Molecular Biology and Evolution

Quantifying Nonvertical Inheritance in the Evolution of *Legionella pneumophila*

Mireia Coscollá,†,[1,2,3], Iñaki Comas,†,[4] and Fernando González-Candelas*,[1,2]

[1]Unidad Mixta de Investigación "Genómica y Salud" CSISP-UV/Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Valencia, Spain

[2]CIBER Epidemiología y Salud Pública, Valencia, Spain

[3]Tuberculosis Research Unit, Swiss Tropical and Public Health Institute, Basel, Switzerland

[4]Division of Mycobacterial Research, National Institute for Medical Research, London, United Kingdom

†These two authors contributed equally to this work.

***Corresponding author:** E-mail: fernando.gonzalez@uv.es.

Associate editor: Daniel Falush

Abstract

The exchange of genetic material among bacterial strains and species is recognized as an important factor determining their evolutionary, population genetic, and epidemiological features. We present a detailed analysis of nonvertical inheritance in *Legionella pneumophila*, a human pathogen and facultative intracellular parasite of amoebas. We have analyzed the exchange of *L. pneumophila* genetic material with other bacteria at three different levels: population genetics, population genomics, and phylogenomics. At the population genetics level, we have analyzed 89 clinical and environmental isolates after sequencing six coding loci and three intergenic regions for a total of 3,923 bp. In the population genomics analysis, we have studied the roles of recombination and mutation in the common portion of the genome sequence of four *L. pneumophila* strains. In the phylogenomic analysis, we have studied the phylogenetic origin of 1,700 genes in the *L. pneumophila* pangenome. For this, we have considered 12 possible phylogenetic alternatives, derived from a reference tree obtained from 104 genes from 41 species, which have been tested under a rigorous statistical framework. The results obtained agree in assigning an important role to nonvertical inheritance in shaping the composition of the *L. pneumophila* genome and of the genetic variation in its populations. We have found a negative correlation between phylogenetic distance and likelihood of horizontal gene transfer. Phylogenetic proximity and increased chances resulting from sharing the ecological niche provided by the amoeba host have likely had a major influence on the rate of gene exchange in *Legionella*.

Key words: horizontal gene transfer, phylogenomics, population genetics, population genomics, intracellular habitat.

Research article

Introduction

Legionella pneumophila is a recently emerged pathogen included in the Gamma-Proteobacteria, where it occupies a basal position according to most phylogenetic analyses (Comas et al. 2006; Wu and Eisen 2008; Gao et al. 2009; Gómez-Valero et al. 2009). It was first associated to an infectious outbreak of legionellosis, a potentially fatal form of pneumonia, in 1976 (Fraser et al. 1977; McDade et al. 1977), and to a milder form of infection known as Pontiac fever. Since its discovery, a large number of outbreaks and community-acquired infections with *L. pneumophila* have been reported all over the world, mainly associated to man-made water-holding installations from where the bacteria are disseminated by aerosols. These bacteria naturally inhabit water environments (Fliermans et al. 1981), and they are often found as components of biofilms (Rogers et al. 1994). Additionally, *Legionella* have developed the ability to infect different amoeba species (Rowbotham 1980; Fields 1996) in whose cytoplasm they complete part of their life cycle after preventing the full development of phagosomes by the infected cells.

Legionella are not the only bacteria infecting amoebas, which host in their cytoplasm a number of bacterial species with different forms and degrees of symbiotic association (Horn and Wagner 2004). There are reports of simultaneous identification of two or even three different bacterial species inside the same amoeba cell (Heinz et al. 2007). Finally, under appropriate conditions, *Legionella* can survive and replicate in man-made aquatic systems from where it can potentially spread by aerosols and infect humans (Fields 1996). *Legionella* is able to invade, replicate, and survive in human macrophages mainly by surrounding itself by a membrane-bound vacuole that prevents its lysis by lysosomes (Fields et al. 2002).

The impact of recombination on the population structure and genetic diversity of *L. pneumophila* is controversial. The population structure of *L. pneumophila* was initially defined as clonal due to high linkage disequilibrium between allozyme markers and the presence of clonal complexes with a global distribution (Selander et al. 1985). Although several studies have provided some evidence for recombination (Ko et al. 2002, 2003; Coscollá and

González-Candelas 2007) and horizontal gene transfer (HGT) (Morozova et al. 2004; de Felipe et al. 2005), Edwards et al. (2008) still sustained that frequent recombination does not play a major role in the diversification of this species.

A clonal view of bacterial population structure became popular at the beginning of the 1980s when the analysis of bacteria samples by pulse-field gel electrophoresis (PFGE) showed that very few types dominated global strains collections (Levin 1981; Selander et al. 1986). A periodic selection model (Atwood et al. 1951) seemed to fit PFGE data, with bacterial populations showing a constant turnover of the most frequent type coupled to the emergence of a new fitter variant. The periodic selection model assumes that bacterial populations are mainly asexual with very little genetic exchange among their members, with the notable exception of the spread of drug resistance. However, a landmark article by Maynard Smith et al. (1993) showed how bacterial populations present a range of recombination rates from almost fully asexual to highly recombinogenic species. In the last decade, a number of schemes have been developed for the typing of bacterial pathogens, mostly through the sequencing of a few genes, resulting in multilocus sequence typing (MLST) (Maiden et al. 1998; Enright and Spratt 1999). The analysis of the publicly accessible databases corroborated the early findings by Maynard Smith et al. (1993) and acknowledged homologous recombination as an important force shaping bacterial population diversity (Vos and Didelot 2009; Didelot and Maiden 2010). Although intraspecies homologous recombination decreases with increasing sequence divergence (Shen and Huang 1986), interphylum transfer of DNA occurs by different processes, leading to HGTs (Lawrence and Retchless 2009).

One of the main outcomes of the genome-sequencing revolution since the late 1990s is the recognition that HGT and gene loss in bacterial genomes are much more frequent than previously thought when only single-gene information was available (Ochman et al. 2000; Dagan and Martin 2007). Strains of the same bacterial species differ not only in the sequence of their genes but also, and often substantially, in the contents of their genes(Tettelin et al. 2005). Despite this, the extent of HGT between bacterial genomes and its consequences, both taxonomically and biologically, are hotly debated issues (Bapteste et al. 2009). However, it is clear that HGT has contributed to the success of many bacterial pathogens (Lawrence 2005), although it remains unclear how many of the observed imported genes are truly adaptive and how many have become fixed by random drift (Marri et al. 2007).

Given our interest in understanding the population structure and evolution of *L. pneumophila*, we have analyzed the relevance of genetic exchange in this species at three different levels. First, we have used clinical and environmental isolates from the Comunitat Valenciana region (Spain) to evaluate the extent of genetic recombination at the population level. Second, we have used the genome sequences of four available *L. pneumophila* strains to evaluate the extent of recombination at the genomic level using isolates from different geographic regions (one American, two French, and one English strain). Finally, we have identified, by phylogenetic analyses, the most likely origin of each gene in the *L. pneumophila* pangenome considering as most likely donors other bacterial species that share the same intracellular niche with *Legionella*.

All these analyses concur in assigning a relevant role to nonvertical inheritance in the evolution and population structure of *Legionella*, and we will argue that the ecological and genetic conditions that currently facilitate genetic exchange among *Legionella* strains have likely been operating along a substantial fraction of the evolutionary history of this species. To our knowledge, this is one of the first evaluations at these three levels of the origin and amount of nonvertical inheritance in a bacterial pathogen.

Materials and Methods

Population Level Analyses

Isolates and DNA Extraction

Forty-seven isolates of *L. pneumophila* were obtained from different human-related environments from which the bacterium is usually spread such as cooling towers, air-conditioning trays, etc. (more details can be found in supplementary table S1, Supplementary Material online). DNA extraction and serogroup assignment were performed as described (Coscollá and González-Candelas 2007). Forty-two clinical samples were also analyzed including DNA from cultured respiratory samples, which was extracted as described (Coscollá and González-Candelas 2007) and directly from respiratory isolates. DNA was extracted with UltraClean BloodSpin Kit (Mobio Laboratories, Inc.) and stored at -20°C. See supplementary table S1, Supplementary Material online, for a detailed list of the isolates analyzed.

Polymerase Chain Reaction Amplification and DNA Sequencing

A seminested approach was used to amplify DNA extracted from respiratory samples. Primers for the six protein-coding genes were described in Gaia et al. (2005) and for the three intergenic regions in Coscollá and González-Candelas (2007). Polymerase chain reaction products were purified using High Pure PCR Product Purification Kit (Roche Diagnostics). Purified DNA was directly sequenced by the dideoxy method using BigDye Terminator v3.0 Ready Reaction Cycle Sequencing Kit and analyzed in an ABI PRISM 3700 sequencer (Applied Biosystems).

Sequence Analysis of Population Genetic Data

Multiple sequence alignments were obtained using ClustalX (Thompson et al. 1997) and refined by visual inspection. For each individual locus and for the concatenated alignment of the nine loci, maximum likelihood (ML) phylogenetic trees were obtained with PHYML 2.4.4 (Guindon and Gascuel 2003) using the most appropriate model of nucleotide substitution for each genome region as determined with the ML approach implemented

in Modeltest v3.7 (Posada and Crandall 1998). Likelihood scores for each model were estimated in PAUP*4.0b10 (Swofford 2002), and the best model was determined by using the Akaike Information Criterion (Akaike 1974). Support for the nodes was evaluated by bootstrapping with 1,000 pseudoreplicates.

Recombination in the Population Data Set

RDP3 (Martin et al. 2005) was used to infer recombination in 89 clinical and environmental samples by analyzing the concatenated multiple alignment of unique haplotypes in the data set. Six methods were employed: two phylogenetic methods, which infer recombination when different parts of the genome result in discordant topologies, RDP, and Bootscan/Recscan (Salminen et al. 1995; Martin and Rybicki 2000), and four nucleotide substitution methods (which examine the sequences either for a significant clustering of substitutions or for a fit to an expected statistical distribution): MaxChi (Smith 1992), Chimaera (Posada and Crandall 2001), GeneConv (Padidam et al. 1999), and SiScan (Gibbs et al. 2000). We only considered recombination events that were identified by more than two methods to avoid dependence on one single methodology. Common settings for all methods were to consider sequences as circular, require phylogenetic evidence, polish breakpoints, and statistical significance was set at the $P < 0.05$ level, with Bonferroni's correction for multiple comparisons as implemented in RDP.

We confirmed each recombination event detected with RDP3 by comparing the ML phylogenetic tree of the region involved in recombination and those corresponding to the flanking regions. We used the expected likelihood weight (ELW) (Strimmer and Rambaut 2002) and the Shimodaira and Hasegawa (1999) tests for the reciprocal topologies and the corresponding alignments.

