Evolutionary Dissection of the Dot/Icm System Based on Comparative Genomics of 58 Legionella Species

Laura Gomez-Valero, Alvaro Chiner-Oms, Iñaki Comas, Carmen Buchrieser

Evolutionary Dissection of the Dot/Icm System Based on Comparative Genomics of 58 *Legionella* Species

Laura Gomez-Valero[1,2,*], Alvaro Chiner-Oms[3], Iñaki Comas[4,5], and Carmen Buchrieser[1,2,*]

[1]Institut Pasteur, Departement of Microbiology, Biologie des Bactéries Intracellulaires, Paris, France

[2]CNRS UMR3525, Paris, France

[3]Unidad Mixta "Infección y Salud Pública" FISABIO-CSISP/Universidad de Valencia, Instituto de Biología Integrativa de Sistemas, Spain

[4]CIBER en Epidemiología y Salud Pública, Valencia, Spain

[5]Instituto de Biomedicina de Valencia, IBV-CSIC, Valencia, Spain

*Corresponding authors: E-mails: lgomez@pasteur.fr; cbuch@pasteur.fr.

Accepted: August 21, 2019

Abstract

The Dot/Icm type IVB secretion system of *Legionella pneumophila* is essential for its pathogenesis by delivering >300 effector proteins into the host cell. However, their precise secretion mechanism and which components interact with the host cell is only partly understood. Here, we undertook evolutionary analyses of the Dot/Icm system of 58 *Legionella* species to identify those components that interact with the host and/or the substrates. We show that high recombination rates are acting on DotA, DotG, and IcmX, supporting exposure of these proteins to the host. Specific amino acids under positive selection on the periplasmic region of DotF, and the cytoplasmic domain of DotM, support a role of these regions in substrate binding. Diversifying selection acting on the signal peptide of DotC suggests its interaction with the host after cleavage. Positive selection acts on IcmR, IcmQ, and DotL revealing that these components are probably participating in effector recognition and/or translocation. Furthermore, our results predict the participation in host/effector interaction of DotV and IcmF. In contrast, DotB, DotO, most of the core subcomplex elements, and the chaperones IcmS-W show a high degree of conservation and not signs of recombination or positive selection suggesting that these proteins are under strong structural constraints and have an important role in maintaining the architecture/function of the system. Thus, our analyses of recombination and positive selection acting on the Dot/Icm secretion system predicted specific Dot/Icm components and regions implicated in host interaction and/or substrate recognition and translocation, which will guide further functional analyses.

Key words: *Legionella*, Dot/Icm system, positive-selection, negative-selection, diversifying-selection, evolution.

Introduction

Legionella are Gram-negative Gammaproteobacteria present in natural and man-made aquatic environments where they are replicating intracelullarly within a wide range of amoeba species belonging to the genera *Acanthamoeba*, *Hartmanella*, *Valkampfia*, and *Naegleria* as well as inside ciliates such as *Tetrahymena*, *Cyclidium*, or *Paramecium* (Newton et al. 2010; Boamah et al. 2017). However, certain *Legionella* species are also well known human pathogens due to their capacity to replicate in mammalian cells like macrophages when accidentally reaching the human lung. In susceptible individuals, infection with *Legionella* can lead to an acute pneumonia known as legionellosis or Legionnaires' disease (Newton et al. 2010).

The capacity of *Legionella* to infect eukaryotic cells relies on the Dot/Icm system, a type IV secretion machinery that is translocating proteins into the eukaryotic host cell. These effectors help this pathogen to subvert host functions to invade, replicate in, and escape from its hosts (Isberg et al. 2009; Hubber and Roy 2010). Based on sequence similarity, the Dot/Icm secretion machinery is classified as type IVB secretion system (T4BSSs) (Christie and Vogel 2000), in contrast to type IVA secretion systems (T4ASS) that are similar to the *Agrobacterium tumefaciens* VirB/VirD4 system. Recently, a classification based on the phylogenetic analysis of VirB4, the only ubiquitous protein in T4SSs, has been proposed (Guglielmini et al. 2013). Accordingly, T4SSs were classified in eight groups called MPF groups in reference to the

mating-pair formation proteins of the secretion systems. Based on this classification, T4ASSs belong to groups MPF_T (based on the T-DNA conjugation system of *A. tumefaciens* plasmid Ti) and MPF_F (based on plasmid F) whereas T4BSSs fall in the MPF_I group (based on the IncI plasmid R64). Both T4ASS and T4BSSs are bacterial multiprotein organelles specialized in the transfer of (nucleo)protein complexes across cell membranes (Christie et al. 2017).

The development of techniques to test experimentally whether a predicted Dot/Icm substrate is indeed secreted through this machinery (Zhu and Luo 2013) together with bioinformatics analyses has revealed that *Legionella pneumophila* is provisioned with a unprecedented number of over 330 Dot/Icm substrates (Burstein et al. 2009; Zhu et al. 2011; Lifshitz et al. 2013; Finsel and Hilbi 2015; Ensminger 2016; Escoll et al. 2016; Qiu and Luo 2017). Recent bioinformatics predictions of effectors in different *Legionella* species added many more proteins to this list (Burstein et al. 2016; Gomez-Valero et al. 2019). It is thought that this huge arsenal of effector proteins allows *Legionella* to adapt to its very large host spectrum of amoeba and ciliated protozoa. Interestingly, this effector repertoire is highly conserved among different *L. pneumophila* strains, but when comparing different *Legionella* species the scenario is very different, as most of the effectors lack orthologs in most or all *Legionella* species analyzed (Gomez Valero et al. 2011; Gomez-Valero et al. 2014, 2019; Burstein et al. 2016). Indeed, of the over 330 substrates described in the species *L. pneumophila*, only 10 are present in all sequenced *Legionella* species, pointing to a very small core set of effectors (Burstein et al. 2016; Gomez-Valero et al. 2019). Nevertheless, bioinformatics predictions suggested that all species carry a large number of effectors. Therefore, this surprisingly small number of core substrates indicates that a very different effector set is present in different *Legionella* species (Gomez-Valero et al. 2019) probably because each *Legionella* species/strain faces a particular range of protozoan hosts in the environment. In contrast to the effector diversity, the Dot/Icm secretion system is highly conserved at interspecies level (Burstein et al. 2016; Gomez-Valero et al. 2019). This raises an intriguing question: which components of the Dot/Icm secretion system allow detection and translocation of such a broad spectrum of different proteins?

Structural and functional analyses of several components of the Dot/Icm system in the past few years have helped to better understand how the Dot/Icm machinery may work, but a detailed knowledge is still missing (Raychaudhury et al. 2009; Ghosal et al. 2017; Kwak et al. 2017; Chetrit et al. 2018; Meir et al. 2018). These studies have defined two major subcomplexes: the core transmembrane subcomplex and the coupling protein subcomplex (fig. 1). The first one is composed of the proteins DotC, DotD, DotF (IcmG), DotG (IcmE), and DotH (IcmK) and crosses the inner and outer membranes of *L. pneumophila* (Vincent, Buscher, et al. 2006). In this complex, the DotH protein forms the outer membrane pore, localized properly thanks to the lipoproteins DotC and DotD, and receives energy from DotG (Sutherland et al. 2013). Additionally, DotF interacts with DotG and regulates energy transducing activity of DotG. The second complex recruits substrates and delivers them to the secretion channel. While only one protein, VirD4 mediates coupling between substrate recruitment and delivery to the secretion channel in type T4ASSs, a coupling complex is needed in T4BSSs. The coupling complex of the Dot/Icm system comprises the inner membrane AAA+ ATPase DotL (IcmO), which is a homolog of VirD4, two inner membrane/cytoplasmic components, DotM (IcmP) and DotN (IcmJ), a complex of two cytoplasmic chaperones, IcmS and IcmW (IcmSW) and the protein LvgA (Kwak et al. 2017). LvgA was originally described as a *Legionella* virulence factor (Edelstein et al. 2003), but later it has been shown that this protein is also part of the Dot/Icm secretion machinery (Vincent and Vogel 2006). Finally, to be functionally complete, the Dot/Icm secretion apparatus contains 16 other proteins in addition to the above-mentioned subcomplexes (table 1).

Given its function and localization, the Dot/Icm system is submitted to strong and competing evolutionary pressures: it needs to (i) avoid host defences, (ii) constantly adapt to new hosts/niches, and (iii) preserve interactions with other components of the Dot/Icm apparatus. Accordingly, some Dot/Icm elements have to remain unaltered to maintain the stability of the system and are thus expected to be highly conserved even among different *Legionella* species. Given that, most of the amino acid changes occurring within these proteins will be deleterious and thus removed by natural selection, the evolutionary pressure acting on these components/regions is negative/purifying selection. In contrast, other components/regions, in particular the ones directly exposed to the host system and/or the substrates, have to adapt constantly to different or new hosts/niches and are therefore evolving fast and show high variability. These components are usually under a positive selective pressure as the amino acid changes have a high probability to provide a selective advantage and thus to be fixed. Therefore, the study of the evolutionary pressures acting on the Dot/Icm system allows predicting which elements/regions are key for effector/host interaction and which remain unaltered, probably for maintaining the integrity of the system.

