The evolution and role of eukaryotic-like domains in environmental intracellular bacteria: the battle with a eukaryotic cell

Jessica Martyn, Laura Gomez-Valero, Carmen Buchrieser

The evolution and role of eukaryotic-like domains in environmental intracellular bacteria: the battle with a eukaryotic cell

Jessica E. Martyn, Laura Gomez-Valero & Carmen Buchrieser*

Institut Pasteur, Biologie des Bactéries Intracellulaires and CNRS UMR 3525, Paris, France,

Keywords: Intracellular bacteria, eukaryotic-like domains, *Legionella*, *Chlamydia*, evolution, horizontal gene transfer, virulence

*For correspondence:

Carmen Buchrieser

Biologie des Bactéries Intracellulaires

Institut Pasteur

28, rue du Dr. Roux, 75724 Paris Cedex 15, France

Tel: (33-1)-45-68-83-72

Fax: (33-1)-45-68-87-86

E-mail: **cbuch@pasteur.fr**

ABSTRACT

Intracellular pathogens that are able to thrive in different environments, such as *Legionella* spp. which preferentially live in protozoa in aquatic environments or environmental Chlamydiae which replicate either within protozoa or a range of animals, possess a plethora of cellular biology tools to influence their eukaryotic host. The host manipulation tools that evolved in the interaction with protozoa, confer these bacteria the capacity to also infect phylogenetically distinct eukaryotic cells, such as macrophages and thus they can also be human pathogens. To manipulate the host cell, bacteria use protein secretion systems and molecular effectors. Although these molecular effectors are encoded in bacteria, they are expressed and function in a eukaryotic context often mimicking or inhibiting eukaryotic proteins. Indeed, many of these effectors have eukaryotic-like domains. In this review we propose that the main pathways environmental intracellular bacteria need to subvert in order to establish the host eukaryotic cell as a replication niche are chromatin remodelling, ubiquitination signalling, and modulation of protein-protein interactions *via* tandem repeat domains. We then provide mechanistic insight into how these proteins might have evolved as molecular weapons. Finally, we highlight that in environmental intracellular bacteria the number of eukaryotic-like domains and proteins is considerably higher than in intracellular bacteria specialised to an isolated niche, such as obligate intracellular human pathogens. As mimics of eukaryotic proteins are critical components of host pathogen interactions, this distribution of eukaryotic-like domains suggests that the environment has selected them.

One-sentence summary

This review aims to dissect the evolutionary processes that may lead to an enrichment of eukaryotic-like proteins in environmental intracellular bacteria through continuous adaptation to multiple eukaryotic hosts. We reveal that chromatin and ubiquitin modulation functions, as well as adaptable tandem repeat domains are crucial for their virulence in protozoa and accidentally also in humans.

INTRODUCTION

Bacteria exploit all niches of the abiotic environment, but also form mutualistic or competitive relationships with other bacteria and archaea. When eukaryotes evolved; bacteria were a food source for unicellular eukaryotes but equally a new environment to be exploited by bacteria that were able to fight the digestion by these unicellular eukaryotes to live parasitically or symbiotically with them. Frequent biological interactions of multiple genomes during co-evolution, results in adaptation of the respective gene repertoire.

To interact with their environment, bacteria use protein secretion systems to secrete molecular effectors that manipulate other organisms (Green & Mecsas, 2016). For bacteria that interact or inhabit eukaryotes, molecular effectors act in a eukaryotic context and have often been found to share homology with eukaryotic domains or proteins, and may thus be described as "eukaryotic-like" (Stebbins & Galan, 2001, Cazalet *et al.*, 2004, de Felipe *et al.*, 2005, Elde & Malik, 2009, Mondino *et al.*, 2020). Indeed, eukaryotic-like domains have been used as a prediction criterion to detect molecular effectors against eukaryotes encoded in a bacterial genome (Burstein *et al.*, 2009, Cazalet *et al.*, 2010, Jehl *et al.*, 2011). Eukaryotic-like domains and proteins are commonly found to retain their predicted eukaryotic function (Mondino *et al.*, 2020). To define a eukaryotic-like domain, Gomez-Valero and colleagues proposed that the domain must predominantly (at least 75 %) be found in eukaryotic genomes (Gomez-Valero *et al.*, 2019). Eukaryotic-like proteins are defined as bacterial proteins that have more than 20 % amino acid identity to eukaryotic proteins, over more than a third of the protein length (Gomez-Valero *et al.*, 2019). When mining the genomes of bacterial genera that contain environmental intracellular bacteria, common themes in eukaryotic-like domains emerge. The most common and enriched eukaryotic-like domains in these genomes are tandem repeat containing domains and domains that mimic or interfere with host ubiquitin signalling and chromatin remodelling domains. Tandem repeat domains include Ankyrin (ANK), Tetratricopeptide repeat (TPR), Pentatricopeptide repeat (PPR) and Leucine rich repeat (LRR). In bacteria, tandem repeat domains provide a surface to mediate protein-protein or nucleic acid interactions for fundamental biological functions and therefore likely directly engage host proteins. Proteins with tandem repeat domains are commonly associated with additional functional domains, such as domains which mediate ubiquitinylation, lipid metabolism or small GTPases (Rolando & Buchrieser, 2012, Burstein *et al.*, 2016).

A correlation between bacterial lifestyle and eukaryotic-like domain enrichment has been observed; environmental intracellular bacteria (*e.g.* amoeba-resistant bacteria) possess significantly more eukaryotic-like domains compared with extracellular bacteria and even

host-restricted intracellular bacteria (*e.g* obligate intracellular human pathogens) (Jernigan & Bordenstein, 2014). Intracellular bacteria that are mainly found in the environment infecting a wide range of different hosts, but that are never, or only accidentally human pathogens are referred to as environmental intracellular bacteria with an open host range. However, host specialism is not a binary term, but on a scale and that host specialisation may exist to different degrees (Amaro *et al.*, 2015, Park *et al.*, 2020). Many environmental intracellular bacteria are frequently exposed to and survive more than one species of eukaryotes, in particular unicellular eukaryotes which are ubiquitous and make up more than twice the biomass on earth as animals (Bar-On *et al.*, 2018). An example of one such environmental, intracellular bacterial genus is *Legionella*, the etiological agents of Legionnaires disease a severe pneumonia (McDade *et al.*, 1977, Mondino *et al.*, 2020). *Legionella* spp. survive and replicate in many species of amoebae as well as human macrophages; and are enriched with eukaryotic-like domains and proteins, harbouring at least 137 different eukaryotic-like domains and more than 200 eukaryotic-like proteins (Escoll *et al.*, 2013, Boamah *et al.*, 2017, Gomez-Valero *et al.*, 2019). In contrast to environmental intracellular bacteria which have an open host range, some intracellular bacteria have a restricted set of host species. The phylum Chlamydiae, which are obligate intracellular bacteria and one of the oldest successful eukaryote-associated prokaryotic lineages, possesses both a host-restricted family (Chlamydiaceae) and open host range environmentally distributed families (*e.g*. Parachlamydiaceae, Simkaniaceae and Waddliaceae) (Horn, 2008, Kostlbacher *et al.*, 2021). Chlamydiaceae encompass well known obligate animal and human pathogens, such as *Chlamydia trachomatis*, *Chlamydia pneumoniae* and *Chlamydia felis* Fe/C-56 (Horn, 2008, Kostlbacher *et al.*, 2021). Environmental Chlamydiae include species, such as *Parachlamydia acanthamoeba, Protochlamydia amoebophila, Estrella lausannensis*, or *Simkania negevensis* (Horn, 2008, Collingro *et al.*, 2020, Kostlbacher *et al.*, 2021). Like *Legionella* spp., environmental Chlamydiae are intracellular bacteria that have been found to survive and replicate in a number of species of amoebae (Horn, 2008). Furthermore, there is growing evidence that some environmental Chlamydiae, such as *S. negevensis* and *P. acanthamoebae* also cause respiratory tract infections in humans (Lamoth & Greub, 2010). Interestingly, like *Legionella* spp., environmental Chlamydiae are enriched in eukaryotic-like domains and proteins, however this enrichment is not observed in pathogenic Chlamydiae (Schmitz-Esser *et al.*, 2010, Collingro *et al.*, 2011).

One hypothesis put forward to explain the presence of eukaryotic-like domains and proteins in environmental intracellular bacteria such as *Legionella* spp. or environmental

Chlamydiae, is that eukaryotic genes undergo interdomain horizontal gene transfer in protozoan hosts (Cazalet *et al.*, 2004, de Felipe *et al.*, 2005, Cazalet *et al.*, 2010, Lurie-Weinberger *et al.*, 2010, Collingro *et al.*, 2011, Bertelli & Greub, 2012, Gomez-Valero L., 2013). In interdomain horizontal gene transfer the bacterium acquires foreign genetic material from eukaryotic cells, which is incorporated into the bacterial genome and retains part of its function. The enriched presence of eukaryotic-like domains and proteins in environmental intracellular bacteria, suggests that frequent interaction with numerous hosts, in particular protozoa and co-evolution with protozoan hosts provide an environment for interdomain horizontal gene transfer. In *Legionella* spp. phylogenetic and sequence analysis has demonstrated interdomain horizontal gene transfer, in particular the sporadic distribution of these eukaryotic-like domains and proteins suggests that there have been different acquisition and selection events (Gomez-Valero *et al.*, 2019). Interdomain horizontal gene transfer has not been experimentally elucidated from eukaryote to prokaryote but has from prokaryote to eukaryote (Pitzschke & Hirt, 2010). The enriched presence of eukaryotic-like domains and proteins in environmental intracellular bacteria in particular is thought to be due to frequent interaction with numerous hosts, in particular protist, which acts as a selection pressure for gene transfer.

In this review we compare and contrast eukaryotic-like domains and proteins found in *Legionella* spp. and environmental Chlamydiae, two environmental intracellular bacteria that can also be pathogenic to humans. We assess functional themes in eukaryotic-like domains that seem essential to support intracellular life of *Legionella* spp. and environmental Chlamydiae. We then focus on how these eukaryotic-like domains or proteins might occur in bacteria and expanded to the advantage of the bacterium. Finally, we explore why the environment might select for more eukaryotic-like domain or proteins in *Legionella* spp. and environmental Chlamydiae.

COMMON HOST SUBVERSION STRATEGIES ARE USED BY ENVIRONMENTAL INTRACELLULAR BACTERIA

As mimics of eukaryotic proteins, eukaryotic-like domains and proteins are key players in bacterial eukaryote interactions (Mondino *et al.*, 2020). They have been found to be involved in nearly every signalling pathway in the eukaryotic cell, such as: (1) disruption of endocytic and autophagic targeting of membrane bound compartments to avoid digestion in the lysosome, (2) transforming the phagosome into a replicative niche (3) and maintaining the integrity of the replicative niche by avoiding bacterial detection by host innate immune

recognition and host cell death (Mondino *et al.*, 2020). As more genome data become available, it becomes evident, that there are common themes in the role of eukaryotic-like domains with enzymatic activity found in environmental intracellular bacteria. When analysing the number of eukaryotic-like domains per Megabase (**Figure 1**), we observed an enrichment in F-box domains, which have a role in host ubiquitination signalling, in the genomes of *Legionella* spp. and environmental Chlamydiae **(Figure 1 & Table 1)** (Domman *et al.*, 2014, Gomez-Valero *et al.*, 2019). Additionally, eukaryotic-like repeat containing domains (ANK, TPR, PPR etc.) which have a role in protein-protein and protein-nucleic acid interactions are also enriched in environmental intracellular bacteria and found in combination with many different eukaryotic-like domains (Domman *et al.*, 2014, Burstein *et al.*, 2016, Gomez-Valero *et al.*, 2019). Here we will explore why eukaryotic-like domains with functions in these processes may be important for environmental intracellular bacteria focusing primarily on protozoa.

- **Chromatin modulation and histone modification**

Multiple bacterial effectors with eukaryotic-like domains target the nucleus and interfere with the host transcriptional machinery including chromatin remodelling, DNA replication and repair (Rolando *et al.*, 2015). Eukaryotic cells, including unicellular eukaryotes, package large amounts of DNA into the nucleus, without compromising replication, repair, transcription, and chromosome segregation. DNA is packed into the cell into a highly organized structure called chromatin, by two classes of proteins, histones and chromatin remodelling proteins (Luger *et al.*, 1997). For gene expression, DNA replication, repair, or recombination, chromatin remodelling complexes use ATP hydrolysis to unwind DNA and/or reposition nucleosomes. They are organised into four classes defined by their ATPase subunit: SWI/SNF, ISWI, Mi-2, and Ino80 (Kallin & Zhang, 2004). Each of these complexes interact with sequence-specific DNA-binding factors at the targeted genes and can have both a positive and a negative role on transcription (Kallin & Zhang, 2004). Additionally, histone post-translational modifications, including phosphorylation, acetylation, methylation, and ubiquitylation, regulate transcription either by affecting the chromatin structure directly, and/or by recruiting non-histone proteins such as transcription factors (Bracha *et al.*, 2003, Kallin & Zhang, 2004, Rolando *et al.*, 2015). In single celled eukaryotes, such as amoebae gene expression is also controlled using these methods. For example, in *Entamoeba histolytica,* rapid silencing occurs at the transcriptional level (RNAi), and stable multi-generation silencing is achieved at post-transcriptional level *via* a loss of H3K4 methylation

and enrichment of H3K27Me2 (Huguenin *et al.*, 2010, Foda & Singh, 2015). Therefore, chromatin structure and histone modifications are key regulators of eukaryotic transcription broadly and thereby also effective targets for intracellular bacteria during an infection.

Bacterial interactions with host cells can induce dynamic transcriptional responses in the host cell. For example, RNA sequencing analysis on *Legionella* infected *Acanthamoeba castellanii* observed upregulation of vesicle transport and small GTPase mediated signal transduction, and downregulation of energy metabolism and cell cycle (Li *et al.*, 2020). These transcriptional responses can be driven by host chromatin modulators. For example, sirtuin proteins which are NAD dependent deacetylases within the class III histone deacetylase family, are upregulated during *Legionella* infection of *A. castellani* (Li *et al.*, 2020). In particular, *sir6f* a member of the class IV siturins which are involved in DNA repair (Li *et al.*, 2020).

