

More than 18,000 effectors in the Legionella genus genome provide multiple, independent combinations for replication in human cells

Laura Gomez-Valero, Christophe Rusniok, Danielle Carson, Sonia Mondino, Ana Elena Pérez-Cobas, Monica Rolando, Shivani S. Pasricha, Sandra Reuter, Jasmin Demirtas, Johannes Crumbach, et al.

▶ To cite this version:

Laura Gomez-Valero, Christophe Rusniok, Danielle Carson, Sonia Mondino, Ana Elena Pérez-Cobas, et al.. More than 18,000 effectors in the Legionella genus genome provide multiple, independent combinations for replication in human cells. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116 (6), pp.2265-2273. 10.1073/pnas.1808016116 . pasteur-02563435

HAL Id: pasteur-02563435 https://pasteur.hal.science/pasteur-02563435

Submitted on 16 Jun2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 More than 18,000 effectors in the *Legionella* genus genome provide multiple,

2

independent combinations for replication in human cells

3	
4 5	Laura Gomez-Valero ^{$1,2$} Christophe Rusnick ^{$1,2$} Danielle Carson ³ Sonia Mondino ^{$1,2$} Ana
6	Elena Pérez-Cobas ^{$l,2$} Monica Rolando ^{$l,2$} Shivani Pasricha ⁴ Sandra Reuter ⁵⁺ Jasmin
7	Demirtas ^{1,2} , Johannes Crumbach ^{1,2} , Stephane Descorps-Declere ⁶ , Elizabeth L. Hartland ^{4,7,8} ,
8	Sophie Jarraud ⁹ , Gordon Dougan ⁵ , Gunnar N. Schroeder ^{3,10} , Gad Frankel ³ , and Carmen
9	Buchrieser ^{1,2,*}
10	
11	¹ Institut Pasteur, Biologie des Bactéries Intracellulaires, 75724 Paris, France, ² CNRS UMR
12	3525, 75724 Paris, France, ³ MRC Centre for Molecular Bacteriology and Infection,
13	Department of Life Sciences, Imperial College, London SW7 2AZ, United Kingdom,
14	Department of Microbiology and Immunology, University of Melbourne, at the Peter
15	Doherty Institute for Infection and Immunity, Melbourne, 3000, Victoria, Australia,
16 17	⁶ Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge,
1/ 10	UK, Institut Pasteur, Center of Bioinformatics, Biostatistics and Integrative Biology (C3BI), 75724 Daris, France ⁷ Centre for Inpute Immunity and Infectious Discesses, Hudson Institute of
10 10	Medical Research, Clayton 3168, Australia ⁸ Department of Molecular and Translational
20	Science Monash University Clayton 3168 Australia ⁹ CIRL International Center for
21	Infectiology Research Inserm U1111 CNRS UMR5308 Université Lyon 1 École Normale
22	Supérieure de Lyon, Lyon, F-69008 France; National Reference Centre of <i>Legionella</i> ,
23	Hospices Civils de Lyon, France, ¹⁰ Centre for Experimental Medicine, Queen's University
24	Belfast, 97 Lisburn Rd. Belfast, BT9 7BL, UK,
25	
26	
27	Running title: Legionellae and the emergence of human pathogens
28	
29	Key words: Legionella, protozoa, co-evolution, horizontal gene transfer, virulence
30 31	⁺ Present address: Medical Center - University of Freiburg Institute for Infection Prevention
32	and Hospital Epidemiology Breisacher Straße 115B 79106 Freiburg Germany
33	and mosphar Dpheemiology, Dielsacher Straße 115D, 75100 Heloarg, Semiany
34	
35	
36	
37	
38	
39	
40 41	*For correspondence
42	Carmen Buchrieser & Laura Gomez Valero
43	Institut Pasteur
44	Biologie des Bactéries Intracellulaires
45	28, rue du Dr. Roux, 75724 Paris Cedex 15, France
46	Tel: (33-1)-45-68-83-72
47	E-mail: <i>cbuch@pasteur.fr, lgomez@pasteur.fr</i>

1 Significance

2 Legionella pneumophila is a bacterial pathogen causing outbreaks of a lethal pneumonia. The 3 genus Legionella comprises 65 species for which aquatic amoebae are the natural reservoirs. Using functional and comparative genomics to deconstruct the entire bacterial genus we 4 5 reveal the surprising parallel evolutionary trajectories that have led to the emergence of 6 human pathogenic Legionella. An unexpectedly large and unique repository of secreted 7 proteins (>18,000) containing eukaryotic-like proteins acquired from all domains of life (plant, 8 animal, fungal, archaea) is contrasting with a highly conserved type 4 secretion system. This 9 study reveals an unprecedented environmental reservoir of bacterial virulence factors, and provides a new understanding of how reshuffling and gene-acquisition from environmental 10 11 eukaryotic hosts, may allow for the emergence of human pathogens.

12

13 Abstract

14 The bacterial genus Legionella comprises 65 species among which Legionella pneumophila is a human pathogen causing severe pneumonia. To understand the evolution of an 15 16 environmental to an accidental human pathogen, we have functionally analyzed 80 Legionella 17 genomes spanning 58 species. Uniquely, an immense repository of 18,000 secreted proteins encoding 137 different eukaryotic-like domains and more than 200 eukaryotic-like proteins is 18 19 paired with a highly conserved T4SS. Specifically, we show that eukaryotic Rho and Rab-20 GTPase domains are found nearly exclusively in eukaryotes and Legionella species. 21 Translocation assays for selected Rab-GTPase proteins revealed that they are indeed T4SS 22 secreted substrates. Furthermore, F/U-box and SET domains were present in >70% of all 23 species suggesting that manipulation of host signal transduction, protein turnover and 24 chromatin modification pathways, respectively are fundamental intracellular replication 25 strategies for Legionellae. In contrast, the Sec-7 domain was restricted to L. pneumophila and 26 seven other species, indicating effector repertoire tailoring within different amoebae. Functional screening of 47 species revealed 60% were competent for intracellular replication 27 28 in THP-1 cells, but interestingly this phenotype was associated with diverse effector 29 assemblages. These data, combined with evolutionary analysis indicate that the capacity to 30 infect eukaryotic cells has been acquired independently many times within the genus and that a highly conserved yet versatile T4SS secretes an exceptional number of different proteins 31 shaped by inter-domain gene transfer. Furthermore we revealed the surprising extent to which 32 33 legionellae have co-opted genes and thus cellular functions from their eukaryotic hosts and provides a new understanding of how dynamic reshuffling and gene-acquisition has led to the 34 35 emergence of major human pathogens.