Such evolutionary forces can be better dissected among different species that diverged sufficiently long ago and live in different environments. Therefore, interspecies analysis provides a unique opportunity to identify elements under adaptive evolution. The genus *Legionella* is a privileged model for such analyses as the quasi entire genus has been sequenced. Although intracellular replication has been shown experimentally only for 35 of the 58 *Legionella* species included here

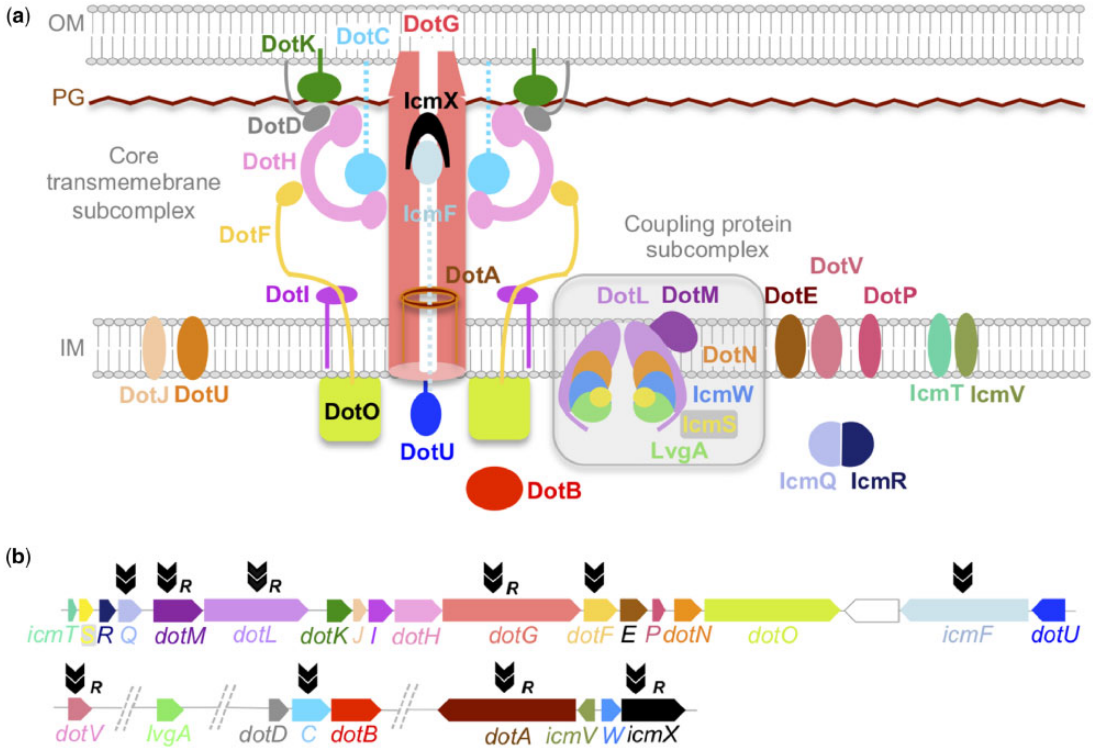

Fig. 1.—Schematic representation of the Dot/Icm secretion apparatus and the gene loci encoding the different Dot/Icm components. (*a*) The representation of the core transmembrane subcomplex is based mainly on the work of Ghosal et al. (2017). The representation of the coupling protein subcomplex is based on studies reported earlier (Kwak et al. 2017; Chetrit et al. 2018; Meir et al. 2018). DotC for which the structure is not known is drawn as circle. DotH and IcmX are represented in the shapes reported from densities seen in the subtomogram averages or difference maps (Ghosal et al. 2017). (*b*) Genes coding the different Dot/Icm components represented according to their size, position, and orientation in the *Legionella pneumophila* Paris genome, and colored according to the schematic presentation shown in (*a*). A double arrow above the gene indicates positive selection acting on it, either on specific sites and/or specific nodes of the phylogeny. If positive selection was detected only with aBSREL and only on one node of the phylogeny, it is indicated only when *P* values are <0.01. An "R" above a gene indicates that more than one sequence was affected by recombination. OM, outer membrane; IN, inner membrane; PG, peptidoglycan layer.

(summarized in Gomez-Valero et al. 2019) it can be assumed that all species replicate in eukaryotic hosts. This idea is also substantiated by a recent phylogenetic analysis of T4BSSs in Gram-negative bacteria that proposes that the acquisition of the T4BSS on the chromosome might be related to the alteration of the life style as intracellular bacterium. Furthermore, they suggest that the genes found only in the Dot/Icm systems of *Legionella* and related bacteria may encode components that are important for life as intracellular pathogens (Nagai and Kubori 2011). Thus, we used here the sequence of 80 *Legionella* strains belonging to 58 species (Gomez-Valero et al. 2019) and applied an evolutionary approach to identify the different functions of the Dot/Icm components.

Materials and Methods

The sequence of the 27 genes that constitute the Dot/Icm secretion system were extracted from the genomes of 80 *Legionella* strains belonging to 58 species, previously sequenced in our lab (Gomez-Valero et al. 2019) (supplementary table S1, Supplementary Material online). BLASTp of each Dot/Icm protein encoded by the *L. pneumophila* Paris genome against the orthologous proteins in each of the other *Legionella* species/strains was used to calculate the average amino acid identity of the Dot/Icm proteins (table 1) (Gomez-Valero et al. 2019). Each gene was translated and the corresponding proteins were aligned using the program PRANK v.170427 (Loytynoja 2014). PRANK uses a phylogenetic tree of the strain/species included in the alignment. Thus, a phylogenetic tree based on the core genome previously published (Gomez-Valero et al. 2019) of these strains was included in the study (supplementary fig. S1, Supplementary Material online). The resulting amino acid alignments were used as a guide for the alignment of the corresponding nucleotide sequences using the software DAMBE v.6.4.40 (Xia 2017). The obtained alignments were cleaned with Gblocks v. 0.91 b (Castresana 2000) with the less astringent conditions (allowing: smaller final blocks, gap positions within the final

Table 1
General Features of Each Dot/Icm Component in the 80 *Legionella* Strains Analyzed in This Study

Protein Name	Protein Label[a]	No. of Sequences	Amino Acid Identity (%)[b]	Gene Length[a]	Aligment Length[c]	Average Identical Nucleotides[b]	% Change[b]	Recombin[d]
IcmT	Lpp0507	80	78–100 (87%)	261	249	196.4	20.5	No
IcmS	Lpp0508	80	74–100 (91%)	345	333	263.6	20.4	No
IcmR	Lpp0509	16	48–99 (94%)[e]	363	357	337.0	6.6	No
IcmQ	Lpp0510	80	49–99 (70%)	576	504	364.7	27.1	No
IcmP /DotM	Lpp0511	80	64–99 (79%)	1,131	1,098	834.8	23.5	Yes (3)
IcmO /DotL	Lpp0512	80	77–100 (89%)	2,352	2,325	1,854.6	19.9	Yes (10)
IcmN /DotK	Lpp0513	80	50–99 (69%)	570	486	341.1	29.0	No
IcmM /DotJ	Lpp0514	80	33–99 (63%)	285	291	188.6	35.0	No
IcmL /DotI	Lpp0515	80	69–100 (90%)	639	633	504.7	19.7	No
IcmK /DotH	Lpp0516	80	65–99 (79%)	1,083	867	672.6	21.9	No
IcmE /DotG	Lpp0517	80	49–100 (71%)	3,147	2,706	1,904.1	27.8	Yes (43)
IcmG /DotF	Lpp0518	80	37–99 (57%)	810	504	362.1	27.4	No
IcmC /DotE	Lpp0519	80	57–100 (76%)	585	525	384.7	25.5	No
IcmD /DotP	Lpp0520	80	57–99 (78%)	399	315	242.0	22.5	No
IcmJ /DotN	Lpp0521	80	72–100 (87%)	627	603	482.8	19.5	No
IcmB /DotO	Lpp0522	80	80–100 (88%)	3,030	3,006	2,366.4	20.6	Yes (1)
IcmF	Lpp0524	80	34–100 (70%)	2,922	2,811	1,985.7	28.2	No
IcmH /DotU	Lpp0525	80	36–100 (71%)	786	747	537.1	27.5	No
DotV	Lpp0537	79	44–100 (69%)	543	462	332.1	28.1	Yes (6)
LvgA	Lpp0590	80	42–100 (69%)	627	528	380.5	27.3	Yes (1)
DotD	Lpp2728	80	69–100 (84%)	492	465	364.2	21.0	No
DotC	Lpp2729	80	69–100 (81%)	912	810	634.4	20.9	No
DotB	Lpp2730	80	86–100 (94%)	1,134	1,101	878.1	19.7	No
DotA	Lpp2740	80	47–99 (60%)	3,108	2,025	1,419.0	28.5	Yes (20)
IcmV	Lpp2741	80	49–99 (67%)	456	429	300.6	29.4	No
IcmW	Lpp2742	80	71–100 (89%)	456	453	361.0	19.8	Yes (1)
IcmX	Lpp2743	78	36–89 (51%)	1,419	711	464.6	34.1	Yes (4)

NOTE.—Amino acid (aa) identity represents the minimum, maximum, and average values (average between parenthesis) of aa identity between each Dot/Icm protein from *L. pneumophila* strain Paris against the corresponding orthologous in other species calculated by BLASTp. The parameters: alignment length, average of identical nucleotides, and the % of changes are calculated from the multiple alignment of each protein after cleaning with Gblocks and taking as a reference the genome of *L. pneumophila* Paris for calculating identity and % of change.

