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: Deep sequencing diverse lineages of *Toxoplasma gondii*.

Authors: August 23, 2009

Primary Investigator Contact:

Name	L. David Sibley
Position	Professor
Institution	Washington University School of Medicine
Address	660 S. Euclid Ave., St. Louis
State	MO
ZIP Code	63110
Telephone	314-362-8873
Fax	314-286-0060
E-Mail	sibley@borcim.wustl.edu

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

Provide an executive summary of the proposal.

Toxoplasma gondii is a wide spread protozoan parasite of animals and an important opportunistic pathogen in humans where it causes disease in congenitally infected infants and in immunocompromised individuals. Toxoplasmosis has been implicated as the third most common source of food borne infection in the USA. Due to the risk of food and water borne infection, *T. gondii* is listed as a category B Biodefense Agent by NIAID. In addition to being an important pathogen, *T. gondii* has emerged as a model for other less tractable apicomplexan parasites. Whole genome sequences have previously been generated for three prototypic strains presenting the major clonal lineages found in North America and Europe. More recent studies have detected additional genetic variation in North America and Europe, suggesting a more complex population structure than previously thought. Additionally, sampling from South America has revealed that strains from this region are highly divergent, and comprise novel groups with both clonal and nonclonal genotypes. Importantly, some of these South American lineages have been associated with severe ocular disease, suggesting that differences in clinical severity may be influenced by the parasite genotype. Comparison of more than 900 strains analyzed to date indicates that they can be grouped into 12 major lineages. Current whole genome sequences are available for only 3 of these, reflecting relatively narrow genetic and biological diversity. The proposed studies will advance research on toxoplasmosis by generating high-coverage, whole genome sequences for 9 new prototypic strains, thus providing annotated genomes for designated references strains for all 12 of the major lineages described to date. RNA sequencing from these same prototypic strains will be used to improve gene models and compare gene expression among the lineages. These reference strains will also be deposited into a public archive, making them available for study by the community. As well, moderate-coverage, whole genome sequences will be generated for an additional 35 members of the 12 lineages to assess genetic diversity, determine population structure, and establish patterns of antigenic and allelic variation. Collectively these studies will foster future studies on population structure, genetic diversity, basic cellular and molecular biology of *T. gondii*. They are also expected to advance efforts to improve diagnostics, advance epidemiological studies, and develop new treatments to combat infection.

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.*

Significance of the organism

Toxoplasma gondii is a wide spread protozoan parasite of animals and an important opportunistic pathogen in humans, causing disease in congenitally infected infants and in immunocompromised individuals. *T. gondii* is transmitted by cats, the definitive host, and infects a wide range of intermediate hosts (Dubey, 2007). Humans are not a natural part of the life cycle but they become infected by ingestion of tissue cysts in under cooked meat, or oocysts that are shed by cats and which can contaminate water and food (Dubey, 2007). Toxoplasmosis has been implicated as the third most common cause of food borne infection in the USA (Mead et al., 1999). Serological studies suggest that chronic infection rates in humans can vary from less than 10% to greater than 70% depending on geographic region and various risk factors (Joynson and Wreghitt, 2001). Although most human infections are often asymptomatic, they predispose individuals to complications of reactivation in the event of impaired immunity. In contrast, genetically diverse strains have been associated with severe disseminated disease in French Guyana (Dardé et al., 1998) and with debilitating ocular toxoplasmosis in southern Brazil (Silveira et al., 2001; Jones et al., 2006); notably these cases occurred in otherwise healthy adults. *T. gondii* is listed as a category B Biodefense Agent by NIAID due to its risk of transmission through contamination of food or water and importance as an opportunistic pathogen. *T. gondii* has also emerged as a model apicomplexan parasite, due to the ease of experimental investigation. Hence, studies performed in *T gondii* are often extrapolated to related but less tractable pathogens such as *Cryptosporidium*, another category B Biodefense Agent that causes severe diarrheal disease.

Previous genome-wide studies

Whole genome sequences have previously been generated for three prototypic strains of *T. gondii re*presenting the three major lineages found in North America and Europe. These studies reveal a genome size of ~65 megabases, containing relatively few repeats, with a slight GC bias, and comprising 14 chromosomes. The assembled and annotated whole genome sequences, as well as details on the funding agencies, sequencing centers, and community organizers involved in these projects are provided in a genome

database for *T. gondii* http://toxodb.org/toxo/ (Gajria et al., 2007). ToxoDB is a component of the Eukaryotic Pathogen Genome Database (EuPathDB), an NIAID-supported Bioinformatics Resource Center emphasizing biodefense pathogens.

Previous sequencing efforts have demonstrated that ~ 10X coverage using conventional sequence reads from ABI 377 sequencers was sufficient to assemble the *T. gondii* genome into large contigs and capture allelic variation with a high degree of confidence. This was originally conducted by whole genome shotgun sequencing of the type II ME49 strain, which serves as the reference genome for *T. gondii*. Importantly, the contigs were then re-assembled into chromosomes using a genetic linkage map (Khan et al., 2005). This genome map was further enhanced by end-reads from cosmid and BAC libraries that were used to close some gaps in the assemblies. The genomes for type I (GT1 strain) and type III (VEG strain) were subsequently obtained by whole genome shotgun sequencing (again using conventional reads), assembled, annotated, and integrated into ToxoDB. Gene annotations based on several different models are provided in ToxoDB, which also houses several genomic-scale datasets including ESTs, SAGEs, expression profiling arrays, proteomic surveys, and CHiP data on chromatin modification. ToxoDB also supports a wide range of queries to facilitate data mining by the community.

