# Step I: White Paper Application

## Application Guidelines

1. *The application should be submitted electronically per requirements via the web site of any of the NIAID Genomic Sequencing Centers for Infectious Diseases. Include all attachments, if any, to the application.*
2. *There are no submission deadlines; white papers can be submitted at anytime.*
3. *GSC personnel at any of the three Centers can assist / guide you in preparing the white paper.*
4. *Investigators can expect to receive a response within 4-6 weeks after submission.*
5. *Upon approval of the white paper, the NIAID Project Officer will assign the project to a NIAID GSC to develop a management plan in conjunction with the participating scientists.*

## White Paper Application

**Project Title:** *Leptospira* Genomics and Human Health

**Authors: Joseph M. Vinetz, M.D. on behalf of the Leptospirosis Research Community as represented by the International Leptospirosis Society; Karen Nelson, Ph.D., J. Craig Venter Institute**

**Primary Investigator Contact:**

| | |
|---|---|
| Name | Joseph M. Vinetz, M.D. |
| Position | Professor of Medicine |
| Institution | University of California, San Diego |
| Address | 9500 Gilman Drive, La Jolla 0741 |
| State | California |
| ZIP Code | CA 92093-0741 |
| Telephone | 858-822-4469 |
| Fax | 858-822-5322 |
| E-Mail | jvinetz@ucsd.edu |

### 1. Executive Summary *(Please limit to 500 words.)*

*Provide an executive summary of the proposal.*

Leptospirosis, caused by more than 250 different serovars of the genus *Leptospira,* is the most common and widespread zoonotic disease worldwide. Infection is primarily spread through contact with water contaminated by urine of infected carrier animals. Leptospirosis is clearly an emerging and reemerging infectious disease. There are newly discovered leptospirosis-related syndromes (pulmonary hemorrhage) and newly discovered species of *Leptospira* that cause human leptospirosis. Accessible and geographically useful leptospirosis diagnostics remain unavailable to diagnose disease efficiently, which prevents the accurate assessment of the burden of disease. Obtaining whole genome sequences of a diverse and representative set of globally-significant *Leptospira* is a major priority of the leptospirosis community that will directly facilitate these goals of improving public health through the judicious and well-considered application of fundamental scientific discovery.

In this project, *Leptospira* strains for whole genome analysis will be chosen to address the following major scientific goals:

1. To obtain whole genome information for all known *Leptospira* species. Currently there are 9 known pathogenic *Leptospira* species*,* 5 intermediate *Leptospira* species, and 6 saprophytic *Leptospira* species. This information will provide the basis for identifying a minimal number of molecular markers for multilocus sequence typing that can differentiate infecting leptospires directly from human samples without the need for bacterial isolation. Conserved protein markers that are the targets of antibody recognition or antigen detection will be identified. Accomplishing these goals depends on obtaining whole genome sequence of a globally diverse and representative set of *Leptospira* strains.
2. To delineate taxonomic and phylogenetic relationships among *Leptospira* species. Current methods to classify new species or serovars, and to identify the emergence of new leptospiral causes of human disease, are cumbersome and insufficiently informative. Genomic-level information will allow us to determine serovar without the need for serological typing, will provide fresh insights into the utility of serovar as a tool for strain identification, and if shown to be robust will facilitate the development of molecular-based serovar typing. Complete genome sequence of reference strains used for serological diagnosis is critical for refining and optimizing efficient diagnosis.
3. To understand the mechanisms of leptospirosis pathogenesis and determinants of clinical outcome. Correlations of genetic polymorphisms and virulence will be identified between isolates at the species, serovar, and strain level. This will require sequencing of isolates associated with distinct clinical presentations and outcomes. This data will provide the fundamental basis for hypothesis-driven research to determine virulence factors and for vaccine development.

Summary of importance of comprehensive *Leptospira* whole genome sequencing to the research community

I. Global diversity: determine the sequence of geographically and molecularly diverse set of isolates that represent all current *Leptospira* spp. and as many serovars as possible. Outcomes:

   A. Will provide true phylogenetic picture based on global isolates and whole genome data. Will help understand taxonomy and possible global phylogeographic trends.
   B. Aid in developing optimal loci/primers for MLST so that more species can be included in MLST scheme and assay can be made more sensitive for direct detection and typing from clinical specimens.
   C. Aid in selection of appropriate targets for diagnostic assay and vaccine development.
   D. Identify genomic differences between pathogenic, intermediate, and non pathogenic-species

II. Regional diversity: determine the sequence of closely related isolates from select geographical regions. Outcomes:

   A. Allow development of high resolution genotyping assays that can be used for molecular epidemiology. Current methods such as MLST and MLVA frequently do not have the discriminatory power to investigate local outbreaks in regions where a single species/serovar predominate.
   B. Also further supports identification of virulence factors in Goal 1
   C. Also further supports identification of targets for diagnostic assay and vaccine development as in Goal 1.

III. Clinical Diversity: Determine sequence of *Leptospira* spp. isolated from cases with different clinical presentations and severity (i.e., mild vs. Weil's disease vs. Severe Pulmonary Hemorrhagic Syndrome). Outcomes:

   A. Identification of specific virulence factors or mutations associated with disease severity
   B. Also further supports identification of virulence factors in Goal 1
   C. Also further supports identification of targets for diagnostic assay and vaccine development as in Goal 1.