[a]*L. pneumophila* strain Paris.

[b]With respect to *L. pneumophila* Paris sequence.

[c]After cleaning with Gblocks.

[d]Number of recombinant sequences.

[e]Identity value based on the comparison of 16 strains (15 belonging to the same species *L. pneumophila* and only one to a different species *L. norrlandica*).

blocks, and less strict flanking positions). Alignment properties were calculated using the tool infoalign from the EMBOSS package v. 6.6.0.0 (Rice et al. 2000) and taking *L. pneumophila* Paris as reference genome. Additionally, Shannon entropy calculation was done for some selected protein alignments using the Protein variability server (Garcia-Boronat et al. 2008), to visualize variability along the protein sequence.

Intragenic recombination detection was carried out using the software 3seq v1.7 (Lam et al. 2018) according to recommendations by Martin et al. (2011). The program was also run on concatenated genes belonging to the same cluster. To confirm identified recombination events, we analyzed the congruence of each Dot/Icm protein alignment with the phylogeny obtained from all alignments and the phylogeny derived from the core genome alignment. The congruence was evaluated applying the Shimodaira–Hasegawa (SH) test (Shimodaira and Hasegawa 1999) implemented in TreePuzzle (Schmidt et al. 2002). Tree-Puzzle was also used to evaluate the phylogenetic information contained in each Dot/Icm protein alignment, using likelihood mapping (Strimmer and von Haeseler 1997) (supplementary table S2, Supplementary Material online).

For selection analyses, recombinant sequences from each alignment were removed, as these may interfere with the positive selection analysis. To ensure that differences among the compared sequences represented real fixation events along independent lineages, only one strain from each species was used in this analysis as artifacts may appear when comparing closely related strains (Kryazhimskiy and Plotkin 2008). Finally, for positive selection detection, we applied several methods implemented in the Hyphy package v. 2.3

(Pond et al. 2005): MEME (Murrell et al. 2012), FUBAR (Murrell et al. 2013), FEL (Pond et al. 2005), and aBSREL (Smith et al. 2015). FEL and FUBAR allow detecting sites under pervasive diversifying selection. Whereas both methods assume that selection pressure for each site is constant along the entire phylogeny, they differ in the way to calculate nonsynoymous (d*N*) and synonymous (d*S*) substitutions. FEL uses a maximum-likelihood approach whereas FUBAR uses a Bayesian approach. Moreover, FUBAR seems to have more power than FEL, in particular when positive selection is present but relatively weak (Murrell et al. 2013). Additionally, we applied MEME to test the hypothesis that individual sites have been subject to episodic diversifying selection. Finally, aBSREL was applied to identify lineages, which have experienced episodic purifying selection independently of the sites affected by this selective pressure. We choose the *P* value according to the recommendations in the HYPHY package; 0.1 for FEL and MEME. For aBSREL, we choose a *P* values <0.01. For FUBAR, we do not have *P* values but posterior probabilities.

To ensure that the positive selection signal, we detected was not due to misalignments, in particular, in the case of the most variable Dot/Icm components, we used an alternative alignment program, MUSCLE v.3.8, and rerun FUBAR/aBSREL to confirm the results. Using this control, the only case that was not corroborated was positive selection acting on several codons at the beginning of DotK when using PRANK alignments. Detailed visual inspection on this region revealed that in fact it could not be aligned with high accuracy. In consequence, we did not take this result into account. Signal IP v.4.1 (Petersen et al. 2011) was used to detect secretion signals. The MOTIF search tool of the Japanese GenomeNet service (Kanehisa 1997) was applied for detecting pentapetide repeats using the Pfam library (Finn et al. 2016). The program IBS v.1.0.3 (Liu et al. 2015) was used for schematic representations of protein domains and iTOL v.4.3 (Letunic and Bork 2016) for representing phylogenies together with partial alignments.

Results and Discussion

Recombination Is Important in the Evolution of the Dot/Icm System

The *Legionella* Dot/Icm system is highly conserved at interspecies level as previously reported (Burstein et al. 2016; Gomez-Valero et al. 2019) with an average amino acid identity of the Dot/Icm proteins ranging from 51% (IcmX) to 94% (DotB) (table 1). Due to this high homology at protein level, it was possible to align the corresponding genes with high accuracy based on protein alignments. Alignments were curated to avoid spurious orthologs codon detection while keeping most of the gene sequence (table 1).

To analyze the role of recombination and natural selection on shaping the interspecies genetic diversity of the Dot/Icm components, we followed a two-step strategy. First, recombination analyses were carried out using the nonparametric method 3seq to detect mosaic structures in the sequences. This analysis identified intragenic recombination in 9 of the 27 *dot/icm* genes with particular high recombination rates in *dotG*, *dotA*, and *icmX* as further detailed below (table 1). As a second approach, we undertook recombination detection on concatenated genes, which was confirming our results (supplementary table S3, Supplementary Material online). Furthermore, we undertook a phylogenetic incongruence test (SH), which compares the likelihood of each protein alignment versus the phylogenetic trees derived from them and the core phylogeny to confirm these results (supplementary table S4 and fig. S2, Supplementary Material online). Combined these methods disclosed that recombination events are important in the evolution of the Dot/Icm system.

We then aimed to discover the Dot/Icm components, and more specifically the codons in these proteins, that may have been subjected to positive selection. After removing the above-detected recombinant sequences to avoid false positive results, amino acids that have been subjected to pervasive positive selection were searched for with the methods FEL and FUBAR. Furthermore, the sites that have evolved under positive selection even when this selection has acted only on a proportion of the branches in the phylogenetic tree were analyzed by MEME. Finally, we applied the aBSREL method (Smith 1992) to detect branches in the phylogeny of *Legionella* that could have been under positive selection for specific Dot/Icm components regardless whether codons affected by this selective pressure had been detected. Among the 27 Dot/Icm proteins, specific sites under positive selection were identified in IcmQ, DotM, DotF, IcmF, DotV, and DotC. Moreover, four genes that evolved through diversifying selection on different nodes of the phylogeny were identified (*dotL*, *dotG*, *dotA*, and *icmX*). No significant signs of positive selection for the remaining 17 genes could be detected.

Analysis of Proteins of the Core Transmembrane Subcomplex

High Intragenic Recombination Rates Support Host Exposure of DotG and DotA

DotG (IcmE/Lpp0517) is an integral membrane protein that was proposed to form a central channel spanning inner and outer membranes (fig. 1) (Kubori et al. 2014). Recent in situ cryo-electron tomography confirmed that DotG couples the outer membrane core complex with the cytoplasmic complex by forming the cylinder domain that constitutes the central channel portion (Chetrit et al. 2018). When analyzing the DotG protein sequence at the genus level, we found that recombination plays a major role in the evolution of this protein as recombination events were identified in 43 of the 80 analyzed DotG sequences (table 1). Despite this result, the DotG protein phylogeny is compatible with the core

phylogeny (supplementary fig. S2 and table S5, Supplementary Material online) when the SH test is applied (*P* value 0.82). This result may be due to the fact that the DotG protein is a large protein (1,048 aa in *L. pneumophila*) and recombination events are affecting only the N-terminal and middle part of the protein but not the C-terminal region (supplementary fig. S3, Supplementary Material online). Consequently, the phylogenetic signal of the alignment is strong enough to recover a tree with a similar topology as the core phylogeny.

The C-terminus of DotG of *L. pneumophila* has been reported to be similar to the TrbI domain of VirB10 family proteins of T4ASSs (Nagai and Kubori 2011). Our sequence analysis shows that this TrbI domain is conserved in the 75 analyzed strains/species for which the complete DotG sequence was available. In contrast to the well-conserved C-terminal region, the middle region is highly variable and contains a variable number of 1–13 pentapeptide repeats (supplementary table S4, Supplementary Material online). For *L. pneumophila*, *Coxiella burnetti*, and *Rickettsiella*, it has been reported that DotG proteins are significantly larger than other homologs due to these repeats (Segal et al. 1998; Nagai and Kubori 2011). Here, we show that the DotG proteins of 58 *Legionella* species vary between 1,000 (cluster of *L. pneumpophila*) and >1,500 amino acids (cluster of *Legionella sainthelensi–Legionella longbeachae* and *Legionella santicrucis*) (supplementary fig. S4, Supplementary Material online). Interestingly, even between strains from the same species a considerable size variation can be present, as seen for the two *Legionella oakridiginesis* strains analyzed here, that have 1,011 and 1,161 amino acids, respectively. The *Helicobacter pylori* HP0527 (CagY) protein, a homolog of DotG, contains also a large number of repeat regions following the well conserved C-terminal part, that can be deleted or extended by intragenic recombination leading to variations between strains. This mechanism has been suggested to help the pathogen to avoid or modulate the host immune response as CagY decorates the bacterial surface (Rohde et al. 2003; Barrozo 2013). Although, in *Legionella*, there is to date no experimental evidence that the variable sequence regions of DotG are surface displayed, our results suggest that some segments are surface exposed, and that DotG function could be similar to its homolog CagY. In addition, although specific sites under positive selection were not detected, several branches on the evolutionary tree of DotG have been subjected to diversifying selection (aBSREL) (table 2). Such a faster evolution supports again an interaction of DotG with the host system and/or with the secreted substrates.