The population recombination rate (ρ) was estimated using the standard likelihood coalescent approach implemented in the LDhat package (McVean et al. 2002) within each gene region. In haploid organisms, ρ can be expressed as $2N_e r$, where N_e is the inbreeding effective population size and r is the recombination rate per locus per generation. For organisms such as viruses and bacteria in which a gene-conversion model for recombination is more appropriate than a crossing-over model, r can be approximated by $r \approx 2c\bar{t}$, where c is the per base rate of initiation of gene conversion and $\bar{t}$ is the average gene conversion tract length. As loci ranged from 200 to 600 bp in size, we performed analyses for different values of $\bar{t}$: 200, 500, 750, and 1,200, and we chose the one with the highest likelihood value. The likelihood permutation test implemented in LDhat was used to test the hypothesis of no recombination ($\rho = 0$).

We used ClonalFrame 1.1 (Didelot and Falush 2007) to estimate parameters in the evolutionary process that led to the observed pattern of nucleotide variation in the 4,923 bp alignment obtained for the nine concatenated loci. Estimates of the ratio of recombination and mutation rates (ρ/θ) and also the ratio of probabilities that a given site is altered through recombination and mutation (r/m) were obtained with this method, with two runs of 100,000 iterations including 50,000 of burn-in. Congruence of parameter estimates between both runs was evaluated according to Gelman and Rubin (1992) criteria.

Recombination versus Total Variation

We derived estimates of the relevance of recombination, as compared with mutation, in the generation of genetic diversity in this sample with the following procedure. First, DnaSP 4.0 (Rozas et al. 2003) was used to calculate $\theta = 2Nu$, where N is the effective population size and u is the mutation rate per nucleotide and per generation (Nei 1987 equation 10.3; Tajima 1983 equation 3), and $R = 2Nr$, where r is the recombination rate between adjacent sites (Hudson et al. 1987). Additionally, a genetic network was obtained using the median-joining method (Bandelt et al. 1999), implemented in Networks 4.2.0.1, to estimate the number of nucleotide substitutions between nodes. By mapping the inferred recombination events onto the median network, we estimated which changes had likely resulted from mutations and which from recombination events. For this, we compared the observed number of changes in each segment in the network and for each locus with the number of expected substitutions, assuming that these followed a Poisson distribution whose rate depended on the branch length and the locus diversity. The significance level was fixed at 0.05, and Bonferroni's correction was used to account for multiple simultaneous testing. We considered that significant deviations, both in excess and in defect, from the expected number of mutational changes between nodes might be the result of recombination. These results were compared with those of RDP and phylogenetic incongruence analyses, and we only considered those confirmed by at least one of these other methodologies. Hence, we considered branches in which we found more or less nucleotide changes than expected, and these were used as an estimate of the proportion of changes due to recombination compared with the total variation in this data set.

Population Genomic Analyses

Genome Alignment of the Legionella Genomes

The complete genome sequences of four *L. pneumophila* strains available in GenBank on June 2008 were downloaded. A whole-genome alignment was obtained with MAUVE version 2.0 (Darling et al. 2004). The resulting set of locally collinear blocks was concatenated, using the genome order in *L. pneumophila* str. Philadelphia, to derive the *Legionella* syntenic genome. Coding regions in this syntenic genome were annotated using the *L. pneumophila* str. Philadelphia as a reference. Positions with gaps were trimmed to remove noise from the ensuing evolutionary analyses.

Incidence of Recombination Along the Legionella Syntenic Genome

We looked for evidence of recombination along the syntenic genome. Each syntenic block obtained from the

genome alignment was analyzed using RDP3 (Martin et al. 2005) as explained above. However, given the small number of taxa included in this analysis (only four strains), we discarded recombination events inferred only with methods based on phylogenetic analysis but we maintained the requirement of detection with at least two different methods to consider a recombination event. This allowed us to obtain a reliable set of recombination events affecting the genomes of *Legionella* and to map them on the syntenic genome. We normalized the number of recombination events per block to the number of recombination events detected every 10 kb in order to compare this estimate with the variation of nucleotide diversity along the syntenic genome. A sliding-window analysis (window size = 10 kb, step size = 500 bp) of nucleotide diversity was obtained using VariScan (Vilella et al. 2005).

We also used ClonalFame 1.1 (Didelot and Falush 2007) to obtain estimates of the ratios of recombination and mutation rates (ρ/θ) and of the probabilities that a given site is altered through recombination or mutation (r/m) with the same parameters described above.

Phylogenomic Analyses

Determination of the Legionella Pangenome

Orthology relationships among the genes in the four *Legionella* genomes were determined with OrthoMCL (Li et al. 2003). Using this program, we performed an all-against-all Blast search using the amino acid sequences and a cutoff with *E* value = 1.0×10^{-5}. Reciprocal best hits found were classified into putative orthologs or recent paralogs when the best hit of a protein was found in the same genome. Next, a similarity matrix was generated by normalizing the *P* values' matrix between all reciprocal best hits, which was then analyzed by means of a Markov Cluster algorithm (MCL) in order to delimit clusters of orthologs/paralogs.

Phylogenetic Origin of the Legionella Pangenome

We looked for putative orthologs of all the genes in the *L. pneumophila* pangenome in 37 additional bacterial genomes (table 4). These genomes were selected on the basis of two criteria. First, we selected complete genomes from bacterial groups known to share the same intracellular niche of amoebas as *L. pneumophila* (supplementary table S4, Supplementary Material online). If the genome of the reported species was not available, we selected the closest related genome sequenced. Second, we selected some additional genomes for underrepresented bacterial families after the selection of these ecological partners. We looked for those genes of the *Legionella* pangenome putatively present in this set of 37 genomes using BLASTP with an "Expected score" $<1.0 \times 10^{-10}$.

Multiple alignments for each set of putative orthologous genes were obtained from the derived amino acid sequences using ClustalW (Thompson et al. 1994). Ambiguous homologous positions and those including gaps were removed using Gblocks 0.91 (Castresana 2000) with default parameters. The 104 multiple alignments from the shared homologous genes of the 41 species considered in the analysis were concatenated to yield a supermatrix (38,697 amino acids). This was analyzed by ML using PHYML in order to derive a reference tree. We used the JTT model (Jones et al. 1992) of evolution with estimated proportion of invariants (I) and rate heterogeneity approximated by a discrete-gamma distribution using eight categories (G). The resulting topology was used to test alternative phylogenetic origins for each gene in the *Legionella* pangenome.

For this, we considered alternative positions of the monophyletic group of *Legionella* sequences in ten groups defined in the reference tree by monophyletic clusters, occasionally including more than one taxonomic group, as indicated in figure 5. These and two additional phylogenies (supplementary fig 1, Supplementary Material online), one corresponding to a fully unresolved tree and the other with only the tip clades resolved, were used to determine the most likely origin and to test the phylogenetic signal of each gene.

To ascertain the origin of a gene and to assign it to 1 of the 12 alternative topologies, we checked the congruence between two different tests for phylogenetic topologies. We applied the ELW (Strimmer and Rambaut 2002) and the Shimodaira and Hasegawa (1999) tests of competing phylogenetic hypotheses to each multiple alignment considering the 12 alternative topologies described. We restricted our analyses of competing phylogenetic hypotheses to those genes that simultaneously presented no duplicates in the *Legionella* pangenome and were present in at least eight genomes of the previously defined set. Only those genes for which the best topologies for both tests were coincident were considered to correspond to fully resolved cases. A summary of the workflow used in this procedure is presented in table 1.

The average 16S rDNA nucleotide distance from the species in each group to the *Legionella* strains was calculated to test the correlation between the number of inferred HGT events and phylogenetic distance. The 16S rDNA sequences were obtained from the GreenGenes and NCBI databases. Nucleotide distances were obtained after alignment with ClustalW and complete deletion of sites with gaps using the composite ML distance implemented in MEGA (Tamura et al. 2007). Correlation between number of HGT events and phylogenetic distance was obtained using STATA (Stata Statistical Software: Release 10. College Station, TX: StataCorp LP).

We used a chi-square to test for differences in the distribution of HGT events (excluding those genes with a Gamma-Proteobacterial origin) in functional categories with respect to the pangenome. We removed in both cases all the genes annotated with more than one category (HGT data set, $n = 170$ removed, pangenome data set, $n = 1891$). For the test, individual categories were grouped into four main categories (cellular processes, information storage and processing, metabolism, and poorly characterized) following COG classification to avoid underrepresented categories in the HGT data set that could affect the results of the test.

Table 1. Workflow Summarizing the Steps Followed in the Identification of HGT Genes in the *Legionella pneumophila* Pangenome. More details Can Be Found in the Text.

1. **Infer the *Legionella pneumophila* pangenome by comparing the genes in the four completely sequenced strains.**
2. **Identify the complete genomes of other intracellular parasites of amoebas, or their closest relatives, as well as representative genomes of additional bacterial phyla.**
3. **Search for orthologous genes in the data set derived in step 2 for each gene in the pangenome obtained in step 1.**
4. **Construct an ML reference tree with the concatenated multiple alignment of the genes derived in step 3 that were present in all the genomes after the corresponding individual alignments had been trimmed of ambiguous and gap-containing positions.**
5. **Construct 12 different topologies considering the 10 well-supported groups in the reference tree and 2 additional topologies used to test the phylogenetic signal for each gene (supplementary fig 1, Supplementary Material online).**
6. **Obtain the multiple alignment for each gene in the pangenome with at least eight orthologs in the data set derived in step 2 using the results from step 3. Remove ambiguously aligned and gap-containing positions.**
7. **For each gene in step 6, perform SH and ELW tests for the 12 topologies described in step 5. When both tests agree in identifying the same topology as the best one, use this as the one representing the most likely origin of the corresponding *L. pneumophila* gene.**

Results

MLST Analysis of *L. pneumophila* Clonality

To ascertain the level of clonality in a bacterial population, it is necessary to characterize the number and location of recombinant fragments in the genome and the relative contribution of recombination to the overall population genetic diversity. In order to detect the footprints of recombination in the MLST data set, we used three approximations: phylogenetic incongruence, the consensus of several tests implemented in RDP, and detection of regions where the number of nucleotide changes deviates significantly from random mutation expectations.

