Am. J. Trop. Med. Hyg., 94(2), 2016, pp. 302–313
doi:10.4269/ajtmh.15-0401

Whole Genome Shotgun Sequencing Shows Selection on *Leptospira* Regulatory Proteins during in vitro Culture Attenuation

Jason S. Lehmann,† Victoria C. Corey,† Jessica N. Ricaldi, Joseph M. Vinetz, Elizabeth A. Winzeler, and Michael A. Matthias*

Division of Infectious Diseases, School of Medicine, University of California, La Jolla, San Diego, California; Biomedical Sciences Graduate Program, University of California, La Jolla, San Diego, California; Department of Pediatrics, School of Medicine, University of California, La Jolla, San Diego, California; Instituto de Medicine Tropical "Alexander von Humboldt," Department of Cellular and Molecular Sciences, Faculty of Sciences and Laboratory of Research and Development, Universidad Peruana Cayetano Heredia, Lima, Peru

Abstract. Leptospirosis is the most common zoonotic disease worldwide with an estimated 500,000 severe cases reported annually, and case fatality rates of 12–25%, due primarily to acute kidney and lung injuries. Despite its prevalence, the molecular mechanisms underlying leptospirosis pathogenesis remain poorly understood. To identify virulence-related genes in *Leptospira interrogans*, we delineated cumulative genome changes that occurred during serial in vitro passage of a highly virulent strain of *L. interrogans* serovar Lai into a nearly avirulent isogenic derivative. Comparison of protein coding and computationally predicted noncoding RNA (ncRNA) genes between these two polyclonal strains identified 15 nonsynonymous single nucleotide variant (nsSNV) alleles that increased in frequency and 19 that decreased, whereas no changes in allelic frequency were observed among the ncRNA genes. Some of the nsSNV alleles were in six genes shown previously to be transcriptionally upregulated during exposure to in vivo-like conditions. Five of these nsSNVs were in evolutionarily conserved positions in genes related to signal transduction and metabolism. Frequency changes of minor nsSNV alleles identified in this study likely contributed to the loss of virulence during serial in vitro culture. The identification of new virulence-associated genes should spur additional experimental inquiry into their potential role in *Leptospira* pathogenesis.

INTRODUCTION

Leptospirosis, caused by pathogenic species of the genus *Leptospira,* is an emerging zoonotic infection of global distribution.[1] Recent estimates by the Leptospirosis Burden Epidemiology Reference Group have placed the number of hospitalized cases at over 500,000 per year[2]; this, more than likely, is an underestimate of the true burden of disease due primarily to inadequate diagnostics, a lack of clinical awareness, and poor surveillance.[3] Transmission to humans occurs via exposure to contaminated water and wet soil or infected tissues and urine from chronically colonized reservoir hosts. Humans living in poverty with poor sanitation are at greatest risk, particularly during seasonal flooding, monsoons, and tropical cyclones.[1,3]

The *Leptospira* genus includes at least 22 species classified into three large subgroups based on 16S rDNA phylogeny, in vitro growth characteristics, and virulence.[4–7] There are 15 recognized pathogenic species. Group I pathogens (Figure 1) comprise > 250 serotypes and cause disease varying in severity, ranging from subclinical infections to severe disease—often associated with renal failure and pulmonary hemorrhage—and death.[8] By contrast, group II species grow better in culture and cause predominantly mild, self-resolving illness without fatal complications. Of the pathogenic species, *Leptospira interrogans*, the leading cause of leptospirosis-associated morbidity and mortality in humans, is the most extensively studied species.[1,2] Nonetheless, the molecular mechanisms underlying *L. interrogans* pathogenesis remain largely unknown primarily because targeted gene knockouts in pathogenic *Leptospira* is inefficient and technically challenging.[9] Despite this barrier to progress in the field, transposon mutagenesis, first reported by Bourhy and others[10] and Murray and others,[11] has been successful. Though technically difficult, targeted gene knockouts have also been described and used to validate a handful of *Leptospira* virulence-related genes (e.g., *fliY*, *colA*, *mce*).[12–14]

Given the difficulty of targeted gene knockouts, systems-based approaches, including transcriptome and comparative genome analysis, have been used to identify potential virulence candidates. Microarrays have been applied to investigate the transcriptional response of pathogenic *Leptospira* to various "host-like" conditions including temperature,[15,16] serum,[17] physiological osmolarity,[18] iron depletion,[19] and host immune cells.[20] Recent RNA-seq experiments have further improved our understanding of global transcriptional responses during *Leptospira* growth in vivo.[21,22] In addition, our group has applied comparative genome analysis to identify 452 conserved pathogen-specific genes that likely play a role in *Leptospira* pathogenesis.[4,23–27] Nonetheless, the contribution of individual genes or combinations of genes to the overall virulence phenotype of pathogenic *Leptospira* remains poorly understood.

In a previous independent study, we used reference-guided assemblies to identify inactivating nonsynonymous single nucleotide variant (nsSNVs) in 11 putative virulence-associated genes that had emerged after passaging a P1 isolate for 18 subcultures including a family of virulence-modifying proteins upregulated during in vivo in an acute hamster infection model.[28] However, in this experiment, we considered only dominant alleles in P1 and P18 isolates. Here, in an independent attenuation experiment, we serially in vitro passaged the P1 isolate ($LD_{50} < 100$ *Leptospira*) into an avirulent derivative (P8A, $LD_{50} > 10^8$). We define the cumulative genome changes accompanying this observed loss of virulence by comparing the genomes of the parental strain and its isogenic, attenuated

*Address correspondence to Michael A. Matthias, Department of Medicine, Division of Infectious Diseases, School of Medicine, University of California, San Diego School of Medicine, 9500 Gilman Drive, BRF 2, Room 4A15, La Jolla, CA 92093-0760. E-mail: mmatthias@ucsd.edu

†These authors contributed equally to this work.

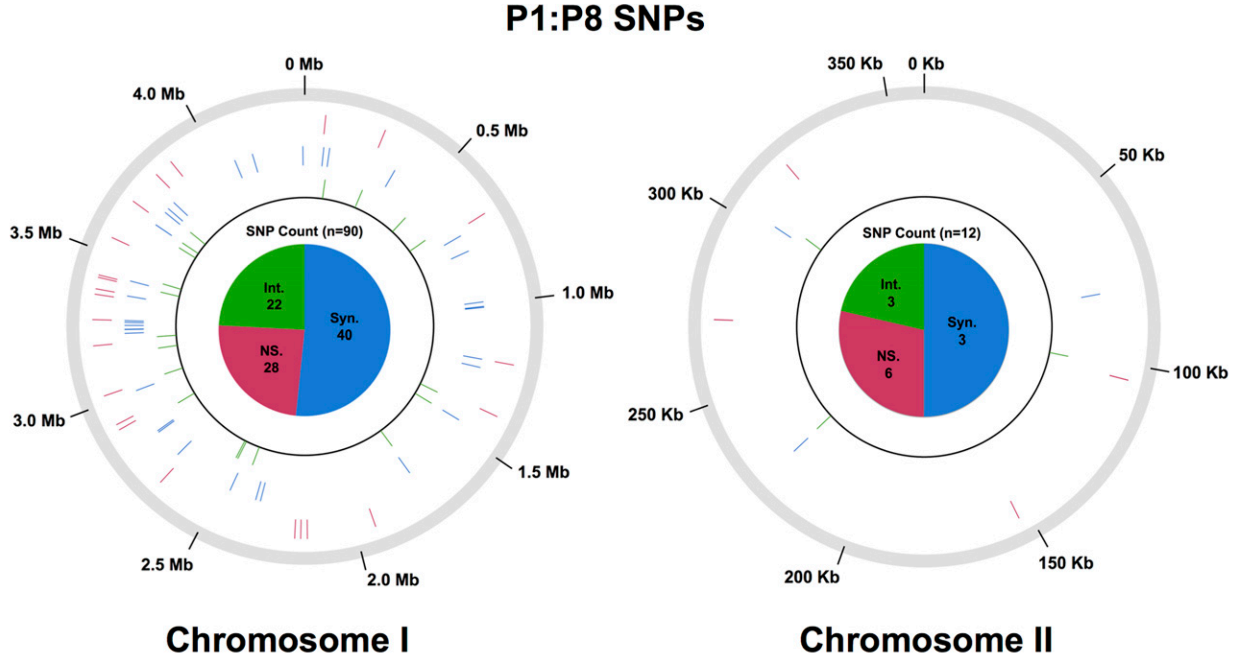


FIGURE 1. Genomic locations of single nucleotide variants (SNVs) that change allelic frequency from P1 to P8A. Genomic location of study identified SNVs in the reference genome of *Leptospira interrogans* serovar Lai strain 56601 that significantly changed in allelic proportionality during culture attenuation from a virulent P1 strain to an attenuated P8A strain. Individual hash marks represent the genomic location of genes containing SNVs, and are color coded in concentric circles. Red = nonsynonymous; blue = synonymous; green = intergenic. The total number of genes containing each type of SNV is represented by the pie chart in the center of each of the chromosome representations.

derivative through the use of next-generation sequencing and a custom SNV calling pipeline.[29]

METHODS

Ethics statement. The experimental animal work was carried out in accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health in Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC)-approved facilities, and was approved by the Institutional Animal Care and Use Committee of the University of California, San Diego under protocol S03128H.