A second Dot/Icm protein showing a high recombination rate is DotA (Lpp2740), an integral cytoplasmic membrane protein (Roy and Isberg 1997), that is essential for the functioning of the T4SS (fig. 1) (Berger et al. 1994). The presence of recombination and frequent nonsynonymous mutations in the *dotA* gene has been reported for different *L. pneumophila* strains (Ko et al. 2003; Costa et al. 2010). Our present analysis of DotA shows also a high degree of interspecies variability (only 60% of average amino acid identity) and a high recombination rate with 20 different *Legionella* species being affected. A SH test confirmed this result as the phylogeny derived from DotA is not compatible with the phylogeny derived from the core genome (*P* value 0.007). The test shows incongruent evolutionary histories between *dotA* and the core genome, a result in agreement with recombination events affecting this protein (supplementary fig. S2, Supplementary Material online). Like for DotG, no specific site under diversifying selection was detected, but many branches in the tree are under positive selection (aBSREL) (table 2 and supplementary fig. S4, Supplementary Material online). It is highly probably that this is due to the fact that sequences affected by recombination had to be removed prior to the analyses, which reduced the power of detection of sites under positive selection for all methods used except for one (aBSREL). Thus, recombination and positive selection have played an important role in the evolution of DotA at the genus level.

DotA is an inner membrane protein but it is also secreted after cleavage of the 19 amino acids long leader peptide (Nagai and Roy 2001). Our results that show fast evolution of DotA further support the idea that DotA directly interacts with the host forcing DotA in the different *Legionella* species to constantly adapt to different hosts.

Positive Selection Analyses Suggest a Role of the Periplasmic Region of DotF in Substrate Recognition

DotF (IcmG/Lpp0518) is an inner membrane protein that is part of the core transmembrane complex of the Dot/Icm apparatus (fig. 1). Our analysis shows that, like DotA, this protein has high variability at interspecies level as the percentage of change is 27.4% (table 1). Moreover, the analysis of selective forces acting on DotF points to at least two residues under diversifying selection (table 2). These two residues reside in the periplasmic domain (fig. 2) adjacent to the transmembrane domain of DotF. This finding suggests a role of this periplasmic region in effector interaction as predicted by Sutherland et al. (2013). Indeed, it has been demonstrated previously that DotF interacts with several Dot/Icm effectors (Luo and Isberg 2004) and that this interaction takes place through the transmembrane and/or periplasmic domain of DotF (Sutherland et al. 2013; Kubori et al. 2014). Additionally, the recent electron cryotomography results suggested that the walls of the Dot/Icm channel crossing the bacterial membrane have openings. The opening between the beta and gamma rings is localized just below DotD and next to DotF, suggesting that DotF plays a role in translocating effectors initially secreted to the periplasm, to the secretion chamber (Ghosal et al. 2017). Interestingly, the C-terminal

Table 2
Results of Negative/Positive Selection Acting on Dot/Icm Genes at Interspecific Level

Gene Name	Gene Label*	No. of Sequences	Methods Used to Infer Selection					
			FEL			MEME (codon under positive selection)	FUBAR (codon under positive selection)	ABSREL (nodes under positive selection)
			No. of Codons Tested	No. of Codons Under Neg. Select.	Codon Under Pos. Select.			
icmT	*lpp0507*	58	83	76	0	0	0	0
icmS	*lpp0508*	58	111	109	0	0	0	0
icmR	*lpp0509*	2	—	—	—	—	—	—
icmQ	*lpp0510*	58	168	161	1 (0.0558)	1 (0.0756)	0	0
icmP /dotM	*lpp0511*	56	366	340	0	22 (0.0153) 362 (0.0594)	0	74 (0.03936)
icmO /dotL	*lpp0512*	56	777	753	0	0	0	2 (0.0168) 94 (0.03696)
icmN /dotK	*lpp0513*	58	162	150	0	0	0	0
icmM /dotJ	*lpp0514*	57	74	66	0	0	0	0
icmL /dotI	*lpp0515*	58	211	203	0	0	0	0
icmK /dotH	*lpp0516*	58	285	266	0	0	0	0
icmE /dotG	*lpp0517*	35	902	873	0	0	0	1 (0.00376) 2 (0.00459) 14 (0.01272)
icmG /dotF	*lpp0518*	58	168	159	124 (0.0152)	124 (0.0238) 127 (0.0348)	124 (0.9729)	0
icmC /dotE	*lpp0519*	58	175	171	0	0	0	0
icmD /dotP	*lpp0520*	58	105	101	0	0	0	6 (0.02037)
icmJ /dotN	*lpp0521*	58	201	194	0	0	0	0
icmB /dotO	*lpp0522*	58	1002	937	0	0	0	0
icmF	*lpp0524*	58	937	908	0	434 (0.0314)	0	4 (0.00270) 69 (0.00397) 95 (0.00504)
icmH /dotU	*lpp0525*	58	248	240	0	0	0	0
dotV	*lpp0537*	52	155	146	0	43 (0.0065)	0	0
lvgA	*lpp0590*	57	176	172	0	0	0	6 (0.03112)
dotD	*lpp2728*	58	155	146	0	0	0	0
dotC	*lpp2729*	58	270	251	2 (0.0004) 5 (0.0630) 6 (0.0017)	2 (0.00090) 5 (0.0840) 6 (0.0032) 250 (0.0678)	2 (0.9978) 5 (0.9314) 6 (0.9975)	4 (0.00039) 36 (0.02733)
dotB	*lpp2730*	58	367	360	0	0	0	15 (0.01046)
dotA	*lpp2740*	48	675	644	0	0	0	50 (0.00047) 34 (0.00173) 63 (0.01155) 1 (0.02190) 29 (0.02565)
icmV	*lpp2741*	58	143	127	0	0	0	0
icmW	*lpp2742*	57	150	143	0	0	0	0
icmX	*lpp2743*	53	237	229	0	0	0	17 (0.00005)

NOTE.—Number in parenthesis indicates either *P* value or posterior probability (FUBAR method) associated to the codon/branch inferred to be subjected to positive selection. Gray areas highlight those genes where positive selection was found.

domain of DotF that contacts this opening is similar to the C-terminal domain of PilP/GspC (Ghosal et al. 2017), which in type II and III secretion systems recruits effectors from the periplasm and delivers them to the translocation channel. DotF was therefore suggested to play a role in effector interaction or alternatively a role in stabilizing the apparatus or triggering conformational changes (Ghosal et al. 2017). Our current results showing diversifying selection acting on DotF support a role of DotF in effector interaction. Moreover, the fact that amino acids of DotF for which we detect positive selection are exposed in the periplasm supports the function of the periplasmic domain of DotF in effector binding. Indeed, from the 168 codons of DotF, 159 were under negative selection pressure and only two are under positive selection (table 2). Thus, the diversifying evolution detected in the periplasmic domain of DotF is probably reflecting adaptation to changes in effector repertories in the different *Legionella*

