4532–4552 *Nucleic Acids Research, 2011, Vol. 39, No. 11* *Published online 8 February 2011*
doi:10.1093/nar/gkr036

A novel immunity system for bacterial nucleic acid degrading toxins and its recruitment in various eukaryotic and DNA viral systems

Dapeng Zhang, Lakshminarayan M. Iyer and L. Aravind*

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

Received December 23, 2010; Revised January 12, 2011; Accepted January 14, 2011

ABSTRACT

The use of nucleases as toxins for defense, offense or addiction of selfish elements is widely encountered across all life forms. Using sensitive sequence profile analysis methods, we characterize a novel superfamily (the SUKH superfamily) that unites a diverse group of proteins including Smi1/Knr4, PGs2, FBXO3, SKIP16, Syd, herpesviral US22, IRS1 and TRS1, and their bacterial homologs. Using contextual analysis we present evidence that the bacterial members of this superfamily are potential immunity proteins for a variety of toxin systems that also include the recently characterized contact-dependent inhibition (CDI) systems of proteobacteria. By analyzing the toxin proteins encoded in the neighborhood of the SUKH superfamily we predict that they possess domains belonging to diverse nuclease and nucleic acid deaminase families. These include at least eight distinct types of DNases belonging to HNH/EndoVII- and restriction endonuclease-fold, and RNases of the EndoU-like and colicin E3-like cytotoxic RNases-folds. The N-terminal domains of these toxins indicate that they are extruded by several distinct secretory mechanisms such as the two-partner system (shared with the CDI systems) in proteobacteria, ESAT-6/WXG-like ATP-dependent secretory systems in Gram-positive bacteria and the conventional Sec-dependent system in several bacterial lineages. The hedgehog-intein domain might also release a subset of toxic nuclease domains through auto-proteolytic action. Unlike classical colicin-like nuclease toxins, the overwhelming majority of toxin systems with the SUKH superfamily is chromosomally encoded and appears to have diversified through a recombination process combining different C-terminal nuclease domains to N-terminal secretion-related domains. Across the bacterial superkingdom these systems might participate in discriminating 'self' or kin from 'non-self' or non-kin strains. Using structural analysis we demonstrate that the SUKH domain possesses a versatile scaffold that can be used to bind a wide range of protein partners. In eukaryotes it appears to have been recruited as an adaptor to regulate modification of proteins by ubiquitination or polyglutamylation. Similarly, another widespread immunity protein from these toxin systems, namely the suppressor of fused (SuFu) superfamily has been recruited for comparable roles in eukaryotes. In animal DNA viruses, such as herpesviruses, poxviruses, iridoviruses and adenoviruses, the ability of the SUKH domain to bind diverse targets has been deployed to counter diverse anti-viral responses by interacting with specific host proteins.

INTRODUCTION

The use of toxins as a defensive, offensive or selfish addictive strategy is observed across the tree of life. Interestingly, a diverse set of protein toxins from distantly related organisms have a propensity to catalyze nucleic acid modifying or cleaving reactions in their target cells. Well-known examples are currently known from across the phylogenetic spectrum: plants deploy toxins such as ricin, abrin and modeccin to protect their seeds, which are RNA N-glycosidases that remove a specific purine

*To whom correspondence should be addressed. Tel: +1 301 594 2445; Fax: +1 301 480 9241; Email: aravind@ncbi.nlm.nih.gov

base from eukaryotic 28S rRNA to render it non-functional (1,2). In a similar vein, the fungal toxin α-sarcin, produced by fungi such as *Aspergillus giganteus*, acts as a specific endonuclease that cleaves the 28S rRNA at a position close to the site of action of the above plant toxins (3). Among animals the use of nucleic acid-targeting enzymes is observed in the venoms of snakes (4). Several animals, including vertebrates, are known to deploy cytotoxic RNases, such as RNase A, which potentially target RNA from bacteria and viruses (5). Bacteria are a particularly rich source of nucleic acid-targeting toxins, which are deployed in various contexts. Pathogenic bacteria secrete RNA N-glycosidases that target the 28S rRNA of eukaryotic hosts similar to the ricin-like plant toxins (6). Bacteria are also known to deploy RNase and DNase bacteriocins in intra- and possibly inter-specific competition that target molecules such as tRNA and genomic DNA (7). The best known are the plasmid-borne toxins of the model bacterium *Escherichia coli,* which kill closely related competing strains. Of these colicins E3, E4 and E6 cleave rRNA, colicins E5 and D cleave tRNA and colicins E2, E7, E8 and E9 cleave DNA (8). Additionally, bacterial genomes are also colonized by systems such as the toxin–antitoxin systems and restriction-modification systems which produce enzymes that function as nucleic acid-targeting toxins (9–12). In these systems the primary function of the toxin is to kill the host bacterial cell if the toxin encoding system is genetically disrupted in some way (10,11). Thus, they act as selfish elements that forcibly 'addict' the host to maintain them in genomes or plasmids. In many of these cases, organisms or genetic elements that produce the toxin also produce an antitoxin or immunity protein that renders the 'self' resistant to the action of the toxin. The study of these toxins and antitoxins or immunity proteins has not only expanded our understanding of the evolution of inter-species competition but also thrown considerable light on the biochemistry of nucleic acids and other molecules that interact with them (9–12). In practical terms these nucleic acid- targeting toxins and antitoxins/immunity proteins are potential reagents that could be utilized in numerous biotechnological contexts ranging from chemical analysis of nucleic acids to bio-defense.

Availability of an enormous wealth of genomic sequence data provides opportunities to identify novel versions of such toxins and associated immunity and delivery systems through computational analysis, thereby opening the door for new investigations on nucleic acid-modifying enzymes. The first step in this process requires detailed case-by-case analysis of protein sequences and structures using the best available methods for detecting sequence and structure similarity. Results from such an analysis of protein structures needs to be further combined with in-depth analysis of genomic contexts and domain architectures to glean novel functional associations. Finally, these results need to be placed in the context of phyletic patterns of the occurrence of various components of the system in order to reconstruct a total picture of their natural history and predict aspects of their biochemistry and biological functions. Indeed, such a strategy has allowed the prediction of novel biochemical activities and has laid the foundations for further systematic investigations of the toxin–antitoxin and peptide-modification systems of prokaryotes (9,13–15).

In this article we present the results of such a strategy that helped us uncover and characterize a remarkable, diverse class of nuclease toxins, whose immunity appears to depend primarily on a protein superfamily prototyped by the *Saccharomyces cerevisiae* protein Smi1/Knr4. The Smi1/Knr4 protein was first recovered in a screen for *S. cerevisiae* mutants that confer resistance to the killer toxin produced by the competing yeast species *Hansenula mrakii* (16,17). Smi1/Knr4 was shown to physically interact with the tyrosyl tRNA synthetase and it appears to functionally interact with the non-ribosomal peptide ligase Dit1, with a tRNA-synthetase-like catalytic domain, in the efficient synthesis of dityrosine a peptide metabolite that is typical of fungal spore-walls (18). Interestingly, it also shows synthetic lethal and physical interactions with a great number of proteins (19). Nevertheless, its exact significance and biochemical action has remained poorly understood (20). Parallel studies recovered other Smi1/Knr4 eukaryotic homologs namely FBXO3, a subunit of a SCF-type E3 ubiquitin ligase in vertebrates (21), and PGs2, a subunit of the tubulin polyglutamylase, which is a non-ribosomal peptide-ligase that links multiple glutamates to the γ-carboxyl group of target proteins (22,23). Exploratory sequence surveys suggested that Smi1/Knr4 homologs are also abundantly represented in bacteria (Smi1/Knr4 domain, Pfam: PF09346). Furthermore, our preliminary contextual analysis of conserved gene neighborhoods of these representatives suggested that they might be functionally linked to potential nucleases. Very recently, a novel contact-dependent inhibitory (CDI) toxin system has been reported in proteobacteria that delivers multiple nuclease toxins into target cells (24,25). Our observations indicated that Smi1/Knr4 homologs are potential immunity proteins in a subset of these CDI systems. Together, these observations prompted us to systematically investigate both the bacterial and eukaryotic Smi1/Knr4 homologs and explore their potential connection to nuclease toxins, their delivery and immunity against them. As a result we were able to identify a diverse group of previously unknown nuclease toxins and immunity proteins that are present across all the major bacterial lineages with considerable significance for intra-specific and host interactions. This investigation also allowed us to uncover diverse, previously unknown nuclease and deaminase domains in bacterial toxins and predict their folds and biochemical mechanisms. We also show that the Smi1/Knr4 homologs, which were ultimately derived from bacterial toxin–immunity systems, have been recruited by eukaryotic double-stranded DNA viruses to perform multiple roles in intracellular survival and morphogenesis of these viruses. Finally, we present evidence that the ability of the conserved domain in the Smi1/Knr4 superfamily of proteins to bind structurally diverse protein partners has been re-used in eukaryotes as a means to recruit targets to peptide-modifying systems such as the ubiquitin and the polyglutamylase systems.

METHODS

Iterative sequence profile searches were run using the PSI-BLAST program (26) against the non-redundant (nr) protein database of National Center for Biotechnology Information (NCBI). Similarity based clustering for both classification and culling of nearly identical sequences was performed using the BLASTCLUST program (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html). The HHpred program was used for profile–profile comparisons (27). Structure similarity searches were performed using the DaliLite program (28). Multiple sequence alignments were built by MUSCLE (29), PROMALS (30), KALIGN (31) and PCMA (32) programs, followed by manual adjustments on the basis of profile–profile and structural alignments. The consensus for alignments were calculated and colored by the Chroma program (33). Secondary structures were predicted using the JPred and PSIPred programs (34,35). For earlier known domains the PFAM database (36) was used as a guide, though the profiles were often augmented by addition of newly detected divergent members that were not detected by the original PFAM models. Clustering with BLASTCLUST followed by multiple sequence alignment and further sequence profile searches were used to identify other domains that were not present in the PFAM database. Signal peptides and transmembrane segments were detected using the TMHMM and Phobius programs (37,38). Contextual information from prokaryotic gene neighborhoods was retrieved by a PERL custom script that extracts the upstream and downstream genes of the query gene and uses BLASTCLUST to cluster the proteins to identify conserved gene-neighborhoods. Phylogenetic analysis was conducted using an approximately-maximum-likelihood method implemented in the FastTree 2.1 program under default parameters (39). The Modeller9v1 program (40) was utilized for homology modeling of the structure of the IRS1 N-terminal domain. Structural visualization and manipulations were performed using VMD (41) and PyMol (http://www.pymol.org) programs. The in-house TASS package, a collection of PERL scripts, was used to automate aspects of large-scale analysis of sequences, structures and genome context (Anantharaman, V., Balaji, S., and Aravind, L., unpublished data).

Species abbreviations used in the figures are: AHV1, Anguillid herpesvirus 1; Aave, *Acidovorax avenae*; Abau, *Acinetobacter baumannii*; Adef, *Abiotrophia defectiva*; Ahyd, *Aeromonas hydrophila*; Ahyd, *Anaerobaculum hydrogeniformans*; Amar, *Acaryochloris marina*; Aory, *Aspergillus oryzae*; Apla, *Arthrospira platensis*; Asp., *Anaeromyxobacter* sp.; Atha, *Arabidopsis thaliana*; Bamb, *Burkholderia ambifaria*; Bamy, *Bacillus amyloliquefaciens*; Bant, *Bacillus anthracis*; Bbac, *Bdellovibrio bacteriovorus*; Bcen, *Burkholderia cenocepacia*; Bcer, *Bacillus cereus*; Bcyt, *Bacillus cytotoxicus*; Bflo, *Branchiostoma floridae*; Bgra, *Bartonella grahamii*; Bmar, *Blastopirellula marina*; Bmcb, *Brevibacterium mcbrellneri*; Bmyc, *Bacillus mycoides*; Bpse, *Bacillus pseudofirmus*; Bpse, *Burkholderia pseudomallei*; Bpum, *Bacillus pumilus*; Bsel, *Bacillus selenitireducens*; Bsp., *Bacillus* sp.; Bsp., *Bacteroides* sp.; Bsp., *Beggiatoa* sp.; Bsub, *Bacillus subtilis*; Btha, *Burkholderia thailandensis*; Bthu, *Bacillus thuringiensis*; Btri, *Bartonella tribocorum*; Bvie, *Burkholderia vietnamiensis*; CHam, Candidatus *Hamiltonella*; CKor, Candidatus *Koribacter*; CV, Crocodilepox virus; Cace, *Clostridium acetobutylicum*; Caci, *Catenulispora acidiphila*; Cbac, *Campylobacterales bacterium*; Cbei, *Clostridium beijerinckii*; Cbol, *Clostridium bolteae*; Cbot, *Clostridium botulinum*; Ccar, *Clostridium carboxidivorans*; Ccel, *Clostridium cellulolyticum*; Ccel, *Clostridium cellulovorans*; Ccol, *Campylobacter coli*; Cdip, *Corynebacterium diphtheriae*; Cgle, *Chryseobacterium gleum*; Cgra, *Campylobacter gracilis*; Chom, *Cardiobacterium hominis*; Cjap, *Cellvibrio japonicus*; Clen, *Clostridium lentocellum*; Clep, *Clostridium leptum*; Cmic, *Clavibacter michiganensis*; Csak, *Cronobacter sakazakii*; Csho, *Campylobacter showae*; Csp., *Chloroflexus* sp.; Csp., *Clostridium* sp.; Csp., *Cyanothece* sp.; Cspu, *Capnocytophaga sputigena*; Ctro, *Candida tropicalis*; Ctur, *Cronobacter turicensis*; Cure, *Corynebacterium urealyticum*; Ddad, *Dickeya dadantii*; Ecol, *Escherichia* coli; Efer, *Escherichia fergusonii*; Erec, *Eubacterium rectale*; Even, *Eubacterium ventriosum*; Exsp, *Exiguobacterium* sp.; FAV1, Fowl adenovirus 10; Faln, *Frankia alni*; Fmor, *Fusobacterium mortiferum*; Fsp., *Fusobacterium* sp.; Fsym, *Frankia* symbiont; GHV2, Gallid herpesvirus 2; Gaur, *Gemmatimonas aurantiaca*; Ghae, *Gemella haemolysans*; Gint, *Giardia intestinalis*; Gsp., *Geobacillus* sp.; Gsp., *Geobacter* sp.; Gvio, *Gloeobacter violaceus*; HHV5, Human herpesvirus 5; HHV7sJ, Human herpesvirus 7 strain JI; Hasp, *Halobacterium* sp.; Haur, *Herpetosiphon aurantiacus*; Hbor, *Halogeometricum borinquense*; Hche, *Hahella chejuensis*; Hoch, *Haliangium ochraceum*; Hpyl, *Helicobacter pylori*; Hsap, *Homo sapiens*; Hsom, *Haemophilus somnus*; Iloi, *Idiomarina loihiensis*; Kalg, *Kordia algicida*; Kfla, *Kribbella flavida*; Krad, *Kineococcus radiotolerans*; Kset, *Kitasatospora setae*; Lara, *Lentisphaera araneosa*; Lgoo, *Leptotrichia goodfellowii*; Ljoh, *Lactobacillus johnsonii*; Lpla, *Lactobacillus plantarum*; Lsph, *Lysinibacillus sphaericus*; Mabs, *Mycobacterium abscessus*; Maur, *Micromonospora aurantiaca*; Mcat, *Moraxella catarrhalis*; Mext, *Methylobacterium extorquens*; Mgil, *Mycobacterium gilvum*; Minf, *Methylacidiphilum infernorum*; Mlep, *Mycobacterium leprae*; Mmar, *Microscilla marina*; Msp., *Micromonospora* sp.; Msp., *Mycobacterium* sp.; Mxan, *Myxococcus xanthus*; Ndas, *Nocardiopsis dassonvillei*; Nmen, *Neisseria meningitidis*; Nmuc, *Neisseria mucosa*; Nmul, *Nakamurella multipartita*; Nsic, *Neisseria sicca*; Oana, *Ornithorhynchus anatinus*; Osin, *Oribacterium sinus*; Patl, *Pseudoalteromonas atlantica*; Patr, *Pectobacterium atrosepticum*; PbCVN, Paramecium bursaria Chlorella virus NY2A; Pcar, Pectobacterium carotovorum; Pcry, Psychrobacter cryohalolentis; Pdag, Pasteurella dagmatis; Plum, *Photorhabdus luminescens*; Pmar, *Planctomyces maris*; Pmar, *Prochlorococcus marinus*; Pmel, *Prevotella melaninogenica*; Pmir, *Proteus mirabilis*; Pmul, *Pasteurella multocida*; Pput, *Pseudomonas putida*; Prum, *Prevotella ruminicola*; Psp., *Paenibacillus* sp.; Psp., *Prevotella* sp.; Pstu, *Providencia stuartii*; Psyr, *Pseudomonas syringae*; Ptim, *Prevotella*

