HAL

open science

Genomic Comparison Among Global Isolates of L. interrogans Serovars Copenhageni and Icterohaemorrhagiae Identified Natural Genetic Variation Caused by an Indel

Luciane A Santos, Haritha Adhikarla, Xiting Yan, Zheng Wang, Derrick E Fouts, Joseph M Vinetz, Luiz C J Alcantara, Rudy A Hartskeerl, Marga G A Goris, Mathieu Picardeau, et al.

▶ **To cite this version:**

Genomic Comparison Among Global Isolates of *L. interrogans* Serovars Copenhageni and Icterohaemorrhagiae Identified Natural Genetic Variation Caused by an Indel

Luciane A. Santos[1,2†], Haritha Adhikarla[1†], Xiting Yan[3], Zheng Wang[3], Derrick E. Fouts[4], Joseph M. Vinetz[5], Luiz C. J. Alcantara[2], Rudy A. Hartskeerl[6], Marga G. A. Goris[6], Mathieu Picardeau[7], Mitermayer G. Reis[2], Jeffrey P. Townsend[3], Hongyu Zhao[3], Albert I. Ko[1,2*] and Elsio A. Wunder Jr.[1,2*]

[1] Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT, United States, [2] Gonçalo Moniz Institute, Oswaldo Cruz Foundation, Salvador, Brazil, [3] Department of Biostatistics, Yale School of Public Health, New Haven, CT, United States, [4] J. Craig Venter Institute, Rockville, MD, United States, [5] Division of Infectious Diseases, Department of Medicine, University of California San Diego School of Medicine, La Jolla, CA, United States, [6] Royal Tropical Institute, KIT Biomedical Research, Amsterdam, Netherlands, [7] Biology of Spirochetes Unit, Institut Pasteur, Paris, France

Leptospirosis is a worldwide zoonosis, responsible for more than 1 million cases and 60,000 deaths every year. Among the 13 pathogenic species of the genus *Leptospira*, serovars belonging to *L. interrogans* serogroup Icterohaemorrhagiae are considered to be the most virulent strains, and responsible for majority of the reported severe cases. Serovars Copenhageni and Icterohaemorrhagiae are major representatives of this serogroup and despite their public health relevance, little is known regarding the genetic differences between these two serovars. In this study, we analyzed the genome sequences of 67 isolates belonging to *L. interrogans* serovars Copenhageni and Icterohaemorrhagiae to investigate the influence of spatial and temporal variations on DNA sequence diversity. Out of the 1072 SNPs identified, 276 were in non-coding regions and 796 in coding regions. Indel analyses identified 258 indels, out of which 191 were found in coding regions and 67 in non-coding regions. Our phylogenetic analyses based on SNP dataset revealed that both serovars are closely related but showed distinct spatial clustering. However, likelihood ratio test of the indel data statistically confirmed the presence of a frameshift mutation within a homopolymeric tract of *lic12008* gene (related to LPS biosynthesis) in all the *L. interrogans* serovar Icterohaemorrhagiae strains but not in the Copenhageni strains. Therefore, this internal indel identified can genetically distinguish *L. interrogans* serovar Copenhageni from serovar Icterohaemorrhagiae with high discriminatory power. To our knowledge, this is the first study to identify global sequence variations (SNPs and Indels) in *L. interrogans* serovars Copenhageni and Icterohaemorrhagiae.

Keywords: *Leptospira*, leptospirosis, Copenhageni, Icterohaemorrhagiae, whole-genome sequencing, SNPs, Indels, phylogeny

BACKGROUND

Leptospirosis is a zoonosis with worldwide distribution and considered endemic in developing countries and tropical regions. This life-threatening disease is caused by pathogenic spirochetes from the genus *Leptospira* (Bharti et al., 2003; Mcbride et al., 2005; Ko et al., 2009). Globally, leptospirosis is conservatively estimated to cause 1.03 million cases and 58,900 deaths each year (Costa et al., 2015). Transmission of leptospirosis requires continuous enzootic circulation of the pathogen among animal reservoirs. This zoonosis is maintained in nature through chronic renal infection of carrier animals, with rodents and other small mammals being the most important reservoirs. Humans and other accidental hosts get infected by direct contact with infected animals or with contaminated water or soil (Bharti et al., 2003; Ko et al., 2009).

Pathogenic *Leptospira* includes 13 species and have been classified into over 250 distinct serotypes (Reis et al., 2008; Lehmann et al., 2014; Picardeau, 2017). In humans, severe leptospirosis is frequently associated with *L. interrogans* serogroup Icterohaemorrhagiae (Bharti et al., 2003). Currently majority of the human cases caused by Icterohaemorrhagiae serogroup are attributed to serovar Copenhageni strains. Although *Rattus norvegicus* is recognized as the main reservoir for serovar Copenhageni, rodents of the *Rattus* spp. are considered as the reservoir for both serovars (Ido et al., 1917; Vinetz, 1996; Ko et al., 1999; de Faria et al., 2008). Furthermore, there is no apparent difference in terms of disease outcome between *L. interrogans* Copenhageni and Icterohaemorrhagiae serovars. In spite of being serologically distinct in serotyping methods, these serovars are still considered to be genetically similar (Kmety and Dikken, 1993; Majed et al., 2005). Therefore researchers are trying to understand the phenotypical and genotypic differences between these serovars with little success.

Genotyping methods such as multilocus variable-number tandem repeat (VNTR), pulsed-field gel electrophoresis (PFGE) and multilocus sequence typing (MLST) employed to distinguish *L. interrogans* Copenhageni and Icterohaemorrhagiae serovars lack enough discriminatory power (Barocchi et al., 2001; Bharti et al., 2003; Majed et al., 2005; Galloway and Levett, 2010; Romero et al., 2011). Multispacer sequence typing (MST) represents an evolved genotyping method for the serovar identification of *L. interrogans* serogroup Icterohaemorrhagiae (Zilber et al., 2014). Though serovar variation has been related to lipopolysaccharide (LPS) structure, specifically involving the biosynthesis locus (rfb cluster) of the O-antigen, none of the above molecular typing methods are directly related to this cluster (Faine, 1994; Llanes et al., 2016). Further comparative genomic analysis of *L. interrogans* serovars Icterohaemorrhagiae and Copenhageni may reveal the differences essential for development of novel molecular serotyping techniques (Moreno et al., 2016).

DNA polymorphisms such as single nucleotide polymorphisms (SNPs), insertions and deletions (Indels), and other larger rearrangements were successfully employed to study sequence diversity among closely related but distinct populations (Gutacker et al., 2002; Joshi et al., 2012). The use of next generation sequencing (NGS) data to detect DNA polymorphisms in the context of whole-genome analysis has been previously reported in pathogenic bacteria like *Salmonella typhi*, *Brucella* spp. and *Bacillus anthracis* (Fournier et al., 2014). Therefore, whole-genome sequencing might serve as a robust and unbiased method to resolve intraspecies relationships in *Leptospira*.

Genome-wide identification of SNPs and Indels in *L. interrogans* serovars Copenhageni and Icterohaemorrhagiae would advance our understanding of genomic diversity of these strains isolated from various geographic locations and their evolution. Identifying the genotypic differences between both serovars can improve our understanding of their evolutionary relationships in diverse epidemiological settings. This study of genomic variations will also facilitate the development of new molecular markers to differentiate pathogenic serovars and will further aid in the leptospirosis prevention strategies. Hence we performed whole-genome sequencing of 67 different strains of *L. interrogans* serovars Copenhageni and Icterohaemorrhagiae and conducted genome-wide analyses to identify serovar-specific differences.

MATERIALS AND METHODS

Leptospira Isolates

A total of 67 strains of *L. interrogans* serogroup Icterohaemorrhagiae, including 55 serovar Copenhageni isolates and 12 serovar Icterohaemorrhagiae isolates, were included in this study. These strains were isolated from different geographic locations and hosts, and the years of isolation ranged from 1915 to 2012 (**Table S1**). To include more strains for validation, eight *L. interrogans* serovar Icterohaemorrhagiae and seven serovar Copenhageni strains, whose genomes were not sequenced, were also included in this study for Sanger sequencing and/or function analyses (**Table S1**).

