
flict rather than against eukaryotes. Their mode of action 2033
in the bacterial context is not entirely clear – it is possible 2034
that they deplete NAD or NADP and interfere with vari- 2035
ous metabolic processes dependent on them. Alternativa- 2036
ly, the cADPr or NAADP generated by them could 2037
have toxin consequences for the target cell, for example 2038
by interfering with NAD-utilizing process such as RNA 2039
metabolism or DNA ligation. The bacterial Tox-ARC 2040
domains show considerably more sequence diversity than 2041
the eukaryotic counterparts and appear to have been the 2042
progenitors of two independent sets of eukaryotic repre- 2043
sentatives in animals and fungi respectively. 2044

2045 **Non-catalytic toxins: Pore-forming and peptidoglycan-** 2046 **binding domains**

2047 Several classical polymorphic and PVC-SS delivered
2048 toxin proteins display unusual C-terminal predicted
2049 toxin domains that do not show any indications of being
2050 enzymes. Further analysis of these predicted toxin
2051 domains suggested that they are likely to operate via
2052 non-catalytic mechanisms. One of these, which is thus
2053 far restricted to proteobacteria is the W-TIP domain that
2054 was named after a conserved tryptophan and TIP tripep-
2055 tide motif (Figure 10F). This small toxin domain is
2056 highly hydrophobic in composition and is predicted to
2057 form two membrane spanning-helices. The first of these
2058 helices bears two absolutely conserved positively charged
2059 residues (RxxR signature), while the second bears the
2060 W-TIP motif. These features suggest that the W-TIP
2061 toxin domain might effect its cytotoxicity by forming a
2062 transmembrane pore similar to pore-forming toxins
2063 from diverse organisms [164,165]. Several PVC-SS deliv-
2064 ered toxins also display a single annexin domain
2065 (Figure 5C); however, this domain is unlikely to be a
2066 stand-alone toxin domain as it is always followed by a
2067 further C-terminal *bona fide* enzymatic toxin domain
2068 (e.g. the anthrax lethal factor-like metallopeptidase and
2069 Ntox3 domains; Figure 5C). The eukaryotic annexins
2070 typically contain four tandem annexin domains and bind
2071 both phospholipids, such as phosphatidylinositol (4,5)-

2072 bisphosphate (Annexin A2) and phosphatidylserine
2073 (Annexin A5), or components of lipid rafts such as chol-
2074 esterol (Annexin A2) [166]. The eukaryotic annexins
2075 also have the unusual capability of apparently traversing
2076 cell membranes despite lacking signal peptides. Hence, it
2077 is conceivable that the annexin domains in bacterial tox-
2078 ins act as accessory domains that aid in the breaching of
2079 target cell membranes to facilitate the delivery of the C-
2080 terminal toxin domain.

2081 One of the most enigmatic toxins is Ntox38
2082 (Figure 10G), which is currently restricted to actinobac-
2083 teria, and might be found in several paralogous copies
2084 per genome (e.g. 7 copies in *Actinosynnema mirum* and
2085 9 copies in *Saccharopolyspora spinosa*). This toxin do-
2086 main is usually linked to a N-terminal WXG domain by
2087 a low-complexity glycine-rich linker, suggesting that it is
2088 secreted via the T7SS. This is further supported by the
2089 frequent presence in their gene neighborhoods of a gene
2090 encoding a subtilisin-like serine peptidase associated with
2091 processing of proteins secreted via the T7SS [126]. The
2092 Ntox38 domain is just 33–43 residues in length and is
2093 predicted to adopt a simple three-stranded fold
2094 (Figure 10G). Its size and lack of potential conserved
2095 catalytic residues suggest that it is unlikely to be an en-
2096 zymatic domain. It shows several, conserved hydropho-
2097 bic residues and an invariant C-terminal PXhhG
2098 signature (where h is a hydrophobic residue). It is one of
2099 the few toxin domains whose mode of action remains ra-
2100 ther elusive, but is likely to involve a physical interaction
2101 with a key cellular component rather than catalytic
2102 modification. It shows a strong association with a single
2103 immunity protein, Imm56.

2104 We uncovered an unusual toxin domain at the C-
2105 termini of giant toxin proteins from arthropod alphapro-
2106 teobacterial and gammaproteobacterial endosymbionts
2107 such as *Wolbachia* and *Rickettsiella grylli* (Figure 10H).
2108 Homologous domains are also found at the C-termini of
2109 the latrotoxins (latrotoxin-CTD) of the black widow
2110 spider (*Latrodectus* species) [167]. The latrotoxins also
2111 display other architectural similarities with the above
2112 bacterial toxins in sharing N-terminal ankyrin repeats.
2113 Interestingly, the latrotoxins are not secreted in a con-
2114 ventional fashion, but released upon disintegration of
2115 the producing cell [167]. Upon release the latrotoxin-
2116 CTD is proteolytically cleaved off to form the mature
2117 latrotoxin [168]. Given that the latrotoxin-CTD is shared
2118 by distantly related bacterial endosymbionts, which
2119 colonize a wide range of arthropods, it appears likely
2120 that the spider latrotoxins were acquired via lateral
2121 transfer from a bacterial endosymbiont. The latrotoxin-
2122 CTD is characterized by a conserved, hydrophobic helix;
2123 hence, it is possible that it associates with the membrane
2124 and might facilitate disintegration of the producing cells
2125 in spiders. Bacterial toxins with latrotoxin-CTDs do not

2126 display any neighboring immunity protein genes; hence,
2127 it is likely that they are primarily used against the
2128 eukaryotic hosts. In this regard, it is interesting to note
2129 that the salivary gland proteins of mosquitoes have been
2130 suggested as being laterally transferred from *Wolbachia*
2131 [169,170]. We found that such proteins are more widely
2132 distributed across arthropods (e.g. the crustacean *Daph-
2133 nia pulex*), and that they are related to endosymbiont
2134 toxin proteins, such as those reported above. However,
2135 in place of a C-terminal toxin domain they contain a
2136 conserved domain termed the SGS domain (for salivary
2137 gland secreted protein), which is not found in any bac-
2138 terial toxin, but only in arthropods (Figure 10H, Addi-
2139 tional File 1). Thus, it appears that following lateral
2140 transfer of a bacterial toxin protein, the toxin domain
2141 was displaced by an arthropod-specific domain. Hence,
2142 the latrotoxin and SGS proteins could represent different
2143 examples of toxins of endosymbiotic bacteria being
2144 coopted for arthropod-specific functions.

2145 Several toxins delivered via the PVC-SS displayed a pu-
2146 tative toxin domain belonging to the OmpA superfamily
2147 of peptidoglycan-binding domains [171-173] (e.g. gi:
2148 171059731 from *Leptothrix cholodnii*; Figure 5C). While
2149 several toxin polypeptides contain domains that might
2150 facilitate extracellular adhesion, including peptidoglycan-
2151 binding domains such as a PGB1 and the LysM domains,
2152 the OmpA domain, unlike those, always occurred at the
2153 extreme C-terminus. This supports the inference that in
2154 these cases the OmpA domain might have a toxin func-
2155 tion. The OmpA domains have been shown to anchor
2156 porins and the T6SS to the peptidoglycan [172-174].
2157 Given that OmpA domains can bind peptide precursors
2158 for peptidoglycan biosynthesis [172], it is possible that
2159 such toxin domains might act by interfering with pep-
2160 tidoglycan synthesis through binding of such peptides.

2161 Lineage-specific expansion of N-terminal domains in toxin 2162 proteins: Novel secretion/anchoring mechanisms?

2163 The N-terminal domains of the full length polymorphic
2164 toxins are usually good predictors of their trafficking
2165 pathways because they contain domains that are specific
2166 to a given secretory pathway (Table 1). We found another
2167 interesting feature in the N-terminal regions of certain
2168 polymorphic toxins and related proteins from endo-
2169 symbionts/parasites secreted via the T2SS, which is thus
2170 far restricted to a few bacteria. This feature is character-
2171 ized by the presence of lineage-specific domains that
2172 occurs downstream of a N-terminal signal peptide in full-
2173 length toxins from certain organisms. The best example
2174 of this is provided by the MAFB group of polymorphic
2175 toxins found in *Neisseria* species (Figure 10I). Here all the
2176 full-length toxin proteins display a globular domain, the
2177 MAFB-N domain (Additional file 1; overlapping but not
2178 identical to the model defined as the domain of unknown

2179 function DUF1020 in the PFAM database), just after their
2180 signal peptide. Across different full length toxins the
2181 MAFB-N domain is highly conserved, which is in sharp
2182 contrast to the C-terminal polymorphism in their toxin
2183 domains (Figure 10I). Furthermore, though the MAFB-N
2184 domain is strongly conserved in the genus *Neisseria*, the
2185 MAFB-N domain is not found outside of it. In terms of
2186 operonic organization, all full-length genes encoding
2187 MAFB-N type polymorphic toxins are accompanied by an
2188 upstream gene which encodes MAFA, a secreted protein
2189 with a lipobox, indicating that it is a lipid anchored sur-
2190 face protein [175]. Like the MAFB domain, the MAFA
2191 domain is restricted to *Neisseria* and shows no poly-
2192 morphism. This suggests that the conserved MAFB do-
2193 main of these polymorphic toxins is likely to interact with
2194 the surface-anchored MAFA protein, thereby anchoring
2195 them to the cell surface. This hinted that certain lineage-
2196 specific N-terminal domains might serve as a surface an-
2197 chor for toxins. A comparable situation was observed in a
2198 group of seven polymorphic toxins in *Microscilla marina*,
2199 which are typified by a conserved N-terminal domain up-
2200 stream of their signal peptides (Microscilla-N). This con-
2201 served globular domain is currently not observed outside
2202 of this species and might again play a specific anchoring
2203 function for these polymorphic toxins. It is also conceiv-
2204 able that homotypic interaction between these “constant”
2205 N-terminal domains help spatial clustering of different
2206 toxins on the cell surface.

2207 Like *Microscilla*, yet another member of the bacteroi-
2208 detes clade, i.e. the *Acanthamoeba* endosymbiont *Amoe-*
2209 *bophilus asiaticus* displays a variety of effectors, which
2210 are predicted to be directed against its eukaryotic host,
2211 that are united by shared conserved N-terminal
2212 domains. We were able to identify two distinct types of
2213 such N-terminal domains that occur immediately down-
2214 stream of a signal peptide and a lipobox, that we termed
2215 Amoeboprodomain 1 (APD1) and 2 (APD2) respectively
2216 (Additional File 1). The presence of the lipobox prior to
2217 APD1 and APD2 suggests that these effectors do not dif-
2218 fuse into the host cytoplasm, but are likely to be
2219 anchored on the surface of endosymbiont. The proteins
2220 bearing the APD1 and APD2 domains show highly con-
2221 served N-termini but extremely polymorphic C-termini,
2222 with several distinct effector domains – thus, they appear
2223 to represent a mechanistic principle similar to the
2224 MAFB-N and *Microscilla* toxin N-terminal domains.
2225 However, unlike the classical polymorphic toxins, where
2226 the C-terminal domains are serially variable due to dis-
2227 placement by alternative toxin domain cassettes, the
2228 *Amoebophilus* effectors with diverse C-termini are likely
2229 to be deployed in parallel at the same time [79]. Among
2230 the variable C-terminal domains of these effectors are
2231 several domains shared with the toxin domains of poly-
2232 morphic toxin systems, such as: 1) papain-like peptidases

of the Otu family; 2) lipase-like α/β hydrolases; 3) The
EDA39C-like nucleases. Additionally, these effectors also
display diverse C-terminal domains that are specifically
related to the ubiquitin system, such as the F-box and
U-box subunits of ubiquitin E3 ligases, SMT4/Ulp1-like
desumoylating and UBCH-like deubiquitinating peptidases,
and other regulatory modules such as the GIMAP-type
GTPase domains, STAND NTPase domains, SecA-like
helicase-related domains and SbcC-like ATPase domains
[79,176,177]. This suggests that over and beyond typical
toxin-like effectors, the *Amoebophilus* effectors also inter-
face with the host via a wide range of catalytic activities
that are typically not encountered in the polymorphic
toxin systems. Indeed, the deployment of effectors inter-
acting with the eukaryotic Ub-system is a common
strategy used by several endo-symbiotic/parasitic bac-
teria as well as exoparasitic bacteria that deliver effec-
tors via different secretory systems [80]. On the other
hand deployment of STAND NTPases and GIMAP-type
GTPases is a strategy limited to endo-symbiotic/parasitic
forms. Nevertheless, the presence of the lineage-specific
APD1 and APD2 domains suggests that, as in the case
of the polymorphic toxin systems, these N-terminal
domains might mediate surface anchoring or homotypic
interactions that allow clustering of effectors to certain
locations on the cell surface. Given the lineage-specific
nature of this feature, it might turn out to be more wide-
spread upon more careful analysis.

Immunity proteins

Our earlier studies had revealed that two major immu-
nity protein superfamilies, namely SUKH and SuFu, domi-
nate the polymorphic toxin systems [17]. The current
study further corroborated this observation – systematic
comparisons revealed that members of the SUKH super-
family act as immunity proteins across the greatest
mechanistic and structural range of toxins. They were
found as immunity proteins for toxin domains belonging
to 18 distinct families of nucleases displaying eight dis-
tinct folds, three families of deaminases, DOC-like pro-
tein AMP/UMPylating enzymes, TIM-barrel aldo-keto
reductase, two types of α/β hydrolases and two mechan-
istically distinct peptidases (Table 3). We extended the
diversity of the SuFu superfamily by identifying a second,
previously unknown clade of SuFu domains (Table 3,
Additional File 1). These domains are extremely diver-
gent with respect to the classical SuFu domain but could
be unified with them by means of profile-profile com-
parisons ($p = 10^{-6}$; probability 86% for matching the clas-
sical SuFu superfamily profile). Together, the two clades
of SuFu domains are immunity proteins for toxins with
six families of nuclease domains of the HNH/EndoVII
fold, the ParB domain, Ntox7 nuclease domain, peptid-
ase domains belonging to two unrelated folds and the

2354 while there are over 50 different types of immunity pro-
2355 teins, with $\alpha + \beta$ domains being preponderant, only a few
2356 of them belong to previously characterized superfamilies
2357 of domains mediating protein-protein interactions in
2358 other sub-cellular contexts. Among these are Imm-
2359 NTF2 and Imm-NTF2-2 (NTF2 fold domain), Imm-
2360 MyosinCBD (related to the cargo-binding domain of the
2361 type VI myosins of animals), Imm-LRR (leucine-rich
2362 repeats), Imm-Ank (Ankyrin repeats) and Imm-HEAT
2363 (HEAT repeats), which display domains that are widely
2364 used in protein-protein interactions across several cellu-
2365 lar systems (Table 3). However, unlike the SUKH or
2366 SuFu superfamilies, none of these immunity proteins
2367 with versions of previously characterized interaction
2368 domains are widely used across different toxin types in
2369 the polymorphic toxin systems. Some otherwise com-
2370 mon protein-protein interaction domains used in other
2371 biological systems, such as the immunoglobulin or β -
2372 propeller domains, have not yet been found among im-
2373 munity proteins. This suggests that, rather than widely
2374 coopting common protein-protein interaction domains
2375 that are prominent in other sub-cellular systems, the
2376 polymorphic toxin systems have selected for their own
2377 unique set of proteins specializing in protein-protein
2378 interactions (Table 3). In the case of the SUKH and the
2379 SuFu superfamilies, evidence from gene neighborhoods
2380 and phyletic patterns suggests that they primarily func-
2381 tion in the context of the polymorphic toxin systems
2382 and were on several occasions secondarily adapted for
2383 other protein-protein interaction functions, especially in
2384 eukaryotes and viruses [17]. Interestingly, most immu-
2385 nity protein superfamilies are entirely absent in archaea
2386 (Table 3). This is consistent with the general paucity of
2387 classical polymorphic toxin systems in most archaea;
2388 though haloarchaea display functionally related PVC-SS
2389 delivered toxin systems (See below for further discus-
2390 sion). These observations also indicate that the poly-
2391 morphic toxin systems have provided a unique niche in
2392 bacteria for the innovation of a great variety of domains
2393 mediating distinctive protein-protein interactions, ma-
2394 jority of which are not utilized elsewhere. Nevertheless,
2395 at least 13 distinct types of immunity proteins have been
2396 transferred on different occasions to eukaryotes (Table 3).
2397 While some of these transfers to eukaryotes are ancient,
2398 the majority of these transfers are to fungi and diverse
2399 amoeboid eukaryotes which share micro-environments
2400 with bacteria. It would be of interest to investigate if
2401 these have been adapted for eukaryote-specific functions
2402 as observed in the case of the SUKH and SuFu super-
2403 families [17]. In conclusion, we suggest that a systematic
2404 structural investigation of the toxin-immunity protein
2405 interactions might offer a unique opportunity to study
2406 the evolutionary constraints acting on protein-protein
2407 interaction interfaces.

Polyimmunity loci and polyimmunity proteins

2408 Our earlier analysis had indicated the presence of tan- 2409
2410 dem arrays of genes encoding several distinct paralogous 2411
2412 immunity proteins of the SUKH superfamily, many of 2413
2414 which are often only distantly related to each other [17]. 2415
2416 We term these “polyimmunity loci”. Such polyimmunity 2417
2418 loci were suggested to function as potential backups that 2419
2420 allow organisms to survive not only their own toxins but 2421
2422 also neutralize a range of toxins that might be delivered 2423
2424 by non-kin strains that are present in the environment 2425
2426 [17]. Further, they might provide reservoirs of immunity 2427
2428 proteins that allow an organism to potentially “cover” 2429
2430 any new toxin it might evolve or acquire through lateral 2431
2432 transfer. In this study we systematically identified several 2433
2434 new polyimmunity loci and further extended this concept 2435
2436 to include homogeneous and heterogeneous poly- 2437
2438 immunity loci (Figure 11A): The homogeneous 2439
2440 polyimmunity loci are defined as those which are domi- 2441
2442 nated by a single type immunity protein e.g. several tan- 2443
2444 dem paralogs of the SUKH superfamily [18]. The most 2445
2446 frequently found homogeneous polyimmunity loci are 2447
2448 those containing tandem SUKH superfamily genes. In 2449
2450 addition, Imm6, Imm11, Imm28, Imm33, Imm36 and 2451
2452 Imm 41 also form prominent homogeneous polyimmu- 2453
2454 nity loci (Additional File 1). In contrast, the heteroge- 2455
2456 neous polyimmunity loci contain a wide range of 2457
2458 structurally unrelated immunity proteins. For example, a 2459
2460 heterogeneous polyimmunity locus from *Bacteroides* sp. 2461
2462 D22 encodes 19 different immunity proteins belonging 2463
2464 to 13 distinct superfamilies, of which the SUKH super- 2465
2466 family alone is represented by 6 distinct versions in this 2467
2468 locus (Figure 11A). As such these polyimmunity loci 2469
2470 represent a unique type of prokaryotic gene cluster – 2471
2472 they differ from other large prokaryotic gene clusters in 2473
2474 concentrating genes that are effectively functionally 2475
2476 equivalent in a certain sense rather than encoding mul- 2477
2478 tiple subunits of a protein complex (e.g. ribosomal or 2479
2480 CRISPR operons) or enzymes catalyzing successive steps 2481
2482 of a complex pathway (e.g. the antibiotic and siderophore 2483
2484 biosynthetic operons) [179,180]. 2485

2448 Examination of both polyimmunity loci reveals several 2449
2450 interesting features (Figure 11A and Additional File 1): 2451
2452 1) The immunity genes in a polyimmunity locus are 2453
2454 never interrupted by intervening toxin genes or toxin 2455
2456 cassettes. Thus, they are distinct from regular poly- 2457
2458 morphic toxin loci, which typically display arrays of tox- 2459
2460 ins or toxin cassettes, often with an adjacent immunity 2461
2462 protein. 2) The intergenic distance between two immu- 2463
2464 nity genes in a polyimmunity locus is typically small and 2465
2466 they are arranged in the same orientation. This implies 2467
2468 that they might be transcribed into a single polycistronic 2469
2470 message, from which multiple immunity proteins are 2471
2472 synthesized at once. This appears to distinguish them 2473
2474 from the immunity proteins located within a regular 2475
2476

2462 polymorphic toxin locus in which only the complete
2463 toxin gene and its adjacent immunity protein are
2464 expressed [181]. 3) The polyimmunity loci show consid-
2465 erable differences in terms of the number and type of
2466 included immunity genes, even between strains of the
2467 same species (Figure 11A). 4) In several cases the poly-
2468 immunity loci are adjacent to genes encoding recombi-
2469 nases, such as the XerC/D recombinase (Additional File
2470 1). It is conceivable that the recombination mediated by
2471 these adjacent elements might play a role in accumula-
2472 tion of immunity genes at polyimmunity loci. 5) Usually
2473 organisms possess only a single polyimmunity locus. A
2474 minority of the organisms possess more than one poly-
2475 immunity locus (~13% of the organisms with polyimmu-
2476 nity loci). 6) Extended polyimmunity loci (i.e. those with
2477 four or more tandem immunity genes) are not found in
2478 all bacterial lineages – thus far, they are only found in
2479 certain lineages of proteobacteria, bacteroidetes, firmi-
2480 cutes and actinobacteria. This suggests that extended
2481 polyimmunity loci are probably selected for only in cer-
2482 tain ecological settings (see below). Some of the above
2483 features indeed suggest that these loci are probably
2484 under selection to provide a preemptive defensive
2485 backup against a constantly changing profile of deployed
2486 toxins in context of frequent, recurrent organismal con-
2487 flicts (see below for further details).

