The Phage-shock-protein (PSP) Envelope Stress Response: Discovery of Novel Partners and Evolutionary History

Janani Ravi[1,2]*, Vivek Anantharaman[3], Samuel Zorn Chen[1], Pratik Datta[2], L Aravind[3*], Maria Laura Gennaro[2]*.

[1]Pathobiology and Diagnostic Investigation, Michigan State University, East Lansing, MI, USA; [2]Public Health Research Institute, Rutgers University, Newark, NJ, USA; [3]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, USA. *Corresponding authors. janani@msu.edu; aravind@nih.gov; marila.gennaro@rutgers.edu

Running title: *Psp evolution across the tree of life*

Abstract

The phage shock protein (PSP) systems orchestrate a conserved stress response function by stabilizing the cell membrane and protecting bacteria from envelope stress. The full repertoire of PSP components remains poorly characterized. We combined comparative genomics and protein sequence-structure-function analyses to systematically identify homologs, phyletic patterns, domain architectures, and gene neighborhoods to trace the evolution of PSP components across the tree of life. This approach showed that the core component PspA/Snf7 (Psp/ESCRT systems) was present in the Last Universal Common Ancestor and that different clades co-opted a diverse range of partners to constitute distinct PSP systems. We identified several novel partners of the PSP system: (i) the Toastrack domain, which likely facilitates assembling diverse sub-membrane stress-sensing and signaling complexes, (ii) the newly-defined HAAS–PadR-like transcription regulator pair system, and (iii) multiple independent associations with ATPase or CesT/Tir-like chaperones, and Band-7 domain proteins that likely mediate sub-membrane dynamics. Our work also uncovered links between the PSP components and diverse SHOCT-like domains, suggesting a role in assembling membrane-associated complexes of proteins with disparate biochemical functions. Tracing the evolution of Psp cognate proteins provides new insights into the functions of the system and helps predict previously uncharacterized, often lineage-specific, membrane-dynamics and stress-response systems. The conservation of PSP systems across bacterial phyla emphasizes the importance of this stress response system in prokaryotes, while its modular diversity in various lineages indicates the emergence of lineage-specific cell-envelope structures, lifestyles, and adaptation mechanisms. The results can be accessed at https://jravilab.shinyapps.io/psp-evolution.

Keywords

Phage shock protein response | Psp | Envelope stress response | Phylogeny and evolution | Domain architectures | Genomic contexts and neighborhoods

Introduction

Membranes are complex dynamical structures enclosing all living cells, composed of bilayer and non-bilayer lipids and proteins (1). These membrane components partake in many critical dynamic processes, including membrane biogenesis, cell division, cell-shape maintenance, transport of small molecules, cell-cell and cell-environment signaling, maintenance of proton motive force (PMF), and processes like motility involving cytoskeletal proteins (2). Despite dramatic variation in composition and structure, most membrane functions are conserved across the tree of life (1). Cellular membranes continually adapt, particularly in unicellular organisms, in response to external stresses (3–7). Membrane damage can ultimately result in cell death; therefore, a specialized suite of lipids and proteins participate in maintaining membrane function under stress (8, 9). Among the mechanisms involved in maintaining membrane integrity are the multiprotein bacterial Phage Shock Protein (PSP) system and the archaeo-eukaryotic ESCRT system (10–19). Fully understanding the phenotypic convergence (20) of membrane maintenance mechanisms emanating from systems such as PSP requires mapping out their evolutionary history.

The PSP system is centered on the peripheral membrane protein PspA and the homologous plant protein Vipp1 (11–16, 21–29). Structurally, both PspA and Vipp1 contain the PspA domain (Pfam: *PspA_IM30*, *IM30*, standing for inner membrane 30KDa domain, henceforth called PspA*)* with a predominantly coiled-coil structure (16, 17, 30–32) [**Table S1**]. The *PspA_IM30* domain is found in bacteria, photosynthetic eukaryotes, and archaea (12, 15, 23, 33, 34) [**Fig. 1**]. PspA performs a stress sensing role (7, 14, 21, 35–37), while Vipp1 is involved in thylakoid biogenesis, vesicular transport and storage of lipids, and membrane shaping (16, 24, 25). However, the common functional denominator role of the PspA domain proteins is to facilitate membrane fusion (16) and prevent leakage of protons through the membrane (2, 21) in response to a wide range of surface stresses, including changes in membrane potential, Δ*pH* (2, 21) and PMF (38–41), (l)antibiotic stress (7), heat/osmotic shock (42) and mechanical stress via stored curvature elastic stress (17, 43–45), among others (42).

The three best-described PSP systems are from proteobacteria, firmicutes, and actinobacteria [**Fig. 1**]. Cognate PSP proteins serve distinct roles such as transcriptional regulation or membrane tethering and relocation. Few partner proteins, PspBC (39, 46–52) or the LiaRS two-component system (11, 37, 53–55), have been implicated in direct stress-sensing functions independent of PspA. However, the classical PSP system initially discovered in Proteobacteria (56) is not universal, *i.e.*, the system is not conserved across all bacterial/archaeal clades. Current studies have mostly focused on lineage-specific PSP systems (3, 12, 15, 23, 34, 57). How PspA has evolved to function in membrane stress-sensing across diverse species, and what the nature of its cognate partners in this pivotal role remain unclear.

We, therefore, conducted a comprehensive evolutionary analysis across the three superkingdoms of life (58, 59) to delineate all possible occurrences of PspA and its partners, their associated genomic

contexts, and phyletic spreads. To do so, we resolved Psp proteins into their constituent domains and used each of the individual domains as queries for our comparative genomics and phylogenetic analyses. This domain-centric approach is best suited to address the following questions regarding the evolutionary and functional significance of the PSP system: 1) Which are the most distinct/conserved PSP themes in terms of domain architectures and conserved gene neighborhoods? 2) Have the Psp proteins evolved functions beyond those discussed above? 3) Are there any other common themes that emerge from the characterization of the phyletic patterns of PSP systems? 4) Finally, do the gene neighborhoods and phyletic patterns of the Psp components imply greater independence of the individual subsystems than initially anticipated? The resulting findings presented here identify novel partners of the PSP envelope stress response system and throw light on both its evolution and function.

Results and Discussion

A major evolutionary and functional question regarding the PSP system concerns the apparent paradox of the universality of PspA while existing in distinct genomic neighborhoods (*i.e.,* working with different functional partners) (15, 28, 57) across the major branches of life. To tackle this question, we first sought to identify and characterize all close and remote homologs of PspA and cognate partner domains across the tree of life.

Using comparative genomics and a computational molecular evolutionary approach, we comprehensively identified Psp protein homologs, characterized their domain architectures, genomic contexts, and their phyletic spreads across the tree of life (see *Methods* for details). We present below the delineation of the core Psp proteins in proteobacteria, actinobacteria, and firmicutes, and describe their underlying domain architectures and phyletic spreads [**Fig. 1**]. Then, we outline the conservation and evolution of the central protein, PspA, spanning all the complete genomes across the tree of life [**Fig. 2**], followed by the characterization of the homologs in terms of domain architecture, genomic context, and phyletic spread [**Fig. 3**]. Finally, we report the findings from our in-depth analysis of each of the Psp cognate partners to identify novel components of the PSP system across the tree of life [**Fig. 4, 5**]. We then summarize our findings in the form of a proximity network of domains within the PSP system along with their phyletic spread and frequencies of occurrence across the tree of life [**Fig. 6**]. All our findings (data, results, and visual summaries) are available in an interactive and queryable web application https://jravilab.shinyapps.io/psp-evolution, referred to henceforth as the 'webapp.' The webapp can be queried with the protein accession numbers cited inline throughout the Results section.

The known components of Psp and their domain architectures

Proteobacteria: We initiated our analysis with the *E. coli* Psp operon that is widely conserved across other gammaproteobacteria (10, 12, 56, 60–62). PspA, PspB, and PspC are each composed of their namesake domains, which, respectively, span almost the entire protein length [**Fig. 1A**; **Table S1**]. In line with previous studies (63), we note that PspF, the transcriptional regulator of the Psp operon in *E.*

coli, contains an enhancer-binding NtrC-AAA (a specialized AAA-ATPase) domain in addition to the Fis-type HTH (helix-turn-helix) DNA-binding domain [**Fig. 1B**].

Actinobacteria: The protein central to the discovery of the Psp operon in *Mycobacterium tuberculosis* (actinobacteria) is PspA (encoded by *rv2744c*) (14), containing the PspA domain (<30% similar to the *E. coli* homolog) [**Fig. 1A**]. In contrast to the PspF transcriptional regulator in *E. coli*, the *M. tuberculosis* operon includes a transcription factor, ClgR, with a cro-like-HTH DNA-binding domain, unique to actinobacteria [**Fig. 1B**]. Moreover, the *M. tuberculosis* operon codes for an integral membrane protein, Rv2743c, distinct from *E. coli* PspC and PspB [**Fig. 1B**]. The last member of the four-gene actinobacterial operon, Rv2742c, is an uncharacterized protein with a "Domain of Unknown Function", DUF3046 [**Fig. 1B**; **Table S1**], which is required for the full activity of the PSP system (14, 15, 34). To remain consistent with the nomenclature of Psp cognate proteins, Rv2743c and Rv2742c were renamed as PspM and PspN, respectively [Datta *et al.,* unpublished].

Firmicutes: We then examined the domain architectures of the well-studied Lia operon in *B. subtilis* and found that, in line with recent studies (11, 54, 55), LiaI is a small TM protein; LiaF is a protein with N-terminal TM helices (DUF2157) and a C-terminal globular domain (DUF2154), and LiaG contains an uncharacterized domain DUF4097 [**Fig. 1B**]. Our analysis of these uncharacterized domains helped clarify relationships and correctly define two new domains of the PSP system: i) the **LiaI-LiaF-TM** sensory domain unifying 4 TM domains (4TM) and DUF2157 from LiaI/LiaF, and ii) **Toastrack** unifying the DUF4097/DUF2154/DUF2807 Pfam models (see **Text S1.1** for domain-specific details) [**Fig. 1B**; **Table S1**]. As documented previously in firmicutes (11, 37), the transcriptional response is elicited by the two-component system, LiaRS, which comprises of LiaR with a receiver domain and a winged HTH (wHTH) DNA-binding domain (REC–wHTH) and a sensor kinase LiaS with a TM region, intracellular HAMP and histidine kinase (HISKIN) signaling domains (64) [**Fig. 1B**; **Table S1**].

Psp components across the tree of life

PspA

To comprehensively identify all the close and remote homologs of PspA, we started with six distinct starting points as queries (from proteobacteria, firmicutes, actinobacteria, cyanobacteria, plants; see *Methods*). Consistent with previous evolutionary studies, we found that the PspA domain is present in most major bacterial and archaeal lineages in one or more copies per organism [**Fig. 1C**; webapp]. Within eukaryota, only the SAR group and Archaeplastida (the greater plant lineage) carry PspA/Vipp1 homologs [**Fig. 1C**]. Our analyses using multi-query sequence searches (see *Methods*) also recovered the eukaryotic Snf7 proteins as homologs [**Fig. 1B**]; these are part of the eukaryotic membrane-remodeling ESCRT systems (18, 65–69). Iterative searches with Snf7 also recovered several pan-eukaryotic (distinct from Vipp1) and archaeal homologs [**Fig. 1B**, **1C**; **Table S3**].

LiaI-LiaF-TM and Toastrack

Iterative domain-centric searches and multiple sequence alignments (MSA) (described in **Text S1.1;** 'MSA' tab, webapp) of LiaIGF proteins helped us identify that LiaI and LiaG are similar to the N- and C-terminal regions of LiaF. We define two novel domains, one with a single-stranded right-handed beta-helix-like structure, called the **Toastrack** domain (unifying DUF2154, DUF4097, DUF2807) and another with a 4TM region called **LiaI-LiaF-TM** (previously Toastrack_N) [**Fig. 1**; **Table S1**; detailed domain definitions in **Text S1.1**]. Toastrack-containing proteins are pan-bacterial with transfers to archaea and eukaryotes [**Fig. 1C**]. We also found a few eukaryotic genomes that encode Toastrack domains [**Fig. 1C**]. The Toastrack domain is frequently coupled to the LiaI-LiaF-TM, PspC, or a combination of both [**Fig. 4**; **Table S4**]. LiaI-LiaF-TM is found mainly in many bacterial clades with transfers to euryarchaeota [**Fig. 1B**].

PspC

PspC, an integral membrane protein ['MSA' tab, webapp] first identified in the proteobacterial PSP system, is critical in sensing the membrane stress response and restoring envelope integrity. Recent studies have shown that PspC may function independently of PspA in some bacterial species in response to membrane stressors such as secretin toxicity (47, 49, 61, 70, 71). We observed that the PspC domain is pan-bacterial and also present in a few archaeal clades [**Fig. 1C**]. Our analysis showed that PspC has two TM helices, the first being a cryptic TM. A conserved amino acid, R, is found between the two TM regions, suggesting a role in membrane-associated sensing of a stimulus. We also identified several novel domain architectures and genomic contexts of PspC (see below).

PspB

PspB is another integral membrane protein, often part of the proteobacterial Psp operon, and is implicated in stress sensing and virulence (61, 71, 72). PspB has an N-terminal anchoring TM helix followed by an α-helical cytoplasmic globular domain. While PspB is rarely found outside proteobacteria [**Fig. 1C**], we discovered previously unrecognized divergent PspB domains fused to PspC [see the section below on *PspC architectures*].

PspM/PspN

In accordance with our previous work, we found no discernible homologs for PspMN (and constituent domains) outside actinobacteria [**Fig. 1C**; webapp] (14, 15, 34). PspM (encoded by *rv2743c*), the corynebacterial integral membrane partner, comprises two TM helices. Our analyses revealed that PspM has minimum variation as shown by the MSA ['MSA' tab, webapp] and a narrow phyletic spread restricted to Mycobacteria and Corynebacteria [**Fig. 1B**] (14, 15, 34).

The fourth member in the Mycobacterial operon, PspN (Rv2742c), contains a C-terminal DUF3046 and an N-terminal domain we call **PspN_N** [**Table S1**; **Text S1.2**; 'MSA' tab in the webapp]. **DUF3046** domain is widely prevalent only in actinobacteria [**Fig. 1C**] (14, 15). Moreover, the *M. tuberculosis* genome carries a second copy of DUF3046 (Rv2738c), located four genes downstream of PspN. We

infer that Rv2738c, rather than PspN, contains the ancestral copy of the DUF3046 [**Text S1.2**]. The DUF3046 domain found as part of PspN is likely a duplicated copy of the domain translocated into the PspN ORF of mycobacteria, especially the *M. tuberculosis* complex. Moreover, unlike the pan-actinobacterial DUF3046, the N-terminal portion, PspN_N, is conserved only in *M. tuberculosis*, with remnants of the coding region existing as potential pseudo-genes or degraded sequences in a few closely related mycobacteria (*e.g., M. avium*).

Evolution of PspA

We delved deeper into the PspA/Snf7 superfamily aided by a structure-informed multiple sequence alignment and phylogenetic tree ['MSA' tab, webapp; **Fig. 2A**; **Table S1**].

Identifying PspA+ and Snf7+ clades

PspA+: The PspA+ clade is defined as all versions of the PspA domain closer to PspA proper than Snf7. We found that most bacterial lineages contain PspA+ clade members [**Fig. 1C**] with a few transfers to archaea. Among eukaryotes, only those containing plastids and two flagella (comprising the SAR/HA supergroup and excavata) have members of this clade [**Fig. 2; 1C**; webapp; (73, 74)].

Snf7+: Snf7, which is part of the ESCRT-III complex required for endosome-mediated trafficking via multivesicular body formation and sorting, has a predominantly archaeo-eukaryotic phyletic pattern [**Fig. 1C**; (18, 65–69)]. Examining the curated set of PspA-like proteins additionally revealed a divergent cluster of proteins belonging to the Snf7+ clade [**Fig. 2**; see *Methods*; webapp].