Attenuation of *L. interrogans* serovar Lai strain 56601. *Generation of the P1 isolate of L. interrogans serovar Lai strain 56601. Leptospira interrogans* serovar Lai strain 56601 was kindly provided by David Haake (University of California Los Angeles, Los Angeles, CA), and was passaged through 3-week-old male Golden Syrian hamsters (N = 3, Charles Rivers Laboratories, Hollister, CA) to ensure a virulent phenotype. The initial three hamsters were each injected intraperitoneally (IP) with approximately 10^7 *Leptospira* in 1 mL of Ellinghausen-McCullough-Johnson-Harris *Leptospira* culture media (EMJH; BD Difco, Franklin Lakes, NJ). Four days post inoculation the animals were killed, the livers were harvested, macerated with a sterile scalpel blade, pooled in 5-mL sterile phosphate-buffered saline, then made into a slurry by repeatedly passing through a 22-gauge needle; 1 mL of this homogenate was then used to inject each of a second group (N = 3) of hamsters IP. The liver homogenization procedure was repeated 4 days later, and a third group (N = 3) of hamsters were injected, also IP. Four days after the IP injection of liver homogenate into the third group, the animals were killed, and livers harvested aseptically. Approximately 10 mg of minced liver tissue was then used to inoculate EMJH semisolid medium supplemented with 5-fl.[30] The semisolid culture was incubated at 25°C and monitored for *Leptospira* growth by dark field microscopy. Once growth occurred, 100 μL of this culture was used to inoculate 20 mL of sterile EMJH media, and the culture was incubated at 28°C on a rotary shaker, and was designated P1.

Genomic DNA isolation of P1 isolate. Approximately 10^7 *Leptospira* from 1 mL of EMJH P1 culture were spun down in a microcentrifuge (10,000 rpm, 5 minutes). Genomic DNA was then isolated from the cell pellet using the DNEasy Blood and Tissue kit (Qiagen, Valencia, CA) according to manufacturer's instructions. Eluted DNA was stored at −20°C for later sequencing.

LD_{50} determination of P1 isolate. Leptospira cells were counted using a Petroff-Hauser counting chamber (Hausser Scientific, Horsham, PA) under dark field microscopy. Challenge doses of 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, and 10^8 *Leptospira*/mL in sterile EMJH were then prepared based on observed counts. For each dilution group, 3-week-old male Golden Syrian hamsters (N = 3, Charles Rivers Laboratories) were each injected IP with 1 mL of the appropriate challenge dose. Animals were monitored for 21 days and euthanized when moribund. The LD_{50} was defined as the last dose in which two-thirds of the animals died after challenge.

In vitro EMJH culture-passage attenuation of the virulent P1 isolate into P8 isolate. The P1 isolate EMJH culture was subcultured by transferring 2 mL into 18 mL of sterile EMJH media (thus becoming P2A), and incubated at 28°C on a rotary shaker for 14 days. This process was repeated iteratively for a total of seven subcultures, with the final subculture being designated P8A (~400 generations from the P1 parent culture). Genomic DNA extraction and LD_{50} determination were then performed exactly as described for the P1 isolate.

Genomic library preparation and assembly. Genomic DNA libraries were normalized to 0.2 ng/μL and prepared for sequencing using the Illumina Nextera XT Kit (Illumina, San Diego, CA) whole genome resequencing library according to manufacturer's instructions, using the Illumina protocol of tagmentation followed by ligation (v. 2013; Illumina, Inc., San Diego). DNA libraries were clustered and run on an Illumina HiSeq 2500 platform (Illumina) with PE100 on Rapid Run mode. Base calls were made using CASAVA v 1.8+ (Illumina).

Sequences were processed though the PLATYPUS pipeline (Winzeler Lab, UCSD, San Diego, CA).[29] In brief, reads were aligned to the reference *L. interrogans* serovar Lai strain 56601 genome (NC_004342 and NC_004343) using Burrows-Wheeler Aligner,[31] and unmapped reads were filtered using SAMtools.[32] SNVs were then initially called using Genome Analysis Toolkit[33,34] and filtered using default filter values in PLATYPUS. Although the filters were initially designed for *Plasmodium falciparum*, they resulted in high sensitivity (93.4%) and specificity (91.2%) for *L. interrogans* as well when screening for known SNVs between the 56601 and IPAV *L. interrogans* serovar Lai strains. After alignment, read depth per nucleotide identity at every position was called using SAMtools *mpileup*, which were then converted into proportional nucleotide identities per base. These proportions were then compared using a custom script testing for multi-comparison significant changes in allelic proportion across the entire genome. For two proportions x_1/n_1 and x_2/n_2 reads, our comparison statistic was

$$z = \frac{\left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right)}{\sqrt{\left(\frac{x_1 + x_2}{n_1 + n_2}\right)\left(1 - \frac{x_1 + x_2}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

This statistic is an expansion of the simple two-proportional *z*-test for differences between two populations. This assumption is reasonable as each read serves as an independent random test of the nucleotide identity of the population, though significant error terms do exist. This number was then converted to a p-statistic using the total read depth and corrected using the Bonferroni method, as assumption about the independence of allelic frequency at multiple polymorphic sites may not hold. A list of sites that underwent statistically significant changes were then exported and annotated using a custom script.

Clusters of orthologous groups' functional category analysis of nsSNV-containing genes. Genes identified as containing nsSNVs with increasing allele frequencies in P8 were assigned to clusters of orthologous groups (COG) categories using the National Center for Biotechnology Information conserved domain webpage (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml), and compared with the genome-wide predicted COG frequencies for *L. interrogans* serovar Lai strain 56601 obtained from the spirochete genome browser webpage (http://sgb.fli-leibniz.de/cgi/index.pl). Statistical significance was assessed via χ^2 analysis using Fisher's exact test with a Bonferroni correction to account for multiple comparisons in Graphpad Prism (GraphPad Software, Inc., La Jolla, CA).

Pan genus comparative genome analysis of study-identified genes. The following genomes, consisting of a representative grouping all 20 *Leptospira* species, were used to analyze the presence of homologs of study-identified genes in other *Leptospira* species:

Leptospira alexanderi sv. Manhoa 3 str. L 60^T (Genbank: AHMT00000000), *Leptospira alstoni* sv. Pingchang str. 80-412 (Genbank: AOHD00000000), *Leptospira biflexa* sv. Patoc str. Patoc I Paris (Genbank: CP000786), *Leptospira borgpetersenii* sv. Javanica str. UI 09931 (Genbank: AHNP00000000), *Leptospira broomii* sv. Hurstbridge str. 5399^T (Genbank: AHMO00000000), *Leptospira fainei* sv. Hurstbridge str. BUT 6^T (Genbank: AKWZ00000000), *Leptospira inadai* sv. Lyme str. 10^T (Genbank: AHMM00000000), *L. interrogans* sv. Copenhageni str. Fiocruz L1-130 (Genbank: AE016823), *L. interrogans* sv. Lai str. 56601 (Genbank: AE010300), *Leptospira kirschneri* sv. Cynopteri str. 3522 C^T (Genbank: AHMN00000000), *Leptospira kmetyi* sv. Malaysia str. Bejo-$Iso9^T$ (Genbank: AHMP00000000), *Leptospira licerasiae* sv. Varillal str. VAR 010^T (Genbank: AHOO00000000), *Leptospira meyeri* sv. Hardjo str. Went 5 (Genbank: AKXE00000000), *Leptospira noguchii* sv. Panama str. CZ 214^T (Genbank: AKWY00000000), *Leptospira santarosai* sv. Shermani str. $1342K^T$ (AOHB00000000), *Leptospira terpstrae* sv. Hualin str. LT $11\text{-}33^T$ (Genbank: AOGW00000000), *Leptospira vanthielii* sv. Holland str. WaZ Holland (Genbank: AOGY00000000), *Leptospira weilii* sv. undetermined str. LNT 1234 (Genbank: AOHC00000000), *Leptospira wolbachii* sv. Codice str. CDC (Genbank: AOGZ00000000), *Leptospira wolffii* sv. undetermined str. Khorat-$H2^T$ (Genbank: AKWX00000000), *Leptospira yanagawae* sv. Saopaulo str. Sao $Paulo^T$ (Genbank: AOGX00000000).

Genes were considered homologs if they were bidirectional best hits[35,36] using Basic Local Alignment Search Tool (BLAST) with cutoff values of 70% query coverage, e-values < $1e^{-3}$, and 30% identity.

Amino acid residue conservation analysis of study-identified nsSNV positions. Domain architecture analysis was performed on the protein sequences for LA_2704 (NP_712885.1), LA_2930 (713110.1), LA_2950 (NP_713130.1), LA_3455 (NP_713635.1), LA_3725 (NP_713905.1), and LA_3834 (NP_714014.1) using Simple Modular Architecture Research Tool (SMART)[37] and protein structure prediction server (PSIPRED),[38,39] and represented graphically at http://prosite.expasy.org/mydomains.

Multiple sequence alignments (MSAs) of the homologs (defined by 70% query coverage, e-values < $1e^{-3}$, and 30% identity BLAST cutoffs) of each of these six genes were constructed by aligning sequences obtained from the Pathosystems Resource Integration Center database (http://patricbrc.org) using the CLUSTAL X alignment program freely available at http://www.clustal.org/clustal2/#Download. The accession numbers used in the alignments (Supplemental Table 1) are a representative collection of homolog sequences from each of the 20 species in the *Leptospira* genus in which homolog sequences could be identified. The LA_2704 alignment contained 40 homolog sequences, LA_2930 had 26 homologs, LA_2950 had 41 homologs, LA_3455 had 41 homologs, LA_3725 had 19 homologs, and LA_3834 had 45 homologs.

MSAs were then used to predict protein residue conservation based on Jensen-Shannon Divergence (JSD).[40] Conservation scores were then graphed using Microsoft Excel (Redmond, WA).

The proportion of sequencing reads from the P8A strain coding for the nsSNV amino acid was compared with the proportion of the same mutant residues in homolog MSAs from the entire pan-*Leptospira* genome using a Fisher's exact test for each of the six genes. Results were considered statistically significant at $P < 0.05$.