2488 Comparable to the polyimmunity loci, are the polyim-
2489 munity proteins, which combine multiple immunity pro-
2490 tein domains into a single polypeptide (Figure 11B).
2491 Thus, they may be viewed as polyvalent immunity pro-
2492 teins that have the ability to neutralize more than one
2493 toxin simultaneously or serially. We first observed such
2494 polyimmunity proteins in the SUKH superfamily,
2495 wherein the same protein contains multiple tandem
2496 repeats of the SUKH domain [17]. Similarly, we observed
2497 that the SUKH domain might also be fused to SuFu and
2498 Imm33 (DUF2185) domains indicating that there are
2499 polyimmunity proteins, which combine structurally un-
2500 related immunity domains in the same polypeptide. A
2501 systematic search for polyimmunity proteins revealed
2502 several additional architectures (Figure 11B). Some of the
2503 largest polyimmunity proteins combine up to 10 distinct
2504 immunity domains in a single polypeptide (e.g., gi:
2505 160893617 from *Clostridium* sp. L2-50; Figure 11B).
2506 Given its prevalence as an immunity domain, not sur-
2507 prisingly, the SUKH domain is a common denominator
2508 in several of these polyimmunity proteins – it is com-
2509 bined with at least 8 structurally unrelated immunity
2510 domains in different polypeptides (Figure 11C). The
2511 other prominent domains in polyimmunity proteins are
2512 SuFu (combined with five other domains), Imm13,
2513 Imm33 and Imm-Ank (combined with four other
2514 domains) and, Imm11 and Imm34 (each with combina-
2515 tions to three other domains) (Figure 11C). The most

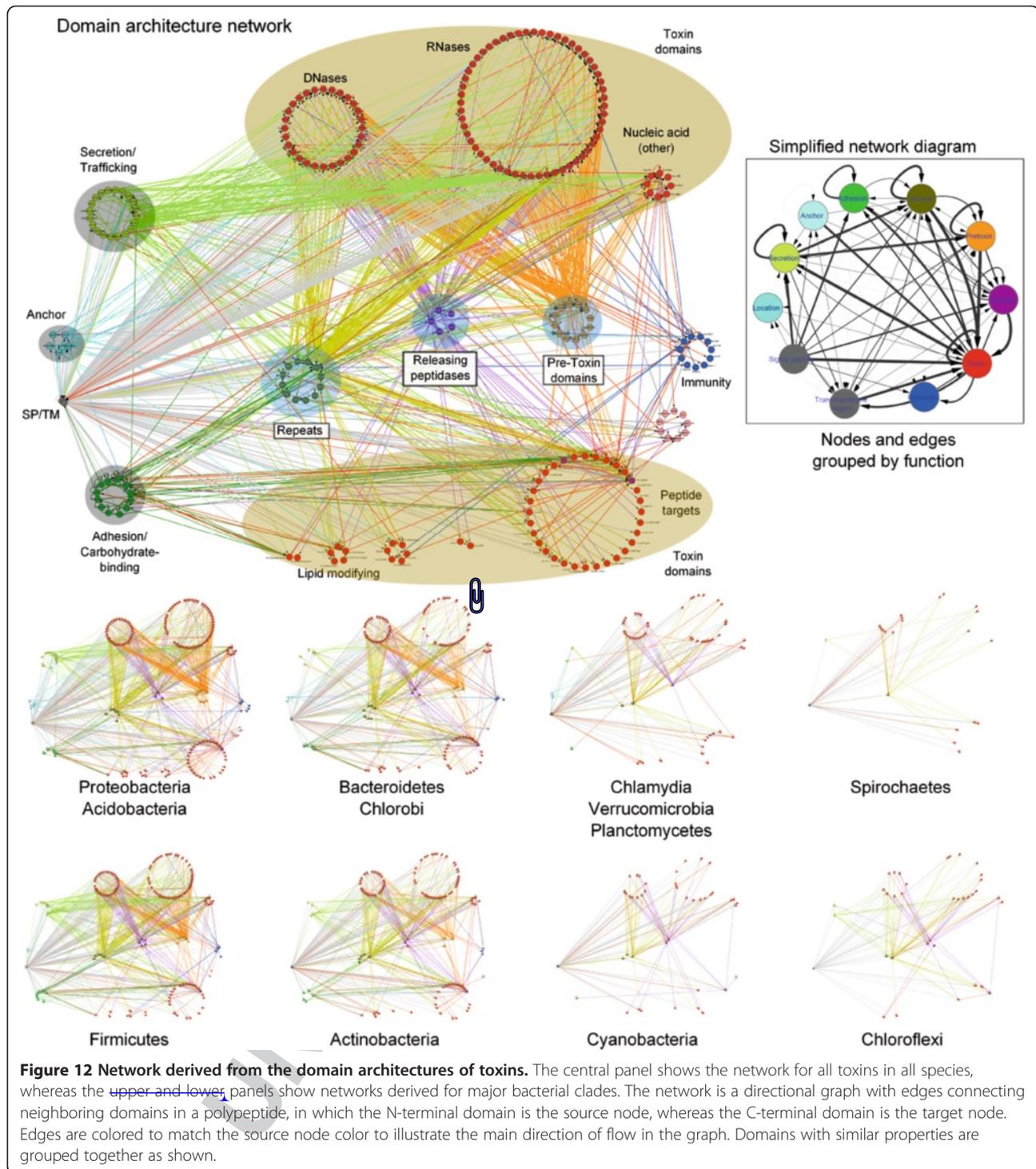
frequently found domain combinations in polyimmunity 2516
proteins with more than one type of immunity domain 2517
involve combinations between one or more of the follow- 2518
ing immunity domains: SUKH, SuFu (including SuFu- 2519
family 2), Imm-Ank, Imm5, Imm33, Imm34, Imm36, 2520
Imm66, Imm67, Imm68 and Imm69. Like the polyimmu- 2521
nity loci, the polyimmunity proteins are encoded in oper- 2522
ons, which usually do not contain associated toxin genes 2523
or cassettes. Interestingly, while polyimmunity proteins 2524
tend to be coded by small polyimmunity loci with two or 2525
three tandem immunity genes, they might not be found 2526
in the same bacteria with extended polyimmunity loci 2527
(see above) suggesting that the two are functionally 2528
related but distinct adaptations. Interestingly, some poly- 2529
immunity proteins have also been transferred to amoe- 2530
bozoan eukaryotes (Table 3, Additional File 1). 2531

Contextual features: Functional implications of gene- 2532 neighborhoods and domain architectures 2533

To better understand the functional aspects of the genomic 2534
organization of the polymorphic toxins and related 2535
toxin systems in terms of genomic organization, recom- 2536
bination, secretion and interactions with immunity pro- 2537
teins, we resorted to a systematic analysis of their gene 2538
neighborhoods and domain architectures of toxins. For 2539
the sake of visualization, we represented the connections 2540
emerging from both these types of analysis as directed 2541
graphs: In the case of domain architectures, the nodes in 2542
the graph are the individual domains and the edges are 2543
connections between two adjacent domains in a poly- 2544
peptide in the N- to C-terminal orientation. Each of the 2545
repetitive structures such as RHS and filamentous 2546
hemagglutinin repeats were treated as a single node 2547
(Figure 12). In the case of gene neighborhoods the nodes 2548 **F12**
are individual genes or toxin cassettes and the edges 2549
indicate their neighborhood relationship in the 5'->3' 2550
orientation (Additional File 1). 2551

Inferences from the gene neighborhoods 2552

The one pervasive feature of polymorphic toxins across 2553
most gene neighborhoods was the predominance of the 2554
toxin-immunity gene (TI) order, wherein the toxin gene 2555
is to the 5' end, while the immunity gene is to the 3' end 2556
of the operon (Figure 13). This tendency holds good for 2557 **F13**
both complete toxin genes encoding all the N-terminal 2558
domains, as well as individual toxin cassettes which only 2559
encode toxin domains. There are several implications of 2560
this gene organization: 1) The toxin is synthesized prior 2561
to the immunity protein during translation. As the toxin 2562
protein is targeted to one of the many secretion systems 2563
for delivery to the cell surface, it is unlikely to cause im- 2564
mediate “self-intoxication”, thereby obviating the need 2565
for a premade immunity protein. This is supported in 2566
experiments with toxins exported by the T5SS, where the 2567



2568 toxin is only activated in the target cell [183]. 2) Because
 2569 polymorphism is achieved by recombining different toxin
 2570 cassettes to a constant 5' gene body coding for trafficking
 2571 and presentation domains, there is the need for the re-
 2572 combination event to not only replace the 3' toxin cas-
 2573 sette [17,45], but also bring in its cognate immunity
 2574 gene. This feature explains why cassettes also occur as TI

pairs: On account of the TI organization of cassettes, a 2575
 single recombination event at the 3' tip of the complete 2576
 toxin gene can replace the existing toxin coding region 2577
 and simultaneously bring in the new immunity gene. Evidence for multiple such recombination 2578
 events is presented by the genomic organization of the full toxin genes. They often have a string of 2581

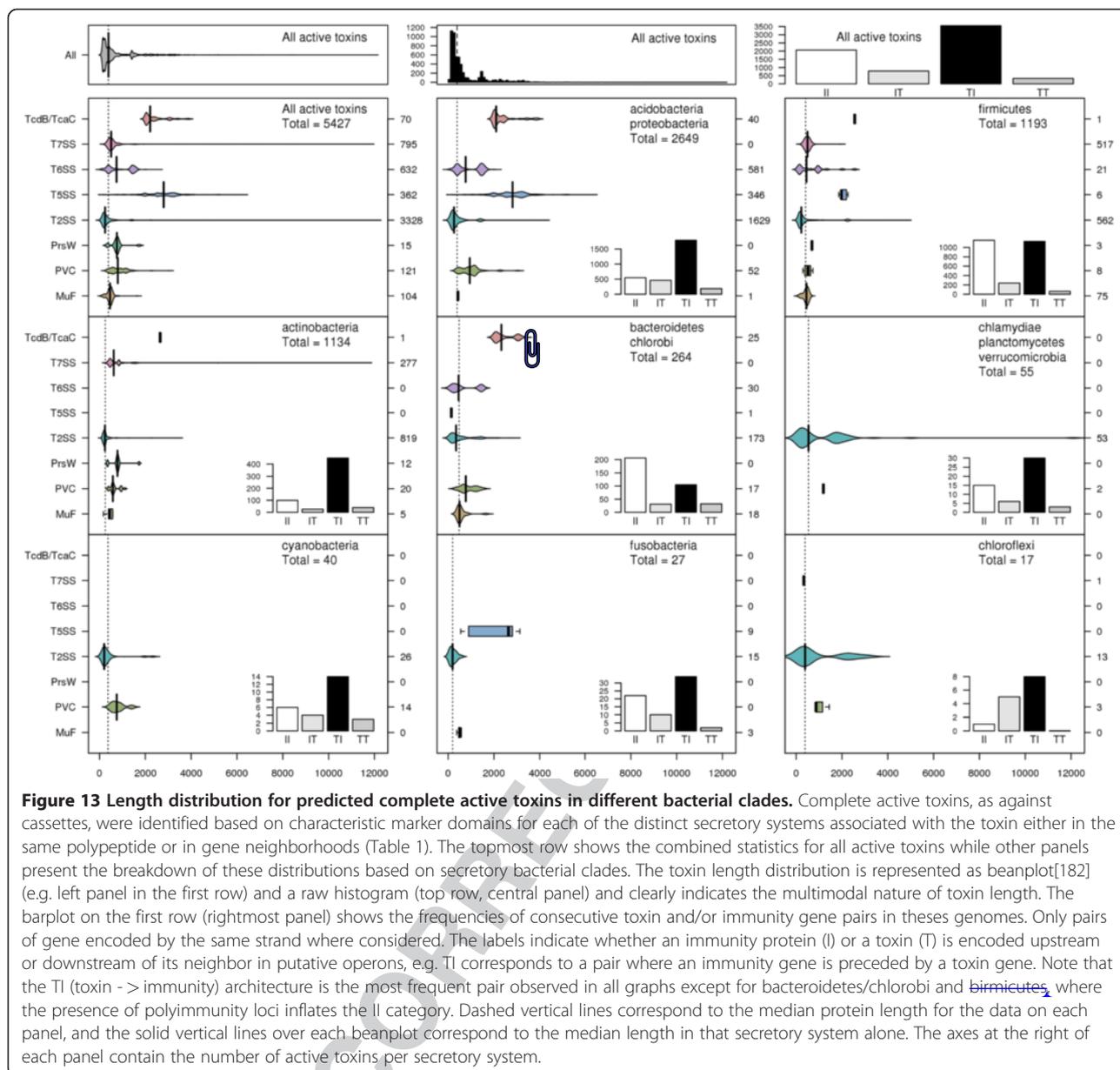


Figure 13 Length distribution for predicted complete active toxins in different bacterial clades. Complete active toxins, as against cassettes, were identified based on characteristic marker domains for each of the distinct secretory systems associated with the toxin either in the same polypeptide or in gene neighborhoods (Table 1). The topmost row shows the combined statistics for all active toxins while other panels present the breakdown of these distributions based on secretory bacterial clades. The toxin length distribution is represented as beanplot[182] (e.g. left panel in the first row) and a raw histogram (top row, central panel) and clearly indicates the multimodal nature of toxin length. The barplot on the first row (rightmost panel) shows the frequencies of consecutive toxin and/or immunity gene pairs in these genomes. Only pairs of gene encoded by the same strand where considered. The labels indicate whether an immunity protein (I) or a toxin (T) is encoded upstream or downstream of its neighbor in putative operons, e.g. TI corresponds to a pair where an immunity gene is preceded by a toxin gene. Note that the TI (toxin -> immunity) architecture is the most frequent pair observed in all graphs except for bacteroidetes/chlorobi and firmicutes, where the presence of polyimmunity loci inflates the II category. Dashed vertical lines correspond to the median protein length for the data on each panel, and the solid vertical lines over each beanplot correspond to the median length in that secretory system alone. The axes at the right of each panel contain the number of active toxins per secretory system.

2582 multiple immunity genes at the 3' end [17]: each of these
 2583 immunity genes is likely to represent a remnant of a
 2584 former recombination even that replaced the tip toxin re-
 2585 gion while inserting a new immunity gene ahead of it.
 2586 Thus, the lack of the need for a premade immunity pro-
 2587 tein due to outward trafficking of the toxin appears to
 2588 have allowed the emergence of the TI gene order. The TI
 2589 gene order in turn seems to have facilitated the emer-
 2590 gence of polymorphism in these systems. Indeed the
 2591 widely distributed simple barnase-barstar gene pairs
 2592 might represent an incipient TI gene order without not-
 2593 able polymorphism, whereas the barnase cassette within
 2594 larger polymorphic systems represents its incorporation
 2595 into the fully developed versions of these systems.

The gene-neighborhood graph also contains the im- 2596
 print of some of the secretory systems utilized for the 2597
 outward trafficking of toxins by the producing cells 2598
 (Additional File 1, Table 1)[18]. The complete toxin 2599
 genes trafficked via the T5SS, T6SS, T7SS and PVC-SS 2600
 often contain neighboring genes whose products medi- 2601
 ate their trafficking. In the case of the T5SS the adjacent 2602
 gene typically codes for CdiB-like proteins belonging to 2603
 the TpsB class of outer-membrane trafficking proteins 2604
 [37]. Such gene neighborhoods are only found in proteo- 2605
 bacteria, bacteroidetes, fusobacteria and the negativicute 2606
 clade of firmicutes (e.g. *Veillonella* and *Selenomonas*) 2607
 and are strong markers indicative of the use of the two- 2608
 partner system (T5SS) for the extrusion of toxins. The 2609

2610 phyletic pattern of this system suggests that it might
2611 have emerged in the proteobacteria-bacteroidetes assem-
2612 blage (members of the group I bacterial division [184])
2613 followed by transfer to a subset of group II lineages such
2614 as negativicutes and fusobacteria. This supports the hy-
2615 pothesis that the negativicutes have secondarily acquired
2616 a “proteobacterial”-type cell wall through lateral transfer
2617 of specific components, and not as a by-product of the
2618 sporulation system as recently proposed [185]. The
2619 T6SS, PVC-SS, and MuF-SS utilizing toxins are typically
2620 marked by the presence of genes for the injection or
2621 capsid packaging apparatus, and a recycling AAA + ATP
2622 in the case of the former two systems [38,39,75,82]. Sev-
2623 eral T6SS operons additionally encode a PspP/MOG1-
2624 like protein. The gene coding for the latter protein is
2625 often adjacent to the toxin gene and is related to the
2626 photosynthetic oxygen-evolving complex protein PspP
2627 ($p = 10^{-17}$; probability 98% in profile-profile searches)
2628 and might represent a novel subunit of the T6SS that
2629 acts as an adaptor between the secreted toxin and the
2630 injection apparatus. The genes of toxins secreted via the
2631 T7SS are occasionally characterized by gene neighbor-
2632 hoods that encode additional T7SS components such as
2633 the YueA-like FtsK/HerA ATPase (the motor driving
2634 T7SS), and EsaC, which contains a bacterial version of
2635 the PH-like fold [33,186]. Toxins associated with T7SS
2636 neighborhoods are found only in firmicutes, actinobac-
2637 teria and chloroflexi, suggesting that toxins with this
2638 secretory mode possibly emerged early in the diversifica-
2639 tion of the group II bacteria (Table 1).

2640 *Inferences from domain architectures*

2641 Comprehensive analysis of domain architectures of
2642 complete toxins reaffirms the results from the more
2643 restricted studies regarding the generally “tripartite
2644 organization” of the polymorphic toxins (Figure 1B): The
2645 N-terminal-most domains are related to trafficking of
2646 the toxin to the cell surface in the producing cell. The
2647 central domains, typically forming filamentous struc-
2648 tures, are related to presentation of the toxin on the cell
2649 surface, and processing and release for delivery into the
2650 host cell. The C-terminal-most domains are the toxin
2651 domains. This architectural blue print might be violated
2652 in certain toxins that lack the central filamentous ele-
2653 ments – these are usually shorter secreted proteins. N-
2654 terminal modules are usually associated with the
2655 secretory pathway taken by the toxin, with specific
2656 domains uniquely characterizing different secretory
2657 pathways (Table 1; Figures 12, 13): 1) The TpsA-like se-
2658 cretion domain (TPSASD) defines the T5SS [37]; 2) the
2659 PVC metallopeptidase is determinant of the PVC-SS; 3)
2660 The WXXG-like helical bundle (including LXG and
2661 LDXD) domains are strictly associated with the T7SS
2662 [187]; 4) the SpvB domain with integrin-like β -propeller

domains are the determinants of the TcdB/TcaC export 2663
pathway [42]; 5) the PrsW peptidase domain defines 2664
the eponymous export system. In the case of the T6SS, 2665
the VgrG module, which form the tip of the injection 2666
apparatus [39], might be fused in certain cases to the 2667
N-terminus of the toxin protein. Although the VgrG 2668
module might be also found in the PVC-SS gene neigh- 2669
borhoods it is never fused to toxins secreted via this 2670
pathway. Additionally, our current analysis indicated 2671
that the conserved PAAR motifs (named after the 2672
eponymous signature found in a subset of these 2673
domains; PFAM: PF05488) with an associated TM helix 2674
is found in toxins strictly associated with T6SS gene 2675
contexts. This suggests that the PAAR motif is a deter- 2676
minant for T6SS-driven export. The PAAR motifs typic- 2677
ally occur as pairs and each motif is predicted to form a 2678
3-stranded element, with the second copy usually dis- 2679
playing conserved cysteines, histidines and an aspartate 2680
that might constitute a stabilizing metal-binding site 2681
(See Additional file 1 for alignment). Given their fixed 2682
N-terminal location in the complete toxins and their 2683
specific gene-context association with components of 2684
the T6SS, it is likely that the PAAR motif represents a 2685
signal recognized by this secretory pathway. The T2SS 2686
(general secretory pathway) is the most prevalent 2687
secretory system for polymorphic toxins (Figure 12, 13). 2688
Of the dedicated secretory systems (i.e. those other than 2689
T2SS) we found that T7SS, T6SS and T5SS are the dom- 2690
inant ones, accounting for 12, 11 and 10 percent respec- 2691
tively of the complete toxins in our collection 2692
(Figure 13). The remaining dedicated secretory systems 2693
accounted for lower numbers of the total number of 2694
complete toxins. With respect to the ~150 distinct types 2695
of toxin domains we identified among polymorphic tox- 2696
ins and related systems, other than the general secretory 2697
pathway, the T7SS, T6SS and T5SS again dominate in 2698
terms of diversity of the C-terminal toxin domains with 2699
which they are associated (Figure 12). They are respect- 2700
ively being combined with 45, 43 and 43 percent of the 2701
total number of different types of toxins. Though the 2702
total number of toxin proteins delivered via the PVC-SS 2703
is much lower than that delivered by the three previ- 2704
ously named systems, it is combined with a considerable 2705
diversity of distinct types of C-terminal toxin domains 2706
(31.5% of the total number of toxin types). 2707

As discussed above, the two distinct positions of the 2708
processing peptidases, i.e., just prior to the toxin domain 2709
(e.g. HINT, papain-like peptidase, caspase) or at the N- 2710
terminus of the toxin protein (e.g. ZU5 and PrsW) ap- 2711
pear to reflect two distinct functional themes in terms 2712
of autoproteolytic cleavage of the toxin protein. The 2713
HINT peptidase is found in association with T2SS, 2714
T5SS, T7SS and the TcdB/TcaC export pathway but 2715
never with the T6SS and PVC-SS (Table 1, Figure 12). 2716

2717 This suggests that proteolytic processing by HINT and
2718 the PVC-metallopeptidase are mutually exclusive. This
2719 supports our above-stated inference that the PVC-
2720 metallopeptidase and the HINT peptidase are function-
2721 ally equivalent. It also suggests that the injection process
2722 of the T6SS probably obviates the need for autoproteo-
2723 lytic action in toxin release. Of the repeats constituting
2724 the central filamentous regions, the filamentous
2725 hemagglutinin repeats are found only in toxins delivered
2726 via the T5SS. In contrast, the RHS repeats are found in
2727 toxins delivered by all the different secretory systems,
2728 except the T5SS. The less-common, central filamentous
2729 modules, which are also promiscuous in terms of secre-
2730 tion systems, include the phage tail-fiber and the alpha-
2731 helical ALF repeats. The HINT peptidase domain is
2732 found in association with representatives of all these dif-
2733 ferent repeat types in classical polymorphic toxins sug-
2734 gesting that autoproteolytic processing to release the
2735 C-terminal toxin is a phenomenon that is independent
2736 of the type of the N-terminal stalk on which it is borne.
2737 A subset of toxin proteins from firmicutes, actinobac-
2738 teria, proteobacteria and bacteroidetes are characterized
2739 by the presence of additional adhesion-related domains in
2740 their architectures (Figure 12). Most are carbohydrate or
2741 peptidoglycan binding and include the LysM, discoidin,
2742 Laminin-G, RicinB, bulb-lectin, PGB (peptidoglycan
2743 binding), CWB (cell wall binding) and SH3 domains
2744 [188-190]. The SH3 and laminin-G domains are usually
2745 found at the N-termini of the complete toxin proteins
2746 delivered by the T2SS and are likely to help in anchor-
2747 ing the toxin to the cell wall of the producing cell by
2748 binding components of the peptidoglycan or cell-surface
2749 carbohydrates. In contrast, RicinB, discoidin and bulb
2750 lectin domains might be found either at the N-termini
2751 or embedded among the RHS repeats or close to the
2752 C-terminal toxin module. This suggests that certain
2753 versions of these domains might also be used to enhance
2754 contact with target cells. Indeed, previously the RHS
2755 repeats have also been proposed to possess carbohydrate
2756 binding ability – hence, the RHS repeats might also di-
2757 rectly participate in the adhesive action of the long toxins
2758 with such stalks [115,191]. The architecture graph also
2759 makes it clear that the nucleic acid-targeting toxins are
2760 the most prevalent type of toxin, far exceeding the
2761 peptide- and lipid- targeting toxins by a large margin
2762 (Figure 12). This is likely to be a reflection of the fact
2763 that a cell can be killed most effectively by disrupting
2764 the two key junctions in the flow of biological informa-
2765 tion, namely by disrupting the genome and by blocking
2766 translation.
2767 Examination of the length distribution of the complete
2768 toxins reveals a multimodal distribution with peaks of
2769 decreasing magnitude (Figure 13). The first peak is
2770 around 400, the second is between 1400–1600, the third

is between 2200–2400 and the fourth is between 3000–
3400 residues in length. The longest toxin recorded in
our set is SACTE_5178 (gi: 345002682), with multiple
toxin domains, from *Streptomyces* sp. SirexAA-E, and
13652 amino acids in length. This suggests that while
the complete toxins cover a wide length range there are
certain preferred lengths. In general terms it suggests
that the polymorphic toxins are of two types: 1) stalked –
those with long N-termini with multiple repetitive ele-
ments, which are likely to be used primarily in the
contact dependent mode as described for the original
CDI systems [17,36]. 2) Unstalked – these toxins lack
a substantial N-terminal extension and are like to be
secreted toxins that possibly act through diffusion into
the environment or through directed delivery into the
target cell [17]. The peaks of the distributions of the
toxins delivered via the PVC-SS, T7SS and phage MuF-
terminase system, are in the short range and these con-
tribute in a major way to the first peak in the overall
length distribution curve (Figure 13). In the case of the
T7SS, while the majority of toxins are short and likely to
be unstalked, there is a smaller set of longer stalked tox-
ins which are also delivered by this system (Figure 13).
The T6SS delivered toxins show a clear bimodal length
distribution, with a shorter variety lacking stalks or fused
to N-terminal HCP1 domains (Figure 13). This type
contributes to the first peak seen in the overall length
distribution curve. The second peak is around 1400–
1500 amino acids in length (matching the second peak
in the overall length distribution curve) and consists of
stalked toxins with RHS repeats. This suggests that the
T6SS delivers both unstalked and stalked toxins. The
former are probably directly delivered into the target cell,
whereas the latter are merely placed on the cell surface
and might act through the contact-dependent mode.
TcdB/TcaC-delivered toxins show a peak at around 2200
amino acids and contribute to the third peak observed
in the overall distribution. The T5SS-delivered toxins
show a peak a little after 3000 residues and contribute
to the 4th peak in the overall distribution (Figure 13).
The toxins with RHS repeats show a peak in their length
distribution around 1400–1600 amino acids (second peak
in the overall distribution), while for the filamentous
hemagglutinin repeats the peak length distribution is
3000–3400 amino acids (the fourth peak in the overall
distribution) (Figure 13). This indicates that the major
types of stalked toxins with different kinds of repeats,
each have their own preferred lengths. This suggests
that contact via such stalked toxins happens at a rela-
tively constant distance from the cell surface. This in
turn probably points to an optimal approach distance
between neighboring cells in colonial aggregates, such
as biofilms, where intra-specific competition would be
expected.