Culture, Genomic DNA Extraction, and Sequencing

The *Leptospira* strains were cultured in liquid Ellinghausen-McCullough-Johnson-Harris (EMJH) (Johnson and Harris, 1967) media incubated at 29°C with moderate shaking at 100 rpm. DNA was extracted from late-log phase cultures using the Maxwell 16 cell DNA purification kit along with the Maxwell DNA extraction system (Promega). The quality and concentration of DNA was measured by spectrophotometry using the NanoDrop system (Thermo Scientific, DE, USA) and by fluorometic assay using the Quanti-iT PicoGreen dsDNA assay kit (Invitrogen).

Genomes were sequenced for all 67 strains described in this study and corresponding datasets were recruited for further analyses. Genomic sequencing was done at the J. Craig Venter Institute (JCVI) using an Illumina/Solexa Genome Analyzer II, and at the Yale Center for Genome Analysis (YCGA) using the Illumina HiSeq 2000. Whole genome reads for each isolate were deposited at NCBI in the Sequence Read Archive (SRA) database (accession numbers in **Table S1**).

Serological Characterization of Isolates

The microscopic agglutination test (MAT) was used to type *Leptospira* isolates. For serogrouping, a standard 19 panel rabbit polyclonal antisera against reference serovars representing 12 different serogroups was used, as previously described (Reis et al., 2008). For serotyping, we used different monoclonal antibodies (F89 C12-6, F70 C14, F70 C24-20, and F12 C3-11 - KIT-Biomedical Research, Amsterdam) to classify isolates of serogroup Icterohaemorrhagiae as serovar Copenhageni or serovar Icterohaemorrhagiae.
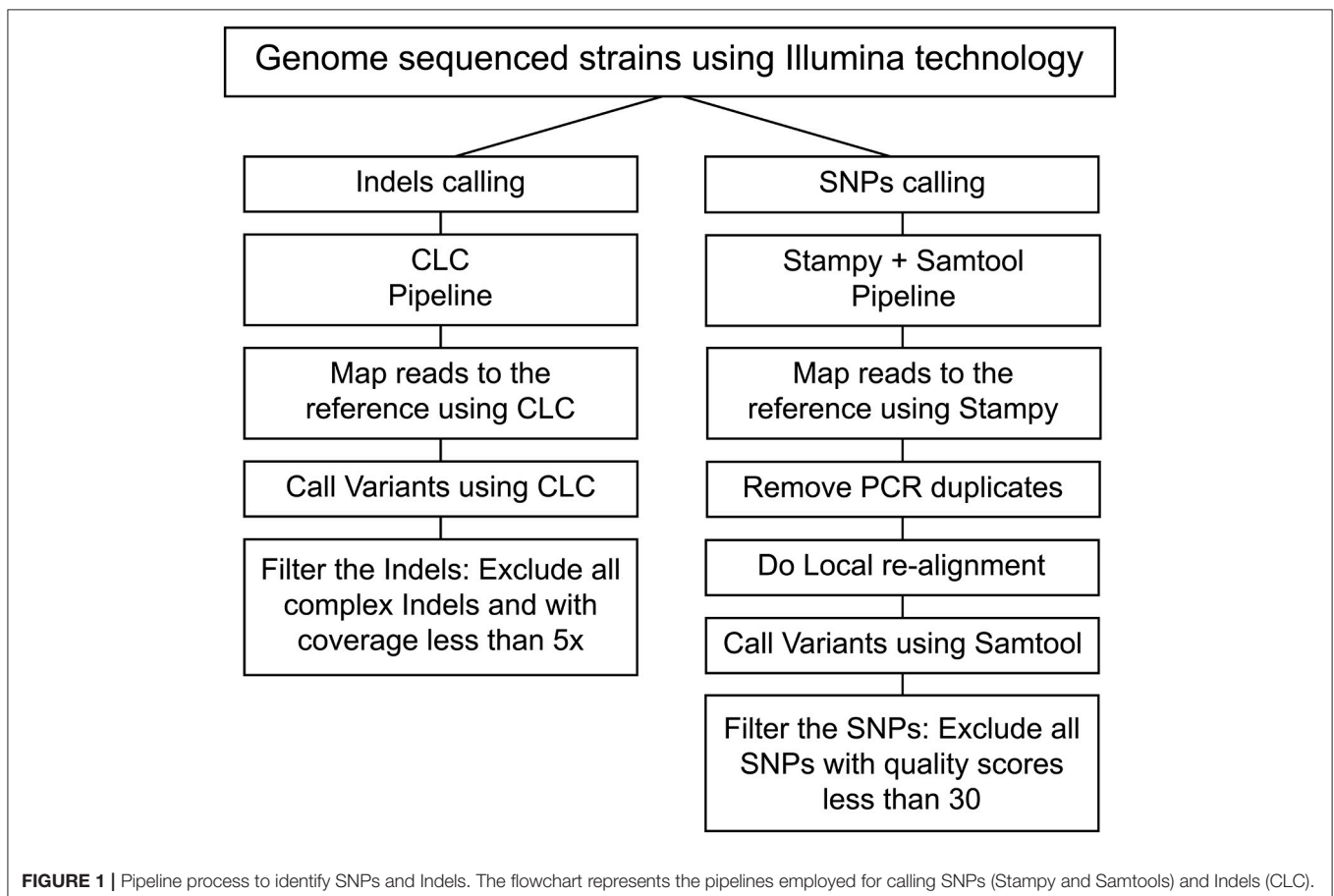
Sequence Analysis Pipeline

Reads were mapped to the *L. interrogans* serogroup Icterohaemorrhagiae serovar Copenhageni L1-130 strain reference genome (Nascimento et al., 2004) (Accession numbers: NC005823 and NC005824) using Stampy tool (Lunter and Goodson, 2011). The replicated alignment removal and local realignment were done using Samtools (Li et al., 2009). The processed mapping results were further analyzed for SNP calling using Samtools. Called SNPs were filtered to have a quality score cut-off >30. CLC Genomics workbench (CLC Genomics 6.0.4) was used to call Indels, and those with coverage lower than 5x were filtered. The Samtools pipeline exhibited better rates of consistency for SNPs calling while CLC was consistent for Indel calling (**Figure 1**, **Tables S2**, **S3**). In our reference-guided approach, we subsequently analyzed the un-mapped reads and did not detect any plasmids or misplaced contigs. To evaluate the presence of complex mutations (heterozygosis), Sanger sequencing and RFLP were performed for 3 regions. All analyses showed that the alternative allele was not present indicating a sequence error in some of the reads. Based on these analyses all the complex SNPs were excluded from this pipeline.

Phylogenetic Analyses

Phylogenetic analyses of SNPs across the whole-genome were used to infer the relationships among the 67 isolates of *L. interrogans* serogroup Icterohaemorrhagiae collected from diverse geographical locations. The reference strain *L. interrogans* serogroup Icterohaemorrhagiae serovar Copenhageni L1-130 (Nascimento et al., 2004) was included in the phylogenetic analysis. *L. interrogans* serogroup Icterohaemorrhagiae serovar Lai (Ren et al., 2003) was used as an outgroup. The length of the sequence alignment consisted of 1,731 variable sites, out of which 1072 variations were detected in this study from the sequences of serovars Copenhageni and Icterohaemorrhagiae. Additional 659 variable sites were included from the outgroup sequence serovar Lai. The Maximum Likelihood (ML) phylogeny was inferred using PAUP* (Rogers and Swofford, 1998) applying the GTR model of nucleotide substitution and gamma shape parameter. Bootstrap analysis (1,000 replicates) was used to calculate the

FIGURE 1 | Pipeline process to identify SNPs and Indels. The flowchart represents the pipelines employed for calling SNPs (Stampy and Samtools) and Indels (CLC).

statistical support of the tree branches. Bayesian trees were also inferred including the years and country of isolation in the tree construction parameters using BEAST v.1.8 package (Drummond and Rambaut, 2007). For this dataset, we only included the 57 sequences that presented complete data from year and country of isolation (**Table S1**). The models tested were the strict molecular clock with constant population size prior and the relaxed molecular clock using the constant population size, Bayesian Skyline Plot (BSP), Skyride and exponential growth priors. The parameters for each model were estimated using the Monte Carlo Markov Chain (MCMC) method (50,000,000 generations with sampling every 5,000 generations). The tested models were compared calculating the Bayes Factor (BF). Using TreeAnnotator v1.4.8, included in the BEAST package (Drummond and Rambaut, 2007), the maximum clade credibility tree was summarized from the posterior tree distribution after a 50% burn-in, for each dataset (Drummond and Rambaut, 2007). MacClade was used to identify the frequency of unambiguous changes among various countries from where isolates originated in the tree (Maddison and Maddison, 1989).