Tracing the roots of the PspA/Snf7 superfamily to the last universal common ancestor (LUCA)

We explored the evolution of the PspA-Snf7 superfamily with a multiple sequence alignment of a comprehensive selection of PspA/Snf7 homologs from the distinct clades across the tree of life (58, 59) as well as different domain architectures [**Fig. 2**; webapp; detailed in the next section; see *Methods*]. Examination of the MSA showed a unique insertion of heptad repeats in actinobacterial PspAs, likely conferring a lineage-specific adaptation. We also found a C-terminal extension in a few cyanobacterial PspA homologs that are more similar to the ancestral plant variant, Vipp (14, 86). We used this alignment to generate a maximum-likelihood PspA/Snf7 phylogenetic tree [**Fig. 2A**; webapp]. In addition to the clade-specific segregation of the PspA/Snf7 homologs, the principal observations from the tree (and the underlying sequence alignment) [**Fig. 2A**] are: i) the homologs from actinobacteria, firmicutes, and proteobacteria form easily distinguishable clusters; ii) the eukaryotic Vipp1 clade is a clear sister group to a clade of cyanobacterial homologs that are closer Vipp1, indicating a plastid-origin; and iii) the Snf7 domain-containing homologs from archaea and eukaryota form a well-defined branch separated by a long internal branch from the PspA containing branches. These observations point to the inheritance of the PspA/Snf7 superfamily from the last universal common ancestor (LUCA).

PspA: Novel architectures and neighborhoods

We leveraged our comprehensive domain-level search to delineate the domain architectures and genomic contexts of the PspA homologs in organisms from diverse clades.

Domain Architectures

Our searches revealed that most PspA homologs (>98%; ~2500 homologs; **Fig. 2B**; webapp) do not show much variation in their underlying domain architecture. In most lineages, PspA homologs only contain the characteristic PspA domain without additional fusions [**Fig. 2B, 3**; **Table S3**]. The remaining small fraction of homologs (<2%) contain multiple tandem PspA domains or fusions with other domains, including NlpC/P60, a novel PspAA domain, and repeated PspA domains [**Figs. 2B, 3**; **Text S2.1**]. The PspA–NlpC/P60 hydrolase fusion is predicted to catalyze the modification of phosphatidylcholine, likely altering membrane composition [**Fig. 3**; **Table S3**; AFZ52345.1; (75)]. Similarly, Snf7 showed minimal variation in domain architecture, with occasional fusions (<5%) in eukaryotes, few of which may have arisen from genome annotation errors [**Fig. 2B**]. Some actinobacteria have an Snf7 homolog fused to an RND-family transporter member [**Fig. 2B**], which transports lipids and fatty acids, and is flanked by two genes encoding a Mycobacterium-specific TM protein with a C-terminal Cysteine-rich domain (76).

Paralogs

In genomes containing multiple copies of PspA, we found that the paralogs do not maintain the same domain architecture and genomic context [**Fig. 3**; **Table S3**; 'Phylogeny –> Paralog' tab, webapp]. In cyanobacteria, dyads/triads of PspA paralogs occur as adjacent genes which might be occasionally fused [**Fig. 3**; **Table S3**; *e.g.,* BAG06015.1, Microcystis]. Consistent with earlier reports (23, 25), we observe that these neighboring PspAs are part of two distinct clusters of homologs, one resembling the bacterial PspA and the other the eukaryotic (plastid) Vipp1 [**Fig. 3**]. Few archaeal and eukaryotic species also carry Snf7 gene clusters (CBY21170.1*,* Oikopleura; **Table S3**), suggestive of tandem gene duplication and possibly relating to their conventional role in membrane stabilization as oligomeric complexes as in the eukaryotic ESCRT systems (77). Further analysis of PspA/Snf7 paralogs, including their likely evolution (gene duplication *vs.* horizontal gene transfer inferred from domain architectures and genomic contexts), can be found in the webapp. Occasionally, PspA/Snf7 co-occur in conserved gene-neighborhoods [**Fig. 2B, 3**; **Table S3**] such as in deltaproteobacteria (AKJ06548.1); in bacteroidetes (CBH24266.1), these neighborhoods contain a third gene also encoding a coiled-coil protein with a structure reminiscent of the PspA coiled-coils; **Table S3**).

Novel variations of known genomic contexts

We next explored the gene-neighborhood themes for each of the Psp members [**Fig. 3**; **Table S3**]. The major themes that are variations of previously characterized Psp operons are summarized below (for the full list of genomic contexts, phyletic spreads, see webapp: 'Data,' 'Genomic Contexts' tabs).

PspFABC operon

In addition to proteobacteria, the classic PspABC is seen in nitrospirae and some spirochaetes [**Fig. 3**; **Table S3**]. We also noted a few variations to this theme (webapp): 1) PspC is fused to a divergent C-terminal PspB in addition to a solo PspB in the operon [**Fig. 3**; **Table S3**; *e.g.,* ANW39986.1, *E. coli*], 2) multiple PspB copies in the operon [**Fig. 3**; **Table S3**; *e.g.,* AOL22920.1, Erythrobacter], 3) PspD occurs along with this operon only in gammaproteobacteria [**Fig. 3**; **Table S3**; *e.g.,* ANW39986.1]. The transcription regulator PspF (NtrC-AAA and HTH) is encoded divergently on the opposite strand but shares a central promoter region with the PspABC operon in most gammaproteobacteria or with just the PspA gene in the remaining organisms that contain it. Key PspF-containing variations include, 1) an operon with additional genes for PspB and PspC fusions and Toastrack-containing proteins in a few spirochaetes; 2) some operons in gammaproteobacteria (*e.g.,* AEN65073.1) with genes encoding NtrC-like transcription factors with N-terminal ligand-binding ACT and PAS domains (78) [**Fig. 3**; **Table S3**], and a further protein of the DO-GTPase family, predicted to play a role in membrane-related stresses (79, 80). These occasionally feature an additional PspF; 3) in some alphaproteobacteria, the PspC in PspFABC has been replaced by multiple Toastrack-containing proteins [**Fig. 3**; **Table S3**; *e.g.,* ANQ40502.1, Gluconobacter].

Associations with Vps4 and other classical AAA$^+$-ATPases

One or more copies of Snf7 (*e.g.,* OLS27540.1; **Table S3**) and a gene encoding a **VPS4-like AAA$^+$-ATPase** (with an N-terminal **MIT** and C-terminal oligomerization domain; **Table S2**) are known to co-occur in archaea; they define the core of an ESCRT complex (81). However, we observed some diversity in Vps4 gene neighborhoods [see **Text S2.2**]: 1) Asgardarchaeota carry a genomic context most similar to eukaryotes comprising the Vps4 AAA$^+$-ATPase and the Snf7 genes along with an ESCRT-II wHTH-domain-containing gene (82); 2) Crenarchaeota contain a three-gene operon with Snf7, Vps4 AAA+-ATPase, and an additional CdvA-like coiled-coil protein with an N-terminal PRC-barrel domain implicated in archaeal cell-division (83). In this case, the Snf7 domain is fused to a C-terminal wHTH domain, which might play a role equivalent to the ESCRT-II wHTH domain. These operons may be further extended with additional copies of Snf7 genes, and other genes coding for a TM protein and an ABC ATPase; 3) related VPS4-like AAA$^+$-ATPase transferred from archaea to several bacterial lineages (*e.g.,* ACB74714.1, Opitutus; **Table S3**), where the Snf7 gene is displaced by an unrelated gene containing TPR repeat and 6TM, again suggesting a membrane-proximal complex. Our analysis also showed that the bacterial PspA (*e.g.,* AEY64321.1, Clostridium; **Table S3**) might occur with a distinct protein containing two AAA$^+$-ATPase domains in various bacterial clades, with the N-terminal version being inactive [detailed in **Text S2.2**]. Both PspA and the membrane-associated Snf7+ clade member, along with AAA$^+$-ATPase, may occur in longer operons with other genes encoding an ABC-ATPase, permease, and a solute- and peptidoglycan-binding protein with PBPB-OmpA domains (*e.g.,* OGG56892.1; **Table S3**).

Snf7 and VPS4 are known to be involved in the ESCRT-III-mediated membrane remodeling in the archaeo-eukaryotic lineage (65–67, 69, 77). The similar associations with AAA$^+$-ATPases, which we

found for certain PspA+ and Snf7+ clade members in bacteria, suggests a comparable functional combination in ATP-dependent membrane-remodelling [**Fig. 3**; **Table S3**]. The additional transporter components seen in some of these systems point to a further link to solute transport, the PbpB-OmpA domain, and ABC transporter proteins. The latter domain interacts with extracellular peptidoglycan, and the former could bind a ligand that is transported by the ABC transporter.

Operons with CesT/Tir-like chaperones and Band-7 domain proteins

We discovered two novel overlapping genomic associations across various bacteria linking PspA with the Band-7 domain and CesT/Tir-like chaperone domain proteins [**Fig. 3**; **Table S2, S3**; 'MSA' tab, webapp]. Band-7 has previously been implicated in a chaperone-like role in binding peptides as part of the assembly of macromolecular complexes (84). Similarly, CesT/Tir chaperone domains have been shown to mediate protein-protein interactions in the assembly and dynamics of the Type-III secretion systems of proteobacteria (85). We established by profile-profile searches that a previously uncharacterized protein encoded by genes linked to the yjfJ family of PspA genes (*e.g.,* ANH61663.1, Dokdonia) is a novel member of the CesT/Tir superfamily [**Fig. 3**; **Table S1, S2**; ANH61662.1]. We also observed that the proteobacterial proteins in the neighborhood of PspA [BAB38581.1, *E. coli*, **Fig. 3**], typified by YjfI from *E. coli* (*e.g.,* BAB38580.1, DUF2170, Pfam; **Table S2**), contained a CesT/Tir superfamily domain (31).

One class of these conserved gene-neighborhoods is centered on a core two-gene dyad, comprising a CesT/Tir gene followed by a PspA gene [**Table S2**], which occurs either as a standalone unit or is further elaborated to give rise to distinct types of larger operons. The first prominent elaboration combines the CesT/Tir–PspA dyad [**Fig. 3**; **Table S3**; ANH61663.1, Dokdonia] with i) a gene coding for a membrane-associated protein with the domain architecture TM+Band-7+Coiled-coil+Flotillin, ii) a novel AAA+-ATPase fused to N-terminal coiled-coil and β-propeller repeats, and iii) a 3TM domain protein prototyped by *E. coli* YqiJ and *B. subtilis* YuaF (previously been implicated in resistance to cell-wall targeting antibiotics (86, 87)). In some proteobacteria, this operon also contains a phospholipase D (HKD) superfamily hydrolase [**Fig. 3**]. Related abbreviated operons coding only for the Band-7/flotillin-containing protein, the YqiJ/YuaF TM protein, and, in some cases, the above-mentioned AAA+-ATPase protein are more widely distributed across bacteria and archaea. These might function with PspA homologs encoded elsewhere in the genome.

The second major elaboration incorporates the CesT/Tir–PspA gene dyad into a larger 6-7 gene operon (AAN56746.1, Shewanella) [**Fig. 3**; **Table S3**]. These operons feature a gene encoding a spermine/spermidine-like synthase domain (88) fused to an N-terminal 7TM transporter-like domain (AAN56744.1, Shewanella; **Fig. 3**; **Table S2, S3**). This operon codes for three additional uncharacterized proteins, including the YjfL family 8TM protein (DUF350, Pfam), a novel lipoprotein (Ctha_1186 domain), and a β-strand rich protein that is predicted to localize to the intracellular compartment (AAN56747.1, Shewanella; DUF4178, Pfam; **Fig. 3**; **Table S3, S4**). In some cases, the

last gene in this operon codes for a ribbon-helix-helix (RHH) domain transcription factor [**Fig. 3**; **Table S2**, **S3**].

The third elaboration that combines the CesT/Tir–PspA gene dyad (AMJ95269.1, Alteromonas) with polyamine metabolism genes encodes an ATP-grasp peptide ligase (AMJ95273.1, Alteromonas) related to the glutathionyl spermidine synthetase [**Fig. 3**; **Table S3**]. This association was also recently noticed in a study of AdoMet decarboxylase gene linkages (89). Additionally, this operon codes for a potassium channel with intracellular-ligand-sensing TrkA_N and TrkA_C domains, a YjfL family 4TM protein (DUF350, Pfam; **Table S2, S3**), metal-chelating lipoprotein (DUF1190, Pfam), and another uncharacterized protein with predicted enzymatic activity (DUF2491, Pfam). The operons containing these polyamine metabolism genes show two variants where the CesT/Tir–PspA dyad is absent. In the first, which also contains DUF4178, the dyad is replaced by a distinct protein occurring as either a secreted or lipoprotein version (DUF4247, CCP45395.1, Mycobacterium) and a predicted enzymatic domain with two highly conserved histidines and glutamate (DUF2617, Pfam; CCP45394.1, Mycobacterium). In the second variant, the dyad is replaced by a protein containing a Band-7 domain fused to C-terminal 2TM and SHOCT domains (see below). These operons encode two further polyamine metabolism genes, an AdoMet decarboxylase, and a polyamine oxidase.

Another class of associations of PspA, typified by the *Bacillus subtilis* ydjFGHL operon, couples the PspA gene (ydjF; ANX09535.1; **Table S1**) with multiple other genes that in firmicutes includes genes coding for 1) a protein (YdjL) with a Band-7 with a C-terminal Zinc Ribbon (ZnR) domain; and 2) a protein (YdjG) with two ZnRs followed by an uncharacterized domain related to YpeB with an NTF2-like fold and a TM segment. However, the conserved partner of PspA in these systems is a membrane-associated protein with a so-called "**TPM phosphatase domain**" (YdjH) (90) [**Fig. 3**; **Table S3**] followed by a low complexity segment or a long coiled-coil. While this domain has been claimed to be a generic phosphatase based on studies on its plant homolog (90), the evidence supporting this activity is limited. Other studies have implicated it in the repair of damaged membrane-associated complexes of the photosystem-II involved in photosynthesis and the assembly of the respiratory complex III (91, 92). Its association with PspA, which parallels its association with the chaperone CesT/Tir and Band-7 involved in macromolecular assembly, is more consistent with the latter role. This supports the alternative hypothesis that the TPM domain might play a chaperone-like role in the assembly of membrane-linked protein complexes along with PspA.

Membrane dynamics with Chaperone-like domains

The association of PspA with the CesT/Tir type chaperone of two distinct families, the Band-7 and the TPM domains, both implicated in the assembly of protein complexes, suggests that these domains might play a role in the assembly of specific membrane-associated complexes together with PspA. These gene-neighborhoods may also feature the AAA^{+}-ATPase fused to C-terminal coiled-coil and β-propeller domains. Concentrations of polyamines like spermine/spermidine have been previously implicated in membrane stability (93). The repeated coupling of polyamine

metabolism genes in these operons may imply that the PspA-based system for membrane dynamics additionally interfaces with changes in polyamine concentration or aminopropylation of other substrates as recently proposed (89) to alter membrane structure and membrane-associated protein complexes. We propose that these systems might link sensing of extracellular stresses to intracellular membrane dynamics based on the repeated coupling with cell-surface lipoproteins. In particular, the flotillin domain-containing proteins are likely to be recruited to lipid-subdomains with special components such as the cardiolipin-rich inner membrane to interface with PspA (87).