## 2. Justification

*Provide a succinct justification for the sequencing or genotyping study by describing the significance of the problem and providing other relevant background information.*

*This section is a key evaluation criterion.*

1. *State the relevance to infectious disease for the organism(s) to be studied; for example the public health significance, model system etc.*
2. *Are there genome data for organisms in the same phylum / class / family / genus? What is the status of other sequencing / genotyping projects on the same organism? Provide information on other characteristics (genome size, GC content, repetitive DNA, pre-existing arrays etc.) relevant to the proposed study. Have analyses been performed on the raw data already generated/published?*
3. *If analyses have been conducted, briefly describe utility of the new sequencing or genotyping information with an explanation of how the proposed study to generate additional data will advance diagnostics, therapeutics, epidemiology, vaccines, or basic*

*knowledge such as species diversity, evolution, virulence, etc. of the proposed organism to be studied.*

**1. Relevance of *Leptospira* strains to human disease and public health**

General Considerations of *Leptospira* and Leptospirosis

Leptospirosis is the most common and widespread zoonotic disease worldwide, caused by more than 250 different serovars of the genus *Leptospira* that are primarily spread through contact with water contaminated by urine of infected animals. Millions of people are estimated to be infected annually, but little is known regarding the true incidence of leptospirosis because of limitations of diagnosis. It is estimated that 0.1 to 1 per 100,000 people living in temperate climates are affected each year, and as many as 10 or more per 100,000 people living in developing countries. During epidemics, the incidence can rise to 100 or more per 100,000 people. In the US, 100-200 cases are reported annually with the approximately 50% of the cases occurring in Hawaii. The disease remains underreported for a variety of reasons, which include difficulty in distinguishing clinical signs from those of other endemic diseases and a lack of effective diagnostic tests and appropriate diagnostic laboratory services. Endemic foci have been well characterized in many developing countries worldwide particularly in rural areas and urban slum communities.

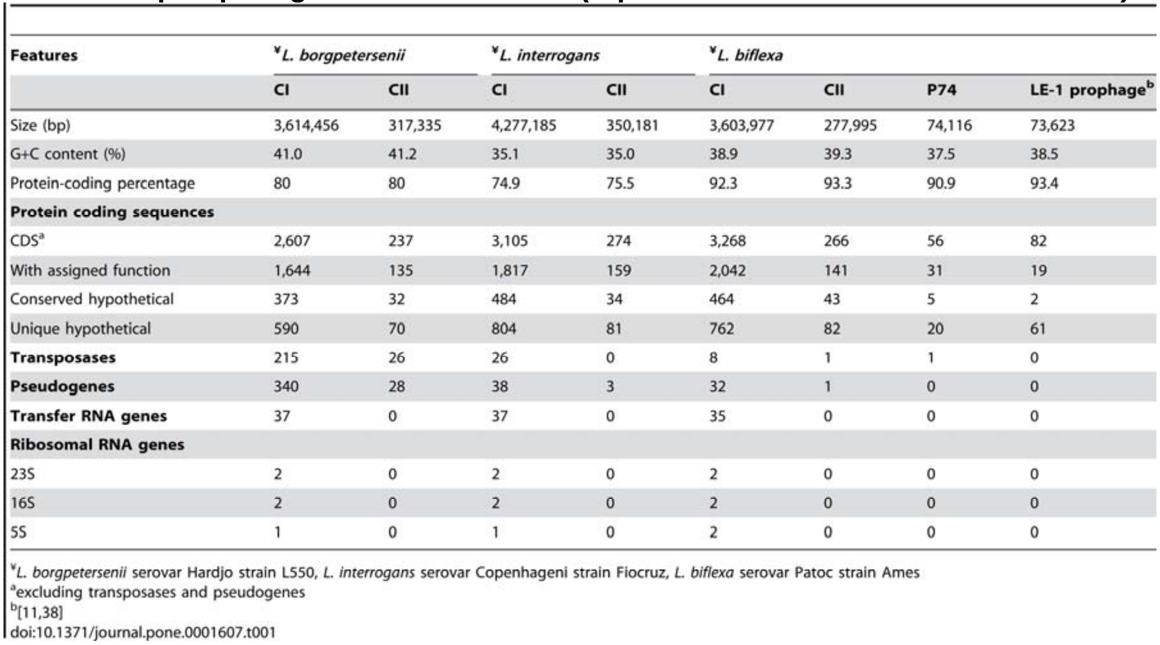
Leptospiral infections are common in tropical environments, in agricultural areas with high densities of livestock and rodents, and in areas supporting large and diverse wildlife populations where warm and humid conditions facilitate survival of the organism in the environment. Clinical manifestations range from asymptomatic infections recognizable only by seroconversion without illness, to severe, potentially fatal infections involving renal, liver and/or respiratory complications (including pulmonary hemorrhage) with case fatality rates as high as 25% in affected locales despite the most advanced medical care. The epidemiological pattern of leptospirosis has been changing in recent years in association with increasing urbanization and changes in the relationship between the environment and human activities. In industrialized countries, recreational exposures are being reported increasingly, particularly among persons participating in adventure tourism or water sports. For example, following a period of heavy rainfall, more than half of 98 triathletes participating in a triathlon in Springlfield, Illinois in 2000 contracted leptospirosis. A similar epidemic affected participants in the EcoChallenge Sabah in Borneo, Indonesia. There are countless other similar examples of epidemic leptospirosis in recent years. More generally, in Asia and Latin America, leptospirosis remains a current and significant public health problem. In Thailand, India, Philippines, Guyana, and Brazil, there are ongoing epidemics with often-fatal disease.

**2. Existing *Leptospira* genome and systems biology data**

Genome sequence information is now available for 6 leptospiral strains, including 2 serovars of the pathogenic species, *Leptospira interrogans* (Lai and Copenhageni), *Leptospira borgpetersenii* (serovar Hardjo, two different isolates), and 2 isolates (same serovar) of the saprophytic species *Leptospira biflexa* (Table 2). In summary, the genomes have a similar GC content ranging from 35% to 41% and possess two circular chromosomes of approximately 4 Mb and 300 kb. In *L. biflexa*, an additional 74 kb replicon has been identified. Plasmids have not been identified in any *Leptospira.* Genome-level gene expression analysis in response to environmental changes has been reported for *L. interrogans* (microarrays have been generated based on the serovar Lai genome but tested strains are usually other than Lai).