The phylogenetic incongruence approach requires a reference tree for the complete data set and an individual tree for each locus analyzed. An ML reference tree was obtained (fig. 1) from the concatenated alignment of the nucleotide sequences for nine loci in 89 *L. pneumophila* samples of clinical and environmental origins from the Comunidad Valenciana region (Spain) (supplementary table S1, Supplementary Material online). The multiple alignment spanned 3,923 bp and included 497 variable positions. ML trees were also obtained for each separate locus (available upon request). Both the reference and the single gene trees revealed that one sample (isolate 366) was clearly different from the other isolates studied. Its average nucleotide divergence with respect to the remaining isolates was 0.0651 (range 0.0552–0.0689), whereas the average value among the other isolates was 0.0085 (0.00–0.0162). Consequently, the ensuing analyses were performed excluding this isolate in order to minimize possible distorting effects of this outlier, especially in genetic variability estimates. Nevertheless, the same analyses (some of which will be detailed below) were carried out including the isolate to corroborate that we were not missing important information.

To investigate the topological congruence among all trees, we used Shimodaira–Hasegawa (SH) and ELW tests. These resulted in all single-gene alignments rejecting the trees obtained from every other locus with both tests (table 2). Furthermore, the tree derived after concatenation of the sequences for the nine loci was also rejected by the ELW test with the nine individual alignments, whereas L2 and *mip* also rejected it with the SH test. This suggests that no pair of loci in this set share the same phylogenetic history in this population sample, that is, there is a strong indication that recombination is widespread in this sample.

We proceeded to explore recombination in more detail by using six different methods implemented in RDP with the 37 different haplotypes obtained after concatenation of the nine loci. Seventeen recombination events were identified by at least two of the six methods in RDP (Table 3). All the events mapped to the bordering regions between loci, which are between 50,077 and 746,767 bp apart in the reference *L. pneumophila* Philadelphia genome. Hence, no cases of intragenic recombination were detected with this approach. All loci except *mompS* were included in at least one recombination event involving from 1 (*asd*) up to 16 haplotypes (*pilE*). Five independent recombination events were detected in locus L2. Only eight haplotypes, which correspond to 25 of the 89 samples studied, were not involved in any recombination event (table 3).

Finally, a median-joining network was obtained to estimate the number of mutations inferred to have occurred in each branch and to infer recombination events by comparing them with the expected number deduced from the levels of variation in the corresponding locus and haplotype (fig. 2). Gaps in the alignment were considered as an additional character state to construct the network, but they were not taken into account when estimating the number of nucleotide substitutions between nodes in the network. We evaluated the contribution of recombination to genetic variation in these samples by comparing the expected and observed numbers of nucleotide substitutions between isolates in the median-joining network for each locus. Of the 838 nucleotide differences observed, 195 (23.27%) could be explained by a recombination event (34% when isolate 366 was removed). This figure is likely an underestimate of the actual number of changes due to recombination as events detected by other methods but not by the network analysis was not considered.

A comparison of the recombination events inferred with the three different methodologies described above (phylogenetic incongruence, RDP, and deviation from expected variation in the median-joining network) is shown in table 3. Most recombination events (11/17) were detected by the three methods, five were detected by two methods, and only one event, in locus *proA*, was detected by RDP but not by the two other methods.

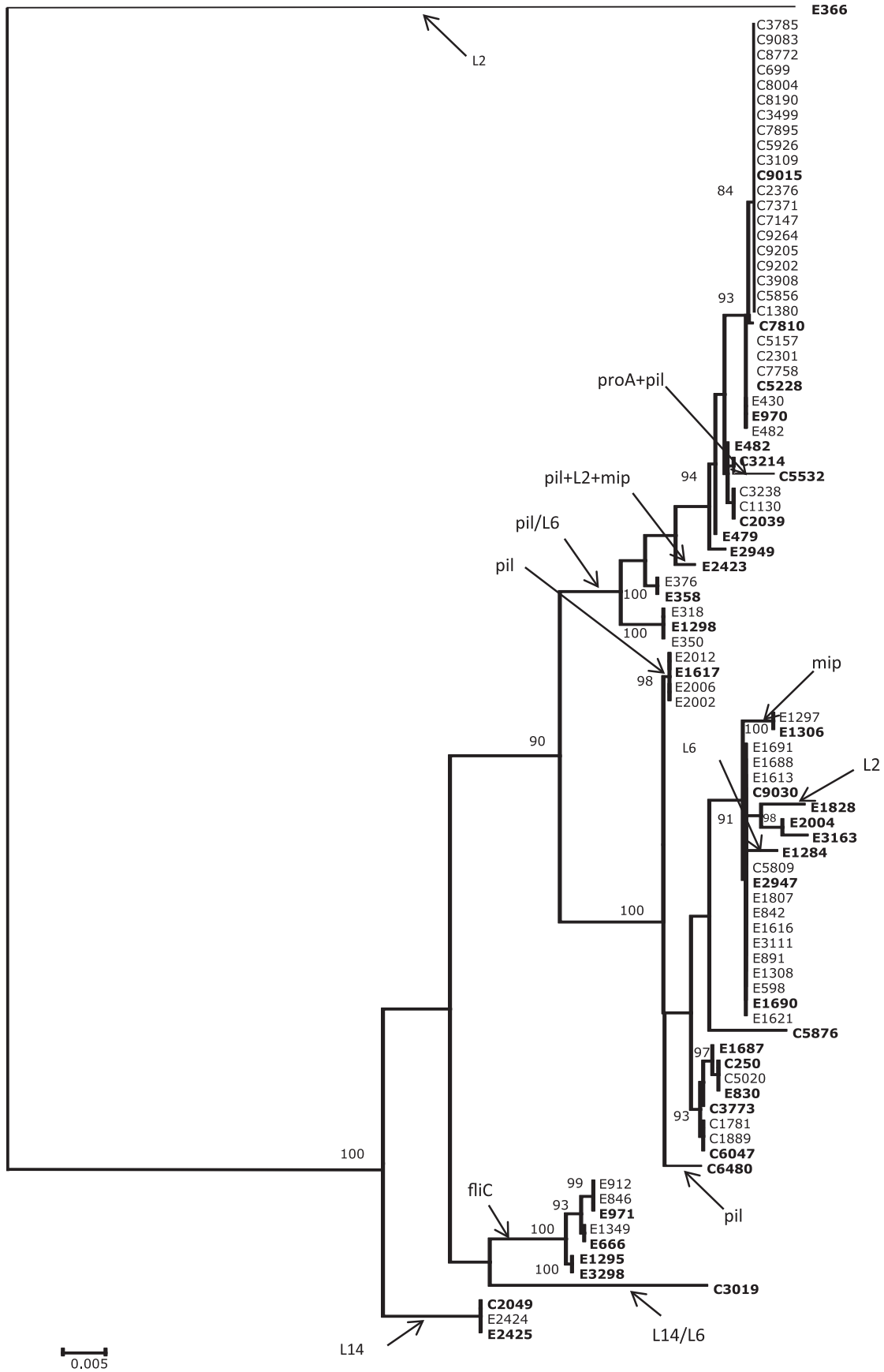

Fig. 1. ML tree obtained from the concatenated alignment of the nucleotide sequences for nine loci in 89 *Legionella pneumophila* samples of clinical and environmental origins. Bootstrap support values (1,000 replicates) higher than 70% are indicated. Recombination events detected by RDP with the corresponding loci are mapped onto the phylogenetic tree.

Table 2. Summary of SH and ELW Tests.

	cat9	*L14*	*proA*	*pilE*	*L2*	*mip*	*fliC*	*L6*	*asd*	*mompS*
cat9	—	22.91	70.30	61.23	169.38	152.76	19.91	139.20	57.05	72.63
L14	1837.51[a]	—	112.51	243.25	476.93	284.98	26.11	314.07	139.94	205.71
proA	3112.19[a]	305.20	—	386.21	452.29	403.50	241.54	375.85	243.06	294.56
pilE	5746.28 [a]	980.12	314.16	—	718.81	457.52	145.59	730.37	274.14	446.89
L2	3175.10	681.44	238.30	178.71	—	416.43	160.39	547.32	195.26	320.50
mip	2787.71	536.32	155.23	306.82	640.71	—	98.76	441.85	201.01	213.35
fliC	3812.77 [a]	620.83	166.64	220.05	544.58	432.18	—	556.03	174.62	268.70
L6	2235.01 [a]	388.07	191.34	241.26	425.34	389.65	52.74	—	137.49	203.77
asd	2440.05 [a]	457.88	104.32	357.87	252.15	441.17	79.11	325.17	—	227.92
mompS	2052.70 [a]	178.42	131.05	254.91	514.09	327.74	26.11	369.46	145.66	—

[a] All the cases were statistically significantly different ($P < 0.01$) by both tests except those labeled, which were significant only by the ELW test.

In order to estimate the relative contribution of recombination to the *L. pneumophila* nucleotide genetic diversity, we used several approaches. DnaSP estimates the recombination parameter (ρ) based on the variance of the average number of nucleotide differences between pairs of sequences (Hudson et al. 1987) and the estimate of the mutation rate according to Watterson (1975). LDhat implements an estimate of the population recombination rate using the approximate likelihood method, which is based on coalescent theory. Finally, we applied a Bayesian procedure, implemented in the program ClonalFrame, which is based on the estimation of the fraction of an alignment that has not undergone recombination.

Table 3. Recombination Events Detected by Phylogenetic Incongruence (column P), RDP (column R), and Deviation of the Variation Expected (column V) for Each Locus and Haplotype Derived after Concatenation of Nine Loci.

	L14		*proA*		*pilE*		*L2*		*mip*		*fliC*		*L6*		*asd*		*mompS*		
Haplotype	**P**	**R**	**V**	**P**	**R**	**V**	**P**	**R**	**V**	**P**	**R**	**V**	**P**	**R**	**V**	**P**	**R**	**V**	**P**	**R**	**V**
E366								1													
E2947																					
E1284														1							
E1690																					
C9030																					
E2004								2													
E3163								2													
E1828								4													
E1306											2										
E1687																					
E830																					
C250																					
C3773																					
C6047																					
E1617							1														
C6480							1														
C5876										3											
E2949				2			1							1							
E479				2			1							1							
C7810							1			1				1							
C9015							1			1				1							
C5228							1			1				1							
E970							1			1				1							
C2039							1						1	1							
E482							1						1	1							
C3214							1						1	1							
C5532				1			1						1	1							
E2423				2			1		5					1			1				
E358				2			1		3					2							
E1298							1		3												
E971																					
E666																					
E1295																					
E3298																					
C3019	2																				
C2049	1																				
E2425	1						1														

Shading indicates that the same recombination event was detected by the corresponding methods. Numbers in the R columns denote different events in the same locus. All the events are mapped onto the phylogenetic tree shown in figure 1.