The PspAA association

The PspA gene-neighborhood analysis identified a new partner in certain archaeal and bacterial phyla, a protein containing a novel trihelical domain occurring in a two-gene cluster with PspA [**Fig. 3**; **Table S3**; *e.g.,* AAM04874.1, Methanosarcina]. This domain mostly occurs by itself but is occasionally fused to an N-terminal PspA in actinobacteria and chloroflexi [**Fig. 3**; **Table S3**; *e.g.,* ACU53894.1, Acidimicrobium]. We call this domain **PspAA** (for **PspA-Associated**; **Fig. 3**; **Table S2**; **Text S1.3**). This PspA–PspAA dyad occurs in two other contexts with 1) a two-component system (CAB51252.1, Streptomyces) and 2) another dyad comprising a membrane-associated Metallopeptidase, a novel domain, we termed **PspAB** (for **PspA-associated protein B**, AAZ55047.1, Thermobifida), and, occasionally, with a third gene encoding a SHOCT-like domain [**Fig. 3**; **Table S2, S3**; webapp].

PspA with PspM (ClgRPspAMN) or Thioredoxin

In cyanobacteria and actinobacteria, certain PspA homologs may occur as an operonic dyad with a gene encoding an active Thioredoxin domain-containing protein (15) [**Fig. 3**; **Table S2**]. The actinobacterial PspA in this dyad is typified by Corynebacterial RsmP (AKK09942.1; **Fig. 3**; **Table S3**; **Text S2.3**), which, when phosphorylated, regulates the rod-shaped morphology of this bacterium (94). This actinobacterial RsmP-type PspA homolog is predominantly found in rod-shaped actinobacteria. These PspA homologs either occur with Thioredoxin (*e.g.,* AKK09942.1, Corynebacterium) or with ClgR-HTH and PspM (*e.g.,* CCP45543.1, Mycobacterium; **Fig. 3, S1**, **Table S3**). The Corynebacterial members of the family have paralogs with both versions of PspA contexts ['Paralog' tab, webapp]. The association of the ClgR-PspM and Thioredoxin and their mutual exclusion in the operon suggest that they act as alternative regulators of the PspA homologs — repressors in the case of the ClgR and a redox regulator in the case of Thioredoxin [**Fig. S1**; **Text S2.3**]. The presence of the same family of Thioredoxin with a different clade of PspA (typically, two copies) in cyanobacteria [AFZ14666.1, Crinalium; **Fig. 3**; **Table S3**] suggests that a similar redox-dependent control of PspA is likely to be a more pervasive feature.

Operons with two-component systems

The association of two-component systems with the PSP system is widespread. Previously, *liaIHGFSR* operon from *B. subtilis* has been studied in the context of lantibiotic resistance [**Fig. 3**; **Table S3**; (11); ANX06812.1, Bacillus]. In a few Paenibacilli, an additional PspA and PspC with a

C-terminal coiled-coil have been inserted into a LiaIHGFSR-like operon [**Fig. 3**; **Table S3**; *e.g.,* APB74393.1, part of eight-gene context]. In actinobacteria [**Fig. 3**; **Table S3**; CAB51252.1, Streptomyces], the dyad of PspA–PspAA occurs with a firmicute-like two-component system.

Classic two-component transcriptional signaling system

Two-component signaling systems linked to either the PspA or PspC in operons are seen in firmicutes, actinobacteria, and other clades [**Fig. 3**; webapp]. In these operons, the classic membrane-bound HISKIN communicates an external stress signal to the response regulator (REC–HTH) protein, triggering a transcriptional response. It is very likely that PspA, LiaI-LiaF-TM, and Toastrack tie into a two-component system to regulate the membrane integrity and permeability in response to the stress signal. In actinobacteria, where the HisKin is fused to PspC, the signal is presumably sensed by PspC. Even when these PspC/Toastrack operons with two-component systems lack PspA in their immediate operonic neighborhood, they likely recruit PspA proteins from elsewhere in the genome to bring about the stress response function.

PspA-free variations of domain architectures and gene-neighborhood

We next investigated the domain architectures and genomic contexts involving Psp components that do not carry PspA. Below we describe some of our significant findings related to PspA-free genomic contexts, including those with the PspC and Toastrack domains novel.

PspC domain architectures and gene-neighborhood

PspC by itself is present in most bacteria and the archaeal clades of euryarchaeota and Asgardarchaeota. Some orthologs of PspC are fused to an N-terminal ZnR domain [**Table S4**; ABC83427.1, Anaeromyxobacter], while most other homologs either show fusions to, or occur in predicted operons with, multi-TM domains such as the LiaI-LiaF-TM and PspB [**Fig. 4**; **Table S4**]. Additionally, PspC is also fused to diverse signaling domains such as i) HISKIN from two-component systems (see above), ii) a novel signaling domain, we term the HTH-associated α-helical signaling domain (**HAAS**, overlaps partly with the Pfam model for DUF1700; **Table S2**), and iii) the Toastrack domain (**Table S1**). Moreover, actinobacteria contain a Lia-like system without PspA. The HISKIN domain in these systems is distinct from the Lia-like systems with PspA; it is fused to N-terminal PspC and LiaF-LiaI-TM domains (*e.g.,* AIJ13865.1, Streptomyces) and accompanied by a REC–HTH-containing protein [**Fig. 5**; **Table S4**]. This core shares a promoter region with a second gene on the opposite strand containing a PspC domain that is fused to additional TMs and, in some cases, a Toastrack domain [*e.g.,* AIJ13866.1, Streptomyces; **Fig. 5**; **Table S4**]. These associations strongly imply that PspC is a sensor domain that likely senses a membrane-proximal signal and communicates it via the associated domains.

Contextual associations of the Toastrack Domain

The **Toastrack** domain repeatedly emerges as a partner to the PSP system. As noted above, genes encoding Toastrack-containing protein widely co-occur with PspA in several conserved

neighborhoods [**Fig. 4**; **Table S4**; see *Methods*], and are fused to PspC when occurring independently of PspA. We found that the Toastrack domains tend to be C-terminal to various TM domains such as PspC, LiaI-LiaF-TM, and other multi-TM/single-TM domains unique to these systems [**Fig. 4**; **Text S2.4**]. These fusions strongly suggest that the Toastrack is an intracellular domain. Further, in several Toastrack architectures, the N-terminal regions contain at least two divergent homologs of the bihelical SHOCT domain (e.g., *yvlB*, CAB15517.1, Bacillus) [**Fig. 4, 5**; **Table S1, S4**] that we call **SHOCT-like** domains (partly detected by the Pfam **DUF1707** model (95); **Text S2.4**). Based on our observation that these SHOCT-like domains could occur at the N- or C-termini of various globular domains, including Toastrack, we predict that this domain plays a role in anchoring disparate domains to the inner leaf of the membrane.

The DNA-binding domain **LytTR** may also be encoded by operons carrying Toastrack fused to LiaI-LiaF-TM in firmicutes (e.g., Clostridium CAL82154.1**; Fig. 5**; **Table S4**). Furthermore, the Toastrack protein (*e.g.,* CAB15517.1, Bacillus), which occurs in an operon with a PspC homolog (CAB15516.1), has been implicated in the activation of the membrane-associated LiaFSR operon [**Table S1**] and protection against membrane permeabilization (52, 96, 97). We also identified variants of the classic Lia operon with Toastrack and the two-component system that lack PspA, typified by *vraT* of *Staphylococcus aureus* [ABD31150.1; **Fig. 5**; **Table S4**]. These systems carry LiaI-LiaF-TM and Toastrack with a two-component system (*vraSR*) and are involved in methicillin and cell-wall-targeting antibiotic resistance (3, 98–100). A similar organization of LiaI-LiaF-TM and Toastrack containing protein with a two-component system is found in a few bacterial species (*e.g.,* proteobacteria, ABD83157.1, Ignavibacteriae, AFH48155.1) [**Table S4**].

We also recovered several other conserved gene neighborhoods centered on Toastrack genes that are likely to define novel functional analogs of the Psp system with potential roles in membrane-linked stress response. One of these analogs is a four-gene context found across diverse bacterial lineages, including a previously uncharacterized protein with hits to the Pfam model DUF2089. We found that this Pfam model **DUF2089** can be divided into an N-terminal ZnR, central HTH, and C-terminal SHOCT-like domains (ADE70705.1, Bacillus). Variants of these might be combined with LiaI-LiaF-TM, B-box, or PspC domains [**Fig. 4, 5**; **Table S2, S4**; **Text S2.4**]. We propose that these Toastrack contexts function similarly to the classical Lia operon in transducing membrane-associated signals to a transcriptional output affecting a wide range of genes, including via an operonically linked sigma factor.

Similarly, we observed operons coupling a membrane-anchored Toastrack-containing protein with an ABC-ATPase, a permease subunit, a **GNTR-HTH** transcription factor with distinct C-terminal α-helical domain [**Fig. 5**; **Table S2, S4**; **Text S2.4**], and a DUF2884 membrane-associated lipoprotein (*e.g.,* AJI31452.1, Bacillus), which might function as an extracellular solute-binding partner for the ABC-ATPase and permease components. We found a comparable operon in actinobacteria with the Ribbon-helix-helix (**RHH**) domain transcription factor replacing **GNTR-HTH** [**Table S2**; **Text S2.4**]. In

some actinobacteria, the Toastrack domain is fused to a SHOCT-like domain or occurs in a transport operon [**Fig. 5**; **Table S4**], likely coupling Toastrack to transcriptional regulation, sensing of membrane-proximal signal and transport [**Fig. 5**; **Table S4**].

Based on these observations, we posit that the two extended β-sheets of the Toastrack domain (formed by the β-helical repeats) present two expansive surfaces that are amenable to protein-protein interactions proximal to the membrane. Hence, we suggest that this domain provides a platform for the assembly of sub-membrane signaling complexes. The docking of proteins to the Toastrack domain could be further transduced to activate transcription via fused or associated HTH or LytTR domains, associated two-component systems, or the HAAS-HTH system.

The HAAS-PadR-like wHTH systems

We discovered a novel signaling system that frequently co-occurs with PspC and Toastrack domains as part of gene neighborhoods that do not carry PspA [**Fig. 4, 5**]. This system consists of two components, the HAAS domain and the PadR-like wHTH [**Table S2**]. The HAAS domain is partly detected by the Pfam models DUF1700 and DUF1129 (Pfam clan: Yip1). An HHpred search further unifies this model with the Pfam profile DUF1048 (PDB: 2O3L). We, therefore, unify them as the **HAAS** superfamily [see HAAS under 'MSA' tab, webapp]. The core HAAS fold has three consecutive α-helices, with the third helix displaying a peculiar kink in the middle corresponding to a conserved GxP motif, predicted to form part of a conserved groove that mediates protein-protein interactions. The HAAS domain always occurs in gene-neighborhoods coupled with a protein containing a standalone PadR-like wHTH DNA-binding domain [**Fig. 5**]. This co-migration suggests that HAAS transduces a signal from domains fused to it to this wHTH transcription factor.

When coupled with PspC, the HAAS domain shows three broad themes: 1) as part of a multidomain TM protein with additional PspC, LiaI-LiaF-TM, and Toastrack domains; 2) fused directly to a Toastrack domain in a gene neighborhood that also codes for a PspC domain protein occurring either by itself or fused to other TM domains; 3) as part of a TM protein fused to conserved multi-TM domains other than the LiaI-LiaF-TM or PspC (*e.g.,* VanZ) [**Fig. 4, 5**; **Table S4**]. These gene neighborhoods code for standalone PspC genes. Additionally, the HAAS domains occur in contexts independent of PspC but are typically coupled to the N-terminus of other multi-TM or intracellular sensory domains. We found conserved fusions to at least 5 distinct multi-TM domains [**Fig. 4, 5**; **Table S4**], such as 1) The FtsX-like TM domains with extracellular solute-binding MacB/PCD domains (ACO32024.1, Acidobacterium); 2) FtsW/RodA/SpoVE family of TM domains (CAC98500.1, Listeria) (101); 3) various uncharacterized 6TM, 4TM and 2TM domains. In addition to Toastrack, the HAAS domain may combine with other intracellular domains such as the Pentapeptide repeat domains [AOH56696.1, Bacillus; **Fig. 5**; **Table S4**] (mostly in firmicutes). While these operons do not contain PspC, the organisms have Psp components elsewhere in the genome. These proteins might be recruited to the stress response system as suggested by the effects of PadR-like wHTH deletion studies in the Listeria Lia systems (101).

A key observation is that the HAAS–PadR-like-wHTH dyad replaces the histidine kinase-receiver domain transcription factor pair in bacteria that do not contain the typical Lia operon (carrying the LiaRS-like two-component system). Hence, we posit that the HAAS-PadR dyad constitutes an alternative to the classical two-component systems. Here, the HAAS domain could undergo a conformational change in response to membrane-linked or intracellular signals detected by the sensor domains (*e.g.,* LiaI-LiaF-TM or PspC) that are fused to or co-occur with the dyad. This conformational change is likely transmitted to the associated PadR-like wHTH via the groove associated with the conserved GxP motif; this change might lead to the release of the transcription factor to mediate a downstream transcriptional response.

A unified view of PSP partners and evolution

In conclusion, we put together all our findings, including genomic contexts and partner domains, to assemble the Psp puzzle. We first built a network of domain architectures for the extended PSP system [**Fig. 6A**], with domains as nodes and co-occurrence within a protein as edges. This network consolidates the PSP system with connections between PspA, Psp cognate partner domains, and other novel domain connections. Next, we summarized the phyletic spread and prevalence of all Psp members and partners in a single heatmap [**Fig. 6B**]. This view recapitulates the findings that of the Psp components (i) only PspA/Snf7 and Toastrack are present in eukaryotes; (ii) while PspC, PspA/Snf7 are present in most archaeal lineages, our analysis identifies occasional transfers of Toastrack, LiaI-LiaF-TM, and SHOCT domains to euryarchaeota; (iii) in bacteria, domains such as Toastrack, PspC, PspA, and LiaI-LiaF-TM tend to co-migrate. Finally, using an 'upset plot,' we unified: i) the relative occurrence of domain architectures from Psp members and their most common neighbors, ii) combinations of these domain architectures, and iii) their frequencies [**Fig. 6C**]. This plot, for example, reveals that singleton PspA/Snf7, PspC, and Toastrack domains are most prevalent, closely followed by HAAS–PadR-like-wHTH dyads with/without Toastrack, the Toastrack/LiaI-LiaF-TM contexts with two-component systems, and PspABC operons [**Fig. 6C**]; these consolidated visualizations can be explored in our webapp.

Discussion

Recent work with the bacterial PSP stress response system suggests that though its function is maintained across phyla, the Psp proteins, the regulatory mechanisms, and the membrane stress response mechanism vary widely across lineages (12, 15, 24, 34). These variations reflect how a common peripheral inner membrane protein is utilized in lineage-specific envelope dynamics and responses to unique environmental stresses. In this study, we report the results of the first comprehensive analysis of the evolution of the PSP system and all its partner proteins, focusing on the phyletic distribution [**Fig. 1**], sequence/structural features of PspA, its known and newly-predicted partners [**Fig. 2–4**], and their conserved gene-neighborhoods [**Fig. 3, 5**] across the tree of life [**Fig. 2, 6**]. We established that PspA/Snf7 goes back to the LUCA [**Fig. 2**], in agreement with a recent study (19). As a corollary, despite the different membrane types (*e.g.,* ether-linked vs.

ester-linked) in the archaeal, bacterial, and eukaryotic lineages, the LUCA already possessed a membrane whose curvature and dynamics were mediated by an ancestral PspA/Snf7-like coiled-coil protein assembled into polymeric structures adjacent to the inner-leaf of the membrane. This hypothesis is supported by studies on signal-recognition particle GTPases (102) and ATP-synthases that imply the presence of a transmembrane secretory and ion-transport apparatus in the LUCA (103, 104), suggesting that this ancient machinery was robust to the subsequent lineage-specific changes in membrane chemistry.