timonensis; Ptor, *Psychroflexus torquis*; RHV1, Ranid herpesvirus 1; RbIV, Rock bream iridovirus; Rcat, *Rana catesbeiana*; Rery, *Rhodococcus erythropolis*; Rfla, *Ruminococcus flavefaciens*; Rinu, *Roseburia inulinivorans*; Rsol, *Ralstonia solanacearum*; Salb, *Streptomyces albus*; Save, *Streptomyces avermitilis*; Sbal, *Shewanella baltica*; Sbin, *Streptomyces bingchenggensis*; Scla, *Streptomyces clavuligerus*; Scoe, *Streptomyces coelicolor*; Sent, *Salmonella enterica*; Sgri, *Streptomyces griseoflavus*; Sgri, *Streptomyces griseus*; Shyg, *Streptomyces hygroscopicus*; Sisp, *Silicibacter* sp.; Slin, *Spirosoma linguale*; Smut, *Streptococcus mutans*; Snas, *Stackebrandtia nassauensis*; Sone, *Shewanella oneidensis*; Spie, *Shewanella piezotolerans*; Spom, *Schizosaccharomyces pombe*; Spri, *Streptomyces pristinaespiralis*; Srot, *Segniliparus rotundus*; Ssal, *Salmo salar*; Ssp., *Streptomyces* sp.; Sspi, *Sphingobacterium spiritivorum*; Sste, *Sagittula stellata*; Ssui, *Streptococcus suis*; Ssvi, *Streptomyces sviceus*; StIV, Soft-shelled turtle iridovirus; Ster, *Sebaldella termitidis*; Stro, *Salinispora tropica*; Svir, *Streptomyces viridochromogenes*; Swol, *Syntrophomonas wolfei*; Taue, *Tolumonas auensis*; Tcur, *Thermomonospora curvata*; Tfus, *Thermobifida fusca*; Tsp., *Thauera* sp.; Tthe, *Tetrahymena thermophila*; Ttur, *Teredinibacter turnerae*; Valg, *Vibrio alginolyticus*; Vcho, *Vibrio cholerae*; Veis, *Verminephrobacter eiseniae*; Vmet, *Vibrio metschnikovii*; Vmim, *Vibrio mimicus*; Vpar, *Vibrio parahaemolyticus*; Vsp., *Vibrio* sp.; Vspl, *Vibrio splendidus*; Vvul, *Vibrio vulnificus*; Wend, *Wolbachia endosymbiont*.

RESULTS AND DISCUSSION

Sequence profile searches and structural comparisons reveal a vast superfamily of Smi1-related proteins

As a first step to computationally characterize the Smi1/Knr4 protein, we analyzed it using the SEG program to identify potential globular regions in it (42). This indicated the presence of a single globular domain that was then used as a seed in iterative sequence profile searches of the nr database with PSI-BLAST and JACKHMMER from the HMMER3 package. In addition to recovering other eukaryotic proteins with a homologous region, such as FBXO3 from animals, SKIP16 from plants and PGs2, a subunit of tubulin polyglutamylase complex, the search also recovered a large number of bacterial proteins such as the *Bacillus subtilis* YobK. Given the great diversity of sequences recovered prior to convergence from bacteria, we initiated transitive sequence profile searches with several distinct bacterial starting points to achieve maximal coverage in terms of detection. We also noted that a crystal structure for YobK has been solved by the joint structural genomics initiative (PDB: 2PRV). We used this structure as a query for structure similarity searches using the DALIlite program and recovered hits to four other homologous structures (3FFV, 2PAG, 2ICG, 3D5P; $Z > 7.5$). Of these, 3FFV was the structure of the earlier characterized protein Syd from *E. coli* which interacts with SecY, a key component of the Sec-dependent protein secretion system that traffics proteins across the bacterial inner membrane (43–45). Consistent with this, we also found that Syd homologs were recovered with borderline *e*-values ($e \sim 0.03$–0.05) in the above JACKHMMER and PSI-BLAST searches. Hence we included the Syd homologs in the profiles to further expand the relationships of the group of proteins homologous to Smi1/Knr4. At convergence, some of these searches also recovered with borderline *e*-values proteins ($e \sim 0.05$) from certain DNA viruses such as FPV250 (gi: 9634920) from the fowl poxvirus, and the US22 family of proteins (e.g. US22, UL26, IRS1 and TRS1) from herpesviruses. To confirm the relationship of these proteins to Smi1 we used them in a profile–profile comparison search with the HHpred program against a library of HMMs created using the sequence of polypeptides in the PDB database as a query. These searches recovered the structures 2PRV, 3FFV and 2ICG as the best hits with significant *P*-values ($P = 10^{-4\ \text{to}\ -8}$). Furthermore, examination of the hits produced by the viral proteins in profile–profile comparisons showed that most of the versions from herpesviruses possessed two tandem repeats of the domain homologous to Smi1. Additional transitive searches with these viral proteins revealed that homologous proteins are present in a number of distantly related or unrelated DNA viruses. Finally, the above searches also recovered hits to two distinct groups of proteins each with over 100 representatives in the nr database, predominantly from bacteria, typified respectively by CA_C3700 (gi: 15896931) from *C. acetobutylicum* and SGR_4389 (gi: 182438182) from *S. griseus*. Profile–profile comparisons with the HHpred program using alignments of each of these groups of proteins also confirmed their relationship to the Smi1-like proteins via recovery of significant hits ($e = 10^{-4\ \text{to}\ -6}$) to HMMs generated using the sequences of 2PRV and 3FFV as best hits. Thus, it became clear that Smi1/Knr4 defines a large superfamily of conserved domains that is widespread in bacteria, eukaryotes and various DNA viruses but practically absent in currently sequenced archaeal genomes. We accordingly named it the SUKH (for Syd, US22, Knr4 homology) domain superfamily.

Structural features and internal diversity of the SUKH domain superfamily

Despite the low average pairwise sequence similarity across this superfamily, all representatives are known or predicted to possess a similar core fold comprising of four conserved helices and six strands (Figure 1, Supplementary Data). Strands 1 and 2 form a β-hairpin and the strands 3–6 form a 4-stranded β-meander; however, the β-hairpin and the β-meander show only limited or no hydrogen-bonding along their length, despite being spatially beside each other. Thus, the structural core of the SUKH domain can be described as a split β-sheet with only weak interaction between its two parts. This structural peculiarity could potentially be critical for the functional interactions of the domain (see below). Based on sequence-similarity-based clustering and phylogenetic analysis five major groups can be recognized within the SUKH domain superfamily (Figure 1,

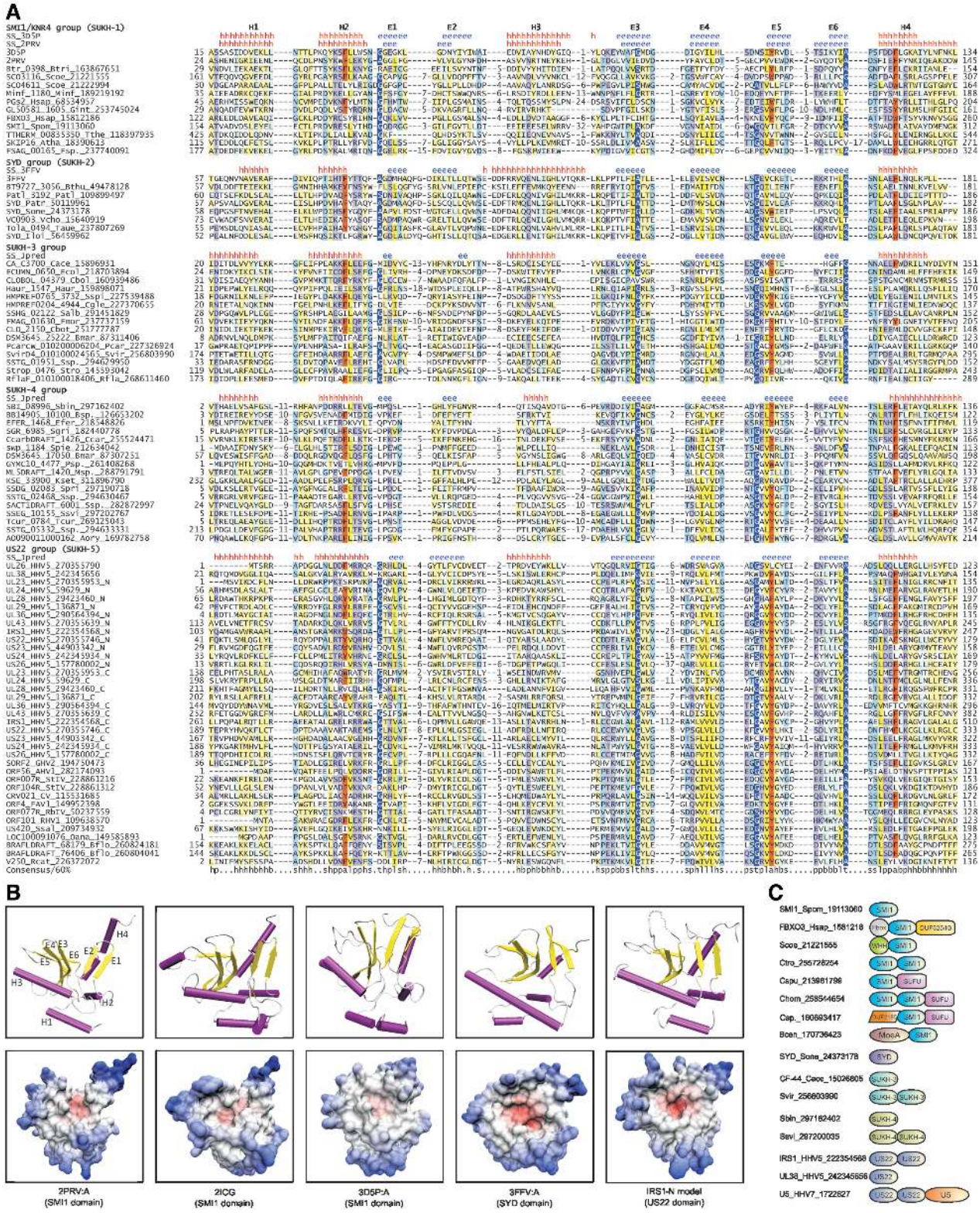

Figure 1. (**A**) Multiple sequence alignment of the different groups of the SUKH superfamily. The consensus was derived using the following amino acid classes: a, aromatic (FHWY, black on orange); b, big (EFHIKLMQRWY, black on light blue); h, hydrophobic (ACFGHILMTVWY, black on yellow); l, aliphatic (ILV, black on dark yellow); p, polar (CDEHKNQRST, blue on gray); s, small (ACDGNPSTV, black on blue); t, tiny (AGS, white on dark blue). Secondary structures derived from PDB structures or predicted using Jpred program are indicated above the alignment ('e' in blue, β-sheet; 'h' in red, α-helix). The numbers in bracket are indicative of the excluded residues from sequences. (**B**) Cartoons of known structures and a homology model of the US22 IRS-N domain made using Modeller are shown in approximately similar orientation. The α-helices are shown in purple, β-sheets in yellow, and loops in gray. Surface diagrams are colored based on their positions relative to the center of the structure (outside to inside: blue to red) to illustrate the cleft. (**C**) Domain architectures of representatives of the SUKH superfamily. Other than the domain abbreviations already provided in the text and the rest of the domains are the Ig-fold domain overlaps with PFAM DUF525; MoeA is a domain found in the MoeA protein of the molybdopterin biosynthesis pathway; U5, herpesvirus U5-like family (PF05999).

Supplementary Data). The first of these, and the most widespread, is the one typified by Smi1/Knr4, FBXO3, SKIP16, PGs2 and YobK (that entirely includes the PFAM model PF09346, 'SMI1/KNR4 family', and additional proteins not detected by that model within it) and is seen in both bacteria and eukaryotes. This ensemble, which we term Smi1-like or SUKH-1 group includes the majority of the SUKH domains. We term the second group, prototyped by Syd, the Syd-like or SUKH-2 group. This group is largely restricted to the gammaproteobacteria and firmicutes. The SUKH-3 group prototyped by CA_C3700 (gi: 15896931) is widely distributed across most bacterial lineages. The group prototyped by SGR_4389 (gi: 182438182), the SUKH-4 group, is again seen in several bacteria and sporadically in fungi. The SUKH-5 or US22-like group is present in fowl adenoviruses, various vertebrate iridoviruses, archosaur poxviruses (Crocodilepox virus and Fowlpox virus), and in multiple copies in several herpesviruses (representatives of the alphaherpesvirus, betaherpesvirus and alloherpesvirus clades). Members of this group are also encoded by genomes of the early-branching chordate *Branchiostoma*, the salmon, the frog *Rana catesbeiana* and the duckbilled platypus, where they appear to have been acquired from the genomes of integrated herpesviruses (46). Phylogenetic analysis of each group, along with the phyletic patterns, strongly suggests that SUKH domain proteins have been widely disseminated both within and across the superkingdoms via extensive lateral transfer (Supplementary Data). In light of this pattern, the near complete absence of this superfamily in archaea suggests that there could be certain specific functional barriers that prevent acquisition of the SUKH domain by that superkingdom. Phylogenetic analysis strongly suggests that the groups SUKH-2–5 are monophyletic clades. The largest group, SUKH-1 is likely to represent the ancestral group from within which the above clades have diversified through rapid sequence divergence.