Identification of potential ncRNAs in the *L. interrogans* serovar Lai strain 56601 genome. To identify novel ncRNA loci within the *L. interrogans* Lai strain 56601 genome, we first aligned the *L. interrogans* Lai 56601 (Genbank: AE010300), *L. kirschneri* Cynopteri 3522 C (Genbank: AHMN00000000), and *L. noguchii* Panama CZ214^T (Genbank: AKWY00000000) genomes using the progressive Cactus algorithm.[41,42] The whole genome alignment was then used as input for RNAz (with default settings: -w 120 and -s 120) for prediction of structural RNAs[43] and then putative ncRNA loci identified and annotated using the nocoRNAc pipeline.[44] Predicted loci that could not be annotated using the Rfam database were considered potentially novel ncRNA genes.

RESULTS

Culture passage-based attenuation of *L. interrogans* serovar Lai strain 56601. The P1 isolate was derived from *L. interrogans* serovar Lai strain 56601 that had been serially passaged 3X in vivo to ensure a virulent phenotype.[28] The LD_{50} was determined to be $< 10^2$ *Leptospira*.[28] The P1 isolate was serially passaged in vitro in liquid *Leptospira* culture medium for ~400 generations (16 weeks) to become P8A. The LD_{50} of the P8A isolate was determined to be $> 10^8$ *Leptospira*, administered IP, indicating a complete loss of virulence. After in vitro passage, genomic DNA was isolated from both the P1 and P8A strains and frozen before sequencing.

Identification of SNV alleles differing in frequency between the attenuated and parental strains. Cumulative changes occurring during adaptation to in vitro growth, and associated with loss of virulence, were studied at the whole genome level. Genomic DNA from the nonclonal parental strain, P1, and from the attenuated isogenic derivative, P8A, was sequenced on an Illumina platform using paired-end 100-bp reads to a mean coverage of greater than 250X. For strain P1, 15,492,436 reads were generated covering 99.4% of the *L. interrogans* serovar Lai reference genome (4.689 Mb), and 15,651,273 reads were generated from P8A covering 99.9% of the reference genome (Table 1). In addition, > 99% of the reads from both the P1 and P8A samples aligned to the *L. interrogans* Lai strain 56601 genome, indicating high sample purity.

Variants were called and compared using a modified automated PLATYPUS genome analysis pipeline.[29] PLATYPUS aligned reads from each sequencing run (P1 and P8A) to the reference Lai genome[27] and identified SNVs using a default list of filters for each set of sequencing files. Given that the bacterial populations were not clonal, an allele frequency was calculated at each polymorphic site using the number of aligned reads metric for the P1 and P8A isolate (Supplemental Table 1). From this analysis, 99 SNVs were identified as having undergone a significant change in allele frequency between P1 and P8A, as determined by a two-proportional z-test before Bonferroni correction. Alternate nucleotides in these positions would result in 43 SNVs encoding synonymous amino acid substitutions, 34 encoding nonsynonymous amino acid substitutions, and 25 intergenic SNVs (Figure 1, see Supplemental Table 1 for complete listing of variants). In the P8A genome, all of these minor alleles had changed allele frequencies by at least 5% compared with the P1 genome and vice versa.

Analysis of nsSNVs with allelic frequencies that increased during attenuation. Since amino acid coding changes can alter overall functionality of the gene in which they reside, and may contribute to the observed loss of virulence in the P8A strain, we further examined the nsSNVs that were identified during our genomic comparisons. There were 15 genes that contained nsSNVs that increased in frequency during the course of the attenuation (Figure 2A). To determine if the genes containing these nsSNVs were biased toward any particular biological function, they were organized by COG category,[45] and the observed proportions were compared with their genome-wide expected frequencies. This approach identified a strong enrichment for genes involved in signal transduction mechanisms (Figure 2B). Of the 3,683 genes in the genome, 233 are annotated as involved in signal transduction and comprised five of the 15 in our set ($P = 0.01$). We also noted that three genes, *rbsK, mgtA*, and *mcm2* (encoding a putative ribokinase, a magnesium transporter, and methylmalonyl-CoA mutase, respectively), contained multiple SNVs. This is a higher number than that would be expected due to chance alone, and because these genes all have functions related to core metabolic pathways and cofactor biosynthesis, their allele frequency increase may be a result of bacterial adaptation to long-term in vitro culture conditions.

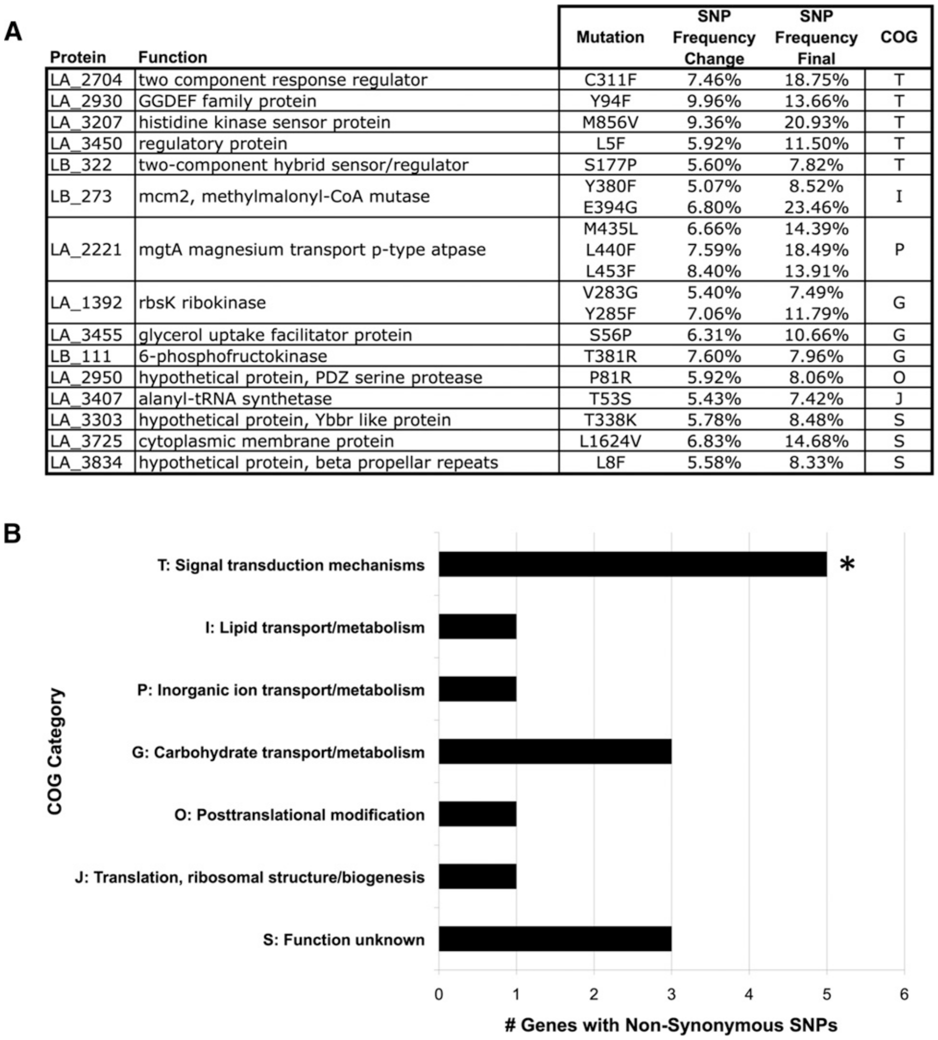
To infer additional possible biological significance of these genes, we performed a meta-analysis of previously published data showing transcriptional responses of *L. interrogans* under several surrogate in vivo conditions including temperature, physiological osmolarity, iron depletion, exposure to host innate immune cells, and peritoneal culture of pathogenic *Leptospira* in dialysis membrane chambers.[15,16,18–21] Of the 15 genes identified by this study as harboring nsSNVs of increasing allele frequency, six (LA_2704, LA_2930, LA_2950, LA_3455, LA_3725, and LA_3834) were previously reported to be upregulated in at least one set of in vitro surrogate experimental conditions.

To gain further insight into how these genes might contribute to the pathogenicity of *Leptospira* and their overall prevalence in the genus, the subcellular locations of the proteins they encode were predicted using PSORTb v. 3.0 (http://psort.org/psortb/index.html),[46] and the prevalence of gene homologs

TABLE 1

Genome alignment statistics for *Leptospira interrogans* serovar Lai strains P1 and P8A

Sample	Median insert size (bp)	Total reads	Aligned reads (%)	Mean coverage	% Bases > 20X	% Bases > 40X	% Bases > 100X	% Bases > 130X	% Bases > 150X
P1	113	15,583,466	15,492,436 (99.4)	250.49	100	99.9	98.8	94.9	89.6
P8	129	15,668,032	15,651,273 (99.9)	263.53	100	99.9	98.4	95.5	91.8



Protein	Function	Mutation	SNP Frequency Change	SNP Frequency Final	COG
LA_2704	two component response regulator	C311F	7.46%	18.75%	T
LA_2930	GGDEF family protein	Y94F	9.96%	13.66%	T
LA_3207	histidine kinase sensor protein	M856V	9.36%	20.93%	T
LA_3450	regulatory protein	L5F	5.92%	11.50%	T
LB_322	two-component hybrid sensor/regulator	S177P	5.60%	7.82%	T
LB_273	mcm2, methylmalonyl-CoA mutase	Y380F E394G	5.07% 6.80%	8.52% 23.46%	I
LA_2221	mgtA magnesium transport p-type atpase	M435L L440F L453F	6.66% 7.59% 8.40%	14.39% 18.49% 13.91%	P
LA_1392	rbsK ribokinase	V283G Y285F	5.40% 7.06%	7.49% 11.79%	G
LA_3455	glycerol uptake facilitator protein	S56P	6.31%	10.66%	G
LB_111	6-phosphofructokinase	T381R	7.60%	7.96%	G
LA_2950	hypothetical protein, PDZ serine protease	P81R	5.92%	8.06%	O
LA_3407	alanyl-tRNA synthetase	T53S	5.43%	7.42%	J
LA_3303	hypothetical protein, Ybbr like protein	T338K	5.78%	8.48%	S
LA_3725	cytoplasmic membrane protein	L1624V	6.83%	14.68%	S
LA_3834	hypothetical protein, beta propellar repeats	L8F	5.58%	8.33%	S

