Computational modeling and experimental validation of the *Legionella* and *Coxiella* virulence-related type-IVB secretion signal

Ziv Lifshitz[a,1], David Burstein[b,1], Michael Peeri[b], Tal Zusman[a], Kierstyn Schwartz[c], Howard A. Shuman[c], Tal Pupko[b,2], and Gil Segal[a,2]

Departments of [a]Molecular Microbiology and Biotechnology and [b]Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel; and [c]Department of Microbiology, University of Chicago, Chicago, IL 60637

Edited by Thomas J. Silhavy, Princeton University, Princeton, NJ, and approved December 28, 2012 (received for review September 5, 2012)

Legionella **and** ***Coxiella*** **are intracellular pathogens that use the virulence-related Icm/Dot type-IVB secretion system to translocate effector proteins into host cells during infection. These effectors were previously shown to contain a C-terminal secretion signal required for their translocation. In this research, we implemented a hidden semi-Markov model to characterize the amino acid composition of the signal, thus providing a comprehensive computational model for the secretion signal. This model accounts for dependencies among sites and captures spatial variation in amino acid composition along the secretion signal. To validate our model, we predicted and synthetically constructed an optimal secretion signal whose sequence is different from that of any known effector. We show that this signal efficiently translocates into host cells in an Icm/Dot-dependent manner. Additionally, we predicted** ***in silico*** **and experimentally examined the effects of mutations in the secretion signal, which provided innovative insights into its characteristics. Some effectors were found to lack a strong secretion signal according to our model. We demonstrated that these effectors were highly dependent on the IcmS-IcmW chaperons for their translocation, unlike effectors that harbor a strong secretion signal. Furthermore, our model is innovative because it enables searching ORFs for secretion signals on a genomic scale, which led to the identification and experimental validation of 20 effectors from** ***Legionella pneumophila***, ***Legionella longbeachae***, **and** ***Coxiella burnetii.*** **Our combined computational and experimental methodology is general and can be applied to the identification of a wide spectrum of protein features that lack sequence conservation but have similar amino acid characteristics.**

pathogenomics | type-IV secretion | translocated substrates | Legionnaire disease | Q-fever

The bacteria *Legionella pneumophila*, the causative agent of Legionnaire disease, and *Coxiella burnetii*, the causative agent of Q-fever, are both human intracellular pathogens that multiply in alveolar macrophages (1, 2). In nature, *L. pneumophila* multiplies in a broad range of amoebae, whereas *C. burnetii* infects various ruminants, such as cattle and sheep (3, 4). The intracellular vacuole formed by these two bacteria has been shown to be completely different; *L. pneumophila* inhibits phagosome-lysosome fusion and resides in a vacuole at almost neutral pH, whereas *C. burnetii* multiplies in an acidic vacuole (5–8). However, these two pathogens have been shown to use a similar Icm/Dot type-IVB secretion system for pathogenesis (9–12).

The Icm/Dot secretion systems of these bacteria have been shown to translocate a large number of effector proteins from the bacterial cytoplasm into the host cell during infection (the current number of effectors in *L. pneumophila* and *C. burnetii* is estimated as 300 and 100, respectively). The effector proteins subvert a wide repertoire of functions in the host cells and direct the establishment of the unique phagosome formed by these two pathogens (5, 6). Most of the effectors translocated by these two pathogens are unique to one of them, but a few were shown to contain similar protein motifs, such as ankyrin domains (13, 14).

In order for an effector to be translocated into the host cell cytoplasm via the Icm/Dot secretion system, the effector must be recognized by components of the secretion system. The first effector identified, RalF, was shown to harbor a C-terminal secretion signal (15), and additional analyses of this effector indicated that one of its three C-terminal amino acids should be hydrophobic in order for it to translocate (16). Further analysis using several effectors indicated that in addition to the hydrophobic amino acid described, the C-terminal secretion signal is enriched with tiny, polar, and charged amino acids (e.g., alanine, serine, threonine, glutamic acid) (17). When the number of known *L. pneumophila* effectors increased to about 100, an analysis of the C-terminal region of effectors indicated there are amino acids that are enriched and/or depleted in this region. Glutamic acid was found to be enriched at positions −17 to −10 (from the C-terminal end) and depleted from the five C-terminal amino acids, which were enriched with hydrophobic amino acids (e.g., isoleucine, leucine, valine). In addition, serine and threonine were enriched at positions −10 to −5 (18). A more recent study (19) described the importance of the glutamic acid stretch (E-block) for effector translocation and used this feature to identify several novel effectors. Notably, all these studies were performed only on *L. pneumophila*.

In addition to the C-terminal secretion signal described above, the IcmS and IcmW proteins were found to function as a chaperon complex that assists in the translocation of effector proteins into host cells (20). Many effectors were shown to have a reduced level of translocation in the absence of this chaperon complex, but others were not affected (16). The current information suggests that both the IcmS–IcmW chaperon complex and the C-terminal secretion signal jointly contribute to the translocation efficiency of the Icm/Dot effectors (21, 22).

The purpose of this research was to characterize the secretion signal of Icm/Dot effectors. To achieve this goal, we have implemented a hidden semi-Markov model (HSMM). The HSMM was trained on all 283 *L. pneumophila* known effectors and a set of noneffectors to identify sequence properties that are unique to

Author contributions: Z.L., D.B., M.P., T.Z., K.S., H.A.S., T.P., and G.S. designed research; Z.L., D.B., M.P., T.Z., and K.S. performed research; Z.L., D.B., M.P., T.Z., K.S., H.A.S., T.P., and G.S. contributed new reagents/analytic tools; Z.L., D.B., M.P., T.Z., K.S., H.A.S., T.P., and G.S. analyzed data; and Z.L., D.B., T.P., and G.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]Z.L. and D.B. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: gils@tauex.tau.ac.il or talp@tauex.tau.ac.il.

See Author Summary on page 2709 (volume 110, number 8).

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1215278110/-/DCSupplemental.

the C terminus of effectors. We have used the trained model to predict and experimentally examine a synthetic "optimal" secretion signal (OSS) that was found to translocate into host cells efficiently in an Icm/Dot-dependent manner. Furthermore, we predicted several mutations according to the HSMM and validated their effect on the translocation of the secretion signal experimentally. Finally, the HSMM was used to identify effectors in *L. pneumophila*, *Legionella longbeachae*, and *C. burnetii*.

Results

To characterize the C-terminal secretion signal of Icm/Dot effectors, we have implemented an HSMM. Based on a set of known *L. pneumophila* effector proteins and a set of noneffectors, two HSMMs were trained: one that best characterizes the C terminus of effectors and one that best characterizes the C terminus of noneffectors. Contrasting these two models allowed characterization of the physicochemical properties of the secretion signal of effectors. We chose the HSMM over position-specific score matrices (23) because the HSMM is more general; it accounts for dependencies among sites, and thus can better capture spatial variation in amino acid composition along the secretion signal (24, 25).

HSMM Analysis for Identification of the Effectors' Secretion Signal. The HSMM we implemented consists of different states, each representing one or more positions of the C-terminal signal of effectors. Based on a given set of effectors, the model learned for each state: (*i*) how many positions it represents and (*ii*) what is the probability of observing each amino acid in that state (*Materials and Methods*). There are two additional important attributes that are not parameters of the model itself: the length of the C-terminal signal and the number of states in the model. To determine the most suitable number of states and signal length, we thus tested the performance of the HSMM on different combinations of signal length and states. The performance of the model with the different parameter values (*Materials and Methods*) is summarized in Table S1. The best-performing HSMM consisted of 22 states that model the 35 C-terminal amino acids of effectors, and these parameters were used in all subsequent analyses.

Using an effector model based on all 283 *L. pneumophila* effectors known to date (Dataset S1) and a background model based on a set of 1,000 *L. pneumophila* noneffector proteins (Dataset S2), the signal scores of all known effectors were computed. This score represents the odds that the C terminus fits the effectors model vs. a null background model. Low-scoring effectors were removed from the set of effectors to ensure that only effectors likely to harbor a strong secretion signal are represented in the final model (*Materials and Methods*). This final model was used for all subsequent analyses.

Effectors That Obtain a Low Signal Score and Translocate Efficiently Are Affected by the IcmS–IcmW Chaperon Complex. The final HSMM scoring of the *L. pneumophila* effectors revealed that 191 effectors received a score above 2 and 92 effectors received a score lower than 2 (Dataset S1), indicating that their signal is less than seven (e^2)-fold more similar to the effectors' signal model in comparison to the background model. This observation led us to examine the published information about this group of effectors regarding their degree of translocation into host cells (17–19, 26, 27). It was found that 54 of these effectors (Dataset S1) have been shown previously to translocate into host cells at low or medium levels, and this result fits with their low signal score. However, 19 of these effectors (Dataset S1) were shown to have high translocation levels, which seem to contradict their low signal score. Therefore, we examined the effect of the IcmS–IcmW chaperon complex on the translocation efficiency of 10 of these 19 effectors. These 10 effectors, which received a signal score between 0.8 and −3.7 (Dataset S1), were examined for translocation from the *L. pneumophila* WT strain in comparison to the *icm*S-*icm*W double-deletion mutant. The results obtained show that the translocation efficiency of all these effectors was strongly reduced in the *icm*S-*icm*W double-deletion mutant (Fig. 1) but that their level of expression was similar to the one in the WT strain (Fig. S1). For comparison, we also examined the effect of the IcmS–IcmW chaperon complex on the translocation efficiency of the two top-scoring effectors (lpg1144-CegC3 with a signal score of 16.5 and lpg1290-Lem8 with a signal score of 13.8). The translocation of these two effectors was not affected by the absence of the IcmS–IcmW chaperon complex (CegC3 and Lem8 in Fig. 1). Altogether, this analysis reveals the involvement of both the C-terminal secretion signal and the IcmS–IcmW chaperon complex in the translocation of Icm/Dot effectors (*Discussion*).