Contextual analysis of the SUKH domain proteins suggest potential functional linkages with nuclease toxins in bacteria

Contextual information gleaned from gene neighborhoods in prokaryotes and domain architectures of proteins, when combined with sequence analysis, can be a powerful means of discerning protein function (47). Indeed, this method has proven particularly effective in both function prediction and identification of new analogous systems, using the organizational syntax of tightly linked genes, in case of toxin–antitoxin and restriction-modification systems (9,13,14,23,48). To better understand the role of the SUKH domain we performed a detailed analysis of the gene-neighborhoods of all bacterial genes encoding a protein with this domain (Figure 2). Consequently, we were able to identify at least three striking themes among the gene-neighborhoods of this superfamily. Firstly, across the bacterial phylogenetic tree we found numerous genomic neighborhoods that linked two or more adjacent genes encoding SUKH domain proteins. In certain cases, e.g. *B. grahamii* (gi: 240850988), we found tandem arrays with up to six paralogous SUKH superfamily genes (Figure 2). We found that in several instances these paralogous versions are not closely related and in certain cases adjacent paralogs might belong to completely different SUKH groups. For example, we found combinations of genes encoding proteins belonging to the Smi1-like (SUKH-1), Syd-like (SUKH-2), SUKH-3 and SUKH-4 groups in the same neighborhood in several bacteria such as *B. cereus MM3* and various *Streptomyces* species (Figure 2). This observation suggested that there appears to be selective pressure for the diversification of the linked SUKH domain proteins encoded in a gene neighborhood either via sequence divergence, or independent assembly of neighborhoods from distantly related paralogs of different groups. This situation, wherein multiple paralogous genes are linked together as tandem arrays in a neighborhood, is relatively rare in bacteria (49). Given that products of genes linked in conserved gene-neighborhoods physically interact, it is possible that these paralogs interact to form a single complex (47). On the other hand, the multiple paralogs could also represent different alternative versions of the same component of a system which is under selection to display diversity. Given the great variability in the numbers and types of paralogous versions of the SUKH superfamily encoded by these neighborhoods, we favor the later explanation in this case (details see below). The second major feature that emerged from the analysis of gene neighborhoods was the linkage of genes encoding diverse SUKH superfamily members to genes encoding different types of nucleases (Figure 2). Among these, we observed multiple linkages in distantly related bacteria, such as *B. thuringiensis* and *M. marina* and *S. griseoflavus,* to genes for nucleases of the metal-dependent NucA family, which includes the well-studied *S. marcescens* secreted endonuclease (50) and the *Anabaena* non-specific endonuclease NucA, which degrades both RNA and DNA (51). Another prominent linkage observed in several bacteria, such as *M.infernorum*, various *Bacillus* species and *N. mucosa*, was to genes encoding proteins with a HNH superfamily nuclease domain (Figure 3). Sequence analysis showed that several of the HNH domains were related to similar nuclease domains found in previously studied bacteriocins such as pyocin AP41 of *P. aeruginosa, Klebsiella* klebicin B and colicin E8 of *E. coli* (52). These linkages involved members of both the Smi1-like and Syd-like groups; thus, despite their diversity, potential functional interactions with different types of nuclease domains are a common feature of the bacterial representatives of the SUKH superfamily.

The third major linkage we observed was between SUKH superfamily genes and those encoding gigantic bacterial surface proteins with repetitive motifs such as the hemagglutinin-repeats, RHS repeats (YD) and another previously uncharacterized α-helical repeat motif. All these proteins showed a characteristic feature of possessing a highly variable but globular domain at the extreme C-terminus of the protein, downstream of the repetitive region. These proteins also usually contain

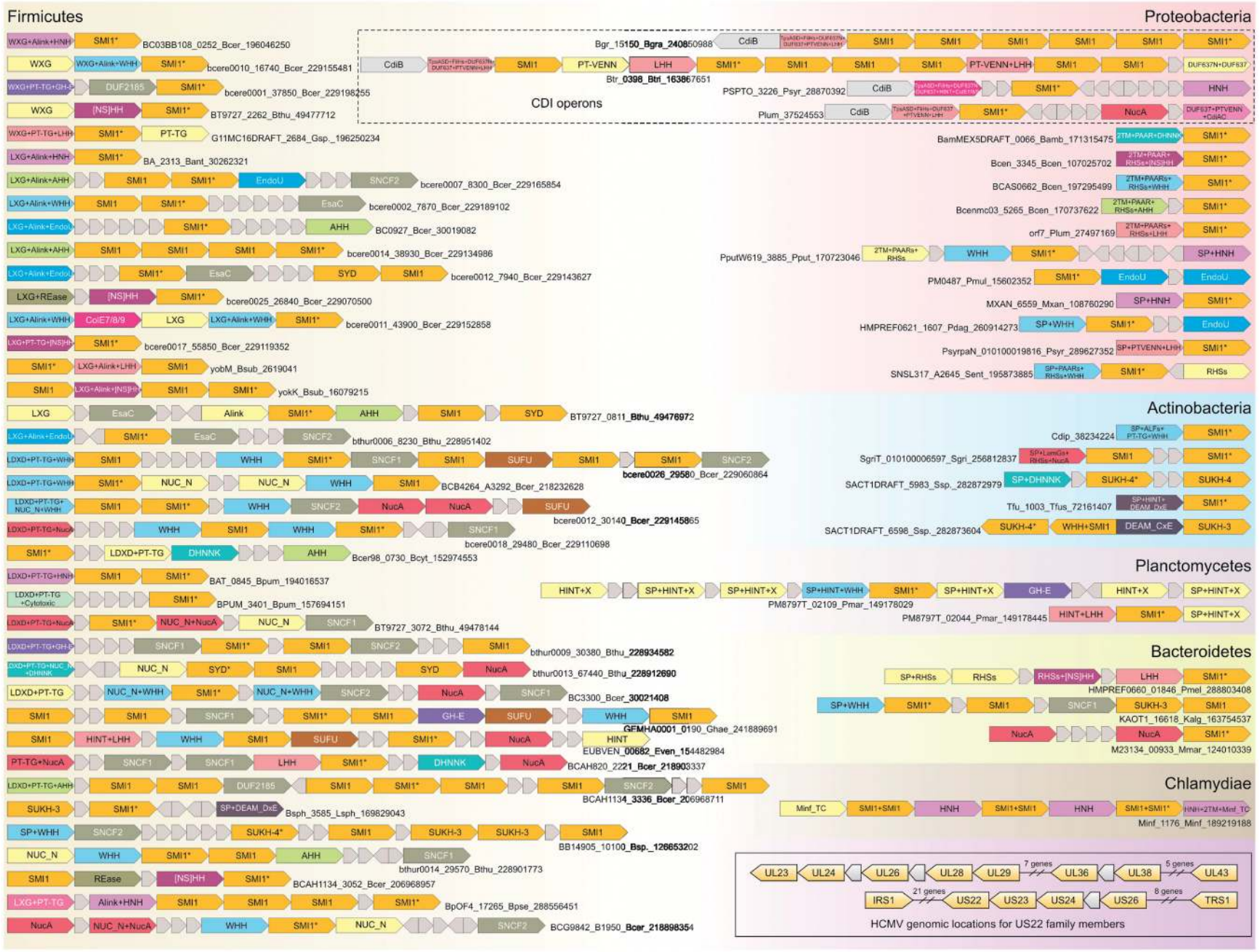

Figure 2. Gene neighborhoods of representative SUKH superfamily genes in bacteria are shown. Individual genes are represented as box arrows pointing from the 5′ to the 3′-end of the coding frame. Genes were named by their domain architectures. For each operon, the gi of the SUKH gene (marked with a star) and species name are indicated. Most ORFs shown as gray boxes are small ones (<80 amino acids) that appear to be false gene predictions, or in a few cases are uncharacterized genes. For species abbreviations see 'Materials and Methods' section.

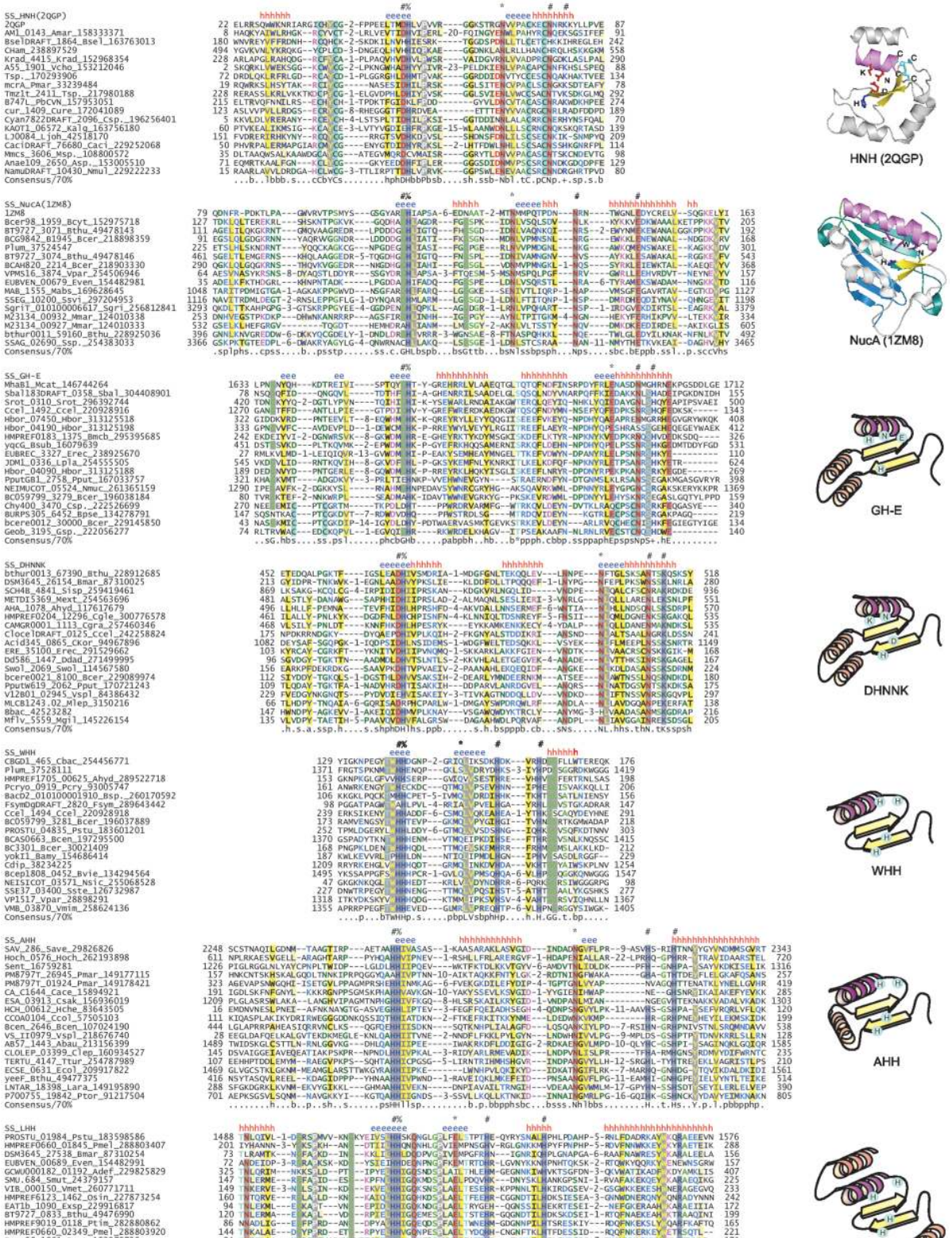

Figure 3. Multiple sequence alignments and structural scaffolds of the distinct families of the HNH/EndoVII fold recovered in SUKH neighborhoods: HNH, NucA, WHH, LHH, AHH, DH-NNK and GH-E. Their secondary structures are indicated above the alignments ('e' in blue, β-sheet; 'h' in red, α-helix). The numbers in bracket are indicative of the excluded residues from sequences. 'hash' indicates the residues involved in metal ion-binding, 'percent' symbol indicates the conserved histidine which is required for activation of the water molecule for hydrolysis and 'asterisk' indicates the conserved asparagines. On the right, structures of HNH and EndoG families are shown as cartoon representations with the central structural core colored by structural element type (α-helices in purple, β-sheets in yellow), and key catalytic residues highlighted. For those newly identified families, inferred topology diagrams of their core nuclease domains are shown with conserved catalytic residues.

certain domains related to adhesion and the two-partner secretory (TPS) system N-terminal to the repetitive region, such as PAAR (PFAM: PF05488) and the TpsA-secretion domain (TpsA-SD, also known as the filamentous hemagglutinin FhaB secretory domain; PFAM: PF05860) with a pectate lyase-like fold (53–55). Some of these proteins with repetitive domains, which were recovered in our analysis of SUKH superfamily neighborhoods, are representatives of toxins of the CDI systems (Figure 2) that were reported even as this study was being prepared for submission (24,25). Like the above proteins, the CDI toxins are characterized by multiple N-terminal TpsA-SD domains and hemagglutinin-repeats combined with polymorphic C-terminal domains that vary greatly between different CDI toxins. In all these CDI proteins the polymorphic C-terminal domain is separated from the repetitive region by either or both of two small α-helical domains annotated as domains of unknown function in the PFAM database (DUF638 or DUF637). Furthermore, it was shown that the protein encoded by the gene following the CDI toxin was an immunity gene, whose product provided resistance against the toxin to the cell that was producing it (25). By this criterion it became clear that the SUKH superfamily genes in the CDI operons were actually immunity proteins for the toxins encoded by the upstream genes. However, in contrast to the pan-bacterial distribution of the SUKH superfamily, the CDI operons were only observed in proteobacteria (25). Furthermore, we observed that polymorphic C-terminal domains of the CDI toxins, which are found linked to the SUKH superfamily immunity proteins in CDI systems, are also seen in bacterial lineages outside of proteobacteria, where too they are linked to SUKH superfamily genes. In these cases they are linked to other N-terminal domains that are distinct from the TpsA-SD and hemagglutinin repeat domains. Studies on CDI systems indicated that the toxin function resides in the polymorphic C-terminal domains and at least two of these domains are nuclease toxins that cleave both tRNAs and DNA (25). Our above observations indicate that outside of CDI systems, the SUKH superfamily genes are linked to genes encoding the HNH and NucA nucleases; hence, it is likely that even these nucleases function as distinct but analogous toxins that cleave nucleic acids in target cells. Together, the above observations raised the possibility that the SUKH superfamily protein might serve as immunity proteins, not just in certain proteobacterial CDI systems, but also more generally function, across all major bacterial lineages, to protect against linked genes, which are predicted to act as toxins.

