Beyond Slam: the DUF560 family outer membrane protein superfamily SPAM has distinct network subclusters that suggest a role in non-lipidated substrate transport and bacterial environmental adaptation

Terra J. Mauer[1¶], Alex S. Grossman[2¶], Katrina T. Forest[1], and Heidi Goodrich-Blair[1,2*]
[1]University of Wisconsin-Madison, Department of Bacteriology, Madison, WI

[2]University of Tennessee-Knoxville, Department of Microbiology, Knoxville, TN

¶These authors contributed equally to this work.

*To whom correspondence should be addressed: hgblair@utk.edu, https://orcid.org/0000-0003-1934-2636

Short title: Bacterial Surface/Secreted Protein Associated Outer Membrane Proteins (SPAMs)

Abstract

In host-associated bacteria, surface and secreted proteins mediate acquisition of nutrients, interactions with host cells, and specificity of host-range and tissue-localization. In Gram-negative bacteria, the mechanism by which many proteins cross, become embedded within, or become tethered to the outer membrane remains unclear. The domain of unknown function (DUF)560 occurs in outer membrane proteins found throughout and beyond the proteobacteria. Functionally characterized DUF560 representatives include NilB, a host-range specificity determinant of the nematode-mutualist *Xenorhabdus nematophila* and the surface lipoprotein assembly modulators (Slam), Slam1 and Slam2 which facilitate surface exposure of lipoproteins in the human pathogen *Neisseria meningitidis*. Through network analysis of protein sequence similarity we show that DUF560 subclusters exist and correspond with organism lifestyle rather than with taxonomy, suggesting a role for these proteins in environmental adaptation. Cluster 1

had the greatest number of representative proteins, was dominated by homologs from animal-associated symbionts, and was composed of subclusters: 1A (containing NilB, Slam1, and Slam2), 1B, and 1C. Genome neighborhood networks revealed that Cluster 1A DUF560 members are strongly associated with TonB, TonB-dependent receptors, and predicted co-receptors such as the Slam1 lipoprotein substrates transferrin binding protein and lactoferrin binding protein. The genome neighborhood network of Cluster 1B sequences are similarly dominated by TonB loci, but typically the associated co-receptors (the presumed DUF560 substrates) are predicted to be non-lipidated. We suggest that these subclusters within the DUF560 protein family indicate distinctive activities and that Slam activity may be characteristic of Cluster 1A members but not all DUF560 homologs. For Cluster 1 DUF560 homologs we propose the name SPAM (Surface/Secreted Protein Associated Outer Membrane Proteins) to accommodate the potential for non-lipoprotein substrates or different activities. We show that the repertoire of SPAM proteins in *Xenorhabdus* correlates with host phylogeny, suggesting that the host environment drives the evolution of these symbiont-encoded proteins. This pattern of selection for specific sequences based on host physiology and/or environmental factors may extend to other clusters of the DUF560 family.

Introduction

All eukaryotes exist in close association with microbial symbionts that can contribute to nutrition, development, protection from threats, and reproductive fitness. In these associations the genotypes of the partners influence the degree to which beneficial or detrimental outcomes occur, and the outcomes in turn directly influence fitness of both partners (examples: [1-6]). Bacteria vary greatly in their genomic potential, even among members of a single genus or species. This variation raises the question of how hosts evaluate potential partners for their symbiotic potential and recruit those most likely to support positive outcomes. One way that

microbes interact with and affect hosts is through surface and secreted proteins. Surface exposed or excreted proteins can have a variety of functions, including mediating acquisition of nutrients (e.g. TonB-dependent transporters in gram-negative host-associated bacteria [7,8]), interactions with host cells and surfaces (e.g. *Acinetobacter baumanii* OmpA binding host desmoplakin [9]), and specificity in host-range and tissue-localization (e.g. swapping pili production genes in *Vibrio fischeri* leads to changes in host specificity [10]).

In gram negative bacteria, proteins are transported to the periplasm or outside the cell through diverse types of well-studied secretion systems (reviews: [11,12]). However, mechanistic details of how proteins become embedded within or tethered to the outer membrane remain unclear. Recently, a mechanism of lipoprotein tethering to the outer leaflet of the outer membrane was identified in *Neisseria meningitidis* and termed the Slam (Surface lipoprotein assembly modulator) machinery [13,14]. In this system, prolipoproteins are initially transported across the inner membrane by the Sec or Tat translocases, and subsequently localized and tethered to either the inner or outer membrane by the Lol (localization of lipoprotein) system [15]. In *N. meningitidis*, Slam proteins containing a beta-barrel domain of the protein family DUF560 are then required in a third step for surface presentation of certain lipoproteins, especially those that capture host-metal chaperones and allow proliferation within a host exhibiting nutritional immunity [13]. In addition, two different Slam proteins from the same strain exhibit differential specificity for lipoprotein targets [13,14].

DUF560 family proteins are present throughout Proteobacteria, and only a few members have been characterized [13,14,16,17]. One characterized DUF560 homolog, NilB, was identified as a host-association and species-specificity factor in the entomopathogenic nematode symbiont *Xenorhabdus nematophila*, a proteobacterium in the family Morganellaceae [16-20]. The ability of a nematode host to associate with a particular bacterial species, strain, or mutant can be

experimentally investigated, since methods exist to generate axenic *Steinernema* nematodes that can be paired with laboratory-grown, genetically tractable *Xenorhabdus* bacteria and easily assessed for colonization [21,22]. Using these methods, a screen for *X. nematophila* nematode intestinal localization (*nil*) mutants defective in colonizing the nematode *S. carpocapsae* revealed a genomic locus, termed Symbiosis Region 1 (SR1) [16,17,23]. This locus contains the genes *nilB* and *nilC*, each of which is necessary for colonization of *S. carpocapsae* nematodes. Curiously, the distribution of the SR1 locus among *Xenorhabdus* species is limited, having been identified in only 3 of the 46 strains for which full or draft genome sequences are available (Unpublished data, [17,23]). The association between *Xenorhabdus* bacteria and their *Steinernema* spp. nematode hosts is species specific; the intestinal tract of each species of *Steinernema* typically is colonized specifically by a single *Xenorhabdus* species [24-28]. The SR1 region is sufficient to confer *S. carpocapsae* colonization proficiency on non-native, non-colonizing *Xenorhabdus* strains, indicating this locus is a determinant of species-specific interactions between *X. nematophila* and *S. carpocapsae* [23]. Biochemical and bioinformatic analyses of the genes encoded on SR1 have established that NilC is an outer-membrane-associated lipoprotein and NilB is a 14-stranded outer-membrane beta-barrel protein in the DUF560 (PF04575) family, with a ~140 aa N-terminal domain (NTD) that contains a tetratricopeptide repeat (TPR) [16,17,29,30].

Slam activity has been demonstrated for DUF560 representatives from *Neisseria meningitidis*, *Pasteurella multocida, Moraxella catarrhalis,* and *Haemophilus influenzae* [13,14] but it remains unclear whether lipoprotein translocation is a universal attribute of DUF560 proteins. Some characterized Slams are encoded adjacent to genes encoding their experimentally-verified lipoprotein targets [13,14]. However, of the 353 DUF560 homologs bioinformatically examined by Hooda *et al.* only 185 were encoded near genes predicted to encode lipidated proteins [14], leaving open the possibility that some DUF560 family members recognize non-lipidated targets.

To begin to understand the range of functions of DUF560 proteins in biology, we assessed their distribution, genomic context, and relatedness. We focused particularly on the distribution and putative substrates of these proteins among *Xenorhabdus* bacteria since the DUF560 repertoire of these symbionts could be placed in the context of host specificity. We experimentally examined the *X. nematophila* DUF560 homolog XNC1_0074, which is not encoded near a predicted lipoprotein-encoding gene. Taken together, our data suggest that the activities of the DUF560 family extend beyond lipoprotein surface presentation. We suggest a more inclusive acronym to describe the family: Surface/Secreted Protein Associated Outer Membrane Proteins (SPAM) proteins.

Results

DUF560 proteins are encoded by diverse Proteobacteria with up to six copies represented in a single genome

The Pfam database, which divides proteins into families based on sequence similarity and hidden Markov models [31] includes PF04575 (http://pfam.xfam.org/family/PF04575) (IPR007655), otherwise known as the domain of unknown function 560 (DUF560) family. This family includes 707 sequences from 509 species in 25 domain architectures (Table S1; Pfam32 database, accessed 5/15/19). The majority (86%) of the Pfam DUF560 sequences are encoded by Proteobacteria, as easily visualized in a taxonomic distribution plot (Fig. 1; Table S2) [32]. Among the proteobacteria the number of DUF560 homologs present in an individual species ranged from 0 to 6, with *Neisseria gonorrhoeae* (taxid 528360) as the only species in the dataset to have 6 homologs. Outside the proteobacteria, DUF560 homologs showed scattered representation among several other phyla. The 707 DUF560 sequences fall into 25 domain architectures, with two domain architectures, DUF560 alone (478 sequences) and TPR_19,

DUF560 (152 sequences), containing the majority of DUF560 sequences (89.1% of the pfam32 dataset), including Slam1, Slam2, and NilB. Both groups have a DUF560 C-terminal domain (CTD) but differ in the N-terminal domain (NTD), with the former being of low complexity and the latter having an annotated TPR motif (Table S2). Twenty of the remaining domain architectures have TPR motifs in the N-terminus with or without additional domains (e.g. ankyrin [protein-protein interaction], SH3 [eukaryotic-motif possibly cell entry or intracellular survival], and ANAPC3 [TPR related]), and the final three domain architectures lack TPR motifs but have other, potentially functionally equivalent domains (ANAPC3 [TPR related], BTAD [TPR related], TIG and Big_9 [both bacterial immunoglobulin-like domains]). Overall, DUF560 architectures indicate a conserved periplasmic NTD that likely mediates protein-protein interactions [29] and a CTD comprising a 14-stranded outer membrane beta barrel.

DUF560 protein sub-families exist and are correlated with lifestyle rather than taxonomy

Using either NilB or Slam proteins as bait, previous work had identified a wide distribution of DUF560 proteins in Gram-negative bacteria, seemingly enriched in those associated with animal mucosal surfaces [14,16,17]. We sought to gain more quantifiable information about whether sub-families of DUF560 homologs exist and whether DUF560 are enriched among mucosal symbionts. A sequence similarity network of DUF560 homologs was generated using the Enzyme Function Initiative (EFI) [33-36] (Fig. 2) and manually annotated using publicly available database entry information to highlight either the environmental source (Fig. 2A, Table S3) or the taxonomic grouping (Fig. 2B) of the isolates containing the DUF560-related-proteins. We identified 10 major clusters of DUF560-related proteins (Fig. 2, Table S3). Cluster 1 contains the majority of nodes in the network (691; 62.4%) and could be visually divided into three subclusters (1A, 1B, and 1C) using force directed node placement (Fig. 3). Consistent with our previous observations, the majority of the 691 Cluster 1 nodes (521, 75%) comprise sequences from animal-associated isolates (including mammal, invertebrate, nematode, and

general animal annotations) and another 20% contain sequences from marine, freshwater, soil, or built-environment isolates (Fig. 3A).

The division of sequence nodes among the three subclusters 1A, 1B, and 1C more strongly reflected the environmental origin than the taxonomy of bacterial isolates. Subcluster 1A almost exclusively comprises mammal- or animal-associated bacteria and contains both *Neisseria* Slam1 and Slam2, though these occupy different regions of the subcluster and separate at higher alignment scores (Fig. 4). The three known NilB *Xenorhabdus* homologs (from *X. nematophila, X. innexi,* and *X. stockiae*) cluster together within Subcluster 1A, and cluster separately from *N. meningitidis* Slam homologs when viewed at higher resolution (Fig. 4). Subclusters 1B and 1C do not have any as-yet-characterized DUF560 members. Subcluster 1B contains a mixture of nodes containing sequences from animal- or plant-associated and free-living bacteria while Cluster 1C contains DUF560 sequences largely from environmental bacterial isolates, and plant-associated representatives are scattered among all the clusters. In contrast, subclusters 1 and 2 do not align as well with phylum-level taxonomy of the isolates. For example, the network Cluster 1 contains 81 nodes with sequences from Moraxellaceae and all but seven of these nodes have exclusively Moraxellaceae in them. Of the Moraxellaceae-containing nodes, 79% are dominated by animal-associated isolates (mammal, animal, nematode, invertebrate) and 12.3% are dominated by environmental isolates (water, plant, soil, built), and these nodes exist in both Subclusters 1A and 1B, but not in 1C. The animal-associated isolates are enriched in the former and environmental-isolates are enriched in the latter: 78% of the Moraxellaceae animal-associated nodes are in Subcluster 1A, while 40% of the environmental isolates are in Subcluster 1B. Overall, the network data indicate that DUF560 protein sub-families exist and are correlated with lifestyle rather than taxonomy. Further, they reveal that thus far the only functionally characterized members of the DUF560 family (Slam1, Slam2, and NilB), each of which is involved in specialized host interactions, occupy the same

Subcluster 1A, which is enriched for host-associated bacteria. Taken together, these data suggest that other subclusters of DUF560 homologs may have divergent functions relative to those already characterized.