Collaborators on this genome sequencing proposal that have been involved in published genome sequencing and systems biology efforts include Matthieu Picardeau, Albert Ko, Ben Adler, David Haake, Rich Zuerner, Ana Nascimento, and Rudy Hartskeerl.

**Table 2: Leptospiral genome information (reproduced from Picardeau *et al*. 2008)**



| Features | [¥]*L. borgpetersenii* | | [¥]*L. interrogans* | | [¥]*L. biflexa* | | | |
|---|---|---|---|---|---|---|---|---|
| | **CI** | **CII** | **CI** | **CII** | **CI** | **CII** | **P74** | **LE-1 prophage**[b] |
| Size (bp) | 3,614,456 | 317,335 | 4,277,185 | 350,181 | 3,603,977 | 277,995 | 74,116 | 73,623 |
| G+C content (%) | 41.0 | 41.2 | 35.1 | 35.0 | 38.9 | 39.3 | 37.5 | 38.5 |
| Protein-coding percentage | 80 | 80 | 74.9 | 75.5 | 92.3 | 93.3 | 90.9 | 93.4 |
| **Protein coding sequences** | | | | | | | | |
| CDS[a] | 2,607 | 237 | 3,105 | 274 | 3,268 | 266 | 56 | 82 |
| With assigned function | 1,644 | 135 | 1,817 | 159 | 2,042 | 141 | 31 | 19 |
| Conserved hypothetical | 373 | 32 | 484 | 34 | 464 | 43 | 5 | 2 |
| Unique hypothetical | 590 | 70 | 804 | 81 | 762 | 82 | 20 | 61 |
| **Transposases** | 215 | 26 | 26 | 0 | 8 | 1 | 1 | 0 |
| **Pseudogenes** | 340 | 28 | 38 | 3 | 32 | 1 | 0 | 0 |
| **Transfer RNA genes** | 37 | 0 | 37 | 0 | 35 | 0 | 0 | 0 |
| **Ribosomal RNA genes** | | | | | | | | |
| 23S | 2 | 0 | 2 | 0 | 2 | 0 | 0 | 0 |
| 16S | 2 | 0 | 2 | 0 | 2 | 0 | 0 | 0 |
| 5S | 1 | 0 | 1 | 0 | 2 | 0 | 0 | 0 |

[¥] *L. borgpetersenii* serovar Hardjo strain L550, *L. interrogans* serovar Copenhageni strain Fiocruz, *L. biflexa* serovar Patoc strain Ames

[a] excluding transposases and pseudogenes

[b] [11,38]

doi:10.1371/journal.pone.0001607.t001

A comparative analysis of the sequenced leptospiral genomes has identified genetic features potentially responsible for the ability of *Leptospira* to survive either within the external environment or closely associated with renal tubular epithelium in reservoir hosts. Comparative genomic analysis suggests that *Leptospira* share a core set of 2,052 proteins. The *L. interrogans* and *L. borgpetersenii* genomes contain approximately 3,400 and 2,800 predicted coding regions, respectively, of which 656 are pathogen-specific and not found in the saprophyte *L. biflexa*. The function of the majority (59%) of these genes is unknown, suggesting the existence of pathogenic mechanisms unique to *Leptospira*. Unsurprisingly, the "free-living" leptospiral species *L. biflexa*, which survives exclusively in the external environment, has many more genes encoding environment-sensing and metabolic proteins than the pathogenic leptospires. While *L. interrogans* and *L. borgpetersenii* share 2,708 genes, there are 627 and 265 genes from *L. interrogans* and *L. borgpetersenii*, respectively, that are unique to either species. While *L. interrogans* has retained more genes from its free-living ancestor, most of which relate to survival in the external environment, *L. borgpetersenii* has a smaller genome (3.9 Mb compared with 4.6 Mb) and a much larger proportion of transposase genes or pseudogenes (20% compared with 2%) than *L. interrogans*. Together, these findings indicate that *L. borgpetersenii* is undergoing a process of genome reduction becoming a specialized parasite incapable of survival in the external environment. Moreover, comparative genomics may also have provided insight into leptospiral host tropism; for example *L. interrogans* serovar Lai strain 56601 appears to have undergone significant genome rearrangement when compared to *L. interrogans* serovar Copenhageni L1-130 and bears a genomic island which may account for its altered host specificity (Lai – *Apodemus* spp; Copenhageni – *Rattus* spp.).

A major goal of this application is to extend these comparisons to include a more representative collection of species (including the intermediately pathogenic species for which no genome sequences are currently available) to better understand leptospiral pathogenetic mechanisms and evolution.

### 3. Proposed analysis and utility of new *Leptospira* genome information

Previous genome sequencing has provided critical insights into genome composition of two pathogenic species of *Leptospira* and one saprophytic species. However, the pathogenic species that have been sequenced represent a tiny proportion of the *Leptospira* that cause human disease. This proposed study will add substantially to this previous work. Anticipated outcomes include the following:

1) Comprehensive genomic information for all currently recognized species of *Leptospira*. There are 9 known pathogenic *Leptospira* species*,* 5 intermediate *Leptospira* species, and 6 saprophytic *Leptospira* species. This will provide basic knowledge of the taxonomic and evolutionary relationships among all clades of *Leptospira.* Current methods to classify new species, serovars and to identify the emergence of new leptospiral causes of human disease are cumbersome and insufficiently informative. Genomic-level information will allow us to determine serovar without the need for serological typing, which can be complicated and time-consuming. will provide fresh insights into the utility of serovar as a tool for strain identification, and if shown to be robust will facilitate the development of molecular-based serovar typing akin to the schemes developed recently for molecular *S. pneumoniae* serotyping. Comparative systems biology analysis will allow us to define what makes a leptospire pathogenic.

