nature genetics
OPEN

Genomic analysis of 38 *Legionella* species identifies large and diverse effector repertoires

David Burstein[1,5], Francisco Amaro[2,5], Tal Zusman[3], Ziv Lifshitz[3,5], Ofir Cohen[1,5], Jack A Gilbert[4], Tal Pupko[1], Howard A Shuman[2] & Gil Segal[3]

Infection by the human pathogen *Legionella pneumophila* relies on the translocation of ~300 virulence proteins, termed effectors, which manipulate host cell processes. However, almost no information exists regarding effectors in other *Legionella* pathogens. Here we sequenced, assembled and characterized the genomes of 38 *Legionella* species and predicted their effector repertoires using a previously validated machine learning approach. This analysis identified 5,885 predicted effectors. The effector repertoires of different *Legionella* species were found to be largely non-overlapping, and only seven core effectors were shared by all species studied. Species-specific effectors had atypically low GC content, suggesting exogenous acquisition, possibly from the natural protozoan hosts of these species. Furthermore, we detected numerous new conserved effector domains and discovered new domain combinations, which allowed the inference of as yet undescribed effector functions. The effector collection and network of domain architectures described here can serve as a roadmap for future studies of effector function and evolution.

Several bacterial pathogens, such as the agents of tuberculosis, typhus, typhoid fever, Q fever and Legionnaires' disease, use specialized secretion systems that translocate into the host's cytoplasm a cohort of proteins, termed effectors, which modulate host cell processes[1]. One such pathogen is *L. pneumophila*, the causative agent of Legionnaires' disease. These bacteria multiply in nature in a broad range of free-living amoebae[2] and cause pneumonia in humans when contaminated water aerosols are inhaled[3]. Besides *L. pneumophila*, more than 50 *Legionella* species have been identified, and at least 20 were associated with human disease[4].

The major pathogenesis system of *L. pneumophila* is composed from a group of 25 proteins called Icm/Dot (intracellular multiplication/defect in organelle trafficking), which constitute a type IVB secretion system[5,6]. Type IV secretion systems are macromolecular devices, evolutionarily related to bacterial conjugation systems, which translocate effector proteins into host cells[7]. All the *Legionella* species studied so far harbor a type IVB Icm/Dot secretion system[8], which is required for intracellular growth[9]. This secretion system is also found in *Coxiella burnetii*, the etiological agent of Q fever[10,11], and in the arthropod pathogen *Rickettsiella grylli*[12].

Thus far, approximately 300 *L. pneumophila* effectors have been experimentally shown to translocate into host cells via the Icm/Dot secretion system. Deletion of a single effector-coding gene rarely causes a detectable defect in intracellular growth[13]. This is commonly explained by functional redundancy involving, for example, multiple effectors that perform similar functions, effectors that target different steps of the same host cell pathway or effectors that manipulate parallel pathways[14]. Notably, only very few Icm/Dot effectors have been identified in other *Legionella* species. The pool of effectors for each species is believed to orchestrate its intracellular lifestyle, and effectors present in different *Legionella* species might modulate different host cell pathways through biochemical activities not mediated by *L. pneumophila* effectors.

De novo sequencing of 38 *Legionella* species allowed us to explore the diversity of effectors used by the *Legionella* genus, to study the evolution of these special proteins and to infer new potential functions mediated by them.

RESULTS

Sequencing, phylogeny and genomic characterization

We sequenced the genomes of isolates from 38 different *Legionella* species (Online Methods and **Supplementary Tables 1** and **2**). These sequences were analyzed along with three publically available *Legionella* genomes (*L. pneumophila*, *Legionella longbeachae* and *Legionella drancourtii*). Protein-coding genes were clustered into 16,416 orthologous groups, of which 1,054 were present in all 41 genomes. We designated these groups as LOGs for *Legionella* orthologous groups (**Supplementary Table 3**).

[1]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel. [2]Department of Microbiology, University of Chicago, Chicago, Illinois, USA. [3]Department of Molecular Microbiology and Biotechnology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel. [4]Biology Division, Argonne National Laboratory and Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, USA. [5]Present addresses: Department of Earth and Planetary Science, University of California Berkeley, Berkeley, California, USA (D.B.), Department of Natural Sciences. Saint Louis University–Madrid Campus, Madrid, Spain (F.A.), Division of Epidemiology and the National Center for Antibiotic Resistance, Tel Aviv Sourasky Medical Centre, Tel Aviv, Israel (Z.L.) and Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA (O.C.). Correspondence should be addressed to H.A.S. (hashuman@uchicago.edu) or G.S. (gils@tauex.tau.ac.il).

Received 22 June 2015; accepted 8 December 2015; published online 11 January 2016; doi:10.1038/ng.3481

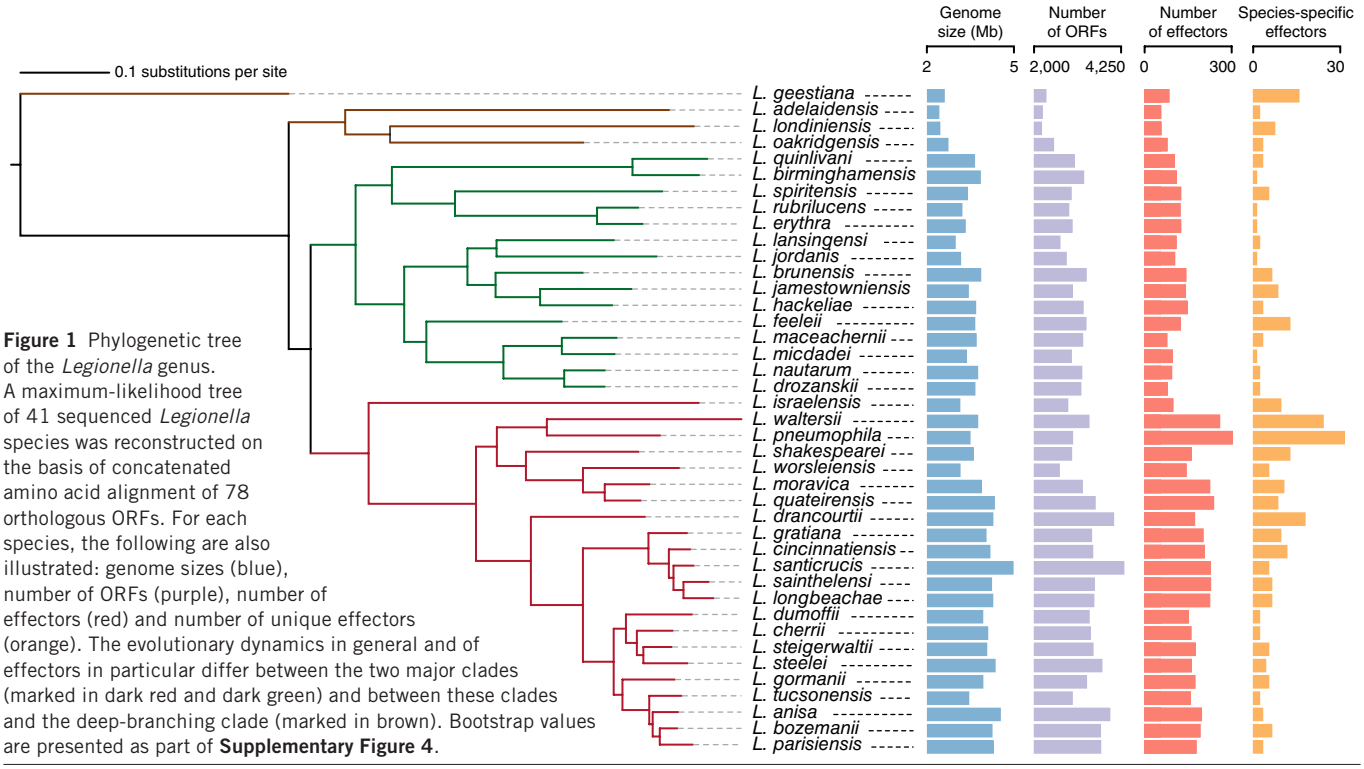

Figure 1 Phylogenetic tree of the *Legionella* genus. A maximum-likelihood tree of 41 sequenced *Legionella* species was reconstructed on the basis of concatenated amino acid alignment of 78 orthologous ORFs. For each species, the following are also illustrated: genome sizes (blue), number of ORFs (purple), number of effectors (red) and number of unique effectors (orange). The evolutionary dynamics in general and of effectors in particular differ between the two major clades (marked in dark red and dark green) and between these clades and the deep-branching clade (marked in brown). Bootstrap values are presented as part of **Supplementary Figure 4**.

The species phylogeny was reconstructed on the basis of a concatenated alignment of 78 nearly universal bacterial proteins[15] that corresponded to single genes in each genome (Online Methods). The phylogenetic tree (**Fig. 1**) indicates a division of the genus into three major clades: (i) a clade containing 22 species including the best studied *Legionella* pathogens—*L. pneumophila*, *L. longbeachae*, *Legionella bozemanii* and *L. dumoffii*—responsible together for more than 97.8% of human cases of *Legionella* infection[16]; (ii) another major clade characterized by long branches that encompasses 15 *Legionella* species, including *Legionella micdadei*; and (iii) a deep-branching clade consisting of *Legionella oakridgensis*, *Legionella londiniensis* and *Legionella adelaidensis*. Finally, as previously reported[17], *Legionella geestiana* is an outgroup to the rest of the *Legionella* genus (on the basis of tree rooting using *C. burnetii*).

The length of the reconstructed genomes ranged from 2.37 Mb in *L. adelaidensis* to 4.82 Mb in *Legionella santicrucis* (**Fig. 1**). Notably, the deep-branching clade is characterized by species with significantly smaller genomes as compared to the rest of the species ($P = 2.3 \times 10^{-8}$, one-sided *t* test). The GC content of the genomes was highly variable, ranging from 36.7% in *L. santicrucis* to 51.1% in *L. geestiana*. The genomes of five species forming a monophyletic group (*Legionella quinlivanii*, *Legionella birminghamensis*, *Legionella spiritensis*, *Legionella erythra* and *Legionella rubrilucens*) had significantly higher GC content (43.0–47.7%; $P = 0.0004$, one-sided Wilcoxon test) than the remaining genomes. Other species with high GC content were spread across the *Legionella* tree (**Supplementary Table 4**).