A Protein Carrying a Synthetic Signal Predicted by the Model Is Efficiently Translocated into Host Cells in an Icm/Dot-Dependent Manner. One of the advantages of the computational approach used in this study is its ability to predict the amino acid sequence of the OSS by inferring the probability of occurrence of each amino acid in each position of the signal (*Materials and Methods*). Examination of the computational analysis revealed that amino acids with similar properties were found at high probability in the same positions (Fig. S2 and Dataset S3). Therefore, we grouped together amino acids that have similar physicochemical properties (ILVF, ED, RK, QN, and TS). For each position along the sequence, the amino acid group that received the highest probability of occurrence was selected, and within each group, the most probable amino acid was selected to represent the group (Dataset S3). The OSS that resulted from this process (Fig. 2*A*) had no obvious sequence similarity to any known effector (the OSS shares only 20% identical and 31% similar amino acids with lpg1144, the effector that received the highest score by the HSMM). The probabilities of occurrence of the group of amino acids in each position are presented in Fig. 2*A*. To determine if the OSS is sufficient to obtain translocation into host cells, it was fused to the CyaA reporter and introduced into the *L. pneumophila* WT and mutant strains. As can be seen in Fig. 2*B*, the OSS showed a high level of translocation, and this translocation was found to be completely dependent on the Icm/Dot secretion system because no translocation was observed from the *icm*T or the *dot*A deletion mutants. These results validate the computational model, they demonstrate the importance of the C-terminal secretion signal for translocation, and they show that the computationally predicted secretion signal is sufficient to direct the CyaA reporter to the Icm/Dot type-IVB secretion apparatus and to result with translocation.

Mutations in Key Residues of the OSS Affect the Level of Translocation. In an effort to identify the most significant amino acids in the OSS, we used our computational model to predict which single amino acid mutation has the most deleterious effect on the translocation efficiency of the OSS. Interestingly, this mutation was predicted to be a leucine-to-glutamic acid change of the most C-terminal amino acid of the OSS (Fig. 3*A*, M1). Examination of translocation into host cells of an OSS CyaA fusion containing this mutation clearly demonstrated that this single change completely abolished translocation (Fig. 3*B*, M1). In line with this result, we found that only a single known effector (lpg2832) contains a glutamic acid as its last amino acid. In addition, only 11 effectors (of 283) contain a glutamic acid in one of the three C-terminal amino acids, strongly indicating that glutamic acids are depleted from these positions.

This result led us to examine the effect of changing additional C-terminal amino acids to glutamic acid. Specifically, we changed the second, third, and fifth positions from the C terminus to glutamic acid (in the OSS, both the forth and fifth amino acids are represented by the same state of the HSMM; Fig. 3*A*, M2, M3, and

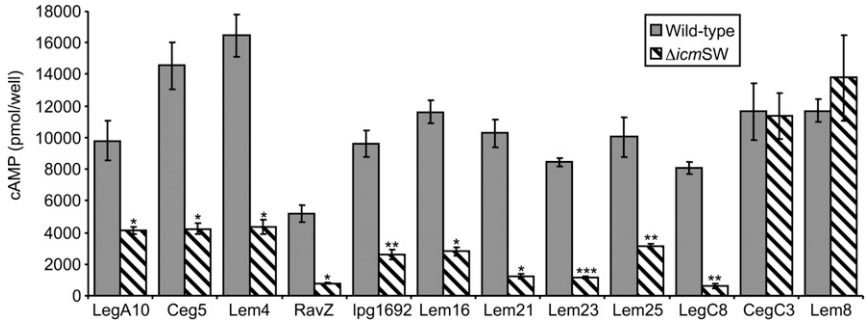

Fig. 1. The IcmS–IcmW chaperon complex affects the translocation of effectors that received a low score by the HSMM. The WT strain JR32 (gray bars) and the *icmS-icm*W double-deletion mutant ED400 (diagonal striped bars) harboring the CyaA fusion proteins indicated below each bar were used to infect HL-60–derived human macrophages, and the cAMP levels of the infected cells were determined as described in *Materials and Methods*. The bar heights represent the means of the amount of cAMP per well obtained in at least three independent experiments; error bars indicate SDs. The cAMP levels of each fusion were found to be significantly different (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$, paired Student *t* test), comparing the WT strain and the *icmS-icm*W double-deletion mutant.

M4, respectively, and Fig. S2). Mutation M3 resulted with lack of translocation, similar to mutation M1, whereas mutation M2 resulted with translocation, but to a low degree (Fig. 3*B*). The decrease in translocation was in agreement with the probabilities of occurrence of the amino acid groups in these positions (Fig. 2*A*). Mutation M4 resulted with an intermediate level of translocation, but when a double mutation of both the forth and fifth C-terminal amino acids (Fig. 3*A*, M5) was examined, it resulted with a low level of translocation (Fig. 3*B*, M5). These results fit the model because both of these residues are represented in the HSMM by the same state and a change of one of them is predicted to have a modest effect on translocation. As mentioned above, the mutations to glutamic acids constructed in the first and third positions from the C terminus resulted with lack of translocation (Fig. 3, M1 and M3). We constructed two additional mutations in these positions (Fig. 3*A*, M6 and M7) that were predicted by the model to have a less severe effect on translocation in comparison to the change to glutamic acid. Indeed, these two mutations had intermediate levels of translocation (Fig. 3*B*, M6 and M7). Collectively, the mutations tested reveal key amino acids of the Icm/Dot type-IVB secretion signal, and their translocation levels are in excellent agreement with the model representation of the OSS. Although glutamic acid residues are highly abundant in specific positions of the signal, their introduction into different positions within the secretion signal C-terminal end results with a dramatic reduction in translocation.

The Location of the Glutamic Acid Stretch Is Critical for Efficient Translocation. As indicated in the introductory section, it has previously been shown that a sequence containing several residues of glutamic acids (called "E-block") is important for effector translocation (19). This motif was also found to be present in the OSS (Fig. 3*A*). We noticed that when a greater number of the most deleterious mutations were predicted by the HSMM (decuple mutations; Fig. 3*A*, M8) a stretch of five glutamic acids was reintroduced to the OSS in the five C-terminal positions and the original glutamic acid stretch was mostly mutated. This observation led us to examine whether such a mutated OSS can direct translocation of the CyaA reporter into host cells because it contains an E-block. The results clearly show that no translocation was obtained with this mutated OSS (Fig. 3*B*, M8). These results convincingly show that the position of the glutamic acid stretch in the secretion signal of the Icm/Dot effectors is critical for translocation.

Scoring of the *Legionella* and *C. burnetii* Genomes Using the HSMM. In addition to the prediction of the OSS, the model implemented allows scoring each ORF in the genome for its likelihood to contain an Icm/Dot secretion signal. The signal scores of the 300 top-scoring ORFs in six bacterial genomes, *L. pneumophila*, *L. longbeachae*, *Legionella drancourtii*, *Legionella dumoffii, C. burnetii*, and, as a control, *Escherichia coli* are presented in Fig. 4. It is evident that in *E. coli*, which does not harbor an Icm/Dot secretion system, the top-scoring ORFs are dramatically lower in comparison to *Legionella* and *C. burnetii*. Of the 4,144 ORFs in

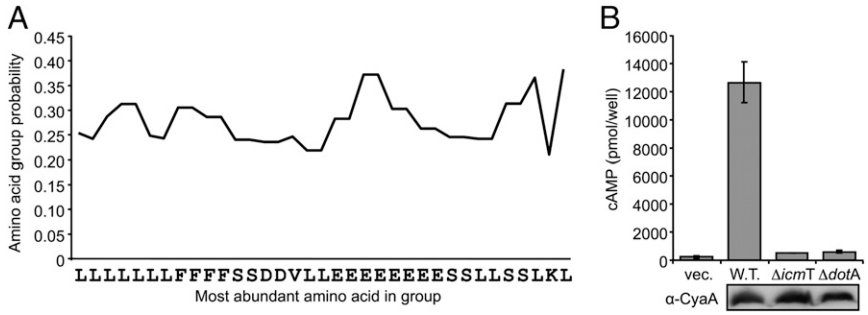

Fig. 2. The synthetic signal predicted by the HSMM translocates a protein into host cells in an Icm/Dot- dependent manner. (*A*) Amino acids of the synthetic signal were divided into five groups (ILVF, DE, RK, ST, and QN) according to their physicochemical properties. Each group of amino acids is represented by the amino acid in the group that received the highest probability of occurrence in each position, and the probability of occurrence of each amino acid group is presented. (*B*) CyaA protein fused to the synthetic signal was examined for translocation into HL-60–derived human macrophages from the WT strain JR32, the *icm*T deletion mutant GS3011, and the *dot*A insertion mutant LELA3118. vec., vector control. The bar heights represent the means of cAMP levels per well obtained from at least three independent experiments; error bars represent SDs. Protein levels were assessed by Western blot using α-CyaA antibody and are shown below each bar.