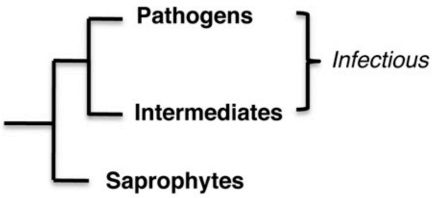
B
T: Signal transduction mechanisms
*
I: Lipid transport/metabolism
P: Inorganic ion transport/metabolism
G: Carbohydrate transport/metabolism
O: Posttranslational modification
J: Translation, ribosomal structure/biogenesis
S: Function unknown
COG Category
0
1
2
3
4
5
6
# Genes with Non-Synonymous SNPs

FIGURE 2. Clusters of orthologous group analysis of genes containing nonsynonymous single nucleotide variants (nsSNVs) that increased allelic frequency from P1 to P8A. The nsSNVs that increased allelic frequency during the attenuation of the virulent P1 strain of *L. interrogans* serovar Lai strain 56601 into the avirulent isogenic P8A strain are listed in (**A**). The nsSNV-containing genes in each group were then organized by clusters of orthologous groups (COG) category.[45] Asterisks denote an enrichment of a particular COG category compared with genome-wide expected percentages of genes in each category (Fisher's exact test, *P* value given in figure) in (**B**). Total number of genes in *Leptospira interrogans* serovar Lai strain 56601 genome in represented COG categories: T = 233, I = 110, *P* = 134, G = 131, O = 118, J = 172, S = 853. Total number of predicted genes = 3,683.

across all 20 species of the *Leptospira* pan-genome was also determined. Genes in other *Leptospira* species were considered homologous to our study-identified genes if they were reciprocal best BLAST hits using filters of 70% query length, e-value < $1e^{-3}$, and 30% identity match (Figure 3). This analysis revealed that five of the six genes (the subcellular location of LA_2950 could not be determined by the algorithm) were predicted to reside inside the bacterial cell, indicating that these proteins are likely not the ultimate effectors of *Leptospira* pathogenesis, like toxins or other secreted factors, but may contribute to upstream signaling processes or metabolic capability. The pan-genus conservation analysis showed that three genes (LA_2930, LA_3725, and LA_3834) are found only in infectious *Leptospira* species and may have particularly relevant pathogenesis-related functions.

Pan-*Leptospira* genomic analysis of amino acid residue conservation at nsSNV positions in homologs of attenuation-identified genes. We conducted a three-part analysis of six genes of interest (Figure 3) to determine if the nsSNVs in these genes caused amino acid changes in evolutionarily conserved residues. First, protein domain architecture was evaluated using SMART[37] and PSIPRED.[38,39] Next, we generated MSAs using homologous sequences from the 20 species pan-*Leptospira* genome for each of these genes. These MSAs were used to generate amino acid conservation scores for each residue in a respective gene based on JSD (scores above 0.8 are considered highly conserved, those less than 0.4 are considered disordered).[40] Finally, we compared the proportion of sequencing reads from the P8A strain coding for the nsSNV amino acid to the proportion of the same mutant residue in homologs from the entire pan-*Leptospira* genome using a Fisher's exact test with the following results.

LA_2704. Diguanylate cyclases participate in the formation of the ubiquitous second messenger, cyclic diguanylate monophosphate (cyclic-di-GMP), involved in bacterial virulence, biofilm formation, and persistence.[47,48] The nonsynonymous C311F substitution in this GGDEF, diguanylate cyclase is C-terminal to the catalytic core of this protein by one amino



				L. interrogans Copenhageni str. Fiocruz L1-130	L. kirschneri Cynopteri 3522 CT	L. noguchii Panama CZ 214T	L. alstoni serovar Pingchang str. 80-412	L. weilii serovar Ranarum str. ICFT	L. alexanderi Manhao 3 L 60T	L. borgpetersenii Javanica UI 09931	L. santarosai serovar Shermani str. 1342KT	L. kmetyi undetermined Bejo-Iso9T	L. fainei Hurstbridge BUT 6T	L. broomii undetermined 5399T	L. wolffii undetermined Khorat-H2T	L. licerasiae Varillal VAR 010	L. inadai Lyme 10T	L. wolbachii serovar Codice str. CDC	L. yanagawae serovar Saopaulo str. Sao Paulo	L. biflexa Patoc Patoc1	L. vanthielii serovar Holland str. Waz Holland	L. terpstrae serovar Hualin str. LT 11-33T	L. meyeri Hardjo Went 5
Protein	**Function**	**Upregulation**	**PSORTb**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**P**	**I**	**I**	**I**	**I**	**I**	**S**	**S**	**S**	**S**	**S**	**S**
LA_2704	two component response regulator	*a*	CM	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
LA_2930	GGDEF family protein	*a*	C	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-
LA_2950	HtrA2, PDZ serine protease	*a*	UNK	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
LA_3455	glycerol uptake facilitator protein	*a*	CM	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
LA_3725	cytoplasmic membrane protein	*a*	CM	+	+	+	-	-	-	-	-	-	+	-	-	-	+	-	-	-	-	-	-
LA_3834	putative lipoprotein	*a,b,c*	CM	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-

FIGURE 3. Homolog identification and characterization of potential virulence-associated genes in other *Leptospira* species. Potential virulence-associated genes identified in this study include LA_2704 (NP_712885.1), LA_2930 (NP_713110.1), LA_2950 (NP_713130.1), LA_3455 (NP_713635.1), LA_3725 (NP_713905.1), and LA_3834 (NP_714014.1). Previous studies have shown these genes to be upregulated by *Leptospira interrogans* during exposure to host-like conditions (*a* = response to host innate immunity[20]; *b* = response to host physiological osmolarity,[18] *c* = response to host cues during in vivo culture in intraperitoneal dialysis cassettes[21]). The PSORTb predicted subcellular locations of each of these proteins are listed. The presence of orthologous genes (defined as bidirectional best Basic Local Alignment Search Tool [BLAST] hits with minimum 70% query coverage, e-values < $1e^{-3}$, and 30% identity) was also determined for each of the 20 species in the *Leptospira* genus. P = pathogenic species; I = intermediate pathogens; S = saprophytic species). A schematic representation of these three clades of the genus was included for clarity.

acid residue (Figure 4A).[49] The wild-type cysteine residue is conserved in every single homolog evaluated in this study, which is reflected by the high conservation score obtained from JSD analysis. The proportion of phenylalanine substitutions observed in the P8A strain represents a highly significant divergence from the genus-wide residue conservation at this position ($P < 0.001$).

LA_2930. The Y94F substitution in this Per-ARNT-Sim-(PAS)-GGDEF predicted signaling protein falls within the PAS sensor domain (Figure 4B). PAS domains detect a large range of chemical and physical signals and then regulate the activity of their covalently linked effector domains, often by promoting the formation of dimers (a process required for proper GGDEF domain function). We could not deduce any insight into the particular ligands to which the PAS domain of this protein may bind, as the range of potential signals is diverse (ranging from oxygen tension to small metabolites and to light itself) and on average, the pairwise identity shared between PAS domains is less than 20%.[50] Nonetheless, conservation analysis revealed that this position is highly conserved in *Leptospira* with significant divergence ($P < 0.03$) away from conservation status in the P8A attenuated strain.

LA_2950. post synaptic density protein (PDZ) serine proteases are a unique family of proteins that form higher order oligmeric structures and have been demonstrated to degrade misfolded proteins in the periplasm of bacteria.[51] The P81R nsSNV in this PDZ serine protease was found to occur in an in-silico predicted coil to sheet transition, indicating that the wild-type proline residue may serve a structural role (Figure 4C). JSD conservation analysis revealed the site to be moderately conserved within the *Leptospira* genus. The P8A arginine substitution at this residue was significant ($P < 0.05$), and was not observed in any of the protein homologs evaluated. Interestingly, domain architecture analysis revealed an N-terminal signal peptide indicating that this protein potentially has extracellular function.

LA_3455. This protein is a transmembrane nonselective transport channel found in the inner membrane of gram-negative bacteria that facilitates the diffusion of glycerol.[52] The S56P substitution in this protein was a significant divergence from genus-wide expected residues ($P < 0.01$) (Figure 4D). The conserved residue position lies at the end of one of the eight α-helical regions of the aquaglyceroporin. The tight spatial arrangement of these helices to one another is essential for the proper function of the protein's glycerol-conducting channel,[53] and the proline substitution in the P8A population of *Leptospira* could conceivably introduce a structural change that would alter its transport efficiency.

LA_3725. Domain analysis of the large LA_3725 protein revealed a single N-terminal transmembrane domain and a pre-toxin Hedgehog/Intein (HINT) domain (Pfam PF07591) nearer the C-terminal end of the coding region. The HINT superfamily belongs to a system of proteases that in bacteria are usually found N-terminal to a toxin module in polymorphic toxin systems,[54,55] and are believed to release the toxin domain via autoproteolysis. The L1624V nsSNV position lies at the in silico predicted transition of an α helix to a coiled secondary structure in a region of low sequence conservation (Figure 4E). MSA analysis revealed that the P8A proportion of nsSNV reads was not statistically significant compared with genus-wide expected ratios.