Both LDhat and ClonalFrame showed very similar estimates for the ratio between the population recombination rate (ρ) and the population mutation rate (θ) ($\rho/\theta = 0.407$ for LDhat and $\rho/\theta = 0.44-0.87$ for ClonalFrame in the 95% credibility region). Contrarily, the DnaSP estimate of this ratio was much lower ($\rho/\theta = 0.1$), although of same order of magnitude (see Discussion for possible causes). We also estimated the ratio of nucleotide changes as the result of recombination relative to point mutations using ClonalFrame, and this yielded an estimate for $r/m = 2.6–5.7$ in the 95% credibility region. This means that it is approximately four times more likely that a site in this multiple alignment has changed by recombination than by mutation. A total of 103 recombination events of an average length of 305 bp and which introduced 755 substitutions were found using this method. The evolutionary tree reconstructed by ClonalFrame (supplementary fig. 2, Supplementary Material online) was coincident with the ML tree shown in figure 1, thus allowing the same mapping of recombination events.

The Syntenic Genome of *L. pneumophila*

The alignment of the four complete *L. pneumophila* genomes considered (strains Philadelphia, Lens, Paris, and Corby) revealed 38 syntenic regions spanning 2,920,612 bp (83% of an average *Legionella* genome). The analysis of these 38 blocks with RDP revealed 706 recombinant fragments that met the filtering criteria (see Material and Methods) with an average length of 1,431 bp (range = 130–33,610 bp; median = 769 bp). The most frequent fragment sizes were found in the 150–450 bp range, which is very similar to the mean recombinant fragment size (305 bp) derived in the population-based approach described above (see supplementary fig. 3, Supplementary Material

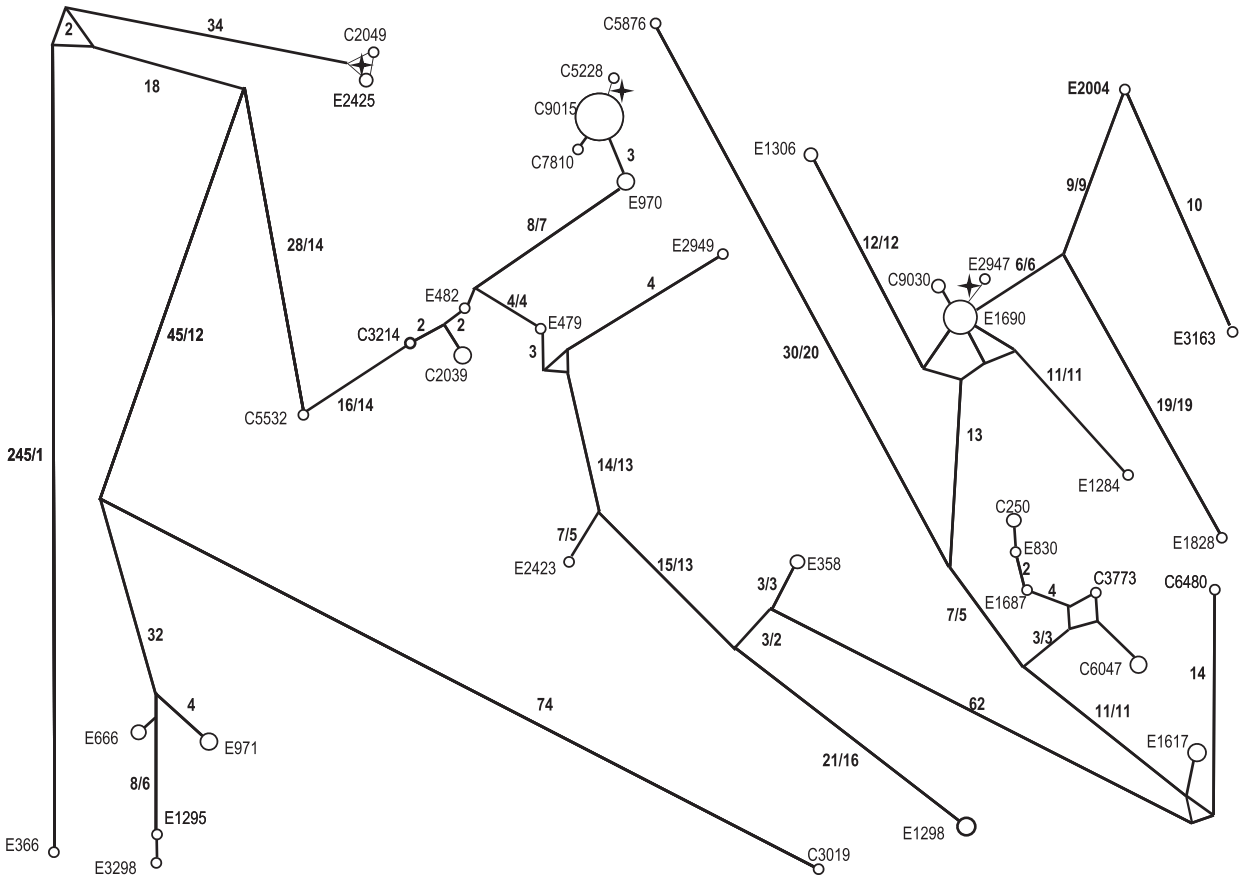

Fig. 2. Median-joining network for the 37 different haplotypes identified among the 89 clinical and environmental *Legionella pneumophila* samples analyzed. Numbers next to each branch indicate the number of nucleotide substitutions between the flanking nodes when larger than 1. Values after a slash represent the number of substitutions inferred to be due to a recombination event. Stars indicate differences between haplotypes due to an insertion/deletion. Circle sizes are proportional to the number of samples with the corresponding haplotype.

online, for the distribution of fragment sizes). The total length of the genome involved in recombination events according to this procedure was 1,010,086 bp, which represents 34.55% of the syntenic genome and involves 42.76% of the genes (1,172/2,741). Several hotspot regions for recombination were detected in the syntenic genome (fig. 3), and a sliding-window scan of genetic diversity in the syntenic genome revealed a moderate but significant positive correlation with the number of recombinant fragments in that window (Spearman's $\rho = 0.5598$, P value < 0.001). We verified the existence of these hotspot regions by testing the number of events detected per fragment of size 4,136 bp against a Poisson distribution with one recombination event per fragment as a null hypothesis. The result was highly significant (chi-square = 901.57, degrees of freedom [df] = 4, P value = 0, supplementary fig. 4, Supplementary Material online), corroborating the existence of hotspot regions for recombination. We observed that several hotspot regions clustered contiguously in the syntenic genome. The two regions with the largest number of contiguous fragments with high numbers of recombination events (R1, 110 kbp, and R2, 50 kbp) are shown in figure 3. A complete list of the genes included in regions R1 and R2 is provided as supplementary table S3, Supplementary Material online.

To compare the estimates of recombination in the four genomes with those from population-based data, we also applied ClonalFrame to the syntenic alignment of the four *L. pneumophila* strains. The analysis retrieved 1,703 recombination events with an average length of 220 bp and which were responsible for 49,136 nucleotide differences among the strains (fig. 4). Using this approach, the proportion of the syntenic genome affected by recombination was found to be around 57%, which is larger than the fraction derived using the RDP approach (34.55%). This was an expected result because, under our criteria, the RDP approach to detect recombination events is more conservative than the Bayesian procedure. Nevertheless, both estimates show the relevance of recombination in shaping *L. pneumophila* diversity at the genome level in agreement with the values observed in the MLST approach. Similarly, the Bayesian estimate of the ratio of nucleotide changes as the result of recombination relative to point mutation (r/m) comprised between 0.280 and 0.313.

The *L. pneumophila* Pangenome and the Impact of HGT

The analysis of single-copy genes present in at least one of the four *L. pneumophila* genomes revealed the pangenome of this species to consist of 3,846 genes (supplementary

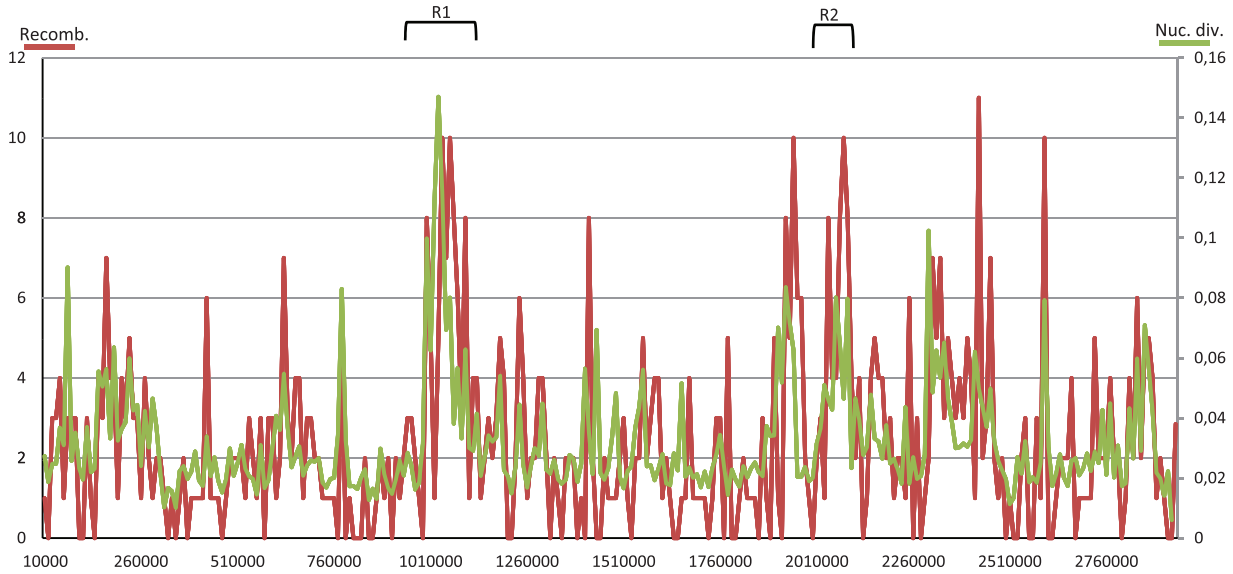

Fig. 3. Distribution of the identified recombination events and estimates of genetic diversity along the syntenic genome of four *Legionella pneumophila* strains. The two hotspot regions for recombination (R1 and R2) commented in the text are indicated.

fig. S5, Supplementary Material online). The core genome, the set of genes that are common to all the strains considered, comprised 2,418 genes (62% of the pangenome). Therefore, the so-called dispensable genome accounted for the remaining 38% of the pangenome (1,428 genes). Among these, 895 were present in only one strain thus indicating the importance of gene gains and losses in shaping the dispensable genome and showing that the rate of gain/loss is larger at the tips of the tree (supplementary fig. 6, Supplementary Material online).