Although the *T. gondii* genomes are not completely "finished", they provide a high degree of sequence coverage (>10X on average) that has enabled numerous research studies relevant to human disease and biodefense concerns. This resource is heavily used by the *Toxoplasma* research community, including more than 200 basic research laboratories worldwide. The contents of ToxoDB are also relevant to the many clinical research laboratories working on toxoplasmosis throughout the world. For example, information from this site has enabled the development of polymorphic markers for genetic linkage studies and strain typing. These resources are highly valuable; however, they are limited by the lack of genome sequence data for many of the newly described lineages, as discussed below.

Genetic diversity of *T. gondii*

T. gondii displays an unusual population structure consisting of three predominant strains that are abundant in North America and Europe (Sibley and Ajioka, 2008). Most investigators currently use one of the three prototypic strains that have been sequenced to date. The extent of divergence within the three lineages is very minor owing to their recent origin and clonal expansion (Sibley and Ajioka, 2008). However, sampling from South America has revealed that strains from this region are highly divergent and comprise novel groups with both clonal and non-clonal genotypes (Lehmann et al., 2006; Khan et al., 2007; Pena et al., 2008). Importantly, some of these South American lineages are associated with severe ocular disease (Khan et al., 2006), suggesting that differences in clinical severity may be influenced by the parasite genotype. Comparison of the >900 strains analyzed to date using RFLP markers has allowed classification of 140 unique genotypes into approximately 12 major groups (Figure 1). Similar results were obtained by sequencing of introns from a smaller set of strains (Khan et al., 2007), and these results are also concordant with studies using microsatellite markers (Ajzenberg et al., 2002a; Ajzenberg et al., 2002b). Thus far, these studies have been focused primarily in Europe and the Americas. Ongoing efforts to further define *T. gondii* population structure in other geographic regions may reveal new genotypes. Regardless of the outcome of these studies, there is presently an urgent need to generate whole genome sequences from major known lineages for which little data is available.

Knowledge gaps to be filled by the proposed study

Research over the past few years has revealed that the genetic diversity of *T. gondii* is far greater than previously appreciated. The population structure is strongly subdivided by geographic region and by the existence of clonal lineages in some regions (Sibley and

Ajioka, 2008). Studies of the basic biology of this parasite are largely based on just a few representative members of the genetic lineages that are common in North America and Europe. Although much has been learned from these studies, the existing genomes are not representative of the global population as a whole. In particular, large numbers of new genetic types have been found in Central and South America. More recent studies have also revealed additional genetic variation in North America and Europe, suggesting a more complex population structure than previously thought. Increasingly these "exotic" genotypes have been associated with human infections, yet their genetic and biological diversity are not well understood. The proposed studies will provide whole genome sequences for members of these additional genetic groups, thus defining their gene content, profiling differences in expression, and expanding our understanding of genetic diversity. A combination of high-coverage sequencing for designated prototypic strains combined with moderate coverage for additional isolates from each group will support further studies on genetic diversity within and between lineages. For comparative purposes, we will obtain genome sequences of several closely related protozoan parasites outside the genus Toxoplasma. Collectively, these data will allow comparison of genetic composition, chromosome organization, gene content, gene expression, synteny, diversity, and estimates of ancestry within and between the major lineages.

Another present deficiency that restricts research on *T. gondii* is the uncertainly of currently available gene models. This is a consequence of historical application of diverse gene finding algorithms, and the shortage of the full-length cDNA sequences needed to refine models for a eukaryotic organism with numerous introns (and differential splicing). Such problems are exascerbated because accurate prediction of first exons – the most difficult part of gene finding – is particularly important for organisms like *T. gondii* that depend heavily on secreted proteins (as targeting information is frequently encoded in the first exon). We will therefore perform deep sequencing of mRNAs from the prototypic strains for each of the major groups, and the resulting transcriptome profiles will be mapped to the genome to better define gene splicing and to support improved gene models. They will also be useful for comparing the relative expression level of genes between different lineages.

Additionally, many aspects of the biology of *T. gondii* are poorly understood in terms of host range, pathogenicity, and transmission. Hence, there is a great need for reference strains that can be used to characterize these biological features and to support studies on basic cell and molecular biology and gene regulation. At present most laboratories use only a few reference strains and these fail to capture the full genetic and biological diversity of *T. gondii*. The proposed project will make available a wider range of characterized strains that will be archived in public repositories and for which whole genomes will be available. This advance will support a variety of functional studies ranging from basic science topics to more applied projects to improve diagnostics, perform drug sensitivity studies, investigate host resistance, and control of infection.

Relevance of the project to toxoplasmosis in humans

There are many outstanding clinical/epidemiological questions that will either be directly addressed by the data generated here, or where future studies that address them will be massively facilitated by these data. Among such issues are the following:

The clinical outcome of human infection with *T. gondii* is highly variable, ranging from asymptomatic to fatal (even in non-immunocompromised adults). Likewise, in the developing fetus or immunocompromised adult, outcomes also range from mild to extreme. While environmental and host factors almost certainly play a role, the strain type is strongly implicated based on epidemiologic and animal studies. Determining a specific role for *T. gondii* strain-type is dependent on better reagents for serological and PCR-based

methods for strain-typing. Both methods would benefit hugely from an extensive collection of genome sequences that identify antigens and sequences that are specific for particular strains.