 $1 \ \body$

2 Introduction

3 Legionnaires' disease or legionellosis is an atypical pneumonia caused by bacteria of the genus Legionella. Shortly after the discovery of L. pneumophila (1) it was reported that this 4 5 bacterium is pathogenic for freshwater and soil amoebae of the genera Acanthamoeba and Naegleria (2). This finding led to a new perception in microbiology, whereby bacteria that 6 7 parasitize protozoa can utilize similar processes to infect human cells. Sequencing and 8 analyses of the L. pneumophila genome substantiated this idea, when it revealed the presence 9 of a large number and variety of eukaryotic-like domains within the predicted proteome (3). Many of these proteins, termed effector proteins, were shown to be secreted into the host cell 10 11 where they facilitate Legionella intracellular replication within a specialized compartment 12 termed the Legionella containing vacuole (LCV) (3, 4). Overall, the type IV secretion system 13 (T4SS), Dot/Icm, secretes more than 300 different effector proteins into the host cell and is 14 indispensable for virulence of L. pneumophila (5-8). The presence of the Dot/Icm T4SS in other *L. pneumophila* strains and in selected *Legionella* species had also been reported (9-12) 15 16 but recent genome scale studies of Legionella (13-15) indicated that the T4SS system is 17 present in every Legionella strain analyzed.

18 Despite high conservation of the Dot/Icm system among different Legionella species, 19 effector repertoires appear to vary greatly. An analysis of putative T4SS effectors of 20 L. longbeachae, the second most frequent cause of Legionnaires' disease, revealed that only 21 about 50% of the virulence factors described in L. pneumophila were also present in the 22 genome of L. longbeachae (16). Recently, Burstein et al. (14) analyzed 38 Legionella species using a machine learning approach to predict T4SS effectors and Joseph et al. (15) examined 23 24 Legionella genome dynamics, both concluding that DNA interchange between different 25 species is rare. However, still little is known about the potential of the different species to 26 cause human disease and about the impact and the specific characteristics of the T4SS 27 effectors on the evolution of new human pathogens within this environmental bacterial genus.

Here we present a comprehensive analysis of the *Legionella* genus genome, covering 80 *Legionella* strains belonging to 58 *Legionella* species and subspecies. We establish a pangenus pool of putative T4SS effectors and show that this comprises over 18,000 proteins and identify more than 200 new eukaryotic-like proteins and 137 eukaryotic domains, including a unique class of putative bacterial Rab GTPases. We confirmed experimentally that a subset of these proteins translocate into the host cell upon infection. We conclude that the T4SS is highly conserved at the sequence level, but the effector proteins secreted are highly diverse.

35

1 Results and discussion

2 The Legionella genus genome is dynamic and characterized by frequent genetic 3 exchange. We sequenced 58 Legionella species of which 16 were newly sequenced, and 4 analyzed them in combination with all publicly available genomes (80 genomes in total) (SI 5 Appendix, Table S1). The *Legionella* genomes were extremely diverse, as the genome size varied from 2.37Mb (L. adelaidensis) to 4.88Mb (L. santicrucis), the GC content from 6 7 34.82% (L. busanensis) to 50.93% (L. geestiana) and the number of clusters of orthologous 8 genes as defined with OrthoMCL was 17,992 of which 5,832 (32%) were strain specific 9 (singletons) (Fig. 1A). Only 1,008 genes (6%) constituted the core genome (Fig. 1B), compared to an earlier analysis of 38 Legionella species, which found 16,416 clusters of 10 11 orthologous and 1,054 core genes (14). The addition of 40 new genomes comprising 16 newly 12 sequenced *Legionella* species in our study increased the number of orthologous gene clusters 13 by over 1,576 and decreased the core genome by 46 genes, underlining the high diversity of 14 the Legionella genus. This difference suggested that the Legionella genus pan-genome is far from fully described and that sequencing of additional Legionella species will increase the 15 16 genus gene repertoire significantly. This was supported by the rarefaction curve that does not 17 reach a plateau (Fig. 1C).

The highly dynamic nature of these genomes is also seen in the analysis of the strain 18 19 specific genes and the accessory genome as it highlights the presence of several mobile 20 genetic elements; often associated with genes encoding for transfer regions/conjugative 21 elements such as the type IVA secretion systems (T4ASS). These T4ASSs (classified as 22 T4SSF, G, I and T (17) are present in each strain to varying degrees indicating that they 23 circulate among the different Legionella strains (SI Appendix, Table S2) and therefore drive 24 genome dynamics and diversification. It has been suggested that the incorporation of foreign 25 DNA via horizontal gene transfer (HGT) is responsible for an increase in the AT content and 26 the increase in genome size (18). Indeed, we found a negative correlation between the genome 27 size and the GC content for the Legionella genomes, which also suggests frequent HGT (Fig. 28 **1D**) (19). Despite the importance of flagella for transmission to new hosts as shown for L. 29 pneumophila, flagella encoding genes were not conserved in all species, but showed a patchy 30 distribution, as 23 of the 80 strains analyzed lacked flagella genes (SI Appendix, Fig. S1). The analyses showed that the Legionella genus genome is highly diverse, dynamic and 31 32 shaped by HGT.

33

The genus *Legionella* encodes proteins with 137 different eukaryotic domains. Interpro scan analysis of all 58 *Legionella* species revealed the presence of 137 different eukaryotic

1 motifs/domains in the genus Legionella (SI Appendix, Table S3) according to the definition that an eukaryotic domain is one that is found in >75% of eukaryotic genomes and <25% in 2 3 prokaryotic genomes. The most abundant eukaryotic domains identified were ankyrin repeats. Interestingly, L. santicrucis and L. massiliensis encoded 41 and 39 ankyrin domains, 4 5 respectively (Fig. 2). Ankyrin motifs were found frequently associated with other eukaryotic motifs and thus constituted modular proteins associated with eukaryotic F-box, U-box, Rab or 6 7 SET domains. Notably, F-box and U-box domains were present in more than two thirds of the 8 species analyzed (Fig. 2) suggesting manipulation of the host ubiquitin-system is a 9 fundamental virulence strategy of Legionella species. Generally, the genomes contained one to three F-box containing proteins with the exception of L. nautarum and L. dronzanskii, 10 11 which contained 18 and 10, respectively. The SET domain containing protein RomA of L. pneumophila that induces a unique host chromatin modification (20) is present in 46 of the 12 13 58 Legionella species suggesting the ability of many Legionella species to manipulate host 14 chromatin (Fig. 2). Interestingly, the Sec-7 domain present in the effector RalF, a bacterial 15 ARF guanine exchange factor and the first described Dot/Icm effector of *L. pneumophila* (21) 16 was present in only eight (L. pneumophila, L. longbeachae, L. feelei, L. sainthelensis, L. santicrucis, L. shakespeari, L. quateirensis L. moravica) of the 58 Legionella species 17 analyzed, suggesting that, different effectors may compensate for RalF activity or that LCV 18 19 biogenesis varies among different species (Fig. 2).