Indeed, eukaryotic-like domains predicted to code for enzymatic functions that modulate chromatin, are present in the majority of *Legionella* spp. and Chlamydiae (**Figure 1 & Table 1**). Some of these proteins also have a nuclear localisation sequence allowing these proteins to be targeted to the nucleus. One of the most characterised eukaryotic-like domains which have a role in chromatin modulation in *Legionella* spp. and Chlamydiae are su(var)3-9, enhancer of zeste, trithorax (SET) methyltransferase domain containing proteins which mimic host histone methyltransferases (Murata *et al.*, 2007, Pennini *et al.*, 2010, Rolando *et al.*, 2013, Rolando *et al.*, 2015). In *L. pneumophila* it was shown that the SET domain encoding protein RomA, methylates host histones thereby repressing gene transcription and promoting intracellular bacterial replication (Rolando *et al.*, 2013). The SET domain is present in ~ 80 % of *Legionella* spp. suggesting the ability of many *Legionella* spp. to manipulate host chromatin (Gomez-Valero *et al.*, 2019). The SET domain is also present in ~ 80 % of environmental Chlamydiae and all pathogenic Chlamydiae sequences analysed (**Figure 1 & Table 1**) (Stephens *et al.*, 1998, Pennini *et al.*, 2010). Other domains with a role in chromatin modulation have been found in *Legionella* spp. and Chlamydiae. For example, pathogenic and environmental Chlamydiae possess the SWIB/MDM2 domain that belongs to the SWI/SNF family of complexes, which are ATP-dependent chromatin remodelling proteins involved in transcriptional activation (**Figure 1 & Table 1**) (Stephens *et al.*, 1998). The SWIB domain and the SET domain often co-occur in Chlamydiae (**Figure 1 & Table 1**) (de Barsy *et al.*, 2019). Recently, a SWIB domain protein was characterised in *Waddlia chondrophila* where it was found to localise to the host nucleus and binds along the genome (de Barsy *et al.*, 2019). Both *Legionella* spp. and Chlamydiae possess Rossmann-fold histone

methylase domains: protein arginine methyltransferases (PRMT) and disruptor-of-telomeric silencing (Dot1), shown to positively or negatively regulate transcription respectively in eukaryotes, however these are uncharacterised in these bacteria **(Figure 1 & Table 1)** (Aravind *et al.*, 2011, Rolando *et al.*, 2015, Gomez-Valero *et al.*, 2019). Interestingly, VipF, one of the effectors found in all *Legionella* spp. possess a GCN5-Related N acetyltransferase (GNAT) domain found in histone acetyltransferase, suggesting that it is also involved in chromatin remodelling (Dutnall *et al.*, 1998, Gomez-Valero *et al.*, 2019).

Taken together, all *Legionella* spp. and Chlamydiae analysed to date, independent of their lifestyle encode at least one, but often several chromatin modulating proteins. Intracellular bacteria therefore require functions allowing them to target this central organelle directly to manipulate the cellular control centre to their advantage.

- **Targeting of the ubiquitination machinery**

Ubiquitin is a highly conserved 76 amino acid protein that regulates the activities of the ATP-dependent ubiquitination activating enzyme (E1), ubiquitin conjugating enzyme (E2) and ubiquitin ligase (E3). Ubiquitination can be reversed by deubiquitinating enzymes (DUBS). Around between 1.5 and 4.5 % of the genes in eukaryotic genomes (including *Dictyostelium discoideum*) are putatively involved in creating, reading and erasing ubiquitin (Pergolizzi *et al.*, 2019, Pohl & Dikic, 2019). Ubiquitin is a post-translational modification that is important in eukaryotes in protein interaction, activity, localisation and degradation (Pergolizzi *et al.*, 2019, Pohl & Dikic, 2019). Genes which are important to regulation in eukaryotic cells, such as transcription factors, signal transduction proteins, cell cycle control proteins, proteins that regulate cell death and apoptosis are subject to ubiquitination (Geng *et al.*, 2012, Pohl & Dikic, 2019). Ubiquitin degradation allows for rapid changes in gene transcription and protein synthesis, including during infection. Ubiquitylation is also involved in the activation of immune responses and antigen presentation when intracellular bacteria are present (Rahman & McFadden, 2011, Hu & Sun, 2016). For example, ubiquitylation functions as a signal for nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB) activation, which triggers a broad range of host inflammatory responses that prevent bacterial proliferation. In addition, ubiquitylation can also trigger the host cell to kill bacterial pathogens *via* proteasome, phagolysosome and autophagosome mediated degradation pathways (Pohl & Dikic, 2019). Given the many different cellular processes that are regulated by the ubiquitination machinery, targeting the host ubiquitination system is key for intracellular bacteria to manipulate the host cell to their advantage.

It has been shown that multiple bacterial effectors target the ubiquitin system, many achieve this *via* eukaryotic-like domains but some even evolve novel structures or new enzymatic activities to modulate various steps of the host ubiquitin pathway (Zhou & Zhu, 2015). Modulation of the ubiquitin system for successful eukaryotic infection is also seen in the multitude of pathogens including *Salmonella* spp.*, Shigella* spp., pathogenic *E. coli* spp., pathogenic Chlamydiae and *Legionella* spp. (Zhou & Zhu, 2015). How *Legionella* spp. interfere with the ubiquitin system has been extensively studied and many enzymes that catalyse ubiquitin through conventional, unconventional and novel mechanisms have been identified, this have been summarised in recent reviews (Kitao *et al.*, 2020, Price & Abu Kwaik, 2021). Here we will focus on eukaryotic-like domains mimicking proteins that are part of the eukaryotic ubiquitin system, such as Really Interesting New Gene (RING) or U-box domains (E3 ligase mimics), BR-C, ttk and bab (BTB)/Pox virus and Zinc finger (POZ) BTB/POZ domains or F-box domains (ubiquitin pathway protein-protein interaction mimics) and Ovarian Tumour (OTU) deubiquitinase or Ubiquitin Like specific Protease 1 (ULP) domains (Deubiquitination mimics) (Perez-Torrado *et al.*, 2006, Price & Kwaik, 2010, Zheng & Shabek, 2017).

Domains which are E3 ligase and ubiquitin pathway protein-protein interaction mimics are found in *Legionella* spp. and environmental Chlamydiae, but are absent in pathogenic Chlamydiae analysed **(Figure 1 & Table 1)**. Nearly all *Legionella* genomes analysed to date contain U-box encoding proteins (1 to 3 per genome) (Gomez-Valero *et al.*, 2019). In *L. pneumophila* U-box containing proteins have been functionally analysed and have been shown to act as E3 ubiquitin ligases in the host cell (Kubori *et al.*, 2008). The U-box domain is found in some but not all environmental Chlamydiae with as many as 14 per genome (Domman *et al.*, 2014). F-box or BTB/POZ domains mediate the ubiquitination of proteins by binding target proteins to the E3 ubiquitin ligase complex (Perez-Torrado *et al.*, 2006, Price & Kwaik, 2010). Nearly all *Legionella* genomes analysed to date contain an F-box domain (1 to 18 per genome) (Gomez-Valero *et al.*, 2019). F-box containing proteins are also notably expanded in environmental Chlamydiae, sometimes encoding 120 copies per genome (Domman *et al.*, 2014). In *L. pneumophila* the F-box recruits targets by binding to the E3 ubiquitin ligase complex and other *Legionella* effectors for different purposes (Kubori *et al.*, 2008, Kubori *et al.*, 2010, Lomma *et al.*, 2010, Price *et al.*, 2010, Price *et al.*, 2011). The BTB/POZ domain, which has been shown to function both as a F-box like domain and to mediate transcriptional repression by interacting with components of histone deacetylase co-repressor complexes in eukaryotes, is notably expanded in environmental Chlamydiae

(**Figure 1 & Table 1**) (Collins *et al.*, 2001, Stogios *et al.*, 2005, Perez-Torrado *et al.*, 2006, Domman *et al.*, 2014). The BTB domain may have a role in ubiquitin–proteasome regulated transcriptional control due to its dual function, however further investigation is required. If true, this dual role maybe very useful for environmental intracellular bacteria. Deubiquitinase (DUB) domains, such as OTU deubiquitinase and ULP1 domains which have been shown to mediate inhibition of autophagy, NF-κB signalling or cell death, during infection, are common to *Legionella* spp., environmental and pathogenic Chlamydiae (**Figure 1 & Table 1**) (Le Negrate *et al.*, 2008, Sheedlo *et al.*, 2015, Fischer *et al.*, 2017). This suggests that the role of DUB domains is beneficial for all intracellular lifestyles.

Host-derived ubiquitin conjugating enzymes and ubiquitin ligases also participate in host pathogen interactions. Recently it has been shown that the E2 ubiquitin conjugating enzyme, UBE2E1 and the E3 ubiquitin ligase, CUL7, both aid in the translocation of *L. pneumophila* effector proteins and influence intracellular persistence and survival of the pathogen (Ong *et al.*, 2021). Furthermore, the analysis of the *Dictyiostelium discoideum* transcriptome upon infection with *L. pneumophila* revealed the genes of the ubiquitination machinery are strongly upregulated after infection with *L. pneumophila* (Farbrother *et al.*, 2006).

These different results provide strong evidence that manipulation of the host ubiquitin system is fundamental for bacteria to replicate intracellularly, in particular in protozoa.

- **Exploitation of tandem repeat containing domains**

Tandem repeat containing eukaryotic-like domains include: HEAT, RCC1, LRR, WD40, PPR, MORN, Sel-1, TPR and ANK (Newton *et al.*, 2007, Coil *et al.*, 2008, Doxey & McConkey, 2013, Domman *et al.*, 2014, Swart *et al.*, 2020). In eukaryotes, tandem repeat proteins mediate protein–protein and protein–nucleic acid interactions that are central to a variety of cellular processes. Tandem repeat proteins contain tandem arrays of small structural motifs, the individual units of tandem repeat proteins are too small, ~ 20-30 amino acids, for them to be independently folded and they are found in arrays of at least three units but as many as 20 units (Grove *et al.*, 2008). For example, ANK and TPR repeats are found in small arrays (typically 3–6 repeats) and LRR and HEAT repeats are found in much larger arrays (typically 10–15 repeats) (Grove *et al.*, 2008). As a consequence of the repeat architecture, tandem repeat domains adopt non-globular, extended structures that present large, highly specific surfaces for ligand binding (Grove *et al.*, 2008). For several classes of repeat proteins, including ANK, TPR, and LRR, analyses of the amino acid variability at

different positions within a single repeat, revealed that the residues that comprise the ligand binding site, are significantly more variable than other positions on the protein surface, showing these domains are adaptable (Grove *et al.*, 2008). Proteins with large curvature and small twist, such as LRRs form concave surfaces that can wrap around a globular protein domain (Grove *et al.*, 2008). Repeat proteins with large twist and small curvature, such as HEAT repeats which have surfaces with grooves that can bind to an extended peptide surface (Grove *et al.*, 2008).

Tandem repeat domains are the most abundant eukaryotic-like domains in environmental intracellular bacteria but relatively absent in pathogenic Chlamydiae **(Figure 1 & Table 1)** (Schmitz-Esser *et al.*, 2010, Domman *et al.*, 2014, Gomez-Valero *et al.*, 2019). Here we will focus on three tandem repeat domains: ANK, LRR and TPR (including Sel-1) **(Figure 1 & Table 1)**. ANK is the most predominant eukaryotic-like repeat in *Legionella* spp., it is also highly promiscuous binding many domains including F-box, Fic and LRR domains (Burstein *et al.*, 2016, Gomez-Valero *et al.*, 2019). ANK is also abundant in environmental Chlamydiae but notably absent in pathogenic Chlamydiae analysed (**Figure 1 & Table 1**). ANK domains are commonly found in the secreted effectors of intracellular replicating bacteria (Voth, 2011). In eukaryotes ANK domains are highly enriched and mediate protein-protein interactions in a wide range of protein families such as those in cell cycle regulation, vesicular trafficking and inflammatory response (Jernigan & Bordenstein, 2014). The degeneracy of the ANK repeat allows for the specificity of individual molecular interactions, whilst the variability in the number of individual repeats in an ANK domain provides a platform for protein interactions (Jernigan & Bordenstein, 2014). LRR is the second most predominant eukaryotic-like repeat in *Legionella* spp., it is also highly promiscuous, binding many domains (Burstein *et al.*, 2016, Gomez-Valero *et al.*, 2019). Again, LRR repeats show the same profile, they are highly abundant in environmental Chlamydiae but are absent in pathogenic Chlamydiae (**Figure 1 & Table 1**). LRR are highly present in receptor-coreceptor complexes and innate immunity in plants, invertebrates and vertebrates. The third repeat, TPR repeat and closely related Sel-1 are abundant in *Legionella* spp. and environmental Chlamydiae, but only the Sel-1 repeat is absent from pathogenic Chlamydiae (**Figure 1 & Table 1**). In eukaryotes the TPR repeat functions in a variety of proteins involved in numerous eukaryotic cellular processes, such as gene regulation, mitosis, regulation of steroid receptor function, and protein import (Mittl & Schneider-Brachert, 2007, Cerveny *et al.*, 2013). In eukaryotes, Sel-1 proteins have a role in the regulation of cell division and the degradation of proteins from the endoplasmic reticulum (Mittl & Schneider-

Brachert, 2007). TPR-containing proteins in bacteria have been found to be directly involved in virulence-associated functions, such as the translocation of virulence factors into host cells, adhesion to host cells, and blocking of phagolysosomal maturation (Cerveny *et al.*, 2013). Sel-1 domains in bacteria were described in association with an effector that influences vacuolar trafficking (Newton *et al.*, 2007).

In addition to protein-protein interaction, these repeats themselves mimic the repetitive architecture of some host proteins. For example, in eukaryotes, LRR repeat proteins are abundant in the extracellular matrix, where they function in cell growth, adhesion, migration and bind with other extracellular matrix components. Some bacteria have been shown to use the LRR domain to mimic these LRR proteins in the extracellular matrix. This exemplified by internalin surface proteins of *Listeria monocytogenes* which have a crucial role in adhesion and invasion into the host cell (Marino *et al.*, 1999). The LRR is also commonly found in host immunity proteins such as NOD-like receptors and Toll-like receptors. The *L. pneumophila* protein Lpl1579 possesses the motif GAKALA in this variable region, which is similar to the sequence conservation pattern of LLRs in *Naegleria* NOD-like receptors (e.g., NLRC3) suggesting that it might mimic NLRC3 of *Naegleria* to subvert the host defence of this amoebae, but the role of Lpl1579 is not known yet (Doxey & McConkey, 2013).