Statistical Analyses

Genotypes of *L. interrogans* serogroup Icterohaemorrhagiae serovars Icterohaemorrhagiae and Copenhageni strains were compared based on the log likelihood ratio test. For a given SNP or indel, two binomial distributions were fitted for the number of alternative alleles observed separately in Icterohaemorrhagiae and Copenhageni strains by maximizing the likelihood. The supposed maximized likelihood of Icterohaemorrhagiae and Copenhageni were denoted as L_I and L_C. Then another binomial distribution was fitted for the number of alternative allele by treating all strains from Icterohaemorrhagiae and Copenhageni as one group. If the maximized likelihood of this binomial distribution is $L_{I\&C}$, then the log likelihood ratio is calculated as $-2\log(\frac{L_{I\&C}}{(L_I \cdot L_C)})$. The *P* value was then calculated by comparing this observed log likelihood ratio to the Chi-squared distribution with 1 degree of freedom. To detect the degree of clustering of the genetic variation, a Principal Component Analysis (PCA) was performed using the SNPs. These analyses were performed using R.

Identification of Internal Indel in *L. interrogans* Serovars Icterohaemorrhagiae

Sanger sequencing confirmed the observed *lic12008* mutation using gene specific primers (forward 5′TAGGTTGGCACG AAGGTTCT3′ and reverse 5′TTTTTCCGGGAACTCCAAC3′) Sequencher 5.2 (Sequencher® Gene Codes Corporation) was used to conduct the sequence analysis, and new sequences were aligned with the reference strain to identify the presence of the mutation. A total of 13 strains belonging to *L. interrogans* serovar Icterohaemorrhagiae (WGS for 5 strains) and 9 strains (WGS for 2 strains) of serovar Copenhageni were also analyzed (**Table S1**). BLAST analyses identified homologous sequences of LIC12008 at nucleotide and protein levels. Domain analysis of LIC12008

protein was performed using NCBI CD-search and Pfam 27.0 sequence search tools (Finn et al., 2008).

Isolation of RNA and Quantitative Reverse Transcription PCR (RT-qPCR)

L. interrogans serovar Icterohaemorrhagiae strains 201000458, 201000456 and serovar Copenhageni strain L1-130 were cultured to a density of 10^8 bacteria per mL at 29°C with shaking. Cultures were harvested via centrifugation at 3,200 *g* and RNA was extracted for two biological replicates using the TRIzol (Invitrogen) method, as previously described. Ambion® TURBO DNA-free™ DNase Treatment kit was employed to remove contaminating DNA from RNA preparations. The concentration of RNA was determined using a Spectrophotometer (NanoDrop). The high capacity cDNA reverse transcription kit (Life Technologies) was employed for conversion of total RNA to single-stranded cDNA. Two primer sets were used to assess the impact of frameshift mutation on expression of LIC12008 (**Table S7**). First primer set (12008 T1) amplified a fragment of 126 bp in region encompassing nucleotides from 29 to 155 bp before the frameshift mutation. A second primer set (12008 T2) amplified a fragment of 133 bp after the mutation (from nucleotide 335–468) (**Table S7**).

The qPCR was carried out on 7500 fast real-time PCR (ABI, USA) using iQ™ SYBR® Green supermix (Biorad) according to manufacturer's instructions. The thermal cycling conditions used in the qPCR were 95°C for 3 min, followed by 40 cycles of 95°C for 5 s and 60°C for 1 min. The specificity of the SYBR green PCR signal was confirmed by melt curve analysis. In RT-qPCR experiments, *flaB* gene was used as an endogenous control and reference strain employed was *L. interrogans* serovar Copenhageni strain L1-130. Relative quantification analysis was performed using the comparative C_t method, and relative gene expression was calculated by using the $2^{-\Delta\Delta}$Ct method (Schmittgen and Livak, 2008).

In Vivo Characterization

To test whether the *lic12008* mutation impacts virulence, *in vivo* experiments were performed in the hamster model of infection. Two strains of *L. interrogans* serovar Icterohaemorrhagiae (201000458, 201000456) were tested and compared to *L. interrogans* serovar Copenhageni strain L1-130. Groups of three 21 day old male Golden Syrian hamsters were infected with doses of 10^2 and 10^8 leptospires by intraperitoneal (IP) and conjunctival routes as described previously (Wunder et al., 2016).

Animals were monitored twice daily for endpoints including signs of disease and death up to 21-days post-infection. Surviving animals 21-days after infection or moribund animals at any time presenting with difficulty moving, breathing or signs of bleeding or seizure were sacrificed by CO_2 inhalation (Wunder et al., 2016).

Ethics Statement

Animal protocols and work were approved and conducted under the guidelines of the Yale Institutional Animal Care and Use Committee (IACUC), under protocol #2017–11424. The Yale

IACUC strictly adheres to all Federal and State regulations, including the Animal Welfare Act, those specified by Public Health Service, and the US Department of Agriculture and uses the *US Government Principles for the Utilization and Care of Vertebrate Animals Used in Testing, Research, and Training* as a guide for animal studies.

RESULTS AND DISCUSSION

Pipeline for Whole-Genome Mapping

In this study, a total of 67 strains were sequenced, out of which 55 strains belonged to *L. interrogans* serovar Copenhageni and 12 strains belonged to *L. interrogans* serovar Icterohaemorrhagiae (**Table S1**). Greater number of *L. interrogans* serovar Copenhageni sequences were included in this study as this serovar is more prevalent than serovar Icterohaemorrhagiae and has greater number of isolates. All the 67 strains that were included in this study were confirmed by MAT using polyclonal sera for confirmation of Icterohaemorrhagiae serogroup (serogrouping) and monoclonal sera for differentiating Icterohaemorrhagiae and Copenhageni serovars. Serogrouping and serotyping by MAT (data not shown) confirmed the relatedness of the isolates to serogroup Icterohaemorrhagiae, and their identity to serovars Copenhageni or Icterohaemorrhagiae (**Table S1**). Stampy and Samtools were used for read mapping and SNP identification, respectively. Indels for both mapping and identification were analyzed using the CLC genome workbench. The pipeline used for identification of SNPs and indels was validated in this study by re-sequencing seven *Leptospira* isolates. Sequences from each of the seven isolates were analyzed and the selection of the best pipeline was made based on identification of the highest overlap percentage of SNPs and/or indels found in both sequences (**Figure 1**, **Tables S2**, **S3**). The information related to reads and assembly quality for these sequenced genomes can be found in **Table S4**.

L. interrogans serovar Copenhageni strain Fiocruz L1-130, sequenced using shotgun technology, was used as the reference genome (Accession numbers: NC005823 and NC005824) (Nascimento et al., 2004). Comparison of genome of *L. interrogans* serovar Copenhageni strain Fiocruz L1-130 strain re-sequenced using Illumina technology with the previously published genome sequence resulted in identification of 66 SNPs and 62 indels. Of these, 45 SNPs and 46 indels had a distribution frequency of 97% or higher in all the 132 Copenhageni strains sequenced by our group (55 included in this study and 77 from a different study) (data not shown). This higher distribution frequency of SNPs and INDELs in all the above strains might be attributed to the propagation of sequencing errors from the reference *L. interrogans* Fiocruz L1-130 sequence. In order to avoid those mutations with high frequency from the reference *L. interrogans* Fiocruz L1-130 sequence we excluded them from our analysis.

For this manuscript, we used a reference-guided approach to identify the basic differences between the genomes. However, subsequently we also employed *de novo* assembly but it did not improve the outcome. Furthermore, we checked the unmapped reads and performed a de novo assembly with those reads but this

approach also did not find any misplaced contigs or reads. In our data, the assembly covered from 97.4 to 99.99% of the genome.