One newly identified motif in Legionella was the ergosterol reductase ERG4/ERG24 20 21 (IPR001171) domain. Ergosterol is the primary sterol in the cell membranes of filamentous 22 fungi, present in membranes of yeast and mitochondria (22). Importantly, it is also the major 23 sterol of amoebae such as A. castellanii and A. polyphaga, the natural hosts of Legionella (23, 24 24). We found that 31 Legionella species encoded one or two proteins with the ERG4/ERG24 25 domain (Fig. 2). The L. longbeachae protein (L101320) containing this domain showed 56% 26 aa identity to that encoded by the amoeba Naegleria gruberi and 30% aa identity to that 27 encoded by A. castellani strain Neff. This domain was also present in other amoebae related 28 bacteria such as Parachlamydia acanthamoebae and Protochlamydia naegleriophila, as well 29 as Coxiella burnetii. Phylogenetic analyses suggested that L. longbeachae acquired this 30 domain from amoeba (SI Appendix, Fig. S2A).

Phylogenetic analyses of the here identified C-terminal alliinase and Caleosin domains
present in *L. beliardensis* and *L. anisa* or the *L. longbeachae* clade (Fig. 2), respectively
further supported acquisition of these domains from plants, amoeba or fungi (SI Appendix,
Fig. S2B-C). They probably help *Legionella* to fight competitor bacteria or fungi in amoebae
or in the environment. Taken together, our analyses highlight key domains preferentially

1 present in protozoa, fungi, plants or animals that have been acquired by different *Legionella*

- 2 species.
- 3

4 A unique case in the prokaryotic world: Legionella encode small GTPase-like domains 5 The Ras-related small GTPase superfamily comprises more than 150 members in humans, which function as key regulators of signal transduction in almost all cellular processes(25). 6 7 These enzymes bind and hydrolyse GTP to GDP and activate downstream effectors when 8 bound to GTP. The first identified member was the p21-Ras protein, an evolutionary 9 conserved small GTPase that controls cell proliferation, survival and migration through its 10 effector binding at RAF/MAPK and PI3K (26). The Ras protein superfamily is subdivided 11 into at least five distinct branches: Ras, Rho, Rab, Arf and Ran (27). Evolutionarily conserved 12 orthologs are found in Drosophila, C. elegans, S. cerevisiae, S. pombe, Dictyostelium and 13 plants (28).

14 The only Rab-like protein in a prokaryotic genome was reported in the L. longbeachae 15 genome sequence (16). However, upon analysis of our 80 Legionella strains, we identified 184 small GTPases of which 104 could be classified with a very high confidence as Rab, Ras 16 17 or Rho like proteins (34 Ras, 71 Rab and one Rho domain) (SI Appendix, Table S4 and Fig. S3). Blastp analysis of these proteins in the NCBI database revealed that 149 of the 184 small 18 19 GTPases of Legionella were exclusively present in Legionella and eukarvotic organisms 20 (Table 1). The Rab domain was localized to different parts of the effector proteins, and a 21 subset of Rab proteins carried additional domains such as U-box domains, ankyrin motifs or 22 F-box domains (Fig. 3A). Alignment of the different Rab domains identified in the Legionella 23 genomes revealed that the structural features of eukaryotic Rab domains were conserved 24 among the Legionella proteins (SI Appendix, Fig. S4).

25 To analyze further the evolutionary history of the Ras-related domains in Legionella 26 we undertook phylogenetic analyses of these proteins. For example, the two L. longbeachae 27 Rab proteins, Llo1716 and Llo3288, were present in all strains closely related to 28 L. longbeachae, suggesting that they and their orthologous share a common origin and 29 evolved from a gene acquired by the ancestor of all these species (SI Appendix, Fig. S5). 30 Further phylogenetic analysis of 16 Rab proteins present in eight different Legionella species showed that these Rab domains were acquired by HGT, mainly from protozoa (Fig. 3B and 31 SI Appendix, Fig. S6A-P). Recently a novel isoform of Rab5D was identified in the 32 33 Acanthamoeba polyphaga mimivirus (APMV) and all group I members of the Mimiviridae 34 (29). Phylogenetic analyses suggested that the Rab GTPase was acquired by an ancestor of 35 the Mimiviridae family and Rabs from Mimiviridae, Plasmodium and few lower eukaryotes

1 form a separate clade (29). Thus, Legionella and APMV that both infect the protozoa Acanthamoeba encode Rab proteins most likely to mimic and subvert host cell function. To 2 3 substantiate that these proteins act in the host cell, we determined whether the Rab containing proteins were bona fide substrates of the Dot/Icm T4SS by creating fusion proteins between 4 5 the 16 different Rab proteins and the catalytic domain of the TEM-1 beta-lactamase (indicated by a star in SI Appendix, Fig. S5). Translocation assays were performed using wild type 6 7 L. pneumophila as a surrogate host and compared with an isogenic Dot/Icm mutant ($\Delta dotA$). 8 All 16 Rab motif-containing proteins were translocated by L. pneumophila but not by the 9 $\Delta dotA$ mutant (Fig. 3C-D).

10

11 More than 250 different eukaryotic like proteins are encoded in Legionella genomes. In 12 addition to modular effectors with eukaryotic domains, the Legionella genome encodes 13 proteins that are similar to eukaryotic proteins, many of which are proven effectors of the 14 Dot/Icm T4SS. A wider search for eukaryotic like proteins in the Legionella genus identified 15 2196 eukaryotic like proteins representing more than 400 different orthologous groups that 16 matched better to eukaryotes than to prokaryotes from a total of 6809 different orthologous 17 proteins that matched with eukaryotic proteins. Among these, we identified 156 proteins with a eukaryotic domain, and 210 new eukaryotic-like proteins (SI Appendix, Table S5). 18 19 Furthermore, 152 eukarvotic like proteins detected possess a higher GC content (40%-62%) 20 than the rest of the genome indicating recent HGT. Phylogenetic analysis of selected, newly 21 identified proteins suggested that these were acquired from eukaryotes. As an example, SI 22 Appendix, Fig. S7 shows the protein LanA0735 from Legionella anisa, a species frequently 23 found in artificial water systems. This protein belongs to the pyridine nucleotide-disulfide oxidoreductase family, a subfamily of the FAD dependent oxidoreductase family. LanA0735 24 25 showed some similarity to thioredoxin reductase that exists as two major ubiquitous 26 isoenzymes in higher eukaryotic cells, one cytosolic and the other one mitochondrial. The 27 cytosolic form has been implicated in interference with the acidification of the lysosomal 28 compartment in C. elegans (30), and thus LanA0735 may help Legionella avoid vacuole 29 acidification during infection.