2825 Comparisons with other toxin systems

2826 The polymorphic toxin systems show several similarities
2827 and differences with other well-studied toxin systems of
2828 bacteria involved in different levels of intra-genomic,
2829 intra-species and inter-species conflicts. We compare
2830 below the polymorphic toxin systems with several of
2831 these systems and discuss the potential importance of
2832 significance of the similarities and differences:

2833 1) *Effectors directed at hosts and distantly related com-*
2834 *petitors*: Mechanistically the polymorphic toxins and the
2835 effectors directed against hosts and distantly related
2836 competitors are closely related. These effectors are usu-
2837 ally chromosomally encoded like classic polymorphic
2838 toxins. As seen from the above discussion (Tables 1, 2),
2839 both these systems share a large number of toxin
2840 domains, processing peptidases, and also common
2841 secretory pathways including T2SS, T5SS, T6SS, T7SS,
2842 PVC-SS and TcdB/TcaC-like export. However, the T3SS
2843 and T4SS do not appear to be used by classical poly-
2844 morphic toxins, even though they are common export
2845 pathways for effectors in specific bacterial lineages
2846 [34,192]. Some of them also have a structure closely re-
2847 sembling conventional polymorphic toxins and are only
2848 distinguished by the lack of associated genes for immun-
2849 ity proteins. Neighboring cassettes for standalone toxin
2850 domains are rare in these systems. However, the
2851 organization of other effector proteins sharing toxin
2852 domains with conventional polymorphic toxins might be
2853 different – the toxin domain is not necessarily located at
2854 the C-terminus and might occur internally or as a stan-
2855 dalone protein. Additionally, these effectors also display
2856 certain toxin domains, such as those pertaining to the
2857 eukaryotic Ub-systems that are not deployed in classical
2858 polymorphic toxin systems used in intraspecific conflict.
2859 This reflects the relative rarity or the relatively limited
2860 functional penetration of sub-cellular systems by the
2861 prokaryotic cognates of the Ub-system [126], making
2862 them less effective targets for interference.

2863 2) *Plasmid-encoded bacteriocins*: The plasmid-encoded
2864 bacteriocins, such as colicins, pyocins and cloacins con-
2865 ceptually resemble the classical polymorphic toxins in
2866 being deployed against closely related target cells. They
2867 also share the general architectural organization with
2868 classical polymorphic toxins – the N-terminal and cen-
2869 tral domains being deployed in trafficking with a toxin
2870 domain at the extreme C-terminus. Likewise, these sys-
2871 tems are also characterized by immunity proteins that
2872 help protect the producing cells [20]. Not only do their
2873 toxin domains share several mechanistic themes, such as
2874 cleaving of DNA, RNA and perforating of membranes,
2875 with the toxin domains of polymorphic toxins, but they
2876 also share certain homologous toxin domains such as
2877 the HNH, ColE3 and BECR-fold nucleases such as the
2878 colicinD and ColicinE5 domains (Table 2). However,

2879 being on plasmids their primary function is to enhance
2880 the fitness of the carrying plasmid. Hence, they usually
2881 do not have dedicated [methods](#) for their export and
2882 depend on inducing lysis of a subset of the producing cells
2883 [20].

2884 3) *Toxin-Antitoxin systems (Type I, II and III TA-*
2885 *systems)*: These systems might be encoded either on the
2886 chromosome or on a plasmid, and resemble the poly-
2887 morphic toxin systems in comprising of a pair of ele-
2888 ments with opposing activities. In the type II systems
2889 both the toxin and antitoxin are proteinaceous and
2890 interact physically with each other, thus being analogs of
2891 the polymorphic systems [22,24,28,193]. In contrast to
2892 the above described TI order of the polymorphic toxin
2893 systems with a 3' immunity gene, in TA systems the
2894 antitoxin is typically the 5' gene [22]. These elements are
2895 primarily intra-genomic selfish elements that are
2896 selected for maintaining themselves, and on occasions
2897 providing incidental advantage to the host cell [24,28].
2898 Thus, they do not have a need for any kind of export
2899 trafficking and delivery apparatus that are encountered
2900 in the other systems. As a consequence both the toxin
2901 and antitoxin from these systems are small proteins, typ-
2902 ically comprised of a single domain [22]. Nevertheless,
2903 certain toxin domains from the TA systems are homolo-
2904 gous to toxin domains of polymorphic toxins. The chief
2905 examples of these are the RNases belonging to the BECR
2906 fold (see above), the RES domain, Ntox24 and [Dcg](#)-like
2907 protein AMP/UMPylyating enzymes. However, we cur-
2908 rently do not have evidence for sharing of any of the
2909 metal-dependent nucleases between these two systems –
2910 the PIN domain nucleases are thus far only known from
2911 TA systems [108], whereas the REase, HNH and URI
2912 fold nucleases of the polymorphic toxin systems are not
2913 seen in the TA systems. On the whole, toxins of TA sys-
2914 tems tend to predominantly target the genome and the
2915 RNAs of the translation apparatus [193], but those from
2916 the polymorphic toxin systems appear to have a much
2917 wider range, though even among them there is prepon-
2918 derance of nucleic acid-targeting activities that target the
2919 above functions (Figure 12). Peptidases are relatively rare
2920 in classical TA systems in comparison to the poly-
2921 morphic toxins and their PVC-dependent relatives.
2922 However, in course of this study we uncovered a previ-
2923 ously unknown TA system, which combines a toxin pep-
2924 tidase of the YabG family with a distinctive antitoxin
2925 which was previously annotated as a “domain of un-
2926 known function” (DUF1021). This adds to the pool of
2927 toxin domains that are shared by these systems. Another
2928 enzymatic domain shared by the toxins of type II TA
2929 systems and polymorphic toxins is the ART domain
2930 [148]. Interestingly, in this case the immunity protein or
2931 the antitoxin in both these systems might be an enzyme
2932 that removes the ADP-ribose modification, such as the

2933 ADP-ribosyl glycohydrolase. The immunity proteins
2934 from the type II TA systems, in addition to physically
2935 binding their cognate toxins, also usually act as tran-
2936 scription factors that regulate the expression of the TA
2937 gene-pair via their common promoter [22]. There is cur-
2938 rently no evidence for any immunity proteins with a
2939 transcription factor function in the polymorphic toxin
2940 systems. In the case of the type I and type III TA sys-
2941 tems the antitoxin is a small RNA that respectively inter-
2942 acts with the toxin transcript or the toxin protein
2943 [24,133]. Currently, there are no known polymorphic
2944 toxin systems with RNA regulators. It appears that the
2945 need for specific physical interactions between the toxin
2946 and antitoxin in most type II and III TA systems places
2947 certain restrictions on the types of toxin domains that
2948 can be incorporated into them – they typically are small
2949 domains that are not vastly different in size from the
2950 antitoxins.

2951 **4) Restriction-Modification systems:** Like the TA sys-
2952 tems, the R-M systems are mobile, intra-genomic selfish
2953 elements that operate in prokaryotic genomes [21].
2954 Comparable to the cell-killing mediated by TA systems
2955 they have means of enforcing addiction by launching
2956 restriction attacks on cell if they are disrupted [194].
2957 They resemble both classical polymorphic toxins and
2958 TA systems in combining a toxin (the restriction en-
2959 zyme) with an antidote (the modification enzyme, typic-
2960 ally a cytosine or adenine DNA methylase). However,
2961 unlike those systems the physical interaction between
2962 the modification enzyme and the restriction enzyme is
2963 not central to the counteraction of the latter's toxic
2964 properties. Rather, since they operate on DNA, the anti-
2965 dote action of the modification enzyme is mediated by
2966 rendering the genome resistant to the restriction en-
2967 zyme by preemptively modifying it. Being purely intra-
2968 genomic selfish elements, like TA systems, but unlike
2969 polymorphic toxin systems, they do not have any fea-
2970 tures related to trafficking or delivery. Instead, R-M sys-
2971 tems display elaborate adaptations that enhance their
2972 target specificity and DNA-binding and manipulation
2973 capabilities in the form of specialized DNA-binding
2974 domains and accessory subunits such as helicases and
2975 MORC ATPases [120,195,196]. Nevertheless, as noticed
2976 above, R-M systems and polymorphic toxin systems ap-
2977 pear to share several enzymatic toxin domains such as
2978 the REase, HNH, URI and ParB domains.

2979 In conclusion, polymorphic toxin systems share certain
2980 key features with each of the other well-characterized
2981 prokaryotic toxin systems. The distinctions appear to
2982 arise from the differences in selective forces shaping each
2983 of these systems. On the whole the greatest mechanistic
2984 diversity of toxin and immunity domains are seen in the
2985 polymorphic toxin systems, which is reflective of the rela-
2986 tively few constraints faced by them in terms of their

2987 targets. However, certain types of catalytic domains are 2987
2988 preponderant across several of these systems due to dis- 2988
2989 ruption of the genome or the translation machinery being 2989
2990 apparently the easiest means of killing a cell. 2990

2991 **Genome-wide distribution of polymorphic toxin systems** 2991 2992 **and ecological implications** 2992

2993 **Differences in distributions and structure of toxins and** 2993 2994 **immunity protein: Phylogenetic and ecological tendencies** 2994

2995 To better understand the ecological significance of poly- 2995
2996 morphic toxins and related systems we systematically 2996
2997 compared their genome-wide prevalence to organismal 2997
2998 phylogeny. Our analysis revealed that all the major 2998
2999 lineages of bacteria with sufficient genomic data had at 2999
3000 least one representative coding for polymorphic toxin 3000
3001 systems. However, the distribution of these systems be- 3001
3002 tween different bacterial lineages shows pronounced dif- 3002
3003 ferences (Figures 13, 14). Among the group-I bacteria 3003 **F14**
3004 [184], polymorphic toxin systems are abundant in the 3004
3005 proteobacteria-like clade (including acidobacteria), bac- 3005
3006 teroidetes, and the clade unifying chlamydiae, verruco- 3006
3007 microbia and planctomycetes, but are relatively rare in 3007
3008 aquificae and spirochaetes. Among the group-II bacteria 3008
3009 [184], such systems are abundant in firmicutes, actino- 3009
3010 bacteria and chloroflexi but are relatively rare in cyano- 3010
3011 bacteria and thermotogae. They are generally absent in 3011
3012 most archaeal lineages, with the rare exception of certain 3012
3013 methanoarchaea and haloarchaea. Of these, *Methanosar-* 3013
3014 *cina acetivorans* displays classical stalked polymorphic 3014
3015 toxins with RHS repeats and cassettes for toxin modules 3015
3016 and immunity proteins, just as in the cognate bacterial 3016
3017 systems. A few other methanoarchaea display simple 3017
3018 barnase-barstar-like systems, whereas haloarchaea like 3018
3019 *Halogeometricum borinquense* display several PVC-SS 3019
3020 delivered toxins with variable C-terminal toxins modules 3020
3021 (Additional File 1). This general rarity of the poly- 3021
3022 morphic toxin systems is in striking contrast to the gen- 3022
3023 eral prevalence of the toxin-antitoxin systems across 3023
3024 archaea [22]. This distribution, with a dominant pres- 3024
3025 ence in most major clades of both group-I and group-II 3025
3026 bacteria, suggests that polymorphic toxin systems could 3026
3027 have been present in the ancestral bacterium. However, 3027
3028 it should be noted that these genes and cassettes are 3028
3029 highly prone to lateral transfer as suggested by the spor- 3029
3030 adic phyletic distribution of both toxin domains and im- 3030
3031 munity proteins [17]. Hence, the distribution of these 3031
3032 systems might also reflect in part the secondary disper- 3032
3033 sion of such systems across diverse bacteria by lateral 3033
3034 transfer. In support of this it may be **note** that in many 3034
3035 organisms the polymorphic toxins are situated on hyper- 3035
3036 variable chromosomal islands that are prone to lateral 3036
3037 transfer [197]. Nevertheless, distributions of the asso- 3037
3038 ciated specialized secretory systems that deliver these 3038
3039 toxins usually follow stricter phylogenetic boundaries, i.e. 3039

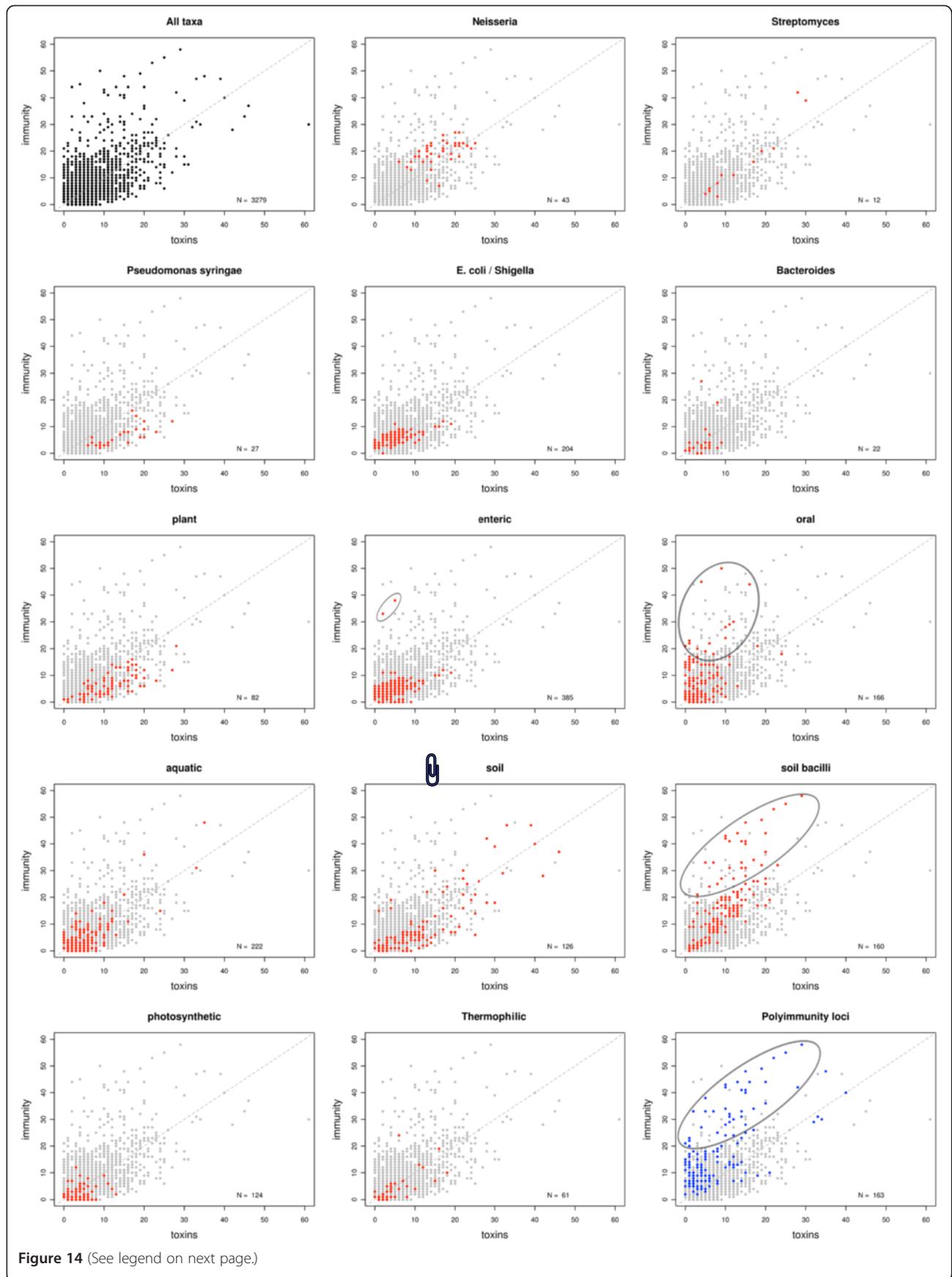


Figure 14 (See legend on next page.)

(See figure on previous page.)

Figure 14 Scatterplots of the number of toxins versus number of immunity proteins per genome. In scatter plots, black or gray dots in the background represent all taxa, and red or blue dots correspond to taxa belonging to the clade or ecological properties described on each plot's title. The dashed line corresponds to the diagonal ($x=y$) and the ellipses encircle taxa that are characterized by an excess of immunity proteins as discussed in the text.

3040 T5SS and T6SS occur primarily in group-I bacteria and
3041 T7SS in group-II bacteria. This suggests that indeed there
3042 might have been an ancestral presence of such poly-
3043 morphic toxin systems in bacteria that selected for differ-
3044 ent dedicated delivery systems in each lineage and
3045 diversified further as these delivery system were fixed.

3046 Certain patterns of distribution of polymorphic
3047 toxin systems appear to transcend phyletic boundaries
3048 (Figure 14): 1) the hyperthermophiles, which are often
3049 chemoautotrophs, from both bacteria and archaea show
3050 a strong tendency to lack such systems. 2) Likewise, the
3051 photosynthetic bacteria across different bacterial clades
3052 have a dearth of such systems (Figures 12, 14; Add-
3053 itional File 1). The relative underrepresentation of such
3054 systems in both these groups of organisms is not related
3055 to their genome sizes because organisms with similar sized
3056 genome with other lifestyles do possess such systems. In
3057 particular, the relative rarity of such systems in cyanobac-
3058 teria is striking when they are compared to other bacteria
3059 with multicellular tendencies and similar complex signal-
3060 ing mechanisms [65], such as deltaproteobacteria and acti-
3061 nobacteria, which in contrast possess abundant arrays of
3062 polymorphic toxin systems (Figures 12, 14). While in the
3063 case of archaea it is possible that the rarity of these sys-
3064 tems is related to their lack of bacterial-type protein up-
3065 take systems [20], it should be noted that bacterial
3066 hyperthermophiles show a similar pattern. The only ex-
3067 ception is the firmicute *Geobacillus thermoglucosidasius*,
3068 which, unlike the rest, is not a classical hyperthermo-
3069 phile, and can survive across a wide temperature range
3070 [198]. It appears that the relative rarity of such systems
3071 might be more related to their phototrophic or chemo-
3072 lithotrophic tendencies. It is possible that that their
3073 relative independence with respect to energy, reducing
3074 equivalents and/or carbon dioxide results in lower levels
3075 of intra-specific competition for resources.

3076 Finally, we also observed strong phylogenetic signals
3077 in the length distributions of complete toxins: 1) The
3078 group- I bacteria with Gram-negative cell walls with
3079 outer membranes (proteobacteria and bacteroidetes) had
3080 a multimodal distribution of complete toxins, showing
3081 both unstalked toxins and stalked toxins of various
3082 modal lengths (Figure 13). This suggested that they are
3083 likely to engage in both contact-dependent inhibition as
3084 well as inhibition via secreted toxins. 2) Firmicutes with
3085 the exception of the negativicute clade showed a largely
3086 unimodal distribution of complete toxin lengths with a

median value of 492 residues. This suggests that the firm- 3087
3088 icutes deploy their toxins either mainly via secretion
3089 or through much closer contact than in the previous
3090 group. 3) The actinobacteria show a bimodal distribution
3091 of toxin lengths (Figure 13). The first peak is around
3092 400–500 amino acids in length and the second is around
3093 1400–1500 amino acids. This suggests that, like proteo-
3094 bacteria, they use both distant contact and secretion or
3095 close contact. The use of both short secreted toxins and
3096 longer contact-dependent toxins suggest that intra-
3097 specific conflict might play out both in the context of
3098 biofilms, where contact is critical, and also in motile
3099 phases and swarming growth, where contact might be
3100 less intense. The distinction in this regard between firm-
3101 icutes and the two other groups raises question as to
3102 whether certain bacterial groups might resort to such
3103 forms of conflict only under specific circumstances.

Differences in the relative numbers of toxins and immunity 3104 proteins: Implications of intra- and inter-specific conflicts 3105

The median number of toxin domains found in organ- 3106
3107 isms that possess such systems is 3, which is the same as
3108 the median number of immunity proteins found per genome
3109 (Additional File 1). The difference in the number of
3110 immunity proteins and toxin domains per organism is
3111 normally distributed with a sharp peak at 0 (Additional
3112 File 1). Furthermore, there is a positive correlation be-
3113 tween the number of toxin domains and number of im-
3114 munity proteins with an approximately linear increase in
3115 the number of immunity proteins with increasing num-
3116 ber of toxin cassettes (Figure 14). These observations in-
3117 dicate that on the whole there is a balance between the
3118 number of toxin cassettes and immunity proteins, which
3119 is consistent with the genomic organization of the poly-
3120 morphic toxin loci and the principle of approximately
3121 one-to-one mapping of immunity proteins with toxins.
3122 The number of active toxins is positively correlated with
3123 the total number of toxin cassettes, suggesting that with
3124 an increase in the number of individual polymorphic
3125 toxin loci the number of toxin cassettes associated with
3126 them increase more or less linearly (Additional File 1).
3127 The median number of active cassettes per organism is 1,
3128 indicating a median 1:3 ratio between active toxins and
3129 associated toxin cassettes.