The Icm/Dot secretion system

The Icm/Dot type IVB secretion system is the major pathogenesis system of *L. pneumophila*. The system components are encoded by 25 genes organized in two separate genomic regions[18]. In both Icm/Dot regions, gene order and orientation were perfectly conserved throughout the genus (**Fig. 2** and **Supplementary Fig. 1**). The main differences in organization were in the presence of gene insertions of variable size that seem to be unrelated to the Icm/Dot secretion system. For example, there is an insertion of seven genes between *icmB* and *icmF* in *Legionella brunensis*, as compared to the insertion of a single gene in *L. pneumophila* and no intervening gene in *L. quinlivanii*. In 15 *Legionella* species, an OmpR family two-component regulatory system is encoded next to the *icmB* gene, the last gene in the large subregion of region II. These 15 species are monophyletic (**Fig. 2**), suggesting that this regulatory system was acquired once and has been preserved since then. Similar patterns of gene insertions were observed in region I (**Supplementary Note**). Notably, the locations of the gene insertions were highly conserved across the genus, suggesting tight co-regulation within subregions composed of sets of *icm*/*dot* genes that were not separated throughout the genus evolution.

The *Legionella* genus effector repertoire

The *L. pneumophila* Icm/Dot type IV secretion system translocates a large cohort of approximately 300 effector proteins[19,20]. To predict new effectors in all available *Legionella* genomes, we first identified in each species proteins highly similar to experimentally validated *Legionella* effectors. These proteins served as training sets for a machine learning procedure that we previously developed and proved high precision rates using experimental validations[21,22]. The machine learning procedure takes into account various aspects of the effector-coding genes, including regulatory information, the existence of eukaryotic motifs, the Icm/Dot secretion signal, and similarity to known effectors and host proteins. Predictions were performed for each genome separately, enabling the recognition of patterns unique to individual *Legionella* species (**Supplementary Fig. 2**).

The number of putative effectors was highly variable, ranging from 52 in *L. adelaidensis* to 247 in *Legionella waltersii* (**Fig. 1**). Species in different clades of the *Legionella* tree significantly differed in the number of predicted effector-coding genes, even when accounting for variance in genome size (ANOVA, $P = 2.77 \times 10^{-6}$). The species from the deep-branching clade contained on average 59 effectors, as

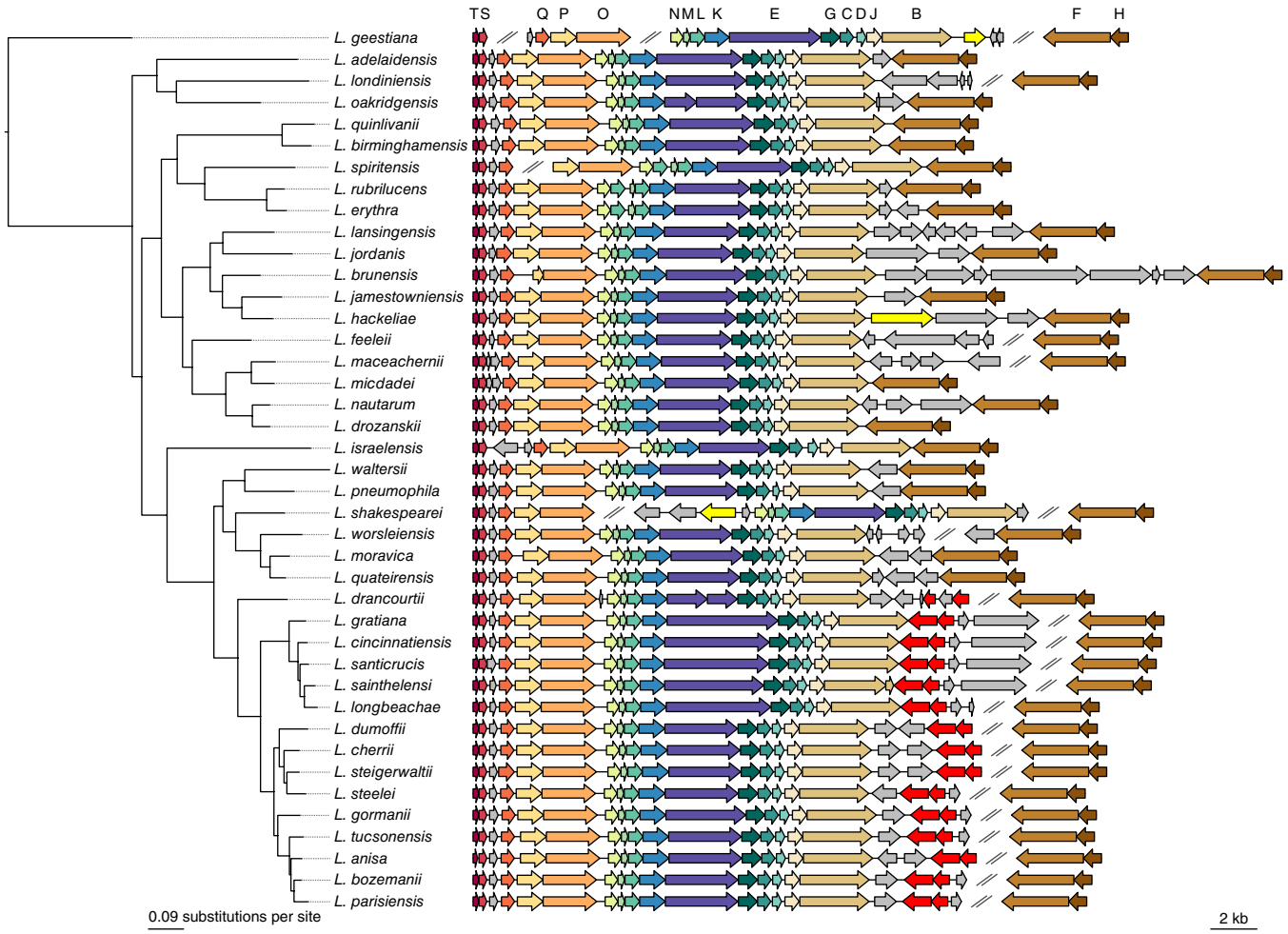

Figure 2 Icm/Dot secretion system region II in 41 *Legionella* species. In 15 genomes, genes encoding an OmpR family two-component system were found (bright red). In three other genomes, putative effectors were found in region II (bright yellow). Genes colored in gray represent non-effector genes found between the two parts of the region. Gene symbols: T, *icmT*; S, *icmS*; R, *icmR*; Q, *icmQ*; P, *icmP*/*dotM*; O, *icmO*/*dotL*; N, *icmN*/*dotK*; M, *icmM*/*dotJ*; L, *icmL*/*dotI*; K, *icmK*/*dotH*; E, *icmE*/*dotG*; G, *icmG*/*dotF*; C, *icmC*/*dotE*; D, *icmD*/*dotP*; J, *icmJ*/*dotN*; B, *icmB*/*dotO*; F, *icmF*; H, *icmH*/*dotU*. A similar analysis of region I is presented in **Supplementary Figure 1**.

compared to an average of 107 in the major *L. micdadei* clade and 183 in the *L. pneumophila* clade. This variability does not seem to be due to bias in the effector identification procedure (**Supplementary Fig. 3** and **Supplementary Note**). In total, we identified in the *Legionella* genus a set of 5,885 putative effectors.

Orthologous groups of *Legionella* genes that consisted of ≥80% predicted effectors were designated *Legionella* effector ortholog groups (LEOGs). We identified 608 LEOGs and found that most of them were shared by a small subset of species. Surprisingly, only seven effectors were 'core effectors', that is, had orthologs in every *Legionella* genome analyzed (**Fig. 3**). Notably, about 63% of the effector repertoire (3,715 effectors in 269 LEOGs) consisted of orthologs of experimentally validated effectors from *L. pneumophila* and *L. longbeachae*. The rest—2,170 effectors in 339 LEOGs—represent new putative effectors, potentially with novel functionality.

We identified 15 cases of clear effector pseudogenization due to nonsense mutations (Online Methods and **Supplementary Table 5**). Our results suggest that some species are more prone to pseudogenization. For example, five pseudogenes in *Legionella anisa* and *L. bozemanii* were homologous to complete genes in *Legionella steelei*, but no pseudogene was identified in *L. steelei* itself. Effector pseudogenization does not necessarily result in a non-functional protein. This process might be part of effector evolution, leading to diversification, as was suggested for effectors in *C. burnetii*[11].

The high number of effectors predicted in the *Legionella* genus allowed us to perform genomic analyses on this extraordinary group of genes. These analyses resulted in intriguing observations regarding the distribution, function and evolution of the *Legionella* genus effector repertoire.

The seven core effectors of the *Legionella* genus

In light of the high number of LEOGs found, the identification of only seven core effectors in the *Legionella* genus was surprising. The evolutionary histories of the core effectors are in good agreement with the species tree (Online Methods and **Supplementary Fig. 4**), and, whereas all core effectors are highly similar across species at the protein level, their GC content is variable and similar to the average genomic GC content of each species (**Supplementary Table 4**). In combination, these findings suggest that these core effectors evolved as part of the *Legionella* genus for an extended period of time.