Characteristics of the Mutations Detected in *L. interrogans* Serogroup Icterohaemorrhagiae Serovars Copenhageni and Icterohaemorrhagiae Strains

Whole genome sequencing enabled us to study the genome-wide variations of *L. interrogans* serovars Copenhageni and Icterohaemorrhagiae. Cumulatively, we identified 1,072 SNPs in 67 isolates, of which 276 were in non-coding region and 796 in coding regions (**Table 1** and **Table S5**). SNPs in coding regions were distributed in 594 different genes, and 115 of those had two or more SNPs in the same gene (**Table 1** and **Table S5**). Of the identified mutations in coding regions, 258 were synonymous and 538 were non-synonymous. The frequency of distribution of SNPs indicated a high proportion of non-synonymous to synonymous changes, at a ratio of 2:1. Previous comparative studies of closely related *B. anthracis*, *M. bovis,* and *Chlamydia* strains also revealed the high frequency of non-synonymous SNPs (Jordan et al., 2002; Read et al., 2002; Garnier et al., 2003). Therefore, a high proportion of non-synonymous mutations in serovars Copenhageni or Icterohaemorrhagiae could be due to their recent emergence suggesting that purifying selection might have had insufficient time to remove these slightly deleterious mutations as observed in other pathogenic bacteria (Harrison, 2013). Alternatively, this high proportion of non-synonymous mutations in both the serovars could also be an indication that some of these genes might be under positive selection.

We also identified 258 indels, of which 191 (107 deletions and 84 insertions) were found in coding regions and 67 (26 deletions and 41 insertions) in non-coding regions. Indels in coding-region were distributed in 153 different genes with 26 genes harboring two or more mutations (**Table 1** and **Table S6**). Of the 191 indels identified in coding region 183 indels caused a frameshift in the reading frames. The resulting amino acid changes and other characteristics have been represented in **Table S6**. However more functional studies are required to understand if these affected genes are related to any phenotypic consequences.

TABLE 1 | Classification and total number of identified mutations among isolates of *L. interrogans* serovars Copenhageni and Icterohaemorrhagiae.

Mutation	Classification	Coding region	Non-coding region	Total
SNPs	Synonymous	258	NA	258
	Non-synonymous	538	NA	538
	Total	796	276	1,072
Indels	Insertion	84	41	125
	Deletion	107	26	133
	Total	191	67	258

NA, Not applied.

Our results further identified genes with higher frequency of SNPs and indels (**Table 2**). However, we could not correlate this high frequency of compound mutations (multiple mutations on the same gene) with any of the phenotypic outcomes. It is possible that some of these multiple indels in the same gene could serve as compensatory indels thereby restoring the translation frame and making it a less deleterious mutation (Liu et al., 2015). Given the caveat that most of the genes in the *Leptospira* genome are annotated as hypothetical proteins or have less homology to the characterized proteins, it might be difficult to phenotypically evaluate the compensatory indels.

To our knowledge, this is the first large-scale study to identify global sequence variations among *L. interrogans* Copenhageni and Icterohaemorrhagiae serovars.

Phylogenetic Analysis

We constructed phylogenetic trees to gain insights into the spatial and temporal diversity of *L. interrogans* serovars Copenhageni and Icterohaemorrhagiae (**Figure 2** and **Figure S1**). The analysis was performed with all 67 strains isolated from different geographic locations, hosts and collected during a broad range of time (**Table S1**). A maximum likelihood (ML) tree was constructed based on the SNPs identified in each of the 67 strains, using *L. interrogans* serogroup Icterohaemorrhagiae serovar Lai (Ren et al., 2003) as an outgroup (**Figure 2**). ML tree presented a topology where both serovars Copenhageni and Icterohaemorrhagiae clustered together with statistic support (**Figure 2** cluster 6). In order to better interpret the internal branches and relation between genomes in **Figure 2** we also constructed a Bayesian phylogenetic tree (**Figure S1**) without the outgroup serovar Lai. In both the trees (**Figure 2** and **Figure S1**) we observed that Icterohaemorrhagiae and Copenhageni genomes clustered together with statistical support (represented by bootstrap for the ML tree and posterior probability for the Bayesian analyses). Therefore, our phylogenetic analyses indicated a genetic relatedness of the *L. interrogans* Copenhageni and Icterohaemorrhagiae serovars.

Despite phylogenetic relatedness, clones of strains from the Icterohaemorrhagiae serogroup seem to have evolved and adapted in different locations. Phylogenetic analysis showed seven geographic clades that corresponded spatially with the potential origins of the isolates (**Table 3**, **Figure 2**). The bootstrap support of higher than 70% was employed based on the general consensus (Hillis and Bull, 1993) and statistical support for these clades was further confirmed by principal components analysis (PCA) (data not shown). These clades include isolates from Brazil (clade 1A to 1F), Hawaii (USA, clade 2), French Polynesia (clade 3), Egypt, Denmark, Netherlands (clade 4), Japan (clade 5), Guadalupe and Slovenia (clade 6) and Colombia (clade 7) (**Table 3**). To substantiate our results for geographical clustering, Bayesian analysis was performed for 57 isolates of *L. interrogans* serovars Copenhageni and Icterohaemorrhagiae, which had specific information related to their location and year of isolation. The temporal and spatial information was incorporated into this analysis and a relaxed clock with Bayesian skyline plot was selected as the best explanatory model. The resulting tree for spatiotemporal analyses was similar to the ML tree, with distinct geographic clades from Brazil, Colombia, Guadalupe, French Polynesia, Hawaii, Egypt, and Japan (**Figure S1**).

Isolates of *L. interrogans* serovar Copenhageni from Salvador, Brazil, were significantly represented in both phylogenies and exhibited a mixed clustering pattern with isolates from other countries (**Figure 2**). The frequency of unambiguous changes between states over the maximum parsimonious trees (MPTs) was estimated in MacClade (**Figure S2**). Significant distribution of *L. interrogans* serovars was observed between European Union to Egypt (25%) and Guadeloupe to Slovenia (12.5%). Other small distribution events observed were from Brazil to Guyana and French Polynesia (1.5%). One of our important interpretations arising from the MacClade analysis was that strains from Egypt could have ancestral origins in Europe (**Figure S2**). Taken together, the MacClade analyses indicated few instances of distribution events within *L. interrogans* serovars across the globe and particularly in Salvador, Brazil.

Human connectivity with infectious rodent host reservoirs (*R. novergicus*) via transcontinental trading by mercantile ships from Europe, Africa and other continents is one plausible explanation for the introduction of various strains of *L. interrogans* serovar Copenhageni strains into Salvador. A recent study explored the global population structure of *R. novergicus* and suggested that brown rats expanded across Asia, Europe and North America through human settlements associated with Silk Road trade routes (Puckett et al., 2016). Therefore, it is possible that strains of *L. interrogans* serogroup Icterohaemorrhagiae might have been spread via rodent host reservoir. However, high level of sequence similarity among serogroup Icterohaemorrhagiae isolates indicates that specific SNPs/Indels observed were essential for their survival and prevalence in diverse habitats.