The Slam acronym was defined on the basis that DUF560 homologs from *N. meningitidis, M. catarrhalis, H. influenzae,* and *P. multocida* facilitate surface exposure of lipoproteins, with *N. meningitidis* Slam1 facilitating TbpB and LbpB surface exposure and Slam2 facilitating hemoglobin/haptoglobin receptor surface exposure [13,14]. In *Neisseriaceae* and *Pasteurellaceae,* TbpB is a co-receptor for the integral outer-membrane TonB-dependent receptor, TbpA, with which it serves to acquire iron from transferrin [37,38]. The possibility that many DUF560 homologs may have Slam function is further supported by the fact that many of them, including *Xenorhabdus nilB,* are genomically associated with genes predicted to encode lipoproteins [14]. Since all DUF560 homologs functionally characterized to date as Slams fall within Subcluster 1A we next considered whether the association with lipoproteins is characteristic of Cluster 1 members. We used the EFI Genome Neighborhood Tool [33-36] to identify co-occurring protein families within the genomic contexts of each DUF560 subcluster member. We retrieved a genome neighborhood network (GNN) for each of the subclusters (1A, 1B, and 1C, although 1C has too many genome neighborhood category members to be shown graphically) (Fig. S1, Table S4). This analysis revealed that genes predicted to encode TonB and TonB-dependent receptors are commonly associated with DUF560 members in Subclusters 1A and 1B, but not in Subcluster 1C. A subset of DUF560 homologs in Subcluster 1A are associated with proteins involved in binding and/or transport of metals [transferrin binding protein B beta-barrel domain (TbpB_B_D) (15/15), putative zinc/iron chelating domain CxxCxxCC (9/15), transporter associated domain (7/15), PhnA_Zn_ribbon (3/15)], protein/peptide modification or transport [EscC T3SS family protein (13/15), metallopeptidase family M24 peptidase (13/15), and aminotransferase class-III (13/15)]. Subcluster 1B, in addition

to being associated with TonB-dependent receptors (240/350), TbpB_B_D (215/350), and TonBC (191/350) are also associated with FecR (118/350) which regulates iron uptake, and major facilitator superfamily MFS_1 (81/350) indicating potential further function in metal efflux. Subcluster 1C has no such obvious patterns of function (Table S4). These association data combined with knowledge of Slam1 function implicate members of Subcluster 1A and 1B as facilitators of metal-binding-protein secretion.

Given the prevalence of TbpB_B_D proteins in the EFI GNN we examined them more closely in terms of their gene structure and their genomic association with DUF560 homologs (Fig. 3C). Using a combination of EFI-GNN co-occurrence data [33], Rapid ORF Description & Evaluation Online (RODEO) software data [39], and manual annotation we analyzed genomic frameworks extending ten to twenty ORFs away from each DUF560 homolog present in Cluster 1 of the network. Within these genomic frameworks proteins predicted to have one or more TbpB_B_D domains were retrieved. Isolated ORFs were analyzed with SignalP-5 to predict the signal peptide of each TbpB_B_D-domain-containing-protein and the list of protein pairs and predictions can be seen in Table S5 [40]. We manually annotated each node within the Cluster 1 network based on the presence or absence of a co-occurring gene predicted to encode a TbpB_B_D domain, and whether this protein was predicted to be secreted and non-lipidated (with signal peptidase 1 [SP1] cleavage sites and lacking a lipobox) or lipidated (with signal peptidase 2 [SP2] cleavage sites and a lipobox) (Fig. 3C). This analysis revealed that the majority of TbpB_B_D proteins genomically associated with Subcluster 1A DUF560 homologs are predicted to be lipidated, hereafter referred to as TbpB_B_D_{lip}. This includes the *H. influenzae* Slam1-dependent transferrin-binding-protein co-receptor TbpB, which has two TbpB_B_D domains, one in the N-lobe and one in the C-lobe (Fig. 5) [14,41,42].

In contrast, the TbpB_B_D proteins associated with Subcluster 1B are almost exclusively predicted to be secreted (hereafter referred to as TbpB_B_D_{sol}) (Fig. 3C; Fig. 5). One example is *X. nematophila* XNC1_0075, encoded adjacent to the DUF560 homolog XNC1_0074 in a TonB-dependent heme-receptor protein/TonB genomic context (Hrp locus). XNC1_0075 is encoded adjacent to XNC1_0074 and in the same orientation, as previously described [17] and it includes a TbpB_B_D "lipoprotein-5" class domain (Fig. 5) [14,17]. However, XNC1_0075 has a SPI-type (rather than SPII) signal sequence and lacks the canonical lipobox necessary for lipidation so is predicted to be a secreted soluble protein. Further, the XNC1_0075 homolog is small (~24.7 kDa), has only one TbpB_B_D domain (versus two found in TbpB), and does not appear to have the canonical TbpB_A or TbpB_C handle domains that pair with a TbpB_B_D domain in the N-lobe and C-lobe of TbpB, respectively (Fig. 5). The *Haemophilus haemolyticus* hemophore, hemophilin also has this domain architecture (Fig. 5). Biochemical and structural evidence support the conclusion that it is a soluble secreted protein that binds free heme and facilitates heme uptake into the cell [43]. Three-dimensional homology modeling (Phyre^2 [44,45]) was used to visualize potential structural similarities between hemophilin and several TbpB_B_D_{sol} proteins including XNC1_0075. Models were built using either the NTD alone (amino acids 22-140 only) (Table S6) or including the TbpB_B_D domain (Table S6; Fig. S5). For all queries Phyre selected protein data bank file C6OM, Hemophilin [43] as the top template and C3HO,TbpB from *Actinobacillus pleuropneumoniae* [46] as the second best template. These comparisons indicate that XNC1_0075 and the other TbpB_B_D_{sol} homologs likely are secreted proteins, similar to hemophilin (EGT80255.1) and other hemophores such as HasA [43,47-49]. The presence of the TbpB_B_D domain may indicate a conserved function in metal acquisition despite the apparent absence of lipidation at the N-terminus. We propose that XNC1_0075 is likely to bind directly to heme or to a metal-chelating protein based on its similar domain architecture and predicted structure with those of the heme-binding hemophilin, as well as the genomic context of the XNC1_0075 ORF among genes predicted to encode heme

uptake mechanisms. Further, TpbB is a substrate of DUF560 Slam1 in *N. meningitidis* [13,14], strengthening the notion that despite the absence of a lipid anchor, TbpB_B_D_{sol} homologs may be substrates for DUF560 homologs for extracellular exposure. As further support for potential TbpB_B_D_{sol} export, *N. gonorrhoeae* FA19 TbpB lacking its lipidation site is found in the supernatant in contrast to WT TbpB which is surface exposed but cell anchored, indicating that even without lipid anchoring, TbpB can be exported and can interact with TbpA [50].

Subcluster 1B DUF560 homologs are associated with single-domain TbpB_B_D proteins that are predicted to be secreted and soluble.

The features of the Hrp locus, including a DUF560 homolog encoded near a gene predicted to encode a non-lipidated, single TbpB_B_D_{sol} domain protein, are conserved in several proteobacteria including *Photorhabdus* species, *Proteus mirabilis*, *Providencia rettgeri, Acinetobacter baumannii* AB0057*, Neisseria gonorrhoeae* FA 1090, and *Haemophilus haemolyticus* (Fig. 5 and data not shown). The TbpB_B_D_{sol} proteins of *A. baumanii* (AB57_0988/WP_001991318.1), *N. gonorrhoeae* (NGO0554), and *H. haemolyticus* (KKZ58958.1) are each flanked by a gene predicted to encode an outer membrane receptor (AB57_0987/WP_085920715.1, NGO0553, and KKZ58957.1, respectively) and a gene predicted to encode a DUF560 protein (AB57_0989/WP_079740461.1, NGO0555, and KKZ58959.1, respectively). Consistent with the predicted role of these TbpB_B_D_{sol} proteins in metal homeostasis, NGO0554 is repressed by iron, upregulated in response to oxidative stress and contributes to resistance to peroxide challenge [51-53]. None of these TbpB_B_D-containing proteins are predicted to be lipidated and the associated DUF560 homologs encoded in the *A. baumanii* and *H. haemolyticus* Hrp loci are members of Subcluster 1B (Fig. 3). These observations reinforce the possibility that if TbpB_B_D_{sol} proteins, such as hemophilin, are substrates for their respective DUF560 proteins and that the activity of DUF560 proteins extends beyond surface presentation of lipoproteins. Further, our data indicate that DUF560

homologs cluster based on specialization for distinctive molecular characteristics of their substrates such as the presence (for Subcluster 1A) or absence (for Subcluster 1B) of a lipid tail. In light of this expanded view of DUF560 potential function we propose a more inclusive name of Surface/Secreted Protein Associated Outer Membrane Proteins (SPAM) to encompass the entirety of DUF560 Cluster 1, of which Subcluster 1A Slams are a subset.

Xenorhabdus genomes encode a conserved Hrp SPAM

Our network analysis revealed that SPAM distribution can vary among bacterial species and corresponds to bacterial environmental niche and that the nematode mutualist *X. nematophila* encodes two SPAM homologs NilB and XNC1_0074, from subgroups 1A and 1B, respectively. All known *Xenorhabdus* species are obligate mutualists of nematodes in the genus *Steinernema*, and the host identity of most isolated *Xenorhabdus* species is known [24-28,54]. The Magnifying Genome (MaGe) platform contains 46 closed or draft *Xenorhabdus* genomes, providing us the opportunity to analyze the distribution of SPAM homologs among *Xenorhabdus* species and its relation to nematode host identity (a key environmental parameter for *Xenorhabdus*). We used sequence similarity searches through BLASTp [55,56] with NilB (SPAM subgroup 1A; SPAM1A) and XNC1_0074 (SPAM subgroup 1B; SPAM1B) as queries to assess the repertoire and genomic contexts of SPAM homologs within the 46 available genomes (Fig. 6; Table 1; Table S7).

We found that all *Xenorhabdus* genome sequences encode a homolog of XNC1_0074 that occurs in the Hrp genomic locus context and is hereafter referred to as SPAM1B$_{hrp1}$ (Fig. 6, Table 1; Fig. S2; Table S7). The conservation of the Hrp locus in sequenced *Xenorhabdus* genomes and its predicted role in metal acquisition or homeostasis suggested a conserved function in the *Xenorhabdus* lifecycle, which is entirely host-associated. To test this possibility we inactivated the *X. nematophila* SPAM1B$_{hrp1}$ gene, *XNC1_0074* through insertion-deletion

mutagenesis, creating *X. nematophila ΔXNC1_0074::kan*. We saw no difference between wild type and the *ΔXNC1_0074::kan* mutant in a battery of standard tests, including growth in minimal medium with casamino acids, lipase activity, protease activity, motility on soft agar surfaces, and biofilm formation in defined medium (data not shown). Unlike NilB (SPAM1A), XNC1_0074 (SPAM1B$_{hrp1}$) is not critical for nematode colonization: *ΔXNC1_0074::kan* colonization of *S. carpocapsae* infective juvenile nematodes was not different from wild type under our standard co-cultivation conditions on lipid agar (Fig. S3). Since the ability to kill insects is a common trait among *Xenorhabdus* species we tested the virulence of the *ΔXNC1_0074::kan* mutant using our standard assays of injection into *Manduca sexta* insects. The *XNC1_0074::kan* mutant displayed a killing phenotype similar to wild type, indicating it does not have a role in virulence in this insect injection model (Fig. S4).

Xenorhabdus **genomes vary in their SPAM occurrence**

In addition to the SPAM1B$_{hrp1}$/TbpB_B_D$_{sol}$ pair at the Hrp locus some *Xenorhabdus* species encode other SPAM homologs and putative targets. We categorized *Xenorhabdus* genomes into 6 classes (I-VI) based on the presence or absence of these additional homologs (Fig. 6; Table 1; Table S6). Class I, including *X. japonica, X. koppenhoeferi, X. vietnamensis,* and all 17 sequenced *X. bovienii* strains have only one copy of SPAM (and its associated TbpB_B_D$_{sol}$) at the Hrp locus, with no additional copies detected. The majority of these genome sequences are drafts, precluding definitive conclusions about the presence or absence of other SPAMs in these strains. However, two SPAM Class I genomes (*X. bovienii* SS-2004 and CS03) are complete, enabling the conclusion that these two *X. bovienii* strains encode only one SPAM [57-59]. Class II genomes have an additional SPAM1B/TbpB_B_D$_{sol}$ pair (Hrp2) in the Hrp locus (Fig. 6, Table 1, Fig. S6, Table S7). Class III, comprises only the *X. hominickii* genome which has an additional SPAM1B/TbpB_B_D$_{sol}$ pair in another genomic context surrounded by multiple

transposes and other mobile genetic elements (Fig. 6, Table 1, Table S7). Finally, another SPAM1B/Tbp_B_D_{sol} pair was detected near a large locus encoding proteins predicted to be involved in cobalt utilization (Cut). *Xenorhabdus* with a Cut locus SPAM1B (SPAM1B_{cut}) were categorized as Class IV (*X. cabanillasii* and *X. budapestensis*) or V (*X. innexi*) depending on the absence or presence, respectively of a NilB(SPAM1A)/NilC$_{lip}$ pair. Finally Class VI genomes of *X. nematophila* and *X. stockiae* encode Hrp SPAM1B/TbpB_B_D_{sol} and NilB(SPAM1A)/NilC$_{lip}$.

As expected, NilB distribution was restricted among *Xenorhabdus* species: other than *X. nematophila,* NilB was found only in *X. innexi* and *X. stockiae* (Fig. 6 and Table 1). The *X. nematophila* SR1 locus (encoding NilB and NilC) is adjacent to a locus encoding a non-ribosomal peptide synthetase responsible for the synthesis of the anti-microbial PAX lysine-rich cyclopeptide [60,61], while *nilB* and *nilC* of *X. innexi* and *X. stockiae* interrupt a locus predicted to encode MARTX proteins, which is intact in *X. nematophila* (Fig. S7) [62]. The varied genomic context of the NilB(SPAM1A)/NilC pairs is consistent with previous suggestions that these genes were horizontally acquired by *X. nematophila* [23]. Based on these data we suggest that the Hrp SPAM homologs have been vertically inherited and that they play a conserved and general role in *Xenorhabdus-Steinernema* interactions, a concept supported by the phylogenetic relationships of TbpB_B_D_{sol} paralogs (Fig. S7). This alignment indicates that the TbpB_B_$D_{solHrp1}$ paralog is basal and was likely present in the common ancestor of *Xenorhabdus* and its close relatives in the *Photorhabdus* genus, which have a similar entomopathogenic and nematode mutualistic lifestyle [63]. Further, we suggest that some *Xenorhabdus* strains have acquired additional SPAMs, through duplication or horizontal gene transfer, that play specialized roles in bacterial cell physiology or host interactions. This is supported by the facts that the additional *Xenorhabdus* TbpB_B_D_{sol} copies (Hrp2, Tn, and Cut) group separately from the Hrp1 paralogs (98 bootstrap value) and that NilB is a host-species specificity determinant in *X. nematophila* ATCC 19061 [23]. In this regard it is notable that *X.*

innexi and *X. stockiae*, while not closely related to *X. nematophila*, do associate with nematodes (*S. scapterisci* and *S. siamkayai)* that are closely related to *S. carpocapsae* [54,64].