2). Comparative genomic analysis of all *Leptospira* species that infect humans will allow for the identification of a minimal number of molecular markers for multilocus sequence typing that can differentiate infecting leptospires directly from human samples without the need for bacterial isolation.

3) Identification of conserved genes in pathogenic spp. which encode proteins that may be targets of antibody recognition or can be used for antigen detection for diagnostics, or for vaccine development.

4) To understand the mechanisms of leptospirosis pathogenesis and determinants of clinical outcome. Correlations of genetic polymorhphims and virulence will be identified between isolates at the species, serovar, and strain level. This will require sequencing of isolates associated with distinct clinical presentations and outcomes. This data will provide the fundamental basis for hypothesis-driven research to determine virulence factors and for future vaccine development..

In summary, the new sequencing information obtained from this project will advance all aspects of the leptospirosis field: therapeutics, epidemiology, vaccines, and basic knowledge including species diversity, evolution, and virulence.

## 3. Rationale for Strain Selection

Leptospirosis as an Emerging Infectious Disease

*Leptospira* were first discovered as an infectious cause of fever, jaundice and renal failure early in the 20th century but only in the past 15 years or so has leptospirosis emerged not only as the most common bacterial zoonosis in the world but as a globally significant cause of epidemic mortality. The landmark Institute of Medicine 1992 report, “Emerging Infections: Microbial Threats to Health in the United States” documented the importance of leptospirosis as a threat to military personnel worldwide. In the 1990s, the new syndrome of pulmonary hemorrhage due to leptospirosis was discovered as the cause of hemorrhagic fever in the Andaman Islands of India. An epidemic of fever and pulmonary hemorrhage affecting rural Nicaragua after flooding mystified the world until pathologists at the CDC found *Leptospira* in kidney tissues at autopsy. Simultaneously, rat-transmitted leptospirosis was rediscovered as the cause of severe febrile illness including jaundice, renal failure, refractory shock and pulmonary hemorrhage in Baltimore, Maryland, and a large (and currently ongoing) urban epidemic of severe leptospirosis in Salvador, Brazil, was reported. Leptospirosis is the leading cause of admission to intensive care units in urban coastal Brazil during the rainy reason. Around the world, epidemic leptospirosis predictably occurs annually largely in the context of hurricanes and flooding. Even at the writing of this white paper, there is an ongoing epidemic of leptospirosis in the Philippines: in the first half of October, 2009, a total of 1887 leptospirosis cases were reported in 15 hospitals in metropolitan Manila, with

138 deaths. Remarkably, the Philippines Ministry of Health has provided prophylactic anti-leptospirosis antibiotics to more than 1 million people with the goal of reaching 4 million. More generally, it is clear that increasing urbanization and anthropogenic environmental changes clearly are driving the reemergence of leptospirosis as a significant public health threat.

In addition to the emergence of new syndromes of leptospirosis, new species of *Leptospira* causing significant human disease have recently been discovered and reported, for example new pathogenic *Leptospira* and the intermediately pathogenic *Leptospira L. fainei, L. broomii,* and *L. licerasiae*

The global health burden of leptospirosis is directly related to its emergence/ reemergence. *Leptospira* are highly diverse both regionally and globally, which has limited the development of readily available and geographically-relevant diagnostics. The lack of sufficient fundamental knowledge of the wide variety of *Leptospira* and mechanisms of molecular targets of anti-leptospiral immunity has limited vaccine development. Based on these considerations, the Director General of the World Health Organization has constituted the Leptospirosis Epidemiology Reference Group (LERG), which is comprised of members from around the world including several contributors to this White Paper.

In summary, the designation of leptospirosis as both an emerging and reemerging infectious disease is well-justified. There are newly discovered leptospirosis-related syndromes (pulmonary hemorrhage) and newly discovered species of *Leptospira* that cause human leptospirosis. Yet accessible and geographically useful leptospirosis diagnostics remain unavailable to diagnose disease efficiently, which prevents the accurate assessment of the burden of disease. Obtaining whole genome sequence of a diverse and representative set of globally-significant *Leptospira* is a major priority of the leptospirosis community that will directly facilitate these goals of improving public health through the judicious and well-considered application of fundamental scientific discovery.

<u>*Leptospira* Taxonomy</u>

*Leptospira*, the only free-living pathogenic spirochete known to cause human disease, is comprised of three major sub-groups comprising 20 named species based on classical DNA-DNA hybridization studies [2] and 16S ribosomal RNA gene phylogeny [3] (Fig 1). Saprophytic leptospires (6 species, see Table 1) are free-living and incapable of infecting mammals, while pathogenic *Leptospira*, which represent the largest and most diverse subgroup (9 species), cause disease of varied severity in both humans and animals. Intermediately pathogenic *Leptospira* comprising 5 genomospecies (Table 1) share genetic and growth characteristics of both the saprophytic and pathogenic subgroups and are thought to be a transitional evolutionary group based upon16S rDNA phylogeny. Currently, there are more than 200 recognized pathogenic antigenic variants (termed serovars) within the nine described species. This antigenic diversity, thought to be largely due to the polymorphic O-polysaccharide antigen of leptospiral LPS, may reflect leptospiral host tropism.

Although leptospiral evolutionary mechanisms are unknown, in light of the correlation between leptospiral and host biodiversity, it is likely that host adaptation is the driving force of leptospiral diversification and/or evolution. Longevity and transmissibility of the organisms within host populations may lead to adaptations, particularly among surface exposed molecules involved in attachment to host renal tubular cells and/or evasion of host immune mechanisms. Selection pressures driving the adaptation of the bacteria to new hosts (extending host-range) may therefore lead to the emergence of novel leptospiral variants.