Tandem repeat domain proteins have a role in localisation. RCC1 is a versatile domain which may perform many different functions, including guanine nucleotide exchange on small GTP-binding proteins, enzyme inhibition or interaction with proteins and lipids (Hadjebi *et al.*, 2008). RCC1 repeats are specifically found in *Legionella* spp. but seem to be absent from environmental and pathogenic Chlamydiae (**Figure 1 & Table 1**). In *Legionella* spp., the RCC1 repeat effectors localise to the pathogen vacuole or the host plasma membrane and target distinct components of the Ran GTPase cycle, including Ran modulators and the small GTPase itself (Swart *et al.*, 2020). Another example is the Sel-1 repeat which may have a role in localisation to Cis-Golgi (Voth *et al.*, 2019).

The importance of protein-protein interaction/ tandem repeat domains was recently demonstrated when four eukaryotic-like ANK repeat proteins from a sponge symbiont, were expressed in *Escherichia coli* were sufficient to modulate phagocytosis by amoebae (Nguyen *et al.*, 2014). This demonstrates that repeat containing proteins alone can be a predictor of host-pathogen interaction proteins. This theme is also emphasised in two of the core eight effectors common to all *Legionella* spp., AnkH (eukaryotic domain: ANK) and Lpg1356 (Eukaryotic domain: Sel-1) which contain protein-protein interaction/ tandem repeat

domains. Lpg1356 is not characterised, it is predicted to be secreted, the Sel-1 domain is present in many substrates of the type IVB system and it was shown to be implicated in host-pathogen interactions (Newton *et al.*, 2007, Lifshitz *et al.*, 2013). Additionally, an intracellular replication defect of a transposon mutant of Lpg1356 was observed in a human lung lymphobast cell line (Park *et al.*, 2020). Further evidence of the importance of repeat domains in host interacting proteins is seen in Chlamydiae as the expanded group of ubiquitination effectors all possess protein–protein interaction domains such as LRR and TPR (Domman *et al.*, 2014).

These different results provide strong evidence that a protein-protein interaction/ tandem repeat proteins system is fundamental for bacteria to replicate intracellularly, in particular in protists.

HOW ARE EUKARYOTIC-LIKE DOMAINS IN THE GENOME OF ENVIRONMENTAL INTRACELLULAR BACTERIA OBTAINED AND FIXED?

When the *L. pneumophila* genome was sequenced first, over 5 % of its genome seemed to be "of eukaryotic origin" (Cazalet *et al.*, 2004). This extraordinary percentage of the genome together with phylogenetic analyses of certain of these genes suggested that interdomain horizontal gene transfer rather than convergent evolution takes place and that the genes derived from eukaryotes are incorporated into the prokaryotic genome (Cazalet *et al.*, 2004). Horizontal gene transfer of exogenous genes in bacteria requires a double strand break to occur which is repaired *via* homologous recombination, homology facilitated illegitimate recombination, non-homologous end joining or alternative end-joining, as known to date (de Vries & Wackernagel, 2002, Chayot *et al.*, 2010, Popa *et al.*, 2011). With sufficient nucleotide sequence similarity between donor DNA and recipient genome integration can occur *via* homologous recombination leading to replacement of genes (homologous recombination) (**Figure 2**). One side homology is sufficient for horizontal gene transfer (homology facilitated illegitimate recombination) (**Figure 2**) (de Vries & Wackernagel, 2002). For integration of foreign DNA with low or no sequence identity three mechanisms have been identified. One relies on short specific nucleotide sequences recognised by cognate enzymes, such as transposases or integrases that cut and paste in these sites (mobile genetic element assisted horizontal gene transfer) (**Figure 2**) (Gillings, 2014, Arnold *et al.*, 2021). There are two non-homologous recombination mechanisms which require direct ligation of DNA ends after minor processing making them compatible for ligation. The first is alternative end joining which requires end processing by RecBCD, then regions of

microhomology are available to synapse and ligate (**Figure 2**) (Chayot *et al.*, 2010). Alternatively, non-homologous end joining requires Ku to protect the double strand break and LigD to repair (**Figure 2**) (Chayot *et al.*, 2010, Dupuy *et al.*, 2019). An *in silico* analysis showed that gene transfer events are more frequent when the recipient encodes non-homologous end joining proteins (Popa *et al.*, 2011). Preliminary analysis revealed that non-homologous end -joining proteins are present in some *Legionella* and environmental Chlamydiae, whereas they are notable absent from obligate intracellular Chlamydiae. As horizontal transfer happens most frequently between related species from the same domain, proposed barriers to interdomain horizontal gene transfer are homology, difference in GC content and physical transfer of DNA (Popa & Dagan, 2011).

The mechanism of how a gene or domain is transferred between a eukaryote and a prokaryote is unclear to the field, but it is likely to be associated with a mobile genetic element. Mobile genetic elements code for mobility related enzymatic functions, which enable breakage and joining of chromosomal DNA and are ubiquitous in both prokaryotes and eukaryotes (Schaack *et al.*, 2010, Arkhipova & Yushenova, 2019). Mobile genetic elements from both prokaryotes and eukaryotes support mobilisation of unrelated fragments of genomic DNA, if these fragments are appropriately positioned between *cis*-acting elements or otherwise placed within genomic DNA segments that will be subject to relocation (Arkhipova & Yushenova, 2019). Of note, intracellular bacteria that possess an enrichment of eukaryotic-like domains, are usually naturally competent and possess a type IVA secretion system (i.e. *Legionella* spp., *Coxiella* spp., *Rickettsia* spp., Chlamydiae (Collingro *et al.*, 2011, Jeffrey *et al.*, 2013, Christie *et al.*, 2017, Gomez-Valero *et al.*, 2019). Indeed, it is thought that type IVB evolved from ancient conjugation machines whose original functions were to disseminate mobile DNA elements within and between bacterial species (Christie *et al.*, 2017). Interkingdom movement of proteins and DNA *via* conjugation has been shown from *Agrobacterium* spp. to plant cells, however interkingdom DNA and protein movement in the opposite direction has not been demonstrated (Pitzschke & Hirt, 2010).

Acquired genes have to adapt within the host genome to be retained, as bacteria readily delete non-functional DNA. This is in particular true for obligate host-associated symbiotic bacteria which readily acquire mutations but are unable to remove them (Gomez-Valero *et al.*, 2007). Fixation is dependent on a new genes function and usefulness to the recipient under selectable conditions. Proposed barriers to eukaryotic gene expression and thereby fixation of a eukaryotic domain in a prokaryote include promoter incompatibility,

incompatible translation initiation signals, incompatible codon usage and the presence of spliceosomal introns in eukaryotic genes (Dunning Hotopp, 2011, Popa & Dagan, 2011, Li & Bock, 2019). Proof in principle experiments have recently been shown for transferred eukaryotic genes becoming functional in bacterial recipients *via* bacterial promoter capture (Li & Bock, 2019). They demonstrate that the eukaryotic promoter is deleted or gene amplification by tandem duplication takes place. In both cases a bacterial promoter is captured by microhomology mediated non-homologous recombination and is sufficient to achieve gene activation (Li & Bock, 2019). Similar molecular mechanisms were described for activation of plastid genes following transfer to the nuclear genome following endosymbiotic gene transfer (Stegemann & Bock, 2006). Single domain proteins composed of just one eukaryotic-like domain are found in *Legionella* spp., for example the SET domain, these may be an example of prokaryotic promoter capture and activation (**Figure 2 and 3**). Biotechnological expression of some eukaryotic proteins through bacterial expression systems has demonstrated that some translation initiation signals, and codon usage affect eukaryotic protein expression (Sahdev *et al.*, 2008). It can thereby be assumed that genes with certain translation initiation signals, and codon usage may be preferentially fixed in a prokaryote over others. Possessing an intron has been shown to not be an insurmountable obstacle to DNA mediated endosymbiotic gene transfer (Fuentes *et al.*, 2012). Using an experimental system that screened for functional endosymbiotic gene transfer of intron containing chloroplast genes to the nuclear genome of tobacco plants; gene activity was gained by using cryptic splice sites within chloroplast intron sequences that resulted in a contiguous reading frame (Fuentes *et al.*, 2012). Lastly, introns may not be a barrier to domain transfer, as there is a correlation between domain boundaries and exons ("cassette-exon") (Liu & Grigoriev, 2004). Therefore, it is possible that only the exon and thereby the functional domain is inherited and that underwent non-homologous recombination into the prokaryote genome. Hypothetically it is also possible that environmental bacteria incorporate spliced RNA from its host (Gomez-Valero & Buchrieser, 2019). Such a mechanism would require a reverse transcriptase present in the bacterial or viral genome and subsequent recombination using one of the aforementioned pathways incorporating it into the bacterial genome (**Figure 2**) (Gomez-Valero & Buchrieser, 2019).

HOW ARE EUKARYOTIC-LIKE DOMAINS INCORPORATED INTO DOMAIN ARCHITECTURES IN THE BACTERIAL GENOME?

Effector proteins are among the fastest evolving proteins in a number of pathogens (Nogueira *et al.*, 2012). Bacterial effectors, including eukaryotic-like domain proteins are very often multi-domain proteins, with each domain performing specific functions or contributing in a specific way to the function of their proteins (Forslund *et al.*, 2019). Domain architecture refers to domains within a protein and their order. Protein domains are regarded as portable units, adding or removing domains in different combinations to make new domain architectures enables new proteins to evolve which can fulfil complex tasks at high evolutionary speed (Bornberg-Bauer & Alba, 2013). Here we analysed the domain architectures of the eukaryotic- like domain containing proteins involved in chromatin modulation, ubiquitin modulation and repeat domains (**Figure 3**). Environmental bacteria exhibit a greater number of domain architectures than obligate intracellular Chlamydiae (**Figure 3**).

Following eukaryotic-like domain gene expression *via* prokaryotic promoter capture and potentially intron removal, what are the potential mechanisms by which new eukaryotic-like fusion domain architectures are built? The following events could lead to new architectures: (1) the domain is duplicated and subject to protein evolution (2) loss of the stop codon between eukaryotic domain and prokaryotic domain rearrangement result in gene fusion (3) non-homologous or mobile genetic element recombination **(Figure 3)** (Stavrinides *et al.*, 2006, Engel *et al.*, 2011, Bornberg-Bauer & Alba, 2013, Li & Bock, 2019).

Gene duplication is a common mechanism used in protein evolution, so the domain or protein is not lost if a recombination is unviable (Nasvall *et al.*, 2012, Adler *et al.*, 2014). Duplicated genes may eventually be non-functional, sub-functionalise (the original function is distributed between the duplicated genes) or undergo neo-functionalisation (a new function evolves in one of the duplicated copies while the old function is maintained in another copy) (Nasvall *et al.*, 2012, Adler *et al.*, 2014). the eukaryotic-like SET domain in *Legionella* spp. as an example, we observe in some species that proteins containing SET are duplicated and appear to show domain divergence (**Figure 4**). Duplication of a domain means that it is more abundant in the genome, more abundant domains have a higher chance of combining with a different partner and being in more multi-domain proteins than less abundant ones (Vogel *et al.*, 2005). For example, *Legionella* spp. ANK repeats are highly abundant and in a number of proteins in the proteome. For example, 35 different protein architectures in 300 effectors have

been identified (Burstein *et al.*, 2016). Gene duplication in intracellular bacteria is more frequent due to low genomic GC content, which leads to an increase of homopolymers which are hotspots for recombination insertion or deletion events due to replication slippage (Gomez-Valero *et al.*, 2008, Moran *et al.*, 2009). Of note, *Legionella* spp. generally have a GC content of ~ 39 % and Chlamydiae generally have a GC content of 42 % (Gomez-Valero *et al.*, 2019, Kostlbacher *et al.*, 2021). Intriguingly, tandem repeat domains seem also to be prone to duplicate in some environments (Coil *et al.*, 2008, Cooley *et al.*, 2010) as it was reported that for example strains isolated in the clinic have a higher repeat copy number than those isolated from hot springs (Coil *et al.*, 2008).

As described above, some eukaryotic-like domains such as ANK are present in various multi-domain protein combinations (Burstein *et al.*, 2016). Changes to domain architectures are more common by the N- and C- termini than internally in the architecture (Bjorklund *et al.*, 2005, Pasek *et al.*, 2006). One mechanism of insertion or deletion of a domain occurs by introducing start codons or removing stop codons, which results in the fusion of two neighbouring domains that are transcribed and translated as a single unit resulting in a domain fusion (Weiner *et al.*, 2006, Bornberg-Bauer & Alba, 2013). Gene fusion is thought to be a major contributor to the evolution of multi-domain bacterial proteins (Pasek *et al.*, 2006). The presence of two versions of the SWIB domain in the *C. trachomatis* genome, one singleton and one that is fused to topoisomerase I, suggests the possible two step evolutionary scenario for the origin of such fusions, which includes transfer of the eukaryotic gene followed by recombination, domain duplication, then loss of the stop codon, followed by fusion with Chlamydial topoisomerase I (Stephens *et al.*, 1998, Wolf *et al.*, 2000).

Another mechanism that could lead to these eukaryotic-like domain architectures is a non-homologous recombination event referred to as “domain/exon shuffling” (Stavrinides *et al.*, 2006, Bornberg-Bauer & Alba, 2013). This can be mediated by non-homologous recombination, the recombination between short homologous sequences or non-homologous sequences or transposon mediated exon shuffling, recombination mediated by transposases (Stavrinides *et al.*, 2006). There are few studies on the mechanisms of how multi-domain proteins in any bacteria shuffle domains. The non-homologous recombinational process named “terminal re-assortment” in type III secretion system (T3SS) effectors of bacteria such as *Pseudomonas* spp. and *Xanthamonas* spp. have been described (Stavrinides *et al.*, 2006). In T3SS effector terminal re-assortment the promoter and 5′ portion of a T3SS effector, which encodes all the regulatory domains required for T3SS-dependent expression, secretion

and translocation known as an ORPHETs (Orphan end terminals), fuse to a new functional domains to provide novel functional combinations (Stavrinides *et al.*, 2006). The above non-homologous recombination likely involves mobile genetic elements such as insertion sequences, integrative conjugative elements, phage, and plasmids, as over 50 % of ORPHETS or chimeric T3SS effectors are associated with one (Stavrinides *et al.*, 2006). In general, the insertion of mobile genetic elements into distinct sites is performed by recombinases, including transposases and integrases which break the DNA, then this break is repaired through the use of the non-homologous end-joining pathway (Stavrinides *et al.*, 2006). Interestingly, transposons have also been described to recombine domains and make new proteins in eukaryotes. As an example, domains such as the SET domain have been captured by transposons to produce host-transposase fusion proteins, leading to the evolution of transcription factors (Cosby *et al.*, 2021).