Interestingly, in addition to gene-neighborhoods with multiple tandem divergent SUKH superfamily genes, in several bacteria, we also observed notable lineage-specific expansions of SUKH domain proteins (e.g. 21 paralogs in *Gemmata obscuriglobus*, 20 paralogs in *C. gingivalis* and 15 in *S. albus*). These observations also make sense in light of the above toxin–immunity protein hypothesis: while the SUKH superfamily gene adjacent to a nuclease or CDI toxin gene is likely to provide immunity to the 'self' toxin, the supernumerary SUKH superfamily genes, which occur as tandem arrays or as isolated versions, might provide immunity against other 'non-self' toxins delivered by competing bacteria in the environment. Such associations of multiple distinct immunity genes have also been observed in the case of plasmid-borne colicin gene operons (8). Other features of the genomics of the SUKH superfamily also support this proposal. Gene neighborhoods encoding SUKH proteins and linked nucleases or CDI toxin are highly variable in terms of being present or absent between different strains of the same species or between different closely related species which share an otherwise similar genomic organization. Secondly, there appear to have been recent duplications of entire loci encompassing these gene-neighborhoods within the same genome in several bacteria (Supplementary Data). This kind of phyletic and genomic polymorphism is also typical of loci involved in inter- and intra-genomic competition such as toxin–antitoxin, restriction-modification and virulence toxin systems (6,9,10,15), suggesting that even systems with SUKH superfamily proteins might have comparable roles. To test this proposal further, as the first line of investigation, we aimed at exploring further the link between nucleases and the SUKH domain proteins. While the polymorphic C-terminal domains of two CDI toxins have been characterized as nucleases, the C-terminal domains of those CDI toxins which are found linked to the SUKH superfamily immunity proteins have not be characterized. We speculated that these domains, along with some of the other uncharacterized domains in proteins encoded by conserved gene-neighborhoods containing a SUKH superfamily gene, might be as yet uncharacterized nuclease domains. As a second line of investigation we sought to uncover those among the associated uncharacterized domains, which might have a role in distinct toxin-trafficking mechanisms, comparable to the two-partner system used by the proteobacterial CDIs. Therefore, to accomplish these two objectives and identify other components of these systems we resorted to systematic sequence analysis of the uncharacterized proteins recovered in the above gene-neighborhood analysis.

Sequence analysis reveals the presence of 11 distinct families of nuclease toxins encoded by genes adjacent to those of the SUKH superfamily

Sequence analysis indicated that at least 11 distinct families of domains recovered in our searches in proteins encoded by genes adjacent to one encoding a SUKH domain protein are potential nucleases. While some of these, as noted above, belong to the earlier characterized families, several of those identified here belong to entirely new families or are highly distinctive previously unrecognized versions of previously known families (Figures 3–5 and Supplementary Data). Identification of this diverse panoply of nuclease domains as being functionally linked to the SUKH domain lends critical support to the proposal that this domain functions primarily as an immunity protein against nucleic acid-targeting toxins in bacteria. We briefly describe below these newly identified nuclease domains.

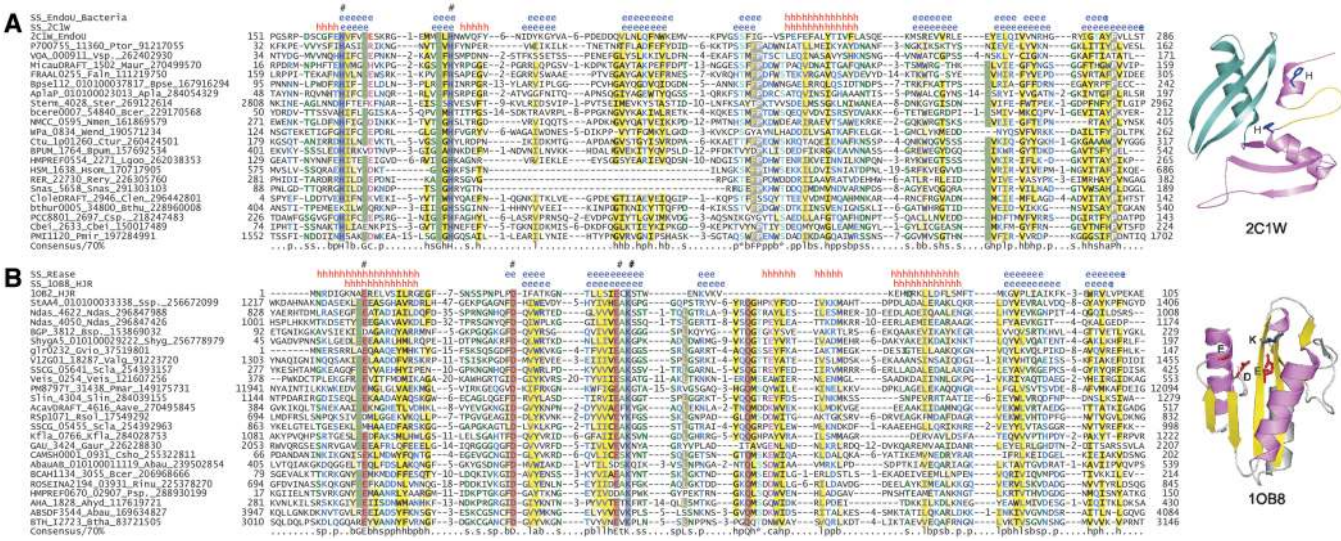

Figure 4. (**A**) Multiple sequence alignment of the EndoU family emphasizing the new bacterial versions found in this study. The eukaryotic EndoU domain (PDB: 2c1w) is shown to the right to indicate the spatial position of the conserved elements and the two units with three-strands each. (**B**) Multiple sequence alignment of the newly identified REase family. The structure of the archaeal Holliday junction resolvase (PDB: 1OB8) is shown to the right to indicate the spatial location of the conserved residues in this fold. Secondary structure elements are indicated above the alignments ('e' in blue, β-sheet; 'h' in red, α-helix). The numbers in brackets represent excluded residues from sequences and 'hash' indicates the catalytic residues.

Nuclease toxins of the HNH/ENDOVII fold. The HNH or the ENDOVII fold is a version of the treble-clef fold. The treble-clef fold is one of the most prevalent Zn-binding motifs across the three superkingdoms of life (56). Classical HNH nucleases, like the restriction endonuclease (REase) McrA and the T4 endonuclease VII, contain the four conserved, Zn-chelating cysteines of the treble-clef fold (52). However, these cysteines are lost in several forms, such as the REase MboII, colicin E8 and the NucA family, but these domains still retain the characteristic structural geometry of the treble-clef (52,56). The active site of these enzymes is formed at the interface of the characteristic helix and β-hairpin and contains a divalent cation, which is chelated by three polar residues usually from the first strand of the β-hairpin and the C-terminal helix of the treble-clef fold. The residues chelating the metal are typically histidine, aspartate and asparagine but their exact configuration can greatly vary between different members of this fold making them difficult targets for identification through sequence analysis (52). Among the nucleases of this fold occurring in the neighborhood of the SUKH superfamily we observed eight distinct families spanning the entire gamut ranging from conventional HNH nucleases to certain highly derived forms that have not be identified before. The conventional HNH versions (e.g.AM1_0143, gi: 158333371 from the cyanobacterium *A. marina*) retain all the four cysteines of the treble-clef fold and a typical arrangement of residues chelating the catalytic metal. Others, like the nuclease domains of the PSPTO_3229 protein from *P. syringae* (gi: 28870395) and some CDI proteins, belong to the colicin E7/E8/E9 family (Figure 2). A highly derived version is represented by the NucA family (57), where structural analysis reveals that a treble-clef domain which has lost the characteristic cysteines is inserted between two copies of a three-stranded domain with distinct loop-like C-terminal extensions (Figure 3). We uncovered several divergent, earlier unrecognized NucA family nuclease domains in both the SUKH superfamily neighborhoods and CDI systems, such as those typified by the *B. subtilis* protein YeeF (gi: 251757354). The structural organization of the NucA domain suggests that it arose from an ancestral HNH/EndoVII domain, which 'carried' these duplicated three-stranded units along with it to form a more complex domain. Consistent with this proposal, we discovered a family of novel HNH fold nucleases in our gene-neighborhoods, which contain an active site similar to the NucA nucleases, but are standalone versions without the two flanking three-stranded units. We called this family GH-E after the three conserved residues associated with the active site typical of these domains. Interestingly, a subset of the GH-E family preserves the conserved cysteines of the treble-clef suggesting that they indeed represent the potential evolutionary intermediate from a classical HNH domain to the derived NucA-like forms (Figure 3).

We also recovered three other novel families of domains, which are respectively typified by nearly absolutely conserved tripeptide sequence motifs LHH, WHH and AHH (Figure 3). Most CDI operons, which encode a SUKH domain immunity protein, have proximal toxin genes with a LHH domain as the polymorphic C-terminal unit of their products (Figure 2). Additionally, the LHH domain is found in products of genes adjacent to the SUKH superfamily gene outside of proteobacteria in several other bacterial lineages such as firmicutes, actinobacteria, bacteroidetes and planctomycetes (Figure 2). Although we also found the WHH domain as the polymorphic toxin unit of a subset

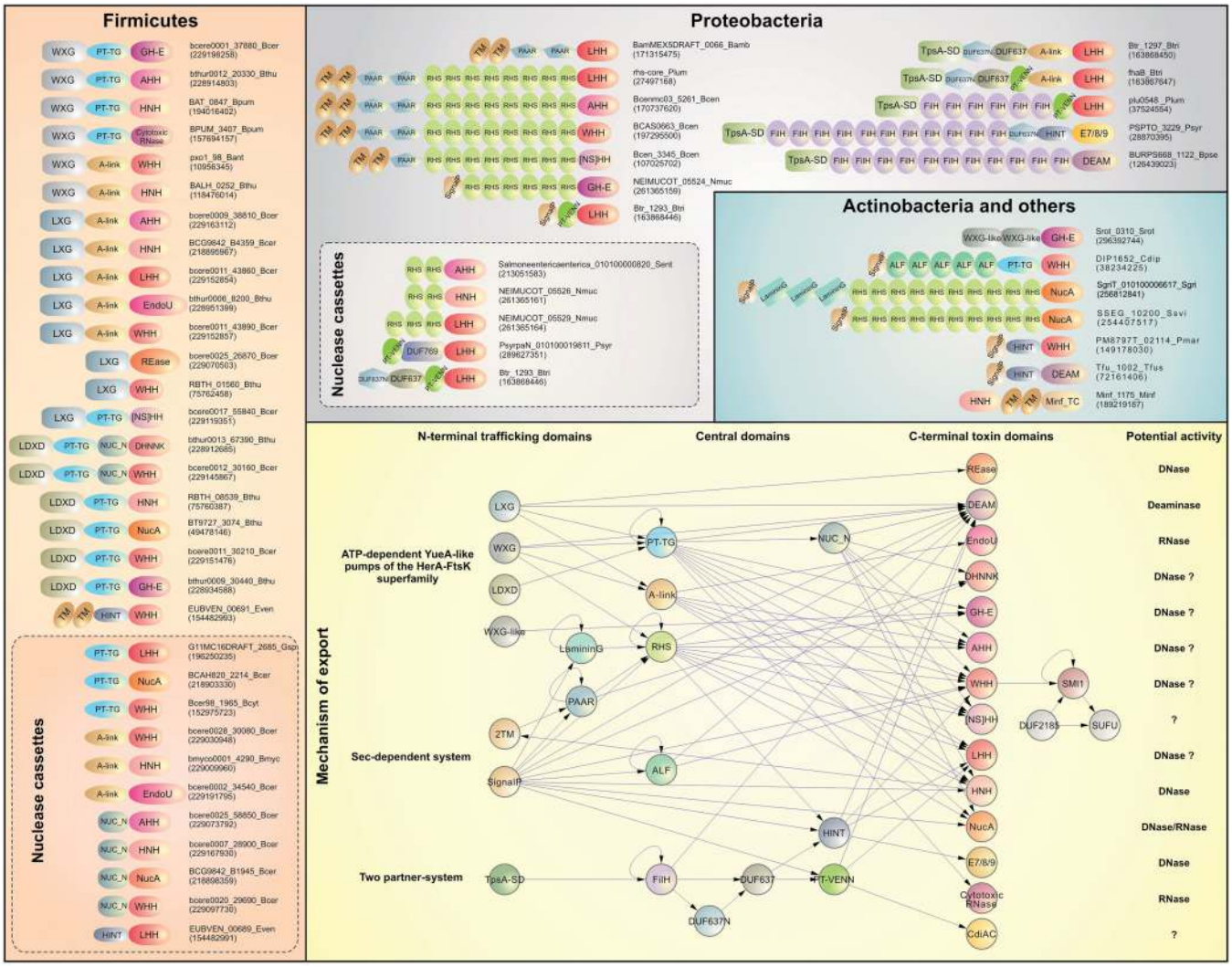

Figure 5. Domain architectures of selected examples of nuclease toxins encoded in the neighborhood of the SUKH superfamily genes. A domain architecture network of these toxins is shown to illustrate the directionality and syntactical features of their organization. Arrows indicate the polarity of domain arrangement in a polypeptide with the arrowhead pointing to the C-terminus. Newly identified domains include DUF637-N, A-link (α-helical PT domain), WXG-like, LDXD, NUC_N, PT-TGE which are non-catalytic domains, and AHH, LHH, WHH, DHNNK, GH-E, EndoU, REase, [NS]HH, DEAM, which are toxin domains. The CdiAC domain is a predicted nucleic acid modifying domain that is present in the C-termini of CdiA proteins of *Photorhabdus* and *E. coli*.

of proteobacterial CDI systems, none of these have a SUKH superfamily immunity protein. However, we found several non-CDI gene neighborhoods, which are likely to define distinct but analogous toxin systems, in proteobacteria, firmicutes, actinobacteria, synergistetes and bacteroidetes that combine genes for WHH and SUKH domain proteins (Figure 2). The AHH domain is also found in similarly organized gene-neighborhoods from the same bacterial lineages as those in which the WHH and LHH domains are found. Profile–profile comparisons with multiple alignments of all these three novel domains indicated that the best matches are families of the HNH fold. Indeed, a visual examination of the conservation patterns of these three domains showed that the HH dyad shared by them corresponds to the HH or DH dipeptide found in the first strand of treble-clef fold of the classical HNH domains (52). The first H forms one of the catalytic metal-chelating ligands and the second H contributes to the active site that directs the water for phosphoester hydrolysis (58). Further, the sequence alignments of the LHH, WHH and AHH motifs revealed two further conserved histidines, which were associated with the helix of the treble-clef fold and aligned with the two C-terminal metal-chelating residues in the profile of the classical HNH domains (58,59). These observations indicated that the LHH, WHH and AHH domains are highly derived versions of the HNH fold. The eighth family of HNH fold enzymes emerging from this analysis comprises of proteins typified by the protein Dd586_1447 (gi: 271499995) from *D. dadanti* found in predicted toxins in SUKH neighborhoods and also in CDI operon products which do not contain a SUKH-type immunity protein (Figure 2). A subset of these domains constitutes the PFAM model for a

'domain of unknown function', DUF1994, that does not define the boundaries of this domain precisely. We were able to define the proper boundaries of this domain by using the diversity of distinct architectural contexts in which we detected it and used the refined alignment for profile–profile comparisons. This comparison revealed the representatives of HNH domains as the best hits and indicated a perfect match between the polar residues conserved in this domain and catalytic and active-site metal chelating residues of the classical HNH domains. We named this family of HNH domains as DH-NNK after the conserved DH dyad in the strand-1 and the two asparagines and lysine which are conserved in the helix of the core treble-clef fold (Figure 3). While all these above versions have lost the cysteines of the ancestral treble-clef, they nevertheless, retain the catalytic configuration typical of those nucleases. Hence, we predict that these domains are likely to be nucleases with a similar catalytic mechanism. Practically all characterized HNH fold nucleases, barring those of the NucA family, which show a distinct active metal chelating site (51), have a preference for DNA substrates. Hence, it is likely that most of these domains are the active components of toxins that hydrolyze DNA in the target cells.