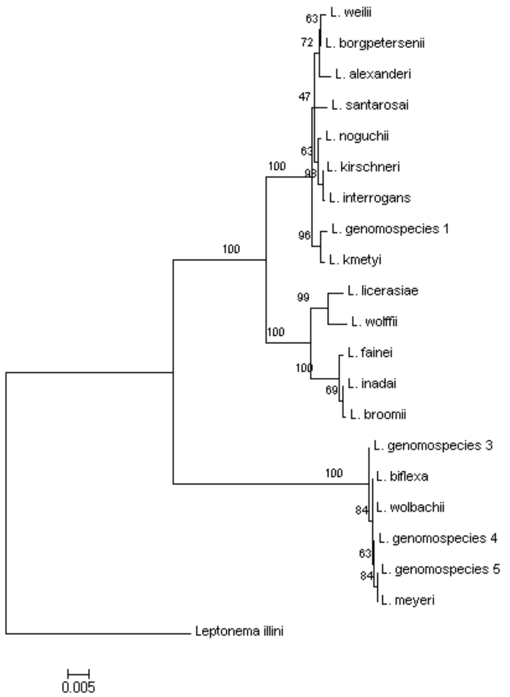


**Figure 1: Dendrogram showing the evolutionary relationships of the 20 characterized leptospiral species**. *NB* L. genomospecies 1, 3, 4, 5 have been recently designated *L. alstonii*, *L. terpstrae*, *L. vanthielii* and *L. yanagawae*, respectively.

Table 1: Leptospiral genomospecies and type strains

| | Species | Serovar | Type strain |
|---|---|---|---|
| Pathogenic (9) | *L. alexanderi* | Manhoa 3 | L 60$^{T}$ |
| | *L. alstonii* | Pingchang | 80-412$^{T}$ |
| | *L. borgpetersenii* | Javanica | Veldrat Batavia 46$^{T}$ |
| | *L. interrogans* | Icterohaemorrhagiae | RGA$^{T}$ |
| | *L. kirschneri* | Cynopteri | 3522 C$^{T}$ |
| | *L. kmetyi* | Undesignated | Bejo-Iso9$^{T}$ |
| | *L. noguchii* | Panama | CZ 214$^{T}$ |
| | *L. santarosai* | Shermani | 1342 K$^{T}$ |
| | *L. weilii* | Ranarum | ICF$^{T}$ |
| Intermediately Pathogenic (5) | *L. broomii* | Undesignated | 5399$^{T}$ |
| | *L. faineii* | Hurstbridge | BUT 6$^{T}$ |
| | *L. inadai* | Lyme | 10$^{T}$ |
| | *L. licerasiae* | Varillal | Var 010$^{T}$ |
| | *L. wolffii* | Undesignated | Korat-H2$^{T}$ |
| Saprophytic (6) | *L. bifelxa* | Patoc | Patoc$^{T}$ |
| | *L. meyeri* | Semaranga | Veldrat Semarang 173$^{T}$ |
| | *L. terpstrae* | Undesignated | H 2$^{T}$ |
| | *L. vanthielii* | Holland | WaZ Holland$^{T}$ |
| | *L. wolbachii* | Codice | CDC$^{T}$ |
| | *L. yanagawae* | Saopaulo | Sao Paulo$^{T}$ |

Multilocus Sequence Typing (MLST) for *Leptospira*

MLST was applied to the microbiological identification of *Leptospira* isolated from humans in a project led by Professor Sharon Peacock in collaboration with Thai collegues at Mahidol University in Bangkok and with Lee Smythe and colleagues at the WHO/FAO/OIE Collaborating Centre for Reference & Research on Leptospirosis, Centre for Public Health Sciences, Queensland Health Scientific Services, Brisbane, Australia. The *Leptospira* MLST database (http://leptospira.mlst.net/misc/info.asp) currently contains information on more than 200 isolates but is limited to *L.interrogans* and *L. kirschneri* because the current primer sets do not reliably amplify other *Leptospira* species. Of these, around two thirds were isolated from patients in Thailand between 2000 and 2005; about one third are reference strains from the collection maintained by the WHO/FAO/OIE Collaborating Center for Reference & Research on Leptospirosis, Australia.

Currently, seven loci used to carry out the PCR-based *Leptospira* MLST scheme for *L.interrogans* and *L. kirschneri* use internal fragments of the following seven house-keeping genes: 1) *glmU* (UDP-N-acetylglucosamine pyrophosphorylase); 2) *pntA* (NAD(P) trans-hydrogenase subunit alpha); 3) *sucA* (2-oxoglutarate dehydrogenase decarboxylase component); 4) *fadD* (probable long-chain-fatty-acid--CoA ligase); 5) *tpiA* (yriosephosphate isomerase); 6) *pfkB* (ribokinase); 7) *mreA* (rod shape-determining protein rodA). Another MLST scheme is available (Ahmed et al., 2006) and can be used on a wide panel of pathogenic species (most but not all alleles can be amplified from pathogenic strains).

Thus a major goal of this genome project is to obtain sufficient information to be able to differentiate and detect *Leptospira* from geographically distinct regions where *Leptospira* are highly divergent.

Role of Strain-Specific Factors in Leptosirosis-Associated Pulmonary Hemorrhage Syndrome (LPHS)

**The reason for the global emergence of LPHS is a major unanswered question facing the field**. Through support from NIH, the Oswaldo Cruz Foundation (Fiocruz) and Weill Medical College of Cornell University have conducted active population-based surveillance for leptospirosis since 1996 in the city of Salvador, Brazil. In this epidemiological setting, transmission of a single serovar agent, Copenhageni, causes annual rainfall-associated epidemics in slum communities. In 2003, The Fiocruz-Cornell team identified the sudden and sustained appearance of LPHS. Case fatality was 52% among 169 LPHS cases identified to date. This phenomenon was not associated with changes in case ascertainment, climate, risk behaviors or underlying host conditions, leading to the working hypothesize that the emergence of LPHS may have been due to the introduction of a strain with enhanced virulence. Roche 454 and Illumina-based sequencing conducted at the CDC found that the genomes of three patient isolates, including one obtained from a LPHS case, demonstrated a high degree of sequence conservation. Less than 150 SNPs were identified which occurred in ORFs. However, SNPs clustered in ORFs with putative pathogenesis-associated function.