Remarkably, only a single core effector (LOG_00212, represented in *L. pneumophila* by lpg2300/LegA3) was found in all the bacteria known to harbor an Icm/Dot secretion system. This effector was present in all the *Legionella* species examined, as well as in *C. burnetii*

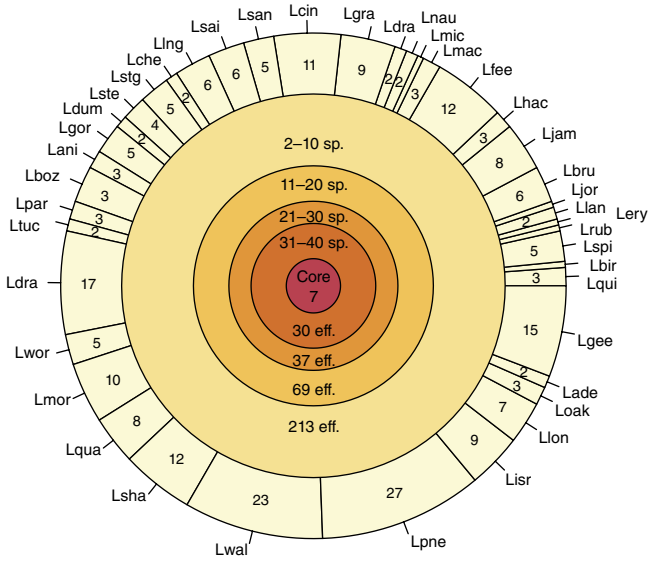

Figure 3 Extent of effector sharing by the *Legionella* species studied. Circles represent sets of effectors (eff.) shared by a different number of *Legionella* species (sp.). The number of species in which these effectors were found is indicated at the top of each circle, and the number of effectors contained in the set is indicated on the bottom. The innermost circle represents the set of core effectors shared by all *Legionella* species studied. The outermost circle depicts the 258 species-specific effectors and is divided on the basis of the number of species-specific effectors found in each *Legionella* species. Notably, only seven effectors are shared by all *Legionella* species, and most effectors (78.5%) are shared by ten or fewer species (two outermost circles). Lpne, *L. pneumophila*; Ldra, *L. drancourtii*; Llng, *L. longbeachae*; species abbreviations for newly sequenced species appear in **Supplementary Table 1**.

(CBU_1292) and *R. grylli* (RICGR_1042). The homologs in these bacteria show a high degree of similarity with their *L. pneumophila* counterpart throughout the length of the protein (BLAST *e* value = 2×10^{-116} and 3×10^{-120}, respectively). All members of this LEOG contain a single ankyrin repeat at their N terminus. Ankyrin repeats are usually found in eukaryotic proteins, where they mediate protein-protein interactions in a wide range of protein families[23,24].

An additional core effector, LOG_00334 (MavN), was found in *R. grylli* (RICGR_0048) but not in *C. burnetii*. The *L. pneumophila* MavN (lpg2815) effector is strongly induced under iron-restricted conditions. Mutants lacking the corresponding gene are defective for growth on iron-depleted solid media, defective for ferrous iron uptake and impaired in intracellular growth within their environmental host, *Acanthamoeba castellanii*[25,26]. These findings suggest that this core effector might be involved in iron acquisition during intracellular growth within the iron-poor milieu of the *Legionella*-containing *vacuole* (LCV).

The five other core effectors were not found in either *C. burnetii* or *R. grylli*, but two—LOG_00341 and LOG_01049 (RavC), represented in *L. pneumophila* by lpg2832 and lpg0107, respectively—had homologs in more distant bacteria. lpg2832 is homologous to proteins in several Rhizobiales such as *Bradyrhizobium oligotrophicum* (BLAST *e* value = 1×10^{-22}). These bacteria are symbionts of leguminous plants that fix atmospheric nitrogen and use a type IVA secretion system for symbiosis[27,28]. lpg0107 (RavC) has homologs in several members of the Chlamydiae phylum, such as *Diplorickettsia massiliensis* (*e* value = 2×10^{-49}), *Protochlamydia amoebophila* (*e* value = 2×10^{-34}) and *Chlamydia trachomatis* (*e* value = 3×10^{-31}). All these bacteria are intracellular human pathogens that use a type III secretion system for intracellular growth. The presence of these two effectors in evolutionarily distant species could be the result of cross-genera horizontal gene transfer (HGT), or, alternatively, these genes might have existed in a common ancestor and been lost in the lineages lacking them. We tested these alternative hypotheses by comparing two models representing these evolutionary scenarios (Online Methods) and conclude that these effectors have been horizontally transferred across genera ($P = 5 \times 10^{-34}$ for LOG_00341 and 7×10^{-60} for LOG_01049, likelihood-ratio test) and were adapted in different pathogens to different secretion systems.

To obtain a first indication regarding the importance of these seven core effectors, knockout mutants were constructed in *L. pneumophila* and tested for intracellular growth in *A. castellanii*. The results indicated that LegA3 and MavN are partially required for intracellular growth in this host (**Supplementary Fig. 5**), with the growth defects completely complemented when each

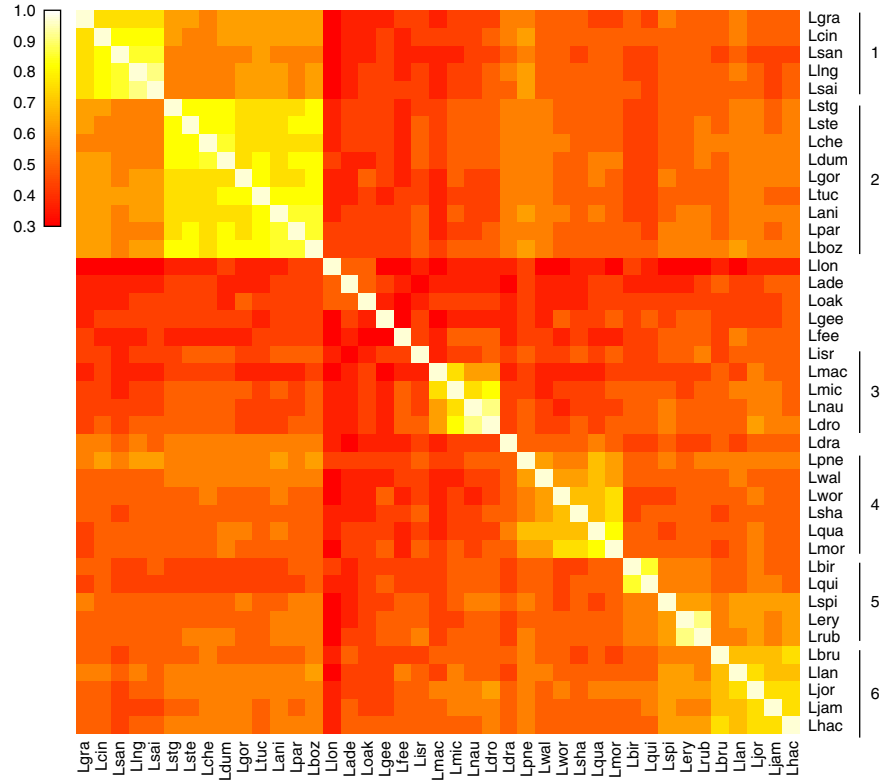

Figure 4 Comparison of the putative effector pools among *Legionella* species. The color gradient represents the similarity between sets of effectors (light colors for high similarity). Clusters defined on the basis of similar effector repertoires (marked on the right) are in agreement with the clades of the phylogenetic tree (**Supplementary Fig. 6**). Lpne, *L. pneumophila*; Ldra, *L. drancourtii*; Llng, *L. longbeachae*; species abbreviations for newly sequenced species appear in **Supplementary Table 1**.

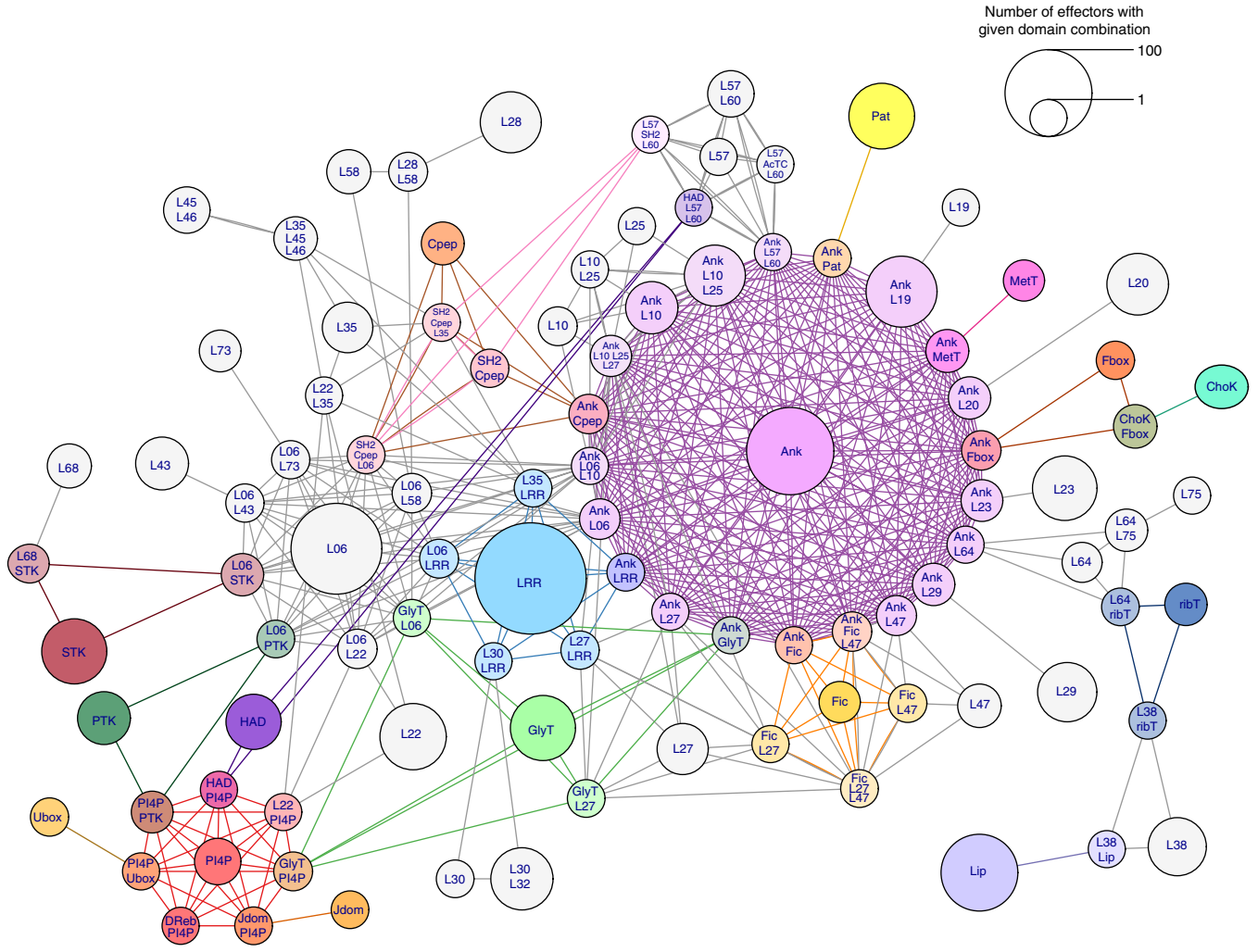

Figure 5 Protein architecture network of effectors. Each node represents a specific protein architecture (combination of effector domains; **Supplementary Table 8**). Node labels indicate domains taking part in the architecture. Edges represent domains shared by architectures. Known domains are colored; new conserved effector domains are in gray. Node size is proportional to the number of ORFs with the architecture represented by the node.

of the effectors was reintroduced on a plasmid. To conclude, although the exact function of these seven *Legionella* genus core effectors is unknown, their high conservation throughout the evolution of the genus and, in some cases, beyond indicates that they perform critical functions during infection. The fact that only two of the core effectors presented an intracellular growth phenotype when deleted suggests that the function of the other core effectors may be redundant, at least for intracellular growth in *A. castellanii*. It is possible, however, that these core effectors carry out essential functions required for growth in other hosts.