New effectors are established by remodelling existing eukaryotic-like domain architectures by further iterative duplication and recombination events. One study that describes this is in *Bartonella* spp., which like *Chlamydiae* and *Legionella* are able to survive and replicate in amoebae (Saisongkorh *et al.*, 2010). A group of *Bartonella* effector proteins (Beps) share a bipartite secretion signal composed of a C-terminal Bep intracellular delivery (BID) domain and a positively charged tail (Schulein *et al.*, 2005, Engel *et al.*, 2011). Mostly the N terminal possesses a FIC domain, however protein architectures with tandem repeated tyrosine phosphorylation motifs or additional BID domains are also described (Engel *et al.*, 2011, Harms *et al.*, 2017). These effectors are described to evolve *via* gene duplication then recombination events (Engel *et al.*, 2011, Harms *et al.*, 2017). In environmental Chlamydiae, a large expansion of ubiquitin ligase associated proteins has been observed (Domman *et al.*, 2014). *Neochlamydia* expansion (NEX1) and *Protochlamydia* expansion (PEX1) have 138 members and 27 members, respectively in two sequenced *Neochlamydia* and *Protochlamydia* genomes (Domman et al., 2014). The architecture in general of these proteins is eukaryotic-like E3 ubiquitin ligase-associated domains paired with repeat domains (Domman *et al.*, 2014). These paralogs show at least 30 % amino acid sequence identity over 60 % of protein length and phylogenetic trees show species specific expansion events (Domman *et al.*, 2014). As more abundant domains have a higher chance of combining with a different partner and being in more multi-domain proteins than less abundant ones, this observation is perhaps an intermediate step in new eukaryotic-like domain containing architectures in the genome of these environmental Chlamydiae (Vogel *et al.*, 2005).

Intriguingly, domains that mediate protein-protein interactions (e.g. LRR and ANK), or are involved in signal transduction, especially in the ubiquitin system (e.g. RING related to the U-box) and in chromatin (e.g. PHD and SET) are commonly found in diverse protein architectures in eukaryotes (Basu *et al.*, 2008, Basu *et al.*, 2009, Zmasek & Godzik, 2012). Domains which show a tendency to occur in diverse domain architectures are considered mobile (or "promiscuous") (Basu *et al.*, 2008). Furthermore, LRR and ANK are among a class of protein domains with a statistically significant correlation between their borders and exon borders known as exon bordering domains (Liu *et al.*, 2005). This property allows these domains to be highly mobile, preferentially expanded in the genome and found to undergo exon shuffling and combine with other domains (Liu *et al.*, 2005). It is interesting to note that the same eukaryotic-like domains described above to be enriched in environmental intracellular bacteria, are abundant and in a variety of domain architectures in eukaryotes including amoebae (**Figure 1 & Table 1**) (Basu *et al.*, 2008, Basu *et al.*, 2009). The high abundance, mobility and potentially interkingdom mobility of some domains is also reflected in the high number of orthologous eukaryotic domains being found in the genome of environmental intracellular bacteria. For example, in the genus *Legionella* the different proteins containing SET domains are classified in at least 7 different orthologous groups (**Figure 4**). Some species such as *L. oakridgensis* Oakridge 10 have acquired three different SET domains (**Figure 4**). Each of the 7 orthologs exhibit different presence/absence/ duplication patterns, a protein distribution compatible with different independent acquisition events (**Figure 4**). The SET domain is either present by itself or with an ANK domain (**Figure 3 & Figure 4**). Possession of an ANK domain could be either due to multi-domain acquisition or a domain shuffling event described above. Duplication events of SET-ANK and SET alone have occurred in a few species, which may result in further protein evolution (**Figure 4**).

WHY DO ENVIRONMENTAL INTRACELLULAR BACTERIA HAVE A HIGH PERCENTAGE OF EUKARYOTIC DOMAIN/PROTEINS IN THEIR GENOME?

- **Intimate contact with the host**

Eukaryotic-like proteins are enriched in intracellular bacteria, which is an environment where they are in close proximity to eukaryotes and eukaryotic DNA. Bacteria such as the obligate intracellular α- proteobacterium *Rickettsia prowazekii*, the progenitor of the mitochondria, possess higher numbers of eukaryotic-like proteins (Merhej & Raoult, 2011). Eukaryotic-like proteins are enriched in particular in bacteria that are able to survive and replicate within

amoebae (de Felipe *et al.*, 2005, Lurie-Weinberger *et al.*, 2010, Schmitz-Esser *et al.*, 2010, Gomez-Valero *et al.*, 2011, Jernigan & Bordenstein, 2014). Bacteria are targets to predation by grazing protozoa in fresh water, sea water, manmade water networks and soil; this is considered one of the oldest prey-predator interactions in nature (Hahn & Hofle, 2001, Matz & Kjelleberg, 2005, Samba-Louaka *et al.*, 2019). Some prokaryotes such as *Legionella* spp. and environmental Chlamydiae are able to avoid digestion in some but not all protozoa and establish an intimate symbiotic contact with their eukaryotic hosts post-ingestion (Rowbotham, 1983, Brown & Barker, 1999, Amaro *et al.*, 2015). Bacteria which are able to evade protozoal mediated death show an increased environmental fitness, as they are not digested and are in a protected niche with a source of nutrients. Therefore, there is a selection pressure for factors that contribute to survival inside protozoa, including bacterial secretion systems and their effectors (*e.g.* ones harbouring eukaryotic-like domains/ proteins).

Legionella have a broad host range spanning multiple phyla, from Amoebozoa (amoebae) to Percolozoa (excavates) to Ciliophora (ciliated protozoa) and at least 20 species of amoebae and protozoa to mammalian macrophages have been experimentally described to support *Legionella* growth (Rowbotham, 1980, Fields, 1996, Boamah *et al.*, 2017). *Legionella* spp. are found to often co-occur with free-living amoebae and believed to be a common environmental niche (Samba-Louaka *et al.*, 2019). There is phylogenetic evidence of tight co-evolution between *Legionella* spp. and free-living amoebae (Gomez-Valero *et al.*, 2019). As an example, genome analyses identified 184 genes that are predicted to encode small GTPases, 71 of which are Rab GTPases with high similarity with proteins from protozoan organisms such as *Entamoeba* or *Tetrahymena*e, indicating acquisition of these genes from protits (Gomez-Valero et al., 2019). Another example is *L. longbeachae* that is found in a different environment than other *Legionella* spp., namely in moist soil and potting soil (O'Connor *et al.*, 2007). Interestingly, the profile of eukaryotic-like domains present in *L. longbeachae* genome is different and seem to reflect the soil environment as for example PPR domain containing proteins are identified (**Figure 1 & Table 1**) (Cazalet *et al.*, 2010, Gomez-Valero *et al.*, 2011). PPR motifs are a family of RNA binding proteins that is greatly expanded in plants, and several other plant symbionts have PPR encoding genes such as *Rhodobacter*, *Ralstonia*, *Simkania* and *Erwinia* (Gomez-Valero *et al.*, 2011). However, the source of PPR repeats may not be the plant as protozoa are present in the soil and PPR proteins have also been found in amoebae such as *A. castelllanii* which has 28 (Schallenberg-Rudinger *et al.*, 2013, Samba-Louaka *et al.*, 2019).

Chlamydiae are strictly intracellular and therefore intimately reliant on their hosts (Horn, 2008). Indeed, the pangenome of Chlamydiae reveals that the core genome comprises the genes required for intracellular life (Kostlbacher *et al.*, 2021). Environmental Chlamydiae have been isolated from a range of locations including fresh water, sea water, fresh water sediment and sea water sediment (Kostlbacher *et al.*, 2021). Contrasting the genomes of pathogenic Chlamydiae and environmental Chlamydiae provides insight into the influence of the lifestyle on eukaryotic-like domain enrichment. Environmental Chlamydiae have a high number of eukaryotic-like domains such as ANK and F-box domain proteins and in abundant numbers **(Figure 1 & Table 1)** (Collingro *et al.*, 2011). In comparison pathogenic Chlamydiae have no ANK domain proteins and no F-box containing proteins **(Figure 1 & Table 1)**. This suggests that lifestyle of environmental Chlamydiae selects for an abundance of ANK and F-box domain proteins. In contrast, some eukaryotic-like domains such as the SET and the ULP1 deubiquitinase domain are found in both environmental and pathogenic Chlamydiae, perhaps indicating that they are key to intracellular life in all environments (amoebae and macrophage) (**Figure 1 & Table 1)**.

Obligate intracellular bacteria of free-living amoebae are also found to be enriched in eukaryotic-like domains suggesting that eukaryotic-like proteins might be a molecular manifestation of the host/symbiont relationship as much as predator/prey (Schmitz-Esser *et al.*, 2010). These bacteria belong to the class of α- and β-proteobacteria and are found among species in *Bacteroidetes* and Chlamydiae phyla (Schmitz-Esser *et al.*, 2008). These bacteria have different intracellular lifestyles living either in the amoebae cytoplasm or enclosed in host-derived vacuoles suggesting different mechanisms of host interaction (Schmitz-Esser *et al.*, 2008). *Candidatus amoebophilus asiaticus* which is a symbiont of amoebae encodes 129 proteins (8 % of all CDSs) with domains predominantly found in eukaryotic proteins, including ANK repeats, TPR/Sel-1 repeats, LRR repeats, or F-box and U-box domains (**Figure 1 & Table 1**) (Schmitz-Esser *et al.*, 2010).

When bacteria are intracellular there is limited possibility of gene acquirement *via* horizontal gene transfer. However, in *Legionella* spp. and environmental Chlamydiae horizontal gene transfer appears to be frequent (Gomez-Valero *et al.*, 2019, Kostlbacher *et al.*, 2021). Proof in principle experiments have shown that protozoa are "hotspots" of horizontal gene transfer. Indeed, *E. coli* strains engulfed by the ciliate, *Tetrahymena pyriformis*, exhibited increased rates of conjugation (Matsuo *et al.*, 2010). Bacteria that are present in the same niche can share eukaryotic-like proteins between each other (Gomez-Valero L., 2013). An example of gene exchange between species is RalF, a protein that

possesses a Sec7 domain that is homologous in *Rickettsia* and *Legionella* spp. Phylogenetic analysis suggests one bacterium acquired it from its host and gave it to the other (Gomez-Valero L., 2013). Intimate contact of *Legionella* spp. and environmental Chlamydiae with amoebae, is further indicated as mimivirus, a virus which infects amoebae, also was found to donate eukaryotic-like proteins to *Legionella* spp. (Moreira & Brochier-Armanet, 2008, Lurie-Weinberger *et al.*, 2010, Watanabe *et al.*, 2018, Pillonel *et al.*, 2019).

- **Open host lifestyle**

A trend has been observed that the stronger the level of host dependency the smaller the bacterial genome (Toft & Andersson, 2010). Pathogenic Chlamydiae follow this trend possessing a small genome of ~ 1 Mb (**Table 1**). In contrast, *Legionella* spp. and environmental Chlamydiae have a larger genome (2-4 Mb), indicating these species do not have as strong a dependency on the host (**Table 1**) (Horn *et al.*, 2004). Furthermore, *Legionella* spp. have an open pangenome as the *Legionella* genus genome is highly diverse and dynamic (Gomez-Valero *et al.*, 2019). Similarly, despite being obligate intracellular bacteria, environmental Chlamydiae showed a significantly more open pangenome compared to the pathogenic Chlamydiae (Kostlbacher *et al.*, 2021). Counter to the trend of obligate intracellular bacteria having reduced metabolic capacity due to the metabolite rich intracellular environment, environmental Chlamydiae also have a more complete metabolic capabilities, although synthesis of amino acids, cofactors, vitamins and nucleotides are still truncated consistent with their obligate intracellular lifestyle (Toft & Andersson, 2010, Omsland *et al.*, 2014, Kostlbacher *et al.*, 2021). This further suggests that environmental Chlamydiae are less reliant on their hosts than pathogenic Chlamydiae are on their animal hosts.

When bacteria are intracellular there is also a general trend of reduction of population size (Toft & Andersson, 2010). Population size is constrained in intracellular bacteria by the number of hosts, the number of infected cells and the cellular space available for growth (Toft & Andersson, 2010). Reduced population size makes selection less efficient in countering deleterious mutations and transmission dynamics among hosts may introduce strong bottlenecks on the effective population size and reduce variability (Toft & Andersson, 2010). However, in *Legionella* spp. and environmental Chlamydiae population size is potentially less constrained as amoebae are omnipresent and more abundant than animal hosts (Bar-On *et al.*, 2018). *Legionella* spp. are facultatively intracellular and in addition possess transmissive properties such as flagellum so have the opportunity to readily move

between environments (Oliva *et al.*, 2018). All Chlamydiae share a characteristic biphasic developmental cycle alternating between two distinct morphological and physiological stages: the elementary body (EB) survives (but cannot replicate) outside eukaryotic host cells and infects new hosts, whereas the reticulate body (RB) replicates inside the host cell within a host derived vacuole (Herrera *et al.*, 2020). Even though Chlamydiae are obligate intracellular bacteria, they can undergo mixed mode transmission: inherited vertically by the host or horizontally through uptake of the EB (Herrera *et al.*, 2020). Horizontal EB in environmental Chlamydiae is a chance for a diverse range of hosts to be infected.

- **Open host range**

Fitness of intracellular bacteria depends on their ability to both, expand and maintain its host interaction gene repertoire. For environmental intracellular bacteria, the host protozon is likely to continually change, fitness will be determined by the breadth and diversity of protozoa (Boamah *et al.*, 2017). *Legionella* spp. possess an extensive eukaryotic-like domain/ protein pangenome of over 18,000 proteins that contain at least 137 different eukaryotic domains and over 200 different eukaryotic proteins have been unveiled (Gomez-Valero *et al.*, 2019). However, only a set of eight conserved core effectors was identified in the genus *Legionella* (Gomez-Valero *et al.*, 2019). A recent study looking at *L. pneumophila* growth in different three amoebae hosts (*A. castellanii*, *H. vermiformis* or *N. gruberi*) showed that indeed only 2 % of *L. pneumophila* effectors were necessary for all hosts, but 15 % were necessary for an individual host (Park *et al.*, 2020). 83 *%* of *L. pneumophila* effectors were not necessary or specialised to amoebae hosts tested, but could be important for other protozoa hosts not tested (Park et al., 2020). In line with this idea is a study where large genomic regions containing each multiple effector proteins were deleted from the *L. pneumophila* genome. The strain that lacked 31 % of the effector repertoire had nearly no defect replicating in mouse macrophages, but deletions of each effector containing region alone had a different effect among amoebal species (O'Connor *et al.*, 2011).