We propose to perform genome sequencing of the Fiocruz-Cornell strain collection to identify genetic polymorphisms which are associated with the development of LPHS and severe outcome from leptospirosis. At present, the collection has 85 patient isolates with well-characterized clinical phenotypes, of which 17 were obtained from LPHS cases. As a caveat, the sample size of isolates in this white paper proposal may not be sufficient to identify significant associations between individual SNPs and clinical outcomes. However genome sequencing will characterize the range of SNPs and enable development of refined detection methods which can be used to evaluate the study hypotheses prospectively at the site in Salvador and in other geographical regions where LPHS is an emerging problem. Furthermore, identification of such polymorphisms will directly aid efforts to perform targeted mutagenesis approaches aimed at characterizing *Leptospira* virulence factors.

**Summary of Rationale for Selection of *Leptospira* Strains for Sequencing**

The guiding principles for selecting *Leptospira* strains for whole genome analysis are as follows:

1. To obtain whole genome information for all known *Leptospira* species. Currently there are 9 known pathogenic *Leptospira* species*,* 5 intermediate *Leptospira* species, and 6 saprophytic *Leptospira* species.
2. To delineate and quantify taxonomic and evolutionary relationships among *Leptospira*, and to define the fundamental set of complete biological parameters that characterize a pathogenic *Leptospira*
3. Understand the genomic-level basis for serovar differences within species. Serovars are defined by antibody reactivity and do not necessarily associate with a single species, suggesting the possibility of putative lateral gene transfer between and within species. There are multiple loci involved in determining serovar (LPS biosynthetic loci and others).
4. Understanding the relationship of clinical pathogenesis to genetic polymorphisms within identified species, serovars, and strains.
5. Identify a minimal number of molecular markers for MLST typing that can differentiate infecting leptospires directly from human samples without the need for bacterial isolation. This will require a globally diverse and representative set of strains to be sequenced.

## 4a. Approach to Data Production: Data Generation

5. *State the data and resources planned to be generated. (e.g draft genome sequences, finished sequence data, SNPs, DNA/protein arrays generation, clone generation etc.)*

We plan to generate nearly-finished genomes for the novel isolates, draft genomes for different strains of similar serovars, and SNP discovery for strains within serovars. It is prohibitively expensive to propose formal closure of all genomes; however, current technologies using paired-end sequencing, will yield, even with automated analyses, very nearly complete genomes that members of the leptospirosis research could close on their own given their own scientific criteria. The comprehensive genome data will be used to construct phylogenetic trees based on whole genomic information using synteny plots. At least two sequencing runs will be performed for each strain, which will produce nucleotide sequences over 30-fold coverage. Adjacent contigs are subsequently linked by additional nucleotide sequencing of PCR-generated DNA fragments using paired end sequencing. Long-range PCR will be used to close any remaining gaps. This analysis will provide far better resolution and sufficient understanding of the evolutionary relationships among *Leptospira* than 16S rRNA analysis or the archaic DNA-DNA hybridization. Genomes will be annotated by powerful software available at JCVI, which also will facilitate sophisticated comparative genome analysis. Gene calling will be done using an automated pipeline customized for *Leptospira* based on the extant genome analyses. BLAST searching will be carried out against extant annotated *Leptospira* genomes and in a reciprocal manner among the new genome sequence. Pseudogenes and polymorphisms (insertions, deletions, SNPs) will be identified. Strains will also be grouped on the basis of important phenotypes (clinical history, virulence phenotype, host response phenotype) and genetic differences will be analyzed to identify putative genotypes that underly phenotypes. Additional custom analyses to be carried out by the JCVI on all of the new genomic information will include 1) lipoprotein prediction based on recent algorithms developed in the Haake laboratory; 2) identification of a non-redundant set of potential diagnostic antigens to be fabricated into diagnostic protein microarrays; and 3) serovar prediction based on predicted LPS biosynthetic loci.

Closed and draft genome sequences will be generated for a comprehensive list of *Leptospira* species, serovars and strains included in the attached table. Using paired-end sequencing and previously obtained scaffolds, we anticipate that next generation sequencing technologies will be able to come very close to closing most or all of the *Leptospira* sequences.

Deep sequencing to assess diversity within the listed *Leptospira* isolates will be obtained.

Furthermore we will sequence 90 serovar Copenhageni isolates from Brazil to evaluate the association between genetic polymorphisms and clinical outcome. During the organization and management plan phase, we will work with JCVI to: 1) define an initial sample of isolates for which whole genome sequencing would be performed; 2) identify the sequencing approach which would be most efficient for identifying SNPs for this specific research question, and 3) determine whether whole genome sequencing or a focused SNP detection assay would be used to characterize genomes for the rest of the isolates in the serovar Copenhageni collection.

All genome information will be deposited by the JCVI to GenBank and the NIAID-supported Bioinformatics Resource Center. Sequenced and anbalyzed *Leptospira* strains will be deposited to the Biodefense and Emerging Infections Research Resources Repository at the American Type Culture Collection, supported by the National Institute of Allergy and Infectious Diseases (NIAID).

Bacterial growth and DNA extraction

DNA for all strains to be sequenced will be provided by the collaborators in this project to the JCVI. Prior to growing up, the identity of each strain will be verified by serological and molecular typing and these data, as well as metadata associated with each strain will be informatically attached to each specimen for depositing to GenBank and the BRC. Bacteria will be grown in appropriate medium (EMJH or other), genomic DNA extracted using TriReagent (Invitrogen) or equivalent and quantified using microfluorimetry. 5-10 micrograms of extracted DNA will sent to JCVI on dry ice. JCVI will further confirm the identity of each strain by 16S rDNA sequencing prior to genome sequencing.