Nuclease toxins of the EndoU fold. The EndoU nuclease domain is typified by the nuclease domain previously identified in the U-specific, metal-dependent endonuclease, which in eukaryotes processes intron-encoded U16 and U86 snoRNAs and generates products with 2′–3′ cyclic phosphate and 5′-OH termini (60). A related endonuclease was identified in nidoviruses, such as the severe acute respiratory syndrome coronavirus where it appears to process RNAs as a part of the replication complex (60,61). Our structural analysis revealed that the catalytic domain of these enzymes contains two elements each comprised of a single helix followed by a three-stranded unit. This suggests that it is likely to have emerged through duplication of the simple helix-three-strand structural element, followed by flipping of the sheet in one of the units (Figure 4A). The catalytic residues, i.e. two histidines, appear to have emerged asymmetrically in a peculiar hairpin insertion within the helix of the first repeat. This hairpin insertion appears to be mobile and adopts different conformations in structures (60,61)—this mobility might have a role in accommodating the substrate between the helix and the sheets formed by the three-stranded units of the repeats (Figure 4A). We found that the bacterial members of the EndoU family are linked to genes of the SUKH superfamily mainly in firmicutes and proteobacteria (Figure 2). Other than SUKH superfamily gene-neighborhoods, related versions also comprise the polymorphic C-terminal domain of the CDI toxins from *Moraxella* and *Mannheimia* that, however, lack a SUKH superfamily immunity gene. A further set of bacterial nucleases of this family are predicted secreted versions encoded by intracellular symbiotic and pathogenic bacteria, such as *Wolbachia* (gi: 310643370) and *Ehrlichia* (gi: 73666818). Most bacterial versions that we identified are extremely divergent relative to the eukaryotic and viral forms and are not recognized by the previously available HMM models for this nuclease (PF09412). Hence, the identification of these relationships represents a significant extension of this superfamily (Figure 4A, Supplementary Data). Versions within these gene-neighborhoods show considerable variability including loss of strands from the first unit. This variability suggests that the EndoU fold is rather flexible to accommodating drastic modification, which in turn might help it recognize a diverse spectrum of substrates. On the precedence of the eukaryotic EndoU and the nidoviral nuclease and their genomic organization we suggest that the majority of the bacterial EndoU homologs are nuclease toxins that cleave RNAs in the competitor cells. Those secreted by intracellular bacteria could be deployed as toxins or regulators to manipulate host physiology by cleaving specific transcripts. With the identification of these new EndoU homologs it becomes clear that the bacteria contain the greatest diversity of this superfamily, with certain versions closer to the eukaryotic and nidoviral versions and others that are more divergent (Supplementary Data). This suggests that the original radiation of this superfamily probably happened within the bacterial toxin systems and were subsequently acquired, perhaps from intracellular symbiotic bacteria, by eukaryotes and viruses. In the latter they appear to have been recruited as RNA processing enzymes.

Nuclease toxins of the REase fold. The REase fold is a highly versatile fold that accommodates considerable structural diversity and has, not surprisingly, been used as the primary fold from which REases of restriction-modification systems are derived (48,52). We also found several proteins with this fold to be encoded by genes that are neighbors of SUKH superfamily genes (Figure 2). These versions were originally identified as a distinct conserved domain of unclear affinities—both PSI-BLAST and HMMER searches failed to identify any relationships with previously known domain. However, we observed that the multiple sequence alignment of this domain showed a characteristic signature of conserved residues of the form GE-D-ExK-Q (Figure 4B) that matched the pattern of similar conserved residues in the lambda exonuclease and the RecB family of the REase fold (52,62). The predicted secondary structure pattern of these domains also closely matched the REase fold with conserved D and ExK motif falling on a β-hairpin as is typical of the REase fold (Figure 4B). These observations induced us to use the alignment of this domain in a profile–profile comparison with the HHpred program, and we recovered a composite profile made of diverse REase fold superfamilies such as the VRR-Nuc, lambda exonuclease, the archaeal Holliday junction resolvase and RecB as the best hits ($P = 10^{-5}$). This suggested that this family defines a novel group of REase-fold nucleases. Given that the majority of the REase-fold enzymes are DNases, we predict that these toxins are likely to cleave the DNA of the target cells.

Nuclease toxins of the cytotoxic RNase fold. The last family of nucleases that we found encoded by genes linked to the SUKH superfamily genes was the cytotoxic

RNase family (63). This nuclease domain was first characterized as the toxin domain of the colicins E3 and E6 and is typified by a conserved active site configuration with an aspartate followed by a glutamate sandwiched between two histidines (Supplementary Data). The version of this domain in colicin E3 has been demonstrated to function as an EndoRNAse that specifically cleaves the phosphoester bond between bases 1493 and 1494 of 16S ribosomal RNA (63). Given that versions detected in systems characterized in our current study are closely related to the version found in colicin E3 and E6, we posit that these nuclease domains act as RNAses that similarly cleave RNA in the target cells.

Other domains with a possible role in nucleic acid modifications. We found three other families of domains in proteins that were encoded by genes which occupied positions adjacent to SUKH superfamily genes in certain predicted operons, equivalent to positions of the genes encoding the above nucleases. Additionally, these families of domains are also found as representatives of the polymorphic C-terminal module of the proteobacterial CDIs. Together these observations hinted that they are potentially uncharacterized enzymatic domains operating on nucleic acids. PSI-BLAST and JACKHMMER searches showed that the first of these families belonged to the nucleotide deaminase superfamily that includes RNA-editing enzymes, such as the APOBECs and DNA-modifying enzyme AID of vertebrates. Hence, like the nucleases, these enzymes are likely to function as toxins that mutate nucleic acids in the target cells. We discuss the natural history of these enzymes in a separate article (Iyer LM, Zhang D, Aravind L, manuscript in preparation). The second of these families prototyped by the *B. cereus* protein bcere0017_55840 (gi: 229119351) is characterized by a conserved signature [NS]HH followed by another conserved histidine (Supplementary Data). Although we were unable to unify this family with any of the other nuclease folds, the presence of the HH motif typical of many of the above families of HNH/EndoVII fold nucleases might point to a divergent relationship with those proteins. The third of these families, typified by the CDI system from *P. luminiscens* (gi: 37524545) includes a globular domain of 170–200 amino acids that might define yet another uncharacterized nucleic acid-modifying domain (CdiAC in Figure 5, Supplementary Data).

Identification of conserved domains with potential roles as trafficking components and auxiliary partner proteins of the SUKH superfamily-toxin systems

Earlier characterized toxin systems such as the classical plasmid-encoded bacteriocins and the recently characterized CDI systems use thematically comparable, albeit biochemically distinct mechanisms for trafficking of nuclease toxins. While these systems have been used as models to understand bacterial protein trafficking, the complete set of events starting from the extrusion of the 'pro-toxin' by the producing cell to its recognition at the target cell surface and delivery into the target cell are only partially understood (8). Classical plasmid-borne colicins and cognate bacteriocins from other bacteria do not have secretory mechanisms and their release appears to occur primarily through cell-lysis mediated by the colicin-release proteins (8). Colicin-like bacteriocins are multidomain proteins with an extreme C-terminal toxin module, which is either a nuclease or a membrane-perforating domain (e.g. colicin E1 and A) (8). They typically possess two additional N-terminal modules, of which the first facilitates translocation across the target cell membrane and the second (i.e. the central module) facilitates binding to a membrane receptor on the target cell. These colicins hijack either the Tol or the Ton-dependent molecular import systems to enter the target cells (8). The chromosomally encoded proteobacterial CDI system toxins do not require lysis; instead they are trafficked out of the cell which produces them via the two-partner-system that depends on the CdiB proteins belonging to the TpsB class of outer-membrane trafficking proteins (25). These latter proteins contain N-terminal periplasmic polypeptide-transport-associated (POTRA) domains linked to a C-terminal β-barrel transmembrane domain. They recognize the secretory domains such as the TpsA-SD in the extreme N-terminal region of the CDI 'pro-toxins' to deliver them across the outer membrane of proteobacteria (64). This N-terminal region is separated from the C-terminal regions by repetitive regions with RHS- or filamentous hemagglutinin-type repeats. Their uptake by the target cell is less-clearly understood. In the well-studied examples, the first step of this process appears to depend on the outer membrane-biogenesis protein BamA recognizing a conserved α-helical domain immediately N-terminal to the toxin module, with a VENN signature that overlaps with the PFAM model termed 'DUF638′. Subsequently the inner-membrane protein AcrB, a transporter, appears to be necessary for uptake into the target cell cytoplasm (25). Additionally, it is posited that a proteolysis step at the cell surface releases just the C-terminal nuclease module for uptake by the target cell (25). Thus, despite the differences between the CDI and classical colicin-like systems they share a common feature of the toxin activity being borne by the extreme C-terminal domain in a multidomain polypeptide. Further, the modules located immediately-N-terminal to the nuclease domain (e.g. the α-helical domain with the VENN motif ∼PFAM DUF638) are involved in association with receptors on the target cell. Hence, we term these domains collectively the pre-toxin (PT) domains. The extreme N-terminal domains appear to play a critical role in export from the host cell in the cases where lysis is not involved, i.e. typically chromosomally borne versions. These observations accordingly presented the organizational logic for these systems, wherein there are usually three functionally distinct sets of modules in the pro-toxin going from the N- to the C- terminus of the protein.

Analysis of the domain architectures of the nuclease domain-containing proteins encoded in the SUKH-superfamily neighborhoods revealed that the majority of the proteins followed an architectural logic which was consistent with the above-described organization of these

earlier studied toxin systems (Figure 5). However, only a relatively small subset of the SUKH domain-associated systems overlaps with the CDI systems. Further the SUKH superfamily proteins and functionally linked toxins are also found outside of proteobacteria, in lineages lacking outer membranes and CdiB-like delivery systems. We reasoned that analysis of these distinct pre-nuclease and extreme N-terminal domains might reveal features pertaining to the trafficking of toxins in non-CDI systems and point to alternative delivery mechanisms.

Identification of multiple distinct trafficking systems for toxins encoded in SUKH superfamily neighborhoods. We observed that in Gram-positive bacteria, proteins with the C-terminal nuclease typically possessed one of a set of several distinct domains at the extreme N-terminus of the protein (Figure 5). A significant subset of these could be unified using sequence profile searches with the PSI-BLAST and JACKHMMER programs to the WXG/ESAT6 superfamily of α-helical domains (65). These domains are a specific signal recognized by the YueA-like ATPases of the HerA-FtsK superfamily that secrete them in an ATP-dependent manner (65,66). This indicated that the WXG/ESAT-6 domain-containing toxins in Gram-positive bacteria are extruded by YueA-like pumps using an ATP-dependent mechanism. A significant subset of toxin proteins from firmicutes possessed a distinctive N-terminal domain that could not be unified with any earlier known domain (a subset of these have been included in the erroneously annotated model Transposase_30 of PFAM; PF04740). Sequence searches showed that this domain possessed a conserved [LF]XG sequence motif and it was predicted to assume an α-helical bundle fold based on the multiple sequence alignment (Supplementary Data). We accordingly termed it the LXG domain (Figure 5) and were able to unify it with the WXG domain by means of profile–profile comparisons with the HHpred program ($P = 10^{-7}$). Contextual analysis indicated that this domain is encoded by certain conserved gene-neighborhood across firmicutes, where it is associated with genes coding for a YueA-like HerA-FtsK superfamily protein pump and a small protein related to the *S. aureus* EsaC protein (gi: 282917938, Supplementary Data). Through profile–profile comparisons we showed that the EsaC-like superfamily is a bacterial version of the eukaryotic EVH1 peptide-binding domains with the PH-like fold (HHpred *P*-value: 10^{-4}) (67). These observations suggest that the LXG domain is comparable to the WXG/ESAT-6 domain, and is likely to utilize the ATP-dependent YueA pumps and the potential peptide-binding EsaC domain as partners for extrusion from the producing cell. The protein Srot_0310 (gi: 296392744) from the actinobacterium *S. rotundus* contains two copies of a distinct domain N-terminal to the GH-E nuclease domain (Figure 5). This domain is also widely found in several actinobacteria at the N-termini of putative cell-surface proteins. Profile–profile comparisons suggested a possible relationship between these N-terminal domains and the WXG domain suggesting that it might be yet another representative of the WXG-like superfamily ($P = 10^{-4}$) and might utilize a similar ATP-dependent mechanism for its extrusion. A fourth group of proteins, restricted to certain firmicutes (e.g. *S. aureus* SACOL0281 protein; gi: 57652555), is typified by yet another N-terminal α-helical domain (LDXD in Figure 5) that is also found in domain architectural contexts very similar to the WXG and LXG domains. It is conceivable that this domain is comparable to them and functions similarly as a mediator of export via the HerA-FtsK superfamily pumps. Thus, a notable mode of export of nuclease toxins in Gram-positive bacteria appears to be via the ATP-dependent extrusion system, which while biochemically distinct from the TPS of the proteobacteria, is thematically comparable.