LA_3834. The nsSNV position identified in the attenuated P8 strain codes for an L8F substitution in the N-terminal

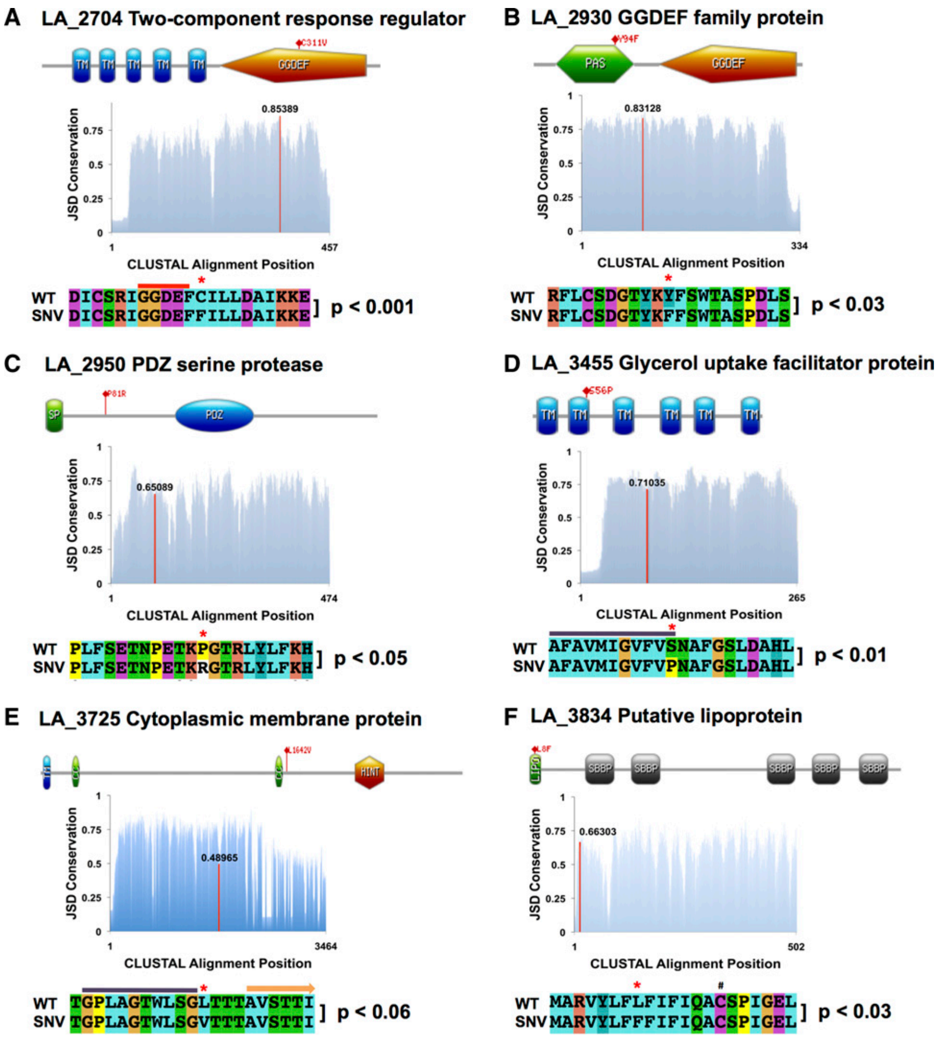


FIGURE 4. Amino acid conservation analysis at nonsynonymous single nucleotide variants (nsSNV) positions in study identified gene homologs across the *Leptospira* genus. Protein domain analysis was conducted for all genes and results are represented as diagrams at the top of panels **A**–**E** that include the nsSNV position for each gene (TM = transmembrane domain; PAS = Per-ARNT-Sim domain; GGDEF = diguanylate cyclases domain; SP = signal peptide; PDZ = PDZ serine protease domain; CC = coiled-coil domain; HINT = hedgehog intein domain; LIPO = *Leptospira* lipobox; SBBP = seven-bladed beta propeller domain). In addition, a Jensen-Shannon Divergence (JSD) estimate of amino acid residue conservation is represented graphically with the nsSNV residue highlighted as a red line with the conservation score indicated above it (scores above 0.8 indicate high conservation, those below 0.4 indicate disorder). Finally, at the bottom of each panel, an alignment schematic with the nsSNV position highlighted by a red asterisk is presented with the probability from a Fisher's exact test comparison of the number of P8A sequencing reads coding for the mutant amino acid to the genome-wide prevalence of that same residue in homologs of that particular gene. In (**A**), the red line above the alignment indicates the position of the catalytic residue of the protein. In (**D**), the purple line represents a conserved α helix. In (**E**), the purple line represents a predicted α helix, and the orange arrow represents a predicted beta-sheet secondary structure. In (**F**), the hash mark denotes the cysteine residue that is lipidated.

lipobox[56] of this protein (Figure 4F). This amino acid substitution occurs at a moderately conserved residue according to JSD analysis, that is, seven residues upstream of the cysteine residue that is lipidated during export through the bacterial inner membrane. Although there appeared to be some flexibility in the amino acid conservation at the SNV position, genus-wide analysis revealed that no homologs contained the mutant phenylalanine at this position, indicating a significant divergence from expected proportions ($P < 0.03$).

Intergenic SNV analysis and novel ncRNA prediction. Analysis of SNV allele frequency differences between the P1 and P8A *L. interrogans* Lai strains revealed 25 intergenic SNVs, 22 on chromosome I, and three on chromosome II (Figure 1, Supplemental Table 1). In previous whole genome surveys, several ncRNA loci were detected in the *L. interrogans* Lai genome,[4] including three cobalamin riboswitches that are expressed both in vivo and in vitro.[21] Because these elements play vital roles in the regulation of gene expression, mutations within predicted ncRNAs could have functional implications potentially affecting virulence.

To evaluate whether any of our study identified intergenic SNVs resided in predicted ncRNA loci, we generated a list of predicted ncRNA in the *L. interrogans* Lai strain 56601 genome using RNAz[43] and the nocoRNAc pipeline.[44] Fifty candidate ncRNA loci were identified on chromosome I (cI replicon), and five on the cII replicon, none of which contained study-identified intergenic SNVs. Of the 55 candidate ncRNA loci identified, 41 were antisense to

Table 2
Predicted ncRNAs in *Leptospira interrogans* serovar Lai

Chromosome	Start (bp)	Stop (bp)	Strand	SIDD value	Terminator confidence
CI	235,751	235,834	−	−0.5	73
CI	422,027	422,128	−	3.5	100
CI	538,202	538,259	+	2.5	72
CI	548,190	548,249	−	3.2	100
CI	1,181,513	1,181,570	+	2.0	74
CI	2,172,984	2,173,177	−	−0.2	76
CI	2,410,576	2,410,725	−	−0.3	78
CI	2,545,112	2,545,233	−	2.9	77
CI	2,823,520	2,823,576	−	−0.5	100
CI	2,876,638	2,876,810	−	−0.5	80
CI	2,935,474	2,935,522	−	−1.0	70
CI	3,382,029	3,382,080	+	2.2	89
CI	4,152,345	4,152,404	+	2.6	79
CII	73,532	73,627	−	2.5	86

In silico predicted ncRNAs in the *L. interrogans* serovar Lai genome. The 14 predicted ncRNAs are listed by chromosome, start/stop positions, as well as the DNA strand the site resides on. Stress-induced duplex destabilization (SIDD) values near zero are highly destabilized states that promote helicase action and transcription. Terminator confidence is a percentage of certainty of transcription stop sites.

protein coding genes and 14 were found in intergenic regions (Table 2). Of these 14 ncRNA loci, none could be annotated using the Rfam database and could represent novel ncRNA genes.

DISCUSSION

This study analyzed genomic changes in a polyclonal population of *L. interrogans* serovar Lai strain 56601 that occur during the culture-based attenuation of a highly virulent parent strain into a nearly avirulent isogenic derivative. This analysis was carried out using a modified PLATYPUS pipeline, originally designed to analyze eukaryotic genomes, which was readily and accurately adapted for the analysis of *Leptospira* genomes (prokaryotic). Novel, potentially virulence-related genes were identified in this study by analyzing nsSNV allele frequency changes accompanying in vitro, culture attenuation of *L. interrogans* serovar Lai. Because of the stochasticity of the underlying processes giving rise to deleterious mutations in virulence-associated genes that are under neutral selection in vitro, future attenuation experiments would be most informative if whole genome sequencing data from several independent attenuated lineages are compared. The data summarized here will provide the foundation for future investigations to determine the role these genes play in the pathogenesis of leptospirosis.

Genome changes that occurred in the bacterial population during long-term in vitro culture passage attenuation of the virulent P1 *L. interrogans* Lai strain 56601 isolate into the avirulent P8A isolate likely occurred as the result of selection for rapid growth in vitro culture, likely to be in a tradeoff with virulence. After the isolation of the P1 strain from hamsters, the only selective pressure on the bacterial population became intrapopulation competition for growth in vitro in EMJH media. Therefore, the process of natural selection under these conditions would be expected to increase the population-level allelic frequencies of mutations beneficial to in vitro growth.[57] Such changes are often accompanied by allele-frequency increases of mutations in genes necessary for growth in vivo (i.e., relaxed selection on virulence genes would lead to the accumulation of deleterious mutations in these genes during growth in vitro). Accordingly, nsSNVs otherwise deleterious to virulence in vivo, which had previously been kept at low frequencies by in vivo selection, for example, immune pressures of the host, would now be selectively neutral in vitro. These mutations would then be free to synchronously move with alleles under positive selection for growth in EMJH media in a type of genetic hitchhiking.[58]

Interestingly, all single nucleotide variations identified in our genomic analysis of the attenuated P8A isolate originated from existing low-frequency subpopulation alleles in the virulent P1 isolate. We did not find any spontaneous mutations arising during the in vitro attenuation process. Only preexisting mutations expanded in frequency based on statistically significant thresholds, a phenomenon also noted previously in the apicomplexan parasite *Babesia bovis*.[59] The original process that generated these mutants appeared to have proceeded in a stochastic manner, SNVs appeared across the *L. interrogans* Lai genome with a nucleotide transition to transversion ratio, Ts/Tv, of approximately 0.5 (Supplemental Table 1), suggesting that at a given position a substitution of one nucleotide was just as likely as any other. All nsSNVs identified existed as minor variants to wild-type alleles in the P8A population (Figure 2A). Surprisingly, nsSNVs diverging from the reference sequence had an allele frequency of only 12% ± 4.97 (mean ± SD). It has been previously demonstrated in several other pathogens that microbial populations may harbor subpopulations that retain pathogenic capacity, despite being attenuated at the population level.[60–63] Similarly, we were able to detect wild-type alleles in the majority of sequencing reads derived from the P8A isolate.

