Gene 511 (2012) 364–370

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Short Communication

Sequence of *Leptospira santarosai* serovar Shermani genome and prediction of virulence-associated genes

Li-Fang Chou [a], Yu-Tin Chen [b], Chia-Wei Lu [b], Yi-Ching Ko [a], Chuan-Yi Tang [c], Ming-Jeng Pan [d], Ya-Chung Tian [a], Cheng-Hsun Chiu [e], Cheng-Chieh Hung [a], Chih-Wei Yang [a,*]

[a] *Kidney Research Center, Chang Gung Memorial Hospital, Linkou, and College of Medicine, Chang Gung University, Taiwan*
[b] *Department of Computer Science, National Tsing Hua University, HsinChu, Taiwan*
[c] *Department of Computer Science and Information Engineering, Providence University, Taichung, Taiwan*
[d] *Institute of Life Sciences, Central Taiwan University of Science and Technology, Taichung, Taiwan*
[e] *Molecular Infectious Diseases Research Center, Chang Gung Memorial Hospital, Linkou, Taiwan*

ARTICLE INFO

Article history:
Accepted 12 September 2012
Available online 3 October 2012

Keywords:
Leptospirosis
Leptospira
Genome
Virulence
Lipoproteins

ABSTRACT

Leptospirosis, a widespread zoonosis, is a re-emerging infectious disease caused by pathogenic *Leptospira* species. In Taiwan, *Leptospira santarosai* serovar Shermani is the most frequently isolated serovar, causing both renal and systemic infections. This study aimed to generate a *L. santarosai* serovar Shermani genome sequence and categorize its hypothetical genes, particularly those associated with virulence. The genome sequence consists of 3,936,333 nucleotides and 4033 predicted genes. Additionally, 2244 coding sequences could be placed into clusters of orthologous groups and the number of genes involving cell wall/membrane/envelope biogenesis and defense mechanisms was higher than that of other *Leptospira* spp. Comparative genetic analysis based on BLASTX data revealed that about 73% and 68.8% of all coding sequences have matches to pathogenic *L. interrogans* and *L. borgpetersenii*, respectively, and about 57.6% to saprophyte *L. biflexa*. Among the hypothetical proteins, 421 have a transmembrane region, 172 have a signal peptide and 17 possess a lipoprotein signature. According to PFAM prediction, 32 hypothetical proteins have properties of toxins and surface proteins mediated bacterial attachment, suggesting they may have roles associated with virulence. The availability of the genome sequence of *L. santarosai* serovar Shermani and the bioinformatics re-annotation of leptospiral hypothetical proteins will facilitate further functional genomic studies to elucidate the pathogenesis of leptospirosis and develop leptospiral vaccines.

1. Introduction

Leptospirosis is a re-emerging infectious disease as a consequence of global warming and flooding, which frequently occurs in warm and humid countries (Adler and de la Pena Moctezuma, 2010; Bharti et al., 2003). In Taiwan, 291 human leptospirosis cases were confirmed out of 7733 suspected cases from 2001 to 2006. However, it is still regarded as an under-reported infectious disease in Taiwan (Lin et al., 2008). The infection is caused by pathogenic *Leptospira* spp., which belong to the spirochete family. *Leptospira* spp. are highly motile bacteria capable of penetrating the human body through abraded mucous surfaces, allowing systemic dissemination (Adler and de la Pena Moctezuma, 2010). The genus *Leptospira* comprises a heterogeneous group of pathogenic and saprophytic species and consists of more than 250 serovars (Cerqueira and Picardeau, 2009).

Pathogenic *Leptospira* spp. comprises *L. interrogans*, *L. borgpetersenii* and *L. santarosai* that are isolated from clinical samples of patients with leptospirosis. *L. biflexa* belongs to saprophyte *Leptospira* spp. isolated from the environment. Normally, *Leptospira* spp., but not *L. borgpetersenii*, can be isolated from fresh water of lakes whereas *L. borgpetersenii* does not survive outside of the host (Evangelista and Coburn, 2010). Genomic studies indicate that genes necessary for survival in low-nutrient environments were lost in *L. borgpetersenii*; therefore, reflecting in its limitation in transmission through direct host to host contact (Bulach et al., 2006). Pathogenic *Leptospira* spp. infect a wide range of hosts, causing a wide range of clinical manifestations from flu-like syndrome to illnesses characterized by a multiorgan complications. In addition, vascular injury and endothelial lesions were observed in all affected organs. Kidney is the preferential site of infection, and the bacteria can accumulate in the