The relevance of HGT in shaping the *L. pneumophila* pangenome was explored using phylogenetic tests. For this, we tried to include in the analyses the genome sequence of potential partners of *L. pneumophila* in horizontal transfers. It has been shown that *Legionella* can coexist with a range of taxonomically different species within several amoeba species. Our survey of the published research revealed more than 40 additional bacterial species found in the cytoplasm of amoebas (supplementary table S4, Supplementary Material online). We used a reference set of 37 bacterial genomes (Table 4), which corresponded either to other intracellular parasites of amoebas (or their closest relative with a completely sequenced genome) or to representative species of other important bacterial families not included in the previous set. To test the phylogenetic origin of the *L. pneumophila* pangenome genes, we generated

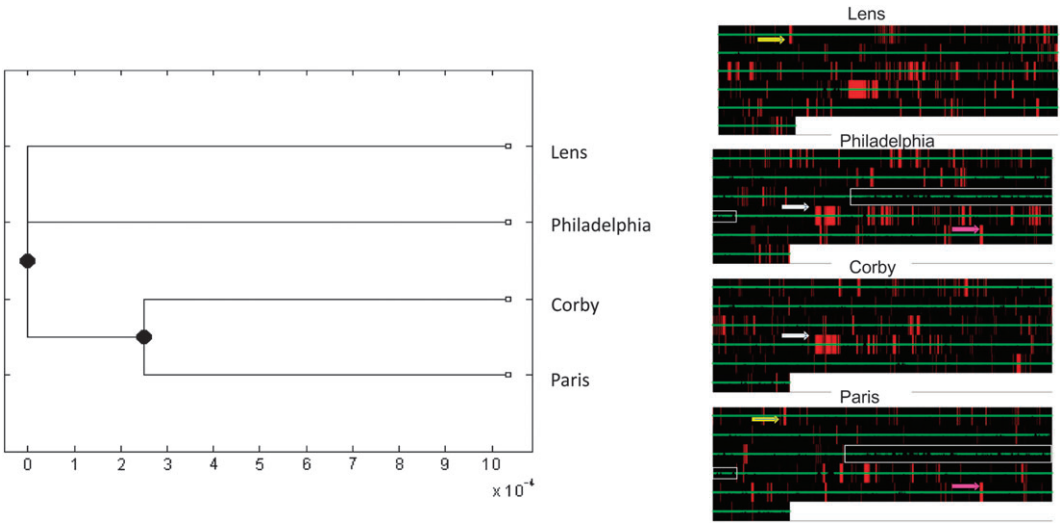

Fig. 4. Analysis with ClonalFrame of the syntenic genome of four *Legionella pneumophila* strains. The genealogy on the left represents the consensus based on the posterior probabilities of two independent runs of the program. The panel on the right depicts the mutation (green) and recombination (red) events inferred in each branch of the genealogy. Each line corresponds to 300 kb in the syntenic genome. White arrows represent a large recombination stretch common to Philadelphia and Corby, and other arrows represent common recombination events of Lens and Paris (yellow) and Philadelphia and Paris (pink) strains. Philadelphia and Corby share a large (white square) stretch with fewer mutations than the rest of their genomes and the homologous area in the two other strains.

Table 4. Completely Sequenced Bacterial Genomes Used in the Analysis of Putative HGT Events in the *Legionella pneumophila* Pangenome.

Species	Abbreviation	Genes	Accession No.
Acidobacteria bacterium Ellin345	abe	1,214	NC_008009
Anabaena variabilis ATCC 29413	ava	1,055	NC_007413
Brucella abortus biovar 1 str. 9-941	bab	977	NC_006932/3
Bdellovibrio bacteriovorus HD100	bbh	1,119	NC_005363
Bacillus cereus ATCC 14579	bc1	1,141	NC_004722
Burkholderia cepacia AMMD	bca	1,351	NC_008391/2
Burkholderia pseudomallei 1710b	bp1	1,301	NC_007434/5
Coxiella burnetii RSA 493	cbr	1,237	NC_002971
Caulobacter crescentus CB15	ccc	1,285	NC_002696
Clostridium perfringens ATCC 13124	cpa	876	NC_008261
Chlamydophila pneumoniae CWL029	cpc	531	NC_000922
Chlorobium tepidum TLS	ctt	941	NC_002932
Deinococcus radiodurans R1	drr	889	NC_001263
Escherichia coli O157:H7 EDL933	ece	1,397	NC_002655
Francisella tularensis subsp. *tularensis* FSC198	ftt	1,067	NC_008245
Gramella forsetii KT0803	gfk	1,033	NC_008571
Haemophilus influenzae 86-028NP	hi8	1,015	NC_007146
Helicobacter pylori 26695	hp2	748	NC_000915
Leptospira interrogans sv. *Copenhageni* str. *Fiocruz* L1-130	lic	976	NC_005823/4
Listeria monocytogenes EGD-e	lme	965	NC_003210
Lactobacillus salivarius subsp. *salivarius* UCC118	lsu	736	NC_007929
Mycobacterium avium 104	ma1	930	NC_008595
Mycobacterium bovis AF2122/97	mba	945	NC_002945
Mycoplasma pulmonis UAB CTIP	mpu	299	NC_002771
Mycobacterium smegmatis str. MC2 155	msm	1,079	NC_008596
Nitrosomonas europaea ATCC 19718	nea	1,238	NC_004757
Neisseria gonorrhoeae FA 1090	ngf	1065	NC_002946
Nostoc sp. PCC 7120	npc	1,038	NC_003272
Pseudomonas aeruginosa PAO1	pap	1,578	NC_002516
Candidatus Protochlamydia amoebophila UWE25	pau	802	NC_005861
Rickettsia bellii RML369-C	rbr	820	NC_007940
Rickettsia felis URRWXCal2	rfu	792	NC_007109
Rhodopseudomonas palustris BisA53	rpb	1,236	NC_008435
Salmonella enterica subsp. *enterica* serovar Typhi str. CT18	sec	1,398	NC_003198
Vibrio cholerae O1 biovar ElTor str. N16961	vco	1,275	NC_002505/6
Xanthomonas axonopodis pv. citri str. 306	xac	1,457	NC_003919
Yersinia pestis KIM	ypk	1,328	NC_004088

The number of genes orthologous to the set of 1,700 *L. pneumophila* genes considered in this study is shown for each genome.

a phylogenetic tree (fig. 5) based on the concatenated alignment of the 104 homologous genes common to these 37 genomes and the 4 *L. pneumophila* strains (fig. 5 and supplementary table S5, Supplementary Material online). The reference tree mostly agreed with the accepted bacterial phylogeny with the exception of *Mycoplasma pulmonis* and *Helicobacter pylori*, which grouped unexpectedly together, although *Mycoplasma* belongs to the phylum Firmicutes. Consequently to avoid potential confounding effects, they were excluded from the phylogenetic origin analysis.

Using the reference phylogeny, we tested 12 possible evolutionary hypotheses for the 1,700 *L. pneumophila* genes that met our criteria for inclusion in this analysis (see Material and Methods and supplementary table S6, Supplementary Material online, for details). These genes encode 599,459 amino acids in *L. pneumophila*, which were reduced to 434,355 sites after the multiple alignments were trimmed of poorly aligned positions with Gblocks. Ten hypotheses corresponded to assigning the origin of the *Legionella* gene to the main bacterial groups considered and the other two hypotheses corresponded to control phylogenies (one fully unresolved and another only resolved within classes). The unresolved phylogenies allowed us to test for the existence of enough phylogenetic signal for evolutionary inference in each gene of this data set, especially after the removal of ambiguously aligned positions using Gblocks with default parameters.

SH and ELW tests of the phylogenetic origins considered for each gene resulted in 182 (10.7%) phylogenies being best supported by one of the two control phylogenies in which the *Legionella* genes grouped with the Gamma-Proteobacteria but without enough signal to resolve the relationships among the ten bacterial groups considered. Of the 1,518 informative genes, 814 (47.88% of the total) grouped with statistical significance with the Gamma-Proteobacteria clade and 704 genes (46.3%) carried a phylogenetic signal from a non-Gamma-Proteobacteria group, therefore being likely involved in HGT events (fig. 5).

We used the average 16S rDNA nucleotide distance between the genomes from each particular group to the *L. pneumophila* strains as a surrogate of their evolutionary distance to estimate the correlation between the number of HGT events and the phylogenetic distance. This

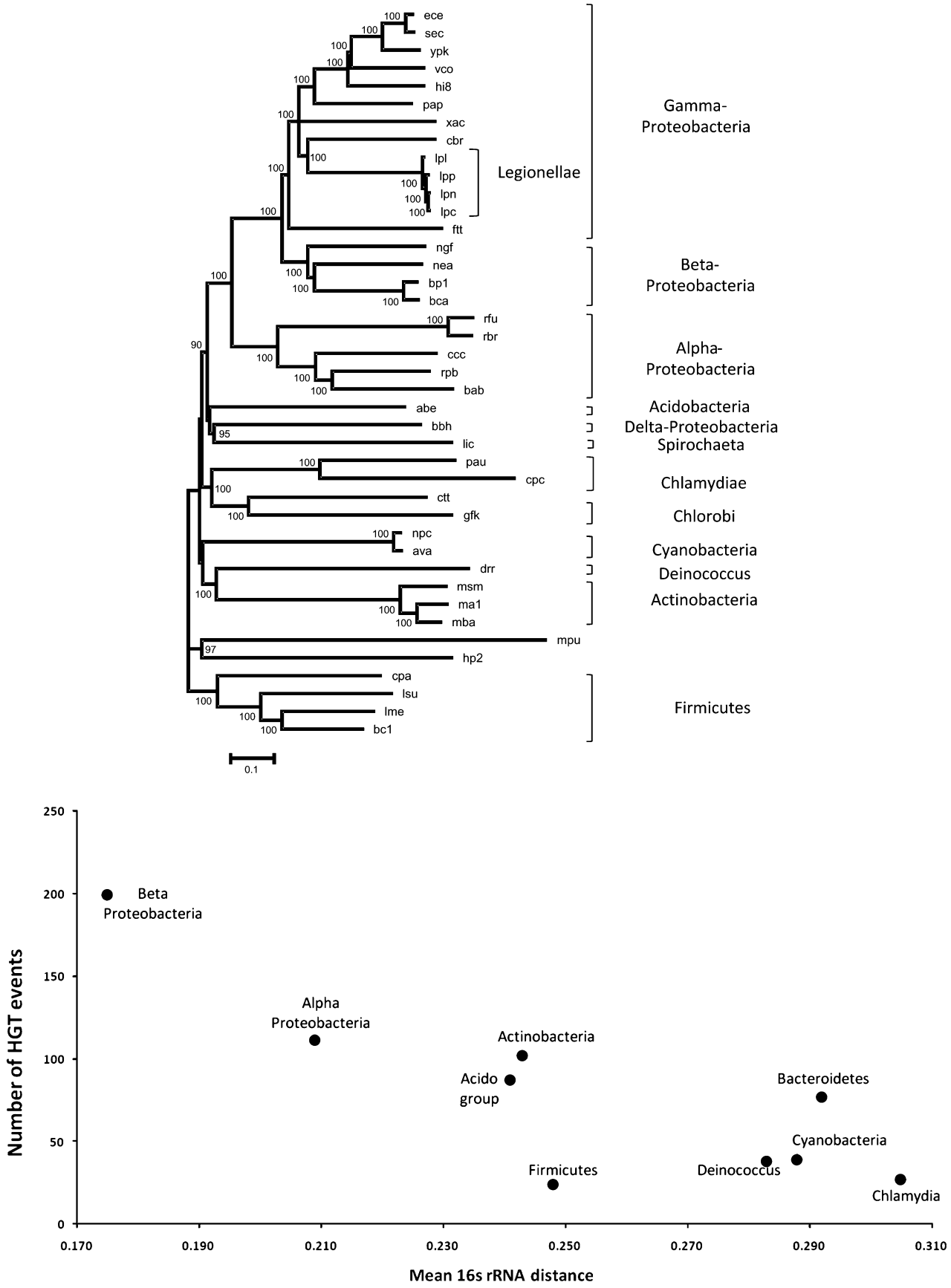

Fig. 5. Reference phylogenetic tree obtained by ML analysis of the concatenated alignment of 104 genes from 41 bacterial species (upper panel). Relationship between the number of detected HGT events and phylogenetic distance estimated from 16S rDNA genes for the non-Gamma-Proteobacteria groups considered and *Legionella pneumophila* strains (lower panel).