We then studied the patterns of relative numbers of 3130
3131 active toxins, cassettes and immunity proteins and their
3132 correlations, if any, with life-style and preferred

3133 ecosystems of the organisms. With exceptions discussed
3134 in the preceding subsection, bacteria across most well-
3135 sampled ecosystems display polymorphic toxin systems.
3136 However, we observed that a subset of organisms show
3137 strong anomalies in terms of the relative distribution of
3138 toxin domains to immunity proteins (Figure 14). We
3139 measured this anomaly using the difference between the
3140 number of immunity proteins and toxin domains and
3141 uncovered some striking ecological correlations. In gen-
3142 eral, in aquatic ecosystems we observed a strong propor-
3143 tionality in the number of toxins domains and immunity
3144 proteins, with roughly equal number of both (Figure 14).
3145 This suggests that in these niches there is a tendency for
3146 “honest” intra-specific conflict, with the polymorphic
3147 toxin systems primarily geared towards discrimination of
3148 non-kin conspecifics. Those organisms that showed sig-
3149 nificantly greater number of toxins than immunity pro-
3150 teins could be grouped into two general ecological
3151 niches: 1) *pathogens- Both extracellular and intracellu-*
3152 *lar pathogens of animals, plants and microbial eukar-*
3153 *yotes.* We interpret the relative abundance of toxins to
3154 immunity proteins in the former group as an adaptation
3155 for pathogenesis – the toxins are primarily used against
3156 hosts, rather than for intra-specific conflict; hence, many
3157 of their toxins do not have corresponding immunity pro-
3158 teins. This situation is especially prominent in intracellu-
3159 lar bacteria such as *Waddlia chondrophila*, *Legionella*
3160 and *Amoebophilus asiaticus*, which have a large number
3161 of toxins but hardly any immunity proteins (Additional
3162 File 1). In general, the notable absence of immunity pro-
3163 teins in intracellular pathogens suggests that in most
3164 cases (barring exceptions like *Odyssella*) they do not en-
3165 gage in competition with conspecifics in their distinctive
3166 niche. In contrast, other pathogens of animals (e.g. *Neis-*
3167 *seria* species), plants (e.g. *Ralstonia* and *Pseudomonas*
3168 *syringae*) and microbial eukaryotes (e.g. *Odyssella*), while
3169 showing a large number of toxins, also have comparable
3170 number of immunity proteins. This suggests that they
3171 are likely to compete actively with conspecific rivals in
3172 course of colonizing niches on or within their hosts. 2)
3173 *Slow growing, heterotrophic bacteria with a degree of*
3174 *“multicellular” organization, mainly actinobacteria and*
3175 *deltaproteobacteria* [65]. Organisms of this group are
3176 also well-known for their production of diverse non-
3177 proteinaceous antibiotics and maintain their slow-
3178 growing life-style by inhibiting competing faster-growing
3179 bacteria [5]. Thus, we see the over-representation of tox-
3180 ins relative to immunity proteins in this group as being
3181 part of their weaponry deployed in inter-specific compe-
3182 tition. Importantly, both these groups are also enriched
3183 in organisms coding for the greatest number of toxin
3184 domains in their genomes. The greatest number of tox-
3185 ins is seen in different *Photorhabdus* species, which are
3186 nematode symbionts that aid nematodes in killing their

insect prey [84]. Indeed, this bacterium is not only
known to kill insects with their toxins, but also com-
petes intra- and inter-specifically with other bacteria
[199]. Thus, a large number of toxins domains might be
a predictor for not just pathogen-host and inter-specific
conflict but also intense intra-specific competition in
certain niches.

On the other end of the spectrum we found several
bacteria with an overrepresentation of immunity proteins
relative to toxins. Especially striking were bacteria which
showed a marked paucity of toxins but had a large num-
ber of immunity proteins, typically occurring in polyim-
munity loci or as polyimmunity proteins. This group of
bacteria is enriched in taxa belonging to the human oral
microbiome (Figure 14; Additional File 1). Interestingly,
this phenomenon was observed across bacteria belonging
to phylogenetically distinct clades in the human oral
microbiome: this group includes representatives of bacte-
roidetes (*Capnocytophaga gingivalis*), betaproteobac-
teria (*Eikenella corrodens*), spirochetes (*Treponema*
denticola), actinobacteria (*Actinomyces* sp.) and firmi-
cutes (*Streptococcus oralis*) (Figure 14; Additional File 1).
This indicates that the oral environment has repeatedly
favored proliferation of immunity proteins relative to tox-
ins in a subset of bacteria across different clades. We in-
terpret this imbalance in terms of the ecology of
microfilms formed in the oral environment, where several
bacteria are often packed in close proximity [200]. In this
situation, non-kin “cheaters” which can invade micro-
films to benefit from cooperative associations with prox-
imal organisms can accrue an increase in fitness. Hence,
we propose that the excess of immunity proteins in these
organisms, particularly in the form of polyimmunity loci
and polyimmunity proteins, is an adaptation to evade at-
tack from a diverse array of toxins while invading non-
kin bacterial assemblages. In support of this, we observed
that there is a second group of taxa from the human oral
microbiome that display relatively balanced ratios of tox-
ins and immunity proteins (Figure 14; Additional File 1).
It is likely that these organisms are the targets for inva-
sion by the lineages with excess immunity proteins. Gen-
eralizing, this observation we propose that the presence
of a large excess of immunity proteins over toxins might
be a predictor for cheating behavior in invading non-kin
bacterial assemblages.

A distinct second group of bacteria with a large excess
of immunity protein differed from the above group in
having a median or above median number of toxins.
This group was greatly enriched in bacilli from soil such
as *Bacillus cereus*, *B. mycoides*, *B. thuringiensis*, *Breviba-*
cillus brevis and *Paenibacillus polymyxa* and representa-
tives of the human colonic microflora (Figure 14;
Additional File 1). Even in this case, the excess of im-
munity proteins were typically associated with the

3241 presence of polyimmunity loci and polyimmunity pro- 3294
3242 teins. Remarkably, we found that even within the same 3295
3243 species (e.g. *B. cereus* and *B. thuringiensis*) different 3296
3244 strains widely differed in the relative number of toxin 3297
3245 domains to immunity proteins – some isolates had a 3298
3246 considerable excess of immunity proteins, while other 3299
3247 had a balanced ratio to toxin domains and immunity 3300
3248 proteins (Figure 14; Additional File 1). This suggests that 3301
3249 the different strains in a given species adopt two general 3302
3250 strategies during intra-specific competition: 1) those 3303
3251 which participate in “honest” cooperation between kin 3304
3252 and discrimination against non-kin. These have similar 3305
3253 numbers of immunity proteins and toxins because they 3306
3254 possess only as many immunity proteins as required to 3307
3255 balance their own toxins. 2) Those which adopt the 3308
3256 strategy of cheating by invading non-kin assemblages. 3309
3257 These varieties could potentially shift to the second 3310
3258 strategy, by expressing their polyimmunity loci or pro- 3311
3259 teins, when there is an excess of “honest players”, be- 3312
3260 cause in these situations cheating might become 3313
3261 profitable. Notably, not all soil bacilli present an excess 3314
3262 of immunity proteins over toxins, e.g. *B. subtilis* does not 3315
3263 show the marked imbalance we observed in the above 3316
3264 species. This predicts that there are likely to be differ- 3317
3265 ences in the social behavior of different soil bacilli, with 3318
3266 species like *B. cereus* possibly engaging in greater degree 3319
3267 of colonial or cooperative behavior throughout their life 3320
3268 history. Further, the observation that the soil bacilli with 3321
3269 an excess of immunity proteins have multiple toxins, un- 3322
3270 like several of the above-described oral taxa which lack 3323
3271 toxins, indicates that the context in which these groups 3324
3272 might adopt a cheating strategy might differ. Among the 3325
3273 oral taxa that lack toxins, it is conceivable that they have 3326
3274 phase in their life history where they do not engage in 3327
3275 interactions with other bacteria. However, when they en- 3328
3276 counter target bacteria that can be invaded, they prob- 3329
3277 ably express their polyimmunity loci to interact with 3330
3278 them while evading their toxins. In general terms, our 3331
3279 findings might also explain how these organisms might 3332
3280 escape collapse of the cheating strategy, which would 3333
3281 happen when the numbers of cooperators are dimin- 3334
3282 ished. By facultatively expressing polyimmunity proteins 3335
3283 or loci only when target cooperators are abundant and 3336
3284 switching them off when they are absent, the deploy- 3337
3285 ment of the cheating strategy might be limited to advan- 3338
3286 tageous circumstances. 3339

3287 **Transfer of components of polymorphic toxins and** 3288 **related systems to eukaryotes and their viruses**

3289 While eukaryotes deploy a wide-range of toxins, some of 3340
3290 which share homologous domains with the polymorphic 3341
3291 toxins and related systems, most of them do not seem to 3342
3292 represent direct counterparts of the bacterial systems. 3343
3293 The eukaryotic systems that come closest to the 3344

bacterial systems described herein are the fungal killer 3294
toxins such as the *Kluyveromyces lactis* γ -toxin and PaT 3295
secreted by *Millerozyma acacia* and *Debaryomyces* 3296
robertsiae [201-203]. Like the bacterial polymorphic tox- 3297
ins, these secreted fungal toxins are primarily used in 3298
conflict with closely related non-self strains and act as 3299
endo-tRNases. However, it should be noted that they are 3300
coded by linear plasmids, which makes them similar to 3301
the classical colicin-like bacteriocins, though, unlike 3302
them, release of the fungal toxins does not entail lysis of 3303
the producing cells. These endo-tRNases currently do 3304
not have any homologs outside of fungi and were not 3305
detected in any bacterial toxin system. Nevertheless, in 3306
this study we observed that at least 13 toxin domains 3307
from polymorphic toxin systems and their relatives have 3308
been laterally transferred to fungi (Table 2). This sug- 3309
gests at least a subset of these toxin domains of bacterial 3310
provenance might also be used by fungi in intra-specific 3311
conflict in a manner comparable to the above- 3312
mentioned, fungi-specific tRNases. Our earlier study of 3313
the deaminase toxins revealed that at least a subset of 3314
these, which were acquired by fungi, are probably used 3315
in intra-specific conflict, counter-selfish element defense 3316
or in phenomena related to heteroincompatibility [18]. 3317
Indeed, a major effector in the apoptosis-like heteroin- 3318
compatibility process of several fungi, namely Het-C, 3319
appears to have originated from a bacterial toxin domain 3320
found in polymorphic toxin systems (see above). 3321

The toxin domains from the bacterial systems also ap- 3322
pear to have been acquired by animals and several other 3323
eukaryotes. At least 14 toxin domains observed in poly- 3324
morphic toxin systems are also present in metazoans, 3325
whereas at least six are present in amoeboid eukaryotes 3326
belonging to the amoebozoan and heterolobosean 3327
lineages (Table 2). Experimental evidence in animals 3328
suggests that at least a subset of these, are deployed in 3329
antiviral defense and apoptosis. The AID/APOBEC de- 3330
aminases are notable in the former context, though it 3331
appears that their role has further expanded in animals 3332
to encompass genome mutagenesis for generating anti- 3333
gen receptor diversity [204]. Like the fungal Het-C, on at 3334
least two occasions in metazoans, executors of apoptosis 3335
have emerged from toxin domains derived from poly- 3336
morphic toxin systems – the DNA-fragmenting nuclease 3337
CIDE (a HNH fold endonuclease domain) [114] and the 3338
pierisin-like ARTs which ADP-ribosylate DNA [205,206]. 3339
The phyletic patterns indicate that the lateral transfer of 3340
these two toxin domains happened at very different 3341
points in animal evolution – the CIDE-like nuclease was 3342
transferred close to the base of the metazoa, whereas the 3343
pierisin appears to have been transferred only into the 3344
lepidopteran insects. Indeed, several of the toxin 3345
domains that have been sporadically transferred to 3346
eukaryotes could have been incorporated as lineage- 3347

3348 specific components of apoptosis or antiviral defense
3349 systems. Of particular interest is the animal version of
3350 the Het-C domain which is currently known from chor-
3351 dates and the rotifer *Adineta vaga*. Like bacterial poly-
3352 morphic toxins, it occurs in a cell-surface protein, which
3353 in vertebrates is encoded by the MHC class III region
3354 [207,208]. Given this architecture it is conceivable that it
3355 is deployed as a defensive toxin against fungal or bacter-
3356 ial pathogens. However, in certain cases, such as the
3357 GHH domain, which was acquired by animals, the toxin
3358 is no longer retained in its catalytic form; instead the
3359 catalytically inactive form is used as an extracellular sig-
3360 naling molecule (i.e. Od-Oz or teneurin). As noted
3361 above, the ADP-ribosyl cyclase appears to have been
3362 acquired by both metazoa and fungi from bacterial poly-
3363 morphic toxin systems. In metazoa this enzyme was
3364 recruited as a signaling enzyme (prototyped by human
3365 CD38 and CD157), which generates two nucleotide mes-
3366 sengers cADPr and NAADP that in turn regulate the in-
3367 flux of calcium via the ryanodine receptor [162,163].
3368 Thus, the origin of multiple metazoan signaling messen-
3369 gers can be traced back to the polymorphic toxin.

3370 Of note is the observation that several toxin domains of
3371 the polymorphic toxin systems are shared with effectors
3372 delivered by endo- parasitic or symbiotic bacteria. Given
3373 the widespread presence of such resident bacteria in cells
3374 of animals, amoeboid eukaryotes and ciliates [78,79,209],
3375 it is probable that such effectors are a major source of
3376 several of the toxin domains transferred to eukaryotes
3377 and their viruses (which might share the host cell with
3378 the intracellular bacterial residents; Tables 2). Indeed the
3379 toxin-like domains of effectors and polymorphic toxins
3380 deployed by several intracellular bacteria, such as *Wolba-*
3381 *chia*, *Orientia*, *Rickettsia*, *Rickettsiella*, *Legionella*, *Odys-*
3382 *sella*, *Amoebophilus*, *Protochlamydia* and *Hamiltonella*
3383 might affect the host evolution at various levels. In a very
3384 direct sense, their action might play a major role in the
3385 manipulation of host behavior, reproduction, sex ratio
3386 and fitness (e.g. defense against parasitoid wasps in
3387 aphids by *Hamiltonella* [100,101,144]). In certain animal
3388 lineages, such as the arthropods, the pervasive presence
3389 of endosymbiotic bacteria might facilitate the routine
3390 transfer of certain toxin genes, and appears to have con-
3391 tributed to the toxins of the arthropods themselves, as
3392 suggested by the latrotoxins of spiders. The acquisition of
3393 certain toxin domains by the mimiviruses (Tox-MCF1-
3394 SHE and Ntox19), iridoviruses (Tox-Otu domain), and
3395 several NCLDVs (Tox-JAB-2) suggests that they might be
3396 used by these viruses to manipulate host behavior in a
3397 manner comparable to the intracellular bacteria. Simi-
3398 larly, several toxin domains are also encountered in bac-
3399 teriophages (Table 2), suggesting these viruses might also
3400 utilize toxin domains as a strategy to interfere with host
3401 physiology.

Certain endosymbiotic bacteria like *Odysella* also contain full-fledged polymorphic toxin systems with both toxins and immunity proteins. Such endosymbionts could possibly explain the occasional acquisition of immunity protein domains by eukaryotes and their viruses (which might share the host cell with the resident bacteria; Tables 2, 3). As previously noted, the SUKH domain proteins observed in several lineages of DNA viruses appear to have originated from immunity proteins of the polymorphic toxin systems [17]. Likewise, we had shown that the SuFu immunity protein has given rise to an intracellular component of the metazoan-specific hedgehog signaling pathway [17]. Our current analysis indicated that the C-terminal cargo-binding domain that is unique to animal type VI myosins is evolutionarily related to the immunity protein Imm-MyosinVICBD [210] ($p = 10^{-7}$ in iteration 4 with JACKHMMER in a search initiated with an immunity protein gi: 332655030) that is predicted to counter certain ADP-ribosyltransferase toxins. Given that in eukaryotes the MyosinVICBD is only found in the animal lineage and in a single association, i.e. with myosin VI, it is likely it was acquired from bacteria through transfer of a gene encoding an immunity protein. Transport of cargo by the myosin VI is unique in that it is directed toward the minus ends of the actin filaments and is required for several key cellular differentiation events in eukaryotes [210]. Other than toxin domains and immunity proteins, processing components such as the HINT peptidase domain, have been acquired by eukaryotes and incorporated into several distinct eukaryote- or even animal-specific regulatory systems such as the hedgehog pathway [17]. Another example of a processing peptidase from polymorphic toxin-like proteins, the ZU5 autopeptidase domain, might have also contributed to the evolution of the animal apoptosis system – the two ZU5 domains are observed in PIDD, the core protein of the PIDDosome, which provides a platform for recognizing molecular patterns that are associated with loss of genomic integrity and genotoxic stress [211]. We observed that related ZU5 domains are also observed in a lineage-specifically expanded group of proteins from sponges, which might have a role in defense against pathogens (Additional File 1).

On a more general note, several endosymbiotic alpha-proteobacteria such as *Wolbachia*, *Rickettsia* and *Odysella* closely resemble the progenitor of the mitochondrion [212]. Thus, such endosymbiotic associations point back to the very origin of the eukaryotes. Similarly, other endosymbiotic associations, such as those with chlamydiae might have played an important role in the origin of the photosynthetic plant lineage [213,214]. Hence, it is conceivable that the origin of some of the eukaryotic systems might be related to acquisition of genes from the toxin systems of these early

3456 bacterial symbionts. We had earlier proposed that the
3457 **key RNase component** of the eukaryotic nonsense-
3458 mediated mRNA decay system might have emerged
3459 from the prokaryotic toxin-antitoxin systems [22]. Simi-
3460 larly, the SUKH, Tad1/ADAR-like deaminase, the SuFu-
3461 associated HNH fold nuclease, ADP-ribosyltransferase
3462 and the ParBL1 domains might be early acquisitions
3463 from polymorphic or related secreted toxin systems of
3464 endosymbiotic bacteria, which were incorporated into
3465 various core function systems of eukaryotes [17,18]. In
3466 this context, it is tempting to suggest that the deubiqui-
3467 tinating peptidases such as those of the Otu clade, the
3468 Zu5 peptidase domain in the nuclear membrane protein
3469 Nup96/98, and the polyADP-ribose transferases (PARPs)
3470 might also be early acquisitions from polymorphic toxins
3471 or related effectors of the earliest endosymbionts in the
3472 associations leading to eukaryogenesis. Hence, it is con-
3473 ceivable that the very origin of certain features of the
3474 eukaryotic cell, and pan-eukaryotic regulatory systems
3475 such as ubiquitination and polyADP-ribosylation might
3476 have depended on domains derived from systems used
3477 in intra- and inter- specific conflict among prokaryotes.
3478 Thus, components derived from polymorphic toxins and
3479 related systems in symbiotic or pathogenic bacteria
3480 might have been critical for more than one major evolu-
3481 tionary transition in eukaryotes.

3482 **Conclusions**

3483 The current work is the first comprehensive analysis of
3484 the recently discovered polymorphic toxin systems. It
3485 builds upon our two earlier studies [17,18] that first
3486 uncovered these systems and revealed that their diversity
3487 was much greater than what was suspected in initial ex-
3488 perimental studies [44]. In this work we have systematic-
3489 ally identified the most prevalent toxin and immunity
3490 protein domains and have classified them based on sen-
3491 sitive sequence and structure analysis. This work thereby
3492 provides a framework for future studies on this exciting
3493 class of toxin systems. By creating an annotated inven-
3494 tory of toxins and immunity proteins it allows for fur-
3495 ther biochemical characterization of these proteins. In
3496 this regard, we offer a number of clear biochemical pre-
3497 dictions in terms of the secretory mechanisms, the mode
3498 and site of action, enzymatic activities, active sites and
3499 possible catalytic mechanisms of toxins and immunity
3500 proteins. The systematic collection of toxins also aids
3501 their investigation as potential biotechnological and
3502 therapeutic reagents – a possibility underscored by the
3503 precedent presented by several other related toxins [4,7].
3504 The pervasive relationship of toxins involved in intra-
3505 specific conflict to those used by bacteria in inter-
3506 specific conflict, such as toxins directed against hosts, is
3507 highlighted in this study. Thus, the results presented
3508 here also help in understanding the pathogenesis of

3509 numerous plant and animal pathogens, as also the inter- 3509
3510 action between unicellular eukaryotes and their abun- 3510
3511 dant intracellular bacterial residents. These findings 3511
3512 might have considerable significance for our future 3512
3513 understanding of the virulence of key pathogens, such as 3513
3514 *Pseudomonas aeruginosa*, *Legionella*, and rickettsiae 3514
3515 among other animal pathogens, and *Pseudomonas syrin-* 3515
3516 *gae*, *Xanthomonas* and *Ralstonia* among plant patho- 3516
3517 gens. The toxins characterized here also provide insights 3517
3518 regarding the biochemical basis for complex multi- 3518
3519 organism interactions, such as the role for *Hamiltonella* 3519
3520 in defense against parasitoid wasps and *Photorhabdus* in 3520
3521 nematode predation of insects[84,100,101,144,199]. 3521

3522 This study offers a platform for understanding certain 3522
3523 key ecological aspects of bacterial interactions. Systems 3523
3524 characterized here suggest, for the first time, possible mo- 3524
3525 lecular determinants for phenomena such as kin versus 3525
3526 non-kin discrimination, cooperation and cheating both in 3526
3527 the context of biofilms and motile growth. The ideas pre- 3527
3528 sented here allow for several testable microbiological hy- 3528
3529 potheses regarding bacterial conflicts. For example, the 3529
3530 proposal regarding cheating in diverse taxa from the oral 3530
3531 microbiome and certain soil bacilli can be tested via rela- 3531
3532 tively straight-forward competition experiments. Indeed, 3532
3533 such experiments can test our proposal if the polyimmu- 3533
3534 nity loci and proteins facilitate a facultative cheating strat- 3534
3535 egy in interactions between conspecifics. The systematic 3535
3536 characterization of these loci also allow for further explor- 3536
3537 ation of the rates of polymorphic transitions of toxins 3537
3538 under different conditions and in different ecosystems. 3538
3539 Some of these studies might have considerable bearing in 3539
3540 human, non-human animal and plant health, because 3540
3541 they might help explaining the preferential colonization 3541
3542 of bodily niches by certain strains as opposed to others 3542
3543 [15,199]. This might be of considerable value in facilitation 3543
3544 of processes such as wound healing and appropriate 3544
3545 re-colonization of bodily niches after antibiotic therapy. 3545

3546 The immunity proteins from these systems also offer a 3546
3547 means for understanding the two contrasting aspects of 3547
3548 the evolution of protein-protein interfaces. Our earlier 3548
3549 study had shown the versatility of the SUKH and SuFu 3549
3550 domain immunity proteins in interacting with a diverse 3550
3551 array of structurally and mechanistically distinct toxin 3551
3552 domains [17]. Thus, they join the previously studied 3552
3553 scaffolds such as the immunoglobulin domain and LRRs 3553
3554 in vertebrate antigen receptors as models to understand 3554
3555 how a single structural scaffold can diversify to accom- 3555
3556 modate an enormous variety in protein-protein interac- 3556
3557 tions [178]. On the other hand, we have also uncovered 3557
3558 numerous immunity proteins that are specific in terms 3558
3559 of the toxins they counter. Furthermore, a notable ma- 3559
3560 jority of these immunity proteins are apparently unique 3560
3561 to these systems. This presents them as models for the 3561
3562 converse aspect of the evolution of interactions, i.e. how 3562

3563 a large number of distinct domains with very specific
3564 interfaces for interaction have emerged apparently *de*
3565 *novo* in these systems. Further investigation of immunity
3566 proteins through a combination of structure determin-
3567 ation studies and biochemical analysis would be of great-
3568 est interest in regard to the evolution of these specific
3569 protein-protein interaction capabilities.

3570 Finally, the analysis of the diversification of compo-
3571 nents from polymorphic toxins and related systems
3572 points to a previously underappreciated evolutionary
3573 principle. Several toxin, immunity protein, structural
3574 modules and secretory components from these systems
3575 have a distinct life beyond their locus of provenance, es-
3576 pecially in eukaryotic regulatory and defense systems. We
3577 have documented that on numerous occasions compo-
3578 nents from these systems were incorporated into regula-
3579 tory systems of eukaryotes, and in many cases might have
3580 played a major role in the very origin of some of these
3581 systems [17,18]. Thus, these systems appear to be par-
3582 ticularly rich sources to draw from for new functional
3583 innovation. We attribute this to the consequences of nat-
3584 ural selection in systems related to inter-organismal or
3585 intra-genomic conflicts. Not surprisingly, such toxin-
3586 immunity systems have a large effect on the fitness of
3587 organisms [15,44], thereby escalating an arms race situ-
3588 ation. This has resulted in a strong selective pressure for
3589 constant diversification of polymorphic toxins and their
3590 immunity proteins. Thus, such systems have acted as a
3591 “nursery” for innovations in the protein world. Given that
3592 such conflicts often extend to the sphere of symbiotic
3593 and parasitic interactions with eukaryotes, the latter have
3594 access to a “readymade” set of molecular innovations
3595 from such systems, which can be recruited to spur the
3596 emergence of new interactions in eukaryotic systems. This
3597 is consistent with the similar diversification seen in other
3598 systems involved in intra-genomic or inter-organismal
3599 conflict [5,127,196,215,216]. These include antibiotic bio-
3600 synthesis systems which are used in inter-specific conflict,
3601 siderophore biosynthesis systems whose diversification
3602 helps prevent siderophore-stealing by “cheaters”, R-M
3603 and TA systems involved in intra-genomic conflict
3604 [5,21,194,217]. Indeed, our earlier studies indicated that
3605 components from each of these conflict systems have
3606 played a major role in contributing components to di-
3607 verse eukaryotic regulatory systems [127,196,215,216].
3608 Thus, organismal and genomic conflicts being the basis
3609 for major molecular innovations, which in turn might
3610 facilitate major evolutionary transitions, can be consid-
3611 ered a general evolutionary principle.