Each species of *Legionella* has a specific arsenal of molecular strategies to survive intracellularly. Indeed, 32 % of the genes found in the pangenome were strain specific (Cazalet *et al.*, 2010, Burstein *et al.*, 2016, Gomez-Valero *et al.*, 2019). Individual species can interact (die/survive/ replicate) with a number of hosts to specialised degrees depending on the genes they possess (Amaro *et al.*, 2015, Park *et al.*, 2020). The individuality of a single species to interact with different hosts was recently highlighted when 13 analysed *Legionella* spp. which lacked 6–15 % genes that have been found to be important for *L. pneumophila* specialised growth in *A. castellanii*, *H. vermiformis* or *N. gruberi* were also tested (Park *et*

al., 2020). Some of these *Legionella* spp. were unable to survive and replicate in any of these amoebae, however some were still able survive and replicate in some amoebae suggesting uncharacterised compensatory genes in these *Legionella* spp. (Park *et al.*, 2020). The effectors present in an individual species are varied and are likely a record of the host interaction encountered by a given species which we term "effector memory". Protozoa appear to have had a big impact on the *Legionella* genome as while 44 genes were important for growth in amoebae only 4 were important in the macrophage (Park *et al.*, 2020). Indeed, some effectors e.g. LegC4 have a beneficial role during infection of protozoan hosts, but have a detrimental effect when this bacterium enters the lungs of mammalian hosts suggesting some effectors promote host bias (Shames *et al.*, 2017, Ngwaga *et al.*, 2019, Park *et al.*, 2020).

The amoebal host range of different environmental Chlamydiae has not been extensively studied; but infection of *Acanthamoeba* spp., *Hartmannella* spp. and *Naegleria* spp. has been recorded (Horn, 2008). The intracellular growth of environmental Chlamydiae was tested in 11 different *Acanthamoeba* strains and demonstrate significant differences in host susceptibilities to infection (Coulon *et al.*, 2012). The NEX1 and PEX1, ubiquitin pathway effector protein expansion is suggested to provide lineage specific functions related to particular host interactions (Domman *et al.*, 2014).

Thus, host specific functions evolved through the selection pressure exerted by diverse ameobal hosts seem to be a general theme of environmental, intracellular, protozoa associated pathogens and symbionts.

CONCLUSION

Over the past decades host-pathogen research has uncovered molecular strategies bacteria use to subvert host functions to survive in a eukaryotic cell, and even revealed unknown aspects of eukaryotic cellular biology. It also revealed that molecular mimicry of eukaryotic proteins *via* eukaryotic-like domains and proteins acquired by interdomain horizontal gene transfer often in the form of protein secretion system effectors, is a key tactic deployed in bacterial-eukaryote interactions. Whilst the function of effectors harbouring eukaryotic-like domains has been and continues to be investigated, the evolution of these proteins has received less attention. Most interestingly, there is a correlation between the number of eukaryotic-like domains encoded in a bacterial genome and the bacterial lifestyle. Bacteria living in complex interactions with bacterial communities in biofilms and with grazing protozoa, such as *Legionella* spp. and environmental Chlamydiae, display an enrichment in eukaryotic-like

domains (**Figure 5**). Although this is a clear indication that these proteins are important for bacteria-protozoa interactions, *Legionella* spp. and environmental Chlamydiae also infect accidentally humans and cause disease, thereby using the same proteins to subvert host functions in human cells. This suggests that bacterial pathogens may emerge from the environment, "trained" through interactions with protozoa **(Figure 5**). Therefore, studying these organisms may allow us to increase our knowledge of bacterial-eukaryote interactions and how bacteria adapt to succeed in this arms race in many different hosts.

Legionella spp. are exciting to study as these bacteria are intracellular pathogens of protozoa but also human macrophages. The Chlamydiae phylum shows two divergent host-bacteria interaction strategies within one group: a host-restricted strategy employed by animal and human pathogenic Chlamydiae and a more open-host strategy employed by environmental Chlamydiae. Comparing eukaryotic-like domains in the pathogenic Chlamydiae versus environmental Chlamydiae allows us to unveil what mimicry is core to eukaryotic intracellular life, and what is useful for an open-host strategy. There are strong themes in these eukaryotic-like domains found in *Legionella* spp. and environmental Chlamydiae genomes, which are chromatin modulation, ubiquitin pathway exploitation and tandem repeat domain proteins useful in protein-protein interactions. This suggests that these domains are particularly important and interfering with these pathways is crucial for survival and replication in a eukaryotic cell (**Figure 1 & Table 1**)**.** Knowledge on which pathways a bacterium needs to subvert to survive and replicate in a eukaryotic cell may deepen our understanding of host-pathogen interaction and allow to find new treatments to fight intracellular pathogens. The extraordinary percentage of eukaryotic-like proteins in *Legionella* spp. and environmental Chlamydiae suggests that interdomain horizontal gene transfer rather than convergent evolution is occurring. However, we still do not know the mechanism by which horizontal gene transfer is taking place in protozoa and experimental evidence is required to uncover how these intracellular bacteria pickpocket useful domains for intracellular survival.

References in figure/table legends

(Li *et al.*, 2003)

(Blum *et al.*, 2021)

(Mistry *et al.*, 2021)

Funding

Work in the CB laboratory is financed by the Institut Pasteur, the Fondation pour la Recherche Médicale (FRM) grant N° EQU201903007847 and Agence nationale de la recherche grant n°ANR-10-LABX-62-IBEID. JM was supported by a postdoctoral fellowship Fondation pour la Recherche Médicale (FRM)

Conflict of Interest

The authors declare there are no conflicts of interests.

Acknowledgements

We thank Christophe Rusniok for his help for the bioinformatics analyses shown in Figure 1 and Table 1 and Suzanna Salcedo for her patience in waiting for this manuscript that was prepared during three consecutive Covid-lockdowns

References

Adler M, Anjum M, Berg OG, Andersson DI & Sandegren L (2014) High fitness costs and instability of gene duplications reduce rates of evolution of new genes by duplication-divergence mechanisms. *Mol Biol Evol* **31**: 1526-1535.

Amaro F, Wang W, Gilbert JA, Anderson OR & Shuman HA (2015) Diverse protist grazers select for virulence-related traits in Legionella. *ISME J* **9**: 1607-1618.

Aravind L, Abhiman S & Iyer LM (2011) Natural history of the eukaryotic chromatin protein methylation system. *Prog Mol Biol Transl Sci* **101**: 105-176.

Arkhipova IR & Yushenova IA (2019) Giant Transposons in Eukaryotes: Is Bigger Better? *Genome Biol Evol* **11**: 906-918.

Arnold BJ, Huang IT & Hanage WP (2021) Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol*.

Bar-On YM, Phillips R & Milo R (2018) The biomass distribution on Earth. *Proc Natl Acad Sci U S A* **115**: 6506-6511.

Basu MK, Poliakov E & Rogozin IB (2009) Domain mobility in proteins: functional and evolutionary implications. *Brief Bioinform* **10**: 205-216.

Basu MK, Carmel L, Rogozin IB & Koonin EV (2008) Evolution of protein domain promiscuity in eukaryotes. *Genome Res* **18**: 449-461.

Bertelli C & Greub G (2012) Lateral gene exchanges shape the genomes of amoeba-resisting microorganisms. *Front Cell Infect Microbiol* **2**: 110.

Bjorklund AK, Ekman D, Light S, Frey-Skott J & Elofsson A (2005) Domain rearrangements in protein evolution. *J Mol Biol* **353**: 911-923.

Blum M, Chang HY, Chuguransky S*, et al.* (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* **49**: D344-D354.

Boamah DK, Zhou G, Ensminger AW & O'Connor TJ (2017) From Many Hosts, One Accidental Pathogen: The Diverse Protozoan Hosts of Legionella. *Front Cell Infect Microbiol* **7**: 477.

Bornberg-Bauer E & Alba MM (2013) Dynamics and adaptive benefits of modular protein evolution. *Curr Opin Struct Biol* **23**: 459-466.
Bracha R, Nuchamowitz Y & Mirelman D (2003) Transcriptional silencing of an amoebapore gene in Entamoeba histolytica: molecular analysis and effect on pathogenicity. *Eukaryot Cell* **2**: 295-305.
Brown MR & Barker J (1999) Unexplored reservoirs of pathogenic bacteria: protozoa and biofilms. *Trends Microbiol* **7**: 46-50.
Burstein D, Zusman T, Degtyar E, Viner R, Segal G & Pupko T (2009) Genome-scale identification of Legionella pneumophila effectors using a machine learning approach. *PLoS Pathog* **5**: e1000508.
Burstein D, Amaro F, Zusman T, Lifshitz Z, Cohen O, Gilbert JA, Pupko T, Shuman HA & Segal G (2016) Genomic analysis of 38 Legionella species identifies large and diverse effector repertoires. *Nat Genet* **48**: 167-175.
Cazalet C, Rusniok C, Bruggemann H*, et al.* (2004) Evidence in the Legionella pneumophila genome for exploitation of host cell functions and high genome plasticity. *Nat Genet* **36**: 1165-1173.
Cazalet C, Rusniok C, Bruggemann H*, et al.* (2004) Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. *Nat Genet* **36**: 1165-1173.
Cazalet C, Gomez-Valero L, Rusniok C*, et al.* (2010) Analysis of the Legionella longbeachae genome and transcriptome uncovers unique strategies to cause Legionnaires' disease. *PLoS Genet* **6**: e1000851.
Cerveny L, Straskova A, Dankova V, Hartlova A, Ceckova M, Staud F & Stulik J (2013) Tetratricopeptide repeat motifs in the world of bacterial pathogens: role in virulence mechanisms. *Infect Immun* **81**: 629-635.
Chayot R, Montagne B, Mazel D & Ricchetti M (2010) An end-joining repair mechanism in Escherichia coli. *Proc Natl Acad Sci U S A* **107**: 2141-2146.
Christie PJ, Gomez Valero L & Buchrieser C (2017) Biological Diversity and Evolution of Type IV Secretion Systems. *Curr Top Microbiol Immunol* **413**: 1-30.
Coil DA, Vandersmissen L, Ginevra C, Jarraud S, Lammertyn E & Anne J (2008) Intragenic tandem repeat variation between Legionella pneumophila strains. *BMC Microbiol* **8**: 218.
Collingro A, Kostlbacher S & Horn M (2020) Chlamydiae in the Environment. *Trends Microbiol* **28**: 877-888.
Collingro A, Tischler P, Weinmaier T*, et al.* (2011) Unity in variety--the pan-genome of the Chlamydiae. *Mol Biol Evol* **28**: 3253-3270.
Collins T, Stone JR & Williams AJ (2001) All in the family: the BTB/POZ, KRAB, and SCAN domains. *Mol Cell Biol* **21**: 3609-3615.
Cooley MB, Carychao D, Nguyen K, Whitehand L & Mandrell R (2010) Effects of environmental stress on stability of tandem repeats in Escherichia coli O157:H7. *Appl Environ Microbiol* **76**: 3398-3400.
Cosby RL, Judd J, Zhang R, Zhong A, Garry N, Pritham EJ & Feschotte C (2021) Recurrent evolution of vertebrate transcription factors by transposase capture. *Science* **371**.
Coulon C, Eterpi M, Greub G, Collignon A, McDonnell G & Thomas V (2012) Amoebal host range, host-free survival and disinfection susceptibility of environmental Chlamydiae as compared to Chlamydia trachomatis. *FEMS Immunol Med Microbiol* **64**: 364-373.
de Barsy M, Herrgott L, Martin V, Pillonel T, Viollier PH & Greub G (2019) Identification of new DNA-associated proteins from Waddlia chondrophila. *Sci Rep* **9**: 4885.
de Felipe KS, Pampou S, Jovanovic OS, Pericone CD, Ye SF, Kalachikov S & Shuman HA (2005) Evidence for acquisition of Legionella type IV secretion substrates via interdomain horizontal gene transfer. *J Bacteriol* **187**: 7716-7726.

de Vries J & Wackernagel W (2002) Integration of foreign DNA during natural transformation of Acinetobacter sp. by homology-facilitated illegitimate recombination. *Proc Natl Acad Sci U S A* **99**: 2094-2099.
Domman D, Collingro A, Lagkouvardos I, Gehre L, Weinmaier T, Rattei T, Subtil A & Horn M (2014) Massive expansion of Ubiquitination-related gene families within the Chlamydiae. *Mol Biol Evol* **31**: 2890-2904.
Doxey AC & McConkey BJ (2013) Prediction of molecular mimicry candidates in human pathogenic bacteria. *Virulence* **4**: 453-466.
Dunning Hotopp JC (2011) Horizontal gene transfer between bacteria and animals. *Trends Genet* **27**: 157-163.
Dupuy P, Sauviac L & Bruand C (2019) Stress-inducible NHEJ in bacteria: function in DNA repair and acquisition of heterologous DNA. *Nucleic Acids Res* **47**: 1335-1349.
Dutnall RN, Tafrov ST, Sternglanz R & Ramakrishnan V (1998) Structure of the histone acetyltransferase Hat1: a paradigm for the GCN5-related N-acetyltransferase superfamily. *Cell* **94**: 427-438.
Elde NC & Malik HS (2009) The evolutionary conundrum of pathogen mimicry. *Nat Rev Microbiol* **7**: 787-797.
Engel P, Salzburger W, Liesch M*, et al.* (2011) Parallel evolution of a type IV secretion system in radiating lineages of the host-restricted bacterial pathogen Bartonella. *PLoS Genet* **7**: e1001296.
Escoll P, Rolando M, Gomez-Valero L & Buchrieser C (2013) From amoeba to macrophages: exploring the molecular mechanisms of Legionella pneumophila infection in both hosts. *Curr Top Microbiol Immunol* **376**: 1-34.
Farbrother P, Wagner C, Na J, Tunggal B, Morio T, Urushihara H, Tanaka Y, Schleicher M, Steinert M & Eichinger L (2006) Dictyostelium transcriptional host cell response upon infection with Legionella. *Cell Microbiol* **8**: 438-456.
Fields BS (1996) The molecular ecology of *Legionellae*. *Trends Microbiol* **4**: 286-290.
Fischer A, Harrison KS, Ramirez Y*, et al.* (2017) Chlamydia trachomatis-containing vacuole serves as deubiquitination platform to stabilize Mcl-1 and to interfere with host defense. *Elife* **6**.
Foda BM & Singh U (2015) Dimethylated H3K27 Is a Repressive Epigenetic Histone Mark in the Protist Entamoeba histolytica and Is Significantly Enriched in Genes Silenced via the RNAi Pathway. *J Biol Chem* **290**: 21114-21130.
Forslund SK, Kaduk M & Sonnhammer ELL (2019) Evolution of Protein Domain Architectures. *Methods Mol Biol* **1910**: 469-504.
Fuentes I, Karcher D & Bock R (2012) Experimental reconstruction of the functional transfer of intron-containing plastid genes to the nucleus. *Curr Biol* **22**: 763-771.
Geng F, Wenzel S & Tansey WP (2012) Ubiquitin and proteasomes in transcription. *Annu Rev Biochem* **81**: 177-201.
Gillings MR (2014) Integrons: past, present, and future. *Microbiol Mol Biol Rev* **78**: 257-277.
Gomez-Valero L & Buchrieser C (2019) Intracellular parasitism, the driving force of evolution of Legionella pneumophila and the genus Legionella. *Genes Immun* **20**: 394-402.
Gomez-Valero L, Silva FJ, Simon JC & Latorre A (2007) Genome reduction of the aphid endosymbiont Buchnera aphidicola in a recent evolutionary time scale. *Gene* **389**: 87-95.
Gomez-Valero L, Rusniok C, Cazalet C & Buchrieser C (2011) Comparative and functional genomics of legionella identified eukaryotic like proteins as key players in host-pathogen interactions. *Front Microbiol* **2**: 208.