Two previous studies have examined genome differences between virulent and avirulent strains of *L. interrogans* serovar Lai to identify mutations that accompany the loss of the virulence of the parental strain. The first, which compared genome differences between *L. interrogans* serovar Lai strain IPAV (avirulent) and a non-isogenic isolate of *L. interrogans* serovar Lai,[64] identified several hundred SNVs in gene-coding regions as well as dozens of insertions and deletions; interestingly many SNVs were found in genes related to signal transduction. The second study, recently reported from our group,[28] identified a set of SNVs in 11 pathogen-specific genes of an attenuated isogenic derivative of *L. interrogans* serovar Lai strain 56601. There was no overlap in the genes identified in these earlier studies with those identified in this work, which mirrors results from another experimental evolution study in *Escherichia coli* that found few of the 115 strain replicates shared similar mutations.[57] This lack of overlap strongly underscores the stochastic nature of SNV expansion in vitro. Nonetheless, it should be noted that the sequencing coverage was approximately 2.5X higher in this study relative to our previous study, reads were substantially longer here (100 bases versus 36) and paired ends were used. Also significant, this analysis detected mixed alleles, whereas our previous study focused only on dominant alleles. Because genes necessary for in vivo growth are under relaxed selection in vitro, the complement of putative virulence genes identified in attenuation experiments can differ substantially (i.e., no convergence), suggesting, that is, approach would be most informative if data derived from several independent attenuated lineages are analyzed.

Pathogenic *Leptospira* have evolved numerous signal transduction proteins to properly respond to environmental as well as in vivo host queues,[26] in contrast to obligate parasites that have far fewer.[65] Because pathogenic *Leptospira* are transmitted by soil and surface water, they must transition between the external and host environment.

The identification of nsSNVs in two GGDEF di-guanylate cyclase (DGC) signal transduction genes (LA_2704 and LA_2930, both previously shown to be upregulated during exposure to in vivo-like conditions; Figure 3) in our study was particularly intriguing. GGDEF domains catalyze the formation of the ubiquitous secondary messenger di-cyclic-GMP[66] through a process of homo-dimerization of two DGC domains from separate proteins.[49] Intracellular concentrations of di-c-GMP have been demonstrated experimentally to regulate several pathogenesis-related bacterial processes related to biofilm formation, motility, and virulence.[67–69] The *L. interrogans* serovar Lai genome contains genes for 14 distinct GGDEF domain containing proteins,[26] and members of the genus are known to produce biofilms both in vitro and in vivo.[70,71] The physiological effects of di-c-GMP levels have been reviewed previously[47]; while intracellular di-c-GMP levels promote biofilm formation, they might have differential effects on (i.e., promote or inhibit) other phenotypes. While there are currently no experimental data regarding DGCs and di-c-GMP in *Leptospira*, it should be noted that di-c-GMP levels appear to positively regulate motility and virulence in other spirochetes.[72–74]

Data from a model biofilm system using *Pseudomonas aeruginosa* have demonstrated that increases in intracellular di-c-GMP levels, through the action of DGCs, lead to secretion of exopolysaccharide components required for biofilm formation.[75] These polysaccharides then act as signals for DGCs in neighboring bacteria to increase their di-c-GMP levels, encouraging further exopolysaccharide secretion in a positive feedback mechanism similar to paracrine signaling in eukaryotes. Whether impaired GGDEF signaling causes a similar nonautonomous trait in *L. interrogans* remains undetermined, it is interesting to consider whether a small percentage of mutant cells (i.e., those harboring LA_2704 or LA_2930 nsSNV mutations) would influence the in vivo survival of the *Leptospira* population as a whole through impaired biofilm production.

Bacterial lipoproteins have been suggested to be involved in pathogenesis including adhesion to host cells, immune modulation, and the translocation of virulence factors into host cells,[76,77] and there are several predicted in the genomes of spirochetes.[56] Thus, the identification of mutations in the putative lipoprotein LA_3834 in this study is intriguing. While the function of this protein has yet to be determined experimentally, several independent lines of evidence point to LA_3834 being a part of the *Leptospira* virulence gene repertoire. In addition to being transcriptionally upregulated during in vivo surrogate experiments,[18,20,21] LA_3834 was recently demonstrated to be under the control of a transcriptional regulator (LB_139) that when knocked out, decreased expression of several genes (including LA_3834) and attenuated virulence in a hamster model of leptospirosis.[22]

Two other study-identified genes (*LA_2950* and *LA_3455*) with nsSNVs at conserved residues may also be important to *Leptospira* virulence and survival in vivo. LA_2950 encodes a protein predicted PDZ serine protease. Studies in *Salmonella typhimurium* have demonstrated that other PDZ serine proteases participate in the in vivo stress response to host microbicidal pressures.[78–80] Bacteria with mutations in these genes were attenuated compared with wild-type parental strains, with decreased tissue burdens (up to a 10^5-fold decrease in one study).[80] LA_3455 encodes the *Leptospira* GlpF glycerol uptake facilitator protein. *L. interrogans* cannot use sugars as carbon sources, but instead, synthesizes sugars with de novo gluconeogenesis from glycerol.[27] Since the nonsynonymous S56P SNV identified in the P8 strain of our study may introduce a strain in the secondary structure of a conserved helix essential for function, *Leptospira* cells harboring this mutation could conceivably experience an impaired acquisition of glycerol in vivo that could have downstream biosynthetic consequences.

To the best of our knowledge, our generation of a list of computationally predicted ncRNAs in *L. interrogans* is the first in the field. Although none of our study-identified intergenic SNVs mapped to these regions, small noncoding RNAs have recently been shown to regulate pathogenic mechanisms in bacteria.[81–84] Thus, it would be important to test whether similar mechanisms exist in pathogenic *Leptospira.*

This study has limitations, primarily in that further work needs to be done both qualitatively and quantitatively to describe the individual contribution of the genes identified here to *Leptospira* pathogenesis. It is likely that the genes identified in this study may be part of a virulence-related transcriptional profile, and the increase in nsSNV alleles seen may collectively reduce the pathogenicity of *Leptospira* with these mutations. The ultimate mechanism of attenuation of *L. interrogans* serovar Lai in this study appears to be the additive effect of multiple mutant alleles, each subtracting from overall population fitness in vivo. The individual contributions of each of these genes to overall virulence is likely to remain hazy until more reliable methods of targeted mutagenesis are established for this important pathogen, until then attenuation-based studies are a reasonable alternative for identifying putative virulence-associated genes.

Received June 1, 2015. Accepted for publication September 8, 2015.

Published online December 28, 2015.

Note: Supplemental table appears at www.ajtmh.org.

Acknowledgments: We would like to thank David Haake (UCLA) for providing an isolate of *Leptospira interrogans* serovar Lai strain 55601.

Financial support: This work was supported by NIH grants T32 GM008666 (NIH predoctoral training grant), R25GM083275 ("Mentoring Young Minds to Increase Diversity in the Biomedical Research"), R21AI064466 ("Microarray Analysis of Leptospiral Genomes"), RO1TW05860 ("Leptospirosis Transmission in the Peruvian Amazon"), and 1D43TW007120 (Fogarty Global Infectious Diseases Training Grant, "Endemic Infectious Diseases of the Peruvian Amazon"). This project was also funded in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under contract number HHSN272200900007C.

Disclaimer: The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' addresses: Jason S. Lehmann, Joseph M. Vinetz, and Michael A. Matthias, Department of Medicine, Division of Infectious Diseases, University of Florida, Gainesville, FL, E-mails: Jason.Lehmann@medicine.ufl.edu, jvinetz@ucsd.edu, and mmatthias@ucsd.edu. Victoria C. Corey and Elizabeth A. Winzeler, Department of Pediatrics, School of Medicine, University of California, La Jolla, San Diego, CA, E-mails: vcorey@ucsd.edu and ewinzeler@ucsd.edu. Jessica N. Ricaldi, Department of Cellular and Molecular Sciences, Faculty of Sciences and Laboratory of Research and Development, Instituto de Medicine Tropical "Alexander von Humboldt," Universidad Peruana Cayetano Heredia, Lima, Peru, E-mail: jessica.ricaldi@upch.pe.