Identification of a Unique Indel in Serovar Icterohaemorrhagiae Strains

To identify the genetic mutation that could differentiate both serovars, a likelihood ratio test (LRT) of the SNP analyses was performed. LRT suggests that no SNPs distinguished these two
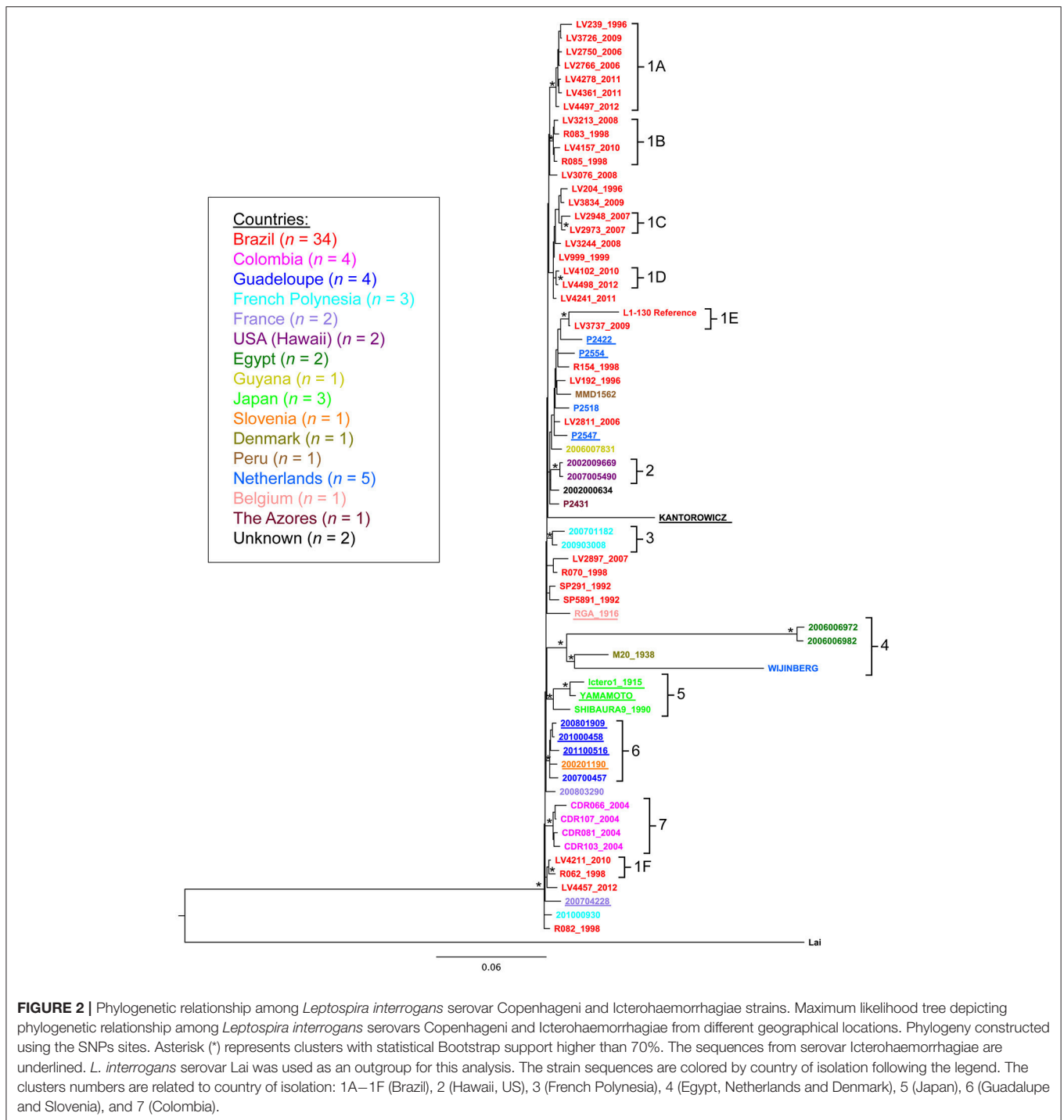
TABLE 2 | Identification of genes showing multiple SNPs and Indels in *L. interrogans* serovars Copenhageni and Icterohaemorrhagiae.

Mutation	Gene ID	Gene Product	nV[a]	Chr[b]	Gene Position[c]	Gene Size (bp)
SNPs	LIC12896	Hypothetical protein	17	1	3498112	9,423
	LIC10502	Cytoplasmic protein	8	1	590325	8,364
Indels	LIC10900	Adenylate/guanylate	6	1	1085410	1,404
	LIC10674	Hypothetical protein	4	1	819364	432
	LIC13017	Acriflavin resistance	4	1	3673377	3,312

[a] *Total number of variants in corresponding gene.*
[b] *Chromosome.*
[c] *Gene position based on the L. interrogans serovar Copenhageni strain Fiocruz L1-130 reference genome (Accession numbers: NC005823 and NC005824).*

FIGURE 2 | Phylogenetic relationship among *Leptospira interrogans* serovar Copenhageni and Icterohaemorrhagiae strains. Maximum likelihood tree depicting phylogenetic relationship among *Leptospira interrogans* serovars Copenhageni and Icterohaemorrhagiae from different geographical locations. Phylogeny constructed using the SNPs sites. Asterisk (*) represents clusters with statistical Bootstrap support higher than 70%. The sequences from serovar Icterohaemorrhagiae are underlined. *L. interrogans* serovar Lai was used as an outgroup for this analysis. The strain sequences are colored by country of isolation following the legend. The clusters numbers are related to country of isolation: 1A−1F (Brazil), 2 (Hawaii), 3 (French Polynesia), 4 (Egypt, Netherlands and Denmark), 5 (Japan), 6 (Guadeloupe and Slovenia), and 7 (Colombia).

serovars of **Icterohaemorrhagiae and Copenhageni**. However, LRT of the indel data was statistically consistent ($p = 0.039$) with the presence of a single base insertion of a thymine nucleotide within a polyT tract (9 bp long) in gene *lic12008* (**Figures 3A,B**) in all the *L. interrogans* serovar Icterohaemorrhagiae isolates. As an orthogonal approach, we ran parsimony informative test using our SNPs and it yielded 30% of parsimony-informative sites (data not shown). However, except for the one indel in *lic12008*

locus we did not identify any statistically significant SNP's in our analyses.

The gene *lic12008* (726 bp) is located at nucleotide positions 2416786 to 2417511 of Chromosome 1 (reading frame of −2) in *L. interrogans* serovar Copenhageni Fiocruz L1-130 genome and was not found to be part of any operon. Sanger sequencing confirmed the presence of this insertion at the 5′ end of *lic12008* gene in all *L. interrogans* Icterohaemorrhagiae strains but its

absence in Copenhageni strains (data not shown). The identified insertion in *lic12008* gene resulted in a frameshift mutation at N terminal region (45th amino acid) of the corresponding protein (**Figure 3C**). This highly conserved change at amino acid level

within the serovar Icterohaemorrhagiae might have evolutionary implications leading to its divergence from the Copenhageni serovar. A significant overrepresentation of frameshifts at N terminal region of proteins was also previously observed in pathogenic *P. aeruginosa* (Harrison, 2013).

Our results indicated that an internal indel in a homopolymeric tract region explains one key aspect of sequence diversity between closely related *L. interrogans* serovars Copenhageni and Icterohaemorrhagiae strains. This observation agrees with previous results where homopolymeric regions were known to be subject to indel mutations as major sources of sequence diversity in various animal, plants, insects and bacteria (Moran et al., 2009). However, indels in homopolymers more often have deleterious effects, and there is an increasing evidence for strong purifying selection against frameshift mutations in coding regions (Williams and Wernegreen, 2013). Therefore, the conserved frameshifting indel in *lic12008* observed in *L. interrogans* serovar Icterohaemorrhagiae isolates is interesting from an evolutionary perspective.

In this study, the power of variant discovery was enhanced by jointly analyzing all the samples (Li and Durbin, 2010; DePristo et al., 2011; Nielsen et al., 2011). Subsequently genotype likelihood-based LRT was assessed to compute a statistically robust association that was able to classify *L. interrogans* serovars Copenhageni and Icterohaemorrhagiae isolates. This test combined with the number of sequences analyzed, demonstrated that both serovars are highly related beyond serovar classification. Thus, the majority of SNPs and indels

TABLE 3 | Correlation of isolates based on the geographical clades from phylogenetic analysis.

Clade	Isolate[a]	Country[b]
1A	LV239, LV3726, LV2750, LV2766, LV4278, LV4361, LV4497	Brazil
1B	LV3213, R83, LV4157, R85	Brazil
1C	LV2948, LV2973	Brazil
1D	LV4102, LV4498	Brazil
1E	L1-130, LV3737	Brazil
1F	LV4211, R62	Brazil
2	2007005490, 2002009669	Hawaii
3	200903008, 200701182	French Polynesia
4	2006006972, 200600682, M20, Wijnberg	Egypt, Denmark, Netherlands
5	Ictero1, Yamamoto, Shibaura	Japan
6	200801909, 201000458, 201100516, 200700457, 200201190	Guadeloupe, Slovenia
7	CIDEIM R006, CIDEIM R107, CIDEIM R081, CIDEIM R103	Colombia

[a,b]*Details of all the isolates and countries of origin are listed in* **Table S1**.