Gomez-Valero L, Latorre A, Gil R, Gadau J, Feldhaar H & Silva FJ (2008) Patterns and rates of nucleotide substitution, insertion and deletion in the endosymbiont of ants Blochmannia floridanus. *Mol Ecol* **17**: 4382-4392.

Gomez-Valero L, Rusniok C, Carson D*, et al.* (2019) More than 18,000 effectors in the Legionella genus genome provide multiple, independent combinations for replication in human cells. *Proc Natl Acad Sci U S A* **116**: 2265-2273.

Gomez-Valero L. NBM, Gribaldo S., Buchrieser C. (2013) *Interdomain Horizontal Gene Transfer Shaped the Genomes of Legionella pneumophila and Legionella longbeachae.* Springer, New York, NY. .

Green ER & Mecsas J (2016) Bacterial Secretion Systems: An Overview. *Microbiol Spectr* **4**.

Grove TZ, Cortajarena AL & Regan L (2008) Ligand binding by repeat proteins: natural and designed. *Curr Opin Struct Biol* **18**: 507-515.

Hadjebi O, Casas-Terradellas E, Garcia-Gonzalo FR & Rosa JL (2008) The RCC1 superfamily: from genes, to function, to disease. *Biochim Biophys Acta* **1783**: 1467-1479.

Hahn MW & Hofle MG (2001) Grazing of protozoa and its effect on populations of aquatic bacteria. *FEMS Microbiol Ecol* **35**: 113-121.

Harms A, Segers FH, Quebatte M*, et al.* (2017) Evolutionary Dynamics of Pathoadaptation Revealed by Three Independent Acquisitions of the VirB/D4 Type IV Secretion System in Bartonella. *Genome Biol Evol* **9**: 761-776.

Herrera P, Schuster L, Wentrup C*, et al.* (2020) Molecular causes of an evolutionary shift along the parasitism-mutualism continuum in a bacterial symbiont. *Proc Natl Acad Sci U S A* **117**: 21658-21666.

Horn M (2008) Chlamydiae as symbionts in eukaryotes. *Annu Rev Microbiol* **62**: 113-131.

Horn M, Collingro A, Schmitz-Esser S*, et al.* (2004) Illuminating the evolutionary history of chlamydiae. *Science* **304**: 728-730.

Hu H & Sun SC (2016) Ubiquitin signaling in immune responses. *Cell Res* **26**: 457-483.

Huguenin M, Bracha R, Chookajorn T & Mirelman D (2010) Epigenetic transcriptional gene silencing in Entamoeba histolytica: insight into histone and chromatin modifications. *Parasitology* **137**: 619-627.

Jeffrey BM, Suchland RJ, Eriksen SG, Sandoz KM & Rockey DD (2013) Genomic and phenotypic characterization of in vitro-generated Chlamydia trachomatis recombinants. *BMC Microbiol* **13**: 142.

Jehl MA, Arnold R & Rattei T (2011) Effective--a database of predicted secreted bacterial proteins. *Nucleic Acids Res* **39**: D591-595.

Jernigan KK & Bordenstein SR (2014) Ankyrin domains across the Tree of Life. *PeerJ* **2**: e264.

Kallin E & Zhang Y (2004) *Chromatin Remodeling*. Elsevier

Kitao T, Nagai H & Kubori T (2020) Divergence of *Legionella* Effectors Reversing Conventional and Unconventional Ubiquitination. *Front Cell Infect Microbiol* **10**: 448.

Kostlbacher S, Collingro A, Halter T, Schulz F, Jungbluth SP & Horn M (2021) Pangenomics reveals alternative environmental lifestyles among chlamydiae. *Nat Commun* **12**: 4021.

Kubori T, Hyakutake A & Nagai H (2008) *Legionella* translocates an E3 ubiquitin ligase that has multiple U-boxes with distinct functions. *Mol Microbiol* **67**: 1307-1319.

Kubori T, Shinzawa N, Kanuka H & Nagai H (2010) *Legionella* metaeffector exploits host proteasome to temporally regulate cognate effector. *PLoS Pathog* **6**: e1001216.

Lamoth F & Greub G (2010) Amoebal pathogens as emerging causal agents of pneumonia. *FEMS Microbiol Rev* **34**: 260-280.

Le Negrate G, Krieg A, Faustin B, Loeffler M, Godzik A, Krajewski S & Reed JC (2008) ChlaDub1 of Chlamydia trachomatis suppresses NF-kappaB activation and inhibits IkappaBalpha ubiquitination and degradation. *Cell Microbiol* **10**: 1879-1892.

Li L, Stoeckert CJ, Jr. & Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**: 2178-2189.

Li P, Vassiliadis D, Ong SY, Bennett-Wood V, Sugimoto C, Yamagishi J, Hartland EL & Pasricha S (2020) Legionella pneumophila Infection Rewires the Acanthamoeba castellanii Transcriptome, Highlighting a Class of Sirtuin Genes. *Front Cell Infect Microbiol* **10**: 428.

Li Z & Bock R (2019) Rapid functional activation of a horizontally transferred eukaryotic gene in a bacterial genome in the absence of selection. *Nucleic Acids Res* **47**: 6351-6359.

Lifshitz Z, Burstein D, Peeri M, Zusman T, Schwartz K, Shuman HA, Pupko T & Segal G (2013) Computational modeling and experimental validation of the Legionella and Coxiella virulence-related type-IVB secretion signal. *Proc Natl Acad Sci U S A* **110**: E707-715.

Liu M & Grigoriev A (2004) Protein domains correlate strongly with exons in multiple eukaryotic genomes--evidence of exon shuffling? *Trends Genet* **20**: 399-403.

Liu M, Walch H, Wu S & Grigoriev A (2005) Significant expansion of exon-bordering protein domains during animal proteome evolution. *Nucleic Acids Res* **33**: 95-105.

Lomma M, Dervins-Ravault D, Rolando M*, et al.* (2010) The Legionella pneumophila F-box protein Lpp2082 (AnkB) modulates ubiquitination of the host protein parvin B and promotes intracellular replication. *Cell Microbiol* **12**: 1272-1291.

Luger K, Rechsteiner TJ, Flaus AJ, Waye MM & Richmond TJ (1997) Characterization of nucleosome core particles containing histone proteins made in bacteria. *J Mol Biol* **272**: 301-311.

Lurie-Weinberger MN, Gomez-Valero L, Merault N, Glockner G, Buchrieser C & Gophna U (2010) The origins of eukaryotic-like proteins in *Legionella pneumophila*. *Int J Med Microbiol* **300**: 470-481.

Marino M, Braun L, Cossart P & Ghosh P (1999) Structure of the lnlB leucine-rich repeats, a domain that triggers host cell invasion by the bacterial pathogen L. monocytogenes. *Mol Cell* **4**: 1063-1072.

Matsuo J, Oguri S, Nakamura S*, et al.* (2010) Ciliates rapidly enhance the frequency of conjugation between Escherichia coli strains through bacterial accumulation in vesicles. *Res Microbiol* **161**: 711-719.

Matz C & Kjelleberg S (2005) Off the hook--how bacteria survive protozoan grazing. *Trends Microbiol* **13**: 302-307.

McDade JE, Shepard CC, Fraser DW, Tsai TR, Redus MA & Dowdle WR (1977) Legionnaires' disease: isolation of a bacterium and demonstration of its role in other respiratory disease. *N Engl J Med* **297**: 1197-1203.

Merhej V & Raoult D (2011) Rickettsial evolution in the light of comparative genomics. *Biol Rev Camb Philos Soc* **86**: 379-405.

Mistry J, Chuguransky S, Williams L*, et al.* (2021) Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**: D412-D419.

Mittl PR & Schneider-Brachert W (2007) Sel1-like repeat proteins in signal transduction. *Cell Signal* **19**: 20-31.

Mondino S, Schmidt S & Buchrieser C (2020) Molecular Mimicry: a Paradigm of Host-Microbe Coevolution Illustrated by Legionella. *mBio* **11**.

Mondino S, Schmidt S, Rolando M, Escoll P, Gomez-Valero L & Buchrieser C (2020) Legionnaires' Disease: State of the Art Knowledge of Pathogenesis Mechanisms of Legionella. *Annual review of pathology* **15**: 439-466.

Moran NA, McLaughlin HJ & Sorek R (2009) The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* **323**: 379-382.
Moreira D & Brochier-Armanet C (2008) Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. *BMC Evol Biol* **8**: 12.
Murata M, Azuma Y, Miura K, Rahman MA, Matsutani M, Aoyama M, Suzuki H, Sugi K & Shirai M (2007) Chlamydial SET domain protein functions as a histone methyltransferase. *Microbiology* **153**: 585-592.
Nasvall J, Sun L, Roth JR & Andersson DI (2012) Real-time evolution of new genes by innovation, amplification, and divergence. *Science* **338**: 384-387.
Newton HJ, Sansom FM, Dao J, McAlister AD, Sloan J, Cianciotto NP & Hartland EL (2007) Sel1 repeat protein LpnE is a Legionella pneumophila virulence determinant that influences vacuolar trafficking. *Infect Immun* **75**: 5575-5585.
Nguyen MTHD, Liu M & Thomas T (2014) Ankyrin-repeat proteins from sponge symbionts modulate amoebal phagocytosis. *Mol Ecol* **23**: 1635-1645.
Ngwaga T, Hydock AJ, Ganesan S & Shames SR (2019) Potentiation of Cytokine-Mediated Restriction of Legionella Intracellular Replication by a Dot/Icm-Translocated Effector. *J Bacteriol* **201**.
Nogueira T, Touchon M & Rocha EP (2012) Rapid evolution of the sequences and gene repertoires of secreted proteins in bacteria. *PLoS One* **7**: e49403.
O'Connor BA, Carman J, Eckert K, Tucker G, Givney R & Cameron S (2007) Does using potting mix make you sick? Results from a Legionella longbeachae case-control study in South Australia. *Epidemiol Infect* **135**: 34-39.
O'Connor TJ, Adepoju Y, Boyd D & Isberg RR (2011) Minimization of the *Legionella pneumophila* genome reveals chromosomal regions involved in host range expansion. *Proc Natl Acad Sci U S A* **108**: 14733-14740.
Oliva G, Sahr T & Buchrieser C (2018) The Life Cycle of L. pneumophila: Cellular Differentiation Is Linked to Virulence and Metabolism. *Front Cell Infect Microbiol* **8**: 3.
Omsland A, Sixt BS, Horn M & Hackstadt T (2014) Chlamydial metabolism revisited: interspecies metabolic variability and developmental stage-specific physiologic activities. *FEMS Microbiol Rev* **38**: 779-801.
Ong SY, Schuelein R, Wibawa RR, Thomas DW, Handoko Y, Freytag S, Bahlo M, Simpson KJ & Hartland EL (2021) Genome-wide genetic screen identifies host ubiquitination as important for L. pneumophila Dot/Icm effector translocation. *Cell Microbiol* e13368.
Park JM, Ghosh S & O'Connor TJ (2020) Combinatorial selection in amoebal hosts drives the evolution of the human pathogen Legionella pneumophila. *Nat Microbiol* **5**: 599-609.
Pasek S, Risler JL & Brezellec P (2006) Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics* **22**: 1418-1423.
Pennini ME, Perrinet S, Dautry-Varsat A & Subtil A (2010) Histone methylation by NUE, a novel nuclear effector of the intracellular pathogen Chlamydia trachomatis. *PLoS Pathog* **6**: e1000995.
Perez-Torrado R, Yamada D & Defossez PA (2006) Born to bind: the BTB protein-protein interaction domain. *Bioessays* **28**: 1194-1202.
Pergolizzi B, Bozzaro S & Bracco E (2019) Dictyostelium as model for studying ubiquitination and deubiquitination. *Int J Dev Biol* **63**: 529-539.
Pillonel T, Bertelli C, Aeby S, de Barsy M, Jacquier N, Kebbi-Beghdadi C, Mueller L, Vouga M & Greub G (2019) Sequencing the Obligate Intracellular Rhabdochlamydia helvetica within Its Tick Host Ixodes ricinus to Investigate Their Symbiotic Relationship. *Genome Biol Evol* **11**: 1334-1344.
Pitzschke A & Hirt H (2010) New insights into an old story: Agrobacterium-induced tumour formation in plants by plant transformation. *EMBO J* **29**: 1021-1032.