SPAM content classes correlate with host nematode taxonomy

Our network analysis suggests that SPAMs have diversified and hint that this diversification may be correlated with environmental features rather than taxonomy. To further assess this possibility, we exploited the *Xenorhabdus* symbiosis for which we have both host environment and SPAM content information. To visualize correlations between these two parameters we constructed Maximum Likelihood and Bayesian phylogenetic trees using genes conserved among 32 *Xenorhabdus* strains for which we had whole genome sequence and a classified SPAM distribution (Table 1, Table S7, Table S8, Fig. S9, Fig. S10). Next, we constructed Maximum Likelihood and Bayesian phylogenetic trees of their *Steinernema* hosts using a multi-locus approach combining Internal Transcribed Spacer, 18S rRNA, 28S rRNA, Cytochrome Oxidase I, and 12S rRNA sequences as available (Table S9, Fig. S11, Fig. S12). Both trees were color-coded according to the bacterial SPAM distribution class (Table S7, Fig. 7). We observed that the phylogenetic placement of the *Steinernema* host was more predictive of the SPAM complement of their *Xenorhabdus* symbionts than was the bacterial phylogenetic placement. For example, the majority of *Xenorhabdus* within SPAM Class II (with two SPAM1B/TbpB_B_D_{sol} homologs at the Hrp locus) are symbionts of nematodes within the phylogenetic Clade IV, suggesting that these nematodes present a distinctive environment in which the additional SPAM1B_{Hrp2} homolog is beneficial. Further support for correspondence of SPAM distribution and host phylogeny occurs in the subset of strains highlighted in Fig. 7B. The presence of a SPAM1B_{Cut} homolog (red lines) is a feature of symbionts of Clade V nematodes. *X. innexi* and *X. stockiae* (orange and purple lines respectively) diverged from this lineage with the gain of a NilB homolog and, in the case of *X. stockiae,* the loss of the SPAM1B_{Cut} homolog. This divergence correlates with a switch in host association with Clade II nematodes. The *X.*

nematophila lineage appears to have independently gained *nilB*, consistent with the distinctive genomic location of this gene relative to its location in the *X. stockiae* and *X. innexi* genomes (Fig. S8). This acquisition correlates with the ability of *X. nematophila* to colonize the Clade II nematode *S. carpocapsae* [65].

Our previous work has established that the SR1 locus, encoding NilB and NilC, is necessary and sufficient to confer colonization of *S. carpocapsae* [16,23]. Two other NilB-encoding *X. nematophila* strains were isolated from the nematodes *S. websteri* and *S. anatoliense*, which are not in the same clade as *S. carpocapsae*. Our hypothesis that SPAM homologs are involved in host-environment adaptations predicts that *X. nematophila* will also require the SR1 locus to colonize these nematodes. To test this idea we generated axenic eggs of *S. anatoliense*, *S. websteri*, and *S. carpocapsae* [21] and exposed them to an *X. nematophila* ATCC19061 SR1 mutant (HGB1495) and an SR1 complemented strain (HGB1496) [17]. The colonization states of individual nematodes were visually monitored using fluorescence microscopy to observe GFP-expressing bacteria at known colonization sites: the anterior intestinal caecum (AIC) of developing nematodes and the receptacle of progeny infective juvenile (IJ) nematodes. In addition, colonization was quantified as the average colony forming units (CFU) per IJ by grinding populations of nematodes and dilution plating the homogenate for CFU. Like *S. carpocapsae*, developing *S. anatoliense* and *S. websteri* are colonized by *X. nematophila* at the AIC (Fig. S13) [27]. We found that the SR1 mutant had a trend towards lower AIC colonization in adult females and juveniles compared to the SR1 complement although there were no significantly different results in the presence or absence of SR1 for a particular stage or species (Fig. S13). The SR1 mutant was deficient in infective juvenile (IJ) receptacle colonization in all three nematode species (Fig. 8) demonstrating that SR1, encoding NilB (SPAM1A) and NilC, is necessary for IJ colonization of nematodes (*S. anatoliense* and *S. websteri*) outside of Clade II and supporting our hypothesis that SPAMs are involved in adaptations to host environments.

Discussion

The Surface/Secreted Protein Associated Outer Membrane Proteins (SPAM) family of proteins

The DUF560 (domain of unknown function) family comprises outer membrane beta-barrel proteins. Its presence in and importance for animal-associated bacteria was first recognized by Heungens *et al.* (2002) who noted that the *X. nematophila* DUF560 homolog NilB host-colonization factor has homologs in *N. meningitidis, M. catarrhalis, P. multocida,* and *H. influenzae* [16]. This correlation was strengthened by subsequent analyses of the phylogenetic distribution of NilB and other DUF560 homologs, as well as by the demonstration that the Slam DUF560 family members from human pathogenic bacteria facilitate the surface presentation of host-metal acquisition systems [13,14,17]. Here, enabled by the availability of extensive new genomic sequences of bacteria from myriad environments and ever-improving bioinformatic computational and visualization tools, we have presented data demonstrating that DUF560-domain containing genes are present in bacteria from a wide range of habitats including water, soil, and plants. This expanded network and our follow up analyses have revealed two major insights into the function and distribution of the DUF560 family of proteins, which we coin SPAM proteins.

First, SPAM homologs can be categorized by their relatedness into clusters, and these clusters correspond with environmental niche. For example, Cluster 1A is dominated by nodes with SPAMs derived from mammal- and animal-associated bacteria, including *N. meningitidis* Slam1 and Slam2 and *X. nematophila* NilB, Cluster 1B is dominated by SPAMs from various host-associated bacteria, including *Xenorhabdus* SPAM1B$_{Hrp}$, while Cluster 1C represents SPAMs from water, soil, and built-environment isolates. The environmental signature of the clusters suggests that SPAMs are important for adaptations that improve bacterial fitness under distinct

environmental pressures, a hypothesis supported by the surface localization of SPAMs. We have provided support for this theory using the *Xenorhabdus*-*Steinernema* mutualism by showing that *Xenorhabdus* species are evolving SPAM repertoires that correspond with the taxonomy of their *Steinernema* nematode host environments.

Second, our analyses suggest that the different SPAM clusters have predictive power for the characteristics of the potential targets of SPAM activity. Work in *N. meningitidis* and several other pathogens has demonstrated that SPAM1A$_{Slam1}$ and SPAM1A$_{Slam2}$ function to present their lipoprotein targets TbpB, LbpB, fHbp, and HpuA on the cell surface [13,14]. TbpB is a co-receptor, with the TonB-dependent receptor TbpA, for host-derived transferrin [66]. Using a genome neighborhood network analysis we identified a similar association of SPAM1B homologs with genes predicted to encode TonB, TonB-dependent receptors, and proteins with a TbpB_B_D domain. Strikingly, these SPAM1B-associated TbpB_B_D proteins, which are found in all *Xenorhabdus* and *Photorhabdus* species as well as human pathogens including *N. gonorrhoeae, A. baumanii, P. rettgeri*, and *P. mirabilis* are distinguishable from the SPAM1A$_{Slam1}$ substrates by virtue of their small size (≈half the TbpB protein) and lack of lipidation signals, suggesting they are soluble proteins. These strong correlations suggest distinctive functions for SPAM1A and SPAM1B members in surface delivery of lipidated and non-lipidated substrates, respectively. Extending this idea further, SPAMs expressed by bacteria that occupy plant, built, soil, water, and other environments likely function in the secretion of diverse proteins that enable adaptation to these environments. Since SPAMs are surface molecules, they represent a potential target for development of treatments for issues deriving from the presence of such bacteria in these environments, such as plant disease and surface biofilms in hospitals and pipes. The network-enabled classification of SPAMs presented here will facilitate the investigation of these proteins in diverse bacteria.

Our work also enables categorization of SPAM homologs within *Neisseria* species that are found in many animal hosts, some as commensal members of the community, and some as pathogens. *N. meningitidis* and *N. gonorrhoeae* are strictly human associated, residing in the oronasopharynx and urogenital tract respectively [67]. As detailed above, *Neisseria* strains can encode up to 6 SPAM paralogs. *N. meningitidis* MC58 has two functionally characterized SPAMs, Slam1 and Slam2, that both fall within Cluster 1A. Our network analysis indicates *N. meningitidis* encodes a third SPAM, NMB1466/ NP_274965, that also falls within Cluster 1A. Overall, *N. gonorrhoeae* SPAMs are represented in more nodes than *N. meningitidis*, particularly in Subcluster 1B: Of the 108 Subcluster 1A nodes, 17.6% have *N. meningitidis* and 20.3% have *N. gonorrhoeae*, with 5 nodes (5.6%) overlapping between them. Of the 31 Subcluster 1B nodes, 6.4% have *N. meningitidis* SPAMs, while 19.3% have *N. gonorrhoeae*, without any overlap. Taken together, these data indicate that *N. gonorrhoeae* strains encompass a somewhat greater diversity of SPAM paralogs in both Subclusters 1A and 1B, which may indicate diversified targets for surface exposure and interaction with the human host environment.

The Hrp locus is conserved among *Xenorhabdus* species and may encode a metal acquisition system

The genomic correlation of SPAM1B with TbpB_B_D_{sol} described above and analogy to the SPAM1A_{Slam1} targeting of TbpB raises the hypothesis that *Xenorhabdus* SPAM1B_{hrp1} facilitates transport of TbpB_B_$D_{solhrp1}$ across the outer membrane. Extending this analogy further, Phyre models [45] indicate that the TbpB_B_D_{sol} proteins fold similarly to the substrate-binding N-lobe of TbpB. *Xenorhabdus* species encode at the Hrp locus a conserved predicted TonB-dependent receptor, akin to TbpA, as well as TonB itself. Taken together, these observations lead to the model that the vertically inherited Hrp locus of *Xenorhabdus* (and *Photorhabdus*) species encodes a metal acquisition system either binding heme, via a hemophilin-like fold, or binding

transferrin via a TbpB-like fold. Members of the *Xenorhabdus* and *Photorhabdus* genera are pathogens of insects and mutualists of nematodes (of the genera *Steinernema* and *Heterorhabditis*, respectively), and transferrin-like molecules would be accessible to these bacteria in either of these environments.

Transferrin homologs are found in the circulatory fluids of some insects, including *Drosophila melanogaster* which encodes three transferrin homologs, one of which (Tsf1) is found in the hemolymph where it plays a role in trafficking iron between the intestine and fat body (analog of the liver) [68-70]. In *Manduca sexta* the hemolymph transferrin (MsTf) is induced upon infection and in its apo-form it can inhibit the growth of *E. coli* in an iron-starvation-dependent manner, indicating that like other animals insects appear to engage in nutritional immunity [71,72]. It is tempting to speculate that the Hrp locus, which is predicted to encode a TonB-dependent metal acquisition system and is ancestral and conserved among the entomopathogens *Xenorhabdus* and *Photorhabdus* (which associate with different nematode hosts), plays a role in countering such a response. We did not observe a virulence defect of the *X. nematophila hrp* locus mutant when injected into *M. sexta* insects. However, *X. nematophila* virulence is multifactorial and it is difficult to observe a dramatic phenotype associated with disruption of a single locus. Further, *M. sexta* may not be fully representative of natural hosts infected by the nematode host of *X. nematophila*. Future experiments will delve more specifically into the role of Hrp machinery in recognizing, binding, and acquiring insect-host metals. While we favor the concept that the Hrp machinery plays a role in the insect host environment it may also function during interactions with the nematode host. Both *Caenorhabditis elegans* and *Steinernema carpocapsae* encode genes of unknown function (CELE_K10D3.4/NP_492020.1 and L596_007732/TKR93238.1, respectively) each with a TR_FER (transferrin) and a Periplasmic_Binding_Protein_Type_2 domains indicative of transferrin protein family members. The *X. nematophila hrp* locus mutant

did not have a colonization defect in *S. carpocapsae* nematodes under standard laboratory conditions, but such a defect may become apparent under limiting metal regimes.

SPAMs as mediators of host adaptation

Using the *Steinernema-Xenorhabdus* symbiosis we have demonstrated that the composition of SPAM copies in the bacterial genome correlates with host environment. *X. nematophila* NilB, a member of the SPAM Subcluster 1A, is necessary for association with Clade II nematode *S. carpocapsae* [16,23,27]. Two other *Xenorhabdus* species, *X. innexi* and *X. stockiae* encode orthologs of NilB. These two species are in a distinct bacterial lineage than is *X. nematophila* but both colonize Clade II nematodes (*S. scapterisci* and *S. siamkayai*). These data indicate that *Xenorhabdus* require NilB to effectively colonize the Clade II nematode environment. Notably, *X. nematophila* cannot colonize the Clade II *S. scapterisci* host of *X. innexi,* nor can *X. innexi* colonize *S. carpocapsae* [23,62], despite the fact that both *X. nematophila* and *X. innexi* encode NilB (and NilC) homologs. This indicates that the different NilB homologs are not interchangeable, consistent with the fact that they each occupy a separate node within Subcluster 1A (Fig. 4).