resulted in a highly significant correlation (Spearman's $\rho = -0.818$, P value $= 0.0038$) revealing a strong effect of phylogenetic distance on horizontal transfer (table 5 and fig. 5). Furthermore, when genes with a Gamma-Proteobacteria phylogenetic origin were removed, thus leaving in the analysis only those genes likely involved in HGT events, the correlation was still significant (Spearman's $\rho = -0.750$, P value $= 0.0199$) (fig. 5). All the groups contributed similarly to this correlation (data not shown) except Firmicutes, which were involved in fewer HGT events than expected (Spearman's $\rho = -0.881$, P value $= 0.0039$ after their removal).

We compared the distribution of functional categories in the HGT-related genes detected with respect to those in the *Legionella* pangenome. The comparison was performed with genes assigned to only one functional category. This resulted in 557 putative HGT genes compared with 2,064 genes in the pangenome. The test for four major functional categories revealed significant differences (chi-square $=$ 10.87, df $= 3$, P value $= 0.012$). As expected, the metabolism category is overrepresented among HGT genes, whereas information-related genes are underrepresented (see supplementary table S6 and fig. 7, Supplementary Material online).

Discussion

Evidence for a Nonclonal Population Structure of *L. pneumophila* from Natural Populations and Whole Genomes

The assumption that bacterial population structure is strictly clonal is based on the prevalence of mutation over recombination in shaping the extent and distribution of genetic variation (Levin 1981; Maynard Smith et al. 1993). The clonal population model has been invalidated for many bacterial species (Feil et al. 1999) in which a nucleotide is several times more likely to change as a result of recombination, or gene conversion, than through mutation. The existence of a worldwide dominant clone among disease-causing *L. pneumophila* and the absence of evidence of intragenic recombination in its populations has lead to the conclusion that *L. pneumophila* is essentially clonal (Edwards et al. 2008) despite possessing molecular mechanisms for homologous and nonhomologous recombination (Mintz 1999; Stone and Kwaik 1999) and inhabiting an ecological niche generally associated with high levels of recombination. Our analysis demonstrates at two levels, that is, population genetic and genomic, not only that recombination is present in *L. pneumophila* but also that it has made and still makes major contributions to its genetic diversity.

The detection of a recombination event is influenced by several factors, such as the genetic diversity of the studied region, the time elapsed since the event took place, the sampling size, and the demographic history of the population(s) (Stephens 1986). So, disparities in the results derived from recombination estimates between the MLST and genomic approaches are probably explained by differences in sample size and genetic diversity and history between the two data sets. The isolates used in the MLST analysis were obtained in a geographically reduced area, although they encompass relatively large genetic variation compared with samples from European countries (Coscollá et al. 2006), whereas the four strains whose genomes we have analyzed here were obtained in the United States, France (two strains), and the United Kingdom. More importantly, the fraction of the genome used in both analyses is remarkably different (3,923 and 2,920,612 bp for MLST and genome analysis, respectively). Also, the different methods (Bayesian and parametric) used in the analysis of these data have different assumptions making the complete congruence among them very unlikely (Martin and Beiko 2010). Therefore, it is more appropriate to compare the results obtained for each data set with other similar analyses.

Vos and Didelot (2009) reviewed the relative impact of recombination and point mutation on genetic variability in several MLST data sets of bacterial species using ClonalFrame. These analyses revealed a wide range of values for the ratio of the two rates that spanned three orders of magnitude, from 0.02 in *Leptospira interrogans* to 63 in *Flavobacterium psychrophilum* and *Pelagibacter ubique* (SAR 11). Estimates of the same ratio from MLST analysis in *L. pneumophila* in this study ($r/m = 2.6$–5.7 and $\rho/\theta = 0.44$–0.87) place this bacteria in a central position of the distribution, with similar values to those of *Campylobacter insulaenigrae*, *M. hyopneumoniae*, *Haemophilus parasuis*, *C. jejuni*, *Halorubrum* sp., *Pseudomonas syringae*, and *P. viridiflava*, most of them also commensal or opportunistic pathogens like *L. pneumophila*. The results obtained in this study also contrast with the value reported for this species ($r/m = 0.9$) by Vos and Didelot (2009), who used only 2 of the 19 loci analyzed in Coscollá and González-Candelas (2007). The difference can be explained by the lower representation of the diversity in this *L. pneumophila* sample as only 2 loci and 12 sequence types were considered by Vos and Didelot.

Perez-Losada et al. (2006) established three groups of bacteria according to their intragenic population mutation rates obtained with LDhat (McVean et al. 2002). The first group, defined by $\rho > 50$, comprised *H. pylori*, *Neisseria gonorrhoeae*, *N. meningitidis*, and *S. pneumoniae*; the second one, with $15 < \rho < 50$, corresponding to *Bacillus cereus*, *H. influenzae*, *Streptococcus galactiae*, and *S. pyogenes*; and finally, *Burkholderia pseudomallei*, *Moraxella catarrhalis*, *Staphylococcus epidermidis*, *Vibrio vulnificus*, *C. jejuni*, *Enterococcus faecium*, *Escherichia coli*, and *St. aureus* showed $\rho < 15$. The population mutation rate obtained in our study places *L. pneumophila* in this third group, with very low values of the intragenic recombination rate ($\rho \cong 0.01$). In consequence, with regard to recombination, *L. pneumophila* might present a population structure similar to that of *E. coli* because both species show high intergenic, even higher than bacteria such as *St. aureus*, but low intragenic recombination rates.

There are few studies analyzing the contribution of recombination and mutation to genetic variation at the population genomic level. Touchon et al. (2009) have performed the so far most comprehensive of such an analysis with the core genome of 20 *E. coli* strains. Despite the widely held view of this as a clonal species, their results reveal a higher relevance of recombination (i.e., gene conversion) over mutation in changing the state at individual nucleotide sites. Didelot et al. (2007) compared the genomes of four *Salmonella enterica* serovar *Typhimurium* using the same Bayesian approach that we have used here. Compared with their results, we have detected in *L. pneumophila* smaller clonal frames (220 vs. 800 bp) but more events (1,703 vs. 50 bp) than in *Sa. enterica*, but the overall contribution of recombination to the generation of genetic variation is similar in both species. The relative impact of recombination versus point mutation derived from the genome analysis of *L. pneumophila* ($r/m = 0.280$ and 0.313) is lower than that for *Clostidium difficile* (0.63 and 1.13) based also on whole-genome sequences but in this case derived from 30 isolates (He et al. 2010).

Our analysis of the clonal fraction in the syntenic genome of *L. pneumophila* has revealed that 34–57% of the genome has been involved in recombination events. These are not distributed evenly and several recombination hotspots were detected. Two regions seemed to accumulate more recombination events than any other. Among the genes involved in recombination events in the hotspot region R1, we detected many proteins related to the type 4A secretion system, which is used by *Legionella* to induce apoptosis of macrophages (Zink et al. 2002). Most of the genes in the hotspot region R2 are annotated as hypothetical genes, but it is interesting that the one involved in more recombination events (lpg1945) is a substrate (and therefore a potential virulence factor) of the *tatB* system which, along with the *sec* system, allows the initial transport of proteins to the periplasm where the type II secretion machinery releases the protein out of the bacterial cell (Rossier and Cianciotto 2005). It is tempting to speculate on a possible relationship between increased recombination and pathogenicity, but more detailed analyses are necessary to verify it.

Altogether, these results indicate that *L. pneumophila* is far from being a clonal species. Recombination among strains of this species is an ongoing process, and it has been operating also in the recent history of the species. This shift in our understanding of the basic population structure of bacterial species has been driven mainly by the increasing use of MLST data and, more recently, by the complete sequencing of genomes from strains and isolates of many species. Recent advances in sequencing technologies (Snyder et al. 2009; Metzker 2010) are anticipating even richer data sets for population genomics analyses (Harris et al. 2010; Holden et al. 2010). When these become widely available, it will be possible not only to have a better picture of the population structure of bacterial species but also to reconcile discordant results between estimates based on a few loci and those from complete genome information.

The nonclonality of *L. pneumophila* has multiple consequences from a public health perspective. Typing of microorganisms is an essential tool to determine the strains responsible for outbreaks and sources of infection and also to evaluate the proportion of infections due to new strains and reinfections from the pre-existing ones (Foxman and Riley 2001). The possibility that recombination eliminates differences among strains can have important consequences for public health interventions. Additionally, little is known about antibiotic resistance in *Legionella* but given that recombination has been associated to the spread of such traits and that even hyperrecombinant strains leading to drug resistance have been described for other bacteria (Hanage et al. 2009), the presence of important levels of recombination in *L. pneumophila* has to be taken into account to monitor possible drug-resistance emergence in this human pathogen. Antibiotic resistance genes usually spread in populations and among species through plasmids, and a recent analysis (Halary et al. 2010) has shown that these genetic elements have a major role in the recent exchange of DNA in *Legionella*, thus enabling the rapid spread in its populations of virulence- and antibiotic resistance–associated traits.