Genome Data Acquisition

The complete genome of the *Leptospira* strains will be sequenced by the JCVI NIAID Genomic Sequencing Center using next-generation DNA sequencing approaches.

## 4b. Approach to Data Production: Data Analysis

6. *Briefly describe the analysis (value-add) envisioned to be performed subsequently by the community and the potential to develop hypotheses driven proposals given the datasets and resources produced by this work.*

After automated sequence annotation is carried out by the JCVI, the leptospirosis research community, represented by the collaborators involved here and others wishing to participate, will be following up in the following ways:

1. Manual curation of genome sequencing and collaboration with JCVI in preparing large scale genome analysis papers including primary analysis, comparative genomics
2. Development of more comprehensive typing schemes (MLST) based on comparative genomic analysis and inclusion of common genetic targets
3. Development of diagnostic antigen discovery based on non-redundant arraying of conserved protein antigens shared across the major leptospiral strains
4. Hypothesis-driven pathogenesis experiments including novel approaches to gene targeting and heterologous gene expression
5. Hypothesis-driven vaccine discovery experimentation
6. Systems biology analysis of complete metabolic functions of leptospires based on comparative genomics; define what makes a leptospire pathogenic
7. Identification of genetic polymorphisms which influence LPHS and disease outcome, which in turn may identify new candidates for virulence factors.

## 5. Community Support and Collaborator Roles:

7. *Provide evidence of the relevant scientific community's size and depth of interest in the proposed sequencing or genotyping data for this organism or group of organisms.*

8. *List all project collaborators and their roles in the project*

9. *List availability of other funding sources for the project.*

**7. Leptospirosis scientific community.**

The leptospirosis community is represented internationally by the International Leptospirosis Society, which convenes biannual meetings that alternate between East and West. Collaborators on this project represent many of the leaders of the ILS. For the past two meetings, the NIH has supported the International Leptospirosis Meetings in Quito, Ecuador and Cochin, India with R13 conference grants 1R13AI075971-01 and 1R13AI084386-01 , respectively. These R13 grants were put together with the active collaboration of the leaders in the field and represented a community effort. Each ILS meeting attracts ~150-200 registered participants. The ILS meetings of 2007 and 2009 included important presentations of genomic and systems biology-level data on *Leptospira* and leptospirosis.

**8.** Project collaborators and role on project

1) University of California, San Diego: Joseph Vinetz, M.D., Michael A. Matthias, PhD., Bernhard Palsson, PhD., Pep Charusanti , Ph.D., Harish Nagarajan. Provision of leptospiral strains from Peru; bioinformatic and systems biology analysis.

2) University of Cambridge, UK, Sanger Center, Cambridge; Mahidol University, Bangkok, Thailand: Sharon Peacock, M.D., Ph.D., Stephen Bentley, Ph.D., Janjira Thaipadungpanit, Ph.D. Provision of leptospiral strains from Thailand, Laos, Sri Lanka; bioinformatics analysis, development and refinement of MLST methods.

3) Weill Medical College of Cornell University, New York, and Centro de Pesquisas Gonçalo Moniz, Fundação Oswaldo Cruz/MS, Salvador, Brazil: Albert I. Ko, M.D., Mitermayer Galvão Reis, M.D., Ph.D., Paula Ristow, D.V.M., Ph.D., Daniele Takahashi, Ph.D., Elsio Wunder, D.V.M.: Provision of leptospiral strains from Brazil; bioinformatics, manual curation, identification of genetic polymorphisms associated with pulmonary haemorrhage syndrome and other clinical phenotypes.

4) Leptospirosis Reference and Research,WHO/FAO/OIE Collaborating Centre, Queensland, Australia: Lee Smythe, Ph.D., Scott Craig, Ph.D. Provision of leptospiral strains from Australia, Asia, reference strains; bioinformatics analysis.

5) Centers for Disease Control, Atlanta, GA: Alex Hoffmaster, Ph.D., Renee Galloway, M.P.H., Robyn Stoddard, DVM, Ph.D. Provision of leptospiral strains, bioinformatics and diagnostic and typing development.

6) WHO/FAO/OIE and National Collaborating Centre for Reference and Research on Leptospirosis, Amsterdam, the Netherlands: Rudy Hartskeerl, Ph.D. Provision of leptospiral strains from Europe, India, Africa, Indonesia, bioinformatics analysis and manual curation.

7) Institut Pasteur, "Biology of Spirochetes" Unit, National Reference Center for

Leptospira, WHO/FAO Collaborating Centre: Mathieu Picardeau, Ph.D. Provision of leptospiral strains and bioinformatic analysis (SpiroScope database, http://www.genoscope.cns.fr/agc/mage); in collaboration with Claudine Médigue Genoscope, Laboratoire de Génomique Comparative, Evry, France

8) University of California, Los Angeles: David Haake, M.D.
Provision of leptospiral strains, bioinformatics analysis and manual curation.

9) Monash University, Melbourne, Australia: Ben Adler, Ph.D. Dieter Bulach, Gerald Murray, Ph.D. and Miranda Lo, Ph.D. Provision of leptospiral strains, bioinformatics analysis and manual curation.

10) Instituto Butantan, Sao Paulo, Brazil: Ana Nascimento, Ph.D. , Gustavo Cerqueira , Ph.D., Bioinformatics analysis, manual curation

11) University College Dublin: Jarlath Nally, Ph.D. Manual curation, downstream proteomics analysis utilizing new genome information

12) Agricultural Research Service, United States Department of Agriculture: Richard Zuerner, Ph.D., Bioinformatics analysis, manual curation

13) Pasteur Institute of Montevideo, Montivideo, Uruguay: Alejandro Buschiazzo, Ph.D., Hugo Naya, Ph.D. Manual curation, bioinformatic analysis and identification of pathogenesis-related genes.