Colonization phenotypes in the *Steinernema-Xenorhabdus* association include bacterial adherence to the anterior intestinal caecum (AIC) of developing juvenile and adult nematodes and the occupation of an intestinal luminal pocket of the soil-dwelling, non-feeding infective juvenile (IJ) stage, which transmits the bacterial symbiont between insect hosts [27]. The *X. nematophila* 3.5-Kb SR1 locus, encoding the SPAM1A NilB and the lipoprotein NilC, is necessary for colonization of the AIC and the IJ of *S. carpocapsae* nematodes. *X. nematophila* SR1 mutants have moderate to severe defects in colonizing the AIC and IJ receptacle. Further, other *Xenorhabdus* species that cannot colonize these tissues in *S. carpocapsae* gain this ability when provided a copy of the SR1 locus [16,17,23,27]. These published data established

that SR1 is a host-range specificity determinant that is necessary and sufficient for colonization of *S. carpocapsae* nematodes [23]. Our phylogenetic analysis (Fig. 7) led us to test the prediction that non-Clade II nematode hosts of *X. nematophila*, *S. anatoliense* and *S. websteri* also require SR1 for AIC and IJ colonization. As expected, *X. nematophila* was able to colonize *S. carpocapsae, S. anatoliense,* and *S. websteri* at both the AIC and in the IJ receptacle (Fig. 8; Fig. S13; [65]. Further, the SR1 locus was essential for colonizing the IJ stage of all three nematode species, confirming a role of NilB in colonization of non-Clade II nematode hosts of *X. nematophila.* The AIC colonization phenotype was less clear: although a general trend was apparent in lower AIC colonization by the SR1 mutant relative to wild type, this trend was not significant. In *S. carpocapsae*, the role of SR1 in colonization of the AIC is most striking during the second generation of nematode reproduction, while at earlier stages the phenotype is less clear. As such, a more detailed temporal analysis comparing wild type and SR1 mutant colonization of *S. anatoliense* and *S. websteri* AIC would be necessary to firmly establish a role for SR1 in this aspect of the association.

Conclusions

Our work indicates that DUF560 homologs, or SPAMs, have diversified for adaptation to specific environments, including specific host species or tissues. This can be seen in the in-depth analysis of *Xenorhabdus* species and their specific *Steinernema* hosts through a co-phylogeny which indicates that presence of specific DUF560 homologs contributes to host-range. The network-based classification system we propose provides a framework for predicting the function and targets of SPAM family members and highlights that current mechanistic knowledge of this family is restricted to a single subcluster.

Materials and methods

DUF560 distribution within bacteria

All TaxIDs encoding DUF560 homologs that were included in the network analysis (Clusters 1-4 and 6-11) shown in Fig. 2 were added to a custom, semicolon separated values file (organized TaxID;DUF560#) (Table S2). Current bacterial TaxIDs were isolated from the latest version of Uniprot_bacteria FTP file (updated 7/10/19) and added to the custom file above with a DUF560 value = 0. Duplicate TaxIDs were removed. This custom file was submitted to Aquerium [32] as a custom input, with background set as only the input file, and taxonomy set to the root taxonomy, with TaxIDs being assigned to taxonomies based on the PFAM (v.28.0 and v.27.0) and TIGRFAM (v.15 and v.14) databases through Aquerium (Fig. 1).

DUF560 sequence similarity network analysis

Enzyme Function Initiative's Enzyme Similarity Tool (EFI-EST) was used to collect all predicted DUF560-domain-containing proteins from the Interpro 73 and the Uniprot 2019-02 databases (accessed 4.24.19) and to perform an all-by-all BLAST to obtain similarities of these protein sequences [33-36]. Initially each sequence was represented by a node on the network, however for ease of analysis representative networks were constructed by collapsing nodes which shared ≥ 40% identity. On an EFI-EST network, edges are drawn according to a database-independent value called alignment score. A greater alignment score requirement means fewer edges are drawn and the network "hairball" is more separated. For separation of the whole protein family of DUF560-domain-containing proteins an alignment score of 38 was chosen (Fig. 2). For separation of the more closely related Subcluster 1A an alignment score of 89 was chosen (Fig. 4A, C) and for separation of Subcluster 1B an alignment score of 100 was chosen (Fig. 4B, D). The EFI-EST Color SSN tool was used to assign cluster numbers to each node. After generation, networks were visualized and interpreted using Cytoscape v3.7.1 [73] and

Gephi v0.9.2 [74]. Nodes were arranged with the Fruchterman-Reingold force-directed layout algorithm which aims for intuitive node placement and equal length edges wherever possible [75].

Within the whole protein family network (Fig. 2), the contents of each cluster were compared to the more conservative Pfam DUF560 (PF04575) to ensure that clusters were actually DUF560 proteins [31]. Any clusters for which fewer than 18% of sequences were present in the Pfam family, or which included fewer than 20 total sequences were excluded from downstream analysis. This filtering removed cluster 5, which was composed almost entirely of *Klebsiella pneumoniae* PgaA sequences, a protein which is often incorrectly annotated as having a DUF560 domain. This generated a sequence similarity network that contains 10 clusters (1-4, 6-11), 1222 nodes and 52190 edges with 1589 TaxIDs represented, some of which have multiple DUF560 homologs represented (Fig. 2). Each node was examined and manually curated for the isolates' environmental origin(s) among the following categories: water, soil, plant, mammal, animal, invertebrate, nematode, built (environments such as sewage, bioreactors, mines etc. where humans have had a large impact), multiple, and unclassified. For each node, all Taxonomy IDs were searched in the NCBI Taxonomy Browser. If a relevant citation was available, it was recorded (Table S3). If no citation was available, the isolation source was searched for in the biosamples or bioprojects records (if available). If neither resource was available, strain isolation source was searched for through other resources (NCBI linkout, google search). A node was assigned an environment if the majority of strains within the node fell into that environment. If a node had no majority environment (or was tied), the category multiple environments was assigned. If no taxonomy id within a node gave a specific environment, the category unclassified was assigned. Animal association was a focus, so any nodes where the majority of sequences could be assigned to mammal, insect, or nematode associated microbes were separated. Any node with animal associated microbes that did not fit

one of those groups, was designated a generic “animal associated”. For fine scale interpretation, Cluster 1 was isolated out for some analyses (Fig. 3). In Subcluster 1A and Subcluster 1B networks all singleton nodes were excluded from analysis.

Three different techniques were used to determine if DUF560 proteins present within our network were genomically associated with TbpB_B_D-domain-containing proteins. First, using Enzyme Function Initiative’s Genome Neighborhood Network Tool (EFI-GNN) GNNs were generated for Subcluster 1A (Alignment Score 89; Rep node 40; 20 ORFs around), Subcluster 1B (Alignment Score 38; Rep node 40; 10 ORFs around), and Subcluster 1C (Alignment Score 38; Rep node 40; 10 ORFs around) [76]. These data sets resulted in 352 DUF560-TbpB_B_D protein pairs. Next, the Rapid ORF Description & Evaluation Online (RODEO) web tool was used to analyze regions upstream and downstream of each DUF560 protein [39] RODEO uses profile Hidden Markov Models to assign domains to nearby ORFs and look at protein co-occurrence. This data set resulted in 712 DUF560-TbpB_B_D protein pairs. Finally, seven additional DUF560-TbpB_B_D protein pairs known in *Xenorhabdus* were manually annotated. All three datasets were combined for a total of 851 non-redundant protein pairs (Table S5) and SignalP-5 [40,77] was used to predict the signal peptides of all TbpB_B_D-domain-containing-proteins. These predictions were used to manually annotate each node of the network (Fig. 3C). In the event that a node was associated with both signal peptide bearing TbpB_B_D proteins and those with no predicted signal peptide it was annotated according to the signal peptide bearing TbpB_B_D proteins.

DUF560 genome neighborhood analysis

Subclusters 1A-C were separated and analyzed in EFI-EST with an alignment score of 38 and a repnode 40 as above. Each network was then analyzed through Enzyme Function Initiative’s Genome Neighborhood Network Tool (EFI-GNN) (Alignment Score 38; Rep node 40; 10 ORFs

around). For each subcluster, the DUF560 co-occurring genes are shown in Table S4, and can be visualized for Subclusters 1A and B in Fig. S1.

Three-dimensional structure similarity predicted through Phyre analysis

TbpB_B_D_{sol} protein sequences from *Xenorhabdus nematophila* (XNC1_0074), *Providencia rettgeri* (PROVETT_08181/PROVETT_05852), and *Proteus mirabilis* (WP_134940027.1) were collected and the first 22 amino acids were trimmed to prevent the signal sequences from influencing the final structure predictions. These sequences were entered into the Phyre[2] Protein Homology/analogy Recognition Engine v. 2.0 to predict potential 3-dimensional structures. (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) [45]. The top predicted structural model output for all three proteins, hemophilin (protein data bank file 6OM5), was used to generate structural models (Fig.S5). Subsequent PDB files were visualized with Protean 3D v15. (Protean 3D®. Version 15.0. DNASTAR. Madison, WI). To elucidate the NTD of these proteins, amino acids 22-140 of each one were isolated and entered into Phyre[2]. The top two models of all protein sequences, alongside the coverage and confidence of those models, are recorded in Table S6.

Isolation of a *ΔXNC1_0074::kanR* mutant

For deletion of the *XNC1_0074* gene, a 1000 bp fragment upstream of XNC1_0074 was amplified from the *X. nematophila* ATCC 19061 (HGB800) genome using the primers XNC1_0074_UP_F (aacgacggccagtgaattcgagctctcaataaattaaataaaataaacaatataaagc) (all primers listed in 5'-3' direction) and XNC1_0074_UP_R (agacacaacgtggtggaagaccaacttacaaag). Likewise, a 1000 bp fragment downstream of XNC1_0074 was amplified from the genome using the primers XNC1_0074_DN_F (tgagtttttctaatcaaatagttatgatttttatttgcgtgatgataacc) and XNC1_0074_DN_R (caagcttgcatgcctgcaggtcgacacggcagatacgggtccag). A kanamycin resistance (*kanR*) cassette was amplified from pEVS107 using primers KanR_F

(agttggtcttccaccacgttgtgtctcaaaatc) and KanR_R (cataactatttgattagaaaaactcatcgagc). The pUC19 backbone was amplified using pUC19_F (gtcgacctgcaggcatgc) and pUC19_R (gagctcgaattcactggccgtc). All 4 fragments were assembled into a circular plasmid using a Hi-Fi assembly kit (New England Biolabs, Ipswitch, MA) according to manufacturer's instructions. The resulting plasmid was transformed into *E. coli* S17**λ***pir* for plasmid production. Subsequently, the *XNC1_0074* upstream region, *kanR* cassette, and *XNC1_0074* downstream region were excised as a single unit by SacI and SalI restriction digestion. Simultaneously, the suicide vector pKR100 was linearized via digestion with SacI and SalI. The deletion cassette region was ligated into pKR100 using T4 DNA ligase, generating pKR100-*XNC1_0074*::*kanR*. The plasmid was transformed back into *E. coli* S17**λ***pir* from which it was conjugated into *X. nematophila* ATCC 19061 according to the method described in Murfin et al., 2012 [21] except that strains were initially subcultured 1:10 into fresh medium. Exconjugants were grown on LB agar containing 50 μg ml^{-1} kanamycin to select for recombinants. Colonies were then tested for sensitivity to 15 μg ml^{-1} chloramphenicol. The presence of the *kanR* cassette and deletion of *XNC1_0074* were confirmed using Sanger sequencing at the University of Tennessee Genomics Core using multiple primer combinations including 0074confirmF (aaccaatccacattgctgtgg) and 0074confirmR (caacaaatcctgcatcaacaacct). All PCRs performed with ExTaq DNA polymerase (Takara Bio USA, Inc., Mountain View, CA) according to the manufacturer's instructions.

***Manduca sexta* survival assay**

Manduca sexta larvae were raised for 9 days prior to bacterial injection. Briefly, eggs were sterilized with 0.6% bleach solution for 2 min. before hatching. *Manduca* larvae were reared at 26°C with a 16-hour light cycle in order to simulate natural conditions. During rearing *Manduca* larvae were fed gypsum moth diet. Injections were carried out as soon as the *Manduca* larvae reached the 4th instar life stage. Bacterial cultures of wild type *X. nematophila* ATCC19061

(HGB800) and a *ΔXNC1_0074::kanR mutant* (HGB2301) were grown at 30°C in Lysogeny Broth (LB) medium that had been stored in the dark or supplemented with 0.1% pyruvate [78,79]. Prior to injection bacterial cultures were brought into logarithmic growth and rinsed in 1X PBS twice. For each treatment 35 insect were injected with ≈300 bacteria suspended in 10μl 1X PBS. As a negative control 35 additional insects were injected with 10μl sterile 1X PBS. All solutions were injected between the first and second set of prolimbs. Insect mortality was observed over a 50-66 h time course. Death was defined as when an insect ceased all movement and response to stimuli.

Biofilm Assay

Cultures of *X. nematophila* ATCC19061 (HGB800), GFP expressing *X. nematophila* ATCC19061 *att*Tn7::*GFP* (HGB2018), and GFP expressing *X. nematophila ΔXNC1_0074::kanR kefA*::*GFP* (HGB2302) were grown in a defined medium [17]. Cultures were equalized to an initial OD_{600} of 0.05 and grown for 72 hours at 30°C on 24 well polystyrene plates. Biofilms were rinsed twice with ddH_20, stained for 15 minutes with 0.1% crystal violet, rinsed 2 additional times, and dried. Crystal violet was re-suspended in 95% ethanol and measured at OD_{590} to measure biofilm biomass.