FIGURE 3 | Genomic organization and domain architecture of gene *lic12008*. **(A)** Organization of *lic12008* loci in *L. interrogans* serovars. **(B)** Domain architecture of LIC12008 protein. **(C)** Details of amino acid fragment (1–55aa) showing the sequence differences in *L. interrogans* Copenhageni (I) and *L. interrogans* Icterohaemorrhagiae (II) serovars.
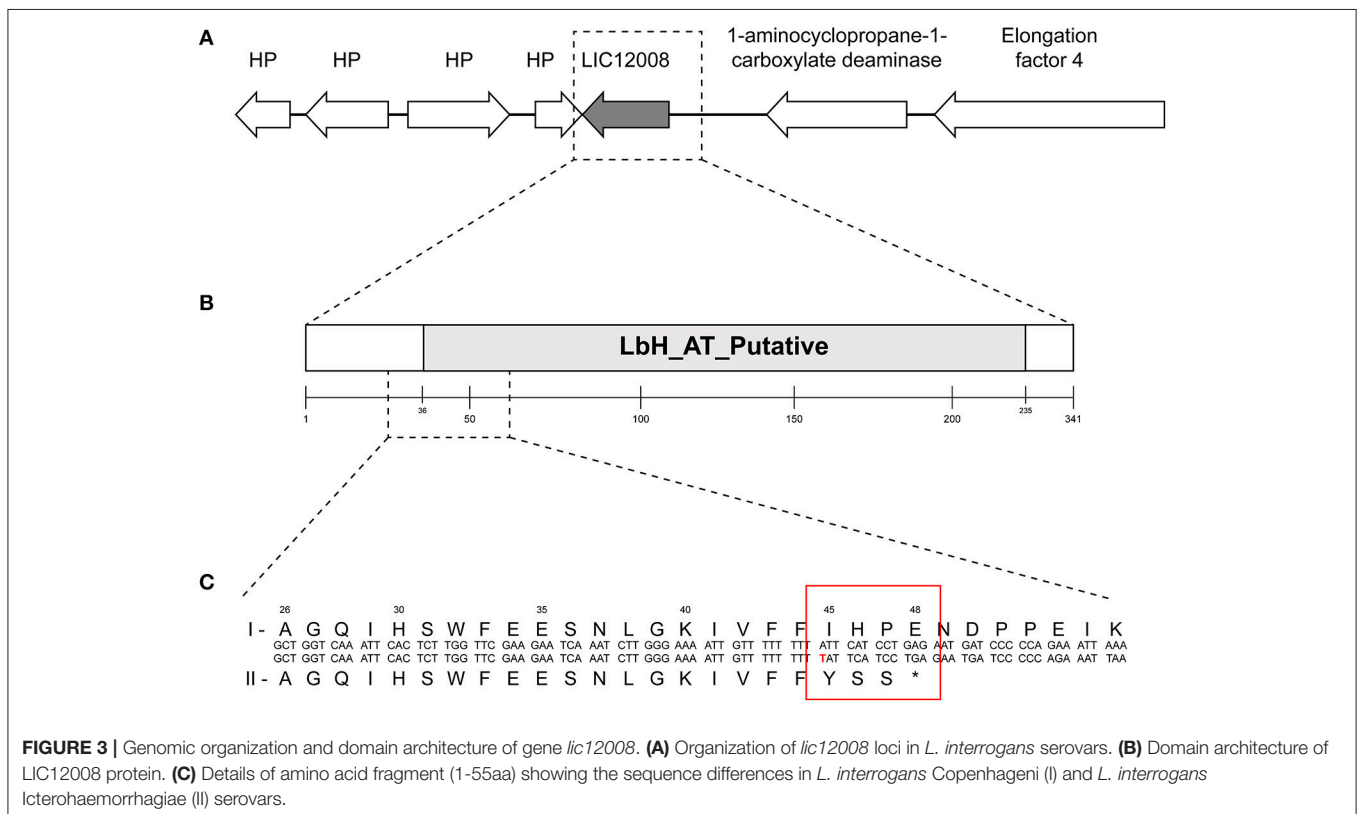
detected among the 67 strains represented the natural diversity of the sequences. In a recent study, MST was employed to differentiate Copenhageni and Icterohaemorrhagiae serovars. However, MST could not define unique profiles for few strains of *L. interrogans* serovars Copenhageni (M20 and Wijnberg strains) and Icterohaemorrhagiae (RGA and Verdun strains) (Zilber et al., 2014). The identified Indel in *lic12008* gene however had a high discriminatory power to distinguish between Copenhageni and Icterohaemorrhagiae serovars without any exceptions.

Analysis of Genomic Region With Predicted Indels

Basic local alignment search tool (BLAST) was used to study the distribution of *lic12008* gene in other *Leptospira* species. Nucleotide analyses demonstrated that this gene was absent in non-pathogenic and intermediate *Leptospira* species and present only in four pathogenic species: *L. interrogans, L. kirschneri, L. noguchii,* and *L. santarosai.* BLAST with the LIC12008 amino acid sequence as query (cutoff: >30% identity) identified UDP-3-O-(3-hydroxymyristoyl) glucosamine N-acyltransferase, sugar O-acyltransferase and sialic acid O-acetyltransferase of NeuD family as closest homologs. In other bacteria, UDP-3-O-(3-hydroxymyristoyl) glucosamine N-acyltransferase is involved in the biosynthesis of lipid A, a phosphorylated glycolipid that anchors the lipopolysaccharide to the outer membrane of the cell (Bartling and Raetz, 2009). Previous studies demonstrated the physiological relevance of *lpxD1* gene (LIC13046) encoding a UDP-3-O-(3-hydroxymyristoyl) glucosamine N-acyltransferase in *L. interrogans* serovar Manilae strain L495 (Eshghi et al., 2015).

We identified a paralog of UDP-3-O-(3-hydroxymyristoyl) glucosamine N-acyltransferase, annotated as an acetyl transferase gene (LIC12184), in *Leptospira interrogans* Copenhageni Fiocruz L1-130, employing LIC12008 as query sequence (with identity of 74% and similarity of 89%). Further genomic characterization of this paralog might unveil the evolutionary mechanisms underlying the development of this gene family.

Domain analysis of the hypothetical protein encoded by *lic12008* showed that the region spanning amino acids 36 to 233 (out of 242aa) is comprised of the putative Acyltransferase (AT) and Left-handed parallel beta-Helix (LbH) domain (E-value: 6.33e-48 and domain accession cd03360) (**Figure 3B**). LIC12008 belongs to LbetaH superfamily proteins composed mainly of acyltransferases [33]. Three imperfect tandem repeats of a hexapeptide repeat motif (X-[STAV]-X-[LIV]-[GAED]-X) were also identified in LIC12008. Thus, the presence of LbHAT domain in LIC12008 (**Figure 3B**) allows us to speculate that this protein has a role in LPS biosynthesis. Previous studies implicated the role of horizontal transfer of genes located within the *rfb* cluster for serological relatedness of genetically similar serovars (de la Peña-Moctezuma et al., 1999; Nalam et al., 2010). In contrast our study identified an indel in *lic12008* gene unrelated to *rfb* cluster but with a presumable role in LPS biosynthesis. Further gene neighborhood analysis of *lic12008* did not show any evidence of horizontal gene transfer events (data not shown).

In clinical strains of *Burkholderia pseudomallei*, accumulation of four indels affecting lipopolysaccharide (LPS) biosynthesis was identified as a mechanism used by this pathogen to evade the immune response (Price et al., 2013). In this context, the presence of a frameshifting indel in *lic12008,* a LPS biosynthesis related gene seems important. Since protective antibody responses for *Leptospira* are against LPS, altered expression of LPS might have an impact on the host immune response, which might provide a plausible explanation for the serological differences found between Copenhageni and Icterohaemorrhagiae strains.

Functional Analysis

Gene expression analysis was performed to identify differences in expression of the *lic12008* transcript in *L. interrogans* serovars Copenhageni and Icterohaemorrhagiae. Two pairs of primers were used to study the impact of mutation on *lic12008* expression. The first primer (12008 T1) encompassed the region of the mutation while the second primer was outside the mutation (12008 T2) (**Figure 3C**, **Table S7**). Transcripts from both regions of *L. interrogans* serovar Icterohaemorrhagiae were downregulated compared to *L. interrogans* Copenhageni Fiocruz L1-130 (**Figure S3**).