In Actinobacteria, but not firmicutes, we observed several large proteins with architectures similar to the CDIs of the proteobacteria. These typically contain RHS repeats; however, their extreme N-terminal domains did not bear any close relationship to the proteobacterial TpsA-SD. Instead they were found to contain an N-terminal signal peptide and some of these proteins (e.g. gi: 256812841, a protein from *S. griseus*) contain multiple laminin G domains embedded within repetitive regions. The protein DIP1652 (gi: 38234225) from *C. diphtheria* shows another distinct low complexity repeat N-terminal to the nuclease domain (Figure 5) and like in the above case it also possesses a conventional signal peptide. Likewise, a distinctive signal peptide, which is highly conserved in multiple proteins only within the genus *Planctomyces,* is seen in predicted nuclease toxins from this organism (e.g. gi: 149178028). Another group of large toxin proteins with RHS repeats, which predominantly occur in proteobacteria, are defined by the presence of repeats of the PAAR domain (PFAM: PF05488) N-terminal to the RHS repeats. All these proteins are typified by the presence of a conserved transmembrane domain with two TM segments (Figure 5 and Supplementary Data) just N-terminal to the PAAR domains. We propose that these TM segments are required for their trafficking to the cell membrane, following which they might be processed in the periplasm for release via the outer membrane in a process that might depend on the PAAR domains. We also noticed a comparable domain with two TM segments in few firmicutes (e.g. gi: 125974537 from *C. thermocellum*) and in chlamydiae (e.g. 189219187 from *M. infernorum*, which is a rare case of the nuclease domain occurring N-terminal to the two TM domain; Figure 5). These proteins lack PAAR domains but the firmicute versions have additional hedgehog-intein (HINT) peptidase domains (see below) that could aid in their release on the cell-surface (Figure 5). These observations suggest that at least some nuclease toxins in bacterial lineages such as actinobacteria, bacteroidetes and planctomycetes with conventional signal peptides, and those in proteobacteria, chlamydiae and firmicutes with two-TM domains are probably delivered to the cell using the conventional Sec-dependent system (68). In the context of the above cases, it is of interest to note that *E. coli* Syd, an archetypal member of the SUKH superfamily, was first

identified as a possible proof-reading component of the Sec-dependent export system (43–45). In this context it is possible that the binding of certain members of the SUKH superfamily (at least the Syd-like group) in the producing cell might not only help in conferring immunity to 'self' but also in guiding the 'pro-toxin' to the Sec-dependent export machinery.

Both actinomycetes and firmicutes do not display proteins with a PT domain with the VENN motif (PT-VENN). However, we observed that in both these lineages there was a conserved α-helical domain that frequently occurred just to the N-terminus of several distinct nuclease modules in different predicted toxins. This domain had a conserved TG motif and we accordingly named it the PT-TG domain (Figure 5, Supplementary Data). The PT-TG domain might play a role similar to the PT-VENN domain in Gram-positive bacteria and mediate interaction of the extruded toxin with cell-surface receptors on the target cells. The complementary distribution of the PT-VENN and PT-TG domains in proteobacteria and Gram-positive bacteria suggests that they are distinct adaptations related to the drastically different cell-surface morphologies of the respective groups. Another domain, which we found frequently associated with several unrelated or distantly related nuclease domains from Gram-positive bacteria, was the Nuclease_N domain (Figure 5, Supplementary Data). It is predicted to be an α-helical domain and might also play a role in the delivery of the toxin module into the host cells. Toxins in the SUKH superfamily neighborhoods, irrespective of the type of the nuclease domain, can also be distinguished into two major architectural groups: one comprised of relatively small proteins with no notable stretches of repetitive sequence separating the N- from the C-terminal regions, and the second in which such repetitive sequences, such as the RHS and the filamentous hemagglutinin are present (Figure 5). This might reflect a mechanistic difference in their mode of action: the smaller proteins could be soluble toxins that diffuse away from the cell producing it. In contrast, the large proteins with repetitive elements might form filamentous appendages that stick out from the cell-surface and depend primarily on contact with target cells for delivery [Hence, the latter group includes the recently characterized CDIs (25)]. Alternatively, this difference might reflect the differences in the cell-wall structures of the bacterial lineages, with the smaller toxin proteins being more prevalent in the firmicutes. A subset of the smaller proteins with nuclease domains lack noticeable trafficking-related (N-terminal) domains. The corresponding genes could represent cassettes for alternative toxin modules that are linked by recombination to the larger full-length genes (Figure 5, see below).

Other auxiliary domains which might play a role in resistance, trafficking or processing of toxins. Several other domain families were found to be encoded by genes having persistent association with the SUKH superfamily neighborhoods across distantly related bacterial species. One of these is the SuFu superfamily (Figure 2 and Supplementary Data) prototyped by the Suppressor of Fused protein from *Drosophila* (69). In addition, we also detected members of this superfamily to be encoded by CDI-like operons, such as the one from *N. gonorrhoeae* that encodes a toxin with a distinct version of the HNH fold nuclease domain (toxin NGO1392, gi: 59801740; Supplementary Data). In these cases the SuFu superfamily gene occupies a position equivalent to that of the SUKH superfamily gene, suggesting that they might be functionally comparable. We also found several examples wherein the SuFu and SUKH domains are combined in the same polypeptide (Figures 1 and 5). Based on these associations we propose that the SuFu domain represents a second widely conserved domain that function as an immunity protein for diverse nuclease toxins. Two other conserved protein families are encoded in the toxin neighborhoods (SUKH-neighborhood conserved family 1 and 2; SNCF1 and SNCF2, Supplementary Data) that occupy positions similar to the SUKH and SuFu superfamily genes (Figure 2). They were not found in multi-domain architectures typical of the nuclease toxins and always occurred as proteins with standalone domains. This suggested that they were unlikely to be novel toxins but act as alternative immunity proteins just like the SuFu and SUKH superfamily proteins. The HINT domain, prototyped by the peptidase domains of the animal hedgehog proteins and protein-splicing inteins, is also frequently associated with SUKH superfamily neighborhoods (70–72). These versions of the HINT domain are closer to those found in several bacterial surface proteins and the secreted animal proteins such as hedgehog and the *C. elegans* Hog proteins (70). When present in a multidomain 'pro-toxin' protein, the HINT domain always occurs sandwiched between the PT domains such as PT-VENN and PT-TG and the nuclease toxin domain. This location of the HINT domain suggests that it is likely to serve as a peptidase that undergoes autoproteolytic cleavage, similar to what is observed in hedgehog and the inteins (70), to release the C-terminal nuclease domain for uptake by the target cell. It is conceivable that this cleavage step is regulated by the interaction of the PT domains with the surface receptor on the target cell.

Eukaryotic/DNA viral members and structure–function analysis of the SUKH superfamily

While SUKH superfamily neighborhoods are very widespread in bacteria, they are largely absent in archaea. Although we uncovered potential extruded nuclease toxins in certain halophilic archaea such as *H. borinquense* (gi: 312291883, with a GH-E nuclease domain), which are delivered by means of a distinctive N-terminal metallopeptidase domain, we did not find any immunity proteins of the SUKH or SuFu superfamilies. Although the exact reason for this exclusion is unclear, it is conceivable that these immunity proteins are ineffective in the context of the distinct archaeal secretory systems. However, several eukaryotes possess one or more SUKH superfamily members. Phylogenetic analysis and phyletic patterns suggest that there are two major eukaryotic lineages of the SUKH superfamily that are nested within the radiation of the bacterial versions (Supplementary

Data). They are respectively prototyped by the polyglutamylase subunit PGs2 (22), and the vertebrate SCF ubiquitin E3 ligase subunit FBXO3 with yeast Smi1/Knr4 (21,73). The PGs2 version is found in basal eukaryotes such as *Giardia* and *Spironucleus*, animals and chlorophyte algae suggesting that it was likely to have been acquired prior to the last eukaryotic common ancestor (LECA) and subsequently lost in several lineages. The FBXO3 lineage is present in animals, fungi, plants, stramenopiles and ciliates. However, it does not group with the PGs2 lineage, instead grouping with other bacterial forms. Hence, it was probably acquired relatively early in eukaryotic evolution via an independent transfer from bacteria. In both plants and animals the FBXO3 version is fused to an N-terminal F-box domain and a distinctive C-terminal immunoglobulin superfamily domain (overlaps with the PFAM model DUF525), suggesting that it was recruited as an E3 subunit prior to the radiation of these eukaryotic groups. In addition to these versions, there appear to have been other sporadic transfers of SUKH superfamily members to eukaryotes. For example, land plants contain a version typified by the *Arabidopsis* protein At3g50340 (gi: 15229727) which seems to have been independently acquired by them from a bacterial source. Another sporadic transfer is seen in certain filamentous fungi, which acquired a version of the SUKH-4 group that has been independently fused to an N-terminal F-box domain (e.g. *A. oryzae* gi: 169782758). DNA viral versions show no specific relationship with eukaryotic forms; instead, they share specific sequence motifs with the SUKH-3 group, recover them as best hits in profile–profile comparisons, and group with them in the phylogenetic tree (Supplementary Data). Within viruses they are most widespread and abundant in herpesviruses, with the versions from adenoviruses, poxviruses and iridoviruses being nested within the herpesviral radiation of the family (Supplementary Data). Thus, they appear to have been acquired first by an ancestral herpesvirus, similar to that inserted in the amphioxus genome (46), from a bacterial source and subsequently disseminated across diverse DNA viruses.

Although there has been gene loss in several eukaryotic lineages, at least the two ancient versions, namely PGs2 and FBXO3 appear to have been largely vertically inherited and show no lineage-specific expansions within eukaryotes. This is in sharp contrast to the high propensity for lateral transfer and for lineage-specific expansions of the SUKH superfamily that is observed in bacteria. This feature, together with the available functional evidence suggests that these conserved eukaryotic versions have acquired a biological role distinct from that in the toxin–immunity systems of bacteria. Nevertheless, there were several features that suggested to us that biochemically the eukaryotic versions might be exploiting an ancient functional template provided by the SUKH domains in bacterial nuclease toxin systems. Firstly, the studies on yeast Smi1/Knr4 have shown that it interacts with a large number of structurally and functionally distinct proteins (19). In FBXO3, and independently in the above-mentioned fungal proteins, it appears in a domain architectural context corresponding to the part of the E3 F-box subunit that recognizes the substrate for ubiquitination (74). This suggests that it might be deployed as a recognition domain to recruit particular substrates for ubiquitination. In bacteria the SUKH superfamily domains are one of the most widespread immunity proteins that appear to function in conjunction with a repertoire of nuclease toxins that are extremely diverse in sequence and structure (Figures 3 and 4). Taken as a whole, these observations indicate that the SUKH domain contains a scaffold that has been adapted to recognize a diverse set of protein partners.

A possible clue for the structural basis of this capability is offered by studies on the *E. coli* Syd protein: it has been shown to contain a prominent negatively charged cleft with which it could interact with partner proteins (45). Examination of the structure of this protein indicates that this cleft is formed by the space between the conserved helix H3 and the fissure in sheet between the two-stranded N-terminal unit and the C-terminal 4-stranded meander (Figure 1). Given that this unusual feature is seen across the fold, we examined the surface renderings of different SUKH superfamily members and a corresponding cleft is observed in most of them (Figure 1). Although this cleft is not necessarily negatively charged as in Syd, and might vary in depth and shape, its widespread presence suggests that it might be the means by which the SUKH superfamily is able to accommodate different protein partners. In support of this hypothesis we observed that in the case of two distantly related members of the SUKH superfamily, namely Syd (PDB: 3ffv) and YobK (PDB: 2prv), this cleft is used in protein–protein interactions. In both these crystal structures one of the monomers is bound in the cleft of the other monomer resulting in an asymmetric dimer (Supplementary Data). These dimers are unlikely to represent biologically native dimeric states, but in any case illustrate the ability of the conserved cleft of the SUKH fold to accommodate other proteins. Interestingly, the SuFu superfamily also shows a comparable kind of sheet with a fissure between two sets of strands (69). Experimental studies on the *Drosophila* SuFu shows that it also functions as protein tether which holds the Zn-finger transcription factor Gli in the cytoplasm in the absence of the hedgehog signal (75). In vertebrates the SuFu ortholog has been shown to bind Gli2 and Gli3 and prevent their degradation due to ubiquitination by F-box E3 ligases (76). Thus, the presence of comparable binding interfaces that have the flexibility to recognize a wide range of protein ligands might be a common feature shared by both the SUKH and the SuFu superfamilies of immunity proteins. It is this feature that appears to have resulted in them being utilized as adaptors for recruiting other proteins in eukaryotic regulatory systems.

The extensive spread of the US22 group of the SUKH superfamily across unrelated or distantly related DNA viruses of animals suggests that it confers an important advantage to these viruses. This is also supported by the lineage-specific expansion in betaherpesviruses of the SUKH superfamily in the form of multigene arrays

similar to what is seen in bacteria (Figure 2). Indeed multiple studies suggest that distinct copies of the proteins in herpesviruses are required for effective survival and replication of the virus in their hosts. For instance, mutagenesis of two SUKH superfamily paralogs M142 and M143 in the murine cytomegalovirus was shown to be essential for survival of the virus itself, whereas mutagenesis of other paralogs M139, M140 and M141 specifically prevents its replication in macrophages (77). Other studies indicated that M142 and M143 form a heterotetrameric complex which counters the action of the host protein kinase R (PKR) in shutting down viral protein synthesis (78–81). The human cytomegalovirus SUKH superfamily proteins TRS1 and IRS1 have been shown to similarly counter the PKR and the dsRNA dependent arm of the anti-viral response (78,82–86). Another paralog UL38 inhibits the host cell stress responses by antagonizing the tuberous sclerosis protein complex in the endoplasmic reticulum (87,88) and counters apoptosis in conjunction with yet another paralog UL36 (89,90). In light of these observations it appears that the viral versions of the SUKH superfamily are deployed to counter different facets of the host anti-viral and stress response. By analogy to the bacterial versions, which function as immunity proteins, we propose that the viral SUKH domain proteins in general bind diverse host proteins that are used against the virus. Here again the special ability of the SUKH scaffold to bind diverse proteins appears to have been exploited by the virus as a flexible binding interface to neutralize a diverse group of host anti-viral defenses.

Evolutionary implications and general considerations

Identification of the SUKH superfamily and associated nucleic acid modifying toxin systems has considerable implications for understanding bacterial genetic conflicts, evolutionary forces acting on strongly linked multi-gene loci, and potential biotechnological applications. We briefly discuss some of these implications that emerge directly from our observations.