Pohl C & Dikic I (2019) Cellular quality control by the ubiquitin-proteasome system and autophagy. *Science* **366**: 818-822.
Popa O & Dagan T (2011) Trends and barriers to lateral gene transfer in prokaryotes. *Curr Opin Microbiol* **14**: 615-623.
Popa O, Hazkani-Covo E, Landan G, Martin W & Dagan T (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res* **21**: 599-609.
Price CT & Kwaik YA (2010) Exploitation of Host Polyubiquitination Machinery through Molecular Mimicry by Eukaryotic-Like Bacterial F-Box Effectors. *Front Microbiol* **1**: 122.
Price CT, Al-Quadan T, Santic M, Jones SC & Abu Kwaik Y (2010) Exploitation of conserved eukaryotic host cell farnesylation machinery by an F-box effector of *Legionella pneumophila*. *J Exp Med* **207**: 1713-1726.
Price CT, Al-Quadan T, Santic M, Rosenshine I & Abu Kwaik Y (2011) Host Proteasomal Degradation Generates Amino Acids Essential for Intracellular Bacterial Growth. *Science*.
Price CTD & Abu Kwaik Y (2021) Evolution and Adaptation of *Legionella pneumophila* to Manipulate the Ubiquitination Machinery of Its Amoebae and Mammalian Hosts. *Biomolecules* **11**.
Rahman MM & McFadden G (2011) Modulation of NF-kappaB signalling by microbial pathogens. *Nat Rev Microbiol* **9**: 291-306.
Rolando M & Buchrieser C (2012) Post-translational modifications of host proteins by Legionella pneumophila: a sophisticated survival strategy. *Future Microbiol* **7**: 369-381.
Rolando M, Gomez-Valero L & Buchrieser C (2015) Bacterial remodelling of the host epigenome: functional role and evolution of effectors methylating host histones. *Cell Microbiol* **17**: 1098-1107.
Rolando M, Sanulli S, Rusniok C, Gomez-Valero L, Bertholet C, Sahr T, Margueron R & Buchrieser C (2013) Legionella pneumophila effector RomA uniquely modifies host chromatin to repress gene expression and promote intracellular bacterial replication. *Cell Host Microbe* **13**: 395-405.
Rowbotham TJ (1980) Preliminary report on the pathogenicity of Legionella pneumophila for freshwater and soil amoebae. *J Clin Pathol* **33**: 1179-1183.
Rowbotham TJ (1983) Isolation of Legionella pneumophila from clinical specimens via amoebae, and the interaction of those and other isolates with amoebae. *J Clin Pathol* **36**: 978-986.
Sahdev S, Khattar SK & Saini KS (2008) Production of active eukaryotic proteins through bacterial expression systems: a review of the existing biotechnology strategies. *Mol Cell Biochem* **307**: 249-264.
Saisongkorh W, Robert C, La Scola B, Raoult D & Rolain JM (2010) Evidence of transfer by conjugation of type IV secretion system genes between Bartonella species and Rhizobium radiobacter in amoeba. *PLoS One* **5**: e12666.
Samba-Louaka A, Delafont V, Rodier MH, Cateau E & Hechard Y (2019) Free-living amoebae and squatters in the wild: ecological and molecular features. *FEMS Microbiol Rev* **43**: 415-434.
Schaack S, Gilbert C & Feschotte C (2010) Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol Evol* **25**: 537-546.
Schallenberg-Rudinger M, Lenz H, Polsakiewicz M, Gott JM & Knoop V (2013) A survey of PPR proteins identifies DYW domains like those of land plant RNA editing factors in diverse eukaryotes. *RNA Biol* **10**: 1549-1556.

Schmitz-Esser S, Toenshoff ER, Haider S, Heinz E, Hoenninger VM, Wagner M & Horn M (2008) Diversity of bacterial endosymbionts of environmental acanthamoeba isolates. *Appl Environ Microbiol* **74**: 5822-5831.
Schmitz-Esser S, Tischler P, Arnold R, Montanaro J, Wagner M, Rattei T & Horn M (2010) The genome of the amoeba symbiont "Candidatus Amoebophilus asiaticus" reveals common mechanisms for host cell interaction among amoeba-associated bacteria. *J Bacteriol* **192**: 1045-1057.
Schulein R, Guye P, Rhomberg TA, Schmid MC, Schroder G, Vergunst AC, Carena I & Dehio C (2005) A bipartite signal mediates the transfer of type IV secretion substrates of Bartonella henselae into human cells. *Proc Natl Acad Sci U S A* **102**: 856-861.
Shames SR, Liu L, Havey JC, Schofield WB, Goodman AL & Roy CR (2017) Multiple Legionella pneumophila effector virulence phenotypes revealed through high-throughput analysis of targeted mutant libraries. *Proc Natl Acad Sci U S A* **114**: E10446-E10454.
Sheedlo MJ, Qiu J, Tan Y, Paul LN, Luo ZQ & Das C (2015) Structural basis of substrate recognition by a bacterial deubiquitinase important for dynamics of phagosome ubiquitination. *Proc Natl Acad Sci U S A* **112**: 15090-15095.
Stavrinides J, Ma W & Guttman DS (2006) Terminal reassortment drives the quantum evolution of type III effectors in bacterial pathogens. *PLoS Pathog* **2**: e104.
Stebbins CE & Galan JE (2001) Structural mimicry in bacterial virulence. *Nature* **412**: 701-705.
Stegemann S & Bock R (2006) Experimental reconstruction of functional gene transfer from the tobacco plastid genome to the nucleus. *Plant Cell* **18**: 2869-2878.
Stephens RS, Kalman S, Lammel C*, et al.* (1998) Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis. *Science* **282**: 754-759.
Stogios PJ, Downs GS, Jauhal JJ, Nandra SK & Prive GG (2005) Sequence and structural analysis of BTB domain proteins. *Genome Biol* **6**: R82.
Swart AL, Gomez-Valero L, Buchrieser C & Hilbi H (2020) Evolution and function of bacterial RCC1 repeat effectors. *Cell Microbiol* **22**: e13246.
Toft C & Andersson SGE (2010) Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat Rev Genet* **11**: 465-475.
Vogel C, Teichmann SA & Pereira-Leal J (2005) The relationship between domain duplication and recombination. *J Mol Biol* **346**: 355-365.
Voth DE (2011) ThANKs for the repeat: Intracellular pathogens exploit a common eukaryotic domain. *Cell Logist* **1**: 128-132.
Voth KA, Chung IYW, van Straaten K, Li L, Boniecki MT & Cygler M (2019) The structure of Legionella effector protein LpnE provides insights into its interaction with Oculocerebrorenal syndrome of Lowe (OCRL) protein. *FEBS J* **286**: 710-725.
Watanabe T, Yamazaki S, Maita C, Matushita M, Matsuo J, Okubo T & Yamaguchi H (2018) Lateral Gene Transfer Between Protozoa-Related Giant Viruses of Family Mimiviridae and Chlamydiae. *Evol Bioinform Online* **14**: 1176934318788337.
Weiner J, 3rd, Beaussart F & Bornberg-Bauer E (2006) Domain deletions and substitutions in the modular protein evolution. *FEBS J* **273**: 2037-2047.
Wolf YI, Kondrashov AS & Koonin EV (2000) Interkingdom gene fusions. *Genome Biol* **1**: RESEARCH0013.
Zheng N & Shabek N (2017) Ubiquitin Ligases: Structure, Function, and Regulation. *Annu Rev Biochem* **86**: 129-157.
Zhou Y & Zhu Y (2015) Diversity of bacterial manipulation of the host ubiquitin pathways. *Cell Microbiol* **17**: 26-34.
Zmasek CM & Godzik A (2012) This Deja vu feeling--analysis of multidomain protein evolution in eukaryotic genomes. *PLoS Comput Biol* **8**: e1002701.

Table 1 Distribution of eukaryotic protein domains among proteomes of *Legionella* spp. and Chlamydiae

	Legionellaceae										Environmental Chlamydiae							Obligate human pathogens Chlamydia							Obligate fish
Species	*L. adelaidensis*	*L. yabuuchiae*	*L. hackeliae*	*L. micdadei* Tatlock	*L. beliardensis*	*L. gresilensis*	*L. longbeacheae* NSW150	*L. santicrucis*	*L. quateirensis*	*L. pneumophila* strain Paris	*Parachlamydia acanthamoebae* UV-7	*Protochlamydia amoebophila* UWE25	*Candidatus Rhabdochlamydia* T3358	*Criblamydia sequanensis* CRIB-18	*Estrella lausannensis*	*Waddlia chondrophila* WSU 86-1044	*Simkania negevensis* Z	*Chlamydia caviae* GPIC	*Chlamydia felis* Fe/C-56	*Chlamydia muridarum* Nigg	*Chlamydia pneumoniae* TW-183	*Chlamydia trachomatis* D/UW-3/CX	*Chlamydia psittaci* 6BC	*Chlamydia suis* 5-27b	*Candidatus Similichlamydia laticola*
Genome Size (Mb)	2,37	2,65	3,57	3,31	3,5	3,87	4,15	4,89	4,2	3,66	3,07	2,44	1,83	2,97	2,82	2,13	2,63	1,17	1,17	1,07	1,23	1,04	1,18	1,1	0,78
SET (IPR001214)	1	2	1	2	0	0	0	0	2	1	1	0	1	3	3	1	2	1	1	1	1	1	1	1	0
DOT1 (IPR025789)	1	1	1	1	1	1	1	2	1	1	0	0	0	1	0	1	0	1	1	0	1	0	1	0	1
SWIB/MDM2 (IPR003121)	0	0	0	0	0	0	0	0	0	0	2	2	2	2	2	2	3	2	2	2	2	2	2	2	2
U-box (IPR003613)	1	1	1	1	2	0	1	2	0	1	3	0	2	1	0	0	0	0	0	0	0	0	0	0	0
F-box (IPR001810)	0	1	3	1	0	2	1	1	1	3	18	11	2	37	6	1	0	0	0	0	0	0	0	0	0
OTU deubiquitinase (IPR003323)	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
ULP1/Peptidase_C48_C (IPR003653)	1	0	1	0	0	0	1	1	0	0	0	0	1	1	0	0	2	1	1	2	0	2	1	3	0
BTB/POZ domain TYPE (IPR000210)	0	0	0	0	0	0	0	0	0	0	1	19	2	9	1	0	0	0	0	0	0	0	0	0	0
HEAT repeat (IPR000357)	0	0	0	2	6	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RCC1 (IPR009091)	0	0	1	1	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LRR (IPR001611)	2	1	1	2	1	3	1	2	1	0	1	48	4	2	9	0	1	0	0	0	0	0	0	0	0
WD40 repeat (IPR001680)	0	0	0	0	0	0	0	0	0	0	3	0	0	13	0	0	1	0	0	0	0	0	0	0	0
PPR (IPR002885)	0	1	0	0	0	0	3	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
MORN (IPR003409)	0	0	0	0	0	0	1	0	0	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Sel1-like repeat (IPR006597)	3	4	4	4	9	6	3	2	5	5	5	7	1	0	13	1	1	0	0	0	0	0	0	0	0
TPR (IPR019734)	3	4	5	4	4	2	5	5	4	4	4	13	4	5	1	3	7	2	1	1	3	1	2	1	1
ANK (IPR020683/IPR002110)	8	15	13	12	25	35	23	41	26	14	15	6	4	7	2	4	8	0	0	0	0	0	0	0	1

The *Legionella* proteins containing eukaryotic-like proteins or domains are those identified in Gomez-Valero *et al*. 2019. The Chlamydia proteins were identified by applying the method used in Gomez-Valero *et al.* 2019, Briefly, protein domains were detected running the Pfam method from Interpro with default parameters over the whole proteomes of the selected genomes (Blum *et al.* 2020).

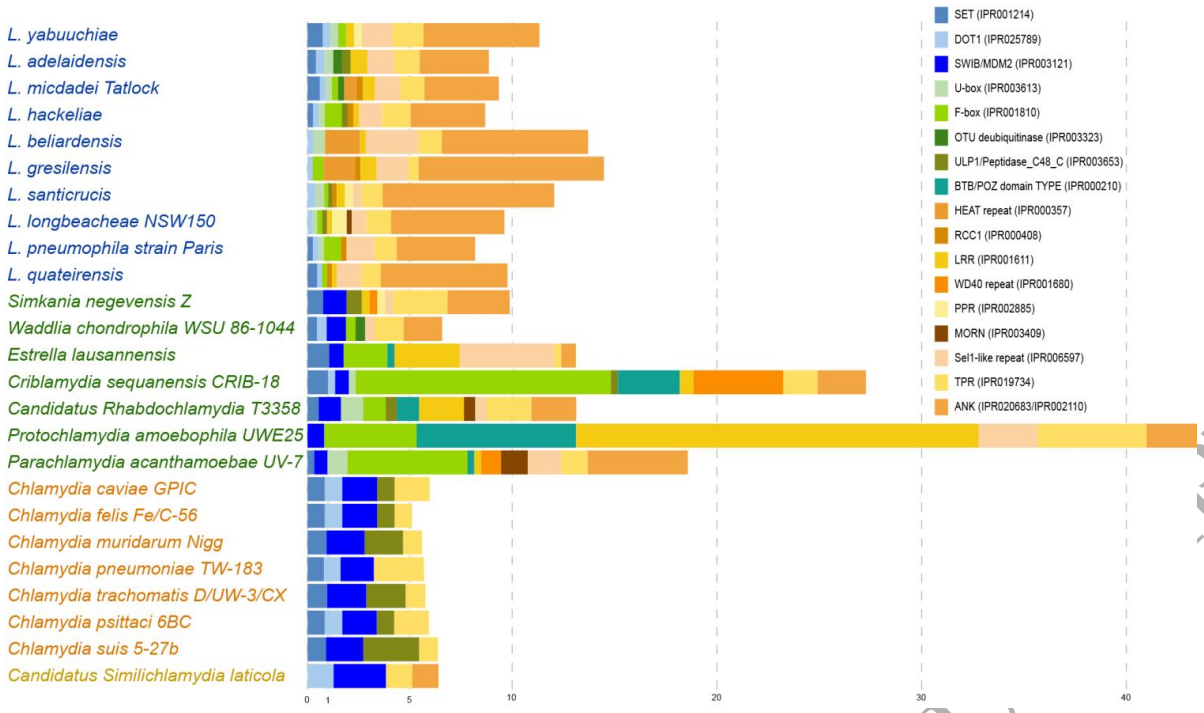

Figure 1: Selected examples of eukaryotic protein domains among proteomes of *Legionella* spp. and Chlamydiae in proportion to genome size. The length of the bars represents the number of selected eukaryotic domains in different *Legionella* and *Chlamydia* genomes per Mb. The dotted lines indicate the number of proteins per Mb (see related Table 1 for further details). Shades of blue, chromatin modulation domains (SET, DOT1, GNAT, SWIB/MDM2). Shades of green, ubiquitin modulation domains (U-box, F-box, OTU deubiquitinase, ULP1/Peptidase_C48_C). Blue-green; chromatin and ubiquitin modulation domains (BTB/POZ type). Shades of yellow and orange, tandem repeat domains (HEAT, RCC1, LRR, WD40 repeat, TPR, MORN, Sel-1, ANK).