The route of transmission is unknown in the vast majority of human infections. The data to be generated here will help address this by providing strain- and stage-specific markers that can be used in combination with epidemiological studies. For example, it is clear that toxoplasmosis is a zoonosis yet the most important animal sources of human infection are presently unknown. Generating an extensive collection of strain sequences will allow the development of methods for rapidly and efficiently determining the strain present in suspected reservoirs of human infection. Even more remarkably, there are almost no data that address whether the majority of humans are infected by ingestion of oocysts (shed in feline feces and ingested through environmental contamination) or tissue cysts (present in meat and other tissues of livestock). The data to be generated here will greatly facilitate development of reagents for determining whether a given infection initiated with the oocyst or tissue cyst stage. As a result of this work, therefore, it may become possible to develop strain- and stage-specific peptides for serological study to distinguish between different origins of an infection.

Treatment strategies for toxoplasmosis have been empirically determined and are the subject of much debate. For example, in ocular toxoplasmosis, some clinicians recommend use of steroids to reduce inflammation while others fear that such can exacerbate the infection. Likewise, there is debate about the effectiveness of current therapies for preventing congenital infection. Recent data in animal studies suggest that, as with other pathogens where the immune response can be a key part of the pathogenesis, cytokine levels are highly variable depending on which strain of *T. gondii* is responsible for a given infection, ranging from relatively low to and extreme "cytokine storm". Hence, it may be that the correct clinical decision will depend on identifying the particular strain of *T. gondii* and the corresponding repertoire of its interaction with the human host. Generation of an extensive database of *T. gondii* sequences and a set of reference strains for further biological study, may provide crucial tools for managing and treating localized outbreaks of toxoplasmosis.

References

Ajzenberg, D., Bañuls, A.L., Tibayrenc, M., Dardé, M.L., 2002a. Microsatellite analysis of *Toxoplasma gondii* shows considerable polymorphism structured into two main clonal groups. Intl. J. Parasitol. 32, 27-38.

Ajzenberg, D., Cogne´, N., Paris, L., Bessieres, M.H., Thulliez, P., Fillisetti, D., Pelloux, H., Marty, P., Dardé, M.L., 2002b. Genotype of 86 *Toxoplasma gondii* isolates associated with human congenital toxoplasmosis and correlation with clinical findings. J. Infect. Dis. 186, 684-689.

Dardé, M.L., Villena, I., Pinon, J.M., Beguinot, I., 1998. Severe toxoplasmosis caused by a *Toxoplasma gondii* strain with a new isotype acquired in French Guyana. J. Clin. Microbiol. 36, 324.

Dubey, J.P., 2007. The history and life cycle of *Toxoplasma gondii*. In: Weiss, L.M.,Kim, K. (Eds.), *Toxoplasma gondii* the model Apicomplexan: perspectives and methods, Academic Press, Elsevier, New York, pp. 1-17.

Gajria, B., Bahl, A., Brestelli, J., Dommer, J., Fischer, S., Gao, X., Heiges, M., Iodice, J., Kissinger, J.C., A.J., M., Pinney, D.F., Roos, D.S., Stoeckert, C.J., Wang, J., Brunk, B.P., 2007. ToxoDB: an integrated *Toxoplasma gondii* database resource. Nucl. Acids Res. 36, D553-556.

Jones, J.L., Muccioli, C., Belfort, R., Jr. , Holland, G.N., Roberts, J.M., Silveira, C., 2006.

Recently acquired *Toxoplasma gondii* infection, Brazil. Emerg. Infect. Dis. 12, 582-587.

Joynson, D.H., Wreghitt, T.J., 2001. Toxoplasmosis: A comprehensive clinical guide. Cambridge University Press.

Khan, A., Jordan, C., Muccioli, C., Vallochi, A.L., Rizzo, L.V., Belfort Jr., R., Vitor, R.W., Silveira, C., Sibley, L.D., 2006. Genetic divergence of *Toxoplasma gondii* strains associated with ocular toxoplasmosis Brazil. Emerg. Infect. Dis. 12, 942-949.

Khan, A., Fux, B., Su, C., Dubey , J.P., Darde, M.L., Ajioka, J.W., Rosenthal, B.M., Sibley, L.D., 2007. Recent transcontinental sweep of *Toxoplasma gondii* driven by a single monomorphic chromosome. Proc. Natl. Acad. Sci. (USA) 104, 14872-14877.

Lehmann, T., Marcet, P.L., Graham, D.H., Dahl, E.R., Dubey, J.P., 2006. Globalization and the population structure of *Toxoplasma gondii*. Proc. Natl. Acad. Sci. (USA) 103, 11423-11428.

Mead, P.S., Slutsker, L., Dietz, V., McCaig, L.F., Bresee, J.S., Shapiro, C., Griffin, P.M., Tauxe, R.V., 1999. Food-related illness and death in the United States. Emerg. Infect. Dis. 5, 607-625.

Pena, H.F., Gennari, S.M., Dubey, J.P., Su, C., 2008. Population structure and mouse-virulence of *Toxoplasma gondii* in Brazil. Intl. J. Parasitol. 38, 561-569.

Sibley, L.D., Ajioka, J.W., 2008. Population structure of *Toxoplasma gondii*: Clonal expansion driven by infrequent recombination and selective sweeps. Ann. Rev. Microbiol. 62, 329-351.

Silveira, C., Belfort, R., Jr., Muccioli, C., Abreu, M.T., Martins, M.C., Victora, C., Nussenblatt, R.B., Holland, G.N., 2001. A follow-up study of *Toxoplasma gondii* infection in southern Brazil. Am. J. Ophthalmol. 131, 351-354.