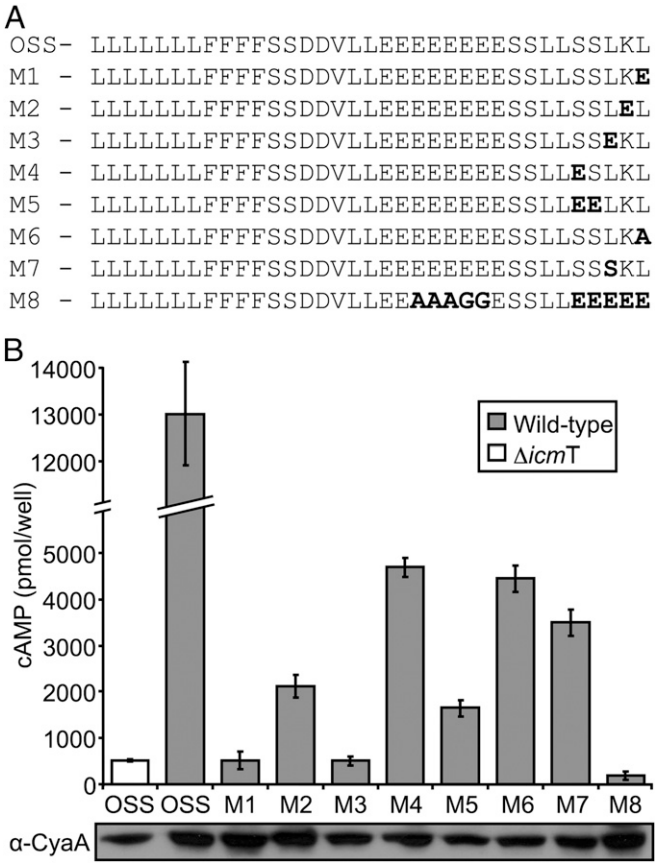

Fig. 3. Mutation analysis of the OSS. (*A*) Amino acid sequences of the OSS and the mutations constructed are shown. The mutated residues are marked in bold. (*B*) CyaA fusions of the OSS and the eight mutated OSSs were examined for translocation from the WT strain JR32 into HL-60–derived human macrophages. The translocation level of the OSS from the *icm*T deletion mutant is also presented. The bar heights represent the means of cAMP levels per well obtained from at least three independent experiments; error bars represent SDs. The cAMP levels were found to be significantly different ($P < 10^{-6}$, paired Student *t* test) when comparing the OSS and each of the mutants. Protein levels were assessed by Western blot using α-CyaA antibody and are shown below each bar.

E. coli, only 8 ORFs received a score above 5 (which indicates that an ORF is e^5 ≈146-fold more similar to the effectors model than to the background model), suggesting that when using this threshold to identify effectors, we should expect a false detection rate lower than 0.2%. The distributions of scores for ORFs from the *Legionella* genus were similar to each other, suggesting that approximately the same number of effectors is encoded in these genomes. The number of *C. burnetii* effectors that received a score above 5 was lower than that of *Legionella*. The relatively small size of the *C. burnetii* genome by itself (2,163 ORFs, compared with 2,943 ORFs in *L. pneumophila*) does not explain this difference in the number of putative effectors (χ^2 goodness-of-fit test, $P < 0.00086$). Based on these results, we provide a list of putative effectors in all these organisms (Dataset S4).

Validation of Predicted Effectors in *L. pneumophila*. According to the signal score analysis described above, we focused the experimental validation step of this study only on the 135 *L. pneumophila* ORFs that obtained a score above 5 (Dataset S4). These ORFs include 118 known effectors, which constitute 87% of the ORFs that received a score above the cutoff of 5. Seven of the remaining 17 ORFs showed homology to known proteins present in many bacteria that are unlikely to encode for effectors (Dataset S4); therefore, they were not examined further. The rest of the ORFs (10 ORFs) were annotated as hypothetical proteins, and they were examined for translocation into host cells using the CyaA translocation system. Seven of these ORFs were found to translocate into host cells in an Icm/Dot-dependent manner (Fig. 5*A* and Fig. S1), and all of them were expressed at similar levels in the WT strain JR32 and in the *icm*T deletion mutant (Fig. S1). The seven effectors identified were designated CetLp for C-terminal signal for effector translocation of *L. pneumophila* (Table S2). The identification of these *L. pneumophila* effectors establishes the ability to use the secretion signal score as a single feature to differentiate *L. pneumophila* Icm/Dot effectors from the rest of the ORFs in the genome.

Validation of Predicted Effectors in *L. longbeachae*. Because the Icm/Dot type-IV secretion system is conserved in all the *Legionella* species examined (28), we were interested to determine whether our computational approach can also predict effectors in other pathogenic *Legionella* species. Therefore, we decided to examine ORFs that received a score above 5 in *L. longbeachae* as a representative *Legionella* species. This pathogenic *Legionella* species has been shown previously to have different intracellular characteristics in comparison to *L. pneumophila* (29). Using the same secretion signal model described above, 128 *L. longbeachae* ORFs obtained a score above 5 (Dataset S4). From these ORFs, 4 were found to contain domains known to be present in *L. pneumophila* effector proteins [e.g., ankyrin domain, Ser/Thr kinase domain (30, 31)], and 40 additional proteins were found to have some homology (E-value less than 0.01) to known *L. pneumophila* effectors. Because we wished to test only ORFs that have no homology to known effectors, these ORFs were not examined further. In addition, 11 ORFs were found to contain homology to known proteins present in many bacteria and 10 additional ORFs were shorter than 100 aa in length (suggesting that they may not encode for functional proteins); these 21 ORFs were not examined further

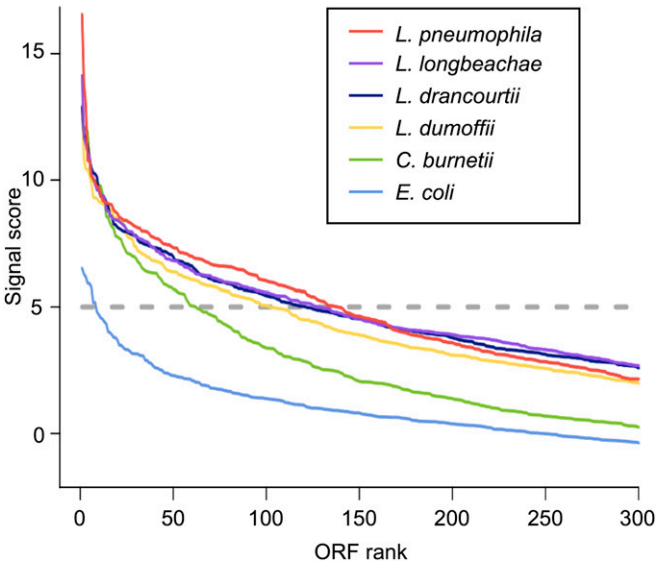

Fig. 4. Top-scoring ORFs from four *Legionella* species and *C. burnetii* according to the HSMM. The HSMM signal score reflects the agreement between the C terminus of each ORF and the effector model, while accounting for the background model trained separately for each bacterial species. The 300 top-scoring ORFs for the indicated *Legionella* species and *C. burnetii* are shown. The scoring of *E. coli* ORFs was used as a negative control for a bacterial species that does not contain an Icm/Dot secretion system. The horizontal gray dashed line represents the cutoff score of 5, which was used for the experimental validation of *L. pneumophila* and *L. longbeachae* ORFs.

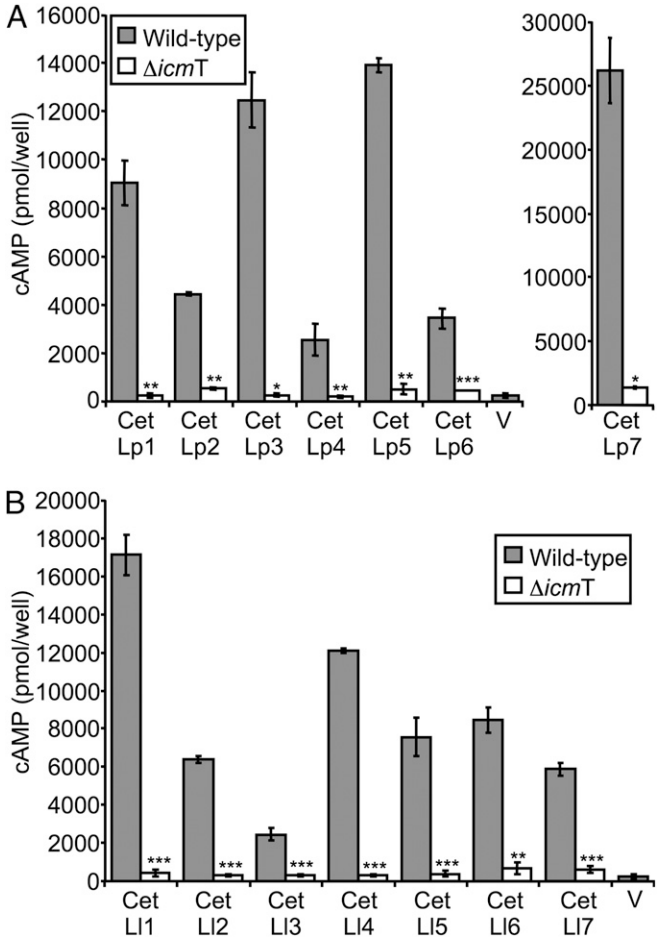

Fig. 5. Icm/Dot-dependent translocation of *Legionella* effector proteins predicted by the HSMM. CyaA fusions of predicted *L. pneumophila* (*A*) and *L. longbeachae* (*B*) effectors were examined for translocation into HL-60–derived human macrophages from the WT strain JR32 (gray bars) and the *icm*T deletion mutant GS3011 (white bars); the effectors examined are indicated below each bar. V, vector control. The cAMP levels of the infected cells were determined as described in *Materials and Methods*. The bar heights represent the means of the amount of cAMP per well obtained in at least three independent experiments; error bars indicate SDs. The cAMP levels of each fusion were found to be significantly different (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$, paired Student *t* test) between the WT strain and the *icm*T deletion mutant.

because it is unlikely that they encode for effectors (Dataset S4). The remaining 63 ORFs were all annotated as hypothetical proteins, and their HSMM score ranged from 14.1 to 5.0. Of these ORFs, we randomly picked 10 ORFs that received a wide range of scores (9.6 to 5.2) and examined them experimentally for translocation into host cells. Seven of these ORFs were found to translocate into host cells in an Icm/Dot-dependent manner (Fig. 5*B* and Fig. S1), and all of them were expressed at similar levels in the WT strain JR32 and in the *icm*T deletion mutant (Fig. S1). The *L. longbeachae* effectors identified were designated CetLl for C-terminal signal for effector translocation of *L. longbeachae* (Table S2). Notably, this study identifies effectors unique to *L. longbeachae*, and this result establishes the applicability of our approach to identify effectors in *Legionella* species other than *L. pneumophila*.