3612 **Methods**

3613 As described in the search strategy, protein sequences
3614 corresponding to predicted toxins, trafficking, presenta-
3615 tion, processing and immunity domains were isolated

using diagnostic domain architectures and gene- 3616
neighborhood templates, that were initially identified in 3617
previous studies [17,18] (Figure 1). The sequences of 3618
representatives of each of the domains from toxins, im- 3619
munity proteins and associated trafficking components 3620
were then used as seeds in iterative profile searches with 3621
the PSI-BLAST [218] and JACKHMMER [219] programs 3622
that run against the non-redundant (NR) protein data- 3623
base of National Center for Biotechnology Information 3624
(NCBI), to identify further homologs. A list of these 3625
search-seeds and the residue ranges for each domain is 3626
provided in Additional file 1. For most searches, which 3627
were used to report the relationships presented in this 3628
work, a cut-off e-value of .01 was used to assess signifi- 3629
cance. In each iteration the newly detected sequences 3630
that had e-values lower than the above cutoff were exam- 3631
ined for being false positives and the search was contin- 3632
ued with the same e-value threshold only if the profile 3633
was uncorrupted. The postulated relationships recover- 3634
ed using such iterative searches were further confirmed 3635
with other aids such as secondary structure prediction 3636
and superposition on known structures, if available. This 3637
resulted in the identification of over 250 toxin and im- 3638
munity domains. Search results for these domains are 3639
provided in Additional file 1. 3640

For each toxin or immunity gene, the gene neighbor- 3641
hood was also comprehensively analyzed using a custom 3642
Perl script of the inhouse TASS package. This script uses 3643
either the PTT file (downloadable from the NCBI ftp site) 3644
or the Genbank file in the case of whole genome shot 3645
gun sequences to extract the neighbors of a given query 3646
gene. Usually we used a cutoff of 5 genes on either side of 3647
the query. The protein sequences of all neighbors were 3648
clustered using the BLASTCLUST program (ftp://ftp. 3649
ncbi.nih.gov/blast/documents/blastclust.html) to identify 3650
related sequences in gene neighborhoods. Each cluster of 3651
homologous proteins were then assigned an annotation 3652
based on the domain architecture or conserved shared do- 3653
main. This allowed an initial annotation of gene neigh- 3654
borhoods and their grouping based on conservation of 3655
neighborhood associations. The remaining gene neigh- 3656
borhoods were examined for specific template patterns 3657
typical of toxin-immunity systems. In this analysis care 3658
was taken to ensure that genes are unidirectional on the 3659
same strand of DNA and shared a putative common pro- 3660
moter to be counted as a single operon. If they were head 3661
to head on opposite strands they were examined for po- 3662
tential bidirection promoter sharing patterns. 3663

Multiple sequence alignments of all domains were 3664
built by the Kalign [220], Muscle [221] and PCMA [222] 3665
programs, followed by manual adjustments on the basis 3666
of profile-profile and structural alignments. Secondary 3667
structures were predicted using the JPred [223] and 3668
PSIPred [224] programs. A comprehensive database of 3669

3670 profiles was then constructed using these multiple align-
3671 ments and was used extensively in the annotation and
3672 analysis of protein domain architectures and gene neigh-
3673 borhoods. For other known domains, the Pfam database
3674 database [189] was used as a guide, though the profiles
3675 were augmented in several cases by addition of newly
3676 detected divergent members that were not detected by
3677 the original Pfam models. Clustering with BLASTCLUST
3678 followed by multiple sequence alignment and further
3679 sequence profile searches were used to identify other
3680 domains that were not present in the Pfam database. Sig-
3681 nal peptides and transmembrane segments were detected
3682 using the TMHMM [225] and Phobius [226] programs.
3683 The HHpred program [227] was used for profile-profile
3684 comparisons to either unify poorly characterized families
3685 to proteins with a known structure in the PDB database
3686 or to group related families of toxins or immunity
3687 domains. Structure similarity searches were performed
3688 using the DaliLite program [228]. Phylogenetic ana-
3689 lysis was conducted using an approximately-maximum-
3690 likelihood method implemented in the FastTree 2.1
3691 program under default parameters [229]. Predicted lat-
3692 eral transfers to eukaryotes were further evaluated for
3693 false positives by ensuring they were embedded in contigs
3694 or complete chromosome sequences with other genes
3695 typical of eukaryotes, comparing exon-intron structure of
3696 the genes, studying their phyletic distribution within
3697 eukaryotes and comparing the protein distances of the
3698 predicted eukaryotic proteins (as measured by bit scores)
3699 with bacterial homologs. Structural visualization and
3700 manipulations were performed using the VMD [230]
3701 and PyMol (<http://www.pymol.org>) programs. Auto-
3702 matic aspects of large-scale analysis of sequences,
3703 structures and genome context were performed by
3704 using the in-house TASS package, which comprises a
3705 collection of Perl scripts. Supplementary material can
3706 also be accessed at [ftp://ftp.ncbi.nih.gov/pub/aravind/
3707 temp/TOXIMM/toxinDBsupplementary.html](ftp://ftp.ncbi.nih.gov/pub/aravind/temp/TOXIMM/toxinDBsupplementary.html)

3708 Additional files

3709 **Additional file 1: Polymorphic toxin systems: comprehensive**
3710 **characterization of trafficking modes, processing, mechanisms,**
3711 **immunity and ecology using comparative genomics.**

3712 Competing interests

3713 The authors declare that they have no competing interests.

3714 Acknowledgements

3715 The authors' research is supported by the intramural funds of the US
3716 Department of Health and Human Services (National Library of Medicine,
3717 NIH).

3718 Author details

3719 ¹National Center for Biotechnology Information, National Library of Medicine,
3720 National Institutes of Health, Bethesda, MD 20894, USA. ²Departamento de

Microbiologia, Instituto de Ciências Biomédicas, Universidade de São Paulo, 3721
São Paulo, Brazil. 3722

3723 Authors' contributions

3724 DZ, LMI and LA designed the study; DZ, LMI, VA and LA obtained the data;
3725 RFdeS wrote the custom scripts for analyzing, managing and interpreting
3726 the data; DZ, LMI, RFdeS and LA performed data analysis and interpretation;
3727 LA wrote the manuscript with inputs from DZ; LMI prepared the tables; DZ
3728 and RFdeS prepared the figures. All authors read and approved the final
3729 manuscript.

3730 Reviewers' comments

3731 **Reviewer 1: Dr. Igor Zhulin (Oak Ridge National Laboratory, USA)**

3732 I have conflicting views on this paper. On one hand, I have read
3733 Introduction, the beginning of Results & Discussion (the authors lost me half
3734 through this section though as it become very descriptive and I had a hard
3735 time connecting the pieces), and Conclusions with a great interest. The topic
3736 is fascinating and the amount of work that has been done is unbelievable.
3737 The authors analyzed an enormous amount of data, both published and
3738 results of their computational research, and presented not only a catalog of
3739 proteinaceous toxin systems, but a multi-scale picture of their roles in
3740 various biological processes. On the other hand, it all came at a high price of
3741 lacking necessary details regarding computational analyses and focus. I
3742 perfectly understand that presenting such a huge amount of information
3743 requires sacrifices in some areas, but I do not think that it should be in
3744 describing "experimental procedures". It is a generally accepted policy in
3745 science that procedures must be presented in a sufficient detail, so
3746 experiments can be independently reproduced. This paper, in my opinion,
3747 does not fulfill this requirement. The section "Search strategy to identify new
3748 toxins and immunity proteins", which serves the purpose of providing such
3749 details, gives only a very general description.

3750 **Authors' response:** *We have altered the Material and Methods to provide more*
3751 *extensive details regarding the procedures we followed with respect to sequence*
3752 *and structure analysis. We do not agree with the referee's statement that*
3753 *experimental procedures have been sacrificed. In essence all the sequence and*
3754 *structure analysis was performed using publically available programs, which*
3755 *have been published and are well-known in the computational biology*
3756 *community, if not more widely. In the current version of the Material and*
3757 *Methods we describe these without omission and any reader with access to*
3758 *appropriate computer resources can use the same. We also disagree with the*
3759 *referee's allegation of the lack of sufficient information for independent*
3760 *reproducibility – see below for further details in this regard.*

3761 Finally, the length and overall organization of this paper makes it very
3762 difficult to follow it through and the lack of page numbers is inexcusable for
3763 a manuscript that has 130 of them. Nearly each of the 38 subchapters of this
3764 paper has its own introduction and reads as a separate story. As a result, we
3765 do have an encyclopedia of polymorphic toxin systems, but its true scientific
3766 quality is hard to estimate.

3767 Personally, I would rather see much smaller pieces of this work presented in
3768 a concise way with all details of searches and analyses clearly shown. The
3769 global view that authors aimed at presenting is much better suited for
3770 review papers. Here we have a lot of original work mixed up with a review
3771 of literature: the number of references in this paper is higher than in many
3772 comprehensive reviews on similar topics. I think the quality of both original
3773 work and review suffers from this mix.

3774 The bottom line is that to me this is a paper that reaches very interesting
3775 conclusions, but which is very difficult to comprehend in its entirety and
3776 some (if not many) of its results cannot be verified (or are very difficult to
3777 verify) independently.

3778 **Authors' response:** *We regret the inconvenience caused by the lack of page*
3779 *numbers, which stems from using a PDF reader which provides the page*
3780 *numbers as against a print version. The referee raises three basic issues which*
3781 *we address below-*

3782 (i) *Length of the article – single long versus multiple short papers: Short articles*
3783 *are useful when a single domain or computational observation needs to be*
3784 *succinctly presented. Indeed, upon our initial discovery of these systems we*
3785 *published two shorter articles outlining just the details of specific aspects of*
3786 *them. However, upon further investigation it became clear that neither those*
3787 *two works nor subsequent experimental studies on these systems really do*
3788 *justice to the magnitude of domain diversity seen in these systems. Unlike many*
3789 *other systems, despite these proteins being around and accumulating in the*

3790 non-redundant protein database for now more than a decade, there has been
3791 hardly any comprehensive study on them. This is testified by the rather
3792 rudimentary annotation borne by most of them in protein databases. This
3793 being the first such treatment on a long-neglected class of highly represented
3794 proteins meant a particularly long paper. Furthermore, the practical aspects of
3795 publication meant it was quite infeasible to prepare numerous separate small
3796 papers and submit each for peer-review. We realized in course of our study that
3797 splitting the individual discoveries into multiple manuscripts would dilute the
3798 big picture emerging from these systems. With respect to shorter works being
3799 easier to read than a comprehensive manuscript as this we opine that it is
3800 largely a matter of taste. It may be noted that referee two, despite finding the
3801 length remarkable, commented regarding its easy readability. The apparent self-
3802 sufficiency of the sub-sections is primarily to help readers who might be more
3803 interested in one or few of toxin or immunity domain families but the text has
3804 been edited to minimize redundancy. Hence there is no repetition of material
3805 between sections.

3806 (ii) Review versus original paper admixture: We disagree with the referee in
3807 saying that it is a mixture of review and original research. The "review" aspect is
3808 limited to the introduction and general conclusions, as is typical of any research
3809 paper. It should be kept in mind that any kind computational analysis work
3810 based on sequence/structure analysis needs to place newly identified domains in
3811 the context of what is already known in order to make new functional
3812 predictions. This is exactly what we do – this necessitates the mention of
3813 previous studies and also precedence of biochemical activities for functional
3814 inference. We do not see this as being a mixture of review with new results but
3815 merely an aspect of building a functional argument. While there are several
3816 domains and ideas presented in this study, we were particular in only
3817 emphasizing those that are novel and discovered in this study. In our
3818 calculation, ~ 85% of our dataset (that has about 250 toxin and immunity
3819 domains) is not found in any domain database. Those that are already present
3820 in protein domain databases like PFAM, they are typically listed as domains of
3821 unknown function (DUFs) and are need of functional annotation.

3822 (iii) Reproducibility: As noted above, we do not accept the claim that our results
3823 are not reproducible. Of course, the ease of reproducibility depends entirely on
3824 the time available to one attempting it. We should emphasize that all the
3825 computational discoveries reported here use standard sequence/structure
3826 analysis techniques laid out in the Material and Methods, as is typical of a
3827 paper in this field. Those cases involving more difficult detections we explicitly
3828 mention in the paper program used and statistical support for the particular
3829 relationship or the Z score cutoffs used by DALI for structural relationships.

3830 Since we have provided Genbank identifiers (gis) for the prototypical proteins of
3831 every group, all the remaining relationships can be reproduced by running
3832 profile searches with PSI-BLAST, HMMsearch3, JACKHmmer or HHpred on the
3833 Web or locally, either in a unidirectional or transitive fashion. Most importantly
3834 we have provided one of the most extensive supplements for a sequence/
3835 structure analysis paper -- alignments for each toxin and immunity domain
3836 have been provided; hence, obtaining starting points for reproducing searches
3837 should not pose any difficulty. The gis of all proteins under consideration are
3838 also provided along with an appropriate classification. This allows for
3839 independent verification of architectures and operonic associations. In addition
3840 to the extensive tables in the body of the article which provide details regarding
3841 active sites and phyletic patterns, the data is also provided in the supplement as
3842 searchable tables, where readers can browse the data by species, domain,
3843 operons, and pathway of secretion. We fear the referee did not peruse the
3844 extensive supplement that provides all the material for reproducing the
3845 presented analysis. In the revised version we have further improved the
3846 presentation of the supplement to improve ease of access to the alignments.
3847 We will also upload all the new alignments to protein databases such as Pfam
3848 making the material available upon publication to facilitate easy reproduction
3849 and use of the presented results.

3850 **Reviewer's response to above:**

3851 I am not persuaded with authors' arguments regarding their description of
3852 "experimental procedures".

3853 Let me consider just the first paragraph of Materials and Methods, which is
3854 shown below (in italics) in its entirety and is fragmented only by my
3855 interjections.

3856 *As described in the search strategy, protein sequences corresponding to*
3857 *predicted toxins, trafficking, presentation, processing and immunity domains*
3858 *were isolated using diagnostic domain architectures and gene-neighborhood*
3859 *templates, that were initially identified in previous studies [17,18] (Figure 1).*
3860 *These domains were then used as seeds in iterative profile searches with the PSI-*

3861 BLAST [217] and JACKHMMER [218] programs that run against the non-
3862 redundant (NR) protein database of National Center for Biotechnology
3863 Information (NCBI), to identify further homologs.
3864 This is a very general statement, which provides very little detail. Clearly, each
3865 PSI-BLAST and JACKHMMER search is carried out not with "domains", but
3866 with one concrete protein sequence, which has a name and coordinates of
3867 the region that was used as a query.

3868 **Authors' response:** We concede that the word domain in this context might be
3869 confusing for some readers. However, it is should be noted that in this context
3870 we obviously imply the amino acid sequence corresponding to a given domain.
3871 This point has been emended.

3872 A search is performed against a specific database of a certain size and
3873 content. The size of NR database has doubled in less than 3 years and is
3874 changing every day. Thus, it is important either to work with a fixed version
3875 of NR or to report which version was used in a given search. Here is the
3876 excerpt from the authors' own work, which provides a good example of
3877 how "experimental procedure" should be described:
3878 "A PSI-BLAST search was initiated with the conserved N-terminal extension of
3879 the SGC (human SGC1 β , gi: 4504215, region 1–360), using an inclusion
3880 threshold of .01, and compositional bias based statistics to eliminate false
3881 positives arising due to peculiarities of sequence composition. Both the N-
3882 and the C-terminal parts of this extension gave several distinct hits to
3883 different bacterial proteins, supporting the presence of two distinct globular
3884 domains in this extension. Based on these hits we divided the extension into
3885 N- and C-terminal parts and initiated separate PSI-BLAST searches with them.
3886 Searches with the N-terminal part of the extension gave significant hits to
3887 bacterial proteins of the length 180–195 residues within the first 3 iterations
3888 (eg. Mdge1313 from Microbulbifer degradans is detected with an expect-
3889 value (e) of 10⁻⁴ in the first iteration). . . ." (LM Iyer, V Anantharaman and L
3890 Aravind 2003 BMC Genomics 2003 4:5)". Although some details are still
3891 lacking and the NR version was not specified (not that critical for the year
3892 2003), this description is thorough enough to reproduce the steps that were
3893 taken during the domain identification process. I regret that ten years later
3894 authors think that providing search details is no longer necessary. Once
3895 again, I understand the reason for not providing details for numerous
3896 searches that they have carried out, and once again I disagree with this
3897 position.

3898 **Authors' response:** We appreciate the referee quoting from a former work of
3899 ours. Obviously we have neither forgotten nor changed our philosophy to
3900 domain discovery or analysis in the past 8 years. We note that the referee states
3901 that he understands why we do not give these details in the same manner as it
3902 is done when reporting the discovery of a single/few domains. We should
3903 reiterate that when such an analysis is scaled up to hundreds of domains
3904 providing descriptions as that pasted by the referee would result in an
3905 extraordinary and tedious prolixity for most readers (users) of the article. Hence,
3906 the report in the actual manuscript focuses on the points of biochemical/
3907 biological interest with only a general description of the search strategy for
3908 most cases. This does not mean that the issues raised by the referee are
3909 inaccessible. They are simply provided in the supplementary material. Herein a
3910 reader might find a collection of the actual saved PSI-BLAST searches for all the
3911 notable domains described herein. The same files should supply the specifics of
3912 the nr database at the point of the run. Furthermore, another file in the
3913 supplement provides the query gi with sequence coordinates of all seeds used
3914 for the domain-specific searches. Yet another file provides the searches with all
3915 the profiles, which we created for this work (either PSI-BLAST or HMM) against
3916 the NR database from May 23rd 2012. The links have been made explicit in the
3917 additional file.

3918 **Referee's comment resumes:** For most searches in which were used to report
3919 the relationships presented in this work a cut-off e-value of .01 was used to
3920 assess significance.

3921 Let us leave alone the fact that something is missing from this sentence
3922 (what were used?) and focus on the main point. This statement means that
3923 for some searches a cut-off E value other than 0.01 was used.

3924 **Authors' response:** This sentence had a typo which we have now corrected and
3925 appreciate the referee pointing the same.

3926 FOR WHICH ONES? WHY? No details provided. Furthermore, 0.01 is already a
3927 "dangerous" level, when it comes to false positives. The description provided
3928 by authors leaves a possibility that some searches were carried out even
3929 with a worse E value. It does not automatically mean the results are
3930 incorrect, but it does mean that a special care must be taken when
3931 considering such relationships and description must be provided.

3932 **Authors' response:** The .01 cutoff is dangerous only in the hands of the
3933 untrained sequence analyst. Obviously we took special care to manually
3934 examine every iteration of searches with every domain reported in this study.
3935 Thus, we ensured that the new sequences being included are unlikely to be false
3936 positives.
3937 **Referee's comment resumes:** This was further confirmed with other aids such
3938 as secondary structure prediction and superposition on known structures, if
3939 available. For each toxin or immunity gene, the gene neighborhood was also
3940 comprehensively analyzed using a custom Perl script of the inhouse TASS
3941 package. The process was carried out iteratively and exhaustively and resulted in
3942 the identification of over 250 toxin and immunity domains.
3943 I am guessing that the first sentence refers to assessing the validity of
3944 multiple sequence alignments (which is described in the next paragraph).
3945 This indeed is a common technical element, which requires no further
3946 description. However, the next sentence makes quite a difference. What is
3947 meant by "comprehensive analysis of the gene neighborhood"? How many
3948 genes in the vicinity of the gene of interest were analyzed? How were they
3949 analyzed: by their RefSeq annotation? COGs? Best BLAST hit? Gene
3950 neighborhood analysis is a very important element of computational
3951 genomics of prokaryotes; however, there is no publicly available, published
3952 program or even a single, commonly accepted approach on how to do this
3953 analysis. Thus, it is important to provide details.
3954 **Authors' response:** The Material and Methods have emended to include further
3955 details on neighborhood analysis.
3956 "The process was carried out iteratively and exhaustively. . ." Which process?
3957 The entire process of domain identification or only the PSI-BLAST searches? I
3958 understand how the latter can be done iteratively and exhaustively, but I can
3959 only guess what it means with respect to the entire process, and certainly
3960 cannot distinguish between these possibilities.
3961 **Authors' response:** The Material and Methods have emended to remove the
3962 potential confusion arising from this statement.
3963 In response to my original critique authors replied that they "do not agree
3964 with the referee's statement that experimental procedures have been
3965 sacrificed. In essence all the sequence and structure analysis was performed
3966 using publicly available programs, which have been published and are
3967 well-known in the computational biology community, if not more widely". In
3968 essence, yes, but in some cases, obviously, no: a custom Perl script of the in-
3969 house package. . . Custom scripts execute specific actions. We do not need
3970 to know what the script is, but we certainly do need to know what the
3971 action was. "Comprehensive analysis of gene neighborhoods" to me is a
3972 prototype example of sacrificing the description of "experimental
3973 procedures". Even when it comes to publicly available and published tools,
3974 procedure details should be provided. In experimental biology, it is not
3975 enough to state that PCR was used to amplify a given gene – exact primers
3976 must be provided. Perhaps, this is not the best analogy, but it illustrates the
3977 point.
3978 **Authors' response:** The Material and Methods have been emended to describe
3979 the action of the script which in essence provides the details pertaining to the
3980 gene-neighborhood analysis raised above.
3981 On the final note, I would like to emphasize that I have an utmost respect
3982 for the authors, who have been leaders in the field for many years now, and
3983 who produced a series of groundbreaking papers in computational
3984 genomics. Without doubts, their results and conclusions are both correct
3985 and important. Furthermore, I applaud their decision to submit all domain
3986 models to the public repository (Pfam). However, I do disagree with their
3987 position on attention to detail in describing "experimental procedures". I can
3988 expand on this point substantially; however, this is not the place for such a
3989 debate.
3990 **Authors' response:** We too believe that this is not the place for a general
3991 debate on methodology.
3992 **Reviewer 2: Dr. Arcady Mushegian (Stowers Institute for Medical
3993 Research, USA)**
3994 The manuscript by Zhang et al. is a magisterial treatment of a large and
3995 heterogeneous group of bacterial complex toxin proteins as well as the
3996 immunity proteins that countervail the action of these toxins. It is a
3997 comprehensive collection of old and new protein families, genome contexts
3998 and phyletic distributions of these important functional modules in
3999 prokaryotes, which also crosses over to partially analyze the sequence
4000 relationships of secretion systems in bacteria. I have no concerns about the
4001 quality of sequence comparison, domain definition and genome context
4002 analysis. This is a catalog of novel predicted functions, which can guide the