Figure 2: Possible mechanisms of interkingdom horizontal gene/domain transfer. Transfer of protozoal DNA or RNA (blue) occurs and recombines *via* unknown mechanisms in bacteria (lilac). Insertion of foreign DNA may occur *via* the following examples: (1) Homology facilitated non-homologous recombination (one-sided or two-sided homology). (2) Mobile genetic element facilitated site specific non-homologous recombination (3) Alternative end-joining (4) Non homologous end joining.

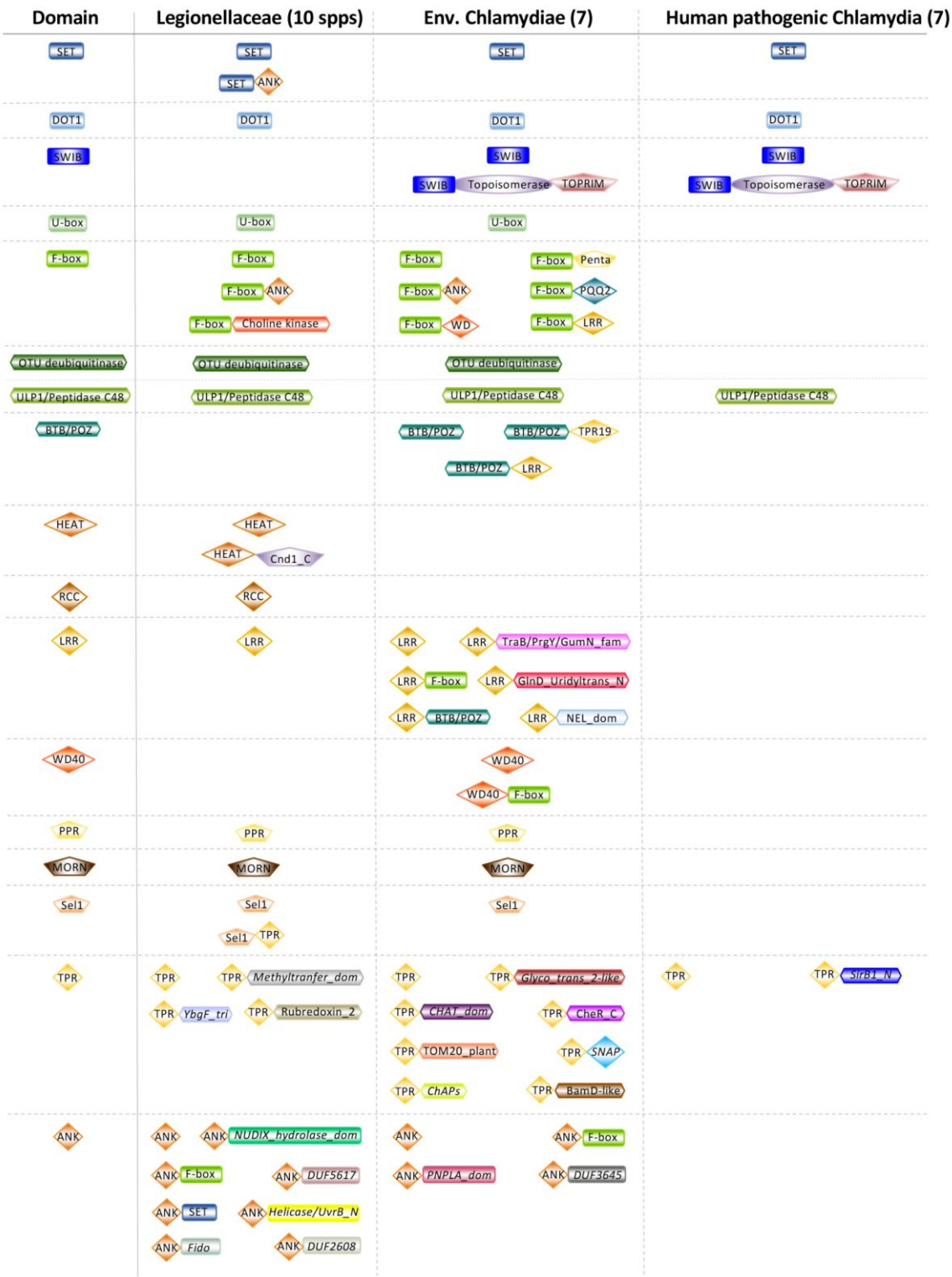

Figure 3: Eukaryotic domain architectures identified in selected *Legionella* spp. and Chlamydiae All domains were detected using the Pfam database **(**Mistry *et al.* 2020) in Interpro**.** The associated domains identified are: DNA topoisomerase (IPR013497); TOPRIM

(IPR006171); Choline_kinase (PF01633); Cnd1 (IPR032682); Rubredoxin_2 (IPR041166); YbgF, trimerisation domain (IPR032519); NUDIX_hydrolase_dom (IPR000086); DUF5617 (IPR041234); Fido (IPR003812); Helicase/UvrB,N-terminal (IPR006935); DUF2608 (IPR022565), PNPLA_dom (IPR002641); DUF3645 (IPR022105); ChAPs (IPR015374); BamD-like (IPR039565), Glyco_trans_2-like (IPR001173); SNAP (PF14938); CHAT_dom (IPR024983); CheR_C (IPR022642); TOM20_plant (PF06552); NEL_dom (PF14496); TraB/PrgY/GumN_fam domain (IPR002816); DUF294/GlnD_Uridyltrans_N (IPR005105) and SirB1_N (IPR032698).

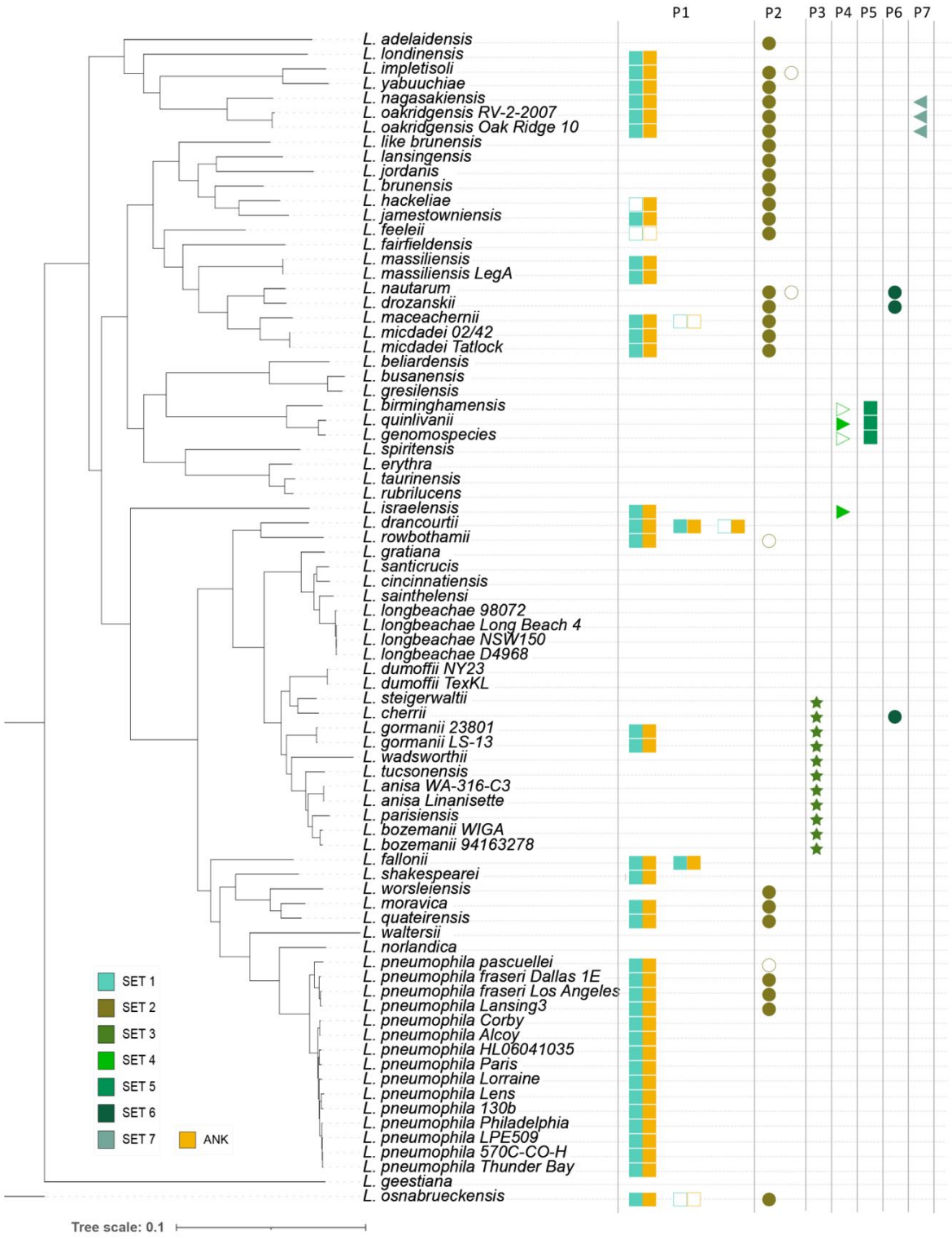

Figure 4 Seven different groups of orthologous proteins containing SET domains were identified in *Legionella* spp. The *Legionella* genus tree was published in Gomez-Valero *et al.* 2019. Due to space limits the branch connecting to the outgroup species *L. osnabrueckensis* has been artificially shortened. For the identification of orthologous

groups OrthoMCL software was used. Orthologous domains are represented with different symbols and/or colours and are named P1 – P7. Filled symbols represent SET domain (in green) and ankyrin (ANK) domain (in yellow) containing proteins. Empty symbols represent an orthologous protein that does not contain the SET and/or ANK domain. Duplicated or triplicated filled symbols represent paralogs of the same protein in the corresponding species.

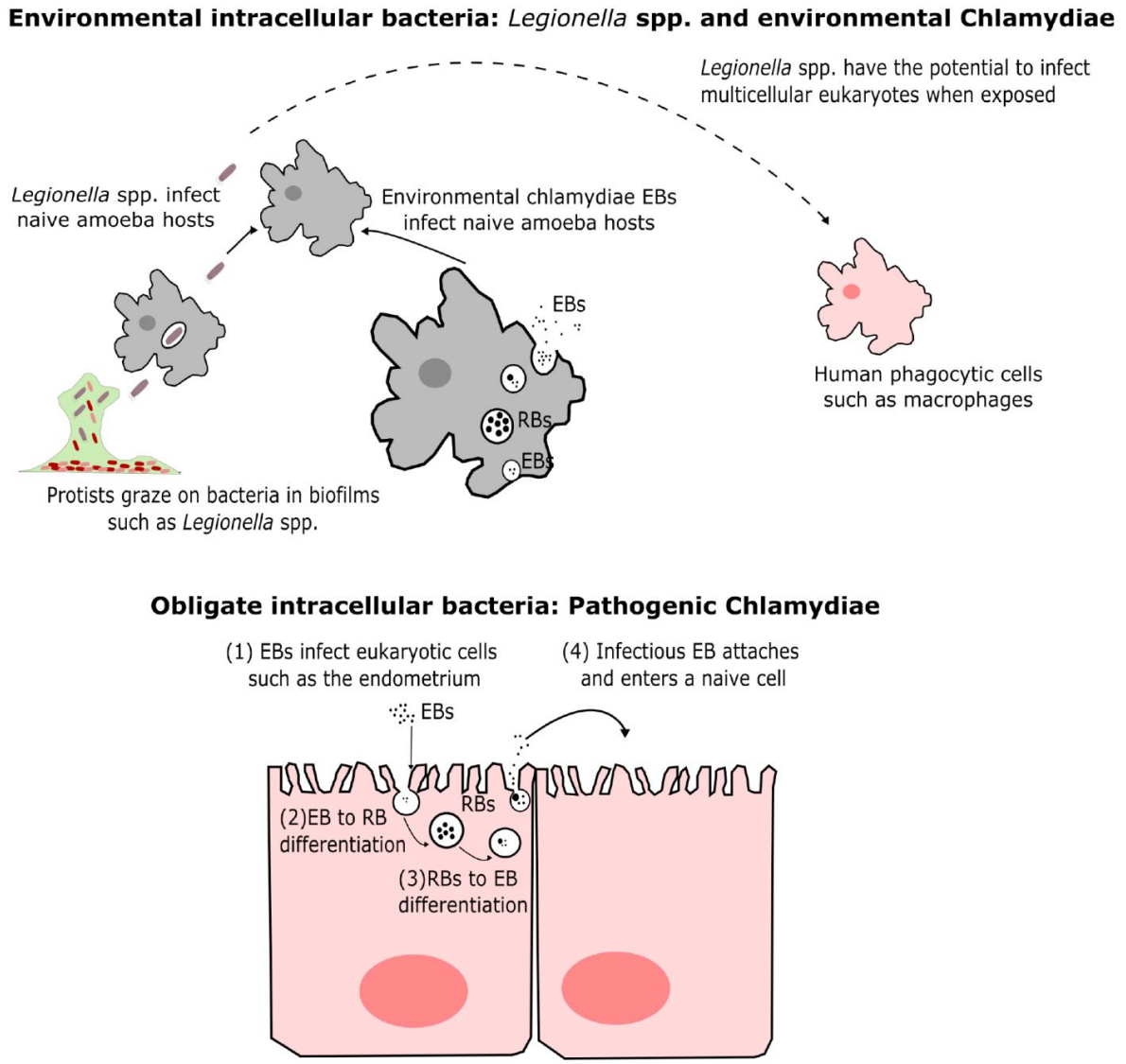

Figure 5: Environmental generalists *vs.* Environmental specialists. The infection cycles of *Legionella* sp. and Chlamydiae. A) Environmental bacteria are targets of predation by grazing protozoa. In response to predation many bacteria, such as *Legionella* spp. and Chlamydiae evolved strategies to survive and replicate within these predators. As environnemental bacteria encounter a large number of protozoan species fitness is determined by their ability to survive and replicate in many of them. **B)** Obligate intracellular bacteria are isolated and only survive and replicate in one host leading to host specialisation as represented here by pathogenic Chlamydiae. Abbreviations: EB, infectious extracellular elementary body, RB, intracellular replicative reticulate body.