Validation of Predicted *C. burnetii* Effectors. Encouraged by the results obtained with the two *Legionella* species, we decided to extend our analysis to the genus *Coxiella*. *C. burnetii* has been shown to contain a functional Icm/Dot secretion system that translocates effector proteins into host cells (10, 32). Most of the *C. burnetii* effectors were validated using *L. pneumophila* as a translocation platform, suggesting that effectors from both bacteria share a similar secretion signal (13, 14, 33, 34). For the experimental examination of *C. burnetii* predicted effectors, a cutoff score of 6 (instead of 5) was used to improve the chances for accurate prediction. Forty-four *C. burnetii* ORFs obtained a score of 6 and above, 15 of which have been shown previously to encode for effector proteins (Dataset S4). Of the remaining 29 ORFs, 11 were annotated as pseudogenes, 2 were shorter than 100 aa, and 1 encodes for a transporter present in many bacteria; therefore, these 14 ORFs were not considered further (Dataset S4). The remaining 15 ORFs were all annotated as hypothetical proteins, and they were examined for translocation into host cells using *L. pneumophila* as a translocation platform. Six of these ORFs were found to translocate into host cells in an Icm/Dot-dependent manner (Fig. 6*A* and Fig. S1), and all were expressed at similar levels in the WT strain JR32 and in the *icm*T deletion mutant (Fig. S1). To determine if the effectors identified are expressed in *C. burnetii*, the levels of their mRNA were measured using RT-quantitative PCR (qPCR) with *C. burnetii* grown in axenic media, as well as in vivo in HEK 293T cells (*Materials and Methods*). All these effector-encoding genes were found to be expressed, and they had a similar expression pattern as two previously identified *C. burnetii* effectors, *cox*DFB1 and CBU2056 (10, 34) (Fig. 6*B*). These *C. burnetii* effectors were designated CetCb for C-terminal signal for effector translocation of *C. burnetii* (Table S2). These results indicate that *C. burnetii* and *L. pneumophila* effectors harbor a similar secretion signal and that this signal can be used as a single characteristic to identify *C. burnetii* effectors. The only other known genus that uses an Icm/Dot secretion system is *Rickettsiella*, which includes entomopathogenic bacteria (35). We examined the genome of *Rickettsiella grylli* using our model, and the high scoring predictions are presented in Dataset S4.

Discussion

Many bacterial pathogens use different types of secretion systems as their main virulence determinants. These systems translocate effector proteins into the host cell, and the effectors subvert numerous host cell processes during infection. In order for an effector to be translocated into host cells, it should be recognized by the secretion apparatus through which it is being translocated. In type-III and type-IV secretion systems, recognition of effectors by the secretion apparatus occurs in different ways, but the effectors in both systems usually contain two characteristics that participate in their recognition by the relevant secretion system. In type-III secretion systems, the signal for translocation of effectors was shown to be located at the N terminus of the effectors, and it constitutes a short amphipathic sequence (36, 37). The translocation of most of the type-III effectors also requires a chaperon, and its binding site on the effector protein was shown to be located adjacent to the secretion signal (38). In addition, type-III effector chaperons were often found to be specific for one or two effectors (39). In *Bartonella henselae*, which uses a type-IVA secretion system, the Bep effectors were also shown to contain a bipartite C-terminal signal. The C terminus of each effector harbors at least one copy of the Bep intracellular delivery domain and a short, positively charged signal sequence (40). In the *L. pneumophila* Icm/Dot type-IVB secretion system, the secretion signal was shown to be located at the C-terminal end of the effectors (16), and some of the effectors also require the IcmS–IcmW chaperon complex for efficient translocation (refs. 20, 21 and this study). The location and properties of the IcmS–IcmW chaperon binding domain in the effectors are currently not known.

The aim of our study was to characterize the properties of the secretion signal of the Icm/Dot effectors. Previously, several studies have found valuable information regarding the Icm/Dot secretion signal. The first indication of the Icm/Dot secretion

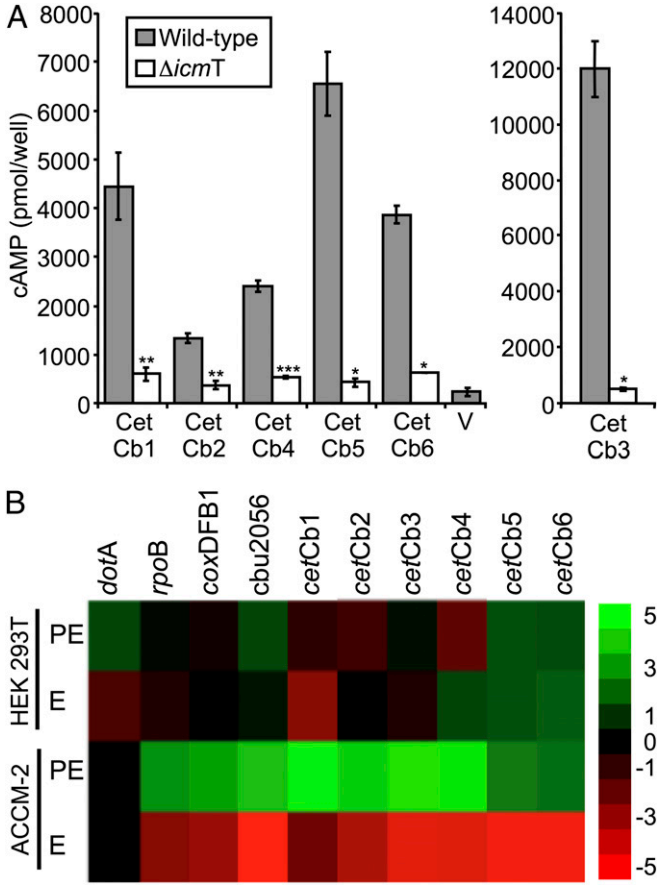

Fig. 6. Icm/Dot-dependent translocation of *C. burnetii* effectors identified by the HSMM and their expression during infection. (*A*) WT strain JR32 (gray bars) and the *icm*T deletion mutant GS3011 (white bars) harboring the CyaA fusion proteins (indicated below each bar) were used to infect HL-60–derived human macrophages, and the cAMP levels of the infected cells were determined as described in *Materials and Methods*. V, vector control. The bar heights represent the means of the amount of cAMP per well obtained in at least three independent experiments; error bars indicate SDs. The cAMP levels of each fusion were found to be significantly different (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$, paired Student *t* test) between the WT strain and the *icm*T deletion mutant. (*B*) Gene expression levels of *C. burnetii* effector genes during growth in ACCM-2 axenic medium and in HEK 293T cells. The heat map shows the logarithm of expression values of specific genes during growth in ACCM-2 media or in HEK 293T cells. Four genes were used as controls: *dot*A and *rpo*B, as well as *cox*DFB1 and cbu2056, two previously identified effectors. The values are the RNA levels normalized to 16S RNA during exponential (E) and postexponential (PE) phases. The RNA levels were measured by RT-qPCR and normalized to the levels of 16S RNA as described in *Materials and Methods*.

signal was the requirement of hydrophobic amino acids at the C-terminal end (16), and the requirement was then widened to include enrichment of tiny, polar, and charged amino acids (e.g., alanine, serine, threonine, glutamic acid) in the last 20 amino acids (17). Later, an analysis of the C-terminal region of effectors indicated that there are amino acids (e.g., glutamic acid, isoleucine, leucine, valine, serine, threonine) that are enriched and/or depleted in different regions of the secretion signal (18). Finally, the importance of glutamic acids (termed "E-block motif") for effector translocation was described (19). Our current model suggests that the previous results published concerning the secretion signal were all parts of the puzzle that constitutes the secretion signal of the Icm/Dot effectors. Our results indicate that as in type-III secretion systems, the secretion signal of Icm/Dot effectors does not contain specific amino acids at specific positions; rather, it contains amino acids with similar physicochemical properties located along the 35 C-terminal amino acids of the effectors. The most important characteristics of the secretion signal were found to be a large glutamic acid stretch at positions −10 to −17 and several hydrophobic residues located at its C-terminal end. Moreover, we found that the location of these amino acids along the secretion signal has a critical functional role. In addition to the characterization of the secretion signal, our results clearly demonstrate that three types of effectors exist in type-IVB secretion systems: (*i*) effectors that use mainly the C-terminal secretion signal for translocation (e.g., CegC3, Lem8 in this study), (*ii*) effectors that use both the C-terminal secretion signal and the IcmS–IcmW chaperon complex for translocation [e.g., LegS2, SidE (21, 22)], and (*iii*) effectors that use mainly the IcmS–IcmW chaperon complex for translocation (e.g., RavZ, LegC8 in this study). The third category of effectors might contain a yet unknown secretion signal; however, in any case, due to the reduced level of their translocation in the *icmS-icmW* double-deletion mutant, it seems that the contribution of such a signal to their translocation is minor. To substantiate the existence of this category of effectors, we examined in the *icmS-icmW* double-deletion mutant, the only effector that does not contain a C-terminal secretion signal [Lpg1368 (Lgt1)] (41), which received a very low score (−3.17) by the HSMM. This effector did not translocate from this mutant (Fig. S1), proving that its translocation is mainly dependent on the IcmS–IcmW chaperon complex. The three groups of effectors described above probably have different efficiencies of translocation, which might determine (together with other factors, such as their level of expression and stability) the hierarchy of their translocation into host cells.