Our findings point to several new likely functional processes of note: 1) Based on its occurrence with membrane partners (such as PspA, LiaI-LiaF-TM, PspC, HAAS, SHOCT-like domains) and transcription regulators/two-component systems (such as LiaRS, PadR-like-wHTH), we propose that Toastrack is an intracellular domain likely proximal to the membrane that provides a platform for the assembly of sub-membrane signaling complexes. 2) The repeated association of PspA homologs with distinct AAA^{+}-chaperones and a non-ATPase chaperone or predicted chaperone-like proteins (*e.g.,* CesT/Tir, Band-7, and TPM) suggests that these associations are involved in the assembly of sub- and trans-membrane complexes presumably in response to specific envelope stresses. These observations suggest a more general role for analogs of ESCRT-like complexes in these bacterial systems. 3) The association of PspA with various transporters and polyamine metabolism systems suggests that the regulation of membrane structure by PspA is associated with changes in the concentration of different solutes and covalent modifications that may affect membrane stability. 4) The finding of several new versions of the SHOCT-like domain suggests that these domains mediate membrane localization of disparate catalytic and non-catalytic activities, which in turn may interface with the PSP system. 5) A diverse array of sub-membrane (*e.g.,* Toastrack), TM, and cell-surface domains (*e.g.,* PspC, LiaI-LiaF-TM) co-occur with two-component or HAAS-PadR-like-wHTH systems, which suggests alternatives to conventional two-component signaling that interface with the PSP system in signaling membrane stress.

All the findings (data, results, and visual summaries) from this work are found in an interactive and queryable web application available at https://jravilab.shinyapps.io/psp-evolution.

Methods

Query and subject selection

All known Psp members — PspA (with eight representatives across Proteobacteria, Actinobacteria, two copies from Firmicutes, cyanobacteria, and archaea); PspM (Rv2743c) and PspN (Rv2742c) (from *M. tuberculosis*); PspB and PspC (from *E. coli*); LiaI, LiaG, and LiaF (from *B. subtilis*) — were queried against all sequenced genomes across the tree of life. Homology searches were run against custom databases of ~6500 completed representative/reference genomes with taxonomic lineages or the NCBI non-redundant database (105, 106) [**Table S1, S2**]. All Psp homologs are listed in the

webapp, with representatives in **Tables S3, S4**. We obtained the phyletic order (sequence) from NCBI taxonomy and PATRIC (105–110).

Identification and characterization of protein homologs

To locate and stratify all Psp homologs, we started with the previously characterized Psp proteins from the best-studied operons *pspF||ABCDE-G* (from *E. coli*), *liaIHGFSR* (from *B. subtilis*), and *clgRpspAMN* (from *M. tuberculosis*), and analyzed their phyletic spread. To ensure an exhaustive search and identification of all related Psp proteins, including remote homologs, we first resolved these Psp proteins into their constituent domains and used each of the individual domains as queries for all our phylogenetic analyses. This approach allows us to find homologies beyond those from full-length proteins only. We then performed homology searches for each constituent domain across the tree of life (approximately 6500 representative/reference genomes). We used a combination of protein domain and orthology databases, iterative homology searches, and multiple sequence alignment to detect domains, signal peptides, transmembrane (TM) regions to help construct the domain architecture of each of the query Psp proteins. Due to the ubiquitous presence of transcription factor Helix-turn-helix (HTH)-type domains and Histidine kinase-receiver domain two-component systems in bacterial phyla, we did not perform dedicated searches with these domains; instead, we noted their occurrence in the predicted operons alongside the core Psp genes to identify transcriptional regulation partners.

To ensure the identification of a comprehensive set of homologs (close and remote) for each queried protein, we performed iterative searches using PSIBLAST (111) with sequences of both full-length proteins and the corresponding constituent domains. More distant relationships were determined using profile-profile searches with the HHpred program. For each protein, searches were conducted using homologous copies from multiple species as starting points. We then aggregated these search results and recorded the numbers of homologs per species and genomes carrying each of the query proteins [webapp]. These proteins were clustered into orthologous families using the similarity-based clustering program BLASTCLUST (112). SignalP, TMHMM, Phobius, JPred, Pfam, and custom profile databases (113–120) were used to identify signal peptides, TM regions, known domains, and the secondary protein structures in every genome. Homolog information, including domain architectures, can be accessed in the webapp ('Data,' 'Domain architectures' tabs).

Neighborhood search

Bacterial gene neighborhoods — ±7 genes flanking each protein homolog — were retrieved using a Perl script from GenBank (105, 106). Gene orientation, domains, and secondary structures of the neighboring proteins were characterized using the same methods applied to query homologs above. Genomic contexts can be accessed in the webapp ('Genomic contexts' tab). We note that eukaryotic components (*e.g.,* Snf7) appearing alone is an artifact of the genomic context analysis being restricted (and relevant) to bacteria and archaea (see webapp).

Phylogenetic analysis

Multiple sequence alignment (MSA) of the identified homologs was performed using Kalign (121) and MUSCLE (122). The phylogenetic tree was constructed using FastTree 2.1 with default parameters (123). MSA and phylogenetic trees can be accessed in the webapp ('Phylogeny' tab).

Network reconstruction

The Psp proximal neighborhood network was constructed based on the domain architectures and genomic contexts of PspA and its cognate partner proteins [**Table S1, S2**]. The nodes represented the domains, and edges indicated a shared neighborhood (domains of the same protein or neighboring proteins). Proximity networks can be accessed in the webapp ('Domain architectures' tab).

Web application

The interactive and queryable web application was built using R Shiny (124, 125). All data analyses and visualizations were carried out using R/RStudio (126, 127) and R-packages (128–141). All the data and results from our study are available on the web application (https://jravilab.shinyapps.io/psp-evolution). The webapp can be queried with the protein accession numbers, domain names, and species that are cited inline throughout the Results section.

Figures

Figure 1. The phyletic spread of known Psp members across all major lineages.

A. The three known PSP systems in *E. coli* (pspF||pspABC), *M. tuberculosis* (clgRpspAMN), and *B. subtilis* (liaIHGFSR) are shown. Boxes indicate genes/proteins, and colors indicate the nature of the protein (black, PspA homolog; teal, transcription factor/regulator; warmer shades of orange/yellow, transmembrane protein). Thin arrows denote the direction of transcription, and block arrows denote promoters. **B. Domain architectures of the known Psp operons** in *E. coli*, *M. tuberculosis*, and *B. subtilis*. Domains are denoted by rectangular segments inside the block arrow representing a protein. The direction of the arrow indicates the direction of transcription. See *Results* for new domain definitions. **C. Phyletic spreads of Psp proteins.** Sunburst plots are shown for the homologs of 'known' domains/protein families of interest: PspA, Snf7, PspB, PspC, PspM, PspN (and DUF3046), LiaI-LiaF-TM, Toastrack. In each plot corresponding to a particular protein, the inner ring corresponds to the proportion of its homologs present in kingdoms, bacteria, archaea, and eukaryota. The outer ring depicts the distribution of the homologs among key lineages (phyla). Interactive sunburst plots for each of the Psp proteins are available in the webapp.

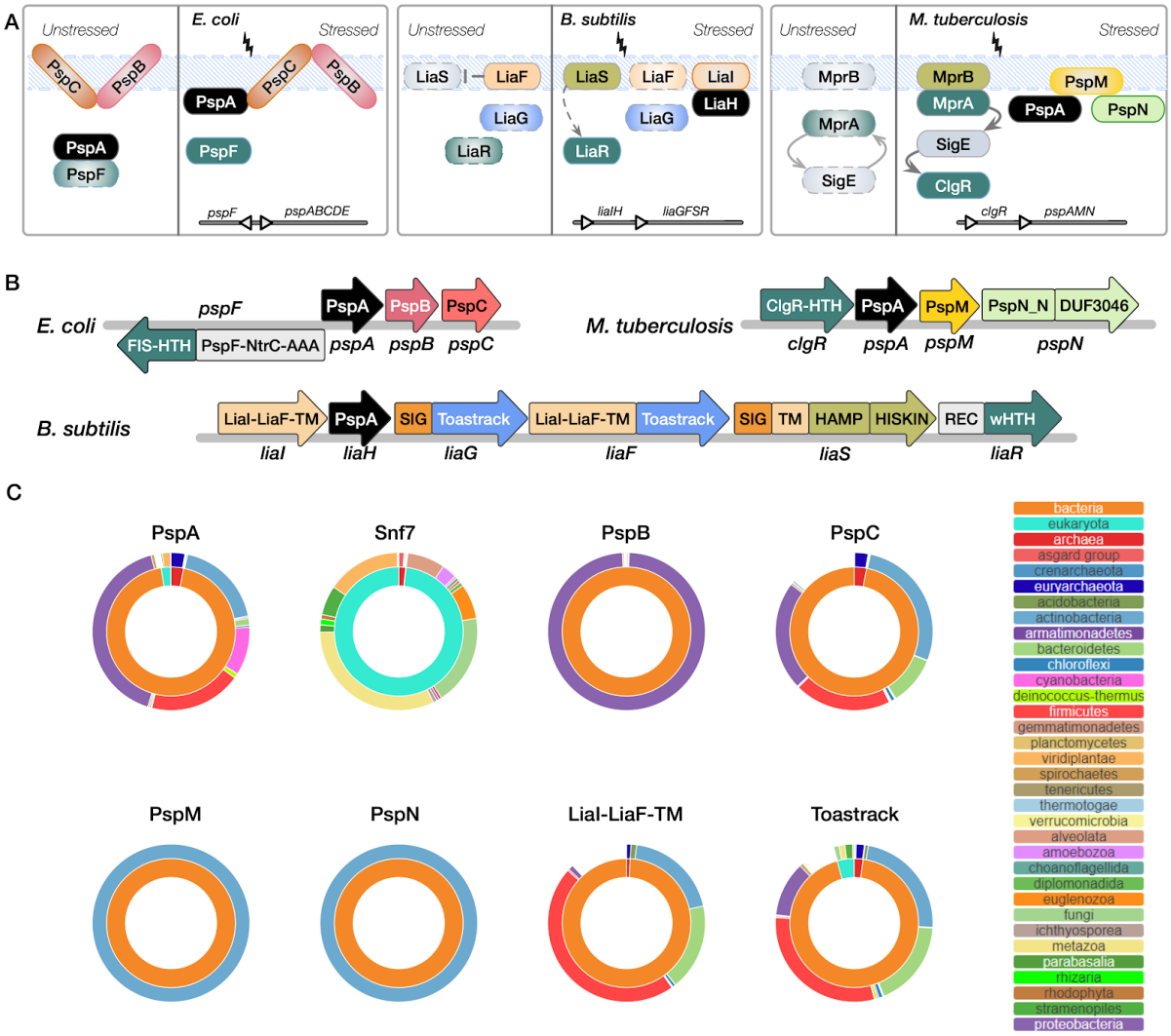

A
Unstressed
E. coli
Stressed
PspC
PspB
PspA
PspF
pspF
pspABCDE
B. subtilis
LiaS
LiaF
LiaI
LiaH
LiaG
LiaR
liaIH
liaGFSR
M. tuberculosis
MprB
MprA
SigE
ClgR
PspM
PspN
clgR
pspAMN
B
E. coli
pspF
FIS-HTH
PspF-NtrC-AAA
pspA
pspB
pspC
M. tuberculosis
ClgR-HTH
PspN_N
DUF3046
pspM
pspN
B. subtilis
LiaI-LiaF-TM
SIG
Toastrack
TM
HAMP
HISKIN
REC
wHTH
liaI
liaH
liaG
liaF
liaS
liaR
C
PspA
Snf7
PspB
PspC
PspM
PspN
LiaI-LiaF-TM
Toastrack
bacteria
eukaryota
archaea
asgard group
crenarchaeota
euryarchaeota
acidobacteria
actinobacteria
armatimonadetes
bacteroidetes
chloroflexi
cyanobacteria
deinococcus-thermus
firmicutes
gemmatimonadetes
planctomycetes
viridiplantae
spirochaetes
tenericutes
thermotogae
verrucomicrobia
alveolata
amoebozoa
choanoflagellida
diplomonadida
euglenozoa
fungi
ichthyosporea
metazoa
parabasalia
rhizaria
rhodophyta
stramenopiles
proteobacteria

Figure 2. PspA/Snf7 homologs across the tree of life.

A. Phylogenetic tree of PspA homologs across the tree of life. The phylogenetic tree was constructed based on a multiple sequence alignment performed using representative PspA homologs across all the major kingdoms/phyla (see *Methods*; webapp). The key lineages are labeled next to distinct clusters of similar PspA proteins. The insets show the 3D structures for PspA (4WHE) and Snf7 (5FD7) from the Protein Data Bank. The tree leaves are labeled by lineage, species, and accession numbers. The outgroup at the bottom of the tree includes Snf7 homologs. **B. Phyletic spreads of PspA/Snf7 domain architectures across the tree of life.** The phyletic spread of the various PspA/Snf7 domain architectures is shown along with their relative frequencies as a stacked barplot. Further details of the domain architectures of all PspA/Snf7 homologs and their phyletic spreads are shown in the webapp, with representative ones shown in **Table S3**.

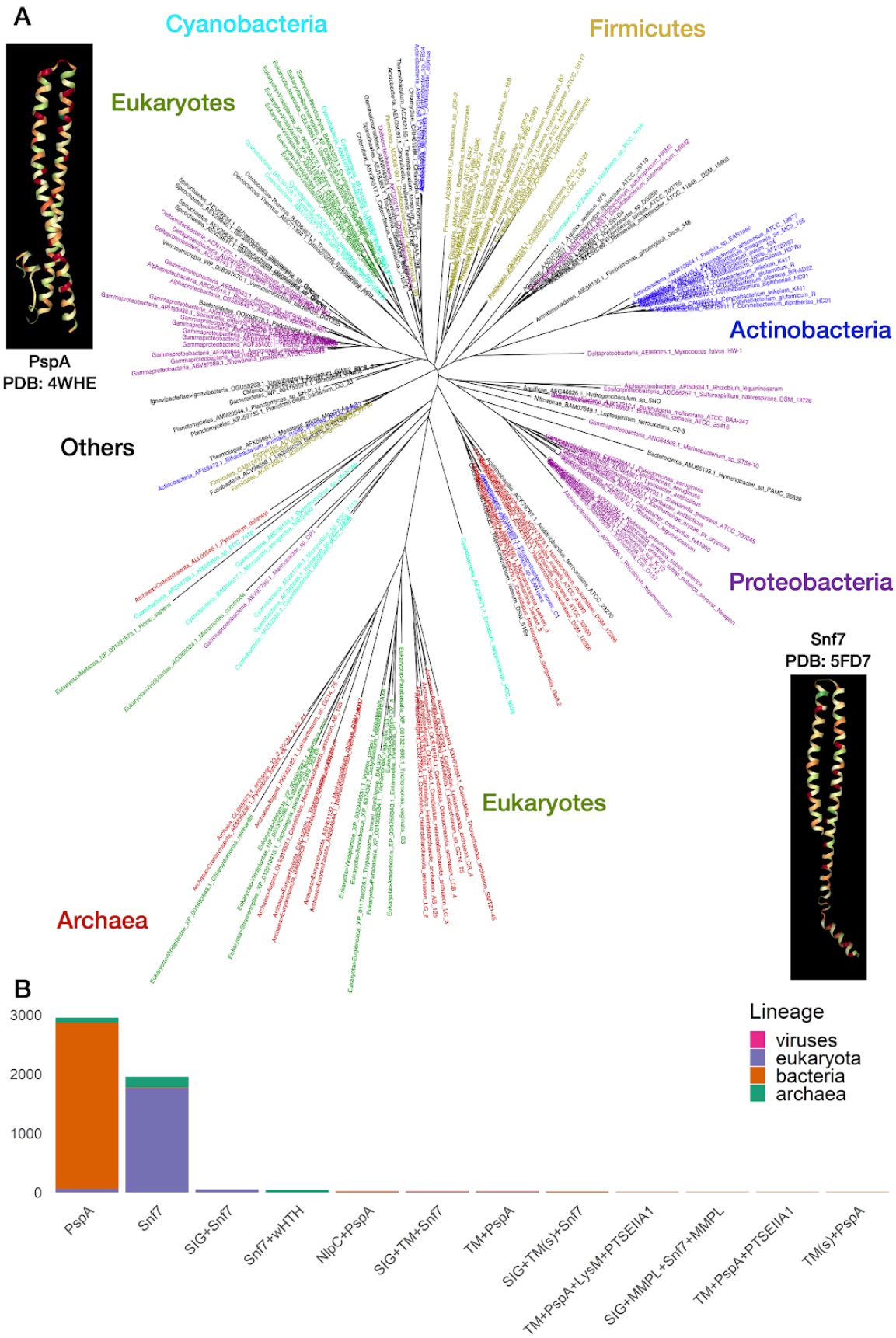

A
Cyanobacteria
Firmicutes
Eukaryotes
PspA
PDB: 4WHE
Actinobacteria
Others
Proteobacteria
Snf7
PDB: 5FD7
Eukaryotes
Archaea
B
3000
2000
1000
0
Lineage
viruses
eukaryota
bacteria
archaea
PspA
Snf7
SIG+Snf7
Snf7+wHTH
NlpC+PspA
SIG+TM+Snf7
TM+PspA
SIG+TM(s)+Snf7
TM+PspA+LysM+PTSEIIA1
SIG+MMPL+Snf7+MMPL
TM+PspA+PTSEIIA1
TM(s)+PspA