To determine the possible phenotypic consequences of the *lic12008* frameshifting indel in *L. interrogans* serovar Icterohaemorrhagiae, we performed an *in vivo* experiment in hamster model of infection. Representative strains from *L. interrogans* serovar Icterohaemorrhagiae were used to infect two groups of three hamsters, via intraperitoneal and conjunctival routes respectively. Both strains of serovar Icterohaemorrhagiae were virulent in the hamster model of infection (**Table S8**) similar to Copenhageni strains which were previously shown to be virulent (Silva et al., 2008). This indicates that *lic12008* frameshifting indel might not be associated with any negative effects on virulence. Alternatively, paralogous gene of *lic12008* might compensate for any functional aberrations caused by this frameshifting indel.

In summary, we analyzed the genome-wide SNPs and indels among the *L. interrogans* serovars Copenhageni and Icterohaemorrhagiae isolates. Analyses of these genome-wide variations revealed that both serovars are genetically similar. Phylogenetic analyses also indicated that *L. interrogans* serovar Copenhageni and Icterohaemorrhagiae strains are highly conserved along time with a distinct geographical clustering. However, our results indicated that an internal indel is the major source of variation in *L. interrogans* serovar Icterohaemorrhagiae. This identified internal indel could provide a plausible explanation for the unexplained antigenic differences between *Leptospira interrogans* serovars Icterohaemorrhagiae and Copenhageni. Previously, indel studies in other species have led to the identification of powerful taxon diagnostics and phylogenetic markers (Baldauf and Palmer, 1993; de Jong et al., 2003; Ajawatanawong and Baldauf, 2013). In this context, the internal indel identified in this study could be possibly validated as a diagnostic marker to differentiate *L. interrogans* Copenhageni and Icterohaemorrhagiae isolates.

AUTHOR CONTRIBUTIONS

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcimb.2018.00193/full#supplementary-material

Figure S1 | Bayesian maximum clade credibility phylogenetic tree. The tree was constructed using the relaxed molecular clock model and Bayesian Skyline Plot population size model. The branch support is given by posterior probability >0.88 and is indicated with an asterisk (*). The tree branches are colored by country.

Figure S2 | MacClade analysis. Identification of the migration relation between strains from different geographic locations. The selected regions were Brazil (BR), United States (US), Slovenia (SL), Egypt (EG) Guyana (GY), Guadalupe (GP), European Union (EU), French Polynesia (FP), Colombia (CO), and Japan (JP).

Figure S3 | Gene expression analysis of *lic12008* in *L. interrogans* serovars. Real-time PCR analysis of *lic12008* transcript in *L. interrogans* serovar Icterohaemorrhagiae strains (201000456 and 201000458) by employing primer pairs (12008 T1 and T2). Fold change was calculated by using the $2^{-\Delta\Delta Ct}$ method. Data was normalized by the amount of *flaB* transcript expressed relative to the corresponding value for *L. interrogans* serovar Copenhageni L1-130 and is expressed as mean ± SD.

Table S1 | General information list of all the *Leptospira interrogans* strains of the serogroup Icterohaemorrhagiae used in this study.

Table S2 | Validation rate for the pipeline call for SNPs.

Table S3 | Validation rate for the pipeline call for Indel.

Table S4 | Sequencing and mapping quality.

Table S5 | List of polymorphic nucleotide sites identified in *L. interrogans* serovars from this study.

Table S6 | List of Insertion and deletions (Indels) identified in *L. interrogans* serovars from this study.

Table S7 | Primers employed for qRT-PCR.

Table S8 | *In vivo* testing of virulence for *L. interrogans* serovar Icterohaemorrhagiae strains.

REFERENCES

Ajawatanawong, P., and Baldauf, S. L. (2013). Evolution of protein indels in plants, animals and fungi. *BMC Evol. Biol.* 13:140. doi: 10.1186/1471-2148-13-140

Baldauf, S. L., and Palmer, J. D. (1993). Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc. Natl. Acad. Sci. U.S.A.* 90, 11558–11562. doi: 10.1073/pnas.90.24.11558

Barocchi, M. A., Ko, A. I., Ferrer, S. R., Faria, M. T., Reis, M. G., and Riley, L. W. (2001). Identification of new repetitive element in leptospira interrogans serovar copenhageni and its application to PCR-based differentiation of leptospira serogroups. *J. Clin. Microbiol.* 39, 191–195. doi: 10.1128/JCM.39.1.191-195.2001

Bartling, C. M., and Raetz, C. R. H. (2009). Crystal structure and acyl chain selectivity of *Escherichia coli* LpxD, the N-acyltransferase of lipid a biosynthesis. *Biochemistry* 48, 8672–8683. doi: 10.1021/bi901025v

Bharti, A. R., Nally, J. E., Ricaldi, J. N., Matthias, M. A., Diaz, M. M., Lovett, M. A., et al. (2003). Leptospirosis: a zoonotic disease of global importance. *Lancet Infect. Dis.* 3, 757–771. doi: 10.1016/S1473-3099(03)00830-2

Costa, F., Hagan, J. E., Calcagno, J., Kane, M., Torgerson, P., Martinez-Silveira, M. S., et al. (2015). Global morbidity and mortality of leptospirosis: a systematic review. *PLoS Negl. Trop. Dis.* 9:e0003898. doi: 10.1371/journal.pntd.0003898

de Faria, M. T., de Calderwood, M. S., Athanazio, D. A., McBride, A. J. A., Hartskeerl, R. A., Pereira, M. M., et al. (2008). Carriage of Leptospira interrogans among domestic rats from an urban setting highly endemic for leptospirosis in Brazil. *Acta Trop.* 108, 1–5. doi: 10.1016/j.actatropica.2008.07.005

de la Peña-Moctezuma, A., Bulach, D. M., Kalambaheti, T., and Adler, B. (1999). Comparative analysis of the LPS biosynthetic loci of the genetic subtypes of serovar Hardjo: leptospira interrogans subtype Hardjoprajitno and Leptospira borgpetersenii subtype Hardjobovis. *FEMS Microbiol. Lett.* 177, 319–326. doi: 10.1111/j.1574-6968.1999.tb13749.x

de Jong W. W., Van, M. D., Poux, J., Kappé, G., Van, T. R., and Madsen, O. (2003). Indels in protein-coding sequences of *Euarchontoglires* constrain the rooting of the eutherian tree. *Mol. Phylogenet. Evol.* 28, 328–340. doi: 10.1016/S1055-7903(03)00116-7

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Drummond, A. J., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7:214. doi: 10.1186/1471-2148-7-214

Eshghi, A., Henderson, J., Trent, M. S., and Picardeau, M. (2015). Leptospira interrogans lpxD homologue is required for thermal acclimatization and virulence. *Infect. Immun.* 83, 4314–4321. doi: 10.1128/IAI.00897-15

Faine, S. (1994). *Leptospira and leptospirosis. Leptospira Leptospirosis.* Available online at: https://www.cabdirect.org/cabdirect/abstract/19952009809 (Accessed Nov 23, 2017). doi: 10.1016/j.vetmic.2009.03.012

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., et al. (2008). The Pfam protein families database. *Nucleic Acids Res.* 36, D281–D288. doi: 10.1093/nar/gkh121

Fournier, P.-E., Dubourg, G., and Raoult, D. (2014). Clinical detection and characterization of bacterial pathogens in the genomics era. *Genome Med.* 6:114. doi: 10.1186/s13073-014-0114-2

Galloway, R. L., and Levett, P. N. (2010). Application and validation of PFGE for serovar identification of leptospira clinical isolates. *PLoS Negl. Trop. Dis.* 4:e824. doi: 10.1371/journal.pntd.0000824

Garnier, T., Eiglmeier, K., Camus, J.-C., Medina, N., Mansoor, H., Pryor, M., et al. (2003). The complete genome sequence of *Mycobacterium bovis*. *Proc. Natl. Acad. Sci. U.S.A.* 100, 7877–7882. doi: 10.1073/pnas.1130426100

Gutacker, M. M., Smoot, J. C., Migliaccio, C. A. L., Ricklefs, S. M., Hua, S., Cousins, D. V., et al. (2002). Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics* 162, 1533–1543.