Because the Icm/Dot complex components are conserved among different *Legionella* species (28, 42, 43), it was possible to use our model to predict effectors in *Legionella* species other than *L. pneumophila*. This analysis made it possible to evaluate whether the exceptionally large number of effectors present in *L. pneumophila* is common to other *Legionella* species or unique to *L. pneumophila*. Our results show that using an HSMM score cutoff of 5, 125 (92%) of the 135 *L. pneumophila* ORFs that received this score indeed encode for effectors. Examination of the three additional *Legionella* species (*L. longbeachae*, *L. drancourtii*, and *L. dumoffii*) for which a genomic sequence is currently available resulted in a similar number of ORFs that received a score higher than 5 (Fig. S3; excluding ORFs that are unlikely to encode for effectors). The genome of *L. pneumophila* encodes for 290 validated effectors, of which 125 received a score higher than 5. If the ratio between effectors that received a high signal score and the total number of effectors is maintained for the other *Legionella* species, our results indicate that a similar number of effectors as in *L. pneumophila* are likely to exist in the other *Legionella* species as well. It is important to note that our analysis was focused on the secretion signal of effectors; therefore, effectors that rely mainly on the IcmS–IcmW chaperon complex for their translocation were not predicted in the three *Legionella* species analyzed.

To explore the putative effectors identified in *L. longbeachae*, *L. drancourtii*, and *L. dumoffii* further, we examined their homology to known effector domains (e.g., ankyrin domain, Ser/Thr kinase domain), as well as for local homology with known *L. pneumophila* effectors (Fig. 7 *A*–*C*). This analysis indicated that in each of these *Legionella* species, about 70% of the predicted effectors have no sequence homology with *L. pneumophila* effectors or domains known to be present in the effectors of *L. pneumophila* (Fig. 7). To determine how many of the predicted effectors that received a score higher than 5 are species-specific, we performed a BLAST search for each of these ORFs against the genomes of the other three *Legionella* species described. A similar search was also performed for all the validated *L. pneumophila* effectors (Fig. 7*D*). To identify truly unique

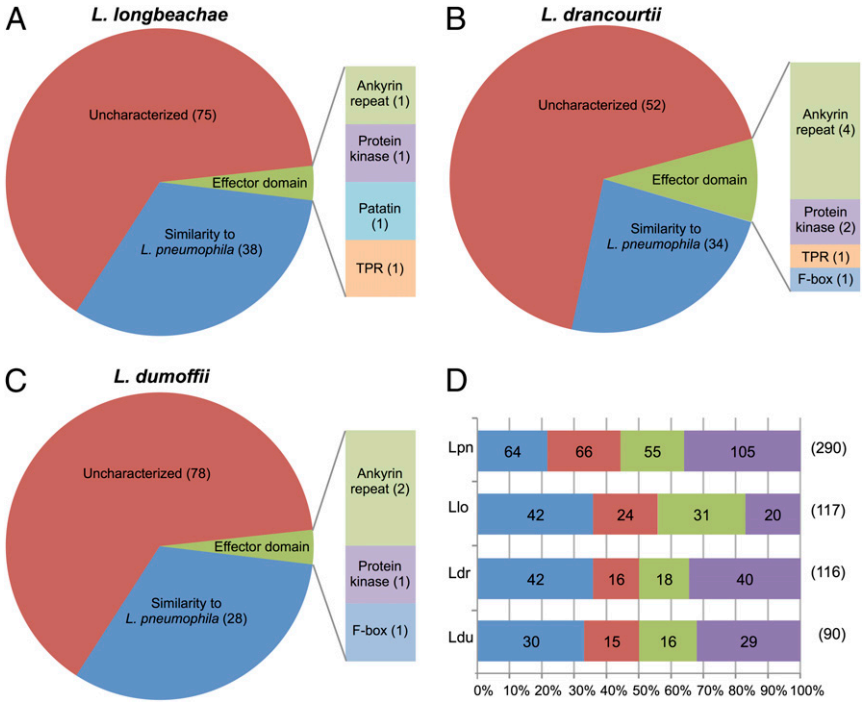

Fig. 7. Analysis of putative effectors in *L. longbeachae*, *L. drancourtii*, and *L. dumoffii*. The ORFs of *L. longbeachae* (*A*), *L. drancourtii* (*B*), and *L. dumoffii* (*C*) that received a signal score above 5 (excluding ORFs that are unlikely to encode for effectors; Dataset S4) were divided into three groups: (*i*) "Effector domain": putative effectors containing one of the domains known to characterize *L. pneumophila* effectors (the bars next to the "effector domain" specify which domains are present in the putative effectors), (*ii*) "Similarity to *L. pneumophila*": putative effectors that do not harbor such a domain but show local similarity to known *L. pneumophila* effectors, and (*iii*) "Uncharacterized": putative effectors containing neither a known effector domain nor similarity to known *L. pneumophila* effectors. The number of effectors in each category is indicated in parentheses. (*D*) Representation of the percentage of putative effectors that are unique (blue), as well as those that have homology in one (red), two (green), or all three (purple) other *Legionella* species. The top bar presents the same analysis for all 290 effectors of *L. pneumophila*. Ldr, *L. drancourtii*; Ldu, *L. dumoffii*; Llo, *L. longbeachae*; Lpn, *L. pneumophila*. The total number of predicted or validated (for *L. pneumophila*) effectors in each species is indicated in parentheses to the right. The number of effectors in each category is indicated inside the bars.

predictions, a tolerant E-score of 0.01 was used as a cutoff to determine similarity. About 30% (20% in *L. pneumophila*) of the predicted effectors in the three *Legionella* species examined were found to be species-specific. Taking into account that our experimental analysis of the *L. longbeachae* predicted effectors resulted in 70% prediction accuracy, we estimate that at least 50 species-specific effectors are present in each of the *Legionella* species. Because 58 different *Legionella* species are currently known (44, 45), an enormous number of bacterial effectors that subvert host cell processes are expected to be found in the *Legionella* genus.

The results presented suggest that effectors encoded by different species from the *Legionella* genus are expected to subvert an enormous number of host cell processes, probably reflecting the adaptation of each *Legionella* species to its specific niche (i.e., specific amoebae). Each *Legionella* species probably contains an arsenal of effectors that partially overlaps the one present in other *Legionella* species but also contains a significant number of species-unique effectors. Using the approach presented here, it will be possible to predict the pool of unique effectors for each *Legionella* species, as well as the pool of the *Legionella* genus conserved effectors, when additional *Legionella* genomic sequences become available.

Materials and Methods

HSMM. We implemented HSMMs (24, 25) to characterize the amino acid secretion signal of Icm/Dot effectors. The models have a linear topology without insertion or gap states (i.e., sequences must pass through all states of the model). The duration distribution used for each state in the HSMM was a discrete approximation of the gamma distribution (with two parameters). This provides the model more flexibility compared with a hidden Markov model (HMM). In an HMM, the duration distribution of each state is always exponential, which does not necessarily fit for modeling regions of the secretion signal. The parameters for the duration distribution and the amino acid emission probabilities for each state were optimized on the training set using Baum–Welch expectation-maximization (24, 25).

To determine the number of model states and the length of the C-terminal region that best characterize effectors, we trained the HSMMs for all valid combinations of number of states between 4 and 28 in steps of two and C-terminal lengths between 15 and 40 amino acids in steps of five (valid combinations are those in which the length is greater than the number of states). For each combination of these parameters, the performance of the HSMM was estimated using 10-fold cross-validation over a set of validated effectors (46). To overcome the possibility of local minima, each model was optimized using at least 400 random initial parameters, and the best model, determined by area under the receiver operating characteristic curve (AUC), was selected. The HSMM with the highest average AUC was the one with 22 states that modeled the 35 C-terminal amino acids (Table S1). These values of the number of states and signal length were used in all subsequent analyses.