REFERENCES

1. Bharti AR, Nally JE, Ricaldi JN, Matthias MA, Diaz MM, Lovett MA, Levett PN, Gilman RH, Willig MR, Gotuzzo E, Vinetz JM, 2003. Leptospirosis: a zoonotic disease of global importance. *Lancet Infect Dis 3:* 757–771.
2. Abela-Ridder B, Sikkema R, Hartskeerl RA, 2010. Estimating the burden of human leptospirosis. *Int J Antimicrob Agents 36 (Suppl 1):* S5–S7.
3. Ko AI, Galvao Reis M, Ribeiro Dourado CM, Johnson WD Jr, Riley LW, 1999. Urban epidemic of severe leptospirosis in Brazil. Salvador Leptospirosis Study Group. *Lancet 354:* 820–825.
4. Ricaldi JN, Fouts DE, Selengut JD, Harkins DM, Patra KP, Moreno A, Lehmann JS, Purushe J, Sanka R, Torres M, Webster NJ, Vinetz JM, Matthias MA, 2012. Whole genome analysis of *Leptospira licerasiae* provides insight into leptospiral evolution and pathogenicity. *PLoS Negl Trop Dis 6:* e1853.
5. Matthias MA, Diaz MM, Campos KJ, Calderon M, Willig MR, Pacheco V, Gotuzzo E, Gilman RH, Vinetz JM, 2005. Diversity of bat-associated *Leptospira* in the Peruvian Amazon inferred by bayesian phylogenetic analysis of 16S ribosomal DNA sequences. *Am J Trop Med Hyg 73:* 964–974.
6. Brenner DJ, Kaufmann AF, Sulzer KR, Steigerwalt AG, Rogers FC, Weyant RS, 1999. Further determination of DNA relatedness between serogroups and serovars in the family Leptospiraceae with a proposal for *Leptospira alexanderi* sp. nov. and four new *Leptospira* genomospecies. *Int J Syst Bacteriol 49:* 839–858.
7. Bourhy P, Collet L, Brisse S, Picardeau M, 2014. *Leptospira mayottensis* sp. nov., a pathogenic *Leptospira* species isolated from humans. *Int J Syst Evol Microbiol 64:* 4061–4607.
8. Gouveia EL, Metcalfe J, de Carvalho AL, Aires TS, Villasboas-Bisneto JC, Queirroz A, Santos AC, Salgado K, Reis MG, Ko AI, 2008. Leptospirosis-associated severe pulmonary hemorrhagic syndrome, Salvador, Brazil. *Emerg Infect Dis 14:* 505–508.
9. Ko AI, Goarant C, Picardeau M, 2009. *Leptospira*: the dawn of the molecular genetics era for an emerging zoonotic pathogen. *Nat Rev Microbiol 7:* 736–747.
10. Bourhy P, Louvel H, Saint Girons I, Picardeau M, 2005. Random insertional mutagenesis of *Leptospira interrogans*, the agent of leptospirosis, using a mariner transposon. *J Bacteriol 187:* 3255–3258.
11. Murray GL, Morel V, Cerqueira GM, Croda J, Srikram A, Henry R, Ko AI, Dellagostin OA, Bulach DM, Sermswan RW, Adler B, Picardeau M, 2009. Genome-wide transposon mutagenesis in pathogenic *Leptospira* species. *Infect Immun 77:* 810–816.
12. Liao S, Sun A, Ojcius DM, Wu S, Zhao J, Yan J, 2009. Inactivation of the *fliY* gene encoding a flagellar motor switch protein attenuates mobility and virulence of *Leptospira interrogans* strain Lai. *BMC Microbiol 9:* 253.
13. Kassegne K, Hu W, Ojcius DM, Sun D, Ge Y, Zhao J, Yang XF, Li L, Yan J, 2014. Identification of collagenase as a critical virulence factor for invasiveness and transmission of pathogenic *Leptospira* species. *J Infect Dis 209:* 1105–1115.
14. Zhang L, Zhang C, Ojcius DM, Sun D, Zhao J, Lin X, Li L, Li L, Yan J, 2012. The mammalian cell entry (MCE) protein of pathogenic *Leptospira* species is responsible for RGD motif-dependent infection of cells and animals. *Mol Microbiol 83:* 1006–1023.
15. Lo M, Bulach DM, Powell DR, Haake DA, Matsunaga J, Paustian ML, Zuerner RL, Adler B, 2006. Effects of temperature on gene expression patterns in *Leptospira interrogans* serovar Lai as assessed by whole-genome microarrays. *Infect Immun 74:* 5848–5859.
16. Qin JH, Sheng YY, Zhang ZM, Shi YZ, He P, Hu BY, Yang Y, Liu SG, Zhao GP, Guo XK, 2006. Genome-wide transcriptional analysis of temperature shift in *L. interrogans* serovar lai strain 56601. *BMC Microbiol 6:* 51.
17. Patarakul K, Lo M, Adler B, 2010. Global transcriptomic response of *Leptospira interrogans* serovar Copenhageni upon exposure to serum. *BMC Microbiol 10:* 31.
18. Matsunaga J, Lo M, Bulach DM, Zuerner RL, Adler B, Haake DA, 2007. Response of *Leptospira interrogans* to physiologic osmolarity: relevance in signaling the environment-to-host transition. *Infect Immun 75:* 2864–2874.
19. Lo M, Murray GL, Khoo CA, Haake DA, Zuerner RL, Adler B, 2010. Transcriptional response of *Leptospira interrogans* to iron limitation and characterization of a PerR homolog. *Infect Immun 78:* 4850–4859.
20. Xue F, Dong H, Wu J, Wu Z, Hu W, Sun A, Troxell B, Yang XF, Yan J, 2010. Transcriptional responses of *Leptospira interrogans* to host innate immunity: significant changes in metabolism, oxygen tolerance, and outer membrane. *PLoS Negl Trop Dis 4:* e857.
21. Caimano MJ, Sivasankaran SK, Allard A, Hurley D, Hokamp K, Grassmann AA, Hinton JC, Nally JE, 2014. A model system for studying the transcriptomic and physiological changes associated with mammalian host-adaptation by *Leptospira interrogans* serovar Copenhageni. *PLoS Pathog 10:* e1004004.
22. Eshghi A, Becam J, Lambert A, Sismeiro O, Dillies MA, Jagla B, Wunder EA Jr, Ko AI, Coppee JY, Goarant C, Picardeau M, 2014. A putative regulatory genetic locus modulates virulence in the pathogen *Leptospira interrogans. Infect Immun 82:* 2542–2552.
23. Picardeau M, Bulach DM, Bouchier C, Zuerner RL, Zidane N, Wilson PJ, Creno S, Kuczek ES, Bommezzadri S, Davis JC, McGrath A, Johnson MJ, Boursaux-Eude C, Seemann T, Rouy Z, Coppel RL, Rood JI, Lajus A, Davies JK, Medigue C, Adler B, 2008. Genome sequence of the saprophyte *Leptospira biflexa* provides insights into the evolution of *Leptospira* and the pathogenesis of leptospirosis. *PLoS One 3:* e1607.
24. Adler B, Lo M, Seemann T, Murray GL, 2011. Pathogenesis of leptospirosis: the influence of genomics. *Vet Microbiol 153:* 73–81.
25. Nascimento AL, Ko AI, Martins EA, Monteiro-Vitorello CB, Ho PL, Haake DA, Verjovski-Almeida S, Hartskeerl RA, Marques MV, Oliveira MC, Menck CF, Leite LC, Carrer H, Coutinho LL, Degrave WM, Dellagostin OA, El-Dorry H, Ferro ES, Ferro MI, Furlan LR, Gamberini M, Giglioti EA, Goes-Neto A, Goldman GH, Goldman MH, Harakava R, Jeronimo SM, Junqueira-de-Azevedo IL, Kimura ET, Kuramae EE, Lemos EG, Lemos MV, Marino CL, Nunes LR, de Oliveira RC, Pereira GG, Reis MS, Schriefer A, Siqueira WJ, Sommer P, Tsai SM, Simpson AJ, Ferro JA, Camargo LE, Kitajima JP, Setubal JC, Van Sluys MA, 2004. Comparative genomics of two *Leptospira interrogans* serovars reveals novel insights into physiology and pathogenesis. *J Bacteriol 186:* 2164–2172.
26. Nascimento AL, Verjovski-Almeida S, Van Sluys MA, Monteiro-Vitorello CB, Camargo LE, Digiampietri LA, Harstkeerl RA, Ho PL, Marques MV, Oliveira MC, Setubal JC, Haake DA, Martins EA, 2004. Genome features of *Leptospira interrogans* serovar Copenhageni. *Braz J Med Biol Res 37:* 459–477.
27. Ren SX, Fu G, Jiang XG, Zeng R, Miao YG, Xu H, Zhang YX, Xiong H, Lu G, Lu LF, Jiang HQ, Jia J, Tu YF, Jiang JX, Gu WY, Zhang YQ, Cai Z, Sheng HH, Yin HF, Zhang Y, Zhu GF, Wan M, Huang HL, Qian Z, Wang SY, Ma W, Yao ZJ, Shen Y, Qiang BQ, Xia QC, Guo XK, Danchin A, Saint Girons I, Somerville RL, Wen YM, Shi MH, Chen Z, Xu JG, Zhao GP, 2003. Unique physiological and pathogenic features of *Leptospira interrogans* revealed by whole-genome sequencing. *Nature 422:* 888–893.
28. Lehmann JS, Fouts DE, Haft DH, Cannella AP, Ricaldi JN, Brinkac L, Harkins D, Durkin S, Sanka R, Sutton G, Moreno A, Vinetz JM, Matthias MA, 2013. Pathogenomic inference of virulence-associated genes in *Leptospira interrogans. PLoS Negl Trop Dis 7:* e2468.
29. Manary MJ, Singhakul SS, Flannery EL, Bopp SE, Corey VC, Bright AT, McNamara CW, Walker JR, Winzeler EA, 2014. Identification of pathogen genomic variants through an integrated pipeline. *BMC Bioinformatics 15:* 63.
30. Faine SAB, Bolin C, Perolat P, 1999. *Leptospira and leptospirosis*. Melbourne, Australia: MedScience.
31. Li H, Durbin R, 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics 25:* 1754–1760.