Unique effectors in the *Legionella* genus

Analysis of the LEOGs determined that 258 (42%) were species specific, meaning that they were observed in only one of the *Legionella* species analyzed. Excluding *L. pneumophila*, the species with the highest number of unique putative effectors was *L. waltersii*, with 23 species-specific effectors. Notably, every genome analyzed had at least a single species-specific effector (**Fig. 1**). The GC content of these effectors was consistently lower than the genomic GC content (**Supplementary Table 4**), suggesting that these genes might have been recently acquired from an exogenous source, possibly from the natural protozoan hosts of *Legionella*, which are typically characterized by low GC content[29].

The 258 species-specific LEOGs included 188 LEOGs that displayed local similarity to at least one other *Legionella* protein (BLAST *e* value $<1 \times 10^{-4}$) and 70 putative effectors with no similarity to any other *Legionella* protein. Of these 70 unique putative effectors, only five were similar to a known protein (*e* value $<1 \times 10^{-4}$, BLAST search against the NCBI non-redundant database; **Supplementary Table 6**). Four of these were similar to proteins encoded by bacteria from different ecosystems, and the fifth (Lmac_0005) was similar to a hypothetical protein from *Candidatus Protochlamydia amoebophila*, an endosymbiont of *Acanthamoeba* species[30]. This hit was not highly significant (*e* value $= 3.09 \times 10^{-6}$), yet it might be the result of HGT occurring inside a common protozoan host. For 62 of the 70 unique effectors, we could find sequence-based support that these unique proteins are indeed genuine effectors (**Supplementary Note**). The fact that none of these 70 unique effectors had significant sequence similarity to another *Legionella*-encoded protein demonstrates the magnitude of functional novelty of the *Legionella* effectors. The low GC content of species-specific effectors (**Supplementary Table 4**) in combination with the fact that most of them contain an Icm/Dot-associated regulatory element or an Icm/Dot secretion signal suggests that recently acquired genes can be adapted to function as effectors in a relatively short evolutionary timeframe.

Dynamics of effector gain and loss

To gain insights into the evolutionary processes that shape the current effector repertoire of each *Legionella* species, we examined the dynamics of effector gain and loss along the phylogenetic tree. The results of this analysis (**Supplementary Fig. 6**) demonstrate that the rate of acquisition and loss of effectors in the *L. micdadei* clade is significantly lower than that in the *L. pneumophila* clade ($P = 1.4 \times 10^{-10}$, one-sided Wilcoxon test)—that is, the species belonging to the latter clade have a more dynamic repertoire of effectors. This finding is in agreement with the number of unique LEOGs found in the different *Legionella* species (**Fig. 1**). These analyses suggest that certain *Legionella* species, including the most pathogenic ones, acquire genetic information from outside the *Legionella* genus more frequently than others and adapt these sequences to encode functional effector proteins.

Comparison of effector repertoires among the *Legionella* species showed that, despite the dynamic nature of the effector repertoire, evolutionarily close species tended to have similar sets of effectors (**Fig. 4**). Phylogenetic analysis of effectors undergoing numerous HGT events suggests that this is probably due to preferential HGT among closely related species (**Supplementary Fig. 7** and **Supplementary Note**).

Effector synteny and co-evolution

Effector-coding genes residing in close proximity on the genome have been shown in some cases to function together in the host cell[31,32]. We hence searched for effectors that were consistently found together (within five ORFs of each other) across multiple genomes. We found 143 pairs of effectors whose members were in close proximity in at least two genomes, with 51 pairs found in five genomes or more. The syntenic genes and their organization in each genome are detailed in **Supplementary Table 7**. We further analyzed which of these pairs displayed statistically significant co-evolution, that is, which constituted syntenic effectors that were gained or lost together during genus evolution more often than expected by chance. The combined analysis identified 19 pairs of effectors that are syntenic and co-evolved (**Supplementary Table 7**), some of which are already known to have related functions. For example, AnkX (LOG_03154) and Lem3 (LOG_032115) both modulate the host GTPase Rab1 (ref. 33) (counteracting each other). Similarly, SidH (LOG_04780) and LubX (LOG_06016) were also shown to function together[34]. Recently, it was reported that SidJ affects the localization and toxicity of effectors from the SidE family[35,36]. Here we found that the gene encoding SdjA (a *sidJ* paralog; LOG_04652) was consistently next to the gene encoding SdeD (encoded by a *sidE* paralog; LOG_04652) in all six genomes where both of them were present, and the gene encoding SdeC (another paralog of *sidE*) was found next to that encoding SidJ in five of the six genomes that encoded both effectors. These results led us to examine additional syntenic effectors that co-evolved. We found five such pairs of effectors in *L. pneumophila* (SidL-LegA11, Lpg2888-MavP, SidI-Lpg2505, Ceg3-Lpg0081 and CetLp7-Lem29) that were not previously described to function as pairs. Notably, SidL and SidI inhibit translation by interacting with the translation elongation factor eEF1A and inhibit yeast growth[37,38]. The LegA11 and Lpg2505 effectors might counteract the activity of SidL and SidI, respectively, as the translational block mediated by SidL and SidI early during *L. pneumophila* infection is probably removed to enable successful infection.

Domain shuffling has a major role in effector evolution

Previous studies of *L. pneumophila* effectors reported that they harbor numerous eukaryotic domains[39,40] as well as effector-specific domains[41,42]. The high number of effectors we found made it possible to identify and analyze conserved effector domains across the *Legionella* genus. We identified *Legionella* effector domains (LEDs) using two separate criteria: (i) similarity to known domains in domain databases and (ii) conservation of effector regions across orthologous groups (Online Methods). Conserved domains were detected in 56% of the putative effectors. In total, 99 distinct domains were identified using the two criteria: 53 well-characterized (mostly eukaryotic) domains and 46 new conserved domains, reported here for the first time (**Supplementary Table 8**). Analyzing the protein architectures

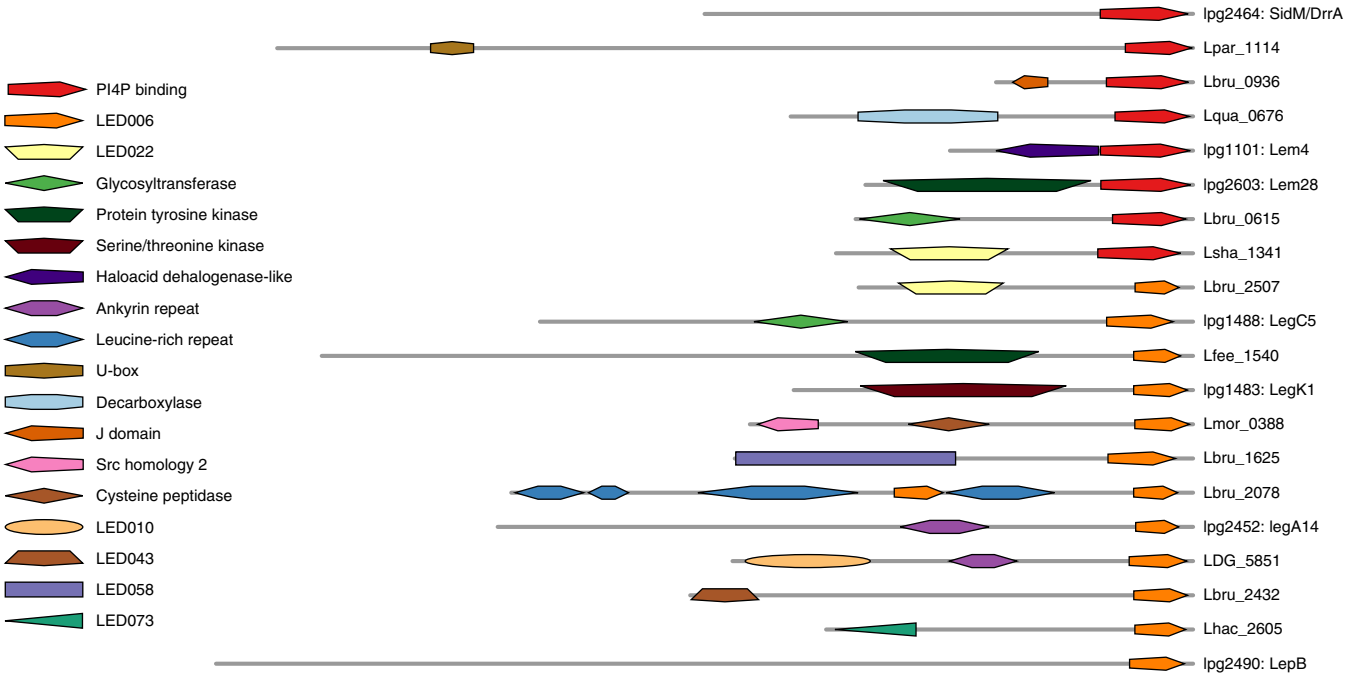

Figure 6 Architectures containing either the PI4P-binding domain or LED006. Each protein architecture containing either the PI4P-binding domain or the novel, uncharacterized LED006 domain is represented by a single putative effector.