Among the proteins defined as eukaryotic like, two previously described phospholipases of *L. pneumophila*, PlcB (Lpp1411/Lpg1455) and PlcA (Lpp0565/Lpg0502) were identified in our analysis as eukaryotic proteins. The only other bacteria encoding these two enzymes are *Pseudomonas* and amoebae-associated bacteria. The two enzymes have phospholipase activity (31), but their role in infection is unknown. Here they were predicted as phosphatidylcholine-hydrolyzing phospholipase C. Phosphatidylcholine is a eukaryotic 1 membrane phospholipid that is present in only about 15% of prokaryotic species, in particular 2 bacteria interacting with eukaryotes (32). L. pneumophila belongs to the phosphatidylcholine-3 containing group of bacteria, which includes Francisella tulurensis or Brucella abortus (33). These pathogens use the phosphatidylcholine synthase pathway exclusively for 4 5 phosphatidylcholine formation and are thought to depend on choline supplied from the host cell (34). Indeed, it has been shown that phosphatidylcholin synthesis is required for 6 7 L. pneumophila virulence (35). Thus, it is tempting to infer that the role of these enzymes may 8 be to help acquire choline from the host cell.

9

10 **Evolutionary history of eukaryotic domains and eukaryotic proteins.** It is intriguing that 11 Legionella species encode such a diverse repertoire of eukaryotic domains and eukaryotic-like 12 proteins. To understand better this unique feature of the genus we analyzed the evolutionary 13 history of these proteins. After phylogenetic reconstruction of the genus Legionella based on 14 the core genome (at least 50% identical) (Fig. 1A), we analyzed the distribution of the 15 eukaryotic motifs and the eukaryotic proteins with respect to the evolution of the genus. For 16 most we found patchy distribution, as the repertoire of these proteins is variable among the 17 different Legionella species (Fig. 2). Such a distribution is indicative of gain and loss events during the evolution of the genus. To analyze further how these proteins may have evolved in 18 Legionella we selected 25 eukarvotic motifs representing 2,837 different proteins in over 800 19 20 orthologous groups and used the program Gloome to analyze the gain and loss events for 21 these proteins. We found that the number of gain events (1,197/69%) considerably exceeded 22 the number of loss events (549/31%), a bias that was even stronger when using parsimony 23 (1,628 gain events versus 89 loss events) (SI Appendix, Fig. S8). These results were 24 confirmed also when using a more conservative approach by taking a probability cut-off for 25 the stochastic model of 0.8 instead of 0.5, and when analyzing each motif separately.

26 An exemplary view of this result is shown for four proteins encoding different motifs (U-box and ankyrin repeat, SET domain and ankyrin repeat, astacin domain and allinase 27 28 domain; Fig. 4). Loss events are indicated by a star and gain events by a dot. The number of 29 gain events exceeds the number of loss events, indicating that in the Legionella genus gene 30 acquisition is dominant. Moreover, gene acquisition seems to be an on-going and frequent process in the genus Legionella given the high number of events we observed and the fact that 31 most of them are localized in the terminal branches of the tree (SI Appendix, Fig. S8). To 32 33 analyse if eukaryotic-like proteins have the same evolutionary history, we took the sphingosine1-phosphate lyse (LpSpl) (36, 37) as an example. Indeed, when running the same 34

analyses this gene also appeared to have been gained multiple times during the evolution of
the genus (Fig. 4).

3 Thus, in comparison to most prokaryotic species analysed to date, more gene gain events are evident than loss events during evolution of the Legionella genus, which is also 4 5 corroborated by the fact that the ancestral genomes were probably smaller (Fig. 1A, cluster I). Indeed, as seen in Fig. 1A, in each of the defined phylogenetic clusters only few genomes 6 7 have a larger size *e.g.* in cluster II *L. massiliensis* is the only species with a big genome, thus 8 the most parsimonious explanation is that the ancestor of this clade had a small genome and 9 in the branch leading to L. massiliensis gene gain occurred. This finding is similar to what was described for the adaptation of louse-borne intracellular pathogens and amoeba 10 11 associated bacteria. It is well known that the specialization of intracellular bacteria is 12 associated with genome reduction, and extreme genome reduction cane be seen in louse-borne 13 human specialists. In contrast, nonspecialized intra-amoebal microorganisms exhibit a 14 genome larger than their relatives due to gene conservation and acquisition (38).

15

16 The Dot/Icm secretion system is a highly conserved machinery secreting thousands of 17 different proteins. The Dot/Icm T4SS is indispensable for intracellular replication of 18 L. pneumophila in both amoeba and macrophages (39). In stark contrast to the high genetic 19 diversity observed in the *Legionella* genomes, the Dot/Icm T4SS is part of the core genome as 20 it is present in all species analyzed and the organization of the constituent proteins is highly 21 conserved, even at the amino acid level. The proteins comprising the secretion machinery 22 show an average amino acid identity of more than 50% and some even more than 90% when 23 compared to the L. pneumophila Dot/Icm components (SI Appendix Fig. S9A and Table S6). The most conserved proteins are DotB, a secretion ATPase (86-100% aa identity) and IcmS, a 24 25 small acidic cytoplasmic protein (74-98% aa identity). This high conservation is even seen 26 with one of the few non-Legionella species that encode a Dot/Icm system, Coxiella burnetti.

The only gene of the Dot/Icm system that is not present in all Legionella species is 27 28 icmR. IcmR interacts with IcmQ as a chaperone preventing IcmQ self-dimerization (40). 29 Although IcmQ is highly conserved, the gene encoding IcmR is frequently replaced by one or 30 two non-homologous genes encoding for proteins that are called FIR because they can 31 functionally replace IcmR (41). When overlapping the occurrence of the different FIR genes 32 with the phylogeny of the species, most phylogenetically closely related species share homologous FIR genes (SI Appendix, Fig. S10). Apart from two conserved regions (SI 33 34 Appendix, Fig. S11), the absence of sequence homology among FIR proteins indicates that 35 *icm*R is an extremely fast evolving gene and therefore probably under positive selection. The reason why this gene is extremely divergent is still unknown but could be also linked to the
 high variety of Dot/Icm effectors described in this genus. Thus, except for the FIR genes, the
 Dot/Icm T4SS is highly conserved and encoded in a very dynamic genetic context.