Phylogenetic tree generation

To generate the bacterial phylogenetic tree, select *Xenorhabdus* and *Photorhabdus* species with whole genome sequencing data were analyzed. MicroScope MaGe's Gene Phyloprofile tool [76,80] was used to identify homologous protein sets which were conserved across all assayed genomes resulting in 1661 initial sets. Homologs which had multiple sets in any assayed organisms (putative paralogs) were excluded from downstream analysis to ensure homolog relatedness, resulting in the 665 homologous sets shown in Table S8. Whole genomes

were downloaded from MicroScope MaGe and indexed via locus tags. Homolog sets were retrieved using BioPython [81], individually aligned using Muscle v3.8.31 [82], concatenated into a single alignment using Sequence Matrix v1.8 [83], and trimmed of nucleotide gaps using TrimAL v1.3 [84]. JmodelTest v2.1.10 [85,86] was used to choose the GTR+γ substitution model for bacterial maximum likelihood and Bayesian analysis.

For nematode phylogenetic analysis select *Steinernema* and *Heterorhabditis* species were analyzed. The Internal Transcribed Spacer (ITS), 18S rRNA, 28S rRNA, Cytochrome Oxidase I, and 12S rRNA loci were collected from Genbank's protein database as available and used as homolog protein sets as shown in Table S9. Nematode species which had fewer than 3 of the 5 loci sequenced were excluded from downstream analysis. Homologues sets were individually aligned, concatenated, and trimmed using the same methods as the *Xenorhabdus* sequences [81-84]. JmodelTest v2.1.10 [85,86] was used to choose the GTR+γ+I substitution model for nematode maximum likelihood and Bayesian analysis.

Maximum likelihood analyses were performed via RAxML v8.2.10 [87] using rapid bootstrapping and 1000 replicates. Maximum likelihood trees were visualized via Dendroscope v3.6.2 [88]. Nodes with less than 60% bootstrap support were deemed low support and collapsed. Bayesian analyses were performed via MrBayes v3.2.6 with BEAGLE [89-91] on the Cipres Science Gateway platform [92]. 500,000 MCMC replicates were performed for the bacterial tree, 4,000,000 were performed for the nematode tree. The first 25% of replicates were discarded as burn-in, and posterior probabilities were sampled every 500 replicates. Two runs were performed, each with 3 heated and one cold chain. The final standard deviation of split frequencies was 0.000000 for the bacterial tree, and 0.002557 for the nematode tree. Bayesian trees were visualized with FigTree v1.4.4 [93]. Bayesian and maximum likelihood methods

generated phylogenies with consistent topologies after collapsing low support maximum likelihood nodes (Figs. S9-S12).

WT vs. ΔSR1 colonization of *S. anatoliense*, *S. carpocapsae*, and *S. websteri*

Axenic eggs of *S. anatoliense*, *S. carpocapsae*, and *S. websteri* were generated [21] and exposed them to an SR1 mutant (HGB1495 *X. nematophila* Δ*nilR* ΔSR1 *kefA:GFP* empty attTn7 site), and an SR1 complemented strain (HGB1496 *X. nematophila* Δ*nilR* Δ*SR1 kefA:GFP* attTn7::SR1) grown on lipid agar plates for 2 days at 25°C [17]. 3 days after adding eggs, colonization was visualized using fluorescence microscopy on the Keyence BZX-700 microscope to observe GFP expressing bacteria at the anterior intestinal caecum (AIC) (Fig. S13). Lipid agar plates were placed into water traps 1 week after adding eggs to collect infective juvenile (IJ) nematodes. To determine the percent of IJs colonized, 2 weeks after adding eggs, IJ nematodes were visualized using fluorescence microscopy on the Keyence BZX-700 microscope to observe GFP expressing bacteria in the receptacle, and greater than 2000 IJs were counted for each biological replicate (3 biological replicates, 2 technical replicates) (Fig. 8A). To determine the number of CFU per IJ, 200 IJs were surface sterilized, resuspended in PBS, and ground for 2 min using a Fisherbrand motorized tissue grinder (CAT# 12-1413-61) to homogenize the nematodes and release the bacteria. Serial dilutions in PBS were performed and plated on LB agar, which were incubated at 30°C for 1 day before enumerating CFUs (Fig. 8B). To calculate the CFU per colonized IJ (Fig. 8C), the percent colonized nematodes was divided by the CFU/IJ for each biological replicate. For all data shown in Fig. 8, the data were analyzed using a one-way Anova with Tukey's multiple comparison's test to compare the mean of each treatment.

Acknowledgments

This work was supported by a grant from the National Science Foundation (IOS- 1353674) to KTF and HGB and the University of Tennessee-Knoxville to HGB. TJM was supported by an NIH National Research Service Award T32-GM07215.

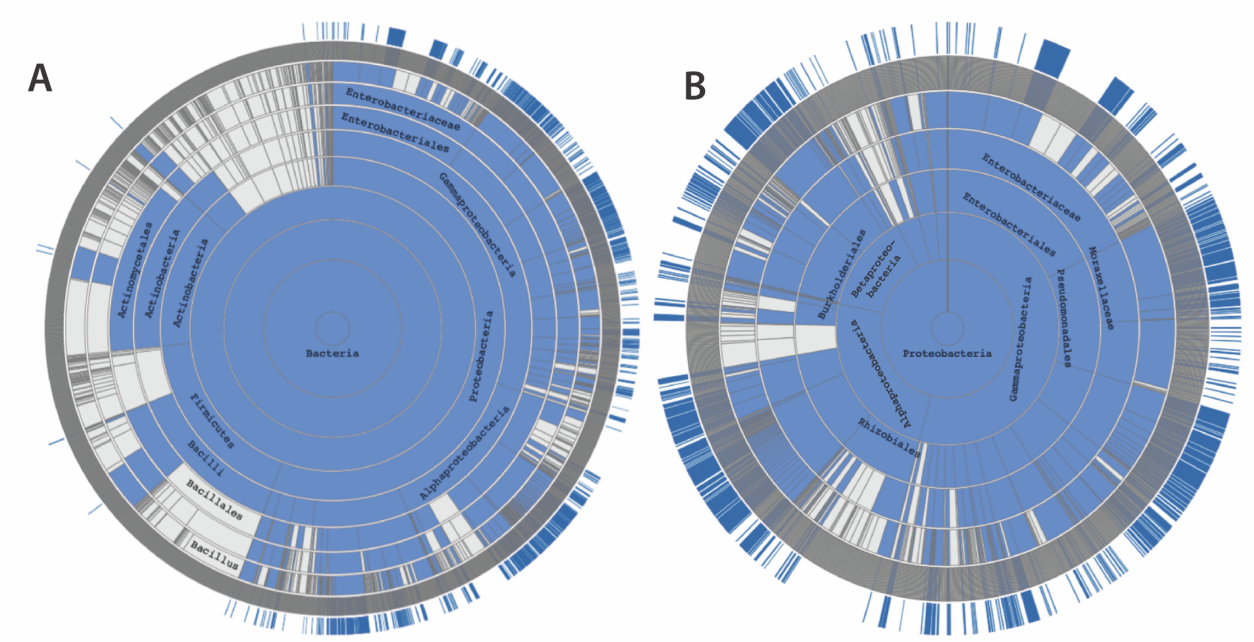

Figure 1. DUF560 homologs occur almost exclusively in genomes of proteobacteria

DUF560 homolog taxonomic distribution is shown. Taxonomic categories increase in resolution from the middle outward, with the most basal taxonomic classification in the center. A blue line in the outermost ring indicates presence of DUF560 encoded in a particular TaxID. A) Bacteria are shown as the basal taxonomic category. B) Proteobacteria are shown as the basal taxonomic category to allow for a more detailed viewing of the distribution. Images were created using the Sunburst option of Aquerium [32].

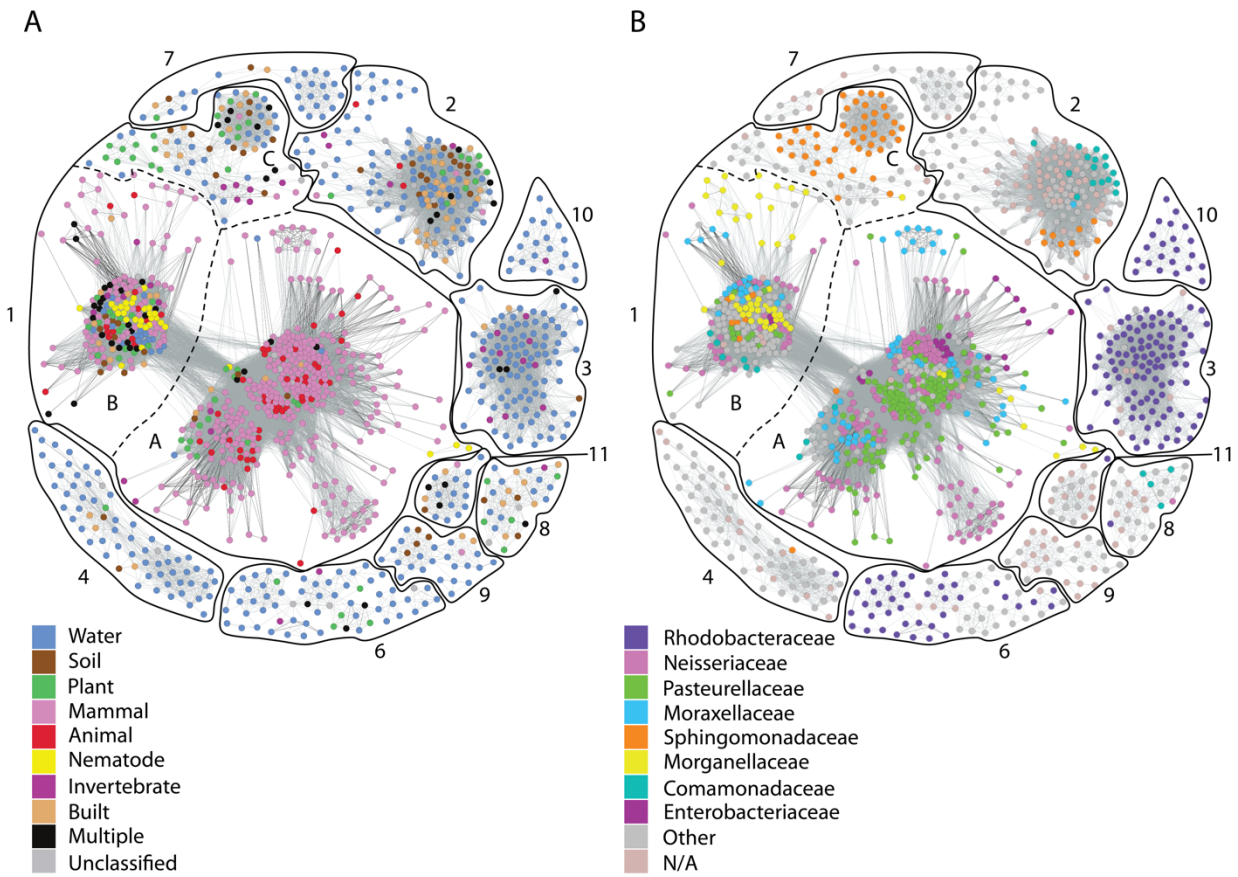

Figure 2. A sequence similarity network shows grouping of DUF560 homologs from animal-associated bacteria by environment rather than taxonomy. A sequence similarity network was generated of DUF560 homologs using the EFI application accessed 4.24.19 [33-36]. The sequences were aligned using an alignment score of 38, and any sequences which shared ≥ 40% identity were collapsed to allow the separation of clusters and prevent noise. Each node is a group of highly similar sequences, with the edge darkness demonstrating similarity, and the distance between nodes determined by using the Fruchterman-Reingold algorithm to optimize edge lengths [75]. Any clusters for which fewer than 18% of sequences were present in the Pfam family, or which included fewer than 20 total sequences were excluded from downstream analysis. After generation of these plots, each node was either automatically assigned or manually curated to show the isolates' environmental origin(s) (A) and taxonomic class (B).

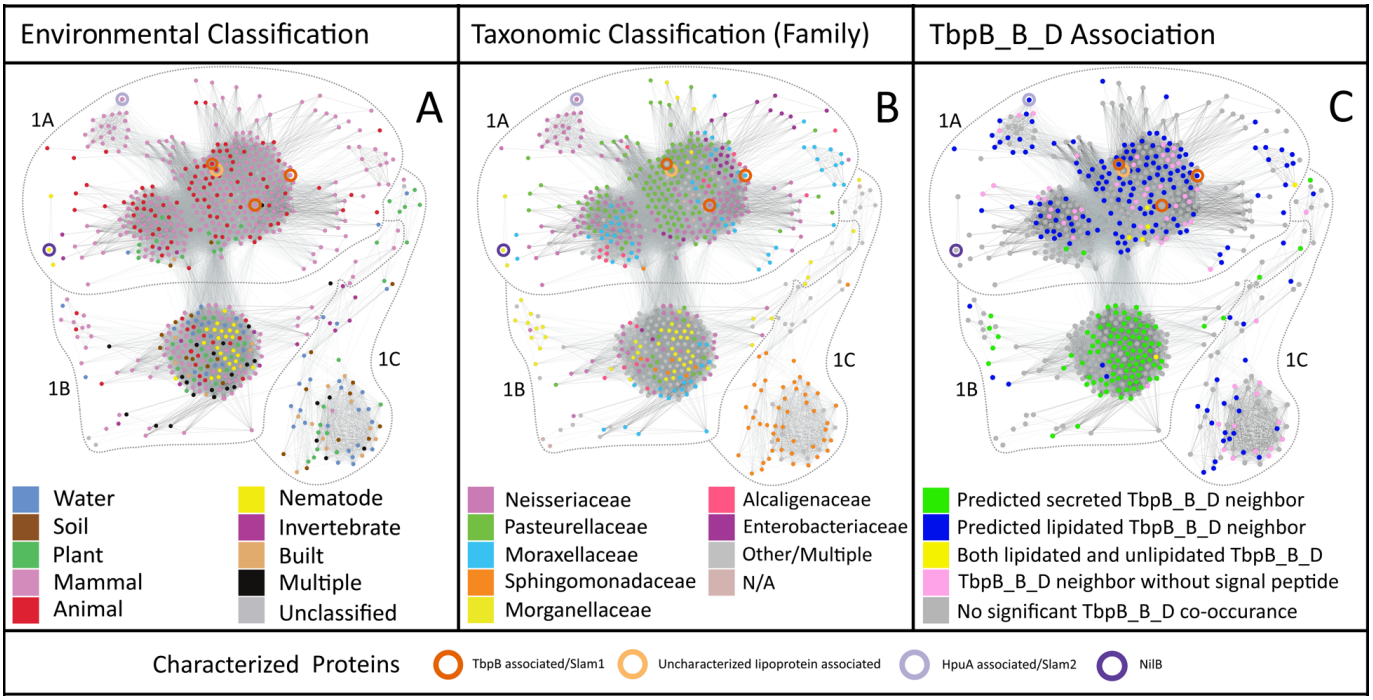

Figure 3. Cluster 1 of the DUF560 sequence similarity network contains a mix of host associated and environmental isolates and its subclusters coincide with the presence of TbpB_B_D domain encoding genes in a genome neighborhood network.