Scoring ORFs Using the HSMM. The score of an ORF is computed as the natural logarithm of the ratio between (*i*) the likelihood of observing the 35 C-terminal amino acids of the ORF according to the HSMM trained on effectors and (*ii*) the likelihood of observing this C terminus according to the background HSMM. The effector HSMM was trained on a set of all 283 *L. pneumophila* effectors known to date (Dataset S1), using pseudocounts based on a prior distribution from the amino acid frequencies in all *L. pneumophila* proteins, with a prior probability of 0.05 (23). The background HSMM was trained with the same parameters using a training set of proteins with known function that are not likely to be effectors. For each bacterial species, a separate background model was trained to account for possible differences of the amino acid content in the C termini. The sets of proteins used for the training of the background models of *L. pneumophila*, *L. longbeachae*, *L. drancourtii*, *L. dumoffii*, *C. burnetii*, *R. grylli*, and *E. coli* are specified in Dataset S2. Both the effector HSMM and the background HSMM for each organism were trained 1,000 times, each with a random set of initial parameters. The model chosen was the one with

the highest likelihood on the relevant training set. The signal score was calculated as follows: Let $L(ORF|eff)$ be the likelihood of observing the C terminus of an ORF given the effector HSMM, and let $L(ORF|bk_G)$ be the likelihood of the same ORF given the background model constructed according to genome G; the signal score of this ORF is $Score(ORF,G) = \ln [L(ORF \mid eff)/L(ORF \mid bk_G)]$.

Scoring all 283 effectors revealed that 7.1% (30 effectors) received a negative score, suggesting they fit the background model better than the effector termini model. We hypothesized that low-scoring ORFs do not contain a C-terminal signal and that they are translocated using a different mechanism (*Results*). Such effectors are expected to introduce noise in the training process of the effector model. To obtain a model that is based solely on effectors likely to harbor a C-terminal signal, we have trained HSMMs using only the 200 effectors that received a score higher than 2. The trained model that received the highest likelihood on the reduced training set was used to score these effectors again. Of the 200 effectors, 192 received a signal score higher than 2 in the new model. This process was iterated until all the effectors used to train the model received a score higher than 2. The model meeting this criterion was based on 190 effectors, and this model was used in all subsequent analyses.

Prediction of the OSS and Most Deleterious Mutations. The sequence of the OSS was inferred from the model by following the most likely path of 35 amino acids through the trained HSMM. The emission probabilities of each amino acid for each state occupying each position along this path were extracted (Fig. S2 and Dataset S3). To account for the physicochemical properties of the amino acids, the emitted amino acids were clustered into five groups: (*i*) positively charged (lysine and arginine), (*ii*) negatively charged (aspartic acid and glutamic acid), (*iii*) polar amino acids bearing a hydroxyl group (serine and threonine), (*iv*) polar amino acids bearing an amide group (asparagine and glutamine), and (*v*) hydrophobic (leucine, isoleucine, valine, and phenylalanine). The OSS was constructed by choosing the group with the highest probability for each state and the amino acid with the highest emission probability from within that group (Fig. S2 and Dataset S3).

The mutations that are most deleterious to the OSS were predicted by scoring different perturbations of the optimal signal using the HSMMs. All the relevant single and double mutations were scored. A recursive heuristic approach was applied to search for the lowest scoring sequence with *i* mutations for *i* between 3 and 10. Specifically, we first computed the 1,000 lowest scoring sequences containing exactly two mutations. From these 1,000 sequences and using the HSMM, we generated and scored all possible sequences with one additional mutation. To avoid testing mutations that are irrelevant to *L. pneumophila* (i.e., mutations that introduce amino acids that are rare in *L. pneumophila* proteins), we calculated the frequency of each amino acid in each position of the C terminus of the 1,000 sequences used to train the background model of *L. pneumophila* (Dataset S2). Relevant mutations were defined as those that appear in at least 5% of the background sequences in the same position relative to the C terminus.

Annotation of Putative Effectors. ORF *x* was annotated as containing local similarity to ORF *y* from genome *G* if the E-value of the alignment of *x* with *y* in a BLAST search against a database of all the ORFs in genome *G* is lower than 0.01, after filtering for low-complexity regions. This permissive threshold was used to ensure that we do not overestimate the number of unique putative effectors. Motif annotations were performed using Pfam (47), by scanning the ORFs against Pfam-A HMMs with default cutoffs.

Genomic Sequences Used for Analysis. The genomic sequences used for the bioinformatic analyses were *L. pneumophila* Philadelphia-1 (NCBI accession no. NC_002942) (48); *L. longbeachae* NSW150 (NCBI accession nos. NC_013861 and NC_014544) (43); *L. drancourtii* LLAP12 (NCBI accession no. NZ_ACUL00000000) (49); *L. dumoffii* Tex-KL (NCBI accession no. CM001371.1); *C. burnetii* RSA493 Nine Mile phase I (MNII; NCBI accession nos. NC_002971 and NC_004704) (50); *E. coli* K-12 MG1655 (NCBI accession no. NC_000913) (51); and *R. grylli* (NCBI accession no. NZ_AAQJ02000000).

Bacterial Strains, Plasmids, and Primers. The *L. pneumophila* WT strain used in this study was JR32, a streptomycin-resistant, restriction-negative mutant of *L. pneumophila* Philadelphia-1, which is a WT strain in terms of intracellular growth (52). In addition, mutant strains derived from JR32, which contain a kanamycin cassette instead of the *icm*T gene (GS3011) (53); the *dot*A gene (LELA3118) (52); and a double-deletion mutant that contains a kanamycin cassette in the *icm*S gene and a nonpolar in-frame deletion of the *icm*W gene (ED400) (22) were used. Plasmids and primers used in this study are listed in Dataset S5.

Construction and Translocation of CyaA Fusions. CyaA fusions were constructed as described previously (18), and the resulting plasmids are described in Dataset S5. To generate the synthetic signal, two ~70-bp primers were used as templates for PCR that overlapped one another by ~20 bp (Dataset S5). Two additional primers, which contained EcoRI and BamHI restriction sites and overlapped the 5′ and the 3′ ends of the template primers, were used for amplification. To generate mutated synthetic signals, suitable sets of primers containing the mutations were used. The CyaA translocation assays were preformed as previously described (18). The levels of cAMP were determined using the cAMP Biotrak enzyme immunoassay system (Amersham Biosciences) according to the manufacturer's instructions.

***C. burnetii* RT-qPCR.** *C. burnetii* MNII was grown in acidified citrate cysteine medium (ACCM-2) axenic medium (54) at 37 °C, with 5% (vol/vol) CO_2 and 2.5% (vol/vol) O_2. Samples were taken at 3 d (exponential phase) or 6 d (postexponential phase), and bacteria were collected by filtration. Filters were flash-frozen and stored at −80 °C. Intracellular infections were performed as described previously, with several modifications (55). HEK 293T cells (CRL-11268; American Type Culture Collection) were infected with ACCM-2–grown *C. burnetii* NMII at a multiplicity of infection of 10. After 14 h, the cells were washed with PBS to remove extracellular bacteria. The infected cells were collected 3 d (exponential) and 7 d (postexponential) after infection. Medium was replaced with RNALater (Ambion), cells were lysed in 1% (wt/vol) saponin (Sigma), and the pellet was flash-frozen and stored at −80 °C. Bacterial RNA was isolated with TRIzol (Invitrogen), and cDNA was produced using the iScript Select kit (BioRad) according to the manufacturer's instructions. The qPCR assay was performed using SYBR Green reagents and the StepOne Plus system (Applied Biosystems). Primers for the qPCR assay were designed using Primer3 (Dataset S5). The ΔC_T (cycle threshold) was calculated for each RT-PCR product with the corresponding 16S value. The starting quantity was calculated by raising 2 to the power of the C_T. Biological triplicates were averaged for each growth condition. The data were $\log_2$-transformed, and the gene expression values were centered using Cluster 3.0. The heat map was made using Java Treeview (56).

ACKNOWLEDGMENTS. This study was supported, in part, by Grant 3-8178 from the Chief Scientist Office of the Ministry of Science, Israel (to G.S. and T.P.), and, in part, by Grant 2009070 from the United States–Israel Binational Science Foundation (to G.S., T.P., and H.A.S.). D.B. is a fellow of the Converging Technologies Program of the Israeli Council for Higher Education. K.S. and H.A.S. acknowledge membership within and support from the Region V Great Lakes Regional Center of Excellence in Biodefense and Emerging Infectious Diseases Research Consortium (National Institutes of Health Award 1-U54-AI-057153).

1. Ghigo E, et al. (2009) Intracellular life of *Coxiella burnetii* in macrophages. *Ann N Y Acad Sci* 1166:55–66.
2. Diederen BM (2008) *Legionella* spp. and Legionnaires' disease. *J Infect* 56(1):1–12.
3. Cutler SJ, Bouzid M, Cutler RR (2007) Q fever. *J Infect* 54(4):313–318.
4. Fields BS (1996) The molecular ecology of *legionellae*. *Trends Microbiol* 4(7):286–290.
5. Newton HJ, Roy CR (2011) The *Coxiella burnetii* Dot/Icm system creates a comfortable home through lysosomal renovation. *MBio* 2(5):e00226–e00211.
6. Isberg RR, O'Connor TJ, Heidtman M (2009) The *Legionella pneumophila* replication vacuole: Making a cosy niche inside host cells. *Nat Rev Microbiol* 7(1):13–24.
7. Voth DE, Heinzen RA (2007) Lounging in a lysosome: The intracellular lifestyle of *Coxiella burnetii*. *Cell Microbiol* 9(4):829–840.
8. Fields BS, Benson RF, Besser RE (2002) *Legionella* and Legionnaires' disease: 25 years of investigation. *Clin Microbiol Rev* 15(3):506–526.
9. Segal G, Feldman M, Zusman T (2005) The Icm/Dot type-IV secretion systems of *Legionella pneumophila* and *Coxiella burnetii*. *FEMS Microbiol Rev* 29(1):65–81.
10. Carey KL, Newton HJ, Lührmann A, Roy CR (2011) 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(5):e1002056.
11. Voth DE, Heinzen RA (2009) *Coxiella* type IV secretion and cellular microbiology. *Curr Opin Microbiol* 12(1):74–80.
12. Shin S, Roy CR (2008) Host cell processes that influence the intracellular survival of *Legionella pneumophila*. *Cell Microbiol* 10(6):1209–1220.
13. Pan X, Lührmann A, Satoh A, Laskowski-Arce MA, Roy CR (2008) Ankyrin repeat proteins comprise a diverse family of bacterial type IV effectors. *Science* 320(5883):1651–1654.
14. Voth DE, et al. (2009) The *Coxiella burnetii* ankyrin repeat domain-containing protein family is heterogeneous, with C-terminal truncations that influence Dot/Icm-mediated secretion. *J Bacteriol* 191(13):4232–4242.
15. Nagai H, Kagan JC, Zhu X, Kahn RA, Roy CR (2002) A bacterial guanine nucleotide exchange factor activates ARF on *Legionella* phagosomes. *Science* 295(5555):679–682.