It has been shown previously, that the more than 300 substrates of the L. pneumophila 4 5 Dot/Icm system are not universally present within the genus Legionella as among 38 Legionella species only seven core effectors had been described (14). Surprisingly, when 6 7 adding the 40 additional genomes and 16 new Legionella species sequenced in this study, we 8 identified 8 core effectors instead of seven. A comparison of the two studies confirmed 9 Lpg0103 (VipF), Lpg0107 (RavC), Lpg2300 (LegA3/AnkH/AnkW), and Lpg2815 (IroT/MavN) as core substrates (14) (SI Appendix, Fig. S9B and Table S7). Three of the 10 11 previously defined core substrates (Lpg0140, Lpg2832, Lpg3000) were present in two 12 genomes as two consecutive genes instead of one, however, this fragmentation might be a 13 sequencing error, and thus we considered these substrates also as core substrates (SI 14 Appendix Table S7). In our study we identified one additional core effector gene, 15 *lpg1356/lpp1310*. This protein has been reported by Lifshitz and colleagues (42) as secreted protein, but had not been included in the Burstein effector search, which explains the different 16 17 result (SI Appendix, Fig. S9B and Table S7). Similarly, to most of the other core substrates, their functions are not known, but Lpg1356 encodes eight eukaryotic Sel-1 motifs similar to 18 19 LpnE, a *L. pneumophila* virulence determinant that influences vacuolar trafficking (43). 20 Furthermore, seven other genes are present in all but one, two or four genomes, thus they 21 might have important functions in host pathogen interactions (SI Appendix Table S7). 22 Interestingly, when the effector repertoire of several strains of one species is compared the 23 conservation of the effectors is very high (between 82 and 97%) (SI Appendix Table S8). 24 However, if more strains than two are available for a species as it is the case for 25 L. pneumophila where 11 strains could be compared, the conservation of the effector pool is 26 only 65% (264 of the 408 different effectors identified in the 11 strains) (SI Appendix Table 27 **S8**). Thus the *L. pneumophila* core effector set is also smaller than previously thought. Taken 28 together, the genus Legionella has 8 core substrates present in all genomes and seven 29 additional ones that are present in nearly all genomes.

Interestingly, whereas the number of core Dot/Icm substrates is extremely small, the number and the diversity of predicted Dot/Icm substrates is extremely high. Indeed, through a machine learning approach, Burstein *et al* predicted that the *Legionella* genus would encode 5,885 effectors (14). Here we extended these analyses and identified 4,767 proteins with eukaryotic motifs that have a high probability to be secreted effectors as shown for the Rablike proteins. If we consider that the orthologous of these proteins in each species are also

1 effectors then the number raises to 7103 (representing 1145 different orthologous proteins) (SI Appendix Fig. S9C). Moreover, we identified 2,196 eukaryotic like proteins representing 2 3 414 different orthologous genes, which form together with the above-mentioned eukaryotic motif carrying proteins 1,400 different putative orthologous substrates of the Dot/Icm T4SS. 4 5 Finally, when adding to the effectors predicted in this study (based on their similarity to eukaryotic domains and proteins), the effectors previously described in L. pneumophila and 6 7 their orthologues (more than 7000 proteins representing about 300 different orthologous), as 8 well as the effectors predicted by the machine learning approach and their orthologous (more 9 than 10 000 proteins representing about 900 different orthologous) (14) the total number of different effectors rises to almost 18.000 proteins (more than 1.600 orthologous groups) (SI 10 Appendix, Table S9 and Fig. S9C). Therefore, the Legionella genus has by far the highest 11 12 number and widest variety of effectors described for an intracellular bacterium. Furthermore, 13 when calculating the growth accumulation curve for Dot/Icm predicted effectors, this number 14 should still increase with the sequencing of new Legionella genomes, as the plateau is not 15 reached yet (SI Appendix, Fig. S9D).

16

17 The ability to infect human cells has been acquired independently several times during the evolution of the genus Legionella. Among the 65 Legionella species known, 18 19 L. pneumophila is responsible for over 90% of human disease, followed by L. longbeachae 20 (2-7% of cases, except Australia and New Zealand with 30% (44)). Certain Legionella species 21 such as L. micdadei, L. dumoffii or L. bozemanii have once or sporadically been associated 22 with human disease (44), and all other species seem to be environmental bacteria only. The reasons for these differences are not known. To explore whether all species are able to 23 24 replicate in human cells we chose the human macrophage like cell line THP-1 as model and 25 tested the replication capacity of 47 different Legionella species. Infections were carried out 26 in duplicates or triplicates and colony-forming units were recorded at 24h, 48h and 72h post 27 infection. Levels of intracellular replication were compared to wild type L. pneumophila 28 strain Paris and an isogenic non-replicating $\Delta dotA$ mutant as reference strains (Fig. 5 and SI 29 Appendix, Fig. S12 and S13). Results were also compared to data previously reported for 30 different Legionella species in THP-1, U937 and A549 cells, Mono Mac 6, mouse and guinea pig derived macrophages, or in guinea pigs (SI Appendix, Table S10). When results at 72 h 31 after infection were analyzed, 28 of the 47 species tested were impaired for intracellular 32 33 replication whereas nine species replicated similarly to L. pneumophila Paris or better (Fig. 5). These nine species were L. gormanii, L. jamestowniensis, L. jordanis, L. like brunensis, 34 35 L. maceachernii, L. micdadei, L. nagasikiensis L. parisiensis, and L. tucsonensis. Interestingly,

1 L. jamestowniensis, for which one human case has been reported (45), replicated better than 2 L. pneumophila Paris. Indeed, L. jamestowniensis productively infects human U937-derived 3 phagocytes. The remaining eight species showed variable replication patterns being 4 significantly different from L. pneumophila Paris only in one or two of the three analyzed 5 time points (SI Appendix, Fig. S12). Broadly, the species most frequently reported from human disease (L. pneumophila, L. longbeachae, L. micdadei, L. bozemanii and L. dumoffii) 6 7 are also those that replicated robustly in THP-1 cells. The only exception was the L. dumoffii 8 strains that were impaired for replication in THP-1 cells but which have been shown to 9 replicate in other cell types and guinea pigs. Taken together, there is a convincing correlation between the frequency of isolation from human disease and the ability to grow in 10 11 macrophage-like cells.