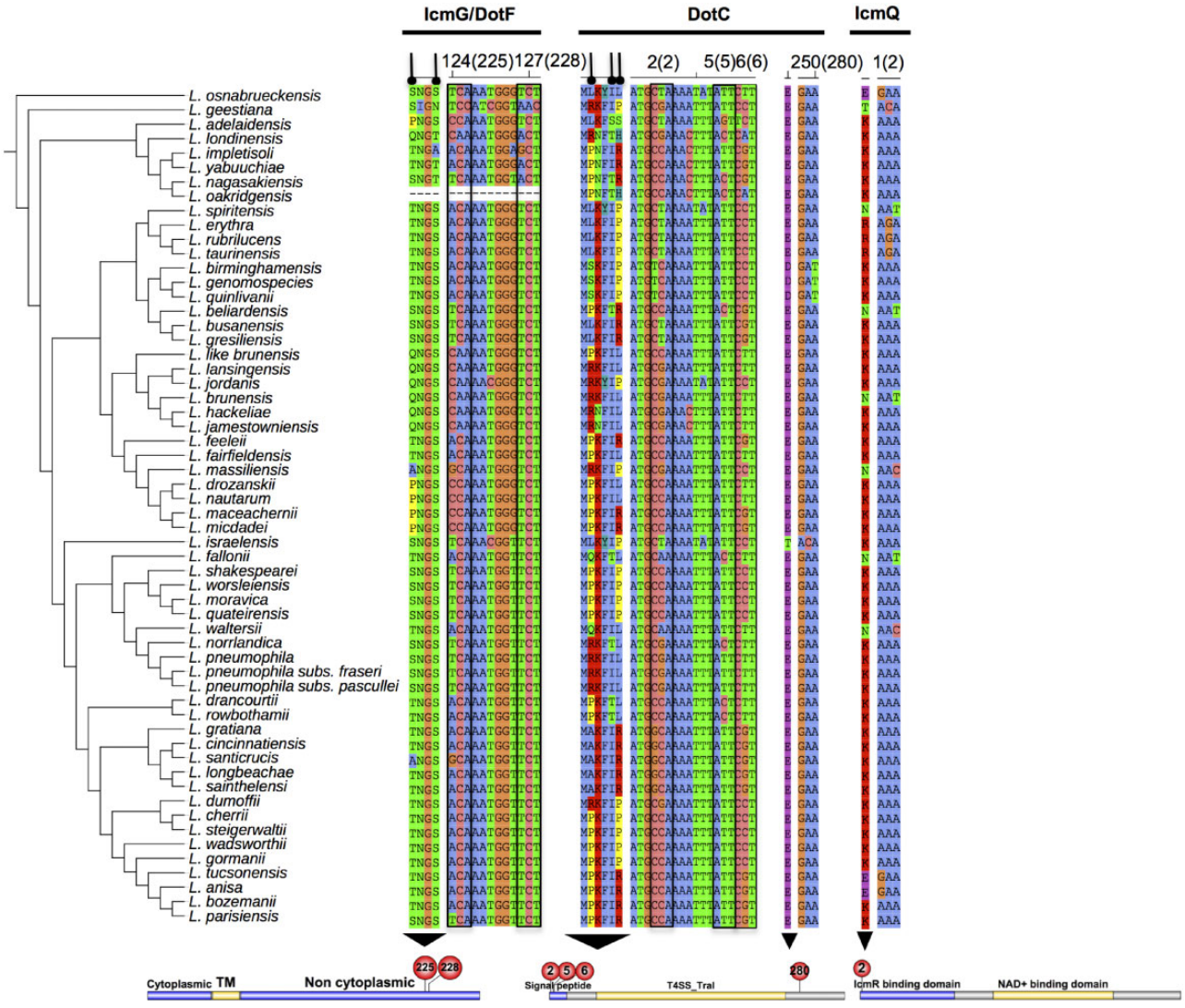

Fig. 2.—Amino acids/codons under positive selection (detected by at least two different methods) in the proteins IcmG/DotF, DotC, and IcmQ. The fragment/s of the alignment containing amino acids/codons under positive selection in IcmG/DotF, DotC, or IcmQ are indicated for each species ordered according to their position in the phylogenetic tree depicted on the left hand side of the figure. The lines with circle heads point to the amino acids that are under positive selection whereas the corresponding codons are framed with a black line. The numbers above highlighted codons correspond to the codon number in the analyzed alignment whereas the corresponding codon number in the sequence from *Legionella pneumophila* strain Paris is shown between brackets. The main domains of the protein and the position of the codons under positive selection are shown in the bottom of the figure taking as reference the *L. pneumophila* Paris sequence.

species and defines the region of DotF involved in effector binding.

High Conservation and Positive Selection Characterize the Outer Part of the Core Complex

Whereas DotG and DotF contact the inner membrane proteins, three other proteins: DotH (Lpp0516), DotC (Lpp2729), and DotD (Lpp2728) constitute the outer part of the core transmembrane complex (fig. 1) (Nakano et al. 2010). Moreover, it was reported that DotH is associated with a fibrous structure that covers the entire bacterial surface under certain conditions enhancing bacteria internalization (Watarai et al. 2001). If exposed on the bacterial surface, we would expect that this protein interacts with the host system and is subjected to an evolutionary arm race. However, except an intrinsically disordered N-terminal region of 50 amino acids (15% of the protein) DotH shows a high degree of interspecies conservation (table 1). Moreover, when removing this region, no signatures of recombination or of positive selection are detectable (table 2). Thus, although we cannot discard that the disordered N-terminal region contains amino acids

under positive selection our results suggest that DotH is not directly exposed to the host system as previously described as no sign of diversifying selection acting on this protein was detected.

DotC and DotD, like DotH show a high degree of conservation (table 1) and contain signal peptides responsible for their secretion through the outer membrane. Our analyses identified three codons in the DotC N-terminal region within the signal peptide that are under positive selection (table 2 and fig. 2). The DotC signal peptide is well conserved as it was detected in 63 of the 80 sequenced DotC proteins (supplementary table S6, Supplementary Material online). In addition to targeting preprotein secretion, it has been shown that the cleaved signal peptides can play additional roles as hormones, neurotransmitters, or self-antigens (Hegde and Bernstein 2006). It is thus tempting to suggest that the released signal peptide of *Legionella* DotC is detected by, for example, the host immune system explaining why positive selection is acting on this sequence. However, alternative explanations such as relaxed selection due for example to alternative start codons cannot be discarded. Additionally, we also detected positive selection acting on amino acid 250 of DotC (fig. 2), however, only MEME is supporting this and very few species are affected by this amino acid change, thus this result needs to be taken with care. In contrast, DotD sequence conservation is high in all species and DotD shows no signs of recombination or of positive selection acting on it. The crystal structure of this protein showed that it has striking structural similarity to the N-terminal subdomain of secretins and N0 domain proteins (Nakano et al. 2010). They found that the DotD/N0/T3S domain is present in outer membrane components of many even distantly related secretion systems indicating that negative selection is acting on it which is in line with our results.

In summary, although highly conserved at the sequence level, the outer membrane side of the core complex shows positive selection pressure acting on the signal peptide sequence of DotC suggesting an additional role of this protein in the host cell after cleavage.

DotB and DotO are Subjected to Strong Negative Selection

DotB (Lpp2730) is a protein that forms stable homohexameric rings and hydrolyses ATP (Sexton, Miller, et al. 2004). The corresponding mutant is defective for growth in macrophages, but Sexton et al. (2005) obtained some *dotB* alleles with partial activity. Two of these were unable to export a subset of T4SS substrates indicating a possible role of DotB in substrate selection. Our comparative analysis shows that DotB is the Dot/Icm protein with the highest degree of conservation (93% average amino acid identity among different species; table 1) and even DotB from *C. burnetti* can complement a *L. pneumophila dotB* mutant (Zamboni et al. 2003). Furthermore, no recombination or sites under positive selection were identified in the DotB proteins. A similar result was found for the ATPase DotO (Lpp0522) where conservation at the sequence level is high and neither branches nor sites under positive selection were detected. Recently, it has been reported that DotB is a dynamic entity as it can be free in the cytosol or associated with the Dot/Icm system through the DotO ATPase (Chetrit et al. 2018) where it constitutes the disc at the base of the cytoplasmic complex of the Dot/Icm system. Our results demonstrate that these proteins are mostly under negative selection and suggest that the basal structure composed by DotB-DotO is highly structurally constrained to allow the passage of all effectors across the DotB-DotO energy complex.

Analysis of Proteins of the Coupling Subcomplex

Recombination and Diversifying Selection Shape DotL and DotM Evolution

The Dot/Icm system-coupling subcomplex (T4CP) contains integral inner-membrane proteins that play a dual role of recruiting substrates and escorting them to the secretion conduit (Gomis-Ruth et al. 2004). DotL and DotM are part of this subcomplex (fig. 1). DotL (IcmO/Lpp0512) is an inner membrane protein related to *Escherichia coli* VirD4 and TrwB (Buscher et al. 2005), both structural prototypes of coupling subcomplexes in T4SS. However, compared with TrwB, DotL contains an additional 200-residue segment at the C-terminus of unknown function that is found also in *Coxiella*, *Yersinia*, and *Pseudomonas* species (Kwak et al. 2017).

Our analysis shows a high degree of interspecies conservation of DotL. This result is in line with the knowledge that DotL is involved in multiple interactions with DotN, IcmS, IcmW, and DotM and therefore under high structural constraints to maintain the architecture of the coupling subcomplex. Moreover, DotL plays an important role in intracellular replication since the corresponding mutants are defective in replication in a variety of host cells (Sutherland et al. 2012). Despite its high conservation several branches are affected by diversifying selection (aBSREL) but we did not detect any site under positive selection. Furthermore, we detected intragenic recombination events affecting mostly *L. pneumophila* strains and the species *Legionella dumoffi* and *Legionella worseilensis*. The regions involved in recombination were always located in the P-loop Ntpase domain (supplementary fig. S5*A*, Supplementary Material online), but were not affecting DotL regions that interact with IcmS-W, DotN, and DotM (Vincent et al. 2012). Taken together, most of the DotL sequence is highly conserved whereas the variability accumulates mainly in the DotM-interacting domain of the transmembrane region and in particular in the segment of the C-terminus (supplementary figs. S5*B* and S6, Supplementary Material online). Although we did not detect specific sites under selection in DotL, the high variability localized in the C-terminal region is a sign of a fast evolution rate of this part of the protein, which is

in line with a possible role in effector binding as recently suggested by analyzing the DotL structure (Chetrit et al. 2018).

Another component of this coupling complex is DotM (IcmP/Lpp0511), a protein that possesses a cytoplasmic domain that has just been crystalized (Meir et al. 2018) and that is thought to interact with DotL through their transmembrane domains (fig. 1) (Vincent et al. 2012). Indeed, we observed that the transmembrane region of DotM is highly variable like the DotM-interacting domain of DotL as mentioned above and that positive selection is acting on codon 22 localized at the end of the first transmembrane helix domain of DotM. These results further suggest coevolution of both proteins. The analyses of the crystal structure of the DotM cytoplasmic domain revealed that it contains large patches of basic residues suggesting that it might form a recruiting platform for Glu-rich motif effectors containing the so-called E-block motif (Huang et al. 2011; Meir et al. 2018). Indeed, Meir et al. (2018) demonstrated that DotM can bind acidic Glu rich peptides which is in agreement with our results that identify a weak positive selection signal on codon 362 of the cytoplasmic domain (table 2 and supplementary fig. S7, Supplementary Material online). Moreover, this codon under positive selection has undergone amino acid replacement alternating between polar and neutrally charged residues suggesting that this impacts DotM-effector interactions.