work of experimentalists for years to come. I do have, however, several small
concerns about data presentation and some comments that have to do with
the broader discussion of bacterial evolution. More specifically:
Authors' response: We thank the reviewer for his positive comments and
suggestions.
p. 21–22: a few homologs of multidomain polymorphic bacterial toxins are
purported to be present in eukaryotes (e.g. gi 321474287 in *Daphnia* and
Tox-REase-8 in a subset of insects), and it is surmised that they have been
horizontally transferred from bacteria. How do we know that these genes
are indeed found in the genomes of these eukaryotes, and do not
represent endosymbiont DNA or other contamination? Have the genomic
contigs been assembled, do these genes display eukaryotic features - e.g.,
introns?
Authors' response: In our analysis, we were particularly careful in eliminating
false assignments of lateral transfer to eukaryotes and used several parameters
to decide if the laterally transferred genes were indeed encoded by the
eukaryotic species. In the simplest scenario, the presence of introns was
indicative of their eukaryotic presence. For example, the gene for gi 321474287
in *Daphnia* contains 11 introns, whereas most Tox-REase-8 genes in insects at
least contain one intron, eliminating the possibility of these genes being
contaminants. Other parameters that were considered include: 1) Elimination of
sequences that were identical or almost identical to bacterial sequences. In our
dataset, none of the proteins assigned as laterally transferred showed any
identities or near identities to bacterial sequences; 2) Most proteins assigned as
laterally transferred to eukaryotes also showed a presence in more than one
eukaryotic species, which further helps in eliminating false lateral transfer
assignments. For e.g. Tox-REase-8 is present in crustaceans, insects and
placozoans. Similarly, Tox-GHH domains are present in five major lineages of
bacteria, while in the eukaryotes they are only found in multiple metazoan
species (TCAP domains of teneurins). In response to this comment and to that
made by Reviewer 3, we have explained this procedure in more detail in the
Materials and Methods.
p. 44–45. The gene neighborhood network shown in Figure 12: I am not
sure what it is supposed to visualize. The authors state that the direction of
the edges is important, i.e., it shows the 5' to 3' order of genes or protein
domains; but the arrowheads are barely visible even in the pdf at
magnification 250%, and will not be seen online. In any case, the edge
density is so high that the main message seems to be 'anything can link to
anything'. The graphs become more sparse when clade-specific connections
are shown - this is more interesting, but perhaps visualization would be
better if the density of connections is modeled by the edges of different
thickness.
Authors' response: We agree with the reviewer that the full view of the domain
architectural network was too dense for a detailed view. We have now added a
simplified graph next to the central graph that further combines all nodes into
metanodes based on their functional type. This simplified graph gives a better
view of the follow on connectivities across all toxin polypeptides. For example, it
clearly shows that toxin domains detected in this study are almost always at
the C-terminus of the protein.
The next several comments have to do with somewhat superficial and
inconsistent discussion of relative plausibility of various evolutionary
scenarios.
To wit:
p. 46 "The phyletic pattern of this system suggests that it might have
emerged in the proteobacteria-bacteroidetes assemblage (members of the
group I bacterial division [183]) followed by transfer to a subset of group II
lineages such as negativicutes and fusobacteria." --- Why not the other
direction, or ancestral origin followed by gene losses (especially given that
these scenarios are discussed later for essentially the same phyletic vectors)?
Authors' response: The above argument is based on parsimony. In this study,
we notice a strict correlation between the occurrence of T5SS and the presence
of an outer membrane. Most lineages from Group I bacteria (including all
proteobacteria and bacteroidetes) contain an outer membrane and also
components of T5SS. In contrast, most lineages of Group II bacteria contain
only one membrane layer around the cell further encapsulated by a cell wall.
Some exceptions include the negativicutes which are a subset of firmicutes that
have an outer membrane. Since the ancestral state of the Group I and Group II
bacteria can be generally reconstructed as possessing an outer membrane in
the former and containing a single membrane layer in the latter, we propose
that the T5SS were laterally transferred to the negativicutes and fusobacteria .
We have added an additional remarks in this regard in the revised manuscript.

4074 **Referee's further response:** The explanation is fine in this case, but compare it
4075 to the following point-counterpoint.
4076 p. 52–53: "This general rarity of the polymorphic toxin systems is in striking
4077 contrast to the general prevalence of the toxin-antitoxin systems across
4078 archaea [22]. This distribution, with a dominant presence in most major
4079 clades of both group-I and group-II bacteria, suggests that polymorphic
4080 toxin systems could have been present in the ancestral bacterium." --- First,
4081 what is meant by "this distribution"? My understanding is that "this
4082 distribution" includes "general rarity" of polymorphic toxins in archaea. How
4083 can rarity of a system in archaea suggest its presence in bacterial stem, as
4084 opposed to later invention in bacteria? I suspect that this is mostly
4085 unfortunate wording that should be edited. In contrast, my second concern
4086 is more fundamental: essentially, any phyletic distribution may be interpreted
4087 as 1. ancestral presence of a gene followed by gene losses, or 2. later
4088 invention in one clade followed by horizontal transfers to the other
4089 clades; or else 3. some combination of ancestral presence, losses and HGT.
4090 To turn these scenarios from mere hand waving to something supported by
4091 the evidence, one has to specify the model of gene gain and gene loss
4092 more explicitly, or to bring in some auxiliary evidence that favors one of the
4093 explanations. I do not see much of this here.
4094 **Authors' response:** We agree that this section was a bit unclear and we have
4095 now revised it. Similar to the previous point, the polymorphic toxin systems that
4096 we report in this study are present in all major lineages of bacteria. While there
4097 is no denial that extensive lateral transfer of these systems occurs, the presence
4098 in the ancestral bacterium with divergence mirroring the evolution of different
4099 secretion systems within the bacterial superkingdom is a parsimonious
4100 argument. In contrast only a few archaeal "species" contain these systems
4101 suggesting that they were probably not present in the ancestral archaeon.
4102 Parsimoniously, this suggests that the few archaeal polymorphic toxin systems
4103 were acquired from bacterial versions, because alternatively it would require a
4104 large number of gene losses in different archaeal lineages.
4105 **Referee's further response:** In the previous exchange, the presence of a gene at
4106 the root of group I only, but not at the root of group II nor at joint root of I+II,
4107 was called "parsimonious". Now, presence at the root of all bacteria is believed
4108 to be parsimonious, when the same set of taxa is examined. What kind of
4109 parsimony is invoked in each case? (I think I can discern the answer from the
4110 next two sentences, but please correct me if I am wrong). The authors appear
4111 to understand parsimony as the explanation that requires the smaller number
4112 of events. I cannot accept this as an always-preferable explanation, when it
4113 does not matter what these events are and how are they counted; in a
4114 moderate form, however, we can use parsimony as a criterion of selecting the
4115 null hypothesis, i.e., "choose the scenario with the smallest number of events,
4116 unless the additional evidence suggests that a more complex scenario has to be
4117 considered". I think that, in this case, however, precisely such additional evidence
4118 is available in the form of evolutionary estimates of the relative rate of gene
4119 gain and gene loss: almost every estimate suggests that on average gene losses
4120 are moderately to highly more frequent than gene gains. So, unweighted
4121 parsimony will not work in these cases – a scenario with 1:1 gain-to-loss ratio
4122 will be actually making an additional assumption of a relative loss rate that is
4123 constrained to be lower than what is observed in nature. Everything is then
4124 hanging on the word "large" – how large the excess of losses in archaea is, so
4125 that this makes the scenario so unlikely?
4126 **Authors' response:** We agree that the general frequencies of gene loss tend to
4127 exceed those of gains. However, with respect to the toxin systems in archaea we
4128 are dealing with the following situation: The non-redundant database has
4129 representatives from over 225 completely sequenced WGS sequences. Classical
4130 polymorphic toxin-like systems are found only in about 15 of them. Thus, there
4131 are approximately 15 times the archaeal genomes which lack these as those
4132 which have these systems. Approximately more 1/3rd of the bacterial genomes
4133 have at least one such system. Hence, although the referee is right in pointing
4134 to the differences in the rates of loss exceeding gain, we believe our original
4135 reasoning based on the parsimony argument is a valid one.
4136 **Referee's further response:**
4137 This is also supported in phylogenetic trees, where the archaeal toxins or
4138 immunity domains group with particular bacterial versions.
4139 **Is this true for the trees of all families, or only some?**
4140 **Authors' response:** Baring the barnases where the relationship is difficult to
4141 ascertain one way or another, consistently the other toxin domains shows the
4142 archaeal branches embedded within the bacterial radiation.
4143 p. 53, the following sentence: "However, it should be noted that these genes
4144 and cassettes are highly prone to lateral transfer as suggested by the

sporadic phyletic distribution of both toxin domains and immunity proteins 4145
[17]. Hence, the distribution of these systems might also reflect in part the 4146
secondary dispersion of such systems across diverse bacteria by lateral 4147
transfer." --- Essentially, this is the same as to say that inheritance of any 4148
genetic element may be either vertical or horizontal. So? 4149
Authors' response: While the sentence might on the surface appear trivial but 4150
needs to be seen in light of the earlier comment on the polymorphic toxins 4151
being inferred present in the stem of the bacterial superkingdom. While that 4152
inference can be made based on the distribution of the toxins and their 4153
corresponding secretion systems, we intended to provide a more realistic picture 4154
(the above sentences), lest it be taken that their evolutionary history was 4155
predominantly vertical since their emergence early in bacterial evolution. 4156
Referee's further response: Once again, in the exchange regarding the 4157
statement on p. 46, the inference was that certain toxin was present in the step 4158
of proteobacteria + Bacteroidetes, but not in the stem of all bacteria. I suppose 4159
the scenarios are really different for different toxins – can this be made more 4160
explicit? 4161
Authors' response: The toxin distributions in bacteria are certainly affected by 4162
lateral transfer so we cannot be certain of the inference of particular toxin in 4163
the common ancestor. Nevertheless, based on the differential distributions, we 4164
can tentatively propose that some of the widespread versions, such as the 4165
barnase, HNH and deaminase domain toxins might have been present in the 4166
stems of the major bacterial clades such as those uniting the group-I bacteria 4167
or group-II bacteria. 4168
p. 53: "Certain patterns of distribution of polymorphic toxin systems appear 4169
to transcend phyletic boundaries. . . 1) the hyperthermophiles, which are 4170
often chemoautotrophs, from both bacteria and archaea show a strong 4171
tendency to lack such systems." --- This seems to be the case of multiple 4172
losses in bacteria, possibly favored by similarity in the habitats, and possibly 4173
ancestral absence in archaea. Ecological adaptations like this 'transcend 4174
phyletic boundaries' more or less by definition - is this the point? 4175
Authors' response: While adaptations directly related to an ecological niche are 4176
indeed obvious in terms of transcending phyletic boundaries, this is not 4177
necessarily the case with inter-organismal conflict systems, which do not directly 4178
relate to the ecological niche. Since we nevertheless found correlations between 4179
these systems and ecology, we felt it would be useful to point them out. This 4180
would help understanding the more subtle effects of ecology of a species on 4181
their interactions with conspecifics and other organisms. 4182
Referee's further response: The correlation has been observed between 4183
hyperthermophily and lack of polymorphic toxins. As the authors imply, this 4184
may in fact be the correlation between chemoautotrophy and lack of toxins – 4185
or is it? Which effects here are gross, and which are subtle? Could it be, for 4186
example, that hyperthermophily is generally correlated with reduced repertoire 4187
of all kinds of secreted proteins, which would be more easily destabilized and 4188
inactivated by adverse environment outside the cell? 4189
Authors' response: We agree that the point raised by the referee regarding 4190
temperature affecting protein stability and thereby placing a selective constraint 4191
on the number of toxins could be in principle a valid alternative explanation. 4192
However, beyond certain compositional and length distribution differences the 4193
total number of secreted and membrane proteins in hyperthermophiles do not 4194
appear to be significantly different from other organisms (e.g. Nilson et al. 4195
Proteins. 2005 Sep 1;60(4):606–16.) Hence, we are not certain if this explanation 4196
might be more relevant than autotrophy, which additionally accounts for the 4197
comparable situation in photosynthetic autotrophs. 4198
p. 56: in the case of oral microbiomes, I am not sure how some species were 4199
assigned to 'biofilm-forming' category and others to 'cheaters' - I think that 4200
at least some species in the latter category are biofilm-forming in their own 4201
right. 4202
Authors' response: As pure cultures, all these species are likely to form biofilms, 4203
but the oral environment is a mixed population of diverse bacterial species, and 4204
it is well known that oral biofilms are comprised of mixed bacterial species 4205
(Paster BJ et al. *Bacterial diversity in human subgingival plaque*, ref 198). In this 4206
context, we hypothesize that the number of toxin and immunity domains 4207
predicts how a species will interact with another one during the formation of a 4208
mixed biofilm. 4209
Reviewer 3: Dr Frank Eisenhaber (Bioinformatics Institute, Singapore) 4210
I agreed to be a reviewer when reading the author list only to find out that 4211
MS is by far the longest that I have ever seen as reviewer in my life. Despite 4212
of the initial horror and of the impressive length, the text is a fine reading - 4213
both as a research paper and as a review of this specific field. One would 4214
not think to shorten it by a page. The thoughts and results are plausible 4215

4216 (there is no hope to repeat the calculations even partially). There is
4217 considerable care for the detail throughout the text, figures and additional
4218 files (except for very minor things such as ref. 144 appearing incomplete).
4219 I find the generous addition of supplementary information especially
4220 notable.
4221 Possibly, this will be of greatest benefit for people creating annotation
4222 pipelines and sequence databases. For practical purposes, the authors might
4223 think to add archives with all the individual alignments in single files and
4224 domain models in several formats such as the HMMER2, HMMER3, etc. ready
4225 made.
4226 I think that the work is a welcome addition to the scientific literature.
4227 **Authors' response:** We thank the reviewer for his positive comments and
4228 suggestions. A more user-friendly supplementary file is now provided with the
4229 alignments of the toxins and immunity domains as separate files in a zipped
4230 format. We will additionally upload all alignments to protein domain databases
4231 such as Pfam, so that researchers can access them more easily. Ref. 144 has
4232 been updated in the revision.

4233 Received: 20 March 2012 Accepted: 25 June 2012

4234 Published: 25 June 2012

4235 References

- 4236 1. Rochat H, Martin-Eaudaire H: *Animal toxins: facts and protocols*. Basel
4237 Boston: Birkhauser Verlag; 2000.
- 4238 2. Keeler RF, Tu AT: *Toxicology of plant and fungal compounds*. New York:
4239 Dekker; 1991.
- 4240 3. Mackessy SP: *Handbook of venoms and toxins of reptiles*. Boca Raton: CRC
4241 Press; 2010.
- 4242 4. Alouf JE, Popoff MR: *The comprehensive sourcebook of bacterial protein toxins*.
4243 3rd edition. Amsterdam; Boston: Elsevier Academic Press; 2006.
- 4244 5. Walsh C: *Antibiotics: actions, origins, resistance*. Washington, D.C.: ASM Press;
4245 2003.
- 4246 6. Prof T: *Microbial toxins: molecular and cellular biology*. Norfolk, England: BIOS
4247 Scientific; 2005.
- 4248 7. Rappuoli R, Montecucco C: *Guidebook to protein toxins and their use in cell*
4249 *biology*. Oxford; New York: Oxford University Press; 1997.
- 4250 8. Dhananjaya BL: **CJ DS: An overview on nucleases (DNase, RNase, and**
4251 **phosphodiesterase) in snake venoms**. *Biochemistry (Mosc)* 2010,
4252 **75**(1):1–6.
- 4253 9. Endo Y, Tsurugi K: **Mechanism of action of ricin and related toxic lectins**
4254 **on eukaryotic ribosomes**. *Nucleic Acids Symp Ser* 1986, **17**:187–190.
- 4255 10. Chakrabarti A, Jha BK, Silverman RH: **New insights into the role of RNase L**
4256 **in innate immunity**. *J Interferon Cytokine Res* 2011, **31**(1):49–57.
- 4257 11. Wiesner J, Vilcinskis A: **Antimicrobial peptides: the ancient arm of the**
4258 **human immune system**. *Virulence* 2010, **1**(5):440–464.
- 4259 12. Li WM, Barnes T, Lee CH: **Endoribonucleases—enzymes gaining spotlight**
4260 **in mRNA metabolism**. *FEBS J* 2010, **277**(3):627–641.
- 4261 13. Rosenberg HF: **RNase A ribonucleases and host defense: an evolving**
4262 **story**. *J Leukoc Biol* 2008, **83**(5):1079–1087.
- 4263 14. Merritt EA, Hol WG: **AB5 toxins**. *Curr Opin Struct Biol* 1995, **5**(2):165–171.
- 4264 15. Russell AB, Hood RD, Bui NK, LeRoux M, Vollmer W, Mougous JD: **Type VI**
4265 **secretion delivers bacteriolytic effectors to target cells**. *Nature* 2011,
4266 **475**(7356):343–347.
- 4267 16. Aoki SK, Poole SJ, Hayes CS, Low DA: **Toxin on a stick: modular CDI toxin**
4268 **delivery systems play roles in bacterial competition**. *Virulence* 2011,
4269 **2**(4):356–359.
- 4270 17. Zhang D, Iyer LM, Aravind L: **A novel immunity system for bacterial**
4271 **nucleic acid degrading toxins and its recruitment in various eukaryotic**
4272 **and DNA viral systems**. *Nucleic Acids Res* 2011, **39**(11):4532–4552.
- 4273 18. Iyer LM, Zhang D, Rogozin IB, Aravind L: **Evolution of the deaminase fold**
4274 **and multiple origins of eukaryotic editing and mutagenic nucleic acid**
4275 **deaminases from bacterial toxin systems**. *Nucleic Acids Res* 2011,
4276 **39**(22):9473–9497.
- 4277 19. Sisto A, Cipriani MG, Morea M, Lonigro SL, Valerio F, Lavermicocca P: **An**
4278 **Rhs-like genetic element is involved in bacteriocin production by**
4279 ***Pseudomonas savastanoi* pv. *savastanoi***. *Antonie Van Leeuwenhoek* 2010,
4280 **98**(4):505–517.
- 4281 20. Cascales E, Buchanan SK, Duche D, Kleanthous C, Lloubes R, Postle K,
4282 Riley M, Slatin S, Cavard D: **Colicin biology**. *Microbiol Mol Biol Rev* 2007,
4283 **71**(1):158–229.
21. Kobayashi I: **Behavior of restriction-modification systems as selfish mobile**
4284 **elements and their impact on genome evolution**. *Nucleic Acids Res* 2001,
4285 **29**(18):3742–3756.
- 4286 22. Anantharaman V, Aravind L: **New connections in the prokaryotic**
4287 **toxin-antitoxin network: relationship with the eukaryotic**
4288 **nonsense-mediated RNA decay system**. *Genome Biol* 2003, **4**(12):R81.
- 4289 23. Engelberg-Kulka H, Glaser G: **Addiction modules and programmed cell**
4290 **death and antideath in bacterial cultures**. *Annu Rev Microbiol* 1999, **53**:43–70.
- 4291 24. Van Melderen L: **Toxin-antitoxin systems: why so many, what for?** *Curr*
4292 *Opin Microbiol* 2010, **13**(6):781–785.
- 4293 25. Aepfelbacher M, Aktories K, Just I: *Bacterial protein toxins*. Berlin; New York:
4294 Springer; 2000.
- 4295 26. Nguyen VT, Kamio Y: **Cooperative assembly of beta-barrel pore-forming**
4296 **toxins**. *J Biochem* 2004, **136**(5):563–567.
- 4297 27. Gilbert RJ: **Pore-forming toxins**. *Cell Mol Life Sci* 2002, **59**(5):832–844.
- 4298 28. Lepplae R, Geeraerts D, Hallez R, Guglielmini J, Dreze P, Van Melderen L:
4299 **Diversity of bacterial type II toxin-antitoxin systems: a comprehensive**
4300 **search and functional analysis of novel families**. *Nucleic Acids Res* 2011,
4301 **39**(13):5513–5525.
- 4302 29. MacIntyre DL, Miyata ST, Kitaoka M, Pukatzki S: **The *Vibrio cholerae* type VI**
4303 **secretion system displays antimicrobial properties**. *Proc Natl Acad Sci USA*
4304 2010, **107**(45):19520–19524.
- 4305 30. Schwarz S, West TE, Boyer F, Chiang WC, Carl MA, Hood RD, Rohmer L,
4306 Tolker-Nielsen T, Skerrett SJ, Mougous JD: **Burkholderia type VI secretion**
4307 **systems have distinct roles in eukaryotic and bacterial cell interactions**.
4308 *PLoS Pathog* 2010, **6**(8):e1001068.
- 4309 31. Linhartova I, Bumba L, Masin J, Basler M, Osicka R, Kamanova J,
4310 Prochazkova K, Adkins I, Hejnova-Holubova J, Sadilkova L, et al: **RTX**
4311 **proteins: a highly diverse family secreted by a common mechanism**.
4312 *FEMS Microbiol Rev* 2010, **34**(6):1076–1112.
- 4313 32. Holberger LE, Garza-Sanchez F, Lamoureux J, Low DA, Hayes CS: **A novel**
4314 **family of toxin/antitoxin proteins in *Bacillus* species**. *FEBS Lett* 2012,
4315 **586**(2):132–136.
- 4316 33. Iyer LM, Makarova KS, Koonin EV, Aravind L: **Comparative genomics of the**
4317 **FtsK-HerA superfamily of pumping ATPases: implications for the origins**
4318 **of chromosome segregation, cell division and viral capsid packaging**.
4319 *Nucleic Acids Res* 2004, **32**(17):5260–5279.
- 4320 34. Alvarez-Martinez CE, Christie PJ: **Biological diversity of prokaryotic type IV**
4321 **secretion systems**. *Microbiol Mol Biol Rev* 2009, **73**(4):775–808.
- 4322 35. Cornelis GR: **The type III secretion injectisome**. *Nat Rev Microbiol* 2006,
4323 **4**(11):811–825.
- 4324 36. Hayes CS, Aoki SK, Low DA: **Bacterial contact-dependent delivery systems**.
4325 *Annu Rev Genet* 2010, **44**:71–90.
- 4326 37. Delattre AS, Clantin B, Saint N, Loch C, Villeret V, Jacob-Dubuisson F:
4327 **Functional importance of a conserved sequence motif in FhaC, a**
4328 **prototypic member of the TpsB/Omp85 superfamily**. *FEBS J* 2010,
4329 **277**(22):4755–4765.
- 4330 38. Bonemann G, Pietrosiuk A, Mogk A: **Tubules and donuts: a type VI**
4331 **secretion story**. *Mol Microbiol* 2010, **76**(4):815–821.
- 4332 39. Basler M, Pilhofer M, Henderson GP, Jensen GJ, Mekalanos JJ: **Type VI**
4333 **secretion requires a dynamic contractile phage tail-like structure**. *Nature*
4334 2012, **483**(7388):182–186.
- 4335 40. Yang G, Dowling AJ, Gerike U, French-Constant RH, Waterfield NR:
4336 **Photorhabdus virulence cassettes confer injectable insecticidal activity**
4337 **against the wax moth**. *J Bacteriol* 2006, **188**(6):2254–2261.
- 4338 41. Hurst MR, Glare TR, Jackson TA: **Cloning *Serratia entomophila* antifeeding**
4339 **genes—a putative defective prophage active against the grass grub**
4340 ***Costelytra zealandica***. *J Bacteriol* 2004, **186**(15):5116–5128.
- 4341 42. Bowen D, Rocheleau TA, Blackburn M, Andreev O, Golubeva E, Bhartia R,
4342 French-Constant RH: **Insecticidal toxins from the bacterium *Photobacterium***
4343 ***luminescens***. *Science* 1998, **280**(5372):2129–2132.
- 4344 43. Ellermeier CD, Losick R: **Evidence for a novel protease governing**
4345 **regulated intramembrane proteolysis and resistance to antimicrobial**
4346 **peptides in *Bacillus subtilis***. *Genes Dev* 2006, **20**(14):1911–1922.
- 4347 44. Aoki SK, Diner EJ, de Roodenbeke CT, Burgess BR, Poole SJ, Braaten BA,
4348 Jones AM, Webb JS, Hayes CS, Cotter PA, et al: **A widespread family of**
4349 **polymorphic contact-dependent toxin delivery systems in bacteria**.
4350 *Nature* 2010, **468**(7322):439–442.
- 4351 45. Jackson AP, Thomas GH, Parkhill J, Thomson NR: **Evolutionary**
4352 **diversification of an ancient gene family (rhs) through C-terminal**
4353 **displacement**. *BMC Genomics* 2009, **10**:584.
- 4354