Figure 3. PspA domain architectures and genomic contexts

The first row contains domain architectures of the most prevalent homologs of PspA (indicated in black throughout). The remaining rows show the predominant genomic contexts of PspA homologs across multiple bacterial and archaeal lineages (identified by neighborhood searches ±7 genes flanking each homolog; see *Methods*). Representative neighborhoods, along with their archetypal lineages and archetypal example proteins (with accession numbers and species), are shown. The PspA contexts are grouped by neighboring domains such as (a) PspF, PspB/PspC; (b) PspAA, PspAB; (c) ClgR, PspM/PspN, Thioredoxin [**Fig. S1**]; (d) chaperones such as Band-7, Flotillin, CesT_Tir, TPM_phosphatase, ZnR, SpermGS-ATPgrasp, and Spermine synthase [**Table S2**]; (e) two-component systems such as the Lia system and Toastrack, and other novel genomic contexts; and (f) cyanobacterial variations. Key: *rectangles*, domains; *arrowheads*, the direction of transcription; *domains enclosed by dotted lines*, absent in the genomic contexts in certain species; *white cross*, substitution with protein(s) mentioned just below; *white triangle*, insertion with one or more proteins; *two vertical lines* '||', indicate a change in the direction of transcription; *small black boxes,* domain repeats co-occurring within a protein/context. Archetypal accession numbers and species are provided mostly on the left. Archetypal lineages are indicated in grey on the right for each of the domain architectures and genomic contexts. Different domains are indicated by different colors and by the same coloring scheme across the figures in this study. Also, similar domains are given similar hues. For example, membrane proteins are in different shades of orange (SIG, predicted signal peptide, dark orange, PspC, orange, other transmembrane domain, light orange); transcription factors/regulators (including HTH, helix-turn-helix domain) are in teal; DUFs, Domains of Unknown Function, and other domains are in grey. Further details of the domain architectures and gene neighborhoods shown are described in the text and **Table S3**, and the full list of PspA homologs, their domain architectures, genomic contexts, and lineages are shown in the webapp (under the 'Data,' 'Domain architectures,' and 'Genomic contexts' tabs).

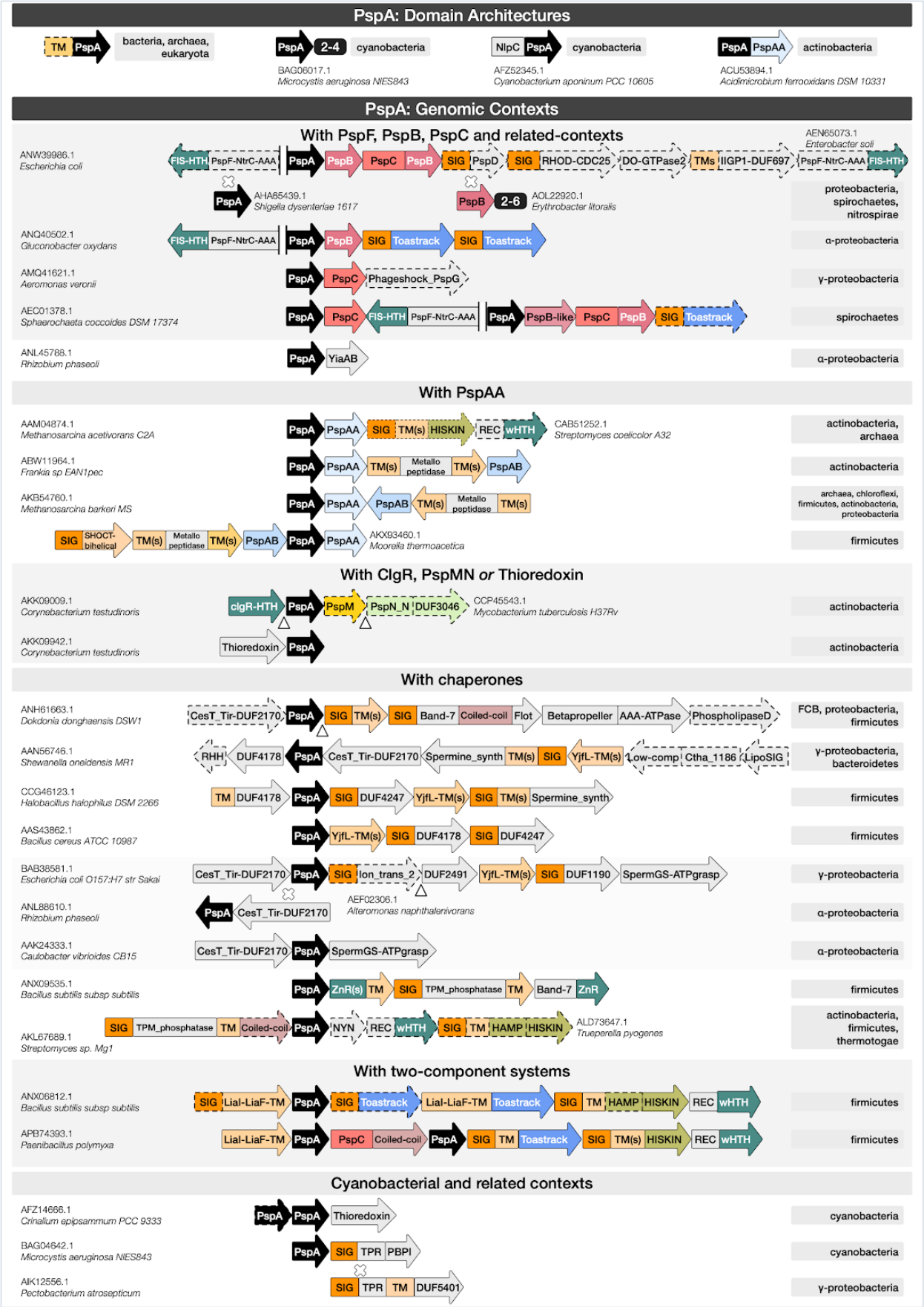

PspA: Domain Architectures
TM PspA
bacteria, archaea, eukaryota
PspA 2-4
cyanobacteria
BAG06017.1
Microcystis aeruginosa NIES843
NlpC PspA
cyanobacteria
AFZ52345.1
Cyanobacterium aponinum PCC 10605
PspA PspAA
actinobacteria
ACU53894.1
Acidimicrobium ferrooxidans DSM 10331
PspA: Genomic Contexts
With PspF, PspB, PspC and related-contexts
ANW39986.1
Escherichia coli
FIS-HTH PspF-NtrC-AAA PspA PspB PspC PspB SIG PspD SIG RHOD-CDC25 DO-GTPase2 TMs IIGP1-DUF697 PspF-NtrC-AAA FIS-HTH
AEN65073.1
Enterobacter soli
PspA
AHA65439.1
Shigella dysenteriae 1617
PspB 2-6
AOL22920.1
Erythrobacter litoralis
proteobacteria, spirochaetes, nitrospirae
ANQ40502.1
Gluconobacter oxydans
FIS-HTH PspF-NtrC-AAA PspA PspB SIG Toastrack SIG Toastrack
α-proteobacteria
AMQ41621.1
Aeromonas veronii
PspA PspC Phageshock_PspG
γ-proteobacteria
AEC01378.1
Sphaerochaeta coccoides DSM 17374
PspA PspC FIS-HTH PspF-NtrC-AAA PspA PspB-like PspC PspB SIG Toastrack
spirochaetes
ANL45788.1
Rhizobium phaseoli
PspA YiaAB
α-proteobacteria
With PspAA
AAM04874.1
Methanosarcina acetivorans C2A
PspA PspAA SIG TM(s) HISKIN REC wHTH
CAB51252.1
Streptomyces coelicolor A32
actinobacteria, archaea
ABW11964.1
Frankia sp EAN1pec
PspA PspAA TM(s) Metallo peptidase TM(s) PspAB
actinobacteria
AKB54760.1
Methanosarcina barkeri MS
PspA PspAA PspAB TM(s) Metallo peptidase TM(s)
archaea, chloroflexi, firmicutes, actinobacteria, proteobacteria
SIG SHOCT-bihelical TM(s) Metallo peptidase TM(s) PspAB PspA PspAA
AKX93460.1
Moorella thermoacetica
firmicutes
With ClgR, PspMN or Thioredoxin
AKK09009.1
Corynebacterium testudinoris
clgR-HTH PspA PspM PspN_N DUF3046
CCP45543.1
Mycobacterium tuberculosis H37Rv
actinobacteria
AKK09942.1
Corynebacterium testudinoris
Thioredoxin PspA
actinobacteria
With chaperones
ANH61663.1
Dokdonia donghaensis DSW1
CesT_Tir-DUF2170 PspA SIG TM(s) SIG Band-7 Coiled-coil Flot Betapropeller AAA-ATPase PhospholipaseD
FCB, proteobacteria, firmicutes
AAN56746.1
Shewanella oneidensis MR1
RHH DUF4178 PspA CesT_Tir-DUF2170 Spermine_synth TM(s) SIG YjfL-TM(s) Low-comp Ctha_1186 LipoSIG
γ-proteobacteria, bacteroidetes
CCG46123.1
Halobacillus halophilus DSM 2266
TM DUF4178 PspA SIG DUF4247 YjfL-TM(s) SIG TM(s) Spermine_synth
firmicutes
AAS43862.1
Bacillus cereus ATCC 10987
PspA YjfL-TM(s) SIG DUF4178 SIG DUF4247
firmicutes
BAB38581.1
Escherichia coli O157:H7 str Sakai
CesT_Tir-DUF2170 PspA SIG Ion_trans_2 DUF2491 YjfL-TM(s) SIG DUF1190 SpermGS-ATPgrasp
γ-proteobacteria
ANL88610.1
Rhizobium phaseoli
PspA CesT_Tir-DUF2170
AEF02306.1
Alteromonas naphthalenivorans
α-proteobacteria
AAK24333.1
Caulobacter vibrioides CB15
CesT_Tir-DUF2170 PspA SpermGS-ATPgrasp
α-proteobacteria
ANX09535.1
Bacillus subtilis subsp subtilis
PspA ZnR(s) TM SIG TPM_phosphatase TM Band-7 ZnR
firmicutes
AKL67689.1
Streptomyces sp. Mg1
SIG TPM_phosphatase TM Coiled-coil PspA NYN REC wHTH SIG TM HAMP HISKIN
ALD73647.1
Trueperella pyogenes
actinobacteria, firmicutes, thermotogae
With two-component systems
ANX06812.1
Bacillus subtilis subsp subtilis
SIG LiaI-LiaF-TM PspA SIG Toastrack LiaI-LiaF-TM Toastrack SIG TM HAMP HISKIN REC wHTH
firmicutes
APB74393.1
Paenibacillus polymyxa
LiaI-LiaF-TM PspA PspC Coiled-coil PspA SIG TM Toastrack SIG TM(s) HISKIN REC wHTH
firmicutes
Cyanobacterial and related contexts
AFZ14666.1
Crinalium epipsammum PCC 9333
PspA PspA Thioredoxin
cyanobacteria
BAG04642.1
Microcystis aeruginosa NIES843
PspA SIG TPR PBPI
cyanobacteria
AIK12556.1
Pectobacterium atrosepticum
SIG TPR TM DUF5401
γ-proteobacteria

Figure 4. Lineage spread of PspA-free domain architectures.

The domain architectures of the most prevalent homologs of PspA partner domains (frequency of occurrence >50 across lineages), including known (Toastrack, LiaI-LiaF-TM, PspBC, PspMN, DUF3046) and other novel neighbors (PspAA, PspAB, HAAS, SHOCT-bihelical, SHOCT-like, AAA^{+}-ATPase domains) are illustrated on the left. The phyletic spread of the various domain architectures is shown along with their relative frequencies as a stacked barplot. Further details of the domain architectures of all PspA partner domain homologs and their phyletic spreads are shown in the webapp, with representative ones shown in **Table S4**.

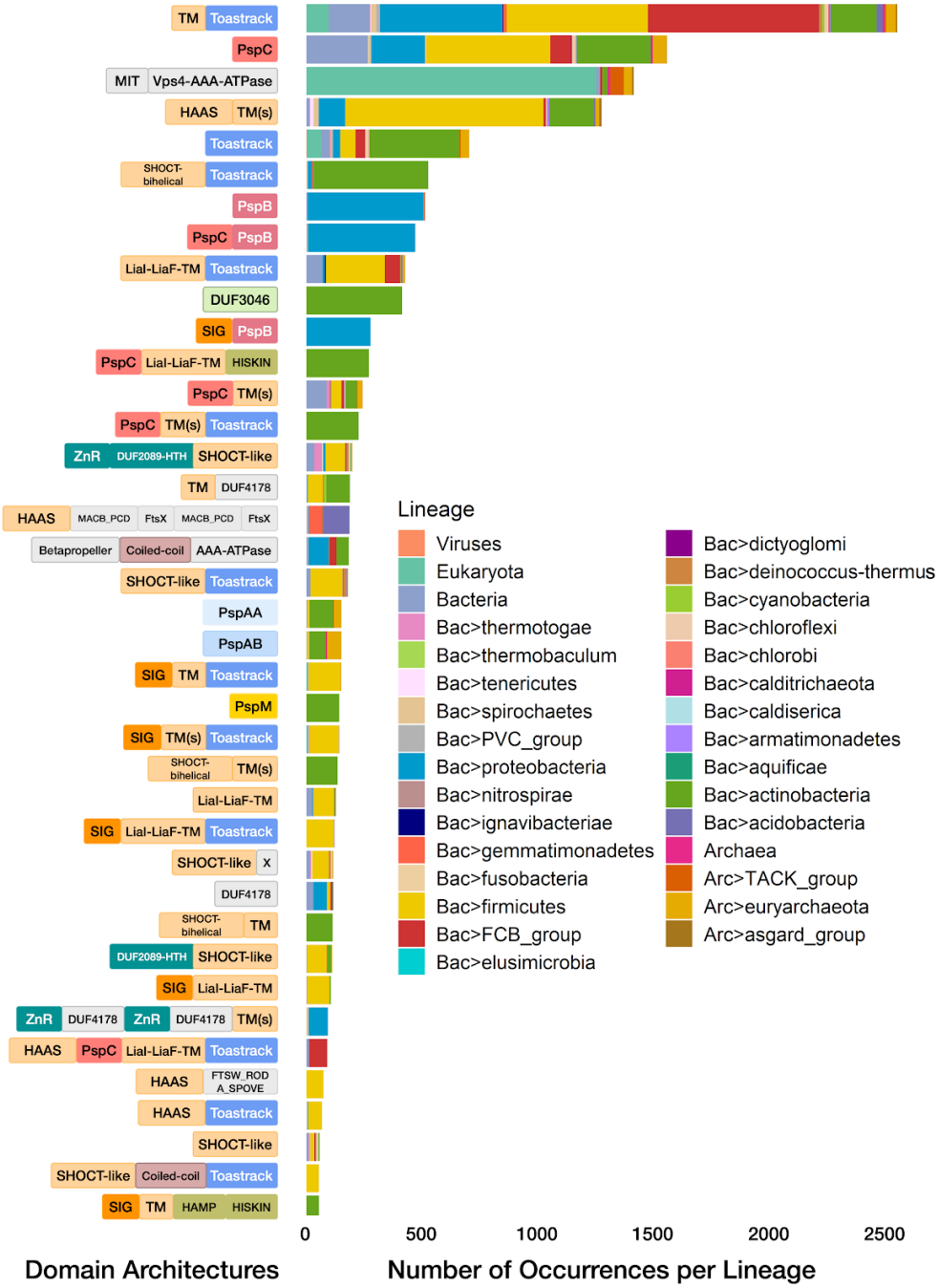


Figure 5. The genomic contexts housing all Psp cognate partner domain homologs.