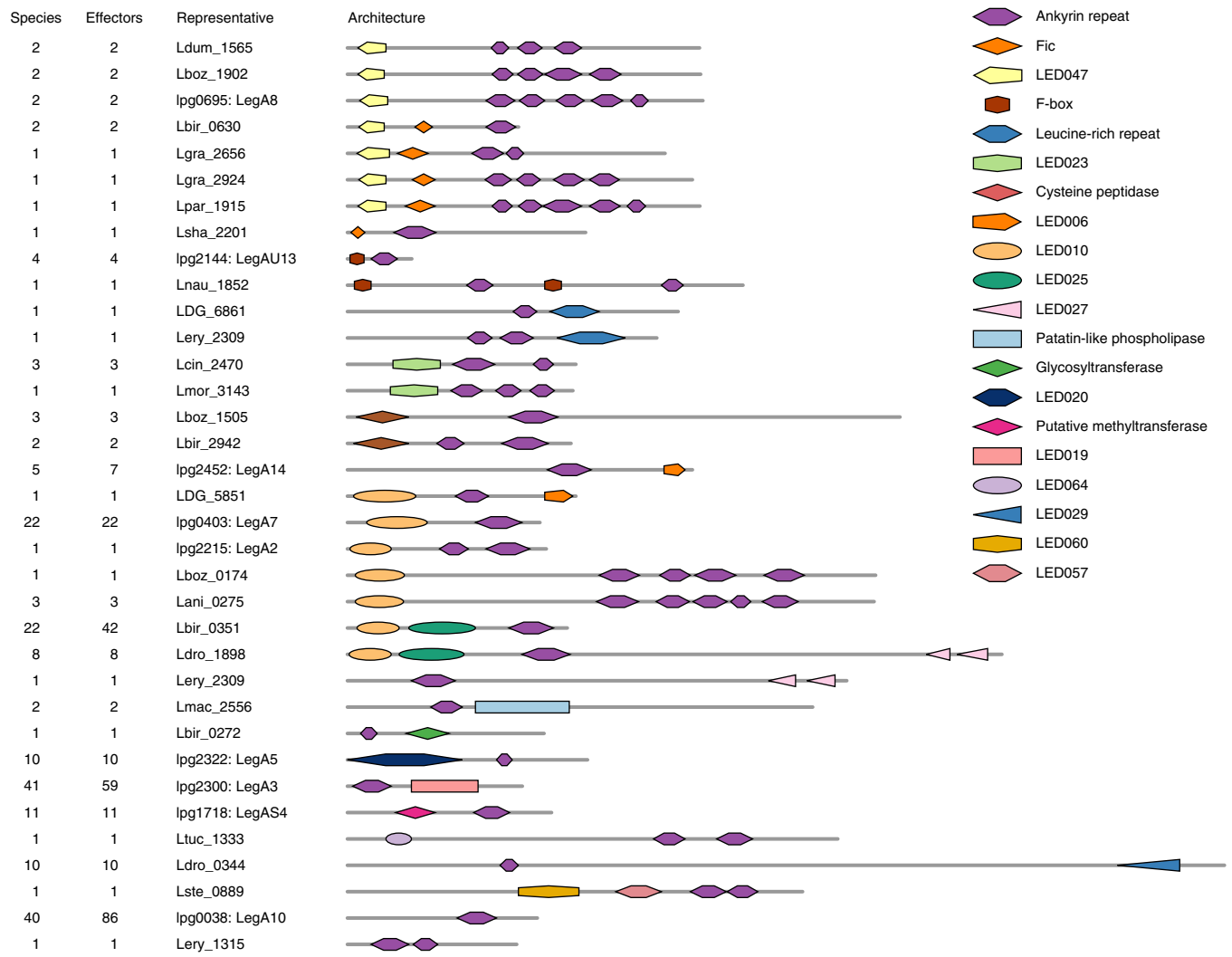

Figure 7 Diversity of ankyrin repeat–containing putative effectors. Each domain configuration that includes an ankyrin domain is represented by a single example (architectures with a different number of ankyrin repeats are represented separately). The number of putative effectors, as well as the number of *Legionella* species in which each configuration occurs, is indicated.

(different domain combinations), we noticed that the same domains were often shared by different architectures. We visualized this phenomenon as a network of protein architectures connected by shared domains. The biggest connected subnetwork of architectures found, which harbored many known effector domains such as ankyrin repeats, the leucine-rich repeat (LRR) domain and the phosphatidylinositol 4-phosphate (PI4P)-binding domain, is displayed in **Figure 5**. The network clearly demonstrates that several domains are present in numerous effectors (indicated by node size in **Fig. 5**), as well as in numerous different architectures (indicated by the number of connected nodes harboring the same LED in **Fig. 5**). The network represents a wealth of domains with known and unknown functions, from which insights regarding possible effector functions can be deduced.

The PI4P-binding domain localizes effectors (including lpg2464-SidM/DrrA, lpg1101 and lpg2603) to the LCV[41,43]. This domain is of special interest because it indicates which functions might be targeted to the LCV in different *Legionella* species. PI4P-binding domains were found as part of eight architectures in 36 putative effectors. In all these architectures, the PI4P-binding domain was located at the C-terminal end, in some cases together with an additional characterized domain (**Fig. 6** and **Supplementary Fig. 8**). For example, in *Legionella parisiensis*, the PI4P-binding domain was found on a putative effector (Lpar_1114) together with a U-box domain. The U-box domain is typically found in ubiquitin-protein ligases where it determines the substrate specificity for ubiquitination (E3 ubiquitin ligases)[44]. In *L. pneumophila*, a functional U-box domain was found in the LegU2/LubX (lpg2830) effector[34], which does not contain a PI4P-binding domain. The presence of a U-box domain together with a PI4P-binding domain in Lpar_1114 suggests that this effector is involved in protein ubiquitination on the LCV. Interestingly, it was previously shown that ubiquitination occurs on the *L. pneumophila* LCV as well, but no ORF containing both PI4P-binding and U-box domains was found in this species. Instead, in *L. pneumophila*, this function is mediated (at least in part) by LegAU13/AnkB (lpg2144), which contains a U-box domain and an ankyrin repeat. This effector anchors to the LCV membrane by host-mediated farnesylation that occurs at the C-terminal end of the protein[45,46]. Collectively, these results demonstrate that *Legionella* species use a variety of molecular mechanisms to direct effectors to the LCV.

An additional domain found together with the PI4P-binding domain was the glycosyltransferase domain, which glycosylates

proteins. In *L. pneumophila*, a functional glycosyltransferase domain was found in the N terminus of the SetA effector (lpg1978), in which a C-terminal phosphatidylinositol 3-phosphate (PI3P)-binding domain is required for proper localization to the early LCV[47] (PI3P- and PI4P-binding domains are not homologous). We found glycosyltransferase domains together with PI4P-binding domains in putative effectors from three species (**Supplementary Fig. 8** and **Supplementary Table 8**), suggesting that these putative effectors also localize to the LCV. The PI4P-binding localization domain was found with numerous additional characterized domains as well as novel domains, which possibly mediate additional functions on the LCV.

Analysis of the architectures containing the PI4P-binding domain provided putative insights into the function of LED006, a prevalent novel domain. Both the PI4P-binding domain and LED006 were found with a protein tyrosine kinase domain, a glycosyltransferase domain and a domain with unknown function—LED022 (**Fig. 6** and **Supplementary Fig. 8**). Similar to the PI4P-binding domain, LED006 was also located at the C-terminal end of all the putative effectors in which it was found. The *L. pneumophila* effector LepB (lpg2490), which contains the LED006 domain, localizes to the LCV[33,48] and functions as a GTPase-activating protein (GAP) for the Rab1 protein (a small GTPase that regulates endoplasmic reticulum (ER)-to-Golgi trafficking[49]). The region overlapping LED006 is required for LepB targeting to the LCV[33,48], and effectors containing the LED006 domain often harbor a functional domain at their N terminus. These observations suggest that LED006 is another domain involved in the targeting of effectors to the LCV. The putative effectors in which only LED006 was identified probably contain an N-terminal domain that was not conserved enough to be identified by the stringent methods we applied.

Ankyrin repeats, which mediate protein-protein interactions and direct effectors to their target proteins in the host, were the most versatile domain according to our analysis. They appeared with a multitude of various domains and architectures in different species. Overall, ankyrin repeats were found in 22 different architectures (35 different architectures when taking into account different numbers of repeats) in more than 300 putative effectors (**Fig. 7** and **Supplementary Fig. 9**). Whereas some of the ankyrin repeat–containing architectures were present only in species-specific effectors, others were widespread and appeared in all 41 species analyzed. Some domains, including F-box, Fic and LRR domains, were found adjacent to varying numbers of ankyrin repeats in different effectors, further demonstrating the degree of domain variability in *Legionella* effectors. Effectors with ankyrin repeats might target a host protein to the LCV, in which case they would be expected to also have a localization domain. Alternatively, effectors with ankyrin repeats might have an additional enzymatic domain that serves to manipulate the targeted host protein.

The various architectures in which the domains are found and the different organizations of shared domains demonstrate the vast functional variability of the effectors in the *Legionella* genus and the important effect that domain shuffling has on the evolution of the virulence system of these intracellular pathogens.

DISCUSSION

Pathogens belonging to the *Legionella* genus cause severe, often fatal disease in humans. This is despite the fact that *Legionella* have not co-evolved with humans: *Legionella* are not transmitted from person to person, and they are thus either defeated by their human hosts or perish with them. They are able to manipulate numerous cellular pathways in humans owing to a large and versatile repertoire of effector proteins acquired during their co-evolution with a variety of protozoan hosts. *De novo* sequencing of 38 *Legionella* species allowed us to predict and extensively analyze an enormous cohort of 5,885 putative effectors belonging to more than 600 orthologous groups. The effectors were predicted using stringent species-specific cutoffs to minimize false detection. Hence, the total number of effectors is expected to be higher than that reported. We estimate, on the basis of a second round of predictions performed on the combined set of genomes (Online Methods), that the total number of effectors in the genomes analyzed might be as high as 9,300. This amounts to 7.2% of all the ORFs, as compared to *L. pneumophila*, where 10% of the genome encodes validated effectors.