3. Rationale for Strain Selection

4. *Provide the rationale behind the selection of strains and the number of strains proposed in the study. The focus of the program is on potential agents of bioterrorism or organisms responsible for emerging or re-emerging infectious diseases. Non-select agents or non-pathogenic organisms will be considered when they can provide insight into these scientific areas.*

Proposed sequencing of strains

Overview: We are proposing to obtain high-coverage, whole genome sequences from additional prototypic strains of *T. gondii*, as well as sampling of several representative members from each major haplogroup. To define the extent of genetic diversity, we have combined published and unpublished data from the main laboratories involved in genetic typing of *T. gondii*. These studies are ongoing, but at present they indicate more than 140 unique genotypes. Clustering by various methods reveals that these individual genotypes belong to approximately 12 major groups, as shown by the network analysis in Figure 1. Within these 12 groups, ~75% of the strains cluster within the top 30 most abundant genotypes. From this set of clustered strains, we have chosen one prototypic strain from each major lineage (denoted in bold in Figure 1). Additionally, we have selected other closely related members of each group, based on abundance, divergence, geographic and host range. In total, we have selected 9 new prototypic strains and 35 related strains as part of the sequencing plan (Table 1, Figure 1). We are proposing a combination of high coverage sequencing of DNA and RNA for the prototypic reference strains with moderate coverage for the remaining strains. These levels of coverage are designed to provide nearly complete

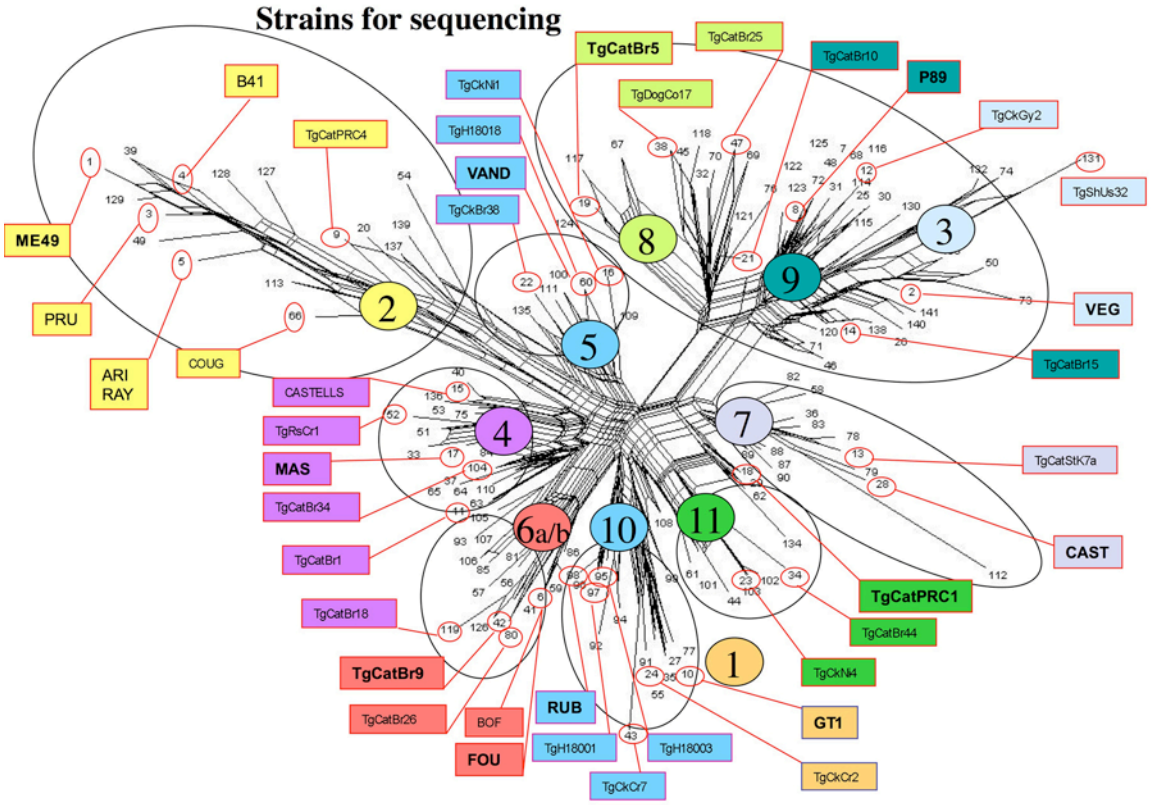

Figure 1 Network analysis of 140 genotypes of *T. gondii* defined by 10 RFLP and sequence-based markers. Candidate strains for sequencing are shown in boxes, prototypic strains denoted in bold. Additional strains are listed in Table 1.

reference genomes for each of the major groups and to capture the extent of genetic diversity within each group. Finally, we are proposing whole genome sequencing of two related apicomplexan species to serve as outgroups. Coverage required for these different groups varies depending on the degree of genetic diversity and project objectives, as outlined below.

Enhanced sequence coverage and reannotation of the reference type II genome

The most abundant cluster of *T. gondii* strains fall into the type II lineage, for which the reference genome is ME49. This is the first genome that was sequenced and average coverage is only ~ 10X, with notable gaps. Importantly, some gene families that are repetitive are poorly represented in the assembly, despite playing critical roles in pathogenesis (i.e. secretory ROP kinases and surface antigens). We are proposing additional pair-end genomic DNA reads from the type II ME49 strain using either Solexa (76 bp paired-end from various length fragments) or paired-end 454 reads from various sized libraries. These increased reads will be used to generate a new higher quality assembly. Alignment of the RNASeq reads described below will be used to improve the gene models and generate an improved annotation, which then be used as the reference for other genomes.

Prototypic strains for each lineage: 9 genomes at high coverage

Currently there are whole genome sequences available for three of the twelve major lineages of *T. gondii* (types 1, 2, and 3, See Table 1). We have selected 9 representative prototypic strains in order to complete whole genome sequencing of the major lineages that are presently known (Table 1, Figure 1). The specific strains to be used have been chosen based on the following criteria: date, source, and host of isolation are known; passage history is known; isolated from human infections, companion or food animals; information on infection in animal models is available; and the strains are available for deposit in public repositories without restrictions.