Phylogenetic Distance and Ecological Opportunity

HGT between species has been recognized as a major process in the evolution of Bacteria (Lawrence and Retchless 2009), leading even to questioning the nature and concept of bacterial species (Bapteste et al. 2009). Now, it is possible to start disentangling the factors that have a relevant role in the distribution of successful HGT events (Ragan and Beiko 2009).

Phylogenetic distance is expected to have a major effect on HGT, as closely related species usually share genome architectures (Hendrickson and Lawrence 2006) and nucleotide composition (Lucchini et al. 2006). However, ecological factors may also have a major impact in shaping bacterial gene contents. For example, intracellular bacteria like some pathogens and endosymbionts seldom show evidence of recent HGT or recombination (Toft and Andersson 2010).

Legionella pneumophila, with its dual lifestyle as a waterborne bacteria associated to biofilms and intracellular parasite of amoebas and as an opportunistic pathogen of humans (Fields et al. 2002), probably has more chances of exchanging genetic material than obligate intracellular pathogens but less than free-living bacteria, including commensal and opportunistic parasites (Vos and Didelot 2009). The ability of *Legionella* to infect amoebas has surely played a key role in the successful gene exchange with other bacteria (Salah et al. 2009).

In order to prevent artifacts in phylogenetic inferences, which might lead to wrong conclusions, we have restricted our analyses to genes that have a strong phylogenetic signal. Even in this subset of the *L. pneumophila* pangenome, there is a substantial, but nor universal, support for the assignment of *Legionella* to the Gamma-Proteobacteria, where it occupies a basal position. This major signal

coexists in the genome with a substantial number of genes (around 41% among the 1,700 genes studied) with strong statistical evidence for having been involved in horizontal transfers with other phyla.

We have previously shown (Comas et al. 2006) that another Proteobacteria group, the Xanthomonadales, also presents a high level of genome mosaicism due to a high rate of gene exchange with other Proteobacteria species which makes practically impossible its assignment to any of the main Proteobacterial divisions. However, unlike Xanthomonadales, in *Legionella*, there is a main Gamma-Proteobacterial phylogenetic signal, present in more than half of its genes, which justifies its inclusion in this class. Although HGT has played an important role in the evolution of *Legionella*, its Gamma-Proteobacterial nature is corroborated by several results. First, the phylogeny obtained after concatenation of the 104 common genes to all the genomes considered (fig. 5) in which the grouping of *Legionella* within the Gamma-Proteobacteria is well supported by bootstrap analyses, and second, the majority of genes tested in the phylogenetic analysis had a Gamma-Proteobacteria origin (996 over 1,700 or almost 58.6%). This does not mean that HGT is restricted only to the remaining 41.4% because we have not tested the possible existence of HGT within the Gamma-Proteobacteria. Finally, the position of Legionellaceae in the 16S rDNA tree also corresponds to a basal group within Gamma-Proteobacteria. Still this Gamma-Proteobacteria signal is present in only 59% of the genes tested, corroborating the high degree of mosaicism observed in *Legionella*.

Our analysis of the phylogenetic origin of *L. pneumophila* genes has focused in bacterial taxa likely involved in transfer events with *Legionella*. Given that we have detected HGT events to/from all known bacterial groups present in amoebas and that other groups have reported extensive HGT among amoeba-associated bacteria (Chien et al. 2004; Ogata et al. 2006) as well as the presence of eukaryote-like genes in *Legionella* most likely of amoeba origin (Lurie-Weinberger et al. 2010; Schmitz-Esser et al. 2010), we think that the most likely explanation for the multiple phylogenetic origins of a significant fraction of *Legionella* genes is the exchange of genetic material in the common amoeba host. Nevertheless, there are other alternative explanations for this observation such as horizontal gene exchange mediated by shared phages that allow to overcome geographic and phylogenetic barriers (Thomas and Nielsen 2005) or interchanges with bacteria in natural aquatic environments, where many of the groups considered in this analysis are also present (Tamames et al. 2010). To discern between these alternatives, more detailed analyses using complete genomes of amoeba-associated and closely related but nonassociated species will be needed.

Our analyses have revealed a strong correlation between the number of detected HGT events and the evolutionary distance between *Legionella* and their partners. This correlation likely arises as a by-product of the many traits shared by common ancestry by closely related species, such as genome architecture (Hendrickson and Lawrence 2006), the control of foreign DNA through methylation (Waldron and Lindsay 2006), or the detection of abnormal G+C composition (Lucchini et al. 2006; Navarre et al. 2006), which along with ecological opportunity play an important role in determining the likelihood of a successful HGT event. Our results indicate that the coexistence in the cytoplasm of amoebas, and possibly also in biofilms, has provided ample opportunities for lateral gene transfer among the diverse array of bacterial species found in these habitats. This is also reflected in the presence, in the *Legionella* genome, of several genes of eukaryotic origin (Cazalet et al. 2004), which likely provide an adaptive advantage for the intracellular parasitic life style of *Legionella*.

Among the genes detected to be involved in HGT, there is an overrepresentation of those involved in secondary metabolism, whereas those involved in information processes are underrepresented. This result corroborates observations in other bacteria (Nakamura et al. 2004; Beiko et al. 2005) for which exchanges between species are less likely when essential genes are involved. Nevertheless, this is only a trend, and the results reported here and in other analyses show that informational genes are frequently involved in HGT events (Sorek et al. 2007; Hao and Golding 2008).

The consequences of the capacity of *Legionella* to incorporate foreign genetic material extend beyond the mosaic composition of its genome. It has been proposed that amoebas have provided a training ground for pathogenic species able to infect human macrophages by allowing them to incorporate eukaryotic genetic material into their genomes (Salah et al. 2009). This hypothesis is supported by the recent sequencing of an amoeba genome (Schmitz-Esser et al. 2010), and experiments with *Mycobacterium avium* that show how a particular pathogenicity island encodes for proteins involved in the invasion of both amoebas and macrophages (Danelishvili et al. 2007). Therefore, life in the interior of amoebas may have increased the capacity of *Legionella*, as well as that of other intracellular pathogens of human macrophages, for invading these cell types using similar mechanisms, which represents an extraordinary case of convergent evolution mediated by HGT among bacterial pathogens.

Supplementary Material

Supplementary figures S1 to S7 and supplementary tables S1 to S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Aida Rodrigo for technical assistance, and Centro de Salut Publica de Alcoi, Centro de Salut Publica de Benidorm, Hospital Virgen de los Lirios de Alcoi, Dirección General de Salud Pública (Conselleria de Sanitat, Generalitat Valenciana), and Sanidad Ambiental (Conselleria de Territori i Habitatge, Generalitat Valenciana) for providing *Legionella* samples.

This research was funded by Conselleria de Sanitat (Generalitat Valenciana) and projects BFU2008-03000 from Ministerio de Ciencia e Innovación and ACOMP/2009/240 and ACOMP/2010/148 from Conselleria d'Educació (Generalitat Valenciana). M.C. benefited from a fellowship from Generalitat Valenciana. I.C. is supported by MRC core funds (MRC_U117588500).

References

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Inform. Tech. *Biomed*. 19:716–723.

Atwood KC, Schneider LK, Ryan FJ. 1951. Periodic selection in *Escherichia coli*. *Proc Natl Acad Sci U S A*. 37:146–155.

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol*. 16:37–48.

Bapteste E, O'Malley M, Beiko R, et al. (11 co-authors). 2009. Prokaryotic evolution and the tree of life are two different things. *Biol Direct*. 4:34.

Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U S A*. 102:14332–14337.

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

Cazalet C, Rusniok C, Bruggemann H, et al. (14 co-authors). 2004. Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. *Nat Genet*. 36:1165–1173.

Chien M, Morozova I, Shi S, et al. (37 co-authors). 2004. The genomic sequence of the accidental pathogen *Legionella pneumophila*. *Science* 305:1966–1968.

Comas I, Moya A, Azad RK, Lawrence JG, Gonzalez-Candelas F. 2006. The evolutionary origin of Xanthomonadales genomes and the nature of the horizontal gene transfer process. *Mol Biol Evol*. 23:2049–2057.

Coscollá M, González-Candelas F. 2007. Population structure and recombination in environmental isolates of *Legionella pneumophila*. *Environ Microbiol*. 9:643–656.

Coscollá M, Gosalbes MJ, Catalán V, Gomzález-Candelas F. 2006. Genetic variation in environmental samples of *Legionella pneumophila* from the Comunidad Valenciana (Spain). *Environ Microbiol*. 8:1056–1063.

Dagan T, Martin W. 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci U S A*. 104:870–875.

Danelishvili L, Wu M, Stang B, Harriff M, Cirillo S, Cirillo J, Bildfell R, Arbogast B, Bermudez LE. 2007. Identification of *Mycobacterium avium* pathogenicity island important for macrophage and amoeba infection. *Proc Natl Acad Sci U S A*. 104:11038–11043.

Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res*. 14:1394–1403.

de Felipe KS, Pampou S, Jovanovic OS, Pericone CD, Ye SF, Kalachikov S, Shuman HA. 2005. Evidence for acquisition of *Legionella* type IV secretion substrates via interdomain horizontal gene transfer. *J Bacteriol*. 187:7716–7726.

Didelot X, Achtman M, Parkhill J, Thomson NR, Falush D. 2007. A bimodal pattern of relatedness between the *Salmonella* Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? *Genome Res*. 17:61–68.

Didelot X, Falush D. 2007. Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175:1251–1266.

Didelot X, Maiden MCJ. 2010. Impact of recombination on bacterial evolution. *Trends Microbiol*. 18:315–322.

Edwards MT, Fry NK, Harrison TJ. 2008. Clonal population structure of *Legionella pneumophila* inferred from allelic profiling. *Microbiology* 154:852–864.

Enright MC, Spratt BG. 1999. Multilocus sequence typing. *Trends Microbiol*. 7:482–487.

Feil EJ, Maiden MC, Achtman M, Spratt BG. 1999. The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Mol Biol Evol*. 16:1496–1502.

Fields BS. 1996. The molecular ecology of *Legionellae*. *Trends Microbiol*. 4:286–290.

Fields BS, Benson RF, Besser RE. 2002. Legionella and Legionnaires' disease: 25 years of investigation. *Clin Microbiol Rev*. 15:506–526.

Fliermans CB, Cherry WB, Orrison LH, Smith SJ, Tison DL, Pope DH. 1981. Ecological distribution of *Legionella pneumophila*. *Appl Environ Microbiol*. 41:9–16.

Foxman B, Riley L. 2001. Molecular epidemiology: focus on infection. *Am J Epidemiol*. 153:1135–1141.

Fraser DW, Tsai TR, Orenstein W, et al. (12 co-authors). 1977. Legionnaires' disease: description of an epidemic of pneumonia. *New England J Med*. 297:1189–1197.