Harrison, F. (2013). Bacterial cooperation in the wild and in the clinic: Are pathogen social behaviours relevant outside the laboratory? *Bioessays* 35, 108–112. doi: 10.1002/bies.201200154

Hillis, D. M., and Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* 42, 182–192. doi: 10.1093/sysbio/42.2.182

Ido, Y., Hoki, R., Ito, H., and Wani, H. (1917). The rat as a carrier of spirochæta icterohæmorrhagiæ, the causative agent of weil's disease (spirochætosis icterohæmorrhagica). *J. Exp. Med.* 26, 341–353. doi: 10.1084/jem.26.3.341

Johnson, R. C., and Harris, V. G. (1967). Differentiation of pathogenic and saprophytic leptospires. I. Growth at low temperatures. *J. Bacteriol.* 94, 27–31.

Jordan, I. K., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2002). Microevolutionary genomics of bacteria. *Theor. Popul. Biol.* 61, 435–447. doi: 10.1006/tpbi.2002.1588

Joshi, D., Harris, N. B., Waters, R., Thacker, T., Mathema, B., Krieswirth, B., et al. (2012). Single nucleotide polymorphisms in the *Mycobacterium bovis* genome resolve phylogenetic relationships. *J. Clin. Microbiol.* 50, 3853–3861. doi: 10.1128/JCM.01499-12

Kmety, E., and Dikken, H. (1993). *Classification of the Species Leptospira Interrogans and History of its Serovars*. University Press Groningen.

Ko, A. I., Reis, M. G., Dourado, C. M. R., Johnson, W. D., and Riley, L. W. (1999). Urban epidemic of severe leptospirosis in Brazil. *Lancet* 354, 820–825. doi: 10.1016/S0140-6736(99)80012-9

Ko, A. I., Goarant, C., and Picardeau, M. (2009). Leptospira: the dawn of the molecular genetics era for an emerging zoonotic pathogen. *Nat. Rev. Microbiol.* 7, 736–747. doi: 10.1038/nrmicro2208

Lehmann, J. S., Matthias, M. A., Vinetz, J. M., and Fouts, D. E. (2014). Leptospiral Pathogenomics. *Pathogens* 3, 280–308. doi: 10.3390/pathogens3020280

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Liu, M., Watson, L. T., and Zhang, L. (2015). Predicting the combined effect of multiple genetic variants. *Hum. Genomics* 9:18. doi: 10.1186/s40246-015-0040-4

Llanes, A., Restrepo, C. M., and Rajeev, S. (2016). Whole genome sequencing allows better understanding of the evolutionary history of Leptospira interrogans serovar Hardjo. *PLoS ONE* 11:e015938710. doi: 10.1371/journal.pone.0159387

Lunter, G., and Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* 21, 936–939. doi: 10.1101/gr.111120.110

Maddison, W. P., and Maddison, D. R. (1989). Interactive analysis of phylogeny and character evolution using the computer program macclade. *Folia Primatol.* 53, 190–202. doi: 10.1159/000156416

Majed, Z., Bellenger, E., Postic, D., Pourcel, C., Baranton, G., and Picardeau, M. (2005). Identification of variable-number tandem-repeat loci in leptospira interrogans sensu stricto. *J. Clin. Microbiol.* 43, 539–545. doi: 10.1128/JCM.43.2.539-545.2005

Mcbride, A. J. A., Athanazio, D. A., Reis, M. G., and Ko, A. I. (2005). Leptospirosis. *Curr. Opin. Infect. Dis.* 18, 376–386. doi: 10.1097/01.qco.0000178824.05715.2c

Moran, N. A., McLaughlin, H. J., and Sorek, R. (2009). The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* 323, 379–382. doi: 10.1126/science.1167140

Moreno, L. Z., Miraglia, F., Lilenbaum, W., Neto, J. S., Freitas, J. C., Morais, Z. M., et al. (2016). Profiling of Leptospira interrogans, *L. santarosai*, *L. meyeri* and *L. borgpetersenii* by SE-AFLP, PFGE and susceptibility testing—a continuous

attempt at species and serovar differentiation. *Emerg. Microbes Infect.* 5:e17. doi: 10.1038/emi.2016.16

Nalam, K., Ahmed, A., Devi, S. M., Francalacci, P., and Baig, M., et al (2010). Genetic affinities within a large global collection of pathogenic Leptospira: implications for strain identification and molecular epidemiology. *PLoS ONE* 5:e12637. doi: 10.1371/journal.pone.0012637

Nascimento, A. L. T. O., Ko, A. I., Martins, E. A. L., Monteiro-Vitorello, C. B., Ho, P. L., Haake, D. A., et al. (2004). Comparative genomics of two leptospira interrogans serovars reveals novel insights into physiology and pathogenesis. *J. Bacteriol.* 186, 2164–2172. doi: 10.1128/JB.186.7.2164-2172.2004

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12, 443–451. doi: 10.1038/nrg2986

Picardeau, M. (2017). Virulence of the zoonotic agent of leptospirosis: still terra incognita? *Nat. Rev. Microbiol.* 15, 297. doi: 10.1038/nrmicro.2017.5

Price, E. P., Sarovich, D. S., Mayo, M., Tuanyok, A., Drees, K. P., Kaestli, M., et al. (2013). Within-host evolution of *Burkholderia pseudomallei* over a twelve-year chronic carriage infection. *mBio* 4:e00388–13. doi: 10.1128/mBio.00388-13

Puckett, E. E., Park, J., Combs, M., Blum, M. J., Bryant, J. E., Caccone, A., et al. (2016). Global population divergence and admixture of the brown rat (*Rattus norvegicus*). *Proc. R. Soc. B* 283:20161762. doi: 10.1098/rspb.2016.1762

Read, T. D., Salzberg, S. L., Pop, M., Shumway, M., Umayam, L., Jiang, L., et al. (2002). Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science* 296, 2028–2033. doi: 10.1126/science.1071837

Reis, R. B., Ribeiro, G. S., Felzemburgh, R. D. M., Santana, F. S., Mohr, S., Melendez, A. X. T. O., et al. (2008). Impact of environment and social gradient on Leptospira infection in urban slums. *PLoS Negl. Trop. Dis.* 2:e228. doi: 10.1371/journal.pntd.0000228

Ren, S.-X., Fu, G., Jiang, X.-G., Zeng, R., Miao, Y.-G., Xu, H., et al. (2003). Unique physiological and pathogenic features of Leptospira interrogans revealed by whole-genome sequencing. *Nature* 422:888. doi: 10.1038/nature01597

Rogers, J. S., and Swofford, D. L. (1998). A fast method for approximating maximum likelihoods of phylogenetic trees from nucleotide sequences. *Syst. Biol.* 47, 77–89. doi: 10.1080/106351598261049

Romero, E. C., Blanco, R. M., and Galloway, R. L. (2011). Analysis of multilocus sequence typing for identification of leptospira isolates in Brazil *J. Clin. Microbiol.* 49, 3940–3942. doi: 10.1128/JCM.01119-11

Schmittgen, T. D., and Livak, K. J. (2008). Analyzing real-time PCR data by the comparative CT method. *Nat. Protoc.* 3, 1101–1108. doi: 10.1038/nprot.2008.73

Silva, É. F., Santos, C. S., Athanazio, D. A., Seyffert, N., Seixas, F. K., Cerqueira, G. M., et al. (2008). Characterization of virulence of leptospira isolates in a hamster model. *Vaccine* 26, 3892–3896. doi: 10.1016/j.vaccine.2008.04.085

Vinetz, J. M. (1996). Sporadic urban leptospirosis. *Ann. Intern. Med.* 125:794. doi: 10.7326/0003-4819-125-10-199611150-00002

Williams, L. E., and Wernegreen, J. J. (2013). Sequence context of indel mutations and their effect on protein evolution in a bacterial endosymbiont. *Genome Biol. Evol.* 5, 599–605. doi: 10.1093/gbe/evt033

Wunder, E. A., Figueira, C. P., Santos, G. R., Lourdault, K., Matthias, M. A., Vinetz, J. M., et al. (2016). Real-time PCR reveals rapid dissemination of leptospira interrogans after intraperitoneal and conjunctival inoculation of hamsters. *Infect. Immun.* 84, 2105–2115. doi: 10.1128/IAI.00094-16

Zilber, A.-L., Picardeau, M., Ayral, F., Artois, M., Demont, P., Kodjo, A., et al. (2014). High-resolution typing of leptospira interrogans strains by multispacer sequence typing. *J. Clin. Microbiol.* 52, 564–571. doi: 10.1128/JCM.02482-13