We found that the vast majority (78.5%) of the 5,885 putative effectors identified in this study are shared by ten or fewer species, and only a handful of effectors are present across the genus. These findings, in combination with the atypically low GC content of species-specific effectors, suggest that these genes have been recently acquired, probably as part of an ongoing process of acquiring genes from hosts and co-infecting pathogens and adapting them to function as effectors. Notably, we identified dozens of conserved effector domains, which are the basic building blocks that, when rearranged during the course of evolution, contribute to the myriad of functions exerted by *Legionella* effectors.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The whole-genome shotgun projects have been deposited at the DNA Data Bank of Japan (DDBJ)/European Molecular Biology Laboratory (EMBL)/GenBank under accessions LNKA00000000, LNXS00000000, LNXT00000000, LNXU00000000, LNXV00000000, LNXW00000000, LNXX00000000, LNXY00000000, LNXZ00000000, LNYA00000000, LNYB00000000, LNYC00000000, LNYD00000000, LNYE00000000, LNYF00000000, LNYG00000000, LNYH00000000, LNYI00000000, LNYJ00000000, LNYK00000000, LNYL00000000, LNYM00000000, LNYN00000000, LNYO00000000, LNYP00000000, LNYQ00000000, LNYR00000000, LNYS00000000, LNYT00000000, LNYU00000000, LNYV00000000, LNYW00000000, LNYX00000000, LNYY00000000, LNYZ00000000, LNZA00000000, LNZB00000000 and LNZC00000000. The data have been deposited with links to NCBI BioProject PRJNA285910.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We wish to thank E. Levy Karin for her kind help with some of the phylogenetic analyses. This work was supported by National Institute of Allergy and Infectious Diseases grant 5RO1 AI23549 (H.A.S.), and the costs of DNA sequencing were supported by startup funds from the Division of Biological Sciences of the University of Chicago (H.A.S.). This work was also supported in part by grant 2013240 from the United States–Israel Binational Science Foundation (G.S. and H.A.S.). T.P. was supported by Israel Science Foundation (ISF) grant 1092/13. D.B. was a fellow of the Converging Technologies Program of the Israeli Council for Higher Education. D.B. and T.P. were also supported by the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. F.A. was supported by a postdoctoral fellowship from the Fulbright Commission and the Ministry of Education of Spain.

AUTHOR CONTRIBUTIONS

G.S., H.A.S., D.B. and F.A. designed the study. F.A., T.Z. and Z.L. performed experiments and analyzed the results. D.B. and O.C. performed the bioinformatic analysis. G.S., H.A.S., T.P. and J.A.G. supervised the research. D.B. and G.S. wrote the manuscript. All authors approved and contributed to the final version of the manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Hayes, C.S., Aoki, S.K. & Low, D.A. Bacterial contact-dependent delivery systems. *Annu. Rev. Genet.* **44**, 71–90 (2010).
2. Fields, B.S. The molecular ecology of legionellae. *Trends Microbiol.* **4**, 286–290 (1996).
3. Fields, B.S., Benson, R.F. & Besser, R.E. *Legionella* and Legionnaires' disease: 25 years of investigation. *Clin. Microbiol. Rev.* **15**, 506–526 (2002).
4. Diederen, B.M.W. *Legionella* spp. and Legionnaires' disease. *J. Infect.* **56**, 1–12 (2008).
5. Segal, G., Purcell, M. & Shuman, H.A. 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. USA* **95**, 1669–1674 (1998).
6. Vogel, J.P., Andrews, H.L., Wong, S.K. & Isberg, R.R. Conjugative transfer by the virulence system of *Legionella pneumophila*. *Science* **279**, 873–876 (1998).
7. Isberg, R.R., O'Connor, T.J. & Heidtman, M. The *Legionella pneumophila* replication vacuole: making a cosy niche inside host cells. *Nat. Rev. Microbiol.* **7**, 13–24 (2009).
8. Feldman, M., Zusman, T., Hagag, S. & Segal, G. Coevolution between nonhomologous but functionally similar proteins and their conserved partners in the *Legionella* pathogenesis system. *Proc. Natl. Acad. Sci. USA* **102**, 12206–12211 (2005).
9. Feldman, M. & Segal, G. A specific genomic location within the *icm/dot* pathogenesis region of different *Legionella* species encodes functionally similar but nonhomologous virulence proteins. *Infect. Immun.* **72**, 4503–4511 (2004).
10. Beare, P.A. *et al.* Dot/Icm type IVB secretion system requirements for *Coxiella burnetii* growth in human macrophages. *MBio* **2**, e00175–11 (2011).
11. Carey, K.L., Newton, H.J., Lührmann, A. & Roy, C.R. The *Coxiella burnetii* Dot/Icm system delivers a unique repertoire of type IV effectors into host cells and is required for intracellular replication. *PLoS Pathog.* **7**, e1002056 (2011).
12. Leclerque, A. & Kleespies, R.G. Type IV secretion system components as phylogenetic markers of entomopathogenic bacteria of the genus *Rickettsiella*. *FEMS Microbiol. Lett.* **279**, 167–173 (2008).
13. O'Connor, T.J., Adepoju, Y., Boyd, D. & Isberg, R.R. Minimization of the *Legionella pneumophila* genome reveals chromosomal regions involved in host range expansion. *Proc. Natl. Acad. Sci. USA* **108**, 14733–14740 (2011).
14. O'Connor, T.J., Boyd, D., Dorer, M.S. & Isberg, R.R. Aggravating genetic interactions allow a solution to redundancy in a bacterial pathogen. *Science* **338**, 1440–1444 (2012).
15. Puigbò, P., Wolf, Y.I. & Koonin, E.V. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. *J. Biol.* **8**, 59 (2009).
16. Muder, R.R. & Yu, V.L. Infection due to *Legionella* species other than *L. pneumophila*. *Clin. Infect. Dis.* **35**, 990–998 (2002).
17. Ko, K.S. *et al.* Application of RNA polymerase β-subunit gene (*rpoB*) sequences for the molecular differentiation of *Legionella* species. *J. Clin. Microbiol.* **40**, 2653–2658 (2002).
18. Segal, G., Feldman, M. & Zusman, T. The Icm/Dot type-IV secretion systems of *Legionella pneumophila* and *Coxiella burnetii*. *FEMS Microbiol. Rev.* **29**, 65–81 (2005).
19. Lifshitz, Z. *et al.* Computational modeling and experimental validation of the *Legionella* and *Coxiella* virulence-related type-IVB secretion signal. *Proc. Natl. Acad. Sci. USA* **110**, E707–E715 (2013).
20. Gomez-Valero, L., Rusniok, C., Cazalet, C. & Buchrieser, C. Comparative and functional genomics of *Legionella* identified eukaryotic like proteins as key players in host-pathogen interactions. *Front. Microbiol.* **2**, 208 (2011).
21. Burstein, D. *et al.* Genome-scale identification of *Legionella pneumophila* effectors using a machine learning approach. *PLoS Pathog.* **5**, e1000508 (2009).
22. Lifshitz, Z. *et al.* Identification of novel *Coxiella burnetii* Icm/Dot effectors and genetic analysis of their involvement in modulating a mitogen-activated protein kinase pathway. *Infect. Immun.* **82**, 3740–3752 (2014).
23. Bork, P. Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally? *Proteins* **17**, 363–374 (1993).
24. Mosavi, L.K., Cammett, T.J., Desrosiers, D.C. & Peng, Z.Y. The ankyrin repeat as molecular architecture for protein recognition. *Protein Sci.* **13**, 1435–1448 (2004).
25. Portier, E. *et al.* *IroT/mavN*, a new iron-regulated gene involved in *Legionella pneumophila* virulence against amoebae and macrophages. *Environ. Microbiol.* **17**, 1338–1350 (2015).
26. Isaac, D.T., Laguna, R.K., Valtz, N. & Isberg, R.R. MavN is a *Legionella pneumophila* vacuole-associated protein required for efficient iron acquisition during intracellular growth. *Proc. Natl. Acad. Sci. USA* **112**, E5208–E5217 (2015).
27. Hennecke, H. Nitrogen fixation genes involved in the *Bradyrhizobium japonicum*–soybean symbiosis. *FEBS Lett.* **268**, 422–426 (1990).
28. Kaneko, T. *et al.* Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. *DNA Res.* **9**, 189–197 (2002).
29. Li, X.-Q. & Du, D. Variation, evolution, and correlation analysis of C+G content and genome or chromosome size in different kingdoms and phyla. *PLoS One* **9**, e88339 (2014).
30. Collingro, A. *et al.* '*Candidatus Protochlamydia amoebophila*', an endosymbiont of *Acanthamoeba* spp. *Int. J. Syst. Evol. Microbiol.* **55**, 1863–1866 (2005).
31. Tan, Y. & Luo, Z.-Q. *Legionella pneumophila* SidD is a deAMPylase that modifies Rab1. *Nature* **475**, 506–509 (2011).
32. Neunuebel, M.R. *et al.* De-AMPylation of the small GTPase Rab1 by the pathogen *Legionella pneumophila*. *Science* **333**, 453–456 (2011).
33. Ingmundson, A., Delprato, A., Lambright, D.G. & Roy, C.R. *Legionella pneumophila* proteins that regulate Rab1 membrane cycling. *Nature* **450**, 365–369 (2007).
34. Kubori, T., Shinzawa, N., Kanuka, H. & Nagai, H. *Legionella* metaeffector exploits host proteasome to temporally regulate cognate effector. *PLoS Pathog.* **6**, e1001216 (2010).
35. Jeong, K.C., Sexton, J.A. & Vogel, J.P. Spatiotemporal regulation of a *Legionella pneumophila* T4SS substrate by the metaeffector SidJ. *PLoS Pathog.* **11**, e1004695 (2015).
36. Havey, J.C. & Roy, C.R. Toxicity and SidJ-mediated suppression of toxicity require distinct regions in the SidE family of *Legionella pneumophila* effectors. *Infect. Immun.* **83**, 3506–3514 (2015).
37. Shen, X. *et al.* Targeting eEF1A by a *Legionella pneumophila* effector leads to inhibition of protein synthesis and induction of host stress response. *Cell. Microbiol.* **11**, 911–926 (2009).
38. Fontana, M.F. *et al.* Secreted bacterial effectors that inhibit host protein synthesis are critical for induction of the innate immune response to virulent *Legionella pneumophila*. *PLoS Pathog.* **7**, e1001289 (2011).
39. Cazalet, C. *et al.* Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. *Nat. Genet.* **36**, 1165–1173 (2004).
40. de Felipe, K.S. *et al.* Evidence for acquisition of *Legionella* type IV secretion substrates via interdomain horizontal gene transfer. *J. Bacteriol.* **187**, 7716–7726 (2005).
41. Brombacher, E. *et al.* Rab1 guanine nucleotide exchange factor SidM is a major phosphatidylinositol 4-phosphate–binding effector protein of *Legionella pneumophila*. *J. Biol. Chem.* **284**, 4846–4856 (2009).
42. Weber, S.S., Ragaz, C., Reus, K., Nyfeler, Y. & Hilbi, H. *Legionella pneumophila* exploits PI(4)P to anchor secreted effector proteins to the replicative vacuole. *PLoS Pathog.* **2**, e46 (2006).
43. Hubber, A. *et al.* The machinery at endoplasmic reticulum–plasma membrane contact sites contributes to spatial regulation of multiple *Legionella* effector proteins. *PLoS Pathog.* **10**, e1004222 (2014).
44. Ardley, H.C. & Robinson, P.A. E3 ubiquitin ligases. *Essays Biochem.* **41**, 15–30 (2005).
45. Price, C.T.D., Al-Quadan, T., Santic, M., Jones, S.C. & Abu Kwaik, Y. Exploitation of conserved eukaryotic host cell farnesylation machinery by an F-box effector of *Legionella pneumophila*. *J. Exp. Med.* **207**, 1713–1726 (2010).
46. Price, C.T.D., Al-Quadan, T., Santic, M., Rosenshine, I. & Abu Kwaik, Y. Host proteasomal degradation generates amino acids essential for intracellular bacterial growth. *Science* **334**, 1553–1557 (2011).
47. Jank, T. *et al.* Domain organization of *Legionella* effector SetA. *Cell. Microbiol.* **14**, 852–868 (2012).
48. Mishra, A.K., Del Campo, C.M., Collins, R.E., Roy, C.R. & Lambright, D.G. The *Legionella pneumophila* GTPase activating protein LepB accelerates Rab1 deactivation by a non-canonical hydrolytic mechanism. *J. Biol. Chem.* **288**, 24000–24011 (2013).
49. Mizuno-Yamasaki, E., Rivera-Molina, F. & Novick, P. GTPase networks in membrane traffic. *Annu. Rev. Biochem.* **81**, 637–659 (2012).