- 4355 46. Kung VL, Khare S, Stehlik C, Bacon EM, Hughes AJ, Hauser AR: **An rhs gene of *Pseudomonas aeruginosa* encodes a virulence protein that activates the inflammasome.** *Proc Natl Acad Sci USA* 2012, **109**(4):1275–1280. 4426
- 4356 47. Yongqiang T, Potempa J, Pike RN, Wijeyewickrema LC: **The lysine-specific gingipain of *Porphyromonas gingivalis*: importance to pathogenicity and potential strategies for inhibition.** *Adv Exp Med Biol* 2011, **712**:15–29. 4427
- 4359 48. Tonello F, Montecucco C: **The anthrax lethal factor and its MAPK kinase-specific metalloprotease activity.** *Mol Aspects Med* 2009, **30**(6):431–438. 4428
- 4361 49. Sheahan KL, Cordero CL, Satchell KJ: **Autoprocessing of the *Vibrio cholerae* RTX toxin by the cysteine protease domain.** *EMBO J* 2007, **26**(10):2552–2561. 4429
- 4362 50. Shao F, Merritt PM, Bao Z, Innes RW, Dixon JE: **A *Yersinia* effector and a *Pseudomonas* avirulence protein define a family of cysteine proteases functioning in bacterial pathogenesis.** *Cell* 2002, **109**(5):575–588. 4430
- 4363 51. Rossetto O, de Bernard M, Pellizzari R, Vitale G, Caccin P, Schiavo G, Montecucco C: **Bacterial toxins with intracellular protease activity.** *Clin Chim Acta* 2000, **291**(2):189–199. 4431
- 4364 52. Pei J, Grishin NV: **Prediction of a caspase-like fold in *Tannerella forsythia* virulence factor PrtH.** *Cell Cycle* 2009, **8**(9):1453–1455. 4432
- 4365 53. Makarova KS, Aravind L, Koonin EV: **A superfamily of archaeal, bacterial, and eukaryotic proteins homologous to animal transglutaminases.** *Protein Sci* 1999, **8**(8):1714–1719. 4433
- 4366 54. Gordon VM, Leppa SH: **Proteolytic activation of bacterial toxins: role of bacterial and host cell proteases.** *Infect Immun* 1994, **62**(2):333–340. 4434
- 4367 55. McNulty C, Thompson J, Barrett B, Lord L, Andersen C, Roberts IS: **The cell surface expression of group 2 capsular polysaccharides in *Escherichia coli*: the role of KpsD, RhsA and a multi-protein complex at the pole of the cell.** *Mol Microbiol* 2006, **59**(3):907–922. 4435
- 4368 56. Hill CW, Sandt CH, Vlazny DA: **Rhs elements of *Escherichia coli*: a family of genetic composites each encoding a large mosaic protein.** *Mol Microbiol* 1994, **12**(6):865–871. 4436
- 4369 57. Lupardus PJ, Shen A, Bogyo M, Garcia KC: **Small molecule-induced allosteric activation of the *Vibrio cholerae* RTX cysteine protease domain.** *Science* 2008, **322**(5899):265–268. 4437
- 4370 58. Tinel A, Janssens S, Lippens S, Cuenin S, Logette E, Jaccard B, Quadroni M, Tschopp J: **Autoproteolysis of PIDD marks the bifurcation between pro-death caspase-2 and pro-survival NF-kappaB pathway.** *EMBO J* 2007, **26**(1):197–208. 4438
- 4371 59. Janssens S, Tinel A: **The PIDDosome, DNA-damage-induced apoptosis and beyond.** *Cell Death Differ* 2012, **19**(1):13–20. 4439
- 4372 60. Ponting CP, Hofmann K, Bork P: **A latrophilin/CL-1-like GPs domain in polycystin-1.** *Curr Biol* 1999, **9**(16):R585–R588. 4440
- 4373 61. Mans BJ, Anantharaman V, Aravind L, Koonin EV: **Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex.** *Cell Cycle* 2004, **3**(12):1612–1637. 4441
- 4374 62. Hurst MR, Glare TR, Jackson TA, Ronson CW: **Plasmid-located pathogenicity determinants of *Serratia entomophila*, the causal agent of amber disease of grass grub, show similarity to the insecticidal toxins of *Photobacterium luminescens*.** *J Bacteriol* 2000, **182**(18):5127–5138. 4442
- 4375 63. Pei J, Mitchell DA, Dixon JE, Grishin NV: **Expansion of type II CAAX proteases reveals evolutionary origin of gamma-secretase subunit APH-1.** *J Mol Biol* 2011, **410**(11):18–26. 4443
- 4376 64. Frias M, Gonzalez C, Brito N: **BcSp1, a cerato-platanin family protein, contributes to *Botrytis cinerea* virulence and elicits the hypersensitive response in the host.** *New Phytol* 2011, **192**(2):483–495. 4444
- 4377 65. Aravind L, Iyer LM, Anantharaman V: **Natural history of sensor domains in bacterial signaling systems.** In *Sensory Mechanisms in Bacteria: Molecular Aspects of Signal Recognition*. Edited by Spiro S, Dixon R. Norfolk, UK: Caister Academic Press; 2010. 4445
- 4378 66. Aravind L, Koonin EV: **Classification of the caspase-hemoglobinase fold: detection of new families and implications for the origin of the eukaryotic separins.** *Proteins* 2002, **46**(4):355–367. 4446
- 4379 67. Barrett AJ, Rawlings ND: **Evolutionary lines of cysteine peptidases.** *Biol Chem* 2001, **382**(5):727–733. 4447
- 4380 68. Kitadokoro K, Kamitani S, Miyazawa M, Hanajima-Ozawa M, Fukui A, Miyake M, Horiguchi Y: **Crystal structures reveal a thiol protease-like catalytic triad in the C-terminal region of *Pasteurella multocida* toxin.** *Proc Natl Acad Sci USA* 2007, **104**(12):5139–5144. 4448
- 4381 69. Zhu M, Shao F, Innes RW, Dixon JE, Xu Z: **The crystal structure of *Pseudomonas* avirulence protein AvrPphB: a papain-like fold with a distinct substrate-binding site.** *Proc Natl Acad Sci USA* 2004, **101**(1):302–307. 4449
- 4382 70. Kagawa TF, Cooney JC, Baker HM, McSweeney S, Liu M, Gubba S, Musser JM, Baker EN: **Crystal structure of the zymogen form of the group A *Streptococcus* virulence factor SpeB: an integrin-binding cysteine protease.** *Proc Natl Acad Sci USA* 2000, **97**(5):2235–2240. 4450
- 4383 71. Anantharaman V, Aravind L: **Evolutionary history, structural features and biochemical diversity of the NlpC/P60 superfamily of enzymes.** *Genome Biol* 2003, **4**(2):R11. 4451
- 4384 72. Pei J, Grishin NV: **The Rho GTPase inactivation domain in *Vibrio cholerae* MARTX toxin has a circularly permuted papain-like thiol protease fold.** *Proteins: Structure, Function, and Bioinformatics* 2009, **77**(2):413–419. 4452
- 4385 73. Wood MW, Williams C, Upadhyay A, Gill AC, Philippe DL, Galyov EE, van den Elsen JM, Bagby S: **Structural analysis of *Salmonella enterica* effector protein SopD.** *Biochim Biophys Acta* 2004, **1698**(2):219–226. 4453
- 4386 74. Richards GP, Watson MA, Crane EJ III, Burt JG, Bushek D: ***Shewanella* and *Photobacterium* spp. in oysters and seawater from the Delaware Bay.** *Appl Environ Microbiol* 2008, **74**(11):3323–3327. 4454
- 4387 75. Burroughs AM, Iyer LM, Aravind L: **Comparative genomics and evolutionary trajectories of viral ATP dependent DNA-packaging systems.** *Genome Dyn* 2007, **3**:48–65. 4455
- 4388 76. Nanao MH, Tcherniuk SO, Chroboczek J, Dideberg O, Dessen A, Balakirev MY: **Crystal structure of human otubain 2.** *EMBO Rep* 2004, **5**(8):783–788. 4456
- 4389 77. Wertz IE, O'Rourke KM, Zhou H, Eby M, Aravind L, Seshagiri S, Wu P, Wiesmann C, Baker R, Boone DL, et al: **De-ubiquitination and ubiquitin ligase domains of A20 downregulate NF-kappaB signalling.** *Nature* 2004, **430**(7000):694–699. 4457
- 4390 78. Birtles RJ, Rowbotham TJ, Michel R, Pitcher DG, Lascola B, Alexiou-Daniel S, Raouf D: ***Candidatus Odysseella thessalonicensis* gen. nov., sp. nov., an obligate intracellular parasite of *Acanthamoeba* species.** *Int J Syst Evol Microbiol* 2000, **50**(Pt 1):63–72. 4458
- 4391 79. Schmitz-Esser S, Tischler P, Arnold R, Montanaro J, Wagner M, Rattei T, Horn M: **The genome of the amoeba symbiont "*Candidatus Amoebophilus asiaticus*" reveals common mechanisms for host cell interaction among amoeba-associated bacteria.** *J Bacteriol* 2010, **192**(4):1045–1057. 4459
- 4392 80. Loureiro J, Ploegh HL: **Antigen presentation and the ubiquitin-proteasome system in host-pathogen interactions.** *Adv Immunol* 2006, **92**:225–305. 4460
- 4393 81. Iyer LM, Leippe DD, Koonin EV, Aravind L: **Evolutionary history and higher order classification of AAA + ATPases.** *J Struct Biol* 2004, **146**(1–2):11–31. 4461
- 4394 82. Bonemann G, Pietrosiuk A, Diemand A, Zentgraf H, Mogk A: **Remodelling of *VipA/VipB* tubules by *ClpV*-mediated threading is crucial for type VI protein secretion.** *EMBO J* 2009, **28**(4):315–325. 4462
- 4395 83. Dhanaraj V, Ye QZ, Johnson LL, Hupe DJ, Ortwine DF, Dunbar JB Jr, Rubin JR, Pavlovsky A, Humblet C, Blundell TL: **X-ray structure of a hydroxamate inhibitor complex of stromelysin catalytic domain and its comparison with members of the zinc metalloproteinase superfamily.** *Structure* 1996, **4**(4):375–386. 4463
- 4396 84. ffrench-Constant RH, Dowling A, Waterfield NR: **Insecticidal toxins from *Photobacterium* bacteria and their potential use in agriculture.** *Toxicon* 2007, **49**(4):436–451. 4464
- 4397 85. Pechy-Tarr M, Bruck DJ, Maurhofer M, Fischer E, Vogne C, Henkels MD, Donahue KM, Grunder J, Loper JE, Keel C: **Molecular analysis of a novel gene cluster encoding an insect toxin in plant-associated strains of *Pseudomonas fluorescens*.** *Environ Microbiol* 2008, **10**(9):2368–2386. 4465
- 4398 86. Rodou A, Ankras DO, Stathopoulos C: **Toxins and Secretion Systems of *Photobacterium luminescens*.** *Toxins (Basel)* 2010, **2**(6):1250–1264. 4466
- 4399 87. Daborn PJ, Waterfield N, Silva CP, Au CP, Sharma S, ffrench-Constant RH: **A single *Photobacterium* gene, makes caterpillars floppy (mcf), allows *Escherichia coli* to persist within and kill insects.** *Proc Natl Acad Sci USA* 2002, **99**(16):10742–10747. 4467
- 4400 88. Wei CF, Kvitko BH, Shimizu R, Crabill E, Alfano JR, Lin NC, Martin GB, Huang HC, Collmer A: **A *Pseudomonas syringae* pv. tomato DC3000 mutant lacking the type III effector HopQ1–1 is able to cause disease in the model plant *Nicotiana benthamiana*.** *Plant J* 2007, **51**(1):32–46. 4468
- 4401 89. Li X, Lin H, Zhang W, Zou Y, Zhang J, Tang X, Zhou JM: **Flagellin induces innate immunity in nonhost interactions that is suppressed by *Pseudomonas syringae* effectors.** *Proc Natl Acad Sci USA* 2005, **102**(36):12990–12995. 4469

- 4497 90. Masuda M, Betancourt L, Matsuzawa T, Kashimoto T, Takao T, Shimonishi Y, Horiguchi Y: **Activation of rho through a cross-link with polyamines catalyzed by Bordetella dermonecrotizing toxin.** *EMBO J* 2000, **19**(4):521–530.
- 4501 91. Dirix G, Monsieurs P, Dombrecht B, Daniels R, Marchal K, Vanderleyden J, Michiels J: **Peptide signal molecules and bacteriocins in Gram-negative bacteria: a genome-wide in silico screening for peptides containing a double-glycine leader sequence and their cognate transporters.** *Peptides* 2004, **25**(9):1425–1440.
- 4506 92. Ishii S, Yano T, Ebihara A, Okamoto A, Manzoku M, Hayashi H: **Crystal structure of the peptidase domain of Streptococcus ComA, a bifunctional ATP-binding cassette transporter involved in the quorum-sensing pathway.** *J Biol Chem* 2010, **285**(14):10777–10785.
- 4510 93. Kelly M, Hart E, Mundy R, Marches O, Wiles S, Badea L, Luck S, Tauschek M, Frankel G, Robins-Browne RM, *et al*: **Essential role of the type III secretion system effector NleB in colonization of mice by Citrobacter rodentium.** *Infect Immun* 2006, **74**(4):2328–2337.
- 4514 94. Wong AR, Pearson JS, Bright MD, Munera D, Robinson KS, Lee SF, Frankel G, Hartland EL: **Enteropathogenic and enterohaemorrhagic Escherichia coli: even more subversive elements.** *Mol Microbiol* 2011, **80**(6):1420–1438.
- 4517 95. Iyer LM, Koonin EV, Aravind L: **Novel predicted peptidases with a potential role in the ubiquitin signaling pathway.** *Cell Cycle* 2004, **3**(11):1440–1450.
- 4520 96. Odagaki Y, Hayashi A, Okada K, Hirotsu K, Kabashima T, Ito K, Yoshimoto T, Tsuru D, Sato M, Clardy J: **The crystal structure of pyroglutamyl peptidase I from Bacillus amyloliquefaciens reveals a new structure for a cysteine protease.** *Structure* 1999, **7**(4):399–411.
- 4524 97. Takamatsu H, Imamura A, Kodama T, Asai K, Ogasawara N, Watabe K: **The yabG gene of Bacillus subtilis encodes a sporulation specific protease which is involved in the processing of several spore coat proteins.** *FEMS Microbiol Lett* 2000, **192**(1):33–38.
- 4528 98. Biarrotte-Sorin S, Hugonnet JE, Delfosse V, Mainardi JL, Gutmann L, Arthur M, Mayer C: **Crystal structure of a novel beta-lactam-insensitive peptidoglycan transpeptidase.** *J Mol Biol* 2006, **359**(3):533–538.
- 4531 99. Bielnicki J, Devedjiev Y, Derewenda U, Dauter Z, Joachimiak A, Derewenda ZS: **B. subtilis ykuD protein at 2.0 Å resolution: insights into the structure and function of a novel, ubiquitous family of bacterial enzymes.** *Proteins* 2006, **62**(1):144–151.
- 4535 100. Degnan PH, Moran NA: **Diverse phage-encoded toxins in a protective insect endosymbiont.** *Appl Environ Microbiol* 2008, **74**(21):6782–6791.
- 4537 101. Oliver KM, Degnan PH, Hunter MS, Moran NA: **Bacteriophages encode factors required for protection in a symbiotic mutualism.** *Science* 2009, **325**(5943):992–994.
- 4540 102. Aravind L, Walker DR, Koonin EV: **Conserved domains in DNA repair proteins and evolution of repair systems.** *Nucleic Acids Res* 1999, **27**(5):1223–1242.
- 4543 103. Aravind L, Makarova KS, Koonin EV: **Holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories.** *Nucleic Acids Res* 2000, **28**(18):3417–3432.
- 4546 104. Mak AN, Lambert AR, Stoddard BL: **Folding, DNA recognition, and function of GIY-YIG endonucleases: crystal structures of R. Eco29kl.** *Structure* 2010, **18**(10):1321–1331.
- 4549 105. Zhao L, Bonocora RP, Shub DA, Stoddard BL: **The restriction fold turns to the dark side: a bacterial homing endonuclease with a PD-(D/E)-XK motif.** *EMBO J* 2007, **26**(9):2432–2442.
- 4552 106. Stoddard BL: **Homing endonuclease structure and function.** *Q Rev Biophys* 2005, **38**(1):49–95.
- 4554 107. Yang W: **Nucleases: diversity of structure, function and mechanism.** *Q Rev Biophys* 2011, **44**(1):1–93.
- 4556 108. Anantharaman V, Aravind L: **The NYN domains: novel predicted RNAses with a PIN domain-like fold.** *RNA Biol* 2006, **3**(1):18–27.
- 4558 109. Carr S, Walker D, James R, Kleanthous C, Hemmings AM: **Inhibition of a ribosome-inactivating ribonuclease: the crystal structure of the cytotoxic domain of colicin E3 in complex with its immunity protein.** *Structure* 2000, **8**(9):949–960.
- 4562 110. Graille M, Mora L, Buckingham RH, van Tilbeurgh H, de Zamaroczy M: **Structural inhibition of the colicin D tRNase by the tRNA-mimicking immunity protein.** *EMBO J* 2004, **23**(7):1474–1482.
- 4565 111. Ghosh M, Meiss G, Pingoud A, London RE, Pedersen LC: **Structural insights into the mechanism of nuclease A, a betabeta alpha metal nuclease from Anabaena.** *J Biol Chem* 2005, **280**(30):27990–27997.
112. Guthrie EP, Quinton-Jager T, Moran LS, Slatko BE, Kucera RB, Benner JS, Wilson GG, Brooks JE: **Cloning, expression and sequence analysis of the SphI restriction-modification system.** *Gene* 1996, **180**(1–2):107–112.
113. Woo EJ, Kim YG, Kim MS, Han WD, Shin S, Robinson H, Park SY, Oh BH: **Structural mechanism for inactivation and activation of CAD/DFP40 in the apoptotic pathway.** *Mol Cell* 2004, **14**(4):531–539.
114. Lugovskoy AA, Zhou P, Chou JJ, McCarty JS, Li P, Wagner G: **Solution structure of the CIDE-N domain of CIDE-B and a model for CIDE-N/CIDE-N interactions in the DNA fragmentation pathway of apoptosis.** *Cell* 1999, **99**(7):747–755.
115. Minet AD, Rubin BP, Tucker RP, Baumgartner S, Chiquet-Ehrismann R: **Teneurin-1, a vertebrate homologue of the Drosophila pair-rule gene ten-m, is a neuronal protein with a novel type of heparin-binding domain.** *J Cell Sci* 1999, **112**(Pt 12):2019–2032.
116. Silva JP, Lelianaova VG, Ermolyuk YS, Vysokov N, Hitchen PG, Berninghausen O, Rahman MA, Zangrandi A, Fidalgo S, Tonevitsky AG, *et al*: **Latrophilin 1 and its endogenous ligand Lasso/teneurin-2 form a high-affinity transsynaptic receptor pair with signaling capabilities.** *Proc Natl Acad Sci USA* 2011, **108**(29):12113–12118.
117. Topf U, Chiquet-Ehrismann R: **Genetic interaction between Caenorhabditis elegans teneurin ten-1 and prolyl 4-hydroxylase phy-1 and their function in collagen IV-mediated basement membrane integrity during late elongation of the embryo.** *Mol Biol Cell* 2011, **22**(18):3331–3343.
118. Qian X, Barysyt-Lovejoy D, Wang L, Chewpoy B, Gautam N, Al Chawaf A, Lovejoy DA: **Cloning and characterization of teneurin C-terminus associated peptide (TCAP)-3 from the hypothalamus of an adult rainbow trout (Oncorhynchus mykiss).** *Gen Comp Endocrinol* 2004, **137**(2):205–216.
119. Aravind L, Koonin EV: **Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system.** *Genome Res* 2001, **11**(8):1365–1374.
120. Aravind L, Iyer LM: **The HARE-HTH and associated domains: novel modules in the coordination of epigenetic DNA and protein modifications.** *Cell Cycle* 2012, **11**(1):119–131.
121. Vosman B, Kuiken G, Kooistra J, Venema G: **Transformation in Bacillus subtilis: involvement of the 17-kilodalton DNA-entry nuclease and the competence-specific 18-kilodalton protein.** *J Bacteriol* 1988, **170**(8):3703–3710.
122. Johnson EP, Mincer T, Schwab H, Burgin AB, Helinski DR: **Plasmid RK2 ParB protein: purification and nuclease properties.** *J Bacteriol* 1999, **181**(19):6010–6018.
123. Jonsson TJ, Murray MS, Johnson LC, Poole LB, Lowther WT: **Structural basis for the retroreduction of inactivated peroxiredoxins by human sulfiredoxin.** *Biochemistry* 2005, **44**(24):8634–8642.
124. Chen S, Wang L, Deng Z: **Twenty years hunting for sulfur in DNA.** *Protein Cell* 2010, **1**(1):14–21.
125. Iyer LM, Tahiliani M, Rao A, Aravind L: **Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids.** *Cell Cycle* 2009, **8**(11):1698–1710.
126. Burroughs AM, Iyer LM, Aravind L: **Functional diversification of the RING finger and other binuclear treble clef domains in prokaryotes and the early evolution of the ubiquitin system.** *Mol Biosyst* 2011, **7**(7):2261–2277.
127. Iyer LM, Burroughs AM, Aravind L: **The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like beta-grasp domains.** *Genome Biol* 2006, **7**(7):R60.
128. Burns KE, Baumgart S, Dorrestein PC, Zhai H, McLafferty FW, Begley TP: **Reconstitution of a new cysteine biosynthetic pathway in Mycobacterium tuberculosis.** *J Am Chem Soc* 2005, **127**(33):11602–11603.
129. Sarkar S, Iyer G, Wu J, Glass NL: **Nonself recognition is mediated by HET-C heterocomplex formation during vegetative incompatibility.** *EMBO J* 2002, **21**(18):4841–4850.
130. Wichmann G, Sun J, Dementhon K, Glass NL, Lindow SE: **A novel gene, phcA from Pseudomonas syringae induces programmed cell death in the filamentous fungus Neurospora crassa.** *Mol Microbiol* 2008, **68**(3):672–689.
131. Hough E, Hansen LK, Birknes B, Jynge K, Hansen S, Hordvik A, Little C, Dodson E, Derewenda Z: **High-resolution (1.5 Å) crystal structure of phospholipase C from Bacillus cereus.** *Nature* 1989, **338**(6213):357–360.
132. Romier C, Dominguez R, Lahm A, Dahl O, Suck D: **Recognition of single-stranded DNA by nuclease P1: high resolution crystal structures of complexes with substrate analogs.** *Proteins* 1998, **32**(4):414–424.