Taken together, we show that despite a general high degree of sequence conservation of both DotM and DotL, specific regions of these two proteins are under fast evolution pointing to their potential role in the interaction with substrates. In contrast to DotM and DotL, the proteins DotN and IcmT that were also suggested to be involved in effector recruitment (Meir et al. 2018) show a high conservation and no recombination nor sites and/or branches under positive selection.

Negative Selection Drives the Evolution of the Chaperones IcmS-W

IcmS (Lpp0508) and IcmW (Lpp2742) are small acidic cytoplasmic proteins that interact with each other while being part of the coupling protein complex (fig. 1). The crystal structure shows that the two C-terminal alpha helices of IcmW interact with IcmS to form a structure with a concave surface containing hydrophobic residues that interact with LvgA (Lpp0590), whereas DotL binds IcmSW and also DotN through its C-terminus (Kwak et al. 2017). Interestingly, IcmS-W and IcmS-LvgA have been involved in substrate recognition in previous studies (Bardill et al. 2005; Ninio et al. 2005; Vincent, Friedman, et al. 2006).

Our analyses show that IcmS and IcmW are highly conserved among species and no signal of recombination or positive selection was detected (table 2). Instead, the large majority of the analyzed codons of IcmS and IcmW are subjected to negative selection (e.g., 109 from 111 analyzed codons of IcmS), which would suggest that IcmSW are not evolving to adapt to different set of effectors (table 2). However, it has been suggested that during interaction with IcmSW, effectors adopt an unfolded conformation (Xu et al. 2017) and that the IcmSW surface that binds effectors interacts also with DotL. Together, these data are suggesting that like most chaperones, IcmS and IcmW have little interaction specificity explaining the lack of positive selection acting on them despite their potential role in effector binding. In contrast, LvgA shows higher variability, especially in the C- and N-terminal regions, and at least one node in the phylogeny is under positive selection. This result fits well with the crystal structure of the coupling subcomplex (Kwak et al. 2017) showing that whereas most of the IcmSW surfaces are interacting with other proteins of the complex, LvgA possesses some loops exposed to the cytoplasmic side. In addition, LvgA is critical for recruitment of certain substrate as only when it is present, the complexes DotL-DotN-IcmSW and IcmSW can bind effectors (Kwak et al. 2017).

Evolution of Cytoplasmic Proteins

Extremely Fast Evolution is Acting on the IcmRQ Complex

IcmQ and IcmR are essential for growth of *L. pneumophila* in macrophages (Coers et al. 2000). These proteins interact in vivo in *L. pneumophila* (Dumenil and Isberg 2001) through the middle region of IcmR and the N-terminal region of IcmQ (Raychaudhury et al. 2009). When not bound to IcmR, IcmQ can insert into the lipid membrane forming pores through its N-terminal part (Dumenil et al. 2004). Our analysis of IcmQ (Lpp0510) shows a moderate conservation (70% average amino acidic identity), no recombination but at least one amino acid under positive selection, localized at the beginning of the protein (table 2 and fig. 2). Curiously, this amino acid is located in the part of the protein that has been defined as the interacting with IcmR (amino acids 1–57 in *L. pneumophila* Paris). The alignment shows that most of the hydrophobic residues previously defined to be involved in this interaction (Raychaudhury et al. 2009) are also conserved in all *Legionella* species analyzed here (supplementary fig. S8, Supplementary Material online). The structure of full-length IcmQ in complex with IcmR revealed that the C-terminal domain of IcmQ contains a NAD^+ binding domain (Farelli et al. 2013). An alignment of IcmR from different bacteria was used to define the essential residues of this IcmR NAD^+ domain (Farelli et al. 2013). Here, our interspecies alignment revealed which of these residues have a higher conservation among species and are therefore potentially essential for IcmR function (supplementary fig. S8, Supplementary Material online). The presence of this module has allowed to suggest that IcmRQ binds to membranes, where it may interact with, or perhaps modify, a protein in the T4SS when NAD(+) is bound (Farelli et al. 2013).

IcmR (Lpp0509) is the Dot/Icm protein with the highest rate of evolution of all Dot/Icm proteins. An *icmR* gene similar to *L. pneumophila icmR* was identified only in the species *Legionella norrlandica*, the phylogenetically closest species to *L. pneumophila* (Gomez-Valero et al. 2019). Thus, IcmR from *L. pneumophila* strain Paris can only be aligned with homologous proteins belonging either to strains from the same species or to the closely related *L. norrlandica*, which explains the high amino acid identity values obtained (table 1) despite the high evolutionary rate of IcmR. Indeed, in all other *Legionella* species analyzed, one or two nonhomologous genes replace this gene in the same position where *icmR* is present in *L. pneumophila* (Gomez-Valero et al. 2019). Feldman et al. (2005) had shown that the genes that replace *icmR* in *Legionella hackeliae* and *Legionella micdadei* are functional homologs of *L. pneumophila icmR* (designated FIR proteins). Despite the lack of sequence homology, two conserved structural regions were predicted in the FIR proteins containing nonidentical, hydrophobic side chains that may contribute to the binding between IcmR and IcmQ (Raychaudhury et al. 2009). We have shown that these two regions in FIR proteins are also conserved in 58 different *Legionella* species (Gomez-Valero et al. 2019). However, the absence of homology at the sequence level for IcmR in the different *Legionella* sp. constitutes a limiting factor for the analysis of diversifying selection. Homology among five or more strains was present only in two subgroups, one containing the species *Legionella gratiana*, *L. sainthelensi*, *L. longbeachae*, *Legionella cincinatiensis*, and *L. santicrucis* and the other one containing *L. pneumophila* and *L. norrlandica* strains. Therefore, we used these two groups to search for positive selection within IcmR. Among *L. pneumophila* and *L. norlandica* strains (16 sequences), codons 39 and 90 were identified as being under positive selection. In contrast, within the *L. longbeachae* cluster, codon 10 was under selection (data not shown). Thus, positive selection seems to act on specific amino acids of IcmR. The reason why this gene is so extremely divergent is not known. Originally, it was suggested that the FIR–IcmQ complex is secreted upon contact with a protozoan host cell what would explain the positive selection acting on it (Dumenil and Isberg 2001). Later, the crystal structures of the N-terminal domain of IcmQ with the interacting region of IcmR suggested that IcmQ is associated with the inner bacterial membrane (Raychaudhury et al. 2009) and consequently not exposed to the host system. Therefore, the diversifying selection acting on these proteins and more specifically the high evolutionary rate of IcmR is probably linked to the large variety of Dot/Icm effectors secreted in the different *Legionella* species by this system.

Our analysis combined with the crystal structures suggest that IcmR may have a central role in substrate interaction and thus needs constantly to adapt to the changing effector repertoire in the different species.

Selection Analysis of Dot/Icm Components of Yet Unknown Function

IcmX (Lpp2743) is a 50-kDa periplasmic protein that is essential for *L. pneumophila* pathogenesis (Sadosky et al. 1993; Edelstein et al. 1999; Matthews and Roy 2000) and required for pore formation in the membrane of the eukaryotic cell (Matthews and Roy 2000). Our analysis reveals that, although present in all analyzed *Legionella* species, IcmX is one of the least conserved proteins of this system (table 1), especially in the N-terminal region. Consequently, many regions of the gene could not be included in our analysis due to uncertainty in their corresponding alignment. However, we still detected intragenic recombination in *icmX* in four *Legionella* species (table 2) and one branch of the tree under positive selection (aBSREL). This diversifying selection in some lineages may reflect a role of IcmX as signal transmitter to the host cell. It has been reported that a truncated IcmX product is secreted into culture supernatants by *L. pneumophila* (Matthews and Roy 2000), although its translocation across eukaryotic cell membranes has not been detected. Additionally, it has been shown that IcmX is a surface exposed protein (Khemiri et al. 2008). Together, these results strongly suggest that IcmX is a protein exposed to the host, and therefore, in different *Legionella* species, it interacts with different protozoan hosts explaining its fast evolution.

The Dot/Icm system comprises also many membrane-associated proteins, such as DotK/IcmN (Lpp0513), IcmF (Lpp0524), DotU (Lpp0525), DotE (Lpp0519), DotV (Lpp0537), DotP (Lpp0520), DotI (Lpp0515), DotJ (Lpp0514), and IcmV (Lpp2741). Among those, IcmF and DotU proteins prevent DotH degradation, stabilize the *L. pneumophila* T4SS (Sexton, Pinkner, et al. 2004) and recruit the DotCH complexes to the poles of the cell (Jeong et al. 2017). Our analysis revealed no recombination or diversifying selection acting on DotU, but IcmF contains one residue under positive selection (MEME) and several branches are affected by diversifying selection (supplementary fig. S9, Supplementary Material online). The amino acid under positive selection is located in a conserved segment among IcmF homologs and outside the C-terminal region predicted to contact DotCH (Ghosal et al. 2017). This is consistent with the fact that the IcmF segment predicted to interact with DotCH is likely subjected to structural constraints that prevent amino acid changes. Among DotK, DotE, DotP, DotI, DotJ, IcmV, and DotV, recombination and positive selection were detected only for DotV and positive selection affects the last codon of one of the predicted transmembrane helices (Nagai and Kubori 2011) (supplementary fig. S11, Supplementary Material online).