16. Nagai H, et al. (2005) A C-terminal translocation signal required for Dot/Icm-dependent delivery of the *Legionella* RalF protein to host cells. *Proc Natl Acad Sci USA* 102(3):826–831.
17. Kubori T, Hyakutake A, Nagai H (2008) *Legionella* translocates an E3 ubiquitin ligase that has multiple U-boxes with distinct functions. *Mol Microbiol* 67(6):1307–1319.
18. Burstein D, et al. (2009) Genome-scale identification of *Legionella pneumophila* effectors using a machine learning approach. *PLoS Pathog* 5(7):e1000508.
19. Huang L, et al. (2011) The E Block motif is associated with *Legionella pneumophila* translocated substrates. *Cell Microbiol* 13(2):227–245.
20. Ninio S, Zuckman-Cholon DM, Cambronne ED, Roy CR (2005) The *Legionella* IcmS-IcmW protein complex is important for Dot/Icm-mediated protein translocation. *Mol Microbiol* 55(3):912–926.
21. Cambronne ED, Roy CR (2007) The *Legionella pneumophila* IcmSW complex interacts with multiple Dot/Icm effectors to facilitate type IV translocation. *PLoS Pathog* 3(12): e188.
22. 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.
23. Durbin R, Eddy S, Krogh A, Mitchison G (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge Univ Press, Cambridge, UK).
24. Levinson SE (1986) Continuously variable duration hidden Markov models for automatic speech recognition. *Comput Speech Lang* 1(1):29–45.
25. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE* 77(2):257–286.
26. de Felipe KS, et al. (2008) *Legionella* eukaryotic-like type IV substrates interfere with organelle trafficking. *PLoS Pathog* 4(8):e1000117.
27. Zhu W, et al. (2011) Comprehensive identification of protein substrates of the Dot/Icm type IV transporter of *Legionella pneumophila*. *PLoS ONE* 6(3):e17638.
28. 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 USA* 102(34):12206–12211.
29. Asare R, Abu Kwaik Y (2007) Early trafficking and intracellular replication of *Legionella longbeachae* within an ER-derived late endosome-like phagosome. *Cell Microbiol* 9(6):1571–1587.
30. Cazalet C, et al. (2004) Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. *Nat Genet* 36(11): 1165–1173.
31. de Felipe KS, et al. (2005) Evidence for acquisition of *Legionella* type IV secretion substrates via interdomain horizontal gene transfer. *J Bacteriol* 187(22):7716–7726.
32. Beare PA, et al. (2011) Dot/Icm type IVB secretion system requirements for *Coxiella burnetii* growth in human macrophages. *MBio* 2(4):e00175–e11.
33. Voth DE, et al. (2011) The *Coxiella burnetii* cryptic plasmid is enriched in genes encoding type IV secretion system substrates. *J Bacteriol* 193(7):1493–1503.
34. Chen C, et al. (2010) Large-scale identification and translocation of type IV secretion substrates by *Coxiella burnetii*. *Proc Natl Acad Sci USA* 107(50):21755–21760.
35. Leclerque A, Kleespies RG (2008) Type IV secretion system components as phylogenetic markers of entomopathogenic bacteria of the genus *Rickettsiella*. *FEMS Microbiol Lett* 279(2):167–173.
36. Lloyd SA, Norman M, Rosqvist R, Wolf-Watz H (2001) *Yersinia* YopE is targeted for type III secretion by N-terminal, not mRNA, signals. *Mol Microbiol* 39(2):520–531.
37. Lloyd SA, Sjöström M, Andersson S, Wolf-Watz H (2002) Molecular characterization of type III secretion signals via analysis of synthetic N-terminal amino acid sequences. *Mol Microbiol* 43(1):51–59.
38. Hueck CJ (1998) Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiol Mol Biol Rev* 62(2):379–433.
39. Galán JE, Wolf-Watz H (2006) Protein delivery into eukaryotic cells by type III secretion machines. *Nature* 444(7119):567–573.
40. Schulein R, et al. (2005) A bipartite signal mediates the transfer of type IV secretion substrates of *Bartonella henselae* into human cells. *Proc Natl Acad Sci USA* 102(3): 856–861.
41. Hurtado-Guerrero R, et al. (2010) Molecular mechanism of elongation factor 1A inhibition by a *Legionella pneumophila* glycosyltransferase. *Biochem J* 426(3): 281–292.
42. Kozak NA, et al. (2010) Virulence factors encoded by *Legionella longbeachae* identified on the basis of the genome sequence analysis of clinical isolate D-4968. *J Bacteriol* 192(4):1030–1044.
43. Cazalet C, et al. (2010) Analysis of the *Legionella longbeachae* genome and transcriptome uncovers unique strategies to cause Legionnaires' disease. *PLoS Genet* 6(2):e1000851.
44. Pearce MM, et al. (2012) *Legionella cardiaca* sp. nov., isolated from a case of native valve endocarditis in a human heart. *Int J Syst Evol Microbiol* 62(Pt 12):2946–2954.
45. Campocasso A, Boughalmi M, Fournous G, Raoult D, La Scola B (2012) *Legionella tunisiensis* sp. nov. and *Legionella massiliensis* sp. nov., isolated from environmental water samples. *Int J Syst Evol Microbiol.* 62(12):3003–3006.
46. Witten IH, Frank E (2005) *Data Mining: Practical Machine Learning Tools and Techniques* (Morgan Kaufmann, San Francisco).
47. Finn RD, et al. (2010) The Pfam protein families database. *Nucleic Acids Res* 38(Database issue):D211–D222.
48. Chien M, et al. (2004) The genomic sequence of the accidental pathogen *Legionella pneumophila*. *Science* 305(5692):1966–1968.
49. Moliner C, Raoult D, Fournier PE (2009) Evidence that the intra-amoebal *Legionella drancourtii* acquired a sterol reductase gene from eukaryotes. *BMC Res Notes* 2:51.
50. Seshadri R, et al. (2003) Complete genome sequence of the Q-fever pathogen *Coxiella burnetii*. *Proc Natl Acad Sci USA* 100(9):5455–5460.
51. Blattner FR, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* 277(5331):1453–1462.
52. Sadosky AB, Wiater LA, Shuman HA (1993) Identification of *Legionella pneumophila* genes required for growth within and killing of human macrophages. *Infect Immun* 61(12):5361–5373.
53. Zusman T, Yerushalmi G, Segal G (2003) Functional similarities between the *icm/dot* pathogenesis systems of *Coxiella burnetii* and *Legionella pneumophila*. *Infect Immun* 71(7):3714–3723.
54. Omsland A, et al. (2011) Isolation from animal tissue and genetic transformation of *Coxiella burnetii* are facilitated by an improved axenic growth medium. *Appl Environ Microbiol* 77(11):3720–3725.
55. Howe D, Shannon JG, Winfree S, Dorward DW, Heinzen RA (2010) *Coxiella burnetii* phase I and II variants replicate with similar kinetics in degradative phagolysosome-like compartments of human macrophages. *Infect Immun* 78(8):3465–3474.
56. Saldanha AJ (2004) Java Treeview—Extensible visualization of microarray data. *Bioinformatics* 20(17):3246–3248.