The genomic contexts are presented using the same schematic as in **Figure 3**. The focus is on Psp partner domains such as Toastrack (blue), PspC, LiaI-LiaF-TM, HAAS, SHOCT-bihelical (in shades of orange), and the various genomic neighborhoods with SHOCT-like proteins, transcription regulators (*e.g.,* PadR-wHTH, SIGMA-HTH, GNTR-HTH), and two-component systems [**Table S2**]. Further details of the genomic contexts of all PspA-free partner domain homologs and their phyletic spreads are in the webapp, and representatives indicated in the figure are shown in **Table S4**.

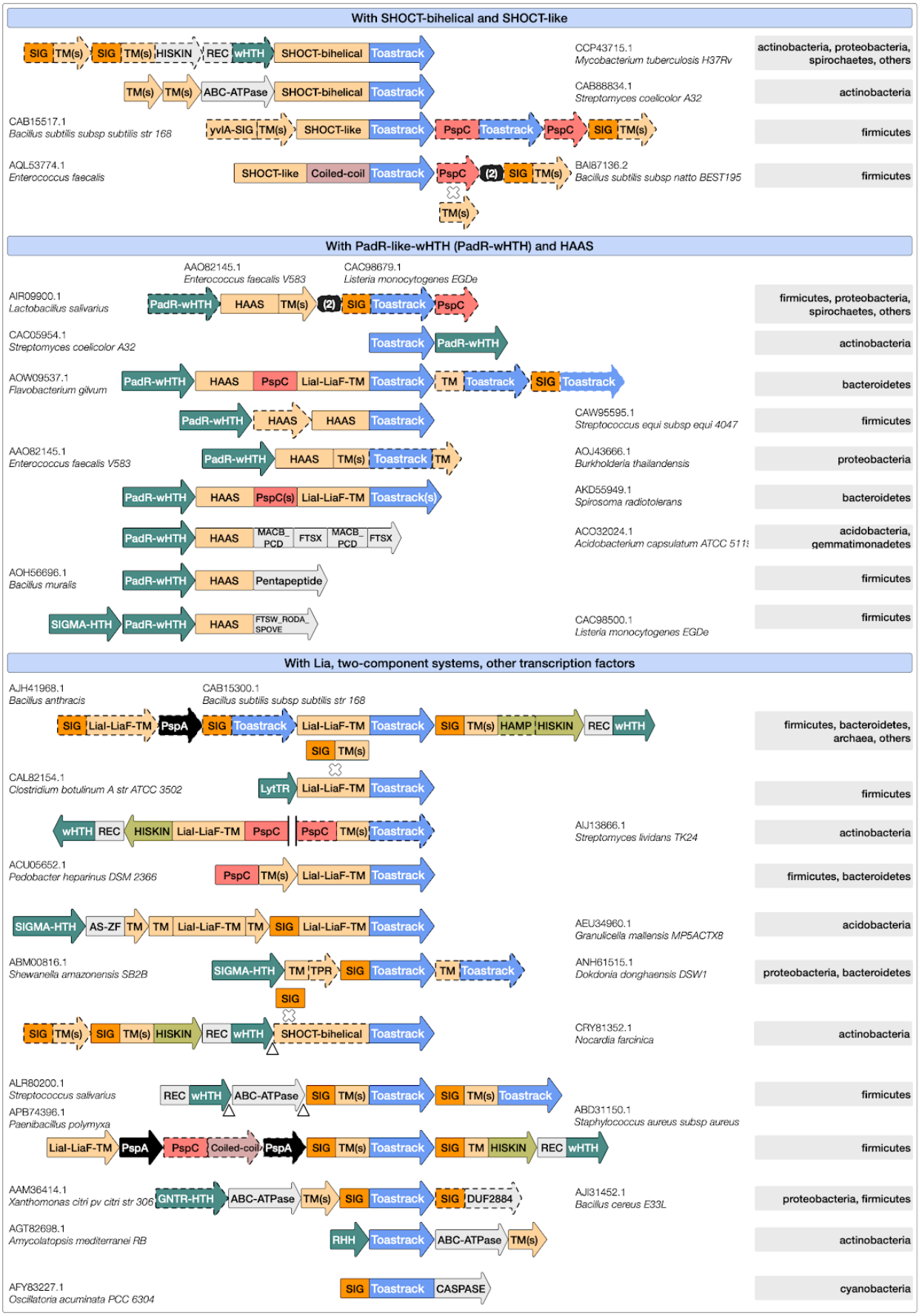
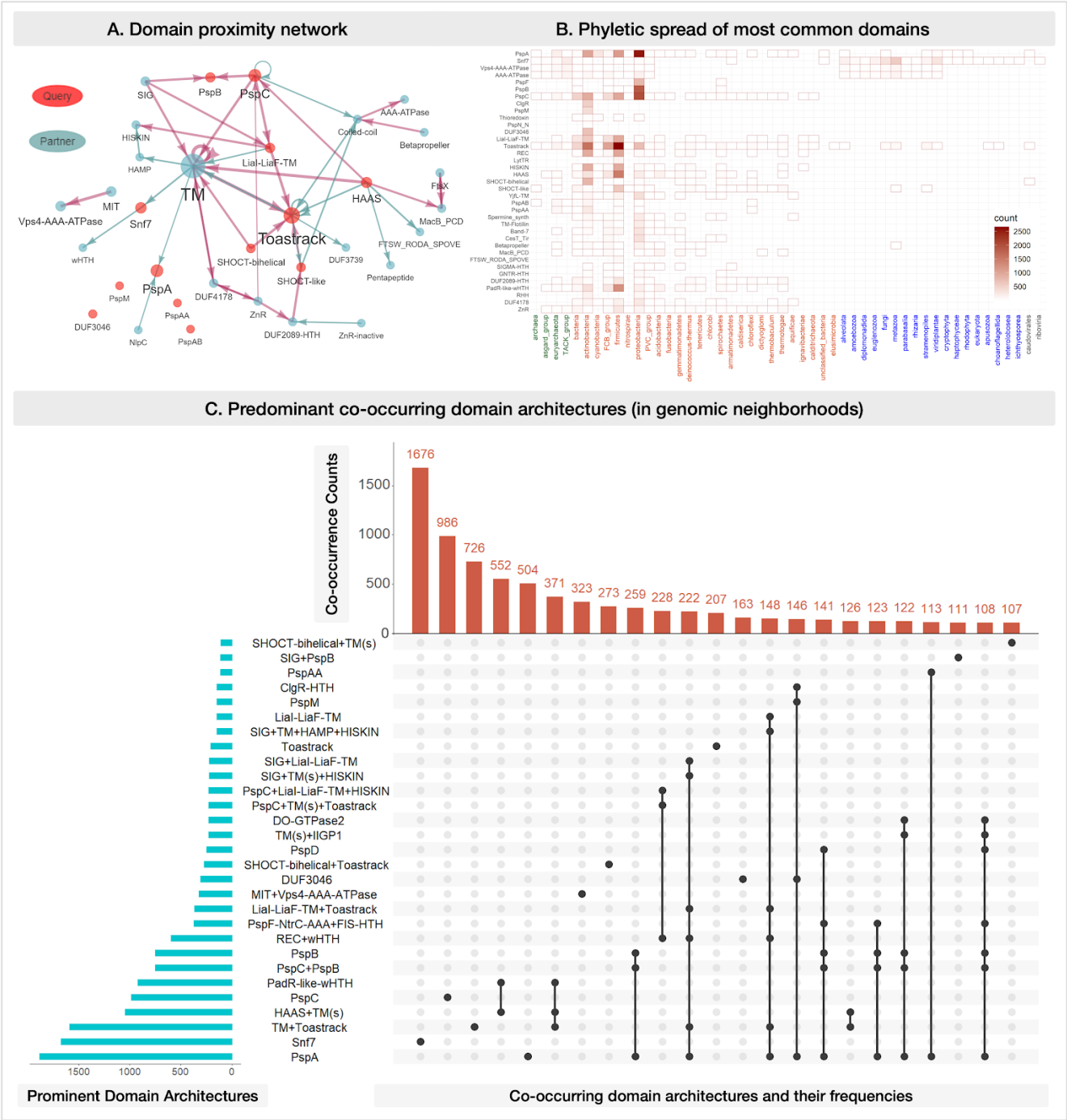

With SHOCT-bihelical and SHOCT-like
SIG TM(s) SIG TM(s) HISKIN REC wHTH SHOCT-bihelical Toastrack
CCP43715.1
Mycobacterium tuberculosis H37Rv
actinobacteria, proteobacteria, spirochaetes, others
TM(s) TM(s) ABC-ATPase SHOCT-bihelical Toastrack
CAB88834.1
Streptomyces coelicolor A32
actinobacteria
CAB15517.1
Bacillus subtilis subsp subtilis str 168
yvlA-SIG TM(s) SHOCT-like Toastrack PspC Toastrack PspC SIG TM(s)
firmicutes
AQL53774.1
Enterococcus faecalis
SHOCT-like Coiled-coil Toastrack PspC (2) SIG TM(s)
TM(s)
BAI87136.2
Bacillus subtilis subsp natto BEST195
firmicutes
With PadR-like-wHTH (PadR-wHTH) and HAAS
AAO82145.1
Enterococcus faecalis V583
CAC98679.1
Listeria monocytogenes EGDe
AIR09900.1
Lactobacillus salivarius
PadR-wHTH HAAS TM(s) (2) SIG Toastrack PspC
firmicutes, proteobacteria, spirochaetes, others
CAC05954.1
Streptomyces coelicolor A32
Toastrack PadR-wHTH
actinobacteria
AOW09537.1
Flavobacterium gilvum
PadR-wHTH HAAS PspC LiaI-LiaF-TM Toastrack TM Toastrack SIG Toastrack
bacteroidetes
PadR-wHTH HAAS HAAS Toastrack
CAW95595.1
Streptococcus equi subsp equi 4047
firmicutes
AAO82145.1
Enterococcus faecalis V583
PadR-wHTH HAAS TM(s) Toastrack TM
AOJ43666.1
Burkholderia thailandensis
proteobacteria
PadR-wHTH HAAS PspC(s) LiaI-LiaF-TM Toastrack(s)
AKD55949.1
Spirosoma radiotolerans
bacteroidetes
PadR-wHTH HAAS MACB_PCD FTSX MACB_PCD FTSX
ACO32024.1
Acidobacterium capsulatum ATCC 511
acidobacteria, gemmatimonadetes
AOH56696.1
Bacillus muralis
PadR-wHTH HAAS Pentapeptide
firmicutes
SIGMA-HTH PadR-wHTH HAAS FTSW_RODA_SPOVE
CAC98500.1
Listeria monocytogenes EGDe
firmicutes
With LiaI, two-component systems, other transcription factors
AJH41968.1
Bacillus anthracis
CAB15300.1
Bacillus subtilis subsp subtilis str 168
SIG LiaI-LiaF-TM PspA SIG Toastrack LiaI-LiaF-TM Toastrack SIG TM(s) HAMP HISKIN REC wHTH
SIG TM(s)
firmicutes, bacteroidetes, archaea, others
CAL82154.1
Clostridium botulinum A str ATCC 3502
LytTR LiaI-LiaF-TM Toastrack
firmicutes
wHTH REC HISKIN LiaI-LiaF-TM PspC PspC TM(s) Toastrack
AIJ13866.1
Streptomyces lividans TK24
actinobacteria
ACU05652.1
Pedobacter heparinus DSM 2366
PspC TM(s) LiaI-LiaF-TM Toastrack
firmicutes, bacteroidetes
SIGMA-HTH AS-ZF TM TM LiaI-LiaF-TM TM SIG LiaI-LiaF-TM Toastrack
AEU34960.1
Granulicella mallensis MP5ACTX8
acidobacteria
ABM00816.1
Shewanella amazonensis SB2B
SIGMA-HTH TM TPR SIG Toastrack TM Toastrack
SIG
ANH61515.1
Dokdonia donghaensis DSW1
proteobacteria, bacteroidetes
SIG TM(s) SIG TM(s) HISKIN REC wHTH SHOCT-bihelical Toastrack
CRY81352.1
Nocardia farcinica
actinobacteria
ALR80200.1
Streptococcus salivarius
REC wHTH ABC-ATPase SIG TM(s) Toastrack SIG TM(s) Toastrack
firmicutes
APB74396.1
Paenibacillus polymyxa
ABD31150.1
Staphylococcus aureus subsp aureus
LiaI-LiaF-TM PspA PspC Coiled-coil PspA SIG TM(s) Toastrack SIG TM HISKIN REC wHTH
firmicutes
AAM36414.1
Xanthomonas citri pv citri str 306
GNTR-HTH ABC-ATPase TM(s) SIG Toastrack SIG DUF2884
AJI31452.1
Bacillus cereus E33L
proteobacteria, firmicutes
AGT82698.1
Amycolatopsis mediterranei RB
RHH Toastrack ABC-ATPase TM(s)
actinobacteria
AFY83227.1
Oscillatoria acuminata PCC 6304
SIG Toastrack CASPASE
cyanobacteria

Figure 6. Psp consolidated

A. Domain proximity network. The network captures co-occurring domains within the top 97% of the homologs of all the 'query' Psp members and their key partner domains (after sorting by decreasing frequency of occurrence). The size of the nodes (domains) and width of edges (co-occurrence of domains within a protein) are proportional to the frequency of their occurrence across homologs. The query domains (original proteins/domains of interest; **Table S1**) and other commonly co-occurring domains [**Table S2**] are indicated in red and grey. *Note*: A few connections may be absent in the displayed network due to low occurrence (fewer than the threshold), *e.g.,* PspA and PspAA, Betapropeller, and AAA-ATPase. The full network, and the domain-centric ones, are available in the webapp. **B. Phyletic spread of the most common domains.** The heatmap shows the presence/absence of homologs of PspA and partner domains across lineages. The color gradient represents the number of homologs identified within each lineage (*e.g.,* the darkest shade indicates the highest number of homologs in a particular lineage). *Rows*: Psp members and their most frequent partners are queried against all sequenced and completed genomes across the three major kingdoms of life. *Columns*: The major archaeal (green), bacterial (orange), eukaryotic (blue), and viral (grey) lineages with representative sequenced genomes are considered. Details of all homologs across the tree of life, their domain architectures, genomic contexts, and their lineage distributions are shown in the webapp (representatives in **Table S3, S4**). **C. Predominant co-occurring domain architectures in genomic neighborhoods.** UpSet plot of the most common neighboring proteins (genomic contexts >100 occurrences are shown) underlying all Psp homologs. *Blue histogram*: Distribution of the predominant domain architectures. *Dots and connections*: Combinations in which these domain architectures come together in the genomic neighborhoods. *Red histogram*: Frequency of occurrences of genomic neighborhoods comprising specific combinations of the predominant domain architectures. Phyletic spreads and UpSet plots of the domain architectures and genomic contexts for the homologs of all Psp member proteins are available in the webapp.