These results demonstrate that IcmX, IcmF, and DotV are subjected to diversifying selection suggesting that they play a role in host/effector interaction and are thus interesting targets for future functional studies.

Concluding Remarks

It is well known that the constant arms race between pathogens and hosts selects for the maintenance of polymorphisms thereby allowing adaptations and counter-adaptations to occur. The *Legionella* Dot/Icm type IVB secretion system or at least a part of it has to contact the host cell for the delivery of effectors. It is thus a target of the pathogen recognition systems and a hot spot of selection. Our study analyzing these evolutionary forces acting on it revealed high rates of recombination and/or positive selection for proteins DotA, DotG, and IcmX suggesting, in line with previous studies, that these proteins are directly interacting with the host system. Moreover, our analysis highlights DotC whose signal peptide is subjected to diversifying selection which may indicate that after cleavage it is released and plays a role in the host cell. In contrast, our results did not support the suggestion that DotH and DotO have a role in host interaction.

In contrast to the high conservation of the Dot/Icm secretion system at interspecific level, the effector repertoire is very variable (Burstein et al. 2016; Gomez-Valero et al. 2019) suggesting that proteins involved in effector binding have to evolve through diversifying selection to adapt to different effector sets in the different species. Indeed, we detected diversifying selection acting on DotL and DotM that had been suggested previously to be involved in effector recruitment. Moreover, we detected amino acids and regions under positive selection in DotF, the cytoplasmic domain of DotM and the C-terminal region of DotL pointing to protein segments probably involved in interacting with Dot/Icm substrates. In the case of IcmR, several amino acids were under positive selection, despite a limited analysis possibility due to lack of conservation of this protein at interspecific level. The positive selection acting on the protein together with its extremely high rate of evolution clearly points to a key role in substrate recruitment. Furthermore, IcmF, DotV, and DotK contain amino acids and/or branches under positive selection suggesting they are involved in effector/host interactions. In contrast, DotO and DotB that constitute the base of the cytoplasmic Dot/Icm complex are under strong negative selection as neither signs of recombination nor of positive selection and a high degree of interspecific conservation were detected suggest that these proteins are essential for the maintenance of the architecture and or function of the T4SS.

In conclusion, our evolutionary studies of the Dot/Icm components allowed identifying those proteins and amino acids of this secretion system that may be functionally important for host/effector binding. Indeed, the detection of diversifying selection acting on pilus proteins of the type III secretion system of *Pseudomonas syringae* (Guttman et al. 2006) or on the type IV secretion system of *Bartonella* (Nystedt et al. 2008) suggested their role in host–pathogen interactions. For the Dot/Icm secretion system, we are far from a complete understanding of its structural and functional mechanism, but the availability of interspecies genome data allows new ways to analyze their components and thereby to predict their role. Functional analyses of the proteins predicted here through the analysis of evolutionary forces acting on them will be exciting to gain further insight into this important secretion system.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We would like to thank Dr Miroslaw Cygler for his constructive criticism on the article. This work was supported by the Institut Pasteur, the Agence National de la Recherche (grant n°ANR-10-LABX-62-IBEID) and the Fondation pour la Recherche Médicale (FRM) (grant N°DEQ20120323697).

Author Contributions

L.G.V., I.C., and A.C.O. did the data analyses and interpretation. The article was written by L.G.V. and C.B. with input from coauthors. The project was conceived, planned, and supervised by L.G.V. and C.B.

Literature Cited

Bardill JP, Miller JL, Vogel JP. 2005. IcmS-dependent translocation of SdeA into macrophages by the *Legionella pneumophila* type IV secretion system. Mol Microbiol. 56(1):90–103.

Barrozo RM. 2013. Functional plasticity in the type IV secretion system of *Helicobacter pylori*. PLoS Pathog. 9(2):e1003189.

Berger KH, Merriam JJ, Isberg RR. 1994. Altered intracellular targeting properties associated with mutations in the *Legionella pneumophila dotA* gene. Mol Microbiol. 14(4):809–822.

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.

Burstein D, et al. 2009. Genome-scale identification of *Legionella pneumophila* effectors using a machine learning approach. PLoS Pathog. 5(7):e1000508.

Burstein D, et al. 2016. Genomic analysis of 38 *Legionella* species identifies large and diverse effector repertoires. Nat Genet. 48(2):167–175.

Buscher BA, et al. 2005. The DotL protein, a member of the TraG-coupling protein family, is essential for viability of *Legionella pneumophila* strain Lp02. J Bacteriol. 187(9):2927–2938.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17(4):540–552.

Chetrit D, Hu B, Christie PJ, Roy CR, Liu J. 2018. A unique cytoplasmic ATPase complex defines the *Legionella pneumophila* type IV secretion channel. Nat Microbiol. 3(6):678–686.

Christie PJ, Gomez Valero L, Buchrieser C. 2017. Biological diversity and evolution of type IV secretion systems. Curr Top Microbiol Immunol. 413:1–30.

Christie PJ, Vogel JP. 2000. Bacterial type IV secretion: conjugation systems adapted to deliver effector molecules to host cells. Trends Microbiol. 8(8):354–360.

Coers J, et al. 2000. Identification of Icm protein complexes that play distinct roles in the biogenesis of an organelle permissive for *Legionella pneumophila* intracellular growth. Mol Microbiol. 38(4):719–736.
Costa J, Tiago I, Da Costa MS, Verissimo A. 2010. Molecular evolution of *Legionella pneumophila dotA* gene, the contribution of natural environmental strains. Environ Microbiol. 12(10):2711–2729.
Dumenil G, Isberg RR. 2001. The *Legionella pneumophila* IcmR protein exhibits chaperone activity for IcmQ by preventing its participation in high-molecular-weight complexes. Mol Microbiol. 40(5):1113–1127.
Dumenil G, Montminy TP, Tang M, Isberg RR. 2004. IcmR-regulated membrane insertion and efflux by the *Legionella pneumophila* IcmQ protein. J Biol Chem. 279:4686–4695.
Edelstein PH, Edelstein MA, Higa F, Falkow S. 1999. Discovery of virulence genes of *Legionella pneumophila* by using signature tagged mutagenesis in a guinea pig pneumonia model. Proc Natl Acad Sci U S A. 96(14):8190–8195.
Edelstein PH, Hu B, Higa F, Edelstein MA. 2003. lvgA, a novel *Legionella pneumophila* virulence factor. Infect Immun. 71(5):2394–2403.
Ensminger AW. 2016. *Legionella pneumophila*, armed to the hilt: justifying the largest arsenal of effectors in the bacterial world. Curr Opin Microbiol. 29:74–80.
Escoll P, Mondino S, Rolando M, Buchrieser C. 2016. Targeting of host organelles by pathogenic bacteria: a sophisticated subversion strategy. Nat Rev Microbiol. 14(1):5–19.
Farelli JD, et al. 2013. IcmQ in the Type 4b secretion system contains an NAD+ binding domain. Structure 21(8):1361–1373.
Feldman M, Zusman T, Hagag S, Segal G. 2005. Coevolution between nonhomologous but functionally similar proteins and their conserved partners in the *Legionella* pathogenesis system. Proc Natl Acad Sci U S A. 102(34):12206–12211.
Finn RD, et al. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44(D1):D279–D285.
Finsel I, Hilbi H. 2015. Formation of a pathogen vacuole according to *Legionella pneumophila*: how to kill one bird with many stones. Cell Microbiol. 17(7):935–950.
Garcia-Boronat M, Diez-Rivero CM, Reinherz EL, Reche PA. 2008. PVS: a web server for protein sequence variability analysis tuned to facilitate conserved epitope discovery. Nucleic Acids Res. 36(Web Server):W35–W41.
Ghosal D, Chang YW, Jeong KC, Vogel JP, Jensen GJ. 2017. In situ structure of the *Legionella* Dot/Icm type IV secretion system by electron cryotomography. EMBO Rep. 18(5):726–732.
Gomez Valero L, Runsiok 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, 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(6):2265–2273.
Gomez-Valero L, et al. 2014. Comparative analyses of *Legionella* species identifies genetic features of strains causing Legionnaires' disease. Genome Biol. 15(11):505.
Gomis-Ruth FX, Sola M, de la Cruz F, Coll M. 2004. Coupling factors in macromolecular type-IV secretion machineries. Curr Pharm Des. 10(13):1551–1565.
Guglielmini J, de la Cruz F, Rocha EP. 2013. Evolution of conjugation and type IV secretion systems. Mol Biol Evol. 30(2):315–331.
Guttman DS, Gropp SJ, Morgan RL, Wang PW. 2006. Diversifying selection drives the evolution of the type III secretion system pilus of *Pseudomonas syringae*. Mol Biol Evol. 23(12):2342–2354.
Hegde RS, Bernstein HD. 2006. The surprising complexity of signal sequences. Trends Biochem Sci. 31(10):563–571.
Huang L, et al. 2011. The E Block motif is associated with *Legionella pneumophila* translocated substrates. Cell Microbiol. 13(2):227–245.
Hubber A, Roy CR. 2010. Modulation of host cell function by *Legionella pneumophila* type IV effectors. Annu Rev Cell Dev Biol. 26(1):261–283.
Isberg RR, O'Connor TJ, Heidtman M. 2009. The *Legionella pneumophila* replication vacuole: making a cosy niche inside host cells. Nat Rev Microbiol. 7(1):13–24.
Jeong KC, Ghosal D, Chang YW, Jensen GJ, Vogel JP. 2017. Polar delivery of *Legionella* type IV secretion system substrates is essential for virulence. Proc Natl Acad Sci U S A. 114(30):8077–8082.
Kanehisa M. 1997. Linking databases and organisms: genomeNet resources in Japan. Trends Biochem Sci. 22(11):442–444.
Khemiri A, et al. 2008. Outer-membrane proteomic maps and surface-exposed proteins of *Legionella pneumophila* using cellular fractionation and fluorescent labelling. Anal Bioanal Chem. 390(7):1861–1871.
Ko KS, Hong SK, Lee HK, Park MY, Kook YH. 2003. Molecular evolution of the *dotA* gene in *Legionella pneumophila*. J Bacteriol. 185(21):6269–6277.
Kryazhimskiy S, Plotkin JB. 2008. The population genetics of d*N*/d*S*. PLoS Genet. 4(12):e1000304.
Kubori T, et al. 2014. Native structure of a type IV secretion system core complex essential for *Legionella* pathogenesis. Proc Natl Acad Sci U S A. 111(32):11804–11809.
Kwak MJ, et al. 2017. Architecture of the type IV coupling protein complex of *Legionella pneumophila*. Nat Microbiol. 2(9):17114.
Lam HM, Ratmann O, Boni MF. 2018. Improved algorithmic complexity for the 3SEQ recombination detection algorithm. Mol Biol Evol. 35(1):247–251.
Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44(W1):W242–W245.
Lifshitz Z, et al. 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(8):E707–E715.
Liu W, et al. 2015. IBS: an illustrator for the presentation and visualization of biological sequences. Bioinformatics 31(20):3359–3361.
Loytynoja A. 2014. Phylogeny-aware alignment with PRANK. Methods Mol Biol. 1079:155–170.
Luo ZQ, Isberg RR. 2004. Multiple substrates of the *Legionella pneumophila* Dot/Icm system identified by interbacterial protein transfer Proc Natl Acad Sci U S A. 2004;101(3):841–846.
Martin DP, Lemey P, Posada D. 2011. Analysing recombination in nucleotide sequences. Mol Ecol Resour. 11(6):943–955.
Matthews M, Roy CR. 2000. Identification and subcellular localization of the *Legionella pneumophila* IcmX protein: a factor essential for establishment of a replicative organelle in eukaryotic host cells. Infect Immun. 68(7):3971–3982.
Meir A, Chetrit D, Liu L, Roy CR, Waksman G. 2018. *Legionella* DotM structure reveals a role in effector recruiting to the Type 4B secretion system. Nat Commun. 9(1):507.
Murrell B, et al. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 8(7):e1002764.
Murrell B, et al. 2013. FUBAR: a fast, unconstrained Bayesian approximation for inferring selection. Mol Biol Evol. 30(5):1196–1205.
Nagai H, Kubori T. 2011. Type IVB secretion systems of *Legionella* and other Gram-negative bacteria. Front Microbiol. 2:136.
Nagai H, Roy CR. 2001. The DotA protein from *Legionella pneumophila* is secreted by a novel process that requires the Dot/Icm transporter. EMBO J. 20(21):5962–5970.
Nakano N, Kubori T, Kinoshita M, Imada K, Nagai H. 2010. Crystal structure of *Legionella* DotD: insights into the relationship between type IVB and type II/III secretion systems. PLoS Pathog. 6(10):e1001129.