Relationship of toxin systems to genetic conflicts in the bacterial world. Classical colicins and earlier characterized CDIs act primarily on related bacterial strains of the same 'species'. Although the systems identified in our studies are abundantly represented in extracellular pathogenic bacteria, they are rare in intracellular symbionts or pathogens. This might be because intracellular bacteria are much less likely to encounter a heavy load of competing cells in the same niche. The bacterial toxin systems which we uncovered in this study and the related CDIs are also different in certain features from the classical colicin-like systems. Classical colicins are in large part encoded on plasmids, which might be either single copy, medium-sized conjugative plasmids or small multi-copy small plasmids that depend on the conjugative plasmids for their transmission (8). Such bacteriocins are relatively rare on chromosomes. In contrast, 99.25% of the systems recovered in our study are chromosomally encoded. Majority of the plasmid-encoded classical colicin-like toxins are accompanied by a gene encoding a lysis protein and their release is concomitant with the lysis of the host cell. However, none of the systems identified in this study or the CDIs have lysis genes in their neighborhoods (25). This difference suggests that, while both the plasmid-borne bacteriocins and these systems might be directed at close relatives, they appear to be geared toward distinct genetic conflicts. The lysis of the cell nullifies the fitness of the chromosome; hence, it would be largely deleterious for the chromosome to encode systems that require lysis. The plasmid being a selfish element is not completely affected by loss of fitness of the host as long as it can offset it by holding on to, or spreading in the host population (i.e. the plasmid's own fitness is enhanced or maintained). Cells of the host type without the bacteriocinogenic plasmid are competitors that affect the plasmid fitness, especially under stationary phase or starvation conditions. Hence, the plasmid-borne colicin would be primarily selected to act against host cells that have lost the plasmid or lack it by default under these stress conditions. Further, the plasmid toxins are unlikely to have ready access to trafficking by the host because, given the large amounts in which the colicins are produced (8), their export is likely to impair host fitness. Further, it has been shown that under starvation only ~3% of the cells produce colicin (91). Although the loss of the cells producing the colicin would endanger the resident plasmid, a relatively small fraction of the host population is affected. By the principle of inclusive fitness of kin (92), the plasmid could still have an enhancement of fitness from the copies in the surviving cell along with the elimination of competitors by the released toxin. On the other hand, the toxin domains of many of the chromosomal versions like the CDIs and those identified in this study appear to be borne on filamentous structures that are primarily geared toward to elimination of competitors that come in physical contact with the cell-surface (25,93). Therefore, these systems are likely to be critical in the context of the formation and organization of biofilms and solid substrate colonies. When bacterial cells are aggregating in the above contexts it would benefit to eliminate resource sharing with non-kin competitors. Hence, presence of a chromosomally encoded toxin that acts at a short range is likely to be selected, resulting in the proliferation of systems such as those described here. Nevertheless, it would also benefit 'cheater cells' to evade such defensive mechanisms. Hence, they would be selected to maintain a wide diversity of immunity proteins to counter different non-self toxins, which might explain the arrays of diverse SUKH genes in several bacterial genomes.

Potential evolutionary processes in diversification of toxins and immunity proteins. Imprints of the evolutionary arms race arising from the above processes are readily observed in our systems. The toxin proteins appear to show a rather peculiar pattern of diversification. The N-termini, which are typically associated with trafficking, tend to be relatively conserved while C-terminal nuclease domains show major diversity (Figure 5). This is consistent with a recent study on the diversification of RHS proteins in

enterobacteria which showed that the RHS proteins undergo C-terminal polymorphism due to rampant recombination with invading cassettes that encode alternative C-terminal modules (94). This type of recombination or gene-conversion with polymorphic C-terminal cassettes might explain the presence of smaller loci found in the gene-neighborhoods characterized here that encode just a nuclease domain by itself or with an additional small N-terminal extension (Figures 2 and 5). Hence, we extend the original proposal for RHS diversification to suggest that, more generally, recombination with cassettes with distinct C-terminal modules is the primary proximal mechanism for diversification of the toxin proteins across all bacterial lineages (Figure 5). Furthermore, the presence of nuclease and nucleic acid deaminase domains as the primary toxin modules of these systems raises the possibility that their nucleic acid cleaving or mutating activity is involved in triggering recombination events. This appears plausible given the observations that most of these nucleases are likely to be endonucleases, which like their counterparts in the restriction-modification systems could cleave at specific sequences. Similarly, deaminase-induced mutations have been implicated in the triggering of class-switching recombination events in vertebrates (95). More generally, this ties in with earlier studies which have demonstrated the role for both recombination and positive selection in the evolution of plasmid-borne bacteriocins (96). It has been proposed that pore-forming versions have predominantly utilized recombination for diversification whereas nucleases have mainly evolved through positive selection. In our systems, the evidence points to both these forces being active at different levels in the evolution of the toxin proteins (96). While the basic architectures evolve through recombination generating C-terminal polymorphism, the C-terminal nucleases themselves show evidence for considerable sequence diversification within each family. Indeed, much of the diversification of the HNH/EndoVII fold appears to have happened within the context of these systems, with several structurally distinct forms evolving amidst the nuclease toxins (Figure 3).

Phyletic and phylogenetic analysis of the SUKH superfamily indicates three salient features, namely rampant lateral transfer between different branches of the bacterial tree, gene loss and lineage-specific expansion followed by divergence of the lineage-specific paralogs (Supplementary Data). This suggests that there is a notable trend for maintaining diversity within the SUKH superfamily that probably arises from selection for recognition of a diverse range of nucleic acid-modifying toxins. Although there are multiple distinct types of immunity proteins known from plasmid-borne bacteriocins and CDI systems, most show very limited phyletic patterns. For example the CdiI toxin seen in several CDI systems is entirely limited to proteobacteria (25). We observed that it is a protein with two TM segments that is likely to form a membrane channel (Supplementary Data) and have a mode of action very distinct from the SUKH superfamily. As only the SUKH superfamily and, to certain extent, the SuFu superfamily show a pattern of wide dissemination across bacteria it is likely that only these scaffolds can support sufficient diversification that goes hand in hand with the polymorphism of the toxin domains.

Implications for eukaryotic and viral functions. Our observations also suggest that the biochemical diversity generated within these bacterial toxin systems has been taken up and utilized for very different functions by eukaryotes and their viruses. Both the SUKH and the SuFu superfamily domains have been utilized as adaptors that regulate recognition of different substrates by protein modification systems such as ubiquitination and polyglutamylation. In a completely different context, the HINT domains derived from such bacterial toxin systems appear to have been used to release peptide messengers in animal signaling pathways, like the hedgehog pathway (70). The nuclease domains ultimately derived from various toxins also appear to have been used for different functions by eukaryotes and their viruses. The EndoU nuclease domain, which ultimately emerged from these toxin systems, has been recruited by the nidoviruses for the replication of their negative-strand RNA genome, whereas a related domain was recruited by eukaryotes for processing of certain snRNAs. We also observed that a HNH/EndoVII fold nuclease found in the bacterial toxin typified by the *N. gonorrhoea* protein NGO1392 is found in several eukaryotic lineages such as animals, plants, stramenopiles and apicomplexans (Supplementary Data). Given its conservation and relatively lower divergence, it is unlikely that the nuclease functions as a toxin in eukaryotes. However, it is possible that it has been recruited as a DNA-repair enzyme, as has been previously observed in the case of certain nucleases of bacterial restriction-modification and phage replication systems (97). In general terms, these observations suggest that the origin of key systems in eukaryotes, including those related to the emergence of certain lineages, such as animals (i.e. the hedgehog pathway), appear to have extensively benefited from the availability of 'pre-adaptations' in the form of components whose ultimate origins lay in these toxin systems.

CONCLUDING REMARKS

The current study points to the remarkable flexibility of SUKH domains in mediating different protein–protein interactions. In a sense, this situation resembles what has earlier been observed with certain scaffolds like the immunoglobulin domain and the leucine-rich repeats of various immunity-related proteins of eukaryotes (98,99). The ability of the SUKH scaffold to accommodate diverse binding partners makes it a potential candidate as a template for protein engineering to generate novel binding capabilities. Likewise, the C-terminal diversification of the toxin domain could also have biotechnological utility as a model for generating secreted proteins that differ extensively in a given module but retain a constant N-terminal part. We hope that this characterization of the SUKH superfamily and identification of the associated nuclease toxin families provides new leads for the future exploration of the manifold implications of the systems discussed here.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Institutes of Health Postdoctoral Visiting Fellowship; intramural funds of the National Library of Medicine at the National Institutes of Health, USA. Funding for open access charge: Intramural funds of the National Institutes of Health, USA.

Conflict of interest statement. None declared.

REFERENCES

1. Stirpe,F., Barbieri,L., Battelli,M.G., Soria,M. and Lappi,D.A. (1992) Ribosome-inactivating proteins from plants: present status and future prospects. *Biotechnology*, **10**, 405–412.
2. Endo,Y. and Tsurugi,K. (1986) Mechanism of action of ricin and related toxic lectins on eukaryotic ribosomes. *Nucleic Acids Symp. Ser*., 187–190.
3. Endo,Y., Huber,P.W. and Wool,I.G. (1983) The ribonuclease activity of the cytotoxin alpha-sarcin. The characteristics of the enzymatic activity of alpha-sarcin with ribosomes and ribonucleic acids as substrates. *J. Biol. Chem*., **258**, 2662–2667.
4. Dhananjaya,B.L. and D'souza,C.J. (2010) An overview on nucleases (DNase, RNase, and phosphodiesterase) in snake venoms. *Biochemistry*, **75**, 1–6.
5. Rosenberg,H.F. (2008) RNase a ribonucleases and host defense: an evolving story. *J. Leukoc. Biol*., **83**, 1079–1087.
6. Alouf,J.E. and Popoff,M.R. (2006) *The comprehensive sourcebook of bacterial protein toxins*, 3rd edn. Elsevier/Academic Press, Amsterdam/Boston.
7. Riley,M.A. (1998) Molecular mechanisms of bacteriocin evolution. *Annu. Rev. Genet*., **32**, 255–278.
8. Cascales,E., Buchanan,S.K., Duche,D., Kleanthous,C., Lloubes,R., Postle,K., Riley,M., Slatin,S. and Cavard,D. (2007) Colicin biology. *Microbiol. Mol. Biol. Rev*., **71**, 158–229.
9. Anantharaman,V. and Aravind,L. (2003) New connections in the prokaryotic toxin–antitoxin network: relationship with the eukaryotic nonsense-mediated RNA decay system. *Genome Biol*., **4**, R81.
10. Kobayashi,I. (2001) Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res*., **29**, 3742–3756.
11. Engelberg-Kulka,H. and Glaser,G. (1999) Addiction modules and programmed cell death and antideath in bacterial cultures. *Annu. Rev. Microbiol*., **53**, 43–70.
12. Jensen,R.B. and Gerdes,K. (1995) Programmed cell death in bacteria: proteic plasmid stabilization systems. *Mol. Microbiol*., **17**, 205–210.
13. Burroughs,A.M., Iyer,L.M. and Aravind,L. (2009) Natural history of the E1-like superfamily: implication for adenylation, sulfur transfer, and ubiquitin conjugation. *Proteins*, **75**, 895–910.
14. Iyer,L.M., Burroughs,A.M. and Aravind,L. (2006) The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like beta-grasp domains. *Genome Biol*., **7**, R60.
15. Kawano,M., Aravind,L. and Storz,G. (2007) An antisense RNA controls synthesis of an SOS-induced toxin evolved from an antitoxin. *Mol. Microbiol*., **64**, 738–754.
16. Yamamoto,T., Hiratani,T., Hirata,H., Imai,M. and Yamaguchi,H. (1986) Killer toxin from Hansenula mrakii selectively inhibits cell wall synthesis in a sensitive yeast. *FEBS Lett*., **197**, 50–54.
17. Hong,Z., Mann,P., Brown,N.H., Tran,L.E., Shaw,K.J., Hare,R.S. and DiDomenico,B. (1994) Cloning and characterization of KNR4, a yeast gene involved in (1,3)-beta-glucan synthesis. *Mol. Cell. Biol*., **14**, 1017–1025.
18. Dagkessamanskaia,A., Martin-Yken,H., Basmaji,F., Briza,P. and Francois,J. (2001) Interaction of Knr4 protein, a protein involved in cell wall synthesis, with tyrosine tRNA synthetase encoded by TYS1 in Saccharomyces cerevisiae. *FEMS Microbiol. Lett*., **200**, 53–58.
19. Basmaji,F., Martin-Yken,H., Durand,F., Dagkessamanskaia,A., Pichereaux,C., Rossignol,M. and Francois,J. (2006) The 'interactome' of the Knr4/Smi1, a protein implicated in coordinating cell wall synthesis with bud emergence in Saccharomyces cerevisiae. *Mol. Genet. Genomics*, **275**, 217–230.
20. Dagkessamanskaia,A., Durand,F., Uversky,V.N., Binda,M., Lopez,F., El Azzouzi,K., Francois,J.M. and Martin-Yken,H. (2010) Functional dissection of an intrinsically disordered protein: understanding the roles of different domains of Knr4 protein in protein-protein interactions. *Protein Sci*., **19**, 1376–1385.
21. Shima,Y., Shima,T., Chiba,T., Irimura,T., Pandolfi,P.P. and Kitabayashi,I. (2008) PML activates transcription by protecting HIPK2 and p300 from SCFFbx3-mediated degradation. *Mol. Cell. Biol*., **28**, 7126–7138.
22. Janke,C., Rogowski,K., Wloga,D., Regnard,C., Kajava,A.V., Strub,J.M., Temurak,N., van Dijk,J., Boucher,D., van Dorsselaer,A. *et al*. (2005) Tubulin polyglutamylase enzymes are members of the TTL domain protein family. *Science*, **308**, 1758–1762.
23. Iyer,L.M., Abhiman,S., Maxwell Burroughs,A. and Aravind,L. (2009) Amidoligases with ATP-grasp, glutamine synthetase-like and acetyltransferase-like domains: synthesis of novel metabolites and peptide modifications of proteins. *Mol. Biosyst*., **5**, 1636–1660.
24. Aoki,S.K., Pamma,R., Hernday,A.D., Bickham,J.E., Braaten,B.A. and Low,D.A. (2005) Contact-dependent inhibition of growth in *Escherichia coli*. *Science*, **309**, 1245–1248.
25. Aoki,S.K., Diner,E.J., de Roodenbeke,C.T., Burgess,B.R., Poole,S.J., Braaten,B.A., Jones,A.M., Webb,J.S., Hayes,C.S., Cotter,P.A. *et al*. (2010) A widespread family of polymorphic contact-dependent toxin delivery systems in bacteria. *Nature*, **468**, 439–442.
26. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
27. Soding,J., Biegert,A. and Lupas,A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res*., **33**, W244–W248.
28. Holm,L., Kaariainen,S., Rosenstrom,P. and Schenkel,A. (2008) Searching protein structure databases with DaliLite v.3. *Bioinformatics*, **24**, 2780–2781.
29. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.
30. Pei,J. and Grishin,N.V. (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, **23**, 802–808.
31. Lassmann,T. and Sonnhammer,E.L. (2005) Kalign—an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, **6**, 298.
32. Pei,J., Sadreyev,R. and Grishin,N.V. (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics*, **19**, 427–428.
33. Goodstadt,L. and Ponting,C.P. (2001) CHROMA: consensus-based colouring of multiple alignments for publication. *Bioinformatics*, **17**, 845–846.
34. Cuff,J.A., Clamp,M.E., Siddiqui,A.S., Finlay,M. and Barton,G.J. (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892–893.
35. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol*., **292**, 195–202.
36. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al*. (2009) The Pfam protein families database. *Nucleic Acids Res*., **38**, D211–D222.
37. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol*., **305**, 567–580.