To capture meaningful differences in gene content and polymorphisms that will be essential for genetic linkage mapping, strain-specific serological testing, population studies, and functional analyses, it will be important that the reference genotypes for the prototypic strains have a high degree of coverage. Hence, for these 9 prototypic *T. gondii* lineages, ~20 X coverage will likely be required, assuming 454 sequencing technology is employed. It will also be desirable to obtain end reads from BAC or fosmid libraries to aid in assembly. The whole genome sequences for these 9 new prototypic strains will be assembled using the existing reference genome of the ME49 strain as a guide.

RNA sequencing: 12 genomes at high coverage

Shotgun transcriptome sequencing of RNA (RNA-Seq) will be performed for all 12 of the prototypic strains. These analyses will include both the mRNA populations as well as small RNAs. Parasite small RNAs have been implicated in transcriptional control of parasite poly A mRNAs. These transcriptome data will be useful for retraining gene models, providing information on stage-specific and lineage-specific alternative splicing, and for comparing gene expression between lineages. RNA-Seq will be conducted on mRNA (cDNAs) the major replicative forms (tachyzoites, bradyzoites, sporozoites (sporulated oocysts)) from the type II strain ME49. Since ME49 is the reference genome for which additional gene models will be developed, coverage for these stages should include 10^7 reads per life cycle stage. These reads will be mapped back to the existing genome of Me49 and used to refine gene models and provide a reference for annotation for other genomes. To examine differences in global gene expression, RNA-Seq will also be conducted on mRNA (cDNAs) from tachyzoites of the prototypic strains for the remaining 11 major lineages (Table 1), including the previously sequenced type III (VEG) and type I (GT1) strains. For these remaining strains, RNAs will be harvested from freshly egressed parasites (tachyzoites) propagated in vitro. To assure adequate coverage for detecting differences in splicing, or expression level, 10^7 RNA-Seq reads will be generated for each of these prototypic strains. Small RNA will be mapped back to the genome and their locations will be included in the genome annotation.

Within lineage variation: 35 genomes at moderate coverage

To accurately capture within group variation, we are proposing moderate coverage DNA sequencing of different representatives from each of the 12 major lineages. It is likely that 10-12x coverage using 76 bp paired-end reads from Solexa or Illumina type technologies will be sufficient to provide adequate coverage for gene identification and SNP identification. Reads from these representative isolates will be mapped to the closest reference genome from the 12 major groups described above. Although members of the I, 2, 3 lineages are highly clonal (Su et al., 2003), some strains within these groups show evidence of recombination with other more divergent strains. Recent evidence also suggests that these mixed isolates may have very different biological traits. For example type II-like strains have been associated with encephalitis in sea otters (Miller et al., 2004), and mixed strains have been associated with more severe ocular infection in humans (Grigg et al., 2001). Several of these mixed strains will be included to estimate the degree of genetic diversity within groups I, 2, and 3 (denoted with a * in Table 1). Given the widespread us of clonal strains like RH, it may also be desirable to obtain moderate coverage of these genomes for comparison to the prototypic type I strains GT-1.

Isolates from other more divergent lineages have been chosen to maximize genetic diversity while sampling the most prevalent genotypes (Figure 1). In particular, we are proposing more extensive sampling of several South American lineages, notably 4 and 9 (Figure 1, Table 1). Each of these lineages has been predicted to be an ancestral group (Khan et al., 2007), and their high degree of diversity suggests they have undergone more extensive recombination in the wild. This feature also suggests they will have more diverse biological traits including those that affect pathogenesis and transmission. Consistent with this, severe ocular strains and strains associated with waterborne outbreaks in Brazil have been associated with the type 4 lineage (Khan et al., 2006). More extensive sampling is planned from these groups (Figure 1, Table 1). We are also including members of the highly divergent types 5 and 10, which have been associated with extremely severe disease in otherwise healthy adults (Dardé et al., 1998; Carme et al., 2002). These isolates come from French Guyana, where the genetic diversity is very high. We are also planning increased sampling of group 6, which is widely distributed in Europe (6a) and South America (6b), and often associated with human infection. Finally, more divergent but not quite as common lineages 7, 8 and 11 will also be included for comparison. Although these groups have not yet been found to be prevalent in humans, they are nonetheless common in animals, including both companion and food animals. Hence, the genetic diversity of these sources is relevant to possible zoonotic infections.

Outgroups: whole genome sequencing of comparator groups

A number of other apicomplexans have been sequenced and these are useful for comparative genomic studies. Included are *Plasmodium* spp., *Cryptosporidium* spp., and *Theileria* spp., all of which are available in EuPathDB. Whole genome sequences have also been generated by the Welcome Trust Sanger Institute for *Eimeria tenella*, *Babesia bigemina*, and *Neospora caninum*. As well, a whole genome project was recently approved for *Sarcocystis neurona*, a pathogen of horses (jointly funded by USDA/NSF and coordinated by Dan Howe). Although these sequences are useful for comparative studies, we are lacking very close relatives of *T. gondii* and also do not have representative of the more distantly related gregarines. We are proposing whole genome sequencing of two additional organisms to fill these important gaps.