To analyze this further, we overlapped the replication results with the phylogeny of 12 13 the genus. Apart from the small cluster containing L. beliardensis, L. gresilensis and 14 L. busanensis, which were all unable to grow in THP-1 cells, replicating and non-replicating 15 strains were mixed in the phylogeny (SI Appendix, Fig. S14). This suggests that the capacity to replicate in human cells has been acquired independently several times during evolution of 16 17 the Legionella genus, possibly as a result of recruiting effectors that allow adaptation to particular niches. To understand whether a specific set of effectors is necessary to infect 18 19 human cells, we further analyzed the combination of effectors present in the strains isolated 20 from human disease and effectors present in strains capable of replicating in THP-1 cells. 21 Surprisingly, no specific set of effectors could be attributed to strains capable of replicating in 22 human cells or isolated from human disease, although among these strains certain conserved 23 motifs always present were identified, such as ankyrin motifs, F-box or SET-domains, 24 suggesting that common pathways need to be subverted to cause human infection. Thus, the 25 capacity to infect human cells has been acquired independently, several times during the 26 evolution of the genus Legionella.

27 In conclusion, the analysis of 80 Legionella strains representing 58 different 28 Legionella species has revealed a contrasting picture of the Legionella genus. It encodes a 29 highly conserved T4SS predicted to secrete more than 18,000 proteins, of which only 8 are 30 conserved throughout the genus. Together the genomes portray an extremely diverse genus 31 shaped by massive inter-domain horizontal gene transfer, circulating mobile genetic elements 32 and eukaryotic like proteins. Our in-depth analyses of eukaryotic features of the Legionella 33 genomes identified 137 different eukaryotic domains of which Rab or Ras domain-containing 34 proteins were quasi unique to the genus Legionella. The secretion assays undertaken for 16 of 35 these Rab or Ras domain-containing proteins confirmed that these were translocated Dot/Icm

1 effectors. In addition to the eukaryotic domains, we identified 210 orthologous groups of 2 eukaryotic like proteins. If all these proteins in the different species and their orthologues are 3 taken into account, we found more than 8,000 proteins that have been shaped by inter-domain horizontal gene transfer in the genus Legionella. Thus, to our knowledge the genus Legionella 4 5 contains the widest variety and highest number of eukaryotic proteins and domains of any prokaryotic genus genome analyzed to date. Analyzing more strains per species will probably 6 7 discover new unknown effectors increasing our knowledge of the set of tools used by 8 Legionella to infect eukaryotic cells. Although eukaryotic proteins and domains were a 9 universal feature of the genus *Legionella*, the repertoire of these proteins for each species was different. Surprisingly, even when the same motif was present in different species, these were 10 11 often present in different proteins with no orthology. In accordance with this finding, our evolutionary analysis of the presence/absence of these domains and proteins suggested that 12 13 these proteins were mostly acquired through gene gain events.

14 When exploring the replication capacity of 47 different Legionella species in human 15 macrophage-like cell line THP-1, we found that the 23 species were capable of replicating in THP-1 cells. However, these did not cluster in the phylogeny, indicating that the capacity to 16 17 replicate in macrophages can be achieved by different combinations of effectors, and this capacity has been acquired several times during the evolution of the Legionella genus. As 18 19 humans are an accidental host for *Legionella*, the capacity to replicate in macrophages may 20 also have been obtained by a coincidental acquisition of different virulence properties initially 21 needed to adapt to a specific natural host, such as amoebae. Indeed, due to the high 22 conservation of key signaling pathways in professional phagocytes such as amoebae and 23 human macrophages, different combinations of effectors may allow Legionella species to 24 infect higher eukaryotic cells by chance.

Here we show that all *Legionella* species have acquired eukaryotic proteins that likely modulate specific host functions to allow intracellular survival and replication in eukaryotic host cells. At a certain point, the evolution of a combination of effector proteins that allow replication in human cells may inadvertently lead to the emergence of new human pathogens from environmental bacteria.

30

31 Material and Methods

The materials and methods are described at length in *SI Appendix*. This includes: Sequencing and assembly, sequence processing and annotation, pan/core genome, ortholog and singleton definition, phylogenetic reconstruction and evolutionary analysis, phylogenetic analyses of Rab and eukaryotic-like proteins, infection assays, statistical analysis, and translocation 1 assays. The raw sequence reads were deposited in the European Nucleotide Archive (study

2 accession number PRJEB24896). The sequences and annotations can be accessed through:

3 https://github.com/bbi-ip/Legionella_genus_proteins.git

4

5 Acknowledgements

We would like to thank Tim P. Stinear for critical reading of the manuscript and helpful
comments and we acknowledge the receipt of 53 different *Legionella* strains from the
Collection of the Institut Pasteur (CIP). Work in the CB laboratory is financed by the Institut
Pasteur, the grant n°ANR-10-LABX-62-IBEID and the Fondation pour la Recherche
Médicale (FRM) grant N° DEQ20120323697.

11

12 Author contributions

- 13 SJ and LGV, ELH contributed to sample collection and strain analyses DNA extraction and
- 14 sequencing; CR, DC, SM, AEPC, MR, SP, SR, JD, JC, SDD, GNS to functional experiments,
- 15 data analyses and interpretation. The manuscript was written by LGV and CB with input from
- 16 co-authors. The project was conceived, planned and supervised by LGV, GD, GF and CB.
- 17

18 References

- Fraser DW, *et al.* (1977) Legionnaires' disease: description of an epidemic of pneumonia.
 N Engl J Med 297(22):1189-1197.
- Rowbotham TJ (1980) Preliminary report on the pathogenicity of *Legionella pneumophila* for freshwater and soil amoebae. *J Clin Pathol* 33(12):1179-1183.
- 23 3. Cazalet C, *et al.* (2004) Evidence in the *Legionella pneumophila* genome for exploitation
 24 of host cell functions and high genome plasticity. *Nat Genet* 36(11):1165-1173.
- 4. Bruggemann H, Cazalet C, & Buchrieser C (2006) Adaptation of *Legionella pneumophila* to the host environment: role of protein secretion, effectors and eukaryotic-like proteins. *Curr Opin Microbiol* 9(1):86-94.
- 5. Komano T, Yoshida T, Narahara K, & Furuya N (2000) The transfer region of IncI1
 plasmid R64: similarities between R64 tra and *Legionella icm/dot* genes. *Mol Microbiol*35(6):1348-1359.
- 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.
- 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.
- 8. Nora T, Lomma M, Gomez-Valero L, & Buchrieser C (2009) Molecular mimicry: an
 important virulence strategy employed by *Legionella pneumophila* to subvert host
 functions. *Future Microbiol* 4:691-701.
- 38 9. Burstein D, *et al.* (2009) Genome-scale identification of *Legionella pneumophila*39 effectors using a machine learning approach. *PLoS Pathog* 5(7):e1000508.
- 40 10. Gomez-Valero L, *et al.* (2011) Extensive recombination events and horizontal gene
 41 transfer shaped the *Legionella pneumophila* genomes. *BMC Genomics* 12:536.
- 42 11. Gomez-Valero L, *et al.* (2014) Comparative analyses of Legionella species identifies
 43 genetic features of strains causing Legionnaires' disease. *Genome Biol* 15(11):505.