- 4639 133. Kawano M, Aravind L, Storz G: **An antisense RNA controls synthesis of an SOS-induced toxin evolved from an antitoxin.** *Mol Microbiol* 2007, **64**(3):738–754.
- 4642 134. SCOP: *Structural Classification of Proteins*. <http://scop.mrc-lmb.cam.ac.uk/scop/index.html>.
- 4644 135. Gioia U, Laneve P, Dlakic M, Arcenci M, Bozzoni I, Caffarelli E: **Functional characterization of XendoU, the endoribonuclease involved in small nucleolar RNA biosynthesis.** *J Biol Chem* 2005, **280**(19):18996–19002.
- 4647 136. Raines RT: **Ribonuclease A.** *Chem Rev* 1998, **98**(3):1045–1066.
- 4648 137. Ng CL, Lang K, Meenan NA, Sharma A, Kelley AC, Kleanthous C, Ramakrishnan V: **Structural basis for 16 S ribosomal RNA cleavage by the cytotoxic domain of colicin E3.** *Nat Struct Mol Biol* 2010, **17**(10):1241–1246.
- 4652 138. Duron O: **Insights beyond Wolbachia-Drosophila interactions: never completely trust a model: insights from cytoplasmic incompatibility beyond Wolbachia-Drosophila interactions.** *Heredity (Edinb)* 2008, **101**(6):473–474.
- 4656 139. Yarbrough ML, Li Y, Kinch LN, Grishin NV, Ball HL, Orth K: **AMPylation of Rho GTPases by Vibrio VopS disrupts effector binding and downstream signaling.** *Science* 2009, **323**(5911):269–272.
- 4659 140. Feng F, Yang F, Rong W, Wu X, Zhang J, Chen S, He C, Zhou JM: **A Xanthomonas uridine 5'-monophosphate transferase inhibits plant immune kinases.** *Nature* 2012, **485**(7396):114–118.
- 4662 141. Goto Y, Li B, Claesen J, Shi Y, Bibb MJ, van der Donk WA: **Discovery of unique lanthionine synthetases reveals new mechanistic and evolutionary insights.** *PLoS Biol* 2010, **8**(3):e1000339.
- 4665 142. You YO, Levensgood MR, Ihnken LA, Knowlton AK, van der Donk WA: **Lactacin 481 synthetase as a general serine/threonine kinase.** *ACS Chem Biol* 2009, **4**(5):379–385.
- 4668 143. Reinert DJ, Jank T, Aktories K, Schulz GE: **Structural basis for the function of Clostridium difficile toxin B.** *J Mol Biol* 2005, **351**(5):973–981.
- 4670 144. Degnan PH, Yu Y, Sisneros N, Wing RA, Moran NA: **Hamiltonella defensa, genome evolution of protective bacterial endosymbiont from pathogenic ancestors.** *Proc Natl Acad Sci USA* 2009, **106**(22):9063–9068.
- 4673 145. Fieldhouse RJ, Turgeon Z, White D, Merrill AR: **Cholera- and anthrax-like toxins are among several new ADP-ribosyltransferases.** *PLoS Comput Biol* 2010, **6**(12):e1001029.
- 4676 146. Otto H, Reche PA, Bazan F, Dittmar K, Haag F, Koch-Nolte F: **In silico characterization of the family of PARP-like poly(ADP-ribosyl)transferases (pARTs).** *BMC Genomics* 2005, **6**:139.
- 4679 147. Bazan JF, Koch-Nolte F: **Sequence and structural links between distant ADP-ribosyltransferase families.** *Adv Exp Med Biol* 1997, **419**:99–107.
- 4681 148. de Souza RF, Aravind L: **Identification of novel components of NAD-utilizing metabolic pathways and prediction of their biochemical functions.** *Mol Biosyst* 2012, **8**(6):1661–1677.
- 4684 149. Jorgensen R, Purdy AE, Fieldhouse RJ, Kimber MS, Bartlett DH, Merrill AR: **Cholix toxin, a novel ADP-ribosylating factor from Vibrio cholerae.** *J Biol Chem* 2008, **283**(16):10671–10678.
- 4687 150. Yates SP, Jorgensen R, Andersen GR, Merrill AR: **Stealth and mimicry by deadly bacterial toxins.** *Trends Biochem Sci* 2006, **31**(2):123–133.
- 4689 151. Reinert DJ, Carpusca I, Aktories K, Schulz GE: **Structure of the mosquitoicidal toxin from Bacillus sphaericus.** *J Mol Biol* 2006, **357**(4):1226–1236.
- 4692 152. Hayashi S, Ishii T, Matsunaga T, Tominaga R, Kuromori T, Wada T, Shinozaki K, Hirayama T: **The glycerophosphoryl diester phosphodiesterase-like proteins SHV3 and its homologs play important roles in cell wall organization.** *Plant Cell Physiol* 2008, **49**(10):1522–1535.
- 4696 153. Kang TS, Georgieva D, Genov N, Murakami MT, Sinha M, Kumar RP, Kaur P, Kumar S, Dey S, Sharma S, et al: **Enzymatic toxins from snake venom: structural characterization and mechanism of catalysis.** *FEBS J* 2011, **278**(23):4544–4576.
- 4700 154. Sandoval-Calderon M, Geiger O, Guan Z, Barona-Gomez F, Sohlenkamp C: **A eukaryote-like cardiolipin synthase is present in Streptomyces coelicolor and in most actinobacteria.** *J Biol Chem* 2009, **284**(26):17383–17390.
- 4703 155. Dowhan W: **Molecular basis for membrane phospholipid diversity: why are there so many lipids?.** *Annu Rev Biochem* 1997, **66**:199–232.
- 4705 156. Nambu T, Minamino T, Macnab RM, Kutsukake K: **Peptidoglycan-hydrolyzing activity of the FlgJ protein, essential for flagellar rod formation in Salmonella typhimurium.** *J Bacteriol* 1999, **181**(5):1555–1561.
- 4708 157. Henrissat B, Callebaut I, Fabrega S, Lehn P, Mornon JP, Davies G: **Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases.** *Proc Natl Acad Sci USA* 1995, **92**(15):7090–7094.
- 4711 158. Copley RR, Bork P: **Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways.** *J Mol Biol* 2000, **303**(4):627–641.
- 4714 159. Aravind L, Koonin EV: **DNA polymerase beta-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history.** *Nucleic Acids Res* 1999, **27**(7):1609–1618.
- 4717 160. Potrykus K, Cashel M: **(p)ppGpp still magical?.** *Annu Rev Microbiol* 2008, **62**:35–51.
- 4719 161. Liu Q, Kriksunov IA, Graeff R, Munshi C, Lee HC, Hao Q: **Crystal structure of human CD38 extracellular domain.** *Structure* 2005, **13**(9):1331–1339.
- 4721 162. Guse AH, Lee HC: **NAADP: a universal Ca²⁺ trigger.** *Sci Signal* 2008, **1**(44):re10.
- 4722 163. Chini EN: **CD38 as a regulator of cellular NAD: a novel potential pharmacological target for metabolic conditions.** *Curr Pharm Des* 2009, **15**(1):57–63.
- 4724 164. Iacovache I, van der Goot FG, Pernot L: **Pore formation: an ancient yet complex form of attack.** *Biochim Biophys Acta* 2008, **1778**(7–8):1611–1623.
- 4727 165. Gonzalez MR, Bischofberger M, Pernot L, van der Goot FG, Freche B: **Bacterial pore-forming toxins: the (w)hole story?.** *Cell Mol Life Sci* 2008, **65**(3):493–507.
- 4729 166. Rescher U, Gerke V: **Annexins—unique membrane binding proteins with diverse functions.** *J Cell Sci* 2004, **117**(Pt 13):2631–2639.
- 4731 167. Rohou A, Nield J, Ushkaryov YA: **Insecticidal toxins from black widow spider venom.** *Toxicon* 2007, **49**(4):531–549.
- 4733 168. Dulubova IE, Krasnoperov VG, Khvotchev MV, Pluzhnikov KA, Volkova TM, Grishin EV, Vais H, Bell DR, Usherwood PN: **Cloning and structure of delta-latroinsectotoxin, a novel insect-specific member of the latrotoxin family: functional expression requires C-terminal truncation.** *J Biol Chem* 1996, **271**(13):7535–7543.
- 4736 169. King JG, Vernick KD, Hillyer JF: **Members of the salivary gland surface protein (SGS) family are major immunogenic components of mosquito saliva.** *J Biol Chem* 2011, **286**(47):40824–40834.
- 4737 170. Klasson L, Kambris Z, Cook PE, Walker T, Sinkins SP: **Horizontal gene transfer between Wolbachia and the mosquito Aedes aegypti.** *BMC Genomics* 2009, **10**:33.
- 4738 171. Aschtgen MS, Gavioli M, Dessen A, Lloubes R, Cascales E: **The SciZ protein anchors the enteroaggregative Escherichia coli Type VI secretion system to the cell wall.** *Mol Microbiol* 2010, **75**(4):886–899.
- 4740 172. Parsons LM, Lin F, Orban J: **Peptidoglycan recognition by Pal, an outer membrane lipoprotein.** *Biochemistry* 2006, **45**(7):2122–2128.
- 4741 173. Neumann U, Schiltz E, Stahl B, Hillenkamp F, Weckesser J: **A peptidoglycan binding domain in the porin-associated protein (PAP) of Rhodospirillum rubrum FR1.** *FEMS Microbiol Lett* 1996, **138**(1):55–58.
- 4742 174. Park JS, Lee WC, Yeo KJ, Ryu KS, Kumarasiri M, Heseck D, Lee M, Mobashery S, Song JH, Kim SI, et al: **Mechanism of anchoring of OmpA protein to the cell wall peptidoglycan of the gram-negative bacterial outer membrane.** *FASEB J* 2012, **26**(1):219–228.
- 4743 175. Babu MM, Priya ML, Selvan AT, Madera M, Gough J, Aravind L, Sankaran K: **A database of bacterial lipoproteins (DOLOP) with functional assignments to predicted lipoproteins.** *J Bacteriol* 2006, **188**(8):2761–2773.
- 4744 176. Leipe DD, Koonin EV, Aravind L: **STAND, a class of P-loop NTPases including animal and plant regulators of programmed cell death: multiple, complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer.** *J Mol Biol* 2004, **343**(1):1–28.
- 4745 177. Schwefel D, Frohlich C, Eichhorst J, Wiesner B, Behlke J, Aravind L, Daumke O: **Structural basis of oligomerization in septin-like GTPase of immunity-associated protein 2 (GIMAP2).** *Proc Natl Acad Sci USA* 2010, **107**(47):20299–20304.
- 4746 178. Velikovskiy CA, Deng L, Tasumi S, Iyer LM, Kerzic MC, Aravind L, Pancer Z, Mariuzza RA: **Structure of a lamprey variable lymphocyte receptor in complex with a protein antigen.** *Nat Struct Mol Biol* 2009, **16**(7):725–730.
- 4747 179. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV: **Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context.** *Genome Res* 2001, **11**(3):356–372.
- 4748 180. Koonin EV, Wolf YI, Aravind L: **Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach.** *Genome Res* 2001, **11**(2):240–252.

- 4781 181. Poole SJ, Diner EJ, Aoki SK, Braaten BA, t'Kint de Roodenbeke C, Low DA, 4782 Hayes CS: **Identification of functional toxin/immunity genes linked to** 4783 **contact-dependent growth inhibition (CDI) and rearrangement hotspots** 4784 **(Rhs) systems.** *PLoS Genet* 2011, **7**(8):e1002217.
- 4785 182. Kampstra P: **Beanplot: A Boxplot Alternative for Visual Comparison of** 4786 **Distributions.** *J Stat Softw* 2008, **28**(1):1–9.
- 4787 183. Diner EJ, Beck CM, Webb JS, Low DA, Hayes CS: **Identification of a target** 4788 **cell permissive factor required for contact-dependent growth inhibition** 4789 **(CDI).** *Genes Dev* 2012, **26**(5):515–525.
- 4790 184. Iyer LM, Koonin EV, Aravind L: **Evolution of bacterial RNA polymerase:** 4791 **implications for large-scale bacterial phylogeny, domain accretion, and** 4792 **horizontal gene transfer.** *Gene* 2004, **335**:73–88.
- 4793 185. Vollmer W: **Bacterial outer membrane evolution via sporulation?.** *Nat* 4794 *Chem Biol* 2012, **8**(1):14–18.
- 4795 186. Simeone R, Bottai D, Brosch R: **ESX/type VII secretion systems and** 4796 **their role in host-pathogen interaction.** *Curr Opin Microbiol* 2009, 4797 **12**(1):4–10.
- 4798 187. Pallen MJ, Chaudhuri RR, Henderson IR: **Genomic analysis of secretion** 4799 **systems.** *Curr Opin Microbiol* 2003, **6**(5):519–527.
- 4800 188. Bateman A, Bycroft M: **The structure of a LysM domain from E. coli** 4801 **membrane-bound lytic murein transglycosylase D (MltD).** *J Mol Biol* 2000, 4802 **299**(4):1113–1119.
- 4803 189. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, 4804 Gunasekaran P, Ceric G, Forslund K, et al: **The Pfam protein families** 4805 **database.** *Nucleic Acids Res* 2010, **38**(Database issue):D211–222.
- 4806 190. Ponting CP, Aravind L, Schultz J, Bork P, Koonin EV: **Eukaryotic signalling** 4807 **domain homologues in archaea and bacteria.** *Ancient ancestry and* 4808 *horizontal gene transfer. J Mol Biol* 1999, **289**(4):729–745.
- 4809 191. Wren BW: **A family of clostridial and streptococcal ligand-binding** 4810 **proteins with conserved C-terminal repeat sequences.** *Mol Microbiol* 1991, 4811 **5**(4):797–803.
- 4812 192. Dean P: **Functional domains and motifs of bacterial type III effector** 4813 **proteins and their roles in infection.** *FEMS Microbiol Rev* 2011, 4814 **35**(6):1100–1125.
- 4815 193. Hayes F, Van Melderden L: **Toxins-antitoxins: diversity, evolution and** 4816 **function.** *Crit Rev Biochem Mol Biol* 2011, **46**(5):386–408.
- 4817 194. Ishikawa K, Fukuda E, Kobayashi I: **Conflicts targeting epigenetic systems** 4818 **and their resolution by cell death: novel concepts for methyl-specific** 4819 **and other restriction systems.** *DNA Res* 2010, **17**(6):325–342.
- 4820 195. Iyer LM, Babu MM, Aravind L: **The HIRAN domain and recruitment of** 4821 **chromatin remodeling and repair activities to damaged DNA.** *Cell Cycle* 4822 **2006**, **5**(7):775–782.
- 4823 196. Iyer LM, Abhiman S, Aravind L: **MutL homologs in restriction-modification** 4824 **systems and the origin of eukaryotic MORC ATPases.** *Biol Direct* 2008, **3**:8.
- 4825 197. Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW: **Genomic islands: tools of bacterial horizontal gene transfer and** 4826 **evolution.** *FEMS Microbiol Rev* 2009, **33**(2):376–393.
- 4828 198. Nazina TN, Tourova TP, Poltarau AB, Novikova EV, Grigoryan AA, 4829 Ivanova AE, Lysenko AM, Petrunyaka VV, Osipov GA, Belyaev SS, et al: **Taxonomic study of aerobic thermophilic bacilli: descriptions of** 4830 **Geobacillus subterraneus gen. nov., sp. nov. and Geobacillus uzenensis** 4831 **sp. nov. from petroleum reservoirs and transfer of Bacillus** 4832 **stearothermophilus, Bacillus thermocatenuatus, Bacillus** 4833 **thermoleovorans, Bacillus kaustophilus, Bacillus thermodenitrificans to** 4834 **Geobacillus as the new combinations G. stearothermophilus, G. th.** *Int J* 4835 *Syst Evol Microbiol* 2001, **51**(Pt 2):433–446.
- 4837 199. Viswanathan VK: **Eating in, eating out.** *Gut Microbes* 2010, **1**(4):207–208.
- 4838 200. Paster BJ, Boches SK, Galvin JL, Ericson RE, Lau CN, Levanos VA, 4839 Sahasrabudhe A, Dewhirst FE: **Bacterial diversity in human subgingival** 4840 **plaque.** *J Bacteriol* 2001, **183**(12):3770–3783.
- 4841 201. Jain R, Poulos MG, Gros J, Chakravarty AK, Shuman S: **Substrate specificity** 4842 **and mutational analysis of Kluyveromyces lactis gamma-toxin, a eukaryal** 4843 **tRNA anticodon nuclease.** *RNA* 2011, **17**(7):1336–1343.
- 4844 202. Klassen R, Paluszynski JP, Wemhoff S, Pfeiffer A, Fricke J, Meinhardt F: **The** 4845 **primary target of the killer toxin from Pichia acaciae is tRNA(Gln).** *Mol* 4846 *Microbiol* 2008, **69**(3):681–697.
- 4847 203. Lu J, Huang B, Esberg A, Johansson MJ, Bystrom AS: **The Kluyveromyces** 4848 **lactis gamma-toxin targets tRNA anticodons.** *RNA* 2005, 4849 **11**(11):1648–1654.
- 4850 204. Conticello SG: **The AID/APOBEC family of nucleic acid mutators.** *Genome* 4851 *Biol* 2008, **9**(6):229.
205. Kanazawa T, Watanabe M, Matsushima-Hibiya Y, Kono T, Tanaka N, 4852 Koyama K, Sugimura T, Wakabayashi K: **Distinct roles for the** 4853 **N- and C-terminal regions in the cytotoxicity of pierisin-1, a putative** 4854 **ADP-ribosylating toxin from cabbage butterfly, against mammalian cells.** 4855 *Proc Natl Acad Sci USA* 2001, **98**(5):2226–2231.
206. Orth JH, Schorch B, Boundy S, French-Constant R, Kubick S, Aktories K: 4857 **Cell-free synthesis and characterization of a novel cytotoxic pierisin-like** 4858 **protein from the cabbage butterfly Pieris rapae.** *Toxicon* 2011, 4859 **57**(2):199–207.
207. van Kooij M, de Groot K, van Vugt H, Aten J, Snoek M: **Genotype versus** 4861 **phenotype: conflicting results in mapping a lung tumor susceptibility** 4862 **locus to the G7c recombination interval in the mouse MHC class III** 4863 **region.** *Immunogenetics* 2001, **53**(8):656–661.
208. Kumanovics A, Lindahl KF: **G7c in the lung tumor susceptibility (Lts)** 4865 **region of the Mhc class III region encodes a von Willebrand factor type** 4866 **A domain protein.** *Immunogenetics* 2001, **53**(11):64–68.
209. Taylor M, Mediannikov O, Raouf D, Greub G: **Endosymbiotic bacteria** 4868 **associated with nematodes, ticks and amoebae.** *FEMS Immunol Med* 4869 *Microbiol* 2012, **64**(1):21–31.
210. Yu C, Feng W, Wei Z, Miyanoiri Y, Wen W, Zhao Y, Zhang M: **Myosin VI** 4871 **undergoes cargo-mediated dimerization.** *Cell* 2009, **138**(3):537–548.
211. Zhang J, Xu LG, Han KJ, Shu HB: **Identification of a ZU5 and death** 4873 **domain-containing inhibitor of NF-kappaB.** *J Biol Chem* 2004, 4874 **279**(17):17819–17825.
212. Georgiades K, Madoui MA, Le P, Robert C, Raouf D: **Phylogenomic** 4876 **analysis of Odysseella thessalonicensis fortifies the common origin of** 4877 **Rickettsiales, Pelagibacter ubique and Reclimonas americana** 4878 **mitochondrion.** *PLoS One* 2011, **6**(9):e24857.
213. Fournier GP, Huang J, Gogarten JP: **Horizontal gene transfer from extinct** 4880 **and extant lineages: biological innovation and the coral of life.** *Philos* 4881 *Trans R Soc Lond B Biol Sci* 2009, **364**(1527):2229–2239.
214. Wolf YI, Aravind L, Koonin EV: **Rickettsiae and Chlamydiae: evidence of** 4883 **horizontal gene transfer and gene exchange.** *Trends Genet* 1999, 4884 **15**(5):173–175.
215. Iyer LM, Abhiman S, de Souza RF, Aravind L: **Origin and evolution of** 4886 **peptide-modifying dioxygenases and identification of the** 4887 **wybutosine hydroxylase/hydroperoxidase.** *Nucleic Acids Res* 2010, 4888 **38**(16):5261–5279.
216. Aravind L, Abhiman S, Iyer LM: **Natural history of the eukaryotic chromatin** 4890 **protein methylation system.** *Prog Mol Biol Transl Sci* 2011, **101**:105–176.
217. Smith EE, Sims EH, Spencer DH, Kaul R, Olson MV: **Evidence for diversifying** 4892 **selection at the pyoverdine locus of Pseudomonas aeruginosa.** *J Bacteriol* 4893 **2005**, **187**(6):2138–2147.
218. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database** 4896 **search programs.** *Nucleic Acids Res* 1997, **25**(17):3389–3402.
219. Eddy SR: **A new generation of homology search tools based on** 4898 **probabilistic inference.** *Genome Inform* 2009, **23**(1):205–211.
220. Lassmann T, Frings O, Sonnhammer EL: **Kalign2: high-performance** 4900 **multiple alignment of protein and nucleotide sequences allowing** 4901 **external features.** *Nucleic Acids Res* 2009, **37**(3):858–865.
221. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced** 4903 **time and space complexity.** *BMC Bioinforma* 2004, **5**:113.
222. Pei J, Sadreyev R, Grishin NV: **PCMA: fast and accurate multiple sequence** 4905 **alignment based on profile consistency.** *Bioinformatics* 2003, **19**(3):427–428.
223. Cole C, Barber JD, Barton GJ: **The Jpred 3 secondary structure prediction** 4907 **server.** *Nucleic Acids Res* 2008, **36**(Web Server issue):W197–W201.
224. Buchan DW, Ward SM, Lobley AE, Nugent TC, Bryson K, Jones DT: **Protein** 4909 **annotation and modelling servers at University College London.** *Nucleic* 4910 *Acids Res* 2010, **38**(Web Server issue):W563–W568.
225. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting** 4912 **transmembrane protein topology with a hidden Markov model:** 4913 **application to complete genomes.** *J Mol Biol* 2001, **305**(3):567–580.
226. Kall L, Krogh A, Sonnhammer EL: **Advantages of combined** 4915 **transmembrane topology and signal peptide prediction—the Phobius** 4916 **web server.** *Nucleic Acids Res* 2007, **35**(Web Server issue):W429–W432.
227. Soding J, Biegert A, Lupas AN: **The HHpred interactive server for protein** 4918 **homology detection and structure prediction.** *Nucleic Acids Res* 2005, 4919 **33**(Web Server issue):W244–W248.
228. Holm L, Kaariainen S, Rosenstrom P, Schenkel A: **Searching protein structure** 4921 **databases with DALI Lite v. 3.** *Bioinformatics* 2008, **24**(23):2780–2781.

- 4923 229. Price MN, Dehal PS, Arkin AP: **FastTree 2—approximately maximum-likelihood trees for large alignments.** *PLoS One* 2010, **5**(3):e9490.
4924
4925 230. Humphrey W, Dalke A, Schulten K: **VMD: visual molecular dynamics.** *J Mol Graph* 1996, **14**(1):33–38.
4926

4927 doi:10.1186/1745-6150-7-18

4928 **Cite this article as:** Zhang *et al.*: Polymorphic toxin systems:
4929 Comprehensive characterization of trafficking modes, processing,
4930 mechanisms of action, immunity and ecology using comparative
4931 genomics. *Biology Direct* 2012 **7**:18.

UNCORRECTED PROOF

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