ONLINE METHODS

Sequencing, assembly and annotation. Thirty-eight *Legionella* isolates of the following species were collected: *L. adelaidensis*, *L. anisa*, *L. birminghamensis*, *L. bozemanii*, *L. brunensis*, *L. cherrii*, *L. cincinnatiensis*, *L. drozanskii*, *L. dumoffii*, *L. erythra*, *L. feeleii*, *L. geestiana*, *L. gormanii*, *L. gratiana*, *L. hackeliae*, *L. israelensis*, *L. jamestowniensis*, *L. jordanis*, *L. lansingensis*, *L. londiniensis*, *L. maceachernii*, *L. micdadei*, *L. moravica*, *L. nautarum*, *L. oakridgensis*, *L. parisiensis*, *L. quateirensis*, *L. quinlivanii*, *L. rubrilucens*, *L. sainthelensi*, *L. santicrucis*, *L. shakespearei*, *L. spiritensis*, *L. steelei*, *L. steigerwaltii*, *L. tucsonensis*, *L. waltersii* and *L. worsleiensis* (**Supplementary Table 1**). DNA was extracted from each sample using the DNeasy kit (Qiagen), including proteinase K and RNase treatments and following the manufacturer's instructions. The DNA was sequenced using the Illumina HiSeq platform, producing a total of 773 Gb of 100-bp paired-ends reads with a target insert size of 200–300 bp. Low-quality reads from each of the samples were trimmed using Trimmomatic[50], and trimmed reads were assembled using Velvet[51]. A combination of different Trimmomatic parameters and Velvet *k*-mer values was used to optimize assembly, as measured by N50 length (**Supplementary Table 2**).

The assembled genomes had a high median coverage of 606×. The number of contigs ranged between 12 and 154, and the median N50 length for the 38 genomes was 386 kb, with eight genomes having N50 >1.4 Mb (see **Supplementary Table 2** for details). The smallest reconstructed *Legionella* genome was that of *L. adelaidensis*, which had a 2.37-Mb genome, and the largest was that of *L. santicrucis*, for which we reconstructed 4.82 Mb. To ensure that the observed variation in genome size was not due to assembly quality or coverage, we tested for the correlation of genome size with coverage, N50 and the number of assembled contigs. None of these presented significant correlation with genome length (**Supplementary Table 2**).

ORFs were predicted using Prodigal[52] with default parameters. The ORFs from the different genomes were clustered into LOGs using OrthoMCL[53] (**Supplementary Table 3**). To assess genome completeness, we examined the presence of 55 genes consisting of 31 single-copy universal bacterial genes[54] and the 24 genes encoding the components of the Icm/Dot secretion system, which is universal in the *Legionella* genus. Of the 38 sequenced genomes, 36 contained 100% of the genes examined, and two genomes (*L. cherrii* and *L. santicrucis*) were each missing a single gene (**Supplementary Table 9**).

ORF annotation was performed on the basis of similarity to available fully sequenced *Legionella* genomes with preference given to *L. pneumophila* Philadelphia-1 (**Supplementary Table 10**). Specifically, annotation was transferred from *L. pneumophila* Philadelphia-1 (NCBI accession NC_002942) to BLAST hits. If no significant hit was found in comparison to Philadelphia-1, than annotation was transferred from the best hit out of the following genomes: *L. longbeachae* D-4968, *L. longbeachae* NSW150, *L. pneumophila* str. Corby, *L. pneumophila* 2300/99 Alcoy, *L. pneumophila* str. Paris, *L. pneumophila* str. Lens, *L. pneumophila* subsp. pneumophila ATCC 43290, *L. pneumophila* subsp. Pneumophila and *L. drancourtii* LLAP12 (NCBI accessions GCF_000176095.1, GCF_000091785.1, GCF_000092545.1, GCF_000092625.1, GCF_000048645.1, GCF_000048665.1, GCF_000239175.1, GCF_000306865.1, and GCF_000162755.2, respectively).

Machine learning approach for effector prediction. Icm/Dot effector prediction was performed using the machine learning approach we have previously described[22]. Briefly, for each ORF in each genome, we calculated an array of features that are expected to be informative for the classification of the ORF as an effector. The features used for learning included ORF length, GC content, similarity to proteins in sequenced *Legionella* hosts (human, *Tetrahymena thermophila* and *Dictyostelium discoideum*), similarity to *C. burnetii* RSA-493 ORFs, similarity to known Icm/Dot effectors, existence of eukaryotic domains typically found in effectors, amino acid composition, similarity of CpxR- and PmrA-binding sites within regulatory regions, presence of CAAX (a myristoylation pattern), information regarding transmembrane domains, presence of coiled-coil domains, similarity of the amino acid profile to that of known Icm/Dot effectors and strength of the Icm/Dot secretion signal[19]. The full list of features used is specified in **Supplementary Table 11**. These features served as input to four different machine learning classification algorithms: (i) naive Bayes[55]; (ii) Bayesian networks[56]; (iii) support vector machine (SVM)[57]; and (iv) random forest[58]. The final prediction score of each ORF was calculated as a weighted mean of the prediction scores for the four classification algorithms, where the weights were based on the estimated performance of each algorithm. These performances were evaluated by the mean area under the precision-recall curve[59] over tenfold cross-validation. The prediction score and the Icm/Dot secretion signal score for each ORF are detailed in **Supplementary Table 10**.

The machine learning prediction procedure was performed separately for each genome to account for the unique effector characteristic in each species. The training set for each species comprised close homologs of validated effectors (selected on the basis of BLAST bit score ≥60; marked by "H" in **Supplementary Table 10**). These effectors were used to train the machine learning scheme that performed predictions for each genome separately. The machine learning score threshold to consider an ORF as a putative effector was selected such that the set of effectors in each genome included at least one-third of the effectors from the training set but no more than one-third of the ORFs were newly predicted effectors. This is a conservative threshold: it increased the set of effectors over the training set by only 22.4% (ORFs marked by "ML" in **Supplementary Table 10**). In addition, ORFs that were part of ortholog groups that included ≥80% effectors were also considered as effectors by orthology (marked by "O" in **Supplementary Table 10**). Only a very small fraction of the effectors (1.5%) were added because of orthology.

To test the significance of the different numbers of effectors encoded by the species across the major clades, an ANOVA test allowing the comparison of means across numerous groups was used. The percentages of ORFs that encoded effectors were compared across the clades represented by different colors in **Figure 1**. The percentage of effectors in each genome was used for this comparison rather than the effector count to account for differences in genome length.

To estimate the total number of effectors in the 41 *Legionella* species, we performed a second round of learning, based on validated effectors and the putative effectors predicted by the strict species-specific thresholds in the first round of learning. This learning was performed on a combined data set including all the genomes to allow effectors detected in one species in the first round to aid in the identification of effectors in other species in this round of prediction (the scores are detailed in **Supplementary Table 10**). Because this learning is based on effectors that have not yet been validated, the predictions should be considered speculative but can be useful in estimating the total number of effectors in the genomes analyzed. The number of novel high-scoring predictions (score >0.99) was 2,643. On the basis of the false positive rate for the training set, we deduced that 744 additional effectors were scored below this threshold. Combining these numbers with the 5,885 putative effectors used as an input to this learning led us to estimate that the total number of effectors is approximately 9,272.

Effector repertoire similarity was calculated as the mean of (i) the fraction of effector ortholog groups shared by species *i* and *j* out of all the effector ortholog groups represented in species *i* and (ii) the fraction of effector ortholog groups shared by species *i* and *j* out of all the effector ortholog groups represented in species *j*.

Effector pseudogenes were identified by performing a strict BLAST search (e value $\leq 1 \times 10^{-10}$) of all putative effector genes against the sequenced genomes. Pseudogenes were identified if they covered at least 80% of the homologous effector and had between one and four stop codons that split the effector into two or more parts such that the second largest ORF was greater than 20% of the length of the homologous gene.