- Morozova I, et al. (2004) Comparative sequence analysis of the *icm/dot* genes in Legionella. Plasmid 51(2):127-147.
- 3 13. Sanchez-Buso L, Comas I, Jorques G, & Gonzalez-Candelas F (2014) Recombination
 4 drives genome evolution in outbreak-related *Legionella pneumophila* isolates. *Nat Genet* 5 46(11):1205-1211.
- Burstein D, *et al.* (2016) Genomic analysis of 38 *Legionella* species identifies large and diverse effector repertoires. *Nat Genet* 48(2):167-175.
- 8 15. Joseph SJ, *et al.* (2016) Dynamics of genome change among *Legionella* species. *Sci Rep*9 6:33442.
- 10 16. Cazalet C, et al. (2010) Analysis of the Legionella longbeachae genome and
 11 transcriptome uncovers unique strategies to cause Legionnaires' disease. PLoS Genet
 12 6(2):e1000851.
- 13 17. Guglielmini J, de la Cruz F, & Rocha EP (2013) Evolution of conjugation and type IV
 14 secretion systems. *Mol Biol Evol* 30(2):315-331.
- 15 18. Bohlin J, Brynildsrud OB, Sekse C, & Snipen L (2014) An evolutionary analysis of
 genome expansion and pathogenicity in *Escherichia coli*. *BMC Genomics* 15:882.
- 17 19. Bohlin J, Sekse C, Skjerve E, & Brynildsrud O (2014) Positive correlations between
 18 genomic %AT and genome size within strains of bacterial species. *Environ Microbiol* 19 *Rep* 6(3):278-286.
- 20. Rolando M, *et al.* (2013) *Legionella pneumophila* effector RomA uniquely modifies host
 21 chromatin to repress gene expression and promote intracellular bacterial replication. *Cell* 22 *Host Microbe* 13(4):395-405.
- 23 21. Nagai H, Kagan JC, Zhu X, Kahn RA, & Roy CR (2002) A bacterial guanine nucleotide
 24 exchange factor activates ARF on *Legionella* phagosomes. *Science* 295(5555):679-682.
- 25 22. Pasanen AL, Yli-Pietila K, Pasanen P, Kalliokoski P, & Tarhanen J (1999) Ergosterol
 26 content in various fungal species and biocontaminated building materials. *Appl Environ* 27 *Microbiol* 65(1):138-142.
- 28 23. Smith FR & Korn ED (1968) 7-Dehydrostigmasterol and ergosterol: the major sterols of
 29 an amoeba. *J Lipid Res* 9(4):405-408.
- 24. Thomson S, *et al.* (2017) Characterisation of sterol biosynthesis and validation of
 14alpha-demethylase as a drug target in *Acanthamoeba*. Sci Rep 7(1):8247.
- 32 25. Wennerberg K, Rossman KL, & Der CJ (2005) The Ras superfamily at a glance. *J Cell* 33 *Sci* 118(Pt 5):843-846.
- Simanshu DK, Nissley DV, & McCormick F (2017) RAS Proteins and Their Regulators
 in Human Disease. *Cell* 170(1):17-33.
- 27. Rojas AM, Fuentes G, Rausell A, & Valencia A (2012) The Ras protein superfamily:
 evolutionary tree and role of conserved amino acids. *J Cell Biol* 196(2):189-201.
- 28. Colicelli J (2004) Human RAS superfamily proteins and related GTPases. *Sci STKE* 2004(250):RE13.
- 29. Zade A, Sengupta M, & Kondabagil K (2015) Extensive *in silico* analysis of Mimivirus
 coded Rab GTPase homolog suggests a possible role in virion membrane biogenesis. *Front Microbiol* 6:929.
- 43 30. Li W, et al. (2012) Two thioredoxin reductases, trxr-1 and trxr-2, have differential
 44 physiological roles in *Caenorhabditis elegans*. *Mol Cells* 34(2):209-218.
- 45 31. Hiller M, Lang C, Michel W, & Flieger A (2017) Secreted phospholipases of the lung
 46 pathogen *Legionella pneumophila*. *Int J Med Microbiol*.
- 47 32. Aktas M, *et al.* (2010) Phosphatidylcholine biosynthesis and its significance in bacteria
 48 interacting with eukaryotic cells. *Eur J Cell Biol* 89(12):888-894.
- 49 33. Geiger O, Lopez-Lara IM, & Sohlenkamp C (2013) Phosphatidylcholine biosynthesis
 50 and function in bacteria. *Biochim Biophys Acta* 1831(3):503-513.
- S1 34. Comerci DJ, Altabe S, de Mendoza D, & Ugalde RA (2006) *Brucella abortus* synthesizes
 phosphatidylcholine from choline provided by the host. *J Bacteriol* 188(5):1929-1934.