Gregarines are deep branching members of the phylum Apicomplexa that infect a wide variety of invertebrate hosts. Currently there are approximately 1,600 named species of gregarines, although this is a conservative estimate. The great diversity of gregarines and their evolutionary position make them a KEY outgroup for comparative genomic analysis. They retain a number of features that are characteristic of apicomplexnas including an apical

complex and actin-myosin based gliding motility. We have chosen a gregarine for which a monotypic culture can be obtained and which has previously been the focus of a small scale EST project conducted at Washington University (Omoto et al., 2004). The genome size of *G. niphandrodes* is estimated to be between 12-17 Mb based on quantitative fluorescence microscopy. We propose to sequence the genome of *G. niphandrodes* to ~ 20X by combination of 454 sequencing, and end sequencing from BAC clones and fosmid libraries. The genome will be assembled and annotated based on BLAST comparison. In addition, RNA-Seq reads will be generated from the trophozoite and gametocyst stages and these will be mapped to the genome to develop appropriate gene models and identify genes.

We also propose sequencing of the genomes of *Hammondia hammondi*, which is more closely related to *T. gondii. H. hammondi* is a parasite primarily of rodents that is transmitted by cats as the definitive host. It differs from *T. gondii* in that this life cycle is obligatory (referred to as heteroxenous). In other words, Hammondia lacks direct transmission by ingestion of tissue stages between intermediate hosts, and also has a more narrow range of intermediate hosts. Although not a major animal or human pathogen, the genome of *H. hammondi* will be extremely useful for comparative studies on the evolution of virulence, host range, and transmission. Additionally, animals that infected with *H. hammondi* develop cross reactive antibodies to *T. gondii*, and the oocysts of these two species are highly similar. Both of these features confound epidemiology studies. Availability of whole genome sequence for *H. hammondi* will facilitate development of more specific diagnostic tools to distinguish these closely related parasites.

In the case of *H. hammondi*, ~20X coverage by 454 technology combined with end reads from BAC or fosmid libraries will likely be required for assembly. Although we do not have a precise estimate of the genome size of *H. hammondi*, it is likely to be similar to *T. gondii* (65 mb haploid size). Genomic DNA will be isolated from oocysts obtained from infected cats, in collaboration with J.P. Dubey.

References

Carme, B., Bissuel, F., Ajzenberg, D., Bouyne, R., Aznar, C., Demar, M., Bichat, S., Louvel, D., Bourbigot, A.M., Peneau, C., Neron, P., Dardé, M.L., 2002. Severe acquired toxoplasmosis in immunocompetent adult patients in French Guiana. J . Clin. Microbiol. 40, 4037-4044.

Dardé, M.L., Villena, I., Pinon, J.M., Beguinot, I., 1998. Severe toxoplasmosis caused by a *Toxoplasma gondii* strain with a new isotype acquired in French Guyana. J. Clin. Microbiol. 36, 324.

Grigg, M.E., Ganatra, J., Boothroyd, J.C., Margolis, T.P., 2001. Unusual abundance of atypical strains associated with human ocular toxoplasmosis. J. Infect. Dis. 184, 633-639.

Khan, A., Jordan, C., Muccioli, C., Vallochi, A.L., Rizzo, L.V., Belfort Jr., R., Vitor, R.W., Silveira, C., Sibley, L.D., 2006. Genetic divergence of *Toxoplasma gondii* strains associated with ocular toxoplasmosis Brazil. Emerg. Infect. Dis. 12, 942-949.

Khan, A., Fux, B., Su, C., Dubey , J.P., Darde, M.L., Ajioka, J.W., Rosenthal, B.M., Sibley, L.D., 2007. Recent transcontinental sweep of *Toxoplasma gondii* driven by a single monomorphic chromosome. Proc. Natl. Acad. Sci. (USA) 104, 14872-14877.

Miller, M.A., Grigg, M.E., Kreuder, C., James, E.R., Melli, A.C., Crosbie, P.R., Jessup, D.A., Boothroyd, J.C., Brownstein, D., Conrad, P.A., 2004. An unusual genotype of Toxoplasma gondii is common in California sea otters (*Enhydra lutris nereis*) and is a cause of mortality. Intl. J. Parasitol. 34, 275-284.

Omoto, C.K., Toso, M., Tang, K., Sibley, L.D., 2004. Expressed sequence tag (EST) analysis of Gregarine gametocyst development. Intl. J. Parasitol. 34, 1265-1271.

Su, C., Evans, D., Cole, R.H., Kissinger, J.C., Ajioka, J.W., Sibley, L.D., 2003. Recent expansion of Toxoplasma through enhanced oral transmission. Science 299, 414-416.

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.)*

Additional genomic reads will be generated for the reference type II genome from the ME49 strain. These reads will be combined with existing data to provide an improved assembly.

RNA-Seq reads will be generated from different life cycle stages of the type II ME49 reference strain and used to improve gene models and annotation for *T. gondii*. This will result in a high quality reference genome.

High-coverage, whole geneome sequences will be obtained for 9 new prototypic strains representing geneticaly distinct lineages of *T. gondii*. These will be assembled and annotated.

RNA-Seq reads will be generated from prototypic members of each of the 12 lineages and used to profile differences in gene expression.

Modeate-coverage whole geneome sequences will be obtained for 35 additional members of the different lineages of *T. gondii*.

Annotated high-coverage geneome sequences will be developed for two outgroup species for comparison to *T. gondii*.

A set of reference strains of biologic strains will be deposited in public repositories.

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.*

Annotations

All sequence reads will be produced and assemblies generated by JCVI, including an improved assembly of the ME49 genome including new sequence data, and refinement of gene models based on RNASeq results. Preliminary annotation for the remaining prototypic genomes will be performed by comparison to the reference ME49 annotated genome. All sequences will be submitted to GenBank, with JCVI and EuPathDB listed as joint owners of the records. These data will also be imported into ToxoDB, and analyzed for DNA and predicted protein features using established EuPathDB workflows, providing graphical views of the genome (e.g. GBrowse tracks showing genes, SNPs, synteny, etc) and a variety of complex queries enabling the research community to mine the data for biologically-relevant information. Extensive manual curation of these genome is not planned, but ME49 gene models provide a point of reference, and the user community

may add annotations as user comment entries in ToxoDB.