DUF560 sequence similarity network Cluster 1 nodes were separated and the node positioning was optimized using the Fruchteman-Reingold algorithm [75]. The plots were annotated according to color legends at the bottom of each panel to categorize nodes according to the A) environmental origin(s) or B) taxonomic class of the isolates encoding the node homolog(s), or C) whether the node homolog(s) co-occur with a TbpB_B_D-domain encoding gene in the genome neighborhood network. Nodes that contain DUF560 homologs that have been functionally characterized are highlighted within the networks using colored circles according to the legend at the bottom of the figure.

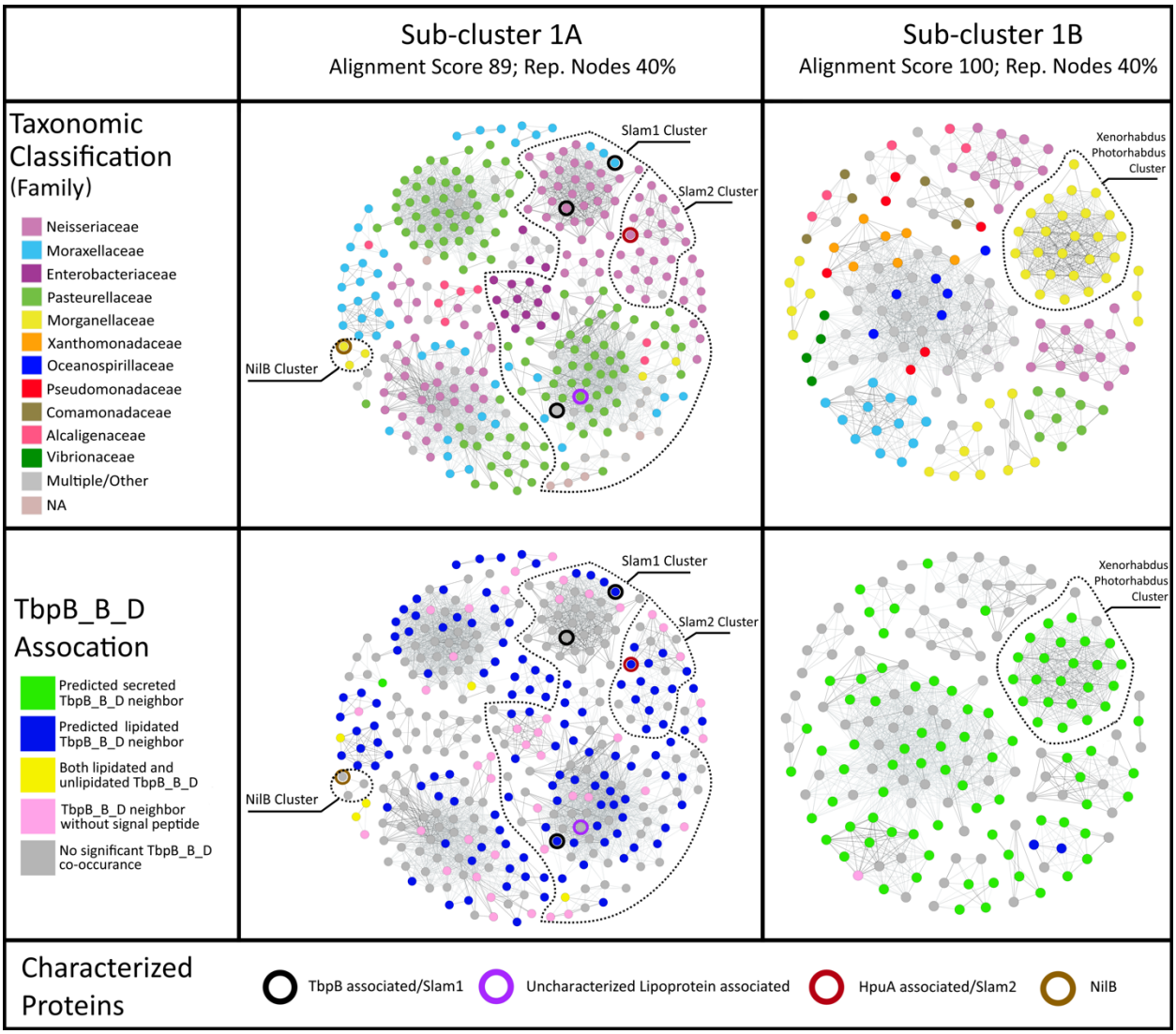

Figure 4: Increased stringency separates subcluster 1A and 1B into more functional groups showcased by clustering of Slam homologs

A series of very stringent EFI-EST sequence similarity networks highlights detail in Cluster 1 of the DUF560 homologs. Node positioning was optimized using the Fruchteman-Reingold algorithm [75]. Dotted lines indicate hypothetical functional clusters. Circled nodes indicate proteins which have been molecularly characterized. Sub-clusters 1A and 1B were analyzed separately to allow fine tuning of alignment score (89 and 100 respectively). Networks were color coded to separately display both taxonomic categories and association with TbpB_B_D proteins.

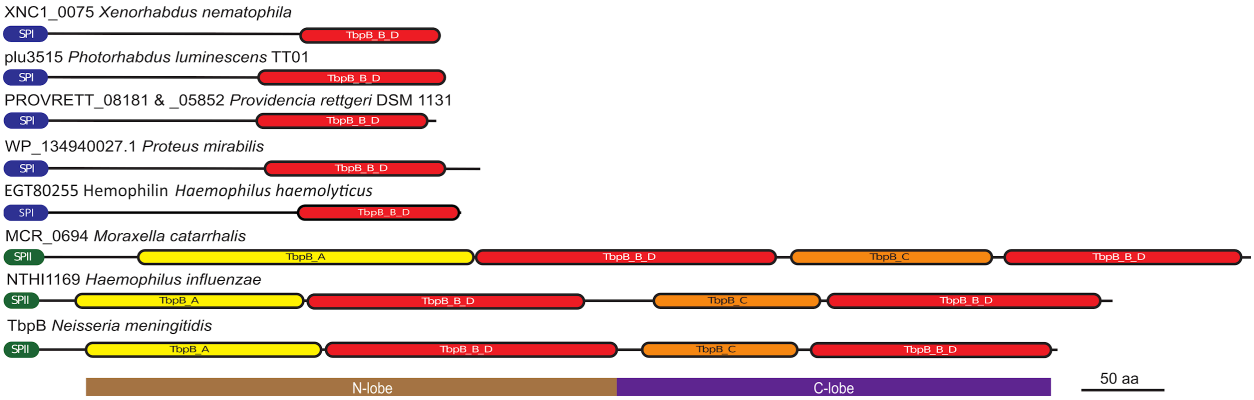

Figure 5: TbpB domain organization for TbpB_B_D$_{sol}$ vs. Slam-dependent TbpB$_{lip}$ homologs

The schematic diagram shows features found in select TbpB_B_D-domain containing proteins encoded on the same locus as a SPAM1B or SPAM1A (Slam1) proteins. SPAM1B-associated proteins from *X. nematophila, P. luminescens, P. rettgeri, P. mirabilis,* and *H. haemolyticus* have signal-peptidase 1 (SPI) type signal sequences, lack annotated “handle” domains, and have a single TbpB_B_D domain. SPAM1A (Slam1)-associated proteins from *M. catarrhalis, H. influenzae, and N. meningitidis* have signal-peptidase 2 (SPII) type signal sequences, indicative of lipidated proteins, N-lobe and C-lobe handle domains (TbpB_A and TbpB_C, respectively) and two TbpB_B_D domains, one each in the N- and C-lobes. Domains are indicated by color-coded rods: SPI (blue); SPII (green); TbpB_B_D (red); TbpB_A (yellow); TbpB_C (orange); N-lobe (brown); C-lobe (purple).

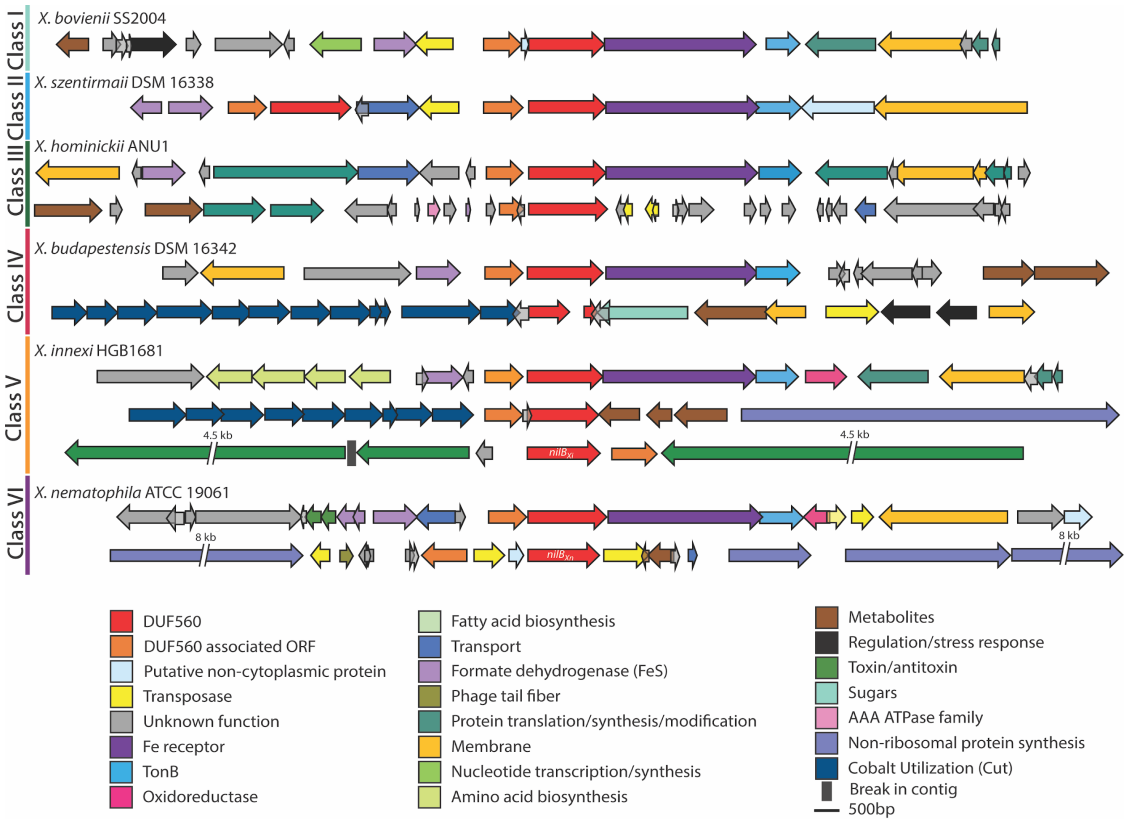

Figure 6: Representative genomes of *Xenorhabdus* SPAM Classes I-VI

Schematic diagrams of *Xenorhabdus* SPAM-encoding genomic loci representing each of the six classes defined in the text. Class is indicated on the left with a vertical line and color coded as in Table 1 and Table S7: Class I (light green); Class II (light blue); Class III (dark green); Class IV (red); Class V (orange); Class VI (purple). One species from each of the classes was selected for presentation: *X. bovienii* SS2004 for Class I, *X. szentirmaii* DSM 16338 for Class II, *X. hominickii* ANU1, for Class III, *X. budapestensis* DSM 16342 for Class IV, *X. innexi* HGB1681 for Class V, and *X. nematophila* ATCC 19061 for Class VI. Box arrows represent open reading frames (ORFs), which are color coded according to predicted annotated function as indicated by the embedded legend. In all cases the DUF560 (SPAM) homolog is shown in red and the DUF560 associated ORF (putative SPAM target) is shown in orange. To more easily visualize loci, large ORFs were not presented in their entirety and the length of the gap is indicated above the break line shown within such ORFs.