- S. Conover GM, *et al.* (2008) Phosphatidylcholine synthesis is required for optimal function
 of Legionella pneumophila virulence determinants. *Cell Microbiol* 10(2):514-528.
- 36. Degtyar E, Zusman T, Ehrlich M, & Segal G (2009) A *Legionella* effector acquired from
 protozoa is involved in sphingolipids metabolism and is targeted to the host cell
 mitochondria. *Cell Microbiol* 11(8):1219-1235.
- 37. Rolando M, *et al.* (2016) Legionella pneumophila S1P-lyase targets host sphingolipid
 metabolism and restrains autophagy. *Proc Natl Acad Sci U S A* 113(7):1901-1906.
- 8 38. Moliner C, Fournier PE, & Raoult D (2010) Genome analysis of microorganisms living
 9 in amoebae reveals a melting pot of evolution. *FEMS Microbiol Rev* 34(3):281-294.
- 39. Segal G & Shuman HA (1999) Legionella pneumophila utilizes the same genes to
 multiply within Acanthamoeba castellanii and human macrophages. Infect Immun
 67(5):2117-2124.
- 40. 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.
- 41. 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.
- 42. 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
 USA 110(8):E707-715.
- 43. Newton HJ, *et al.* (2007) Sel1 repeat protein LpnE is a *Legionella pneumophila* virulence
 determinant that influences vacuolar trafficking. *Infect Immun* 75(12):5575-5585.
- 44. Yu VL, *et al.* (2002) Distribution of *Legionella* species and serogroups isolated by
 culture in patients with sporadic community-acquired legionellosis: an international
 collaborative survey. *J Infect Dis* 186(1):127-128.
- 45. Prochazka B, *et al.* (2016) Draft Genome Sequence of *Legionella jamestowniensis*Isolated from a Patient with Chronic Respiratory Disease. *Genome announcements* 4(5).
- 29
- 30 31

1 Figure 1: The Legionella genomes are diverse in size and gene content. A) Phylogeny of 2 the genus based on the core genome, genome size, GC content and number of singletons of 3 each species are depicted. Numbers represent bootstrap values. Branches are coloured according to the clade they belong to. Genome size and GC content include plasmids if 4 5 present in the corresponding species. The number of singletons is based on the results of OrhtoMCL (takes into account orthologs and paralogs). Each species has been compared to 6 7 the others without taking into account strains from the same species to avoid bias due to the 8 number of strains sequenced within a species. B) Occurrence of genes within the 80 analysed 9 Legionella genomes. Left end of the x-axis, genes present in a single genome (strain specific genes: 5832 \approx 32% of the pangenome); right end of the x-axis, genes present in all 80 10 11 genomes (core-genome; 1008 genes $\approx 6\%$ of the pan-genome) C) Gene accumulation curve 12 for the total number of proteins of the 80 genomes. **D**) Negative correlation between genome 13 size and GC content indicating high acquisition of foreign genes (Pearson's correlation 14 coefficient equal to -0.46 with a p-value<0.0001)

15

16 Figure 2: Eukaryotic domains have a diverse distribution within the genus Legionella 17 suggesting multiple acquisition events. The number and distribution of the 41 most frequently identified eukaryotic motifs within the genus Legionella are shown. Numbers 18 19 represent the number of proteins containing this eukarvotic motif. Abbreviations used: ANK 20 (ankyrin), F-box, U-box), SET domain, Pkinases (protein kinases), Sec-7 domain, LLR 21 (leucine rich repeats), Miro (Mitochondrial Rho domain), TTL (tubulin-tyrosine ligase), SH2 22 (The Src homology 2), PAM2 (ataxin-2, C-terminal), PPR (pentatricopeptide repeat), I-set 23 (immunoglobulin I-set), NP (nucleoside phosphatase gda1/cd39), HAD (HAD-superfamily 24 hydrolase), DH (Dbl homology domain), Mit. Substrate (mitochondrial substrate/solute 25 Rho **GTPases-activating** protein domain, T-complex (T-complex10/11), carrier), 26 PC65 (Peptidase C65 otubain), Ergosterol (Ergosterol biosynthesis), Flavin (flavin 27 monooxygenase-like), Astacin (Peptidase M12A, astacin), Cyt:P450 (Cytochrome P450), 28 Cytokine FAD (Cytokinin dehydrogenase 1, FAD/cytokinin binding domain), PQ loop repeat, 29 Peptidase C2 (calpain, catalytic domain), LR glioma (Leucine-rich glioma-inactivated, EPTP 30 repeat), Ovarian (Ovarian tumour, otubain), Papain (Peptidase C1A, papain C-terminal, DOT1 (Histone methylation DOT1), Rab small GTPases, DUF155, C/C (Clathrin/coatomer 31 32 adaptor, adaptin-like), RCC1 (Regulator of chromosome condensation). 33

Figure 3: Domain organization of small GTPases in *Legionella* and phylogenetic
 analyses of the Llo3288 Rab proteins suggests eukaryotic origin. A) Domain organization

1 of the different small GTPases proteins identified. B) Unrooted tree of Llo3288 and 2 homologues recruited by blastp constructed using likelihood. Local support values are 3 represented with circles on the corresponding branches and size of circles is proportional to the values (only local support of at least 0.7 are shown). C) Transloctation of selected proteins 4 5 using the beta-lactamase transloctation assay and infection of Raw264.7 cells for 1h with Lp wild type or $Lp \Delta dotA$ expressing BlaM-effector fusions analysed with a microplate reader. 6 7 Three independent experiments (n=9) were done. Statistical significance was determined by 2-way Anova with multiple comparisons test (*, P < 0.05; **, P < 0.01; ***, P < 0.001. **D**) 8 9 Transloctaion of selected proteins using the beta-lactamase transloctaion assay and infection of THP-1 cells at an MOI of 50 during 1h 30min with Lp and Llo strains in before addition of 10 CCF4-AM and analyses by flow cytrometry. Histograms show the frequency of BlaM-11 translocated, blue fluorescence-emitting cells as means ± SD of three independent 12 experiments (n=12). Statistical significance was determined by Wilcoxon matched pairs test 13 (**, P<0.01; ***, P<0.001).). Lp, L. pneumophila wild type; Llo, L. longbeachae wild type; 14 *Lp* $\Delta dotA$, *L*. *pneumophila* $\Delta dotA$; *Llo* Δ , *L*. *longbeachae* $\Delta dotA$. 15

16

Figure 4. Gain/loss prediction for selected eukaryotic proteins and domain containing
proteins. Circles on the branches represent gain events, crosses loss events. The full squares,
circles, triangles or stars indicate the presence of the respective protein; the empty squares,
circles, triangles or stars indicate that the protein is absent in this species.

21

22 Figure 5: The replicative capacity of the different Legionella species in THP-1 cells 23 correlates with their epidemiological features. Replication of each strain at the time point 24 72h after infection of THP-1 cells is shown (24h and 48h of infection are shown in SI 25 Appendix, Fig. S14. Intracellular replication was determined by recording the number of 26 colony-forming units (CFU) after plating on BCYE agar. L. pneumophila Paris, representative of a replicating strain (blue box); L. pneumophila $\Delta dotA$, representative of non-replicating 27 28 strain (red box). The strains are ordered according to the mean replication values. A) 29 Legionella species replicating like or significantly better than L. pneumophila Paris. B) 30 Species with no or significantly lower replication capacities than L. pneumophila Paris.

31 32



0.06