Supporting Information

Lifshitz et al. 10.1073/pnas.1215278110

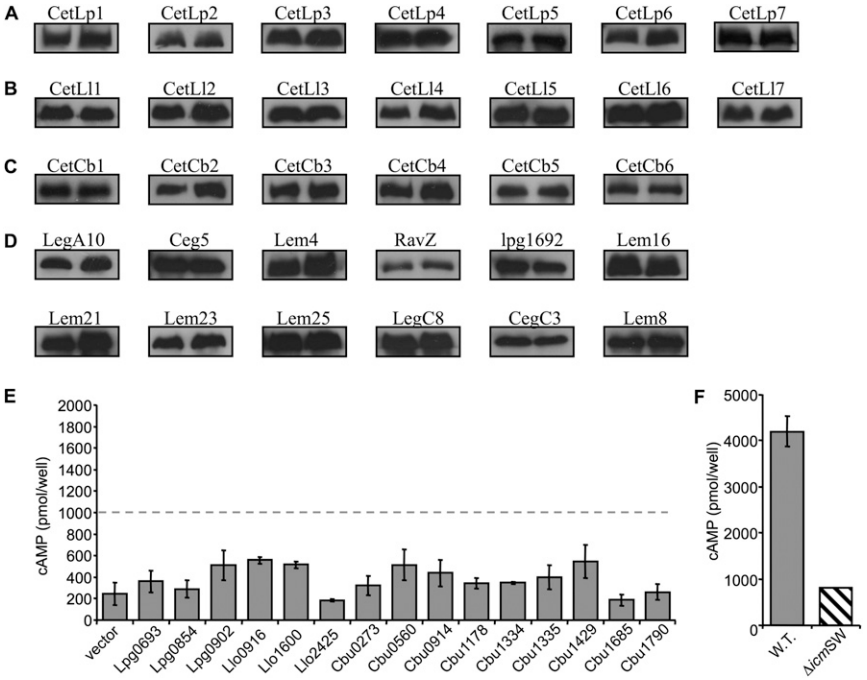

Fig. S1. Expression of *Legionella* and *Coxiella* effectors, translocation of Lpg1368, and examination of noneffectors. The CyaA fusions of *Legionella pneumophila* (*A*), *Legionella longbeachae* (*B*), and *Coxiella burnetii* (*C*) effectors identified were examined for their expression in the WT strain JR32 (*Left*) and the *icm*T deletion mutant GS3011 (*Right*). (*D*) Expressions of CyaA fusions of effectors with a low hidden semi-Markov model (HSMM) score were examined in the WT strain JR32 (*Left*) and the *icmS-icmW* double-deletion mutant ED400 (*Right*). (*E*) CyaA fusions of predicted *L. pneumophila* (Lpg), *L. longbeachae* (Llo), and *C. burnetii* (Cbu) ORFs were examined for translocation into HL-60–derived human macrophages from the WT strain JR32; the effectors examined are indicated below each bar, and the gray dashed line indicates the minimal cAMP level required for a protein to be considered an effector. (*F*) CyaA fusion of Lpg1368 (Lgt1) was examined for translocation into HL-60–derived human macrophages from the WT strain JR32 (gray bars) and from the *icmS-icmW* double-deletion mutant ED400 (diagonal striped bars). The cAMP levels of the infected cells were determined as described in *Materials and Methods*. The bar heights represent the means of the amount of cAMP per well; error bars indicate SDs.

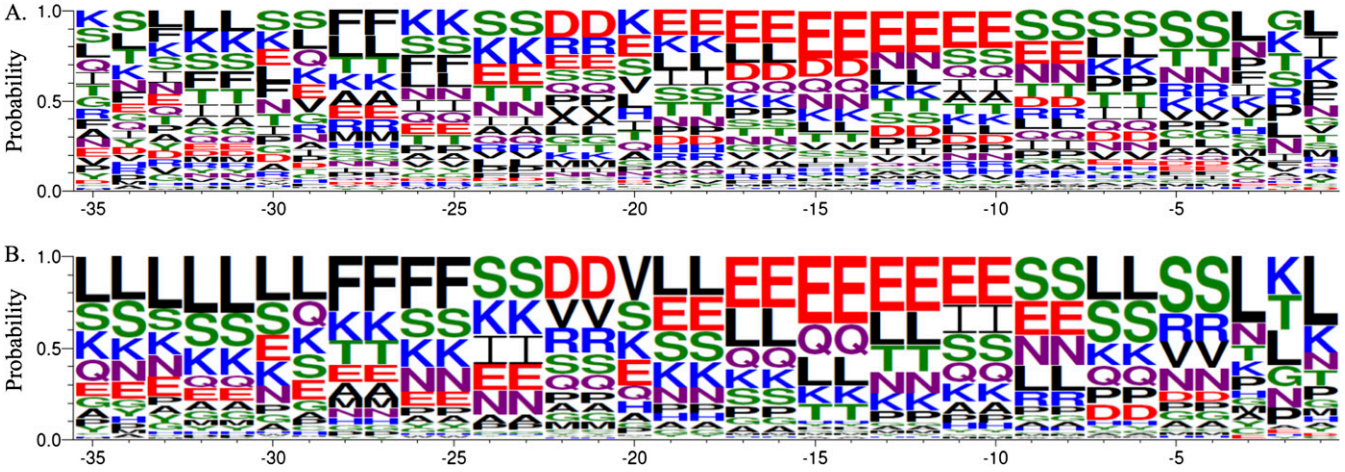

Fig. S2. Probability of occurrence of amino acids in the optimal secretion signal (OSS) according to the hidden semi-Markov model (HSMM). (*A*) Sequence of the OSS was inferred from the model by extracting the emission probabilities of each amino acid along the most likely path of the model with the relevant length. (*B*) To account for the physicochemical properties of the amino acids, the emitted amino acids were clustered into five groups: (*i*) positively charged (lysine and arginine), (*ii*) negatively charged (aspartic acid and glutamic acid), (*iii*) polar amino acids bearing a hydroxyl group (serine and threonine), (*iv*) polar amino acids bearing an amide group (asparagine and glutamine), and (*v*) hydrophobic (leucine, isoleucine, valine, and phenylalanine). Each group in each position is represented by the amino acid with the highest emission probability, as shown in Fig. 2*A*.

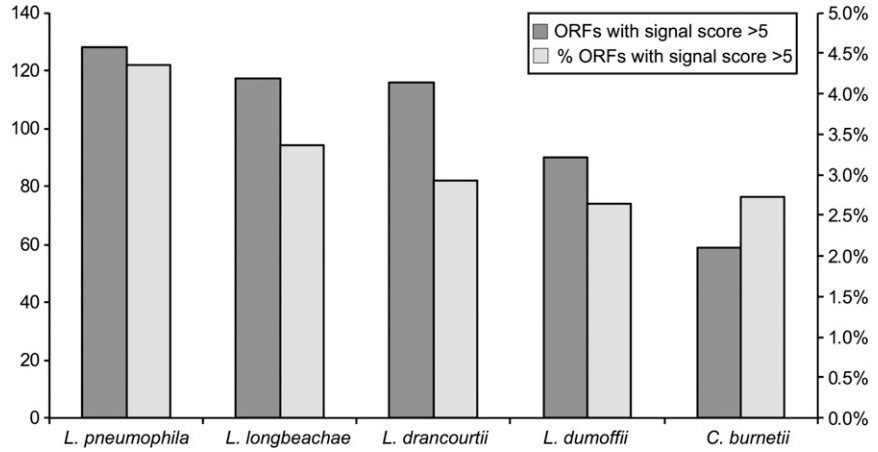

Fig. S3. Number and percentage of putative effectors according to their signal score in four *Legionella* species and *Coxiella burnetii*. The number of putative effectors (dark gray) was computed as the number of ORFs with a signal score higher than 5, excluding ORFs with annotations suggesting they probably do not encode for effectors (Dataset S4). The percentage of the putative effectors of the total number of ORFs in the relevant genome indicated below the bars is shown in light gray.

Table S1. Performances of the hidden semi-Markov model over different combinations of C terminus lengths and number of states

No. of amino acids	No. of states												
	4	6	8	10	12	14	16	18	20	22	24	26	28
15	0.832	0.837	0.839	0.844	0.843	0.834	—	—	—	—	—	—	—
20	0.833	0.837	0.846	0.848	0.845	0.86	0.859	0.844	—	—	—	—	—
25	0.827	0.838	0.845	0.854	0.851	0.854	0.855	0.856	0.859	0.856	0.845	—	—
30	0.842	0.851	0.854	0.86	0.865	0.867	0.871	0.868	0.87	0.877	0.872	0.868	0.858
35	0.84	0.855	0.866	0.867	0.872	0.874	0.875	0.878	0.88	**0.881**	0.879	0.873	0.873
40	0.84	0.854	0.863	0.867	0.872	0.871	0.873	0.874	0.872	0.879	0.879	0.874	0.872

The performances are measured as area under the receiver operating characteristic curve (AUC). The highest AUC value that led to the choice of 22 states and 35 amino acids is marked in bold.

Table S2. Effectors identified in this study

ORF	Name	Size, no. of amino acids	Homologs in			
			Legionella longbeachae	*Legionella drancourtii*	*Legionella dumoffii*	*Coxiella burnetii*
Lpg0140	CetLp1	444	LLO_3260	LDG_6044	FdumT_00670	—
Lpg0393	CetLp2	287	LLO_2832	LDG_8440	FdumT_13152	—
Lpg1663	CetLp3	168	LLO_1992	LDG_5734	FdumT_14047	—
Lpg1822	CetLp4	339	—	—	—	—
Lpg2244	CetLp5	980	—	—	—	—
Lpg2283	CetLp6	132	—	—	—	—
Lpg2806	CetLp7	460	LLO_0161	LDG_8759	FdumT_01030	—
LLO_0105	CetLl1	262	—	LDG_6606	—	—
LLO_0118	CetLl2	240	—	—	—	—
LLO_1403	CetLl3	127	—	—	—	—
LLO_1506	CetLl4	324	—	—	—	—
LLO_2179	CetLl5	536	LLO_2985	—	—	—
LLO_2598	CetLl6	368	—	—	—	—
LLO_3390	CetLl7	322	—	—	—	—
CBU_0021	CetCb1	809	—	—	—	—
CBU_0388	CetCb2	1,392	—	—	—	—
CBU_0626	CetCb3	695	—	—	—	—
CBU_0885	CetCb4	388	—	—	—	CBU_1676
CBU_1686	CetCb5	773	—	—	—	—
CBU_1724	CetCb6	747	—	—	—	—

None of the validated effectors has a homolog in *Legionella pneumophila*.

Other Supporting Information Files

Dataset S1 (XLS)
Dataset S2 (XLS)
Dataset S3 (XLS)
Dataset S4 (XLS)
Dataset S5 (XLS)