Abbreviations: MAT, microscope agglutination test; LPS, lipopolysaccharide; Lig proteins, leptospira immunoglobulin-like proteins; EMJH, Ellinghausen–McCullough–Johnson–Harris; MAQ software, Mapping and Assembly with Qualities software; iCORN, Iterative Correction of Reference Nucleotides; PGAAP, Prokaryotic Genomes Automatic Annotation Pipeline; MEGA, Molecular Evolutionary Genetics Analysis; ORF, open reading frames; COG, clusters of orthologous groups; MLST, multilocus sequence typing; STs, sequence types; LRR, leucine-rich repeat; MCPs, methyl-accepting chemotaxis proteins; CDS, coding sequences.

* Corresponding author at: Kidney Research Center, Chang Gung Memorial Hospital. 5-7, Fusing St., Gueishan Township, Taoyuan County 333, Taiwan. Tel.: +886 3 3281200x2506; fax: +886 3 3282173.
E-mail address: cwyang@ms1.hinet.net (C.-W. Yang).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.09.074

interstitium and renal proximal tubule epithelial cells. In Taiwan, acute tubulointerstitial nephritis is primarily attributed to leptospiral infection and the major etiological agent is *Leptospira santarosai* serovar Sherman as confirmed by microscope agglutination test (MAT) (Yang, 2007; Yang et al., 1997, 2001).

At present, the molecular mechanisms of the pathogenesis of leptospirosis remain to be elucidated. Several candidate virulence factors have been identified including LPS, hemolysins, outer membrane proteins and other surface proteins as well as adhesion molecular (Cinco, 2010; Haake, 2000). Leptospiral endotoxins appear to be a major antigen affecting host immunity and may be responsible for renal dysfunction (Barnett et al., 1999; Haake and Matsunaga, 2002). A major outer membrane lipoprotein, LipL32, can induce tubulointerstitial nephritis in mouse through mediating proinflammatory cytokines gene expression in the proximal tubule cells (Yang et al., 2002). Toll-like receptor 2 has been reported to involve in the LipL32-stimulated chemokine secretion (Hung et al., 2006; Yang et al., 2006).

So far very few leptospiral virulence factors have been identified, and the mechanisms by which the bacteria invade and adhere to the hosts are poorly understood. Recent completion of several leptospiral genomic sequences has fostered the identification of virulence factors, and been elucidated the pathogenic mechanisms in this group of bacteria (Adler et al., 2011; Hartwig et al., 2011; Xue et al., 2009). Comparative genomics has revealed approximately 2050 core leptospiral genes and more than 900 genes that are only found in the pathogenic *Leptospira* spp. (*e.g.*, *L. interrogans* and *L. borgpetersenii*), but not in saprophyte *Leptospira* spp. (*e.g.*, *L. biflexa*). A number of genes unique to either *L. interrogans* or *L. borgpetersenii* are initially thought to encode virulence-associated proteins (Adler et al., 2011). These genes include those encoding lipopolysaccharide (LPS) biosynthesis proteins, lipoproteins (LipL32, LipL36, LipL41, and LipL45), surface leptospira immunoglobulin-like (Lig) proteins, and Len proteins (Nascimento et al., 2004a). However, inactivation of the genes encoding lipoproteins and Lig proteins did not result in attenuation of virulence, suggesting that additional pathogenic mechanisms are used in the pathogenic *Leptospira* spp. Further explorations of the yet identified virulence factors will allow us better understand the leptospiral pathogenesis and accelerate vaccine development.

Although the genome sequences of several pathogenic *Leptospira* species have been described to date (Bulach et al., 2006; Nascimento et al., 2004b; Picardeau et al., 2008; Ren et al., 2003; Zhong et al., 2011), the genome sequence of *L. santarosai* serovar Shermani remains lacking. This paper describes the *L. santarosai* serovar Shermani genome and its comparison with previously sequenced *Leptospira* genomes. In accordance to the same strain of *Leptospira* used for MAT serology testing, *L. santarosai* serovar Shermani str. LT821 from ATCC was taken as the standard reference strain for genomic study. We believe that the unique features identified for *L. santarosai* serovar Shermani in this study will provide information for our understanding of the biology and pathogenesis of this important pathogen in Taiwan.