38. Kall,L., Krogh,A. and Sonnhammer,E.L. (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*, **21(Suppl. 1)**, i251–i257.
39. Price,M.N., Dehal,P.S. and Arkin,A.P. (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.*, **26**, 1641–1650.
40. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
41. Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD: visual molecular dynamics. *J. Mol. Graphics*, **14**, 33–38.
42. Wootton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.*, **266**, 554–571.
43. Shimoike,T., Taura,T., Kihara,A., Yoshihisa,T., Akiyama,Y., Cannon,K. and Ito,K. (1995) Product of a new gene, syd, functionally interacts with SecY when overproduced in *Escherichia coli*. *J. Mol. Biol.*, **270**, 5519–5526.
44. Matsuo,E., Mori,H., Shimoike,T. and Ito,K. (1998) Syd, a SecY-interacting protein, excludes SecA from the SecYE complex with an altered SecY24 subunit. *J. Mol. Biol.*, **273**, 18835–18840.
45. Dalal,K., Nguyen,N., Alami,M., Tan,J., Moraes,T.F., Lee,W.C., Maurus,R., Sligar,S.S., Brayer,G.D. and Duong,F. (2009) Structure, binding, and activity of Syd, a SecY-interacting protein. *J. Mol. Biol.*, **284**, 7897–7902.
46. de Souza,R.F., Iyer,L.M. and Aravind,L. (2010) Diversity and evolution of chromatin proteins encoded by DNA viruses. *Biochim. Biophys. Acta*, **1799**, 302–318.
47. Ye,Y., Osterman,A., Overbeek,R. and Godzik,A. (2005) Automatic detection of subsystem/pathway variants in genome analysis. *Bioinformatics*, **21(Suppl. 1)**, i478–i486.
48. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2010) REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **38**, D234–D236.
49. Osbourn,A.E. and Field,B. (2009) Operons. *Cell. Mol. Life Sci.*, **66**, 3755–3775.
50. Shlyapnikov,S.V., Lunin,V.V., Perbandt,M., Polyakov,K.M., Lunin,V.Y., Levdikov,V.M., Betzel,C. and Mikhailov,A.M. (2000) Atomic structure of the Serratia marcescens endonuclease at 1.1 A resolution and the enzyme reaction mechanism. *Acta Crystallogr. D Biol. Crystallogr.*, **56**, 567–572.
51. Ghosh,M., Meiss,G., Pingoud,A., London,R.E. and Pedersen,L.C. (2005) Structural insights into the mechanism of nuclease A, a betabeta alpha metal nuclease from Anabaena. *J. Mol. Biol.*, **280**, 27990–27997.
52. Aravind,L., Makarova,K.S. and Koonin,E.V. (2000) Holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. *Nucleic Acids Res.*, **28**, 3417–3432.
53. Makhov,A.M., Hannah,J.H., Brennan,M.J., Trus,B.L., Kocsis,E., Conway,J.F., Wingfield,P.T., Simon,M.N. and Steven,A.C. (1994) Filamentous hemagglutinin of Bordetella pertussis. A bacterial adhesin formed as a 50-nm monomeric rigid rod based on a 19-residue repeat motif rich in beta strands and turns. *J. Mol. Biol.*, **241**, 110–124.
54. Beckmann,G., Hanke,J., Bork,P. and Reich,J.G. (1998) Merging extracellular domains: fold prediction for laminin G-like and amino-terminal thrombospondin-like modules based on homology to pentraxins. *J. Mol. Biol.*, **275**, 725–730.
55. Jacob-Dubuisson,F., Locht,C. and Antoine,R. (2001) Two-partner secretion in Gram-negative bacteria: a thrifty, specific pathway for large virulence proteins. *Mol. Microbiol.*, **40**, 306–313.
56. Krishna,S.S., Majumdar,I. and Grishin,N.V. (2003) Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res.*, **31**, 532–550.
57. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) The SCOP database. *Nucleic Acids Res.*, **36**, D419–D425.
58. Aravind,L., Makarova,K.S. and Koonin,E.V. (2000) Survey and summary: holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. *Nucleic Acids Res.*, **28**, 3417–3432.
59. Sokolowska,M., Czapinska,H. and Bochtler,M. (2009) Crystal structure of the beta beta alpha-Me type II restriction endonuclease Hpy99I with target DNA. *Nucleic Acids Res.*, **37**, 3799–3810.
60. Renzi,F., Caffarelli,E., Laneve,P., Bozzoni,I., Brunori,M. and Vallone,B. (2006) The structure of the endoribonuclease XendoU: From small nucleolar RNA processing to severe acute respiratory syndrome coronavirus replication. *Proc. Natl Acad. Sci. USA*, **103**, 12365–12370.
61. Ricagno,S., Egloff,M.P., Ulferts,R., Coutard,B., Nurizzo,D., Campanacci,V., Cambillau,C., Ziebuhr,J. and Canard,B. (2006) Crystal structure and mechanistic determinants of SARS coronavirus nonstructural protein 15 define an endoribonuclease family. *Proc. Natl. Acad. Sci. USA*, **103**, 11892–11897.
62. Wang,J., Chen,R. and Julin,D.A. (2000) A single nuclease active site of the Escherichia coli RecBCD enzyme catalyzes single-stranded DNA degradation in both directions. *J. Mol. Biol.*, **275**, 507–513.
63. Carr,S., Walker,D., James,R., Kleanthous,C. and Hemmings,A.M. (2000) Inhibition of a ribosome-inactivating ribonuclease: the crystal structure of the cytotoxic domain of colicin E3 in complex with its immunity protein. *Structure*, **8**, 949–960.
64. Delattre,A.S., Clantin,B., Saint,N., Locht,C., Villeret,V. and Jacob-Dubuisson,F. (2010) Functional importance of a conserved sequence motif in FhaC, a prototypic member of the TpsB/Omp85 superfamily. *FEBS J.*, **277**, 4755–4765.
65. Pallen,M.J. (2002) The ESAT-6/WXG100 superfamily—and a new Gram-positive secretion system? *Trends Microbiol.*, **10**, 209–212.
66. Iyer,L.M., Makarova,K.S., Koonin,E.V. and Aravind,L. (2004) Comparative genomics of the FtsK-HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. *Nucleic Acids Res.*, **32**, 5260–5279.
67. Peterson,F.C. and Volkman,B.F. (2009) Diversity of polyproline recognition by EVH1 domains. *Front Biosci.*, **14**, 833–846.
68. Pallen,M.J., Chaudhuri,R.R. and Henderson,I.R. (2003) Genomic analysis of secretion systems. *Curr. Opin. Microbiol.*, **6**, 519–527.
69. Das,D., Finn,R.D., Abdubek,P., Astakhova,T., Axelrod,H.L., Bakolitsa,C., Cai,X., Carlton,D., Chen,C., Chiu,H.J. *et al.* (2010) The crystal structure of a bacterial Sufu-like protein defines a novel group of bacterial proteins that are similar to the N-terminal domain of human Sufu. *Protein Sci.*, **19**, 2131–2140.
70. Burglin,T.R. (2008) The Hedgehog protein family. *Genome Biol.*, **9**, 241.
71. Perler,F.B. (1998) Protein splicing of inteins and hedgehog autoproteolysis: structure, function, and evolution. *Cell*, **92**, 1–4.
72. Hall,T.M., Porter,J.A., Young,K.E., Koonin,E.V., Beachy,P.A. and Leahy,D.J. (1997) Crystal structure of a Hedgehog autoprocessing domain: homology between Hedgehog and self-splicing proteins. *Cell*, **91**, 85–97.
73. Cenciarelli,C., Chiaur,D.S., Guardavaccaro,D., Parks,W., Vidal,M. and Pagano,M. (1999) Identification of a family of human F-box proteins. *Curr. Biol.*, **9**, 1177–1179.
74. Bai,C., Sen,P., Hofmann,K., Ma,L., Goebl,M., Harper,J.W. and Elledge,S.J. (1996) SKP1 connects cell cycle regulators to the ubiquitin proteolysis machinery through a novel motif, the F-box. *Cell*, **86**, 263–274.
75. Tukachinsky,H., Lopez,L.V. and Salic,A. (2010) A mechanism for vertebrate Hedgehog signaling: recruitment to cilia and dissociation of SuFu-Gli protein complexes. *J. Cell. Biol.*, **191**, 415–428.
76. Wang,C., Pan,Y. and Wang,B. (2010) Suppressor of fused and Spop regulate the stability, processing and function of Gli2 and Gli3 full-length activators but not their repressors. *Development*, **137**, 2001–2009.
77. Menard,C., Wagner,M., Ruzsics,Z., Holak,K., Brune,W., Campbell,A.E. and Koszinowski,U.H. (2003) Role of murine cytomegalovirus US22 gene family members in replication in macrophages. *J. Virol.*, **77**, 5557–5570.
78. Valchanova,R.S., Picard-Maureau,M., Budt,M. and Brune,W. (2006) Murine cytomegalovirus m142 and m143 are both required to block protein kinase R-mediated shutdown of protein synthesis. *J. Virol.*, **80**, 10181–10190.
79. Budt,M., Niederstadt,L., Valchanova,R.S., Jonjic,S. and Brune,W. (2009) Specific inhibition of the PKR-mediated antiviral response

by the murine cytomegalovirus proteins m142 and m143. *J. Virol.*, **83**, 1260–1270.
80. Child,S.J. and Geballe,A.P. (2009) Binding and relocalization of protein kinase R by murine cytomegalovirus. *J. Virol.*, **83**, 1790–1799.
81. Child,S.J., Hanson,L.K., Brown,C.E., Janzen,D.M. and Geballe,A.P. (2006) Double-stranded RNA binding by a heterodimeric complex of murine cytomegalovirus m142 and m143 proteins. *J. Virol.*, **80**, 10173–10180.
82. Hakki,M., Marshall,E.E., De Niro,K.L. and Geballe,A.P. (2006) Binding and nuclear relocalization of protein kinase R by human cytomegalovirus TRS1. *J. Virol.*, **80**, 11817–11826.
83. Child,S.J., Hakki,M., De Niro,K.L. and Geballe,A.P. (2004) Evasion of cellular antiviral responses by human cytomegalovirus TRS1 and IRS1. *J. Virol.*, **78**, 197–205.
84. Hakki,M. and Geballe,A.P. (2005) Double-stranded RNA binding by human cytomegalovirus pTRS1. *J. virol.*, **79**, 7311–7318.
85. Marshall,E.E., Bierle,C.J., Brune,W. and Geballe,A.P. (2009) Essential role for either TRS1 or IRS1 in human cytomegalovirus replication. *J. Virol.*, **83**, 4112–4120.
86. Cassady,K.A. (2005) Human cytomegalovirus TRS1 and IRS1 gene products block the double-stranded-RNA-activated host protein shutoff response induced by herpes simplex virus type 1 infection. *J. Virol.*, **79**, 8707–8715.
87. Xuan,B., Qian,Z., Torigoi,E. and Yu,D. (2009) Human cytomegalovirus protein pUL38 induces ATF4 expression, inhibits persistent JNK phosphorylation, and suppresses endoplasmic reticulum stress-induced cell death. *J. Virol.*, **83**, 3463–3474.
88. Moorman,N.J., Cristea,I.M., Terhune,S.S., Rout,M.P., Chait,B.T. and Shenk,T. (2008) Human cytomegalovirus protein UL38 inhibits host cell stress responses by antagonizing the tuberous sclerosis protein complex. *Cell Host Microbe*, **3**, 253–262.
89. Terhune,S., Torigoi,E., Moorman,N., Silva,M., Qian,Z., Shenk,T. and Yu,D. (2007) Human cytomegalovirus UL38 protein blocks apoptosis. *J. Virol.*, **81**, 3109–3123.
90. McCormick,A.L., Roback,L., Livingston-Rosanoff,D. and St Clair,C. (2010) The human cytomegalovirus UL36 gene controls caspase-dependent and -independent cell death programs activated by infection of monocytes differentiating to macrophages. *J. Virol.*, **84**, 5108–5123.
91. Mulec,J., Podlesek,Z., Mrak,P., Kopitar,A., Ihan,A. and Zgur-Bertok,D. (2003) A cka-gfp transcriptional fusion reveals that the colicin K activity gene is induced in only 3 percent of the population. *J. Bacteriol.*, **185**, 654–659.
92. Dugatkin,L.A. (2007) Inclusive fitness theory from Darwin to Hamilton. *Genetics*, **176**, 1375–1380.
93. Hayes,C.S., Aoki,S.K. and Low,D.A. (2010) Bacterial contact-dependent delivery systems. *Annu. Rev. Genet.*, **44**, 71–90.
94. Jackson,A.P., Thomas,G.H., Parkhill,J. and Thomson,N.R. (2009) Evolutionary diversification of an ancient gene family (rhs) through C-terminal displacement. *BMC Genomics*, **10**, 584.
95. Conticello,S.G. (2008) The AID/APOBEC family of nucleic acid mutators. *Genome Biol.*, **9**, 229.
96. Tan,Y. and Riley,M.A. (1997) Nucleotide polymorphism in colicin E2 gene clusters: evidence for nonneutral evolution. *Mol. Biol. Evol.*, **14**, 666–673.
97. Iyer,L.M., Babu,M.M. and Aravind,L. (2006) The HIRAN domain and recruitment of chromatin remodeling and repair activities to damaged DNA. *Cell Cycle*, **5**, 775–782.
98. Hamill,S.J., Cota,E., Chothia,C. and Clarke,J. (2000) Conservation of folding and stability within a protein family: the tyrosine corner as an evolutionary cul-de-sac. *J. Mol. Biol.*, **295**, 641–649.
99. Velikovsky,C.A., Deng,L., Tasumi,S., Iyer,L.M., Kerzic,M.C., Aravind,L., Pancer,Z. and Mariuzza,R.A. (2009) Structure of a lamprey variable lymphocyte receptor in complex with a protein antigen. *Nat. Struct. Mol. Biol.*, **16**, 725–730.