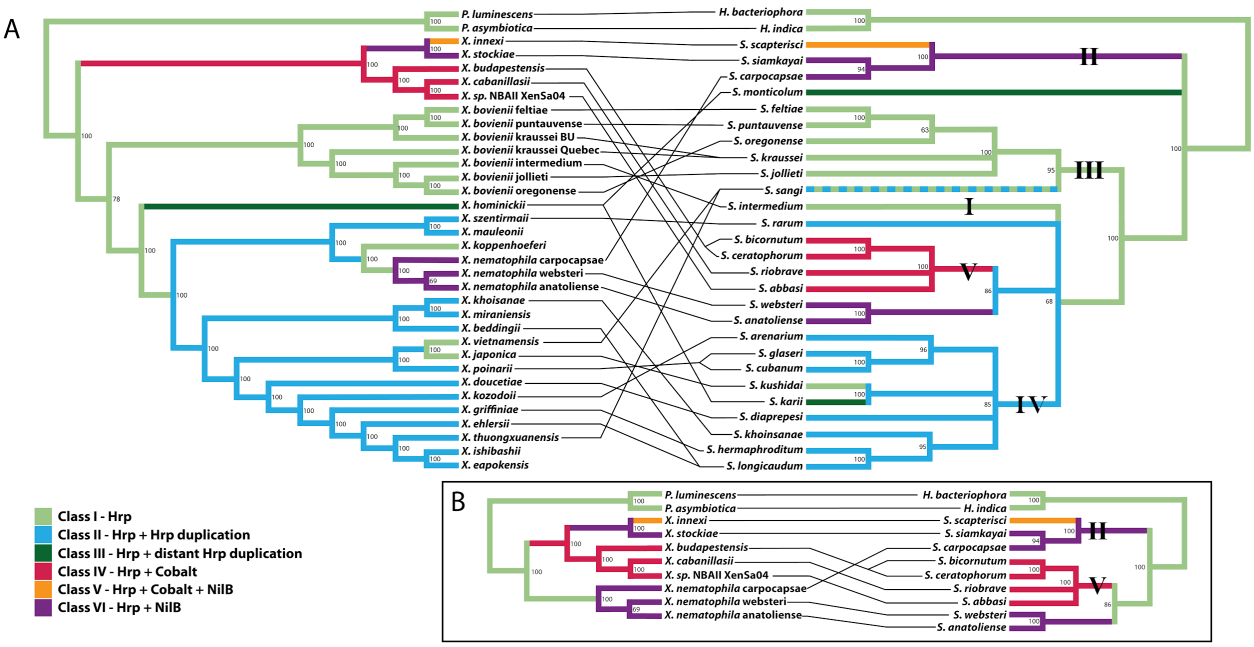

Figure 7: Phylogenies of *Xenorhabdus* and *Steinernema* color coded based on *Xenorhabdus* DUF560 Class

Co-phylogeny of nematode species and their colonizing mutualist bacteria. Numbers on branches indicate bootstrap support values. Bootstrap values below 60% were contracted. Lines connecting the phylogenies indicate mutualist pairs. Large roman numerals highlight the 5 *Steinernema* clades originally described by Spridinov *et al.*, 2004 [94], [54]. For bacteria, colored overlays indicate the complement of SPAMs they possess. For nematodes, colored overlays indicate the complement of SPAMs possessed by their colonizing bacterial partner. Color-coding is as presented in Table 1 and Table S7: Class I (light green); Class II (light blue); Class III (dark green); Class IV (red); Class V (orange); Class VI (purple).

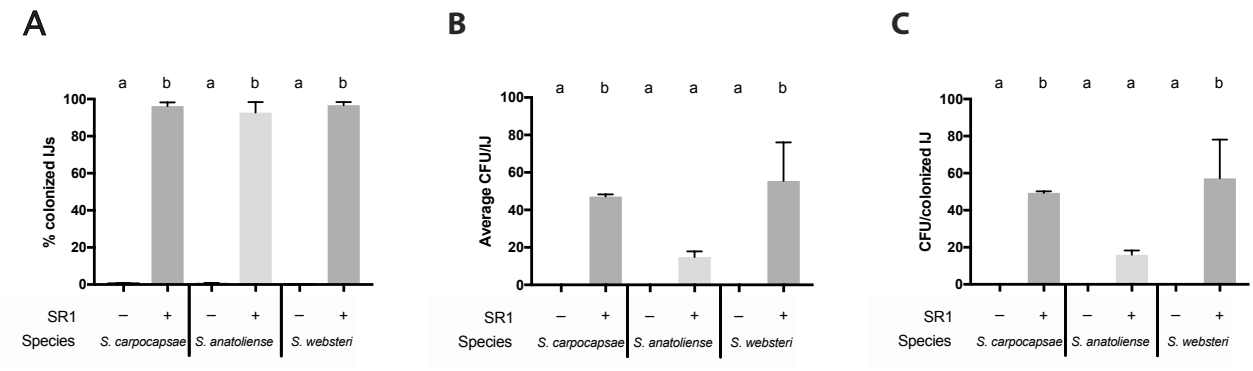

Figure 8: SR1 (encoding NilB and NilC) is necessary for IJ receptacle colonization of *S. anatoliense, S. carpocapsae,* and *S. websteri*

Three *Steinernema* species: *S. carpocapsae, S. anatoliense,* and *S. websteri* were cultivated on lawns of either *X. nematophila* ATCC19061 SR1 mutant (SR1 -) (HGB1495 *X. nematophila* Δ*nilR* ΔSR1 *kefA:GFP att*Tn*7::*empty vector) or an SR1 complemented strain (SR1 +) (HGB1496 *X. nematophila* Δ*nilR* Δ*SR1 kefA:GFP att*Tn*7*::SR1). The resulting progeny infective juveniles (IJs) were monitored for colonization by these bacteria either by microscopy (A) or by surface sterilization, sonication, and plating for average CFU/IJ (B). The average CFU per colonized IJ was then calculated by multiplying the average CFU/IJ (shown in panel B) by the fraction of nematodes that were colonized in that treatment (shown in panel A). Treatments with the same letters within each panel are not significantly different from each other when analyzed with one-way ANOVA with multiple comparisons using Tukey's post-hoc test.

Table 1. Loci and nomenclature for *Xenorhabdus* SPAM classes I-VI and their predicted targets

SPAM Cluster[b]	1B				1A
Locus[c]	Hrp1	Hrp2	Tn	Cut	Pax or MARTX
Class I	SPAM1B$_{Hrp1}$ TbpB_B_D$_{solHrp1}$				
Class II	SPAM1B$_{Hrp1}$ TbpB_B_D$_{solHrp1}$	SPAM1B$_{Hrp2}$ TbpB_B_D$_{solHrp2}$			
Class III	SPAM1B$_{Hrp1}$ TbpB_B_D$_{solHrp1}$		SPAM1B$_{Tn}$ TbpB_B_D$_{solTn}$		
Class IV	SPAM1B$_{Hrp1}$ TbpB_B_D$_{solHrp1}$			SPAM1B$_{Cut}$ TbpB_B_D$_{solCut}$	
Class V	SPAM1B$_{Hrp1}$ TbpB_B_D$_{solHrp1}$			SPAM1B$_{Cut}$ TbpB_B_D$_{solCut}$	SPAM1A$_{nilB}$ NilC$_{lip}$
Class VI	SPAM1B$_{Hrp1}$ TbpB_B_D$_{solHrp1}$				SPAM1A$_{nilB}$ NilC$_{lip}$

[a]Class I (light green); Class II (light blue); Class III (dark green); Class IV (red); Class V (orange); Class VI (purple).

[b]SPAM Cluster 1A or 1B as indicated by the network analysis shown in Fig. 3.

[c]Genomic location of the SPAM protein and predicted target; see text for details.

References

1. Wendlandt CE, Regus, J. U., Gano-Cohen, K. A., Hollowell, A. C., Quides, K. W., Lyu, J. Y., Adinata, E. S., Sachs, J. L. (2019) Host investment into symbiosis varies among genotypes of the legume acmispon strigosus, but host sanctions are uniform. New Phytol 221: 446-458.

2. Gruntenko NE, Ilinsky, Y. Y., Adonyeva, N. V., Burdina, E. V., Bykov, R. A., Menshanov, P. N., Rauschenbach, I. U. (2017) Various wolbachia genotypes differently influence host drosophila dopamine metabolism and survival under heat stress conditions. BMC Evol Biol 17.

3. Rivera-Pérez JI, González, A. A., Toranzos, G. A. (2017) From evolutionary advantage to disease agents: Forensic reevaluation of host-microbe interactions and pathogenicity. Microbiol Spectr 5: 10.1128/microbiolspec.EMF-0009-2016.

4. Nanda AM, Thormann, K., Frunzke, J. (2015) Impact of spontaneous prophage induction on the fitness of bacterial populations and host-microbe interactions. J Bacteriol 197: 410–419.

5. Ford SA, Kao, D., Williams, D., King, K. C. (2016) Microbe-mediated host defence drives the evolution of reduced pathogen virulence. Nat Commun 7.

6. Kopac SM, Klassen, J. L. (2016) Can they make it on their own? Hosts, microbes, and the holobiont niche Front Microbiol 7.

7. Noinaj N, Guillier, M., Barnard, T. J., Buchanan, S. K. (2010) Tonb-dependent transporters: Regulation, structure, and function. Annu Rev Microbiol 64: 43-60.

8. Chu BC, Garcia-Herrero, A., Johanson, T. H., Krewulak, K. D., Lau, C. K., Peacock, R. S., Slavinskaya, Z., Vogel, H. J. (2010) Siderophore uptake in bacteria and the battle for iron with the host; a bird's eye view. Biometals 23: 601-611.

9. Schweppe DK, Harding, C., Chavez, J. D., Wu, X., Ramage, E., Singh, P. K., Manoil, C., Bruce, J. E. (2015) Host-microbe protein interactions during bacterial infection. Chem Biol 22: 1521-1530.

10. Chavez-Dozal AA, Gorman, C., Lostroh, C. P., Nishiguchi, M. K. (2014) Gene-swapping mediates host specificity among symbiotic bacteria in a beneficial symbiosis. PLoS ONE 9.

11. Green ER, Mecsas, J. (2016) Bacterial secretion systems – an overview. Microbiol Spectr 4.

12. Rapisarda C, Tassinari, M., Gubellini, F., Fronzes, R. (2018) Using cryo-em to investigate bacterial secretion systems. Annu Rev Microbiol 72: 231-254.

13. Hooda Y, Lai, C. C., Judd, A., Buckwalter, C. M., Shin, H. E., Gray-Owen, S. D., Moraes, T. F. (2016) Slam is an outer membrane protein that is required for the surface display of lipidated virulence factors in neisseria. Nat Microbiol 1: 16009.

14. Hooda Y, Lai, C. C. L., Moraes, T. F. (2017) Identification of a large family of slam-dependent surface lipoproteins in gram-negative bacteria. Front Cell Infect Microbiol 7: 207.

15. Okuda S, Tokuda, H. (2011) Lipoprotein sorting in bacteria. Annu Rev Microbiol 65: 239-259.

16. Heungens K, Cowles, C. E., Goodrich-Blair, H. (2002) Identification of *xenorhabdus nematophila* genes required for mutualistic colonization of *steinernema carpocapsae* nematodes. Mol Microbiol 45: 1337-1353.

17. Bhasin A, Chaston, J. M., Goodrich-Blair, H. (2012) Mutational analyses reveal overall topology and functional regions of nilb, a bacterial outer membrane protein required for host association in a model of animal-microbe mutualism. J Bacteriol 194: 1763-1776.

18. Poinar GO (1966) The presence of *achromobacter nematophilus* in the infective stage of a *neoaplectana* sp. (steinernematidae: Nematoda). Nematologica 12: 105-108.

19. Akhurst RJ (1983) Neoplectana species: Specificity of association with bacteria of the genus xenorhabdus. Exp Parasitol 55: 258-263.

20. Adeolu M, Alnajar, S., Naushad, S., Gupta, R. S. (2016) Genome-based phylogeny and taxonomy of the 'enterobacteriales': Proposal for enterobacterales ord. Nov. Divided into the families enterobacteriaceae, erwiniaceae fam. Nov., pectobacteriaceae fam. Nov., yersiniaceae fam. Nov., hafniaceae fam. Nov., morganellaceae fam. Nov., and budviciaceae fam. Nov. Int J Syst Evol Microbiol 66: 5575-5599.

21. Murfin KE, Chaston, J., Goodrich-Blair, H. (2012) Visualizing bacteria in nematodes using fluorescence microscopy. Journal of visualized experiments : JoVE 68: e4298.

22. Yadav S, Shokal, U., Forst, S., Eleftherianos, I. (2015) An improved method for generating axenic entomopathogenic nematodes. . BMC Res Notes 8.

23. Cowles CE, Goodrich-Blair, H. (2008) The *xenorhabdus nematophila nilabc* genes confer the ability of *xenorhabdus* spp. To colonize *steinernema carpocapsae* nematodes. J Bacteriol 190: 4121-4128.

24. Bird AF, Akhurst, R. J. (1983) The nature of the intestinal vesicle in nematodes of the family steinernematidae. Int J Parasitol 13: 599-606.

25. Martens EC, Heungens, K., Goodrich-Blair, H. (2003) Early colonization events in the mutualistic association between *steinernema carpocapsae* nematodes and *xenorhabdus nematophila* bacteria. J Bacteriol 185: 3147-3154.

26. Snyder HA, Stock, S. P., Kim, S. K., Flores-Lara, Y., Forst, S. (2007) New insights into the colonization and release process of *xenorhabdus nematophila* and the morphology and ultrastructure of the bacterial receptacle of its nematode host, *steinernema carpocapsae*. Appl Environ Microbiol 73: 5338-5346.

27. Chaston JM, Murfin, K. E., Heath-Heckman, E. A., Goodrich-Blair, H. (2013) Previously unrecognized stages of species-specific colonization in the mutualism between *xenorhabdus* bacteria and *steinernema* nematodes. Cellular microbiology 15: 1545-1559.

28. Kampfer P, Tobias, N. J., Ke, L. P., Bode, H. B., Glaeser, S. P. (2017) Xenorhabdus thuongxuanensis sp. Nov. And xenorhabdus eapokensis sp. Nov., isolated from steinernema species. Int J Syst Evol Microbiol 67: 1107-1114.

29. Blatch GL, Lassle M (1999) The tetratricopeptide repeat: A structural motif mediating protein-protein interactions. Bioessays 21: 932-939.

30. Cowles CE, Goodrich-Blair, H. (2004) Characterization of a lipoprotein, nilc, required by *xenorhabdus nematophila* for mutualism with its nematode host. Mol Microbiol 54: 464-477.

31. El-Gebali S, Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., Finn, R. D. (2019) The pfam protein families database in 2019. Nucleic Acids Res 47: D427–D432.