2. Materials and methods

2.1. Bacterial culture and genomic DNA extraction

L. santarosai serovar Shermani str. LT821 (ATCC number 43286) was purchased from the American Type Culture Collection (Manassas, VA). The bacterium was propagated at 28 °C under aerobic conditions in liquid Ellinghausen–McCullough–Johnson–Harris (EMJH) medium (BD Diagnostics). Genomic DNA was extracted using the Qiagen Genomic DNA kit, according to the manufacturer's instructions.

2.2. Genome sequencing, assembly, and annotation

The genome of *L. santarosai* serovar Shermani was sequenced using the Illumina/Solexa paired-end sequencing technology. Sequencing raw reads was aligned onto reference sequences using the MAQ (Mapping and Assembly with Qualities; estimated error rate $<10^{-400}$) software with the strategy in Iterative Correction of Reference Nucleotides (iCORN) and assembled into contigs by using SOAPdenovo (Li et al., 2008, 2009; Otto et al., 2010). The assembled contigs were annotated from the Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP, http://www.ncbi.nlm.nih.gov/genomes/static/Pipeline.html).

2.3. Comparative genomic analysis

Previously published *Leptospira* spp. sequences were downloaded from the NCBI. GeneBank accession number AE010300.2 for the large chromosome (CI) and AE010301.2 for the small chromosome (CII) of *L. interrogans* serovar Lai str. 56601; CP001221.1 for CI and CP001222.1 for CII of *L. interrogans* serovar Lai str. IPAV; AE016823.1 for CI and AE016824.1 for CII of *L. interrogans* serovar Copenhageni str. Fiocruz L1-130; CP000348.1 for CI and CP000349.1 for CII of *L. borgpetersenii* serovar Hardjo-bovis L550; CP000350.1 for CI and CP000351.1 for CII of *L. borgpetersenii* serovar Hardjo-bovis JB197; CP000777.1 for CI , CP000778.1 for CII and CP000779.1 for plasmid p74 of *L. biflexa* serovar Patoc strain "Patoc 1 (Ames)"; CP000786.1 for CI, CP000787.1 for CII and CP000788.1 for plasmid p74 of *L. biflexa* serovar Patoc strain "Patoc 1 (Paris)". Phylogenetic analyses were performed with Molecular Evolutionary Genetics Analysis (MEGA) software version 5.0 (Tamura et al., 2011). Neighbor-joining trees were constructed based on concatenated gene sequences in the following order: *glmU*, *pntA*, *sucA*, *tpiA*, *mreA* and *lipL41* for each strain (Ahmed et al., 2006; Thaipadungpanit et al., 2007). The Tamura–Nei genetic distance model and the bootstrap method were selected in this study. To conduct genome sequence comparisons, we used wgVISTA, available through the VISTA Browser at http://genome.lbl.gov/vista/index.shtml. The protein sequences deduced from open reading frames (ORF) were used to search the GenBank databases for orthologous genes using the BLASTX analysis with a cutoff E value of $\leqq 1\times10^{-5}$. Cellular localization of the proteins was predicted by the PSORT program (Nakai and Horton, 1999). Sequence motifs, including signal peptides and lipoprotein cleavage sites were evaluated by LipoP (Juncker et al., 2003) and SignalP programs (Bendtsen et al., 2004). PFAM (Finn et al., 2006) and InterProScan (Quevillon et al., 2005) web servers were used to predict functional and structural domains within the putative amino acid sequences. The protein sequences were subjected to comparative homology protein modeling via SWISS-MODEL and searched for functional associations using the STRING database (Arnold et al., 2006; Kiefer et al., 2009; Szklarczyk et al., 2011).

2.4. Nucleotide sequence accession number

The genome sequence of *L. santarosai* serovar Shermani has been deposited at DDBJ/EMBL/GenBank under the accession ADOR00000000. The version described in this paper is the first version, ADOR01000000.

3. Results and discussion

3.1. Assembly and annotation of the L. santarosai serovar Shermani genome sequence

The *L. santarosai* serovar Shermani genome was sequenced using high-throughput Illumina–Solexa sequencing platforms, and the data set was made up of 7,987,143 paired-end reads with a 75 bp average length and a raw coverage depth of 150×. Because of genomic diversity in *Leptospira* spp., we first performed un-gapped alignment of these reads to the reference sequence using the MAQ software with iCORN algorithm for the efficient mapping and the improvement in the assembly quality. One way to evaluate the similarity of *L. santarosai* serovar Shermani with other leptospiral species