Newton HJ, Ang DK, van Driel IR, Hartland EL. 2010. Molecular pathogenesis of infections caused by *Legionella pneumophila*. Clin Microbiol Rev. 23(2):274–298.
Ninio S, Zuckman-Cholon DM, Cambronne ED, Roy CR. 2005. The *Legionella* IcmS-IcmW protein complex is important for Dot/Icm-mediated protein translocation. Mol Microbiol. 55(3):912–926.
Nystedt B, Frank AC, Thollesson M, Andersson SG. 2008. Diversifying selection and concerted evolution of a type IV secretion system in *Bartonella*. Mol Biol Evol. 25(2):287–300.
Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 8(10):785–786.
Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21(5):676–679.
Qiu J, Luo ZQ. 2017. *Legionella* and *Coxiella* effectors: strength in diversity and activity. Nat Rev Microbiol. 15(10):591–605.
Raychaudhury S, et al. 2009. Structure and function of interacting IcmR-IcmQ domains from a type IVb secretion system in *Legionella pneumophila*. Structure 17(4):590–601.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16(6):276–277.
Rohde M, Püls J, Buhrdorf R, Fischer W, Haas R. 2003. A novel sheathed surface organelle of the *Helicobacter pylori* cag type IV secretion system. Mol Microbiol. 49(1):219–234.
Roy CR, Isberg RR. 1997. Topology of *Legionella pneumophila* DotA: an inner membrane protein required for replication in macrophages. Infect Immun. 65(2):571–578.
Sadosky AB, Wiater LA, Shuman HA. 1993. Identification of *Legionella pneumophila* genes required for growth within and killing of human macrophages. Infect Immun. 61(12):5361–5373.
Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18(3):502–504.
Segal G, Purcell M, Shuman HA. 1998. Host cell killing and bacterial conjugation require overlapping sets of genes within a 22-kb region of the *Legionella pneumophila* genome. Proc Natl Acad Sci U S A. 95(4):1669–1674.
Sexton JA, Pinkner JS, et al. 2004. The *Legionella pneumophila* PilT homologue DotB exhibits ATPase activity that is critical for intracellular growth. J Bacteriol. 186(6):1658–1666.
Sexton JA, Miller JL, Yoneda A, Kehl-Fie TE, Vogel JP. 2004. *Legionella pneumophila* DotU and IcmF are required for stability of the Dot/Icm complex. Infect Immun. 72(10):5983–5992.
Sexton JA, Yeo HJ, Vogel JP. 2005. Genetic analysis of the *Legionella pneumophila* DotB ATPase reveals a role in type IV secretion system protein export. Mol Microbiol. 57(1):70–84.
Shimodaira H, Hasegawa H. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 16(8):1114–1116.
Smith JM. 1992. Analyzing the mosaic structure of genes. J Mol Evol. 34(2):126–129.
Smith MD, et al. 2015. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol Biol Evol. 32(5):1342–1353.
Strimmer K, von Haeseler A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc Natl Acad Sci U S A. 94(13):6815–6819.
Sutherland MC, Binder KA, Cualing PY, Vogel JP. 2013. Reassessing the role of DotF in the *Legionella pneumophila* type IV secretion system. PLoS One 8(6):e65529.
Sutherland MC, Nguyen TL, Tseng V, Vogel JP. 2012. The *Legionella* IcmSW complex directly interacts with DotL to mediate translocation of adaptor-dependent substrates. PLoS Pathog. 8(9):e1002910.
Vincent CD, Buscher BA, et al. 2006. Identification of non-Dot/Icm suppressors of the *Legionella pneumophila* Delta*dotL* lethality phenotype. J Bacteriol. 188(23):8231–8243.
Vincent CD, Friedman JR, et al. 2006. Identification of the core transmembrane complex of the *Legionella* Dot/Icm type IV secretion system. Mol Microbiol. 62(5):1278–1291.
Vincent CD, Friedman JR, Jeong KC, Sutherland MC, Vogel JP. 2012. Identification of the DotL coupling protein subcomplex of the *Legionella* Dot/Icm type IV secretion system. Mol Microbiol. 85(2):378–391.
Vincent CD, Vogel JP. 2006. The *Legionella pneumophila* IcmS-LvgA protein complex is important for Dot/Icm-dependent intracellular growth. Mol Microbiol. 61(3):596–613.
Watarai M, Andrews HL, Isberg RR. 2001. Formation of a fibrous structure on the surface of *Legionella pneumophila* associated with exposure of DotH and DotO proteins after intracellular growth. Mol Microbiol. 39(2):313–329.
Xia X. 2017. DAMBE6: new tools for microbial genomics, phylogenetics, and molecular evolution. J Hered. 108(4):431–437.
Xu J, et al. 2017. Structural insights into the roles of the IcmS-IcmW complex in the type IVb secretion system of *Legionella pneumophila*. Proc Natl Acad Sci U S A. 114(51):13543–13548.
Zamboni DS, McGrath S, Rabinovitch M, Roy CR. 2003. *Coxiella burnetii* express type IV secretion system proteins that function similarly to components of the *Legionella pneumophila* Dot/Icm system. Mol Microbiol. 49(4):965–976.
Zhu W, et al. 2011. Comprehensive identification of protein substrates of the Dot/Icm type IV transporter of *Legionella pneumophila*. PLoS One 6(3):e17638.
Zhu W, Luo ZQ. 2013. Methods for determining protein translocation by the *Legionella pneumophila* Dot/Icm type IV secretion system. Methods Mol Biol. 954:323–332.

Associate editor: Ruth Hershberg