32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup, 2009. The sequence alignment/map format and SAMtools. *Bioinformatics 25:* 2078–2079.
33. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ, 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet 43:* 491–498.
34. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA, 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res 20:* 1297–1303.
35. Altenhoff AM, Dessimoz C, 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLOS Comput Biol 5:* e1000262.
36. Wolf YI, Koonin EV, 2012. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol Evol 4:* 1286–1294.
37. Schultz J, Milpetz F, Bork P, Ponting CP, 1998. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci USA 95:* 5857–5864.
38. Jones DT, 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol 292:* 195–202.
39. Buchan DW, Minneci F, Nugent TC, Bryson K, Jones DT, 2013. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res 41:* W349-57.
40. Capra JA, Singh M, 2007. Predicting functionally important residues from sequence conservation. *Bioinformatics 23:* 1875–1882.
41. Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D, 2011. Cactus: algorithms for genome multiple sequence alignment. *Genome Res 21:* 1512–1528.
42. Paten B, Diekhans M, Earl D, John JS, Ma J, Suh B, Haussler D, 2011. Cactus graphs for genome comparisons. *J Comput Biol 18:* 469–481.
43. Gruber AR, Findeiss S, Washietl S, Hofacker IL, Stadler PF, 2010. RNAz 2.0: improved noncoding RNA detection. *Pac Symp Biocomput 15:* 69–79.
44. Herbig A, Nieselt K, 2011. nocoRNAc: characterization of non-coding RNAs in prokaryotes. *BMC Bioinformatics 12:* 40.
45. Tatusov RL, Koonin EV, Lipman DJ, 1997. A genomic perspective on protein families. *Science 278:* 631–637.
46. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, Brinkman FS, 2010. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics 26:* 1608–1615.
47. Romling U, Galperin MY, Gomelsky M, 2013. Cyclic di-GMP: the first 25 years of a universal bacterial second messenger. *Microbiol Mol Biol Rev 77:* 1–52.
48. Ryan RP, 2013. Cyclic di-GMP signalling and the regulation of bacterial virulence. *Microbiology 159:* 1286–1297.
49. Chan C, Paul R, Samoray D, Amiot NC, Giese B, Jenal U, Schirmer T, 2004. Structural basis of activity and allosteric control of diguanylate cyclase. *Proc Natl Acad Sci USA 101:* 17084–17089.
50. Moglich A, Ayers RA, Moffat K, 2009. Structure and signaling mechanism of Per-ARNT-Sim domains. *Structure 17:* 1282–1294.
51. Singh N, Kuppili RR, Bose K, 2011. The structural basis of mode of activation and functional diversity: a case study with HtrA family of serine proteases. *Arch Biochem Biophys 516:* 85–96.
52. Maurel C, Reizer J, Schroeder JI, Chrispeels MJ, Saier MH Jr, 1994. Functional characterization of the *Escherichia coli* glycerol facilitator, GlpF, in Xenopus oocytes. *J Biol Chem 269:* 11869–11872.
53. Fu D, Libson A, Miercke LJ, Weitzman C, Nollert P, Krucinski J, Stroud RM, 2000. Structure of a glycerol-conducting channel and the basis for its selectivity. *Science 290:* 481–486.
54. Zhang D, de Souza RF, Anantharaman V, Iyer LM, Aravind L, 2012. Polymorphic toxin systems: comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. *Biol Direct 7:* 18.
55. Zhang D, Iyer LM, Aravind L, 2011. A novel immunity system for bacterial nucleic acid degrading toxins and its recruitment in various eukaryotic and DNA viral systems. *Nucleic Acids Res 39:* 4532–4552.
56. Setubal JC, Reis M, Matsunaga J, Haake DA, 2006. Lipoprotein computational prediction in spirochaetal genomes. *Microbiology 152:* 113–121.
57. Tenaillon O, Rodriguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS, 2012. The molecular diversity of adaptive convergence. *Science 335:* 457–461.
58. Lang GI, Rice DP, Hickman MJ, Sodergren E, Weinstock GM, Botstein D, Desai MM, 2013. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature 500:* 571–574.
59. Mazuz ML, Molad T, Fish L, Leibovitz B, Wolkomirsky R, Fleiderovitz L, Shkap V, 2012. Genetic diversity of *Babesia bovis* in virulent and attenuated strains. *Parasitology 139:* 317–323.
60. Bawden FC, 1958. Reversible changes in strains of tobacco mosaic virus from leguminous plants. *J Gen Microbiol 18:* 751–766.
61. Callow LL, Mellors LT, McGregor W, 1979. Reduction in virulence of *Babesia bovis* due to rapid passage in splenectomized cattle. *Int J Parasitol 9:* 333–338.
62. Wong MM, Karr SL Jr, Chow CK, 1977. Changes in the virulence of *Naegleria fowleri* maintained in vitro. *J Parasitol 63:* 872–878.
63. Ebert D, 1998. Experimental evolution of parasites. *Science 282:* 1432–1435.
64. Zhong Y, Chang X, Cao XJ, Zhang Y, Zheng H, Zhu Y, Cai C, Cui Z, Zhang Y, Li YY, Jiang XG, Zhao GP, Wang S, Li Y, Zeng R, Li X, Guo XK, 2011. Comparative proteogenomic analysis of the *Leptospira interrogans* virulence-attenuated strain IPAV against the pathogenic strain 56601. *Cell Res 21:* 1210–1229.
65. Galperin MY, Nikolskaya AN, Koonin EV, 2001. Novel domains of the prokaryotic two-component signal transduction systems. *FEMS Microbiol Lett 203:* 11–21.
66. Ausmees N, Mayer R, Weinhouse H, Volman G, Amikam D, Benziman M, Lindberg M, 2001. Genetic data indicate that proteins containing the GGDEF domain possess diguanylate cyclase activity. *FEMS Microbiol Lett 204:* 163–167.
67. Tischler AD, Camilli A, 2004. Cyclic diguanylate (c-di-GMP) regulates *Vibrio cholerae* biofilm formation. *Mol Microbiol 53:* 857–869.
68. Simm R, Morr M, Kader A, Nimtz M, Romling U, 2004. GGDEF and EAL domains inversely regulate cyclic di-GMP levels and transition from sessility to motility. *Mol Microbiol 53:* 1123–1134.
69. Tischler AD, Camilli A, 2005. Cyclic diguanylate regulates *Vibrio cholerae* virulence gene expression. *Infect Immun 73:* 5873–5882.
70. Ristow P, Bourhy P, Kerneis S, Schmitt C, Prevost MC, Lilenbaum W, Picardeau M, 2008. Biofilm formation by saprophytic and pathogenic leptospires. *Microbiology 154:* 1309–1317.
71. Brihuega B, Samartino L, Auteri C, Venzano A, Caimi K, 2012. In vivo cell aggregations of a recent swine biofilm-forming isolate of *Leptospira interrogans* strain from Argentina. *Rev Argent Microbiol 44:* 138–143.
72. He M, Zhang JJ, Ye M, Lou Y, Yang XF, 2014. Cyclic di-GMP receptor PlzA controls virulence gene expression through RpoS in *Borrelia burgdorferi. Infect Immun 82:* 445–452.
73. Bian J, Liu X, Cheng YQ, Li C, 2013. Inactivation of cyclic di-GMP binding protein TDE0214 affects the motility, biofilm formation, and virulence of *Treponema denticola. J Bacteriol 195:* 3897–3905.
74. Novak EA, Sultan SZ, Motaleb MA, 2014. The cyclic-di-GMP signaling pathway in the Lyme disease spirochete, *Borrelia burgdorferi. Front Cell Infect Microbiol 4:* 56.
75. Irie Y, Borlee BR, O'Connor JR, Hill PJ, Harwood CS, Wozniak DJ, Parsek MR, 2012. Self-produced exopolysaccharide is a signal that stimulates biofilm formation in *Pseudomonas aeruginosa. Proc Natl Acad Sci USA 109:* 20632–20636.
76. Kovacs-Simon A, Titball RW, Michell SL, 2011. Lipoproteins of bacterial pathogens. *Infect Immun 79:* 548–561.
77. Ristow P, Bourhy P, da Cruz McBride FW, Figueira CP, Huerre M, Ave P, Girons IS, Ko AI, Picardeau M, 2007. The OmpA-like protein Loa22 is essential for leptospiral virulence. *PLoS Pathog 3:* e97.

78. Chatfield SN, Strahan K, Pickard D, Charles IG, Hormaeche CE, Dougan G, 1992. Evaluation of *Salmonella typhimurium* strains harbouring defined mutations in *htrA* and *aroA* in the murine salmonellosis model. *Microb Pathog 12:* 145–151.
79. Sinha K, Mastroeni P, Harrison J, de Hormaeche RD, Hormaeche CE, 1997. *Salmonella typhimurium aroA, htrA*, and *aroD htrA* mutants cause progressive infections in athymic (nu/nu) BALB/c mice. *Infect Immun 65:* 1566–1569.
80. Johnson KS, Charles IG, Dougan G, Miller IA, Pickard D, O'Goara P, Costa G, Ali T, Hormaeche CE, 1990. The role of a stress-response protein in bacterial virulence. *Res Microbiol 141:* 823–825.
81. Toledo-Arana A, Repoila F, Cossart P, 2007. Small noncoding RNAs controlling pathogenesis. *Curr Opin Microbiol 10:* 182–188.
82. Papenfort K, Vogel J, 2010. Regulatory RNA in bacterial pathogens. *Cell Host Microbe 8:* 116–127.
83. Storz G, Opdyke JA, Zhang A, 2004. Controlling mRNA stability and translation with small, noncoding RNAs. *Curr Opin Microbiol 7:* 140–144.
84. Gong H, Vu GP, Bai Y, Chan E, Wu R, Yang E, Liu F, Lu S, 2011. A *Salmonella* small non-coding RNA facilitates bacterial invasion and intracellular replication by modulating the expression of virulence factors. *PLoS Pathog 7:* e1002120.