Gene Models
RNA-Seq reads will be mapped to the reference genome of ME49 and used to improve the gene models and for predicting splicing and 5' and 3' ends. The resulting improved gene model will also be used to predict genes for the prototypic strains that are representative of the major lineages. Comparison of RNA-Seq reads will also be used to identify alternative splicing or expression level differences between life cycle stages or strains. Workflows already in place at ToxoDB provide a variety of automated analyses, including BLAST analysis and the identification of targeting signals and motifs.

SNP identification
SNPs will be defined by comparison of short sequence reads to the closest reference genome. SNPs will be displayed in ToxoDB in order to illustrate these as tracks against a reference genome. In additional, queries will be provided in ToxoDB to analyze SNPs in genes of interest between strains.

Gene expression differences
RNA reads will be mapped to the genome to generate relative expression values (e.g. mean coverage per 1000 unique bp / gene), and displayed on gene record pages in ToxoDB and genome browser tracks showing coverage of RNA reads across the genome. Such tracks provide evidence for gene models including 5' and 3' UTRs, facilitating analysis of exon boundaries, identifying differentially processed transcripts etc. Queries based on these expression values will be integrated with existing queries against microarray and SAGE tag datasets already available in ToxoDB.

Additional research areas that will be fostered by the availability of the data:
Availability of whole genome sequence data from a large collection of *T. gondii* isolates will foster further studies on genetic diversity, population structure, and ancestry of lineages. These studies are highly relevant to the patterns of spread of genes that influence pathogenesis within hosts and transmission between hosts. Sequence polymorphism data for antigenic genes is expected to support more advances studies to detected human infections based on strain-specific serological responses. As well, several sets of polymorphic antigens have previously been implicated in virulence and host defense in animal models. Availability of polymorphism data on these and other genes from a wide collection of strains may suggest specific functional differences that could be pursued by hypothesis driven approaches. Availability of the prototypic strains used in the sequencing effort will also enable studies in animal models for monitoring pathogenesis and/or host defense. These strains will also be useful for testing drug susceptibility among a wider range of isolates. Finally, the comparison of outgroups may enable hypothesis driven studies designed to investigate the basis of host range and transmission among this group of parasites.

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.*

Community involvement:
The Toxoplasma research community is highly collaborative and has worked closely to advance common technology platforms and biological resources to improve research opportunities. Consistent with these previous efforts, this white paper has been assembled with input from the entire community. We recently held the *10th International Toxoplasmosis* meeting, and this project was discussed among the scientists in attendance. The participants at the meeting represented more than 200 worldwide laboratories that are engaged in basic research in *T. gondii*. Nearly all of these laboratories are active users of genomic data obtained through previous sequencing efforts, which are housed in ToxoxDB, a component of the EuPathDB Bioinformatics Resource Center for biodefense pathogens. There was overwhelming support for the value of both whole genome and RNA sequencing efforts. Capturing additional genetic diversity of this organism and improving the prediction of genes and estimating their degree of polymorphism in the population will be critical for understanding *T. gondii* virulence and pathogenesis in human infection. To meet these needs of the research community, the following organizing committee was assembled to guide the proposed project.

Dr. James W. Ajioka, Cambridge University, UK
Dr. Daniel Ajzenberg, University of Limoges, Grance
Dr. John C. Boothroyd, Stanford University, USA
Dr. Brian P. Brunk, University of Pennsylvania, USA
Dr. J.P. Dubey, USDA, USA
Dr. Marie Laure Dardé, University of Limoges, France
Dr. Michael E. Grigg, NIH, USA
Dr. Daniel K. Howe, University of Kentucky, USA
Dr. Kami Kim, Albert Einstein College of Medicine, USA
Dr. Charlotte Omoto, Washington State University, USA
Dr. Benjamin M. Rosenthal, USDA, USA
Dr. David S. Roos, University of Pennsylvania, USA
Dr. L. David Sibley, Washington University, USA
Dr. Chunlei Su, University of Tennessee, USA
Dr. Michael W. White, University of South Florida, USA

Drs. Ajzenberg, Dardé, Dubey, Sibley, and Su will primarily oversee the comparison of the genotypes of strains and choice of strains for sequencing. Drs. Roos and Brunk will coordinate data analysis and integration of sequences into ToxoDB.

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?*

Note: If samples are prospectively prepared the GSC can provide protocols and recommendation based on the Centers past experiences. The samples must however meet minimum quality standards as established by the Center for the optimal technology platform (sequencing/ genotyping) to be used in the study.