A. Domain proximity network
Query
Partner
SIG
PspB
PspC
AAA-ATPase
Coiled-coil
Betapropeller
HISKIN
HAMP
Lial-LiaF-TM
TM
HAAS
FtsX
MIT
Vps4-AAA-ATPase
Snf7
MacB_PCD
Toastrack
FTSW_RODA_SPOVE
wHTH
SHOCT-bihelical
DUF3739
SHOCT-like
Pentapeptide
PspA
PspM
DUF4178
PspAA
ZnR
DUF3046
NlpC
PspAB
DUF2089-HTH
ZnR-inactive
B. Phyletic spread of most common domains
count
2500
2000
1500
1000
500
C. Predominant co-occurring domain architectures (in genomic neighborhoods)
Co-occurrence Counts
1676
986
726
552
504
371
323
273
259
228
222
207
163
148
146
141
126
123
122
113
111
108
107
SHOCT-bihelical+TM(s)
SIG+PspB
PspAA
ClgR-HTH
PspM
Lial-LiaF-TM
SIG+TM+HAMP+HISKIN
Toastrack
SIG+Lial-LiaF-TM
SIG+TM(s)+HISKIN
PspC+Lial-LiaF-TM+HISKIN
PspC+TM(s)+Toastrack
DO-GTPase2
TM(s)+IIGP1
PspD
SHOCT-bihelical+Toastrack
DUF3046
MIT+Vps4-AAA-ATPase
Lial-LiaF-TM+Toastrack
PspF-NtrC-AAA+FIS-HTH
REC+wHTH
PspB
PspC+PspB
PadR-like-wHTH
PspC
HAAS+TM(s)
TM+Toastrack
Snf7
PspA
1500
1000
500
0
Prominent Domain Architectures
Co-occurring domain architectures and their frequencies

Declarations

Acknowledgments

We are incredibly grateful to Arjun Krishnan and Krishnan Raghunathan for detailed comments on the manuscript and web application, guidance, and support. We also thank Srinand Sreevatsan, Beronda Montgomery, Deborah Johnson, and the MSU Diversity Research Network for support and writing spaces. Finally, we thank members of the JRaviLab for feedback on the web application.

Funding

The following sources supported this work: National Institutes of Health (NIH) grants (R01AI104615, R01HL106788, and R01HL149450) awarded to M.L.G.; NIH Intramural Research Program awarded to V.A. and L.A; Oak Ridge Institute for Science and Education scholarship for the Visiting Scientist program to NIH, the Michigan State University (MSU) College of Veterinary Medicine Endowed Research Funds, and MSU start-up funds awarded to J.R.

Author Contributions

J.R. and M.L.G. conceived the study; J.R., V.A., L.A., and M.L.G. designed the study; J.R. and V.A acquired the data, performed all the analyses, and made the figures and tables. J.R., V.A., L.A., and M.L.G. interpreted the results and wrote the manuscript. S.Z.C built the web application with all the results (data summarization and visualization) with J.R.; S.Z.C. also contributed to making figures and linking identifiers in the manuscript to reference databases. P.D. contributed to renaming some domains.

Data Availability and Reuse

All the data, analyses, and visualizations are available in our interactive and queryable web application: https://jravilab.shinyapps.io/psp-evolution/. Text, figures, and the webapp are licensed under Creative Commons Attribution CC BY 4.0.

References

1. van den Brink-van der Laan E, Killian JA, de Kruijff B. 2004. Nonbilayer lipids affect peripheral and integral membrane proteins via changes in the lateral pressure profile. Biochim Biophys Acta 1666:275–288.
2. Kobayashi R, Suzuki T, Yoshida M. 2007. Escherichia coli phage-shock protein A (PspA) binds to membrane phospholipids and repairs proton leakage of the damaged membranes. Mol Microbiol 66:100–109.
3. Jordan S, Hutchings MI, Mascher T. 2008. Cell envelope stress response in Gram-positive bacteria. FEMS Microbiol Rev 32:107–146.
4. Raivio TL. 2005. Envelope stress responses and Gram-negative bacterial pathogenesis. Mol Microbiol 56:1119–1128.
5. Rowley G, Spector M, Kormanec J, Roberts M. 2006. Pushing the envelope: extracytoplasmic stress responses in bacterial pathogens. Nat Rev Microbiol 4:383–394.
6. Omardien S, Drijfhout JW, van Veen H, Schachtschabel S, Riool M, Hamoen LW, Brul S, Zaat SAJ. 2018. Synthetic antimicrobial peptides delocalize membrane bound proteins thereby inducing a cell envelope stress response. Biochim Biophys Acta Biomembr 1860:2416–2427.
7. Wenzel M, Kohl B, Münch D, Raatschen N, Albada HB, Hamoen L, Metzler-Nolte N, Sahl H-G, Bandow JE. 2012. Proteomic response of Bacillus subtilis to lantibiotics reflects differences in interaction with the cytoplasmic membrane. Antimicrob Agents Chemother 56:5749–5757.
8. Guest RL, Wang J, Wong JL, Raivio TL. 2017. A Bacterial Stress Response Regulates Respiratory Protein Complexes To Control Envelope Stress Adaptation. J Bacteriol 199.
9. Gottesman S. 2017. Stress Reduction, Bacterial Style. J Bacteriol 199.
10. Brissette JL, Weiner L, Ripmaster TL, Model P. 1991. Characterization and sequence of the Escherichia coli stress-induced psp operon. J Mol Biol 220:35–48.
11. Mascher T, Zimmer SL, Smith T-A, Helmann JD. 2004. Antibiotic-inducible promoter regulated by the cell envelope stress-sensing two-component system LiaRS of Bacillus subtilis. Antimicrob Agents Chemother 48:2888–2896.
12. Huvet M, Toni T, Sheng X, Thorne T, Jovanovic G, Engl C, Buck M, Pinney JW, Stumpf MPH. 2011. The evolution of the phage shock protein response system: interplay between protein function, genomic organization, and system function. Mol Biol Evol 28:1141–1155.
13. McDonald C, Jovanovic G, Ces O, Buck M. 2015. Membrane Stored Curvature Elastic Stress Modulates Recruitment of Maintenance Proteins PspA and Vipp1. mBio 6:e01188-01115.
14. Datta P, Ravi J, Guerrini V, Chauhan R, Neiditch MB, Shell SS, Fortune SM, Hancioglu B, Igoshin OA, Gennaro ML. 2015. The Psp system of Mycobacterium tuberculosis integrates envelope stress-sensing and envelope-preserving functions. Mol Microbiol 97:408–422.
15. Ravi J, Anantharaman V, Aravind L, Gennaro ML. 2018. Variations on a theme: evolution of the phage-shock-protein system in Actinobacteria. Antonie Van Leeuwenhoek 111:753–760.
16. Heidrich J, Thurotte A, Schneider D. 2017. Specific interaction of IM30/Vipp1 with cyanobacterial and chloroplast membranes results in membrane remodeling and eventually in membrane fusion. Biochim Biophys Acta Biomembr 1859:537–549.

17. McDonald C, Jovanovic G, Wallace BA, Ces O, Buck M. 2017. Structure and function of PspA and Vipp1 N-terminal peptides: Insights into the membrane stress sensing and mitigation. Biochim Biophys Acta Biomembr 1859:28–39.
18. Lee I-H, Kai H, Carlson L-A, Groves JT, Hurley JH. 2015. Negative membrane curvature catalyzes nucleation of endosomal sorting complex required for transport (ESCRT)-III assembly. Proc Natl Acad Sci USA 112:15892–15897.
19. Liu J, Tassinari M, Souza DP, Naskar S, Noel JK, Bohuszewicz O, Buck M, Williams TA, Baum B, Low HH. 2020. Bacterial Vipp1 and PspA are members of the ancient ESCRT-III membrane-remodelling superfamily. bioRxiv 2020.08.13.249979.
20. Horinouchi T, Suzuki S, Hirasawa T, Ono N, Yomo T, Shimizu H, Furusawa C. 2015. Phenotypic convergence in bacterial adaptive evolution to ethanol stress. BMC Evol Biol 15:180.
21. Kleerebezem M, Crielaard W, Tommassen J. 1996. Involvement of stress protein PspA (phage shock protein A) of Escherichia coli in maintenance of the protonmotive force under stress conditions. EMBO J 15:162–171.
22. Huvet M, Toni T, Tan H, Jovanovic G, Engl C, Buck M, Stumpf MPH. 2009. Model-based evolutionary analysis: the natural history of phage-shock stress response. Biochem Soc Trans 37:762–767.
23. Vothknecht UC, Otters S, Hennig R, Schneider D. 2012. Vipp1: a very important protein in plastids?! J Exp Bot 63:1699–1712.
24. Bultema JB, Fuhrmann E, Boekema EJ, Schneider D. 2010. Vipp1 and PspA: Related but not twins. Commun Integr Biol 3:162–165.
25. Westphal S, Heins L, Soll J, Vothknecht UC. 2001. Vipp1 deletion mutant of Synechocystis: a connection between bacterial phage shock and thylakoid biogenesis? Proc Natl Acad Sci USA 98:4243–4248.
26. Heidrich J, Wulf V, Hennig R, Saur M, Markl J, Sönnichsen C, Schneider D. 2016. Organization into Higher Ordered Ring Structures Counteracts Membrane Binding of IM30, a Protein Associated with Inner Membranes in Chloroplasts and Cyanobacteria. J Biol Chem 291:14954–14962.
27. Theis J, Gupta TK, Klingler J, Wan W, Albert S, Keller S, Engel BD, Schroda M. 2019. VIPP1 rods engulf membranes containing phosphatidylinositol phosphates. 1. Scientific Reports 9:8725.
28. Zhang L, Kato Y, Otters S, Vothknecht UC, Sakamoto W. 2012. Essential role of VIPP1 in chloroplast envelope maintenance in Arabidopsis. Plant Cell 24:3695–3707.
29. Armstrong RM, Adams KL, Zilisch JE, Bretl DJ, Sato H, Anderson DM, Zahrt TC. 2016. Rv2744c Is a PspA Ortholog That Regulates Lipid Droplet Homeostasis and Nonreplicating Persistence in Mycobacterium tuberculosis. J Bacteriol 198:1645–1661.
30. Aseeva E, Ossenbühl F, Eichacker LA, Wanner G, Soll J, Vothknecht UC. 2004. Complex formation of Vipp1 depends on its alpha-helical PspA-like domain. J Biol Chem 279:35535–35541.
31. Jovanovic G, Mehta P, McDonald C, Davidson AC, Uzdavinys P, Ying L, Buck M. 2014. The

N-terminal amphipathic helices determine regulatory and effector functions of phage shock protein A (PspA) in Escherichia coli. J Mol Biol 426:1498–1511.

32. Ranea JAG, Sillero A, Thornton JM, Orengo CA. 2006. Protein superfamily evolution and the last universal common ancestor (LUCA). J Mol Evol 63:513–525.
33. Joly N, Engl C, Jovanovic G, Huvet M, Toni T, Sheng X, Stumpf MPH, Buck M. 2010. Managing membrane stress: the phage shock protein (Psp) response, from molecular mechanisms to physiology. FEMS Microbiol Rev 34:797–827.
34. Manganelli R, Gennaro ML. 2017. Protecting from Envelope Stress: Variations on the Phage-Shock-Protein Theme. Trends Microbiol 25:205–216.
35. Helmann JD. 2016. Bacillus subtilis extracytoplasmic function (ECF) sigma factors and defense of the cell envelope. Curr Opin Microbiol 30:122–132.
36. Eiamphungporn W, Helmann JD. 2008. The Bacillus subtilis sigma(M) regulon and its contribution to cell envelope stress responses. Mol Microbiol 67:830–848.
37. Jordan S, Junker A, Helmann JD, Mascher T. 2006. Regulation of LiaRS-dependent gene expression in bacillus subtilis: identification of inhibitor proteins, regulator binding sites, and target genes of a conserved cell envelope stress-sensing two-component system. J Bacteriol 188:5153–5166.
38. van der Laan M, Urbanus ML, Ten Hagen-Jongman CM, Nouwen N, Oudega B, Harms N, Driessen AJM, Luirink J. 2003. A conserved function of YidC in the biogenesis of respiratory chain complexes. Proc Natl Acad Sci USA 100:5801–5806.
39. Maxson ME, Darwin AJ. 2006. Multiple promoters control expression of the Yersinia enterocolitica phage-shock-protein A (pspA) operon. Microbiology (Reading, Engl) 152:1001–1010.
40. Weiner L, Model P. 1994. Role of an Escherichia coli stress-response operon in stationary-phase survival. Proc Natl Acad Sci USA 91:2191–2195.
41. Jovanovic G, Lloyd LJ, Stumpf MPH, Mayhew AJ, Buck M. 2006. Induction and function of the phage shock protein extracytoplasmic stress response in Escherichia coli. J Biol Chem 281:21147–21161.
42. Mitchell AM, Silhavy TJ. 2019. Envelope stress responses: balancing damage repair and toxicity. Nat Rev Microbiol 17:417–428.
43. Engl C, Beek AT, Bekker M, de Mattos JT, Jovanovic G, Buck M. 2011. Dissipation of proton motive force is not sufficient to induce the phage shock protein response in Escherichia coli. Curr Microbiol 62:1374–1385.
44. Weiner L, Brissette JL, Model P. 1991. Stress-induced expression of the Escherichia coli phage shock protein operon is dependent on sigma 54 and modulated by positive and negative feedback mechanisms. Genes Dev 5:1912–1923.
45. Jovanovic G, Engl C, Buck M. 2009. Physical, functional and conditional interactions between ArcAB and phage shock proteins upon secretin-induced stress in Escherichia coli. Mol Microbiol 74:16–28.
46. Yamaguchi S, Reid DA, Rothenberg E, Darwin AJ. 2013. Changes in Psp protein binding

partners, localization and behaviour upon activation of the Yersinia enterocolitica phage shock protein response. Mol Microbiol 87:656–671.