**9.** F**unding for studies of leptospirosis research by the collaborators includes the following:**

NIH grants to Vinetz, Matthias, Palsson: 1D43TW007120 , 1R21AI067745, 1K24AI068903, 1R01AI067727, 1U01AI075420, 2R01GM068837, 5R01GM057089, 5R01GM062791

NIH grants to Ko: 2R01AI052473, 5D43TW000919, 2 R44 AI072856

NIH and VA grants to Haake: 5R01AI034431, 1I01BX000119

M. Picardeau, coordinator of a "Maladies Infectieuses et leur environnement" program from Agence Nationale de la Recherche, « *Leptospira*: from genetics to pathogenesis».

Ben Adler, Australian Research Council

## 6. Availability & Information of Strains:

10. Indicate availability of relevant laboratory strains and clinical isolates. Are the strains/isolates of interest retrospectively collected, prepared and ready to ship?

Please see the availability of the relevant strains in the attached document.

All isolates listed are present in the identified collaborator's collection. On approval of this project, the strains will be grown up and DNA prepared for shipment and sequencing.

What supporting metadata and clinical data have been collected or are planned on being collected that could be made available for community use?

Comprehensive clinical data and results of host response studies on patients and strains as described in the previous sections are available for the human isolates and will be made available to the scientific community as part of the metadata for the project. For animal strains (relevant to human health because of the zoonotic nature of the disease), species and conditions of obtaining isolates will be made available for each isolate. For all strains, serovar testing results will be made available as part of the metadata.

IRB approval and regulatory requirements including shipping

All human isolates to be studied in this project were obtained after appropriate Human Subjects Protocol review.

International shipment of strains out of endemic countries was done with national approvals where legally required.

Shipment of live Leptospira strains into the United States was approved by the CDC and USDA under approved import permits for already received isolates and will be for any new ones.

Metadata management, progress of work, timeline and prioritization of isolates for sequencing

Given the large number of isolates for analysis, the Vinetz laboratory at UC San Diego will be responsible for managing the metadata associated with each isolate provided to the JCVI. Metadata will be attached to each DNA sample and will be published in the appropriate databases with each genome sequence.

Whole genome analysis is anticipated to proceed at the rate of 8 strains per week (at 2-fold sequencing per strain). At this rate, raw sequence data from 200 strains can be obtained in 6 months. It is estimated that 5-10 ug of DNA per strain can be obtained in 4 weeks for typical slow-growing pathogenic Leptospira strains, and in 2 weeks from intermediate and saprophytic Leptospira.

With approval, DNA from strains at the rate of 8 per week can be available as soon as January 2010, which should allow completion before the end of FY2010.

Prioritization:

1. Reference strains representing known leptospiral species (pathogenic, intermediate, saprophytic), including geographic representation.

2. New Leptospira species as identified by 16S rDNA sequencing and novel PFGE pattern.

3. New serovars within known species as identified by 16S rDNA sequencing and novel PFGE pattern.

4. Strains with known species and serovar with new/novel alleles as identified by multi-

locus sequence typing.

5. Variants of Leptospira within known leptospiral species, serovars, and MLST types that have implications for pathogenetic and virulence differences.

Notes on the approaches for the Leptospirosis Community to analyze and annotate sequence data and prepare manuscripts for publication.

We anticipate that a major outcome of this proposal will be one or a few papers that will report a global genome analysis based on analyses performed by white paper collaborators and JCVI. Given the large scope of the comparative analysis (perhaps unprecedented in bacterial genomics), the organization and implementation of the research will be done via scheduled teleconferences between white paper collaborators and JCVI scientists. The group of collaborators will have access to and interact with JCVI to contribute to development of the automatic annotation protocol (for example, adding lipobox prediction algorithms to the pipeline, as developed by the Haake laboratory) and large scale computational analyses that will be performed as part of the annotation step. Manual curation of genomic information will be done primarily by the collaborators. Beyond the scope of comparative genome analysis, scientists within this group as well as in the general scientific community will have specific research questions which they would like to address, some of which are described in the white paper.

## 7. Compliance Requirements:

### 7a. Review NIAID's Reagent, Data & Software Release Policy:

*http://www3.niaid.nih.gov/research/resources/mscs/data.htm*
*http://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-013.html*

Accept ☒ Decline □

### 7b. Public Access to Reagents, Data, Software and Other Materials:

State plans for deposit of starting materials as well as resulting reagents, resources, and datasets in NIAID approved repositories.

Sequence data along with associated metadata will be submitted to Genbank and to the NIAID-supported Bioinformatics Resource Center as they become available per GSC policy. The metadata will be coordinated with the genomic data by the Vinetz laboratory at UC San Diego.

All *Leptospira* isolates will be deposited with the ATCC via the NIAID-supported Biodefense and Emerging Infections Research Resources Repository (BEI Resources). Alternatively, if strains cannot be approved by the U.S. Department of Agriculture for immediate importation because of the presence of bovine serum albumin in the culture medium, strains can be deposited into the Australian and/or Amsterdam WHO Reference Center with the explicit understanding that the strains will be passaged into USDA-approved *Leptospira* culture medium and then later deposited into the BEI Resources Repository. This arrangement has been agreed to by the Reference Center Directors.

**7c. Research Compliance Requirements**

*Upon project approval, NIAID review of relevant IRB/IACUC documentation is required prior to commencement of work. Please contact the GSC Principal Investigator(s) to ensure necessary documentation are filed for / made available for timely start of the project.*

**Investigator Signature:**

Joseph M. Vinetz

**Investigator Name: Joseph M. Vinetz, M.D.** **Date: 11/5/09**

Blank Last Page