Phylogeny reconstruction and evolutionary analyses. An initial evolutionary tree was reconstructed on the basis of concatenated alignments of proteins belonging to 93 ortholog groups that had one ortholog per *Legionella* species and have been reported to be nearly universal in Bacteria[15]. For each one of these 93 proteins, a separate tree was also reconstructed and compared to the concatenated tree using the AU test[60] on the protein's multiple-sequence alignment. The AU test (approximately unbiased test) determines whether an observed multiple-sequence alignment is significantly better supported by one of two maximum-likelihood phylogenies. Fifteen proteins with gene trees that were significantly different from the combined phylogeny (AU test, $P < 0.01$, after false detection rate (FDR) correction for multiple testing[61]) were filtered out. The final phylogeny was achieved by reconstructing an evolutionary tree

doi:10.1038/ng.3481

based on the concatenated alignment of the remaining 78 nearly universal single-copy proteins (marked in **Supplementary Table 3**).

To test whether the gene trees of the seven core effectors agreed with the genus phylogeny, each of these trees was compared with the species tree (**Supplementary Fig. 4**). With the exception of LOG_01106, the core effector trees did not differ significantly from the species tree (AU test[60], $P < 0.01$, after FDR correction). Close examination of the LOG_01106 gene tree indicated that the evolutionary history of this effector agreed with the species tree in general, and the incongruent splits were not well supported by bootstrap values (**Supplementary Fig. 4**).

The evolutionary rates of gain and loss of effectors were computed as described in Cohen *et al.*[62]. Briefly, the phyletic pattern of effectors was coded as a gapless alignment of 0's and 1's representing, respectively, the absence or presence of specific effector families in each of the analyzed genomes. The gain and loss rates for effectors were inferred using a maximum-likelihood framework allowing for a gene-specific variable gain/loss ratio. Branch-specific gain and loss events were inferred by a stochastic mapping approach that accounts for tree topology, branch lengths, effector-specific evolutionary rates and the posterior probability of the presence of the effector at each node of the tree.

To determine whether the presence of the core effectors LOG_00341 (Lpg2832) and LOG_01049 (RavC) in a few bacterial species outside of the genus was the result of HGT events or multiple loss events, we compared the likelihood of two probabilistic models: one model allowing both gains (HGTs) and losses and a simpler model allowing only loss events. We tested these two models across an extensive set of 1,165 microbial genomes, with phylogeny based on the MicrobesOnline species tree[63], by running the probabilistic GLOOME algorithm[64] with a 'gain and loss' model and a 'loss-only' model. The statistical significance of the difference between the models was calculated by a likelihood-ratio test comparing the maximum log likelihoods obtained for the two alternative models.

Co-evolution of syntenic effectors was estimated on the basis of the extent to which pairs of effectors were gained and lost together during their evolutionary histories among the *Legionella* genomes. We used the algorithm CoPAP[65] that is based on a probabilistic framework in which gain and loss events are stochastically mapped onto the phylogeny. The significance of co-evolutionary interaction is estimated by computing the degree of shared gain and loss events in the evolutionary histories of the effector pair analyzed while accounting for the overall evolutionary dynamics of gain and loss.

The evolutionary histories of putative effectors that underwent HGT within the *Legionella* genus were analyzed by reconstructing the gene trees for 96 effectors that were present in 4–40 species, that had no paralogs and that were inferred to be gained at least twice during genus evolution. These trees were compared to the species tree using the AU test[60], and the *P* values obtained were corrected for multiple testing using FDR[61].

All phylogenetic trees were reconstructed using RAxML[66] under the LG + GAMMA + F evolutionary model with 100 bootstrap resamplings. AU test *P* values were calculated using CONSEL[67]. *C. burnetii* was used as an outgroup to root the tree.

Plasmid construction. To generate deletion substitutions in the seven *L. pneumophila* core effectors, a 1-kb DNA fragment on each side of the planned deletion was amplified by PCR using the primers listed in **Supplementary Table 12**. The primers were designed to contain a SalI site at the position of the deletion. The two fragments amplified for each gene were cloned into pUC-18 digested with suitable enzymes to generate the plasmids listed in **Supplementary Table 12**. The resulting plasmids were digested with suitable enzymes, and the inserts were used for a four-way ligation containing the kanamycin resistance cassette (Pharmacia) digested with SalI and the pUC-18 vector digested with suitable enzymes. The desired plasmids were identified by plating transformed bacteria on plates containing ampicillin and kanamycin, and after plasmid preparation the desired clones were identified by restriction digests. The plasmids generated (**Supplementary Table 12**) were digested with PvuII (which cuts on both sides of the pUC-18 polylinker), and the resulting fragments were cloned into the pLAW344 allelic exchange vector digested with EcoRV to generate the plasmids that were used for allelic exchange (listed in **Supplementary Table 12**), as described previously[68].

To construct isopropyl β-D-1-thiogalactopyranoside (IPTG)-inducible effectors, the *L. pneumophila legA3* and *mavN* genes were amplified by PCR using the primers listed in **Supplementary Table 12**. The PCR products were then digested with BamHI and SalI for *legA3* and with EcoRI and BamHI for *mavN* and cloned into pMMB207C downstream of the P*tac* promoter to generate the plasmids listed in **Supplementary Table 12**. These plasmids contain the effectors under P*tac* control, and they were used for intracellular growth complementation assays. Intracellular growth assays of *L. pneumophila* strains in *A. castellanii* were performed as previously described[69].

Identification of *Legionella* effector domains. Each LEOG was represented by a hidden Markov model (HMM), which was constructed as follows. The proteins belonging to a given ortholog group were aligned by MAFFT[70] version v7.164b using the 'einsi' strategy. HMMs were constructed from the multiple-sequence alignments using hmmbuild from the HMMER suite[71] version 3.1b1.

Characterized domains were identified by comparing LEOG HMMs to domain databases using hhsearch version 2.0.15 from the HH-suite[72]. Specifically, an *e*-value threshold of 1×10^{-5} for the hhsearch was used to identify similarities between the LEOG HMMs and HMMs derived from the following databases: (i) the NCBI Conserved Domain Database (CDD)[73], (ii) Pfam[74] and (3) SMART[75], downloaded from the HH-suite ftp site. Resulting hits were manually curated to filter out domains of unknown function and non-informative domains. Additional characterized domains were identified during the process of novel domain detection.

Novel domains were identified as follows. All-against-all BLAST[76] search of all 5,885 putative *Legionella* effectors was performed with an *e*-value cutoff of 0.001. From the BLAST hits that received a bit score >40, we extracted maximal joined segments longer than 50 amino acids that were nearly non-overlapping (overlap of <10 amino acids). The extracted segments were searched using BLAST against the putative effector data set with a threshold bit score of 40. Segments that had four or more hits were aligned and used to construct HMMs (as described above). These HMMs, representing conserved domains, were compared to each other using hhsearch. HMMs with a homology probability score ≥95% and an *e* value <0.01 across at least 50% of their length were designated as describing the same domain. The HMMs for the detected domains were scanned for coiled-coil domains using COILS[77], and domains that were ≥80% covered by coiled-coil domains were labeled as coiled-coiled domains. The domain HMMs were further scanned against the HMM databases of CDD[73], Pfam[74] and SMART[75], and those with a homology probability score ≥95% and an *e* value <0.01 across at least 50% of their length were annotated according to the characterized domain (after excluding non-informative hits). The domain HMMs were used to scan the data set of putative effectors. A domain was considered to be a novel *Legionella* effector domain if it did not overlap any characterized domain and appeared in at least 80% of the members of two different ortholog groups, each composed of at least two putative effectors.

In the effector domain network, each node represents an architecture, that is, a combination of domains that was present in the same effector. An edge between two architecture nodes represents a domain that is shared by the two architectures. The size of each node is proportional to the number of putative effectors that had the architecture represented by the node. The network was visualized using the igraph package[78] of R (ref. 79). The mapping of domain architectures on the species tree was visualized using iTOL[80].

50. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
51. Zerbino, D.R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
52. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
53. Li, L., Stoeckert, C.J. Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
54. Wu, M. & Eisen, J.A. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* **9**, R151 (2008).
55. Langley, P., Iba, W. & Thompson, K. in *Proc. 10th Natl. Conf. Artificial Intelligence* 223–228 (AAAI Press, 1992).
56. Heckerman, D., Geiger, D. & Chickering, D.M. Learning Bayesian networks: the combination of knowledge and statistical data. *Mach. Learn.* **20**, 197–243 (1995).

57. Burges, C.J.C. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **2**, 121–167 (1998).
58. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
59. Davis, J. & Goadrich, M. in *Proc. 23rd Int. Conf. Machine Learning* 233–240 (ACM, 2006).
60. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).
61. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
62. Cohen, O. & Pupko, T. Inference and characterization of horizontally transferred gene families using stochastic mapping. *Mol. Biol. Evol.* **27**, 703–713 (2010).
63. Dehal, P.S. *et al.* MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.* **38**, D396–D400 (2010).
64. Cohen, O., Ashkenazy, H., Belinky, F., Huchon, D. & Pupko, T. GLOOME: gain loss mapping engine. *Bioinformatics* **26**, 2914–2915 (2010).
65. Cohen, O., Ashkenazy, H., Levy Karin, E., Burstein, D. & Pupko, T. CoPAP: coevolution of presence-absence patterns. *Nucleic Acids Res.* **41**, W232–W237 (2013).
66. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
67. Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246–1247 (2001).
68. Segal, G. & Shuman, H.A. Characterization of a new region required for macrophage killing by *Legionella pneumophila*. *Infect. Immun.* **65**, 5057–5066 (1997).
69. Segal, G. & Shuman, H.A. *Legionella pneumophila* utilizes the same genes to multiply within *Acanthamoeba castellanii* and human macrophages. *Infect. Immun.* **67**, 2117–2124 (1999).
70. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
71. Eddy, S.R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
72. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2012).
73. Marchler-Bauer, A. *et al.* CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* **30**, 281–283 (2002).
74. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–D301 (2012).
75. Letunic, I., Doerks, T. & Bork, P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* **40**, D302–D305 (2012).
76. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
77. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164 (1991).
78. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* **1695**, 1–9 (2006).
79. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2013).
80. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**, W475–W478 (2011).