32. Adebali O, Zhulin, I. B. (2017) Aquerium: A web application for comparative exploration of domain-based protein occurrences on the taxonomically clustered genome tree. Proteins 85: 72-77.

33. Gerlt JA, Bouvier, J. T., Davidson, D. B., Imker, H. J., Sadkhin, B., Slater, D. R., Whalen, K. L. (2015) Enzyme function initiative-enzyme similarity tool (efi-est): A web tool for generating protein sequence similarity networks. Biochemica et Biophysica Acta 1854: 1019-1037.

34. Gerlt JA (2017) Genomic enzymology: Web tools for leveraging protein family sequence−function space and genome context to discover novel functions. Biochemistry 56: 4293-4308.

35. Zallot R, Oberg, N. O., Gerlt, J. A. (2018) 'Democratized' genomic enzymology web tools for functional assignment. Curr Opin Chem Biol 47: 77-85.

36. Zallot R, Oberg N, Gerlt JA (2019) The efi web resource for genomic enzymology tools: Leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. Biochemistry 58: 4169-4182.

37. Mickelsen PA, Sparling, P. F. (1981) Ability of neisseria gonorrhoeae, neisseria meningitidis, and commensal neisseria species to obtain iron from transferrin and iron compounds. Infect Immun 33: 555-564.

38. Cornelissen CN, Anderson, J. E., Boulton, I. C., Sparling, P. F. (2000) Antigenic and sequence diversity in gonococcal transferrin-binding protein a. Infect Immun 68: 4725-4735.

39. Choe K, Mitchell Laboratory. (2018) Rodeo (rapid orf description & evaluation online) is an algorithm to help biosynthetic gene cluster (bgc) analysis, with an emphasis on ribosomal natural product (ripp) discovery. University of Illinois at Urbana-Champaign. pp. http://ripp.rodeo/advanced.html.

40. Armenteros JJA, Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., von Heijne, G., Nielsen, H. (2019) Signalp 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 37: 420-423.

41. Stevenson P, Williams, P., Griffiths, E. (1992) Common antigenic domains in transferrin-binding protein 2 of neisseria meningitidis, neisseria gonorrhoeae, and haemophilus influenzae type b. Infect Immun 60: 2391-2396.

42. Adamiak P, Calmettes, C., Moraes, T. F., Schryvers, A. B. (2015) Patterns of structural and sequence variation within isotype lineages of the neisseria meningitidis transferrin receptor system. Microbiology Open 4: 491-504.

43. Latham RD, Torrado M, Atto B, Walshe JL, Wilson R, Guss JM, et al. (2019) A heme-binding protein produced by haemophilus haemolyticus inhibits non-typeable haemophilus influenzae. Mol Microbiol.

44. Kelley LA, Sternberg, M. J. (2009) Protein structure prediction on the web: A case study using the phyre server. Nat Protoc 4: 363-371.

45. Kelley LA, Mezulis, S., Yates, C. M., Wass, M. N., Sternberg, M. J. E. (2015) The phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10: 845-858.

46. Moraes TF, Yu, R. H., Strynadka, N. C. J., Schryvers, A. B. (2009) Insights into the bacterial transferrin receptor:The structure of transferrin-binding protein bfromactinobacillus pleuropneumoniae. Mol Cell 35: 523-533.

47. Yukl ET, Jepkorir, G., Alontaga, A. Y., Pautsch, L., Rodriguez, J. C., Rivera, M., Moënne-Loccoz, P. (2010) Kinetic and spectroscopic studies of hemin acquisition in the hemophore hasap from pseudomonas aeruginosa. Biochemistry 49: 6646-6654.

48. Kumar R, Qi, Y., Matsumura, H., Lovell, S., Yao, H., Battaile, K. P., Im, W., Moënne-Loccoz, P., Rivera, M. (2016) Replacing arginine 33 for alanine in the hemophore hasa from pseudomonas aeruginosa causes closure of the h32 loop in the apo-protein. Biochemistry 55: 2622-2631.

49. Dent AT, Mouriño, S., Huang, W., Wilks, A. (2018) Post-transcriptional regulation of the pseudomonas aeruginosa heme assimilation system (has) fine-tunes extracellular heme sensing. J Biol Chem 294: 2771-2785.

50. Ostberg KL, DeRocco, A. J., Mistry, S. D., Dickinson, M. K., Cornelissen, C. N. (2013) Conserved regions of gonococcal tbpb are critical for surface exposure and transferrin iron utilization. Infect Immun 81: 3442-3450.

51. Quillin SJ, Hockenberry, A. J., Jewett, M. C., Seiferta, H. S. (2018) Neisseria gonorrhoeae exposed to sublethal levels of hydrogen peroxide mounts a complex transcriptional response. mSystems 3: e00156-00118.

52. Jackson LA, Ducey, T. F., Day, M. W., Zaitshik, J. B., Orvis, J., Dyer, D. W. (2010) Transcriptional and functional analysis of the neisseria gonorrhoeae fur regulon. J Bacteriol 192: 77-85.

53. Stohl EA, Criss, A. K., Seifert, H. S. (2005) The transcriptome response of neisseria gonorrhoeae to hydrogen peroxide reveals genes with previously uncharacterized roles in oxidative damage protection. Mol Microbiol 58: 520-532.

54. Lee MM, Stock, S. P. (2010) A multigene approach for assessing evolutionary relationships of *xenorhabdus* spp. (gamma-proteobacteria), the bacterial symbionts of entomopathogenic *steinernema* nematodes. J Invertebr Pathol 104: 67-74.

55. Mcginnis S, Madden, T. L. (2004) Blast: At the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 32: W20-W25.

56. Boratyn GM, Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L., Matten, W. T., McGinnis, S. D., Merezhuk, Y., Raytselis, Y., Sayers, E. W., Tao, T., Ye, J., Zaretskaya, I. (2013) Blast: A more efficient report with usability improvements. Nucleic Acids Res 41: W29-W33.

57. Latreille P, Norton, S., Goldman, B. S., Henkhaus, J., Miller, N., Barbazuk, B., Bode, H. B., Darby, C., Du, Z., Forst, S., Gaudriault, S., Goodner, B., Goodrich-Blair, H., Slater, S. (2007) Optical mapping as a routine tool in bacterial genome sequencing. BMC Genomics 8: 321.

58. Chaston JM, Suen, G., Tucker, S. L., Andersen, A. W., Bhasin, A., Bode, E., Bode, H. B., Brachmann, A. O., Cowles, C. E., Cowles, K. N., Darby, C., de Leon, L., Drace, K., Du, Z., Givaudan, A., Herbert Tran, E. E., Jewell, K. A., et. al. (2011) The entomopathogenic bacterial endosymbionts *xenorhabdus* and *photorhabdus*: Convergent lifestyles from divergent genomes. PLoS ONE 6: e27909.

59. Ogier JC, Pages, S., Bisch, G., Chiapello, H., Medigue, C., Rouy, Z., Teyssier, C., Vincent, S., Tailliez, P., Givaudan, A., Gaudriault, S. (2014) Attenuated virulence and genomic reductive evolution in the entomopathogenic bacterial symbiont species, *xenorhabdus poinarii*. Genome Biol Evol 6: 1495-1513.

60. Gualtieri M, Aumelas, A., Thaler, J. O. (2009) Identification of a new antimicrobial lysine-rich cyclolipopeptide family from *xenorhabdus nematophila*. J Antibiot 62: 295-302.

61. Dreyer J, Malan, A. P., Dicks, L. M. T. (2018) Bacteria of the genus xenorhabdus, a novel source of bioactive compounds. Front Microbiol 9: 3177.

62. Kim IH, Ensign, J., Kim, D. Y., Jung, H. Y., Kim, N. R., Choi, B. H., Park, S. M., Lan, Q., Goodman, W. G. (2017) Specificity and putative mode of action of a mosquito larvicidal toxin from the bacterium *xenorhabdus innexi*. J Invertebr Pathol 149: 21-28.

63. Goodrich-Blair H, Clarke DJ (2007) Mutualism and pathogenesis in *xenorhabdus* and *photorhabdus*: Two roads to the same destination. Molecular Microbiology 64: 260-268.

64. Dillman AR, Macchietto, M., Porter, C. F., Rogers, A., Williams, B., Antoshechkin, I., Lee, M. M., Goodwin, Z., Lu, X., Lewis, E. E., Goodrich-Blair, H., Stock, S. P., Adams, B. J., Sternberg, P. W., Mortazavi, A. (2015) Comparative genomics of *steinernema* reveals deeply conserved gene regulatory networks. Genome Biol 16: 200.

65. Veesenmeyer JL, Andersen, A. W., Lu, X., Hussa, E. A., Murfin, K. E., Chaston, J. M., Dillman, A. R., Wassarman, K. M., Sternberg, P. W., Goodrich-Blair, H. (2014) Nild crispr rna contributes to *xenorhabdus nematophila* colonization of symbiotic host nematodes. Mol Microbiol 93: 1026-1042.

66. Pogoutse AK, Moraes, T. F. (2017) Iron acquisition through the bacterial transferrin receptor. Crit Rev Biochem Mol Biol 52: 314-326.

67. Tonjum T, van Putten, J. (2017) Neisseria. In: Johathan Cohen WP, Steven Opal, editor. Infectious diseases 4th ed Elsevier.

68. Yoshiga T, Georgieva, T., Dunkov, B. C., Harizanova, N., Ralchev, K., Law, J. H. (1999) Drosophila melanogaster transferrin: Cloning, deduced protein sequence, expression during the life cycle, gene localization and up-regulation on bacterial infection. Eur J Biochem 260: 414-420.

69. Lambert LA, Perri, H., Halbrooks, P. J., Mason, A. B. (2005) Evolution of the transferrin family: Conservation of residuesassociated with iron and anion binding. Comp Biochem Physiol: 129-141.

70. Geiser DL, Winzerling, J. J. (2012) Insect transferrins: Multifunctional proteins. Biochemica et Biophysica Acta 1820: 437-451.

71. Brummett LM, Kanost, M. R., Gorman, M. J. (2017) The immune properties of manduca sexta transferrin. Insect Biochem Mol Biol 81: 1-9.

72. Sheldon JR, Skaar, E. P. (2019) Metals as phagocyte antimicrobial effectors. Curr Opin Immunol 60: 1-9.

73. Shannon P, Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498-2504.

74. Bastian M, Heymann, S., Jacomy, M. (2009) Gephi : An open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

75. Fruchterman TMJ, Reingold, E. M. (1991) Graph drawing by force-directed placement. Softw Pract Exp 21: 1129-1164.

76. Vallenet D, Calteau, A., Cruveiller, S., Gachet, M., Lajus, A., Josso, A., Mercier, J., Renaux, A., Rollin, J., Rouy, Z., Roche, D., Scarpelli, C., Medigue, C. (2017) Microscope in 2017: An expanding and evolving integrated resource for community expertise of microbial genomes. Nucleic Acids Res 45: D517–D528.

77. Nielsen H, Engelbrecht, J., Brunak, S., von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 10: 1-6.

78. Miller JH (1972) Experiments in molecular genetics. Cold Spring Harbor, N. Y: Cold Spring Harbor Laboratory Press. 466 p.

79. Xu J, Hurlbert, R.E. (1990) Toxicity of irradiated media for *xenorhabdus* spp. Appl Environ Microbiol 56: 815-818.

80. Vallenet D, Belda, E., Calteau, A., Cruveiller, S., Engelen, S., Lajus, A., Le Fevre, F., Longin, C., Mornico, D., Roche, D., Rouy, Z., Salvignol, G., Scarpelli, C., Smith, A. A. T., Weiman, M., Medigue, C. (2012) Microscope—an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. Nucleic Acids Res 41: D636-D647.

81. Cock PJA, Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M. J. L. (2009) Biopython: Freely available python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422-1423.

82. Edgar RC (2004) Muscle: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.

83. Vaidya G, Lohman, D. J., Meier, R. (2010) Sequencematrix: Concatenation software for the fast assembly of multi-gene datasets with character set and codon information. Cladistics 27: 171-180.

84. Capella-Gutiérrez S, Silla-Martínez, J. M., Gabaldón, T. (2009) Trimal: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972-1973.

85. Darriba D, Taboada, G. L., Doallo, R., Posada, D. (2012) Jmodeltest 2: More models, new heuristics and highperformance computing. Nat Methods 9: 772.

86. Guindon S, Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood Syst Biol 52: 696-704.

87. Stamatakis A (2014) Raxml version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312-1313.

88. Huson DH, Scornavacca, C. (2012) Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Syst Biol 61: 1061-1067.

89. Ronquist F, Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M. A., Huelsenbeck, J. P. (2012) Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61: 539-542.

90. Ayres DL, Darling, A., Zwickl, D. J., Beerli, P., Holder, M. T., Lewis, P. O., Huelsenbeck, J. O., Ronquist, F., Swofford, D. L., Cummings, M. P., Rambaut, A., Suchard, M. A. (2011) Beagle: An application programming interface and high-performance computing library for statistical phylogenetics. Syst Biol 61: 170-173.

91. Altekar G, Dwarkadas, S., Huelsenbeck, J. P., Ronquist, F. (2004) Parallel metropolis coupled markov chain monte carlo for bayesian phylogenetic inference Bioinformatics 20: 407-415.

92. Miller MA, Pfeiffer, W., Schwartz, T. Creating the cipres science gateway for inference of large phylogenetic trees; 14 Nov. 2010; New Orleans, LA. Proceedings of the Gateway Computing Environments Workshop (GCE). pp. 1-8.

93. Rambaut A (2018) Figtree v1.4.4 pp. http://tree.bio.ed.ac.uk/software/figtree/.

94. Spiridonov SE, Reid, A.P., Podrucka, K., Subbotin, S.A., Moens, M. (2004) Phylogenetic relationships within the genus *steinernema* (nematoda: Rhabditida) as inferred from analyses of sequences of the its1-5.8s-its2 region of rdna and morphological features. Nematology 6: 547-566.