Gaia V, Fry NK, Afshar B, Luck PC, Meugnier H, Etienne J, Peduzzi R, Harrison TJ. 2005. Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of *Legionella pneumophila*. *J Clin Microbiol*. 43:2047–2052.

Gao B, Mohan R, Gupta RS. 2009. Phylogenomics and protein signatures elucidating the evolutionary relationships among the Gammaproteobacteria. *Int J Syst Evol Microbiol*. 59:234–247.

Gelman A, Rubin DB. 1992. Inference from iterative simulation using multiple sequences. *Stat Sci*. 7:457–472.

Gibbs MJ, Armstrong JS, Gibbs AJ. 2000. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* 16:573–582.

Gómez-Valero L, Rusniok C, Buchrieser C. 2009. *Legionella pneumophila*: population genetics, phylogeny and genomics. *Infect Genet Evol*. 9:727–739.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52:696–704.

Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E. 2010. Network analyses structure genetic diversity in independent genetic worlds. *Proc Natl Acad Sci U S A*. 107:127–132.

Hanage WP, Fraser C, Tang J, Connor TR, Corander J. 2009. Hyper-recombination, diversity, and antibiotic resistance in *Pneumococcus*. *Science* 324:1454–1457.

Hao W, Golding GB. 2008. Uncovering rate variation of lateral gene transfer during bacterial genome evolution. *BMC Genomics*. 9:235.

Harris SR, Feil EJ, Holden MTG, et al. (15 co-authors). 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 327:469–474.

He M, Sebaihia M, Lawley TD, et al. (22 co-authors). 2010. Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc Natl Acad Sci U S A*. 107:7527–7532.

Heinz E, Kolarov I, Kastner C, Toenshoff ER, Wagner M, Horn M. 2007. An *Acanthamoeba* sp. containing two phylogenetically different bacterial endosymbionts. *Environ Microbiol*. 9:1604–1609.

Hendrickson H, Lawrence JG. 2006. Selection for chromosome architecture in bacteria. *J Mol Evol*. 62:615–629.

Holden MTG, Lindsay JA, Corton C, Quail MA, Cockfield JD, Pathak S, Batra R, Parkhill J, Bentley SD, Edgeworth JD. 2010. Genome sequence of a recently emerged, highly transmissible, multi-antibiotic- and antiseptic-resistant variant of methicillin-resistant *Staphylococcus aureus*, sequence type 239 (TW). *J Bacteriol*. 192:888–892.

Horn M, Wagner M. 2004. Bacterial endosymbionts of free-living amoebae. *J Eukaryot Microbiol*. 51:509–514.

Hudson RR, Kreitman ME, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153–159.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comp App Biosci*. 8:275–282.

Ko KS, Hong SK, Lee HK, Park MY, Kook YH. 2003. Molecular evolution of the *dotA* gene in *Legionella pneumophila*. *J Bacteriol*. 185:6269–6277.

Ko KS, Lee HK, Park MY, Park MS, Lee KH, Woo SY, Yun YJ, Kook YH. 2002. Population genetic structure of *Legionella pneumophila* inferred from RNA polymerase gene (*rpoB*) and DotA gene (*dotA*) sequences. *J Bacteriol*. 184:2123–2130.

Lawrence JG. 2005. Common themes in the genome strategies of pathogens. *Curr Opin Genet Dev*. 15:584–588.

Lawrence JG, Retchless AC. 2009. The interplay of homologous recombination and horizontal gene transfer in bacterial speciation. In: Gogarten MB, Gogarten JP, Olendzenski L., editors. Horizontal gene transfer: Genomes in flux. New York: Humana Press. p. 29–53.

Levin BR. 1981. Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. *Genetics* 99:1–23.

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

Lucchini S, Rowley G, Goldberg MD, Hurd D, Harrison M, Hinton JCD. 2006. H-NS mediates the silencing of laterally acquired genes in bacteria. *PLoS Pathog*. 2:e81.

Lurie-Weinberger MN, Gomez-Valero L, Merault N, Glöckner G, Buchrieser C, Gophna U. 2010. The origins of eukaryotic-like proteins in *Legionella pneumophila*. *Int J Med Microbiol*. 7:470–481.

Maiden MC, Bygraves JA, Feil E, et al. (13 co-authors). 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A*. 95:3140–3145.

Marri P, Hao W, Golding GB. 2007. The role of laterally transferred genes in adaptive evolution. *BMC Evol Biol*. 7:S8.

Martin D, Rybicki E. 2000. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* 16:562–563.

Martin DP, Williamson C, Posada D. 2005. RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* 21:260–262.

Martin DP, Beiko RG. 2010. Genetic recombination and bacterial population structure. In: Ashley RD, Falush D, Feil EJ, editors. Bacterial population genetics in infectious diseases. New Jersey: Wiley-Blackwell. p. 61–85.

Maynard Smith J, Smith NH, O'Rourke M, Spratt BG. 1993. How clonal are bacteria? *Proc Natl Acad Sci U S A*. 90:4384–4388.

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

McVean G, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160:1231–1241.

Metzker ML. 2010. Sequencing technologies—the next generation. *Nat Rev Genet*. 11:31–46.

Mintz CS. 1999. Gene transfer in *Legionella pneumophila*. *Microbes Infect*. 1:1203–1209.

Morozova I, Qu X, Shi S, Asamani G, Greenberg JE, Shuman HA, Russo JJ. 2004. Comparative sequence analysis of the *icm/ dot* genes in *Legionella*. *Plasmid* 51:127–147.

Nakamura Y, Itoh T, Matsuda H, Gojobori T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet*. 36:760–766.

Navarre WW, Porwollik S, Wang Y, McClelland M, Rosen H, Libby SJ, Fang FC. 2006. Selective silencing of foreign DNA with low GC content by the H-NS protein in *Salmonella*. *Science* 313:236–238.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299–304.

Ogata H, La Scola B, Audic S, Renesto P, Blanc G, Robert C, Fournier PE, Claverie JM, Raoult D. 2006. Genome sequence of *Rickettsia bellii* illuminates the role of amoebae in gene exchanges between intracellular pathogens. *PLoS Genet*. 2:e76.

Padidam M, Sawyer S, Fauquet CM. 1999. Possible emergence of new geminiviruses by frequent recombination. *Virology* 265:218–225.

Perez-Losada M, Browne EB, Madsen A, Wirth T, Viscidi RP, Crandall KA. 2006. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. *Infect Genet Evol*. 6:97–112.

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics*. 14:917–918.

Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci U S A*. 98:13757–13762.

Ragan MA, Beiko RG. 2009. Lateral genetic transfer: open issues. *Phil Trans Royal Soc B: Biol Sci*. 364:2241–2251.

Rogers J, Dowsett AB, Dennis PJ, Lee JV, Keevil CW. 1994. Influence of temperature and plumbing material selection on biofilm formation and growth of *Legionella pneumophila* in a model potable water system containing complex microbial flora. *Appl Environ Microbiol*. 60:1585–1592.

Rossier O, Cianciotto NP. 2005. The *Legionella pneumophila tatB* gene facilitates secretion of phospholipase C, growth under iron-limiting conditions and intracellular infection. *Infect Immun*. 73:2020–2032.

Rowbotham TJ. 1980. Preliminary report on the pathogenicity of *Legionella pneumophila* for freshwater and soil amoebae. *J Clin Pathol*. 33:1179–1183.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19:2496–2497.

Salah IB, Ghigo E, Drancourt M. 2009. Free-living amoebae, a training field for macrophage resistance of mycobacteria. *Clin Microbiol Infect*. 15:894–905.

Salminen MO, Carr JK, Burke DS, McCutchan FE. 1995. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Res Hum Retrovir*. 11:1423–1425.

Schmitz-Esser S, Tischler P, Arnold R, Montanaro J, Wagner M, Rattei T, Horn M. 2010. The genome of the amoeba symbiont '*Candidatus* Amoebophilus asiaticus' reveals common mechanisms for host cell interaction among amoeba-associated bacteria. *J Bacteriol*. 192:1045–1057.

Selander RK, McKinney RM, Whittam TS, Bibb WF, Brenner DJ, Nolte FS, Pattison PE. 1985. Genetic structure of populations of *Legionella pneumophila*. *J Bacteriol*. 163:1021–1037.

Selander RK, Caugant DA, Ochman H, Musser JM, Gilmour MN, Whittam TS. 1986. Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Appl Environ Microbiol*. 51:873–884.

Shen P, Huang HV. 1986. Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. *Genetics* 112:441–457.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol*. 16:1114–1116.

Smith JM. 1992. Analyzing the mosaic structure of genes. *J Mol Evol*. 34:126–129.

Snyder LA, Loman N, Pallen MJ, Penn CW. 2009. Next-generation sequencing–the promise and perils of charting the great microbial unknown. *Microb Ecol*. 57:1–3.

Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318:1449–1452.

Stephens JC. 1986. On the frequency of undetectable recombination events. *Genetics* 112:923–926.

Stone BJ, Kwaik YA. 1999. Natural competence for DNA transformation by *Legionella pneumophila* and its association with expression of type IV pili. *J Bacteriol*. 181:1395–1402.

Strimmer K, Rambaut A. 2002. Inferring confidence sets of possibly misspecified gene trees. *Proc R Soc Lond Ser B*. 269:137–142.

Swofford DL. 2002. PAUP*. Phylogenetic analysis using parsimony (* and other methods) [4.0beta]. Sunderland, MA: Sinauer Associates.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.

Tamames J, Abellan JJ, Pignatelli M, Camacho A, Moya A. 2010. Environmental distribution of prokaryotic taxa. *BMC Microbiol*. 10:85.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol*. 24:1596–1599.

Tettelin H, Masignani V, Cieslewicz MJ, et al. (45 co-authors). 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A*. 102:13950–13955.

Thomas CM, Nielsen KM. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol*. 3:711–721.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*. 25:4876–4882.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 22:4673–4680.

Toft C, Andersson SGE. 2010. Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat Rev Genet*. 11:465–475.

Touchon M, Hoede C, Tenaillon O, et al. (41 co-authors). 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet*. 5:e1000344.

Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J. 2005. VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics* 21:2791–2793.

Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. *ISME J*. 3:199–208.

Waldron DE, Lindsay JA. 2006. Sau1: a novel lineage-specific type I restriction-modification system that blocks horizontal gene transfer into *Staphylococcus aureus* and between *S. aureus* isolates of different lineages. *J Bacteriol*. 188:5578–5585.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*. 7:256–276.

Wu M, Eisen JA. 2008. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol*. 9:R151.

Zink SD, Pedersen L, Cianciotto NP, Abu-Kwaik Y. 2002. The Dot/Icm type IV secretion system of *Legionella pneumophila* is essential for the induction of apoptosis in human macrophages. *Infect Immun*. 70:1657–1663.