47. Kleine B, Chattopadhyay A, Polen T, Pinto D, Mascher T, Bott M, Brocker M, Freudl R. 2017. The three-component system EsrISR regulates a cell envelope stress response in Corynebacterium glutamicum. Mol Microbiol 106:719–741.
48. Weiner L, Brissette JL, Ramani N, Model P. 1995. Analysis of the proteins and cis-acting elements regulating the stress-induced phage shock protein operon. Nucleic Acids Res 23:2030–2036.
49. Srivastava D, Moumene A, Flores-Kim J, Darwin AJ. 2017. Psp Stress Response Proteins Form a Complex with Mislocalized Secretins in the Yersinia enterocolitica Cytoplasmic Membrane. mBio 8.
50. DeAngelis CM, Nag D, Withey JH, Matson JS. 2019. Characterization of the Vibrio cholerae Phage Shock Protein Response. J Bacteriol 201.
51. Anh Le TT, Thuptimdang P, McEvoy J, Khan E. 2019. Phage shock protein and gene responses of Escherichia coli exposed to carbon nanotubes. Chemosphere 224:461–469.
52. Kingston AW, Liao X, Helmann JD. 2013. Contributions of the σ(W) , σ(M) and σ(X) regulons to the lantibiotic resistome of Bacillus subtilis. Mol Microbiol 90:502–518.
53. Wolf D, Kalamorz F, Wecke T, Juszczak A, Mäder U, Homuth G, Jordan S, Kirstein J, Hoppert M, Voigt B, Hecker M, Mascher T. 2010. In-depth profiling of the LiaR response of Bacillus subtilis. J Bacteriol 192:4680–4693.
54. Suntharalingam P, Senadheera MD, Mair RW, Lévesque CM, Cvitkovitch DG. 2009. The LiaFSR system regulates the cell envelope stress response in Streptococcus mutans. J Bacteriol 191:2973–2984.
55. Jordan S, Rietkötter E, Strauch MA, Kalamorz F, Butcher BG, Helmann JD, Mascher T. 2007. LiaRS-dependent gene expression is embedded in transition state regulation in Bacillus subtilis. Microbiology (Reading, Engl) 153:2530–2540.
56. Brissette JL, Russel M, Weiner L, Model P. 1990. Phage shock protein, a stress protein of Escherichia coli. Proc Natl Acad Sci USA 87:862–866.
57. Flores-Kim J, Darwin AJ. 2016. Interactions between the Cytoplasmic Domains of PspB and PspC Silence the Yersinia enterocolitica Phage Shock Protein Response. J Bacteriol 198:3367–3378.
58. Castelle CJ, Banfield JF. 2018. Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. Cell 172:1181–1197.
59. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. Nat Microbiol 1:16048.
60. Darwin AJ, Miller VL. 2001. The psp locus of Yersinia enterocolitica is required for virulence and for growth in vitro when the Ysc type III secretion system is produced. Mol Microbiol 39:429–444.
61. Maxson ME, Darwin AJ. 2006. PspB and PspC of Yersinia enterocolitica are dual function

proteins: regulators and effectors of the phage-shock-protein response. Mol Microbiol 59:1610–1623.
62. Becker LA, Bang I-S, Crouch M-L, Fang FC. 2005. Compensatory role of PspA, a member of the phage shock protein operon, in rpoE mutant Salmonella enterica serovar Typhimurium. Mol Microbiol 56:1004–1016.
63. Elderkin S, Bordes P, Jones S, Rappas M, Buck M. 2005. Molecular determinants for PspA-mediated repression of the AAA transcriptional activator PspF. J Bacteriol 187:3238–3248.
64. Aravind L, Ponting CP. 1999. The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins. FEMS Microbiol Lett 176:111–116.
65. Caillat C, Maity S, Miguet N, Roos WH, Weissenhorn W. 2019. The role of VPS4 in ESCRT-III polymer remodeling. Biochem Soc Trans 47:441–448.
66. Schöneberg J, Pavlin MR, Yan S, Righini M, Lee I-H, Carlson L-A, Bahrami AH, Goldman DH, Ren X, Hummer G, Bustamante C, Hurley JH. 2018. ATP-dependent force generation and membrane scission by ESCRT-III and Vps4. Science 362:1423–1428.
67. Weiss P, Huppert S, Kölling R. 2009. Analysis of the dual function of the ESCRT-III protein Snf7 in endocytic trafficking and in gene expression. Biochem J 424:89–97.
68. Godinho RM da C, Crestani J, Kmetzsch L, Araujo G de S, Frases S, Staats CC, Schrank A, Vainstein MH, Rodrigues ML. 2014. The vacuolar-sorting protein Snf7 is required for export of virulence determinants in members of the Cryptococcus neoformans complex. Sci Rep 4:6198.
69. Webster BM, Colombi P, Jäger J, Lusk CP. 2014. Surveillance of nuclear pore complex assembly by ESCRT-III/Vps4. Cell 159:388–401.
70. Flores-Kim J, Darwin AJ. 2012. Phage shock protein C (PspC) of Yersinia enterocolitica is a polytopic membrane protein with implications for regulation of the Psp stress response. J Bacteriol 194:6548–6559.
71. Gueguen E, Savitzky DC, Darwin AJ. 2009. Analysis of the Yersinia enterocolitica PspBC proteins defines functional domains, essential amino acids and new roles within the phage-shock-protein response. Mol Microbiol 74:619–633.
72. Wallrodt I, Jelsbak L, Thomsen LE, Brix L, Lemire S, Gautier L, Nielsen DS, Jovanovic G, Buck M, Olsen JE. 2014. Removal of the phage-shock protein PspB causes reduction of virulence in Salmonella enterica serovar Typhimurium independently of NRAMP1. J Med Microbiol 63:788–795.
73. Burki F, Shalchian-Tabrizi K, Pawlowski J. 2008. Phylogenomics reveals a new “megagroup” including most photosynthetic eukaryotes. Biol Lett 4:366–369.
74. Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AGB, Roger AJ. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic “supergroups.” Proc Natl Acad Sci USA 106:3859–3864.
75. Anantharaman V, Aravind L. 2003. Evolutionary history, structural features and biochemical

diversity of the NlpC/P60 superfamily of enzymes. Genome Biol 4:R11.
76. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, Parsons LM, Pym AS, Samper S, van Soolingen D, Cole ST. 2002. A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proc Natl Acad Sci U S A 99:3684–3689.
77. Tang S, Henne WM, Borbat PP, Buchkovich NJ, Freed JH, Mao Y, Fromme JC, Emr SD. 2015. Structural basis for activation, assembly and membrane binding of ESCRT-III Snf7 filaments. Elife 4.
78. Anantharaman V, Koonin EV, Aravind L. 2001. Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. J Mol Biol 307:1271–1292.
79. Kaur G, Burroughs AM, Iyer LM, Aravind L. 2020. Highly regulated, diversifying NTP-dependent biological conflict systems with implications for the emergence of multicellularity. Elife 9.
80. Tsai JT, Sung N, Lee J, Chang C, Lee S, Tsai FTF. 2019. Crystal Structure of the YcjX Stress Protein Reveals a Ras-Like GTP-Binding Protein. J Mol Biol 431:3179–3190.
81. Wollert T, Wunder C, Lippincott-Schwartz J, Hurley JH. 2009. Membrane scission by the ESCRT-III complex. Nature 458:172–177.
82. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU, Stott MB, Nunoura T, Banfield JF, Schramm A, Baker BJ, Spang A, Ettema TJG. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541:353–358.
83. Kaur G, Iyer LM, Subramanian S, Aravind L. 2018. Evolutionary convergence and divergence in archaeal chromosomal proteins and Chromo-like domains from bacteria and eukaryotes. Sci Rep 8:6196.
84. Burroughs AM, Aravind L. 2016. RNA damage in biological conflicts and the diversity of responding RNA repair systems. Nucleic Acids Res 44:8525–8555.
85. Delahay RM, Shaw RK, Elliott SJ, Kaper JB, Knutton S, Frankel G. 2002. Functional analysis of the enteropathogenic Escherichia coli type III secretion system chaperone CesT identifies domains that mediate substrate interactions. Mol Microbiol 43:61–73.
86. Lee YH, Kingston AW, Helmann JD. 2012. Glutamate dehydrogenase affects resistance to cell wall antibiotics in Bacillus subtilis. J Bacteriol 194:993–1001.
87. Jovanovic G, Mehta P, Ying L, Buck M. 2014. Anionic lipids and the cytoskeletal proteins MreB and RodZ define the spatio-temporal distribution and function of membrane stress controller PspA in Escherichia coli. Microbiology (Reading, Engl) 160:2374–2386.
88. Lee M-J, Huang C-Y, Sun Y-J, Huang H. 2005. Cloning and characterization of spermidine synthase and its implication in polyamine biosynthesis in Helicobacter pylori strain 26695. Protein Expr Purif 43:140–148.
89. Li B, Kurihara S, Kim SH, Liang J, Michael AJ. 2019. A polyamine-independent role for S-adenosylmethionine decarboxylase. Biochem J 476:2579–2594.
90. Cerutti ML, Otero LH, Smal C, Pellizza L, Goldbaum FA, Klinke S, Aran M. 2017. Structural and

functional characterization of a cold adapted TPM-domain with ATPase/ADPase activity. J Struct Biol 197:201–209.
91. Wegener KM, Bennewitz S, Oelmüller R, Pakrasi HB. 2011. The Psb32 protein aids in repairing photodamaged photosystem II in the cyanobacterium Synechocystis 6803. Mol Plant 4:1052–1061.
92. Sirpiö S, Allahverdiyeva Y, Suorsa M, Paakkarinen V, Vainonen J, Battchikova N, Aro E-M. 2007. TLP18.3, a novel thylakoid lumen protein regulating photosystem II repair cycle. Biochem J 406:415–425.
93. Ballas SK, Mohandas N, Marton LJ, Shohet SB. 1983. Stabilization of erythrocyte membranes by polyamines. Proc Natl Acad Sci USA 80:1942–1946.
94. Fiuza M, Letek M, Leiba J, Villadangos AF, Vaquera J, Zanella-Cléon I, Mateos LM, Molle V, Gil JA. 2010. Phosphorylation of a novel cytoskeletal protein (RsmP) regulates rod-shaped morphology in Corynebacterium glutamicum. J Biol Chem 285:29387–29397.
95. Eberhardt RY, Bartholdson SJ, Punta M, Bateman A. 2013. The SHOCT domain: a widespread domain under-represented in model organisms. PLoS ONE 8:e57848.
96. Miller C, Kong J, Tran TT, Arias CA, Saxer G, Shamoo Y. 2013. Adaptation of Enterococcus faecalis to daptomycin reveals an ordered progression to resistance. Antimicrob Agents Chemother 57:5373–5383.
97. Wang L, Xiao S, Chen X, Chen S, Wang S, Wang C, Tang Y, Dong F. 2019. ytiB and ythA Genes Reduce the Uranium Removal Capacity of Bacillus atrophaeus. Int J Mol Sci 20.
98. Boyle-Vavra S, Yin S, Jo DS, Montgomery CP, Daum RS. 2013. VraT/YvqF is required for methicillin resistance and activation of the VraSR regulon in Staphylococcus aureus. Antimicrob Agents Chemother 57:83–95.
99. Gardete S, Wu SW, Gill S, Tomasz A. 2006. Role of VraSR in antibiotic resistance and antibiotic-induced stress response in Staphylococcus aureus. Antimicrob Agents Chemother 50:3424–3434.
100. Wolf D, Domínguez-Cuevas P, Daniel RA, Mascher T. 2012. Cell envelope stress response in cell wall-deficient L-forms of Bacillus subtilis. Antimicrob Agents Chemother 56:5907–5915.
101. Hauf S, Möller L, Fuchs S, Halbedel S. 2019. PadR-type repressors controlling production of a non-canonical FtsW/RodA homologue and other trans-membrane proteins. Sci Rep 9:10023.
102. Leipe DD, Wolf YI, Koonin EV, Aravind L. 2002. Classification and evolution of P-loop GTPases and related ATPases. J Mol Biol 317:41–72.
103. Lane N, Martin WF. 2012. The origin of membrane bioenergetics. Cell 151:1406–1416.
104. Sojo V, Pomiankowski A, Lane N. 2014. A bioenergetic basis for membrane divergence in archaea and bacteria. PLoS Biol 12:e1001926.
105. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. Nucleic Acids Res 41:D36-42.
106. Tatusova T, Ciufo S, Federhen S, Fedorov B, McVeigh R, O'Neill K, Tolstoy I, Zaslavsky L. 2015. Update on RefSeq microbial genomes resources. Nucleic Acids Res 43:D599-605.
107. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix

D, Kenyon R, Machi D, Mao C, Nordberg EK, Olson R, Overbeek R, Pusch GD, Shukla M, Schulman J, Stevens RL, Sullivan DE, Vonstein V, Warren A, Will R, Wilson MJC, Yoo HS, Zhang C, Zhang Y, Sobral BW. 2014. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res 42:D581-591.

108. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, Conrad N, Dietrich EM, Disz T, Gabbard JL, Gerdes S, Henry CS, Kenyon RW, Machi D, Mao C, Nordberg EK, Olsen GJ, Murphy-Olson DE, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Vonstein V, Warren A, Xia F, Yoo H, Stevens RL. 2017. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res 45:D535–D542.
109. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733-745.
110. Federhen S. 2012. The NCBI Taxonomy database. Nucleic Acids Res 40:D136-143.
111. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.
112. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.
113. Söding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33:W244-248.
114. Nielsen H. 2017. Predicting Secretory Proteins with SignalP. Methods Mol Biol 1611:59–73.
115. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580.
116. Käll L, Krogh A, Sonnhammer ELL. 2004. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338:1027–1036.
117. Käll L, Krogh A, Sonnhammer ELL. 2007. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Res 35:W429-432.
118. Cole C, Barber JD, Barton GJ. 2008. The Jpred 3 secondary structure prediction server. Nucleic Acids Res 36:W197-201.
119. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: The protein families database in 2021. Nucleic Acids Res 49:D412–D419.
120. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity

searching. Nucleic Acids Res 39:W29-37.
121. Lassmann T. 2019. Kalign 3: multiple sequence alignment of large data sets. Bioinformatics https://doi.org/10.1093/bioinformatics/btz795.
122. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797.
123. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS ONE 5:e9490.
124. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. 2019. shiny: Web Application Framework for R.
125. Winston Chang RCT Joe Cheng, JJ Allaire, Yihui Xie, Jonathan McPherson, RStudio, jQuery Foundation, jQuery contributors, jQuery UI contributors, Mark Otto, Jacob Thornton, Bootstrap contributors, Twitter, Inc, Alexander Farkas, Scott Jehl, Stefan Petre, Andrew Rowls, Dave Gandy, Brian Reavis, Kristopher Michael Kowal, es5-shim contributors, Denis Ineshin, Sami Samhuri, SpryMedia Limited, John Fraser, John Gruber, Ivan Sagalaev. 2020. Web Application Framework for R.
126. R Core Team. 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
127. RStudio Team. 2020. RStudio: Integrated Development Environment for R. RStudio, PBC., Boston, MA.
128. Mike Bostock CY Kerry Rodden, Kevin Warne, Kent Russell, Florian Breitwieser. 2020. Sunburst “Htmlwidget.”
129. Hao ZhuThomas Travison DM Timothy Tsai, Will Beasley, Yihui Xie, GuangChuang Yu, Stéphane Laurent, Rob Shepherd, Yoni Sidi, Brian Salzer, George Gui, Yeliang Fan. 2020. Construct Complex Table with “kable” and Pipe Syntax.
130. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. 2014. UpSet: Visualization of Intersecting Sets. IEEE Transactions on Visualization and Computer Graphics 20:1983–1992.
131. Enrico Bonatesta UB Christoph Horejs-Kainrath. 2020. Multiple Sequence Alignment.
132. Emmanuel Paradis D de V Simon Blomberg, Ben Bolker, Joseph Brown, Santiago Claramunt, Julien Claude, Hoa Sien Cuong, Richard Desper, Gilles Didier, Benoit Durand, Julien Dutheil, RJ Ewing, Olivier Gascuel, Thomas Guillerme, Christoph Heibl, Anthony Ives, Bradley Jones, Franz Krah, Daniel Lawson, Vincent Lefort, Pierre Legendre, Jim Lemon, Guillaume Louvel, Eric Marcon, Rosemary McCloskey, Johan Nylander, Rainer Opgen-Rhein, Andrei-Alin Popescu, Manuela Royer-Carenzi, Klaus Schliep, Korbinian Strimmer. 2020. Analyses of Phylogenetics and Evolution. R.
133. Fellows I. 2018. Word Clouds.
134. Dawei Lang GC. 2018. Create Word Cloud by ’htmlwidget.
135. Jake Conway NG. 2019. A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets.
136. Müller K. 2017. here: A Simpler Way to Find Your Files.
137. Wickham H, Mara Averick HY Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan,

Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo. 2019. Welcome to the tidyverse. Journal of Open Source Software 4:1686.

138. Martin Morgan MR. 2019. Access the Bioconductor Project Package Repository.
139. Perry M. 2016. Flexible Heatmaps for Functional Genomics and Sequence Features.
140. Xie Y, Allaire JJ, Grolemund G. 2018. R Markdown: The Definitive Guide. Chapman and Hall/CRC, Boca Raton, Florida.
141. Allaire JJ, Xie, Y, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W, Iannone R. 2020. rmarkdown: Dynamic Documents for R.