11. *Attach relevant information, if available in an excel spreadsheet for multiple samples: e.g*
    - *Name*
    - *Identifier*
    - *Material type (DNA/RNA/Strain)*
    - *Genus*
    - *Species*
    - *Specimen / Strain*
    - *Isolation source*
    - *Isolated from*
    - *Select agent status*
    - *International permit requirement*
    - *BEIR/ATCC repository accession number*
    - *Other public repository location*
    - *Other public repository identifier*
    - *Sample provider's name*
    - *Sample provider's contact*

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

The list of proposed strains for sequencing is found on the following page. All of the *T. gondii* strains to be sequenced under this proposal are currently available from laboratories within the USA. They are not restricted by import requirements, intellectual property or ongoing IRB approval. Most have been in the laboratory setting for 5 or more years and efforts have been made to preserve them as low passage isolates. The majority of *T. gondii* strains are available in the laboratory of Dr. David Sibley, Washington University. The remaining strains will be obtained from Dr. J.P. Dubey at the USDA, or from Dr. Marie Laure Dardé, Biological Resource Center for Toxoplasma (BRC Toxoplasma) in France. The Sibley lab will propagate strains and produce DNA and RNA samples that will be sent to the JCVI for sequencing. Protocols for assuring QC standards will be worked out jointly between the participant laboratories and JCVI. Samples for *H. hammondi* will be prepared by Dr. J.P. Dubey at the USDA. Samples for sequencing of *G. niphandrodes* will be provided by Charlotte Omoto. For most *T. gondii* isolates, we have information on their lethality in the murine model, ability to undergo differentiation, and limited data on transmission. These data, along with a summary of the host of origin, geographic location, and date of isolation will be included in the description forms when the strains are submitted to appropriate public archives. They will also be posted in a haplotype – strain database maintained by the Sibley Lab (http://toxomap.wustl.edu/). As well, strain histories and phenotypes will be incorporated into ToxoDB, where users can update phenotypic data as it becomes available.

Table 1 Toxoplasma strains for whole genome sequencing

gDNA	cDNA	Strains	Host	Geographic	Year	Haplogroups
Done		**GT1**	Goat	USA-MD	1980	1
		TgCkCr2	Chicken	Costa Rica	2006	1*
Done		**ME49**	Sheep	USA-CA	1965	2
		PRU	Human (congenital)	France	1964	2*
		ARI	Human (transplant)	USA	1992	2*
		RAY	Human (congenital)	USA	1993	2*
		B41	Bear	USA	1994	2*
		TgCatPRC4	Cat	China	2007	2*
		COUG	Cougar	Canada-BC	1996	2*
Done		**VEG**	Human (AIDS)	USA-CA	1988	3
		TgShUs32	Sheep	USA	2008	3*
		TgCkGy2	Chicken	Guyana	2007	3*
		MAS	Human (congenital)	France (Nice)	1991	4
		CASTELLS	Sheep	Uruguay	1993	4
		TgCatBr1	Cat	Brazil	2006	4
		TgCatBr34	Cat	Brazil	2006	4
		TgCatBr18	Cat	Brazil	2006	4
		TgRsCr1	Toucan	Costa Rica	2006	4
		BRC TgH 26044	Human (congenital)	Europe	2007	(4)
		BRC TgH 21016	Human (congenital)	Europe	2007	(4)
		BRC TgH 20005	Human (congenital)	Europe	2002	(4)
		GUY-2004-JAG1	Jaguar	French Guiana	2004	(4)
		RUB	Human adult	French Guiana	1992	5
		BRC TgH 18018	Human adult	French Guiana	2002	5
		BRC TgH 18003	Human adult	French Guiana	2002	5
		TgCkBr38	Chicken	Grazil	2008	5
		TgCkNi1	Chicken	Nicaragua	2006	5
		FOU	Human (transplant)	France (Brest)	1992	6a
		BOF	Human (AIDS)	Belgium	1993	6a
		TgCatBr9	Cat	Brazil	2006	6b
		TgCatBr26	Cat	Brazil	2006	6b
		CAST	Human (AIDS)	USA-CA	1988	7
		TgCatStK7a	Cat	St. Kitts	2009	7
		TgCatBr5	Cat	Brazil	2006	8
		TgDogCo17	Dog	Columbia	2006	8
		TgCatBr25	Cat	Brazil	2006	8
		P89	Pig	USA-IA	1991	9
		TgCatBr15	Cat	Brazil	2006	9
		TgCatBr10	Cat	Brazil	2006	9
		VAND	Human adult	French Guiana	1996	10
		BRC TgH 18001	Human adult	French Guiana	1997	10
		TgCkCr7	Chicken	Costa Rica	2006	10
		BRC TgH 18021	Human adult	French Guiana	2006	(10)
		BRC TgH 18009	Human adult	French Guiana	2004	(10)
		TgCatPRC1	Cat	China	2007	11
		TgCatBr44	Cat	Brazil	2008	11
		TgCkNi4	Chicken	Nicaragua	2006	11

(tentative)

Medium coverage (i.e 10X) N= 35
High coverage (i.e 25-30x) N=9
RNA Seq (20X) N=12

Note: in rare circumstances it may be necessary to substitute other closely related strains, for example in the event that a particular strain cannot be revived from cryopreservation.

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 X Decline ☐

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

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

All of the prototypic strains, for which have high-coverage genomes will be generated, will be deposited in the Biodefense and Emerging Infections Repository (BEIR) subsection of the American Type Culture Collection. Independently (and not funded by this initiative) our French colleagues will submit strains to the Biological Resource Center for Toxoplasma, maintained in France. Passage history, genotype and phenotype data will be included with these descriptions. Where available, clinical data on the infection and/or disease symptoms in the original host will be provided along with the strain description.

All original sequence reads will be deposited to the NCBI's Archive sequence databases.

All original sequence reads will be deposited to the NCBI's Archive sequence databases. Whole genome sequence assemblies and corresponding annotations will also be deposited in GenBank NCBI, listing JCVI and EuPathDB as joint owners of the records in accordance with NIAID data release policies. *Toxoplasma* genomes and annotation will be made available to the research community via ToxoDB, a component of the NIAID supported EuPathDB Bioinformatics Resource Center, under the supervision of Dr. Brian Brunk and David Roos, both members of the Organizing Committee.

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:

Investigator Name: **L. David Sibley** **Date: 28 September 2009**

Blank Last Page